

MULTIPLE SEQUENCE ALIGNMENT

ANALORENA CIFUENTES



Published online March 19, 2004

1792-1797 *Nucleic Acids Research*, 2004, Vol. 32, No. 5
DOI: 10.1093/nar/gkh340

MUSCLE: multiple sequence alignment with high accuracy and high throughput

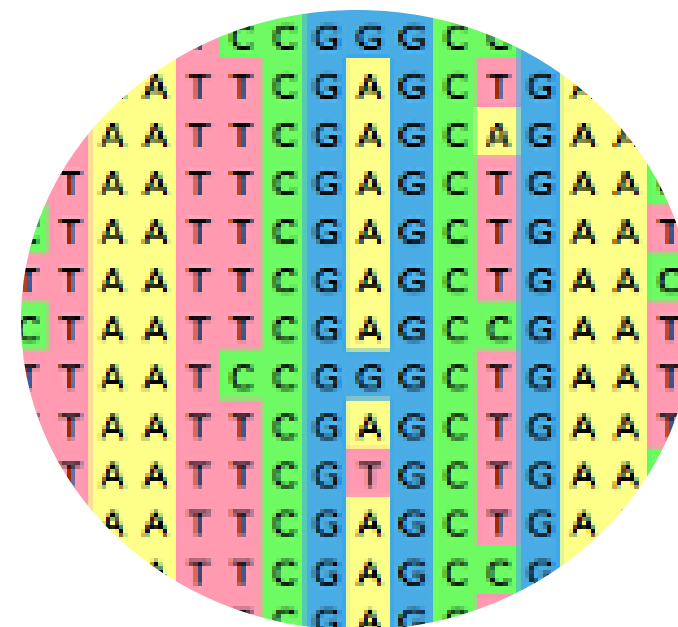
Robert C. Edgar*

195 Roque Moraes Drive, Mill Valley, CA 94941, USA

Received January 19, 2004; Revised January 30, 2004; Accepted February 24, 2004

MUSCLE


Stage 1, Draft progressive
Stage 2, Improved progressive
Stage 3, Refinement



MUSCLE Algorithm

kmer distance (unaligned)
Kimura distance (aligned)

$$\text{LE}^{xy} = (1 - f_G^x) (1 - f_G^y) \log \sum_i \sum_j f_i^x f_j^y p_{ij} / p_i p_j \quad \mathbf{1}$$

frequency of gaps *frequency of i in column x of the first profile*


This is a modified version of the log-average function (15):

$$\text{LA}^{xy} = \log \sum_i \sum_j \alpha_i^x \alpha_j^y p_{ij} / p_i p_j \quad \mathbf{2}$$

The estimated probability a_{xi} of observing amino acid i in position x can be derived from f_x

