Computational Chemistry Approaches for Understanding how Structure Determines Properties*

Alan R. Katritzky[a], Svetoslav Slavov[a], Maksim Radzvilovits[a,c], Iva Stoyanova-Slavova[a], and Mati Karelson[b,c]

[a] Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611, USA
[b] Institute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia
[c] Molcode, Ltd., Turu 2, Tartu 51014, Estonia

Reprint requests to A. R. Katritzky. E-mail: katritzky@chem.ufl.edu

Z. Naturforsch. **2009**, *64b,* 773 – 777; received April 18, 2009

Dedicated to Professor Gerhard Maas on the occasion of his 60th birthday

The establishment of quantitative relationships between numerous molecular properties and chemical structures is now of great importance to society in understanding and improving environmental, medicinal and technological aspects of life. Quantitative structure-activity (property) relationships (QSA(P)R) relate physical, chemical, physico-chemical, technological and biological properties of compounds to their structure. A major factor driving the widespread use of QSP(A)R models is the rational estimation of properties of new compounds, without first synthesizing and testing them. Some of our recent findings in the field are briefly discussed below.

Key words: QSPR, QSAR, CODESSA Pro, Computational Chemistry, Structural Predictions

Overview

All properties of a compound – physical, chemical, biological, and technological – depend on the way its atoms, (the "building blocks"), are connected to form the individual molecule. Theory provides insight on how the molecular structure (composition) determines the behavior of substances: *e. g.* hydrocarbon molecules containing from one to four carbon atoms are gases at r. t., but as more carbons are added, they exist as liquids (starting at C_6H_{14}) and finally as solids (starting at $C_{18}H_{38}$). With the advance of computational techniques, it is now possible to calculate a wide range of physicochemical characteristics: ionization energies, polarizabilities, heats of formation, *etc*. However, it should be noted that in most cases such calculations relate to isolated individual molecules rather than to bulk matter, which corresponds to real experimental situations.

The vastly increased computational power of modern computers has enabled the application of new powerful alternative methods to model and understand more complex physicochemical, chemical, technological, and biochemical properties. In particular, this applies to the quantitative structure-activity (property) relationship approach, abbreviated QSA(P)R. The methodology used to generate QSPR for predicting a physical property (such as solubility in water) or QSAR for biochemical property (such as insect repellency) is similar.

Progress in QSA(P)R methodology has led to the development of various software products aimed at the automation of the modeling procedures. Among the packages currently available are: SYBYL [1], CODESSA (later CODESSA PRO [2]), DRAGON [3], AUTODOCK [4], OPENEYE [5], TSAR [6], and others. For more than 15 years now, our group has actively contributed to this field by developing, supporting and applying the methodology encoded in CODESSA and CODESSA PRO software. Some recent successful applications are discussed below.

Validation of QSA(P)R Methodology

A recent criticism [7] stating that a "chance" correlation was involved in our earlier work was rebutted [8]. Using sets of "natural" and generated "ran-

* Based on a lecture to be presented by A. R. Katritzky at the 8th IMSAT Meeting, September 2009.

dom" descriptor values, applying comprehensive statistical and comparative techniques, we demonstrated that large descriptor pools could be used legitimately and advantageously for QSPR/QSAR modeling. The ability of our Best Multi-Linear Regression algorithm to produce robust correlations was discussed in detail.

Physical and Physico-chemical Properties

UV spectral absorbance

High performance liquid chromatography (HPLC) combined with ultraviolet (UV) spectrophotometric detection is the method most applied in organic chemistry for analyzing reaction products. UV is also considered a nearly universal detector for drug-like molecules: 85 % of the structures in the MDDR (a database of drugs and candidate drugs [9]) contain an aromatic group and most of the remaining 15 % contain an alternative chromophore. The NIST Chemistry WebBook database was used to extract the UV absorption intensities for a diverse set of 805 organic compounds at 260 nm and 25 °C in water. CODESSA PRO descriptors were utilized to generate a five-parameter multilinear model with $R^2 = 0.692$ [10]. Concurrently, a neural networks approach was used to develop a corresponding nonlinear model. The UV absorption intensity is mostly determined by the overlap of the excited and ground state wavefunctions of the molecule. Most of the descriptors were related to the symmetry of the molecule, the degree of unsaturation and the HOMO-LUMO energy gap. Since mixtures of compounds identified by HPLC UV method usually have $\Delta\varepsilon$ values bigger than $1-2$ log units, we believe that even this imperfect correlation ($R^2 = 0.692$) could be useful for identification purposes.

