Essentials of Biomolecules: Nucleic Acids, Peptides and Carbohydrates

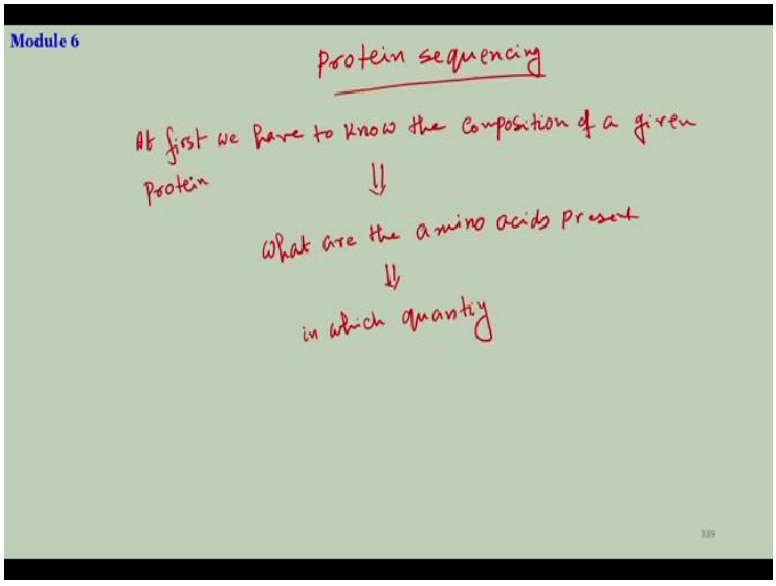
Prof. Dr. Lal Mohan Kundu
Department of Chemistry
Indian Institute of Technology-Guwahati

Lecture-24
Protein Sequencing and Solid Phase Peptide Synthesis (SPPS)

Hello, everybody, and welcome back to the lecture. So today we will start a new module, module 6. So, after an extensive amount of biology and biological processes of how proteins are synthesized in biological cells, we are now back again with chemical tools, how chemistry can be used in order to study or in order to know understand some of the biological molecules, biomolecules. So in this module we will talk about protein sequencing and solid phase peptide synthesis.

So this module is actually a combination of 2 topics. One is the protein sequencing. And the second one is how can you synthesize short peptide chains or short peptide molecules in your laboratory, and both of these techniques would involve chemical tools, we will use the knowledge of chemistry or the methods of chemical reactions in order to study these 2 subjects. So, let us start first with the protein sequencing.

(Refer Slide Time: 01:56)



So sequence of a protein means which amino acids are connected together in a row. So that you will understand the composition of protein as well as the specific linear chain of the or the

primary structure of the protein. Of course, this is a very important aspect to know, because that is how you can understand or you can kind of predict what kind of function of certain protein will assume.

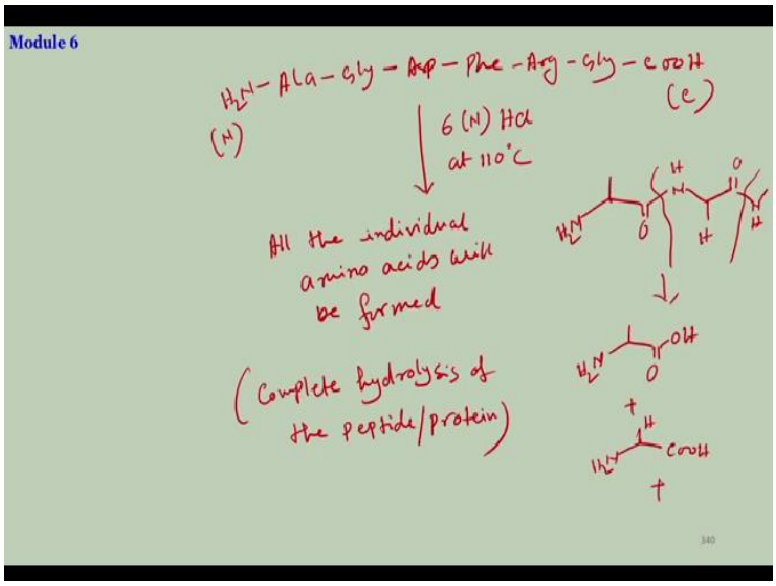
We have so many enzymes in our biological systems, both eukaryotic cells and prokaryotic cells. And all those biological processes are catalyzed by enzymes which are basically the proteins. So, it is of course, very important to understand or very important to find out the sequence of a protein. So, there are kind of several techniques are 3 to 4 different techniques that can be adapted for protein sequencing.

So, before we have seen that, how to do the DNA sequencing, we have seen the Sanger method, we have seen the Maxam Gilbert method that were being developed to know the sequence of a given DNA or gene. And DNA sequencing was kind of pretty much straightforward. So, in Sanger sequencing, or also the Maxam Gilbert method, we have seen that if you have if you just read the gel, if you see that image of the gel, you can pretty much find out or you can pretty much write down the sequence of the DNA or the gene.

Protein sequencing methods are not that straightforward and it involves a combination of methods. So, before going to the actual sequencing of the protein, what most people do is to find out first, what is the composition of the protein, what amino acids are involved and in how many quantity. So, at first we need to know the composition of the protein. We have to know the composition of a given protein or it can be protein or enzyme.

