Running Head: STBase

STBase: One billion species trees for comparative biology

Michelle M. McMahon[1], *Akshay Deepak*[2], *David Fernández-Baca*[2], *Darren Boss*[3] *and Michael J. Sanderson*[3, 4]

[1]School of Plant Sciences, University of Arizona, Tucson, AZ 85721

[2]Department of Computer Science, Iowa State University, Ames IA 50011

[3]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85721

[4]Corresponding author: email: sanderm@email.arizona.edu; phone: 520-626-6848; fax: 520-621-9190. Full address: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721

Abstract.—A new database, STBase, lets comparative biologists quickly retrieve phylogenetic hypotheses about species relationships. The database consists of 1 million single- and multi-locus phylogenetic data sets, each with a confidence set of 1000 species trees, computed from GenBank sequence data for 413,000 eukaryotic taxa. Two bodies of theoretical work are leveraged to aid in the assembly of multi-locus concatenated data sets for species tree construction. First, a novel "multree" reduction algorithm is used to prune multiply labeled gene trees to conflict-free, conservative, singly-labeled species trees that can be combined between loci. Second, impacts of missing data in multi-locus data sets are ameliorated by assembling only *decisive* data sets [*sensu* Sanderson et al. BMC Evol. Biol. 2010]. Using theses approaches, and in contrast to many phylogenetic databases that archive gene trees, STBase explicitly aims to construct and archive hypotheses about species relationships. Data sets overlapping with the query are ranked according to a scoring scheme that weighs tree quality and taxonomic overlap of the tree with the query. An efficient inverted indexing scheme lets us scale the database to ~1 billion trees with retrieval times independent of the size of the database, typically on the order of a few seconds. Tree quality is assessed by a real-time evaluation of bootstrap support on just the overlapping subtree. Associated sequence alignments, tree files and metadata can be downloaded for subsequent analysis. STBase may serve as a prototype for future species tree oriented databases and as a resource for assembly of larger species phylogenies from precomputed trees.

Phylogenetic trees have greatly altered comparative biology by rearranging the context for comparison, enhancing statistical power of comparative tests, and broadening taxonomic scope (Felsenstein 2004; Baum and Smith 2012). In recent years the demand for phylogenetic trees has been so high that comparative biologists themselves have frequently turned to heuristic or even non-algorithmic methods for assembling trees comprehensive enough to contain the taxa in which they are interested (e.g., Pringle et al. 2007 use of the Phylomatic Project: Webb and Donoghue 2005). This reflects one basic impediment to phylogenetic comparative studies: the mismatch between the set of taxa present in published or databased phylogenetic trees and the set of taxa for which comparative data are available. For example, the Royal Botanic Gardens, Kew, maintains a database of morphological and biochemical data on seeds of angiosperms (Flynn et al. 2006), which has been used in comparative analyses such as Moles et al.'s (2005) study of the correlates of seed size variation. Currently, of the 2572 eudicot species having data for the trait "percent oil content," some 36% have no sequences in GenBank, even though eudicots are arguably one of the best sampled species-rich taxonomic groups in the tree of life (the overall species level sequence coverage across described eukaryotes is closer to 10%: Sanderson 2008). Moreover, the eudicots that *are* represented in GenBank are not all sequenced for the same set of homologous loci; instead taxon coverage is patchy among various loci, so that phylogenetic trees assembled from GenBank sequence are more limited in their taxon coverage than the count of species in GenBank suggests.

One strategy to overcome this mismatch is assembly of ultra-large, dense phylogenies of particular clades (Bininda-Emonds et al. 2007; Nyakatura and Bininda-Emonds 2011; Peters et al. 2011; Smith et al. 2009, 2011; Aliscioni et al. 2012; Jetz et al. 2012), or particular regions of the world (Forest et al. 2007; Lanfear and Bromham 2011; Saslis-Lagoudakis et al. 2012; Holt et

al. 2013), depending on the biological question. However, scaling up phylogenetic inference presents numerous computational challenges (Bader et al. 2006; Goldman and Yang 2008; Liu et al. 2009; Izquierdo-Carrasco et al. 2011), especially in handling the patchy coverage across multiple sparsely sampled loci (Sanderson et al. 2010, 2011; Roure et al. 2013). An alternative strategy, which should be useful in the near term, is to assemble a very large collection of phylogenetic trees of small to medium scale, and optimize the delivery of these trees via efficient search and retrieval. This is the strategy we have employed here. As larger trees are needed, data sets and/or trees can be pieced together by other algorithms (see Discussion). One clear advantage of this is that it allows relatively robust estimation of reliability (yet another computational problem that does not scale well), and these estimates of reliability can be returned to the user.

In addition to the frequent mismatch between taxon sets of interest and taxon sets that are in published trees, a second basic impediment to harnessing available phylogenetic trees in comparative biology is that many are gene trees. More generally, many are "multrees", that is, trees having multiple sequences with the same taxon name. This can arise because of multiple sampling of individuals within a species, multiple alleles at the same locus, or multiple paralogs in the same gene family. Several tree databases implicitly allow such trees, including TreeBASE (Piel et al. 2002), and the PhyLoTA database (Sanderson et al. 2008), in addition to genomic databases that literally set out to archive gene trees instead of species trees (e.g., PFAM; Bateman et al. 2004; TreeFam: Li et al. 2006). However, it is not straightforward to undertake comparative biology of structure, function, ecology, etc., on multrees, especially those riddled with gene duplications, losses, or lateral transfer. The construction of species trees from gene trees is an active area of research, with an extensive and long-standing literature (Goodman et al.

1979; Page and Charleston 1997; Knowles 2009; Scornavacca et al. 2011; Anderson et al. 2012). We take an extremely conservative view of the problem, and implement a new method to ameliorate this impediment which we hope will at least expose some of the problems that must be resolved in future database efforts.

In this paper we describe a new database of precomputed phylogenetic trees of eukaryotes, STBase ("Species Tree Database"), optimized for use by comparative biologists. In it we deposit one billion pre-computed phylogenies built from single- and multi-locus datasets assembled from GenBank. Selection of taxa and loci for data set assembly is guided by recent theory on optimal multigene data set construction and treatment of multrees. We join this with a scalable search engine that accepts lists of taxon names and efficiently returns a ranked list of trees, the subtrees that overlap with the taxa of interest, and support values.

