Dallwitz, M.J. 1992. A comparison of matrix-based taxonomic identification systems with rule-based systems. In 'Proceedings of IFAC Workshop on Expert Systems in Agriculture', pp. 215–8. (Ed. F. L. Xiong.) (International Academic Publishers: Beijing.) Also available at http://delta-intkey.com

A comparison of matrix-based taxonomic identification systems with rule-based systems

1992

M. J. Dallwitz

Abstract

Taxonomic identification systems based on character-taxon matrices usually perform better than rule-based systems. Also, matrix data can be used for other purposes, such as description writing, classification, and information retrieval. Most matrix-based systems do not use probabilities, but this is seldom a significant limitation.

Introduction

In biological terminology, *classification* is the process of defining and naming classes of organisms. These classes are called *taxa*. *Identification* is the process of assigning a specimen to a (pre-existing) taxon. The name of the taxon can then be used as an index to find known information about the taxon, and therefore about the specimen itself (e.g. whether it is a pest, and, if so, how it can be controlled). Alternatively, information about the specimen can be added to the body of knowledge about the taxon. The science of biological description, classification, and identification is called *taxonomy*.

Both classification and identification must be based on *comparative* descriptions. The descriptions are framed (implicitly or explicitly) in terms of a list of characters. A *character* is a set of *states* that describe some aspect of the organisms to which the character is to be applied. The number of states may be finite (e.g. colour of petals: 1. white; 2. yellow; 3. red) or (in principle) infinite (e.g. width of head, in millimetres). Descriptions of taxa (or specimens) can be represented as a table with each row corresponding to a taxon and each column to a character. The entry in each cell of the table consists of the value or values that the character takes for the taxon. Entries may be missing if the value is unknown or the character is inapplicable to the taxon. The table is generally called a *data matrix*.

There are many kinds of aids to identification[1]. The traditional one, and still the most commonly used, is the *diagnostic key*, which has been in use for more than 200 years. Table 1 shows the first few lines of a key to the grass genera of the Australian Capital Territory. Each number at the left of the key labels a set of character states. To identify a specimen, the key user starts at label 1, and selects the character state that best describes the specimen. This state points to another label, or to a taxon name. In the former case, the states at the new label are examined, and the appropriate one selected. This process is continued until a name is reached.

Computers can be used to produce conventional diagnostic keys[2–6]. However, they can be more effective aids to identification when used interactively. In a conventional key, a predetermined set of characters must be used to identify a given specimen, whereas, in a well-designed interactive system, characters can be avoided if they are difficult or impossible to use. Also, an error in the construction or use of a conventional key almost inevitably leads to a wrong identification, whereas an interactive system can be made tolerant of errors, both those in the data matrix and those made by the user. Using an interactive identification system is similar to using a key, in that character states exhibited by the specimen are selected until a name is reached. The differences lie in the flexibility with which characters can be selected, and the other facilities that may be offered to assist the process.

Table 1. Part of a computer-generated key to the grass genera of the Australian Capital Territory[45]

1.	Female-fertile spikeletsw ithp roximali ncompletef lorets	2
	Female-fertile spikeletsw ithout proximal incomplete florets	25
2.	Rachilla prolongedb eyond the uppermost female-fertile floret	3
	Rachilla notp rolongedb eyondt he uppermost female-fertile floret ...	5
3.	Ovarya pexg labrous;s tylesf used basally	4
	Ovarya pexh airy;s tylesf reet ot heir bases	*Arrhenatheru m*
4.	Hilums hort;l eafb ladesb road	*Phragmite s*
	Hiluml ong-linear; leaf blades narrow	*Ehrharta*
5.	Spikeletsw ithb ractiformi nvolucres	*Themeda*
	No bractiform involucres	6

Matrix-Based Versus Rule-Based Systems

Two main approaches have been tried in constructing interactive identification systems: programs which directly use a data matrix[7–24]; and rule-based expert systems[25–28]. Frame-based expert systems do not seem to have been much used for identification, although Edwards, Morse and Fielding[29] consider that they would be superior to rule-based systems. It is not clear whether frames offer significant advantages over data matrices in this context.

Rule-based expert systems are usually designed to mimic the methods of human experts. Expert taxonomists can certainly carry out identifications more quickly and reliably than non-experts, so it would seem logical to try to capture their knowledge and methods for use by others. There are two basic difficulties with this approach.