It is the same thing basically. So this will tell you this means what are the amino acids present in that particular protein sample and in which quantity, how many of them are there. So, that is called the composition of the protein, because protein is a very long molecule. So, before going to the sequencing, you need to have some ideas about them that makes the process easier. So, how do we do it. Usually, this is done by complete hydrolysis of the protein or complete hydrolysis of the peptide sequence.

(Refer Slide Time: 06:15)



For example, if we take to write down a sequence of a protein, alanine, glycine, aspartic acid, phenylalanine, arginine, glysine, this arbitrary sequence, this is a short peptide actually I have not drawn the whole protein structure or whole protein sequence. So, this is your N-terminal which means it has a free amine group N-terminal and this has a free carboxylic group which is C terminal. Of course they can exist in Zwitterionic form.

Now competition of a protein means to know this is a sequence I have written, but if you isolate a sample you do not know what is the sequence. So, composition of a protein means just to find out which amino acids are present. So, to do that, what is done is you treat the protein with a 6 normal hydrochloric acid, 6 normal HCl and heat at around 110 degrees Celsius is a pretty rigorous condition actually.

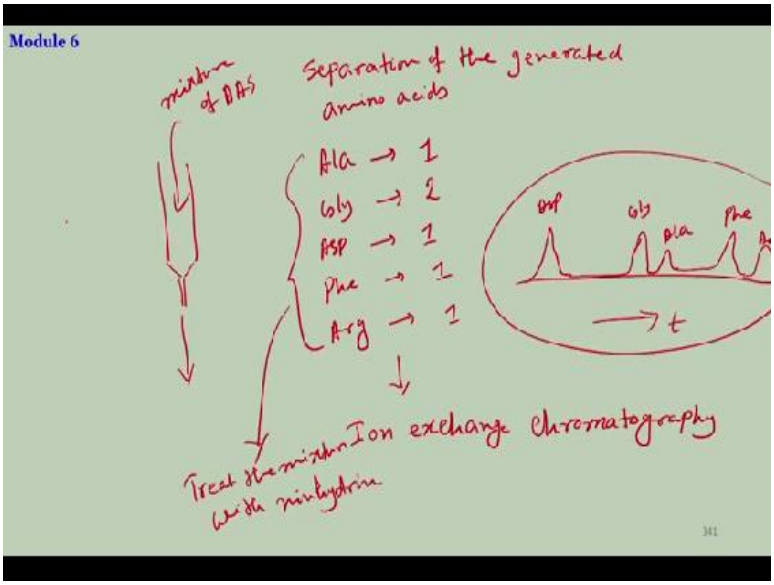
So, what does it do. It breaks all the peptide bonds, it breaks all the amide bonds present in the protein. So, if you have if I just write the alanine I am not drawing the stereochemistry if this is your N-terminal, this would be CO and this is your glycine with H here N H CO and this will be your aspartic acid and it will go on. I have missed something right CO here I am sorry, CO N H this is your glycine CO N H and so on.

So if you treat these with 6 normal hydrochloric acid at high heat, then it will cleave all the peptide bonds and what you will get is individual amino acids this last a glycine plus all the

others. So, after treatment with 6 normal hydrochloric acid at 110 degrees Celsius, what you will get is individual amino acids, it will break all the peptide bonds. So, that is what is done to understand the composition.

So, if we treat this with this then all the individual amino acids will be formed So, this is known as complete hydrolysis of the peptide or protein. So, once you have all the amino acids then what we do is you try to separate them individually.

(Refer Slide Time: 10:26)



So, next day is the separation of the generated amino acids. So, in this sequence, what would you get, what you are expected to get, you should get alanine in 1 mole quantity, or 1 equivalent quantity, you should get glycine individually 2 of them, if you look at the sequence here, you have 1 glycine here, you have 1 glycine there. So, glycine should be of 2 equivalent and then you have aspartic acid 1, phenylalanine 1 and arginine 1.

Aspartic acid 1, phenylalanine 1 and arginine that is also 1 equivalent. So, this would be the competition of the amino acid present in the mixture, but you still do not know their identity because they are still in the mixture. Now, what you do is you try to separate these individual amino acids using chromatographic technique. So usually a chromatography means it is a chromatographic technique is a separation technique where you can separate individual molecules.

If you have a mixture of number of molecules, then you can separate the molecules into the individual ones using the chromatographic technique. Now, there are a variety of chromatographic techniques a separation techniques available. So, I will talk more about those in detail when we will be in talking about the modern techniques in that module. So, roughly, there are many chromatographic techniques based on the separations can be based on the property physical properties of the molecules that are present in your mixture.

So those properties can be based on the polarity of the molecules. So if you have a mixture of a non polar compound and a polar compound, then you can separate them using a chromatography based on their polarity difference. Similarly, you can separate the molecules based on their charge or charge density. As we have seen in the electrophoretic diagram that DNA shorter DNA with less negative charge, longer DNA with more negative charge can be separated using the gel electrophoresis.

So, in chromatography also those kinds of techniques are there where you can separate the molecules based on their charges. So here in this case, and all of course on their pH, based on their pH also, because some molecules will be eluted out at certain pH level. That is another way of separating the molecules using the chromatography. We can also separate molecules based on their size which is known as size exclusion chromatography.

