Computer-Aided Approaches
to Identification
I. Expert Systems

R. FROESE
W. SCHÖFER
A. RÖPKE
U. PIATKOWSKI
D. SCHNACK
Institut für Meereskunde
Düsternbrooker Weg 20
D-2300 Kiel 1
Federal Republic of Germany

Introduction

Identification of aquatic organisms is a difficult and time-consuming job. Our working group (Computer-aided Identification of Aquatic Organisms, CIAO) tried to find out how computers could provide solutions to this problem. Because some kind of expert knowledge seemed to be required, one idea was to build an expert system for identification. The result of about one year's work was IDEXSYS, our IDentification EXpert SYStem. Another idea was to use the possibilities of modern databases (i.e., query-by-forms and limited-choice fields) for identification. Finally we tried to use a numerical approach, i.e., perform several measurements on the organisms and let the computer identify the species. The results of these different methods will be presented in a series of three FISHBYTE articles, starting with the present one on IDEXSYS.

IDEXSYS, the questioner

Expert systems are the first viable products of artificial intelligence research. They are typically built around one question, e.g., "to which species does my fish larva belong", or "what disease do my fish suffer from", or "is my site suited for aquaculture", or "what is the best species (strain) for my aquaculture system", etc. Expert systems consist of (i) an inference engine and (ii) a knowledge base. The inference engine puts a question to the user, receives the answer(s), and decides which question to ask next, until it reaches an advice (e.g., "your larva is probably *Trachurus trachurus*") or runs out of questions. The knowledge base normally consists of a semantic net, i.e., a well-structured tree which contains the questions and final responses. Such a knowledge base is very similar to a printed identification key; thus, we implemented a key for fish larvae of the Northeast Atlantic recently developed at our institute (Halbeisen 1988).

With IDEXSYS, the user must answer a sequence of questions, as when working with a printed key. The main advantages of the system are:

besides "yes" and "no" the user can answer "probably yes" or "probably not" (see Fig. 1). If an identification fails (the final question is answered "no"), the program will return to the questions that were answered with "probably ...", thus giving the user a chance to select a different answer;

every question is supported by a picture illustrating the subjects of decision (see Fig. 2 and 3);

Fig. 1. Working screen of IDEXYS. The upper left window shows the groups considered, some of which are already excluded. The lower window shows the last questions, and the answers given. The upper right window shows the possible answers, which can be selected with the cursor keys or a mouse.
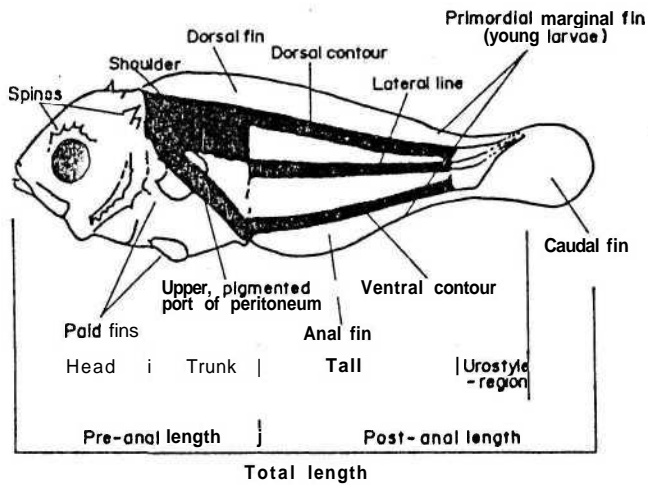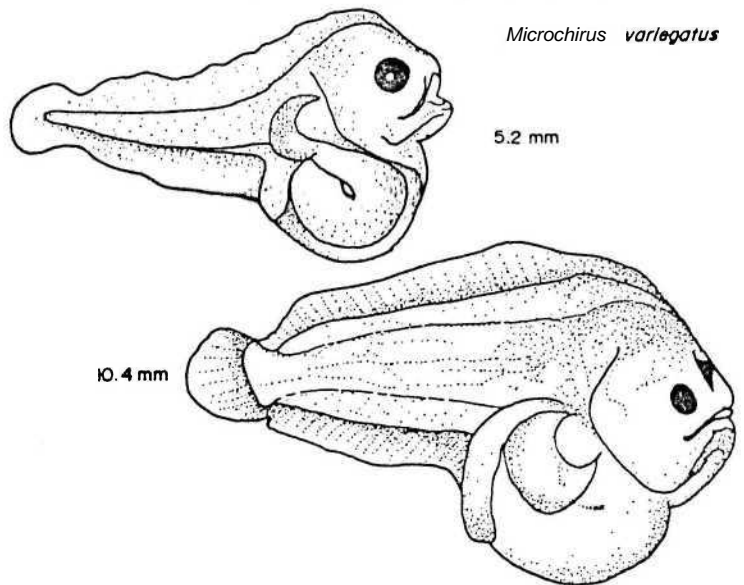
Nomenclature

