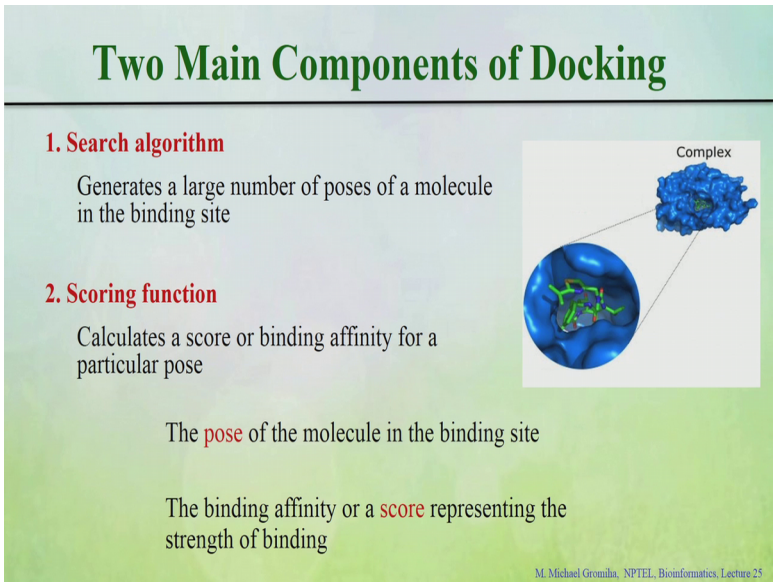
Bioinformatics
Dr. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 25b
Computer Aided Drug Design II

When you dock for example, if you have a protein and the ligand, what are the various aspects we need to think, there are two different aspects right.

(Refer Slide Time: 00:26)



These are the major limitations one is the search algorithm right what is the search algorithm?

Student: It will find search poses.

Poses right because the protein can form large number of poses, they involve active sites even the change of dihedral angles right you can see large amount of conformations.

So, how to see which, see which one is the best one. So that the ligand can interact and the second aspect the scoring function, when your ligand interacts with the protein how to define or how to quantify this ligand one interacts better than ligand 2. So, we need to have a scoring function to decide the binding affinity right between the protein as well as the ligand.

So, two aspects one is pose and the second one is a score right which give you the strength of the binding right then how to do these right.

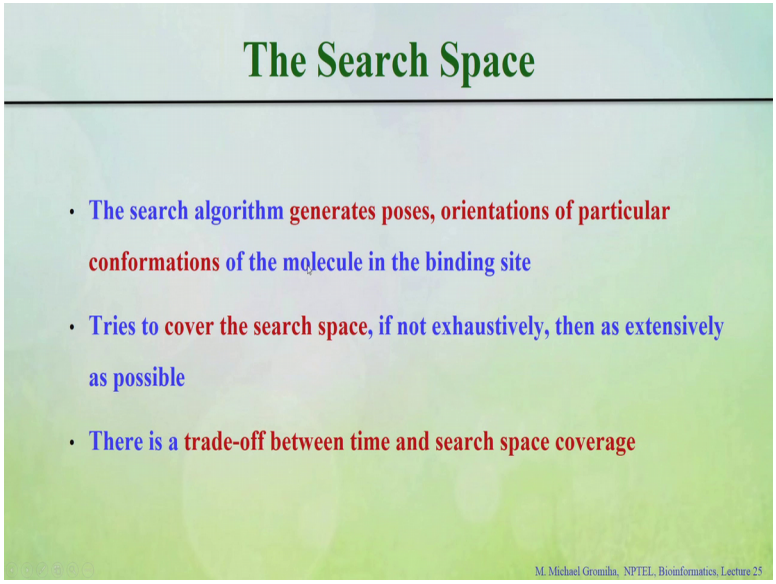
(Refer Slide Time: 01:20)

If you look into the search algorithm, there are various number of search space. If you take the protein ligand docking right it involves many degrees of freedom right what are the various degrees of freedom when you protein interacts with the ligand, you can see at the translation, how many degrees of freedom for the translation? 3 degrees of freedom right because you can see x y z the 3 degrees of freedom.

So, likewise you can see 3 rotations shortly we have 6 degrees of freedom right. Then if you see the conformation degrees of freedom right, because you have the rotations right you can see the conformation changes, there are several conformation changes depending upon where or how many places where we can have the rotations.

Then if you look in to the solvent, the solvent also may play a significant role in the protein ligand geometry. So, in this case we need to have a large search space.

(Refer Slide Time: 02:12)



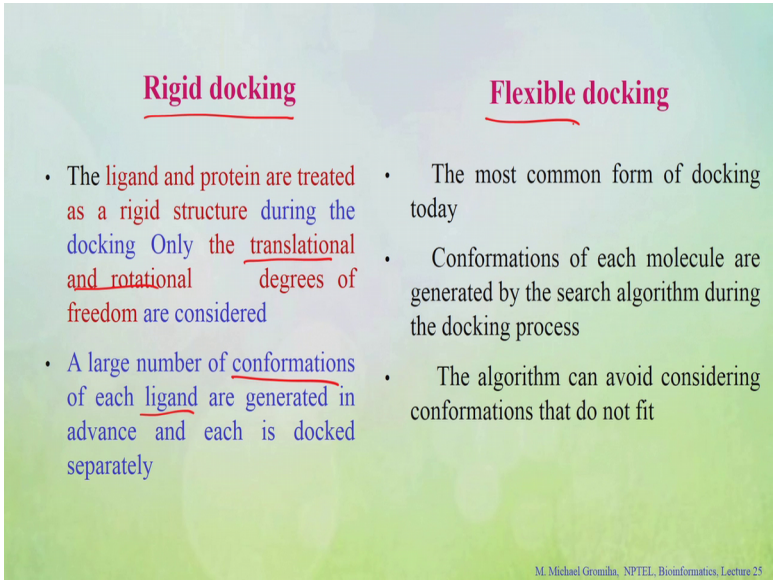
So, then if you need to derive a search algorithm, which generates various poses and different orientations right for any particular conformation of a molecule at the binding state. So, it allows to generate various orientations and it has to cover the whole search space for example, if we see simple amino acids right and in one rotation they can put 0 to 360 degrees, if there are two one will rotate second one will also rotate right you can systematically do it right. So, it takes long time right.

