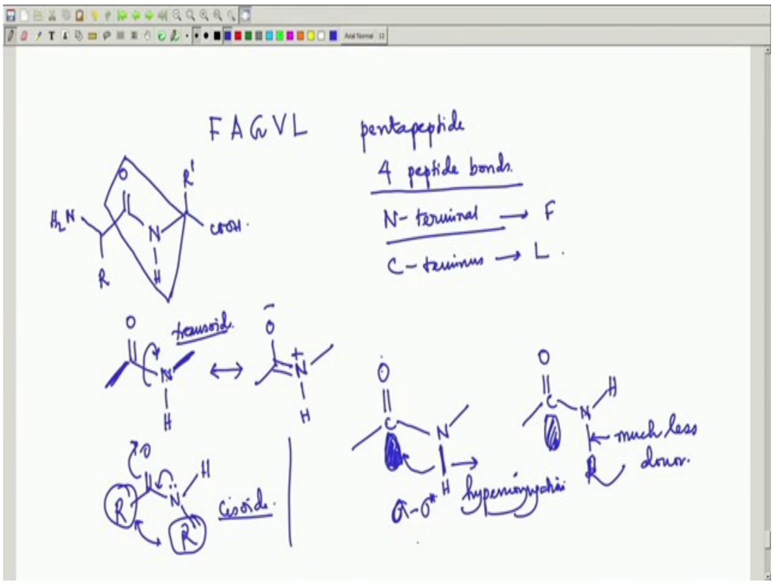
Organic Chemistry In Biology And Drug Development
Prof. Amit Basak
Department of Chemistry
Indian Institute of Technology, Kharagpur

Lecture - 05
Method of Determination of Amino Acid Sequence: Primary Structure of Polypeptide/Protein

In the last session, we have discussed two important issues in the protein chemistry: 1) Formation of proteins, which deals with how amino acids can be combined to make peptide bonds. 2) For a given protein, how can we know the sequence of amino acids that are linked one after another.First, I will talk about the sequencing; sequencing means what are the amino acids and the sequence in which they are connected.

(Refer Slide Time: 01:26)



Now, I told you how we write a peptide. Suppose I write FAGVL. Now, what does it mean? F is phenylalanine, A is alanine, G is glycine, V is valine, L is leucine. So, this is a pentapeptide. How many peptide bonds it has? 4 peptide bonds; which is the N terminal amino acid? It is F, that means, phenylalanine; Which is the C-terminus? That is L. Now, what is the nature of this peptide bond? Peptide bond is nothing but an amide bond NHCO. So, you have NH and a CO bond. When do we call an amide bond, a peptide bond? When it is between two amino acids (that is the only thing). So, suppose there are two amino acids, Thus it is a dipeptide.. So, this one is the peptide bond.

Now, you have always noticed that we write the peptide bond in such a way that the carbonyl is anti to the hydrogen attached to the nitrogen. The question is why is that? And secondly, if I write in this fashion, I could see a single bond between the nitrogen and the carbonyl carbon. The question is that, what is the real character of this bond? Because if it is a single bond then there will be free rotation between this (carbon) carbonyl carbon and the nitrogen; so these are some of the issues.

First issue is that why this is assuming this type of form (the NH and the carbonyl anti); Second issue is that, whether there is any rotation between this carbon and the nitrogen. The answer to the second question is that this carbon nitrogen bond has sufficient double bond character because of the existence of the resonance that is taking place here..

What is resonance? Resonance is nothing but the rearrangement of the positioning of the electrons. It follows certain rules like the octet rule and the charge concept etc.. But in this case, what happens is that this is the neutral form and this is the resonating structure, this is another canonical form. So, these are the two canonicals of the peptide bond between amino acids. Now, if that is the structure, then there is sufficient double bond character between this carbon and nitrogen, so that explains that this carbon nitrogen bond is actually quite rigid and consequently very little rotation is allowed between this carbon and nitrogen, so that answers the first question.

