

**Algorithms for Protein Modelling and Engineering**  
**Professor Doctor Pralay Mitra**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**  
**Lecture 32**  
**SCOP (contd.), Symmetry in proteins**

Welcome back. So, we are discussing SCOP structural classification of proteins. So, that we will continue. So, after finishing the discussion on the SCOP we plan to show you the website from where we can get this information, kind of a tutorial, short tutorial and after that we will move on to symmetry in proteins. That is an interesting topic.

(Refer Slide Time: 00:34)




So, the concept will cover, the SCOP that we are continuing and the symmetry in proteins.

(Refer Slide Time: 00:40)

### SCOP – new developments

- **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;



Pralay Mitra

### SCOP – new developments

- **Superfamily**, bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor.
- **Folds**, structurally similar superfamilies with different characteristic features;
- **Class**, mainly on their secondary structure content and organization.

*Handwritten notes:*  
 $\alpha$   
 $\alpha \cdot \beta$   
 $\beta$   
 $\alpha + \beta$

*Handwritten diagram:*  
N — (H<sub>1</sub>) — (H<sub>2</sub>) — (H<sub>3</sub>) — C  
with "long loop" written above H<sub>2</sub>



Pralay Mitra

So, we finished our discussion on this particular topic in the last lecture, that SCOP new developments. So, species which was incorporated, so initially the first structure of the SCOP was, at the bottom there was domain. After that one that was superfamily, family, fold. That way it goes.

Now, additional things with the new development, since lot of protein structures are coming up and they are deposited in the Protein Data Bank, so different variations has been noted. A lot of variations are there. So, we try to cover most of them. And among them one is a species, that

represents a distinct protein sequence and its naturally occurring or artificially created variants. So, based upon that one we can categorize.

Then based upon the protein, grouping together similar sequences of essentially the same functions that either originate from different biological species or represents different isoforms within the same organism. So, this is important also, because when say, we identify two proteins it may possible that they are coming from the two different biological species but their functions are same. So, if it is then it is always a good idea, put them together because they are functionally also going to be the same.

So, that is why at the protein level, and along with that one family is always there, that containing proteins with related sequences but typically distinct functions. Then our superfamily bridging together protein families with a common functional and structural features inferred to be from a common evolutionary ancestor, folds that we discussed on the last lecture, structurally similar superfamilies with different characteristic feature, and the class mainly on their secondary structures content, and the organization means topology.

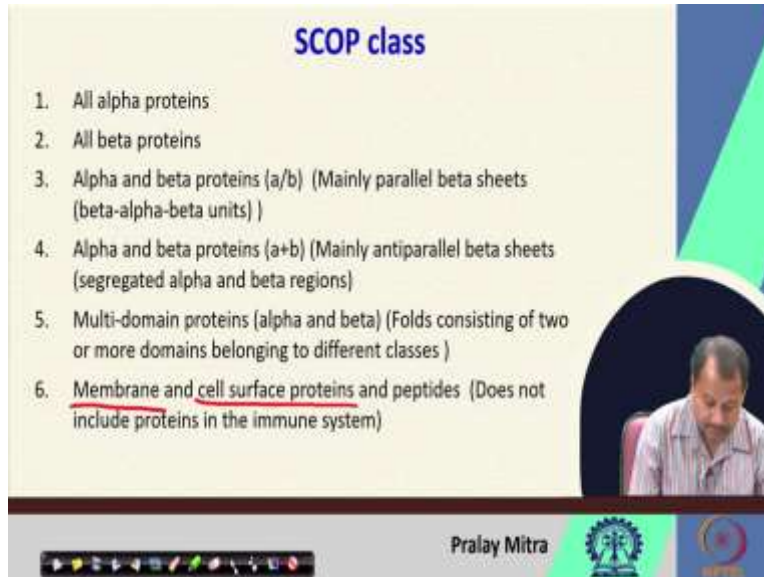
What is their organization? When they will fold and which kind of structure it will take? Say, for example at the fold level or say, at the secondary level, structural level you can mention on the last lecture also, I, okay, sorry. So, on the last lecture also I mentioned that, so secondary structure level there can be H, there can be only, there can be H dot E or say if I write in that form that I mentioned last time that is alpha and beta form that is preferable even in the context of the SCOP. Do one is alpha, another is alpha dot beta, then beta only, alpha plus beta. So, along with C or coil will be there definitely.

But, so mainly secondary structure content and the organization, the organization says that when it is alpha also, so the organization is. What is their organization? Means if I identify there are three helices, so H1, so H1 H2 H3. Now, in one organization it can be this H2 is the largest or say longest as per the sequence. In another organization it could be the H1 is going to be the longest or largest.

Now, if I say that H1 is connected with H2 and then it is connected with H3 then this is going to be my C-terminus, this is going to be the N-terminus. So, H1 H2 H3, in one situation H1 is the largest and its length is very high compared to H2 and H3. In another case it is different, like H2

is the longest and its length or its structure is huge compared to say H1 or H3. So, from that point of view some organization changes are there and that will be captured at the class level.

