Syst. Biol. 53(5):822–834, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490522313

Metacanthomorpha: Essay on a Phylogeny-Oriented Database for Morphology—The Acanthomorph (Teleostei) Example

AGNES DETTAI,[1] NICOLAS BAILLY,[2] RÉGINE VIGNES-LEBBE,[3] AND GUILLAUME LECOINTRE[1]

[1]*Equipe "phylogenie," UMR 7138 "Systematique, adaptation, evolution," Departement "Systematique et Evolution," Muséum National d'Histoire Naturelle, CP26, 57 rue Cuvier, case postale n° 26, 75231 Paris cedex 05, France; E-mail: lecointr@mnhn.fr (G.L.)*

[2]*Équipe FishBase, Département Systématique et Évolution, Muséum National d'Histoire Naturelle, CP26, 57 rue Cuvier, 75231 Paris cedex 05, France*

[3]*Laboratoire Informatique et Systématique (UMR 5143 Paléobiodiversité et Paléoenvironnements), Université Pierre et Marie Curie, 12 rue Cuvier, 75005 Paris*

For more than a century, researchers have been trying to reconstruct the history of taxa and their relationships, using morphological, behavioral, ecological, physiological, and lately, cytological, karyological, and molecular characters. Over the last decades, phylogenies based on molecular data have grown to become the largest percentage of the publications in this field, and have even been considered by some to hold the keys to the history of life. Some people have even considered the acquisition and study of morphological data obsolete in phylogenetic inference and, unfortunately, the specialized expertise needed to work on the morphoanatomy of a particular group is slowly disappearing. One of the reasons why morphological phylogenies have experienced such a backlash, despite their past successes and the amount of information already available in the literature, might be that the amount of data has surpassed our ability to manage and use it. In most taxonomic groups, and especially in some of the larger ones for which the internal phylogeny is not known (Eukaryota, Acanthomorpha, Aves), these data have accumulated to such a point that they became unmanageable by a single human mind long ago. More often than not, however, the analysis of molecular data has stressed the need for the reassessment of other types of data. The newly obtained relationships are often totally unexpected, and before reconciliation (or an explanation of incongruence) are possible among the various types of data, these must be surveyed for a larger set of taxa.

An example of this is the new molecular phylogeny of eutherian orders (Murphy et al., 2001; Madsen et al., 2001). These results would have been totally unexpected 5 years ago, and because of this incongruency, a complete reassessment of mammalian anatomical characters is needed. A similar example is the new phylogeny recently published for clades of Acanthomorpha (Teleostei), independently recovered by several teams using different molecular markers (Miya et al., 2001, 2003; Chen et al., 2003; Dettaï and Lecointre, 2004, submitted) that also brought several surprises. But part of this surprise effect might come from the fact that data about some of those groups had never been brought together in a morphological matrix.

The molecular data have out-paced the morphological for two main reasons: an ever increasing speed of data acquisition, and several databases like GENBANK, that are easily searchable. Although it is not possible to increase much the speed of acquisition of morphological data, the amount of data to consider is also very important. Parts of the work could be automated in order to increase analytical power. The present paper describes the basis for a database of morphological characters intended specifically to simplify everyday work in phylogenetic reconstruction based on morphological data. The database will enable cross-referencing and comparisons of specimens and character states across large numbers of taxa, even with high variability in attributes. We compare our database to other current databases and explain the similarities and considerable differences.

The Need for Phylogeny-Oriented Databases

The need for powerful ways to manage information flow, make comparisons across data sources, and summarize the result has always been present in the branch of systematics called phylogenetic reconstruction. For a time the medium mattered little: summarizing data can be done without computers, as has been repeatedly demonstrated by numerous review papers and books. But now the amount of data has increased to a point where nondigital media are insufficient, particularly if all characters and taxa, not just a subset, are incorporated in a single study. In other domains, the management of large amounts of data has been under development for some time already.

Databases have developed into a flexible, reliable way not only to store data, but also to manage it, to search for connections and establish correlations at a much larger scale than possible by mental manipulation alone. As databases became more sophisticated, they could no longer be considered only as a powerful tool for information management and retrieval, and began to fully participate to discoveries. In systematics, databasing is a natural and legitimate activity, and its importance has been stressed in a number of papers in the last decade (Sanderson et al., 1993; Lebbe, 1996, among many others). It has already been used with success to keep track of biodiversity, automatically generate identification keys, store sequence data, and many other applications.

Representing Structure of Knowledge: Links between Data

There are many relationships among data used for phylogeny that must be taken into account to give an accurate representation of expressed views in the community and get an accurate retrieval of them. An examination of the processes of coding and interpretation of the information in modern systematics, as well as of the way systematists think in their everyday work, has been conducted in order to facilitate and enlarge the scope of systematics research by providing automatization of the most repetitive tasks.

Objectives

Our relational database, METACANTHOMORPHA, is primarily intended to represent the structure of our knowledge of homologies in spiny bony fish (Teleostei: Acanthomorpha) and help in the retrieval of character states and the search for new primary homologies, by giving the researchers access to enough supplementary information to appraise the data published in the literature before use. It is designed for the storage and management of data for phylogenetic purposes, permiting diachronic links between hypotheses formulated by different authors at different times, sometimes with a different vocabulary. It supports searching for data and bibliographical references for a given taxon and character, evaluating names given to character states and characters, and matrix edition. It has been conceived specifically for comparative work, and allows complex requests and data intersections. This database is intended to be cooperative, with data entry directly by the researchers over the Web.

Other Database Projects

To this day, only a few highly experienced specialists and/or a long and patient exploration of decades of publications can retrieve all the information necessary to generate valid matrices for phylogenetic reconstruction. A database of character states would allow that knowledge to be shared and surveyed more efficiently. The submission of complete matrices for publication of phylogenetic analyses as in TREEBASE (Sanderson et al., 1994) is a first step toward the easy recovery of more strictly controlled results. But it is not possible in TREEBASE to search for a particular taxon or character, nor to summarize results across publications. Each researcher using it must therefore still search through the whole scale of publications in order to obtain complementary information. The concept of a database making these possible is very different, and offers exciting new possibilities for research.

