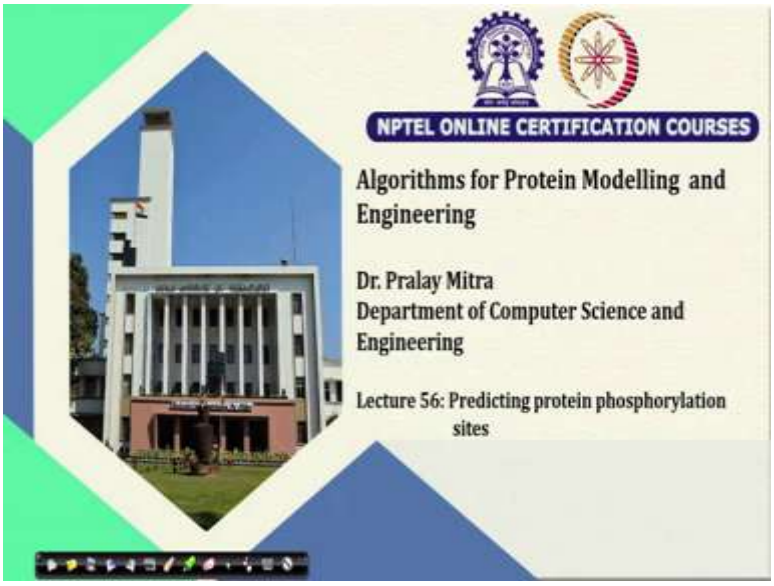
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur
Lecture 56
Predicting Protein Phosphorylation Sites

Welcome back to the course on Algorithms for Protein Modeling and Engineering. We have reached to the last week of this course.

(Refer Slide Time: 00:32)



And in this week, we will start from where we left on the last week, that is we started discussion on post translational modification, in short PTM, and I also mentioned a lot of work in terms of the competition is still pending in this particular area.

So, some fascinating or interesting experimental work done by several scientists over the last few decades has been displayed to you or I presented that to you, so which indicates that a lot of post translational modifications are there and in a nutshell. So post translational modifications is actually occur after the translation of the protein through that DNA to RNA and then protein. So, after the translation is over, then only post translation modification will take place.

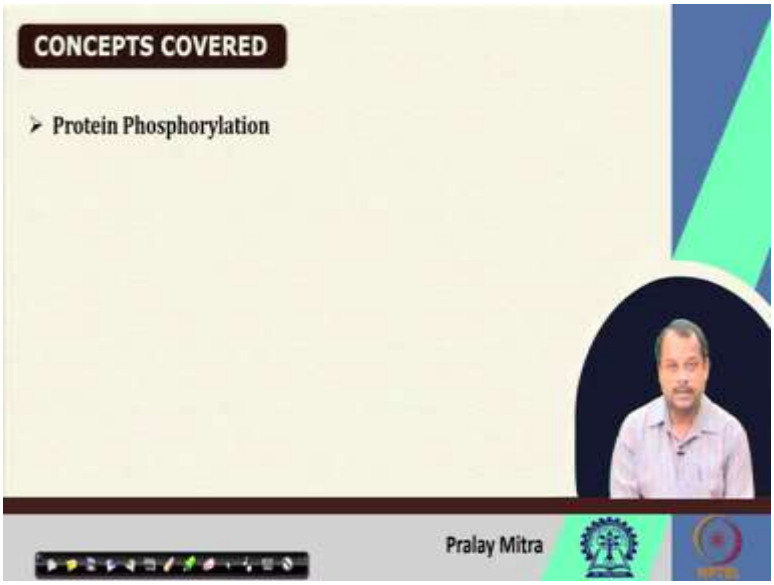
Now, among those post translation modifications, so few post translation modifications or PTM is responsible for stability, stabilizing the structure. Say, for example, if one disulfide bond is formed between two cysteine molecule, then basically it indicates that the stability will be

increased. So, there are say two cysteine atoms as, two cysteine molecule where each part is having one sulfur here and there and if sequentially they are not neighbor, but they are placed at a distance position, but during the protein folding process after the translation is over, they may come closer to each other.

Now, you think about a situation where there is no disulfide bond, then although they come close to each other, then the interaction or the stability because of the closeness will not be that much because of their closeness one disulfide bond kind of one constraint is placed at that position. So, if you can predict that kind of post translation modification, then you can comment or you can infer out of that whether the protein is going to be more stable or say whether functionally it will be having some importance or not.