Fig. 3. Example of *a* screen display of a larva *(Microchirus variegatus).*

- with every question, the group considered is displayed, and the user can see which characters characterize which group (see Fig. 1);
- if the user feels she/he is on the wrong track, she/he can backtrack through the key;
- experienced users can select any starting point within the key, e.g., "flatfish" or "eel-like larvae";
- the user can easily update the knowledge base;
- the user can create new keys for other objects; and
- the program can produce a formatted printout of the knowledge base which can be used like a conventional printed key.

Thus, although IDEXSYS offers nothing that could not be done with a printed key, it provides a very comfortable way of identifying organisms. Its main advantage is, probably, the ability to continuously update existing keys and to easily construct new ones. At the moment, we are working on a key for fish larvae of the Mediterranean in cooperation with the Instituto de Ciencias del Mar, Barcelona and the University of Athens, and on a key to identify fish diseases.

IDEXSYS runs on IBM compatibles with 640 K RAM. A HERCULES graphics card is required to display the pictures. IDEXSYS is available from the ICLARM Software Project in return for five 5 1/4" blank diskettes. For a more detailed description of IDEXSYS, see Froese and Schofer (1987).

References

Froese, R. and W. Schofer. 1987. Computer-aided identification of fish larvae. ICES CM. 1987/L:23.

Halbeisen, H.W. 1988. Bestimmungsschlüssel fur Fischlarven der Nordsee und angrenzender Gebiete. Berichte des Instituts fur Meereskunde an der Universität Kiel No. 178:76 p.
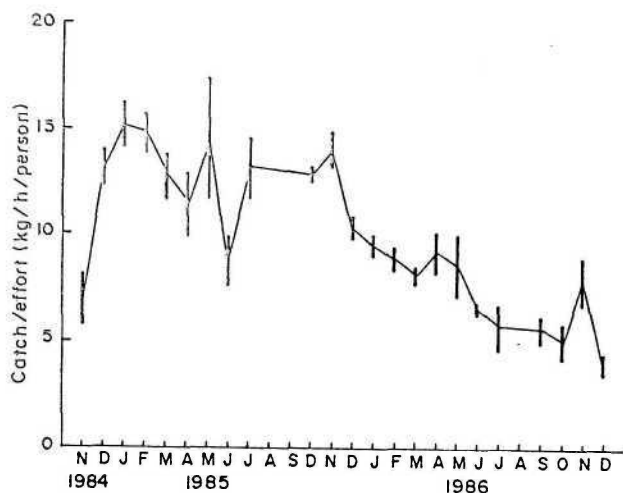
Fig. 5. Time series of catch/effort of yellow clam (± *si*. dev.) November 1984 and December 1986.

Uruguayan yellow clam fishery, the limited area along which the resource is distributed simplifies the identification and enforcement of low-cost management measures. This makes this fishery an attractive experimental unit to analyze the impact of different management schemes.

Without prejudice to the management issues discussed above, the increase in demand from the domestic and the potential of foreign markets make it necessary to initiate restocking experiments - either by "sowing" or colonization (*sensu* Castilla 1987) - in appropriate areas along the Uruguayan Atlantic coast. Indeed, spontaneous clam restocking

activities carried out by fishermen in Uruguayan beaches have been successful.