A thorough review of the representation of systematic data and its extractability would deserve its own paper, so we only cite here some of the most relevant examples (Table 1 summarizes some of the properties and differences among some of the main formats and applications). Formats have already been established for several purposes, among which is the representation of descriptive data that allows generation of natural language descriptions and of identification keys from matrices. Characters used for description and those used for phylogenetic purposes are of the same kind, and can be stored in the same type of databases. But the value for description of one particular character and its value for phylogeny are two different things. This must be indicated in any database mixing the two. The main difference between descriptive data and phylogenetic data is that descriptive data needs no hypothesis of homology. It is generally used in an identification purpose, that is, the assignment of an individual object to a concept, most often a taxon (Lebbe and Vignes, 1998). Phylogenetic data are used to aggregate objects into groups, and, as the criterion of aggregation is common descent, the concept of homology is central to it. Descriptive data, although much easier to use when presented in batches, can also stand alone, because they often appeal to external sorting criteria like shape, length, etc., whereas homology is a relational concept (Rieppel and Kearney, 2000) and every hypothesis needs to be replaced among the other formulated at the same time, in order to be assessed critically (contextual validity). There are also differences in the way both are used. Descriptive data generation is labor intensive, but once it is arranged in a computer-aided identification system, it can be used to generate very easily a wide range of ready to use applications, like identification keys or automatic diagnoses. Its purpose is often to provide stable statements about the objects. On the other hand, phylogenetic data generation is also labor intensive, but even with an automated system it requires heavy interpretation and evaluation work from the end-user. To perform innovative work and criticism, the availability of the data in itself is not sufficient, and additional details are required. It is especially interesting in this context to have access to earlier reexaminations and critics, and this is one of the functionalities that was lacking in the previously available databases. Therefore, there is a need for the creation of a database specifically addressing the issues particular to phylogeny. Other teams have seen this need, and several projects (MORPHOBANK, O'Leary et al., 2001, and VIRTUAL FLORAS, Gilbert et al., among others) are currently under development. However, they are mainly oriented towards picture repositories, each having a slightly different purpose. From the documents currently accessible, MORPHOBANK was originally designed to provide access to morphological characters through a vast collection of images and information about these images, in order to help investigations without referring directly to specimens. As we will explain further in the present paper, our aim is rather different. Moreover, MORPHOBANK is devoted to all organisms and is not restricted to a particular group. Project PALEOBANK (Kaesler et al., 2001), which integrates the Treatises for Invertebrate Paleontology series, will mainly present characters for higher rank taxa. Similar projects, focusing even more on being a picture repository, are already online, like the project MORPHBANK (Buffington et al., 1997–1998) and the DIGITAL MORPHOLOGY LIBRARY project (Digital Morphology Group). As for description formats, numerous projects have been or

TABLE 1. Comparative table of the some databases and morphological data standards discussed in this paper. METACANTHOMORPHA, MORPHOBANK, and SDD are still in development, so the information is based on prospective documents available to the authors at the time of publication of this paper (see references), and might be incomplete, or change.

Database		Main goal	Cooperative web based for the entries	Possible interrogations through the web	Importance of illustrations	A posteriori comments and discussion on characters	Link to references	Integrated taxon/character tables
DELTA (1978 onwards)	Format	Flexible language for encoding taxonomic descriptions for computer processing	Not yet available	Depending on application	Entry of illustrations possible	Full text	Possible	Yes, generated on request
SDD (development 2000 onwards)		XML schema for descriptive data exchange	Possible, but needs an additional program	Depending on application	Entry of illustrations possible	Full text	Possible	Possible, but needs an additional program
TREEBASE(1993 onwards)		Phylogenetic trees and associated data matrices repository	Yes	Yes, but limited No taxonomic arborescence, no search by character	Trees only	No	Mandatory. Does not accept unpublished data	No, stored as entered
METACANTHO--MORPHA (development 2002 onwards)	Application	Presentation of homology hypotheses for phylogenetic study with context	Planned	Planned by any criterion necessary for phylogenetic analysis	Advisable	Planned structured comments and references	Mandatory but will also accepts unpublished data references	Planned, will be generated on request
MORPHOBANK (development 2001 onwards)		Morphology pictures repository with context, and homology hypotheses	Planned	Planned by any criterion necessary for phylogenetic analysis	Central compulsory	Planned structured comments and references	Mandatory but will also accepts unpublished data references	Planned, will be generated on request
MORPHBANK (1997 onwards)		General purpose web based picture repository	Yes	Yes, but not by character	Central compulsory	No	Possible	No

are currently available, for example XPER (http://lis.snv.jussieu.fr/apps/xper/doc/XPER.html; Lebbe, 1984) used for example by the CIPA (Computer-aided Identification of Phlebotomine sandlies of America) international program for descriptive data and CAI (http://lis.snv.jussieu.fr/productions/cipa_iao); the LucID package of the Centre for Biological Information Technology of the University of Queensland, or the DELTA format (DEscription Language for Taxonomy, Dallwitz, 1980; Dallwitz et al., 1993), which has been widely used as an interchange format (e.g., VIDE Virus Identification Data Exchange, Boswell and Gibbs, 1986; LIAS Global Information System for Lichenized and Non-Lichenized Ascomycetes, 1995; DEEMY DEtermination of EctoMYcorrhizae, Agerer and Rambold, 1996; crustacea.net, Ahyong et al., 1999 onwards; among a great number of other projects), but will probably be replaced by the Taxonomic Database Working Group (TDWG) Standard of Descriptive Data (SDD) XML descriptive standard currently under elaboration. There are also various databases aiming at the representation of descriptive data/classifications like PANDORA (Pankhurst, 1993), ALICE (White et al., 1993), HICLAS (Beach et al., 1993), TAXON-OBJECT (Saarenmaa et al., 1995), or, more recently, PROMETHEUS (Paterson et al., 2004).

Although these databases cannot be used to represent all the refinements of phylogenetic data as conceived here, the new SDD project of the TDWG, aiming at establishing an XML standard for descriptive data, might be more interesting for this. The SDD format could probably be used to represent also phylogenetic information, but it is an exchange format, not directly adapted to data integration, and it would need a specific interface for the entry and recovery of data. Our database, with an additional functionality for extracting data, could be used to generate XML files compliant with the future SDD standard. This would have one additional advantage: the representation of homology hypotheses supposes the correct representation of rather complex links between the different data. Any forgotten link implies an important loss of information. Generating the files from a structure specifically designed for the storage and manipulation of phylogenetic data would ensure a better reliability of the completeness of links between files and data, as well as of data package integrity that is necessary for correct comparisons. These will have to be considered for future development.

A Suitable Group

Our work is based on the example of the acanthomorphs. This clade of spiny teleostean fishes is currently divided into 314 families and more than 15 000 species. It presents a wide variety of morphological traits. Relationships within the acanthomorphs are far from being resolved, despite the fact that a great amount of data has already been published. Constructing a phylogenetic tree implies a comparison of entities (whatever the name we give to them: taxa, terminals, individuals, species, operational taxonomic units, ...). However, in the current state of the art, it is almost impossible to make a consistent synthesis for their morphological comparison on a large scale, because of the dispersion of the information, its very important heterogeneity, and the isolation of specialists of the subgroups. It is not even possible to focus on a subgroup, because the delineation of such a subgroup requires some knowledge about their broad interrelationships, which are, for most, not known. The group is too vast, too diverse, to have been embraceable by a handmade single morphological data matrix, especially because several researchers have expressed different opinions about most characters. These three problems, added to the need to reassess the congruence between morphological data and the recent results of molecular phylogenies (Wiley et al., 2000; Chen, 2001; Miya et al., 2001, 2003; Chen et al., 2003; Dettai et al., submitted) stimulated the creation of a specifically phylogeny-oriented database, METACANTHOMORPHA (Fig. 1).

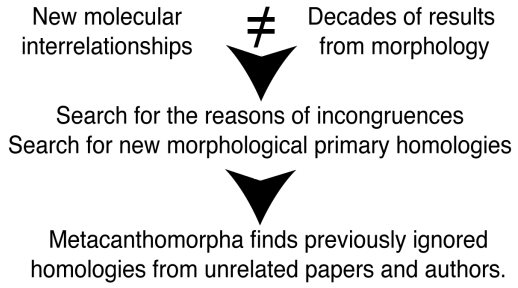

FIGURE 1. Importance of the reassessment of morphological characters for recently revised groups.

