Introduction
1. **Introduction to Cancer**¹ ²

Cancer is caused in all or almost all instances by *mutation* or by some other *abnormal activation* of cellular genes that control cell growth and cell mitosis. The abnormal genes are called oncogenes. A growing number of genes have been identified that contribute to the cancerous state when they mutate. These cancer-related genes fall into two classes: dominant and recessive. The dominant cancer producing genes are called *oncogenes* (Greek, *onkos*, mass, tumor; the branch of medicine that deals with cancer is known as *oncology*). Oncogenes arise as mutations of normal genes known as *proto-oncogenes*. For example, some oncogenes code for abnormal forms of cell surface receptors that bind growth factors, producing a state in which the altered receptor produces a continuous growth signal in the absence of bound growth factor. The oncogenes are considered dominant since only one of the two homologous proto-oncogenes needs to be mutated for the mutation to contribute to the cancerous state. As many as 100 different oncogenes have been discovered. Also present in all cells are *antioncogenes*, which suppress the activation of specific oncogenes. Therefore, loss of or inactivation of antioncogenes can allow activation of oncogenes that lead to cancer. Only a minute fraction of the cells that mutate in the body ever lead to cancer. There are several reasons for this. First, most mutated cells have less survival capability than normal cells and simply die. Second, only a few of the mutated cells that do survive become cancerous, because even most mutated cells still have normal feedback controls that prevent excessive growth. Third, those cells that are potentially cancerous are often, if not usually, destroyed by the body’s immune system before they grow into a cancer. This occurs in the following way: Most mutated cells form abnormal proteins within their cell bodies because of their altered genes, and these proteins activate the body’s immune system, causing it to form antibodies or sensitized lymphocytes that react against the cancerous cells, destroying them. In support of this is the fact that in people whose immune systems have been suppressed, such as in those taking immunosuppressant drugs after kidney or heart transplantation, the probability of a cancer’s developing is multiplied as much as fivefold. Fourth, usually several different activated oncogenes are required simultaneously to cause a cancer. For instance, one such gene might promote rapid reproduction of a cell line,
but no cancer occurs because there is not a simultaneous mutant gene to form the needed blood vessels. But what is it that causes the altered genes? Considering that many trillions of new cells are formed each year in humans, a better question might be why is it that all of us do not develop millions or billions of mutant cancerous cells? The answer is the incredible precision with which DNA chromosomal strands are replicated in each cell before mitosis can take place, and also the proofreading process that cuts and repairs any abnormal DNA strand before the mitotic process is allowed to proceed. Yet, despite all these inherited cellular precautions, probably one newly formed cell in every few million still has significant mutant characteristics. Thus, chance alone is all that is required for mutations to take place, so we can suppose that a large number of cancers are merely the result of an unlucky occurrence. However, the probability of mutations can be increased manyfold when a person is exposed to certain chemical, physical, or biological factors, including the following:

1. It is well known that ionizing radiation, such as x-rays, gamma rays, and particle radiation from radioactive substances and even ultraviolet light can predispose individuals to cancer. Ions formed in tissue cells under the influence of such radiation are highly reactive and can rupture DNA strands, thus causing many mutations.

2. Chemical substances of certain types also have a high propensity for causing mutations. It was discovered long ago that various aniline dye derivatives are likely to cause cancer, so that workers in chemical plants producing such substances, if unprotected, have a special predisposition to cancer. Chemical substances that can cause mutation are called carcinogens. The carcinogens that currently cause the greatest number of deaths are those in cigarette smoke. They cause about one quarter of all cancer deaths.

3. Physical irritants also can lead to cancer, such as continued abrasion of the linings of the intestinal tract by some types of food. The damage to the tissues leads to rapid mitotic replacement of the cells. The more rapid the mitosis, the greater the chance for mutation.

