PHARMACOPHORE CONTINUED

Screening Strategy for Lead Optimization

Introduction

The goal of small molecule drug discovery for pharmaceutical companies is to turn an organic compound into a highly valued drug candidate in a timely fashion. This is an extremely complex endeavor with numerous hurdles that have to be overcome. The late stage attrition of drug candidates in development and beyond is costly [1], therefore such failures must bekept to a minimum by setting in place a rigorous assessment of drug candidates at the earliest stage of drug discovery: Lead optimization. During this stage, the initial hits are optimized into lead series by knowledge-driven decisions. It is important to increase the knowledge applied to the design of compounds through each synthesis cycle by providing early, complete, and parallel structure-activity-relationship (SAR) data.

Lead Optimization Process

During lead optimization, there are many unanticipated scientific, medical, and business challenges to every drug discovery program. Therefore, we must increase our understanding of the properties of each leading compound and set up criteria for selection of viable leading series (see Figure 1). Collecting as much information early on will provide invaluable guidance for lead optimization. In this paper, we will discuss the in vitro screening strategies for lead optimization.

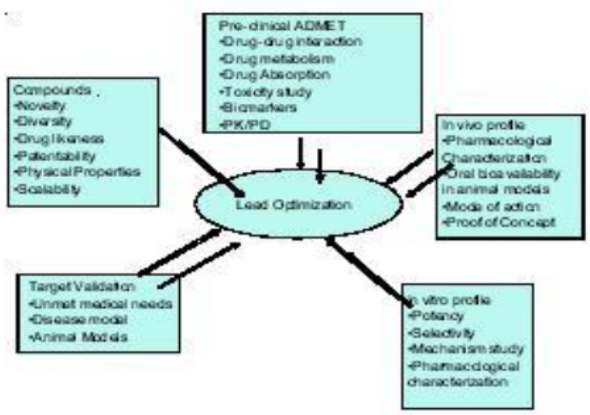


Figure 1. Criteria for selection of lead compounds

An Integrated Screening Approach for Lead Optimization

The screening challenges for lead optimization are significantly different from the hit identification. In hit identification, one bioassay tests many hundreds of thousands of compounds in a short space of time. During lead optimization, smaller numbers of compounds, typically 10 to 100, are put through an array of assays over a longer time-frame. We commonly modify in vitro screening paradigms to respond and solve issues that occur during the lead optimization. A series of advances in bioassay technologies, automation, and miniaturization have enabled the efficiency of hit identification during the past decade [2]. The assays that are used for lead optimization demands connectivity between in vitro and in vivo studies. Therefore, the factors that dictate the selection of the assays for hit identification and lead optimization are different. Results from lead optimization assays need to direct chemists and biologists to make more informed decisions on which compound to synthesize and which compound to move into in vivo efficacy tests. Providing a full array of safety and selectivity profiles on each and every project compound, as quickly as possible after synthesis or discovery, allow scientists to rank and compare compounds against a multiparameter matrix, not just potency. In our experiences, the application of appropriate HTS technologies and selection of suitable biologically relevant in vitro assays for lead optimization will enable project teams to track a full package of information throughout the life of drug discovery programs. Here, we propose an integrated screening approach for lead optimization, which is an efficient process from compound preparation, reagent selection,assay execution, and data reporting for each drug discovery program (Figure 2).

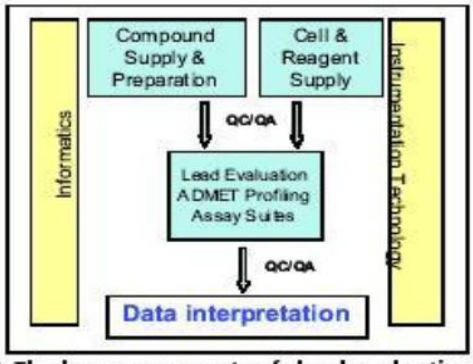


Figure 2. The key components of lead evaluation and lead profiling process: This integrated process allows rapidly delivering high quality, multiparameter data on every compound produced within a discovery program

Lead Evaluation and Lead Profiling

The in vitro bioassays that are used in lead optimization can be divided into lead evaluation and lead profiling assays (see Figure 2). The lead evaluation assays assess potency and selectivity for a given lead compound. The assays used for lead profiling is to provide ADMET liability data that gives the researcher confidence in chemical series as they progress toward drug development. Together, these data direct a project team to identify the most promising preclinical drug candidates.

