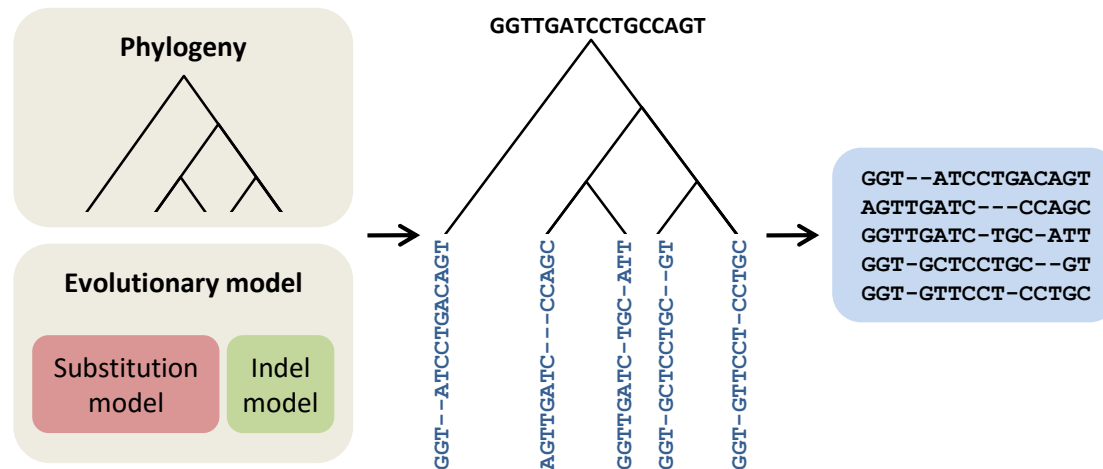


Towards Realistic *Sequence Simulations* for the Reconstruction of Mega Phylogenies

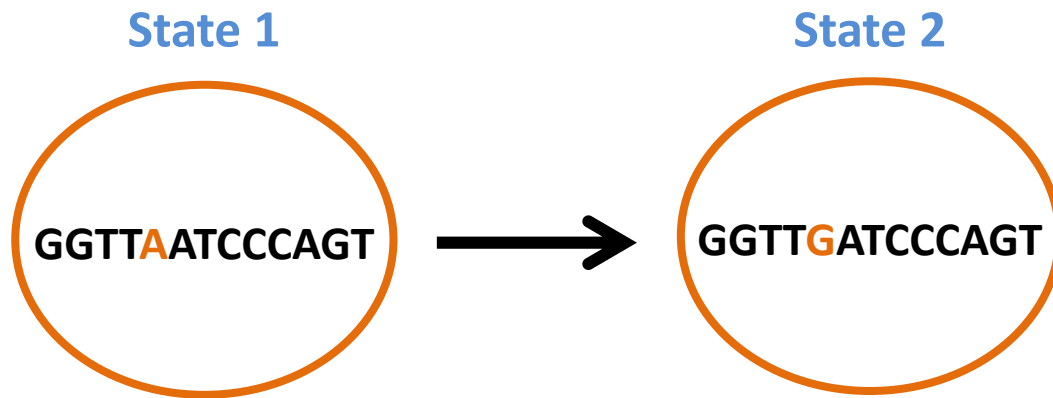
Thesis submitted towards the M.Sc. degree in Bioinformatics
in Tel-Aviv University
by
Nomi Hadar

Schema of typical sequence simulator



Markov chain

State of the chain = whole sequence



Simulating Markov chains using
Gillespie's algorithm



$\exp(\lambda)$

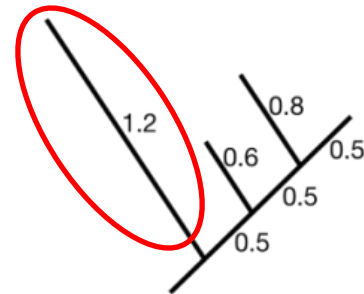


INDELible simulator – a closer look

Given:

GGTTGATCCTGCCAGTAGTCATCT

sequence



branch length (t)

Denote:

