

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high-quality, educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at [ocw.mit.edu](https://ocw.mit.edu).

**BOGDAN**  
**FEDELES:**

Hi, and welcome to 5.07 Biochemistry Online, the biochemistry course on MIT OpenCourseWare. I'm Dr. Bogdan Fedeles. Let's metabolize some problems.

Now today we're going to do a really nice problem. This is Problem 1 in Problem Set 2. Now, this is a problem about elucidating the primary structure of a protein. Problems like this one that we're about to discuss are a lot of fun because they're really biochemistry puzzles. We're given a number of pieces of data or clues, if you want, and then we'd have to use, not only our biochemical sense, but also deduction and elimination process in order to come up with a final answer.

Well, in practice, elucidating the primary structure of a protein is now a largely automated process and utilizes high resolution mass spectrometry. Some of the traditional chemical methods that we're discussing here are still occasionally useful. But most importantly, the logical step-wise process through which we analyze and use each piece of data to construct a big picture result is really representative of the process by which we make discoveries in biochemistry.

Before we begin, let me just say that this problem assumes familiarity with the structures and abbreviations of the 20 natural amino acids. Feel free to pause this video and review the relevant chapters in the book and the lecture notes before continuing.

One important tool that we have for elucidating the primary structure of proteins is proteases. Proteases are enzymes that can hydrolyze the peptide bonds of a polypeptide chain. Now, proteases that can cleave in the middle of a polypeptide chain are also called endopeptidases.

Now you notice a lot of these names that end in "ase" denote enzymes, and peptidase means an enzyme that acts on a peptide. It's the enzyme that hydrolyzes the peptide bond. Endo, in this case, refers to the fact that it acts in the middle of a polypeptide chain.

Now, we're going to be learning in this problem about trypsin and chymotrypsin. These are proteases that cleave in the middle of a polypeptide chain. They are endopeptidases. One

important feature of proteases is that they are specific. They don't just cut any which one peptide bond, but rather they recognize a specific sequence of amino acids.

In the case of trypsin, for example, we are told that it cleaves adjacent to positively charged amino acid. As you know, positively charged amino acids would be lysine or arginine. So trypsin will always be cutting after arginine or lysine. Let's take a look.

Now, here is a polypeptide chain with amino acid residues, R1, R2, R3, R4. Now, let's look at this peptide bond right here in the middle of the chain. Now let's say in order for trypsin to cut, to hydrolyze this peptide bond, it means that R2, the amino acid residue adjacent to it, should be a positively charged one.

So if R2 is lysine or arginine, then this bond here becomes a good substrate for trypsin. And what it's going to do, it's going to use a water molecule-- it's going to put here trypsin-- and it's going to hydrolyze forming a carboxyl end and an amine end of this peptide bond. So we're getting this is a carboxyl end and this is the amine end of that original peptide bond.

So what I want you to remember is that if we're having a reaction with trypsin on a polypeptide chain, then the carboxy end of the peptide that results, which is this particular amino acid, is going to be one of either lysine or arginine. So in other words in a trypsin digest, all the smaller peptides that we obtain are going to end in arginine or lysine except, of course, the very end of the chain, which might have a different amino acid at its carboxyl end.

Now, this consideration we've just made about the trypsin digest, in fact, answers part 1 of the problem, which is asking what is common about all these peptides generated by a trypsin digest. And as we've just explained, all these peptides should be ending with a positively charged amino acid such as lysine or arginine.

This problem also mentions another protease chymotrypsin, which is a protease that has a different specificity from trypsin. It actually cleaves after amino acids that are either very hydrophobic and large or aromatic. So let's write that down and remember.

So if we were to do a digest with chymotrypsin, then our R2, the residue that's recognized by the protease, is going to have to be either aromatic, so phenylalanine, tyrosine, or tryptophan, or something that's large and hydrophobic such as leucine, isoleucine, or even valine sometimes. Put this in parentheses.

So if any of these residues are at this position, R2, then this protease, chymotrypsin, is going to generate two peptides. And once again, the resulting peptide, R2 at the carboxy end is going to be one of these amino acids. That's going to be the signature of a chymotrypsin digest.

Now, keeping these things in mind, we're ready to tackle the rest of the problem. Question 2 asks about the use of DTT, or dithiothreitol. Now, DTT is a commonly used reducing agent, which can reduce disulfide bridges in proteins. There are many reasons why we want to use DTT.

For example, when proteins form disulfide bridges, you may shield certain amino acids from being accessible by proteases, and therefore, they're not going to be cleaved and we'll get a mixture of peptides. So it makes our analysis and our results very difficult.

