Synthesis Lectures on
Mathematics & Statistics

SYNTHESIS
COLLECTION OF TECHNOLOGY

Subhash C. Basak *Editor*

Mathematical Descriptors of Molecules and Biomolecules

Applications in Chemistry, Drug Design, Chemical Toxicology, and Computational Biology

Springer

Synthesis Lectures on Mathematics & Statistics

This series includes titles in applied mathematics and statistics for cross-disciplinary STEM professionals, educators, researchers, and students. The series focuses on new and traditional techniques to develop mathematical knowledge and skills, an understanding of core mathematical reasoning, and the ability to utilize data in specific applications.

Subhash C. Basak
Editor

Mathematical Descriptors of Molecules and Biomolecules

Applications in Chemistry, Drug Design, Chemical Toxicology, and Computational Biology

Springer

Editor
Subhash C. Basak
University of Minnesota
Duluth, MN, USA

ISSN 1938-1743 ISSN 1938-1751 (electronic)
Synthesis Lectures on Mathematics & Statistics
ISBN 978-3-031-67840-0 ISBN 978-3-031-67841-7 (eBook)
https://doi.org/10.1007/978-3-031-67841-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Only those who will risk going too far can possibly find out how far one can go.

—T. S. Eliot

…shall we stay our upward course? In that blessed region of Four Dimensions, shall we linger at the threshold of the Fifth, and not enter therein? Ah, no! Let us rather resolve that our ambition shall soar with our corporal ascent. Then, yielding to our intellectual onset, the gates of the Six Dimension shall fly open; after that a Seventh, and then an Eighth…

—Edwin Abbott, In: Flatland

In science there is and will remain a Platonic element which could not be taken away without ruining it. ***Among the infinite diversity of singular phenomena science can only look for invariants****.*

—Jacques Monod

Upon this gifted age, in its dark hour,
Rains from the sky a meteoric shower
Of facts… they lie unquestioned, uncombined.
Wisdom enough to leech us of our ill
Is daily spun; but there exists no loom
To weave it into fabric…

—Edna St. Vincent Millay

We are now living in an age when numerous spheres of science, technology, and life are experiencing an explosion of huge amounts of data. The buzzword is "*data is the new oil*." The same is true in the advancing frontiers of chemoinformatics and bioinformatics. More and more physicochemical, biological, omics (genomics, proteomics, etc.), and sequence

data are coming to the public domain almost daily. Just as the refinement of crude oil is necessary for our daily use, the raw data in chemoinformatics and bioinformatics needs to be systematized and used in the formulation of robust predictive models to aid basic research and to assist decision support systems in socially and economically important fields like new drug discovery, and regulation of chemicals for the protection of human health and ecology.

In many cases, experimental data come with associated information about the structure of molecules or biomolecules. Unless descriptors used in model development are all totally experimentally determined, nonempirical descriptors calculated from structure without the input of any other experimental data are used for modeling. Such descriptors are mainly topological, three-dimensional (3-D), or quantum chemical in nature. Mathematical methods like matrix theory, graph theory, and information theory are some frequently used tools in molecular descriptor calculation. *The use of descriptors in predictive models arises from the dictum: Function (property) follows form (structure).* These days, often the number of predictors/descriptors (p) is much larger than n, the number of data points to be modeled. In such rank-deficient cases, proper statistical and machine learning methods need to be used to develop robust predictive models. *Therefore, it is evident that mathematical and statistical/ML methods constitute an indispensable link between structural/property data on the one hand and implementable models in the end user's computer, on the other.*

