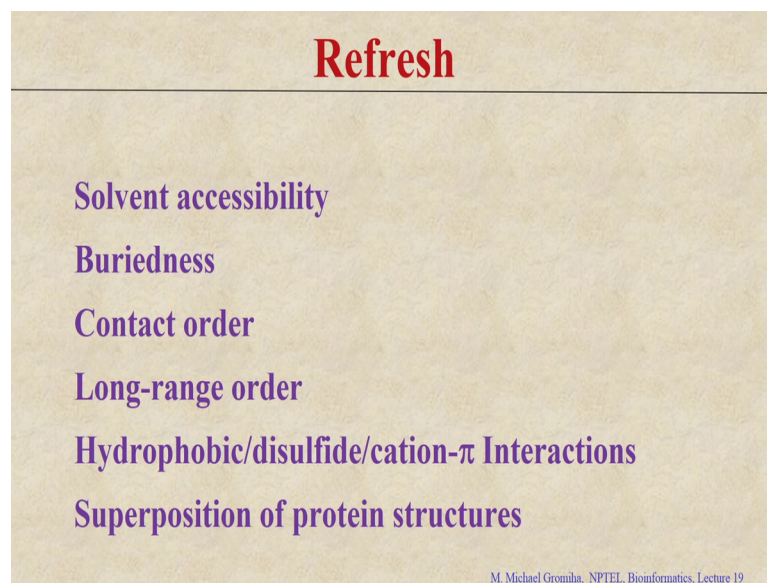


BioInformatics: Algorithms and Applications
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 19a
Protein Structure Prediction I

In this lecture, we will discuss about the applications of the bioinformatics tools or algorithms to structure prediction and folding rates and folding stability and so on. So, in this class I will mainly focus on predicting the 3 D structures of proteins for amino acid sequence and using that bioinformatics tools. In the previous class, we discussed about various parameters or the properties, which can be derived from protein three dimension structures right what are the various parameters we discussed in the earlier classes.

(Refer Slide Time: 00:50)



Student: (Refer Time: 00:52).

Contact maps.

Student: Contact maps solvent accessibility.

Solvent accessibility.

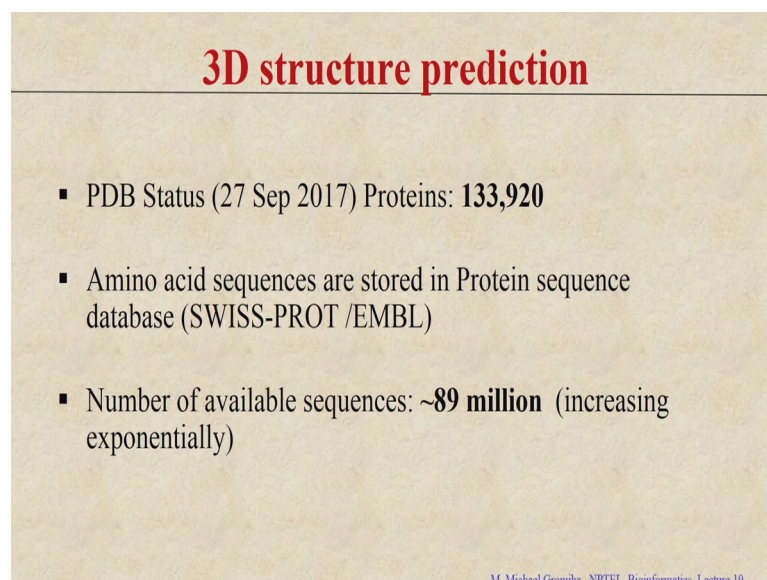
Student: Buriedness.

Buriedness solvent accessibility reduction ratio.

Student: Transfer free energy.

Transfer free energy and based on the contacts you can see the contact order, long range order, multiple contact index then other parameters like as hydrophobicity and the interaction between the 2 different residues, what are the potential interactions right between 2 different residues, based on the type of residues like the hydrophobic, interactions or cation pi interactions or electrostatic interactions and so on. And later on we discussed about applications of the contact maps to superimpose protein structures, if the 2 different structures you can see how far they are similar in 3 D structures, and we discussed about the super imposition of protein structures based on contact maps.

