SciForum
Mol2Net

Two QSAR Paradigms- Congenericity Principle *versus* Diversity Begets Diversity Principle- Analyzed Using Computed Mathematical Chemodescriptors of Homogeneous and Diverse Sets of Chemical Mutagens

Subhash C. Basak [1]* 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, USA

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

Published: 4 December 2015

Abstract: The age old paradigm of quantitative structure-activity relationship (QSAR) is the congenericity principle which states that similar structures usually have similar properties. But these days a lot of large and structurally diverse data sets of chemicals with the same experimental data (dependent variable) are available. Starting with the same classes of descriptors we extracted the two subsets of the most significant predictors for the formulation of QSARs for two sets of chemicals: A homogeneous set of 95 amine mutagens and a diverse set of 508 structurally diverse mutagens. The predictors included calculated topostructural (TS), topochemical (TC), geometrical, and quantum chemical (QC) indices. Whereas for the homogeneous amines, a small group of descriptors were sufficient for QSAR development, for the 508 diverse set we needed a large and diverse set of indices for effective QSAR formulation. This empirical study thus vindicates the DIVERSITY BEGETS DIVERSITY paradigm of QSAR.

Keywords: quantitative structure-activity relationship (QSAR); Congenericity principle; Diversity begets diversity paradigm; Aromatic amine mutagens; Structurally diverse mutagens; Topostructural (TS) indices; Topochemical (TC) indices; Geometrical indices; Quantum chemical indices.

1. Introduction

Quantitative structure-activity relationships (QSARs) pertaining to the prediction of physicochemical, pharmacological, and toxicological properties of chemicals are mathematical models developed for the prediction of properties of chemicals from their physical properties or structural descriptors [1-10]. The basic idea underlying QSAR development can be conveniently expressed by the following equation:

P = f (S) ………….. Eq. 1

where P is any physical, biological, medicinal or toxicological property of interest and S represents the relevant aspect of the structure that determines the property. A look at the recently published literature would show that various classes of calculated properties, viz., topological, geometrical, quantum chemical, substructural, are used routinely in QSAR formulation [2-15].

A perusal of QSAR literature would indicate that in many cases QSARs are developed based on properties of a set of structurally related molecules. This is based on the "congenericity principle" or the "structure-property similarity principle" which states that similar structures usually have similar properties [11]. But in many practical situations one has to develop models for the prediction of property/ bioactivity of sets of chemicals which are structurally diverse instead of being homogeneous [12]. In the course of carrying out principal components analysis (PCA) of homogeneous and diverse data sets, Basak et al [13, 14] noted that a larger number of PCs are required to explain an analogous percentage of variance for diverse data sets as compared to homogeneous collection of molecules. From such QSAR studies Basak [15] formulated the

"diversity begets diversity principle," which states that we need a diverse collection of descriptors independent variables) if we want to develop a QSAR for structurally diverse sets of chemicals. We have tested this hypothesis using two data sets: a) Mutagenicity of a homogeneous set of 95 aromatic and heteroaromatic amines, and b) Mutagenic activity of a large diverse set of 508 chemicals.

2. Results and Discussion

The results of QSAR based on the 95 aromatic amine mutagens [16] and 508 diverse mutagens [17] are shown in Table 1 based on calculated descriptors described in Table 2. After QSAR development of the two sets of mutagens, one congeneric and the other structurally diverse, we sorted the descriptors based on their significance as measured by [t] values which are given in Table 3. As evident from data in Table 3, for the set of congeneric mutagens, only seven molecular descriptors of limited class diversity were sufficient to give a reasonably good QSAR. Starting from the same set of calculated descriptors (Table 2), the significant descriptors for the 508 diverse mutagens needed 42 descriptors for good QSAR development. Whereas the indices needed for 95 amines fall into some narrow classes, those needed for 508 chemical set need not only higher number of descriptors, but also heterogeneous types of descriptors. For example, the diverse set of mutagens needed triplet indices (ASV1), information theoretic indices of neighborhood complexity (IC, SIC, CIC indices of different orders), and the quantum chemical descriptors (HOMO, LUMO) which were not selected for the congeneric amine data set. This supports the dichotomy in the QSAR paradigms: Congenericity principle for congeneric data set and diversity begets diversity principle for structurally diverse situations.

