J. Eng. Sci. Manage. Educ. Vol-7 (III) 166-, 2014

Journal of Engineering, Science & Management Education

Molecular Similarity and Hazard Assessment of Chemicals: A Comparative Study Arbitrary and Tailored Similarity Spaces

Subhash C. Basak

International Society of Mathematical Chemistry,

1802 Stanford Avenue, Duluth, MN 55811 and UMD-NRRI, 5013

Miller Trunk Highway, Duluth MN 55811, USA; sbasak@nrri.umn.edu

Corresponding Author

Article History : Received: 1 Revised: 5th Accepted: 1

Abstract- *This review discusses the utility of various types of quantitative molecular similarity analysis (QMSA) methods derived from experimental properties as well as computed chemodescriptors in the selection of analogs. Such analogs have been used in the estimation of property, bioactivity/ toxicity and modes of action of chemicals. Comparative studies of arbitrary versus tailored (property-specific) QMSA methods have been reviewed with special reference to studies carried out by Basak and coworkers since 1985 until the present time. The use of dissimilarity based clustering of databases for drug discovery has been discussed.*

Keywords: *quantitative molecular similarity analysis (QMSA), tolerance relation, molecular similarity, arbitrary similarity, tailored similarity, topostructural indices (TSIs), topochemical indices (TCI), principal components analysis (PCA), principal components (PCs), analog selection, mode of action (MOA), molecular dissimilarity, mutagenicity, boiling point, structure space, property space, clustering of databases, Euclidean distance*

"All cases are unique and very similar to the others"
T. S. Eliot, In: The Cocktail Party

1- INTRODUCTION

Molecular similarity/dissimilarity has important applications in chemistry, new drug discovery, and hazard assessment of chemicals[1-19]. Human beings have used the intuitive notion of similarity of objects from time immemorial. For example, in drug design if a new compound discovered by serendipity is found to possess a useful therapeutic profile, the drug designer would like to know whether that chemical's analogs (structurally similar molecules) also possess similar biological properties. In the hazard estimation of industrial chemicals, one has to operate in a data poor situation. The Toxic Substances Control Act (TSCA) Inventory of the United States Environmental Protection Agency (USEPA) currently has over 86,000 chemicals. Most of the TSCA chemicals have very little or no experimental data required for their toxicity estimation. In the face of this lack of available data, two approaches are used by the regulators: a) class-specific quantitative structure-activity relationship (QSAR) models and b) QMSA based modeling of properties using structural analogs [20].

The fundamental notion behind the use of molecular similarity arises from the structure-property similarity principle which states that similar structures usually have similar properties [1]. This principle is based on the notion that the relationships between the structures of molecules and their properties are guided by smooth, although unknown, mathematical functions. To find analogs of a candidate chemical, one might search large public domain or proprietary databases. For that one needs methods that can select analogs of a chemical fast and using properties which can be calculated directly from molecular structure without the input of any other experimental data because such data are expensive and often not available.

2- TOLERANCE RELATION AND MOLECULAR SIMILARITY

As pointed out by Basak and Grunwald [11], from a mathematical point of view similarity or resemblance among elements of a set of objects is characterized by the tolerance relation.

Definition: The relation A on a set M is called tolerance (or tolerance relation) if it is reflexive and symmetric [21, 22].

When a relation defined on a set is an equivalence relation, such a relation is reflexive, symmetric, and transitive. The absence of transitivity in the tolerance relation has important consequences for similarity. To demonstrate this point, let us take a set of five-letter English words and call any two words similar if they differ at most by one letter. Let us consider the following sequence of words:

White, while, whale, shale, shave, stave, stare, stark, stack, slack, black

Any two consecutive words in the above sequence are similar

J. Eng. Sci. Manage. Educ. Vol-7 (III) 166-, 2014

by our definition of similarity. It is clear that distinct "similar" neighbors of any specific word are mutually dissimilar. This arises out of the fact that the tolerance relation is not transitive. What happens with the above sequence of words may also happen in the practical case of analog selection by molecular similarity methods if we are not careful. One may start with a chemical, find its near neighbor, and find, in turn, an analog of the near neighbor and so on. If this process continues, it is not difficult to imagine that at some stage the chosen analog will be structurally quite dissimilar to the starting chemical, just like the example above where we started with the word "white" and ended with the word "black" through a sequence of similar neighbors.

