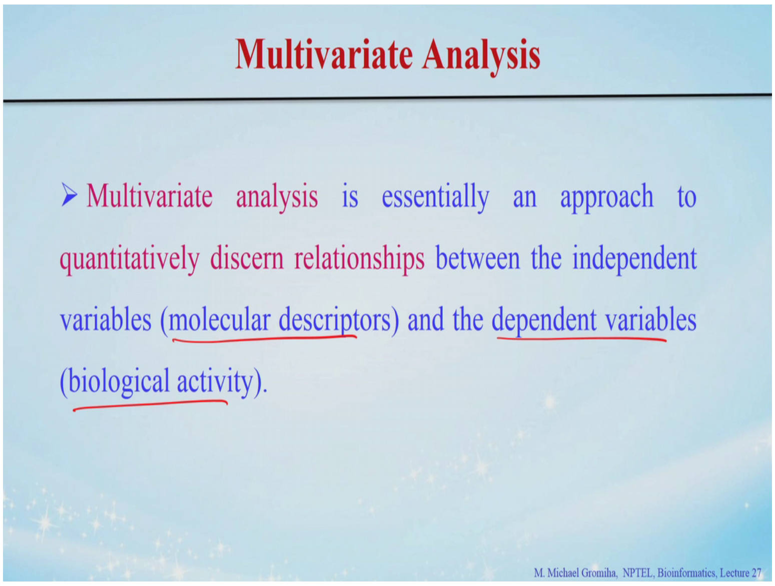
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 27b
QSAR II

(Refer Slide Time: 00:19)



Multivariate Analysis

- Multivariate analysis is essentially an approach to quantitatively discern relationships between the independent variables (molecular descriptors) and the dependent variables (biological activity).

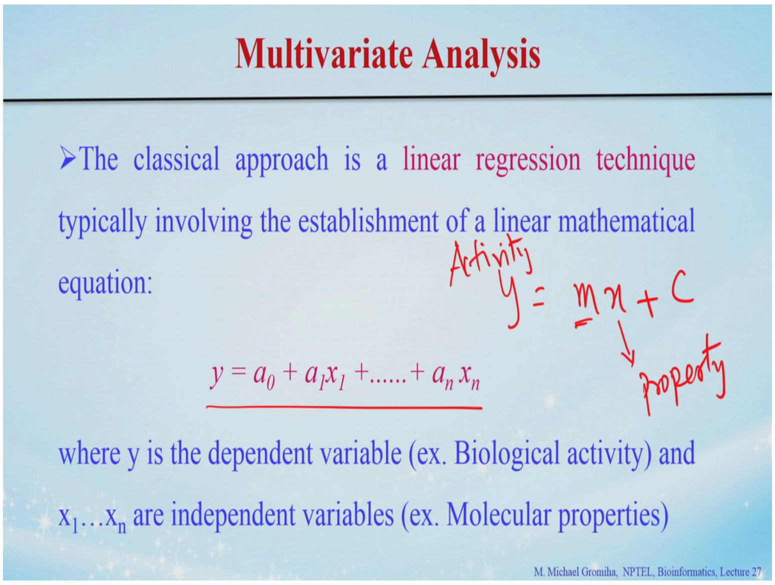
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So how to build a model? So, in this case mainly we can use the multivariate analysis to relate the structures and the molecular descriptors with the biological activity. So, here we have the independent variable as molecular descriptors and the dependent variable this is the biological activity.

Because we need to predict the activity here this is the different ones. So, then what are the other variables that we can choose you can change? But the biological activity, we cannot change because biological activity obtained from experimental data; so that is a known value, so that is a different one. So, the independent variables that you can change; how to select the variables and what are the variables which can better represent with the biological activity. So, then we can use different types of multivariate analysis to define

These different descriptors and how we can link these descriptors to explain the biological activity. So, the classical approach as we know is the linear regression technique for example, if you have only one variable.

(Refer Slide Time: 01:26)



Multivariate Analysis

- The classical approach is a linear regression technique typically involving the establishment of a linear mathematical equation:

$$y = a_0 + a_1x_1 + \ldots\ldots + a_nx_n$$

where y is the dependent variable (ex. Biological activity) and $x_1 \ldots x_n$ are independent variables (ex. Molecular properties)

Activity
y = mx + C
property

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, you can easily do the y equal to?

Student: mx plus.

mx plus c; so, what is y?

Student: Dependent variable.

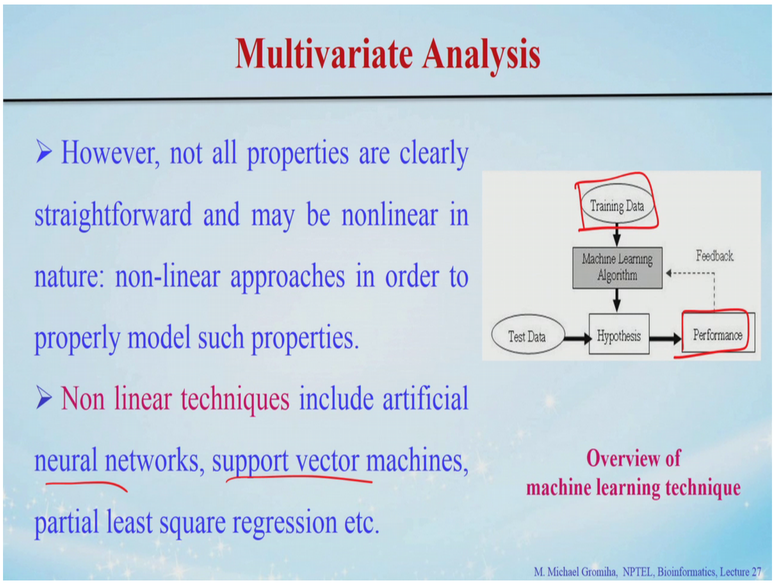
This is the biological activity. So, this m or this x? this is any property; so, what is m?

Student: slope.

So, m is the slope, c is the constant. So, you can if you there is only one variable; if you like to add more variables, in this case you can develop multiple regression technique y equal to a0 plus a1 x1 up to an, xn. So, for each compound we get one equation because we know the activity and we know all these properties, but only the coefficients we do not know; then we use principle of least squares.

We try to build this model; so, get the coefficients and then you use the coefficients to get the activity of a new compound; now that is you can easily do with this a multiple regression technique. Then sometimes is not, properties are not straightforward; so, you cannot fit with the linear regression. In this case you can also use some kind of non-linear techniques and you can also use the different types of machine learning, I will discuss later about the different types of machine learning.

(Refer Slide Time: 02:45)



Multivariate Analysis

- However, not all properties are clearly straightforward and may be nonlinear in nature: non-linear approaches in order to properly model such properties.
- Non linear techniques include artificial neural networks, support vector machines, partial least square regression etc.

Training Data
Machine Learning Algorithm
Feedback
Test Data
Hypothesis
Performance

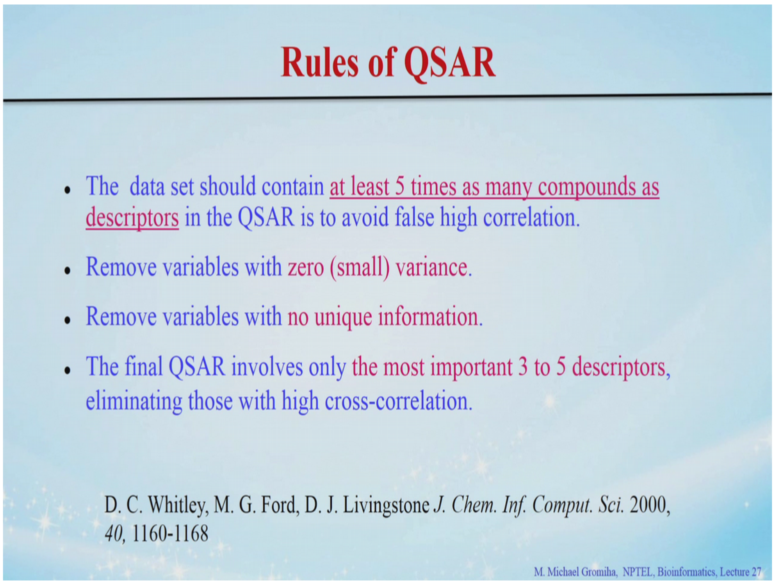
Overview of machine learning technique

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, for example, neural networks, support vector machines. So, we can have the training data here and here this is the machine learning. So, put in these machine learning techniques; so, this will try to fit the data for all the features plus the activity and then we will get the data. So, then evaluate the performance then see whether it is fine or not; otherwise based on the performance you can give the feedback for these performance obtained with these type of features.

Then machine learning will optimize could the get the optimal features so that you can get the best performance. We will discuss the machine learning techniques may be in the later classes. So, then what are the various rules of QSAR? For example, if you have a 100 data; can we use 50 variables to relate the activity? We cannot do it. So, are there any rules?

(Refer Slide Time: 03:42)



Rules of QSAR

- The data set should contain at least 5 times as many compounds as descriptors in the QSAR is to avoid false high correlation.
- Remove variables with zero (small) variance.
- Remove variables with no unique information.
- The final QSAR involves only the most important 3 to 5 descriptors, eliminating those with high cross-correlation.

D. C. Whitley, M. G. Ford, D. J. Livingstone *J. Chem. Inf. Comput. Sci.* 2000, *40*, 1160-1168

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, generally you can use the data set can contain at least 5 times as many compounds as their descriptors. Otherwise you can fit the data that will cause over fitting due to the inclusion of various number of descriptors. And this will work for the training set, but it will fail for the test set or any blind set. So, this case you have to maintain that the descriptors or the features should be as less as possible; compared with the experimental data available that to use to do for developing the model.