The second point is that why this is in the anti-form? As I can also draw the other form where they are in syn-form. The syn-form does not hinder the resonance that was earlier operating here. Because this lone pair is perpendicular to this nitrogen. And this is the p-orbital of the carbon, this is the p-orbital of the oxygen. So, it does not matter whether the hydrogen attached to the nitrogen is anti or syn to the carbonyl the resonance is still there. So, resonance is not affected.

Suppose one substituent (R) is here which is syn to another substituent R (in the peptide backbone). As a result, there will be steric interaction between these two, because they are in the cisoid form. Now, this is what is called (where this carbon-carbon bond are anti to each other) the transoid form; this is not really cis and trans because you do not have a formal double bond between the carbon and the nitrogen, so that is why it is called transoid and this syn form is called the cisoid form.

Firstly, the cisoid form is less stable because of the steric steric problems. Secondly there was another interesting argument that also can be placed in favor of this transoid form. Some people question that how big is the steric effect; because for all amino acids this steric factor cannot be same because R differs for different amino acids. But still in most of the amino acids, the NH is still assuming this transoid form. So, there must be some other factor which is giving stability to the transoid form. And what is that? In the carbonyl, there is one π bond and one σ bond. So, in the σ bond, the sp^3, sp^2 hybridized carbon orbital of the carbon is combining with the sp^2 hybridized orbital of the oxygen. Now, when two orbitals approach each other, there are two scenarios one is the bonding scenario, another is the antibonding scenario.

σ, σσ bond is in the plane of the paper by a plane of this screen. But the π bond is actually forming between the orbitals which are perpendicular to the plane of this screen.

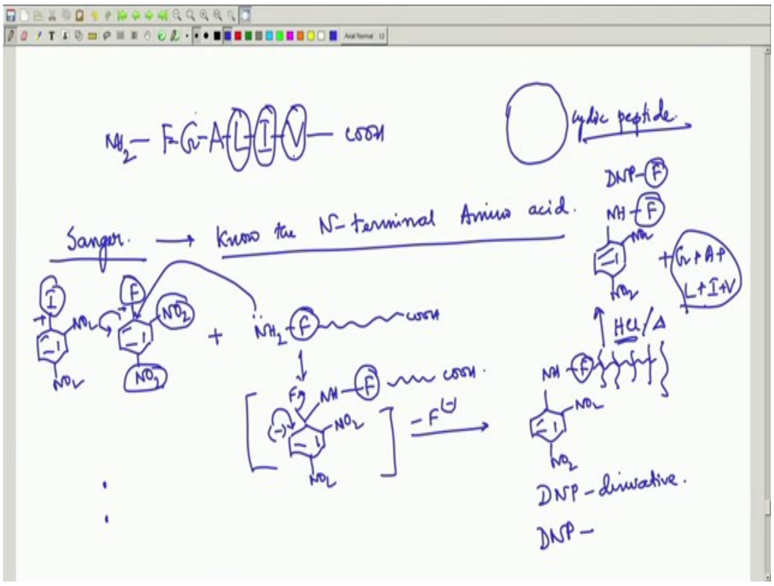
Now, σ bond is in the plane. So, when there is antibonding scenario between the carbon and the oxygen for the formation of the σ bond, the antibonding will have a bigger lobe on this side and which is empty. This is because the two electrons are occupying the bonding scenario. Now, this anti bonding orbital is empty and it has got a bigger lobe on the backside and that is in the plane of the screen.

And now if we write the nitrogen along with the hydrogen, we will see a beautiful correlation that this nitrogen-hydrogen (bonding orbital) is now interacting with the vacant antibonding orbital of this carbonyl-oxygen σ bond. If you draw the cisoid form, you do not get this, because this electronic donation (a kind of hyperconjugation) is effective for nitrogen-hydrogen σ bond only. This hyper conjugation is absent when H is replaced by R now.

So, this has got much less donor power. This also has this vacant orbital and this is the nitrogen carbon bond, but this is a much less donor; whereas hydrogen can donate much better. σσ

So, these are the two reasons. Yes, we are not discounting this steric factor. So, this steric factor as well as this additional electronic stability due to this donation of the N-H σσ to the σ^* of the carbonyl σ bond. Thus $\sigma^*\sigma$ of the carbonyl stabilizes this amide. So, the peptide bond remains in the transoid form. That is all about the peptide bond. Now, the question is that when I have a peptide, how do I know the sequence of amino acids in an unknown peptide? How do I know that which amino acid is linked to which amino acid?.

