Internet On-Ramp

Phylogenetics—A Web of Trees

Tim Littlejohn, eBioinformatics,
Sydney, Australia

The quantity and diversity of genome sequence data continues to grow rapidly, thanks to ever-cheaper data production technologies and the relentless drive of the global genome effort. As a consequence, opportunities continue to expand around the use of these resources for richer comparative sequence analysis to discover the mode and nature of evolutionary change. In its latest release (June 15, 2000), the GenBank® database contained over 8.6 billion bases of sequence from over 75 000 different organisms (<http://www.ncbi.nlm.nih.gov>). While this information is by no means equally distributed across the species represented in the database, it does represent a rich repository of data ripe for phylogenetic analysis. The tools for such analysis are not just of interest to evolutionary biologists; the patterns of similarities between sequences they can reveal are also used in population genetics and molecular diagnostics, for example. Numerous "anonymous" Internet resources (sites that can be accessed without a username or password) exist for phylogenetic analysis. Also, Internet-based application service providers (ASPs) are particularly suited to phylogenetic analysis because they store the data and the results of analyses—along with the tools and databases necessary to analyze the data—with the often intense computing resources needed to run phylogenetic investigations.

Collecting Data

What are the best ways to collect this data so that evolutionary information can be mined from it? It is critical that the data being analyzed are homologous, that is, related by evolutionary descent. Database records can be retrieved using powerful text-searching engines such as Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>) and SRS (<http://srs.ebi.ac.uk/>) on the Web, or similar systems in bioinformatics' ASP systems such as the TextSearch function in BioNavigator (<http://www.bionavigator.com>) or similar academic ASPs such as those in the EMBnet (<http://www.embnet.org>). The pitfall with this approach is that it can be difficult to identify homologous sequences using keyword searches because databases frequently contain errors and inconsistencies in how homologous genes are named. Some databases, such as the Organelle genome database GOBASE (<http://megasun.bch.umontreal.ca/gobase/gobase.html>), Hobacgen (<http://biom3.univ-lyon1.fr/databases/hobacgen.html>) and Hovergen (<http://biom3.univ-lyon1.fr/databases/hovergen.html>) are ideal for phylogenetic analysis as they provide collections of homologous sequences for organelles, bacteria and vertebrates for such in-depth research.

Sequence-similarity searching methods can also be used to retrieve homologous sequences. Tools like BLAST are ideally suited to this function (see BLAST on the Web by Bruno A. Gaëta, *BioTechniques*, Vol. 28, No. 3, March 2000). This tool can be accessed at sites such as the NCBI (<http://www.ncbi.nlm.nih.gov>), the Japanese GenomeNet server (<http://www.blast.genome.ad.jp>), the European Bioinformatics Institute (<http://www.ebi.ac.uk>) and ASPs such as <http://www.bionavigator.com>. Organism-specific collections of

sequence data—for example, the *Arabidopsis* sequence database at <http://www.arabidopsis.org>—are an excellent way to collect specific data ready for phylogenetic analysis.

Multiple Sequence Alignments

Once collected, the next step for the analysis of phylogenetic information in molecular sequence collections is to perform a multiple sequence alignment (MSA). This step is critical in phylogenetic analysis because data is typically analyzed down columns of residues across aligned taxa under the assumption that residues in each column are homologous, that is, they share a common history. (See *BioTechniques* Vol. 29, No. 1, July 2000, for the article, *Multiple Sequence Alignment Tools on the Web*, by Giles J. Gaskell). The ClustalW multiple sequence alignment tool is one of the most popular and can be reached on the Internet at several sites including the European Bioinformatics Institute (EBI) (<http://www2.ebi.ac.uk/clustalw/>), the Baylor College of Medicine (<http://dot.im-gen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html>) and the Institut Pasteur Software Environment (PISE) in France (<http://bioweb.pasteur.fr/intro-uk.html>) or ASPs like BioNavigator. Other software tools for aligning and refining alignments can be found elsewhere (see Gaskell's article in Vol 29, No. 2, cited above). Frequently, MSA files need manual fine-tuning as none perform a perfect job at aligning homologous sites. Internet-based tools like CINEMA (<http://www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/>) and JalView (<http://www2.ebi.ac.uk/~michele/jalview/contents.html>) offer Java tools for such functions.

