NUMERICAL TAXONOMY

THE PRINCIPLES AND PRACTICE OF NUMERICAL CLASSIFICATION

Peter H. A. Sneath

MEDICAL RESEARCH COUNCIL MICROBIAL SYSTEMATICS UNIT
UNIVERSITY OF LEICESTER

Robert R. Sokal

STATE UNIVERSITY OF NEW YORK AT STONY BROOK

W. H. FREEMAN AND COMPANY
San Francisco

A SERIES OF BOOKS IN BIOLOGY

Editors: *Donald Kennedy*
Roderic B. Park

Library of Congress Cataloging in Publication Data

Sneath, Peter H A
Numerical taxonomy.

Bibliography: p.
1. Numerical taxonomy. I. Sokal, Robert R., joint author. II. Title.
QH83.S58 574'.01'2 72-1552
ISBN 0-7167-0697-0

Printed in the United States of America

The cover design of this book is adapted from a contour diagram of the relationships of dermanyssid mites. Reproduced with permission of Dr. W. Wayne Moss of The Academy of Natural Sciences of Philadelphia.

2 3 4 5 6 7 8 9

To JOAN and JULIE

whose endless forbearance
has sustained us through
yet another venture.

Contents

A Tribute

We are indebted to those numerous younger colleagues all over the world who, stimulated by our first book on this subject, expanded and improved our ideas and methods and built new conceptual constructs on the earlier foundations. We also acknowledge the many established systematists and specialists in other sciences who were willing to make the effort to try out our methodology. We have benefited immeasurably from the activities and enthusiasm of these persons and we hope that their efforts and the intellectual excitement that they have engendered are faithfully reflected in the pages that follow.

Preface

From the time of Linnaeus to our own, a weak point in biological science has been the absence of any quantitative meaning in our classificatory terms. What is a Class, and does Class A differ from Class B as much as Class C differs from Class D? The question can be put for the other classificatory grades, such as Order, Family, Genus, and Species. In no case can it be answered fully, and in most cases it cannot be answered at all. . . . Until some adequate reply can be given to such questions as these, our classificatory schemes can never be satisfactory or "natural." They can be little better than mnemonics—mere skeletons or frames on which we hang somewhat disconnected fragments of knowledge. Evolutionary doctrine, which has been at the back of all classificatory systems of the last century, has provided no real answer to these difficulties. Geology has given a fragmentary answer here and there. But to sketch the manner in which the various groups of living things arose is a very different thing from ascribing any quantitative value to those groups.
C. Singer, *A History of Biology* (1959), p. 200.

The rapid, almost explosive development of numerical taxonomy and the increasingly wide interest in this field made it clear to us that a new edition of our *Principles of Numerical Taxonomy*, first published in 1963, was necessary. We soon realized that a mere updating of the contents would not do justice to our task. Not only had much new material been published, which required incorporation, but changes in emphasis and in the theoretical framework of the science have taken place that needed to be presented in proper perspective. In view of this, an entirely new book has been written.

The past decade has witnessed a marked change in outlook and methodology in the fields of biological systematics and population biology. Many of the new approaches are quantitative, employing a variety of mathematical disciplines. In taxonomy there has been a considerable development of numerical methods,

many of them implemented by computers. These methods have influenced other disciplines as well, which have in turn provided numerous new concepts and techniques for systematics.

Numerical taxonomy—the grouping by numerical methods of taxonomic units based on their character states—has been aided in its present rapid development by the simultaneous development of computer techniques. Numerical taxonomy aims to develop methods that are objective, explicit, and repeatable, both in evaluation of taxonomic relationships and in the erection of taxa. Moreover, numerical methods have opened up a wide field in the exact measurement of evolutionary rates and in phylogenetic analysis. The success of the intuitive approach of the past lay in the ability of the mind to recognize swiftly, though inexactly, overall similarity in morphological detail. Such recognition is not easy with the ever increasing data bases in taxonomy, now often in tabular form, as is true of microbiological, chemical, or physiological characters; use of numerical methods with these characters becomes a necessity.

