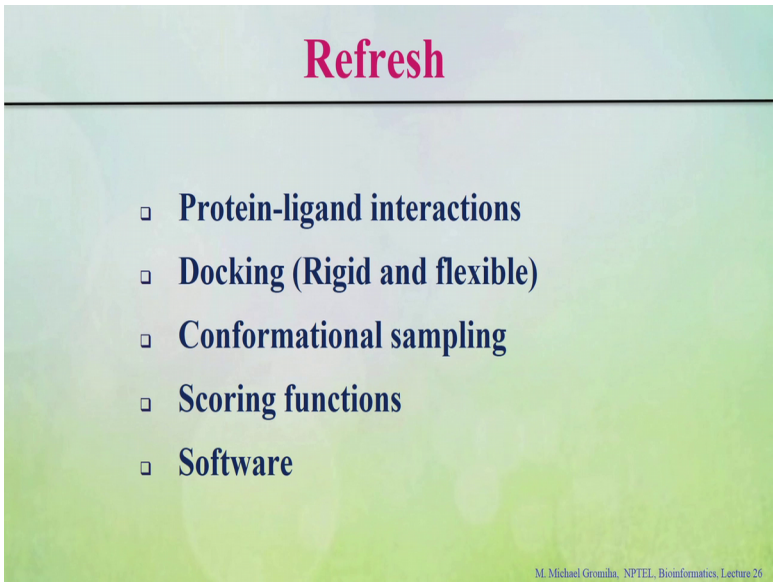
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 26a
Virtual screening I

In this lecture, we will discuss about the virtual screening of compounds to identify lead compounds from your pool of different ligands.

(Refer Slide Time: 00:29)



So, in the earlier lecture we discussed about protein ligand interactions, what is a ligand?

Student: a drug like small molecule.

Small molecule, it can trigger the activity of the proteins when they interact with the active site or the binding site. So, it could be a activator or the inhibitors or neurotransmitters etc. So, now when they interact with the protein, so you can identify different types of inhibitors for any drug targets. So, what are the different types of docking that we discuss?

Student: Rigid and flexible.

Rigid and flexible docking; what is the rigid docking?

Student: So, protein is rigid

You can see the binding site of protein as rigid, ligand also a one conformation you can dock. You can make different conformation for ligand, you can drag to the protein; we get the score, what is the flexible docking?

Student: There will be some.

So, here the proteins they are also flexible, so you can allow different degrees of freedom. In this case, you will get the different sampling; so for these sampling you can get this scoring function. So, you will find the best score to identify the target for any particular active site. So, to do the docking we need proteins mainly the active site and different poses, as well as the ligands, then you can add with types of docking. So, then we do the docking what are two different aspects you need to consider?

Student: Search algorithm.

One is a conformation sampling and second one is scoring function. What are different ways to sample the conformation of the protein will active sites?

Student: Systematic and stochastic.

Systematic or stochastic; in systematic you need to systematically carry out all the possible conformations; its time consuming, but you can get the data; it is good for the small compounds. So, if you have the stochastic one; so in this case randomly it can choose; the different types of algorithms we discussed; which algorithms we discussed for the sampling?

Student: simulated annealing

simulated annealing and the genetic algorithm; so, use different algorithm to sample the proteins conformation, then regarding scoring function; so what is scoring function?

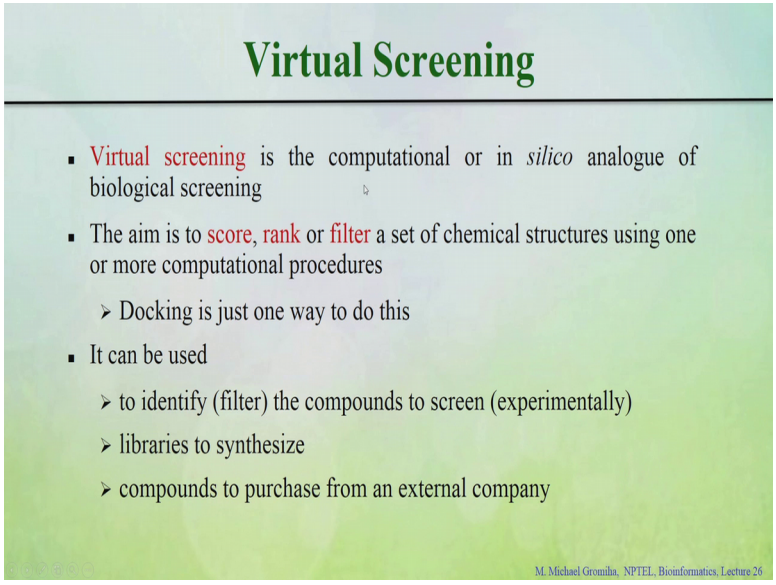
Student: In which binding strength.