If you have a small molecule and if you have a very large molecule, obviously their sizes are very different and you can separate these 2 molecules based on their size difference using a certain chromatographic technique. So, in this case, what we use is ion exchange chromatography. Ion exchange chromatography is a separation technique, which separates molecules based on their charge and the pH.

So, that is why it is called the ion exchange. So the ions basically. So, if you do this, then you can make, you can have versus time, then you can have the individual amino acids appearing at different time interval. So, for example, I am not going details into the methods how the

chromatography is done. So, it can be based on the change of ionic value, you can do based on the pH difference.

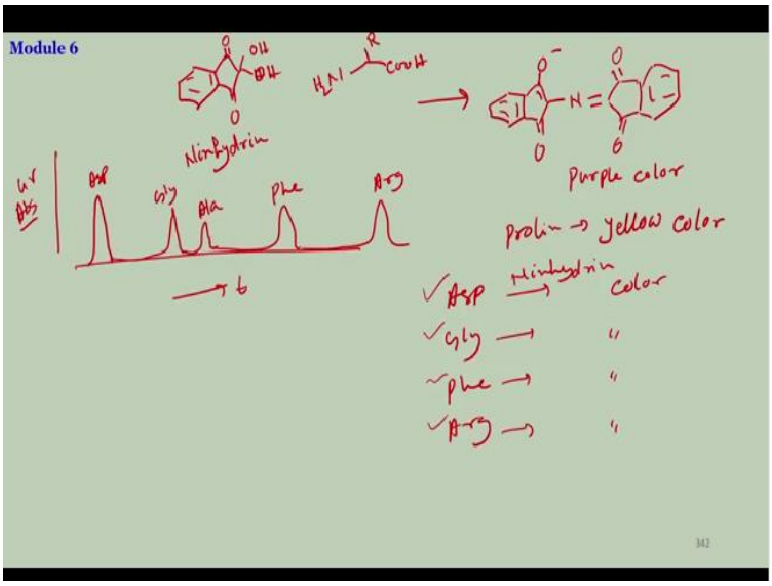
So, ideally what you can remember is that the time. So, if you are running your mixture of compounds in a chromatography column. Here is your mixture of the amino acids individual AAS and then if your column is packed with the ion exchange material, then based on their ions or ionic charge or pH this molecules will come out at the end of the column at different time. So, if you draw a time versus the appearing profile, it will look something like this.

Here you will see 1 peak. So, this would be might be for aspartic acid, this would be for glycine, this would be for alanine, glycine and alanine are not very different only one methyl group. So, they will be coming somewhat close, phenylalanine and here a little bit further would be your arginine. So, at different time they will elute out. So, obviously there is a gap in between, so you can collect them that will give you the separated amino acids in individual purity.

So, that is one aspect, second is fine you have got your amino acids. Now, how to see them, all these amino acids do not have any color and they do not most of the amino acids do not absorb UV radiation also, UV light except the tryptophan and some of the others. So, most of these amino acids are not sensitive to UV radiation. So, you cannot see them. So, what you do is, you actually do a reaction here.

Once you have all the amino acids after the hydrolysis, you treat these, treat the mixture of amino acids with ninhydrin. So do you remember the structure of ninhydrin and we have used ninhydrin to find out the amino acids that if you have amino acid to determine or how to characterize the presence of an amino acid that was done by ninhydrin test right. So the same thing is done here.

(Refer Slide Time: 18:28)



I will draw again the structure of ninhydrin OH, so this is your ninhydrin. Ninhydrin if you treat with free amine group which is present in the amino acids, I am writing R in general and then you have the amino acid. Then if you go back and look at the mechanism of these reactions, what is the product finally you get is this 6 membered I guess it was something like this. So this one was your product and that has intense color.

Those intense purple color bright red except prolene which prolene gives yellow color orange to yellow color. So, basically, you get colored compounds, when you treat amino acids with ninhydrin you get the colored components. So, what is done is once you have this profile and separate them each of them I will write the profile again. So, you have separated aspartic acid. Now, you have all the separations there, you have glycine, you have phenylalanine, you have arginine and all this.

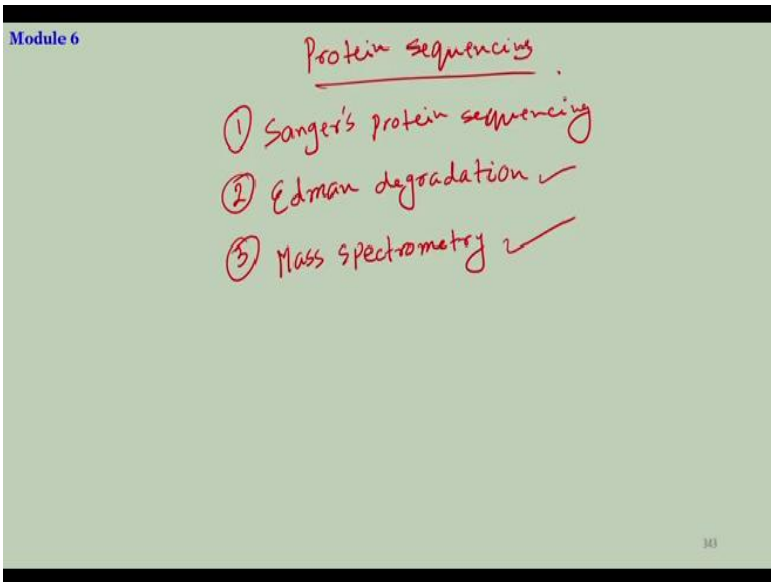
So, each of them are now treated with ninhydrin and you get a color. So, now by measuring this color of course is that will absorb UV light. So, you can see the UV absorbance and can quantify it. So, the y axis I was writing here, the intensity of this peak would now be visible when you have if you now measure y axis the UV absorbance ABS I am writing for absorbance, there you can see the peaks properly.