(Refer Slide Time: 04:42)

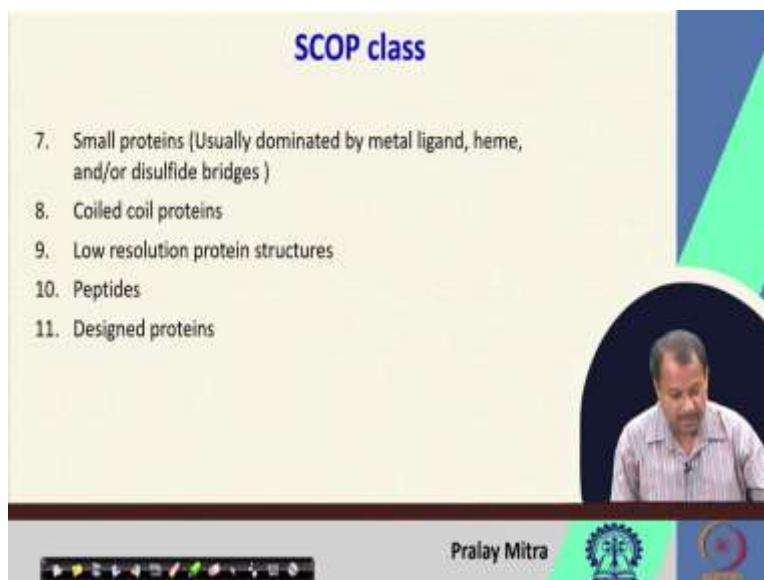


**SCOP class**

1. All alpha proteins
2. All beta proteins
3. Alpha and beta proteins (a/b) (Mainly parallel beta sheets (beta-alpha-beta units) )
4. Alpha and beta proteins (a+b) (Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) (Folds consisting of two or more domains belonging to different classes )
6. Membrane and cell surface proteins and peptides (Does not include proteins in the immune system)

Pralay Mitra

This slide shows the first six classes of the SCOP classification. It includes a list of protein classes and a small video inset of the speaker, Pralay Mitra, in the bottom right corner. The slide has a navigation bar at the bottom with various icons and the name 'Pralay Mitra'.



**SCOP class**

7. Small proteins (Usually dominated by metal ligand, heme, and/or disulfide bridges )
8. Coiled coil proteins
9. Low resolution protein structures
10. Peptides
11. Designed proteins

Pralay Mitra

This slide shows the last five classes of the SCOP classification. It includes a list of protein classes and a small video inset of the speaker, Pralay Mitra, in the bottom right corner. The slide has a navigation bar at the bottom with various icons and the name 'Pralay Mitra'.

Now, the SCOP class, redefine SCOP class. Initially what I mentioned? All alpha class, all beta class, alpha and beta, alpha plus beta so those four classes; so alpha, beta, alpha dot beta, alpha plus beta. Now, the extended or new development incorporating that one; all alpha it is there, all beta it is there, alpha and beta so that slash or dot mainly parallel beta sheets, beta alpha beta units, another is alpha and beta proteins a plus b, mainly anti parallel beta sheets segregated

alpha and beta regions, multi-domain proteins that also we discussed, alpha and beta folds consists of two or more domains belonging to two different classes.

Next, membrane and cell surface proteins and peptides, does not include proteins in the immune system. So, this membrane protein and the cell surface proteins, so they are of special type or kind of proteins. So, specifically these membrane proteins it contains lot of hydrophobic residues and sometimes it is difficult to manage.

So, throughout our discussion mostly we will focus on globular proteins rather than the membrane protein. But for the completion of the structural classification of proteins or SCOP class it is included and also you can see from here they are classified into a separate class because their properties are completely different from the properties of the globular proteins.

Small proteins usually dominated by metal ligand, heme and, or disulfide bridges. Small proteins indicates you can consider those are generally called as the peptide whose length you can consider, say atmost 20. So, 10 to 20, sometimes small, small proteins.

Now, you can understand that if there are only, say 5-6 amino acids or when I say that say it is a peptide of length, say 10, means 10 amino acids then it might be difficult, and yes it is, to have a particular fold corresponding to that 10 amino acids. And also you can understand that when the minimum number of residues which are required to form a pattern either as a alpha helix or beta sheet is not present then we cannot classify that as part of say, first four categories; all alpha, all beta, alpha and beta, alpha or plus beta. So, we cannot classify that one.

So, since we cannot classify that one so it might be good idea that if we take those and keep in a separate class like we have done for, say multi-domain protein or main, main protein because they are of separate class. So, keep them separate. So, that is the small proteins. Along with that one, coiled coil proteins which consist of no regular secondary structures. So, that is also separate class. Take it out and keep separate.

Low resolution protein structures, as I mentioned that structures are mostly deposited in the Protein Data Bank. And yes, if you wish to decide, if you wish to experimentally determined some structure and wish to get one publication out of that one, first thing you need to do, you have to deposit that structure in the RCSB or Protein Data Bank. So, if you deposit, then from

there you will get one accession ID which is basically the PDB ID. So, using that only you can go for publication.

Now, if you look at the Protein Data Bank and if you look for different experimental techniques you will find variety of experimental techniques are there in order to determine the protein structures at the atomic level resolution. So, most widely and mostly the structures are determined through X-ray crystallography. But solid crystallography, NMR, solid NMR and nowadays lot of structures; not nowadays, it has started long back actually, cryo-electron microscopy Cryo EM. So, using that also a number of structures has been deposited.

But when I am talking about say, this crystallography then the resolution with which I determine the structure varies. It varies a lot. So, in earlier days, say 80s or 90s when sophisticated instruments and software was not there to get the structure and then refine that structure with the atomic coordinate so the resolutions was very poor. So, 5 Angstrom, 6 Angstrom, etc. So, with that resolution the structure which was determined may have some artifacts. So, it is not possible to have a very good quality structure.

So, you know this is very much related with the precision of some experimental technique I am talking about. Now, if we do not have a high resolution or good resolution structure so keep them aside.

Nowadays a number of alternative techniques are used to get more accurate structure of those and when somebody will get that one then the previous or old data has been replaced. But until it is replaced you keep them aside because they are prone to some error and since through some visual inspection or if I go for some semi-automated technique also, if I go and get the structure then for the low resolution structure some error may happen. That is why you keep them separate.

Next is peptide. This is also similar to the 7 but it is rather very, very small compared to that small protein. And finally the designed protein. For the last 20 years people are designing new proteins with new functionalities and they are also depositing that one in the Protein Data Bank.