This is a kind of energy function; so, energy function which can provide the binding affinity in terms of some values that called scores. This will tell you for a different pose of any protein, the active site which ligand can fit; the best with this particular pose.

So, different types of scoring functions you can decide which ligand is the best. For example, the shape or chemical complementarity; you can see empirical potentials energy, then force field. Then we discussed about the propensity values, we calculate the propensity then convert into potentials; knowledge based method and the consensus, you can see the use of various combination different potentials.

So, there are several software available in the literature; there is some of them are publicly available for example, of dock and the glide; this is available in the part of Schrodinger, you can use for the docking. So, now the question is docking is only way to identify which ligand can fit with this particular pose. So, we have a pool of compounds for example, 5 million compounds. How to identify the potential 8 compounds? This the technique called virtual screening.

(Refer Slide Time: 04:01)

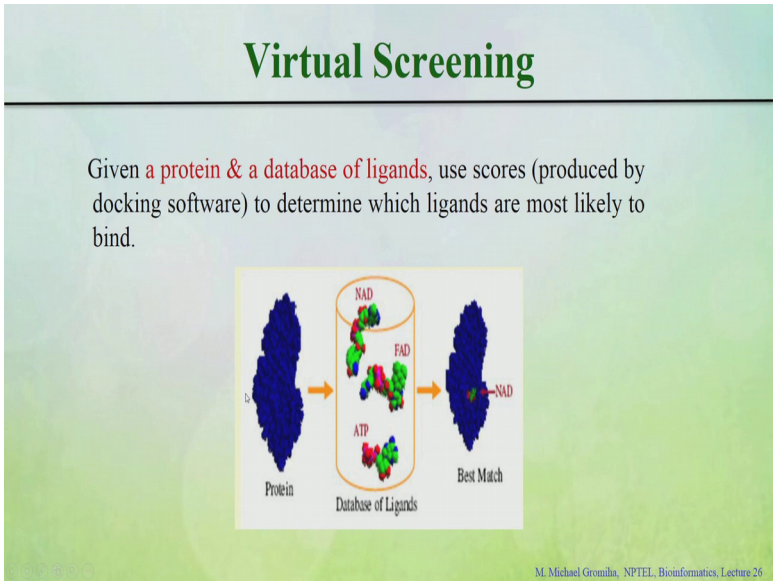


So, this is the computational or in silico algorithm for screening this compounds. So, what is the aim of the virtual screening? We have the proteins and the ligand this will give you score and this can rank these ligands and also filter a set of structures; whether this can fit or this is not fitting with any particular conformation in the protein site.

So, you can use docking to do this; so in this case you can fit it to the compounds. You can and get some libraries to synthesize and you can also suggest to purchase some components for the experiments. Then in case final experimental validation is required

even if you identify any lead compounds using virtual screening; the computational technique.

(Refer Slide Time: 04:51)