DESCRIPTION

Overview

The goal of STBase is to provide a tool that accepts a user's query list of taxon names and returns a ranked list of good "hits" to a database of phylogenetic trees. To quantify what "good hit" means (the term is meant to be analogous to BLAST searches, Altschul et al. 1990), we construct a scoring function that increases with the quality of the trees found and the amount of taxonomic overlap between the tree and the query. We assume that tree quality can be characterized by including a confidence set of trees in the database, computed, for example, by bootstrapping (as here) or by sampling the posterior distribution (Felsenstein 2004). Let A, be the query list, and α be a user supplied preference indicating the relative importance of tree quality

vs. taxon overlap. For any tree, T, let $L(T)$ be the taxa in the tree; $T \mid A$ be the subtree restricted to just the query taxa, and $L(T \mid A)$ be the taxa shared between the query and the tree. Then define $\omega\ (L(T \mid A))$ to be some increasing function of this overlap. Let $q\ (T \mid A)$ be some increasing function of the quality of the subtree. The score of a "hit" on tree A is then

$$S = \alpha \cdot \omega\ (L(T \mid A)) + q\ (T \mid A)$$

We normalize the score to range from 0 to 100 (1) using a quality score consisting of the product of the average bootstrap support for nodes above 50% and the fraction of nodes resolved on the majority rule tree (MRT) of the overlapping subtrees, (2) using an overlap function given by the number of overlapping taxa divided by the number of query taxa that are in the database (rather than the larger set of query taxa that might include taxa not found in GenBank at all) x 100, and finally, (3) dividing by $1 + \alpha$. Higher values of α make overlap increasingly important relative to quality.

Tree Construction

Single-locus data sets.—Figure 1 illustrates our tree construction pipeline. A pool of 160,972 single-locus data sets (Table 1) was assembled from GenBank rel. 184 largely according to the PhyLoTA pipeline described elsewhere (Sanderson et al. 2008). A set of 517 taxonomic groups in the NCBI hierarchy was selected so as not to exceed 35,000 sequences (excluding model organisms; cf. Sanderson et al. 2008 for details). We refer to these as "hub groups". These corresponded in practice very roughly to the rank of Linnean "orders" according to NCBI's taxonomy. Within each hub group clusters of homologous sequences were identified by all-

against-all BLAST searches and single-linkage clustering following 50% minimal overlap requirements as described. This operation was then repeated for each descendant group of the hub group in the NCBI hierarchy, inducing a set of parent-child relationships among clusters. From an original pool of 5,798,234 sequences among 413,628 distinct taxa, a set of 343,888 taxa were retained in phylogenetically informative clusters. For each cluster multiple sequence alignments using MUSCLE (Edgar 2004), ML optimal trees using default options in RAxML (Stamatakis 2006), and 1000 "fast" parsimony bootstrap trees using PAUP* (Swofford 2002) were obtained. Many (69%) of these clusters included at least one taxon ID multiple times; such taxonomically redundant sequences could be due to sampling of multiple individuals, or they could represent multiple alleles or even paralogous loci. The trees from such data sets are called "multrees" (Huber and Moulton 2006). We exploited a novel multree reduction algorithm (Deepak et al. 2012) to extract from each of these multrees a singly labeled "reduced" tree that retains the maximum amount of conflict-free species-level information (see Fig. 2). This is a conservative procedure that limits the number of false positive species relationships (see also Scornavacca et al. 2011 for a comparable algorithm aimed specifically at trees with gene duplications only). The user can retrieve either the multree or reduced tree for a single-locus data set.

Multi-locus datasets.—Assembly of multi-locus concatenated data sets is problematic in cases in which multiple sequences are present in the same taxon (Scornavacca et al. 2011). We therefore used the reduced set of taxa obtained from the multree reduction as the source of sequence data for assembly of supermatrices. This results in a loss of some taxa on average, but it also reduces the conflict within a gene tree arising from biological processes such as gene duplication and loss or incomplete lineage sorting. Although we have not built species trees

using any methods aside from concatenation, our collection of reduced loci/trees could be used as inputs to species tree inference methods using consensus (Degnan et al. 2009), reconciliation (e.g., Wehe et al. 2008; Akerborg et al. 2009) or explicit likelihood or bayesian methods exploiting the sequence data proper (e.g., Liu et al. 2009).

Two protocols were used to guide selection of subsets of taxa and loci for assembly of multi-locus supermatrices from the single-locus reduced data sets in each NCBI hub group and all its descendant groups. Both generate multi-locus data sets with a desirable property, "decisiveness", which can help limit the impact of missing entries in the supermatrix (Steel and Sanderson 2010; Sanderson et al. 2010, 2011; Soltis et al. 2011; Xi et al. 2012). A supermatrix, *M*, is *decisive* for tree, *T*, if and only if the subtrees, t_i, for each locus *i*, obtained by restricting *T* to only those taxa that have sequence data at locus *i*, uniquely define *T*. If, instead, the subtrees are consistent with more than one overall tree, the supermatrix may be unable to distinguish between those trees for certain reconstruction methods (e.g., parsimony or partitioned likelihood analysis: Sanderson et al. 2011). A particularly strong form of decisiveness, which holds for some patterns of missing data, is that *M* may be decisive *for all possible trees, T*. Our first protocol assembles maximal *complete* supermatrices, meaning every taxon is sampled for each locus, by finding all so-called maximal bicliques in an associated graph data structure (Sanderson et al. 2003; Driskell et al. 2004). Since any supermatrix in which one locus includes sequence from all taxa is decisive, these are decisive for all trees. Our second protocol also guarantees decisiveness but allows some missing entries in the supermatrix. It builds a supermatrix using one locus as a reference locus. Taxa restricted only to those in the reference locus, and any other locus with at least 33.3% taxon overlap with the reference locus, are allowed to join the supermatrix. Because of the reference locus, this supermatrix is also decisive for all trees, even

though it contains missing data, and we refer to it as a *decisive quasi-biclique* (*dqbc*). For a given collection of loci, one *dqbc* can be constructed using each locus as a reference in turn. Figure 3 illustrates these kinds of data sets, including the trivially decisive case of single-locus data sets.

The collection of maximal bicliques or decisive quasi-bicliques built at some node in the NCBI hierarchy can overlap with one another. It can also, in some cases with relatively dense taxon coverage, be a large collection. We found, for example that within mammals there were hundreds of thousands of primate and carnivore bicliques (more than all the number of bicliques for all other taxa combined, in fact); we therefore sampled from these collections only 2% and 20% of bicliques, respectively. Various checks and filters were run on the results. We checked whether there were duplicate data sets within or between nodes in the NCBI hierarchy and whether any decisive quasi-bicliques were actually bicliques (which occurs rarely when the taxon coverage pattern is conducive). In addition we used a BLAST protocol to check that all loci in a data set are independent from each other, sharing no local homologies (these can arise occasionally for a variety of reasons upstream in the pipeline) which might lead to redundant inclusion in the same supermatrix (e.g., Smith et al. 2011, corrigendum).

