TreeBASE: A database of phylogenetic information.

William H. Piel[1], Michael Donoghue[2], Mike Sanderson[3]) (Contact Person)
[1]*Institute of Evolutionary and Ecological Sciences, Leiden University 2311 GP Netherlands, and Naturalis Nationaal NatuurHistorisch Museum, Leiden Netherlands.*
[2]*Harvard University Herbaria, 22 Divinity Avenue, Cambridge Massachusetts 02138, USA*
[3]*Section of Evolution and Ecology, University of California, Davis, California 95616, USA*

Abstracts

Phylogenetic systematics brings added value to the species lists and taxonomic inventories that form the groundwork of our understanding of biodiversity. But this added value is lost without a central database to store what we know about historical patterns and evolutionary relationships. TreeBASE was developed to harness this information and to provide a tool to study the evolution of biodiversity. Access to phylogenetic trees, and to the data underlying them, is needed for a wide variety of purposes, including comparative studies of morphological and molecular evolution, biogeography, coevolution, and studies of congruence of results based on different sources of evidence. Such data are also needed to monitor progress in phylogenetic research, to test new methods of analysis, and to address immediate practical problems in conservation of biodiversity. TreeBASE stores published phylogenetic trees, character and molecular data matrices, bibliographic information, and some details on taxa, characters, algorithms used, and analyses performed. The database is designed to be explored interactively and to allow retrieval and recombination of trees and data from different studies. TreeBASE therefore provides a means of assessing and synthesizing knowledge of phylogenetic information and biodiversity. The URL is: *http://phylogeny.harvard.edu/treebase*.

Key Words: TreeBASE, database, phylogeny

Introduction

In the early 1980's, the advent of personal computers and PCR techniques together precipitated an explosion of phylogenetic knowledge. Personal computers made it easier, faster, and more affordable for biologists to analyze phylogenetic data, and DNA sequencing made large amounts of raw phylogenetic data available to the non-specialist. By 1989, publications of phylogenetic trees were growing by 15 to 20% per year without any sign of abatement (Pagel, 1997, Sanderson, et al., 1993). In addition, trees were increasingly used to answer questions outside of systematics, such as in coevolution and adaptation (Coddington, 1988, Mitter and Brooks, 1983, Mitter, et al., 1988, Mitter, et al., 1991).

For reasons similar to those that drove the molecular biology community to develop DNA databases, Sanderson et al. proposed that the systematic community develop a phylogenetic database to harness this rapidly-growing field (Sanderson, et al., 1993). In addition to providing a central repository of tree and character data, a phylogenetic database promised to become an important tool for biologists studying coevolution, biogeography, conservation, phylogenetic methods, or character congruence. In 1994, M. Donoghue, T. Eriksson, W. Piel, K. Rice, and M. Sanderson began work on a prototype phylogenetic database called TreeBASE, with support from Harvard University Herbaria, University of California at Davis, and a SGER NSF grant (DEB9318325). Since then, TreeBASE grew from a prototype to a more mature database. At present it is endorsed by a growing number of journals as a site for prospective authors to deposit their data.

TreeBASE Database Model

TreeBASE assumes that published phylogenetic works can boil down to a basic plan: each publication should contain one or more distinct phylogenetic analyses, each of which applies a particular algorithm using a particular weighting scheme on a set of data matrices to produce a set of trees. A tree, therefore, is always the product of one analysis, but matrices can be used by several different analyses. Authors invariably publish papers that are more complex than this simple plan envisions, but usually what is presented can be reduced to approximate it. Since TreeBASE only paraphrases the original work, we strongly recommend that the scientific community not use TreeBASE without consulting the author's published paper.

