Letter to the Editor

Evidence from *Aeromonas* for Genetic Crossing-Over in Ribosomal Sequences

In a recent issue of the *International Journal of Systematic Bacteriology*, Martinez-Murcia and colleagues (8) report complete 16S rRNA sequences from *Aeromonas*, each consisting of 1,502 bp which offer no problems of homology. There is no exact correspondence of sequence with species or with strains, and so these entities are referred to by the sequence numbers in their Fig. 1. These authors noted several inconsistencies with the relatedness of aeromonads as obtained from the results of DNA-DNA pairing. Examination of their data suggests yet another type of inconsistency, a departure from the expected pattern of evolution in ribosomal sequences. Ribosomal sequences, particularly 16S rRNA, are much used in reconstructions of bacterial phylogeny. It has been pointed out (6) that the numbers of differences between sequences may be insufficient to give satisfactory resolution within bacterial genera. This may be termed lack of sensitivity at the species level. The present contribution points out that there is also evidence for another problem, the occurrence of hybrid events in the form of gene crossovers.

The reconstruction of bacterial phylogeny by these authors from 16S sequences (their Fig. 4) shows a tree with very uneven tips. Molecular evolution, both of protein and nucleic acid sequences, is broadly divergent and constant in rate in different lineages (3, 5). The comparison of sequences that have been aligned properly so as to give convincing homologies between the sites therefore leads ideally to phylogenetic trees whose tips lie at the same level, as in phenograms. Such relationships are described as ultramet-

ric. Departures from this relationship are caused by factors such as inconstant evolution rates and parallel or convergent evolutionary changes. Their figure is poorly ultrametric and seemed to me to be more discrepant than chance effects would imply. The data in their Table 2 also suggested poor ultrametricity, in that some triads (e.g., sequences 1, 4, and 9) do not fit well into an ultrametric pattern (see below).

There are 48 variable sites in *Aeromonas* (30 are in pairing regions of the molecule, which reduces the potential for variation to $48 - 15 = 33$ independent changes, but this is not, I believe, germane to the present analysis). When the 48 variable sites were copied and represented graphically (Fig. 1), there appeared to be a difference between the patterns of the left part and the right part if the sequences were divided between bases 264 and 457, as indicated by the dashed line. Thus, sequences 4 and 5 (*Aeromonas caviae*) were identical to sequence 1 (*A. hydrophila*) in the left part but very different in the right part. Conversely, sequences 2 and 3 (*A. salmonicida*) were very different from sequence 1 (*A. hydrophila*) in the left part but quite similar in the right part. This suggested that hybrid events might have occurred in the ancestry of *Aeromonas* species.

The sequences fall into five major patterns, A to E (Fig. 2a). Figure 2b shows how a gene crossover between a parent of pattern A and a parent of pattern D would yield a hybrid with a B-like pattern. Indeed, if the representatives of patterns A and D are sequence 1 (*A. hydrophila*) and sequence 9 (*A. jandei*), respectively, the resulting crossover would give a sequence identical, except for two positions, to

FIG. 1. Diagram of the variable sites in *Aeromonas* 16S rRNA from Fig. 1 of Martinez-Murcia et al. (8). Sequence numbers with abbreviated names are shown on the left. The lightly stippled areas are nucleotides in sequence 1; the hatched areas are the alternative sites, except where there are three alternatives, in which case the rarest is darkly stippled. The sites are numbered according to the *Escherichia coli* scheme. The nucleotides are indicated against the key at the bottom. The species names are as follows: 1, *A. hydrophila*; 2, *A. salmonicida* subsp. *salmonicida*; 3, *A. salmonicida* subsp. *achromogenes* and subsp. *mausocida*; 4, *A. caviae* group 1; 5, *A. caviae* group 2; 6, *A. media*; 7, *A. eucrenophila*; 8, *A. sobria*; 9, *A. jandei*; 10, *A. veronii*; 11, *Aeromonas* sp. strain ATCC 35941; 12, *A. schubertii*; 13, *A. trota*. The main regions among these variable sites that are involved in secondary structure are as follows: sites 131 to 132 pair with sites 231 to 230, sites 154 to 156 pair with sites 167 to 165, and sites 457 to 462 pair with sites 475 to 470, respectively. There are also single pairs: 76 with 93, 839 with 847, 1011 with 1018, 1309 with 1329, and 1355 with 1367. There are also numerous pairing regions between sites that do not vary within *Aeromonas* and which are not shown in this figure. In particular, there are no differences between the *Aeromonas* strains in sites 265 to 456, so that crossovers within this region would not alter the resulting sequence from sites 265 to 456 and would not affect the secondary structure.

