Research Paper

Volume : 4 | Issue : 1 | January 2015 • ISSN No 2277 - 8179

# QSAR Modeling for Drug Discovery and Development: Applications and Methodology

Computer Science

**KEYWORDS :** QSAR Modeling, structure-activity relationship, regression modeling, Molecular Descriptors

**Bhagavati Parekh** BCA & PGDCA College, Junagadh

## ABSTRACT

*QSARs, or quantitative structure–property relationships (QSPRs), are mathematical models that attempt to relate the structure-derived features of a compound to its biological or physicochemical activity. 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. There are different methods and techniques to devise the QSAR models. There are a lot of applications available for developing in-silico QSAR Models and other related tasks. This paper provides an insight into QSAR Modeling concepts and the related software.*

**1. Introduction:**

QSARs, or quantitative structure–property relationships (QSPRs), are mathematical models that attempt to relate the structure-derived features of a compound to its biological or physicochemical activity. These models are used to predict the biological activities or class of compounds before the actual biological testing takes place. They can be used even in the structural characterization that can reveal the properties of interest.

**2. Applications and usefulness:**

QSAR techniques basically aim at predicting the biological activities, hence they prove to be useful in a lot many areas. The below mentioned is the list of potential uses of various QSAR applications:

- The optimization of various activities in the field of pharmacology, biology and pesticides
- The rational identification of new leads with all above activities
- The identification of hazardous compounds in the early stage of the discovery and development.
- The inventory screening of the existing compounds
- The identification of Toxicity and other such hazardous properties in the new compounds
- The prediction of various properties including Toxicity through various types of exposures
- The selection of compounds with the optimal pharmacokinetic as well as ADME properties

**Various types of QSAR Models:**

QSAR models can be classified in two categories: QSAR regression models or QSAR classification models. Just like other regression models, QSAR regression models also relate a set of predictor variables (X) to the potency of the response variable (Y). In case of QSAR classification models, the predictor variables are related to the categorical value of the response variable.

**Predictor variables:** In QSAR modeling, predictor variables are the physico-chemical properties of theoretical molecular descriptors of chemicals/compounds.

**Response variables:** The response variables could be the biological activities of the chemical.

| Sr. No. | Type of Model | Description |
|---|---|---|
| 1. | Fragment based (group contribution) | There are two main ways, the "partition coefficient"—a measurement of differential solubility and itself a component of QSAR predictions—can be predicted: either by using the atomic methods, in which case, it is known as "XLogP" or "ALogP", or by implementing the chemical fragment methods in which case, it is known as "CLogP" or other related variations |
| 2. | 3D QSAR | These approaches use the three dimensional properties of the legands to predict their biological activities using robust chemometric techniques such as PLS, ANN etc. |
| 3. | Chemical descriptor based | Under this approach, various descriptors quantifying various electronic, geometric, or steric properties of a catalyst are computed and used to develop a QSAR |

**Various methods for developing QSAR Models:**

| Sr. No. | Category | Methods |
|---|---|---|
| 1. | Methods for regression problems | 1. Multiple Linear Regression<br>2. Partial Least Squares<br>3. Feedforward backpropagation neural network<br>4. General regression neural network<br>5. Gaussian processes |
| 2. | Methods for classification problem | 1. Logistic regression<br>2. Linear discriminant analysis<br>3. Decision tree and random forests<br>4. k-nearest neighbor<br>5. Probabilistic neural network<br>6. Support vector machine |

**Various software for developing *in-silico* QSAR Models and other related tasks:**

As, the QSAR is serving as an important modeling tool, there is a huge development in the relevant software tools, both commercial and freeware, which are available on various resources. These include specialized software for drawing chemical structures; inter converting chemical file formats, generating 3D structures, calculating chemical descriptors, developing QSAR models, and general-purpose software that have all the necessary components for QSAR development.

**Software for drawing Chemical Structures/converting files:**

**Chemdraw:** It offers a cluster of software which can run on various platforms. It offers various software, which provides the robust, built in applications and features to scientists, chemists and biologists with an up to date collection of scientifically intelligent applications for chemical structure drawing and analysis combined with biological pathway drawing. Different software are: ChemDraw for ipad, ChemBioDraw, ChemDro Pro, ChemBioDraw Ultra etc. last two software are made available on the free trial basis

**ACD/ChemSketch:**

ACD/ChemSketch Freeware is a drawing package that allows you to draw chemical structures including organics, organometallics, polymers, and Markush structures.

OpenBabel:

Open Babel is an open-source program that enables users to search, convert files, analyze or store data from molecular modeling projects

**Software for 3D Structure Generation:**

**CORINA:** It can process a data set of 1,00,000 small to medium sized molecules with very high speed and accuracy. Its java based GUI contains the molecular editor as well as the 3D structure viewer

**PYMOL:** PyMOL is a powerful and comprehensive molecular visualization product for rendering and animating 3D molecular structures. It runs on Linux/OSX/Windows.