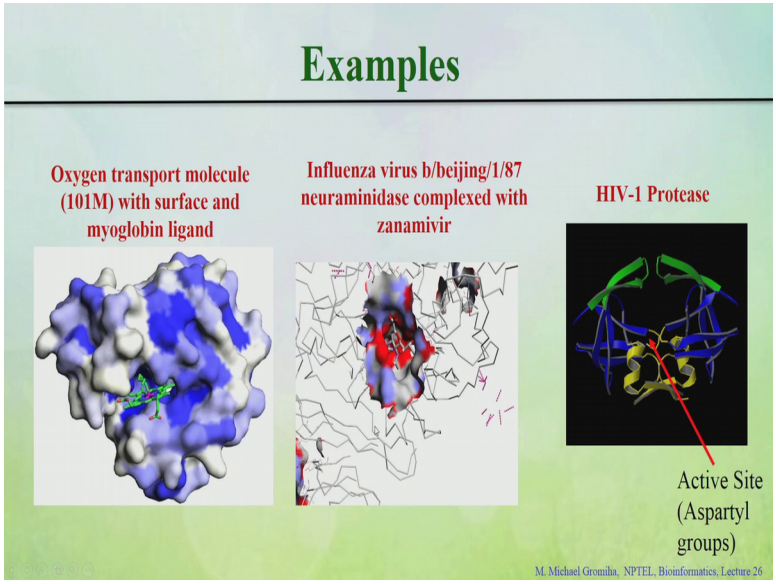


So, if I give the protein here; so where is active site probably? You can see it is fitting here. So, you can see the some cavities and here there are many ligands; you can say database of ligands. So, what the virtual screen does? It say given the protein and database of ligands; we can use several software. So, see from among all these ligands, which ligand will fit with this particular protein? In this case which one is the best fit?

Student: NAD.

NAD; you can see the NAD is found to be best match until for this combination for this protein from this list of ligands.

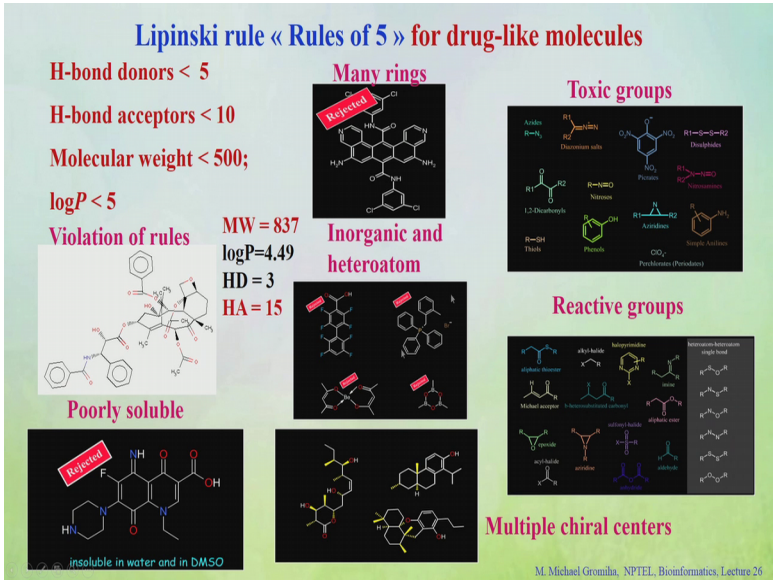
(Refer Slide Time: 05:31)



As a some more examples, this is a data for the oxygen transport molecule; the surface in the myglobin and the ligand. And here you can see influenza virus B, the neuraminidase complex with this zanamivir. This is zanamivir here; this is the protein site they how the fit.

This another HIV protease here also you can aspartyl groups, this can be acting as this active sites so that the ligands can go on interact at this site. So, several examples available in literature for the protein ligand complexes, then if you look into these available ligands; many structures you can directly reject based on some of these chemical structures; as well as the availability of the rings or the lipinski rule of five.

(Refer Slide Time: 06:19)



For example, I showed several rejected compounds based on various aspects for example, take this one.

So, here is a molecule weight is 837; hydrogen bond acceptor is 15. So, these two violated the lipinski rule of 5; what is the lipinski rule of 5?

Student: Molecular weight; it should be less than 500.

Less than 500.

Student: In hydrogen bond acceptor should be less than 10.

Less than 10.

Student: And donor should be less than 5.

Less than 5.

Student: And log.

And the partition coefficient this less than 5; so, here if you see compound number 1; this compound, molecular weight is 837 and hydrogen bond acceptor is 15. So, in this case you violated the two aspects; so, this may not be drug like molecule. Here one is rejected

mainly because it is insolvable in water because solvability is important because it is not soluble; in this case if it is drug; it is not soluble; in this case it rejected.

