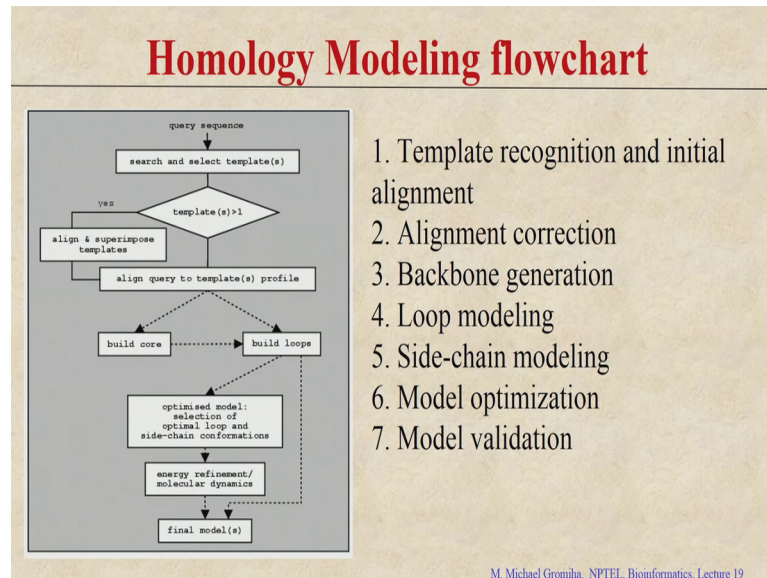


**BioInformatics: Algorithms and Applications**  
**Prof. M. Michael Gromiha**  
**Department of Biotechnology**  
**Indian Institute of Technology, Madras**

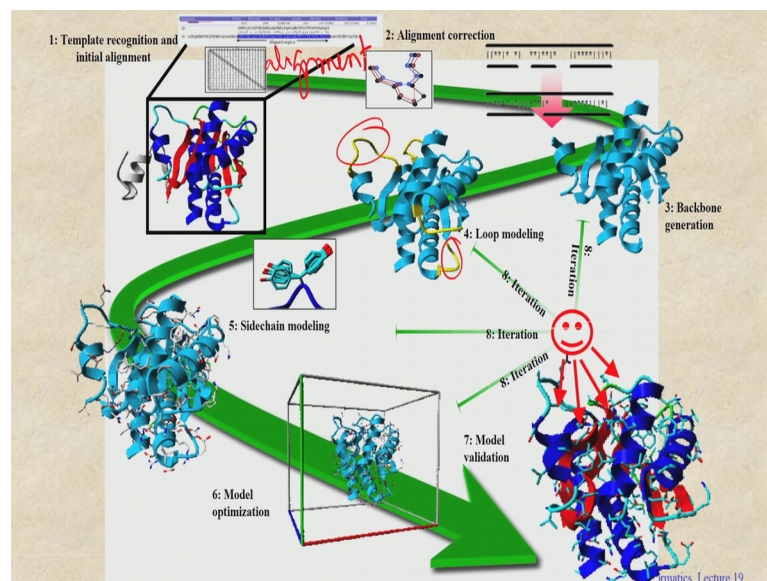
**Lecture - 19b**  
**Protein Structure Prediction II**

(Refer Slide Time: 00:16)



I will explain the steps now one by one.

(Refer Slide Time: 00:18)



Here this is the overall summary of the homology modelling technique; first one, if we see here this is the structure we need. So, we have the 2 sequences. So, identify the template right you check the alignment and get the template once we get the templates, then go with the alignment corrections right here with the; this is the alignment right, then check the alignment show that there will be less gaps in the alignment once we have the alignment correction, then we generate the backbones right because you take the same backbone and get the backbone and then when the backbone is set, they look for the loops here. So, you can see the several loops here and model these loops right when the loops are done, then go with the side chains, there are several libraries you can check the side chains and when we model the side chain.

Then finally, optimize the model right you can also try to use molecular dynamic simulations to see whether there is any fluctuations in these amino acids residues and finally, we validate the model, if you are happy with the model and it satisfies the requirements then it is fine, if not you have to iterate again and again with the different steps we change the loops and you change the template or change the side chains right and finally, see whether if the model is valid or not. First we need to take the template; how to get the templates?

You check with the blast your sequence will give and then blast and finally, you will get lot of hits with the different identities. So, your accuracy of the model depends on the template and you can see it is very important to identify the best template if you get several hits with high sequence identity what the other criteria you use.

Student: (Refer Time: 02:13).

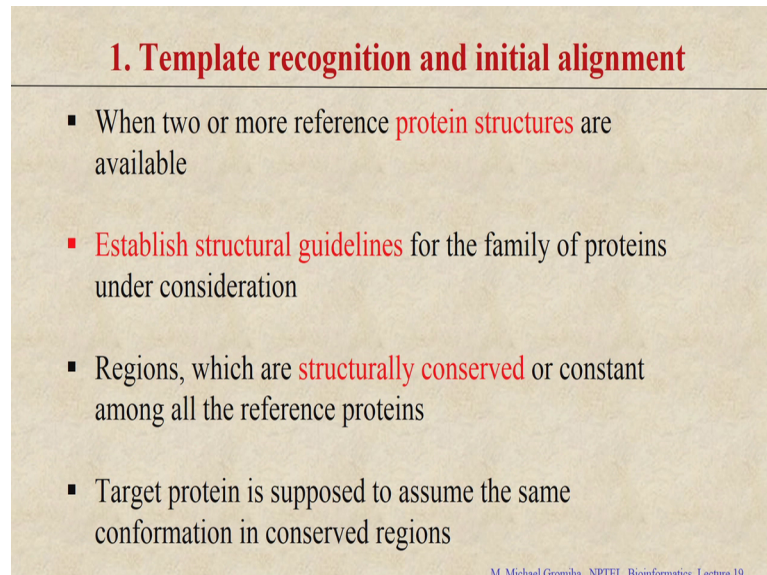
You can see sequence coverage and you can see the less number of gaps and see whether the proteins belong to similar families, right and all the aspects you need to consider right and second is see if these proteins have some similar folds and the regions are properly aligned.

So, all these things you need to consider before you select the template, then the second one; you look into these different proteins and see the important residues, whether the residues are same in these 2 different proteins then based on all these aspects then you can choose as template because it is very important to choose the template otherwise you end up with a wrong structure.



So, now if you see the structures are known because you search your query with the known structures.

(Refer Slide Time: 03:05)



A presentation slide with a light beige background and a dark red title. The title is '1. Template recognition and initial alignment'. Below the title is a bulleted list of four points. The first point is 'When two or more reference protein structures are available'. The second point is 'Establish structural guidelines for the family of proteins under consideration'. The third point is 'Regions, which are structurally conserved or constant among all the reference proteins'. The fourth point is 'Target protein is supposed to assume the same conformation in conserved regions'. At the bottom right, there is a small blue text credit: 'M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19'.

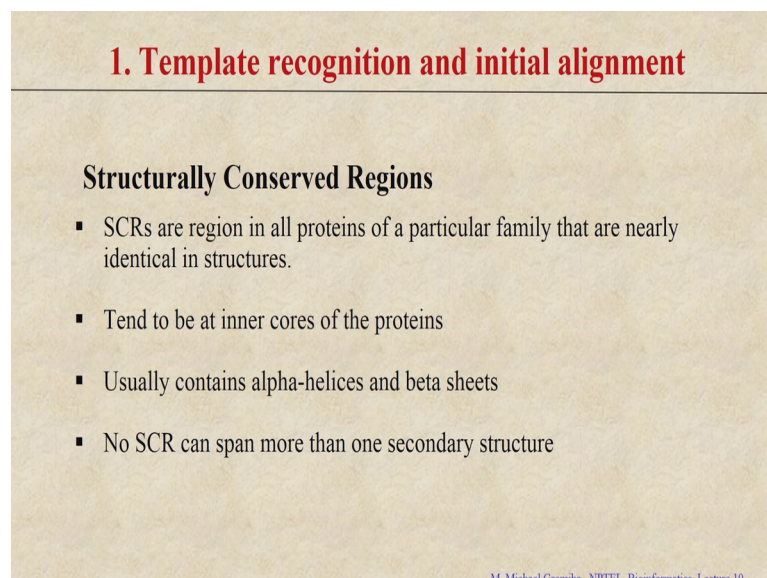
### 1. Template recognition and initial alignment

- When two or more reference **protein structures** are available
- **Establish structural guidelines** for the family of proteins under consideration
- Regions, which are **structurally conserved** or constant among all the reference proteins
- Target protein is supposed to assume the same conformation in conserved regions

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, then we get the structural guidelines with the templates, we will get different structural information because the PDB is known and see whether you can see the family of the proteins.