4. In many families, there is a strong hereditary tendency to cancer. This results from the fact that most cancers require not one mutation but two or more mutations before cancer occur. In those families that are particularly predisposed to cancer, it is
presumed that one or more cancerous genes are already mutated in the inherited genome. Therefore, far fewer additional mutations must take place in such family members before a cancer begins to grow.

5. In laboratory animals, certain types of viruses can cause some kinds of cancer, including leukemia. This usually results in one of two ways. In the case of DNA viruses, the DNA strand of the virus can insert itself directly into one of the chromosomes and thereby cause a mutation that leads to cancer. In the case of RNA viruses, some of these carry with them an enzyme called reverse transcriptase that causes DNA to be transcribed from the RNA. The transcribed DNA then inserts itself into the animal cell genome, leading to cancer.

If a cancer is detected in the early stages of its growth, before it has metastasized, the tumor may be removed by surgery. Once it has metastasized to other organs, curative surgery is no longer possible. Drugs and radiation can be used to inhibit cell multiplication and destroy malignant cells, both before and after metastasis, although these treatments unfortunately also damage the growth of normal cells. Some cancer cells retain the ability to respond to normal growth signals, such as the growth of breast tissue in response to the hormone estrogen. Blocking the action of the hormones on hormone-dependent tumor cells can inhibit their growth.

2. **Cell Division**

Starting with a single fertilized egg, the first cell division produces 2 cells. When these daughter cells divide, they each produce 2 cells, giving a total of 4. These 4 cells produce a total of 8, and so on. Thus, starting from a single cell, 3 division cycles will produce 8 cells ($2^3$), 10 division cycles will produce $2^{10} = 1024$ cells, and 20 division cycles will produce $2^{20} = 1,048,576$ cells. If the development of the human body involved only cell division and growth without any cell death, only about 46 division cycles would be needed to produce all the cells in the adult body. However, large numbers of cells die during the course of development, and even in the adult many cells survive only a few days and are continually replaced by the division of existing cells. The time between cell divisions varies considerably in different types of cells, with the most rapidly growing cells dividing about once every 24 h. During most of this period, there is no visible evidence that the cell will divide. For example, in a 24-
h division cycle, changes in cell structure begin to appear 23 h after the last division. The period between the end of one division and the appearance of the structural changes that indicate the beginning of the next division is known as **interphase**. Since the physical process of dividing one cell into two cells takes only about 1 h, the cell spends most of its time in interphase, and most of the cell properties described in this book are properties of interphase cells. One very important event related to subsequent cell division does occur during interphase, namely, the replication of DNA, which begins about 10 h before the first visible signs of division and lasts about 7 h. This period of the cell cycle is known as the S phase (synthesis) (**Figure 1**).

Following the end of DNA synthesis, there is a brief interval, G₂ (second gap), before the signs of cell division begin. The period from the end of cell division to the beginning of the S phase is the G₁ (first gap) phase of the cell cycle. In terms of the capacity to undergo cell division, there are two classes of cells in the adult body. Some cells proceed continuously through one cell cycle after another, while others seldom or never divide once they have differentiated. The first group consists of the stem cells, which provide a continuous supply of cells that form the specialized cells to replace those (such as blood cells, skin cells, and the cells lining the intestinal tract) that are continuously lost. The second class includes a number of differentiated, specialized cell types, such as nerve and striated-muscle cells, that rarely or never divide once they have

---

**Figure 1:** Phases of the cell cycle with approximate elapsed time in a cell that divides every 24 h. A cell may leave the cell cycle and enter the G₀ phase where division ceases unless the cell receives a specific signal to reenter the cycle. Synthesis, there is a brief interval, G₂ (second gap), before the signs of cell division begin. The period from the end of cell division to the beginning of the S phase is the G₁ (first gap) phase of the cell cycle.
differentiated. Also included in this second class are cells that leave the cell cycle and enter a phase known as \(G_0\) (Figure 1) in which the process that initiates DNA replication is blocked. A cell in the \(G_0\) phase, upon receiving an appropriate signal, can reenter the cell cycle, begin replicating DNA, and proceed to divide.

