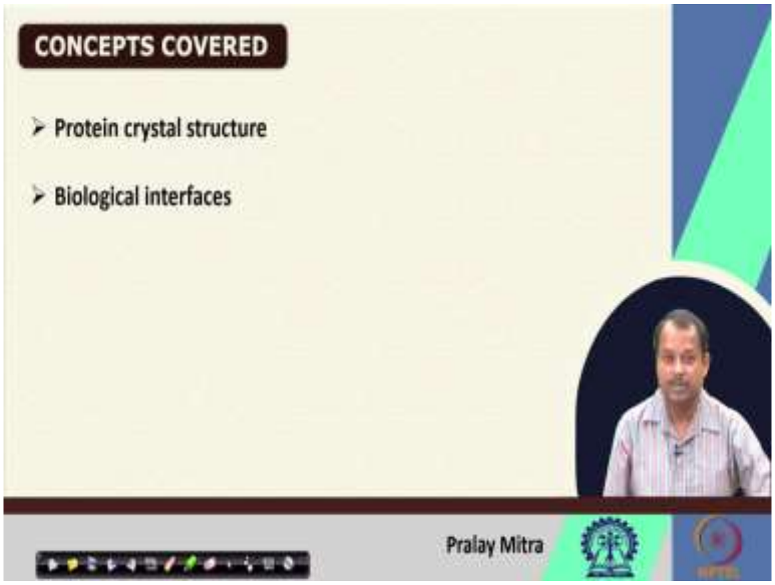
Algorithms for Protein Modelling and Engineering
Professor Doctor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 35
Discriminating Biological Protein Interfaces from Crystal Artifacts (Contd.)

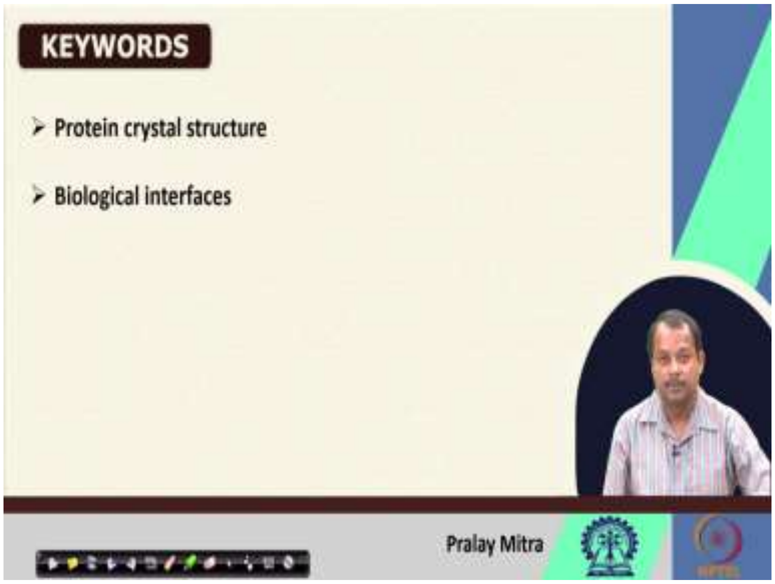
Welcome to everybody. So we are continuing with the Discriminating Biological Protein Interfaces from Crystal Structures. So on the last lecture I mentioned about 10 features that will be useful for us to integrate into some machine learning technique so that we can discriminate or infer that whether the interface is a crystal lattice or not, I mean, in other way, whether it is biologically relevant or not from the crystal lattice.

(Refer Slide Time: 00:41)