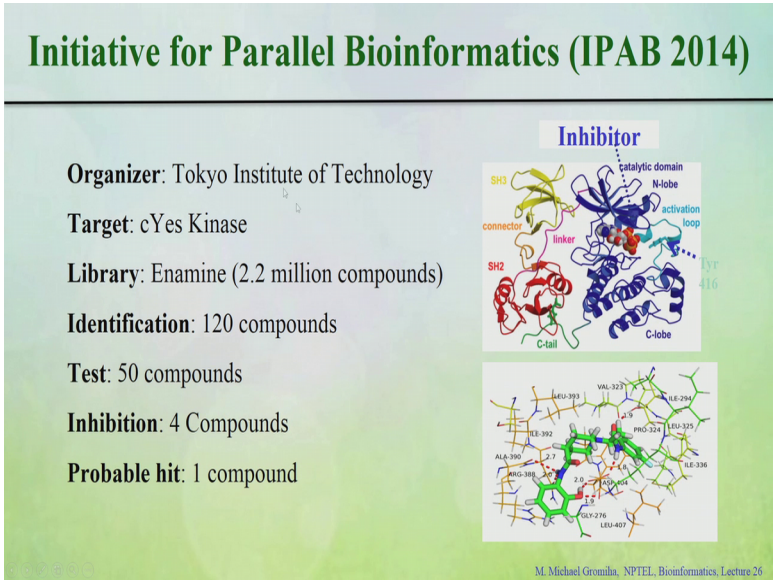
So, here this one is having many rings; so, because of that they rejected this compound. Here you can this are the depending up on the inorganic and the heteroatom. So, in this because of the region; they rejected this compound. So, here are several toxic groups and here you can see several reactive groups; some cases multiple chiral centers. So, these are the various aspect how we rejected some molecules; which may not be regulate molecule.

(Refer Slide Time: 07:43)

So, then what do we need in the virtual screening? Several set of compounds; so, several actives, some of them are inactives. So, we need to identify the actives and inactives and reject inactives and get the positives. Then we have to get the quality novel lead because get we should be quality; then this can be potentially it could be a drug.

So, you get reproduce a binding mode and predict the binding affinity and rank the diverse compounds. Then we get the hit rate by database mining; then you have to reduced false positives and false negatives. And then it should be fast to get this a structure based drug design. So, there is various steps you need to do for screening the compounds from the pool of compounds; to identify a lead like molecule; I show one example.

(Refer Slide Time: 08:36)



For here you can see the cyes kinase protein; this is involved in this colorectal cancer. So, the target is the cyes kinase; how to identify the inhibitors for this particular protein? So, Tokyo Institute of Technology in 2014; they organized a competition because there are several methods available for docking. For example, we discussed about various software; autodock and glide and many software; likewise there are hundreds of software available then different ways to get the probable compounds.

So, it is very important to identify whether these any method can potentially identify the probable compounds. Second question is; are there any mismatches or matches or over lapping of the different techniques; which can potentially identify same compounds. And whether we can get the diverse variety of compounds, if you have different methods and you get the diverse compounds; then you can reduce the pool of compounds to limited set and that we can go for the experimental validation. Based on that background, they organized a competition like the casp; what is casp?

Student: Critical assesment of

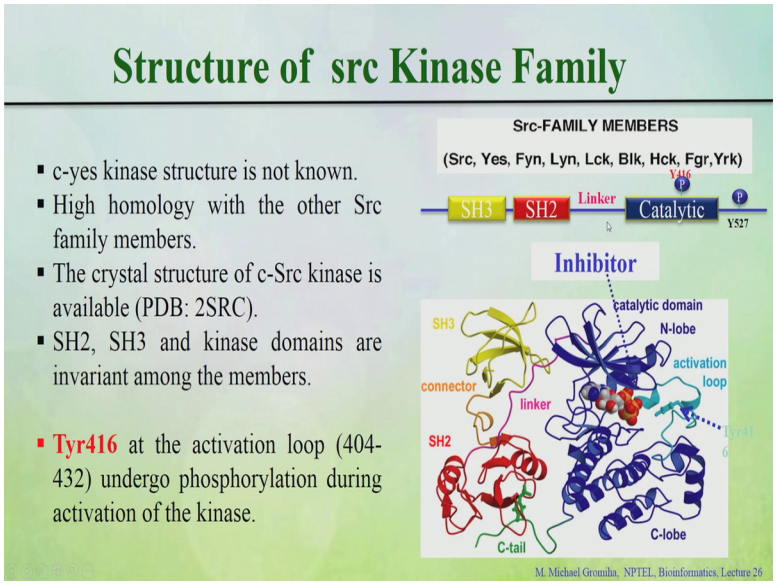
Critical assessment of?

Student: prediction of protein structures.

Protein prediction of; protein structures, so likewise here they listed of given of 2.2 million compounds from enamine. Why they used enamine compounds? Because there

all the compounds readily available, so do not have to synthesis and take time, then we can identify the inhibitors with known structures and once we submitted the compound then they will do experiments. And then compare whether any of these compounds are potentially inhibiting the activity. This is the major role for this initiative of parallel bioinformatics; so, I will explain.

