PROBABILISTIC PHYLOGENETIC INFERENCE WITH INSERTIONS AND DELETIONS

E.R.
Sean Eddy

Generative birth-death model of sequence evolution

based on a (N+1) Markov process

→ no memory effect (alig/seqs of arbitrary length)

→ no fixed gap frequency

Assumes insertions occur one at the time

→ Allows an efficient $\Theta(n \cdot L)$ algorithm for P(alignment | Tree, Model)

Pratical application to phylogenetic reconstruction

(phylip):

DNAML ⇝ DNAMLE

ignores gaps

includes gaps as per above model

Same time complexity
same results for ungapped alignments

Test on real rRNA alignments

CONCORDANCE TEST

The (N+1) Markov process

Durbin et al. '98

extended rate:

$$R^{\varepsilon} \equiv \begin{array}{c|cc|c} & 1 \ \dots \ \kappa & & - \\ \hline & R-\mu\delta_{ij} & \mu & 1 \\ & & \vdots & \vdots \\ & & \mu & \kappa \\ \hline \lambda\pi_1 \ \dots & \lambda\pi_\kappa & -\lambda & - \end{array}$$

$$e^{tR^{\varepsilon}} = \begin{array}{c|cc} & P_t^{\varepsilon}(j|i) & \gamma_t \\ & & \vdots \\ & & \gamma_t \\ \hline \xi_t\pi_1 \ \dots & \xi_t\pi_\kappa & 1-\xi_t \end{array}$$

If R reversible:

Prob INSERTION: $\xi_t = \frac{\lambda}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right)\Big|_{\lambda=\mu=0} = 0$

Prob DELETION: $\gamma_t = \frac{\mu}{\lambda+\mu}\left(1-e^{-(\lambda+\mu)t}\right)\Big|_{\mu=\lambda=0} = 0$

$$\left[e^{tR}\right]_{ij} = P_t(j|i) = \pi_j + \sum_{a=1}^{A} Q_a(i,j)\, e^{-\varepsilon_a t}$$

R eigenvalues $\begin{cases} 0 \longrightarrow 0 \\ -\varepsilon_a \longrightarrow -(\varepsilon_a+\mu) \\ \longrightarrow -(\lambda+\mu) \end{cases}$ R^{ε} eigenvalues

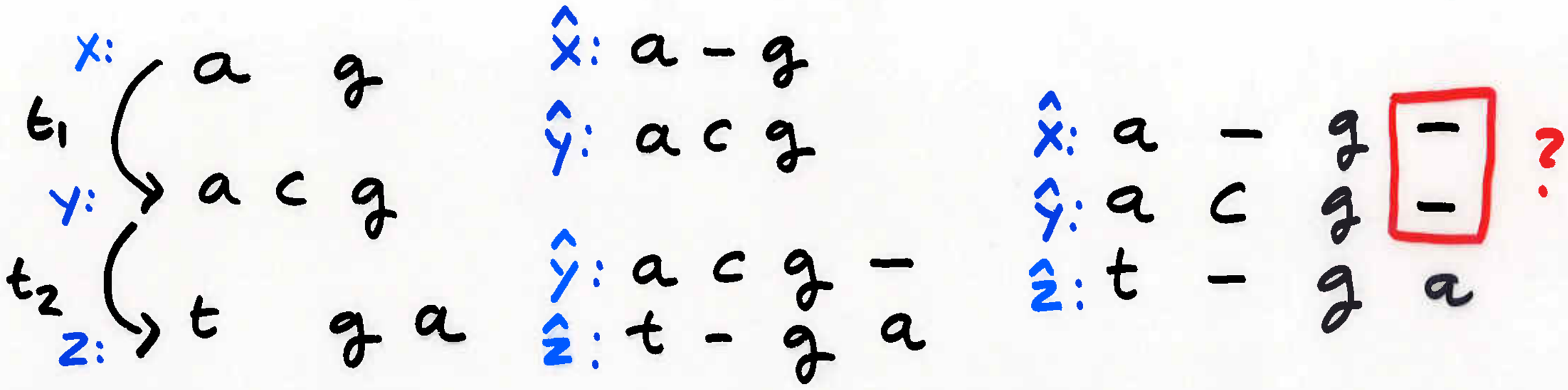
Prob SUBSTITUTION

$$P_t^{\varepsilon}(j|i) = \frac{\lambda}{\lambda+\mu}\pi_j + \sum_{a=1}^{A} Q_a(i,j)\, e^{-(\varepsilon_a+\mu)t} + \frac{\mu}{\lambda+\mu}\pi_j e^{-(\lambda+\mu)t}\Big|_{\lambda=\mu=0} = P_t(j|i)$$

McGuire et al. '03 [R=FBM, λ=β, rev]

The Generative Model

Ignore the (=) transitions $(1-\xi_t)$



$$\rightarrow P_{t_1+t_2}(z \mid x,y) = P_{t_1}(y, \hat{x}\hat{y} \mid x) \cdot P_{t_2}(z, \hat{y}\hat{z} \mid y)$$

Borrowed from Markov model

$$P_{t_1}(y, \hat{x}\hat{y} \mid x) \propto \overbrace{\xi_{t_1}\pi_c}^{ins}\ P^{\varepsilon}_{t_1}(g \mid g)\ P^{\varepsilon}_{t_1}(a \mid a)$$

$$P_{t_2}(z, \hat{y}\hat{z} \mid y) \propto \underbrace{\gamma_{t_2}}_{del}\ \underbrace{\xi_{t_2}\pi_a}_{ins}\ \underbrace{P^{\varepsilon}_{t_2}(g \mid g)}_{subs}\ P^{\varepsilon}_{t_2}(t \mid a)$$

Normalization ?

s-substitutions
d-deletions
i-insertions

$$P_t(\ell' = s+i \mid \ell = s+d) \propto \frac{(s+d+i)!}{s!\,d!\,i!}\,\gamma_t^{\,d}\,(1-\gamma_t)^{s}\,\xi_t^{\,i}$$

$$\sum_{s=0}^{\infty}\sum_{d=0}^{\infty}\sum_{i=0}^{\infty} P(\ell' = s+i \mid \ell = s+d) = 1 \Rightarrow$$

$$P(\ell' \mid \ell) = (1-\xi_t)^{\ell+1}\,\frac{(\ell+i)!}{s!\,(\ell-s)!\,(\ell'-s)!}\,\gamma_t^{\,\ell-s}\,(1-\gamma_t)^{s}\,\xi_t^{\,\ell'-s}$$

The Generative Model

$$P_t(Y_{\ell'}, \hat{x}\hat{y} \mid X_\ell) = (1-\xi_t)^{\ell+1}\, \gamma_t^{\ell-s}\, \xi_t^{\ell'-s}\, \pi_{I_1} \cdots \pi_{I_{\ell'-s}} \prod_{k=1}^{s} P_t^{\varepsilon}(Y_{\sigma_k} \mid X_{\sigma_k})$$

Column Factorization

$$P_t(y\ \hat{x}\ \hat{y} \mid x) = (1-\xi_t)\Big(\prod_{i=1}^{\ell'-s} P_t^{q}(I_i \mid -)\Big)\Big(\prod_{d=1}^{\ell-s} P_t^{q}(- \mid x_{\sigma_d})\Big)\Big(\prod_{k=1}^{s} P_t^{q}(y_{\sigma_k} \mid x_{\sigma_k})\Big)$$

