3D PHAMARCOPHOMORE MODELLING PART III

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 modelling, the predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be a biological 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. Different properties or behaviors of chemical molecules have been investigated in the field of QSPR.

Some examples are quantitative structure–reactivity relationships (QSRRs), quantitative structure–chromatography relationships (QSCRs) and, quantitative structure–toxicity relationships (QSTRs), quantitative structure–electrochemistry relationships (QSERs), and quantitative structure–biodegradability relationships (QSBRs). 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.

Essential steps in QSAR studies

Principal steps of QSAR/QSPR include;

(i) Selection of Data set and extraction of structural/empirical descriptors
(ii) variable selection,

(iii) model construction and
(iv) validation evaluation.

SAR and the SAR paradox

The basic assumption for all molecule-based hypotheses is that similar molecules have similar activities. This principle is also called Structure–Activity Relationship (SAR). The underlying problem is therefore how to define a small difference on a molecular level, since each kind of activity, e.g. reaction ability, biotransformation ability, solubility, target activity, and so on, might depend on another difference. In general, one is more interested in finding strong trends. Created hypotheses usually rely on a finite number of chemicals, so care must be taken to avoid overfitting: the generation of hypotheses that fit training data very closely but perform poorly when applied to new data.

The *SAR paradox* refers to the fact that it is not the case that all similar molecules have similar activities.

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. 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. 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.

An advanced approach on fragment or group-based QSAR based on the concept of pharmacophore-similarity is developed. 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

The acronym **3D-QSAR** or **3-D QSAR** refers to the application of force field calculations requiring three-dimensional structures of a given set of small molecules with known activities (training set). The training set needs to be superimposed (aligned) by either experimental data (e.g., based on ligand-protein crystallography) or molecule superimposition software. 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. The first 3-D QSAR was named Comparative Molecular Field Analysis (CoMFA) by Cramer et al. It examined the steric fields (shape of the molecule) and the electrostatic fields which were correlated by means of partial least squares regression (PLS).

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. 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).

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.

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.

Modelling

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 and induction in one step.

Data mining approach

Computer SAR models typically calculate a relatively large number of features. Because those lack structural interpretation ability, the pre-processing steps face a feature 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, artificial 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

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 modelling 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 of

molecular structure or properties. QSARs are being applied in many disciplines, for example: risk assessment, toxicity prediction, and regulatory decisions in addition to drug discovery and lead optimization. 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 modelling and for validation. Any QSAR modelling should ultimately lead to statistically robust and predictive models capable of making accurate and reliable predictions of the modelled 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 modelling 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 include 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. 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 REACH regulation, where "REACH" abbreviates "Registration, Evaluation, Authorisation and Restriction of Chemicals". Regulatory application of QSAR methods includes *in silico* toxicological assessment of genotoxic impurities. Commonly used QSAR assessment software such as DEREK or CASE Ultra (MultiCASE) is used to genotoxicity of impurity according to ICH M7. 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.

Examples of machine learning tools for QSAR modelling include:

S.No.	Name	Algorithms
1.	R	RF,SVM, Naïve Bayesian, and ANN
2.	libSVM	SVM
3.	Orange	RF, SVM, and Naïve Bayesian
4.	RapidMiner	SVM, RF, Naïve Bayes, DT, ANN, and k-NN
5.	Weka	RF, SVM, and Naïve Bayes
6.	Knime	DT, Naïve Bayes, and SVM
7.	AZOrange	RT, SVM, ANN, and RF
8.	Tanagra	SVM, RF, Naïve Bayes, and DT
9.	Elki	k-NN
10.	MALLET	
11.	MOA	
12.	Deep Chem	Logistic Regression, Naive Bayes, RF, ANN, and others
13.	alvaModel	OLS, k-NN

References

1. Kim, K.H., Greco, G. and Novellino, E., *A critical review of recent CoMFA applications*, In Kubinyi, H., Folkers, G., and Martin, Y.C., (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 257–316.
2. Dunn III, W.J. and Hopfinger, A.J., *3D QSAR of flexible molecules using tensor representation*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 167–182.

3. Hahn, M. and Rogers, D., *Receptor surface models*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 117–134.
4. Heritage, T.W., Ferguson, A.M., Turner, D.B. and Willett, P., *EVA — a novel theoretical descriptor for QSAR studies*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht.

5. Klebe, G., *Comparative molecular similarity indices analysis — CoMSIA*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 87–104.
6. Walters, D.E., *Genetically evolved receptor models (GERM) as a 3D QSAR tool*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 159–166.
7. Wade, R.C., Ortiz, A.R. and Gago, F., *Comparative binding energy analysis*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 19–34.
8. Holloway, M.K., *A priori predicition of ligand affinity by energy minimization*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 63–84.
9. Todeschini, R. and Gramatica, P., *New 3D molecular descriptors: The WHIM theory and QSAR applications*, In Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 355–380.
10. Silverman, B.D., Platt, D.E., Pitman, M. and Rigoutsos, I., *Comparative molecular moment analysis (COMMA)*, in Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.) 3D QSAR in drug design: Vol. 3, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998, pp. 183–196.
11. Jain, A.N., Koile, K. and Chapman, D., *Compass: Predicting biological activities from molecular surface properties — performance comparisons on a steroid benchmark*, J. Med. Chem., 37 (1994) 2315–2327.
12. Martin, Y.C., Kim, K.-H. and Lin, C.T., *Comparative molecular field analysis: CoMFA*, In Charton, M. (Ed.) Advances in quantitative structure property relationships, JAI Press, Greenwich, CT, 1996, pp. 1–52.
13. Greco, G., Novellino, E. and Martin, Y.C., *Approaches to 3D-QSAR*, In Martin, Y.C. and Willett, P. (Eds.) Designing bioactive molecules: Three-dimensional techniques and applications, America Chemical Society, Washington, DC, 1997 (in press).
14. Ajay and Murcko, M.A., *Computational methods to predict binding free-energy in ligand–receptor complexes*, J. Med. Chem., 38 (1995) 4953–4967.
15. Kollman, P.A., *Advances and continuing challenges in achieving realistic and predictive*

simulations of the properties of organic and biological molecules, Acc. Chem. Res., 29 (1996) 461–469.
16. Bush, B.L. and Nachbar Jr., R.B., *Sample-distance partial least-squares — PLS optimized for many variables, with application to CoMFA*, J. Comput.-Aided Mol. Design, 7 (1993) 587–619.
17. Burger, A., *Medical chemistry — the first century*, Med. Chem. Res., 4 (1994) 3–15.
18. Willett, P., *Similarity and clustering techniques in chemical information systems*, Research Studies Press, Letchworth, 1987.
19. Hodgkin, E.E. and Richards, W.G., *Molecular similarity based on electrostatic potential and electric field*, Int. J. Quantum Chem., 14 (1987) 105–110.
20. Kier, L.B., *Molecular orbital theory in drug research*, Academic Press, New York, 1971, p. 258.