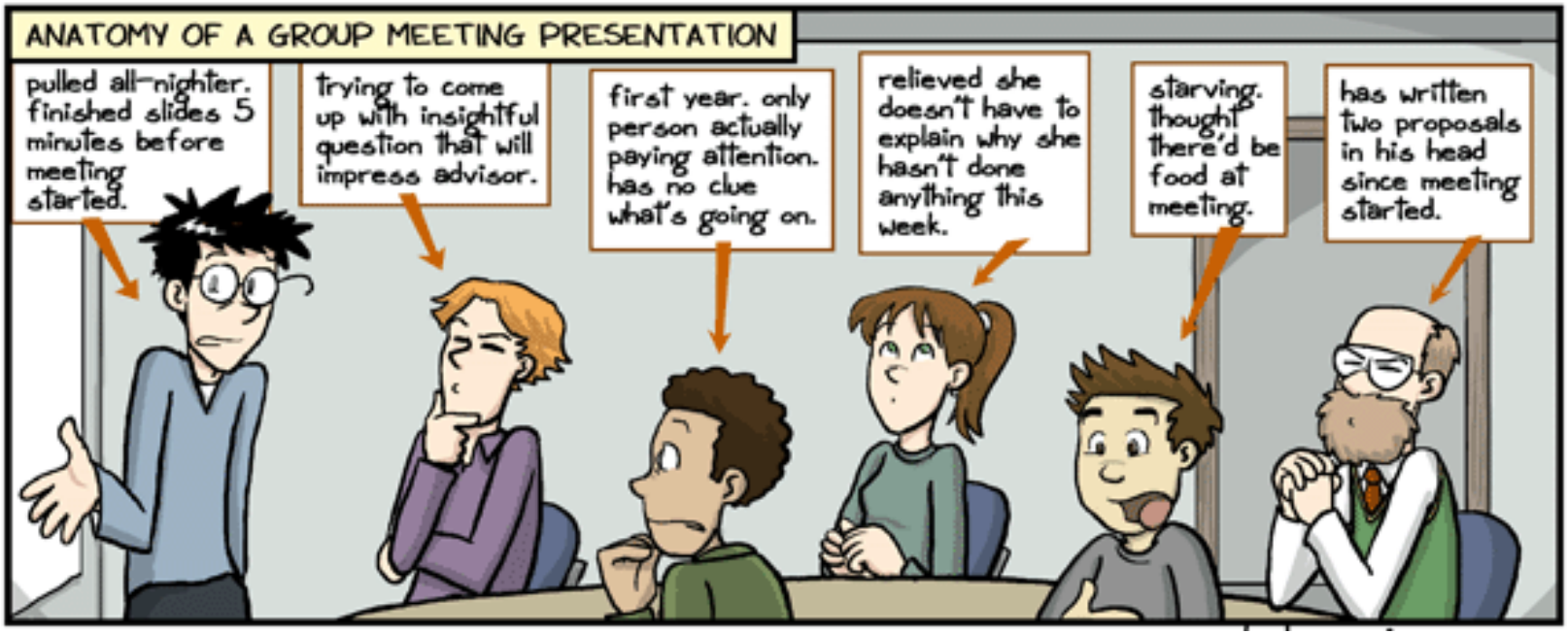
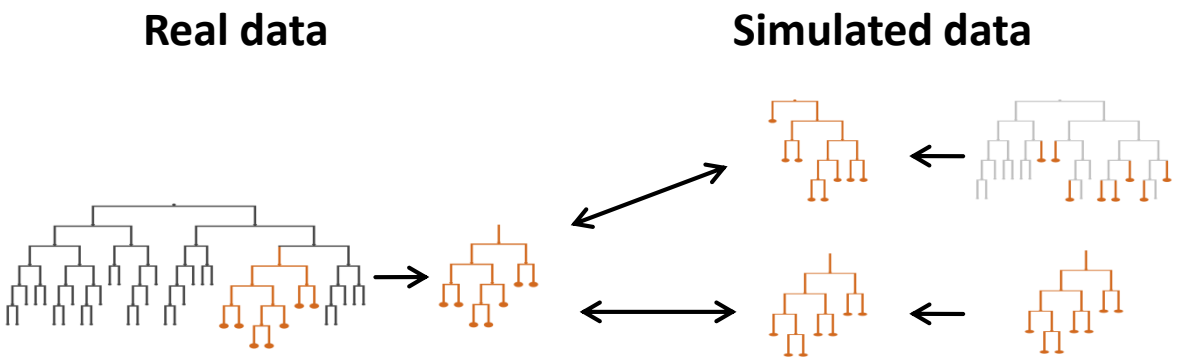
Lab meeting 07.12.17

