SciForum
Mol2Net

Hierarchical Quantitative Structure-Activity Relationships (HiQSARs) for the Prediction of Physicochemical and Toxicological Properties of Chemicals Using Computed Molecular Descriptors

Subhash C. Basak[1*] and Subhabrata Majumdar[2]

[1] University of Minnesota Duluth-Natural Resources Research Institute (UMD-NRRI) and Department of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA; sbasak@nrri.umn.edu

[2] School of Statistics University of Minnesota Twin Cities Minneapolis, MN 55414

* Author to whom correspondence should be addressed; E-Mail: sbasak@nrri.umn.edu; Tel.: +1-218-727-1335

Published: 4 December 2015

Abstract: Attempts have been made to formulate quantitative structure=activity relationships (QSARs) for the prediction of property/ bioactivity of chemicals from their experimental test data as well as properties that can be computed directly from molecular structure without the input of any other experimental property. Because both in drug design and hazard assessment of chemical scenarios relevant experimental data for property/ bioactivity estimation are not available for the majority of candidate chemicals, QSARs based on computed molecular descriptors are emerging as methods of choice for property/ bioactivity estimation in many cases. Numerical graph invariants or topological indices, viz., topostructural (TS) indices, topochemical (TC) indices, as well as three-dimensional (3-D) descriptors, and quantum chemical (QC) indices have been used for QSAR formulation based on computed descriptors. In the 1990s, Basak et al formulated the concept of hierarchical quantitative structure=activity relationships (HiQSAR) in which TS, TC, 3-D, and QC descriptors were used in a graduated manner, the more computationally demanding descriptors being used only if the simpler ones did not give acceptable QSAR models. Our experience with a substantial number of HiQSARs for physical, pharmacological, and toxicological properties of different congeneric as well diverse sets chemicals indicate that the combinations of TS + TC descriptors are capable of giving good quality QSARs in most situations. The addition of 3-D or QC descriptors make marginal or no improvement in model quality after the use of TS+ TC descriptors. At this age of "big data screening and analysis" this is a good news because QSARs derived from the less expensive and practically useful TS+ TC combination can be effective tools in the screening of large chemical libraries.

Keywords: keyword; keyword; keyword (3–10 keywords separated by semi colons)
Quantitative structure-activity relationship (QSAR); Hierarchical quantitative structure-activity relationship (HiQSAR); Topostructural (TS) indices; Topochemical (TC) indices; Three-dimensional (3-D) descriptors; Quantum chemical (QC) indices; Vapor pressure; Mutagenicity; Ridge Regression; Interrelated two-way clustering (ITC).

1. Introduction

A contemporary trend in quantitative structure-activity/ property relationship (QSAR/ QSPR) studies is the use of properties which can be computed from structure without the input of any other data [1-7]. The underlying reason for this is that for the majority of candidate chemicals that need to be screened for both new drug discovery and hazard assessment of environmental pollutants, experimental properties needed for QSAR formulation are not available [4-7]. Table 1 gives a partial list of physical and bi9ochemical/ toxicological properties needed for the prediction of bioactivity/ toxicity of chemicals. In the realm of hazard assessment of industrial chemicals currently listed in the Toxic Substances Control Act (TSCA) Inventory of the United States Environmental Protection Agency (USEPA), Auer et al [8] reported that for most of the chemicals under investigation the majority of the properties needed for hazard estimation were not available. Over the years, after the publication of this summary by Auer et al [8] in 1990, the availability of good quality experimental data needed for the risk assessment of chemicals has probably became worse with time. Therefore, quantitative structure-activity/ property relationships (QSAR/ QSPR) remain one important source of property data itemized in Table 1. Because property-property relationships (PPRs) are not practical in many situations arising out of the paucity of the predictor property, QSARs derived from computed molecular descriptors have emerged as useful tools in the screening of chemicals.

