10 Mathematical Chemodescriptors and Biodescriptors: Background and Their Applications in the Prediction of Bioactivity/Toxicity of Chemicals

Subhash C. Basak

10.1 Introduction

> At quite uncertain times and places,
> The atoms left their heavenly path,
> And by fortuitous embraces,
> Engendered all that being hath.
> And though they seem to cling together,
> And form 'associations' here,
> Yet, soon or late, they burst their tether,
> And through the depths of space career.
>
> – James Clerk Maxwell
> In: "Molecular Evolution," Nature, 8, 1873.
> In Lewis Campbell and William Garnett, The Life of James Clerk Maxwell (1882), 637

Many physiological, pathological, toxicological, and biomedicinal processes are determined by interactions of small molecules such as endogenous ligands, drugs, xenobiotics, and substrates as well as inhibitors of enzymes related to metabolic pathways with their appropriate biological targets. The maintenance of the integrity and continuity of such key ligand-biotarget interactions is critical for the smooth functioning of biological systems ranging from the single-celled organism to the complex ecosystems. A large number of drugs are small molecules that interact with specialized enzymes/receptors in appropriate physiological compartments and thereby produce effect(s) that bring a pathologically perturbed biological system back to a healthy state [1–4]. Biological properties of molecules, beneficial or deleterious, can be looked upon as the result of ligand-biotarget interactions and can be expressed by the relationship:

$$BR = f\left(S, B\right) \tag{10.1}$$

where *BR* represents the normal biological or pathological/toxicological response produced by the ligand (drug or toxicant) in the target biological system and *B* represents the relevant biochemical part of the target system which is perturbed by ligand to produce the measurable effect. It is believed that a major determinant of *BR* is the nature or structure (S) of the ligand. The structure becomes the sole determinant of the variation of the measured BR from one chemical to another when the biological system, *B*, remains practically the same during the course of the experiment and there is alternation only in the structure of the ligands. Eq. 10.1 under such a condition approximates to:

$$BR = f\left(S\right) \tag{10.2}$$

A lot of research conducted in drug discovery, toxicology, environmental sciences, and biochemistry follows the paradigm expressed in Eq. 10.2, and using this relationship researchers

S.C. Basak (✉)
International Society of Mathematical Chemistry, University of Minnesota Duluth-Natural Resources Research Institute, Duluth, MN, USA

Department of Chemistry and Biochemistry,
University of Minnesota Duluth,
5013 Miller Trunk Highway, Duluth,
MN 55811, USA
e-mail: sbasak@nrri.umnj.edu

© Springer India 2016

S. Singh (ed.), *Systems Biology Application in Synthetic Biology*,
DOI 10.1007/978-81-322-2809-7_10

attempt to decipher the effects as well as the modes and mechanism(s) of action of molecules on some selected biotargets, which are assumed not to change significantly during the course of the experiment.

When we embark on the characterization of *BR* based on chemical structure alone following Eq. 10.2, we really attempt to understand which characteristics of the chemical structure are recognized by the biomolecular target. What are the factors involved in recognition: molecular size, shape, chirality, stereo-electronic nature, or charge? Which ones are more important and which have a marginal impact on *BR*? This is often accomplished by the development of molecular descriptors, referred to by us as chemodescriptors, which quantify various aspects of molecular structure such as shape, size, symmetry, chirality, stereo-electronic nature, etc. using various mathematical techniques.

10.2 Mathematical Characterization of Structure: Molecules and Biomolecules

> Ostensibly there is color, ostensibly sweetness, ostensibly bitterness, but actually only atoms and the void.
>
> Galen
> In: Nature and the Greeks, Erwin Schrodinger, 1954

> In order to describe an aspect of holistic reality we have to ignore certain factors such that the remainder separates into facts. Inevitably, such a description is true only within the adopted partition of the world, that is, within the chosen context.
>
> Hans Primas
> Chemistry, Quantum Mechanics and Reductionism [5]

10.2.1 The Molecular Structure Conundrum: Simple Graph to Quantum Chemical Hamiltonians

The structure of an assembled entity is the pattern of relationship among its parts. Molecular structure can be looked upon as the representation of the relationship among its various constituents. The term *molecular structure* represents a set of nonequivalent and probably disjoint concepts [5]. There is no reason to believe that when we discuss diverse topics, e.g., chemical synthesis, reaction rates, spectroscopic transitions, chemical reaction mechanisms, and *ab initio* calculations, using the notion of molecular structure, the different meanings we attach to the single term “molecular structure” originate from the same fundamental concept [6, 7]. In the context of molecular science, the various concepts of molecular structure, e.g., classical valence bond representations, various chemical graph theoretic representations, ball and spoke model of a molecule, representation of a molecule by minimum energy conformation, and representation of chemical species by Hamiltonian operators, are model objects [8–15] derived through different abstractions of the same chemical reality. In each instance, the *equivalence class* (concept or model of molecular structure) is generated by selecting certain aspects while ignoring some unique properties of those actual entities. This explains the plurality of the concept of molecular structure and their autonomous nature, the word “autonomous” being used here in the same sense that one concept is not logically derived from the other [7].

10.2.2 The Philosophical Basis of Modeling in Mathematical Chemistry

The process of modeling arises out of abstraction from sense data derived from reality. As put forward by Albeit Einstein [8] in his remarks on the philosopher Bertrand Russell’s theory of knowledge:

> The more, however, we turn to the most primitive concepts of everyday life, the more difficult it becomes amidst the mass of inveterate habits to recognize the concept as an independent creation of thinking. It was thus that the fateful conception -fateful, that is to say, for an understanding of the here-existing conditions – could arise, according to which the concepts originate from experience by way of “abstraction,” i.e., through omission of a part of its content.

As pointed out by Basak [8] regarding the philosophy of modeling [9] of molecular structure:

> Any concept of molecular structure is a hypothetical sketch of the organization of molecules. Such a model object is a general theory and remains empirically untestable. A model object has to be grafted onto a specific theory to generate a theoretical model. A theoretical model of an object can be empirically tested. For example, when it was suggested by Sylvester [12] in 1878 that the structural formula of a molecule is a special kind of graph, it was an innovative general theory without any predictive potential. When the idea of combinatorics was applied on chemical graphs (model objects), it could be predicted that "there should be exactly two isomers of butane (C_4H_{10})" because "there are exactly two tree graphs with four verüces" when one considers only the non-hydrogen atoms present in C_4H_{10}. This is a theoretical model of limited predictive potential. Although it predicts the existence of chemical species, given a set of molecules, e.g. isomers of hexane (C_6H_{14}), the model is incapable of predicting any property. This is because of the fact that any empirical property P maps a set of chemical structures into the set R of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, we need a nonempirical (structural) ordering scheme which closely resembles the empirical ordering of structures as determined by P. This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariant(s).

10.2.3 Mathematical Chemodescriptors: Topological Indices, 3D Descriptors, and Quantum Chemical Indices

One of the important goals of structural chemistry, biomedicinal chemistry, and computational toxicology is the "optimal characterization" of molecular structure for the purpose of predicting their properties. As discussed in Sect. 10.2.1, optimal characterization of structure has remained elusive. Different groups of researchers have used different methods for the representation and quantification of molecular structure. In our quantitative structure-activity relationship (QSAR) and quantitative molecular similarity analysis (QMSA) research, we have used mainly three classes of descriptors for the quantification of structure, viz., (a) graph invariants defined on molecular graphs, also known as *topological indices*, (b) three-dimensional (3D) or geometrical descriptors, and (c) quantum chemical descriptors.

In our research, we have also used *atom pairs* (APs), which are fragment-based descriptors. The method of Carhart et al. [10] 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:

$$\text{<atom descriptor}_i\text{>} - \text{<separation>} - \text{<atom descriptor}_j\text{>}$$

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.

Graph theory was discovered by Euler [11] in 1736. Sylvester [12] in 1878 saw the clear-cut relationship between graph theory and molecular structure. He also commented on the connection between chemistry and mathematics in general, as evident from the following [13]:

> Chemistry has the same quickening and suggestive influence upon the algebraist as a visit to the Royal Academy, or the old masters may be supposed to have on a Browning or a Tennyson. Indeed it seems to me that an exact homology exists between painting and poetry on the one hand and modem chemistry and modem algebra on the other. In poetry and algebra we have the pure idea elaborated and expressed through the vehicle of language, in painting and chemistry the idea is enveloped in matter, depending in part on manual processes and the resources of art for its due manifestation.

Applications of graph theory to chemical problems are part of a fast developing field of science called mathematical chemistry or, more correctly, discrete mathematical chemistry. Although Sylvester [12] saw the connection between molecular structure and chemistry as back as 1878, modern research in chemical graph theory had its humble beginning at the middle of the twentieth century probably with the publication of the seminal paper by Harry Wiener [14] on the calculation of structural indices for the prediction of molecular properties. Invariants of graphs associated with molecules and biomolecules quantify certain aspects of their structure and have been used in the characterization and comparison of such structures as well as prediction of their properties Specifically, such invariants and orthogonal factors like *principal components* (PCs) derived from them have found applications in QSAR studies [15–18], QMSA research [18–22], clustering of large libraries of structures into smaller subsets [20, 21], and in the discrimination of pathological structures like isospectral graphs [15].

The author of this chapter (Basak) and his coworkers have been involved since the early 1970s in the development of novel numerical graph invariants or topological indices (TIs) [16–19, 22–26] as well as biodescriptors derived from DNA/RNA sequences [16, 27] and proteomics maps [28]. It may be mentioned here that graph theoretical numerical indices were called "topological indices" by Hosoya [29] for the first time in a paper published in 1971.

Many topological indices can be conveniently derived from various matrices including the *adjacency matrix A* (*G*) and the *distance matrix D* (*G*) of a *chemical graph G*. These matrices are usually constructed from labeled graphs of hydrogen-suppressed molecular skeletons. For details of theoretical basis and calculation of topological indices, see refs [17, 18, 23–29].

Basak et al. have divided the topological indices (TIs) into two major groups: topostructural (TS) indices and topochemical (TC) indices. TS indices are calculated from skeletal graph models of molecules which 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 indices quantify information regarding the connectivity, adjacency, and distances between vertices 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 and chemically informative as compared to the TS descriptors.

The geometrical or 3D parameters quantify the volume, size, and shape of molecules from various models. We have used van der Waals' volume as a measure of gross size of molecules. The three-dimensional Wiener indices calculated on the hydrogen-suppressed and hydrogen-filled graphs are also quantifiers of molecular shape and size. With respect to calculation of quantum chemical descriptors, we have used both the AM_1 semiempirical method and *ab initio* calculations based on the STO-3G, 6-31G(d), 6-311G, 6-311G(d), and aug-cc-pVTZ basis sets. For chemodescriptors used by Basak group in their studies, see [18, 29–35]. Table 10.1 gives the symbols and definition of molecular chemodescriptors.

10.2.4 Hierarchical Classification of Descriptors

The combination of topological, geometrical, and quantum chemical chemodescriptors, and biodescriptors (*vide infra*) derived from proteomics, genomics, and DNA sequence characterization, leads to a hierarchy of descriptors that begins with the simplest graph invariants and ends with the biodescriptors, which require expensive and time-intensive laboratory test data (Fig. 10.1). It should be clearly stated here that descriptors in the higher levels of the hierarchy are not necessarily superior to those placed at lower levels. The scheme simply shows a gradation based on the need for computational and laboratory resources.

The molecular descriptors itemized in Table 10.1 are calculated by Basak's team using Molconn-Z [30], POLLY [31], APProbe [32], and Triplet [33], MOPAC [34], and Gaussian [35].

