Translational Medicine Research

Series Editors: Zhu Chen · Xiaoming Shen
Saijuan Chen · Kerong Dai

Dong-Qing Wei
Yilong Ma
William C.S. Cho
Qin Xu
Fengfeng Zhou *Editors*

Translational Bioinformatics and Its Application

上海交通大學出版社
Shanghai Jiao Tong University Press

Springer

Translational Medicine Research

Translational medicine converts promising laboratory discoveries into clinical applications and elucidates clinical questions with the use of bench work, aiming to facilitate the prediction, prevention, diagnosis and treatment of diseases. The development of translational medicine will accelerate disease control and the process of finding solutions to key health problems. It is a multidisciplinary endeavor that integrates research from the medical sciences, basic sciences and social sciences, with the aim of optimizing patient care and preventive measures that may extend beyond health care services. Therefore, close and international collaboration between all parties involved is essential to the advancement of translational medicine. To enhance the aforementioned international collaboration as well as to provide a forum for communication and cross-pollenation between basic, translational and clinical research practitioners from all relevant established and emerging disciplines, the book series "Translational Medicine Research" features original and observational investigations in the broad fields of laboratory, clinical and public health research, aiming to provide practical and up-to-date information on significant research from all subspecialties of medicine and to broaden readers' horizons, from bench to bed and bed to bench. Produced in close collaboration with National Infrastructures for Translational Medicine (Shanghai), the largest translational medicine research center in China, the book series offers a state-of-the-art resource for physicians and researchers alike who are interested in the rapidly evolving field of translational medicine. Prof. Zhu Chen, the Editor-in-Chief of the series, is a hematologist at Shanghai Jiao Tong University, China's former Minister of Health, and chairman of the center's scientific advisory board.

More information about this series at http://www.springer.com/series/13024

Dong-Qing Wei • Yilong Ma • William C.S. Cho
Qin Xu • Fengfeng Zhou
Editors

Translational Bioinformatics and Its Application

上海交通大學出版社
Shanghai Jiao Tong University Press

Springer

Editors
Dong-Qing Wei
State Key Laboratory of Microbial Metabolism and School of Life Sciences and Biotechnology
Shanghai Jiao Tong University
Shanghai, China

William C.S. Cho
Department of Clinical Oncology
Queen Elizabeth Hospital
Kowloon, Hong Kong, China

Fengfeng Zhou
College of Computer Science & Technology
Jilin University
Changchun, Jilin, China

Yilong Ma
Center for Neurosciences
The Feinstein Institute for Medical Research
New York, USA

Department of Molecular Medicine
Hofstra Northwell School of Medicine
New York, USA

Qin Xu
State Key Laboratory of Microbial Metabolism and School of Life Sciences and Biotechnology
Shanghai Jiao Tong University
Shanghai, China

ISSN 2451-991X ISSN 2451-9928 (electronic)
Translational Medicine Research
ISBN 978-94-024-1043-3 ISBN 978-94-024-1045-7 (eBook)
DOI 10.1007/978-94-024-1045-7

The print edition is not for sale in China Mainland. Customers from China Mainland please order the print book from: Shanghai Jiao Tong University Press.

Library of Congress Control Number: 2017932746

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media B.V.
The registered company address is: Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands

Preface

It was May 2015 when I was invited to join the editorial team of the "Translational Medicine Publication Project." I proposed to edit a book entitled *Translational Bioinformatics (TBI)*. I was happy to have invited a few colleagues from China and the USA who are experts in the field to join me as coeditors, Profs. Yilong Ma, William C.S. Cho, and Fengfeng Zhou. Prof. Qin Xu from my research team and my PhD student Huiyuan Zhang spent much time in managing the project. It has been many years since I started to collaborate with Springer. Our proposal was approved quickly as a collaboration project with the Shanghai Jiao Tong University Press.

TBI is an emerging field in the study of health informatics, focused on the convergence of molecular bioinformatics, biostatistics, statistical genetics, medical imaging, and clinical or medical informatics. Its focus is on applying sound informatics methodology to the increasing amount of biomedical and genomic data to formulate knowledge, disease models, and medical tools, which can be utilized by scientists, clinicians, and patients. TBI employs data mining and analytical biomedical informatics in order to generate clinical knowledge for a wide array of applications. Furthermore, it involves cross-disciplinary biomedical research to improve human health through the use of computer-based information systems. This new field has achieved great success in the recent decade by synergic integration of the molecular and genetic footprints in tissue cultures, animal models, and patients with various diseases.

Our book tries to cover, but not limited to, the following topics:

Biomedical knowledge integration
Data-driven view of disease biology
Biological knowledge assembly and interpretation
Human microbiome analysis
Pharmacogenomics
Mining electronic health records in the genomics era
Small molecules and disease

Protein interactions and disease
Network biology approach to complex diseases
Structural variation and medical genomics
Analyses using disease ontologies
Mining genome-wide genetic markers
Genome-wide association studies
Cancer genome analysis
Medical bioinformatics: biomarkers and medical imaging
Neuroinformatics of neurological and psychiatric disorders
Neuroimaging genetics

It is a challenging task that these topics are quite diversified and involved scientists with various expertise. Finally, we tried our best to summarize these diverse topics into five Parts, as in the Contents, with the chapters 2, 6, 10, 14, 16 and 17 edited by Yilong Ma, the chapters 3, 8, 11 and 13 edited by William C.S. Cho, the chapters 5, 6 and 7 edited by Qin Xu, as well as the chapters 1, 4, 9, 11, 12, 14, 15 and 16 edited by Fengfeng Zhou. My assistants Mrs. Ruili Zhao and Ms. Qiuyuan Hu made great efforts in soliciting manuscripts. Mrs. Becky Jinan Zhao from Springer and Mrs. Min Xu and Zhufeng Zhou from the Shanghai Jiao Tong University Press give us a lot of help in formulating this book and applying for funding.

In 2015, we enter the era of "precision medicine," which integrates two major contemporary developments including various omics (e.g., genomics, proteomics and metabolomics) and Big Data. I believe the TBI would play an important role in the endeavor for precision and personalized medicine.

Shanghai, China Dong-Qing Wei
2016-11-13

Contents

Part I Computer-Aided Drug Discovery

Part II Protein-Ligand Interactions and Drug Development

Part III Omics for Precision Medicine

Part IV Biostatistics, Bioinformatics, and System Biology Approaches to Complex Diseases

Part V Bioimaging and Other Applications of Informatics Techniques in Translational Medicine

Part I
Computer-Aided Drug Discovery

Chapter 1
Drug Discovery

Geetha Ramakrishnan

Abstract An understanding of the process of drug discovery is necessary for the development of new drugs and put into clinical practice, to alleviate the diseases prevalent in modern era. This chapter covers the basic principles of how new drugs can be discovered with emphasis on target identification, lead optimization based on computer-aided drug design methods and clinical trials. The drug design principles in the pharmaceutical industry are explained based on the target and chosen ligand using molecular docking, pharmacophore modelling and virtual screening methods. The drug design is illustrated with specific examples. The clinical trials are necessary to introduce the drugs into market after due validation.

