

Algorithms for Protein Modelling and Engineering
Professor Doctor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 31
Structural Classification of Proteins (SCOP)

Welcome back. So, today in this week we have couple of topics to cover. So, to start with, structural classification of proteins then we will discuss the symmetry in protein functional form. Then we will continue our discussion on discriminating crystal artifacts from the biological interfaces.

So, let us start this lecture with structural classification of protein in short which is called as the SCOP. Although there are two different ways to do this kind of classification, one is manually curated, another is automatic one, but we will cover manually curated one which is also hosted at the MRC laboratory at the University of Cambridge, and that database is also called as the SCOP that is the structural classification of proteins.

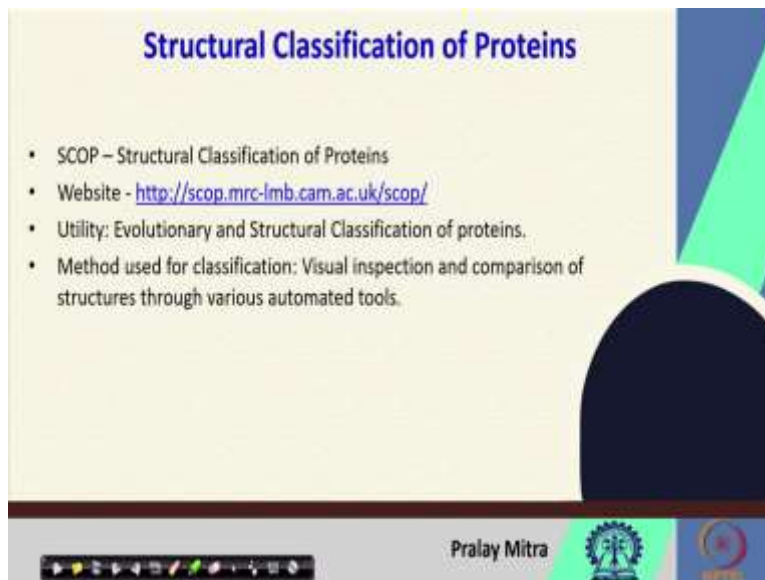
(Refer Slide Time: 01:08)





So, the topics we have planned to discuss is, the topic we plan to discuss is SCOP and that is why I am also picking the keyword as SCOP.

(Refer Slide Time: 01:19)



So, this structural classification of proteins are basically the classification of evolution information and structural similarity. What I wish to mean, that give a protein structure, so you will have a lot of, say, you will have a lot of different or variety of protein structures. Then the question or the motivation came long back that if we can group them together based upon some information or some knowledge.

Now, you know that nowadays with the advent of the machine learning technique and lot of computer-based algorithms and techniques so it is probably easy to classify the proteins. Definitely, you have to know what are the patterns and of course if you wish to apply the deep learning technique then you need not have to bother mostly about the features but you have to concentrate or focus on the architecture of your deep learning technique.

But when the idea came long back that we need to go for the classification then that much machine learning was not there, and also people trying to classify this not only based upon the structure but probably based upon the function. Now, this probably, the word I have used because of the fact that, as you know that there is a direct correspondence or direct mapping between the structure of the protein and the function of a protein, and the relationship is not that much direct between the sequence and the structure of a protein. So, we need to classify basically mostly based upon the structure and then we expect that then it will go for the classification by the function also.

Now, as of now we learnt number of techniques like TM-align or say rotation about an arbitrary axis. Based upon that we can align given two protein structures. Although we discussed that in the context of two protein structures or say two arbitrary structures but it can be, say customized for the purpose when in your protein molecule there are multiple, multiple chains. I mean that when there are multiple components which are not connected with each other then also we can apply that technique.

But the problem with those alignment is that when two protein structures and their underlying sequences has some differences. Specifically when sequences are not similar but the structures are tend to be similar then some problem may arise. What is the problem? Very simple. Say, sequences are not similar but the structures are similar.

Now, as I mentioned that there is a direct relationship between the structure and the functions, so the functions of those two proteins structures who, which are very similar in terms of the structure is going to be the same. But if their sequences are different; so in order to compute that whether they are similar or not after that alignment either based upon rotation about an axis or TM-align or any other technique if you come up with, then also you have to measure that whether they are deviating too much or not.