Critical micelle concentrations

Surfactants are amphiphilic molecules that contain a nonpolar segment, named "tail", and a polar segment, called "head". When the surfactant concentration is low, the molecules exist as individual entities, but when the concentration increases the presence of two very different substructural features (tail and head) causes aggregation. The simplest of such aggregates, having approximately spherical shape, are called micelles.

The transition from premicellar to micellar solutions occurs at a concentration called the "critical micelle concentration" (CMC). Many important properties of a surfactant solution undergo sharp change at the CMC including surface tension, interfacial tension, conductivity, osmotic pressure, detergency, emulsification, and foaming. The CMC is therefore a very useful parameter for characterization of surfactants and can be correlated with many industrially important properties.

Non-ionic surfactants

Employing the general QSPR approach encoded in CODESSA, our group [11] proposed a three-parameter logCMC model for a set of 77 non-ionic surfactants ($R^2 = 0.983$, $F = 1433$, $s^2 = 0.0313$) using topological descriptors calculated for the hydrophobic "tail" and constitutional descriptors for the hydrophilic "head". Two of the three descriptors represent contributions from the size and structural complexity of the hydrophobic group; the third is related to the size of the hydrophilic group.

Later [12] we updated and extended this model by applying linear and nonlinear modeling techniques to a larger dataset of 162 nonionic surfactants. The descriptors in the derived models were again related to the molecular shape and size and to the presence of heteroatoms participating in donor-acceptor and dipole-dipole interactions. The steric hindrance in the hydrophobic domain was also identified as important for the micellization phenomena.

Anionic surfactants

In our early studies we reported a three-parameter QSPR correlation ($R^2 = 0.940$, $F = 597$, $s^2 = 0.0472$) for the logCMC values of 119 anionic surfactants (sulfates and sulfonates) [13]. The Kier and Hall index (0^{th} order) calculated for the hydrophobic tails, the relative number of the carbon atoms in the head, and the total dipole of the molecule were found to control the CMC.

We again updated and extended this work in a recently conducted research [14]. A larger and more diverse dataset of 181 diverse anionic surfactants was used to relate the logarithm of CMC to the molecular structure using CODESSA PRO software. A fragment approach produced a five-parameter linear QSPR model of superior statistical characteristics and predictive ability ($R^2 = 0.897$). The regression equation allowed insight into the structural features that influence the CMC of surfactants. The contributions from the hydrophobic fragments were expressed by topological and geometrical descriptors, while the hydrophilic

fragment was represented by constitutional, geometrical, and charge-related descriptors. Topological, solvational, and charge-related descriptors were significant in the preferred QSPR models representing the driving force of the intermolecular interactions between anionic surfactants and water.

Cationic surfactants

We also examined cationic surfactants, utilizing a dataset of 50 ammonium and quaternary pyridinium derivatives [15]. Multilinear models were developed for both the first CMC at which spherical micelles ($R^2 = 0.977$) were formed and the second ($R^2 = 0.965$) CMC formation of larger aggregates. A general ANN model for the first CMC with $R^2 = 0.974$ was also proposed. Most of the descriptors in these models were related to the size and charge distribution of the hydrophobic tail and to the size of the head. The multilinear model for the second CMC was more closely related to the hydrophobic domain of the surfactant than that of the first CMC.

Flash points

Recent work by our group [16] reported MLR and ANN QSPR models for the flash points of a data set of 758 diverse organics. The best four-descriptor linear model was characterized by $R^2 = 0.849$ and an average error of 13.9 K. Descriptors appearing in the model mostly reflect the electrostatic and hydrogen bonding interactions in the bulk compound as well as the elaborate molecular shape. The ANN modelling produced slightly better statistical parameters: $R^2 = 0.878$ and an average error of 12.6 K. This work extended studies previously conducted by Zefirov and coworkers [17].

Universal solvation equation

Understanding factors determining solubility in water is important both for its own sake, and because solvation interactions play a crucial role in the rational modeling of various physicochemical processes. The construction of reliable theoretical models for the quantitative estimation of partition processes has been of particular interest. We participated with the groups of Oliferenko and Zefirov in the development [18] of a Universal Solvation Equation (USE) using a dataset consisting of 525 diverse small organic molecules including mono-, di-, and polyfunctional aliphatic and aromatic species, heterocycles, amino acids, nucleotides, and pharmaceuticals. This incorporates descriptors which were fashioned to represent (i) the hydrogen bond acidity A, (ii) the effective atomic basicity B, (iii) the total molecular polarizability α, (iv) the polarity P, (v) the hydrophobicity H, (vi) the steric correction S, and (vii) the total energy of the π-system E_π:

$$\begin{aligned} \mathrm{Log}\,SP = \mathrm{Const} + \beta_1 A + \beta_2 B + \beta_3 \alpha + \beta_4 P \\ + \beta_5 H + \beta_6 S + \beta_7 E_\pi \end{aligned} \quad (1)$$

Coefficients β_n of Eq. 1 are found by the method of multiple linear regression, and SP means some solvation-related property. Applications of USE were demonstrated to estimate diverse properties, including gas/water and octanol/water partition coefficients, aqueous solubilities, and solvation of ionic species.

