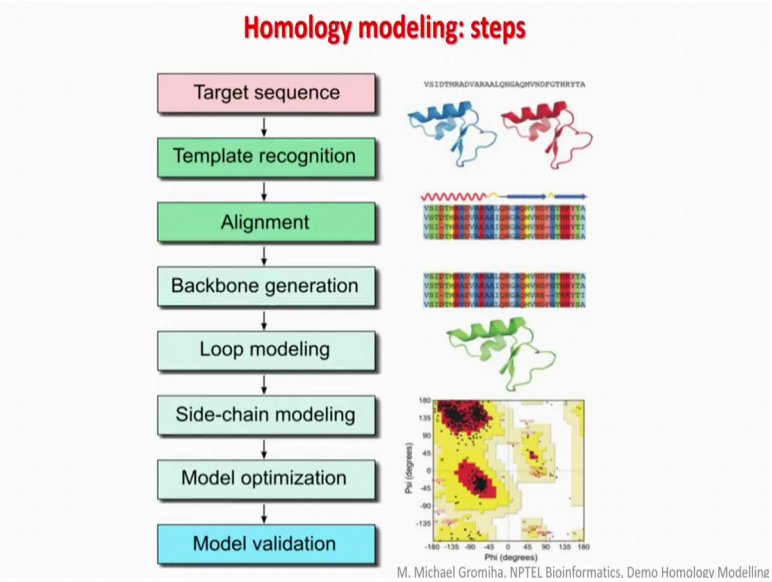
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 44
Demo on Homology Modeling

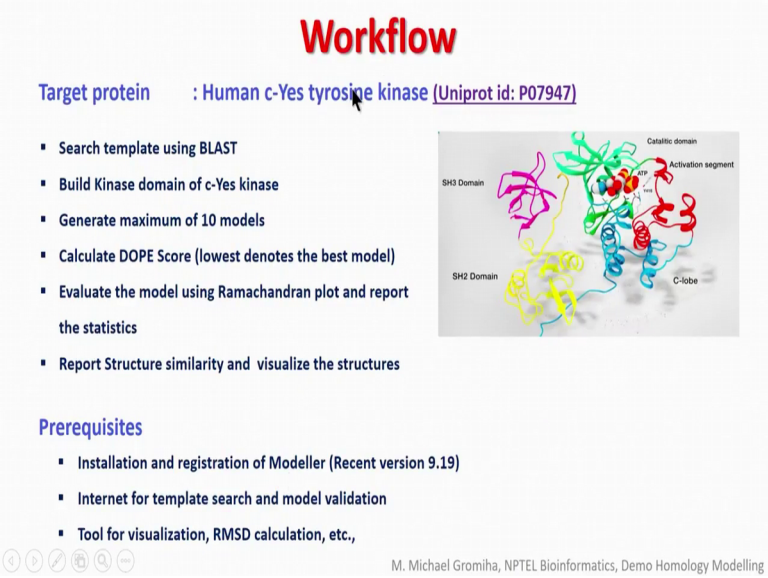
Demo on Homology Modeling. So, here we will demonstrate how to build a model for a protein structure from any specific amino acid sequence.

(Refer Slide Time: 00:38)



As I discussed earlier the homology model requires several steps, such as target sequence. This sequence which one you like to get the structure and one scenario get that completes, then alignment backbone generation, loop modeling, side chain modeling, model optimization and validation.

(Refer Slide Time: 00:54)



For example if you have a protein called human c-Yes kinase tyrosine kinase that the Uniprot id, uniprot is the database for protein sequences P07947. So, in our sequence now the task is to get the structure. So, there are various steps involved to get the structure, first you have to get the template right using the blast, then build kinase domain of the c-Yes kinase and we generate a maximum 10 models and calculate (Refer Time: 01:26) score, which these lowest denotes the best model. Then evaluate the model using Ramachandran plot and report the structure similarity and visualize the structures.

So, to get the structures their prerequisites are first install and register modeller this software used for modeling the structure cc homology modeling and we had internet for the template search as well as validating the model, and we meet the tools suggest primal for visualization and RMSD calculation and so on. Now, we will see how to build a model from the sequence of the human c-Yes tyrosine kinase.

Now, I will give demonstration on prediction of protein structure using modeller software. Using modeller software we can do a homology modeling and here the task is predicting the structure of human c-Yes tyrosine kinase and its uniprot sequence id is P07947.

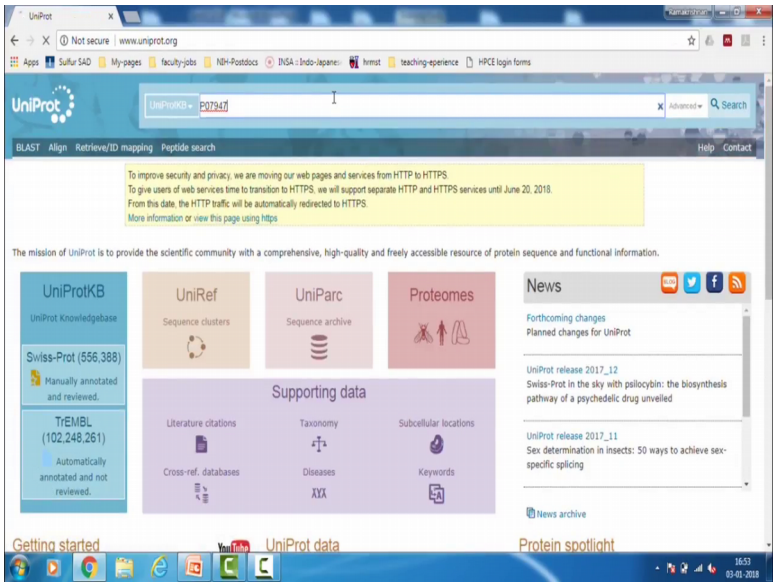
So, it includes few steps like retrieving protein sequence from uniprot, and searching template using blast and building the structure based on the template.

So, these are all the 3 essential steps which I am going to give demo and to dos these we have to understand how the structure structural organization of the protein tyrosine kinase. Here the protein c-Yes kinase which is belongs to the SRC kinase family has 3 domains, one is the kinase domain and then SH 2 and SH 3 domain. But the task given is building the kinase in domain alone. So, accordingly we have to retrieve the sequence and start doing modeling.

So, if the prerequisites are the model software installed and then the internet facility for template search and model validation and the tools visualization tools for visualizing the target and templates and to make a comparative statements on it using matrix likes RMSD.

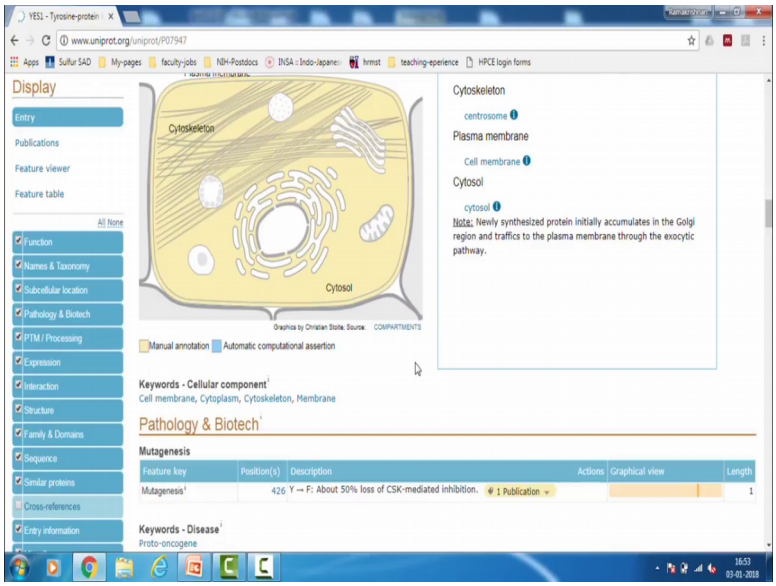
So, first I will explain how to start searching the template for given protein sequence.

