Letter to the Editor

The Order of Sequence Alignment Can Bias the Selection of Tree Topology[1]

James A. Lake
Molecular Biology Institute and Biology Department, University of California, Los Angeles

Sequential pairwise alignment of multiple sequences is a widely used procedure (Kruskal 1983). It is useful and generally successful when sequences within a set differ by relatively few substitutions. Although it is well known that differential substitution rates can artifactually bias the assessment of tree topology (Felsenstein 1978), it is not generally known that the order in which sequences are aligned can bias tree selection.

To test the effect of alignment order, the classical four-taxon test has been applied to the "tree of life" (Lake et al. 1984; Woese and Olsen 1986) by using alternative alignments and three reconstruction algorithms [maximum parsimony (Fitch 1971), transversion parsimony (Brown et al. 1982), and evolutionary parsimony (Lake 1987)]. There is enormous interest in this tree because it relates all known organisms and because its topology is expected to provide insight into the evolution of modern organisms. Because the tree spans large evolutionary distances, its topology has been difficult to establish.

By means of sequences from elongation factor Tu (EF-Tu), the most conserved protein sequence known to span the tree of life, it is shown that specific alignment orders systematically favor alternative trees. In particular, if taxa A and B are pairwise aligned and if C and D are pairwise aligned, the resulting alignment of the EF-Tu sequences more often gives the tree that has A and B as topological neighbors and C and D as topological neighbors, regardless of the tree reconstruction algorithm used. Because all three reconstruction algorithms produced the same tree for any particular alignment, unequal rate effects appear to be secondary for EF-Tu sequences. This indicates that order-dependent alignment biases are distinct from unequal rate effects and that, for some data, they could be as important as unequal rate effects.

Pairwise alignments of protein sequences were performed with the ALIGN program available in the Dayhoff package (Dayhoff et al. 1983). The penalty for a break was 6, and the mutation data matrix corresponded to 250 accepted point mutations with a bias of +2. These are reasonable values for the weights and correspond to those used in the examples in the description of the ALIGN program. [For an insightful discussion of alignment weights, see the paper by Fitch and Smith (1983); also see Waterman and Perlwitz (1984).] EF-Tu sequences were aligned as protein sequences to obtain more robust alignments and were back-translated into nucleic acid sequences (e.g., phe was translated as UUY, leu as YUN, arg as NGN, and ser as NNN) so that the maximum-, transversion-, and evolutionary-parsimony methods could be compared by equivalent data. Only positions consisting of a single nucleotide (i.e., U, C, A, or G but not R, Y, or N) in each of the four sequences were scored. These uniquely defined replacement sites are presumed to correspond to the most conserved nucleotide positions.

A multiple alignment of four sequences can be achieved by successively aligning

[1] Key words: parsimony, alignment, sequences.

Address for correspondence and reprints: James A. Lake, Molecular Biology Institute and Biology Department, University of California, Los Angeles, Los Angeles, California 90024.

Mol. Biol. Evol. 8(3):378–385. 1991.

0737-4038/91/0803-0009$02.00

three pairs of sequences. Let A, B, C, and D represent amino acid (or nucleotide) sequences, and let AB represent the alignment (sensu Kruskal 1983) of A with B, etc. [In this definition, an alignment consists of a matrix of two rows in which a match is indicated by a column with the same element above and below, a replacement (or substitution) is indicated by a column with different elements above and below, and a deletion in A (or an insertion in B) is represented by a column with a gap (−) above and with a nongap element below, etc.] If one sequence is common to all three pairs, I call these *star* alignments. If no sequence is common to more than two pairs, I call these *linear alignments;* there are 12 of them represented by a linear notation such as ABCD, where the four letters represent the four sequences and where the three adjacent pairs of letters—AB, BD, and DC—represent the three pairwise alignments used to generate the total alignment.

Pairwise-alignment algorithms do not distinguish between the order of the two sequences being aligned. Thus, the alignment of sequence A with B, AB, is equivalent to the alignment of B with A, BA. There are 4! = 24 orderings of four letters (ABCD represents the alignment of A with B, B with C, and C with D) corresponding to four-taxon pairwise alignments, but because only the neighbors—and not their order—count, the alignment ABDC is equivalent to the alignment CDBA. This explains why it was stated earlier that there are only 12 (=24/2) independent linear, pairwise alignments of four sequences.