CONCEPTS COVERED

- Protein crystal structure
- Biological interfaces



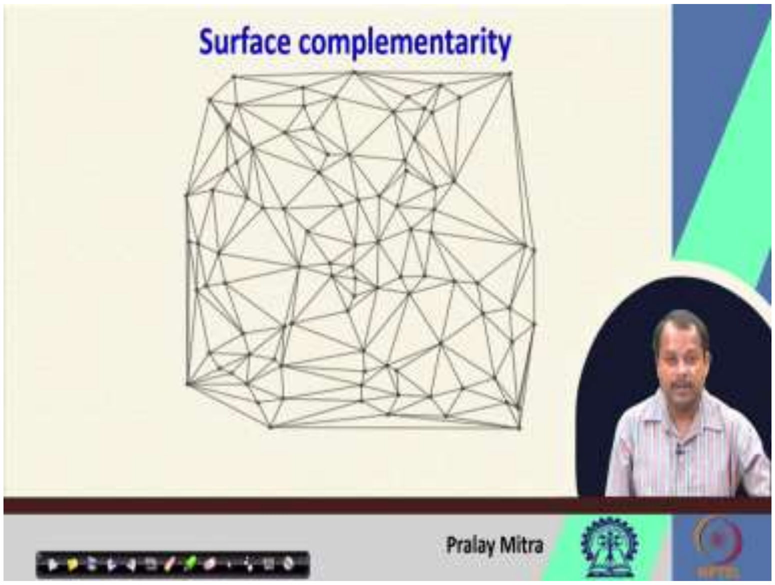
KEYWORDS

- Protein crystal structure
- Biological interfaces

Pralay Mitra

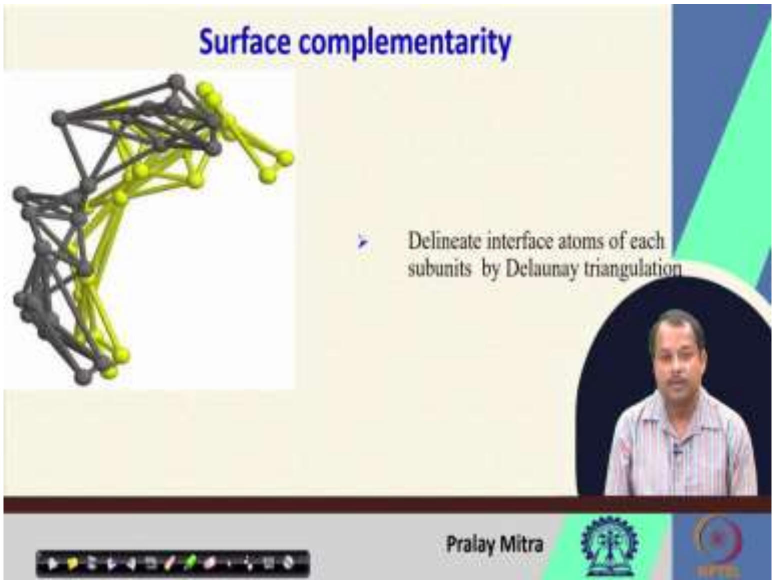
So, the concepts we will be covering, protein crystal structure and the relevance as a biological interfaces. We are only working on that one. Accordingly I picked up the keywords and I keep them, I keep them same.

(Refer Slide Time: 00:55)



Surface complementarity

Pralay Mitra



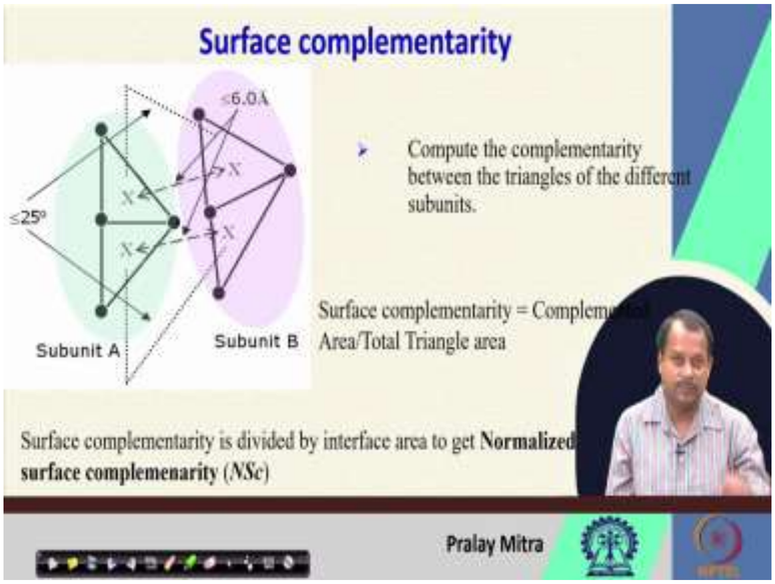
Surface complementarity

- Delineate interface atoms of each subunits by Delaunay triangulation

Pralay Mitra

So, on the last lecture we were discussing regarding the surface complementarity, where after identifying the interface atoms we provide some transformations, and after the transformations we compute the Delaunay triangulations of the atoms for one, for each subunit separately. And then what we are getting is basically the set of triangles which are formed here. Now gray color indicates the set of triangles from one subunit and yellow indicates the, and yellow indicates the Delaunay triangulated atoms from another subunit.

(Refer Slide Time: 1:41)



Surface complementarity

- Compute the complementarity between the triangles of the different subunits.

Surface complementarity = Complem
Area/Total Triangle area

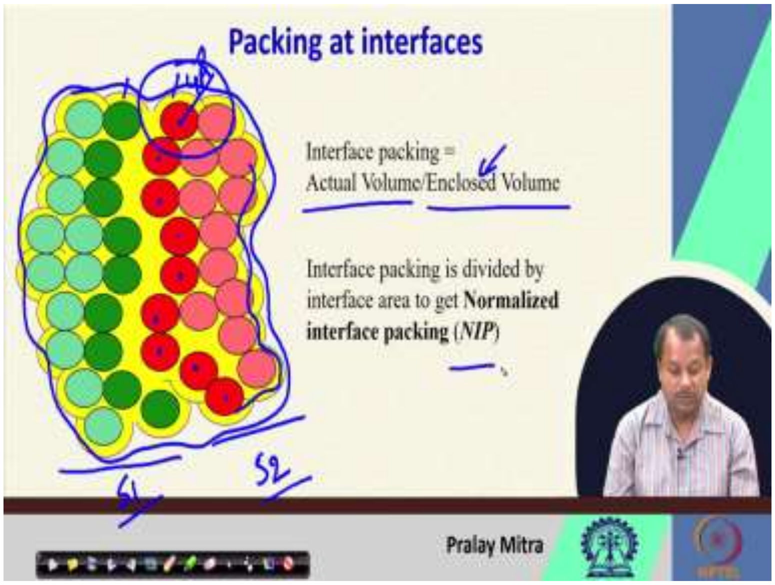
Surface complementarity is divided by interface area to get **Normalized surface complemenarity** (*NSc*)