And their intensity, how much the intensity, what is the absorbance, how much is the absorbance of the individual amino acids, aspartic acid will be here and then you can see the glycine, then you can see the alanine, you can see the phenylalanine and you can see the arginine, Asp, glysine, alanine, phenylalanine and arginine. So, this is based on your time or pH variations. So, now, by measuring the intensity of the absorbance ideally it is the whole peak area calculation area of the peak.

If you calculate, then from there you can find out which amino acid is present in how much quantity and from the spectra, you can find out that glycine will have the 2 equivalent quantity or the intense peak area would be double compared to all the other ones because glycine are present in the whole peptide that I have written. So, that is how you can actually find out about the composition of the amino acids present in the peptide or in the protein.

And you can also know the quantity of the individual amino acids, that is the first step. Now, once you know that, after that you go for sequencing method.

(Refer Slide Time: 23:26)



So, now starts the real protein sequencing. So, as I mentioned, sequencing is about which amino acids are connected to each other or the primary structure of the protein. Now, protein is really big molecule, a large molecule contents 200, 300 amino acids in a protein chain and the methods that are developed, I will first describe the methods and then actually will, this thing I will talk

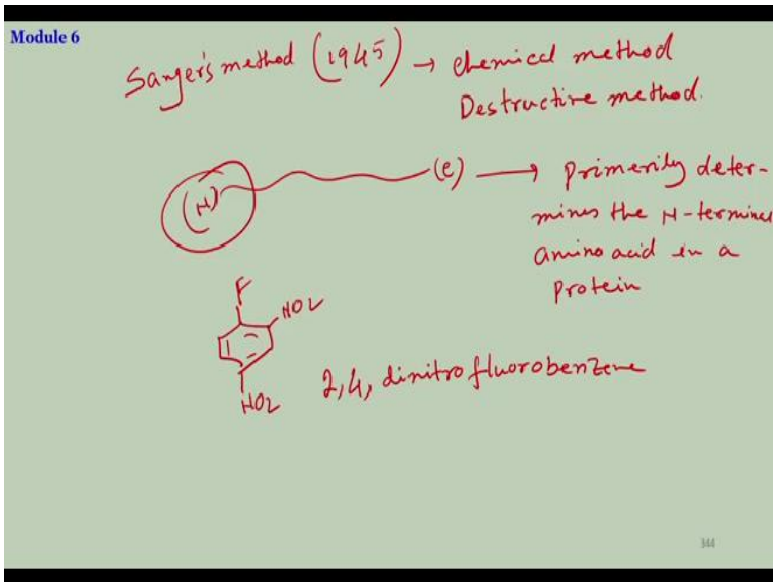
later that you have to actually cut the proteins into the shorter peptide versions to use them for the peptide sequencing or protein sequencing.

So, there are 3 major methods or chemical methods used to understand the sequencing of a protein. The first one is again, Professor Sangers came here for our rescue. The first one is called the Sanger's method. Sanger's protein sequencing. The second one is called Edman degradation and the third one is not a chemical tool, it is an analytical tool using the mass spectrometry. So, nowadays, Sanger's degradation method is the most widely used for the protein sequencing.

You also have the automated machine just like the PCR, just like DNA sequence I have talked about for protein sequencing also, there is automated machine pre programmed, but based on the chemistry of the Edman degradation. So, this is the most widely used method for protein sequencing. Nowadays, mass spectrometry is also used as the major technique to understand the sequence of a protein.

So, this is of course that using the high end mass spectrometry okay. So let us talk one by one Sanger's method.

(Refer Slide Time: 26:15)



The method was developed close to 1950, 1945 I think, long back. In fact, Sanger's protein sequencing method was one of his first work, this was done much before the DNA sequencing