So, that disulphide bond formation is following the stability of the protein, whereas if there is some methylation or phosphorylation, so which will take place then that mostly goes with the functionality of the protein. So, both has some effect. And also, we discussed that if the required post translation modification will not take place, then that may lead to unstability or say misfolding of the protein or say it can generate some toxic protein or may lead to some protein related disease condition. So, here in this lecture, we will start discussing regarding predicting one such PTM which is phosphorylation.

(Refer Slide Time: 03:28)

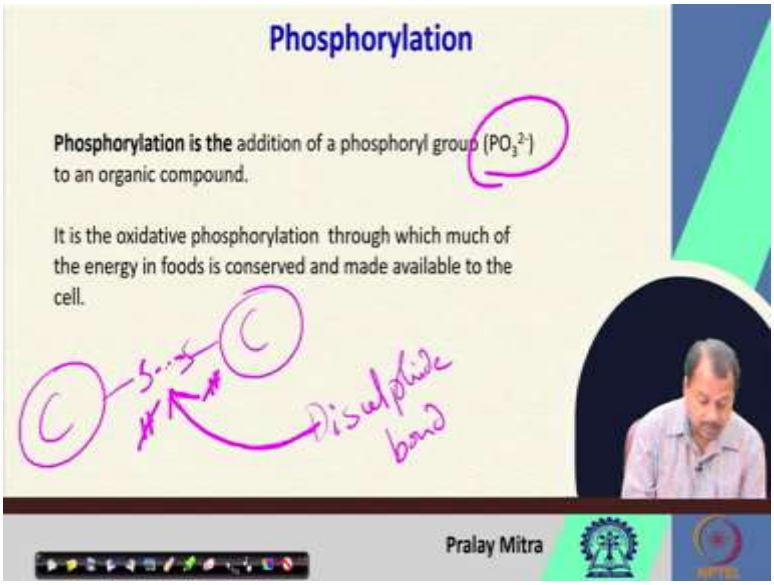


So, we are planning to cover the protein phosphorylation here. So, first we will start discussing regarding the phosphorylation then we will discuss a little bit of chemistry, not much. But after discussing that phosphorylation, we will see whether the knowledge of the chemistry and some domain knowledge of the biology can be useful for us to prune down our search space or not that we will discuss, and then we will go for developing some machine learning based tools regarding this protein phosphorylation site prediction. Again, when I say that it is prediction which means I am using some algorithm, either machine learning or any other algorithm to predict what is going to be the correct one.

(Refer Slide Time: 04:18)

So, keyword that is why I picked the same, protein phosphorylation here.

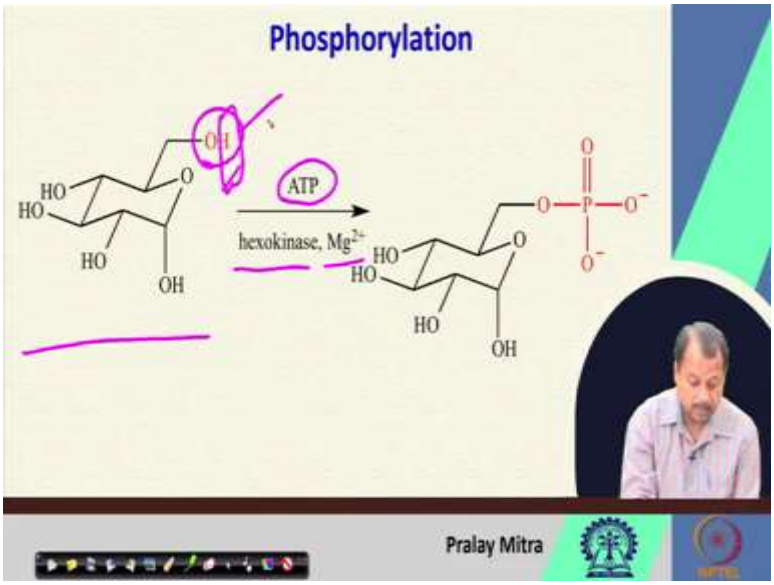
(Refer Slide Time: 04:22)