Why a Taxonomically Limited Database?

It could be argued that it would be interesting to have all known taxa in a single database. But, even in teleosts, the structural plasticity is such that only a limited number of characters can be compared across the whole taxonomic range. This problem expands as the taxonomic sampling expands, and most characters end up being scored as nonapplicable outside a restricted group, making vizualization of the data more difficult. The carrying out of a database convenient for all groups also requires accounting for all necessities for very different groups, and aside the huge amount of planning work this would require, one could well end up with a base fitting all but not really adapted to any group.

The Species 2000 project (www.sp2000.org) or the GBIF initiative (www.gbif.org) are examples of an alternative to a database covering all groups. General searches are possible through a metadatabase of all databases available to query, and providing a web portal that can redirect users to the relevant database. See Gewin (2002) for a listing of the main meta projects of this type.

Exportability of Such a Database to Other Taxa

Theoretically, conceptual and technical structures of the present database can be applied to any taxonomic

group (though some adaptations may be required for groups with very different biological characteristics like bacteria, viruses or even plants). A lot of digitized information is already available for fishes. But for other large groups, where less data are available online, the entry of a complete list of taxa and synonyms, information about the biology, and complete specimen information might still be required.

THE DATABASE

In the present paper we describe the database and discuss issues and solutions. We relate this to the entity relationship diagram of the database (Fig. 2) currently under construction.

The Complexity of Characters Requires Tracing Back Their Authors

Pointing to a part of a specimen and naming it is very different from applying the same name to two parts of two different individuals: one is a definition, and therefore cannot be questioned, and the other is a hypothesis of sameness. This second case is the one that is widespread in comparative anatomy. Before phylogenetic analysis, a complex process of formalization and coding takes place, leading to the data matrix. The various hypotheses underlying this process are summarized in Figure 3. In a column there are two different, hierarchical, putative homologies. First, homologies among character states, formalized through the grouping together of several different codings under the same column (character homology, CH). Second, homologies among different observations of the "same" character state from different specimens and species (character state homology, CSH), formalized though the attribution in several cells of the same coding number to several observations. Also, depending on the step in the procedures used by the researchers, characters can be understood as primary homologies (hypotheses of homology not yet tested by a tree), or of secondary homologies (homologies confirmed as synapomorphies by a tree; de Pinna, 1991). Neither characters, nor character states, nor even the attribution of an observation to a character state are "raw observations," but hypotheses of homology. As such they are *statements* (hypotheses proposed by an author at a given time, in a given context, that have to be presented in that context with as many precisions as possible to be fully usable). They must therefore retain a link to the author who voiced them. In classical publications, the author of the information within a cell of a data matrix is either given in the accompanying text when the "observation" was taken up from a previous study, or else, implicitly, is the author of the publication. The link between author and/or publication and the data they describe must be as direct and precise as possible, as each cell in a matrix can come from a different source (author and/or publication).

Moreover, each character state in a data matrix has been identified on the basis of specimen(s) that can be different according to authors, on a sampling of specimens that embraces different ranges of biodiversity depending on the authors, and last but not least, associated with a taxon name of different circumscriptions and meanings according to different authors.

Therefore the minimal information contained in each cell of a matrix corresponds to a specimen, to the character state under which the observation has been catalogued (CSH), to the character under which the character state is recorded (CH), and to an author (and publication). *The core and central articulation of the database is therefore the character state/specimen statements cross-table* (Fig. 2). This table links the specimen to the character state to which the statement refers. This table also keeps track of the author's name each time a given character state is scored for a specimen. Links to the bibliographical origin of the data have been represented in our entity relationship diagram (Fig. 2) in an abbreviated form (a star in each table where present) to avoid overcrowding the diagram. Additionally, a description and possibly one or more comments on the observation can help to clarify the meaning of the researcher making the description, and also allows specifying searches in the database.

Precise references to the origin of data have the additional advantage of managing diverging opinions about characters, character states or observations. By adding comments to render the different interpretations of observations supposedly the same (alternative, mutually exclusive codings: CH) or of different observations (CSH: "counter-observations"), researchers can give their opinion and share their observations with high precision. The goal is to have them directly argue in the database their alternative point of view on a character with supporting data. By extending this, it is also possible to represent dependency between characters, as well as opinions about that dependency.

What Data?

Using a standardized terminology for data allows a less ambiguous representation of the information and simpler sorting processes. But it also complicates data entry, as the information extracted from publications has to be translated to that language, and a reverse translation has to be made to allow a user to go back to the original reference. Therefore the decision of not imposing a terminology has been made, so the data can be entered exactly as published, but the disadvantages of this approach have to be circumvented in another way. All characters proposed by an author must be kept, even if they are considered dubious by other researchers, and if their homology is not ascertained. Discarding data is a perilous exercise at best because the grounds for selecting which data to discard are difficult to justify. First it would need a universal competence for determining the quality of the data; second, it is the combination of knowledge from different sources itself that provides the tool for exploring the quality of particular characters. Also, the situation is rarely clear-cut; most often only a few characters in a publication are questionable. Data under doubt and scrutiny are the most interesting to represent, as this allows discussion and clear indication of the points and