10.3 Quantitative Structure-Activity Relationship (QSAR) Using Chemodescriptors

> Those alone are wise who act after investigation.
>
> Charaka
> In Sutrasthana, 10:5

> We haven't got the money, so we've got to think
>
> Ernest Rutherford

Table 10.1 Symbols, definitions, and classification of structural molecular descriptors

	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
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 rth ($r=0$–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth ($r=0$–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for rth ($r=0$–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 $h=0$–6
${}^h\chi^v_C$	Valence cluster connectivity index of order $h=3$–6

(continued)

Table 10.1 (continued)

	Topostructural (TS)
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order $h=3$–6
${}^{h}\chi^{v}_{PC}$	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
${}^{h}\chi^{v}$	Valence path connectivity index of order $h=7$–10
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order $h=7$–10
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	Wienerp
Pf	Plattf
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

(continued)

Table 10.1 (continued)

	Topostructural (TS)
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 CH 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)
$SHBint_{y}$	E-State descriptors of potential internal hydrogen bond strength ($y=2$–10)
	Electrotopological state index values for atoms 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*, *SsssssP*, *SsSH*, *SdS*, *SssS*, *SaaS*, *SdssS*, *SddssS*, *SssssssS*, *SsCl*, *SsGeH3*, *SssGeH2*, *SsssGeH*, *SssssGe*, *SsAsH2*, *SssAsH*, *SsssAs*, *SdsssAs*, *SsssssAs*, *SsSeH*, *SdSe*, *SssSe*, *SaaSe*, *SdssSe*, *SddssSe*, *SsBr*, *SsSnH3*, *SssSnH2*, *SsssSnH*, *SssssSn*, *SsI*, *SsPbH3*, *SssPbH2*, *SsssPbH*, *SssssPb*
Geometrical (3D)/shape	
kp_{0}	Kappa zero

(continued)

Table 10.1 (continued)

	Topostructural (TS)
kp_1-kp_3	Kappa simple indices
ka_1-ka_3	Kappa alpha indices
V_W	Van der Waals volume
^{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
	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

Modern society routinely uses a large number of natural and man-made chemicals in the form of drugs, solvents, synthetic intermediates, cosmetics, herbicides, pesticides, etc. to maintain the lifestyle. But in many cases, a large fraction of these chemicals do not have the experimental data necessary for the prediction of their beneficial and deleterious effects [36]. Table 10.2 gives a partial list of properties, both physical and biochemical/pharmacological/toxicological, needed for the effective screening of chemicals for new drug discovery and protection of human as well as ecological health. Because determination of such properties for so many chemicals in the laboratory is prohibitively costly, one solution of this quagmire has been the use of QSARs and molecular similarity-based analogs to obtain acceptable estimated values of properties.

10.3.1 Statistical Methods for QSAR Model Development and Validation

> In God we trust. All others must bring data.
>
> W. Edwards Deming

> To call in the statistician after the experiment is done maybe no more
> than asking him to perform a post-mortem examination:
> he may be able to say what the experiment died of.
>
> Ronald Fisher: http://www.brainyquote.com/quotes/authors/r/ronald_fisher.html

In the early 1970s, when this author (Basak) started carrying out research on the development and use of calculated chemodescriptors in QSAR, only a few such descriptors were available. But now, with the availability of various software [30–35, 37, 38], the landscape of availability and calculation of molecular descriptors is very different. The four major pillars [18] of a useful QSAR system development are:

(a) Availability of high-quality experimental data (veracity of dependent variable)
(b) Data on sufficient number of compounds (volume or reasonably good sample size)
(c) Availability of relevant descriptors (independent variables of QSAR) which quantify aspects of molecular structure relevant to the activity/toxicity of interest
(d) Use of appropriate methods for model building and validation

The various pathways for the development of structure-activity relationship (SAR) and property-activity relationship (PAR) models either from calculated molecular descriptors or from experimentally determined as well as calculated properties as independent variables may be expressed by the scheme provided in Fig. 10.2.

The use of computed molecular descriptors and experimental property data in PAR/SAR/QSAR may be illuminated through a formal exposition of the structure-property similarity principle – the central paradigm of the field of SAR [39]. Figure 10.2 depicts the determination of an experimental property, e.g., measurement of octanol-water partition coefficient of a chemical in the laboratory, as a function α: C → R which maps the set C of compounds into the real line R. A nonempirical QSAR may be looked upon as a composition of a description function β_1: C → D mapping each chemical structure of C

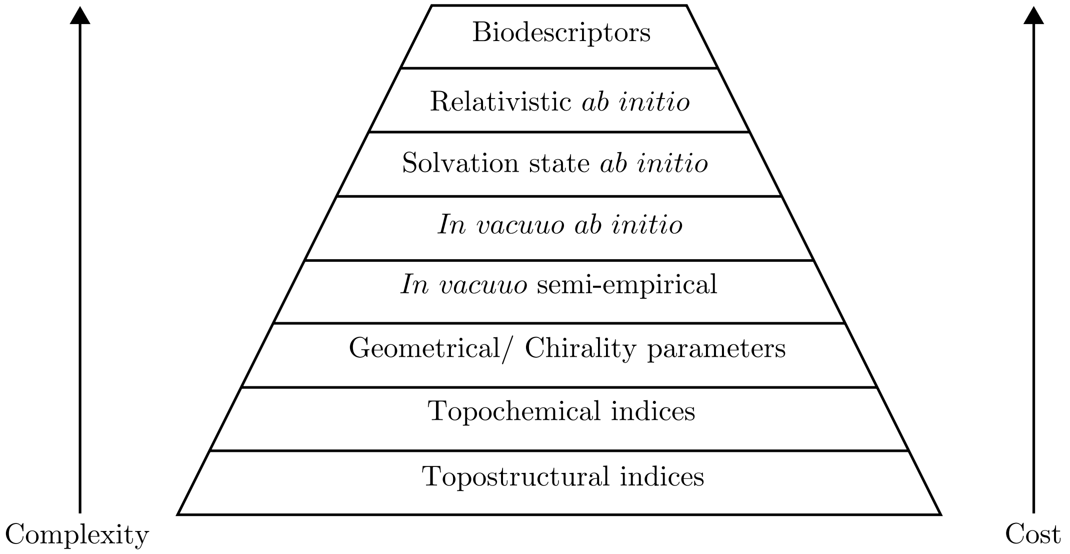


Fig. 10.1 Hierarchical classification of chemodescriptors and biodescriptors used in QSAR (Source: Basak [18]. With permission from Bentham Science Publishers)

Table 10.2 List of properties needed for screening of chemicals

Physicochemical	Pharmacological/toxicological
Molar volume	Macromolecular level
Boiling point	Receptor binding (K_d)
Melting point	Michaelis constant (K_m)
Vapor pressure	Inhibitor constant (K_i)
Water solubility	DNA alkylation
Dissociation constant (pKa)	Unscheduled DNA synthesis
Partition coefficient	Cell level
Octanol-water (log P)	Salmonella mutagenicity
Air-water	Mammalian cell transformation
Sediment-water	Organism level (acute)
Reactivity (electrophilicity)	LD_{50} (mouse, rat)
	LC_{50} (fathead minnow)
	Organism level (chronic)
	Bioconcentration factor
	Carcinogenicity
	Reproductive toxicity
	Delayed neurotoxicity
	Biodegradation

into a space of nonempirical structural descriptors (D) and a prediction function β_2: D → R which maps the descriptors into the real line. One example can be the use of Molconn-Z [30] indices for the development of QSARs. When [$\alpha(C) - \beta_2 \circ \beta_1 (C)$] is within the range of experimental errors, we say that we have a good QSAR model.

On the other hand, PAR is the composition of θ_1: C → M which maps the set C into the molecular property space M and θ_2: M → R mapping those molecular properties into the real line R. Property-activity relationship seeks to predict one property (usually a complex physicochemical property) or bioactivity of a molecule in terms of other (usu-

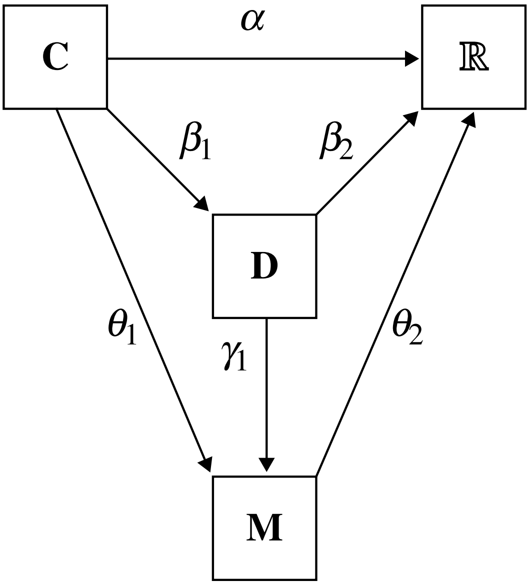


Fig. 10.2 Composition functions of various mappings for structure-activity relationship (SAR) and property-activity relationship (PAR) (Source: Basak and Majumdar [46]. With permission from Bentham Science Publishers)

ally simpler or easily determined experimentally) properties.

Basak group uses the following generic method in the validation of QSAR models: In the process of formulating a scientifically interpretable and technically sound QSAR model, we need to keep in mind some important issues. First and foremost, one has to check whether a specific method is the best technique in modeling a specific QSAR scenario. In a regression set up, for example, when the number of independent variables or descriptors (p) is much larger than the number of data points (dependent variable, *n*), i.e., $p>>n$, the estimate of the coefficient vector is nonunique. This is also the case when predictors in the study are highly correlated with one another to the extent that the "design matrix" is rank-deficient. Both of these factors are relevant to QSARs. In many contemporary QSAR studies, the number of initial predictors typically is in the range of hundreds or thousands, whereas more often than not, mostly to keep cost of generation of experimental data under control, the experimenter can collect data on only a much smaller number (tens or hundreds) of samples. This effectively makes the problem high dimensional and rank-deficient ($p>>n$) in nature. Also, when a large number of descriptors on a set of chemicals are used to model their activity, one should expect that some predictors within a single class, e.g., TC descriptors, or even predictors belonging to apparently different classes are highly correlated with one another. Such situations can be tackled either by attempting to pick important variables through model selection or "sparsity"-type approaches (e.g., forward selection, LASSO [40], adaptive LASSO [41]), or finding a lower-dimensional transformation that preserves most of the information present in the set of descriptors, e.g., principal component analysis (PCA) and envelope methods [42].

We need to check the ability of a model to give competent predictions on "similar" data sets via validation on out-of-sample test sets. For a relatively small sample, i.e., a small set of compounds, this is achieved by carrying out a **leave-one-out (LOO) cross-validation**. For data sets with a large number of compounds, a more computationally economical way is to do a **k-fold cross-validation**: split the data set randomly into k (previously decided by the researcher) equal subsets, take each subset in turn as test set, and use the remaining compounds as training sets and use the model to obtain predictions. Comparing cross-validation with the somewhat prevalent approach in QSAR research of **external validation**, i.e., choosing a single train-test split of compounds, it should be pointed out that in external validation, the splits of data sets are carried out only once using the experimenters' *a priori* knowledge or some subjectively chosen ad hoc criterion. But in cross-validation, the splits are chosen randomly, thus providing a more unbiased estimate of the generalizability of the QSAR model. Furthermore, Hawkins et al. [43] proved theoretically that compared to external validation, LOO cross-validation is a better estimator of the actual predictive ability of a statistical model for small data sets, while for large sample size both perform equally well. To quote Hawkins et al. [43], "The bottom line is that in the typical QSAR setting where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly." Specific drawbacks of holding out only one test set in the external validation method include: (1) structural features of the held out chemicals are not included in the modeling process, resulting in a loss of information; (2) predictions are made on only a subset of the available compounds, whereas the LOO method predicts the activity value for all compounds; (3) there is no scientific tool that can guarantee similarity between chemicals in the training and test sets; and (4) personal bias can easily be introduced in selection of the external test set.

In the rank-deficient situation of QSAR formulation, special care should be taken in combining conventional modeling with the additional step of variable selection or dimension reduction. An intuitive, but frequently misunderstood and wrong, procedure would be to perform the first stage of preprocessing first, selecting important variables or determining the optimal transformation, and then use the transformed data/selected variables to build the predictive QSAR models and obtain predictions for each train-test split. The reason why this is not appropriate is that the data is split only after the variable selection/dimension reduction step is already completed. Essentially this method ends up using information from the holdout compound/split subset to predict activity of those very samples. This *naïve cross-validation* procedure causes synthetic inflation of the cross-validated q^2, hence compromises the predictive ability of the model [44, 45] (Fig. 10.3). A two-step procedure (referred in Fig. 10.3 as *two-deep CV*) helps avoid this tricky situation. Instead of doing the pre-model building step first and then taking multiple splits for out-of-sample prediction, for each split of the data the initial steps are performed only using the training set of compounds each time. Since calculations on two different splits are not dependent on each other, for large data sets the increased computational demand arising out of the repeated variable selection can be tackled using substantial computer resources like parallel processing. It should be emphasized that the naïve cross-validation (naïve CV) method gives **naïve or wrong q^2** values, whereas the two-deep cross-validation (two-deep CV) approach gives us the correct or **true q^2**.

For recent reviews and research on this topic of proper cross-validation, please see the recent publications of Basak and coworkers [46–52].

