QSAR – A Tool for Drug Discovery and Development

Kunal Roy

Drug Theoretics and Cheminformatics Lab
Division of Medicinal and Pharmaceutical Chemistry
Department of Pharmaceutical Technology
Jadavpur University, Kolkata 700 032 (India)

Email: kunalroy_in@yahoo.com
URL: http://www.geocities.com/kunalroy_in
Tel: +91-9831594140, Fax: +91-33-2837 1078
Background

• Exploration for lead and its exploitation have been the two mainstays of medicinal chemistry.

• The principal aim of a medicinal chemist is to discover novel drugs with greater potency and reduced toxicity which may be achieved by molecular modification or tailoring of existing drugs, optimization of various lead compounds, isolation of active constituents from natural sources or syntheses of new series of compounds.

Background

• Identification of a lead compound for a particular activity is a real problem in drug design.

• Recognition of biochemical principles of drug action is a prerequisite for drug discovery process.
Background

A rational explanation of drug action is often limited by our ability to correlate the observed physiological effects with a reasonable hypothesis or concept.

Various structural, physicochemical, and biological parameters are used to correlate those with biological activity and the observed relations are used to predict activity of a new compound and this information is exploited to develop newer molecules of optimum activity.

Such correlation may also help in exploring the mechanistic features of biological activity.

Background

• Biological activity is an aftermath of various interactions of a bioactive substance at critical reaction site in the biological system that occurs simultaneously with complex series of events encompassing pharmacokinetics of the substance.

• Drugs have varied structures, diverse biological activities, and multifarious modes of action. Despite the certainties of chemical structures of drugs and their biological activities, their mechanistic aspects are overshadowed by a great degree of uncertainties of the intervening steps between drug administration and response that make the drug response phenomenon a complicated multistep process.

**Background**

- Drug molecules have to confront the uncertainties of absorption, transport, metabolism, excretion and, above all, the ‘random walk’ to the critical reaction site and their subsequent adsorption and binding with the receptor.

- To circumvent the complexity, the biological aspect of the disease and/or drug action should be understood to the finest level, as far as possible.

**Background**

Scheme I: Schematic representation of various pharmacokinetic processes
Background

The various factors regulating pharmacokinetics and pharmacodynamics of a bioactive substance are considered for mechanistic interpretations of the activity. Once the factors are identified and their relationships with the intervening steps between drug administration and biological response are established, the process of drug design and tailoring and/or modifying structures of drugs becomes easier to be carried out.

Drug Discovery and Development

Earlier, drugs were designed by systemic modification of chemical precursors using standard tools of medicinal chemistry. But, the approach of Edisonian research for synthesizing organic molecules with an objective of obtaining medicinal compounds with desired biological activities is not effective in the present days, in view of heavier demands to be met by the new molecules. Random synthesis is quite time consuming and expensive, and also failure rate is often high in this approach, as it does not adequately and properly utilise the information obtainable from the compounds already synthesized or available.

Drug Discovery and Development

As in the process, the prospective drug substances have to cross long and rigid methodologies of tests and should satisfy all the requirements, which makes the probability of success very little. Thus, pharmaceutical research and drug discovery involves a gamble at a very high stake. The rational approach of drug designing is, therefore, a natural choice to enhance probability of success as well as to minimize labour, time, and cost. The quantitative aspects of the biological activity and the mathematical relationships existing between the biological activity (BA), chemical structure (C) and physicochemical properties (P) must be understood for the rational drug design (RDD).

Drug Discovery and Development

The process of drug development is time consuming and costly affair that can no longer be satisfied by classical and empirical mode of research. It requires approx. 750 million dollars and about 12 to 15 y to bring a drug to market. The chance of discovery of a new agent has diminished to 1 in 10000 and the situation is even more unfavourable with anticancer and antiviral agents.

The time and cost requirement of drug development process are due to thoroughness and caution prescribed. The various stages of classical drug development process are as follows:

Drug Discovery and Development

i) Synthesis of compounds and their initial screening for pharmacological activity: thousands of compounds are synthesized and subjected to *in vitro* and *in vivo* pharmacological screening in search of the best candidates for the subsequent step.

(ii) The requisite preclinical animal studies of the selected compounds (about dozens) for both short-term and long-term toxicity.

**Drug Discovery and Development**

(iii) Phase I clinical trials in healthy volunteers of the few compounds (two or three) selected from step (ii).

(iv) Phase II clinical trials in limited cohort of patients with target disease: some of the compounds under clinical trial are usually eliminated from further consideration when unforeseen side effects occur.

(v) Phase III clinical trial in broad population of affected patients.

