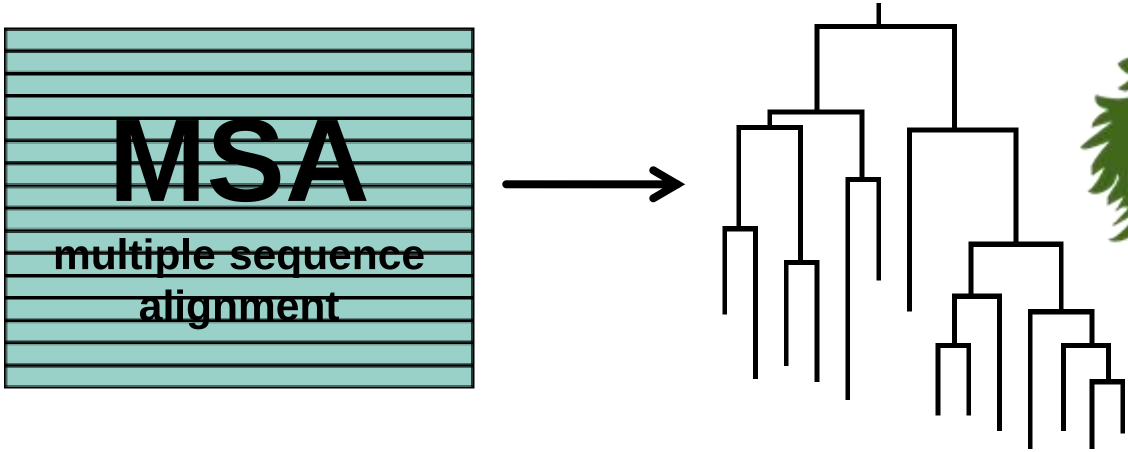
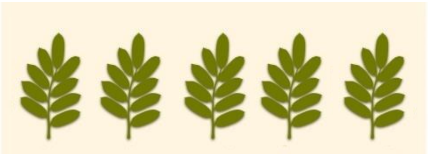
My research progress

A quick remainder of what I am doing



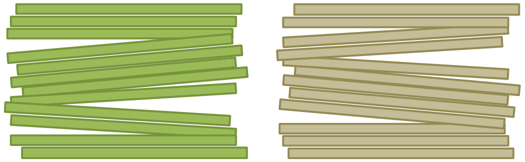
MSA

multiple sequence alignment



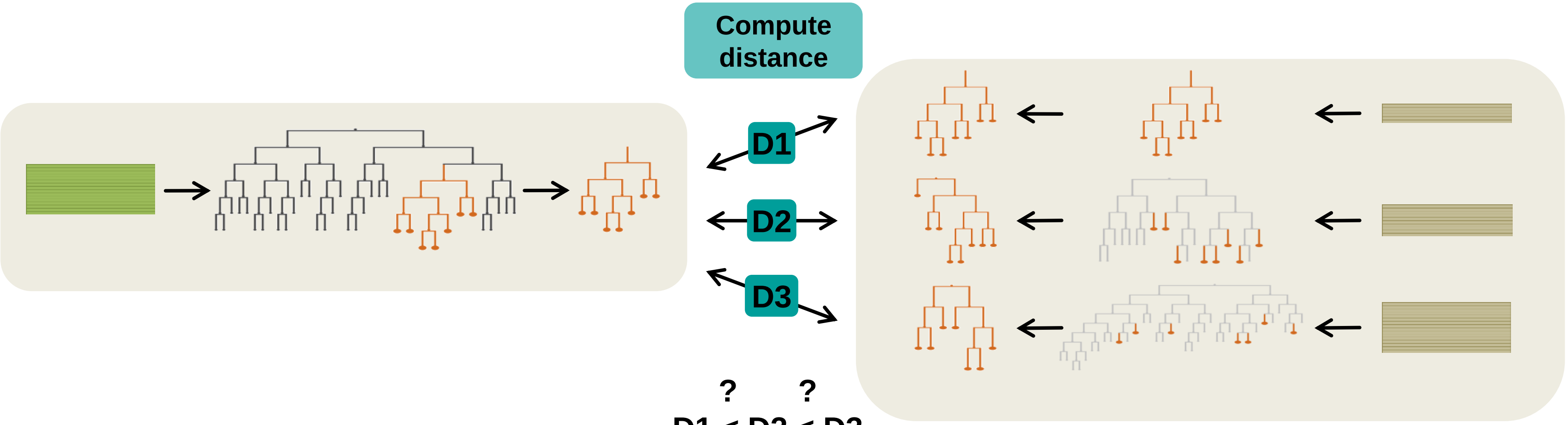
Examining the effect of sequence divergence

Sequences



Real
(NCBI)

Simulated
(INDELible)