The present state of the resource suggests that there is presently no need to improve the harvesting technology (Defeo, in press). On the other hand, improving the hygienic and sanitary conditions of post-harvest processing would result in greater value added, which would directly benefit the artisanal fishing community.

Bearing the above considerations in mind, improvements in clam processing and marketing techniques should be promoted both for the present fishery and for newly stocked area within the context of a policy aimed at the development of organized artisanal fishery community centers.

References

Castilla, J.C. 1987. La problemática de la repoblacion de mariscos en Chile: diagnóstico, estrategias y ejemplos. Simposio "Presente y Futuro de la repoblacion de recursos marinos en Chile", Ancud, Chile, 26-29 de Octubre, 1987. mimeo.

Defeo, O. Consideraciones sobre la ordenacion de una pesqueria en pequena escala. Biología Pesquera (Chile). in press.

Defeo, O., C. Layerle and A. Masello. 1986. Spatial and temporal structure of the yellow clam *Mesodesma mactroides* (Deshayes 1854) in Uruguay. Medio Ambiente (Chile) 8(1): 48-57.

Hall, M. 1983. A spatial approach to the population dynamics of the Manila clam *(Tapes philippianarum)*. Ph.D. thesis, Univ. of Washington, U.S.A. 224 p.

Panayotou, T. 1983. Conceptos de la ordenacion para las pequerias en pequena escala: aspectos económicos y sociales. FAO, Documento Técnico de Pesca 228. 60 p.

Computer-Aided Approaches to Identification
II. Numerical Taxonomy

R. FROESE
Institut für Meereskunde
Universitat Kiel
D2300 Kiel 1
Federal Republic of Germany

Introduction

This is the second in a series of three Fishbyte articles dealing with computer-based methods to facilitate identification of aquatic organisms. In the preceding article (Froese et al. 1989), IDEXSYS, an IDentification EXpert SYStem for fish larvae of the Northeast Atlantic. IDEXSYS was presented which relies on a text-based approach and which, apart

from the fact that it is very comfortable to use and easy to update, functions as would a printed key.

Morphometric measurements in combination with discriminant analysis have been successfully used to identify and separate stocks (Ihssen et al. 1981; Meng and Stocker 1983; Misra and Ni 1983; Misra 1985; Maccrimmon and Claytor 1986; Reddin 1986). This led to the idea of using morphometric measurements and discriminant analysis to build a numerical identification key.

Materials and Methods

The study is based mainly on larvae sampled with R/V Poseidon in the Celtic Sea in April 1986 (Röpke 1988). An interactive image-analysis system was used for quickly and accurately measuring the larvae, of which 781 were included in this study. Two video images of a larva were captured, one in lateral and one in dorsal view and the following parameters were measured: (in dorsal view) standard length, prepectoral length, width between pectorals, width at anus; (in lateral view) preanal length, preorbital length, diameter of eye, depth above eye, depth above pectorals, depth of tail behind anus (Fig. 1). About 50 individuals of each species listed in Table 1, covering the size range from early to late postlarvae (i.e., excluding yolk sac larvae and larvae in the process of metamorphosis) were measured. To approximate multivariate normality and linear relationships, all measurements were transformed to (base 10) logarithms (Bliss 1967; Pimentel 1979). Because what is important are differences in body shape rather than the actual size of body parts, all measurements were corrected for length. This is often done by expressing the measurements as ratios of body length, but according to Pimentel (1979), ratios have unusual distributions and are subject to various errors. Therefore, analysis of covariance was used, as suggested by Ihssen et al. (1981), Misra and Ni (1983) and others, which adjusts each of the morphometric characters to the overall mean standard length according to the formula:

$$AM = OM - (RC * (SL - MSL))$$

where AM is the measurement adjusted for the covariate, OM is the original measurement, RC is the overall regression coefficient between character and standard length, SL is the correlated standard length, and MSL is the overall mean standard length.