Now, disulfide bridges can also hold two peptides together that have no other covalent attachment between them. So in that case, we get one fragment instead of two fragments, and once again, it complicates our analysis of the protein. But more importantly, because we're typically purifying proteins in the air, in an oxygen atmosphere, proteins can acquire disulfide bonds, which weren't there in the beginning.

So in that case, we can get very unusual results, unreproducible and artefactual. That's why using DTT can prevent formation of spurious disulfide bridges. And finally, DTT is also useful to tell if there were any disulfide bridges in the protein to begin with. Because if we're looking at the analysis before and after using DTT, we can tell if the results change, and that will tell us whether disulfide bridges were there to begin with.

Now, let's take a look at the DTT chemistry. Here we have a disulfide bond or bridge in a protein, and we're going to treat this with DTT, which looks like this. This is DTT or dithiothreitol.

Now, if we substitute the SH groups with OH's, you notice we're going to have four OH's and four carbons. That's just an alcohol derived from a sugar, which is called threose That's why the threitol part of the name.

All right. So when we do this chemical reaction, this disulfide bridge is going to transfer between the two sulfur atoms of DTT. They're going to form a intramolecular sulfide bond. So our cysteines are going to get reduced, and from DTT, we're going to get this intramolecular

sulfide.

And because entropical considerations, make the formation of the six-member ring very, very easy, then this equilibrium will typically shift to the right. Of course, we're also using vast quantities of DTT to make sure that the entire protein becomes reduced, and then our agent is going to pick up this [INAUDIBLE] off the bond. This sums up the question 2 of the problem.

Next, we're going to look at distinguishing between a couple of different peptides that we generate during our analysis of our mystery protein. So we're told that we're isolating by HPLC peptides of the following composition. One has tryptophan, phenylalanine, valine, aspartate, lysine, cysteine. That's peptide one. Another one has phenylalanine, serine, cysteine, and an unknown amino acid. Finally, the third one has alanine and lysine.

All right. Now, obviously, these peptides have different compositions, so if you could just put them through mass spectrometry, we're going to get different masses, so we can tell very quickly which one is which. But if we don't have mass spec available, we can also tell them apart only using UV-Vis spectroscopy.

So all you need to remember is that an amino acid like tryptophan that we have here, W, has a very strong absorption in the 280 to 300 nanometers. Whereas, amino acids like phenylalanine, present here or here, they absorb only around 260 nanometers. Most of the other amino acids don't absorb in this range at all.

So if we were to plot the UV-Vis spectra of these peptides, this is going to be our absorption, and this is going to be lambda in nanometers, the wavelength. So let's look in the range from 200 to 400. 300 is about here. 250 is about here.

And let's label these. Let's say this peptide is red, this peptide is blue, and this one is green. All right. So now the red peptide, as I told you, contains both tryptophan and phenylalanine, so it's going to absorb both around 260 and both around 280 to 300. So it's going to have a pretty big hump in this area from 250 to 300.

Now, the blue peptide only has phenylalanine, so above 280 or so, nanometer is going to drop off. So it's going to look more like this, whereas, the green one, well, it has neither phenylalanine or tryptophan, so below 240 or so, it's going to have no UV absorption at all. So it's going to drop off right here.

So basically, if we're just looking, say, around 260 nanometers, we should see a stronger peak

from the red peptide and a weaker peak from the blue one. But if we're going to look around 300, we should only see the red peptide. So by just using the UV-Vis absorption, we can tell these three peptides apart.

Now we're ready to figure out the structure of our mystery protein. This is exactly what the last part of the problem is asking. So we're going to take each one of the clues provided-- A, B, C, and D-- and analyze and see what information we can derive from each one of them.

The first piece of information that we're given is the result of a total hydrolysis of our peptide with six molar HCl. So we're told we get the following amino acid composition. We get two phenylalanines, one methionine, alanine, valine, two lysines, one serine, two cysteine, and one aspartate.

Now, recall this is not a comprehensive list because the hydrolysis and six smaller HCl may destroy some of the amino acids. And specifically, we know for sure something like tryptophan, threonine, and tyrosine. They're going to be destroyed, and we're not going to see them here. So these amino acids may still be present in our protein, but we're not going to be able to see them in this situation.

Now, let's analyze these amino acids and see what we can derive from this. So first of all, we have cysteines. So two cysteines, it means our protein can form disulfide bridges. So from the get go, two cysteine, it means we can form an internal disulfide bridge or our protein contains a disulfide bridge between two otherwise unconnected peptides.