(Refer Slide Time: 10:30)



So, what is c-yes kinase? You can see the c-yes kinase; this is the from the kinase family. So, here this is also involved in the colorectal cancer; so it is one of the major targets. It has several domains; for example, SH3 domain; it is 2 domains. These domains are similar to the other family members like Src; yes, fyn, lyn and so on.

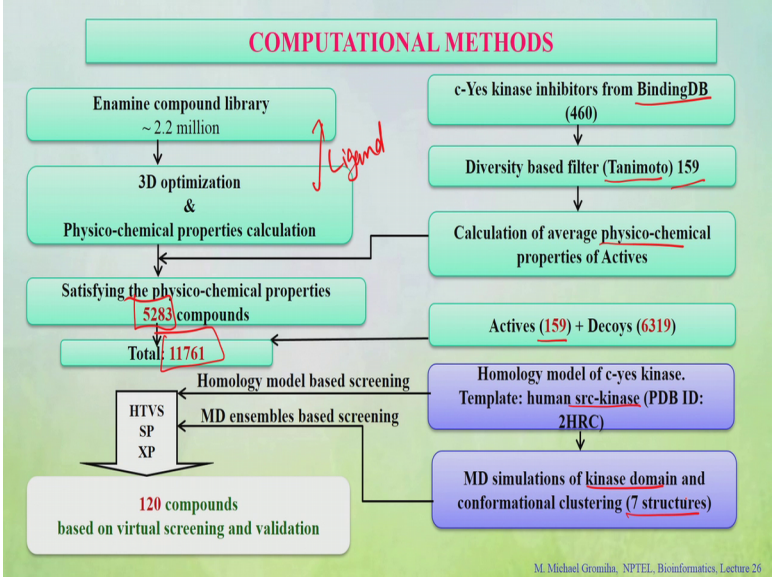
So, it is a catalytic site; here you have the two phosphorylation site, one is a Y 416; another one is Y 527; here you can see the SH3 domain, this is SH2 domain and here this is the linkage. And here is a activation loop; this loop, this part is very important for this inhibitor.

So, if we look in the literature; so crystal structure of cSrc kinase is known, but cyes kinase is not known. But the homology is very high with the other members; so tyrosine 416. So, it is in the activation loop this is the activation loop here, so here is the tyrosine 416. This undergo phosphorylation during the activation of kinase, then how to identify the probable inhibitors. So, to have docking, we need the protein as well as ligand; here the protein, what is the protein here?

Student: C-yes kinase.

C-yes kinase is a protein; so protein we know, but protein structure we do not know. Then the ligand; what is the library of ligand? Enamine library. So, in the last class when we discussed about the docking; before docking first we need to prepare the protein as well as prepare the ligand. So, now we have the enamine library; to prepare the ligands and you have the protein. So, protein also we have to prepare right, I will explain what we will do?

(Refer Slide Time: 12:18)



So, first we get the enamine library 2.2 million compounds; we need to optimize and do or check the all the parameters where it is a stable on the this fine or not. Once you optimize then you can calculate the physicochemical properties like various physicochemical properties. For example, molecular weight, hydrogen bond donors, acceptors and the molar refractivity.

So, various physicochemical properties; you can calculate these properties; so this is with the ligands. Then among these 2.2 million ligands; several ligands we can reject; because I discussed about the compounds which you can reject at the beginning.

So, we do not go for the screening and the final checking, so we can reject based on the some of these characteristic features. So, for that purpose what we did? So, we get these c-yes kinase inhibitors from this experimental data called the binding DB. This has some

data for the cyes kinase inhibitors; known inhibitors, but that is not available in this enamine library, if available in enamine library, they remove that one.

So, now all the 460 then some of them are similar, some of them are diverse. So, we need to have the diverse compounds; so get the diverse compounds using tanimoto coefficient; I will explain is soon. In this case we can reduce is 460 into 159 ligands; then among the 153, you calculate some of the physical chemical properties. This are called actives because already known the do not binding affinity; so they known as inhibitors.

So, now we compare this physicochemical properties and the physicochemical properties of this 2.2 million ligands, you compare with the deviation then we can remove several compounds finally, we ended up with only 5200 compounds. So, we can drastically reduced from 2.2 million to 5000 compounds. Then we added some decoys because to see whether decoys are correctly eliminated and the actives are correctly identified.