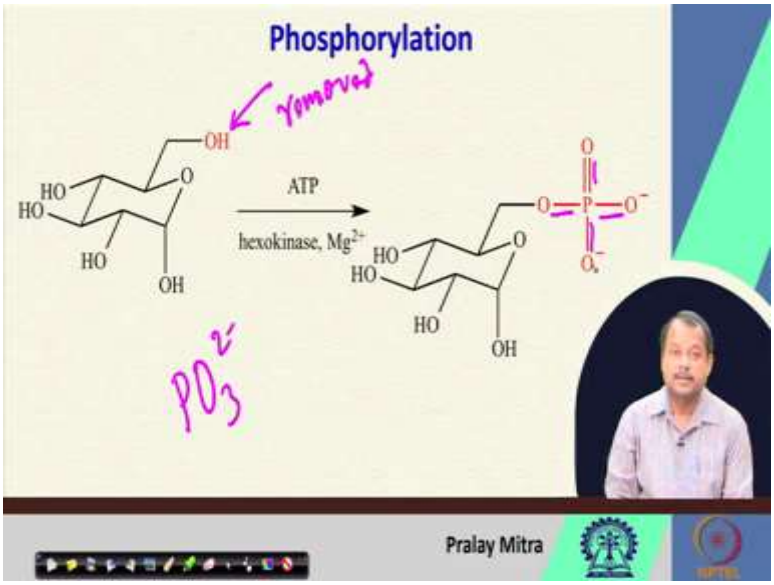


So, as per the definition, so phosphorylation indicates, phosphorylation is the addition of a phosphoryl group PO3 2 minus to an organic compound. It is an oxidative phosphorylation through which much of the energy in foods is conserved and made available to the cell. So, in this case, when I am defining this phosphorylation process, I am not considering that it is the phosphorylation of the protein. It is in general, what is the definition of the phosphorylation, and what it used to do.

Now, here you can see that, so in case of say disulfide bond when I say that it is disulfide bond, so in case of disulfide bond, there are two cysteine molecule S and S from both the side will form the bond. So, dotted line indicates the disulfide bond, so this one. Let me ease a little bit. So, this one is indicating the disulfide bond. Initially what was there is along with this single bond with some previous carbon one H was here and H was here those two H was eliminated and one bond is formed here. So, that is the disulfide bond formation. Similar to that in case of phosphorylation group also we are expecting that this PO3 double minus actually will go and bind.

(Refer Slide Time: 06:17)

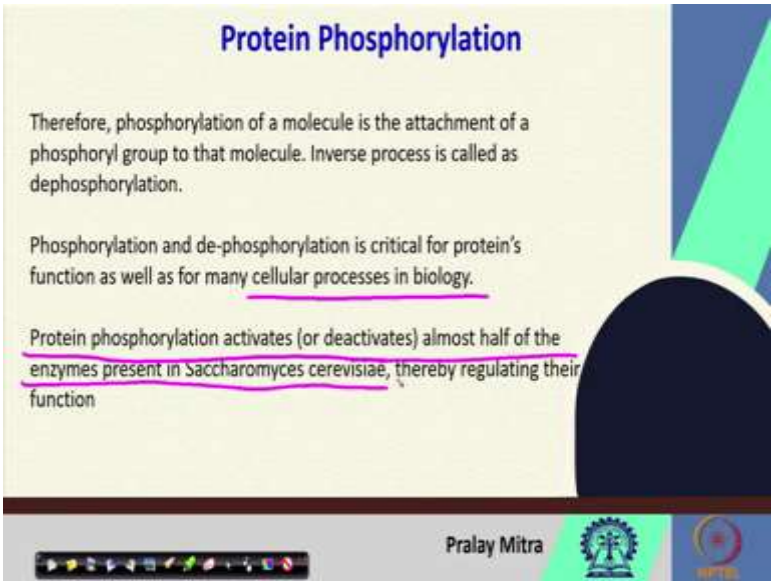




Now, chemically it shows like this. Let us assume that on the left hand side there is one molecule. So, what is the molecule, I am not going to detail. So, here I see that there is one OH here. Now, if ATP, we heard the name ATP, adenosine triphosphate, which actually donates one phosphate and then it net shell become ADP adenosine diphosphate from triphosphate to diphosphate it changes and as a result of phosphorylation occurs. And when the opposite occurs, ADP consumes one more phosphate and converts to ATP. So, that is called dephosphorylation with respect to that particular molecule.

So, from which molecule the phosphate is going out, so phosphorylation and dephosphorylation, but our interest is phosphorylation and we are going to discuss this phosphorylation part. So hexokinase or, so this ATP in presence of hexokinase or magnesium, and the magnesium atom it will transfer its phosphate to this guy. So, when this phosphate will be transferred you see that this H, so this H actually will be removed, this H will be removed and in place of that one actually this phosphate, what I mentioned on the last slide, it is PO3 2 minus so that will be added here. So, with this one, so 1, 2, 3 and this and 2 negative that has been added here. So, that is the phosphorylation stage.