In the first chapter "Approaching Modeling in Chemoinformatics and Bioinformatics Using Mathematical Descriptors: Some Comments on the Emerging Landscape and Future Directions" of the book, Subhash C. Basak discusses the fundamental philosophy behind the formulation of chemodescriptors and biodescriptors. He also explains the methodologies for the creation of robust predictive models under rank-deficient situations. In the second chapter "Hierarchy of Descriptors: From Topology to Bio-descriptors", Marjan Vračko and Subhash C. Basak discuss the hierarchy of molecular descriptors derived from diverse mathematical techniques. The third chapter "Chirality Descriptors for Numerical Characterization of Enantiomers and Diastereomers", authored by Ramanathan Natarajan, Subhash C. Basak, and Claudiu N. Lungu, summarizes the recent development of a set of novel chirality descriptors to mathematically characterize relative chirality of molecules containing one or more asymmetric centers. The fourth chapter "QSAR Modeling Using Molecular Fragment Descriptors" by Suman K. Chakravarti delves into the use of structural fragments in the formulation of models useful in predictive pharmacology and toxicology. The fifth chapter "Quantitative Structure-Activity Analysis Using Conceptual DFT and Information Theory-based Descriptors" of the book, authored by Arpita Poddar, Ranita Pal, Shanti Gopal Patra, and Pratim Kumar Chattaraj, discusses the calculation and utility of quantum chemical density functional theory (DFT) descriptors in Quantitative Structure-Activity Analysis (QSAR) model development. The sixth chapter of the book entitled "Complexity of Molecular Ensembles with Basak's Indices: Applying Structural Information Content", jointly authored by Denis Sabirov, Alexandra Zimina, and Igor Shepelevich, discusses the utility of the novel information theoretic

topological index, structural information content (SIC), originally developed by Subhash C. Basak, in the characterization of the complexity of molecular ensembles. Chapter seven entitled "Descriptors from Calculated Stereo-Electronic Properties and Molecular Electrostatic Potentials (MEPs) May Provide a Powerful "Interaction Pharmacophore" for Drug Discovery", by Apurba K. Bhattacharjee, discusses calculation and practical use of computed interaction pharmacophore in the design of new life-saving drugs. The eighth and final chapter of the book "Network-Based Molecular Descriptors for Protein Dynamics and Allosteric Regulation", by Ziyun Zhou, Lorenza Pacini, Laurent Vuillon, Claire Lesieur, and Guang Hu, deals with the use of network theory in the characterization of allosteric properties of proteins from their sequence and structural data.

I would like to specially mention that currently the fields of chemoinformatics and bioinformatics are witnessing a huge explosion of data describable by four V's: volume, velocity, variety, and veracity. However, the data per se is of very limited use unless we complete the three-step process: data is first converted to information which is then transformed into useful knowledge.

The real challenge lies in the implementation of the last two steps of the three-step pathway in the creation of actionable knowledge.

The chapters in this book involving chemodescriptors and biodescriptors expand in two distinct directions. On the one hand, the authors delve into more subtle and novel aspects of the structure of molecules and biomolecules using innovative methods ranging from topology and graph theory, and network theory to high-level quantum chemistry. Such research may lead to a new understanding of behaviors of molecules and biomolecules at the molecular and submolecular levels. On the other hand, following the "*diversity begets diversity paradigm*", in the day-to-day practical applications by end users, the collection of descriptors taken from diverse sources may be useful in the evaluation of chemicals and biochemicals, real or hypothetical, facilitated by computer algorithms and statistical/machine learning-based software.

We sincerely hope that this book will enrich the expanding frontiers of chemoinformatics and bioinformatics in the areas of both basic and applied research.

Duluth, Minnesota, USA Subhash C. Basak

Contents

Approaching Modeling in Chemoinformatics and Bioinformatics Using Mathematical Descriptors: Some Comments on the Emerging Landscape and Future Directions 1
Subhash C. Basak

Hierarchy of Descriptors: From Topology to Bio-descriptors 13
Marjan Vračko and Subhash C. Basak

Chirality Descriptors for Numerical Characterization of Enantiomers and Diastereomers 27
Ramanathan Natarajan, Subhash C. Basak, and Claudiu N. Lungu

QSAR Modeling Using Molecular Fragment Descriptors 75
Suman K. Chakravarti

Quantitative Structure-Activity Analysis Using Conceptual DFT and Information Theory-based Descriptors 91
Arpita Poddar, Ranita Pal, Shanti Gopal Patra, and Pratim Kumar Chattaraj

Complexity of Molecular Ensembles with Basak's Indices: Applying Structural Information Content 113
Denis Sabirov, Alexandra Zimina, and Igor Shepelevich

Descriptors from Calculated Stereo-Electronic Properties and Molecular Electrostatic Potentials (MEPs) May Provide a Powerful "Interaction Pharmacophore" for Drug Discovery 123
Apurba K. Bhattacharjee

Network-Based Molecular Descriptors for Protein Dynamics and Allosteric Regulation 151
Ziyun Zhou, Lorenza Pacini, Laurent Vuillon, Claire Lesieur, and Guang Hu

Approaching Modeling in Chemoinformatics and Bioinformatics Using Mathematical Descriptors: Some Comments on the Emerging Landscape and Future Directions

Subhash C. Basak

Abstract

This chapter briefly discusses the history, status, and use of different classes of chemodescriptors and biodescriptors, both experimentally determined and computed from structure, in the formulation of predictive models. Possible future uses of the various classes of descriptors in model building at this age of big data analytics are highlighted.

