Chapter 2
Translational Bioinformatics and Drug Discovery

Pramodkumar Pyarelal Gupta

Abstract With drug pipelines running dry and a slew of blockbuster medicines about to lose patent protection, the voices arguing that the traditional drug development process is too expensive and inefficient to survive are getting louder. To overcome the cost and accelerate the discovery of novel drug, in silico methodologies have made an enormous contribution. This chapter discusses the paradigm of bioinformatics and its translational approaches in drug discovery. Public domain database and efficient data mining approaches are the most optimum criteria for identification and selection of data, whereas genomic technologies such as microarray and next-generation sequencing (NGS) stand for its target identification and validation process. The use of molecular docking and QSAR techniques under the structure- and ligand-based discovery helps in screening the chemical data from nonfunctional to functional ones in terms of activity and toxicity. However, pharmacokinetic and pharmacodynamic (PKPD) simulation can help produce desired concentrations and least side effects with an approximately computed dose regimen.

Keywords Chemical database • Drug discovery • NGS • QSAR • Translational bioinformatics

2.1 Introduction

2.1.1 Translational Bioinformatics

Translational bioinformatics is the evolution of conventional in silico science that deals with storage, analysis, and knowledge extraction from voluminous genomic, proteomic, sequence, and structural data. Translational bioinformatics takes account of research in the development of novel techniques for the integration of

P.P. Gupta (✉)
School of Biotechnology and Bioinformatics, D Y Patil University, Plot 50, Sector 15, CBD Belapur, Navi Mumbai, Maharashtra, India
e-mail: pramodkumar785@gmail.com; pramod.gupta@dypatil.edu

D.-Q. Wei et al. (eds.), *Translational Bioinformatics and Its Application*, Translational Medicine Research, DOI 10.1007/978-94-024-1045-7_2

clinical and biological data that serves as a source input to designed algorithms and includes the methodology to transform the biological observations into desired knowledge that benefits the scientists, clinicians, and patients that we will see in this chapter. Complicated biological network mechanisms of disease and structure of molecules involved pose several experimental challenges in the drug discovery processes. These complications arise from independent operation of the different parts involved in drug development process with little interaction between clinical practitioners, academic institutions, and pharmaceutical industries (Portela and Soares-da-Silva 2015). Specially, the research in drug development is purpose specific and performed by highly specialized scientists and researchers in their respective fields considering few inputs from clinicians and medical practitioners in strategy design for future therapies (Portela and Soares-da-Silva 2015). Translational research is a road map in which novel therapies will link the experimental discoveries with computational techniques in delivering the clinical needs to the market. Theoretical/computational techniques offer valuable visions in experimental discoveries with pharmacological and pathophysiological mechanisms and virtual development of new prospects in designing and synthesis of novel and better molecular entities with time and cost-effectiveness (Raza 2006).

2.2 Supporting Resources

2.2.1 Online Database

Sequence database such as NCBI, EMBL, or UniProt imparts a mammoth contribution to disease, diagnosis, and drug development industry. Structure database such as Protein Databank incorporates structures evaluated by the 3D crystallography, NMR, and hybrid technology and plays a key role in the structural bioinformatics (Berman 2008). SCOP (Hubbard et al. 1999) and CATH (Oreng et al. 1997) classify the structure on the basis of structural and domain features, whereas PDBsum describes the graphical overview of the deposited 3D structure in a more precise form (Laskowski et al. 1997).

Database that handles reaction and kinetics between the genes, proteins, enzymes, and chemical components with their signal activity is known as metabolic pathway database. MetaCyc (http://metacyc.org) holds experimentally identified biochemical pathways which can be used as a reference data set for the metabolism design and analysis (Zhang et al. 2005). KEGG (http://www.genome.jp/kegg/) is a database for understanding complex functions of the biological system such as cell, organism, and ecosystem by combining the knowledge from genomic and molecular information. KEGG executes a computational representation of the biological system in a wired network diagram (system information) consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) (Kanehisa et al. 2002). The BioCyc database data

sets contain a group of organism-specific pathway/genome databases (PGDBs). They provide reference to genome and metabolic pathways of a few thousand organisms (Caspi et al. 2011). BRENDA (BRaunschweig ENzyme DAtabase) is an enzyme database established in 1987 at the Helmholtz Centre for Infection Research, formerly known as German National Research Centre for Biotechnology, and is currently maintained by the Department of Bioinformatics and Biochemistry at the TU Braunschweig. BRENDA is supplemented by enzyme-specific data classified by their biochemical reaction (Scheer et al. 2011). Other databases are also available such as Panther (Thomas et al. 2003), Reactome (Croft et al. 2010), HumanCyc (Miles et al. 2010), Mint (Licata et al. 2012), etc.

2.2.2 Small Chemical Structure Database

The online free access chemical databases assist the scientific community in identifying the previous experimental and nonexperimental chemical entities which can be an auxiliary/further tested for similar or different therapeutic applications. Online publically available small chemical structure databases such as PubChem (Bolton et al. 2008), DrugBank (Wishart et al. 2006), ZINC database (Irwin and Shoichet 2005), eMolecules (https://www.emolecules.com/), etc., listed in Table 2.1 regularly share their information on the basis of knowledge exposure. More than thousands of structures are deposited annually in these public databases with millions of compounds tested for known or unknown activities (http://depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/).

2.3 Chemical Data Mining Strategies

The exhaustive and fast designed algorithms compete in the identification of structurally similar compounds. Methodology including structural similarity searching and clustering of small molecules plays an important role in screening of compounds with identical or common scaffold in drug discovery pipelines. To search, analyze, and assemble the diverse compounds from a public database is critical to enable the full utilization of existing resources. However, most of the software in this area is only commercially available, and open source is at high demand with optimum accuracy and precision. The long-term goal of the *ChemmineR* project is to narrow this resource gap by providing free access to a flexible and expandable open-source framework for the analysis of small molecule data from chemical genomics, agrochemical, and drug discovery screens (Cao et al. 2008). Based on screening data from PubChem BioAssay database, Pouliot et al. used reported adverse event data with experimental molecular data and generated a logistic regression model to correlate and predict post-marketing ADRs (Shah and Tenenbaum 2012; Pouliot et al. 2011). In a similar way, an existing data mining

Table 2.1 List of chemical structure database

Sr no	Database	Link
1	ChEMBL	https://www.ebi.ac.uk/chembl/
2	ChemDB/Chemical Search	http://cdb.ics.uci.edu/cgibin/ChemicalSearchWeb.py
3	ChemSpider	http://www.chemspider.com/
4	ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/
5	CoCoCo	http://cococo.unibo.it/
6	Comparative Toxicogenomics Database (CTD)	http://ctdbase.org/
7	DNP (Dictionary of Natural Products)	http://dnp.chemnetbase.com/intro/index.jsp
8	DrugBank	http://www.drugbank.ca/
9	e-Drug3D	http://chemoinfo.ipmc.cnrs.fr/MOLDB/index.html
10	GLL (GPCR Ligand Library)	http://cavasotto-lab.net/Databases/GDD/
11	GLIDA (GPCR-Ligand Database)	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
12	Glide Fragment Library	http://www.schrodinger.com/Glide/Fragment-Library
13	Glide Ligand Decoys Set	http://www.schrodinger.com/Glide/Ligand-Decoys-Set
14	KEGG DRUG	http://www.genome.jp/kegg/drug/
15	KKB (Kinase Knowledgebase)	http://www.eidogen.com/kinasekb.php
16	Ligand Expo	http://ligand-expo.rutgers.edu/
17	MMsINC	http://mms.dsfarm.unipd.it/MMsINC/search/
18	Mcule database	https://mcule.com/pricing/
19	PubChem	https://pubchem.ncbi.nlm.nih.gov/
20	PubChem Mobile	https://play.google.com/store/apps/details?id=com.bim.pubchem
21	SPRESIweb	http://www.spresi.com/
22	The Cambridge Structural Database (CSD)	https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/
23	SuperDrug database	http://bioinf.charite.de/superdrug/
24	TCM	http://tcm.cmu.edu.tw/
25	Virtual Library Repository	http://nbcr.ucsd.edu/wordpress2/
26	ZINClick	http://www.symech.it/index.asp?catID=31&lang=en
27	Zinc database	http://zinc.docking.org/

algorithm was enhanced by using molecular fingerprints with chemical information that codifies the structural features or functional groups to augment the ADE signals generated from adverse event reports (Shah and Tenenbaum 2012; Vilar et al. 2011).