Firstly, some aspects of the knowledge and methods cannot be captured as rules, or easily conveyed to others. An important component of expert identification is recognizing general appearance. (We all do this in identifying familiar objects (e.g. faces) and taxa (e.g. models of cars).) This process may be rationalized as a set of rules, but, clearly, this is not the way it is actually done. Also, experts are familiar with the meanings and delimitations of character states (especially if they have defined the characters themselves!). We can attempt to convey this information by detailed explanations and illustrations, but this cannot entirely take the place of experience.

Secondly, the expert knowledge and methods that can be readily captured are not necessarily optimal. The limitations of the human mind make it necessary to organize knowledge as classifications and rules. Computers, on the other hand, can easily store, interrogate, and manipulate large quantities of raw data, and may be able to quickly derive from them optimal procedures to suit any situation.

An important advantage of basing an identification system on a data matrix is that the matrix can also be used for other purposes: for generating descriptions, keys (which are still *de rigueur* for taxonomic publications), and classifications, and for information retrieval[19–24,30–52]. Also, a data matrix can help the expert to gain insights, not only from cladistic and phenetic classification programs, but also from sufficiently versatile identification programs; whereas a rule-based system cannot provide new information to the expert. In fact, the rules used in expert identification systems are sometimes derived from data matrices[26].

Taxonomic data are usually gathered and presented as descriptions, keys, or matrices, not as rules. The rules used in expert identification systems have usually been derived from these other forms of data. Atkinson and Gammerman[25] derived their rules from published descriptions, and also used keys directly in their system. Woolley and Stone[28] derived their rules from consultations between a domain expert and a knowledge engineer, but the former had already constructed a complete data matrix. Fermanian and Michalski[26] automatically induced sets of rules from a data matrix. The rules used in the expert system were selected from these sets, and modified, if necessary, on the basis of the expert's opinion or the results of preliminary tests of the system.

The construction of rules has tended to be a difficult and time-consuming process, involving collaboration between a domain expert and a knowledge engineer[26,28], but software is becoming available to make this easier[26]. This difficulty is greatly compounded by the fact that when new taxa need to be added to a system (which is almost inevitable), existing rules may need to be revised. (There is a similar problem with the manual construction of identification keys.)

The main function of the rules in an expert identification system is to decide which character or characters should be used in various circumstances in the course of identifications. The number of circumstances that can be covered in this way is limited by the effort required to generate the rules, and the computer resources required to store and manipulate them. Thus, the expert-system knowledge base tends to represent a sparse subset of the information present in a complete data matrix. (This is also true of the information contained in diagnostic keys.) In matrix-based systems such as ONLINE[12] and INTKEY[20], this function is carried out by an algorithm that can be applied at any stage of an identification. The merit of a character is calculated partly from its separating power, and partly from its 'weight' — a measure of reliability or ease of use, assigned by the expert. The algorithm is, in effect, generating rules as required. Characters are displayed in order of merit, but the user is free to choose any character. The choice will be influenced by the accessibility of the character in particular circumstances (e.g. whether parts such as flowers and fruits are present, whether it is convenient to use a microscope or a chemical test); by the user's own assessment of the separating power of the character for a particular specimen (an experienced user is often aware that some attributes of a specimen are unusual, and are therefore likely to lead to a quick identification); and by the user's confidence in his ability to understand and use the characters.

An expert can often improve the automatic selection of rules for interactive systems (or of characters for key generation) by applying knowledge not incorporated in the matrix. For example, a character may be generally difficult to use and error prone, but may give clear-cut results when applied to certain taxa. This kind of information could be made accessible to character-selection algorithms in the form of weights for individual cells of the matrix, in addition to the character weights, but, as far as I know, this has not yet been done. Another way of incorporating such information in matrices is via special characters, which are coded only for those taxa to which they can be easily applied. This method requires no special programming, and is already in use.

The ability to make use of probabilities has been cited as an advantage of rule-based identification systems[28]. It is true that most expert-system shells incorporate probabilistic reasoning, and many matrix-based identification systems currently do not. However, this is not an intrinsic limitation of matrix-based systems. For example, the key-generation program of Payne[6] uses probabilities, as do some programs for identification of micro-organisms[53–55]. One of the reasons that probabilities are not more widely used in identification is the lack of suitable information. Where possible, taxonomists tend to avoid characters that require the use of probabilities: they would rather try to find alternative characters.