So, one technique maybe RMSD calculation or may be TM score calculation and based upon that you can compute that one. But you know that RMSD computation will actually give you some error and it is not a good idea to infer based upon that RMSD value. Say if a majority of the part is similar but few parts are very dissimilar in nature then that will bias the result. Also TM score can give you at the four level, and so that is just a measure. But based upon that one classification may not be that much easy.

So, that is why a long back one person at the MRC Cambridge started to look at the protein structure and manually curate the structures into different groups. And when he is dividing that into different groups, since it is the function which is of our interest so he also divided into hierarchical way which means it is not that, say K number of clusters are there and you are putting all the protein structures into K number of clusters. So, first let us divide by some patterns. Then inside the pattern there will be sub-patterns, like that way it will go, so in a hierarchical way. So, that hierarchical stuff I will also discuss.

Now, its utility is evolutionary and structural classification of protein. So, you will know that how they are related from evolution point of view or structural similarity is there. Now, the method used for classification is just a visual inspection and compression of structures through various automated tools.

So, first I would like to pinpoint this one. It is visual inspection. It is the visual inspection but this comparison of structures using automated tool came later. So, first it was started with the visual inspection.

(Refer Slide Time: 06:56)

The slide is titled "Structural Classification of Proteins". It contains a bullet point defining a protein family: "Family: Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%." Handwritten notes in red ink include "5 → 0/1" at the top right, "Mis match or Match" below it, "Visual inspection" on the left, and "Human interpretation" at the bottom left. A video inset shows a man, Pralay Mitra, speaking. The slide footer includes the name "Pralay Mitra" and two logos.

First it started with the concept of family. So, proteins are clustered together into families on the basis of one of two criteria that simply have, that imply they are having a common evolutionary origin. So, common evolutionary origin means you know that as per our definition that we discussed on the last week they are going to be homologous in nature.

So, first all proteins that have residue identities of 30 percent and greater. Second, proteins with lower sequence identities but whose functions and structures are very similar, for example, globins with sequence identity of 15 percent. So, first, all proteins that have residue identities of 30 percent.

Now, we demonstrated the dynamic programming algorithm through which either the global alignment or local alignment, but in this case I would prefer for global alignment. I can compute what will be the sequence identity, and since I am interested to compute the identity, so in my dynamic programming so it will be, the score value will be 0 or 1. So, this is mismatch or match. So, I am not considering (0)(8:23). So, the initial demonstration of the dynamic programming, so where score is either 0 or 1. So, that way if I compute then I will get that how much identity is there, and if the identity is greater than 30 percent it is fine.

So, then I will put them together and considered that part of family. Otherwise if proteins with lower sequence identity, I mean that less than 30 percent but their structures are similar again through the visual inspection then I can also put them together into the cluster. So, some manual

or visual or human interpretation is involved at the second part, when from the sequence point of view we are not getting it correctly.

(Refer Slide Time: 09:28)

Structural Classification of Proteins

- **Superfamily:** Families, whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, actin, the ATPase domain of the heat-shock protein and hexokinase.

HSP

Pralay Mitra


Next will come superfamily. Families whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable are placed together in super families. For example, actin, then ATPase domain of the heat-shock proteins and hexokinase. This HSP, this heat-shock protein also called as the HSP, it has a huge role in lot of application including the cancer.

So, after looking at the families then we are looking for a bit higher in the hierarchy where proteins with low sequence identity but whose structures, and in many cases, functional features suggest that a common evolutionary origin is probable. So, if it is so then we are putting them together into superfamily.


(Refer Slide Time: 10:30)

Structural Classification of Proteins

- **Common folds:** Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement with the same topological connections.
- **Class:** Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they composed.