Keywords Lead compound • Computer-aided drug design • Molecular docking • Scoring functions • Virtual screening • Pharmacophore modelling • Quantitative structure-activity relationship (QSAR) • Clinical trials

1.1 Introduction

Drug discovery process deals with the root cause of the disease finding relevant genetic/biological components (i.e. drug targets) to discover lead compounds. Currently specialists in various fields, such as medicine, biochemistry, chemistry, computerized molecular modelling, pharmacology, microbiology, toxicology, physiology and pathology, contribute their research capability to achieve this goal**.** The drug discovery process (Fig. 1.1) in general is divided into three parts, namely, target identification, lead discovery and clinical trials.

The target identification will normally require a detailed assessment of the pathology of the disease and in some cases basic biochemical research such as study of the basic processes of life, body biochemistry and the use of metabolic analogues; study and exploitation of differences in molecular biology, differential

G. Ramakrishnan (✉)
Department of Chemistry, Sathyabama University, Rajiv Gandhi Salai, Chennai 600 119, Tamil Nadu, India
e-mail: icget2011@gmail.com

D.-Q. Wei et al. (eds.), *Translational Bioinformatics and Its Application*,
Translational Medicine Research, DOI 10.1007/978-94-024-1045-7_1

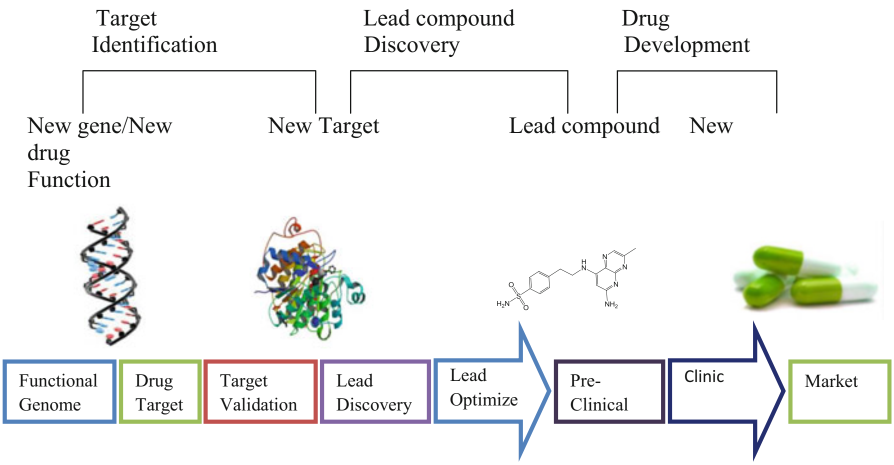


Fig. 1.1 Drug discovery process

cytology, biochemistry and endocrinology; and study of the biochemistry of diseases which will be necessary before initiating a drug design investigation.

The *lead compound* design is the most decisive step in the process of drug discovery. Methods used in lead compound design include folk/ethno-pharmacy and therapeutics, massive pharmacological screening, modification of bioactive natural products, exploitation of secondary or side effects of drugs, an approach through the molecular mechanism of drug action, drug metabolism and chemical delivery systems (Drews 1999, Bodor 1982, 1987). Numerous methods have been invented for the quantification of electronic, hydrophobic and steric effects of functional groups (Franke 1984). Statistical methods, pattern recognition/principal components analysis and cluster analysis can lead to the prediction and optimization of activity and ultimately to the design of newer drugs.

The structure of the proposed lead compound allows the medicinal/organic chemist to prepare the sample by synthetic route, and the lead compound undergoes initial pharmacological and toxicological testing. The selected lead compounds are given to animals for preclinical trials. When the lead compound has been found to be effective and safe in animal testing, it is used for human clinical trials. The lead compound is required to pass three phase clinical trials in human beings. In phase I, studies on healthy subjects are conducted to confirm safety. In phase II, studies are conducted on patients to confirm efficacy. Finally in phase III, large studies on patients are conducted to gather information about safety and efficacy at the population level.

The results of these tests enable the team to decide whether it is profitable to continue development by preparing a series of analogues, measure their activity and correlate the results to determine the drug with optimum activity.

Because of the strict prerequisites of drug authorities, which are becoming ever more demanding, the cost of drug discovery is steadily increasing. Thus, rational drug design becomes the main objective of medicinal chemistry today. Based on rational design, new structures can be developed with a high probability of possessing the required properties and biological activity.

1.1.1 Need for Drug Design

Drug discovery is a time-consuming and costly process. The process takes 12–15 years to release a new drug into market, and average cost for the development of a new drug is about 600–800 million dollars (Adams and Brantner 2006). Among 10,000 drugs that are applied on animals, only ten of them are tested for human clinical trials, in which one or two of the drugs only are put into the market (Hughes 2009). In order to reduce the research timeline and cost, various computational methods were used. The computer-aided drug design process is fast, automatic and less expensive with high success rate and fruitful with respect to intellectual property rights. The problems encountered for this procedure with possible solutions (Kubinyi 1999) are given in Table 1.1.

The strategies to be followed in the drug design include structure-based design of ligands with affinity and selectivity using molecular docking, virtual screening of favourable drug properties and bioavailability and pharmacophore modelling.

1.2 Target Identification

This process involves identification of relevant molecular target based on the known pathology of the disease due to an enzyme, receptor, ion channel or transporter. The next step is to determine the responsible DNA and protein sequence with their function and its mechanism of action (Ryan et al. 2000; Silverman 2004). The mechanism of action can be obtained by the earlier study done on animals as proof and a suitable choice for the target from earlier investigations. Based on the mechanism of drug action, the associated disease and status of the drug are given in Table 1.2.

1.3 Computer-Aided Drug Design

Computer-aided drug design (CADD) is a specialized discipline that uses computational knowledge-based methods to aid the drug discovery process. It is estimated that the computational methods could save up to 2–3 years and $300 million (Price waterhouse coopers 2005). There are several areas where CADD plays an important role in the traditional drug discovery. Genomics and bioinformatics support genetic methods of target identification and validation. Cheminformatics enables

Table 1.1 Problems faced by drug industry with its possible solutions

Sl.No.	Problems	Possible solutions
1.	Target search	Genome information
2.	Target validation	Knockouts, RNA silencing
3.	Lead search	In vitro test models, high-throughput screening
4.	Lead optimization	Parallel syntheses, chemogenomics
5.	Absorption, permeability	Lipinski rules, Caco cells, prodrugs
6.	Metabolism	Liver microsomes
7.	Toxicity	Ames test, hERG models
8.	Drug-drug interactions	CYP inhibition/induction

Table 1.2 Targets with their mechanism, associated disease and status of the drug

Sl. No.	Drug targets	Mechanisms of drug action	Disease	Status of the drug
1.	Enzymes	Reversible and irreversible inhibitors		
	Angiotensin-converting enzyme	Renin-Ang system	Hypertension	Launched
	Tryptase	Phagocytosis	Inflammation, asthma	Clinical phase III
	Cathepsin K	Bone resorption	Osteoporosis	Clinical phase I
2.	Receptors	Agonists and antagonists	Chronic pain	Dopamine, epinephrine, morphine-known drugs
3.	Ion channels	Blocker and opener Ca^{+2}, Na^+ and K^+ channel blockers, K^+ channel openers	Renal Problems	Cyclosporine – launched
4.	Transporters	Uptake inhibitors	H^+/K^+-ATPase (proton pump)	Omeprazole – as known drug
5.	DNA	Alkylating agents, minor groove binders, intercalating agents	DNA duplication, tumours	Distamycin A, netropsin as known drugs

researchers to process virtual screening for selection of lead compounds for synthesis and screening. This allows researchers to make fast decision on lead compound identification and optimization. In silico ADMET (absorption, distribution, metabolism, excretion and toxicology) modelling aids researchers to identify a bioavailable drug with suitable drug metabolism properties.

