WILEY-BLACKWELL

Cladistics 27 (2011) 1–9

Cladistics

10.1111/j.1096-0031.2011.00355.x

MorphoBank: phylophenomics in the ''cloud''

Maureen A. O'Leary[a,*] and Seth Kaufman[b]

[a]*Department of Anatomical Sciences, HSC T-8 (040), Stony Brook University, Stony Brook, NY 11794-8081, USA;*
[b]*Whirl-i-Gig, 220 5th Street, Greenport, NY 11944, USA*

Accepted 3 March 2011

Abstract

A highly interoperable informatics infrastructure rapidly emerged to handle genomic data used for phylogenetics and was instrumental in the growth of molecular systematics. Parallel growth in software and databases to address needs peculiar to phylophenomics has been relatively slow and fragmented. Systematists currently face the challenge that Earth may hold tens of millions of species (living and fossil) to be described and classified. Grappling with research on this scale has increasingly resulted in work by teams, many constructing large phenomic supermatrices. Until now, phylogeneticists have managed data in single-user, file-based desktop software wholly unsuitable for real-time, team-based collaborative work. Furthermore, phenomic data often differ from genomic data in readily lending themselves to media representation (e.g. 2D and 3D images, video, sound). Phenomic data are a growing component of phylogenetics, and thus teams require the ability to record homology hypotheses using media and to share and archive these data. Here we describe MorphoBank, a web application and database leveraging software as a service methodology compatible with ''cloud'' computing technology for the construction of matrices of phenomic data. In its tenth year, and fully available to the scientific community at-large since inception, MorphoBank enables interactive collaboration not possible with desktop software, permitting self-assembling teams to develop matrices, in real time, with linked media in a secure web environment. MorphoBank also provides any user with tools to build character and media ontologies (rule sets) within matrices, and to display these as directed acyclic graphs. These rule sets record the phylogenetic interrelatedness of characters (e.g. if X is absent, Y is inapplicable, or X–Z characters share a media view). MorphoBank has enabled an order of magnitude increase in phylophenomic data collection: a recent collaboration by more than 25 researchers has produced a database of > 4500 phenomic characters supported by > 10 000 media.

Reconstructing the tree of life (or parts of it) is the research focus of this journal and one of the most important challenges facing contemporary science (Cracraft, 2002; Wilson, 2004; Seife, 2005). Wilson (2004) emphasized that estimates of total species diversity vary widely but that there may be as many as 4–100 million living species on Earth, only a small percentage of which have been studied. To name and diagnose all of Earth's species, much less to embark on the detailed comparative phenomic and genomic research required for phylogenetics, is a major frontier for contemporary biologists and palaeontologists. For the small percentage of species described so far, only a fraction of descriptive information (mostly genomic) is currently archived in and shared via web-accessible databases.

Phylophenomics (phylogeny reconstruction incorporating data from phenotypes) has exhibited three critical trends in recent years. First, teams increasingly conduct the work (e.g. large-scale, NSF-sponsored Assembling the Tree of Life projects). Team members are also geographically dispersed, whether down the hall or across the world, they are not physically entering data into the same computer. Secondly, documentation of phenomic homology with media has come to enhance scientific communication and to increase the repeatability of phenomic observations made in the service of phylogeny reconstruction. Thirdly, investigators need rapid access to detailed electronic copies of published (legacy) comparative phenomic data sets, particularly matrices, in formats that can be instantly read by tree

*Corresponding author:
E-mail address: maureen.oleary@stonybrook.edu

search programs. Building supermatrices (de Queiroz and Gatesy, 2007) for robust tests of character congruence often requires adding new taxa and characters to published work. It is inefficient to recollect (or even retype) published phenomic matrix data from scratch, just as it is inefficient to re-sequence a gene previously used for molecular phylogenetics, unless there is reason to suspect error or a need to increase sampling. Legacy genomic data can be rapidly culled from GenBank, and alignments from Treebase (2010). Legacy phylophenomic data, however, often do not reside in any database, and if in Treebase, have no web-enabled tools to demonstrate homology.

Limited growth of phenomic matrices on the desktop

Historically, systematists studying comparative phenomics have organized their data in one of several landmark desktop data management programs [e.g. MacClade (Maddison and Maddison, 1992); Mesquite (Maddison and Maddison, 2005); Winclada (Nixon, 1999)]. Operational problems emerge, however, with the continued use of single-user, file-based software systems as teams assemble. Teams can work more efficiently if provided simultaneous access to a matrix changing in real time as it is being assembled by their collaborative group. Teams need: (i) Online space for commentary and discussion of homology. (ii) Ability to display, label and share media documenting hypotheses as they develop, and (iii) A place to archive matrices, taxonomy, character argumentation, and digital media, even when the last-named may consume large amounts of storage space.

No existing desktop software package delivers all of these services because in a desktop environment, each collaborator works on his or her own private version of project data. Changes made by one participant have no way of automatically propagating to others, preventing investigators from seeing a collaborator's data edits until changes are manually (and due to the effort involved, often only periodically) merged into a single "true" dataset. In all but the smallest and most disciplined of teams, file version control issues and the reconciliation of changes made on multiple copies of the data quickly emerge as significant drags on productivity. Lack of simultaneous viewing of a single dataset during data collection also impedes discussion of homology concepts.

Treebase (2010) and journal websites have been the primary archives for phenomic (and other) matrices, if matrices are electronically archived at all. Such storage is important but lacks dynamic viewing tools specialized for phenomic matrices, tools to link media and metadata to matrices. Furthermore, the user must reopen data from these sources in one of the desktop programs if he/she wishes to expand the matrix with new research. In the case of some journals (e.g. *Nature*) cladistic data on the journal website are often presented in electronic formats (e.g. pdf, Thewissen et al., 2007) that are not directly readable by any tree-searching program; this is a significant drawback for efficient expansion of phylophenomic research. Contemporary phylogenetic methods have been around long enough that most systematists must build on some prior study when embarking on new data collection.