Nomi



A little bit “history”

1

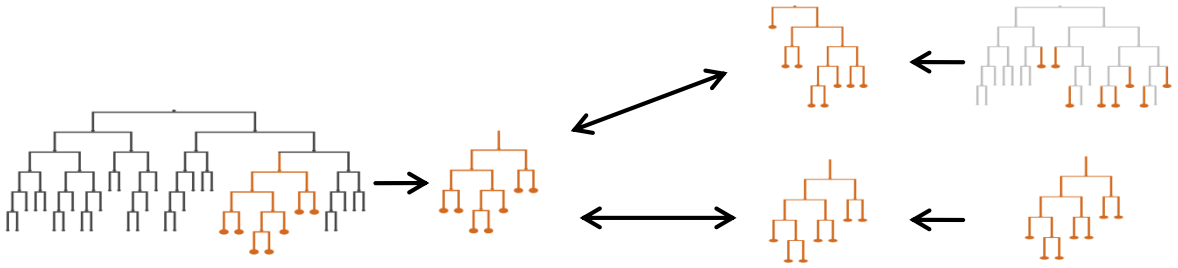


A little bit “history”

1

Real data

Simulated data

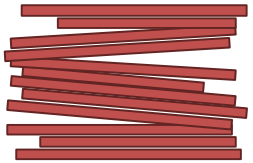


2

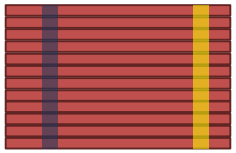
Are simulations realistic or “too easy”?

GUIDANCE as a measure of aligning difficulty

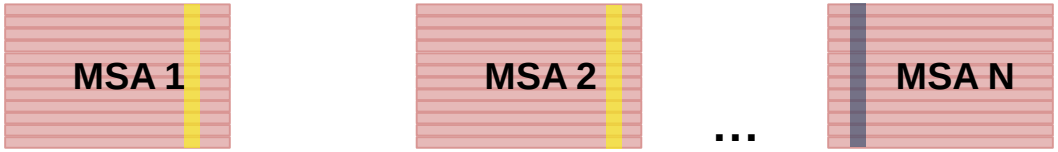
Input:
sequences



Reference
alignment



Alternative
alignments



Column score: 2/3, Column score: 1/3

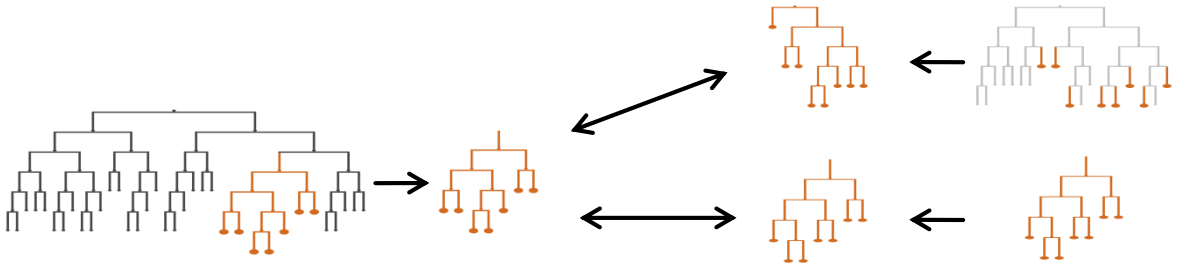
Total columns score: (2/3 + 1/3)/2 = 1.5

