International Journal of Pharma and Bio Sciences V1(1)2010

QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR)

Abhilash M

Department of Biotechnology,The Oxford college of Engineering,Bangalore,INDIA.

Corresponding Author abhibiotek@gmail.com

Abstract

Quantitative structure—activity relationship (QSAR) represents an attempt to correlate structural descriptors of compounds with activities. The physicochemical descriptors include numerical parameters to account for physical and electronic properties, steric effect, topology fragment compositions, hydrophobicity of analogous compounds, as well as calculated properties of the three-dimensional (3D) structures of the compounds. The 3D properties include scalar parameters like solvent-accessible surface area or hydrophobic surface area. They also include field-type reductions of the structure that
represent steric interactions, electrostatic potentials, hydrogen-bonding potential, hydrophobic interactions, and so on.

Key words

QSAR, Types, 3D QSAR and Application.

Introduction

Quantitative structure-activity relationship (QSAR) (sometimes **QSPR**: quantitative structure-property relationship) is the process by which chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity.For example, biological activity can be expressed quantitatively as in the concentration of a substance required to give a certain biological response. Additionally, when physicochemical properties or structures are expressed by numbers, one can form a mathematical relationship, or quantitative structure-activity relationship, between the two. The mathematical expression can then be used to predict the biological response of other chemical structures.QSAR's most general mathematical form is:

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

QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR)

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 **S**tructure-**A**ctivity **R**elationship (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. A good example was given in the bioisosterism review of Patanie/LaVoie.[1].In general, one is more interested in finding strong trends. Created hypotheses usually rely on a finite number of chemical data. Thus, the induction principle should be respected to avoid overfitted hypotheses and deriving overfitted and useless interpretations on structural/molecular 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)

It has been shown that the logP of compound can be determined by the sum of its fragments. Fragmentary logP values have been determined statistically. This method gives mixed results and is generally not trusted to have accuracy of more than ±0.1 units.[2]

3D-QSAR

3D-QSAR refers to the application of force field calculations requiring three-dimensional structures, e.g. based on protein crystallography or molecule superposition. 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) and the electrostatic fields based on the applied energy function.[3].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.[4].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

For the coding usually a relatively large number of features or molecular descriptors is calculated, which can lack structural interpretation ability. In combination with the later applied learning method or as preprocessing step occurs a feature selection problem.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.[5][6]

Planning a QSAR study

When starting a QSAR study it is important to decide which physicochemical parameters are going

to be studied and to plan the analogues such that the parameters under study are suitably varied. For example, it would be pointless to synthesize analogues where the hydrophobicity and steric volume of the substituents are correlated,if these two parameters are to go into the equation.It is also important to make enough structures to make the results statistically meaningful. As a rule of thumb, five structures should be made for every parameter studied. Typically, the initial QSAR study would involve the two parameters π and σ, and possibly *Es*.Certain substituents are worth avoiding in the initial study since they may have properties other than those being studied. For example, substituents which might ionize (CO2H, NH2, SO2H) should be avoided. Groups which might easily be metabolized should be avoided if possible (e.g. esters or nitro groups).If there are two or more substituents, then the initial equation usually considers the total π and σ contribution.As more analogues are made, it is often possible to consider the hydrophobic and electronic effect of substituents at specific positions of the molecule. Furthermore, the electronic parameter a can be split into its inductive and resonance components *(F* and *R).* Such detailed equations may show up a particular localized requirement for activity. For example, a hydrophobia substituent may be favoured in one part of the skeleton, while an electron withdrawing substituent is favoured at another. This in turn gives clues about the binding interactions involved between drug and receptor.

Judging the quality of QSAR models

QSARs represent predictive models derived from application of statistical tools correlating biological activity (including desirable therapeutic effect and undesirable side effects) of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or properties. QSARs are being applied in many disciplines for example risk assessment, toxicity prediction, and regulatory decisions[7] in addition to drug discovery and lead optimization.[8] Obtaining a good quality QSAR model depends on many factors, such as the quality of biological data, the choice of descriptors and statistical methods. Any QSAR modeling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of biological activities of new compounds.

