Phylogenetics on the Web

Stuart M. Brown, Ph.D., NYU School of Medicine, New York, NY, USA (browns02@mcrcr.med.nyu.edu)

Anyone who spends a little bit of time working with DNA and protein sequences quickly starts wondering how these sequences are related. This often leads to the more fundamental question of how the organisms from which these sequences come are related, which is the basis for molecular taxonomy or phylogenetics.

The essence of taxonomy is choosing heritable traits (characters) that accurately represent groups of organisms, i.e., all organisms with feathers and wings are classified as birds. Taxonomy has the mystique of an old discipline, practiced in musty rooms in the sub-basement of museums where white-haired scientists spend their time staring through magnifying glasses at jars of pickled specimens. However, taxonomists were early adopters of the new technologies of molecular biology because DNA (and protein) sequences have many advantages over more classical types of visible taxonomic characters.

- Character states can be scored unambiguously (genotypes can be read directly rather than relying on phenotypes);
- Huge numbers of characters can be scored for each individual;
- Information can be obtained on both the extent and the nature of divergence between sequences (nucleotide substitutions, insertion/deletions or genome rearrangements);
- Comparisons can be made between groups of organisms (populations) that have minimal phenotypic differences;
- Phylogenies can be built for groups of organisms that are so widely diverged that they do not share many morphological traits.

It seems logical that for a given gene, closely related organisms have similar sequences and more distantly related organisms have more dissimilar sequences. However, it is necessary to consider phylogenetic analysis, conducted with molecular data, in the context of the full body of taxonomic and biological knowledge. If calculated correctly, relationships derived from sequence data actually represent the relationships between genes, which is not necessarily the same as relationships between whole organisms. The sequences that were used for data may not have had the same phylogenetic history as the species within which they were contained. Different genes (and different species) evolve at different speeds, and there is the possibility of horizontal gene transfer either by hybridization, vector-mediated DNA movement or direct uptake of DNA.

Within the field of phylogenetics, there are two different, yet hotly contested and philosophically opposed methods of building trees—phenetics and cladistics. These two different approaches to the interpretation of sequence data lead to different computational algorithms and, ultimately, to different phylogenies computed from the same data.

The phenetic approach constructs trees by considering the overall phenotypic similarities among species without trying to understand the evolutionary history that has brought the species to their current phenotypes. Sequence data is used to classify organisms into trees based on the absolute number of characters that they share. Distances between DNA sequences are simply the sum of all base pair differences between two aligned sequences. Distances between aligned protein sequences are calculated using both amino acid identities and similarities (calculated from a scoring table such as the PAM or BLOSUM matrix).

Computer phylogeny programs based on the phenetic model use a matrix of pairwise distances and simple clustering algorithms such as the unweighted pair group method using arithmetic averages (UPGMA) and Neighbor Joining to build trees. These methods build clusters by first grouping pairs of sequences with the fewest differences, adding the next most distant and then repeating this process until all sequences are included.

The cladistic approach is based on the study of clades, which are groups of related organisms. Evolutionary trees are “reconstructed” by considering the various possible pathways of evolution (branching points where groups diverged from a common ancestor) and choosing the best possible tree from among these. Known ancestral relationships are considered as well as current data. Instead of looking at all the phenotypic characters of organisms, cladistics uses only a subset of characters that are uniquely shared by a group of organisms and not present in their ancestors. Computer algorithms based on the cladistic model generally rely on parsimony (the least number of evolutionary changes) or maximum likelihood (the most probable ancestors) methods for the calculation of relationships and the building of trees.

For data such as classical morphological characters and for higher (or perhaps we should say deeper) levels of taxonomy, the cladistic approach is almost certainly superior. However, cladistic methods are often difficult to implement, and require assumptions that are not always satisfied with molecular data. Molecular data make pure “phenotypic” characters since there is no inherent reason for a researcher to think that one mutation is more important or more “ancestral” than another. Phenetic approaches lead to faster algorithms and often have nicer statistical properties for molecular data.

