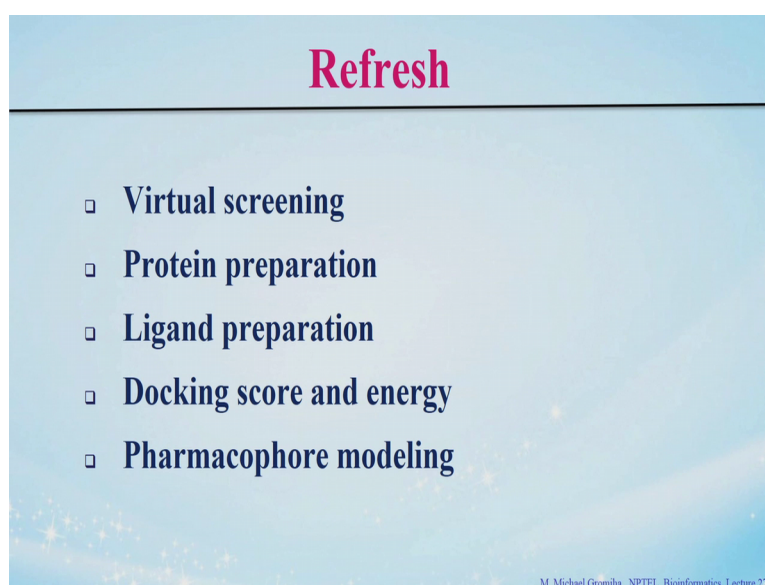


Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 27a
QSAR I

In this lecture we will discuss about the QSAR analysis that is quantitative structure activity relationship.

(Refer Slide Time: 00:23)



In the last class we discussed about virtual screening and various aspects, to screen a specific potential compounds from your pool of compounds right what did we do.

Student: select hit from library of ligands.

Yeah.

Student: identify hit in library of molecules

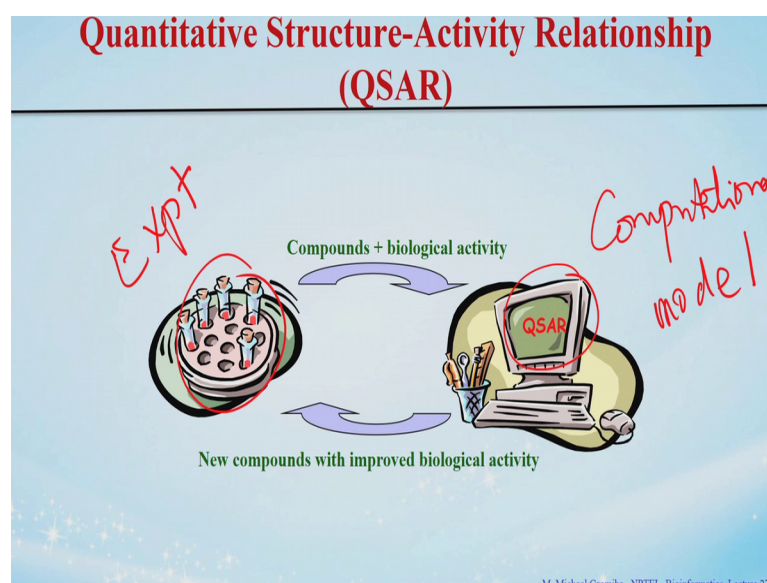
Right we have library of molecules right we identify the potential lead compounds right we will to say fit between the ligand as well as target. So, we discussed about the different types of docking, and then the virtual screening first we need to prepare the protein right under different aspects as well as the homology modeling if the structure is not available, and we simulated some structures for the various conformations.

And for the ligand we also analyzed the different types of ligands and made the charges and other optimization we did, and then we calculate the properties and then from the properties of the ligands with known activities right we compare the properties right and then we finally, we filter the compounds. Then we did the docking virtual screening of the compounds with the selected compounds along with actives and decoys, to see whether or model could correctly identify the actives and reject the decoys right. Then we from the docking score and the energy right we identify the potential lead compounds.

We also discussed about the pharmacophore modeling and this based on the conformation of the particular protein, why are the ligands could probably bind then we can identify the compounds, then this subject to experimental verifications, then we say that some of the compounds right they showed the good activation right the for example, inhibition of more than 60 percent.

So, now we discussed about the docking and screening. So, this is another approach to identify the potential lead compound is quantitative structure activity relationship. This is a kind of ligand based approach. So, here if you have a different ligands for a particular target, and if we know the activity then we use this information to identify some properties of these ligands to see which can relates the binding affinity or these IC₅₀ values.

(Refer Slide Time: 02:40)



So, first you get the data from the experiment, here this is the data from the experiment. So, we use this information right this binding affinity or this inhibitive constant, and see whether these experimental data right are related with any of the properties of the particular compounds, if this could be related with any compounds right. So, then we develop some models using computationally. So, here this is the computational model say fit the data, then we try to use the model right to identify new compounds and the new compounds again this go back to the experiments for verification.