$\ell'-s$ INS $\ell-s$ DEL s SUBS

extra term

$$\begin{cases} P^q(i \mid -) = \xi_t\, \pi_i \\ P^q(- \mid i) = (1-\xi_t)\,\gamma_t \\ P^q(j \mid i) = (1-\xi_t)\, P_t^{\varepsilon}(j \mid i) \end{cases}$$

Sequence Generator

* a * g *

(ℓ+1) places: $*$ —$(1-\xi_t)\,\xi_t^{n}$→ $* I_1 * \ldots * I_n *$

ℓ residues: a —$P_t^{\varepsilon}(a' \mid a)$→ a' ; —γ_t→ death of right link (after it emits)

$$\langle \ell' \mid \ell \rangle = \ell(1-\gamma_t) + (\ell+1)\,\frac{\xi_t}{1-\xi_t}$$

$t=0$ → ℓ

$t \uparrow \infty$ → $\ell\,\frac{\lambda}{\lambda+\mu} + (\ell+1)\,\frac{\lambda}{\mu}$

$\mu=0$ → ∞ ; $\lambda=0$ → 0

Probability ancestral/descendant alignment

$$P(\hat{x}, \hat{y}) = P(y, \hat{x}\hat{y} \mid x) \cdot P(x)$$

generative model

distribution of ancestral segs.

arbitrary: $(1-p)\,p^{\ell}$

length distributions

$$P(\ell') = (1-q_t)\,q_t^{\ell'} \quad \text{descendant segs} \quad // \quad q_t = \frac{\xi_t + p(1-\xi_t)(1-\gamma_t)}{1-p\gamma_t(1-\xi_t)}$$

$$P(L) = (1-r_t)\,r_t^{L} \quad \text{alignments} \quad // \quad r_t = \xi_t + p(1-\xi_t)$$

expected frequencies

$$f_i(t) \equiv \frac{\langle L\rangle - \langle \ell\rangle}{\langle L\rangle} = \frac{\xi_t}{\xi_t + p(1-\xi_t)} \quad \text{insertions}$$

$$f_d(t) \equiv \frac{\langle L\rangle - \langle \ell\rangle'}{\langle L\rangle} = \frac{p\gamma_t(1-\xi_t)}{\xi_t + p(1-\xi_t)} \quad \text{deletions}$$

(i) Model is non-reversible : $f_i \neq f_d \wedge P(\ell) \neq P(\ell')$

(ii) Model is non-stationary : $P(\ell')$ is time dependent

$\lambda = 0$	$f_i = 0$	$\lambda \neq 0$	$f_i = 1 - e^{-\lambda t}$
$\mu \neq 0$	$f_d = 1 - e^{-\mu t}$	$\mu = 0$	$f_d = 0$

arbitrary rates λ, μ require nonreversible models for gaps (no indels)

Comparison to TKF91

Similar set-up:

i) probabilistic modelling ins/del

ii) a generative birth-death model

iii) $\lambda \geq 0$, $\mu \geq 0$ rates ins/dels

Departure:

TKF91 is reversible

$$P(\text{ancestral seq } \ell) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{\ell}$$

$\lambda < \mu$

all seqs have same length on average

TKF91 not invariant by column rearrangements

adds realism ~ increases algorithmic complexity

Actual correspondence:

$$P\left(\underbrace{\begin{matrix}\# & - & & - \\ \# & \# & \cdots & \#\end{matrix}}_{n}\right) = \frac{\lambda}{\mu} p_1^{t} \beta_t^{n-1}$$

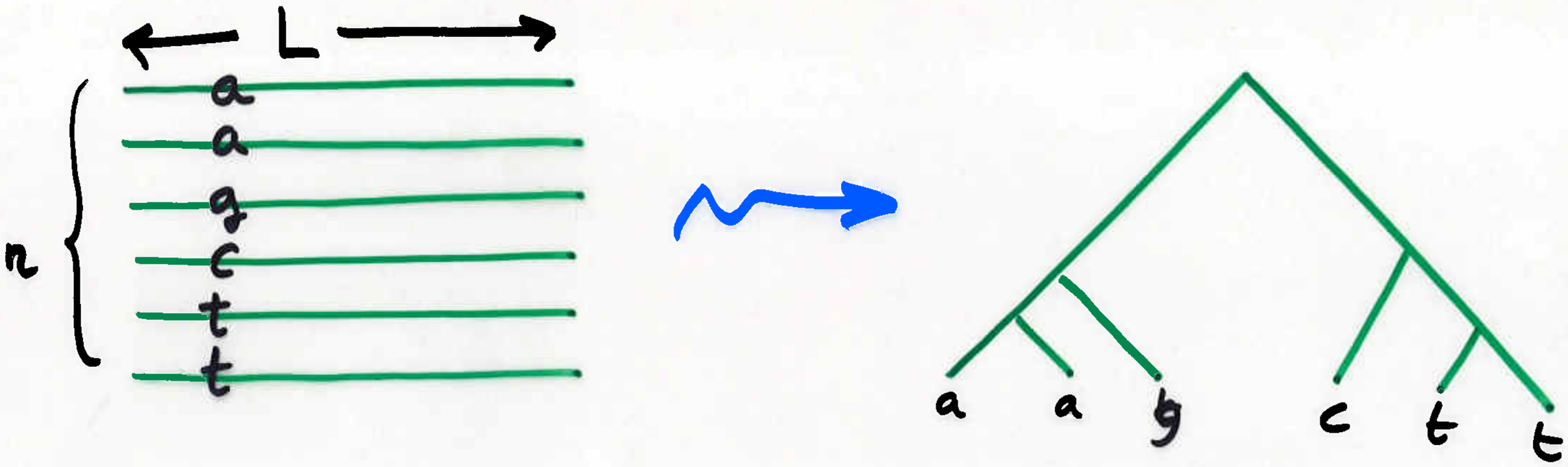
$$P\left(\underbrace{\begin{matrix}\# & - & & - \\ - & \# & \cdots & \#\end{matrix}}_{n}\right) = \begin{cases} \frac{\lambda}{\mu} q_1^{t} \beta_t^{n-1} & n > 0 \\ \beta_t & n = 0 \end{cases}$$

$$P\left(\underbrace{\begin{matrix}- & & - \\ \# & \cdots & \#\end{matrix}}_{n}\right) = \beta_t^{n}$$

extra term $= \left(1 - \frac{\lambda}{\mu}\right)(1 - \beta_t)$

TKF91	ERATE
$P(1-\varepsilon_t)(1-\gamma_t)\,\varepsilon_t^{n-1}$	
$P(1-\varepsilon_t)\,\gamma_t\,\varepsilon_t^{n}$	
$(1-\varepsilon_t)\,\varepsilon_t^{n}$	
$(1-P)(1-\varepsilon_t)$	

Phylogenetic Inference



$P(\text{alignment} \mid T, t's, R) \sim n \cdot L$

Felsenstein peeling algorithm

At the core of most phylogenetic inference packages:
PHYLIP, MOLPHY, PAUP*, PHYML, Palm, Paxml, Mr Bayes, ...

