5

A General Structure for Biological Databases

J. Diederich[1], R. Fortuner[2] and J. Milton[1]

[1]Department of Mathematics, University of California, Davis, CA 95616, USA; [2]4 rue des Jardins, 17130 Montendre, France (Correspondant du Muséum National d'Histoire Naturelle, Paris) Fax: +33 546 703659/E-mail: fortuner@wanadoo.fr

Introduction

The Project NEMISYS (Nematode Identification System) was launched in 1987 to create an identification system that would not be restricted to a particular identification method but would have the potential to use and support all identification approaches. The need for a multi-method approach to identification was argued by Fortuner (1989, 1993). A tool-based expert workstation which can meet this requirement was described by Diederich and Milton (1989, 1993a), and a NEMISYS prototype was developed using several identification tools as examples (Diederich and Milton, 1993b). NEMISYS later became GENISYS (General Identification System) when it became apparent that the concepts we developed applied outside nematology.

One of the most critical lessons learned in NEMISYS was the importance of laying a solid foundation for data and knowledge representation. Methods employed by others in developing databases for identification systems seemed inadequate in addressing the difficulties we experienced with nematological data. In recent years, the construction of a database has subsumed the entire project and given it a different orientation.

From the very beginning, the standard data decomposition of entity/property/value used by database designers (and often used for biological databases, e.g. Lebbe, 1991; Lebbe and Vignes, Chapter 4, this volume) was used for the definition of a list of characters. However, biological characters do not easily and comfortably fit into this form. Early attempts at defining a list of characters revealed that this alone would not solve the problem of a lack of homogeneity in the data. It was necessary to define new concepts such as basic properties,

 Information Technology, Plant Pathology and Biodiversity (eds P. Bridge, P. Jeffries, D.R. Morse and P.R. Scott)

name extensions, general/specific states, and state-based relationships (Diederich, 1997) and to propose guidelines to ensure that these concepts would be used correctly (Diederich *et al.*, 1997) in creating a list of characters that could be consistently and uniformly expressed. These concepts are not only important for creating morpho-anatomical databases in nematology, but they can also play a central role in other areas.

Here, we assume that it is accepted that it is better to have access to several different approaches to identification than to be restricted to a single one, and that it is better to have a database that can be used for multiple purposes. Ideally, the data in any database is used through a database management system (DBMS). However, there are limitations on what can be done in existing commercial DBMSs to support biological data and their applications. This is not surprising because DBMSs are typically business oriented and generally ill-suited to the complexity of biological data. What is needed is a Bio-DBMS; a DBMS that is created specifically for biological data and applications. In fact, part of our project will ultimately involve the definition of the architecture of a Bio-DBMS, and what we present here are some preliminary requirements for it. Naturally, a Bio-DBMS would be built on top of a commercial DBMS since it would not be practical to create a Bio-DBMS from scratch.

In this context we will: (i) describe some of the problems with traditional characters and demonstrate the need to develop and enforce a general discipline in building biological databases; (ii) give some examples of information, other than simple data, that will be needed in any application; (iii) give examples of non-morphological data that can be captured using our methodology; and (iv) discuss some key requirements for a Bio-DBMS.

Definitions

Here, the word 'character' will be used in its widest sense to mean either an abstract concept that can be used to describe and differentiate a species or other taxa, or a characteristic in an actual specimen and that can be used to identify this specimen. A character will be called 'traditional' if it is in the form found in traditional keys, e.g. organ-X cylindrical, or if it is composed of a characteristic that can be used to differentiate two taxa, e.g. shape of organ-X, and a set of values taken by this characteristic in the taxa or specimens considered, e.g. cylindrical, ovoid, spherical, etc. In many computer identification applications, characters are: (i) given in their traditional form; (ii) numerically coded and stored in a data matrix; or (iii) coded using the DELTA code and stored in text files (Dallwitz, 1980; Webster, 1988).