Keywords

Molecular structure • Model object • Theoretical model • Chemodescripors • Biodescriptor • Quantitative structure–activity relationship (QSAR) • Statistical and machine learning methods • Big data analytics • Graph invariant • Topological index

1 Introduction

The most beautiful thing we can experience is the mysterious. It is the source of all true art and all science. He to whom this emotion is a stranger, who can no longer pause to wonder and stand rapt in awe, is as good as dead: his eyes are closed.

— Albert Einstein

It's all a series of serendipities

with no beginnings and no ends.

S. C. Basak (✉)
Department of Chemistry and Biochemistry, Retired Adjunct Professor, University of Minnesota Duluth, Duluth, MN 55811, USA
e-mail: sbasak@d.umn.edu

S. C. Basak (ed.), *Mathematical Descriptors of Molecules and Biomolecules*, Synthesis Lectures on Mathematics & Statistics, https://doi.org/10.1007/978-3-031-67841-7_1

Such infinitesimal possibilities

Through which love transcends.

— Ana Claudia Antunes, The Tao of Physical and Spiritual

[T]he prepared mind requires unfettered opportunity to recognize and follow unplanned paths . . . when we pursue our passion to master what was once unknowable, we move from a plodding struggle with nature to an ongoing, enlightening conversation.

Joshua Lederberg, 21stC: Research at Columbia, 1995

Computed mathematical descriptors of chemical and biological systems, called chemodescriptors and biodescriptors [1–4, 6, 7, 16, 17, 21, 23, 31] play a key role in many basic and applied aspects of chemical and biological research related to new drug discovery and computational toxicology. In the realms of pharmaceutical drug design and assessment of toxicity of environmental pollutants as well as industrial chemicals we must screen many thousands of candidate substances for their potential beneficial properties, adverse effects, and bioactivities [3, 8]. The exhaustive testing of all these large number of chemicals in the laboratory is prohibitively costly and will necessitate the sacrifice of a huge number of test organisms. Under such circumstances, property/ bioactivity prediction models based on existing test data can act as a decision support system in the allocation of scant resources in the necessary lab testing phase [3, 6, 7, 16]. During the last half century or so, numerous software for the calculation of molecular descriptors (Table 1) have become available to us, e.g., POLLY [11, 12], Triplet [10], MolConnZ [39], Dragon [25], MOPAC [40], Gaussian [27]. Fortunately, during the past few decades a lot of property/ bioassay databases have come to the public domain (Table 2). Such data can be used for model building and prediction of properties of untested and even unsynthesized chemicals. The available computing power has been steadily increasing following Moore's law [32].

As highlighted by Basak [3], the four important pillars of quantitative structure–activity relationship (QSAR) studies supporting predictive pharmacology and computational toxicology are:

(a) Good quality experimental physicochemical and biological test data,
(b) A reasonable number of experimental data prerequisite to good model building and validation pertinent to extrapolation of models to a structurally diverse set of chemical structures [13],
(c) Availability of software that can compute properties of chemicals ***from their structure only without the input of any other experimental data*** so that such properties can be calculated for any molecule, real or hypothetical (Table 1),
(d) Robust statistical and machine learning methods necessary for model building and validation.