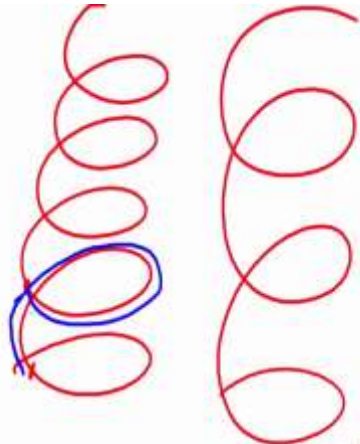


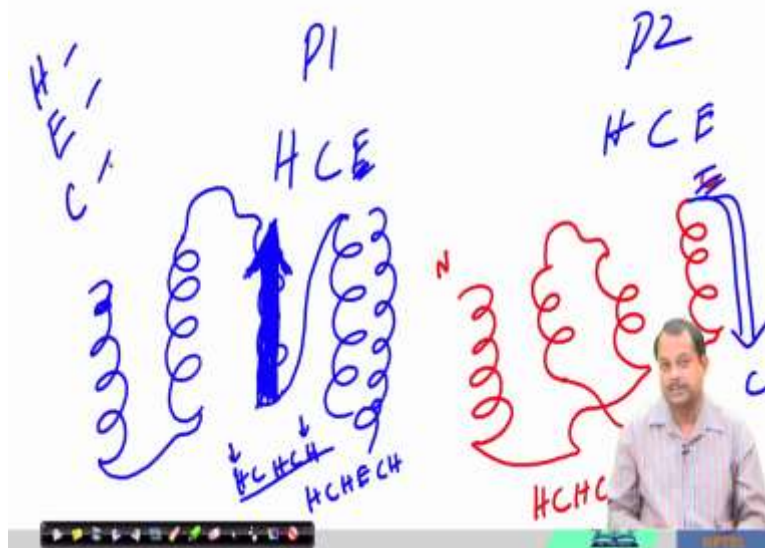
Pralay Mitra



Handwritten notes and diagrams illustrating protein secondary structures:

- Handwritten text: **Regular Secondary Structure**
- Handwritten text: **Helix (H)**
- Handwritten text: **Sheet (E)**
- Handwritten text: **Coil (C)**
- Handwritten text: **360°**





What next? So, common folds, super families and families are defined as having a common fold if their proteins have some major secondary structure in same arrangement with the same topological connections. So, please note it down, in the same, if their proteins having same major secondary structures in same arrangement with the same topological connections.

Let me elaborate little bit on this. So, how many different secondary structures we discussed? That is helix, sheet and coil, in short HEC. Now, among these, helix and sheet, they are called as a regular. Why? Because for helix and sheet we find some similarity and that similarity is in the pattern of the hydrogen bond formation. Since in helix and sheet there is some pattern in the hydrogen bond formation so some regularity exists. So, we call that as a regular secondary structure. Otherwise it is called as the coil which is non-regular.

Now, there are several variations of helix, say alpha helix, Pi helix, etc. For sheet also there are some variations. Even for coil, so turn, bridge, so based upon that there are several variations. But we will not go into details of that one.

So, mostly the variations arise because of the fact that when I say pattern, and then say for the helix also when I say pattern then the pattern can be something like this; it is a helix or it can be something like this. So, how many number of turns, per say, 360 degree, or say, starting from, this is a spiral, so if I consider this item and this item, so both are having same degree. So, there is a, there is a turn, sorry, it will be not this one. So, so from here it is, so from here it is taking this turn, so how many amino acids will be involved in this complete one 360 degree turn. Based

upon that one, whether it is a α helix or β helix that will help. So, that variations is based upon the kind of pattern through which we are working.

Now, if I go back to the definition it says that if the proteins have same major secondary structures in same arrangement with the same topological connections. So, first of all between two, in order to be in the same common fold, so what we need to do? First of all between two structures same major secondary structures should be there. Major means helix, sheet, coil so whatever is there, those common secondary structure will be there.

Now, if I consider that there are 2, say, there are 2, there are 2, so I need H, E, C. Now, if say I consider that there are two proteins, say P1 and P2, so one situation maybe all the secondary structures in P is either helix or coil, either helix or coil. If in both the cases if it is either helix and coil then that indicates that at the major secondary structure level these two proteins are same.

