Towards a Taxonomically Intelligent Phylogenetic Database

Roderic D. M. Page
Division of Environmental and Evolutionary Biology
Institute of Biomedical and Life Sciences
University of Glasgow
Glasgow, G12 8QQ, Scotland
r.page@bio.gla.ac.uk [*]

Abstract

This note outlines some of the key intellectual obstacles that stand in the way of creating a usable phylogenetic database. These challenges include the need to accommodate multiple taxonomic names and classifications, and the need for tools to query trees in biologically meaningful ways. Until these problems are addressed, and a taxonomically intelligent phylogenetic database created, much of our phylogenetic knowledge will languish in the pages of journals.

1. Introduction

The last decade has seen an explosion in biological data, and much of the success of the bioinformatics community has derived from having ready access to that data. Databases such as GenBank and EMBL have themselves spawned a growing number of secondary databases tailored to specific questions (such a protein and RNA structure, gene families, metabolic pathways, polymorphisms, whole genomes, etc.). Integrating these diverse databases has become a major challenge for the bioinformatics community [18]. Ontologies (controlled vocabularies) [2] and web services are at the heart of projects such as BioMoby (biomoby.org) and MyGrid (www.mygrid.org.uk), which aim to provide descriptions of bioinformatics services and how to invoke them [18]. The Science Environment for Ecological Knowledge (SEEK) project has similar ambitions for ecological data (EcoGrid) (seek.ecoinformatics.org).

The tree-building community has been noticeably absent from most of these developments. Although the phylogenetists have made considerable strides in the development of sophisticated methods of phylogenetic inference, much of the community's output of data sets and trees is languishing in the pages of journals, rather than in openly accessible databases. Consequently, the rapid growth of published evolutionary trees is not matched by the availability of those trees in databases (Fig. 1). These trees and their supporting data form a tremendous resource for biologists, with applications in genomics, evolutionary biology, parasitology, biodiversity, and public health [4]. The full potential of this resource will only be realised when the phylogenetic community makes the results of their work more easily available.

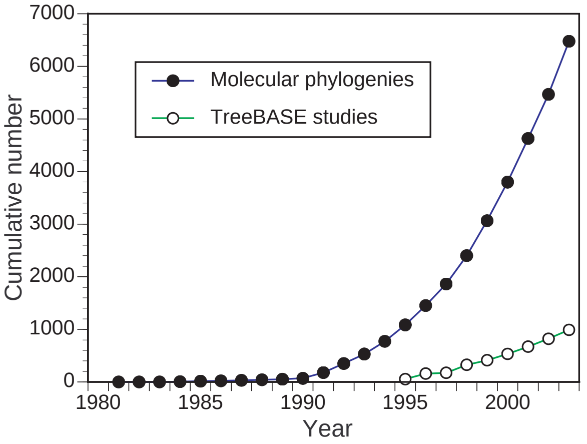

Figure 1: Cumulative growth of publications on phylogenetics, based on the number of papers found in the Web of Science by searching on the key words “molecular” and “phylogenetic” since 1981 [10], updated to 2003 and compared with the growth of the phylogenetic repository TreeBASE, which launched in 1996 (a study in TreeBASE is equivalent to a single paper).

Central to creating a useful phylogenetic database is having ontologies for taxonomic names and characters, so that meaningful queries can be constructed. Because trees are complex structures (when compared, say, to the string of letters representing a DNA sequence), we also need appropriate ways to ask biologically interesting questions on trees. This note explores some of these issues.

[*] Technical Reports in Taxonomy 04-01. Paper to be presented at the Workshop on Database Issues in Biological Databases (DBiBD), National e-Science Centre, Edinburgh, Scotland, January 8-9, 2005

2 The problem

There are several reasons why phylogenetic databases are languishing behind sequence databases. Most journals make submission of sequences to public databases a prerequisite for publication, but very few impose a similar requirement for trees and alignments (mycological journals are a notable exception). There are also issues of computational infrastructure, staff, and long term funding that need to be addressed. However, my goal here is to outline some of the key intellectual obstacles that stand in the way of creating a usable database [6, 9, 8]. These challenges include: (i) the lack of consistent taxonomic names; (ii) the absence of standardised character names; (iii) the dearth of tools for querying trees, and (iv) the lack of tools for synthesising trees and datasets for large-scale analysis.