Pralay Mitra

And the threshold we are using that between in two triangles, one triangle is taken from one subunit and another triangle is taken from another subunit. In this diagram actually the greenish is one and pink is another subunit so two triangles facing with each other. If the plane containing the two triangles are within 25 degree and the centroid between the two triangles are within 6 Angstrom then we declare that two triangles are complemented with each other.

We identify how many such complementary triangles are there. All the areas of such complemented triangles are summed up. When it is normalized by the total area of all the triangles then basically we will get the surface complementarity. And if I divide it further by the interface area then we will get normalized surface complementarity, in short it is NSc. That is our third feature in the 10 feature list.

(Refer Slide Time: 02:40)



Packing at interfaces

Interface packing = Actual Volume/Enclosed Volume

Interface packing is divided by interface area to get **Normalized interface packing** (*NIP*)

Pralay Mitra

Second feature that we considered is the packing at the interface. Now regarding that one again we are, say I am suggesting you to look at the interface. Now you see our main intention is that, so there are several contact areas, few are biologically relevant, few are crystal artifacts. Now when it is biologically relevant the fitting is better like this, when it is crystal artifact the fitting is less. Now in order to check the fitting, so we identified all the features. So packing at the interface is one of them.

So, what we have done that we identified the interface atoms, and for computing the surface complementarity the same set of atoms we are using. In this case so green and greenish indicate one protein or one subunit or one chain. Say that is my, or say S1 and red and reddish indicates another subunit that is my S2. This dark green and dark red color indicates those atoms are at the interface.

So, what I did that, so what basically you have to do that this red color, with respect to this red color you take one slice. So regarding this slice you can consider, so keeping this atom at centre I will draw one sphere whose radius is, say 4 Angstrom degree. And if that is within that one then we will include that. So similar to this here, here, here, here so I will draw one sphere and all those which are within this sphere I will include that.

Now, when we are including, so red will include reddish, means red will include the atoms from subunit 2. Green will include the atoms from the subunit 1. So it will not be like, say red will include the atoms from green, and green will include the atoms from the red. That way if we consider then what will happen? That we will have one slice something like this.

Next what is our intention? In order to check the packing, so we know that these many number of atoms are there. We know what is the van der Waal radii of these atoms. Since we know the van der Waal radii of these atoms so we can compute that how much space is occupied by all these atoms.

Now, you consider this situation. One is I am considering as the better fitting, another is the loosely fitting. So if it is better fitting then the intermediate spaces in between say red and green, in this region will be less. If it is loosely fitting there will be more because the fitting is not proper in nature. So that is why, what we will do next is that on this slice actually we will roll a probe using the concept of calculating the surface area of a molecule.

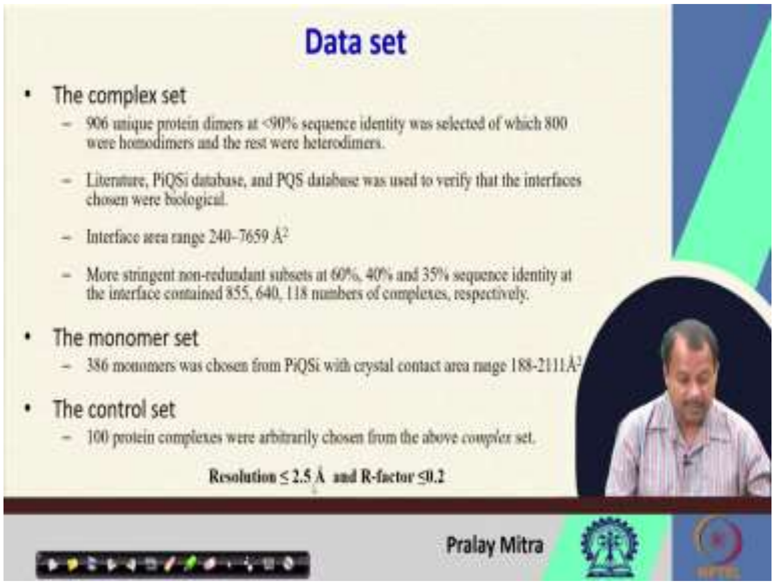
In this case it is not the surface area of a molecule rather it is a surface area of this slice, green and greenish, red and reddish. So only that part I have taken and I will roll a probe on over that one. When I will role the probe over that one then what I will get basically? That I will get one area something like this.

That is nothing but the area mostly covered by yellow color. That is actually our following the trajectory of the probe, now if the probe size is very small then it will almost capture the information of fitting the surface properly. That way we are having two volume, one is the volume of the yellow color. Another is the sum of volume of green, greenish, red, reddish.