When published matrices have been stored with affiliated media their data become even more useful to future projects. Many phenomic features can be described by multiple semantic expressions, thus accompanying labelled media often assist in clarifying terminological confusion. As the number of comparative observations expands into the thousands or tens of thousands of matrix cells, media, rather than words alone, often become essential for the repeatable and unambiguous documentation of homology. Given the complexity of phenotypes, no single type of media will describe all types of phenomic characters.

When sharing of labelled media is germane to clarification of phenomic homology, limitations of the desktop environment become even more acute. Phylogenetics software must then be versatile enough to support and display a variety of media. Although some existing desktop applications (e.g. Mesquite: Maddison and Maddison, 2005) do support referenced images, they rely upon keeping image files in a static location relative to matrix data (Ramirez et al., 2007). This arrangement is not only prone to failure (simply moving or renaming a directory may break the link between images and a matrix) but creates the inelegant side effect that all media must be physically transmitted with a matrix when it is shared. These problems simply do not exist if the team simultaneously accesses its project data in a networked repository. Even with the addition of version control systems [e.g. Subversion (Collabnet, 2006); Git (2011)], which have not, to date, been integrated into existing desktop programs, traditional desktop software cannot provide the ease of use of a combined database and web application environment. For high-resolution media including video or sound, each of which may document unique aspects of phenomic homology, a second problem develops, namely that individual media file sizes may swell to hundreds of megabytes. Serving such large files from a single online repository, rather than from desktop software, facilitates delivery to a team of project investigators and contributors, and, eventually, to the scientific community.

Finally, in recent years, ontology development has emerged in several taxonomic areas as a means of making concepts of organismic hierarchy explicit, machine readable, and linked to phylogenetic matrices (Mabee et al., 2007a,b; Ramirez et al., 2007; Dahdul et al., 2010; Vogt et al., 2010). Ontologies, as applied in

systematics, are sets of rules that describe relationships among concepts. Construction of such rules does not, however, currently transpire during a large majority of primary systematics research projects. Currently, phylophenomic research proceeds day to day more typically in what has been a called "free-text" format (Dahdul et al., 2010, p. 369), meaning character and state descriptions are not designed to be machine readable. As systematics moves towards a more structured vocabulary, ontologies are likely to emerge and grow in a variety of ways: top down, from organized teams of experts and bottom up from isolated studies. It is conceivable that there will be differences and disagreements about ontological rule building in practice. It may enhance broader ontology development in this transition period to put basic tools for certain kinds of rule building into the hands of any systematist collecting phenomic data. These might include tools available during primary assembly of character-by-taxon matrices that permit one to record basic dependency of one character on another (i.e. one character is "inapplicable" if another is scored "absent").

Web applications—software innovations that can aid phylophenomics

With the increasing sophistication of web-browser software, a new category of software has emerged, the "web application" (Chaffee, 2000). Web applications integrate existing web-based data services and enhanced user interface techniques into software that combines many of the best features of desktop applications with the collaborative capabilities of the Internet. A web application differs from a desktop application in that the former is accessed via a web browser without being installed directly on the user's computer. Investigators log on to a central server (or servers) using only the web browser software already installed on their computers (regardless of platform) to carry out application functions and access data.

Desktop software has historically been able to provide performance that the web could not; however, new web technologies, including Ajax (Garrett, 2005), are breakthroughs that permit nearly the same functionalities over the web as have existed on the desktop. Hundreds of web applications are in wide use for such purposes as personal information management, software helpdesks, games, and education, such as Backpack (Thirty-seven Signals, 2004), Google Docs (Google, 2005a), and Google Maps (Google, 2005b). Web applications can manage a single copy of a collaborative project's data that always reflects the current state of the project and that tracks changes made to it. This permits true team collaboration with an archived audit trail.

Web applications present several technical advantages over desktop software: they can be upgraded at a central source, they are relatively independent of any particular operating system, they have near-ubiquity, and, for virtually all modern operating systems, require no additional software to run beyond what comes with the computer's operating system by default. Development of web applications for systematics research represents an opportunity to permit collaboration by investigators located worldwide or down the hall, in real time.

MorphoBank: a web application and database moving phylophenomics into the "cloud"

Here we describe MorphoBank (O'Leary and Kaufman, 2007), a web application and database in its tenth year that allows users to build matrices linked to comparative phenomic data through a web browser interface. Within MorphoBank, users, and teams of users, can label and score media within phylogenetic matrices, specify relationships (rules) among characters, and download the matrices for analysis. All data are archived. Although the organization and presentation of phenomic matrices with media is the most complex task MorphoBank accomplishes, the site is also a key repository for more straightforward storage of phenomic media simply associated with metadata (e.g. taxonomy, collection), such as might be part of a specimen voucher or an expanded description from a publication. Data models laying the groundwork for this innovation have been discussed over the last decade (Nixon et al., 2001; O'Leary et al., 2001); however, MorphoBank is the first publicly available, functioning implementation of this idea.

Leaving the desktop

MorphoBank moves phenomic data matrix construction out of the relatively inefficient sphere of the desktop and onto the Internet. This shift can be described as taking phylophenomics into the realm of "cloud computing" (Knorr and Gruman, 2009) where software is provided as a service to users upon request. The "cloud" is a loosely defined term describing the practice of working in a network-based application hosted on opaque (and often geographically dispersed) server infrastructure. The user accesses programs from any internet-connected computer via a web browser as easily as he or she once accessed programs installed locally on his/her desktop. The user is also no longer concerned with how the application is implemented, run, or maintained. From the user's point of view the software is simply a service to be used. "Cloud" technology is widely in use in business and is often analogized to the electricity grid (Danielson, 2008).

MorphoBank might be most appropriately described as a "community cloud" (Wikipedia, 2010) because the software it provides for systematics is housed on university and museum-based servers (see Supporting information, Appendix S1) accessible to a specific research community and the public.

MorphoBank extends the matrix-editing capabilities found in single-user desktop programs: (i) It enables similar data management functions over the web for single or multiple users. (ii) It supplements these functions with shared, web-enabled discussion of characters and homology among collaborators, and manipulation of media affiliated with matrices, and (iii) It maintains matrices produced, and their associated media (see below), as an archived and publicly searchable product in a database after publication of the research (Appendix S1: Technical Description). Along with media, MorphoBank can record a variety of metadata such as author of submission, publications, critical commentary, names of species and higher taxa, and descriptions of characters and character states. MorphoBank ensures that all team members are always working with the same data, such that changes made by one member are instantly propagated to the entire team. All work is recorded in an event log, making it possible to determine how any element of a project's dataset arrived in its current state, and what work was directly contributed by each member of the project.

