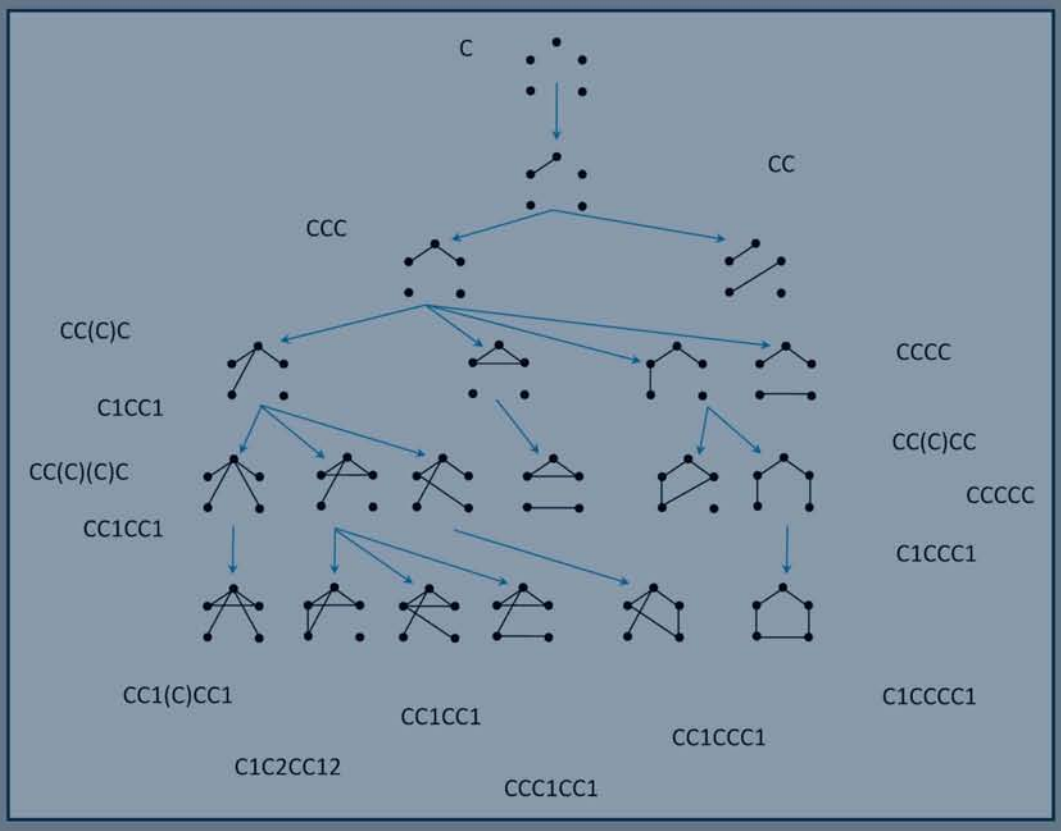
Chapman & Hall/CRC
Mathematical and Computational Biology Series

Handbook of Chemoinformatics Algorithms



Edited by
Jean-Loup Faulon
Andreas Bender

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC Mathematical and Computational Biology Series

Handbook of Chemoinformatics Algorithms

Edited by
Jean-Loup Faulon
Andreas Bender

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CHAPMAN & HALL/CRC
Mathematical and Computational Biology Series

Aims and scope:
This series aims to capture new developments and summarize what is known over the entire spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical, and computational methods into biology by publishing a broad range of textbooks, reference works, and handbooks. The titles included in the series are meant to appeal to students, researchers, and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Proposals for the series should be submitted to one of the series editors above or directly to:
CRC Press, Taylor & Francis Group
4th, Floor, Albert House
1-4 Singer Street
London EC2A 4BQ
UK

Published Titles

Algorithms in Bioinformatics: A Practical Introduction
Wing-Kin Sung

Bioinformatics: A Practical Approach
Shui Qing Ye

Biological Sequence Analysis Using the SeqAn C++ Library
Andreas Gogol-Döring and Knut Reinert

Cancer Modelling and Simulation
Luigi Preziosi

Cell Mechanics: From Single Scale-Based Models to Multiscale Modeling
Arnaud Chauvière, Luigi Preziosi, and Claude Verdier

Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R
Gabriel Valiente

Computational Biology: A Statistical Mechanics Perspective
Ralf Blossey

Computational Neuroscience: A Comprehensive Approach
Jianfeng Feng

Data Analysis Tools for DNA Microarrays
Sorin Draghici

Differential Equations and Mathematical Biology, Second Edition
D.S. Jones, M.J. Plank, and B.D. Sleeman

Engineering Genetic Circuits
Chris J. Myers

Exactly Solvable Models of Biological Invasion
Sergei V. Petrovskii and Bai-Lian Li

Gene Expression Studies Using Affymetrix Microarrays
Hinrich Göhlmann and Willem Talloen

Glycome Informatics: Methods and Applications
Kiyoko F. Aoki-Kinoshita

Handbook of Chemoinformatics Algrithms
Jean-Loup Faulon and Andreas Bender

Handbook of Hidden Markov Models in Bioinformatics
Martin Gollery

Introduction to Bioinformatics
Anna Tramontano

An Introduction to Systems Biology: Design Principles of Biological Circuits
Uri Alon

Kinetic Modelling in Systems Biology
Oleg Demin and Igor Goryanin

Knowledge Discovery in Proteomics
Igor Jurisica and Dennis Wigle

Meta-analysis and Combining Information in Genetics and Genomics
Rudy Guerra and Darlene R. Goldstein

Modeling and Simulation of Capsules and Biological Cells
C. Pozrikidis

Niche Modeling: Predictions from Statistical Distributions
David Stockwell

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems
Qiang Cui and Ivet Bahar

Optimal Control Applied to Biological Models
Suzanne Lenhart and John T. Workman

Pattern Discovery in Bioinformatics: Theory & Algorithms
Laxmi Parida

Python for Bioinformatics
Sebastian Bassi

Spatial Ecology
Stephen Cantrell, Chris Cosner, and Shigui Ruan

Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation
Horst Malchow, Sergei V. Petrovskii, and Ezio Venturino

Stochastic Modelling for Systems Biology
Darren J. Wilkinson

Structural Bioinformatics: An Algorithmic Approach
Forbes J. Burkowski

The Ten Most Wanted Solutions in Protein Bioinformatics
Anna Tramontano

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-8299-9 (Ebook-PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

The field of handling chemical information electronically—known as Chemoinformatics or Cheminformatics—has received a boost in recent decades, in line with the advent of tremendous computer power. Originating in the 1960s in both academic and industrial settings (and termed by its current name only from around 1998), chemoinformatics applications are today commonplace in every pharmaceutical company. Also, various academic laboratories in Europe, the United States, and Asia confer both undergraduate and graduate degrees in the field.

But still, there is a long way to go. While resembling its sibling, bioinformatics, both by name and also (partially) algorithmically, the chemoinformatics field developed in a very different manner right from the onset. While large amounts of biological information—sequence information, structural information, and more recently also phenotypic information such as metabolomics data—found their way straight into the public domain, large-scale chemical information was until very recently the domain of private companies. Hence, public tools to handle chemical structures were scarce for a very long time, while essential bioinformatics tools such as those for aligning sequences or viewing protein structures were available at no cost to anyone interested in the area. More recently—luckily—this situation changed significantly, with major life science data providers such as the NCBI, the EBI, and many others also making large-scale chemical data publicly available.

However, there is another aspect, apart from the actual data, that is crucial for a scientific field to flourish—and that is the proper documentation of techniques and methods, and, in the case of informatics sciences, the proper documentation of algorithms. In the bioinformatics field, and in line with a tremendous amount of open access data and tools available, algorithms were documented extensively in reference books. In the chemoinformatics field, however, a book of this type is missing until now. This is what the editors, with the help of expert contributors in the field, are attempting to remedy—to provide an overview of some of the most common chemoinformatics algorithms in a single place.

The book is divided into 15 chapters. Chapter 1 presents a historical perspective of the applications of algorithms and graph theory to chemical problems. Algorithms to store and retrieve two-dimensional chemical structures are presented in Chapter 2, and three-dimensional representations of chemicals are discussed in Chapter 3. Molecular descriptors, which are widely used in virtual screening and structure–activity/property predictions, are presented in Chapter 4. Chapter 5 presents virtual screening methods from a ligand perspective and from a structure perspective including docking methods. Chapters 6 and 7 are dedicated to quantitative structure–activity relationships (QSAR). QSAR modeling workflow and methods to prepare the data are presented in Chapter 6, while the development and validation of QSAR models are discussed in Chapter 7. Chapter 8 introduces algorithms to enumerate and sample chemical structures, with applications in combinatorial libraries design. Chapters 9 and 10 are

dedicated to computer-aided molecular design: from a ligand perspective in Chapter 9, where inverse-QSAR methods are reviewed, and from a structure perspective in Chapter 10, where *de novo* design algorithms are presented. Chapter 11 covers reaction network generation, with applications in synthesis design and biological network inference. Closing the strictly chemoinformatics chapters, Chapter 12 provides a review of Open Source software and database technologies dedicated to the field. The remaining chapters (13–15) present techniques developed in the context of bioinformatics and computational biology and their potential applications to chemical problems. Chapter 13 discusses possible applications of sequence alignment algorithms to tree-like structures such as glycans. Chapter 14 presents classical machine learning algorithms that can be used for both bioinformatics and chemoinformatics problems. Chapter 15 introduces a systems biology approach to study the kinetics of metabolic networks.

While our book covers many aspects of chemoinformatics, our attempt is ambitious—and it is probably impossible to provide a complete overview of "all" chemoinformatics algorithms in one place. Hence, in this work we present a selection of algorithms from the areas the editors deemed most relevant in practice and hope that this work will be helpful as a reference work for people working in the field.

MATLAB® and Simulink® are registered trademarks of The Math Works, Inc. For product information, please contact:

The Math Works, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098, USA
Tel: 508 647 7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Jean-Loup Faulon, *Paris, France*
Andreas Bender, *Leiden, the Netherlands*

Acknowledgments

The editors would like to first thank Robert B. Stern from the Taylor & Francis Group for giving them an opportunity to compile, for the first time, an overview of chemoinformatics algorithms. They also thank the authors for assembling expert materials covering many algorithmic aspects of chemoinformatics. Jean-Loup Faulon would like to acknowledge the interest and encouragement provided by Genopole's Epigenomics program and the University of Evry, France, to edit and coauthor chapters in this book.

The authors of Chapter 2 would like to thank Ovidiu Ivanciuc for providing relevant literature references. They also acknowledge the permission to reprint Algorithm 2.1 [Dittmar et al. *J. Chem. Inf. Comput. Sci.*, 17(3): 186–192, 1977. Copyright (1977) American Chemical Society]. Milind Misra acknowledges funding provided by Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Markus Meringer would like to thank Emma Schymanski for carefully proofreading Chapter 8.

Shawn Martin would like to acknowledge funding (to write Chapter 14) provided by Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Finally, Fangping Mu, Amy L. Bauer, James R. Faeder, and William S. Hlavacek acknowledge funding support (to write Chapter 15) provided in part by the NIH, under grants GM080216 and CA132629, and by the DOE, under contract DE-AC52-06NA25396. They also thank P.J. Unkefer, C.J. Unkefer, and R.F. Williams for helpful discussions.

Contributors

Tatsuya Akutsu
Institute for Chemical Research
Kyoto University
Uji, Japan

Amy L. Bauer
Theoretical Biology and Biophysics Group
Theoretical Division, Los Alamos National Laboratory
Los Alamos, New Mexico

Pablo Carbonell
Institute of Systems and Synthetic Biology
University of Evry
Evry, France

Robert D. Clark
Biochemical Infometrics and School of Informatics
Indiana University
Bloomington, Indiana

James R. Faeder
Department of Computational Biology
University of Pittsburgh School of Medicine
Pittsburgh, Pennsylvania

Jean-Loup Faulon
Department of Biology
University of Evry
Evry, France

Nikolas Fechner
Department of Computer Architecture
University of Tübingen
Tübingen, Germany

Alexander Golbraikh
Division of Medicinal Chemistry and Natural Products
University of North Carolina
Chapel Hill, North Carolina

Rajarshi Guha
NIH Chemical Genomics Center
Rockville, Maryland

Georg Hinselmann
Department of Computer Architecture
University of Tübingen
Tübingen, Germany

William S. Hlavacek
Theoretical Biology and Biophysics Group
Theoretical Division, Los Alamos National Laboratory
Los Alamos, New Mexico

Ovidiu Ivanciuc
Department of Biochemistry and Molecular Biology
University of Texas Medical Branch
Galveston, Texas

Shawn Martin
Sandia National Laboratories
Albuquerque, New Mexico

Markus Meringer
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany

Milind Misra
Sandia National Laboratories
Albuquerque, New Mexico

Fangping Mu
Theoretical Biology and Biophysics Group
Theoretical Division, Los Alamos National Laboratory
Los Alamos, New Mexico

Diana C. Roe
Department of Biosystems Research
Sandia National Laboratories
Livermore, California

Alexander Tropsha
Division of Medicinal Chemistry and Natural Products
University of North Carolina
Chapel Hill, North Carolina

Donald P. Visco, Jr.
Department of Chemical Engineering
Tennessee Technological University
Cookeville, Tennessee

Jörg Kurt Wegner
Integrative Chem-/Bio-Informatics
Tibotec (Johnson & Johnson)
Mechelen, Belgium

Egon L. Willighagen
Department of Pharmaceutical Biosciences
Uppsala University
Uppsala, Sweden

1 Representing Two-Dimensional (2D) Chemical Structures with Molecular Graphs

Ovidiu Ivanciuc

CONTENTS

1.1 INTRODUCTION

Graphs are used as an efficient abstraction and approximation for diverse chemical systems, such as chemical compounds, ensembles of molecules, molecular fragments, polymers, chemical reactions, reaction mechanisms, and isomerization pathways. Obviously, the complexity of chemical systems is significantly reduced whenever they are modeled as graphs. For example, when a chemical compound is represented as a molecular graph, the geometry information is neglected, and only the atom connectivity information is retained. In order to be valuable, the graph representation of a chemical system must retain all important features of the investigated system and has to offer qualitative or quantitative conclusions in agreement with those provided by more sophisticated methods. All chemical systems that are successfully modeled as graphs have a common characteristic, namely they are composed of elements that interact between them, and these interactions are instrumental in explaining a property of interest of that chemical system. The elements in a system are represented as graph vertices, and the interactions between these elements are represented as graph edges. In a chemical graph, vertices may represent various elements of a chemical system, such as atomic or molecular orbitals, electrons, atoms, groups of atoms, molecules, and isomers. The interaction between these elements, which are represented as graph edges, may be chemical bonds, nonbonded interactions, reaction steps, formal connections between groups of atoms, or formal transformations of functional groups. The chapter continues with an overview of elements of graph theory that are important in chemoinformatics and in depicting two-dimensional (2D) chemical structures. Section 1.3 presents the most important types of chemical and molecular graphs, and Section 1.4 reviews the representation of molecules containing heteroatoms and multiple bonds with weighted graphs and molecular matrices.