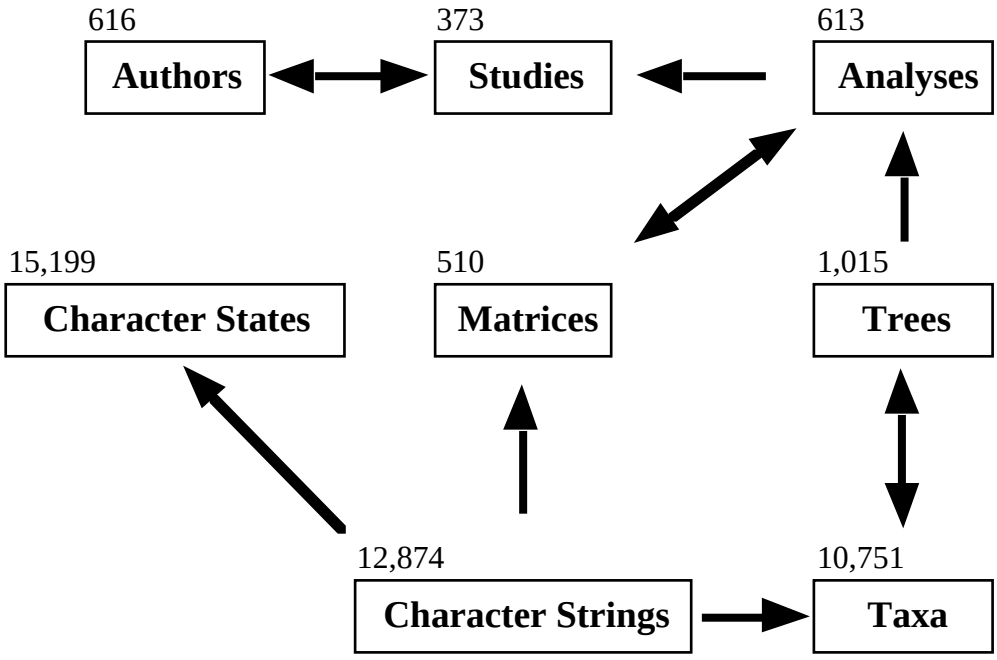

Fig. 1 Schematic of TreeBASE relational tables. Relations between tables are indicated by arrows from the "many" table to the "one" table. In other words, there can be many records of analyses for each study record, many trees for each analysis, etc. Two headed arrows indicate many-to-many relations, thus each tree can have many taxa and each taxon can exist in many trees. The number above each table indicates the number of records in TreeBASE as of November 1999.

In order to build this basic epistemological plan into the database model, we chose to use a relational database, which means that different types of data are stored in separate tables, each linked to another in a hierarchical chain. If a set of records in one table links to a single record in another table, the relations between the tables are said to be in a "many-to-one" relationship. Alternatively, a non-hierarchical "many-to-many" relationship is when a record from either table can relate to many records in the other (Fig. 1). TreeBASE's model is fairly complex and forms a circular loop: two separate paths link the table of studies with the table of taxa, one by way of the trees and one by way of the matrices (Fig. 1). Separating the data into numerous relational tables gives us more power over the stored data. For example, it allows us to combine matrices or recreate matrices with different sets of taxa—a procedure that would be much more difficult with a flat-file database. In addition, the separate tables approximately mirror the nexus phylogenetic data format, in which different types of data are

stored in their respective data blocks (Maddison et al., 1997). Figure 2 illustrates how the modular segmentation of data in the nexus format parallels the database model in TreeBASE.

Central to TreeBASE's model is the *Studies* table, which contains the citation of the published data and the paper's abstract. The names of authors are stored in a separate table with a many-to-many relation with the *Studies* table. Each study is linked to one or more records in the *Analyses* table, which stores information on the cladistic algorithm and the software used to perform it. The *Analyses* table also acts to associate matrices with resulting trees. Thus, each *Analyses* record simultaneously points to one or more *Trees* records and one or more *Matrices* records. The *Trees* table stores the phylogenetic tree in parenthetical notation (i.e. newick notation), the name of the figure in the publication, a generalized title for the tree, and the type of tree (consensus or single).

