

Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 36

Discriminating Biological Protein Interfaces from Crystal Artifacts (Contd.)

Welcome to everybody. So, in this week, we are planning to continue the discrimination, discriminating biological protein surfaces from crystal artifacts. So, the discussion has started on the last week, and we are continuing here. So, as I mentioned on the last week that we will integrate the classifier, the binary classifier that we discussed on the last class along with the symmetry information in order to have one computational framework, which will take one protein structure which is solved experimentally using X-ray crystallography as an input and will output its functional form or the quaternary structure which is possible.

(Refer Slide Time: 01:02)

The slide is titled "CONCEPTS COVERED" in a black box. Below the title, there is a list of two items: "Protein crystal structure" and "Biological interfaces", each preceded by a blue arrowhead. To the right of the text is a hand-drawn diagram in blue ink showing several interconnected shapes: squares, circles, and rectangles, some with hatched patterns, representing protein interfaces. The initials "TCR" are written in blue in the top right corner of the slide area. In the bottom right corner, there is a circular video inset showing a man (Pralay Mitra) speaking. At the bottom of the slide, there is a toolbar with various drawing tools, the name "Pralay Mitra", and logos for IIT Kharagpur and NPTEL.

So, to continue the concept that we will be covering is crystal structure, protein crystal structure and biological interfaces. So, to give a recapitulation of what we have discussed on the last week. So, basically, the problem in crystal structure is that sometimes it does not give you the biological or functional form. So, whatever is deposited in the protein databank may not give you the correct interface.

The reason is very simple that when say, we are solving or I mean the scientists are solving the protein structure using X-ray crystallography then definitely first step is to get the solution then they crystallize that solution, and the structure which is determined is the structure which is available in the crystal. And also in the crystal, it forms lattice that is why instead of only one contact, we will get a series of neighbors. So, that neighbor information.

So, from that neighbor information, definitely, we have to get which is the biological form either experimentally or computationally. Most of the cases the experimentalists who determine the protein structure using X-ray crystallography perform some sort of computational analysis, and then mention that which is the most likely biological form in the remark section.

So, we discussed in the protein databank structure file format there is one remark section, remark 1, 2 up to like that way. So, in remarks section they mentioned what is the biological form. Optionally, they also provide the transformation matrix, which will be applied on the structure provided in the protein databank in order to get that biological form. But it is not true or sometimes it is not correct also for all the crystal structures, which are deposited in the protein databank So, we have to come up with some idea or some computational technique which is fast enough, as well as, can able to predict that what is the biological form. Whether it is Dimer, Trimer, Tetramer, etc that we have to know.

Now, in order to do that one fast step, we discussed, that is, when we have say several number of contact areas among those we discriminated that which is the correct interaction or interface and which is not. That way, if I look on the crystal structure then say the same square and circle example that I gave to you.

If I consider only in 2D, now, out of this four let me tell you that probably this is one correct form. And probably this has also been identified as another correct form. And this has identified as another correct form, this has identified as another correct form. Since it is a crystal so first thing you have to know that from this lattice, so how many contact areas are there. So, that way from one say small unit you can expand it in all the directions and you can identify all the contact areas. Then among those contact areas which are relevant so I am assuming that these are relevant.

Now, out of this relevance. So, up to this we have discussed. So, what we have discussed that when this lattice we have generated in some way actually we can do that one. If we know that what is the space group and what is the dimension of the asymmetric unit and what are the coordinates of the atom then using orthogonalization and de-orthogonalization we can convert that information to say within the actual protein structure into that unit cell then we translate that unit cell, and then, we de-orthogonalize that one in order to generate the lattice. After generating that one, we can identify the contact area.

While we are identifying the contact area one fortunate question may come that how do I know whether two contact areas are equivalent or not? That is a very fortunate question. I agree with you. Say, for example, in this case, so this interface is similar to this interface, but that is my visual inspection. So, when in number of coordinate is given to you and the chains are say separated by TER or TER then TER. Then how do I know that what is, which interfaces are equivalent.

Now, this question we are deferring right now, we will discuss in more detail in this week itself, but when we will discuss protein docking again. Now, let us assume that it is not there. So, what I can do, I you will generate so spurious or extra information like say, if these are equivalent then both will be cancelled out because, if they are not say biological, if they are crystal artifacts or if they are biological then both will be selected like that way say few will be selected.