1.2 ELEMENTS OF GRAPH THEORY

This section presents the basic definitions, notations, and examples of graph theory relevant to chemoinformatics. Graph theory applications in physics, electronics, chemistry, biology, medicinal chemistry, economics, or information sciences are mainly the effect of the seminal book *Graph Theory* of Harary [1]. Several other books represent essential readings for an in-depth overview of the theoretical basis of graph theory: *Graphs and Hypergraphs* by Berge [2]; *Graphs and Digraphs* by Behzad, Chartrand, and Lesniak-Foster [3]; *Distance in Graphs* by Buckley and Harary [4]; *Graph Theory Applications* by Foulds [5]; *Introduction to Graph Theory* by West [6]; *Graph Theory* by Diestel [7]; and *Topics in Algebraic Graph Theory* by Beineke and Wilson [8]. The spectral theory of graphs investigates the properties of the spectra (eigenvalues) of graph matrices, and has applications in complex networks, spectral embedding of multivariate data, graph drawing, calculation of topological indices, topological quantum chemistry, and aromaticity. The major textbook in the spectral theory of graphs is *Spectra of Graphs. Theory and Applications* by Cvetković, Doob, and Sachs [9]. An influential book on graph spectra applications in the quantum chemistry of conjugated systems and aromaticity is *Topological Approach to the Chemistry of Conjugated Molecules* by Graovac, Gutman, and Trinajstić [10]. Advanced topics

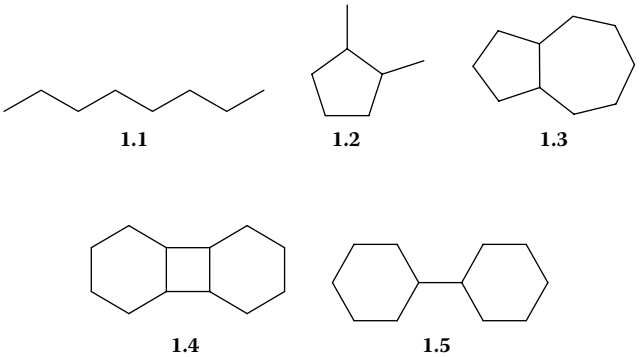
of topological aromaticity are treated in *Kekulé Structures in Benzenoid Hydrocarbons* by Cyvin and Gutman [11]; *Introduction to the Theory of Benzenoid Hydrocarbons* by Gutman and Cyvin [12]; *Advances in the Theory of Benzenoid Hydrocarbons* by Gutman and Cyvin [13]; *Theory of Coronoid Hydrocarbons* by Cyvin, Brunvoll, and Cyvin [14]; and *Molecular Orbital Calculations Using Chemical Graph Theory* by Dias [15]. The graph theoretical foundation for the enumeration of chemical isomers is presented in several books: *Graphical Enumeration* by Harary and Palmer [16]; *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds* by Pólya and Read [17]; and *Symmetry and Combinatorial Enumeration in Chemistry* by Fujita [18]. A comprehensive history of graph theory can be found in the book *Graph Theory 1736–1936* by Biggs, Lloyd, and Wilson [19].

The first edited book on chemical graphs is *Chemical Applications of Graph Theory* by Balaban [20]. Several comprehensive textbooks on chemical graphs are available, such as *Chemical Graph Theory* by Trinajstić [21], *Mathematical Concepts in Organic Chemistry* by Gutman and Polansky [22], and *Handbook of Chemoinformatics* by Gasteiger [23]. Applications of topological indices in quantitative structure–activity relationships (QSAR) are presented in *Molecular Connectivity in Chemistry and Drug Research* by Kier and Hall [24], *Molecular Connectivity in Structure–Activity Analysis* by Kier and Hall [25], *Molecular Structure Description. The Electrotopological State* by Kier and Hall [26], *Information Theoretic Indices for Characterization of Chemical Structure* by Bonchev [27], and *Topological Indices and Related Descriptors in QSAR and QSPR* by Devillers and Balaban [28]. A comprehensive text on reaction graphs is *Chemical Reaction Networks. A Graph-Theoretical Approach* by Temkin, Zeigarnik, and Bonchev [29], and a graph-theoretical approach to organic reactions is detailed in *Synthon Model of Organic Chemistry and Synthesis Design* by Koča et al. [30]. Graph algorithms for drug design are presented in *Logical and Combinatorial Algorithms for Drug Design* by Golender and Rozenblit [31]. Graph theory concepts relevant to chemoinformatics are introduced in this section, together with examples of graphs and graph matrices.

1.2.1 Graphs

A *graph* $G(V, E)$ is an ordered pair consisting of a vertex set $V(G)$ and an edge set $E(G)$. Each element $\{i, j\} \in E$ (where $i, j \in V$) is said to be an edge joining vertices i and j. Because each edge is defined by an unordered pair of vertices from V, the edge from vertex i to vertex j is identical with the edge from vertex j to vertex I, $\{i,j\} = \{j, i\}$. The number of vertices N defines the order of the graph and is equal to the number of elements in $V(G)$, $N = |V(G)|$, and the number of edges M is equal to the number of elements in $E(G)$, $M = |E(G)|$. Several examples of graphs relevant to chemistry are shown in Graphs **1.1** through **1.5**.

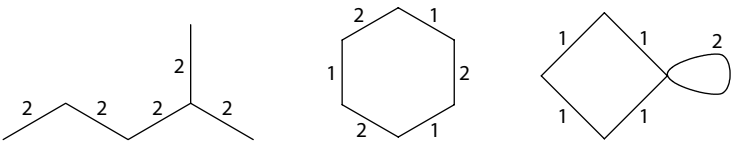
Vertices and edges in a graph may be labeled. A vertex with the label i is indicated here as v_i. An edge may be denoted by indicating the two vertices that define that edge. For example, the edge connecting vertices v_i and v_j may be denoted by e_{ij}, $e_{i,j}$, $\{i,j\}$, or $v_i v_j$. Usually, graph vertices are labeled from 1 to N, $V(G) = \{v_1, v_2, \ldots, v_N\}$, and graph edges are labeled from 1 to M, $E(G) = \{e_1, e_2, \ldots, e_M\}$. There is no special rule in labeling graphs, and a graph with N vertices may be labeled in $N!$ different ways.



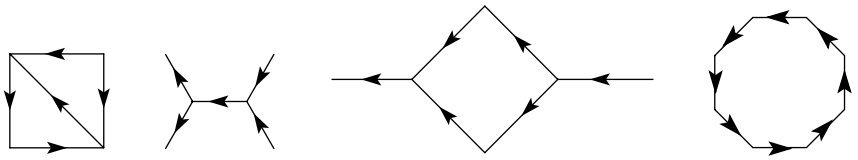
A *graph invariant* is a number, sequence of numbers, or matrix computed from the graph topology (information contained in the V and E sets) that is not dependent on the graph labeling (the graph invariant has the same value for all $N!$ different labelings of the graph). Two obvious graph invariants are the number of vertices N and the number of edges M. Other invariants of molecular graphs are topological indices, which are used as structural descriptors in quantitative structure–property relationships (QSPR), QSAR, and virtual screening of chemical libraries (cf. Chapters 4 and 5).

Graphs that have no more than one edge joining any pair of vertices are also called *simple graphs*. A *multigraph* is a graph in which two vertices may be connected by more than one edge. A *multiedge* of multiplicity m is a set of m edges that connects the same pair of distinct vertices. A *loop* $e_{ii} \in E$ is an edge joining a vertex v_i with itself. A loopgraph is a graph containing one or more vertices with loops.

Simple graphs cannot capture the complexity of real life systems, such as electrical circuits, transportation networks, production planning, kinetic networks, metabolic networks, or chemical structures. In such cases it is convenient to attach weights to vertices or loops, weights that may represent current intensity, voltage, distance, time, material flux, reaction rate, bond type, or atom type. A graph $G(V, E, w)$ is a *weighted graph* if there exists a function $w : E \rightarrow R$ (where R is the set of real numbers), which assigns a real number, called weight, to each edge of E. Graph **1.6** has all edge weights equal to 2, whereas in Graph **1.7** the edge weights alternate between 1 and 2. In the loopgraph **1.8** all edges have the weight 1 and the loop has the weight 2. Alkanes and cycloalkanes are represented as molecular graphs with all edges having a weight equal to 1, whereas chemical compounds containing heteroatoms or multiple bonds are represented as vertex- or edge-weighted molecular graphs. Section 1.4 reviews in detail the representation of chemical compounds with weighted graphs.



In many graph models, such as those of kinetic, metabolic, or electrical networks, it is useful to give each edge a direction or orientation. The graphs used to model such oriented systems are termed *directed graphs* or *digraphs*. A graph $D(V, A)$ is an ordered pair consisting of two sets $V(D)$ and $A(D)$, where the vertex set V is finite and nonempty and the arc set A is a set of ordered pairs of distinct elements from V. Graphs **1.9** through **1.12** are several examples of digraphs. A comprehensive overview of reaction graphs is presented by Balaban [32], and graph models for networks of chemical reactions are reviewed by Temkin et al. [29].

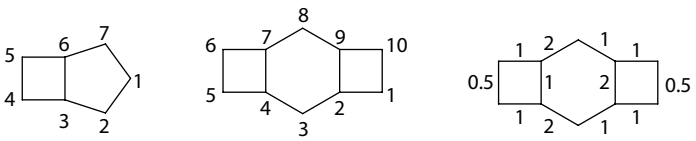


1.9 **1.10** **1.11** **1.12**

1.2.2 Adjacency, Walks, Paths, and Distances

Two vertices v_i and v_j of a graph G are adjacent (or neighbors) if there is an edge e_{ij} joining them. The two adjacent vertices v_i and v_j are said to be incident to the edge e_{ij}. The neighborhood of a vertex v_i is represented by the set of all vertices adjacent to v_i. Two distinct edges of G are adjacent if they have a vertex in common.

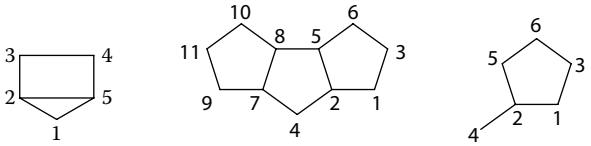
The degree of a vertex v_i, denoted by $\deg_i$, is equal to the number of vertices adjacent to vertex v_i. The set of degree values for all vertices in a graph gives the vector $\mathrm{Deg}(G)$ whose ith element represents the degree of the vertex v_i. In a weighted graph $G(V, E, w)$, the valency of a vertex v_i, $\mathrm{val}(w, G)_i$, is defined as the sum of the weights of all edges e_{ij} incident with vertex v_i [33,34]. The set of valencies for all vertices in a graph forms the vector $\mathrm{Val}(w, G)$ whose ith element represents the valency of the vertex v_i. From the definition of degree and valency it is obvious that in simple, nonweighted graphs, the degree of a vertex v_i, $\deg_i$, is identical to the valency of that vertex, val_i. Consider the simple labeled graph **1.13**. A simple count of the neighbors for each vertex in **1.13** gives the degree vector $\mathrm{Deg}(\mathbf{1.13}) = \{2, 2, 3, 2, 2, 3, 2\}$. The second example considers a weighted graph with the labeling given in Graph **1.14** and with the edge weights indicated in **1.15**. The degree vector of **1.14** is $\mathrm{Deg}(\mathbf{1.14}) = \{2, 3, 2, 3, 2, 2, 3, 2, 3, 2\}$, and the valency vector is $\mathrm{Val}(\mathbf{1.14}) = \{1.5, 4, 3, 4, 1.5, 1.5, 4, 3, 4, 1.5\}$. Both degree and valency are *graph invariants*, because their numerical values are independent of the graph labeling.