3- ARBITRARY AND TAILORED SIMILARITY METHODS

"From the words of the poet, men take what meanings please them; yet their last meaning points to thee."
Rabindranath Tagore,
Poem #75, Gitanjali

The majority of currently known QMSA techniques belong to the category of arbitrary similarity methods. Basak et al. called such method arbitrary because the attributes used to measure intermolecular similarity in these methods are selected subjectively by the practitioner based on his/ her intuitive view of similarity. Because the various properties/ bioactivities that we need to assess using QMSA techniques are not often mutually strongly correlated, one arbitrary QMSA protocol cannot be relevant to the various properties we need to estimate for chemistry, drug design, and predictive toxicology. To correct this situation, Basak *et al*. [18] developed the idea of tailored QMSA method. In this approach, one starts by selecting the set of attributes to measure intermolecular similarity based on a specific property of interest. The tailored structure space is then used in selecting analogs and estimating properties of chemicals from their chosen analogs. Investigation by Basak and coworkers [14, 16-18] on the relative effectiveness of arbitrary versus tailored QMSA methods showed that tailored QMSA methods outperformed the arbitrary QMSA models.

4- ATTRIBUTES USED TO MEASURE INTERMOLECULAR SIMILARITY

Whenever we talk about the quantification of molecular similarity, we are confronted with the question: Similarity with respect to what? In other words, which experimental or computed properties can we use for the computation of similarity of two molecules in a database? Use of experimental properties is not often very useful. First, relevant experimental data are not available for many chemicals of interest in drug design and environmental protection. Second, experimental properties are available only for those substances which already exist [23]. But both drug designer and risk assessor may be interested in structures which are hypothetical. As a result, most QMSA methods use properties which can be computed from some structural representation of chemicals. Topological, geometrical, and quantum chemical descriptors fall in this category. Basak et al. [24] used three classes of molecular descriptors in their QMSA research: a) Topostructural indices (TSIs), topochemical indices (TCIs), and c) Atom pairs (APs).

Numerical invariants defined on simple molecular graphs which represent only the adjacency and distance relationship of atoms (vertices) and bonds (edges) are called topostructural indices (TSIs). On the other hand, invariants derived from weighted molecular graphs that represent both the chemical nature of vertices (atoms) and bonds (edges) are called topochemical indices (TCIs). Collectively, the TSI and TCI descriptors are known as topological indices (TIs).

Basak et al. [11] also used atom pairs (APs), which are fragment-based descriptors as opposed to TSIs and TCIs which characterize the structure of an entire molecule numerically. The method of Carhart et al. [5] was used to calculate the atom pairs, which defines an atom pair as a substructure consisting of two non-hydrogen atoms i and j and their interatomic separation:

<atom descriptori> – <separation> – <atom descriptorj>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of π electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms.

Currently available software like DRAGON [25], MolConn Z [26], POLLY [27] and APProbe [28] can calculate a large number of chemodescriptors from chemical structure. Basak group of researchers have been using all classes of descriptors in the formulation of QSAR models. These descriptors can be computed fast and, consequently, may be applied to analyze medium-sized and large databases.

In their arbitrary and tailored QMSA method development and applications the Basak group of researchers at the University of Minnesota Duluth/ Natural Resources Research Institute (UMD-NRRI) used different sets of TSI+TCI combination, available experimental properties and APs since the early 1980s to the present time. When discussing the individual applications (vide infra) the summary of the individual chemodescriptor sets used for the individual research project will be given.

5- FAILURE OR PILLAR OF SUCCESS?