National Cancer Institute (NCI) database is one of the first amalgamated public efforts in distributing the large data sets according to their bioactivity information

in a searchable database format for the cancer and HIV research community (Voigt et al. 2001; Ihlenfeldt et al. 2002; Couzin 2003). ChemBank, PubChem, ZINC, and other public databases have also joined the race in screening the database on the basis of structure similarity and biological activity. Online and open-sources software are useful resources in cheminformatics software development (Girke et al. 2005).

Liu et al. (2012) demonstrated the ability to predict adverse drug reactions (ADRs) by integrating chemical, biological, and phenotypic properties of drugs. They showed that data fusion approaches are promising for large-scale ADR predictions in both preclinical and post-marketing phases (Shah 2012).

2.4 Genomic Technologies

The completion of human (*Homo sapiens*) and mouse (*Mus musculus*) genome sequence projects has increased the number of gene annotations and made it possible for bioinformaticians to develop new approaches that help experimental researchers tackle biological problems (Jin et al. 2004).

Microarray technique also known as chip-based technique was launched in the early 1990s which helped the scientists to monitor the expression of many genes concurrently, and this technology became a powerful and gold standard tool for analyzing, studying, and understanding the expression and regulation of a number of genes in parallel (Tavera-Mendoza et al. 2006). Analyzing multiple genes at the same time revealed detailed genomic and proteomic information which may lay the foundation for identification of novel target or receptor. The outputs from the microarray analysis strengthen the translational research in drug discovery and development method by generating the results from chip-based technology. Microarrays have been used to slice up nuclear receptor functions both in normal and disease states, in tissues, and in cell models. Numerous studies on nuclear receptor gene regulation for identification of downstream signaling pathways have been carried out in an experiment (Tavera-Mendoza et al. 2006). In a similar experiment, activation of PPAR is studied in a high cholesterol context trailed by microarray studies and results in a potential target gene of triglyceride-lowering drugs (Tavera-Mendoza et al. 2006; Frederiksen et al. 2004).

2.4.1 Next-Generation Sequencing (NGS)

The main application of sequencing technology is to sequence out biological data from an organism, including molecular cloning, gene identification comparative studies, and evolutionary studies. The first-generation sequencing method such as "Sanger sequencing" has been estimated to cost US$2.7 billion for the Human

Genome Project (HGP), whereas the identical procedure costs only US$1.5 million with the next-generation sequencing (NGS) method (Morini et al. 2015).

In the past few years, the NGS-based procedure has expanded its growth and application by attracting the attention from the most cutting-edge technologies. Technological advancement and increased automation, in the field of benchtop sequencing and high-throughput sequencing, have also decreased the cost and facilitated the use of sequencing technology by laboratories of all sizes involved in studies ranging from plants to human diseases (Benjamin 2015). NGS refers to those DNA sequencing methods that came after capillary-based Sanger sequencing (first generation) back in 2005. Current next-generation DNA and RNA sequencing companies include Illumina (TruSeq, HiSeq), Life Technologies (Ion Torrent, SOLiD), Complete Genomics (DNA nanoball sequencing), 454 Sequencing (pyrosequencing), and Oxford Nanopore Technologies (GridION) (Carlson 2012).

2.4.2 NGS and Personalized Medicine

Sudden cardiac death (SCD) is commonly defined as a natural death from unexplained cardiac causes. Young athlete's community is the most affected group by SCD. The most common factor identified is the adrenergic stress during the competitive sports activity for arrhythmias and SCD in the presence of inherited cardiac disease such as cardiomyopathy, primary arrhythmia syndrome, or vascular diseases. Hence, study and molecular analysis of cardiac channelopathies and cardiomyopathies would allow early diagnosis and prevention of SCD in a significant percentage of young individuals. To gain a fruitful result, one should design an appropriate and well-defined NGS diagnostic protocol and must verify in a validation phase that all the details such as mutation identified in a previous group of individuals by Sanger sequencing method must also be detectable by new advanced sequencing techniques. By contrast, novel variants identified by NGS must also be confirmed by Sanger sequencing to evaluate the reproducibility of the NGS approach (Fig. 2.1) (Morini et al. 2015).

Research published in *Nature Medicine* reports that NGS sequencing has revealed genomic alterations directly associated with clinically available therapeutics or a relevant clinical trial of a targeted therapy in 72% of 24 non-small cell lung cancer (NSCLC) tumors and in 52.5% of 40 colorectal cancer (CRC) tumors. Two novel gene fusions, KIF5B-RET in NSCLC and C2orf44-ALK in CRC, were among the alterations that might be treated by drugs. The fusion of C2orf44 and ALK produces an overexpression of anaplastic lymphoma kinase (ALK), the target of crizotinib (Xalkori), approved for the treatment of ALK-positive NSCLC, which suggests the possibility that ALK-positive CRC patients may respond to ALK-inhibitor treatment (Fig. 2.2) (Carlson 2012).

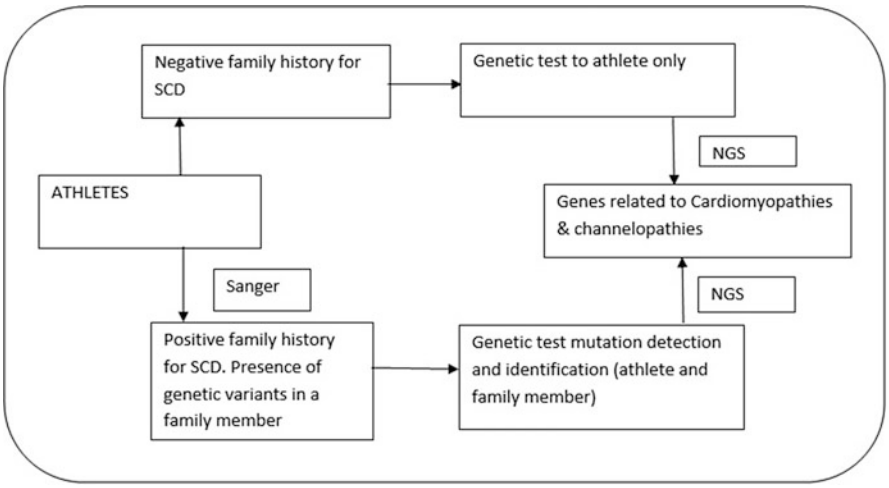

Fig. 2.1 NGS protocol for sudden cardiac death conditions

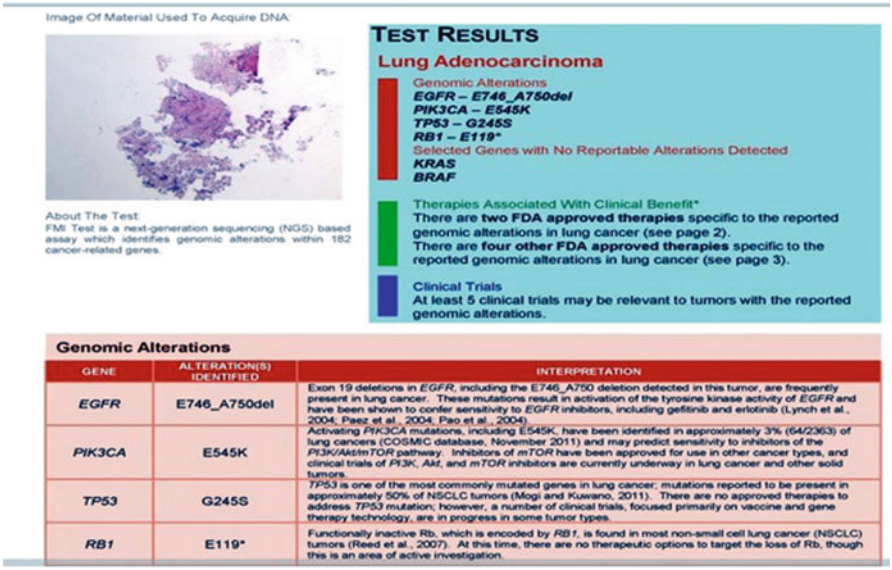

Fig. 2.2 Test result for genomic alterations (Carlson 2012)

2.5 Structure-Based Drug Discovery

In recent years, structure-based drug discovery (SBDD) is a rapidly rising methodology in overall drug discovery and development industry. The boom of genomic, proteomic, and related structural data has delivered a number of novel targets and

future prospects in lead discovery. In early 1980s the capability of rational drug design with protein structure was an unidentified object to structural biologists. The first success stories of SBDD were published in the early 1990s, and it now becomes an integral and major subject of inquiry in many research and academic organizations (Amy 2003; Roberts et al. 1990; Erickson et al. 1990; Dorsey et al. 1994).