(Refer Slide Time: 04:32)



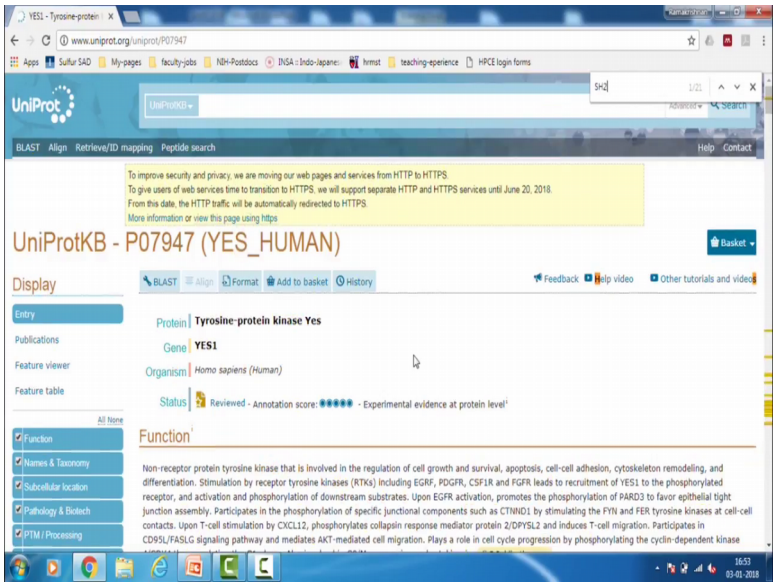
So, to retrieve the sequence go to uniprot and type the id and it will show the full annotation.

(Refer Slide Time: 04:37)



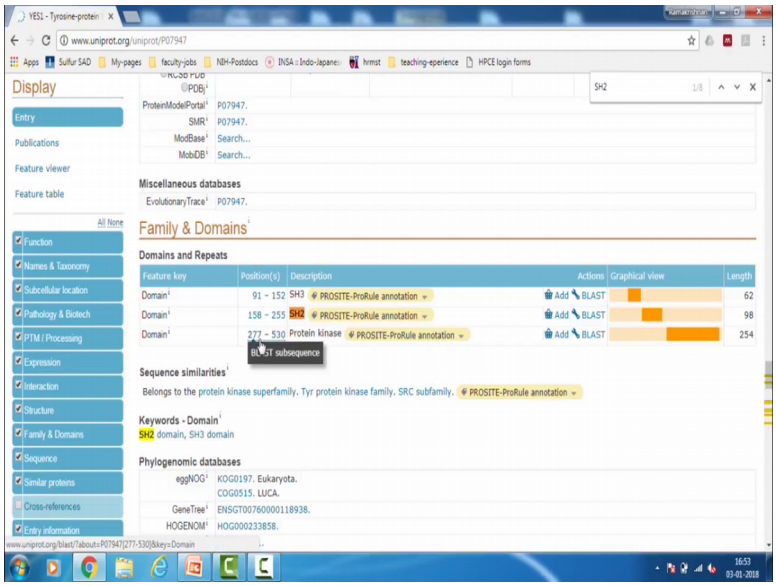
And you search for the sequence belong to the kinase domain.

(Refer Slide Time: 04:57)



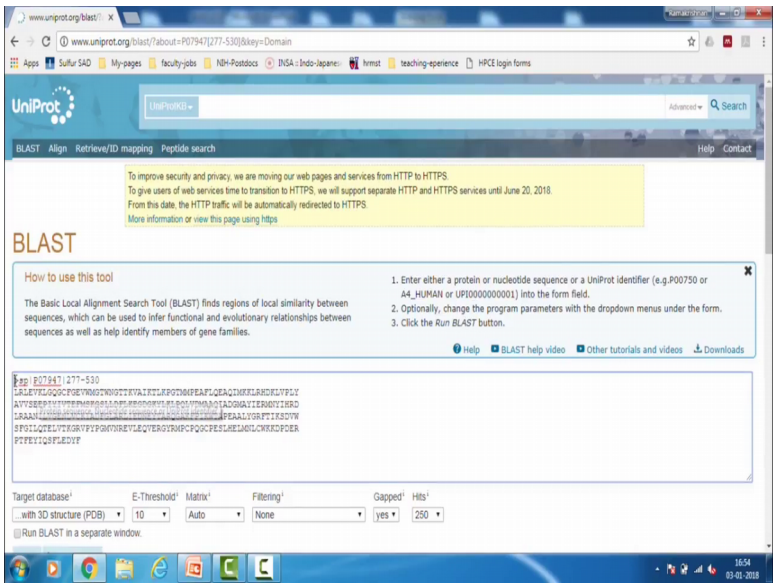
So, for that if you search kinase domain or SH 2 or SH 3 whatever it is.

(Refer Slide Time: 05:08)



So, here you can see the region the sequence belongs to protein kinase that starts from 277 to 530.

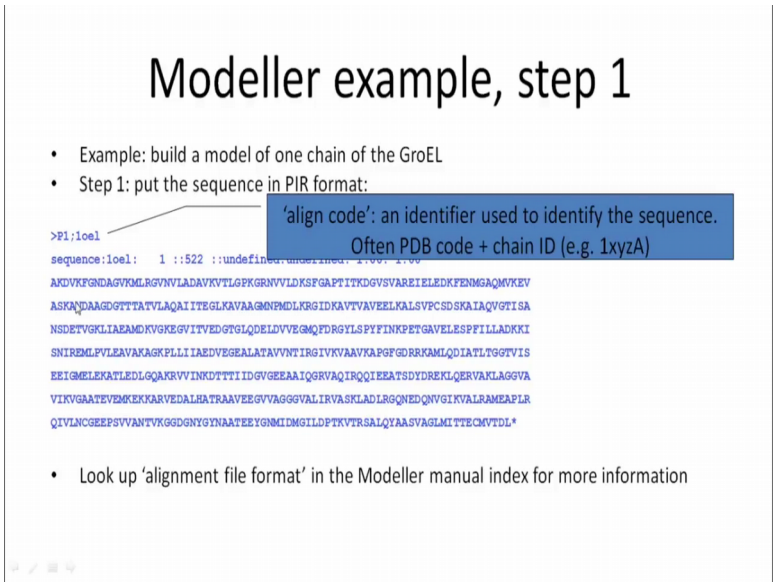
(Refer Slide Time: 05:23)



If you click on this link, you will get the sequence exactly that is belongs to the kinase domain of the protein kinase. You can copy this and save it as the fasta file save as. So, use the same id as a file name dot f a s t a fasta and save it in your preferred working directory, just make homology or modeler.

So, now we have retrieve the kinase domain of the sequence of kinase domain and now as per the need for the modeller software, this fasta file should be converted into the PIR format.

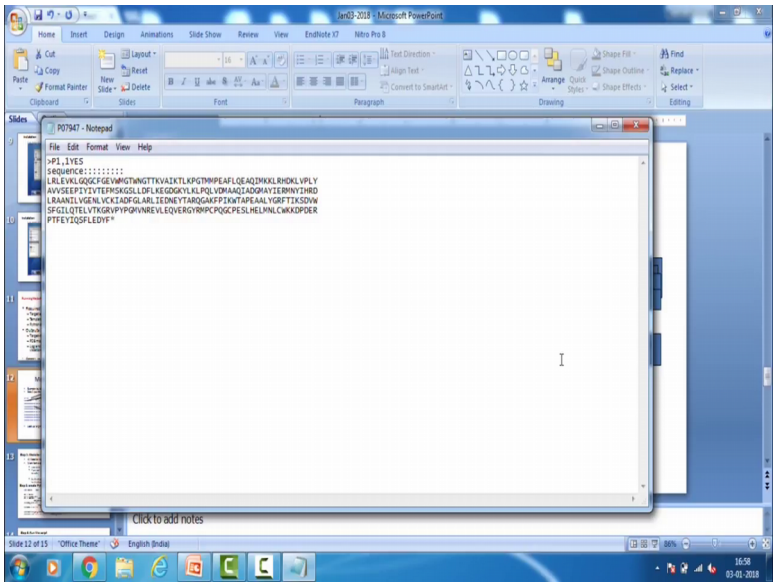
(Refer Slide Time: 06:41)



So, you know. So, this is the typical PIR file format shown in blue color, it starts with the comment line followed by the sequence description and the actual sequence and it ends with the star. Actually the first belongs to the line gives the detail about the align code, which is very essential for modeller to read the sequence.