1.13 **1.14** **1.15**

A *walk* W in a graph G is a sequence of vertices and edges $W(G) = \{v_a, e_{ab}, v_b, e_{bc}, v_c, e_{cd}, v_d, e_{de}, v_e, \ldots, v_i, e_{ij}, v_j, \ldots, v_m, e_{mn}, v_n\}$ beginning and ending with vertices, in which two consecutive vertices v_i and v_j are adjacent, and each edge e_{ij} is incident with the two vertices v_i and v_j preceding and following it, respectively. A walk may also be defined as a sequence of vertices $W(G) = \{v_a, v_b, \ldots, v_n\}$ in which two consecutive vertices v_i and v_{i+1} are adjacent. Similarly, a walk may be defined as a sequence of edges $W(G) = \{e_{ab}, e_{bc}, \ldots, e_{mn}\}$ in which two consecutive edges e_{ij} and e_{jk} are adjacent. In a walk any edge of the graph may appear more than once. The length of a walk is equal to the total number of edges that define the walk. A walk in which the initial and the terminal vertices coincide is called a closed walk. A walk in which the initial and the terminal vertices are different is called an open walk. A *trail* is a walk in which no edge is repeated. A certain vertex may appear more than once in a trail, if the trail intersects itself. A *path* P is a walk in which all vertices (and thus necessarily all edges) are distinct. The length of a path in a graph is equal to the number of edges along the path.

A *graph cycle* or *circuit* is a closed walk in which all vertices are distinct, with the exception of the initial and terminal vertices that coincide. In Graph **1.16** there are three cycles: $C_1(\mathbf{1.16}) = \{v_1, v_2, v_5, v_1\}$, with length three; $C_2(\mathbf{1.16}) = \{v_1, v_2, v_3, v_4, v_5, v_1\}$, with length five; and $C_3(\mathbf{1.16}) = \{v_2, v_3, v_4, v_5, v_2\}$, with length four. In Graph **1.17** there are three cycles of length five: $C_1(\mathbf{1.17}) = \{v_1, v_2, v_5, v_6, v_3, v_1\}$, $C_2(\mathbf{1.17}) = \{v_2, v_4, v_7, v_8, v_5, v_2\}$, and $C_3(\mathbf{1.17}) = \{v_7, v_9, v_{11}, v_{10}, v_8, v_7\}$.



1.16 **1.17** **1.18**

The *cyclomatic number* μ represents the number of cycles in the graph, $\mu = M - N + 1$. For Graph **1.16** we have $\mu(\mathbf{1.16}) = 6 - 5 + 1 = 2$, for Graph **1.17** we have $\mu(\mathbf{1.17}) = 13 - 11 + 1 = 3$, and for Graph **1.18** we have $\mu(\mathbf{1.18}) = 6 - 6 + 1 = 1$.

In a simple (nonweighted) connected graph, the *graph distance* d_{ij} between a pair of vertices v_i and v_j is equal to the length of the shortest path connecting the two vertices (i.e., the number of edges of the shortest path). The distance between two adjacent vertices is 1. The graph distance satisfies the properties of a metric:

a. The distance from a vertex v_i to itself is zero:

$$d_{ii} = 0, \quad \text{for all } v_i \in V(G). \tag{1.1}$$

b. The distance between two distinct vertices v_i and v_j is larger than 0:

$$d_{ij} > 0, \quad \text{for all } v_i, v_j \in V(G). \tag{1.2}$$

c. The distance between two distinct vertices v_i and v_j is equal to the distance on the inverse path, from v_j and v_i:

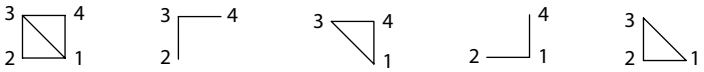
$$d_{ij} = d_{ji}, \quad \text{for all } v_i, v_j \in V(G). \tag{1.3}$$

d. The graph distance satisfies the triangle inequality:

$$d_{ik} + d_{kj} \geq d_{ij}, \quad \text{for all } v_i, v_j, v_k \in V(G). \tag{1.4}$$

The *eccentricity* $\mathrm{ecc}(v_i)$ of a vertex v_i is the maximum distance from the vertex v_i to any other vertex v_j in graph G, that is, max [35] for all $v_j \in V(G)$. The *diameter* **diam**(G) of a graph G is the maximum eccentricity. If the graph G has cycles, then the *girth* of G is the length of a shortest cycle, and the *circumference* is the length of a longest cycle.

A graph G may be transformed into a series of *subgraphs* of G by deleting one or more of its vertices, or by deleting one or more of its edges. If $V(G')$ is a subset of $V(G)$, $V(G') \subseteq V(G)$, and $E(G')$ is a subset of $E(G)$, $E(G') \subseteq E(G)$, then the subgraph $G' = (V(G'), E(G'))$ is a subgraph of the graph $G = (V(G), E(G))$. A subgraph $G - v_i$ is obtained by deleting from G the vertex v_i and all its incident edges. A subgraph $G - e_{ij}$ is obtained by deleting from G the edge e_{ij}. Graph **1.19** has four subgraphs of the type $G - v_i$, **1.20** through **1.23**, which are obtained by deleting, in turn, one vertex and all its incident edges from Graph **1.19**.

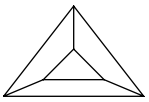


1.19 **1.20** **1.21** **1.22** **1.23**

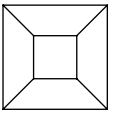
1.2.3 Special Graphs

A *tree*, or an *acyclic graph*, is a connected graph that has no cycles (the cyclomatic number $\mu = 0$). Alternative definitions for a tree are the following: a tree is a connected graph with N vertices and $N-1$ edges; a tree is a graph with no cycles, N vertices, and $N-1$ edges. A graph that contains as components only trees is a *forest*. A k-tree is a tree with the maximum degree k. Alkanes are usually represented as 4-trees. A *rooted tree* is a tree in which one vertex (the root vertex) is distinct from the other ones.

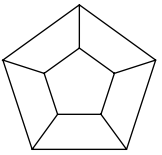
A graph with the property that every vertex has the same degree is called a *regular graph*. A graph G is called a k-regular graph or a regular graph of degree k if every vertex from G has the degree k. A ring R_N with N vertices is a 2-regular graph with N vertices, that is, a graph with all vertices of degree 2. The cycloalkanes cyclopropane, cyclobutane, cyclopentane, cyclohexane, cycloheptane, and cyclooctane are examples of 2-regular graphs. The 3-regular graphs, or *cubic graphs*, **1.24** through **1.27**, represent as molecular graphs the polycyclic hydrocarbons triprismane, tetraprismane (cubane), pentaprismane, and hexaprismane, respectively. Fullerenes are also represented as cubic graphs.



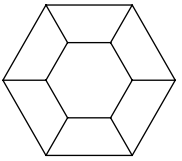
1.24



1.25



1.26



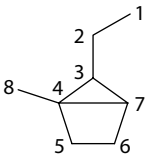
1.27

1.2.4 Graph Matrices

A graph is completely determined by indicating its adjacency relationships or its incidence relationships. However, the algebraic properties of graphs are easier studied by representing a graph as a matrix, such as adjacency matrix, incidence matrix, cycle matrix, path matrix, Laplacian matrix, distance matrix, and detour matrix. Graph matrices of chemical systems are used to investigate the spectral properties of molecular graphs [9], to apply the Hückel molecular orbitals method to conjugated molecules [10], to compute various topological indices for QSAR models [36,37], and to study the topology of biological networks [38]. In presenting graph matrices we consider only labeled, connected, simple graphs.

1.2.4.1 Adjacency Matrix

The *adjacency matrix* $\mathbf{A}(G)$ of a vertex labeled graph G with N vertices is a square $N \times N$ symmetric matrix in which $[\mathbf{A}]_{ij} = 1$ if vertex v_i is adjacent to vertex v_j and $[\mathbf{A}]_{ij} = 0$ otherwise. The adjacency matrix is symmetric, with all elements on the main diagonal equal to zero. The sum of entries over row i or column i in $\mathbf{A}(G)$ is the degree of vertex v_i, $\deg_i$. As an example we consider Graph **1.28** labeled from 1 to 8 and its adjacency matrix $\mathbf{A}(\mathbf{1.28})$.



1.28

$$\mathbf{A}(\mathbf{1.28}) = \begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 7 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 8 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline \end{array}$$

From the definition of the adjacency matrix, it follows that if $[\mathbf{A}]_{ij} = 1$ then there is a walk of length one between vertices v_i and v_j. Higher powers of the adjacency matrix can be used to count the number of closed or open walks of a certain length between two vertices. The element $[\mathbf{A}^k]_{ij}$ of the kth power of the adjacency matrix $\mathbf{A}$ is the number of walks of length k between vertices v_i and v_j [1]. If $i = j$ then the element $[\mathbf{A}^k]_{ii}$ is the number of closed walks of length k that start and end at the same vertex v_i. Similarly, when $i \neq j$, the element $[\mathbf{A}^k]_{ij}$ is the number of open walks of length k starting from vertex v_i and ending at vertex v_j. Because the kth power of the adjacency matrix is symmetric, it follows that the number of walks of length k from v_i to v_j is equal to the number of walks of length k from v_j to v_i, that is, $[\mathbf{A}^k]_{ij} = [\mathbf{A}^k]_{ji}$. $\mathbf{A}^k$ matrices can also be used to determine the distances between vertices in simple graphs. If in a sequence of $\mathbf{A}^k$ matrices all elements $[\mathbf{A}^{k-1}]_{ij} = 0$ and $[\mathbf{A}^k]_{ij} \neq 0$, it follows that the distance between vertices v_i and v_j is k (the two vertices are separated by k edges). A general procedure for computing graph distances, which can be applied to general graphs, is presented in the section on the distance matrix.

Randić suggested the use of the closed walk counts of different lengths originating from a vertex to describe the environment of that vertex [39]. He defined the closed walk atomic code of the vertex v_i, CWAC_i, as the sequence $\{[\mathbf{A}^1]_{ii}, [\mathbf{A}^2]_{ii}, \ldots, [\mathbf{A}^k]_{ii}, \ldots, [\mathbf{A}^N]_{ii}\}$. The count of closed walks is also related to the graph spectrum and spectral moments. The complete set of graph eigenvalues $x_1, x_2, \ldots, x_N$ of the adjacency matrix $\mathbf{A}(G)$ forms the spectrum of a graph G, $\mathrm{Sp}(\mathbf{A},G) = \{x_i, i = 1, 2, \ldots, N\}$. The kth spectral moment of $\mathbf{A}(G)$, $\mathrm{SM}(\mathbf{A}, G)_k$, is defined as the sum of the kth power of $\mathrm{Sp}(\mathbf{A}, G)$. Finally, the sum of the diagonal elements of $\mathbf{A}^k$ (the trace of the kth power of the adjacency matrix which is equal to the count of closed walks of length k) equals $\mathrm{SM}(\mathbf{A}, G)_k$. Spectral moments represent a powerful theoretical tool in correlating structural features with various properties of chemical systems. Burdett used spectral moments to estimate the electronic properties of solids [40,41]. Spectral moments of conjugated compounds are correlated with the presence of certain subgraphs [42–44], thus making possible the calculation of the resonance energy per electron (REPE) from subgraph contributions [42]. A similar approach was proposed by Schmalz, Živković, and Klein for the decomposition of the π-electron energy of conjugated acyclic hydrocarbons in terms of various substructures [45].

1.2.4.2 Laplacian Matrix

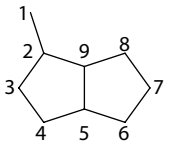
Consider a simple graph G with N vertices and M edges, and its adjacency matrix $\mathbf{A}(G)$. We define the diagonal matrix $\mathbf{DEG}(G)$ with the diagonal elements $[\mathbf{DEG}]_{ii} = \deg_i$ (the degree of vertex v_i) and with the nondiagonal elements $[\mathbf{DEG}]_{ij} = 0, i \neq j$. The *Laplacian matrix* of the simple graph G, $\mathbf{L}(G)$, is the difference between $\mathbf{DEG}$ and $\mathbf{A}$ [46–48]:

$$\mathbf{L}(G) = \mathbf{DEG}(G) - \mathbf{A}(G). \tag{1.5}$$

The most significant chemoinformatics applications of the Laplacian matrix are in computing topological indices [48,49], defining the resistance distance matrix [50], and interpolating QSAR models based on molecular networks [51–54].

1.2.4.3 Distance Matrix

The *distance matrix* $\mathbf{D}(G)$ of a simple graph G with N vertices is a square $N \times N$ symmetric matrix in which $[\mathbf{D}]_{ij} = d_{ij}$, where d_{ij} is the distance between vertices v_i and v_j, that is, the length of the shortest path that connects vertices v_i and v_j [1,4]. The distance matrix is symmetric, with all elements on the main diagonal equal to zero. Applications of the distance matrix to chemical graphs may be found in several reviews [37,55]. As an example we consider Graph **1.29** labeled from 1 to 9 and its distance matrix $\mathbf{D}(\mathbf{1.29})$.



1.29

$$\mathbf{D}(\mathbf{1.29}) = \begin{array}{c|ccccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 1 & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 3 & 2 \\ 2 & 1 & 0 & 1 & 2 & 2 & 3 & 3 & 2 & 1 \\ 3 & 2 & 1 & 0 & 1 & 2 & 3 & 4 & 3 & 2 \\ 4 & 3 & 2 & 1 & 0 & 1 & 2 & 3 & 3 & 2 \\ 5 & 3 & 2 & 2 & 1 & 0 & 1 & 2 & 2 & 1 \\ 6 & 4 & 3 & 3 & 2 & 1 & 0 & 1 & 2 & 2 \\ 7 & 4 & 3 & 4 & 3 & 2 & 1 & 0 & 1 & 2 \\ 8 & 3 & 2 & 3 & 3 & 2 & 2 & 1 & 0 & 1 \\ 9 & 2 & 1 & 2 & 2 & 1 & 2 & 2 & 1 & 0 \end{array}$$

In a simple graph, the distances between one vertex and all other vertices may be computed with the algorithm proposed by Dijkstra [35], which may also be applied to graphs with non-negative edge weights. Unlike the Dijkstra algorithm, the Floyd–Warshall algorithm [56,57] may be applied to graphs that have some edges with negative weights, as long as all cycle weights are non-negative.