A little bit “history”

1

Real data

Simulated data



2

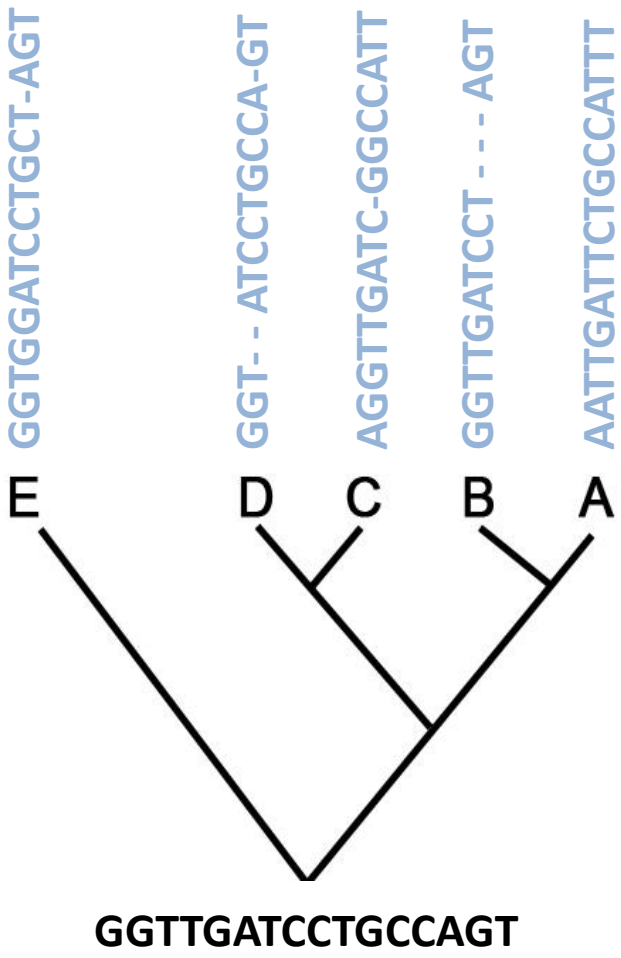
Are simulations realistic or “too easy”?

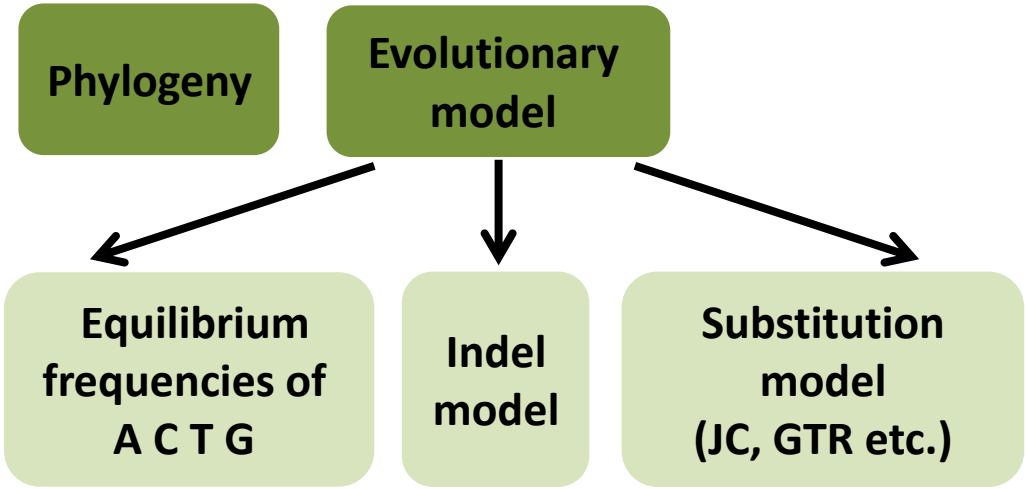
3

Explore simulations.