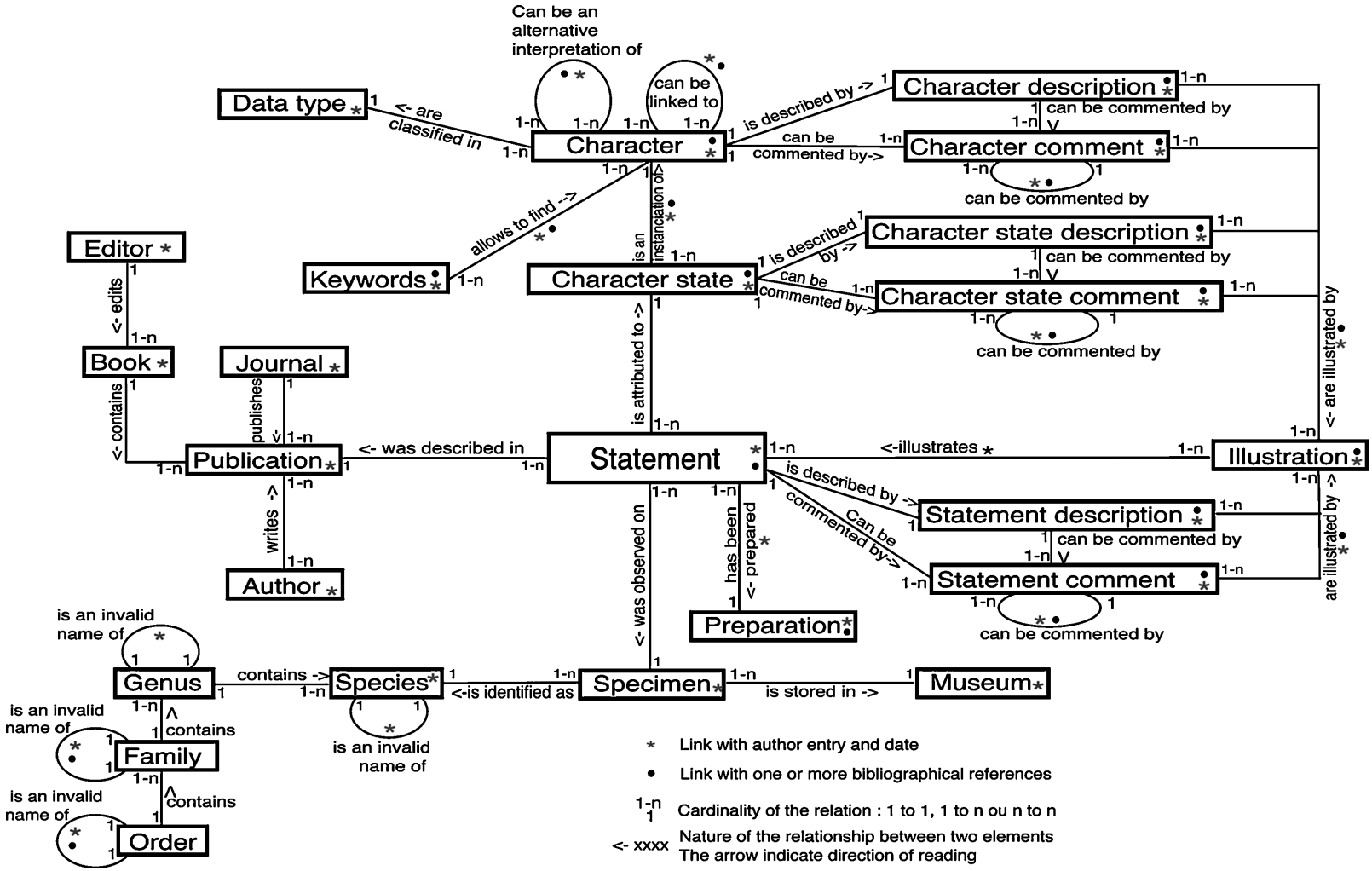

FIGURE 2. Simplified entity-relationship diagram for the exemplar Metacanthomorpha database. This model may change during implementation, but an updated model is available at http://ichtyonb1.mnhn.fr/metacant.

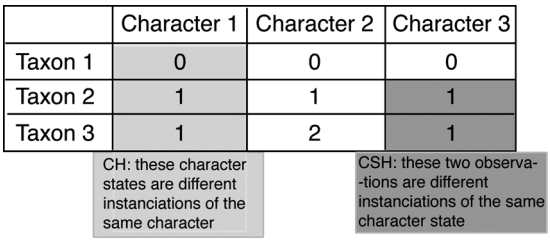

FIGURE 3. The putative hierachical homologies in character and character state in a matrix under the primary homology.

reasons of the dissent, giving an assessment of the reliability that can be at a very precise level (on an observation) or more general (on the validity of a character or a character state). This brings more information than the elimination of the questionable data, and gives to future studies an insight of the data that have already been under scrutiny but did not yield any appreciable result so as not to loose time by reevaluating their applicability again. These comments and criticisms can be gathered from the bibliography or entered directly to the database. Attaching these comments to confirming/disconfirming evidence (i.e., other characters or alternative characters) is also necessary to complete the critics. Comments on characters can also include confirmation of whether analysis has shown them to be secondary homologies. The precision of such an entry can range from a pointer to a reference in literature to a complete description of the analytical context of the study and trees involved. This will provide the researcher with as much data and earlier critics as possible to allow them to evaluate the quality and relevance of the data. The addition of a discussion list on topics related to the database is a desideratum, and would allow an even more dynamic interaction and communication on the data among researchers.

Two Types of Entry

Types of data.—Two very different types of information are present in the database, and they have been partitioned in different tables: tables containing the data per se, that is, information entered by the researchers, and tables containing categories that will allow manipulation and sorting of the data ("metadata" on the data). An example of why such additional entries are needed can be understood by the following example: the database information system might have great difficulties, or even be unable to decide whether a comment made on a datum is confirming it or criticizing it, whereas it is very easy for the author of the comment to decide and indicate this in an adapted field. Another example is the need for being able to sort characters by categories in order to retain the capacity to ask for the display of a limited group of characters. Categories (caudal skeleton characters, neurocranium characters, myological characters, etc.) must be created, as any automatic sorting system will be overly complex as well as not entirely reliable. Other examples of such sorting categories are supra-specific (higher rank) taxa and keywords (see Fig. 2).

Types of data per se.—Table 2 summarizes the various possible sources of the data that can be included in the database and the degree of precision necessary for each, depending on the accessibility of the source of the data. The references to not easily available sources such as theses that have not been published digitally, or other unpublished data must be more detailed than well-known publications. If the user can be reasonably certain that it contains data useful to him or her before

TABLE 2. Summary of the various possible sources for the data that can be included in the database, and the degree of precision necessary for each depending on the accessibility of the source of the data.

	Availability of the data outside the database	Examples of such data	Role	Completeness of the information in the database
Preexisting data	Data already present in other databases	Specimen collection numbers	Pointer to other databases	Must be kept to the lowest, as it can be source of errors, but enough must be present to be able to find the data in the other databases and insure an efficient sorting of the data in this database
		Taxon names Bibliographical references, author names		
	Data already present in papers or thesis	Characters	Direct use or as a pointer to the original source	Can be expanded to be precise enough for direct use, but can also be sketchier and mainly indicative. The degree of precision could be dependant of the availability of the original source
		Character states Descriptions, comments Observations		
New data	Submitted paper	Same as above	Same as above	Same as above
	Not present elsewhere than in its authors mind or data	Discussions	Direct use	Must be complete, as there is no way to refer to another source
		Comments Negative results		

making the effort to get the original, one of the goals of the database will have been accomplished. Yet even if the source is easily available, an amount of data sufficient to give a fair idea of its content and allow sorting has to be present. We will also endeavor to reference data from existing databases, so as not to provide a source of errors by providing copied, but out-of-date data. The update of replicated data is notoriously problematic even when performed very carefully and frequently. This is especially important to keep in mind in fishes, because they are well covered by nomenclature and biological databases such as Fishbase (Froese and Pauly, 2003) and the California Academy of Sciences Catalog of Fishes (http://www.calacademy.org/research/ichthyology/catalog). Also, many museum collections are now digitized and accessible on the Web, so specimen information can be kept to a minimum. Lastly, much of the recent literature is already accessible under digital form. However some redundancy remains a necessity, or else the sorting of the data would be difficult. Therefore some compromise has to be found.

Specimens and Names

The only link of the statement with the real world is the specimen on which it was based. Therefore character states are linked to the taxa only through specimens ("sample description" type of concept representation; Lebbe and Vignes, 1998), as taxa are conceptual human constructions associated with causal explanations of biodiversity (i.e., descent with modification, etc.), and the membership of a specimen in a given taxon is a hypothesis, except when the specimen is the name-bearer of the taxon (holotype). However, in the database each specimen has to be associated to the taxon it is placed in, even if it is not the semaphoront: a matrix where only specimen names are displayed would not be very convenient for every day work and manipulation. Linking the observation directly to the specimen avoids the issue of overgeneralization (Dayrat and Tillier, 2000). Any user of the database can see clearly the number of specimens on which a character was observed, compare it to the degree of generalization the author of the publication used, and decide whether he or she deems that number sufficient or not, making the decision on straightforward data. Although being the most rigorous solution, this is not without problems, as there are cases when there is no specimen information (see part on general problems).