The starting point of a phylogenetic analysis is usually a set of related proteins. It is possible to make phylogenies from DNA sequence, but unless you are dealing with either noncoding sequences or a set of very closely related sequences (intra-species), it is much more informative to work with proteins; much more distant relationships can be analyzed.

Once a list of related proteins is collected, the next step is to build a multiple alignment. It is important to align only homologous regions of the sequences. In some cases,
BioBit: Taxonomy and Phylogenetics Information on the Web

There is a wealth of information about taxonomy and molecular evolution available on the Web.

For general background reading, explore the University of California Museum of Paleontology’s Journey Into Phylogenetic Systematics at http://www.ucmp.berkeley.edu/clad/clad4.html

The Tree of Life Project at the University of Arizona gives a good overview of the overall process of taxonomy and is found at http://phylogeny.arizona.edu/tree/phylogeny.html

Joe Felsenstein, the author of PHYLIP, maintains an extensive Web site that offers free downloads of PHYLIP software for all computing platforms, PHYLIP documentation and FAQs and information about 160 other phylogeny software packages from http://evolution.genetics.washington.edu/phylip.html

The Willi Hennig Society maintains a Web site that contains an index of all articles from the journal Cladistics, a searchable database of phylogenetics literature and an extensive list of phylogenetics software at http://www.vims.edu/~mes/hennig/hennig2.html

The Society of Australian Systematic Biologists has posted two excellent review articles titled “Introduction to Phylogenetic Systematics” by Peter H. Weston and Michael D. Crisp, and “Introduction to Some Computer Programs Used in Phylogenetics” by Michael D. Crisp. They are reachable at


An excellent introductory manual to phylogenetic analysis (using cladistics) by Diana Lipscomb, George Washington University is at http://www.gwu.edu/~clade/faculty/lipscomb/Cladistics.pdf


distantly related genes may share homology in some regions, but not in others. Phylogenetic analysis programs require a multiple alignment with no large gaps. It is also extremely important to have all the sequences in the alignment be of the same length. If a list of proteins includes some partial sequences, the decision will have to be made whether to eliminate these shorter sequences or to truncate all of the other sequences and make an alignment of only the region that is shared.

Once the sequences have been aligned, the multiple alignment file becomes the input for a phylogenetic analysis program. Phylogenetics programs are available for both desktop computers (Macs and PCs) and for mainframes. A few computing centers have recently begun to offer free phylogenetics servers on the Web. This is quite brave, considering the potential for these programs to consume huge amounts of computing power if large numbers of long sequences are used. The amount of computation is also dependent on the type of algorithm used: distance methods (UPGMA and Neighbor Joining) are quick, parsimony methods take longer, and maximum likelihood methods are very slow.

The Institut Pasteur, Paris has a PHYLIP server at http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

Joaquin Dopazo at the Centro Nacional de Biotecnologia, Madrid, Spain has made a nice server interface for PHYLIP at http://www.cnb.uam.es/cgi-bin/dopazo/PHYLIP/phylip

The Belozersky Institute at Moscow State University in Russia has its own “GeneBee” phylogenetics server at http://www.genebee.msu.su/services/phtree_reduced.html

The last challenge for phylogenetic analysis is to produce a publication-quality printout of the results. The PHYLIP drawtree program is rather poor in this regard, both in terms of the user interface and in the quality of the final product. The Phylodendron Web site is a tree drawing program with a better user interface and many formatting options, including PDF. Whatever program is used, the final tree may need to be redrawn in a graphics program: http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

Phylogenetic analysis of protein families is one of the most challenging and interesting tasks in bioinformatics. This type of analysis is rarely done and, therefore can often make a valuable contribution to a research area. For example, new genes are being discovered on a daily basis, but little is being done to establish the evolutionary relationships between these genes and their homologs. A biologist with an interest in phylogenetic analysis could make a significant contribution to the annotation of these sequences.