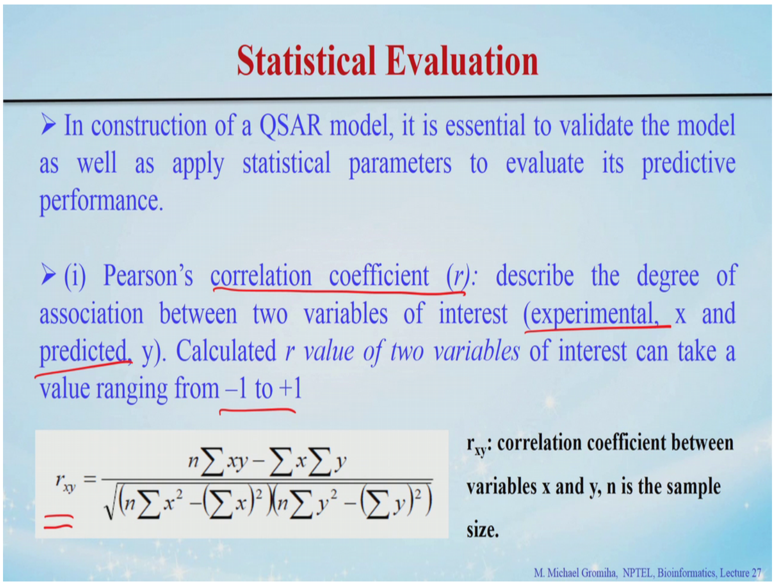
Then you have to remove the variables with the 0 variance and also you have to remove the variables with no unique information. If the variable sometimes we have some features that has no information or no relationship with the activities. So, in this case you need to remove that variable because it is not possible to explain why the specific feature is selected to relate the biological activity.

So, if you look into this QSAR models in most of the case if you see the final model contains the most important 3 to 5 descriptors, you can have a large number of descriptors, make a subsets and train then again take the best features make a small subset; finally, if you end up with the 3 to 5 descriptors eliminating with high correlation among these properties; then you can see whether a model can explained well with the biological activity.

So, there are several rules for the QSAR when you are developing any specific models. Then how to evaluate? How to evaluate the QSAR model? So, if you develop a model;

so you need to evaluate the performance of the model. So, in this QSAR here there are various ways to analyze or evaluate the performance; one is the correlation coefficient.

(Refer Slide Time: 05:28)



Statistical Evaluation

- In construction of a QSAR model, it is essential to validate the model as well as apply statistical parameters to evaluate its predictive performance.

- (i) Pearson's correlation coefficient (*r*): describe the degree of association between two variables of interest (experimental, x and predicted, y). Calculated *r value of two variables* of interest can take a value ranging from –1 to +1

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

r_{xy}: correlation coefficient between variables x and y, n is the sample size.

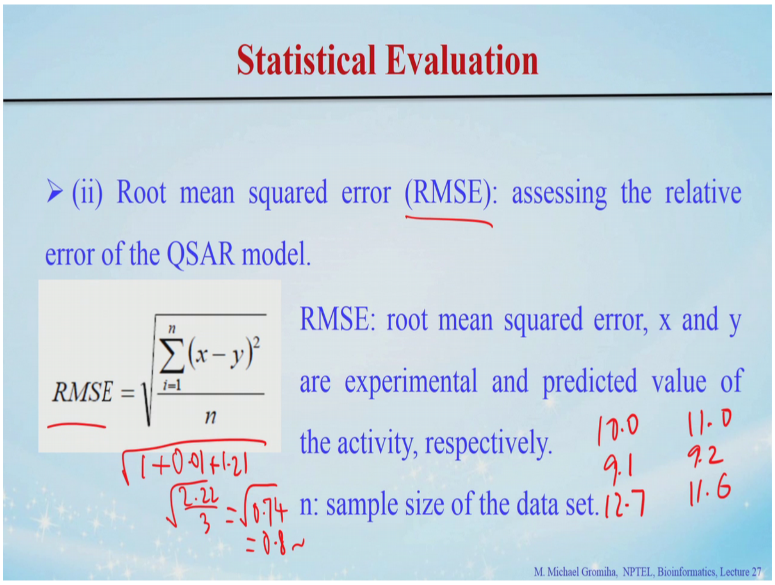
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

This will tell you how far the two variables are related with the each other for example, one is x experimental and y is a predicted; then you can use this formula to see whether they are related with the each other; what is normally the r value; the range of r values?

Student: Minus 1to plus 1.

Ranges from minus 1 to plus 1, if it is minus 1 they are inversely, related and plus 1 it is positively related and if it is 0; it is not related; then depending upon the numbers for example, 0.7 or 0.8; we can check how far the predicted values are related with these experimental values. This will tell you whether your model is correct or not; how far you can rely your new model?

(Refer Slide Time: 06:15)



Statistical Evaluation

➢ (ii) Root mean squared error (RMSE): assessing the relative error of the QSAR model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x-y)^2}{n}}$$

$$\sqrt{1+0.01+1.21}$$

$$\sqrt{\frac{2.22}{3}} = \sqrt{0.74} = 0.8\sim$$

RMSE: root mean squared error, x and y are experimental and predicted value of the activity, respectively.

n: sample size of the data set.

10.0	11.0
9.1	9.2
12.7	11.6

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Then you can see with the root mean square error; how to get root mean square error? You can see the predicted one experimental one for example, the predicted values are 10.0 and 11.0 and the experimental is 9.1 and it is a 9.2 and this is a 12.7 and here it is 11.6; so how to get this root mean square deviation?

Student: Calculate the error.

Calculate the; what is a error for the first one? This is a 10 minus 1; this is 1 squared; 1 plus second one is 0.1; so, 0.01 and the third one 1.1; 1.21 so, add up 2.22; divided by 3.

Student: 3.

This equal to 0.74; so this equal to 0.9, 0.8.

Student: 0.8.

Yeah.

Student: 0.87.

So, you can see if this is because this is 1; this is the 1.1; so you can see a balance among all the data which used in your model. So, from this you can calculate the RMSE and you can have a cut off values; for example, it is less than 1 or less than 0.5; depending upon your data size or depending upon the values actual values. For example, the actual values around 10, 9 with this 0.8; then how many percentage of deviation? 10 percent.

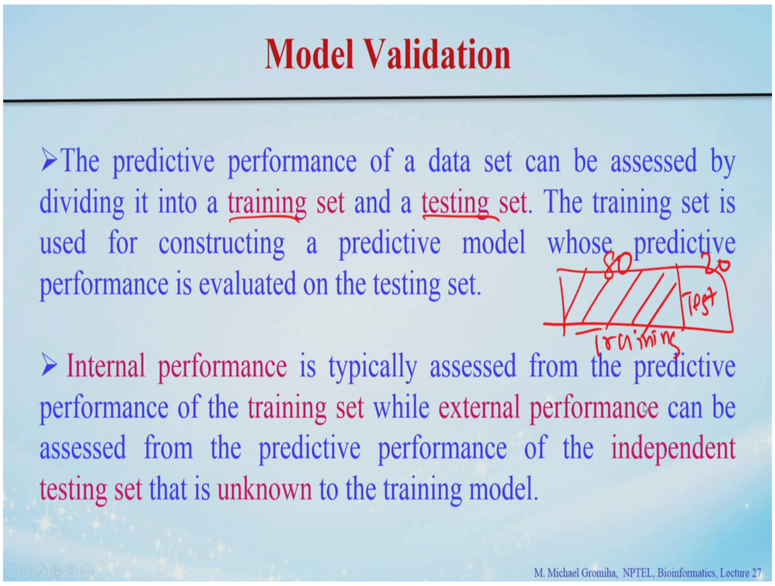
Student: 8 percent.

Or 8 percent deviation; so, you can see for that is in this case we need to be careful because sometimes these values are less; then whatever value you get this deviation then if you take the percent that will be high. So, in this case we need to use different measures to see whether your predicted values are related with this experimental data.

So, you can see n sample size and here x and y are the predicted experimental values. So, you can get the root mean square error; this will tell you how far your method could fits well with this experimental data. Then how to validate the performance? You can use various ways to validate the performance.

Either you can divide the dataset into training set and the testset.

(Refer Slide Time: 08:53)



Model Validation

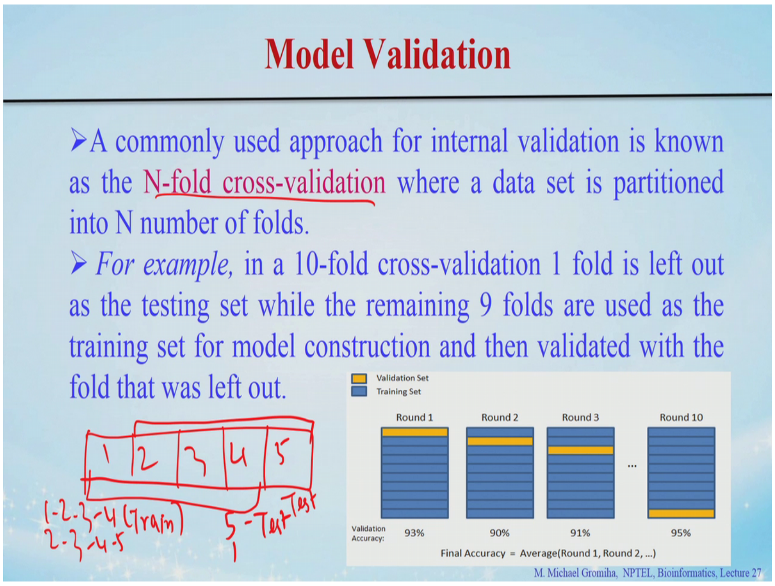
- The predictive performance of a data set can be assessed by dividing it into a training set and a testing set. The training set is used for constructing a predictive model whose predictive performance is evaluated on the testing set.

