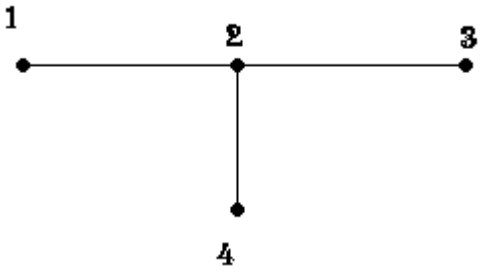
CHAPTER 3

Methodology

This chapter describes various classical as well as quantum indices and various physicochemical parameters in terms of which we would be presenting our analysis.

3.1. Topological indexes

A topological index also known as connectivity index is a numerical parameter that characterized molecular structure using graph theoretic formalism. The molecular graph is generally represented as G =(V,E) where V is a set of vertexes which represents the set of atoms in a molecule and E is a unordered pairs of elements of the set V which symbolizes covalent bonds between adjacent atoms. Most of the topological indices are derived from adjacency matrix A(G) and the distance matrix D(G) of the graph G. The hydrogen suppressed graph of isobutene is given below (Figure 3.1):



G

Figure 3.1. Hydrogen suppressed graph of isobutene

$$\text{Adjacency matrix A(G)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \end{array}$$

$$\text{Distance matrix D(G)} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix} \end{array}$$

3.1.1. Wiener index

This was the first index based on graph theory to model the boiling point of hydrocarbons. The Wiener index (W) is calculated as half-sum of all the elements d_{ij} of the distance matrix [1]:

$$W = \frac{1}{2}\sum_{ij} d_{ij} = \sum_h h.g_h \quad \ldots\ldots. (3.1)$$

Where g_h is the number of unordered pairs of vertices whose distance is h.

3.1.2. Harary index

Harary index (H) is derived from the reciprocal of the distance matrix. It measures the molecular compactness as it is increases with increasing molecular size and branching [2].

$$H = \frac{1}{2}\sum_{ij} d_{ij}^{-1} \quad \ldots\ldots. (3.2)$$

Here d_{ij} is equal to distance between vertices v_i and v_j in G.

3.1.3. Randić connectivity index

The degree of the ith vertex (δ_i) is calculated as the sum of all entries in the ith row of the adjacency matrix with n vertices:

$$\delta_i = \sum_{j=1}^{n} a_{ij} \quad \ldots\ldots. (3.3)$$

Zero order connectivity index ($^0\chi$) is defined as [3]

$$^0\chi = \sum_i (\delta_i)^{-1/2} \quad \ldots\ldots. (3.4)$$

Randić connectivity index ($^1\chi$) is defined as [4]

$$^1\chi = \sum_{all\ edges} (\delta_i \delta_j)^{-1/2} \quad \ldots\ldots. (3.5)$$

A generalized connectivity index ($^h\chi$) can be defined as [3]:

$$^h\chi = \sum (\delta_{v_o} \delta_{v_1} \ldots \delta_{v_h})^{-1/2} \quad \ldots\ldots.. (3.6)$$

Where the summation is taken over all possible path of lengths 0,1, …,h.

3.1.4. Information-theoretic topological indices

Basak et al, developed information theoretic indices which take into account all atoms including hydrogens in the constitutional formula [5]. There are three types of informational indices such as IC (mean information content), CIC (complementary information content) and SIC (structural information content). An appropriate set A of n-elements is derived from a molecular graph G depending on various classes of atoms in their topological neighborhood. The set A is partitioned into equivalence classes Ai of order n_i (i=1, 2,… h; $\Sigma_i n_i$=n). A probability distribution is then assigned to the set of equivalence classes:

$A1, A2, \ldots, A_h$

$p_1, p_2, \ldots, p_h$

Where $p_i = n_i/n$, n_i and n are the cardinalities of Ai and A respectively.

On the basis of Shannon information theory, the mean information content is defined as [6]:

$$IC = -\sum_{i=1}^{h} p_i \, log_2 p_i \quad \ldots\ldots. (3.7)$$

The binary logarithm is taken to measure the information content in bites. The total information content is then n times IC.

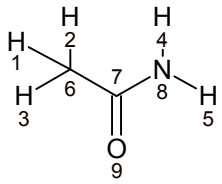
The division of atoms into different order of neighborhoods depends upon the coordination sphere taken into account [7, 8]. This leads to the indices of different order r.

$$IC_r = -\sum_{i} p_i \, log_2 p_i \quad \ldots\ldots. (3.8)$$

$$SIC_r = \frac{IC_r}{log_2 n} \quad \ldots\ldots.. (3.9)$$

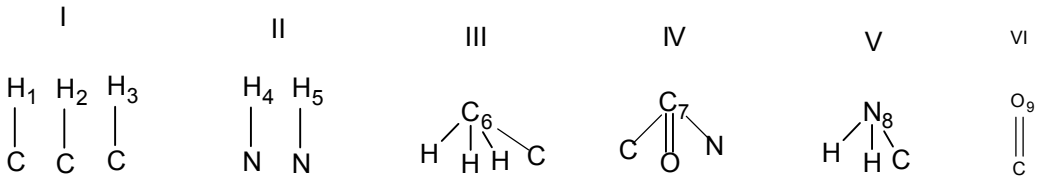
$$CIC_r = log_2 n - IC_r \quad \ldots\ldots.. (3.10)$$