So, this 5000 compounds we added this 159 actives and 6000 decoys so that this will be 11761 compounds. Now the ligand side, we fixed this type of compound; at the protein side, so we know the target is cyes kinase, structure is not known. So, we can have the template called the Src kinase; with sequence identity of more than 90 percent. So, if it is more than 90 percent; what can we do?

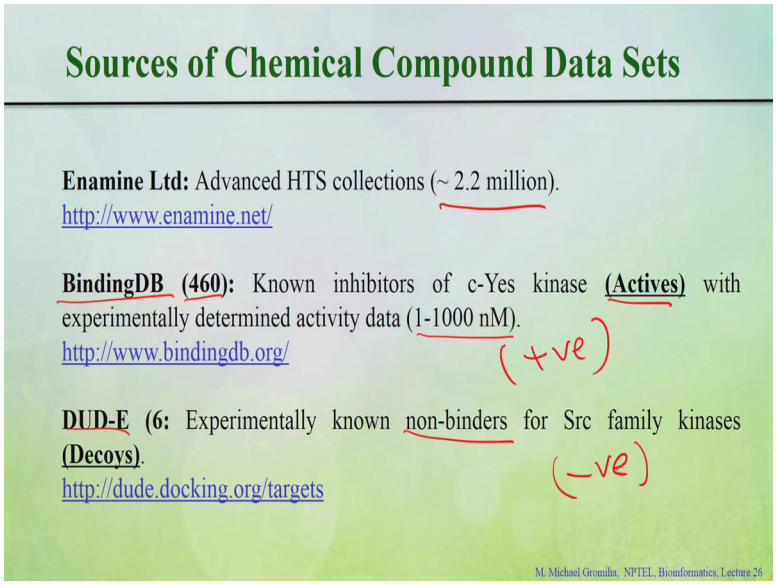
We can model using homology modelling; so you can homology model you can get the homology model.

So, now, we get the homology modeling of this c-yes kinase, but this way only one static structure. In this case we get only one conformation; so, protein also have different conformation not only active sites, but other place also they have different conformations. Accordingly the conformation of actives it also change; so in this case you can do MD simulations.

This simulation mainly these kinase domain and take the different conformations and cluster the conformations and finally, we get some 7 structures. This 7; one is homology model and one is the Src kinase; take this as proteins. Then with this ligands and these proteins you can do docking, we can now we significantly reduce the ligands. You can use the high throughput virtual screening or simple precession or the extra precession; you will do all these screening methods to get the final compounds say 120 compounds.

So, first we prepare the ligand and compare the ligand with the known actives and even we reduce the total number of ligands based on the physicochemical properties. They included actives and decoys for the test purpose; then we prepare the protein side using the homology modeling and the MD simulations and we make the docking and finally, we get the compounds and we can screen it.

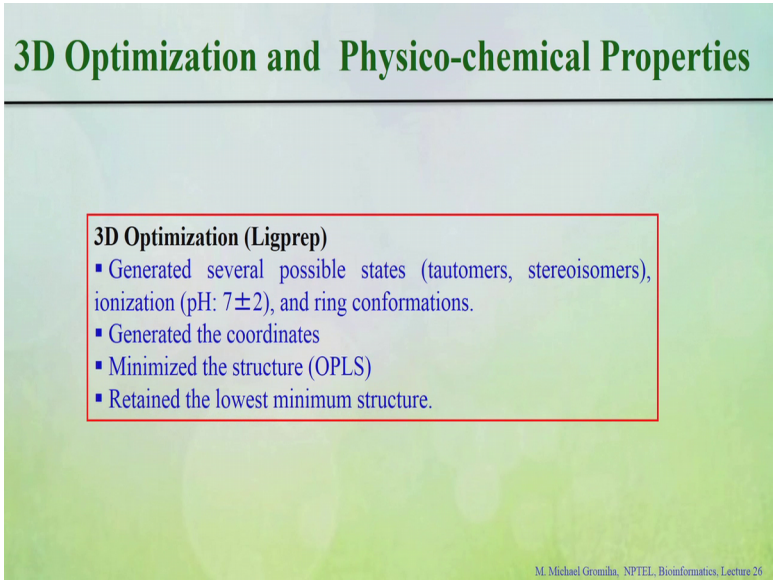
(Refer Slide Time: 16:26)



So, how to do this? First we get the ligands; so now I mean we 2.2 million compounds; then to see whether this compound show any similarity or how we can identify the inhibitors for this c-yes kinase; we checked the the binding DB. So, in this case we have the actives already reported like the 1 to 1000 nanomolar range. So, in the have consider 460 compounds.