- Internal performance is typically assessed from the predictive performance of the training set while external performance can be assessed from the predictive performance of the independent testing set that is unknown to the training model.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

If you have the different data, you can use this as training and these are the test set. Then you can use the data in the training to develop your model and you can use the values to test the new data. Whether this can be able to predict or not, so you can see these numbers for example, it can use various percentages. For example, you can take 80 percent here or 20 percent here or 90; 10 or 70; 30 you can use a various proportions and see whether even less number of training, you are able to capture the features and whether it is possible to predict the remaining cases with the high reliability and this is this you can do that. So, then you can see this is the then also you can do with the external test and evaluate whether your model works fine or not.

(Refer Slide Time: 09:53)



Model Validation

- A commonly used approach for internal validation is known as the N-fold cross-validation where a data set is partitioned into N number of folds.
- *For example,* in a 10-fold cross-validation 1 fold is left out as the testing set while the remaining 9 folds are used as the training set for model construction and then validated with the fold that was left out.

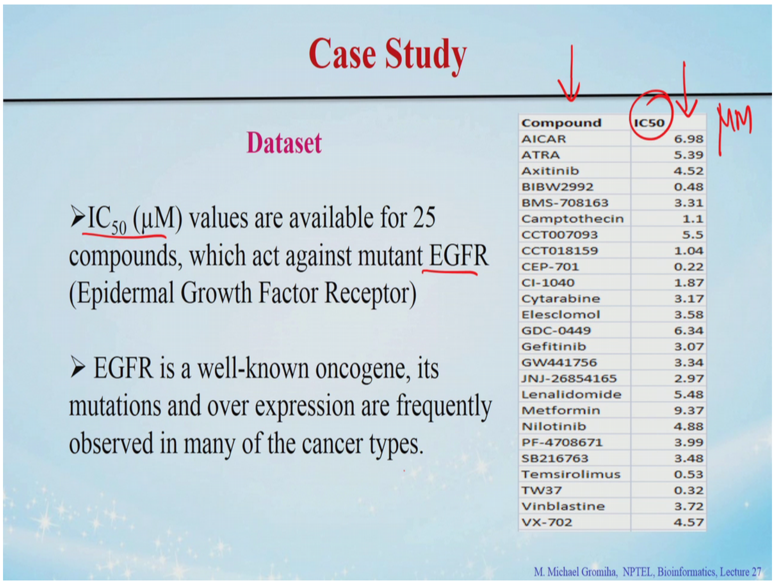
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

This is the commonly used cross validation this called the N fold cross validation. In this case, we classify the data set in the N groups and take the N minus 1 group as the training and the left out for the test. So, in this case if it is the 10-fold cross validation; so we can make into 10 groups or if you have 5-fold cross validation for example, 1, 2, 3, 4 5.

So, we use this for training and this for test and the second case you use this for training; first case is 1, 2, 3, 4 training and 5 as test. The second case, we take 2, 3, 4, 5 for training and one as test; likewise how many times you need to repeat? 5 times you need to repeat and then take the average this will give you; so, how far your method that could predict. Then we need to sample resample again and take another 5-fold cross validation. So, the samples will be different and see how can you do that; so, repeat several times and almost all the time.

We get similar level of performance then you can see that your method is reliable and you can use your model for predicting any of these compounds; now I show the some case studies I show you.

(Refer Slide Time: 11:18)



Case Study

Dataset

- ➢ IC_{50} (μM) values are available for 25 compounds, which act against mutant EGFR (Epidermal Growth Factor Receptor)
- ➢ EGFR is a well-known oncogene, its mutations and over expression are frequently observed in many of the cancer types.

Compound	IC50
AICAR	6.98
ATRA	5.39
Axitinib	4.52
BIBW2992	0.48
BMS-708163	3.31
Camptothecin	1.1
CCT007093	5.5
CCT018159	1.04
CEP-701	0.22
CI-1040	1.87
Cytarabine	3.17
Elesclomol	3.58
GDC-0449	6.34
Gefitinib	3.07
GW441756	3.34
JNJ-26854165	2.97
Lenalidomide	5.48
Metformin	9.37
Nilotinib	4.88
PF-4708671	3.99
SB216763	3.48
Temsirolimus	0.53
TW37	0.32
Vinblastine	3.72
VX-702	4.57

μM

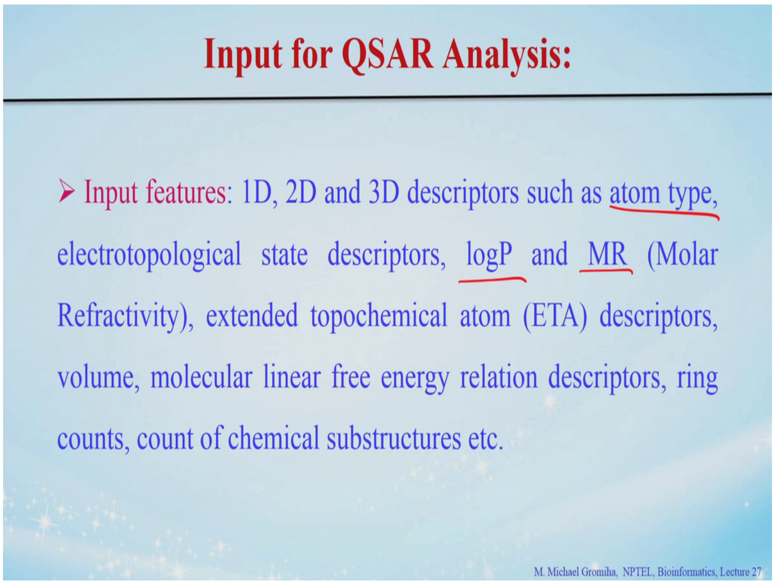
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, here we have the different compounds; these are the different 25 compounds IC50 value are known in micromolar; it is against the epidermal growth factor receptor. This is a well known oncogene, and its mutations and over expression are frequently observed in many of the cancer types; this is very important.

So, it is very important to design different types of inhibitors. So, we have the values for the 25 compounds against EGFR. Now these are the 25 compounds and this is IC50 value in micromolar, now can we predict these IC50 values using any properties?

Or if you have the new compounds because many compounds available in the literature; can you identify any better compound other than the one which reported in the literature.

(Refer Slide Time: 12:05)



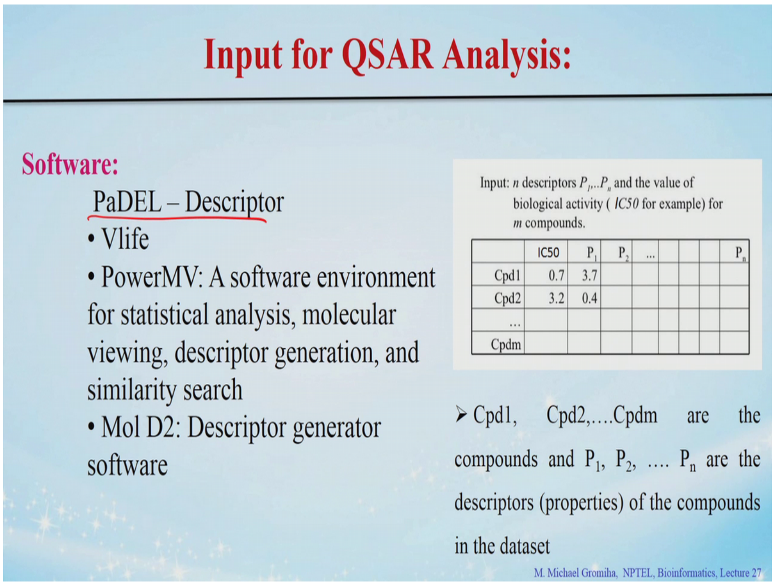
Input for QSAR Analysis:

- Input features: 1D, 2D and 3D descriptors such as atom type, electrotopological state descriptors, logP and MR (Molar Refractivity), extended topochemical atom (ETA) descriptors, volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures etc.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

First we see that; so, we have to derive the features you can see the different types of QSAR's that is 1D descriptor, 2D descriptor, 3D descriptors of the atom type logP molecular refractivity; different types of parameters you can derive.

(Refer Slide Time: 12:25)



Input for QSAR Analysis:

Software:

PaDEL – Descriptor

- Vlife
- PowerMV: A software environment for statistical analysis, molecular viewing, descriptor generation, and similarity search
- Mol D2: Descriptor generator software

Input: *n* descriptors P_1,..P_n and the value of biological activity (*IC50* for example) for *m* compounds.