Now, when they are designing, of course after designing that protein they are getting the structure of that one experimentally, then only they are depositing to the PDB data bank but when it is designed which means it is not naturally occurring. Some mutations has done or some

insertion deletion operation that I mentioned in detail regarding this protein design we will discuss later.

But it is not available in the nature, and in some computational way or some experimental way first the sequence has been realized. And then based upon that sequence, the experimental technique like crystallography or NMR or other technique, so it has been identified, I mean the atomic resolution of that particular protein structure has been identified. So, those are designed proteins. They are separate category.

So, personally I do not feel that they should go separately in some another category because they are indeed the same structure like any naturally occurring proteins. But you know sometimes, a single point mutation may create some lot of changes in the function may not be affected in the protein structure.

So, in order to take care of those facts, so it might be good idea that first you will take since it is designed, so you take it in separate class. And when you will look at those proteins in that particular class then you redistribute that among other classes. So, that can be done.

(Refer Slide Time: 12:19)

The slide features a title "SCOP class - an instance" at the top. Below it is a bulleted list: "The seven main classes in the release 1.73 contain" followed by four sub-points: "92,927 domains organized into", "3,464 families,", "1,777 superfamilies and", and "1086 folds." Below the list, it states "The SCOP domains correspond to 34,495 entries in the Protein Data Bank (PDB)." A video feed of a man speaking is visible in the bottom right corner of the slide area. At the bottom of the slide, there is a navigation bar with the name "Pralay Mitra" and two logos.

So, one, an instance, so do not consider this as a final, so this stat was taken long back so it is release 1.73. Now, a lot of release has come even the SCOP is now with SCOP 2 actually. The release 1.73 content 92927 domains, organized into 3464 families, 1777 superfamilies and 1086

folds. The SCOP domains corresponds to 34495 entries in the Protein Data Bank. Now, this number has gone up.

So, when I am recording this lecture it is about 165000 and it will keep on gradually increasing. And accordingly this will also increase. I mean the entry at the SCOP will also keep on increasing but the rate at which your PDB is increasing may not be same at which SCOP is increasing because it is manually curated. Although nowadays a semi-automated technique has come up to take care of some of the new structure proteins, so who is homologous to some existing proteins available in the SCOP class as well as the PDB.

(Refer Slide Time: 13:44)

The slide features a title "SCOP - current semi automatic update protocol" at the top. Below the title, there are several handwritten notes in red ink. On the left, "CATH" is written and underlined. In the center, "PSI BLAST" is written and underlined, with a downward arrow pointing to "Position Specific Iterative". To the right, there are two numbered items: "1) Sequence Alignment" and "2) Structure Alignment", both underlined. A small video inset in the bottom right corner shows a man speaking. At the bottom of the slide, there is a name "Pralay Mitra" and two logos.

So, the current semi-automated update protocol, what it used to do? That given one protein structure, so sequence is also known to you, so it first tries to check using some sequence alignment. So, one popular sequence alignment technique you can consider as PSI-BLAST, this is the position specific BLAST. So, this BLAST is basically one heuristic for sequence alignment that we did not discuss. So, this PSI-BLAST is one variation of that.

So, using that one you first check that whether there is any sequence similarity of a new candidate which is deposited in the Protein Data Bank with the already curated data in the SCOP. So, if yes, then you try to put it in that category. So, based upon one threshold that you will mention, that say 70 percent or so. If not then, if it is in the gray region, means maybe something like this but not sure, say between 20 to 70 percent then you do, say structural alignment.



After doing the structural alignment you will get at the fold level accuracy, say you can use TM-Score or RMSD in order to know that what is the fold level accuracy. So, based upon that one also you can infer that way, and using this, so one is the sequence alignment, then structural alignment, so all these we have discussed.

You will get, so most of the new structures as, which are similar to already curated data in the SCOP database. So, that way you can reduce your new data significantly for your manual curation. But who will not pass this sequence alignment or structural alignment, then keep them separate so that you can basically, keep them separate so that you can basically manually curate and classify them into proper class in the SCOP.

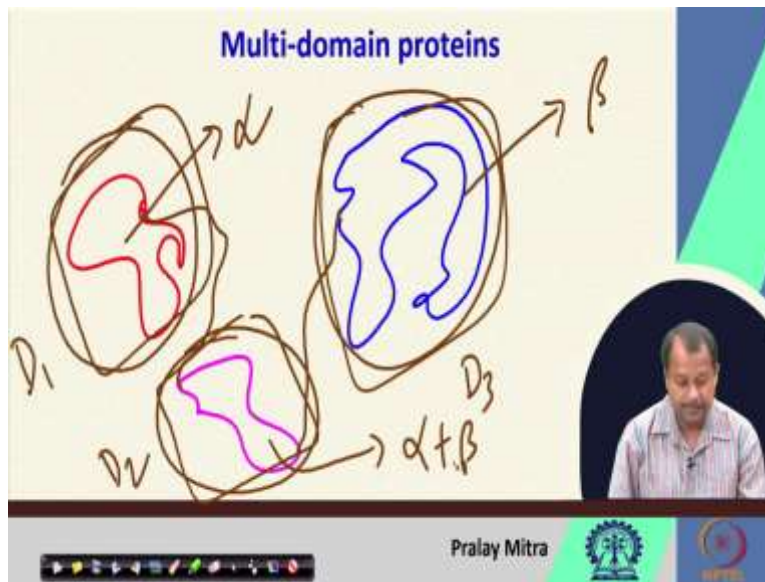
So, there is one semi-automated update protocol in place. So, I am not going to detail that one. Rather I leave it up to you to come up with your own idea which can be incorporated so that who will take one new protein structure from the PDB as the input and will output what will be the SCOP class.

I understand that there may be some cases where there will be one exit option, I mean that I cannot able to classify that into a proper SCOP class. So, let us come out of that one. So, once it will come out of that one then definitely I will take care of that manually. But as you understand that lesser and lesser number of such exits will happen, lesser and lesser manual intervention or human intervention is required. But rest of the part will be done automatically.