ALGORITHM 1.1 FLOYD–WARSHALL

01. Consider the labeled, weighted graph G with N vertices, M edges, the vertex set $V(G)$, the edge set $E(G)$, and with a weight w_{ij} for each edge $e_{ij} \in E(G)$.
02. Define the cost matrix ${}^1\mathrm{Co} = {}^1\mathrm{Co}(G)$ of the labeled graph G as the square $N \times N$ symmetric matrix in which $[{}^1\mathrm{Co}]_{ii} = 0, [{}^1\mathrm{Co}]_{ij} = w_{ij}$ if $e_{ij} \in E(G)$, and $[{}^1\mathrm{Co}]_{ij} = \infty$ otherwise.
03. For each $k \in \{1, 2, \ldots, N\}$ do

```
04.     For each i ∈ {1,2,...,N} do
05.        For each j ∈ {1,2,...,N} do
06.          Update the cost matrix ^kCo:
             [Co]_ij = min{[^(k-1)Co]_ij, [^(k-1)Co]_ik + [^(k-1)Co]_kj}
07. End do
08. D = ^NCo
```

Step 06 in the Floyd–Warshall algorithm is based on the triangle inequality mentioned in Equation 1.4. If a graph contains cycles with negative weights, then the cost matrix Co has some negative numbers on the main diagonal. If $\mathbf{Co}_{ii} < 0$, then the vertex v_i belongs to at least one cycle with negative weight. The distance matrix is used to compute many important topological indices, such as Wiener index W [58], Balaban index J [59,60], Kier–Hall electrotopological indices [26,61], information theory indices [62], and molecular path code indices [63]. The distance matrix is the source of several molecular matrices [37,64], namely the reciprocal distance matrix [65], the distance-valency matrix [33], the distance complement matrix [66], the reverse Wiener matrix [67], the distance-path matrix [68,69], and the Szeged matrix [70,71]. These distance-related molecular matrices are used to compute topological indices and related graph descriptors for QSPR and QSAR.

1.3 CHEMICAL AND MOLECULAR GRAPHS

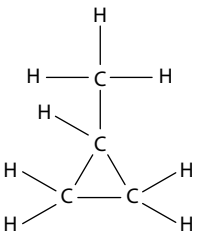
Chemical compounds are usually represented as molecular graphs, that is, nondirected, connected graphs in which vertices correspond to atoms and edges represent covalent bonds between atoms. The molecular graph model of the chemical structure emphasizes the chemical bonding pattern of atoms, whereas the molecular geometry is neglected. Among other applications, molecular graphs are used in chemoinformatics systems, chemical databases, design of combinatorial libraries, reaction databases, computer-assisted structure elucidation, molecular design of novel chemicals, and computer-assisted organic synthesis. Molecular graphs are the basis for computing the structural descriptors used in QSPR and QSAR models to predict physical, chemical, biological, or toxicological properties. The molecular graph representation of chemical structure reflects mainly the connectivity of the atoms and is less suitable for modeling those properties that are determined mostly by molecular geometry, conformation, or stereochemistry.

1.3.1 Molecular Graphs

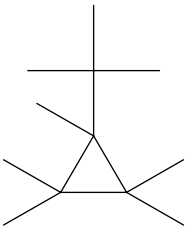
A chemical structure may be represented by a large number of different molecular graphs, depending on the translation rules for depicting atoms and chemical bonds. The translation rules, that is, "atom → vertex" and "bond → edge," should preserve the features of the molecular structure that are relevant for the scope of the modeling, for example, database search, reaction representation, molecular design, or property prediction. Cayley introduced the concept of molecular graphs in 1874, as "plerograms" and "kenograms," in which graph edges correspond to covalent bonds [72]. In

a plerogram all atoms (including hydrogen atoms) are represented as vertices, whereas in a kenogram only non-hydrogen atoms are represented, because the hydrogen atoms can be reconstructed from the skeleton of a molecule. In modern terminology a plerogram is a hydrogen-included molecular graph, and a kenogram is a hydrogen-excluded molecular graph (called also hydrogen-depleted or hydrogen-suppressed molecular graph).

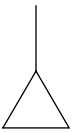
Using different rules for converting a chemical structure into a molecular graph, methylcyclopropane can be represented by Graphs **1.30**, **1.31**, and **1.32**. Graph **1.30** is a hydrogen-included molecular graph with labeled vertices, Graph **1.31** is a hydrogen-included molecular graph in which hydrogen and carbon atoms are not differentiated, and Graph **1.32** is a hydrogen-excluded molecular graph.



1.30

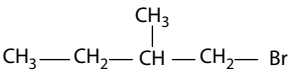


1.31

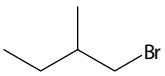


1.32

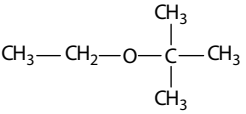
The usual graph representation of an organic chemical compound is as a nondirected, connected multigraph in which vertices correspond to non-hydrogen atoms and edges represent covalent bonds between non-hydrogen atoms. For hydrocarbons, the vertices in the molecular graph represent carbon atoms. Using this convention, alkanes are represented as 4-trees, that is, acyclic graphs with the maximum degree 4. Several studies compared structural descriptors (topological indices) computed from hydrogen-included and hydrogen-excluded molecular graphs of alkanes, and found that the topological indices are correlated [73,74]. These results support the preponderant use of hydrogen-excluded molecular graphs. To accommodate the presence of heteroatoms, a molecular graph has vertex labels corresponding to the atomic symbol of the heteroatoms, as shown for 2-methyl-1-bromobutane **1.33** (molecular graph **1.34**) and for ethyl *tert*-butyl ether **1.35** (molecular graph **1.36**).



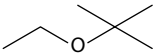
1.33



1.34

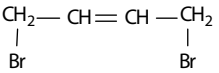


1.35



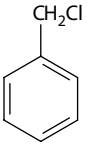
1.36

Multiple bonds are represented as multiedges, as shown for 1,4-dibromo-2-butene **1.37** (molecular graph **1.38**). Conjugated systems may be represented with the usual pattern of alternating double and single bonds, or with two lines, one continuous and the second broken, as shown for the aromatic system of benzyl chloride **1.39** (molecular graph **1.40**). The differences between these two representations of conjugated systems are significant when computing topological indices that have special parameters for aromatic bonds, and in chemical database registration, search, and retrieval.

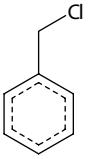


1.37

1.38



1.39



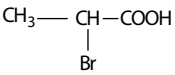
1.40

1.3.2 Molecular Pseudograph

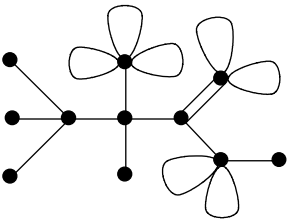
There are a multitude of molecular graph models, each one developed with a specific set of rules, and fit for particular applications, such as structure elucidation, chemical synthesis design, or structure–property relationships. Koča et al. defined a mathematical model of organic synthesis design based on the graph theory formalism [30]. In this model, a chemical compound is represented by a molecular pseudograph (or general graph, containing multiedges and loops) $G(V, E, L, \varphi, \upsilon)$, where V is a vertex set, E is an edge set, L is a loop set, and φ is a mapping of the vertex set into the vocabulary υ of vertex labels. A single bond is represented by an edge, a double bond is represented by a multiedge of double multiplicity, and a triple bond is represented by a multiedge of triple multiplicity. A pair of free, unshared valence electrons of an atom is represented as a loop. Nitrogen is represented by a vertex with a loop, oxygen is represented by a vertex with two loops, whereas a halogen atom is represented by a vertex with three loops, as shown for 2-bromopropanoic acid **1.41** (molecular graph **1.42**) and for morpholine **1.43** (molecular graph **1.44**).

1.3.3 Molecular Graph of Atomic Orbitals

Toropov introduced the molecular graph of atomic orbitals (GAO) as a source of structural descriptors for QSPR and QSAR [75–77]. GAO is based on the hydrogen-included molecular graphs, in which each atom is substituted by the corresponding set

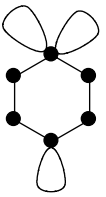


1.41



1.42

1.43



1.44

of atomic orbitals: H, $1s^1$; C, $1s^2$, $2s^2$, $2p^2$; N, $1s^2$, $2s^2$, $2p^3$; O, $1s^2$, $2s^2$, $2p^4$; F, $1s^2$, $2s^2$, $2p^5$; S, $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^4$; Cl, $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^5$; Br, $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $3d^{10}$, $4s^2$, $4p^5$. Using this convention, C is represented in GAO by three vertices, Cl is represented by five vertices, and Br is represented by eight vertices. A covalent bond between atoms i and j is represented in GAO by $n_i \times n_j$ edges between the n_i atomic orbitals of atom i and the n_j atomic orbitals of atom j. As example we show the GAO of fluorobenzene (Figure 1.1). Another example of atomic orbitals graphs are the molecular graphs proposed by Pogliani, based on the hydrogen-excluded pseudograph augmented with information regarding the inner-core electrons [78–82].

1.3.4 Markush Structures

A major branch of chemoinformatics is represented by the development of efficient algorithms for the computer storage and retrieval of generic chemical structures. Using special topological representations, generic chemical structures encode into a single chemical graph an entire family of structurally related compounds. Among the different generic chemical structure representations, Markush structures have a special place because of their use in representing generic structures in patents. In a 1925

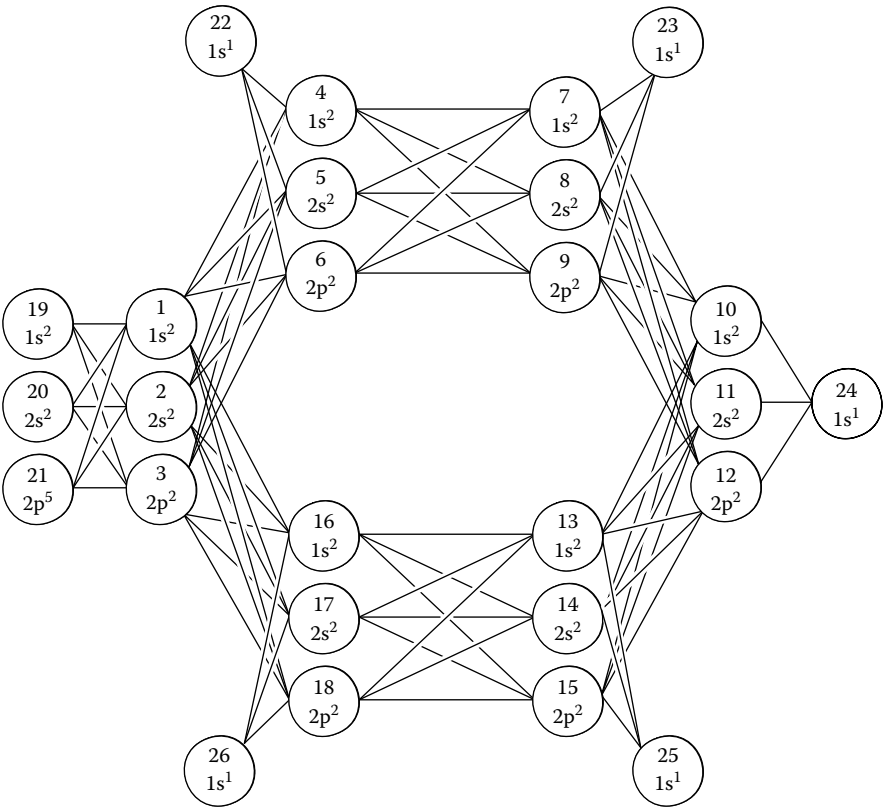
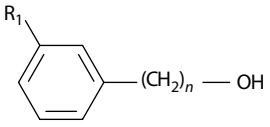
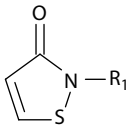


FIGURE 1.1 GAO of fluorobenzene.

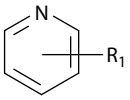
court case Eugene Markush put forward such structures, which were later accepted in patent claims by the US Patent Office. Several approaches for the implementation of Markush structures are in use [83]. Among them, the Chemical Abstracts Service [84,85] and the Questel.Orbit [86] systems are more prominent. Markush structures **1.45** through **1.47** represent several examples of generic chemical structures.



1.45



1.46



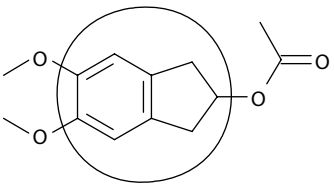
1.47

The Sheffield University group led by Lynch [87,88] developed graph representations for generic chemical structures, together with the GENSAL language [89] that is used to encode patent information into a computer-readable form [90]. The system developed by Lynch is a comprehensive collection of algorithms and procedures for the utilization of generic chemical structures: connection table representation [91], generation of fragment descriptors [92–94], computer interpreter for GENSAL [95,96], substructure search algorithm [97], reduced chemical graphs [98,99], algorithm to find the extended set of smallest rings [100], chemical ring identification [101], chemical graph search [102,103], and atom-level structure matching [104].