And the same ID should be used throughout the process and so, accordingly the user can specify any id in 4 letter alpha numeric 4 letter to identify the sequence. And followed by the second line the first term sequence are denotes this is belongs to the target sequence and in case of the template, it will be structure x that is explained here. And followed by this it has 10 fields hills related to the structural detail of the protein, but for sequence it will be null.

(Refer Slide Time: 08:04)

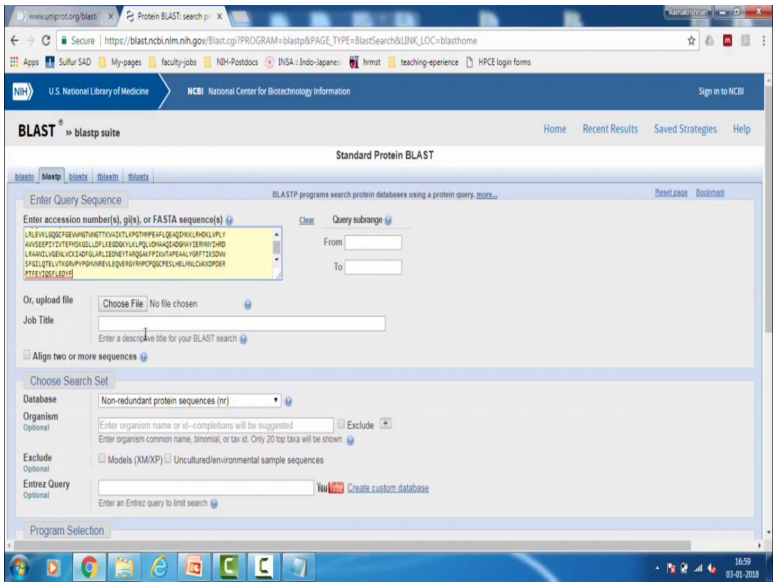


And so, accordingly we have to modify the fasta file to make the PIR. So, you just modify this P 1 comma you can give this CYES as a keyword or 1YES followed by this is the sequence target sequence.

So, it has to be specified with sequence, sequence followed by 10 fields 1 2 3 4 5 6 9 9 10 fields separated by 9 colons. So, we can leave these fields blank because for which the structure is not known and it has to be ended up with the asterisks then you can save it as PIR.

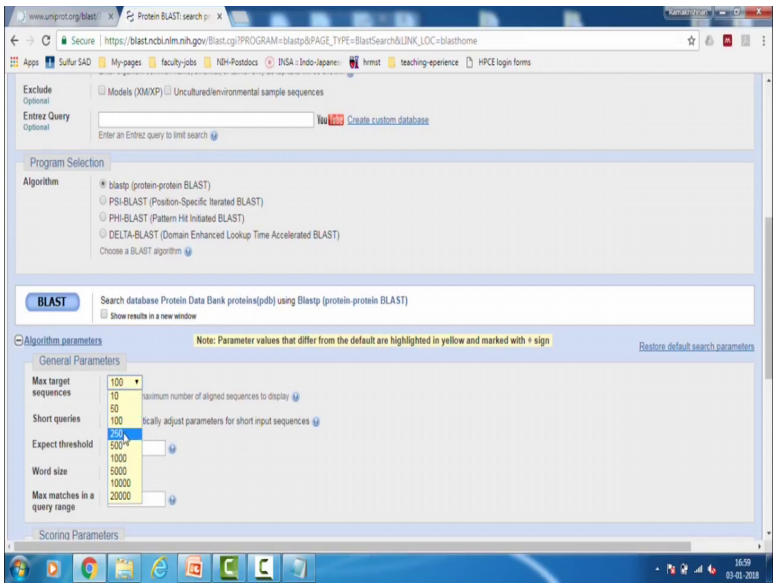
So, this is the first essential step to get ready with the target sequence. So, now, we can start with the typical workflow involved in homology modeling.

(Refer Slide Time: 09:46)

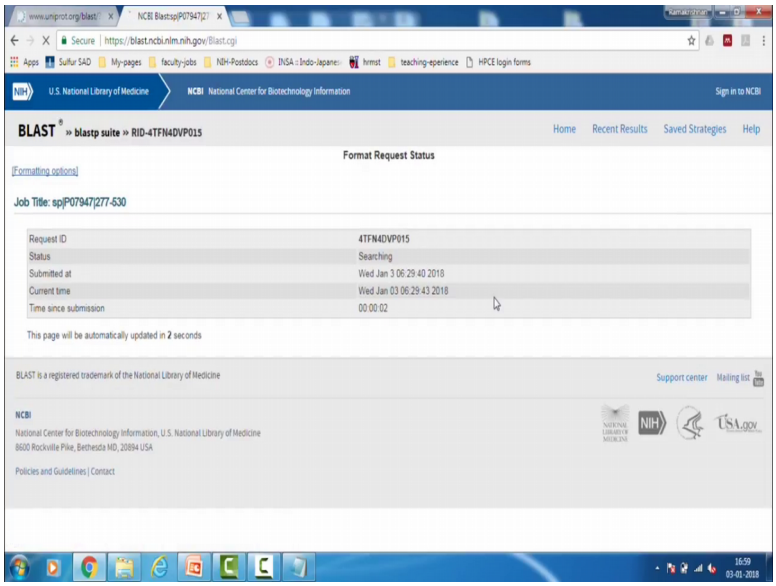


Now, in order to do the first step, we have to search template using the blast do protein blast against the bpdb database. So, that the homology will be searched in the pdb database use the fasta formatted sequence, and here you have to select the database the structure database pdb and do a blast. And you play around this parameters also in order to get some more number of hits, and do blast.

(Refer Slide Time: 10:33)

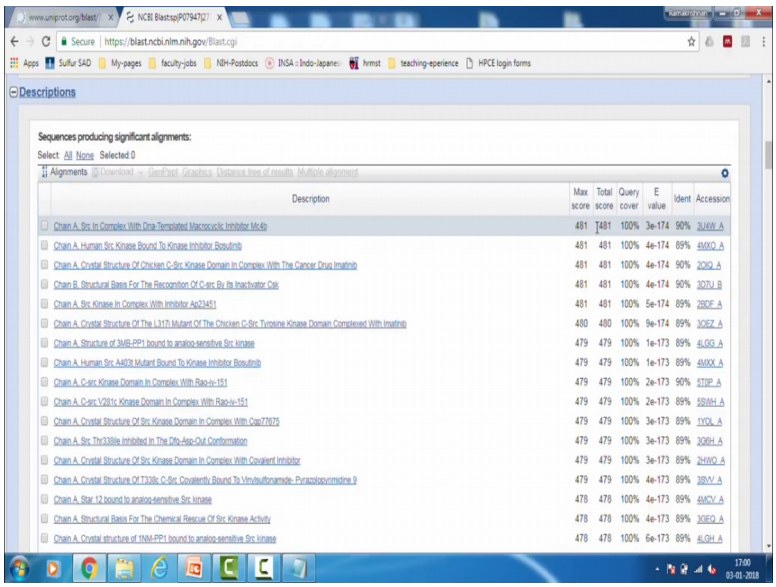


(Refer Slide Time: 10:44)



So, as a result you will get the number of homology the homologues proteins, with protein structure existing the pdb.