Lead Evaluation Assays

Lead evaluation assays for lead optimization use both cellfree and cell-based assays. Leading compounds with balanced property profiles will be chosen for further in vivo tests. Here, we refer a group of lead evaluation assays as an assay panel. The assay panel screening approach is determined by 1) target class families, 2) relevant in vitro assays to reflect disease models, and 3) animal models. In this paper, we will exemplify the assay panel screening strategies for target families including G-couple receptors (GPCR), nuclear hormone receptors (NHR), and kinases.

- GPCRs mediate the majority of transmembrane signal transduction in living cells and many marketed drugs target these receptors. About 600 GPCR genes [3] have been identified from genomic sequencing, and 200 GPCRs have known ligands.Three criteria are critical to determine the screening strategy for this target family - Ligand Specificity: when molecules are developed against one GPCR, they often show a cross-reactivity with other members of the same family. For example, there are 17 known chemokine receptors. Some chemokine receptors are closely related, i.e., CR2/CCR5 and CCR1/CCR3. Specificity for a selected chemokine target is a common hurdle that needs to be overcome for the design of highly selective antagonists (unpublished internal observation). Consequently, we assay a number of closely related targets concurrently during lead optimization. There are several benefits of

using this strategy, including identifying selective leading series, assessing a number of potential drug targets simultaneously, encompassing large range of disease-relevant biology, and serendipitous discovery drug-like compounds for another target.
- Target-Specific Liability: the leading compounds for a given GPCR target may also have cross-reactivity with members of different GPCR families. For example, in our experience, GPCR antagonists often show cross-reactivity with the monoamine GPCR receptors includingdopamine, serotonin, and adrenergic receptors. This cross-activity can limit the potential therapeutic use of these antagonists. Therefore, emphasis on the early detection of liability issues related with the monoamine GPCR leads to increased probability of success for drug candidates. - - GPCR signal transduction pathway: GPCRs are modulated by a three-component system including receptors, G proteins, and downstream effectors. Together, these determine which assay technology will be used for drug discovery. The fluorescent-based and radio-labeled, ligandbinding assays have led to the study of GPCR-ligand interactions. These assay techniques have also been enormously useful in defining basic properties of GPCR systems. The cell-based technologies including Ca2+, cAMP, and IP3 quantitation, etc. have been used as functional readouts for GPCRs. These assays require GPCRs to be expressed in multiple cellular backgrounds for pharmacological analysis. During lead optimization, we need both cell-free and cell-based GPCR assays because the binding assays are based on the displacement of labeled ligands and can not differentiate between agonists, antagonists, or inverse agonists. For cell-based GPCR assays, we must keep in mind that the receptor expression levels may have a huge impact on the efficacies of partial agonists and inverse agonists. GPCR signaling in vivo is influenced by tissue location and signaling complexes can be cell type specific. A GPCR can interact with different proteins when over expressed in a cell system compared with its interaction partners in native tissue.