An alternative strategy would have been to store the tree as a set of parent-offspring records, where each parent-offspring record represents an internode segment within the tree (as implemented in HICLAS, http://aims.cps.msu.edu/hiclas/). Trees are then recreated by following the branching chain of parent-offspring records, starting from the most ancestral node and ending with the terminal taxa. This method would have allowed for more sophisticated manipulation of trees, such as combining them into supertrees, or skeletonizing them by pruning branches. However, we determined that this method would slow down the speed of tree retrieval and excessively burden the database with additional records—storing the tree in a single coded text field proved more effective, albeit less elegant.

The *Trees* table has a many-to-many relationship with the *Taxa* table (i.e. each *Trees* record can have several *Taxa* records, and each *Taxa* record can link to more than one *Trees* record). Each *Matrices* record stores the name of the matrix, the data type, and the format of the data. In addition, it has a one-to-many relationship with the *Character Strings* table, which stores each row of character data in the matrix. As of November 1999, TreeBASE has 510 matrices, which amount to 12,874 rows of character data (Fig. 1). Rather than store the columns of the matrix as separate records, we chose to put each row in a single text field in the *Character Strings* table. For the database program 4th Dimension, this limits each matrix to no more than 32,000 characters per taxon—the maximum number of characters in a text field. We felt safe with this strategy, seeing as the entire mitochondrion is about half this number.

Browsing and Searching Records

The scientific community can access TreeBASE using an HTML web browser pointing to the following URL: *http://phylogeny.harvard.edu/treebase*. The browser interacts with the web server through the Internet, which in turn passes instructions to and from the 4D database engine that is running on the same computer as the server. Users can initiate a search using five criteria:

1. *Author Search*—searches on the last name of an author of a publication.
2. *Citation Search*—searches on the full citation of the paper.
3. *Study Accession Number*—searches on a unique code that is assigned to each study.
4. *Matrix Accession Number*—searches on a unique code that is assigned to each matrix.
5. *Taxon Search*—searches on a taxonomic name that appeared in either the data matrix or the phylogenetic tree.