The probabilistic information actually used in rule-based identification systems often seems to be a degraded form of information which may originally have been in an exact or an alternative probabilistic form. For example, in the system of Fermanian and Michalski[26], rules such as 'Weed is Bentgrass if ligule is round; confidence level 65%' have apparently been derived, at least in part, from knowledge of the ligule shape for each taxon in the data set. Atkinson and Gammerman[25] use rules of the form 'If an unknown British umbellifer is found in sand dunes then it is 95% probable that it belongs to the set {*Crithmum maritimum*, *Daucus carota*, *Eryngium maritimum*, *Pimpinella saxifraga*}'. They claim that this type of information requires a 'weaker commitment' from the expert who formulates it than knowing, for example, what proportion of specimens of *Crithmum maritimum* live in dunes as opposed to other habitats. I think that most taxonomists would find it far easier to estimate the latter type of information. It is implicit in the distribution and other properties of the specimens on which each taxon description is based, whereas the former information would require the gathering of statistics across all of the taxa.

Another advantage claimed for rule-based systems is their ability to explain their reasoning processes; in particular, the reason why a particular rule (character) has been selected to be the next used[25,28]. The example given by Woolley and Stone[28] has the form 'The rules used so far, *x* and *y*, have reduced the likely taxa to *a* and *b*, and rule *z* will distinguish them'. Matrix-based systems can be made to supply

even more detailed information of this kind. For example, INTKEY can display a list of the taxa remaining in contention; the information entered so far about the specimen; a list of characters with numbers indicating their separating power for the remaining taxa; full, partial, or diagnostic descriptions of the remaining taxa; the differences between the remaining taxa; the differences between the specimen and the remaining taxa; and the differences between the specimen and any of the eliminated taxa.

Matrix-based identification systems are capable of providing good performance with quite large data sets. For example, an INTKEY data set for the grass genera of the world[14,15] contains 787 taxa and 495 characters. The data files occupy about 1MB of disk space. On an IBM-compatible PC, the program occupies about 300kB of disk space, and requires about 550kB of available memory. The response time for a simple operation, such as the processing of an attribute in an identification, is less than 1 second on a 33MHz 386 machine.

Rule-based systems (apart from medical systems) seem to have been tested only on small data sets. Atkinson and Gammerman[25] used 4 characters for the rule-based part of their system, and about 50 taxa. Woolley and Stone[28] used 22 characters and 12 taxa, and state that memory and speed constraints (on an IBM XT) would limit the size of the system to about double this. Fermanian and Michalski[26] used 11 characters and 37 taxa.

Summary

Some important aspects of the identification skills used by experts cannot be captured in computerized identification aids, and those that can are not necessarily optimal. Matrix-based identification systems have already reached a high standard of performance, and the data matrices are also valuable for other purposes. Most taxonomic information is gathered and published in a form more akin to data matrices than to rules, so rule-based systems are more difficult to construct, and the information they contain tends to be sparse. However, rule-based systems can readily represent and use qualifying information, such as probabilities and special cases, which are not used in most current matrix-based systems. An empirical comparison of the performance of the two types of system (similar to the comparison by Fermanian and Michalski[26] of a key and a rule-based expert system) would be interesting. This would need to be a substantial project, involving experts in the creation of both types of system, in order to avoid setting up straw men.

References

1. Pankhurst, R. J. (1978). Biological identification — the principles and practice of identification methods in biology. 104 pp. (Edward Arnold: London.)
2. Dallwitz, M. J. (1974). A flexible computer program for generating identification keys. Syst. Zool. 23, 50–57.
3. Hall, A. V. (1970). A computer-based system for forming identification keys. Taxon 19, 12–18.
4. Morse, L. E. (1971). Specimen identification and key construction with time-sharing computers. Taxon 20, 269–282.
5. Pankhurst, R. J. (1970). A computer program for generating diagnostic keys. Comput. J. 13, 145–151.
6. Payne, R. W. (1975). Genkey: a program for constructing diagnostic keys. In ‘Biological Identification with Computers’. (Ed. R. J. Pankhurst.) pp. 65–72. (Academic Press: London.)
7. Aiken, S. G., and Dallwitz, M. J. (1991). Festuca (Poaceae) of North America: interactive identification and information retrieval. Flora Online 26.
8. Colosimo, A., Rota, E., and Omodeao, P. (1991). A Hypercard program for the identification of biological specimens. Comput. Applic. Biosc. 7, 63–69.
9. Duncan, T., and Meacham, C. A. (1986). Multiple-entry keys for the identification of angiosperm families using a microcomputer. Taxon 35, 492–494.
10. Forget, P. M., Lebbe, J., Puig, H., and Hideux, M. (1986). Microcomputer-aided identification: an application to trees from French Guiana. Bot. J. Linn. Soc. 93, 205–223.
11. Lebbe, J., Nilsson, S., Praglowski, J., and Vignes, R. (1987). A microcomputer-aided method for identification of airborne pollen grains and spores. Grana 26, 223–229.
12. Pankhurst, R. J., and Aitchison, R. R. (1975). An on-line identification program. In ‘Biological Identification with Computers’. (Ed. R. J. Pankhurst.) pp. 181–185. (Academic Press: London.)