FIG. 2. (a) The major patterns of sequences (see text). The sets of sequences are as follows: A, sequences 1 and 6; B, sequences 4, 13, and 5; C, sequences 2, 3, 7, 8, and 11; D, sequences 9 and 10; E, sequence 12. Shading corresponds to that in Fig. 1. Sites where sequences vary considerably are indicated with a question mark. (b) Illustration of how a crossover at the dashed line (as shown by arrows) between a sequence of pattern A and one of pattern D could yield a B-like sequence.

that of sequence 4 (one form of *A. caviae*). There are some other plausible crossovers (mentioned later), and so there may have been more than one crossover event, and if these events occurred some time ago in the evolution of the genus, the picture could have become somewhat obscured by later evolutionary changes. First it is convenient to consider the statistical evidence for a crossover as exemplified by the case of sequences 1 and 9 yielding sequence 4.

If one tabulates separately the numbers of nucleotide differences in the left and right parts one obtains for the comparisons 1 and 4, 1 and 9, and 4 and 9 the following values, respectively: for the left part, 0, 13, and 13; for the right part, 14, 12, and 2. If one considers the left part and clusters the sequences, 1 joins 4 first (no differences) and sequence 9 joins with linkage levels of 13 and 13 differences, which is a perfect ultrametric relationship. The values from the right part, however, first join sequences 4 and 9 (two differences), and then sequence 1 joins at linkage levels of 14 and 12 differences, which is almost perfect ultrametric behavior. However, if the total differences for both left and right parts are considered, the values for the comparisons of 1 and 4, 1 and 9, and 4 and 9 are 14, 25, and 15, respectively. These values do not fit well to an ultrametric relationship. If 1 is first joined to 4 (14 differences), then 9 links to them by reason of two very different values, 25 and 15, whereas an ultrametric relationship would require both values to be the same (i.e., the average, 20).

These relationships also imply that if a minimum-length tree is constructed from the total differences, then sequence 4 is carried on a very short stem from a fork that is near the presumed ancestral root, and so the tree is obviously poorly ultrametric. Furthermore, if left and right subsequences are shown separately, one obtains a diagram resembling Fig. 3a. From this it is evident that the major evolutionary changes occur in different subsequences in different lineages. Most of the change in the left subsequence occurs in the lineage from the fork to sequence 9, whereas, conversely, most of the change in the right subsequence occurs in the lineage from the fork to sequence 1. This is another expression of the way evolution here is not constant, and this is true wherever the ancestral root is placed. Figure 3a may be viewed as the orthodox evolutionary hypothesis, in which phylogeny is solely a branching process.

FIG. 3. Two evolutionary hypotheses for the sequences 1, 4, and 9. The left and right parts of the sequences are shown by the thin and thick lines, respectively. The numbers of nucleotide differences for each branch of the evolutionary triad are shown: changes in the left subsequence are in lightface type, and those in the right subsequence are in boldface numerals. The root of each tree is presumed to be near the curved lower part of the lineages. (a) Orthodox interpretation of a branching phylogeny. The total changes in left and right subsequences are shown by numerals in parentheses. (b) Interpretation as a genetic crossover. Sequence 4 arises as a hybrid event in which a fairly recent ancestor of sequence 1 donates the left part of the ribosomal sequence to a recent ancestor of sequence 9, and a recent ancestor of sequence 9 donates the right part.

The alternative hypothesis is that of a crossover, shown in Fig. 3b. The relationships in the subsequences are almost perfectly ultrametric. The changes in the subsequences are evenly distributed considering the small numbers involved. How strong is the evidence for Fig. 3b in comparison to Fig. 3a? The numbers of changes are small. If each were a hundred times as large, then there would be no doubt that Fig. 3b would be the more probable situation. But by using simplifying assumptions on evolution (5, 9) it seems likely that distortion as great as that in Fig. 3a due to chance would occur only about one time in fifty. Crossing-over should therefore be considered seriously. Also, as noted below, it seems likely that other hybrid events have occurred in the genus.