Compute distance

D1

D2

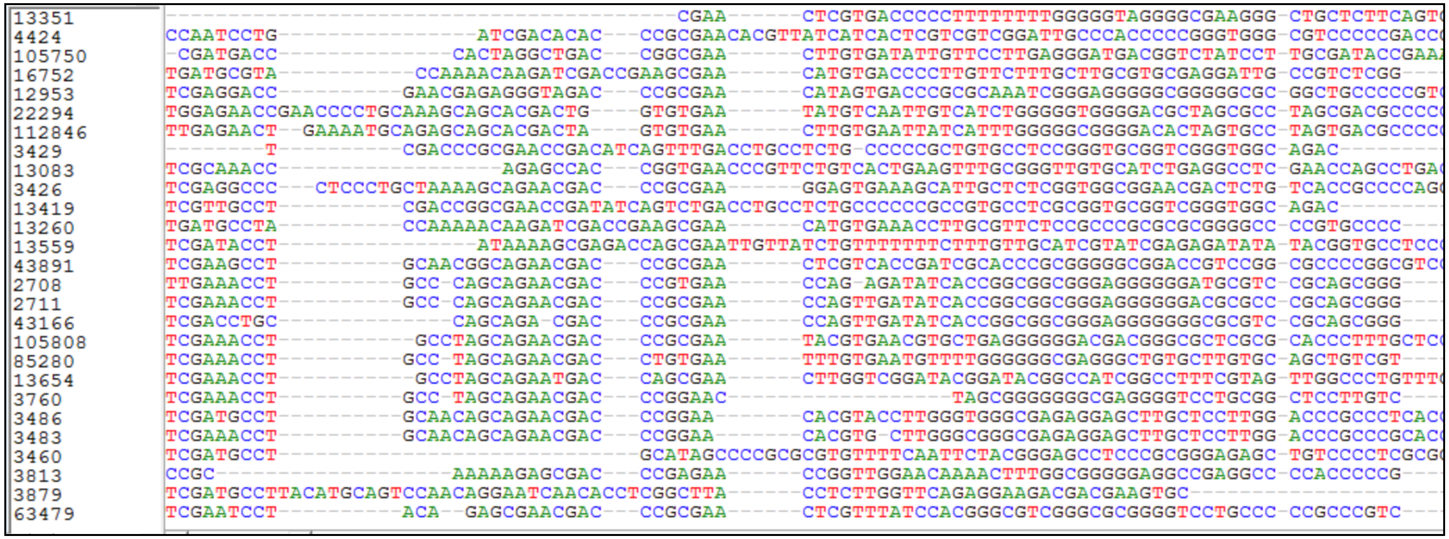
D3

? ?
D1 < D2 < D3

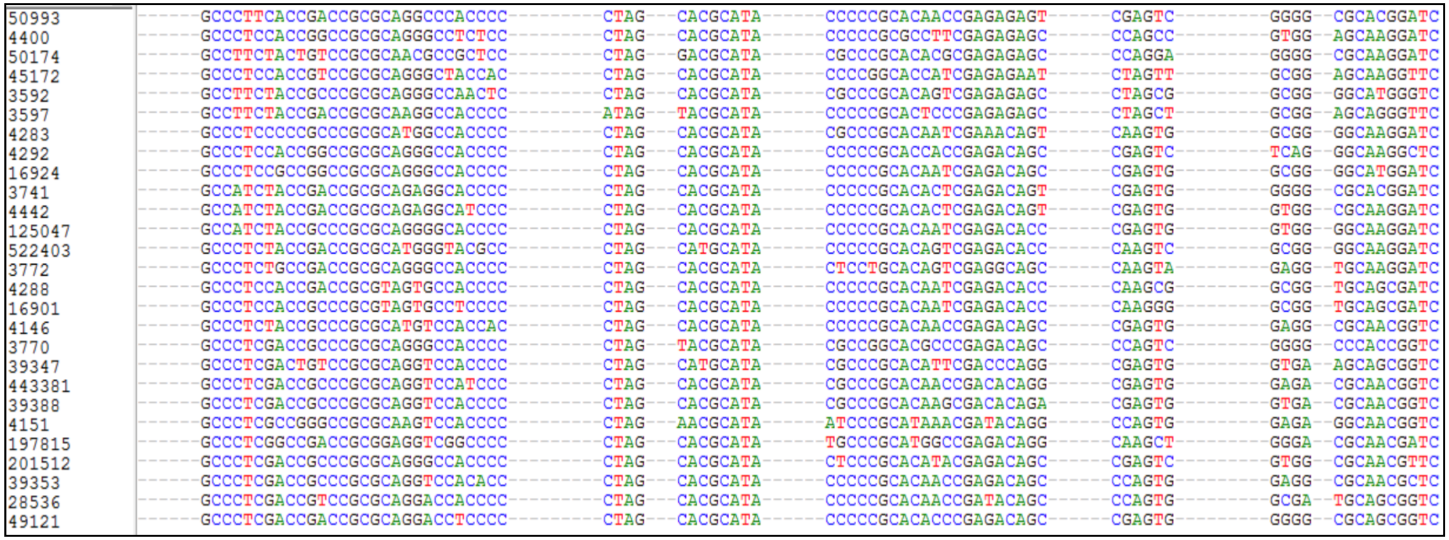
MSA size: 100, 800, 2000, 5000
Distance: 0.38 < 0.43 < 0.441 < 0443

D1 < D2 < D3 < D4

Which sequences are simulated?