For validation of QSAR models usually four strategies are adopted:[9]

1. internal validation or cross-validation;
2. validation by dividing the data set into training and test compounds;
3. true external validation by application of model on external data and
4. data randomization or Y-scrambling.

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.[10] Leave one-out cross-validation generally leads to an overestimation of predictive capacity, and even with external validation, no one can be sure 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,[11] setting training set size[12] and impact of variable selection[13]

for training set models for determining the quality of prediction.Development of novel validation parameters for judging quality of QSAR models is also important.[14]

Application

Chemical

One of the first historical QSAR applications was to predict boiling points.[15].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.[16]

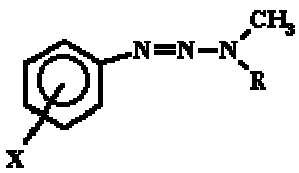
Biological

The biological activity of molecules is usually measured in assays to establish the level of inhibition of particular signal transduction or metabolic pathways. Chemicals can also be biologically active by being toxic. 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.[17].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 also MVUE). In general all QSAR problems can be divided into a coding[18] and learning.[19]

Drug Design

Researchers have attempted for many years to develop drugs based on QSAR. Easy access to computational resources was not available when these efforts began, so attempts consisted primarily of statistical correlations of structural descriptors with biological activities. However, as access to high-speed computers and graphics workstations became commonplace, this field has evolved into what is often termed rational drug design or computer-assisted drug design.

We will discuss the application of QSAR to drug design, some examples of which relied primarily on statistical correlation and some, on computer-based visualization and modeling. An early example of QSAR in drug design involves a series of 1-(X-phenyl)-3,3-dialkyl triazenes.



QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR)

These compounds were of interest for their anti-tumor activity, but they also were mutagenic. QSAR was applied to understand how the structure might be modified to reduce the mutagenicity without significantly decreasing the anti-tumor activity. Mutagenic activity was evaluated in the Ames test, and from those data, the following QSAR was developed:

$$\log\left(\frac{1}{C}\right) = 1.09\log P - 1.63\sigma^{+} + 3.06$$

where C is the molar concentration required to give 30 revertants per 10**8 bacteria and $\sigma^{+}$is a "through resonance" electronic parameter [8,9]. From the equation, it is seen that factors that favor mutagenicity are increased lipophilicity and electron-donating substituents.

Studies of the anti-tumor activity were done against L1210 leukemia in mice. From the data, the following QSAR was developed:

$$\log\left(\frac{1}{C}\right) = 0.10\log P - 0.042(\log P)^2 - 0.31\sigma^{+} - 0.18MR + 0.39E_s R + 4.12$$

where C is the molar concentration of compound producing a 40% increase in life span of mice, MR is molar refractivity, which is a measure of molecular volume, and EsR is a steric parameter for the R group **[10]**. Based on these equations, mutagenicity is more sensitive than anti-tumor activity to the electronic effects of the substituents. Thus, electron-withdrawing substituents were examined, as illustrated in the example below:

	log(1/C)	
4-X substituent	mutagenicity	antitumorigenicity
-H	5.75	3.58
$-SO_2HN_2$	3.15	3.48

By substituting a sulfonamide group at the para position, the anti-tumor activity was reduced 1.2-fold, whereas the mutagenicity was reduced by about 400-fold.

Applicability domain

As the use of (Q)SAR models for chemical risk management increases steadily and is also used for regulatory purposes (in the EU: Registration, Evaluation, Authorisation and Restriction of Chemicals), it is of crucial importance to be able to assess the reliability of predictions. The chemical descriptor space spanned by a particular training set of chemicals is called Applicability Domain. It offers the opportunity to assess whether a compound can be reliably predicted.