Although there is an separate algorithm or method in order to automatically annotate the data which is called as the CATH, we are not discussing that one because accuracy of SCOP is far better compared to that one. That is purely automated in nature but this SCOP requires some manual intervention.

But what I am interested to do, that come up with your own idea so that taking one protein you try to classify into SCOP classes. If you fail to do that one you keep it separate. But that separate set must be small enough so that minimum amount of human intervention is required.

(Refer Slide Time: 17:50)



So, the multi-domain proteins so mostly I discussed, so but I wish to draw one structure for you, say something one domain, another domain, another domain and say they are connected something like this. So, this is one domain. This is one domain. This is one domain. So, D1, say D2. Now, from here also you can see that this may have some specific function. This may have some specific function. This may have some specific function, so separate functions these may have but collectively they can, or they may, may not be any particular function.

Now, when I am classifying that one it may not be a good idea to put everything as one class, because this part can be very much consist of only beta, this can be only alpha, this can be say, alpha plus beta. It may be possible. So, that way they can be of separate domains.

The major challenge with this multi-domain protein is that identifying the number of domains and what is the starting point and what is the end point of the domain. So, it is very difficult. So, few algorithms are coming up which tries to predict the multi-domain proteins but sometimes they are not sufficient enough to classify, so to identify.

Now, if you can identify for surety then it will go to your previous slide where I requested you for a semi-automated tool which will take one such domain as an input and we will classify what is its SCOP class.

(Refer Slide Time: 19:49) 21:36



So, let us go for a tutorial. So, as of now what we understand that there are different classes. And I also mentioned that, if you remember correctly, that at the bottom actually there are domains, family, superfamily and folds. So, for each 1, 2, 3, 4 I will have one classification ID. So, this will be an ID. This will be one ID. This will be one ID. This will be one ID. So, please make a note of this.

Now, if this ID is at the very beginning what was my proposal? That alpha, beta, alpha plus beta, alpha dot beta, so four categories. So, I can write that way as one of the classes. Then for superfamily I can have some naming convention for family, for domains. And each one I can write in a separate way.

So, this dot this dot this dot this. Then this is going to be my entire SCOP class corresponding to one domain or one protein if it is single domain. So, this is one. That is my fold. This is my superfamily, family and domain. So, fold, superfamily, family and then domain. So, this you remember.

Now, we are going back to the tutorial. So, online I already opened one, I already opened one page for this SCOP. But what you can do, that you can basically, what you can do? That you can Google it.

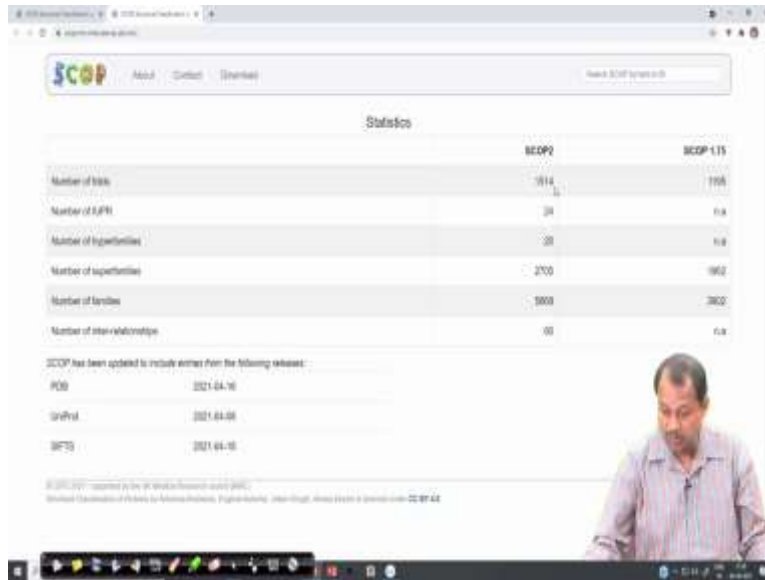
(Refer Slide Time: 22:07)

The screenshot shows the SCOP 2 website interface. At the top, the title "SCOP 2" is displayed in a large font, with a "Learn More" button to its right. Below the title, the text "SCOP: Structural Classification of Proteins" is shown. A paragraph of introductory text follows, explaining the database's purpose. A key statistic is highlighted: "Latest update on 2021-04-23 includes 68,482 non-redundant domains representing 765,687 protein structures. Folds, superfamilies and families statistics here." Below this is a search bar with the text "Browse and search" and a "Go" button. Two main navigation sections are visible: "Browse by structural class" with a list of categories (All alpha proteins, All beta proteins, Alpha and beta proteins(αβ), Alpha and beta proteins(αβ), Small proteins) and "Browse by protein type" with a list (Globular proteins, Membrane proteins, Fibrous proteins, Non-globular intrinsically unstructured proteins). A small inset video of a man speaking is visible in the bottom right corner of the webpage. At the bottom of the screenshot, a Windows taskbar is visible with the system clock showing 12:47 PM on 02/04/2021.

So, in Google what you have to write? SCOP MRC, because it is actually in the MRC, Cambridge. And if you note it down scop dot mrc hyphen lmb dot cam dot ac dot uk. So, that is basically the ID.

Now, this is SCOP 2 class. Now, if I scroll down so there is an option for download from here directly from contact. So, here and about. Now, directly let us come down here. One thing you should note. Let us update on 2021 23rd of April, so quite updated it is. Now, it includes 68482 non-redundant domains representing 765687 protein structures. So, this is the current statistics corresponding to the SCOP class. Now, folds, superfamilies, families statistics are here. If you are interested about that statistics you can open that.