The iterative process of SBDD principally initiates with identification, cloning, purification, and 3D structure determination of the target protein or nucleic acid by any of the following methods: X-ray crystallography, NMR, homology modeling, or various hybrid technologies. Known or calculated active sites are positioned by the computer-based algorithms and targeted by known and unknown 3D chemical compounds, ligands, or drugs identified by specific industry, organization, academic, and research groups from private and public databases. The generated complexes are ranked on the basis of binding energy, pharmacophoric interaction points, and types of interaction such as hydrogen bonding, electrostatic interaction, van der Waals interaction, etc., given in Eq. 2.1. The optimum-generated complexes are then tested with the suitable biochemical assay and knowledge is generated for further evaluation. One with the least micromolar inhibition in in vitro conditions reveals a path to scientists that the compound can be optimized to increase its potency. A repeated cycle of design, synthesis, testing, and evaluation process to a lead compound may produce a patentable market product in binding and specificity to the target (Fig. 2.3) (Amy 2003).

Binding energy:

$$\Delta G = \left(V^{L-L}\ \text{bound} - V^{L-L}\ \text{Unbound}\right) + \left(V^{P-P}\ \text{bound} - V^{P-P}\ \text{Unbound}\right) + \left(V^{P-L}\ \text{bound} - V^{P-L}\ \text{Unbound} + \Delta S_{\text{conf.}}\ \ldots\right) \tag{2.1}$$

where P refers to the protein, L refers to the ligand, V represents the pair-wise evaluations mentioned above, and ΔS_{conf} denotes the loss of conformational entropy upon binding (Ruth et al. 2007).

In comparative docking analysis between known and unknown compounds with respect to a common target, ideally, the generated ligand poses (conformations) that are closest to the experimental or known structure conformation should be ranked highest. In order, the analysis could be achieved by quantifying the similarity between a native ligand and a generated ligand pose, where root-mean-square deviation (RMSD) can be calculated between both the ligand structures (Raschka 2014):

$$\text{RMSD}\ (a,b) = \sqrt{\frac{1}{n}\ \sum_{i=1}^{n}\ (a_{ix} - b_{ix})^2\ + \left(a_{iy} - b_{iy}\right)^2\ + (a_{iz} - b_{iz})^2} \tag{2.2}$$

where a_i refers to the atoms of molecule 1 and b_i to the atoms of molecule 2. The subscripts x, y, and z denote the x-y-z coordinates for every atom.

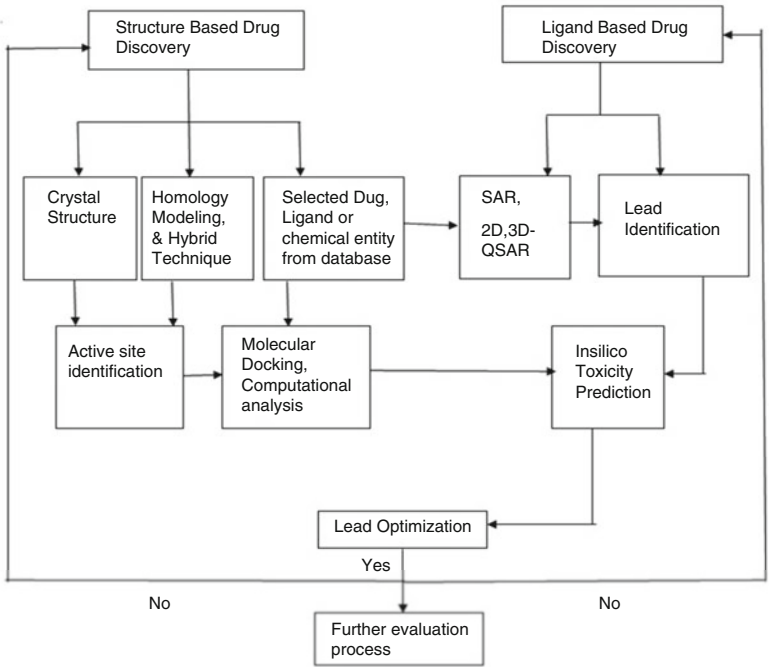

Fig. 2.3 Diagrammatic representation of a structure- and ligand-based drug discovery pipeline

2.5.1 Molecular Docking

The molecular docking is a computational technique to model the interaction between a protein macromolecule known as receptor or target and a small chemical entity/ligand/drug molecule/a protein macromolecule depending on the type of study a scientist carries out. It elucidates the behavior of a ligand molecule with the active site of a receptor protein and its fundamental biochemical process. The docking process involves two basic steps: prediction of ligand conformation within the active site of receptor protein and finally the assessment of binding energies (Meng et al. 2011; McConkey et al. 2002).

Fischer originally proposed a docking mechanism for ligand-receptor binding studies, which is the lock-and-key model, where a ligand fits into a receptor as a key and the receptor behaves as a lock. The primary early docking studies were based on this theory and receptor and ligand were considered as rigid bodies. Koshland put forward an "induced-fit" theory that takes the lock-and-key model a step further and suggests that there is a continuous change in the receptor protein conformation because of the interaction between the ligand and the protein. The theory proposes to treat both ligand and receptor as a flexible entity during docking that could describe the binding events more accurately than under rigid conditions (Fischer 1894; Kuntz et al. 1982; Koshland 1963; Hammes 2002).

Site-specific docking strategies significantly increase the docking efficiency. In many conditions the binding site is unknown. One can predict the putative binding site using commercial software such as SYBYL-X Suite (SYBYL-X-SuiteS: YBYL 8.0), SiteMap – Schrodinger (Halgren 2007), BioPredicta – VLife Molecular Design Suite (MDS) (www.vlifesciences.com), Discovery Studio (Dassault Systèmes BIOVIA 2015), FLEXX (Rarey et al. 1996), Molegro Virtual Docker System (Thomsen and Christensen 2006), ICM-Pro – Molsoft (An et al. 2005), etc. This can also be performed using online servers, e.g., Cast P (Dundas et al. 2006), GRID (Goodford 1985; Kastenholz et al. 2000), POCKET (Levitt and Banaszak 1992), SurfNet (Laskowski 1995; Glaser et al. 2006), PASS (Brady and Jr Stouten 2000), and MMC (Mezei 2003). Docking without any assumption about the binding site is called blind docking.

The main application of molecular docking lies in the structure-based virtual screening for identification of new active compounds for a particular target protein. Molecular docking technique takes a path of translational science and combines the computational output and experimental data in analyzing various biochemical reactions and interactions to study the biological system (Kubinyi 2006; Kroemer 2007; Venhorst et al. 2003; Williams et al. 2003; Meng et al. 2009).

High-throughput screening (HTS) has low rates of success to identify the optimum novel inhibitors of DNA gyrase. Boehm et al. applied de novo design methodology and successfully obtained several new inhibitors (Boehm et al. 2000). Initially, 3D complex structure of DNA gyrase with known inhibitors, ciprofloxacin and novobiocin, was analyzed and patterns of common residual interactions were calculated. Both inhibitors donate one hydrogen bond to Asp 73 and accept one hydrogen bond from a conserved water molecule. In addition, lipophilic fragments are required in the molecule to have lipophilic interaction with the receptor protein. Based on the existing knowledge, LUDI and CATALYST were employed to search and identify similar chemical structure in the Available Chemical Directory (ACD) and Roche Compound Inventory (RIC), resulting in 600 compounds. Close structural analogs of these compounds were considered and 3000 compounds were tested using biased screening. One hundred fifty compounds were selected and clustered into 14 classes of which 7 classes were proved to be the novel and true inhibitor. Succeeding hit optimization was strongly dependent on 3D structures of the binding site and generated a potent DNA gyrase inhibitor (Boehm et al. 2000).