So, the main principle or main aspects for the QSAR analysis is whether we can explain the biological activity of the compounds in terms of other properties of the ligands physical, chemical or any electronic properties and so on.

(Refer Slide Time: 03:45)

Quantitative Structure-Activity Relationship (QSAR)

- QSAR approach attempts to identify and quantify the physicochemical properties of a drug and to see whether any of these properties has an effect on the drug's biological activity by using a mathematical equation.
- To find consistent relationship between biological activity and molecular properties, so that these "rules" can be used to evaluate the activity of new compounds.

Biological activity = f (molecular or fragmental properties)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

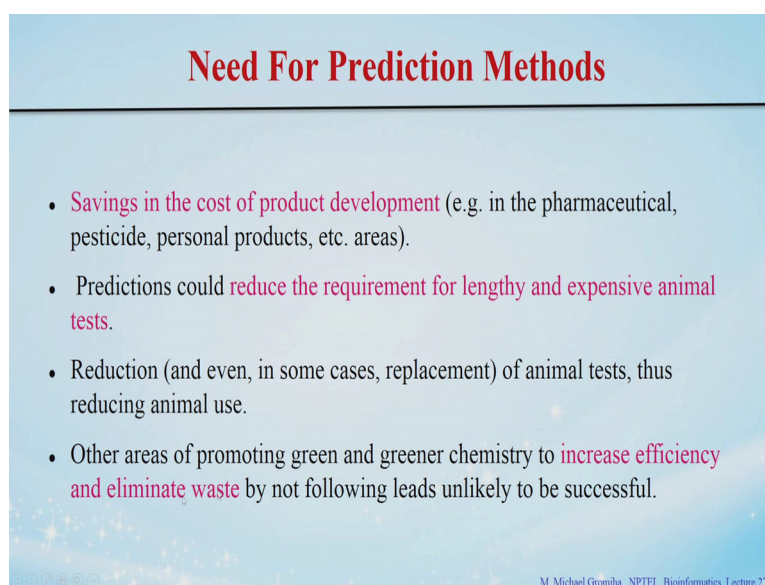
So, this QSAR approach right it try identify and quantify the physicochemical properties of a drug right. And whether these properties right can affect their drugs biological activity using any set of equations. With the linear equations or non-linear equations where we can relates these properties with the biological activity.

So, to identify the consistency of this activity and the properties we can set a kind of rules and evaluate the activity of these new compounds. With a new known compounds if you make some rules, then we can adopt these rules to identify the new compounds. This is essentially you can see the biological activity there is a function of the molecular or the fragmental properties. If you have the ligands either the molecule properties of that

a individual fragments and sum of everything, or the whole or say whole you can see there is a function of the different properties of the ligands can explain the biological activities, this is the principle mainly used in the QSAR studies right.

So, why do you need the QSAR analysis right as we know if you want to do any experiments, it is a time consuming involves lot of man power or also it is very expensive. Because every chemicals, we need to try by trial and error method and it is very expensive then it also require the animal tests right.

(Refer Slide Time: 05:07)



Need For Prediction Methods

- Savings in the cost of product development (e.g. in the pharmaceutical, pesticide, personal products, etc. areas).
- Predictions could reduce the requirement for lengthy and expensive animal tests.
- Reduction (and even, in some cases, replacement) of animal tests, thus reducing animal use.
- Other areas of promoting green and greener chemistry to increase efficiency and eliminate waste by not following leads unlikely to be successful.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, if you could identify the potential compounds from the two million compounds, and we identify ten thousand compounds, then we can reduce the number of experiments right. In this case can increases efficiency right of identifying these compounds right and also you can promote this green chemistry right avoid the experiment.

(Refer Slide Time: 05:33)

Need For Prediction Methods

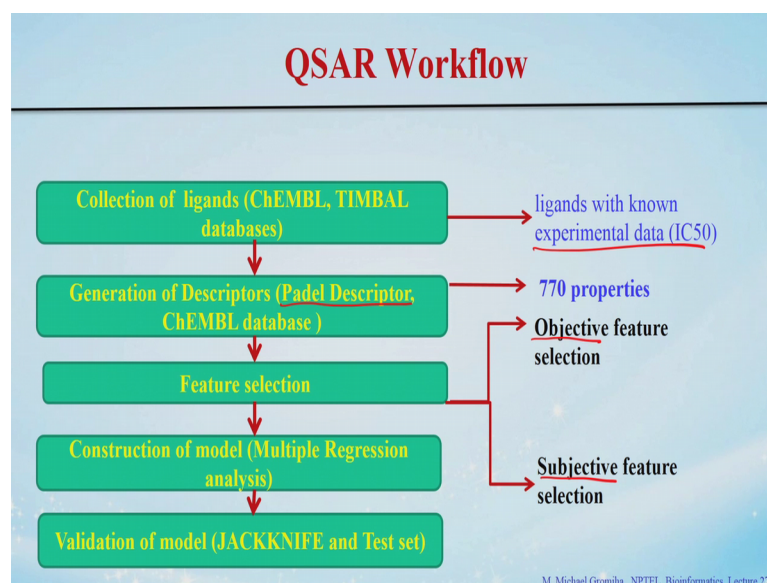
- ♦ The *in vitro* and *in vivo* potencies of a drug vary drastically due various factors:
 - Intrinsic reactivity of the drug,
 - Solubility in water,
 - Ability to pass the blood-brain barrier etc.,
- ♦ Differential behavior of a drug is attributed to electronic and steric properties of its structure.
- ♦ QSAR research is concerned with methods for quantifying differential behavior of drug correlating with its structure.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