CADD methods offer significant benefits for drug discovery. One of them is time and cost savings for lead identification, optimization and ADMET predictions for implementing experimental research. Only the most promising drug candidates will be tested based on the results of CADD. CADD provides deep insight to drug-receptor interactions. Molecular models of drug compounds can reveal intrinsic, atomic scale binding properties that are difficult to envisage. It is classified as structure-based drug design and ligand-based drug design.

1.3.1 Structure-Based Drug Design (SBDD)

The preliminary step in structure-based drug design is to determine the three-dimensional structure of a target molecule (usually protein). This can be achieved by X-ray crystallography or NMR spectroscopy experiments or by approximated computational methods such as comparative modelling (homology modelling uses previously solvated structure as starting point to determine the three-dimensional structure of protein) and ab initio modelling (this method seeks to build three-dimensional protein models based on physical principles rather than previously solved model). The next step in this process is to identify the location of the binding site of a target molecule (receptor). The actual binding site can be located by comparing with known protein-ligand complexes or homology comparisons to related complexes. With well-defined binding site, a ligand (lead) can be determined. Usually, leads can be determined either through de novo design or through large database search for a molecule that matches the binding site. Docking methods are then used to evaluate the quality of ligand.

The molecular docking process mainly involves three steps:

Characterizing the binding site
Positioning the ligand into the binding site
Evaluating the strength of interaction for a specific ligand-receptor complex

Structure-based drug design includes molecular docking methods as a main tool, and certain researchers employ molecular dynamics also, if drug action is known.

1.3.1.1 Molecular Docking

When the structure of protein and its binding site are available, molecular docking techniques are used to identify lead compound. This technique is also used in lead optimization, when modification to known active molecule structure can quickly be tested by CADD before compound synthesis.

Molecular docking is useful in the identification of low-energy binding mode of a molecule or ligand in the active binding site of protein or receptor. A molecule or ligand which binds strongly through hydrogen bonds, van der Waal bonds or any possible electrostatic attractions with receptor or protein associated with disease may inhibit the function and thus acts as a drug. Hydrogen bonds are local electrostatic interaction between the atoms which plays a significant role in recognition of ligand binding with the target. Calculating the accurate protein-ligand interactions is the key principle behind structure-based drug discovery (Cramer et al. 1988).

1.3.1.2 Types of Docking

Three options for docking are available.

Rigid docking – where a suitable position for the ligand in receptor environment is obtained while maintaining its rigidity

Flexible docking – where a favoured geometry for receptor-ligand interaction is obtained by changing internal torsions of ligand into the active site while receptor remains fixed

Full flexible docking – where the ligand is freely rotated via its torsion angles and the side chain of active site residues (selected active site residues within a user-specified radius around the ligand) is freely rotatable.

Most of the docking methods used at the present moment in academic and industrial research employ a rigid target/protein. The algorithms used in docking are given in Appendix I.

The two components of molecular docking are:

(i) Prediction of binding conformation of the ligand in the binding site
(ii) Binding free energy prediction of the ligand (Leach A.R. and Gillet V.J., 2003)

1.3.1.3 Scoring Functions

There are mathematical methods used to predict the strength of the non-covalent interaction called binding affinity between two molecules after docking. The scoring functions have also been developed to predict the intermolecular interaction between two proteins, protein-DNA and protein-drug. The objective of any scoring function is to estimate the free energy change of binding for a ligand in a given binding pose. This can be expressed by the fundamental thermodynamic Eq. (1.1):

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where ΔG is the free energy change of binding, ΔH is the enthalpy change, T is the temperature of the system in Kelvin and ΔS is the entropy change.

Scoring functions are categorized into (i) force field and (ii) empirical (Stahl and Rarey 2001; Perola et al. 2004) (Table 1.3).

Force field scoring functions rely on the molecular mechanics methods. In this method it calculates both the protein-ligand interaction energy and ligand internal energy by van der Waals energy and electrostatic interactions. Advantages of force field-based scoring functions include accounting of solvent, and disadvantages include overestimation of binding affinity and arbitrarily choosing of non-bonded cutoff terms (Kitchen et al. 2004; Moitessier et al. 2008).

Empirical scoring functions – Empirical scoring functions weigh contributions from the different energetic terms in order to make a binding affinity prediction. These terms may include hydrogen bonding using geometric measures as well as force field-based physical potentials. However, the linear weighing of the terms is

Table 1.3 Major docking tools utilized in industrial and academic research institutes

Docking tool	Algorithm/method (Appendix I)	Scoring function
FlexX	Incremental construction	Boehm empirical scoring function
FlexX-Pharm	Incremental construction	Boehm empirical scoring function
Auto Dock	Genetic algorithm	Force filed-based empirical scoring
Dock	Incremental construction	Force filed-based scoring
ICM	Simulated annealing	Force filed-based scoring
GOLD	Genetic algorithm	Empirical knowledge-based scoring
Surflex-Dock	Incremental construction	Empirical Hammerhead scoring
Glide	Simulated annealing/incremental search	Empirical knowledge-based scoring
LigandFit	Shape matching	Empirical knowledge-based scoring

derived from regression methods that fit binding affinity terms to experimental affinities using experimental data and structural information (Teramoto and Fukunishi 2007).

1.3.1.4 Limitations and Challenges

Some key challenges in molecular docking and scoring are discussed based on protein flexibility and role of solvent and scoring function.

Protein flexibility: Docking programmes usually use protein as rigid and ligand as flexible; in this case receptor has one conformation, while the ligands have different conformations. The fundamental goal of virtual screening is to identify molecules with the proper complement of shape, hydrogen bonding and electrostatic and hydrophobic interactions for the target receptor; the complexity of the problem is far greater in reality. For example, the ligand and receptor may exist in different conformations when in free solution, which is different from the conformation when ligand is bound to protein (Koh 2003).