Actual volume that is green greenish, red, reddish divided by enclosed volume that is the volume as measured by the yellow color is basically defined as the interface packing. Now in this case, since I am rolling the probe then you can very much understand that I am basically getting the surface area, the surface area of all the surface protein atoms. I am getting that one.

If I consider that surface area is the surface area of a sphere then actually from there I can calculate that what will be the volume of that sphere and that is my enclosed volume. So that sort of small approximation is done in order to check the packing at the interfaces. Finally interface packing is divided interface area to get normalized interface packing that is my NIP. So that way I finish discussion of 10 features that I mentioned will be useful for my purpose.

(Refer Slide Time: 08:05)



Data set

- The complex set
  - 906 unique protein dimers at <90% sequence identity was selected of which 800 were homodimers and the rest were heterodimers.
  - Literature, PiQSi database, and PQS database was used to verify that the interfaces chosen were biological.
  - Interface area range 240–7659 Å^2
  - More stringent non-redundant subsets at 60%, 40% and 35% sequence identity at the interface contained 855, 640, 118 numbers of complexes, respectively.
- The monomer set
  - 386 monomers was chosen from PiQSi with crystal contact area range 188-2111Å^2
- The control set
  - 100 protein complexes were arbitrarily chosen from the above *complex* set.

Resolution ≤ 2.5 Å and R-factor ≤0.2

Pralay Mitra

But one interesting fact I wish to draw is that these two features, this normalized surface complementarity and normalized interface packing actually has a very good correlation among them. And the correlation is 0.95. That is huge. And on which dataset it is done? The dataset is 906 unique protein dimers at less than 90 percent sequence identity was selected of which 800 was homodimer and rest were heterodimers.

Please note it down that since we are trying to identify whether the crystal interfaces are the biological one or crystal artifact so we will not rely on the PDB interfaces. Rather we have to consult to other databases like PISA, PiQSi or literature and through which you need to verify.

S,o that is why when you are, say developing this, this kind of algorithm or, say machine learning tools or techniques you have to be very much careful about the biasness of the dataset. I believe you are aware of this biasness of the dataset in the machine language. Here also in order to avoid the bias, so we are not considering the data from the PDB directly because we are admitting that in PDB the data is not correct.

So, initially the data has been taken from the PDB but they are verified with respect to the other databases which infers correct biological interfaces like literature or say PiQSi or PISA, sometimes PQS also. And in the dataset interface area range 240 to 7659 Angstrom was present.

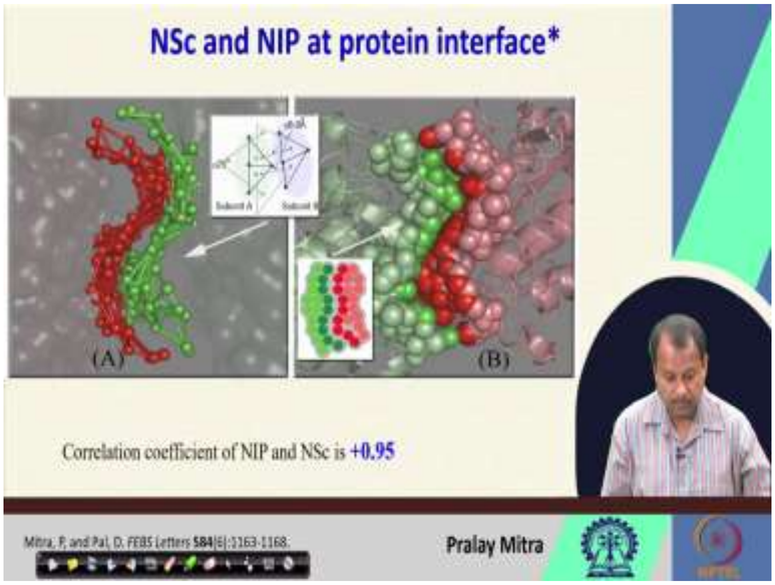
So, this particular information is important to inform you that, to inform you that we will demonstrate that interface area is a factor but it is not the sole factor to determine or to conclude that whether the interface is biological one or not. So more stringent non-redundant subsets at 60 percent, 40 percent and 35 percent sequence identity at the interface contained 855, 640, 118 numbers of complexes respectively are there.

So, we are also using some control set and the monomer set. Why? This is important. So you are basically using the complex that is your positive cases. But when you are using the positive cases how do you know that whether there is a bias or not, whether always you are getting the same value? So when it is a complex so it is not possible.

That it is why three different datasets you should use in case of benchmarking this kind of algorithm. One is the correct positive one, another is the wrong which is the monomer, another is the control which will be, say created taking, starting with the positive dataset and arbitrarily removing some information, I mean the atoms from the interface, so making one artifact.

Now, this control sometimes will be treated as a positive, sometimes it will be treated as a negative. But positive will be in one side, negative will be in another side. I mean dimer will be in one side, monomer will be another side, and in between it will be filled with the control. So that way you have to benchmark your method once you will design something like this.

