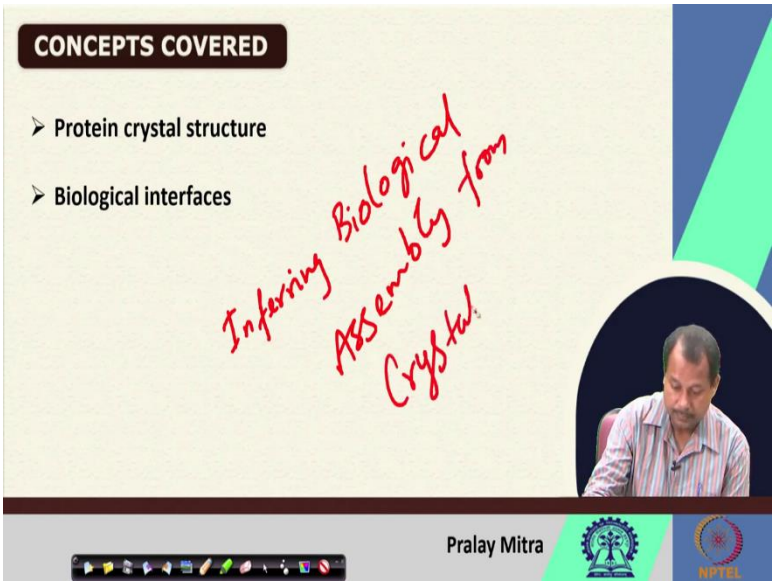
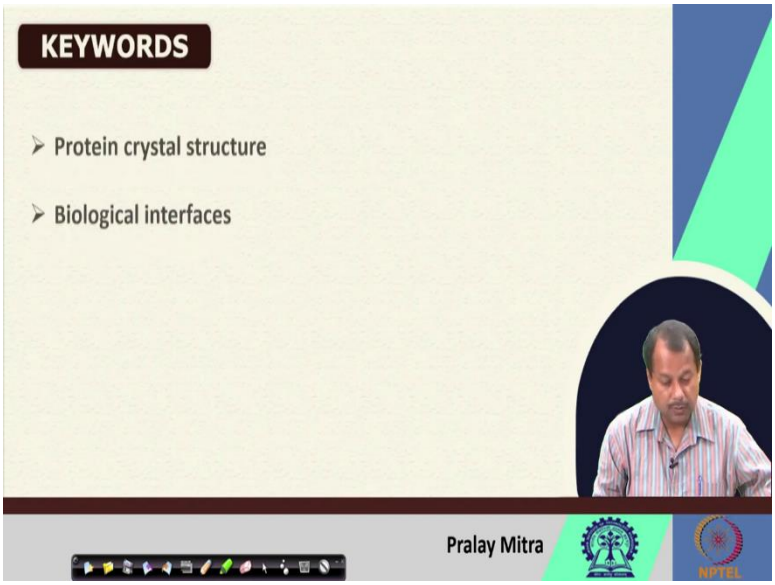
Algorithms For Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 37
Discriminating Biological Protein Interfaces from Crystal Artifacts (Contd.)

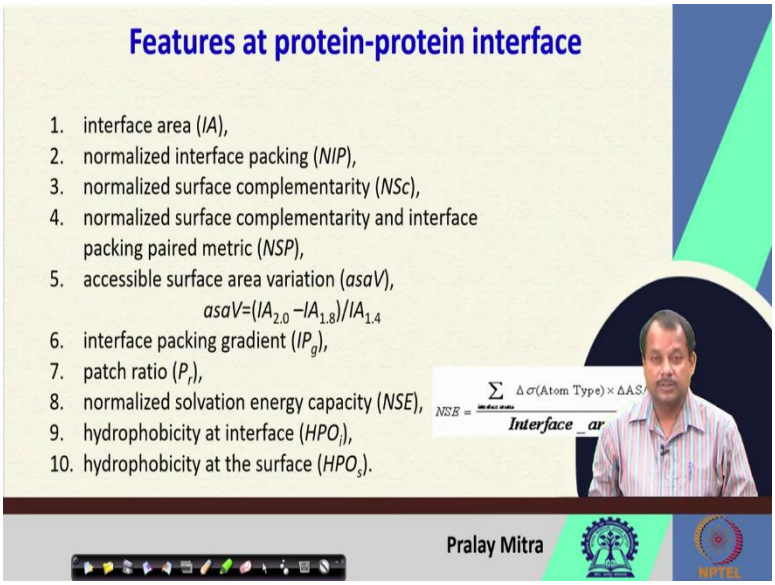
(Refer Slide Time: 00:15)

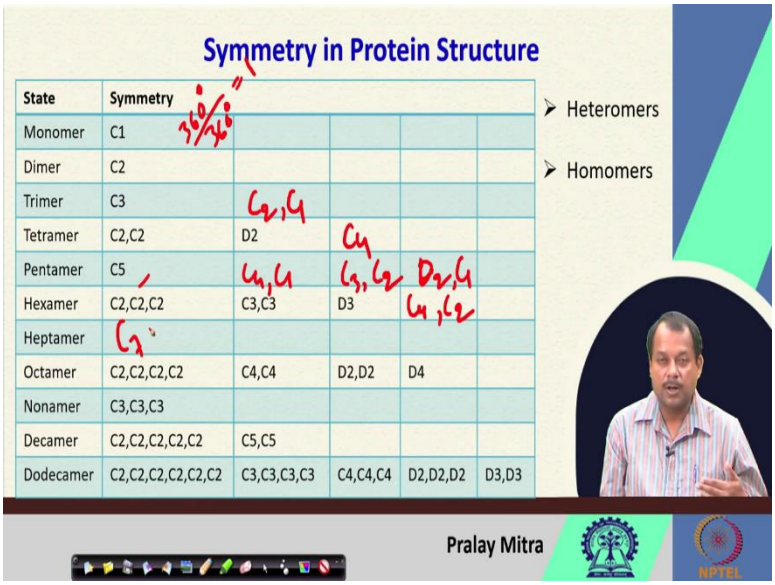




Welcome back. So, we are continuing with discriminating biological protein interfaces from crystal artifacts. And also, I would like to add that in this lecture, I will be covering protein crystal structure, biological interfaces along with that one inferring biological assembly from crystal structure that I will be covering in today's class lecture. So, keywords are same.

(Refer Slide Time: 00:59)





State	Symmetry				
Monomer	C1				
Dimer	C2				
Trimer	C3				
Tetramer	C2,C2	D2			
Pentamer	C5				
Hexamer	C2,C2,C2	C3,C3	D3		
Heptamer					
Octamer	C2,C2,C2,C2	C4,C4	D2,D2	D4	
Nonamer	C3,C3,C3				
Decamer	C2,C2,C2,C2,C2	C5,C5			
Dodecamer	C2,C2,C2,C2,C2,C2	C3,C3,C3,C3	C4,C4,C4	D2,D2,D2	D3,D3

So, these are the features using which we have designed one binary classifier. And I promise that this binary classification will be combined with the symmetry information in protein structure in order to generate the biologically functional forms. Now, just to give you one a recap of the symmetry in protein structure.

So, we mostly mentioned that when there is a symmetry the symmetry will be between the homomeric chains, not with the heteromer, but there is a possibility that there can be heteromer starting from dimer. So, in dimer there may be two different chains. If it is, then we will not find any symmetry there. If say, it is trimer then maybe two chains are same another is different, then, we are suggesting that the symmetry will be C2 between the known chains and C1 with the hetero chain.

Now, the C1 actually for the monomer is basically no symmetry exists. Because as per the cyclic symmetry definition it is 360 cross 260, which gives you one. Which means, if I give 360 rotation then you will get 1. So, it is easy to get or identify one symmetric axis about which if you give 360-degree rotation then you will get the same structure. So, that is trivial. So, that is why that C1 I am not discussing much.

Now, the dimer if it is say homodimer then only I will get C2 symmetry. If it is a trimer then I will get C3 symmetry if it is homotrimer. If it is a heterotrimer then there are three possibilities So, heterotrimer two possibilities sorry. One is that all three chains are different then there is no symmetry at all, other is two chains are homo one is hetero. So, in that case actually I will have C2 between the homo, and C1 with the hetero that is written here C2 and C1.

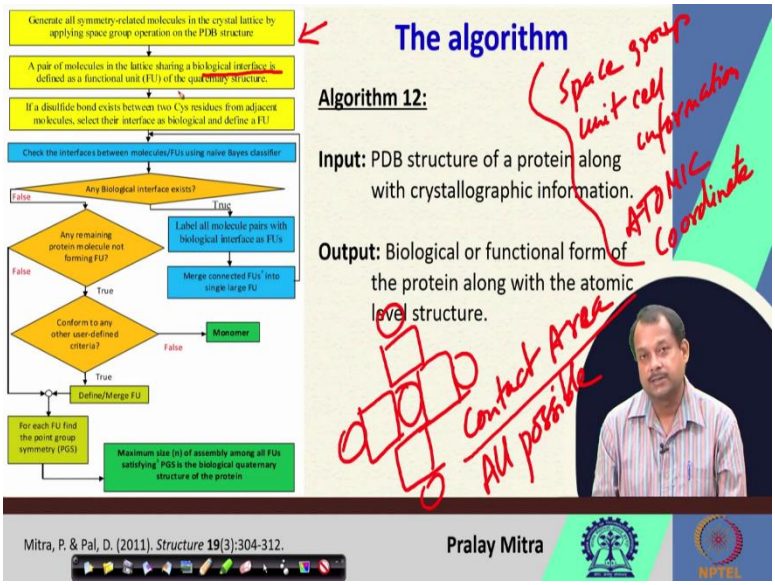
If it is tetramer and again so if it is hetero all chains are different, no question of any symmetry operation. But if say, it is consisting of all four same chains then there can be D2 or C4. And if it is say consisting of two chains which are homomer, two another chains which are homomer but these two and these two are different like hemoglobin 2 alpha, 2 beta, so alpha and beta is not same.