Retinoblastoma (RB), a cancer of the eye, occurs in young children. Researchers have reported their lab findings that fatty acid synthase (FASN) is a promising diagnostic/prognostic and therapeutic target for retinoblastoma. Three inhibitors that target various domains of FASN and are potential anticancer drugs (i.e., cerulenin, triclosan, and orlistat) were considered in the previous studies (Vandhana et al. 2011; Kuhajda et al. 1994; Steven et al. 2004). The experimental data for cerulenin, triclosan, and orlistat gave an IC50 of 3.54 μg/ml, 7.29 μg/ml, and 145.25 μM, respectively, with a dose-dependent decrease in the viability of retinoblastoma cancer cells (Deepa et al. 2010). The crystal structure KS-MAT didomain of human FASN [PDB ID: 3HHD] was also used for docking with cerulenin (Pappenberger et al. 2010) and revealed the binding energy of −5.82 kcal/mol.

As there are no data available for enoyl reductase from human FASN in public database, the crystallized structure of ER domain [PDB ID: 2VZ8] was considered as a template for human ER domain. Furthermore, this model was subjected for docking with triclosan and exhibited a binding energy of −5.73 kcal/mol (Deepa et al. 2010). Pemble et al. considered crystallized 3D complex structure of the human TE domain with orlistat (PDB-ID: 2PX6) in his experiment. Based on the crystal structure, data re-docking was performed using auto dock and binding energy was found to be −2.97 kcal/mol. All these findings have indicated the predictive accuracy of the in silico methods adopted (Pemble et al. 2007).

2.6 Ligand-Based Drug Discovery

The identification of new lead molecule from millions of compound via traditional approach is time consuming and very costly. Since the 1960s, scientists from diverse life science background have put enormous efforts to identify the quantitative parameters that determine the biological activity, in what is known as QSAR/QSPR studies (Nantasenamat et al. 2009). The origin of QSAR was long back in 1863 by Cros in the field of toxicology, where he proposed the relationship between toxicity of primary aliphatic alcohol with their water solubility (Nantasenamat et al. 2009). Crum-Brown and Fraser hypothesized the relationship between chemical constitution and physiological action in 1968 (Crum-Brown and Fraser 1868). A separate discovery was led by Richet (1893), Meyer (1899), and Overton (1901) and showed a linear correlation between lipophilicity (e.g., oil-water partition coefficients) and biological effects (e.g., narcotic effects and toxicity) (Nantasenamat et al. 2009). Hammett (1935, 1937) presented a method to account for substituent effects on reaction mechanisms through the use of an equation which took two parameters into consideration, namely, (i) the substituent constant and (ii) the reaction constant (Nantasenamat et al. 2009; Crum-Brown and Fraser 1868).

Hammett quantified the effect of substituents on any reaction by defining an empirical electronic substituent parameter (σ), which is derived from the acidity constants, Ka's of substituted benzoic acids (Fig. 2.4) (https://web.viu.ca/krogh/chem331/LFER%20Hammett%202012.pdf).

$$\log\left(\frac{KX}{KH}\right) = \rho\sigma \text{ or } pKH - pKX = \rho\sigma \tag{2.3}$$

For the ionization of benzoic acid in pure water at 25 °C (the reference reaction), the constant ρ is defined as 1.00. Thus, the electronic substituent parameter (σ) is defined as

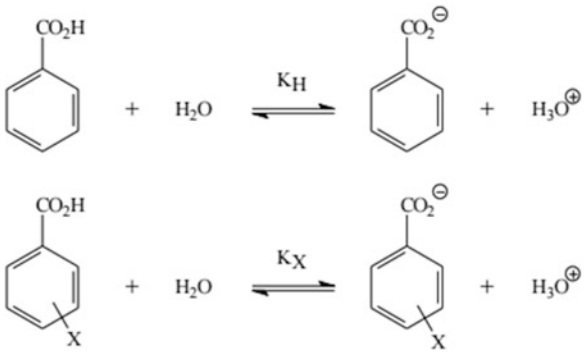

Fig. 2.4 The Hammett equation relates the relative magnitude of the equilibrium constants to a reaction constant ρ and a substituent constant σ Eq. 2.3

$$\sigma = \log\left(\frac{KX}{KH}\right) \tag{2.4}$$

The reaction constant is a measure of how sensitive a particular reaction is to changes in electronic effects of substituent groups (1–5). The reaction constant depends on the nature of the chemical reaction as well as the reaction conditions (solvent, temperature, etc.). Both the sign and magnitude of the reaction constant are indicative of the extent of charge buildup during the reaction progress. Reactions with $\rho > 0$ are favored by electron-withdrawing groups (i.e., the stabilization of negative charge). Those with $\rho < 0$ are favored by electron-donating groups (i.e., the stabilization of positive charge). The greater the magnitude of ρ, the more sensitive the reaction is to electronic substituent effects (Nantasenamat et al. 2009).

In 1956 Taft proposed an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds (Nantasenamat et al. 2009). In 1964 Hansch and Fujita put forward their linear Hansch equation using the contributions of Hammett and Taft that stood as a mechanistic basis of QSAR/QSPR development. Hansch et al. in late 1960s identified the nonlinear (parabolic) dependence of biological activity with log P and gave the following equation:

$$\log(1/C) = a\log P - b\left(\log P^2\right) + c \tag{2.5}$$

where $1/C$ = measure of biological activity, log P = log of octanol-water partition coefficient, and a, b, and c = regression coefficients (Nantasenamat et al. 2009; Corwin and Toshio 1964).

2.6.1 *Quantitative Structural Activity Relationship (QSAR)*

The discovery of clinically germane inhibitors is a challenging assignment, and the quantitative structural activity relationship (QSAR) methodology has become a very expedient and principally widespread technique for ligand-based drug design

and discovery. More than 1000 2D and 3D molecular descriptors are discovered and identified by the scientific community; a few are listed here such as Individual (Mol. Wt, Volume, H-AcceptorCount, H-DonorCount, RotatableBondCount, XlogP, slogp, smr, polarizabilityAHC, and polarizabilityAHP), Retention Index (chi), Atomic valence connectivity index (chiv), Path Count, Chi Chain, Chiv Chain, Chain Path Count, Cluster, Path Cluster, Kappa, Element Count, Dipole Moment, Electrostatic, Distance Based Topological, Estate Numbers, Estate Contributions, Information Theory Index, Semi Empirical, Hydrophobicity XlogpA, Hydrophobicity XlogpK, Hydrophobicity SlogpA, Hydrophobicity SlogpK, and Polar Surface Area (http://www.vlifesciences.com/support/QSAR_Descriptor_Definations_faqs_Answer.php).

2.6.1.1 Model Development

QSAR is among the most extensively used computational technique for ligand-based design, and Bohari et al. have recently reviewed the application of a variety of molecular descriptors like quantum chemical, molecular mechanics, conceptual density functional theory (DFT), and molecular docking-based descriptors for predicting biological activity (Bohari et al. 2011). A summary of relevant data analysis method, regression analysis, and model validation process is provided below along with some examples.

2.6.1.2 Data Analysis Method

Principal components analysis (PCA) and cluster analysis are two widely used methods in 2D and 3D QSAR data analysis. PCA was first invented by Karl Pearson in 1901 and is one of the most popular and primary data reduction techniques. PCA aims at data transformation from large multidimensions to low-dimensional representation, known as data reduction (Pearson 1901; http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture15.pdf). Cluster analysis technique is used to partition the data set (with typical molecular properties) into class and categories.

2.6.1.3 Regression Method

Regression analysis is a statistical process for estimating the relationships among dependent and independent variables by the use of modeling techniques implementing on several variables.

Partial least square (PLS) regression technique is used when the number of descriptors (independent variables) is greater than the number of compounds (data points) and/or there are any factors leading to correlations between variables (Martens and Naes 1989; Höskuldsson 1988; Eriksson et al. 2001).

Multiple linear regression (MLR) is an easily interpretable mathematical expression and primary method to construct QSAR/QSPR models, but it often fails in modeling highly correlated data sets. A few new methods have been developed using MLR such as best multiple linear regression (BMLR), heuristic method (HM), genetic algorithm-based multiple linear regression (GA-MLR), stepwise MLR, factor analysis MLR, and so on. Other methods such as self-learning and machine learning algorithms have also been developed to fit the data into the equations such as neural network (NN), support vector machine (SVM), and its variants: least square support vector machine (LS-SVM), grid search support vector machine (GS-SVM), potential support vector machine (P-SVM), and genetic algorithms support vector machine (GASVM) (Liu and Long 2009).