The thesis that the evolution of the left and right parts has not been concordant is supported by other evidence. First, the correlation of the differences between the sequences on the basis of left and right parts is quite low ($r = 0.20$). Three random divisions of the sites into 18 and 32 (the same numbers as the variable sites in the left and right subsequences, respectively, and subject to constraints imposed by the pairing regions) gave r values of 0.45, 0.50, and 0.68. The value of 0.20 is significantly lower than the mean of the random samples at $P = 0.10$ by one-tailed t and Fisher's z tests (with 2 degrees of freedom). Second, the scatterplot of differences from left and right subsequences showed more outlying points than scatterplots from random sampling, and the outlying points refer to sequences that may be involved in crossing-over (e.g., sequences 1 or 6 and 9 or 10 yielding 4, 5, or 13). Third, significance tests for evolutionary rate differences in left and right subsequences in plausible crossovers show low probabilities that the differences are due to chance. Thus, for sequences 1, 9, and 4 the probability that rate differences are due to chance is less than 0.01, and this test is independent of a test for ultrametricity, which yielded a P value of 0.02. Some other plausible crossovers are sequences 6 and 9, yielding a sequence similar to 4, and sequences 4 and 10, yielding a sequence similar to 9 (it may not be certain in such examples just which are the parental forms).

These phenomena are brought out clearly by cluster

FIG. 4. UPGMA phenograms (11) obtained from the left and the right parts of the ribosomal sequences. *D* is the number of nucleotide differences. The three main clusters in each are labelled (*a*, *b*, *c*, I, II, III).

analysis of the left and right subsequences separately (Fig. 4). Cluster analysis is not the same as reconstructing phylogenetic trees; it does not require the additional assumptions about the way in which evolution occurs that are needed for phylogenetic reconstruction. It is important to avoid such assumptions when (as is the case here) one is questioning the assumptions themselves. Instead, cluster analysis seeks to find the main patterns of variation whatever their cause (e.g., the halogens would be a cluster of similar chemical elements but not a phylogenetic group), and the UPGMA method has the advantage of being reliable and robust. There are three tight clusters in the UPGMA phenogram from the left subsequence, labelled *a*, *b*, and *c*. There are also three clusters in that from the right subsequence: I, II, and III. But the composition of the clusters differs in the two phenograms. If one lists the tight subgroups common to both phenograms and adds their genomic composition, one obtains six subgroups each with a symbol for the left and the right subsequence (pattern C becomes subdivided). A highly speculative evolutionary reconstruction based upon this is shown (Fig. 5). This shows three crossover events (though the lineages might be rearranged to some degree to give different directions of genetic transfer). The figure fits reasonably well to the data, but it must be emphasized that it is mainly speculation. When Fig. 4 and 5 are represented in graphic form to show phylogenetic confidence or uncertainty (9, 10), they are shown as plausible but not confirmed at high significance levels; chance effects seem unlikely to explain the findings, but this is not ruled out.

FIG. 5. A speculative reconstruction of evolution in *Aeromonas* 16S rRNA based on the phenograms of Fig. 4. The sequences fall into six tight groups, and above each group is indicated their genetic composition (*a*, *b*, or *c* from the left part and I, II, or III from the right part of the ribosome sequence).