(Refer Slide Time: 03:25)



A presentation slide with a light beige background and a dark red title. The title is '1. Template recognition and initial alignment'. Below the title is a section header 'Structurally Conserved Regions' followed by a bulleted list of four points. The first point is 'SCRs are region in all proteins of a particular family that are nearly identical in structures.' The second point is 'Tend to be at inner cores of the proteins'. The third point is 'Usually contains alpha-helices and beta sheets'. The fourth point is 'No SCR can span more than one secondary structure'. At the bottom right, there is a small blue text credit: 'M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19'.

### 1. Template recognition and initial alignment

#### Structurally Conserved Regions

- SCRs are region in all proteins of a particular family that are nearly identical in structures.
- Tend to be at inner cores of the proteins
- Usually contains alpha-helices and beta sheets
- No SCR can span more than one secondary structure

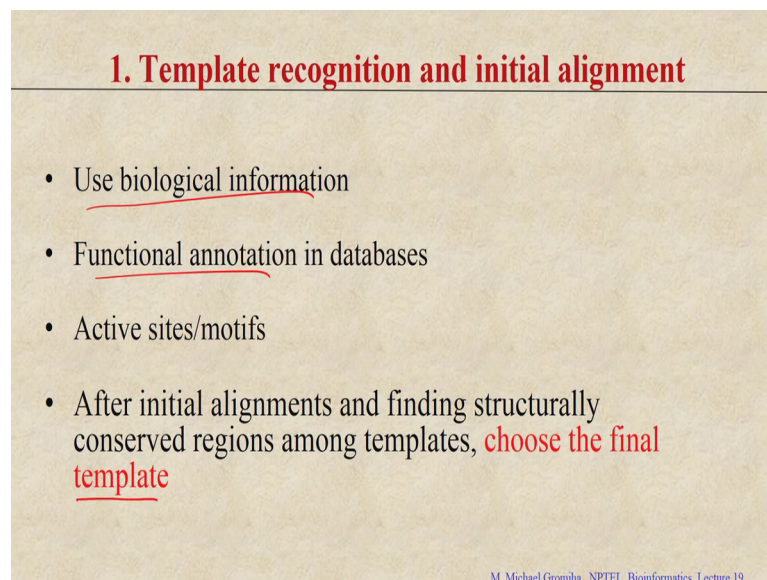
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

How about the structurally conserved regions and then if you see all these information then you can choose the template; how to check the conserved regions because these regions; what are conserved regions.

Student: Conserved (Refer Time: 03:33).

And the protein; they have the nearly identical structures right because the conserved regions they have the identical structures and they are might be in the interior core and mostly they contains alpha helixes and beta strands. So, look into these details right and also the location of this residues, right, in the conserved regions because we know the structures right you need to take into consideration all these aspects.

(Refer Slide Time: 03:56)



**1. Template recognition and initial alignment**

- Use biological information
- Functional annotation in databases
- Active sites/motifs
- After initial alignments and finding structurally conserved regions among templates, **choose the final template**

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

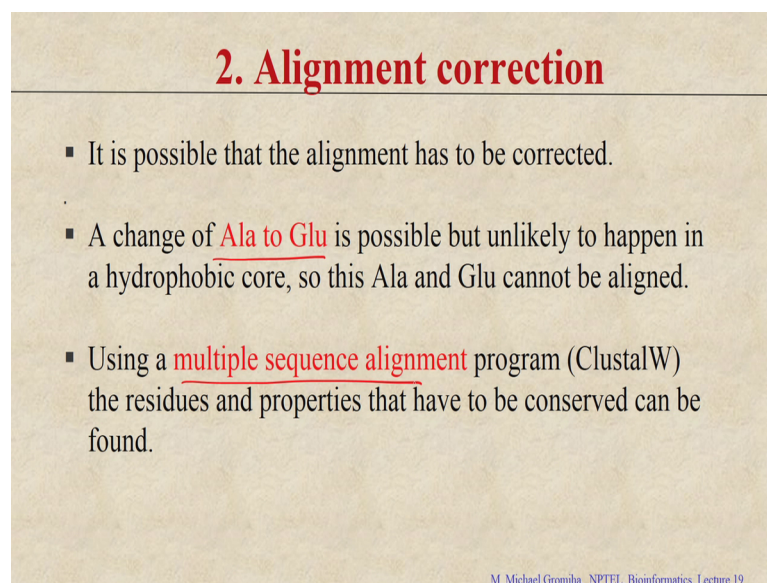
Then we use the biological information and the known functional annotation because the several databases you can see the residues different functions right as active sites or motif and so on, and using all these information.

Finally you can choose the templates. Usually what we do you take the sequence and just paste the sequence in Swiss model or any software you will get the structure. So, the question is whether you get the reliable structure or not. So, if you spend some time right in each steps then you can be confident that you get the good structures the most important one is the template you need to give proper template for the modelling.

So, in that case you can get at least more than 60 percent confident that you can get the reliable structure. So, these are the methods we use for the template you use the blast and then you use the known structure database right and finally, you get the structures from the PDB right this is fine, ok.

Now, what to do you get the templates right you have your query sequence and you get the template you check the literature and other information finally, you have decided that this your template then if you see the alignment is correct or not.

(Refer Slide Time: 05:08)



## 2. Alignment correction

- It is possible that the alignment has to be corrected.
- A change of Ala to Glu is possible but unlikely to happen in a hydrophobic core, so this Ala and Glu cannot be aligned.
- Using a multiple sequence alignment program (ClustalW) the residues and properties that have to be conserved can be found.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

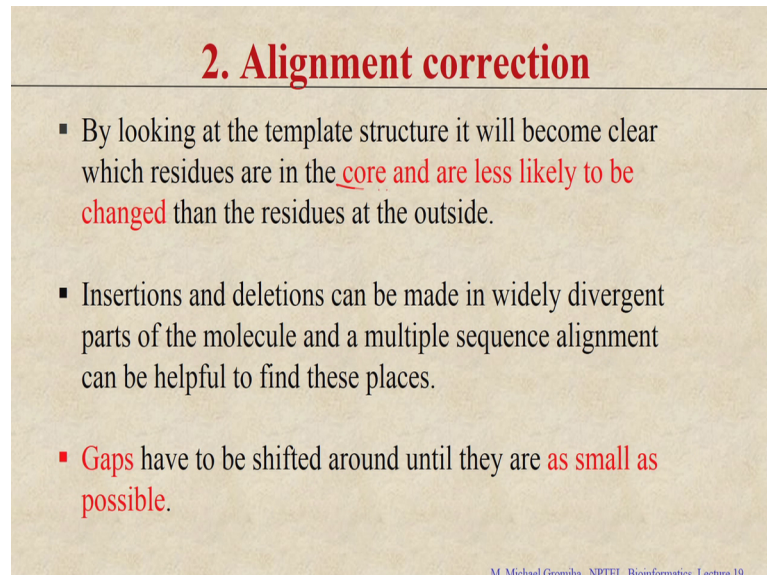
For example, if you see an alignment, right, alanine is mutated glutamic acid. So, you can align any residues you can mutate any residues that is possible, but this is reliable or not right if it is alanine is the buried region and we convert it; the mutate to glutamic acid this is what unlikely happen because it will destabilize the protein right because alanine is a hydrophobic residue; this is a hydrophobic core and if you convert; mutate to alanine to glutamic acid what will happen?

Student: (Refer Time: 05:36).

