1. INTRODUCTION

1. Introduction

For many years our society is faced with dreadful diseases like AIDS, cancer, drug resistance, tuberculosis and malaria. Thus, there is an urgent need to design and develop new drugs with improved therapeutic efficacy and diminished side effects. Faster, efficient and more economic techniques are required to design and develop newer drug molecules. Furthermore, there is also a demand on scientific methods that replace or at least refine and reduce the use of laboratory animals[1,2].

Although, the Chemical Abstracts Service registers an always increasing number of pharmaceutical lead compounds every year, still there is a tremendous need to develop new drugs for curing human diseases quickly. For a long time, medicinal chemists have systematically modified lead compounds with the driving force of synthetic feasibility, experience and intuition. It takes a long time to synthesize a new compound, biological evaluation of the new molecule and further identifying the lead. However, over the last decade, important contributions to the design of biologically active lead compounds reducing experimental research costs, came from rational molecular designing strategies[3-5], such as bio-structural research, computer-assisted data handling, data storage, retrieval and processing from chemical databases[6-8] employing Computer Aided Drug Design (CADD) methodologies etc.

1.1 Drug discovery process

Drug discovery is an expensive and time consuming process. Retrospective analyses of the pharmaceutical industry in past decades estimate that each new drug in the market takes an average of 14 years to develop, costing approximately $800 million. In addition, only one in ten compounds that enter clinical trials, make it to the market[9,10]. The attrition rate in discovery research is similarly high. Modern target-oriented drug discovery is usually organized into a series of stages as shown in **Figure 1**.

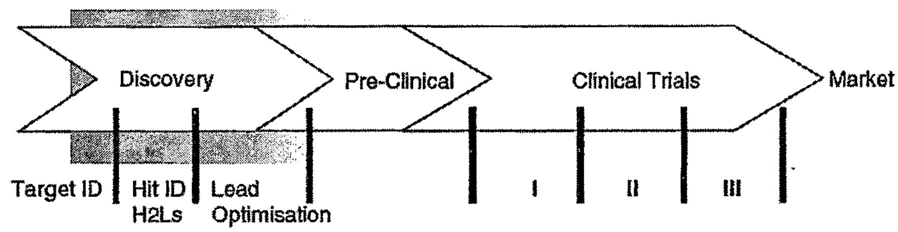

Figure 1 The drug discovery process. The light shaded box emphasizes where CADD techniques can play a significant role.

1.1.1 Target identification

Clearly, the starting point for a target-oriented drug-discovery project is to identify a relevant target. In the pre-genomic era, targets were discovered through cellular and protein biochemistry methods, where a detailed understanding of the origin of a disease led to isolation and characterization of key protein molecules. The nature and significance of these targets were established before much of the modern science of molecular biology and genomics methods were available. The approach to biological research has undergone dramatic changes in the past decade, with successions of 'omics' technologies becoming available. Genomics has recorded the sequence of nucleic acid bases in many genomes, and bioinformatics analyses are identifying the coding regions. Comparing the genomes of both, pathogen and host organism, can help to identify potential target genes. Transcriptomic methods monitor the identity and levels of RNA transcribed for each gene, and there have been high hopes that comparison of "normal" and diseased cells will identify targets. There is a vast literature in these areas[11] which provide an insight into these methods. There has been considerable interest (and investment) in applying these methods to find new targets for different disease conditions. Whatever the mechanism of identifying a target, there needs to be some level of validation before nominating it for a drug-discovery project. The phrase "target validation" is much misused - a target cannot be said to be truly validated until a drug that uniquely affects that target is in the market. In general, the requirements for a target are to establish a biological rationale, for why after binding the target a compound will have the desired therapeutic benefit. This can include assessing the viability of the organisms produced with a particular gene removed, either through knock-out technology or through RNA interference techniques. These are not ideal methods for emulating the actual effect of a drug with gene knock-outs. There is much redundancy and subtlety in biological pathways and the removal of a gene can often be compensated in other ways as the organism differentiates and grows. An example here is the attempts to discover a function for the beta form of the estrogen receptor. Once the gene had been identified, there were intense efforts to ascribe a function to the gene, with considerable investment in producing and characterizing knock-out animals[12]. The best case for a target is to have a compound available that can provide the biological proof of concept. This is a compound that is sufficiently specific for the target of interest and can be studied either in cellular assays or in animal models of disease, to demonstrate that modulating a particular target will have the desired therapeutic benefit, *in vivo*. In addition to biological validation, targets also need to be considered for what is termed, druggability. That is, does the target have a binding site which can accommodate a drug-like compound with sufficient affinity and specificity?

In summary, enzyme active sites tend to be highly druggable consisting of a distinct cleft designed to bind small substrates and with defined shape and directional chemistry. In contrast, most protein-protein interactions are less druggable as they cover quite large areas of protein surface with only a few shape or chemical features that a small molecule could bind to, very selectively. Unless particular "hot-spots" of activity can be identified, they are generally regarded as unsuitable drug targets.

