Biological Identification with Computers

edited by
R.J. Pankhurst

THE SYSTEMATICS ASSOCIATION
SPECIAL VOLUME No. 7

BIOLOGICAL IDENTIFICATION WITH COMPUTERS

Proceedings of a Meeting held at
King's College, Cambridge
27 and 28 September, 1973

Edited by
R. J. PANKHURST
British Museum (Natural History), London, England

Published for the
SYSTEMATICS ASSOCIATION
by
ACADEMIC PRESS
LONDON · NEW YORK · SAN FRANCISCO

ACADEMIC PRESS INC. (LONDON) LTD.
24/28 Oval Road,
London NW1

United States Edition published by
ACADEMIC PRESS INC.
111 Fifth Avenue
New York, New York 10003

Library of Congress Catalog Card Number: 75–15349
ISBN: 0 12 544850 3

PRINTED IN GREAT BRITAIN BY ROBERT MACLEHOSE AND COMPANY LIMITED
PRINTERS TO THE UNIVERSITY OF GLASGOW

Contributors

AITCHISON, R. R., *Department of Botany, University of Cambridge, Downing Street, Cambridge, CB2 3EA, England.*

BEAMAN, J. H., *Department of Botany and Plant Pathology, Michigan State University, E. Lansing, Michigan 48824, U.S.A.*

GÓMEZ-POMPA, A., *Apartado Postal 70–268, University of Mexico, Mexico 20, D.F.*

GOWER, J. C., *Rothamsted Experimental Station, Harpenden, Hertfordshire AL5 2JQ, England.*

GYLLENBERG, H. G., *The Academy of Finland, Lauttasaarentie 1, 00200 Helsinki 20, Finland.*

HALL, A. V., *Bolus Herbarium, University of Cape Town, Rondebosch C.P., South Africa.*

LAPAGE, S. P., *National Collection of Type Cultures, Central Public Health Laboratory, Colindale Avenue, London NW9 5HT.*

MCNEILL, J., *Research Branch, Plant Research Institute, Canada Department of Agriculture, Ottawa, Canada KIA OC6.*

MORSE, L. E., *Gray Herbarium, Harvard University, 22 Divinity Avenue, Cambridge, Massachusetts 02138, U.S.A.*

MOSS, W. W., *Academy of Natural Sciences, 19th and The Parkway, Philadelphia, Pennsylvania 19103, U.S.A.*

NIEMELÄ, T. K., *Department of Microbiology, University of Helsinki, Helsinki 71, Finland.*

PANKHURST, R. J., *Department of Botany, British Museum (Natural History), Cromwell Road, London SW7 5BD.*

PAYNE, R. W., *Rothamsted Experimental Station, Harpenden, Hertfordshire AL5 2JQ, England.*

ROSS, G. J. S., *Rothamsted Experimental Station, Harpenden, Hertfordshire AL5 2JQ, England.*

RYPKA, E. W., *Lovelace-Bataan Medical Center, 5200–5400 Gibson Boulevard, S.E., Albuquerque, New Mexico 87108, U.S.A.*

SHETLER, S. G., *Smithsonian Institution, Washington, DC 20560, U.S.A.*

WALTERS, S. M., *Botanic Gardens, Cambridge, CB2 1JF, England.*

WILKINSON, C., *Department of Biological Sciences, Portsmouth Polytechnic, King Henry 1 Street, Portsmouth PO1 2DY, England.*

WILLCOX, W. R., *National Collection of Type Cultures, Central Public Health Laboratory, Colindale Avenue, London NW9 5HT.*

Preface

While I was serving on the Council of the Systematics Association in 1971, the suggestion was made that I should organize a meeting to discuss identification of biological specimens by computer. This was held at King's College, Cambridge, on 27th and 28th September, 1973, with about 60 delegates attending. The proceedings of this meeting, presented here, constitute one of the first volumes to be published on this subject.

From the point of view of computer science, our subject comes under what is generally called Pattern Recognition. It must be stated at once, however, that nearly all the techniques which seem useful in this context at present require a description of the object by a human observer. Most of the subjects that biologists want to identify are too complex for automatic description by current methods. This situation could always change in the future. Also, it must be made clear that the final decision of which identification to accept, if any, remains in the hands of the biologist. Machines are not taking over a human role here, but just modelling or mimicking our decision processes. Identification is typically a skill held by just a few individuals, gained after many years of practice. Perhaps the automatic methods will make identifications easier, or simply feasible, for the many for whom identification of specimens is, quite rightly, just a means, and not an end in itself.

