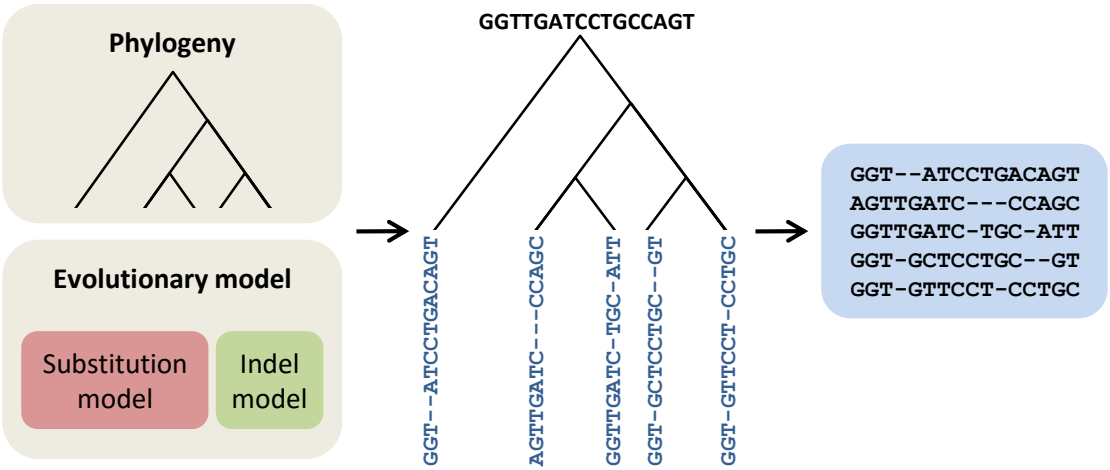
Towards Realistic *Sequence Simulations* for the Reconstruction of Mega Phylogenies

Thesis submitted towards the M.Sc. degree in Bioinformatics
in Tel-Aviv University
by

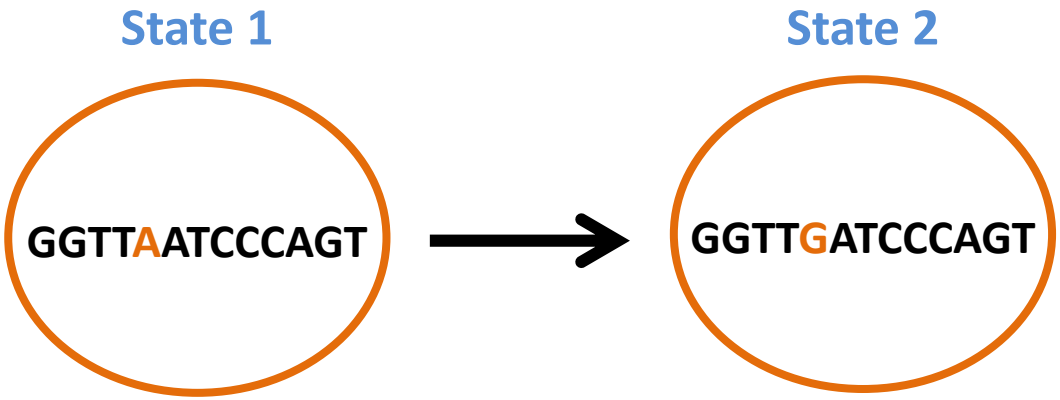
Nomi Hadar

Schema of typical sequence simulator

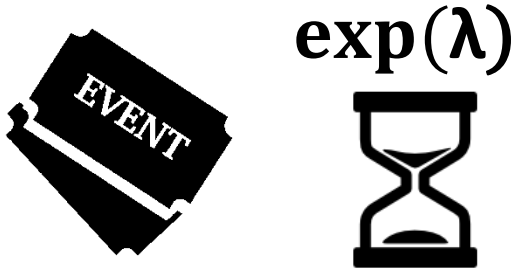


Markov chain

State of the chain = whole sequence



Simulating Markov chains using Gillespie's algorithm

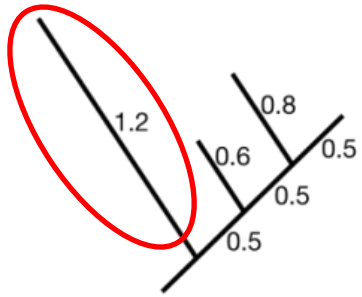


Mol. Biol. Evol. 2009

INDELible simulator – a closer look

Given: GGTTGATCCTGCCAGTAGTCATCT

sequence



branch length (t)

Denote: $\lambda = I + D + S$

Insertion rate
Deletion rate
Substitution rate

Waiting time: $s1$

$s \sim \exp(\lambda)$, mean=$1/\lambda$

INDELible simulator – a closer look

$s1 > t$

no event

$s1 < t$

draw an event with probabilities

$\frac{I}{\lambda'}$ $\frac{D}{\lambda'}$ $\frac{S}{\lambda}$

draw a location

indel

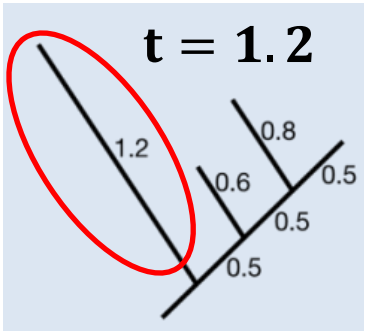
substitution

uniformly (L+1 positions)