**CONCORD:** Concord is available as one of SYBYL applications. It is commercial software that converts 2D inputs into 3D structures rapidly.

**FROG:** Frog is a tool that results from a collaborative work involving several teams at RPBS. Frog is based on Frowns, a chemoinformatics toolkit written in python, to which several features have been added.

**Software for Descriptor calculation:**

**Molinspiration Suit Of Software:** This software suit containing numerous software offers a variety of freeware for supporting molecule manipulation and processing, including SMILES and SDfile conversion, normalization of molecules, generation of tautomers, molecule fragmentation, calculation of various molecular properties needed in QSAR, molecular modeling and drug design, high quality molecule depiction, molecular database tools supporting substructure and similarity searches.

**JSME Molecular Editor:** It supports various functionalities, which range from the property calculation to prediction of the Bioactivity and 3D Image generation

**ADRIANA.CODE:** It comprises a unique combination of methods for calculating molecular structure descriptors on a sound geometric and physicochemical basis. These descriptors can be used for a wide range of applications in all areas of chemistry, in particular in drug design.

**DRAGON:** It is an application for the calculation of molecular descriptors. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high-throughput screening of molecule databases. It is widely used in scientific studies as well as part of several QSAR suites. It calculates 4885 molecular descriptors in total. It keeps itself updating over the time.

**E-DRAGON:** E-DRAGON is the electronic remote version of the well known software DRAGON, which is an application for the calculation of molecular descriptors developed by the Milano Chemometrics and QSAR Research Group of Prof. R. Todeschini.

**Malconn-Z:** Molecular Descriptors Data Base is dedicated to be used for molecular descriptors applications in scientific research. The molecules that constitute the MOLE db - Molecular Descriptors Data Base are mainly collected from the NCI database, while the molecular descriptors have been calculated by means of DRAGON software.

**Software for Modeling:**

| | |
|---|---|
| 1. | KNIME: KNIME is a freeware. KNIME products include additional functionalities such as shared repositories, authentication, remote execution, scheduling, SOA integration and a web user interface as well as world-class support. Big data extensions are available for distributed frameworks such as Hadoop. KNIME is used by over 3000 organizations in more than 60 countries. |
| 2. | RapidMiner: Hundreds of data loading, data transformation, data modeling, and data visualization methods with access to a comprehensive list of data sources like Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files, and more. |
| 3. | WEKA : Weka is a collection of machine learning algorithms for data source software issued under the GNU General Public License mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Weka is open. |
| 4. | Orange : Orange is an Open source data visualization and analysis for novice and experts. Data mining using this software is carried out through the visual programming or Python scripting. It contains the Components for machine learning. It has a rich set of Add-ons for bioinformatics and text mining. The software package is also packed with features for data analytics. |
| 5. | Tanagran : TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license |
| 6. | MATLAB: It is a high-level language and interactive environment for numerical computation, visualization, and programming. Using MATLAB, one can analyze data, develop algorithms, and create models and applications. The language, tools, and built-in math functions enable users to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages, such as C/C++ or Java. MATLAB can be used for a range of applications, including signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology. More than a million engineers and scientists in industry and academia use MATLAB, which is popularly known as the language of technical computing. |

Volume : 4 | Issue : 1 | January 2015 • ISSN No 2277 - 8179

**General purpose software for QSAR:**

| | |
|---|---|
| 1. | SYBYL-X : Molecular Modeling from Sequence through Lead Optimization is made possible using this software suit. It offers a lot of tools for drug design and other molecular discovery projects, from HTS through Lead Optimization. All of the components for life science research are included as standard with the SYBYL-X suite including the library design, scaffold hopping, structure based design, ligand based design, some basic cheminformatics tools, or tools to build a protein model etc. One of the widely used QSAR techniques, Comparative Molecular Field Analysis (CoMFA), can be found as an integrated module in SYBYL. Besides ligand-based design, users may choose to integrate other SYBYL applications for receptor-based design, structural biology, library design or cheminformatics |
| 2. | Discovery Studio : It is the comprehensive predictive science application for the life sciences. It is possible with this interface to make Investigation and test hypotheses in silico prior to costly experimental implementation, thus reducing the time and expense involved in bringing products to market. It allows to drive the scientific exploration from target identification to lead optimization with a wealth of trusted life science modeling and simulation tools. |
| 3. | MOE : Molecular Operating Environment is a commercial package with fully integrated drug discovery software package. The Cheminformatics and (HTS) QSAR suite comes with pipeline tools to process SD files and calculation of over 600 molecular descriptors, model building, similarity searching and combinatorial library design. |
| 4. | CODESSA : The acronym CODESSA stands for COmprehensive DEscriptors for Structural and Statistical Analysis. It is the commercial software that combines various mathematical and computational systems to build QSAR models. CODESSA is also used for developing various models, and making cluster analysis of molecular descriptors. It also offers some tools to make model interpretation and compound property prediction from its chemical structure. |

**Conclusions:**

QSAR modeling produces predictive models derived from application of statistical tools correlating biological activity. 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.