Now, after having that one, next point it says that in same arrangement, arrangement will be same and same topological connections which says that for P1, say if this is my secondary structure and there is a connection something like this, then in this case also it will be something like this. So, in the same C, with the same topological connections. So, arrangement is same, two helices, and topological connection is something like this.

But you see that at the length-wise there is some, there are some differences between these two proteins P1 and P2. But in this case we are interested to know that whether the major secondary structure-wise they are similar? Yes, they contains only helix or coil. And whether their arrangements are same? Yes, there are two helix on the two sides and in those two helices are connected by one coil. Whether their topology is same? Yes, helix coil helix is the topology.

Now, it can be complex in the sense that there can be, say another helix like this, there can be another helix like this, and then the connection can be say something like here and another connection may be like here. If it is then you see as per the connection, so first of all this is my N-terminus, and this is my C-terminus and between N and C, the topology says helix coil helix coil helix.

Now, similar to that in this case if I wish to draw so I have two, I have to draw three, two more. Now, say something like this. Now, if, now you see that if I connect similar to like this red color protein P2 then their topology will also be same, like H C H C H major connections. But if the situation is something like, so this is connected with this and this is connected with, say this, then the question will come that whether their topology is same? Because you look at the spiral. So, it is like clockwise this is anti clockwise. This is anti clockwise. This is anti clockwise. Here clockwise, anticlockwise, anticlockwise and anticlockwise, so that way it is same.

But when I say this H C H C H then this one is anticlockwise, this one is clockwise, and this is anti, anti, anti. Here clockwise, anti, anti, anti. Now, if it is the case that this guy is not like this but like this then also you will get H but this guy then will be say, clockwise. Then from the topology point of view there will be some differences between P1 and P2. Whether that will hinder them to be part of the same classes or not, that is a different question. But from topology point of view there will be some differences.

Also if say along with this H C, some E is introduced here then definitely at the major secondary structure level also there is some changes and at the topology level also there might be some changes. Say for example because of this E and let us assume that this is actually not helix, it is rather a sheet. The sheet is denoted something like this. So, it is a thick arrow. So, if it is then you see that it will be say H C H E C H, whereas it is H C H C H. So, From topology point of view it is different, as well as, so one, say sheet has been incorporated.

Now, if I write, say here E also and, say I did one modification that after, this is not C-terminus, after this there is a sheet here, and this is my C-terminus, not this one. Then this topology will be H C H C ; H C H C H E and here it is H C H E C H. So, their topology will be different. And if it is, then at the fold level also they might be different. Accordingly their structure will, accordingly their function will also is going to change.

Now, this we are looking at the secondary structure level, where it is represented by H C or E. Now, if I replace it or if I look from the atom point of view then perhaps we will see that probably their structures are same, but not from the topology in this case.

(Refer Slide Time: 20:01)

Structural Classification of Proteins

- **Common folds:** Superfamilies and families are defined as having a common fold if their proteins have same major secondary structures in same arrangement with the same topological connections.
- **Class:** Most of the folds are assigned to one of the five structural classes on the basis of the secondary structures of which they composed.

Pralay Mitra

Next is class. So, most of the folds are assigned to one of the five structural classes on the basis of the secondary structure of which they composed of. So, basic secondary structures that we mentioned as helix, sheet or coil. Mostly the emphasis will be given on helix and sheet because there are a lot of variations and a number of organizations may you give different possibilities. So, at the class level most of the folds are assigned to one of the five structural classes on the basis of secondary structure of which they composed of.

So, what we are getting now? Family first, super family, common folds and fold. So, four different levels we are considering now, and based upon that one shortly will see that the classification will have also 4 different levels.

(Refer Slide Time: 21:07)

SCOP class

1. All alpha-(for proteins whose structure is essentially formed by α -helices), $\alpha \rightarrow a$ H
2. All beta (for those whose structure is essentially formed by β -sheets), $\beta \rightarrow b$ E
3. Alpha and beta (for proteins with α -helices and β -strands that are largely interspersed), $\alpha \cdot \beta \rightarrow a \cdot b$
4. Alpha plus beta (for those in which α -helices and β -strands are largely segregated) $\alpha + \beta \rightarrow a + b$

Pralay Mitra

So, SCOP class, the first category is all alpha for proteins whose structure is essentially formed by alpha helices. So, that is considered as one class. Next is all beta for those whose structure is essentially formed by beta sheets. Now, please note it down that again we are not differentiating between different kind of helices like pi helix, alpha helix etc. So, we are considering all are similar helix, and as per our understanding we mentioned this is H, the first one and second one is E.