But adding the convenience of taxon names for sorting poses a problem, as names attached to specimens do change. This is either because the identification of the specimen changed, or because the taxon name did. The database must also be able to deal with those changes, not by comprising all existing synonymies, as there are already databases dealing with this for fishes, but by accommodating the few synonymies encountered in the publications. Many papers address the synonymy problem in databasing, and a considerable number of different studies are still under way (Pullan et al., 2000, and many others). But the system adopted here will be a simplified one, where a specimen will be linked to several names, only one of which is at the same time valid from a nomenclatural point of view and also corresponding to the current identification of the said specimen. The other names presented in the database will be those used in the original publication, so it allows the user to see whether there were any changes, and the original taxonomic context in which the specimen was selected for the study.

The terminology of fish muscular and skeletal elements sometimes varies among papers, and can lead to mistakes when the same name designates different elements or when different names designate the same element. A thesaurus stating the synonymies and currently admitted version is therefore necessary. The vocabulary used by the author of the original reference will be entered as such, because the use of a different terminology can reflect two different things. Either a "mistake" (because of an error or of the use of an older terminology) that could be corrected without modifying what the authors meant, or else a different hypothesis of homology, and in that case, any "correction" would alter the original message. This is especially true in fishes, a group where the number of bones is high compared to other vertebrates (see Harder, 1975), and where bone fusions, reductions, and disappearances are commonplace, increasing the difficulty of identifying homologs. For example, did extrascapulars really fuse with the parietals in clupeomorphs? Or did the supratemporal commissural canal change its trajectory into another bone? What is a parietal bone made of (see review in Zaragueta-Bagils et al., 2002)? To which rodlike pectoral radials of percids do correspond each of the three plate-like pectoral radials of notothenioids (Lecointre et al., 1997; Balushkin, 2000)? Keywords using the thesaurus will be used to attach a standard word (name currently in use for the element) to the original description, so the character can be found using a "modern" name as well as the originaly used one.

Peer Review and Data Reliability

One of the major problems in communication among scientists, and especially on the World Wide Web, is that of reliability. When one aims at building a cooperative work where researchers can participate directly, this is even more of a problem. Several means can be combined to offer greater security and reliability. But there is no way to totally eliminate risks.

Restriction of data entry.—Restriction of the entry of data to competent professionals is one such mean. It is made possible by the fact that the number of professionals at any given moment is not too large. Request for a login and password can be evaluated in numerous ways. Sponsored membership is well known in many scientific societies: an acknowledged member of the society, who stakes his or her reputation on the sponsorship, must recommend a new member. A submission of a list of publications in peer-reviewed journals could also serve as a passport. Although more labor intensive, a live administrator rather than an automatic authorization process helps prevent hacking. Yet it is never possible to

completely circumvent it. The key there is to have very frequent backups, as well as a careful maintenance to spot problems rapidly, which allows replacement of corrupted copies with their backup.

Data entry authorship.—For all entries, the author of the entry and the date of the entry, whether comments or data, will be recorded. This has several advantages, for instance permitting a differed release of the data to the general public until the results are published, or the suppression data entered in the database during a security breach. The feeling of responsibility and therefore the seriousness of the entry are augmented by the fact that the author name is attached to each and every entry. This also permits an indirect assessment of the reliability of the entry, as the name can be displayed next to the entry. When the author of the data himself is the person who enters it in the database, there is little chance of a misunderstanding during entry. It is obviously less so when a person not related to a study enters the data, no matter how much attention it is given.

Critics and peer review.—Finally, the ability to add comments directly will serve as a kind of peer review and give the end-user a more direct way to assess the quality of the data. This will give a post-publication peer review, as is already the case with classical publications, only with faster and more accurate feedback, as problems can be pinpointed very precisely.

The Addition of Pictures

A written precise definition and description are necessary at all levels, whether character, character state in general, or description of a character state for a given specimen. But a well-chosen drawing or photograph can bring even more precision. Yet a picture has no meaning by itself, it needs to be replaced in a context, either by the knowledge of the user or by an attached statement of the homology it illustrates. We will not provide pictures without that statement, as the hypothesis of homology is the important and searchable part, the picture being just an accessory to the description.

Policy relevant to pictures.—We do not think that pictures alone will really help to avoid the need to return to the specimens, but they will certainly help to clarify the delineation of the characters and character states. They must be added whenever possible. The room being limited in most journals, and the publication of pictures leading to significant extra costs, the collection of published pictures is often not as complete as the author(s) of the publication would have liked. Additional pictures can be added in the database, and therefore supplement the published work. But copyrights of the published pictures remain generally with the publishers. Until a proper solution can be devised, only pictures with expired copyrights and unpublished pictures provided by their author (who will retain copyright) will be available.

Requests

Data extraction.—A wide variety of data combinations can be extracted from such a database. The standard

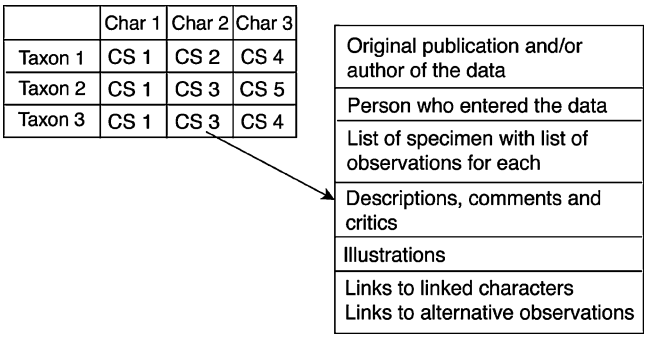

FIGURE 4. Overall view of the homology hypotheses under the form of a character matrix. Each cell of this view is linked to additional and more precise information about its content.

view would comprise a taxon/character matrix, as this is the display mode that allows users to see a wide sampling of taxa and characters at a glance. Yet the necessary additional information discussed earlier must also be available, so that each character, character state, taxon, and coded observation is accessible through a hyperlink (Fig. 4). The content of the displayed matrix can also be selected and restricted, either by taxon, type of character (osteology, soft anatomy, caudal skeleton, neurocranium, etc.), author (either a given publication or all published data), or other more complex requests like implication of a given bone, presence of a matrix in the original publication, or existence of an illustration, etc. Of course, a combination of those criteria will be possible. Such a system allows uses ranging from quick checks of what character states are present in a given group, to finding a publication according to its content, or to assessments of how well-studied various groups have been. In addition to the standard matrix output, more complex requests are also possible, combining different elements of the database to obtain more precise and better-delimited results.

Export.—Exportation of matrices in various standard formats will also be available, as well as the exportation of a list of all references used for the construction of the corresponding matrix. The exportation of a matrix as it was published will be possible, but also the exportation of composite matrix, i.e., a matrix being composed of different sets of characters and codings from different authors and sources. Of course, in this case, many characters will be scored as unknown for a consequent number of taxa.

DISCUSSION

Usefulness of Such a Searchable Data Repository

Obvious possible applications of the database range from easier access to bibliographic information for neophytes (i.e., students) and professionals, to a better direct access to huge amounts of data that will become tractable much more easily for all researchers. Through this, such a database has the potential to boost the discovery of primary homologies, as do wide scale surveys.