For the equation (3.8), (3.9) and (3.10) the summation spans the range from i=1 to i=r, where r = 0, 1, 2… ρ, ρ is the radius of the molecular graph G and n is the total number of vertices of the graph i.e. total number of atoms in the molecule. Figure 3.2 gives a sample calculation of IC_1, SIC_1 and CIC_1.



Acetamide

First order neighbors:



	I	II	III	IV	V	VI
Subset :	$(H_1\text{-}H_3)$	$(H_4\text{-}H_5)$	C_6	C_7	N_8	O_9
Probability :	3/9	2/9	1/9	1/9	1/9	1/9

$$IC_1 = 4 * \frac{1}{9} * log_2 9 + \frac{2}{9} * log_2 \frac{9}{2} + \frac{3}{9} * log_2 \frac{9}{3} = 2.419\ bits$$

$$SIC_1 = \frac{IC_1}{log_2 9} = 0.763\ bits$$

$$CIC_1 = log_2 9 - IC_1 = 0.751\ bits$$

Figure 3.2. Labeled graph of acetamide and sample calculation of IC_1, SIC_1 and CIC_1

3.2. Molar refraction

H. A. Lorenz and L. V. Lorentz independently deduced a relation between density (d) and refractive index (n) of a substance [9].

$$R_s = \frac{n^2-1}{n^2+2} * \frac{1}{d} \quad \text{.....(3.11)}$$

The Rs is called the specific refraction of the substance and it is independent of temperature. Multiplying both side of equation (3.11) by mol. wt. M of the substance, we get molar refraction (R_M)

$$R_M = M * R_s = \frac{n^2-1}{n^2+2} * \frac{M}{d} \quad \text{..... (3.12)}$$

Here M/d is the molar volume and the molar refraction is a unit of volume since refractive index (n) is a dimensionless quantity. The molar refraction depends on the number and nature of atoms present and binding between atoms. However it is almost independent of pressure, temperature and the state of aggregation of the substance. Hence the molar refraction is partly additive and partly and partly constitutive. Table 3.1 shows molar refraction (R_M) of some atoms and structures.

Table 3.1. Molar refraction (R_M) at 589 nm, (cm^3mol^{-1})

H	1.100	O (carbonyl)	2.211	N (Tertiary amines)	2.840
C	2.418	O (ether)	1.644	Double bond (C=C)	1.733
Cl	5.967	O (hydroxyl)	1.522	Triple bond (C≡C)	2.398
Br	8.865	N (Primary amines)	2.322		
I	13.900	N (Secondary amines)	2.499		

3.3. Molar volume

The molar volume (M_V) may be defined as the volume of a gm mole of a substance at a given temperature and pressure [9]. By definition

$$M_V = \frac{M}{d} \ldots\ldots. (3.13)$$

Molar volume (M_V) can be calculated from additive increments. The additive atomic increments were achieved using a database of density (d) and calculated mol. wt. M.

3.4. Solvent accessible surface area

Molecules are often represented as a set of overlapping spheres of the constituent atoms. Lee and Richards defined solvent accessible surface and solvent exclude surface as [10]:

"The solvent accessible surface (SAS) is traced out by the center of the probe representing a solvent molecule. The solvent excluded surface (SES) is the topological boundary of the union of all possible probes which do not overlap with the molecule".

Accurate molecular surface areas are calculated using formulae given by Connolly [11] and these surface computations are based on the use of the reduced surface introduced by Sanner [12]. The algorithms which have been implemented in MSMS program are (i) computation of the reduced surface of a molecule (ii) analytical representation of the solvent excluded surface which may be self intersecting (iii) removing of all self-intersecting parts (iv) the last algorithm produces a triangulation of the SES [13].

3.5. Partition coefficient

The partition coefficient is defined as the ratio of concentrations of unionized solute between two immiscible solvents at equilibrium. Normally one of the solvent is water and the second one is octanol. Partition coefficient (P) is normally calculated in the form of its logarithm to base 10, because it ranges from 10^{-4} to 10^{8}.

$$logP = log \frac{(concentration\ of\ solute\ in\ octanol)}{(concentration\ of\ solute\ in\ water)} \quad \ldots\ldots\ldots (3.14)$$

Partition coefficient measures the hydrophilicity or hydrophobicity of a compound. The hydrophobic drugs are localized in the hydrophobic environment such as lipid bilayers of cells while hydrophilic drugs are distributed in the hydrophilic environment, such as blood serum [14]. LogP can be determined experimentally or predicted from structural data. The standard experimental procedure for logP estimation is the shake flask method, ranging from -2 to 4 logP values. High performance liquid chromatography (HPLC) may be used for more hydrophobic compounds ranging from 0 to 6 logP values. However logP values can be calculated computationally by five major methods: substituent methods, fragments methods, methods based on atomic contribution and/or surface areas, methods based on molecular properties, and, finally, methods based on solvatochromic parameters [15].