proportional to the substitution rate at each site

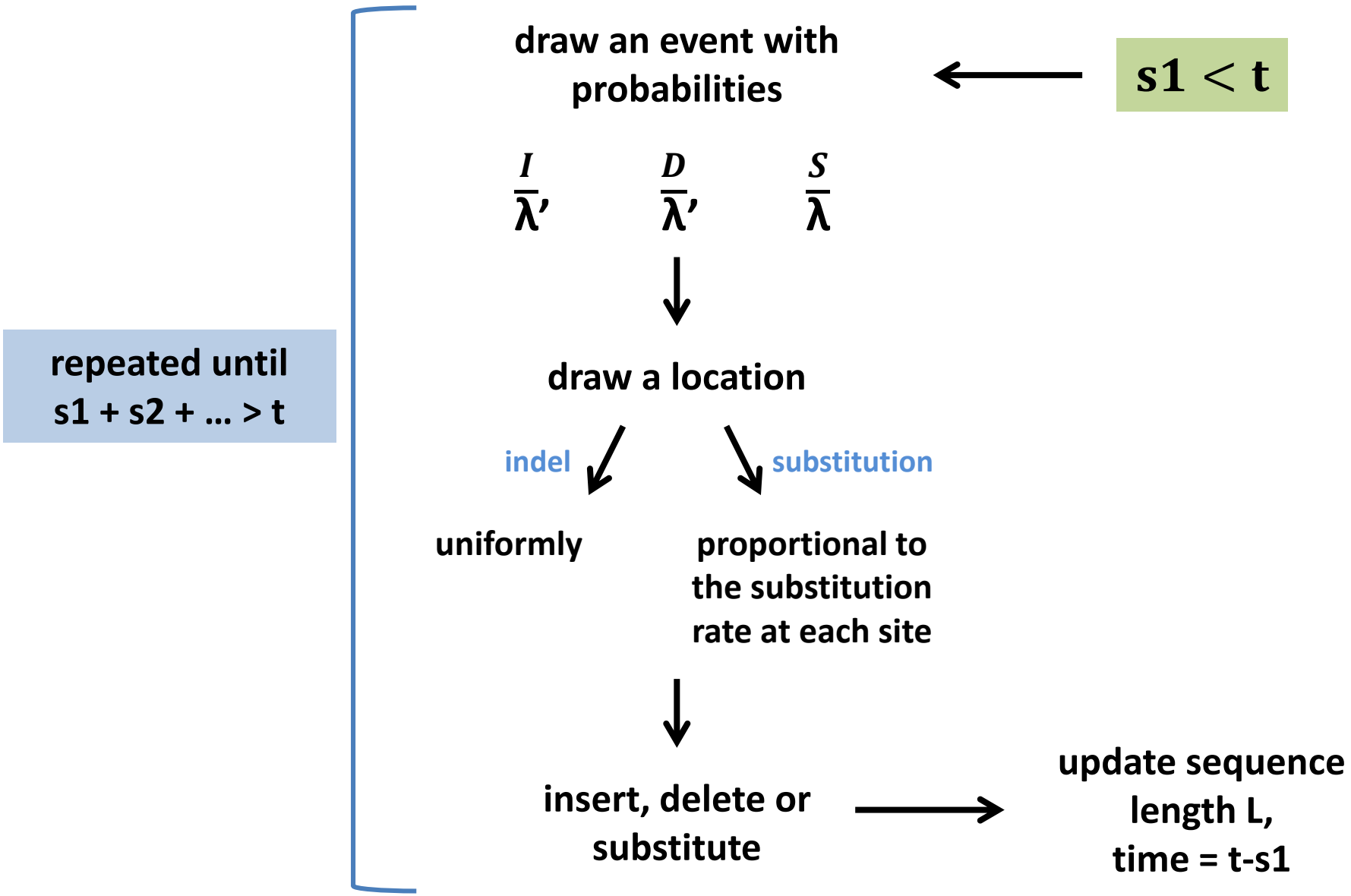
insert, delete or substitute

update sequence length L, time = t-s1



$\lambda = I + D + S$

INDELible simulator – a closer look



Estimating the parameters for sequence simulations

'IR'

Indel Rate

$IR \in [0,1]$

'RL'

Root Length

'a'

Shape parameter of power-law distribution, which controls the indel lengths

How to estimate?

Estimating the parameters for sequence simulations: existing tools

Values of indel parameters inferred by three methods

		Lambda			SPARTA			SpartaABC		
	Marker	IR	a	RL	IR	a	RL	IR	a	RL
Large data set	*18S*	0.002	1.353	1,654	-	-	-	0.001	1.838	226
	26S	0.004	1.336	1,997	-	-	-	0.001	1.739	571
	ITS	0.051	1.354	617	-	-	-	0.002	1.387	275
	atpB	0.001	1.222	1,412	-	-	-	0	1.446	779
	matK	0.005	1.288	1,540	-	-	-	0	1.491	1,117
	rbcL	0.002	1.214	1,295	-	-	-	0	1.42	918
	trnL-F	0.028	1.341	820	-	-	-	0.002	1.104	440
Small data set	*18S*	0.005	1.466	1,689	0.061	1.566	1,805	0.001	1.067	937
	26S	0.011	1.428	2,350	0.137	1.265	3,365	0.001	1.056	1,346
	ITS	0.048	1.298	633	0.083	1.464	1,155	0.003	1.043	492
	atpB	0.003	1.26	1,440	0.146	1.55	1,456	0.001	1.068	691
	matK	0.007	1.335	1,529	0.145	1.539	1,791	0.001	1.064	1,369
	rbcL	0.002	1.287	1,381	0.094	1.55	1,427	0.001	1.069	869
	trnL-F	0.049	1.377	836	0.019	1.508	633	0.005	1.065	631

Estimating the parameters for sequence simulations: existing tools

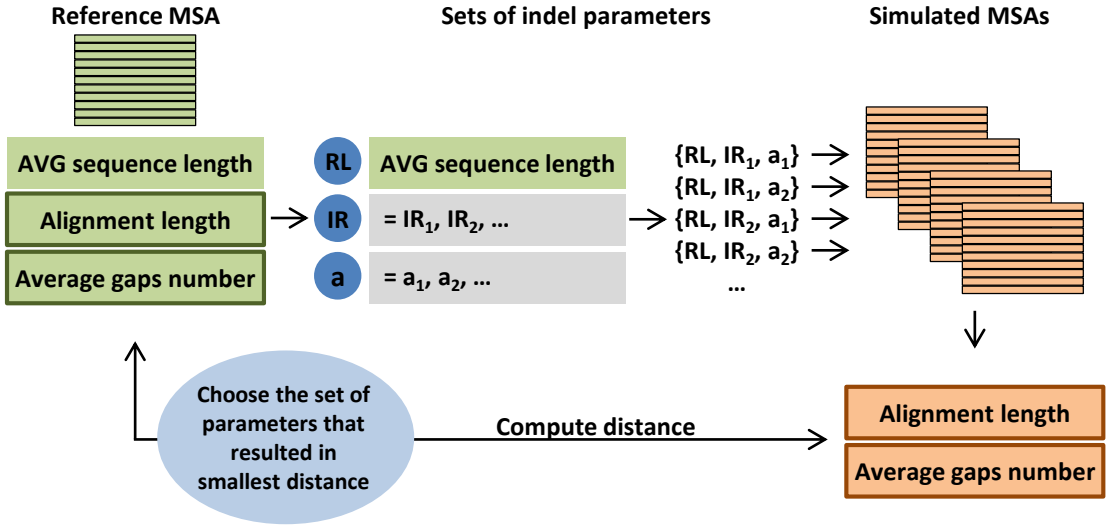
Length and average number of gaps of reference and simulated MSAs

		Reference MSA		Simulated MSAs					
				Lambda		Sparta		SpartaABC	
	Marker	Len	Gaps	Len	Gaps	Len	Gaps	Len	Gaps
Large data set	*18S*	1,836	26	4,795	399	-	-	344	33
	26S	3,715	60	7,595	645	-	-	848	68
	ITS	2,205	87	55,075	592	-	-	1,154	107
	atpB	1,521	4	3,506	227	-	-	779	0
	matK	2,164	29	10,623	806	-	-	1,116	0
	rbcL	1,433	5	5,713	432	-	-	917	0
	trnL-F	1,938	104	33,623	753	-	-	1,748	117
Small data set	*18S*	1,813	10	2,694	162	13,854	1,163	1,176	21
	26S	3,494	31	5,927	495	81,962	2,227	1,776	40
	ITS	1,041	26	5,463	341	13,761	806	846	16
	atpB	1,517	3	2,135	87	25,549	1,110	1,064	19
	matK	1,891	16	3,053	198	31,752	1,357	1,658	30
	rbcL	1,427	3	1,876	65	16,444	1,029	862	33
	trnL-F	1,326	55	6,807	476	2,083	204	1,352	64

Do not resemble the reference!

Estimating the parameters for sequence simulations: OPTIMIM

Scheme of OPTIMIM method to infer indel parameters



$$d(Ref, Sim) = \frac{\left| Len_{Ref} - Len_{Sim} \right|}{Len_{Ref}} + \frac{\left| GAP_{Ref} - GAP_{Sim} \right|}{GAP_{Ref}}$$