(Refer Slide Time: 11:54)



NSc and NIP at protein interface*

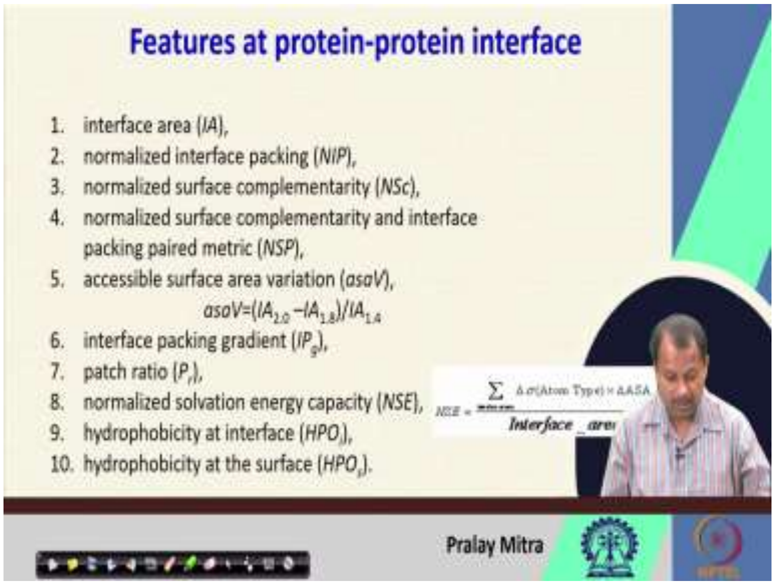
Correlation coefficient of NIP and NSc is +0.95

Mitra, P. and Pal, D. FEBS Letters 584(6):1163-1168.

Pralay Mitra

So, this I talked about.

(Refer Slide Time: 11:55)



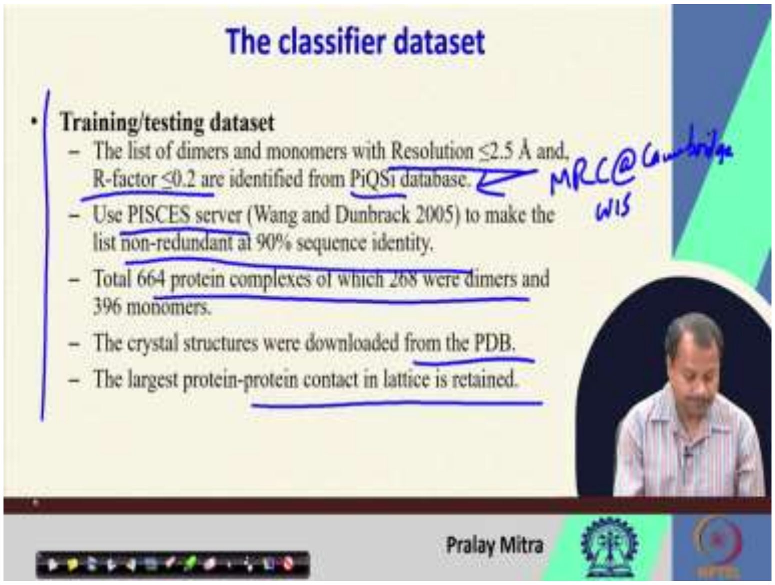
Features at protein-protein interface

1. interface area (*IA*),
2. normalized interface packing (*NIP*),
3. normalized surface complementarity (*NSc*),
4. normalized surface complementarity and interface packing paired metric (*NSP*),
5. accessible surface area variation (*asaV*),
   $asaV=(IA_{2.0}-IA_{1.8})/IA_{1.4}$
6. interface packing gradient (IP_g),
7. patch ratio (P_r),
8. normalized solvation energy capacity (*NSE*), $NSE = \frac{\sum \Delta\sigma(\text{Atom Type}) \times \Delta ASA}{Interface\_are}$
9. hydrophobicity at interface (HPO_i),
10. hydrophobicity at the surface (HPO_s).

Pralay Mitra

So, this way 10 different features at the protein-protein interface we discussed.

(Refer Slide Time: 12:03)



The classifier dataset

- **Training/testing dataset**
  - The list of dimers and monomers with Resolution ≤2.5 Å and, R-factor ≤0.2 are identified from PiQSi database.
  - Use PISCES server (Wang and Dunbrack 2005) to make the list non-redundant at 90% sequence identity.
  - Total 664 protein complexes of which 268 were dimers and 396 monomers.
  - The crystal structures were downloaded from the PDB.
  - The largest protein-protein contact in lattice is retained.

MRC@Cambridge
WIS

Pralay Mitra

Now, we are going to classify or discriminate biological interfaces from the crystal artifacts. How we can do that one? So the training and testing dataset, the list of dimers and monomers with resolution less than 2.5 Angstrom and R-factor 0.2 are identified from PiQSi databases.