(Refer Slide Time: 23:02)



	SCOP2	SCOP 1.75
Number of folds	1514	1195
Number of AFMs	38	n/a
Number of superfamilies	38	n/a
Number of superfamilies	2705	1962
Number of families	3800	2802
Number of interrelationships	60	n/a

SCOP has been updated to include entries from the following releases:

PDB	2021-04-16
UniProt	2021-04-08
SWISS	2021-04-16

© 2021 SCOP, created by the Swiss Research Institute for Protein Sciences, University of Zurich, All rights reserved. SCOP is licensed under CC BY 4.0

So, in a new tab I am opening. Number of folds SCOP 2, so 1514; in SCOP 1.75 it was 1195. Actually I, in my lecture I shown you 1.73 version that was little lesser than this. Now, then superfamily 2705, families 5668, and the number of interrelationships it is now 60 but it was not present there. SCOP has been updated to include entries from the following release. So, PDB, so 16th April PDB up to that one it is updated.

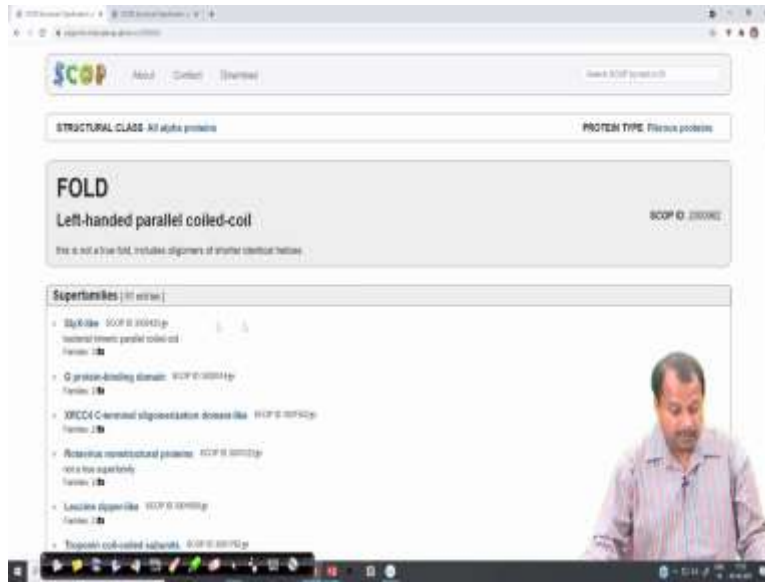
Now, let us browse, browse by structural classes. As I mentioned, so on the top, at the fold level you have all alpha proteins, all beta proteins, alpha and beta proteins so this slash or dot you can use, alpha and beta proteins, alpha plus beta and small proteins. So, protein type; globular, membrane, fibrous, non-globular, etc is there. We are not interested on the right hand side, on the left hand side only. So, let us open all alpha proteins.

(Refer Slide Time: 24:07)

The image displays two screenshots of a web browser showing the SCOP database interface. The top screenshot shows the 'STRUCTURAL CLASS' page for 'All alpha proteins' (SCOP ID: 18000). The page lists 455 entries under the heading 'Fold | 455 entries'. The bottom screenshot shows a list of protein folds, with 'Hemerythrin-type up-and-down 4-helical bundle' (SCOP ID: 0000001) selected. The list includes various fold types such as 'Left-handed parallel coiled-coil', 'Single transmembrane helix', 'Left-handed antiparallel coiled-coil', 'Long alpha helix', 'alpha-alpha superhelix', 'Hemerythrin-type up-and-down 4-helical bundle', 'DRD3-type 3-helical bundle', 'Specific repeat-like', 'SAM domain-like', 'Strand-domain-like', 'Interpenetrating alpha-helical bundle-like', 'Non-globular alpha-helical subunits of globular proteins', 'USA-type 3-helical bundle', 'Transmembrane helix bundle', and 'Coiled beta-helical up-and-down bundle'.

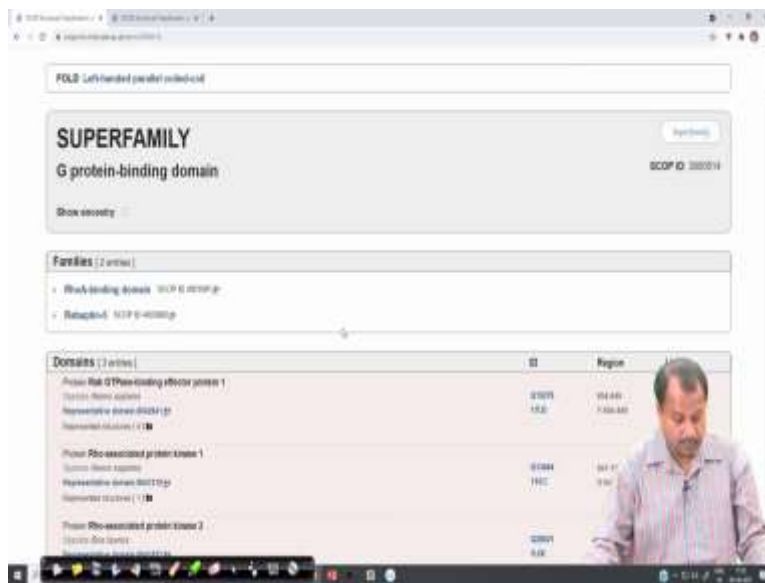
In this case as I mentioned it is going to be consisting of only helix connected by some, connected by some coil it may be. Now, all alpha proteins at the fold level, so there are 455 entries that you can see. Now, we are opening one such, so with the superfamily ID 61 here.

(Refer Slide Time: 24:33)



Then at the superfamily we have 61 entries means 61 different superfamily classes under this all alpha protein class and protein type fibrous proteins is there. Then G protein-binding domain I am opening.