3.6. Quantum chemical descriptors

Quantum chemical techniques are generally used to get accurate molecular properties such as energy of the highest occupied molecular orbital (E_{HOMO}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), dipole moment (μ), electronegativity (χ) etc. The three main

approaches to calculating molecular properties are semi-empirical methods, the density functional method and ab-initio method.

The starting point of any quantum chemical discussion is the time-independent Schrödinger equation which is often written in a compact form as,

$$H\psi=E\psi \quad \ldots\ldots. (3.15)$$

Where H is the Hamiltonian operator corresponding to the total energy of the system i.e. sums of kinetic and potential energies, ψ is the wave function and E is the energy of the molecule. A number of solutions exist for equation (3.15) and the solution of lowest energy represents the ground state. The equation (3.15) is an eigenvalue equation.

The important idea of Hartree–Fock theory is to consider the wavefunction as a series of molecular orbitals with differing electronic occupations and one of these sets of molecular orbitals will be the lowest energy ground state. i.e.

$$\psi = \varphi 1 \varphi 2 \varphi 3 \ \ldots \varphi n \quad \ldots\ldots. (3.16)$$

$$H(\varphi 1 \varphi 2 \varphi 3 \ldots \varphi n) = E(\varphi 1 \varphi 2 \varphi 3 \ldots \varphi n) \quad \ldots\ldots. (3.17)$$

where φi is the ith molecular orbital. Hartree–Fock theory uses the so-called basis functions which are one-electron mathematical functions representing the atomic orbitals. Both semi-empirical and density functional methods make use of basis functions which is called basis sets. The incomplete treatment of exchange–correlation effects is the major drawback in the Hartree–Fock formalization when evaluating the energy of the wavefunction.

The time consuming steps in Hartree–Fock formalization is the manipulation of the mathematical representations of the molecular orbitals. In contrast, semi-empirical AM1 method (Austin model

1) [16] deals only with the valence electrons, thus reducing the computation time. Again the use of parameterized functions for some of the terms in the Hamiltonian reduces the computational time in AM1 method. These parameterized functions are derived using experimental data. In 1989, Stewart gives PM3 method (parametric method 3) [17-19] in which one-center electron repulsion integrals are taken as parameters to be optimized rather than being found from atomic spectra data.

In density functional theory (DFT), the molecular electronic energy is calculated from the molecular electron probability density. DFT provides a more complete electronic structure description than that from Hartree–Fock theory and is also more complete than semi-empirical methods. The commonly used functional in DFT calculation is B3LYP [20, 21].

An ab initio calculation uses Hamiltonian with complete representation of all nonrelativistic interactions between the nuclei and electrons in a molecule [22]. Hence ab initio calculations are limited by the types of atoms and size of molecules [23].

HOMO and LUMO energies are important quantum chemical descriptors that play a major role in governing many chemical reactions and determining electronic band gaps in solids. The energies of HOMO and LUMO are directly related to the ionization potential and electron affinity respectively. HOMO energy characterizes the susceptibility of the molecule toward attack by electrophiles while LUMO energy characterizes the susceptibility of the molecule toward attack by nucleophiles [24]. The gap energy i.e. energy difference between HOMO and LUMO is an important parameter used in QSAR study as it describes the stability of molecules [25].

Electronegativity is defined as the negative of the chemical potential (δ) i.e. negative of the partial derivative of energy (E) with respect to the number of electrons (N) at constant external potential (V) of an atomic or molecular system [26].

$$\chi = -\delta = -(\frac{\partial E}{\partial N})_V \quad \ldots\ldots\ldots (3.18)$$

By combining the work of Iczkowski and Margrave [27] with equation (3.18) and assuming a quadratic relationship between E and N, we get

$$\chi_{koopmans} = \frac{E_{HOMO}+E_{LUMO}}{2} \quad \ldots\ldots\ldots (3.19)$$

The polarity of the molecule is often described by the term dipole moment and it is an important parameter used in QSAR study. The classical expression for the electric dipole moment (μ) of a set of discrete charges Q_i is given by [28]

$$\mu = \sum_i Q_i r_i \quad \ldots\ldots (3.20)$$

where r_i is the position vector from the origin of the *i*th charge.

The quantum mechanical quantity that corresponds to the dipole moment (μ) of the system in the absence of an applied electric field is given by [28]

$$\mu = \int \psi^{(0)*} \hat{\mu} \psi^{(0)} d\tau \quad \ldots\ldots\ldots (3.21)$$

where $\psi^{(0)}$ is the unperturbed wave function and $\hat{\mu}$ is the electric dipole moment operator.