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

Insertion rate

Deletion rate

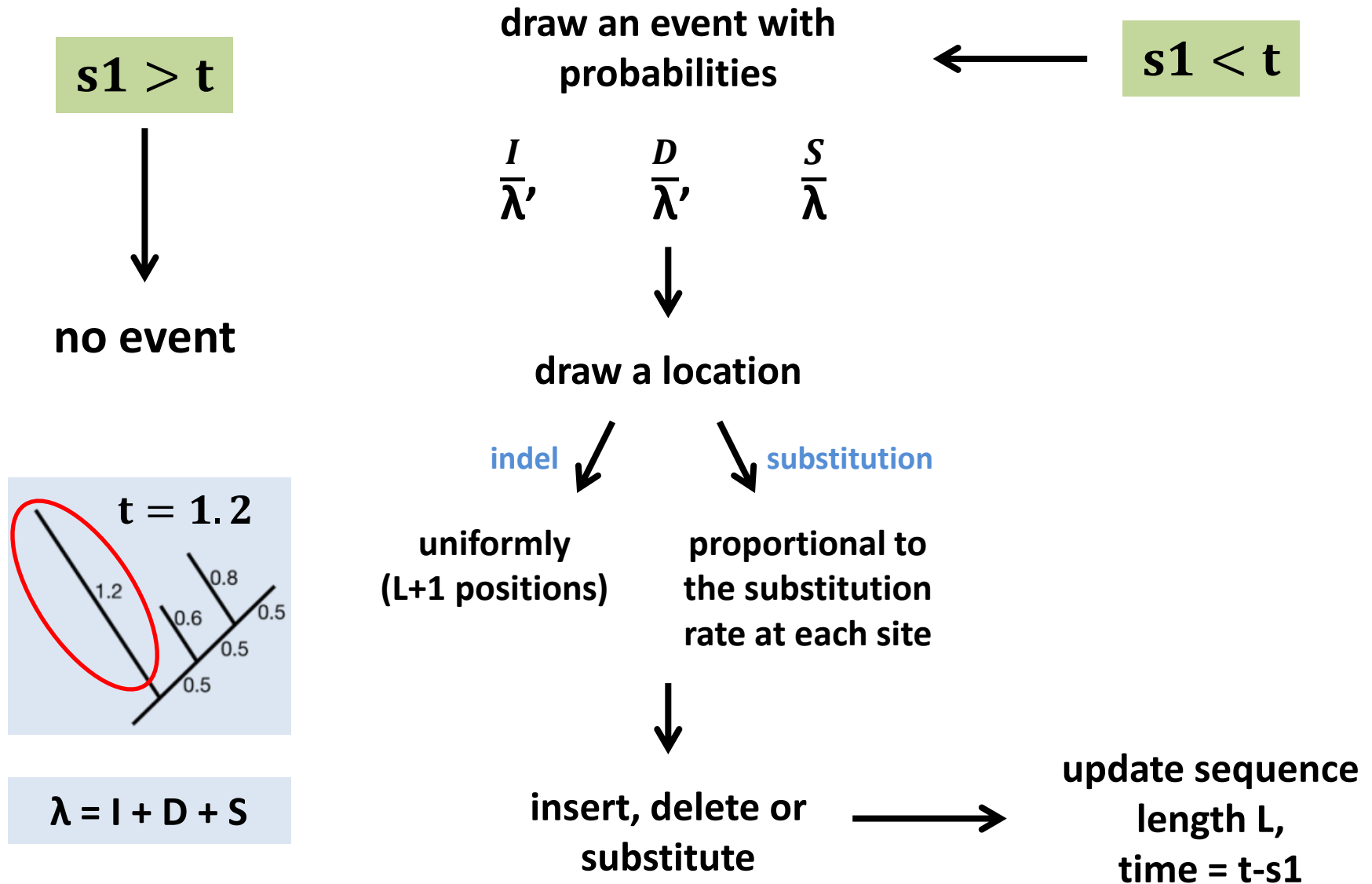
Substitution rate

**Waiting
time:**

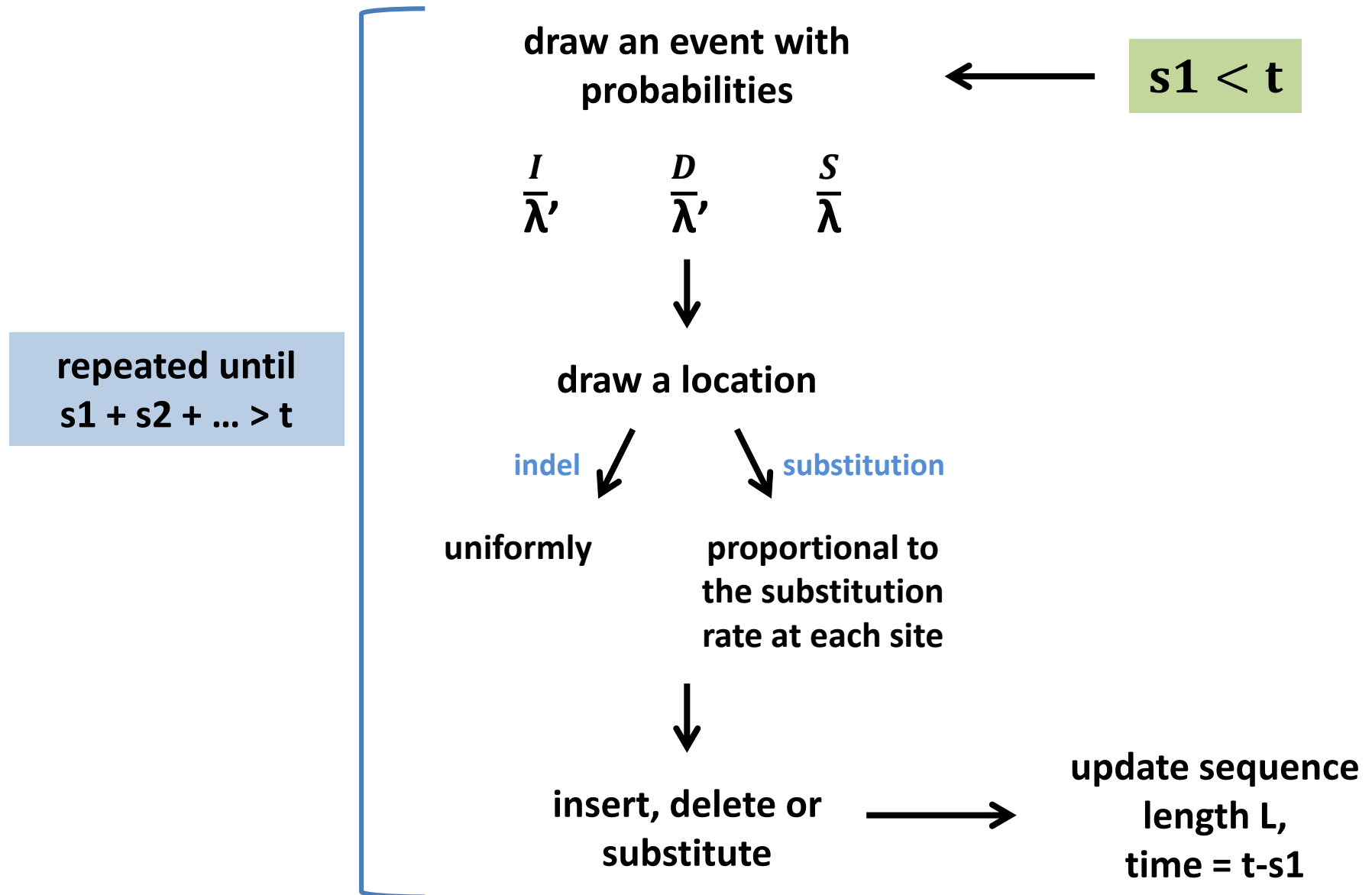
s_1

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

INDELible simulator – a closer look



INDELible simulator – a closer look



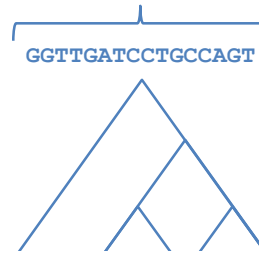
Estimating the parameters for sequence simulations

'IR'
Indel Rate

$$IR \in [0,1]$$

IR ↓	IR ↑
G-TAAGTGTGCGG	GT-CAGTGT--GG
G-TAAGT-TGCGG	G-TAA--G-GCGG
GTCAAGTGTG--G	A--AA-TGTG---

'RL'
Root Length



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

GG--TGCATCC-----TGA
GGAATTGC-----TCCGCAGT



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

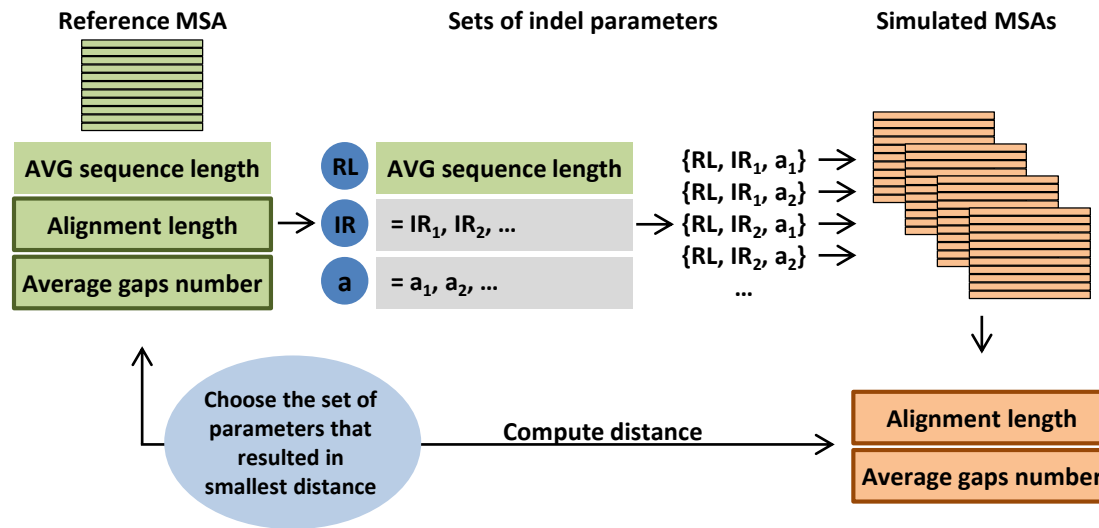
Estimating the parameters for sequence simulations: *existing tools*

Length and average number of gaps of reference and simulated MSAs

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

Do not resemble the reference!

Scheme of OPTIMIM method to infer indel parameters



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

Do resemble the reference!

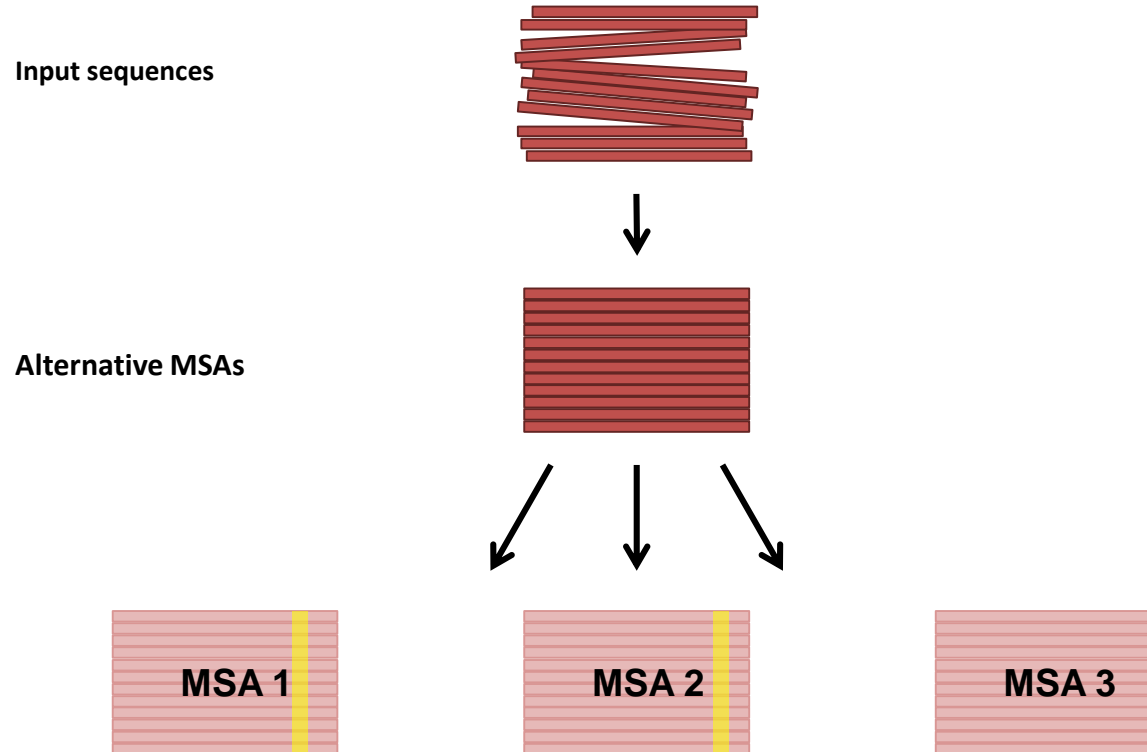
Which sequences are simulated?