(Refer Slide Time: 24:53)



Then when I open, then under that, so, so fold, superfamily is G protein-binding domain. Then family, under that one there are two entries or two classes for the family; Rho A binding domain and Rabaptin-5; then after that one, inside that there are 3 domains; Rab GTPase-binding effector

protein 1, Rho-associated protein kinase 1, Rho-associated protein kinase 2. Now, if I open this one then I will get the details.

(Refer Slide Time: 25:27)

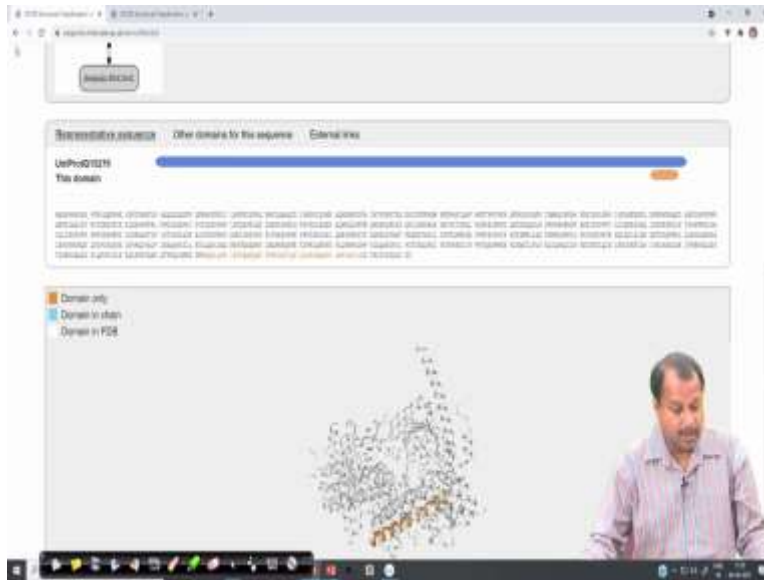
The screenshot shows a web browser displaying a protein domain page. The main heading is "DOMAIN" with the identifier "1TU3 F:804-849" and "OCOP ID: 334291". Below this, there is a section for "Representative sequences" with a blue bar representing the domain. A "Show Ancestry" button is visible. A video inset in the bottom right corner shows a man speaking.

So, what is the domain, how it looks like? And if I click here for showing the ancestry so what I can see?

(Refer Slide Time: 25:33)

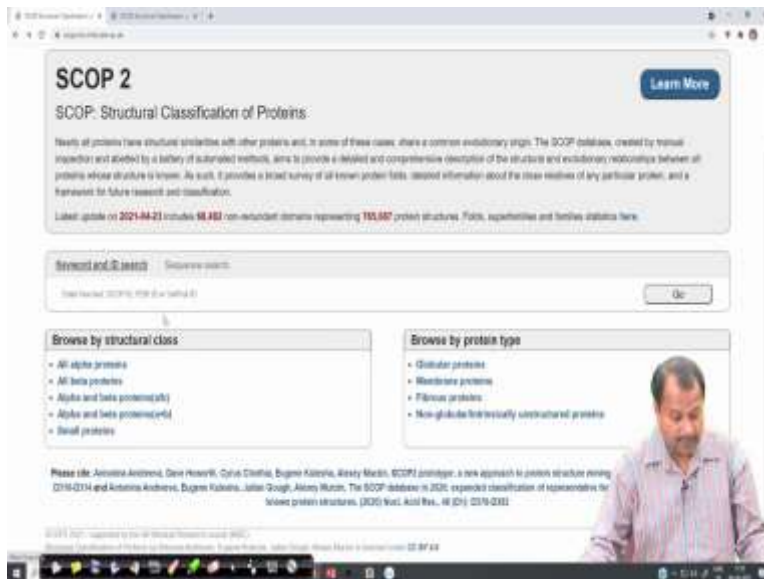
This screenshot shows the same protein domain page as above, but with the "Show Ancestry" button clicked. An ancestry tree is displayed, showing a sequence of domains: "1TU3 F:804-849" at the top, followed by "1TU3 F:804-849", "1TU3 F:804-849", "1TU3 F:804-849", and "1TU3 F:804-849". A video inset in the bottom right corner shows the same man speaking.





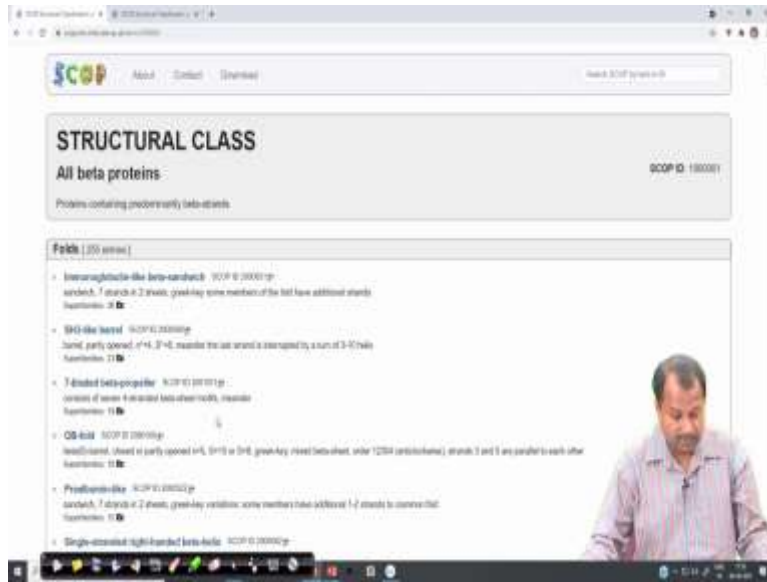
It is domain, superfamily, fold, class. So, those things that I discussed, four levels, that is here now. And if I look at this structure it is all alpha. It is all alpha. That you can see here, and they are connected by some coils. So, definitely there will be some coils. Otherwise how the connection will be?

