MorphoBank

WORKSHOP REPORT

AMERICAN MUSEUM OF NATURAL HISTORY, NOVEMBER 10 AND 11, 2001

Conveners: Maureen O'Leary, Janine Caira, Michael Novacek

Supported by the National Science Foundation Divisions of Earth Sciences, Biological Infrastructure, and Environmental Biology

1. The scientific problem and the need for MorphoBank

Taxonomists have used comparative anatomy to describe and categorize organisms for centuries. The meaning and impact of their work is recorded in images that capture hypotheses of homology. These images document the fundamental consequences of evolution: the staggering diversity of life around us. The recent infusion of comparative gene data notwithstanding, outlines of biological classification remain largely based on morphological evidence.

For the first time there is an opportunity to create an enormous interactive image-based repository of comparative morphological data. The information technology revolution has now made possible the large-scale integration of comparative anatomical images and codified cladistic data. This workshop considered initial steps towards the foundations of an interactive database of this kind, here designated, MorphoBank. This database would greatly enhance research efforts in comparative biology and many related disciplines such as medicine, conservation, and education.

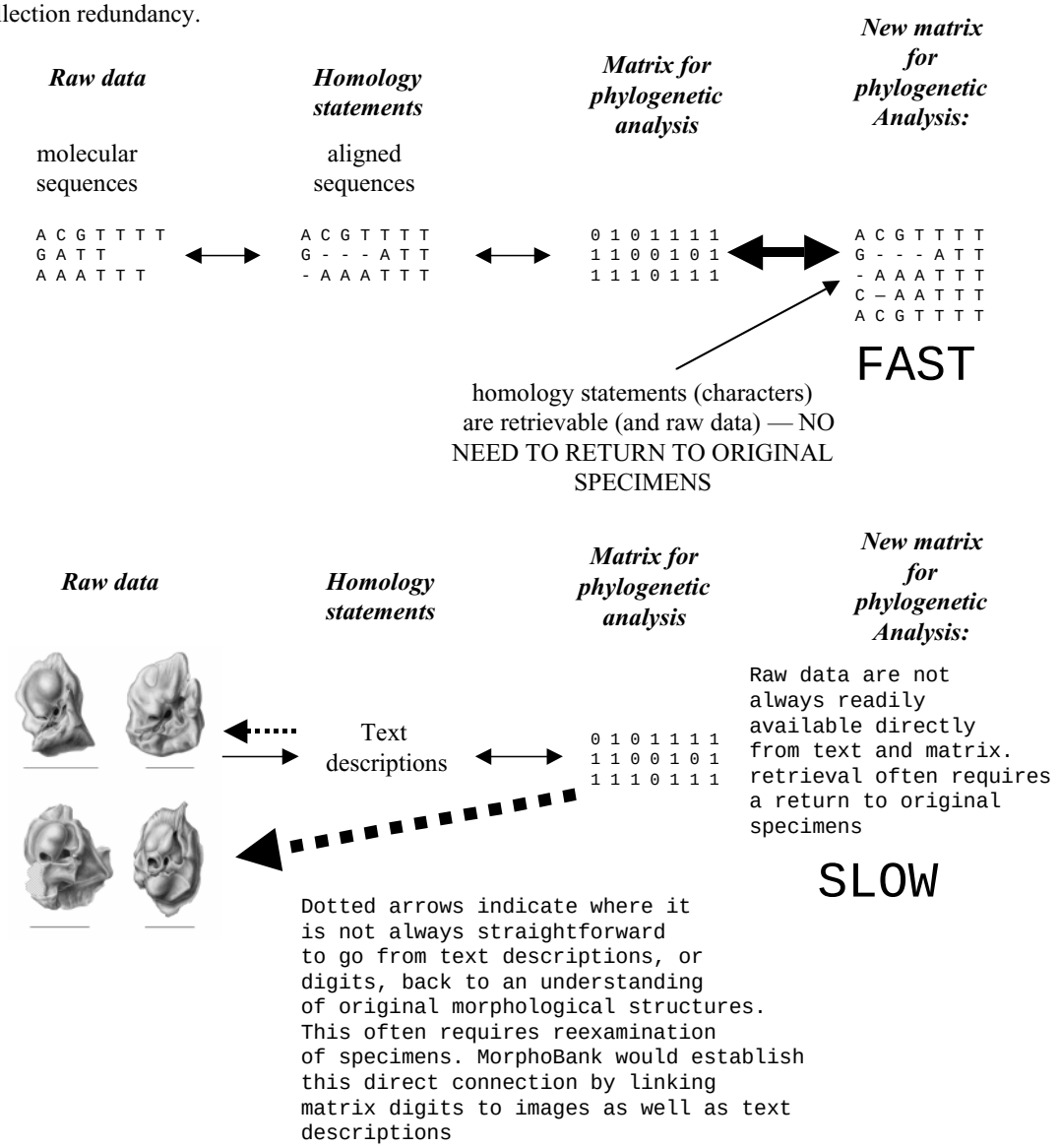
The widespread application of phylogenetic systematics has created a high standard for evidence in contemporary comparative biology. By requiring that systematists divulge specific homology hypotheses (characters) fundamental to their notions of phylogeny, the connection between hypotheses of relationship and evidence becomes very immediate. Before morphological data for systematics research can be used, however, they must become translated into digital form (i.e., coded as 0's, 1's, etc ; Figure 1).