probability of i

joint probability of i and j being aligned to each other

$i, j = \text{amino acid types}$

Alignment estimation methods

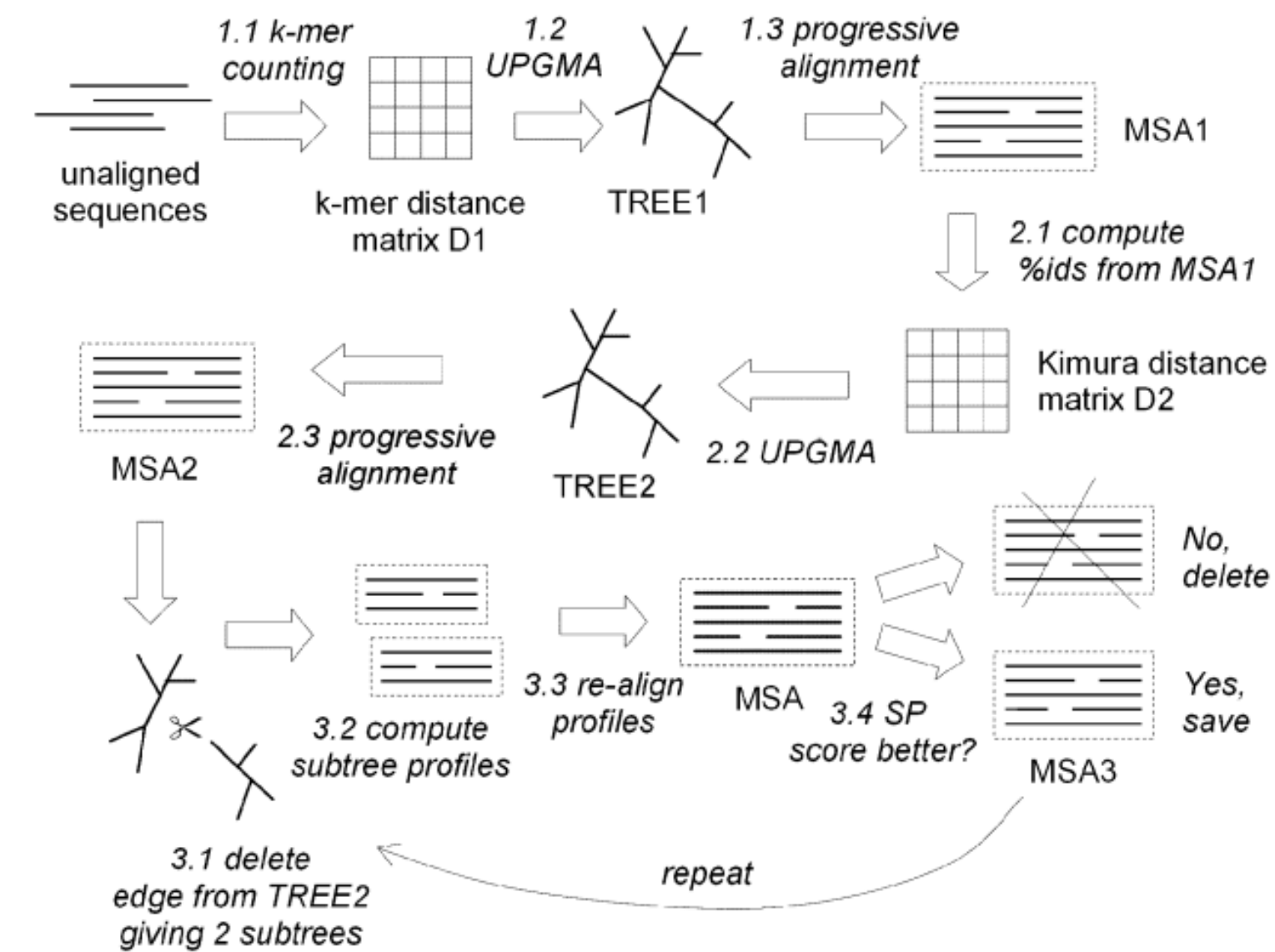


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

accuracy measures

Table 1. BALiBASE scores and times

Method	Q	TC	CPU
MUSCLE	0.896	0.747	97
MUSCLE-p	0.883	0.727	52
T-Coffee	0.882	0.731	1500
NWNSI	0.881	0.722	170
CLUSTALW	0.860	0.690	170
FFTNS1	0.844	0.646	16

quality total column score time



MUSCLE = 7 minutes (5000sq)

FFTNS1 = 10 minutes
CLUSTAL W = 1 year

MUSCLE software, freely available at:
<https://www.drive5.com/muscle/>

Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees

Kevin Liu,¹ Sindhu Raghavan,¹ Serita Nelesen,¹ C. Randal Linder,² Tandy Warnow^{1*}

SATé



Simultaneous alignment and tree estimation

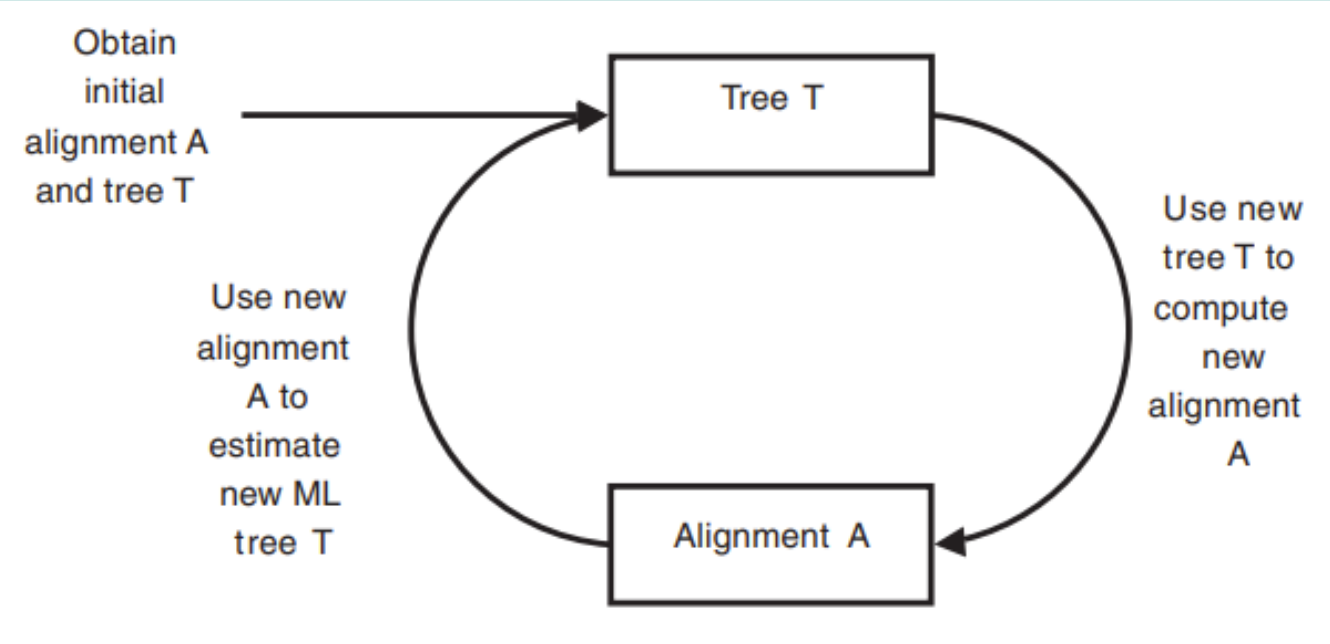


Fig. 1. SATé’s second stage. Beginning with the current best tree/alignment pair, SATé iterates between realigning on the current tree and estimating a RAxML tree for each new alignment. At the end of each iteration, the tree/alignment pair optimizing ML under the GTR+Gamma model of evolution is saved.