(Refer Slide Time: 08:17)



Now, if I go to the protein, then in protein also I have to look for the occurrence of such amino acids. So, here we are now focusing only on the essential amino acids. Apart from that one if there are some moieties, small molecule, ligand etc., which generally we consider as heteroatom as per the PDB file format, so we are not considering that one. Then we have to look for similarity in the protein sequence or among the 20 essential amino acids who are capable of forming this phosphorylation.

So, for example, for disulfide bond it is only the cysteine, although there exists methionine and cysteine which has sulfur atom as their, in their side chain, but methionine does not have the capability to form the disulfide bond. It is only the cysteine who has the capability to form that disulfide bond. So, for protein phosphorylation also we have to see out of 20 different essential amino acids, so which are capable of forming the phosphorylation.

But before that few definitions for you. So, therefore, phosphorylation of a molecule is the attachment of a phosphoryl group to the molecule. Inverse process is called as de-phosphorylation. So, phosphorylation and dephosphorylation is critical for proteins function as well as for many cellular processes in biology. So, protein phosphorylation activates or deactivates almost half of the enzymes present in Saccharomyces cerevisiae thereby regulating their function.

Also, I would like to mention you one more biological process in biology which actually is controlled by this phosphorylation. So, you know that chemotactic movement of a lot of say unicellular organisms, say for, so if I assume that there is one chemotactic moment of say E. coli then it actually senses that the chemical gradient. So, chemotactic movement means it follows the chemical gradient and it tries to go to that region.

Now, if there is a concentration gradient in the environment and say one E. coli is here actually then what it will do that if it finds that there is sufficient amount of say chemical, in this case glucose, then it will try to go to that direction. And when it will find that say there is not such concentration available in the environment then instead of say going to a particular direction so it will basically move around. So, it will make some random move or it will toggle.

So, when it will make some random move during that process it will try to sense that whether there exists any gradient of the chemical concentration increase or not. If yes, then it will go

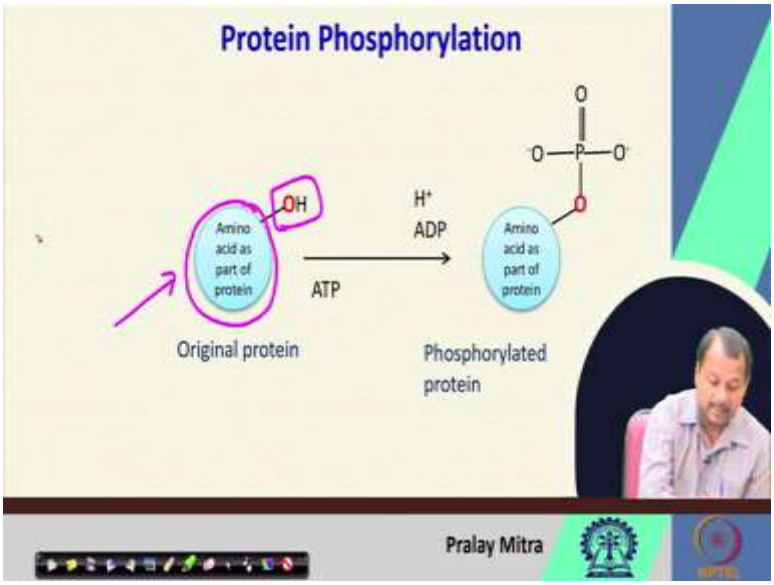
straight and follow that gradient. Now, this movement going to one particular say direction, straight direction, straight or tumbling or tumble around in order to search for a particular or tumble around in order to search for a chemical gradient or concentration is basically dictates by its flagellum movement.