The purpose of this book is to present an up-to-date theoretical basis for numerical taxonomy, to acquaint readers with its procedures, to illustrate its advantages over conventional taxonomy, and to report on the status of the field so far.

We cannot treat all forms of numerical analysis that have been used in taxonomy, for which numerous texts on the use of statistical and mathematical methods in biology can be consulted easily. We have restricted the scope of this book instead to methods that demonstrate taxonomic relationships and create taxonomic groupings, although we treat some other techniques briefly for completeness. We have, however, attempted to treat our topics as broadly as possible and to make the book of value to zoologists, botanists, microbiologists, and paleontologists, as well as to scientists in related fields.

Readers familiar with *Principles of Numerical Taxonomy* will note that the sequence of chapters and topics has been reorganized to form, we believe, a better integrated whole. Some earlier sections have been dropped or much abbreviated—others have been liberally increased. For example, there seems by now little need for a detailed criticism and polemic against conventional practices in taxonomy. The point has been made and widely accepted, and the present need is for an expansion and elaboration of the theory and methodology of newer views. By contrast, numerical methods of phylogenetic analysis—undeveloped when the earlier book was written, required considerable space for a balanced treatment in the present text. We have tried to consolidate the discussion of the theoretical foundations of systematics and taxonomy into a cohesive whole, no longer separating our critical review of conventional systematics from the views we advocate.

Considerable changes will be found in the sections on the theory and assumptions of numerical taxonomy. The work of the last few years has led us to realize that

some of our earlier hopes and expectations of rapid, clearcut solutions to the problems of taxonomy were premature. Thus, for example, we still do not know a method for an optimal taxonomy (or if one exists) and therefore cannot advocate one. Nevertheless, it is clear to all but the most conservative workers in the field that the taxonomy of the future will be greatly aided if not entirely carried out by computers, and though it is too early in the development of numerical taxonomy to provide "cookbook" recipes for all problems, the systematist who ignores numerical taxonomic methods in his own work does so at his own loss.

Rewriting sections on methods has provided the greatest challenge. A complete account of all that has been done or proposed would result in a book twice the size of this volume, obsolescent at the time of publication, and confusing for the novice. We have had to be eclectic in our choice of methods to be presented, reporting here the most frequently used approaches and providing references to others that have been less often used or that we feel are of less general interest. Although, as we shall point out from time to time, the statistical and conceptual foundations of various aspects of numerical taxonomy are insufficient, we present here even those aspects we know lack rigor because we wish to report on the current state of the art as well as to encourage further work in these areas.

Of necessity, the level of mathematical knowledge needed for mastering the subject matter has increased somewhat since *Principles of Numerical Taxonomy* was published. Fortunately, the increasing complexity of the subject has been matched by the increasing mathematical sophistication of young biologists in systematics and other fields of biology in recent years. The time is rapidly approaching, if not at hand already, when a thorough knowledge of biometric analysis and some acquaintance with computer processing will be an ineluctable prerequisite for the aspiring taxonomist. We would expect readers of this book to have some knowledge of statistics and of elementary set and graph theory, as well as of matrix algebra. Numerous books are now on the market that provide such introductions; a selection of these is given in Appendix B.

In *Principles of Numerical Taxonomy* we attempted to review to the best of our knowledge all applications of numerical taxonomy in the field of biology. The number of papers in the field has since become so numerous that we have been hard-put to keep our fingers on the entire literature and have not tried to furnish an exhaustive bibliography in this book. By referring to a number of key review papers we have tried to point the reader to sources where comprehensive reviews of the literature in particular branches of numerical taxonomy can be found. However, in Appendix A we do furnish a list of all recent papers known to us in which numerical taxonomy has been applied to classifications of various organisms, listing these by broad taxonomic groupings. We hope the list will be of value to taxonomists who wish to survey the field in their area of specialization.

In our account of the applications of numerical taxonomy to fields other than biological systematics, we have had to be more restrictive. Applications

have been astonishing in number and diversity. Methods proposed here, and similar ones, have been applied in fields ranging from archaeology to political science, from materials classification to linguistics, and from television programming to biogeography, and though we have tried to give some coverage to the applications of numerical taxonomy to these subjects, limitations of space and our own circumscribed competence in these fields require brevity.