Coestimation

24 hours

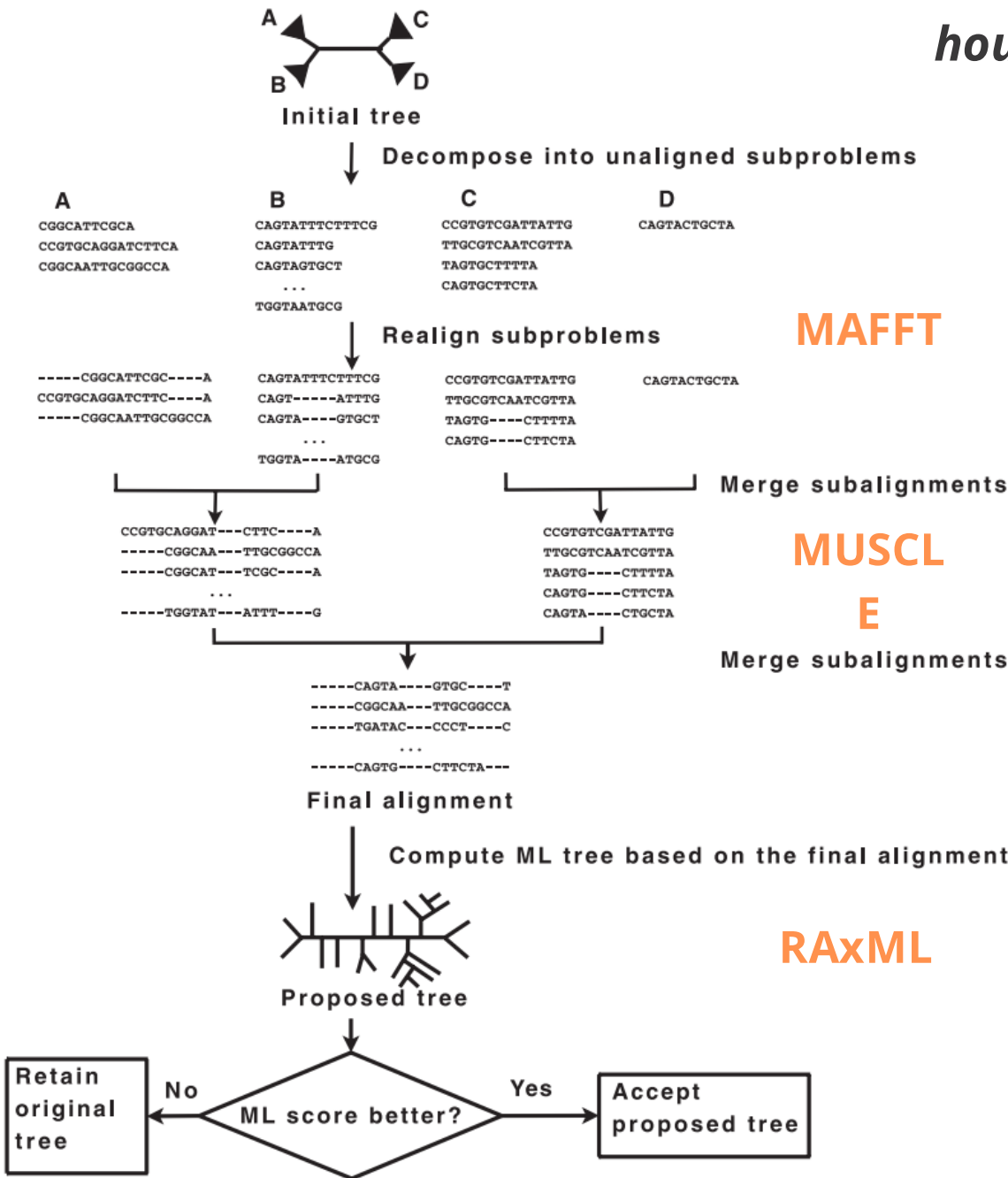


Fig. 2. SATé’s divide-and-conquer strategy, illustrated with a CT-2 decomposition. A branch in the initial tree is selected, and the subtrees—A, B, C, D—around the branch are determined. The sequences for each of the four subtrees are realigned by MAFFT. These realigned subproblems are then aligned with one another, two at a time, using Muscle, until an alignment on the full data set is obtained. RAxML then computes a tree on the alignment. SATé iterates this process, as shown in Fig. 1.

