Computer-Aided Approaches to Identification. III (Conclusion). Modern Databases

R. FROESE
ICLARM
MC P.O. Box 1501
Makati, Metro Manila 1299
Philippines

I. ACHENBACH
Institut für Meereskunde
Universität Kiel
D-2300 Kiel 1
Fed. Republic of Germany

C. PAPASISSI
Department of Biology
Section of Animal Biology
University of Patras
Patras, Greece

Abstract

Modern databases can be successfully used to develop computer-based identification systems. In a first case study, fish larvae were identified with an average of three easily obtained characters. In a second case study, 17 fish diseases out of 20 were diagnosed directly, using an average of six gross signs of a disease. A comparison of computer-based identification systems reveals that (i) expert systems can be viewed as a very comfortable modern version of traditional identification keys, (ii) numerical methods open the way to automatic identification but depend on expensive hard- and software, and (iii) modern databases represent powerful, yet easy to use identification systems.

Introduction

This is the third and last in a series of three Fishbyte articles presenting computer-based methods for identification. In the preceding articles, an expert system (Froese et al. 1989) and a numerical method (Froese 1989) were introduced. Here, we will have a closer look at modern databases, as commonly used in botany and microbiology for the identification of plants and bacteria, respectively. DELTA, for example, is a sophisticated database on grasses containing 300 morphological, physiological and cytological characters as well as the geographical distribution of 712 genera (Abott et al. 1985). The International Center for Living Aquatic Resources Management (ICLARM) is in the process of developing a large database called FISHBASE to summarize information comparable in scope to that normally provided in the species synopses published by the Food and Agriculture Organization (FAO) of the United Nations. FISHBASE provides not only fast access to information on a given species or population but also allows for species identification (Froese and Papasissi 1990) as well as for diagnosis of diseases (Achenbach and Froese 1990).

This article gives a short description of the database approach to identification and then concludes this series with a comparison of the three methods presented.

What is special about modern databases?

In order to efficiently use a database such as, e.g., dBASE III PLUS, the user must know the query language of the software package, the structure of the database, the field names used, and the type and format of the possible entries in a field. A typical query command (adapted from p. 159 in Jones (1988)) would look like this:

REPORT FORM INCOME FOR DTOC(EXPDATE) > "07/31/85" .AND. DTOC(EXPDATE) < "09/01/85" TO PRINT

In a large database with several hundred fields in several interlinked tables, this approach is clearly inappropriate. In contrast, modern databases such as, e.g., DataEase 4.2 provide two powerful features for searching:

"query by forms" (also known as "query by example"), which allows the user to enter the characters he or she is looking for into an empty form on the screen (Fig. 1); and

"choice fields", which can display a window with a list of all possible entries for a given field. The user can pick an entry from the list, thus (i) knowing which entries are possible and (ii) reducing the danger of typographical errors.

Boolean operators such as "AND", "OR", "NOT", ">", and "<" as well as wildcards such as "?" for one character and "*" for any number of characters can be used in every field. This allows the use in a search, of characteristics of which the user may be unsure.

- All larvae could be identified;
- Half of the larvae could be identified with only three characters (minimum one character, maximum six characters);
- Four identifications were performed using morphometric charcters only, five using descriptive characters only, and eleven using both morphometric and descriptive characters;
- Overall, 29 morphometric and 32 descriptive characters were used for identification.

For the diagnosis of fish diseases the results were:

- Of the 20 fish diseases tested, 17 were diagnosed directly. The other three were identified as one of two possible diseases;

Discode	:	00017			Main Ref. : 10006
Disease I	:	virus disease			
Disease II	:	SVC		Pic :	
Other names	:	Spring Viraemia of Carp, Dropsy, Acute Dropsy, Chronic Dropsy			
Caus. agent	:	Rhabdovirus	carpio		

Behavior or part of body afflicted by disease

Skin and/or fins	:	yes	Anus	:	yes
Gills	:	yes	Swimming behavior	:	yes
Eyes	:	yes	Other behavior	:	yes
Belly	:	yes	Other symptoms	:	no
Viscera	:	yes			

Fig. 1. Screen display of page 1 of the form used to diagnose fish diseases.

How can modern databases be used for identification?

The major scientific input involved in developing an identification system structured around a database is the determination of the appropriate characters in all their possible presentations. For example, for the diagnosis of fish diseases we had to identify which parts of the body could be afflicted by a disease and what their possible appearance would then be. This information then was structured into fields and possible choices. The result is a self-explanatory form that can be used for data entry as well as for diagnosis of diseases (Fig. 1).

To date, two studies have tested the utility of FISHBASE for identification purposes, one pertaining to fish larvae (Froese and Papasissi 1990), the other to fish diseases (Achenbach and Froese 1990). The results for fish larvae can be summarized as follows:

Differential diagnosis would have been possible using microscopic examination or histological methods;
- The maximum number of symptoms required for diagnosis was 11, while the minimum was three. On the average, six symptoms were needed for a diagnosis;
- All symptoms used referred to external, gross features of the diseases.

Both studies point out that it is important to ensure a diagnostic or an identification using traditional methods, i.e., check against a complete description of the disease or of the species, respectively (Fig. 2).