Problems with Traditional Characters

Selection of a subset of characters

In all the cases we have observed so far, identification aids use a subset of characters selected by an 'expert' (in the knowledge engineering sense of the term) out of all possible characters as being 'good identification criteria'. The states taken by these selected criteria in different taxa are also defined by the expert. We will not discuss the methods of character and state selection but emphasize the fact that there is selection. On the one hand, this character selection and pre-processing by the expert is compulsory for the creation of the matrix used by taxonomic programs or the text files used by DELTA applications (Dallwitz *et al.*, Chapter 19, this volume). On the other hand, it eliminates from the application database all the characters not selected by the expert and makes them unavailable to: (i) another expert who does not agree with the first one about the list of 'good identification criteria'; (ii) an expert using a different identification approach; and (iii) an expert who wants to use the unselected characters outside of identification, e.g. for taxonomic studies. This results in much duplication of effort, as each expert has to build a database from scratch.

Use of complex characters

Adding to this basic difficulty is the fact that many traditional characters, and many computer-coded characters, are complex entities that group several simpler concepts. For example, the character 'disk with a central zone densely covered with flat, heart-shaped, epidermal denticles' found in a database on lampreys (Cruette, Paris, 1991, personal communication) indicates that a central zone is present, that it is densely covered with epidermal denticles, and that these denticles have a particular outline (heart-shaped) and cross-section (flat). This type of character is well suited to traditional or computer keys (either an unknown has such a disk or not), but it is not suited to other approaches. For example, it would be difficult to compute a coefficient of similarity between a species with the character above and one with a central zone sparsely covered with flat, heart-shaped, epidermal denticles.

Classes for quantitative characters

A final problem is that many coding methods transform quantitative values into classes before storage. Often, lengths are recorded, not as the actual value in a specimen, a population or a taxon, but as a class value. This again is well suited to keys (if a taxon belongs to the same class as the unknown it will be retained; it will be eliminated if it does not), but not to other approaches, such as computation of the NEMAID similarity coefficient (Fortuner and Wong, 1985), which uses actual values.

The NEMISYS approach

Description

The solution adopted for NEMISYS and now GENISYS was to store as many characters as possible, in particular all the characters found in published descriptions of taxa. Actual numerical values and qualitative states of the characters, as they were recorded in the literature, were also stored. Complex characters were decomposed into elementary characters using the entity/property/value decomposition. Any selection of characters for particular uses or applications would be done *a posteriori*, from this general pool of characters.

It was quickly discovered that this was not enough to ensure a homogeneous representation of the characters. This problem was partly solved by enforcing strict guidelines for the decomposition of characters. Using morphological characters only, we defined an entity as the parts composing an organism, starting with the organism down to organs, tissues, cells, etc., all the way down to molecules. Doing this, we found that most of the properties used in the original nematode list of characters come from a very short list of basic properties which are given in Table 5.1 (Diederich, 1997). To these properties are attached the traditional states or values, including synonyms, general states, and fuzzy states for measurements and quantities.

Typical examples of the main types of characters (i.e. real numbers, integers, and qualitative states) with actual values from the description of a nematode (*Helicotylenchus dihystera*) are:

Body	length	725 µm
Lateral field lines	number	4
Tail	shape	dorsally curved

Note that these are individual values, but the database includes fields for range, mean and standard deviation for an individual population and for a composite

Table 5.1. Basic properties for morpho-anatomical data. The list of properties is broken down into four groups relating to: appearance, location, dimensions and quantity of the morpho-anatomical structure described by the properties.

Appearance	Dimension	Placement/Location	Quantity
posture	length	position relative to*	presence
shape	height	distance to*	quantity
kind	width	orientation	number
texture	diameter	angle	
arrangement	depth		
symmetry	ratio of*		
	size		

*Relational properties.

description of the corresponding species. The complex lamprey character given above would be decomposed into three parts linked together in a hierarchy of parts (disk has a central zone; central zone has denticles), and each part would be described by basic properties and the appropriate states. For example, the denticles would have the following properties and states:

Denticles	presence	present
	arrangement	dense
	width	flat
	shape	heart

The guiding principle here is that it is easier to construct complex entities from simple ones rather than the other way around. In addition, there is a greater chance to record all the important information.