13351				CGAA	CTCGTGACCCCTTTTTTTTGGGGTAGGGGCGAAGGG	CTGCTCTTCAGT
4424	CCAATCCTG	ATCGACACAC	CCGCGAACACGTTATCATCACTCGTCGTCGGATTGCCACCCCGGGTGGG	CGTCCCCCGACCC		
105750	-CGATGACC	CACTAGGCTGAC	CGGCGAA	CTTGATATTGTTCCCTTGAGGGATGACGGTCTATCCT	TGCGATACCGGAA	
16752	TGATGCGTA	CCAAAAACAAGATCGACCGAAGCGAA	CATGTGACCCCTTGTCTTTGCTTGCCTGCGTAGGATTG	CCGTCTCGG		
12953	TCCAGGACC	GAAACGAGAGGGTAGAC	CCGCGAA	CATAGTGACCCCGCGCAAAATCGGGAGGGGGCGGGGGCGC	GGCTGCCCCCGGT	
22294	TGGAGAACCGAA	CCCCCTGCAAAGCAGCACGACTG	GTGTGAA	TATGTCATTTGTCACTGGGGGTGGGGACGCTAGCGCC	TAGCGACGCCCC	
112846	TTGAGAACT	GAAAAATGCAGAGCAGCACGACTA	GTGTGAA	CTTGTGAATTTATCATTTGGGGGCGGGGACACTAGTGCC	TAGTGACGCCCC	
3429	T	CGACCCGCGAACCGGACATCAGTTTGACCTGCCTCTG	CCCCCTCGGTC	CGGGTCCGGTCCGGTGGC	AGAC	
13083	TCGCAAAACC	AGAGCCAC	CGGTGAACCCGTTCTGTCACTGAAGTTTGGGGTTGTGCATCTGAGGCCTC	GAAACGACCTGGA		
3426	TCCAGGCCC	CTCCCTGCTAAAAAGCAGAACGAC	CCGCGAA	GGAGTGAAAGCATTGCTCTCGGTGGCGGAACTGCTG	TCACGCCCCAG	
13419	TCGTTCCT	CGACCCGCGAACCCGATATCAGTCTGACCTGCCTCT	GCCCCCGCCGTCGCTC	CGGGTCCGGTCCGGTGGC	AGAC	
13260	TGATGCCCTA	CCAAAAACAAGATCGACCGAAGCGAA	CATGTGAACCTTGCCTTCTCCGCCCCGCGCGGGGCC	CCGTGCCCC		
13559	TCGATACCT	ATAAAAGCGAGACCGAGAAATTGTTATCTGTTTTTTTCTTTGTTCATCGATATCGAGAGATATA	TACGGTGCCTCC			
43891	TCGAAGCCT	GCAACGGCAGAACGAC	CCGCGAA	CTCGTACCCGATCGCACCCCGGGGGCGGACCGTCCGG	CGCCCCGGCGTC	
2708	TTGAAACCT	GCC CAGCAGAACGAC	CCGTGAA	CCAG AGATATCACCGGCGCGGGAGGGGGGATCGCTC	CGCAGCGGG	
2711	TCGAAACCT	GCC CAGCAGAACGAC	CCGCGAA	CCAGTTGATATCACCGGCGCGGGAGGGGGGACCGGCC	CGCAGCGGG	
43166	TCGACCTGC	CAGCAGA CGAC	CCGCGAA	CCAGTTGATATCACCGGCGCGGGAGGGGGGCGCTC	CGCAGCGGG	
105808	TCGAAACCT	GCCTAGCAGAACGAC	CCGCGAA	TACGTGAACGTGCTGAGGGGGGACGACGGGCGTCGCG	CACCTTTGCTC	
85280	TCGAAACCT	GCC TAGCAGAACGAC	CTGTGAA	TTTGTGAATGTTTTGGGGGCGAGGGCTGTGCTTGTGC	AGCTGTCTGT	
13654	TCGAAACCT	GCCTAGCAGAACGAC	CAGCGAA	CTTGGTCCGATACGGATACGGCCATCGGCCCTTTCGTAG	TTGGCCCTGTTT	
3760	TCGAAACCT	GCC TAGCAGAACGAC	CCGGAAC	TAGCGGGGGGCGAGGGGTCCCTCGGG	CTCCTTTGTC	
3486	TCGATGCCT	GCAACAGCAGAACGAC	CCGGAAC	CACGTACCTTTGGGTGGGCGAGAGGAGCTTGCTCCTTGG	ACCCGCCCTCAC	
3483	TCGAAACCT	GCAACAGCAGAACGAC	CCGGAAC	CACGTG CTTGGGCGGGCGAGAGGAGCTTGCTCCTTGG	ACCCGCCCTCAC	
3460	TCGATGCCT	GCAATAGCCCCGCGTGTGTTTTCAATTCTACGGGAGCCTCCCGCGGGAGAGC	TGTCCCCCTCGCG			
3813	CCGC	AAAAAGAGCGAC	CCGGAAC	CCGGTTGGAAACAAACCTTTGGCGGGGGAGGCCGAGGCC	CCACCCCGC	
3879	TCGATGCCTTACATG	CAGTCCCAACAGGAATCAACACCTCGGCTTA		CCCTTTGGTTTACAGGGAAGACGACGAAAGTGC		
63479	TCGAATCCT	ACA GAGCGAACGAC	CCGCGAA	CTCGTTTATCCACGGGCGTCGGGCGCGGGGTCCCTGCC	CCGCCCGTC	
50993		GCCCTTCCCGACCGCGCAGGCCACCC	CTAG CACGCATA	CCCCCGCACAAACGAGAGAGT	CGAGTC	GGGG CGCAGGATC
4400		GCCCTCCACCGGCCCGCAGGGCCTCTCC	CTAG CACGCATA	CCCCCGCGCCTTCGAGAGAGC	CCAGCC	GTGG AGCAAGGATC
50174		GCCCTTCTACTGTCCGCGCAACGCCGCTCC	CTAG CACGCATA	CGCCCGCACACCGCAGAGAGC	CCAGGA	GGGG CGCAAGGATC
45172		GCCCTCCACCGTCCGCGCAGGGATACCC	CTAG CACGCATA	CCCCCGCACCATCGAGAGAA	CTAGTT	GGGG AGCAAGGTTT
3592		GCCCTTCCACCGCCCGCGCAGGGCCAACTC	CTAG CACGCATA	CGCCCGCACAGTCGAGAGAGC	CTAGCC	CGGG GGCATGGGTC
3597		GCCCTTCCACCGCCCGCGCAAGGCCACCC	ATAG TACGCATA	CCCCCGCACCTCCCGAGAGAGC	CTAGCT	CGGG AGCAGGGTTC
4283		GCCCTCCCCCGCCCGCGCATGGCCACCC	CTAG CACGCATA	CGCCCGCACAACTCGAAGCAGT	CAAGTG	CGGG GGCAGGATC
4292		GCCCTCCACCGGCCCGCGCAGGGCCACCC	CTAG CACGCATA	CCCCCGCACCCACCGAGACAGC	CGAGTC	TGAG GGCAGGGCTC
16924		GCCCTCCGCGGGCCCGCGCAGGGCCACCC	CTAG CACGCATA	CCCCCGCACAACTCGAGACAGC	CGAGTG	CGGG GGCATGGATC
3741		GCCATCTACCGACCGCGCAGAGGCCACCC	CTAG CACGCATA	CCCCCGCACACTCGAGACAGT	CGAGTG	GGGG CGCAGGATC
4442		GCCATCTACCGACCGCGCAGAGGCCATCC	CTAG CACGCATA	CCCCCGCACACTCGAGACAGT	CGAGTG	GTGG GGCAGGATC
125047		GCCATCTACCGCCCGCGCAGGGGCCACCC	CTAG CACGCATA	CCCCCGCACAACTCGAGACACC	CGAGTG	GTGG GGCAGGATC
522403		GCCCTCTACCGACCGCGCATGGGTACGCC	CTAG CATGCATA	CCCCCGCACAGTCGAGACACC	CAAGTC	CGGG GGCAGGATC
3772		GCCCTTGCGACCGCGCAGGGGCCACCC	CTAG CACGCATA	CTCCTGCACACTCGAGGCAGC	CAAGTA	GAGG TGCAGGATC
4288		GCCCTCCACCGACCGCGTAGTGGCCACCC	CTAG CACGCATA	CCCCCGCACAACTCGAGACACC	CAAGCG	CGGG TGCAGCGATC
16901		GCCCTCCACCGCCCGCGTAGTGCTCCCC	CTAG CACGCATA	CCCCCGCACAACTCGAGACACC	CAAGGG	CGGG TGCAGCGATC
4146		GCCCTCTACCGCCCGCGCATGTCCACAC	CTAG CACGCATA	CCCCCGCACAAACCGAGACAGC	CGAGTG	GAGG CGCAACGGTC
3770		GCCCTCGACCGCCCGCGCAGGGCCACCC	CTAG TACGCATA	CGCCCGCACCGCCGAGACAGC	CCAGTC	GGGG CCCCACGGTC
39347		GCCCTCGACTGTCCGCGCAGGTCCACCC	CTAG CATGCATA	CGCCCGCACACTCGAGCCAGG	CGAGTG	GTGA AGCAACGGTC
443381		GCCCTCGACCGCCCGCGCAGGTCCATCC	CTAG CACGCATA	CGCCCGCACCAACCGACACAGG	CGAGTG	GAGA CGCAACGGTC
39388		GCCCTCGACCGCCCGCGCAGGTCCACCC	CTAG CACGCATA	CGCCCGCACCAACCGACACAGA	CGAGTG	GTGA CGCAACGGTC
4151		GCCCTCGCCGGGCCCGCGCAAGTCCACCC	CTAG AACGCATA	ATCCCGCATAAACGATACAGG	CCAGTG	GAGA GGCACCGTC
197815		GCCCTCGGGCGACCGCGGAGGTGCGGCC	CTAG CACGCATA	TGCCCCGATGGCCGAGACAGG	CAAGCT	GGGA CGCAACGGTC
201512		GCCCTCGACCGCCCGCGCAGGGTCCACCC	CTAG CACGCATA	CTCCCGCACATACTGAGACAGC	CGAGTC	GTGG GGCACCGTTC
39353		GCCCTCGACCGCCCGCGCAGGTCCACAC	CTAG CACGCATA	CCCCCGCACCAACCGAGACAGC	CCAGTG	GAGG CGCAACGGTC
28536		GCCCTCGACCGTTCGCGCAGGACCCACCC	CTAG CACGCATA	CCCCCGCACCAACCGATACAGC	CCAGTG	GCGA TGCAGCGGTC
49121		GCCCTCGACCGACCGCGCAGGACCTCC	CTAG CACGCATA	CCCCCGCACACCCGAGACAGC	CGAGTG	GGGG CGCAGCGGTC