So, there is a trade off between time and search space coverage right. If you have more time then you can search more space if less time we will get less search space. So, in this case we need to have a smart algorithm right to get different search space and various poses right.

So, that also depends upon the type of docking what do you want to do. Sometimes if you know the ligand I said discuss earlier if you know the protein and if we know the ligand right in this case the docking is symbol right here this we used to a rigid docking.

In this case protein is treated as rigid right we have the pose protein pose right. So, we do not make any changes in protein pose, we keep the pose as it is. Then the ligand we make different orientations and different conformations you can make right and for each conformation right we can treat ligand as is they also rigid and the protein also rigid and you can see the docking.

(Refer Slide Time: 03:48)



So, we get one energy. So, we will again conformation also we can decide. So, make the different conformation based on the translational rotational degrees of freedom right and we can see how they fit right with the protein that is called rigid docking.

In this case protein is fixed; if it is one ligand that is also fixed, but ligand we can make different conformations and each conformation right conformation you can get this scoring.

Then the case of flexible docking what is flexible docking? In this case name itself tells it is flexible right in this case you can change the conformation of protein at the active site, you can make various conformations right in this case takes enormous time because we have the different conformations of ligand and also we have different configuration of protein compared with ligand and protein which takes more time?

Student: Protein.

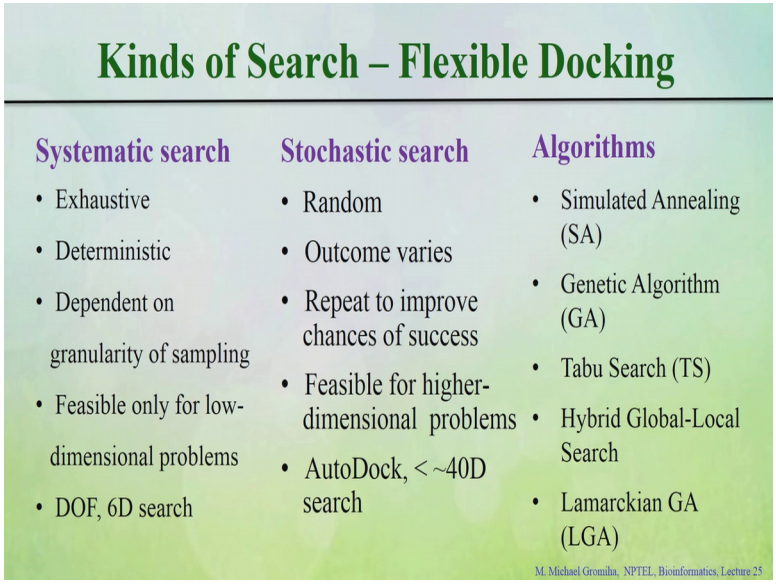
Protein takes more time because ligand take less time because it is a very small molecule and there are many rings. So, in this case there is now much rotations, but you take the proteins right even each amino acid residue right even it is a 5 residues in the active site right there is several degree degrees of freedom.

So, in this case there are various conformations right. In this case you need to general algorithm we need have an algorithm which can sample all these poses or all these

conformations. So, there are various types of algorithms, we can systematically search for example, if you have 3 rotations right and we will systematically we go with one degrees how many times we times you have to do with to cover all the space? For one case you should go 360 degrees, second also 360 degrees or 360 into 360 multiplied by 360 right 360 into 360 3 times into two to complete the entire space or you can do instead of one degree you can use 5 degrees or 10 degree, systematically you can do instead of one you can is 10.

Then 360 you will get 36 times 36 into 36 into 36 we have 3 rotations. If you enormous 4 or 5 then automatically the number will increase, but this is systemic.

(Refer Slide Time: 06:04)



But it is very excess exhaustive right you can determine the exactly which conformation fits well with ligand we have one ligand, change the conformation and fit it and get their scoring. You can see these are the possible cases and we can exactly determine this is the pose where the ligand can interact with the particular conformation of the protein.

That depend upon the granularity of the sampling right how much times you want to do right what is the deviation would be the how much degrees you want to change right. So, that depends on the number of times you need to do the searching right. And all it is feasible only for low dimension problems. If there are two cases you can do it, 3 rotations you can do. If you have many rotations then it is not possible because it is a

very time consuming you can do that because you can see the various degrees of freedom.

So, in order to avoid that right for example, if you want to systematically do that and randomly also you can do right 1 then 10, 20, 50, 60. If you look into these stochastic search randomly you can add fit change and find this is possible or not second you do it and then see its possible or not.

Between these two you can this is a third one, this third one could be probably lay on this place then based on the 3 you can this is the fourth one. So, in this case instead of doing millions of times you can reduce it thousands of times right this is called the stochastic search. If you look into these results between systemic and stochastic search right we cannot find much deviation, but of course, there is a deviation, but not much difference between the poses obtained with systematic search and the stochastic search.

So, here this is random and outcome varies because if you use 100 conformation we will get some data get the binding. If you use another 100 close to this not be the same there will be different. So, depending upon the sampling you do, depending out the space you walk that will give you the outcome and the outcome varies right and you can repeat to improve chance of success right for example, instead of one million times go to 10000 times, with the result is similar to the one thou one million times then that is fine if you not you can change another chance right also among these times they can choose one and 3 are not good. So, you will go to the fourth one fourth one is in between 3 and 4 then 5 one you can decide we can assign some other conformation between these 4 and 5 right they can do that.

And here you can solve the feasible for higher dimension problems, if you have several rotations then also possible right to map different conformations in the case of these stochastic search fine. So, the difference between systematic search and stochastic search are, what are the differences?

Student: It is random.

This random because stochastic, this is exhaustive because systematic. Systematic we can get the exact pose. So, this is the one the ligand prefers to bind, but the case of

stochastic, outcome will change and the systematic will depend on the granularity of sampling right what is the angle you do you require to change right.