2.1 Where we are now: TreeBASE

The only viable phylogeny database currently available is TreeBASE [12], which is housed at the University of Buffalo, New York. TreeBASE contains nearly 1000 studies covering 39,000 taxa. In addition to a simple query interface, TreeBASE provides some simple analytical tools, including supertree construction using the modified mincut supertree algorithm [7] hosted on a machine at Glasgow. TreeBASE has also served as a test bed for a number of theoretical developments in tree querying [15, 13, 20]. However, TreeBASE is greatly weakened by lacking a taxonomic framework.

2.2 Why taxonomy matters

Taxon names in a phylogenetic database should meet four criteria: (i) internal consistency; (ii) external consistency; (iii) synonymy, and (iv) hierarchy (10). The first criterion (internal consistency) is an obvious requirement. If multiple names are used for the same taxon, then a simple search for all data relevant to a given taxon cannot be guaranteed to have found all those data – some might be associated with an alternative name for that taxon.

The second criterion of external consistency assumes that we want to be able to apply knowledge obtained from the phylogenetic database to other domains. For example, a user wanting to use phylogenetic methods to analyse the evolutionary ecology of a group of organisms should be able to use the same scientific name to obtain phylogenetic and ecological data (for example through EcoGrid). This task is complicated because the same taxon may have multiple names (synonyms). A phylogeny database should be able to translate between different names for the same taxon.

The final criterion of hierarchy is equivalent to requiring an ontology that specifies the relationships between terms.

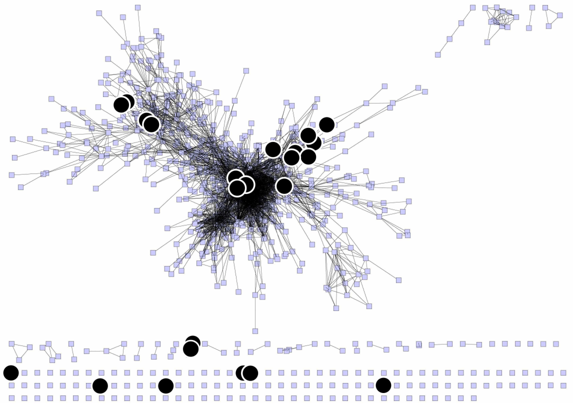

Figure 2: Graph of treespace in TreeBASE. Each node represents a study, and a pair of studies is connected by an edge if the corresponding studies have at least one taxon in common. Note that the graph is not connected. Studies containing birds are indicated by •.

For example, as text strings, “Gallus gallus” and “Struthio camelus” have no obvious connection, but both are names of birds (class Aves). If we query a phylogenetic database using the term “Aves”, we should be able to retrieve all studies containing a bird, regardless of whether those studies actually contain a taxon labelled “Aves.”

2.3 TreeBASE and taxonomy

TreeBASE has no mechanism for ensuring consistency of names, nor does it have an ontology of taxonomic names. These criteria are difficult to satisfy, and were especially so at the time TreeBASE was designed (it went live in 1996). TreeBASE tries to circumvent the need for a taxonomic hierarchy using the notion of “tree surfing. The user is presented with a tree, and can “surf to neighbouring trees that share at least one taxon in common with the original tree, in the processes finding all the taxa of interest. However, for tree surfing to work tree space in TreeBASE must be connected, which it is not [11] (Fig. 2). Furthermore, even if it were connected, there is no guarantee that all members of a taxon will be contained in a set of neighbouring trees. In the case of birds, the 24 studies in TreeBASE that contain one or more birds lie in different components of the tree space graph (Fig. 2), and so tree surfing by itself wont find all bird phylogenies in TreeBASE.