2.6.1.4 2D QSAR (Girgis et al. 2015)

Girgis and his team synthesized a total of 19 dispiro [3H-indole-3,2′-pyrrolidine-3′,3″-piperidines] (Fig. 2.5) of which 11–19 analogs were screened against HeLa (cervical). Compounds 13, 14, and 16 reveal higher potency (IC50 = 4.87, 5.75, and 7.25 μM, respectively) against HeLa (cervical) cell line than the standard reference cisplatin (IC50 = 7.71 μM) (clinically used against cervical carcinoma). See Table 2.2.

Structure–activity relationships (SAR) based on the experimental antitumor activity against HeLa (cervical carcinoma) reveal that the nature of the substituent attached to the phenyl group at C-4′ and consequently the exocyclic olefinic linkage seem to be a controlling factor governing the antitumor properties. Substitution of this phenyl group by fluorine atom enhances the observed antitumor properties more than two chlorine atoms, as exhibited in pairs 11, 13 (IC50 = 16.69, 4.87 μM, respectively) and 12, 14 (IC50 = 12.71, 5.75 μM, respectively) (Tables 2.3 and 2.4).

The basic idea behind QSAR is to generate a relationship between the chemical structure of an organic compound and its physiochemical properties. In the partial pharmacologically active data set mentioned in the present study, external data points were also considered. Spiro-alkaloids with similar scaffold are considered as an external data point and their biological activities were determined, but the same standard technique is earlier followed in the present study.

For the QSAR model development, compounds 11, 13, 15–17, and 19 were considered from Table 2.2 in addition to compounds 20–44 from Table 2.3. Thirty-one derivatives of spiro-alkaloids were used as a training set. The test set (external data set for validation) from synthesized analogs was considered representing high and low potent antitumor active agents 12, 14, and 18 (Table 2.2). Selected compounds geometry is optimized using molecular mechanics force field (MM^+), followed by a semiempirical AM1 method implemented in the Hyperchem. A total of 728 two-dimensional molecular descriptors were calculated using CODESSA-Pro software including constitutional, topological, geometrical, charge-related, semiempirical, molecular-type, atomic-type, and bond-type descriptors for the training set (Table 2.3) and test set (Table 2.4) data. Log property (1/log) and

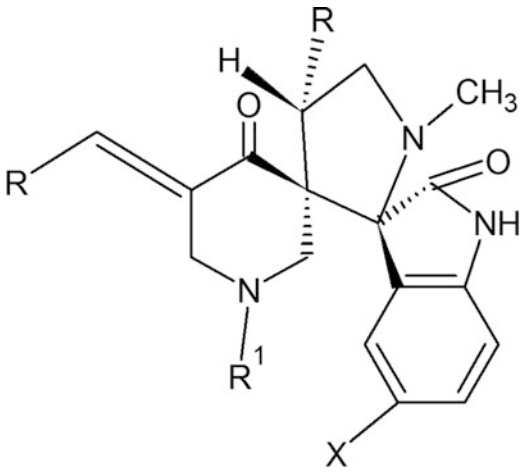

Fig. 2.5 Synthesized dispiro [3H-indole-3,20-pyrrolidine-30,300-piperidines] derivatives (Girgis et al. 2015)

Table 2.2 Antitumor properties of the synthesized compounds 11–19 (tested against HeLa)

No	*R*	R^1	*X*	IC50[a] at, μg/ml (μM) HeLa
11	2,4-$Cl_2C_6H_3$	Et	H	10.27 (16.69)
12	2,4-$Cl_2C_6H_3$	Et	Cl	8.26 (12.71)
13	4-FC_6H_4	Et	H	2.50 (4.87)
14	4-FC_6H_4	Et	Cl	3.15 (5.75)
15	2-Thienyl	Et	H	5.33 (10.89)
16	2-Thienyl	Et	Cl	3.80 (7.25)
17	3-Pyridinyl	Me	H	9.35 (20.08)
18	3-Pyridinyl	Et	H	5.16 (10.76)
19	3-Pyridinyl	Et	Cl	11.58 (22.53)
*	Doxorubicin hydrochloride	–	–	4.19 (7.22)
**	Cisplatin	–	–	4.19 (7.71)

[a]IC50 = concentration required to produce 50% inhibition of cell growth compared to control experimental data
Girgis et al. (2015)
* and ** stands for standard drug

biological activity/IC 50 value were considered for all the training and test sets against HeLa (cervical) cell lines of the training set compounds for the present QSAR modeling.

Best multi-linear regression (BMLR) was utilized which performs a stepwise search for the best *n*-parameter regression equations (where *n* stands for the number of descriptors used), based on the highest R^2 (squared correlation coefficient), $R_{cv}{}^2$OO (squared cross-validation "leave-one-out (LOO)" coefficient), $R_{cv}{}^2$MO (squared cross-validation "leave-many-out (LMO)" coefficient), Fisher statistical significance criteria (F) values, and standard deviation (S^2). Statistical characteristics of the QSAR models are presented in Table 2.5.

Table 2.3 Observed and predicated values of training set compounds 11, 13, 15–17, and 19–44 according to the multi-linear QSAR models

Entry	Comp	*R*	R^1	*X*	HeLa cervical cell line		
					Observed IC50 (μM)	Estimated IC50 (μM)	Error
1	11	2,4-Cl2C6H3	Et	H	16.69	12.26	4.43
2	13	4-FC6H4	Et	H	4.87	5.94	1.07
3	15	2-Thienyl	Et	H	10.89	10.48	0.41
4	16	2-Thienyl	Et	Cl	7.25	7.86	0.61
5	17	3-Pyridinyl	Me	H	20.08	26.07	5.99
6	19	3-Pyridinyl	Et	Cl	22.53	20.89	1.64
7	20	Ph	Me	H	6.21	5.92	0.29
8	21	Ph	Me	Cl	5.92	5.41	0.51
9	22	4-ClC6H4	Me	H	6.74	6.3	0.44
10	23	4-ClC6H4	Me	Cl	5.08	5.72	0.64
11	24	4-ClC6H4	Et	Cl	4.96	5.28	0.32
12	25	4-ClC6H4	Me	OMe	5.78	5.9	0.12
13	26	4-ClC6H4	Et	OMe	5.2	5.43	0.23
14	27	4-FC6H4	Me	H	6.51	5.95	0.56
15	28	4-FC6H4	Me	Cl	5.15	5.71	0.56
16	29	4-FC6H4	Me	OMe	5.44	6.21	0.77
17	30	4-H3CC6H4	Me	H	8.64	7.09	1.55
18	31	4-H3CC6H4	Me	Cl	6.65	6.71	0.06
19	32	4-H3CC6H4	Et	Cl	5.55	7.78	2.23
20	33	4-H3CC6H4	Me	OMe	6.96	7.68	0.72
21	34	4-H3COC6H4	Me	H	6.45	7.17	0.72
22	35	4-H3COC6H4	Et	H	7.22	6.54	0.68
23	36	4-H3COC6H4	Me	Cl	11.2	6.53	4.67
24	37	4-H3COC6H4	Et	Cl	8.74	6.27	2.47
25	38	4-H3COC6H4	Me	OMe	6.1	6.94	0.84
26	39	4-H3COC6H4	Et	OMe	5.51	7.84	2.33
27	40	4-Me2NC6H4	Me	Cl	24.36	20.24	4.12
28	41	2-Thienyl	Me	H	8.94	8.18	0.76
29	42	2-Thienyl	Me	Cl	6.86	7.98	1.12
30	43	2-Thienyl	Me	OMe	9.65	10.77	1.12
31	44	5-Methyl-2-furanyl	Me	Cl	9.88	8.46	1.42

Girgis et al. (2015)

Descriptors enlisted in the table are the chief contributors in the model development. Above all Min # HA and # HD molecular-type descriptor explaining the bioactive agent as hydrogen acceptor/donor is important in governing the QSAR model with its t-criterion (9.200) and minimum coefficient with (0.247). The second largest contributing molecular descriptor is FNSA-2 fractional PNSA (PNSA-2/TMSA), which is a charge-related descriptor with t-criterion (5.546)

Table 2.4 Observed and predicated values of external test set compounds 12, 14, and 18 according to the multi-linear QSAR models