So, between 2 alpha there will be C2, between 2 beta there will be C2 so, that is C2, C2. Therefore, pentamer it is C5 or based upon, if it is a heteromer then based upon how many homo structures are there, so, it can be C4 and C1, it can be C3, C2. So, possibilities are C4, C1, C3, C2. It can be D2, C1 so these many possibilities are there.

So, if it is a hexamer then all C2, C2, C2, C3, C3, D3 or can be C4, C2, so that can be possible. If it is heptamer ideally, C7 can be one symmetry, rarely, we will find any C7. So, the reason I mentioned also during their symmetry discussion, so even number of symmetry, I mean, the even number of biological assembly dimer, hexamer, heptamer, dimer, tetramer, hexamer, octamer, decamer, those are occurring in nature more frequently compared to the odd number of assembly.

But in this odd number of assembly trimer occurs smore, so for trimer we will not consider that one and monomer of course, is not an assembly, so we will exclude. But from pentamer, heptamer, nonamer, so their occurrences are very less.

(Refer Slide Time: 05:16)



Now, let us go to that algorithm for which we waited long. So, here is the total flow diagram of the algorithm, not stepwise but as a flow diagram. Input is, PDB structure of a protein along with crystallographic information output biological or functional form of the protein along with the atomic level structure.

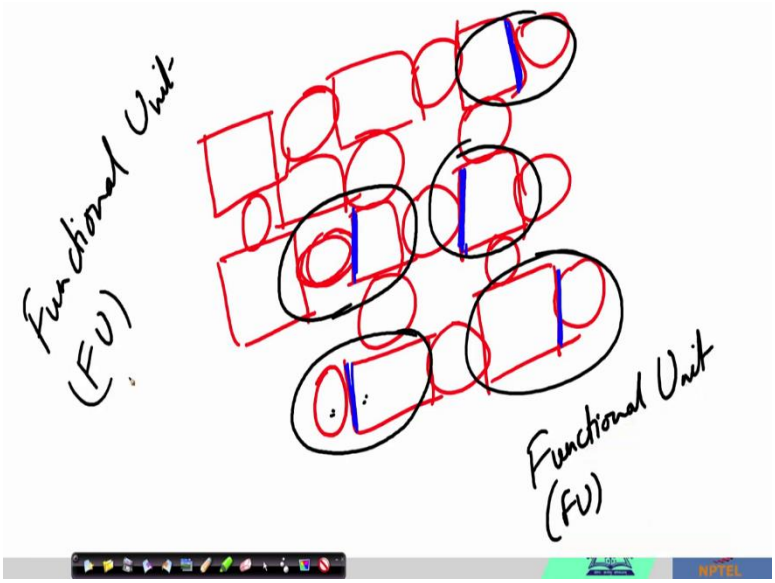
Now, let me go through the algorithm step by step. So, first you need to generate all symmetry-related molecules in the crystal lattice by applying space group operation on that PDB structure. So, it is possible. So, you need to have space group and unit cell information. Along with, of course, the atomic coordinate. If you have all this then definitely you can generate symmetry-related in the crystal lattice by applying the space group operation.

So, in one of the lectures I demonstrated you that using pimol, I generated that one, but it can be generated using program also. Once you will generate then a pair of molecules in the lattice sharing a biological interface is defined as a functional unit of the quaternary structure. So, there is a gap between these two steps. So, read between the lines. So, what do you need to do once you generate the lattice basically, so you need to identify the contact area, which contact area, all possible.

So, all possible contact area you need to identify. Then, on all possible contact area you need to run that binary classifier that we discussed along. 10 features Naive Bayes classifiers, but I also suggested you can increase the number of features, you can redefine the features, recalculate the features and instead of base you can apply some random for rest or other classification technic also, it is fine. So, but the requirement is that you have to go for one binary classification. Once you will have that binary classification then what will happen

among all those contact area few will be identified as the biological interfaces. Now, a pair of molecules in the lattice sharing a biological interface is defined as a functional unit of the quaternary structure.

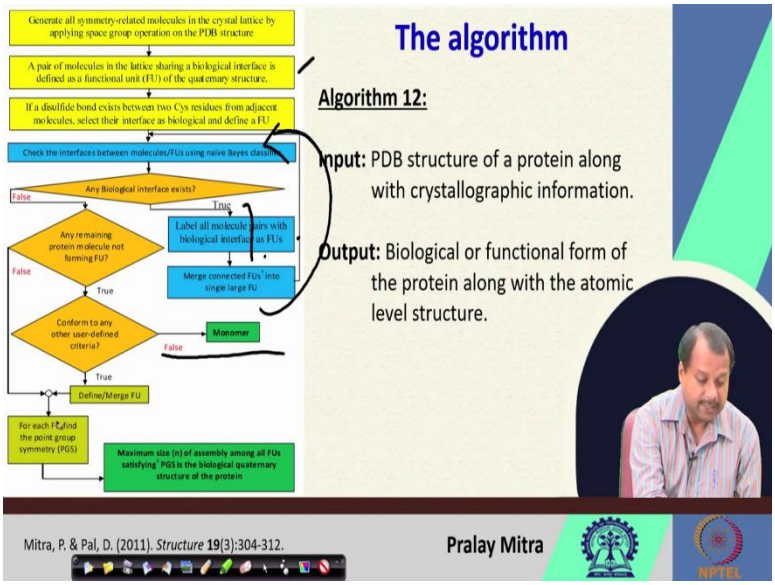
(Refer Slide Time: 08:02)



So, what I am suggesting now or the algorithm says that, if say, let me take a separate structure on this one. So, that one was better actually. In 2D I am drawing, in 3D I believe you will be able to understand. So, the crystal structure is there. Now, let us assume that blue color is indicating the crystal interface, the interface which are biologically relevant.

Then, as per the definition let me pick pink or black. This is one functional unit, this is one functional unit this is one, this is one, and this is one. What I am doing here now, first, I am generating the crystal lattice. After generating the crystal lattices all possible contact area has been identified. I run the binary classifier to check whether they are biologically relevant or not. If yes, then all those interferes and the and the corresponding subunits or the chains, which are actually contributing to that interface is explaining me that it is going to be one functional unit. In short FU, it is going to be the functional unit. In short it is FU. So, that FU I am considering.

(Refer Slide Time: 10:37)



Now, this is my FU here. Along with this, if a disulfide bond exists between two cysteine residues from adjacent molecules select their interface as biological and define as FU. So, what it is suggesting that okay you identify the FU using binary classifier it is fine. But if you find that there exists some disulfide bond because of its interaction potential.

And since disulfide bonds are one of the most strong, strongest bond, which determines a non-covalent bond, which determines mostly the fold or the structure then you declare that also even if your binary classifier fails to identify that interface, as a correct biological, but you please define that one also as biological one and consider that as a FU. That way you will have in number of FU.

Now, check the interfaces between molecules are FU Naive Bayes classifier that we did. Then any biological interface exists, yes. If it is true, label all molecule payers with biological interfaces as FUs. So, you combine and grow your FU. Look, these FUs are going to be your biological unit.

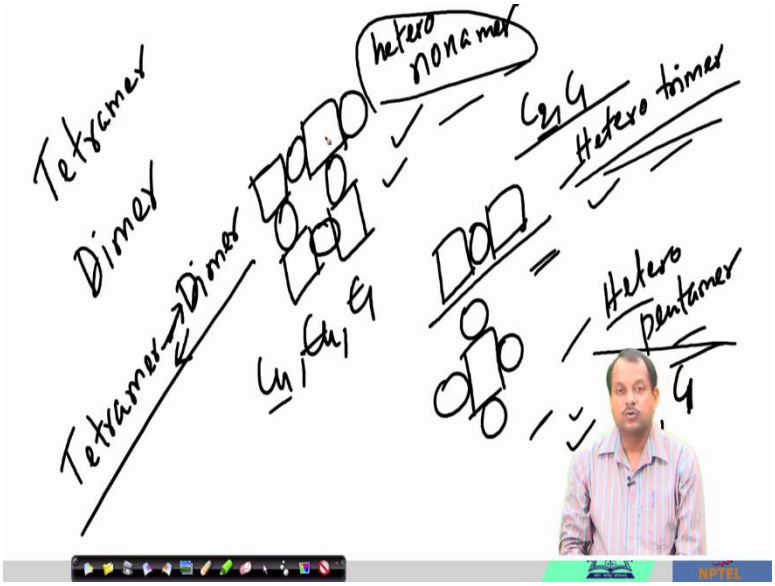
So, you started with the binary classification and each interfaces you are identifying as biological relevant. If they are biologically relevant then you are considering these two as a correct one. So, these are dimer, if these two are dimers then what I am going to do is that this is my FU.

