A SHORT HISTORY OF QSAR EVOLUTION

by Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology,

DBSF, Insubria University, Varese, Italy

Paola.gramatica@uninsubria.it; http://www.qsar.it

QSARs (Quantitative) Structure–Activity relationships) are based on the assumption that the structure of a molecule (i.e. its geometric, steric and electronic properties) must contain the features responsible for its physical, chemical, and biological properties, and on the ability to represent the chemical by one, or more, numerical descriptor(s). By QSAR models, the biological activity (or property, reactivity, etc.) of a new or untested chemical can be inferred from the molecular structure of similar compounds whose activities (properties, reactivities, etc.) have already been assessed. The QSPR (Quantitative Structure–Property relationship) acronymous is used when a property is modeled.

It has been nearly 40 years since the QSAR modeling firstly was used into the practice of agrochemistry, drug design, toxicology, industrial and environmental chemistry . Its growing power in the following years may be attributed also to the rapid and extensive development in methodologies and computational techniques that have allowed to delineate and refine the many variables and approaches used in this modelling approach.

QSAR modelling is born in toxicology field. In fact, attempts to quantify relationships between chemical structure and acute toxic potency have been part of the toxicological literature for more than 100 years. In the defense of his thesis entitled “Action de l’alcohol amylique sur l’organisme” at the Faculty of Medicine, University of Strasbourg, Strasbourg, France on January 9, 1863, Cros noted that a relationship existed between the toxicity of primary aliphatic alcohols and their water solubility. This relationship demonstrated the central axiom of structure–toxicity modeling— the toxicity of substances is governed by their properties, which in turn are determined by their chemical structure. Therefore, there are interrelationships between structure, properties, and toxicity.

More than a century ago, Crum-Brown and Fraser[1] expressed the idea that the physiological action of a substance in a certain biological system (Φ) was a function (f) of its chemical constitution C:

$$\Phi = f\,C \qquad \text{Equation [1]}$$

Thus, an alteration in chemical constitution, ΔC, would be reflected by an alteration in biological activity ΔΦ.

At the turn of the 20th century, Meyer and Overton[2, 3] independently suggested that the narcotic (depressant) action of a group of organic compounds paralleled their olive oil/water partition coefficients.

In following years on the physical organic front, the seminal work of Hammett gave rise to the "σ–ρ" culture[4,5] in the delineation of substituent effects on organic reactions, while Taft devised a way for separating polar, steric, and resonance effects and introducing the first steric parameter, *ES*[6] .

In 1962 Hansch et al. published their study on the structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity[7]. Using the octanol/water system, a whole series of partition coefficients were measured, and thus a new hydrophobic scale was introduced. The parameter π, which is the relative hydrophobicity of a substituent, was defined in a manner analogous to the definition of sigma (18).

$$\pi = \log P_X - \log P_H \qquad \text{Equation [2]}$$

P_X and P_H represent the partition coefficients of a derivative and the parent molecule, respectively.

The contributions of Hammett and Taft together laid the basis for the development of the QSAR paradigm by Hansch and Fujita, which combined the hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms.

There is a consensus among current predictive toxicologists that Corwin Hansch is the founder of modern QSAR. In the classic article[8] it was illustrated that, in general, biological activity for a group of 'congeneric' chemicals can be described by a comprehensive model:

$$\text{Log } 1/C_{50} = a\,\pi + b\,\varepsilon + cS + d \qquad \text{Equation [3]}$$

in which C, the toxicant concentration at which an endpoint is manifested (e.g. 50% mortality or effect), is related to a hydrophobicity term, p, (this is a substituent constant denoting the difference in hydrophobicity between a parent compound and a substituted analog, it has been replaced with the more general molecular term the log of the 1-octanol/water partition coefficient, log Kow), an electronic term, 1, (originally the Hammett substituent constant, s) and a steric term, S, (typically Taft's substituent constant, ES). Due to the curvilinear, or bilinear, relationship between $\log 1/C_{50}$ and hydrophobicity normally found in single dose tests the quadratic π^2 term was later introduced to the model.