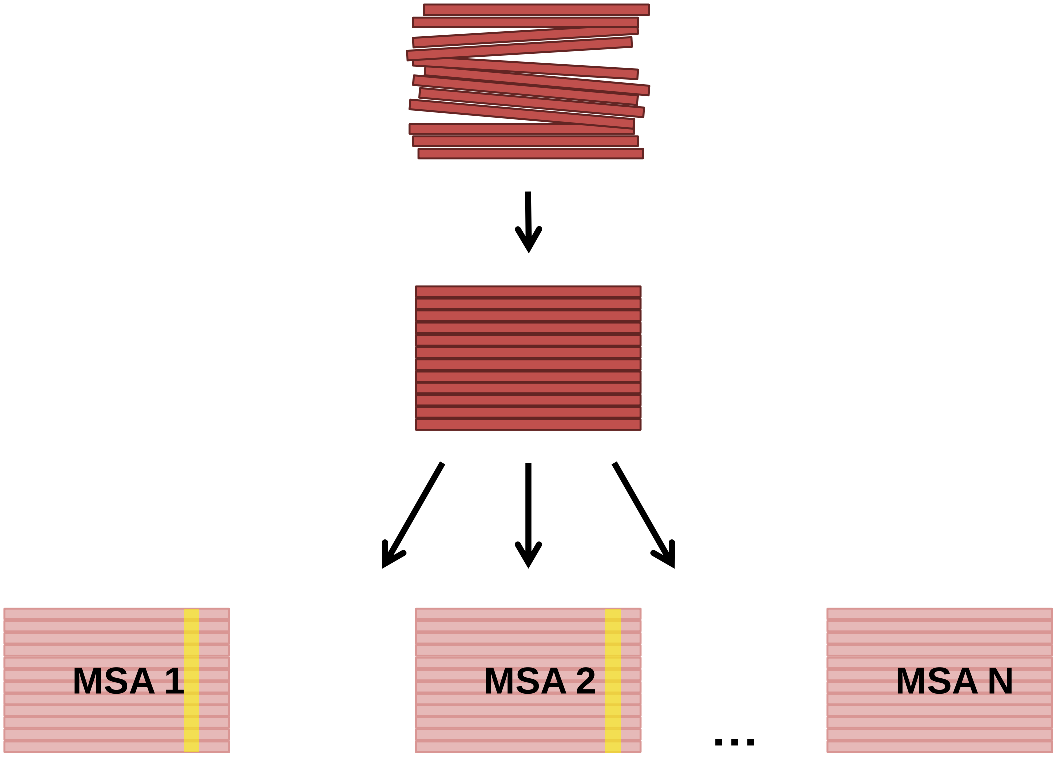
```
13351   -----------------------------------------CGAA------CTCGTGACCCCCTTTTTTTTGGGGGTAGGGGCGAAGGG-CTGCTCTTCAGT
4424    CCAATCCTG----------------ATCGACACAC---CCGCGAACACGTTATCATCACTCGTCGTCGGATTGCCCACCCCCGGGTGGG-CGTCCCCCGACC
105750  -CGATGACC--------------CACTAGGCTGAC---CGGCGAA------CTTGTGATATTGTTCCTTGAGGGATGACGGTCTATCCT-TGCGATACCGAA
16752   TGATGCGTA-----------CCAAAACAAGATCGACCGAAGCGAA------CATGTGACCCCTTGTTCTTTGCTTGCGTGCGAGGATTG-CCGTCTCGG----
12953   TCGAGGACC----------GAACGAGAGGGTAGAC---CCGCGAA------CATAGTGACCCGCGCAAATCGGGAGGGGGCGGGGGCGC-GGCTGCCCCCGT
22294   TGGAGAACCGAACCCCTGCAAAGCAGCACGACTG----GTGTGAA------TATGTCAATTGTCATCTGGGGGTGGGGACGCTAGCGCC-TAGCGACGCCCC
112846  TTGAGAACT--GAAAATGCAGAGCAGCACGACTA----GTGTGAA------CTTGTGAATTATCATTTGGGGGCGGGGACACTAGTGCC-TAGTGACGCCCC
3429    --------T----------CGACCCGCGAACCGACATCAGTTTGACCTGCCTCTG-CCCCCGCTGTGCCTCCGGGTGCGGTCGGGTGGC-AGAC---------
13083   TCGCAAACC------------------AGAGCCAC---CGGTGAACCCGTTCTGTCACTGAAGTTTGCGGGTTGTGCATCTGAGGCCTC-GAACCAGCCTGA
3426    TCGAGGCCC---CTCCCTGCTAAAAGCAGAACGAC---CCGCGAA------GGAGTGAAAGCATTGCTCTCGGTGGCGGAACGACTCTG-TCACCGCCCCAG
13419   TCGTTGCCT----------CGACCGGCGAACCGATATCAGTCTGACCTGCCTCTGCCCCCCGCCGTGCCTCGCGGTGCGGTCGGGTGGC-AGAC---------
13260   TGATGCCTA----------CCAAAAACAAGATCGACCGAAGCGAA------CATGTGAAACCTTGCGTTCTCCGCCCGCGCGCGGGGCC-CCGTGCCCC----
13559   TCGATACCT----------------ATAAAAGCGAGACCAGCGAATTGTTATCTGTTTTTTTCTTTGTTGCATCGTATCGAGAGATATA-TACGGTGCCTCC
43891   TCGAAGCCT----------GCAACGGCAGAACGAC---CCGCGAA------CTCGTCACCGATCGCACCCGCGGGGGCGGACCGTCCGG-CGCCCCGGCGTC
2708    TTGAAACCT----------GCC-CAGCAGAACGAC---CCGTGAA------CCAG-AGATATCACCGGCGGCGGGAGGGGGGATGCGTC-CGCAGCGGG----
2711    TCGAAACCT----------GCC-CAGCAGAACGAC---CCGCGAA------CCAGTTGATATCACCGGCGGCGGGAGGGGGGACGCGCC-CGCAGCGGG----
43166   TCGACCTGC--------------CAGCAGA-CGAC---CCGCGAA------CCAGTTGATATCACCGGCGGCGGGAGGGGGGGCGCGTC-CGCAGCGGG----
105808  TCGAAACCT-----------GCCTAGCAGAACGAC---CCGCGAA------TACGTGAACGTGCTGAGGGGGGACGACGGGCGCTCGCG-CACCCTTTGCTC
85280   TCGAAACCT----------GCC-TAGCAGAACGAC---CTGTGAA------TTTGTGAATGTTTTGGGGGGCGAGGGCTGTGCTTGTGC-AGCTGTCGT----
13654   TCGAAACCT-----------GCCTAGCAGAATGAC---CAGCGAA------CTTGGTCGGATACGGATACGGCCATCGGCCTTTCGTAG-TTGGCCCTGTTT
3760    TCGAAACCT----------GCC-TAGCAGAACGAC---CCGGAAC------------------TAGCGGGGGGGCGAGGGGTCCTGCGG-CTCCTTGTC----
3486    TCGATGCCT----------GCAACAGCAGAACGAC---CCGGAA-------CACGTACCTTGGGTGGGCGAGAGGAGCTTGCTCCTTGG-ACCCGCCCTCAC
3483    TCGAAACCT----------GCAACAGCAGAACGAC---CCGGAA-------CACGTG-CTTGGGCGGGCGAGAGGAGCTTGCTCCTTGG-ACCCGCCCGCAC
3460    TCGATGCCT-----------------------------GCATAGCCCCGCGCGTGTTTTCAATTCTACGGGAGCCTCCCGCGGGAGAGC-TGTCCCCTCGCG
3813    CCGC-------------------AAAAAGAGCGAC---CCGAGAA------CCGGTTGGAACAAAACTTTGGCGGGGGAGGCCGAGGCC-CCACCCCCG----
3879    TCGATGCCTTACATGCAGTCCAACAGGAATCAACACCTCGGCTTA------CCTCTTGGTTCAGAGGAAGACGACGAAGTGC---------------------
63479   TCGAATCCT----------ACA--GAGCGAACGAC---CCGCGAA------CTCGTTTATCCACGGGCGTCGGGCGCGGGGTCCTGCCC-CCGCCCGTC----
```



```
50993   ------GCCCTTCACCGACCGCGCAGGCCCACCCC---------CTAG---CACGCATA------CCCCCGCACAACCGAGAGAGT------CGAGTC---------GGGG--CGCACGGATC
4400    ------GCCCTCCACCGGCCGCGCAGGGCCTCTCC---------CTAG---CACGCATA------CCCCCGCGCCTTCGAGAGAGC------CCAGCC---------GTGG--AGCAAGGATC
50174   ------GCCTTCTACTGTCCGCGCAACGCCGCTCC---------CTAG---GACGCATA------CGCCCGCACACGCGAGAGAGC------CCAGGA---------GGGG--CGCAAGGATC
45172   ------GCCCTCCACCGTCCGCGCAGGGCTACCAC---------CTAG---CACGCATA------CCCCGGCACCATCGAGAGAAT------CTAGTT---------GCGG--AGCAAGGTTC
3592    ------GCCTTCTACCGCCCGCGCAGGGCCAACTC---------CTAG---CACGCATA------CGCCCGCACAGTCGAGAGAGC------CTAGCG---------GCGG--GGCATGGGTC
3597    ------GCCTTCTACCGACCGCGCAAGGCCACCCC---------ATAG---TACGCATA------CCCCCGCACTCCCGAGAGAGC------CTAGCT---------GCGG--AGCAGGGTTC
4283    ------GCCCTCCCCCGCCCGCGCATGGCCACCCC---------CTAG---CACGCATA------CGCCCGCACAATCGAAACAGT------CAAGTG---------GCGG--GGCAAGGATC
4292    ------GCCCTCCACCGGCCGCGCAGGGCCACCCC---------CTAG---CACGCATA------CCCCCGCACCACCGAGACAGC------CGAGTC---------TCAG--GGCAAGGCTC
16924   ------GCCCTCCGCCGGCCGCGCAGGGCCACCCC---------CTAG---CACGCATA------CCCCCGCACAATCGAGACAGC------CGAGTG---------GCGG--GGCATGGATC
3741    ------GCCATCTACCGACCGCGCAGAGGCACCCC---------CTAG---CACGCATA------CCCCCGCACACTCGAGACAGT------CGAGTG---------GGGG--CGCACGGATC
4442    ------GCCATCTACCGACCGCGCAGAGGCATCCC---------CTAG---CACGCATA------CCCCCGCACACTCGAGACAGT------CGAGTG---------GTGG--CGCAAGGATC
125047  ------GCCATCTACCGCCCGCGCAGGGGCACCCC---------CTAG---CACGCATA------CCCCCGCACAATCGAGACACC------CGAGTG---------GTGG--GGCAAGGATC
522403  ------GCCCTCTACCGACCGCGCATGGGTACGCC---------CTAG---CATGCATA------CCCCCGCACAGTCGAGACACC------CAAGTC---------GCGG--GGCAAGGATC
3772    ------GCCCTCTGCCGACCGCGCAGGGCCACCCC---------CTAG---CACGCATA------CTCCTGCACAGTCGAGGCAGC------CAAGTA---------GAGG--TGCAAGGATC
4288    ------GCCCTCCACCGACCGCGTAGTGCCACCCC---------CTAG---CACGCATA------CCCCCGCACAATCGAGACACC------CAAGCG---------GCGG--TGCAGCGATC
16901   ------GCCCTCCACCGCCCGCGTAGTGCCTCCCC---------CTAG---CACGCATA------CCCCCGCACAATCGAGACACC------CAAGGG---------GCGG--TGCAGCGATC
4146    ------GCCCTCTACCGCCCGCGCATGTCCACCAC---------CTAG---CACGCATA------CCCCCGCACAACCGAGACAGC------CGAGTG---------GAGG--CGCAACGGTC
3770    ------GCCCTCGACCGCCCGCGCAGGGCCACCCC---------CTAG---TACGCATA------CGCCGGCACGCCCGAGACAGC------CCAGTC---------GGGG--CCCACCGGTC
39347   ------GCCCTCGACTGTCCGCGCAGGTCCACCCC---------CTAG---CATGCATA------CGCCCGCACATTCGACCCAGG------CGAGTG---------GTGA--AGCAGCGGTC
443381  ------GCCCTCGACCGCCCGCGCAGGTCCATCCC---------CTAG---CACGCATA------CGCCCGCACAACCGACACAGG------CGAGTG---------GAGA--CGCAACGGTC
39388   ------GCCCTCGACCGCCCGCGCAGGTCCACCCC---------CTAG---CACGCATA------CGCCCGCACAAGCGACACAGA------CGAGTG---------GTGA--CGCAACGGTC
4151    ------GCCCTCGCCGGGCCGCGCAAGTCCACCCC---------CTAG---AACGCATA------ATCCCGCATAAACGATACAGG------CCAGTG---------GAGA--GGCAACGGTC
197815  ------GCCCTCGGCCGACCGCGGAGGTCGGCCCC---------CTAG---CACGCATA------TGCCCGCATGGCCGAGACAGG------CAAGCT---------GGGA--CGCAACGATC
201512  ------GCCCTCGACCGCCCGCGCAGGGCCACCCC---------CTAG---CACGCATA------CTCCCGCACATACGAGACAGC------CGAGTC---------GTGG--CGCAACGTTC
39353   ------GCCCTCGACCGCCCGCGCAGGTCCACACC---------CTAG---CACGCATA------CCCCCGCACAACCGAGACAGC------CCAGTG---------GAGG--CGCAACGCTC
28536   ------GCCCTCGACCGTCCGCGCAGGACCACCCC---------CTAG---CACGCATA------CCCCCGCACAACCGATACAGC------CCAGTG---------GCGA--TGCAGCGGTC
49121   ------GCCCTCGACCGACCGCGCAGGACCTCCCC---------CTAG---CACGCATA------CCCCCGCACACCCGAGACAGC------CGAGTG---------GGGG--CGCAGCGGTC
```