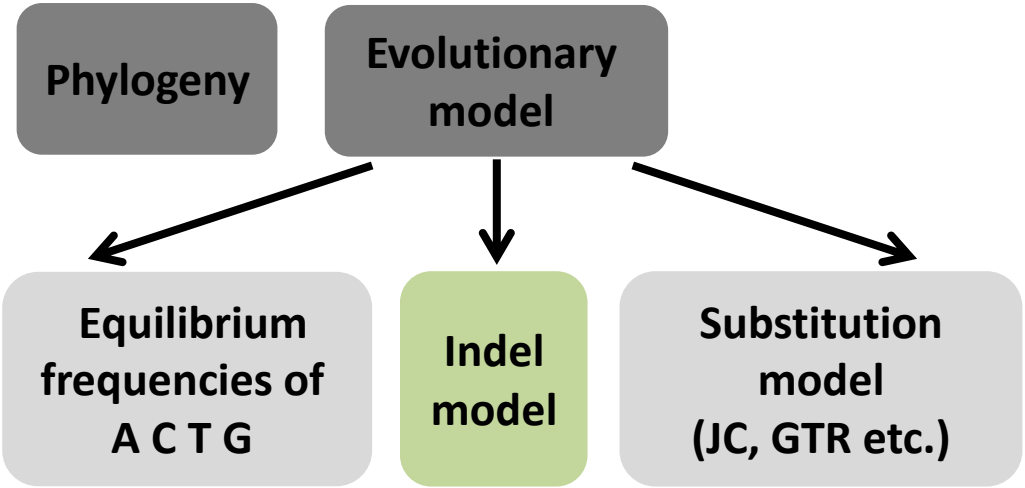
	GUIDANCE2 columns score	
marker	Real data	Simulated data
18S	0.878	0.966
26S	0.049	0.853
ITS	0.132	0.670
atpB	0.956	0.986
matK	0.627	0.984
rbcL	0.990	0.921
trnLtrnF	0.142	0.457