all treat gaps as missing data

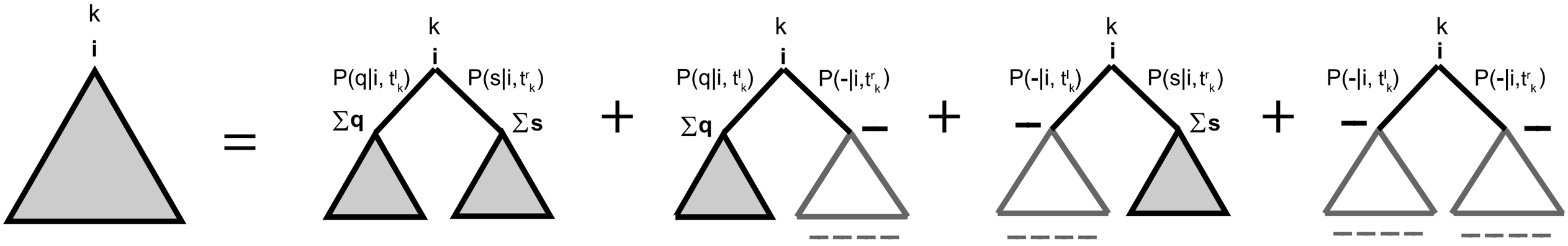
Practical application for ERate

~ the substitutions-only Felsenstein algorithm is extendable to erate model with same time complexity $n L$

~ can calculate the probability of missing data (i.e. events not observed in given sequences) by marginalization

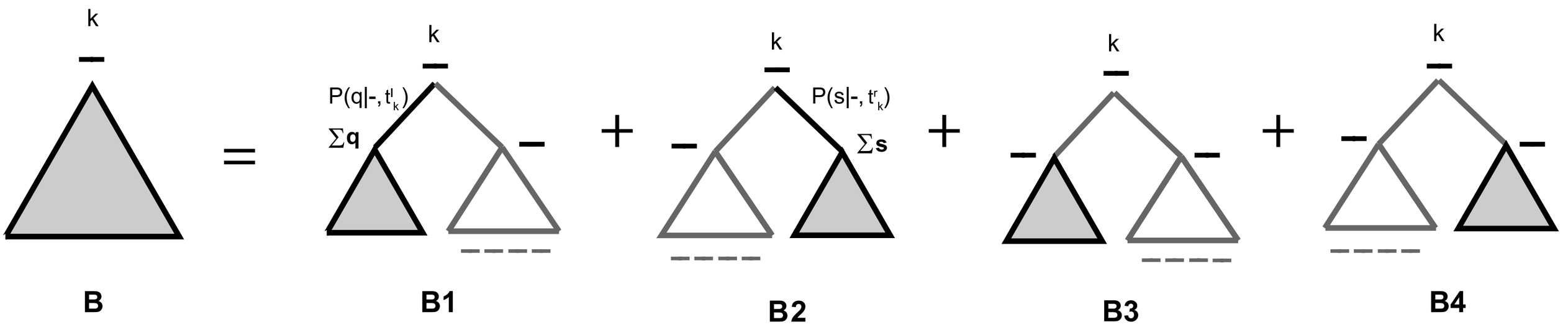
important for ancestral seq reconstruction

Felsenstein algorithm extended to gaps



A **A1** **A2** **A3** **A4**

$P_u(L_k, i)$ **includes only substitutions and deletions**



B **B1** **B2** **B3** **B4**

$P_u(L_k, -)$ **includes one and only one insertion**

Generative model

$$P(j \mid i,\, t) = (1 - \xi_t)\, P_t^{\varepsilon}(j \mid i)$$

$$P(- \mid i,\, t) = (1 - \xi_t)\, \gamma_t$$

$$P(j \mid -,\, t) = \xi_t\, \pi_j$$

Probability of a column u

$$P(u \mid T, R^{\varepsilon}, p) = \underset{\text{(B)}}{P_u(L_{root}, -)} + \; p \sum_{1 \le i \le K} \underset{\text{(A)}}{P_u(L_{root}, i)}\, \pi_i$$

Probability of the alignment **after marginalizing** all unobserved (all gaps) columns

extra-term contribution

$$P(L \mid T, R^{\varepsilon}, p) = \frac{P(\star \mid T, R^{\varepsilon}, p)}{1 - P(u_{\text{gap}} \mid T, R^{\varepsilon}, p)} \prod_{1 \le u \le L} P(u \mid T, R^{\varepsilon}, p)$$

prob of a column with all gaps

DNAML ε

Extension to saps of DNAML (PHYLIP)

DNA alignment + evolutionary model → "maximum likelihood" → phylogenetic tree + branch lengths

a DNAML/DNAMLε comparism is a exclusive test of the effect of including ins/del

~ DNAMLε optimizes rates λμ

~ For unsapped alignments optimizes to λ=μ=0 and produces same results as DNAML

DNAMLε reports practical times

used in Rfam 09 for seed alignments containing up to 64 sequences ~ 92% of families

Tests on Simulated data

for each av. branch length (subs only) (0.005 - 2.0 subs/s.

use F84 to evolve residues

Select method to evolve sequences (ROSE / ERATE)

for each "gap parameters"/length distribution (geometric / poisson / simple

for each n (1, ..., 100)

Generate random 8-Taxon tree T_n (Kuhn-Felsenstein)

for each L (50, ..., 1000 nts) $\Delta L = 5$ nts

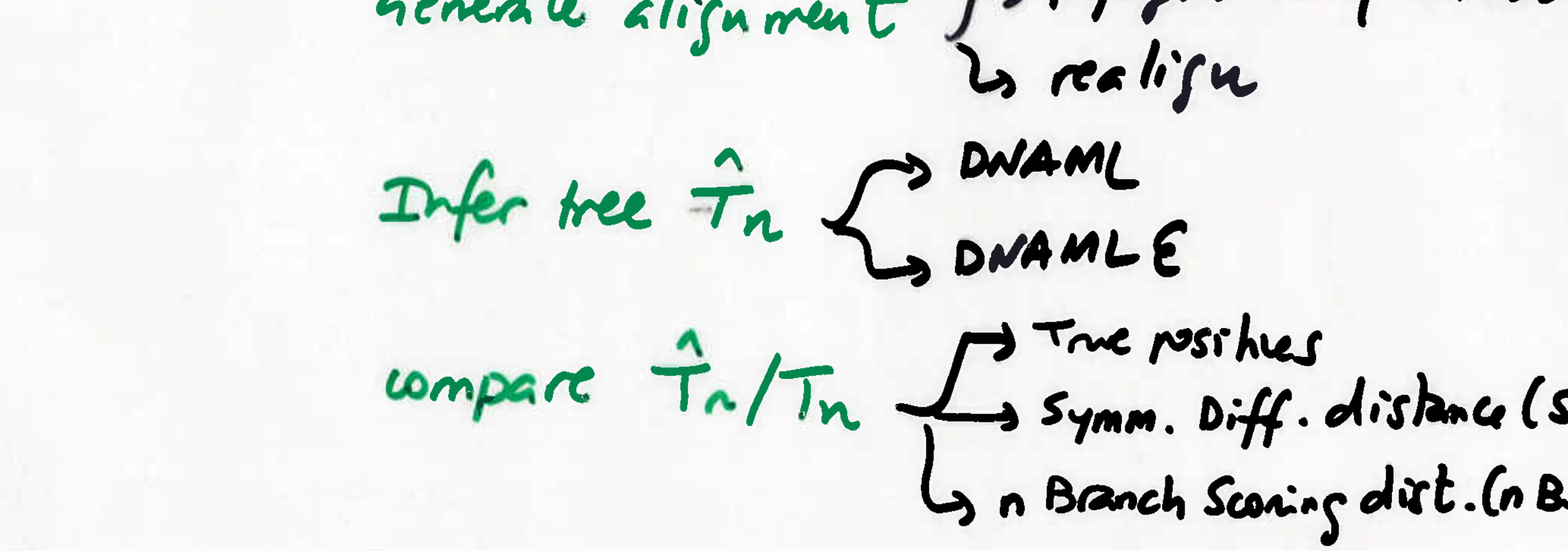
Generate ancestral seq (L) (uniform dist)