The quality of the model, in terms of its predictive ability, is evaluated based on the associated q^2 value, which is defined as:

$$q^2 = 1 - (\text{PRESS} / \text{SSTotal}) \qquad (10.3)$$

where PRESS is the prediction sum of squares and SSTotal is the total sum of squares. Unlike R^2 which tends to increase upon the addition of any descriptor, q^2 will decrease upon the addition of irrelevant descriptors, thereby providing a reliable measure of model quality.

In order to illustrate practically the inflation of q^2 associated with the use of improper statistical techniques, we deliberately developed a wrong model using stepwise ordinary least squares (OLS) regression, which is commonly used in many QSAR studies but often results in overfitting and renders the model unreliable for making predictions for chemicals similar to those used to calibrate the model. The REG procedure of the SAS statistical package [53] was used to develop stepwise regression model. For details see [45]. Rat fat/air partition coefficient values for a diverse set of 99 organic compounds were used for this study. It should be noted that two compounds with fewer than three non-hydrogen atoms, for which we could not calculate our entire suite of structure-based descriptors, were omitted from our study. A total of 375 descriptors were calculated using software packages including POLLY v2.3, Triplet, Molconn-Z v 3.5, and Gaussian 03W v6.0. This is clearly a rank-deficient case with the number of compounds ($n=97$) being much smaller than the number of predictors ($p=375$). The ridge regression (RR) approach [45, 51] in which the Gram-Schmidt algorithm was used to properly thin the descriptors yielded a four-parameter model with an associated q^2 of 0.854. Each of the four descriptors was topological in nature; none of the three-dimensional or quantum chemical descriptors were selected. An inflated q^2 of 0.955 was

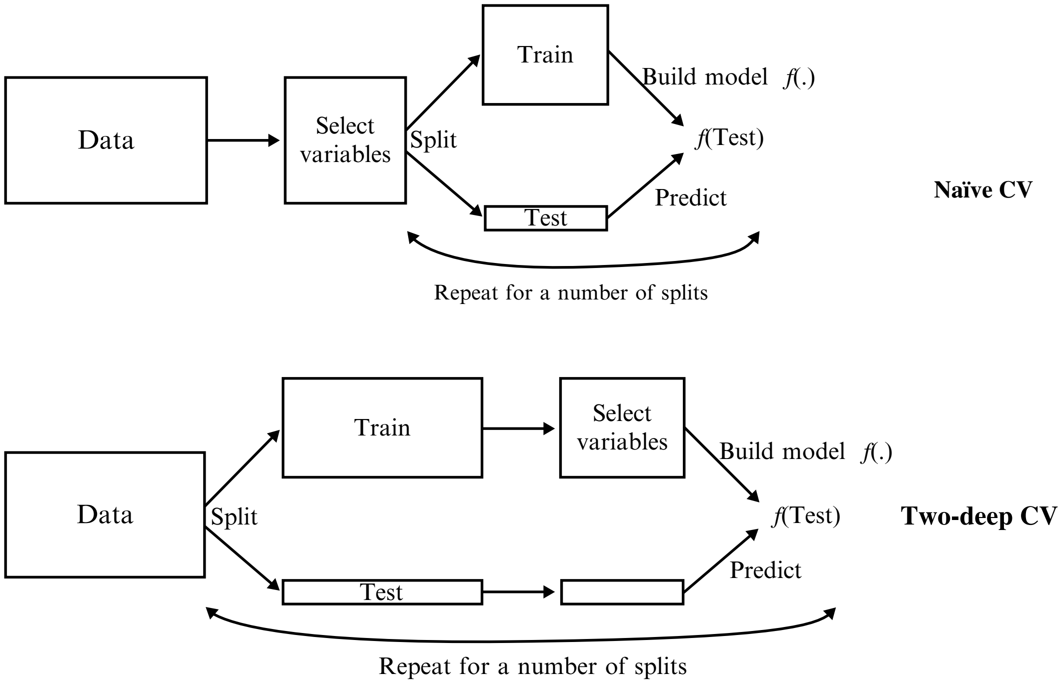


Fig. 10.3 Difference between naïve and two-deep cross-validation (CV) schemes (Source: Basak and Majumdar [46]. With permission from Bentham Science Publishers)

obtained from the stepwise regression approach which yielded a 24-parameter model.

10.3.2 Intrinsic Dimensionality of Descriptor Spaces: Use of Principal Component Analysis (PCA) as the Parsimony Principle or Occam's Razor

> शैले शैले न माणिक्यं मौक्तिकं न गजे गजे ।
> साधवो न हि सर्वत्र चंदनं न वने वने ॥
>
> shaile shaile na maanikyam mauktikam na gaje gaje
> saadhavo naahi sarvatra chandanam na vane vane
> (In Sanskrit)
>
> Not all mountains contain gems in them, nor does every elephant has pearl in it, noble people are not found everywhere, nor is sandalwood found in every forest.
>
> Chanakya
>
> You gave too much rein to your imagination. Imagination is a good servant, and a bad master. The simplest explanation is always the most likely.
> – Agatha Christie

As discussed earlier, these days we can calculate a large number of molecular descriptors using the available software. But **all descriptors are not created equal and each descriptor is not needed for all modeling situations**. In the QSAR scenario, we need to use proper methods for the selection of relevant descriptors. Methods like principal component analysis (PCA) [19, 54, 55] and interrelated two-way clustering (ITC) [56] can be used for variable selection or descriptor thinning.

When *p* molecular descriptors are calculated for *n* molecules, the data set can be viewed as *n* vectors in *p* dimensions, each chemical being represented as a point in R^p. Because many of the descriptors are strongly correlated, the *n* points in R^p will lie on a subspace of dimension lower than p. Methods like principal component analysis can be used to characterize the *intrinsic dimensionality* of chemical spaces. Since the early 1980s, Basak and coworkers have carried out PCA of various congeneric and diverse data sets relevant to new drug discovery and predictive toxicology. Principal components (PCs) derived from mathematical chemodescriptors have been used in the formulation of quantitative structure-activity relationships (QSARs), clustering of large combinatorial libraries, as well as quantitative molecular similarity analysis (QMSA), the last one to be discussed later. This section of the article will discuss PCA studies on characterization and visualization of chemical spaces of two data sets, one congeneric and one structurally diverse: (1) a large and structurally diverse set of 3692 chemicals which was a subset of the Toxic Substances Control Act (TSCA) Inventory maintained by the US Environmental Protection Agency (USEPA) and (2) a virtual library of 248,832 psoralen derivatives,

In the early 1980s, after Basak joined the University of Minnesota Duluth, the software POLLY [31] was developed and large-scale calculation of TIs for QSAR and QMSA analyses was initiated. In one of the earliest studies of its kind, Basak et al. [19, 57] used the first version of POLLY for the calculation of 90 TIs for a collection of 3692 structurally diverse chemicals which was a subset of the Toxic Substances Control Act (TSCA) Inventory of USEPA. The authors carried out PCA on this data set and asked the question: **What is the intrinsic dimensionality of chemical structure measured by the large number of TIs**? As shown in the summary in Table 10.3, first ten PCs with eigenvalues greater than or equal to 1.0 explained 92.6 % of the variance in the data of the calculated descriptors, and first four PCs explained 78.3 % of the variance [19, 57]. For a recent review of our research in this line, see Basak et al. [58].

It is clear from the data in Table 10.3 that PC_1 is strongly correlated with those indices which are related to the size of chemicals. It is noteworthy that for the set of 3692 diverse chemicals PC_1 was also highly correlated with molecular weight ($r = 0.81$) and K_0 (0.95) which is the number of vertices in hydrogen-suppressed graphs. PC_2 was interpreted by us as an axis of molecular complexity as encoded by the higher-order information theoretic indices developed by Basak group [23, 59]. PC_3 is most highly related to the cluster/path-cluster-type molecular connectivity indices which quantify structural aspects regarding molecular branching. The data in Table 10.3 clearly show that PC_4 is strongly correlated with the cyclicity terms of the connectivity class of topological indices [19].

Table 10.3 Correlation of the first four PCs with the original variables in the 90 topological indices, [19, 57]

PC_1	PC_2	PC_3	PC_4
K_1 (0.96)	SIC_3 (0.97)	${}^{4}\chi^{b}{}_{C}$ (0.69)	${}^{4}\chi_{CH}$ (0.85)
${}^{2}\chi$ (0.95)	CIC_4 (−0.96)	${}^{4}\chi^{b}{}_{C}$ (0.69)	${}^{4}\chi^{b}{}_{CH}$ (0.84)
${}^{3}\chi$ (0.95)	CIC_3 (−0.95)	${}^{5}\chi^{b}{}_{C}$ (0.68)	${}^{4}\chi^{v}{}_{CH}$ (0.80)
K_2 (0.95)	SIC_4 (0.95)	${}^{4}\chi_{C}$ (0.68)	${}^{3}\chi_{CH}$ (0.75)
K_0 (0.95)	SIC_2 (0.94)	$3\chi v_{C}$ (0.67)	${}^{3}\chi^{b}{}_{CH}$ (0.75)
${}^{1}\chi$ (0.94)	CIC_5 (−0.94)	${}^{5}\chi_{C}$ (0.64)	${}^{4}\chi^{b}{}_{CH}$ (0.74)
${}^{3}\chi^{b}$ (0.94)	CIC_6 (−0.92)	${}^{6}\chi_{C}$ (0.64)	${}^{3}\chi^{v}{}_{CH}$ (0.72)
${}^{4}\chi$ (0.94)	SIC_5 (0.92)	${}^{3}\chi_{C}$ (0.61)	${}^{5}\chi_{CH}$ (0.71)
${}^{4}\chi^{b}$ (0.93)	SIC_6 (0.89)	${}^{6}\chi^{b}{}_{C}$ (0.60)	${}^{5}\chi^{v}{}_{CH}$ (0.67)
${}^{0}\chi$ (0.93)	CIC_2 (−0.87)	${}^{5}\chi^{v}{}_{C}$ (0.60)	${}^{6}\chi^{b}{}_{CH}$ (0.47)

The symbols and definitions of the indices shown in this Table can be found in Table 10.1. The bonding connectivity indices were defined for the first time by Basak et al. [19]

Some of the TIs used in this study, e.g., Randic's [60] first-order connectivity index (${}^{1}\chi$) and the information theoretic indices developed by Bonchev and Trinajstić [61] and Raychaudhury et al. [24], were used to discriminate the set of congeneric structures including alkanes. In the case of 18 octanes, the molecules do not vary much from one another with respect to size, but primarily in terms of branching patterns. Therefore, these indices were rightly interpreted based on those data as reflecting molecular branching. But when PCA was carried out with a diverse set of 3692 chemical structures, the *results entered an uncharted territory and were counterintuitive, to say the least*. As shown from the correlation of the original variables with PC_1, ${}^{1}\chi$ and related indices were now strongly correlated with molecular size in the large and diverse set, not to molecular branching. PC_3 emerged as the axis correlated with indices that encoded branching information, the cluster-type molecular connectivity indices in particular. *This result shows that the structural meaning of TIs that we derive intuitively or from correlational analyses is dependent on the nature and relative diversity of the structural landscape under investigation*. Further studies of TIs computed for both congeneric and diverse structures are needed to shed light on this important issue.

A virtual library of 248,832 psoralen derivatives [21] was created and analyzed using PCs derived from calculated TIs. *This set may be called congeneric because although it is a large collection of structures, it is derived from the same basic molecular skeleton*: *psoralen*. For this study, 92 topological indices were calculated by POLLY. In this set, the top 3 PCs explained 89.2 % of the variance in the data; first 6 PCs explained 95.5 % of the variance of the originally calculated indices. The PCs were used to cluster the large set of chemicals into a few smaller subsets as an exercise of managing *combinatorial explosion* that can happen in the drug design scenario when one wants to create a large pool of derivatives of a lead compound. For details of the outcome of clustering of the 248,832 psoralen derivatives, please see [21].