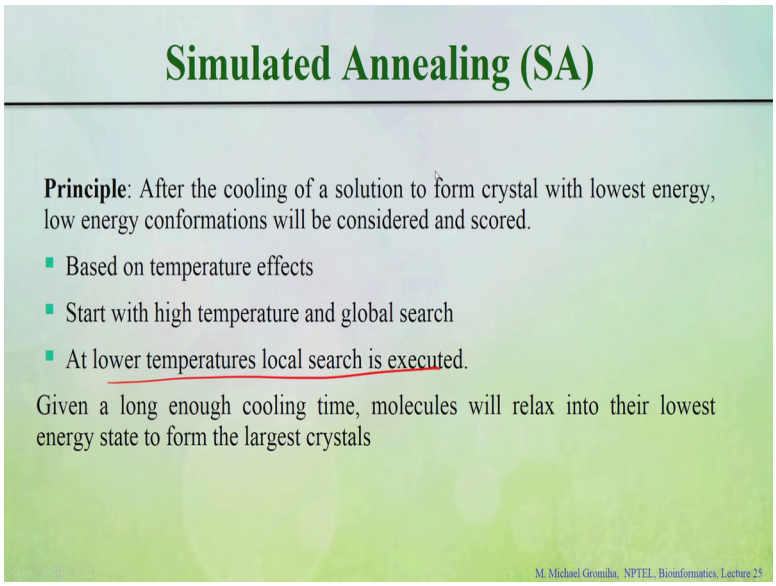
So, here this will repeat to improve the chance of success you can change again, and the systematic is feasible only at the lower dimension right, but here the stochastic you can go with the higher dimensions right mainly for example, autodock is about forty dimension search.

So, there are various algorithms have been proposed right to do the sampling right for the case of stochastic search right aimulated annealing, aenetic algorithm, tabu search right hybrid global-local search and Lamarckian GA algorithm so on. So, various algorithms have been proposed right to do this the stochastic search.

So, I will explain only two of them, first one is simulated annealing here it is mainly based on temperature. So, the assumption is when we have these high temperature. So, the molecule have more degrees of freedom in this case we have lot of conformations, but all the conformations, they won't pass through these fit any scoring function, it would not pass through this scoring function in the case most of them are rejected .

Then go to the accepted once then we decrease the temperature when we cool down you consider as less degrees of freedom in this case you can reduce a conformation sampling space.

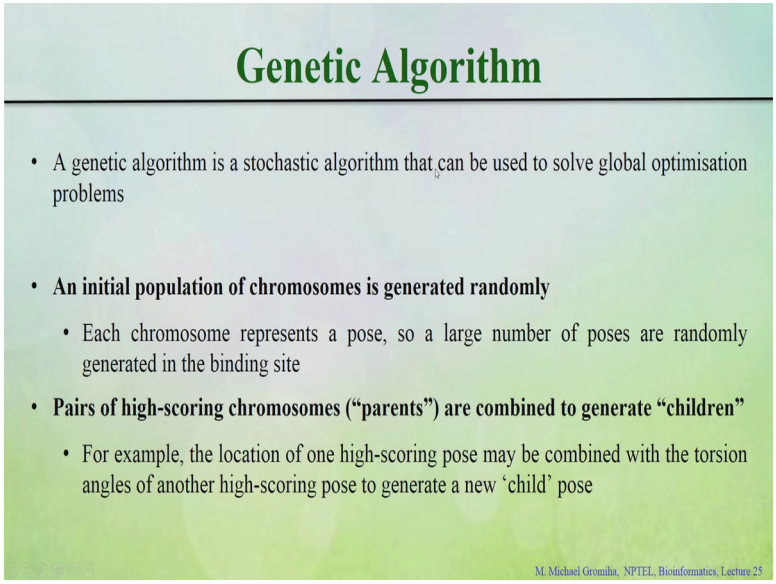
(Refer Slide Time: 10:40)



And finally, lower search you can lower temperature you can get less number of sampling.

So, do the local search and you can identify right the probable conformations right using these a simulated annealing method.

(Refer Slide Time: 10:57)

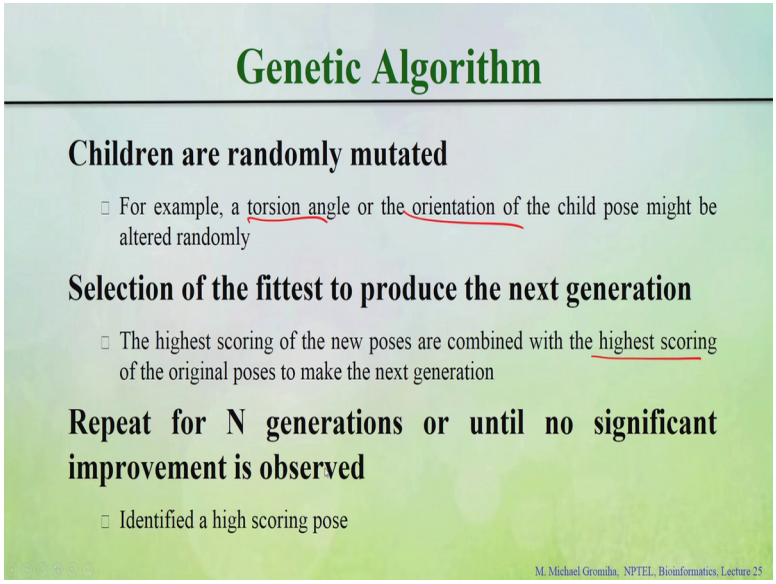


And the another one this is slightly used genetic algorithm right the genetic algorithm is a stochastic algorithm to sample the data right they mainly used in the global optimization issues.

So, what they you first take a pose. The initial population is generated the randomly generated initial population. So, each chromosome represent a pose. So, this case you can have a random number of many poses right that generate the binding states, then you check there are some scoring functions based on torsional angles and the other energy potentials right see what are the locations of high scoring. Take the high scoring ones and you merge the high scoring poses to generate new child pose.

So, take randomly generated poses some of them are having high score, some of them are having low score and then merge this high scoring to generate a child pose.

(Refer Slide Time: 11:57)



Then the child pose they do some mutations, they can change the torsional angle or then check the orientation. So, make new conformation and then see whether the score increases or decreases. Then if it is increases then you make the this new child ones with the previous ones original pose to make next generation. Do it again and again and again right up to the n generations or n significant improvements are is observed.

