Phylogenetic analysis & comparative genomics

Molecular sequence analysis is an evolving field important to phylogenetic analysis and comparative genomics. Here, some of the pitfalls and practical solutions in this young but increasingly important area of comparative genomics, together with its relationship with bioinformatics, are discussed.

James A. Lake and Jonathan E. Moore

232 Molecular Biology Institute
and MCD Biology, University of
California Los Angeles, Los Angeles,
CA 90095, USA.

lake@mbi.ucla.edu
moore@mbi.ucla.edu

Everyone has an idea of what a phylogenetic tree is, but there are two specialized concepts that are often applied to trees. These are the concepts of **rooted** and **unrooted** trees. A rooted tree corresponds, more or less, to everyone's idea of a tree. Typically, the ancestral state of the organisms, or genes, being studied are shown at the bottom of the tree, and the tree branches, or bifurcates, until it reaches the terminal branches (or tips, or leaves) at the top of the tree. There is nothing sacred about this convention, however, and trees can also be drawn with the tips at the bottom, at the left, at the right or even at a 45° angle. An unrooted tree is a less-intuitive, more-abstract concept. Unrooted trees represent the branching order, but do not indicate the root, or location, of the last common ancestor. Ideally, rooted trees are preferable, but, in practice, virtually every phylogenetic reconstruction **algorithm** provides an unrooted tree; thus, one needs to become familiar with them.

Pitfalls of comparative genomics

It is not generally appreciated that molecular sequence analysis is a field in its infancy. It is an inexact science in which there are few analytical tools that are truly based on general mathematical and statistical principles. Consequently, many, perhaps most, phylogenetic trees reconstructed from molecular sequences are incorrect and frequently conflict with common sense. This is mainly caused by one or more of the three pitfalls of sequence analysis: (1) incorrect sequence alignments, caused by inadequate mathematical models and often related specifically to biases created by progressive alignment algorithms when they are used to align more than three **taxa**; (2) the failure to account properly for site-to-site variation (all sites within sequences can evolve at different rates); and (3) unequal rate effects (the inability of most tree-building algorithms to produce good phylogenetic trees when genes from different taxa in the tree evolve at different rates). All three pitfalls can produce the same artefact – **long branch attraction**. In these artefactually produced trees, rapidly evolving sequences (represented by long branches on phylogenetic trees) will be placed with other rapidly evolving sequences, even if the sequences are only distantly related. In comparison with most problems in molecular biology, which can be solved by acquiring more data, long branch attractions are much more complex. If longer sequences are used when long branch attractions are present, the incorrect solution will be even more strongly supported. Of the three pitfalls, alignment artefacts are potentially the most serious, because even if the second and third problems are solved, the misalignments can still produce incorrect trees. A new algorithm, **paralinear (logdet) distances**[1,2], provides a simple, but rigorous, mathematical solution for the third pitfall. This particular algorithm is now available in some of the phylogenetic packages described below. (For a discussion of many other useful algorithms that are available, including maximum parsimony, maximum likelihood and other distance methods, see Ref. 3.)

Practical suggestions

Although it was not realized until recently, long branch artefacts are very common. However, a few simple precautions can be taken to help reduce their frequency. First, always examine the sequence alignments before calculating evolutionary trees. If the alignments have lots of gaps, this could indicate that quite diverse sequences have been included in the alignment and that the alignment might be incorrect. In this case, the results should be viewed with extreme caution. A check should be made to see if the gaps are caused by all the sequences or by just one or a few deviant sequences. In the latter case, the deviant sequences should be removed.

Second, if a sequence really must be included in a tree and the sequence does not align well, consider obtaining the sequence of the same gene from a closely related, but more slowly evolving, relative. Recently, in

0167-7799/98/$ – see front matter

a collaboration[4] between our laboratory and those of Garey, Raff and Turbeville, the 18S ribosomal RNAs of nearly 20 nematodes were sequenced to find a slowly evolving 18S nematode gene. As a result, we were able to demonstrate that *Drosophila* and *Caenorhabditis* are close relatives, as are all molting animals. For years, nematodes had been considered to branch deeply within the metazoan animals, because the rapidly evolving 18S ribosomal RNA genes from *Caenorhabditis* had artefactually led to an incorrect tree. As a result, we now have a much better idea of the relationships among metazoan animals[5], which is important when relating the results from the fly and worm genome projects to the human genome.