(Refer Slide Time: 26:18)



Now, if I go back and check for something else, say if it is not, say all alpha classes, say it is all beta protein.

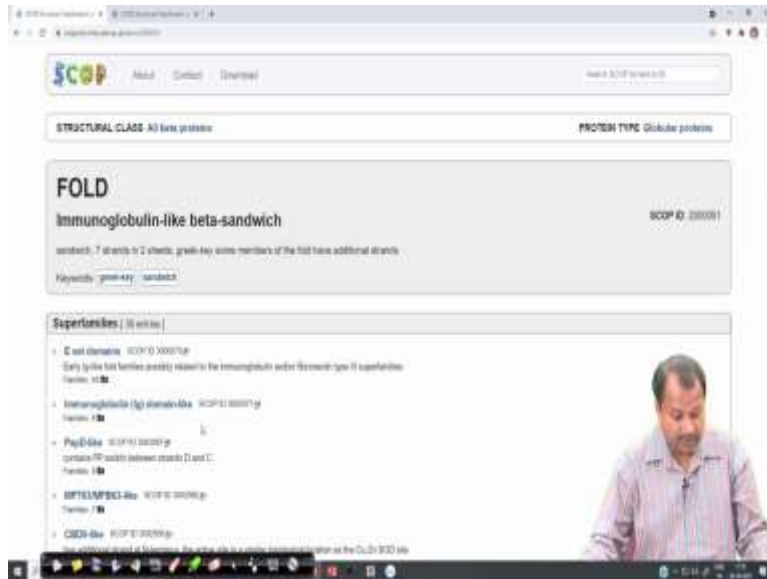
(Refer Slide Time: 26:22)



The screenshot shows the SCOP database interface. At the top, there is a search bar and navigation links. The main heading is 'STRUCTURAL CLASS' with the sub-heading 'All beta proteins' and 'SCOP ID: 100001'. Below this, there is a list of folds under the heading 'Folds (25 entries)'. The first entry is 'Immunoglobulin-like beta-sandwich' with SCOP ID 100001p, described as 'sandwich, 7 strands in 2 sheets, green-key: some members of the fold have additional strands'. Other entries include 'SH3-like barrel', '7 stranded beta-barrel', 'OS-fold', 'Proteinase-like', and 'Single-stranded right-handed beta-helix'. A video player is visible in the bottom right corner of the screenshot, showing a man speaking.

Then Immunoglobulin-like beta-sandwich I am looking for at the fold level.

(Refer Slide Time: 26:27)



The screenshot shows the SCOP database interface for a specific fold. The heading is 'FOLD' with the sub-heading 'Immunoglobulin-like beta-sandwich' and 'SCOP ID: 100001'. Below this, there is a list of superfamilies under the heading 'Superfamilies (38 entries)'. The first entry is 'C set domains' with SCOP ID 100001p, described as 'Early Ig-like beta families possibly related to the immunoglobulin and/or the mouse-type II superfamily'. Other entries include 'Immunoglobulin (Ig) domain-like', 'PlyD-like', 'BPTLMP/EC1-like', and 'CSD1-like'. A video player is visible in the bottom right corner of the screenshot, showing a man speaking.

Superfamily say, Immunoglobulin (Ig) domain-like.

(Refer Slide Time: 26:32)

**SUPERFAMILY**  
Immunoglobulin (Ig) domain-like  
SCOP ID: 3880071

Show identity ...

**Functions** ( 6 items )

- C1 set domain-like SCOP ID: 40002P
- T cell receptor ectodomain-like SCOP ID: 40046P  
composed and composed of 3 & 4 Repeat immunoglobulin-like domains (the first domain corresponds to FF/K30 D,0071) and 2 EGF-like domains, split at the superfamily level
- D1M01 domain-like SCOP ID: 40047P  
FF1036 D,0071 domain vs V set domain
- C2 set domain SCOP ID: 40070P
- V set domain-like SCOP ID: 40073P
- I set domain SCOP ID: 40074P
- Antibody variable domains SCOP ID: 40075P
- Antibody constant domains SCOP ID: 40076P

Fetching domains ...

© 2011-2017, licensed to the US National Institutes of Health  
National Center of Biotechnology Information, National Library of Medicine, Bethesda, MD 20894

Then here I am looking for antibody variable domains, this one I am looking for.

(Refer Slide Time: 26:39)

**FAMILY**  
Antibody variable domains  
SCOP ID: 40075W

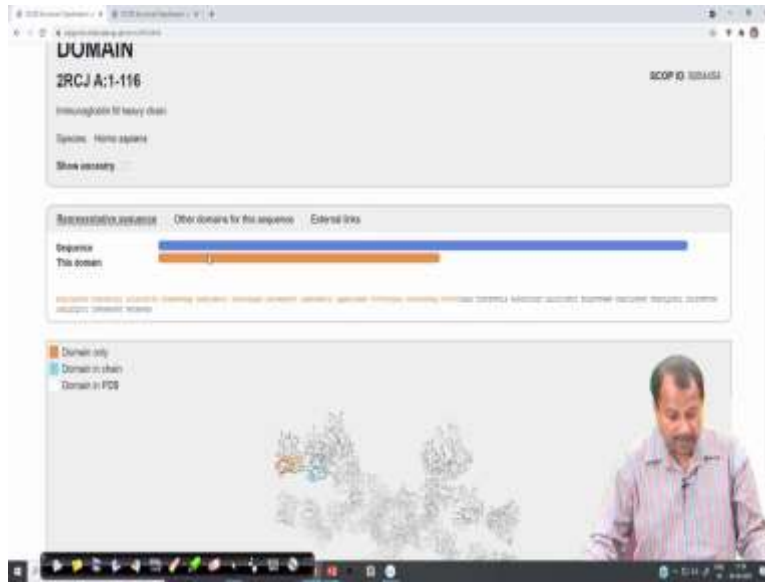
Function: Evolutionarily related complex structures (determined)  
Show identity ...

