Use of Mathematical Chemodescriptors and Biodescriptors for New Drug Discovery, Environmental Protection, and Surveillance of Emerging Global Pathogens

Subhash C. BASAK

Natural Resources Research Institute and Department of Chemistry & Biochemistry, University of Minnesota Duluth,
5013 Miller Trunk Highway, Duluth, Minnesota 55811, USA
e-mail: sbasak@nrri.umn.edu

(Received: February 27, 2017; Accepted for publication: March 30, 2017; Online publication: June 6, 2017)

This article reviews results of research on the development of graph theoretical chemodescriptors, topological indices in particular, and proteomics as well as DNA/RNA sequence based biodescriptors and their applications in predicting property/bioactivity of chemicals as well as viruses. Use of biodescriptors in the characterization of emerging pathogens like the Zika virus (ZIKV) has been discussed. The use of proper statistical methods in model building is emphasized with special reference to research carried out by the author of this review.

Keywords: Mathematical chemodescriptors, Biodescriptors, New drug discovery, Environmental protection, Emerging global pathogens, Graph invariants, Topological index, Quantitative structure-activity/property relationship (QSAR/QSPR)

“The perfection of chemistry might be secured and hastened by the training of the minds of chemists in the mathematical spirit [...]. Besides that, mathematical study is the necessary foundation of all positive science, it has a special use in chemistry in disciplining the mind to a wise severity in the conduct of analysis: and daily observation shows the evil effects of its absence.”

Auguste Comte

1 Introduction

During the past six decades there has been an increased interest in the use of numerical graph invariants or topological indices in the formulation of quantitative structure-activity/property relationship (QSAR/QSPR) models [1–4]. In a molecular graph $G = (V, E)$, V represents the non-empty set of atoms and E usually represents the set of covalent bonds present in the molecule. A generic approach in the development of chemodescriptors (for small molecules) and biodescriptors (for biomolecules

like DNA/RNA/protein sequences and proteomics maps, etc.) from graph theoretic models consists of the following steps:

- (a) Choose a structural model
- (b) Associate a graph or matrix to the selected structural model
- (c) Calculate the corresponding invariant (s) for use as chemo- or biodescriptors

This article will review applications of topological indices to new drug discovery, environmental protection, and surveillance of emerging global pathogens with special reference to research carried out by Basak and collaborators.

2 The Structure-activity Relationship (SAR) Paradigm

The central paradigm of structure-activity relationship (SAR) can be expressed by the following relationship:

$$BR = f(S) \quad (1)$$

In equation (1) BR represents the measured biological re-

Table 1. HiQSAR model (RR) for a diverse set of 508 chemical mutagens/non-mutagens. All four means the model used TS+TC+3D+QC descriptors.

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
	All four	307	77.17	84.38	69.84

sponse and S represents the structural attributes or indices calculated for molecules. Topological indices derived from small molecules or biomolecules may be used to represent S in equation (1).

3 Calculation of Mathematical Descriptors of Molecules and Biomolecules

Harry Wiener [5] was the first to put forward the idea of a structural index for the estimation of properties of molecules from their structure. This index is popularly known as the Wiener index, W . As shown by Hosoya [6] for the first time, the index W can be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed graph G as the sum of entries in the upper triangular submatrix:

$$W = \frac{1}{2} \sum_{ij} d_{ij} = \sum_h h \cdot g_h \quad (2)$$

where g_h is the number of unordered pairs of vertices whose distance is h . It is worth mentioning that Professor Hosoya [6] used the term "topological index" to designate numerical graph invariants for the first time.

In our research during the last four decades we have frequently used connectivity, valence connectivity, electrotopological, information theoretic, and Triplet indices calculated by the software POLLY [7], MolconnZ [8], and Triplet [9]. The graph theoretic chemodescriptors fall into two major categories: a) Numerical invariants defined on simple molecular graphs which represent only the adjacency and distance relationship of atoms (vertices) and bonds (edges); such invariants are called topostructural (TS) indices and b) Topological indices derived from weighted molecular graphs are called topochemical (TC) indices. Collectively, the TS and TC descriptors are known as topological indices (TIs).

4 Quantitative Structure-Activity/Toxicity Relationships Using Chemodescriptors

"Three paths through the high spring grass—One is quicker and I take it."