- Nuclear hormone receptors (NRs) [4] constitute a large super family of intracellular ligand-dependent transcription factors. The mode of action of nuclear receptors consists of three steps, including: 1) repression, where NRs recruit a co-repressor complex; 2)de-repression,

where the NR- corepressor complex binds ligand, resulting in chromatin decondensation, which is believed to be necessary but not sufficient for activation of the target gene; and 3) transcription, where a second coactivator complex is assembled that is able to establish contact with the basal transcription machinery, resulting in transcription activation of the target gene. This mechanism is not ubiquitous, since some NRs may act as activators without a ligand, whereas others are unable to interact with target gene promoter in the absence of ligands. The panel screening strategy is often defined by the following criteria:
- Ligand Pharmacology: Agonist and antagonists of NRs often describe ligands that either activate or repress transcription in certain tissues. Cell-free binding and cellbasedtranscription reporter assays are important methods for lead evaluation.
- Gene profile: Depending on the disease indications, discovery programs may require the monitoring of gene profiles in mutiple cell types through transcriptional profiling of compounds.
- Species selectivity: Commonly, there are efficacy differences between rodent and human NRs. Therefore, species selectivity NR panels are necessary to progress leading compounds and to interpret in vivo data.

- Protein kinases regulate significant aspects of cell life. There are about 518 identified protein kinases in the human genome [5]. Therefore, it is not surprising that selectivity is one major obstacle for the kinase drug discovery. Several protein kinase inhibitors that have been approved or entered human clinical trials are also not very specific for a single kinase. This leads to the concept that it is probably impossible to design completely selective ATP site-based inhibitors. It is perhaps more practical to design an inhibitor with a preferred kinase profile rather than a specific kinase inhibition profile. Protein kinase panel screeningstrategies are critical to the development of protein-kinase inhibitors. The following two factors have impact on establishing the kianse panel for lead optimization.
  - Protein kinases are bisubstrate enzymes, binding of ATP to protein kinases will affect binding of protein substrates (or peptides) and vice versa. Therefore, it is important todefine the appropriate kinetic parameters and understand the limitations of assay technologies that are chosen for a kinase lead evaluation panel.
  - Discovery of kinase inhibitors has been focused on three different types of inhibitors; ATP-competitive inhibitors, non-ATP or allosteric inhibitors, and inhibitors that bind inactivated kinases, and therefore the assays need to reflect these mechanistic approaches.

Lead Profiling Assays

The type of lead profiling assays used is very dependent on the stage of the compounds in the drug discovery process. In the early stages of drug discovery there are usually multiple chemotypes, and the in vitro ADMET assays need to be able to assay relatively large numbers of compounds to help drive the early SAR. Under these circumstances, high throughput methods are critical to provide timely progression. Later in the process where single lead series has been identified, the ADMET assays are more focused on predicting in vivo effects and potential clinical liabilities.

Molecular docking

Molecular docking is a well established computational technique which predicts the interaction energy between two molecules. Molecular docking studies are used to determine the interaction of two molecules and to find the best orientation of ligand which would form a complex with overall minimum energy. The small molecule, known as ligand usually fits within protein's cavity which is predicted by the search algorithm. These protein cavities become active when come in contact with any external compounds and are thus called as active sites.

The results are analyzed by a statistical scoring function which converts interacting energy into numerical values called as the docking score; and also the interacting energy is calculated. The 3D pose of the bound ligand can be visualized using different visualizing tools like Pymol, Rasmol etc which could help in inference of the best fit of ligand. Predicting the mode of protein-ligand interaction can assume the active site of the protein molecule and further help in protein annotation. Moreover molecular docking has major application in drug discovery and designing.

This technique mainly incorporates algorithms like molecular dynamics, Monte Carlo stimulation, fragment based search methods.

Different types of Interactions

Interactions between particles can be defined as a consequence of forces between the molecules contained by the particles. These forces are divided into four categories:

• **Electrostatic forces** - Forces with electrostatic origin due to the charges residing in the matter. The most common interactions are charge-charge, charge-dipole and dipole-dipole.

• **Electrodynamics forces**-The most widely known is the Van der Waals interactions.

• **Steric forces** - Steric forces are generated when atoms in different molecules come into very close contact with one another and start affecting the reactivity of each other. The resulting forces can affect chemical reactions and the free energy of a system.

