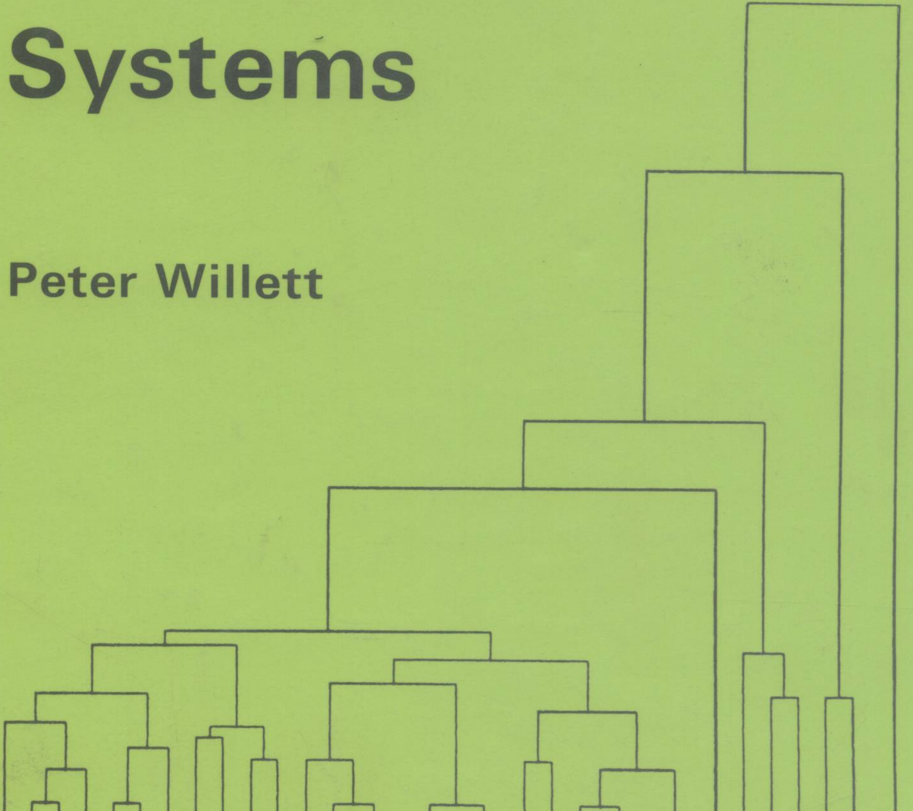
Similarity and Clustering in Chemical Information Systems

Peter Willett



RSP RESEARCH STUDIES PRESS

Similarity and Clustering in Chemical Information Systems

Peter Willett

*Department of Information Studies
University of Sheffield, England*

RESEARCH STUDIES PRESS LTD
Letchworth, Hertfordshire, England

JOHN WILEY & SONS INC.
New York · Chichester · Toronto · Brisbane · Singapore

RESEARCH STUDIES PRESS LTD
58B Station Road, Letchworth, Herts. SG6 3BE, England

Marketing and Distribution
Australia, New Zealand, South-east Asia:
Jacaranda-Wiley Ltd, Jacaranda Press
JOHN WILEY & SONS INC.
GPO Box 859, Brisbane, Queensland 4001, Australia

Canada:
JOHN WILEY & SONS CANADA LIMITED
22 Worcester Road, Rexdale, Ontario, Canada

Europe, Africa:
JOHN WILEY & SONS LIMITED
Baffins Lane, Chichester, West Sussex, England

North and South America and the rest of the world:
JOHN WILEY & SONS INC.
605 Third Avenue, New York, NY 10158, USA

Library of Congress Cataloging-in-Publication Data

Willett, Peter, 1953–
Similarity and clustering in chemical information systems.