The output of this pipeline is nearly one million "phylogenetically informative" data sets (i.e., having at least four taxa), among which 351,212 distinct taxa recognized by NCBI are distributed. Computing time required is approximately 6 weeks on a 300 core linux cluster for the analyses described above. We estimate that repeating this with full maximum likelihood bootstrap analyses with default options in RAxML would require between 5-50 years on the same hardware.

The Database

Schema, search and retrieval.— The STBase database has a very simple schema aimed at maximizing search and retrieval efficiency. Essentially it consists of five entities: taxa, sequences, clusters, data sets and confidence sets of trees. A taxon consists of a species or subspecific name and its NCBI taxon ID (both following NCBI's taxonomy). A taxon can have multiple synonymous names mapped to the same taxon ID. Each sequence – represented by an NCBI GI number as its ID – is associated with a taxon and there can be multiple sequences associated with the same taxon. A cluster is a collection of homologous sequences, loosely interpreted as a "locus". A data set is a collection of one or more aligned clusters/loci, concatenated into a supermatrix (if more than one), from which trees were constructed. Each data set is mapped to a set of one thousand bootstrapped trees. To map efficiently among these entities, STBase employs universal hash functions (Motwani and Raghavan 1995; Cormen 2001) and string-specific hash functions (Jenkins 1997), which are capable of inserting and deleting a random element in constant time irrespective of the size of the collection.

The user inputs a list of taxon names and/or genus names. Genus names are replaced by a list of all taxon names in that genus. This is followed by five steps: (1) retrieval of corresponding taxon IDs, (2) finding the data sets having the desired overlap with the set of query taxa and reading them from disk, (3) processing each data set to restrict each of its thousand trees to the taxa that overlap with the query, (4) summarizing the restricted trees for each cluster as a majority rule tree (MRT), with support values, and returning these MRTs to the user. Finally, (5) in the case of multrees, a singly-labeled reduced tree is computed on demand (this only applies to single-locus data sets – for multi-locus data sets, redundant sequences are handled prior to concatenation).

Because of the collective storage requirements of the trees (over 200GB), trees from all data sets cannot be kept in RAM, which poses several challenges to achieving fast query processing. Given a set of taxon IDs, identifying overlapping clusters and reading them from disk memory is the most time consuming part of the query process, as there are nearly one million data sets, with 4 to nearly 10,000 taxa each, covering more than 340,000 total taxa (Table 1). However, STBase identifies overlapping clusters in time that is independent of size of the database by using inverted indexing (Zobel and Moffat 2006; Manning et al. 2008). Given a large collection of documents (e.g., web pages, or, here, data sets), an inverted index allows one to search and retrieve the subset of documents containing one or more words from the query set. It does so by maintaining a mapping from a predefined set of keywords to the documents in the collection that contain them. In STBase, the goal is to find the data sets containing taxa that map to the list of taxa supplied by the user. STBase's inverted index therefore stores exactly *which* data sets (cf. documents) contain taxon names (cf. keywords) and *where* those data sets are located on the hard drive.

Majority rule tree generation.— A query typically results in 100-200 data sets having sufficient overlap with the taxon names provided as input. Each of these results consists of a thousand bootstrapped trees that are then restricted to the query overlap and summarized as an MRT. To generate the MRT efficiently, we used Amenta et al.'s (2003) randomized linear time MRT algorithm, which uses hash codes – a constant size object – to represent bipartitions and a clever method to construct the MRT using only these hashed bipartitions. This results in a linear-time (i.e., optimal) algorithm.

As a result of these techniques and some careful preprocessing of the data, STBase answers queries in time that is linear in the total size of the query plus the output, and

independent of the size of the underlying tree repository. Retrieval times for queries of ~50 names on our database of 1 billion trees typically require 5 - 15 seconds. However, because the time is linear in the size of the output, query time can be significantly longer when the number of hits is very large. For example, a query on the genus name *Felis* (alone) or *Drosophila* (alone) finds a very large number of hits that must be retrieved, ranked and processed. The search engine by default limits output to 1000 records, each computed from the first 100 bootstrap replicates only. For queries returning longer lists, this is not guaranteed to return an optimal ranking, and modifying the defaults is a good idea. In the future we want to explore a "seeding" strategy, in which particularly poor candidates are immediately filtered out by computation on very small number of bootstrap replicates. The binomial probability of a well-supported clade (say 75%) dropping to 50% or less because of sampling error in a sample of 10 replicates is only 0.078, which may be an acceptable risk to speed up searches.

User interface.— Figure 4 shows a screenshot of the user interface. The user can enter a list of up to 10,000 species or subspecific taxon names as multinomials, following the NCBI taxonomy. Any uninomial is assumed to represent a genus name, and all species in that genus are added to the query. On the search page the user can optionally increase the required minimum taxonomic overlap to reduce the number of trees returned and can also select the format of taxon names for subsequent download after retrieval.

The output consists of a simple table layout of ranked hits, each row corresponding to a data set and its confidence set of trees. To orient the user to the phylogenetic scope of the trees returned, the least common ancestor (LCA) of the data set (within the NCBI hierarchy) is computed and returned (using an efficient LCA implementation as described in Sanderson et al. 2008). The weighting parameter for overlap vs. tree quality can be adjusted on this page with a

slider bar, which instantly re-orders the retrieved trees. A variety of data sets can be accessed from this page as well, including a nexus formatted multiple sequence alignment, a nexus formatted tree file for the overlapping subtree, and the multree reduction of any single cluster trees for which this has been computed. Metadata for the loci included in multi-locus data sets are embedded in the sequence alignment file. Single-locus trees are rooted (provisionally and no doubt approximately) by reference to a midpoint-rooted (Hess and de Moraes Russo 2007) optimal ML tree of the entire source tree, which is constructed and stored elsewhere in the database.