13. Rhoades, F. M. (1987). Synoptic keys for introducing students to bryophytes and lichens of northwestern USA, and a computer program, ASKATAXA, to access them. Flora Online 11.
14. Watson, L., and Dallwitz, M. J. (1988). 'Grass Genera of the World: Illustrations of Characters, Classification, Interactive Identification, Information Retrieval.' With microfiches, and floppy disks for MS–DOS microcomputers. (Research School of Biological Sciences, Australian National University: Canberra.)
15. Watson, L., and Dallwitz, M. J. (1991). Grass genera of the world: an INTKEY package for automated identification and information retrieval of data including synonyms, morphology, anatomy, physiology, cytology, classification, pathogens, world and local distribution, and references. 2nd edition. Flora Online 22.
16. Watson, L., and Dallwitz, M. J. (1991). The families of angiosperms: automated descriptions, with interactive identification and information retrieval. Aust. Syst. Bot. 4, 681–695.
17. Watson, L., Gibbs Russell, G. E., and Dallwitz, M. J. (1989). Grass genera of southern Africa: interactive identification and information retrieval from an automated data bank. S. Afr. J. Bot. 55, 452–463.
18. Wilson, J. B., and Partridge, T. R. (1986). Interactive plant identification. Taxon 35, 1–12.
19. Bruhl, J. J., Watson, L., and Dallwitz, M. J. (1992). Genera of Cyperaceae: interactive identification and information retrieval. Taxon 41, 225–235.
20. Dallwitz, M.J. (in press). DELTA and INTKEY. In 'Proceedings of Workshop on Artificial Intelligence and Modern Computer Methods for Systematic Studies in Biology'. (University of California: Davis.)
21. Gibbs Russell, G. E., Watson, L., Koekemoer, M., Smook, L., Barker, N. P., Anderson, H. M., and Dallwitz, M. J. (1990). 'Grasses of Southern Africa. An identification manual with keys, descriptions, classification and automated identification and information retrieval from computerized data.' Memoirs of the Botanical Survey of South Africa No. 58. 437 pp. (Botanical Research Institute: Pretoria.)
22. Morse, L. E. (1974). Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. Publs. Mich. St. Univ. Mus., biol. ser. 5, 1–128.
23. Pankhurst, R. J. (1986). A package of computer programs for handling taxonomic databases. Comput. Applic. Biosc. 2, 33–39.
24. Watson, L., Dallwitz, M. J., Gibbs, A. J., and Pankhurst, R. J. (1988). Automated taxonomic descriptions. In 'Prospects in Systematics'. (Ed. D. L. Hawksworth.) pp. 292– 304. (Clarendon Press: Oxford.)
25. Atkinson, W. D., and Gammerman, A. (1987). An application of expert systems technology to biological identification. Taxon 36, 705–714.
26. Fermanian, T. W., and Michalski, R. S. (1989). WEEDER: an advisory system for the identification of grasses in turf. Agron. J. 81, 212–216.
27. Shortliffe, E. H. (1976). Computer-based medical consultations: MYCIN. (American Elsevier/North-Holland: New York.)
28. Woolley, J. B., and Stone, N. D. (1987). Application of artificial intelligence to systematics: SYSTEX— a prototype expert system for species identification. Syst. Zool. 36, 248–267.
29. Edwards, M., Morse, D. R., and Fielding, A. H. (1987). Expert systems: frames, rules or logic for species identification? Comput. Applic. Biosc. 3, 1–7.
30. Boswell, K. F., Dallwitz, M. J., Gibbs, A. J., and Watson, L. (1982). 'Plant Viruses: Descriptions and Keys from VIDE.' 2 microfiche. (Research School of Biological Sciences, Australian National University: Canberra.)
31. Boswell, K. F., Dallwitz, M. J., Gibbs, A. J., and Watson, L. (1986). The VIDE (Virus Identification Data Exchange) project: a data bank for plant viruses. Review of Plant Pathology 65, 221–231.
32. Brunt, A., Crabtree, K., and Gibbs, A. (eds) (1990). Viruses of tropical plants. Descriptions and lists from the VIDE database. (CAB International: Wallingford.)
33. Dallwitz, M. J. (1980). A general system for coding taxonomic descriptions. Taxon 29, 41–46.
34. Dallwitz, M. J. (1984). Automatic typesetting of computer-generated keys and descriptions. In 'Databases in Systematics'. (Eds R. Allkin and F.A. Bisby.) Systematics Association Special Volume No. 26. pp. 279–290. (Academic Press: London.)
35. Dallwitz, M. J. (1989). Diagnostic descriptions from INTKEY and CONFOR. DELTA Newsletter 3, 8–13.