In this case you can identify a high scoring pose. So, starting with the several random conformation. So, random poses you get the one which is the high score, then merge the high scores and form the child then you do the mutations. Mutations means the change in torsional angles and the rotations and find this scoring then again the used with these the higher high scoring ones and finally, we continue until unless we get this significance no significant improvements right then you can identify this pose.

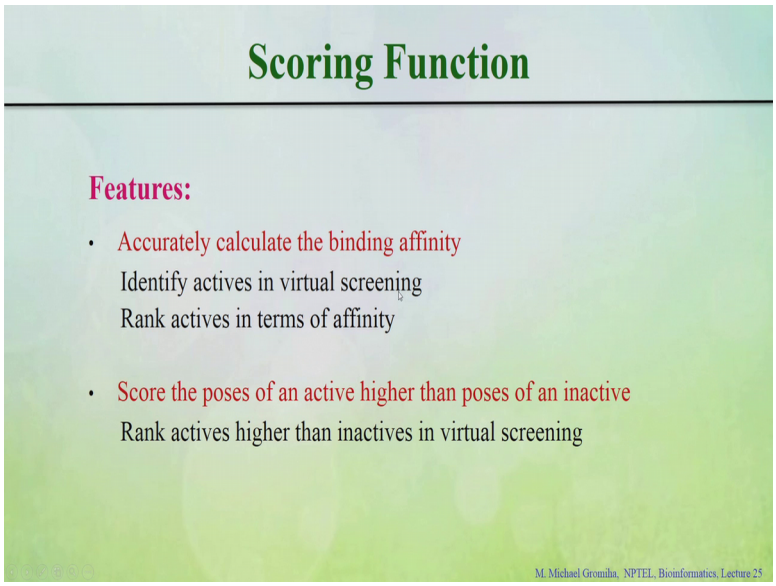
So, likewise there are various other algorithms available in the literature to search right the conformation space and see; what are the probable conforming space for the ligands to interact at the protein site, fine. So, this is a one aspect what are the two major aspects in docking? One is a sampling

Student: Sampling and scoring.

Another one is scoring function right what is the meaning of scoring function what information you can get from scoring function.

Student: difference in different binding force which one is the best

(Refer Slide Time: 13:33)



Best one right you can see it is accurately calculated binding affinity. So, for example, if you have a set of compounds or you do the virtual screening that select some compounds from you pole up compounds, it can identify the actives. So, then also it can rank the actives. If 10 compounds you can use this scoring function right it will give the values for all the 10, and see the best ones you can rank the actives in terms of this affinity right.

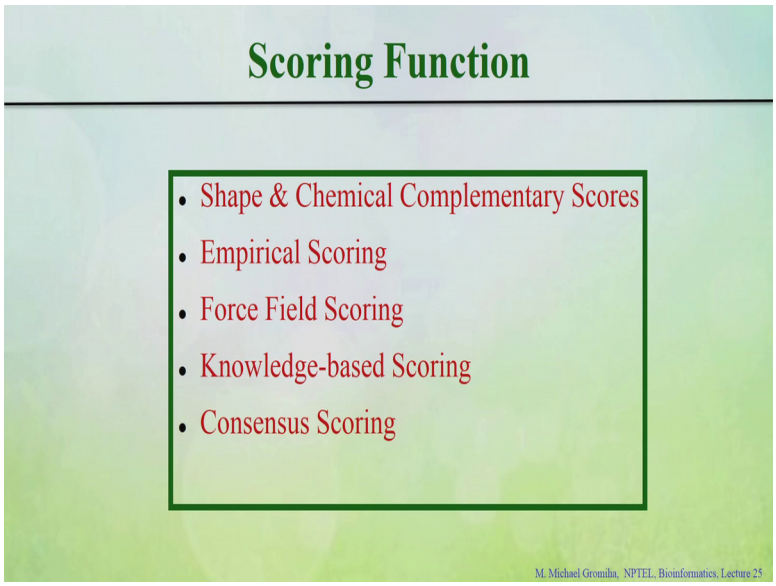
Then also you can score the poses of actives higher than poses of inactive right. For example, you have the 10, 100 actives and one thousand inactives and if you have scoring function you can make a scoring function in such a way that the poses of the active should be should be the high score compared with these inactives in the case of virtual screening. Likewise you can adjust the scoring functions then this scoring function can be useful to identify the actives compared with the inactives right.

So, then how to generate how to develop the scoring functions right there are various ways right. The one is simple one you can say that you have a protein and you have a ligand be in the scoring function, generally if you talk what do we expect if you have your protein and the ligand to interact.

Student: the shape compliment.

The shape compliment it should interact right this should have the proper position. So, where we have the convex or concave in the protein which should be opposite in the case of this ligand, then they have good interactions and then they can make the complex with the tight binding.

(Refer Slide Time: 15:01)

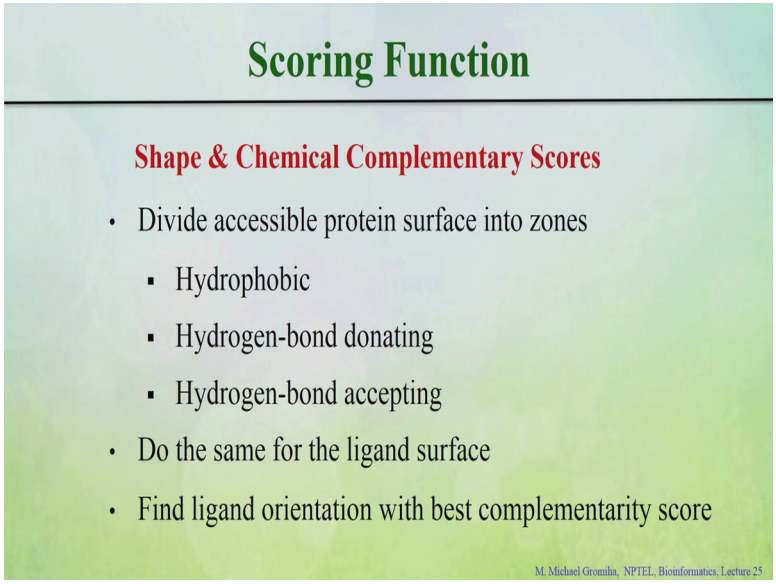


So, one is a shape and chemical complementary score. So, how far their complementarity then we can do some empirical scoring depending upon the known structures they can calculate the interaction energies; from that you can check with the known values activity values right and then see some empirical scoring right. Then we can see a force field scoring that depending upon the force field interaction energies. We can calculate the interaction and using the interaction whether you can see this could be the probable ligand right probable potential compound.