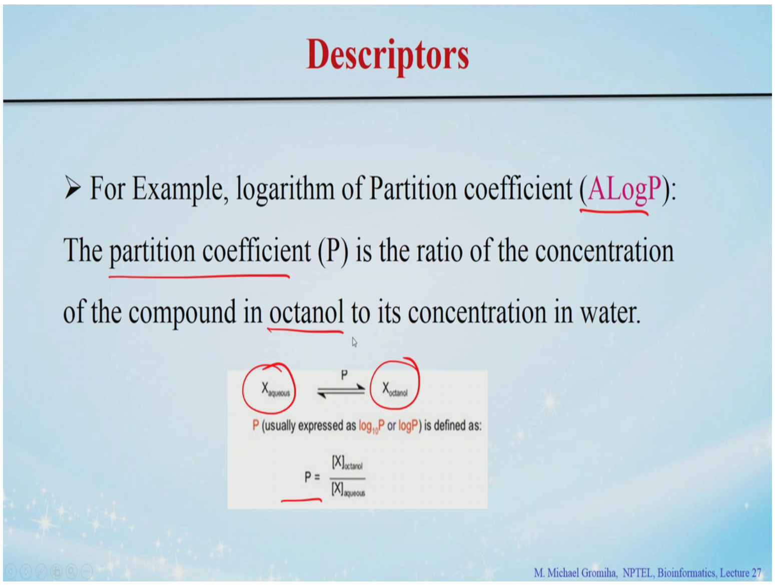
	IC50	P_1	P_2	...				P_n
Cpd1	0.7	3.7						
Cpd2	3.2	0.4						
...								
Cpdm								

- Cpd1, Cpd2,….Cpdm are the compounds and P_1, P_2, …. P_n are the descriptors (properties) of the compounds in the dataset

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, we discussed about the different methods like paDEL or the Vlife or the power MV; these are the various is software available. If we give your compound in its file format or with the molecular formula or with the structure formula; it will give you all the data, it will provide you all the data. So, here if you see these are the experimental values; so then we can have the P1, P2 are the different property values.

(Refer Slide Time: 12:51)



Descriptors

- For Example, logarithm of Partition coefficient (ALogP):

The partition coefficient (P) is the ratio of the concentration of the compound in octanol to its concentration in water.

$$X_{aqueous} \underset{}{\overset{P}{\rightleftharpoons}} X_{octanol}$$

P (usually expressed as $\log_{10}P$ or logP) is defined as:

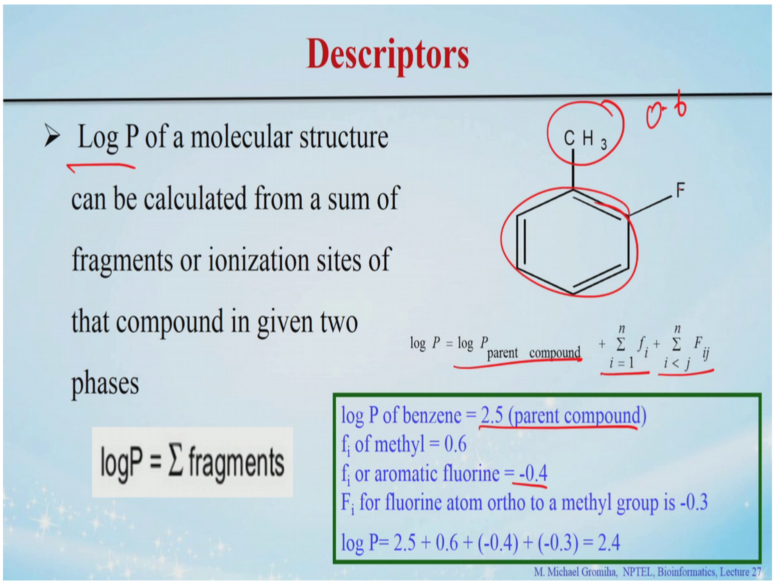
$$P = \frac{[X]_{octanol}}{[X]_{aqueous}}$$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Then take the first property; I will tell you the partition coefficient how to calculate this. So, here is this a case X in aqueous region and this is octanol; this is the octanol aqueous partition coefficient this how we define the log P values like the kind of hydrophobicity.

The partition coefficient is the ratio of the constitutional compound in octonol to its water; this is in the organic solvent octanol, ethanol we can use and with respect to the water; this will give you the partition coefficient.

(Refer Slide Time: 13:26)



Descriptors

- Log P of a molecular structure can be calculated from a sum of fragments or ionization sites of that compound in given two phases

logP = Σ fragments

$$\log P = \log P_{\text{parent compound}} + \sum_{i=1}^{n} f_i + \sum_{i<j}^{n} F_{ij}$$

log P of benzene = 2.5 (parent compound)
f_i of methyl = 0.6
f_i or aromatic fluorine = -0.4
F_i for fluorine atom ortho to a methyl group is -0.3

log P= 2.5 + 0.6 + (-0.4) + (-0.3) = 2.4

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

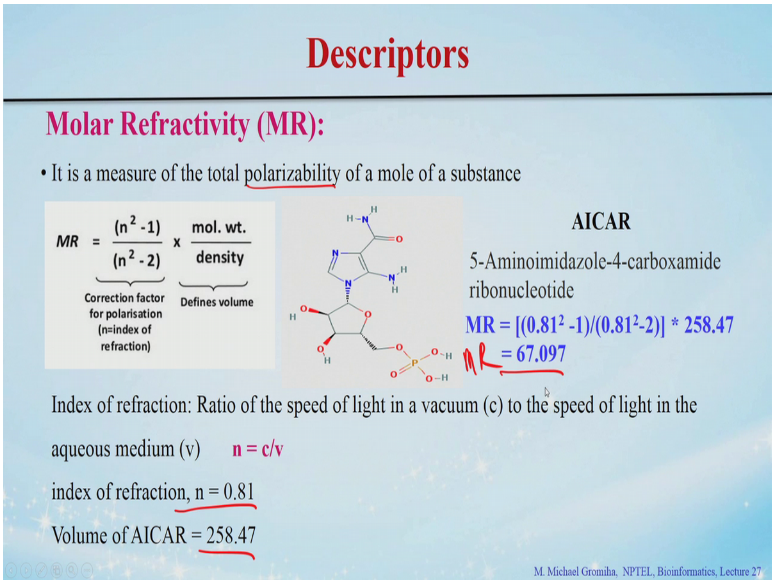
How to calculate this? For example, here I give one a compound, this is the benzene ring with the CH3 is attached here and the fluorines attached here. So, how to calculate log P? You can get the values from the some of the fragments; this a parent compound and we can see the other fragments the CH3 or the fluorine and all. Now you see the benzene this is a parent compound this is 2.5; log P is 2.5; experimentally known and the methyl group.

This is the CH3; this group this equal to 0.6 and the fluorine it is minus 0.4 and the fi for the fluorine atom in the different conformation ortho to a methyl group; this is the log P or the parent compound and this is the for the individual fragments. And this is the influence with these different groups; parent compound is a benzene; so this is given as 2.5; and individual compounds CH 3 plus F.

So, we give the 0.6 plus 0.4 and because of the Fij; this is a minus 0.3; finally, we get the value of 2.4 in the case of this particular structure. Likewise for any compounds you have you can make as a core.

The core you have the values then all the constituents you can get the fragments; then the you can get the Fij values look among these different groups. And finally, to get the log P, the summation of all these values give you the log P for the particular compound; so you can calculate.

(Refer Slide Time: 14:59)



Descriptors

Molar Refractivity (MR):

- It is a measure of the total polarizability of a mole of a substance

$$MR = \underbrace{\frac{(n^2-1)}{(n^2-2)}}_{\text{Correction factor for polarisation (n=index of refraction)}} \times \underbrace{\frac{\text{mol. wt.}}{\text{density}}}_{\text{Defines volume}}$$

AICAR

5-Aminoimidazole-4-carboxamide ribonucleotide

$MR = [(0.81^2 -1)/(0.81^2-2)] * 258.47$

MR $= 67.097$

Index of refraction: Ratio of the speed of light in a vacuum (c) to the speed of light in the aqueous medium (v) $n = c/v$

index of refraction, n = 0.81

Volume of AICAR = 258.47

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

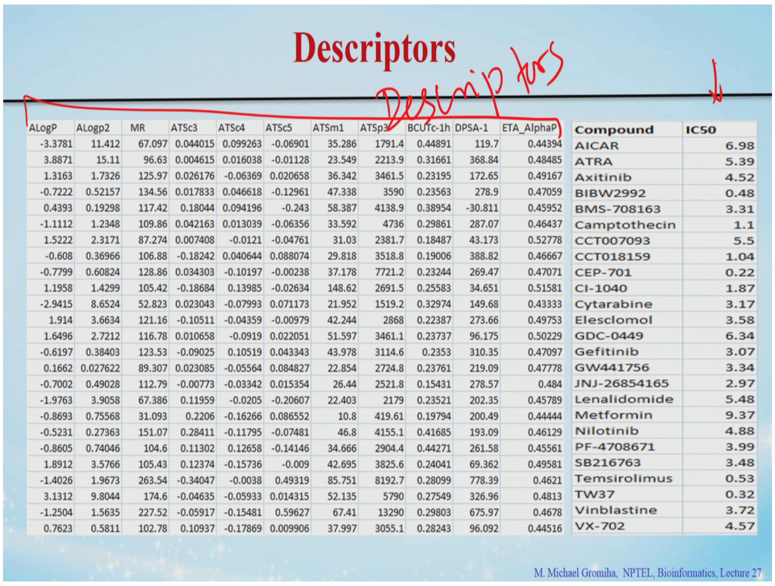
This is another example get the molar refractivity; how to get the model refractivity? So, is the measure of the total polarizability of a molecule. So, you can use this equation MR molar refractivity; this n square minus 1 divided by n square minus 2 multiplied by the molecular weight by density; this is the volume, this is a correction factor for the polarization you can say n is the index of refraction.

Because refraction index is known for the compounds and here this weight by density you can give, use the volume. So, give this compound this is the AICAR compound; the Aminoimidazole 4 carboxamide ribonucleotide. So, you can see this n equal to 0.81 because it is a index of refraction; it is known for all the compounds we know the refractive index.