In: Spring Wind on the Riverbank at Kema

By Yosa Buson (The essential HAIKU)

4.1 Model Development Techniques

Proper statistical methods need to be used for the development of QSAR models particularly when the case is rank deficient, *i.e.*, the number of predictors (p) is much larger than the number of cases (n). For details on proper modeling methods in QSAR development, see [10]

4.2 QSAR for Estimation of Ames' Mutagenicity of 508 Diverse Chemicals

Mutagenic potential of chemicals is important both of drug design and environmental protection. Many chemicals which are mutagenic or genotoxic are also carcinogenic. Ames' mutagenicity is one important endpoint for estimating the potential genotoxicity of chemicals. Ridge regression (RR) method was used to develop hierarchical QSAR (HiQSAR) model for a set of 508 chemically diverse mutagens and non-mutagens. Results in Table 1 show that a TS + TC combination gave the best QSARs for predicting mutagenicity of the 508 chemicals. The addition of 3-D and QC descriptors to the set of independent variables made minimum or no improvement in the quality of the models. Of the 508 chemicals, 256 were mutagens and 252 were non-mutagens based on Ames' Salmonella/microsome mutagenicity assay. For the number of indices in the various classes calculated for this data set, the make-up was: TS (103); TS + TC (298); TS+TC+3D+QC (307).

Such good quality predictive models based on easily calcu-

lated topological TS and TC descriptors may find application in the estimation of chemical mutagenicity [1].

4.3 Prediction of Octanol-Water Partition Co-efficient (Kow) using Topological Indices

Octanol-water partition coefficient (Kow) is a physical property which is

important both for new drug discovery and hazard assessment of environmental pollutants. Starting with a database of measured Kow values of over four thousand chemicals, topological indices were used to develop predictive models for the property. The set of chemicals were first divided into 14 groups based on the number of hydrogen bonds and calculated descriptors were used to develop estimation models for each subset. The results indicated that the predictive power of these models were comparable to other models developed for Kow prediction [11].

4.4 Characterization and Surveillance of Emerging Global Pathogens using Mathematical Sequence Descriptors

In recent years we witnessed a global upsurge in the spread of lethal vector borne diseases like dengue, encephalitis, West Nile virus, Chikungunya (CHIKV) and Zika virus (ZIKV) which fall into the category of emerging viral infections. Recently ZIKV has come into worldwide prominence because this virus, spread mainly by the *Aedes aegypti* mosquito, causes microcephaly in children whose mothers were exposed to the virus while they were pregnant. Our team has used mathematical sequence descriptors for the characterization of the genome [12] of ZIKV isolated from the African and Asian region as well as for the computational prediction of potential peptide vaccines [13] which could be effective against ZIKV. In the genomics approach [12], invariants of graphs of RNA sequences of the ZIKV strains were used to compare sequences. A new sequence, if that is different from the ones already known, can be visualized and detected from the plot of the sequences against the mathematical descriptors. In the vaccinomics area [13], conserved gene sequences are detected using calculated sequence descriptors and then those sequences are used as the starting point for the computational design of peptide vaccines with high potential for success. For further information on this

line of research, please see [14,15].

5 Discussion

In this article we have given a short overview of our research in the use of topological chemodescriptors and biodescriptors. In QSAR development we used a hierarchical approach which used more resource intensive chemodescriptors in a later stage of model development. It was found from various HiQSAR studies that in most cases topological indices, a combination of TS and TC descriptors, gave very good models. Addition of 3-D or QC descriptors after the use of TS+TC combination did very little improvement in model quality [1]. This is a good news because the era of big data has arrived [16]. Now one can develop QSAR models and use those in the screening of databases like Maybridge [17] which has tens of millions of structures/chemicals which are currently commercially available for procurement and laboratory testing. Models based on expensively calculated descriptors like QC indices or experimentally determined properties are not suitable for use or fast enough in such cases. Topological index based QSARs, if they work well, may help in solving such quagmires becoming a fast and cheap *in silico* alternative to the more expensive screening protocols.

The unintended consequences of the use of myriads of chemicals for various societal purposes is that some of them will be released to the environment and we will be exposed to them. Such chemicals are large in number and structurally very diverse. Yet we need to develop models whereby we can assess hazard posed by them because experimental testing of so many candidate chemicals is prohibitively costly and will need the sacrifice of a very large number of test organisms. In the area of protection of the human and ecological health from the toxic effects of such chemicals, QSARs based on easily calculated descriptors can play a useful role. We have shown that many properties needed for the prediction of fate and toxic effects of pollutants as well as their toxic modes of action (MOA) can be estimated well using easily calculated topological indices [1].