3 Matching taxon names

3.1 Simple matching

A naive approach is to match names in TreeBASE to names in an external database, such as the NCBI taxon-

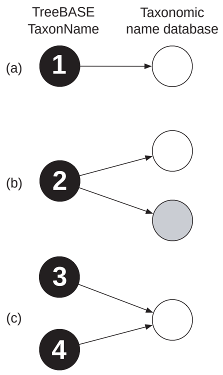

Figure 3: Examples of matching names in TreeBASE to TaxonNames in one or more taxonomic name databases. In (*a*) the TaxonName matches a name in an external database; in (*b*) the TaxonName occurs in two different databases. (*c*) shows a case where two different TaxonNames match the same name in a taxonomic database (for example, the TaxonNames comprise a taxon name concatenated with a specimen number).

omy (www.ncbi.nlm.nih.gov) used in GenBank. Preliminary work shows that only around 45% of TreeBASE names have an exact match to names in the NCBI taxonomy. Because phylogenetic data come sources other than nucleotide sequences, other taxonomic databases must be used in addition to GenBank, such as ITIS (www.itis.usda.gov), Species 2000 (www.sp2000.org), uBio (www.ubio.org), IPNI (www.ipni.org), and Mammal Species of the World (www.nmnh.si.edu/msw).

Some names in TreeBASE are spelt differently to names in external databases, are informal names, or comprise scientific names with voucher codes or GenBank accession numbers tacked on the end. Hence, names will also have to be matched using techniques such as approximate string matching [21]. Where the name includes an accession number, this can be used to extract the organism name from GenBank.

We can model the problem of matching TreeBASE names to taxonomic names using a bipartite graph, where the nodes are partitioned into two disjoint sets, one representing names in TreeBASE, the other representing names in taxonomic databases (Fig. 3). The components of the resulting graph correspond to the sets of names that are equivalent. For example, in this figure, the four TreeBASE TaxonNames belong in three components: 1, 2, and 3,4.

3.2 Synonyms

The mapping shown in Fig. **??** becomes more complicated once we consider taxonomic synonyms and how they are stored in different databases. For example, the harp seal is known by two different scientific names, *Phoca groenlandica* Erxlebben, 1777, and *Pagophilus groenlandicus* (Erxleben, 1777). These names are nomenclatural synonyms: they refer to the same species, but place it in two different genera. TreeBASE has both names and treats them as different taxa, as does GenBank; however ITIS correctly recognises that the two names are synonyms (Fig. **??**).

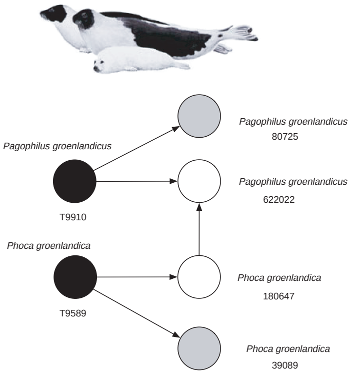

Figure 4: Graph depicting the relationships between different names for the harp seal. TreeBASE (●) and GenBank (○) contain two different scientific names for this seal, and different data is associated with each name. However ITIS (○) correctly links *Phoca groenlandica* and *Pagophilus groenlandicus* as synonyms.

Given cases like that in Fig. 4, we can no longer treat our matching graph as bivariate (cf. Fig. 3), although the components of that matching graph will still define sets of equivalent names. However, the issue then arises – "what name to use for the harp seal? For internal use we can simply assign a unique numerical identifier to each component, but if the user wants to view information on a particular taxon (or download that information into external software for further analysis), they will need an "accepted name for each component. We could represent the graph shown in Fig. 4 using an ontology language such as RDF, and develop a set of logical rules by which a reasoner (a program that explores logical relationships) could compute an accepted name for external use.

3.3 BLASTing TreeBASE

For those TreeBASE taxa with sequence data, an alternative approach to matching names is to look up that sequence in GenBank and compare the name of the corresponding taxon in GenBank with that in TreeBASE. This involves using BLAST [1] to find the best hit for each sequence in TreeBASE. Care needs to be taken in the BLAST search because some sequences in TreeBASE have not been deposited in GenBank, but will still return hits if GenBank contains sequences from closely related taxa. Furthermore,

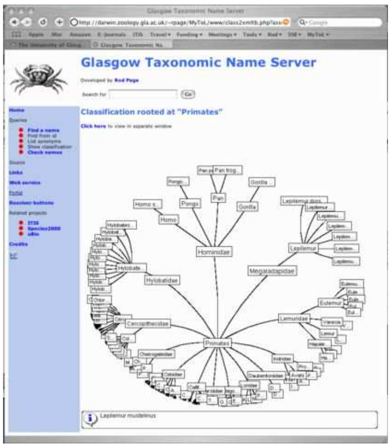

Figure 5: Screenshot of the Glasgow Name Server (darwin.zoology.gla.ac.uk/ rpage/MyToL/www) displaying a classification.

some TreeBASE datasets comprise concatenated genes, and so only one of those genes will be represented in the top hit.