To conclude this section on the exploration of intrinsic dimensionality of structural spaces using PCA and calculated chemodescriptors, the data on the congeneric set of psoralens and the diverse set of 3, 692 TSCA chemicals appear to indicate that as compared to congeneric collections of structures, diverse sets need a higher number of orthogonal descriptors (dimensions) to explain a comparable amount of variance in the data. The fact that PCA brings down the number of descriptors from 90 or 92 calculated indices to 10 or 6 PCs keeping the explained variance at above 90 % level reflects that the intrinsic dimensionality of the structure space is adequately reflected by a small number of orthogonal variables. Thinking in terms of the philosophical idea known as the **Ockham's razor or the parsimony principle** – **it is futile to do with more what can be done with fewer** – PCA helps us to select a *useful and smaller subset of factors from a collection of many more*. To quote Hoffmann et al. [62]:

> Identifying the number of significant components enables one to determine the number of real sources of variation within the data. The most important applications of PCA are those related to: (a) classification of objects into groups by quantifying their similarity on the basis of the Principal Component scores; (b) interpretation of observables in terms of Principal Components or their combination; (c) prediction of properties for unknown samples. These are exactly the objectives pursued by any logical analysis, and the Principal Components may be thought of as the true independent variables or distinct hypotheses.

It is noteworthy that Katritzky et al. used PCA for the characterization of aromaticity [63] and formulation of QSARs [64] in line with the parsimony principle.

10.3.3 Some Examples of Hierarchical QSAR (HiQSAR) Using Calculated Chemodescriptors

10.3.3.1 Aryl Hydrocarbon (Ah) Receptor Binding Affinity of Dibenzofurans

Dibenzofurans are widespread environmental contaminants that are produced mainly as undesirable by-products in natural and industrial processes. The toxic effects of these compounds are thought to be mediated through binding to the aryl hydrocarbon (*Ah*) receptor. We developed HiQSAR models based on a set of 32 dibenzofurans with *Ah* receptor binding affinity values obtained from the literature [65]. Descriptor classes used to develop the models included the TS, TC, 3D, and the STO-3G class of *ab initio* QC descriptors. Statistical metrics for the ridge regression (RR), partial least square (PLS), and principal component regression (PCR) models are provided in Table 10.4. We found that the RR models were superior to those developed using either PLS or PCR. Examining the RR metrics, it is evident that the TC and the TS + TC descriptors provide high-quality predictive models, with R^2_{cv} values of 0.820 and 0.852, respectively. The addition of the 3D and STO-3G descriptors does not result in significant improvement in model quality. When each of these classes viz., 3-D and STO-3G quantum chemical descriptors, is used alone, the results are quite poor. This indicates that the topological indices are capable of adequately representing those structural features which are relevant to the binding of dibenzofu-

Table 10.4 Summary statistics for predictive *Ah* receptor binding affinity models

	$R^2_{\text{c.v.}}$			PRESS		
Independent variables	RR	PCR	PLS	RR	PCR	PLS
TS	0.731	0.690	0.701	16.9	19.4	18.7
TS+TC	0.852	0.683	0.836	9.27	19.9	10.3
TS+TC+3D	0.852	0.683	0.837	9.27	19.9	10.2
TS+TC+ 3D + STO-3G	0.862	0.595	0.862	8.62	25.4	8.67
TS	0.731	0.690	0.701	16.9	19.4	18.7
TC	0.820	0.694	0.749	11.3	19.1	15.7
3D	0.508	0.523	0.419	30.8	29.9	36.4
STO-3G	0.544	0.458	0.501	28.6	33.9	31.3

rans to the *Ah* receptor. Comparison of the experimentally determined binding affinity values and those predicted using the TS + TC RR model is available in Table 10.5. The details of this QSAR analysis has been published [66].

10.3.3.2 HiQSAR Modeling of a Diverse Set of 508 Chemical Mutagens

TS, TC, 3D, and QC descriptors for 508 chemical were calculated, and QSARs were formulated hierarchically using these four types of descriptors. For details of calculations and model building, see [67]. The method interrelated two-way clustering, ITC [56], which falls in the unsupervised class of approaches [68], was used for variable selection. Table 10.6 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.

Recent review of results of HiQSARs carried out by Basak and coworkers [46, 69–71] using topostructural, topochemical, 3-D, and quantum chemical indices for diverse properties, e. g., acute toxicity of benzene derivatives, dermal penetration of polycyclic aromatic hydrocarbons (PAHs), mutagenicity of a congeneric set of amines (heteroaromatic and aromatic), and others, indicates that in most of the above mentioned cases, TS+ TC combination of indices gives reasonable predictive models. The addition of 3-D and quantum chemical indices after the use of TS and TC descriptors did very little improvement in model quality.

How do we explain the above trend in HiQSAR? *One plausible explanation is that for the recognition of a receptor*, e.g., *the interaction of dibenzofuran with Ah receptor*, *discussed in* Sect. 10.3.3.1, *the dibenzofuran derivatives probably need some specific geometrical and stereo-electronic factors or a specific pharmacophore. But once the minimal requirement of this recognition is present in the molecule*, *the alterations in bioactivities from one derivative to another in the same structural class are governed by more general structural features which are quantified reasonably well by the TS and TC indices derived from the conventional bonding topology of molecules and features like sigma bond*, *π bond*, *lone pair of electrons*, *hydrogen bond donor acidity*, *hydrogen bond acceptor basicity*, etc. More studies with different groups of molecules with diverse bioactivities are needed to validate or falsify this hypothesis in line with the falsifiability principle of Sir Karl Popper [72], a basic scientific paradigm in the philosophy of science which defines the inherent testability of any scientific hypothesis.

Table 10.5 Experimental and cross-validated predicted *Ah* receptor binding affinities, based on the TS + TC ridge regression model of Table 10.4

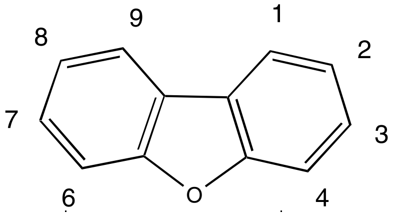
No.	Chemical	Experimental pEC_{50}	Predicted pEC_{50}	Exp. – Pred.
				
1	2-Cl	3.553	3.169	0.384
2	3-Cl	4.377	4.199	0.178
3	4-Cl	3.000	3.692	–0.692
4	2,3-diCl	5.326	4.964	0.362
5	2,6-diCl	3.609	4.279	–0.670
6	2,8-diCl	3.590	4.251	–0.661
7	1,2,7-trCl	6.347	5.646	0.701
8	1,3,6-trCl	5.357	4.705	0.652
9	1,3,8-trCl	4.071	5.330	–1.259
10	2,3,8-trCl	6.000	6.394	–0.394
11	1,2,3,6-teCl	6.456	6.480	–0.024
12	1,2,3,7-teCl	6.959	7.066	–0.107
13	1,2,4,8-teCl	5.000	4.715	0.285
14	2,3,4,6-teCl	6.456	7.321	–0.865
15	2,3,4,7-teCl	7.602	7.496	0.106
16	2,3,4,8-teCl	6.699	6.976	–0.277
17	2,3,6,8-teCl	6.658	6.008	0.650
18	2,3,7,8-teCl	7.387	7.139	0.248
19	1,2,3,4,8-peCl	6.921	6.293	0.628
20	1,2,3,7,8-peCl	7.128	7.213	–0.085
21	1,2,3,7,9-peCl	6.398	5.724	0.674
22	1,2,4,6,7-peCl	7.169	6.135	1.035
23	1,2,4,7,8-peCl	5.886	6.607	–0.720
24	1,2,4,7,9-peCl	4.699	4.937	–0.238
25	1,3,4,7,8-peCl	6.699	6.513	0.186
26	2,3,4,7,8-peCl	7.824	7.479	0.345
27	2,3,4,7,9-peCl	6.699	6.509	0.190
28	1,2,3,4,7,8-heCl	6.638	6.802	–0.164
29	1,2,3,6,7,8-heCl	6.569	7.124	–0.555
30	1,2,4,6,7,8-heCl	5.081	5.672	–0.591
31	2,3,4,6,7,8-heCl	7.328	7.019	0.309
32	Dibenzofuran	3.000	2.765	0.235

Table 10.6 HiQSAR model (RR and ITC + RR) for a diverse set of 508 chemical 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
ITC	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
	All four	307	72.05	76.17	67.86

10.3.4 Two QSAR Paradigms: Congenericity Principle Versus Diversity Begets Diversity Principle Analyzed Using Computed Mathematical Chemodescriptors of Homogeneous and Diverse Sets of Chemical Mutagens

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 statistically 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 TS, TC, geometrical, and QC indices. Whereas for the homogeneous amines, a small group of only seven descriptors were found to be significant in model building, for the 508 diverse set 42 descriptors were found to be statistically significant [73]. This preliminary and empirical study supports the ***DIVERSITY BEGETS DIVERSITY*** principle of QSAR formulated for the first time by Basak [18].

10.3.5 Applicability Domain of QSAR Models

A very important issue in the development of a QSAR model is that of defining the applicability domain (AD) of the model. This is necessary for any valid implementable QSAR model according to OECD principles [74]. There are a few methods of defining the AD of statistical models which can be roughly divided into two classes: (a) AD methods that define the active predictor space through some method like bounding box, PCA, or convex hulls and (b) distance-based methods which compute the similarity/dissimilarity of a new compound to the set of compounds which have been used in formulating the training QSAR model. To obtain predictions for any incoming sample set using the model, the first group of methods is used to ensure that the compounds are within the so-called active subspace: which essentially means we are actually performing interpolation, not extrapolation [75, 76]. For the distance-based approach, a predefined statistic is calculated to quantify the proximity of the test compounds to the training set, and based on whether that statistic is above or below a certain cutoff value, predictions for that compound are considered reasonable or not [75, 77].

10.3.6 Practical Applications of QSAR

> Knowledge is of no value unless you put it into practice.
>
> Anton Chekhov

Practical applications of good quality QSARs, particularly those based on easily calculable molecular descriptors, can be very useful tools in pharmaceutical drug design and specialty chemical design.

The journey of identified lead molecules in the drug discovery pipeline is a long and risky one. Average cost of developing a drug (including the cost of failures) during 2000s to early 2010s was US $2.6 billion [78]. One important contributing factor to this astronomical cost is that the drug developer has to produce and test a large number of derivatives of the lead structure for their beneficial and toxic side effects before one marketable drug is found. QSAR plays a very important role in drug design providing a cheaper and fast alternative to the medium throughput *in vitro* and low throughput *in vivo* screening of chemicals, which are generally used more frequently in the later stages of the discovery cascade. It has been noted that currently no drug is developed without going through the prior evaluation by QSAR methods [79].

In Fig. 10.4, a generic scheme is presented for the use of QSAR in drug discovery. Starting with a “lead,” modern combinatorial chemistry can produce millions, even billions, of derivatives. Such real or hypothetical chemicals must be evaluated in real time to prioritize them for synthesis and testing. QSARs based on easily calculated descriptors can help us in accomplishing this task.

The era of “Big Data” has arrived in the realm of drug discovery. For a concise description of trends in this realm, please see Basak et al. [80].

10.4 Molecular Similarity and Tailored Similarity Methods

> Like substances react similarly and similar changes in structure produce similar changes in reactivity
>
> L P. Hammett

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

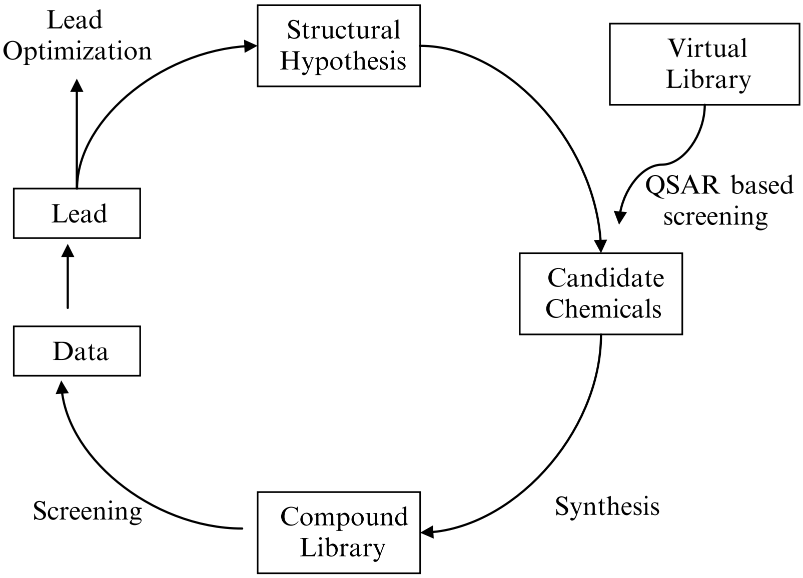


Fig. 10.4 A generic scheme for the use of QSARs in drug discovery protocols

Molecular similarity is a well-known concept, which is intuitively understood by many researchers. There is a tacit consensus among molecular similarity researchers that *similar structures usually have similar properties*. In a broader scope, this "*structure-property similarity principle*" includes the notion that similar "structural organizations" of objects lead to similar observable properties. In the realms of chemistry, biology, and toxicology, the natural extension of this structure-property similarity principle is that atoms, ions, molecules, and macromolecules with similar structures will have similar physicochemical, biological, and toxicological properties. This principle is vindicated by a vast majority of facts at varying levels of structural organization.