"Romans, though you're guiltless, you'll still expiate
your fathers' sins, till you've restored the temples,
and the tumbling shrines of all the gods,
and their images, soiled with black smoke."
Horace (65- 8 BC), in Moral Decadence

"I know that cancer often kills,
But so do cars and sleeping pills."
J. B. S. Haldane

At this juncture a personal history and motivation of Subhash C. Basak in the development of QMSA methods seems advisable. He arrived at the University of Minnesota Duluth (UMD) at the beginning of 1982 to work with the UMD and the Duluth's USEPA lab in predicting the aquatic toxicity of a diverse set of chemicals which was a subset of TSCA. As opposed to drugs most of which act via some specific enzymes or receptors, the TSCA chemicals were not designed to target any particular biomacromolecule, but they were used for various industrial purposes over the past few centuries. Therefore, these chemicals are structurally very diverse. These industrial chemicals belong to a few modes of action (MOAs): Narcotic, polar narcotic, acetylcholine esterase inhibitors (AChE-I), uncouples of oxidative phosphorylation, etc. Initially, when properties like calculated log P or indices like connectivity or information theoretic indices were used to develop QSARs of congeneric sets of chemicals, statistically significant correlations were found. For the aquatic toxicity (LC_{50} in fathead minnow) of esters more than one class of descriptors were needed [29]. When attempt was made to correlated LC_{50} of more diverse data sets, every class of descriptor failed.

A lot of research in QSAR is based on the structure-property similarity principle [1] or the congenericity principle which state that similar compounds usually have similar properties or bioactivities. This similarity or congenericity is a subjective concept which is not often well defined. The tacit assumption underlying this notion is that the relationships between the structures of molecules and their properties are guided by smooth, although unknown, mathematical functions. In other words, small alterations in molecular structure lead to small perturbations in the magnitude of property. A useful concept in organic chemistry is the idea of the homologous series. In therapeutics and toxicology, molecules having the same pharmacophore or toxicophore usually have similar pharmacological or toxicological properties.

There are also a large number of examples that run apparently counter to this notion of similarity... On the other hand, there is a large body of literature concerning bioisosteric molecules that lack any apparent structural similarity but are recognized by biological receptors as similar [30].

From the various QSAR studies, hierarchical QSAR (HiQSAR) research in particular, on a large and diverse sets of molecules, we concluded that for structurally homogeneous sets a limited collection of molecular descriptors can give good quality QSARs [31, 32]. But for progressively more diverse data sets we need a more diverse collection of descriptors. This principle may be called the "***diversity begets diversity principle***." It is possible that the complex interactions between the ligands and the biotarget in structurally diverse sets are captured more efficiently by a broad range of calculated descriptors.

In the middle of 1980s, it was realized that a new approach was called for. One idea was to develop molecular similarity methods extracting useful information from a large number of computed descriptors and apply QMSA in the toxicity estimation of chemicals from their selected analogs. In one of the first studies of its kind, Basak et al. [33] carried out principal components analysis (PCA) on 90 calculated topological indices calculated for a set 3,692 chemicals taken from the Duluth USEPA database. We asked the question: How many molecules in the database (>30,000 structures) have at least one good quality melting point, boiling point or vapor pressure data? The answer came out as 3,692. Because these physical properties are related to the fate and effect environmental pollutants, this database was taken for study.

The above data is viewed as n = 3,692 vectors (chemicals) in p = 90 dimensions (indices).

Each chemical is represented by a point inR^{90}. Because many of the 90 indices are highly interrelated, the 3692 points inR^{90}will lie nearly on a subspace of dimension lower than 90. The method of principal components analysis (PCA) or the Karhunen-Loeve transformation is a standard linear method for reduction of dimensionality. Although there are other methods, PCA is the logical starting point in terms of simplicity, ease of interpretation,and ease of computation.

The fundamental question that was asked in this investigation was: What is the intrinsic dimensionality of the apparent 90-dimensional space? The results of PCA showed that the first four principal components (PCs) explained 78.3% of variance in the data; the first ten PCs explained 92.6% of variance in data [8, 33]. So, the reduction of dimensionality from the 90-dimensional calculated index space to the 10-dimensional PC space did not result in much loss of information.