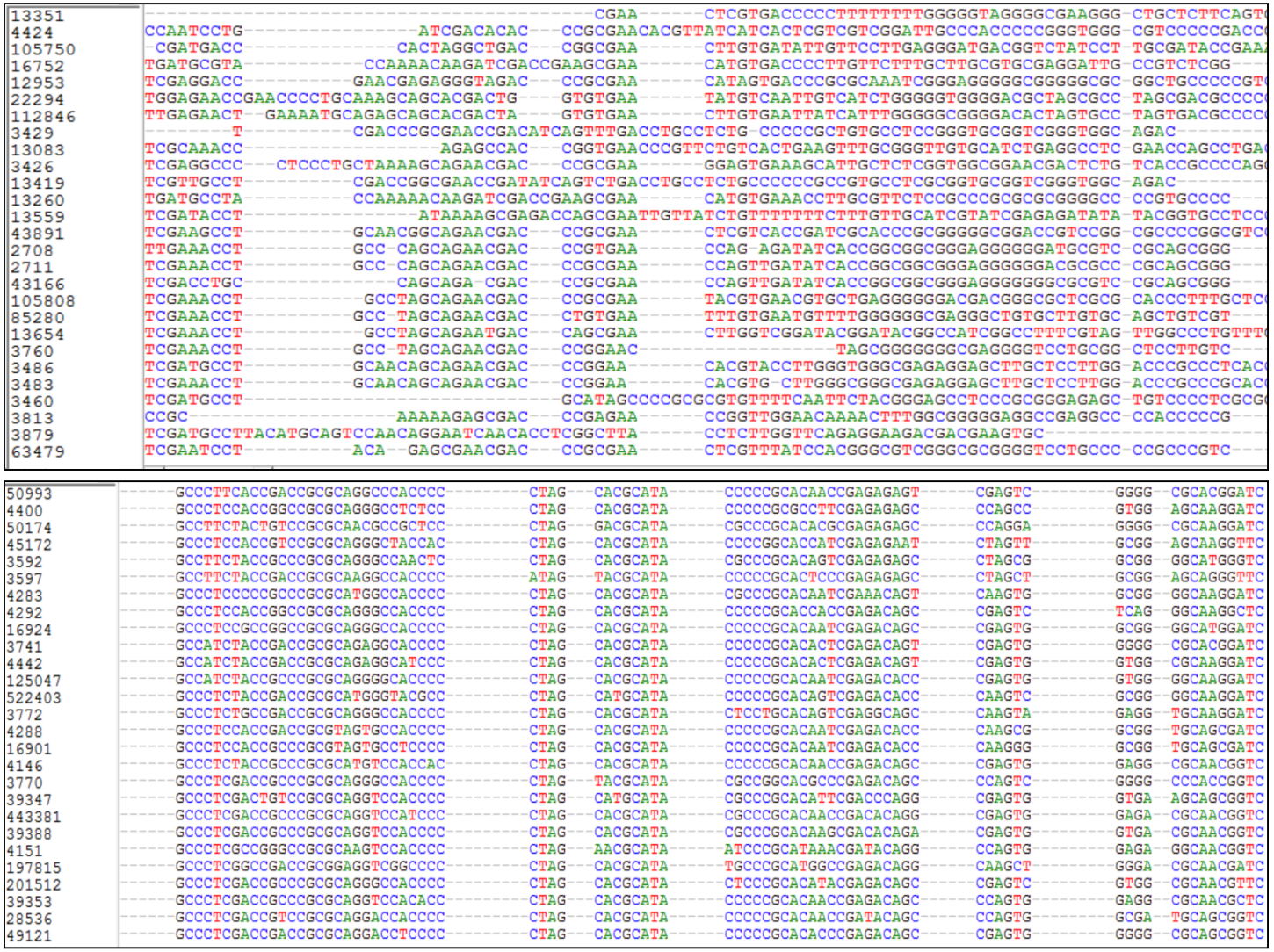
Estimating the parameters for sequence simulations: OPTIMIM

Inference of indel parameters using the OPTIMIM method, and length and average number of gaps of the simulated MSAs using these parameters

	Marker	Inferred parameters			Reference MSA		Simulated MSAs	
		IR	a	RL	Len	Gaps	Len	Gaps
Large data set	***18S***	0.00012	1.00000041	1,659	1,836	26	1,986	30
	26S	0.00030	1.00000012	2,002	3,715	60	2,744	65
	ITS	0.00079	1.00000259	618	2,205	87	1,850	102
	atpB	0.00001	1.00000001	1,415	1,521	4	1,450	3
	matK	0.00012	1.00000481	1,542	2,164	29	1,866	30
	rbcL	0.00057	1.00003067	821	1,433	5	1,345	4
	trnL-F	0.00012	1.00000041	1,659	1,938	104	1,877	89
Small data set	***18S***	0.0002	1.00000002	1,698	1,813	10	1,794	9
	26S	0.0006	1.00000022	2,358	3,494	31	2,668	29
	ITS	0.0020	1.00000022	634	1,041	26	951	28
	atpB	0.0001	1.00000075	1,448	1,517	3	1,471	2
	matK	0.0006	1.00067269	1,530	1,891	16	1,733	18
	rbcL	0.0001	1.00001654	1,383	1,427	3	1,410	3
	trnL-F	0.0028	1.05073443	837	1,326	55	1,409	50

Do resemble the reference!

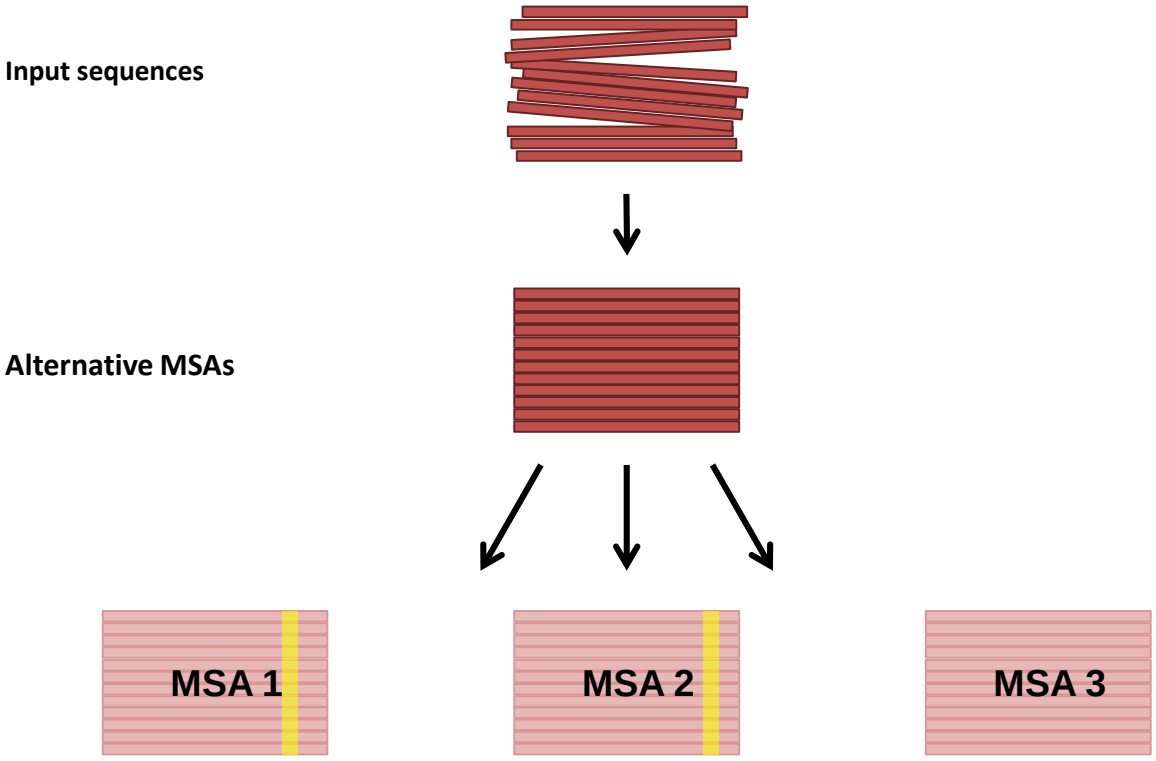
Which sequences are simulated?