The distributed collaborative research MorphoBank supports occurs in a password-protected environment and can be done by autonomous, self-assembling teams of researchers building phylogenetic matrices with affiliated image data, or simply archiving images with annotations. Site architecture is documented in Appendix S2.

The web-based matrix editor

The centrepiece of the Morphobank web application, and its most complex software component, is a full-featured, multi-user matrix editor. The matrix editor displays dynamic phylogenetic matrices of phenomic characters with labelled, high-resolution, zoomable media documenting homology statements (Fig. 1) as character state exemplars or within cells. Collaboration features prominently in the matrix editor. Users can leave comments attached to cells, characters and character states for others to see. All changes are logged and tagged with the name of the editor, the date, time, and details of the change, documenting the history of the discussion of character argumentation and scoring. Editing access can be restricted both at the project level (e.g. a user may only edit characters but not score in all matrices) and at the taxon level (e.g. a user may only score cells for a particular row of taxa). These organizational tools facilitate the inclusion of contributors with different responsibilities (e.g. students and scientific assistants). Investigators can, for example, have full editing access to data while assistants and interns can be granted restricted access.

The matrix editor also provides tools to maintain data quality and to ensure that contributors do not "step on" each other. For example, one software feature tracks whether a user has scored a given character, and issues a warning if there has been a change to the character definition by other users since the score was made. If the

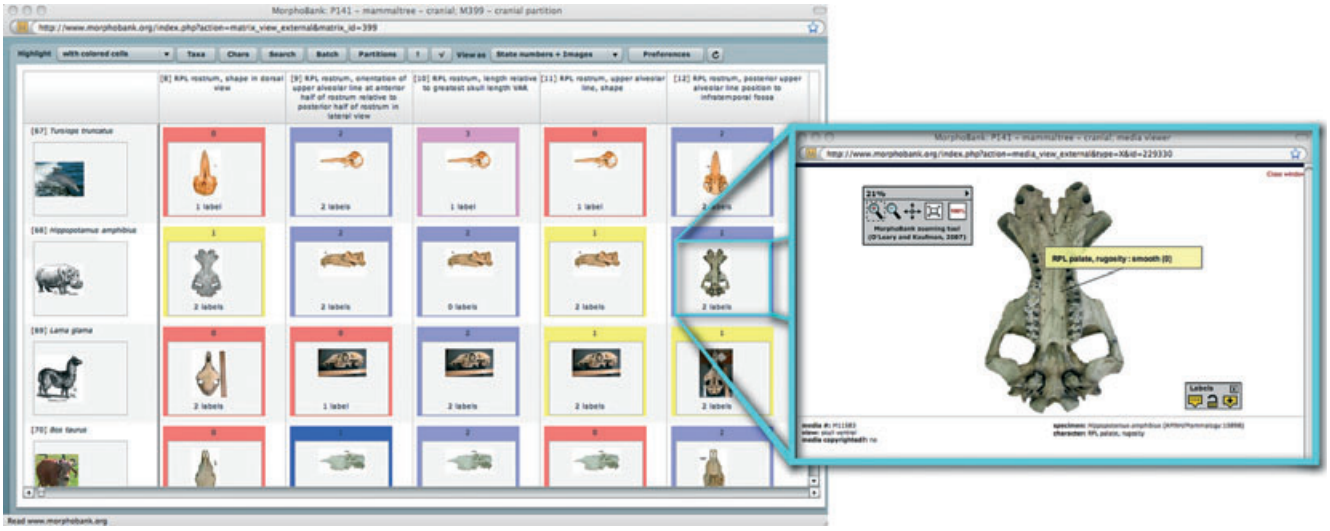

Fig. 1. Screen shots from an example project displayed in the MorphoBank matrix editor. Right side shows the MorphoBank zoom tool, which provides panning and zooming of cell media, as well as labelling of homology statements (intersection of character and character state). Data courtesy of the *NSF-Assembling the Tree of Life* project for Mammalia (http://mammaltree.informatics.sunysb.edu/).

character has been edited since the user's last score, an alert can be generated warning the user to check his or her scores against the character revisions.

Feedback gained from site use by large-scale projects [e.g. Assembling the Tree of Life for Mammals project (Novacek and the Mammal ATOL Team, 2008)] has inspired the development of several features to streamline typical workflows. These include batch scoring and batch media association tools, and a refined in-editor search tool. The matrix editor is continuously upgraded centrally to handle increasingly large (e.g. > 2000 characters and 300 taxa) phenomic data sets effectively as they develop.

Media to support phenomic homology

The program and site catalogue a range of media submissions (2D and 3D images, e.g. drawings, photos, and computer tomography scans), and allow contributors to place annotations on these media. MorphoBank was explicitly designed to enable researchers to use a range of visual media to describe phenomic characters, because unlike molecular sequence data, many phenomic homologies are elucidated by a range of media-rich descriptions, rather than by words and numbers alone.

MorphoBank implements a pan and zoom tool (Fig. 1) that gives users the ability to expand an image to examine it at close range. We have also shared this tool with other online digital initiatives (e.g. Morphbank). If the user wishes, MorphoBank automatically generates labels on media within a matrix cell, with the character name and state intersection. This association of text description with media greatly enhances the clarity of phenomic homology hypotheses being described, just as such a label would be a fundamental part of any anatomical atlas. Importantly, as the user pans and zooms, the MorphoBank label automatically resizes to remain readable at different magnifications.

Tools to build basic ontologies among phenomic characters

Aiming to be maximally useful to the large majority of systematists who work in free-text format, MorphoBank does not require character information to be structured in ontologies in order to be in the database, yet the software is compatible with character descriptions that adhere to an ontology. MorphoBank brings to the ontology revolution a toolkit that allows any user building a matrix to create a type of ontology that establishes rules relating characters.