3.7. Molar entropy

Statistical mechanics provides molar entropy of an ideal gas as the sum of translational, rotational, vibrational and electronic contributions. The contribution of translational term

depends only on the molar mass of the gas. The rotational term depends on the principal moments of inertia and symmetry number. The vibrational contribution depends on the molecular vibrational frequencies and the electronic contribution depends on the ground state electronic degeneracy and in a few cases on the energies of any low-lying electronic states [29].

3.8. Molecular docking

Molecular docking is an important tool for drug discovery in which the interaction between a ligand and a protein at the atomic level are noticed to characterize the behavior of ligand in the protein binding site [30]. Molecular docking involves two main steps: prediction of the ligand pose in the active site of the protein and assessment of the binding affinity. These two steps are related to the sampling methods and scoring schemes respectively.

Examples of some popular protein-ligand docking systems include AutoDock [31], GOLD [32], DOCK [33], GLIDE [34], ICM [35] and FlexX [36]. In our case, the molecular docking simulation was carried out using the Autodock 4.2.

New released AutoDock has three search methods: simulated annealing, Monte Carlo simulated annealing (genetic algorithm) and the Lamarckian genetic algorithm and the Lamarckian genetic algorithm is the most efficient, and reliable. AutoDock predicts the binding free energies from an empirical binding free energy force field which has been calibrated using a large set of diverse protein–ligand complexes.

The free energy of binding is estimated in two steps. At the beginning ligand and protein have unbound conformation. The first step measures the intramolecular energetics of the transition from these unbound states to the bound conformation for each of ligand and protein separately. The second step calculates the intermolecular energetics of combining the ligand and protein

together into the bound complex. The force field contains six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} - \Delta S_{conf}) \text{ ….. (3.22)}$$

where L and P refer to the ligand and protein respectively in a protein-ligand complex. In the unbound state ligand and protein are sufficiently distant from one another i.e. $V_{unbound}^{P-L} = 0$. As we did not allow motion in a protein, the difference of intramolecular energy between bound and unbound state of the protein is zero.

For two atoms i, j, the pair-wise atomic terms include evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation:

$$V = W_{vdw} \sum_{ij} (\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}) + W_{hbond} \sum_{ij} E(t) (\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}) + W_{elec} \sum_{ij} \frac{q_i q_j}{\in(r_{ij}) r_{ij}} + W_{sol} \sum_{ij} (S_i V_j + S_j V_i) e^{-(\frac{r_{ij}^2}{2\sigma^2})} \text{ …… (3.23)}$$

W are weighted factors for calibrate the empirical free energy based on a set of experimentally characterized complexes. Parameters A and B were collected from the Amber force field [37] and parameters C and D are assigned to give a maximal well depth of 5 kcal/mol at 1.9 Å for hydrogen bonds with oxygen and nitrogen, and a depth of 1 kcal/mol at 2.5 Å for hydrogen bonds with sulpher. The first term is a 12-6 dispersion/repulsion term and the second term is a directional hydrogen bond term on a 10/12 potential where E(t) is a directional weight based on the angle, *t*, between the probe and the target atom. The third term is a screened Coulombic electrostatic potential. The calculation of the final term i.e. desolvation potential was most challenging and is based on the volume (V) of the atoms surrounding a given atom being

weighted by a solvation parameter (S) and a distance based exponential term . The distance weighting factor σ is set to 3.5 Å.

The loss of torsional entropy upon binding (ΔSconf) is directly proportional to the number of rotatable bonds in the molecule (Ntors):

$$\Delta S_{conf} = W_{conf} N_{tors} \quad \ldots\ldots\ldots \quad (3.24)$$

Rotatable bonds in the molecule contain all torsional degrees of freedom, including rotation of polar hydrogen atoms on hydroxyl groups and the like [31, 38].

3.9. Regression analysis

The statistical method to find out mathematical models that depicts relationships between two or many variables and the use of these relationships thus modeled for the purpose of prediction and other statistical inferences [39]. Historically, the word "regression" was first used by sir Francis Galton, who studied the parent and son height relationship. He published the results of his studies in a paper "Regression toward mediocrity in hereditary stature". Today the word regression is used in many areas of scientific investigationswithout any reference to biostatistic [40, 41].

Linear relationship between two variables is represented by a straight line. The line of average relationship is another name for a regression line. When a regression equation is to be specified, n paired observations are plotted, setting the vertical scale for dependent variable Y and horizontal scale for independent variable X. This diagram is called scatter diagram. The scatter diagram may be considered as a basis of deciding the type of regression equation, suited for the relationship between the two variables, Y and X. In natural science, social science and

economics, the number of parameters necessarily is not confined to only two variables. A large number of domains exist where involvement of more than two variables would be dictated by the problem. In those areas of study, we often need to give actual relationship between three or more variables. For such domains, multivariate regression and correlation are important tools. If we want to establish the relationship between dependent and independent variables, a mathematical equation can be formulated for tackling the situation. The equation pertaining to such a relationship may be of various types. But here we will only deal with a linear relationship which represents a plane according to the number of variables involved.