Role of solvent and scoring function: Protein and ligands are surrounded by solvent molecules, usually water. If the water mediation is ignored during docking, then the calculated interaction energy may be low, and favourable interactions with water may be lost (Moitessier et al. 2008). Several methods are now available to predict the binding energy accurately by accounting entropic and solvation effects (Reynolds et al. 1992; Zhang et al. 2001). These methods need greater amount of computational time and inappropriate to use in screening large databases. The molecular docking process is shown in Fig. 1.2.

1.3.2 Ligand-Based Drug Design (LBDD)

The ligand-based drug design starts with a database containing set of ligands with known activity interaction with the same receptor. The first step in this process is to

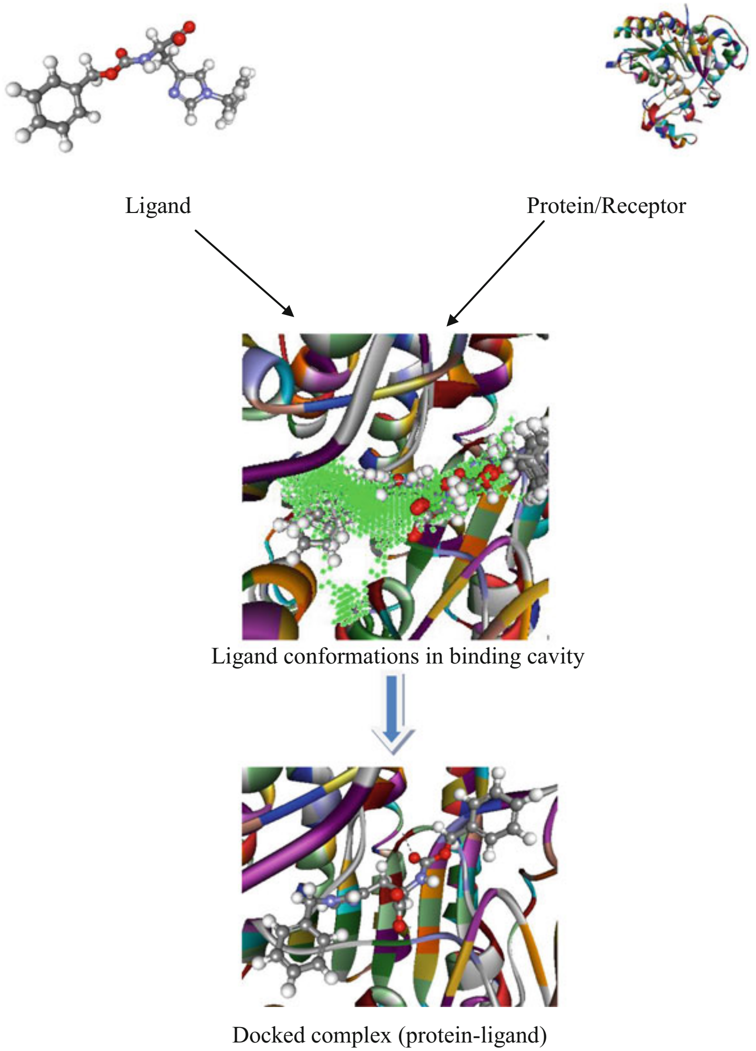

Ligand
Protein/Receptor
Ligand conformations in binding cavity
Docked complex (protein-ligand)

Fig. 1.2 Molecular docking flow chart using a benzamide derivative (MS-275) with HDAC2 protein (Naresh Kandakatla and Geetha Ramakrishnan 2014a, b)

divide the set of ligands into training and test set, and the second step in this process is molecular modelling. Ligand-based approach commonly considers descriptors based on chemistry, shape and electrostatic and interaction points (e.g. pharmacophore points) to assess similarity. A pharmacophore is an explicit geometric hypothesis of the critical features of a ligand (Leach and Gillet 2003). Features usually include hydrogen-bond donors and acceptors, charged groups and hydrophobic patterns. The hypothesis can be used to screen databases for candidate compounds and also can be used to refine existing leads. Another method in ligand-based drug design is quantitative structure-activity relationship (QSAR) modelling method and used for identifying a lead molecule and optimization. The concept of QSAR is based on the fact that the biological properties of a compound can be expressed as functions of its physicochemical parameters. The goal of the QSAR model is to predict the activity of the new molecules (optimized leads). The third step in ligand-based design involves identification of the most promising molecule as lead compound for further experimental investigation.

1.3.2.1 Pharmacophore Modelling

A *pharmacophore* describes a set of interactions required to bind given receptor. The pharmacophore is usually derived from three-dimensional computed conformations of a molecule and is an abstract representation of the molecule.

Common pharmacophore feature types are hydrophobic, hydrogen-bond acceptor, hydrogen-bond donor, aromatic rings and positively ionizable and negatively ionizable groups. The pharmacophore features describe the target binding site, e.g. a hydrophobic feature corresponds to hydrophobic region in the protein and hydrogen-bond acceptor feature as hydrogen bond donating counterpart in the protein. Hydrogen-bond acceptor and donor features usually have direction as parameter. The spatial relationship between the pharmacophore features is defined by interpoint distances between the features.

Pharmacophore modelling is widely used in drug design for identifying novel scaffolds or leads for various targets. Pharmacophore model is classified into two categories as (i) structure-based pharmacophore modelling and (ii) ligand-based pharmacophore modelling.

Structure-Based Pharmacophore Modelling

Structure-based pharmacophore modelling uses a 3D structure of protein co-crystallized with ligand or 3D structure of protein. The structure-based pharmacophore model is further subdivided into two types as protein-ligand complex and protein/receptor without ligand contribution. The protein-ligand-based approach locates the ligand binding sites of the protein target and determines the key interaction points between the protein and ligand. Automated tools for the jobs are LigandScot, Pocket v.2 and GBPM (Wolber and Langer 2005; Chen and Lai 2006; Ortuso et al. 2006). For protein-based approach, Discovery Studio (LUDI) was employed, where LUDI converts the interaction points in the binding site into catalyst pharmacophore features such as H-bond acceptors, H-bond donors and hydrophobe (Bohm 1992). In general structure-based pharmacophore, the generated interaction points consist of a large number of unprioritized pharmacophore features, which complicate further virtual screening process. To overcome this problem, a fast knowledge-based approach, hotspot-guided receptor-based pharmacophores (HS-Pharm) and Apo protein-based approach were used. Hotspot analysis is employed to identify the binding sites, where the ligand forms strong interactions (Barillari et al. 2008). In the second approach, the binding cavity embedded in a GRID and molecular interaction fields of GRID node and protein is calculated using a set of probes; the minimum energy found can be converted into pharmacophore feature (Tintori 2008; Goodford 1985).

