Information Theoretic Indices for Characterization of Chemical Structures

Danail Bonchev

RSP RESEARCH STUDIES PRESS

Information Theoretic Indices for Characterization of Chemical Structures

Professor Danail Bonchev
Higher School of Chemical Technology, Burgas, Bulgaria

RESEARCH STUDIES PRESS
A DIVISION OF JOHN WILEY & SONS LTD
Chichester · New York · Brisbane · Toronto · Singapore

RESEARCH STUDIES PRESS

Editorial Office:
58B Station Road, Letchworth, Herts. SG6 3BE, England

Library of Congress Cataloging in Publication Data:

Bonchev, Danail.
Information theoretic indices for characterisation of chemical structures.
(Chemometrics research studies series; no. 5)
Includes index.
1. Chemical structure. 2. Information theory in chemistry. I. Title. II. Series.
QD471.B76 1983 541.2'2 82-16696
ISBN 0 471 90087 7

British Library Cataloguing in Publication Data:

Bonchev, Danail.
Information theoretic indices for characterisation of chemical structures—(Chemometrics research studies series; 5)
1. Chemistry—Information services
2. Chemical literature
I. Title
540'.7 Q08.3

ISBN 0 471 90087 7

Printed in Great Britain

Information Theoretic Indices for Characterization of Chemical Structures

CHEMOMETRICS SERIES

Series Editor: **Dr. D. Bawden**
Pfizer Central Research, Sandwich, Kent, England

1. Distance Geometry and Conformational Calculations
   G. M. Crippen

2. Clustering of Large Data Sets
   Jure Zupan

3. Multivariate Data Analysis in Industrial Practice
   Paul J. Lewi

4. Correlation Analysis of Organic Reactivity: With particular reference to multiple regression
   John Shorter

5. Information Theoretic Indices for Characterization of Chemical Structures
   Danail Bonchev

Editorial Preface

The concepts of information theory have been steadily infiltrating the physical and biological sciences for several decades now. Although originally conceived as a means of modelling communications channels, information theory has held out the prospect of the introduction of quantitative measures of "order", "complexity", and "information content" into scientific thought. Chemistry, with its hierarchy of nuclei, atoms, functionabilities, molecules, and polymers, and its highly varied, though ordered, patterns of bonding and structure, is a particularly attractive area for the application of information theory.

Professor Bonchev here presents such an application, the derivation of numerical indices based on information theoretic formulae to represent the topological structure of atoms and molecules. Indices based on a graph theoretical analysis of molecular structure have a relatively long history, whereas the information theoretic indices are of more recent introduction. The number of such indices is quite large, to the uninitiated confusingly so. Professor Bonchev, in thoroughly reviewing and comparing these indices, sets them clearly into content and, by including the results of his more recent work, shows the directions in which the field is moving.

The representation of atomic and molecular structure by numeric indices has great attraction, particularly when these indices may be related to concepts such as "complexity". The practical applications, particularly in the quantitative relating of structure to atomic or molecular properties, are beginning to become established, and are likely to increase in importance in the future.

Finally, it may be, as Profesor Bonchev notes, that this area of study may be a forerunner of an increasing tendency towards the unification of presently diverse scientific disciplines, through the common language of cybernetics and information theory. This is indeed an intriguing prospect.

David Bawden

Preface

Contemporary science is characterized by a high degree of differentiation. The enormous number of new scientific disciplines, which has arisen in the 20th century, hinders the contacts between the different branches of science threatening it with the ghost of Babel. Fortunately, integrating tendencies have also been in progress during the last few decades including here many general sciences such as cybernetics, theory of systems, artificial intelligence, information theory, etc.

Information theory can be regarded as a universal language for describing systems, allowing useful analogies or common laws between systems of different nature to be inferred. In this way the laws in a certain scientific area, properly translated into information language, could be projected into other, not well developed scientific areas. The information approach however is not simply a translation from less general languages to a universal one. It provides additional insight into systems and phenomena, allowing new results to be obtained.

The applicability of information theory to chemistry is mainly based on the possibilites offered by it for the quantitative analysis of various aspects of chemical structures. Related to this, the present book is concerned with the so-called theoretic information indices which are structural indices expressing in a quantitative form the degree of complexity of molecules and atoms. These indices are a powerful tool for discrimination between similar structures which make them fairly suitable for various classificational aims, as well as for computer processing and retrieval of chemical information. They may also be helpful when a choice between competing approximations of hypotheses has to be made. The most promising field of application, however, is the correlation between structure and properties (including biological activity) of chemical systems, due to the potential possibility of translating any structural features into a numerical descriptor, a theoretic information index. Finally, new results and generalizations may be expected on the basis of the extremal values of the information function.

This monograph is a survey of the information theoretic indices introduced during the last three decades to characterize molecules and atoms. To a large extent it reports original results obtained in the theoretical chemistry group in Burgas, Bulgaria, where systematic studies on the chemical applications of Information Theory have been carried-out ever since 1968. An outline of the development of the notion of information is given in the Introduction. Other chemical applications, which are not explored further, are also reviewed there. The Second Chapter deals briefly with the basic equations of Information Theory and the ways they are applied to structural problems in Chemistry. The various atomic and molecular information-theoretic indices are discussed in detail in Chapters III to V, while their specific applications in Chapter VI. The book has many examples and illustrations. The mathematical formalism is reduced to a reasonable degree. This makes the information theoretic approach accessible to every chemist, in accord with the purpose of the book to stimulate the interest of the chemical community towards this new branch

of mathematical chemistry.

