Evolution

Perils of molecular introspection

Joe Felsenstein

THE rising flood of nucleic-acid sequence data has reopened old controversies on the evolutionary affinities of organisms, controversies that once seemed unresolvable. Sequences are being collected to resolve century-old tangles in the relationships of groups such as the lower invertebrates, the angiosperms and the protists, and even to see back 3,000 million years to the interconnections of the oldest bacterial groups. The phylogeny of the apes has generated the most controversy of all. Holmquist and colleagues[1,2] now back up the claim, discussed by Diamond in a recent News and Views article[3], that chimpanzees are more closely related to humans than to gorillas. They do this by analysing three data sets totalling 10,393 bases — sufficient to raise the question of whether new methods of analysis, such as Lake's method[4] of "evolutionary parsimony", could resolve the issue.

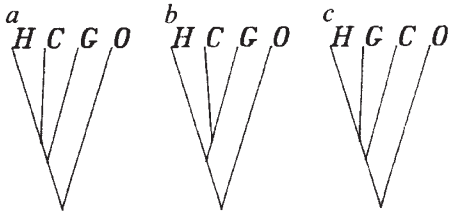
Two computational methods have dominated the reconstruction of molecular phylogenies: parsimony and distance. The parsimony method finds the evolutionary tree that requires the fewest changes of nucleotides to explain evolution of the observed sequences. Distance methods compute a table of pairwise numbers of differences between sequences and try to fit this to expected pairwise distances computed from the tree.

Both of these methods look at only part of the information in the data. For example, parsimony considers sites uninformative unless a different number of nucleotide changes are required on different trees. To use the full information present, maximum-likelihood methods have been developed[5]. These, however, are computationally difficult and are tied to a precise probability model of sequence change.

Lake's method is intended to cut the Gordian knot. It is simple to carry out and aims to avoid the problems which have plagued the other approaches. His application of it[6] to resolving the affinities of the Archaebacteria has been reviewed in the News and Views article by Penny[7]. Holmquist *et al.*[1,2] use it to attack the difficult problem of ape phylogeny and find that humans and chimpanzees are the most closely related, the data having only a 3 per cent probability of showing a pattern this strong by chance.

Although Lake calls his method evolutionary parsimony, it is more closely related to likelihood methods. A similar method was independently developed by Cavender[8], who coined the more precise term invariants. Invariant methods start with nucleotide sequences for four species. Assuming that sites are evolving independently, they classify each site in the sequence into one of 256 possible patterns. In alphabetical order, these would be AAAA, AAAC, . . . TTTT where the four letters are the bases seen in the four species.

Any probability model of independent evolution at different sites predicts frequencies for these 256 classes. The form of the expressions depends on the shape of the true tree and the values depend on parameters including the branch lengths in the tree. Maximum likelihood methods are in effect ones which try to fit these 256 frequencies as closely as possible. Invariants try instead to detect certain of the regularities in the expected pattern frequencies.



The three phylogenetic trees that are in contention for the African apes (human, H, chimpanzee, C and gorilla, G). Most interest centres on a *and* b. *The orang-utan, O, is included for comparison.*

Lake's invariants are sums and differences of expected frequencies of some of these patterns, expressions which will each be zero under two of the trees in the figure and not under the third. They achieve the status of invariance by being zero whatever the branch lengths in the tree.

Holmquist *et al.* do this tabulation for three nucleotide sequences for orang-utan, gorilla, chimpanzee and human. Of two classes of sites that are expected to have equal frequency of occurrence under the trees *b* and *c* in the figure, they find six sites in one class and none in the other. The occurrence of all six sites in one class is an event that has only 1 chance in 64 of occurring if trees *b* or *c* are correct, so that these alternatives to the human–chimp tree can be rejected.

Conventional parsimony would pay attention to a somewhat different collection of sites, the 'phylogenetically informative' sites that have only two nucleotides, each occurring twice. These fall naturally into three groups that each support one of the three trees. There are found to be 25, 13 and 16 of these, respectively. This again favours the human–chimp tree but less strongly than do the invariants. Various statistical tests based on conventional parsimony[9-11] fail to find statistical significance in this part of the data.

One advantage of Lake's invariants is that they are not misled by unequal rates of change at different sites and by unequal amounts of evolution in different branches of the tree. Their computational simplicity also makes them attractive to researchers weary of computers. But they too have their restrictive assumptions. It is critical to Lake's algebra that when a site which is now A undergoes a transversion, it must be equally likely to end up as a C or a T. It is not clear whether modest violations of this will be difficult for Lake's method.

Terms such as 'phylogenetically informative' mislead us by implying that all the relevant information is in a few sites. In fact, all sites contribute information. For example, if the human and chimpanzee are related and equally diverged from their ancestor, we would expect more sites with patterns like AACA than with patterns like ACAA, which is what is found. Both Lake's invariants and parsimony ignore this information, but likelihood and distance approaches do not.

We can either use all the information with a highly specific evolutionary model, as likelihood methods do, or trade some of that information for robustness by looking at a smaller subset of the data, as invariants, parsimony and distance methods each does in different ways. The addition of invariants to the phylogenetic arsenal at least seems to be persuading molecular evolutionists to take a broader look at the assumptions of their methods.

Even with the new analysis provided by Holmquist *et al.*[1,2], the ape issue is still far from being resolved. It is becoming clear that the phylogenetic tree of the African apes (human, chimpanzee and gorilla) is nearly a three-way split. Few physical anthropologists will be bowled over by a single test that merely reaches the 3 per cent level of significance. Nevertheless, that is the best yet done with sequence data. We are perhaps trying to squeeze blood out of stones but at least the stones are getting more numerous. Holmquist *et al.* show that they can be made somewhat moister as well. □

1. Holmquist, R., Miyamoto, M.M. & Goodman, M. *Molec. biol. Evol.* **5**, 201–216 (1988).
2. Holmquist, R., Miyamoto, M.M. & Goodman, M. *Molec. biol. Evol.* **5**, 217–236 (1988).
3. Diamond, J.M. *Nature* **332**, 685–686 (1988).
4. Lake, J. A. *Molec. biol. Evol.* **4**, 167–191 (1987).
5. Felsenstein, J. *J. molec. Evol.* **17**, 368–376 (1981).
6. Lake, J. A. *Nature* **331**, 184–186 (1988).
7. Penny, D. *Nature* **331**, 111–112 (1988).
8. Cavender, J.A. & Felsenstein, J. *J. Classif.* **4**, 57–71 (1987).
9. Felsenstein, J. *Syst. Zool.* **34**, 152–161 (1985).
10. Felsenstein, J. *Evolution* **39**, 783–791 (1985).
11. Templeton, A.R. *Evolution* **37**, 221–224 (1983).

Joe Felsenstein is a professor in the Department of Genetics, University of Washington, Seattle, Washington 98195, USA.