3.4 Taxonomy trees

Matching names is not enough to provide an intelligent taxonomic interface to TreeBASE. A user searching TreeBASE for bird phylogenies using the term “Aves would find 4 studies, yet there are 24 bird studies in TreeBASE. Given that tree surfing also fails to find all these studies (Fig. 2), we need a different approach. Given a classification of all taxon names in TreeBASE we would be able to compute the taxonomic content of any name (for example, we would be able to generate the list of all birds in TreeBASE).

A complication is that different taxonomic databases employ different classifications. I have developed a pilot database (the Glasgow Name Server) for storing, querying and displaying classifications [8] using techniques described by [3] (Fig. 5).

We need to explore different approaches to the problem of combining multiple classifications, such as finding the agreement between two classifications, or merging two or more classifications into a consensus. These problems are related to unordered subtree inclusion [19] and supertree construction with internally labelled nodes [14], respectively. Alternatively, perhaps multiple classifications could be merged into a single graph that is no longer a tree.

4 Querying trees

There are two basic kinds of tree-based queries relevant to phylogenetic databases: pattern matching, and finding trees that resemble another tree. Tree pattern matching techniques include ATreeGrep [13], and the related TreeSearcher tool developed at Glasgow (darwin.zoology.gla.ac.uk/ rpage/TreeSearcher/).

To date all tree pattern matching queries have used existing labels in trees, that is, the queries make no use of taxonomy. As a consequence, a simple query such as “find all trees that have birds and crocodiles as sister taxa is difficult to formulate, unless the user knows the names of all the birds and crocodilians in the database, and constructs a suitable query tree (which may have hundreds of nodes). To avoid this problem, we need to develop a query rewriting mechanism that uses a taxonomic classification and a mapping of tree labels to taxon names o rewrite the query. Hence, the user could enter a query such as find phylogenies matching the tree ((Aves,,Crocodilia),Testudines), and get a meaningful answer, even if each tree in the database is labelled only with species names.

4.1 Finding trees that resemble a query tree

Given a tree, such as one in TreeBASE, a natural question to ask is “are there any other trees that look like this? To answer this, we need a tree comparison measure, and a means of quickly comparing our tree with the other trees in TreeBASE. There are numerous tree comparison metrics available [17], some of which have been developed with searching tree databases in mind [20].

The simple approach to finding similar trees is to do a linear search, that is, compare the query tree with every tree in the database. In practice this approach may perform reasonably well, especially if the database is small and we can filter out lots of trees (for example, those that have no taxa in common with the query tree). However, as the database grows this method is unlikely to scale.

Another approach is to define a metric on tree space, and use a metric-space index to quickly find similar trees. Metric-space indices are a very attractive means of searching complex data types, such as sequences, images, and trees [5]. Although there is an extensive literature on tree comparison measures, there are no published metrics for trees where the trees have different, possibly overlapping leaves. However, we can use a symmetric difference as a starting point. Given two trees, T_1 and T_2, then

$$d(T_1, T_2) = |T_1| + |T_2| - 2|T_1 \cap T_2|$$

where $|T_1|$ is the number of leaves in the tree T_1. We can normalise this by dividing it by $|T_1| + |T2|$, so that it ranges from 0 to 1. This measure regards trees as similar if they have taxa in common, but makes no use of tree topology. We can incorporate tree topology using maximum agreement substrees (MAST) [16]

$$d(T_1, T_2) = |T_1| + |T_2| - 2 \times \mathrm{MAST}(T_1, T_2)$$

where $\mathrm{MAST}(T_1, T_2)$ is the number of taxa in the maximum agreement subtree.

Implementations of metric indices exist (e.g., the C++ library M-tree library: www-db.deis.unibo.it/Mtree/), and it would be interesting to evaluate the performance of this approach to searching for trees.