As systematists investigate increasingly complex morphological problems with many species and many characters, research can slow dramatically due to the need to observe directly, from specimens, the majority of published morphological characters in order to add new data with those same features. For example, if a systematist wishes to add a new taxon to an existing phylogenetic study, it is often essential for that systematist to return to museum collections or resection an organism in order to make the meaningful comparisons required to add a new taxon to the analysis. This is because accurate, high-quality images documenting all of the morphological characters in a cladistic analysis are virtually never published or disseminated with original papers. The duplication of effort required to expand a morphological matrix greatly inhibits the rapid advance of contemporary morphological systematics.

By contrast, molecular sequence data, also widely used in systematics research, are inherently digital (Figure 1), raw sequences of A, T, C and G (translated into 0, 1, 2, 3 for analysis) can be readily databased for this reason. Original sequences can also be deduced from typical alignments (Figure 1), thus there is no loss of information when translating molecular sequence characters into numbers because both are digital. This ease of access to raw data facilitates rapid growth in the size of molecular matrices for

phylogenetic analysis because an investigator does not have to resequence taxa already studied (but always can if he or she desires).

Figure 1. Comparison of the relative retrievability of molecular sequence data and morphological data used in phylogenetic analysis. Sequence data are fully retrievable from a data matrix. New molecular phylogenetics analyses that build on preexisting ones can proceed much faster than new morphological analyses. This occurs, in part, because original morphological observations are not as directly retrievable from data matrices and images with labeled characters are not available on line. An investigator working with molecular sequence data has a greatly reduced requirement to return to the original specimens, even if there is a homology dispute. Ability to retrieve labeled images of characters from an on line database such as MorphoBank would greatly enhance large scale morphological analyses and would reduce data collection redundancy.



The practical success of GenBank, the international database for molecular sequences, including its usefulness beyond phylogenetic systematics, suggests that a morphological equivalent would enjoy spectacular success among systematists, and other researchers in a diversity of disciplines outside of this field including functional morphology, developmental biology, ecology, physiology, conservation biology, forensics and education. Furthermore, in contrast to the data in GenBank, the data in MorphoBank, with its inherently visual nature, has the potential to appeal to nonspecialists of all ages who are interested in questions about the phenotype.

2. The concept and design: MorphoBank as a portal to comparative anatomy

MorphoBank will be a repository for any type of morphological image, as well as information about these images. The underlying organizational structure will link images to a series of informative tables (e.g., taxon names, anatomical labels, cladistic character states; Figure 2). The number of tables within MorphoBank to which an image must be linked will be explored by the workshop steering committee (specified below) and is anticipated to grow with the actual use of the database. Importantly, because we wish the database to be highly inclusive, an image is not required in all cases to be linked to all tables. Indeed the pattern of linkages to particular tables may vary greatly among higher taxa.