Drug Discovery and Development

After stringent criteria have been satisfactorily met in each of the above steps, FDA approval is given so that a compound can be marketed. A significant cost built into the drug development process is the expense of synthesis and testing of the unsuccessful drug candidates. To reduce the cost and time requirement, the probability of obtaining potential and prospective agents should be increased.

Structural specificity in drug action

Chlorpromazine (Antipsychotic)

Promethazine (Antihistaminic)

Ethopropazine (Antiparkinsonian)

Tolbutamide (shorter acting hypoglycemic)

Chlorpropamide (longer acting hypoglycemic)
Structural specificity in drug action

Epinephrine (Hypertensive)

Isoproterenol (Hypotensive)

Diethylstilboestrol
Structural specificity in drug action

(-)Epinephrine

R = -CH₂-CH₂-CH₂CO

4-(4-Hydroxypiperidino)-4'-fluoro-butyrophenone
Quantitative Structure-Activity Relationship (QSAR) Studies

This non-experimental part of drug design encompassing study of both structure-activity and structure-property relations in broad sense is an intellectual exercise of assembling, manipulating and examining data obtained from physical, chemical and biological experiments and correlating these to biological activity. Biological activity of a drug depends on the types and magnitude of interactions between the receptor and the drug molecule. Various structural attributes of the drug molecule like electronic distribution, steric feature, etc., are the determining factors regulating the interactions. All Quantitative Structure-Activity Relationship (QSAR) studies are based on the notion that $BA$ is function of $C$ and/or $P$.

$$BA = f (C, P) \quad ....(1)$$
Quantitative Structure-Activity Relationship (QSAR) Studies

The goals of QSAR studies include better understanding of the modes of actions, prediction of new analogs with better activity, and optimization of the lead compound to reduce toxicity and increase selectivity.

The knowledge of biological system, various factors regulating physiological processes and those contributing to pathological states, a thorough examination of molecular structures of drugs and their properties and unearthing of the factors modulating biological activity of drugs are required to find out the biochemical rationale of drug action. Such understanding helps to develop more effective drugs in a scientific way potentially reducing the cost of drug discovery, time, and manpower requirement (Table 1).
Any physicochemical property which can be related to an individual molecule, can be predicted by SAR. One-to-one correspondence input of the property value and respective molecular structures.
Flow Chart of QSAR Model

1. Preparation of Input DATA (Retention value, Structures)
2. Calculation of Descriptors
3. Statistical Analysis (Feature selection, regression)
4. Interpretation, validation And Prediction
5. 3D Geometry Optimization (conformation, alignment)
6. QSAR Model Building

Data Matrix

<table>
<thead>
<tr>
<th>Data Matrix</th>
<th>descriptor1</th>
<th>descriptor2</th>
<th>descriptor3</th>
<th>...</th>
<th>descriptor m</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule 1</td>
<td>x11</td>
<td>x12</td>
<td>x13</td>
<td></td>
<td>x1m</td>
<td>ln 1/C</td>
</tr>
<tr>
<td>Molecule 2</td>
<td>x21</td>
<td>x22</td>
<td>x23</td>
<td></td>
<td>x2m</td>
<td>ln 1/C</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Molecule n</td>
<td>x n1</td>
<td>x n2</td>
<td>x n3</td>
<td></td>
<td>x nm</td>
<td>ln 1/C</td>
</tr>
</tbody>
</table>

Experiment

Predicted

Interpretation, validation And Prediction
\[
R^2 = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{calc}})^2}{\sum (Y_{\text{obs}} - \bar{Y})^2}
\]
\[ R^2_a = \frac{(n-1)R^2 - p - 1}{n - p - 1} \]
\[ F = \frac{\frac{p}{\sum (Y_{obs} - Y_{cal})^2}}{\frac{n - p - 1}{\sum (Y_{cal} - \bar{Y})^2}} \]
$$Q^2 = 1 - \frac{\sum (Y_{obs} - Y_{pred})^2}{\sum (Y_{obs} - \bar{Y})^2}$$
\[ R^2_{pred} = 1 - \frac{\sum (Y_{pred\ (Test)} - Y_{(Test)})^2}{\sum (Y_{(Test)} - Y_{training})^2} \]
## QSAR