Ligand-Based Pharmacophore Modelling

Ligand-based pharmacophore modelling is a key computational strategy in drug discovery in the absence of 3D structure of protein. Pharmacophore model generation extracts common chemical feature from a set of known molecules (usually training set) as a representative of essential interaction between the ligand and target protein of interest. This method involves two steps: the first step involves conformational analysis of training set molecules that allows conformational flexibility of each molecule, and the second step is alignment – aligning of training set molecules to determine the essential common chemical feature to construct pharmacophore models. Currently various commercial and academic computational softwares are available for pharmacophore model development – such as Hip Hop (Barnum et al. 1996), HypoGen (Li et al. 2000) (Accelrys Inc., http://www.accelrys.com), PHASE (Dixon et al. 2006) (Schordinger Inc., http://www.schrodinger.com), MOE (Chemical Computing Group, http://www.chemcomp.com), DISCO (Martin 2000), GASP (Jones and Willet 2000) and GALAHAD (Tripos Inc., http://www.tripos.com). Challenges to overcome are conformational ligand flexibility and molecular alignment. Conformational ligand flexibility problem is solved by computing multiple conformers for each molecule and creating a database. The second method is on-the-fly method, in which the conformational analysis is carried out in the pharmacophore modelling process; it does not need mass storage but requires higher CPU time (Poptodorov et al. 2006). A good conformer should satisfy low-energy configuration which interacts with the receptor. Molecular alignment is another challenging issue in ligand-based pharmacophore modelling. Alignment method can be classified into two categories as point-based and property-based approaches (Wolber et al. 2008). In point-based approach, pair of atoms or fragments or chemical feature points is superimposed using least square fitting. The biggest problem in this approach is to identify anchor points in dissimilar ligands. Property-based approach makes use of molecular descriptors to generate alignment.

Once pharmacophore model is generated, it can be used for virtual screening of small or large databases. Many tools such as ligand-based pharmacophore mapping, search 3D database (Accelrys Inc., http://www.accelrys.com), PHASE (Schordinger Inc., http://www.schrodinger.com), ChemDBS (VLife MDS., http://www.vlifesciences.com/), etc. are available for virtual screening. The full framework of pharmacophore modelling is illustrated in Fig. 1.3.

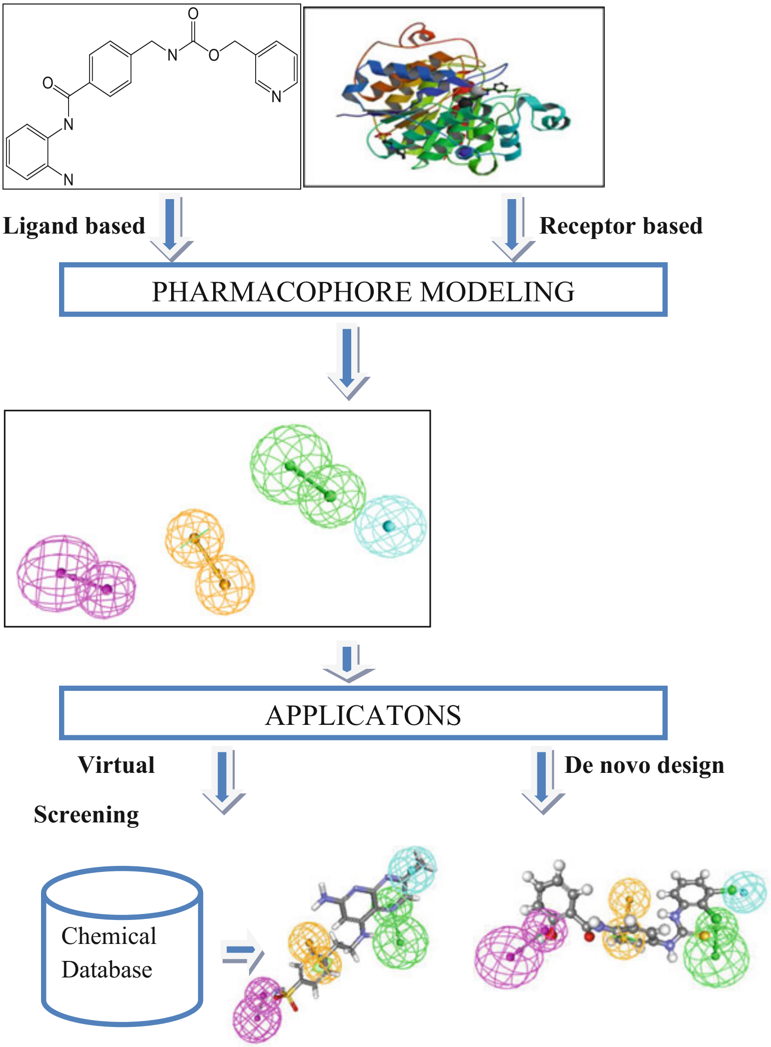


Fig. 1.3 The full framework of pharmacophore modelling

1.3.2.2 Virtual Screening

In silico screening of chemical compound database for identification of novel chemotype is termed as virtual screening. Virtual screening is generally performed on the commercial, public or privately available 2D/3D chemical structural databases. Virtual screening is employed to reduce the number of compounds to be tested in experimental laboratories, thereby focussing on more reliable entities for lead discovery and lead optimization (Rester 2008). The costs and time associated with virtual screening of chemical compounds are significantly lower when compared to screening of compounds in experimental laboratories. Thus virtual screening reduces the size of the haystack by selecting compounds or libraries that are either lead-like or drug-like properties with the potential of oral bioavailability. Virtual screening is divided into two types as (a) ligand-based virtual screening (LBVS) and (b) structure-based virtual screening (SBVS) (refer to Appendix II).

Lipinski Rule

The selection criteria of lead compounds using the rule are referred to as Lipinski analysis (Lipinski et al. 1997).The use of upper and/or lower bounds on quantities such as molecular weight (MW) or logP helps to vary the in vivo properties of drugs. The rule of *5* developed by Lipinski predicts that good cell permeation or intestinal absorption is more probable when there are less than *5* H-bond donors, 10 H-bond acceptors, MW is less than 500 and the calculated logP is lower than *5*. Property ranges for lead-like compounds can be defined: 1–5 rings, 2–15 rotatable bonds, MW less than 400, up to 8 acceptors, up to 2 donors and a logP range of 0.0 to 3.0. The average differences in comparisons between drugs and leads include 2 less rotatable bonds, MW 100 lower and a reduction in logP of *0.5* to 1.0 log units. Thus, one of the key objectives in the identification of lead-like compounds for screening, either by deriving subsets of corporate, or commercial, compound banks or through the design of libraries, is the need for smaller, less lipophilic compounds that, upon optimization, will yield compounds that still have drug-like properties. Figure 1.4 gives the different approaches used in virtual screening process. Further using Lipinski bioavailability rules, neural nets (e.g. drug-like character), pharmacophore analyses, similarity analyses, scaffold hopping and docking and scoring functions, lead compounds can be selected. The example given for selecting the compounds based on the virtual screening method of data bases is illustrated in Sect. 1.3.3.

1.3.2.3 Quantitative Structure-Activity Relationship (QSAR)

In ligand-based drug design, a computational model is needed for further identification of promising molecule as a lead molecule for further experimental investigation. QSAR modelling techniques are used for further lead optimization. It is a mathematical relationship between a biological activity of a molecular system and its geometric and chemical characteristics. QSAR attempts to find consistent relationship between biological activity and molecular properties, so that these "rules" can be used to evaluate the activity of new compounds.