**Domains** ( 5 items )

	ID	Region	Links
Protein Immunoglobulin M heavy chain Typical heavy chain Representative domain: 014441.p1	382	1-179	PROLY K001082P
Protein Ig lambda chain MOC Variable heavy region Representative domain: 011051.p1	376	1-179	PROLY
Protein Ig gamma-1 heavy chain variable domain, VH Variable heavy region Representative domain: 001534.p1	368	1-179	
Protein Ig kappa chain variable domain, VL kappa Variable heavy region Representative domain: 001534.p1	368	1-179	
Protein Constant IgM domain (C1H1) Variable domain (nonvariable)	368		

And at the domain, so many domains are there, I am interested in Immunoglobulin M heavy chain.

(Refer Slide Time: 26:49)

A screenshot of a web browser displaying a protein domain page. The page title is "DOMAIN" and the ID is "2RCJA:1-116". The description is "Immunoglobulin-like heavy chain". The species is "Homo sapiens". There is a "Show ancestry" button. Below this, there are tabs for "Ancestry's sequence", "Other domains for this sequence", and "External links". The "Ancestry's sequence" tab is active, showing a sequence bar with a highlighted orange segment. Below the sequence bar, there are several small text links. At the bottom, there is a section for "Domain only" with sub-sections for "Domain in chain" and "Domain in PDB". A 3D protein structure is visible in the background, and a small video inset shows a man speaking.

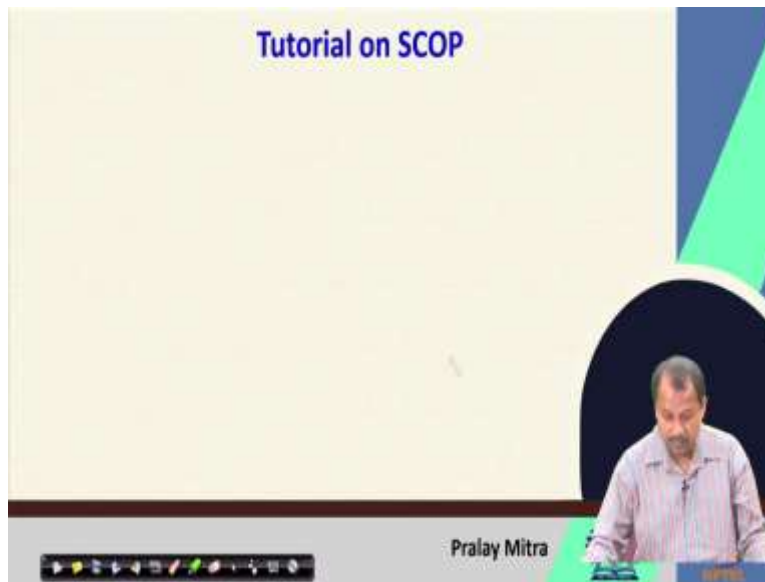
So, this is the protein structure that is shown. And here only which part is highlighted is being classified by this. So, this is, so sequence this domain, so that is there because it is a very large protein actually and if I go for ancestry.

(Refer Slide Time: 27:10)

A screenshot of a web browser displaying a protein domain page, similar to the previous one. The "Show ancestry" button is now active, and a vertical flowchart is displayed. The flowchart starts with "Seq 2RCJA:1-116 (All beta proteins)" at the top, followed by "Seq 2RCJA:1-116 (Immunoglobulin-like heavy chain)", "Superfamily 108071 (Immunoglobulin-like heavy chain)", "Fold 48774 (Immunoglobulin-like heavy chain)", and "Class 108071" at the bottom. Below the flowchart, there are tabs for "Ancestry's sequence", "Other domains for this sequence", and "External links". The "Ancestry's sequence" tab is active, showing a sequence bar with a highlighted orange segment. A small video inset shows a man speaking.

Showing the ancestry domain, one domain is this with this ID, family is with antibody variable domains, then superfamily Immunoglobulin domain-like, fold Immunoglobulin-like beta-sandwich, class all beta proteins. So, it is there.

(Refer Slide Time: 27:38)



So, so going back to our slide, so what we have discussed as of now is basically the SCOP classes. Now, in the SCOP classes the need or the purpose of the SCOP class is to classify the proteins. So, as of now several proteins are, several proteins are coming up. So, in the UniProt the sequences are deposited; in the Protein Data Bank the structures has been deposited. So, classifying them in such a way that I can attach one function corresponding to one class is required.

Now, classification problem is difficult although there is some automated tool like CATH but manual intervention is required sometimes. So, initially started with only by manual observations the SCOP structure classification of proteins which is housed at MRC University of Cambridge, MRC laboratory at the University of the Cambridge now uses some semi-automated tools in order to screen out a lot of new structures, because it is the concept that no new folds are coming up.

So, very limited number of new folds are coming up. And most of the structures which are deposited now in the Protein Data Bank is, similar to that some structure has been already been deposited. So, using some sequence alignment or some structural alignment technique if it is possible to classify them automatically, it is fine, do that. Go ahead and do that. Otherwise visual inspections are required.

While classifying, so one flat classification is not going to be the case for protein function classification. So, the suggestion is that, why not divide it into different, different parts? So, different parts means, so at the domain level divide, at the family level, superfamily level divide, then family level, then at the fold level or the class level you divide.

If you go that way then you will have clearly four different layers as SCOP is following. So, you divide by class whether it is all alpha, all beta, alpha and beta, alpha plus beta, small proteins etc. After that classification you check at the family level, at the superfamily level and at the domain level. That way all the functionality more or less you can cover and you can pin down to one particular function which is relevant for your particular protein. So, that is it for this SCOP class. In the next class, next lecture actually we will start symmetry in protein structure. Thank you very much.