DISCUSSION

Tree quality.— Our pipeline was designed to limit upstream errors due to multiple sequence alignment problems in highly divergent taxa by restricting data set assembly to occur within but not between 500+ subtrees of eukaryotes. Within hub groups, we used "fast" parsimony heuristics. Although these tend to produce conservative tree estimates with bootstrap scores lower than those using more exhaustive heuristics (Müller 2005), the quality of trees was quite good on average. Table 1 reports the average fraction of nodes resolved in the bootstrap MRT, which is an aggregate indication of tree quality. Efforts to engineer decisive multi-locus data sets may explain the higher values in those data sets, but presumably these values are also due to presence of multiple loci and the smaller average size of trees, which is correlated with increasing bootstrap proportions (Sanderson and Wojciechowski 2000). The first release of STBase relied on fast parsimony algorithms for computational reasons. The 5-50 years of computing that would currently be required for full (not fast) ML runs is obviously out of bounds, and fast bootstraps in RAxML were not conservative.

Rationale for tree construction strategy.— The data sets in STBase overlap with one another. In mathematical terms, they represent a "cover" of the underlying sequence data, rather than a "partition" of it. Each data set in STBase comprises a different collection of sequences, taxa and loci, but these *collections* can overlap partially with each other. The effect of this is somewhat analogous to coverage in genome sequence assembly, where multiple reads allow evaluation of mistakes, except in phylogenetic inference the error(s) arise for many inferential reasons. For example, suppose we are interested in a set of taxa, U, common to two different data sets, having taxa sets X_1 and X_2: thus $U = X_1 \cap X_2$. After building alignments and trees from X_1 and X_2 separately, we might well discover that the subalignments and subtrees corresponding to just our taxa of interest, U, are different for any number of reasons. Both the optimal alignment and optimal tree given the alignment for the taxa in U can depend on the context, that is, the other taxa in X_1 or X_2. This is part and parcel of the longstanding debate over adding taxa vs. loci in phylogenetic analysis (Hedtke et al. 2006). By assembling data sets with many different contexts, including different phylogenetic scales (levels in the NCBI hierarchy), different numbers of loci, and different patterns of missing data (bicliques vs. decisive quasi-bicliques), we hope the database exposes sensitivity to these factors. By listing these different data sets ranked by quality in the output, the interface naturally encourages exploration of these effects. The important *caveat emptor* is that users should not be tempted to take *multiple* data sets returned in a search and perform subsequent phylogenetic or statistical analyses on them assuming they are statistically independent. Any *one* data set, however, reflects a non-redundant sample of sequence data.

Species trees and gene tree conflict.— One hallmark of STBase is that it archives estimates of species trees. More precisely it reports singly-labeled trees in which labels

correspond to NCBI taxa at the lowest rank to which they have been identified. We do this, not optimally, but conservatively. In other words, each multree from a single-locus data set is reduced to a singly labeled subtree in such a way that it does not introduce any conflict with the original tree (Deepak et al. 2012). This often entails loss of resolution and/or loss of taxa. These reduced sets of taxa then form the basis of multi-locus data set assembly. This multree reduction is not optimal because it does not exploit all of the information present in the original multrees, some of which (such as numbers of gene duplications, or deep coalescence events), can be helpful in inferring a more complete species tree (Maddison 1997). The structure of the database can accommodate other methods of assembling species trees, but at the moment the number of alternative methods is quite large, and we leave this to future work.

Utility in comparative biology and large tree construction.— Many, though not all, questions in comparative biology have a phylogenetic scope limited to major clades, and can thus be addressed by the trees within hub groups in STBase. Many problems in comparative physiology, functional morphology, developmental biology and comparative genomics are largely within the scope of species in the same taxonomic genus, family or order. On the other hand, STBase does not currently contain trees that span between our 500+ hub groups of eukaryotes, so, for example, studies that look at phylogenetic structure of all plants in a community or regional flora (cf. Forest et al. 2007) would not directly benefit from the database. However, we were reluctant to transcend our current scale for several reasons. The problems of scaling data set assembly, multiple sequence alignment, and tree inference beyond thousands to 10s of thousands of taxa are daunting (Smith et al. 2011). Moreover, we suspect that the exploitation of a small number of idiosyncratic high quality scaffold data sets will be necessary to tie together trees between major groups. For example, Soltis et al.'s (2011) analysis of 17 loci

for 640 angiosperms was a decisive multi-locus data set that could form the scaffold for our smaller trees among angiosperms, as could others. However, how these data should be incorporated with a large collection of smaller data sets is unclear, and no doubt raises many issues about supertree vs. supermatrix construction, as well as the proper handling of missing data.

Big data.— Given the combination of vast sequence data resources and computationally intractable inference problems, few would doubt the assertion that phylogenetics is "big data" science. A few observations gleaned in the construction of STBase adds some support for this notion. First, it became clear that storage of an entire confidence set of trees rather than a single tree (or a few alternative optimal trees) let us build a tool for exploration of the statistical support for phylogenetic hypotheses tailored to the user's taxon list, in real time. However, a consequence of this is the need to store 2-3 orders of magnitude more phylogenetic trees in the database. Although compression of these trees is possible (Stockham et al. 2002; Ané and Sanderson 2005), there will be a tradeoff between decompression speed and savings in database storage. Second, we selected only a small number of protocols for assembling data sets for tree construction. These were guided by some theoretical results on the impact of missing data in multi-locus data sets. Many other protocols could be designed to emphasize different aspects of data set structure, such as ones taking note of other measures of information content (e.g., Townsend et al. 2012), or to exploit the many different available species tree inference methods (Knowles 2009). Given sufficient computing resources, the number of data sets might easily be increased by 1-2 orders of magnitude by including such protocols. Finally, the 6 million sequences used to build STBase represent only a few percent of GenBank, the "taxonomically enriched" part, largely neglecting the vast quantities of high throughput sequence data that are

available (still) for a relatively limited number of taxa. Including these data would scale up analysis in our pipeline by two orders of magnitude, although perhaps not the number or size of trees to the same degree. However, if metagenomic data sets were ultimately included, the size of something like STBase might well approach 10^{13} -10^{15} trees. Databases of that size or larger exist now (e.g., NCBI's Sequence Read Archive, or Shutterfly's image database: Olavsrud 2012), but tailoring database tools to handle tree collections of this size while allowing efficient tree-based queries and other operations specific to phylogenetic analysis may well require new algorithms and engineering.

AVAILABILITY

The database can be accessed at http://searchtree.org. Software for the database retrieval engine and the code implementing the Deepak et al. (2012) multree reduction algorithm is available at http://code.google.com/p/search-tree/. All code is distributed according to a GNU GPL v3 license (http://www.gnu.org/licenses/gpl.html).

ACKNOWLEDGMENTS

This research was supported by NSF grant DEB-0829674 to MJS, MM, and DFB.