Finally, be aware when the tree being studied conflicts with trees from other genes. If another gene, which appears to be evolving more slowly, gives a different tree, then perhaps long branch attraction is causing the discrepancy. In which case, consider looking at a third gene.

How do you know if the results are statistically significant?

Bootstrapping is a commonly used procedure for estimating the statistical significance of individual branches within a tree, yet few people are clear about exactly what it is or how to interpret the results. First, it is necessary to understand the concept of sampling with replacement. This is related to the practice of estimating the number of fish in a lake by collecting a certain number of them, tagging them, releasing them and then, several weeks later, repeating the process and measuring the fraction of tagged fish. By knowing the fraction of fish that were previously tagged, one can estimate the total number of fish in the lake. Bootstrapping uses a set of aligned sequences to estimate what the sequence would be like if it had been infinitely long. In this instance, each column in the alignment (referred to as a sequence pattern) plays the role of a fish, and sampling with replacement is used to create a number of artificial sequence alignments (usually 100). Trees are calculated from each of these, and the frequency with which various branching patterns are observed within these trees is noted. If a particular branching pattern is observed 70% of the time, this branching pattern is said to have 70% bootstrap support. The exact statistical interpretation of bootstrap results is still an active subject of study, but the 'rule of thumb' is that internal tree branches that have >70% bootstrap support are likely to be correct at the 95% level[6]. (Even so, a high bootstrap percentage still does not guarantee that long branch attractions have not biased the results.)

The programs

There are several excellent programs to help calculate trees from genome data. The addresses of Web sites providing more information on these programs are listed in the URLs box. The best known software for reconstructing trees is the program PAUP (phylogenetic analysis using parsimony). This program is extremely user friendly and comprehensive and has recently been released as part of the GCG sequence analysis package. In addition, PAUP can perform paralinear distance (logdet) analyses and thus can produce results free of the third of the three analysis pitfalls. PHYLIP is another well-known package that contains a large variety of routines, including several that incorporate the latest theoretical developments. One might also like to look at Hennig86. MEGA/METREE contains innovative approaches for handling pitfall number two – site-to-site variation. GAMBIT (short for bootstrappers gambit[7]) is a new program that is available as a beta test version; it can handle both site-to-site variation, using a new general algorithm (pattern filtering)[8], and pitfall number three, using paralinear distances. Finally, for manipulating trees, MacClade has impressive capabilities. I suggest that you try them and join in the fun.

References

1 Lockhart, P.J. *et al.* (1994) *Mol. Biol. Evol.* 11, 605–612

2 Lake, J.A. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 1455–1459

3 Stewart, C-B. (1993) *Nature* 361, 603–607

4 Aguinaldo, A.M.A. *et al.* (1997) *Nature* 387, 489–493

5 Balavoine, G. and Adoutte, A. (1998) *Science* 280, 397–398

6 Hillis, D.M. and Bull, J.J. (1993) *Sys. Biol.* 42, 182–192

7 Lake, J.A. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 9662–9666

8 Lake, J.A. *Mol. Biol. Evol.* (in press)

URLs...

PAUP

http://onyx.si.edu/PAUP/

GCG package

http://www.gcg.com/

PHYLIP

http://evolution.genetics.washington.edu/phylip.html

Hennig86

http://www.vims.edu/~mes/hennig/software.html

MEGA/METREE

http://www.bio.psu.edu/faculty/nei/imeg

GAMBIT

http://www.lifesci.ucla.edu/mcdbio/Faculty/Lake/Research/Programs/

MacClade

http://phylogeny.arizona.edu/macclade/macclade.html