Readers who are not primarily interested in biological systematics but would like to use this book as an introduction to clustering and the principles of numerical classification in general, can limit their reading to the following chapters and sections without losing much by way of continuity: Sections 1.1–5, 1.9; 2.1–2, 2.8; 3.1–3, 3.8–9; 4.1–8, 4.10, 4.12–13; 5.1–2, 5.4–10; 6.4; Chapter 8; and Chapter 11.

We have not attempted to illustrate techniques by means of simple examples, as we did in an appendix to *Principles of Numerical Taxonomy*. It would not have been possible to illustrate the many more numerous methods now extant in taxonomy and still keep the present volume reasonable in size and price. Also, complexity of many of the methods makes computer handling of the data virtually a necessity. To be useful, a discussion of the details of methodology would require explicit advice on computer handling (as well as some mention of character coding and initial processing of the data). This is being done in a forthcoming book by F. J. Rohlf and P. M. Neely (W. H. Freeman and Company).

Limiting ourselves to a brief discussion of computational aspects, we have listed some sources in Appendix B for workers who need to go more deeply into this area and provided general advice and information as well. We have also listed there a number of books and reviews on various other aspects of numerical taxonomy.

Preparation of this book started during the fall semester of 1967 when P.H.A.S. was a Visiting Professor at the University of Kansas. Early drafts of several chapters were read before the Biosystematics and the Numerical Taxonomy Luncheon Groups at that institution and benefited from constructive criticism by their members. Collaboration continued during a visit by P.H.A.S. to the State University of New York at Stony Brook made possible by the Medical Research Council of the United Kingdom. The authors were fortunate to be able to complete the final editing during a summer course in numerical taxonomy at the Institute of Advanced Studies in Oeiras, Portugal, sponsored by the Gulbenkian Foundation. We are indebted to Dr. N. Van Uden for putting all necessary facilities at our disposal.

We are glad to acknowledge three colleagues who have read the entire draft and given us the benefit of extensive constructive criticism. Drs. D. H. Colless (Division of Entomology, C.S.I.R.O., Canberra), J. C. Gower (Rothamsted Experimental Station, Harpenden), and F. J. Rohlf (State University of New York at Stony Brook) contributed greatly to improving our exposition. The chapter on phylogeny was read by Dr. J. S. Farris, the section on kinds of characters by

Dr. J. A. W. Kirsch. Dr. J. H. Strauss read and improved the section on psychiatric classification. Dr. T. J. Crovello contributed to numerous discussions throughout the text. These colleagues were of great assistance in editing the material they reviewed. Various students in several courses on numerical taxonomy at the State University of New York at Stony Brook and at the Estudos Avançados de Oeiras made suggestions for improvements in the text. We are grateful for all this assistance from friends, colleagues, and students and beg their indulgence if we have not always followed their advice.

Our efforts were greatly aided by the meticulous and professional secretarial work of Mrs. Ethel Savarese who saw the manuscript through its many stages of preparation. We are very much in her debt. We would also like to acknowledge the assistance of Mrs. Brenda Jones who typed the Appendixes and Bibliography. The students of the numerical taxonomy course at Stony Brook cooperated in getting a draft copy of the book reproduced in record time. We should single out Richard Stone and Mary Mickevich, who proofread most of the typed copy, and Irving Kornfield, who supervised the assembling. Barbara Torraca and Che Nu Paul were most helpful in checking the indexes.

We are indebted to Abelard-Schuman Ltd., Publishers, for permission to cite a passage from C. Singer, *A History of Biology*. Authors of all sources are identified in the text and are cited in the Bibliography.

We would like to acknowledge here the support given by the Medical Research Council of the United Kingdom (to P.H.A.S.) and by the National Science Foundation of the United States (to R.R.S.) for research in numerical taxonomy. The help of these organizations was crucial during the early development of the subject and their continued support is deeply appreciated.

February 1973

Peter H. A. Sneath
Robert R. Sokal