(Refer Slide Time: 10:49)



So, let us again take a peptide similar to the one I wrote earlier there, let’s say F G A L I V. From now on, we will try to write it in only letters. Phenylalanine, glycine, alanine, leucine, isoleucine, valine. V has got the carboxy free and F has got the NH_2 free. Now, suppose this is not known to me and I want to know the sequence. Now, the history of development of the sequence technology goes like this. We must know the chronological development of these biological sciences. It all started as I said with Wohler’s urea synthesis which first broke down the concept that was a remarkable achievement that organic compounds can be obtained from inorganic materials. Wohler proposed his synthesis of urea from ammonium cyanate.

So, in this case of peptide chemistry, people were faced with this challenge that how to know the amino acid sequence. Sequence refers to the primary structure of proteins, because proteins have different hierarchical forms of structures: Quaternary structure, tertiary structure, secondary structure and primary structure. We will come to this secondary, tertiary and quaternary later on, but now we will talk about the primary structure. Primary structure means only the formation of this peptide bond; primary structure is obtained by the combination of amino acids to form a peptide bond.

So, if you want to know the primary structure, you have to know the sequence of the amino acids, you have to know how they are linked together. So, the scientists faced this big challenge, how to know this.. Why this was a challenge? The challenge was basically there because all these amino acids are linked through the same functionality that is the peptide bond.

So, if you want to really know one after another, that which amino acids are linked, you have to break selectively one of the peptide bond and then keep the remaining intact. And then from the remaining, you break again another one and get whatever is the next one. So that was the challenge; it was very difficult because all the bonds are same as they are all amide bonds.

So, how to selectively cleave these bonds to know the amino acid sequence? The first development took place when the scientist Frederick Sanger from Cambridge University in England (who won the Nobel Prize twice), developed a method which was meant only to know to know the N-terminal amino.

That means, if I use Sanger's method, I only get to know that what is the first amino acid from the N-terminus end; that means, from the left side. But this method is not applicable to know the other amino acids. So, that is the drawback, but anyway this was a significant achievement at that point that what is the first amino acid to begin with. What is Sanger's method? He basically used a reagent which is called 2, 4-dinitro fluorobenzene. And he reacted this reagent with whatever peptide (which ends up with the carboxy and starts with NH_2) he had. . Now, what happens with this 2, 4-dinitro fluorobenzene? We know that in aromatic chemistry, usually electrophilic substitution takes place as it is facile. But nucleophilic substitution can also happen if there are very good electron withdrawing groups present in the aromatic system ok.

And provided you have a good leaving group also, because you have to also kick out a group from the system. Now, in this case, the two electron withdrawing groups are these two nitro groups, and the leaving group is the fluoride. So, the reaction that takes place is basically the NH_2 attacking this carbon bearing the fluorine to form the negatively charged σ-σcomplex. And you can always realize why you need those electron withdrawing groups; that is because this negatively charged σ-complex is now stabilized by the nitro.

But this negatively charged σ-complex is not very stable, this is the intermediate; so it regains aromaticity by now expelling the fluoride anion.

And what you get is basically a derivative (a 2, 4-dinitro phenyl derivative) of the first amino acid that is present in the peptide; this is what is called DNP derivative, Organic chemists have a DNP derivative that is dinitro phenyl hydrazine derivative, but this is different; this DNP is dinitro phenyl derivative. So, in this case you will a phenyl alanine linked to this dinitro phenyl moiety.

Now, what you do? Now, you hydrolyze it with acid; amide bonds can be hydrolyzed by base or by acid. In Sanger's method, you treated it with HCl, and then heat it. So, what will happen, all the peptide bonds that are present here, they will all be broken. And ultimately you will end up with NH that is attached to phenylalanine, plus all other amino acids which were there: G, A, L, I, V. So, all these are free amino acids and only one amino acid is attached to the dinitro phenyl group. So, now, this will be called DNP phenylalanine. Now, this is a very good UV chromophore, it has got a color. And you can detect it very easily by doing HPLC. If you do HPLC, and then you see where the peak is coming from the retention time, then you can identify the amino acid present in the N-terminus. It has got a very good chromophore. So, you can have a UV detector and you can immediately visualize it.