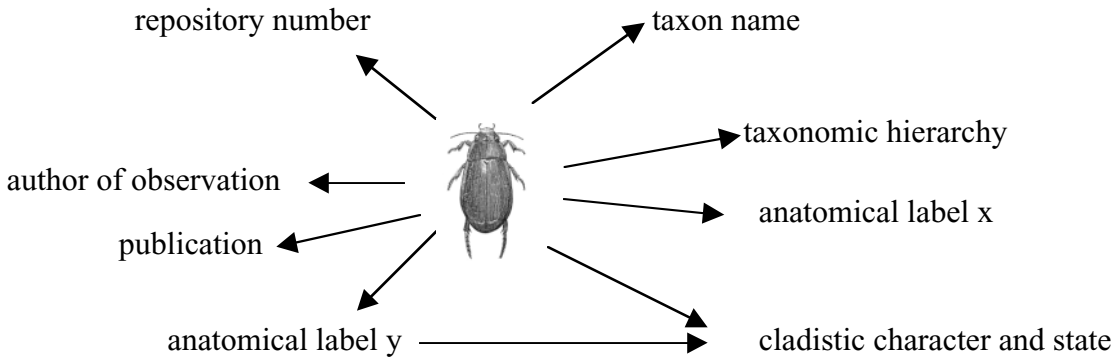


Figure 2. Simplified conceptual framework for MorphoBank. An image is linked to many tables, but not necessarily all tables. Tables are linked to each other (not shown).

The images should be of various kinds, that is classical drawings, photographs, sections, or computed tomography images to name a few examples, and documentation of particular characters, both qualitative and quantitative, need not be restricted to one image only. Images would be submitted to the database, much in the manner that sequences are submitted to GenBank, either as a singular contribution or en masse as the result of a large-scale phylogenetic analysis. One important difference between GenBank and MorphoBank would be that MorphoBank will be able to accommodate more information than GenBank typically does, namely specific documentation of homology hypotheses from cladistic analyses. Images could also be downloaded selectively if an investigator wished to combine characters from different submitted analyses.

The quality of the data submitted may vary greatly therefore other tables of information (Figure 2), such as the source of data, will be archived and linked to the image. Tables may even include information pointing to debates on the nature of particular characters. The vouchering of specimens for which images are submitted is highly recommended but may not be achievable in all cases. Ideally, an image should be linked to a repository number and taxon name tables. However, if the image is historical and this information is not available, it may nonetheless be included if it is useful and in such cases the image can be flagged as undocumented. A contemporary investigator may even document the same structure later and both images can be maintained. Navigation of the database by information in tables, such as any of those listed in Figure 2, should be fully possible. Key tables will be those that link labels to structures in the image, even labels that describe particular characters and character states. The database may require underlying hierarchical structure; one suggestion was the Linnaean hierarchy. Provisions may also be included to add information on such variables as body size, temperature tolerance and behavior because we consider this to be a repository for phenotypic data. Morphology is also dynamic, and changes during the course of ontogenetic development. It is important to plan to capture static images of developmental stages but to work ultimately towards the documentation of visually dynamic developmental information. Advances in digital imaging are an important breakthrough in the traditionally time consuming practices of collecting image data that are only beginning to be explored and implemented. Some new image capture methods (e.g., computed tomography) actually require computer-enhanced support that would be available through MorphoBank, underscoring one area where MorphoBank's capacity will exceed the capabilities of traditional printed documentation of characters.

One aspect of the final interface for viewers should be a cladistic matrix on the web for which every cell representing a morphological character is hyperlinked to at least one labeled image, and whenever possible, many images. Practitioners will need to be able to export to the web their desktop phylogenetics work, collected through such programs as Winclada, MacClade, and Nexus Data Editor. Software to facilitate this must be developed as part of the MorphoBank initiative.