### Table 1 - Representative examples of correct predictions from QSARs

<table>
<thead>
<tr>
<th>Types of compounds</th>
<th>Biological activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzothiadiazines</td>
<td>Antihypertensive</td>
</tr>
<tr>
<td>Clonidine analogs</td>
<td>Antihypertensive</td>
</tr>
<tr>
<td>β-Carbolines</td>
<td>Inhibition of monoamine oxidase</td>
</tr>
<tr>
<td>Sulphonamides</td>
<td>Inhibition of carbonic anhydrase</td>
</tr>
<tr>
<td>Carbamoylpiperidines</td>
<td>Inhibition of cholinesterase</td>
</tr>
<tr>
<td>Pyrazoles</td>
<td>Antivirals</td>
</tr>
<tr>
<td>Nitrosoureas</td>
<td>Cytostatic</td>
</tr>
<tr>
<td>Mytomycins</td>
<td>Cytostatic</td>
</tr>
<tr>
<td>Sulfonamides</td>
<td>Antibacterial</td>
</tr>
<tr>
<td>Erythromycins</td>
<td>Antibacterial</td>
</tr>
<tr>
<td>Quinoxaline 1,4-dioxides</td>
<td>Antibacterial</td>
</tr>
<tr>
<td>Promazines</td>
<td>Neuroleptic</td>
</tr>
<tr>
<td>Benzothiepine derivatives</td>
<td>Neuroleptic</td>
</tr>
<tr>
<td>Thyroxine analogs</td>
<td>Thyroxine</td>
</tr>
<tr>
<td>Azapurine-6-ones</td>
<td>Immunosuppresives</td>
</tr>
<tr>
<td>Triazines</td>
<td>Inhibition of dihydrofolate reductase</td>
</tr>
<tr>
<td>Adenosine analogs</td>
<td>Antihypertensives</td>
</tr>
<tr>
<td>Trimethoprim analogs</td>
<td>Antibacterial</td>
</tr>
</tbody>
</table>
Table 2 - Representative examples of correct predictions from QSARs

<table>
<thead>
<tr>
<th>Names of compounds</th>
<th>Biological activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norfloxacin</td>
<td>Antibacterial</td>
</tr>
<tr>
<td>Metconazole</td>
<td>Fungicide</td>
</tr>
<tr>
<td>Ipconazole</td>
<td>Fungicide</td>
</tr>
<tr>
<td>Flobufen</td>
<td>Antiinflammatory</td>
</tr>
<tr>
<td>Bifenturin</td>
<td>Pesticide</td>
</tr>
</tbody>
</table>
Applications of QSAR

- Prediction of activity
- Diagnosis of mechanism of drug action
- Classification
- Lead compound optimization
- Modeling of ADME parameters
- Ecotoxicological modeling
- Property prediction
**QSAR methodologies**

**Classical QSAR**
Free-Wilson analysis

Free and Wilson model is a true structure-activity relationship model. It is *de novo* mathematical model that finds out contributions of various substituents and the parent ring to the biological activity through regressional method. Its limitation is that it cannot predict contribution of any substituent that is absent in the original data set.

\[ pC = a_o + \sum a_i x_i \]

Fujita-Ban Analysis
A mathematical model that describes presence and absence of certain structure features, e.g., those groups that are chemically modified, by values of 1 or 0 and correlates the resulting structural matrix with biological activity values.

\[
\text{Biological Activity} = \sum a_i + u
\]

Biological Activity = \(\log \frac{1}{C}\), \(C\), drug concentration causes \(EC_{50}\), \(I_{50}\), etc.

\[
\log \frac{1}{C} = -0.301 (\pm 0.50)[m-F] + 0.207 (\pm 0.29)[m-Cl]
\]
\[
+ 0.434 (\pm 0.27)[m-Br] + 0.579 (\pm 0.50)[m-I]
\]
\[
+ 0.454 (\pm 0.27)[m-Me] + 0.340 (\pm 0.30)[p-F]
\]
\[
+ 0.768 (\pm 0.30)[p-Cl] + 1.020 (\pm 0.30)[p-Br]
\]
\[
+ 1.429 (\pm 0.50)[p-I] + 1.256 (\pm 0.33)[p-Me]
\]
\[
+ 7.821 (\pm 0.27)
\]

\((n=22, r=0.969, s=0.194, F=16.99)\)
**QSAR methodologies**

**Classical QSAR**

Hansch analysis

Hansch model is one of the most successfully applied methods in the field of QSAR and RDD. It was developed, based on the following postulates:

(i) Drug reaches near the receptor site by “random walk”, i.e., crossing various lipid barriers by passive diffusion process.

(ii) Drug binds with the receptor (critical reaction site) forming a complex.

(iii) The drug-receptor complex may undergo chemical reaction or conformational changes for the desired activity.