The rationale for Eq. (3) was given by McFarland[9]. He hypothesized that the relative activity of a biological active molecule, such as a toxicant, is dependent on: (1) the probability (Pr1) that the toxicant reaches its site of action, (2) the probability (Pr2) that the toxicant will interact with the target at this site, and (3) the external concentration or dose.

The delineation of these models led to explosive development in QSAR analysis and related approaches[10].

Besides the Hansch approach, other methodologies were also developed to tackle structure-activity questions. The Free-Wilson approach addresses structure-activity studies in a congeneric series as described in Equation 4:

$$BA = \Sigma\, a_i\, x_i + u \qquad \text{Equation [4]}$$

Where BA is the biological activity, u is the average contribution of the parent molecule, and a_i is the contribution of each structural feature; x_i denotes the presence ($x_i = 1$) or absence $x_i = 0$ of a particular structural fragment. structure.

In the years after the 1960s, the need to solve new problems, together with the contributions of many other investigators, generated thousands of variations of the Hansch approach to QSAR modelling, as well as approaches that are formally completely new.

Hans Konemann[11] and Gilman Veith[12] who in the early 1980s developed multi-class-based, hydrophobic- dependent models for industrial organic chemicals, must share credit for the revival of QSAR.

At present, the QSAR science, founded on the systematic use of mathematical models and on the multivariate point of view, is one of the basic tools of modern drug and pesticide design and has an increasing role in environmental sciences.

QSAR models exist at the intersection of chemistry, statistics and biology, in toxicological studies. The development of a QSAR model requires these three components: 1) a data set that provides experimental measures of a biological activity for a group of chemicals; 2) molecular structure and/or property data (i.e. the descriptors, variables, or predictors) for this group of chemicals; and 3) statistical methods, to find the relationship between these two data sets.

The limiting factor in the development of QSARs is the availability of high quality experimental data. In QSAR analysis, it is imperative that the input data be both accurate and precise to develop a

meaningful model. In fact, it must be realized that any resulting QSAR model that is developed is only as valid statistically as the data that led to its development.

Data used in QSAR evaluations are obtained either from the literature or generated specifically for QSAR-type analyses. These data can consist of congeneric series of chemicals or assure structural diversity even within a chemical class. This diversity has allowed the generalization of more robust QSARs, applicable in an extended way. A structure– activity model is defined and limited by the nature and quality of the data used in model development and should be applied only within the model's applicability domain.

The ideal QSAR should: (1) consider an adequate number of molecules for sufficient statistical representation, (2) have a wide range of quantified end-point potency (i.e. several orders of magnitude) for regression models or adequate distribution of molecules in each class (i.e. active and inactive) for classification models, (3) be applicable for reliable predictions of new chemicals (validation and applicability domain) and (4) allow to obtain mechanistic information on the modelled end-point. Chemical descriptor(s) include empirical, quantum chemical, or non-empirical parameters. Empirical descriptors may be measured or estimated and include physico-chemical properties (such as for instance logP). Non-empirical descriptors can be based on individual atoms, substituents, or the whole molecule, they are typically structural features. They can be based on topology or graph theory and, as such, they are developed from the knowledge of 2D structure, or they can be calculated from the 3D structural conformations of a molecule.

A variety of properties have been also used in QSAR modeling, these include physico-chemical, quantum chemical, and binding properties. Examples of molecular properties are electron distribution, spatial disposition (conformation, geometry, and shape), and molecular volume. Physicochemical properties include descriptors for the hydrophobic, electronic, and steric properties of a molecule as well as other properties including solubility and ionization constants. Quantum chemical properties include charge and energy values. Binding properties involve biological macromolecules and are important in receptor-mediated responses.