(Refer Slide Time: 01:43)



3D structure prediction

- PDB Status (27 Sep 2017) Proteins: **133,920**
- Amino acid sequences are stored in Protein sequence database (SWISS-PROT /EMBL)
- Number of available sequences: **~89 million** (increasing exponentially)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, when we will be look into the availability of 3 D structures in protein data bank at the movement we have around a hundred and 33 thousand structures in protein data bank. On the other hand the available information regarding protein sequences, in used data base uniprot database right. So, we required it is currently about 89 million sequences right.

So, if you compare the availability of amino acid sequences and structures right available sequences are about 700 fold than the availability of the structures. And the protein structures will be helpful on understanding function identifying the active sites and

antigenic sites right or the different binding sites with other complexes and so on. So, this is very important to have the structures of proteins right than sequence information. Then due to the availability of more number of sequences and less number of structures, it is important to predict the structures of proteins from this amino acid sequence information. So, we look into to this structural determination. So, what are the various methods used to determine 3 D structures of proteins?

Student: Xray crystallography (Refer Time: 02:57).

Xray crystallography, NMR spectroscopy right, even with the latest information as well as the refined instruments right and well sophisticated apparatus. So, you can see the structures available structures are only 133000 and to determine the structures of a single protein right it may take about 10000 15000 dollars, and at least 2 months to one year if the structures simple. When you go with a complex structures like protein-DNA complexes or protein-protein complexes right in this case you will get more time for example, three to four years right to understand the structure and function right as well as the expenses also very high. So, comparing the time to determine the structures of the proteins and the complexes right as well as the cost of getting the structures, it is very important to predict the 3 D structures of proteins from sequence.

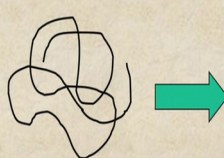
(Refer Slide Time: 03:46)

3D structure prediction

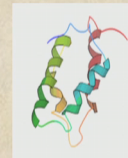
Approximate cost
1 structure: \$10,000-15,000


Duration: 2 months – 1 year

Deciphering the native conformation of a protein (3D structure)
from its amino acid sequence



Bioinformatics





Protein Folding Problem.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

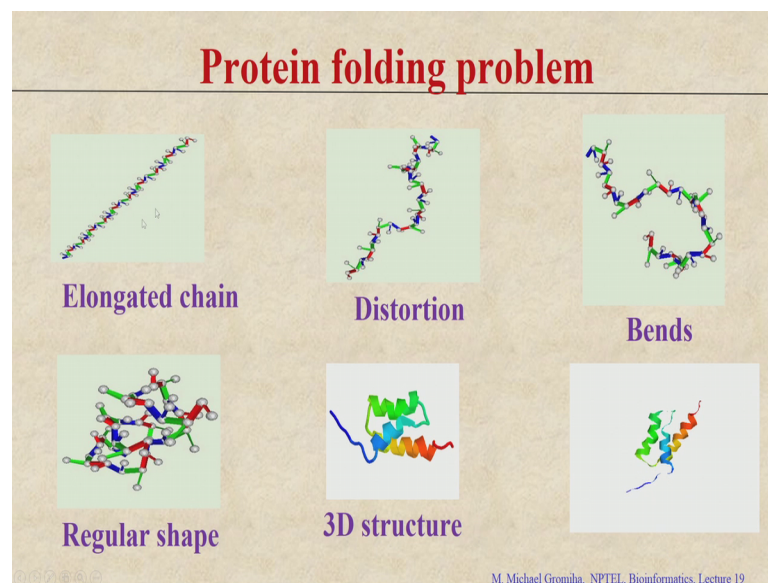
So, deciphering the native conformation of protein from this amino acid sequence rights this is called protein folding problem. In 1963 Anfinsen stated that the amino acid

sequence contains the information regarding the structures, and in this case it may be possible to get the 3 D structures from just amino acid sequence. So, what is the protein folding problem?

Student: (Refer Time: 04:18).

It is getting the 3D structures of a protein right from just its amino acid sequence, how to get that if you took the 3 D structures?