• **Solvent-related forces** - These are forces generated due to chemical reactions between the solvent and the protein or ligand. Examples are Hydrogen bonds (hydrophilic interactions) and hydrophobic interactions.

• A common characteristic of all these forces is their **electromagnetic** nature.

• Other physical factors - **Conformational changes** in the protein and the ligand are often necessary for successful docking.

Molecular docking

Molecular docking can be divided into two separate sections.

1) **Search algorithm** – These algorithms determine all possible optimal conformations for a given complex (protein-protein, protein-ligand) in a environment i.e. the position and orientation of both molecules relative to each other. They can also calculate the energy of the resulting complex and of each individual interaction.

The different types of algorithms that can be used for docking analysis are given below.

- Molecular dynamics
- Monte Carlo methods
- Genetic algorithms
- Fragment-based methods
- Point complementary methods
- Distance geometry methods
- Systematic searches

2) **Scoring function** – These are mathematical methods used to predict the strength of the non-covalent interaction called as binding affinity, between two molecules after they have been docked. Scoring functions have also been developed to predict the strength of other types of intermolecular interactions, for example between two proteins or between protein and DNA or protein and drug.

These configurations are evaluated using scoring functions to distinguish the experimental binding modes from all other modes explored through the searching algorithm.

For example:

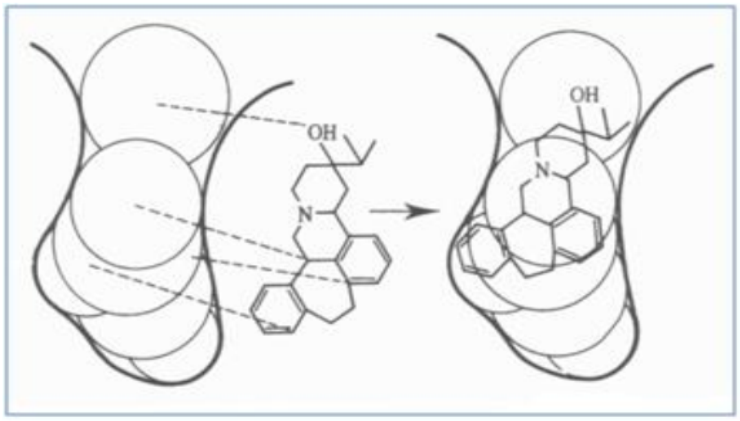
- Binding Energy

$\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elect} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol}$

General concept of the algorithm:

1) A 'negative' image of the binding site is made - a collection of spheres of varying radii, each touching the molecular surface at just 2 points.

2) Ligand atoms are then matched to sphere centers where at least four distances between ligand atoms are matched to sphere center distances.



3) Proper orientation is achieved by a least squares fit of ligand atoms to the sphere centers.

4) Orientation is checked for any steric clashes between ligand and receptor.

5) If acceptable, then interaction energy is computed as a 'score' for that binding mode

6) New orientations are obtained by matching different sets of atoms and sphere centers

7) Top-scoring orientations are retained for subsequent analysis

Types of Docking -

The following are majorly used type of docking are-

- **Lock and Key or Rigid Docking** – In rigid docking, both the internal geometry of the receptor and ligand is kept fixed during docking.
- **Induced fit or Flexible Docking -** In this model, the ligand is kept flexible and the energy for different conformations of the ligand fitting into the protein is calculated. Though more time consuming, this method can evaluate many different possible conformations which make it more reliable.

Major steps in molecular docking:

Step I – Building the Receptor

In this step the 3D structure of the receptor should be downloaded from PDB; and modified. This should include removal of the water molecules from the cavity, stabilizing charges, filling in the missing residues, generation the side chains etc according to the parameters available. After modification the receptor should be biological active and stable.

Step II – Identification of the Active Site

After the receptor is built, the active site within the receptor should be identified. The receptor may have many active sites but the one of the interest should be selected. Most of the water molecules and heteroatoms if present should be removed.