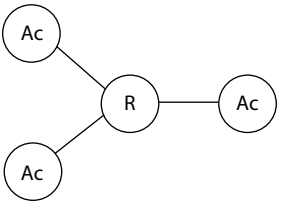
1.3.5 Reduced Graph Model

A more abstract representation of chemical structures is achieved with reduced graphs, in which each vertex represents a group of connected atoms, and an edge links two such vertices if in the original molecule there is a bond between an atom within one group and an atom in the second group [98,99]. A vertex in a reduced graph may represent a ring system, aromatic rings, aliphatic rings, or functional groups. There are several systems to transform a molecule into a reduced graph, by highlighting and grouping together different substructures in a chemical compound. We demonstrate here four types of reduced graphs that start from the same molecular graph and end up with different simplified representations.

Type 1. Vertices in the reduced graph correspond to ring systems (R) and connected acyclic components (Ac). The ring system R from compound **1.48** (shown inside a circle) corresponds to the central vertex in the reduced graph **1.49**.

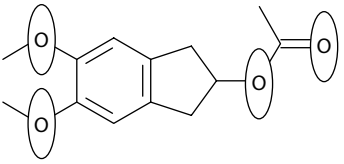


1.48

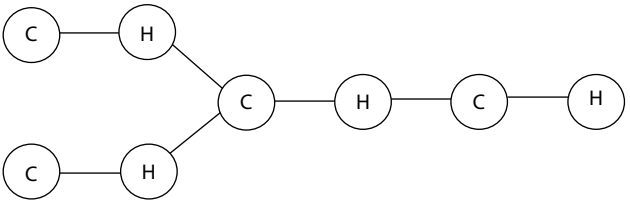


1.49

Type 2. Vertices in the reduced graph correspond to connected carbon components (C) and connected heteroatom components (H). Each heteroatom component in **1.50** is depicted inside an ellipse, and the corresponding reduced graph is shown in **1.51**.

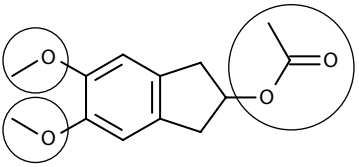


1.50

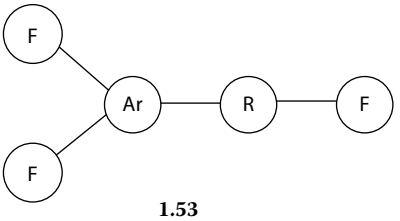


1.51

Type 3. Vertices in the reduced graph correspond to aromatic rings (Ar), aliphatic rings (R), and functional groups (F). Each functional group from molecular graph **1.52** is depicted inside a circle, with the final reduced graph depicted in **1.53**.



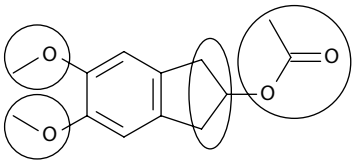
1.52



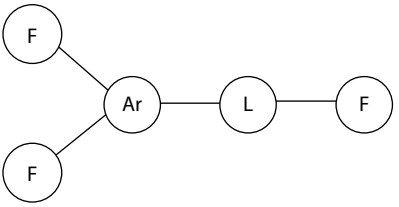
1.53

Type 4. Vertices in the reduced graph correspond to aromatic rings (Ar), functional groups (F), and linking groups (L). Each functional group from molecular graph **1.54** is depicted inside a circle, and the linker group is shown inside an ellipse. The corresponding reduced graph **1.55** has the same topology as reduced graph **1.53**, but with a different fragment type for the vertex between vertices labeled Ar and F.

When using a reduced graph to screen chemical libraries, different molecules may generate the same reduced graph, thus clustering together chemicals that have the same topological distribution of various types of subgraphs. The value of this approach is given by the fact that chemicals with similar bioactivities are translated into identical or highly similar reduced graphs. Several experiments show that reduced



1.54



1.55

graphs may identify bioactive compounds that are missed with a fingerprint similarity search [105–108]. As expected, across a large spectrum of bioactivities, there is no definite advantage of using only reduced graphs, but these studies demonstrate the complementary nature of reduced graph similarity compared to fingerprint similarity.

1.3.6 Molecule Superposition Graphs

The molecular alignment of chemicals in a QSAR dataset is a characteristic of three-dimensional (3D) QSAR models. Similarly, the topological information encoded into the molecular graph may be used to obtain a 2D alignment of all molecules in a QSAR dataset. Such a molecule superposition graph, which is obtained from structurally related compounds by superposing the molecules according to a set of rules, may be considered as a supermolecule with the property that any molecule in the QSAR dataset is its subgraph. An early 2D alignment model is represented by the DARC (description, acquisition, retrieval, correlation) system, which applies the supermolecule approach by considering that molecules are composed of a common skeleton and a variable collection of substituents [109–114]. The contribution of the variable part of the structure to the overall property value of a molecule is determined by regression analysis to predict various physical, chemical, and biological properties.

An example of a DARC supermolecule is demonstrated for the prediction of ^{13}C nuclear magnetic resonance (NMR) chemical shift in acyclic alkenes [113]. In Figure 1.2, the topo-stereochemical description of the environment of the α-sp^2 resonating carbon atom considers all sp^3-hybridized carbon neighbors of types A, B, C, and D situated at 1, 2, 3, and 4 bonds away from the resonating atom. The use of an environment with a larger sphere of atoms does not add much information because the influence on the chemical shift of atoms situated at a distance greater than four bonds can be neglected. In a DARC supermolecule some sites collect a group of atoms that have similar influence on the modeled property, such as site ΣC that collects all carbon atoms situated three bonds away from C*, and site ΣD that collects all carbon atoms situated four bonds away from C.

Simon developed the minimal topological difference (MTD) QSAR model by superposing all molecules from the training set into a supermolecule [115]. Special vertices and edges are then created to embed the substituents by maximizing the superposition of their non-hydrogen atoms, and each molecule is embedded in a unique way into the MTD supermolecule. The MTD map has three types of vertices, namely with a positive contribution (increasing the bioactivity), with a negative contribution (decreasing the bioactivity), and neutral (no influence on the bioactivity). The type

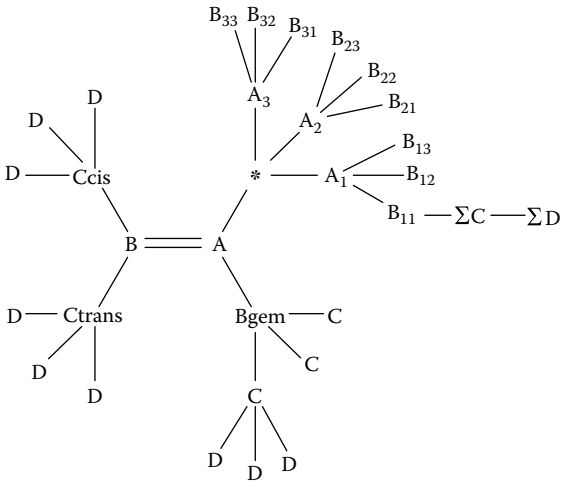


FIGURE 1.2 DARC-type map for the topo-stereochemical environment of α-sp^2 carbon atoms. The ^{13}C NMR chemical shift is predicted for the carbon atom labeled with *.

of each site in the MTD map is determined in an iterative process by embedding the training molecules on the MTD supermolecule and by minimizing the regression error between the experimental and calculated bioactivity. Minailiuc and Diudea extended the MTD supermolecule method by assigning vertex structural descriptors to vertices from the MTD supermolecule that are occupied for a particular molecule [116]. This QSAR model, called topological indices-minimal topological difference (TI-MTD), is very versatile in modeling QSAR properties and can be extended to other atomic properties, such as atomic charge or electronegativity. Recent studies show that the MTD method may be improved by using partial least squares (PLS) instead of multiple regression [117,118].

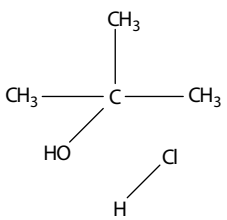
A similar supermolecule is generated in the molecular field topology analysis (MFTA) model introduced by Palyulin et al. [119]. The atomic descriptors associated with each vertex of the MFTA map are atomic charge, electronegativity, van der Waals radius, and atomic contribution to lipophilicity. The contribution of each site is determined with PLS.

1.3.7 Reaction Graphs

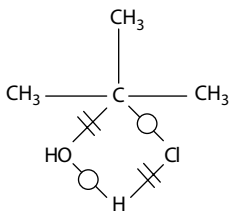
The utilization of reaction databases relies heavily on efficient software for storage and retrieval of reactions and reaction substructure search. Although very useful in suggesting individual reaction steps, reaction databases offer little help in devising strategies for complex reactions. A major accomplishment of chemoinformatics is the development of computer-assisted synthesis design systems and reaction prediction systems (cf. Chapter 11).

The storage and retrieval of reactions in databases, the extraction of reactivity knowledge, computer-assisted synthesis design, and reaction prediction systems are

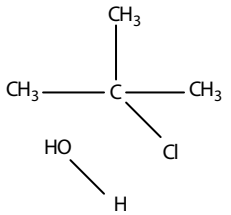
usually based on chemoinformatics tools that represent chemical reactions as a special type of graph [120–122]. As an example we present here the imaginary transition structure (ITS) model proposed by Fujita [120,123,124]. The ITS is a special type of reaction graph that is obtained by superposing reagents and products, and in which the bond rearrangement is indicated with special symbols. The reaction graph of an ITS has three types of bonds: par-bonds, which are bonds that are not modified in the reaction; out-bonds, representing bonds that are present only in reagents; and in-bonds, which are bonds appearing only in products. The diagram of an ITS graph contains distinctive symbols for each bond type: par-bonds are shown as solid lines; out-bonds are depicted as solid lines with a double bar; and in-bonds are depicted as solid lines with a circle. The ITS model is demonstrated here for nucleophilic substitution, with reactants **1.56**, ITS **1.57**, and products **1.58**.



1.56



1.57



1.58

As can be seen from the above reaction, in which *tert*-butyl alcohol reacts with hydrogen chloride to generate *tert*-butyl chloride, reaction mechanism details are not encoded into ITS. The role of ITS is to describe only bond rearrangements that transform reactants into products. The ITSs are not intended to represent reaction mechanisms, but the definition of the ITS may be easily extended to encode them.

The ITS reaction graphs represent a comprehensive framework for the classification and enumeration of organic reactions. The storage and retrieval of chemical reactions are reduced to graph manipulations, and the identification of a reaction type is equivalent to a subgraph search of an ITS database. A unique numerical representation (canonical code) of an ITS can be easily obtained [125,126] with a procedure derived from the Morgan algorithm of canonical coding [127]. The canonical representation of ITS graphs is an effective way of searching and comparing chemical reactions and of identifying reaction types.

1.3.8 Other Chemical Graphs

Many molecular graph models cannot handle systems with delocalized electrons, such as diborane or organometallic complexes, and several special graph models were proposed to encode these systems. Stein extended the bond and electron (BE) matrices introduced by Dugundji and Ugi [128–130] with new bond types for delocalized electrons [131]. Konstantinova and Skorobogatov proposed molecular hypergraphs to depict delocalized systems [132]. Dietz developed a molecular representation for computer-assisted synthesis design systems and for chemical database systems [133].

This molecular representation encodes the constitution, configuration, and conformation of a chemical compound. The constitution is represented as a multigraph describing the unshared valence electrons and the bonding relationships in a molecule, including valence electron sharing and electrostatic interactions. The chemical model suggested by Bauerschmidt and Gasteiger defines a hierarchical organization of molecular systems, starting from the electron system and ending with aggregates and ensembles [134]. Multicenter bonds are described as a list of atoms, type (σ or π), and number of electrons. This molecular representation is implemented in the reaction prediction program elaboration of reactions for organic synthesis (EROS) [135].

Chemical graphs may also be used to model systems in which the interaction between vertices represents hydrogen bonds, especially water, which consists of a large number of locally stable structures with various arrangements of the constituent water molecules. Each water cluster $(H_2O)_n$ is represented by a graph in which vertices are water molecules and bonds represent hydrogen bonds between two water molecules. Although weaker than covalent bonds, hydrogen bonds can form long-lived structures of water clusters for which the thermodynamic properties are determined by the hydrogen bonding patterns. The number of possible configurations of a cluster $(H_2O)_n$ increases very rapidly with n, which makes the identification of all possible local minima on the potential surface of a water cluster difficult [136–139].

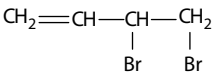
1.4 WEIGHTED GRAPHS AND MOLECULAR MATRICES

Simple graphs lack the flexibility to represent complex chemical compounds, which limits their application to alkanes and cycloalkanes, and many widely used topological indices were initially defined for such simple molecular graphs (cf. Chapter 4). The main chemical application of topological indices is that of structural descriptors in QSPR, QSAR, and virtual screening, which requires the computation of these indices for molecular graphs containing heteroatoms and multiple bonds. Such molecular graphs use special sets of parameters to represent heteroatoms as vertex weights, and multiple bonds as edge weights. Early applications of such vertex- and edge-weighted (VEW) molecular graphs were initially developed for the Hückel molecular orbitals theory [140] and were subsequently extended to general chemical compounds [141]. In this section we present selected algorithms for the computation of weighted molecular graphs that are general in scope and can be applied to a large range of structural descriptors. The application of these weighting schemes is demonstrated for a group of molecular matrices that are frequently used in computing topological indices. Other weighting schemes were proposed for more narrow applications, and are valid only for specific topological indices such as Randić–Kier–Hall connectivity indices [24,25], electrotopological indices [26,142], Burden indices [143], and Balaban index J [60].