The reliability of USE was proven by Oliferenko who applied it to the "challenge" [19] posed by the Journal of Chemical Information and Modeling (Issue 48 of 2008) to the modeling community. Among 99 models entered for this competition to predict the aqueous solubilities of 32 drugs, based on a training set of 100 drug and drug-like molecules, USE was ranked as second best in terms of predictive power. The average prediction of S from the 99 attempts was $R^2 = 0.158$ (standard deviation 0.184). The best prediction for S (mg/mL) had $R^2 = 0.642$ while USE predicted $R^2 = 0.631$.

Biological Properties

Attractants and repellents

The quest to make humans less attractive to mosquitoes has fueled decades of scientific research on mosquito behavior and control. In the United States, most mosquito bites are merely a nuisance. Worldwide, however, mosquitoes transmit disease to more than 700 million people annually, and statistics suggest that they will be responsible for 6 % of the deaths of the current world population.

The factors involved in attracting mosquitoes to the host are complex and not yet fully understood. Mosquitoes use visual, thermal, and olfactory stimuli to locate the host. Of these, olfactory cues are probably most important. It was found that lactic acid – one of the natural odorants produced by the human body – attracts mosquitoes by activating the chemoreceptors on their antennae. These same receptors may be inhibited by synthetic (DEET) or plant derived (Citronella) chemicals called repellents.

The repellents currently in use may cause skin irritation and stinging sensation when in contact with eyelids or lips. Thus, there is a long standing interest in the design of chemicals that would be effective repellents against a wide range of insects, with as few as possible adverse effects.

Our continuing interest to develop more efficient mosquito repellents in a long-term collaboration with the USDA (US Department of Agriculture) led to the development of several QSAR models describing repellency in terms of molecular descriptors.

A successful QSAR model for a dataset of 31 amides with $R^2 = 0.80$, $F = 26$, and $s^2 = 0.47$ was reported [20]. The descriptors involved were related to the duration of repellent action, the geometrical complimentarity of the molecule to the binding site and to the formation of hydrogen bonds between the receptor (possibly AgOr7) and the ligand.

Further studies on a dataset of 200 *N*-acylpiperidines resulted in the development of a predictive ANN model [21], which was able to estimate the efficiency of the most active 55 repellents in the dataset with an accuracy exceeding 70 %. The input neurons reflected the geometrical and topological features of the *N*-acylpiperidines, and the distribution of charge in these molecules which is mainly related to the ability of non-covalent bond (*e. g.* hydrogen bonding) formation important in ligand-receptor interactions. Using the above model we proposed a set of 34 acylpiperidine derivatives, which were synthesized and provided for testing at the USDA facilities. At the lower concentration screened (2.5 μmol/cm^2) all but 3 compounds were found to be more active than DEET, with some having even five times higher activity.

Antifungal activity

Since the 1980s, complications from fungal infections have been recognized as a major cause of morbidity and mortality in immunocompromized patients. Thus, the development of new and effective antifungal agents is highly sought. Our team investigated [22] antifungal activity against the dimorphic fungus *Candida albicans* of a series of 83 cyanoboranes, fluconazoles, carbonylaminobenzoxazoles and imidazolylmethylindoles, obtaining a multilinear QSAR model with $R^2 = 0.788$, $F = 47.144$, $s^2 = 0.130$, $R^2_{cv} = 0.749$. The six descriptors which entered the final equation were: "relative number of C atoms", "hydrogen donors charged surface area", "average valency of an H atom", "RNCG relative negative charge", $(LogP)^2$ and "average electrophilic reactivity index for atom C". These are related to the transport properties and the binding affinity of the compounds.

Anti-invasive activity

A major challenge in current cancer research is the development of anti-invasive and anti-metastasis drugs. The anti-invasive activity index (I index) measures the anti-tumor cell activity at given concentration (μM). A set of 139 structures [23] was split into 4 categories (low, fair, good, and active) each assigned a discrete value (1, 2, 3, and 4, respectively) according to the anti-invasive activity level of the compounds. The BMLR method implemented in CODESSA PRO was used to pre-select a set of relevant descriptors which were further utilized as inputs for the construction of a nonlinear artificial neural network (ANN) model. The descriptors employed in the model relate to the essential electrostatic, conformational interactions and hydrogen acceptor/donor abilities of a compound in the biological system. The resulting ANN QSAR model predicted the class precisely for 66 (71 %) of the training set of 93 compounds and 32 (70 %) of the validation set of 46 compounds.