Destabilization, right because it is a charge residue right it will destabilize the protein. So, this is less possible to have this type of alignment in this case the alignment you need to change. So, introduce a gap or right and you can make these residues not align you have to align properly then you can use the multiple sequence alignment and see the

conserved residues and your template whether the conserved residues are properly aligned or not if it is not properly aligned you have to make the corrections.

(Refer Slide Time: 06:12)



**2. Alignment correction**

- By looking at the template structure it will become clear which residues are in the core and are less likely to be changed than the residues at the outside.
- Insertions and deletions can be made in widely divergent parts of the molecule and a multiple sequence alignment can be helpful to find these places.
- Gaps have to be shifted around until they are as small as possible.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

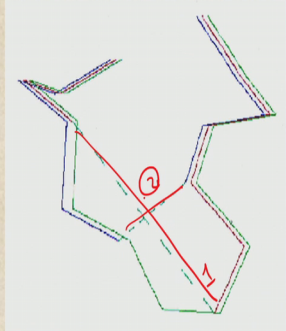
So, that it is properly, then you can see the template and see whether the a core residues which are less likely to be changed because these are the very important residues right then the residues are the outside of the protein. Then also you can see the insertion deletions and the gaps right you have to make the gaps as less as possible because if you introduce more number of gaps it is less possible right. In this case, you check that the gap should be as small as possible ok.



(Refer Slide Time: 06:39)

## 2. Alignment correction

- Template structure (green) with the best aligned target (red) with a large gap
- Target after shifting several residues (blue).
- The gap is much smaller now.



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

I will show one example that here you can see a template structure this in green this is the green one this is the template structure. So, we aligned with 2 types of structures right one is you can see in red and one is in blue if you see the red one this is aligned here this is the red and here this is up to here and this is the gap here this is the large gap if you use the red one and there is another one you can make the corrections using the blue line you can see the blue is up to here and again, it goes from here and this is the gap this is the gap for the one alignment.

This is a gap for the second alignment in this case which one you have to choose the second one right in this case you can align with less gaps, then if you now you have that templates ready and you change the corrections alignment corrections and the alignment is also fine right now what is the next step.

Student: Backbone generation.

Backbone generation, right.

(Refer Slide Time: 07:37)

### 3. Backbone generation

When the alignment is correct, the backbone of the target can be created.

The coordinates of the template-backbone are copied to the target.

When the residues are identical, the side-chain coordinates are also copied.

Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.

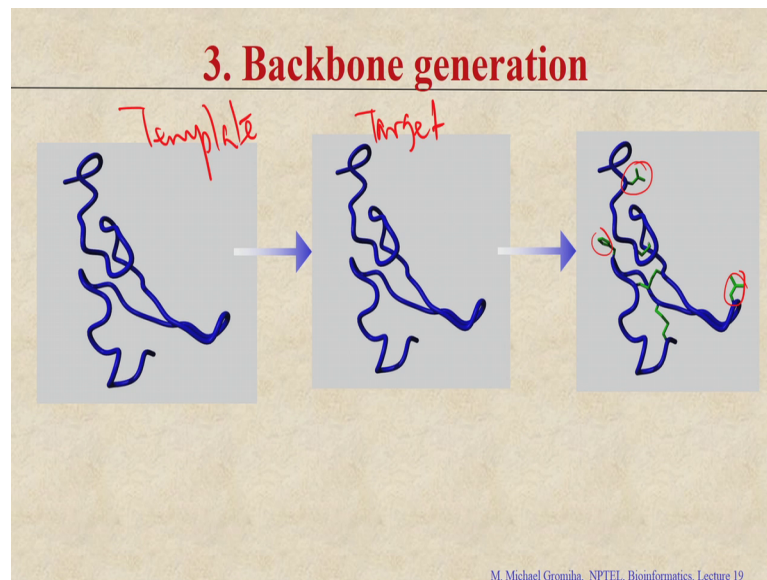
Query: A I K L T V R T A  
Template: A I R L S Y R T V ...  
Backbone: N-C-alpha-C-N-C-alpha-C-N...  
Side chains: KA, RI ...

Because if you have the sequence like for example, you have the sequence right and this is your query sequence and you have the template you get the template right for example, you have a template then if you see in this one. So, backbone is the same because if you draw this right the sequence you can get this one right; N C alpha C N C alpha C N. So, here you have the hydrogen and this is R group if you take the query sequence or the template the backbone is the same what is the difference? Difference is side chain for the take the first one you can see AI and it goes on and the second one AIR.

So, backbone is the same in this case you can just copy the backbone and also if you see some cases the side chains are the same in this case, you can also copy the side chains. This is how we can generate the backbones.



(Refer Slide Time: 08:44)



See, this is the backbone, right of your template right, this is backbone of the template for the target you copy the same right. Now for a target, right, you just copy the backbone that is fine and here what do you do? So, use the same backbone and here you can see some side chains; what we resemble from this side chain.

Student: Similar (Refer Time: 09:10).

The side chains are the same between the template and the target. So, same amino acid residue for example, as I showed here this A and I, they are same. So, this case you can just copy the side chains.

(Refer Slide Time: 09:23)

## 4. Loop modeling

- Check the loops on the basis of steric overlaps
- A specified degree of overlap can be tolerated
- Check the atoms within the loop against each other
- Then check loop atoms against rest of the atoms in the protein

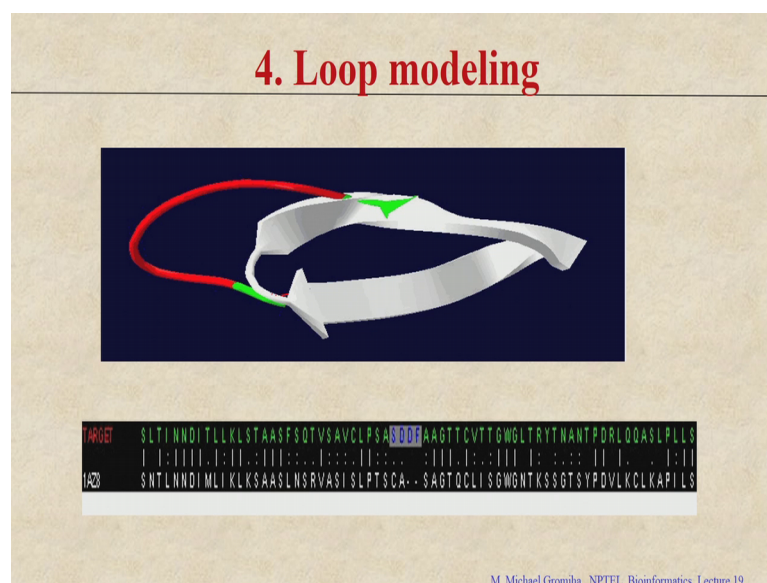
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

And then next is go with the loop modelling because this is the important step because the loops are the basis of them steric overlaps sometimes we have the small loops sometimes you have the longer loops right, if you take the helixes and strands and the confirmation they are restricted right helixes take some specific confirmation right, what is the range in the ramachandran plots.

Student: Something minus (Refer Time: 09:48).

That is minus 50 minus 60 that range right you can see the confirmation restricted even the strands are restricted, but the loop they can take any confirmation. So, in this case they are highly flexible, right, it takes different confirmation. So, in this case you can see a specific degree of overlap right between this different loop regions, then you have to check the atoms right within the loops and see how these atoms fit in the protein ok.

(Refer Slide Time: 10:16)



I will show an example. So, here this is the actual loop right if you take the template and your target you may see the longer once. So, we can see this region here the target you can see a longer loop than the one which are in the your template.