The concept of QSAR was first introduced in 1968 (Selassie et al. 2003), and the model of QSAR is related by the following equation (Crum-Brown and Fraser 1968):

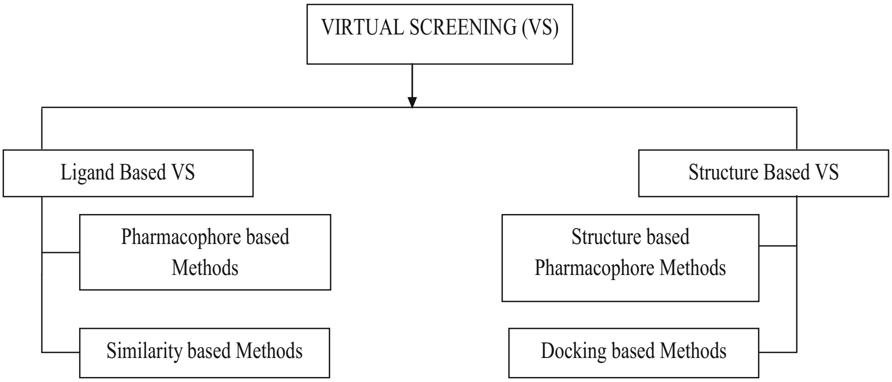


Fig. 1.4 Different approaches to virtual screening process

$$\delta = f(C) \tag{1.2}$$

where the physiological activity δ was expressed as a function of the chemical structure.

Later quantitative approaches combine different physicochemical parameters in a linear additive manner. Free and Wilson proposed structure-activity dependencies by equation

$$AB = u + \Sigma^{i} a_i x_i \tag{1.3}$$

where *AB* is the biological activity, *u* is the average contribution of the unsaturated parent molecule of a particular series (training set compounds), the a_i values are contributions of various structural features and the x_i values denote the presence or absence of particular fragments (Free and Wilson 1964). Since then QSAR has remained a thriving research area in drug design.

More recently developed QSAR modelling approaches include HQSAR (Lowis 1997), inverse QSAR (Cho et al. 1998) and binary QSAR (Gao et al. 1999). The accuracy of QSAR modelling is greatly improved by using sophisticated statistical and machine learning methods, for example, partial least square (PLS) (Dunn and Rogers 1996) and support vector machines (SVM).

QSAR models are regression models used in the chemical and biological sciences; QSAR regression relates a set of physicochemical properties or theoretical molecular descriptors of chemicals to the potency of the biological activity (most often expressed by logarithms of equipotent molar activities) of chemicals. It is a technique that quantifies the relationship between structure and biological data and is useful for optimizing the groups that modulate the potency of the molecule and also predict the activity of newly designed molecules (Hansch 1990).

There are different types of computational methods in QSAR depending upon the data complexity. They are two-dimensional (2D), three-dimensional (3D) and higher methods (Livingstone 2004). 2D QSAR is insensitive to the conformational

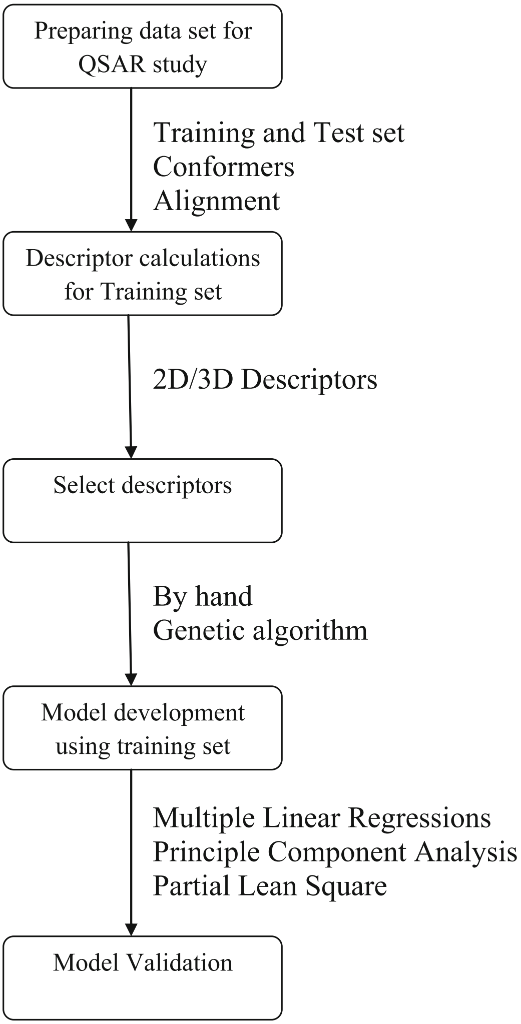

Preparing data set for QSAR study
Training and Test set
Conformers
Alignment
Descriptor calculations for Training set
2D/3D Descriptors
Select descriptors
By hand
Genetic algorithm
Model development using training set
Multiple Linear Regressions
Principle Component Analysis
Partial Lean Square
Model Validation

Fig. 1.5 Various stages of QSAR model development

arrangement of atoms in space, while in 3D QSAR needs information on the position of the atoms in three spatial dimensions. In 4D QSAR for each molecule, a set of automatically docked orientations and conformations are developed by genetic algorithms. Induced-fit scenarios of ligands upon binding to the active site and solvation models can be thought of as the fifth (protein flexibility) and sixth (entropy) dimensions in 5D and 6D QSAR, respectively.

The QSAR model development generally is divided into three stages: data preparation, data analysis and model validation. The development of good quality QSAR model depends on many factors like data set and their biological data, selection of descriptors, statistical methods and model validation. The process of QSAR development was given in the flow chart (Fig. 1.5).