Number of characters

Typically, systematic databases include a relatively small set of characters selected for a particular purpose, e.g. 30 or 50 characters selected for identification (i.e. characters which are easy to observe, clearly differentiate existing species, not variable, etc.). Thiele (1993) used 108 characters for his analysis of *Banksia*. Presumably, these characters were selected because of their value in representing systematic relationships. Compared to these very small numbers, our current nematode (order Tylenchida) list of characters includes 272 biological structures described by over 1000 characters (1 character = 1 structure, 1 property). The potential for growth is staggering as these 272 structures could be described by 20 properties each, which represents 5440 potential characters, and this number would be far greater if states were included in the count. Instead of including only a subset of characters selected for a particular purpose, a GENISYS database aims at storing every possible character to serve as a general pool of characters available for any purpose.

Requirements

Using GENISYS characters in DELTA applications

The data stored in the form described can be used by applications specifically created for it, such as the tools defined within the NEMISYS and GENISYS projects (Diederich and Milton, 1993b). However, storing data for use by *ad hoc* applications only would be an insufficient justification for such an effort. A critical factor in the general acceptance of a new method or approach in data management is its compatibility with existing databases and applications. With morphological data, particularly in botany, much of the data has been recorded in DELTA databases (Dallwitz *et al.*, Chapter 19, this volume), and it is essential that GENISYS-style data can be used with existing DELTA applications.

DELTA applications such as ONLINE (Pankhurst, 1991 and Chapter 26, this volume), INTKEY (Dallwitz, 1993), etc., are generic programs that can be used with any data, provided they are presented as DELTA codes in specially constructed files. Obviously, these applications cannot use GENISYS data directly. Fortunately, characters stored in the GENISYS format can be converted to the complex characters used in DELTA applications. If an expert wants to use the data stored in a GENISYS database with a DELTA identification program, a mapping from characters in the former to those in the latter is required. This can be done by creating the three text files required by DELTA applications (Webster, 1988), i.e. the CHARS file, a list of characters and character states selected by the expert, the ITEMS file, with the species specific values, and the character specifications file (SPECS). The CHARS file will need to be created in the usual way. For example, starting with the GENISYS lamprey data above, the expert will decide that, out of, say, 20 characters considered to be suitable for identification of lampreys, character number 9 refers to the disk, its central zone and denticles with the following codes, corresponding to states seen in known taxa:

#9. Disk/
1. with a central zone densely covered with flat, heart-shaped, epidermal denticles/
2. with a central zone sparsely covered with flat, heart-shaped, epidermal denticles/
3. with a central zone densely covered with thick, barrel-shaped, epidermal denticles/

Then, if a particular species is recorded in a GENISYS database with Denticles / presence = present; arrangement = dense; width = flat; and shape = heart, it will be presented to a DELTA application in the guise of an entry in the ITEMS file such as '... 9,1 ...'

It should be kept in mind that this is only an intermediate solution, as it is not the seamless mediation between database and application that could be provided by a Bio-DBMS. This mediation between the simple characters of GENISYS and the complex characters used in applications in the DELTA format is just one aspect of the set of requirements for large multipurpose biological databases. Here we are simply stating 'what' needs to be done, not 'how' it should be accomplished within a Bio-DBMS. An extension of the concept of a 'view', an existing mechanism within a commercial DBMS, may provide one way to do it. A view can be loosely defined as a virtual table, defined on top of a database, using the DBMS language. For the mapping of characters, using views of the simple characters would be a reasonable approach. Note that commercial DBMS have limited view mechanisms when applied to schemas, i.e. to the table definitions, so an extension of the view concept would be required in a Bio-DBMS.

Examples from systematics

The creation of a general database will be a major effort in time and money, and it needs to be used by as many people and in as many ways as possible. In particular, a general morphological database, including all the known characters of the included taxa, could be used as a source of systematic, as well as identification, characters. In fact, these two categories of characters differ in the criteria applied to select them for a particular application, but the characters would all come from the same general pool. Obviously, a GENISYS database with all existing characters can be used as such a pool.

A view mechanism, similar to the one evoked for DELTA applications, could be used to feed the proper data into other taxonomic applications. In cladistics, PAUP is one of the most used programs. As an example, here are a few characters from Thiele (1993), coded for the PAUP program:

1. Habit: erect (0); repent (1)
2. Lignotuber: absent (0); present (1)

...