ITS gene

Evaluating the complexity of the simulated alignments

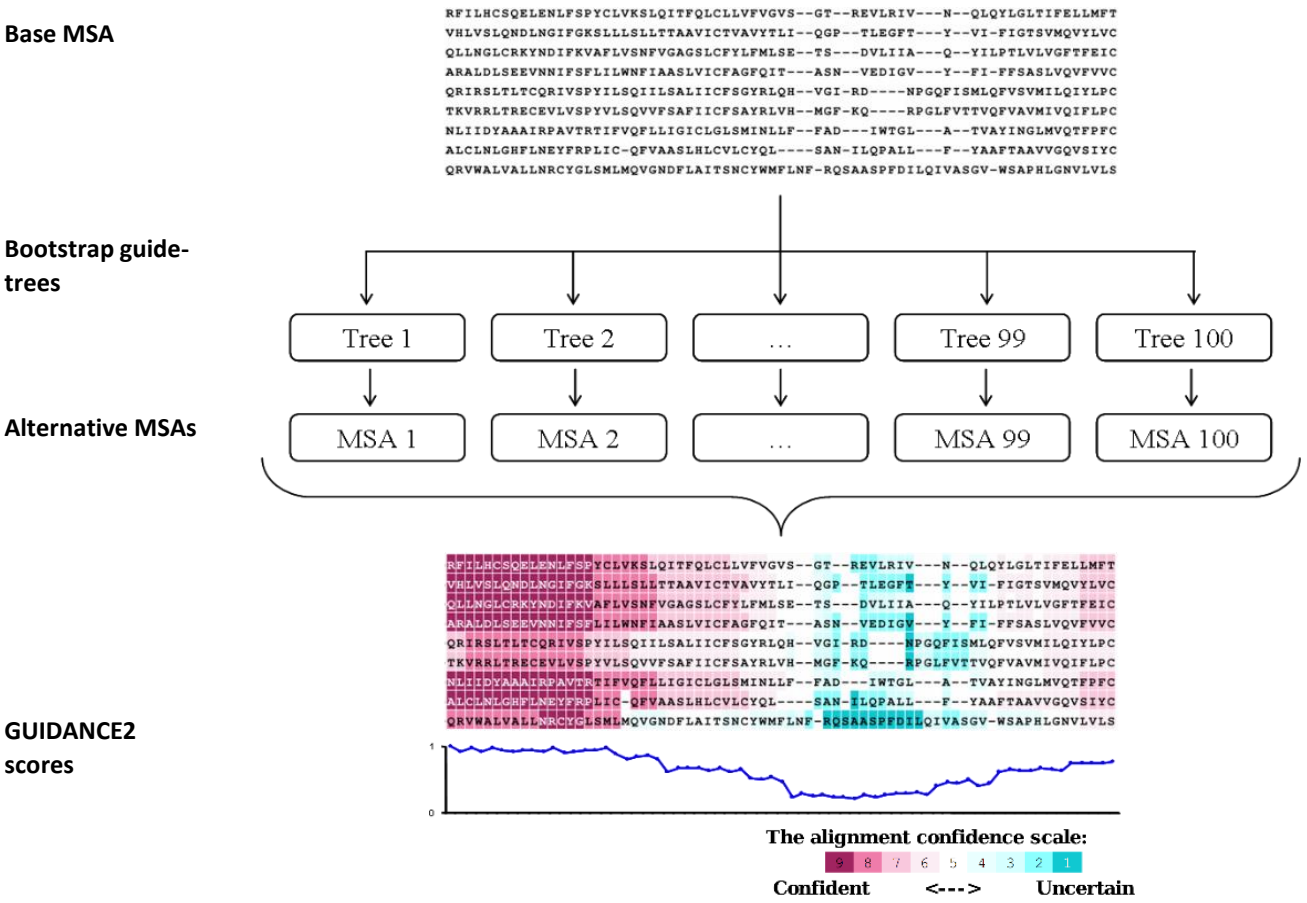
GUIDANCE2 score



Column score: 2/3

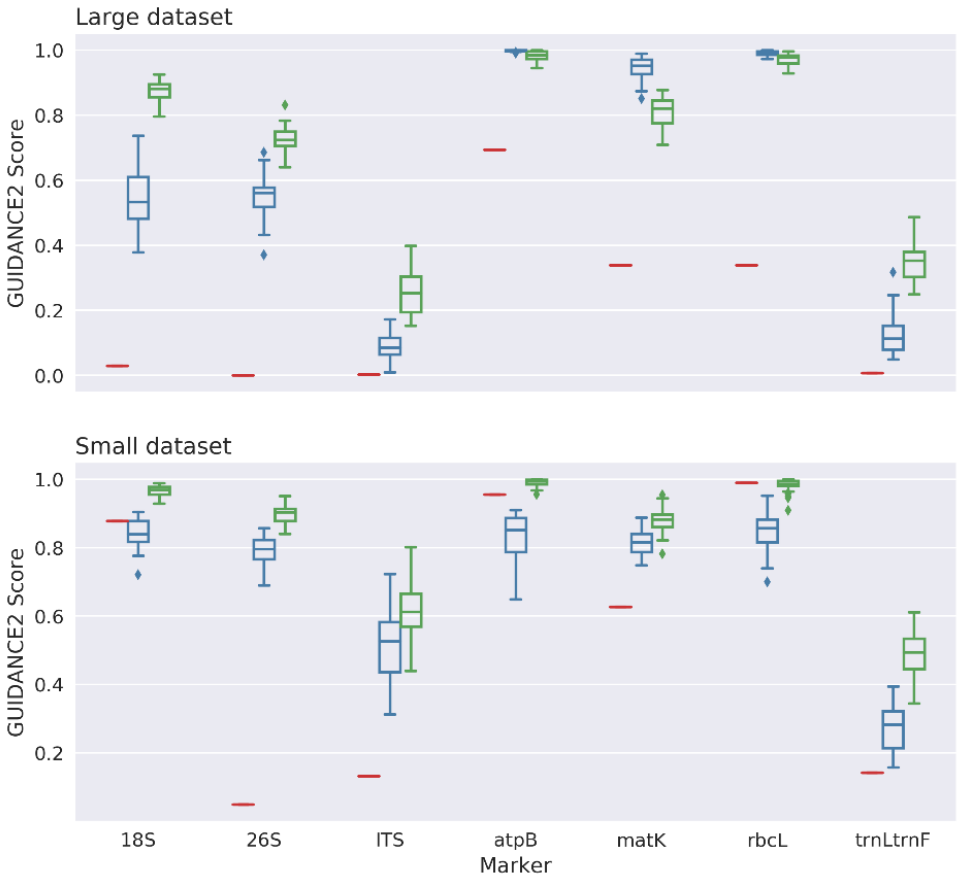
Evaluating the complexity of the simulated alignments

A schematic flowchart of the GUIDANCE2 algorithm



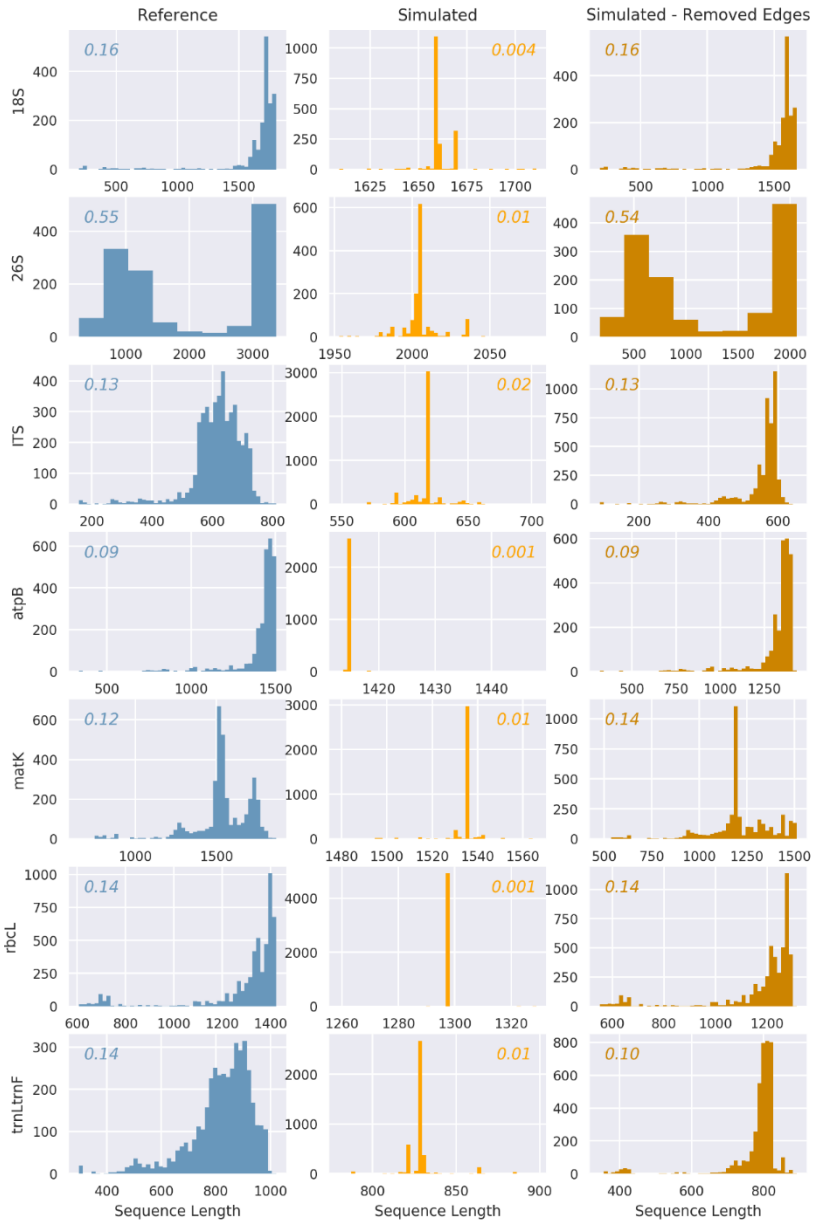
Evaluating the complexity of the simulated alignments