The author would like to acknowledge his great indebtedness to Prof. Balaban (Bucharest), Prof. Trinajstić (Zagreb), Prof. Randić (Ames), Prof. Zhdanov (Rostov na Don), Prof. Dosmorov (Omsk), Dr. Roy (Calcutta) and Dr. Basak (Duluth, Minnesota), for offering him the opportunity to read their papers prior to publication, to Prof. Polansky (Muelheim) and Prof. Rouvray (London) for some useful criticism, to Dr. Nikolov (Sofia) for correcting the English of the manuscript, and to his co-workers Dr. Ovanes Mekenayn, Dr. Verginia Kamenska, and Dr. Brezitsa Rousseva for contributing to the development of some information-theoretic indices.

The author is particularly indebted to Dr. David Bawden of Pfizer Central Research, the Editor of this Series, who stimulated the writing of this monograph and encouraged its preparation.

Danail Bonchev

Table of Contents

CHAPTER 1
The Notion of Information and its Applications

The notion of information appears as one of the most fundamental notions in the 20-th century science, a notion of no less importance than that of matter and energy. This assertion follows from the very definition of information. According to Norbert Wiener (1948) "information is neither matter, nor energy". Ashby (1956) treats information as "a measure of the variety in a given system". Following Glushkov (1964) *"information is a measure of the nonhomogeneity in the distribution of matter or energy in space and time."* On this basis it becomes more and more evident that besides their substance and energy essence, the objects and phenomena in nature and technology also have an information character. Moreover, the centre of scientific research is expected by some prognoses to move towards the information nature of the processes as the major field of study in 21st century. These prognoses are essentially based on the possibility of systems and processes to be controlled by information, which is the major function of information in cybernetics. As a perspective these ideas could lead to a technology in which every atom or molecule is controlled by information, a possibility that is realized in living nature.

The year 1948 is usually considered as the birth date of Information Theory, since this was the year in which Claude

Shannon published his fundamental work (Shannon and Weaver, 1948). The concept of information as a quantity related to entropy is, however, much older. Boltzmann (1894) stated that every piece of information obtained for a physical system is related to the decrease in the number of its possible states, therefore the increase of entropy means "loss of information". Szillard (1929) developed this idea for the general case of information in physics. Later, Brillouin (1956, 1964) generalised the concept of entropy and information in his negentropy principle of information. He specified information as a negative entropy (negentropy) and extended the definition of the Second Law of Thermodynamics so as to encompass the information as well. The possible interplay between Information Theory and Thermodynamics and in particular between entropy and information, is a subject of permanent interest. (A list of selected references to this field is given elsewhere (Bonchev and Lickomannov, 1977). New horizons have been revealed for statistical thermodynamics on the basis of the Jaynes (1957 a,b) maximum entropy hypothesis (or Jaynes'principle). From the latest development of the problem the works of Kobozev (1971) on thermodynamics of thinking are of considerable interest where the concept of the anti-entropy character of thinking is proposed.

Arising as "a special theory of communications", Information Theory soon exceeded its initial limits and found application in numerous scientific and technical areas: physics, chemistry, biology, medicine, linguistics, psychology, aesthetics et al. (A detailed survey is given by Mathai and Rathie, 1975). The role of information has been recognized first in biology. Some important problems of conservation, processing and transmission of information in living beings have been solved, such as coding of genetic information (Quastler, Editor, 1953; Hasegawa and Yano, 1975; Seybold, 1976; Gatlin, 1972), estimation of the possibility of spontaneous self-generation of life on Earth (Rashevsky, 1960: Eigen and Winkler, 1975; Eigen and Schuster, 1979), Formulation of the fundamental laws of biological thermodynamics (Trincher, 1964), analysis of the problems of bioenergetics (Bykhovskii, 1968), etc. The information content of systems has been used as a quantitative criterion of evolution (Ursul, 1966; Bonchev, 1970). The information character of food-consuming processes has been pointed out to dominate over their substance and energy nature (Schroedinger, 1944; Brekhman, 1976). Works on the information theory of disease diagnostics and therapy are in progress.

The interest of the chemical community in Information Theory has been permanently growing. Levine, Bernstein et al. developed an information-theoretic approach to molecular dynamics which describes the behaviour of interacting molecular systems far from equilibrium, in vibrationally and rotationally excited states (Bernstein and Levine, 1972; Ben-Shaul et al., 1972; Levine and Bernstein, 1974, 1976). A quantitative measure of the information content, or entropy deficiency, of different classes of experiments has been introduced, e.g. molecular beam scattering, IR chemoluminescence etc. This approach found various applications including the determination of branching ratios for alternative reaction paths, the study on the operational characteristics of lasers, etc. Equivalence has been demonstrated to exist between the maximum entropy approach to molecular collisions and their description by means of the dynamical equations of motion (Alhassid and Levine, 1977).