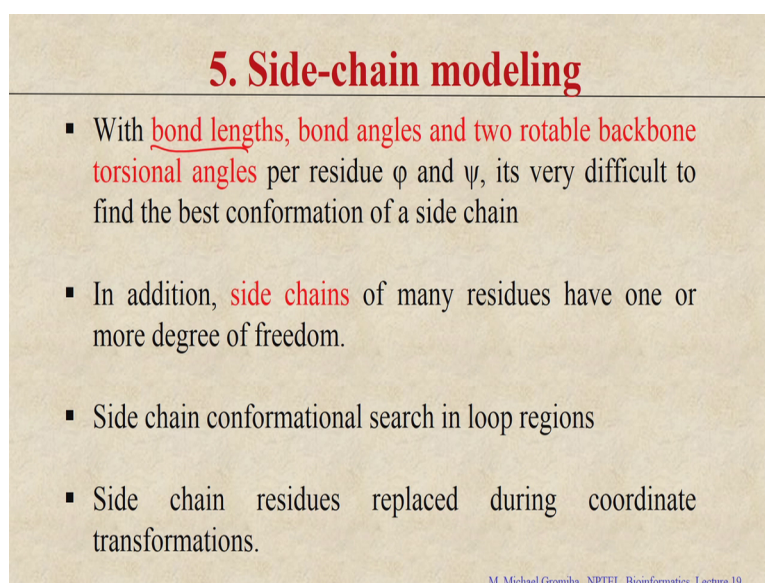
So, in this case you need to model these loops replace these residues there are various options to loop the modelling. Either you can check the same residues in different protein structures which are all start up in the loop regions and then see you can try to connect with that residues and finally, you can make the corrections and energy minimization and to make the conformational changes. So, likewise this loop modelling is an important one because this is highly flexible right in this case you can get the data from these known structures and loop regions and you can make these loops right the conformation of this particular loops.

So, now, you have done the templates to make the alignment correction and you generated the backbone and we fit the loop regions for loop regions are not highly stable right, but we made these loops right within these regular secondary structures. Now we have to change the side chains right if the sequence identity is very high, then it is mostly done; for example, if it is 90 percent right the 90 percent of this chain side chain we can already generated and only a 10 percent of this protein, we need to generate side chain this is the reason why we require the high sequence identity or similarity in this case

almost the conformations almost fixed. So, the now for the rest of the protein where the amino acids are different right, then we need to replace the side chains.

We do the mutations and change the side chains in this case you need to check the conformation of the side chain; for example, what makes a conformation of the different chains right.

(Refer Slide Time: 12:11)



**5. Side-chain modeling**

- With bond lengths, bond angles and two rotatable backbone torsional angles per residue  $\phi$  and  $\psi$ , its very difficult to find the best conformation of a side chain
- In addition, side chains of many residues have one or more degree of freedom.
- Side chain conformational search in loop regions
- Side chain residues replaced during coordinate transformations.

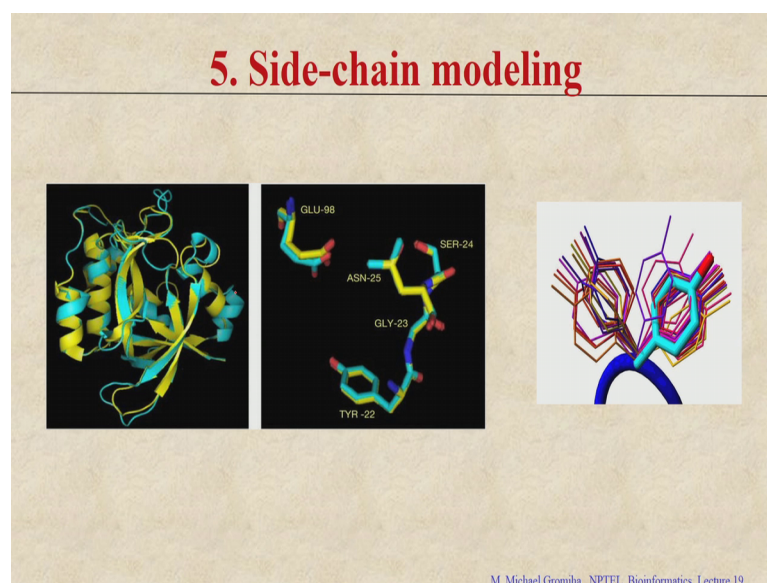
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, we see the bond length you have the bond angle and the torsion angle is phi and psi right because that is mainly from the main chain and the chi angles. So, because if you go with the side chains, it is important to search the side chain conformation, we look into the side chains many residues have more degrees of freedom right for example, if you have these alanine how many rotations are possible.

Student: (Refer Time: 12:36).

If the main chain; you can see one CH<sub>2</sub> group. So, you can have the rotations if you go for this serine or if you have the Threonine you can see depending upon this atom the type right you can see different rotatable ones. So, it can take different types of takes different conformation right.

(Refer Slide Time: 12:53)



So, in this case. So, if we take this is a protein right here I show the tyrosine 22, it has various conformations, we check the all possible conformations and then if you see the libraries and check the one which one is the most probable one.

Based on the type of this specific residue and where it is located. So, whether it is in helix or it is strand or whether, it is a buried or exposed right you can see different conformations and you can fix these conformations for each side chains now we have the backbone and the side chains are the fixed if they have properly aligned with the same residues and if the residues are different then we mutate the residues right and you can build this model based on this available libraries.

So, in this case it is very important because we need the libraries right because if you have the rotations how much degrees one can rotate 300-360 degree you can rotate, but if you do it different angles and different rotations it is computationally very expensive because it search for different conformations. So, for this case you check for the available libraries right there are several rotamer libraries are available how they develop the libraries.

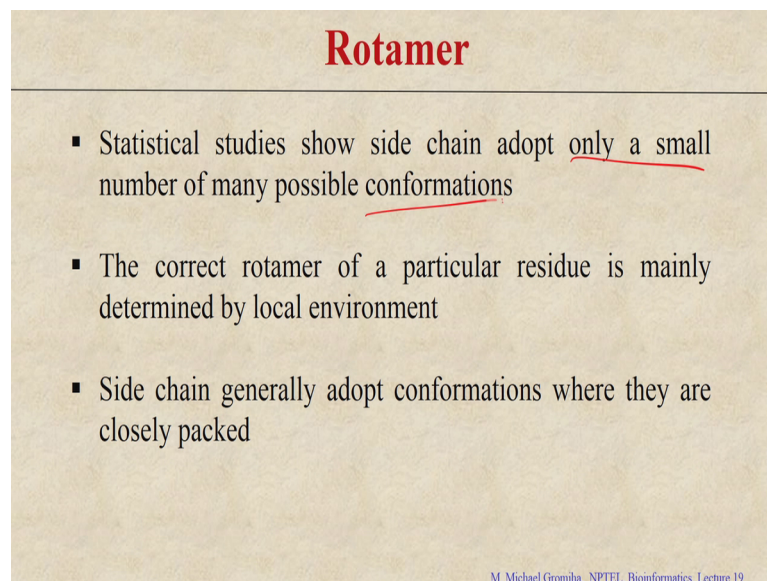
Student: From the investigation.

right from this distance and also from the known structures right. If there is alanine or there is a valine or you can see, what is a probable conformation each amino acid can



adapt and also depending upon these neighbouring residues or the 3 different residues. So, all these can be a probable libraries in this case if you find your residues right in different secondary structures and the different locations you can get the proper libraries from the rotamers and we can use that conformation to build your model otherwise if you do systematic sampling it takes long time and it is very expensive based on the computational time. So, we use the libraries.

(Refer Slide Time: 14:55)



**Rotamer**

- Statistical studies show side chain adopt only a small number of many possible conformations
- The correct rotamer of a particular residue is mainly determined by local environment
- Side chain generally adopt conformations where they are closely packed

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

And then see which conformation each side chain can adopt because each side chain can adopt only a small conformation right then all possible conformations. You can check with the environment and you can use the particular conformation for any side chain.



(Refer Slide Time: 15:07)

### Observations:

- In homologous proteins, corresponding residues virtually retain the same rotameric state
- Within a range of  $\chi$  values, 80% of the identical residues and 75% of the mutated residues have the same conformations
- Certain rotamers are almost always associated with certain secondary structure

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, if you look into this data available in literature in homologous proteins mainly they retain the same rotameric state and if you take the chi values for the 80 percent of the identical and the 75 percent of the mutated residues they have same conformation they do not change much. So, if we look into this original structure template structure and the target structure the difference of the conformation is less right more than 75 percent of the residues they have similar conformations.

