FRUIT FLY EXPERT IDENTIFICATION SYSTEM

Introduction

By F. Christian Thompson

Information is stored and retrieved by names. Scientific information is stored with scientific names. To obtain the scientific name of an organism, one identifies the organism. Identifications are made by matching characteristics of unknowns with knowns. Traditionally taxonomic keys have been used for this matching. The earliest keys were just text, with relatively few characters. Keys have been improved over the years by adding more characters, as well as illustrations. But the main problem with keys is inflexibility. There is a set pathway (often a long one) through the key to each species; a single mistake may lead to an erroneous identification; a single missing character may leave the user at a dead end. Verification of the identification requires reading complete descriptions to find all the characters to check.

An expert system is much more flexible. Many taxa can be eliminated immediately by restricting the data set according to geographic location or host data. Any character of any sex or stage in the life cycle can be chosen in any order which seems best to the user. Or the expert system can select the best characters for use, based on their ability to separate the remaining taxa under consideration.

Characters are accompanied by illustrations, and multiple states are allowed. This speeds up the identification process in two ways: by enabling direct comparison of images with the specimen (rather than reading text), and by reducing the total number of decisions which must be made, because more than the traditional two possibilities can be efficiently evaluated at one time.

Characters are also accompanied by help files which can be accessed at any time. Even so, the possibility of error (e.g., a poor or aberrant specimen) can be accommodated by having the expert system tolerate an error or two before rejecting a taxon. Errors, once detected, can be corrected easily, without stepping through all characters again.

The verification process is also much easier. Although complete descriptions are available, just as in traditional taxonomic references, the expert system can also give the differences between the specimen and another taxon, or between any two taxa. Or the expert system can list all the diagnostic characters for a particular taxon. With the identification process complete, the database can be queried for complete nomenclatural and distributional data, as well as pertinent references.

Expert systems are not a panacea. Unusual specimens, those outside the domain of the expert system or with distorted features will still have to be sent to the systematist.

Fruit Fly Expert System

Our Expert System is designed to be totally self-contained. All one needs to do is to run the program and then follow the instructions on the screens. All questions should be answered by the included help files. A tutorial is also included. If you want a quick start, just run the tutorial.

To run the Expert System one needs an MS-DOS (PC) computer with a VGA monitor and sufficient memory. See below for details on the memory requirements. The Expert System will also run on MacIntosh computers with MS-DOS or Windows emulation software, such as SoftPC from Insignia. The Expert System may be run directly from the CD-ROM or the files may be copied onto the hard disk and run.

TO RUN the program one only needs to set a MS-DOS variable (SET PANKEY=D:/FRUITFLY, where D: is the letter for the CD-ROM drive or the disk drive where the fruitfly files are) and go to the FRUITFLY directory and enter ONLIN7 at the DOS prompt. **To run the tutorial**, one only needs to switch to the appropriate directory where the files are and enter the name of the program and data file at the DOS prompt (D:\demo1\rdemo2t learn1, where D: is the letter for the CD-ROM drive or the disk drive where the fruitfly files are). These tasks are most easily done with a MS-DOS batch file. Samples of such batch files are in the directory/folder called BAT on the CD-ROM.

The Expert System has **three components**: The program, data sets and images. Users need to be aware of these different components and how our design was shaped by them. We have assumed that our users are professional identifiers, such as those working for APHIS-PPQ. Hence, they are already familiar with traditional identification aids, such as keys, and are familiar with their organisms.

The **program** presents character data to users who then make selections which ultimately may lead to an identification. The program differs from traditional identification tools by allowing for random and varied access to character data. The user is free to choose any of the available characters in any order, whereas the traditional key allows only for the use of specific characters in a rigid sequence. Users can also request comparisons between taxa, descriptions and/or diagnoses of taxa, functions not available in traditional keys. So, our objectives in designing the program were to maximize the access to character data and to present those data in the most effective manner. Naturally, our objectives were constrained by the data format used and computer resources available. To build our Expert System, we worked with Richard Pankhurst, the world's authority on computerized biologi-

cal identification. The basic program, known as ONLINE, was his work to which he added some significant new features at our direction. So, when you are using the program, reading the menus and general help screens, you are using ONLINE.

Data sets are what determine the identification capabilities of the expert system. The adage "garbage in, garbage out" is true of the expert system. These data sets are the wisdom of the experts, so the program can only be as effective as the experts were in expressing their wisdom in a set of characters and values for taxa. For data sets our objective was to use a data format which the systematics community endorses and widely uses, so there would be the maximal number of data sets available that could be compiled and used by our expert system. We also wanted a data format which could encode all kinds of character data and was not proprietary, so data sets could be shared. The DELTA data format, which was established by CSIRO was the only available one which matched our criteria. The DELTA data format imposed some limitations on the Expert System, but these are less than the advantages gained. Also, our data sets can be used with other computer identification systems, such as INTKEY.

Two **fruit fly data sets** were developed. The adult data set by the leading Tephritidae experts: Amnon Freidberg, Tel Aviv University, Israel; Ian White, CAB Institute of Entomology, London; and Allen Norrbom, Systematic Entomology Laboratory, Washington. Lynn Carroll worked closely with these specialists, adding her experience and knowledge of DELTA to ensure a uniform and consistent data set. She developed the larval data set. So, when one reads the text of the characters and related help screen, one is using the data set provided by these experts. And when one gets an identification it is because these experts selected the best characters.

Images help users understand character data. They are, therefore, a useful if not necessary adjunct to the data set. However, images are not required by the program. The program was designed so that images were independent of the data set because images are expensive, the most expensive component beyond the data set. Also linking images to the data set and using such technologies as touch screens or mice to select images would have been more costly as each data set would have required special coding. To keep costs within budget, existing images were re-used wherever possible and only the minimal number of new ones were created. However, many images were improved, for instance, black and white habitus figures were colored. So, don't be surprised if these images look familiar!

Authorship: The adult data sets are by Carroll, White, Freidberg and Norrbom; the larva data set by Carroll; the ONLINE application is by Richard Pankhurst; the tutorial was done by Jennifer Fairman; and I did all the little things necessary to tie it all together. So, for example, to cite the larval data set, the following is recommended:

Carroll, L. E.
1998 Larval Character Data Matrix. 76 characters for 81 taxa in DELTA format. *In* Thompson, F. C. (ed.), Fruit Fly Expert System and Biosystematic Database. *Diptera Data Dissemination Disk* 1.

Memory limitations

Depending on the memory resources of one's computer, different versions of the fruit fly identification data may have to be used. These data sets differ only in the number of species treated. The full data set provides information on 197 species, the small data set on only 84 species. The small data set contains the most important pest species and a few other ones for diversity. Otherwise, they are the same.

To use the full data set, your computer needs at least 580KB of conventional DOS memory and 1 MB of expanded memory. If your computer has at least 580KB of conventional DOS memory, then the small data set should be used. If a data set fails to load or the program runs erratically, then the amount of memory available should be checked and increased.

Memory can be increased, but how depends on the computer's resources. Memory may be devoted to various other programs and/or devices. If so, by merely changing the configuration and start-up files enough DOS memory may be released to load at least the small data set. Computers with a 386 or better microprocessor probably already have the extra memory needed (especially if WINDOWS is being used) and how the computer uses that memory needs to be changed. The memory is probably set as EXTENDED instead of EXPANDED. For computers with 286 microprocessor or a classic 8088/8086 microprocessor, plug-in cards with expanded memory can be purchased. A handy reference on PC memory is Goodman (1993, Memory Management for all of us. Sams Publs). The Deluxe Edition ($39.95) includes software, such as diagnostic utilities and memory managers.