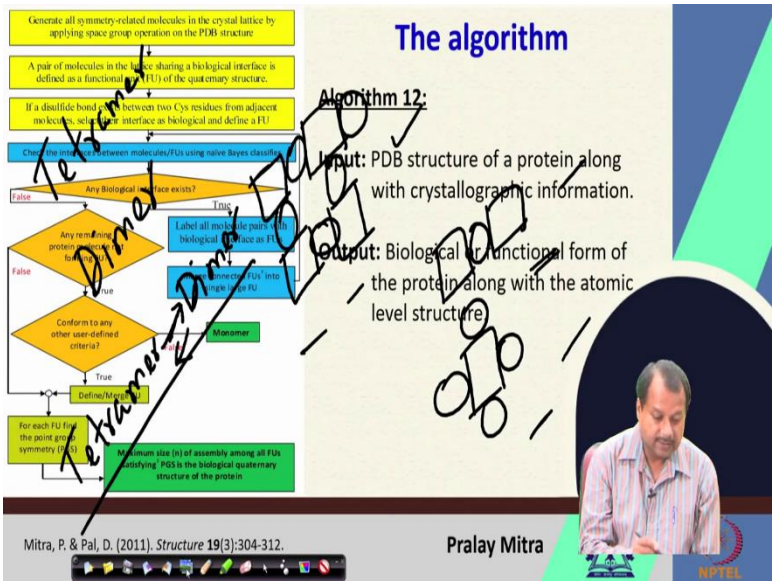
After concerning this FU, with this, then it will be one single unit. With this FU anybody else is interacting. Ad whether that interface is biological if, yes, then you consider that one also part of this FU. That way you are growing the FUs. From 1 then 2, 3, 4, 5, 6, that way you are

growing, and that rowing is done in this true part. So, label the biological FU merge connected FUs into single large FU, go back. And keep on iterating as long as you can combine that FUs. Any biological interface exists, when you will not find anyone then any remaining protein molecule not for me FU so then, true, conform to any other user defined criteria.

If you say that like the disulfide bond, I have another criteria to identify that whether it will be forming a FU or not then basically you can add that one, otherwise, you declare that as a monomer. And if it is then define the margin FU for each FU find the point group symmetry. So, before this finding the point group symmetry let me show you here.

(Refer Slide Time: 13:33)





So, I combined, say, this is one. And after the communism I have grown. And because of my growth so, I got this one, and this one, say, I got this one. And I can also get, so I am going and I get this one, I can also get this one, I can also get. So, it may possible that from one crystal lattice, I may have one large FU, and small, small FUs, it may possible. If, yes, then you instead of picking the largest one you report everything.

Because it is known that a protein in different environmental condition may behave differently. So, in some situation, it can have a tetramer, in some situations and the same protein may act as a dimer, in some situation or in solution it can exist as a drimer, trimer both. So, where he will see, that trimer is converting to dimer and dimer is converting to tetramer. So, in equilibrium, they exist.

So, my suggestion will be during this prediction process if you come up with a number of FUs whose number of subunits are, it is greater than 1. So, it is 1 that is monomer, so you can erase that one. If that is the only FU then it is fine then you report as a monomer, otherwise, if you have different FUs functional units of which one is with one chain size then you can exclude that one, but do not exclude rest. So, you keep all the FUs. This is one, this is one, this is one, you keep all of them.

After keeping all of them, then what it suggests the theory, for each if you find the point group symmetry. So, for each FU, so this, this, this. What is the point of symmetry in this case? In this case, it is going to be a heterotrimer so there is C2 and C1 so hetero trimer. In this case what I am getting, it is hetero pentamer, C4, C1. In this case, it is 1, 2, 3, 4, 5, 6, 7, 8, 9, nonamer.

Now, this hetero nonamer, hetero trimer, hetero pentamer, among these this hetero nonamer is unlikely. Because as I mentioned if you go odd number of assembly state then a nonamer etc then it is not much likely. However, if you find that there exist two quaternary states which can make some equilibrium like say tetramer and dimer kind of thing. Say, in that case say tremor and hexamer or say pentamer and decamer then that may perhaps possible.

But in this case say, hetero nonamer, hetero trimer, hetero pentamer, and as such there is no symmetry related information because here what I can see, C4. C4, C1, for hetero nonamer. For hetero trimer, C2, C1, for hetero pentamer, C4, C1. So, C2, C4, C4, so this with this one and this one there is some similarity, and with this one and this one there is no such similarity.

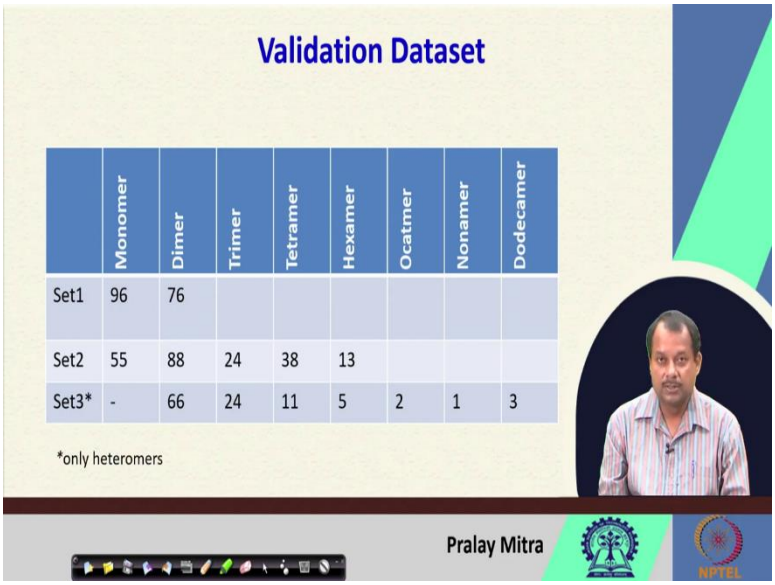
So, if heterotrimer is possible then, perhaps, hetero pentamer is not possible okay. But if you look at the structure, so C2 and C1. Here C2 and C1 here C2 and C1, here C2 and C1 so that way perhaps hetero nonamer maybe a solution. So, this kind of fine tuning you have to do after you identify the symmetry in formation.

So, maximum size of assembly among all FU satisfying PGS point group symmetry is the biological quaternary structure of the protein. So, that is the final output. But there are some scope to improve, say, the way you are identifying the symmetry of information like whether it is cyclic symmetry, dihedral symmetry, so there is a huge scope to improve that one. And as of now, the approach is like growing.

So, where you are starting with one single subunit and then you are appending by using the binary classifier whether it will be connected that way you are growing. Another approach may be shrinking. So, you consider the large size of lattice say I start with say decamer or do decamer that is the functional unit.

Then, you go on excluding that okay this is not going to be as per either symmetry or by using some machine learning or say deep learning information, you exclude some say internal classification information, you exclude some of them, and then, you reduce the steps.

(Refer Slide Time: 19:53)

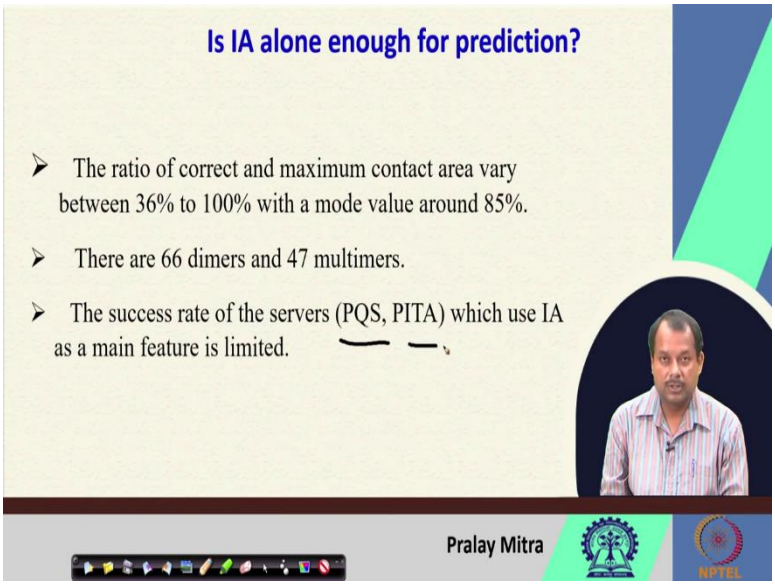


So, for the validation, you can use three different sets, so which are readily available. So, among this third set is consisting of only heteromers. So, these are my used validation data set, but you can come up with your own data set also. For the set 1 so, 96 monomers and 76 dimers are there. As you understand that for this only one binary classifier is enough. You

need not have to go for this computational framework for symmetry-related information etc. Then for set 2, 55 monomer, 88 dimer, 24 timer, 38 tetramer and 13 hexamer. Look, so as I keep on increasing then this pentamer is missing, and then, nonamer is there, but heptamer is missing. So, which means odd number of mers are usually not very frequent.

In the set 3, there is no monomer but 66 dimer, 24 trimer, 11 tetramer and 5 hexamer, 2 octamer, 1 nonamer and 3 dodecamer so 12 size, dodecamer means 12 size, so 12 sizes are also there. So, that is the validation data sets you can use.