ITS gene

Evaluating the complexity of the simulated alignments

GUIDANCE2 score



Column score: 2/3

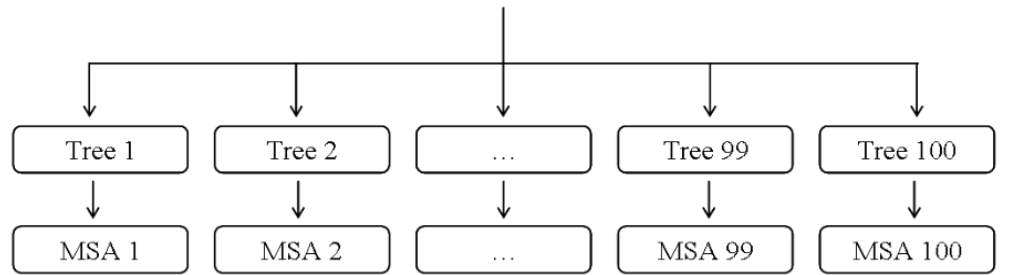
Evaluating the complexity of the simulated alignments

A schematic flowchart of the GUIDANCE2 algorithm

Base MSA

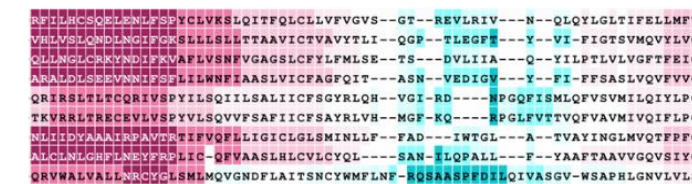
```
RFILHCSQELNLFSPYCLVKSQITFQLCLLVFVGVS--GT--REVLRIV---N--QLQYLGLTIFELLMFT
VHLVSLQNDLNGIFGKSLLSLLTTAAVICTVAVYTLI--QGP--TLEGFT---Y--VI-FIGTSVMQVYLVC
QLNLGLCRKYNDFKVAFLVSNFVGAGSLCFYLFMLSE--TS---DVLIIA---Q--YILPTLVLVGFTFEIC
ARALDLSEEVNIFSFILWNFIAASLVICFAGFQIT---ASN--VEDIGV---Y--FI-FFSASLVQVFVVC
QRIRSLTLCQRIVSPYILSQIILSALIICFSGYRLQH--VGI-RD----NPGQFISMLQFVSMILQIYLP
TKVRRLTRECEVLVSPYVLSQVVFSAFIICFSAYRLVH--MGF-KQ----RPGLFVTTVQFVAVMIVQIFLPC
NLIIDYAAAIRPAVTRTIFVQFLLIGICLGLSMLNLLF--FAD---INTGL---A--TVAYINGLMVQTFPFC
ALCLNLGHFLNEYFRPLIC-QFVAASLHLCVLCYQL---SAN-ILQPALL---F--YAAFTAAVVGQVSIYC
QRVWALVALLNRCYGLSMLMQVGNDFLAITSNCYWMFLNF-RQSAASPFIDILQIVASGV-WSAPHLGNVLVLS
```