Step III – Ligand Preparation

Ligands can be obtained from various databases like ZINC, PubChem or can be sketched using tools like Chemsketch. While selecting the ligand, the LIPINSKY'S RULE OF 5 should be applied. The rule is important for drug development where a pharmacologically active lead structure is optimized stepwise for increased activity and selectivity, as well as drug-like properties, as described.

For the selection of a ligand using LIPINSKY'S RULE:

- Not more than 5 –H bond donors.
- Molecular Weight NOT more than 500 Da.
- Log P not over 5 for octanol water partition coefficient.
- NOT more than 10 H bond acceptors.

Step IV- Docking

This is the last step, where the ligand is docked onto the receptor and the interactions are checked. The scoring function generates scores depending on which the ligand with the best fit is selected.

Software available for Molecular Docking:

SCHRODINGER

DOCK

AUTOLOCK TOOLS.

DISCOVERY STUDIO.

iGemDock

Applications

A binding interaction between a small molecule ligand and an enzyme protein may result in activation or inhibition of the enzyme. If the protein is a receptor, ligand binding may result in agonism or antagonism. Docking is most commonly used in the field of drug design — most drugs are small organic molecules, and docking may be applied to:

hit identification – docking combined with a scoring function can be used to quickly screen large databases of potential drugs in silico to identify molecules that are likely to bind to protein target of interest (see virtual screening).

lead optimization – docking can be used to predict in where and in which relative orientation a ligand binds to a protein (also referred to as the binding mode or pose). This information may in turn be used to design more potent and selective analogs.

Bioremediation – Protein ligand docking can also be used to predict pollutants that can be degraded by enzymes.[12]

Quantitative structure–activity relationship

Quantitative structure–activity relationship models (**QSAR** models) are regression or classification models used in the chemical and biological sciences and engineering. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable.

In QSAR modeling, the predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be abiological activity of the chemicals. QSAR models first summarize a supposed relationship between chemical structures and biological activity in a data-set of chemicals. Second, QSAR models predict the activities of new chemicals.

Related terms include *quantitative structure–property relationships* (*QSPR*) when a chemical property is modeled as the response variable.

As an example, biological activity can be expressed quantitatively as the concentration of a substance required to give a certain biological response. Additionally, when physicochemical properties or structures are expressed by numbers, one can find a mathematical relationship, or quantitative structure-activity relationship, between the two. The mathematical expression, if carefully validated can then be used to predict the modeled response of other chemical structures.

A QSAR has the form of a mathematical model:

- Activity = *f*(physiochemical properties and/or structural properties) + error

The error includes model error (bias) and observational variability, that is, the variability in observations even on a correct model.

Types

Fragment based (group contribution)

Analogously, the "partition coefficient"—a measurement of differential solubility and itself a component of QSAR predictions—can be predicted either by atomic methods (known as "XLogP" or "ALogP") or by chemical fragment methods (known as "CLogP" and other variations). It has been shown that the logP of compound can be determined by the sum of its fragments; fragment-based methods are generally accepted as better predictors than atomic-based methods.[9] Fragmentary values have been determined statistically, based on empirical data for known logP values. This method gives mixed results and is generally not trusted to have accuracy of more than ±0.1 units.

Group or Fragment based QSAR is also known as GQSAR.[11] GQSAR allows flexibility to study various molecular fragments of interest in relation to the variation in biological response. The

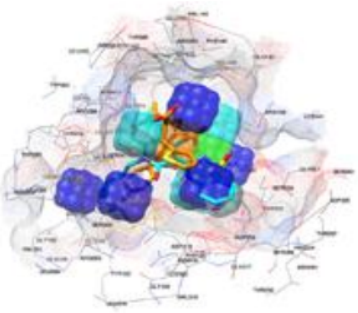
molecular fragments could be substituents at various substitution sites in congeneric set of molecules or could be on the basis of pre-defined chemical rules in case of non-congeneric sets. GQSAR also considers cross-terms fragment descriptors, which could be helpful in identification of key fragment interactions in determining variation of activity. Lead discovery using Fragnomics is an emerging paradigm. In this context FB-QSAR proves to be a promising strategy for fragment library design and in fragment-to-lead identification endeavours.[12]