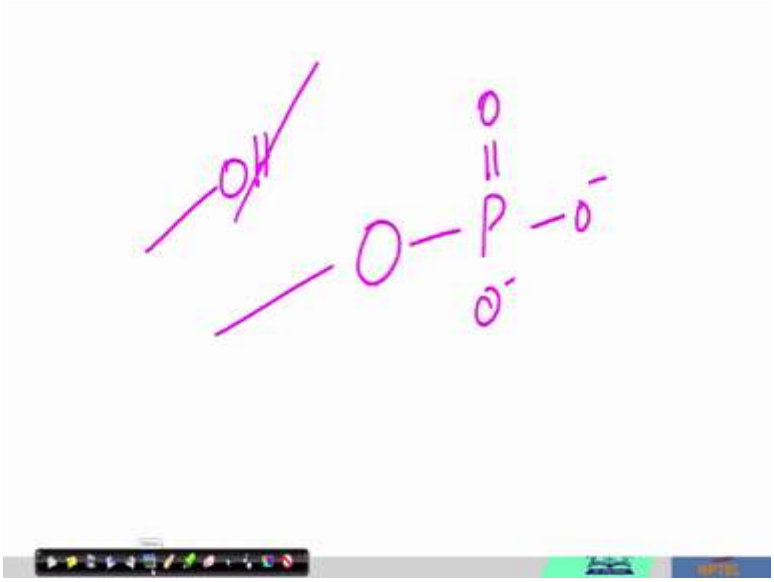
Now, if I say flagellum, so which say is attached with it, so if I assume that this is my flagellum, so this flagellum then you can see that this flagellum will have only two movements either, so that movement actually this flagellum will not have any translation movement, translation of this body or the unicellular organism will be guided by the rotational movement of this flagellum. So, this flagellum as you can see can move either clockwise or counterclockwise.

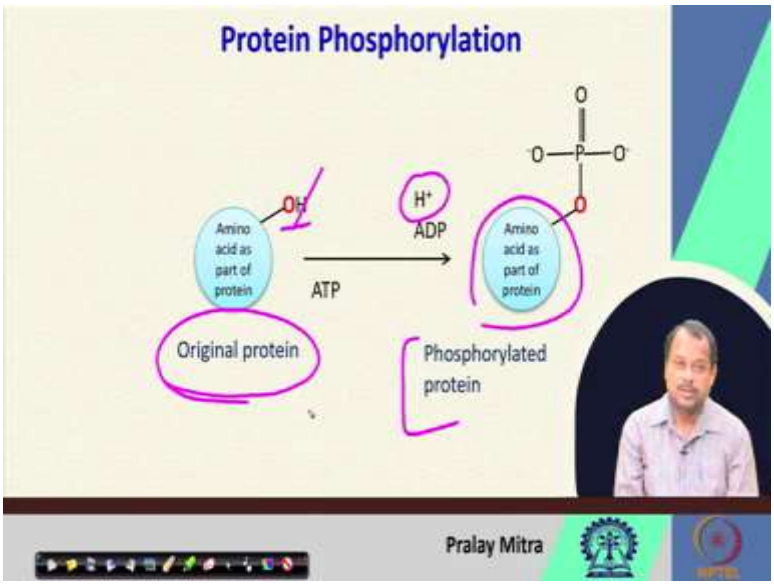
So, this CW, I mean clockwise or counterclockwise, the change in the rotation also the energy which will be generated and will guide this movement is following some phosphorylation which will take part at the motor of the unicellular bacteria, where this particular flagellum goes and binds there. So, there is motor kind of thing. So, there are say motor mechanism. And spin of the motor either clockwise or anticlockwise, counterclockwise that is basically dictated by some phosphorylation technique.

So, that is one example for this biological process, so or cellular process in biology that we have mentioned. So, similar to this some other or several such applications of the phosphorylation will be there. And it is correctly said that protein phosphorylation activates or deactivates almost half of the enzymes present in Saccharomyces cerevisiae, thereby regulating their function.

(Refer Slide Time: 14:14)







So, here in this protein phosphorylation, so let us as you that one protein molecule is there. So, for the brevity or simplicity of our discussion what we did that we consider some circle like this, which actually indicates that one protein structure. So, this is my original protein structure. Now, it has taken some form and as you understand only the amino acids or residue which are on the surface is capable of phosphorylation after the folding will take place.

And if I consider that phosphorus and relates to some function, then and not the stability, so when it is about the stability, then at the core also some PTM may occur. Say, for example disulfide bond formation. But if it is related to the function then it must be on the surface of that protein.

Now, phosphorylation if I assume that it leads to some functional changes, then it will have some function and definitely the amino acid which is supposed to get phosphorylated will be on the surface. So, that is why I am considering say one circle representing the protein in the folded structure and that particular folded structure will have some amino acids on the surface and with that amino acid the phosphorylation will take place.

Now, why I have taken this OH? Because if we draw the similarity with the definition of the phosphorylation that I have shown you few slides back that indicates that, that indicates, so there will be OH and if there is any OH then this H will be replaced and in that position the phosphorylation group will be attached. So, I am considering, so probably there are some amino

acids which are having the side chain something like this that is an important observation. I will go to that detail later.