1.1.2 Hit Identification

A hit is a compound that binds to the target and has the desired effect. The conventional method for identifying hits is by screening collection of compounds which could consist of natural products or substrate mimetics, legacy compounds in a company's collection, compounds synthesized as potential hits against a particular class of target (focused library) or commercially available compounds. A majority of large pharmaceutical companies have invested considerably in automating this initial phase of hit identification, both in the generation of suitable target libraries and in the initial/preliminary assays. The High Throughput Screening (HTS) approach places considerable constraints on the robustness of the assay and the availability and properties of the available compound collection[13]. HTS is also very expensive, consuming large quantities of targets and compounds and requiring significant investment in robotic screening devices. Smaller companies or research groups that rely on screening, usually work with smaller libraries of compounds, and depend on a particular "edge" over the larger companies. That distinctiveness could be either in some detailed knowledge or expertise with the biology of the target class and thus configuring the assay more appropriately, or through a small library of compounds for that particular class of targets.

In most cases, the hit-identification phase relies on configuring a particular assay to monitor binding or inhibition. Usually, a large number of compounds are being screened, so the first experiment is to measure compounds that exhibit activity (above a certain percentage of inhibition) at a set concentration. This is usually followed by confirming the hits, that is where an *in vitro* assay is run at varying concentrations to determine the IC_{50}, the K_i or K_d for the compound and the quality of the compound sample checked. An HTS campaign can require significant resources (compound, target, manpower) and lasts 6–12 months, depending on how long it takes to configure a robust assay. Where smaller collections of compounds are being used, or structure-based methods applied, the hit-identification phase usually lasts around 6 months. The output from a hit-identification campaign is a set of compounds whose chemical structures have been checked and whose activity has been shown to be reproducible. This phase of hit identification is also one of the crucial phases in drug discovery process. These hits need to be converted to lead compound where the molecule possesses potent biological activity as per the target disease.

1.1.3 Hits to Leads

The hit to leads (H2L) phase is where some of the crucial decisions are made in a research project, establishing which chemical series has the potential to be optimized into a drug candidate. This is an important decision as H2L is followed by lead optimization (the next phase) is where significant resources and efforts are put in, optimizing the properties of the compounds. For these reasons, there is a need to set quite stringent criteria for entering lead optimization for each target and reflecting the projected property requirements of the final drug candidate.

The detailed work during the H2L phase varies with the origin of the hit compounds. Wherever the compounds come from, it is usual to re-synthesize the compounds for complete validation of the hit and to either purchase or synthesize close analogs of the compounds. In general, it is during the H2L phase that dramatic changes in chemical template are made and the essential core of the lead series established. The usual aims are to establish preliminary structure-activity relationships (SAR) within one or more series, to explore the indicative physicochemical and ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicological) properties of the compounds and to consider the chemical tractability or synthetic accessibility of the compounds. Entry into lead optimization can be gated by demonstrating some *in vivo* activity in the series. Setting the right barriers for entry into lead optimization is one of the most challenging aspects of medicinal chemistry. This phase usually takes around 6 months, depending on the requirements for biological testing and the degree of syntheses required to establish a lead series with appropriate properties.

There are many ways in which a lead compound might be discovered, few of which are given below[14]:

- Screening of natural materials
- Screening of synthetic 'banks'
- Folklore medicine

- Existing drugs
- Starting from natural ligand or modulator
- Combinatorial synthesis
- Serendipity and the prepared mind
- Computerized searching of structural databanks
- Designing lead compounds by NMR
- Computer-aided design

Once a lead compound is identified, the next most important step in drug discovery is lead optimization which involves moulding the lead candidate to a drug like molecule.

1.1.4 Lead Optimization

This is the most resource-intensive component in drug discovery, requiring considerable inputs from synthetic chemistry, molecular modeling, biology of the disease and assay design. The main challenge is to develop one or more compounds with the desired drug-like properties, as well as having sufficient affinity for the target. Also, the compound needs to have an appropriate selectivity profile and be able to get to the site of action in sufficient concentration.

The early stages of the lead optimization process are usually focused on achieving the desired affinity and selectivity. Selectivity requirements vary from target to target and, in particular, between different therapeutic areas. In these early stages, there can still be some modest changes in the central core of the compound. However, as lead optimization progresses, the main changes are on the periphery of the molecule. The main driver is the biology - it is remarkable to know how small changes in the chemistry can have a large effect on the biological activity, particularly *in vivo*. Lead optimization typically takes 18–30 months, depending on the complexity of the target biology, the resources deployed and the chemistry of the lead series. The real challenge in lead optimization is balancing when certain properties need to be introduced and deciding when to abandon a particular lead series. The output from the lead optimization is a compound (or a set of compounds) that meets the required criteria of *in vivo* efficacy in animal models.

1.1.4.1 Lead optimization techniques

Various computer-aided drug design strategies can be applied for lead optimization. The most commonly employed technique is establishing a Quantitative Structure Activity Relationship (QSAR) amongst the candidates under study. Two dimensional (2D) and Three Dimensional (3D) QSAR techniques are the most widely used techniques for this purpose. These analog-based drug design techniques are receptor independent and have proven to be useful for lead optimization.