Table 1: Results of QSAR based on the 95 aromatic amine mutagens and 508 diverse mutagens

508 compound	No of predictors	% Correct classification	Sensitivity	Specificity
TS+ TC Model (Ridge Regression)	298	76.97	83.98	69.84
TS + TC Model (using ITC+ Ridge Regression)	298	73.23	77.34	69.05
95 Aromatic amine mutagens				
TS+ TC Model (Ridge Regression)	266	84.21	77.36	92.86
TS + TC Model (using ITC+ Ridge Regression)	266	89.47	92.45	85.71

Table 2: Molecular Descriptors used for QSAR development

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance *h*
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order *h* = 0-10
${}^h\chi_C$	Cluster connectivity index of order *h* = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order *h* = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order *h* = 3-10
P_h	Number of paths of length *h* = 0-10
J	Balaban's *J* index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph

DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
	Topochemical (TC)
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
$^{h}\chi^{b}$	Bond path connectivity index of order h = 0-6
$^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3-6
$^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3- 6
$^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4-6
$^{h}\chi^{v}$	Valence path connectivity index of order h = 0-10
$^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3-6
$^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3-10
$^{h}\chi^{v}_{PC}$	Valence path-cluster connectivity index of order h = 4-6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5

AZN$_y$	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ$_y$	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ$_y$	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^{2}Z$_y$	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index *t*
sumI	Sum of the intrinsic state values *I*
sumdelI	Sum of delta-*I* values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp_1* kp_2/nvx)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wiener *p*
Pf	Platt *f*
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
NumHBd	Number of hydrogen bond donors
NumHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =*CH-*
SHtvin	E-State of C atoms in the terminal vinyl group, =CH_2
SHavin	E-State of C atoms in the vinyl group, =*CH-*, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –*OH*, =*NH*, -NH_2, -*NH-*,-*SH*, and #*CH*
SHwHBd	Weak hydrogen bond donor index, sum of *C-H* Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the *E*-State values for –*OH*, =*NH*, -NH_2, -*NH-*, >*N*, -*O*-, -*S*-, along with –F and –Cl
Qv	General Polarity descriptor
NHBint$_y$	Count of potential internal hydrogen bonders (y = 2-10)
SHBinty	E-State descriptors of potential internal hydrogen bond strength (*y =2-10*)
ka_1-ka_3	Kappa alpha indices

Electrotopological State index values for atom types:

	SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax, Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
	Geometrical (3-D)
^{3D}W	3D Wiener number based on the hydrogen-suppressed geometric distance matrix
$^{3D}W_H$	3D Wiener number based on the hydrogen-filled geometric distance matrix
V_W	Van der Waal's volume
	Quantum Chemical (QC)
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO-1}	Energy of the second highest occupied molecular
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO+1}	Energy of the second lowest unoccupied molecular orbital
ΔHf	Heat of formation
μ	Dipole moment

Table 3: Most significant descriptors for the two sets based on *t*-ratio.

508 compound dataset (42 descriptors have significant t-ratios)

t-ratio	*Descriptor name*	*t-ratio*	*Descriptor name*
-30.78	ASV_1	4.75	kp_2
-21.26	$SHBint_6$	-4.09	*StN*
-19.59	*S1*	-3.76	*SddC*
-17.50	*HF*	-3.67	$\overline{IC}$
11.49	*S2*	-3.47	SIC_4
-10.80	SIC_3	3.39	kp_3
-10.51	E_{HOMO}	-3.32	*totop*
9.35	E_{LUMO+1}	-3.14	*SHdNH*
-8.75	E_{LUMO+1}	3.06	*SsF*
8.53	IC_6	-3.05	S_6
7.93	*SHssNH*	2.96	*nelem*
7.83	*SHHBa*	2.96	CIC_4
-7.72	ka_1	-2.61	CIC_6
-7.68	*SHarom*	2.53	*SssssC*
7.24	*SaaaC*	-2.35	*SaasC*
6.64	$DN^2 1_2$	-2.33	$DS1_1$

-6.35	*DN^2I_3*	*-2.33*	*$NHBint_3$*
6.00	*μ*	*-2.22*	*SsNH2*
-5.50	*SIC_0*	*2.19*	*DSZ_2*
-5.39	*kp_1*	*2.04*	*knotpv*
4.85	*SssH*	*2.03*	*SaaCH*

95 compound dataset
(7 descriptors have significant t-ratios)

t-ratio	**Descriptor name**
-8.89	**SsNH2**
-5.72	**NHBint3**
4.65	**NHBint9**
-3.63	**NumHBd**
-3.20	**NumHBd**
-3.17	**NumHBd**
2.54	**SssNH**

3. Materials and Methods

A machine learning method called Interrelated Two-way clustering (ITC), originally developed for application in gene microarray data [72], is used for variable selection, and resulting predictors are fed into a ridge regression model to get final predictions. The ITC algorithm involves the following steps:

i. Predictors are clustered into separate functional groups, say $G_1, G_2, \ldots, G_k$, which are substituted by several types of descriptors in QSAR;

ii. After that samples are clustered into two classes using each functional group, Say $S_{i,a}$ and $S_{i,b}$; $i = 1,2, \ldots, n$;

iii. All possible intersections of the 2^k clusters are taken. For example, for $k = 2$ the intersections are:

$$C_1 = S_{1,a} \cap S_{2,a};\ C_2 = S_{1,b} \cap S_{2,a};$$
$$C_3 = S_{1,a} \cap S_{2,b};\ C_4 = S_{1,b} \cap S_{2,b}$$

iv. These are divided into heterogeneous groups: pairs of intersections with no common elements, e.g. $H_{14} = (C_1, C_4)$ and $H_{23} = (C_2, C_3)$ above;

v. For each $H_{st} = (C_s, C_t)$, cosine distances of subvectors with predictors from this heterogeneous group are calculated with the two model vectors: one with C_s zeros and C_t ones, and another with C_s ones and C_t zeros. Each distance vector is sorted in decreasing order, top one-third of predictors are taken from each of these vectors and are merged.