Cell division involves two processes: nuclear division, or mitosis, and cytoplasmic division, or cytokinesis. Although mitosis and cytokinesis are separate events, the term mitosis is often used in a broad sense to include the subsequent cytokinesis, and so the two events constitute the M phase (mitosis) of the cell cycle. Nuclear division that is not followed by cytokinesis produces multinucleated cells found in the liver, placenta, and some embryonic cells and cancer cells. When a DNA molecule replicates, the result is two identical chains termed sister chromatids, which initially are joined together at a single point called the centromere (Figure 2). As a cell begins to divide, each chromatid pair becomes highly coiled and condensed, forming a visible, rod-shaped body, a chromosome. In the condensed state prior to division, each of the 46 chromosomes, each consisting of 2 chromatids, can be identified microscopically by its length and position of its centromere.

As the duplicated chains condense, the nuclear membrane breaks down, and the chromosomes become linked in the region of their centromeres to spindle fibers (Figure 2c). The spindle fibers, composed of microtubules, are formed in the region of the cell known as the centrosome. The centrosome, which contains two centrioles and associated proteins, is required for microtubule assembly.

---

**Figure 2:** Mitosis and cytokinesis. (Only 4 of the 46 chromosomes in a human cell are illustrated.) (a) During interphase, chromatin exists in the nucleus as long, extended chains. The chains are partially coiled around clusters of histone proteins, producing a beaded appearance. (b) Prior to the onset of mitosis, DNA replicates, forming two identical sister chromatids that are joined at the centromere. A
second pair of centrioles is formed at this time. As mitosis begins, the chromatids become highly condensed and (c) become attached to spindle fibers. (d) The two chromatids of each chromosome separate and move toward opposite poles of the cell (e) as the cell divides (cytokinesis) into two daughter cells.

When a cell enters the mitotic phase of the cell cycle, the two centrioles divide, and a pair of centrioles migrates to opposite sides of the cell, thus establishing the axis of cell division. One centrosome will pass to each of the daughter cells during cytokinesis. Some of the spindle fibers extend between the two centrosomes, while others connect the centrioles to the chromosomes. The spindle fibers and centrosomes constitute the **mitotic apparatus**. As mitosis proceeds, the sister chromatids of each chromosome separate at the centromere and move toward opposite centrioles (**Figure 2d**). Cytokinesis begins as the sister chromatids separate. The cell begins to constrict along a plane perpendicular to the axis of the mitotic apparatus, and constriction continues until the cell has been pinched in half, forming the two daughter cells (**Figure 2e**), each having half the volume of the parent cell. Following cytokinesis, in each daughter cell, the spindle fibers dissolve, a nuclear envelope forms, and the chromatids uncoil. The forces producing the movements associated with mitosis and cytokinesis are generated by (1) contractile proteins similar to those producing the forces generated by muscle cells and (2) the chemical kinetics associated with the elongation and shrinkage of microtubular filaments. There are two critical checkpoints in the cell cycle, at which special events must occur in order for a cell to progress to the next phase (see **Figure 1**). One is at the boundary between G₁ and S, and the other between G₂ and M. For example, if some of the chromosomes have not completed their DNA replication during S phase, the cell will not begin mitosis until the replication is complete. To take another example, if DNA has been damaged, by x-rays for example, the cell will not enter M phase until the DNA has been repaired.