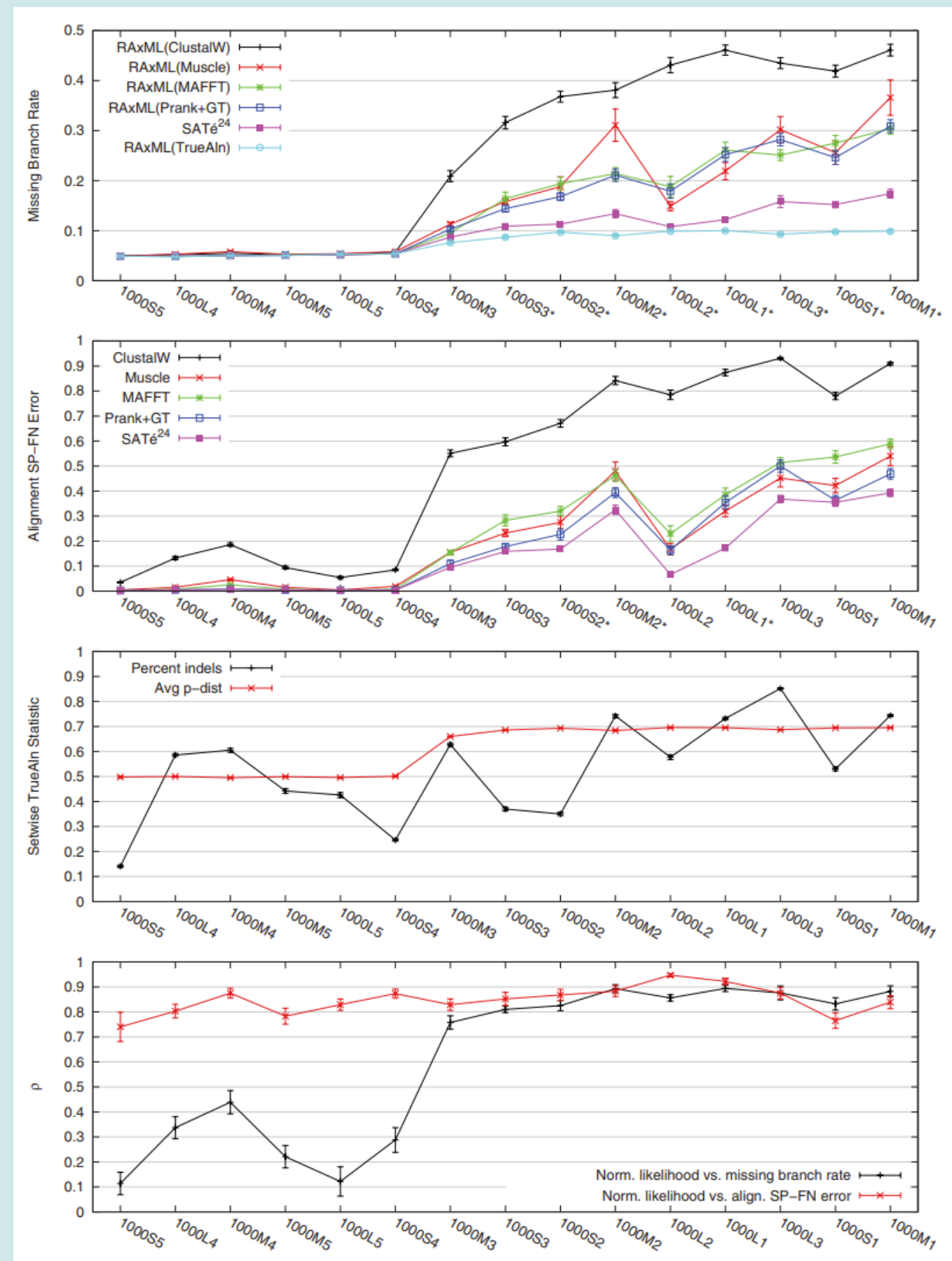


Fig. 3. The 1000-taxon model results. All x axes show the 15 1000-taxon models, from easy to hard, based on missing branch rates. From top-to-bottom panels, the y axes are missing branch rate, alignment SP-FN error, true alignment setwise statistics, and Spearman rank correlation coefficients (ρ). All data points include SE bars. For the top two panels, models on the x axis followed by an asterisk indicate that SATé's performance was significantly better than the nearest two-phase method (paired t tests, setwise $\alpha = 0.05$, $n = 40$ for each test).

Method	All models % (SE)	Moderate to difficult		Easy	
		1000 % (SE)	500 % (SE)	1000 % (SE)	500 % (SE)
Missing branch rate					
RAxML(TrueAln)	7.4 (0.1)	9.3 (0.1)	8.4 (0.1)	5.1 (0.1)	5.3 (0.1)
SATé	*9.1 (0.2)	*13.1 (0.3)	*10.4 (0.2)	*5.1 (0.1)	5.4 (0.1)
RAxML(Prank+GT)	13.0 (0.3)	21.0 (0.6)	15.2 (0.4)	*5.1 (0.1)	*5.3 (0.1)
RAxML(MAFFT)	12.6 (0.4)	21.6 (0.6)	13.5 (0.3)	*5.1 (0.1)	*5.3 (0.1)
RAxML(Muscle)	15.3 (0.5)	22.9 (1.0)	20.5 (0.9)	5.4 (0.1)	5.8 (0.1)
RAxML(ClustalW)	21.6 (0.6)	38.7 (0.7)	21.6 (0.6)	5.3 (0.1)	5.7 (0.1)
Alignment SP-FN error rate					
SATé	*14.2 (0.6)	*25.0 (1.3)	*21.3 (0.7)	*0.4 (0.02)	*1.0 (0.1)
Prank+GT	18.5 (0.8)	30.6 (1.2)	29.9 (1.1)	*0.4 (0.02)	1.1 (0.1)
MAFFT	20.6 (0.8)	38.6 (1.3)	28.5 (1.0)	0.9 (0.1)	1.7 (0.1)
Muscle	22.7 (0.9)	34.0 (1.4)	38.5 (1.5)	1.8 (0.1)	2.9 (0.2)
ClustalW	46.9 (1.3)	77.1 (1.1)	64.1 (1.2)	9.8 (0.5)	12.9 (0.6)

*Improved tree & alignment accuracy
(rapid & accurate)*

- ✓ SATé & MAFFT alignments were of better than other alignments (accuracy)
- ✓ Recommend multiple runs of SATé using more than one starting pair, e.g., using SATéBML if time permits.
- ✓ SATé = Fast, effective, fully on large datasets, and rapid evolving sequences.

