Making family trees from families

Sarah A. Teichmann and Graeme Mitchison

MRC Laboratory of Molecular Biology,
Hills Road, Cambridge CB2 2QH, UK

Since the time of Linnaeus, phylogenetic trees have been constructed using similarity of morphological characters to assess the relationships of organisms. From the early sixties onwards, another method has gradually taken precedence; this uses the bases in nucleic acids or amino acids in proteins as the characters for tree-building (Zuckerkandl & Pauling, 1962). Today's genome projects provide vast amounts of sequence data, and might therefore be expected to give a solid foundation for this kind of molecular taxonomy. The organisms whose genomes are now complete are all bacteria (with the exception of yeast), so one might suppose that bacterial taxonomy in particular would have become more secure. Oddly enough, this is far from true. More data has meant more confusion, because different genes often define different trees. Snel *et al*. (this issue of Nature Genetics) have tried to tackle this problem using a new approach based on gene families in completely sequenced genomes.

One reason for the confusion is that some of the events of phylogenetic interest in bacterial evolution happened as long as 3 thousand million years ago (Feng *et al*., 1997). The molecular traces of such ancient events may have faded to the extent that they are difficult to detect. The other reason is horizontal transfer. Bacteria are scavengers and opportunists, and they can pick up DNA from other organisms and incorporate it into their genomes. This could have occurred (and could still be occurring) on a large scale. One estimate says that *E. coli* acquires DNA from transfer at about the same rate as new DNA is introduced by point mutation, namely about 30kb per million years (Lawrence & Ochman 1998). Can one still talk of genetic identity under these conditions?

One answer is that the genes may have hopped about, but the way an individual bacterial species *uses* the genes is what constitutes an organism's real identity. After all, we are 98% identical to chimpanzees at the DNA level, and so our genes are all very similar; yet the ways those genes are used lead to striking differences (to our eyes, at any rate). Can we identify such higher patterns of organisation in bacteria? Unfortunately, the precise structures and functions of bacterial proteins, as well as the control pathways and networks, are still little known at the level of specific complete bacterial genomes. Snel *et al*. therefore propose a more feasible strategy which still captures something of the grand plan: they look at the extent to which genes are shared in different organisms. The definition of "sharing" needs some care.

The sequence produced by genome projects does not come with its genes labelled, of course. Various computational methods identify likely genes, and these can be clustered into families of similar genes. For instance, a bacterium might have twenty sugar transporters, all quite closely related in sequence. Suppose two bacteria each have a cluster of the same family. Snel *et al*. define two genes in two organisms to be shared if they form a closest pair, with neither gene having a closer match in the other organism (Figure 1). "Closest" here means highest-scoring, according to a pairwise alignment algorithm like Smith-Waterman (Smith & Waterman, 1981). There is a further condition that the matches should be sufficiently high-scoring to be reliable.

Now imagine that the common ancestor of both bacteria possessed several genes in the family. If one followed evolution forward for a short time from the division of that ancestral line into two species, each gene in the descendant organisms would be likely to be shared; pairs of genes coming from the same gene in the ancestor would have diverged in the descendants, but not too far not to be each other's closest match. As time progressed, genes would be lost, or duplicated, and then the proportion of genes in each family that are shared would decrease. Thus the fraction

$$f=1-(\text{genes shared})/(\text{total number of genes})$$

should measure the evolutionary distance since speciation, which is what is needed to build a tree.

Is the resulting tree less likely to suffer from the faults of traditional methods, which measure distances between sequences in terms of the substitutions of one residue by another? If long evolutionary times are the cause of a weakening signal in traditional methods, it is not obvious that counting shared genes (as defined by Snel *et al.*) will be more reliable, since confident detection of matches needs good preservation of sequence homology.

Horizontal transfer, on the other hand, might affect the number of shared genes less than traditional methods. The argument would go as follows: when a gene duplicates, it is likely that evolution will act on the duplicate to confer on it some new function, as a result of which it will rapidly diverge. Thus the shared genes may be those with common functions. If horizontal transfer occurs mostly between genes with the same function, then it will replace a shared gene by another shared gene, (provided the transferred gene is close enough to be recognised). Thus horizontal transfer will not affect the shared-genes distance. However, it is possible that horizontal transfer often brings in genes with new functions to organisms, in which case it will alter the count of shared genes and will affect trees much as it does with traditional methods. The fact that the shared-genes tree broadly agrees with the traditional 16S rRNA tree (Olsen *et al.*, 1994) suggests that horizontal transfer either occurs mostly between genes with the same function or is largely confined to closely related organisms and occurs infrequently between distantly related organisms. This is a valuable finding, as it throws some light on recent speculations about horizontal transfer (Woese, 1998).

Does the tree one gets from the shared-genes distance f provide any answers as to bacterial phylogeny? Yes and no. The traditional molecular method uses 16S rRNA, a molecule that is perhaps less readily transferred between organisms because it is essential, exists in many copies in the genome and has specific interactions with other molecules. Nonetheless, the rRNA molecule carries only limited information, and does not give a reliable tree for distantly related bacteria. For most of the bacterial phyla, the shared-genes tree also has low confidence and hence these uncertain regions of the rRNA tree are not resolved in the shared-genes tree either. There is a difference in the archaebacterial part of the tree, where the shared-genes tree has well separated branch nodes with high bootstrap values (a measure of statistical confidence) which differ from the corresponding region of the rRNA tree; perhaps the new method is giving extra information here.

What are the prospects for this new approach? In so far as shared genes reflect shared functions, they are telling us something about the overall makeup of the organism and therefore should in principle give a more reliable measure of relatedness over long evolutionary times. It is not clear that the current method can exploit this advantage for the reasons given earlier, namely that the criterion for matching proteins requires the sequences to be quite closely related (E value< 0.01). One would like to be able to identify more distantly related shared genes, and this will become possible as massive structure-determining projects ("structural genomics" projects) get under way. When many of the proteins in a genome can be compared structurally (Teichmann *et al.*) and their likely functional similarity assessed, the shared-genes method and its intellectual descendants will become part of the phylogeneticists' kit.

Figure Legend

The lines represent genes in a family, and the two-headed arrows indicate shared genes. Note that not every gene has a match. Suppose gene G in organism 1 has a closest match G' in organism 2, but the closest match of G' is not G. Then G and G' do not satisy the condition of being each other's closest macht. Another reason for lack of a match for G in organism 2 is that it may be too diverged from all the genes in organism 2 to make a reliable match (satisfying E value < 0.01).

References

Zuckerkandl, E. & Pauling, L. *Molecular disease, evolution and genetic heterogeneity.* In Marsha, M. and Pullman, B., eds., *Horizons in Biochemistry*, Academic Press, pp. 189-225 (1962).

Feng, D.-F., Cho, G & Doolittle, R.F. *Proc. Natl. Acad. Sci. USA* **94**, 13028-13033 (1997).

Lawrence, J.G. & Ochman, H. *Proc. Natl. Acad. Sci. USA* **95**, 9413-9417 (1998).

Snel, B., Bork, P. & Huynen, M.A. *Nature Genet.* (1999).

Smith, T.F. & Waterman, M.S. *J. Mol. Biol.* **147**, 195-197 (1981).

Olsen, G.J., Woese, C.R. & Overbeek, R. *J. Bacteriol.* **176**, 1-6 (1994).

Woese, C.R. *Proc. Natl. Acad. Sci. USA* **95**, 6854-6859 (1998).

Teichmann, S.A., Park, J. & Chothia, C. *Proc. Natl. Acad. Sci. USA* **95**, 14658-14663 (1998).

Figure 1: Mapping shared genes