You can know all the 20 amino acids since dinitrophenyl derivative of all the 20 amino acids will have different retention times in the HPLC column and then from there, you can match that with the literature reported values. If you see that it is matching with phenylalanine, then you can know that the first amino acid is phenyl alanine. So, this is the Sanger's method of knowing N-terminus amino acid.

Now there are two questions that come here; firstly why it was used? Why it was considered to be useful? Because knowing only the first amino acid does not reveal the whole sequence. You do not get the sequence of the other amino acids because you have already broken all the peptide bonds. If could you have kept the rest of the peptide chain intact, then you could have again applied the Sanger's reaction and got the second one, but that is not possible.

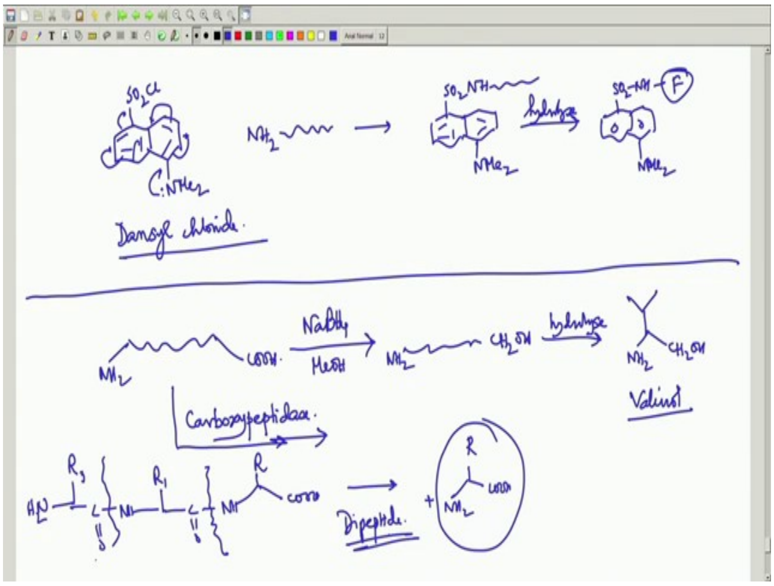
Sanger's method is basically cleaving all the peptide bonds in one shot, but you know only the first amino acid. You know why it is important? It is important when a peptide is cyclic or if the amine is somehow protected with something else. There are two cases: If the peptide is cyclic, remember for a cyclic peptide, there is no N-terminus or there is no C-terminus. A cyclic peptide does not have N-terminus or C-terminus. So, Sanger's method can immediately tell, whether a peptide is cyclic or it has got a nitrogen, which is blocked, which does not react with this dinitro fluorobenzene; so that is why it is very important.

So, it is still a milestone in the progress towards the primary structure determination of peptide. And always you know that after Sanger's method was discovered, people started thinking about how to cleave the amide bond selectively one after another. Till Sanger developed this method, it was very difficult to rationalize or to realize that it is a matter of challenge to break one amide bond selectively and thus improve the Sanger's method.

Before going into that there is another second question. The second question is that why did Sanger took the fluoro 2,4-dinitrobenzene and not the corresponding iodo compound? Because people might think, we all know that C-I bond is much weaker than C-F bond. So, we might think that it would have been better to take the iodo derivative of 2,4-dinitro fluorobenzene, but that is not the case. Sanger took the flouro not the iodo, chloro, or the bromo derivative. And what is the reason for that, the answer which an organic chemist must be knowing is that for this reaction there are two steps: one is first attack by the nucleophile and the second is the regeneration of the aromaticity. And between these two, the first step is the rate determining step and not the second one where the C-F bond is broken. So, it does not matter, whether it is F for I.