Let a mathematical model with dependent variable Y and k independent variables $X_1, X_2, \ldots\ldots X_k$, be,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + - - - - + \beta_k X_k + e \quad \ldots\ldots\ldots (3.25)$$

This type of regression equation is also known as multiple regression equation or prediction equation, where Y is predictant and $X_1, X_2, \ldots X_k$ are predictors. e is the error in formulating the linear model. This error is distributed normally with mean 0 and variance σ^2 , i.e. e $\approx N(0, \sigma^2)$. We have to estimate the parameter $\beta_0, \beta_1, \beta_3, \ldots \beta_k$ on the basis of n sample observations in which each observation is (k+1)-touple, n composite sample observations can be presented in the following format.

Composite observation no.	Variables		
	Y	X_1	X_2……..X_3……..X_k
1	Y1	x11	x21….xj1…….xk1
2	Y2	x12	x22….xj2…….xk3
.			
.			
I	Yi	x1i	x2i….xji…….xki
.			
.			
n	Yn	x1n	x2n….xjn…….xkn
Total	$\sum_i yi$	$\sum_i yi$	$\sum_i x1i$ $\sum_i xji$ xki

Estimation of β's by least squre method: For the i-th-touple the regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + e_i \quad \ldots\ldots\ldots (3.26)$$

Thus

$$e_i^2 = (yi - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \cdots - \beta_k x_{ki})^2$$

Taking the sum overall n-touples, we obtain

$$\sum_i e_i^2 = \sum_i (yi - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \cdots - \beta_k x_{ki})^2$$

For minimization, we have considered the square of errors summed over all observation. Let $\sum_i e_i^2 = Q$. To minimize Q, the overall squared error, we partially differentiate Q with respect to $\beta_0, \beta_1, \beta_3, \ldots \beta_k$ respectively and equate them to zero. Let the estimated values of $\beta_0, \beta_1, \beta_3, \ldots \beta_k$ be b_0, b, b_3. In this way we get (k+1) normal equations in (k+1) unknowns. Solving

these equations, we get the expressions for b_0, b, b_3 in terms of observed values. By substituting these estimates, we obtained the estimated equation.

Normal equations are

$$\sum_i y_i = \sum_1 b_0 + b_i \sum_i x_{1i} + b_2 \sum_i x_{2i} + \ldots + b_k \sum_i x_{ki}$$

$$\sum_i x_{1i} y_i = b_0 \sum_i x_{1i} + b_i \sum_i x_{1i}^2 + b_2 \sum_i x_{1i} x_{2i} + \ldots + b_k \sum_i x_{1i} x_{ki}$$

$$\sum_i x_{2i} y_i = b_0 \sum_i x_{2i} + b_i \sum_i x_{1i} x_{2i} + b_2 \sum_i x_{2i}^2 + \ldots + b_k \sum_i x_{2i} x_{ki}$$

……

$$\sum_i x_{ki} y_i = b_0 \sum_i x_{ki} + b_i \sum_i x_{1i} x_{ki} + b_2 \sum_i x_{2i} x_{ki} + \ldots + b_k \sum_i x_{ki}^2 \quad \ldots\ldots\ldots (3.27)$$

where i=1,2,3,…n.

If we replace the observed data in the form of vectors and matrices, the set of normal equations given by (3.26) can be written as follows.

$$Y_{nX1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad ; \qquad \beta_{(k+1)X1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} ; \qquad B_{(k+)X1} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ . \\ . \\ . \\ b_n \end{bmatrix}$$

$$\text{And } X_{nX(k+1)} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} \\ 1 & x_{12} & x_{21} & x_{k1} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix}$$

And $e' = (e_1, e_2 \ldots e_n)$

The regressions for i= 1,2,…n in matrix notation are,

$Y = X\beta + e$ …… (3.28)

And its estimated equation is

$$Y = XB$$

Whereas the set of normal equations (3.28) in matrix notation is

$$X'Y = X'XB \quad \ldots\ldots (3.29)$$

Or $B = (X'X)^{-1}X'Y$ …… (3.30)

Provided $(X'X)$ is a nonsingular matrix.

$X'Y$ is the left hand side of (3.29) will be as

$$X'X = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2j} \ldots \ldots \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{1i} \ldots \sum x_{1i}x_{1i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \ldots\ldots x_{2i}x_{ki} \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ \sum x_{1i} & \sum x_{1i}\, x_{ki} & \sum x_{2i}\, x_{ki} \ldots\ldots \sum x_{ki}^2 \end{bmatrix}$$

The above matrix is of the order (k+1)X(k+1).