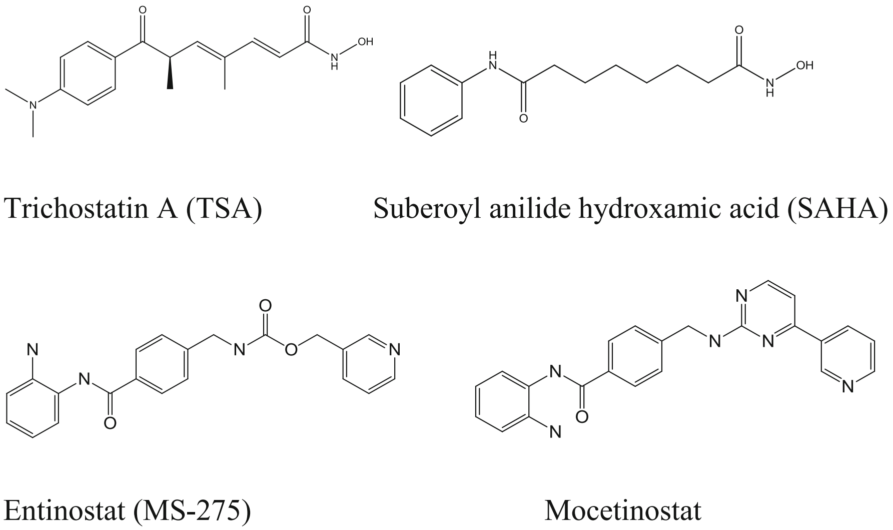
The developed models were useful in prediction of untested compounds. In QSAR model development, the main challenge is the selection of data set and group of descriptors, which describes structural physicochemical features associated with the biological activity. The developed QSAR models were validated by (i) cross-validation, (ii) randomization, (iii) bootstrapping and (iv) external validation. The validation methods are needed to establish the predictiveness of a model on unseen data and to help determine the complexity of an equation that the amount of data justifies. The internal validation uses data set that creates model and a separate data set for external validation. Internal methods for validation of models are least square fit (R^2), cross-validation (Q^2), adjusted R^2 (R^2adj), root mean-squared error (RMSE), bootstrapping and scrambling (Y-randomization). The external validation is a best method to validate the model, such as evaluating QSAR model on a test set of compounds. These are statistical methods used to select the best QSAR model.

1.3.3 Illustrated Examples Using CADD

HDAC proteins have been associated with basic cellular events and disease states, including cell growth, differentiation and cancer formation because of their role in gene expression. Several HDAC inhibitors (HDACi) are in clinical trials, namely, benzamide derivatives (Fig. 1.6), hydroxamic acids, cyclic peptides and short-chain fatty acids (Wagner et al. 2010). SAHA (suberoylanilide hydroxamic acid or vorinostat (Zolinza®)) which is structurally similar to trichostatin A (TSA) was the first HDACi approved for the treatment of refractory cutaneous T-cell lymphoma by the Food and Drug Administration (FDA) in October 2006 (Walkinshaw and Yang 2008). SAHA compound inhibits all zinc-dependent HDACs in the low nanomolar range, and recent studies suggested that it has weak inhibitory effect on the class IIa HDACs (Bradley et al. 2009).

Entinostat (SNDX-275, MS-275) belongs to benzamide class HDACi and inhibits HDAC1 and 2, 3 and 9 and has low effect against HDAC4, 6, 7 and 8 (Khan et al. 2007). Entinostat is in phase II clinical trial for treatment of Hodgkin’s lymphoma and advanced breast cancer (in combination with aromatase inhibitors) and metastatic lung cancer (in combination with erlotinib). Mocetinostat (MGCD0103) is class I selective HDAC inhibitor and is undergoing phase I and II clinical trials for hematologic malignancies and solid tumours (Blum et al. 2009).

The crystal structure of the HDAC2 protein (PDB ID: 3 MAX) was downloaded from the protein data bank (http://www.rcsb.org/pdb). The crystal structure of histone deacetylase 2 (HDAC2) protein has three chains, which are A, B and C. The reference compounds SAHA and MS-275 (Entinostat) were docked into active sites of all three chains using LigandFit programme in Discovery Studio; out of three chains, chain A has given the best docking score and higher H-bond interactions than chains B and C. The docking score of all three chains with SAHA and Entinostat was shown in Table 1.4. Chain A was selected as active



Trichostatin A (TSA)

Suberoyl anilide hydroxamic acid (SAHA)

Entinostat (MS-275)

Mocetinostat

Fig. 1.6 Chemical structures of benzamide HDACi

chain, and the optimized benzamide compounds were docked into active site of 3MAX-A. The docking score along with binding orientations and hydrogen bonds were considered for choosing the best pose of the docked compounds. The docking score of the SAHA compound was 40.8 with three hydrogen-bonding interactions with Arg39(2), Gly305 and Gly142(2), and for Entinostat the docking score was 42.6, with four hydrogen-bonding interaction with Arg39, Cys156, Gly305 and His183 and the configurations are given in Fig. 1.7. The designed compounds that scored docking score above than reference compounds with greater interaction with the crucial amino acids were considered as effective HDAC2 inhibitors.

Virtual screening studies were used to find potential lead molecules with increased inhibitory activity against HDAC2 inhibitors. The pharmacophore model Hypo1 (Fig. 1.8) from benzamide compounds was used as 3D query in database screening of the National Cancer Institute (NCI) database containing 265,242 molecules and Maybridge database containing 58,723 molecules. Ligand pharmacophore mapping protocol was used with flexible search option to screen the database. Hit compounds from the database with estimated activity less than 0.1 μM were selected, and further screening of compounds using Lipinski rule of five compounds has (i) molecular weight less than 500, (ii) hydrogen donors less than 5, (iii) hydrogen acceptors less than 10 and (iv) an octanol/water partition coefficient (Log P) value less than 5.

The pharmacophore model development was performed with Discovery Studio (DS) and Schrodinger softwares. Benzamide pharmacophore model was developed by HypoGen algorithm in DS. Hypo1 of HBD, HBA, RA and HY pharmacophore features were selected based on cost difference and correlation coefficient

Table 1.4 The docking score of SAHA and MS-275 with HDAC2 protein

HDAC2 (3MAX)	Chain A		Chain B		Chain C	
	Docking score	H-bond interaction	Docking score	H-bond interaction	Docking score	H-bond interaction
SAHA	40.8	ARG39(2), GLY305, GLY142(2)	22.66	Tyr308, His146, Gly142, Ala141	39.96	Arg39, Gly142
MS-275 (Entinostat)	42.65	Arg39, Cys156, Gly305, His183	39.07	Tyr308, tyr29	36.9	Tyr308, tyr29

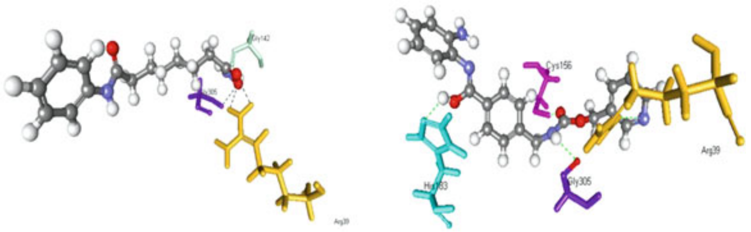


Fig. 1.7 Binding mode of reference compounds SAHA and MS-275

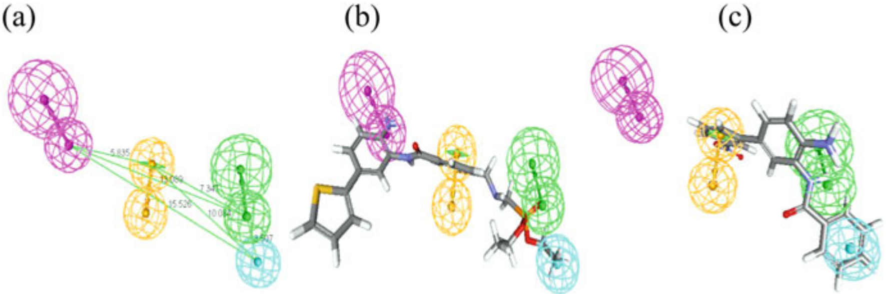


Fig. 1.8 The best pharmacophore model (Hypo1) of HDAC2 inhibitors generated by the HypoGen module: (**a**) the best pharmacophore model Hypo1 represented with distance constraints (Å), (**b**) Hypo1 mapping with one of the active compounds, and (**c**) Hypo1 mapping with one of the least active compound. Pharmacophoric features are coloured as follows: hydrogen-bond acceptor (*green*), hydrogen-bond donor (*magenta*), hydrophobic (*cyan*) and ring aromatic (*orange*) (Naresh Kandakatla and Geetha Ramakrishnan 2014a, b)

(Fig. 1.8). The pharmacophore model can be validated by three methods, such as cost analysis, test set prediction and Fisher's randomization test.