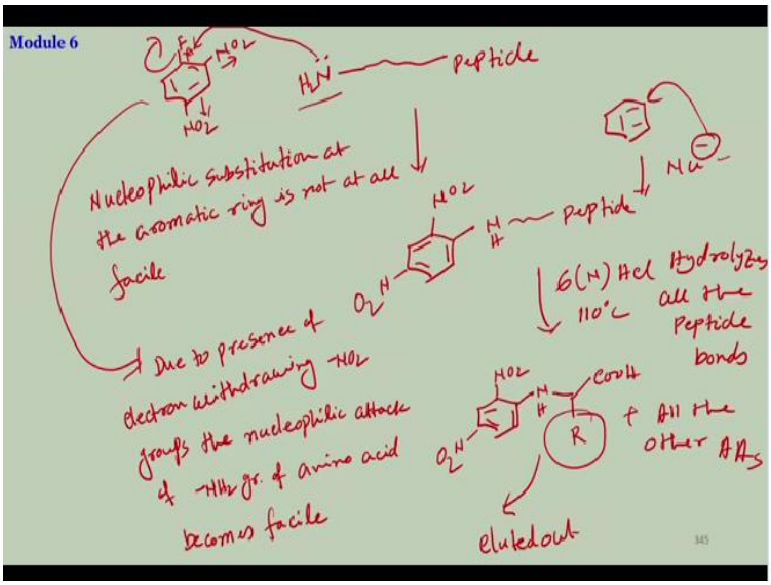
was developed. So, it is purely a chemical method. In DNA sequencing we had seen the Sanger's method was a constructive method, it was not touching your parent sequence of your of the gene.

But here the protein sequencing method is a destructive method, you do not get back your compound. It is a destructive method. So, you cannot rerun the system primarily, Sanger's method is developed to know the N-terminal sequence of a protein. So, if you have a protein which has the N-terminal this is the C terminal. So, Sanger's method primarily tells you primarily determines the N-terminal amino acid in a protein.

And this is the first method that was developed or who has made any attempt that was the first attempt made towards knowing the sequence of a protein where it was of course, at the beginning, it was only for understanding the N-terminal sequence of the protein. So, as I have said that it is a chemical technique. What he has used is a chemical molecule. This molecule fluoride, nitro, nitro. This is an aromatic compound.

The name is this is 1 2 3 4. So it is 2 4 dinitrofluorobenzene. This is the molecule Sanger has used to know the N-terminal amino acid. So how was it done.

(Refer Slide Time: 29:33)



I could draw the structure again. Why he has used this molecule. The molecule was used to label the amino acid, if you have a peptide or the protein so I am not drawing that, this is your protein

or peptide which have the N-terminal amino acid will have the free amine group. So what this will do is this is an nucleophile, so it is basically a nucleophile substitution reaction to the aromatic ring.

Now, obviously you know that the nucleophilic substitution to aromatic compounds is not very facile is not at all facile actually, the reason is if you have a nucleophile which means that it has excess electrons that would be attacking. Now, your aromatic ring also is rich in electron. So, there are 6 electrons, 6 pi electrons in the aromatic ring. So, it is already reached in electron. Therefore, reaction with the another nucleophilic center and another electron rich center will not be facile.

So, there is no reactivity here. What you need to create is an electron deficient center. So, usually in the aromatic ring nucleophilic substitution is not facile. So, if you have just benzene, if you try to react with a nucleophile with a distinct minus charge this reaction will not occur. Because this is electron rich, this is also electron rich, but here this molecule undergoes quite rapid nucleophilic substitution.

The reason is since you have the nitro group, nitro group are highly electron withdrawing groups. So, they will pull the electrons towards themselves that would create the electron deficiency in the aromatic ring. And therefore, you have also the fluoride which is also have minus I that pulls the electron density towards itself. So, this carbon here becomes electron deficient or overall the aromatic ring is also electron deficient.

That is why this nucleophilic substitution can occur. So, it is basically the amino would react amine would attack this carbon and fluoride would be eliminated, so nitro here is your amine and the rest of the peptide. So, for this given compound due to presence of electron withdrawing nitro groups the nucleophilic attack of the NH 2 group of amino acid which is obviously the N-terminal amino acid becomes facile.

That is why this reaction occurs, now once you do that once you get your compound. Now, how to know the identity of the N-terminal, you again go back to the previous method, 6 normal HCl

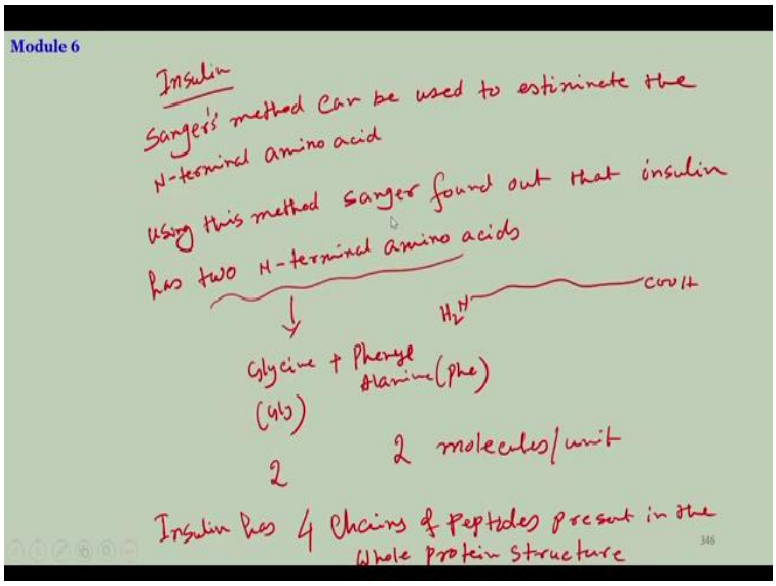
complete hydrolysis at 110 degree Celsius. So, if you treat this molecule with 6 normal HCl under heating condition, then it will hydrolyze all the peptide bonds. So, that will hydrolysis all the peptide bonds.

