CHAPTER 2
LITERATURE SURVEY

The relationship between molecular structure and physical, chemical properties was first introduced in 1935 by Hammett [3] and further developed by Hansch (who related biological activity to the molecules electronic characteristics and hydrophobicity) and Fujita in 1964[4] The QSAR paradigm commenced with Hammett's breakthrough in 1935 which, for the first time, made clear that the relative reaction rates of sets of congeners could be discussed in numerical terms. Many thousands of articles since then have extended his concepts to every area of organic chemistry.

QSAR dates back to the 19th century. In 1863, A.F.A. Cros at the University of Strasbourg observed that toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased [5]. In the 1890's, Hans Horst Meyer of the University of Marburg and Charles Ernest Overton of the University of Zurich, working independently, noted that the toxicity of organic compounds depended on their lipophilicity. The precursor to QSAR models were linear free energy relationships such as the Hammett equation [88], which was originally defined as a relationship between the electronic properties of acids (and bases) and their dissociation constants and reactivity. The equation is defined as

\[ \log \frac{K}{K_0} = \rho \log \frac{K'}{K'_0} \]

where, K and K' represent the dissociation constants for a set of substituted aromatic acids and K₀ and K₀' are the constants for the unsubstituted acids. \(\rho\) is the slope of the best fit line from the model fitted to the observed constants. The term \(\log (K'/K'_0)\) is denoted by \(\sigma\) and describes the substituents.

Hansch originally tried to develop QSAR models using the Hammett \(\sigma\) parameter but this did not lead to good results. He thus considered other parameters such as the
lipophilicity and molecular size as represented by molar refractivity. In order to predict biological activities, the substituent constant $\pi$ (octanol/water partition coefficient) was introduced in Hammett's equation which led to the use of physico-chemical parameters (lipophilicity $\pi$, the Taft parameters $E_s$ and $\sigma$, etc.). This sort of approach in combination with a multiple linear regression technique has been widely recognized as the famous Hansch analysis and it played a pivotal role in QSAR studies since decades.

By considering a set of molecules, a predictive model is developed that can then be used to predict the activity of other molecules. The approach depends on being able to represent the structure of a molecule in numerical form. This is in contrast to the use of empirical parameters ($\sigma$) in the case of linear free energy relationships. The numerical representations of molecules are termed descriptors, and a wide variety of descriptors can be calculated. These include simple forms such as molecular weight and atom counts or more complex types such as partition coefficients and surface-property descriptors.

Given a set of descriptors, a QSAR model can be built by defining a relationship between these descriptors (also known as the independent variables) and the observed property (termed the dependent variable). The first QSAR models, developed by Hansch, specified linear relationships. Even now, linear models are widely used owing to their simplicity and ease of development. However, developments in the field of statistics have produced many new methods of building predictive models. These include nonlinear regression techniques and algorithmic techniques [89]. Other fields such as pattern recognition and machine learning have also developed methods that have been used successfully in QSAR modeling. These include neural networks[90][91], decision trees[92][93] and so on.

In fact a statistical regression model can be built to predict any type of physical property of biological activity, given a set of observations and molecular structures. Examples of the prediction of physical properties include boiling points [6-7], aqueous solubility [8-9], glass transition temperatures [10] and ion mobility [11]. In the area of biological activities, biostatistic regression models have been developed to
predict genotoxicity [12], carcinogenicity [13] and mutagenicity [14]. Furthermore, the use of such models is not restricted to their role in screening large libraries of compounds.

In some cases a series of compounds may be synthesized and assayed. The development of a QSAR model for these compounds would provide the synthetic chemist some idea of what types of compounds could be synthesized to exhibit better activity. In other cases, the structural features highlighted by a QSAR model can provide insight into the mode of action of a drug molecule, which might be otherwise difficult to ascertain by experimental means.

2.1 QSAR STUDIES ON CDK2

Based on the studies that explored various CDK2 bound inhibitors and its role in cell cycle progression and proliferation, CDK2 acts as a potential therapeutic target in several proliferative diseases, including cancer[94]. As kinases within the cell share a high degree of sequence similarity at the active site, much effort has been devoted by many scientific groups to find more specific, potent and selective ATP competitive CDK2 inhibitors. Several kinds of compounds were reported as CDK2 inhibitors[95]. Though many of them possess potent inhibitory activities against CDK2, still many diverse compounds are being developed to increase potency and selectivity. Therefore, in the search for more specific CDK2 inhibition, a number of inhibitors have so far been described, such as, indenopyrazoles[96], pyrrolo[3,4-c]pyrazoles[97][98], indirubin[99], quinazolines[100], beta-carbolines[101], aminopyrimidine derivatives[102][103], anilino pyrimidines [104]and others[105][106], prominent among which are pyrimidine based inhibitors[107-113]. It has been shown that these inhibitors exhibit competitive inhibition with respect to ATP and SAR studies showed the importance of various substitutions on pyrimidine rings enhance potency and selectivity.