SATé

Joint Bayesian Estimation of Alignment and Phylogeny

Benjamin D. Redelings and Marc A. Suchard

Joint Bayesian Framework

- Objective: Simultaneously estimate multiple sequence alignment & phylogenetic tree.
- Key Innovation: Integrates over all possible alignments to incorporate alignment uncertainty.

Traditional Approach Issues

- Overconfidence from using a single “best” alignment.
- Bias from guide tree–dependent progressive alignment

Probabilistic Model

likelihood calculated from the substitution model over the alignment

$$P(Y, A, \tau, T, \Theta, \Lambda) = \boxed{P(Y|A, \tau, T, \Theta)} \times \boxed{P(A|\tau, \Lambda)} \times P(\tau, T) \times P(\Theta) \times P(\Lambda)$$

Y = the observed unaligned sequence data

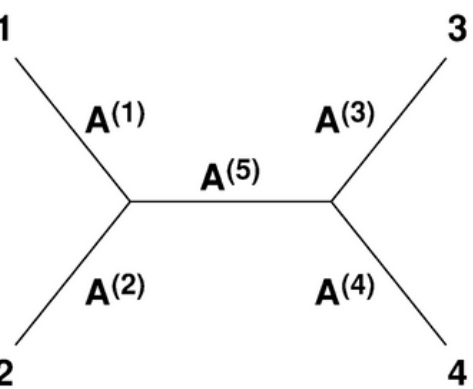
A = multiple sequence alignment

τ = unrooted tree topology

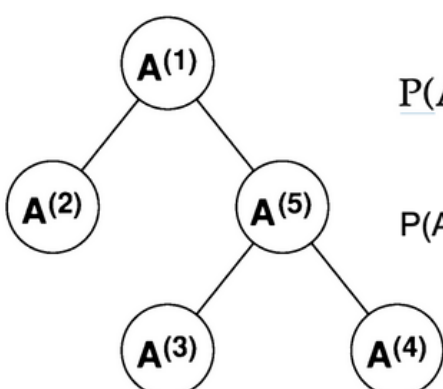
T = branch lengths

Θ = the substitution model parameters

Λ = comprises the indel (gap) process parameters



(a) Pairwise branch alignments



(b) DAG

$$P(A | \tau, \Lambda) = P(A^{(1)} | \Lambda) \prod_{b=2}^B P(A^{(b)} | A^{(\rho(b))}, \Lambda). \quad (6)$$

$$P(A | \tau, \Lambda) = P(A^1 | \Lambda) \times P(A^2 | A^1 \Lambda) \times P(A^5 | A^1 \Lambda) \times P(A^3 | A^5 \Lambda) \times P(A^4 | A^5 \Lambda)$$

$$P(A | \tau, \Lambda) = P(A^{(1)} | \Lambda) \prod_{b=2}^B P(A^{(b)} | |A_{n(b)}^{(b)}| = |a_{n(b)}^{(\rho(b))}|, \Lambda), \quad (7)$$