An advanced approach on fragment or group-based QSAR based on the concept of pharmacophore-similarity is developed.[13] This method, pharmacophore-similarity-based QSAR (PS-QSAR) uses topological pharmacophoric descriptors to develop QSAR models. This activity prediction may assist the contribution of certain pharmacophore features encoded by respective fragments toward activity improvement and/or detrimental effects.

3D-QSAR

3D-QSAR refers to the application of force field calculations requiring three-dimensional structures, e.g. based on protein crystallography or moleculesuperimposition. It uses computed potentials, e.g. the Lennard-Jones potential, rather than experimental constants and is concerned with the overall molecule rather than a single substituent. It examines the steric fields (shape of the molecule), the hydrophobic regions (water-soluble surfaces), and the electrostatic fields.

The created data space is then usually reduced by a following feature extraction (see also dimensionality reduction). The following learning method can be any of the already mentioned machine learning methods, e.g. support vector machines.[16] An alternative approach uses multiple-instance learning by encoding molecules as sets of data instances, each of which represents a possible molecular conformation. A label or response is assigned to each set corresponding to the activity of the molecule, which is assumed to be determined by at least one instance in the set (i.e. some conformation of the molecule).



3-D QSAutogrid-R MPGRS example image

On June 18, 2011 the Comparative Molecular Field Analysis (CoMFA) patent has dropped any restriction on the use of GRID and partial least-squares (PLS) technologies and the Rome Center for Molecular Design (RCMD) team (www.rcmd.it) has opened a 3D QSAR web server (www.3d-qsar.com) based on the 3-D QSAutogrid/R engine.

GOLPE stands for Generating Optimal Linear PLS Estimations. 3-D QSAutogrid/R covers all the main features of CoMFA and GRID/GOLPE with implementation by multiprobe/multiregion variable selection (MPGRS) that improves the simplification of interpretation of the 3-D QSAR map. The methodology is based on the integration of the molecular interaction fields as calculated by AutoGrid and the R statistical environment that can be easily coupled with many free graphical molecular interfaces such as UCSF-Chimera, AutoDock Tools, JMol and others.

Chemical descriptor based

In this approach, descriptors quantifying various electronic, geometric, or steric properties of a molecule are computed and used to develop a QSAR. This approach is different from the fragment (or group contribution) approach in that the descriptors are computed for the system as whole rather than from the properties of individual fragments. This approach is different from the 3D-QSAR approach in that the descriptors are computed from scalar quantities (e.g., energies, geometric parameters) rather than from 3D fields.

An example of this approach is the QSARs developed for olefin polymerization by half sandwich compounds.

Modeling

In the literature it can be often found that chemists have a preference for partial least squares (PLS) methods, since it applies the feature extraction andinduction in one step.

Data mining approach

Computer SAR models typically calculate a relatively large number of features. Because those lack structural interpretation ability, the preprocessing steps face afeature selection problem (i.e., which structural features should be interpreted to determine the structure-activity relationship). Feature selection can be accomplished by visual inspection (qualitative selection by a human); by data mining; or by molecule mining.

A typical data mining based prediction uses e.g. support vector machines, decision trees, neural networks for inducing a predictive learning model.

Molecule mining approaches, a special case of structured data mining approaches, apply a similarity matrix based prediction or an automatic fragmentation scheme into molecular substructures. Furthermore there exist also approaches using maximum common subgraph searches or graph kernels.