The use of computers to help in identification is quite recent, and first becomes discernable in the efforts of several bacteriologists in the early 1960s. The next noticeable development is the appearance of a number of computer programs for constructing diagnostic keys around 1970, and at the present time experiments are being made with a wide variety of different methods. It is interesting that the first impulse to develop numerical methods in classification, as opposed to identification, also came from bacteriologists. The volume of effort in classification by computer far exceeds that put into identification. When one reflects that many biologists carry out identifications daily, and that hardly any complete a biological career and avoid this task, and that the proportion of biologists engaged in classifying things is relatively speaking very small, then this distribution of effort may seem odd. It has been suggested that classification is much more challenging a subject than identification, even if the latter has more practical importance. Readers might like to re-assess this situation after studying these proceedings.

It was decided not to include medical diagnosis within the scope of the meeting. Nonetheless it is interesting to compare the developments in medical diagnosis by computer from the late 1950s onwards with biological identification. Medical work has taken a different course, with much prominence being

given to probabilistic methods. It might be that the problems in the two fields are not as fundamentally different as is often thought. Some references to medical diagnosis are given in the bibliography.

The meeting was planned so that the technical programme was not overcrowded, and included an exhibition and a number of demonstrations of computer programs. Consequently, not every paper which is published here was formally presented at the time. However, the material covered here is all directly ralated to papers presented, discussions, exhibits or demonstrations which took place at the meeting. One paper, which was presented by Dr M. Freudenthal of the National Geological Museum at Leiden, concerning identification applications with a geological data base, was not submitted for publication.

A short film, entitled "Computer Graphics in Fungal Identification", by B. Kendrick, was shown during the meeting. No account of this is given, except for inclusion in the bibliography. I am indebted to Prof. Kendrick for the loan of this film.

My thanks are due to Prof. V. H. Heywood, President of the Systematics Association, for encouragement and assistance throughout, and to Rosemary Aitchison for acting as organizing secretary. We were pleased to welcome Mr J. Gilmour as a session chairman. A grant made by the Royal Society towards speakers' travel expenses is gratefully acknowledged.

R. J. P.

May, 1975

Contents

Historical Introduction

1 | Traditional Methods of Biological Identification

S. M. WALTERS

University Botanic Garden, Cambridge, England

Abstract: Identification of biological specimens can be defined as the practice of assigning the specimens to known, named taxa. It is a necessary activity which most biologists undertake at some time, but in which taxonomists are specially concerned since they produce both the classifications and the tools for identification. Some of these tools, especially the artificial dichotomous key, have achieved particular prominence, but surprisingly little attention has been paid to the relative merits of different tools. The advent of computers and numerical taxonomy have been beneficial in stimulating taxonomists to ask some of these practical questions. The paper is illustrated by reference to the classification of the Umbelliferae.

Key Words and Phrases: identification, dichotomous keys, Umbelliferae, history of taxonomy, taxonomic description

The practice of identification is necessarily such a common experience for all biologists who are working with whole organisms that it seems very appropriate to begin a meeting devoted to automatic identification with a brief outline of the methods actually employed and the history of their use. My only reluctance to do this arises from the way in which my own knowledge of the subject is restricted to the higher plants, as I am aware that we have at this meeting specialists in several other groups. I believe, however, that the important aspects of the traditional method can be conveniently illustrated from the history of botanical classification, and that much of my thesis could have been similarly illustrated by zoological examples.

Identification of biological specimens can be defined as the practice of assigning a given specimen to a known, named taxon. All biologists are involved in identification (if only as laymen in the daily round of affairs), but the taxonomist is specially concerned, since he produces or alters the classifications of organisms as well as providing the tools by which his fellow-scientists can

Systematics Association Special Volume No. 7, "Biological Identification with Computers", edited by R. J. Pankhurst, 1975, pp. 3–8. Academic Press, London and New York.

identify their specimens. It is very instructive to look at the history of taxonomy, and to attempt to trace the interrelations of naming, classification, and identification.