Simulations with INDELible



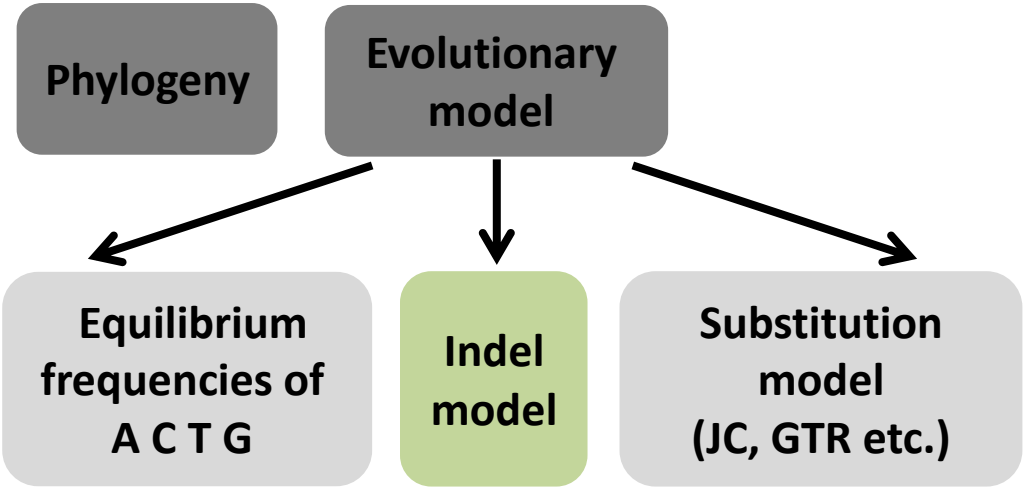


Simulations with INDELible

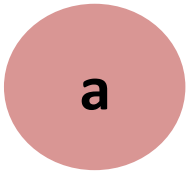


- **ir** — indel rate
- **a** — indel size
- **rl** — root length

Simulations with INDELible



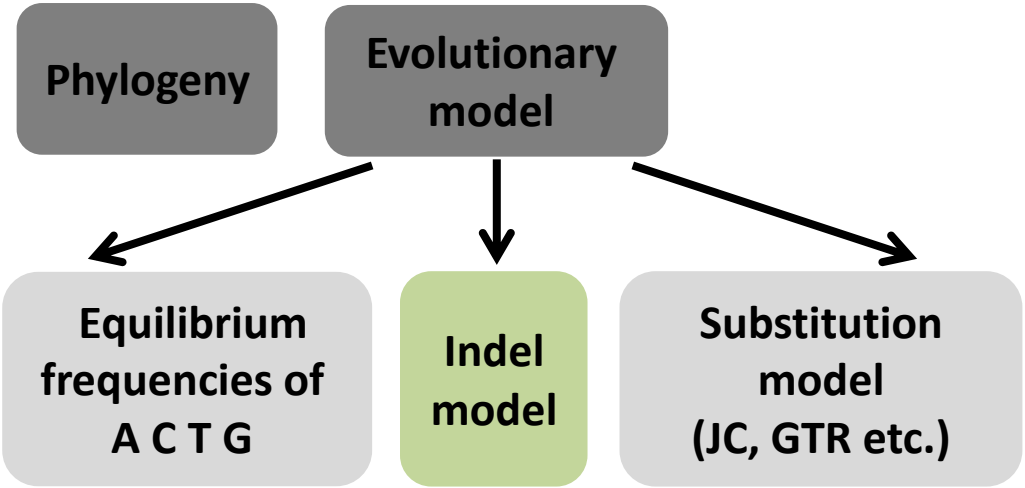
indel rate



indel size

root length

Simulations with INDELible



ir — indel rate

a — indel size – what is the distribution?