We used two measures of intermolecular similarity in our QMSA research: a) Euclidean distance on the descriptor space derived from orthogonal factors computed from molecular descriptors like TIs, and b) Tanimoto type association coefficient to measure the similarity/dissimilarity of molecules when they were represented by binary descriptors like atom pairs

These techniques have been used to: 1) define a variety of structure spaces to quantify molecular similarity, 2) select analogs, 3) carry out comparative studies of spaces derived from experimental physicochemical properties versus topological descriptors, 4) select a number of neighbors of a chemical in various structure spaces to estimate properties of the target chemical, and 5) estimate toxic modes of action (MOA) of environmental toxicants from the MOA of their analogs.

6- ANALOG SELECTION ON THE SET OF 3,692 TSCA CHEMICALS

Once the 10-dimensional PC space was formulated, it was of interest to check what type of analogs of chemicals are selected by the PC based QMSA method. To this end, ten chemicals were randomly selected from the database of 3,692 molecules and for each of them five nearest neighbors were chosen based

on the Euclidean distance (ED) on the PC space [8]. An inspection of the selected structures showed that there are substantial structural similarity between each of the ten query chemicals and their selected neighbors.

7- ESTIMATION OF PROPERTY/ BIOACTIVITY USING QMSA METHODS

The implication of structure-property similarity principle is that one should be able to estimate properties of chemicals from the properties of their near neighbors. We used the k nearest neighbor (KNN) method to test this tacit assumption. We give below such results with two properties: a) Normal boiling point and Ames mutagenicity using arbitrary QMSA methods [111].

7.1 Normal Boiling Point Estimation

The normal boiling point database consisted of 2966 compounds taken from the US EPA's collection. Compounds selected were those that have measured boiling point values. Calculated Tis were divided into disjoint subsets or clusters using the correlationmatrix by applying the VARCLUS procedure of SAS. This method divides the TIs into disjoint clusters so that each cluster is essentially unidimensional. The output of this procedure is the first principal component (PC) from each cluster of TIs. These PCs were subsequently used in the similarity measures based on the Euclidean distance method. Various QMSA methods based both on Tis and APs were used for the estimation of boiling point, viz., associative measure using APs and four Euclidean distance (ED) measures using: a) scaled TIs (TI_s); b) unscaled TIs (TI_u); c) scaled variable clusters (PC_s); and d) unscaled variable clusters (PC_u). The two methods using scaled dimensions involved rescaling the variables to have mean equal to zero and variance equal to one prior tocalculating ED. For more details see Basak and Grunwald [11]. In Table 1 below we give summary results for the TI_s and TI_u methods.

Table 1. Comparison of two similarity methods for prediction of normal boiling point (^oC) using K nearest neighbors

Similarity method	N	K	r	s. e. (0C)
TI_s	2926	10	0.874	40.0
TI_u	2926	7	0.872	40.3

In Table 1 above, N represents the number of data points, K is the number of selected neighbors' used for K nearest neighbor (KNN) based estimation of boiling point, r is the correlation coefficient for the correlation of the experimental boiling point with the estimated value, and s. e. is the standard error.

7.2 Estimation of Ames Mutagenicity of Aromatic and Heteroaromatic Amines

Various QMSA methods and different values of K (1-7) were used to estimate mutagenicity of a set of 95 aromatic and heteroaromatic amines. The mutagenic potency was taken from the literature. The mutagenic activity of these compounds in S. typhimurium TA98 + S9 microsome was collected. . Mutagenic potency is expressed as the mutation rate, Ln (R), in log (revertants/nmol). We give in Table 2 below the results of QMSA based estimation of mutagenicity using three methods, viz., AP, TI_s and TI_u.

Table 2. Correlation (r) and standard error (s.e.) of estimating Ln(R) of 95 aromatic amines using three different similarity methods

Similarity Method	N	K	r	s. e.
AP	95	7	0.830	1.08
TI_s	95	7	0.834	1.06
TI_u	95	7	0.811	1.13

8- COMPARISON OF SIMILARITY METHODS BASED ON PROPERTY SPACES AND STRUCTURE SPACES

We were interested to see what kind of analogs are selected based on physicochemical properties vis-à-vis topological chemodescriptors. Because of the paucity of available experimental data, one of the data sets to which this method was applied had only seventy-six chemicals for which six experimental properties, viz., lipophilicity ($logK_{OW}$), boiling point, melting point, molar volume (V/100), hydrogen bond donor acidity (α), hydrogen bond acceptor basicity (β), and polarizability (π*), were available.