Logically, it would seem that this order of activity must be operating: a taxon is recognized by naming, its position is decided in a hierarchical classification, and then specimens can be assigned to it by a procedure of identification. The history of biological taxonomy does not, however, reveal this process in such a logical sequence, and a little thought soon tells us why it cannot be so. The most obvious complicating factor in the process is that taxonomic knowledge is increasing all the time, so that attempts at identification continually show inadequacies in the existing system of names and hierarchical groups, and new taxa and systems are made to accommodate the new knowledge. For those who are interested in pursuing implications of these thoughts, I unhesitatingly recommend an excellent paper by E. G. Voss (1952) and the many references given in it.

To illuminate our subject, I have selected a single, very familiar group of flowering plants, the members of the carrot family Umbelliferae. There are several reasons for my choice; but a very special reason is that we have just celebrated the tercentenary of Robert Morison's monograph on this family, published in Oxford in 1672, and conveniently described by Hedge (1973).

Morison's monograph is an impressive work which reminds us that, in cases where a modern flowering plant family has many common European representatives, the naming and classification of the family took shape in Medieval Europe, long before Linnaeus and the eighteenth century standardization of nomenclature and classification. The implications of this "European bias" for Angiosperm taxonomy in general I have discussed elsewhere (Walters 1961, 1962), and these general themes lie outside the present field of discussion. The Umbelliferae, however, are not only common, but have for centuries been known for their culinary, medicinal and even poisonous properties; for these reasons also they were the subject of description and illustration in Classical and Medieval writings (Fig. 1), and their correct identification was a matter of some practical importance (see French, 1971). It is, therefore, not surprising to find that Morison provided detailed illustrations of the "seeds" (actually the fruits) of many different kinds of umbellifers (Fig. 2). What is perhaps less expected is that the monograph also includes bracketed diagrams which function to some extent both as a classificatory device ("conspectus") and as an identification tool or "key" (Fig. 3).

According to Voss, several biological writers in the second half of the seventeenth century used such diagrams, and Nehemiah Grew described their

use for identification as early as 1676. Curiously enough, the term "clavis" (key) was apparently not used in connexion with such diagrams until Linnaeus so used it in 1736 (and then with reference to a diagram in which he was classifying, *botanists*, not plants!). The credit for explicit and systematic use of modern artificial dichotomous keys *for identification* is usually given to Lamarck in his "Flore Française" (1778), and after this pioneer work most nineteenth century Floras supplied such keys as a matter of course.

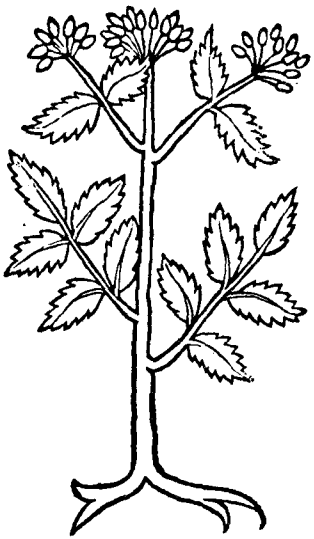

FIG. 1. "Carvi" (Caraway), illustration from Herbal "Ortus Sanitatis", Mainz 1491 (taken from Arber (1938) p. 165).

It would therefore appear from a comparative study of botanical writings over the seventeenth, eighteenth and nineteenth centuries that the *description* and the *illustration* were the earliest identificatory aids, and that the modern, standard, artificial key gradually developed from a diagram which, by grouping and differentiating the different "kinds" of plants (in the case of Umbelliferae most of these "kinds" correspond to the modern genera), served the purposes both of classification and identification. The rigid, logical separation between a synopsis of classification or *conspectus* on the one hand and an artificial *key* on the other seems to have been relatively late in developing. Indeed, examples could still be found in recent Floras where the "keys" provided seem to be uncomfortably attempting to satisfy both these requirements and achieving neither aim as a result.