Examples of such rules are that a user can designate that whenever character 1 is scored absent, characters 2–6 must be scored inapplicable. Similarly, one can create rules stipulating that media associated with one character should be automatically associated with other characters. In addition to capturing important information about the interdependency of characters, these relationships enable the automation of scoring actions according to the user-defined rule set. MorphoBank also implements a graphic tool that permits a user to draw a picture of the rules specified in the ontologies he or she designs. These pictures, known as directional acyclic graphs (Bang-Jensen and Gutin, 2008), or DAGs, are diagrams that specify the interrelatedness of different characters (Fig. 3; see also Dahdul et al., 2010; Fig. 2). Taken together these tools enable significant boost in productivity and reduce opportunities for scoring error.

Character argumentation that takes place during the construction of phylophenomic matrices is one of the key places where hypotheses of character dependencies emerge. Practising systematists building matrices bring extensive anatomical expertise to the establishment of such relationships. In this way, different research groups can contribute to aspects of ontology building, facilitating a grassroots approach to the development of concepts of organismic hierarchy. This bottom-up approach to organizing homology allows certain types of controlled vocabularies to originate within separate phylogenetic research projects. Rules that can be made in MorphoBank differ from those described, for example, in the Teleost Anatomical Ontology (TAO; Dahdul et al., 2010). The TAO describes relationships such as "part of" and is "made of", whereas MorphoBank describes (i) Character 1 is inapplicable if character 2 is absent, and (ii) If character 3 is described by certain media, so is character 7. These latter relationships directly impact phylogenetic hypothesis testing.

Examples in practice

Users have loaded over 45 000 media of both fossil and living species to MorphoBank associated with a variety of professional phylogenetic and biodiversity (neontological and palaeontological) publications. These media reside in private workspaces (Projects) until scientists choose to make their data available for use in scientific research and education, typically in conjunction with a peer-reviewed publication. MorphoBank is currently in version 2.7 with over 275 active projects and contributions from over 675 researchers internationally.

Non-matrix-based project examples in MorphoBank include video of the recently described basal land iguana (Gentile and Snell, 2009), and data such as Malchus's (2010) study of shell tubules in bivalves, which are stored in MorphoBank Folios, annotated, web-viewable, booklets of media and metadata. Boyer et al. (2011) have also used MorphoBank to store media contributing to specimen vouchers that form part of a DNA bar-coding project.

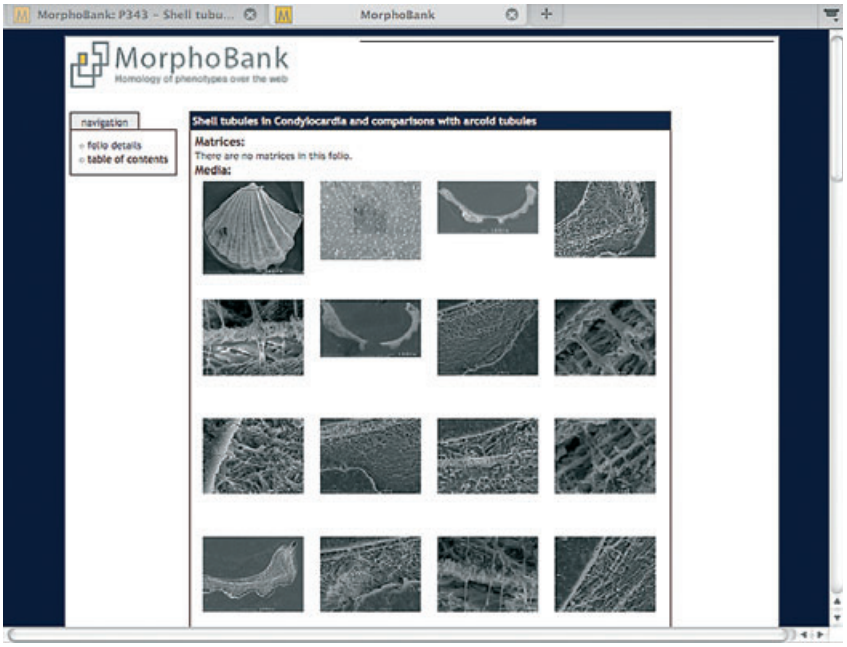

Fig. 2. Screen shot from a published MorphoBank Project (Malchus, 2010) showing an example of a MorphoBank folio, an online booklet of annotated media often used to accompany a journal article. All media can be zoomed and labelled with the character and character state intersection.

Expansion of phenomic characters for Mammalia by an order of magnitude

One of the major projects that has developed in step with the MorphoBank software has been the NSF-supported Assembling the Tree of Life project for Mammalia (http://mammaltree.informatics.sunysb.edu/). Key hypotheses of mammalian interrelationships, including the timing of origin of Placentalia and interordinal relationships (e.g. Wible et al., 2007), remain to be tested by simultaneous analysis of a supermatrix of phenomic and genomic data integrating fossil and living taxa. With this goal in mind, members of the Assembling the Tree of Life Project for Mammalia have assembled and organized over 4000 logically independent phenomic characters within MorphoBank. These come from a variety of systems—cranial, dental, postcranial, soft tissue, developmental, and behavioural (Novacek and the Mammal ATOL Team, 2008). All characters vary across Mammalia and were drawn from over 20 source matrices.

In the project, source matrices, redundancies and all, were concatenated in MorphoBank and the entire team was given access to one copy of the data. Characters were then debated, edited, reorganized, and refined to yield matrices ready for data collection by the team; matrices could also be viewed at any time by any member of the project wishing to monitor data collection. Images (over 20 000) have been associated with each cell in the matrix (and often labelled) as data were collected (Fig. 1). Anatomical synonyms have been placed in the Notes field of character descriptions. These synonyms are fully searchable, providing an important comparative anatomy resource for questions about terminology. An enormous bibliography of over 1000 citations is also part of the Project data on MorphoBank, providing depth and documentation for many homology statements and primary anatomical descriptions. Figure 3 shows a sample of the ontology that underlies the cranial partition from this collaboration. Work by the team in the matrix was checked against these rules to identify scoring errors. MorphoBank recorded work by each team member as they entered particular scores and tallied this work in a Project Details page.