(iv) The drugs in a congeneric series act through same mechanism of action.

\[
\log 1/c = k_1\pi - k_2\pi^2 + k_3\sigma + k_4E_s + k_5
\]
**QSAR methodologies**

Classical QSAR

Bilinear models

\[ \log 1/c = a \log P - b \log(\beta P + 1) + c. \]

\[ \log 1/c = a \pi - b \log(\beta 10^\pi + 1) + c. \]
QSAR methodologies

Classical QSAR

Hansch analysis

(i) Hydrophobicity - Lipids being important constituents of all kinds of membranes, hydrophobicity of drug is an important parameter influencing absorption of drug from the site of administration and its partitioning to different compartments of the body (distribution pattern) and finally interaction with the receptor site which may have lipophilic area for hydrophobic interaction with the drug. However, optimum lipophilicity is required to maintain sufficient concentration of drug in extracellular fluid. Lipid solubility also plays an important role during elimination process and in determining half life of a drug. Hydrophobicity is mostly expressed in terms of partition coefficient (log P) using n-octanol-water system. Various chromatographic parameters like $R_M$, $\log K'$, etc., also have been used instead of log P.
**QSAR methodologies**

### Classical QSAR

**(ii) Electronic Influence** - Various electronic influences like dispersion forces, charge transfer interactions, electrostatic interactions, hydrogen bonding, polarization effects, and acid-base catalysis influence biological activity. Many interactions may occur through multiplicity of mechanisms.

**(iii) Steric Influence** - Various steric effects like intramolecular steric influences of substitutions on molecular properties, specific influence on the fitting to the receptor connected with the bulk and spatial arrangements of the substituents, conformational influence, and receptor requirement for specific steric configuration play an important role in drug action.
**QSAR methodologies**

**Classical QSAR**
Topological schemes

These are based on graph theoretic approach and mostly deal with hydrogen suppressed graphs. Topological consideration includes number and types of atoms and bonds, interatomic connections (adjacency count), paths, branching, molecular size, shape, functionality, etc.

*Wiener*

*Balaban J*

*Hosoya Z*

*Zagreb*

*Molecular connectivity indices*

*Kappa shape index*

*E-state index*
Topological schemes

- These are based on graph theoretic approach and mostly deal with hydrogen suppressed graphs. Topological consideration includes number and types of atoms and bonds, interatomic connections (adjacency count), paths, branching, molecular size, shape, functionality, etc.

\[
\text{CH}_3\text{C}\equiv\text{CH}_3
\]
QSAR methodologies

Statistical methods

- Methods of least squares
- Partial least squares
- Discriminant analysis
- Cluster analysis
- Genetic algorithm
- Factor analysis
- Neural network
Molecular Modeling

In recent few years, drug research has witnessed explosive growth of the field of molecular modelling and computer aided drug design (CADD). Nowadays, it a critical component of RDD. Molecular modeling is a visual interface between the computer and the scientists and it attempts to rationalize the behaviour and activity of bioactive agents. Its components are:

(i) Molecular Graphics - It represents drug molecules and associated molecular properties in a visual way.

(ii) Computational Chemistry - It involves simulation of atomic or molecular properties of compounds of medicinal interest through equations and solving these through computer by either molecular mechanic or quantum mechanical approach.

(iii) Statistical Modeling - It encompasses the QSAR and QSPR studies.

(iv) Molecular Data and Information Management - It includes compilation of databases of properties and synthesis strategies of a large number of compounds, capable of being searched by an user according to his need.
### Computational Chemistry

Quantum chemical methods
Molecular mechanical methods

#### Table 3 - List of selected popular software packages for molecular modelling

<table>
<thead>
<tr>
<th>Molecular mechanics</th>
<th>Ab initio quantum mechanics</th>
<th>Semiempiric quantum mechanics</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>GAUSSIAN</td>
<td>MOPAC</td>
</tr>
<tr>
<td>CHARMm</td>
<td>GAMESS</td>
<td>AMPAC</td>
</tr>
<tr>
<td>Discover</td>
<td>HONDO</td>
<td>PCILIO</td>
</tr>
<tr>
<td>MM2 / MM3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(i) **Comparative Molecular Field Analysis (CoMFA)** - CoMFA is most often employed receptor independent 3-D QSAR approach. The development of CoMFA is based on the concept that biological activity is sensitive to spatially localized differences in molecular field intensities.

(ii) **Molecular Shape Analysis (MSA)** - The common overlap steric volume (COSV) between a pair of superimposed molecules can be used as a global measure of molecular shape similarity. The goals of MSA are to identify the biologically relevant conformation without knowledge of receptor geometry and then in a quantitative fashion explain the activity of a series of analogs using structure-activity table.