50. Style color: red, yellow, golden or purplish-black (0); always yellow (1)
52. Pollen-presenter shape fusiform (0); acicular (1); linear (2); ovoid (32); conical (4); awl-shaped (5)

...

95. Adult leaf blade length (with a note saying that morphometric data is recorded as sample size, log10-transformed mean, and coded state).

The first number corresponds to a column number in a data matrix. The numbers between parentheses are the coded states entered in the rows of the matrix representing the taxa studied. In a GENISYS database, the same characters would be present as follows, where a '–' separates distinct states in the list:

Whole organism	posture	erect–repent
Lignotuber	presence	present–absent
Style	color	red–yellow–golden–purplish-black
Pollen-presenter	shape	fusiform–acicular–linear–ovoid–conical–awl-shaped
Leaf-Blade (adult)	length	actual value (sample size, mean)

Here, a 'PAUP View' would present GENISYS data in the guise of a data matrix, which could be used by any other program that uses a similar matrix of characters. The view would have to take into account the fact that Thiele lumped into a single state (50:0) what would be recorded as four separate states in the GENISYS database.

Of course, Thiele's data could have come straight from a DELTA database, if one had been created for identification within the same taxon. However, it is very likely that the set of characters selected for any identification database would have been different from the set of cladistic characters chosen by Thiele.

(Again, this is because the criteria for character selection are not the same in cladistics and identification.) Even a slight difference would have defeated the transfer of the data from one database to another. For example, if an existing DELTA identification database has:

#50. Style color/
1. red/
2. purplish-black/
3. yellow, golden or always yellow/

Thiele would have been obliged to split the DELTA character state (3) between the two states of his character 50. There is no way he could have done that from the DELTA database and he would have had to go back to the original data. Note that we are not saying that DELTA forces the user to use these particular characters. Quite the opposite, the point is that DELTA leaves the user free to use any character. Freedom is fine, but then you are stuck with something that might be very difficult to use in a different application. In the GENISYS approach, each colour would be stored as a separate state, with yellow as a general state. A general state is a global expression, such as yellow, which may be divided into more specific states, such as golden (Diederich, 1996).

More generally, any coded data must be specifically defined by an expert who chooses the characters and defines the coded values, whereas our system records all the characters found in the descriptions, each in its simplest possible form and without any coding. Any selection is done later by selecting particular characters and states out of the complete pool of data.

Other Bio-DBMS requirements

The complexity of support needed within a Bio-DBMS can be appreciated in terms of the various relationships that must be defined for biological data. This support goes well beyond the task of storing the data and mapping between characters for the different purposes discussed above and it should facilitate efficient and intelligent use of the data. By efficient we mean that the Bio-DBMS should allow applications to use the data with ease, thus minimizing effort. By intelligent use we mean that the data should be properly retrieved and manipulated for the intended purpose. This does not carry the connotation that a Bio-DBMS will provide expert system capabilities, but will retrieve the intended data relative to the context of the request. Relationships play a key role in meeting this objective.

State-based character relationships

Using a view mechanism to define complex characters from simple characters is one requirement for a Bio-DBMS. Biologists also need a way to define relationships between data or between taxa. We briefly mention these as they have been presented and discussed elsewhere (Diederich and Milton, 1993b;

Pankhurst, 1993; Diederich, 1996). An example of relationships is given by dependent and summary characters. A limited concept of a dependent character is given by Pankhurst (1993): 'if character 1 (stem presence) is equal to state 1 (absent) then characters 2 through 4, which are various properties of the stem, are impossible.' We use a wider concept (Diederich and Milton, 1993b) that encompasses properties other than presence of an organ: a dependent character is applicable only if another character has a particular state. For example, the property diameter can be used only if the organ to which it refers has 'shape equal round'. Another concept is that of a summary character. Each state of a summary character implies that a number of other characters have particular states. For example, if the reproductive system of a nematode is described as amphidelphic, it implies that the number of genital branches is 2 and that each branch is directed toward opposite extremities of the worm. State-based relationships need to be defined for improved operation of identification applications, but they are not taxon dependent.