So here are some possibilities. For example, our peptide is one chain, and our cysteines are not actually connected by a disulfide bridge. Another possibility is we do have a disulfide bridge between them, like that. Or yet another possibility is that we have two pieces, two polypeptide chains, and the disulfide bridge is the only thing that's holding them together.

Now, in each one of these cases, obviously, these polypeptide chains are oriented. So in one end, they're going to have the amine group. Let me draw it here,  $\text{NH}_3$ . Here we're going to have two, and the other end is going to be the carboxyl group,  $\text{COO}^-$ . And once again, in this case, we're going to have two of them.

All right. Now, how can we narrow down from these couple of possibilities? Well, the fact that we're getting two lysines here is a very important clue. Now remember, we're told that our peptide, this mystery protein, it actually comes as a result of a trypsin digest. And if you recall

our discussion earlier, we said that upon trypsin digest, all the smaller peptides that were obtained will end in lysine or an arginine.

The fact that we have two lysines here, it means both of them they need to be at the carboxy ends of peptides. Because if they were in the middle of a chain, the trypsin would have cut that chain in half.

So two lysines means we've got to have two carboxy ends. So therefore, this seems to be the only possibility in which we can accommodate two lysines. Basically, the two carboxy ends that we see here, each one has to be a lysine.

This possibility only has one carboxy end, so the other lysine that we have to place will not be at the end, so would not be compatible with a trypsin digest. Same applies for this case. So these possibilities are not consistent with our data. So we know already that our protein must look like this, two polypeptide chains, each one ends with lysine, and there must be a disulfide bridge.

The second clue we're given is that the Edman degradation of the protein yields valine. Now, as you know, Edman degradation is a chemical reaction by which we can digest the protein from the N terminus, from the amino terminus of a polypeptide chain.

Now as you recall, we have established in the first part that we have two amino termini in our protein. So the fact that we only get one amino acid and that is valine, it says that the other amino terminus might be blocked or somehow unavailable for the Edman degradation.

So let's update the structure of our protein to take into account the second clue. So we said we have two polypeptide chains. There's a disulfide bond in between them. Now, each one of them ends with a lysine. This is a carboxy end, and now we know from the Edman degradation that amino end of one of them has to be valine.

The other one-- we're going to put a box like that-- is a blocked end. So the amino terminus is not available. So this is as much as we can tell from these first two clues.

We're given the products of the chymotrypsin digest of our protein after it was previously treated with DTT. So we know first DTT is going to cleave the disulfide bond, and then chymotrypsin, as we talked previously, is going to cut after large, hydrophobic, or aromatic amino acids.

Now, we're told we're getting five smaller peptides. Let's take a look. Here are the five fragments. One is tryptophan, valine; one is cysteine, phenylalanine, another one is aspartate lysine, another one is methionine, alanine, cysteine, and lysine; and finally, the last one has serine, phenylalanine, and something that's not an amino acid. It's going to make it like a small x.

All right. Now, from what we know about chymotrypsin digest, we should be able to orient these peptides basically to tell which amino acid is at the N terminus and which amino acid is at the C terminus of each one of them. Now the first one. Well, we know from the second clue that valine is at the N terminus, so that makes it pretty easy. Then the sequence has to be V-W. So valine is the N terminus, W is the C terminus, and as we said, tryptophan is one of the amino acids recognized by chymotrypsin, so it's going to end up at the carboxy terminus.

C-F, that has to go C and F, phenylalanine, another chymotrypsin amino acid that's left of this carboxy terminus. D-K. Now, neither of these amino acids is recognized by chymotrypsin, but we remember from clue one that D-K must be the carboxy terminus. So the sequence can only be D-K.

Now, here M, A, C, and K, none of these is actually on our recognition list for chymotrypsin, but once again, we know that K must be in the carboxy end. So for now we're going to have M, A, and C in a particular order, which we cannot establish just yet and K at the carboxy end.

And finally, we do an S, F and x. Well, x is not even an amino acid, and F is an amino acid that's left of the carboxy terminus by chymotrypsin, so it's probably x, S, and F. Now, what about x? We're told that x is not an amino acid and is hydrophobic and it has a molecular weight of 256 Daltons.

All right. So in order to figure out what x is, we have to read back the beginning of the problem, which tells us that we're looking at a protein that's associated with a plasma membrane. And one way for proteins to associate with a plasma membrane is to be modified, to incorporate a fatty acid, such as palmitate. So perhaps x, which is blocking the N terminus of this peptide is a fatty acid.