1.1.4.1.1 Two-dimensional Quantitative Structure Activity Relationship (2D QSAR)

Little additional development in QSAR occurred until the work of Louis P. Hammett within the physical organic chemistry field. He is considered to be the father of Linear Free Energy Relationships (LFER). In fact, QSAR methodology as applied nowadays is attributed to the independent contemporary publications on the Free-Wilson model[15] and the Hansch model[16].

Linear Free Energy Relationship

In the mid-1930's, Hammett observed that the addition of substituents to the aromatic ring of benzoic acid had an orderly and quantitative effect on the dissociation constant. He also observed similar effect on the dissociation of other organic acids and bases[17,18]. From empirical observation, he consequently derived the following linear relationship, the so called Hammett equation (1953):

$$\log K/K_o = \rho\sigma$$

where, the slope (ρ) is a proportionality reaction constant pertaining to a given equilibrium that relates the effect of substituents on that equilibrium to the effect on the benzoic acid equilibrium. σ is a parameter that describes the electronic properties of aromatic substituents , i.e. electron-withdrawing or donating power. K and K_o are the dissociation constants for the substituted and unsubstituted benzoic acids, respectively.

These relationships are termed as linear free energy relationships because they recall the equation relating the free energy, ΔG, to an equilibrium constant, K or rate constant, k. In other words, the energy of the reaction is related to concentration measurements by logarithmic relationships.

$$\Delta G = - RT \ln K$$

The Hammett's correlation to describe the reactivity of aromatic systems was instrumental to the ultimate derivation of QSAR as a discipline by Hansch and his colleagues in the late 1960s. Besides, it was the first study that partitioned the molecule and explained its activity from its fragments, instead referring to its totality.

Closely related to Hammett's equation, Taft worked on the steric effects (Es) and derived the so called Taft equation[19,20]:

$$\log k / k_o = \rho^* \sigma^* + \delta E_s$$

Working in the same direction Swain studied the effects of field and resonance in the electronic parameters. He investigated the variation of reactivity of a given electrophilic substrate towards a series of nucleophilic reagents, deriving the linear free-energy relation, called Swain-Scott equation[21]:

$$\log k / k_o = sn$$

where, n is the measure of the nucleophilicity characteristic of the reagent and s is a measure of the sensitivity to the nucleophilicity of the reagent characteristic of the substrate. He also derived the Swain-Lupton equation, a dual parameter approach to the correlation analysis of substituent effects, which involves a field constant (F) and a resonance constant (R)[22].

Free and Wilson visualized the molecule in a different manner. They postulated that the biological activity of a molecule is due to the positive or negative contributions of individual groups present in different positions of that molecule. The basic unsubstituted skeleton is assumed to possess a discrete value of biological activity. Thus, they formulated an additive model, where the activity is discretized as a simple sum of contributions[15]:

$$B.\ A = \Sigma\, a_i + \mu$$

where, a_i the group contribution of the substituent X_i and μ the biological activity of the parent structure. In this additive model, the descriptors, also called indicator variables, codify the presence or absence of particular structural characteristics. They are assigned the binary values of 1 and 0, accordingly.

Regression analysis by modification of Free-Wilson technique was applied by Fujita and Ban[23] to phenylethylamines as substrates of enzymes associated with biosynthetic pathway of sympathetic neurotransmitters.

Besides, in the middle 50's, centering on the year 1954, the work from several laboratories came together to offer a quantitative explanation and relationship to a biological activity. The work focused on the carcinogenicity of polycyclic aromatic hydrocarbons. Daudels, Pullmans and Coulson studies used valance bond theory and molecular orbital theory to quantify the involvement of certain bonds in a event initiating the onset of a carcinogenic outcome[24,25].

Hansch Analysis

The origin of QSAR as practiced today began with the research of Robert Muir, a botanist at Pomona College, who was studying the biological activity of plant growth regulators. In attempting to correlate the structures of the compounds with their activities, he consulted his colleague, Corwin Hansch[26]. Later on, Hansch and Fujita published an LFER related model, considered to be the formal beginning of QSAR[27,28].

Their fragment and additive group contribution theory added two things to what have been done before - the use of calculated properties to correlate the biological activities and the recognition that the multiple properties may influence the biological activity.

Starting from the hypothesis that substituents on a parent molecule have a quantitative relationship with the biological activity, they used Hammett sigma parameters to account for the electronic effect of substituents but that did not lead to meaningful QSAR. Hansch recognized the importance of lipophilicity, expressed as the octanol-water partition coefficient, on biological activity. This parameter provides a measure of the bioavailability of compounds, which determines, in part, the amount of compound that gets in the target site. This so called Hansch equation[29] was developed to correlate physicochemical properties with biological activities which could be expressed in a general form as:

$$\log 1/C = a(\log P)^2 + b\log P + c\sigma + \mathrm{dEs} \ldots\ldots\ldots + k$$

where, C is the molar concentration that produces a particular quantum of biological effect; *P* is the octanol/water partition coefficient, σ is the electronic Hammett constant and Es is the Taft's steric constant.