There can if you see the *in vitro* and *in vivo* potencies of a drug it varies with different factors for example, the intrinsic reactivity and the solubility in water because its too soluble right and the ability to pass the blood brain barrier and so on right. So, differential behavior of these aspects, it also attributed with the electronic and steric properties of the structure.

For example the reactivity of the drug or the solubility or this ability to pass the blood brain barrier and lot of other properties that depends on the electronic and steric properties. In this case if you relate these properties with a function right in this case you can better make a develop a model, and this model will be helpful to narrow down the available ligand molecules.

(Refer Slide Time: 06:26)



So, this is the work flow for the QSAR. So, first what are the requirements of QSAR analysis, but the case of docking what are the requirements?

Student: Protein.

We need to protein and we need the.

Student: Ligand.

Ligand then we will we will dock that, but here first we need the collection of ligands, we can with known experimental data for example, I used the ChEMBL right or the Timbal databases, this will provide the experimental affinity data or the inhibition constant for several ligands right with respect to different targets. So, you can use any specific target, you can get all the ligands this for targeting a specific protein.

Is there are one side and the other side we need to have the descriptors right. There are various descriptors we can define for example, molecule weights, hydrogen bond donor acceptor right several features you can select. So, either you can use the known compounds and develop the features or there is one a program called paddle descriptor; this will provide if you give the compound, this will provide a set of features they put seven hundred to eight hundred features will give.

Then these so many features are repetitive. In this case you have to use feature selection method either objective features selection or the subjective feature selection. Subjective ones you can see there are several properties which are closely related to each other, in this case you can use correlation coefficient approach to discard the properties, if they are having a high correlation for example, say 0.9 or more than that right.

Likewise we can do the objective feature selection and if sometimes if we know some properties are important right. In this case we need to keep these properties for fitting the model, there is called the subjective feature selection. So, you can do both ways for selecting the important features to develop your model. And once we select the features then we construct model there are various ways to construct your model either the multiple regression analysis or you can developing non-linear models right to relate the experimental IC50 values right you can get from the databases plus these properties right then we need to validate the model right. So, using the jackknife test or you can use any blind test set to see whether this can predict reliably the activity of any new compound right.

So, I have discussed about some of the databases right I say mentioned earlier, there are several databases which contain the record of the ligands and some of them with affinity data or the inhibitory constant data.

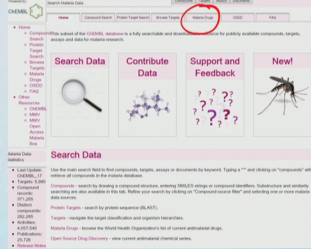
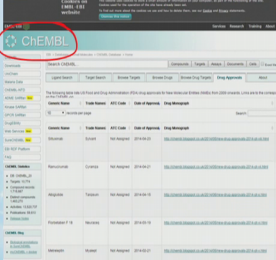
(Refer Slide Time: 09:15)

Databases

*Ligand
target
Drugs*

DB: ChEMBL_20
Targets: 10,774
Compound records: 1,715,667
Distinct compounds: 1,463,270
Activities: 13,520,737
Publications: 59,610

Malaria



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

For example the ChEMBL chembl it has more than 10,000 targets right and the target records more than one million target records right. Also they have the 13 million right the activity. They did the detailed literature search because the included about 60,000 publications right in the database. So, this is website have this a database chembl chembl right you can utilize these database and you have various options to get the data.

For example you can search with the ligand, search with the target right and also you can search with the drugs approved drugs that also you can search. So, you can search the data or you can contribute the data or also you can give the feedbacks, and also they give all the different information right from this particular database.

They are also focusing on some specific diseases for example, the malaria drugs right. So, specifically the drugs they used for this malaria. So, we can use this information to get these your ligands and which drugs and which compounds are used for different diseases or they identified as the drug related molecules or it can give the activity with respect to a target. So, you can get a set of data from a chembl database.

(Refer Slide Time: 10:45)

Databases

Tubulin

1-7 bands

ZINC database: 35 million

Enamine: 2.2 million

Chinese medicinal compounds: 45,000

ChEMBL: 1.72 million

Malaria: 371,000

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, is another one is called the timbal database, here also we can get the different ligands based on the target. For example, the tubulin I use right the tubulin search right. So, these are the various ligands right which are related with this tubulin. So, they interact with the tubulin. So, you can know the binding mode and the binding site as well as the accurate data for all these small molecules.