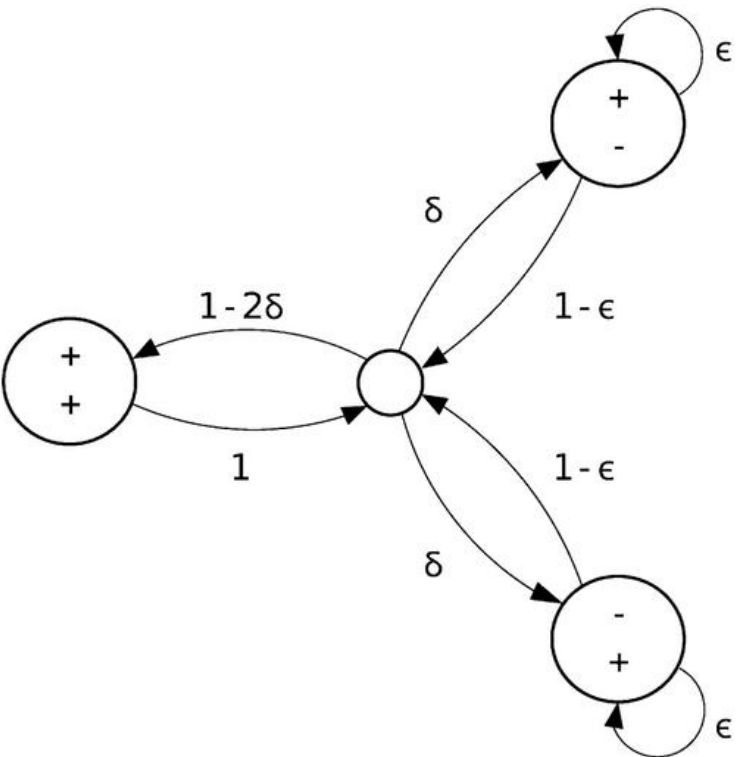


FIGURE 3. Hidden Markov model for pairwise alignments. The start and end states of the model are not shown. Every state emits (+) or does not emit (-) a residue in each of the two sequences. After a match (+/+) or a gap (+/- or -/+) ends, the chain returns to the silent state in the center. From there, a gap in either sequence opens with probability δ. Existing gaps are extended with probability, resulting in geometrically distributed gap lengths with mean length 1/(1 - ?). Transition probabilities shown are conditional on not moving to the end state. The silent state is shown here for clarity, but it can be removed, resulting in transitions only between nonsilent states.

Alignment Uncertainty (AU) Plots

- Point estimate of the alignment
- Each position is color-coded to represent the probability that it's correctly aligned.
- Darker colors indicate higher confidence, while lighter colors highlight ambiguous regions.
- This allows researchers to quickly identify areas of uncertainty in the alignment.

$$\begin{aligned} P(A | \tau, \Lambda) &= P_v(A^{(1)}) \prod_{b=2}^B P_v(A^{(b)} | |A_{n(b)}^{(b)}| = |a_{n(b)}^{(\rho(b))}|) \\ &= P_v(A^{(1)}) \prod_{b=2}^B \frac{P_v(A^{(b)}) \times 1\{|A_{n(b)}^{(b)}| = |a_{n(b)}^{(\rho(b))}|\}}{P_v(|A_{n(b)}^{(b)}| = |a_{n(b)}^{(\rho(b))}|)}, \quad (8) \end{aligned}$$

Empirical Findings

5S rRNA Data:

- AU plots reveal well-resolved vs. ambiguous regions.
- Fixed alignments can lead to biased phylogenetic inference.

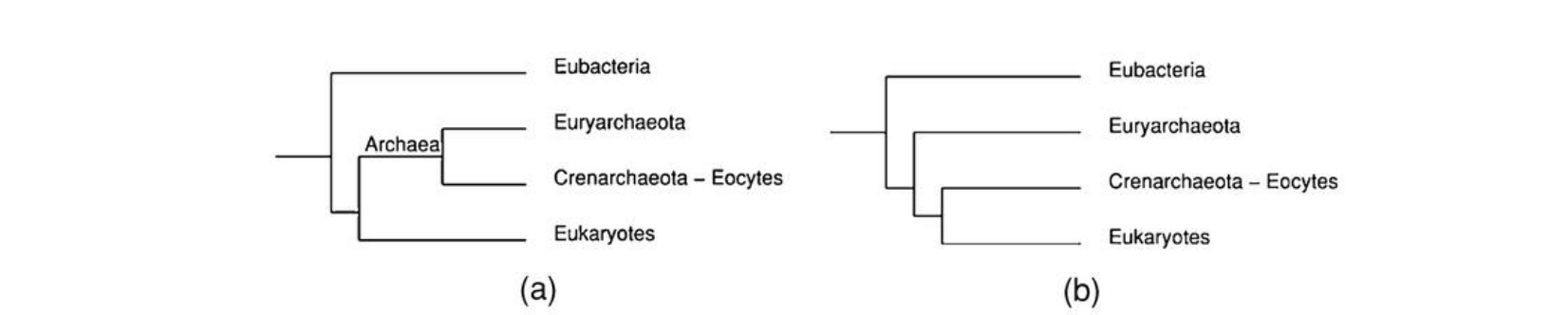


FIGURE 5. Two possibilities for early branching in the Tree of Life. (a) The archaeal tree implies that Archaea form a monophyletic group. (b) The eocyte tree implies that Archaea are paraphyletic (Eocytes and Euryarchaeota) and that the eocyte Archaea are more closely related to Eukaryotes than to other Archaea.