In the realm of cellular biochemistry, the inhibition of succinic dehydrogenase by malonate *in vitro* is explained in terms of the competition by malonate for the active sites of the enzyme succinic dehydrogenase, arising from the structural similarity between the substrate succinic acid and malonic acid [81, 82]. This is probably one of the earliest observations of the inhibition of an enzyme by an analog of its substrate. Another well-known example is that the structural similarities between p-amino benzoic acid and sulfanilic acid allow both compounds to interact with a specific bacterial biosynthetic enzyme. This "case of mistaken identity" is the basis for the antibacterial activity of sulfonamide antimicrobials [1].

There is no consensus regarding the optimal quantification of molecular similarity. In most cases, measures of molecular similarity are defined by the individual practitioner, generally based on his/her experience in a particular research area or some intuitive notion. If the researcher selects n different attributes for the molecules under investigation, then the molecules can be looked upon as points in some type of n-dimensional space. A distance function can then be used to measure the distance between various objects (chemicals) in that space, and the magnitude of distance serves as a measure of the degree of similarity or dissimilarity between any pair of molecules in this n-dimensional similarity space. Difficulties arise from two major factors: (1) the selection of appropriate axes for developing the similarity space and (2) the relevance of the selected axes to the property under investigation. Many molecular similarity scientists have their own favorite measures, but the axes selected might be multicollinear or may encode essentially the same information multiple times. One popular solution for this problem is the use of orthogonal axes derived from the original axes using techniques such as PCA mentioned above. A more serious concern is whether or not the subjectively chosen axes are relevant to the property under investigation. This is a more difficult problem to address. One potential solution to this issue, pursued by our research group, is the use of the *tailored similarity* method (*vide infra*).

One practical application of molecular similarity in pharmaceutical drug design, human health hazard assessment, and environmental risk analysis is the selection of analogs. Once a lead structure with interesting properties is found, the drug designer often asks "*Is there a chemical similar in structure to the lead, which also has analogous properties*?" In contemporary drug discovery research, scientists usually search various proprietary and public domain databases for chemical analogs. Analogs can be selected based on the researcher's intuitive notion of chemical similarity, their similarity with respect to measured properties, or calculated molecular descriptors. Since most of the chemicals in many databases have very little available experimental property data, similarity methods based on calculated properties or molecular descriptors are used more frequently for analog selection. In environmental risk analysis, analogs of suspected toxicants or newly produced industrial chemicals are used in hazard assessment when the molecule is so unique or so complex that class-specific QSARs cannot be applied in toxicity estimation [36]. The flip side of similarity is dissimilarity. This concept can be applied to both drug discovery and predictive toxicology to reduce the number of compounds in the database from a combinatorial explosion to a manageable number that can be handled through laboratory testing. One such example was discussed above in Sect. 10.3.2 for the case of a large virtual library of 248,832 psoralen derivatives which were clustered using PCs extracted from 92 computed POLLY indices.

10.4.1 Arbitrary or User-Defined Similarity Methods

In *arbitrary similarity methods*, one subjectively defines the similarity measure. In essence, the experienced practitioner says "*My personal experience with data or my intuitive notion tells me that the prescribed similarity measures will lead to useful grouping of chemicals with respect to the property of interest*." This might work out in narrowly defined cases, but in complex situations where a large number of parameters are needed to characterize the property, intuition is usually less accurate. Also, one may want to select analogs which are ordered with respect to widely different properties of the same chemical, e.g., carcinogenicity versus boiling point. The same intuitive measure cannot give "good analogs" for properties that are not mutually correlated. Various authors have used apparently diverse, arbitrary similarity measures in an effort to select mutually dissimilar analogs, but the rational basis of such selections has never been clear. The tailored approach to molecular similarity may help solve this issue.

10.4.1.1 Probing the Utility of Five Different Similarity Spaces

A wide variety of chemical information can and have been used in developing molecular similarity spaces. Many researchers contend that similarity spaces derived from physicochemical property data are inherently better, since the results are much more readily interpretable. However, as was stated earlier, physicochemical property data is not widely available for many chemicals, thus necessitating the use of calculated descriptors. One interesting aspect of research in the field of molecular similarity has been the comparison of arbitrary similarity spaces derived from physicochemical properties with spaces derived from calculated molecular descriptors. For a recent review on the topic of quantitative molecular similarity analysis studies carried out by Basak and coworkers, please see [22].

In a 1995 study, Basak and Grunwald [83] developed five distinct similarity spaces and tested those on a set of 73 aromatic and heteroaromatic amines with known mutagenicity (ln Rev/nmol) data. The derived similarity spaces were based on quantum theoretical descriptors believed to correlate well with mutagenicity (property), principal components derived from those descriptors (PC_{Prop}), atom pairs (APs), principal components derived from a set of topological indices (PC_{TI}), and principal components derived from the combined set of quantum theoretical descriptors and topological indices (PC_{All}). While the similarity spaces derived from the quantum theoretical descriptors resulted in the best correlations with mutagenicity, spaces derived from atom pairs and the combined set of topological and quantum theoretical descriptors estimated mutagenicity nearly as well. The results for the five similarity spaces are summarized in Table 10.7, where *r* is the correlation coefficient, *s .e.* is the standard error, *n* is the number of dimensions or axes in the similarity space, and *k* is the number of selected "nearest neighbors" used to estimate mutagenicity for each chemical within the space.

10.4.1.2 Molecular Similarity and Analog Selection

As mentioned earlier, many times a researcher's goal is to select a set of analogs for a chemical of interest from a large, diverse data set based on similarity spaces derived solely from calculated

Table 10.7 Comparison of five similarity methods in the estimation of mutagenicity (In Rev/nmol in *S. typhimurium* TA100 with metabolic activation) for 73 aromatic and heteroaromatic amines

Similarity method	*r*	*s.e.*	*n*	*k*
AP	0.77	0.88	na	4
PC_{TI}	0.72	0.96	6	5
Property	0.83	0.77	3	5
PC_{Prop}	0.84	0.75	3	5
PC_{All}	0.79	0.85	7	4

descriptors of molecular structure. We described above in Sect. 10.3.2 our PCA analysis of the diverse set of 3692 industrial chemicals [19]. As part of this study, analogs were selected based on *Euclidean distance* within the ten-dimensional similarity space derived from the ten major principal components. Figure 10.5 presents an example of the five nearest neighbors (or analogs) selected for one chemical from the set of 3692 molecules.

A look at the five selected structures, particularly the ones closest to 4-hydroxybenzene acetic acid (the probe or query chemical), shows that there is sufficient degree of similarity of the query structure with the selected analogs in terms of the number and type of atoms, degree of cyclicity, aromaticity, etc.

10.4.1.3 The K-Nearest Neighbor (KNN) Approach in Predicting Modes of Action (MOAs) of Industrial Pollutants

Different domains of chemical screening use different model organisms for the assessment of bioactivity of chemicals. In aquatic toxicology and ecotoxicology, *fathead minnow* is an important model organism [84–86]. Numerous QSARs have been developed with subsets of fathead minnow toxicity (LC_{50}) data, many such models being developed using small, structurally related or congeneric sets. But, following the *diversity begets diversity principle* discussed above, one will need a diverse collection of molecular descriptors for the QSAR formulation of diverse collection of chemicals. Another possibility is to develop different subsets of chemicals from a large and diverse set based on their *mode of action* (MOA) first and then treat chemicals with the same MOA as *biological congeners* as opposed to structural classes which may be called *structural congeners*. Basak et al. [87] undertook a classification study based on acute toxic MOA of industrial chemicals. At that time the US Environmental Protection Agency's Mid-Continent Ecology Division-Duluth, Minnesota, fathead minnow database had LC_{50} data on 617 chemicals. But out of that list, only 283 chemicals were selected by us because our experimental cooperators had good confidence about the MOAs of that subset only. Such evidence consisted of concurrent information from joint chemical toxicity studies, physicochemical and behavioral response, information published in peer-reviewed literature, and toxicity over time [88]. Such caution in the selection of good subsets of data for modeling is in line with the *veracity attribute* mentioned above while discussing the major pillars of QSAR and issues regarding Big Data [80].

Acute toxic mode of action of the chemicals was predicted using molecular similarity method, neural networks of the Learning Vector Quantization (LVQ) type, and discriminant analysis methods. The set of 283 compounds was broken down into

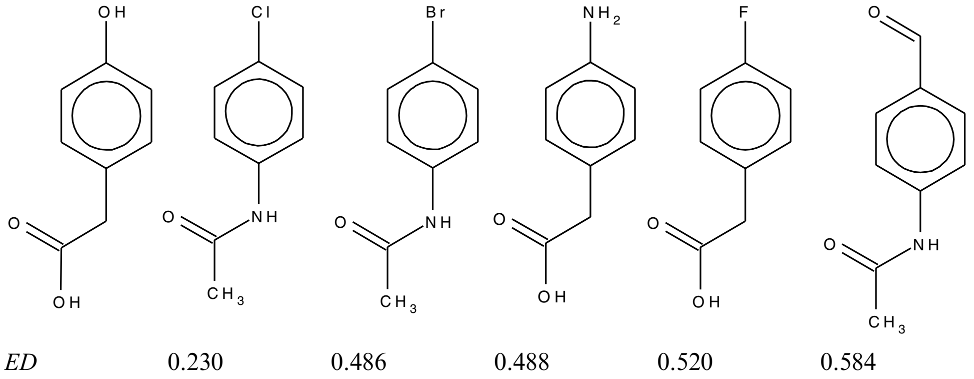


Fig. 10.5 Molecularstructuresfor4-hydroxybenzeneacetic acid and its five analogs selected from a database of 3692 chemicals. The numbers below each structure are the Euclidean distances (*ED*) between 4-hydreoxybenzeneacetic acid (the left-most structure) and its analogs

a training set of 220 compounds and a test set of 63. Computed topological indices and atom pairs were used as structural descriptors for model development. The five MOA classes represented included:

1. Narcosis I/II and electrophile/proelectrophile reactivity (NE)
2. Uncouplers of oxidative phosphorylation (UNC)
3. Acetylcholinesterase inhibitor (AChE-I)
4. Neruotoxicants (NT)
5. Neuordepressants/respiratory blockers (RB/ND)

In the molecular similarity approach, similarity between chemicals i and j was defined as

$$S_{\mathrm{ij}} = 2C/\left(T_{\mathrm{i}} + T_{\mathrm{j}}\right) \qquad (10.4)$$

where C is the number of atom pairs [10] common to molecules i and j. $T_{\mathrm{i}} + T_{\mathrm{j}}$ are the total number of atom pairs in i and j, respectively. The five nearest neighbors (i.e., $K=5$) were used to predict the mode of action of a probe or query chemical.

In the neural network analysis, LVQ classification network was used, consisting of a 60-node input layer, a 5-node hidden layer, and a 5-node output layer.

Linear models utilizing stepwise discriminant analysis were developed in addition to the neural network and similarity models.

All three methods gave good results for training and test sets, with the success ranging from 95 % for the K-nearest neighbor method to 87 % for the discriminant analysis technique. This consistency of results obtained using topological descriptors in different classification methods indicates that the graph theoretical parameters used in this study contain sufficient structural information to be capable of predicting modes of action of diverse chemical species. Table 10.8 provides the classification results obtained using the K-nearest neighbor method, in which 90 % of the training set chemicals and 95 % of the test set chemicals were classified correctly.