(Chemometrics series; 12)
Bibliography: p.
Includes indexes.
1. Chemical structure—Data processing. 2. Information storage and retrieval systems—Chemical structure.
3. Cluster analysis. I. Title.
III. Series
QD471.W578 1987 025′.065412′2 86-31433
ISBN 0 86380 050 5
ISBN 0 471 91463 0 (Wiley)

British Library Cataloguing in Publication Data

Willett, Peter, 1953–
Similarity and clustering in chemical information systems.—(Chemometrics series; 12)
1. Chemistry, Analytic—Statistical methods 2. Cluster analysis
I. Title II. Series
543′.001′51953 QD75.4.S8

ISBN 0 86380 050 5
ISBN 0 471 91463 0 Wiley

ISBN 0 86380 050 5 (Research Studies Press Ltd)
ISBN 0 471 91463 0 (John Wiley & Sons Inc)

Printed in Great Britain by Galliard (Printers) Ltd, Great Yarmouth

Similarity and Clustering in Chemical Information Systems

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

6. Logical and Combinatorial Algorithms for Drug Design
   V. E. Golender *and* **A. B. Rozenblit**

7. Minimum Steric Difference
   The MTD method for QSAR studies
   Z. Simon, A. Chiriac, S. Holban, D. Ciubotaru *and* **G. I. Mihalas**

8. Analytical Measurement and Information
   Advances in the information theoretic approach to chemical analyses
   K. Eckschlager *and* **V. Štěpánek**

9. Molecular Connectivity in Structure-Activity Analysis
   Lemont B. Kier *and* **Lowell H. Hall**

10. Potential Pattern Recognition in Chemical and Medical Decision Making
    D. Coomans *and* **I. Broeckaert**

11. Chemical Pattern Recognition
    O. Štrouf

12. Similarity and Clustering in Chemical Information Systems
    Peter Willett

For Marie and Barbara

Preface

The advent of computers has brought about a revolution in the ways in which information of all kinds is stored and retrieved. Some of the most striking developments have appeared in computerised chemical information systems, such as are run by organisations concerned with the storage, retrieval and processing of information pertaining to chemical structures. This book discusses computational techniques which may further increase the effectiveness of such information systems; specifically, searching and clustering methods are described which are based upon the calculation of measures of similarity between chemical structures in machine-readable files.

The first chapter of the book presents a broad over-view of the current state of development of computerised chemical information systems. After a discussion of the means by which molecules may be represented in machine-readable form, the chapter concentrates upon two major application areas, these being chemical structure and substructure searching and the study of structure-activity relationships, since it is in these areas that similarity and clustering techniques may most usefully be applied. Chapter 2 reviews these two classes of technique. The discussion includes the measurement of similarity between objects, and the wide range of clustering methods which have been suggested for the grouping of data. Clustering methodologies have been developed and used in a very wide range of application areas, and an attempt is made to summarise what is already a widely disparate literature. Particular attention is paid to criteria which can be used to evaluate the merits of different clustering techniques.

The next three chapters form the heart of the book, and present the results of an extended investigation into the application of similarity and clustering techniques to structure-based chemical information systems. Chapter 3 compares different approaches to the calculation of inter-molecular structural similarity, and describes the use of similarity measures to provide powerful new retrieval mechanisms for chemical structure and substructure searching. Chapter 4 follows a similar pattern with the first part devoted to a systematic comparison of a wide range of different types of clustering method; this is followed by a description of the use of one such method in an operational pharmaceutical information system. Clustering methods were first developed for use in the life sciences where a data set may contain only a few tens of objects, this implying that even the most computationally demanding clustering algorithms may be implemented with little difficulty. This is certainly not the case with chemical structure files, which may contain thousands or tens of thousands of molecules, and it is accordingly necessary to devise exceedingly efficient implementations if the methods are to be applied in a chemical context. This problem is discussed in Chapter 5, where algorithms are presented in sufficient detail to allow readers of this book to implement the methods for themselves.

A summary of the major results of the previous chapters and suggestions for further work are presented in the closing chapter.