Matched molecular pair analysis

Main article: Matched molecular pair analysis

Typically QSAR models derived from non linear machine learning is seen as a "black box", which fails to guide medicinal chemists. Recently there is a relatively new concept of Matched molecular pair analysis or Prediction driven MMPA which is coupled with QSAR model in order to identify activity cliffs

Evaluation of the quality of QSAR models

QSAR modeling produces predictive models derived from application of statistical tools correlating biological activity (including desirable therapeutic effect and undesirable side effects) or physico-chemical properties in QSPR models of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative ofmolecular structure or properties. QSARs are being applied in many disciplines, for example: risk assessment, toxicity prediction, and regulatory decisions[25] in addition to drug discovery and lead optimization.[26] Obtaining a good quality QSAR model depends on many factors, such as the quality of input data, the choice of descriptors and statistical methods for modeling and for validation. Any QSAR modeling should ultimately lead to statistically robust and predictive models capable of making accurate and reliable predictions of the modeled response of new compounds.

For validation of QSAR models, usually various strategies are adopted:

1. internal validation or cross-validation (actually, while extracting data, cross validation is a measure of model robustness, the more a model is robust (higher q2) the less data extraction perturb the original model);
2. external validation by splitting the available data set into training set for model development and prediction set for model predictivity check;
3. blind external validation by application of model on new external data and
4. data randomization or Y-scrambling for verifying the absence of chance correlation between the response and the modeling descriptors.

The success of any QSAR model depends on accuracy of the input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose; for QSAR models validation must be mainly for robustness, prediction performances and applicability domain (AD) of the models.

Some validation methodologies can be problematic. For example, *leave one-out* cross-validation generally leads to an overestimation of predictive capacity. Even with external validation, it is difficult to determine whether the selection of training and test sets was manipulated to maximize the predictive capacity of the model being published.

Different aspects of validation of QSAR models that need attention includes methods of selection of training set compounds, setting training set size and impact of variable selection for training set models for determining the quality of prediction. Development of novel validation parameters for judging quality of QSAR models is also important.

Application

Chemical

One of the first historical QSAR applications was to predict boiling points.

It is well known for instance that within a particular family of chemical compounds, especially of organic chemistry, that there are strong correlations between structure and observed properties. A simple example is the relationship between the number of carbons in alkanes and their boiling points. There is a clear trend in the increase of boiling point with an increase in the number carbons, and this serves as a means for predicting the boiling points of higher alkanes.

A still very interesting application is the Hammett equation, Taft equation and pKa prediction methods.

Biological

The biological activity of molecules is usually measured in assays to establish the level of inhibition of particular signal transduction or metabolic pathways. Drug discovery often involves the use of QSAR to identify chemical structures that could have good inhibitory effects on specific targets and have low toxicity (non-specific activity). Of special interest is the prediction of partition coefficient log *P*, which is an important measure used in identifying "druglikeness" according to Lipinski's Rule of Five.

While many quantitative structure activity relationship analyses involve the interactions of a family of molecules with an enzyme or receptor binding site, QSAR can also be used to study the interactions between the structural domains of proteins. Protein-protein interactions can be quantitatively analyzed for structural variations resulted from site-directed mutagenesis.

It is part of the machine learning method to reduce the risk for a SAR paradox, especially taking into account that only a finite amount of data is available (see alsoMVUE). In general, all QSAR problems can be divided into coding and learning.

Applications

(Q)SAR models have been used for risk management. QSARS are suggested by regulatory authorities; in the European Union, QSARs are suggested by the REACHregulation, where "REACH" abbreviates "Registration, Evaluation, Authorisation and Restriction of Chemicals".

The chemical descriptor space whose convex hull is generated by a particular training set of chemicals is called the training set's applicability domain. Prediction of properties of novel chemicals that are located outside the applicability domain uses extrapolation, and so is less reliable (on average) than prediction within the applicability domain. The assessment of the reliability of QSAR predictions remains a research topic.

The QSAR equations can be used to predict biological activities of newer molecules before their synthesis.