There are several copies of the genes for ribosomal sequences on the bacterial chromosome that are usually identical or nearly so (7, 13), but one could envisage that they were sometimes different or that genes from other bacteria might be brought in by plasmids. If so, crossing-over could occur, and selective pressures for new crossover configurations could be exerted by antibiotics that affect ribosome function. Spratt (12) among others has adduced evidence of hybrid genes in bacteria, and Ambler (1) predicted that ribosomal hybrids would be discovered. We do not yet know what effect crossovers might have on ribosome function, although there must be strong functional limits on their positions and extent. By analogy with similar genera, there are probably six to eight copies of the 16S gene in *Aeromonas*. Earlier techniques with reverse transcriptase would have recovered only the expressed copy. The polymerase chain reaction used here recovers a consensus of the sequences, but because there were no ambiguities in the sequences (8) the copies are probably all identical. This raises the question of how a crossover could become fixed in all copies during evolution. Yet the same problem arises with fixation of mutations; these also must eventually replace their alternatives in all copies. There are many mechanisms of genome turnover (gene conversion, concerted evolution, etc.) which could lead to all copies becoming the same, particularly if there are selective pressures (4). Crossover events may also explain the very short or very long stems sometimes seen in phylogenetic trees (2, 13) that are usually attributed to different rates of evolution. It may be thought that these findings in *Aeromonas* are relevant only to evolutionary structure within a genus. The genomic differences between species of *Aeromonas* are very small as judged by 16S ribosomal sequences (similarities are greater than 97.8%). At this level, exchange of genetic material is more likely to occur than within a genus whose species show a higher degree of diversity. Yet presumably genera themselves arise from earlier species, and so they would seem relevant to generic relationships too.

We still lack good methods to discover phenomena of the kind discussed here. If one suspects that two subsets of the data have different evolutionary histories, a single hybrid event may become evident from simple analysis. If there is more than one event, the tabulation of the genetic composition from two phenograms, as in Fig. 5, may be helpful. But if one does not suspect anything untoward, the phenomena may go unrecognized. One useful point is that poor ultrametric properties give low cophenetic correlation of a phenogram (11). The cophenetic correlation of UPGMA clustering for the total nucleotide differences is only 0.80, compared to 0.95 and 0.87 for the left and right subsequences separately, respectively. Low cophenetic correlation may thus point to situations that need investigating, and the importance of such checks on the quality of phenograms or phylogenetic trees is thus underlined. These observations also emphasize the need for systematic recording and comparison of different kinds of sequences such as 16S and 23S ribosomal sequences.

I am grateful for critical comments from M. D. Collins and D. Jones.

REFERENCES

1. Ambler, R. P. 1985. Protein sequencing and taxonomy, p. 307-335. In M. Goodfellow, D. Jones, and F. G. Priest (ed.), *Computer-assisted bacterial systematics*. Academic Press, London.
2. Collins, M. D., S. Wallbanks, D. J. Lane, J. Shah, R. Nietupski, J. Smida, M. Dorsch, and E. Stackebrandt. 1991. Phylogenetic

- analysis of the genus *Listeria* based on reverse transcriptase sequencing of 16S rRNA. *Int. J. Syst. Bacteriol.* **41**:240–246.
3. Dayhoff, M. O. (ed.). 1972. Atlas of protein sequence and structure 1972. National Biomedical Research Foundation, Washington, D.C.
 4. Dover, G. A. 1982. Molecular drive: a cohesive mode of species evolution. *Nature (London)* **299**:111–117.
 5. Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Q. Rev. Biol.* **57**:379–404.
 6. Fox, G. H., J. D. Wisotzkey, and P. Jurtshuk, Jr. 1992. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.* **42**:166–170.
 7. Krawiec, S., and M. Riley. 1990. Organization of the bacterial chromosome. *Microbiol. Rev.* **54**:502–539.
 8. Martinez-Murcia, A. J., S. Benlloch, and M. D. Collins. 1992. Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. *Int. J. Syst. Bacteriol.* **42**:412–421.
 9. Sneath, P. H. A. 1986. Estimating uncertainty in evolutionary trees from Manhattan-distance triads. *Syst. Zool.* **35**:470–488.
 10. Sneath, P. H. A., and J. C. Radbourn. 1991. DOTDND: a FORTRAN-77 program for showing graphically the confidence or uncertainty in phylogenetic trees. *Computers Geosciences* **17**:689–718.
 11. Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W.H. Freeman, San Francisco.
 12. Spratt, B. G. 1988. Hybrid penicillin-binding proteins in penicillin-resistant strains of *Neisseria gonorrhoeae*. *Nature (London)* **332**:173–176.
 13. Woese, C. R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**:221–271.

P. H. A. Sneath
Microbiology Department
Leicester University
Leicester LE1 9HN
United Kingdom