The combination of Hansch and Free-Wilson analysis in a mixed approach widens the applicability of both of the QSAR methods. However, there are few problems associated with these 2D QSAR studies, as enumerated below:

- Only molecules with similar structure can be studied.
- The validity of numerical descriptors is open to doubt.
- The tabulated descriptors may not include entries for unusual substituents.
- It is necessary to synthesize a range of molecules where substituents are varied to evaluate a particular property. However, synthesizing such a range of compounds is not feasible.
- 2D QSAR equations do not directly suggest new compounds to be synthesized.
- 2D QSAR considers the molecule to be a planar entity and not in its three dimensional form.

These problems are avoided by the 3D QSAR approaches offering the following advantages:

- In 3D QSAR the properties of the test molecules are calculated individually by computer programme. There is no reliance on tabulated or experimental factors. There is no need to confine the study to molecules of similar structure. As long as one is confident that all the compounds in the study share the same pharmacophore and interact in the same way with the target, they can all be analyzed in a 3D QSAR study.
- Favorable and unfavorable interactions are represented graphically by 3D contours around a representative molecule. A graphical picture such as this is easier to visualize than a mathematical equation.
- The graphical representation of beneficial and non-beneficial interactions allows medicinal chemist to design new compounds.

1.1.4.1.2 Three-dimensional Quantitative Structure Activity Relationships (3D-QSAR)

In 3D QSAR the 3D properties of a molecule are considered as a whole rather than considering individual substituents or moieties. The philosophy of 3D QSAR revolves around the assumption that the most important features about a molecule are its overall size and shape (steric fields) and its electronic properties (electrostatic fields). There are several approaches to 3D QSAR, like Comparative Molecular Field Analysis (CoMFA), Comparative Molecular Similarity Indices (CoMSIA), Receptor Surface Analysis (RSA) and Molecular Surface Analysis (MSA).

Comparative Molecular Field Analysis

The rationale behind CoMFA is the observation that non-covalent intermolecular interactions producing biological properties of ligands can be approximately described by steric and electrostatic fields[30,31]. In other words, differences in biological activity of molecules are related to local changes in the molecular field intensities. In the CoMFA procedure, aligned dataset molecules are placed into a cubic lattice and interaction energies between the molecule and a defined probe are calculated for each lattice point (usually grid spacing is 2 A°). The steric van der Waals interactions are derived from the Lennard-Jones potential (Equation 1) and the electrostatic forces, utilizing a simple Coulombic interaction term (Equation 2). These potentials contribute only to the enthalpic term of the free energy of binding, although it may be important to take into account the entropic effects (e.g. hydrophobicity) as well[32,33].

$$E_{vdW} = \Sigma\, (A_{ij} / r_{ij}^{12} - B_{ij} / r_{ij}^{6}) \qquad \text{.......... (1)}$$

Where, A_{ij} = repulsive term coefficient; B_{ij} = attractive term coefficient; r_{ij} = distance between atoms i and j

$$E_{elec} = q_1 q_2 / \varepsilon r \qquad \text{.......... (2)}$$

Where, q_1 and q_2 = atomic charges of interacting atoms; ε = dielectric constant; r = interatomic distance.

Since the number of field energy values greatly exceeds the number of compounds, a multivariate statistical technique called partial least square (PLS)[34,35] has to be utilized to evaluate the relationships between the biological activities (Y variables) and the calculated interaction energies (X variables). In the iterative process of PLS, new components (latent variables) are extracted so that at each time, the degree of commonality between the X and Y variables is maximized[30]. Usually a maximum of five to six components is enough to generate a realistic model[32]. The optimum number of components is traditionally determined by cross-validation, a technique that assesses the ability of QSAR model to predict the biological data. In this technique, one or more compounds are left out from the model and their biological activities are predicted on the basis of the model derived with the remaining compounds. This procedure is repeated until each compound has been predicted once[30]. The resulting statistical values are q^2 (square of the cross-validated correlation coefficient, Equation 3) and S_{PRESS} (standard deviation of the error of predictions).

$$q^2 = 1 - PRESS / \Sigma (Y-Y_{mean})^2 \quad \text{.......... (3)}$$

where, PRESS = the sum of the squared error of predictions; Y = experimental value; Y_{pred} = predicted value; Y_{mean} = mean value of the experimentally observed activities.

Values of q^2 can range from 1.0 (perfect predictions) through zero to even negative (poor models). In general, the risk of chance correlation has been considered negligible with a q^2 greater than 0.3[36,37]. However, robust and predictive models should have q2 of at least 0.4-0.5 according to current experience[32,37]. The number of components used in the final model is recommended to be taken as that in which the S_{PRESS} value is the smallest or it no longer decreases significantly[38]. This helps to avoid adding noise to the model. In fact, it has been shown that an excessive amount of noise (i.e., irrelevant grid points that do not contribute to explaining the biological activity) can be detrimental to the PLS prediction. Therefore, different variable selection techniques have been developed to eliminate irrelevant X-variables. The so-called minimum-sigma filter eliminates those gird points that express lower variance in the field values than the user-defined cut-off value. More advanced variable selection approaches include, for example the GOLPE procedure[39] and the smart region definition method[40].