Structure activity relationship studies delineate the structural requirements for potency of inhibitors. QSAR studies have been investigated on the basis of the fact that the biological activity of the compound is a function of its physicochemical properties. From literature it was observed that several attempts were made to build
QSAR models of various CDK2 inhibitors such as indenopyrazoles[114], 6oxindoles[115][116], bisarylmaleimide[117], 3-aminopyrazoles[118], aminothiazoles[119] and purine analogs[120][121]. Moreover, none of the QSAR studies were reported on structures that covered two or more different kinds of ligands. Hence, a QSAR study on dihedral angles of active site amino acid residues with observable structure diversity, if possible, will definitely lead to more universal and robust QSAR models in studying the mode of interactions of various ligands towards CDK2 binding region.

2.3 PREDICTING PROTEIN FUNCTION

Biologists strive to understand the function of a protein. Ultimately, a laboratory experiment is needed to confirm the function of a protein, but a computational prediction can be useful both in itself and in suggesting an appropriate experiment. A variety of computational techniques have been studied, ranging from sequence comparison, to machine learning, to structure analysis and simulation. As well, the techniques range from local predictors to global predictors, either of which may or may not exploit knowledge of the hierarchy.

Protein ligand interactions

A variety of ligands interact with proteins in many biological processes such as enzyme catalysis, cell signaling etc. and regulate these processes. Information on protein-ligand interactions are accumulating at a faster pace through different experimental techniques such as X-ray crystallography, NMR, calorimetry, and equilibrium binding studies, and distributed in different sources of biological literature. Also in recent years an increasing number of small molecular ligands are designed and tested as potential inhibitors of enzymes involved in disease processes. With the advent of combinatorial chemistry and high throughput screening, the synthesis and testing of a larger number of ligands has become faster and this has led to a rapid accumulation of data. Knowledge about these ligands and their target proteins will be valuable in understanding the fundamental process of molecular recognition and will aid in the design of novel ligands and potent drugs [122]
Molecular Recognition Principles in Protein-Ligand Interactions as a Prerequisite for the Design of Specific and Selective Leads

In order to bind to a protein, a ligand has to exhibit correct shape and interaction properties complementary to the residues exposed towards the binding pocket of a target protein. Since protein-ligand binding is a process of mutual molecular recognition, rational drug design is greatly concerned with understanding the principles of molecular recognition. The statistical analysis of geometries of protein-ligand complexes provides a powerful tool to retrieve and correlate information about recognition patterns with respect to protein binding[123][124]

The function of proteins is almost invariably linked with the specific recognition of substrates and ligands in well-defined binding pockets. In consequence, proteins of related function should share comparable recognition properties exposed to these pockets. Cavbase was developed as new module for Relibase that stores protein cavities in terms of simple surface-exposed physicochemical properties. These descriptors allow for fast retrieval of proteins with functional relationships independent of a particular sequence or fold homology. The approach also allows to detect unexpected cross-reactivity of ligands among unrelated proteins. Via the alignment of binding pockets across protein family members, the consensus pattern representative for individual protein families can be extracted and mutually compared. By decomposing binding pockets into elementary sub-pocket motifs the analysis of preferred ligand occupants can be achieved [125].

Modeling Protein/Ligand Interactions in Water

Interactions between proteins in aqueous solutions affect their biological functions and determine the stability of the solutions with respect to phase transformations. Questions abound regarding the role of small solution components in these protein interactions. While some argue that small ions merely screen electrostatic interactions between proteins, others favor the view that the ions play a more active role and change protein solvent structures, thus altering the protein intermolecular interactions. We model the interaction between protein fragments and small ions using a combined explicit and implicit approach to study the effect of small ions on protein properties
A variety of ligands interact with proteins in many biological processes such as enzyme catalysis, cell signaling etc. and regulate these processes. Information on protein-ligand interactions are accumulating at a faster pace through different experimental techniques such as x-ray crystallography, NMR, calorimetry, and equilibrium binding studies, and distributed in different sources of biological literature. Also in recent years an increasing number of small molecular ligands are designed and tested as potential inhibitors of enzymes involved in disease processes. With the advent of combinatorial chemistry and high throughput screening, the synthesis and testing of a larger number of ligands has become faster and this has led to a rapid accumulation of data. Knowledge about these ligands and their target proteins will be valuable in understanding the fundamental process of molecular recognition and will aid in the design of novel ligands and potent drugs.

2.4 CELL CYCLE AND CANCER

Cancer is frequently considered to be a disease of the cell cycle. As such, it is not surprising that the deregulation of the cell cycle is one of the most frequent alterations during tumor development. Cell cycle progression is a highly ordered and tightly-regulated process that involves multiple checkpoints that assess extracellular growth signals, cell size, and DNA integrity. Cyclin-dependent kinases (CDKs) and their cyclin partners are positive regulators or accelerators that induce cell cycle progression; whereas, cyclin-dependent kinase inhibitors (CKIs) that act as brakes to stop cell cycle progression in response to regulatory signals are important negative regulators. Cancer originates from the abnormal expression or activation of positive regulators and functional suppression of negative regulators. Therefore, understanding the molecular mechanisms of the deregulation of cell cycle progression in cancer can provide important insights into how normal cells become tumorigenic, as well as how new cancer treatment strategies can be designed.