Bootstrap guide-trees

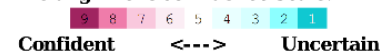


Alternative MSAs

GUIDANCE2 scores

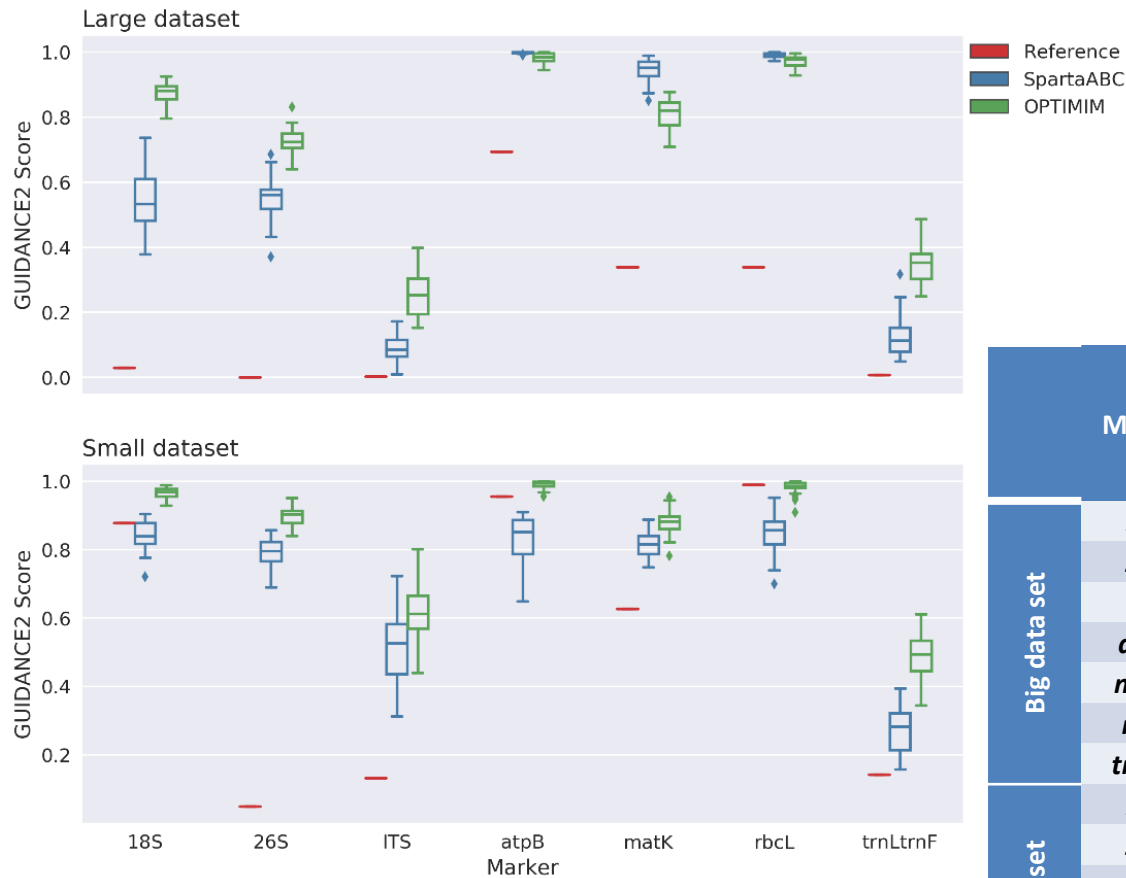


The alignment confidence scale:



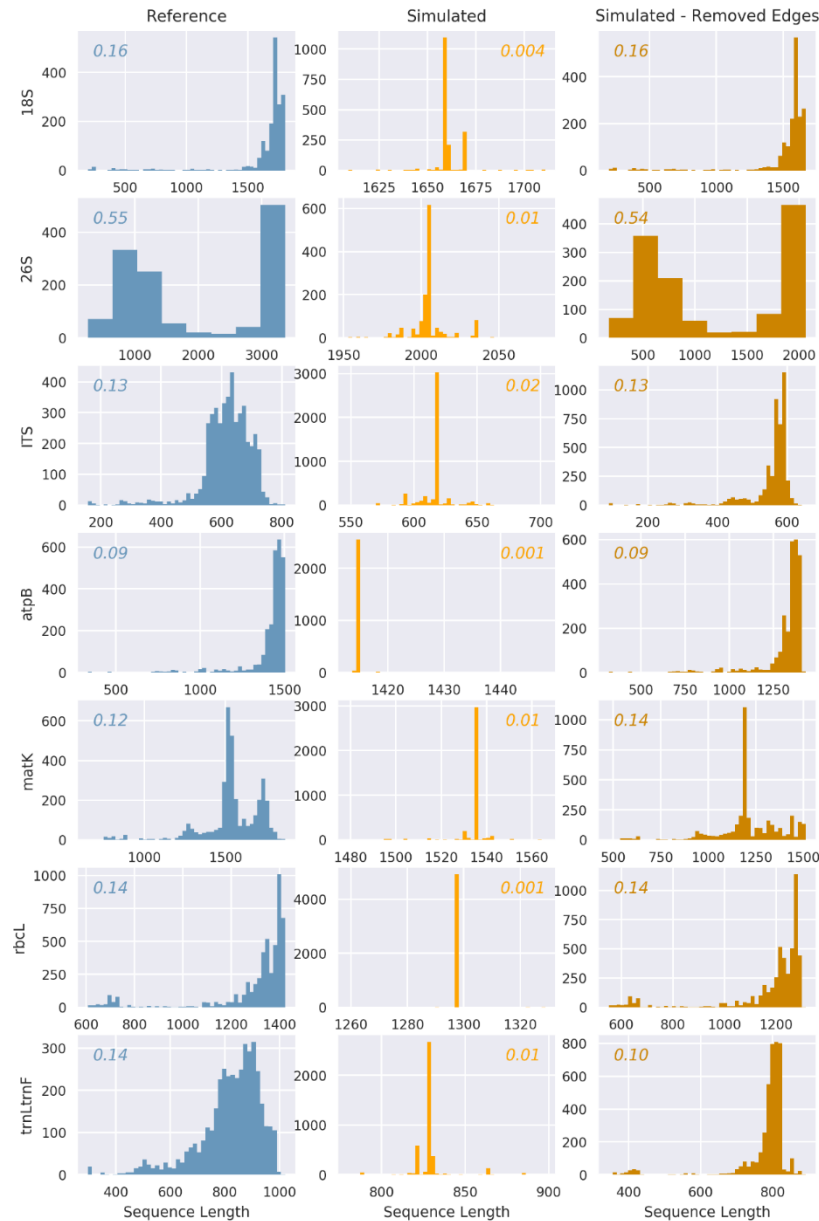
Evaluating the complexity of the simulated alignments

Alignment complexity of real and simulated data



	Marker	Reference MSAs	Simulated MSAs	
			<u>SpartaABC</u>	<u>OPTIMIM</u>
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

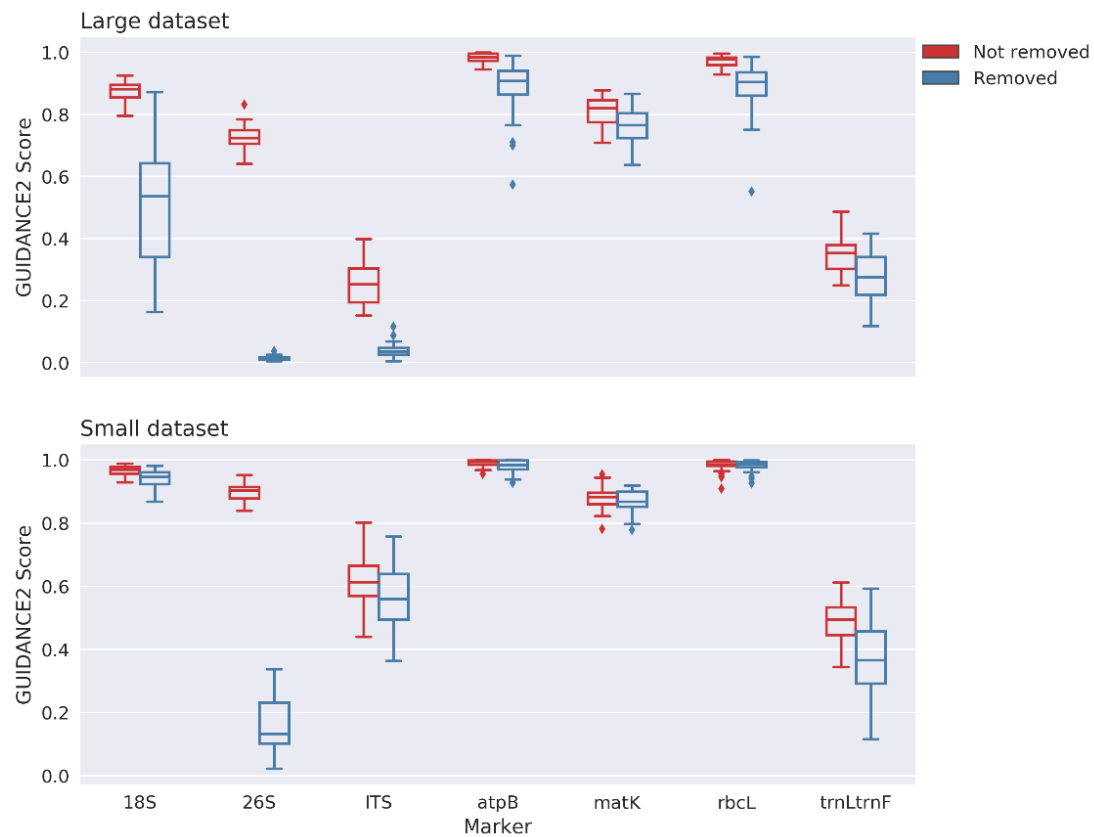
Increasing simulation complexity - examining simulated sequence lengths