MorphoBank will require an editorial or supervisory staff to oversee the concatenation of submissions. The database may also include text fields for dissent, commentary and discussion of homology hypotheses such that MorpoBank can accommodate the presentation of different homology interpretations by different contributors. This forum should enhance the precision of homology statements. Because morphological data lack the universal code inherent in molecular data, a glossary of character definitions may be required if miscommunications about homology statements, particularly among disparate taxa, are to be minimized. This may require separate tables of synonymy and even extensive translation of literature. A layer of peer review within MorphoBank may be required as numerous matrices are submitted and joined.

Levels of interpretation added to the images will make MorphoBank important to educational initiatives directed at nonspecialists who are interested in learning about organisms. Portals to more detailed information using such key words as 'head' or 'foot' could quickly lead a nonspecialist to more detailed data. The expert review of anatomical structures provided by MorphoBank should encourage nonspecialists to participate in the

collection of rare data. Full implementation of the educational possibilities will require a partnership between educators and MorphoBank developers and staff.

Simple prototypical programs for searching, character labeling, and image storage based on a matrix format for both the desktop and the web were presented at the workshop. These demonstrations revealed some of the immediate benefits of an interactive, image-enhanced, morphological matrix.

3. Impact of MorphoBank on the publishing world

The success of MorphoBank will depend on partnerships with the many scientific journals that publish morphological images. Unlike gene sequence data, published images are almost always subject to copyright of their publisher, and written permission is normally required for their reuse. Hence, it is essential that a mutually beneficial relationship with scientific publishers is established. For MorphoBank, the benefit is obvious: once agreement is reached that published images be reposited, MorphoBank's success and centrality in morphological research is virtually guaranteed. There are equally strong benefits for authors and publishers, however. A parallel practice by many molecular journals has been to require authors to submit nucleotide sequences to GenBank.

MorphoBank's emergence may be particularly timely for changes underway in the scientific publishing world. Just as the process of imaging is becoming nearly entirely digital, so too is the publication process. Financial limits are placed on publishers, and very restrictive limits on scope of illustration are placed on authors, by the costs of paper publication of high-quality images. These costs are greatly reduced in the digital realm, and virtually all journals are beginning the process of transition to digital publication. Digital publishing, however, does not eliminate costs. Many not-for-profit professional societies lack the technical expertise and resources for digital publishing, and as long as there is substantial pressure to retain paper publication, electronic publication becomes an added financial and technological burden. MorphoBank could effectively eliminate this burden by providing the technological infrastructure to store, organize, and disseminate images. Authors would benefit via a greatly expanded (indeed, effectively unlimited) scope of illustration. Paper publication is an onerous bottleneck on communication of morphological data and an investigator typically makes many more images than he or she is permitted to publish. These restrictions and limitations could disappear with the full implementation of MorphoBank.

A particular benefit from MorphoBank could be the resuscitation of large-scale systematic expertise via a paradigm for virtual monographs. Systematic study, both neontological and paleontological, is approaching a crisis point due to a long-term decline in specialist training. Paper monographs are difficult to publish due to page costs, and are very time-consuming often with limited immediate professional benefit to the investigator (i.e., they often result in a level of professional reward grossly disproportionate to the effort required). Virtual monographies could replace this situation with a dynamic model in which data can be used as they are assembled, work can be cited as parts of it are completed, and professional benefit is not deferred for years and undervalued. MorphoBank could reinvigorate the development of basic phenotypic data for systematics.

4. Technology

Because of the staggering task of developing or collecting images of so many organisms we discussed means of enhancing the investigator s efficiency in collecting documented images that could become part of MorphoBank. One suggestion was that any means of automating image collection (including the use of robots or artificial intelligence) to more quickly and cheaply generate images would be useful. Secondly, images need to be dynamically scalable for ease of use on the web. As noted above, the storage of so many images on the web as part of MorphoBank will require formulation of policies on copyright and intellectual property that will likely be modeled on some currently established systems in libraries and research networks for shared digital image and text information. These issues will have to be addressed as the project advances. International support, such as that behind GenBank, is also critical if this project is to succeed in its overall objectives. MorphoBank should also serve as an effective node that links with related databases such as GenBank, All Species Initiative, GIS databases, and Treebase, as well as digital library catalogues, virtual monographs, conservation databases and many others.