Two classes of proteins are the major players in timing cell division and the progression through these checkpoints—**cell division cycle kinases** (cdc kinases) and **cyclins**. Cyclins act as modulator molecules to activate the cdc kinases. The concentration of cyclins progressively increases during interphase and then rapidly falls during mitosis. Once activated, the kinase enzymes phosphorylate, and thus activate or inhibit a variety of proteins necessary for division, including an enzyme
that digests cyclin and thus prepares the cell to begin the next division cycle. Signals generated by DNA damage or its failure to replicate inhibit cdc kinases, thus stopping the division process. As we have noted, different types of cells progress through the cell cycle at different rates, some remaining for long periods of time in interphase. In order to progress to DNA replication, most cells must receive an external signal delivered by one or more of a group of proteins known as growth factors. Growth factors bind to their specific receptors in the cell membrane to generate intracellular signals; these signals activate various transcription factors that control the synthesis of key proteins involved in the division process and the checkpoint mechanisms. At least 50 growth factors have been identified. Many are secreted by one cell and stimulate other specific cell types to divide; others stimulate division in the cell that secretes them. Growth factors also influence various aspects of metabolism and cell differentiation. In the absence of the appropriate growth factor, most cells will not divide.
3. Introduction to QSAR

Most molecular discoveries today are the results of an iterative, three-phase cycle of design, synthesis and test. Analysis of the results from one iteration provides information and knowledge that enables the next cycle of discovery to be initiated and further improvement to be achieved. A common feature of this analysis stage is the construction of some form of model which enables the observed activity or properties to be related to the molecular structure. Such models are often referred to as Quantitative Structure Activity Relationships (QSAR).

Quantitative structure-activity relationship (QSAR) studies unquestionably are of great importance in modern chemistry and biochemistry. The concept is to transform searches for compounds with desired properties using chemical intuition and experience into a mathematically quantified and computerized form. QSAR methods are characterized by two assumptions with respect to the relationship between chemical structure and the biological potency of compounds. The first is that one can derive a quantitative measure from the structural properties significant to the biological activity of a compound. The properties are assumed to be physicochemical such as partition coefficient or sub structural such as presence or absence of certain chemical features. The other assumption is that one can mathematically describe the relationship between biological property (one can wishes to optimize and the molecular property calculated from the structure) \(4\). QSAR’s general mathematical form is represented by the following equation.

\[
\text{Biological Activity} = f(\text{Physicochemical Property}) 
\]

3.1. Objectives of QSAR

QSAR attempts to correlate structural, chemical and physical properties with biological activity by various statistical approaches \(^5\). QSAR models are scientific credible tool for predicting and classifying biological activities of untested chemicals. QSAR is an essential tool for lead development (optimization). The growing trend is to use QSAR early in drug discovery process as a screening and enrichment tool to eliminate development of those chemicals lacking “drug like” properties or those predicted to elicit a toxic response \(^6\).
3.2. Historical development

QSAR modeling born in toxicology field. In 1863, A.F.A Cros observed that the toxicity of alcohols in mammals increased as the solubility of alcohols in water decreased. Crum-Brown and Fraser expressed the idea that the physiological action of substance was function of its chemical composition and constitution and published equation (4) in 1868 which is considered to be the first formulation of a quantitative structure-activity relationship: the “physiological activity” $\Phi$ was expressed as a function of the chemical structure $C$.

$$\Phi = f(C) \ldots \ldots (4)$$  

A few decades later, in 1893, Richet showed that the cytotoxicities of a diverse set of simple organic molecules were inversely related to their water solubility’s. In 1900’s, Meyer and Overton independently suggested that narcotic (depressant) action of a group of organic compounds paralleled their oil/water partition coefficient. Hammet in mid 1930s correlated the electronic properties of organic acids and bases with their equilibrium properties. He found that a relationship resulted when different substitutions were made to aromatic compounds. The technique was introduced by Hansch et al. in the early 1960s. The approach stemmed from linear free-energy relationship in general and the Hammett equation in particular. It is based on the assumption that differences in physicochemical properties accounts for the differences in biological activities of compounds. According to this approach, the changes in physicochemical properties that affect the biological activities of a set of congeners are of major three types viz. electronic, steric, and hydrophobic.