REFERENCES

Akerborg O., Sennblad B., Arvestad L., Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. U. S. A. 106:5714-5719.

Aliscioni S., Bell H.L., Besnard G., Christin P.-A., Columbus J.T., Duvall M.R., Edwards E.J., Giussani L., Hasenstab-Lehman K., Hilu K.W., Hodkinson T.R., Ingram A.L., Kellogg E.A., Mashayekhi S., Morrone O., Osborne C.P., Salamin N., Schaefer H., Spriggs E., Smith S.A., Zuloaga F. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol. 193:304-312.

Altschul S., Gish W., Miller W., Myers E.W., Lipman D. 1990. A basic local alignment search tool. J. Mol. Biol. 215:403-410.

Amenta N., Clarke F., St. John K. 2003. A linear-time majority tree algorithm. In: Proc. 3rd Workshop Algs. in Bioinformatics (WABI'03). pp. 216-227.

Anderson C.N.K., Liu L., Pearl D., Edwards S.V. 2012. Tangled trees: the challenge of inferring species trees from coalescent and noncoalescent genes. Methods Mol. Biol. 856:3-28.

Ané C., Sanderson M.J. 2005. Missing the forest for the trees: phylogenetic compression and its implications for inferring complex evolutionary histories. Syst. Biol. 54:146-157.

Bader D.A., Roshan U., Stamatakis A. 2006. Computational grand challenges in assembling the tree of life: Problems and solutions. In: Advances in Computers, Vol 68: Computational Biology and Bioinformatics. p. 127-176. DOI: 10.1016/S0065-2458(06)68004-2

Bateman A., Coin L., Durbin R., Finn R., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E., Studholme D., Yeats C., Eddy S. 2004. The PFAM protein families database. Nucleic Acids Res. 32: D138-D141.

Baum D.A., Smith S. 2012. Tree thinking: an introduction to phylogenetic biology. Roberts and Co.

Bininda-Emonds O.R.P., Cardillo M., Jones K.E., MacPhee R.D.E., Beck R.M.D., Grenyer R., Price S.A., Vos R.A., Gittleman J.L., Purvis A. 2007. The delayed rise of present-day mammals. Nature 446:507-512.

Cormen, T. 2001. Introduction to algorithms. MIT Press.

Deepak, A., Fernández-Baca, D., McMahon, M. 2012. Extracting conflict-free information from multi-labeled trees. In: Proc. 12th Workshop Algs. Bioinformatics (WABI'12). p. 81-92.

Degnan J.H., DeGiorgio M., Bryant D., Rosenberg N.A. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58:35-54.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172-1174.

Edgar R.C. 2004. Muscle: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:1-19.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Press, Sunderland, MA.

Flynn S., Turner R.M., Stuppy W.H. 2006. Seed information database (release 7.0, Oct. 2006) http://www.Kew.Org/data/sid.

Forest F., Grenyer R., Rouget M., Davies T.J., Cowling R.M., Faith D.P., Balmford A., Manning J.C., Proches S., van der Bank M., Reeves G., Hedderson T.A.J., Savolainen V. 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. Nature 445:757-760.

Goldman N., Yang Z. 2008. Introduction. Statistical and computational challenges in molecular phylogenetics and evolution. Philos. Trans. R. Soc. B 363:3889-3892.

Goodman M., Czelusniak J., Moore G.W., Romeroherrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132-163.

Hedtke S.M., Townsend T.M., Hillis D.M. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522-529.

Hess P.N., De Moraes Russo C.A. 2007. An empirical test of the midpoint rooting method. Biol. J. Linn. Soc. 92:669-674.

Holt B.G., Lessard J.-P., Borregaard M.K., Fritz S.A., Araujo M.B., Dimitrov D., Fabre P.-H., Graham C.H., Graves G.R., Jonsson K.A., Nogues-Bravo D., Wang Z., Whittaker R.J., Fjeldsa J., Rahbek C. 2013. An update of Wallace's zoogeographic regions of the world. Science 339:74-78.

Huber K.T., Moulton V. 2006. Phylogenetic networks from multi-labelled trees. J. Math. Biol. 52:613-632.

Izquierdo-Carrasco F., Smith S.A., Stamatakis A. 2011. Algorithms, data structures, and numerica for likelihood-based phylogenetic inference of huge trees. BMC Bioinformatics 12:470.

Jenkins, B. 1997. ALGORITHM ALLEY-What makes one hash function better than another? Bob knows the answer, and he has used his knowledge to design a new hash function that may be better than what you're using now. Dr. Dobb's Journal 22:107-110.

Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature 491:444-448.

Knowles L.L. 2009. Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58:463-467.

Lanfear R., Bromham L. 2011. Estimating phylogenies for species assemblages: A complete phylogeny for the past and present native birds of NewZealand. Mol. Phylogenet. Evol. 61:958-963.

Li H., Coghlan A., Ruan J., Coin L.J., Heriche J.K., Osmotherly L., Li R.Q., Liu T., Zhang Z., Bolund L., Wong G.K.S., Zheng W.M., Dehal P., Wang J., Durbin R. 2006. TreeFam: A curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 34:D572-D580.

Liu L., Yu L.L., Kubatko L., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320-328.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Manning C., Raghavan P., Schutze H. 2008. Introduction to information retrieval. Cambridge University Press.

Moles A., Ackerly D., Webb C., Tweddle J., Dickie J., Westoby M. 2005. A brief history of seed size. Science 307:576-580.

Motwani R., Raghavan P. 1995. Randomized algorithms. Cambridge: Cambridge Univ.

Muller K. 2005. The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and bremer support. BMC Evol. Biol. 5:58.

Nyakatura K., Bininda-Emonds O.R.P. 2011. Updating the evolutionary history of carnivora (Mammalia): A new species-level supertree complete with divergence time estimates. BMC Biol. 10.

Olavsrud T. 2012. How to implement next-generation storage infrastructure for big data. *CIO*. http://www.cio.com/article/704354/How_to_Implement_Next_Generation_Storage_Infrastructure_for_Big_Data.

Page R.D.M., Charleston M.A. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 7:231-240.

Peters R.S., Meyer B., Krogmann L., Borner J., Meusemann K., Schutte K., Niehuis O., Misof B. 2011. The taming of an impossible child: A standardized all-in approach to the phylogeny of hymenoptera using public database sequences. BMC Biol. 9.