10.4.1.4 The *Tailored Approach* to Developing Similarity Spaces

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

As mentioned above, user-defined or arbitrary molecular similarity methods perform reasonably well in narrow, well-defined situations. But the relationship between structural attributes and biomedicinal or toxicological properties are not always crisp; they are often messy. Human intuition often fails in such circumstances. Similarity methods based on objectively defined relationships are needed, rather than those derived from subjective or intuitive approaches. In a multivariate space, this should be accomplished using robust statistical methods. The *tailored similarity method* starts with an appropriate number of molecular descriptors [89–91]. These descriptors are run through *ridge regression* analysis modeling the property of interest, and a small number of independent variables with high |*t*| values are selected as the axes of the similarity space. In this way, we select variables which are strongly

Table 10.8 MOA classification results using the K-nearest neighbor ($K=5$) method

	Training set	
	$n=220$	% Correct
NE	180/183	98 %
UNC	6/10	60 %
AChE-I	7/14	50 %
NT	0/7	0 %
RB/ND	5/6	83 %
	Overall	**90 %**

	Test set	
	$n=63$	% Correct
NE	53/54	98 %
UNC	2/2	100 %
AChE-I	3/3	100 %
NT	1/2	50 %
RB/ND	1/2	50 %
	Overall	**95 %**

related with the property of interest instead of a subjectively selected group of descriptors. Needless to say, human intuition will be hard pressed to match the objective relationship developed by ridge regression techniques.

In one tailored similarity study [91], we examined the effects of tailoring on the estimation of logP for a set of 213 chemicals and on the estimation of mutagenicity for a set of 95 aromatic and heteroaromatic amines. In this study we utilized a much larger set of topological indices than have been used in many of our *earlier* studies. Three distinct similarity spaces were constructed, though two were "overlapping" spaces. The overlapping spaces were derived using principal component analysis on the set of 267 topological indices. The PCA created 20 orthogonal components with eigenvalues greater than one. These 20 PCs were used as the axes for the first similarity space. The second similarity space was derived from the principal components. In examining the PCs, we selected the index most correlated with each cluster as a representative of the cluster. One of the arguments against using PCA to reduce the number of variables for modeling is that PCs, being linear combinations of the indices, are not easily interpretable. So, by selecting the most correlated single TI from each PC, we have a set of easily interpretable topological indices to use in modeling.

Finally, the third set of indices was selected based on a ridge regression model developed from all 267 indices to predict mutagenicity. From the modeling results, *t*-values were extracted and the 20 indices with the highest absolute [t] values were selected as axes for developing the similarity space. A summary of the correlation coefficients for estimating mutagenicity from the three similarity spaces for varying numbers of neighbors using the KNN method is presented in Fig. 10.6.

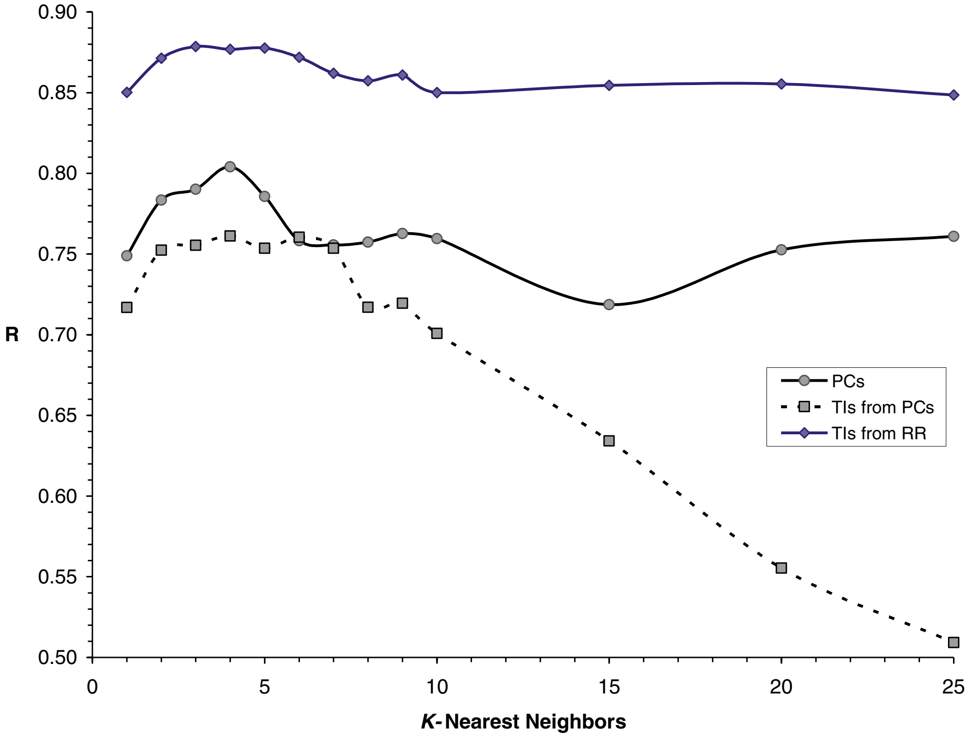


Fig. 10.6 Plot of the pattern of correlation coefficient (*R*) from *k* = 1–10, 15, 20, and 25 for the estimation of mutagenicity (ln Rev/nmol) for 95 aromatic and heteroaromatic amines using a 20 principal component space derived from 267 topological indices (PCs), a 20 topological index space selected from the principal components (TIs from PCs), and a 20 topological index based on space derived from ridge regression (TIs from RR)

It is clear from Fig. 10.6 that tailoring the selected set of indices significantly improved the estimative power of the model, resulting in roughly a 10 % increase to the correlation coefficient. These results, as with all of the results we have seen from tailored similarity spaces, are promising, and we believe that **tailored similarity methods will be very useful both in drug discovery and toxicological research**.

10.5 Formulation of Biodescriptors from DNA/RNA Sequences and Proteomics Maps: Development and Applications

If your chromosomes are XYY,
And you are a naughty, naughty guy,
Your crimes, the judge won't even try,
'Cause you have a legal reason why
He'll raise his hands and gently sigh!
"I guess for this you get a by."
By Carl A. Dragstedt
In: Perspectives in Biology and Medicine
Vol. 14, # 1, autumn, 1970

10.5.1 Mathematical Biodescriptors from DNA/RNA Sequences

After the completion of the *Human Genome Project*, a lot of data for DNA, RNA, and protein sequences are being generated. In line with the idea of representation and mathematical characterization of chemicals (see Fig. 10.2 above), various authors have developed such representation-cum-characterization methods for DNA/RNA sequences [16, 92–96]. In the past few years, a lot of papers have been published in this area. Here, we give a brief history of the recent growth spurt of this exciting field beginning in 1998. Dilip K. Sinha and Subhash C. Basak started the Indo-US Workshop Series on Mathematical Chemistry [97] in 1998, the first event being held at the Visva Bharati University, Santiniketan, West Bengal, India. Raychaudhury and Nandy [98] gave a presentation on mathematical characterization of DNA sequences using their graphical method. This caught the attention of Basak who later developed a research group on the mathematical characterization of DNA/RNA sequences supported by funds from the University of Minnesota Duluth-Natural Resources Research Institute (UMD-NRRI) and University of Minnesota. This led to the publication of the first couple of papers on DNA sequence invariants [99, 100]. The rest of the development of DNA/RNA sequence graph invariants and mathematical descriptors is clear from the hundreds of papers published on this topic subsequently by authors all over the world. More recently Nandy and Basak applied this method in the characterization of the various bird flu sequences, e.g., *H5N1 bird flu* [101] and *H5N2 pandemic bird flu* [102], the latter one causing havoc in the turkey and poultry farms of the Midwest of the USA in 2015. Numerous other theoretical developments and practical applications of DNA/RNA mathematical descriptors are not discussed here for brevity.

10.5.2 Mathematical Proteomics-Based Biodescriptors

Proteomics may be looked upon as a branch of Functional Genomics that studies changes in protein-protein and protein-drug/toxicant interactions. Scientists are studying proteomics for new drug discovery and predictive toxicology [103–105]. A typical 2D gel electrophoresis (2DE)-derived proteomics map provided to us by our collaborators at Indiana University is provided in Fig. 10.7.

The 2DE method of proteomics is capable of detecting and characterizing a few thousand proteins from a cell, tissue, or animal. One can then study the effects of well-designed structural or mechanistic classes of chemicals on animals or specialized cells and use these proteomics data to classify the molecules or predict their biological action. But with 1000–2000 protein spots present per gel, the difficult question we face is: **How do we make sense of the chaotic pattern of the large number of proteins as shown in Fig. 10.7?**

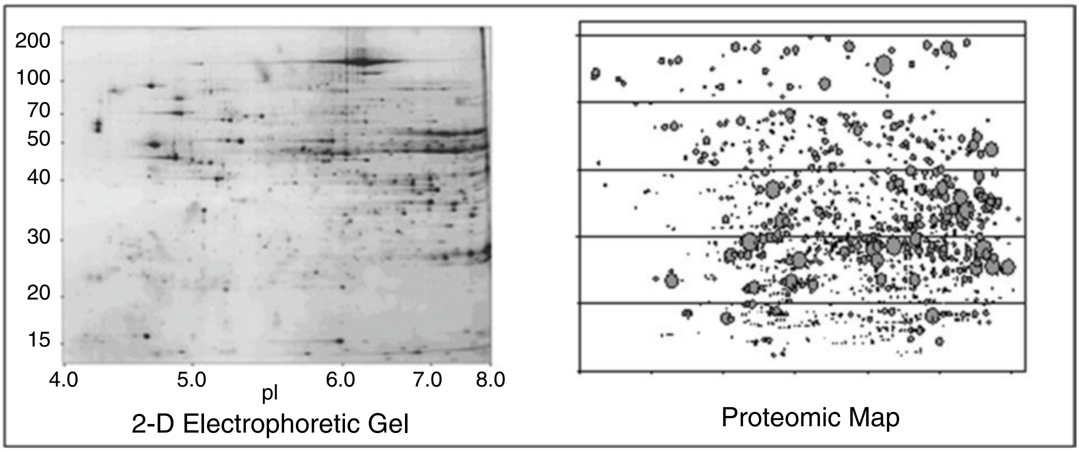


Fig. 10.7 Location and abundance of protein spots derived from 2D gel electrophoresis (Courtesy of Frank Witzmann of Indiana University, Indianapolis, USA)

We have attacked this problem through the formulation of biodescriptors applying the techniques of *discrete mathematics* to proteomics maps. Described below are three major approaches developed by our research team at the Natural Resources Research Institute and its collaborators for the quantitative calculation of biodescriptors of proteomics maps, the term **biodescriptor** being coined by the Basak group for the first time:

(a) In each 2D gel, the proteins are separated by charge and mass. Also associated with each protein spot is a value representing abundance, which quantifies the amount of that particular protein or closely related class of proteins gathered on one spot. Mathematically, the data generated by 2DE may be looked upon as points in a three-dimensional space, with the axes described by charge, mass, and spot abundance. One can then have projections of the data to the three planes, i.e., XY, YZ, and XZ. The *spectrum-like data* so derived can be converted into vectors, and similarity of proteomics maps can be computed from these map descriptors [106].

(b) In a second approach, viz., the graph invariant biodescriptor method, different types of embedded graphs, e.g., zigzag graphs neibhborhood graphs, are associated with proteomics maps, with the set of spots in the proteomics maps representing the vertices of such graphs. In the zigzag approach, one begins with the spot of the highest abundance and draws an edge between it and the spot having the next highest abundance and continues this process. The resulting zigzag curve is converted into a *D*/*D matrix* where the (i, j) entry of such a matrix is the quotient of the Euclidean distance and the through-bond distance. For details on this approach, please see [107].

(c) A proteomics map may be looked upon as a pattern of protein mass distributed over a 2D space. The distribution may vary depending on the functional state of the cell under various developmental and pathological conditions as well as under the influence of exogenous chemicals such as drugs and xenobiotics. Information theoretic approach has been applied to compute biodescriptors called *map information content* (*MIC*) from 2D gels [108].

10.6 Combined Use of Chemodescriptors and Biodescriptors for Bioactivity Prediction

We told above in Eq. 10.2 that in many cases, the property/bioactivity/toxicity of chemicals can be predicted reasonably well using their structure (S) alone. But in many complex biological situations, e.g., induction of cancer by exposure to chemical carcinogens, we need to use both structural features of such chemicals and biological test data to make sense of such endpoints. Arcos [109], for example, suggested the use of specific biological data, e.g., degranulation of endoplasmic reticulum, peroxisome proliferation, unscheduled DNA synthesis, antispermatogenic activity, etc., as biological indicators of carcinogenesis. Such biochemical data not only bring direct and relevant biological observations into the set of predictors, they also bring independent variables which are closer to the endpoint in the scale of complexity than the chemical structure. In line with this *structural*-cum-*functional approach* in predicting bioactivity of chemicals, we have used a combination of chemodescriptors and proteomics-based biodescriptors for assessing toxicity of priority pollutants [28, 110].