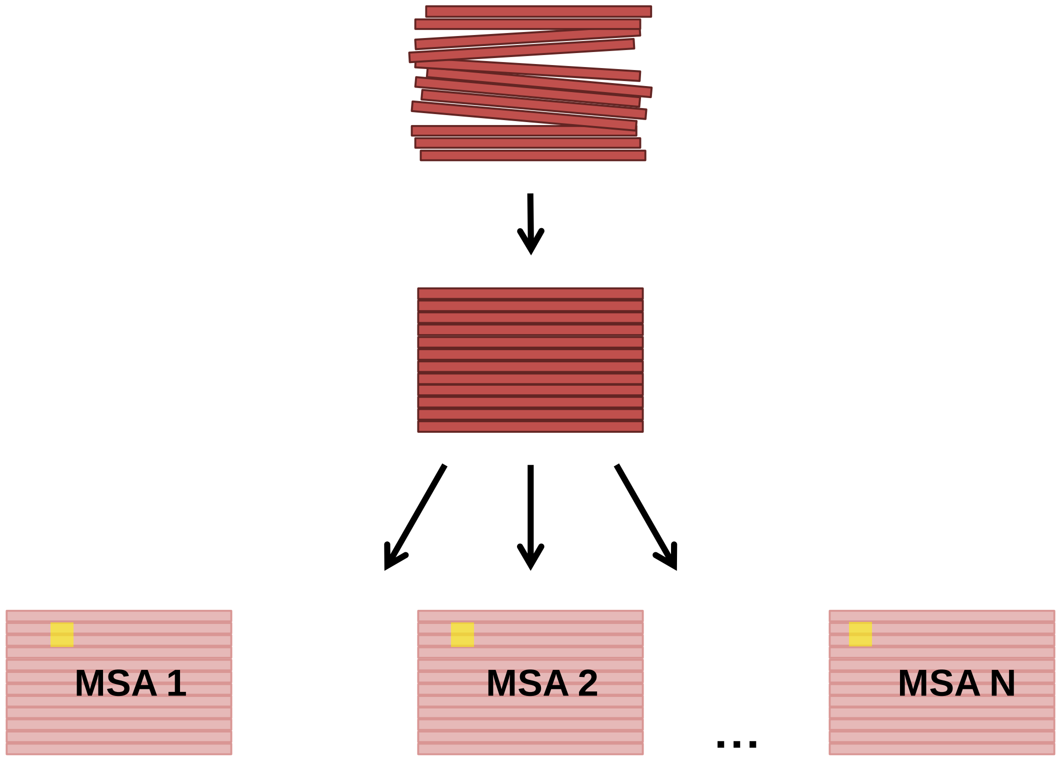
ITS gene

GUIDANCE as a measure of aligning difficulty



Column score: 2/3

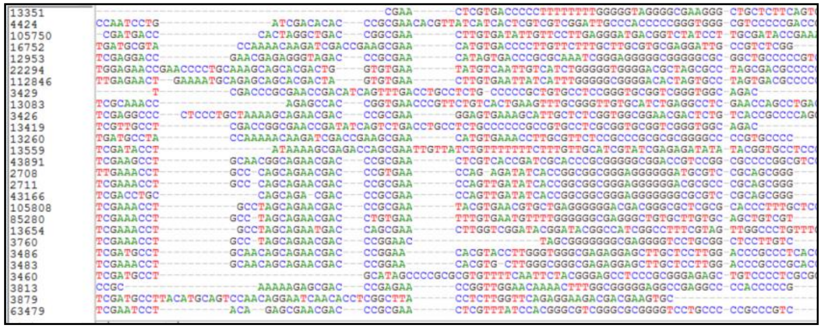
GUIDANCE as a measure of aligning difficulty