Qualitatively, our results on the selection of analogs using physicochemical and topological spaces show that for any particular query chemical the selected set of neighbors is essentially the same with some minor variation, though the order of neighbor selection differs. The details of this analysis can be found in Basak *et al.* [16].

9- TAILORED QMSA METHODS

Basak *et al.* [18] developed for the first time the idea of tailored similarity. The fundamental difference between tailored and arbitrary QMSA techniques arises out of the selection of descriptors. Whereas arbitrary QMSA methods select descriptors that best characterize the variance in the descriptor set, the tailored approach uses the statistical method of ridge regression (RR) to select a subset of descriptors optimal for the property of interest. The ridge regression method is useful in cases where the descriptors are highly multicollinear and where the number of descriptors is substantially larger than the number of observations. The details of our research on tailored QMSA are not given here for brevity. We noted that in each case of our comparative study of arbitrary versus tailored QMSA methods, the tailored approach outperformed the arbitrary ones [14, 16-18].

10- MOLECULAR DISSIMILARITY AND CLUSTERING OF DATABASES

The various similarity spaces discussed above can also be used for cluster-analysis. This type of method is useful in scanning large real or virtual chemical libraries in looking for new pharmaceutical leads or for other testing problems in which the number of compounds is very large, and therefore it is too expensive, to exhaustively subject the entire set to toxicological bioassay. In this situation, representative subsets from each of the clusters can be tested on the assumption that since the compounds within each cluster are similar, their properties should also be similar.

In the 1980s, the software POLLY [27] and the arbitrary QMSA method derived from the PCs was implemented at the Upjohn Company (now part of Pfizer) and they used this method, called the Basak method, in clustering their proprietary databases for the discovery of many new drug leads [34].

11- PREDICTION OF TOXIC MODES OF ACTION (MOA) FROM QMSA METHODS

It may be argued from the structure-property similarity principle that similar chemicals should have analogous biochemical modes of action (MOA). To test this idea a large and structurally diverse set of 283 chemicals were selected for which MOA of aquatic toxicity was known with a high degree of confidence. Thereafter, the MOA of these toxicants were estimated from the MOA of their five nearest neighbors. The MOA data was categorized as: narcosis I (baseline narcosis), narcosis II (polar narcosis), mixed narcosis I/II, oxidative phosphorylation uncouplers, acetylcholinesterase (AchE) inhibition, electrophile/proelectrophile, and chemicals affecting the central nervous system (neurotoxicants and respiratory blockers). kNN estimation, as well as neural network and discriminant analysis techniques, were used to attempt to correctly classify these chemicals. The results showed that KNN method based on calculated PCs derived from topological indices could correctly classify the MOAs of the set of toxicants satisfactorily [35].

12- DISCUSSION AND OBSERVATIONS

"Computers are incredibly fast, accurate, and stupid. Human beings are
Incredibly slow, inaccurate, and brilliant. Together they are powerful beyond imagination."
Albert Einstein

Chemists, pharmacologists, and toxicologists have routinely used the concept of molecular similarity in their day to day analyses of data. In the past few decades, the landscape of attributes used to quantify intermolecular similarity has undergone a dramatic change. Whereas experimental properties useful for measuring similarity of chemicals are often not available, the currently available software can compute a lot of structural properties of chemicals or chemodescriptors which are finding useful applications in lieu of experimental data. The upsurge of chemodescriptor research, particular those based on chemical graph theory, have been fueled by two major factors: a) development of a plethora of novel concepts, and b) availability of high speed computers and associated software whereby hypothesis driven as well as discovery oriented research on large data sets could be carried out fast [36]. A dissimilarity based analysis of a virtual library of almost 250,000 psoralen derivatives carried out by Basak et al. [37] cannot be imagined to be done without the use of high speed computers and efficient software capable of calculating molecular descriptors in a reasonable time scale.