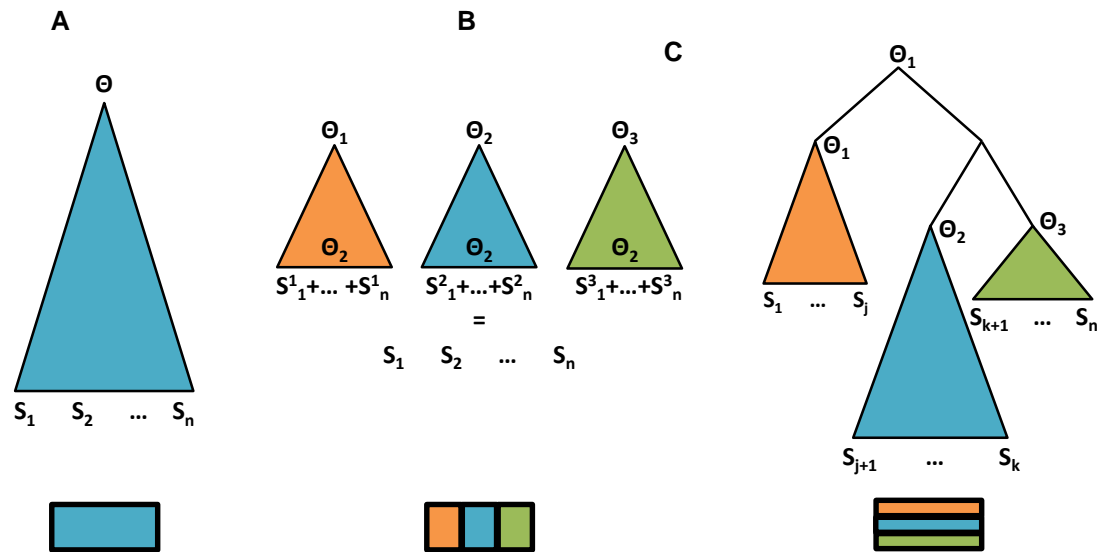
**Histograms of
sequence lengths for
reference and
simulated alignments
of large dataset**

Increasing simulation complexity - examining simulated sequence lengths

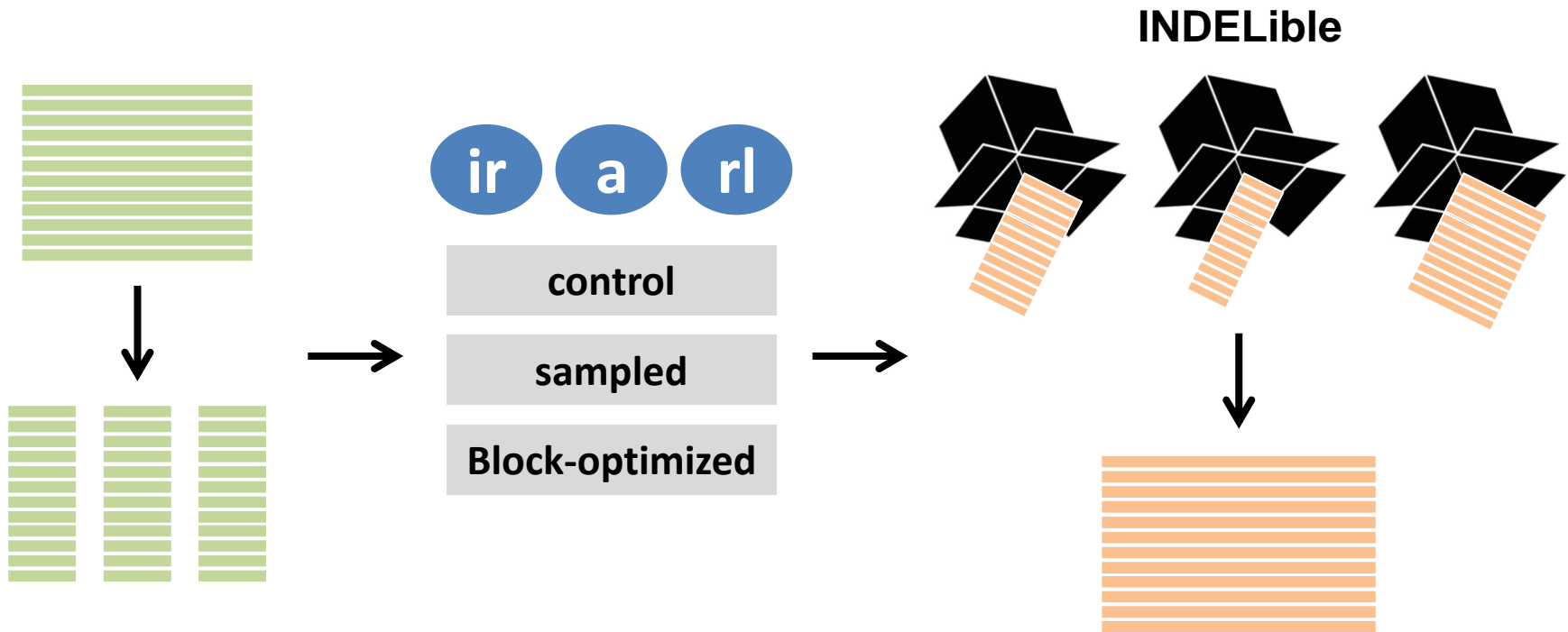
Alignment complexity of simulated data following the removal of edges



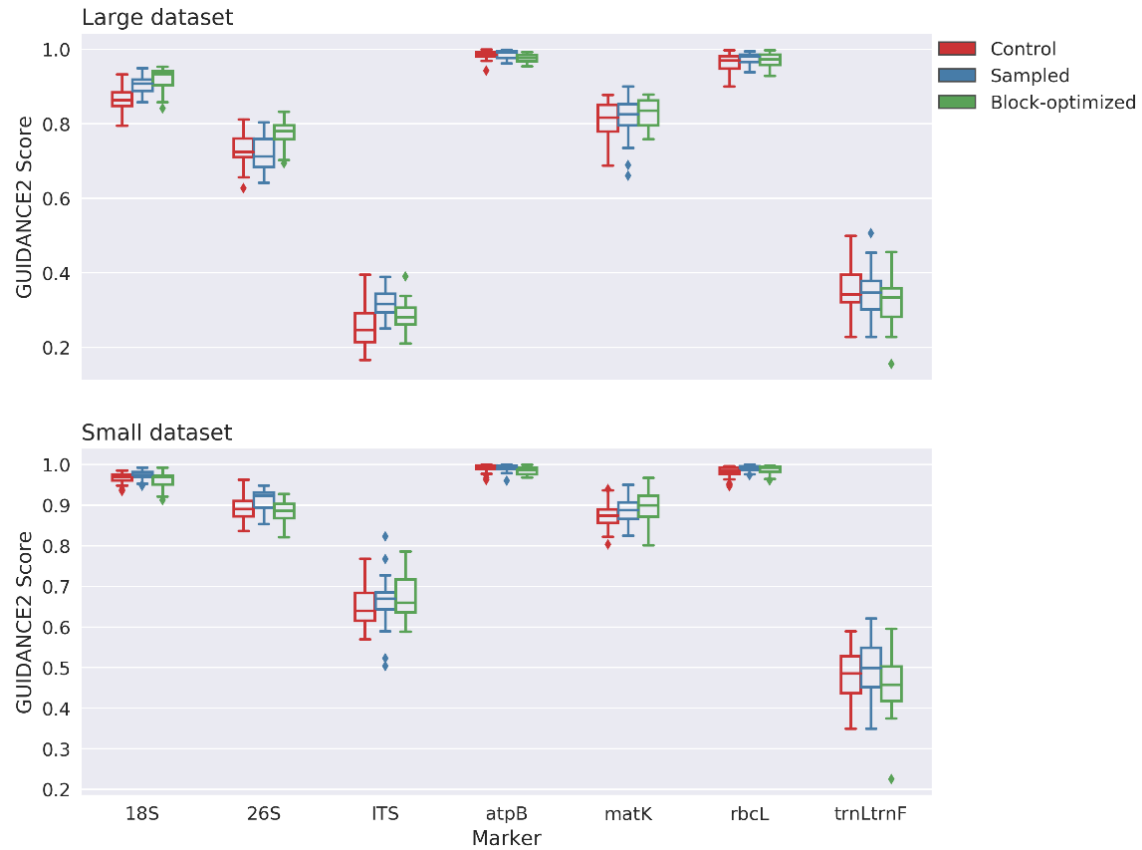
Increasing simulation complexity - non-homogeneous indel model