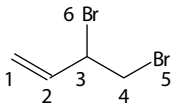
1.4.1 Weighted Molecular Graphs

A VEW molecular graph $G(V, E, Sy, Bo, Vw, Ew, w)$ is defined by a vertex set $V(G)$, an edge set $E(G)$, a set of chemical symbols for vertices $Sy(G)$, a set of topological bond orders for edges $Bo\ (G)$, a vertex weight set $Vw(w, G)$, and an edge weight set

$Ew(w, G)$, where the elements of the vertex and edge sets are computed with the weighting scheme w. Usually, the weight of a carbon atom is 0, whereas the weight of a carbon–carbon single bond is 1. In the weighting schemes reviewed here, the topological bond order Bo_{ij} of an edge e_{ij} takes the value 1 for single bonds, 2 for double bonds, 3 for triple bonds, and 1.5 for aromatic bonds. As an example of a VEW graph, consider 3,4-dibromo-1-butene **1.59** and its corresponding molecular graph **1.60**.



1.59



1.60

Graph distances represent the basis for the computation of almost all topological indices, and their computation in VEW graphs is shown here. The length of a path p_{ij} between vertices v_i and v_j, $l(p_{ij}, w, G)$, for a weighting scheme w in a VEW graph G is equal to the sum of the edge parameters $Ew(w)_{ij}$ for all edges along the path. The length of the path p_1(**1.60**) $= \{v_1, v_2, v_3, v_6\}$ is $l(p_1) = Ew_{1,2} + Ew_{2,3} + Ew_{3,6}$. The topological length of a path p_{ij}, $t(p_{ij}, G)$, in a VEW graph G is equal to the number of edges along the path, which coincides with the path length in the corresponding unweighted graph. In a VEW graph, the distance $d(w)_{ij}$ between a pair of vertices v_i and v_j is equal to the length of the shortest path connecting the two vertices, $d(w)_{ij} = \min(l(p_{ij}, w))$.

1.4.2 Adjacency Matrix

The adjacency matrix $\mathbf{A}(w, G)$ of a VEW molecular graph G with N vertices is a square $N \times N$ real symmetric matrix with the element $[\mathbf{A}(w, G)]_{ij}$ defined as [34,144]

$$[\mathbf{A}(w, G)]_{ij} = \begin{cases} Vw(w)_i & \text{if } i = j, \\ Ew(w)_{ij} & \text{if } e_{ij} \in E(G), \\ 0 & \text{if } e_{ij} \notin E(G), \end{cases} \tag{1.6}$$

where $Vw(w)_i$ is the weight of vertex v_i, $Ew(w)_{ij}$ is the weight of edge e_{ij}, and w is the weighting scheme used to compute the parameters Vw and Ew. The valency of vertex v_i, $\mathrm{val}(w,G)_i$, is defined as the sum of the weights $Ew(w)_{ij}$ of all edges e_{ij} incident with vertex v_i [49]:

$$\mathrm{val}(w)_i = \sum_{e_{ij} \in E(G)} Ew(w)_{ij}. \tag{1.7}$$

1.4.3 Distance Matrix

The distance matrix $\mathbf{D}(w, G)$ of a VEW molecular graph G with N vertices is a symmetric square $N \times N$ matrix with the element $[\mathbf{D}(w, G)]_{ij}$ defined as [144,145]

$$[\mathbf{D}(w, G)]_{ij} = \begin{cases} d(w)_{ij} & \text{if } i \neq j, \\ Vw(w)_i & \text{if } i = j, \end{cases} \tag{1.8}$$

where $d(w)_{ij}$ is the distance between vertices v_i and v_j, $Vw(w)_i$ is the weight of vertex v_i, and w is the weighting scheme used to compute the parameters Vw and Ew. The distance sum of vertex v_i, $\mathbf{DS}\,(w, G)_i$, is defined as the sum of the topological distances between vertex v_i and every vertex in the VEW molecular graph G:

$$\mathbf{DS}(w, G)_i = \sum_{j=1}^{N} [\mathbf{D}(w, G)]_{ij} = \sum_{j=1}^{N} [\mathbf{D}(w, G)]_{ji}, \tag{1.9}$$

where w is the weighting scheme. The distance sum is used to compute the Balaban index J [59] and information on distance indices [62].

1.4.4 Atomic Number Weighting Scheme *Z*

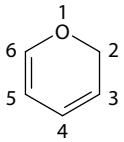
Based on the definitions of adjacency and distance matrices introduced above, we demonstrate here the calculation of molecular matrices for weighted graphs. Barysz et al. proposed a general approach for computing parameters for VEW graphs by weighting the contributions of atoms and bonds with parameters based on the atomic number Z and the topological bond order [141]. In the atomic number weighting scheme Z, the parameter $Vw(Z)_i$ of a vertex v_i (representing atom i from a molecule) is defined as

$$Vw(Z)_i = 1 - \frac{Z_C}{Z_i} = 1 - \frac{6}{Z_i}, \tag{1.10}$$

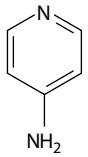
where Z_i is the atomic number Z of atom i and $Z_C = 6$ is the atomic number Z of carbon. The parameter $Ew(Z)_{ij}$ for edge e_{ij} (representing the bond between atoms i and j) is defined as

$$Ew(Z)_{ij} = \frac{Z_C Z_C}{(Bo_{ij} Z_i Z_j)} = \frac{6 \times 6}{(Bo_{ij} Z_i Z_j)}, \tag{1.11}$$

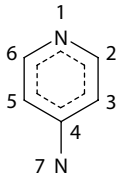
where Bo_{ij} is the topological bond order of the edge between vertices v_i and v_j. The application of the Z parameters is shown for the adjacency matrix of 2*H*-pyran **1.61** and for the distance matrix of 4-aminopyridine **1.61** (molecular graph **1.63**).



1.61



1.62



1.63

$$\mathbf{A}(Z, \mathbf{1.61}) = \begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0.250 & 0.750 & 0 & 0 & 0 & 0.750 \\ 2 & 0.750 & 0 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0.500 & 0 & 0 \\ 4 & 0 & 0 & 0.500 & 0 & 1 & 0 \\ 5 & 0 & 0 & 0 & 1 & 0 & 0.500 \\ 6 & 0.750 & 0 & 0 & 0 & 0.500 & 0 \end{array}$$

$$\mathbf{D}(Z, \mathbf{1.63}) = \begin{array}{c|ccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0.143 & 0.571 & 1.238 & 1.905 & 1.238 & 0.571 & 2.762 \\ 2 & 0.571 & 0 & 0.667 & 1.333 & 1.810 & 1.143 & 2.190 \\ 3 & 1.238 & 0.667 & 0 & 0.667 & 1.333 & 1.810 & 1.524 \\ 4 & 1.905 & 1.333 & 0.667 & 0 & 0.667 & 1.333 & 0.857 \\ 5 & 1.238 & 1.810 & 1.333 & 0.667 & 0 & 0.667 & 1.524 \\ 6 & 0.571 & 1.143 & 1.810 & 1.333 & 0.667 & 0 & 2.190 \\ 7 & 2.762 & 2.190 & 1.524 & 0.857 & 1.524 & 2.190 & 0.143 \end{array}$$

1.4.5 Relative Electronegativity Weighting Scheme *X*

The extension of the Balaban index J to VEW molecular graphs is based on relative electronegativity and covalent radius [60]. First, the Sanderson electronegativities of main group atoms are fitted in a linear regression using as parameters the atomic number Z and the number of the group Ng in the periodic system:

$$S_i = 1.1032 - 0.0204\, Z_i + 0.4121\, \mathrm{Ng}_i. \tag{1.12}$$

Taking as reference the calculated electronegativity for carbon $S_C = 2.629$, the relative electronegativities X are defined as

$$X_i = 0.4196 - 0.0078\, Z_i + 0.1567\, \mathrm{Ng}_i. \tag{1.13}$$

This weight system, developed initially for J, was extended as the relative electronegativity weighting scheme X, in which the vertex parameter $Vw(X)_i$ is defined as [36,146]

$$Vw(X)_i = 1 - \frac{1}{X_i}. \tag{1.14}$$

The edge parameter $Ew(X)_{ij}$ that characterizes the relative electronegativity of a bond is computed with the equation

$$Ew(X)_{ij} = \frac{1}{(Bo_{ij}X_iX_j)}. \tag{1.15}$$

From its definition, the weighting scheme X reflects the periodicity of electronegativity and can generate molecular descriptors that express both the effect of topology and that of electronegativity. A related set of parameters, the relative covalent radius weighting scheme Y, was defined based on the covalent radius [36,146].

1.4.6 Atomic Radius Weighting Scheme *R*

The atomic radius computed from the atomic polarizability is the basis of the atomic radius weighting scheme R, in which the vertex parameter $Vw(R)_i$ is defined as [144,147]

$$Vw(R)_i = 1 - \frac{r_C}{r_i} = 1 - \frac{1.21}{r_i} \tag{1.16}$$

and the parameter $Ew(R)_{ij}$ of the edge e_{ij} representing the bond between atoms i and j is equal to

$$Ew(R)_{ij} = \frac{r_C r_C}{(Bo_{ij}r_ir_j)} = \frac{1.21 \times 1.21}{(Bo_{ij}r_ir_j)}, \tag{1.17}$$

where $r_C = 1.21$ Å is the carbon radius and r_i is the atomic radius of atom i. Similar sets of parameters for VEW graphs were obtained with other atomic parameters, namely the atomic mass weighting scheme A, the atomic polarizability weighting scheme P, and the atomic electronegativity weighting scheme E [144,147].

1.4.7 Burden Matrix

The Burden molecular matrix is a modified adjacency matrix obtained from the hydrogen-excluded molecular graph of an organic compound [143]. This matrix is the source of the Burden, CAS, and University of Texas (BCUT) descriptors, which are computed from the graph spectra of the Burden matrix **B** and are extensively used in combinatorial chemistry, virtual screening, diversity measure, and QSAR [148–150]. An extension of the Burden matrix was obtained by inserting on the main diagonal of **B** a vertex structural descriptor VSD, representing a vector of experimental or computed atomic properties [151]. The rules defining the Burden matrix **B**(VSD, G) of a graph G with N vertices are as follows:

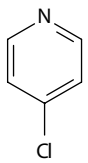
a. The diagonal elements of **B**, $[\mathbf{B}]_{ii}$, are computed with the formula

$$[\mathbf{B}(\mathrm{VSD}, G)]_{ii} = \mathrm{VSD}_i, \tag{1.18}$$

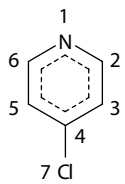
where VSD_i is a vertex structural descriptor of vertex v_i, that reflects the local structure of the corresponding atom i.

b. The nondiagonal element $[\mathbf{B}]_{ij}$, representing an edge e_{ij} connecting vertices v_i and v_j, has the value 0.1 for a single bond, 0.2 for a double bond, 0.3 for a triple bond, and 0.15 for an aromatic delocalized bond.
c. The value of a nondiagonal element $[\mathbf{B}]_{ij}$ representing an edge e_{ij} connecting vertices v_i and v_j is augmented by 0.01 if either vertex v_i or vertex v_j have degree 1.
d. All other nondiagonal elements $[\mathbf{B}]_{ij}$ are set equal to 0.001; these elements are set to 0 in the adjacency matrix **A** and correspond to pairs of nonbonded vertices in a molecular graph.

Examples of the vertex structural descriptor VSD for the diagonal of the Burden matrix are parameters from the weighting schemes A, E, P, R, X, Y, Z, various atomic properties (Pauling electronegativity, covalent radius, atomic polarizability), or various molecular graph indices, such as degree, valency, valence delta atom connectivity δ, intrinsic state I, electrotopological state S, distance sum **DS**, or vertex sum **VS**. An example of the Burden matrix is shown for 4-chloropyridine **1.64** (molecular graph **1.65**) with the Pauling electronegativity EP on the main diagonal.



1.64



1.65

$$\mathbf{B}(EP, \mathbf{1.65}) = \begin{array}{c|ccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 3.040 & 0.150 & 0.001 & 0.001 & 0.001 & 0.150 & 0.001 \\ 2 & 0.150 & 2.550 & 0.150 & 0.001 & 0.001 & 0.001 & 0.001 \\ 3 & 0.001 & 0.150 & 2.550 & 0.150 & 0.001 & 0.001 & 0.001 \\ 4 & 0.001 & 0.001 & 0.150 & 2.550 & 0.150 & 0.001 & 0.110 \\ 5 & 0.001 & 0.001 & 0.001 & 0.150 & 2.550 & 0.150 & 0.001 \\ 6 & 0.150 & 0.001 & 0.001 & 0.001 & 0.150 & 2.550 & 0.001 \\ 7 & 0.001 & 0.001 & 0.001 & 0.110 & 0.001 & 0.001 & 3.160 \end{array}$$

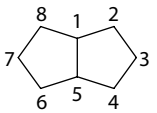
1.4.8 Reciprocal Distance Matrix

Starting with the Wiener index W, graph distances represented a prevalent source of topological indices. A possible drawback of using graph distances directly is that pairs of atoms that are separated by large distances, and thus have low interaction between them, have large contributions to the numerical value of the index. Because physical interaction between two objects decreases with increasing distance, the reciprocal distance $1/d_{ij}$ was introduced. Using the reciprocal distance, it is possible to define graph descriptors in which the contribution of two vertices decreases with increase

of the distance between them [152]. The reciprocal distance matrix of a simple graph G with N vertices $\mathbf{RD}(G)$ is a square $N \times N$ symmetric matrix whose entries $[\mathbf{RD}]_{ij}$ are equal to the reciprocal of the distance between vertices v_i and v_j, that is, $1/d_{ij} = 1/[\mathbf{D}]_{ij}$, for nondiagonal elements, and is equal to zero for the diagonal elements [65,153,154]:

$$[\mathbf{RD}(G)]_{ij} = \begin{cases} \dfrac{1}{[\mathbf{D}(G)]_{ij}} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{1.19}$$

The reciprocal distance matrix of octahydropentalene **1.66** is shown as an example.