rl — root length

The Zipfian distribution

$$P(k) = \frac{k^{-a}}{\zeta(a)}$$

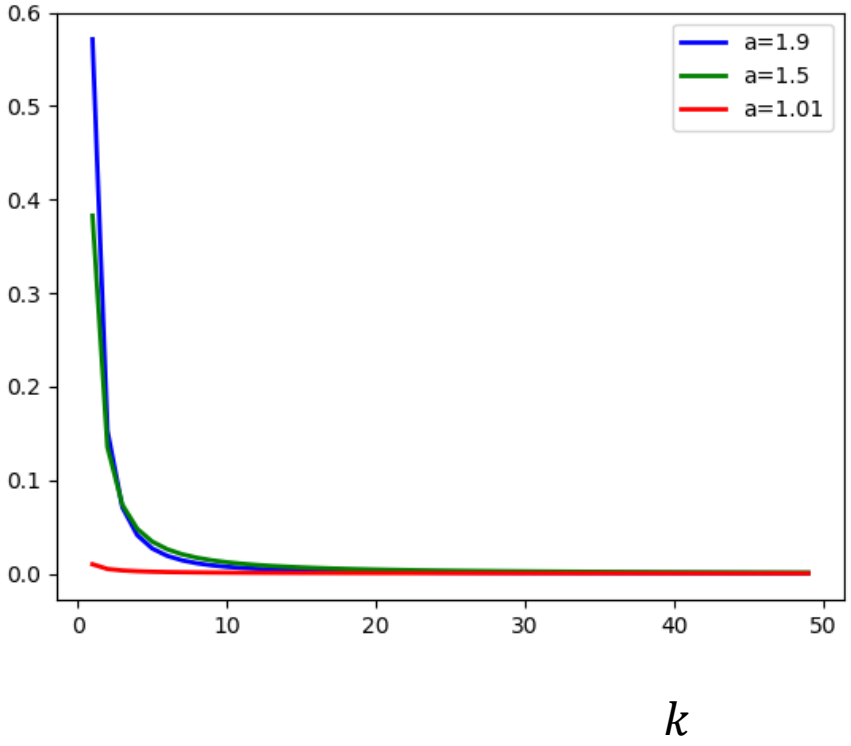
$k \in \{1, 2, ..\}$

$\zeta(a) = \sum_{n=1}^{\infty} \frac{1}{n^a}$

$$P(2) = \frac{2^{-1.9}}{\zeta(1.9)} = 0.57$$

$$P(5) = \frac{5^{-1.9}}{\zeta(1.9)} = 0.02$$

$$P(10) = \frac{2^{-1.9}}{\zeta(1.9)} = 0.007$$



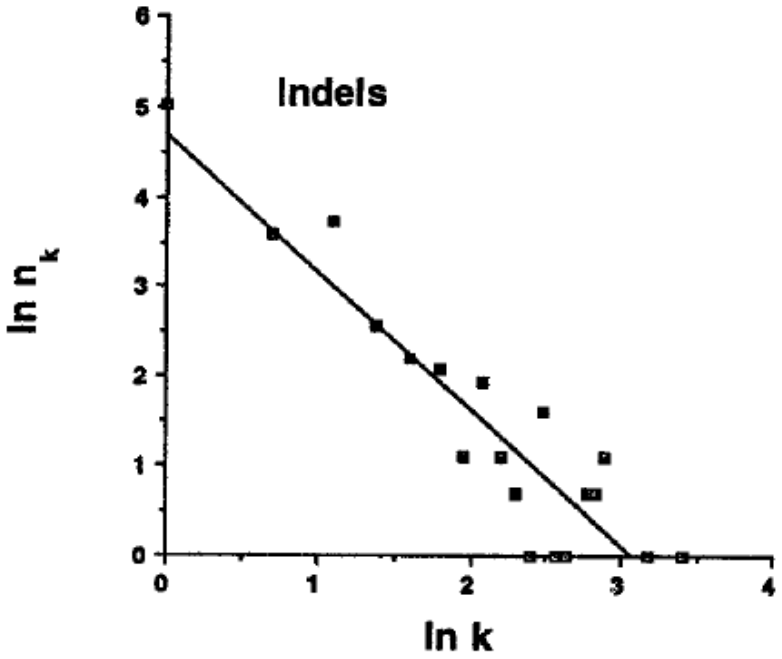
k ↑ probability ↓

Fit of Zipfian distribution

Table 4

Fit of a Zipfian distribution to data from aligned homologous sequences

Gap length	Number of occurrences	Approximation	Cumulative number of occurrences	Approximation
1	796	767·7	796	767·7
2	313	389·6	1109	1157·3
3	231	218·4	1340	1375·7
4	162	136·2	1502	1511·9
5	83	92·2	1585	1604·1
6	85	66·2	1670	1670·2
7	53	49·7	1723	1720·0
8	54	38·7	1777	1758·6
9	31	30·9	1808	1789·5
10	40	25·2	1848	1814·7
11	29	21·0	1877	1835·7
12	13	17·7	1890	1853·5
13	19	15·2	1909	1868·6
14	18	13·1	1927	1881·8
15	20	11·5	1947	1893·2
16	12	10·1	1959	1903·4
17	13	9·0	1972	1912·3
18	14	8·0	1986	1920·4
19	9	7·2	1995	1927·6
20	4	6·5	1999	1934·1
21	5	5·9	2004	1940·0
22	6	5·4	2010	1945·4
23	7	4·9	2017	1950·4
24	2	4·5	2019	1954·9
25	3	4·2	2022	1959·1
26	3	3·9	2025	1963·0
27	3	3·6	2028	1966·6
30	1	2·9	2029	1976·0
31	2	2·7	2031	1978·8
32	3	2·6	2034	1981·3
33	5	2·4	2039	1983·8
35	2	2·2	2041	1988·2
36	5	2·0	2046	1990·2
39	2	1·7	2048	1995·7
40	1	1·7	2049	1997·4
48	1	1·1	2050	2008·1
49	1	1·1	2051	2009·3
50	1	1·1	2052	2010·3
61	1	0·7	2053	2019·7
63	1	0·7	2054	2021·1
86	1	0·4	2055	2032·2
90	1	0·3	2056	2033·5
91	1	0·3	2057	2033·9
103	1	0·3	2058	2037·2
104	1	0·2	2059	2037·5
135	1	0·1	2060	2043·3
138	1	0·1	2061	2043·7
146	1	0·1	2062	2044·8
251	1	0·0	2063	2052·3