(Refer Slide Time: 21:12)

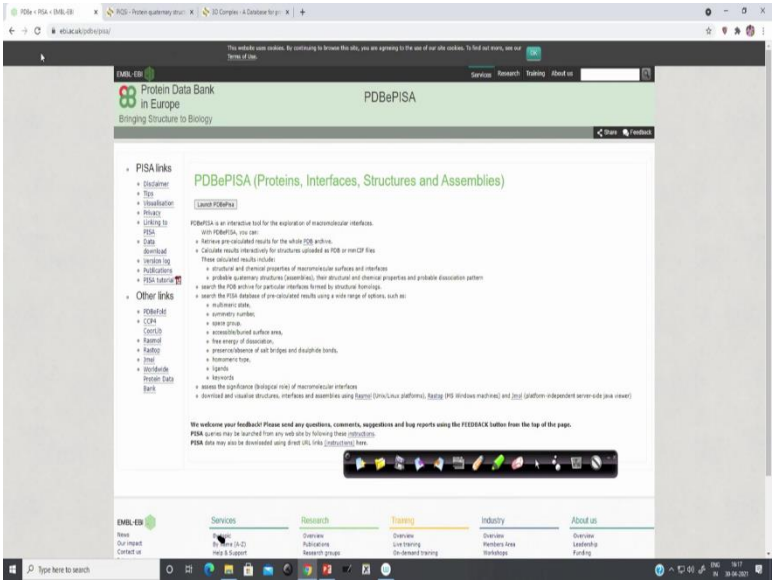


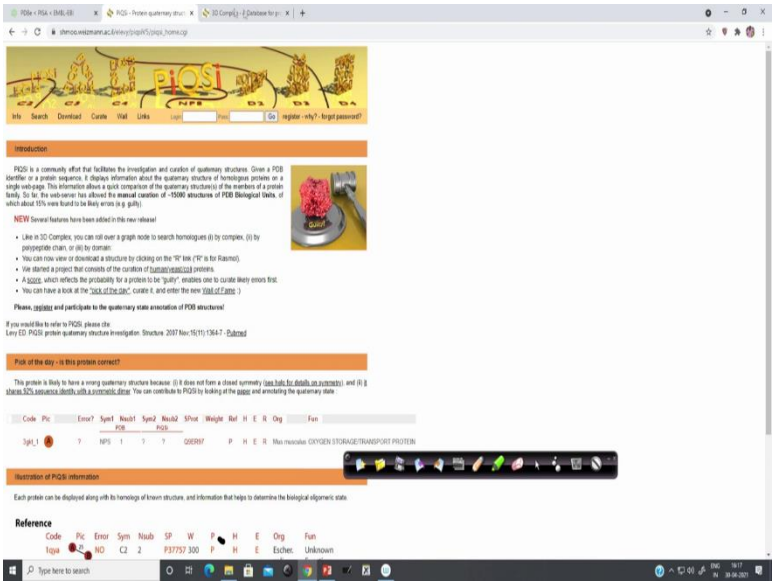
Then, the question is, is IA which means interface area alone enough for prediction? The ratio of correct and maximum contact area vary between 36 percent to 100 percent with a mode value around 85 percent. There are 66 dimer and 47 multimers. The success rate of the servers like PQS, PITA, which use IA as a main feature is limited, and that is perhaps, one of the reason why the support for these PQS and PITA has stopped.

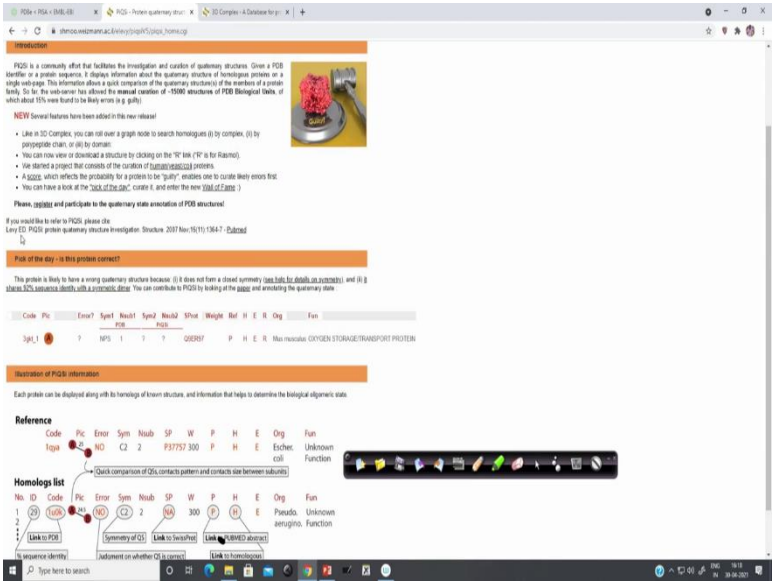
(Refer Slide Time: 21:51)

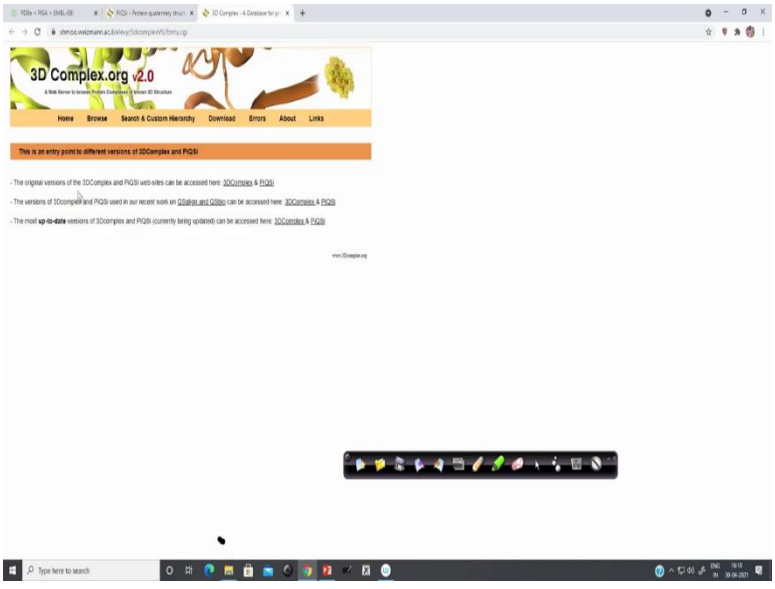
Instead of that two other databases, which has come up is, one is the PISA, which is hosted at the ebiembl another is PiQSi originally it was at named as 3D complex, originally it was named as 3D complex, and at was at MRC, Cambridge but now since the author has moved so it has moved to Wiseman. Another is at the ISC that is the IPACdb.

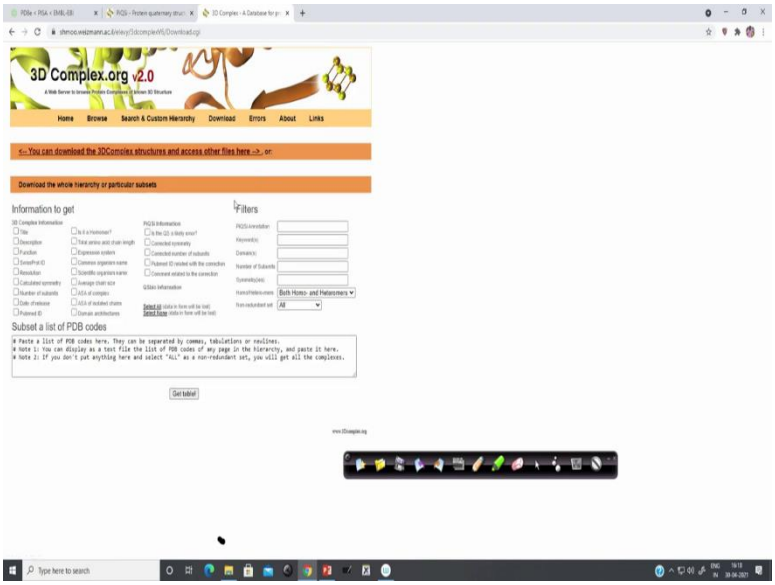
(Refer Slide Time: 22:35)




PiQSi
Info Search Download Curate Wall Links
Introduction
PiQSi is a community effort that facilitates the investigation and curation of quaternary structures. Given a PDB identifier or a protein sequence, it displays information about the quaternary structure of homologous proteins on a single web-page. This information allows a quick comparison of the quaternary structure(s) of the members of a protein family. So far, the web-server has allowed the manual curation of ~15000 structures of PDB Biological Units, of which about 15% were found to be likely errors (e.g. guilty).
NEW Several features have been added in this new release!
Like in 3D Complex, you can roll over a graph node to search homologues (i) by complex, (ii) by polypeptide chain, or (iii) by domain
You can now view or download a structure by clicking on the "R" link ("R" is for Rasmol).
We started a project that consists of the curation of human yeast ccd proteins.
A score, which reflects the probability for a protein to be "guilty", enables one to curate likely errors first.
You can have a look at the "pick of the day", curate it, and enter the new Wall of Fame :)
Please, register and participate to the quaternary state annotation of PDB structures!
If you would like to refer to PiQSi, please cite:
Levy ED. PiQSi: protein quaternary structure investigation. Structure. 2007 Nov;15(11):1364-7. PubMed
Pick of the day – is this protein correct?
This protein is likely to have a wrong quaternary structure because: (i) it does not form a closed symmetry (see help for details on symmetry), and (ii) it shares 52% sequence identity with a symmetric dimer. You can contribute to PiQSi by looking at the paper and annotating the quaternary state.
Code Pic Error? Sym1 Nsub1 Sym2 Nsub2 SProt Weight Ref H E R Org Fun
PDB PiQSi
3jet_1 ? NPS 1 ? ? Q9ER97 P H E R Mus musculus OXYGEN STORAGE/TRANSPORT PROTEIN
Illustration of PiQSi information
Each protein can be displayed along with its homologs of known structure, and information that helps to determine the biological oligomeric state.
Reference
Code Pic Error Sym Nsub SP W P H E Org Fun
1qya NO C2 2 P37757 300 P H E Escher. Unknown
Type here to search