(Refer Slide Time: 04:28)



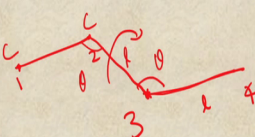
First you can see this elongated chain, then it will be distorted here and there, after that you can see the probable bends right and finally, it will get the regular shape right and kind of 3 D structures right this is the animation you can see that how from the unfolded state of a protein finally, folds into the stable three dimensional structure. So, how to obtain this structure from the amino acid sequence? There are various methods to predict the 3 D structures just from this amino acid sequence.

(Refer Slide Time: 04:57)

3D structure prediction

Methods

- Homology modeling
- *ab initio* method (energetic approach)
- Fold recognition
- Hybrid models



Critical Assessment of Structure Prediction of Proteins (CASP)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

One of the most popular methods right that is called homology modelling, it is based on the principle that if two sequences share high sequence identity or sequence similarity then the assumption is that the structures are also similar based on that principle .

So, they developed the methodology called the homology modelling or comparative modelling. If there are no significant homology is available, then the other method is whether it can recognize the folds, whether they you can make any specific folds; and if that also fails then you can try from the beginning that is called *ab initio* method right you can start from scratch starting from this single atom and build the second atom based on the bond length and that third atom based on the bond length and bond angle, right and the fourth one with the bond length bond angle and torsion angle right you can build up a method that is called *ab initio* method. For example, this atom number one, you can fix the second atom right this one this 2, just with this bond length.

If you know this atom is c and this is also c you can you have the bond length of c and cc. So, you can fix this second position of the second atom; then for the third one for example, if you take here. So, what determines the position of this third atom?

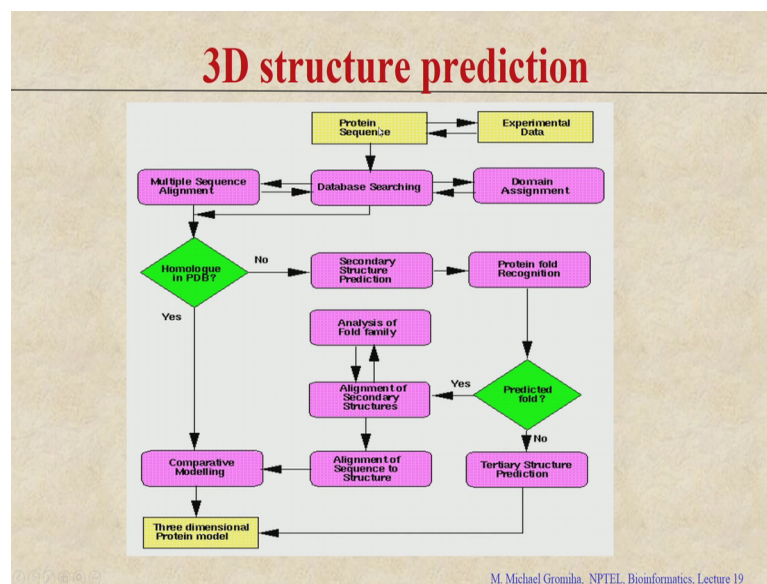
Student: Angle.

One this this length and second is the angle right this theta and length.

So, this will determine the location of this third one. So, the fourth one, here we require three parameters one is the length and here we need the angle, and then we need the torsion angle right they take different conformations. Then we go to the fifth one you have three information you get three measures like length, angle and torsion angle with respect to these 2 3 4 you can fix the fifth atom like this we can grow and finally, make the energy minimization to get the final stable structures I will discuss about this details in later.

Now, recently there are several methods called hybrid method right these combines different techniques, wherever we have the significant similarity or the homology. So, you can take the structures using homology modelling and if the structures are not available or the sequence identity is less, then we do the ab initio modelling and finally, combine everything together and make a single structure using energy minimization. So, this model also works fine right recently these hybrid models. The ability of these techniques can be assessed using a competition called the critical assessment of protein structure prediction of proteins it is called casp, this is conduct once in 2 years right you can enter in to the competition and you can also try to build our models and see how far our method can be accurate compared with the other methods in the literature. So, let us start with the prediction technique.

(Refer Slide Time: 07:52)



So, here the protein sequence I will get the experimental data you can get the protein sequence. So, you want to predict the structure, first see the database searching right available database sequence search. which database you need to search?