Increasing simulation complexity - non-homogeneous indel model

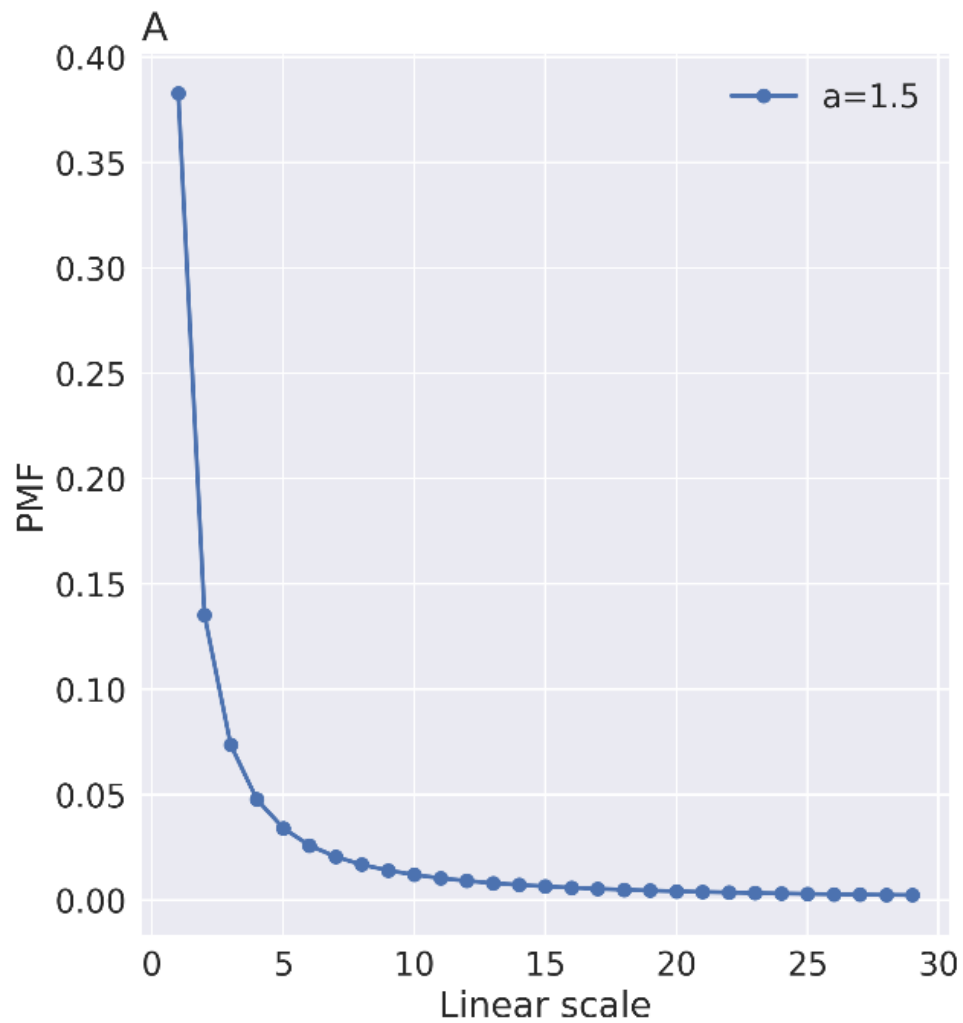


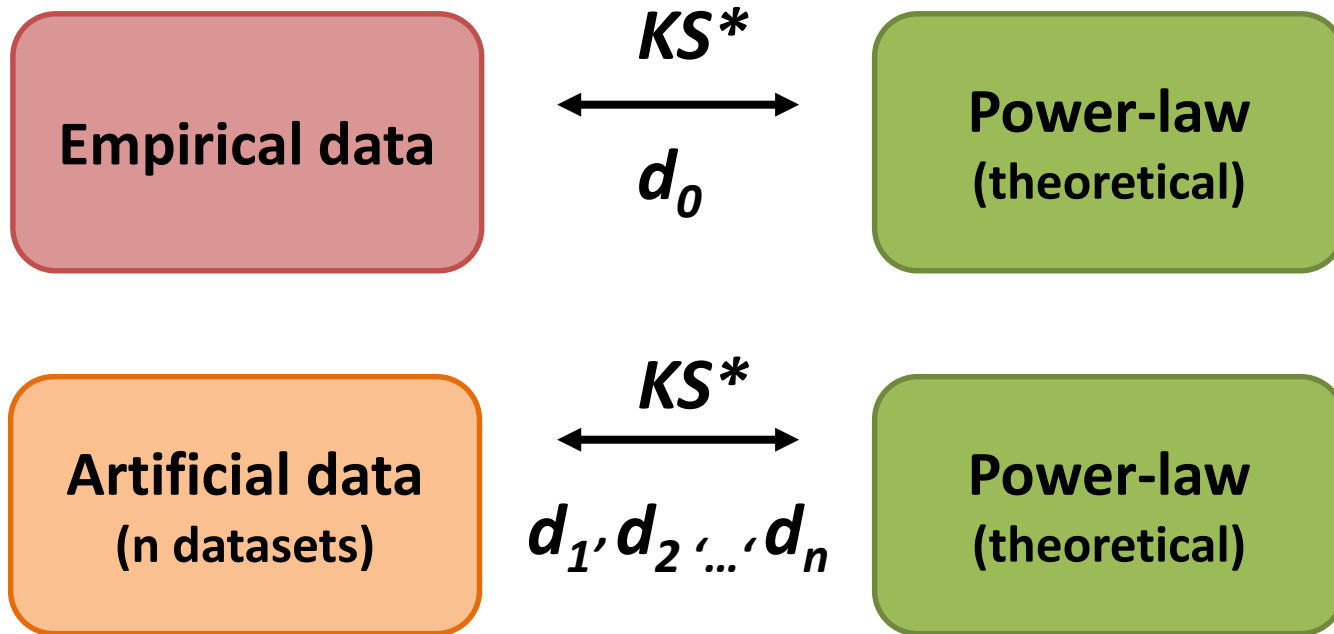
Increasing simulation complexity - non-homogeneous indel model