Then you can use a knowledge based scoring. In this case if you have the ligand and the protein you can see the contacts and convert the contacts into potentials right and this is this potentials you can get the scoring function whether this can identify the actives right compared with the in actives.

Then the consensus scoring as we discussed in the structure secondary structure prediction you check the different types of scoring functions and pick up from here and there finally, make the another scoring function, right consensus from different functions I will explain the details now how to do that?

(Refer Slide Time: 16:11)



First one is the shape and chemical complementary score. So, in this case either you can see the cavity of the protein site, active site and the ligand you can see where this can fit you can find the complementarity. So, in this case. So, the protein surface right, they are the active sites you divided into various zones and from this you can see how many hydrophobic interaction it could make and whether in the ligand side also the part the partner ligand we have that much atoms are available.

Likewise hydrogen bond donator donor, how many donors in the protein side and how many acceptors in the ligand side and so, other way around. How many acceptor in the protein side and how ligands donor acceptors in the ligand side. See if you have the surface right we can divide the surfaces several zones and each zones because ligand is small one right. So, whether this can fits with the binding site or not.

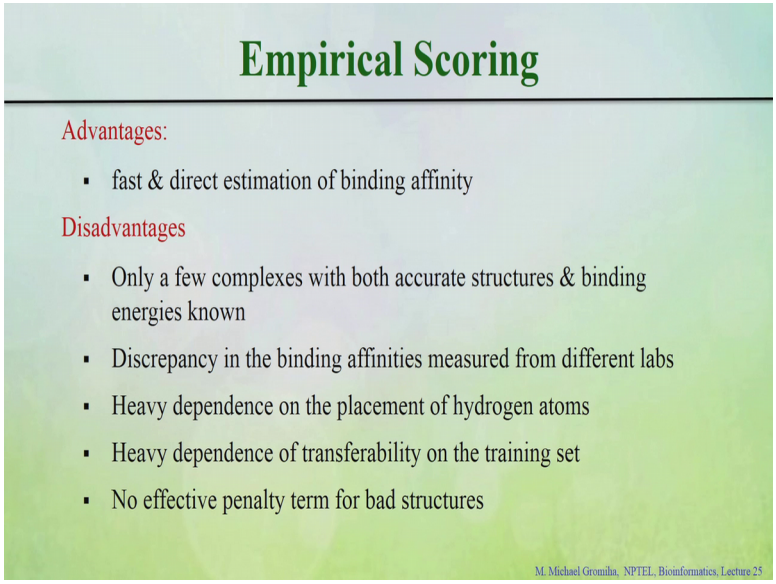
So, we see the complementarity how what is the how many hydrogen bond donors, if there is 10 hydrogen bond donors in the proteins, and the ligand also there are several hydrogen bond donors, but no hydrogen bond acceptor. Then can we fit? no right because if we have the hydrogen bond donor here, there we need the acceptors. So, make the hydrogen bonds right.

So, they have different types of interactions hydrophobic, hydrogen bond donors, and hydrogen bond acceptors at the protein side and get the same from the ligand side and

see the complementarity whether these numbers will match or not. So, if they match then you can say this ligand one is better than the ligand two right

So, likewise you can see the scoring based on the shape or the chemical complementarity scores right you can do that.

(Refer Slide Time: 18:02)



And the second one is the empirical scoring right how the empirical scoring will work?

Student: Energy.

Right in this case if we first we get the binding affinity of non-ligands right for example, if you have 100 ligands, with the specific target right. So, we know the binding affinity experimentally known right.

Then the right said we can calculate various parameters for example, torsion number of torsions, number of hydrogen bonds, energy due to hydrogen bonds.

Student: (Refer Time: 18:43).

Energy due to the charges, energy due to the aromatic interactions and the hydrophobic other hydrophobic interactions, we can calculate various parameters right. Because we have the protein and we have the ligand right the affinities known and the complex is available right then we can calculate all the parameters.

So, when we get the all the parameters take the compound one right you can see deltaG1 this is known this is equal to a into deltaG torsion 1 plus b into delta G hydrogen bond 1 and so on right. So, you take the compound two you can get the value delta G 2 you can use this right.

So, now you here if you see the right hand side these delta G or delta Gs you can calculate from the complex or if we take a free protein you can calculate this all the values are known. So, what is the variable here what is the constant here?

Student: A.

The coefficients are come this for example, this a b plus a you can say constant. So, that we do not know. So, here we know the delta G this experimentally known right. So, now, fit the fit this equations for example, we have 100 compounds, you will get the 100 equations then how we fit these equations.

Student: Principles of least square

using principle of.

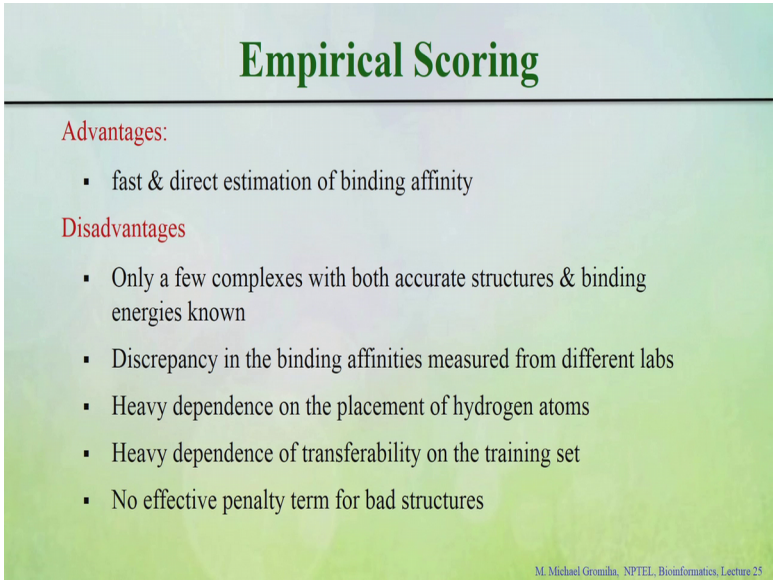
Student: Least square.