And you will be left with I should draw this properly. This amino acid plus all the other amino acids AA. All the other amino acids present in this peptide would be in their free form after the hydrolysis. The first one the N-terminal one would be attached to this 2 4 dinitro aromatic compound and this is colored. So, this can be separated out, eluted out. So, this will have a distinct appearance compared to all other present in the mixture.

And of course it can be easily separated also. So, once you separate this, you can know from your standard curve, if you have a standard that which amino acid will appear at which position if you have done individually with all 20 amino acids, they acted separately with this, then you will know exactly where in the elution that particular amino acids will appear. So, from that comparison, you can find out the nature of the R or you can find out what is the amino acid present in this molecule.

So, that is how you can determine the presence of the N-terminal amino acid. For all the other amino acids you cannot at this moment, you can know the composition of course, but you cannot know the sequence. Since the NH 2 was free at the N-terminal end, you can be certain that this molecule is your N-terminal amino acid. So that is how Sanger's method was used to find out about the N-terminal amino acid of a given protein. Now, the question is, is it important.

(Refer Slide Time: 37:24)



In that time, actually, it was very important and it was, as I have said, it was one of the first attempt ever made to find out a sequence of a protein using this technique. So, if you remember Sanger has solved the structure of insulin and found out the complete structure of insulin. This is one of the techniques that he has used. One of the prominent method that was used to understand the insulin structure.

So, if you have insulin, so what you know now is Sanger's method can be used to estimate the N-terminal amino acid, using this method Sanger found out that insulin has 2 N-terminal amino acids. So if you write a peptide sequence usually if it is a single chain that will have only 1 N-terminal amino acid which has a free amine group and that will have only a single carboxylic terminal.

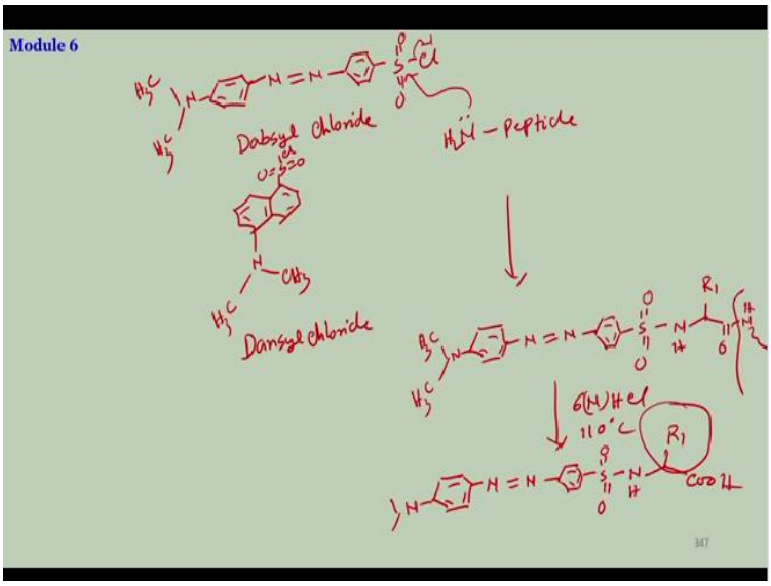
But in insulin what Sanger has found out that instead of getting 1 amino acid here attached with the aromatic ring, there are 2 components that were attached with the aromatic ring. And upon characterization, he came to know that these 2 amino acids are so these 2 N-terminal amino acids. Those were glycine and phenylalanine phe, glycine, and this is phe 2 components right there, which obviously means that the insulin has not a single chain of protein, it should have at least 2 protein chains combined together to form the whole protein.

Not only that, Sanger has also found out that if you estimate their quantity glycine would be was in 2 equivalent in number. phenylalanine was also 2 equivalent in number 2 molecules per unit. That means insulin has 4 N-terminal amino acids, 2 of this N-terminal or glycine, 2 of the N-terminal are phenylalanine which essentially means that insulin has 4 chains of peptides 4 long peptide chain or 4 protein chains present in the whole protein structure.

So, it has to be a combination of 4 subunits of protein combined together, that is insulin. And that was really a big discovery, big development in that time in understanding the structure of the insulin, just using this method using this chemical compound, which is 2 4 dinitrofluorobenzene. So, now it is actually so, as I have said that chlorobenzene is it absorbs UV light. So, you can see it in the UV chamber or you can you can estimate its quantity if you do the UV spectroscopy.

So, nowadays, we do not use the chlorobenzene that is being replaced by fluorescence molecule So instead using that 2 4 dinitrochlorobenzene, now the fluorescence molecules are being used to tag the N-terminal amino acid.