Likewise there is very shorter database available right sometimes they have these experimental data, sometimes they have the information regarding the small molecules; when the protein data bank you can see several complex structures right over the proteins with ligands. So, there you can get the mode of binding for all the ligands with respect to any specific protein, so that you can understand right how the ligands interact with the particular protein.

Some of the examples you can see Zinc database it is about 35 million compounds, now it is growing right. So, enamine 2.2 million compounds, and this database for the Chinese medicinal compounds about 45,000 structures and chembl is one point seven million and specifically and the malaria they have data for 371,000 ligands right.

So, we can utilize the information available in different databases right to collect the data of the ligands on various aspects; either the ligand structures or we can get different properties of the ligands or the experimental data on the binding affinity or the inhibitory constant or you can see binding mode right how the ligand interacts with the particular protein right.

So, using these information we have the ligands, you can derive various properties there is some of the important properties such we discussed right last class we will discuss about lipinski's rule of 5 what is lipinski's rule of 5?

Student: Molecular weight.

Molecular weight less than 500.

Student: less than 5 hydrogen bond donor

Hydrogen bond donor.

Student: Less than 5 hydrogen bond donor.

Less than 5.

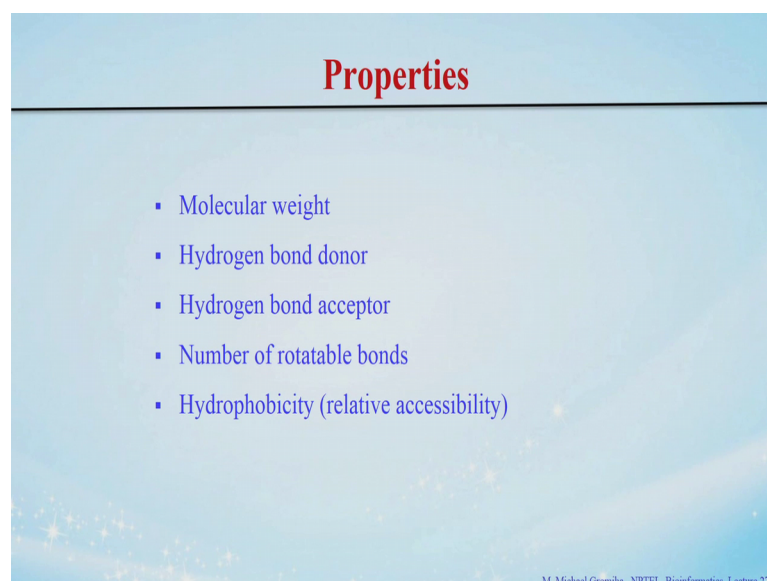
Student: Acceptor.

Hydrogen acceptor less than 10 and log p.

Student: less than 5

Less than 5 right the partition coefficient is less than 5.

(Refer Slide Time: 12:50)



So, likewise you can derive the various properties right for example, molecular weight hydrogen bond donors or acceptors, number of rotatable bonds right and the hydrophobicity, these are the important features which can relate the ligands with the protein. Because the case of the target protein, you can see different amino acid residues they are having different characteristics. So, they are like to perform different types of interactions right.

For example they can make the hydrogen bonds, if the donor is the protein side right then ligand side if have acceptor then they will high tendency to interact right. Numbers of rotatable bonds this can make various conformation, based on the conformation right the ligand how they can interact. Then the cavity depending upon the cavity size the ligand can bind or not this is a reason why we considered about the molecular weight right depending upon the cavity.

So, then hydrophobicity. with the hydrophobic groups rights in the protein side, then we requires the hydrophobic side chains right in the ligand side to have these interactions. So, likewise you can derive various properties, and you can make the selection based on subjective selection as well as the objective selection right. Based on the computationally you can calculate the similarities, and you can select or you can use the important properties subjectivity.

Then how to get the experimental data? They have various experimental techniques used to get the IC50, one is the functional antagonist assay.

(Refer Slide Time: 14:12)

Experimental Determination of IC50

Functional antagonist assay: IC₅₀ values can be calculated for a given antagonist by determining the concentration needed to inhibit half of the maximum biological response of the agonist from a dose response curve.

Competitive binding assay:

- A single concentration of radio-ligand (usually an agonist) is used in every assay tube.
- Determine specific binding of the radio-ligand in the presence of competing non-radioactive compounds (usually antagonists) for measuring the potency.
- Competition curves may also be computer-fitted to a logistic function as described under direct fit.