Multiple alignments were generated from sequential pairwise alignments as illustrated in table 1. In the upper example the AB alignment is aligned to the BC alignment by requiring the common (or guide) sequence B to have its two amino acid sequences aligned perfectly, leading to the introduction of an additional gap (*) in each of the AB and BC alignments. Once this has been done, the ABBC alignment on the left may be reduced to the ABC alignment on the right. [The ABC alignment is defined only within the length (or range) of B because B contains only gaps outside this range. Hence, B can not be used to relate the A and C sequences in regions where B does not exist. Additional rules could be devised to extend this range, but this seems

Table 1
Sequential Alignment of Two Alignments

Alignment of Two Alignments	Reduced Representation
AB aligned with BC:	
A: KN-ITGTS*QA	
B: KNMIT-AS*QA	A: KN-ITGTS-QA
	B: KNMIT-AS-QA
B: KNMIT*AS-QA	C: K-MIT-AAKQM
C: K-MIT*AAKQM	
BC aligned with CA:	
B: KNMITAS-QA	
C: K-MITAAKQM	B: KNMITAS-QA
	C: K-MITAAKQM
C: K*MITAAKQM	A: K-NITGTSQA
A: K*NITGTSQA	

NOTE.—Two pairwise alignments of sequences are aligned by reference to a common sequence. At the top left the AB pair is aligned with respect to the BC pair through the common B sequence. At the bottom left the BC pair is aligned with the CA pair through the common C sequence. On the left, hyphens (-) represent gaps introduced when the initial pairs (AB and BC upper, or AB and CA lower) were aligned, and asterisks (*) represent gaps introduced when two alignments were aligned with each other. Asterisks have been changed to hyphens on the right. The final result is shown in a reduced form at the right. A triple alignment ABC is commonly not the same as a triple alignment BCA.

an unnecessary complication for this paper, since positions containing gaps will not be scored.]

One can easily show that the alignment ABC is equivalent to the alignment CBA, since B is used as the guide sequence for both alignments. Furthermore, alignments are associative; that is, the alignment (AB)(CD) is equivalent to (ABC)(D), where the brackets indicate the order in which alignments are combined. [The alignment (ABC)(D) is equivalent to the alignment (ABC)(CD) since the left and right C sequences are identical. Likewise, (AB)(CD) is equivalent to (ABC)(CD). Hence (ABC)(D) is equivalent to (AB)(CD).] Although sequential alignments are associative, they are not in general commutative. This is shown by the example in the bottom half of table 1, where the alignment BCA is calculated. It is clear that the alignment BCA *is not equivalent* to ABC. A collection of alignments is thus a semigroup under alignment, as pointed out by a reviewer.

The four-taxon tree is the traditional vehicle for testing reconstruction algorithms, and the best-known four-taxon test concerns the tree of life, which relates all known groups of organisms. Hence it will be used to illustrate the effects that alignment has on tree selection. In its unrooted form, the tree of life relates the Halobacteria (H), the Eubacteria (B), the Eukaryotes (K), and the Eocytes (E). The best-studied proteins that are found in all known organisms are the DNA-dependent RNA polymerases, the ATP synthetases, and the protein synthesis factor EF-Tu (EF-1 alpha in eukaryotes). The EF-Tu sequences are the least divergent of the three proteins and were used in the present study. This reduced complications introduced by unequal rate effects.

Amino acid sequences were taken from each of four representative taxa. *Escherichia coli* was selected as the traditional eubacterium (Yokota et al. 1980), but analogous results were obtained with EF-Tu sequences from *Spirulina platensis* (cyanobacterium) and *Thermotoga martima* (thermophilic eubacterium). *Halobacterium marismortui* was chosen as the H sequence (Baldacci et al. 1990). *Thermococcus celer* was selected as an E sequence (Auer et al. 1990). *Saccharomyces cerevisiae* was chosen as the K representative (Nagata et al. 1984) because of its central phylogenetic position and because its sequence appears to have undergone relatively fewer substitutions than have either other single-celled E or most metazoans. For sequences B, E, H, and K, the 12 linear alignments are EKBH, KEBH, EKHB, KEHB, EHBK, HEBK, EHKB, HEKB, HKBE, KHBE, HKEB, and KHEB. In reduced form, EHBK is an alignment of four rows.

Four taxa may be related in only three unrooted trees, as shown in figure 1. In the archaebacterial tree in figure 1A, K are not topologically closest to either E or H. In the halobacterial tree, shown in figure 1B, K are topologically closest to H, and in the eocyte tree, shown in figure 1C, the K are topologically closest to E.

For the six alignments shown in table 2A, the order of alignment strongly influences the topology. In particular, (1) those alignments in which E is aligned with H *and* in which K is aligned with B support the archaebacterial tree, (2) those alignments in which H is aligned with K *and* in which B is aligned with E support the halobacterial tree, and (3) those in which E is aligned with K *and* in which B is aligned with H support the eocyte tree. The order of alignment dominates the topology.