The steering committee should act: 1) as the force behind the emergence of MorphoBank (carrying out the recommendations below) and 2) as a direct link to the scientific community to promote the project and respond to questions about this initiative. One initial task for the steering committee is an examination of existing archives of morphological data, in particular, those on-line, to discover how these can be integrated into the overall MorphoBank effort.

5. Recommendations to NSF

We recommend:

1) that the National Science Foundation provide initial funds to structure the database, particularly funds for programmers working with practicing systematists to develop both the desktop and web software necessary for MorphoBank. Such software should facilitate beta testing with real datasets and by several investigators to develop the initial programs. The trial data sets should reflect different programming and databasing challenges.

2) that the National Science Foundation lead the initiative to involve other governmental agencies in the long-term endowment of MorphoBank in the same way GenBank has been supported. We recommend, for example, that NCBI should ultimately come to support MorphoBank as it supports many other centralized biological databases. MorphoBank must be envisioned and sustained as a vital and frequently upgraded system in order to build incentive for authors to contribute.

3) that the National Science Foundation encourage submission of data to MorphoBank by grantees whose research relates to comparative anatomy and other phenotypic applications in conservation, education, ecology and systematics.

4) that the National Science Foundation assist in the development of effective 3D digitizing techniques for a wide range of biological tissues, for manipulating these datasets, and for enhancing the integration of these technologies in education and research

5) that the National Science Foundation encourage computer science specialists to address challenges presented by the MorphoBank initiative

6. Activities postworkshop

A steering committee was identified to carry out the recommendations of the workshop (see list of participants below). Following the workshop the domain names morphobank.org, .com and .net were registered to serve as the virtual meeting ground for the development of early software ideas. These sites will hold the proceedings of the workshop and provide a log-in area for the steering committee to submit data in a trial fashion as the software designs for both the web and the desktop develop and improve.

Timetable: It is the hope of the workshop attendees that work on a prototype of MorphoBank can begin in early 2002 as an electronic extension of the report from the workshop. The steering committee should then identify funds to develop the software more fully, extending the few trial cases and in shaping applications for this funding. Should funding be granted, a three-year timetable for launching protypic software to organize morphological information on line such as is proposed here should be realistic.

MorphoBank Workshop

Participants
*Jonathan Adrain, University of Iowa
William Bemis, University of Massachusetts
Meredith Blackwell, Louisiana State University
*Mark Boguski, Johns Hopkins University
Janine Caira, University of Connecticut
Jim Carpenter, American Museum of Natural History
*Jonathan Coddington, Smithsonian Institution
Peter Crane, Kew Gardens
*Alfonso Delgado, Universidad Nacional Autonoma de Mexico
William Fink, University of Michigan
Kristian Fauchald, Smithsonian Institution
Darlene Judd, Oregon State University
*Seth Kaufman, Whirl-i-Gig
*Maureen Kearney, Field Museum
John Kress, Smithsonian Institution
*Denis Lynn, University of Guelph
*Paula Mabee, University of South Dakota
Lucinda McDade, Academy of Natural Sciences
Paula Mikkelsen, American Museum of Natural History
Kevin Nixon, Cornell University
*Michael Novacek, American Museum of Natural History
*Maureen O Leary, Stony Brook University
*Lynne Parenti, Smithsonian Institution
Norman Platnick, American Museum of Natural History
Kathlene Pryer, Duke University
*Tim Rowe, University of Texas
Petra Sierwald, Field Museum
Mark Siddall, American Museum of Natural History
John Van Couvering, American Museum of Natural History
Mike Whiting, Brigham Young University
Jim Woolley, Texas A & M University

Attending as observers from the National Science Foundation
H. Richard Lane, National Science Foundation
Larry Page, National Science Foundation
Quentin Wheeler, National Science Foundation

Workshop Assistants
Karen Claeson, Stony Brook University
Kirsten Jensen, University of Connecticut
Lynn Merrill, American Museum of Natural History

* = member of the postconference steering committee