(Refer Slide Time: 11:02)



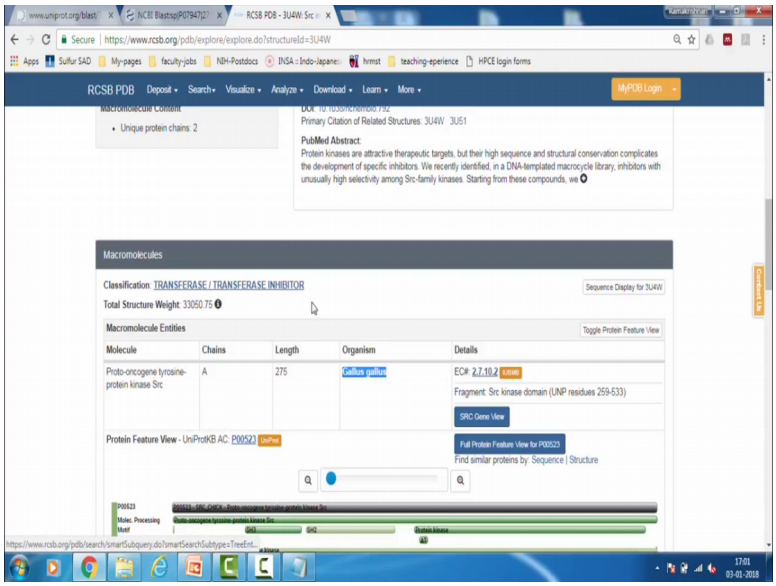
So, in this list you can see the first hit that comes from the Src family even the c-Yes kinases also belongs to the same family and it has the score and query overage good score query coverage of 100 percent and identities about 90 percent.

So, you can note down this pdb id and see how the structure exist. Go to the pdb and give it. The template search is a very important task because the modeller will build based on

template only. If template is not chosen properly the result will also be ended up with error.

So, it is very important to choose the template appropriately. So, even though the score and query coverages are good for the first hit the ranked one based on the A value.

(Refer Slide Time: 12:28)

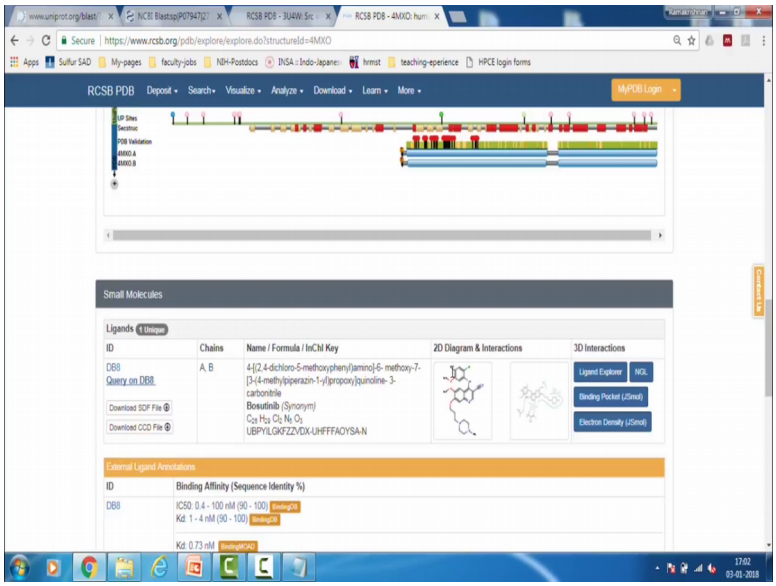


This is belongs to Gallus gallus, but the task given for us is building structure for the human protein kinase human c-Yes kinase.

So, in this case we have to see carefully whether any of the other hits the blast belongs to human are exist. If you see the second hit which is also equivalently good, but only the identity percentage identity is less, but still the 89 percentage is good.

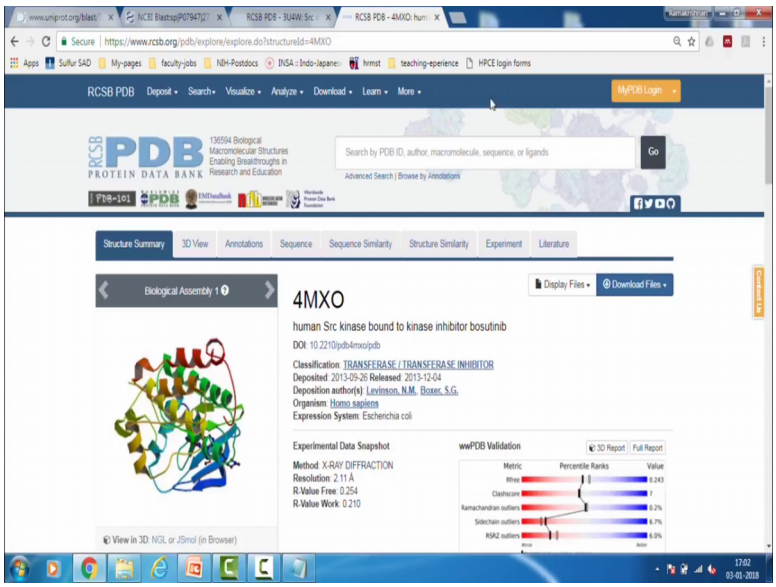
So, if you go to MXO and search go to PDB and see how this structure exists. See here it has two chains A and B belongs to the protein kinase domain and it has the ligand bounder conformation and this is also equivalently good.

(Refer Slide Time: 13:26)



So, in this case which one we will take whether the template has the good identity the ranked first as per the blast result or the second one which is belongs to the human.

(Refer Slide Time: 13:34)

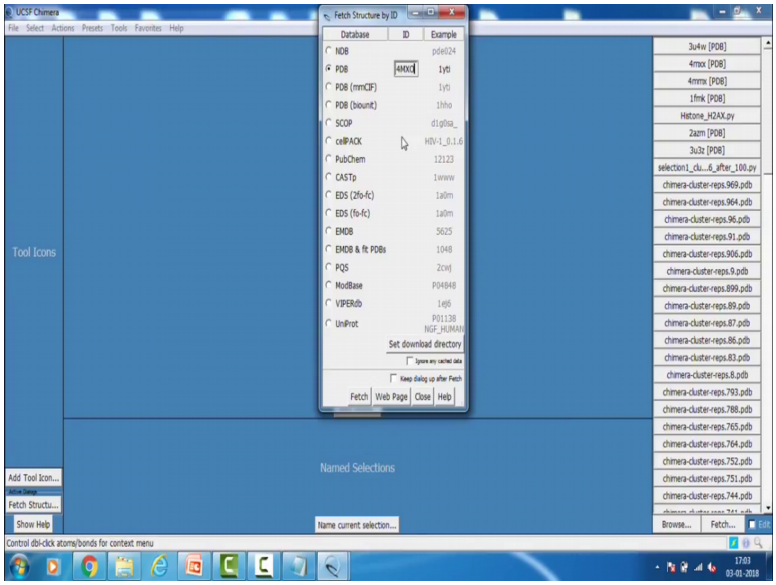


So, both are equivalently good not for model building, but still if you visualize using some visualizer tool for example, chimera if you take.

The visual inspection is very very important and in addition to the blast search results. Because this sequence this is blast search is based on the sequence alignment, while even the PDB the chances for having missing residual PDB structure is high. So, in that case if

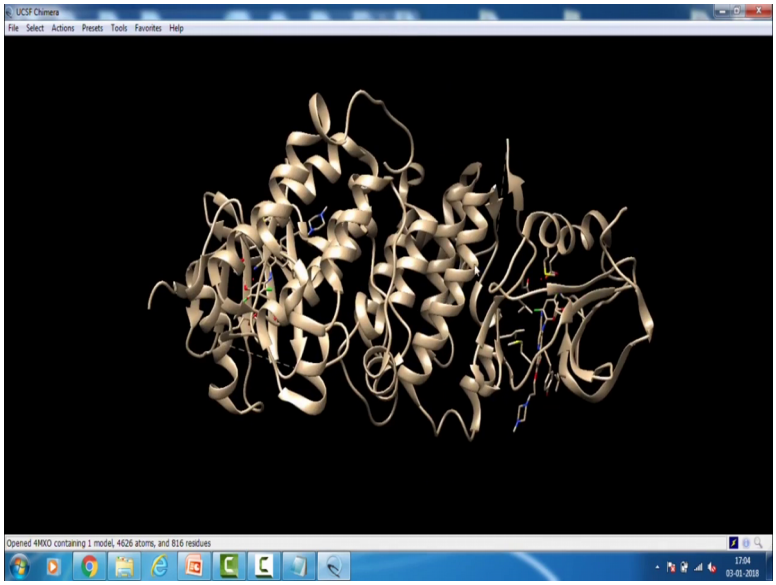
you visualize the structure, you can see whether the structure structural information for the given template is complete or not.