This 0.81 square minus 1 then you can see the volume these to 58.47; so, you divide by this volume here. Finally, we get the value as 67.097; so n we know that n equal 0.81. So, then we can use substitute values and you can use the volume of the compound, so, finally, you can get the MR equal to 67.097. Likewise you can see any parameters; if you know the compound, you can calculate.

So, it is better to understand how to calculate? Then you can use the software to get the derive features. Otherwise you see a kind of black box to get some numbers; numbers have no meaning. So, to understand how they get the numbers at least you need to derive for some compounds and see whether this matches with the data which you get from the software. Then we are sure that the values are high or low; why it is high? And why it is low? Depending on that you can interpret your QSAR equations.

(Refer Slide Time: 16:51)



Descriptors

Descriptors

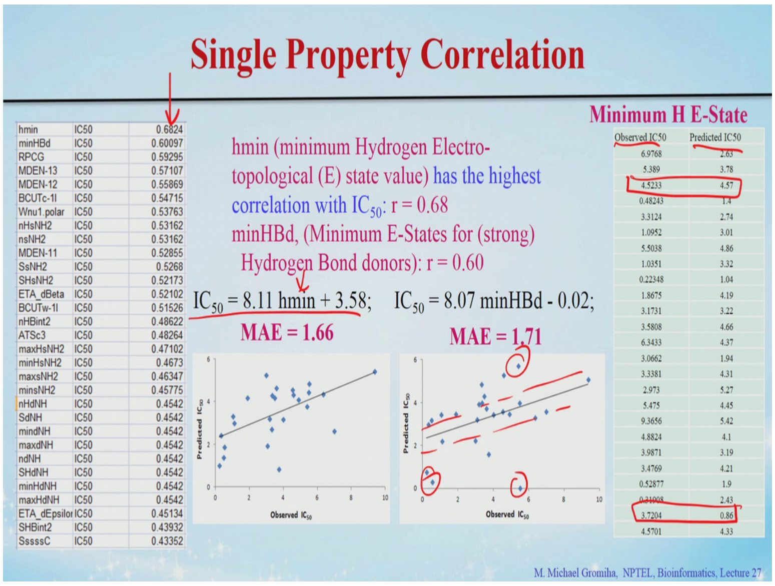
ALogP	ALogp2	MR	ATSc3	ATSc4	ATSc5	ATSm1	ATSp3	BCUTc-1h	DPSA-1	ETA_AlphaP
-3.3781	11.412	67.097	0.044015	0.099263	-0.06901	35.286	1791.4	0.44891	119.7	0.44394
3.8871	15.11	96.63	0.004615	0.016038	-0.01128	23.549	2213.9	0.31661	368.84	0.48485
1.3163	1.7326	125.97	0.026176	-0.06369	0.020658	36.342	3461.5	0.23195	172.65	0.49167
-0.7222	0.52157	134.56	0.017833	0.046618	-0.12961	47.338	3590	0.23563	278.9	0.47059
0.4393	0.19298	117.42	0.18044	0.094196	-0.243	58.387	4138.9	0.38954	-30.811	0.45952
-1.1112	1.2348	109.86	0.042163	0.013039	-0.06356	33.592	4736	0.29861	287.07	0.46437
1.5222	2.3171	87.274	0.007408	-0.0121	-0.04761	31.03	2381.7	0.18487	43.173	0.52778
-0.608	0.36966	106.88	-0.18242	0.040644	0.088074	29.818	3518.8	0.19006	388.82	0.46667
-0.7799	0.60824	128.86	0.034303	-0.10197	-0.00238	37.178	7721.2	0.23244	269.47	0.47071
1.1958	1.4299	105.42	-0.18684	0.13985	-0.02634	148.62	2691.5	0.25583	34.651	0.51581
-2.9415	8.6524	52.823	0.023043	-0.07993	0.071173	21.952	1519.2	0.32974	149.68	0.43333
1.914	3.6634	121.16	-0.10511	-0.04359	-0.00979	42.244	2868	0.22387	273.66	0.49753
1.6496	2.7212	116.78	0.010658	-0.0919	0.022051	51.597	3461.1	0.23737	96.175	0.50229
-0.6197	0.38403	123.53	-0.09025	0.10519	0.043343	43.978	3114.6	0.2353	310.35	0.47097
0.1662	0.027622	89.307	0.023085	-0.05564	0.084827	22.854	2724.8	0.23761	219.09	0.47778
-0.7002	0.49028	112.79	-0.00773	-0.03342	0.015354	26.44	2521.8	0.15431	278.57	0.484
-1.9763	3.9058	67.386	0.11959	-0.0205	-0.20607	22.403	2179	0.23521	202.35	0.45789
-0.8693	0.75568	31.093	0.2206	-0.16266	0.086552	10.8	419.61	0.19794	200.49	0.44444
-0.5231	0.27363	151.07	0.28411	-0.11795	-0.07481	46.8	4155.1	0.41685	193.09	0.46129
-0.8605	0.74046	104.6	0.11302	0.12658	-0.14146	34.666	2904.4	0.44271	261.58	0.45561
1.8912	3.5766	105.43	0.12374	-0.15736	-0.009	42.695	3825.6	0.24041	69.362	0.49581
-1.4026	1.9673	263.54	-0.34047	-0.0038	0.49319	85.751	8192.7	0.28099	778.39	0.4621
3.1312	9.8044	174.6	-0.04635	-0.05933	0.014315	52.135	5790	0.27549	326.96	0.4813
-1.2504	1.5635	227.52	-0.05917	-0.15481	0.59627	67.41	13290	0.29803	675.97	0.4678
0.7623	0.5811	102.78	0.10937	-0.17869	0.009906	37.997	3055.1	0.28243	96.092	0.44516

Compound	IC50
AICAR	6.98
ATRA	5.39
Axitinib	4.52
BIBW2992	0.48
BMS-708163	3.31
Camptothecin	1.1
CCT007093	5.5
CCT018159	1.04
CEP-701	0.22
CI-1040	1.87
Cytarabine	3.17
Elesclomol	3.58
GDC-0449	6.34
Gefitinib	3.07
GW441756	3.34
JNJ-26854165	2.97
Lenalidomide	5.48
Metformin	9.37
Nilotinib	4.88
PF-4708671	3.99
SB216763	3.48
Temsirolimus	0.53
TW37	0.32
Vinblastine	3.72
VX-702	4.57

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Now, we can have different values moral refractivity a log P and all the details; so, these are the various descriptors, so here we have IC50 values.

(Refer Slide Time: 17:04)



Single Property Correlation

hmin	IC50	0.6824
minHBd	IC50	0.60097
RPCG	IC50	0.59295
MDEN-13	IC50	0.57107
MDEN-12	IC50	0.55869
BCUTc-1l	IC50	0.54715
Wnu1.polar	IC50	0.53763
nHsNH2	IC50	0.53162
nsNH2	IC50	0.53162
MDEN-11	IC50	0.52855
SsNH2	IC50	0.5268
SHsNH2	IC50	0.52173
ETA_dBeta	IC50	0.52102
BCUTw-1l	IC50	0.51526
nHBint2	IC50	0.48622
ATSc3	IC50	0.48264
maxHsNH2	IC50	0.47102
minHsNH2	IC50	0.4673
maxsNH2	IC50	0.46347
minsNH2	IC50	0.45775
nHdNH	IC50	0.4542
SdNH	IC50	0.4542
mindNH	IC50	0.4542
maxdNH	IC50	0.4542
ndNH	IC50	0.4542
SHdNH	IC50	0.4542
minHdNH	IC50	0.4542
maxHdNH	IC50	0.4542
ETA_dEpsilor	IC50	0.45134
SHBint2	IC50	0.43932
SssssC	IC50	0.43352

hmin (minimum Hydrogen Electro-topological (E) state value) has the highest correlation with IC_{50}: $r = 0.68$

minHBd, (Minimum E-States for (strong) Hydrogen Bond donors): $r = 0.60$

$IC_{50} = 8.11\ \text{hmin} + 3.58$; $IC_{50} = 8.07\ \text{minHBd} - 0.02$;

MAE = 1.66

MAE = 1.71

Minimum H E-State

Observed IC50	Predicted IC50
6.9768	2.63
5.389	3.78
4.5233	4.57
0.48243	1.4
3.3124	2.74
1.0952	3.01
5.5038	4.86
1.0351	3.32
0.22348	1.04
1.8675	4.19
3.1731	3.22
3.5808	4.66
6.3433	4.37
3.0662	1.94
3.3381	4.31
2.973	5.27
5.475	4.45
9.3656	5.42
4.8824	4.1
3.9871	3.19
3.4769	4.21
0.52877	1.9
0.31908	2.43
3.7204	0.86
4.5701	4.33

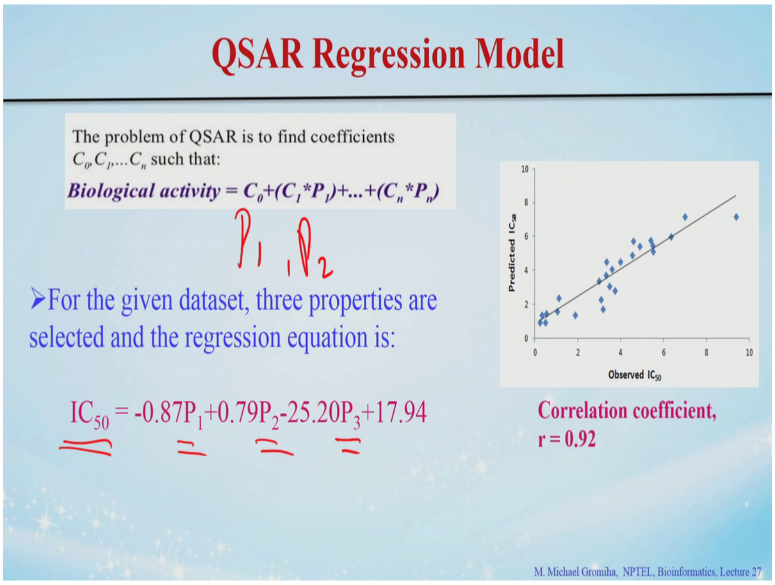
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Now, we can compare for example, if you take a only 1. So, this is one property if you see that then we can relate any property; here I used one of these electronic properties or the topological values. So, you can fit this experimental values using any of these properties, then use this equation for any new compound, you can predict the IC50 values.