(Refer Slide Time: 43:09)



One of them there are 2 of them which are popular CH 3 C H 3, this component is known as dabsyl chloride, this was fluorescents or there is another one called dansyl which is the structure of the dansyl chloride yeah. Similar functionality, here is the S-Cl in this case here is there S-Cl

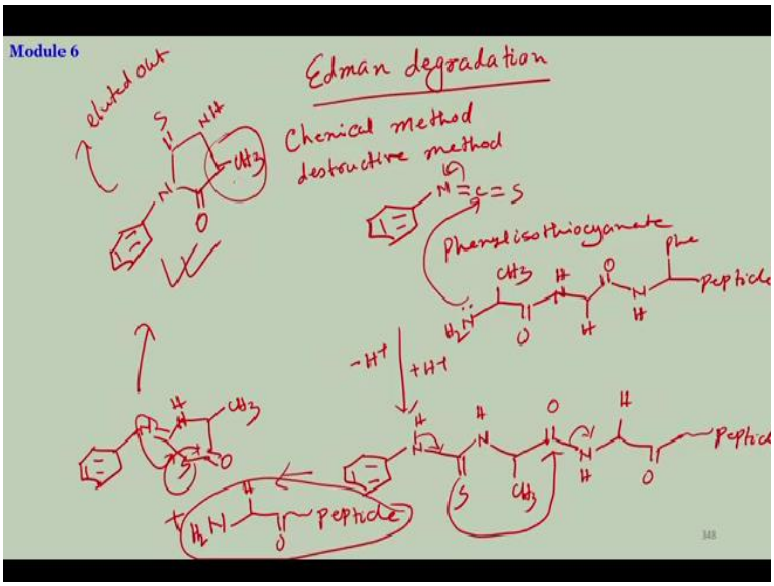
and this is the secondary amine. So this is known as dansyl chloride, both dabsyl chloride and dansyl chloride, so fluorescents.

And both of them react with the free amine present as the N-terminal amino acid simple reaction nucleophilic substitution chloride out. I am just writing this CH 3 CH 3 N double bond N N H. Now you can draw the other peptide. So R1 CO NH and so on, followed by hydrolysis that will cleave your peptide bond here and you will be left with CH 3 CH 3 NH R 1 COOH, the first amino acids, the N-terminal amino acids.

And then all the individual amino acids will be there. So, your first N-terminal amino acids would be attached to the fluorophore unit. And you can see when just by reading the fluorescence, you can identify the structure of the compound or you can also find out what is the amino acid present at N-terminal. So, that is about the Sanger's method, which is used for the determination of the N-terminal amino acids.

But it does not tell you about what are the sequence of the other part of the peptide or other parts of the protein.

(Refer Slide Time: 47:00)



For that comes the Edman degradation. Edman degradation is the most widely used protein sequencing method that tells you about the complete sequence of the protein. Most often, the

protein sequencing is done in combination of Sanger's method plus the Edman degradation method. Sanger method is used to find out the N-terminal amino acid once this is done, then the Edman degradation will come into the picture.

So, this is also a chemical method and a destructive one which means your compound or your protein sample is destroyed you do not get it back. How is it done. It also uses a specific chemical reagent, which is this aromatic. This is phenyl isothiocyanate. This is the molecule or the reagent that is used to react with the protein. So, let me show you the reaction first and then we will go for the sequence wise how is it happening.

So, if you have let us take a peptide sequence CO, this is NH, this is R 1, this is R 2, R 1 I have taken as alanine it is can be R 2 let us say this is a glycine with H CO NH and then CO the rest of the peptide, why I am writing a little bit sequence you will know shortly because the reaction happened here. So first thing first S is a nucleophilic attack of the NH N-terminal amino acid here that will ultimately comes here and then it comes back whatever you get N minus, which will abstract the proton from here.

So it is a minus H plus and plus H plus. So, this becomes NH, this carbon double bond S, this carbon is connected covalently to this nitrogen. So, it would be NH here, I am not drawing the stereochemistry of the amino acids. So, they are all L configuration and then it will be here NH. This is your glycine CO NH, CO NH this should be the I am sorry glysine this should be something else, let us say phenylalanine.

And then the rest of the peptide CO, I think we can go here peptide do not need more. So, this is formed. Now, once this component is formed, there is another internal reaction that occurs here is your NH, this lone pair or this bond breaks down, it will form a double bond that makes this sulpher electronegative. And these reacts with this carbon compound, reacts here and it accurately breaks this one. So, this reaction or this cyclization would break the first peptide bond that is present here. This bond would be chopped off.

And if we can write here it will be this, this will be N here is the double bond, this should form a 5 membered ring, this should be your S reacting with CO O, This is chopped out. So now S CH 3 now comes to NH here yeah, this is N, this is H, this would be formed plus it is cleaving here the rest of the peptide is it will make another this NH 2 free NH you had glycine H CO and the peptide. So, this is the essential part of it.