(Refer Slide Time: 14:47)



So, here for a for example, if you see MXO that is belongs to the human, you can retrieve you can retrieve the structure 4 MXO 4 MXO the fetch.

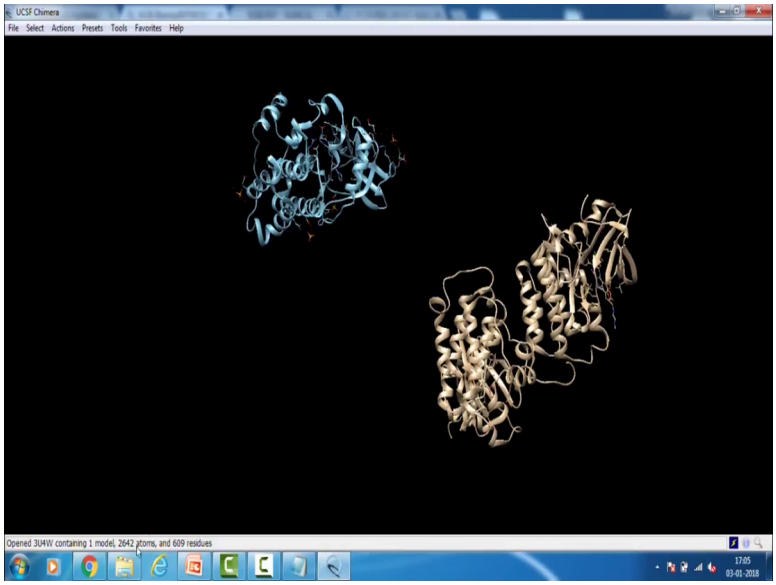
(Refer Slide Time: 15:12)



This is the simplest way to see whether the structure is having any missing residues. See in the chimera will easily which will tell you through the dotted lines see here you can see the dotted lines that is that is belongs to the missing segments.

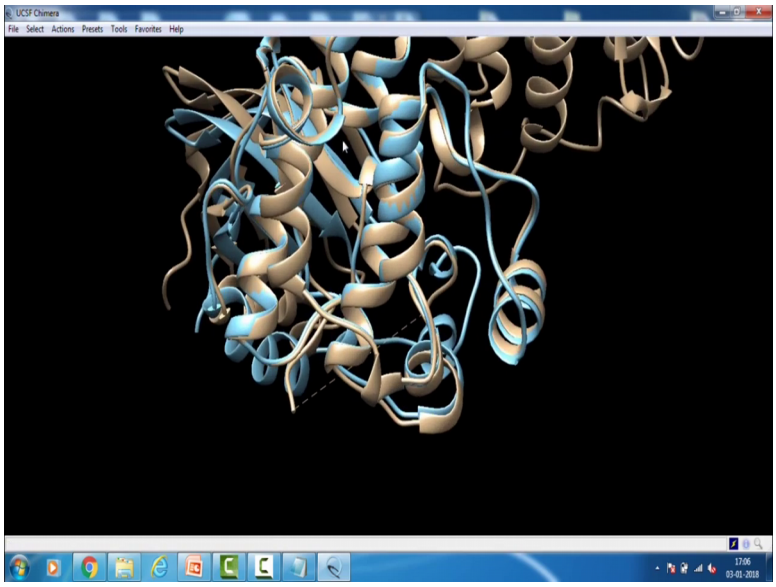
So, this 4 MXO is not the complete one. So, we do not choose it. If you go if you check for 3 u 4 w even though that is belongs to the Gallus gallus, fetch.

(Refer Slide Time: 16:08)



See you can compare now, see here for in Gallus gallus you do not see any missing residues, there is no dotted lines. Here you can now compare and see both the structures how identity both are using the chimera tool matchmaker. This is the reference structure and one another one can be the structure to match you just click ok.

(Refer Slide Time: 16:46)



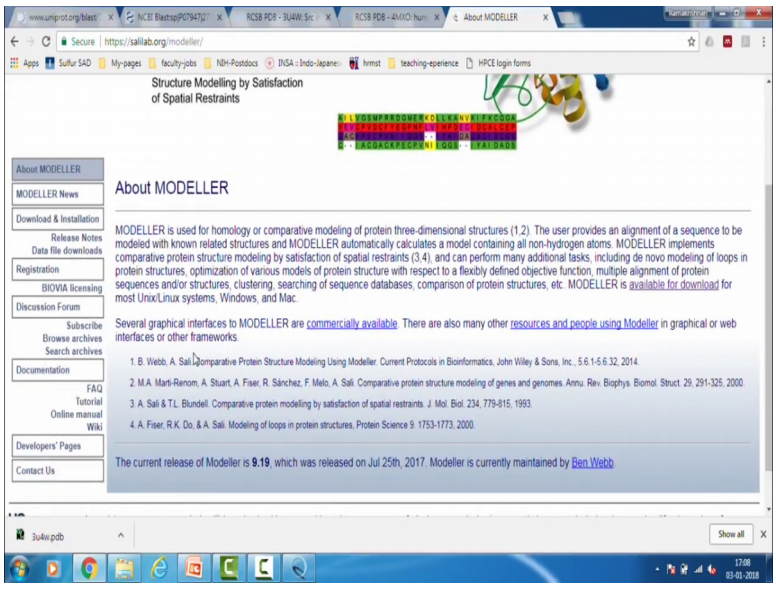
See here the both the structures are identical and you can see here the segment, which is missing in human protein is present in the chicken protein.

So, in this case we do not mind the one percent difference in the identity, here the structural information is very important for model building. So, we will take the c-Yes kinase of Gallus gallus instead of human for this task, and start doing the homology modeling.

So, now we have collected retrieved sequence from uniprot and we have retrieved the appropriate template from PDB. Now target and template both are ready to do the homology modeling.

So, if you go to your working directory, you will be having three files first one is corresponding to the structure in PDB file format and the two files fasta and PIR for formats belongs to the sequence. Now, you can start working with the modular software, which we have already installed in our computer using the procedure given in the modular website.

(Refer Slide Time: 18:49)

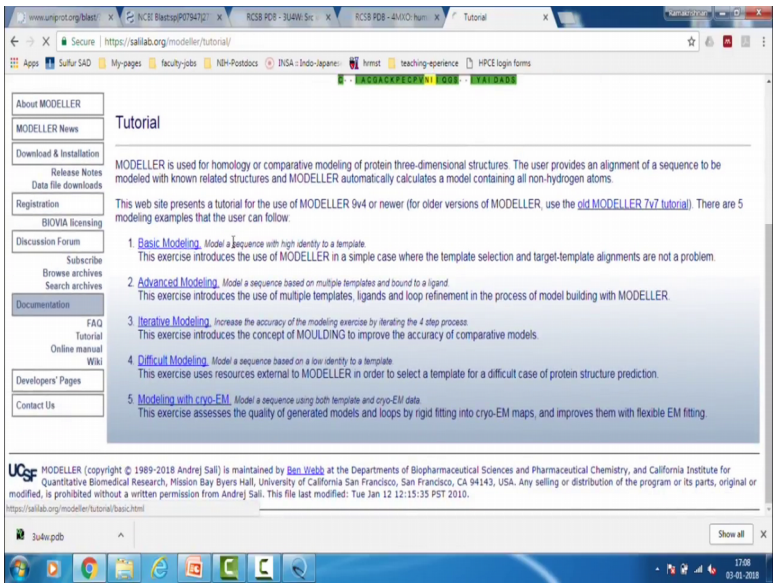


If you go to the modular website you can see the documentation for downloading and installing the software, and to do registration also. Here in the documentation download and installation instruction is given here you can download the software which is so, compatible for your computer and install it

