Running head: PHYLOTA BROWSER

The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research

MICHAEL J. SANDERSON[1], DARREN BOSS[1], DUHONG CHEN[2],

KAREN A. CRANSTON[1] , AND ANDRE WEHE[2]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson 85721, USA; E-mail: sanderm@email.arizona.edu (M.J.S.); dboss@email.arizona.edu (D.B.); cranston@email.arizona.edu (K.C.)

[2]Department of Computer Science, Iowa State University, Ames, IA, USA; E-mail: duhong@iastate.edu (D.C.); andre@wehe.us (A.W.);

Author for proofs: Michael Sanderson; Department of Ecology and Evolutionary Biology, University of Arizona, Tucson 85721, USA; E-mail: sanderm@email.arizona.edu; Phone: 520-626-6848

Abstract.—As an archive of sequence data for over 165,000 species, GenBank is an indispensable resource for phylogenetic inference. Here we describe an informatics processing pipeline and online database, the PhyLoTA Browser (http://loco.biosci.arizona.edu/cgi-bin/pb.cgi), which offers a view of GenBank tailored for molecular phylogenetics . The first release of the Browser is computed from 2.6 million sequences representing the taxonomically enriched subset of GenBank sequences for eukaryotes (excluding most genome survey sequences, ESTs, and other high throughput data). In addition to summarizing sequence diversity and species diversity across nodes in the NCBI taxonomy, it reports 87,000 potentially phylogenetically informative clusters of homologous sequences, which can be viewed or downloaded, along with provisional alignments and coarse phylogenetic trees. At each node in the NCBI hierarchy, the user can display a "data availability matrix" of all available sequences for entries in a subtaxa-by-clusters matrix. This matrix provides a guidepost for subsequent assembly of multigene data sets or supertrees . The database allows for comparison of results from previous GenBank releases, highlighting recent additions of either sequences or taxa to GenBank, and letting investigators track progress on data availability worldwide. Although the reported alignments and trees are extremely approximate, the database reports several statistics correlated with alignment quality to help users choose from alternative data sources.

Keywords: Phyloinformatics; phylogenomics; phylogenetic database; GenBank

As phylogeneticists' use of sequence data has increased in problems ranging from phylogeny reconstruction to species delimitation to conservation biology, so too has their reliance on GenBank and its associated tools. GenBank is the single largest biodiversity database that is both widely accessible and easily queried. Nonetheless its origins and "target" audience have channeled its development in ways that are not always optimized for phylogenetics research. This paper discusses the issues involved in parsing and processing this database to optimize its utility for molecular phylogenetics research and describes the first release of a web-enabled database server designed with this goal in mind.

The quantity of sequence data in GenBank is legendary. Release 159 contains 75 billion nucleotides in 72 million sequences. Distributed separately from these are another 93 billion nucleotides in the WGS (whole genome shotgun) sequence division stemming from 787 registered genome projects. The combination of community-wide commitment to deposition of sequence data, and the free and open access of the database to users everywhere has fostered a remarkably diverse constellation of empirical and computational analyses of GenBank's data across many disciplines. Phylogenetics has been no exception: studies at many levels have exploited the database to varying degrees, from simple searches to check for PCR contamination, to finding homologs for a given sequence among outgroups, to more comprehensive studies based on data mining large numbers of taxa or loci (McMahon and Sanderson, 2006; Ciccarelli et al., 2006; Bininda-Emonds et al., 2007; Li et al., 2007).

GenBank also maintains a "taxonomy tree" for all taxa with sequences in the

database (Federhen 2003). Remarkably, with over 165,000 taxa represented, this tree touches on nearly 10% of described biodiversity. The classification is periodically updated to reflect changes in phylogenetic knowledge. As a phylogeny, it has shortcomings. Most phylogeneticists have probably observed problematic aspects of the NCBI tree in the clades with which they are familiar (e.g., the “tribes” of legumes do not reflect recent findings about legume relationships: c.f., Lewis et al., 2005); but, on the other hand, it is the most comprehensive and readily accessible classification of a significant subset of all life presently available.

In addition to serving "raw" sequence data via conventional text-based queries, NCBI's web server for the BLAST algorithm (Altschul et al., 1990) allows fast database searches based on local homology between a query sequence and GenBank databases. BLAST searches against GenBank are probably one of the most common exercises in molecular phylogenetic studies at all scales, both for validation of primary sequence data and exploration of related sequences. However, the collation and summary of information about sequence homologies is not especially tailored to phylogenetic inquiry. More often than not, a phylogeneticist is interested in the broad taxonomic context of sequence data. It is often useful to assemble or have at hand a set of related sequences, or “clusters”, constructed by global searches across the entire database and reported en masse, rather than piecemeal in response to specific queries of the user. A BLAST search using a single query sequence reports sequences related to the query, but will often miss other sequences that are transitively related to some of the target sequences found in the search. To some extent all of these are candidates for phylogenetic analysis. More useful still would be the ability to examine the entire “cluster set”, or collection of potentially

informative sequence clusters in GenBank at several levels of taxonomic specificity—say, for all turtles, or all Fabales. This would allow efficient construction of multigene data sets and could direct sequencing activities toward filling in missing data. Such tools are widely used in the protein structure/function and evolution communities (e.g., PFAM, Bateman et al., 2004) but not for the broad spectrum of sequence data that ultimately are useful in species tree construction. Finally, the all important exploration of outgroup diversity could be greatly facilitated by tools that examine "parent" clusters of homologs deeper in the tree of life. Of course, the wish list for tools to interoperate between the systematics community and GenBank is quite large, as efforts to link biodiversity informatics resources to GenBank atest (Page, 2005), but in this paper we focus on using high performance computing techniques to parse sequence data from GenBank into pieces—data sets—that are useful in phylogenetic analysis. Before describing the data pipeline to implement this, however, we discuss some of the general issues involved in computing on large collections of GenBank sequences.