Table 1 A list of important classes of chemodescriptors and biodescriptors

	Chemodescriptors					Biodescriptors			
Physical property (experimental or calculated)	LFER/ LSER descriptors	3-D descriptors	Pharmacophoric descriptors	Graph theoretic descriptors	Quantum chemical descriptors	GC content	Proteomics descriptors	AFSDs of of DNA/RNA graphs	Descriptors of amino acid/ protein network
Boiling point/ vapor pressure; log P (octanol–water); aqueous solubility;	Hammet sigma; logP; hydrogen bond donor/ acceptor property	Molar volume; van der Waals's volume	H-bonding, hydrophobic, and electrostatic interaction sites, defined by atoms, ring centers, etc.	Different types of topostructural and topochemical indices; substructures	QC indices calculate by various levels of theory	G + C content of a genome	Mathematical descriptors d4rived from 2-D gel electrophoresis and GC/MS data	Quantitative descriptors from graphs of DNA/ RNA sequences	Protein structure descriptors from primary to the higher levels of organization

Table summary

Chemodescriptors: Linear free energy related (LFER) descriptors are derived from physical organic chemistry model systems and include hydrophobicity (e.g. Log P, octanol–water), electronic descriptors li Hammett's sigma and steric descriptors like Es. Linear solvation energy related (LSER) descriptors mainly consist of volume, polarizability, hydrogen bond donor acidity and hydrogen bond acceptor basicity. 3-D or geometrical descriptors include van der Waals' volume, surface area etc. Pharmacophore descriptors are derived from the characterization of stereo-electronic factors needed for the specific recognition of ligands by specific enzymes or recepto5rs. Graph theoretical descriptors can be either real numbers like the topological indices or substructures derived from the molecular graph model of chemical structure. Quantum chemical indices of molecules are derived by different semiempirical and *ab initio* methods

Biodescriptors: GC (Guanine-Cytosine) content or percentage of GC in the total base content [(G + C)/(A + T + G + C)]. Quantitative proteomics descriptors which are computed mathematically from proteomics patterns derived from 2-dimensional gel electrophoresis (2DE) or GC/ MS analysis of cellular/ tissue proteins. Alignment-free sequence descriptors (AFSDs) are mathematical descriptors derived from operations on graphs corresponding to the DNA/RNA sequences and are not related to any alignment-based approach. Quantitative descriptors of amino acid sequences of protein and 3-D structure can be used in predicting their properties

Table 2 Some available chemical and biological databases

Name	Web address	Type of information
PubChem	https://pubchem.ncbi.nlm.nih.gov	Contains data on a large number of bioassays, providing bioactivity and toxicity endpoints of chemicals
ChEMBL	ChEMBL: https://www.ebi.ac.uk/chembl/	A database with a focus on drug discovery
ToxCast	https://www.epa.gov/chemical-research/exploring-toxcast-data	A database of high-throughput screening chemical toxicity data by the US environmental protection agency
Tox 21- Toxilology in the Twenty First Century	https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html	A high-throughput screening database of toxicity data by the US national institutes of health and the US environmental protection agency
Toxicity reference database (ToxRefDB	https://catalog.data.gov/dataset/toxicity-reference-database	A database of toxicity data by the US environmental protection agency
ECHA dossiers	https://echa.europa.eu/information-on-chemicals/registered-substances	European chemicals agency's (ECHA) system for registration, evaluation, authorization, and restriction of chemicals. Including various in vivo toxicity data
OECD QSAR toolbox	https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm	A collection of QSAR models and toxicity data, developed by the organization for economic co-operation and development (OECD)
EURL ECVAM database	https://data.jrc.ec.europa.eu/dataset/b7597ada-148d-4560-9079-ab0a5539cad3	Data on alternative methods to animal testing (DB-ALM): a database of alternative methods for toxicity testing
UC Irvine (UCI) machine learning repository	https://archive.ics.uci.edu/ml/index.php	A good collection of data of all kinds, ready for model building

(continued)

Table 2 (continued)

Name	Web address	Type of information
GenBank and WGS database	https://www.ncbi.nlm.nih.gov/genbank/statistics/	Whole genome shotgun (WGS) database contains data on chromosomes of prokaryotes or eukaryotes that are generally being sequenced by whole genome shotgun strategy
RCSB protein data bank (RCSB PDB	https://www.rcsb.org/	Contains experimentally derived and computed protein structure data
GISAID (global initiative on sharing avian influenza data)	https://gisaid.org/	Provides open access genomic data of influenza viruses and the coronavirus responsible for the COVID-19 pandemic
DRIGBAML	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238889/	Provides detailed data with comprehensive drug target and drug action information

Today, a combination of the above-mentioned four factors, in particular, gives us the opportunity of developing powerful predictive models for practical use in various areas of basic and applied research. As more experimental test data on collections of highly diverse chemical structures are being available to us, in QSAR research there is a necessary paradigm shift, ***a transition from the "Congenericity Principle" toward the "Diversity begets Diversity" principle*** [13].