Then DUD-E 6; this will give you the number of decays these are these are known non binders for Src family; these are negative database. So, we have the negative we have the positives, we can use this was a control and based on that we can identify the new inhibitors and you can just validate using experiments.

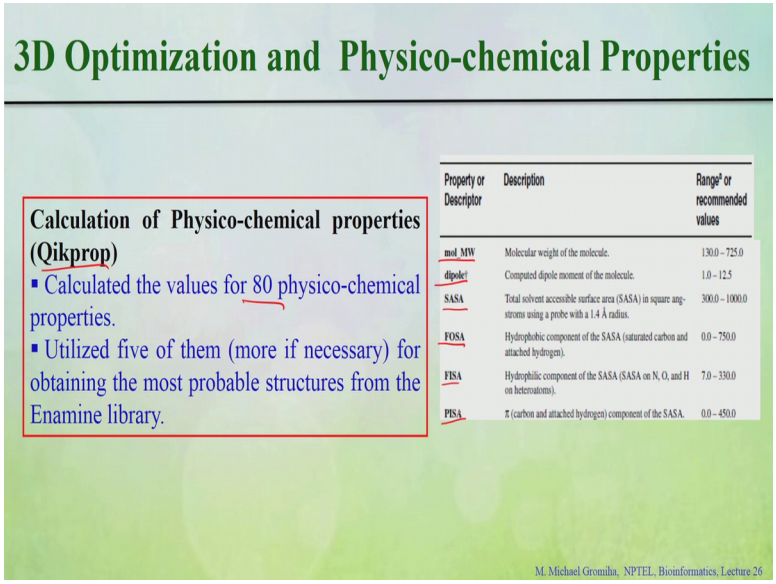
(Refer Slide Time: 17:08)



Now, first we get the ligands; first we do the optimization we discussed earlier like we need to the stereoisomers and the different tautomers, and you can do the ionization and so on; make all these things perfect. And we generate the coordinates using all these various possible states, we generate different coordinates.

Then you minimize the structure using the OPLS force field that is the kind of force field which calculate the energy. And finally, retained lowest minimize structure; for each ligand, we tried to get different conformations based on the possible states and we generate the coordinates and minimize the structure; we retained the minimized structure; likewise you can generate; prepare the ligands.

(Refer Slide Time: 17:50)



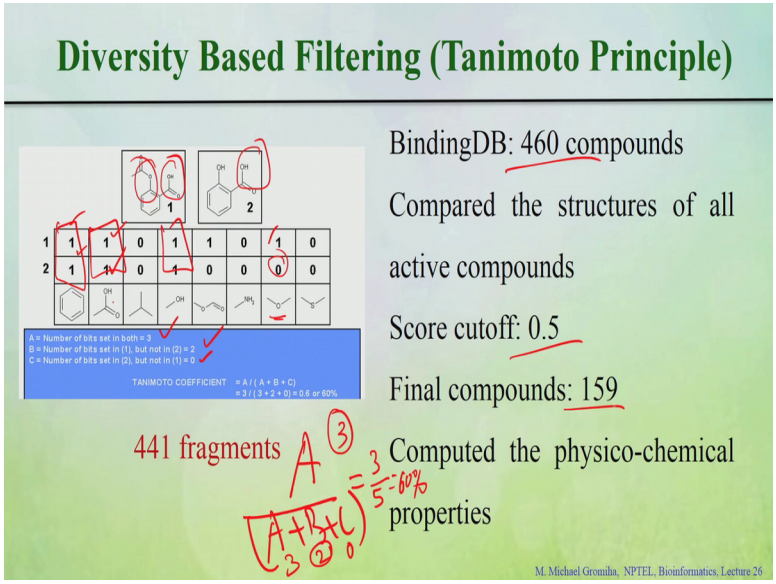
Once the ligands are prepared; then we can use the ligand for calculating the physicochemical properties. For example, in this program Qikprop; we calculate more than 80 physicochemical properties. For example, molecular weight, dipole moment of the molecule and accessible surface area, and the hydrophobic component of this area, hydrophilic component and pi component and so on.