Here I showed the values observed IC50 and the predicted IC50; some of them are having good relationship like this is having a good agreement and some cases it is not very good. For example, if in this case this is not able to predict well because if actually experimentally is 3.7 and the predicted one is 0.8; you can see several cases which are very close in this line for example put a line; this is close but some of them are very far these are very far, they are outliers they are not able to fit with the particular property, so you can use any properties.

So, each property can tend only the partial information because the whole compound it is not depending upon any specific property. This is what we can observe from this one; so only each property they can captured some type of information. So, in these property at least you can get the deviation of 1.6, but some of the cases yes, but some cases no; in this case what can we do?

(Refer Slide Time: 18:30)



QSAR Regression Model

The problem of QSAR is to find coefficients $C_0, C_1, \ldots C_n$ such that:

Biological activity $= C_0+(C_1*P_1)+\ldots+(C_n*P_n)$

P_1 , P_2

- For the given dataset, three properties are selected and the regression equation is:

$IC_{50} = -0.87P_1+0.79P_2-25.20P_3+17.94$

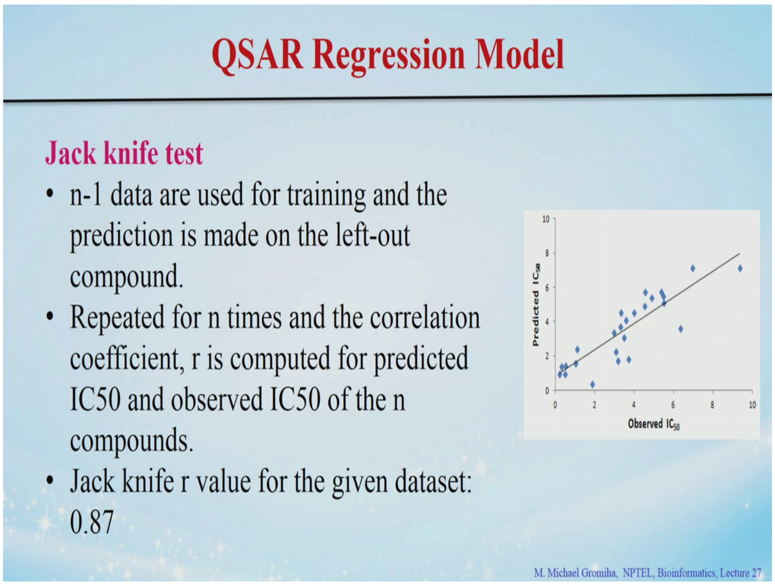
Correlation coefficient, r = 0.92

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

You can combine different properties because for example, property P1 have some information and P2 can capture some other information in particular compound; then if you add up these two properties then can better predict the activity rather than using a single one.

So, you can use different property values and see whether you can combine together to check whether this IC50 can be expressed in terms of the combination of different properties. Here I put P1, P2, and P3 three different properties; so with some coefficients. If you see this equation then here this is the observed and predicted IC50 values; most of them are very near to the line, this is the trend line. So, compared with this one you see we can improve very significantly, when you combine few properties.

(Refer Slide Time: 19:27)



QSAR Regression Model

Jack knife test

- n-1 data are used for training and the prediction is made on the left-out compound.
- Repeated for n times and the correlation coefficient, r is computed for predicted IC50 and observed IC50 of the n compounds.
- Jack knife r value for the given dataset: 0.87

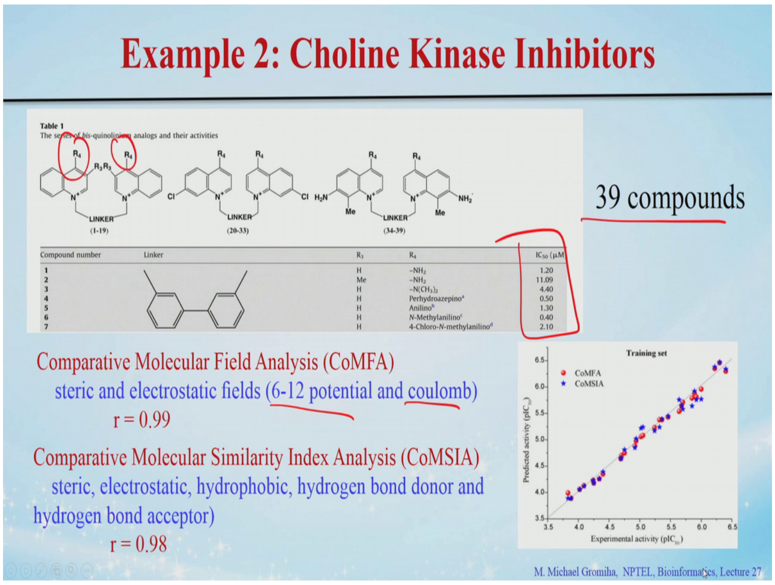
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Then you have to validate whether it is fine or not you do with the jack knife test. In this case if you have n data n minus 1 are used for the training and the left out is used for the test. Do it till 10 times and see how far is the correlation for the test data or obtained with the jack knife tests.

So, if you see this one; this is the training this is very close, but if we go to the test set some of them are close, but even then you can see several compounds which are far away from the line. That means, if you use all the data; it is able to fit because the coefficients will fit the data, but if you take some of the one compound. For example, if one of the compound which are highly variable then it is not able to fit.

So, in this case we need to find what are the other parameters, which you can also capture this particular compound and define your model. So, in this case you can get the better equations.

(Refer Slide Time: 20:26)



Example 2: Choline Kinase Inhibitors

Table 1
The series of *bis*-quinolinium analogs and their activities

Compound number	Linker	R_3	R_4	IC_{50} (μM)
1		H	$-NH_2$	1.20
2		Me	$-NH_2$	11.09
3		H	$-N(CH_3)_2$	4.40
4		H	Perhydroazepino[a]	0.50
5		H	Anilino[b]	1.30
6		H	*N*-Methylanilino[c]	0.40
7		H	4-Chloro-*N*-methylanilino[d]	2.10

39 compounds

Comparative Molecular Field Analysis (CoMFA)
steric and electrostatic fields (6-12 potential and coulomb)
r = 0.99

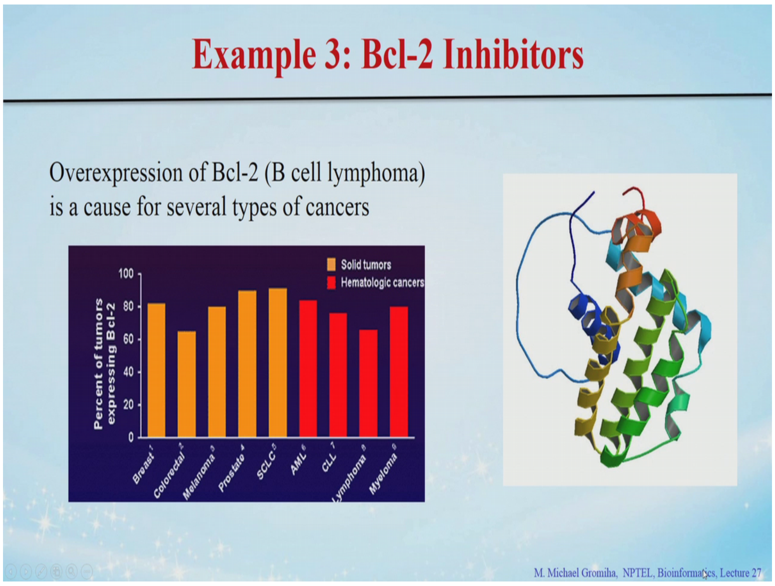
Comparative Molecular Similarity Index Analysis (CoMSIA)
steric, electrostatic, hydrophobic, hydrogen bond donor and hydrogen bond acceptor)
r = 0.98

Training set
CoMFA
CoMSIA
Predicted activity (pIC_{50})
Experimental activity (pIC_{50})
3.5
4.0
4.5
5.0
5.5
6.0
6.5

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, this is another example; so this is the choline kinase inhibitors, here they what they did; they have the common core. Core is common and there are different R groups; so various R groups for 39 compounds; we have the IC50 values. How to do this? In this case they try to develop a force field based methods; say for example, they use the electrostatic potentials 6-12 plus as the coulomb potential and from this one; they are able to predict the experimental activity with respect to the predicted activity with the high accuracy; in this type of inhibitors.