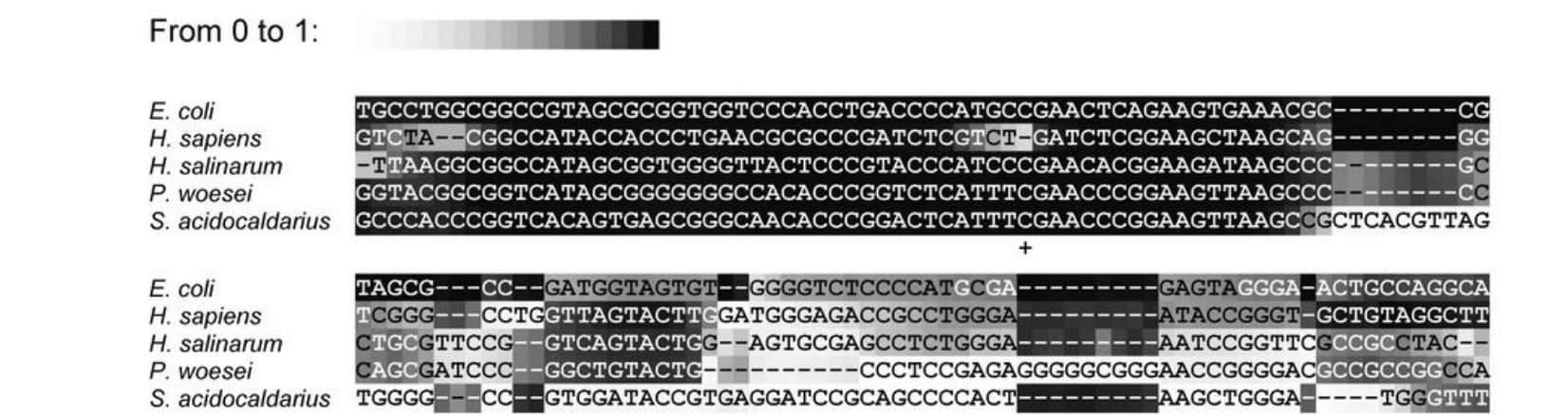


FIGURE 6. Alignment uncertainty plot for the 5S rRNA. Dark shading indicates that an entry is well resolved; light shading indicates that the position is not well resolved. The first half of the alignment is fairly well resolved, but the second half is much less well resolved. Uncertainty in the exact position of gaps is visible as light shading in adjacent letters. *S. acidocaldarius* does not align well with the other sequences.

TABLE 2. Support for the most probable 5-taxon 5S rRNA topologies and partitions. Columns report the posterior probability (PP) and \log_{10} odds (LOD) with their 95% Bayesian credible interval (BCI) in favor of each hypothesis under three different models. These models are the joint estimation model presented in this paper (JE), a joint model constrained to a fixed alignment (Indels), and a traditional model based on a fixed alignment (NoIndels). Fixed alignments are estimated using ClustalW. Although alignments are fixed, the second model does make use of indel information while inferring the phylogeny, whereas the latter model does not. Taxa abbreviations are given in Table 1 and BCIs are estimated using a block bootstrap (Suchard et al., 2003b).

Topologies and partitions	JE			Indels			NoIndels		
	PP	LOD	95% BCI	PP	LOD	95% BCI	PP	LOD	95% BCI
((EC,HS),(SA,(PW,HA)))	0.308	−0.35	(−0.38, −0.32)	0.996	+2.38	(+2.33, +2.43)	0.172	−0.68	(−0.69, −0.67)
((EC,HS),((SA,PW),HA))	0.208	−0.58	(−0.63, −0.54)	0.002	−2.71	(−2.77, −2.65)	0.700	+0.37	(+0.36, +0.38)
(EC,((HS,SA),(PW,HA)))	0.120	−0.86	(−0.90, −0.83)	0.001	−2.89	(−3.02, −2.80)	<0.001	−4.88	< −4.70
((EC,PW),((SA,HS),HA))	0.088	−1.02	(−1.08, −0.97)	<0.001	<−5.17	<−5.17	<0.001	<−5.17	<−5.17
((EC,SA),(HS,(PW,HA)))	0.066	−1.15	(−1.21, −1.10)	0.001	−3.12	(−3.30, −3.11)	<0.001	−4.03	(−4.40, −3.86)
((EC,HS),(PW,(SA,HA)))	0.037	−1.42	(−1.47, −1.37)	<0.001	−3.47	(−3.62, −3.37)	0.127	−0.84	(−0.84, −0.83)
EC,HS HA,PW,SA	0.553	+0.09	(+0.06, +0.13)	0.998	+2.72	(+2.64, +2.82)	>0.999	+3.78	(+3.50, +4.06)
EC,HS,SA PW,HA	0.494	−0.01	(−0.04, +0.02)	0.998	+2.64	(+2.59, +2.70)	0.173	−0.68	(−0.69, −0.67)

EF-1 α /Tu Protein Data:

- Analysis (both 5-taxon and 12-taxon datasets) supports the “eocyte” hypothesis.
- AU plots highlight both clear regions and ambiguous gaps (e.g., due to a divergent taxon).