Student: PDB.

You can say PDB you can use blast with the database the PDB database with known structures right. So, if you have the significant homology significance identity with any of the structures, you can make the multiple sequence alignment, and see the conserved regions and if you find significant homologous structures in the PDB, then you can go with the homology modelling. So, which type of identity is good for the homology modelling with I will explain a very soon. So, if you can find significant homology then you can go with the comparative modelling or homology modelling fine. If there is no homologous in the PDB. Now it is no; then we go with the next step you can try to do with the secondary structure prediction, if you have a secondary structures then based on the secondary structures you can try to predict the fold which type of fold it can make right.

If you can predict any specific fold, then we can align with these structures and then based on that information you can go this homology modelling at the particular alignment with that particular positions. Even that fails then you go with the ab initio modelling we start from scratch and then try to get the 3 D structures the performance and the accuracy depends on various factors for example, if you have highly homologous structures then will get you can try to achieve the highest accuracy right close to the experimental values, less homologous you can do the ab initio technique that also you can predict with reasonably good accuracy. So, now, let us discuss about the homology modelling, what is the principle used in homology modelling?

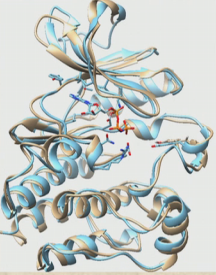
Student: Similar (Refer Time: 09:46).

(Refer Slide Time: 09:49)

Protein Homology Modelling

Prediction of a 3D structure of a protein from its sequence with an accuracy that is comparable to the best results achieved experimentally.

Score	Expect	Method	Identities	Positives	Gaps
815.0	0.0	Compositional matrix adjust.	379/451(84%)	418/451(92%)	0/451(0%)
Query 83	VTTFULVCHAEETTELSPKXKERQITWTEODAAKSLATENDTPSYVAPQDS	152			
Subject 2	VTTFULVCHAEETTELSPKXKERQITWTEODAAKSLATENDTPSYVAPQDS	61			
Query 153	IQAEVPPQVQNDKAEILLQPKQDEPLVSEETTSAGVLSIQDEQSDQNDQV	212			
Subject 62	IQAEVPPQVQNDKAEILLQPKQDEPLVSEETTSAGVLSIQDEQSDQNDQV	121			
Query 213	KTNLQNGEYVTTAGQDTQLVNVYTHAQDLCHLTTCPTVPTQGLAGDAIEI	272			
Subject 122	KTNLQNGEYVTTAGQDTQLVNVYTHAQDLCHLTTCPTVPTQGLAGDAIEI	181			
Query 273	PRESLKLEVLQDQCEVAMETMTTVAJLTLPTPTSPFAFLQEAQNKILBKDL	332			
Subject 182	PRESLKLEVLQDQCEVAMETMTTVAJLTLPTPTSPFAFLQEAQNKILBKDL	241			
Query 333	VPLVAVSEPTVTVTHSKESLLDFLQDEGVKLLQGVPAHQAGQWYVERNY	392			
Subject 242	VPLVAVSEPTVTVTHSKESLLDFLQDEGVKLLQGVPAHQAGQWYVERNY	301			
Query 393	THQDLKAGELVERLVCKADGRLAEIEMETARQAFPTKUTAPFAALVPTTK	452			
Subject 301	THQDLKAGELVERLVCKADGRLAEIEMETARQAFPTKUTAPFAALVPTTK	361			
Query 453	SNMSPGELLTELTVGNPPQVNVRELVENRVRQPCPCPELHQLCKQDK	512			
Subject 361	SNMSPGELLTELTVGNPPQVNVRELVENRVRQPCPCPELHQLCKQDK	421			
Query 513	PREPTTEVQGLDHTFATSPQSEIL	543			
Subject 421	PREPTTEVQGLDHTFATSPQSEIL	452			