3.4. Hansch Analysis

Corwin Hansch may be the best known as the father of concept of quantitative structure activity relationship, described the quantitative correlation of the physicochemical properties of molecules with their biological activities. It is also known as linear free energy relationship (LFER) or extra thermodynamic approach. Its basic assumption is that the effect of substituents on the strength of interactions between drug and its receptor and other biomolecules is an additive combination of the effect of the substituents on various types of simpler model intermolecular interactions. From physical chemistry, the interactions were assumed to be
electrostatic, steric and hydrophobic in nature. He suggested the linear and non-linear dependence of biological activity on different parameters.

(4) \( \log BA = a \log P + b \sigma + cEs + d \) ..........................linear

(5) \( \log BA = a(\log P)^2 + b(\log P) + c\sigma + dEs + e \) ........non-linear

Where \( \log BA \) is the logarithmic form of biological activity, a-e are constants determined for particular biological activity by multiple linear regression analysis. Log P, \( \sigma \), Es represents lipophilic, electronic and steric properties, respectively. Equation (4) was developed from the two assumptions as proposed by the Hansch.

1. The transport of a drug from the site of application to its site of action depends on the lipophilicity of the drugs (in a non-linear manner).

2. The binding affinity (interaction) of drug to its biological counterpart, such as an enzyme or receptor depends on the lipophilicity, the electronic properties and other linear free energy related properties.

### 3.5. Free Wilson analysis

The Free-Wilson approach is truly a structure-activity based methodology because it incorporates the contribution made by various structural fragments to the overall biological activity. Indicator variables are used to denote the presence or absence of a particular structural feature. It is represented by equation (6).

\[
BA = \sum aiXi + \mu \ ................. \ (6)
\]

Where BA is the biological activity, \( \mu \) is the overall activity; \( ai \) is the contribution of each structural feature, \( xi \) denotes the presence \( xi = 1 \) or absence \( xi = 0 \) of a particular structural fragment. This mathematical model incorporated symmetry equation to
minimize linear dependence between variables. This approach was easy to apply; it had its drawbacks, mostly centered on the large number of parameters and subsequent loss of the statistical degree of freedom. In 1971, in an attempt to deal with limitations of this approach, Fujita and Ban proposed a simplified approach that solely focused on the additivity of group contribution.

\[ \log \frac{A}{A_0} = \sum G_i X_i \quad \text{……………….(7)} \]

where \( A \) and \( A_0 \) represents the biological activity of the substituted and unsubstituted compounds respectively, while \( G_i \) was the activity of the ith substituent and \( X_i \) had the value of 1 or 0 that corresponded to the presence or absence of that substituent’s.

3.6. Molecular descriptors used in QSAR

Molecular descriptors can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure. This mathematical representation has to be invariant to the molecule’s size and number of atoms to allow model building with statistical methods.

The information content of structure descriptors depends on two major factors:

(1) The molecular representation of compounds.

(2) The algorithm which is used for the calculation of the descriptor.