GenBank and Molecular Phylogenetics

Model Organisms, Genome Projects and Taxonomic Diversity

GenBank's taxonomic diversity is spectacular but uneven. GenBank has 10.8+ million sequences for humans (not even counting WGS and some other high-throughput divisions), but also has 64,000 terminal taxa (usually species) with just one sequence. Most of the sequences in the database are the product of genome surveys, EST library construction and other high-throughput projects (**Fig. 1**), and with the exception of mostly recent environmental survey projects, most of these high throughput sequences

are focused on relatively few model taxa. This intense taxonomic sample bias in the database presents challenges in deciding what subset to parse as phylogenetically most useful. Although ultimately all of it may find application in phylogenomic analyses, the taxonomically defined divisions in GenBank are the most immediately useful. These are the divisions into which almost all sequences derived from systematics research are deposited as well as much of the sequence data on model organisms, including much (though not all) of the ultimately highly processed and annotated data from genomics projects. These data can be regarded as the "taxonomically enriched" subset of the database and are the target for this project.

Annotations and their Limitations

The two most relevant annotations of a GenBank sequence record for phylogenetic studies are the taxon name and the name for the gene or sequence region, and each presents significant problems. Taxon names are only as reliable as the original specimen identification and a nontrivial error rate has been noted (Vilgalys, 2003; McMahon and Sanderson, 2006). Fortunately the practice of adding voucher specimen data to GenBank records has become more prevalent. The taxon names themselves are coordinated with NCBI's taxonomy of accepted scientific names and synonyms. Many of these are linked to taxonomic names databases in the biodiversity informatics community (e.g., GBIF, IPNI, TROPICOS), but inevitably some interoperability issues have been noted (Page, 2005). Ironically, standardization of names for the sequences themselves seems to be in more difficult straits. GenBank maintains a list of accepted scientific names, but because sequences can have many different "features" (genes, introns, exons,

etc.), there is often no easy way to retrieve by name (i.e., by text query) exactly the sequences that might be of interest. A viable solution to this problem, which we have adopted, is to abandon annotations for sequence names entirely and retrieve collections of phylogenetically related sequences using database search algorithms like BLAST. The added computational burden is compensated for by reduction in errors or inconsistencies in sequence name annotations.

Constructing Optimal Phylogenetic Data Sets—Clusters

As the availability of sequence data has grown, so too have discussions of criteria for constructing "good" data sets for phylogenetic inference. Among the more widely discussed are prevalence of missing data (Wiens, 1998, 2005), sampling more loci or more taxa (Rosenberg and Kumar 2003, Hillis et al 2003), optimal levels of sequence divergence (Yang, 1998; Bininda-Emonds et al., 2001), taxon density (Mossel and Steel, 2005), strategies for avoiding long branch attraction artifacts such as breaking up long branches (Poe, 2003; Wiens, 2005), computational feasibility, and many issues related to data heterogeneity between loci (de Quieroz et al., 1995; Cunningham, 1997; Rokas et al., 2003; Vogl et al., 2003; Burleigh and Mathews, 2007). Unfortunately there are still few hard and fast rules to guide this kind of experimental design. In attempting to parse millions of sequences in GenBank into bite-sized nuggets for further phylogenetic analysis, there are thus few ground rules and many equally justifiable strategies.

Experience in assembling data sets at large scale in our previous work (Driskell et al., 2004; McMahon and Sanderson, 2006) suggests the importance of "alignability" as a criteria for selecting phylogenetic clusters. If sequences are too divergent, too

heterogeneous in length, or have too many insertions or deletions, alignment programs have difficulty finding accurate alignments, especially with noncoding nucleotide sequence data that are so important in many phylogenetic studies. Misalignment can have dramatic impacts on character-based phylogenetic inference, especially when progressive alignment algorithms (e.g., Clustal W; Thompson et al., 1994) insert gaps which propogate over many sequences and are then interpreted as synapomorphies by phylogenetics programs.

Upstream of multiple sequence alignment, even, are the criteria used to find sets of sequence homologs that are candidates for a cluster. Many databases of protein clusters have been built using a variety of strategies (Krause et al., 1999; Bateman et al., 2004; Tian and Dickerman, 2006). One widely used procedure uses database search algorithms such as BLAST to find homologs of one or more sequences. In extreme implementations, all sequences in a database are BLASTed against all others at some reasonably stringent cutoff level. This provides a list of pairwise "hits" between sequences. Then an algorithm to transform this list into a partition of the original sequence set is invoked. A simple agglomerative procedure is single-linkage clustering, which puts into the same cluster any sequence that has a hit to any sequence already in the cluster (Dondoshansky, 2002). Many other more stringent clustering schemes have been described, as have stochastic clustering methods that are sometimes computationally faster (Enright et al., 2002). In all cases, the size distribution and composition of the clusters is often quite sensitive to the criteria used in construction of the original hit list (e.g., BLAST E-values) and to the type of clustering undertaken.

An issue of special concern in phylogenetics is controlling the quantity of missing

data (Yan et al., 2005). Because homologous sequences in the database are often of very different lengths, and because database search algorithms nearly all identify hits based only on local homologies, alignments built from such clusters often of necessity have many missing pieces (owing to end-gap regions, or missing introns, protein domains, etc.). Although few hard and fast rules have been proposed for the amount of missing data that is tolerable (Sanderson et al., 2007), most phylogeneticists would probably prefer a sequence data matrix that is nearly complete, all else being equal. To enforce this in parsing the database requires attention to length heterogeneity, which can be controlled by requiring a certain level of overlap between sequences reported as BLAST hits.