But here to complete this, so ATP is going to bind with that particular amino acid where there is a side chain OH. It replaces H or substitute H. After substituting that one it place this PO3 double minus at that position. Then on the right hand side what I am getting this I am calling as the phosphorylated protein. And as a result of that one ATP will be transformed to ADP so adenosine triphosphate to diphosphate because one phosphate is donated to this particular protein and since it substituted this H, so this H will also come out from there. So, that is the total reaction which will take place.

Now, our job in this case is that given this original protein and hence we know that which amino acids are there, then is it possible for us to predict the protein phosphorylation site or not. Again, in the context of the protein molecule, the moment I will say that I am going to predict something an input is a protein then the first question probably will ask what do you mean by protein, is it a protein sequence or protein structure, because two things are completely different. So, this we demonstrated several times specifically on the last week also.

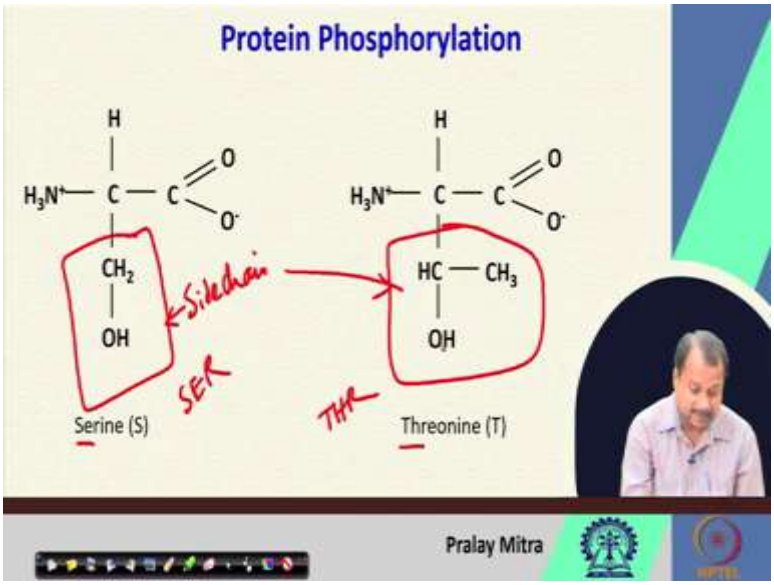
I mentioned that when protein structure is given to you then determining or assigning the secondary structure on the protein structure is not that much difficult. So, there is a geometric criteria, following that I can compute the hydrogen bond, and I can also compute the phi-psi angle and combining those two information I can predict or say I can basically assign what will be the secondary structure.

However, if it is sequence, then the problem is not that much easy. Because if it is sequence, then you do not know that to who can form the hydrogen bond with whom. Because I am, this hydrogen bonding thing in the context of the secondary structure I am talking in the context of main chain hydrogen bonding. So, if it is the main chain so all the amino acids are uniform. So, only it is at the side chain the amino acids differ or varies. So, from that point of view I will not get any advantage.

So, in this case also if I know the protein structure then I can able to identify which amino acids are on the surface and from there probably I can identify the protein phosphorylation site, but how if the protein sequence is given to you as an input, which means I do not know who is on the

surface. Shall I go for a protein folding and then I will go for protein phosphorylation. No, that will be costly. So, is there any alternative other computational way, we will search for that.

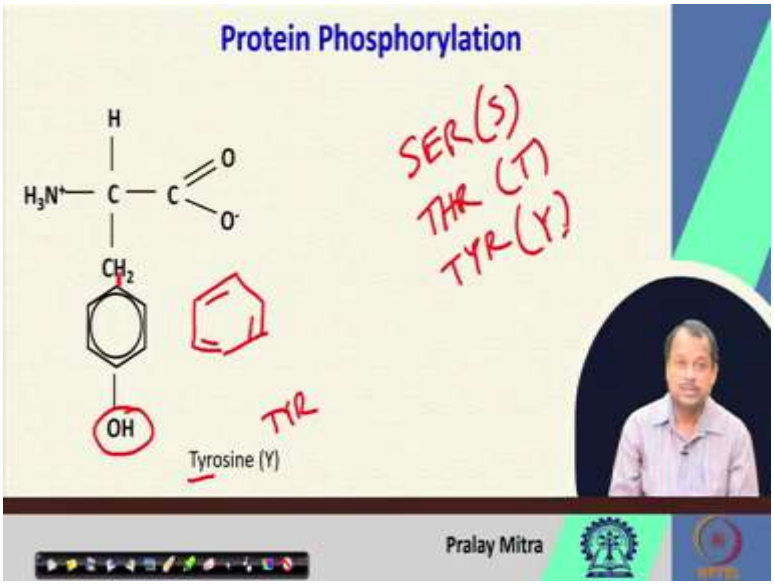
(Refer Slide Time: 19:44)



So, here I got two amino acids serine S and threonine T. So, if I check for the three letter code then it is ser, it is thr, so thr and ser. Now, this is the structure. Now, here what you can see at the side chain, so this is my side chain and this is my side chain. Now, at the side chain so for serine I see one methyl group and then OH and for threonine I see so one methyl and after that one, so it is attached with another methyl and there is one OH. So, OH is present for both, for serine and threonine.