The conventional r^2 (square of correlation coefficient) of a non-cross-validated final model shows how well this particular model fits the input data. In an analogous way with q^2 and S_{PRESS}, r^2 and the corresponding standard deviation can be calculated by replacing Y_{pred} with Y_{calc} (activity value calculated by model) in Equation 3. The predictive r^2 (r^2_{pred}) is obtained if the model is used to predict the activities of a set of compounds not included in the model[30].

The CoMFA QSAR equation can usefully be represented as a 3D coefficient contour map in which the colored polyhedra denote the 3D locations of the steric and electrostatic fields that significantly contribute to the model[30,31]. Usually two different contour levels are displayed for each CoMFA field, positive and negative. For example, a positive steric contour shows a region that increases the activity of a molecule if occupied, whereas a negative contour reveals a sterically unfavorable region. Even though these contour maps may not exactly mirror the corresponding protein binding sites, a ligand alignment based on a protein-bound conformation can produce CoMFA contours that correspond to some extent to the steric and electrostatic environment of the binding site[32].

Comparative Molecular Similarity Indices

CoMSIA was developed as an alternative approach to perform 3D QSAR[41]. Instead of field descriptors based on Lennard-Jones and Coulomb-type potentials, molecular descriptors are computed based on similarity indices of aligned molecules. The energy potentials used in CoMFA are very steep near the van der Waals surface of molecules and they produce singularities at the atomic centers. In order to avoid too large energy values, arbitrary cut-off values have to be defined and evaluation of the potentials is restricted to regions outside the molecules. To overcome such problems, CoMSIA utilizes a Gaussian-type function for the distance dependence between the probe and the atoms of the data set molecules. Therefore, no cut-off limits are needed and the similarity indices can be calculated at all grid points inside and outside the molecules. This makes CoMSIA relatively insensitive to changes in grid spacing or orientation of the aligned molecules with respect to the lattice[42]. In addition to the steric and electrostatic similarity fields, hydrophobic as well as H-bond donor and acceptor properties are presently included in this approach[33,42]. The statistical evaluation of the similarity indices is done by PLS in the same way as in CoMFA. The resulting CoMSIA contour maps directly show the areas important for activity within the space occupied by the molecules[41].

1.1.4.1.3 Validation of QSAR models

As described above cross-validation has traditionally been used to evaluate the predictive power of CoMFA/CoMSIA PLS models. In the most commonly applied leave-one-out (LOO) procedure, each molecule is separately left out from the model and its biological activity is predicted on the basis of the data from the molecules remaining in the model. Unfortunately, a high degree of redundancy in the chemical structures results in too optimistic q^2 values[43]. Additionally, it has been shown that there is no real relationship between q^2 and the external predictivity of the model[44-46]. This phenomenon in which internally predictive models result in low external predictivity and vice versa has been called 'Kubinyi paradox'[47-49]. High q^2 values, although necessary, but are not sufficient for endowing high predictive power and therefore, one should not blindly rely on (LOO) q^2 [46]. Cross-validation in random groups (leave-some-out, LSO) may sometimes produce reliable q^2 values. In particular, a more stringent leave-half-out (LHO) cross-validation has been reported to give a good estimate of model predictivity[50,51]. A random group cross-validation procedure should be repeated several times as different distribution of the compounds to the groups during each PLS analysis affects the q^2 value.

Golbraikh and Tropsha (2002) recommended the use of an external test set as the only way to establish a reliable QSAR model. However, Hawkins et al[52] have argued that when validating a typical QSAR model with a small or modest sample size, it is better to use properly conduced cross-validation than to waste valuable information by holding back compounds for a test set.

Response (Y) randomization has also been used to assess the robustness of LSO models. In this approach, the biological activities are scrambled among the training set molecules and novel PLS models are generated with the scrambled data. If the new model gives q^2 values comparable with the original model, it is likely that the original model is fitting only noise. Therefore, a loss of statistical significance in cross-validation after response scrambling is considered to be a necessary condition for the robustness of the model. However, it is not a sufficient condition particularly in the case of redundant datasets[37].

Scoring functions estimate the free energy of binding of a ligand in a protein-ligand complex. They are used to optimize the placement of the ligands during the docking process and after docking they are applied to rank the resulting ligand poses with respect to the other poses of the ligand[64]. Scoring functions are grouped into force-field based, empirical or knowledge-based functions[65]. It is a well known fact that these fast scoring methods do not perform so accurately as the time-consuming free energy perturbation technique[66]. Wang et al[67] explored the performance of eleven popular scoring functions and concluded that if the conformational sampling of ligands works, scoring also performs better. However, no correlation with the binding affinity is possible at present. In general, using a combination of one or more scoring functions (i.e. consensus scoring) has been reported to improve results[67-69]. Docking and scoring is an integral part of structure-based virtual screening. Additionally, molecular docking to target protein structures has successfully been used as an automated alignment method in creating 3D QSAR models. Docking at both protein crystal structures[70-73] and comparative models[74-76] has resulted in statistically significant and predictive QSAR models. In the absence of a precise pharmacophore, automated docking can be a useful tool for generating alignment that is not biased on subjectivity. Also, creating docking alignments is often more straightforward than aligning ligands manually, especially for flexible molecules and when no common pharmacophore is available.