(Refer Slide Time: 19:06)

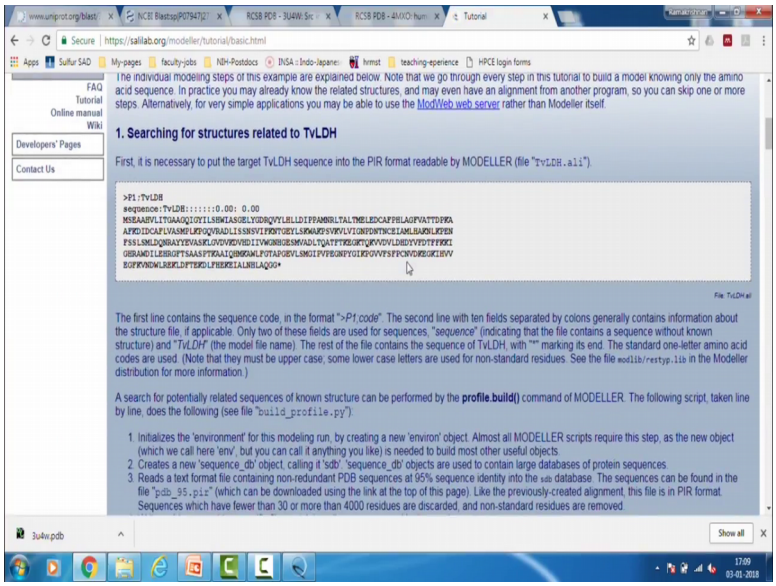
And the current version of the modular is 9.19 and this is command mode. And in the documentation you can see the help manual to perform the modeling.

(Refer Slide Time: 19:39)



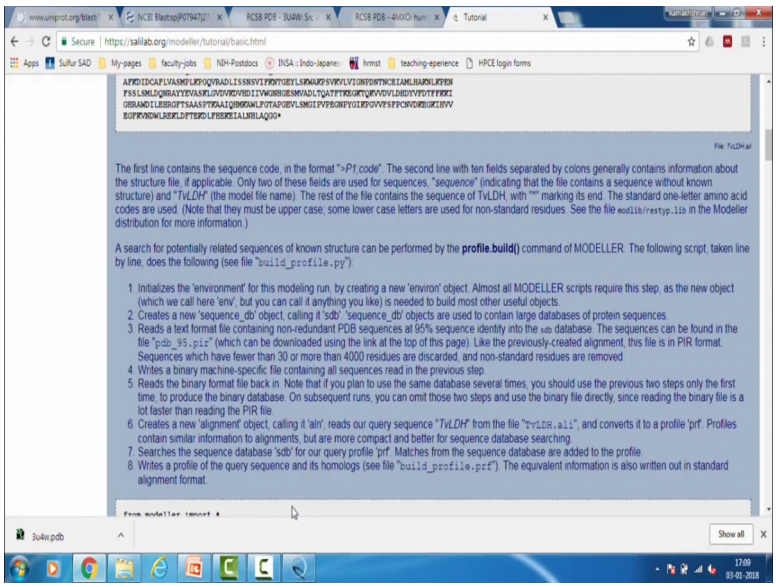
You can follow this simple instruction to perform the homology modeling as we are going to discuss now.

(Refer Slide Time: 19:43)



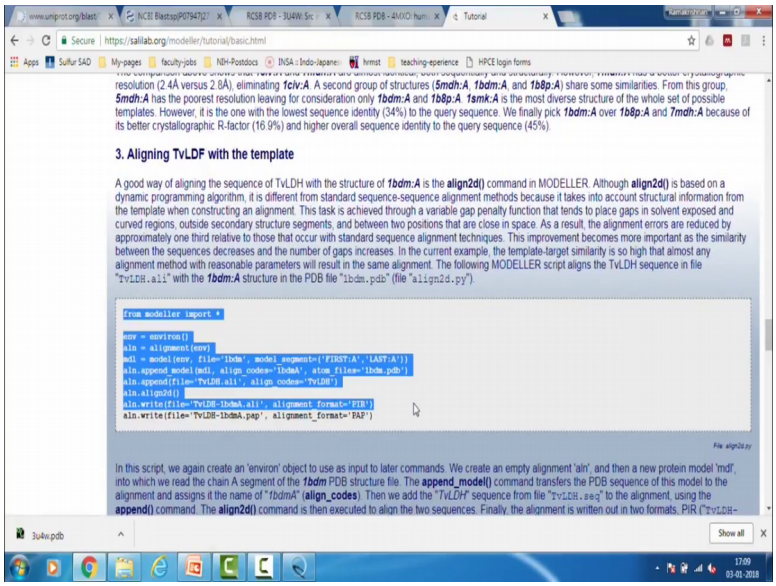
So, in model building there are two essential steps, which is which is very important to discuss here is first the alignment of protein sequence with the template structure.

(Refer Slide Time: 20:07)



We will call it as target template alignment, to do this we can get the script modular script from their website here.

(Refer Slide Time: 20:26)

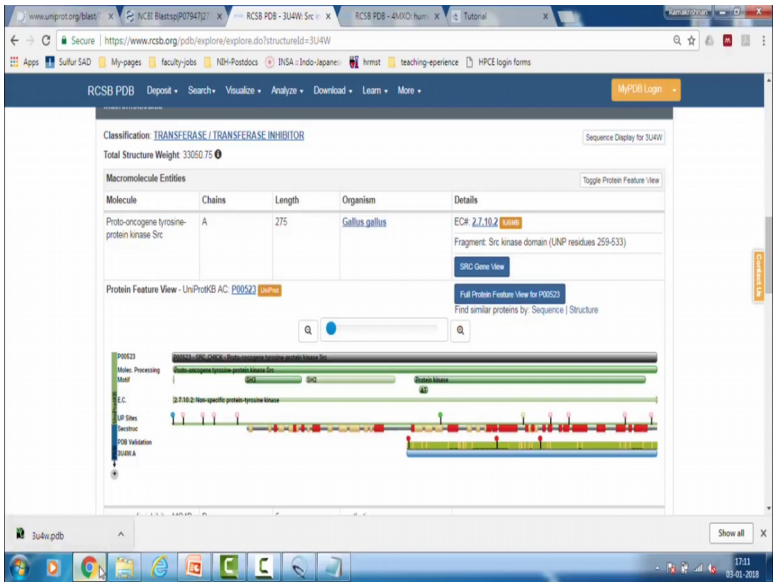


This is the script we have to use to perform the target template alignment; so, that the sequence of our target sequence will be aligned to those sequence of the template structure.

So, you can save it save this script as the model dot py because this is the python step the first step is more alignment dot py. So, if you look at this script, the first 3 4 lines corresponds to setting up the environment variable and improve importing the necessary modules from modular and action to perform the alignment.

So, here in the next line if you see the value to the variable file, that is belongs to the template. So, here what template we have chosen is 3 u 4 w that has to be given here carefully 3 u 4 w and this has only chain a if you go to the pdb side there you can see you can confirm how many chains the protein this protein has.

(Refer Slide Time: 22:04)



It has only one chain A so, you have to specify the starting and ending chain names so, here only one. So, in both first and last will have the same value A and in the in the alignment. This the next line corresponds to the sequence where you have to specify the align code for the sequence of the template structure. So, here we will give the same id 3 u 4 w followed by A. In order to identify the sequence that is belongs to the chain A of the protein tyrosine kinase structure, and followed by the atom file name is 3 u 4 w dot pdb.