From the equation set we can obtain (3.31),

$\sum_i b_0 = \sum_i y_i - b_1 \sum_i x_{1i} - b_2 \sum_i x_{2i} - \ldots - b_k \sum_i x_{ki}$

Since $\sum_i x_{ij} = n\bar{x}_j$ for j=1,2,…,k

And $\sum_i y_{ij} = n\bar{y}_j$,

$nb_0 = n(\bar{y} - b_1\overline{x_1} - b_2\overline{x_2} - \ ... \ - \ b_k\overline{x_k})$

Or $\quad b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2} - \ ... \ - \ b_k\overline{x_k}$ …… (3.31)

Substituting the value of $\sum_i b_0$ in the second equation of the set of normal equations, we get

$\sum_i x_{1i}y_i = n\overline{x_1}\,(\bar{y} - b_1\overline{x_1} - b_2\overline{x_2} - \ ... \ - \ b_k\overline{x_k}) + b_1\sum_i x_{1i}^2 + \ b_2\ \sum_i x_{1i}x_{2i} + \ b_k\ \sum_i x_{1i}x_{ki}$

Or,

$\sum_i x_{1i}y_i - n\overline{x_1}\,y \ = -\,nb_1\overline{x^2{}_1} - nb_2\overline{x_1x_2} - \ ... \ - \ nb_k\overline{x_1x_k} + b_1\sum_i x_{1i}^2 + \ b_2\ \sum_i x_{1i}x_{2i} +$

$b_k\ \sum_i x_{1i}x_{ki}$

$\sum_i(x_{1i} - \bar{x}_i)(y_i - \bar{y}) = \ b_1\sum_i(x_{1i} - \bar{x}_1)^2 + \ b_2\sum_i(x_{1i} - \bar{x}_1)(\ x_{2i} - \bar{x}_2) + \ldots + b_k\sum_i(x_{1i} -$

$\bar{x}_1)(\ x_{ki} - \bar{x}_k)$

Suppose $x_{ji} - x_j = u_{ji}, y_i - \bar{y} = v_i$

For i=1,2,…,n and j=1,2,…,k.

The above equation is,

$\sum_i u_{1i}v_i = \ b_1\sum_i u_{1i}^2 + \ b_2\sum u_{1i}u_{2i} + \cdots + b_k\sum u_{1i}u_{ki}$

Similarly the other equations of the set are

$$\sum_i u_{2i}v_i = b_2\sum_i u_{2i}^2 + b_1\sum u_{1i}u_{2i} + \cdots + b_k\sum u_{ki}u_{ki}$$

……

$$\sum_i u_{ki}v_i = b_k\sum_i u_{ki}^2 + b_2\sum u_{1i}u_{ki} + \cdots + b_1\sum u_{1i}u_{ki} \quad \ldots\ldots\ldots (3.32)$$

In the matrix notation the set of normal equations can be represented by

$$\begin{bmatrix} \sum_i u_{1i}v_i \\ \sum_i u_{2i}v_i \\ . \\ . \\ . \\ . \\ \sum_i u_{ki}v_i \end{bmatrix} = \begin{bmatrix} \sum_i u_{1i}^2 & \sum u_{1i}u_{2i} \ldots\ldots\ldots\ldots \sum u_{1i}u_{ki} \\ \sum u_{1i}u_{2i} & \sum_i u_{2i}^2 \ldots\ldots\ldots \sum u_{2i}u_{ki} \\ & . \\ & . \\ & . \\ & . \\ \sum u_{1i}u_{ki} & \sum u_{2i}u_{ki} \qquad\qquad \sum_i u_{ki}^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ . \\ b_k \end{bmatrix} \quad \ldots\ldots\ldots (3.33)$$

$Y_{k\times 1}$ $\qquad\qquad$ $A_{k\times k}$ $\qquad\qquad$ $B_{k\times 1}$

Thus the set of above equations (3.33) may be written as follows

Y=AB

In the above equations,

$$\sum_i u_{1i}^2 = \sum_i x_{ji}^2 - \left(\sum_i x_{ji}\right)^2/n$$

$$\sum_i u_{ji}u_{j'j} = \sum_i x_{ji}x_{j'j} - \left(\sum_i x_{ji}\right)\left(\sum_i x_{j'j}\right)/n$$

For j≠ j'

$$\sum_i u_{ji}v_i = \sum_i x_{ji}y_i - \left(\sum_i x_{ji}\right)\left(\sum_i y_i\right)/n$$

The matrix A is known as the coefficient matrix. From Y=AB, we can get the solution $B = A^{-1}Y$

Where A^{-1} *is* the inverse of the non-singular matrix A. Let $[c_{jj}]$ be the inverse matrix of A and j,j' =1,2,…,k.

$$A^{-1}=\begin{bmatrix} c_{11} & c_{12} & \cdot\cdot & c_{1k} \\ c_{21} & c_{22} & \cdot\cdot & c_{2k} \\ . & . & \cdot\cdot & . \\ . & . & \cdot\cdot & . \\ . & . & \cdot\cdot & . \\ c_{k1} & . & \cdot\cdot & c_{kk} \end{bmatrix}$$

Matrix A^{-1} is a symmetric matrix.