Previous large-scale studies of phylophenomics have rarely exceeded several hundred characters. Working in MorphoBank has facilitated an order of magnitude increase in the number of phenomic characters to > 4000 that can be managed, scored, and studied. The rich documentation of homology has also enhanced the clarity and repeatability of these characters.

Discussion

DeSalle et al. (2005) recently reviewed why integration of phenomic and genomic data is key to both species delimitation and to discovery of the hierarchy of Life. Phenomes include a variety of levels of biological

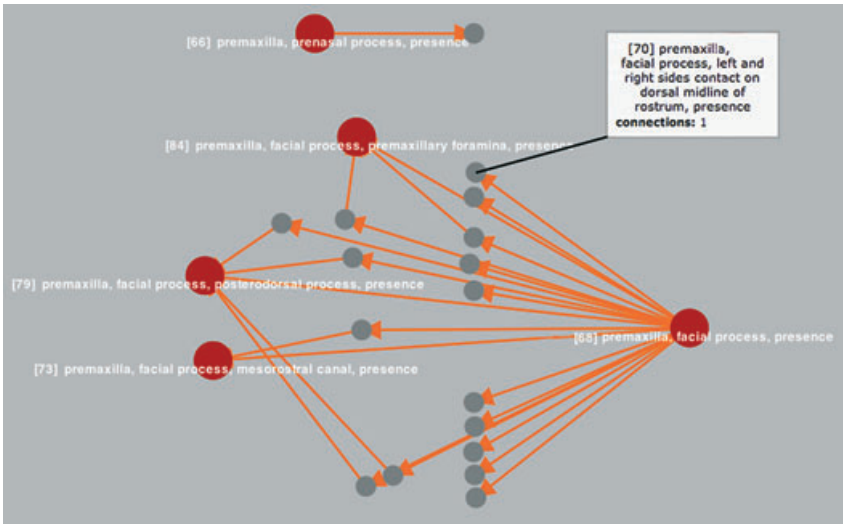

Fig. 3. Directed acyclic graph (DAG) generated within MorphoBank using its ontology tools. Graph describes rules that specify relationships among characters in a matrix. MorphoBank allows scientists to specify rules such as, if a designated parent character (red circles) is scored for a particular state (e.g., absent), that means that several other dependant characters (grey circles) are automatically inapplicable. This bookkeeping tool saves time, builds consistency in scoring, particularly in large team-based projects, and can be shown as graphic record of the hypothesized interdependency of characters. Data courtesy of the *Assembling the Tree of Life for Mammalia*.

organization such as biochemical reactions, developmental pathways, gene expression, histology, neuroscience, anatomy, and behaviour, any of which may yield character data for phylogenetic analysis. Extensive documentation of phenomic homology is necessary for understanding a variety of important evolutionary questions including the role of molecular biology in producing phenotypes. Although the amount of phenomic data collected for systematics is currently numerically smaller than the amount of molecular sequence data collected for the same purpose (understanding the historical branching pattern of species), those numbers may change in the future. The "amount of [phenomic] data has surpassed our ability to manage and use it" (Dettai et al., 2004, p. 822), demanding a new approach to data collection and data sharing.

The extensive informatics infrastructure that has supported molecular biology has been essential to a surge in sequence-based phylogenetics research seen over the last few decades. Phylogenetic research on phenomes rich in character data has lagged behind molecular work in scope, and arguments about the relative superiority of different classes of data have not considered this imbalance in infrastructure. Increasingly large supermatrices integrating phenomic and genomic data (e.g. Goloboff et al., 2010) are a key means of testing phylogenetic hypotheses (de Queiroz and Gatesy, 2007) and rely on the large-scale growth of phylophenomics.

The web application and online archive MorphoBank represents a significant component of the specialized cyberinfrastructure required to facilitate large-scale phylophenomic collaboration. The site moves data collection and analysis from individual desktops to servers that create a collaborative community "cloud". MorphoBank provides password-protected database services, automatic conversion of images to web-viewable form, advanced image viewing and annotation tools, online viewing and editing of phylogenetic matrix data, extensive storage of media describing homology, persistent URLs for hosted data, and rapid online publishing of data in association with the appearance of peer-reviewed research articles. With this tool, a team of systematists studying mammal phylogeny has been able to increase the number of phenomic characters under study by an order of magnitude. Without software of this kind, aggressive large-scale investigations of the impact of phenomic data on phylogenetic hypotheses cannot be seriously undertaken. Many phenomic homologies are best explained with media, creating a need for faster and simpler image sharing technologies in phylogenetics implemented in web applications. Such applications are fundamental to collaboration and data sharing, and to the expansion of phenomic data collection.

Phylogenetic data stored in MorphoBank provide a more immediate distribution of information about prior phenomic homology statements, and new taxa can be scored for those homology statements relatively quickly. This reduces currently repetitious work required to restudy previously published homology statements on original specimens (yet leaves the option open should such study be necessary). Over the long term, databasing and web application initiatives will lead to a much broader comparative database of phenomic data. This will be necessary if we are ever to investigate hypotheses

about relative levels of phenomic vs. genomic homoplasy, and correlation among phenomic and molecular characters. We have highlighted an example that happens to come from mammalian systematics, yet MorphoBank is not taxon specific. It currently supports a large number of invertebrate projects, other vertebrate projects, and plant projects on both extinct and extant taxa.

MorphoBank does not require scientists to frame their phylogenetic characters according to ontologies, which are not yet used in a large majority of phylophenomic projects appearing every year (see recent issues of *Systematic Biology* and *Cladistics*). Requiring users to implement ontologies would have the unwanted effect of barring from the database most of the highest quality contemporary systematics research. Nonetheless, MorphoBank's ontology tools allow teams to build rule sets about character interrelatedness across multiple taxa as they conduct phylophenomics research. This might be viewed as a grass-roots contribution to ontology construction that can contribute to other more comprehensive efforts.