10.7 Discussion

> We are all agreed that your theory is crazy. The question which divides us is whether it is crazy enough to have a chance of being correct. My own feeling is that it is not crazy enough.
>
> Niels Bohr

> Everything should be made as simple as possible,
> but not simpler.
> – Albert Einstein

Major objectives of this chapter have been to review our research in the use of mathematical chemodescriptors and biodescriptors in the prediction of bioactivity/toxicity of chemicals, quantification of similarity/dissimilarity among chemical species from their chemodescriptors, and similarity-based clustering, as well as estimation of toxicologically relevant properties of diverse groups of molecules.

In the chemodescriptor area, our major goal has been to review the utility of graph theoretical parameters, also known as topological indices, in QSAR and QMSA studies. We studied the intercorrelation of major topological indices in an effort to identify subsets that are minimally correlated [57, 111]. We have also used principal components derived from TIs and all TIs simultaneously (e.g., ridge regression models) in QSAR formulation. At present a large number of descriptors can be calculated for chemicals using available software. If the number of experimental data points (dependent variables) for QSAR model building is much smaller than the number of descriptors, i.e., the situation is rank-deficient, one needs to be cautious. We have discussed the variable selection methods including ITC [56] which, to our knowledge, has been brought to QSAR from the genomics/ genetics area for the first time in our research. In the calculation of q^2 in the rank-deficient case, one must follow the *two-deep cross-validation* procedure; otherwise the calculated q^2 will reflect overfitting [43–45, 51, 52, 55]. We have demonstrated this using one example where we deliberately used the wrong ordinary least square (OLS) approach in a rank-deficient case and compared the results with the correct approach to show the difference between them [45]. In HiQSAR modeling, we found that of the four types of calculated molecular descriptors, viz., TS, TC, 3-D, and QC indices, in the majority of cases a TS + TC combination gave good quality models; the addition of 3-D or QC descriptors after the use of TS and TC combination did not improve much the model quality. This is a good news in view of the fact that we are already at the age of *Big Data* [80] and easily calculated indices like TS and TC descriptors, if they give good models in many areas, could find wide applications in the *in silico* screening of chemicals. The *congenericity principle* has been a major theme of QSAR whereby there has been a tendency in developing QSARs of congeneric sets of chemicals. When the same property, viz., mutagenicity, of congeneric versus diverse sets was used to develop QSAR models, the congeneric set of 95 amines had much lower number of significant descriptors as compared to the diverse set of 508 molecules. This gives support to the *diversity begets diversity principle* formulated by us [18].

When a large number of descriptors are calculated for a set of chemicals, the data set becomes high dimensional. The use of PCA can derive a much smaller number of orthogonal variables which reflect the *parsimony principle* or *Occam's razor* [62].

Molecular similarity is used both in drug design and hazard assessment of chemicals [36, 39, 112]. We used calculated TIs and atom pairs to generate similarity spaces following different methods and used both Euclidean distance derived from PCs and Tanimoto coefficient based on atom pairs to select analogs. The structures of analogs selected from the structurally diverse set of 3692 industrial chemicals indicated that the calculated property-based QMSA methods are capable of selecting analogs of query chemicals that look reasonably structurally similar to them. We also used our QMSA method in selecting analogs of environmental pollutants for which the modes of action are known with high confidence from experimental toxicology. The results of the MOA prediction study show that selected analogs of chemicals with specified MOA fall in similar toxicological categories.

In the post-*genomic era*, the *omics* technologies are generating a lot of data on the effects of chemicals on the genetic system, viz., transcription, translation, and posttranslational modification, of the cell and tissue. We have been involved in the development of biodescriptors from DNA/RNA sequences and two-dimensional gel electrophoresis (2DE) data derived from cells/tissue exposed to drugs and toxicants. Results of our research in this area show that the biodescriptors developed from proteomics maps are capable of characterizing the pharmacological/toxicological profiles of chemicals [106–108]. Some preliminary studies have been done on the use of the combined set of chemodescriptors and biodescriptors in predicting bioactivity. Further research are needed to test the relative effectiveness of the two classes of descriptors, chemodescriptors versus biodescriptors, in predictive pharmacology and toxicology [28, 110].

At this juncture, after reviewing results of a large number of QSAR studies using chemodescriptors and biodescriptors, we may ask ourselves: *Quo Vadimus*? We have seen that calculated chemodescriptors are capable of predicting and characterizing bioactivity and toxicity as well as toxic modes of action of chemicals. Research using biodescriptors of different types also shows that such descriptors derived from proteomics maps have reasonable power of discriminating among structurally closely related toxicants. Can we, at this stage, opt for either chemodescriptor or biodescriptors alone? The answer is *no*, as is evident from our experience in predictive toxicology. This indicates that in the foreseeable future, we will need an integrated approach consisting of chemodescriptors and biodescriptors in order to obtain the best results (Fig. 10.8).

As discussed by this author [113] in a recent book on Advances in Mathematical Chemistry and applications:

> Mathematical chemistry or more accurately discrete mathematical chemistry had a tremendous growth spurt in the second half of the twentieth century and the same trend is continuing now. This growth was fueled primarily by two major factors: (1) Novel applications of discrete mathematical concepts to chemical and biological systems, and (2) 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 in a timely manner. This led to

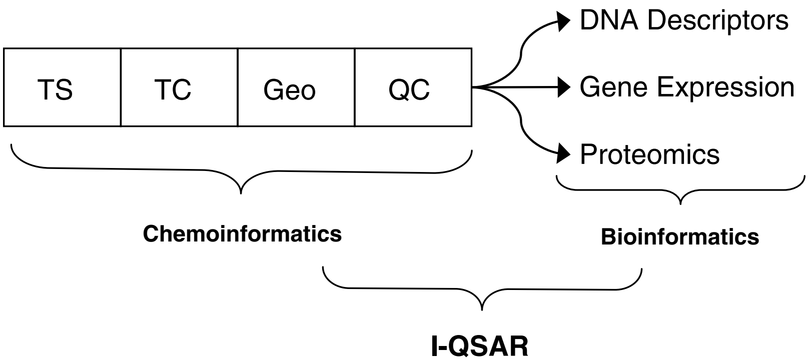


Fig. 10.8 Integrated QSAR, combining chemodescriptors and biodescriptors

the development of not only a plethora of new concepts, but also various useful applications to such important areas as drug discovery, protection of human as well as ecological health, bioinformatics, and chemoinformatics. Following the completion of the Human Genome Project in 2003, discrete mathematical methods were applied to the "omics" data to develop descriptors relevant to bioinformatics, toxicoinformatics, and computational biology.

The results of various types of research using chemodescriptors and biodescriptors [16–21, 28, 108, 114] derived through applications of discrete mathematics on chemical and biological systems give us hope that an exciting future is in front of us.

Acknowledgments I am thankful to Kanika Basak, Gregory Grunwald, Douglas Hawkins, Brian Gute, Subhabrata Majumdar, Denise Mills, Dilip K. Sinha, Ashesh Nandy, Frank Witzmann, Kevin Geiss, Krishnan Balasubramanian, Ramanathan Natarajan, Gerald J. Niemi, Alexandru T. Balaban, the late Alan Katritzky, Milan Randic, Nenad Trinajstic, Sonja Nikolic, Marjan Vracko, Marjana Novic, Xiaofeng Guo, Terry Neumann, Qianhong Zhu, late Gilman D. Veith, Marissa Harle, Vincent R. Magnuson, Donald K. Harriss, Chandan Raychaudhury, Samar K. Ray and Lester R. Drewes for collaboration in my research.

References

1. Hardman JG, Limbird LE, Gilman AG (2001) Goodman and Gilman's the pharmacological basis of therapeutics. McGraw- Hill, New York
2. Hoffman DJ, Ratner BA, Burton GA Jr, Cairns J Jr (1995) Handbook of ecotoxicology. CRC Press, Boca Raton
3. Nogrady T (1985) Medicinal chemistry: a biochemical approach. Oxford University Press, New York
4. Rand G (ed) (1995) Fundamentals of aquatic toxicology: effects, environmental fate and risk assessment, 2nd edn. Taylor and Francis, New York
5. Primas H (1981) Chemistry, quantum mechanics and reductionism. Springer, Berlin
6. Woolley RG (1978) Must a molecule have a shape? J Am Chem Soc 100:1073–1078
7. Basak SC, Veith GJ, Niemi GD (1991) Predicting properties of molecules using graph invariants. J Math Chem 7:243–272
8. Einstein A (1954) Remarks on Bertrand Russell's theory of knowledge. In: Einstein A (ed) Ideas and opinions. Ed. Carl Seelig, (Based on MEIN WELTBILD, edited by Carl Seelig, and other sources; New translations and revisions by Sonja Bargmann), Crown Publishers, New York, pp 18–24
9. Bunge M (1973) Method, model and matter. Reidel, Dordrecht
10. Carhart RE, Smith DH, Venkataraghavan R (1985) Atoms pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73. doi:10.1021/ci00046a002
11. Euler I (1736) Solutio problematis ad geometriam situs pertinentis. Comment Acad Sci U Petrop 8:128–140
12. Sylvester JJ (1878) On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. Am J Math 1:105–125
13. http://www.goodreads.com/quotes/926286-chemistry-has-the-same-quickening-and-suggestive-influence-upon-the
14. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20
15. Balasubramanian K, Basak SC (1998) Characterization of isospectral graphs using graph invariants and derived orthogonal parameters. J Chem Inf Comput Sci 38:367–373
16. Nandy A, Harle M, Basak SC (2006) Mathematical descriptors of DNA sequences: development and application. Arkivoc 9:211–238
17. Basak SC (2013) Philosophy of mathematical chemistry: a personal perspective. HYLE Int J Philos Chem 19:3–17
18. Basak SC (2013) 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 9:449–462
19. Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988) Determining structural similarity of chemicals using graph-theoretic indices. Discrete Appl Math 19:17–44
20. Lajiness M (1990) Molecular similarity-based methods for selecting compounds for screening. In: Rouvray DH (ed) Computational chemical graph theory. Nova, New York, pp 299–316
21. Basak SC, Mills D, Gute BD, Balaban AT, Basak K, Grunwald GD (2010) Use of mathematical structural invariants in analyzing, combinatorial libraries: a case study with psoralen derivatives. Curr Comput Aided Drug Des 6:240–251
22. Basak SC (2014) Molecular similarity and hazard assessment of chemicals: a comparative study of arbitrary and tailored similarity spaces. J Eng Sci Manag Educ 7:178–184
23. Basak SC (1987) Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. Med Sci Res 15:605–609
24. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) Discrimination of isomeric structures