Orthology and Paralogy

A gene tree can be constructed from an aligned cluster of sequences, but some are easier to interpret as species trees than others. Among the many processes that cause incongruence between gene trees and species trees, gene duplication is one of the most common in GenBank. Many protocols have been proposed to detect gene duplications in sequence data (Cotton, 2005; Koonin et al., 2005), ranging from those relying only on homology search strategies such as reciprocal best BLAST hits (Wall et al., 2003), to those using phylogenetic inference explicitly (Zmasek and Eddy, 2002; Storm and Sonnhammer, 2002; Sanderson et al., 2003; Wapinski et al., 2007). Many workers have sought means of detecting paralogous clusters in order to exclude them further phylogenetic analysis of species relationships, but it is also possible to explicitly take duplication histories into account in building species trees, as with “gene tree parsimony” methods (Goodman et al., 1979; Page and Charleston, 1997), or full evolutionary models

(Arvestad et al., 2003).

The PhyLoTA Browser

The PhyLoTA Browser is a web-accessible database that stores results of extensive cluster set computations using sequences from GenBank, aimed at constructing phylogenetic data sets of diverse size and taxonomic composition. Unlike most comparable efforts, which are targeted toward cluster sets of protein families, often using amino acid sequence comparisons, the PhyLoTA Browser works at the nucleotide level and includes noncoding data, which forms a huge component of the phylogenetically relevant subset of the database. Summary statistics on sequence and cluster set diversity across this tree are reported, along with provisional sequence alignments and rapid (i.e., relatively crude) estimates of phylogenetic trees. We anticipate that most users will wish to exert far more extensive manual control over this last stage of analysis, and we report these trees for reference purposes only.

Terminology

Because the sequence data are extracted from GenBank, we use the NCBI taxonomy tree as a convenient hierarchical framework for browsing phylogenetic aspects of sequence diversity (see Discussion for a critical assessment of this decision). The tree consists of *terminal* and *internal nodes*, each of which is named and has a *taxon ID*. A node's immediate descendant node is its *child* and the child's immediate ancestor is its *parent*. Sequences, each of which also has a unique ID called the *GI* number, are associated with nodes in the tree, which may be internal. This means there may be

sequences associated, for example, with "*Oryza sativa*", an internal node, and also with "*Oryza sativa* (japonica cultivar group)", which is a child node of "*Oryza sativa*" and a terminal node. To avoid confusion at internal nodes, sequences associated with exactly and only that node are referred to as *node-only* sequences; whereas sequences associated with any and all nodes descended from that node (including the node itself) are referred to as *subtree* sequences.

A *cluster* is a set of sequences that exhibit some kind of sequence homology (see below). A collection of sequences can be partitioned into a set of mutually exclusive clusters, or *cluster set*. Node-only sequences can be used to build *node clusters*; subtree sequences to build *subtree clusters*. A *phylogenetically informative cluster* is a cluster in which either the number of sequences is more than four, or, more strictly, if the number of taxon IDs is greater than four. In either case, this is a necessary condition for the construction of an unrooted phylogenetic tree. In the second case, it may potentially provide information about taxon relationships in addition to sequence relationships. A *parent cluster* is a cluster at a node's parent in the NCBI tree that contains all the sequences of the node's cluster in addition to others.

A *model organism* is defined in two ways. First, any node that has greater than an arbitrarily large number of sequences (in this release, 20,000, found in exactly 11 taxa) is considered a model organism. In addition, when clustered, any node found to contain more than 100 clusters itself is considered a model organism. Using this second criteria adds an additional 412 nodes to the list of model organisms in the database. This distinction allows the user to obtain different views of the database to distinguish between those species that have many sequences for only a few genes (e.g., from

population genetic surveys) and those that truly have been sequenced for a large diversity of genes or noncoding regions (true "model" organisms).

Database Backend

Data parsed.--All nucleotide sequences from the VERT, ROD, PRI, MAM, INV, and PLN GenBank divisions were parsed. These divisions span eukaryotic diversity but specifically exclude EST, HGT, and GSS and WGS projects. Moreover, a small fraction of these sequences are “long”, greater than 25,000 nt, and are not submitted to the analysis pipeline (though no doubt an important potential source of phylogenetic information, these long sequences pose special problems). In the present GenBank release, of 2.6 million sequences, 87,230 were excluded from cluster set calculations for length reasons. These are mainly complete assembled chromosomes from genome projects, whole chloroplast genomes and complete sequences from large clones such as BACs. Viral, bacterial and archeal diversity is not presently included in the database. The special qualities of the available sequence data for each are not readily accommodated by the same database schema used for eukaryotes (e.g., large number of complete genomes for bacteria). For an alternative, see for example, the GeneTrees phylogenomics system for prokaryotic sequence diversity (Tian and Dickerman 2006).

Cluster sets and parent clusters.--The primary computation undertaken on the sequence data is the construction of clusters representing potential phylogenetic data sets of related sequences. Clusters are assembled from a set of sequences by building a list of pairwise homologies (“hits”) identified in all-against-all BLAST (Altschul et al., 1990) searches, and putting those sequences into clusters using single linkage clustering.