Many people have helped in the work which led to the writing of this book. I owe a deep debt of gratitude to Prof. Michael Lynch and Dr. George Vladutz for introducing me to the study of chemical information systems, and for encouraging my researches for some ten years. Much of the work described here has been carried out by masters and doctoral research students and research associates under my supervision and I would like to thank in this context Abdelmoula El-Hamdouchi, Alan Griffiths, Hazel Morton, Shirley Perry, Lesley Robinson, Ian Rowlands and, in particular, Vivienne Winterman. Funding for their endeavours has come from a variety of sources including the British Council, the British Library Research and Development Department, the Department of Education and Science, Pfizer Central Research, and the Science and Engineering Research Council. I am also pleased to acknowledge the assistance of the staff of the Computing Services Department, University of Sheffield, for providing the computational resources for the work reported here. Dr. David Bawden has been a strong advocate of the utility of my research work, and encouraged me in the writing of this book in his role as the editor of the Research Studies Series.

Much of the material in this book has appeared previously in a range of journal articles and is reprinted with permission from Anal. Chim. Acta 136, 29-37 (1982), 138, 339-342 (1982) and 151, 161-166 (1983) (Copyright: Elsevier Science Publishers); J. Chem. Inform. Comput. Sci. 23, 22-25 (1983), 24, 29-33 (1984), 25, 78-80 (1985), 26, 36-41 (1986) and 26, 109-118 (1986) (Copyright:

American Chemical Society); J. Docum. 40, 175-205 (1984) (Copyright: ASLIB, the Association for Information Management) Quant. Struct. Activ. Relat. 5, 18-25 (1986) (Copyright: VCH Verlagsgesellschaft).

Peter Willett, July 1986.

Contents

CHAPTER 1: CHEMICAL INFORMATION SYSTEMS

CHAPTER 2: SIMILARITY AND CLUSTER ANALYSIS

CHAPTER 3: SIMILARITY IN CHEMICAL INFORMATION SYSTEMS

CHAPTER 4: CLUSTERING IN CHEMICAL INFORMATION SYSTEMS

CHAPTER 5: NEAREST NEIGHBOUR SEARCHING ALGORITHMS

CHAPTER 1
Chemical Information Systems

1.1 INTRODUCTION

Over the years, many schemes have been developed to describe and categorise chemical compounds (Rouvray, 1977). By the start of this century, the two-dimensional chemical structure diagram had established itself as the prime means of communication between chemists, and it is thus hardly surprising that it has formed the basis for the computerised systems that deal with the storage and retrieval of chemical information. This mode of representation is very closely related to the actual nature of chemical compounds, as determined by the wave equations that describe them, and there are thus few of the complexities and ambiguities that characterise, for example, the complex representations that are required for the computer processing of natural language. In addition, it has proved possible to map the chemical structure diagram onto the hardware and software components of modern day computer systems to provide a user-friendly, graphically-based man-machine interface. Given these factors, it is not surprising that information systems in chemistry are probably better developed than those in any other major discipline.

This chapter provides an overview of the facilities which are available in current computerised systems for the storage and processing of chemical structure information. The material is intended to provide the necessary background for an understanding of the similarity and clustering techniques that form the heart of this book. Accordingly, particular attention is paid to the retrieval of chemical structures and to the correlation of molecular properties with structure since it is in these areas that similarity and clustering methods would seem to be most immediately applicable. More detailed treatments of the topics considered in this chapter are presented in the review volumes of Ash and Hyde (1975) and Ash et al. (1985).

1.2 THE REPRESENTATION OF CHEMICAL STRUCTURES

Four types of molecular representation have been used extensively in chemical information systems, these being systematic nomenclatures, fragmentation codes, line notations and connection tables. Of these, systematic nomenclatures, specifically those devised by Chemical Abstracts Service and by the International Union of Pure and Applied Chemistry, are primarily used in manual information retrieval systems, such as printed indexes, and to support database creation activities (Vander Stouw et al., 1976); however, the lack of flexibility in the representation means that nomenclature often needs to be translated automatically into another type of representation if it is to be of practical use in computerised chemical information systems (Vander Stouw et al., 1974; Willett, 1980). The connection table has proved to be by far the most flexible and generally useful representation, and it forms the basis for most present-day systems, as well as for the experimental work that is described in Chapters 3 to 5 of this monograph. Such systems have been designed primarily for use with two-dimensional (2-D) representations of completely specified molecules, i.e., the classical structure diagram, and the contents of the chapters reflect this bias. Rather different approaches are required for the handling of compounds in chemical patent claims and for three-dimensional (3-D) chemical structures: these topics are mentioned in Sections 1.2.1 and 1.4.4 respectively.