The observations are consistent with the following simple explanation: When one aligns the E sequence with the K sequences and then aligns the B and H sequences, the sequences are, in effect, being fit to the tree that has E and K on one side of the central branch and B and H on the other side. In this instance the eocyte topology will be emphasized. Similarly, when one fits E with H and fits B with K, the archaebacterial tree is favored, and so forth.

When the remaining six alignments are examined (table 2B), a similar but decreased effect is found. Although these alignments primarily support the eocyte tree, the greatest support for a given topology (or a tie for greatest support) occurs when

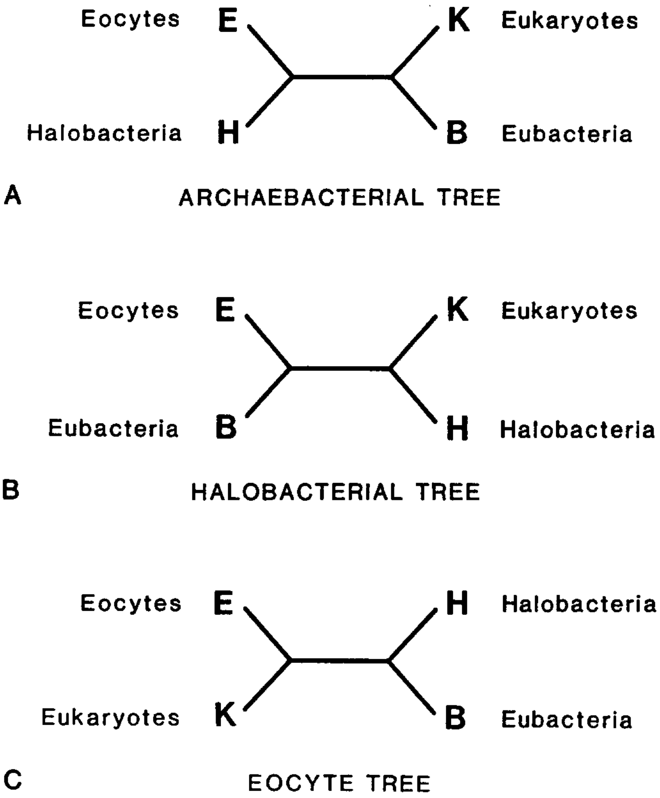

FIG. 1.—Three possible four taxon topologies for "tree of life." The archaebacterial tree, shown in A, relates the eocytes with halobacterial and relates the eukaryotes with eubacteria. The halobacterial tree, shown in B, relates the eocytes with eubacteria and relates the eukaryotes with halobacteria. The eocyte tree relates the eocytes with eukaryotes and relates the halobacteria with eubacteria.

its favored alignment order is used. Thus, if the EKHB or KEHB alignments is used, support for the eocyte tree is increased beyond the level that was found by using the EHKB, HEKB, HKEB, and KHEB alignments.

The differences between table 2A and table 2B suggest that the central alignment pair influences the strength of the effect. In all the six alignments in table 2A the B sequence is part of the central aligned pair. Because B is the most divergent of the four sequences [The length of the peripheral branch leading to B is consistently the longest for all alignments and topologies when measured by operator metrics (Lake 1988).], it appears that the effect is stronger if the two central sequences are highly diverged—and is weaker if they are less diverged. This effect was also noted for polymerase and ATP synthetases (J. A. Lake, unpublished results). Thus, when the divergent B sequence is part of the central pair, alignment effects dominate and determine the tree topology.

Whether analyzed by maximum parsimony, by transversion parsimony, or by evolutionary parsimony (see tables 2A and B), tree selection is principally determined by the alignment. Because the three methods have different sensitivities to unequal rates, these effects are probably not biasing the results. Hence, alignment effects are distinct from unequal rate effects. This implies that all the algorithms studied (including evolutionary parsimony, the least affected by unequal rates) are sensitive to sequential alignment effects.