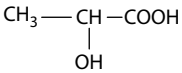
1.66

$$\mathbf{RD}(\mathbf{1.66}) = \begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 0 & 1 & 0.500 & 0.500 & 1 & 0.500 & 0.500 & 1 \\ 2 & 1 & 0 & 1 & 0.500 & 0.500 & 0.333 & 0.333 & 0.500 \\ 3 & 0.500 & 1 & 0 & 1 & 0.500 & 0.333 & 0.250 & 0.333 \\ 4 & 0.500 & 0.500 & 1 & 0 & 1 & 0.500 & 0.333 & 0.333 \\ 5 & 1 & 0.500 & 0.500 & 1 & 0 & 1 & 0.500 & 0.500 \\ 6 & 0.500 & 0.333 & 0.333 & 0.500 & 1 & 0 & 1 & 0.500 \\ 7 & 0.500 & 0.333 & 0.250 & 0.333 & 0.500 & 1 & 0 & 1 \\ 8 & 1 & 0.500 & 0.333 & 0.333 & 0.500 & 0.500 & 1 & 0 \end{array}$$

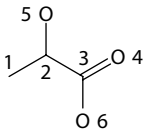
Formula 1.19 can be easily extended to weighted molecular graphs. The reciprocal distance matrix $\mathbf{RD}(w, G)$ of a VEW molecular graph G with N vertices is a square $N \times N$ symmetric matrix with real elements [144,145]:

$$[\mathbf{RD}(w, G)]_{ij} = \begin{cases} \dfrac{1}{[\mathbf{D}(w, G)]_{ij}} & \text{if } i \neq j, \\ Vw(w)_i & \text{if } i = j, \end{cases} \tag{1.20}$$

where $[\mathbf{D}(w)]_{ij}$ is the graph distance between vertices v_i and v_j, $[\mathbf{D}(w)]_{ii}$ is the diagonal element corresponding to vertex v_i, and w is the weighting scheme used to compute the parameters Vw and Ew. The reciprocal distance matrix of 2-hydroxypropanoic acid (lactic acid) **1.67** (molecular graph **1.68**) computed with the atomic electronegativity weighting scheme E is presented as an example.



1.67



1.68 *Source:* From *Encyclopedia of Chemoinformatics.* With permission.

$$\mathbf{RD}(E, \mathbf{1.68}) = \begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1 & 0.500 & 0.425 & 0.587 & 0.370 \\ 2 & 1 & 0 & 1 & 0.740 & 1.420 & 0.587 \\ 3 & 0.500 & 1 & 0 & 2.839 & 0.587 & 1.420 \\ 4 & 0.425 & 0.740 & 2.839 & 0.296 & 0.486 & 0.946 \\ 5 & 0.587 & 1.420 & 0.587 & 0.486 & 0.296 & 0.415 \\ 6 & 0.370 & 0.587 & 1.420 & 0.946 & 0.415 & 0.296 \end{array}$$

1.4.9 Other Molecular Matrices

We have presented here a selection of molecular matrices that are used as a source of topological indices and other graph descriptors. Other types of molecular matrices are investigated with the goal of exploring novel procedures for translating graph topology into a matrix [64]. The search for new structural descriptors based on molecular graphs is the catalyst that prompted the development of many molecular matrices, such as the edge Wiener matrix $\mathbf{W_e}$ [155], the path Wiener matrix $\mathbf{W_p}$ [155], the distance-valency matrix **Dval** [34], the quasi-Euclidean matrix $\rho_{\mathbf{q\varepsilon}}$ [156,157], the distance complement matrix **DC** [66], the complementary distance matrix **CD** [145,158], the reverse Wiener matrix **RW** [67], the distance-path matrix $\mathbf{D_p}$ [68], the Szeged matrix **Sz** [70], the Cluj matrix **Cj** [70], and the resistance distance matrix Ω [50], which is based on a novel distance function on graphs introduced by Klein and Randić and inspired by the properties of electrical networks.

1.5 CONCLUDING REMARKS

This chapter reviewed the applications of graph theory in chemistry. Many objects manipulated in chemistry, such as atomic orbitals, chemical compounds, and reaction diagrams, can be represented as graphs. Graph operations, such as generating reduced graphs, and the calculation of various matrices derived from the connectivity of the graph can thus be applied to chemicals with applications including virtual screening, topological indices calculations, and activity/property predictions such as spectra predictions. Many algorithms have been and are being developed to solve graph problems, and some of these can be applied to chemistry problems. The goal of the next chapter is to present graph algorithms applied to chemicals.

REFERENCES

1. Harary, F., *Graph Theory*. Adison-Wesley: Reading, MA, 1994.
2. Berge, C., *Graphs and Hypergraphs*. Elsevier: New York, 1973.

3. Behzad, M., Chartrand, G., and Lesniak-Foster, L., *Graphs and Digraphs*. Wadsworth International Group: Belmont, CA, 1979.
4. Buckley, F. and Harary, F., *Distance in Graphs*. Adison-Wesley: Reading, MA, 1990.
5. Foulds, L. R., *Graph Theory Applications*. Springer: New York, 1992.
6. West, D. B., *Introduction to Graph Theory*, 2nd edition. Prentice-Hall: Englewood Cliffs, NJ, 2000.
7. Diestel, R., *Graph Theory*, 3rd edition. Springer: Heidelberg, Germany, 2005.
8. Beineke, L. W. and Wilson, R. J., *Topics in Algebraic Graph Theory*. Cambridge University Press: Cambridge, UK, 2005.
9. Cvetković, D. M., Doob, M., and Sachs, H., *Spectra of Graphs. Theory and Applications*, 3rd edition. Johann Ambrosius Barth Verlag: Heidelberg, Germany, 1995.
10. Graovac, A., Gutman, I., and Trinajstić, N., *Topological Approach to the Chemistry of Conjugated Molecules*. Springer: Berlin, 1977.
11. Cyvin, S. J. and Gutman, I., *Kekulé Structures in Benzenoid Hydrocarbons*, Vol. 46. Springer: Berlin, 1988.
12. Gutman, I. and Cyvin, S. J., *Introduction to the Theory of Benzenoid Hydrocarbons*. Springer: Berlin, 1989.
13. Gutman, I. and Cyvin, S. J., *Advances in the Theory of Benzenoid Hydrocarbons*, Vol. 153. Springer: Berlin, 1990.
14. Cyvin, S. J., Brunvoll, J., and Cyvin, B. N., *Theory of Coronoid Hydrocarbons*, Vol. 54. Springer: Berlin, 1991.
15. Dias, J. R., *Molecular Orbital Calculations Using Chemical Graph Theory*. Springer: Berlin, 1993.
16. Harary, F. and Palmer, E. M., *Graphical Enumeration*. Academic Press: New York, 1973.
17. Pólya, G. and Read, R. C., *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*. Springer: New York, 1987.
18. Fujita, S., *Symmetry and Combinatorial Enumeration in Chemistry*. Springer: Berlin, 1991.
19. Biggs, N. L., Lloyd, E. K., and Wilson, R. J., *Graph Theory 1736–1936*. Clarendon Press: Oxford, 1976.
20. Balaban, A. T., *Chemical Applications of Graph Theory*. Academic Press: London, 1976.
21. Trinajstić, N., *Chemical Graph Theory*. CRC Press: Boca Raton, FL, 1992.
22. Gutman, I. and Polansky, O. E., *Mathematical Concepts in Organic Chemistry*. Springer: Berlin, 1986.
23. Gasteiger, J., *Handbook of Chemoinformatics*. Wiley-VCH: Weinheim, 2003.
24. Kier, L. B. and Hall, L. H., *Molecular Connectivity in Chemistry and Drug Research*. Academic Press: New York, 1976.
25. Kier, L. B. and Hall, L. H., *Molecular Connectivity in Structure–Activity Analysis*. Research Studies Press: Letchworth, UK, 1986.
26. Kier, L. B. and Hall, L. H., *Molecular Structure Description. The Electrotopological State*. Academic Press: San Diego, CA, 1999.
27. Bonchev, D., *Information Theoretic Indices for Characterization of Chemical Structure*. Research Studies Press: Chichester, UK, 1983.
28. Devillers, J. and Balaban, A. T., *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers: Amsterdam, the Netherlands, 1999.
29. Temkin, O. N., Zeigarnik, A. V., and Bonchev, D., *Chemical Reaction Networks. A Graph-Theoretical Approach*. CRC Press: Boca Raton, FL, 1996.
30. Koča, J., Kratochvíl, M., Kvasnička, V., Matyska, L., and Pospíchal, J., *Synthon Model of Organic Chemistry and Synthesis Design*, Vol. 51. Springer: Berlin, Germany, 1989.

31. Golender, V. E. and Rozenblit, A. B., *Logical and Combinatorial Algorithms for Drug Design*, p. 289. Research Studies Press: Letchworth, UK, 1983.
32. Balaban, A. T., Reaction graphs. In: D. Bonchev and O. Mekenyan (Eds), *Graph Theoretical Approaches to Chemical Reactivity*, pp. 137–180. Kluwer Academic Publishers: Amsterdam, the Netherlands, 1994.
33. Ivanciuc, O., Design of topological indices. Part 11. Distance-valency matrices and derived molecular graph descriptors. *Revue Roumaine de Chimie* 1999, 44(5), 519–528.
34. Ivanciuc, O., Design of topological indices. Part 14. Distance-valency matrices and structural descriptors for vertex- and edge-weighted molecular graphs. *Revue Roumaine de Chimie* 2000, 45(6), 587–596.
35. Dijkstra, E., A note on two problems in connection with graphs. *Numerische Mathematik* 1959, 1, 269–271.
36. Ivanciuc, O., Ivanciuc, T., and Balaban, A. T., Vertex- and edge-weighted molecular graphs and derived structural descriptors. In: J. Devillers and A. T. Balaban (Eds), *Topological Indices and Related Descriptors in QSAR and QSPR*, pp. 169–220. Gordon and Breach Science Publishers: Amsterdam, the Netherlands, 1999.
37. Ivanciuc, O. and Ivanciuc, T., Matrices and structural descriptors computed from molecular graph distances. In: J. Devillers and A.T. Balaban (Eds), *Topological Indices and Related Descriptors in QSAR and QSPR*, pp. 221–277. Gordon and Breach Science Publishers: Amsterdam, the Netherlands, 1999.
38. Higham, D. J., Kalna, G., and Kibble, M., Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics* 2007, 204(1), 25–37.
39. Randić, M., Random walks and their diagnostic value for characterization of atomic environment. *Journal of Computational Chemistry* 1980, 1(4), 386–399.
40. Burdett, J. K., Lee, S., and McLarnan, T. J., The coloring problem. *Journal of the American Chemical Society* 1985, 107(11), 3083–3089.
41. Burdett, J. K., Topological control of the structures of molecules and solids. *Accounts of Chemical Research* 1988, 21(5), 189–194.
42. Jiang, Y. and Zhang, H., Aromaticities and reactivities based on energy partitioning. *Pure and Applied Chemistry* 1990, 62(3), 451–456.
43. Wu, Y., Zhao, H. G., Liu, X., Li, J., Yang, K., and He, H. B., Evaluation of molecular moments by three methods. *International Journal of Quantum Chemistry* 2000, 78(4), 237–244.
44. Marković, S., Marković, Z., and McCrindle, R. I., Spectral moments of phenylenes. *Journal of Chemical Information and Computer Sciences* 2001, 41(1), 112–119.
45. Schmalz, T. G., Živković, T., and Klein, D. J., Cluster expansion of the Hückel molecular energy of acyclics: Applications to pi resonance theory. In: R. C. Lacher (Ed.), *MATH/CHEM/COMP 1987. Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Science, Dubrovnik, Yugoslavia, 22–26 June 1987*, Vol. 54, pp. 173–190. Elsevier: Amsterdam, the Netherlands, 1988.
46. Mohar, B., Laplacian matrices of graphs. In: A. Graovac (Ed.), *MATH/CHEM/COMP 1988. Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences, Dubrovnik, Yugoslavia, 20–25 June 1988*, Vol. 63, pp. 1–8. Elsevier: Amsterdam, the Netherlands, 1989.
47. Ivanciuc, O., Chemical graph polynomials. Part 3. The Laplacian polynomial of molecular graphs. *Revue Roumaine de Chimie* 1993, 38(12), 1499–1508.
48. Trinajstić, N., Babić, D., Nikolić, S., Plavšić, D., Amić, D., and Mihalić, Z., The Laplacian matrix in chemistry. *Journal of Chemical Information and Computer Sciences* 1994, 34(2), 368–376.