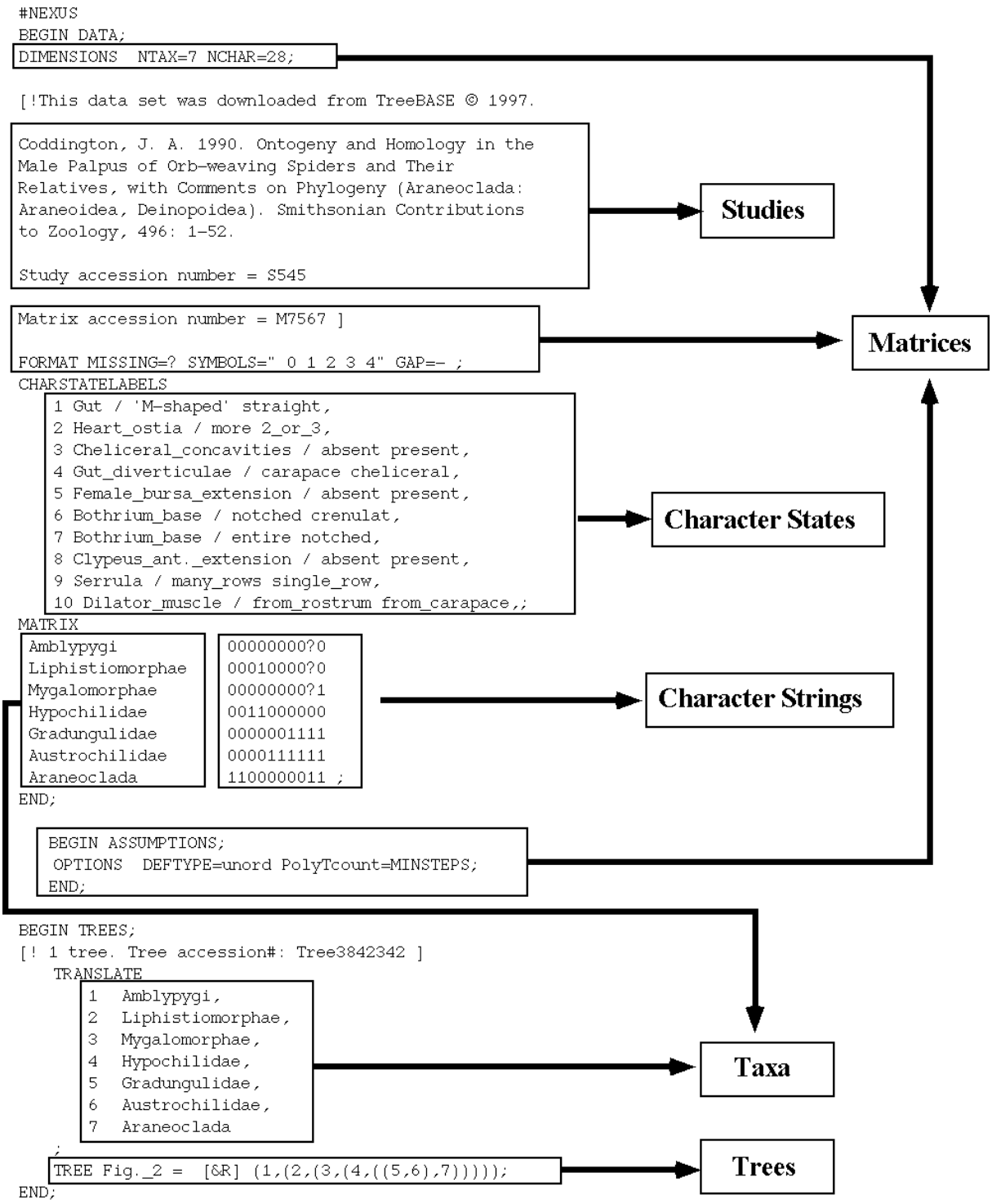

Fig. 2 Schematic illustrating where nexus elements are stored in TreeBASE tables. The boxes on the right represent TreeBASE tables; the boxes on the left represent blocks of nexus-formatted phylogenetic data. Arrows point to where the blocks of nexus data are stored in TreeBASE.

An initial search leads to a preliminary short-list of results. The user can then select from this list to display the resulting studies. A study is presented following the basic plan outlined above, i.e. the reference is followed by a list of analyses, and under each analysis is listed the

matrices used and the trees that result. Selecting a listed matrix downloads the data to the user's local computer; selecting a listed tree displays it in a separate frame.

Interacting with Trees

Displayed trees are interactive—clicking below a branch zooms in on a clade, clicking above a branch returns all descendant taxa to the shortlist to be searched on again. The first feature is designed to make it easier to explore larger trees; the second allows users to jump from one tree to another by searching on what is known about taxa on a particular branch. For large trees, users can also choose to see them in hyperbolic style, thanks to a Java program called Hyperbolic Tree™, provided by Inxight (*http://www.inxight.com/)*. Hyperbolic trees compress and prune branches that are more distant from a given point, removing the clutter and confusion that would otherwise occur when displaying large trees in the more usual way.

Tree surfing in a "small-world" network—The tree surfing feature allows users to locate "neighboring trees" by searching for other trees that contain one or more taxa found in the starting tree. The resulting set of trees is said to be related to the starting tree by one degree of separation; if newly encountered taxa in these related trees are searched on again, another set of trees will result—these are now related to the starting tree by two degrees of separation. As this task is repeated, more distantly related trees are discovered, and soon the entire island of trees has been explored. The associations among trees by way of shared taxa are neither completely random nor completely regular, but somewhere in-between. Consequently, collections of trees, even if spanning a large area of phylogenetic space, are all within relatively short "reach" of one another. This phenomenon is known as a small-world network (Barbási and Albert, 1999, Watts and Strogatz, 1998).

Various tendencies and characteristics of modern systematic methodology and productivity affect the small-world dynamics of a tree database. For example, a linked series of generic-level analyses might together form a "loose" island that requires many degrees of separation to traverse; but if just one family-level tree were added, this addition might suddenly "tighten" the island, markedly reducing the average degrees separating trees. The tendency of systematists to produce deep trees as compared to shallow trees, or the use of many outgroups as compared to just one or two, will shape the dynamics of taxon connectivity.