Thus we can write in the expanded form as

$$\begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdot & \cdot & c_{1k} \\ c_{21} & c_{22} & \cdot & \cdot & c_{2k} \\ . & . & \cdot & \cdot & . \\ . & . & \cdot & \cdot & . \\ . & . & \cdot & \cdot & . \\ . & . & \cdot & \cdot & . \\ c_{k1} & . & \cdot & . & c_{kk} \end{bmatrix} \begin{bmatrix} \sum_i u_{1i} v_i \\ \sum_i u_{2i} v_i \\ . \\ . \\ . \\ . \\ \sum_i u_{ki} v_i \end{bmatrix}$$

From the above relation partial regression coefficients are

$$b_1 = c_{11} \sum_i u_{1i} v_i + c_{12} \sum_i u_{2i} v_i + \ldots + c_{1k} \sum_i u_{ki} v_i$$

$$.b_2 = c_1 \sum_i u_{1i} v_i + c_{22} \sum_i u_{2i} v_i + \ldots + c_{1k} \sum_i u_{ki} v_i$$

……

$$b_j = c_{j1} \sum_i u_{1i} v_i + c_{j2} \sum_i u_{2i} v_i + \ldots + c_{jk} \sum_i u_{ki} v_i$$

……

$$b_k = c_{k1} \sum_i u_{1i} v_i + c_{12} \sum_i u_{2i} v_i + \cdots + c_{kk} \sum_i u_{ki} v_i$$

Thus by obtained values of $b_1, b_2, \ldots$ the final regression equation will be

$$(Y - \bar{y}) = b_1(X_1 - \bar{x}_1) + b_2(X_2 - \bar{x}_2) + \cdots + b_k(X_k - \bar{x}_k)$$

$$Y = R + b_1X_1 + b_2X_2 + \cdots + b_kX_k \quad \ldots\ldots\ldots (3.34)$$

$$\text{And} \quad R = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_k\bar{x}_k$$

Where the estimated value of Y will be obtained by substituting the given values $X_1, X_2, \ldots\ldots\ldots, X_k$ in the prediction Equation (3.34).

3.10. The statistical features used for predicting the best model

3.10.1. Co-efficient of correlation

The extent or degree of relationship between the two variables is measured interms of another parameter called co-efficient of correlation. It is a measure of the closeness between the two variables. It lies between -1 to +1. The correlation is perfect and positive if r=1 and it is perfect and negative if r=-1. If r=0 then there is no correlation between the two variables and said to be independent.

3.10.2. Cross validated coefficient

The cross validate coefficient is the leave one out scheme, a model is build with n-1 compounds and the n^{th} compound is predicted. Each compound is left out of the model derivation and predicted in turn. An indication of the performance of the model is obtained from the cross validated coefficient.

$$r_{CV}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_p)^2}{\sum_{i=1}^{n}(y_i - y_a)^2} \quad \ldots\ldots\ldots (3.35)$$

Where y_i is the actual experimental activity, y_p is the predicted activity of compound i and $y_a$is the average of actual experimental activity.

3.10.3. F test

Fischer statistics is the ratio between explained and unexplained variance for a given number of degree of freedom and is defined as:

$$F = \frac{\{n-(m+1)\}*r^2}{(1-r^2)*m} \quad \ldots\ldots\ldots (3.36)$$

Where n is the number of data set, m is number of descriptors used in the QSAR equation and r is the co-efficient of correlation. Higher values of the F test indicate the significance of the QSAR model.

3.10.4. Quality factor

Quality factor is defined as

$$Q = \frac{r}{s} \quad \ldots\ldots\ldots (3.37)$$

Where r is the co-efficient of correlation and s is the standard deviation. High value of Q indicate high predictive power of the QSAR model.

3.11. References

[1] H. Wiener, Structural determination of paraffin boiling points, J Am Chem Soc. 69 (1947) 17-20.

[2] F. Harary, Graph theory, Addison-Wesley, Reading, MA, 1969.

[3] L.B. Kier, L.H. Hall, Molecular connectivity in structure activity analysis, Research studies press Letchworth, Hertfordshire, U. K., 1986.

[4] M. Randić, Characterization of molecular branching, J Am Chem Soc. 97 (1975) 6609-6615.

[5] S.C. Basak, Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach, Med Sci Res. 15 (1987) 605-609.

[6] C.E. Shannon, A mathematical theory of communication, Bell Syst Tech J. 27 (1948) 379-423.

[7] S.C. Basak, S. Bertelsen, G.D. Grunwald, Application of graph theoretical parameters in quantifying molecular similarity and structure-activity studies, J Chem Inf Comput Sci. 34 (1994) 270-276.