A big problem related to molecular descriptors is their reproducibility: experimental values can differ greatly even when referred to the same compound[13]. Several approaches have been developed for the theoretical calculation of logP[14-18], but also in these calculations it is not uncommon to have differences of several orders of magnitude[19].

In modern QSAR approaches, it is becoming quite common to use a wide set of theoretical molecular descriptors of different kinds, able to capture all the structural aspects of a chemical to

translate the molecular structure into numbers. Different descriptors are different ways or perspectives to view a molecule, taking into account the various features of its chemical structure, not only mono-dimensional as the simple counts of atoms and groups, but also bi-dimensional from the topological graph or three-dimensional from a minimum energy conformation. A lot of software calculates wide sets of different theoretical descriptors, from SMILES, 2D-graphs to 3D-x,y,z-coordinates. Some of the more used are mentioned here: ADAPT[20,21] OASIS[22], CODESSA[23], MolConnZ [24], and DRAGON [25]. It has been estimated that more than 3000 molecular descriptors are now available, and most of them have been summarized and explained[26-28]. The great advantage of theoretical descriptors is that they can be calculated homogeneously by a defined software for all chemicals, even those not yet synthesized, the only need being a hypothesized chemical structure, thus they are reproducible.

Modeling methods used in the development of QSARs are of two types in relation to the modelled response: a potency of an end-point (a defined value of EC50) or a category/class (for instance Mutagen/Not mutagen).

For the potency modelling, the most widely used mathematical technique is multiple regression analysis (MRA). Regression analysis is a simple approach that leads to a result that is easy to understand and, for this reason, most QSARs are derived using regression analysis.
Regression analysis is a powerful means for establishing a correlation between independent variables (molecular descriptors X) and a dependent variable Y, such as biological activity:

$$Y = b + aX_1 + c\,X_2 + \ldots.. \qquad \text{Equation [5]}$$

For the modelling of categories, different quantitative models of classification can be applied. A wide range of classification methods exists, including: discriminant analysis (DA; linear , quadratic, and regularized DA), SIMCA (Soft Independent Modeling of Class Analogy), k-NN (k-Nearest Neighbours), CART (Classification And Regression Tree), Artificial Neural Network, Support Vector Machine, etc. In these tecniques, the term “quantitative” is referred to the numerical value of the variables (the molecular descriptors) necessary to classify the chemicals in the qualitative classes.

It is evident from the literature analysis that the QSAR world has undergone profound changes since the pioneering work of Corvin Hansch, considered the founder of modern QSAR modeling[7,8,10].

The main change is reflected in the growth of a parallel and quite different conceptual approach to the modeling of the relationships among a chemical's structure and its activity/properties.

In the Hansch approach, still applied widely and followed by many QSAR modelers, for instance Schultz et al.[29] Veith and Mekenyan[30], Benigni[31], molecular structure is represented by only a few molecular descriptors (typically log Kow, Hammett constants, HOMO/LUMO, some steric parameters) selected personally by the modeler and inserted in the QSAR equation to model a studied end-point. Alternatively, in a different approach chemical structure is represented, in the first preliminary step, by a large number of theoretical molecular descriptors which are then, in a second step, selected by different chemometric methods as the best correlated with response and included in the QSAR model (the algorithm). The fundamental aim is the optimization of model performance for prediction.

According to the Hansch approach, descriptor selection is guided by the modeler's conviction to have *a priori* knowledge of the mechanism of the studied activity/property, and the presumption to assign mechanistic meaning to any used molecular descriptor selected by the modeler from among a limited pool of potential modeling variables, normally well known and repeatedly used (for instance: logKow is a universal parameter miming cell membrane permeation, thus it is used in a lot of toxicity models, but it is also related to various partition coefficients such as bioconcentration/bioaccumulation, soil sorption coefficient, etc.; HOMO/LUMO are always selected for modeling chemical reactivity, etc.).