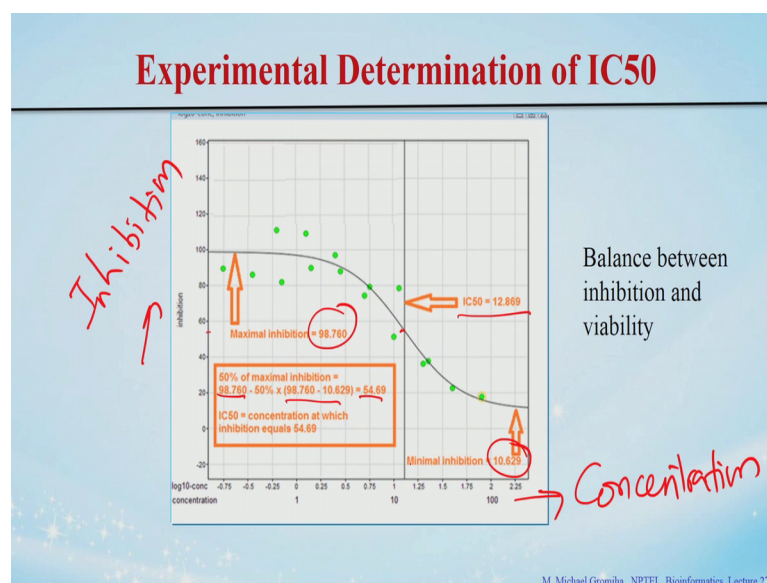
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, you can get the IC50 values right for any antagonist, by determining the some concentration needed to inhibit half of the maximum biological response right of the agonist from the dose response curve. Then get the dose response curve and then you can see how much the concentration is required to inhibit of the half the maximum biology response right you can do it.

Likewise there is another assays for example, competitive binding assay, here also you can see a single concentration of this any radio ligand usually an agonist in in every assay tube, then you determine the specific binding of the radio ligand in the presence of competing non-radioactive compounds usually antagonist, then there is a balance right for measuring the potency, then you can see the competition curves right there can be fitted using computer models like logistic function, to see the fit.

So, different ways to get the IC50 values right you can use the competitive binding assays or you can see the functional antagonist assays.

(Refer Slide Time: 15:16)



I will show you how to get this one. So, here this is a curve here x axis is the concentration, this is the concentration and y axis is the inhibition. So, if you see this is the 98 percent of maximal inhibition, you can see the curve like this is a usual curve we will get to for the inhibition this fit different concentration, then if you goes to the lowest one the minimal inhibition is 10.629.

So, then we get the 50 percent of maximal inhibition right why is IC50 we mentioned? Because we need to have a balance between the inhibition and the viability, this the reason we take the is 50 percent inhibition. So, we calculate the 50 percent of the inhibition. So, these is the maximum 98.76 right minus 50 percent of the difference how much they are the inhibition. The difference is 98 minus 10 this is the maximum 98, minimum is 10 there is a difference. So, from the maximum we get that is 50 percent of the difference this will give you 54.69 right the inhibition.

Then we go with this 54.69 percentage. So, is somewhere here right. So, if we look at here this is the place where we the touch the curve. So, this concentration this is the concentration this is about this is 10, this is equal to 12.869 right. So, this go we get that IC 50 values. So, for this compound the IC 50 value is 12.869. If we take the log value this will be one point something right.

So, here if we see relate to the concentration with the inhibition. So, you get the maximum concentration inhibition and the minimal inhibition, at various concentrations

and using this information we can get the 50 percent of maximum inhibition right. So, for the 50 percent what is the concentration related to 50 percent, that you can see this will be the IC₅₀ value for the particular compound. So, experimentally we know that using different assays right.

So, now we have these experimental values experimental data. So, we need to relate with these computational models right. So, in this case we can have various types of QSAR models based on dimensionality.

(Refer Slide Time: 17:33)

Classification Of QSAR Models

Based on dimensionality - Most often the QSAR methods are categorized into following classes, based on the structural representation or the way by which the descriptor values are derived:

- 1D-QSAR (correlating activity with global molecular properties like pKa, log P etc.)
- 2D-QSAR (correlating activity with structural patterns like connectivity indices etc., without taking into account the 3D-representation of these properties)
- 3D-QSAR (correlating activity with interaction fields surrounding the molecules)
- 4D-QSAR (additionally including ensemble of ligand configurations)
- 5D-QSAR (explicitly representing different induced-fit models)
- 6D-QSAR (further incorporating different solvation models)

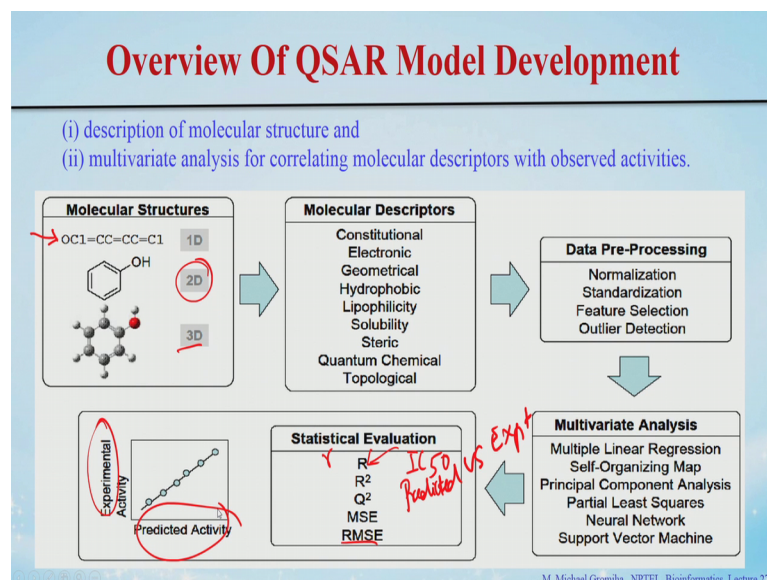
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