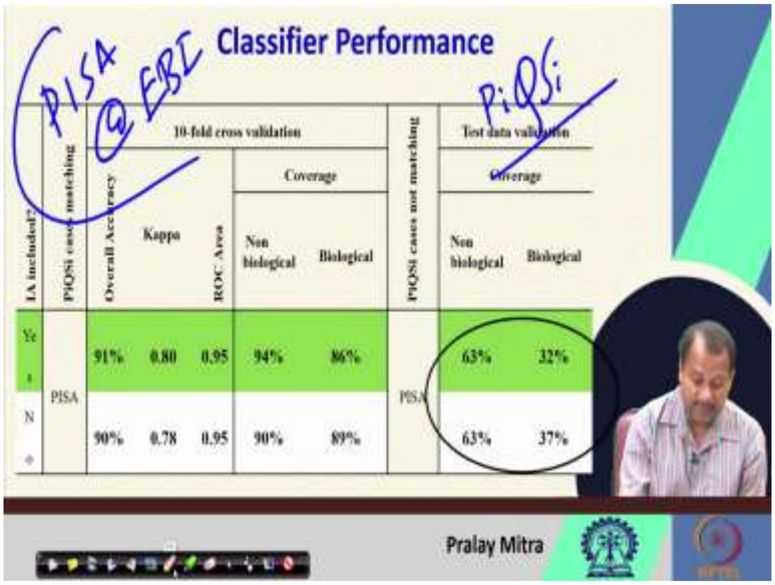
So, PiQSi is a database which was initially hosted at MRC Cambridge but now since the author has moved out so now it is in Weizmann Institute of Science. It is initially done by Emmanuel Levy. So this PiQSi, now this resolution and R-factor is also important. These two factor are, will tell you that whether the structure you are considering is with the good quality structure or not, because regarding the resolution I also mentioned that if, say resolution is 6 Angstrom, 7 Angstrom for protein crystal structure then there is no guarantee that the structure is going to be very good.

So, this kind of restriction on the resolution and R-factor is used mostly everywhere. And there is a very nice web server. So that is called as the PISCES web server develop by Wang and Dunbrack in 2005 to make the list non-redundant. There are several options but for our purpose what you can do that you make that list non-redundant at 90 percent sequence identity.

Then total 664 protein complexes of which 268 were dimers and 396 was monomer is the final dataset for our classification. The crystal structures were downloaded from the PDB and the largest protein-protein contact in lattice is retained. So as of now we are also focusing on the interface which, we are focusing on the contact area which are with the larger size. But later we will change it and we will demonstrate that if it is not then also it will work.

So, this is the dataset. You have to be very careful in selecting the dataset because if there is any bias or if there is any, say wrong representation then you will not be able to judge whether your classifier is doing good or not. `

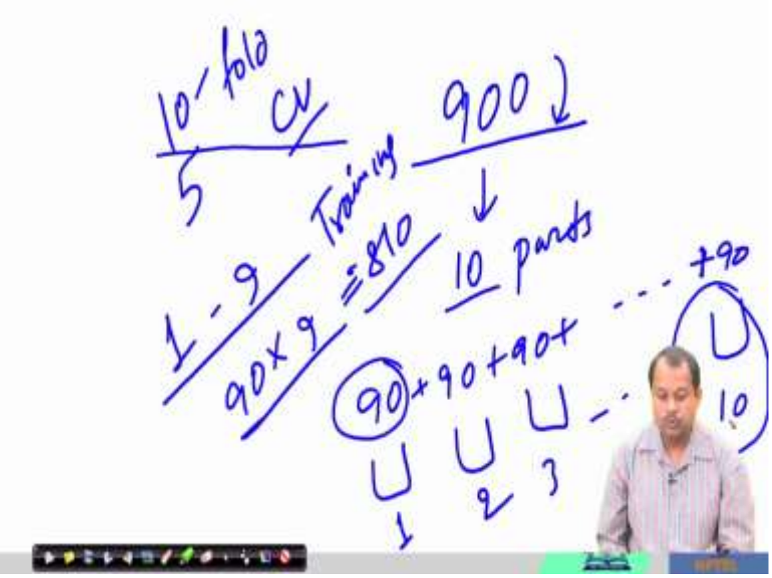
(Refer Slide Time: 14:30)



Classifier Performance

IA included?	PIQSi cases matching	10-fold cross validation					PIQSi cases not matching	Test data validation	
		Overall Accuracy	Kappa	ROC Area	Coverage			Coverage	
					Non biological	Biological		Non biological	Biological
Yes	PISA	91%	0.80	0.95	94%	86%	PISA	63%	32%
No		90%	0.78	0.95	90%	89%		63%	37%

Pralay Mitra



So, if you use the classifier, so before showing you the performance so there are several machine learning techniques, so to, in which you can basically feed your features and get the classification result. So, if you start with the simplest one that Naive Bayes then what will be the accuracy? That I will show that first.

So in 10-fold cross validation, so the overall accuracy is 91 percent, 90 percent. Kappa value is 0.80. 0.78. ROC is radii under the curve, so ROC area is so 0.95, sorry 0.95. And coverage for the non-biological and biological is listed here. And on the test set which is the PiQSi cases not matching PISA, so what is the coverage? That is presented here.