For highly divergent sequences, the order of the alignments can dominate phy-

Table 2
Order of Alignment as Biasing Topology

	MAXIMUM PARSIMONY			TRANSVERSION PARSIMONY			EVOLUTIONARY PARSIMONY		
ALIGNMENT	Archaebacterial Tree	Halobacterial Tree	Eocyte Tree	Archaebacterial Tree	Halobacterial Tree	Eocyte Tree	Archaebacterial Tree	Halobacterial Tree	Eocyte Tree
A: Central alignment including most divergent sequence (B):									
EHBK	27	6	21	16	2	9	11	0	5
HEBK	31	5	16	14	1	8	8	0	5
HKBE	15	22	20	5	14	9	5	11**	4
KHBE	13	28	21	5	18	10	5	12**	5
EKBH	12	7	37	5	4	21	2	2	11
KEBH	8	8	37	3	2	21	2	−1	15***
B: Central alignment not including most divergent sequence (B):									
EHKB	18	6	18	8	3	9	6	3	3
HEKB	17	5	19	7	2	9	3	2	3
HKEB	13	8	20	5	3	10	1	2	7
KHEB	16	10	19	7	4	10	3	3	5
EKHB	14	8	27	5	4	11	3	3	8*
KEHB	15	8	26	6	3	11	2	2	7
C: Order-invariant alignment	7	2	13	3	0	6	3	1	4
D. Star alignments:									
StarB	13	13	22	6	7	12	5	4	5
StarE	13	8	20	5	2	10	0	0	6
StarH	15	11	25	6	5	11	4	5	7
StarK	16	12	21	6	3	9	4	3	4

NOTE.—Data are scores for the respective trees. In all three parsimony methods used to analyze aligned sequences, each of the three trees is associated with particular patterns of nucleotide occurrence. The scores for maximum parsimony and for transversion parsimony are the number of sequence positions at which the nucleotide pattern supports a particular topology. The scores for evolutionary parsimony are the number of sequence positions that support minus the number that oppose a topology. The tree with the greatest support is deemed "most parsimonious." Only nucleotide positions without gaps were used in the tree constuction analyses. The tree topology supported by the most counts is underlined (ties are not indicated). The six, of 12 possible, alignments shown in A correspond to those in which the central alignment includes the most divergent sequence, (B). In B the central alignment pair does not include the most divergent sequence, (B). Two alternative alignment strategies that are less influenced by tree topologies are shown in C and D. In C the order-independent alignment includes only the portion of the alignment that is common to three diverse linearly sequential alignments. Star alignments are analyzed in D. The StarB alignment uses the B sequence for reference; the StarE alignment uses the E sequence; etc.

* $P < .05$, by χ^2 test (as in Lake 1987).

** $P < .03$, by χ^2 test (as in Lake 1987).

*** $P < .01$, by χ^2 test (as in Lake 1987).

logenetic reconstructions. Furthermore, the alignment artifact is likely to have wide-ranging consequences, since almost all alignments are constructed by first aligning the pairs of sequences that are most similar. Whether calculated by computer or by eye, these types of alignments predispose the algorithms toward the tree that has the least divergent taxa as neighbors. Even true multiple-taxon alignment algorithms (Sankoff and Cedergren 1983) suffer from this distortion, which can enter through the choice of distances to be minimized (J. A. Lake, unpublished results). Until we can understand more completely the subtle relationships between sequence alignment and topology determination, some suggestions for obtaining multiple alignments seem useful.

A direct solution to the alignment problem is to search for alignments that are *independent* of the sequential alignments. This procedure can be computationally intensive but is potentially useful. If one can find subsets of an alignment that are common to all of the possible sequential alignments, then it can be argued that the subset is reasonably free of topological alignment biases. The subset that is common to the EKBH, HKBE, and EHBK alignments is shown in the Appendix. I call this an *order-independent alignment,* and the analysis of it is shown in table 2C. Although this alignment is easy to calculate for four taxa, for large data sets this calculation can become computationally intensive. For example, in one study of the tree of life (Lake 1988), ~1,200 individual four-taxon trees were analyzed. Since each four-taxon tree requires 12 independent alignments, one would need to calculate some 14,000 four-taxon alignments in order to use this alignment method. Nevertheless, this is still a feasible computation.

Another type of pairwise, sequential alignment—the star alignment—requires less computation (Lake 1988). In this alignment, one selects a reference sequence and aligns all other sequences to it. If K were selected as a reference, then one would calculate BK, EK, and HK and combine them. Four star alignments are possible for four taxa, and their analyses are shown in table 2D. For them, all three methods are consistent with the same topology found for the order-independent alignment. This suggests that topological distortions are less for the star and order-independent alignments than for the linear sequential alignments, for these data. The results in table 2D are not significantly different when the divergent B sequence is used as a reference, but in general it would seem unwise to use a divergent sequence as a reference. An obvious benefit is that star alignments require substantially less calculation, since all taxa can be referenced to a single sequence. For the tree of life (Lake 1988), only 31 separate pairwise alignments were required for their combination into the complete 32-taxon star alignment.