(Refer Slide Time: 21:02)



Example 3: Bcl-2 Inhibitors

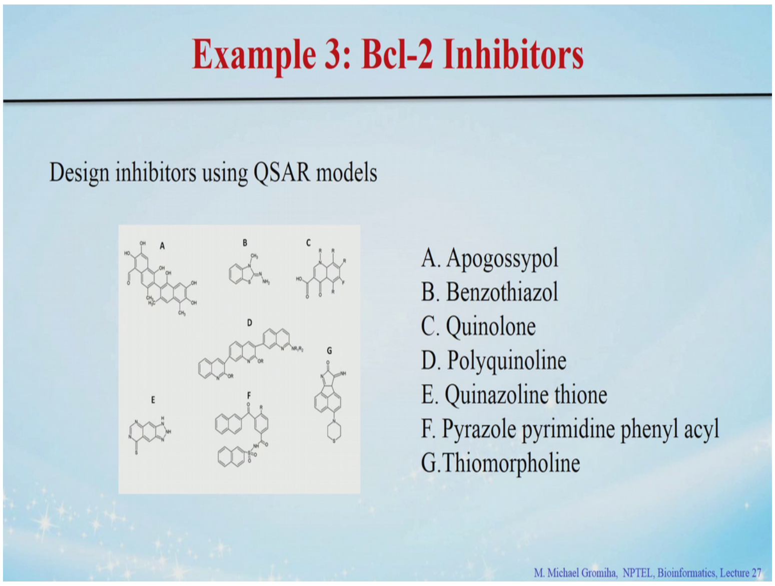
Overexpression of Bcl-2 (B cell lymphoma) is a cause for several types of cancers

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Next one another one important example the Bcl 2 inhibitors; this is the work done in our lab. So, here the why Bcl 2; this is stands for B cell lymphoma; this is also a cause for several types of cancers. So, we see in this graph you have the different types of cancers; a breast cancer, colorectal cancer or the prostate cancer and so on. And you can see the percentage of tumors which express the Bcl 2; this is the is very high. So, it is very important to identify to some drug compounds as an inhibitors; so this is the structure.

This is mainly alpha helixes you can see the structure of this Bcl 2 and how to get these models?

(Refer Slide Time: 21:49)



Example 3: Bcl-2 Inhibitors

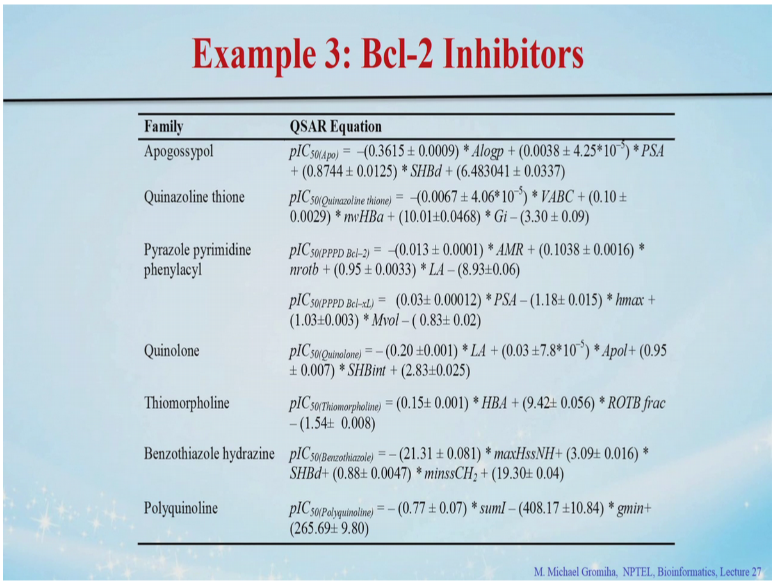
Design inhibitors using QSAR models

A. Apogossypol
B. Benzothiazol
C. Quinolone
D. Polyquinoline
E. Quinazoline thione
F. Pyrazole pyrimidine phenyl acyl
G.Thiomorpholine

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, then if you look into the structures in the active sites and you can see there are 7 core groups; so, these are a 7 core group A, B, C D, E, F, G different core groups. So, they belongs to the different core like apogossypol or the quinolone and so, on. So, based on that they find different types of compounds which showed the activity for example, it is the class A; there are several compounds which have the activity against these compounds.

(Refer Slide Time: 22:16)



Example 3: Bcl-2 Inhibitors

Family	QSAR Equation
Apogossypol	$pIC_{50(Apo)} = -(0.3615 \pm 0.0009) * Alogp + (0.0038 \pm 4.25*10^{-5}) * PSA + (0.8744 \pm 0.0125) * SHBd + (6.483041 \pm 0.0337)$
Quinazoline thione	$pIC_{50(Quinazoline\ thione)} = -(0.0067 \pm 4.06*10^{-5}) * VABC + (0.10 \pm 0.0029) * nwHBa + (10.01 \pm 0.0468) * Gi - (3.30 \pm 0.09)$
Pyrazole pyrimidine phenylacyl	$pIC_{50(PPPD\ Bcl-2)} = -(0.013 \pm 0.0001) * AMR + (0.1038 \pm 0.0016) * nrotb + (0.95 \pm 0.0033) * LA - (8.93 \pm 0.06)$
	$pIC_{50(PPPD\ Bcl-xL)} = (0.03 \pm 0.00012) * PSA - (1.18 \pm 0.015) * hmax + (1.03 \pm 0.003) * Mvol - (0.83 \pm 0.02)$
Quinolone	$pIC_{50(Quinolone)} = -(0.20 \pm 0.001) * LA + (0.03 \pm 7.8*10^{-5}) * Apol + (0.95 \pm 0.007) * SHBint + (2.83 \pm 0.025)$
Thiomorpholine	$pIC_{50(Thiomorpholine)} = (0.15 \pm 0.001) * HBA + (9.42 \pm 0.056) * ROTB\ frac - (1.54 \pm 0.008)$
Benzothiazole hydrazine	$pIC_{50(Benzothiazole)} = -(21.31 \pm 0.081) * maxHssNH + (3.09 \pm 0.016) * SHBd + (0.88 \pm 0.0047) * minssCH_2 + (19.30 \pm 0.04)$
Polyquinoline	$pIC_{50(Polyquinoline)} = -(0.77 \pm 0.07) * sumI - (408.17 \pm 10.84) * gmin + (265.69 \pm 9.80)$

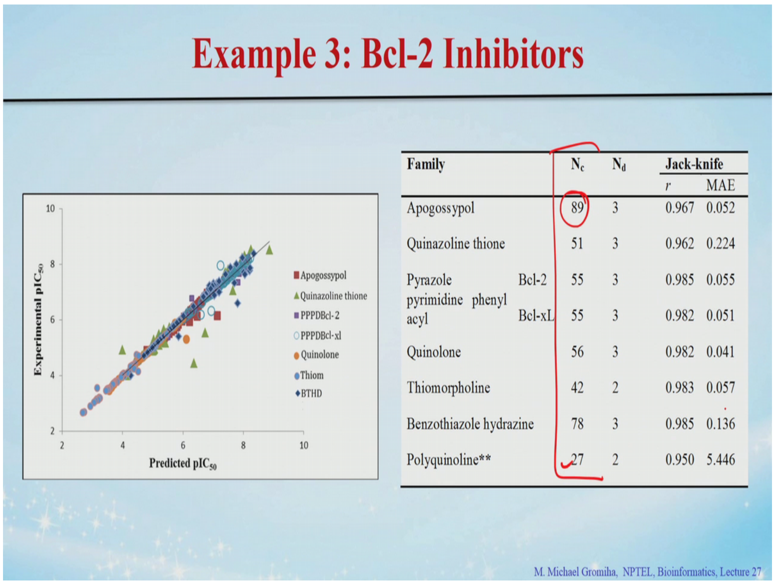
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, now we take the compound a set of compounds for example, a 100 compounds or 50 compounds in each group. So, then we derive the equations, but if you combine all these groups together; it may not work fine because they have a different course with a properties are different, but the activities also different.

In this case, sometimes similar compounds; different activities and the different compound has dissimilar activities. So, it is very important to take the core and derive equations. So, here I show example equations for the different types of families and you have the different equations. Some of them you can see the properties are similar and some cases properties are totally different.

In this case, the compounds which describe any particular family depend upon a specific properties.

(Refer Slide Time: 23:06)



Example 3: Bcl-2 Inhibitors

Family		N_c	N_d	Jack-knife	
				r	MAE
Apogossypol		89	3	0.967	0.052
Quinazoline thione		51	3	0.962	0.224
Pyrazole pyrimidine phenyl acyl	Bcl-2	55	3	0.985	0.055
	Bcl-xL	55	3	0.982	0.051
Quinolone		56	3	0.982	0.041
Thiomorpholine		42	2	0.983	0.057
Benzothiazole hydrazine		78	3	0.985	0.136
Polyquinoline**		27	2	0.950	5.446

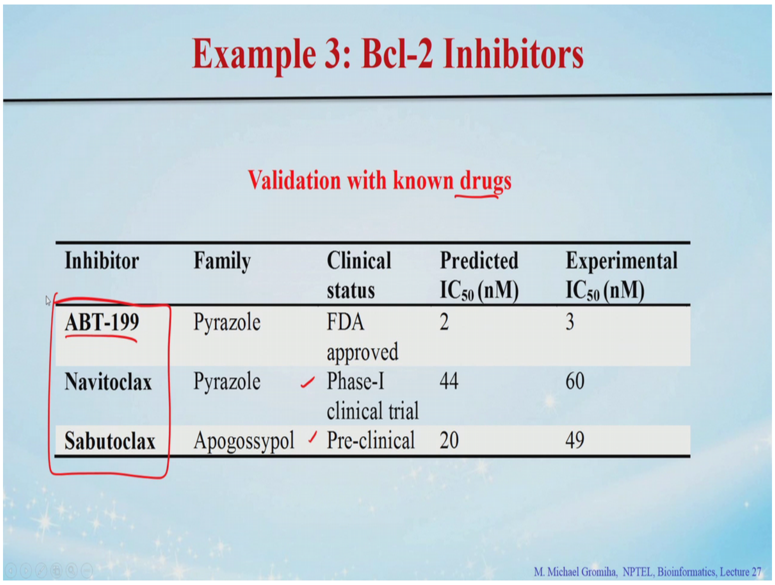
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, here I show the results for the different families this is the number of data; it ranges from 27 to 89 and you get number of descriptors. So, what is the one of the rules in the QSAR? We should finally develop the model with less number of descriptors go with the 800 and 700 descriptors finally, if you ended up with this less number of descriptor. Here the use only 3 or 2 descriptors and see the correlation it is very high.