MorphoBank welcomes and encourages investigators to deposit legacy data (i.e. matrices or media published without citing MorphoBank or published prior to its construction). Such matrices are often in demand for new projects expanding on phylogenetic research on the same taxon. Standards of good scholarship and ethics require that complete and readily reuseable files (i.e. digital, not printed, and in Nexus or TNT format) capturing the state of a phylogenetic dataset at the time of publication be archived for the sake of scientific repeatability. Ultimately, the growth of phenomic databases will need the collaboration of journal editors [see recent editorials, Berta and Barrett (2011), Fairbairn (2011)] who should require that publication coincide with deposition of data into public archives (which should not prevent data from also being stored in duplicate on journal or personal websites, if desired). Without policies at journals whereby databasing is a precursor to and part of the publication process, only a minority of authors may follow through. Databasing via GenBank has long been required by journals for molecular sequence data; treatment of phenomic data should meet no lower standard.

Acknowledgements

Several grants supported this research: NSF-DEB-9903964, NSF-DEB-0210956, NSF-EAR 0116517, NOAA NA04OAR4700191, and a grant from NESCENT. Comments from K. Claeson, M. Kearney, M. Miller, M.J. Novacek, M.E. Siddall, D. Stevenson, A. Wetterer, and two anonymous reviewers greatly improved this manuscript. Development of the software and site design benefited from and continues to grow with the contributions of K. Alphonse, L. Chan, N. Milbrodt, R. Monk, M. Passarotti, J. Salcedo, S. Villaverde, and A. Waller. We are grateful for ongoing systems support from M. Eisenberg, M. Purcell, R. Reeder, and S. Shrestha of Stony Brook University, and R. Choi and F. Lees of the American Museum of Natural History. We are enormously grateful to the MorphoBank user community as whole, and to the AToL Mammal Team, particularly N. Simmons and M. Weksler, for input and ideas on software development.

References

Bang-Jensen, J., Gutin, G., 2008. Digraphs: Theory, Algorithms, and Applications. Springer, London.

Berta, A., Barrett, P.M., 2011. Editorial. J. Vert. Paleontol. 31, 1.

Boyer, S., Howe, A.A., Juergens, N.W., Hove, M.C., 2011. A DNA barcoding approach to identifying juvenile freshwater mussels (Bivalvia, Unionidae) recovered from naturally-infested fishes. J. North Am. Benthol. Soc. 30, 182–194.

Chaffee, A. 2000. What is a web application (or "webapp")? http://www.jguru.com/faq/view.jsp?EID=129328. Accessed 8 October 2010.

Collabnet. 2006. Subversion. http://subversion.tigris.org./.

Cracraft, J., 2002. The seven great questions for systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. Ann. Mo. Bot. Gard. 89, 127–144.

Dahdul, W.M., Lundberg, J.G., Midford, P.E., Balhoff, J.P., Lapp, H., Vision, T., Haendel, M., Westerfield, M., Mabee, P.M., 2010. The teleost anatomy ontology: anatomical representation for the genomics age. Syst. Biol. 59, 369–383.

Danielson, K., 2008. Distinguishing cloud computing from utility computing. Ebiz. http://www.ebizq.net/blogs/saasweek/2008/03/distinguishing_cloud_computing/.

DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 360, 1905–1916.

Dettai, A., Bailly, N., Vignes-Lebbe, R., Lecointre, G., 2004. Metacanthomorpha: essay on a phylogeny-oriented database for morphology—the Acanthomorph (Teleostei) example. Syst. Biol. 53, 822–834.

Fairbairn, D.J., 2011. The advent of mandatory data archiving. Evolution 65, 1–2.

Garrett, J.J. 2005. Ajax: a new approach to web applications. http://www.adaptivepath.com/publications/essays/archives/000385.php. Adaptive Path.

Gentile, G., Snell, H., 2009. *Conolophus marthae* sp. nov. (Squamata, Iguanidae), a new species of land iguana from the Galápagos archipelago. Zootaxa 2201, 1–10.

Git., 2011. http://git-scm.com/.

Goloboff, P.A., Catalano, S.A., Mirande, J.M., Szumik, C.A., Arias, J.S., Källersjö, M., Farris, J.S., 2010. Phylogenetic analysis of 73,060 taxa corroborates major eukaryotic groups. Cladistics 25, 211–230.

Google. 2005a. Google Docs. http://docs.google.com/.

Google. 2005b. Google Maps. http://maps.google.com.

Knorr, E., Gruman, G. 2009. What cloud computing really means. InfoWorld. http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031?page=0,0.

Mabee, P.M., Arratia, G., Coburn, M., Haendel, M., Hilton, E.J., Lundberg, J.G., Mayden, R.L., Rios, N., Westerfield, M., 2007a. Connecting evolutionary morphology to genomics using ontologies: a case study from Cypriniformes including zebrafish. J. Exp. Zool. 308B, 1–14.

Mabee, P.M., Ashburner, M., Cronk, Q., Gkoutos, G.V., Haendel, M., Segerdell, E., Mungall, C., Westerfield, M., 2007b. Phenotype ontologies: the bridge between genomics and evolution. Trends Ecol. Evol. 22, 345–350.

Maddison, W.P., Maddison, D.R., 1992. MacClade: Analysis of Phylogeny and Character Evolution, ver. 3.04. Sinauer Associates, Sunderland, MA.

Maddison, W.P., Maddison, D.R., 2005. Mesquite: A Modular System for Evolutionary Analysis, Ver. 1.05 Build g24. http//mesquite project.org/mesquite/mesquite.html.

Malchus, N., 2010. Shell tubules in Condylocardiinae (Bivalvia: Carditoidea). J. Molluscan Stud. 76, 401–403.

Nixon, K. 1999. Winclada (BETA), ver. 0.9.9. Published by the author. http://www.cladistics.com/about_winc.htm.

Nixon, K.C., Carpenter, J.M., Borgardt, S.J., 2001. Beyond NEXUS: universal cladistics objects. Cladistics 17, S53–S59.

Novacek, M.J., and the Mammal ATOL Team. 2008. A team-based approach yields a new matrix of 4,500 morphological characters for mammalian phylogeny. J. Vert. Paleontol. 28(Suppl. 3), 121A.

O'Leary, M.A., Kaufman, S., 2007. MorphoBank 2.5: web application for morphological systematics and taxonomy. www.morphobank.org.

O'Leary, M.A., Caira, J., Novacek, M.J., 2001. MorphoBank: November 2001 workshop report.

de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.