Generate alignment → phylogenetically correct
→ realign

Infer tree $\hat{T}_n$ → DNAML
→ DNAML E

compare $\hat{T}_n / T_n$ → True positives
→ Symm. Diff. distance (SDD)
→ n Branch Scoring dist. (nBSD)

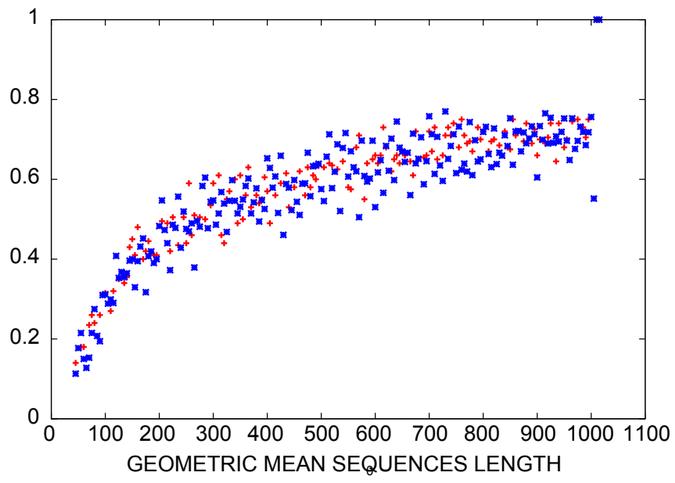
collect stats for all 100/L comparisons

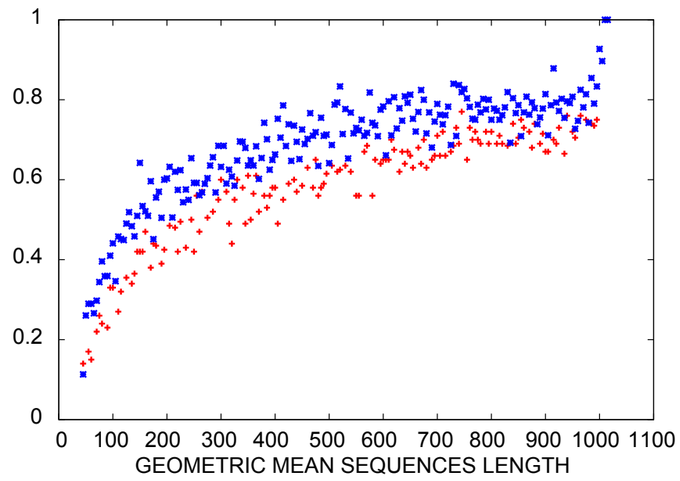


DNAML | 100 RANDOM 8 TAXON TREES | **DNAMLε**

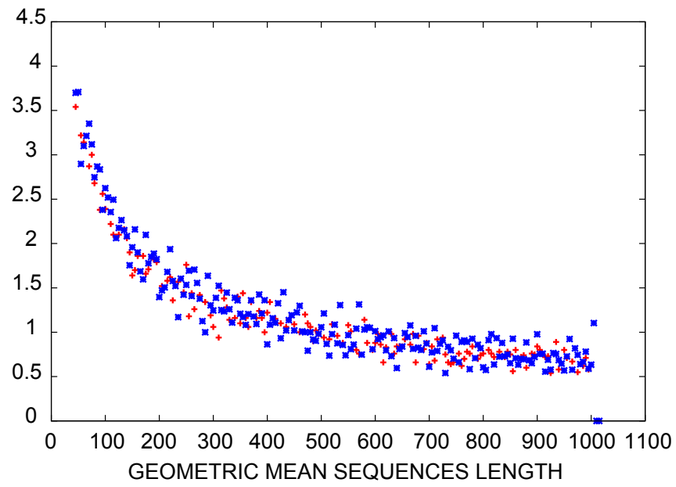
\+ RED ALIGNMENTS: NO GAPS %SUBS: 21 +/- 6 %ID: 79 +/- 6
* BLUE ALIGNMENTS: %GAPS: 13 +/- 4 %SUBS: 20 +/- 5 %ID: 66 +/- 8