You can see 1D QSAR, 2D QSAR, 3D QSAR and so on right in the 1D QSAR right. So, we can correlate with some of the global properties of these molecules for example, you can see the pKa we can only one value right are the logP or different a properties. So, here you will get only one value since one dimension, this called the 1D QSAR when you go with the 2D QSAR we go with the one step further in this case you see the connectivity right. So, you see the patterns and how the interactions are connected.

So, you give you the connectivity this will give you 2D QSAR, when you consider the surrounding environment right for actually different types of interactions surrounding that particular position. Then you can go with a 3D QSAR then along with that you can see the ligand configuration or the induced fit models or you can use solvent models solvation models. So, accordingly you can increase the complexity right to build up the

different dimensionalities of this QSAR right mainly you can see this 2D QSAR, 3D QSAR commonly used in the developing the models along this 1D QSAR.

(Refer Slide Time: 18:49)



So, here you can see the how they develop the model right. So, for example, if you see a different molecular structures right here you can see this is the you can see the 1D right the first 1D QSAR model because we can get some properties using these specific compounds right based on the atoms present in the particular compound. Then second we see the connectivity. So, which residue which atom is connected which position. So, here these will get to 2D QSAR. Then if we go to 3D structure level so what are the neighboring current surrounding residues and how there are interactions made between these different atoms we give that information then we go the 3D QSAR.

So, various molecule descriptors right for example, electronic properties, hydrophobic properties, solubility parameters and steric parameters. So, there topological parameters right. So, various parameters we derive from this 1D information or 2D information or 3D information right. So, once we derive the descriptors, then we need to preprocess the data by the data set is fine or you need a discrepancy data set right or we need to standardize a data right then the normalize the data, if it is a wide range near the normalize the data then we need the feature selection.

So, and see any features are out layers and so on. So, do all these preprocessing right once we have the descriptors. Then we need to feed this in do some type of techniques

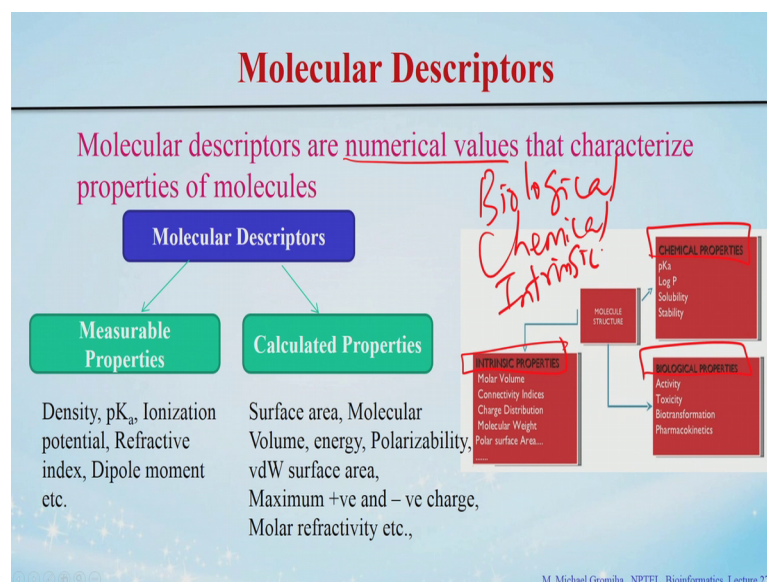
like statistical methods or machine learning techniques right for example, a multiple regression technique or support vector machine neural network right.

So, when we feed this in any of the methods, then you can evaluate right we have the values for the experimental data, and we have the QSAR model right you use any model to calculate the values, then you can evaluate whether the model could correctly predicts the values of the this any new compounds right the any IC50 or this any activity see and use r right this say correlation coefficient between the IC50 values right.

So if you have the IC50 predicted versus experimental right then you can calculate the correlation see how they can fit or you can calculate the mean square error and see how far the actual values right or the predicted values deviate from the actual values right you can see from that you can see whether this model fits well or not. This is the one example this is a almost perfect fit right this we do not expect this type of fit I always write, but if you are like enough then will you are able to get a good fit between the experimental as well as the predicted activities right.

Now, will explain the details how to do this.

(Refer Slide Time: 21:40)



So, we go for the molecular descriptors. So, what do you mean by molecular descriptors?