Now, up to this we have done, and for that, we have discussed some machine learning technique and we also discussed that machine learning technique again can be evolved or it can be improved for the better accuracy. Now, after we got that information then, how that information will be combined along with the symmetry information so that I can have one computational framework that using this as an input can tell me, okay, this will create one tetramer and this is going to be their structure. So, how is it possible? So, that we will see.

(Refer Slide Time: 07:48)

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),
 $NSE = \frac{\sum \Delta \sigma(\text{Atom Type}) \times \Delta ASA}{\text{Interface area}}$
9. hydrophobicity at interface (HPO_i),
10. hydrophobicity at the surface (HPO_s).

+0.95

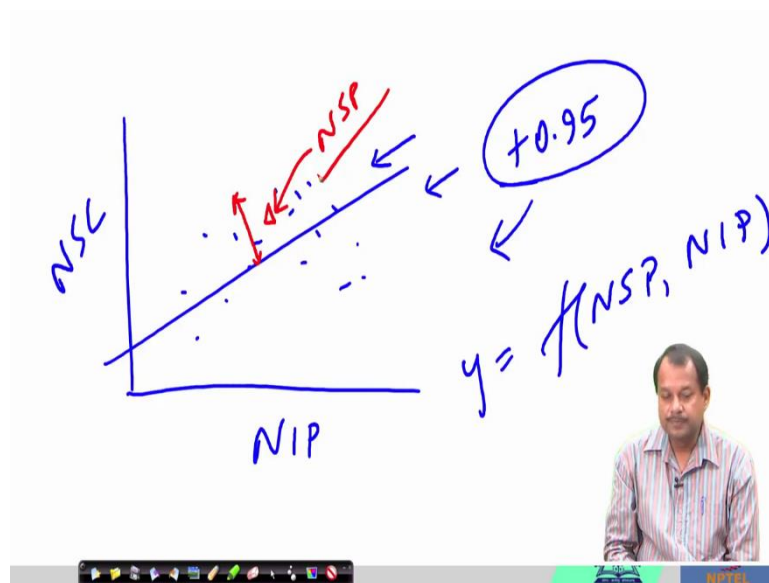
Pralay Mitra

But before looking at that one we need to we need to do some analysis on the feature set. So, these are the 10 features that we have considered. And among these 10 features, I demonstrated these two features are highly correlated. The correlation among NIP and NSc is

denoted as 0.95. And you remember that when there is a correlation then you can consider that as if along the X axis I have NIP along the Y axis I have NSc and the points are indicating their values.

So, if I try to fit one straight line, then I will get one equation from the straight line. So, that equation of the straight line will give me this NSP, which is Normalized Subspace Complementarity and interface packing paired metric. That metric will tell that when we fit one straight line, then, with respect to that straight line, if a new point will come how far it is, that deviation we will compute.

(Refer Slide Time: 09:03)



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),

$$NSE = \frac{\sum \Delta \sigma(\text{Atom Type}) \times \Delta ASA}{\text{Interface area}}$$
9. hydrophobicity at interface (HPO_i),
10. hydrophobicity at the surface (HPO_s).

Pralay Mitra

What I am trying to say is, if I make a plot say NIP and NSc. Now, on the known data set for which I specifically know what is monomer what is Trimmer and a control is there that we

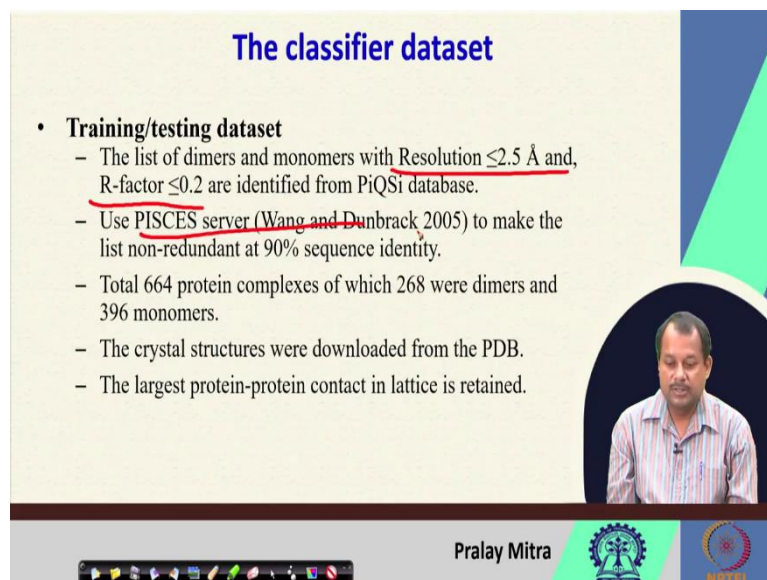
have discussed on the last week that, monomer is one side dimer is another side, and in between is control set and a good classifier can able to classify them correctly.