TreeBASE allows submitting authors to provide higher taxon names for internal nodes. While this feature has the advantage of increasing the connectivity among trees, authors do not always use it. The extent to which authors provide higher taxon names strongly affects the small-world dynamics of the database, since higher taxa are usually well connected with many distant trees. Consequently, future tree connectivity can easily change depending on the attitude and habits of submitting authors.

Presently, TreeBASE has a single super-island comprising over 70% of trees in the database—the remainder being distributed among 114 small islands (see figure 3 in Sanderson et al. 1998). Moreover, trees in the super-island are fairly well dispersed, with a mean distance of 11.5±3 iterations. While we can expect the super-island to tighten as more trees are added, we hope that the mean distance will not decrease too much, as the tree surfing feature will loose its effectiveness if it traverses the database too quickly. If this happens we may need to raise the stringency of the definition of a neighbor.

Eventually all trees in TreeBASE will be linked together to form a single super-island. But at present there are too few trees in TreeBASE, relative to the whole tree of life, for them all

to overlap. The number of additional trees needed in order that the entire database join the super-island depends, again, on the characteristics of systematic productivity. A simple simulation where subsets of randomly selected trees are tested for their degree of connectivity, demonstrates a distinct pattern of island growth and conglomeration. The lowest amount of connectivity occurs when the database has only 70 to 180 trees. At around 250 trees, the rate of new island formation begins to drop and the average island size surges. The number of islands no longer grows after 500 trees, where presumably the rate of conglomeration of separate islands, due to the addition of well-connected trees, matches the rate of new island creation. Since the database is not much larger than 1000 trees, we cannot say when the number of islands will begin to drop until there is just one super-island. However, judging by the fact that island growth already flattened at just 500 trees, we may achieve complete coverage surprisingly soon.

Building Supertrees using TreeBASE—One of the applications of tree surfing is to build collections of related trees that can then be fused together to form a supertree. Building supertrees is a powerful way of summarizing and synthesizing current systematic knowledge for a group of taxa. There are several different methods of supertree construction (Sanderson, et al., 1998)—matrix representation with parsimony (MRP) being the most popular for trees with incomplete overlap of taxa (Baum, 1992, Purvis, 1995, Ragan, 1992, Ronquist, 1996). Currently we are developing an MRP supertree construction tool to work together with our tree surfing tool. This combination will make it fairly easy to summarize and synthesize phylogenetic knowledge using TreeBASE.

Submitting Data

Authors of published phylogenies are strongly encouraged to submit their data to TreeBASE. The submission process is automated and fairly flexible, allowing authors to interrupt their submission and return to it at a later time. Author, reference, and abstract information are entered in HTML forms and processed by a common gateway interface program. Data are uploaded to TreeBASE as nexus-formatted data, and entered into a HTML form using the client browser's copy-paste feature. We strongly recommend that submitters use MacClade (Maddison and Maddison, 1992) to prepare their data because this program insures correspondence in spelling and accuracy in nexus syntax. Authors who do not have access to nexus-editors can contact our staff for assistance.

Submission Policy—TreeBASE will only accept data used in scientific work that has been or will be published in a peer-reviewed publication, such as a journal or book. Authors can initiate a submission prior to the manuscript being accepted for publication, but the data will only be made available to the public once the paper is treated as "accepted," "in press," or "published" by the journal's editorial board. Normally, trees in TreeBASE directly correspond to figures in the publication, however up to twenty additional unpublished trees may be included if their existence is mentioned in the paper's text. For example, they might be the set of most parsimonious trees, even if only a single consensus tree appeared as a figure in the publication.

Initially, authors are given a temporary submission tracking number (e.g. "SN123"). Later, when the paper has been accepted, the data are made public and TreeBASE issues a permanent study accession number (e.g. "S123") and a permanent matrix accession number for each data matrix (e.g. "M456"). The author can insert these numbers into the last version of the manuscript or in the galley proofs so that readers can easily locate the data in TreeBASE.