Piel W.H., Donoghue M.J., Sanderson M.J. 2002. Treebase: A database of phylogenetic knowledge. In: Shimura J., Wilson K.L., Gordon D., editors. Tsukuba, Japan: To the interoperable "Catalog of

Life", National Institute for Environmental Studies. p. 41-47.

Pringle E.G., Alvarez-Loayza P., Terborgh J. 2007. Seed characteristics and susceptibility to pathogen attack in tree seeds of the peruvian amazon. Plant Ecol. 193:211-222.

Roure B., Baurain D., Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol. Biol. Evol. In press.

Sanderson M.J., Driskell A.C., Ree R.H., Eulenstein O., Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036-1042.

Sanderson M.J. 2007. Construction and annotation of large phylogenetic trees. Aust. Syst. Bot. 20:287-301.

Sanderson M.J. 2008. Phylogenetic signal in the eukaryotic tree of life. Science 321:121-123.

Sanderson M.J., Boss D., Chen D., Cranston K.A., Wehe A. 2008. The PhyLoTA browser: Processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335-346.

Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: The limits to inference. BMC Evol. Biol. 10.

Sanderson M.J., McMahon M.M., Steel M. 2011. Terraces in phylogenetic tree space. Science 333:448-450.

Sanderson M.J., Wojciechowski M.F. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from neo-astragalus (leguminosae). Syst. Biol. 49:671-685.

Saslis-Lagoudakis C.H., Savolainen V., Williamson E.M., Forest F., Wagstaff S.J., Baral S.R., Watson M.F., Pendry C.A., Hawkins J.A. 2012. Phylogenies reveal predictive power of traditional medicine in bioprospecting. Proc. Natl. Acad. Sci. U. S. A. 109:15835-15840.

Scornavacca C., Berry V., Ranwez V. 2011. Building species trees from larger parts of phylogenomic databases. Information and Computation 209: 590-605.

Smith S.A., Beaulieu J.M., Donoghue M.J. 2009. Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. BMC Evol. Biol. 9.

Smith S.A., Beaulieu J.M., Stamatakis A., Donoghue M.J. 2011. Understanding angiosperm diversification using small and large phylogenetic trees. Am. J. Bot. 98:404-414.

Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.-L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704-730.

Stamatakis A. 2006. Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Steel M., Sanderson M.J. 2010. Characterizing phylogenetically decisive taxon coverage. Applied Mathematics Letters 23:82-86.

Stockham C., Wang L.S., Warnow T. 2002. Statistically based postprocessing of phylogenetic analysis by clustering. Bioinformatics 18:S285-S293.

Swofford D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), 4.0 edition. Sinauer, Sunderland, MA.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835-849.

Webb C.O., Donoghue M.J. 2005. Phylomatic: Tree assembly for applied phylogenetics. Mol. Ecol. Notes 5:181-183.

Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. Duptree: A program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24:1540-1541.

Xi Z.X., Ruhfel B.R., Schaefer H., Amorim A.M., Sugumaran M., Wurdack K.J., Endress P.K., Matthews M.L., Stevens P.F., Mathews S., Davis C.C. 2012. Phylogenomics and a posteriori data partitioning resolve the cretaceous angiosperm radiation malpighiales. Proc. Natl. Acad. Sci. U. S. A. 109:17519-17524.

Zobel J., Moffat A. 2006. Inverted files for text search engines. ACM Computing Surveys (CSUR) , 38 (2), 6–es.

Table 1. Summary statistics for the three kinds of data sets.

	Number of data sets	Loci (mean and range)	Taxa[1] (mean and range)	Data set size[2] (mean and range)	Mean support (fraction of resolved nodes on MRT)
Single-locus clusters	160,801[3]	1 (1-1)	63.1 (4-8767)	63.1 (4-8767)	0.51
Bicliques	762,529	9.8 (2-91)	15.6 (4-510)	142.3 (8-1526)	0.84
Decisive quasi-bicliques	67,103	12.4 (2-386)	27.8 (5-1406) [4]	234.7 (10-9516)	0.68
Total database	990,433	8.5 (1-386)	24.1 (4-8767)	135.7 (4-9516)	0.79

[1]We require a minimum of four taxa in a data set, required for potentially informative relationships in an unrooted tree.

[2]Product of number of loci and number of taxa.

[3]Of these, 111,433 were multrees. Some 11,358 data sets had fewer than 4 taxa after multree reduction, so only 149,443 were used to build multi-locus data sets.

[4]Because we require four taxa for minimal potential phylogenetic informativeness, a decisive quasi-biclique data set, which has some entries missing, must have a minimum of five taxa (else it would be a biclique, proper).

Figure Legends.

Figure 1. Pipeline for tree construction.

Figure 2. Illustration of the multree reduction algorithm (Deepak et al. 2012). Numbers in parentheses indicate multiplicity and are not part of the labels themselves. The singly-labeled tree on the right is the maximally reduced form of the multree on the left and contains only conflict-free quartets from the original multree, i.e., quartets that are not topologically contradicted by any other quartet displayed by the original tree over the same set of four leaves.

Figure 3. Construction of single- and multi-locus data sets for STBase. Schematic of partial taxon coverage pattern for three loci. Three kinds of data sets are shown: single-locus data sets, which are complete but restricted to a single-locus; biclique data sets, which are complete for a subset of loci and taxa; decisive quasi-biclique loci, which are complete for one locus, but have partial coverage for the other loci. All types are guaranteed to be phylogenetically decisive for all trees on the full label set.

Figure 4. Screenshot of the STBase user interface. Query taxa are shown in the query box at upper left. Top hits are ranked in the list at lower right. One of these is selected for viewing at upper right. Lower left tree shows the larger source tree from which the overlapping subtree was extracted.