The specialization of research produces a kind of isolation, whatever the field. In systematics, a specialist

knows about the occurrence of a morphological trait in his group of interest, but may be less certain of its distribution in other groups. The example of Acanthomorpha is very good in this respect. To perform a good survey of acanthomorph anatomical diversity requires a high level of competence, experience, and huge collections. As in any other group, a specialist of a suborder who tries to understand the phylogeny of his group needs to find an outgroup in order to polarize his characters. Instead of having a single suborder (or two or three) to deal with, there are dozens of potentially relevant outgroups because the phylogeny is unknown and most orders are poorly defined (e.g., Scorpaeniformes, Perciformes). Consequently, the totality of groups cannot be investigated in a reasonable time span, and by selecting a limited number of outgroups, the risk of polarizing the characters in the wrong way is large, finally producing a meaningless tree (Nixon and Carpenter, 1993). The following example demonstrates this. Balushkin (1992) thought that the absence of predorsal bones was a derived character among Antarctic Perciformes (suborder Notothenioidei), because he observed these bones in one outgroup, the Ammodytidae (sand lances, Trachinoidei). Yet if he had covered a higher number of families of the polyphyletic Trachinoidei, he would have seen that presence or absence of that bone is variable among trachinoid families (Pietsch, 1989: 256–257). Had he chosen zoarcoids (eelpouts) as an outgroup, he would have obtained an exactly reverse polarization of the trait (see Lecointre et al., 1997). So, such a database is particularly indicated for monophyletic groups that include a high species richness, a high number of taxonomic entities with unknown interrelationships, and poorly defined medium-sized taxa.

Databases of systematic information can contribute to the robustness and transparency of systematic hypotheses, by forcing users to describe all the hypotheses involved. If systematics is to produce objective knowledge, assumptions and observations of systematists must be explicit, instead of relying on authority. For example, the use of archetypes to polarize traits, among other problems (Bryant, 1997), often lacks character coding also (as in Balushkin, 2000: 75), and makes the work nonreproducible. Such a database of character states helps to make clear the level of generalization permitted for a character state. Dayrat (2000) and Dayrat and Tillier (2000) have described in detail the tremendous impact of "overgeneralizations" of character states on phylogenetic reconstruction, a priori generalizations (when character state instability within a terminal taxon is considered negligible or simply ignored, but see also Wiens 1998a and 1998b) and a posteriori generalizations (assignation of an attribute to the whole group in the conclusions). If a researcher must precise in a database on what specimen the trait was observed, the extent to which the character state can be generalized comes under control.

Cooperation.—This project will be a cooperation, although an indirect one, between all researchers wanting to enter or use data. It might even trigger a new research dynamic by allowing to pinpoint the areas of expertise (even past expertise), of each participant, and ultimately provide new synergetic effects both in information retrieval and in the dispersal of taxonomic expertise, by enhancing cooperation and communication on precise problems.

General Problems

Many problems still remain, as they are external to the database.

There is one very important pitfall to such a database: it can be used to export data and analyze it right away, without going back to the original references and checking the various reliabilities, dependencies, and alternative interpretations of observations, character states, and characters. This risk already exists with matrix repositories, and it is the researcher's responsibility to avoid the blind use of "black box" technology that can only lead to skewed results.

METACANTHOMORPHA, as any such database, is a tool that must be considered as improving the availability of references and links across the data, not as a kit for quick and dirty matrices, even if it has the potential to be used in that way.

But there are also other problems that have more to do with the choices made for data representations, the transposition of data in a form compatible with the database, and the limitations of the application itself.

Data and pictures are sometimes presented in the literature without a reference to a specimen, especially in older publications. The reliability and use of such data are suspect, even though these represent a great amount of work whose exclusion would erase dozens of years of research, but at the same time the risk of over interpretation while entering them in the database is much greater than for more recent, explicit publications, as the person entering them will have to fit those generalizations into a more precise framework. This meets the concerns expressed in several recent and older publications about the need to have voucher specimens for all publications and the drawbacks of such policy (Barkworth and Jacobs, 2001; Griffith and Bates, 2002; Wheeler, 2003, among many others). This is not an easily solved question, and is quite out of the scope of the present paper. In many older publications, there is no indication of the specimens that were studied. Although the database can and will integrate this kind of direct link (though the use of "ghost" specimen references), this information is always a generalization, and should be considered more carefully than recent publications referring to voucher placed specimens. The technical solution adopted for this allows to discriminate between the two types of data, the precise sample based one and the generalized one.

Another problem met in numerous older publications is the lack of data matrix. When there is no matrix, the increased number of implicit assumptions renders the reproducibility of results more difficult. Such data should be entered with as much detail as possible, and indications should be provided as to its context. An option to sort/eliminate these unexplicit/imprecise data from the

analyses will also be included. Yet although this poses a problem at data entry, it is also an important asset of the database. In these cases, the data in a text form will have to be converted into a matrix of statements. This is what happens in everyday work by anatomists, when they try to make sense of more ancient publications and use the data for their work. Although this conversion is a risky approach, the presentation of the result of this conversion to a large community of anatomists is the best way to discuss the local problems, ending up either on an agreement or at least on a pinpointing of the zones of dissent. For instance, the clade Euteleostei was considered as an important group in the phylogeny of teleostean fishes for 20 years (1973–1993; but see Lauder and Liem, 1983). However, a survey of the literature shows that the Euteleostei has never been obtained from a data matrix, morphological or molecular. Its acceptance by ichthyologists was based on tradition rather than hard data (for a critical view of euteleostean traits, see Lecointre and Nelson, 1996). When procedures became explicit, that group disappeared (Lê et al., 1993; Arratia, 1997, 1999; Zaragueta-Bagils et al., 2002).

This problem is linked to the one of the integration of several data matrices from different studies in a single matrix. Equating characters in different works is not a trivial problem, and in some cases might not be possible. In these cases, the characters will not be lumped in a single column, they will be considered as different hypotheses, that might be reunited later on. But researchers also often re-use characters defined in earlier publications, considering the character they use to be the same. In these cases, they will be integrated as a single character altough there are several sources.

The interoperability of databases is nowadays under focus, but in this case direct links to other databases are not absolutely necessary. Whenever additional information is needed, a unique reference to the database (or publication) containing it and a precise reference are sufficient to find it again.

Finally, one persistent problem with any such wide computerization program is the speed of data entry. The more data present, the more interesting a database is, with, ultimately, a database containing all of the available information as a goal. This database is designed to be progressively filled up by the community of ichthyologists. The direct entry of the data by the authors should be reasonably rapid, but even so, some time will go by before it catches up with decades of research. One must be aware that in its early stages the database will lag behind the state of the art in the ichthyological knowledge of acanthomorphs.

CONCLUSION

Toward a New Research Dynamic

The importance of data accessibility in biology has been stressed many times and numerous projects are under way to render taxonomy more accessible (McCabe, 1999; Stein, 2002; Godfray, 2002; Gewin, 2002; etc.). The database described here is different from other databases such as MORPHOBANK that are designed to provide access to morphological characters through a vast collection of images and information about these images. Other databases, some of which briefly described in the text, share some aspects with ours but do not serve the same purpose. Our database is dedicated to fish characters for a specifically phylogenetic purpose and should help researchers working on this group to save time on data entry and bibliographical research, allowing them to select publications by their content in taxa or characters of interest. Above all, it might *reveal unexpected primary homologies* that will support confirmed or new relationships by allowing large-scale comparisons that have not been conducted before.