Quadratic discriminant analysis is particularly suited to handle different group sizes and different within-group covariances. The method has not yet been used for identification purposes, possibly because of the restriction, emphasized by some authors, that the number of groups should not exceed the number of parameters used. This might be viewed as a serious obstacle, since there are about 120 fish species in the North Sea and adjacent waters, but no more than about ten measurable parameters that can be used for identification of fish larvae. This problem was overcome by a two-step approach, in which similar species were assigned to larger groups. An unknown larva is first classified into one of these groups and then into a species within the group.

The SAS statistics software was used to process the data (Anon. 1985). A cluster analysis (method = WARD) was performed on the arithmetic means of the adjusted measurements for each species to
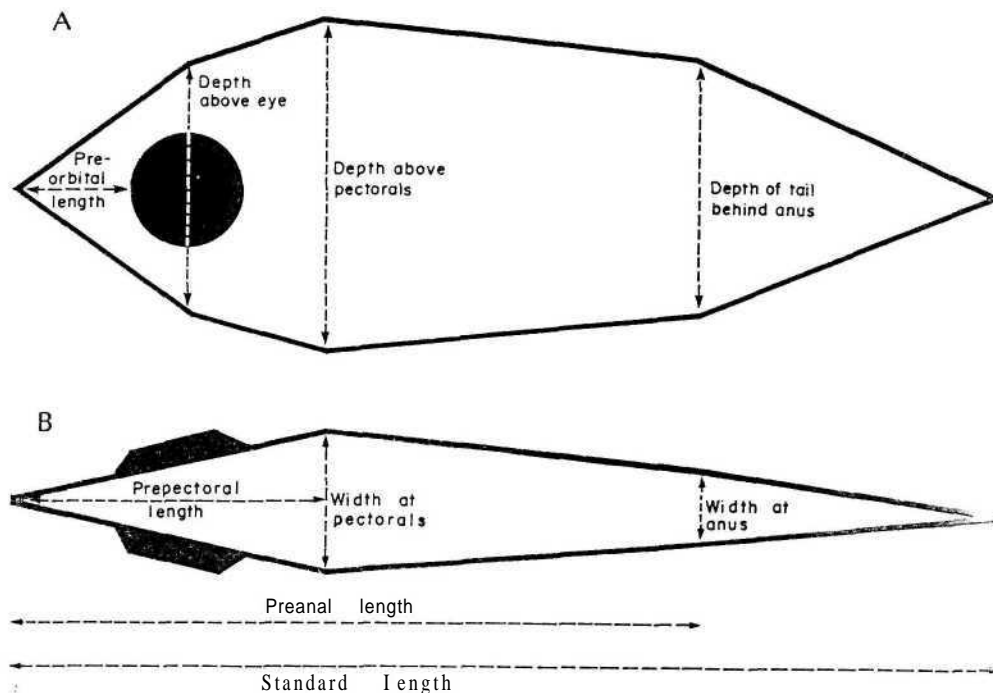

Fig. 1. Measurements performed on the fish larvae (A) lateral and (B) dorsal view.

Table 1. Results of cluster and discriminant analyses (n = number of specimens measured; min SL = minimum standard length; max SL = maximum standard length; Clus. = cluster number; Class. = percentages of observations classified correctly into a) clusters and b) species in the clusters; Prob. = probability of correct identification).

Cluster No.		Species Name	n	min SL mm	max SL mm	Clus. no.	Class. a %	Class. b %	Prob. %
1	1	*Gadiculus argenteus than*	52	2.48	7.33	1	83	89	74
	2	*Merlangius merlangus*	52	2.35	8.01	1	83	71	59
	3	*Merluccius merluccius*	45	3.08	7.20	1	83	91	76
	4	*Micromesistius potassiou*	30	3.47	9.63	1	83	90	75
	5	*Pollachius pollachius*	51	3.12	11.73	1	83	88	73
	6	*Trisopterus* sp.	51	4.52	11.88	1	83	90	75
2	7	*Lepidorhombus boscii*	42	3.47	8.04	2	96	76	73
	8	*Lepidorhombus whiffiagonis*	44	3.55	12.66	2	96	91	87
	9	*Microchirus variegatus*	32	2.27	7.02	2	%	100	96
	10	Triglidae	37	4.15	15.16	2	96	100	96
3	11	*Benthosema glaciale*	52	4.01	8.07	3	81	100	81
	12	*Callinymus* sp.	84	1.94	9.94	3	81	100	81
	13	*Molva molva*	38	3.16	6.68	3	81	100	81
	14	*Scomber scombrus*	45	2.92	9.16	3	81	100	81
4	15	*Argentina sphyrena*	48	4.87	18.47	4	98	98	96
	16	*Clupea harengus*	50	6.80	19.45	4	98	100	98
5	17	*Glyptocephalus cynoglossus*	28	3.31	12.41	5	96	100	96
		Total	781	1.94	19.45	5	87	93	81