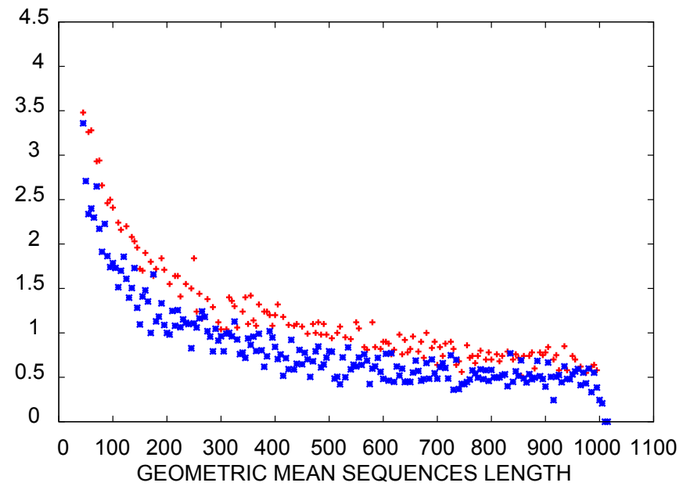
A



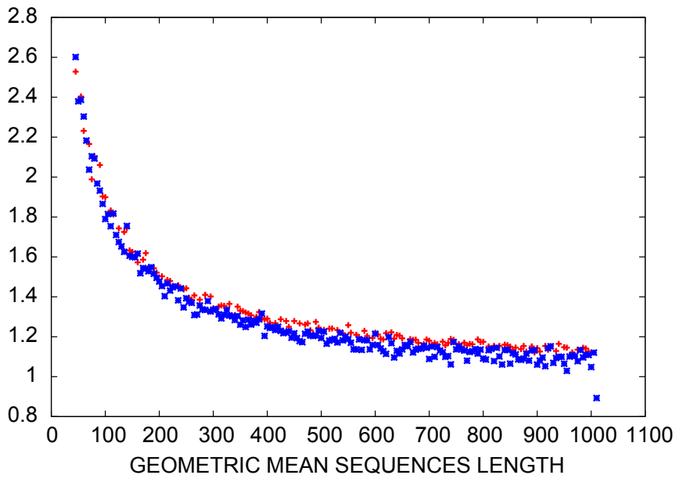


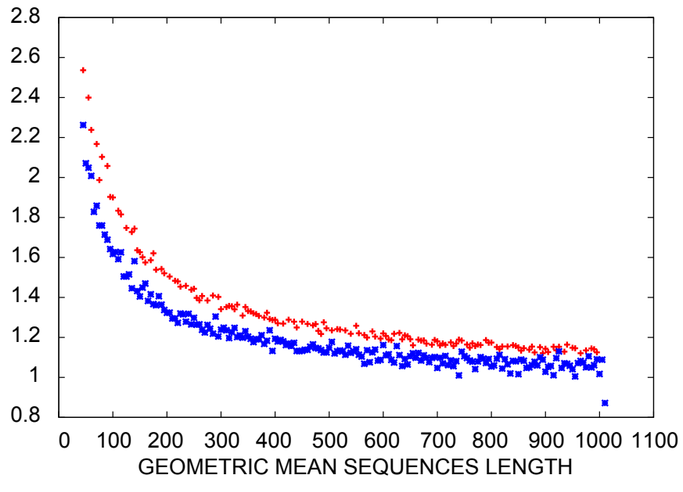
B



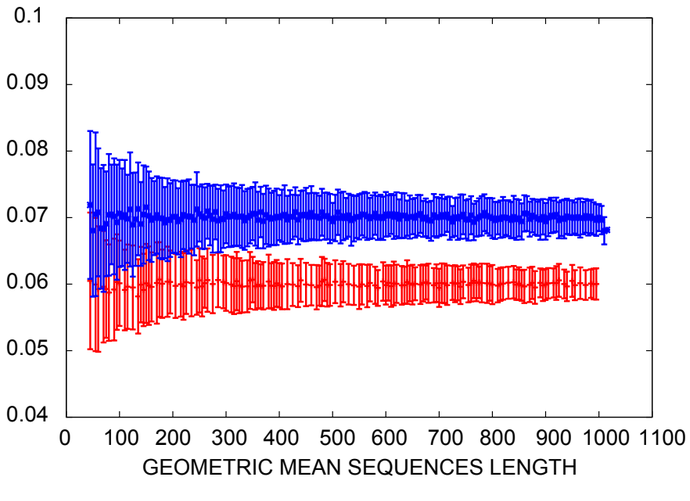


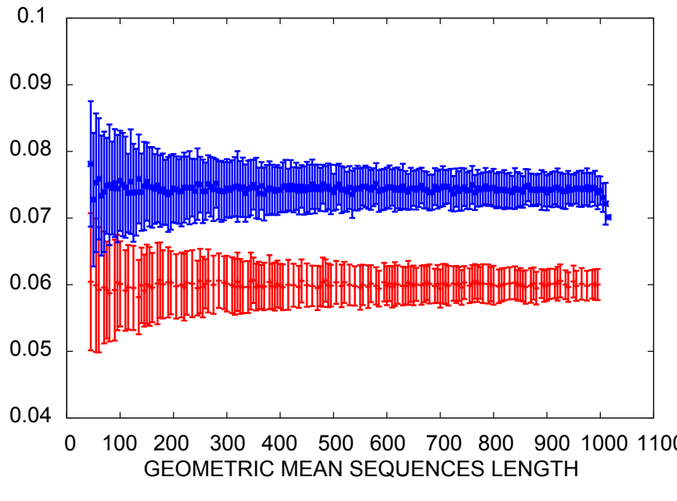
C



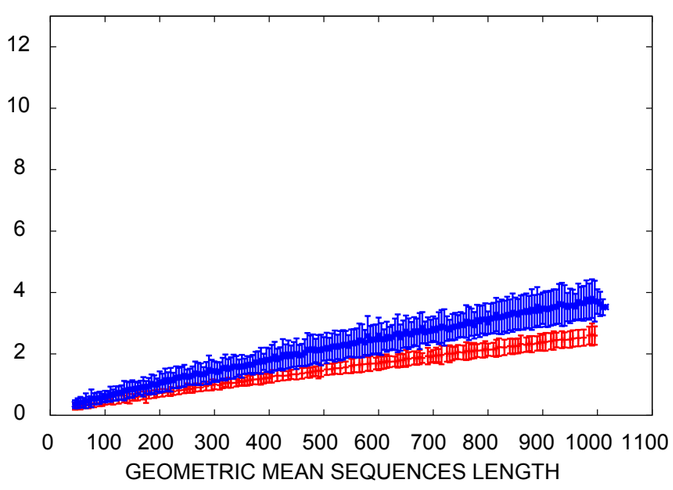


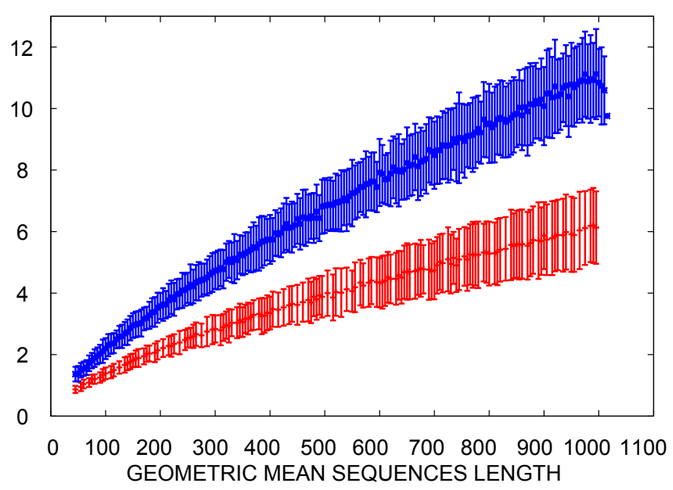
D





E





TEST on rRNA alignments

Concordance test

Break alignment in two pieces

Do both produce the same inferred tree?

rRNA alignment statistics

SSU	Archaea	Chloroplasts	Bacteria	Eukarya	Mitochondria
# seqs	74	75	1601	900	617
Alignment length	1756	2456	3094	7160	4716
Geometric mean seqs	1446	1446	1497	1788	1011
pairwise % ID	71 ± 7	74 ± 11	68 ± 6	56 ± 14	49 ± 20
pairwise % SUBS	23 ± 5	19 ± 7	22 ± 4	20 ± 5	25 ± 6
pairwise % GAP	6 ± 4	7 ± 5	11 ± 5	24 ± 12	26 ± 20
Total % Gaps	17.6	39.0	53.2	74.8	78.0

LSU	Archaea	Chloroplasts	Bacteria	Eukarya
# seqs	26	32	120	89
Alignment length	3346	4311	4555	9055
Geometric mean seqs	2994	2936	2918	3623
pairwise % ID	65 ± 9	70 ± 10	67 ± 6	50 ± 12
pairwise % SUBS	29 ± 7	20 ± 5	25 ± 4	22 ± 5
pairwise % GAP	6 ± 2	11 ± 5	9 ± 4	28 ± 11
Total % Gaps	10.5	31.8	35.9	59.6