Virtual screening (VS) uses computational methods to search for novel ligands or lead structures from large chemical libraries[77]. It is a complementary alternative to the experimental high-throughput screening (HTS) approach[78]. The ligand-based VS methodologies utilize, for example, exact match or substructure search and similarity or pharmacophoric searching. Structure-based VS employs molecular docking and scoring. Also pharmacophoric queries can be constructed on the basis of the target binding site or features of known ligands bound/docked at this site. In addition to the accuracy of scoring functions in docking, the quality and conformation of the target protein affects the results of the screening. It has been shown that ligand-bound (holo) crystal structures yield meaningful enrichment of known ligands from a database containing mostly decoy molecules[79]. Also apo (without ligand) and modeled structures may in some case perform better than the holo structures. Successful results from many groups suggest that in VS of GPCR ligands, it is beneficial to utilize information derived from known ligands or mutational studies when constructing the 3D models[80-84].

The virtual hits (top rank compounds) found from screened databases are usually submitted for further filtering to reduce the number of compounds to be chosen for biological testing. The Lipinski 'rule of five' has frequently been used as crude criteria for drug-like properties[85,86]. According to this rule, a compound that has molecular weight > 500 Da, calculated log P > 5, the number of H-bond donors > 5 and the number of H-bond acceptors > 10, then it is likely to be poorly absorbed and will be unlikely to permeate through cell membranes. Recently, a concept of lead-likeness has been defined as the properties of lead molecules are usually surprisingly different from the actual drug candidates[87-89]. Other filtering criteria can include requirements for chemical diversity of the protein-ligand complexes, absence of reactive functional groups, and ease of synthesis or patentability[65,90].

1.1.6 References

1. QSAR for prediction of fate and effects of chemicals in the environment. Environmental technologies RTD Programme (DGXII/D-1). Contract EV5V-CT92-0211. Commission of the European Union: Brussels, **1995**.

2. Fate and activity modeling of environmental pollutants using structure-activity relationships (FAME) Contract number ENV4-CT96-0221. Environment and climate programme of the European Union: Brussels, **1999**.

3. Bables, L. M. *Guide to Rational (Computer-Aided) Drug Design*. Research Triangle Institute: Research Triangle Park, **1998**.

4. Van de Waterbeemd, H. *Quant-Struc-Act. Relat.* 11, **1992**, 200.

5. Van de Waterbeemd, H. *Drug Des. Disc.* 9, **1993**, 277.

6. 3D Virtual Chemistry Library, Imperial College of Science, Technology and Medicine. Available at http://www.ch.ic.ac.uk/vchemlib/mol/mol.html.

7. LIGAND: Ligand chemical database for chemical reactions and chemical compounds. Kyoto University. Available at http://www.genome.ad.jp/htbin/www_bfind?ligand

8. The Cambridge Structural Database System - from crystallographic data to protein-ligand applications. Stephen J. Maginn, available in via CrystalWeb, ICSD-WWW, ConQuest, Reaction Web or ISIS. Available at http://www.ccdc.cam.ac.uk/products/csd/

9. Kola, I. and Landis, J. *Nat. Rev. Drug Discov.*, 3, **2004**, 711.

10. Dickson, M and Gagnon, J. P. *Nat. Rev. Drug Discov.*, 3, **2004**, 417.

11. Egner, U.; Kratzschmar, J.; Kreft, B.; Poblenz, H. D and M. Chneider *Chembiochem.* 6, **2005**, 468.

12. Koehler, K. F.; Helgnero, L. A.; Haldosen, L. A.; Warner, M and J.A. Gustafsson, *Endocr. Rev.* 26, **2005**, 465.

13. Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P and Tinker, A. C. *Curr. Top. Med. Chem.* 5, **2005**, 421.

14. An Introduction to Medicinal Chemistry, Graham L Patrick, 2nd Edition, Oxford University Press, **2001**, Chapter - Drug discovery and development pp.153.

15. Free, Jr. S. M.; Wilson, J. W. *J. Med. Chem.* 7, **1964**, 395.

16. Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* 86, **1964**, 1616.

17. Hammett, L. P. *J. Am. Chem. Soc.* 59, **1937**, 96.

18. Hammett, L. P. Physical Organic Chemistry. McGraw-Hill: New York, **1940**.

19. Taft, R. W. *J. Am. Chem. Soc.* 74, **1952**, 3120.

20. Taft, R. W. Steric effects in organic chemistry, Ed. Newman M. S, Wiley: New York, **1956**.

21. Swain, C. G and Scott, C. B. *J. Am. Chem. Soc.* 75, **1953**, 141.

22. Lupton, Jr. E. C. *J. Am. Chem. Soc.* 90, **1968**, 4328.

23. Fujita, T and Ban, T. *J. Med. Chem.* 14, **1971**, 148.

24. Pullman, A and Pullman, B. *Rev. Sci.* 84, **1946**, 145.

25. Pullman, A.; Pullman, B. *Ad. Cancer Res.* 3, **1955**, 117.

26. Hansch, C.; Maloney, P. P.; Fujita, T and Muir, R. M. *Nature*, 194, **1962**, 178.

27. Fujita, T.; Isawa, J. and Hansch, C. *J. Am. Chem. Soc.* 86, **1964**, 5175.

28. Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* 86, **1964**, 1616.