Entry	Comp	R	R^1	X	HeLa cervical cell line		
					Observed IC50 (μM)	Estimated IC50 (μM)	Error
1	12	2,4-Cl2C6H3	Et	Cl	12.71	8.99	3.72
2	14	4-FC6H4	Et	Cl	5.75	5.64	0.11
3	18	3-Pyridinyl	Et	H	10.76	23.7	12.94

Girgis et al. (2015)

Table 2.5 Descriptor of the best multi-linear QSAR model for the HeLa (cervical) tumor cell line active agents

$N = 31$, $n = 3$, $R^2 = 0.815$, $R_{cv}{}^2\text{OO} = 0.738$, $R_{cv}{}^2\text{MO} = 0.776$, $F = 39.615$, $s^2 = 0.008$					
Entry	ID	Coefficient	s	T	Descriptor
1	0	0.141	0.185	0.763	Intercept
2	D1	0.247	0.027	9.2	Min.(#HA, #HD) (MOPAC PC)
3	D2	0.596	0.107	5.546	FNSA-2 fractional PNSA (PNSA-2/TMSA) (MOPAC PC)
4	D3	0.426	0.096	4.424	HASA-2/SQRT(TMSA) (Zefirov PC) (all)

Girgis et al. (2015)

and has the highest coefficient value of 0.596 controlling the QSAR model that is given by

$$\text{FNSA2} = \frac{\text{PNSA2}}{\text{TMSA}} \tag{2.6}$$

The third and last molecular descriptor of HeLa QSAR is depicted with t-criterion (4.424), and the second most effective parameter controlling the QSAR model based on its coefficient (0.426) is HASA-2/SQRT(TMSA), which is also a charge-related descriptor. The area-weighted surface charge of hydrogen-bonding acceptor atoms (HASA2) is determined by

$$\text{HASA2} = \sum_{A} \frac{q_A \sqrt{S_A}}{\sqrt{S_{tot}}} \quad A \epsilon X_{H-\text{acceptor}} \tag{2.7}$$

2.6.1.5 QSAR Model Validation

The reliability and statistical validity of QSAR model solely depend on the internal and external validation procedures. In the present QSAR model, the internal validation is assessed by CODESSA-Pro technique employing both leave one out (LOO) and leave many out (LMO). The observed correlations from the internal

validation are R_{cv}^2OO = 0.738 and R_{cv}^2MO = 0.776. The squared correlation coefficient of the designed QSAR model is $R^2 = 0.815$, the standard deviation of the regression is $S^2 = 0.008$, and the Fischer test value is $F = 39.615$ that reflects the ratio of the variance explained by the model and the variance due to their errors. The most potent synthesized analog 13, from the training set, exhibited an IC50 of 5.94 μM on the HeLa QSAR model with an experimental value of 4.87 μM and an error of 1.07. The other compounds from the training data set 16, 20–29, 31, 33–35, 38, and 42 relative to cisplatin standard reference clinically used against cervical carcinoma (IC50 = 7.71 μM) showed predicted experimental values with an error range of 0.06–1.12. Compounds 32 and 39 were considered potent analogs against cervical carcinoma (IC50 = 5.55, 5.51) and had predicted values (IC50 = 7.78, 7.84) with a greater error range of 2.23 and 2.33, respectively. Among the mild antitumor active agents against HeLa cell line, compounds 15, 30, 37, 41, 43, and 44 (IC50 range = 8.64–10.89 μM) revealed predicted potency (IC50 range = 6.27–10.77 μM) with a relatively larger error range (0.41–2.47) than the high potent analogs. Among the low potent analogs against HeLa cell lines, compounds 11, 17, 19, 36, and 40 (IC50 range = 11.20–24.36 μM) revealed large deviation in the predicted potency (IC50 range = 6.53–26.07 μM) with an error range of 1.64–5.99 (Table 2.5). From all the above statistical observations, the attained HeLa QSAR model can be considered a good predicative model to produce high potent HeLa antitumor hits compared to those of mild or low potency.

Compounds 12, 14, and 18 were selected for the purpose of validating and examining the predictive ability. The selected test set exhibited experimentally high or low potency against the tested cell line. Table 2.4 reveals the experimental and predicted IC50 values of the test set. Compound 14, considered as high potent against the HeLa cell line relative to the standard reference (cisplatin), had an experimental value of IC50 = 5.75 μM and a predicted value of IC50 = 5.64 μM with a minimum error of 0.11. However, compounds 12 and 18, considered low potent activity against HeLa cell line, had experimental values of IC50 = 12.71 and 10.76 μM and predicted IC50 values of 8.99 and 23.70 μM along with much greater error values of 3.72 and 12.94, respectively.

2.7 Pharmacokinetic and Pharmacodynamic (PKPD) Simulation (Nielsen and Friberg 2013)

Rowland and Tozer state in 2011 that pharmacokinetic (PK) has been defined as "how the body handles the drug" and pharmacodynamic (PD) has been defined as "how the drug affects the body." PK and PD are the vital mechanisms of the modern drug development process. Characterization of PKPD effectively suggests that the concentration that leads to desired effects and least side effects, with an appropriate dose regimen, can be computed.

2.7.1 Pharmacokinetics

Being a central part of clinical pharmacology, PK designates the link between drug dosing and drug concentration-time profile in the body. The determination of drug concentration (C) in plasma and its change from an initial concentration (C_0) with respect to time (t) is given by an exponential function:

$$C(t) = C_0 * e^{-k_e * t} \tag{2.8}$$

Equation 2.8 describes the single PK model with decline in concentration by single distribution phase. Considering the elimination rate for a given system, the change over the time points is directly proportional to the concentration or amount remaining in the system and elimination rate constant (k_e), which is of the first order and has a unit of per time (h^{-1}):

$$\frac{dc}{dt} = -k_e * C \tag{2.9}$$

where k_e is the parameter to be estimated based on the data and is inversely related to half-life ($t_{1/2}$) of the drug. From Eqs. 2.8 and 2.9, it follows that once k_e is known, the drug concentration can be predicted at any time point for a given C_0.

k_e is determined by the apparent volume of distribution (V_d) as well as clearance (CL) that describe the elimination capacity, which is typically governed by liver and kidney function. For a drug with immediate distribution and a CL value independent of concentration, k_e can be described as

$$k_e = \frac{CL}{V_d} \tag{2.10}$$

Often the nature of a drug is more complex because the distribution of the drug inside the body is not immediate due to the effect of its surrounding environment. Hence, the concentration-time course of drug distribution can be better explained by two or more compartments. The differential equations for a two-compartment model can be written as

$$\frac{dA_c}{dt} = -\frac{CL}{V_c} * A_c - \frac{Q}{V_c} * A_c + \frac{Q}{V_p} * A_p \tag{2.11}$$

$$\frac{dA_p}{dt} = -\frac{Q}{V_p}^{*} A_p + \frac{Q}{V_c} * A_C \tag{2.12}$$

where A_c and A_p are the amounts in the central and peripheral compartments and V_c and V_p are the corresponding volumes of distribution. Q represents intercompartmental clearance. An intravenously administered dose would be given into the central compartment.

The total exposure is often described as the area under the concentration-time curve (AUC). AUC is obtained by integrating the drug concentration-time profile and can also be computed as the systemically available dose over *CL*. The bioavailability, *F*, determines the fraction of an extravascular dose that reaches the systemic circulation and is thereby a measure of the extent of absorption. The rate of absorption is often characterized by a first-order rate constant, k_a.

2.7.2 Pharmacodynamics

Pharmacodynamics/PD designates the association among concentration and both the desired and undesirable effects by the given drug. The mathematical function describing the PKPD relationship is a sigmoidal. $E_{\max}$ model given by

$$E(t) = E_0 + \frac{E_{max} * C\ (t)^{\gamma}}{EC_{50}^{\gamma} + C\ (t)^{\gamma}} \tag{2.13}$$

where $E_{\max}$ is the maximum effect that can be achieved by the drug in the investigated system and EC50 is the drug concentration that results in half of the maximum effect. EC50 is inversely related to the potency. γ is the Hill or sigmoidicity factor that determines the steepness of the relationship but is in many cases not statistically significant from 1.