using information-theoretic topological indices. J Comput Chem 5:581–588

25. Balaban AT, Mills D, Ivanciuc O, Basak SC (2000) Reverse wiener indices. Croat Chim Acta 73:923–941
26. Nikolic S, Trinajstic N, Amic D, Beslo D, Basak SC (2001) Modeling the solubility of aliphatic alcohols in water. Graph connectivity indices versus line graph connectivity indices. In: Diudea MV (ed) QSAR/QSPR studies by molecular descriptors. Nova, Huntington, pp 63–81
27. Randic M, Vracko M, Nandy A, Basak SC (2000) On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf Comput Sci 40:1235–1244
28. Basak SC, Gute BD (2008) Mathematical descriptors of proteomics maps: background and applications. Curr Opin Drug Discov Dev 11:320–326
29. Hosoya H (1971) Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Soc Jpn 44:2332–2339
30. MolconnZ (2003) Version 4.05. Hall Ass. Consult. Quincy
31. Basak SC, Harriss DK, Magnuson VR (1988) POLLY v. 2.3. Copyright of the University of Minnesota, USA
32. Basak SC, Grunwald GD (1993) APProbe. Copyright of the University of Minnesota, USA
33. Filip PA, Balaban TS, Balaban AT (1987) A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. J Math Chem 1:61–83
34. Stewart JJP (1990) MOPAC Version 6.00, QCPE #455, Frank J Seiler Research Laboratory, US Air Force Academy, CO
35. Frisch MJ et al (1998) Gaussian 98 (Revision A.11.2). Gaussian, Inc., Pittsburgh
36. Auer CM, Nabholz JV, Baetcke KP (1990) 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 87:183–197
37. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474
38. Todeschini R, Consonni V, Mauri A, Pavan M. (2006) DRAGON – Software for the calculation of molecular descriptors, version 5.4, Talete srl. Milan.
39. Johnson M, Basak SC, Maggiora G (1988) A characterization of molecular similarity methods for property prediction. Math Comput Mod 11:630–634
40. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
41. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
42. Cook RD, Li B, Chiaromonte F (2010) Envelope models for parsimonious and efficient multivariate linear regression. Stat Sin 20:927–1010
43. Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. J Chem Inf Comput Sci 3:579–586
44. Hawkins DM, Basak SC, Mills D (2004) QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. Environ Toxicol Pharmacol 16:37–44
45. Basak SC, Mills D, Hawkins DM, Kraker JJ (2007) Proper statistical modeling and validation in QSAR: a case study in the prediction of rat fat-air partitioning. In: Simos TE, Maroulis G (eds) Computation in modern science and engineering, proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007). American Institute of Physics, Melville, pp 548–551
46. Basak SC, Majumdar S (2016) 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: Basak SC, Restrepo G, Villaveces JL (ed) Advances in mathematical chemistry and applications, vol 1. Bentham eBooks, Elsevier & Bentham Science Publishers, Sharjah, U. A. E, pp 251–281
47. Basak SC, Majumdar S (2015) Hierarchical quantitative structure-activity relationships (HiQSARs) for the prediction of physicochemical and toxicological properties of chemicals using computed molecular descriptors, Mol2Net Conference. http://sciforum.net/email/validate/49668c1bf65ab8520f721a84f7d84e05
48. Majumdar S, Basak SC, Grunwald GD (2013) Adapting interrelated two-way clustering method for quantitative structure-activity relationship (QSAR) modeling of mutagenicity/non-mutagenicity of a diverse set of chemicals. Curr Comput Aided Drug Des 9:463–471
49. Basak SC, Majumdar S (2015) Prediction of mutagenicity of chemicals from their calculated molecular descriptors: a case study with structurally homogeneous versus diverse datasets. Curr Comput Aided Drug Des 11:117–123
50. Basak SC, Majumdar S (2015) The importance of rigorous statistical practice in the current landscape of QSAR modelling (editorial). Curr Comput Aided Drug Des 11:2–4
51. Kraker JJ, Hawkins DM, Basak SC, Natarajan R, Mills D (2007) Quantitative structure-activity relationship (QSAR) modeling of juvenile hormone activity: comparison of validation procedures. Chemometr Intell Lab Syst 87:33–42
52. Hawkins DM, Kraker JJ, Basak SC, Mills D (2008) QSPR checking and validation: a case study with hydroxy radical reaction rate constant. SAR QSAR Environ Res 19:525–539
53. SAS Institute, Inc (1988) In SAS/STAT user guide, release 6.03 edition. Cary
54. Hoskuldsson A (1995) A combined theory for PCA and PLS. J Chemom 9:91–123

55. Hawkins DM, Basak SC, Shi X (2001) QSAR with few compounds and many features. J Chem Inf Comput Sci 41:663–670
56. Tang C, Zhang L, Zhang A, Ramanathan M (2001) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: Bilof R, Palagi L (eds) Proceedings of BIBE 2001: 2nd IEEE international symposium on bioinformatics and bioengineering, Bethesda, Maryland, November 4–5, 2001. IEEE Computer Society, Los Alamitos, pp 41–48
57. Basak SC, Magnuson VR, Niemi GJ, Regal RR, Veith GD (1987) Topological indices: their nature, mutual relatedness, and applications. Math Mod 8:300–305
58. Basak SC, Grunwald GD, Majumdar S (2015) Intrinsic dimensionality of chemical space: characterization and applications, Mol2Net conference. http://sciforum.net/email/validate/49668c1bf65ab8520f721a84f7d84e05
59. Basak SC (1999) Information theoretic indices of neighborhood complexity and their applications. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam, pp 563–593
60. Randic M (1975) Characterization of molecular branching. J Am Chem Soc 97:6609–6615
61. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67:4517–4533
62. Hoffmann R, Minkin VI, Carpenter BK (1997) Ockham's razor and chemistry. HYLE Int J Philos Chem 3:3–28
63. Katritzky AR, Putrukhin R, Tathan S, Basak SC, Benfenati E, Karelson M, Maran U (2001) Interpretation of quantitative structure-property and -activity relationships. J Chem Inf Comput Sci 41:679–685
64. Katritzky AR, Putrukhin R, Tathan S, Basak SC, Benfenati E, Karelson M, Maran U (2001) Interpretation of quantitative structure-property and -activity relationships. J Chem Inf Comput Sci 41:679–685
65. So SS, Karplus M (1997) Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. J Med Chem 40:4360–4371
66. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding potency of dibenzofurans: a hierarchical QSAR approach. Ind J Chem 42A:1385–1391
67. Basak SC, Majumdar S (2015) 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: Basak SC, Restrepo G, Villaveces JL (eds) Advances in mathematical chemistry and applications, vol 1. Bentham eBooks, Bentham Science Publishers, pp 251–281
68. Ben-Dor A, Friedman N, Yakhini Z (2001) Class discovery in gene expression data. In: Proceedings of the fifth annual international conference on computational molecular biology (RECOMB 2001), New York
69. Gute BD, Basak SC (1997) Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. SAR QSAR Environ Res 7:117–131
70. Gute BD, Grunwald GD, Basak SC (1999) Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): a hierarchical QSAR approach. SAR QSAR Environ Res 10:1–15
71. Basak SC, Mills DR, Balaban AT, Gute BD (2001) Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: a hierarchical QSAR approach. J Chem Inf Comput Sci 41:671–678
72. Popper K (2005) The logic of scientific discovery. Taylor & Francis e-Library, London and New York
73. Basak SC, Majumdar S (2015) Two QSAR paradigms- congenericity principle versus diversity begets diversity principle- analyzed using computed mathematical chemodescriptors of homogeneous and diverse sets of chemical mutagens. Mol2Net Conference. http://sciforum.net/email/validate/49668c1bf65ab8520f721a84f7d84e05
74. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17:4791–4810
75. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim 33:445–459
76. Preparata FP, Shamos MI (1991) Convex hulls: basic algorithms. In: Preparata FP, Shamos MI (eds) Computational geometry: an introduction. Springer, New York, pp 95–148
77. Worth AP, Bassan A, Gallegos A, Netzeva TI, Patlewicz G, Pavan M, Tsakovska I, Vracko M (2005) The characterisation of (quantitative) structure-activity relationships: preliminary guidance, ECB Report EUR 21866 EN. European Commission, Joint Research Centre, Ispra, p 95
78. Pharmaceutical Research and Manufacturers of America (2014) Biopharmaceutical research industry profile. Available from: http://www.phrma.org/sites/default/files/pdf/2014_PhRMA_PROFILE.pdf. Accessed on 11 Dec 2015
79. Santos-Filho OA, Hopfinger AJ, Cherkasov A, de Alencastro RB (2009) The receptor-dependent QSAR paradigm: an overview of the current state of the art. Med Chem (Shariqah) 5:359–366
80. Basak SC, Bhattacharjee AK, Vracko M (2015) Big data and new drug discovery: tackling "Big Data" for virtual screening of large compound databases. Curr Comput Aided Drug Des 11:197–201
81. Crawford MA (1963) The effects of fluoroacetate, malonate and acid-base balance on the renal disposal of citrate. Biochem J 8:115–120
82. Quastel JH, Wooldridge WR (1928) Some properties of the dehydrogenating enzymes of bacteria. Biochem J 22:689–702

83. Basak SC, Grunwald GD (1995) Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. Chemosphere 31:2529–2546
84. Reuschenbach P, Silvani M, Dammann M, Warnecke D, Knacker T (2008) ECOSAR model performance with a large test set of industrial chemicals. Chemosphere 71:1986–1995
85. Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung W, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmeider PK, Serrano JA, Tietge J, Villeneuve DL (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 29:730–741
86. Ankley GT, Villeneuve DL (2006) The fathead minnow in aquatic toxicology: past, present and future. Aquat Toxicol 78:91–102
87. Basak SC, Grunwald GD, Host GE, Niemi GJ, Bradbury SP (1998) A comparative study of molecular similarity, statistical and neural network methods for predicting toxic modes of action of chemicals. Environ Toxicol Chem 17:1056–1064
88. Russom CL, Bradbury SP, Broderius SJ, Hammermeister DE, Drummond RA (1997) Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (pimephales promelas). Environ Toxicol Chem 16:948–967
89. Gute BD, Grunwald GD, Mills D, Basak SC (2001) Molecular similarity based estimation of properties: a comparison of structure spaces and property spaces. SAR QSAR Environ Res 11:363–382
90. Gute BD, Basak SC, Mills D, Hawkins DM (2002) Tailored similarity spaces for the prediction of physicochemical properties. Internet Electron J Mol Des 1:374–387. http://www.biochempress.com/
91. Basak SC, Gute BD, Mills D, Hawkins DM (2003) Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. J Mol Struct THEOCHEM 622:127–145
92. Hamori E, Ruskin J (1983) H Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J Biol Chem 258:1318–1327
93. Gates MA (1986) A simple way to look a DNA. J Theor Biol 119:319–328
94. Nandy A (1996) Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. Curr Sci 70:661–668
95. Leong PM, Morgenthaler S (1995) Random walk and gap plots of DNA sequences. Comput Appl Biosci 11:503–507
96. Randić M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D (2011) Graphical representation of proteins. Chem Rev 111:790–862
97. Indo-US Workshop on Mathematical Chemistry. http://www.nrri.umn.edu/indousworkshop
98. Raychaudhury C, Nandy A (1998) Indexation schemes and similarity measures for macromolecular sequences. Paper presented at the Indo-US Workshop on Mathematical Chemistry, Shantiniketan. 9–13 January 1998
99. Randić M, Vracko M, Nandy A, Basak SC (2000) On 3–D representation of DNA primary sequences. J Chem Inf Comput Sci 40:1235–1244
100. Guo X, Randić M, Basak SC (2001) A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem Phys Lett 350:106–112
101. Nandy A, Sarkar T, Basak SC, Nandy P, Das S (2014) Characteristics of influenza HA-NA interdependence determined through a graphical technique. Curr Comput Aided Drug Des 10:285–302
102. Nandy A, Basak SC (2015) Prognosis of possible reassortments in recent H5N2 epidemic influenza in USA: implications for computer-assisted surveillance as well as drug/vaccine design. Curr Comput Aided Drug Des 11:110–116
103. Steiner S, Witzmann FA (2000) Proteomics: applications and opportunities in preclinical drug development. Electrophoresis 21:2099–2104
104. Witzmann FA, Monteiro-Riviere NA (2006) Multi-walled carbon nanotube exposure alters protein expression in human keratinocytes. Nanomedicine Nanotechnol Biol Med 2:158–168
105. Basak SC, Gute BD, Monteiro-Riviere N, Witzmann FA (2010) Characterization of toxicoproteomics maps for chemical mixtures using information theoretic approach. In: Mumtaz M (ed) Principles and practice of mixtures toxicology. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 215–232
106. Vracko M, Basak SC, Geiss K, Witzmann FA (2006) Proteomics maps-toxicity relationship of halocarbons studied with similarity index and genetic algorithm. J Chem Inf Model 46:130–136
107. Randic M, Witzmann FA, Vracko M, Basak SC (2001) On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: application to peroxisome proliferators. Med Chem Res 10:456–479
108. Basak SC, Gute BD, Witzmann FA (2006) Information-theoretic biodescriptors for proteomics maps: development and applications in predictive toxicology. Conf Proc WSEAS Trans Inf Sci Appl 7:996–1001
109. Arcos JC (1987) Structure–activity relationships: criteria for predicting the carcinogenic activity of chemical compounds. Environ Sci Technol 21:743–745
110. Hawkins DM, Basak SC, Kraker JJ, Geiss KT, Witzmann FA (2006) Combining chemodescriptors and biodescriptors in quantitative structure-

activity relationship modeling. J Chem Inf Model 46:9–16

111. Basak SC, Gute BD, Balaban AT (2004) Interrelationship of major topological indices evidenced by clustering. Croat Chem Acta 77:331–344
112. Johnson M, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
113. Basak SC (2016) Mathematical structural descriptors of molecules and biomolecules: background and applications. In: Basak SC, Restrepo G, Villaveces JL (ed) Advances in mathematical chemistry and applications, vol 1. Bentham eBooks, Elsevier & Bentham Science Publishers, Sharjah, U. A. E. pp 3–23
114. Zanni R, Galvez-Llompart M, Garcıa-Domenech R, Galvez J (2015) Latest advances in molecular topology applications for drug discovery. Expert Opin Drug Discov 10:1–13