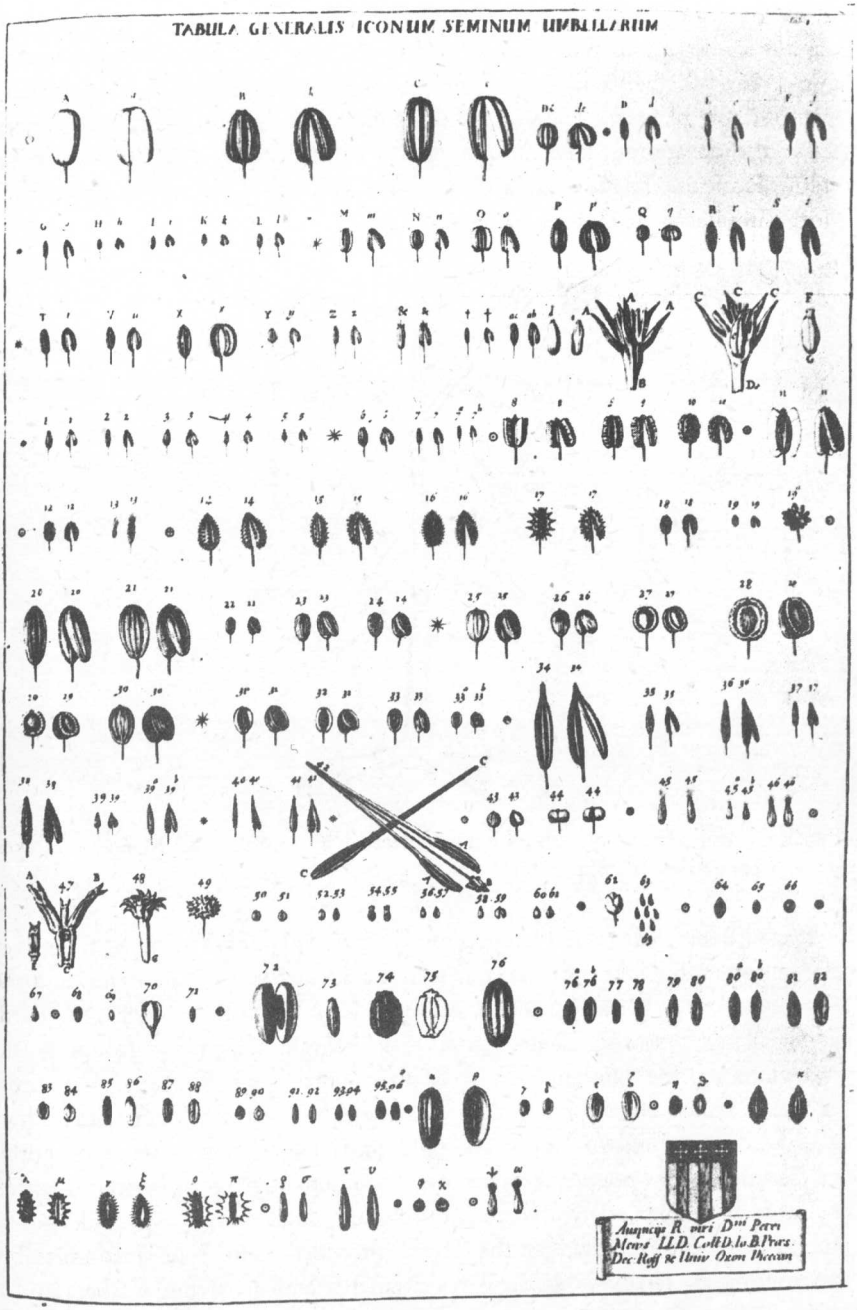

FIG. 2. Fruits of Umbelliferae, from Morison (1672).

To conclude my survey, I can turn to the treatment of the Umbelliferae in volume 4 of the "Flora of Turkey" (Davis *et al.*, 1972), published exactly three centuries after Morison. The first thing to say is that the continuity (some would