Residue score: 3/3

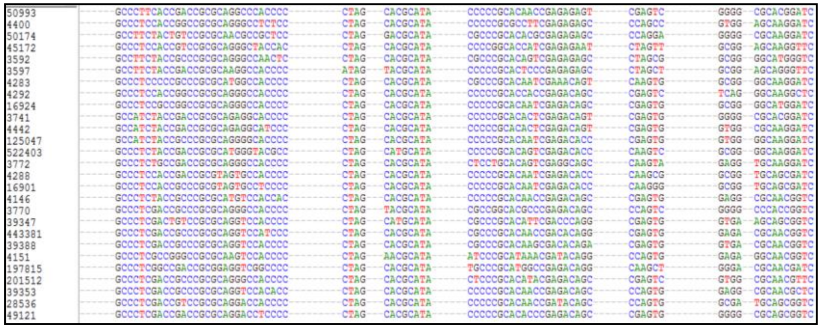
GUIDANCE as a measure of aligning difficulty

Real sequences



Gene	Residues score	Column score
18S	0.99955	0.854151
26S	0.999351	0.742893
atpB	0.999953	0.947041
matK	0.999471	0.959492
ITS	0.893573	0.050596

Simulated sequences

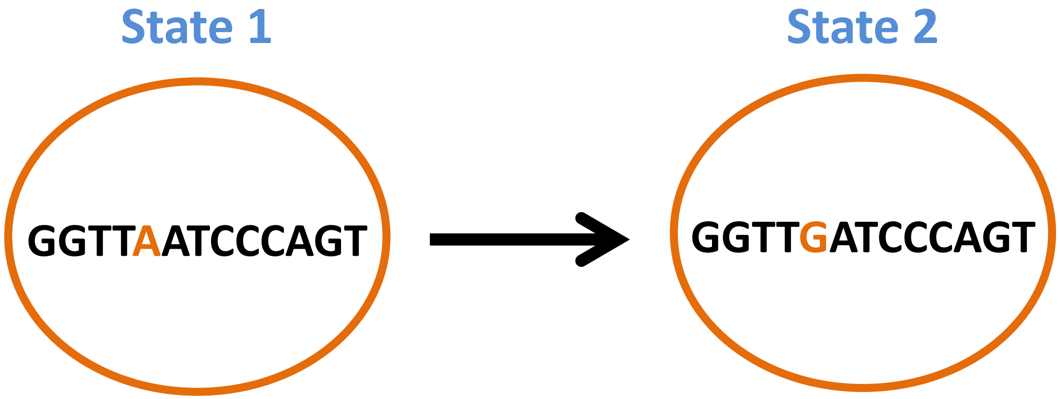


Gene	Residues score	Column score
18S	0.987769	0.001852
26S	0.862116	0.000057
atpB	0.995953	0.350238
matK	0.997872	0.075541
ITS	0.492671	0.00093

INDELible – a deeper insight

Markov chain

State of the chain = whole sequence



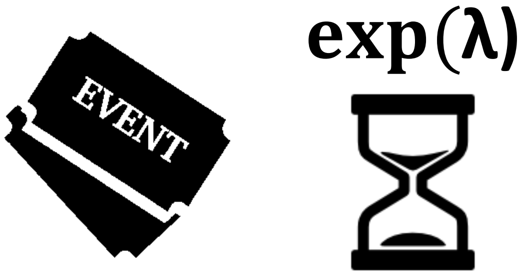
State 1

GGTTAATCCCAGT

State 2

GGTTGATCCCAGT

Simulating Markov chains using Gillespie's algorithm

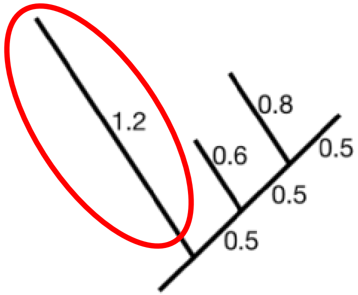


$\exp(\lambda)$

Mol. Biol. Evol. 2009

INDELible – a deeper insight

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/λ

INDELible – a deeper insight

$s1 > t$

no event

$s1 < t$

draw an event with probabilities

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

draw a location

indel

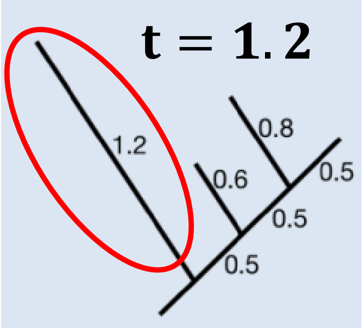
uniformly (L+1 positions)