Benner et al. 1993, Xun Gu et al 1995

Distribution of indel length

Do indel lengths come from the Zipfian distribution?

Observed data: gap lengths

E.g: 6, 2, 1, 2, 1, …

```
A------A
A--TT-CA
A--TTGCA
ACGTT-CA
ACGTTGCA
```

Assumption: Gap lengths are correlated to indel events

Unique indel events

```
GG---①---CCTT----③-    GG
GG---①---CCTT----
GGAAA--②---TT-④-A
GGAAA--②---TT-④-AC
GGAAACCCCCCTTCAAACCCCAAAAGG
```

Edges /no edges

Distribution of indel length

Do indel lengths come from the Zipfian distribution?

Observed data: gap lengths

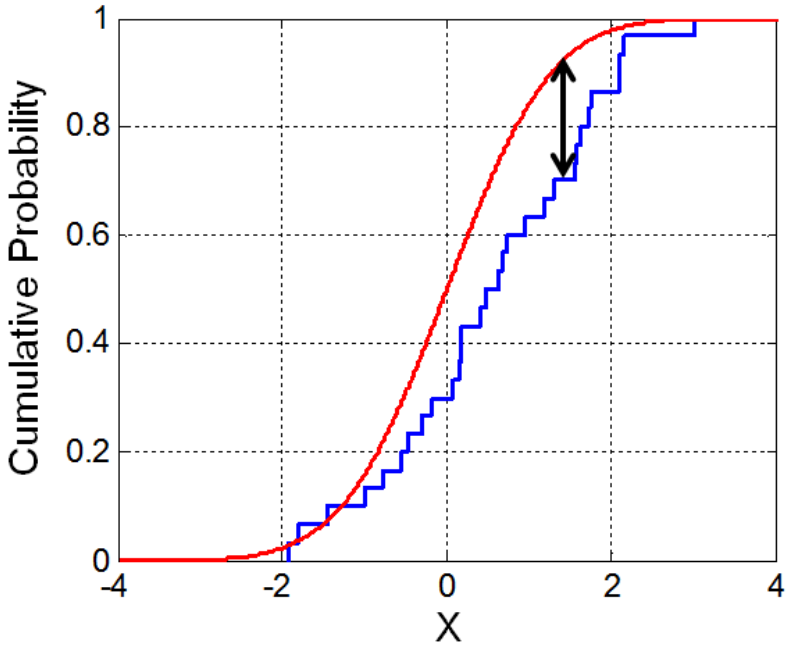
E.g: 6, 2, 1, 1, …

```
A------A
A--TT-CA
A--TTGCA
ACGTT-CA
ACGTTGCA
```

→

Kolmogorov–Smirnov test

Kolmogorov–Smirnov test



Goal:
Compare a sample with a reference probability distribution.