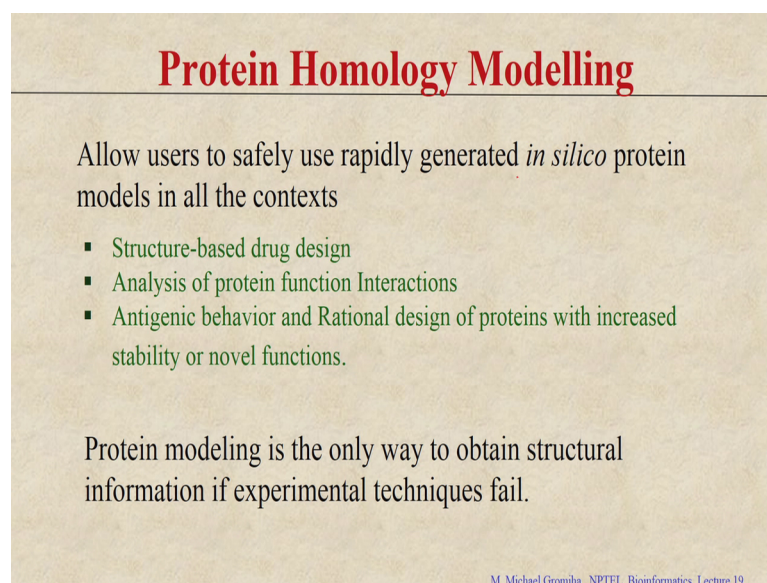
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

Right if you predict the 3 D structure of a protein right from the sequence, with an accuracy which is comparable to the experimental data, if the identity between these proteins are too high. For example, here I give 2 different proteins, here the identity is 84 percent with 92 percent similarities, and if you get highly homologous sequences you can get a model which is similar to the template what you choose. So, you can assume that the 2 sequences they share very common residues or high sequence identity or sequence similarity, then these 2 proteins have similar structures. So, if you see these structures they can show in the 2 different colours one is in blue and one is in gray. So, they also aligned well in the sequence you can see the similar structures with their less RMSD that is about one angstrom.

So, this is a principle used in homology modelling; there are also instances in which the sequence identity may be less, but you they can have similar structures for example, the tim barrel fold, these proteins they share very less sequence identity like less than 30 percent.

But they have similar structures right with the high similar structures you can get in the case of tim barrel folds. So, how this protein homology modelling works and why we need to model the protein using homology modelling; because if you want to do structure based drug design or to understand the important functionally important information for a protein, it is good to generate models using in silico protein structure prediction.

(Refer Slide Time: 11:29)



Protein Homology Modelling

Allow users to safely use rapidly generated *in silico* protein models in all the contexts

- Structure-based drug design
- Analysis of protein function Interactions
- Antigenic behavior and Rational design of proteins with increased stability or novel functions.

Protein modeling is the only way to obtain structural information if experimental techniques fail.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 19

So, here is list of some of the important applications, one is for any target if the structure is not known to identify the small molecules, which can potentially interact to these targets, we require the structures, the structure is not available then it is very important to have a structure right.

In this case if you generate a model, then this model could be used as a potential target for identifying the lead compounds for any drugs. In the structure based drug design it is required to have a model and also to find the protein functions which residues are important for the several protein functions right we require the structures then also you can see the antigenic behaviour and the rational design of proteins with increased ability or functions and so on. For all these aspects modelling is the only way in which you get the structure information right in if the experimental techniques fail to get the structures right. To use a crystal xray crystallography which is required for the xray crystallography?

Student: Crystals.

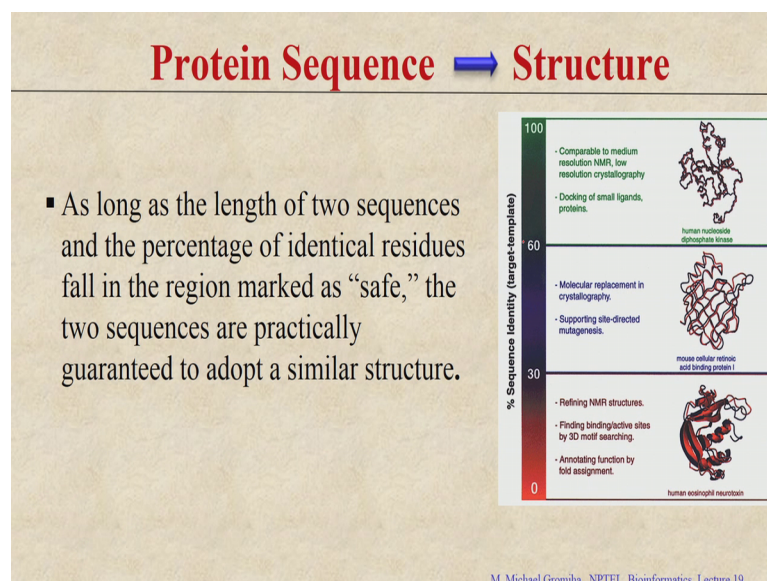
Crystal right if protein is not able to crystallize, then you cannot use xray crystallography for determining the structures right and if you have a give big proteins right sometimes you cannot use NMR spectroscopy. The experimental techniques fail then the only option is modelling, in this case we have to use the modelling techniques for predicting the structure for example, membrane proteins it is very difficult to crystallize, well in this