So, if it is and corresponding to each interface, I have some points. And since I know that between NIP and NSc, there exists a very high correlation, so I can fit a straight line between this. So, when I fit a straight line, then I can have one function in terms of this NIP and NSc. I can have one function. Now, let us assume one new point has come.

I mean, one contact area has come for which we computed what will be the NSc and NIP. Now, the dots in this plot is indicating each dot is indicating corresponding to one contact area what is the value for NIP and NSc? Now, if one new contact area will come then corresponding to that contact area, I can compute what is the value for NIP and NSc. Let us plot it here.

So, this is my the contact area, the NIP and NSc corresponding to that contact area. Then I can compute what is the deviation. Means, this blue straight line is my theoretical line. So, how far it is going from that theoretical prediction. So, that delta will be computed in NSP. So, that deviation will tell me how good is the are in terms of the normalized surface complementarity and normalized interface backing. So, that is my that is my fourth that is my fourth parameter and rest of the parameters we have discussed.

(Refer Slide Time: 11:30)



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.

Pralay Mitra


Logo of IIT Delhi and IITM

Input:
A set of n protein sequences

Output:
Retain m ($m \leq n$) protein sequences
where between any two protein
sequences the sequence identity
is at most 90%

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.



Pralay Mitra

So, we see some parameters, which are correlated with each other. Now, there may be some other parameters also they are correlated. Now, if they are correlated with each other then it might be a good idea to check they are individual discriminative power. Say, instead of 10 different features if I take only one then how good they are in discriminating the interface between crystal artifact and the biological interfaces. So, these kinds of analysis are very much required when you are identifying some features to feed it into some machine learning technique, and to get a machine learning a prediction system.

So, in order to do that one, I mean, to check that individuals have discriminative power and also to ensure, that the features we are calculating is not kind of biased, which means, only correlated features are not there. So, if the correlated features are there then there is a possibility that the prediction or discrimination will be biased.

If I assume that one, then, we have to analyze my feature and its predictive power. So, in order to do that one, so, we designed one data set first. So, the list of the Dimers and monomers marks with resolution less than 2.5 angstrom and R factor less than 0.2 are identified from PiQSi database. So, the reason is now clear to you that why resolution is less than 2.5 and R factor 0.2 because it ensures that we are getting good quality structures.

After that one, we are running it in the using the PISCES web server where we can actually from that set, we can retain all the sequences where between two pair of sequences the maximum sequence identity is 90 percent. If there exists more than one sequences with which where the sequence identity is more than 90 percent one of them will be removed. That way, we make it non-redundant at 90 percent secure sequence identity by running it through PISCES web services. But by this time, I guess, you are also enough, you also have enough experience or exposure through this course to write it by yourself.

Say, for example, you can consider this as a homework that if I give you a set of n protein sequences output is retain m protein sequences, where, between any two protein sequences the sequence identity is at most 90 percent. Now, this 90 percent I put in a box indicating that you can change this. So, this 90 percent, based upon your requirement you can make it to 80 percent to 70 percent, as per your wish.

Now, if you do that one then you did not have to use these PISCES web service, but you can have your own. Next, total 664 protein complexes of which 268 white dimers and 396 was where monomers has been identified. Now, the crystal structures were downloaded from the PDB that is the primary source for our proteins with the structural information, the largest food input in contact in lattice is returned. Again, we are keeping the largest protein-protein contact.

(Refer Slide Time: 16:29)

Classifier Performance

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

So, this is the performance of our classifier that we have discussed on the last week. So, I am not going to details of that one. But before going to individual or feature wise analysis, let us discuss a little bit about the different measures in the context of machine learning or in the context of classification.

(Refer Slide Time: 16:50)

Measure for classifier

		True condition	
		Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive (TP)	False positive (FP), Type I error
	Predicted condition negative	False negative (FN), Type II error	True negative (TN)

TP → hit
 TN → correct rejection
 FP → false alarm, type I error or underestimation
 FN → miss, type II error or overestimation

So, first, at the beginning, we should discuss about the confusion matrix. So, here it is actually a 2 cross 2 matrix, so, this is the size of the matrix. This is also called as the confusion matrix. So, where there are, along the columns there are two conditions, and along the rows there are predicted conditions. Now, there will be since it is a binary classifier there will be two outcome for the true conditions, either condition is positive or condition is

negative. So, here positive or negative and the predicted condition can be positive or predicted condition can be negative.