Notably, it appears that the four sequences analyzed here tend to support the eocyte tree. Whether this is an effect observed for these four sequences or a statement about the tree of life would take us beyond the scope of the present paper and require the analysis of additional data. Nevertheless, additional EF-Tu genes from eocyes are being sequenced. If order-independent alignments of these sequences—the most conserved protein sequences yet found—should also support the eocyte tree, this would argue strongly for it.

Most important, the present work shows that alignment order introduces topological distortions that are distinct from—and, in the present example, more significant than—unequal rate effects. As additional, longer sequences—and even complete genomes—become available, our attempts to reconstruct the past will become even more ambitious. Almost certainly, consideration of the artifacts introduced by alignment order will play a major role in these studies.

APPENDIX

Some alignments of the EF-Tu sequences used in the present paper are listed in fig. A1. The sequences are referenced by the following code: B (eubacterium), E (eocyte), H (halobacterium), and K (eukaryote). This is followed by four letters describing the alignment that

```
K  STAR    MGKEK----- SHINVVVIGH VDSGKST--- -T---TGHLI ------YKC- GGIDKRTIEK FE---KEAAE LGKGSFKYAW V---LDKLKA
E  EKBH    MAKEK----- PHINIVFIGH VDHGKST--- -T---IGRLL ------FDT- ANIPENIIKK FE----EMGE KGK-SFKFAW V---MDRLKE
B  EKBH    MSKEKFERTK PHVNVGTIGH VDHGKTT--- -L---TAAIT TVLAKTY--- GGAAR----A FD---Q---- ---------- ----IDNAPE
H  EKBH    MSDEQ----- -HQNLAIIGH VDHGKST--- -L------VG RLLYETG--- SVPEH----V IE---QHKEE AEEKGKGGFE FAYVMDNLAE
E    HKBE  MAKEK----- PHINIVFIGH VDHGKSTTIG RLLFDTANIP ENIIKKFEEM GEKGK----S FKFAWV---- ---------- ----MDRLKE
B    HKBE  MSKEKFERTK PHVNVGTIGH VDHGKTT--- -L---TAAIT TVLAKTY--- GGAAR----A FD---Q---- ---------- ----IDNAPE
H    HKBE  MSDEQ----- -HQNLAIIGH VDHGKST--- -L---VGRLL ------YET- GSVPEHVIEQ HK---EEAEE KGKGGFEFAY V---MDNLAE
E      EHBK MAKEKP---- -HINIVFIGH VDHGKST--- -T------IG RLLFDTA--- NIPEN----I IK---KF-EE MGEKGKS-FK FAWVMDRLKE
B      EHBK MSKEKFERTK PHVNVGTIGH VDHGKTT--- -L---TAAIT TVLAKTY--- GGAAR----A FD---Q---- ---------- ----IDNAPE
H      EHBK MSDEQ----- -HQNLAIIGH VDHGKST--- -L------VG RLLYETG--- SVPEH----V IE---QHKEE AEEKGKGGFE FAYVMDNLAE
K COMMON   MGKEK----- -HINVVVIGH VDSGKST--- ---------- ---------- ---------- ---------- ---------- ----LDKLKA

K  STAR    ERERGITIDI ALWKFET-P- KYQVTVIDAP GHRDFIKNMI TGTSQADCAI LIIAGGVGEF EAGISKDGQT REHALLAFTL GVRQLIVAVN
E  EKBH    ERERGITIDV AHTKFET-PH RY-ITIIDAP GHRDFVKNMI TGASQADAAV LVVA--VTD- --GVMP--QT KEHAFLARTL GINNILVAVN
B  EKBH    EKARGITINT SHVEYDT-P- TRHYAHVDCP GHADYVKNMI TGAAQMDGAI LVVAATDGPM P-------QT REHILLGRQV GVPYIIVFLN
H  EKBH    ERERGVTIDI AHQEFST-D- TYDFTIVDCP GHRDFVKNMI TGASQADNAV LVVAADDGVQ P-------QT QEHVFLARTL GIGELIVAVN
E    HKBE  ERERGITIDV AHTKFET-P- HRYITIIDAP GHRDFVKNMI TGASQADAAV LVVAVTDGVM P-------QT KEHAFLARTL GINNILVAVN