49. Ivanciuc, O., Design of topological indices. Part 26. Structural descriptors computed from the Laplacian matrix of weighted molecular graphs: Modeling the aqueous solubility of aliphatic alcohols. *Revue Roumaine de Chimie* 2001, 46(12), 1331–1347.
50. Klein, D. J. and Randić, M., Resistance distance. *Journal of Mathematical Chemistry* 1993, 12(1–4), 81–95.
51. Ivanciuc, T., Ivanciuc, O., and Klein, D. J., Posetic quantitative superstructure/activity relationships (QSSARs) for chlorobenzenes. *Journal of Chemical Information and Modeling* 2005, 45(4), 870–879.
52. Ivanciuc, T., Ivanciuc, O., and Klein, D. J., Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative superstructure/activity relationships (QSSAR). *Molecular Diversity* 2006, 10(2), 133–145.
53. Ivanciuc, T., Ivanciuc, O., and Klein, D. J., Prediction of environmental properties for chlorophenols with posetic quantitative super-structure/property relationships (QSSPR). *International Journal of Molecular Sciences* 2006, 7(9), 358–374.
54. Klein, D. J., Ivanciuc, T., Ryzhov, A., and Ivanciuc, O., Combinatorics of reaction-network posets. *Combinatorial Chemistry & High Throughput Screening* 2008, 11(9), 723–733.
55. Mihalić, Z., Veljan, D., Amić, D., Nikolić, S., Plavšić, D., and Trinajstić, N., The distance matrix in chemistry. *Journal of Mathematical Chemistry* 1992, 11(1–3), 223–258.
56. Floyd, R. W., Algorithm 97: Shortest path. *Communications of the ACM* 1962, 5(6), 345.
57. Warshall, S., A theorem on boolean matrices. *Journal of the ACM* 1962, 9, 11–12.
58. Wiener, H., Structural determination of paraffin boiling points. *Journal of the American Chemical Society* 1947, 69, 17–20.
59. Balaban, A. T., Highly discriminating distance-based topological index. *Chemical Physics Letters* 1982, 89(5), 399–404.
60. Balaban, A. T. and Ivanciuc, O., FORTRAN 77 computer program for calculating the topological index J for molecules containing heteroatoms. In: A. Graovac (Ed.), *MATH/CHEM/COMP 1988. Proceedings of an International Course and Conference on the Interfaces Between Mathematics, Chemistry and Computer Sciences, Dubrovnik, Yugoslavia, 20–25 June 1988*, Vol. 63, pp. 193–211. Elsevier: Amsterdam, the Netherlands, 1989.
61. Hall, L. H., Mohney, B., and Kier, L. B., The electrotopological state: Structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Computer Sciences* 1991, 31(1), 76–82.
62. Balaban, A. T. and Balaban, T.-S., New vertex invariants and topological indices of chemical graphs based on information on distances. *Journal of Mathematical Chemistry* 1991, 8(4), 383–397.
63. Balaban, A. T., Beteringhe, A., Constantinescu, T., Filip, P. A., and Ivanciuc, O., Four new topological indices based on the molecular path code. *Journal of Chemical Information and Modeling* 2007, 47(3), 716–731.
64. Ivanciuc, O., Graph theory in chemistry. In: J. Gasteiger (Ed.), *Handbook of Chemoinformatics*, Vol. 1, pp. 103–138. Wiley-VCH: Weinheim, Germany, 2003.
65. Ivanciuc, O., Balaban, T.-S., and Balaban, A. T., Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *Journal of Mathematical Chemistry* 1993, 12(1–4), 309–318.
66. Randić, M., Linear combinations of path numbers as molecular descriptors. *New Journal of Chemistry* 1997, 21(9), 945–951.
67. Balaban, A. T., Mills, D., Ivanciuc, O., and Basak, S. C., Reverse Wiener indices. *Croatica Chemica Acta* 2000, 73(4), 923–941.

68. Diudea, M. V., Wiener and hyper-Wiener numbers in a single matrix. *Journal of Chemical Information and Computer Sciences* 1996, 36(4), 833–836.
69. Diudea, M. V., Ivanciuc, O., Nikolić, S., and Trinajstić, N., Matrices of reciprocal distance, polynomials and derived numbers. *MATCH Communications in Mathematical and in Computer Chemistry* 1997, 35, 41–64.
70. Diudea, M. V., Indices of reciprocal properties or Harary indices. *Journal of Chemical Information and Computer Sciences* 1997, 37(2), 292–299.
71. Ivanciuc, O., Design of topological indices. Part 27. Szeged matrix for vertex- and edge-weighted molecular graphs as a source of structural descriptors for QSAR models. *Revue Roumaine de Chimie* 2002, 47(5), 479–492.
72. Cayley, A., On the mathematical theory of isomers. *Philosophical Magazine* 1874, 67, 444–446.
73. Gutman, I., Vidović, D., and Popović, L., Graph representation of organic molecules. Cayley's plerograms vs. his kenograms. *Journal of the Chemical Society, Faraday Transactions* 1998, 94(7), 857–860.
74. Gutman, I. and Vidović, D., Relations between Wiener-type topological indices of plerograms and kenograms. *Journal of the Serbian Chemical Society* 1998, 63(9), 695–702.
75. Toropov, A. A. and Toropova, A. P., QSPR modeling of the formation constants for complexes using atomic orbital graphs. *Russian Journal of Coordination Chemistry* 2000, 26(6), 398–405.
76. Toropov, A. A. and Toropova, A. P., Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants. *Journal of Molecular Structure (Theochem)* 2001, 538, 287–293.
77. Toropov, A. A. and Toropova, A. P., QSAR modeling of mutagenicity based on graphs of atomic orbitals. *Internet Electronic Journal of Molecular Design* 2002, 1(3), 108–114.
78. Pogliani, L., From molecular connectivity indices to semiempirical connectivity terms: Recent trends in graph theoretical descriptors. *Chemical Reviews* 2000, 100(10), 3827–3858.
79. Pogliani, L., Algorithmically compressed data and the topological conjecture for the inner-core electrons. *Journal of Chemical Information and Computer Sciences* 2002, 42(5), 1028–1042.
80. Pogliani, L., Complete graph conjecture for inner-core electrons: Homogeneous index case. *Journal of Computational Chemistry* 2003, 24(9), 1097–1109.
81. Pogliani, L., Encoding the core electrons with graph concepts. *Journal of Chemical Information and Computer Sciences* 2004, 44(1), 42–49.
82. Pogliani, L., The evolution of the valence delta in molecular connectivity theory. *Internet Electronic Journal of Molecular Design* 2006, 5(7), 364–375.
83. Barnard, J. M., A comparison of different approaches to Markush structure handling. *Journal of Chemical Information and Computer Sciences* 1991, 31(1), 64–68.
84. Fisanick, W., The chemical abstracts service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *Journal of Chemical Information and Computer Sciences* 1990, 30(2), 145–154.
85. Ebe, T., Sanderson, K. A., and Wilson, P. S., The chemical abstracts service generic chemical (Markush) structure storage and retrieval capability. 2. The MARPAT file. *Journal of Chemical Information and Computer Sciences* 1991, 31(1), 31–36.
86. Benichou, P., Klimczak, C., and Borne, P., Handling genericity in chemical structures using the Markush Darc software. *Journal of Chemical Information and Computer Sciences* 1997, 37(1), 43–53.

87. Lynch, M. F., Barnard, J. M., and Welford, S. M., Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy. *Journal of Chemical Information and Computer Sciences* 1981, 21(3), 148–150.
88. Holliday, J. D., Downs, G. M., Gillet, V. J., Lynch, M. F., and Dethlefsen, W., Evaluation of the screening stages of the Sheffield research project on computer storage and retrieval of generic chemical structures in patents. *Journal of Chemical Information and Computer Sciences* 1994, 34(1), 39–46.
89. Barnard, J. M., Lynch, M. F., and Welford, S. M., Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL, a formal language for the description of generic chemical structures. *Journal of Chemical Information and Computer Sciences* 1981, 21(3), 151–161.
90. Welford, S. M., Lynch, M. F., and Barnard, J. M., Computer storage and retrieval of generic chemical structures in patents. 3. Chemical grammars and their role in the manipulation of chemical structures. *Journal of Chemical Information and Computer Sciences* 1981, 21(3), 161–168.
91. Barnard, J. M., Lynch, M. F., and Welford, S. M., Computer storage and retrieval of generic structures in chemical patents. 4. An extended connection table representation for generic structures. *Journal of Chemical Information and Computer Sciences* 1982, 22(3), 160–164.
92. Welford, S. M., Lynch, M. F., and Barnard, J. M., Computer storage and retrieval of generic chemical structures in patents. 5. Algorithmic generation of fragment descriptors for generic structure screening. *Journal of Chemical Information and Computer Sciences* 1984, 24(2), 57–66.
93. Holliday, J. D., Downs, G. M., Gillet, V. J., and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 14. Fragment generation from generic structures. *Journal of Chemical Information and Computer Sciences* 1992, 32(5), 453–462.
94. Holliday, J. D., Downs, G. M., Gillet, V. J., and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 15. Generation of topological fragment descriptors from nontopological representations of generic structure components. *Journal of Chemical Information and Computer Sciences* 1993, 33(3), 369–377.
95. Barnard, J. M., Lynch, M. F., and Welford, S. M., Computer storage and retrieval of generic chemical structures in patents. 6. An interpreter program for the generic structure description language GENSAL. *Journal of Chemical Information and Computer Sciences* 1984, 24(2), 66–71.
96. Dethlefsen, W., Lynch, M. F., Gillet, V. J., Downs, G. M., and Holliday, J. D., Computer storage and retrieval of generic chemical structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system. *Journal of Chemical Information and Computer Sciences* 1991, 31(2), 233–253.
97. Gillet, V. J., Welford, S. M., Lynch, M. F., Willett, P., Barnard, J. M., Downs, G. M., Manson, G., and Thompson, J., Computer storage and retrieval of generic chemical structures in patents. 7. Parallel simulation of a relaxation algorithm for chemical substructure search. *Journal of Chemical Information and Computer Sciences* 1986, 26(3), 118–126.
98. Gillet, V. J., Downs, G. M., Ling, A., Lynch, M. F., Venkataram, P., Wood, J. V., and Dethlefsen, W., Computer storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical structure retrieval. *Journal of Chemical Information and Computer Sciences* 1987, 27(3), 126–137.
99. Gillet, V. J., Downs, G. M., Holliday, J. D., Lynch, M. F., and Dethlefsen, W., Computer storage and retrieval of generic chemical structures in patents. 13. Reduced

graph generation. *Journal of Chemical Information and Computer Sciences* 1991, 31(2), 260–270.

100. Downs, G. M., Gillet, V. J., Holliday, J. D., and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 9. An algorithm to find the extended set of smallest rings in structurally explicit generics. *Journal of Chemical Information and Computer Sciences* 1989, 29(3), 207–214.
101. Downs, G. M., Gillet, V. J., Holliday, J. D., and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 10. Assignment and logical bubble-up of ring screens for structurally explicit generics. *Journal of Chemical Information and Computer Sciences* 1989, 29(3), 215–224.
102. Dethlefsen, W., Lynch, M. F., Gillet, V. J., Downs, G. M., Holliday, J. D., and Barnard, J. M., Computer storage and retrieval of generic chemical structures in patents. 12. Principles of search operations involving parameter lists: Matching-relations, user-defined match levels, and transition from the reduced graph search to the refined search. *Journal of Chemical Information and Computer Sciences* 1991, 31(2), 253–260.
103. Holliday, J. D. and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 16. The refined search: An algorithm for matching components of generic chemical structures at the atom-bond level. *Journal of Chemical Information and Computer Sciences* 1995, 35(1), 1–7.
104. Holliday, J. D. and Lynch, M. F., Computer storage and retrieval of generic chemical structures in patents. 17. Evaluation of the refined search. *Journal of Chemical Information and Computer Sciences* 1995, 35(4), 659–662.
105. Gillet, V. J., Willett, P., and Bradshaw, J., Similarity searching using reduced graphs. *Journal of Chemical Information and Computer Sciences* 2003, 43(2), 338–345.
106. Barker, E. J., Gardiner, E. J., Gillet, V. J., Kitts, P., and Morris, J., Further development of reduced graphs for identifying bioactive compounds. *Journal of Chemical Information and Computer Sciences* 2003, 43(2), 346–356.
107. Barker, E. J., Buttar, D., Cosgrove, D. A., Gardiner, E. J., Kitts, P., Willett, P., and Gillet, V. J., Scaffold hopping using clique detection applied to reduced graphs. *Journal of Chemical Information and Modeling* 2006, 46(2), 503–511.
108. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D., Training similarity measures for specific activities: Application to reduced graphs. *Journal of Chemical Information and Modeling* 2006, 46(2), 577–586.
109. Dubois, J. E., Laurent, D., and Viellard, H., Système DARC. Principes de recherches des corrélations et équation générale de topoinformation. *Comptes Rendus de l'Académie des Sciences Paris* 1967, 264C, 1019–1022.
110. Dubois, J. E. and Viellard, H., Système DARC. Théorie de génération: Description I. *Bulletin de la Société Chimique de France* 1968, 900–904.
111. Dubois, J. E. and Viellard, H., Système DARC. Théorie de génération: Description II. *Bulletin de la Société Chimique de France* 1968, 905–912.
112. Dubois, J. E. and Viellard, H., Système DARC. Théorie de génération: Description III. *Bulletin de la Société Chimique de France* 1968, 913–919.
113. Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A., and Doucet, J. P., 13C NMR chemical shift prediction of the sp^3 carbon atoms in the a position relative to the double bond in acyclic alkenes. *Journal of Chemical Information and Computer Sciences* 1997, 37(3), 587–598.
114. Dubois, J. E., Doucet, J. P., Panaye, A., and Fan, B. T., DARC site topological correlations: Ordered structural descriptors and property evaluation. In: J. Devillers and A. T. Balaban