Alignment complexity of real and simulated data



	Marker	Reference MSAs	Simulated MSAs	
			SpartaABC	OPTIMIM
Big data set	***18S***	0.03	0.55	0.87
	26S	0.00003	0.55	0.72
	ITS	0.003	0.09	0.26
	atpB	0.69	1.00	0.98
	matK	0.34	0.95	0.81
	rbcL	0.34	0.99	0.97
	trnL-F	0.01	0.12	0.35
Small data set	***18S***	0.88	0.84	0.97
	26S	0.05	0.79	0.85
	ITS	0.13	0.52	0.67
	atpB	0.96	0.83	0.99
	matK	0.63	0.82	0.98
	rbcL	0.99	0.85	0.92
	trnL-F	0.14	0.27	0.46

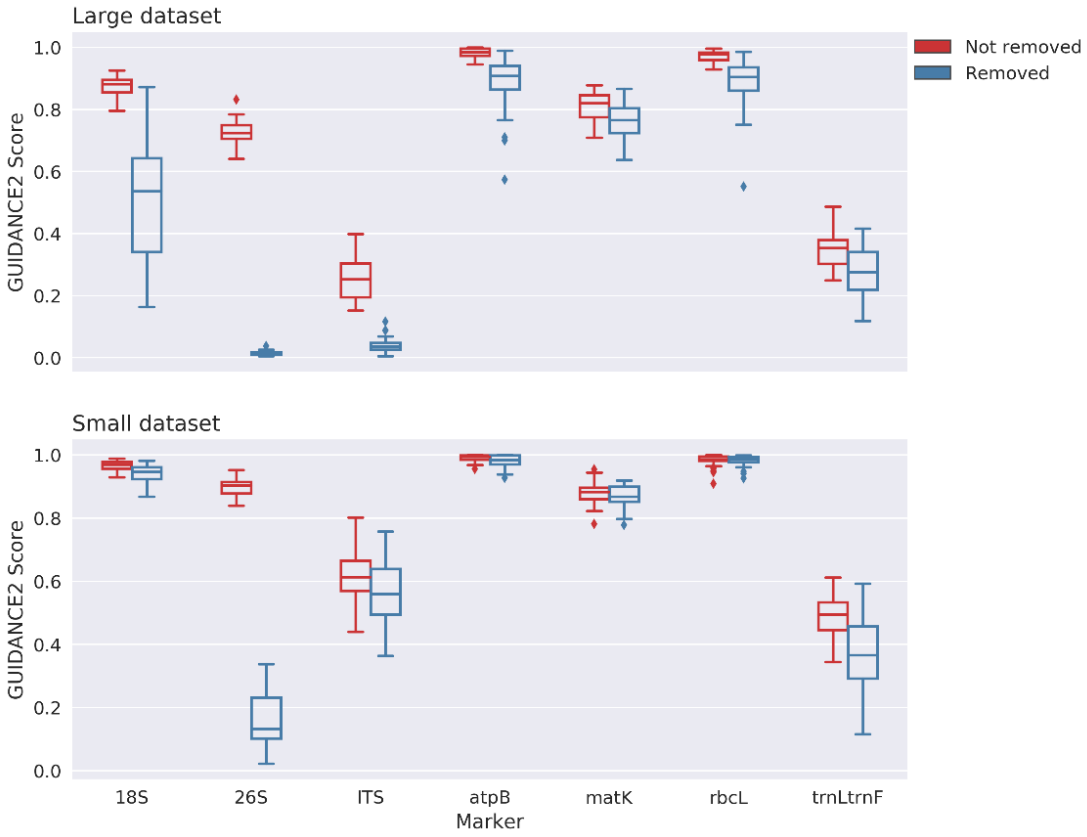
Increasing simulation complexity - examining simulated sequence lengths



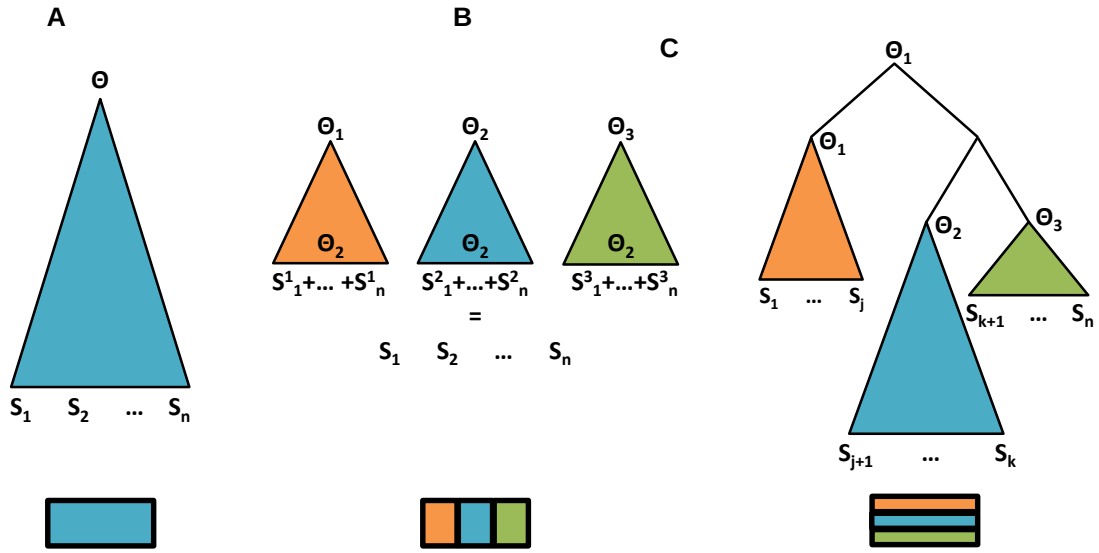
Histograms of sequence lengths for reference and simulated alignments of large dataset