So few things I would like to tell you that what is this 10-fold cross validation? 10-fold cross validation says that if you have a dataset, if you have a dataset which consists of, say 900 or, say 900 is also fine, so 900 data. Then in 10-fold cross validation what you need to do, that you have to divide it into 10 parts, which means 90, 90, 90. This division will be random in nature. So what you will do, at as if you will consider that there are 10 buckets, 1, 2, 3, 10; 10 buckets.

Now out of these 900 randomly you pick one or, say you pick one not randomly. So you pick one and randomly you put in one of the 10 buckets. And at most 90 will be there so evenly it will be distributed. So if it is not divided by 10, so then extra 1, 2 you can, say add in one of the buckets.

After doing that one you consider 1 through 9 buckets, which means 90 cross 9 or 810 data or instances for training, training your machine learning model. And only one bucket, maybe the 10th bucket you keep aside for your testing purposes and you report that what is your accuracy or what is your other measures. So that is called as the 10-fold cross validation.

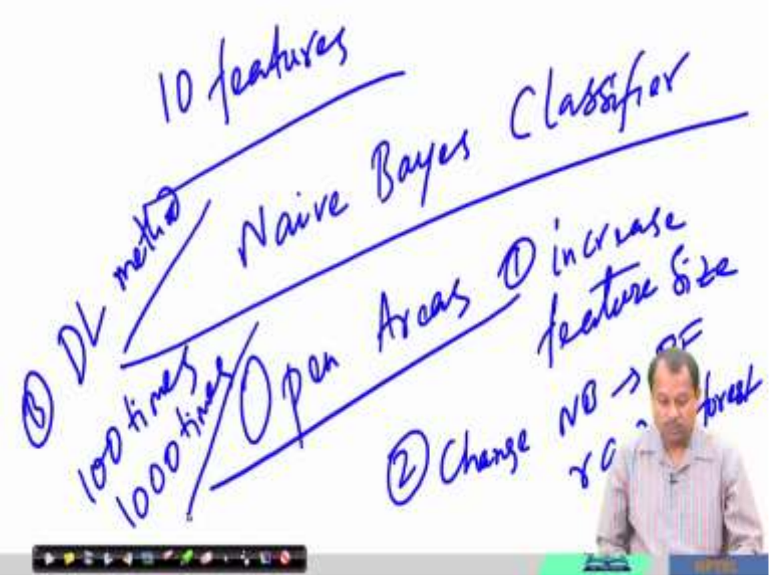
Now if it is 5-fold, accordingly what will happen? Instead of 10 buckets there will be 5 buckets. Randomly you put on 5 buckets and then you basically finished the calculation. Now during the implementation it may possible that picking one from 900 and then putting that in one of the buckets randomly may not give, say even distribution. So what you can do that you pick serially, so one from 900, sorry you pick randomly from 900 and put one.

So after 90 such iterations or 90 random picking then another 90 picking for bucket 2, another 90 picking for bucket 3, that way you will go. So the second one will be better compared to the first one from my point of view, and from implementation point of view also. So that is 10-fold and 5-fold cross validation.

So after this cross validation you can go for the actual validation. So here the cross validation is done and then when the model is developed then what is done, that test data validation is also done, and the accuracy of the test data is also shown here. Now while you are doing this one then two database you can consider for your purpose, one is the PISA which is hosted at the EBI, the link I will provide on the next class or next lecture.

Another is PiQSi; initially it was at MRC Cambridge. Now it has moved to the author's place. That also, you can Google actually PiQSi then you will get it readily. So part of this PiQSi is manually curated whereas PISA is totally automatic in nature. So you can classify that one.

(Refer Slide Time: 19:30)



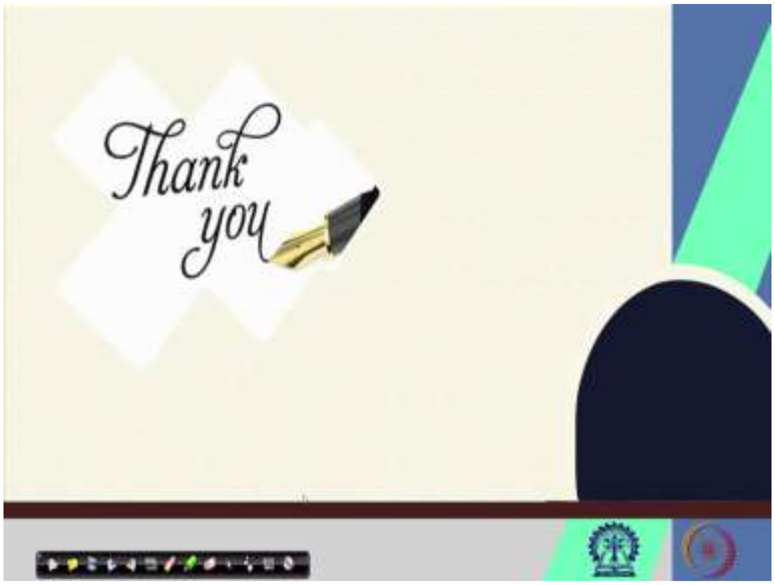
Now, the point is, as of now what I have discussed is based upon 10 features, and Naïve Bayes classifier. But the modification what you can do? Open areas, increase feature size, change Naïve Bayes to, say random forest or use deep learning method. So those changes you can do, and based upon those changes actually the performance will also change.