Now, the general formula for fatty acids is something like this,  $\text{CH}_3 \text{CH}_2$  repeated, say, n times, and then  $\text{COOH}$ . So let's see what would be the fatty acid that has the molecular weight 256 Daltons? Well, the mass of a methyl group is 15. The mass of this carboxylate is 45, so we have about  $60 + 14n$  equals 256 or  $14n$  equals 196. Therefore, n is 14.

So the fatty acid that would fit these criteria-- it's hydrophobic, it has a mass of 256-- will be  $\text{CH}_3$ ,  $\text{CH}_2$  14 times  $\text{COOH}$ , or the fatty acid was 16 carbons. This is palmitic acid.

Now, the answer that we got, palmitic acid, was anticipated in the text of the problem because it gave us an example one way to associate proteins to the plasma membrane is to form a covalent linkage with a fatty acid such as palmitate. But how does a protein associate with a plasma membrane when it has a palmitic acid residue as part of it? Well, let's take a look at a diagram.

Here we have a representation of the plasma membrane, where we have the phospholipids like the hydrophilic head pointing inside and outside the cell and the hydrophobic tails of the fatty acids lined up to each other. Now, imagine we have a protein that's modified to contain one of these fatty acids. Then this fatty acid can just insert right next to the phospholipids of the plasma membrane. And that way it tethers the protein right at the plasma membrane.

The final clue that we're given in order to figure out the structure of our mystery protein is the digest with an inorganic agent this time. It's called cyanogen bromide. Cyanogen bromide reacts quite specifically with methionines, and it cleaves the peptide bond after methionine leaving behind an unnatural amino acid called a homoserine lactone.

Now, let's take a look at what happens when we treat our protein with cyanogen bromide. So we're told we're getting the following peptides, A, K and W, F, V, D, K, C and F, S, C, and an unnatural amino acid.

As I just explained to you, the unnatural amino acid probably is this homoserine lactone. So it's most likely, instead of this amino acid in the actual sequence, we had a methionine. So we know these four residues-- F, S, C, and M-- go together.

All right. Now, we can start putting together the clues from part 3 and 4 and try to figure out the final structure of our protein. So the peptide A, K, we know has to have K at the C terminus. Now, from part 3, we found out that there was a peptide that contained M, A, C, and K, after the chymotrypsin digest.

So now we know K has to be the carboxy end and A has to be right before K, so in order to get an A-K peptide, then A has to be right after methionine because that's where cyanogen bromide is going to cleave the peptide bond. So the only sequence that we can have here is

going to be C followed by M followed by A followed by K. So when we treat the cyanogen bromide, we're going to be cleaving this bond between M and A, generating M as a homoserine lactone.

Now, we also know that M has to be in this small peptide, and we know it has to be C, M as a sequence. So therefore, S and F have to be on the amino terminus of this peptide. And we also have a clue from part 3, which said that S, F, and this non amino acid moiety were in the same peptide.

So from here we said that, well, the sequence there, it's probably x modifying the amino S and then F with a carboxy end. So putting these three things together, we can come up with a sequence for this strand, which is going to be x, S, F, then C, M and then A and K. So this is probably one of the peptide chains in our mystery protein. Then, of course, the other is going to be composed of these amino acids.

Now, we already know something about the sequence of these. For example, we know V goes before W. We also know D goes before K, and also know C goes before F. Now, we know this has to be the carboxy end of the peptide chain. Let's write it here, COO minus. And V has to be the amino end.

So then C-F, there's no other way, has to be in the middle. So the only possible sequence for our second chain is going to be V, W, C, F, D, and K.

Now that we've established the exact sequence of each one of these peptide chains, then we can put together the final structure of our mystery protein. So I'm going to transcribe these here, the first chain V, W, C, F, D, and K. And we know the cysteine is going to have our disulfide bridge to the other cysteine, which goes C, M, A, K, and then F, and then S.

And now we know the N terminus of this peptide, we're going to have our fatty acid residue CH<sub>2</sub> 14 times CH<sub>3</sub>. So let's just mark, once again, the carboxy ends here and the amino end here. So this is the final answer for our problem and the structure of our mystery peptide.

Well, that's it for this problem. I hope you enjoyed this little protein mystery hunt. Now, remember that the strategy that we used here in which we logically string together pieces of data to build a big picture, it's really the same strategy that has been used and is being used right now to advance our knowledge about living systems and their underlying biochemistry.