Cancer cells differ from normal cells in many important characteristics. These include the loss of differentiation, self-sufficiency in growth signals, limitless replicative potential, increased invasiveness, and decreased drug sensitivity (Hanahan and Weinberg, 2000). These differences do not arise simply from uncontrolled
cellular growth, but rather from a cellular evolution. The increased incidence of cancer as a function of age has long been interpreted to suggest that the progressive acquisition of mutations and epigenetic abnormalities in the expression of multiple genes that have highly diverse functions are required for tumorigenesis. An important group of these genes is involved in cell cycle checkpoints, which are positions of control that ensure the order of events in the cell cycle, and that integrate DNA repair with cell cycle progression. Cell cycle transition is an ordered, tightly-regulated process that involves multiple checkpoints that assess extracellular growth signals, cell size, and DNA integrity. The somatic cell cycle is divided into four distinct phases (Fig. 1). During two of these phases, the cells execute the basic events in cell division like generation of a single and faithful copy of its genetic material (synthetic or S phase) and partitioning of all the cellular components between the two identical daughter cells (mitosis or M phase). The two other phases of cell cycle represent gap periods (G1 and G2), during which the cells prepare themselves for the successful completion of the S and M phases, respectively. When the cells cease proliferation, due either to specific antimitogenic signals or to the absence of proper mitogenic signaling, then they exit the cycle and enter a non-dividing, quiescent state, known as G0. In addition, the cell cycle may be arrested at the G1 or G2 checkpoints that assess cell size, extracellular growth signals, and DNA integrity. The molecular analysis of human tumors has shown that cell cycle regulators are frequently mutated in human tumors, which underscores how important the maintenance of cell cycle commitment is in the prevention of human cancer. This review will focus on the abnormalities of the cell cycle control protein and their potential impact on cancer treatment. But, to understand the abnormalities of the cell cycle regulatory protein in cancer, we first need to consider their role in the normal cell cycle. [126]
2.5 PROTEIN DOCKING

Molecular docking is a study of how two or more molecular structures, for example drug and enzyme or receptor of protein, fit together. In other words, the problem is like solving a 3-dimensional puzzle. For example, the action of a harmful protein in human body may be prohibited by finding an inhibitor, which binds to that particular protein. Molecular docking soft wares are mainly used in drug research industry. The most important application of docking software is virtual screening. In virtual screening the most interesting and promising molecules are selected from an existing database for further research. This places demands on the used computational method; it must be fast and reliable. Another application is the research of molecular complexes.

Molecular docking can be divided into two separate problems. The search algorithm should create an optimum number of configurations that include the experimentally determined binding modes. These configurations are evaluated using scoring functions to distinguish the experimental binding modes from all other modes explored through the searching algorithm. A rigorous searching algorithm would go through all possible binding modes between the two molecules. However, this is impractical due to the size of the search space. Consider a simple system comprised of a ligand with four rotatable bonds and six rigid-body alignment parameters and a cubic active site measuring 103 Å3. The translational and rotational properties add up to six degrees of freedom. If the angles are considered in 10 degree increments and translational parameters on a 0.5 Å grid there are approximately 4 × 10^8 rigid body degrees of freedom to sample, corresponding to 6 × 10^14 configurations to be searched. This would require approximately 2 000 000 years of computational time at a rate of 10 configurations per second. As a consequence only a small amount of the total conformational space can be sampled, and so a balance must be reached between the computational expense and the amount of the search space examined.

2.6 CROSS-DOCKING OF INHIBITORS INTO CDK2 STRUCTURES

Predicting protein/ligand binding affinity is one of the most challenging computational chemistry tasks. Numerous methods have been developed to address
this challenge, but they all have limitations. Failure to account for protein flexibility has been a shortcoming of many methods. In this cross-docking study the data set comprised 150 inhibitor complexes of the protein kinase CDK2. Gold and Glide performed well in terms of docking accuracy. The chance of cross-docking a ligand within a 2 Å RMSD of its experimental pose was found to be 50%. Relative binding potency was not properly predicted from scoring functions, even though cross-docking of each inhibitor into each protein structure was performed and only scores of correctly docked ligands were considered. An accompanying paper [128] covers cross-docking and docking accuracy from the perspective of using multiple protein structures.[127]

[129] The accuracy of docking and affinity predictions of the Gold and Glide programs were investigated using single protein conformations spanning 150 CDK2/inhibitor crystallographic complexes. High docking accuracy was observed with both methods; furthermore, Glide showed modest log(IC50)/score correlations. In this part of the study, the effect of combining docking results from multiple protein conformations in a consensus fashion was probed. This approach enhanced docking accuracy only for Glide, which was attributed to the nature of its scoring function. For log(IC50)/score correlations, particular emphasis was placed on considering only scores from correctly docked poses. Using multiple instead of single protein structures showed an improvement in the correlations. Validation sets and scrambling experiments were used to examine the statistical significance and predictivity of these correlations. Rather than actual improvements in scoring accuracy, docking to multiple protein conformations produced overfitting artifacts.