29. Hansch, C. *Acct. Chem. Res.* 2, **1969**, 232.

30. Cramer, III R. D.; Depriest, S. A.; Patterson, D. E.; Hecht, P: The developing practice of Comparative Molecular Field Analysis. In: 3D QSAR in Drug Design: Theory, Methods and Applications, PP. 443-485. Ed. Kubinyi H, ESCOM, Leiden, **1993**.

31. Cramer, III R. D.; Patterson, D. E and Bunce, J. D. *J. Am. Chem. Soc.* 110, **1988**, 5959.

32. Holtje, HD, Sippl, W, Rognan, D, Fokers, G: Molecular modeling: basic principles and applications. 2nd ed. Weinheim, Germany: Wiley-VCH Verlag Gmbh & Co. KgaA; **2003**, pp. 228.

33. Klebe, G and Abraham, U. *J. Comput. Aided Mol. Des.* 13, **1999**, 1.

34. Wold S, Johansson E, Cocchi M: PLS- Partial Least-Squares Projections to Latent Structure. IN: 3D QSAR I Drug Design: Theory, Methods and Applications, pp. 523-550. Ed. Kubinyi H, ESCOM, Leiden, **1993**.

35. Wold, S.; Ruhe, A.; Wold, H and Dunn, W, III *J. Sci. Stat. Comput.* 5, **1984**, 83.

36. Clark M, Cramer III RD, *Quant. Struct-Act. Relat* 12, **1993**, 137.

37. Clark RD, Sprous DG, Leonard JM: Validating models based on large data sets. In: Rational Approaches to Drug Design, pp. 475-485.ED. Holtje H-D, Sippl W, Prous Science, Barcelona, **2001**.

38. Kubinyi H, Abraham U: Practical problems in PLS analyse. In: 3D QSAR in Drug Design: Theory, Methods and Applications, pp. 717-728. Ed. Kulbinyi H, ESCOM, Leiden, **1993**.

39. Baroni M, Costantino G, Cruciani G, Riganelli D. Valigi R. Clementi S *Quant. Struct-Act. Relat.* 12, **1993**, 9.

40. Pastor, M.; Cruciani, G and Clementi, S. *J. Med. Chem.* 40, **1997**, 1455.

41. Klebe, G.; Abraham U. and Mietzner T *J. Med. Chem.* 37, **1994**, 4130.

42. Bohm, M.; Sturzebecher, J and Klebe G. *J. Med. Chem.* 17, **1999**, 1226.

43. Thibaut U, Folkers G, Klebe G, Kubinyi H, Merz A, Rognan D: Recommendations for CoMFA studies and 3D QSAR publications. In: 3D QSAR in Drug Design: Theory, Methods and Application, pp.711-716, Ed. Kubinyi H. ESCOM, Leiden, **1993**.

44. Doweyko, A. M. *J. Comput. Aided Mol. Des.* 18, **2004**, 587.

45. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H and Tropsha, A. *J. Comput. Aided Mol. Des.* 17, **2003**, 241.

46. Golbraikh, A and Tropsha, A *J. Mol. Graph. Model* 20, **2002**, 269.

47. Kubinyi H: Validation and productivity of QSAR models 15th Euro-QSAR Symposium: **2004**; http://www.kubinyi.de/istanbul-2004-lecture.pdf;P.

48. Kubinyi, H.; Hamprecht, F. A and Mietzner, T. *J. Med. Chem.* 41, **1998**, 2553.

49. Van Drie, J. H. *Curr. Pharm. Des.* 9, **2003**, 1649.

50. Oprea, T and Garcia, A. E. *J. Comput. Aided Mol. Des.* 10, **1996**, 186.

51. Tervo, A. J.; Nyronen, T. H.; Ronkko, T and Poso, A *J. Comput. Aided Mol. Des.* 17, **2003**, 797.

52. Hawkins, D. M.; Basak, S. C and Mills, D. *J. Chem. Inf. Comput. Sci.* 43, **2003**, 579.

53. Clark, R.D and Fox, P. C. *J. Comput-Aided Mol. Des.* 17, **2004**, 1.

54. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A and Brown, F. K. *Biopolymers* 68, **2003**, 76.

55. Rarey, M.; Kramer, B.; Lenguer, T and Klebe, G. *J. Mol. Biol.* 261, **1996**, 470.

56. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R and Taylor, R. *J. Mol. Biol.* 267, **1997**, 727.