In some instances, authors can have the data withheld from public release until the journal has reached the library shelves.

Growth of TreeBASE—Submissions are vital for helping TreeBASE keep up with the tremendous growth of phylogenetic knowledge. An excellent source of data comes from those journals that require or recommend that authors submit data to TreeBASE, and we urge more journals to do the same. Recently we have seen an increase in the rate of new fungal data in TreeBASE, largely because several journals in this field are now directing prospective authors to submit their data to us.

For sequence data, the principle advantage of TreeBASE over GenBank (*http://www.ncbi.nlm.nih.gov/*) is that TreeBASE stores aligned datasets, as well as weighting schemes and step matrices that are peculiar to phylogenetics. It is critical that the readership of journals be able to recover the exact alignment used in a published analysis, particularly for non-coding sequences that are difficult to align (Cohen, et al., 1998). Without the alignment, the researcher's original hypotheses of homology are forever lost to the scientific community. Other on-line services, such as EMBL (*http://www.ebi.ac.uk/)*, accept aligned data but only as non-searchable, passive files. Unlike TreeBASE, none of these sites accept other types of phylogenetic data, such as morphological characters, RFLPs, allozymes, and the trees themselves. Consequently, we hope that TreeBASE offers a unique and valuable service for both scientists and scientific journals.

References

Barbási, A.-L. and R. Albert. (1999) Emergence of scaling in random networks. Science. 286:509-512.

Baum, B. R. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41:3-10.

Coddington, J. (1988) Cladistic tests of adaptational hypotheses. Cladistics. 4:3-22.

Cohen, B. L., J. A. Sheps and M. Wilkinson. (1998) Archiving molecular phylogenetic alignments as NEXUS files. Systematic Biology. 47:495-496.

Maddison, D. R., D. L. Swofford and W. P. Maddison. (1997) NEXUS: An extensible file format for systematic information. Systematic Biology. 46:590-621.

Maddison, W. P. and D. R. Maddison. (1992) *MacClade: Interactive analysis of phylogeny and character evolution.* Sinauer Assoc, Sunderland, Mass.

Mitter, C. and D. R. Brooks. (1983) Phylogenetic aspects of coevolution. In *Coevolution*. D. Futuyma and M. Slatkin (ed.). Sinauer Press, Sunderland, Massachusetts. pp. 65-98.

Mitter, C., B. Farrel and B. Wiegmann. (1988) The phylogenetic study of adaptive zones: Has phytophagy promoted insect diversification? American Naturalist. 132:107-128.

Mitter, C., B. Farrell and D. J. Futuyma. (1991) Phylogenetic studies of insect-plant interactions: insights into the genesis of diversity. Trends in Ecology and Evolution. 6:290-293.

Pagel, M. (1997) Inferring evolutionary processes from phylogenies. Zoologica Scripta. 26:331-348.

Purvis, A. (1995) A modification of Baum and Ragan's method for combining phylogenetic trees. Systematic Biology. 44:251-255.

Ragan, M. A. (1992) Phylogenetic inference based on matrix representation of trees. Molecular Phylogenetics and Evolution. 1:53-58.

Ronquist, F. (1996) Matrix representation of trees, redundancy, and weighting. Systematic Biology. 45:247-253.

Sanderson, M. J., B. G. Baldwin, G. Bharathan, D. Ferguson, P. J. M., C. Von Dohlen, M. F. Wojciechowski and M. J. Donoghue. (1993) The rate of growth of phylogenetic information, and the need for a phylogenetic database. Systematic Biology, 42(4) 562-568.

Sanderson, M. J., A. Purvis and C. Henze. (1998) Phylogenetic supertrees: Assembling the trees of life. Trends in Ecology and Evolution. 13:105-109.

Watts, D. J. and S. H. Strogatz. (1998) Collective dynamics of 'small-world' networks. Nature. 394:440-442.