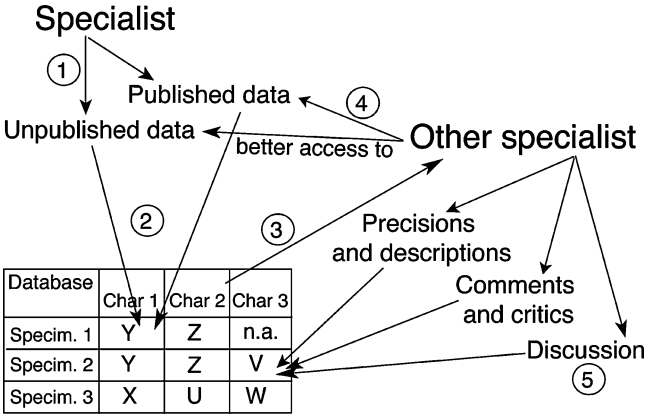

FIGURE 5. A new dynamic of interaction among researchers. The data produced (1) by one specialist is referenced (2) in the database, which advertises to another specialist (3). Any other specialist can then easily pinpoint the publications that are of interest to him or her, and go check the original source (4). The new work or analysis may result in direct comments and other additions (5) to the database on precise parts of the data. This could enhance a more dynamic and precise interaction among researchers.

It will allow a broad view of many publications and taxa and will illustrate almost instantly what taxa or characters can be of interest for a given study, what groups have already been well studied and what groups have not. It will also enable communication among scientists in more precise and more dynamic ways (i.e., discussions can be aimed at a single litigious character or observation), allowing for direct discussion (Fig. 5). The database will allow the systematic community to have a better view of who has worked on what, and help direct requests for expert assistance, promoting an innovative new dynamic of research in comparative biology (Fig. 5).

ACKNOWLEDGMENTS

We thank Paulo Brito, Sara Scharf, Chris Simon, and an anonymous reviewer for useful comments on the manuscript. The whole team of the Laboratory of Informatics and Systematics of Paris 6 University, the ichthyologists, palaeoichthyologists, and database people of the Muséum National d'Histoire Naturelle of Paris, and Istvan Dettaï are warmly thanked for their help in computing and conception. We also want to thank for funding the first author the Comité National des Sciences Biologiques (August 2002 Willi Hennig Meeting in Helsinki, Finland) and the Society of Integrative and Comparative Biology (January 2003 meeting in Toronto, Canada).

REFERENCES

Agerer, R., and G. Rambold. 1996. DEEMY: A DELTA-based system for characterization and DEtermination of EctoMYcorrhizae. CD-ROM. Institute for Systematic Botany, München.

Ahyong, S., M. Berggren, N. Bruce, L. Buhl-Mortensen, D. Camp, P. Davie, P. Firminger, S. Gerkin, E. Gonzalez, T. Haney, P. Haye, C. Hof, D. Jones, J. Just, S. Keable, B. Kensley, K. Larsen, S. LeCroy, J. Lowry, R. Maddocks, P. McLaughlin, K. Meland, A. Myers, A. Parker, R. Peart, G. Poore, R. Rios, J. Thomas, G. Walker-Smith, L. Watling, G. Wilson, and J. Yager. 1999 onwards. crustacea.net: an information retrieval system for crustaceans of the world. http://crustacea.net/

Arratia, G. 1997. Basal teleosts and teleostean phylogeny. Palaeoichthylogica 7:5–168.

Arratia, G. 1999.The monophyly of Teleostei and stem-group teleosts. Consensus and disagreements. Pages 265–334 *in* Mesozoic fishes 2—Systematics and the fossil record (G. Arratia and H.-P. Schultze, eds.). Verlag Dr. Friedrich Pfeil, München.

Balushkin, A. V. 1992. Classification, phylogenetic, and origins of the families of the suborder Notothenioidei (Perciformes). J. Ichthyol. 32:90–110.

Balushkin, A. V. 2000. Morphology, classification, and evolution of notothenioid fishes of the Southern Ocean (Notothenioidei, Perciformes). J. Ichthyol. 40:S74–S109.

Barkworth, M. E., and S. W. Jacobs. 2001. Valuable research or short stories: What makes the difference? Hereditas. 135:263–270.

Beach, J., S. Pramanik, and J. H. Beaman. 1993. Hierarchic taxonomic databases. Pages 241–256 *in* Advances in computer methods for systematic biology: Artificial intelligence, databases, computer vision. John Hopkins University Press, Baltimore, Maryland.

Boswell, K. F., and A. J. Gibbs. 1986. The VIDE (Virus Identification Data Exchange) project: A data bank for plant viruses. Pages 283–287 *in* Development and applications in virus testing (R. A. C. Jones and L. Torranceeds, eds.). Association of Applied Biologists, UK.

Bryant, H. N. 1997. Hypothetical ancestors and rooting in cladistic analysis. Cladistics 13:337–348.

Buffington, M., J. L. Nieves-Aldrey, J. Pujade-Villar, F. Ronquist, F. Fontal-Cazalla, J. Liljeblad, and P. Ros-Farre. 1997–1998. Morphbank (www.morphbank.com). MORPHBANK announcement 2003. Syst. Biol. 52:568.

Chen, W. J. 2001. La répétitivité des clades comme critère de fiabilité: Application à la phylogénie des Acanthomorpha (Teleostei) et des Notothenioidei (acanthomorphes antarctiques). PhD thesis Paris VI University, Paris.

Chen, W. J., C. Bonillo, and G. Lecointre. 2003. Repeatability of clades as a criterion of reliability: A case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. Mol. Phyl. Evol. 26:262–288.

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

Dallwitz, M. J., T. A. Paine, and E. J. Zurcher. 1993 onwards. User's guide to the DELTA System: A general system for processing taxonomic descriptions, 4th edition. http://biodiversity.uno.edu/delta/

Dayrat, B. 2000. L'échantillonnage des caractères et des taxons en phylogénie: Systématique et évolution des Gastéropodes euthyneures (Mollusca). PhD thesis, University of Paris 7, Paris.

de Pinna, M. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7:367–394.

Dettai, A., and G. Lecointre. New data confirm the recent clades obtained by multiple molecular phylogenies in the acanthomorph bush. Submitted to J. Int. Comp. Biol.

DIGITAL MORPHOLOGY GROUP. Digital Morphology Library (http://www.digimorph.org), retrieved from the WWW 04/04/04.

Froese, R., and D. Pauly. 2003. FishBase. World Wide Web electronic publication, http:// www.fishbase.com. Accessed in 2003.

Fujita, K. 1990. The caudal skeleton of teleostean fishes. Tokyo University Press, Tokyo.

Gilbert, E., C. Gries, L. Landrum, and P. McCartney. Virtual Floras: A database integration Poster presentation. http://ces.asu.edu/bdi/Subjects/virtual_floras/. Retrieved from the www on the July 16, 2003.

Gewin, V. 2002. All living things, online. Nature 418:362–363.

Godfray, H. C. J. 2002. Challenges for taxonomy. Nature 417:17–19.

Griffiths, C. S., and J. M. Bates. 2002. Morphology, genetics and the value of voucher specimens: An example with *Cathartes* vultures. J. Raptor Res. 36:183–187.