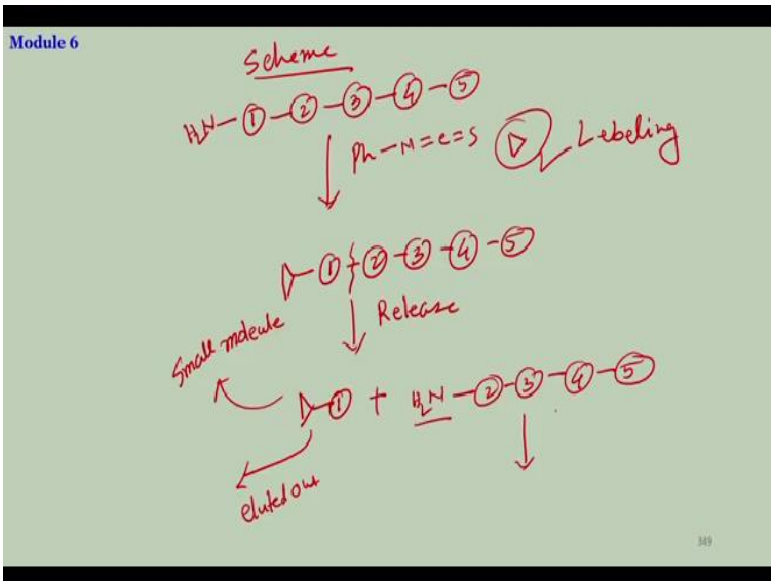
Because of this internal cyclization, it cleaves the first peptide bond that results in formation or exposure of a new amine free group, which is attached to the rest of the peptide plus this molecule with the first N-terminal amino acid which is alanine in this case and this is a small molecule, which can be this actually this compound further gets rearranged into a stable product. Write this down in N S NH.

This is CH 3 and double bond O here. What happens is this double bond to the nitrogen that reacts here and this sulfur double bond, this bond gets back So, you form the carbon sulfur double bond, you attack this that gives you this component. So, as you get this stable ring cyclase product plus the peptide attached with the other group except the first one. Now, this is eluted out because it is a small molecule.

This version is a long molecule still present and it can easily be separated. So, this will be eluted out and from its characterization, you can know what is the first amino acid that was present. And now you are left with the other part of the peptide with an exposure of another amine, now once this is elected out, this one will start reacting with the amine reagent again, and it will continue. So one by one, you will get a cleavage now you are getting sequence.

So it is happening one way one not all at the same time. So first this is out, second would be glycine would be out, the third amino acids would be out something like that, it will go on.

(Refer Slide Time: 55:15)

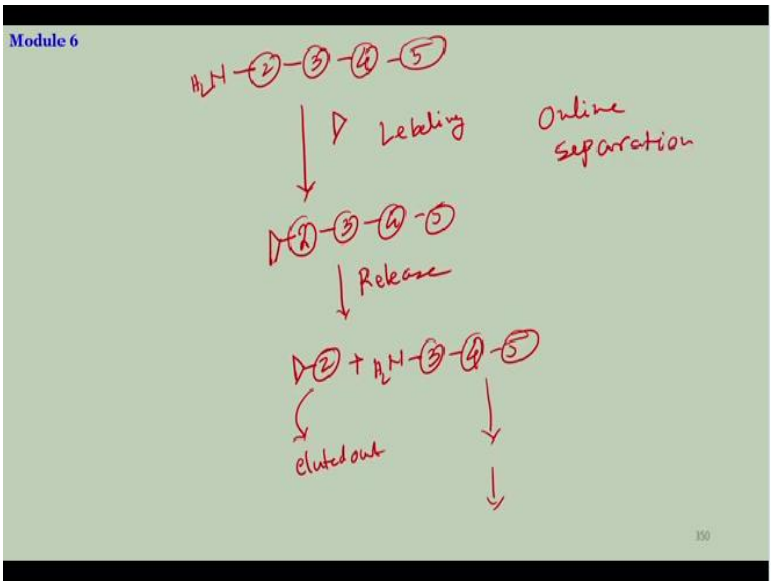


So I will quickly show you the sequence, the scheme, how this is happening. Let us say this is the first amino acid, this is the second amino acid. This is the third, the fourth, fifth. This is the fourth one which has amine free last one. Now, you will react with the pH N double bond C double bond S phenyl isothiocyanate which I am representing by this symbol, in book the symbol is given.

So whatever I am teaching for this protein sequencing, you can find it out in this trial book biochemistry written by Stryer. So, these then what will happen 2 3 4 5 and here you will have this conjugated to the amine group followed by, so this step is known as labeling because you are labeling your protein with the phenyl isothiocyanate a marker group. So, once the labeling is done, the second step is called the release.

Because this is where you are breaking it, it will break this bond, this is formed, you can see here, here the first one is formed and the rest is there. And this is called released because you are breaking this bond keeping the rest of the peptide releasing out the first amino acid plus you get the amine free of the second one 3 4 5. This is eluted out because this is a small molecule compared to the large protein, this is a very small molecule. So, it is easy to take it out and you will lift up with this. Now again this is exposed.

(Refer Slide Time: 58:03)



So in the next scheme, I will start here N H 2, this is 2 3 4 and 5. Again you will react to this, it is all happening in the same mixture, you do not have to do it separately. Once the other part is out the smaller part is out the rest of the thing you already have phenyl isothiocyanate into a solution. So it will go on. So it will be 2 that is connected, that is level 3 4 5. So this is leveling step again followed by release.

That will release number 2 plus you will have free amine after 3, this would be eluted out again and characterized and then it will continue okay. So, all these things are done online and write this term online separation, the moment that small molecule is formed, it will be released or it will be eluted out and the rest will go for the next step of the reactions. So, you are at the same time you are characterizing the small molecule what is the identity of this amino acid plus the other part of the reaction is going on at the same time.

So, that is how sequentially you are breaking or your sequence like cleaving individual peptide bonds one by one and you can understand their nature by the online separation and the characterizations. So that is about the Edman's degradation, which will tell you the full sequence of a protein. Thank you.