The three major types of parameters initially suggested were,

(1) Hydrophobic

(2) Electronic

(3) Steric

Table 1: Molecular descriptors used in QSAR

<table>
<thead>
<tr>
<th>Type</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrophobic Parameters</td>
<td>Partition coefficient (log P)</td>
</tr>
<tr>
<td></td>
<td>Hansch’s substitution constant (( \pi ))</td>
</tr>
<tr>
<td></td>
<td>Hydrophobic fragmental constant (( f, f' ))</td>
</tr>
<tr>
<td></td>
<td>Distribution coefficient (log D)</td>
</tr>
<tr>
<td></td>
<td>Apparent log P</td>
</tr>
<tr>
<td></td>
<td>Capacity factor in HPLC (log ( k' ), log ( k'W ))</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Solubility parameter (log S)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Retention factor (RM)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Electronic Parameters</strong></td>
<td></td>
</tr>
<tr>
<td>Moran Autocorrelation (MATS5m)</td>
<td></td>
</tr>
<tr>
<td>Gaery Autocorrelation (GATS5m)</td>
<td></td>
</tr>
<tr>
<td>Shape Profile (SP19)</td>
<td></td>
</tr>
<tr>
<td>Hammett constant ($\sigma$, $\sigma^+$, $\sigma^-$)</td>
<td></td>
</tr>
<tr>
<td>Taft’s inductive (polar) constant ($\sigma^*$)</td>
<td></td>
</tr>
<tr>
<td>Swain and Lupton field parameter</td>
<td></td>
</tr>
<tr>
<td>Ionization constant (pKa, $\Delta$pKa)</td>
<td></td>
</tr>
<tr>
<td>Chemical shifts: IR, NMR</td>
<td></td>
</tr>
<tr>
<td>Dipole moment</td>
<td></td>
</tr>
<tr>
<td><strong>Steric Parameters</strong></td>
<td></td>
</tr>
<tr>
<td>Taft’s steric parameter (Es)</td>
<td></td>
</tr>
<tr>
<td>Molar volume (MV)</td>
<td></td>
</tr>
<tr>
<td>Van der waals radius</td>
<td></td>
</tr>
<tr>
<td>Van der waals volume</td>
<td></td>
</tr>
<tr>
<td>Molar refractivity (MR)</td>
<td></td>
</tr>
<tr>
<td>Molecular weight (MW), Parachor, Sterimol</td>
<td></td>
</tr>
<tr>
<td><strong>Quantum chemical descriptors</strong></td>
<td></td>
</tr>
<tr>
<td>Atomic net charge ($Q\sigma$, $Q\pi$)</td>
<td></td>
</tr>
<tr>
<td>Superdelocalizability</td>
<td></td>
</tr>
<tr>
<td>Energy of highest occupied molecular orbital (HOMO)</td>
<td></td>
</tr>
<tr>
<td>Energy of lowest unoccupied molecular orbital (LUMO)</td>
<td></td>
</tr>
<tr>
<td><strong>Topological descriptors</strong></td>
<td></td>
</tr>
<tr>
<td>Gaery Autocorrelation (GATS8e)</td>
<td></td>
</tr>
<tr>
<td>Maximal Electro topological Negative Variation (MAXDN)</td>
<td></td>
</tr>
<tr>
<td>Connectivity indices chi ($\chi$), kappa ($\kappa$) shape indices</td>
<td></td>
</tr>
<tr>
<td>Winner index, Zagreb index</td>
<td></td>
</tr>
<tr>
<td><strong>Spatial Descriptor</strong></td>
<td></td>
</tr>
<tr>
<td>Jurs descriptors, Shadow indices, Radius of Gyration,</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Classification of descriptor based on the dimensionality of their molecular representation

<table>
<thead>
<tr>
<th>Molecular representation</th>
<th>Descriptor</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>0D</td>
<td>Atom count, bond counts, molecular weight, sum of atomic properties</td>
<td>Molecular weight, average molecular weight number of - atoms, hydrogen atoms, carbon atoms, hetero-atoms, non-hydrogen atoms, double bonds, triple bonds, aromatic bonds, rotatable bonds, rings, 3-membered ring, 4-membered ring, 5-membered ring, 6-membered ring</td>
</tr>
<tr>
<td>1D</td>
<td>Fragments counts</td>
<td>Number of: primary C, secondary C, tertiary C, quaternary C, secondary carbon in ring, tertiary carbon in ring, quaternary carbon in ring, unsubstituted aromatic carbon, substituted carbon, number of H-bond donor atoms, number of H-bond acceptor atoms, unsaturation index, hydrophilic factor, molecular refractivity</td>
</tr>
<tr>
<td>2D</td>
<td>Topological descriptors</td>
<td>Zagreb index, Wiener index, Balaban index, connectivity indices chi ($\chi$), kappa ($K$) shape indices</td>
</tr>
<tr>
<td>3D</td>
<td>Geometrical descriptors</td>
<td>Radius of gyration, E-state topological Parameters, 3D Wiener index, 3D Balaban index</td>
</tr>
</tbody>
</table>
3.7. Methods of QSAR