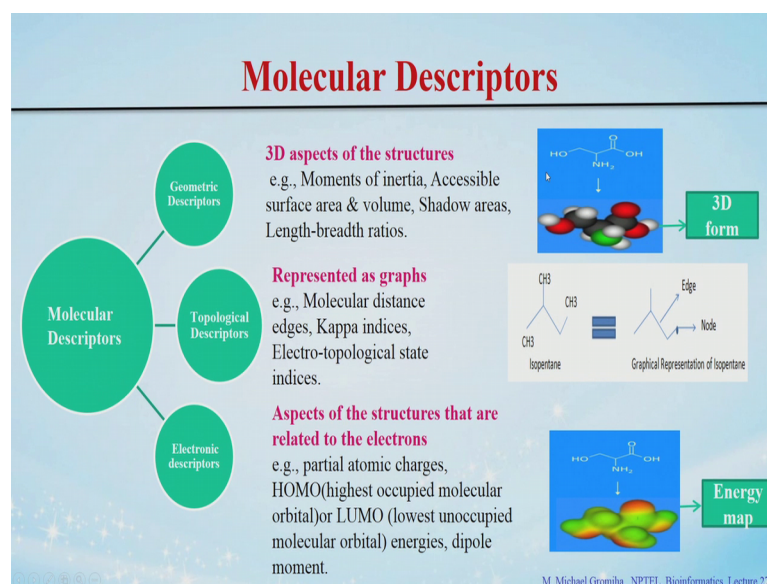
Student: Property.

These are the numerical values then property values numerical values that characterize the properties of the molecules right in various type of properties. For example, some properties you can measure right, density you can measure if you have the compound and pKa and the refractive index and we can see some properties of measurement properties. Some properties you can calculate for example, surface area, van der Waals surface area more are refractivity all these a properties you can calculate then there are different types of properties like chemical properties or the biological properties are the intrinsic properties.

For example chemical properties you can see pKa log P solubility and so on. In the biological properties toxicity activities and so on, and the intrinsic properties like charge distribution, polar surface area, molecular weight on this area. So, you can see intrinsic properties of any ligand chemical properties and the biological properties right. So, you can have the biological, chemical and intrinsic properties right.

So, then you can also get some of the calculated properties, and some of them you can have the measured properties.

(Refer Slide Time: 23:08)



Then we can see these descriptors in different groups, like you can say geometry descriptors or you can have topological descriptors or the electronic descriptors for example, if you see the 3D aspect such has the moment of inertia like we discussed earlier in the calculating the moment of inertia from the 3D structures right and also

accessible surface area right how to get accessible surface area? While rolling a water molecule instead how for the which atom is accessible right you can get this one you can get them for information.

Then also you can get some type of graphs right this is some of the nodes, where they have the different protein or the molecular interact and some adjust right. So, for example, this is the isopentane is here this is CH₃ CH₃ and the CH₂ this can graphically represented right.

So, you can see here this is a node, and you can see there are some edges right they can graphically represent. Then also you can get some electronic properties right because these properties are also important right to define the activity. So, you can get the energy map you can see some of them more green they are accessible and the red one right which are having the high energy system.

So, you can see the different types of molecular orbital energies, and type of movements you can calculate for any compounds right these are a you need to understand the concept to calculate I will explain some of the parameters how to calculate, and if you have this software you can automatically give you all the properties. If you give your a compound right in smile format or you can give you the structure formula right automatically you will give all the a information regarding a particular a ligand right.

So, now if we have these data and calculate the all the properties right, then to reliable QSAR models right we need to ensure the data are of high quality.

(Refer Slide Time: 25:02)

Data Pre-processing

- To obtain reliable QSAR models it is important to handle the data with great care.
- Data pre-processing: ensure the integrity of the data set before proceeding further with the analysis. It includes:
 - Data cleaning
 - Data transformation
 - Feature selection

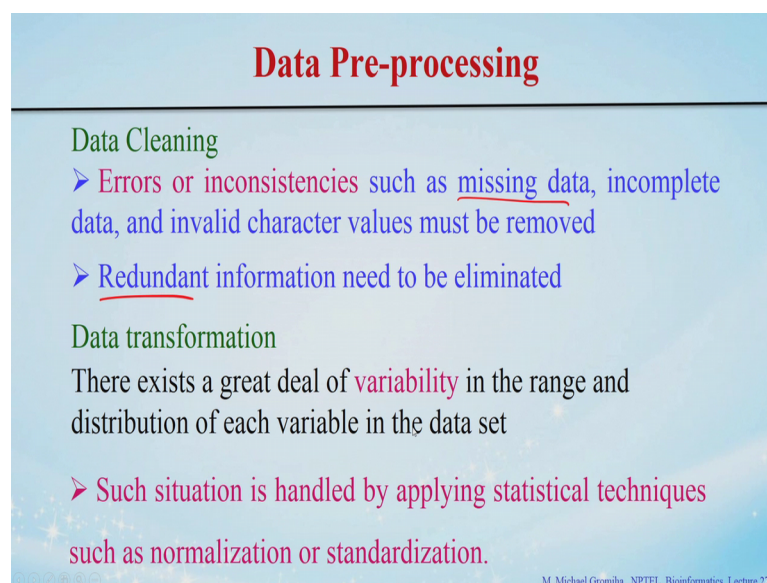
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, in this case we need to preprocess the data, to ensure that the data are high quality they are testable and there are no duplication of data right and all the data which resembles uniquely for any specific compounds, with respect to specific target right.