BLAST searches were run with NCBI BLAST using an expectation value of 1.0e-10, "dust" filter off (to prevent fragmentation of hits in noncoding sequences caused by small motifs of low complexity) and requiring same-strand matches (-*S* 1 switch; this to preclude assembly of reverse complemented sequences into the same cluster). Moreover, hits were required to satisfy a coverage criterion of at least 51%, meaning the portion of sequence involved in the local homology hit(s) had to comprise at least that fraction of the original sequence length, with respect to both sequences (see McMahon and Sanderson, 2006, for additional discussion of this criterion). This criterion tends to produce clusters with an acceptable minimal level of length homogeneity. However, it also means that the pairwise homologies between, say, a complete mitochondrial genome of one species and a single mitochondrial gene of another will not put those two sequences into the same cluster.

Single linkage clustering is a fast clustering procedure that adds a sequence to an existing cluster if it hits any member of that cluster. Here it was implemented using the program 'blink' used in several previous studies (Driskell et al., 2004).

Cluster sets were constructed for each node and subtree of the NCBI tree (**Fig. 2**) until the depth of the node in the tree was such that the number of sequences in the subtree exceeded 20,000. This arbitrary cutoff was not chosen simply as a matter of computational convenience (in fact, all-all BLAST on the entire set of 2.6 million sequences was possible). It also reflects the observation that many clusters assembled via single linkage clustering begin to exhibit odd characteristics when they are extended too deeply in the tree. Enough bridges of small regions of local homology (even though greater than the 51% cutoff) eventually allow assembly of very heterogeneous clusters.

Although the cluster sets are built at many levels in the approximate hierarchy provided by the NCBI taxonomy tree, they are not conditional on the "correctness" of that hierarchy. The user is free to subsample from clusters or combine data across clusters whenever experimental design in the phylogenetic analysis calls for it (e.g., if an NCBI group is suspected to be paraphyletic). The cluster sets merely provide one trial partition of a very large database into a large but hopefully more manageable subset of data sets.

One of the nice properties of single linkage clusters is that a node or subtree cluster at one node in the NCBI tree will be a subset of some subtree cluster at its parent node (**Fig. 2**). This makes it easy to examine sets of related sequences for a group of taxa and then browse deeper into the tree to examine sequences at the parent node, which will often contain relevant outgroup taxa.

Alignment and tree reconstruction.--Alignments were constructed by two progressive alignment programs, Clustal W 1.83 (Thompson et al., 1994; with default parameters) and MUSCLE 3.6 (Edgar 2004; also with defaults). The latter performs alignment "polishing" after progressive alignment to try to improve its alignment scores. To assess similarity between Clustal W and MUSCLE alignments, we calculated the "Sum of Pairs" score (SP) often used in sequence alignment studies (Thompson et al., 1999), using the program 'qscore' (http://drive5.com/). The SP score on the fraction of pairs of nucleotides between sequences in a column that match in both alignments.

We also constructed "rapid assessment" phylogenetic trees for each alignment using the PAUP's "fast bootstrap" option, which builds trees very quickly and probably conservatively (Mort et al., 2000; Sanderson and Wojciechowski, 2000). To assess similarity between trees constructed from the two alignment methods, we calculated the

the consensus fork index between the two bootstrap trees (the fraction of clades present in both trees out of the maximum possible).

Orthology-paralogy assessment.--This assessment is based on a phylogenetic procedure outlined previously (Sanderson et al., 2003). In brief, in clusters in which multiple sequences exist for the same terminal taxon, a statistical test for whether or not these sequences are each other's closest relatives is undertaken. The assessment requires construction of two phylogenetic trees for each cluster, one optimal tree, and one tree constrained to have all conspecific sequences as clades, followed by a statistical tree comparison test like the K-H test (Felsenstein, 2004). To implement this on 87,000+ clusters, we used maximum parsimony trees (simple addition sequence, TBR swapping and a time limit of 60 minutes per tree) based on the Clustal W alignments.

This test can err in both directions. Small gene trees tend to have only a single sequence per taxon, and these are automatically regarded as not having duplications. On the other hand, some large trees sampled from a truly orthologous locus (such as a frequently sampled plastid gene) may show evidence of duplication somewhere in the tree, simply owing to a mistaken taxon name annotation, an introgression event, ancestral polymorphism, or other processes that cause some sequences with the same taxon name to be dispersed on the tree. It may be too severe to discard such clusters altogether, but a warning is indicated in the browser. This explicitly phylogenetic approach tries to overcome the limitations inherent in reciprocal best hit strategies (cf., Koski and Golding, 2001) but also does not require a “trusted” species tree as some recent phylogenetic approaches do (Zmasek and Eddy, 2001; Wapinksi et al., 2007).

Computational environment.--Computations are managed by PERL scripts that

implement an informatics pipeline (**Fig. 3**) that includes parsing the GenBank flat files, constructing the cluster sets and summarizing information about sequence and cluster diversity. One complicated step is cluster set construction. This requires a mixture of recursive code to traverse this tree and spawning of jobs to our Linux compute cluster (64 cores running under 'Rocks 4.3' Linux cluster distribution: http://www.rocksclusters.org) to undertake the actual BLAST runs and associated processing. Although we implemented this serially in the first release of the database, we have designed a more efficient parallel strategy for future releases that traverses the tree once to spawn all the node cluster set construction jobs (which, in the case of model organisms return the result that their sequences should not be used deeper in the tree); traverses it again to spawn the subtree cluster set construction jobs; and finally traverses it again to summarize all results (**Fig. 3**). This way we can manage tens of thousands of parallel jobs according to optimal settings in our job submission protocol rather than devising some protocol that attempts to subdivide the NCBI tree itself.

Alignments and tree reconstruction were distributed across the compute cluster using Sun Grid Engine 'array jobs'.

All results of cluster set construction and summary statistics are stored in a MySQL database.