Second step is not the rate determining step, so bond energy does not play any role. Because of the presence of the fluorine, this carbon is more electrophelic or you can say because of the presence of the fluorine, this anion is more stabilized by the electron withdrawing effect of the fluorine, so that is why Sanger, even after not being an organic chemist, could still realize that which compound has to be taken and accordingly he took the fluoro compound and not the iodo.

(Refer Slide Time: 23:34)



The next method for N-terminus amino acid identification is the similar to the Sanger's method. This method involves a reagent called dansyl chloride. What is dansyl chloride? Dansyl chloride is 5-Naphthalene-1-sulfonyl chloride. This is a very good chromophore because you see the kind of resonance it can have. So, here you have SO_2Cl, so what will happen? The N-terminus, will react with the dansyl chloride and it will form SO_2NH and then the peptide; and you have the dansyl system. It is a naphthalene system with NMe_2.

And now you hydrolyze like Sanger's method; hydrolysis will get you SO_2NH and the first amino acid that is the phenyl alanine. This is very similar to the Sanger's method, but one can say that why then people wanted to devise another method similar to Sanger? The answer is that, what is the limitation of detection of N terminus amino acid? See the stronger the chromophore, the greater will be the limit of detection, so that is why actually this dansyl system is a better chromophore.

So, if you have very tiny amount of the peptide then it is better to use dansyl chloride if your aim is only to know that N-terminus amino acid and you can get the dansyl derivative, which is a stronger chromophore than the one which was used by Sanger. So that is for N-terminus amino acid. Now, at the same time, people started thinking that what about the C-terminus? So, is there any method for C-terminus? Now, again I write the peptide like this and there is a free CO_2H here. The principle behind identifying the amino acid is that the amino acid that you want to know whether, it is present or not, you have to put a tag to that amino acid like 2,4-dinitro fluoro benzene (as a tag to the amino acid) or dansyl.

Dansyl becomes a tag to the N-terminus amino acid. Similarly, for the C-terminus amino acid, you have to do something to put a tag or do some reaction making it different from the other amino acids. So, the first one that was tried was sodium borohydride in methanol.

What does sodium borohydride do? It actually reduces carboxylic acid into alcohol. You cannot use lithium aluminium hydride because lithium aluminium hydride will also reduce the amide bonds, but sodium borohydride in methanol selectively reduces only the carboxy group. So, the C-terminus amino acid will become an alcohol. And the remaining amide bonds will stay intact because sodium borohydride cannot reduce the amide bonds.

So, now, if you hydrolyze, what you will get? You will get the last amino acid. In our example, the last one was valine. So, what will I get? I will not get valine, but after hydrolysis I will get valinol, because that will be CH_2OH. So, my duty is to just check that which of the amino acid is converted into the alcohol form, so that will be the C-terminus amino acid.

But this process is a little complicated as you are not producing any chromophoric system, if you are not producing chromophoric system then isolation of this is very difficult. If the chromophoric system is not present, then the detection limit is very is very low. Detection limit is not satisfactory. So, then people changed the method. What people did, they took this peptide and they took help of not chemical reagents, but bio-chemical reagents like the enzymes.

Now, there is an enzyme which is called carboxy-peptidase. We will talk about carboxy-peptidase later on; it is a metal dependent enzyme. There is a mechanism of hydrolysis;what it does? Hydrolyzes, the peptide bond at the C-terminus end; that means, the C-terminus end has a CO_2H at the end. And then you have an R here and before that what you have NH and then you have a CO and then you have another amino acid R^1 and then you have again NH and you have CO and then you have R^3. Suppose, a tripeptide and it ends up with NH_2. So, this is the tripeptide.

So, what carboxypeptidase does, it hydrolyzes the terminal peptide bond which is terminal from the C-terminus side. So, it will hydrolyze this one ok. So, this amino acid will come out R-CH(CO_2H)(NH_2) plus you get the dipeptide because we started with a tripeptide. Now, what is your task? You just see which amino acid has come out. And the amino acid that will come out will be your C-terminus amino acid.

Now, you can raise this question that why do not you repeat this? You have the dipeptide. Now, you have a method of actually selectively cleaving the peptide bond from the terminus side. So, you repeat it for dipeptides which in turn will be converted into the individual amino acids, if this bond is broken then that will also be converted. So, if you allow more time, so what you do? You do a time dynamics, you have a peptide you add carboxy peptidase and you start doing detection by HPLC and see which amino acid is coming out.