So, now, you have the it is the almost the similar structures for example, if you have the template and you have the backbone and you have a side chain you have the loops. So, you have the crude model right now what we need to see, during these steps we introduced several artifacts or the various artifacts we used.

(Refer Slide Time: 15:51)

## 6. Model optimization

- Many **structural artifacts** can be introduced while the model protein is being built
  - Substitution of large side chains for small ones
  - Strained peptide bonds between segments taken from different reference proteins
  - Non optimum conformation of loops

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

We just replace several amino acid residues for example, large side chains for small ones like alanine to valine right just we put it right whether they are sterically allowed or not that we do not know; we have not checked and the second one we take some peptides from different-different proteins like loop modelling right if you do not get the proper data you take from different-different structures we do not check with this the similar proteins and the similar folds we take from different proteins then also we do not know the conformational loops; loops can take any conformation right in this case we use whatever available in known structures, how to rectify this problem right we need to optimize it.

(Refer Slide Time: 16:37)

## 6. Model optimization

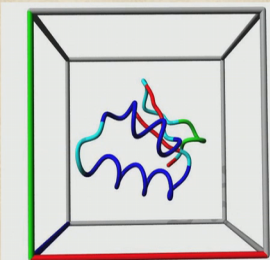
☐ **Energy Minimization** is used to produce a chemically and conformationally reasonable model protein structure

☐ Molecular Dynamics is used to explore the conformational space a molecule could visit

**Molecular dynamics simulation**

Remove big errors

Structure moves to lowest energy conformation



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, in this case, you can do the energy minimization right and see whether this protein or this model is chemically and conformationally reasonable or not. So, there are different methods available to minimize energies right and the use the energy minimization technique to see whether a protein is reasonably good energetically favourable or not then also we can see the md dynamic simulations.

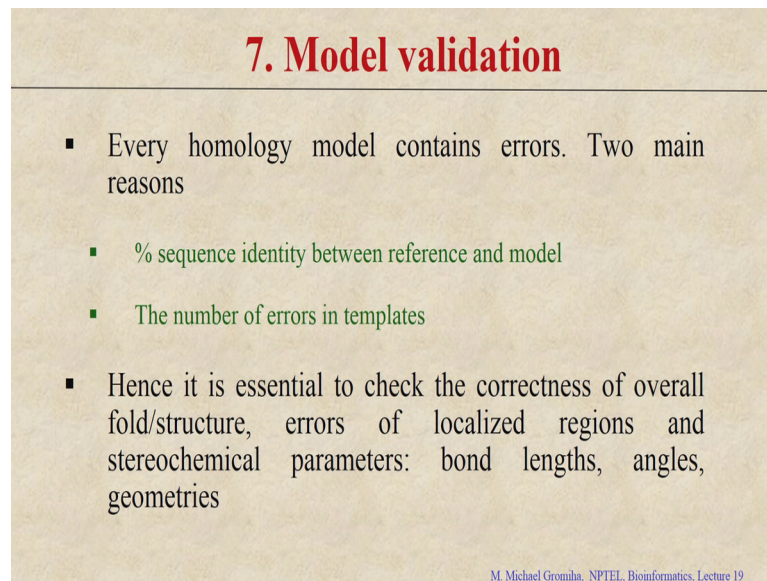
You see whether what are the possible confirmations your protein can visit during different period of time right this can also help to remove the large errors also it can have the low energy conformation. So, you can see this is the backbone and if you do the dynamic finally, we will get the stable 3 dimensional structure. So, once you develop your model is a completely crude model you need to do energy minimization to see and to get the energetically favourable structure and for more aspect if you do the md simulations you can see what the probable confirmations each residue can make and whether it is possible to better confirmation with lowest energy; energy state. So, if you model the structure and you optimized it and you got the structure now your structure is done right.

Then the next question is whether this model is reliable or not, then we do all the experiments we check the proper template and minimise the energy and you are done everything and whether finally, they are going to validate. How to validate this? first we see whether the bond lengths, bond angles and torsion angles right all these

stereochemical properties they are properly set or not and the second one whether the specific important residues for example, active site residues or any antigenic site residues. So, these residues have the proper conformation right for configure the active site residues right.

They tend to interact with the other molecules you should have some type of pocket and so on right, you need to see whether all these things are correct or not.

(Refer Slide Time: 18:37)



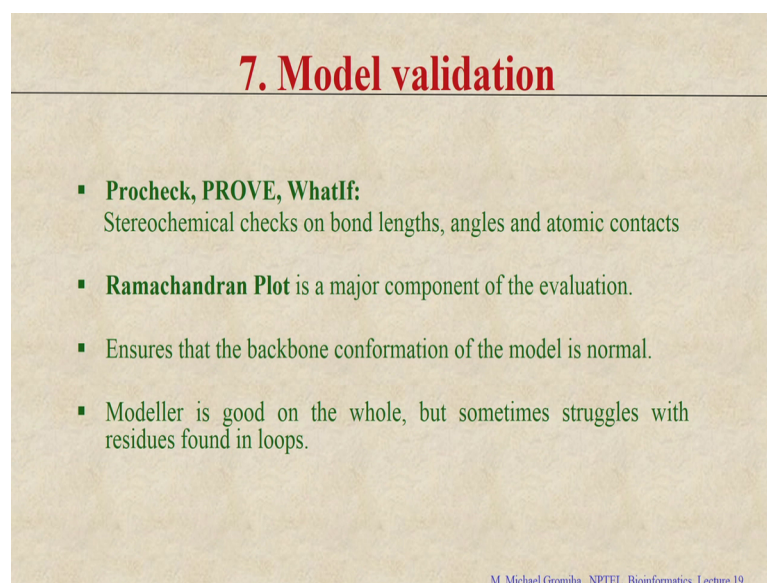
**7. Model validation**

- Every homology model contains errors. Two main reasons
  - % sequence identity between reference and model
  - The number of errors in templates
- Hence it is essential to check the correctness of overall fold/structure, errors of localized regions and stereochemical parameters: bond lengths, angles, geometries

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

And check the correct fold whether it is properly folded because if you see the sequences you can assign the fold and final structures is in proper fold or not. There are various methods available in the literature to validate the structure based on the stereochemical properties or based on the energy, right and based on the experimental information that there are various methods available in literature to validate whether your model is correct or not.

(Refer Slide Time: 19:02)



## 7. Model validation

- **Procheck, PROVE, WhatIf:**  
Stereochemical checks on bond lengths, angles and atomic contacts
- **Ramachandran Plot** is a major component of the evaluation.
- Ensures that the backbone conformation of the model is normal.
- Modeller is good on the whole, but sometimes struggles with residues found in loops.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, here for example, I give an example for example, procheck or this whatIf this will check the stereochemical conformations for example, bond lengths bond angle torsion angles and the atomic contacts.

Whether the proper contacts between different amino acids or not then second one is the Ramachandran plots this is also a one of the major components for evaluating the structure whether the structure is mainly in the allowed region or not. Once you do this it ensures that the backbone conformation of the model is normal; because if the most of the residues are within the allowed region of the Ramachandran plot then you can see that the backbone conformation is almost correct that you can use several servers to get the model, but even then you have to check the and validate their specific models. So, we see the Ramachandran plot then if you have the template and the a query you will get almost similar why it is similar?

Student: Because backbone here.

Because backbone we are just copying from this template. So, in this case almost you can see the similar structures, but during the minimization, we change the confirmation and finally, you need to check whether they are properly aligned.



(Refer Slide Time: 20:16)

## Ramachandran Plot

- The results of the Ramachandran plot will be very similar to that of the template.
- A Good template is therefore key!
- Most residues are mainly found on the left-hand side of the plot.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, you can see; this is a template gives a good structure and you can see most of this the residues which are left side of the Ramachandran plot for the alpha helixes as well as for the beta strands.