Drug transfer into human breast milk

Many women need to take various types of medications while breast feeding. The bio-accumulation of a specific medication in milk is associated with a risk to the infant that can exceed the benefits of breast feeding. Because of the significant role of breast milk, the investigation of transfer of contamination from a mother's medication into her breast milk is important. The milk to plasma concentration ratio (M/P ratio) of a drug is an attempt to quantify the equilibrium concentration between breast milk and blood and is generally used to estimate the infant's exposure to drugs through breast milk. A set of experimentally measured M/P ratio values was collected from the literature for 115 widely used pharmaceuticals of different chemical nature. Based on the dataset, a satisfactory ($R^2 = 0.791$) seven-parameter QSAR model was derived [24]. The descriptors appearing in the model were primarily related to the electrostatic and hydrogen bonding interactions between the drug molecule and the surrounding media.

Conclusions

In conclusion, we believe that the QSA(P)R techniques will continue to expand with many applications

and will help us understand better how chemical structure determines properties.

[1] http://www.tripos.com.
[2] http://www.codessa-pro.com.
[3] http://talete.mi.it.
[4] http://autodock.scripps.edu.
[5] http://www.eyesopen.com.
[6] http://accelrys.com.
[7] A. R. Katritzky, D. Fara, M. Karelson, *Bioorg. Med. Chem*. **2004**, *12*, 3027 – 3035.
[8] A. R. Katritzky, D. A. Dobchev, S. Slavov, M. Karelson, *J. Chem. Inf. Model*. **2008**, *48*, 2207 – 2213.
[9] http://www.akosgmbh.de/Symyx/software/databases/mddr.htm.
[10] A. R. Katritzky, S. Slavov, D. Dobchev, M. Karelson, *J. Comput.-Aided Mol. Des*. **2007**, *21*, 371 – 377.
[11] P. D. T. Huibers, V. S. Lobanov, A. R. Katritzky, O. D. Shah, M. Karelson, *Langmuir* **1996**, *12*, 1462 – 1470.
[12] A. R. Katritzky, L. Pacureanu, S. Slavov, D. Dobchev, M. Karelson, *Ind. Eng. Chem. Res*. **2008**, *47*, 9687 – 9695.
[13] P. D. T. Huibers, V. S. Lobanov, A. R. Katritzky, O. D. Shah, M. Karelson, *J. Colloid Int. Sci*. **1997**, *187*, 113 – 120.
[14] A. R. Katritzky, L. Pacureanu, D. Dobchev, M. Karelson, *J. Chem. Inf. Model*. **2007**, *47*, 782 – 793.
[15] A. R. Katritzky, L. Pacureanu, S. Slavov, D. Dobchev, D. Shah, M. Karelson, *Comput. Chem. Eng*. **2009**, *33*, 321 – 332.
[16] A. R. Katritzky, I. Stoyanova-Slavova, D. Dobchev, M. Karelson, *J. Mol. Graph. Model*. **2007**, *26*, 529 – 536.
[17] N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Russ. Chem. Bull. Int. Ed*. **2003**, *52*, 1885 – 1892.
[18] P. Oliferenko, A. Oliferenko, G. Poda, V. Palyulin, N. Zefirov, A. R. Katritzky, *J. Chem. Inf. Model*. **2009**, in press.
[19] http://pubs.acs.org/userimages/ContentEditor/1226959448617/jcim-solubility-findings.pdf.
[20] A. R. Katritzky, D. A. Dobchev, I. Tulp, M. Karelson, D. A. Carlson, *Bioorg. Med. Chem. Lett*. **2006**, *16*, 2306 – 2311.
[21] A. R. Katritzky, Z. Wang, S. Slavov, M. Tsikolia, D. Dobchev, N. G. Akhmedov, C. D. Hall, U. R. Bernier, G. G. Clark, K. J. Linthicum, *Proc. Natl. Acad. Sci*. **2008**, *105*, 7359 – 7364.
[22] A. R. Katritzky, S. Slavov, D. Dobchev, M. Karelson, *Bioorg. Med. Chem*. **2008**, *16*, 7055 – 7069.
[23] A. R. Katritzky, M. Kuanar, D. Dobchev, B. Vanhoecke, M. Karelson, V. Parmar, C. Stevens, M. Bracke, *Bioorg. Med. Chem*. **2006**, *14*, 6933 – 6939.
[24] A. R. Katritzky, D. Dobchev, E. Hür, D. Fara, M. Karelson, *Bioorg. Med. Chem*. **2005**, *13*, 1623 – 1632.