Ramirez, M.J., Coddington, J.A., Maddison, W.P., Midford, P.E., Prendini, L., Miller, J., Griswald, C.E., Hormiga, G., Sierwald, P., Scharff, N., Benjamin, S.P., Wheeler, W., 2007. Linking of digital images to phylogenetic data matrices using a morphological ontology. Syst. Biol. 56, 1–12.

Seife, C., 2005. What is the universe made of? Science 309, 78–102.

Thewissen, J.G.M., Cooper, L.N., Clementz, M.T., Bajpai, S., Tiwari, B.N., 2007. Whales originated from aquatic artiodactyls in the Eocene epoch of India. Nature 450, 1190–1195.

Thirty-seven Signals. 2004. Backpack. http://www.backpackit.com/.

Treebase. 2010. Treebase. http://www.treebase.org/.

Vogt, L., Bartolomaeus, T., Giribet, G., 2010. The linguistic problem of morphology: structure versus homology and the standardization of morphological data. Cladistics 26, 301–325.

Wible, J.R., Rougier, G.W., Novacek, M.J., Asher, R.J., 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. Nature 447, 1003–1006.

Wikipedia. 2010. Clouding Computing, Wikipedia, the free encyclopedia. Available at http://en.wikipedia.org/wiki/Cloud_computing [Accessed August 2010].

Wilson, E.O., 2004. The meaning of biodiversity and the Tree of Life. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling the Tree of Life. Oxford University Press, New York, pp. 539–542.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. MorphoBank 2.7 technical description.

Appendix S2. MorphoBank 2.7 architecture.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix 1: MorphoBank 2.7 Technical Description

Database Structure

The MorphoBank web application is structured around the concept of "Projects," which consist of all matrices, metadata, and media in use by a scientist or team of scientists during a given research program (e.g., a published paper in a peer-reviewed journal). The application is implemented around a relational database that stores all aspects of hosted projects. Primary tables in the database are: *Taxonomic names*; *Characters* (e.g., phenomic descriptive phrases) with associated character states and ontology statements (e.g., rules); *Specimens* (including collection information, voucher number and taxonomic name, as well other Darwin Core 2 (Darwin Core, 2009) compatible fields, see below); *Media* (still imagery, video, sound); and *Matrices* (taxa and characters united by character state assertions). MorphoBank allows annotations on these entities. These entities are each contained within independent workspaces called "Projects". Naming collisions among Projects do not occur because each Project is self-contained. Once a Project is published, the site search engine can make connections among like-named entities in formerly separate projects.

MorphoBank supports the most widely used media formats: for still imagery, it stores the original uploaded file, JPEGs in several sizes, and a TilePic-format (Berkeley Digital Library Project, 2000) version, equivalent in

resolution to the original, used for the pan-and-zoom image viewer. General website functionality is implemented using accepted web standards and operates in any reasonably modern web-browser (e.g., currently Internet Explorer 7 and 8 for Windows, FireFox 3.5 or higher on any platform, Google Chrome 9.0 and Safari 4 or higher for Mac OS X). MorphoBank does not presently display molecular or amino acid sequence data in its interactive, graphic matrix editor specifically designed for discrete phenomic characters, because the database requires characters and character states to have names. Sequence data in various formats (Nexus, TNT, PDF, Excel) can, however, be scored in the Documents folder ("Docs" tab) associated with all Projects (e.g., Demere et al., 2008). Examples of such usage include Demere et al., (2008: Project 182) and Spaulding et al. (2009: Project 321).

Registration

Registered users can enter data and access MorphoBank via a password-protected login. When a scientist starts a new Project he/she is the Project Administrator and directly invites collaborators who then have password protected access that Project. Workgroup members may be designated full editing access, character annotator, bibliography maintainer or read-only "observer" access. It is possible to grant anonymous read-only access to reviewers of an unpublished project.

Media Uploading

Registered users can upload and catalogue multimedia submissions in a variety of file formats: 2D images in: JPEG, JPEG 2000, TIFF, PNG, GIF and Photoshop formats; 3D images or video in QuickTime, WindowsMedia, Flash Video [FLV], MPEG-2 and MPEG-4; audio in: MP-3, AIFF and WAV Formats. Support for other formats, such as vector formats, is planned; vector-based media can presently be uploaded by simply converting them to JPGs or TIFFs on the desktop first. All project group members can view all uploaded media. Project members can place annotations (labels) on these media, and can label them. Along with media, MorphoBank captures and records metadata such as author of submission, related publications, critical commentary, names of species and higher taxa, and descriptions of characters.

Matrix Loading

MorphoBank supports Nexus (Maddison et al., 1997) and TNT (Goloboff et al., 2004) formats for upload, display (i.e., the matrices can be displayed starting from zero or from one) and download. Information such as Nexus blocks supplying special commands implemented in a particular analysis are preserved in upload such that anyone downloading the matrix in the future will download the block intact. The site currently supports only discrete characters, however, plans

are underway to support continuous characters as such (e.g., Goloboff et al., 2006).

*Dynamic Character Editing and Matrix Creation***.**

For phylogenetic research, MorphoBank displays editable matrices of phenomic characters (homology statements). A new project can be initiated by copying a matrix, loading a matrix, or by creating a matrix within the site itself. Relevant media (e.g., 2D and 3D media, video, sound) can be associated with cells in matrices and labeled (Fig. 1). All data are private until authors release data as "published" on MorphoBank. Multiple authors can collaborate on a single matrix simultaneously over the web, anatomical structures can be labeled on images, and other types of metadata (e.g., notes) can be associated with media and cells. Once images have been made public, they can be downloaded from the web, with or without associated metadata.

Annotations

The usefulness of media to researchers is greatly enhanced by linking text to those media, which can be done in MorphoBank. Characters and cells are accompanied by a comment system that allows project members to debate homology by exchanging comments online. In MorphoBank, images and annotations are also stored separately and can be retrieved separately.

Information about an image (author, date, size, original format), as well as embedded descriptive and technical metadata in International Press Telecommunications (IPTC), Exchangeable Image (EXIF), and Extensible Media Platform formats (XMP), is captured at the time of upload.