1.2.1 Fragmentation codes

Fragmentation codes were the first structural representation to be widely used for chemical retrieval applications and remain in use to this day, albeit in a very different form, as a component of systems based upon other types of representation.

A fragmentation code comprises a set of pre-defined substructural attributes, the presence or absence of which is used to characterise each of the molecules in a file for retrieval, or other, purposes. It can thus be regarded as a pre-coordinate indexing system, and its degree of success will depend in large part upon the discriminatory abilities of the substructures chosen for inclusion in the indexing vocabulary, and upon the extent to which they reflect the overall structural characteristics of the file of compounds that is to be indexed.

Early fragmentation codes were developed on the basis of subjective impressions of chemical importance, and the attributes which were chosen reflected the typical pre-occupations of the organisation responsible for the creation of the file of compounds. Thus a retrieval system for IR spectroscopy data would adopt a very different fragmentation code from a system designed for the retrieval of organo-phosphorus compounds. The general approach involved the identification of structural features that occurred frequently throughout a collection so that these could be described in some detail, while less common features might be assigned less discriminating notational codes: thus a file containing many potential antibiotics might have several codes for the presence of particular types of penicillin or cephalosporin rings. As we shall see, similar considerations are reflected in modern approaches to the selection of retrieval keys.

A fragment code representation has several limitations. The coding of new compounds as they are added to the file needs to be done manually, and all of the molecules need to be reprocessed if the code is changed in any way (Craig and Ebert, 1969). Furthermore, the use of a fragmentation code results in an ambiguous molecular representation. The representation is ambiguous in that the set of codes assigned to a molecule does not serve to characterise the structure fully, as the fragments might be inter-connected in very many ways, or might not represent the complete set of structural features present in the compound. Accordingly, a set of codes serves to characterise, but not necessarily delimit, an incompletely specified class of compounds, rather than characterising an individual molecule. This feature has several advantages when the generic structures that characterise chemical patents are considered, and fragment codes are hence still used extensively in patent information services. A single such generic, i.e. partially specified, structure in a patent may well correspond to many thousands, or even an infinite number, of individual specific molecules, and this presents severe problems to conventional chemical structure handling systems. However, the great importance of patent information in chemical information (Ash et al., 1985) is now leading to a proliferation of research into such problems (Barnard, 1984), and substantial progress in this field may be expected within the not-too-distant future.

Fragmentation codes were developed originally for use with punched card sorters or optical coincidence cards. However, the advent of linear notations and connection tables, which are described below and which are unambiguous, machine-readable representations, led to a rapid decrease in the use of fragmentation codes as the primary means of structural description. Instead, they are now used for the initial screening stage of substructure searching, as described in Section 1.3.2 and as molecular descriptors for the structure-property correlation studies that are described in Section 1.4.3.

1.2.2 Linear notations

A linear notation is a coding mechanism which allows a chemical structure to be represented by a string of alphanumeric characters: notations are thus related in many ways to systematic chemical nomenclatures, although the latter are of less practical use owing to the implicit representation of many of the important chemical features in a molecule. A large number of notational schemes have been described, and are still being described, in the literature, but only one of these has been used extensively within computerised chemical information systems (Rush, 1976). This is the Wiswesser Line-formula Notation (WLN) (Smith and Baker, 1975) which has formed the basis for many in-house chemical information systems over the last two decades. An example of such a system is the well-known CROSSBOW package (Ash and Hyde, 1975), a