(Refer Slide Time: 20:26)

## Ramachandran Plot

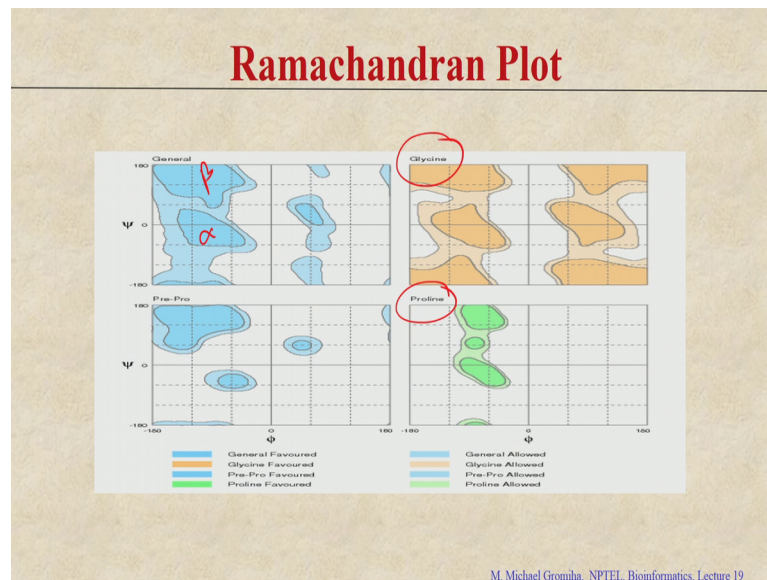
- Glycine is found more randomly within plot (orange), due to its small sidechain (H) preventing clashes with its backbone.
- Proline can only adopt a Phi angle of  $\sim -60^\circ$  (green) due to its side chain.
- This also restricts the conformational space of the pre-proline residue.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

Then you check the specific residues for example, glycine or because glycine it has different conformation in Ramachandran plot that I will show now and the proline the where the proline are located in the Ramachandran plot.

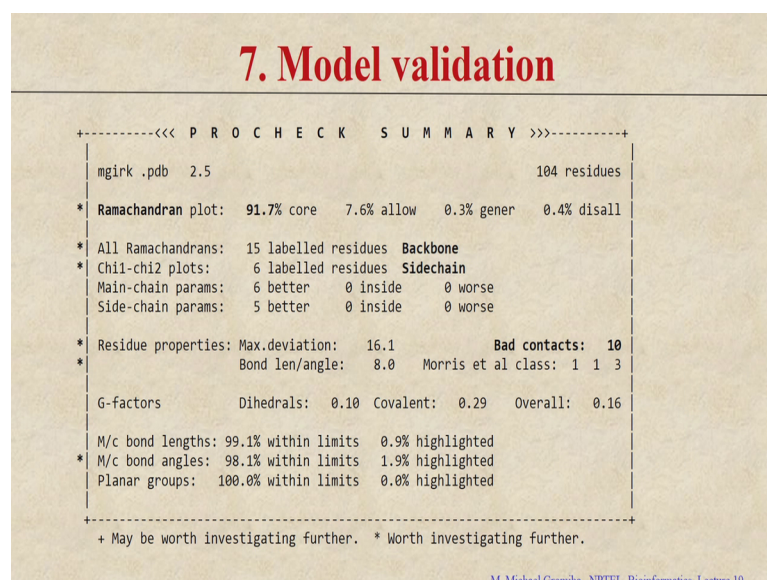


(Refer Slide Time: 20:38)



So, here this is the plot for the general case now this is your alpha helix and you can see the beta strand here and some cases you can see glycine this for the glycine right and here these for the proline. So, check whether the glycines are proline they are also right placed properly they have the a conformational angles which can be seen from this Ramachandran plot.

(Refer Slide Time: 21:03)



So, this is one example you get from the pro check right you get the structure and you submit your structure in the pro checks server it will identify the bond lengths bond angles and torsion angle and how many of them are within the limit.

For example, we see the length 99.1 percent are within the limit. So, these are bond angles you can see the 98.1 percent and the planar groups they are 100 percent within the limit. So, in this case, the model is reasonably good right fine. Then see the Ramachandran plots you can see they are use the allowed region right these are the allowed region or you can say the partially allowed regions. So, it is also fine around 98 percent which are in the allowed regions in this case the model works fine.

Now, I give you one example right as a case study. So, I have one sequence this is the tyrosine kinase protein C-YES kinase.

(Refer Slide Time: 21:54)

### Case Study

```
UniProtKB - P07947 (YES_HUMAN)
>sp|P07947|YES_HUMAN Tyrosine-protein kinase Yes OS=Homo sapiens
MGCIKSKENKSPAICYRPENTPEPVSTSVSHYGAEPITVSPCPSSSAGTAVNFSSLSMT
PFGGSSGVT PFGGASSFSVVPSSYPAGLTGGVTIFVALDYEARITTEDLSFKKGERFQI
INNTEGDWNEARSIAITGKNGYIPSNYVAPADSIQAEWYFGKMGKDAERLLNPGNQRG
IFLVRESETTKGAYSLSIDWDEIRGDNVKKHYKIRKLDNGGYITTRAQFDTLQKLVKHY
TEHADGLCHKLTTCPTVKPQTQGLAKDAWEIPRESLRLEVKLGQCGFGEVWMGTWNGTT
KVAIKTLKPGTMMPEAFLEAQIMKKLRHDKLVPLYAVVSEEPYIVTEFMSKGSLLDFL
KEGDGKYLKLPQLVDMAAQIADGMAYIERMNYIHRDLRAANILVGENLVCKIADFLARL
IEDNEYTARQGAKFPIKNTAPEAALYGRFTIKSDVWSFGILQTELVTGGRVPYPGMVNRE
VLEQVERGYRMPCPQGCPESELHLMNLCKWKDPDERPTFEYIQSFLEDYFTATEPQVQPG
ENL
```

BLAST

→ Same family  
→ Complete structure

**Target Sequence: P07947 (Human – YES Kinase)**  
**Template Structure: PDB ID: 2SRC (Chain A, Human Src Kinase)**  
**Resolution: 1.5 Å**

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

This is important for the target as a colorectal cancer, right. So, to develop a model right first what you have to do?

Student: (Refer Time: 22:02).

You have to find the templates. So, we have to do blast right first we blast. So, we will get different term different sequences of different identities say 90 percent 98 percent 72 percent are different identities then how to choose your template first you see the same family and you see the complete structure is known sometimes you get the 98 percent

sequence identity but structure is not complete if you only know the partial structure. So, you should have the complete structure for this the PDB, sequence wise is align; align and tell it is 98 percent right, but if you see the structure it is only partial structure. So, you need to make sure that the complete structure is available.

So, in this case if we have the human kinase, we choose the 2SRC they belong the same kinase family and the this is c-YES kinase and here you can see this SRC kinase and the look at the resolution what is the resolution of the template you choose right there should be a high resolution right not the low resolution structure. So it is 1.5 angstrom.

So, it is fine and the another aspect is even if you get less identical templates you see they are structurally similar or not right that is also possible for example, if we have these binding sites . So, sites are properly aligned that regions are properly aligned. Are the only specific motifs? So, see that specific motifs that they are properly aligned that they also even the sequence identity is less you can choose the structures in this case your structure will be similar resembled the functional important regions that is also possible fine.