2 The Enormous Descriptor Landscape

It is worth mentioning that In the post-genomics era, catapulted by the completion of the Human Genome Project [29], a lot of data on the macromolecular DNA, RNA, protein sequences as well as expression of genetic information in normal and chemically/ biologically/clinically affected cells/tissues are being collected in different publicly available databases (Table 2). Such data are good starting points in the development of mathematical and computational biodescriptors for the characterization of biological systems. Table 1 gives a short list of popularly used chemodescriptors and biodescriptors.

The term chemodescriptors is used for descriptors derived for small organic molecules or more complex entities like peptides [3]. These may include both experimentally determined properties like melting point, boiling point and hydrophobicity (e.g., Octanol–water partition coefficient) of chemicals. These may also include the different classes of molecular descriptors which can be computed from the structure of chemicals, e.g., topostructural

(TS), topochemical (TC), geometrical or three-dimensional (3-D) and various types of quantum chemical (QC) indices. Biodescriptor development is aimed at the characterization of complex biological objects like DNA/ RNA/ protein sequences, proteomics maps or complex macromolecular networks in cells [2]. ***A descriptor, chemo- or biodescriptor, quantifies certain aspects of the structure of the entity which it characterizes***.

The fundamental philosophy of descriptor formulation consists of two major stages: (a) Representation of the entity (chemical or biological object like the molecule or DNA sequence) under investigation into a model object and (b) Characterization of the model object using mathematical models [14, 20].

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, or symbolic representation of chemical species by Hamiltonian operators) are model objects derived through different methods of 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 objects. This explains the plurality of the concept of molecular structure and their autonomous nature, the word "autonomous" being used in the sense that one concept is not logically derived from the other [14, 36, 42].

Figure 1 gives a short overview of the development of physicochemical molecular descriptors since 1868:

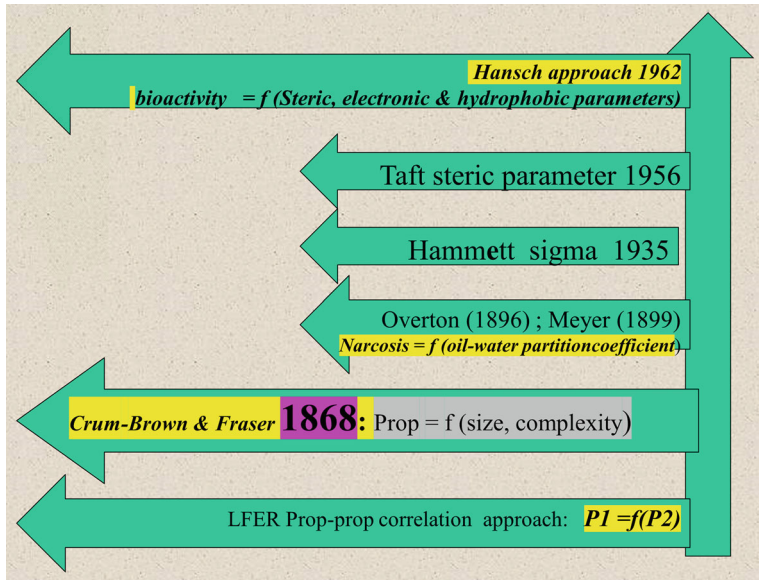

Fig. 1 Brief history of Hansch approach: 1868–to date

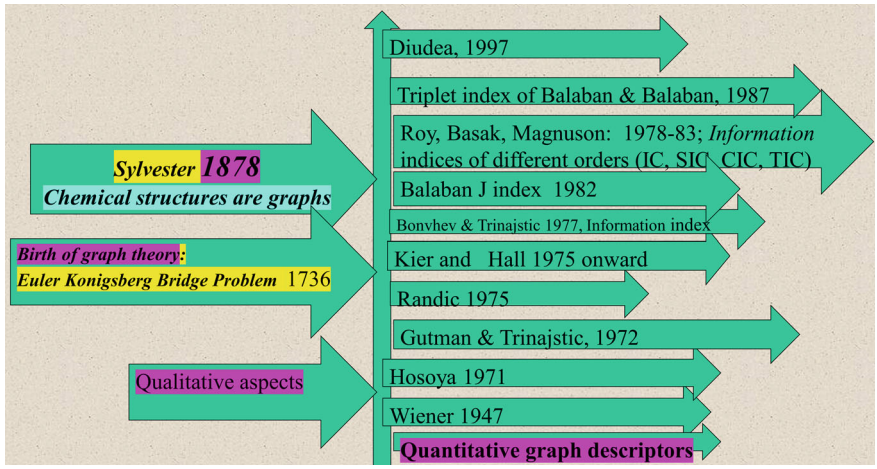

Fig. 2 A brief history of graph theory based indices: 1736–to date

Legend to Fig. 1: Fig. 1 gives a brief chronological order of the development of LFER (linear free energy related) molecular descriptors. For more details, please see Basak [6].

Figure 2 gives a bird's eye view of the history of the development of graph theory-based descriptors starting with the initial discovery by Euler in 1736.

Legend to Fig. 2: Fig. 2 gives a brief overview of the development of graph theory-based descriptors starting with the discovery of graph theory by Euler in 1736.

3 Use of Descriptors in Model Building

In recent years graph theoretical descriptors, both numerical graph invariants and sub-structural descriptors, have found wide applications in the prediction of physical property, pharmacological properties and potential toxicity of chemicals using QSAR models [6, 21, 31]. They have also been used in QMSA (quantitative molecular similarity analysis) models in the quantification of intermolecular similarity/ dissimilarity of chemicals for drug design and predictive toxicology [5]. Figure 2 above gives a brief history of the development and use of numerical graph theoretical descriptors, also known as topological indices, starting from the discovery of graph theory by Euler in 1736.

In many cases mathematical descriptors provide a numerical scale for some qualitative attributes of the structure. Basak [3, 4]. For example, the connectivity index developed by [37] was supposed to represent molecular branching. This was a topostructural (TS) class of descriptor. The inclusion of chemically relevant information into molecular graphs by [30] resulted in a wide variety of indices of the connectivity type which had a lot of success in QSAR.

In the realm of information theoretic topological indices [15, 18, 19], the initial crop of indices was topostructural which quantified mainly the topological/ vertex connectivity information of the molecules under consideration. Subsequent development by [38] used the topochemical (involving both molecular topology or atomic connectivity as well as bonding and electronic properties of vertices/ atoms) approach to formulate a family of novel information theoretic graph invariants based on different orders of topological neighborhoods of atoms in the molecular graph.

A similar picture is true for chirality descriptors which numerically characterize the structural complexity arising out of the presence of one or more asymmetric centers in a molecule. The Basak group and collaborators [35] have developed a family of diverse chirality descriptors which can be applied in the QSAR of many different types of chiral molecules. *The simultaneous use of different classes of chemodescriptors, including the recently developed chirality indices may be looked upon as a pragmatic fusion or synthetic approach in QSAR model building in line with the diversity begets diversity principle put forward by Basak group* [7, 13].

In the realms of biodescriptors-mathematical quantifiers of structural aspects of biomolecules like DNA/RNA sequences, proteomics maps, etc.- various indices have been developed using multiple approaches which found useful applications in the characterization of emerging global pathogens [9, 34, 41], prediction of toxicity of chemicals [2] to cells and tissues as well as design of new drugs and peptide vaccines for new viral diseases [33]. Becasuse in many cases of chemical-biological interactions only chemoinformatics cannot model the complex situation effectively, an integrated QSAR (I-QSAR) approach involving knowledge derived both from chemoinforamtics and bioinfoermtics domains seems to be the right technique for effective model formulation [6, 7, 13, 17]. The pluarlistic I-QSAR approach that uses phyhsicochemical proptrries, computed structural descriptors and biodescriptors may evolve as a viable model for predicting complex emergent properites of chemical and biiological systems [2].

4 Conclusion

Nature knows no pause in progress and development and attaches her curse on all inaction.

Johann Wolfgang von Goethe

See first, think later, then test. But always see first. Otherwise, you will only see what you were expecting. Most scientists forget that.

Douglas Adams

Data matures like wine, applications like fish.

—James Governor