B    HKBE  EKARGITINT SHVEYDT-P- TRHYAHVDCP GHADYVKNMI TGAAQMDGAI LVVAATDGPM P-------QT REHILLGRQV GVPYIIVFLN
H    HKBE  ERERGVTIDI AHQEFST-D- TYDFTIVDCP GHRDFVKNMI TGASQADNAV LVVAA----- DDGV--QPQT QEHVFLARTL GIGELIVAVN
E      EHBK ERERGITIDV AHTKFETPH- RY-ITIIDAP GHRDFVKNMI TGASQADAAV LVVAVTDGVM P-------QT KEHAFLARTL GINNILVAVN
B      EHBK EKARGITINT SHVEYDT-P- TRHYAHVDCP GHADYVKNMI TGAAQMDGAI LVVAATDGPM P-------QT REHILLGRQV GVPYIIVFLN
H      EHBK ERERGVTIDI AHQEFST-D- TYDFTIVDCP GHRDFVKNMI TGASQADNAV LVVAADDGVQ P-------QT QEHVFLARTL GIGELIVAVN
K COMMON   ERERGITIDI ALWKFET--- ---VTVIDAP GHRDFIKNMI TGTSQADCAI LIIA------ --------QT REHALLAFTL GVRQLIVAVN

K  STAR    KMDSVKW-DE SRFQEIVKET SNFIKKVGYN PKTVPF--VP ISGWNGDNMI E--ATTNA-- ------P--- WYKGWEKETK AGVVKGKTLQ
E  EKBH    KMDMVNY-DE KKFKAVAEQV KKLLMMLGY- -KNFPI--IP ISAWEGDNVV K--KSDKM-- ------P--- WYNG------ ------PTLL
B  EKBH    KCDMVD--DE ELLELVEMEV RELLSQYDFP GDDTPI--VR -----GSALK A--LEGDA-- ------E--- W------EAK --ILELAGFL
H  EKBH    KMDLVDYGES EYKQVVE-EV KDLLTQVRFD SENAKF--IP -----VSAFE GDNIAEES-- ------EHTG WY-----DGE --IL-LEALN
E    HKBE  KMDMVNY-DE KKFKAVAEQV KKLLMMLGY- -KNFPI--I- -------PIS A--WEGDNVV KKSDKMP--- WYN----GPT --LIEA---L
B    HKBE  KCDMVD--DE ELLELVEMEV RELLSQYDFP GDDTPI--VR -----GSALK A--LEGDA-- ------E--- W------EAK --ILELAGFL
H    HKBE  KMDLVDY-GE SEYKQVVEEV KDLLTQVRFD SENAKF--IP VSAFEGDNIA E--ESEHT-- ------G--- WYDG------ ------EILL
E      EHBK KMDMVNYDEK KFKAVAE-QV KKLLMMLGYK N----FPIIP -----ISAWE GDNVVKKS-- ------DKMP WY-----NGP --TL-IEALD
B      EHBK KCDMVD--DE ELLELVEMEV RELLSQYDFP GDDTPI--VR -----GSALK A--LEGDA-- ------E--- W------EAK --ILELAGFL
H      EHBK KMDLVDYGES EYKQVVE-EV KDLLTQVRFD SENAKF--IP -----VSAFE GDNIAEES-- ------EHTG WY-----DGE --IL-LEALN
K COMMON   KMDSVKW--- --------ET SNFIKKVGY- ---------- ---------- ---------- ---------- ---------- ----------

K  STAR    EAIDAIEQPS RPTDKPLRLP LQDVYKIGGI GTVPVGRVET GVIKPGMVVT F--APAG--- ---VTTE--V KS-------V EMHHEQLEQA
E  EKBH    EALDQMPEPP KPTDKPLRIP IQDVYSIKGV GTVPVGRVET GVLRVGDVVI F--EPASTIF HKPIQGE--V KS-------I EMHHEPMQEA
B  EKBH    DSY--IPEPE RAIDKPFLLP IEDVFSISGR GTVVTGRVER GIIKVGEEVE I--V--G--- ---IKET--Q KSTCTG---V EMFRKLLDEG
H  EKBH    E----LPAPE PPTDAPLRLP IQDVYTISGI GTVPVGRVET GILNTGDNVS FQPS--D--- ---VSGE--V KT-------V EMHHEEVPKA
E    HKBE  DQ---MPEPP KPTDKPLRIP IQDVYSIKGV GTVPVGRVET GVLRVGDVV- I--F--E--- ---PASTIFH KPIQGEVKSI EMHHEPMQEA
B    HKBE  DSY--IPEPE RAIDKPFLLP IEDVFSISGR GTVVTGRVER GIIKVGEEVE I--V--G--- ---IKET--Q KST---CTGV EMFRKLLDEG
H    HKBE  EALNELPAPE PPTDAPLRLP IQDVYTISGI GTVPVGRVET GILNTGDNVS F--QPSD--- ---VSGE--V KT-------V EMHHEEVPKA
E      EHBK Q----MPEPP KPTDKPLRIP IQDVYSIKGV GTVPVGRVET GVLRVGDVVI FEPA--STIF HKPIQGE--V KS-------I EMHHEPMQEA
B      EHBK DSY--IPEPE RAIDKPFLLP IEDVFSISGR GTVVTGRVER GIIKVGEEVE I--V--G--- ---IKET--Q KSTCTG---V EMFRKLLDEG
H      EHBK E----LPAPE PPTDAPLRLP IQDVYTISGI GTVPVGRVET GILNTGDNVS FQPS--D--- ---VSGE--V KT-------V EMHHEEVPKA
K COMMON   -----IEQPS RPTDKPLRLP LQDVYKIGGI GTVPVGRVET GVIKPGMVV- ---------- ---------- ---------V EMHHEQLEQA