Taxon-dependent character relationships

Homologies and convergences are relations that exist between data for particular taxa. For example, the coiling of the lumen of the anterior part of the oesophagus that occurs in unrelated nematode families (Criconematidae, Belonolaimidae, Dolichodoridae) is a convergence. This fact will need to be recorded as a taxon-dependent relationship at the family level.

Given the dependency on taxa, these relationships are a form of data to be represented in the database, in addition to pure relationships built upon the list of characters. Even then, additional support for homologies and convergence is required. Whatever form of recording is used for such relationships, the standardized decomposition of characters will simplify this operation since the homologous or converging character will always be stored in the same manner, even when it refers to taxa recorded in separate databases.

Properties of characters: metadata

Characters themselves have properties. These are data about data, or metadata for short. Some are well-known such as the type of a character, e.g. real number (lengths and widths), integer, unordered multistate (nominal), and ordered multistate (ordinal), all of which, except ordinal, are generally supported by existing DBMSs. At the beginning of the NEMISYS project, we defined other kinds of metadata (Diederich *et al.*, 1989; Diederich and Milton, 1991) that are useful in identification since they characterize the qualities of the character for purposes of identification, including conspicuity, ambiguity, and variability of characters, besides their type (entered as range and scale). These metadata are taxon dependent.

In a proper database, fields must be provided for metadata with only one type of metadata per field. This would not be possible with DELTA, where all metadata and relationships have to be stored in the comment field. One major

problem with this would concern the use of the information in *ad hoc* queries, such as in building indexes for the queries.

Metadata make it possible to develop the concept of endorsement, which was first presented as a 'pie in the sky' wish (Fortuner, 1993), but has recently moved into the realm of the possible (Diederich and Fortuner, 1996). Using metadata and fuzzy logic, it is possible for the system to judge the reliability of the data entered by the user.

In GENISYS databases, frequency (the percentage of specimens having a particular character state in a population) is entered as metadata when it is given in the description. It can be used for probabilistic applications (Horvitz, 1993), but it has many other uses. For example, state 1 of character 50 of Thiele (1993) given above is 'always yellow'. This means that this state would be present in 100% of the specimens considered. Obviously, the frequency metadata of GENISYS can be used to recreate Thiele's character state.

Capture of Non-morphological Data

Physiological data

So far, we have been talking of morphological data, but identification and systematics also use other kinds of characters. In the list of characters in Thiele (1993), character # 4 is:

4. Terminal buds of flowering stems transform into inflorescence axes: directly (0); after resting period (1).

Character 4 is a physiological character and it should be decomposed and stored as such. Physiology is the study of the working of organs, tissues, cells and molecules, and these are biological structures that already exist in the morphological database. This should make it possible to use a decomposition where the physiological entities would be connected to the existing morphological entities:

Flowering stems
  Terminal bulbs
    Resting period
      presence
        present / absent

where one recognizes Entity (at three levels of decomposition, two morpho-anatomical levels and one physiological level), basic property, and states.

The presence in the entity hierarchy of the same biological structures as in the morphological tables will provide an easy way to link the various kinds of data, and the same view methods will apply.

Other types of data

Similarly, biochemical data can be so decomposed with the entity being specific molecules, which constitute the bottom level of the morphological entities. In fact, the biological activity of each molecule could be described as physiological data, and ecological data could represent the relationship between an organism and its environment, depending on the molecules it emits into, or receives from this environment, emission and reception being made by particular morphological organs.

Extrapolating from this, it is possible to imagine a general database incorporating all kinds of data – morphological, molecular, physiological, ecological, etc. – which would represent all our knowledge about a particular organism. The organism, its organic structure and functions, and its relations with its environment would be captured and arranged in this huge database, and a network of views, relationships, and metadata could be defined between the various types of data to give life to this electronic monster.

The cost in time and effort for such a project would be staggering, and it takes us far from the subject of this meeting, which is identification. However, it does not cost much to keep such possibilities in mind when we start a limited project. We believe that any database should be created with a possibility of being extended to a wider use and that the data decomposition we propose is naturally suited to such an expansion. We also believe that such a database could attain its full potential only if a Bio-DBMS can be created to support a wide range of uses.