Test on a large number of random 8 taxa subalignments

SSU rRNA

38,000 - EIGHT TAXON ALIGNMENTS

geometric mean seqs 1433 +/- 298

DNAMLε results *

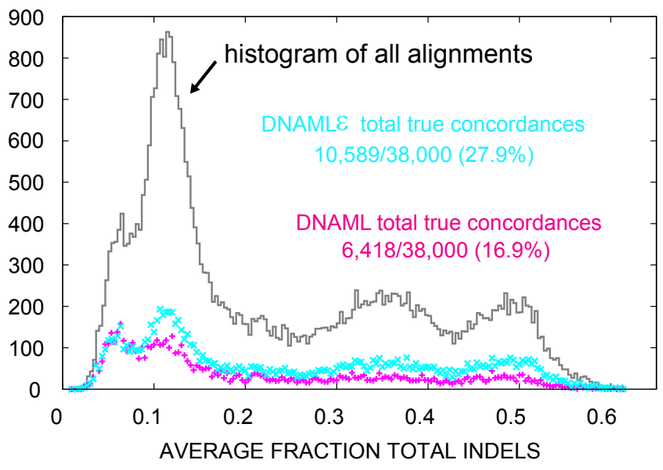
DNAML results +

LSU rRNA

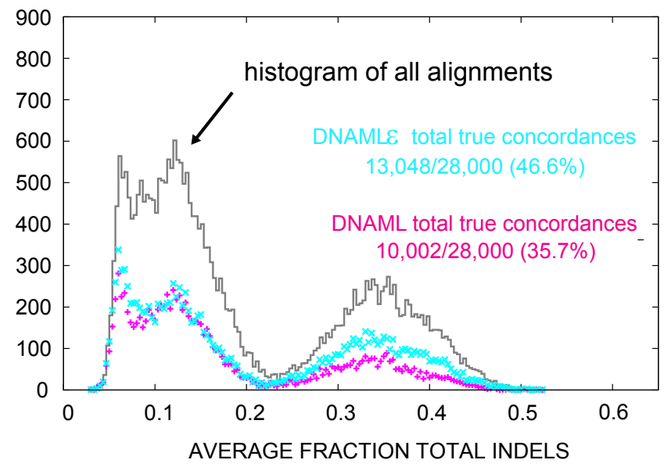
28,000 - EIGHT TAXON ALIGNMENTS

geometric mean seqs 3220 +/- 344

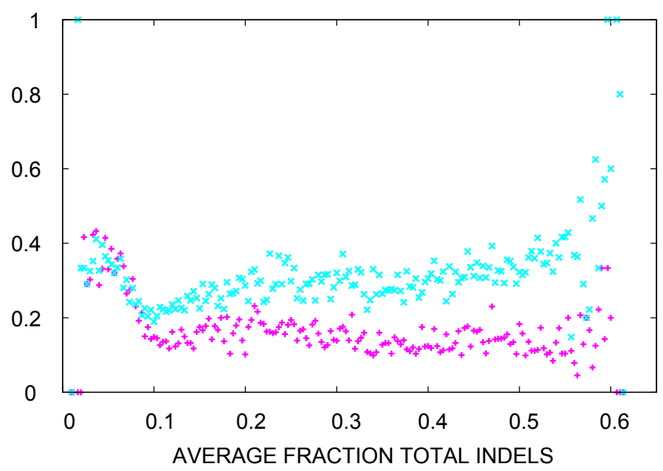
A



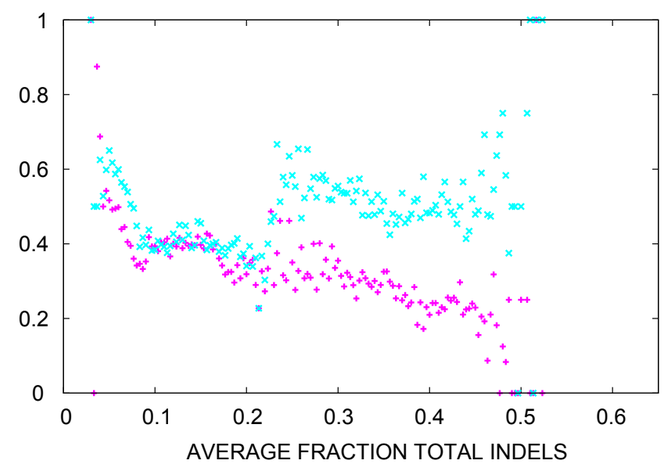
NUMBER TRUE CONCORDANCES



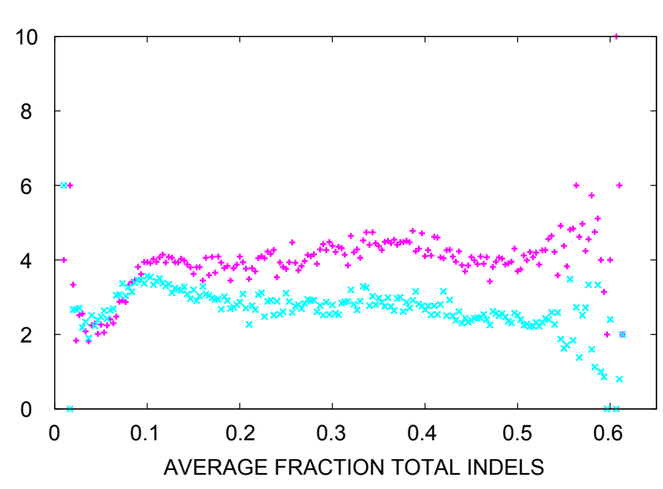
B



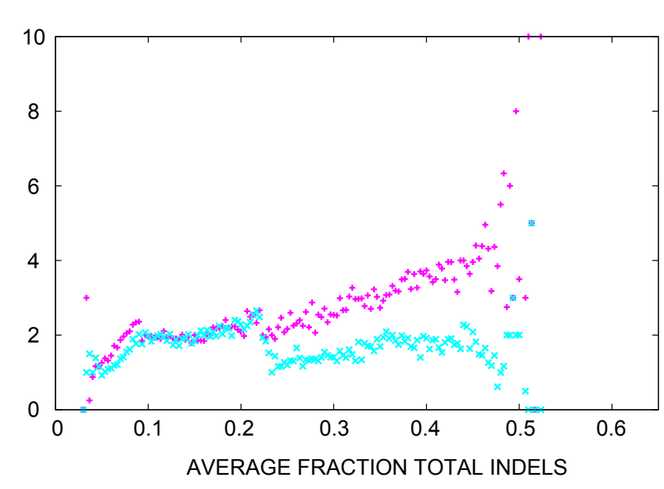
FRACTION TRUE CONCORDANCES



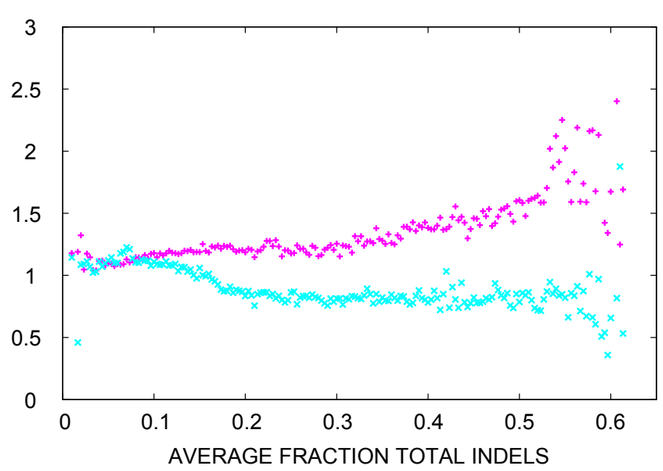
C



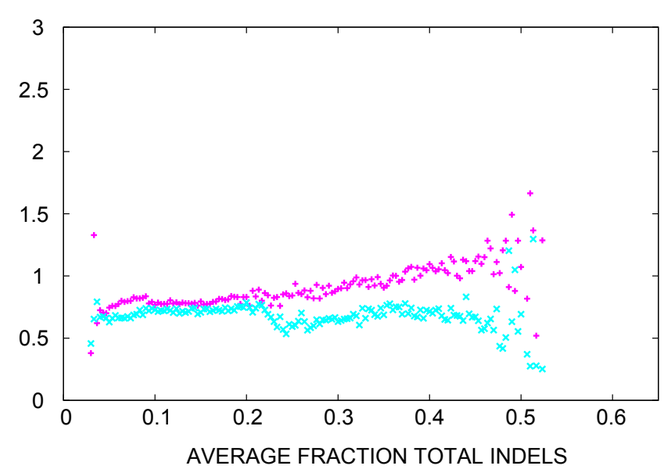
avg SDD DISTANCE



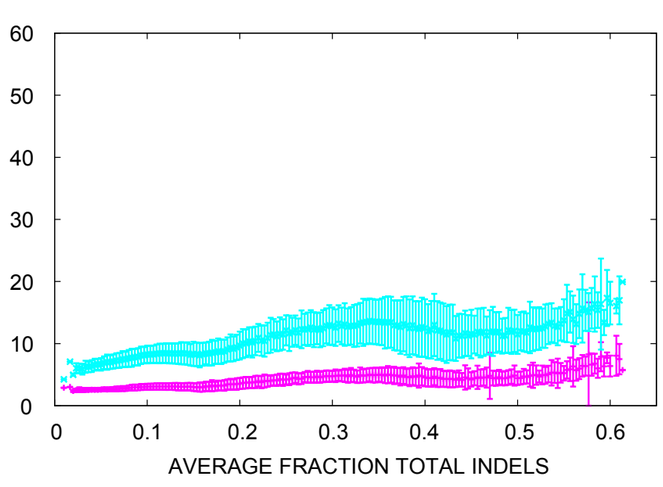
D



avg nBSD DISTANCE

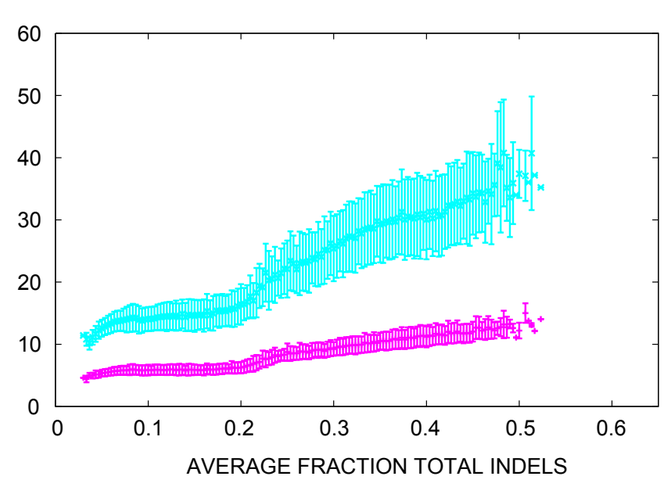


E



RUN TIME (sec)

(avg +/- std)



Comparison of different gap distributions

8-Taxon alignments

gap length distribution	ave subs per site and branch	gap parameter	pairwise % ID	pairwise % SUBS	pairwise % GAPS	$\Delta^{\text{AUC}}_{\%}$(TP)	$\Delta^{\text{AUC}}_{\%}$(SDD)	$\Delta^{\text{AUC}}_{\%}$(nBSD)
εRATE	0.005	0.30	96 ± 2	2 ± 1	2 ± 1	**28.9**	**−27.5**	**−24.9**
POISSON $\lambda = 0.5$	0.005	0.0020	96 ± 2	2 ± 1	2 ± 1	**24.6**	**−19.8**	**−16.8**
CODING $c/t = 100$	0.005	0.0020	96 ± 2	2 ± 1	2 ± 2	**4.8**	**−2.5**	11.8
CODING $c/t = 6$	0.005	0.0010	96 ± 4	2 ± 1	2 ± 4	−4.5	5.0	36.8
CODING $c/t = 1$	0.005	0.0005	96 ± 4	2 ± 1	2 ± 4	−8.1	7.0	41.3
εRATE	0.020	0.30	85 ± 4	7 ± 2	8 ± 3	**22.3**	**−40.6**	**−18.0**
POISSON $\lambda = 0.5$	0.020	0.0020	85 ± 5	7 ± 2	8 ± 3	**20.5**	**−32.9**	**−12.3**