Arbitrary QMSA methods are based on the subjective intuition of the practitioner. The philosopher Bertrand Russell [38] pointed out : “Intuition, in fact, is an aspect and development of instinct, and, like all instincts, is admirable in those customary surroundings which have moulded the habits of the animal in question, but totally incompetent as soon as the surroundings are changed in a way which demands some non-habitual mode of action.”

Any arbitrary similarity method should be looked upon as a subjective conjecture or inductive inference which needs to be validated or falsified by additional data [39]. But statistically derived tailored similarity methods seem to be more powerful in predicting properties as compared to the arbitrary QMSA methods.

13- QUO VADIMUS

Good tests kill flawed theories; we remain alive to guess again.

Karl Popper

In this review article, we mainly discussed QMSA methods based on chemodescriptors. But in the post-genomic era a lot of data are being generated on the omics side. The biodescriptors derived from such data may be analyzed using techniques like PCA or RR to gain insight into the MOA and to aid in drug design and predictive toxicology. In particular, our Natural Resources Research Institute team and collaborators have been involved in developing numerical descriptors for DNA sequences [40] and proteomics patterns [41] which, analogous to calculated chemodescriptors, characterize biological molecules or systems such as DNA sequences or the patterns of distribution of proteins in the 2-D gels. It is tempting to speculate that such “biodescriptors” will gradually find application in similarity/dissimilarity analyses of molecular systems and provide new insight about their biological function.

14- ACKNOWLEDGMENT

I am grateful to a large number of global collaborators, collectively called my “virtual team” for their sustained collaboration in carrying out the research reviewed in this paper. I would like to specially mention Gregory D. Grunwald, Gerald J. Niemi, Gilman D. Veith, Brian D. Gute and Denise Mills for very useful collaboration.

References

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

[2] Johnson, M.; Maggiora, G. M. Concepts and Applications of Molecular Similarity; John Wiley & Sons, Inc.: New York, 1990.

[3] Willett, P.; Barnard, J.; Downs, G. Chemical similarity searching. J. Chem. Inf. Comput. Sci. 1998, 38, 983-996.

[4] Carbo-Dorca, R.; Mezey, P. G., Advances in Molecular Similarity. Vol. 2; Eds.; JAI Press: Stamford, CN., 1998

[5] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. J. Chem. Inf. Comput. Sci. 1985, 25, 64-73.

[6] Stumpfe, D.; Bajorath, J. Similarity searching Wiley Interdisc. Rev.: Comp. Mol. Sci. 2011, 1, 260-282.

[7] Willett, P. Similarity methods in chemoinformatics. Ann. Rev. Inf. Sci. Technol. 2009, 43, 3-71.

[8] Basak, S. C.; , Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices, Discrete Appl. Math.,1988, 19, 17–44, (1988); Special volume: Applications of Graph Theory in Chemistry and Physics, J.W. Kennedy and L.V. Quintas (Eds.).

[9] Basak, S. C.; Grunwald, G. D. Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. SAR QSAR Environ. Res. 1994, 2, 289–307.

10] Basak, S. C.; Grunwald, G. D. Molecular similarity and estimation of molecular properties. J. Chem. Inf. Comput. Sci. 1995, 35, 366–372.

[11] Basak, S. C.; Grunwald, G. D. Tolerance space and molecular similarity. SAR QSAR Environ. Res. 1995, 3, 265–277.

[12] Gute, B. D.; Grunwald, G. D.; Mills, D.; Basak, S. C. Molecular similarity based estimation of properties: A comparison of structure spaces and property spaces. SAR QSAR Environ. Res. 2001, 11, 363–382.

[13] Gute, B. D.; Basak, S.C. Molecular similarity-based estimation of properties: A comparison of three structure spaces. J. Mol. Graph. Model. 2001, 20, 95–109 (2001).

[14] Gute, B. D.; Basak, S. C Optimal neighbor selection in molecular similarity: comparison of arbitrary versus tailored prediction spaces, SAR QSAR Environ. Res. 2006, 17, 37–51.

[15] Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the molecular similarity of chemicals using topological invariants, , In: Advances in Molecular Similarity, Ramon Carbo-Dorca, R. and Mezey, P. G., Eds., pp. 171–185, volume 2, JAI Press, Stanford, Connecticut (1998).