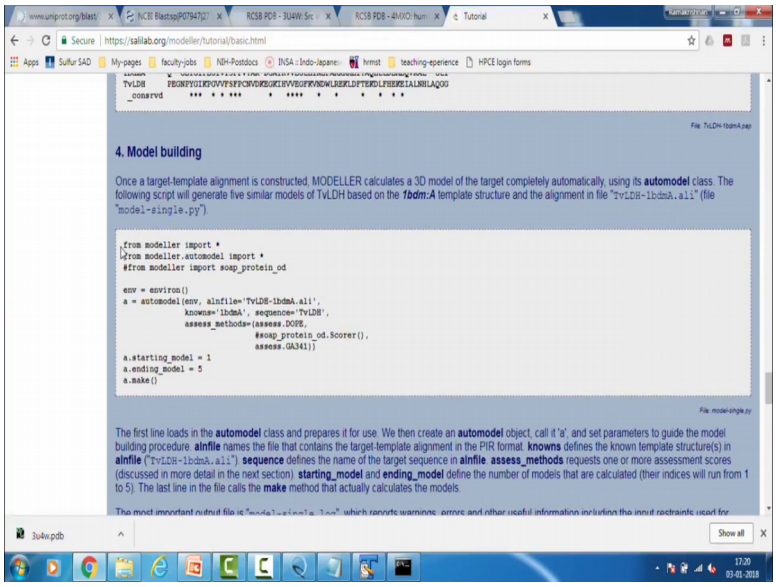
So, that these two lines until modular to consider the chain A of the template structure and for the alignment, and in the next line we will give the target sequence information that is P07497 dot PIR and the align code corresponding to the target protein is 1YES that is what we have mentioned here the same thing should be given here 1YES, and here also you specify the same.

So, that modeller will read these two codes and identify which is the template and which is the target. So, the output alignments file will be like target template target template dot ali is the output file in PIR format we can save this.

So, now, the script for target template alignment is ready and we can perform it using modeller. So, start modeller command mode and change the working directory using the cd command in windows. Now here you can use the command mod 9.18 this is the command here for this version followed by one dot alignment the python script file.

So, it will start doing the alignment.

(Refer Slide Time: 25:52)

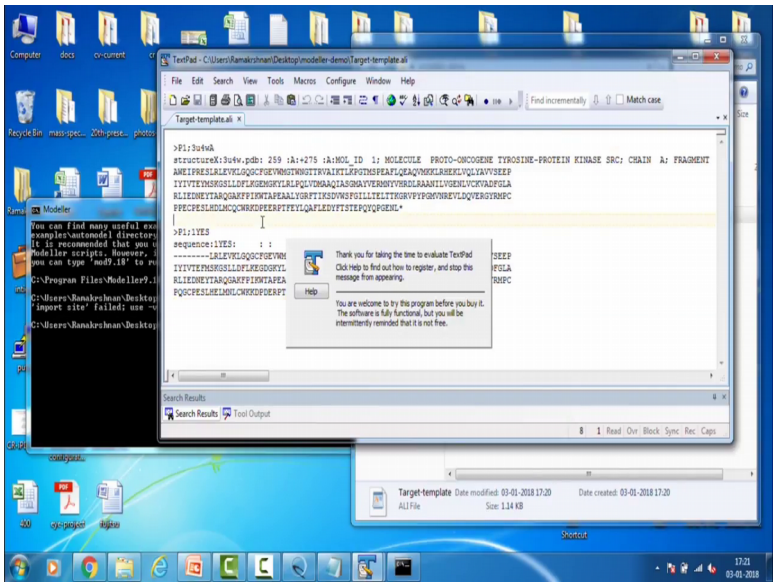


So, once the alignment is completed we can start doing with the model building. In the same help page you can find the second script that is belongs to model building and you can copy the script and save it similarly.

So, in this script if you see similarly the first 4 line corresponding to the modular libraries and the next line belongs to the specification of the target template alignment file, and then the corresponding the align codes. So, from the result file of step 1 we can note down the align codes and then specify here to perform model building.

So, the target template alignment is the output file we go it now.

(Refer Slide Time: 26:42)



And if you open and see here in the alignment file you can see the alignment in the PIR format, and the align codes that is corresponding to the structure and the alignment align codes that is corresponding to the sequence and both the codes you can note down or even for say if you can copy and paste it to the new script the next script.

Here the knowns is corresponding to the structure id. So, here what we used is 3 u 4 w a and for sequence that is target that is 1YES. And the next line tells the what method is going to be used for model building and scoring. Here the dope scoring scheme is used to evaluate the models and here this line corresponds to the number of models to be generated.

So, this is the file called the second step model building, dot py yeah and we have to specify the correct alignment file also, that is target template dot ali. This is the file actually we have we have got from the previous run target template dot ali that is specified now and now you can perform model building, using the same command two underscore.

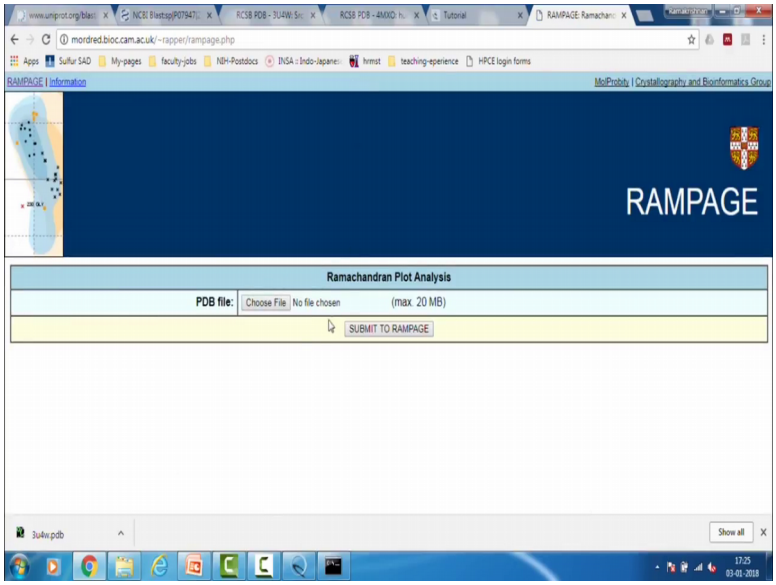
So, these are the two important scripts to be edited carefully in order to perform target template alignment and model building. Once this is done and you will be getting the 10 kinase structures for the given target sequence and you can evaluate the correctness of the model in the next step. So, we will wait for a while until this generates the 10 structures.

So, now, you can see the progress, it generated the two structures so far, that has the file name starts with the align code of the target sequence followed by some alphanumeric numbers. So, once all these 10 structures are built, we can start evaluating the models based on dope score, that will be written in the log file of the present running presently running script.

So, so far we have completed the model building, in the workflow we have retrieved the target sequence and retrieved the template structure and perform the alignment, and perform the model building also that includes backbone generation, loop modeling everything will be taken care by model build step that is called auto model next we have to evaluate the model.

So, to perform this we need a software to generate the Ramachandran plot, we can use some online tools called rampage for example.

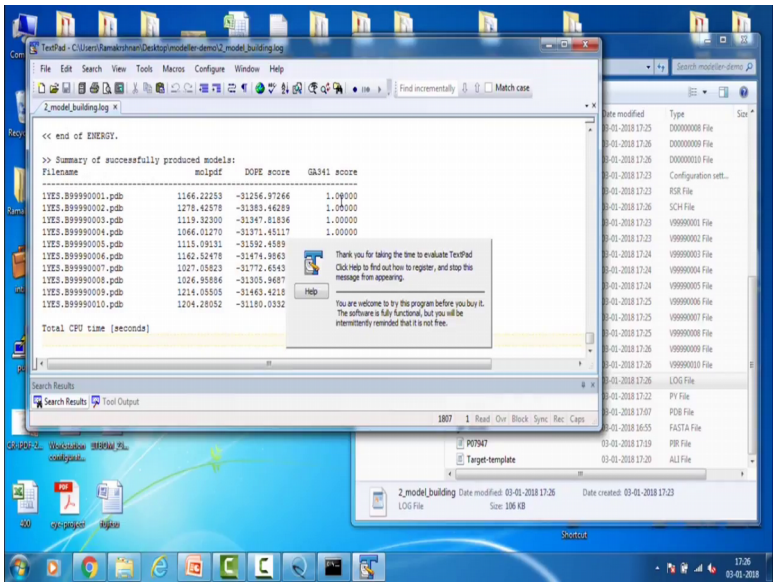
(Refer Slide Time: 31:36)



So, here we have to upload the selected model based on dope score and see how many residues are in allowed and disallowed regions in order to finalize the model for further work.