57. Baxter, C. A.; Nurray, C. W.; Clark, D. C.; Westhead, D. R and Eldridge, M. D. *Proteins* 33, **1998**, 367.

58. Hart, T. N.; Read, R. J. *Proteins* 13, **1992**, 206.

59. Davis, A. M.; Teague, S. J. *Angew. Chem.* 38, **1999**, 737.

60. Fritz, T. A.; Tondi, D.; Finer-Moore, J. S.; Costi, M. P and Stround, R. M. *Chem. Biol* 8, **2001**, 981.

61. Clausen, H.; Buning, C.; Rarey, M.; Lengauer, T and Flex, E. *J. Mol. Biol.* 308, **2001**, 377.

62. Anderson, A. C.; O'Neil, R. I.; Sum, T. S and Stroud, R. M. *Chem. Biol.* 8, **2001**, 445.

63. Schnecke, V.; Kuhn, L. A. *Perspect. Drug Discov. Des.* 20, **2000**, 171.

64. Holtje, H. D, Sippl W, Rognan D, Fokers G: Molecular modeling: basic principles and applications. 2nd ed. Weinheim, Germany: Wiley-VCH Verlag Gmbh & Co. KgaA; **2003**, pp. 228.

65. Lyne, P. D. *Drug Discov. Today* 7, **2002**, 1047.

66. Kollman, P. A. *Chem. Rev.* 93, **1993**, 2395.

67. Wang, R.; Lu, Y and Wang, S. *J. Med. Chem.* 46, **2003**, 2287.

68. Bissantz, C.; Folkers, G and Rognan, D.; *J. Med. Chem.* 43, **2000**, 4759.

69. Charifson, P. S.; Corkery, J. J.; Murcko, M. A and Walters, W. P. *J. Med. Chem.* 42, **1999**, 5100.

70. Tervo, A. J.; Nyronen, T. H.; Ronkko, T and Poso, A. *J. Comput. Aided Mol. Des.* 17, **2003**, 797.

71. Sippl, W. *J Comput. Aided Mol. Des.* 14, **2000**, 559.

72. Soderholm, A. A.; Lehtovuori, P. T and Nyronen, T. H. *J. Med. Chem.* 48, **2005**, 917.

73. Tervo, A. J.; Nyronen, T. H.; Ronkko, T and Poso, A. *J. Chem. Info. Comput. Sci.* 44, **2004**, 807.

74. Kunick, C.; Lauenroth, K.; Wieking, K.; Xie, X.; Schultz, C.; Gussio, R.; Zaharevitz, D.; Leost, M.; Meijer, L.; Weber, A.; Jorgensen, F. S and Lemoke, T. *J. Med. Chem.* 47, **2004**, 22.

75. Tikhonova, I. G.; Baskin, I.; Palyulin, V. A and Zefirov, N. S. *J. Med. Chem.* 46, **2003**, 1609.

76. Xu, Y.; Liu, H.; Niu, C.; Luo, C.; Luo, X.; Shen, J.; Chen, K and Jiang, H. *Bioorg. Med. Chem.* 12, **2004**, 6193.

77. Shoichet, B. K. *Nature* 432, **2004**, 862.

78. Davis, A. M.; Teague, S. J and Kleywegt, G. J. *Angew. Chem. Int.* 42, **2003**. 2718.

79. Miller, M. A. *Nat. Rev. Drug Discov.* 1, **2002**, 220.

80. MeGovern, S. L.; Shoichet, B. K. *J. Med. Chem.* 46, **2003**, 2895.

81. Becker, O. M.; Marantz, Y.; Shacham, S.; Inbal, B.; Heifetz, A.; Kalid, O.; Bar-Haitn, S.; Warshaviak, D.; Fichman, T and Noiman, S. *Proc. Natl. Acad. Sci. U S A.* 101, **2004**, 11304.

82. Bissantz, C.; Bernard, P.; Hibert, M and Rognan, D. *Proteins* 50, **2003**, 5.

83. Evers, A.; Klabunde, T. *J Med. Chem.* 48, **2005**, 1088.

84. Evers, A.; Klebe, G. *Angew. Chem. Int. Ed.* 43, **2004**, 248.

85. Shacham, S.; Marantz, Y.; Bar-Haim, S.; Kalid, O.; Warshaviak, D.; Avisar, N.; Iqbal, B.; Heifetz, A.; Fichman, M.; Topf, M.; Naor, Z.; Noiman, S and Becker, O. M. *Proteins* 57, **2004**, 51.

86. Lipinski, C. A.; Lombardo, F.; Dominy, B.W and Feeney, P. J. *Adv. Drug Deliv. Rev.* 23, **1997**, 3.

87. Lipinski, C. A.; Lombardo, F.; Dominy, B. W and Feeney, P. J. *Adv. Drug Deliv. Rev.* 46, **2001**, 3.

88. Hann, M. M.; Oprea, T. I. *Curr. Opin.Chem. Biol.* 8, **2004**, 255.

89. Oprea, T. I.; Davis, A. M.; Teague, S. J and Leeson, P. D. *J. Chem. Inf. Comput. Sci.* 41, **2001**, 1308.

90. Rishton, G. M. *Drug Discov. Today* 8, **2003**, 86.