[8] S.C. Basak, G.D. Grunwald, Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study, Chemosphere. 31 (1995) 2529-2546.

[9] R. Todeschini, V. Consonni, Handbook of molecular descriptors, Wiley, New York, 2008.

[10] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J Mol Biol. 55 (1971) 379-400.

[11] M.L. Connolly, Analytical molecular surface calculation, J Appl Cryst.16 (1983) 548-558.

[12] M.F. Sanner, Ph.D. dissertation thesis: Modeling and applications of molecular surfaces, Université de Haute-Alsace, France, 1992.

[13] M.F. Sanner, A.J. Olson, Reduced surface: an efficient way to compute molecular surfaces, Biopolymers. 38 (1996) 305-320.

[14] J. Kujawski, H. Popielarska, A. Myka, B. Drabińska, M.K. Bernard, The log P Parameter as a molecular descriptor in the computer-aided drug design – an overview, CMST. 18 (2012) 81-88.

[15] D. Eros, I. Kovesdi, L. Orfi, K. Takacs-Novak, G. Acsády, G. Kéri, Reliability of logP predictions based on calculated molecular descriptors: a critical review, Curr Med Chem. 9 (2002) 1819-1829.

[16] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, J Am Chem Soc. 107 (1985) 3902-3909.

[17] J.J.P. Stewart, Optimization of parameters for semiempirical methods.1. method, J Comput Chem. 10 (1989) 209-220.

[18] J.J.P. Stewart, Comments on a comparison of AM1 with the recently developed PM3 method- reply, J Comput Chem. 11 (1990) 543-544.

[19] J.J.P. Stewart, Optimization of parameters for semiempirical methods. III Extension of PM3 to Be, Mg, Zn, Ga, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, Tl, Pb, and Bi, J Comput Chem. 12 (1991) 320-341.

[20] J.B. Foresman, A. Frisch, Exploring chemistry with electronic structure methods, Gaussian Inc., Pittsburgh, 1996.

[21] W. Koch, M.C. Holthausen, A chemist's guide to density functional theory, Wiley-VCH, Weinheim, 2000.

[22] C. Gruber, V. Buss, Quantum-mechanically calculated properties for the development of quantitative structure-activity relationships (QSAR'S). pKa-values of phenols and aromatic and aliphatic carboxylic acids, Chemosphere. 19 (1989) 1595-1609.

[23] T. Sotomatsu, Y. Murata, T.J. Fujita, Correlation analysis of substituent effects on the acidity of benzoic acids by the AM1 method, J Comput Chem. 10 (1989) 94-98.

[24] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem Rev. 96 (1996) 1027-1043.

[25] D.F. Lewis, C. Ioannides, D.V. Parke, Interaction of a series of nitriles with the alcohol-inducible isoform of P450: computer analysis of structure-activity relationships, Xenobiotica. 24 (1994) 401-408.

[26] R.G. Parr, R.A. Donnelly, M. Levy, W.E. Palke, Electronegativity: the density functional viewpoint, J Chem Phys. 68 (1978) 3801-3807.

[27] R.P. Iczkowski, J.L. Margrave, Electronegativity, J Am Chem Soc. 83 (1961) 3547-3551.

[28] I.N Levine, Quantum chemistry, Pearson Education, Singapore, 2000.

[29] I.N Levine, Physical chemistry, Mc-Graw Hill, New York, 2011.

[30] B.J. McConkey, V. Sobolev, M. Edelman, The performance of current methods in ligand-protein docking, Curr Sci. 83 (2002) 845-855.

[31] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J Comput Chem. 19 (1998) 1639-1662.

[32] M.L. Verdonk, J.C. Cole, M.J. Hartshorn, C.W. Murray, R,D. Taylor, Improved protein-ligand docking using GOLD, Proteins. 52 (2003) 609-623.

[33] T.J. Ewing, S. Makino, A.G. Skillman, I.D. Kuntz, DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases, J Comput Aided Mol Des. 15 (2001) 411-428.

[34] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1 Method and assessment of docking accuracy, J Med Chem. 47 (2004) 1739-1749.

[35] R. Abagyan, M. Totrov, D. Kuznetsov, ICM-A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem. 15 (1994) 488-506.

[36] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J Mol Biol. 261 (1996) 470-489.

[37] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, A new force field for molecular mechanical simulation of nucleic acids and proteins, J Am Chem Soc. 106 (1984) 765-784.

[38] R. Huey, G.M. Morris, A.J. Olson, D.S. Goodsell, A semiempirical free energy force field with charge-based desolvation, J Comput Chem. 28 (2007) 1145-1152.

[39] T.W. Anderson, An introduction to multivariate analysis, John Wiley, New York, 1958.

[40] A.M. Kshirsagar, Multivariate analysis, Marcel Dekker, New York, 1972.

[41] C.R. Rao, Advanced statistical inference and its applications, Wiley Eastern, New Delhi, 1973.