Over the years Basak and coworkers have used different combinations of topostructural (TS) indices, topochemical (TC) indices, 3-D descriptors as well as and quantum chemical (QC) indices for QSAR formulation in a hierarchical manner (Figure 1). In the hierarchical QSAR (HiQSAR) approach [4-7], TS, TC, geometrical, and quantum chemical descriptors are used for model building in a graduated manner, the latter and more complex levels being used when the earlier ones fail to give reasonable QSAR. Basak et al [4-7] divided the topological indices (TIs) into two major groups: Topostructural (TS) indices and topochemical (TC) indices. TS descriptors are indices which are calculated from skeletal graph models of molecules that do not distinguish among different types of atoms in a molecule or the various types of chemical bonds, e.g.; single bond, double bond, triplet bond, etc. Thus, TS descriptors quantify information regarding the connectivity, adjacency, and distances between vertices of molecular graphs, ignoring their distinct chemical nature. TC indices, on the other hand, are sensitive to both the pattern of connectedness of the vertices (atoms), as well as their chemical/bonding characteristics. Therefore, the TC indices are more complex than the TS descriptors. Figure 1 represents the full hierarchical scheme of QSAR formulation involving different levels of chemodescriptors and biodescriptors, the latter being derived from the omics data. Basak et al used various combination of TS, TC, 3-D, and QC indices for the development of QSAR over the years. For a review please see recent references [1, 4-7]. The indices used by Basak and coworkers have been calculated by the software POLLY [9], MolConnZ [10], APProbe [11], Triplet [12, 13], MOPAC [14], and Gaussian [15].

In this paper we discuss our HiQSAR approach for two sets of chemicals: Vapor pressure of a set 476 diverse molecules and Ames' mutagenicity of a heterogeneous group of 508 chemicals.

.

2. Results and Discussion

.

2.1 QSAR for vapor pressure) for a diverse set of 476 chemicals.

The HiQSAR results provided in Table 2 show that of all classes of molecular descriptors the TC class of indices gave the most effective models. The TS+TC combination makes some improvement in model quality over the TC only QSAR. The model developed using all indices which consisted of (TS+ TC+ 3-D) combination plus dipole moment calculated by Sybyl [18] as well as a hydrogen bonding descriptor HB_1 [19, 20] could not outperform the model derived from the TS+TC combination. For details for this analysis, see [17].

2.2 HiQSAR modeling of a diverse set of 508 chemical mutagens

TS, TC, 3D, and QC descriptors for 508 chemical were calculat4ed and QSARs were formulated hierarchically using the four types of descriptors. For details of calculations and model building, see ref. [7]. The method Interrelated two way clustering, ITC [21], was used for variable selection. Table 3 gives results of ridge regression (RR) alone as well as those where RR was used on descriptors selected by ITC. For both RR only and ITC+ RR analysis the TS + TC combination gave the best models for predicting mutagenicity of the 508 diverse chemicals. The addition of 3-D and QC descriptors to the set of independent variables made minimum or no improvement in model quality.

.

.

Table 1. Important properties needed for evaluation of chemicals

Physicochemical	Pharmacological / Toxicological
Molar volume	Macromolecule level
Boiling point	: Receptor binding
Melting point	: Michaels constant (Km)
Vapor pressure	: Inhibition constant (Ki)
Water solubility	: DNA alkylation
Dissociation constant (pK_a)	: Unscheduled DNA synthesis
Partition coefficient	Cell level
: Octanol-water (log P)	: Salmonella mutagenicity
: Air-water	: Mammalian cell transformation
: Sediment-water	Organism level (acute)
Reactivity (electrophile)	: Algae
	: Invertebrates
	: Fish
	: Birds
	: Mammals
	Organism level (chronic)
	: Bioconcentraton factor
	: Biodegradation
	: Carcinogenicity
	: Reproductive toxicity
	: Delayed neurotoxicity

Table 2: Summary of the Regression Results for the Training Set and the Prediction Results for the Test Set for the Hierarchical Analysis of log VP

	Training Set (342)			Test Set (134)		
Parameter Class	F	R^2	S	F	R^2	S
Topostructural (TS)	104.6	48.1	0.56		57.9	0.46
Topochemical (TC)	126.3	79.2	0.36		85.8	0.27
Geometrical	168.9	51.8	0.53		62.2	0.44
TS+ TC	112.5	80.4	0.35		84.7	0.28
All Indices	117.4	79.6	0.35		84.2	0.28