A total of 6130 compounds from NCI and 1379 from Maybridge were mapped using the features of Hypo1. The biological activity IC_{50} (inhibitory concentration

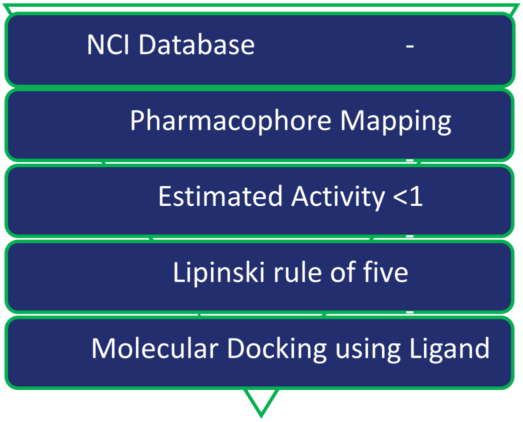


Fig. 1.9 Schematic representation of virtual screening process implemented in the identification of HDAC2 inhibitors

for 50% in µM) was converted to negative logarithmic dose in moles (pIC_{50}) for analysis. The pIC_{50} values of the molecules spanned a wide range from 5 to 8. A total of 1198 and 440 compounds from NCI and Maybridge showed HypoGen estimated value of less than 1 µM for their biological activity and were considered for further studies, and these compounds were screened for Lipinski rule of 5. A total of 625 (382 NCI, 243 Maybridge) compounds obeyed the rule and were subjected to molecular docking studies. The flow chart in Fig. 1.9 was a schematic representation of virtual screening process.

A total of 625 compounds with estimated activity less than 1 µM and favourable Lipinski rule were chosen from NCI and Maybridge databases, and 571 compounds from natural database were subjected to molecular docking studies using LigandFit and LibDock docking programmes. Based on docking score and H-bond interactions, 30 hits were selected from three databases (Naresh Kandakatla and Geetha Ramakrishnan 2014b), and the structure of few of the lead compounds with the respective codes (NSC108392, NSC127064, MFCD01935795, MFCD00830779, ZINC4089202, ZINC4000330) was selected based on structural diversity and stability. These novel compounds can be used for experimental studies for the inhibition of HDAC2 with suitable pharmaceutical formulation.

1.4 Clinical Trials

For a bioactive compound to succeed as a drug, it should pass many selective filters during development like toxicity and in the body including metabolism, uptake, excretion and distribution.

1.4.1 Preclinical Trials

After a lead compound is identified, the medicinal chemist/organic chemist has due interest to prepare them and put into clinical trials. The ability to predict absorption, distribution, metabolism, excretion and toxicology (ADMET) properties from molecular structure has a tremendous impact on the drug discovery process both in terms of cost and the amount of time required to bring a new compound to market. For example, different stereoisomers will exhibit differences in physiochemical properties, such as absorption, metabolism and elimination.

Toxicologists use experimental animals to identify hazardous substances for humans. The main disadvantage is the need for large amounts of substance, several years for the animal studies and relatively expensive. This type of study is of limited value in mechanistic understanding of toxicity. This type of research accounts for 60–65% of the total cost of introduction of a drug into the market. In a nut shell the preclinical activities in the order follows six different sequences as listed below.

Synthesis and purification of the new drug
Pharmacology of the new drug
Pharmacokinetics: absorption, distribution, metabolism, excretion and half-life
Pharmacodynamics: mechanism of action and estimates of therapeutic effects
Toxicology including carcinogenicity, mutagenicity and teratogenicity
Efficacy studies on animals

1.4.2 Human Clinical Trials

To be able to estimate the hazardous risk of humans, additional studies on the mechanism of action, species extrapolation and effects in the low and human-relevant dose range need to be followed. Generally, dose-dependent studies are done for production volume greater than 1000 tons per year in the chemical industry. But drug safety evaluation of pharmaceutical agents is complex as drug exposure to humans is intentional and mechanism of toxicity should be pursued.

An assessment of toxicity requires a broad and interdisciplinary research and development strategy, which includes system biology and case studies on the liver, kidney, cardiovascular, endocrine and in vitro teratogenicity. Further haemotoxicity and peripheral blood cell studies and investigations are done to find their consequences in the drug-induced toxicity (Jurger Borlak 2005).

1.4.3 Types of Clinical Trials

Phase I Trial

In this procedure, how well a drug or procedure can be tolerated in humans acting as healthy volunteers, aged between 18 and 55 years, males and females (however, no females who could be or could become pregnant) of normal weight, no smokers and no alcohol (ab)use will be assessed. The volunteers are given the drug taken with 150 ml water accompanied by standard food, no other therapy and no intake of fruit juices or illegal drugs. The outcome will be to determine a reasonable dose or technique.

Phase II Trial

The phase II trial includes estimation of biological activity or effect (efficacy) and to assess rate of adverse events (toxicity).

Phase III Trial

The phase III trial finds out the effectiveness in comparison to standard treatment or placebo.

Phase IV Trial

Phase IV trial includes long-term surveillance (monitoring) and assesses long-term morbidity and mortality.

Clinical trials provide a systematic framework within which scientific research in human subjects can be carried out efficiently and ethically.

Experimental conclusions are reached in a manner that is statistically defensible.

1.5 Conclusions

Drug discovery process involves target identification, lead compound design and clinical trials. Target identification involves identification of the root cause of the disease. In the case of lead compound selection, virtual screening is a powerful tool to enrich libraries and compound collections. A proper preprocessing of the compound database is of utmost importance in drug design. Further experimental data and theoretical investigations are needed for better p*K*a estimations and better scoring functions. Stepwise procedures (filters, pharmacophore searches, docking and scoring, visual inspection) are most efficient in drug designing. Fragment-based approaches are a promising new strategy in lead structure search and optimization.

The new opportunities in medicinal formulations include genotyping of drug targets and metabolic enzymes which enables cost savings in drug development through better design of clinical trials. The selection of the best drug for a certain patient with individual dose ranges (variance in target sensitivity reduced or increased metabolism) and fewer toxic side effects and drug-drug interactions.