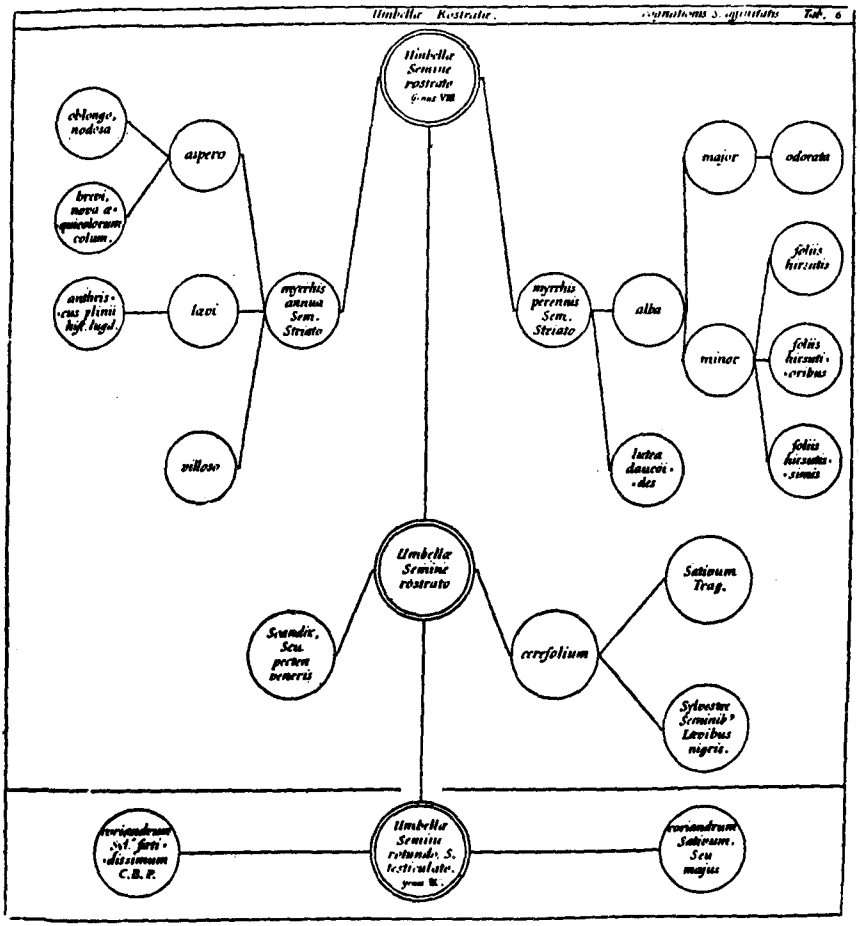

FIG. 3. Diagram of Umbelliferae (Part), from Morison (1672).

say conservatism) of botanical taxonomy is such that Robert Morison, if he were to come alive again, could use this book without much difficulty. Many of the generic names are the same; the nomenclature remains in Latin; there are detailed illustrations of the fruits, and there are "diagrams" or keys to aid identification. (The text is in English, which should also cause him little difficulty!) The "Flora of Turkey" authors have, however, made one significant

departure: they have produced a "multi-access key" which enables a much freer use of characters for generic identification. I feel sure Morison would have approved; it would seem that the rigid orthodoxy of the dichotomous key, which developed long after Morison's time, is at last being challenged and re-thought. I believe that we need to look at all methods of biological identification again, objectively and practically, and choose the ones best suited to our task. The dichotomous key has many advantages, but it is not necessarily the only or the appropriate device in every case. One of the purposes of this meeting would be, I hope, to explore afresh these practical questions.

REFERENCES

ARBER, A. (ed. 2, 1938). "Herbals: Their Origin and Evolution". Cambridge University Press.

DAVIS, P. H., ed. (1972). "Flora of Turkey". Edinburgh University Press.

FRENCH, D. H. (1971). *In* "Biology and Chemistry of the Umbelliferae" (V. H. Heywood, ed.), pp. 385–412. Academic Press, London and New York.

HEDGE, I. C. (1973). Umbelliferae in 1672 and 1972. *Notes from the Royal Botanic Garden, Edinburgh.* **32**(2), 151–160.

LAMARCK, J. B. P. (1778). "Flore Française". Paris, Imp(rimerie) Royale.

MORISON, R. (1672). "Plantarum Umbelliferarum Distributio Nova". Oxonii, e Theatro Sheldoniano.

VOSS, E. G. (1952). The history of keys and phylogenetic trees in systematic biology. *J. Scient. Labs. Denison Univ.* **43**, 1–25.

WALTERS, S. M. (1961). The shaping of Angiosperm taxonomy. *New Phytol.* **60**, 74–84.

WALTERS, S. M. (1962). Generic and specific concepts and the European flora. *Preslia* **34**, 207–226.

Survey