On the other hand, the 'statistical' or chemometric approach, an approach parallel to the previous so-called 'mechanistic' one, is based on the fundamental conviction that the QSAR modeler should not influence, *a priori* and personally, the descriptor selection through mechanistic assumptions, but should apply unbiased mathematical tools to select, from a wide pool of input descriptors, those descriptors most correlated to the studied response. The number and typology of the available input descriptors must be as wide and different as possible in order to guarantee the possibility of representing any aspect of the molecular structure. Different descriptors are different ways or perspectives to view a molecule, however the models must be developed taking into account the principle of parsimony, named the Ockham's Razor : "entities should not be multiplied beyond necessity" or "avoid complexity if not necessary" . This principle is often paraphrased as " The simplest solution is the best."

Thus, descriptor selection must be performed by applying mathematical approaches (such as for instance evolutionary techniques, Genetic Algorithms, etc) with the final and crucial aim to maximize, as an optimization parameter, the predictive power of the QSAR model, as the real utility of any model is considered its predictivity.

Regarding the interpretability of the descriptors, it is important to take into account that modeled response is frequently the result of a series of complex biological or physico-chemical mechanisms, thus it is very difficult and reductionist to ascribe too much importance to the mechanistic meaning of the molecular descriptors used in a QSAR model. Moreover, it must also be highlighted that in multivariate models such as MLR models, even though the interpretation of the singular molecular descriptor can be certainly useful, it is only the combination of the selected set of descriptors that is able to model the studied end-point. If the main aim of QSAR modeling is to fill the gaps in available data, the modeler attention should be focused on model quality. In relation to this point, Livingstone states[32]: "The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the "mechanism" in chemical terms, but it is often not necessary, per se". Zefirov and Palyulin[33] took the same position, differentiating predictive QSARs, where attention essentially concerns the best prediction quality, from descriptive QSARs where major attention is paid to descriptor interpretability.

In fact, the first aim of any modeler should be validation for the predictive application of the QSAR model, for both the mechanistic approach and the statistical one. The famous "Kubinyi Paradox"[34,35], emphasized also by Tropsha et al. in their famous papers: Beware of Q^2 [36] and The Importance of being Earnest[37] is that: The „best fit" models are not the best ones for prediction! In fact, a QSAR model must, first of all, be a real model, robust and predictive, to be considered a reliable model[38] ; only a stable and predictive model can be usefully interpreted for its mechanistic meaning, even so this is not always easy or feasible.

QSAR model validation has been recognized by specific OECD expert groups as a crucial and urgent point in recent years, and this has led to the development, for regulatory purposes, of the "OECD principles for the validation of (Q)SAR models"[39]. The need for this important action was mainly due to the recent new chemicals policy of the European Commission (REACH: Registration, Evaluation and Authorization of Chemicals)[40], that explicitly states the need to use (Q)SAR models to reduce experimental testing (including animal testing). Obviously, to meet the requirements of the REACH legislation it is essential to use (Q)SAR models that produce reliable estimates, i.e., validated (Q)SAR models. Thus, reliable QSAR model must be associated with the following information: 1) a defined endpoint; 2) an unambiguous algorithm; 3) a defined domain of applicability; 4) appropriate measures of goodness-of–fit, robustness and predictivity; 5) a mechanistic interpretation, if possible.

The need for interpretability depends on the application, as a validated mathematical model relating a target property to chemical features may be all that is necessary, particularly when predicted data are needed for screening of large libraries of chemicals, though it is obviously desirable to attempt some explanation of the 'mechanism' in chemical terms[32,33].