Now, since it is a true condition, so, let me phrase it like this then it will be easy also for discussion. Sorry. So, this is my true condition and this is my predicted condition. So, there are two outcome, as I mentioned, since it is a binary classifier, so, either outcome is positive or negative. And for the prediction also either I will predict, as positive or negative. Now, two for the true case and two for the predicted cases. Combining this I will get four different conditions.

When I am telling that, okay, the prediction says that it is true condition positive, predicted condition positive and true condition also says the condition is positive. So, that is called as that true positive or TP. When I am predicting as condition positive, but actually the condition is negative then, I mean, the true condition is negative then it is called as a false positive. So, I am predicting predicated condition is positive, but the true condition is actually negative, which means false so it is false positive. That is also called as that type 1 error.

Now, type 2 error will be just reverse. Means, when I will predict as a negative, but actually it is the true case is positive then that is called as a false negative FP, and that is type 2 error. Another situation, when I am predicting that condition is negative, also the group two condition is negative and that is called Is that true negative or TN. Now, you see, that these two diagonal cases when I am predicting that the condition is positive, and the true condition is also positive that is called true positive. When I am predicting the condition is negative and the true condition is also negative then that is true negative.

So, in TP and TN true positive and true negative my prediction is matching with the true condition. In other cases it is not matching. So, when I am predicting it is positive but actually it is negative then it is a type 1 error or false positive case. Another, when I am predicting the condition is negative but actually the condition is positive, which means, it is a false negative and that it is also called as a type 2 error.

So, the TP indicates hit number of times. I am predicting positive and actually it is positive number of TN indicates correct rejection. So, I am telling that it is negative, remove it, and that true case is also negative, remove it, then FP false alarm type 1 error or underestimation, and FN it means type 2 error or overestimation. So, FP is false alarm, which means, that I am predicting it is positive actually it is negative. So, I am raising okay, it is good it is correct, actually it is negative, so that is a false alarm what type 1 error or underestimation. And

predicted condition negative I am telling no, no, no, it is not correct, but actually it is correct which means I am overestimating, so it is an error for overestimation and type 2 error. So, this is actually miss. It is true case, but I am missing the prediction. So, based upon these TP, TN, FP, and FN other measures will be developed.

(Refer Slide Time: 22:08)

Measure for classifier

Sensitivity, Recall, Hit rate, or True Positive Rate (TPR) = $\frac{TP}{TP+FN}$


Specificity, Selectivity or True Negative Rate (TNR) = $\frac{TN}{TN+FP}$

Precision or Positive Predictive Value = $\frac{TP}{TP+FP}$

Negative Predictive Value (NPV) = $\frac{TN}{TN+FN}$

Miss rate or False Negative Rate (FNR) = $\frac{FN}{FN+TP}$

Fall-out or False Positive Rate (FPR) = $\frac{FP}{FP+TN}$



Pralay Mitra

Measure for classifier


		True condition	
		Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive (TP)	False positive (FP), Type I error
	Predicted condition negative	False negative (FN), Type II error	True negative (TN)

TP → hit

TN → correct rejection

FP → false alarm, type I error or underestimation

FN → miss, type II error or overestimation



Pralay Mitra

So, first, sensitivity or recall or hit rate or true positive rate in short TPR. So, it is TP divided by TP plus FN. So, TP is what, true positive. TP in the denominator is true positive and FN is false negative. If I go back then you will see FP is false positive and FN is false negative. So, FN is type 2 error. Which means, I am predicting negative but actually it is positive. Now that is called as the sensitivity or recall.

Now, specificity or selectivity or true negative rate indicates that TN divided by TN plus FP. So, TN is true negative so actually I am predicting negative, and I am predicting negative and actually it is negative that is true negative, and I am predicting positive but actually it is negative so that is TN divided by TN plus TP that is specificity.

Now, the precision or positive predictive rate is TP divided by TP plus FP. So, the call cases I am predicting as positive with respect to that how many cases is actually positive. Which means, when I am predicting positive and with respect to that the number of cases I am predicted at positive actually how many cases are positive, so that is my precision or positive prediction value.