Increasing simulation complexity - indel-length distribution

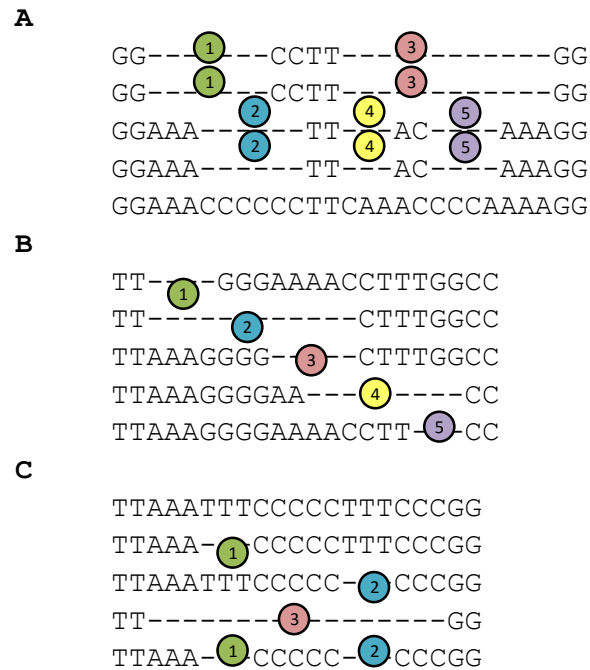
Power-law



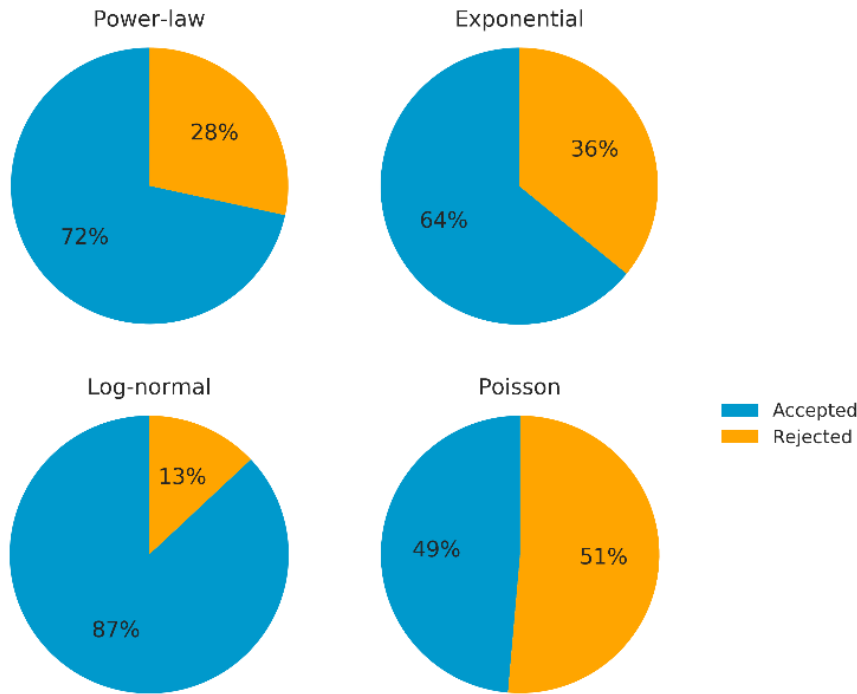


$$p - value = \frac{\{d_i > d_0\}_{1 \leq i \leq n}}{n}$$

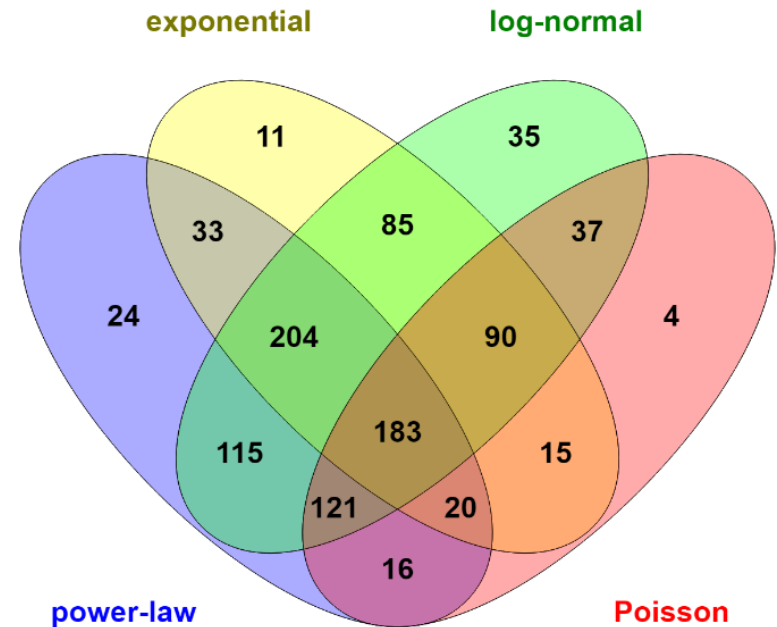
Increasing simulation complexity - indel-length distribution



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
raw pair-wise distance

ACGGTCATGTATAC
TCTCTCCTGTGAAT
 $d_1 = 0.5$

d_2
Jukes-Cantor distance

ACGGTCATGTATAC
TCTCTCCTGTGAAT
 $d_2 = 0.8$

- $x_i = |d_1 - d_2|$ ($x = 0.3$)
- Measure of dispersion is based on the median absolute deviation (MAD):

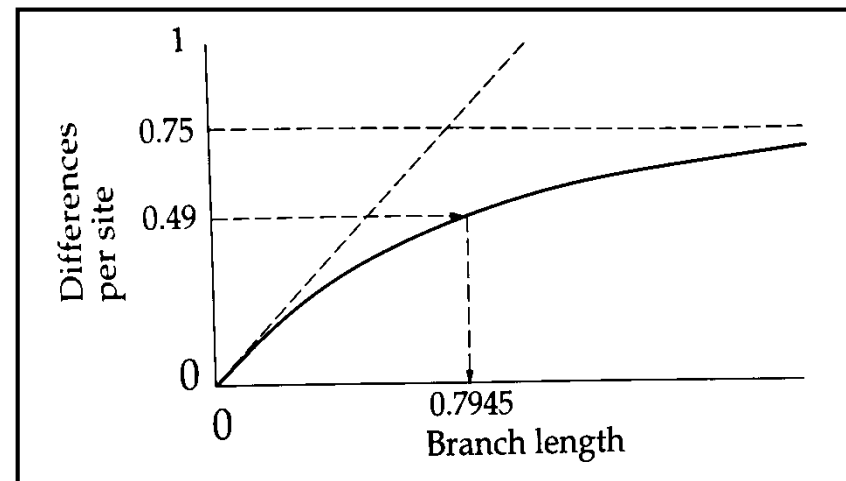
$$\text{MAD} = \text{Med} (|x_i - \text{Med}(x)|)$$

E.g.:

$x = (9, 6, 4, 2, 2, 1, 1)$, $\text{MED}(x) = 2$

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

$\rightarrow (7, 4, 2, 0, 0, 1, 1) \rightarrow (7, 4, 2, 1, 1, 0, 0) \rightarrow \text{MAD} = 1$



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 ----->

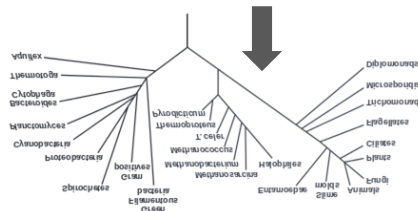
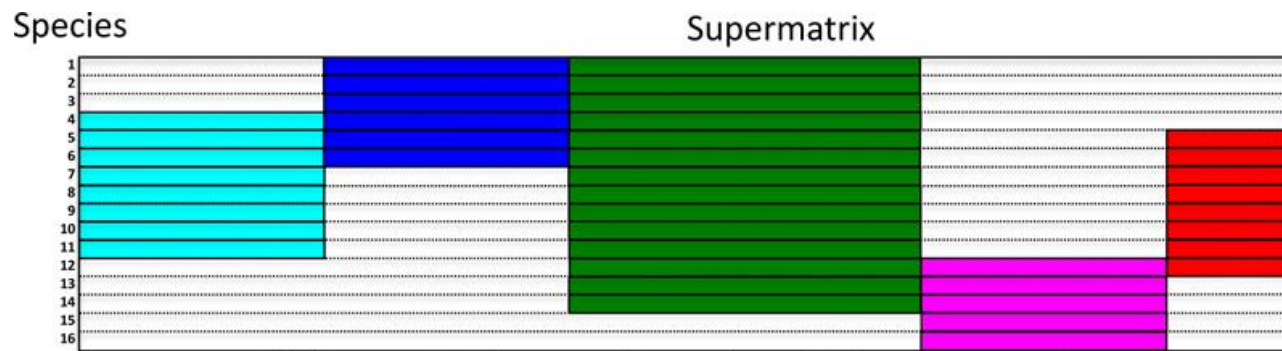
Jukes – Cantor model

WRONG!

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

Supermatrix method

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





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.