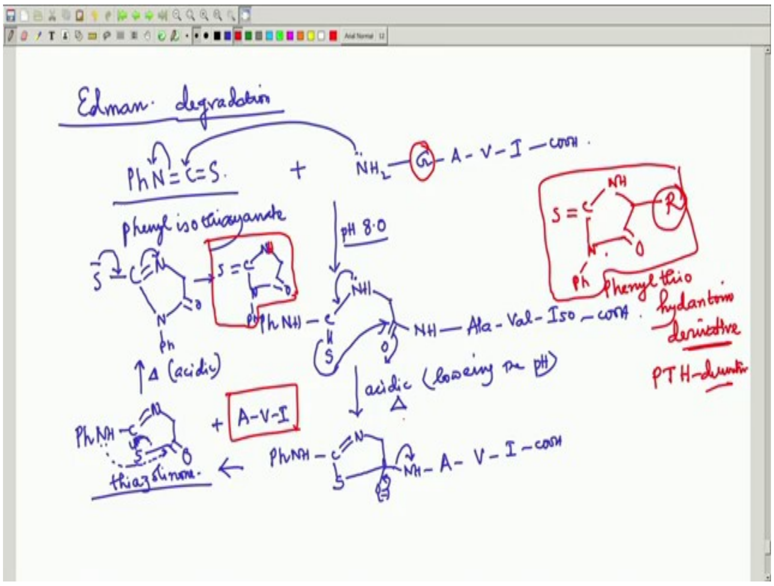
So, you see that valine has come out so; that means, valine is the first amino acid. After 5 minutes if you again inject, you see that another amino acid has started coming out. I am just going to that type page, which I showed our representative peptide F-G-A-L-I-V; so the first amino acid that will come is valine. And after some time, I will see isoleucine is started coming.

So, I can say that the second one from the C-terminus side is isoleucine and if I wait little further, I will say that the third one is also coming; that means, the leucine from the third from the carbon side. The third amino acid is leucine, but there is a limitation; after some time it becomes so complicated to do these dynamics; identifying the amino acids at the positions 1, 2, 3 are ok, but if you want to know 10 amino acids, that is not possible.

So, this carboxy peptidase works to some extent in identifying; the first one is definitely sure, but from the second, third onwards it was difficult. However it is a very beautiful way of knowing the C-terminus amino acid; at least the C-terminus and the penultimate one also ok. The principle is that it hydrolyzes selectively the peptide bond from the C-terminus only.

Like carboxypeptidase, we have an enzyme called amino peptidase which hydrolyzes the first peptide bond from the amine side. So that also can be used, but people have not used amino peptidase because already Sanger’s method is there which is quite useful for the same purpose. So, this are the two methods of determination of N-terminus and C-terminus amino acids. Now, let us go to the more challenging problem that how to know the, the entire sequence now, we know the N and C-terminus amino acids. But what about the middle ones, what is the method?

(Refer Slide Time: 33:40)



There is a scientist whose name is Edman. He developed a method which is called Edman degradation and that method was discovered more than 60 to 70 years back, but still that is the method which is used till today to determine the sequence of amino acids of proteins. How did he achieve that selective protection because he used a reagent which is called phenyl isothiocyanate PhNCS. Now, if you add phenyl isothiocyanate to any peptide(say, G, A, V, I)

What will happen? This reaction was done little at a higher pH. Say around pH 8; if you do this reaction with this peptide and isothiocyanate, then the NH_2 being nucleophilic, and this carbon being electrophilic (as it is flanked between two electronegative groups)

this nitrogen will react and this will take up the hydrogen. As a result, you will get this PhNH and then C double bonded to S; then you have NH. This NH is from the amino acid N-terminus and then you have glycine. I have said glycine. So, it takes only G and then you have alanine. So, there is a peptide bond between these two.