CODING $c/t = 100$	0.020	0.0020	85 ± 5	7 ± 2	7 ± 4	**2.9**	**−0.5**	11.3
CODING $c/t = 6$	0.020	0.0010	84 ± 8	7 ± 2	9 ± 7	−11.6	19.9	43.3
CODING $c/t = 1$	0.020	0.0005	85 ± 8	7 ± 2	7 ± 8	−15.1	20.9	51.6
εRATE	0.040	0.30	73 ± 7	12 ± 3	15 ± 4	**23.4**	**−46.5**	**−13.9**
POISSON $\lambda = 0.5$	0.040	0.0020	73 ± 7	12 ± 3	15 ± 4	**20.7**	**−37.2**	**−10.6**
CODING $c/t = 100$	0.040	0.0020	73 ± 8	13 ± 4	14 ± 5	**0.7**	0.2	8.3
CODING $c/t = 6$	0.040	0.0010	72 ± 10	12 ± 4	16 ± 10	−21.9	31.3	40.4
CODING $c/t = 1$	0.040	0.0005	73 ± 11	13 ± 4	14 ± 11	−26.9	36.4	50.1
εRATE	0.070	0.30	58 ± 9	18 ± 4	24 ± 6	**24.1**	**−47.3**	**−13.2**
POISSON $\lambda = 0.5$	0.070	0.0020	58 ± 10	18 ± 4	24 ± 7	**23.7**	**−41.4**	**−10.8**
CODING $c/t = 100$	0.070	0.0020	60 ± 10	18 ± 4	22 ± 7	**2.4**	**−2.4**	5.5
CODING $c/t = 6$	0.070	0.0010	57 ± 12	17 ± 5	26 ± 12	−28.6	34.5	38.0
CODING $c/t = 1$	0.070	0.0005	59 ± 13	18 ± 5	23 ± 13	−39.4	50.1	48.9

$$\Delta^{\text{AUC}}_{\%}(f) = \frac{\text{AUC}\langle f(\text{DNAML}\varepsilon)\rangle - \text{AUC}\langle f(\text{DNAML})\rangle}{\max\{\,\text{AUC}\langle f(\text{DNAML}\varepsilon)\rangle\,,\,\text{AUC}\langle f(\text{DNAML})\rangle\,\}} \times 100.$$

For a given method M, the area under the curve (AUC): $\text{AUC}\langle f(\text{M})\rangle = \sum_{\text{L}=50}^{\text{L}=1000} \langle f(\text{L}\,|\,\text{M})\rangle \times \Delta\text{L}$,

where:

$\langle f(\text{L}\,|\,M)\rangle = \frac{1}{N}\sum_{n=1}^{N} f(A^n_{\text{L}}\,|\,\text{M})$, for alignments $\{A^n_{\text{L}}\}_{n=1}^{100}$

with geometric mean of sequence length L, and $\Delta\text{L} = 5$ nts, for these experiments.

nomenclature:

AUC = area under the curve

TP = fraction of true positive trees. Higher is better

SDD = Symmetric Difference Distance. Lower is better

nBSD = normalized Branch Score Distance. Lower is better

ROSE - Poisson $\lambda = 0.5$ gap distribution

8-Taxon alignments

ave subs per site and branch	ROSE gap parameter	pairwise % ID	pairwise % SUBS	pairwise % GAPS	$\Delta^{\text{AUC}}_{\%}$(TP)	$\Delta^{\text{AUC}}_{\%}$(SDD)	$\Delta^{\text{AUC}}_{\%}$(nBSD)	mean MBL DNAML	mean MBL DNAMLε	time (L=1000) (secs) DNAML	time (L=1000) (secs) DNAMLε
0.005	0.0020	96 ± 2	2 ± 1	2 ± 1	**23.4**	**−18.5**	**−17.6**	0.005	0.007	2.9 ± 1.7	4.7 ± 0.8
		96 ± 2	2 ± 1	2 ± 1	**3.8**	**−5.0**	**−16.3**	0.006	0.008	2.5 ± 1.1	4.6 ± 0.6
0.010	0.0010	94 ± 2	4 ± 1	2 ± 1	**14.8**	**−16.1**	**−7.9**	0.010	0.012	2.4 ± 0.8	5.3 ± 0.7
		94 ± 2	4 ± 2	2 ± 1	−4.8	2.3	**−6.7**	0.010	0.013	2.4 ± 0.9	5.2 ± 0.8
0.010	0.0020	92 ± 3	4 ± 1	4 ± 2	**22.7**	**−26.1**	**−15.7**	0.010	0.014	2.4 ± 0.6	5.9 ± 0.8
		91 ± 3	5 ± 2	4 ± 2	−4.9	4.3	**−14.2**	0.012	0.018	2.4 ± 0.4	5.8 ± 0.8
0.020	0.0010	88 ± 4	8 ± 2	4 ± 2	**12.9**	**−19.1**	**−6.7**	0.020	0.024	2.4 ± 0.2	6.0 ± 0.7
		87 ± 4	9 ± 3	4 ± 2	−8.7	12.3	**−5.3**	0.023	0.027	2.4 ± 0.2	5.9 ± 0.8
0.020	0.0020	85 ± 5	7 ± 2	8 ± 3	**20.5**	**−32.9**	**−12.3**	0.020	0.028	2.8 ± 0.6	7.0 ± 1.0
		83 ± 6	10 ± 4	7 ± 2	−5.2	7.1	**−11.3**	0.030	0.039	2.9 ± 0.6	7.1 ± 0.8
0.030	0.0005	86 ± 4	11 ± 3	3 ± 1	**7.4**	**−14.7**	**−2.7**	0.030	0.032	2.5 ± 0.7	6.0 ± 0.7
		85 ± 4	12 ± 3	3 ± 1	−12.8	15.9	**−1.4**	0.032	0.035	2.7 ± 0.6	6.1 ± 0.7
0.030	0.0010	83 ± 5	11 ± 3	6 ± 2	**14.3**	**−22.4**	**−6.4**	0.030	0.035	2.7 ± 0.5	7.0 ± 1.0
		81 ± 5	13 ± 4	5 ± 2	−10.5	13.3	**−4.8**	0.036	0.042	2.8 ± 0.5	6.9 ± 0.9
0.030	0.0020	78 ± 6	10 ± 3	11 ± 4	**18.4**	**−35.6**	**−12.1**	0.030	0.042	2.9 ± 0.6	8.2 ± 1.2
		75 ± 8	16 ± 5	9 ± 3	−5.5	5.8	**−9.2**	0.050	0.062	3.1 ± 0.4	8.6 ± 1.0
0.040	0.0005	82 ± 5	14 ± 4	4 ± 2	**9.6**	**−17.2**	**−2.7**	0.040	0.042	2.7 ± 0.4	6.5 ± 0.7
		81 ± 5	15 ± 4	4 ± 2	−16.4	19.3	**−0.7**	0.043	0.047	2.7 ± 0.3	6.7 ± 0.6
0.040	0.0010	78 ± 6	14 ± 4	8 ± 3	**15.3**	**−26.1**	**−6.2**	0.040	0.045	3.0 ± 0.5	8.0 ± 0.9
		76 ± 7	17 ± 5	7 ± 2	−10.1	12.3	**−3.7**	0.051	0.058	3.2 ± 0.8	7.9 ± 0.7
0.040	0.0015	75 ± 7	13 ± 4	12 ± 4	**16.9**	**−33.1**	**−8.4**	0.040	0.048	3.1 ± 0.4	8.9 ± 1.2
		72 ± 8	19 ± 6	9 ± 3	−7.1	7.8	**−6.2**	0.061	0.071	3.4 ± 0.6	9.1 ± 1.2
0.040	0.0020	73 ± 7	12 ± 3	15 ± 4	**20.7**	**−37.2**	**−10.6**	0.040	0.052	3.4 ± 0.5	9.4 ± 1.3
		68 ± 9	21 ± 7	11 ± 3	−4.7	4.6	**−8.4**	0.072	0.085	3.7 ± 0.4	10.8 ± 1.3
0.070	0.0005	71 ± 7	22 ± 5	7 ± 3	**12.6**	**−21.3**	**−3.0**	0.070	0.071	3.3 ± 0.4	9.6 ± 1.0
		70 ± 8	24 ± 6	6 ± 2	−11.6	15.0	**−0.3**	0.080	0.085	3.4 ± 0.5	9.3 ± 0.9
0.070	0.0010	66 ± 8	20 ± 5	13 ± 4	**16.0**	**−31.2**	**−6.3**	0.070	0.074	3.7 ± 0.4	10.8 ± 1.1
		62 ± 10	28 ± 8	10 ± 3	−9.2	9.3	**−3.2**	0.101	0.109	3.9 ± 0.3	11.6 ± 1.2
0.070	0.0015	62 ± 9	19 ± 5	19 ± 5	**20.2**	**−39.6**	**−8.2**	0.070	0.078	4.0 ± 0.6	12.5 ± 1.7
		57 ± 11	31 ± 8	12 ± 3	−7.1	5.0	**−9.1**	0.126	0.135	4.7 ± 0.6	14.6 ± 1.8
0.070	0.0020	58 ± 10	18 ± 4	24 ± 7	**23.7**	**−41.5**	**−10.8**	0.070	0.083	4.3 ± 0.7	13.5 ± 1.7
		52 ± 12	35 ± 9	13 ± 4	−4.4	3.1	**−11.9**	0.154	0.162	5.1 ± 0.7	16.8 ± 2.4
0.100	0.0005	63 ± 9	28 ± 6	10 ± 3	**17.1**	**−29.7**	**−3.6**	0.100	0.100	4.0 ± 0.7	12.1 ± 1.3
		60 ± 10	32 ± 8	8 ± 3	−11.7	12.7	**−0.4**	0.121	0.125	4.1 ± 0.5	12.0 ± 1.5
0.100	0.0010	57 ± 9	25 ± 6	18 ± 6	**20.6**	**−40.3**	**−6.3**	0.100	0.100	4.5 ± 0.7	14.5 ± 1.3
		52 ± 12	36 ± 9	11 ± 3	−7.9	6.6	**−6.3**	0.160	0.165	4.7 ± 0.5	15.4 ± 2.3
0.100	0.0015	52 ± 10	23 ± 5	25 ± 6	**25.3**	**−44.1**	**−8.8**	0.100	0.105	4.9 ± 0.6	16.6 ± 2.2
		47 ± 12	40 ± 10	13 ± 4	−4.0	2.0	**−12.2**	0.201	0.202	5.7 ± 1.0	19.0 ± 3.3
0.100	0.0020	48 ± 11	21 ± 4	31 ± 7	**27.2**	**−50.7**	**−11.4**	0.100	0.110	5.4 ± 0.7	18.3 ± 2.2
		43 ± 12	43 ± 10	14 ± 4	**0.3**	**−0.5**	**−16.9**	0.242	0.234	6.0 ± 1.0	20.7 ± 3.1

$$\Delta^{\text{AUC}}_{\%}(f) = \frac{\text{AUC}\langle f(\text{DNAML}\varepsilon)\rangle - \text{AUC}\langle f(\text{DNAML})\rangle}{\max\{\,\text{AUC}\langle f(\text{DNAML}\varepsilon)\rangle\,,\,\text{AUC}\langle f(\text{DNAML})\rangle\,\}} \times 100.$$

For a given method M, the area under the curve (AUC): $\text{AUC}\langle f(\text{M})\rangle = \sum_{\text{L}=50}^{\text{L}=1000} \langle f(\text{L}\,|\,\text{M})\rangle \times \Delta\text{L}$,

where:

$\langle f(\text{L}\,|\,M)\rangle = \frac{1}{N}\sum_{n=1}^{N} f(A^n_{\text{L}}\,|\,\text{M})$, for alignments $\{A^n_{\text{L}}\}_{n=1}^{100}$

with geometric mean of sequence length L, and $\Delta\text{L} = 5$ nts, for these experiments.

nomenclature:

TP = fraction of true positive trees.

SDD = Symmetric Difference Distance.

nBSD = normalized Branch Score Distance.

MBL = mean branch length.