For example this is 0.9, 0.9 and so on mostly 0.9; this is the MAE; MAE is also very high. The last one is 5 because they give in percentage; so in this case it is 5 percentage this is the unit of the micro molar. So, in this case the MAE also less and the correlation also very high.

Here the figure I show the predicted values pIC50 is logarithmic IC50 values to the experimental values. So, you can see is the good correlation between the predicted and the experimental IC50 values; you can see most of them are on the line.

(Refer Slide Time: 24:13)



Example 3: Bcl-2 Inhibitors

Validation with known drugs

Inhibitor	Family	Clinical status	Predicted IC_{50} (nM)	Experimental IC_{50} (nM)
ABT-199	Pyrazole	FDA approved	2	3
Navitoclax	Pyrazole	Phase-I clinical trial	44	60
Sabutoclax	Apogossypol	Pre-clinical	20	49

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

Now, if you to do this, then you are able to assess the performance the method and this you can predict new compounds and then you can see whether this could be a potential compounds or not. So, we need some known drugs; so, these are the some of the FDA approved drugs; this one is a approved drug ABT 199 and the phase one clinical trial and a preclinical trail; so, from different families.

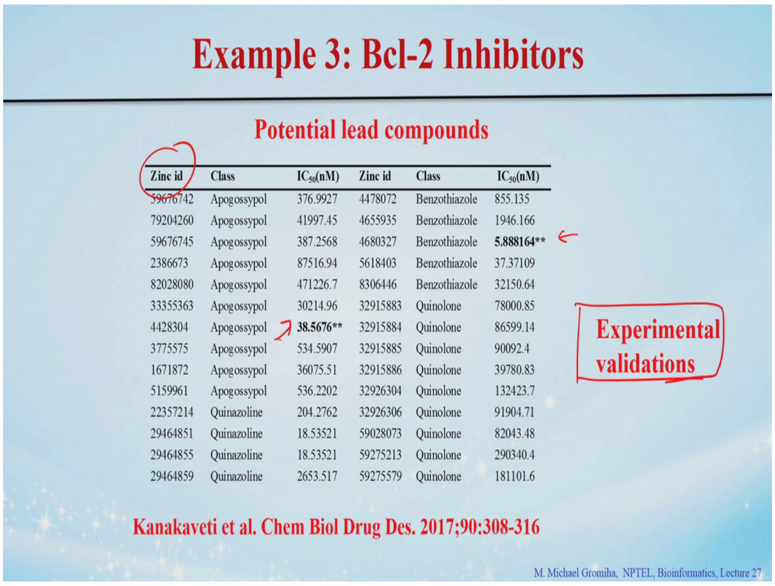
Pyrazole family or appgossypol family and so, on these are the different different inhibitors. So, we have these compounds with these compounds we can get the properties you can get the necessary properties here and use this equation to predict the IC50 values.

This is the predicted IC50 values; so this is experimental IC50 values. So, it is very close this is 2 and 3; say 44 and 60 and 20 and 49; it is not bad one. So, you can see the predicted values; agree well with these experimental values even for the known drugs. This is not used in the development of the model; so this works. So, now we can predict predict the Zinc how many data in the Zinc database?

Student: 37 million.

Around 35 million 40 million sequences; there are various database in the various number of compounds. So, for each family or for each core, we can get the compounds from the zinc database and any other databases; the core should be the same and you can have the different moieties; it can be different.

(Refer Slide Time: 25:48).



Example 3: Bcl-2 Inhibitors

Potential lead compounds

Zinc id	Class	IC_{50}(nM)	Zinc id	Class	IC_{50}(nM)
59676742	Apogossypol	376.9927	4478072	Benzothiazole	855.135
79204260	Apogossypol	41997.45	4655935	Benzothiazole	1946.166
59676745	Apogossypol	387.2568	4680327	Benzothiazole	**5.888164****
2386673	Apogossypol	87516.94	5618403	Benzothiazole	37.37109
82028080	Apogossypol	471226.7	8306446	Benzothiazole	32150.64
33355363	Apogossypol	30214.96	32915883	Quinolone	78000.85
4428304	Apogossypol	**38.5676****	32915884	Quinolone	86599.14
3775575	Apogossypol	534.5907	32915885	Quinolone	90092.4
1671872	Apogossypol	36075.51	32915886	Quinolone	39780.83
5159961	Apogossypol	536.2202	32926304	Quinolone	132423.7
22357214	Quinazoline	204.2762	32926306	Quinolone	91904.71
29464851	Quinazoline	18.53521	59028073	Quinolone	82043.48
29464855	Quinazoline	18.53521	59275213	Quinolone	290340.4
29464859	Quinazoline	2653.517	59275579	Quinolone	181101.6

Experimental validations

Kanakaveti et al. Chem Biol Drug Des. 2017;90:308-316

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 27

So, if you see this is the zinc id and you have the different families, you can see the IC50 values you can predict. So, the some of the cases you could see very a less IC50 values; in this case it could be a lead compounds. Then you can also do something more; if you take these compounds and play around the chemical groups. So, we can change the chemical group and then again predicted predict the IC50 values and you can design your own compounds and see whether you can identify any new compounds with better activity.

If you do that then finally, if subjected to experimental validations and see whether you could identify any new lead compounds which could be potentially inhibitor for this particular target. So, currently this we were trying to fuse with a new chemical groups and we designed several compounds and now we are doing these experiments to validate whether these new compounds could behave better than the existing ones the deposited in the different databases.

So, now, we go once we get these validations; get the inhibitory constants IC50 values then go for next steps and see whether we can get reliable results with a particular compounds to the next levels; to be in the clinical trials. So, what do we discuss today in summarizing?

Student: QSAR.

This QSAR. What is QSAR?

Student: Quantitative Structural Activity Relationship

Quantitative Structural Activity Relationship. So, how to relate? So, we relate the?

Student: Activity.

Activity.

Student: Descriptor.

With respect to descriptors what are the different types of descriptors?

Student: 1D

So, 1D case or here of 2D; one is the molecular weight and get the some shape, size and so on. And we do the connectivity or for the 3D; you can go with this 3D structure information then. So, there are various descriptors then you can have several databases to get the ligands; what are the different databases available to get the information regarding ligands?

Student: ChEMBL

ChEMBL

Student: zinc database.

Zinc database.

Student: Timbal

Timbal and the remaining databases, we can get the data with reliable, different types of information. Either you get the compounds or the structures or the activity or you can get the pose for example, if you take the PDB; so, what is PDB?

Student: Protein.

Protein Data Bank; you can see several identities with the protein with ligands. So, if you look into these ligands with the proteins, and you can see how the ligands interact with the proteins and how about the active sites and how about the interactions; see any shape complementarity or the chemical complementarity between the hydrogen bond donor and acceptors or the hydrophobicity and the pockets and so on, so you get the information.

Based on that you can derive the descriptors and then you have to do the feature selection. We do the subjective feature selection or the objective feature selection; how about the objective feature selection?

Student: correlation or.

You can see the correlation and if it is highly correlated; we can discard one, in the subjective one we know some properties are important.

Student: hydrogen bond

For example?

Student: Log P.

Hydrogen bond, log P; hydrogen bond donors and acceptors. So, we can keep this as IC 50 descriptor we can do that; so what are the rules of QSAR here?

Student: data set size

Data set there exist sufficient number of data compared with the descriptors; you can at least more than 5 times; with 10 times it is advisable. And then we need to what are the other rules of QSAR? So, you have to combine the variability of this data and there should not be the any repeated data sets. So, we need to get all these rules then finally, we ended up with the 3 to 5 important descriptors to define this any biological activity.

So, then we discussed about the various types of validations and they develop, model development. So, then using that information you can develop a model like that you can develop new models. So, from that the molecular descriptors can be explained can be used to predict the biological activity say IC50 values and then you can derive this equation; then you can use the new compounds you can use these particular equations wheather linear or non-linear for the identifying the new compounds.

So, once we get the new compound then finally, all the compounds are subjected to experimental validations. In this case you can get the good results then you can go for the next level of experiments and go for these clinical trials and so on. So, till now we discussed about the bioinformatics aspects on the different protein sequences and protein structures. And the folding stability and interactions as well as the structure based drug design and design inhibitors.

The next few classes, I will discuss about the some sort of the programming; the awk programming to kind of script to get the; manipulate the PDB data or any data, obtain the database or from the programming and different output of the programs. And also we see the what are the various commonly available program available problems in the literature and how to deal with the problems, how to do with that one and also see the statistical methods and machine learning techniques and case studies; to see how to systematically carry out in any project. And what would be the a result and how to interpret the data and so on.

Thanks for your kind attention.