In the light of these results, modern databases prove to be an appropriate tool for building identification and/or diagnostic systems. The average of three easily obtained characters for an identification is a distinct advantage over traditional
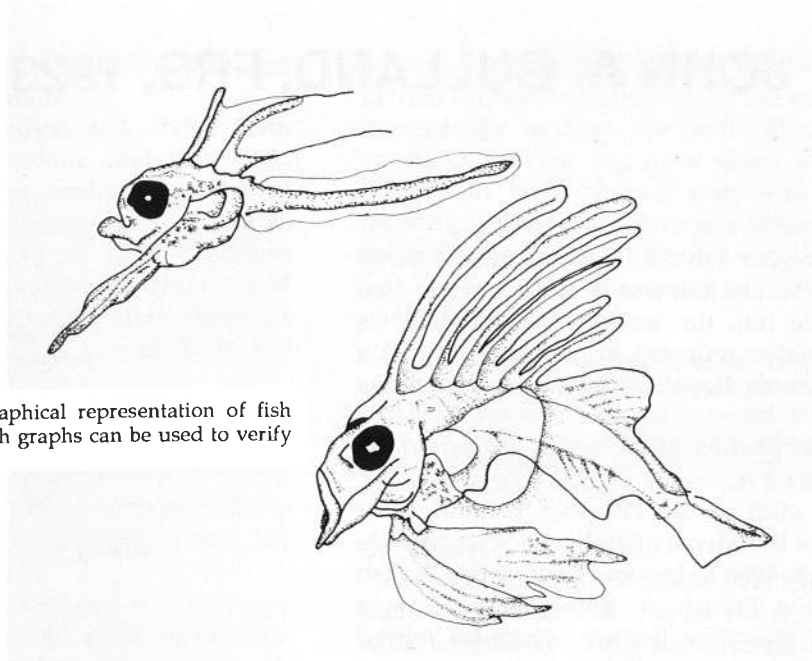
Fig. 2. Some examples of the graphical representation of fish larvae included in FISHBASE. Such graphs can be used to verify an identification.

identification keys. The average of six gross signs of a disease for a diagnosis is a remarkable advantage for a discipline which had failed, to date, to produce simple "diagnostic keys".

Comparing expert systems, numerical methods and databases

Expert systems, numerical methods and data-bases can all be successfully used to identify aquatic animals or diseases. There are, however, essential differences:

After an enthusiastic start the limitations of expert systems are now seen more realistically. Although a lot of prototypes have been developed, only very few systems are used in a professional context (Feigenbaum et al. 1988; Mertens et al. 1988). The main reason for this are difficulties in establishing and maintaining the "knowledge base", a complex set of facts and rules that are prone to error. In addition, expert systems are not really different from traditional identification keys in that the user still has to answer a long series of more or less complicated questions. Thus, expert systems can be viewed as a very comfortable modern version of traditional identification keys.

Numerical methods will be of growing importance in the future because they lead to automatic identification. This will, however, remain a "high tech" approach based on expensive hardware combined with sophisticated image analysis routines. The main problem with numerical methods is the need to measure at least 50 individuals of each species involved in order to estimate the variance in shape. Thus, numerical

methods for identification will have their niche, but that will probably not be in the tropics, where fish and other taxa tend to be speciose, but where individual species may not be abundant.

Modern databases represent powerful, yet easy to use identification systems. Once appropriate forms have been designed, it is very easy to extend a system to other geographical areas. Also, as in the case of the two studies mentioned above, it is possible to extract the information needed from drawings and descriptions already in the literature. Moreover, this approach is not only useful for taxonomic identification, but also allows the development of powerful information systems such as FISHBASE, which will be described in the next issue of Fishbyte.

References

Abott, L.A., F.A. Bisby and D.J. Rogers. 1985. Taxonomic analysis in biology, Columbia University Press, New York. 336 p.

Achenbach, I. and R. Froese. 1990. Presentation of a database system for information on and diagnosis of fish diseases. ICES C.M. 1990/F: 72. 13 p.

Feigenbaum, E., P. Mccorduck, and P. Nii. 1988. The rise of the expert company. Times Books, New York.

Froese, R. 1989. Computer-aided approaches to identification. II. Numerical taxonomy. Fishbyte 7(3):25-28.

Froese, R. and C. Papasissi. 1990. The use of modern relational data bases for identification of fish larvae. J. Appl. Ichthyol. 6:37-45.

Froese, R., W. Schöfer, A. Röpke, U. Piatkowski and D. Schnack. 1989. Computer-aided approaches to identification. I. Expert Systems. Fishbyte 7(2):18-19.

Jones, E. 1988. Using dBASE III PLUS. Osborne MacGraw-Hill, Berkeley, California. 516 p.

Mertens, P., K. Allgeyer, H. Däs. 1986. Betriebliche Experten-systeme in deutschsprachigen Ländern - Versuch einer Bestandsaufnahme. Arbeitsberichte des Instituts für mathematische Maschinen und Datenverarbeitung 19/6, Univ. Erlangen-Nürnberg.