Introduction
PiQSi is a community effort that facilitates the investigation and curation of quaternary structures. Given a PDB identifier or a protein sequence, it displays information about the quaternary structure of homologous proteins on a single web-page. This information allows a quick comparison of the quaternary structure(s) of the members of a protein family. So far, the web-server has allowed the manual curation of ~15000 structures of PDB Biological Units, of which about 15% were found to be likely errors (e.g. guilty).
NEW Several features have been added in this new release!
Like in 3D Complex, you can roll over a graph node to search homologues (i) by complex, (ii) by polypeptide chain, or (iii) by domain
You can now view or download a structure by clicking on the "R" link ("R" is for Rasmol).
We started a project that consists of the curation of human yeast ccd proteins.
A score, which reflects the probability for a protein to be "guilty", enables one to curate likely errors first.
You can have a look at the "pick of the day", curate it, and enter the new Wall of Fame :)
Please, register and participate to the quaternary state annotation of PDB structures!
If you would like to refer to PiQSi, please cite:
Levy ED. PiQSi: protein quaternary structure investigation. Structure. 2007 Nov;15(11):1364-7. PubMed
Pick of the day – is this protein correct?
This protein is likely to have a wrong quaternary structure because: (i) it does not form a closed symmetry (see help for details on symmetry), and (ii) it shares 52% sequence identity with a symmetric dimer. You can contribute to PiQSi by looking at the paper and annotating the quaternary state.
Code Pic Error? Sym1 Nsub1 Sym2 Nsub2 SProt Weight Ref H E R Org Fun
PDB PiQSi
3jet_1 ? NPS 1 ? ? Q9ER97 P H E R Mus musculus OXYGEN STORAGE/TRANSPORT PROTEIN
Illustration of PiQSi information
Each protein can be displayed along with its homologs of known structure, and information that helps to determine the biological oligomeric state.
Reference
Code Pic Error Sym Nsub SP W P H E Org Fun
1qya NO C2 2 P37757 300 P H E Escher. coli Unknown Function
Quick comparison of QSs, contacts pattern and contacts size between subunits
Homologs list
No. ID Code Pic Error Sym Nsub SP W P H E Org Fun
1 29 1a0k NO C2 2 NA 300 P H E Pseudo. aeruginо. Unknown Function
2
3
4
5
Link to PDB
Symmetry of QS
Link to SwissProt
Link to PUBMED abstract
% sequence identity
Judgment on whether QS is correct
Link to homologous
Type here to search

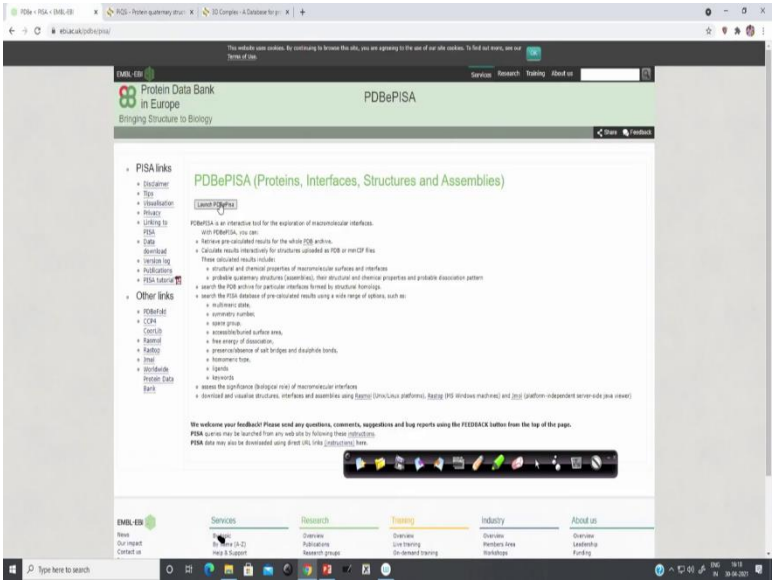

3D Complex.org v2.0
Home Browse Search & Custom Hierarchy Download Errors About Links
This is an entry point to different versions of 3DComplex and PiQSi
- The original versions of the 3DComplex and PiQSi web-sites can be accessed here: 3DComplex & PiQSi
- The versions of 3Dcomplex and PiQSi used in our recent work on QSalign and QSbio can be accessed here: 3DComplex & PiQSi
- The most up-to-date versions of 3Dcomplex and PiQSi (currently being updated) can be accessed here: 3DComplex & PiQSi
Type here to search

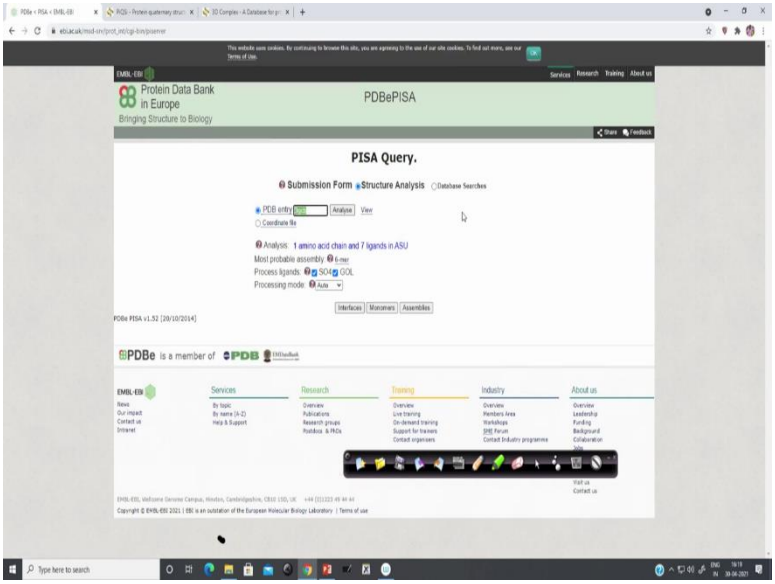


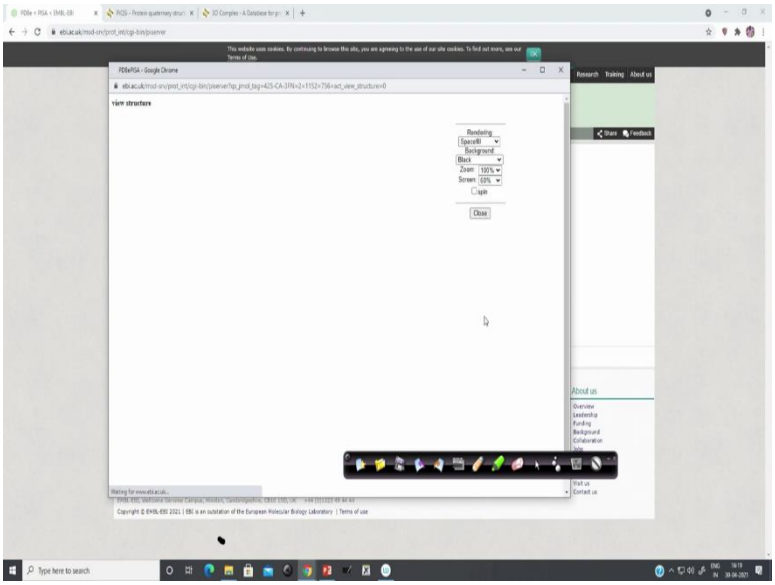
So, let me show you the servers. And what is the use that you can see from there. So, I already opened this. So, in order to open so, here is my suggestion for the PISA you can google search you can write PISA, then EBI or MBL, then this interface will come. If you search for PiQSi then you will get two links. One is the PiQSi dot, so one is that piqsi.org and another is at the Weizmann, but both will redirect to the same page, so this is 3D complex page, and this is the PiQSi page.