(Refer Slide Time: 23:42)

**Target – Template Sequence Alignment (BLASTp)**

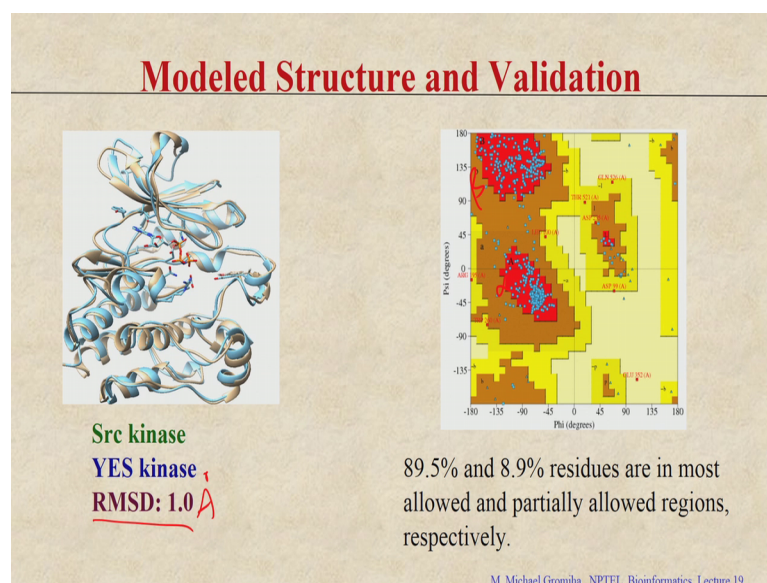
Score	Expect	Method	Identities	Positives	Gaps
815 bits(2105)	0.0	Compositional matrix adjust.	379/451(84%)	418/451(92%)	0/451(0%)
Query 93	VTFPVALVDEARTTDLSTFKKGERPOTINTEGDAHEARSTATGKNGVTPSNVAPADG	154			
VT PVALVDE+RT DLSTFKGER QI-MITEGDA A S++Te+ GYPSNIVAPADG					
Sojct 2	VTFPVALVDESRTEITDLSFKKGERLQIWNTEGDAHLSTQGTQYPSNIVAPADG	61			
Query 153	IQAEHVPFGKQKQAEILLIPQKQGTFLVRESETTGAYSLSTRONDEIRGKWHY	212			
IQAEHVPFGK Q++ERLLLI N NG FLVRESETTGAY LS+ D+ +S 100KH					
Sojct 62	IQAEHVPFGKTRRESERLLNHNPGTFLVRESETTGAYCLVSDFNAGLWVHY	121			
Query 213	KIRKLDGGVITRAQDTQLVWYHTEHAGLCHLITVCPVQDTGLAKDAET	272			
KIRKLD+GG+VIT+R (P++L+V +++)HAGLCH+LITVCPV KPTQGLAKDAET					
Sojct 122	KIRKLDGGVITRTQINSLQQLVAVYSHAGLCHLITVCPVSKPTQGLAKDAET	181			
Query 273	PRESLRLVKGQCFGEVWNTGTTVAIKTLKPETHPEAFLOEADQNKLRHKL	332			
PRESLRLVKGQCFGEVWNTGTTVAIKTLKPETH PEAFLOEADQNKLRHKL					
Sojct 182	PRESLRLVKGQCFGEVWNTGTTVAIKTLKPETHPEAFLOEADQNKLRHKL	241			
Query 333	VLYAVVSEEPVTVTFHSGSLDPLKEGSGVYKLPLVQVHAQIADGAVYERWY	392			
V LYAVVSEEPVTVTFHSGSLDPLK GYKL+LPLVQVHAQIA @VAV+ERWY					
Sojct 242	VQLYAVVSEEPVTVTFHSGSLDPLKGETGYKLRLPLVQVHAQIASGAVYERWY	301			
Query 393	IHDRLAAILVGEILVCKIADGFLARLIEHVEYARQAKFPKUTAPEAALVPTIK	452			
IHDRLAAILVGEILVCKIADGFLARLIEHVEYARQAKFPKUTAPEAALVPTIK					
Sojct 302	IHDRLAAILVGEILVCKIADGFLARLIEHVEYARQAKFPKUTAPEAALVPTIK	361			
Query 453	SDVSGEILLTEITKGRVPPGVNIREVLQVERGVIRPCPCPESLHNLCKIKD	512			
SDVSGEILL TEL TKGRVPPGVNIREVL+QVERGVIRPCP CRESLH+LI Ch+K					
Sojct 362	SDVSGEILLTEITKGRVPPGVNIREVLQVERGVIRPCPCPESLHNLCKIKRKE	421			
Query 513	PDERTFEYDQSLDYPTATEQVQGEIL	543			
P+ERTFEY+Q+FLDYPT+TEQV QGEIL					
Sojct 422	PEERTFTEYQSLDYPTATEQVQGEIL	452			

Identities: 84%  
Positives: 92%  
Gaps: 0%

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, if we take this structure right here this has the identity of 84 percent right with the positives 92 percent right here. They are similarity and there is no gap because all residues are properly aligned and we have the complete structure. So, with 84 percent sequence identity and you can treat this as your template.

(Refer Slide Time: 23:58)



When you make a template then finally you model the structure right. Like the loop modelling and you can use the backbone generation and the side chain modelling right. Finally, you get the energy minimized structure this is a structure you can see the blue one this is for the c-YES kinase and the grey one that is for SRC kinase and the RMSD is around 1 angstrom. So, it is a reasonably good structure, we can get this one then we validate our model this is one of the examples you can do the Ramachandran plot you can see the dots right which are mainly in the alpha helixes and in the beta strands and we see about 90 percent of these residues on the allowed region and about nine percent are in the partially allowed region right in this case you can see the model is reasonably good and you can touch this model for the other applications. So, here the sequence identity is very high. So, you can see the RMSD is 1 angstrom. In this case we can also use these for understanding the different size as well as use this model for identifying the lead compounds in structure based drug design. It is possible because it gives similarly high resolution similar to the structures which you can obtain from experiments.



(Refer Slide Time: 25:17)

## Web-servers for Homology Modeling

- SWISS Model : <http://www.expasy.org/swissmod/SWISS-MODEL.html>
- WHAT IF : <http://www.cmbi.kun.nl/swift/servers/>
- The CPHModels Server : <http://www.cbs.dtu.dk/services/CPHmodels/>
- 3D Jigsaw : <http://www.bmm.icnet.uk/~3djigsaw/>
- SDSC1 : <http://cl.sdsc.edu/hm.html>
- EsyPred3D : <http://www.fundp.ac.be/urbm/bioinfo/esypred/>

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, these are different servers for homology modelling, that you can use swiss model or the WHAT IF different servers you can do it.

(Refer Slide Time: 25:24)

## Tools

- COMPOSER  
<http://www.tripos.com/sciTech/inSilicoDisc/bioInformatics/matchmaker.html>
- MODELER <http://salilab.org/modeler>
- InsightII <http://www.msi.com/>
- SYBYL <http://www.tripos.com/>

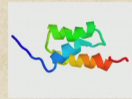
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

And here this is the one of the widely used tools like modeller it has various options to do that. So, you can download this software install this software and you can get the structure using homology modelling. So, now, it did this homology modelling if you do not have any sequence identity then what you have to do you have to do from the scratch right in this case you have any sequence.

(Refer Slide Time: 25:52)

## Ab-initio Prediction

Prediction from sequence using first principle

AVVTW...GTTWVR → 

- *Ab initio* protein structure prediction methods build protein 3D structures from sequence based on physical principles.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

You can get the structure, but we need to do from the sequence based on physical principles. What are the bond length, bond angle, torsion angle and what types of interactions it can make one four interactions one five interactions right different types of van der waals interactions right do all these things it is time demanding because computational expensive why it is computational expensive because we need to do from the scratch and is mainly based on the physical models right we need to calculate the energy and we need to fix the different positions for each atom right there are we have to do lot of conformational sampling. So, it takes lot of time to get a get a structure.

(Refer Slide Time: 26:29)

## Ab-initio Prediction

- Importance
  - The *ab initio* methods are important even though they are computationally demanding
  - *Ab initio* methods predict protein structure based on physical models, they are indispensable complementary methods to Knowledge-based approach

Knowledge-based approach would fail in following conditions:

- Structure homologues are not available
- Possible undiscovered new fold exists

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

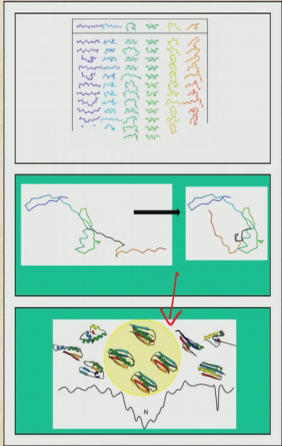


And this is a reason ab initio prediction we can use only for small proteins.