Browser Front-end

Hierarchical display of diversity.-- In response to various user queries, the Browser displays a table of information for a node in the NCBI tree and all its immediate child nodes. Each row reports sequence diversity, taxonomic diversity, and cluster set

diversity for that group, as well as links back to NCBI's taxonomy pages, which in turn provide links to many biodiversity databases. Because model organisms have the potential to dramatically skew the user's perspective on sequence diversity in the data base, the user can optionally exclude these sequences and associated clusters from reports and summary statistics. Even in the "nongenomic" and "taxonomically enriched" subset of GenBank parsed by the PhyLoTA Browser, many of the model organisms still have thousands of sequences.

Cluster sets.--Information reported for each cluster includes summary statistics on the sequences it contains, such as a report on length variation and taxonomic diversity, as well as a rough assessment of the cluster's sequence length heterogeneity. The latter is reported as a "maximum alignment density" (MAD) which is the maximum fraction of cells in the final alignment that *could* have bases if the alignment algorithm did not need to insert any gaps. This is calculated as $\Sigma l_i/(nl_{\max})$, where n is the number of sequences, l_i is the length of the ith sequence, and $l_{\max}$ is the length of the longest sequence. Most alignments will have a density lower than the MAD value, but if the reported MAD value is low, it is likely to be a poor candidate for use in phylogenetic analysis. The user can download a FASTA formatted version of the cluster for subsequent alignment and tree reconstruction. Each cluster is annotated with an assessment of whether it is likely to contain gene duplications (or other sorts of gene tree incongruence that mimic this).

Queries.--Cluster sets (and their associated provisional alignments and trees) and summary statistics can be retrieved for nodes in the NCBI tree via three kinds of queries using taxon names. Each of these queries is implemented optimally, so that searches are very fast, despite the number of clusters and the size of the NCBI tree. The first query

uses single taxon name or GenBank taxon ID number and reports the cluster set statistics for it and all its siblings at the same hierarchical level in the NCBI tree. This is easily implemented via a MySQL query to the indexed taxon name field.

The second query takes a list of taxon names and returns all phylogenetically informative clusters containing any (or all) of the entire list of names. This seemingly elementary query is actually more challenging to implement efficiently. The browser relies on an inverted indexing scheme described in more detail elsewhere (Chen et al., submitted), which allows fast lookups of a list of taxon names. Once the indexes are precomputed, the running time of the query is $O(k \log m)$, where k is the number of clusters that contain the query taxa and m is the number of names in the query; note k does *not* depend on the total number or size of the clusters, important in this instance of 87,000 clusters, the largest of which has 4000+ taxon names.

Finally, the third query option finds the least common ancestor (LCA) in the NCBI taxonomy tree. Given a list of two of more taxon names, the browser returns the clusters of the LCA node and its child nodes. Because the tree has many nodes with high degree (i.e., nodes which have many children, a polytomy), this kind of query is useful for returning results that span a broad set of relatives of taxa of interest. However, because we currently do not compute clusters at all nodes back to the root of the NCBI tree, it is possible to submit a query (e.g., *Homo*, *Arabidopsis*) that will return the LCA but not provide any cluster information. Given that the NCBI tree has several hundred thousand nodes, an efficient algorithm for finding the LCA is needed. The browser uses an implementation of an elegant and efficient algorithm developed by Bender and Farach-Colton (2000). Let N be the size of the tree, and m be the number of taxon names

in the query ($m \geq 2$). The algorithm consists of an $O(N)$ precomputation step (implemented as a background daemon), which then allows a fast, $O(m)$ lookup of the LCA for any query. In other words, as in the second query, the running time does not depend on the size of the tree.

Tracking GenBank changes.--GenBank changes from release to release. The most obvious difference between releases is the increase in the number of sequences and taxa. However, GenBank's taxonomy itself changes with taxonomic progress. The PhyLoTA Browser displays changes in the numbers of sequences and clusters from the present release compared to the previous release, which provides a quick view of where taxa of interest have received attention from biologists somewhere in the world. It also visually highlights new taxon names not present in the previous release. This can quickly reveal very useful information about phylogenetic efforts. For example in the Genisteae, a clade of legumes, 293 new sequences and 19 new species were added between release 157 and 159, resulting in two new phylogenetically informative sequence clusters.

Data availability matrices.--A useful view of the subtree cluster set at a node in the NCBI tree is the data availability matrix, the rows of which correspond to taxa and the columns to clusters. A cell has an 'X' if at least one sequence is present in the cluster set for that taxon and cluster. This view gives an immediate impression of the density and evenness of the distribution of phylogenetic data and can provide a useful guide to supermatrix or supertree construction (Sanderson and Driskell, 2003; Sanderson et al., 2007). In the browser the user can select dense submatrices that have a minimum number of taxa and/or clusters. Formal algorithms also have been described for selecting complete or nearly complete submatrices with no or little missing data (Sanderson et al.,

2003; Yan et al., 2005), and these will be implemented in a future release of the PhyLoTA browser.

Alignment and tree visualization.--Alignments can be retrieved as FASTA files. Trees for clusters of <1000 sequences can be visualized directly from the browser, either by browsing cluster sets or retrieving cluster sets with queries. In the latter case, the query taxa are highlighted in the image along with their paths back to their LCA. Newick text descriptions stored in the database are used to dynamically create PNG graphics files. These in turn can be displayed in HTML pages. The image map associated with the PNG file lets us implement links out to NCBI for taxon and sequence IDs shown in the tree. All trees can also be downloaded as Nexus tree files.