substitution

proportional to the substitution rate at each site

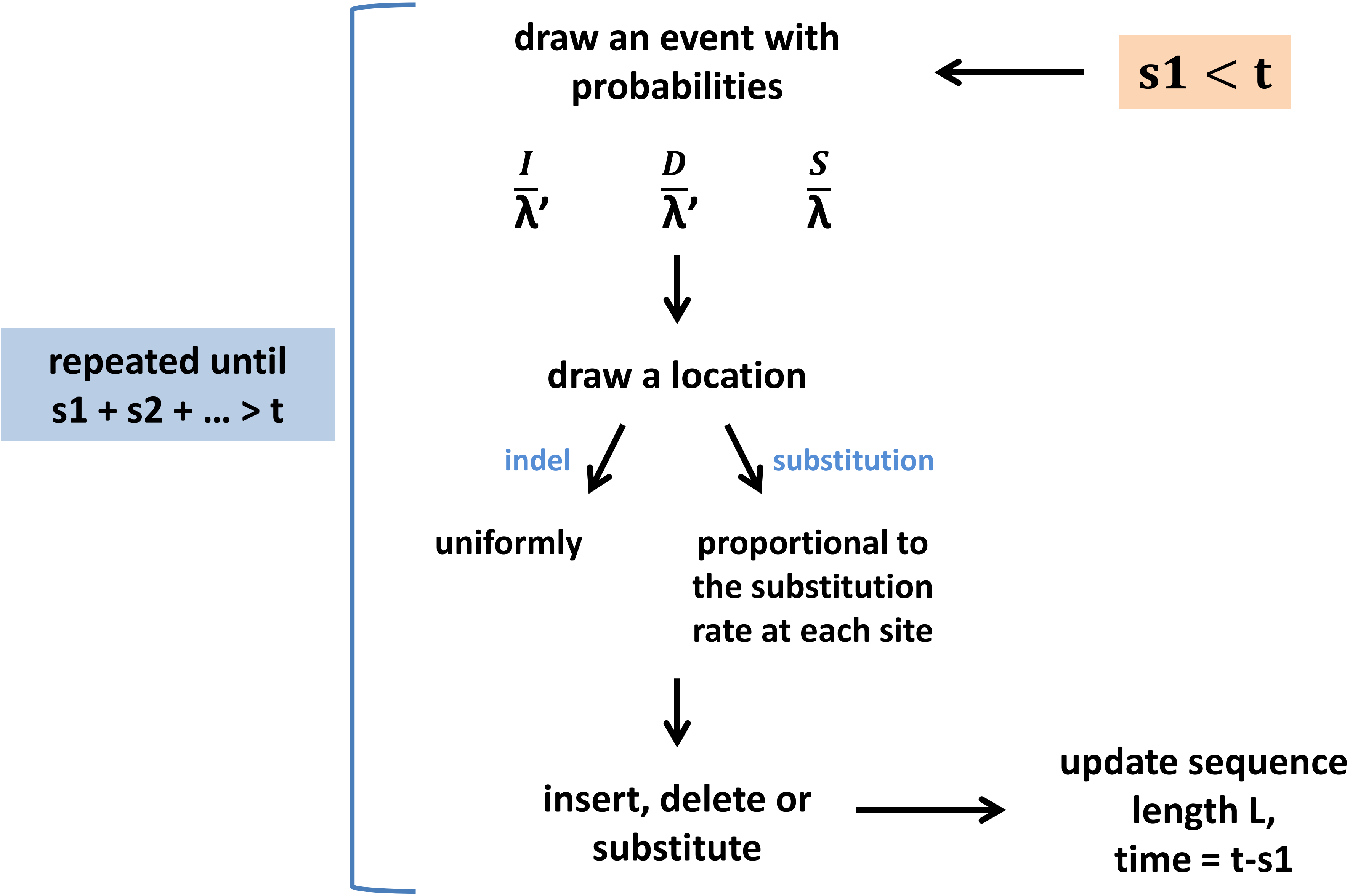
insert, delete or substitute

update sequence length L, time = t-s1



λ = I + D + S

INDELible – a deeper insight



draw an event with probabilities

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

s1 < t

draw a location

indel

uniformly

substitution

proportional to the substitution rate at each site

insert, delete or substitute

update sequence length L, time = t-s1

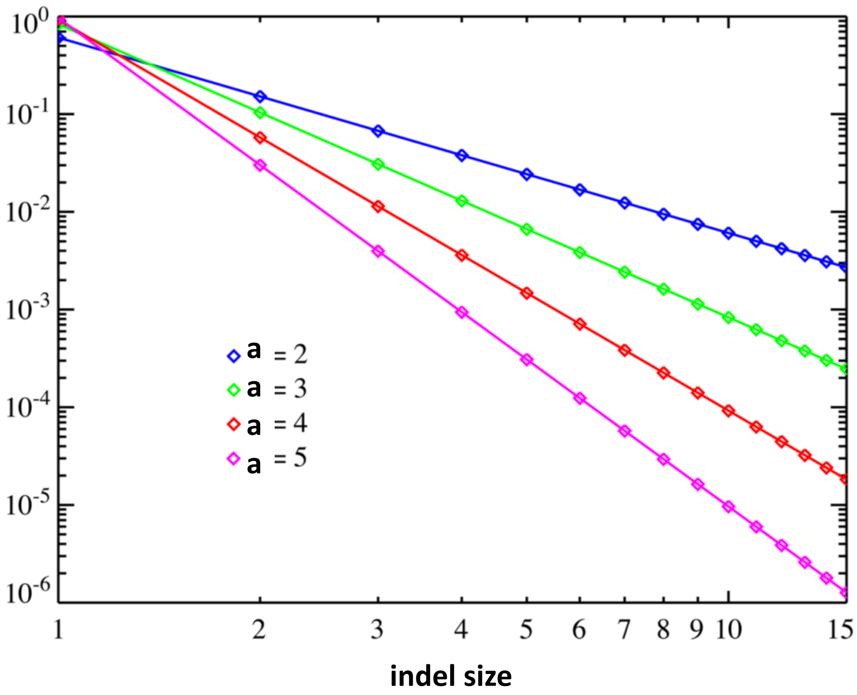
repeated until s1 + s2 + … > t

INDELible – a deeper insight

Insertion-deletion model

ir **indel rate** $ir \in [0,1]$

a **indel size** $a > 1$