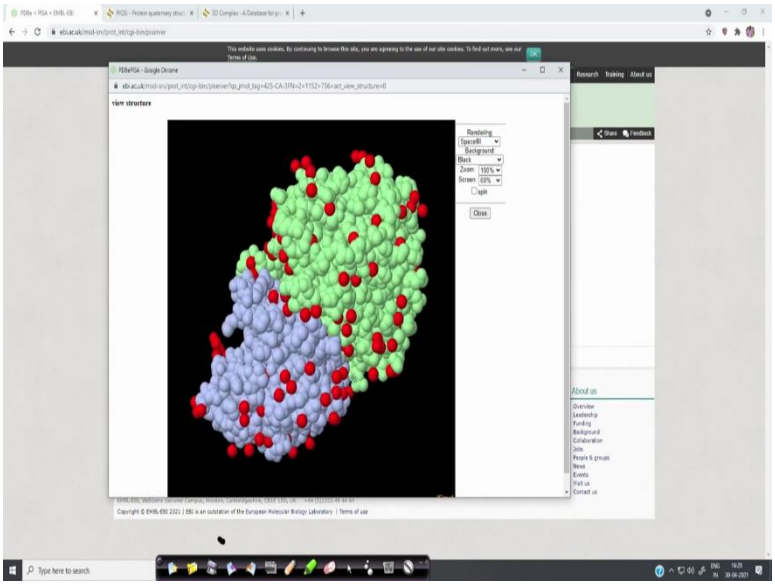
Now, in the PiQSi, so initially they have started to the they have started to manually curate this one. So, it is the work by Emanuel levy. So, PiQSi protein quaternary structure investigation, it was published in structure in 2007. And 3D complex is, so here you will get search and custom hierarchy, download option is there for you, and then errors. So, you can download the structure, so error information all those impulses are there. PiQSi information is also there. So, corresponding to one structure and how you will read this one that information is also provided here.

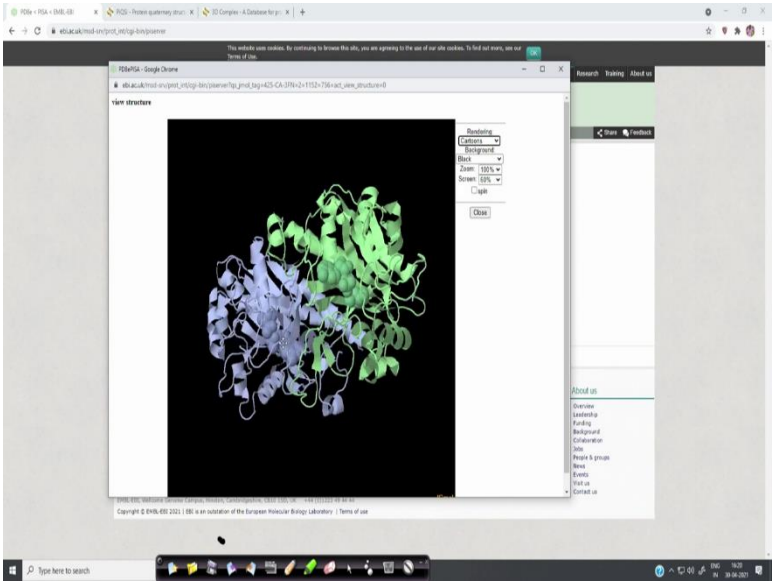
(Refer Slide Time: 23:59)

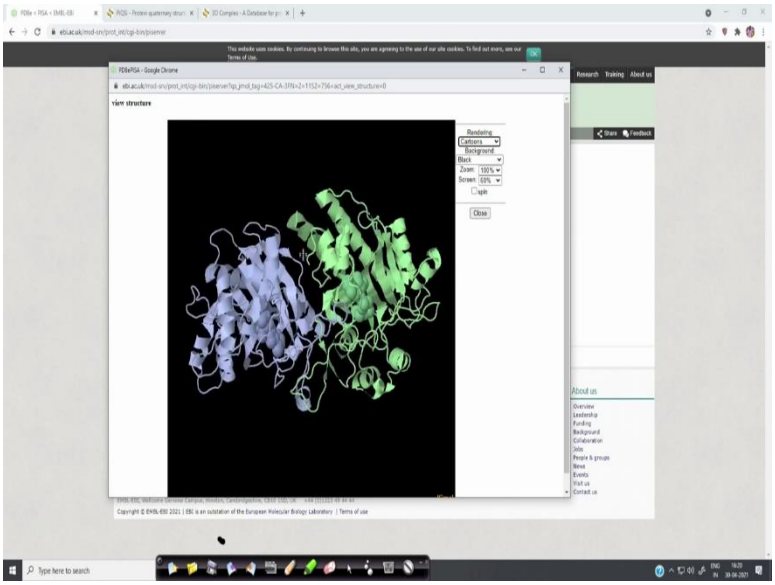










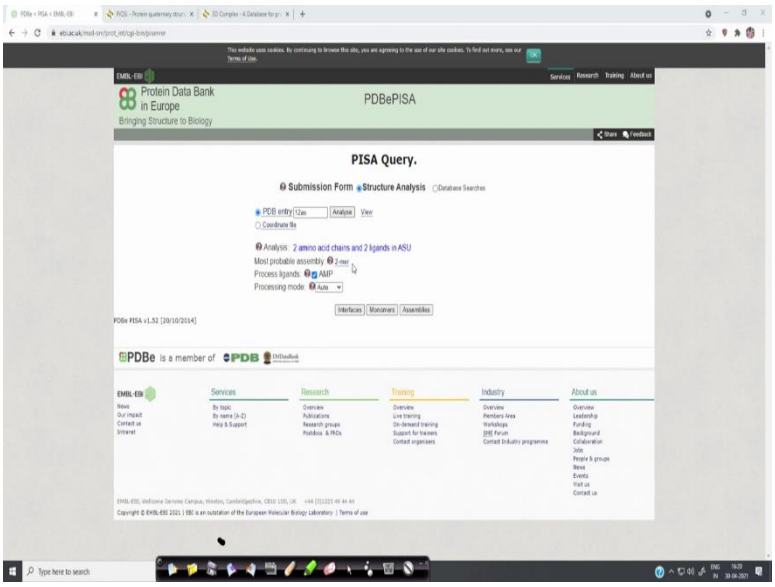


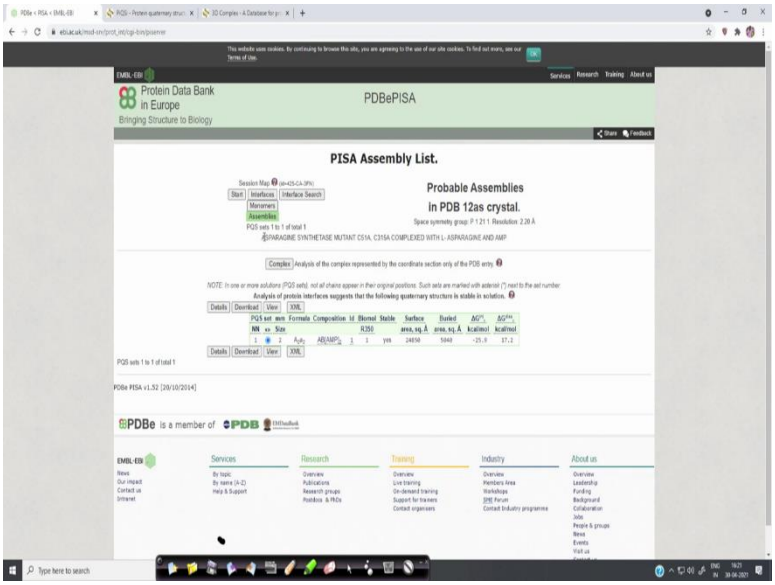
Now, interesting is this PISA. Let me go to this in detail. Launch PDBePISA. So, when you will be here then so you can give one PDB entry. So, one entry could be say 1 2 as. So, automatically when it will give then it will analyze that one. So, after analysis two amino acid chains it is mentioning to me acid chains and two ligands in ASU, asymmetric unit ASU asymmetric unit. Most probable assembly is 2-mar, 2-mar means dimer it suggest.

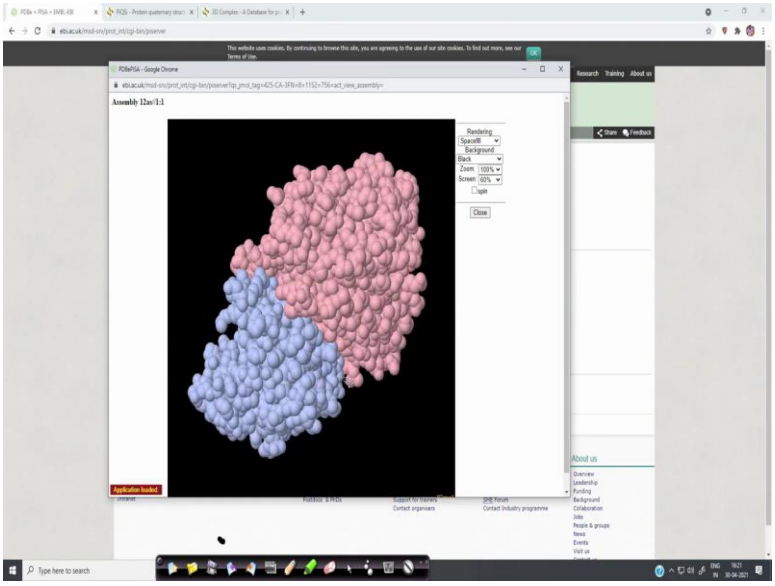
So, there are some ligands AMP. If you want you can process otherwise you need not have to process. Now, you can analyze usually it analyzed automatically but if you want to analyze that one then you can go for the view. So, this is the structure. So, this is the spacefill structure rendering you can go to the cartoon So, if you go to the cartoon so hetero items are in the spacefill. It is here, you can see that one.