Least squares you can do right. So, this is the principle of least squares you can fit this equation then you can calculate all these constants.

Now, for any new compound you have the you can make the ligand and you can see the protein, when you are make the complex when we have the all the constants, we calculate all the energy terms and fit and then we get the value of delta G. This will help you to score if there are another 100 compounds or 200 compounds, we calculate all the energy terms and the constants we know. So, fit this equation right you can fit this equation to get the value for these delta G.

Then we have finite compounds, you get the finite compounds we get your delta G and you can rank or this is one this is two based on this scoring function right. So, what are advantages and disadvantages of we are having this empirical scoring? The advantage it is very fast because we need to calculate only this energy these energy terms.

(Refer Slide Time: 21:12)



And we can also estimate directly we can estimate the binding affinity, because we have the equation get the values substitute the values and they can get delta G very simple right.

So, what are the disadvantages?

Student: enough data.

Because when we for developing this empirical scoring, we need the value for the delta G, but delta is available for thousands of compounds then you can get better results, but delta G is available only for few compounds right for example, if you take any specific target you do not get many many compounds you get only few compounds.

In this case you can, the accuracy of getting these structures and binding right its difficult because the fitting is not proper. Then we can see the discrepancy in binding affinities, sometimes we have the binding affinity reported by the different labs are different in this case is difficult to fit with this particular equation.

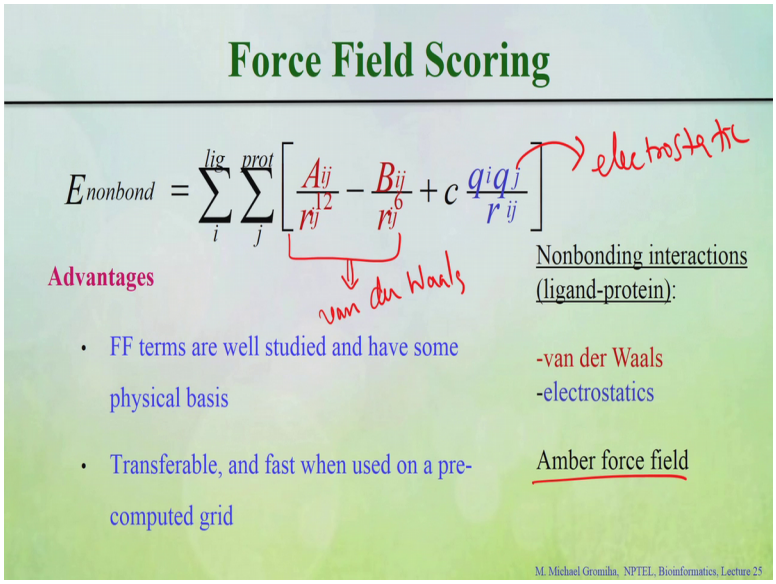
We can also its depend on the placement of hydrogen atoms, because they are making the different interactions. So, it depends upon where the hydrogen atoms are placed. If we wrongly placed then we will have the issues with the hydrogen bond formation right. Then also depending upon the transferability of the training set, then the next one is the

there is no penalty time for the bad structures if the structures are not good even then if they interact right make it interact, in this case there will be good score

So, these are the several disadvantages of having the empirical scoring function, but the good part is if you get good number of affinity values and good number of structures are available, then you can easily get good score using this empirical function So, this works fine for some cases.

Then the next one is a force field scoring right what is a meaning of force field calculate the different types of energies.

(Refer Slide Time: 22:59)



So, here this is this mainly this for the.

Student: van der Waals (Refer Time: 23:01).

van der Walls energy right and this term

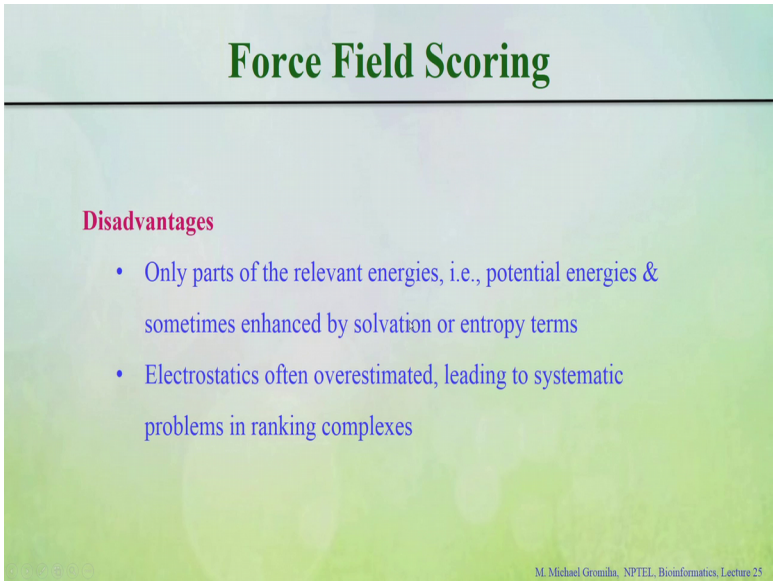
Student: electrostatic.

This electrostatic. So, this is the major terms used in amber force field is widely used force field. So, this is well studied and also this is based on physical basis because calculate the energy.

Then we can see this energy and calculate the energy for any complex, we have protein and we have the ligand right you can calculate the energy because we know the distance right and directly you can relate the energy is low this is the probable compound the energy is high this is not a probable compound.

So, for all the pairs you can calculate energy you can do it, it is easy.

(Refer Slide Time: 23:43)



So, what are the disadvantages? Here it use only some specific terms of energies for example, electrostatic and the Van Der Waals energies right. In this case they may it increase, require the solvation as well as the other types of entropy terms thats missing.

So, you need to refine this force field right to get the better results sometimes the electrostatic is over estimated. So, in this case it will be difficult to rank the complexes the electrostatic over estimated, then if there is the only positive negative charged atoms then there will rank very high they ignoring this hydrogen bonds and all.