References

[1] S.F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] J.B.L. Bard and S.Y. Ree. Ontologies in biology: design, applications, and future challenges. *Nature Reviews Genetics*, 5:213–222, 2004.

[3] J. Celko. *Joe Celko's Trees and Hierarchies in SQL for Smarties*. Morgan Kaufmann, San Francisco, 2004.

[4] J. Cracraft, M.J. Donoghue, J. Dragoo, D. Hillis, and T. Yates. *Assembling the Tree of Life: Harnessing Life's History to Benefit Science and Society*. American Museum of Natural History, 2003.

[5] D. Miranker. Metric-space indexes as a basis for scalable biological databases. *Omics: A Journal of Integrative Biology*, 7:57–60, 2003.

[6] L. Nakhleh, D. Miranker, F. Barbancon, W.H. Piel, and M.J. Donoghue. Requirements of phylogenetic databases. In *Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)*, pages 141–148, 2003.

[7] R.D.M. Page. Modified mincut supertrees. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics (WABI2002) (Lecture Notes in Computer Science 2452)*, Rome, Italy, 2002.

[8] R.D.M. Page. Phyloinformatics: Towards a phylogenetic database. In J. T. L. Wang, M.J. Zaki, H.T.T. Toivonen, and D. Shasha, editors, *Data Mining in Bioinformatics Data Mining in Bioinformatics*, Advanced Information and Knowledge Processing, pages 219–241. Springer-Verlag, 2004.

[9] R.D.M. Page. Taxonomy, supertrees, and the tree of life. In O. Bininda-Emonds, editor, *Phylogenetic supertrees: Combining information to reveal the Tree of Life*, volume 4 of *Computational Biology*. Kluwer Academic Publishers, 2004.

[10] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401:877–884, 1999.

[11] W. Piel, M.J. Sanderson, and M.J. Donoghue. The small-world dynamics of tree networks and data mining in phyloinformatics. *Bioinformatics*, 19:1162–1168, 2003.

[12] W.H. Piel, M.J. Donoghue, and M.J. Sanderson. TreeBASE: a database of phylogenetic knowledge. In *To the Interoperable Catalogue of Life with partners – Species 2000 Asia Oceania – Proceedings of 2nd International Workshop of Species 2000 (Research Report for the National Institute of Environmental Studies, R-171-2002)*, Tsukuba, Japan, 2002. http://www.nies.go.jp/kanko/kenkyu/pdf/r-171-2002.pdf.

[13] D. Sasha, J.T.L. Wang, H. Shan, and K. Zhang. Atreegrep: approximate searching in unordered trees. In J. Kennedy, editor, *Scientific and Statistical Database Management (SSDBM 2002)*, pages 89–98, 2002.

[14] C. Semple, P. Daniel, W. Hordijk, R.D.M. Page, and M. Steel. Supertree algorithms for ancestral divergence dates and nested taxa. *Bioinformatics*, in press.

[15] H. Shan, K.G. Herbert, W. H. Piel, D. Sasha, and J.T.L. Wang. A structure-based search engine for phylogenetic databases. In *14th International Conference on Scientific and Statistical Database Management (SSDBM'02)*, pages 7–10, 2002.

[16] M. Steel and T. Warnow. Kaikoura tree theorems: computing the maximum agreement subtree. *Information Processing Letters*, 48:77–82, 1993.

[17] M.A. Steel and D. Penny. Distributions of tree comparison metrics – some new results. *Systematic Biology*, 42:126–141, 1993.

[18] Lincoln Stein. Integrating biological databases. *Nature Reviews Genetics*, 4:337–345, 2003.

[19] G. Valiente. *Algorithms on trees and graphs*. Springer-Verlag, Heidelberg, 2002.

[20] J.T.L. Wang, H. Shan, and D. Sasha abd W.H. Piel. Treerank: a similarity measure for nearest neighbor searching in phylogenetic databases. In *Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003)*, pages 171–180, 2003.

[21] S. Wu and U. Manber. Agrep – a fast approximate pattern-matching tool. In *Proceedings USENIX Winter 1992 Technical Conference*, pages 153–162, San Francisco, CA, 1992.