case you can use this model techniques to get these 3 D structures. So, how we get the conclusion that you can get the structure from this sequence; because if you have the sequence the residues are accommodated in different specific range right also you can see the specific combination of this amino acid residues.

Then we discussed about the different parameter structural parameters right you can see even they are specific combination, some residues which are far away the sequence they are also close in space. Eventually if you see that the sequence contains the information regarding a structure for example, there are several hydrophobic residues. This hydrophobic residues tend to form the core in the 3 D structures; and if you have one positive charge residue and negative charge residues which are very close in sequence.

They have a tendency to form the ion pairs right or salt bridge right. So, you can see that the sequence contains the information regarding the structure, and it is possible right to predict the structure from its amino acid sequence.

(Refer Slide Time: 14:20)



So, now the question is if 2 proteins have different sequence identity for example, 80 percent or 60 percent or 50 percent how accurate you can obtain the model? Here you can see that if it is more than 60 percent, in this case you can get a model which is similar to the medium resolution NMR structures and you can also the low resolution crystallography structures.

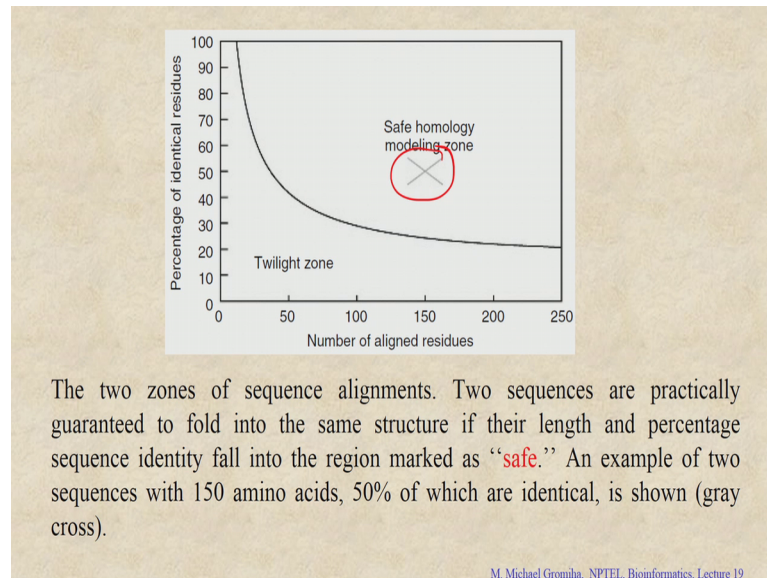
You can get up to that level that that is very high accurate, you can get the accuracy similar to the low resolution structures. And if you get these structures you can get the atomic coordinates and this can be used for docking of small ligands or you can see the interactions, because you will get the high quality structures. And if the sequence identities is less for example, 30 to 60 percent; in this case you can get the accuracy similar to molecular replacement in crystallography how the molecular replacement crystallography works? Because that is based on homology, if there are similar electron density maps, they can see that they may have similar structures. Like that level you can get the structures if the sequence identity is about 30 to 60 percent.

So, you want to do site directed mutagenesis studies this model you can use. If the sequence identity is less you can get a model which is very crude model, in this case it is not good for the very high resolution like high resolution structures. So, we cannot get to all the interactions, but we can try to use it for assigning the fold or you can see the contact between residues if you consider c alpha atoms right because the models are very crude models, we do not get the well-defined structures, but you can use it to understand what could be the probable fold or which residues are possible to interact with each other and so on right. Now the question is, what is a sequence identity we require for homology modelling right where you get 30 percent or 50 percent and 60 percent and so on right what is the average percent accuracy the identity required for homology modelling?