Now, you see that there is a C2. The first thing is that you have to identify what will be my axis of rotation. For C2, mostly it is easy to identify say this can be one axis at this point passing perpendicular to this plane. If I look from here then also here, I can have one axis passing perpendicular to this page like that way. Then, this is one structure.

(Refer Slide Time: 26:01)



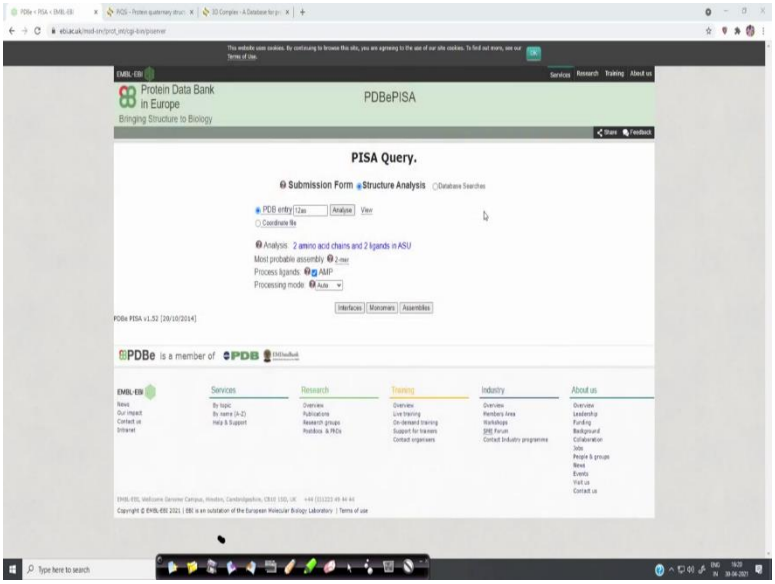


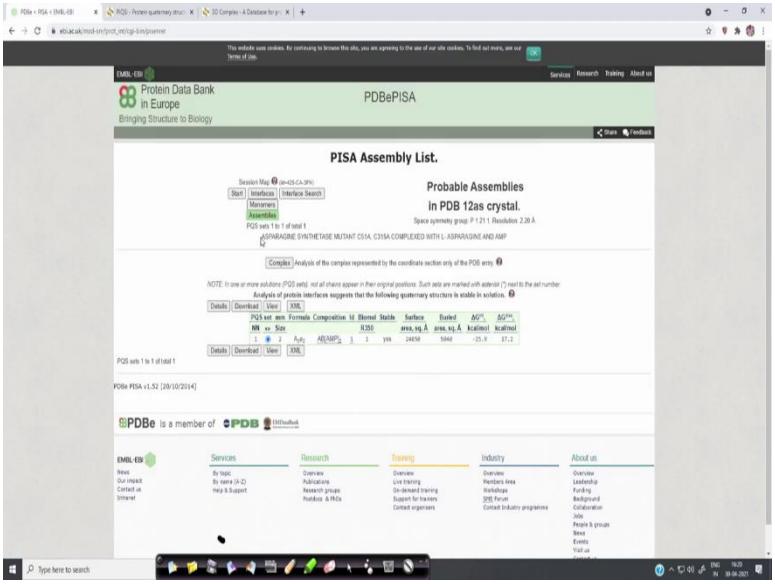


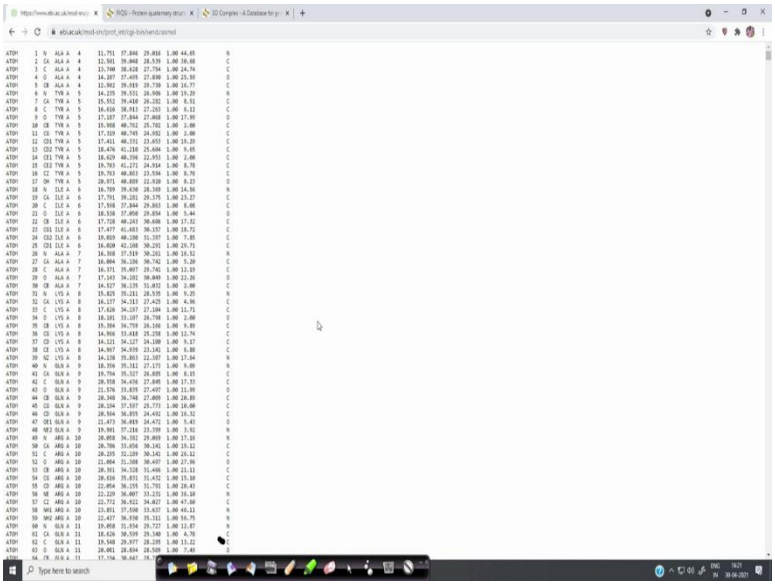
Now, if you wish to look at the interface so, then you have to go to assemblies. So, here in the assembly what you are getting, so, the asparagine synthesis mutant C51a, C315a complexed with L asparagine and AMP that is given. What is the space symmetry group P121 1 that is given, it is a crystal structure resolution is 2.20 angstrom.

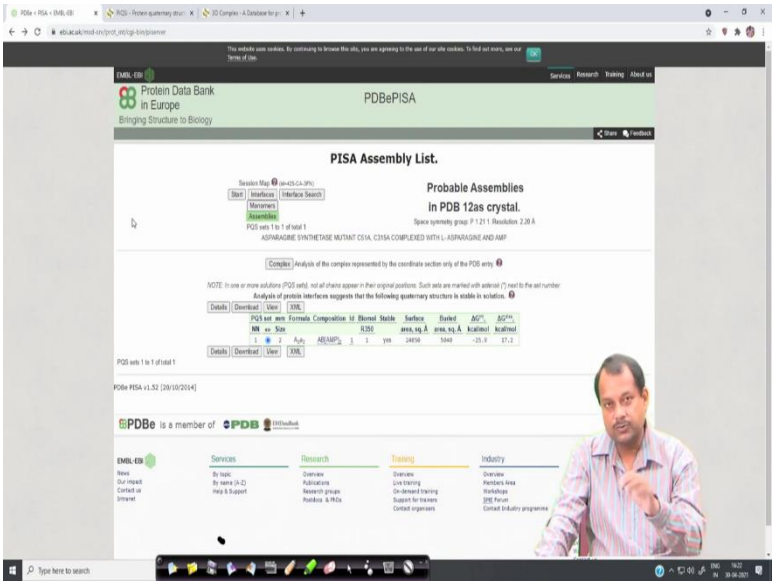
Complex analyze of the complex represented by the coordinate section only of the PDB entry. Now, analysis or PDB, analysis of protein interfaces suggested that suggests that the following quaternary structure is stable in solution. So, download details, view, XML all are here. So, if you wish to view it is the same link that we opened. So, in JSmol it will be open so, this is the for your view.

(Refer Slide Time: 26:59)