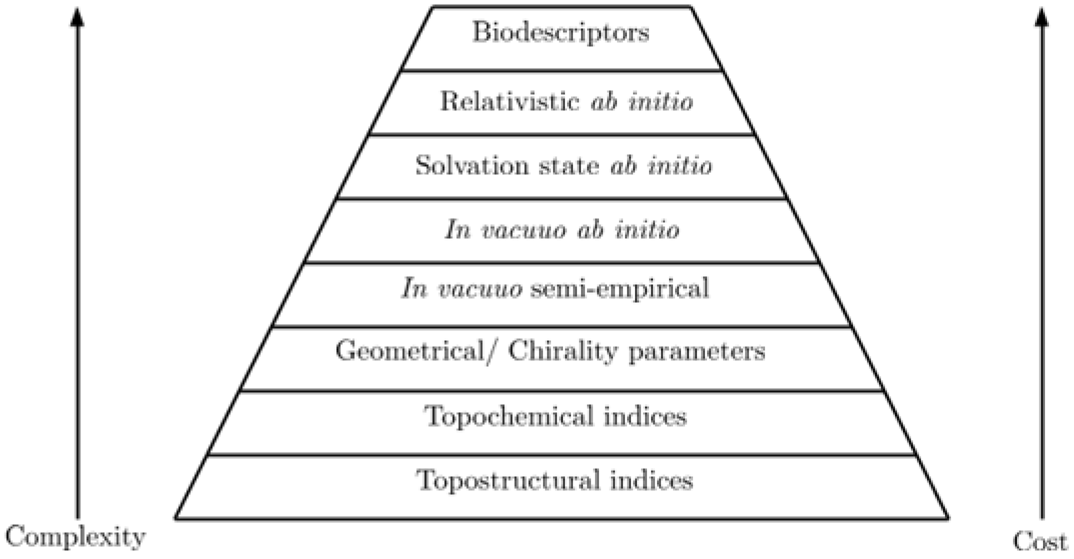
Table 3. HiQSAR model (RR and ITC+RR) for a diverse set of 508 chemical mutagens

Model type	Predictor Type	Predictor Number	% Correct classification	Sensitivity	Specificity
RR	TS	103	53.14	52.34	53.97
	TS+TC	298	76.97	83.98	69.84
	TS+TC+3D+QC	307	77.17	84.38	69.84
ITC+ RR	TS	103	66.34	73.83	58.73
	TS+TC	298	73.23	77.34	69.05
	TS+TC+3D	301	74.80	77.34	72.22
	TS+TC+3D+QC	307	72.05	76.17	67.86

Table 4. Major chemical classes (not mutually exclusive) within the 508 mutagen/non-mutagen database.

Chemical class	Number of compounds
Aliphatic alkanes, alkenes, alkynes	124
Monocyclic compounds	260
Monocyclic carbocycles	186
Monocyclic heterocycles	74
Polycyclic compounds	192
Polycyclic carbocycles	119
Polycyclic heterocycles	73
Nitro compounds	47
Nitroso compounds	30
Alkyl halides	55
Alcohols, thiols	93
Ethers, sulfides	38
Ketones, ketenes, imines, quinones	39
Carboxylic acids, peroxy acids	34
Esters, lactones	34
Amides, imides, lactams	36
Carbamates, ureas, thioureas, guanidines	41
Amines, hydroxylamines	143
Hydrazines, hydrazides, hydrazones, traizines	55
Oxygenated sulfur and phosphorus	53
Epoxides, peroxides, aziridines	25

Figure 1. Hierarchical QSAR development scheme involving different levels of chemodescriptors and biodescriptors, the latter being derived from the omics sciences



3. Materials and Methods

3.1 QSAR for vapor pressure) for a diverse set of 476 chemicals.

Measured vapor pressure (VP) values for 476 subset of the Toxic Substances Control Act (TSCA) Inventory were obtained from the ASTER (Assessment Tools for the Evaluation of Risk) database [16]. Due to the size of the dataset being used in this study, the VP data for these chemicals will not be listed in this paper. The set of 92 TIs was partitioned into 38 topostructural indices and 54 topochemical indices. For details of this study see [17]. Because the number of data points was reasonably large, the data was split into a training set (342 compounds) and a test set (134 compounds), an approximately 75/25 split. Models were developed using the training set of chemicals and then used to predict the VP values of the test chemicals, the results being shown in Table 2.