The negative predictive value NPV is, when I am predicting its wrong, I mean, its false and actually it is false divided by that TN plus FN. Which means, that when I am predicting that it is negative then out of which how many cases are truly negative, so that is negative predictive value. Miss rate or false negative rate is a FN divided by a FN plus TP. So, TP is true positive cases correctly predicted one, FN is the type 2 error, and so I am taking the division. Which means, that how many premiums out of the true cases how many cases are actually missing missed by me during my prediction. So that is that miss rate. Fall out or false positive rate FPR is the FP, the number of cases. I predicted as positive, but actually it is negative divided by a FP plus TN. So, that is my positive, false positive rate.

(Refer Slide Time: 25:17)

Measure for classifier

$$\text{Accuracy (acc)} = \frac{TP+TN}{TP+TN+FP+FN}$$
$$\text{F1 score (is the harmonic mean of precision and sensitivity)} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Mathews correlation MCC

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Pralay Mitra

Next, accuracy is defined as TP plus TN divided by TP plus TN plus FP plus FN. So, TP and TN indicates that, my prediction is correct. So, TP indicates that my prediction is positive, and it is actually positive, and TN indicates that my prediction is negative and actually it is negative. So, in how many cases out of total predictions, in how many cases my prediction is correct, that is my accuracy.

F1 score is the harmonic mean of precision and sensitivity and it is in short written as 2 multiplied with TP divided by 2 multiply the TP plus FP plus FN. Now the Matthews correlation coefficient which used a lot in order to check the discriminating power of a feature is defined as MCC equals to TP multiplied with TN minus FP multiplied with FN, FP multiply with FN, divided by square root of TP plus FP, TP plus FN, TN plus FP and TN plus FN.

Now, this value clearly varies from say negative to positive minus 1 to plus 1 where higher the value towards the positive indicates that highly correlated. If it is a in the negative side then that indicates that it is oppositely correlated. Now, we will start with this MCC in order to win some that are in order to analyze our features.

(Refer Slide Time: 26:46)

MCC based analysis on the binary classifier

Feature	Threshold value	TPR	FPR	TNR	FNR	MCC
1 IA	<700.0 Å ²	0.48	0.08	0.32	0.11	+0.60
2 NSc	>0.35	0.52	0.13	0.27	0.07	+0.57
3 asaV	>0.08	0.52	0.14	0.26	0.08	+0.54
4 NIP	>0.35	0.41	0.07	0.34	0.19	+0.52
5 NSP	>0.4	0.43	0.14	0.26	0.16	+0.37
6 IP _r	>0.4	0.49	0.15	0.25	0.10	+0.46
7 P _r	<0.97	0.35	0.09	0.31	0.25	+0.36
8 NSE	SE _c /3>1.0	0.25	0.09	0.31	0.35	+0.20
9 HPO ₁	HPO ₁ /0.2<1.0	0.37	0.07	0.22	0.33	+0.26
10 HPO ₂	0.8 ≤ HPO ₂ ≤ 1.5	0.42	0.10	0.30	0.18	+0.44

IA 2.0 - IA 1.8
 IA 1.4
 NACCESS

So, MCC-based analysis on the binary classifier. So, features we have, how many features we have, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 so 10 features that we consider. Now, the interface area we see that if we give a threshold. So, please note that this threshold are given manually. This is set as manual threshold by looking at the data or through our app services.

So, if I give 700 then TPR is 0.48 FPR 0.08 TNR 0.32 FNR 0.11 and MCC is 0.60, so which is not that much bad. So, if I use NSc alone and if we give a threshold of 0.35 then it will give me MCC or 0.57. If I use asaV variation, accessible surface area variable so that equation you remember $IP\ 2.0 - IA\ 1.8$ divided by $IA\ 1.4$ where all our interface area calculated using the program NACCESS with varied probe size. 2.0, 1.8, 1.4 are indicating the different probe sizes.

Now when IP if it is 0.35 same as NSc, then its discriminating power is shown here and the MCC is 0.32, but for this NSP for this NSP and what I can see for this two rather it is three. So, PR, NSc, HPOi, and for NSP. So, for these four features actually the MCC values are not that much impressive. So, it is rather low.

So, that indicates that this feature do not have much discriminating power. However, if we combine them and what will be the total effect that we have discussed because we noted the classifiers accuracy as a binary classifier. So, up to this we have discussed about the binary classifier, and after the discussion on a binary classifier definitely you have to go for feature level analysis in order to demonstrate that the individual power of the feature. So, that is it for this lecture. In the next lecture, we will continue to put together everything in order to have a computational framework for identifying the biological functional forms. Thank you very much.