However, there are often situations where sufficiently high concentrations cannot be achieved to estimate $E_{\max}$, and simplifications can be made to estimate fewer parameters. When C « EC50, the $E_{\max}$ model collapses to a linear model ($\gamma = 1$) or a power function ($\gamma \neq 1$) with coefficient slope as shown below:

$$E\ (t) = E_0 + \text{Slope} * C(t)^{\gamma} \tag{2.14}$$

The underlying E_0 is not always constant over the study period. For example, the effect variable may vary because of an underlying disease, such as fluctuations in glucose in the event of diabetes or a diurnal rhythm in blood pressure.

2.8 Conclusion

Translational science in bioinformatics and drug discovery provides a powerful method especially when used as a tool within an armamentarium for discovering new target, drug leads, and novel approach in diagnostic and treatment for the betterment of society. Genomic technologies and NGS methods have proven to be the keystone of advanced research. The identification of genes' role in disease and disorder makes it possible to design personalized medicine approach, where a single or a few genes can be targeted or may act as a biomarker in the diagnosis

and treatment of disease and disorder. Data from public domain chemical libraries selected for appropriate target with structure-based and ligand-based discovery can create a very promising lead which may continue to clinical trials. Simulation study of pharmacokinetic and pharmacodynamic behavior of a chemical compound helps us estimate the concentration and dose value in computed form that can significantly reduce the overconcentration and dosing effects. As bioinformatics develops further, it is expected that genomics, proteomics, drug discovery, and computational power will continuously explode with new advances in therapeutic applications; new targets and leads may be brought to marketplace more rapidly each year.

References

Amy CA. The Process of structure-based drug design. Chem Biol. 2003;10:787–97. doi:10.1016/j.chembiol.2003.09.002.

An J, Totrov M, Abagyan R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics. 2005;4:752–61. doi:10.1074/mcp.M400159-MCP200.

Benjamin B. Next generation sequencing and translational research: from bench to bedside. 2015. http://www2.mlo-online.com/features/201208/lab-management/next-generation-sequencing-and-translational-research-from-bench-to-bedside.aspx. Accessed on 10 Sept 2015.

Berman HM. The protein data bank: a historical perspective. Acta Crystallogr Sect A Found Crystallogr. 2008;A64(1):88–95. doi:10.1107/S0108767307035623.

Boehm HJ, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, et al. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. J Med Chem. 2000;43:2664–74. doi:10.1021/jm000017s.

Bohari MH, Srivastava HK, Sastry GN. Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR Models. Org Med Chem Lett. 2011;1:3. doi:10.1186/2191-2858-1-3.

Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Chapter 12. In: Wheeler RA, Spellmeyer DC, editors. Annual reports in computational chemistry. Oxford: Elsevier; 2008. p. 217–41. doi:10.1016/S1574-1400(08)00012-1.

Brady GP, Jr Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des. 2000;14:383–401. doi:10.1023/A:1008124202956.

Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics. 2008;24:1733–4. doi:10.1093/bioinformatics/btn307.

Carlson B. Next generation sequencing: the next iteration of personalized medicine: next generation sequencing, along with expanding databases like the cancer genome atlas, has the potential to aid rational drug discovery and streamline clinical trials. Biotechnol Healthc. 2012;9(2):21–5.

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, et al. The Meta Cyc database of metabolic pathways and enzymes and the Bio Cyc collection of pathway/genome databases. Nucleic Acids Res. 2011;40:742–53. doi:10.1093/nar/gkr1014.

Corwin H, Toshio F. p-σ-π analysis. a method for the correlation of biological activity and chemical structure. J Am Chem Soc. 1964;86:1616–26. doi:10.1021/ja01062a035.

Couzin J. NIH dives into drug discovery. Science. 2003;302:218–21. doi:10.1126/science.302.5643.218.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2010;39:691–7. doi:10.1093/nar/gkq1018.

Crum-Brown A, Fraser TR. On the connection between chemical constitution and physiological action. Pt 1. On the physiological action of the salts of the ammonium bases, derived from Strychnia, Brucia, Thebia, Codeia, Morphia, and Nicotia. T Roy Soc Edin. 1868;25:151–203.

Dassault Systèmes BIOVIA. Discovery studio modeling environment, Release 4.5. San Diego: Dassault Systèmes; 2015.

Deepa PR, Vandhana S, Muthukumaran S, Umashankar V, Jayanthi U, Krishnakumar S. Chemical inhibition of fatty acid synthase: molecular docking analysis and biochemical validation in ocular cancer cells. J Ocul Biol Dis Infor. 2010;3:117–28. doi:10.1007/s12177-011-9065-7.

Dorsey BD, Levin RB, McDaniel SL, Vacca JP, Guare JP, Darke PL, et al. L-735,524: the design of a potent and orally available HIV protease inhibitor. J Med Chem. 1994;37:3443–51. doi:10.1021/jm00047a001.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. 2006;34(Web Server issue):W116–8. doi:10.1093/nar/gkl282.

Erickson J, Neidhart D, VanDrie J, Kempf D, Wang X, Norbeck D, et al. Design, activity and 2.8 A° crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. Science. 1990;249:527–33. doi:10.1126/science.2200122.

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Chapter 18, process analytical technology (PAT) and quality by design (QbD) multi- and megavariate data analysis: principles and applications. Umetrics: Umeå; 2001.

Fischer E. Einfluss der configuration auf die wirkung der enzyme. Ber Dt Chem Ges. 1894;27:2985–93.

Frederiksen KS, Wulff EM, Sauerberg P, Mogensen JP, Jeppesen L, Fleckner J. Prediction of PPAR-α ligand-mediated physiological changes using gene expression profiles. J Lipid Res. 2004;45:592–601. doi:10.1194/jlr.M300239-JLR200.

Girgis AS, Panda SS, Aziz MN, Steel PJ, Dennis Hall C, Katritzky AR. Rational design, synthesis, and 2D-QSAR study of anti-oncological alkaloids against hepatoma and cervical carcinoma. RSC Adv. 2015;5:28554–69. doi:10.1039/C4RA16663A.

Girke T, Cheng L-C, Raikhel N. ChemMine. A compound mining database for chemical genomics1. Plant Physiol. 2005;138:573–77. doi: http://dx.doi.org/10.1104/pp.105.062687

Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM. A method for localizing ligand binding pockets in protein structures. Proteins. 2006;62:479–88. doi:10.1002/prot.20769.

Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem. 1985;28:849–57. doi:10.1021/jm00145a002.

Halgren T. New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des. 2007;69:146–8. doi:10.1111/j.1747-0285.2007.00483.x.

Hammes GG. Multiple conformational changes in enzyme catalysis. Biochemistry. 2002;41(26):8221–8. doi:10.1021/bi0260839.

Hammett LP. Some relations between reaction rates and equilibrium constants. Chem Rev. 1935;17:125–36.

Hammett LP. The effect of structure upon the reactions of organic compounds. Benzene derivatives. J Am Chem Soc. 1937;59:96–103.

Höskuldsson A. PLS regression methods. J Chemomet. 1988;2:211–28. doi:10.1002/cem.1180020306.

http://depth-first.com/articles/2011/10/12/sixty-four-free-chemistry-databases/. Accessed on 25 Aug 2015.

http://www.vlifesciences.com/support/QSAR_Descriptor_Defination s_faqs_Answer.php. Accessed on 22 Sept 2015.