Many different approaches to QSAR have been developed since Hansch’s seminal works. QSAR methods can be analyzed from two viewpoints:

(1) The types of structural parameters that are used to characterize molecular identities starting from different representation of molecules, from simple chemical formulas to 3D conformations,

(2) The mathematical procedure that is employed to obtain the quantitative relationship between these structural parameters and biological activity.

3.7.1. 2D QSAR

Methods of 2D QSAR have been divided principally into the following:

a. Hansch linear free energy relationship (LFER) model


3.7.2. 3D QSAR (Affinities correlated with 3-dimensional structure)

Molecular shape analysis (MSA)

Molecular shape analysis utilizes matrices which include common overlap steric volume and potential energy fields between pairs of superimposed molecules which were successfully correlated to the activity of series of compounds. The MSA, using common volumes, also provides some insights regarding the receptor-binding site, shape and size.

Molecular topological difference (MTD)

Simons and his coworkers developed a quantitative 3D-approach, the minimal steric (topologic) difference approach. Minimal topological difference uses a ‘hypermolecule’ concept for molecular alignment which correlated vertices (atoms) in the hypermolecule (a superposed set of molecules having common vertices) to activity differences in the series.

Comparative molecular moment analysis (COMMA)

COMMA is a unique alignment independent approach. The 3D QSAR analysis utilizes a succinct set of descriptors that would simply characterize the three dimensional information contained in the movement descriptors of molecular mass and charge up to and inclusive of second order.

Hypothetical Active Site Lattice (HASL)
Inverse grid based methodology developed in 1986-88, that allows mathematical construction of a hypothetical active site lattice which can model enzyme-inhibitor interaction and provides predictive structure-activity relationship for a set of competitive inhibitors. Computer-assisted molecule to molecule match which makes the use of multidimensional representation of inhibitor molecules. The result of such matching are used to construct a hypothetical active site by means of a lattice of points which is capable of modelling enzyme-inhibitor interactions.

Self Organizing Molecular Field Analysis (SOMFA)

SOMFA utilizes a self-centered activity, i.e., dividing the molecule set into actives (+) and inactives (-), and a grid probe process that penetrates the overlaid molecules, the resulting steric and electrostatic potentials are mapped onto the grid points and are correlated with activity using linear regression.

Comparative Molecular Field Analysis (COMFA)

The comparative molecular field analysis a grid based technique, most widely used tools for three dimensional structure-activity relationship studies, was introduced in 1988, is based on the assumption that since, in most cases, the drug-receptor interactions are non covalent, the changes in biological activities or binding affinities of sample compound correlate with changes in the steric and electrostatic fields of these molecules. These field values are correlated with biological activities by partial least square (PLS) analysis.

Comparative Molecular Similarity Indices (COMSIA)

COMSIA is an extension of COMFA methodology where molecular similarity indices can serve as a set of field descriptors in a novel application of 3D QSAR.

3. 8. Statistical Methods

Statistical methods are the mathematical foundation for the development of QSAR models. The application of multivariate analysis, data description, classification, and regression modeling, are combined with the ultimate goal of interpretation and prediction of non-evaluated or non-synthesized compounds.

Discriminate Analysis

The aim of discriminate analysis is to try and separate the molecules into their constituent classes. Discriminate analysis finds a linear combination of factor that best
discriminate between different classes. Linear discriminate analysis was used for the analysis rather than multiple linear regressions since the biological activity data were not on a continuous scale of activity but rather were classified into two groups: active and inactive. It is used to obtain a qualitative association between molecular descriptor and biological property 29.