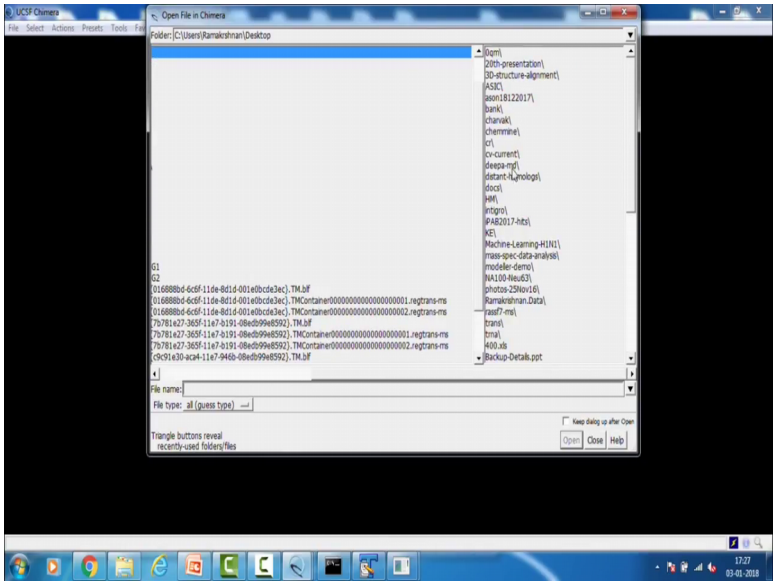
So, the model building is almost done now we will see the log file of the model building and choose the model based on the dope score.

(Refer Slide Time: 32:13)



That is written at the end of this log file, here you if you see for the given for the 10 structures generated the least dope score is the best one, that is somehow 3 5 7 yeah the 7th model is the best as per the dope score. If you take the 7th structure now to visualize using primal or chimera whatever you have, you can see the model built using primal here and you can compare it with the template already we have the template loaded in the chimera.

(Refer Slide Time: 33:32)

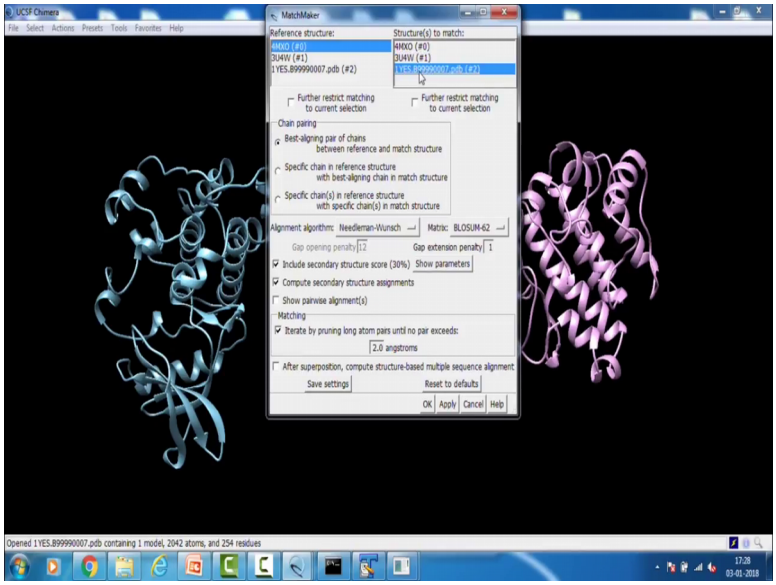


You just visualize our template alone and load the selected build structure based on the dope score.

(Refer Slide Time: 33:45)

The seventh one is the best one open it and align using a chimera the reference structure is your template and the structure to match must be your the build structure.

(Refer Slide Time: 33:52)



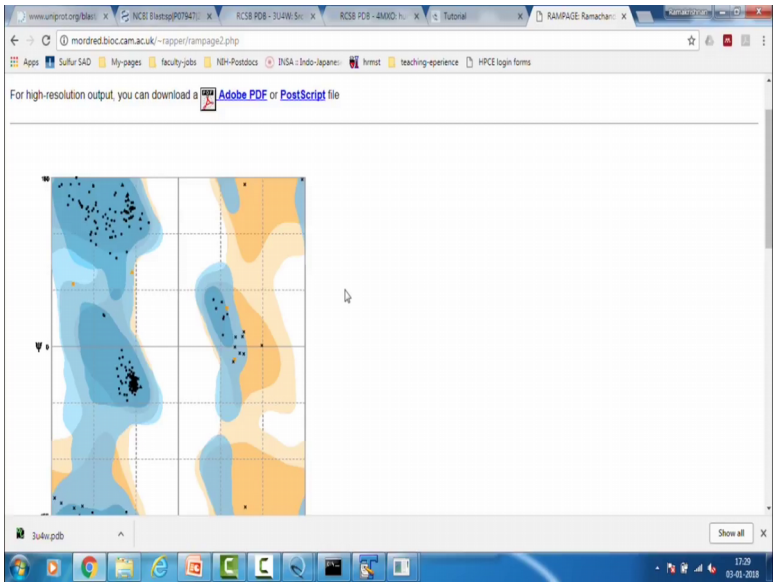
Now, you can see the alignment how perfectly the order built, the structure of your target based on the template with the RMSD of 0.1, which is very low and means the both the structures are not deviating much.

(Refer Slide Time: 34:18)

So, once this is confirmed the structure the model building is confirmed, you can evaluate the same using the Ramachandran plot also.

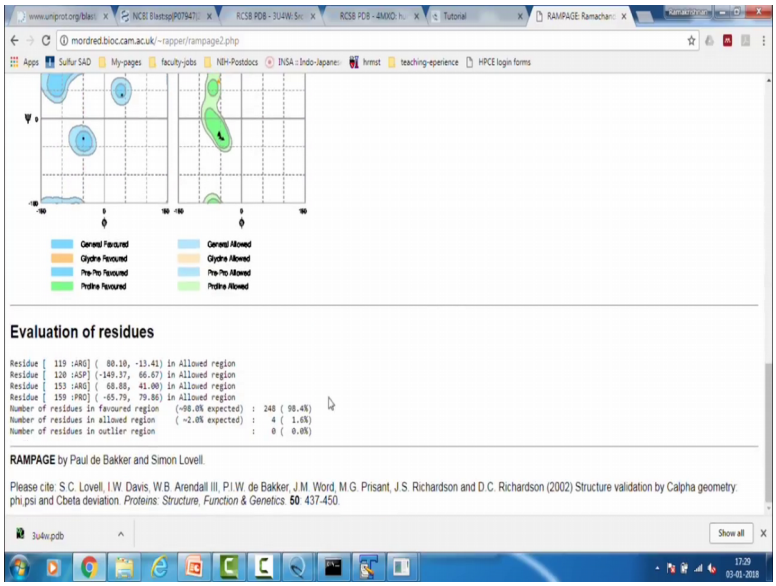
The modeller demo you can load the 7th model and submit.

(Refer Slide Time: 34:56)



Here you can see the Ramachandran plot with the phi and psi angles specified for each and every amino acids and the statistics.

(Refer Slide Time: 35:07)



So, here the model was built with 98 percent 0.4 percent of the residues are in favored region and 4 percent are in 4 residues are in; that means, 1.6 percent are in allowed region and there is no even single residue in the disallowed region means, the model built the 10 confirmations and we have selected the correct one and evaluate and evaluated the correctness using the Ramachandran plot.

So, it means the model what we built is very perfect and we can take it for further structure based web designing.