We developed novel DNA/RNA based biodescriptors and applied them for the characterization and classification of emerging pathogens like Bird Flu, Swine, Flu, Zika virus, etc. Sequence descriptors have also been used to zero in on "conserved" part of the genome of pathogens like ZIKV and utilize those in the computer-assisted design of peptide vaccines [13].

A few words about the practical utility of topological indices is needed at this juncture. Many scientists have expressed the opinion, mainly privately, that topological indices have very little practical utility even if they produce good predictive models in narrowly selected chemical classes. In this article we have shown that for important properties like mutagenicity and octanol-water partition coefficient, topological models had high predictive power. We give here a couple of successful examples of practical new drug discovery using topological indices. In the 1980s our topological calculation software POLLY [7] and quantitative molecular similarity analysis (QMSA) method [18] derived from principal components (PCs) calculated from 90 POLLY indices were installed at the Upjohn Co. (now part of Pfizer). They used this method – called the Basak method [19] – for the discovery of numerous new lead compounds for drugs. Another good example is from Garcia-Domenech et al. [20] who used topological indices to design and experimentally validate many different classes of drugs. From the above we can safely conclude that practical applications of topological models in chemoinformatics and bioinformatics have an exciting future.

References

- [1] S. C. Basak, *Curr Comput Aided Drug Des*, **9**, 449 (2013). PMID:24138422, DOI:10.2174/15734099113096660041
- [2] L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999
- [3] J. Devillers, Balaban, A.T. (Eds.) *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, 1999.
- [4] N. Trinajstić, *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
- [5] H. Wiener, *J. Am. Chem. Soc.*, **69**, 17 (1947). PMID:20291038, DOI:10.1021/ja01193a005
- [6] H. Hosoya, *Bull. Chem. Soc. Jpn.*, **44**, 2332 (1971). DOI:10.1246/bcsj.44.2332
- [7] Basak, S.C.; Harriss, D.K.; Magnuson, V.R. POLLY v. 2.3: 1988; Copyright of the University of Minnesota, USA.
- [8] MolconnZ, Version 4.05, 2003; Hall Ass. Consult.; Quincy, MA.
- [9] Basak, S.C.; Grunwald, G.; Balaban, A. T. TRIPLET; Copyright of the University of Minnesota, 1993.
- [10] S. C. Basak, S. Majumdar, *Curr. Comput. Aided Drug Des.*, **12**, 1 (2016).
- [11] G. J. Niemi, S. C. Basak, G. D. Veith, G. D. Grunwald, *Environ. Toxicol. Chem.*, **11**, 891 (1992).
- [12] A. Nandy, S. Dey, S. C. Basak, D. Bielińska-Wąż, P. Wąż, *Aided Drug Des. To Curr. Comput. -Aided Drug Des.*, **12**, 87 (2016). DOI:10.2174/1573409912666160401115812
- [13] A. Nandy, S. C. Basak, *Int. J. Mol. Sci.*, **17**, 666 (2016). DOI:10.3390/ijms17050666
- [14] A. Nandy, M. Harle, S. C. Basak, *ARKIVOC*, **9**, 211 (2006).
- [15] M. Randić, M. Vracko, A. Nandy, S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **40**, 1235 (2000). DOI:10.1021/ci000034q
- [16] S. C. Basak, M. Vracko, A. K. Bhattacharjee, *Curr Comput Aided Drug Des*, **11**, 197 (2015). PMID:26643819
- [17] Maybridge screening collections and fragment libraries for drug discovery. Available from: <http://www.maybridge.com/>. (Accessed 23 December, 2016).
- [18] S. C. Basak, V. R. Magnuson, G. J. Niemi, R. R. Regal, *Discrete Appl. Math.*, **19**, 17 (1988). DOI:10.1016/0166-218X(88)90004-2
- [19] M. Lajiness, In: *Computational Chemical Graph Theory*, Rouvray, D. H.; Ed.; Nova, New York, 1990, pp. 299–316.
- [20] R. Garcia-Domenech, R. Zanni, M. Galvez-Llompart, J. Galvez, *Mol. Divers.*, **19**, 357 (2015). DOI:10.1007/s11030-015-9575-5