36. Dallwitz, M. J., and Paine, T. A. (1986). User's guide to the DELTA system: a general system for processing taxonomic descriptions. 3rd edition. CSIRO Aust. Div. Entomol. Rep. No. 13, 1–106.
37. Dallwitz, M. J. and Zurcher, E. J. (1988). User's guide to TYPSET. A computer typesetting program. 2nd edition. CSIRO Aust. Div. Entomol. Rep. No. 18, 1–25
38. Pankhurst, R. J. (1978). The printing of taxonomic descriptions by computer. Taxon 27, 35–38.
39. Partridge, T. R., Dallwitz, M. J., and Watson, L. (1988). A primer for the DELTA system on MS-DOS and VMS. 2nd edition. CSIRO Aust. Div. Entomol. Rep. No. 38, 1–17.
40. Watson, L., Aiken, S. G., Dallwitz, M. J., Lefkovitch, L. P., and Dubé', M. (1986). Canadian grass genera: keys and descriptions in English and French from an automated data bank. Can. J. Bot. 64, 53–70.
41. Watson, L., Clifford, H. T., and Dallwitz, M. J. (1985). The classification of the Poaceae: subfamilies and supertribes. Aust. J. Bot. 33, 433–484.
42. Watson, L., and Dallwitz, M. J. (1980). 'Australian Grass Genera: Anatomy, Morphology, and Keys.' 209 pp. (Research School of Biological Sciences, Australian National University: Canberra.)
43. Watson, L., and Dallwitz, M. J. (1981). An automated data bank for grass genera. Taxon 30, 424–9 + 2 microfiche.
44. Watson, L., and Dallwitz, M. J. (1983). 'Genera of Leguminosae-Caesalpinioideae: Anatomy, Morphology, Classification and Keys.' 95 pp. (Research School of Biological Sciences, Australian National University: Canberra.)
45. Watson, L., and Dallwitz, M. J. (1985). 'Australian Grass Genera: Anatomy, Morphology, Keys and Classification.' 2nd edition. 165 pp. (Research School of Biological Sciences, Australian National University: Canberra.)
46. Watson, L., and Dallwitz, M. J. (1986). 'Grass Genera of the World.' 1st edition, 3 microfiche. (Research School of Biological Sciences, Australian National University: Canberra.)
47. Watson, L., and Dallwitz, M. J. (1989). 'Grass Genera of the World.' 4th edition, 5 microfiche. (Research School of Biological Sciences, Australian National University: Canberra.)
48. Watson, L., and Dallwitz, M. J. (in press). 'The Grass Genera of the World.' 1000 pp. (approximately). (CAB International: Wallingford.)
49. Watson, L., Dallwitz, M. J. and Johnston, C. R. (1986). Grass genera of the world: 728 detailed descriptions from an automated database. Aust. J. Bot. 34, 223–230.
50. Watson, L., Damanakis, M., and Dallwitz, M. J. (1988). 'The Grass Genera of Greece: Descriptions, Classification, Keys.' (In Greek.) (University of Crete: Heraklion.)
51. Webster, R. D. (1987). 'The Australian Paniceae (Poaceae).' (J. Cramer: Berlin & Stuttgart.)
52. Xu Zhu and Dallwitz, M. J. (in preparation). English, Chinese, Japanese and Russian descriptions and keys for 13 species of *Elymus* in China.
53. Gyllenberg, H. G., and Niemelä, T. K. (1975). New approaches to automatic identification of micro-organisms. In 'Biological Identification with Computers'. (Ed. R. J. Pankhurst.) pp. 121–136. (Academic Press: London.)
54. Möller, F. (1962). Quantitative methods in the systematics of Actinomycetales. IV. The theory and application of a probabilistic identification key. Giorn. Microbiol. 10, 29–47.
55. Willcox, W. R., and Lapage, S. P. (1975). Methods used in a program for computer-aided identification of bacteria. In 'Biological Identification with Computers'. (Ed. R. J. Pankhurst.) pp. 103–119. (Academic Press: London.)