For example less than 100 residues, if it is very high proteins because it is very difficult to do all the minimization in this case what they do. So, they cut into pieces based on the domain information and model the proteins for each domains and then they combine everything together and do the minimization to get the final structure. So, because of the knowledge based approach it may fail because the homologous structures are not available and if there is new folds because they look for these similar folds right to do this knowledge based approach.

(Refer Slide Time: 27:05)

### Structure Prediction with Rosetta



- Select fragments consistent with local sequence preferences
- Assemble fragments into models with native-like global properties
- Identify the best model from the population of decoys

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, Rosetta is one of the widely used methods for the ab initio modelling right. So, what they do they start from the beginning and currently they implemented the Homology modelling along with ab initio right. They take some fragments right based on the local sequence preference and you can see the library of the fragments for the different cases and they think that these have the native like globular properties then they combine all these things together from the populations.

And for the smaller for their unknown cases, they do the modelling and then finally, they generate the model and finally, go with the different possible structures and select the best one this is how they use even the IIT Madras, IIT Delhi, Professor Jairam developed a server called Bhagirath that is also based on the ab initio modelling technique. That is also have reasonably good accuracy, if the residues are less than 100 residues your

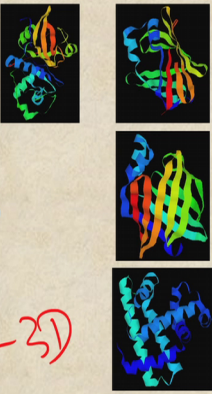
protein is small in size. So, there is another method that is called a threading or the fold recognition.

(Refer Slide Time: 28:11)

## Threading - Fold Recognition

Identify “best” fit between target sequence & template

1. Develop energy function
2. Develop template library
3. Align target sequence with each template & score
4. Identify best scoring template (1D to 3D alignment)
5. Refine structure as in homology modeling



1D-3D

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

In this case, you can see the best fit between the sequence and template for example, if we have a sequence, from the sequence you can derive some information you have energy function and see how this can fit with the structures this is a kind of 1D to 3D alignment. So, what to do we first get the energy function that is very important that is difficult step and template libraries you can see the difference structures available in protein data bank you make libraries and see which energy function fits with which type of proteins right.

And then align the sequence with the template using score. The score they depend on the statistical potentials how far the residues are close to each other or what is the environment based on physicochemical properties and how about the gap based on that they developed energy functions and then compare right this is the kind of 1D to 3D alignment, once the whole fold is set, then they can use the homology modelling and the energy minimization to refine the structure how to do this for example, this is your target sequence right and we have the different templates first you have taken the energy function, right to align the sequences with the structures and then they look for the goodness of fit which one will fit the different templates and the align based on the ranks.

And finally, they get the best one based on the rank how to use the scoring function ok.


(Refer Slide Time: 29:42)

**Protein Threading: Typical Energy Function**

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

What is "probability" that two specific residues are in contact?

- **Contact potential** based on contact statistics from PDB



How well does a specific residue fit structural environment?

physicochemical properties of amino acids

Alignment gap penalty?

**Total energy:  $E_p + E_s + E_g$**

Find a sequence-structure alignment that minimizes the energy function

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

This is a sequence they use 3 a different aspects one is what is the probability that 2 residues are in contact. For example, if we know the sequence what are 2 residues which are high possibility that they are in contact with each other for example, this lysine and another aspartic acid they can be close to each other, for example this and this can be close to each other or this leucine and this tyrosine can be close to each other, then see the structures and they have the scoring function on the second one they see how well they fit with the structure environment they see a physicochemical properties of amino acids right in the sequence and different secondary structures right and the solvent accessibility whether they are buried are they exposed and look into the structures whether it is buried or exposed as same secondary structures and based on that they can define the function then see the gaps. Giving the gap penalty. So, considering these 3 aspects they get this total energy this is one is the probability of the potential and.

This is the structural environment and the gap and they combined everything and map if you have this sequence this would have the energy function of this much values. So, then take all the templates then which one has the best one then check all these aspects and rank the models and check for this sequence this could be the probable model this is a kind of 1D, 3D model right get the sequence and then find the structure this is called the

protein threading based on these fold recognition. So, we discuss about 3 different types of modelling techniques what are 3 different types of modelling techniques?

Student: Homology modelling.

Homology modelling.

Student: Ab initio.

Ab initio.

Student: Fold recognition.

And fold recognition right. So, now, there is a competition called the CASP, right.

(Refer Slide Time: 31:28)

**Critical Assessment of Structure Prediction (CASP)**

- A Biennial *competition* that has run since 1994.
- The next competition will be in 2018 (CASP13)
- <http://predictioncenter.org/>
- The goal is to advance the methods for predicting protein structure from sequence.
- Protein structures yet to be published are used as blind targets for the prediction methods, with only sequence information released.
- Competitors may use Homology Modelling, Fold recognition or Ab Initio structural prediction methods to propose the structure of the protein

**Protein Structure Prediction Center**

Welcome to the Protein Structure Prediction Center

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been seven previous CASP experiments. The twelfth experiment is planned to start in May 2015. Description of these experiments and the full data targets, predictions, interactive tables with numerical evaluation results, domain graphs and prediction visualization tools can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014)

Raw data for the experiments held so far are webbed and stored in our [data archive](#).

In November 2012 we have opened a new rolling CASP experiment for all-year-round testing of all blind prediction methods.

**CASP 10**

Details of this experiment have been published in a scientific journal (protein structure prediction and bioinformatics). CASP proceedings include papers describing the structure and conduct of the experiments, the numerical evaluation of methods, reports from the experiment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structures. Summary of current evaluation of methods used in the latest CASP experiment can be found at: [CASP10 2012](#). The main assessment measures used in evaluations are described in the papers [1], [2]. The latter paper also contains explanation of data handling procedures and guidelines for comparing the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used to predict protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

CASP 10 FORCASP

[Prediction Center](#)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

The critical assessment of a structure prediction, they conduct once in 2 years, right. So, what they do. So, they get the experimental structures from the crystallographers or the from NMR spectroscopy.

Before publishing this into PDB, they keep the structures and give the sequences open to competition. So, you can register and you can try to develop models or you can use the different meta servers. So, there are meta servers available there you can get the models with the different models you can choose which among the 50 models which are the best 10 right you can use that. So, then you submit your model and they evaluate and once the

structure is known they can compare. So, yesterday we discussed about the alignment of structures. So, they use the alignment and see the structurally conserved regions right and then see how far your model can fit with the known structures and accordingly they will tell you ok.

We work under this category this method works fine is with less RMSD right they give different types of proteins with simple proteins or complex structures and so on right. Then you cannot evaluate which method fits very well with which type of these sequences for example, your sequence identity is high we expect the homology modelling work fine.

And other cases ab initio can work well some can do well with fold recognition and so on, right. So, this is the blind prediction method anybody can register and work on that right or maybe you can also try next year we have the competition. So, see whether their method can predict right the structures right fine with these other available existing methods right summarizing what we discussed today.

Student: We started with structure prediction.

Yeah mainly with the predicting the 3D structures from sequence; so what is the name of the problem?

Student: Protein folding.

Protein folding problem because comparing the sequences and structures right known sequences are around 700 to 800 fold compared to the structures, it is very important to build a model right also structures are important to understand the function. So, then we discussed about the different types of modelling right if you the homology modelling we need to think about see the coverage plus the sequence identity and then you can choose the template.

Template is very important and then do the alignment corrections and then do the backbone generation loop modelling and side chain modelling and finally, optimize your model and validate right and finally, if you are happy you should obey the all the rules mainly stereo chemical properties and the energies and then we discussed about the fold recognition that is 1D, 3D, alignment from the energy functions you can check identify



what are the structures which can fit with your sequence. Then we discussed about the ab initio based on the physical energy functions right if the homology modelling or the fold recognition fails then it is very important to do this ab initio. So, you can do with these ab initio structure prediction in the next classes, I will discuss about some of the applications mainly how the parameters we derive from the sequence and structures can be used for understanding the protein folding rates or protein stability or the protein interactions right and the structure based drug design and so on.

Thank you for your kind attention.