References

Dallwitz, M.J. (1980) A general system for coding taxonomic descriptions. *Taxon* 29, 41–46.

Dallwitz, M.J. (1993) DELTA and INTKEY. In: Fortuner, R. (ed.) *Advances in Computer Methods for Systematic Biology – Artificial Intelligence, Databases, Computer Vision*. The Johns Hopkins University Press, Baltimore and London, pp. 287–296.

Diederich, J. (1997) Basic properties for biological databases: character development and support. *Journal of Mathematical and Computer Modelling* 25(10), 109–127.

Diederich, J. and Fortuner, R. (1996) Endorsement of observations in identification. *IEEE International Conference on Fuzzy Systems, Sept. 8–11, 1996, New Orleans, Louisiana, USA*, pp. 175–179.

Diederich, J. and Milton, J. (1989) NEMISYS: an expert system for nematode identification. In: Fortuner, R. (ed.) *Nematode Identification and Expert-system Technology*. Plenum Publishing Corp., New York, pp. 45–63.

Diederich, J. and Milton, J. (1991) Creating domain specific metadata for scientific data and knowledge bases. *IEEE Transactions Knowledge Data Engineering* 3, 421–434.

Diederich, J. and Milton, J. (1993a) Expert workstations: a tool based approach. In: Fortuner, R. (ed.) *Advances in Computer Methods for Systematic Biology – Artificial*

Intelligence, Databases, Computer Vision. The Johns Hopkins University Press, Baltimore and London, pp. 103–123.

Diederich, J. and Milton, J. (1993b) NEMISYS: a computer perspective. In: Fortuner, R. (ed.) *Advances in Computer Methods for Systematic Biology – Artificial Intelligence, Databases, Computer Vision*. The Johns Hopkins University Press, Baltimore and London, pp. 165–179.

Diederich, J., Fortuner, R. and Milton, J. (1989) Building a knowledge base for plant-parasitic nematodes: description and specification of metadata. In: Fortuner, R. (ed.) *Nematode Identification and Expert-system Technology*. Plenum Publishing Corp., New York, pp. 65–76.

Diederich, J., Fortuner, R. and Milton, J. (1997) Construction and integration of large character sets for nematode morpho-anatomical data. *Fundamental and Applied Nematology* 20(5), 409–424.

Fortuner, R. (1989) A new description of the process of identification of plant-parasitic nematode genera. In: Fortuner, R. (ed.) *Nematode Identification and Expert-system Technology*. Plenum Publishing Corp., New York, pp. 35–44

Fortuner, R. (1993) The NEMISYS solution to problems in nematode identification. In: Fortuner, R. (ed.) *Advances in Computer Methods for Systematic Biology – Artificial Intelligence, Databases, Computer Vision*. The Johns Hopkins University Press, Baltimore and London, pp. 137–163.

Fortuner, R. and Wong, Y. (1985) Review of the genus *Helicotylenchus* Steiner, 1945. 1. A computer program for identification of the species. *Revue de Nématologie* 7, 385–392.

Horvitz, E.J. (1993) Automated reasoning for biology and medicine. In: Fortuner, R. (ed.), *Advances in Computer Methods for Systematic Biology – Artificial Intelligence, Databases, Computer Vision*. The Johns Hopkins University Press, Baltimore and London, pp. 3–27.

Lebbe, J. (1991) Représentation des concepts en biologie et en médecine. Introduction à l'analyse des connaissances et à l'identification assistée par ordinateur. Thèse de doctorat; Université Pierre et Marie Curie, Paris, xii + 282 + xxiv pp.

Pankhurst, R.J. (1991) *Practical Taxonomic Computing*. Cambridge University Press, 202pp.

Pankhurst, R.J. (1993) Principles and problems of identification. In: Fortuner, R. (ed.) *Advances in Computer Methods for Systematic Biology – Artificial Intelligence, Databases, Computer Vision*. The Johns Hopkins University Press, Baltimore and London, pp. 125–136.

Thiele, K. (1993) The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* 9, 275–304.

Webster, R.D. (1988) *A Beginner's Guide to* DELTA (2nd edn). Agricultural Research Service, USDA, Beltsville, MD, 52 pp.