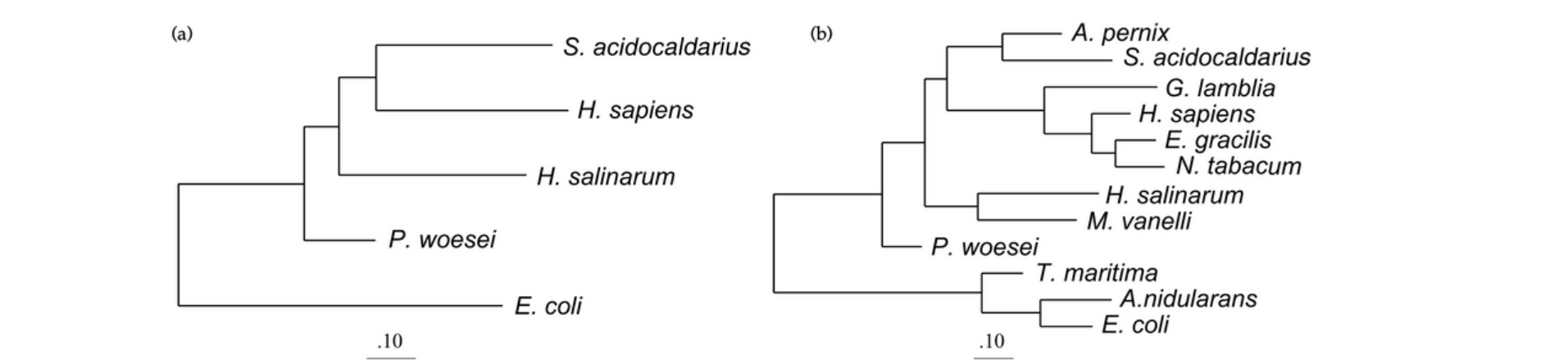
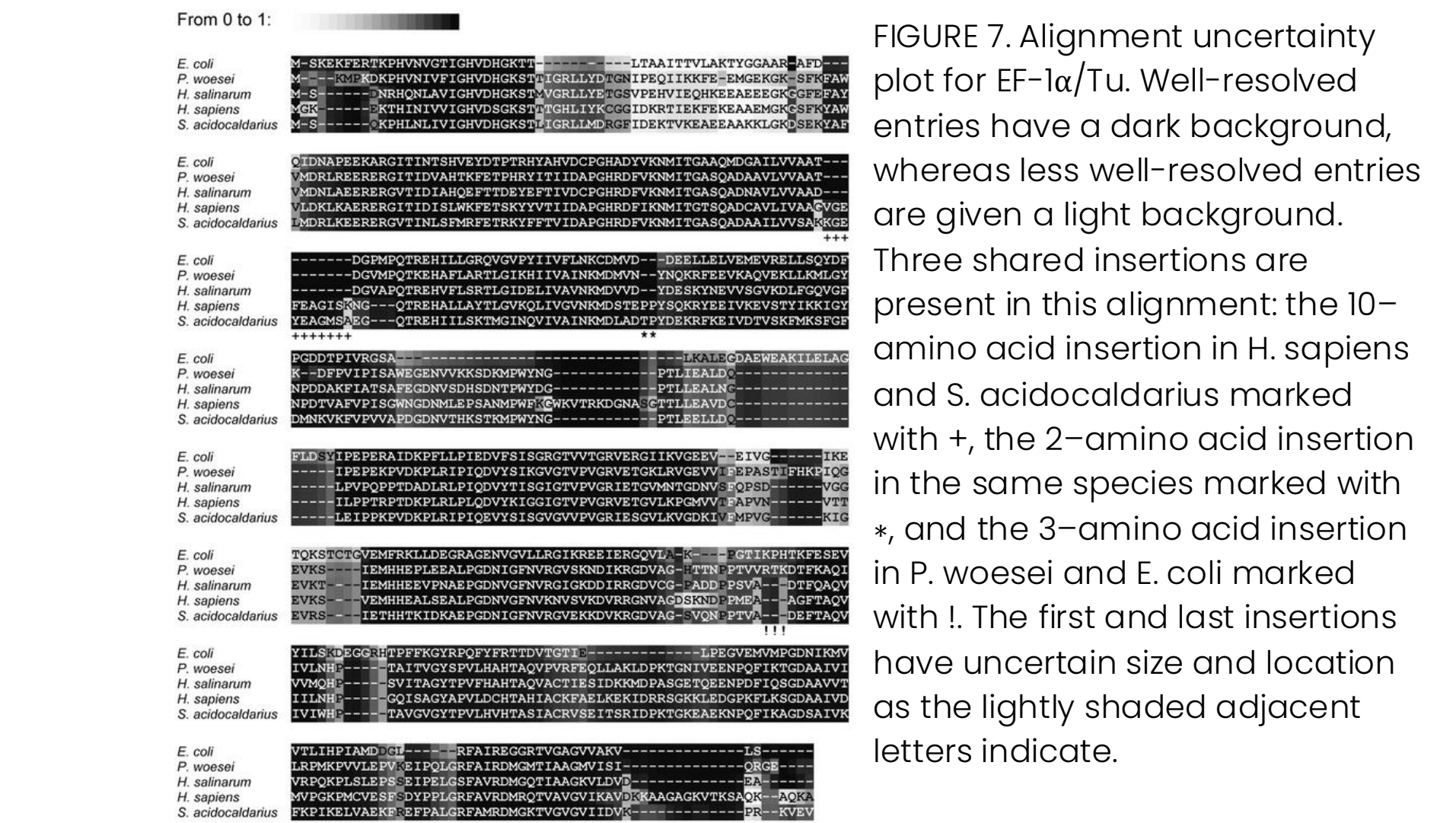


FIGURE 8. MAP topologies for EF-1 α /Tu 5- and 12-taxon data sets. Reported branch lengths are posterior means. The 12-taxon tree (b) is consistent with the 5-taxon tree (a). Both trees support the eocyte hypothesis. Both trees also place *P. woesei* closer to the root and in a separate clade from *H. salinarum*.



Thank you!

Quiz:

1. What is the difference between SATé and muscle methods?
2. ----- distance and ----- distance are the 2 distances measures that MUSCLE use.
3. Why MSA are important in phylogenetics?
4. What assumptions do Redelings and Suchard make about indels along the branches of the phylogenetic tree?
5. What is an AU plot?