RESULTS: PHYLOTA BROWSER (GENBANK RELEASE 159)

GenBank release 159 (April 15, 2007) was parsed to a subset consisting of 2.59 million nucleotide sequences. This version of the NCBI taxonomy tree contains 240,708 nodes for eukaryotes, for which clusters were constructed at 236,023 of these nodes, the remaining nodes being too deep, or in rare cases having no sequences (for terminal nodes). The total number of clusters constructed was 1.59 million, of which 1.19 million have only a single sequence. Of the remaining clusters, 182,399 had at least four sequences and hence were potentially informative about its gene tree; 87,087 had at least four distinct taxon IDs and hence were potentially informative about species relationships. Many clusters were large: the largest with respect to number of sequences and taxa had 7479 sequences for 4079 taxa (containing ITS sequences for the angiosperm clade Asteraceae). Eleven taxa were treated as model organisms because they each

contained at least 20,000 sequences. An additional 419 taxa were treated as model organisms because they had at least 100 clusters. For example *Xenopus leavis* had 17,471 sequences (few enough to proceed with clustering) but 11,266 clusters, clearly a reflection of its status as a model organism.

Clusters are variable in the number and length of their constituent sequences. The latter can be characterized roughly by their maximum alignment density (MAD) scores. The distribution of MAD scores among clusters reveals some interesting patterns (**Fig. 4**). Among all clusters, MAD scores are quite high, with over 65% above 0.9. However, many of these reflect the high scores of small clusters. Among the 1000+ clusters with over 500 sequences, MAD scores are much lower, peaking at an intermediate value between 0.4-0.5. This suggests large clusters will offer challenges not only because of their size, but also because of their length heterogeneity, a problem for all alignment programs.

Clustal W alignments were successfully constructed for all but one of the 87,000 phylogenetically informative clusters. A small number (49) of the largest clusters caused memory allocation errors in MUSCLE and were not completed despite considerable efforts both at the software and hardware levels to finish the set. Calculation of the SP scores between Clustal W and Muscle alignments was surprisingly compute intensive (apparently at least quadratic in the number of sequences). They were positively correlated with MAD scores ($\rho = +0.35$) and negatively correlated with number of sequences ($\rho = -0.33$). Orthology/paralogy assessments were successfully completed for all the Clustal W alignments. These tended to agree only roughly with expectations based on whether or not the locus is "known" to be single copy, such as for chloroplast

sequences. However, many factors can cause this test to infer paralogy even for true single copy loci, including paraphyly of species and misidentifications.

"Fast" bootstrap trees were successfully built for all alignments. For 4000+ clusters, neither of the trees built from Clustal W or Muscle alignments had any resolved clade, indicating that this fraction of clusters is essentially phylogenetically useless. On the other hand, a sizable number, 17727, had perfect resolution for both alignments and complete agreement with each other. However, these tended to be small with a mean number of sequences of 5.4 and the largest having only 33 sequences (the mean CFI of trees larger than 33 sequences was only 0.31). There was negligible correlation (ρ = ~0.02) between the SP alignment quality score and the CFI of either individual bootstrap tree or the strict consensus tree.

DISCUSSION

The Cluster Sets

The primary resource provided by the database, aside from summaries of information about sequence and taxonomic diversity in GenBank, are its cluster sets. Cluster set construction has been a mainstay of the protein family field for many years (Krause et al., 1999; Bateman et al., 2004), but its techniques have rarely been applied to non-protein-coding data like those that dominate many phylogenetic studies. Our previous work (Sanderson et al., 2003; Driskell et al., 2004; McMahon and Sanderson, 2006; Sanderson and McMahon, 2007) documented many factors that influence the number, size and composition of cluster sets, including BLAST parameters and filters, and requirements regarding fractional coverage or overlap between query and target

sequence. Thus there are many alternative cluster sets that can be constructed from any collection of sequences.

Our goal in choosing among all possible ways to build cluster sets is to aim for those that are phylogenetically useful, which is, sadly, a poorly defined concept. Some criteria for "useful" include having at least four sequences and being homogeneous enough that multiple sequence alignment programs will generate something serviceable. One of the conundrums in this analysis is the need to use local homology search algorithms to find potential homologs in the same cluster early in the informatics pipeline, but to satisfy the eventual desire to use well-proven global sequence alignment algorithms to build robust phylogenetic data sets later. Although local multiple sequence alignment algorithms exist, they are trying to solve a much harder problem than global alignment and do not generally perform as well on arbitrary inputs (Thompson et al., 1999). This has many implications for our design of the cluster sets.

Our experimental construction of alignments and trees supports some of these expectations. The overall quality of alignment as measured by the SP score between Clustal W and Muscle alignments decreased with decreasing MAD scores (increasing heterogeneity of sequences) and with larger numbers of sequences. Users of these alignments or trees should expect their quality to vary as a function of the reported MAD and SP alignment scores, as well as the CFI of the strict consensus tree relative to the original CFIs of the bootstrap trees. If the latter CFI drops significantly it is another indication of alignment problems. Irrespective of the alignment and tree quality, however, the cluster sets themselves are guaranteed to satisfy the criteria of local homology and overlap with which they were assembled. We hope these provide a resource for further

phylogenetic study at diverse phylogenetic scales.

Integration of Genomic Data

We processed only a subset of GenBank's sequence data, setting aside for now most of the sequence data gathered in a high throughput fashion from genome projects or model organisms. Large data sets associated with these data have been, and no doubt will continue to be, used for phylogenetic inference (Rokas et al., 2003; Ciccarelli et al., 2006; Sanderson and McMahon, 2007; Drosophila Genome Consortium, 2007) but their idiosyncracies and scale pose challenges for integration with the taxonomically enriched data we have assembled in the PhyLoTA Browser.