Next, alpha and beta for proteins with alpha helices and beta strands that are largely interspersed, so please note it down this word because one new thing again will come. So, this says alpha and beta. And the next thing says alpha plus beta, so for those in which alpha helices and beta strands are largely segregated. So, one is interspersed, another is segregated.

So, what, how you can represent that one? So, since it is alpha helix, so you can represent as alpha. This you can represent as beta. This you can represent as alpha and beta, so borrowing the notation of, say Boolean logic alpha dot beta and alpha plus beta. It is simple. Now, this thing you can further change because you see that I wrote as alpha beta, so these are Greek letters. Now, when I was to type so then it may be convenient that if I convert these Greek letters to some, say, to some English characters. So, this alpha I can convert to A because alpha starts with a, beta b. Then accordingly it will be a dot b, this will be a plus b. This also you can think of and you can write.

So, at the first level I am dividing the total structures into four classes where in a protein only the alpha helices are present. Please note that it will not be possible if the protein size is large that only the helices will be there. So, multiple helices may be connected by coils so that non-regular part we are not considering explicitly but we are assuming it will be there. So, that is kind of connectors and those connectors will be there.

But the first category says that there is no helix, sorry there is no beta. So, it is only the helix. The second category says that there is no alpha. There is only helix, there is only beta. So, first one is consist, one protein structure that consists of only alpha helices or helices, but they are connected, so if there are multiple helices they will be connected by some coils.

Now, all beta, all beta indicates that in that protein structure only the beta sheets are present and they are connected by coil if required. But in first category there will be no beta sheet. In the second category there will be no helix, alpha helices. The third and fourth category indicates that alpha beta both will be present there and again they will be connected by some coils which I mentioned as connectors in our case.

Now, in third case alpha and beta are interspersed, and in the fourth case alpha and beta are largely segregated. That way one is alpha dot beta. Another is alpha plus beta. When it is segregated alpha is in one side, beta is another side so it is plus. Another interspersed so alpha dot beta.

(Refer Slide Time: 25:41)

The slide is titled "SCOP class" in blue text. It contains a numbered list of five categories:

1. All alpha-(for proteins whose structure is essentially formed by α -helices),
2. All beta (for those whose structure is essentially formed by β -sheets),
3. Alpha and beta (for proteins with α -helices and β -strands that are largely interspersed),
4. Alpha plus beta (for those in which α -helices and β -strands are largely segregated)
5. Multi-domain (for those with domains of different fold and for which no homologues are known at present).

On the right side of the slide, there is a circular video inset showing a man in a light-colored shirt speaking. At the bottom of the slide, there is a navigation bar with icons and the name "Pralay Mitra" next to a logo.

Final thing is multi-domain. We will not discuss much about this multi-domain. It says that when we will find there are more than one domains, so the definition of domain is a different issue that we will discuss later, so if there are multi-domains and clearly you can visualize that, say one part is here, another part is here and they are loosely connected through some coil so that they can make some move and if I consider the topology between these two domains or when there are multiple domains.

So, in other way when one protein structure is given to you and I can specifically identify, okay this part is probably, this part is in one region, this is another region and these two regions are connected by some coils and the coil is very long in nature and that is why they can take some different shapes.

So, based upon the fluctuation of the coil so that indicates the multi-domain. In that case we need to consider that separately because it may possible that one domain, say consist of only alpha, another will consists of only beta or here alpha beta is interspersed, here only beta is present, so different variations may exist. And from biology point of view when it is multi-domain it says that that particular protein must have multiple functions. And function is attached with some domain. So, it is not like corresponding to one protein. So, it is not true.