Student: More than 30.

All right; that you can see the number of align residues, right.

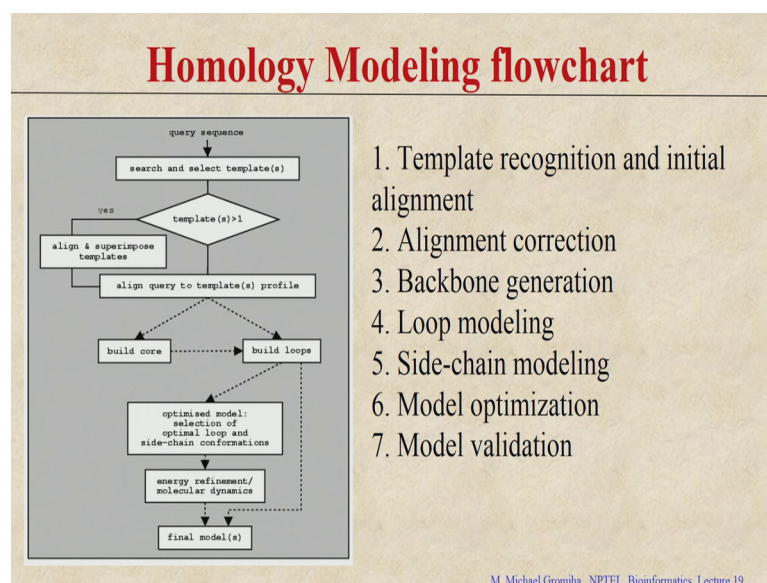
(Refer Slide Time: 16:19)



If we have 150 residues which are aligned, and if the identity is more than 50 percent likewise like see here, this you can say as a safe zone. If it is less number of aligned residues, even the sequence identity is very high and aligned residues are high, because we will get same sequence different structures. So, in this case the homology modelling would not work because it will adopt different types of structures. So, you have the hundred residues then you need at least 60 percent sequence identity, depending upon the number of aligned residues and the sequence identity right you we can select the cut off identity for homology modelling.

Here I mentioned that 150 residues with the 50 percent is safe zone, this you can use for homology modelling and you will get reliably good structures right using the structure prediction algorithms. So, when you make the homology modelling right what are the various steps one has to be consider for homology modelling? The first one is we need to identify a template; your query sequence right you have to identify a template because your model you build based on the template. So, it is very important to identify a proper template right. So, because your structure finally, depends on the selection of templates.

(Refer Slide Time: 17:39)



So, it is very important to take their proper templates right and then we did need to do the alignments how to make the alignments.

Student: Structure (Refer Time: 17:48).

You can do the initial sequence alignment using the blasts right and if you look into the alignments so you check any specific residues which are conserved or any residues which are having important functions these residues should be properly aligned right. If these residues are not properly aligned finally, end up with a structure right, which is different from what you require. So, you need to do the proper alignment and check the alignment whether the important residues right with the experimental information are properly aligned or not. Once you do that you can make the corrections and finally, make sure that your alignment is correct right with respect to the experimental information available for the particular protein. how to get the experimental information for a specific proteins?

Student: (Refer Time: 18:33) literature.

Either use the literature or you can also get from this uniprot database and so on. Once we get the alignment correction, then you go with backbone generation because we get the template your template is ready, then for your target you can just copy the backbone

because the backbone is same right the target and the template, and use this as your to build a model.

Then if you see there are proper alignment of residues, the residues are same, then you can copy the side chains right and other cases you can replace this amino acid residues. Then you took the loop modelling right. For there are several loops then you have to get the modelling the loops right and then you do the side chain modelling, the side chains can take several confirmations that you can use the available confirmations to model side chain, and once the model is built then you need to optimize. Whether you check the model is energetically favourable right you can optimize the model. So, finally, if you are happy with the model then you validate, whether this model is valid based on the energy or the bond length bond angle torsion angle and so on.