In this case we need to be cautious when you use this force field type scoring. Second aspect is they also use various constants, the constants also important because they derive from the a known structures right this is also plays a role when we get the non bond interactions.

So, we discuss about 3 different types, what is different types of scoring functions? shape and chemical complementarity

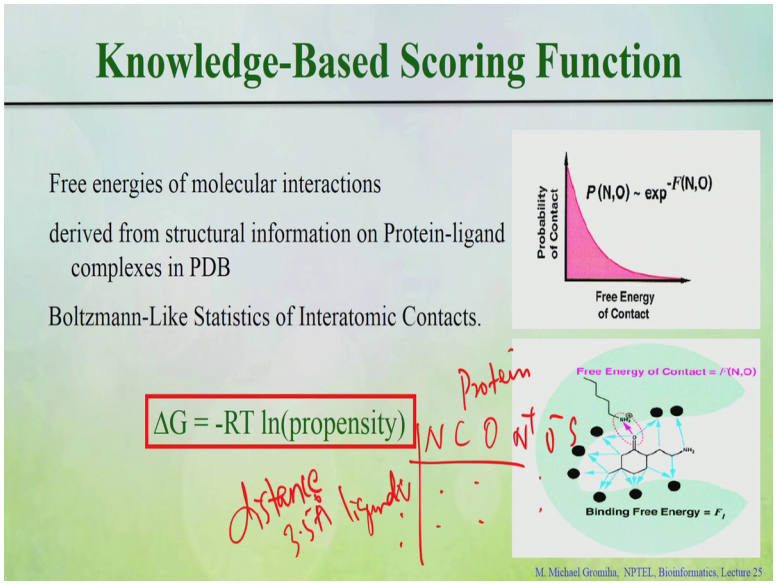
Student: Empirical.

Empirical scoring function

Student: Force field.

And the force field right now the next one is knowledge based scoring function the knowledge based scoring function, that is depends upon how the atoms are in contact right in the ligand and the protein. So, what they do.

(Refer Slide Time: 25:02)



So, they make a matrix right one side is for the protein other side is for the ligand, and see the possibility or probability of contacts between two atoms, what are different atoms in protein?

Student: CH.

NC.

Student: H.

O N plus O minus S right likewise ligand also you can have the different, types of atoms and see how far the atoms are distributed. Take any a space say 3.5 angstrom or 4 angstrom and within the distance they can see the pair preference right. Whether the ligands right there is a N and this N are in contact or not N and O are in contact or not

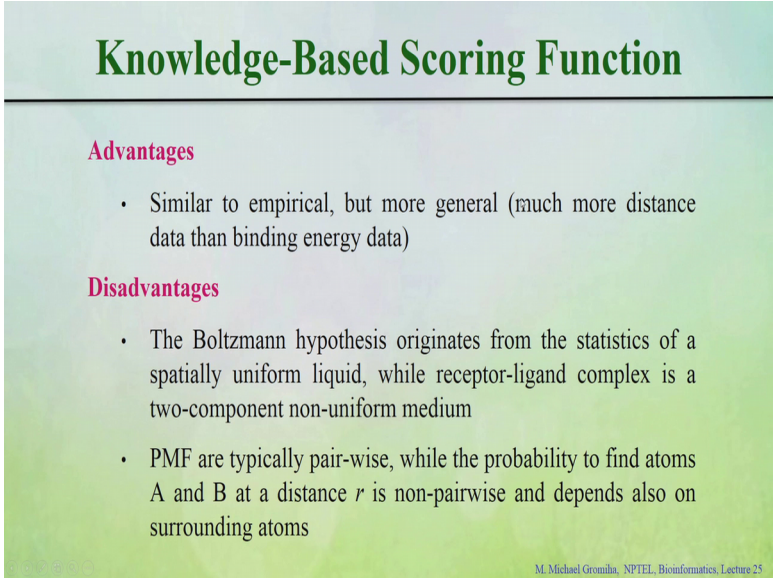
So, they get these preference based on distance right then you see the any specific distance for example, 3.5 angstrom right. So, totally how many residue pairs among the residue pairs how many residue pairs in different specific groups right normalize the values and finally, we can get this a propensity right take any distance get the atom contacts, any pairs. So, based on atom contacts you normalize the total number of contacts this can help, help to get this propensity.

Then we convert the propensity right into the energy potential right using this equation minus RT logarithmic of this propensity propensity is treated as a kind of partition coefficient; that means, data from several sampling right then this propensity is converted to the free energy using minus RT into logarithmic propensity

So, now you can get these values then we have the new protein ligands we have the protein and the ligand, we can get the distance right for each pairs we have the delta G values and substitute the values and some of the data and finally, you get the score.

So, now for any complex we can say get this score and add up these scores. So, this is based on knowledge based scoring function how this works? If any specific pairs of atoms they are preferred and that preferred residue atoms are also present in your new complex, then it will have high score. If it is not preferred right if there is highly preferred in this new complex it will get low score. Depending upon the availability of these contacts in known protein ligand complexes right this will score the new complex.

(Refer Slide Time: 27:37)

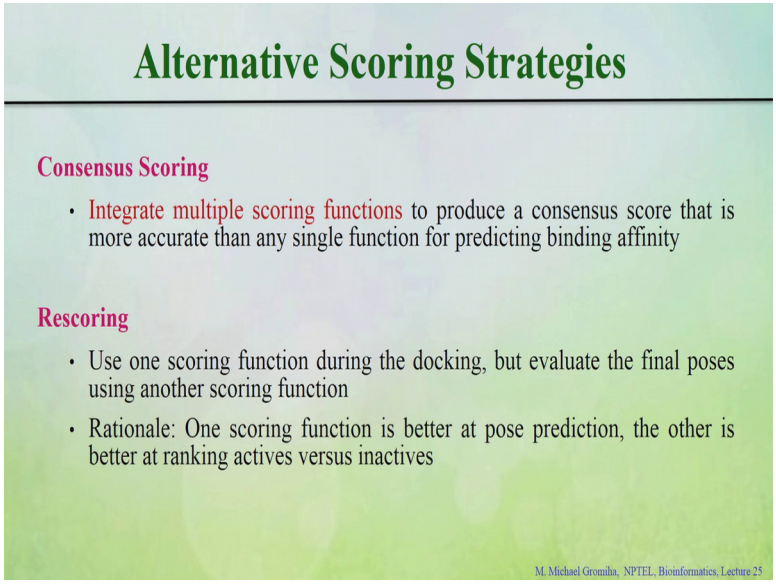


So, this is kind of empirical, but it is more general because it is more distance that the binding data, because we get more number of data because their binding energy data is not required we need the complexes that that that are available. So, you have plenty of complexes we can do that.