Now, taking the analogy or similarity with the phosphorylation interaction which is supposed to take place when there is OH, hydroxyl group, then probably for the serine and threonine it will be easy for the phosphorylation or the phosphorylation will take place in serine and threonine. But is it only serine and threonine or anybody else? Let us see next.

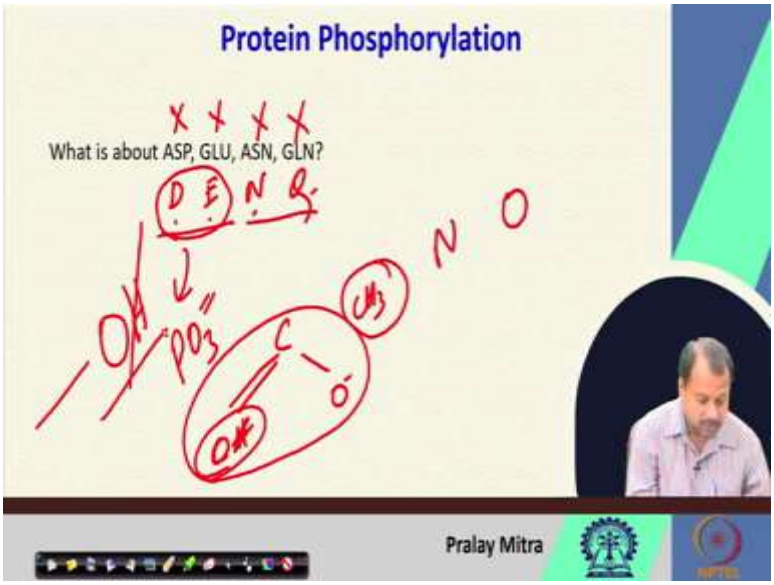
(Refer Slide Time: 21:13)



Yes, there is one more and that is my tyrosine. So, in case of tyrosine the side chain, so after the methionine there is one basically. So, here I miss to give one bond here actually. After that one there will be an aromatic one. So, this aromatic you can draw like this way or by giving a circle. So, I prefer to give some circle here. But this you can also draw like this. So, after that aromatic in one position so there is OH, so I got OH there.

Next, is there anybody else apart from this? This is TYR. So far what I have got serine, threonine and tyrosine, S, T and Y. So, apart from this S, T and Y is there anybody else with which the phosphorylation can take place. So, is it the presence of just an oxygen on the side chain atom. So, let us see. So far apart from these three, so there are 17 others. So, among those 17 amino acids, so who else are having the oxygen at the side chain?

(Refer Slide Time: 22:33)



It is aspartic acid ASP, glutamic acid GUE, asparagine ASN or in short N and glutamine so GLN. So, for all these cases, so here it is D, E, N, Q. So, for all these cases actually there is an oxygen at the side chain. So, for the acid, so D and E so what I will get see, so, OH and O. So, if I remove this O, then accordingly, so one double bond will be created here and then one negative will come here. So, like that way so it is not actually the OH which can able to form the phosphorylation, but this combination for the acid D and E is very negative in nature and can form the salt bridge with the base.

On the other hand, for asparagine and glutamine for both the cases, so similar to D and E, so actually this aspartic acid, this is asparagines, this is glutamic acid, then this is glutamine, so the variation between D and E is by only one CH3. So, variation between asparagine and glutamine also by one CH3. Now, in case of N and Q at the side chain, we have the N as well as O, but we do not have that particular OH which is actually required where this H will be replaced or substituted by some PO3 double minus or 2 minus. So, that kind of situation is not there.

So, this four is not qualified for protein phosphorylation. Three was there, now four in total seven, rest 13 do not have any oxygen in their side chain. So, if I think of say glycine, alanine, then cysteine, methionine, then say phenylalanine, then arginine, lysine, then tryptophan, then who else is remaining, yeah, mostly we have covered those. So, all those to not have any such OH kind of side chain so that that can be utilized for the protein phosphorylation.