K  STAR    VPGDNVGFNV KNVSVKEIRR GNVCGDAK-- --NDP----P KGCA-----S FNATVIVL-- NHPGQISA-- --GYSPVLDC HTAHI-----
E  EKBH    LPGDNIGFNV RGVGKNDIKR GDVAGHTN-- --NPPTVVRP KD-------T FKAQIIVL-- NHPTAITV-- --GYTPVLHA HTLQV-----
B  EKBH    RAGENVGVLL RGIKREEIER GQV--LAK-- --PGT----I KPHT-----K FESEVYILSK DEGGRHTPFF K-GYRPQFYF RTTDV-----
H  EKBH    EPGDNVGFNV RGVGKDDIRR GDV---CG-- --PAD----D PPSVAE---T FQAQIVVM-- ----QHPSVI TEGYTPVFHA HTAQV-----
E    HKBE  LPGDNIGFNV RGVGKNDIKR GDV--AGHTN NPPTV----V RPKD-----T FKAQIIVLN- ----HPTAIT V-GYTPVLHA HTLQVAVRFE
B    HKBE  RAGENVGVLL RGIKREEIER GQV--LAK-- --PGT----I KPHT-----K FESEVYILSK DEGGRHTPFF K-GYRPQFYF RTTDV-----
H    HKBE  EPGDNVGFNV RGVGKDDIRR GDVCGPA--- --DDP----P SVAE-----T FQAQIVVM-- QHPSVITE-- --GYTPVFHA HTAQV-----
E      EHBK LPGDNIGFNV RGVGKNDIKR GDV---AG-- --HTN----N PPTVVRPKDT FKAQIIVL-- ----NHPTAI TVGYTPVLHA HTLQV-----
B      EHBK RAGENVGVLL RGIKREEIER GQV--LAK-- --PGT----I KPHT-----K FESEVYILSK DEGGRHTPFF K-GYRPQFYF RTTDV-----
H      EHBK EPGDNVGFNV RGVGKDDIRR GDV---CG-- --PAD----D PPSVA---ET FQAQIVVM-- ----QHPSVI TEGYTPVFHA HTAQV-----
K COMMON   VPGDNVGFNV KNVSVKEIRR GNV------- ---------- ---------S FNATVIVL-- ---------- --GYSPVLDC HTAHI-----

K  STAR    ---------A CRFDELLEKN DRRSGKKL-- ----EDHPKF -------LKS GDAALVKFVP S-------KP MCVEAFSEYP P-----LGRF
E  EKBH    ---------A VRFEQLLAKL DPRTGNIV-- ----EENPQF -------IKT GDSAIVVLRP T-------KP MVIEPVKEIP Q-----MGRF
B  EKBH    ---------T GTIE------ -------L-- ----PEGVEM -------VMP GDN--IKMVV TLI-----HP IAMDD----- G-----L-RF
H  EKBH    ---------A CTVE------ -------SID KKIDPSSGEV AEENPDFIQN GDA--AVVTV RPQ-----KP LSIEP----- SSEIPELGSF
E    HKBE  QLLAKLDPRT GNI------- -------V-- ----EENPQF -------IKT GDS--AIVVL RPTKPMVIEP VKEIP----- Q-----MGRF
B    HKBE  ---------T GTIE------ -------L-- ----PEGVEM -------VMP GDN--IKMVV T-----LIHP IAMDD----- G-----L-RF
H    HKBE  ---------A CTVESIDKKI DPSSGEVA-- ----EENPDF -------IQN GDAAVVTVRP Q-------KP LSIEPSSEIP E-----LGSF
E      EHBK ---------A VRFE------ -------QLL AKLDPRTGNI VEENPQFIKT GDS--AIVVL RPT-----KP MVIEP----- VKEIPQMGRF
B      EHBK ---------T GTIE------ -------L-- ----PEGVEM -------VMP GDN--IKMVV TLI-----HP IAMDD----- G-----L-RF
H      EHBK ---------A CTVE------ -------SID KKIDPSSGEV AEENPDFIQN GDA--AVVTV RPQ-----KP LSIEP----- SSEIPELGSF
K COMMON   ---------- ---------- ---------- ---------- -------LKS GDA------- ---------- ---------- ------LGRF

K  STAR    AVRDMRQTVA VGVIKSVDKT EKAAKVTKAA QKAAKK        458
E  EKBH    AIRDMGQTVA AGMVISIQKA E--------- ------        428
B  EKBH    AIREGGRTVG AGVV------ ---AKVLS-- ------        394
H  EKBH    AIRDMGQTIA AG-------- ----KVLGVN ER----        421
E    HKBE  AIRDMGQTVA AGMV------ ---ISIQKAE ------        428
B    HKBE  AIREGGRTVG AGVV------ ---AKVLS-- ------        394
H    HKBE  AIRDMGQTIA AGKVLGVN-- ER-------- ------        421
E      EHBK AIRDMGQTVA AG-------- ----MVISIQ KAE---        428
B      EHBK AIREGGRTVG AGVV------ ---AKVLS-- ------        394
H      EHBK AIRDMGQTIA AG-------- ----KVLGVN ER----        421
K COMMON   AVRDMRQTVA VG-------- ---------- ------        237
```