Increasing simulation complexity - examining simulated sequence lengths

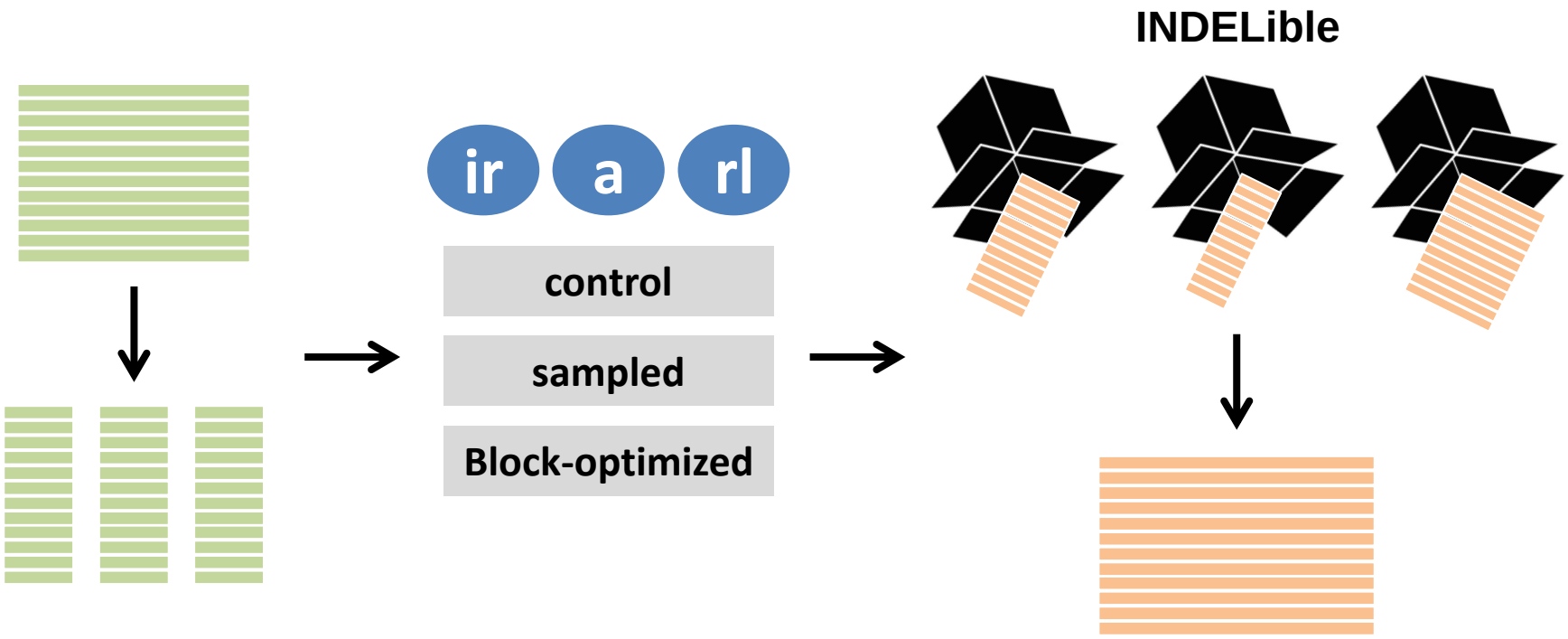
Alignment complexity of simulated data following the removal of edges



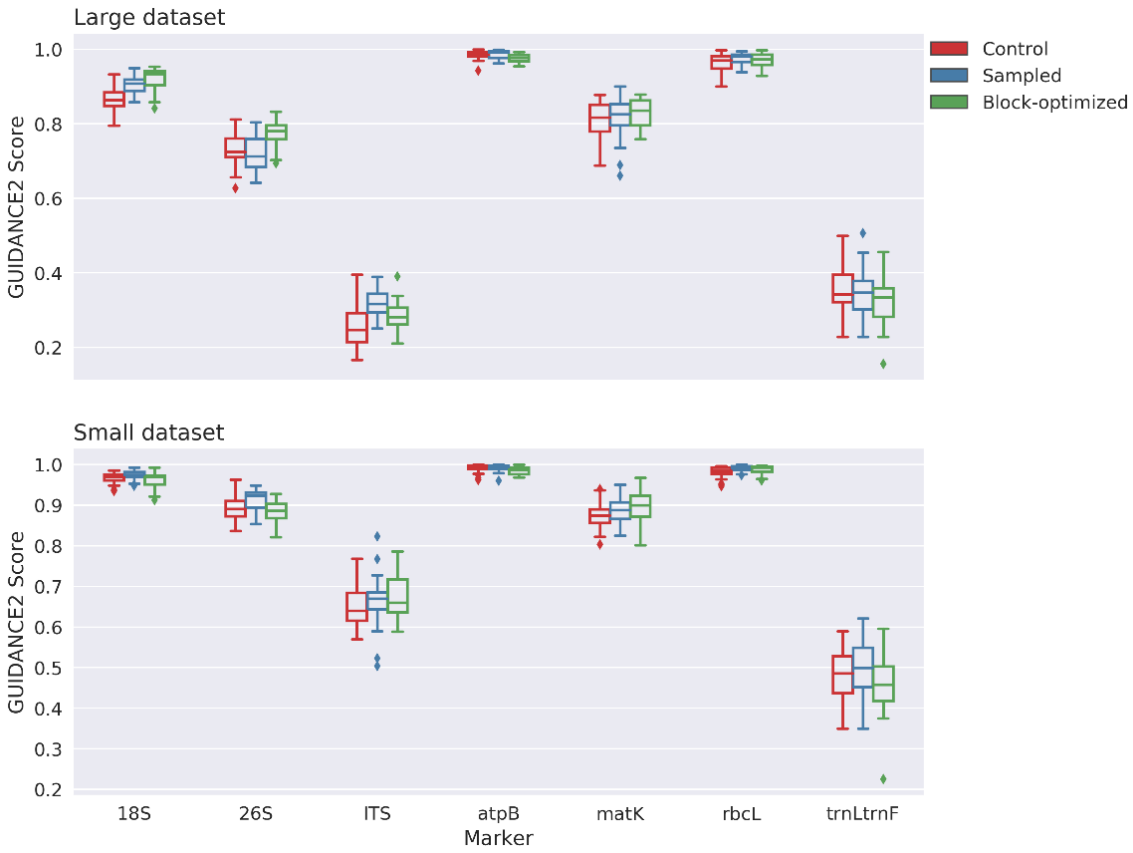
Increasing simulation complexity - non-homogeneous indel model