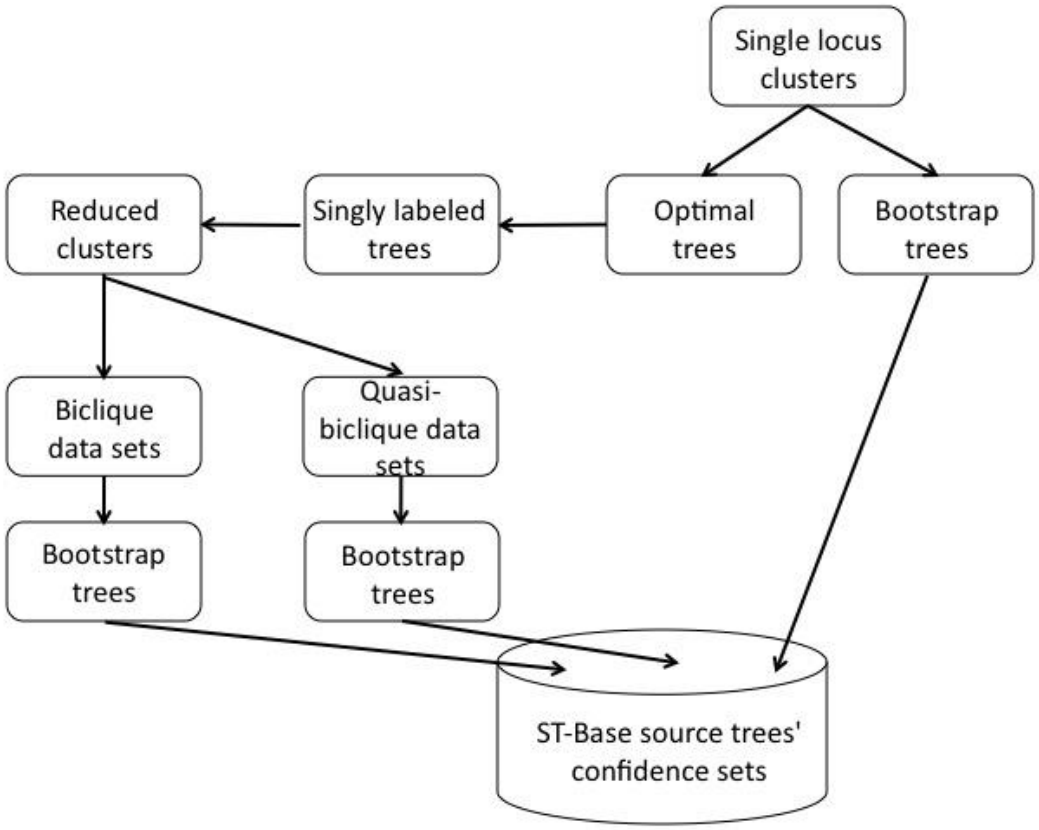
Single locus clusters
Optimal trees
Bootstrap trees
Singly labeled trees
Reduced clusters
Biclique data sets
Quasi-biclique data sets
Bootstrap trees
Bootstrap trees
ST-Base source trees' confidence sets

Fig. 1.

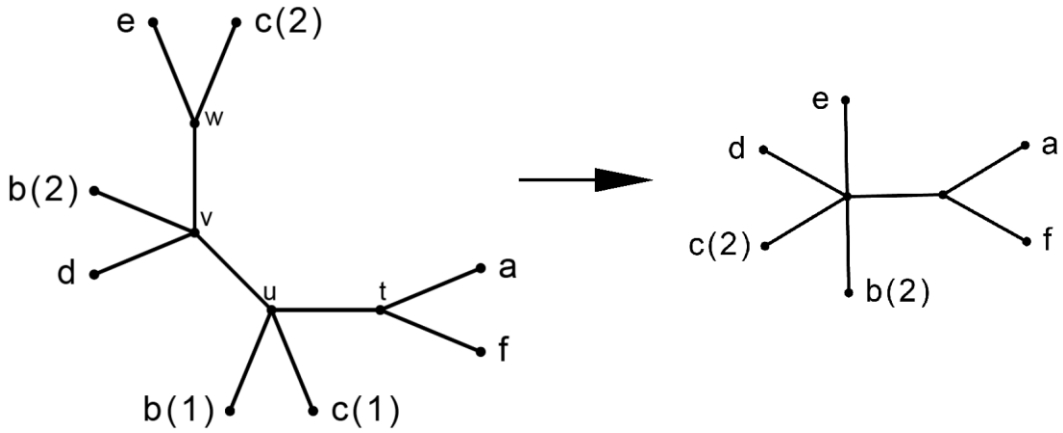

Fig. 2.

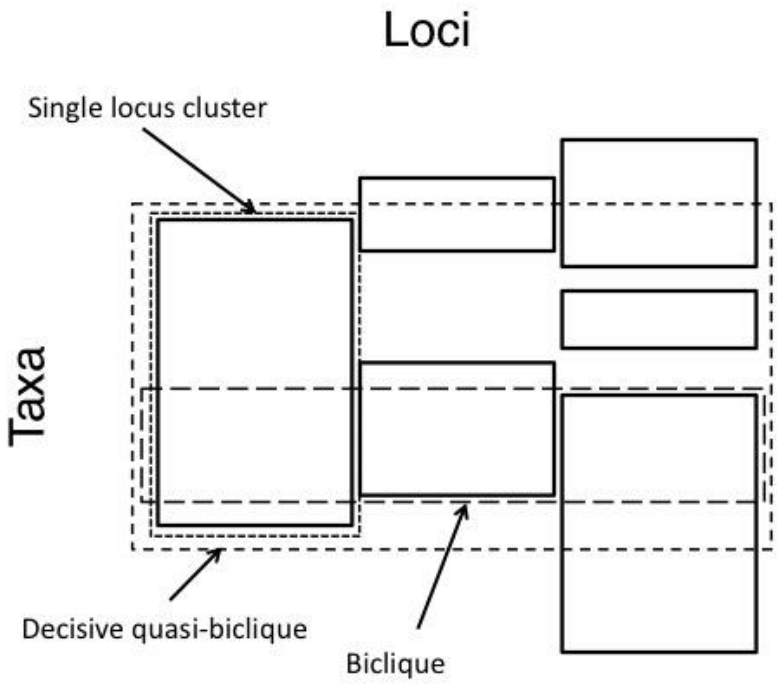
Loci
Single locus cluster
Taxa
Decisive quasi-biclique
Biclique

Fig. 3.