(iii) **Molecular Similarity Matrices** - This approach is similar to CoMFA and assumes that the alignments and conformations used in the analysis are correct ones. It is based on comparing each molecule in the training set with each other. Various indices used in the construction of similarity matrices are Carbo index, Meyer shape index, Hodgkin index, and Spearman rank correlation coefficient.
(iv) **Distance Geometry** - The use of interatomic distances as representative of molecular shape has also shown success in 3-D QSAR. The four important methods of 3-D QSAR based on distance geometry are Ensemble distance geometry, site pocket model, REMOTEDISC and Veronoi site modelling.

(v) **Hypothetical Active-site Lattice Model (HASL)** - It is related to both CoMFA and MSA. The two aims of the HASL approach are the prediction of activities of untested compounds and identification of substructures influencing observed activities.

(vi) **Receptor surface analysis** - Receptor surface analysis (RSA) provides compact and quantitative descriptors that capture three-dimensional information about a putative receptor site.
Other topics under Molecular Modeling

1. Protein modeling

2. De novo ligand design

3. Receptor mapping and Pharmacophore search
Alignment of Molecules

• RMS atoms alignment
  – pair-wise model alignments based on the superposition of a set of matching atoms
  – require that for each pair of modes the corresponding atoms be identified.
  – relatively fast and accurate for models where a strong correspondence between specific set of atoms can be readily identified.

• Moments alignment
  – model alignment using either electrostatic moments or principal moments of inertia
  – a field similarity calculation is made to determine the best alternative orientation for these moment vectors
  – the fastest, but least accurate of the alignment methods
  – useful for tasks such as placing models in protein cavities

• Field alignment
  – model alignment by maximizing the overlap between the steric and electrostatic fields calculated about the models using a probe potential.
  – slower and less accurate than RMS
  – advantage of not requiring definition of any matching atoms
Comparative Molecular Field Analysis

CoMFA

- Align the molecules in the training set using some suitable alignment strategy.
- Create a cubic lattice of points around the molecules.
- Compute steric and electrostatic interaction energies using a probe such as a pseudo methyl group with a unit positive charge.
- Fit partial least square model to the biological response and the interaction energies.
- Make predictions for a test set, and visualize the results as contour plots on displays of the individual molecules in the set.
Grid maps for CoMFA and Field Calculation

Electrostatic field

$$E_C = \sum_{i=1}^{Natoms} \left[ \frac{Q_i Q_j}{D_{ij} R_{ij}} \right]$$

Steric field

$$E_{vdw} = \sum_{i=1}^{Natoms} \left[ \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} \right]$$

Biological Activity = $a \cdot \text{Steric0001} + b \cdot \text{Steric0002} + \ldots + m \cdot \text{Steric4913}$

+ $n \cdot \text{Elec0001} + \ldots + y \cdot \text{Elec4913} + \text{intercept}$
Additional Field in 3D QSAR

- Interaction energies with functional groups
  - GRID software
- Hydrophobic field
  - HINT software
- Molecular Lipophilicity Field
  - CLIP software
- Hydrogen bond donor and acceptor similarity field
  - CoMSiA in Sybyl software
### Applications of Molecular Modeling

**Table 4 - Representative examples of successful applications of molecular modeling**

<table>
<thead>
<tr>
<th>Types of compounds</th>
<th>Biological activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ro 46-6240</td>
<td>Thrombin inhibition</td>
</tr>
<tr>
<td>BCX-34</td>
<td>Purine nucleoside phosphorylase inhibition</td>
</tr>
<tr>
<td>Thymitaq (AG337)</td>
<td>Thymidylate synthetase inhibition</td>
</tr>
<tr>
<td>Trusopt (MK-507)</td>
<td>Carbonic anhydrase inhibition</td>
</tr>
<tr>
<td>Tolrestat</td>
<td>Aldose reductase inhibition</td>
</tr>
<tr>
<td>Ritonavir</td>
<td>HIV-1 protease inhibition</td>
</tr>
<tr>
<td>DX-9065a</td>
<td>Factor Xa inhibition</td>
</tr>
<tr>
<td>Marimastat</td>
<td>Matrix metalloproteinase inhibition</td>
</tr>
</tbody>
</table>
### Applications of Molecular Modeling

#### Table 5 - Representative examples of successful applications of molecular modelling

<table>
<thead>
<tr>
<th>Types of compounds</th>
<th>Biological activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Losartan</td>
<td>Angiotensin II receptor antagonist</td>
</tr>
<tr>
<td>Zolmitriptan</td>
<td>5HT-1 antagonist</td>
</tr>
<tr>
<td>Indinavir</td>
<td>HIV-1 Protease inhibitor</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>HIV-1 Protease inhibitor</td>
</tr>
</tbody>
</table>
Please visit

http://www.geocities.com/kunalroy_in/