The disadvantages is they use a typically a pairwise one this is mean force, potential mean force, but the probability of the; A and B to be in contact mainly depends upon the surrounding atoms right. In this case difficult to map right all the contacts in the case of the protein ligand complex.

So, second one is the hypothesis originated from the statistics of the uniform liquid medium, but here this is totally non uniform medium right in the case of the protein ligand complexes right in the environment.

(Refer Slide Time: 28:28)



Then the next one is the consensus. So, now, we have a different scoring function,some scoring function performs good in some aspects.

So, they try to integrate multiple scoring functions to get the consensus score. Either they use all the scoring functions or they use the data from different scoring functions with different weightage right this will see there can be any they better than any single scoring function for predicting this binding affinity.

So, there is various scoring functions, we discussed what are the various scoring functions.

Student: shape complementary.

Shape complementarity and chemical complementarity.

Student: Empirical.

Empirical based scoring function.

Student: Force field.

Force field based scoring function

Student: Knowledge based.

Knowledge based scoring function.

Student: And the consensus.

And the consensus how the consensus works why the consensus is important?

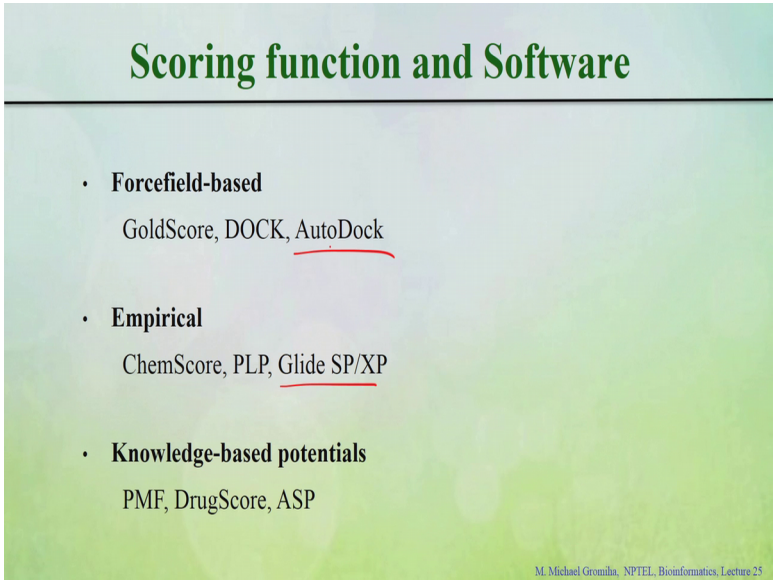
Student: It includes many.

Yeah because the rational is they think, take any scoring function that give you some data. But if you compare that could be better than at you are using any of the single ones.

So, in this case they try to use experimental data, and see different scoring functions and they merges scoring functions either they use all the scoring functions or take some information from different scoring functions and finally, they try to get the consensus one right in this case you can get the better scoring function right. So, that it can account the binding affinity of various protein ligand complexes.

Now, when using these such sampling space and the scoring functions right we can use to screen the compounds, that is called virtual screening right we will discuss in the next class right. So, get the compounds. So, there are various scoring functions and the software available in the literature right.

(Refer Slide Time: 30:18)



For example force field based software the GoldScore, the DOCK and AutoDock right autodock is the widely used public domain program that is available for free in the academic usage.

Then there is a empirical based fore field like the ChemScore or the PLP or the glide extra precision right or standard precision, and knowledge based potential like the PMF or DrugScore or asp and so on. So, several software available in the literature right and glide is the commercial one and the autodock is a for free right you can use any of these software right to get this docking with the protein with any ligand.

Then once we have the docking then if we have a set of compounds right for example, in the enamine there are two million compounds and the zinc 35 million compounds, you can use this information to pick up the probable ligands right which can be a potential lead compound for a, in the case of in the discovery.

So, in summarizing what did we discuss in this class.

Student: Computer aided drug discovery.

Right. So, what are the various aspects of this computer aided web design right. So, if you have the; what are the information required for this docking.

Student: Binding data.

We need the protein side as well as the ligand.

Student: Ligand.

Right. So, and then if you have this one right. So, you have two types of docking we discussed one is the rigid docking and the flexible docking, what difference between rigid docking and flexible docking.

Student: Both are rigid.

Yeah rigid docking of protein side is rigid ligand also we can make the conformation and then try to dock right. In the flexible docking it is proteins we have different conformations right in this case you can generate various conformations right which will take time.

Student: Flexible.

Flexible docking right it because you have search through various conformations right.

So, what are the difference two different major aspects in docking?

Student: search algorithm.

Search algorithm.

Student: scoring function.

And the scoring function right. So, what are the various search algorithms you can use?

Student: Systematic.

Systematic search and the stochastic search, what are difference between systematic and stochastic search?

Student: one is random.

Takes all systematic you have to do, but is time consuming right stochastic it can randomly sampled right and you can get the probable scoring right the pose.

Right then the energy functions what are the different types of energy functions?

Student: scoring functions.

So, scoring functions right different types scoring function based on the complementarity or the empirical or you can see the energy or the using the propensity values and the consensus ones right.

So, we have these sampling and then if you have the scoring function, then you can dock the compounds and if you have a set of compounds in library you can do the virtual screening right to identify the probable compounds as the potentially it compounds right.

In the next class I will discuss about the virtual screening of compounds right, if we have a set of compounds. How to identify the best ones right and to get the probable compounds which are for example, the lead compounds right and then we will see the how the QSAR models are also used right to identify the novel potential inhibitors in the next few classes.

Thanks for your kind attention.