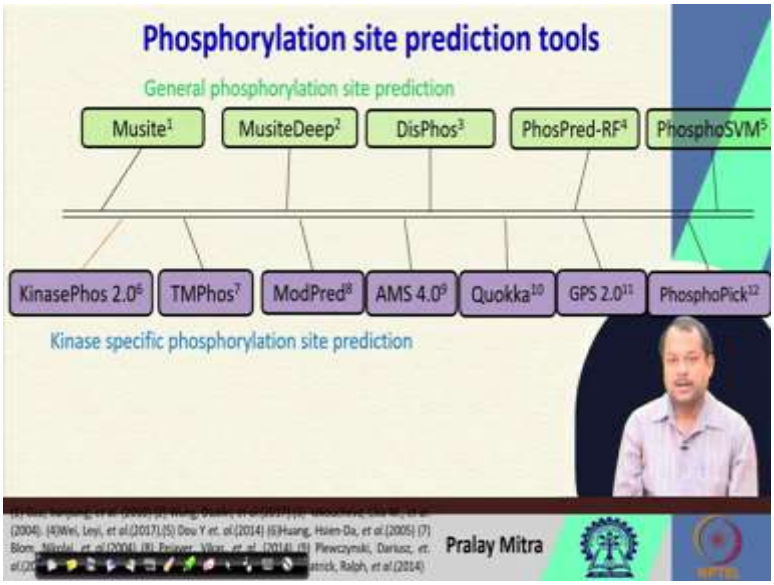
So, the biological insight that we got from here for this, from this discussion is that out of 20 amino acids, it is only the serine, threonine and tyrosine who are capable of protein phosphorylation or who can act as a site for the phosphorylation. Now, you see the biological and the insight from the biology and the chemistry combining together readily I am getting what, pruning down the search space. So, even say, if I am given say a protein molecule whose number of amino acids are say 100.

Now, if I assume that sequence is already given and from that sequence I need to predict the phosphorylation site, then you may go with the equal probability for all the 100 amino acids. That way if you go and try to design some computational framework, then that accuracy must be less than with the situation if I am using this particular domain knowledge from the chemistry and the biology that it is only the serine, threonine and tyrosine who is going to be the protein phosphorylation site.

Also, if somehow for the sequence itself I can predict that whether they will be on the surface or not, bingo, so added advantage. So, I can take a conjunction of those two or I can combine those two conditions and then I can further reduce my search space then probably out of the 100 amino acids for a protein so it can be say less than 10 or 15 based upon the kind of protein it is. So, then I have to search for 10 or 15 amino acids to check what is the phosphorylation site. That way you see a huge amount of noise you are reducing.

So, it has a huge advantage as you understand if you do it by yourself. So, I will tell you one algorithm. And based upon that one, if you wish to extend it without incorporating this domain knowledge then you will see the performance is bound to degrade. So, that is why we need to know little bit of chemistry and biology and if required then physics also along with the algorithm, knowledge of algorithm and implementation, then only we will go for a very good implementation which gives so accuracy and which will be fast enough. So, let us see what we can do with this one.

(Refer Slide Time: 28:22)



There exists a number of phosphorylation site prediction tools. So, it is not the new thing I am going to tell you, because people understood that there is a need for such a phosphorylation tool and that is why they started to work on this. It is not only phosphorylation tool, so there is a need of some computational alternative for almost everything if I can design that one, because for computation it is, once we will write that algorithm, implement that one, then the job will be run or program will be running in the system.

And with the advent of the high end computers, with the advent of the clusters, supercomputers and access for us to all those facilities will speed up our computation, and finally, allow us to give some filtered or some pruned solution on which the experimentalists can do the experiment and save a lot of experimental cost and time involved.

Now, when I am going for phosphorylation site prediction tool grossly they are categorized into two parts, general phosphorylation site prediction and kinase specific phosphorylation site prediction. In the kinase specific what are the different kinase that information is also integrated. And if I combine that then you can see that further I can prune down my search space. Those are say KinasePhos 2.6, TMPhos 7, ModPred 8, AMPhos 4.9, Quokka 10, so that superscript actually is the reference.

So, that is given at the end of the slide you can see. And GPS 2.0, PhosphoPik are the some of the kinase specific phosphorylation site prediction. On the other hand, general phosphorylation

site prediction tools are like Musite, MusiteDeep, DisPhos, Phosphate Pred-RF, and PhosphoSVM. So, that is it for this lecture. We will continue this to the next lecture. Thank you very much.