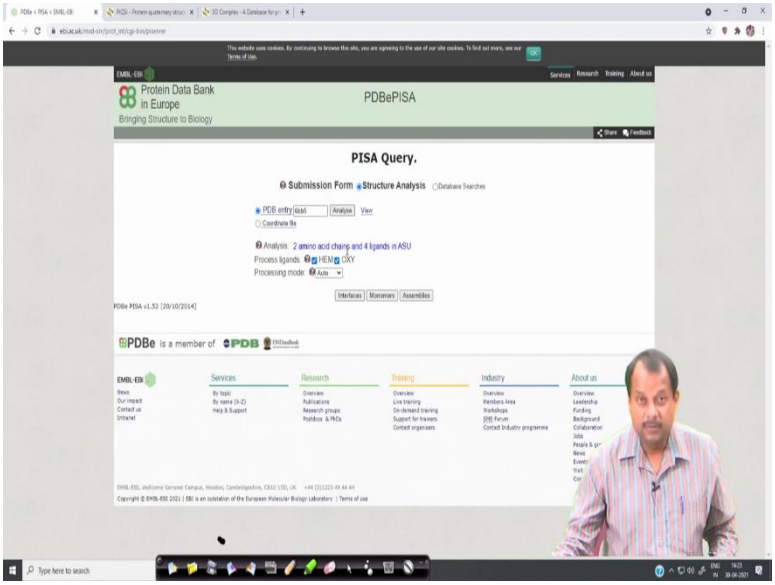


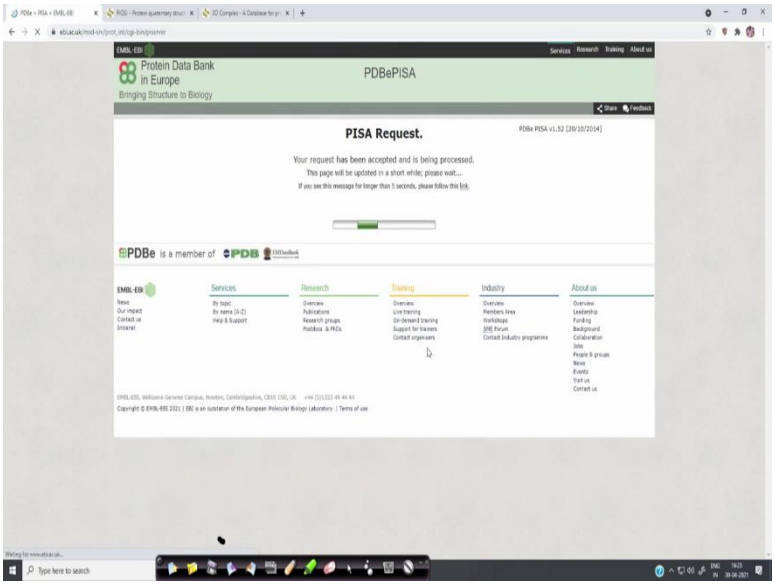


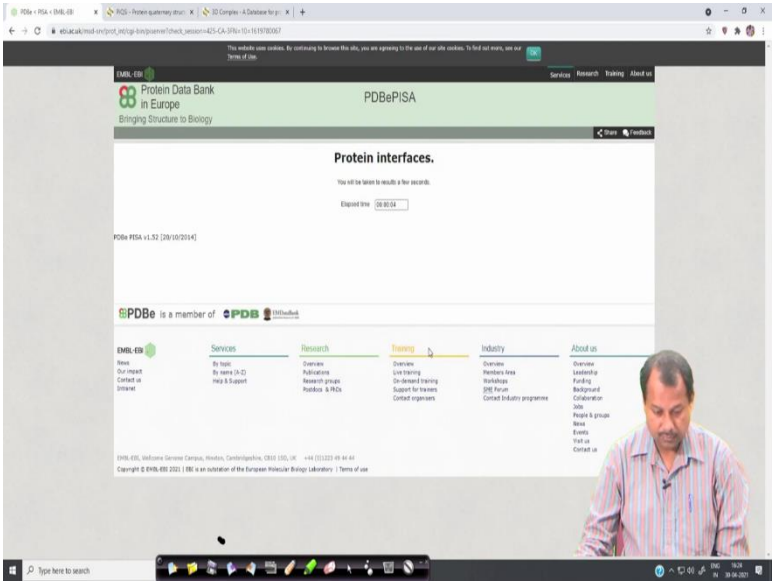
Now, if you wish to download say will get in PDB format. Now, the point is here that if so, incidentally this 1 to AS contains only two chains, A and B, it is a homodimer and that homodimer presents in the PDB. So, as such, there will be no difference between the structure you are downloading and original PDB structure.

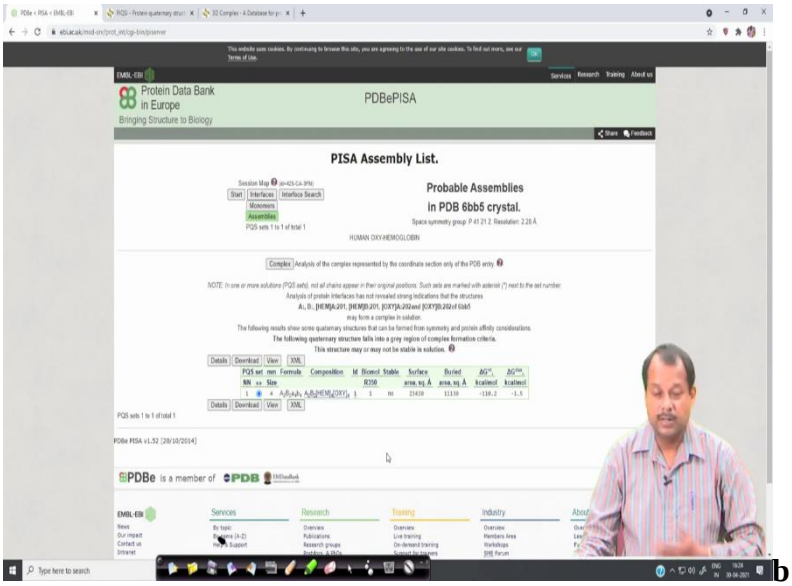
But if it is the case that the original structure is a monomer and then basically you are then basically you are analyzing that one using PISA then PISA suggests that it may be dimer, it may be tetramer then corresponding to that analysis if you download then you will get the symmetry-related information and also the that particular say if it is predicted as dimer or tetramer then that dimer you can able to download, tetramer you can able to download, so those things you can able to download. So, those information's are possible.

(Refer Slide Time: 28:04)

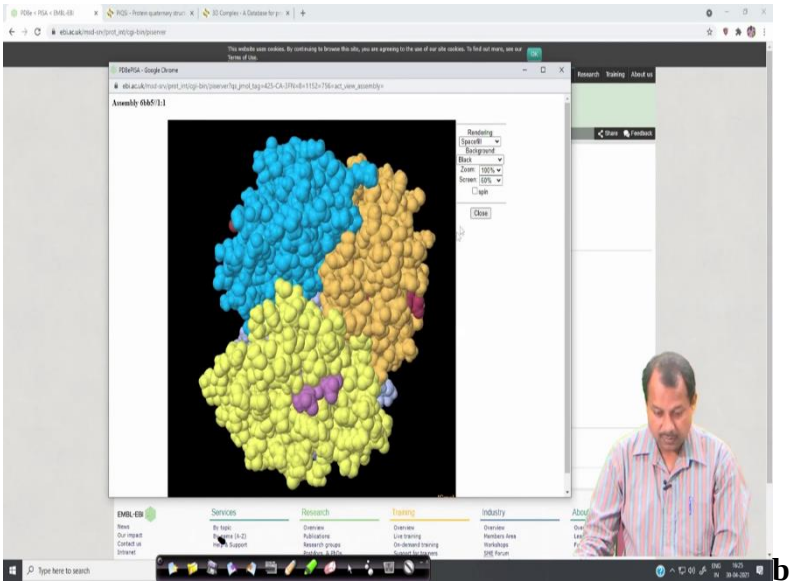



Protein Data Bank in Europe
Bringing Structure to Biology
PDBePISA
PISA Request.
Your request has been accepted and is being processed.
This page will be updated in a short while; please wait...
PDBe is a member of PDB

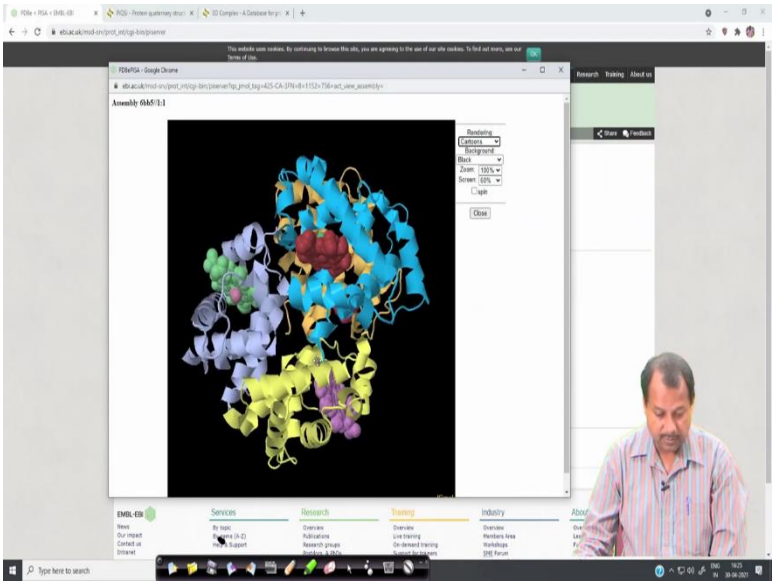

Protein Data Bank in Europe
Bringing Structure to Biology
PDBePISA
Protein interfaces.
You will be taken to results a few seconds.
PDBe is a member of PDB


Protein Data Bank in Europe
Bringing Structure to Biology
PDBePISA
PISA Assembly List.
Probable Assemblies
in PDB 6bb5 crystal.
HUMAN OXY-HEMOGLOBIN
PDBe is a member of PDB

b



b



So, let me see any other protein structure. So, 2 amino acid chains in PDB and 4 ligands in asymmetric unit. And processing ligands heme and oxygen do you wish to process, if no, then I am unchecking this, and then I am going to the assemblies. So, it is taking a little bit time because it is a web service. So based upon the workload and the load of this particular PDB ID that I have selected, so it will take some time. So, it is telling that 6 processors are being utilized. CPU 1 through 6.

I got the result. So, you remember that initially, the suggestion there are two chains in a PDB. So, its PDB ID is 6 BV5. It is human oxyhemoglobin. So, two chains was there initially now it is predicting that actually the size is 4, so it is A2B2 which means 2 A chains and 2 B chains. Now, if I view, I will see there are four chains. See, there are four chains with four

different colors. Now, if I go to the which cartoons you can see, the solid spheres are indicating the hetero atoms. So, no need to look at that one.

So, you see there are four chains. But actually, in PDB there are two chains. So, you may be mislead, by the information at the PDB. That is why, we are basically discussing this particular topic. And also, we are proposing that the kind of work you can do and the algorithm which exists to take care of this fact.

Means, say you take the information of space group you need cell atomic coordinate for the crystal structures, and the new predict what are the most likely biological assembly and what are the coordinates for that one. So, you generate the total quaternary structure out of the protein crystals. So that is what we have discussed.

(Refer Slide Time: 31:25)

Now, going back to the slide, actually. So, similar to that IPACdb you can check it out in the website. So, this also provides a nice interface, but the update has stopped for a long for this IPACdb. So that is it. Thank you very much.