3.2 HiQSAR modeling of a diverse set of 508 chemical mutagens

The data were taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [22]. The response variable is Ames mutagenicity, the sample available being 508 compounds classified as not mutagenic (scored 0) or mutagenic (scored 1). The set of 508 is comprised of 256 mutagens and 252 non-mutagens. Table 4 gives an idea regarding the diversity of the chemicals in this database in terms of chemical types and functional groups. Ridge regression was used for model building because it is a sound method, in the rank deficient case in particular. For the ITC+RR modeling, ITC was first used for variable selection and then RR was employed for model building.

.

.

4. Conclusions

The objective of HiQSAR research reported in this paper was to study the relative effectiveness of topological (TS, TC), geometrical, and quantum chemical descriptors in the development of useful QSAR models. Results derived for two large data sets, viz. vapor pressure of a group of 476 diverse chemicals and a structurally diverse set of 508 mutagens, show that the computationally less expensive TS and TC descriptors give QSARs of reasonable quality. The addition of 3-D or QC descriptors after the use of TS+TC combination does not make any improvement in model quality. We previously observed this trend in different properties of other data sets [23-30]. At this age of "big data screening and analysis" [31], this is a good news because QSARs derived from the less expensive TS+ TC combination can be effective tools in the fast and effective screening of large chemical libraries. Further QSAR research is in progress to validate the broad applicability of the HiQSAR paradigm based on mathematical structural descriptors [32].

Acknowledgments

One of the authors (Basak) would like to thank Gregory D. Grunwald, Brian D. Gute, and Denise Mills for sustained technical support and collaboration in his research.

.

Author Contributions

Subhash C. Basak formulated the HiQSAR concept in the 1990s and applied it to QSAR of various congeneric and diverse sets of chemicals. Subhabrata Majumdar has been carrying out collaborative research with Basak in QSAR during the past five years.

Conflicts of Interest

"The authors declare no conflict of interest".

References and Notes

[1] Basak, S. C., Role of Mathematical Chemodescriptors and Proteomics-Based Biodescriptors in Drug Discovery, Drug Develop. Res., 2010, 72, 1-9.

[2] Kier, L.B.; Hall, L. Molecular Structure Description: The Electrotopological State; Academic Press: San Diego, CA, 1999.

[3] Devillers, J.; Balaban, A.T., Eds. Topological Indices and Related Descriptors in QSAR and QSPR; Gordon and Breach: Amsterdam, 1999.

[4] Basak, S. C., MATHEMATICAL STRUCTURAL DESCRIPTORS OF MOLECULES AND BIOMOLECULES: BACKGROUND AND APPLICATIONS, in Advances in Mathematical Chemistry and Applications, volume 1, pp. 3-23, Basak, S. C., Restrepo, G. and Villaveces, J. L., Editors, Bentham eBooks, Bentham Science Publishers, 2015.

[5] Basak, S. C., Gute, B. D., Grunwald, G. D., Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, , in Quantitative Structure-activity Relationships in Environmental Sciences VII, F. Chen and G. Schuurmann, Eds., SETAC Press, Pensacola, FL., pp. 245–261 (1998).

[6] Basak, S. C., Mathematical Descriptors for the Prediction of Property, Bioactivity, and Toxicity of Chemicals from their Structure: A Chemical-Cum-Biochemical Approach, Current Computer-Aided Drug Design, 2013, 9, 449-462.

[7] Basak, S. C.; Majumdar, S. Current landscape of hierarchical QSAR modeling and its applications: Some comments on the importance of mathematical descriptors as well as rigorous statistical methods of model building and validation, in Advances in Mathematical Chemistry and Applications, volume 1, pp. 251-281, Basak, S. C., Restrepo, G. and Villaveces, J. L., Editors, Bentham eBooks, Bentham Science Publishers, 2015.