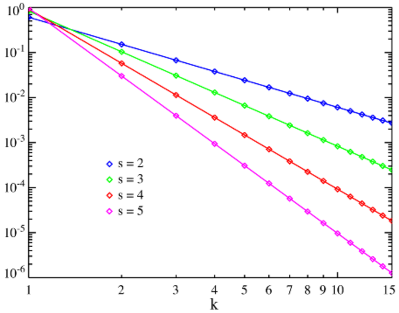
Zeta distribution

INDELible – a deeper insight

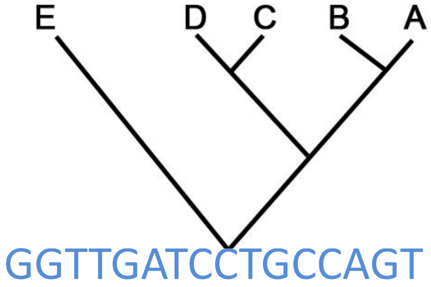
Insertion-deletion model

ir **indel rate** $ir \in [0,1]$

a **indel size** $a > 1$



rl **root length**



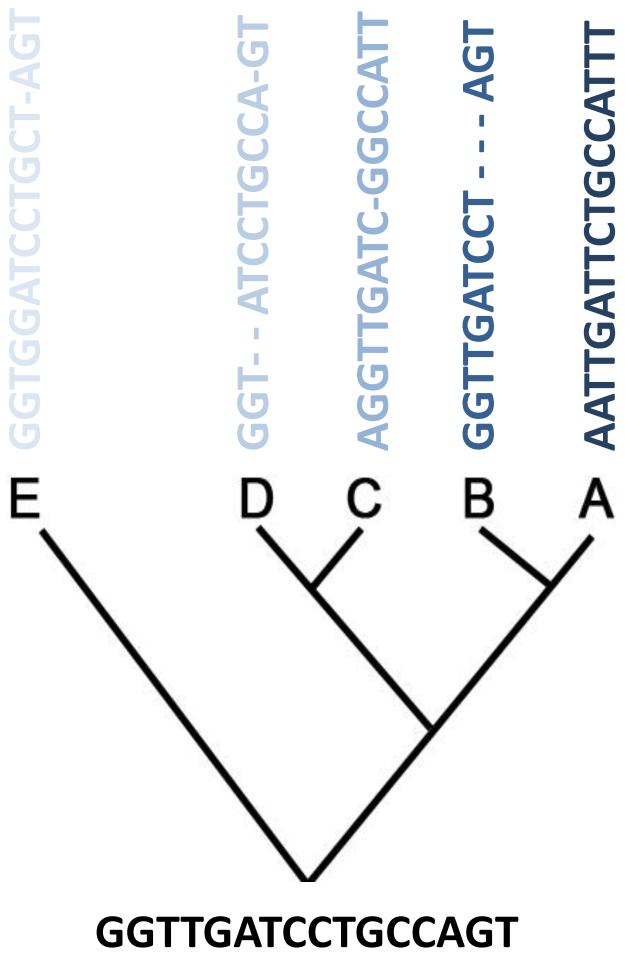
INDELible – a deeper insight

Insertion-deletion model

ir	indel rate
a	indel size
rl	root length

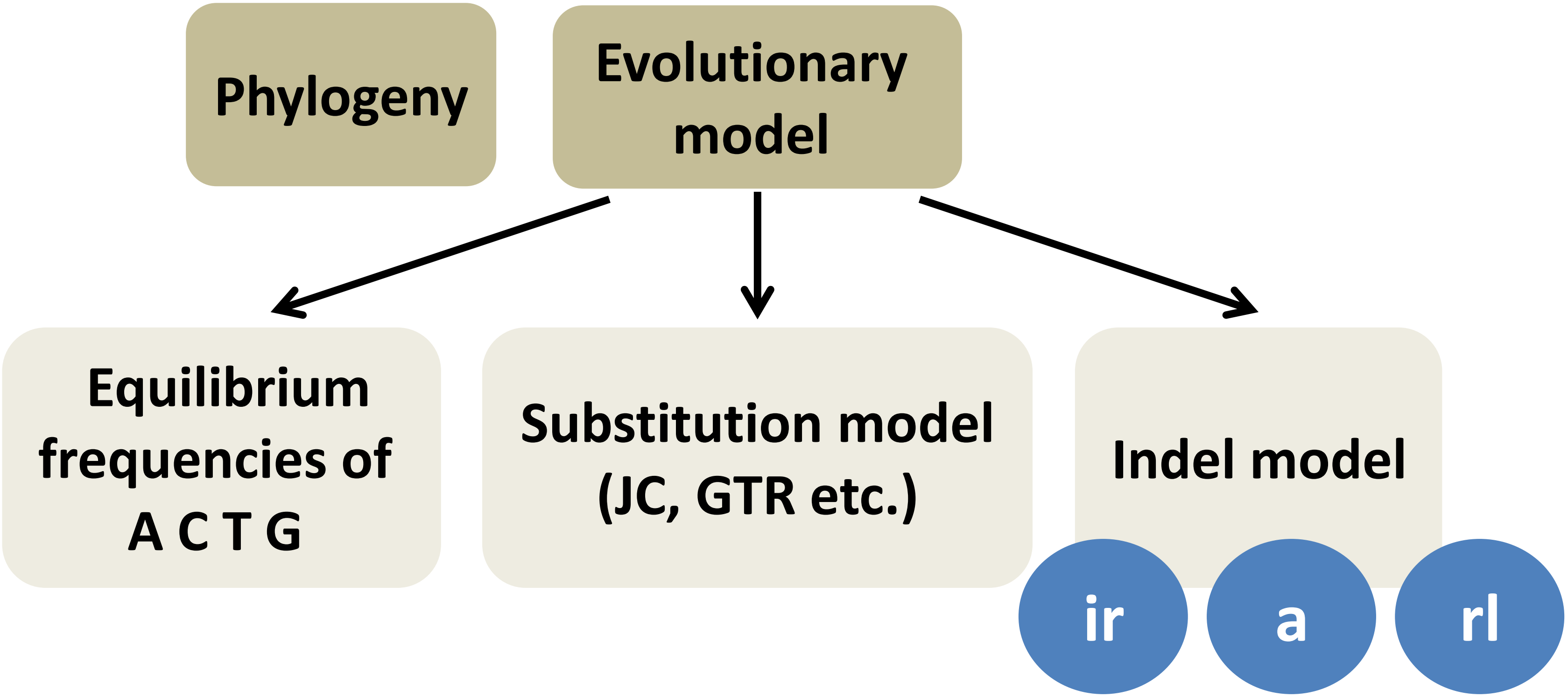
INDELible – a deeper insight

Gillespie algorithm



True alignment

INDELible – estimating parameters



Phylogeny

Evolutionary model

Equilibrium frequencies of A C T G

Substitution model (JC, GTR etc.)

Indel model

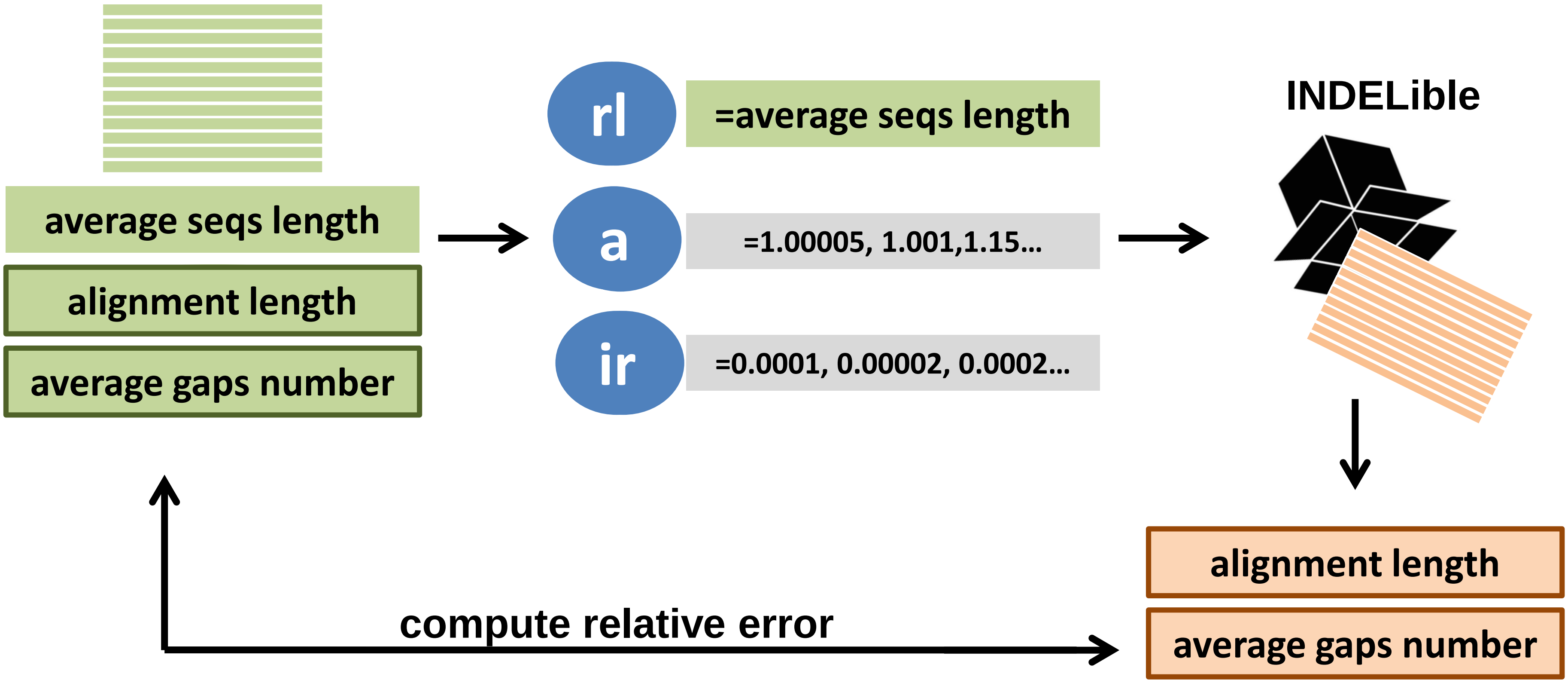
ir

a

rl

How to estimate?