https://web.viu.ca/krogh/chem331/LFER%20Hammett%202012.pdf.
https://www.emolecules.com/. Accessed on 14 Sept 2015.
Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res. 1999;27(1):254–6. doi:10.1093/nar/27.1.254.
Ihlenfeldt WD, Voigt JH, Bienfait B, Oellien F, Nicklaus MC. Enhanced CACTVS browser of the open NCI database. J Chem Inf Comput Sci. 2002;42:46–57. doi:10.1021/ci010056s.
Irwin JJ, Shoichet BK. J Chem Inf Model. 2005;45(1):177–82. doi:10.1021/ci049714+.
Jin VX, Leu Y-W, Liyanarachchi S, Sun H, Fan M, Nephew KP, et al. Identifying estrogen receptor a target genes using integrated computational genomics and chromatin immunoprecipitation microarray. Nucleic Acids Res. 2004;32:6627–35. doi:10.1093/nar/gkh1005.
Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42–6. doi:10.1093/nar/30.1.42.
Kastenholz MA, Pastor M, Cruciani G, Haaksma EE, Fox T. GRID/CPCA: a new computational tool to design selective ligands. J Med Chem. 2000;43:3033–44. doi:10.1021/jm000934y.
Koshland Jr DE. Correlation of structure and function in enzyme action. Science. 1963;142:1533–41. doi:10.1126/science.142.3599.1533.
Kroemer RT. Structure-based drug design: docking and scoring. Curr Protein Pept Sci. 2007;8:312–28.
Kubinyi H. Success stories of computer-aided design. In: Ekins S, Wang B, editors. Computer applications in pharmaceutical research and development, Wiley series in drug discovery and development. New York: Wiley-Interscience; 2006. p. 377–424.
Kuhajda FP, Jenner K, Wood FD, Hennigar RA, Jacobs LB, Dick JD, et al. Fatty acid synthesis: a potential selective target for antineoplastic therapy. Proc Natl Acad Sci. 1994;91:6379–83. doi:10.1073/pnas.91.14.6379.
Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982;161:269–88. doi:10.1016/0022-2836(82)90153-X.
Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995;13:323–30. doi:10.1016/0263-7855(95)00073-9.
Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci. 1997;22(12):488–90. doi:10.1016/S0968-0004(97)01140-7.
Lecture 15: Principal component analysis. DOC493: intelligent data analysis and probabilistic inference lecture. http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture15.pdf. Accessed on 18 Aug 2015.
Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph. 1992;10:229–34. doi:10.1016/0263-7855(92)80074-N.
Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40(Database issue):D857–61. doi:10.1093/nar/gkr930.
Liu P, Long W. Current mathematical methods used in QSAR/QSPR studies. Int J Mol Sci. 2009;10:1978–98. doi:10.3390/ijms10051978.
Liu M, et al. Large-scale prediction of adverse drug reactions by integrating chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc. 2012;19(e1):e28–35. doi:http://dx.doi.org/10.1136/amiajnl-2011-000699.
Martens H, Naes T. Multivariate calibration. Chichester: Wiley; 1989.
McConkey BJ, Sobolev V, Edelman M. The performance of current methods in ligand-protein docking. Curr Sci. 2002;83:845–85.
Meng XY, Zheng QC, Zhang HX. A comparative analysis of binding sites between mouse CYP2C38 and CYP2C39 based on homology modeling, molecular dynamics simulation and

docking studies. Biochim Biophys Acta. 2009;1794:1066–72. doi:10.1016/j.bbapap.2009.03.021.
Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011;7:146–57. doi:10.2174/157340911795677602.
Meyer H. Zur Theorie der AIkoholnarkose. Arch Exp Pathol Pharm. 1899;42:109–18.
Mezei M. A new method for mapping macromolecular topography. J Mol Graph Model. 2003;21:463–72. doi:10.1016/S1093-3263(02)00203-6.
Miles T, Tomer A, Carol AF, Ron C, Markus K, Suzanne P, et al. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. Genome Biol. 2010;11(Suppl 1):O12. doi:10.1186/gb-2010-11-s1-o12.
Molecular docking, estimating free energies of binding, and AutoDock's semi-empirical force field – written by Sebastian Raschka July 26, 2014. http://sebastianraschka.com/Articles/2014_autodock_energycomps.html. Accessed on 28 Sept 2015.
Morini E, Sangiuolo F, Caporossi D, Novelli G, Amati F. Application of next generation sequencing for personalized medicine for sudden cardiac death. Front Genet. 2015;6:55. doi:10.3389/fgene.2015.00055.
Nantasenamat C, Isarankura-Na-Ayudhya C, Thanakorn Naenna T, Prachayasittikul VA. Practical overview of quantitative structure-activity relationship. EXCLI J. 2009;8:74–88.
Nielsen EI, Friberg LE. Pharmacokinetic-pharmacodynamic modeling of antibacterial drugs. Pharmacol Rev. 2013;65:1053–90. doi:10.1124/pr.111.005769.
Oreng CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure. 1997;5(8):1093–108. doi:10.1016/S0969-2126(97)00260-8.
Overton CE. Studien über die Narkose. Jena: Fischer; 1901.
Pappenberger G, Benz J, Gsell B, Hennig M, Ruf A, Stihle M, et al. Structure of the human fatty acid synthase KS-MAT didomain as a framework for inhibitor design. J Mol Biol. 2010;397:508–19. doi:10.1016/j.jmb.2010.01.066.
Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2:559–72.
Pemble CW, Johnson LC, Kridel SJ, Lowther WT. Crystal structure of the thioesterase domain of human fatty acid synthase inhibited by orlistat. Nat Struct Mol Biol. 2007;14:704–9. doi:10.1038/nsmb1265.
Portela C, Soares-da-Silva P. The translational approach between computational chemistry and clinical expertise in drug development. 2015. http://sigarra.up.pt/fmup/pt/publs_pesquisa.show_publ_file?pct_gdoc_id=42752. Accessed on 15 Aug 2015.
Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available PubChem BioAssay data. Clin Pharmacol Ther. 2011;90:90–9. doi:10.1038/clpt.2011.81.
Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol. 1996;261:470–89. doi:10.1006/jmbi.1996.0477.
Raza M. A role for physicians in ethnopharmacology and drug discovery. J Ethnopharm. 2006;104:297–301. doi:10.1016/j.jep.2006.01.007.
Richet MC. Note sur le rapport entre la toxicité et les propriétes physiques des corps. Compt Rend Soc Biol (Paris). 1893;45:775–6.
Roberts N, Martin J, Kinchington D, Broadhurst A, Craig J, Duncan I, et al. Rational design of peptide-based HIV proteinase inhibitors. Science. 1990;248:358–61. doi:10.1126/science.2183354.
Ruth H, Morris GM, Olson AJ, Goodsell DS. A semiempirical free energy force field with charge-based desolvation. J Comput Chem. 2007;28:1145–52. doi:10.1002/jcc.20634.
Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. Nucleic Acids Res. 2011;39:D670–6. doi:10.1093/nar/gkq1089.
Shah NH. Survey: translational bioinformatics embraces big data. Yearb Med Inform. 2012;7:130–4.

Shah NH, Tenenbaum JD. The coming age of data-driven medicine: translational bioinformatics' next frontier. J Am Med Inform Assoc. 2012;19:e2–4. doi:10.1136/amiajnl-2012-000969.

Steven JK, Fumiko A, Natasha R, Jeffrey WS. Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity. Cancer Res. 2004;64:2070–5. doi:10.1158/0008-5472.CAN-03-3645.

SYBYL-X-SuiteS: YBYL 8.0. Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

Tavera-Mendoza LE, Mader S, White JH. Genome-wide approaches for identification of nuclear receptor target genes. Nucl Recept Signal. 2006;4:e018. doi:10.1621/nrs.04018.

Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. 2003;31:334–41. doi:10.1093/nar/gkg115.

Thomsen R, Christensen MH. MolDock: a new technique for high-accuracy molecular docking. J Med Chem. 2006;49:3315–21. doi:10.1021/jm051197e.

Vandhana S, Deepa PR, Jayanthi U, Biswas J, Krishnakumar S. Clinico-pathological correlations of fatty acid synthase expression in retinoblastoma: an Indian cohort study. Exp Mol Pathol. 2011;90:29–37. doi:10.1016/j.yexmp.2010.11.007.

Venhorst J, ter Laak AM, Commandeur JN, Funae Y, Hiroi T, Vermeulen NP. Homology modeling of rat and human cytochrome P450 2D (CYP2D) isoforms and computational rationalization of experimental ligand-binding specificities. J Med Chem. 2003;46:74–86. doi:10.1021/jm0209578.

Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc. 2011;18 (Suppl 1):i73–80. doi:http://dx.doi.org/10.1136/amiajnl-2011-000417 i73-i80.

VLifeMDS: Molecular Design Suite, VLife Sciences Technologies Pvt. Ltd., Pune, India, 2010 (www.vlifesciences.com)

Voigt JH, Bienfait B, Wang S, Nicklaus MC. Comparison of the NCI open database with seven large chemical structural databases. J Chem Inf Comput Sci. 2001;41:702–12. doi:10.1021/ci000150t.

Williams PA, Cosme J, Ward A, Angove HC, Matak Vinkovic D, Jhoti H. Crystal structure of human cytochrome P450 2C9 with bound warfarin. Nature. 2003;424:464–8. doi:10.1038/nature01862.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(suppl 1):D668–72. doi:10.1093/nar/gkj067.

Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY. MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol. 2005;138(1):27–37. doi:http://dx.doi.org/10.1104/pp.105.060376.