Cluster Analysis
Cluster analysis is the process of dividing a collection of objects (molecules) into groups (or cluster) such that the objects within a cluster are highly similar whereas objects in different clusters are dissimilar. When applied to a compound dataset, the resulting clusters provide an overview of the range of structural types within the dataset and a diverse subset of compounds can be selected by choosing one or more compounds from each cluster. Clustering methods can be used to select diverse subset of compounds from larger dataset. The clustering methods most widely applied to compound selection include k-means clustering, non-hierarchiral clustering and hierarchial clustering 30.

Principle Component Analysis
The dimensionality of a data set is the number of variables that are used to describe each object. Principle Components Analysis (PCA) is a commonly used method for reducing the dimensionality of data set when there are significant correlations between some or all of the descriptors 18. PCA provides a new set of variables (the principle component) which represent most of the information contained in the independent variables 31.

Quantum Mechanical Methods
Quantum mechanical techniques are used to obtain accurate molecular properties such as electrostatic potential or polarizabilities that are only available with much lower resolution from classical mechanical techniques or those (ionization potential or electron affinities etc.) that can be obtained only quantum mechanically. The methods used commonly are divided into three categories: semi-empirical molecular orbital theory, density functional theory (DFT) and \textit{ab-initio} molecular orbital theory 32. Quantum chemical methods can be applied to quantitative structure-activity relationship by direct derivation of electronic descriptors from molecular wave
function. There is no single method that works best for all problems. Besides above mentioned methods, statistical modelling techniques aims to develop correlation models between independent variables (molecular descriptors) and dependent variable (biological property) which include simple linear regression, multiple linear regression, principle component regression, partial least squares (PLS) regression, genetic function approximation (GFA) and genetic partial least squares (GPLS).

3.9. Recent advances in structure-activity relationships

3.9.1. 3D Pharmacophore modeling
Pharmacophore modeling is powerful method to identify new potential drugs. Pharmacophore models are hypothesized on the 3D arrangement of structural properties such as hydrogen bond donor and acceptor properties, hydrophobic groups and aromatic rings of compounds that bind to the biological target. The pharmacophore concept assumes that structurally diverse molecules bind to their receptor site in a similar way, with their pharmacophoric elements interacting with the same functional groups of the receptor. Pharmacophore generation utilizes Hiphop and Hypogen hypothesis (Discovery Studio 2.1, Accelrys Inc. USA). Hiphop is common feature-based pharmacophore modeling and Hypogen is activity-based pharmacophore modeling, it is used in virtual screening of databases in lead identification/lead optimization phase.

3.9.2. 4D QSAR
4D QSAR can be interpreted as a feasible extension of 3D QSAR to address uncertainties during the alignment process. 4D QSAR concept approaches the alignment issue by incorporating molecular and spatial variety by representing each molecule in different conformation, orientations, tautomers, stereoisomers or protonation states. Two different class of 4D concept have been developed: one class of QSAR makes use of large ensemble of structurally similar conformations. In the second, QSAR approach a small set of diverse ligand configurations represents independent alternatives for the QSAR modeling.

3.9.3. 5D QSAR (Induced fit modeling)
Flexible-protein docking involves allowance for the flexibility of the binding pocket, while docking a small molecule-ligand is now-a-days considered as state of the art.
The adaptation of this philosophy to the area of QSAR is still in its infancy. To simulate induced fit in an explicit manner, simulation of a topological adaptation of the model of the binding site surface to the individual ligand molecule has been devised where the surface of the binding site model can slightly shrink or expand depending on the size and topology of the ligand binding to it. As the identification of the correct magnitude and mechanism of induced fit is not possible in absence of the structure of the true target protein, different induced fit protocols (e.g. magnitude dependent on steric, electrostatic, hydrogen bond or lipophilic potential) are presented as alternative scenarios (5D QSAR) to the QSAR 36.
4. REFERENCES