The algorithm is then repeated with selected predictors, and terminated when 90% of total number of samples is covered by the largest heterogeneous group, or maximum number of iterations reached. This is done because through the algorithm the functional groups become more and more similar, so sample classifications using them become more and more similar, thus heterogeneous groups cover an increasing proportion of total number of samples.

To tackle high collinearity among different predictors, after variable selection through ITC we use ridge regression to build the predictive models. Given n samples and p variables, the $n \times p$ data matrix of predictors $\mathbf{X}$ and $n \times 1$ vector of 0/1 responses $\mathbf{Y}$, the vector of coefficients obtained by ridge regression is defined as:

$$\mathbf{b} = (\mathbf{X'X} + k\mathbf{I})^{-1}\mathbf{Y} \qquad (1)$$

Where $k > 0$ is the ridge constant, chosen by cross-validation.

The aim of research in this paper was to test the hypothesis that congeneric data sets need a structurally narrow set of descriptors for QSAR formulation as compared to diverse sets which require diverse collection of independent variables for model building. The results derived from two data sets, viz. a congeneric set of 95 amines and a diverse set of 508 structurally diverse mutagens appear to support the "diversity begets diversity" hypothesis. Further QSAR studies on other congeneric and diverse data sets are necessary to test the validity of this hypothesis.

3. Conclusions

Acknowledgments.

Author Contributions
Subhash C. Basak developed the hypothesis and Subhabrata Majumdar carried out the data analysis for the paper.

Conflicts of Interest
"The authors declare no conflict of interest".

References and Notes

[1] Hansch, C.; Leo, A., Exploring QSARs: Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC 1995.-

[2] Kier, L.B.; Hall, L. Molecular Structure Description: The Electrotopological State; Academic Press: San Diego, CA, 1999, pp. 245.

[3] Devillers, J.; Balaban, A.T., Eds. Topological Indices and Related Descriptors in QSAR and QSPR; Gordon and Breach: Amsterdam, 1999, pp. 811.

[4] Diudea, M.V., Ed. QSPR / QSAR Studies by Molecular Descriptors; Nova: Huntington, N.Y., 2001, pp. 438.

[5] Karelson, M. Molecular Descriptors in QSAR/QSPR; Wiley-Interscience: New York, 2000, pp. 448.

[6] Balaban, A.T., Ed. From Chemical Topology To Three-Dimensional Geometry; Plenum Press: 1997,

pp. 420.

[7] Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics; Wiley-VCH: Weinheim, 2009,Vol. I, pp. 967; Vol. II, pp. 257.

[8[Hawkins, D. M.; Basak, S. C.; Kraker, J. J.; Geiss, K. T.; Witzmann, F. A., Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling, J. Chem. Inf. Model., 2006, 46, 9–16.

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

[10] Basak, S. C.; Mills, D.; Hawkins, D. M., Predicting allergic contact dermatitis: A hierarchical structure-activity relationship (SAR) approach to chemical classification using topological and quantum chemical descriptors. J. Comput. Aided Mol. Des., 2008, 22, 339-343.

[11] Johnson, M, Basak, S. C.;, Maggiora, G., A characterization of molecular similarity methods for property prediction. Mathl. Comput. Modelling 1988, 11, 630–634.

[12] Basak, S. C.; Grunwald, G. D.; Host, G.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical and neural network methods for predicting toxic modes of action of chemicals, Environ. Toxicol. Chem., 1998, 17, 1056–1064.

[13] Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of Mathematical Structural Invariants in Analyzing, Combinatorial Libraries: A Case Study with psoralen Derivatives, Current Computer-Aided Drug Design, 2010, 6, 240-251.

[14] Basak, S. C.; Grunwald, G. D., A COMPARATIVE STUDY OF GRAPH INVARIANTS, TOTAL SURFACE AREA AND VOLUME IN PREDICTING BOILING POINTS OF ALKANES, Math Modelling & Sci. Computing, 1993, 2, 735-740.

[15] 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.

[16] Debnath, A.K.; Debnath, G.; Shusterman, A.J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella typhimurium TA98 and TA100. Environ. Mol. Mutagen. 1992, 19, 37-52.

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