Although solving the computational problems associated with perhaps two orders of magnitude more data than we processed will be challenging (e.g., scaling all-against-all BLAST searches), far more difficult may be the issues of data assembly at this scale for phylogenetic inference. Complete genome sequences for eukaryotes are too complex and large to be aligned or analyzed phylogenetically en masse, so partitioning them into smaller data sets will be necessary. Yet, the complexity of these genomes, the presence of vast numbers of transposable elements between genes (for example, in plants) with little detectable homology between species, the difficulty of finding coding regions, the problem of gene duplication, exon shuffling, and larger genome rearrangements, and the inevitable fragmentation of phylogenetic data sets that these factors induce (Sanderson et al., 2007) will all exacerbate this challenge. Targeting exons in rapidly evolving protein coding regions is attractive if only because they are discrete entities. Phylogenetic methodology currently emphasizes analysis of such discrete entities. It is not yet

equipped to handle optimally genome scale sequences that contain mixtures of alignable and unalignable sequences, except via heuristics such as the widely used Gblocks alignment filterer (Castresana, 2000). In the meantime, it is likely that phylogeneticists will continue to gather sequences across many taxa, "one gene region at a time" and therefore the structure of this database will continue to accommodate these new data.

NCBI's Tree as a Framework for the Database

NCBI's taxonomy tree is not the tree of life. It is a hierarchical organizing scheme for the very large database it serves, and its authors take pains to convey this. Since we intend the PhyLoTA Browser to be a tool for phylogenetic research, it may seem disingenuous to make that same utilitarian claim here. However, the NCBI tree is also an extremely handy way to tunnel into a phylogenetic database of 165,000 taxa and 87,000 potential phylogenies. Generally speaking, it guides users to cluster sets in the phylogenetic neighborhood of the taxa of interest. However, the taxonomic composition of these clusters (though not their alignments or phylogenies) is obviously determined and limited to some extent by the composition of groups laid out by the NCBI hierarchy. Many of these groups are monophyletic; some or not. For example, there is a cluster set for Galegeae (a group of legumes), a taxon recognized by NCBI but now known to be polyphyletic (Sanderson and Wojciechowski, 1996). The next deeper group in the hierarchy, Papilionoideae, on the other hand, is indeed a clade. The PhyLoTA Browser does not assess the monophyly of groups on the NCBI tree. It is up to the user to be aware that outgroups according to the NCBI tree may actually nest within the ingroup and that any group may be paraphyletic or polyphyletic. This is usually a caveat to

consider at the early stages of all phylogenetic analyses.

The taxonomic composition of the cluster sets and trees derived from them should be viewed as *potential subtrees* of a larger tree; that is, a tree obtained by pruning some branches from a larger tree. A subtree *can* be a clade, but, if some taxa are missing, it can also be paraphyletic with respect to the taxonomic composition of the larger tree. Note that this view of the composition of clusters in the database would be necessary even if all NCBI taxa were perfectly monophyletic, because even then many clusters would be missing sequences from some taxon because of incomplete sampling.

Eventually, as methodologies for assembling cluster sets into large synthetic phylogenies mature, it might be hoped that these syntheses can replace the NCBI scheme as an organizing framework.

Integration with Other Databases

Seamless integration and interoperability between biological databases is the holy grail of biological informatics (Stein, 2003). NCBI provides many links from its sequence and taxon records to other databases. From taxon records, it frequently generates links to both taxonomic names databases like GBIF, or IPNI and TROPICOS for plants, and a few natural history collections databases, such as ARCTOS. The keystone problem for much of biodiversity informatics, including phylogenetics, is taxonomic name resolution (Page, 2005). Among the obstacles to establishing nomenclatural connections between databases are (1) multiple formal codes of nomenclature for different parts of the tree of life; (2) synonyms and homonyms; (3) differing taxonomic *concepts* for the same name (independent of formal taxonomic synonymy); (4) as yet incomplete electronic databases

of taxonomic names and concepts (although much progress has been made in the past few years: e.g., uBIO, GBIF, etc.); (5) and the more prosaic problems associated with data cleaning (i.e., mistaken spellings of names or other data). Page (2007) has illustrated many of these issues with respect to the phylogenetic tree database, TreeBASE (Piel et al., 2002), although most are emblematic of all biological database systems.

The particular stumbling block when thinking about connecting the PhyLoTA Browser to other databases is that it currently uses NCBI's list of accepted taxonomic names, which is a small subset of all names in the taxonomic literature. This will provide only limited interoperability capabilities until we can map these names onto larger lists of names. Some taxonomic names efforts are headed in the direction of providing easily usable web services for this purpose. For example, the uBio system will accept queries from external programs and reply with XML-encoded objects containing the relevant taxonomic name data. These and like-minded efforts will become indispensable tools to forge connections between the PhyLoTA Browser and the broader world of biodiversity databases.

ACKNOWLEDGMENTS

We thank Oliver Eulenstein, David Fernandez-Baca, Junhyong Kim, Michelle McMahon, Brian O'Meara, and Travis Wheeler for discussion of this project, and NSF's AToL program for funding.

REFERENCES

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

Arvestad, L., A.-C. Berglund, J. Lagergren, and B. Sennblad. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19 (suppl. 1):i7-i15.

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

Bininda-Emonds, O. R. P., S. G. Brady, J. Kim, and M. J. Sanderson. 2001. Scaling of accuracy in extremely large phylogenetic trees. Pacific Symposium on Biocomputing 6:547-558.

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

Burleigh, J. G., and S. Mathews. 2007. Assessing among-locus variation in the inference of seed plant phylogeny. Int. J. Plant Sci. 168:111-124.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540-552.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283-1287.