Phylogenetic Tree Inference

When a reliable multiple sequence alignment (MSA) file has been generated, the real phylogenetic analysis can begin. The three major methods for phylogenetic inference include maximum likelihood, parsimony and distance methods. Maximum likelihood methods calculate the probability that a given tree will result in the observed MSA. Programs using this approach are typically computer-intensive and have not been implemented on the Internet as widely as those using other methods. Parsimony methods determine the minimum number of changes required to produce the observed MSA. Unlike maximum likelihood and distance trees, parsimony trees do not include branch lengths. Distance methods work by pairwise comparisons of sequences, which infers an evolutionary distance between them, and then derive a tree based on these distances using methods like neighbor joining.

Some MSA programs produce dendrograms (trees indicating how the sequences are clustered, indicating their overall relationships with each other). In fact, ClustalW even allows the user to create Neighbor Joining trees with a bootstrap option. As described above, many Internet resources allow access to ClustalW. Other sequence alignment tools include TreeAlign, a multiple sequence alignment application that builds trees even as it aligns DNA or protein sequences using a combination of distance matrix and approximate parsimony methods. TreeAlign can be accessed through the PISE system (<http://bioweb.pasteur.fr/seqanal/phylogeny/intro-uk.html>).

The next step in phylogenetic analysis is the application of these MSA files as input data to phylogenetic tree construction software. Integrated ASP systems platforms such as BioNavigator (<http://www.bionavigator.com>) allow all these analyses to be conducted on one site using a seamless flow

Internet On-Ramp

of data through pipelines offering data collection from public and private data collections, pruning these sets, aligning and then phylogenetic tree inference. Other sites such as the Pasteur Institute's PISE system (<http://bioweb.pasteur.fr/intro-uk.html>) also facilitate sequence collection and alignment and then pass on these alignments for phylogenetic analysis. The BioNavigator system (<http://www.bionavigator.com>) further permits these analyses to be recorded as histories and then re-run automatically as autoprotocols, using other query sequences.

There are over 175 phylogenetic programs available (further discussion is beyond the scope of this review) that can be downloaded from the Internet or run from their Web sites directly. Many of these programs are specific to Macintosh® or Microsoft® Windows, and rely heavily on their operating system-specific user interfaces and so are not suitable for integration into a standard Web interface. However, as pointed out in an earlier *On-Ramp* article, all tree-building programs make a certain number of assumptions and it is critical that users understand these limitations before launching into any analyses.

Some packages (for example, PHYLIP) contain applications that allow analysis using all the major methods for molecular (nucleotide and protein) sequence data but also other data types. The PHYLIP package allows analysis of restriction site, gene frequency, quantitative characters and discrete characters, and can be found at <http://evolution.genetics.washington.edu/phylic/programs.html>. An excellent summary of these resources can be found at the University of Washington site at <http://evolution.genetics.washington.edu/phylic/software.html>. Web-based sites allowing online analyses through the anonymous Web are listed at <http://evolution.genetics.washington.edu/phylic/software.serv.html>. Other lists for phylogenetic software exist at EBI in the BioCatalog (<http://corba.ebi.ac.uk/Biobcatalog/Phylogeny.html>), the "Tree of Life" Project (<http://phylogeny.arizona.edu/tree/programs/programs.html>) and at the Willi Hennig Society site (<http://www.cladistics.org/education.html>).