So, you have to write the peptide bond, because now we are entering the selective cleavage of the peptide bonds. So, this is your first peptide bondNow you have to write alanine and valine followed by isoleucine and then the carboxylic acid. So, then you lower the pH. So, you make it acidic lowering the pH and slightly heat the solution.

Now, a reaction takes place. What is that reaction? This nitrogen lone pair comes here there (it is an intramolecular reaction that takes place); this sulphur becomes more negative and the lone pair flies here; the sulphur becomes more negative and you know that sulphur is a very good nucleophile, because it is less electronegative compared to nitrogen or oxygen. So, now, it comes and intramolecularly attacks this carbonyl.

So, this carbonyl then withdraws the electron towards itself. So, what you get is PhNH then C double bond N; the nitrogen will definitely lose the hydrogen. This is now O minus, this is your O minus and there is a sulphur carbon bond. Now, what happens is this O minus comes back and this peptide bond is broken, this amine leaves.

So, now you have a tripeptide (A-V-I) that leaves. So, along with the tripeptide, the other compound fromed from the first amino acid is PhNH then C double bond N then you have S and this is CO. This is what is called thiazolinone derivative; but this is not very stable. Now, it rearranges, on further heating under the acidic condition. It rearranges as nitrogen attacks the carbonyl and there is a rearrangement and the sulphur comes out of the ring.

So, basically this nitrogen attacks this carbonyl and this carbon sulphur bond is broken that comes here. So, what you get is C double bond N then CO then you have NPh (that will lose the hydrogen) and then you have S minus. This nitrogen is attacking here and this is coming to the sulphur. So, S becomes negatively charged but this is not very stable, the S minus will now donate electron back to the carbon and the nitrogen becomes N minus and it picks up the hydrogen.

In the final product, C double bond S, then NH then this actually comes from the glycine part. Remember this NCH_2CO comes from the glycine part; then you have CO and then NPh and that one. If it is not glycine, suppose any other amino acid, then you have a substituent here depending on the different type of amino acids here. This final product is called phenyl thiohydantoin derivative; or in abbreviation PTH derivative.

This PTH derivative will have different retention time in the column depending on the nature of R. The beauty of this is that the first amino acid you can locate by doing the

HPLC; at the same time you are separating, when you do the HPLC, you are separating this tripeptide. So, now, you take the tripeptide and repeat the process again. So, if you repeat the process, next hydantoin that will come out will belong to alanine and you get the dipeptide which is valine and isoleucine. And then again you repeat, so you take the valine out and you the only one that is remaining is isoleucine. So, in this way, you can tell the entire sequence of amino acids in a protein.

However, every answer of a problem gives rise to other questions. And this also is very similar to that. The question is how many I can do in one shot? Suppose, I have a peptide containing hundred amino acids and the question is how many I can do, I can recycle these peptides doing Edman degradation, Edman degradations can be recycled at a time? The answer is about 20 to 30 amino acids, you can detect, still depending on the amount of peptide that we have started. But even if you have sufficient peptide, even then your detectable limit is that you can detect up to 20 to 30 amino acids.

Then what are you going to do if it is a 100 amino acid containing peptide? And other problem is that if you have a very big protein, suppose 1000 amino acids and if it is folded like this, of course, you can denature it you can unfold it, but the problem is if it is a very big molecule, then bringing this NH_2 and the final isothiocyanate close together, will be also a very big challenge since a remotely located NH_2 group that has to react with phenyl isothiocyanate. So, that is also very challenging which is entropically not favoured at all ok. So, they have you have to bring these two.