FIG. A1.—Various sequence alignments of EF-Tu

corresponds to the code used in the text. The K STAR sequence is the yeast sequence used as a "star" reference to combine the three different alignments. The K COMMON sequence is the yeast sequence at only those positions where the EKBH, HKBE, and EHBK alignments are identical.

Acknowledgments

I thank Walter Fitch for encouraging this work and thank a reviewer for many helpful comments and pointing out that a collection of alignments is a semigroup under alignment. This work was supported by grants from the NSF, NIH, and Sloan Foundations.

LITERATURE CITED

AUER, J., B. SPICKER, and A. BOCK. 1990. Nucleotide sequence of the gene for elongation factor EF-1 alpha for the extreme thermophilic archaebacterium *Thermococcus celer.* Nucleic Acids Res. **18**:3989.

BALDACCI, B., F. GUINET, J. TILLIT, G. ZACCAI, and A.-M. DE RECONDO. 1990. Functional implications related to the gene structure of the elongation factor EF-Tu form *Halobacterium marismortui.* Nucleic Acids Res. **18**:507–511.

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

DAYHOFF, M. O., W. C. BARKER, and L. T. HUNT. 1983. Establishing homologies in protein sequences. Methods Enzymol. **91**:524–545.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

FITCH, W. M., and T. F. SMITH. 1983. Optimal sequence alignments. Proc. Natl. Acad. Sci. USA **80**:1382–1386.

KRUSKAL, J. B. 1983. An overview of sequence comparison. Pp. 1–45 *in* D. SANKOFF and J. B. KRUSKAL, eds. Time warps, string edits, and macromolecules. Addison-Wesley, Reading, Mass.

LAKE, J. A. 1987. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol. Biol. Evol. **4**:167–191.

———. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature **331**:184–186.

LAKE, J. A., E. HENDERSON, M. W. CLARK, and M. OAKES. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. USA **81**:3786–3790.

NAGATA, S., K. NAGASHIMA, Y. TSUNETSUGU-YOKATA, K. FUJIMURA, M. MIYAZAKI, and Y. KAZIRO. 1984. Polypeptide chain elongation factor 1 alpha from yeast: nucleotide sequence of one of the two genes for EF-1 alpha from *Saccharomyces cerevisiae.* EMBO J. **3**:1825–1830.

SANKOFF, D., and R. J. CEDERGREN. 1983. Simultaneous comparison of three or more sequences related by a tree. Pp. 253–263 *in* D. SANKOFF and J. B. KRUSKAL, eds. Time warps, string edits, and macromolecules. Addison-Wesley, Reading, Mass.

WATERMAN, M., and M. D. PERLWITZ. 1984. Line geometries for sequence comparison. Bull. Math. Biol. **46**:567–577.

WOESE, C. R., and G. J. OLSEN. 1986. Archaebacterial phylogeny: perspectives on the urkingdoms. S-yst. Appl. Microbiol. **7**:161–177.

YOKOTA, T., H. SUGISAKI, M. TAKANAMI, and Y. KAZIRO. 1980. The nucleotide sequence of the cloned *tufA* gene of *Escherichia coli.* Gene **12**:25–31.

WALTER M. FITCH, reviewing editor

Received August 28, 1990; revision received December 13, 1990

Accepted December 17, 1990