A hallmark of most phylogenetic programs is that they are computationally intensive: as a consequence, there are few public sites that allow phylogenetic analysis because of the large computational requirements. Indeed, as Joe Felsenstein, author of the popular PHYLIP package (<http://evolution.genetics.washington.edu/phylic.html>) says, "Some sites have also been bold enough to make PHYLIP available on a server basis... I think they're very brave considering how slow some of our programs are". ASP Web providers such as BioNavigator offer the PHYLIP and other packages (e.g., GCG: <http://www.gcg.com> and fastDNAmI:ftp at <ftp://rdp.life.uiuc.edu>) for phylogenetic analysis and have no restrictions on usage of these packages. PHYLIP is also available on the Web through the WebPhylip interface developed at the National University of Singapore (<http://sdmc.krdl.org.sg:8080/~lxzhang/phylic>). The PISE system allows access to PHYLIP, fastDNAmI, MOLPHY, BIONJ, TREE-PUZZLE, Seq-Gen, TreeAlign, QR2, PhyloQuart, LVB, and BAMBE. Other sites such as the Computational Biochemistry Group of the ETH in Zürich, Switzerland (<http://cbrg.inf.ethz.ch/>) has the TreGen, AllAll, Multalign and Darwin programs available.

Tree Evaluation and Visualization of Results

When a phylogenetic tree is created, the next step is to evaluate the tree for robustness. The most common method for

doing this is bootstrap analysis. Some sites allow bootstrapping using the PHYLIP package (e.g., BioNavigator, WebPhylip and PISE), which essentially involves resampling the dataset and then analyzing the resampled data. Phylogenies that are well supported by the alignment will tolerate variations in the data brought about by the resampling process, although alignments that do not have a consistent evolutionary history along their length will give rise to different trees when small changes are introduced through resampling. In addition, it is often useful to provide a user tree as a guide for phylogenetic analyses. This permits a narrowing down of the tree space that has to be evaluated, an advantage that greatly speeds up the analysis.

User trees can also be used to compare different trees and to evaluate whether they are significantly different from each other, given the experimental evidence. This is particularly important when the "best" tree found by a program appears to contradict another tree derived from different evidence. In such a case, it is important to compare the two trees to ensure that the data clearly supports one tree over the other using a Kishino-Hasegawa-Templeton (KHT) test. The BioNavigator system's Web implementation of PHYLIP allows phylogenetic analysis with user trees as well as tree comparison using the KHT test.

While some phylogeny programs create rooted trees in which the location of the common ancestor is inferred from the data, this inference requires assumptions that are often regarded as unacceptable, such as that of a molecular clock. Without assumptions like these, most programs will generate unrooted trees where only the relative topology of the taxa is known. After producing such a tree it is often necessary to "root" it, based on alternative evidence (for example, morphological evidence).

When trees are created they need to be manipulated and visualized. The FragTree (<http://193.50.234.246/~beaudoin/ans/Tree.html>) and Phylodendron (<http://iubio.bio.indiana.edu/treeapp/treeprint-form.html>) servers allow such visualizations, as does the BioNavigator system. BioNavigator allows online graphics viewers, relabeling and drawing of both rooted and unrooted trees in a range of formats, and basic manipulation of trees (including flipping branches and re-positioning of the root of the tree). BioNavigator also allows the launching of helper applications such as TreeView (<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>), which can be launched directly from the browser to add significantly improved interactive tree viewing and manipulation.

Conclusion

There are many steps to phylogenetic analysis requiring databases, local data repositories and applications. Phylogenetic analysis tends to be an iterative, complicated process, and it is essential that the researcher doing the analysis has a good understanding of the methods to use and their potential pitfalls. One of the strengths of the PHYLIP package is its extensive documentation that provides a detailed description of the assumptions that are behind the various approaches. Integrated systems such as BioNavigator are ideal for phylogenetic analysis because they allow a flow of information all the way from database query through to a refined, visualized tree. In addition, sites that provide step-by-step protocols such as those found on BioNavigator greatly assist the new or infrequent user in identifying and choosing the right approach for such complicated analyses.