Conclusion

The QSAR approach attempts to identify and quantify the physicochemical propertiesof a drug and to see whether any of these properties has an effect on the drug's biological activity. If such a relationship holds true, an equation can be drawn up which quantifies the relationship and allows the medicinal chemist to say with some confidence that the property (or properties) has an important role in the distribution or mechanism of the drug. It also allows the medicinal chemist some level of prediction. By quantifying physicochemical properties, it should be possible to calculate in advance what the biological activity of a

novel analogue might be. There are two advantages to this. Firstly, it allows the medicinal chemist to target efforts on analogues which should have improved activity and thus cut down the number of analogues which have to be made. Secondly, if an analogue is discovered which does not fit the equation, it implies that some other feature is important and provides a lead for further development.

References

1. Patani GA, LaVoie EJ (December 1996). "Bioisosterism: A Rational Approach in Drug Design". *Chemical Reviews* **96** (8): 3147–3176. PMID 11848856.
2. Wildman SA, Crippen GM (1999), "Prediction of physicochemical parameters by atomic contributions", *J. Chem. Inf. Comput. Sci* **39** (5): 868–873,
3. Leach, Andrew R. (2001). *Molecular modelling: principles and applications*. Englewood Cliffs, N.J: Prentice Hall. ISBN 0-582-38210-6.
4. Vert, Jean-Philippe; Schlkopf, Bernhard; Schölkopf, Bernhard; Tsuda, Koji (2004). *Kernel methods in computational biology*. Cambridge, Mass: MIT Press. ISBN 0-262-19509-7.
5. Gusfield, Dan (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge, UK: Cambridge University Press. ISBN 0-521-58519-8.
6. Helma, Christoph (2005). *Predictive toxicology*. Washington, DC: Taylor & Francis. ISBN 0-8247-2397-X.
7. Tong W, Hong H, Xie Q, Shi L, Fang H, Perkins R (April 2005). "Assessing QSAR Limitations – A Regulatory Perspective". *Current Computer-Aided Drug Design* **1** (2): 195–205.
8. Dearden JC (2003). "In silico prediction of drug toxicity". *Journal of Computer-aided Molecular Design* **17** (2–4): 119–27. PMID 13677480.
9. Wold S, Eriksson L (1995). "Statistical validation of QSAR results". in Waterbeemd, Han van de. *Chemometric methods in molecular design*. Weinheim: VCH. pp. 309–318. ISBN 3-527-30044-9.
10. Roy, K. (2007), "On some aspects of validation of predictive quantitative structure-activity relationship models", *Expert Opin. Drug Discov.* **2** (12): 1567–1577,
11. Leonard JT, Roy K (2006), "On selection of training and test sets for the development of predictive QSAR models", *QSAR & Combinatorial Science* **25** (3): 235–251
12. Roy PP, Leonard JT, Roy K (2008), "Exploring the impact of size of training sets for the development of predictive QSAR models", *Chemometrics and Intelligent Laboratory Systems* **90** (1): 31–42,
13. Roy PP, Roy K (2008), "On some aspects of variable selection for partial least squares regression models", *QSAR & Combinatorial Science* **27** (3): 302–313
14. Roy PP, Paul S, Mitra I, Roy K (April 2009). "On Two Novel Parameters for Validation of Predictive QSAR Models". *Molecules* **14** (5): 1660–1701.
15. Rouvray, D. H.; Bonchev, Danail (1991). *Chemical graph theory: introduction and*

fundamentals. Tunbridge Wells, Kent, England: Abacus Press. ISBN 0-85626-454-7.

16. Fraczkiewicz R (2007). "In Silico Prediction of Ionization". *Comprehensive medicinal chemistry II*. Amsterdam: Elsevier. ISBN 0-08-044518-7.
17. Freyhult EK, Andersson K, Gustafsson MG (April 2003). "Structural modeling extends QSAR analysis of antibody-lysozyme interactions to 3D-QSAR". *Biophysical Journal* **84** (4): 2264–72.PMID 12668435.
18. Hendrik Timmerman; Todeschini, Roberto; Viviana Consonni; Raimund Mannhold; Hugo Kubinyi (2002). *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH. ISBN 3-527-29913-0.
19. Strok, David G.; Duda, Richard O.; Hart, Peter W. (2001). *Pattern classification*. Chichester: John Wiley & Sons. ISBN 0-471-05669-3.