generate clusters of morphometrically similar species. Quadratic discriminant analysis with within-group covariance matrices was used to develop the equations to classify an unknown larva into one of the clusters in a first step, and into one of the species within the cluster in a second step. All 781 measured larvae were used to test the system, i.e., every larva was treated as 'unknown' and classified by the system into a cluster (= group) and a species, resulting in a percentage of correct classification for every cluster and species.

Results

The results of the cluster and discriminant analysis are summarized in Table 1. The result of the cluster analysis is convincing: all gadiforms except for *Molva molva* were grouped into one cluster (#1). Three flatfish and the larvae of the Triglidae family (which resemble the flatfish in lateral view and which are here treated as one "species") were assigned to cluster #2. Cluster #3 contains four roundfish larvae not related to each other. Cluster #4 consists of eel-like larvae with long guts, and cluster #5 contains only one flatfish species (white sole).

The first step of the test put each larva in one of the five clusters. The percentages of correct classifications range from 81% (i.e., 19% belonged to one of the other clusters) for the most heterogeneous cluster (#3) to 98% for the "eel-like" cluster (#2). The percentages of correct classifications in the second step, in which species are identified within each cluster, ranged from 71% for *Merlangius merlangus* to 100% for 7 other species, with an average of 93%. Overall, the probability of correct identification was 81% (the random probability for 17 species would be about 6%).

Discussion

In the light of these results, our two-step discriminant analysis of morphometric measurements seems to be a promising identification tool. The method uses twice the full power of discriminant analysis: while in the first step it might be, e.g., the relation between preanal length and body depth that produces a clear separation between flatfish and eel-like larvae, other parameters might lead in the second step to the segregation of the very similar species within the 'gadiform' cluster. Of course, the user must verify the program's suggestions using traditional methods, e.g., by comparing the larva with a picture and check the most significant characters. The purpose of the system thus would be to guide the

user very fast to the most probable species without the need to answer a long list of questions.

Although the method looks promising, it must be emphasized that it requires at least 50 well preserved specimens, covering the whole size range of each species for estimating the discriminant function. Even in the well-explored Northeast Atlantic it was impossible to get enough larvae for more than about 30 species.

This led us to the idea of using descriptions and images from the literature together with a modern database for identification. This will be described in third and last contribution to this series, to be published in the next issue of Fishbyte.

References

Anon. 1985. SAS user's guide: Statistics, Version 5 edition. SAS Institute Inc., Cary, NC. 1985. 956 p.

Bliss, C.I. 1967. Statistics in biology, Vol. 1. MacGraw-Hill Company, New York, NY. 558 p.

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

Ihssen, P.E., D.O. Evans, W.J. Christie, J.A. Reckahzn and R.L. Desjardine 1981. Life history, morphology and electrophoretic characteristics of five allopatric stocks of the lake whitefish (Coregonus clupeaformis) in the Great Lakes region. Can. J. Fish. Aquat. Sci. 38: 1790-1807.

Maccrimmon, H.R. and R.R. Claytor. 1986. Possible use of taxonomic characters to identify Newfoundland and Scottish stocks of Atlantic salmon, Salmo salar L. Aquaculture and Fisheries Management 17:1-7.

Meng, H.J. and M. Stocker 1984. An evaluation of morphometrics and meristics for stock separation of pacific herring (Clupea harenguspallasi). Can. J. Fish. Aquat. Sci. 41: 414-422.