Searching

MorphoBank's search engine is capable of searching all aspects of a Project's data and returning taxonomic records, specimen data, characters, media and matrices, any of which can be downloaded. The search engine implements Boolean operators, exclusion, wildcards, stemming, and parenthetical grouping. The search returns image and metadata from all projects that have been added to the MorphoBank public archive by investigators, as well as any unpublished projects to which a registered user is currently contributing. Public projects can also be browsed. Within a matrix several search tools also exist to identify quickly key words, empty cells and other isolated pieces of information

Hardware and Backup.

MorphoBank is distributed by Stony Brook University's Departments of Medical Informatics and Information Technology; it is also supported by offsite backup at the American Museum of Natural History. The primary MorphoBank

site resides on a DELL PowerEdge 2650 server with a 3.06GHz quadcore processor and 16GB of RAM and an 8tb RAID-5 disk sub-system. The American Museum backup system provides Sun Fibre Channel SAN- and SATA-attached disk. Stored information is always online, always shared and backed up nightly. Database services are provided, along with DBA and System administration expertise to run these systems.

References

Berkeley Digital Library Project, 2000. Tilepic code, http://wiki.collectiveaccess.org/index.php?title=TilePic. University of California, Berkeley.

DarwinCore, 2009. Biodiversity Information Standards TDWG, http://rs.tdwg.org/dwc/index.htm.

Demere, T. A., McGowen, M. R., Berta, A., Gatesy, J., 2008. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. Syst. Biol. 57, 15-37.

Goloboff, P. A., Farris, J. S., Nixon, K., 2004. TNT, a free program for phylogenetic analysis. Version 1.1. Cladistics 24, 774-786.

Goloboff, P. A., Mattoni, C. I., Quinteros, A. S., 2006. Continuous characters analyzed as such. Cladistics 22, 589-601.

Maddison, D. R., Swofford, D. L., Maddison, W. P., 1997. Nexus: and extensible file format for systematic information. Syst. Biol. 46, 590-621.

Spaulding, M., O'Leary, M. A., Gatesy, J., 2009. Relationships of Cetacea

(Artiodactyla) among mammals: increased taxon sampling alters interpretations of key fossils and character evolution. PloS One 4, 1-14.

Appendix 2: MorphoBank 2.7 Architecture

The MorphoBank web application conforms to a standard 3-tier architecture model. The site's view and logic layers are implemented in the PHP programming language (Holmes, 2007). This choice means that we have adopted a lightweight environment that emphasizes a maximally functional aesthetic. PHP, a mature, stable language, provides many useful features for data access and sharing, and avoids the large overhead required by enterprise web applications built on platforms (e.g., Java Struts/Enterprise Java Beans [Burke and Monson-Haefel, 2006]).

General web site functionality is implemented using HTML/Javascript pages accessible via standard web browsers. Highly interactive and data intensive user interfaces, such as the matrix editor (Fig. 1) and image viewer/annotation tool, however, are implemented as "Rich Internet Applications" developed in the Adobe Flex framework and compiled to SWF (Shock Wave Flash) format executables that run within the Adobe Flash Player browser plug-in. The near ubiquitous Flash plug-in is currently available to the vast majority of Internet users on Windows, Mac OS X and Linux. Use of Flex/SWF makes possible fast-loading and responsive cross-platform tools with complex user interfaces that are tightly integrated with the surrounding HTML-based web-application. All such applications communicate with the database server via AMF-based web services. AMF is an open binary protocol developed by Adobe that is optimized for bandwidth and performance.

The database tier of MorphoBank consists of a relational database for project data. Media (e.g., images, video, sound) are stored in controlled, web-accessible file systems. The relational database contains only metadata extracted from the media and the information necessary to link to images on web-accessible file systems. A simple object-relational mapping system, based upon the Active Record pattern links the physical model of the relational database to the application's runtime object model. Mapping is managed by libraries making it feasible to extend the application data model without wholesale redesign of database schema.

The relational database is implemented in MySQL (MySQL, 2010) a widely deployed, open source, relational database management system that uses Structured Query Language [SQL]. The relational database stores all metadata associated with hosted projects. The primary tables in the database are listed above. These entities are contained within independent Project workspaces, preventing naming collisions between projects.

MorphoBank architecture includes a robust facility for media handling. A plug-in interface enables support for new media formats to be added without modification of the MorphoBank core. Currently, open-source media handling libraries are used to process and convert various media formats: 1) ImageMagick (ImageMagicStudioLLC, 1999-2011) image processing library provides handling of over 90 image formats including TIFF, JPEG, GIF, JPEG2000, and PNG; 2) ffmpeg handles various video and audio file formats and compression schemes, including QuickTime, WindowsMedia, MPEG-2, MPEG-4, MP3, AIFF, WAV and

FLV; and 3) built-in libraries parse SWF multimedia, QuickTime VR panoramic imagery and Tilepic format imagery. MorphoBank automatically creates derivative versions of media files at selected resolutions, including a Tilepic format version for use with the pan-and-zoom image viewer. The original uploaded file is retained as well. When processing of an incoming media file is expected to take more than a few seconds, MorphoBank automatically queues jobs for background processing avoiding a long wait for the user. The background processing queue is designed for distributed operation on multiple servers, for maximum scalability.

The MorphoBank database incorporates structures for various domain specific entities as well as a configurable metadata schema that supports mapping to widely used schemas like Darwin Core 2 (Darwin Core, 2009) and Dublin Core (Dublin Core Metadata Initiative, 1999-2011). Entities such as taxonomic names, characters and matrices that are invariant across all use cases are implemented using domain-specific entities. Other types of data are stored as project-specific metadata with an explicit mapping to a project-defined standard.

References

Burke, B., Monson-Haefel, R., 2006. Enterprise Java Beans 3.0 5th Edition. O'Reilly, Cambridge.

DarwinCore, 2009. Biodiversity Information Standards TDWG, http://rs.tdwg.org/dwc/index.htm.

Dublin Core Metadata Initiative, 1999-2007. http://dublincore.org/ .

Holmes, J., 2007. Struts: the complete reference. McGraw Hill, New York.

MySql, 2010. http://www.mysql.com/?bydis_dis_index=1.