So, you need to check all the data this is data cleaning is very important right check all the data and see whether any data are missing are the compounds are a known with known structures right and the experimental data are obtained reliably right all the information you have to collect.

Then we need to change the transformation data transformation this is a different a properties right and then from these a properties you need to select the features right for each a molecule then how to get this data cleaning.

(Refer Slide Time: 25:53)



Data Pre-processing

Data Cleaning

- Errors or inconsistencies such as missing data, incomplete data, and invalid character values must be removed
- Redundant information need to be eliminated

Data transformation

There exists a great deal of **variability** in the range and distribution of each variable in the data set

- Such situation is handled by applying statistical techniques such as **normalization or standardization**.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Sometimes some error and inconsistencies, the missing data or the incomplete data, are sometimes invalid characters if available then you want to remove. In the second accept you need to remove the redundancy, if same information available right in the list of compounds, then we need to remove the redundancy right and eliminate the redundant want. Already we discussed about the removal redundancy in the case of protein sequences, and the protein structures likewise ligands also we discussed right what is the measure used to remove the redundancy in the ligands.

Student: tanimoto.

Tanimoto coefficient right yesterday we discuss. So, we have a different groups and for each group right we see whether this type of for example, ring. The ring is available in compound one and compound two right if it is both are available then put one. Likewise we do for the different parameters and we use it to remove the vibration right a by a plus b plus c right then if it is more than 50 percent right then you can see they are redundant likewise you need to remove the redundancy for any group of ligands.

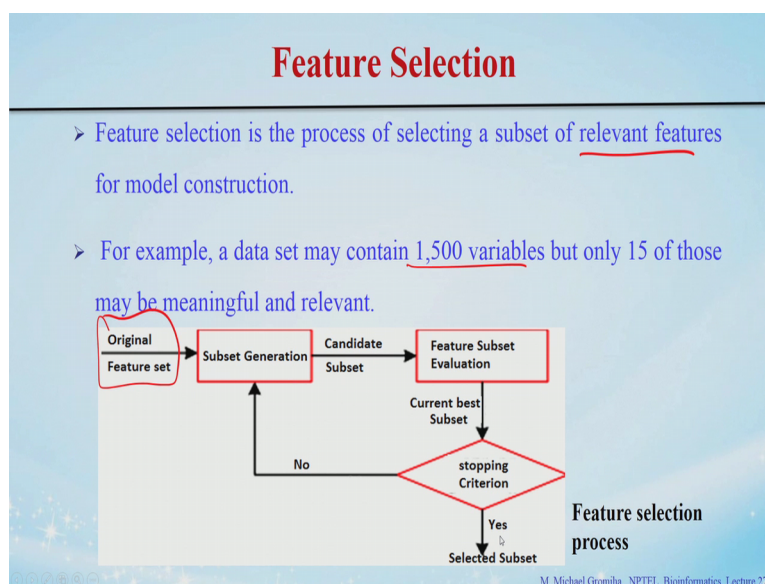
Then the second one we need to do the data transformation for example, whether the variability of data right there exist a great deal of variability in the range right and the distribution of each variable the data set right. We should not a remove the variable once you know to keep as much as possible the variables data, and whether our model could identify all these a ligands with different inhibitory constant right that is very important.

Then how to do this you can handle by using say some kind of normalization or the standardization. Sometimes you see the variables very high value deviation, then you can use the normalization are some data standardization procedures so that you can see a specific range right and then you can handle the data.

For example in the QSAR studies, most of the inhibitor constants right they range from mlili molar or micro molar or nano molar in this case we the wide range. So, take this logarithmic of the scale right in that case we can be a standardized values in specific range, and we can accommodate almost all the a ligands right in your dataset. Then the next step is feature selection this is very important, because what are the features you select in your model that will give you the reliability of your model because if you choose wrong features right then you can reproduce the data, but in you go with the real test set or the new data in this case your method will fail. So, you have to be careful about the feature selection, there should be at least explain the activity of your particular compound.

So, if we choose the feature right take some of the features, this is where we need the subjective selection right. If you need the hydrogen bond donor or hydrogen bond acceptor or the hydrophobicity right or molecular weight these features are important to explain the property of in ligand. So, we need to be very careful took the feature selection right what are the relevant features for the constructive model.

(Refer Slide Time: 29:02)



For example if you have a data set can 150 1500 variables right only if 15 of those meaningful relevant right. So, we need to take the only the relevant information. Here we have the original feature set here right then first we need the subset, from this original features we can get the subsets and from any candidate subset we evaluate the features right and if this is this explains well the your activity data then its go with these selected subset.

If not go back and they get that generate another subset and from the subset you collect the candidates, and then again evaluate whether this can be able to explain the activity and then if yes then go the selected ones or repeat again unless you are satisfied with the a features selections as well as the relationship with the experimental data. So, now, we can set the features right then we need to build a model.