But while you are doing that one be careful about choosing the dataset, so positive dataset and the negative dataset and accordingly you pick that one. But you will see that this classification, the accuracy will increase, so everything is fine, but along with this one if you incorporate the symmetry information that we have discussed then your accuracy will increase very much. So we will integrate that accuracy information, sorry we will integrate that symmetry information along with this classifier.

So that we will discuss on the next lecture, but before that, before closing this lecture I wish to point out one more fact in front of you that regarding the 10-fold cross validation or, say 5-fold cross validation, in order to make your system unbiased it may be a good idea, that instead of one run of 10-fold cross validation or, say 5-fold cross validation you go for several runs, say 100 times or 1000 times, 100 times or 1000 times run for 10-fold cross validation and in each step, what is your accuracy?

You take the average and standard deviation and you report that one. That is much better. And since you are going for random picking etc so that way you will get more, more converged result or unbiased result compared to just one run. Again the random number is coming here. So you understand the importance of having one library function with a very good random number generator. So do not miss that one.

(Refer Slide Time: 22:27)



Now, before we conclude, so to summarize what we discussed in this week is that we started with the SCOP class, the structural classification of proteins that we discussed. In that discussion we included that different layers of classifications are required or hierarchy of the classifications are required instead of one flat level classification, and the classification should aim to functional classification rather than just sequence or structure level classification, because at the end of the day the function of the protein is of our interest, not anything else.

And since function is directly related to structure and to some extent sequence so we will not keep our eyes closed when there is a sequence or structure. That is why the recent semi-automated classification techniques is adopted in SCOP although initially it was started based upon purely manual curation, but now, so first sequence level analysis if there is a 100 percent hit, so I know for sure that it is going to be in that particular category or class.

If there is some structural alignment, so based upon that one I know perhaps this is going to be the correct class. That way, after the classification if we do not find any similar sequence or structure which is already classified and is in SCOP database then for that we need to go for some manual curation.

Now during this process it is encountered that number of, there are several membrane proteins are there, there are several designed proteins are there, small proteins are there without any valid fold and since fold also is one part of the classification, so those proteins should be kept aside and those will be discussed separately.

Now after the scope we discussed about the symmetry of the protein. So there are three different point group symmetry which are there, cyclic, dihedral and then cubic. But cubic occurs for large organization of the protein when the number of chains are 12 or more. So that is why we do not include that in our discussion because we are focusing mostly on globular proteins of, say up to octamer or hardly decamer, I mean 10 chains are there.

Now while looking at the symmetry, so we understand the major concern of symmetry is identifying the symmetric axis and once you will identify the symmetric axis then giving a rotation about that symmetric axis and then after aligning that one, we know that alignment algorithm, so getting that what is the amount of error, based upon that one we can conclude whether the symmetry exists or not.

Cyclic is probably the simplest one but dihedral also exists. If I go higher organization, higher assembly like tetramer or, say octamer, or hexamer D2 D3 D4 is possible. Now in the dihedral symmetry along with the cyclic, reflexive may also be present about the plane or with respect to the plane. About means where there is there is another cyclic, or with respect means that when one plane is there and you are looking for the reflexive symmetry. We discussed that one.

Then we move on to another interesting topic. We started. We will continue in the next week, that is discriminating the crystal artifacts from the biological one, specifically from the crystal structures. So crystal structures are mostly present in the Protein Data Bank. So I mean that you will see that 80 to 90 percent structures in the Protein Data Bank is crystal structures.

So it is a very good idea to have one software in place which will take one crystal structure as an input, I mean the atomic co-ordinate information, its space group information, its unit cell information it will take as an input. It will generate the whole lattice. It will identify how many contact areas are there and then it will discriminate that which interfaces are biologically relevant and which are not. That way it will identify a subset of interfaces.

But we are not yet done with that because we do not know what will be the organization, means whether it will be biologically function as a dimer or trimer or tetramer that we did not discuss. That we will discuss in our next class.

While building the machine learning based classifier for discriminating the biological interface from the crystal artifacts, so there are 10 features we are concentrating, which will look after the complementarity, compactness, hydrophobicity, hydrophilicity along with the solvation energy and the contact area.

But you can include more features in there to make it more accurate. Also instead of the Naïve Bayes, you can use a random forest or any other machine learning, classification, binary classification technique for identifying the crystal, for discriminating the crystal artifacts from the biological one.

So this is the first step. The next step will be, so we will have an algorithm which will combine these two informations, say I have identified this is biological and I have this symmetry information. How to combine that one in order to know that what is the biological form, I mean a dimer, trimer, tetramer, etc. So that is it for today. Thank you very much.