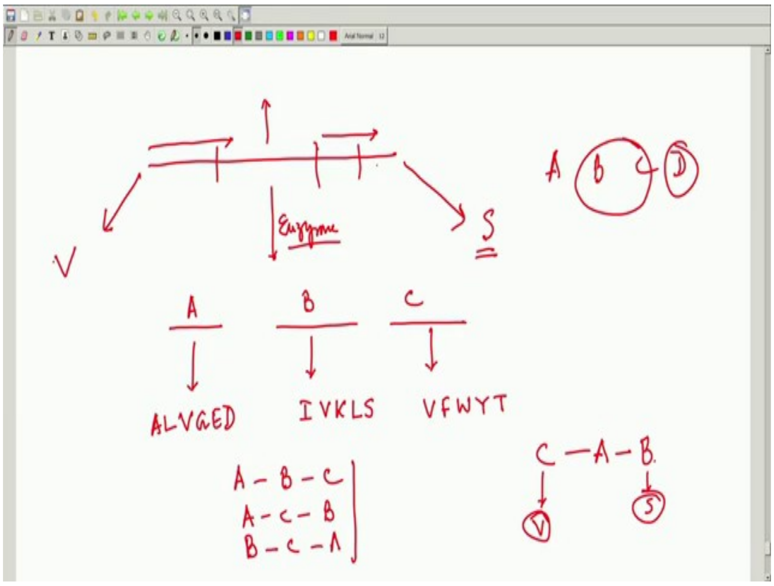
So, there are the two challenges. How to solve these two problems? In one shot, you can solve these problems. The trick is that if you have a big protein, you cut the protein into small pieces. How to do that? You do a digestion with HCl, all the proteins with peptide bonds will be hydrolyzed, so that is not a solution. If you hydrolyze with alkali, again all the peptide bonds will hydrolyze.

So, there must be something and here again nature offers the help. Nature provides us with enzymes having tailor made specificities. There are enzymes which will cleave the peptide bond only when certain amino acids are present ther;e like trypsin, chymotrypsin, pepsin.

So, what trypsin does? If there is a peptide bond involving a basic amino acid like lysine or arginine, then it hydrolyzes that peptide bond. So, now suppose you have a 100 amino acid containing peptide, where there are three lysines. So, there will be three cleavages because of the three lysines. So, you get three peptide fragments, suppose each peptide fragment is about say 1 is 40, another is 20, another is 40, so then you can do the Edman degradation and try to determine the sequence.

So, different enzymes have different selectivities. And using enzymes, you can cut it where the enzyme acts like an artificial scissor; thus the enzymes are called artificial scissor. You cut the peptide bond at different places to get smaller peptides. And smaller peptides are isolated and then subjected to the Edman degradation.

(Refer Slide Time: 46:38)



However, there is still one more problem. Suppose, this is your peptide a long peptide. Now, you have an enzyme. You cut it into pieces. So, these are the pieces. Suppose, this is piece A, this is piece B, this is piece C. Now, you know the sequence suppose you know the sequence of this be say A L V G E (something arbitrary). You know the sequence of this. Suppose, this is I V K L S and the sequence of this is say V F W Y T. So, these are the sequences from Edman degradation.

Now, the issue is now you have to join these three and then you should know what your primary structure of the peptide is. But the problem is now you do not know that whether A is connected to B and B is connected to C or A is connected to C followed by B or it is B is first, then C, then A.

So, you see lot of combinations is possible now. So, you do not know which one comes first. So, now, you see the success of the Sanger’s method. See, if you first have determined what is the terminus amino acid, then you know which one is the first one whether it is A or B or C.

And if you really know what is the C-terminus from the very beginning and then do this peptide chopping and do the Edman degradation, then if you know this two suppose it says that S is the last amino acid (at the C-terminus) and suppose this says that V is the first amino acid (at the N-terminus), then you know that if V is the first amino acid, so it has to be C followed by A followed by B; because in B, -S is the last amino acid and in C, - V is the first amino acid. So, in this way, you can do it.

But if the fragments are more, then there is a problem; if you have four fragments say A, B, C, D, then the challenges are more. Even if you know the A-terminus and C-terminus, you do not know whether B and C are connected by B-C or C-B. How to know that? And in real scenario you have say 40 fragments, how to do that? So, there are different techniques. Then you have to use another enzyme which cuts at different points. And then there is the way that you can get overlapping regions, from overlapping regions you can again reconstruct it.

And finally, form the actual sequence of the peptide. And that has led to the formation of the branch of bioinformatics. Bioinformatics tools will arrange by looking at the overlapping regions and reveal the sequence., It will join the fragments and then tell that this is the sequence of the amino acids.

So, that is for today. So, what we have learned today is that how to determine the primary sequence of amino acids.

Thank you.