[8] Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environ. Health Perspect. 1990, 87,183–197.

[9] Basak, S. C.; Harriss, D. K.; Magnuson, V. R. 1988. POLLY v. 2.3: 1988; Copyright of the University of Minnesota.

[10] MolconnZ, Version 4.05, 2003; Hall Ass. Consult.; Quincy, MA..

[11] Basak, S. C.; Grunwald, G. D., APProbe. 1993; Copyright of the University of Minnesota.

[12] Filip, P. A.; Balaban, T. S..; Balaban, A. T. A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability.1987, J. Math. Chem.1, 61-83.

[13] Basak, S.C.; Grunwald, G.D.; Balaban, A.T.TRIPLET: Copyright of the Regents of the University of Minnesota, 1993.

[14] Stewart, J.J.P. MOPAC Version 6.00, QCPE #455, Frank J Seiler Research Laboratory, US Air Force Academy, CO, 1990.

[15] Frisch, M. J. et al. Gaussian 98 (Revision A.11.2). 1998. Pittsburgh, PA, Gaussian, Inc.

[16] Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An Integration of the AQUIRE Data Base and the QSAR System for Use in Ecological Risk Assessments. Sci. Total Environ. 1991, 109/110, 667-670.

[17] Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach, 1997, J. Chem. Inf. Comput. Sci., 37, 651-655.

[18] SYBYL Version 6.2; Tripos Associates, Inc.: St. Louis, MO, 1994.

[19] Basak, S. C. H-Bond; Copyright of the University of Minnesota, 1988.

[20] Gu, Y-C.; Cuyang, Y.; Lien, E.J. (1986). Examination of quantitative relationship of partition coefficient (log P) and molecular weight, dipole moment and hydrogen bond capability of miscellaneous compounds. 1986, J. Mol. Sci., 4, 89- 95.

[21] Tang, C.; Zhang, L.; Zhang, A.; Ramanathan, M. In: Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis, Proceedings of BIBE 2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, November 4-5, 2001; Bilof, R.; Palagi, L., Eds.; IEEE Computer Society: Los Alamitos, CA, 2001; pp. 41-48.

[22] Soderman, J.V. CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database, Boca Raton, Florida, 1982.

[23] Gute, B. D.; Basak, S. C. Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, SAR QSAR Environ. Res., 1997, 7, 117–131.

[24] Gute, G. D.; Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach, B.D. SAR QSAR Environ. Res., 1999, 10, 1–15.

[25] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach, , J. Chem. Inf. Comput. Sci., 2001, 41, 671–678.

[26] Hawkins, D. M.; Basak, S. C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics, Environ. Toxicol. Pharmacol., 2004, 16, 37–44.

[27] Gute, B. D.; Basak, S. C.; Balasubramanian, K.; Geiss, K.; Hawkins, D. M. Prediction of halocarbon toxicity from structure: A hierarchical QSAR approach, Environ. Toxicol. Pharmacol., 2004, 16, 121–129.

[28] Basak, S. C.; Natarajan, R.; Mills, D. Structure-activity relationships for mosquito repellent aminoamides using the hierarchical QSAR method based on calculated molecular descriptors, Conference proceedings, WSEAS Transactions on Information Science and Applications, 2005, 7, 958–963.

[29] Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. Prediction of tissue: air partition coefficients: A comparison of structure-based and property-based methods, SAR QSAR Environ. Res., 2002, 13, 649–665.

[30] Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding potency of dibenzofurans: A hierarchical QSAR approach. Indian. J. Chem., 2003, 42A, 1385–1391.

[31] Basak, S. C.; Bhattacharjee, A. K.; Vracko, M. Big Data and New Drug Discovery: Tackling "Big Data" for Virtual Screening of Large Compound Databases. Current Computer-Aided Drug Design, 2015, 11, 197-201.

[32] Basak, S. C. Philosophy of Mathematical Chemistry: A Personal Perspective,
HYLE--International Journal for Philosophy of Chemistry, 2013, 19, 3-17.