[16] Basak, S. C.; Gute, B. D.; Mills, D. Quantitative molecular similarity analysis (QMSA) methods for property estimation: A comparison of property-based, arbitrary, and tailored similarity spaces, SAR QSAR Environ. Res., 2002,, 13, 727–742.

[17] Gute, B. D.; Basak, S. C.; Mills, D.; Hawkins, D. M. Tailored similarity spaces for the prediction of physicochemical properties. Internet Elect. J. Mol. Des. 2002, 1, 374–387.

[18] Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. Quantitative molecular similarity methods in the property/ toxicity estimation of chemicals: A comparison of arbitrary versus tailored similarity spaces. J. Mol. Struct. THEOCHEM, 2003, 622, 127–145.

[19] Basak, S. C. Similarity methods in analog selection, property estimation and clustering of diverse chemicals. ARKIVOC 2006, 157–210.

[20] 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.

[21] Schreider, J.A. Equality, Resemblance, and Order; Mir Publishers: Moscow, 1975.

[22] Zeeman, E. C., The topology of the brain and visual perception, Topology of 3-manifolds and related topics (Proc. The University of Georgia Institute, 1961), 240-256, Prentice-Hall, 1962.

[23] Vracko, M., Mathematical (Structural) Descriptors in QSAR: Applications in Drug Design and Environmental Toxicology, In: , In: Advances in Mathematical Chemistry and Applications, eBook volume 1, Bentham Science Publishers, pp. 222-250, in press, 2014.

[24] Basak, S. C. Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications, , In: Advances in Mathematical Chemistry and Applications, eBook volume 1, Bentham Science Publishers, pp. 3-23, in press, 2014.

[25] DRAGON – Software for the Calculation of Molecular Descriptors, Version 5.4, 2006; Todeschini, R.; Consonni, V.; Mauri, A. et al., Talete srl.; Milan, Italy

[26] MolConnZ, Version 4.05, 2003; Hall Ass. Consult.; Quincy, MA.

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

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

[29] Basak, S. C.; Gieschen, D. P.; Magnuson, V. R. A quantitative correlation of the LC_{50} values of esters in Pimephales promelas using physicochemical and topological parameters. Environ. Toxicol. Chem. 1984, 3, 191–199.

[30] Thornber, C. W. Isosterism and molecular modification in drug design. Chem. Soc. Rev. 1979, 8, 563-580.

[31] 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. J. Chem. Inform. Comput. Sci. 1997, 37, 651–655.

[32] Basak, S. C., Mathematical descriptors for the prediction of property, bioactivity, and toxicity of chemicals from their structure: a chemical-cum-biochemical approach. Curr. Comput.-Aided Drug Des. 2013, 9, 449-462.

[33] Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological indices: their nature, mutual relatedness, and applications, Mathematical Modelling1987, 8, 300–305.

[34] Lajiness, M.S. (1990). Molecular similarity-based methods for selecting compounds for screening. In, Computational Chemical Graph Theory (D.H. Rouvray, Ed.). Nova, New York, pp. 299-316.

[35] Basak, S. C.; Grunwald, G. D.; Host, G. E.; 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.

[36] Basak, S. C. Philosophy of Mathematical Chemistry: A personal perspective. HYLE—Int. J. Phil. Chem. 2013, 19, 3–17.

[37] Basak, S. C.; Mills, D.; Gute, B.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of mathematical structural invariants in analyzing combinatorial libraries: a case study with psoralen derivatives. Curr. Comput.-Aided Drug Des. 2010, 6, 240-251.

[38] Russell, B. Mysticism and Logic. George Allen & Unwin, Ltd.; London 1950.

[39] Popper, K. "The Logic of Scientific Discovery", Hutchinson, London 1959.

[40] Nandy, A.; Harle, M.; Basak, S.C. Mathematical descriptors of DNA sequences: development and applications. Arkivoc,2006, 9, 211-238.

[41] Basak, S. C.; Gute, B. D. Mathematical descriptors of proteomics maps: Background and applications. Curr. Opin. Drug Discov. Develep. 2008, 11, 320–326.