Increasing simulation complexity - non-homogeneous indel model

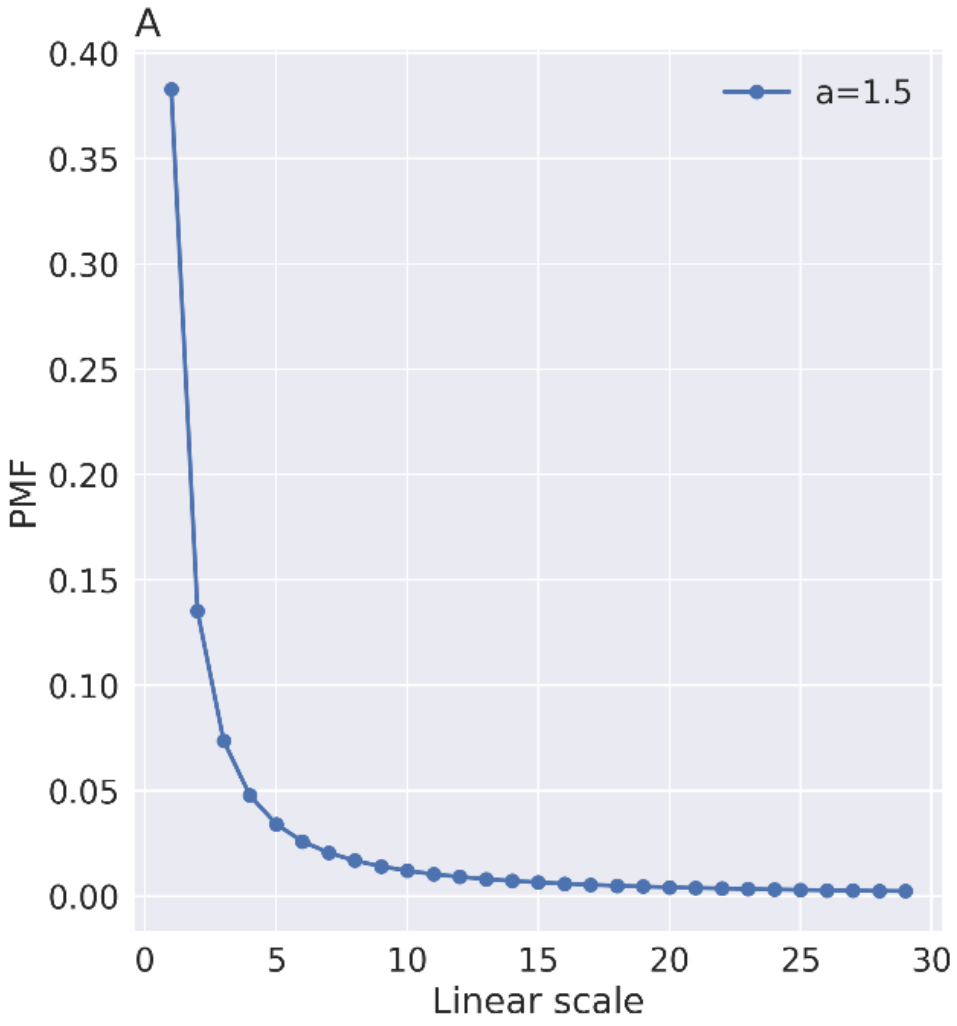


Increasing simulation complexity - non-homogeneous indel model

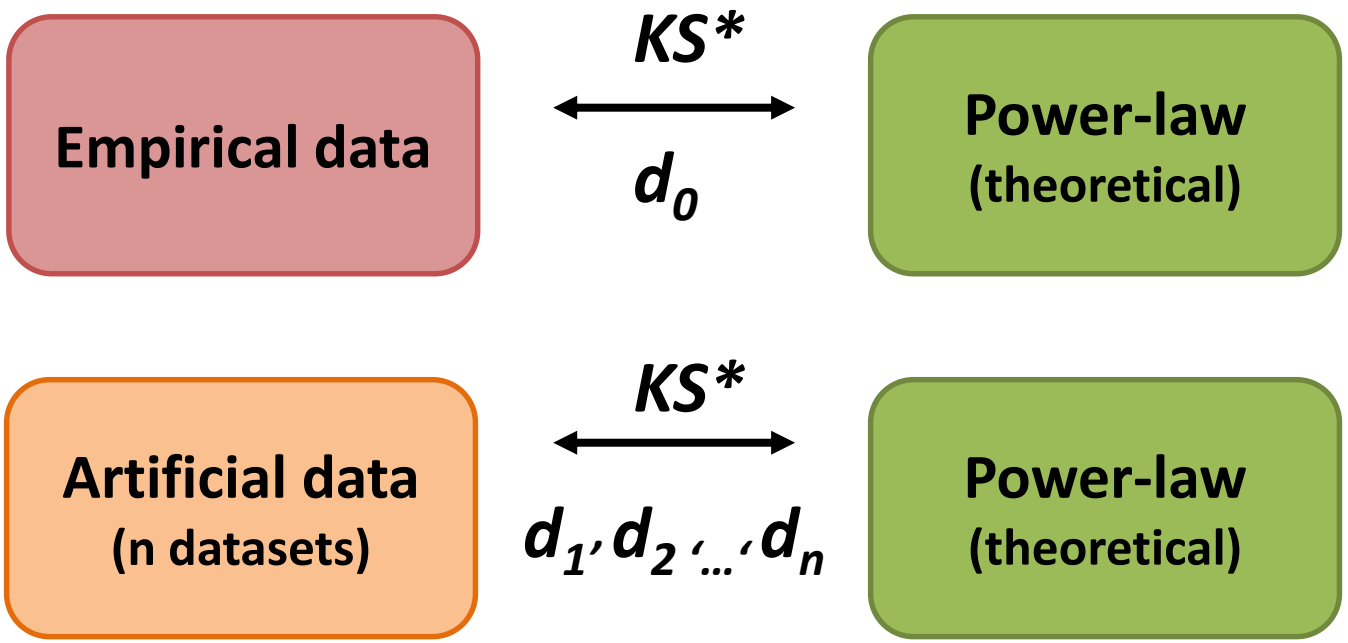


Increasing simulation complexity - indel-length distribution

Power-law



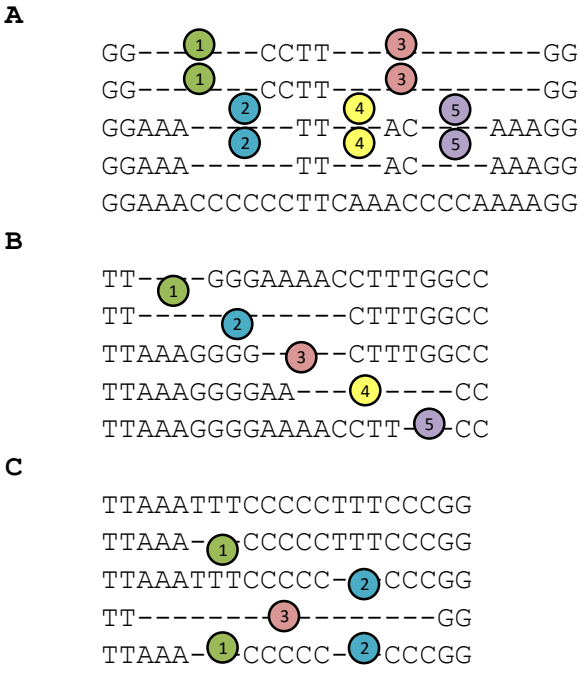
Increasing simulation complexity - indel-length distribution



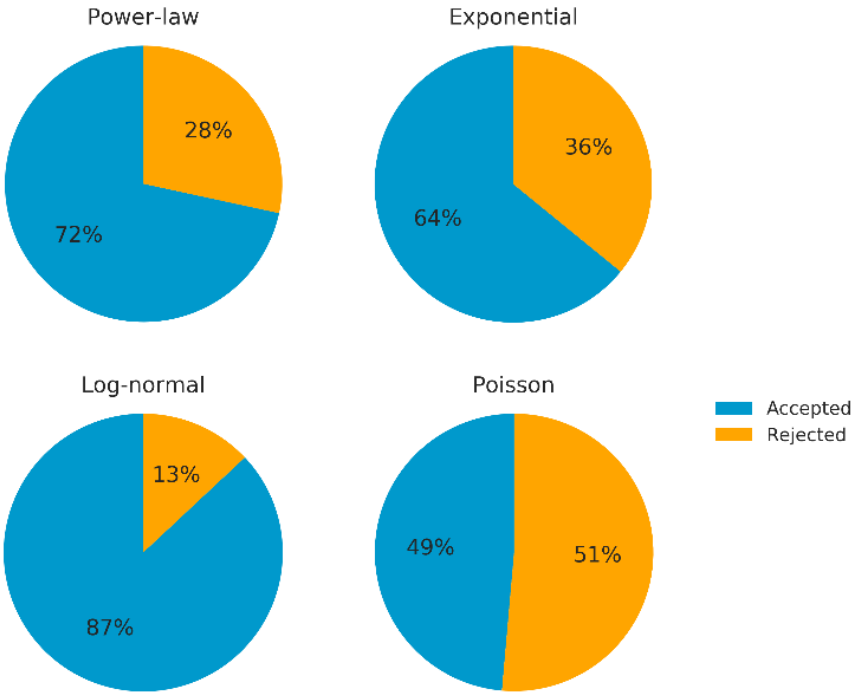
$$p - value = \frac{\{d_i > d_0\}_{1 \le i \le n}}{n}$$