Harder, W. 1964. Anatomie der Fische, 1st edition. E. Schweizerbartsche Verlagsbuchhandlung (Nagele u. Obermiller), Stuttgart.

Inoue, J. G., M. Miya, K. Tsukamoto, and M. Nishida. 2003. Basal actinopterygian relationships: A mitogenomic perspective on the phylogeny of the "ancient fishes." Mol. Phyl. Evol. 26:110–120.

Kaesler, R. L., J. W. Krebs, and D. L. Miller. 2001. The role and design of databases in paleontology. Pages 377–395 *in* Fossils, phylogeny and form (J. M., Adrain, G. D. Edgecombe, and B. S. Lieberman, eds.). Topics in geobiology 19. Kluwer Academics/Plenum Publishers, New York. http://www.ukans.edu/~paleo/paleobank.html

Lauder, G. V., and K. F. Liem. 1983. The evolution and interrelationships of actinopterygian fishes. Bull. Mus. Comp. Zool. 150:95–197.

Le, H. L. V., G. Lecointre, and R. Perasso. 1993. A 28S rRNA based phylogeny of the Gnathostomes: First steps in the analysis of conflict and congruence with morphologically based cladograms. Mol. Phyl. Evol. 2:31–51.

Lebbe, J. 1984. Manuel d'utilisation du logiciel XPER. Micro Application, Rueil Malmaison.

Lebbe, J. 1996. Quelques réflexions sur l'informatique appliquée à la systématique en France. Biosystema 14:5–10

Lebbe, J., and R. Vignes. 1998. State of the art in the computer-aided identification in biology. Océanis 24:305–317.

Lecointre G. 1995. Molecular and morphological evidence for a Clupeomorpha-Ostariophysi sister-group relationship (Teleostei). Geobios Spec. Pub. 19:205–210.

Lecointre, G., C. Bonillo, C. Ozouf-Costaz, and J. C. Hureau. 1997. Molecular evidence for the origins of Antarctic fishes: Paraphyly of the Bovichtidae and no indication for the monophyly of the Notothenioidei (Teleostei). Polar Biol. 18:193–208.

Lecointre, G., and G. J. Nelson.1996. Clupeomorpha, sistergroup of Ostariophysi. Pages 193–207 *in* Interrelationships of fishes II. (M. L. J. Stiassny, L. Parenti, and G. D. Johnson, eds.). Academic Press, San Diego.

LIAS. 1995–2004. A global information system for lichenized and non-lichenized Ascomycetes: www.lias.net.

McCabe, H. 1999. Vast database offers vision of biodiversity. Nature 400:5.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptative radiation in two major clades of placental mammals. Nature 409:610–614.

Miya, M., A. Kawaguchi, and M. Nishida. 2001. Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial sequences. Mol. Biol. Evol. 18:1993–2009.

Miya, M., H. Takeshima, H. Endo, N. B. Ishiguro, J. G. Inoue, T. Mukai, T. P. Satoh, M. Yamaguchi, A. Kawaguchi, K. Mabuchi, S. M. Shirai, and M. Nishida. 2003. Major patterns of higher teleostean phylogenies: A new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phyl. Evol. 26:121–138.

Monod, T. 1968. Le complexe urophore des poissons téléostéens. IFAN-Dakar, Dakar, Senegal.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origin of placental mammals. Nature 409:614–618.

Nixon, K. C., and Carpenter, J. M. On outgroups. Cladistics 9:413–426.

O'Leary, M., J. Caira, and M. Novacek (eds.). 2001. Morphobank November 2001 workshop report. http://www.morphobank.net/morphobank.pdf. Retrieved from the World Wide Web in June 2003.

Pankhurst, R. J. 1993. Taxonomic databases: The PANDORA system. Pages 229–240 *in* Advances in computer methods for systematic biology: Artificial Intelligence, databases, computer vision (R. Fortuner, ed.). John Hopkins University Press, Baltimore, Maryland.

Paterson, T., J. B. Kennedy, M. R. Pullan, A. Cannon, K. Armstrong, M. F. Watson, C. Raguenaud, S. M. McDonald and G. Russell. 2004. A universal character model and ontology of defined terms for taxonomic description. Pages 63–78 *in* Lecture notes in bioinformatics 2994 (E. Rahm, ed.). Proceedings of Data Integration in the Life Sciences (DILS 2004). Springer Verlag, Berlin.

Pietsch, T. W. 1989. Phylogenetic relationships of trachinoid fishes of the family Uranoscopidae. Copeia 1989:253–303.
Pullan, M. R., M. F. Watson, J. Kennedy, C. Raguenaud, and R. Hyam. 2000. The Prometheus taxonomic model: A practical approach to representing multiple taxonomies. Taxon 49:55–75.
Rieppel, O., and Kearney, M. 2002. Similarity. Biol. J. Linn. Soc. 75:59–82.
Saarenmaa, H., Leppäjärvi, S., Perttunen, J. and Saarikko, J. 1995. Object-oriented taxonomic biodiversity databases on the World Wide Web. Pages 121–128 *In* Internet applications and electronic information resources in forestry and environmental sciences (A. Kempf and H. Saarenmaa. eds). Workshop at the European Forest Institute, Joensuu, Finland, August 1–5, 1995. EFI Proceedings 10.
Sanderson, M. J., B. G. Baldwin, G. Bharathan, C. S. Campbell, D. Ferguson, J. M. Porter, C. Von Dohlen, M. F. Wojciechowski, and M. J. Donoghue. 1993. The growth of phylogenetic information and the need for a phylogenetic database. Syst. Biol. 42:562–568.
Sanderson, M. J., M. J. Donoghue, W. Piel, and T. Eriksson. 1994. TREEBASE: A prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am. J. Bot. 81:183.
Stein, L. 2002. Creating a bioinformatics nation. Nature 417:119–120.
Wheeler, T. A. 2003. The role of voucher specimens in validating faunistic and ecological research. Biological Survey of Canada (Terrestrial Arthropods) Document Series 9:1–21.
White, R. J., R. Allkin, and P. J. Winfield. 1993. Systematic databases: The BAOBAB design and the ALICE system. Pages 297–312 *in* Advances in computer methods for systematic biology: Artificial Intelligence, databases, computer vision. (R. Fortuner, ed.). John Hopkins University Press, Baltimore, Maryland.
Wiens, J., 1998a. Testing phylogenetic methods with tree congruence: Phylogenetic analysis of polymorphic morphological characters in phrynosomatid lizards. Syst. Biol. 47:427–444.
Wiens, J. 1998b. The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis: A simulation study. Syst. Biol. 47:397–413.
Wiley, E. O., G. D. Johnson, and W. W. Dimmick. 2000. The interrelationships of acanthomorph fishes: A total evidence approach using molecular and morphological data. Biochem. Syst. Ecol. 28:319–350.
Zaragueta-Bagils, R., S. Lavoué, A. Tillier, C. Bonillo, and G. Lecointre. 2002. Assessment of otocephalan and protacanthopterygian concepts in the light of multiple molecular phylogenies. C. R. Acad. Sci. 325:1191–1207.

First submitted 25 July 2003; reviews returned 7 March 2004; final acceptance 17 June 2004
Associate Editor: Chris Simon