Null hypothesis:
Sample comes from the reference distribution

If *p-value* < 0.05,
sample does NOT come from the reference distribution

Distribution of indel length

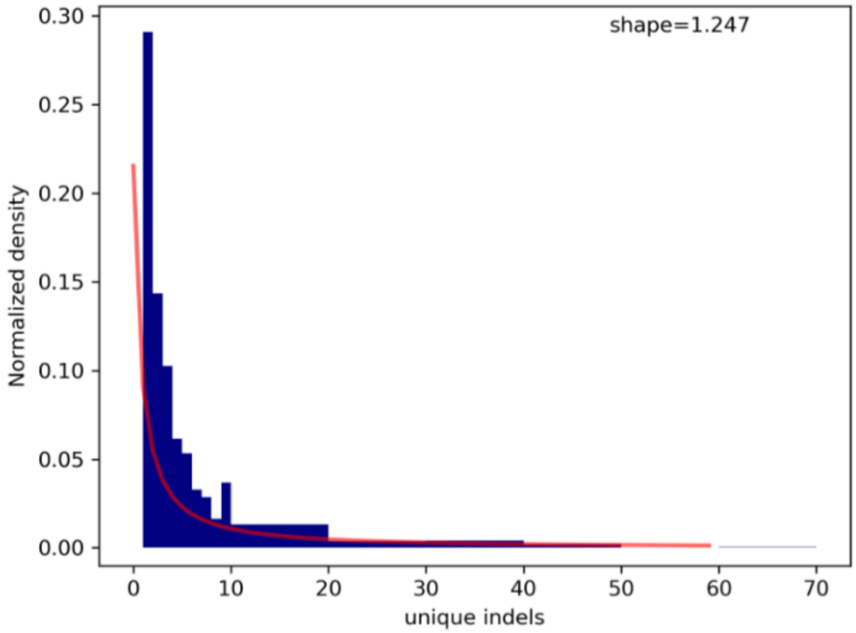
Kolmogorov–Smirnov test

Observed data:
gap lengths

E.g.: 1, 6, 22, 3, 1, 34, …

Zipfian distribution

$$P(k) = \frac{k^{-a}}{\zeta(a)}$$



Distribution of indel length

Kolmogorov–Smirnov test

Observed data:
gap lengths

E.g.: 1, 6, 22, 3, 1, 34, …

Zipfian distribution

$$P(k) = \frac{k^{-a}}{\zeta(a)}$$

How to estimate shape parameter?

1) Maximum likelihood
2) spartaABC
3) Try all possible values

Distribution of indel length

Kolmogorov–Smirnov test

Do gaps lengths come from the Zipfian distribution?

Data: ~100 genes
Observations:
gaps lengths for each gene

GeneA: 1, 6, 22, 3, 1, 34, …
GeneB: 2, 13, 2, 2, 5, 4, …

Results:
> 4 genes come from the Zipfian distribution

(for all methods of estimating shape parameter, and include / not include edges)

But…

test Kolmogorov–Smirnov test

But…
do we know to say that observations come from the Zipfian distribution when they indeed come from it?

1. Simulate data with the Zipfian distribution as indel lengths model. For the shape parameter pick a random number.
2. Get the unique indels lengths from the simulated TRUE alignments.
3. Perform Kolmogorov–Smirnov test with the Zipfian distribution. Use the shape parameter from 1.

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

2,4,1,1,7,36,1,3…

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

test Kolmogorov–Smirnov test

But…
do we know to say that observations come from the Zipfian distribution when they indeed come from it?

Results:
Only 6 genes come from the Zipfian distribution

(< 25 genes with other shape parameters)

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

2,4,1,1,7,36,1,3…

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

test Kolmogorov–Smirnov test

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

2,4,1,1,7,36,1,3…

$$P(k) = \frac{k^{-1.5}}{\zeta(1.5)}$$

So…
maybe the way we define an observation is not good?

In simulations we do know the unique indel events

Simulate with observed indel lengths

GUIDANCE2 columns score			
marker	Real data	Simulated data with Zipfian model	Simulated data with observed indel lengths
18S	0.878	0.966	0.946
26S	0.049	0.853	0.545
ITS	0.132	0.670	0.426
atpB	0.956	0.986	0.945
matK	0.627	0.984	0.773
rbcL	0.990	0.921	0.926
trnLtrnF	0.142	0.457	0.272

Find yourself…