Cotton, J. A. 2005. Analytical methods for detecting paralogy in molecular datasets. Methods Enzymol. 395: 700-724.

Cunningham, C. W. 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. 14:733-740.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. Ann. Rev. Ecol. Syst. 26:657-681.

Dondoshansky, I. 2002. Blastclust (NCBI Software Development Toolkit), 6.1 edition. NCBI, Bethesda, MD.

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

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575-1584.

Farach, M., T. M. Przytycka, and M. Thorup. 1995. On the agreement of many trees. Inf. Processing Letters 55: 297-301.

Federhen, S. 2003. The taxonomy project *in* The NCBI handbook http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch4.

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

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romeroherrera, and G. Matsuda. 1979. Fitting the Gene Lineage into Its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Syst. Zool. 28:132-163.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124-126.

Koski, L. B., and G. B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52:540-542.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39:309-338.

Krause, A., P. Nicodeme, E. Bornberg-Bauer, M. Rehmsmeier, and M. Vingron. 1999. WWW access to the SYSTERS protein sequence cluster set. Bioinformatics 15:262-263.

Lewis, G., B. Schrire, B. Mackinder, and M. Lock. 2005. Legumes of the World. Royal Botanic Gardens, Kew.

Li, C. H., G. Orti, G. Zhang, and G. Q. Lu. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evolutionary Biology 7.

McMahon, M. M., and M. J. Sanderson. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55:818-836.

Mitchell, A., C. Mitter, and J. C. Regier. 2000. More taxa or more characters revisited: combining data from nuclear protein-coding genes for phylogenetic analyses of Noctuidea (Insecta Lepidoptera). Syst. Biol. 49:202-224.

Mort, M. E., P. S. Soltis, D. E. Soltis, and M. L. Mabry. 2000. Comparison of three methods for estimating internal support on phylogenetic trees. Syst. Biol. 49:160-171.

Mossel, E., and M. Steel. 2005. How much can evolved characters tell us about the tree that generated them? Pages 384-412 *in* Mathematics of Evolution and Phylogeny (O.

Gascuel, and M. Steel, eds.). Oxford University Press, New York.

Page, R. D. M. 2005. A taxonomic search engine: Federating taxonomic databases using web services. BMC Bioinformatics 6.

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

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423-428.

Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119-124.

Sanderson, M. J., C. Ane, O. Eulenstein, D. Fernandez-Baca, J. Kim, M. M. McMahon, and R. Piaggio-Talice. 2007. Fragmentation of large data sets in phylogenetic analysis *in* Reconstructing evolution: new mathematical and computational advances (O. Gascuel, and M. Steel, eds.). Oxford University Press, Oxford.

Sanderson, M. J., and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8:374-379.

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

Sanderson, M. J., and M. F. Wojciechowski. 1996. Diversification rates in a temperate legume clade: Are there "so many species" of Astragalus (Fabaceae)? Am. J. Bot.

83:1488-1502.

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

Stein, L. D. 2003. Integrating biological databases. Nat. Rev. Genet. 4:337-345.

Storm, C. E. V., and E. L. L. Sonnhammer. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92-99.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27:2682-2690.

Tian, Y. Y., and A. W. Dickerman. 2007. GeneTrees: a phylogenomics resource for prokaryotes. Nucleic Acids Res. 35:D328-D331.

Vogl, C., J. Badger, P. Kearney, M. Li, M. Clegg, and T. Jian. 2003. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. J. Mol. Evol. 56:330-340.

Wall, D. P., H. B. Fraser, and A. E. Hirsh. 2003. Detecting putative orthologs. Bioinformatics 19:1710-1711.

Wapinski, I., A. Pfeffer, N. Friedman, and A. Regev. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23:I549-I558.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease

phylogenetic accuracy? Syst. Biol. 47:625-640.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54:731-742.

Vilgalys, R. 2003. Taxonomic misidentification in public DNA databases. New Phytol. 160:4-5.

Yan, C. H., J. G. Burleigh, and O. Eulenstein. 2005. Identifying optimal incomplete phylogenetic data sets from sequence databases. Mol. Phylogenet. Evol. 35:528-535.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47:125-133.

Zmasek, C. M., and S. R. Eddy. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821-828.

FIGURE 1. Pie chart showing the fraction of nucleotides in GenBank release 159 distributed among its various divisions. Highlighted in the upper right is the taxonomically enriched subset used to populate the PhyLoTA Browser database. Abbreviations are the standard GenBank division names: BCT=Bacteria; PLN=Plants/Fungi; INV=invertebrates; VRT=vertebrates; MAM=mammals; ROD=rodents; PRI=primates; GSS=genome survey sequences; PAT=patented sequences; HTG=high throughput genomic; EST=expressed sequence tag. Some smaller divisions are not annotated on this figure.

FIGURE 2. Diagram showing the assembly of clusters and parent clusters at different nodes in the NCBI tree. DNA symbol represents a collection of sequences, usually associated with a terminal node, but occasionally also associated with an internal node, as here. Node clusters are assembled by analysis of sequences found only at that node. Subtree clusters are assembled from all sequences associated with that node *and* all its descendant nodes. Arrows indicate a parent-child relationship between clusters. That is, a parent cluster contains all the sequences found in a child cluster (plus more usually).

FIGURE 3. Flowchart of the informatics pipeline used to construct the PhyLoTA Browser database of sequence clusters.

FIGURE 4. Distribution of Maximum Alignment Density (MAD) scores among all phylogenetically informative clusters (diamonds) and among large (>500 taxa) phylogenetically informative clusters (boxes). MAD scores are positioned along the

horizontal axis at the lower boundary of their bin.