REFERENCES

1. A. Crum-Brown, T.R. Fraser, *Trans. R. Soc. Edinburgh* **1868–1869**, *25*, 151.
2. H. Meyer, *Arch. Exp. Pathol. Pharmakol*. **1899**, *42*, 109.
3. C.E. Overton, *Studien Uber die Narkose*, Fischer, Jena, Germany, **1901**.
4. L. P. Hammett, *Chem. Rev*., **1935**, *17*, 125.
5. L. P. Hammett, *Physical Organic Chemistry*, 2nd ed., McGraw-Hill, New York, **1970**.
6. R. W. Taft, *J. Am. Chem. Soc*., **1952**, *74*, 3120.
7. C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, *Nature*, **1962**, *194*, 178.
8. C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616.
9. J.W. McFarland, *J. Med. Chem.* **1970**, *13*, 1092.
10. C. Hansch, A. Leo, *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology,* ACS Professional Reference Book, American Chemical Society, Washington, DC, **1995**.
11. H. Ko¨nemann, *Toxicolog,* **1981**, *19*, 209.
12. G.D. Veith, D.J. Call, L.T. Brooke, *Can. J. Fish Aquat. Sci*. **1983,** *40*, 743.
13. R. Renner, *Environ. Sci. Technol.* **2002**, *36*, 410A.
14. R. Mannhold, A. Petrauskas, *QSAR Comb. Sci.* **2003**, *22*, 466.
15. http://clogP.pomona.edu/medchem/chem/clogP/index.html;
16. http://www.syrres.com/esc/kowwin.htm;
17. http://www.vcclab.org;
18. G. Klopman, J. K. Li, S.Wang, M. Dimayuga, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752.
19. E. Benfenati, G. Gini, N. Piclin, A. Roncaglioni, M. R. Varı`, *Chemosphere*, **2003**, *53*, 1155.
20. A.J. Stuper P.C., Jurs, *J Chem Inf Comput Sci*, **1976,** *16*, 99.
21. http://research.chem.psu.edu/pcjgroup/ADAPT.html
22. O. Mekenyan, D. Bonchev, , *Acta Pharm Jugosl.,* **1986**, *36*, 225.

23. A.R., Katritzky, V.S., Lobanov, CODESSA, Version 5.3, University of Florida, Gainesville, **1994**.
24. MolConnZ, Ver. 4.05, **2003**, Hall Ass. Consult., Quincy, MA
25. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON—Software for the calculation of molecular descriptors. Ver. 5.4 for Windows, **2006**, Talete srl, Milan, Italy.
26. J. Devillers, A.T Balaban, (Eds.) *Topological Indices and Related Descriptors in QSAR and QSPR*, Amsterdam: Gordon Breach Sci. Pub., **1999**.
27. M. Karelson, *Molecular Descriptors in QSAR/QSPR*. New York: Wiley-InterScience, **2000**.
28. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors,* Wiley-VCH, Weinheim (Germany), **2000**.
29. T.W Schultz, M.T.D.Cronin, T.I. Netzeva , A.O. Aptula, , *Chem. Res. Toxicol*., **2002,** *15*, 1602.
30. G.D Veith, O.G. Mekenyan, , *Quant. Struct-Act. Rel.,* **1993**, *12*, 349.
31. R. Benigni, *Chem. Rev*., **2005,** *105*, 1767.
32. D.J. Livingstone, *J. Chem. Inf. Comput. Sci.*, **2000,** *40*, 195..
33. N.S. Zefirov and V.A. Palyulin, *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1022.
34. J. H. van Drie, *Curr. Pharm.Des.* **2003**, *9*, 1649.
35. J. H. van Drie, in: Computational Medicinal Chemistry for Drug Discovery, P. Bultinck et al., Eds., Marcel Dekker, **2004**.
36. A. Golbraikh, A. Tropsha, *J. Mol.Graph Mod.* **2002**, *20*, 269.
37. A. Tropsha, P. Gramatica, V.J. Gombar, *QSAR Comb. Sci.,* **2003,** *22*, 69.
38. P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb.Sci.* **2007,** 26(5), 694-701.
39. http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html
40. http://europa.eu.int/comm/environment/chemicals/reach.htm