INDELible – estimating parameters



average seqs length

alignment length

average gaps number

rl =average seqs length

a =1.00005, 1.001,1.15…

ir =0.0001, 0.00002, 0.0002…

INDELible

alignment length

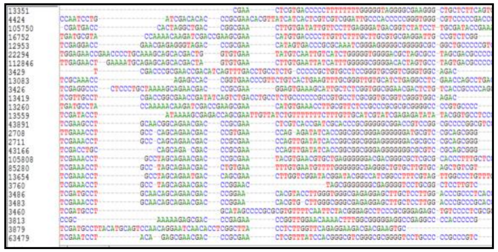
average gaps number

compute relative error

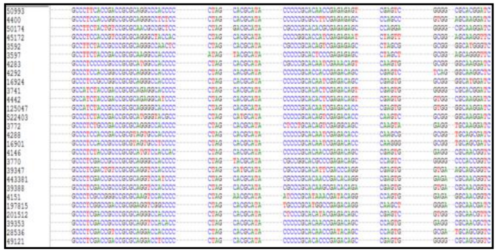
Choose the parameters a, ir which gave the lowest error

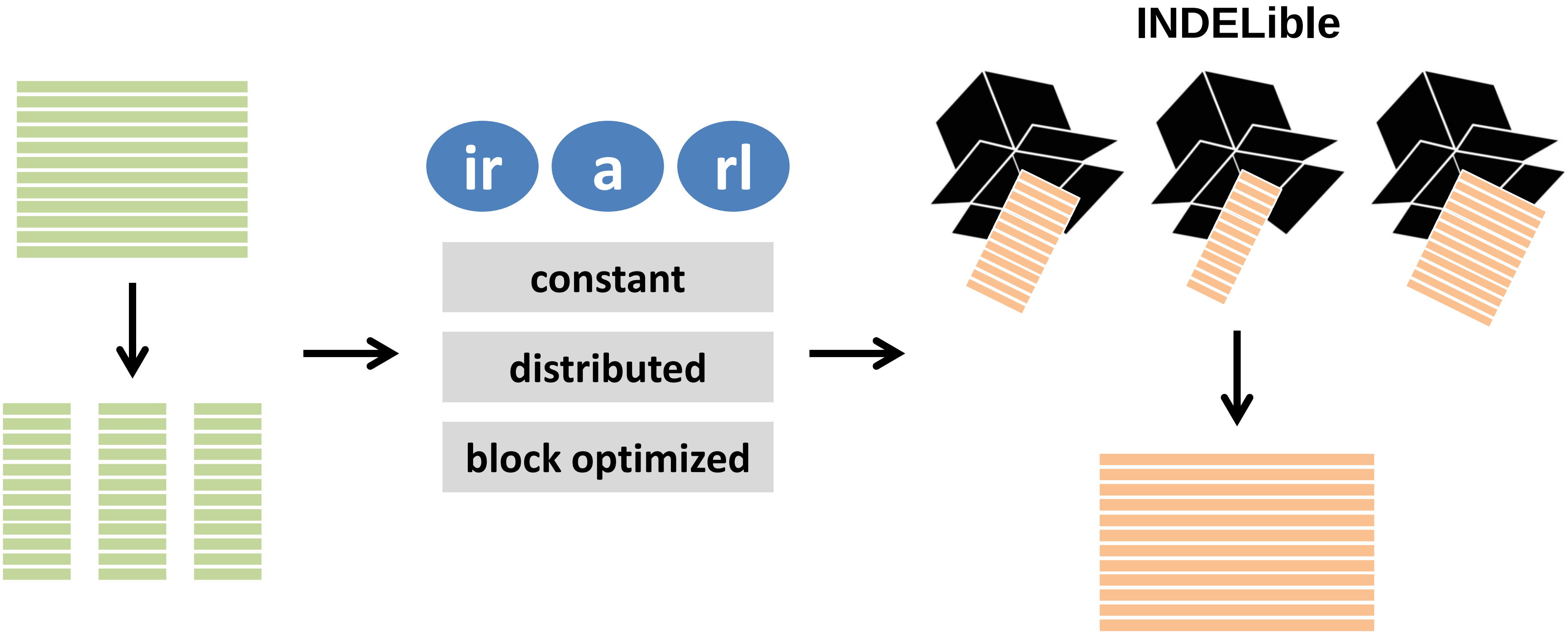
Getting more realistic simulations

Real sequences



Simulated sequences





INDELible

ir

a

rl

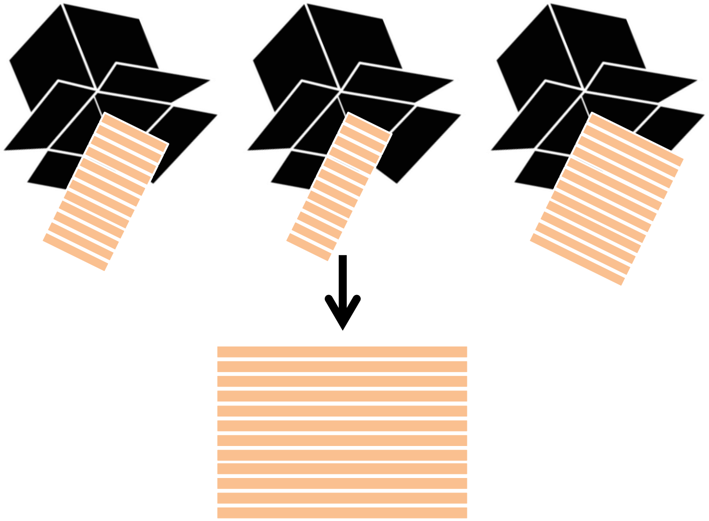
constant

distributed

block optimized

Getting more realistic simulations

INDELible



	constant	distributed	block optimized
column score (GUIDANCE)	0.97901696	0.97916526	0.97672856
	0.91552168	0.94023076	0.91669352
	0.9883448	0.98878618	0.98792382
	0.99055344	0.99287678	0.98774598
	0.94787124	0.9599763	0.9362862
	0.76490788	0.79192566	0.74172564
	0.51005486	0.52628762	0.60550184

Changing route?