*Kolmogorov–Smirnov test

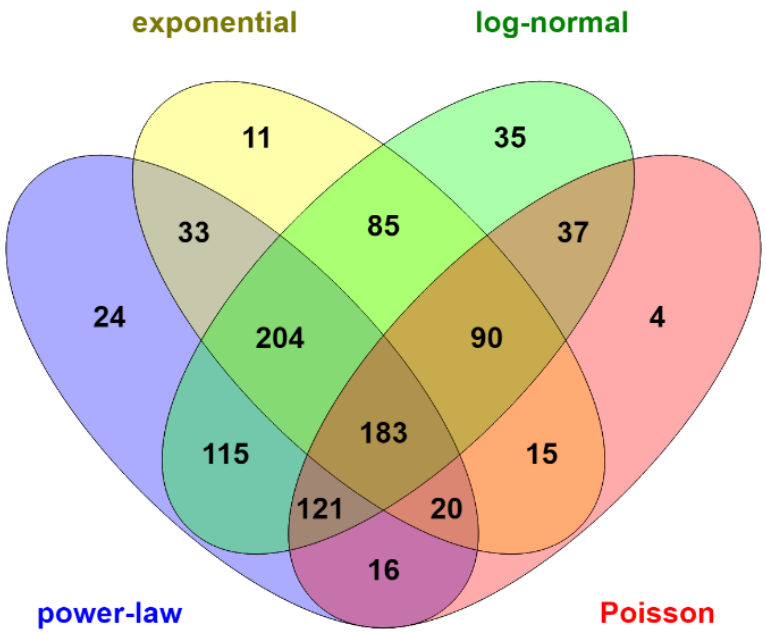
Increasing simulation complexity - indel-length distribution



Increasing simulation complexity - indel-length distribution



Pie chart of results of goodness-of-fit tests



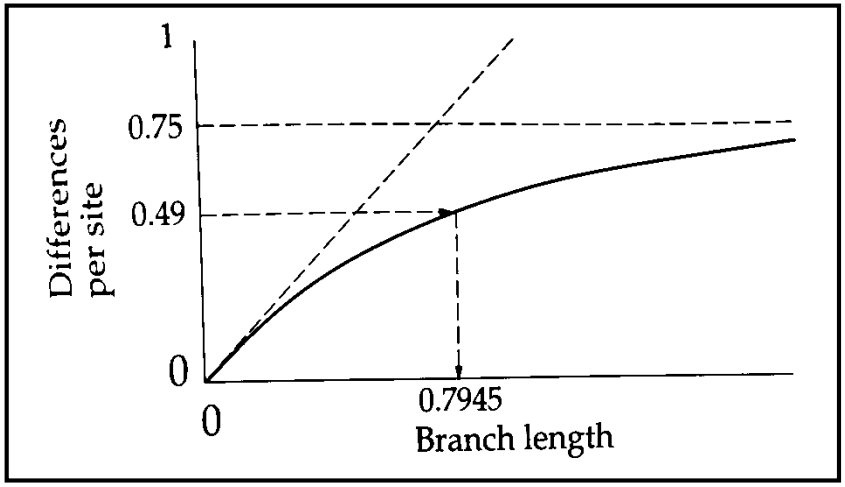
Venn diagram of datasets intersection

Appendices

PHLAWD: Detect saturation across a set of sequence data

d_1	d_2
raw pair-wise distance	Jukes-Cantor distance
ACGGTCATGTATAC	ACGGTCATGTATAC
TCTCTCCTGTGAAT	TCTCTCCTGTGAAT
$d_1 = 0.5$	$d_2 = 0.8$

- $x_i = |d1 - d2|$ *(x = 0.3)*



- **Measure of dispersion is based on the median absolute deviation (MAD):**

$$MAD = Med\,(|x_i - Med\,(x)|)$$

E.g.:

x=(9, 6, 4, **2**, 2, 1, 1), MED(x)=2

absolute deviations: (7-2, 4-2, 2-2, 0-2, 2-2, 1-2, 1-1)

->(7, 4, 2, 0, 0, 1, 1) ->(7, 4, 2, **1**, 1, 0 ,0)-> **MAD=1**

PHLAWD: Detect saturation across a set of sequence data

- **The larger the MAD -> the larger the overall spread of distances.**

- **That is, above a certain value the assumed nucleotide substitution model is no longer adequately accounting for the rate variation exhibited by pair-wise distances among species.**

Large MAD ---->

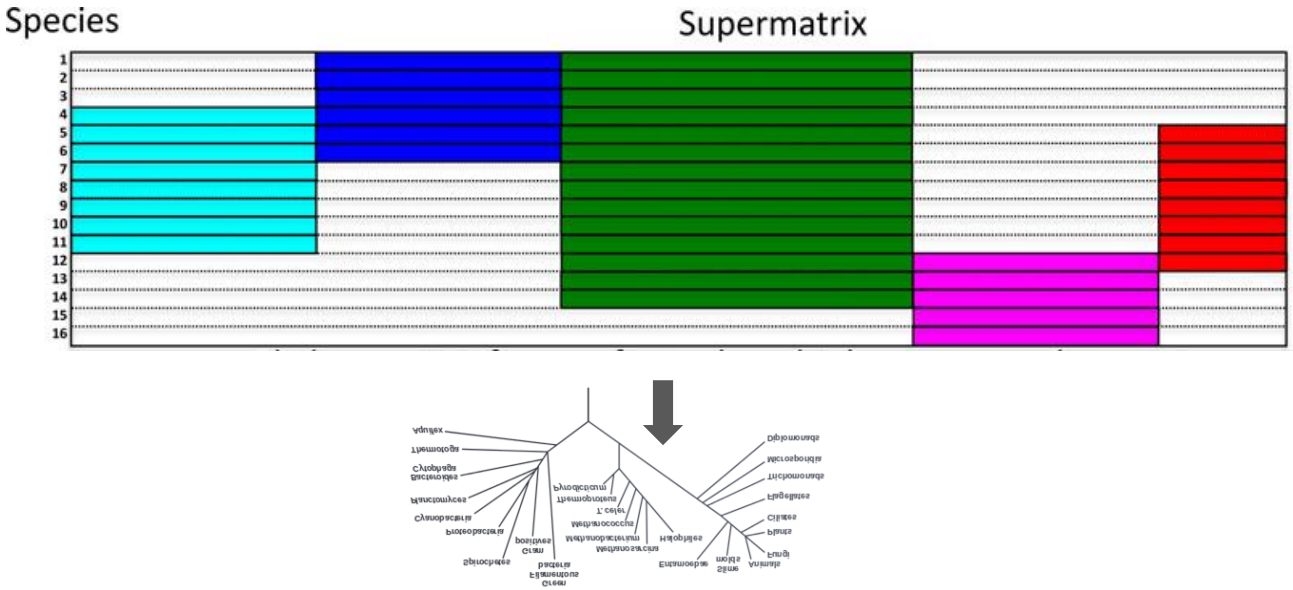
Jukes – Cantor model

WRONG!

$$Q = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

Supermatrix method

- Concatenates sequences from multiple genes into a single, giant phylogenetic matrix.
- Gaps in place of missing genes.
- Analyzing all the characters simultaneously.



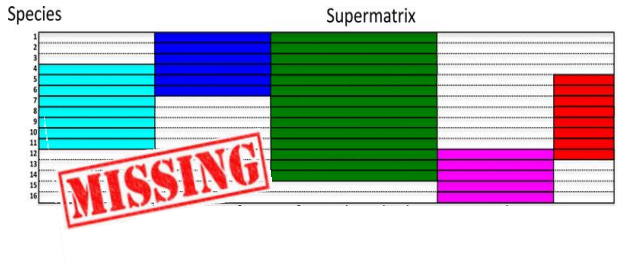
The supermatrix approach to systematics Alan et al. (2006)

Supermatrix method

Strengths

- Uses character evidence more fully in estimating the tree than do supertree methods.
- In supertree data is lost when sets of characters are summarized as trees.
  More information is better than less information!
- The phylogenetic signal from the analysis of the full, direct data can be different from what is apparent in the trees from the separate analyses.
- This is because the combined analysis enables the signal to assert itself more strongly over noise.

Missing Data



- Missing data are empty cells in a phylogenetic data matrix, and are often viewed as a liability.
- However, the crucial issue is whether or not the characters of the taxon are sufficiently informative, rather than the proportion of missing characters.
- Argument for using supertree: each of the separately analyzed data might have few or no empty cells.
- However, each taxon in the supermatrix is actually coded for as many or more characters than the same taxon in any of the separate data sets.
- Thus, overall the supermatrix is much more complete than any of the component data sets.