Misra, R.K. 1985. Quadratic discriminant analysis with covariance for stock delineation and population differentiation: a study of beaked redfishes (Sebastes mentella and S. fasciatus) by a recent developed QDF. Can. J. Fish. Aquat. Sci. 42: 1672-1676.

Misra, R.K. and I-H. Ni. 1983. Distinguishing beaked redfishes (deepwater redfish, Sebastes mentella and Labrador redfish, S. fasciatus) by discriminant analysis (with covariance) and multivariate analysis of covariance. Can. J. Fish. Aquat. Sci. 40: 1507-1511.

Pimentel, R.A. 1979. Morphometrics. Kendall/Hunt Publishing Company, Dubuque, IA. 276 p.

Reddin, D.G. 1986. Discrimination between Atlantic salmon (Salmo salar L.) of North American and European origin. J. Cons. int. Explor. Mer 43: 50-58.

Ropke, A. 1988. Small-scale vertical distribution of ichthyoplankton in the Celtic Sea in April 1986. Meeresforsch. 32(3): 192-203.

A Theory of Fishing for a Two-Dimensional World

DANIEL PAULY
*International Center for Living
Aquatic Resources Management (ICLARM)
MC P.O. Box 1501
Makati, Metro Manila
Philippines*

Introduction

In 1884, Edwin A. Abbott, a schoolmaster with a passion for theology and literature, published - pseudonymously - a little book titled "Flatland: a romance of many dimensions" in which he explored some of the implications of living in a two-dimensional world.

He described a world in which there is Left and Right, and Back and Forth, but no Up and Down, and dealt with issues such as the climate and houses, the inhabitants (especially the women, who, as opposed to the round males, were pointed and hence, had to be treated with great respect), the problems of color recognition and other issues illustrating the differences between Flatland and a three-dimensional world such as ours.

Mainly, however, he dealt with moral and theological issues - this was the thing to do in the Victorian era. So, the emphasis of "Flatland" was devoted to the conflicts between the local clergy (who were "Administrators of all business, art and science"), and those Flatlanders, philosophers and mystics, who were spreading seditious notions, such as "third dimension", "cube" or "upward".

A.K. Dewdney published in 1984 "The Planiverse: computer contacts with a two-dimensional world", in which the idea of Flatland was carried further

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 fur Meereskunde
Universitat 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 modem 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;

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
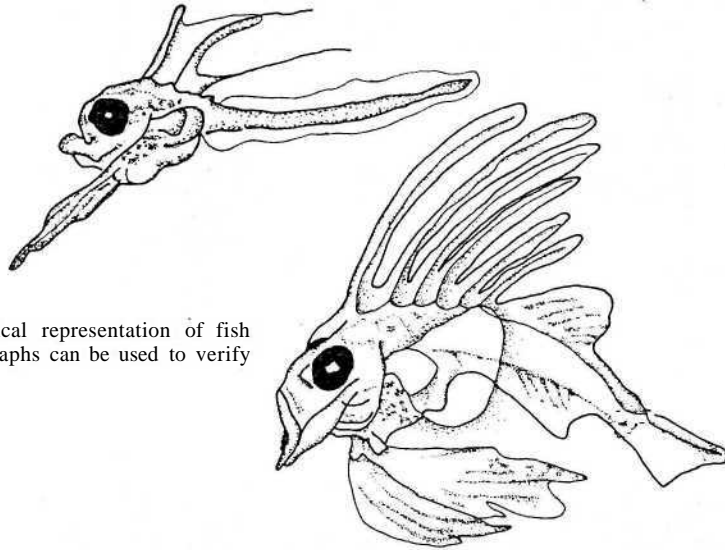
Fig. 2. Some examples of the graphical representation of fish larvae included in FISHBASE. Such graphs can be used to verify an identification.

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 databases 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 modem 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 CM. 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. Ropke, 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. Das. 1986. Bctriebliche Expcrten-systeme in deutschsprachigen Ländern - Versuch einer Bestandsaufnahme. Arbeitsberichte des Instituts fur mathematische Maschinen und Datenverarbeitung 19/6, Univ. Erlangen-Nürnberg.