So, it is noted also during the experiment and people understand now that one protein may have multiple functions because of multiple environment. So, if the, say solution changes, if the

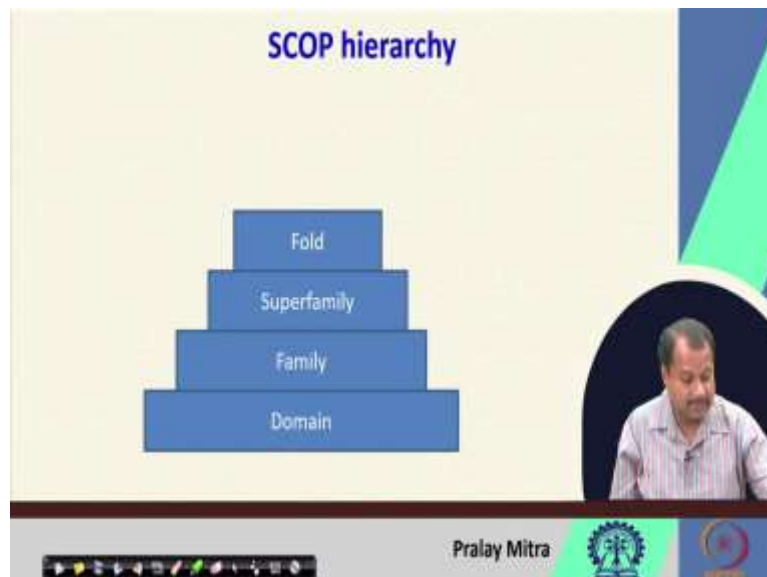
temperature changes, if pressure changes, if say ionic concentration in the solution changes, based upon those, you remember those factors or features we discussed in the context of protein folding.

Now, in the context of the protein folding when we discussed this, when we discussed, then all those features are responsible for the protein folding; which means that if I change one feature then it will take some another fold or another structure. So, that way, so it is possible that their function will be also different.

Now, for the multi-domain so it is not like that only. So, apart from that each part I can visually understand and see that, okay they are different, and because they are different they have some different function. So, those two things are different. So, one protein may have multiple functions based upon different environment features; that is one situation.

Another situation is that one protein may have multiple functions because it has multiple domains inside and each domain will have one function. Third kind is also possible. So, based upon that say, apart from the environment so one protein may have multiple functions at different time frame, so simultaneously or mutually exclusive, in a mutually exclusive way those things are possible.

(Refer Slide Time: 29:09)



So, in the SCOP hierarchy, so at the bottom it is the domain. So, if it is multi-domain so as I mentioned that it may possible one domain will have, say alpha. Another domain may have beta or it may have alpha beta interspersed, it may have alpha beta segregated. So, since those possibilities are there, so it is always a good idea that either you consider multi-domain separately, or for each multi-domain you chop it, or so you consider each domain separately.

But if it is not multi-domain it is a single domain, so only one domain is there. That is going to be at the bottom. On top of that one there will be the family classification. Now, definitely the number of proteins which has the same, which, for which we have identified the domain, so when we classify into family then it will be reduced in the numbers, I mean number of families will be less. Number of super families will be more less. Number of folds will be more less and that is the different hierarchy. So, domain, family, superfamily and fold; that is what we discussed.

(Refer Slide Time: 30:14)

The slide is titled "SCOP - new developments" in blue text. It contains three bullet points defining the terms:

- **Species**, representing a distinct protein sequence and its naturally occurring or artificially created variants;
- **Protein**, grouping together similar sequences of essentially the same functions that either originate from different biological species or represent different isoforms within the same organism;
- **Family**, containing proteins with related sequences but typically distinct functions;

On the right side of the slide, there is a circular video inset showing a man in a striped shirt, identified as Pralay Mitra. At the bottom of the slide, there is a navigation bar with various icons and the name "Pralay Mitra" next to two logos.

So, there are few new developments also for the SCOP class, species which represents a distinct proteins sequence and its naturally occurring or artificially created variants. Proteins, grouping together similar sequences of essentially the same functions that either originate from different biological sequences or represent different isoforms within the same organism. And family, it containing proteins with related sequences but typically distinct functions. So, thank you. We will continue this discussion in our next lecture also. Thank you very much.