So, we can calculate various properties more than 80 properties and among the different properties; we can select the best properties which can fit or which are important for binding. What are the various properties which are mainly important for binding?

Student: Shape complementary

Right you can see the hydrogen bonds, then you can see the log P, you can see molecular weight some properties are very important; along with some other properties. You can take a five or more; depending upon the properties which are important for the binding with the target.

(Refer Slide Time: 18:55)



So, now the filter the ligands for example, if you have the ligands 1 and 2; either you take both or you take only one. If you have the sequence; non redundant sequence structures, sequence identity how to get the non redundant sequences?

Student: Cluster.

Cluster we have the clustering; see similarity we take the similarity of 40 percent or 30 percent. Even they are similarly then maybe you can take one; you get not similarity to in both this you should non redundancy sequences you can make. Likewise the ligands; if you see there are different types of elements; say a ring, ring is present in a first compound; yes ring is present a second compound.

Student: Yes.

Yes then the second one like this OHO is present here?

Student: No.

Yes here; OHO the same.

Student: Yeah.

Right this is yes; this is here and this is here; this is yes. Then this one it is not available both then take this type of group; this O, this is here; here this is here yes, but here no; so you put 0.

So, you put 1 for any groups, you put 1; if it is present and if it is 0, if it is not present 1 and 2. Now you can see comparing this 1 and 2; number of bits set in both, how many you can see is available in both? 1.

Student: 1, 2, 3.

2.

Student: 3.

3; so B is a number of bits set in 1, but not in 2; how many cases present only in 1 and not in 2?

Student: 2.

2; how many bits set only in 2, but not in 1?

Student: 0.

0; this is a case you can calculate the tanimoto coefficient using the equation A by A plus B plus C. So, A equal to 3, B equal to 2, C equal to 0; so in this case this is equal to?

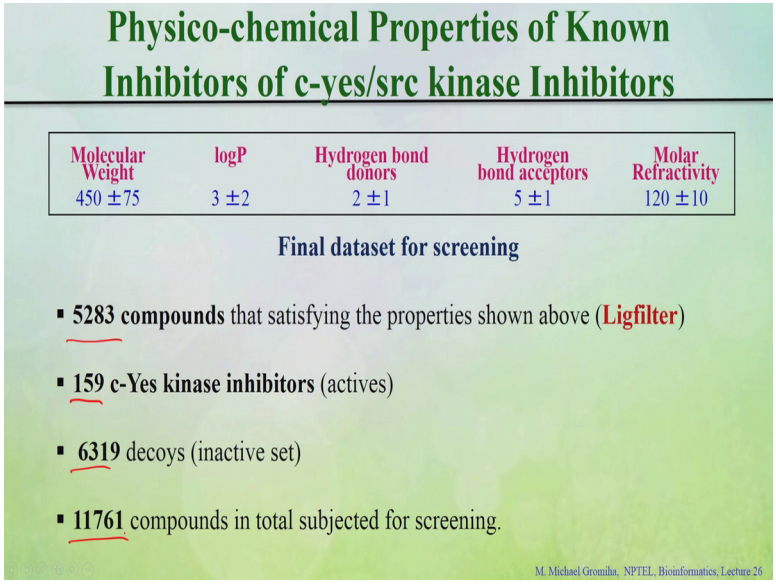
Student: 3 by 5.

3 by 5; 3 plus 2, this is 3 plus this is equal to 60 percent or 0.6.

So, they take the 460 compounds and compare these structures of all the active compounds using the tanimoto coefficient and say the cut off of 0.5. We will do this, you ended up with 159 compounds from this 460. Like the sequence analysis; structural analysis use RMSD; sequence analysis if we look the similarity of the similarity matrix and here also we use the tanimoto coefficient to get the non redundant structures.

So, from this 159 you can calculate the physicochemical properties; different properties you can calculate. For example, we take the five important ones molecular weight this is 450, plus or minus 75.

(Refer Slide Time: 21:55)



In this case; 375 to 525; log P 3 plus or minus 2; hydrogen bond donors 2 plus or minus 1; acceptors 5 plus or minus 1 and the molar refractivity 120 plus or minus 10.

So, we use both these ranges; minus and plus and the screen the initial 2.2 million compounds. Finally, we ended up with this 5000 plus compounds; then we have the actives 159; 6000 decoys. Finally, we get 11000 compounds that is fine, so then we need this is a ligand side that is fine; I said with the ligands.