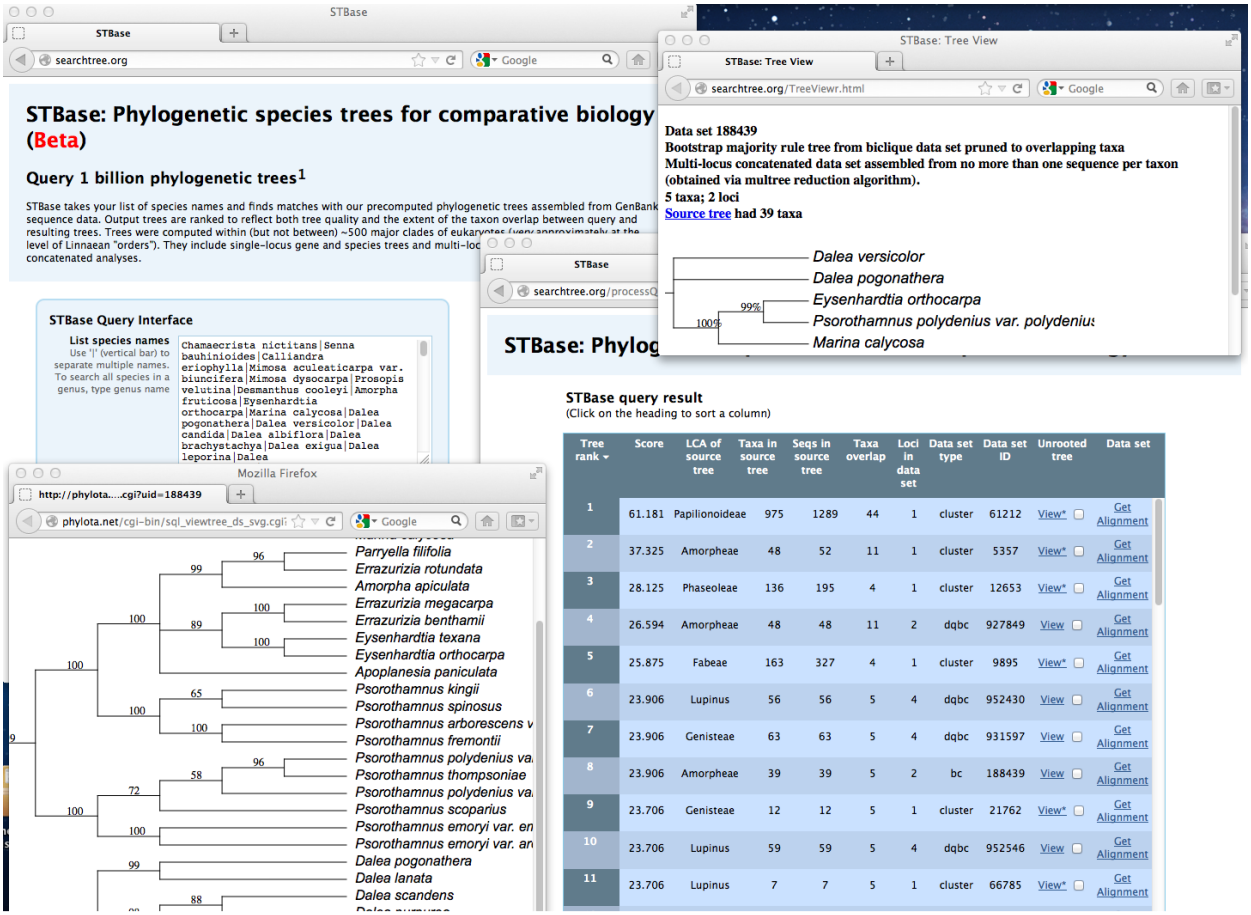
STBase
searchtree.org
STBase: Phylogenetic species trees for comparative biology (Beta)
Query 1 billion phylogenetic trees[1]
STBase takes your list of species names and finds matches with our precomputed phylogenetic trees assembled from GenBank sequence data. Output trees are ranked to reflect both tree quality and the extent of the taxon overlap between query and resulting trees. Trees were computed within (but not between) ~500 major clades of eukaryotes (very approximately at the level of Linnaean "orders"). They include single-locus gene and species trees and multi-locus concatenated analyses.
STBase Query Interface
List species names
Use '|' (vertical bar) to separate multiple names. To search all species in a genus, type genus name
Chamaecrista nictitans|Senna bauhinioides|Calliandra eriophylla|Mimosa aculeaticarpa var. biuncifera|Mimosa dysocarpa|Prosopis velutina|Desmanthus cooleyi|Amorpha fruticosa|Eysenhardtia orthocarpa|Marina calycosa|Dalea pogonathera|Dalea versicolor|Dalea candida|Dalea albiflora|Dalea brachystachya|Dalea exigua|Dalea leporina|Dalea
STBase: Tree View
searchtree.org/TreeViewr.html
Data set 188439
Bootstrap majority rule tree from biclique data set pruned to overlapping taxa
Multi-locus concatenated data set assembled from no more than one sequence per taxon (obtained via multree reduction algorithm).
5 taxa; 2 loci
Source tree had 39 taxa
Dalea versicolor
Dalea pogonathera
Eysenhardtia orthocarpa
Psorothamnus polydenius var. polydenius
Marina calycosa
99%
100%
Mozilla Firefox
http://phylota....cgi?uid=188439
phylota.net/cgi-bin/sql_viewtree_ds_svg.cgi
Parryella filifolia
Errazurizia rotundata
Amorpha apiculata
Errazurizia megacarpa
Errazurizia benthamii
Eysenhardtia texana
Eysenhardtia orthocarpa
Apoplanesia paniculata
Psorothamnus kingii
Psorothamnus spinosus
Psorothamnus arborescens v
Psorothamnus fremontii
Psorothamnus polydenius va
Psorothamnus thompsoniae
Psorothamnus polydenius va
Psorothamnus scoparius
Psorothamnus emoryi var. en
Psorothamnus emoryi var. an
Dalea pogonathera
Dalea lanata
Dalea scandens
STBase query result
(Click on the heading to sort a column)
Tree rank
Score
LCA of source tree
Taxa in source tree
Seqs in source tree
Taxa overlap
Loci in data set
Data set type
Data set ID
Unrooted tree
Data set
1 61.181 Papilionoideae 975 1289 44 1 cluster 61212 View* Get Alignment
2 37.325 Amorpheae 48 52 11 1 cluster 5357 View* Get Alignment
3 28.125 Phaseoleae 136 195 4 1 cluster 12653 View* Get Alignment
4 26.594 Amorpheae 48 48 11 2 dqbc 927849 View Get Alignment
5 25.875 Fabeae 163 327 4 1 cluster 9895 View* Get Alignment
6 23.906 Lupinus 56 56 5 4 dqbc 952430 View Get Alignment
7 23.906 Genisteae 63 63 5 4 dqbc 931597 View Get Alignment
8 23.906 Amorpheae 39 39 5 2 bc 188439 View Get Alignment
9 23.706 Genisteae 12 12 5 1 cluster 21762 View* Get Alignment
10 23.706 Lupinus 59 59 5 4 dqbc 952546 View Get Alignment
11 23.706 Lupinus 7 7 5 1 cluster 66785 View* Get Alignment

Fig.4.