Estimation of Phylogenetic Trees

The data: A morphological character matrix

Morphological data matrix for Carnivora

Taxa	Character-state scoring											
	# 1	2	3	4	5	6	7	8	9	10	11	12
Lemur (outgroup)	0	0	0	0	0	0	0	0	0	0	0	0
Cat	0	1	0	1	0	0	1	1	1	0	0	0
Hyena	0	1	0	1	0	0	1	0	1	0	0	0
Civet	0	1	0	0	0	0	0	0	1	0	0	0
Dog	1	0	0	0	1	0	0	0	0	0	0	0
Raccoon	1	0	0	0	1	0	0	0	0	0	0	0
Bear	1	0	0	0	1	1	0	0	0	1	0	0
Otter	1	0	0	0	1	0	0	0	0	1	0	0
Seal	1	0	1	0	1	1	0	0	0	1	1	1
Walrus	1	0	1	0	1	1	0	0	0	1	1	1
Sea lion	1	0	1	0	1	1	0	0	0	1	0	0

All character states in outgroup are ancestral

This does NOT mean that the outgroup lacks derived traits!

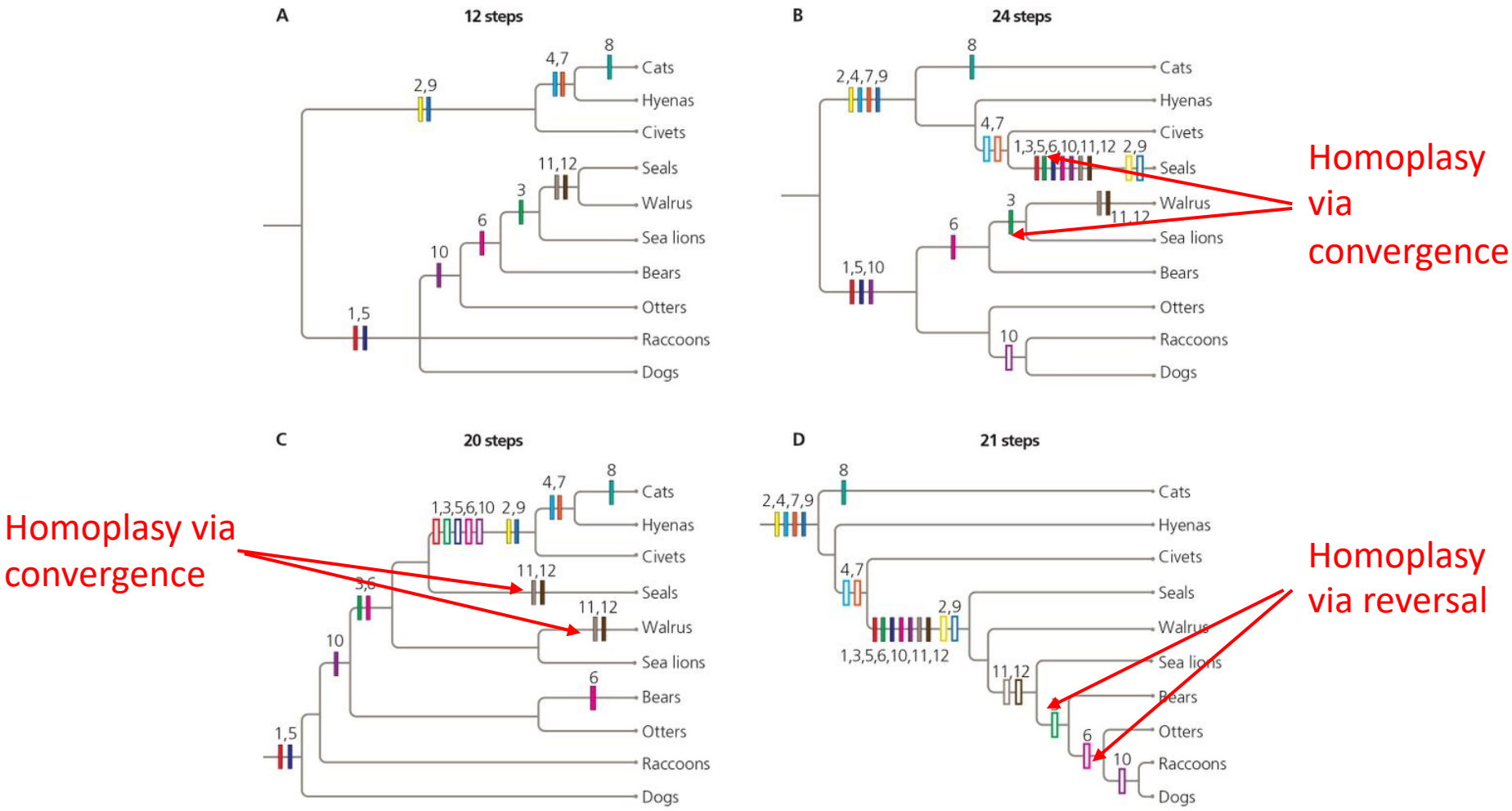
Lots of these…

1 1 1 1 0 0 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 1

Not informative because they are not synapomorphies in ingroup.

Outgroups help us identify **shared derived** states (**synapomorphies)**

How to choose a tree that best explains the data



Bars = synapomorphies (shared, derived traits)

Open bars = reversion to ancestral-like state

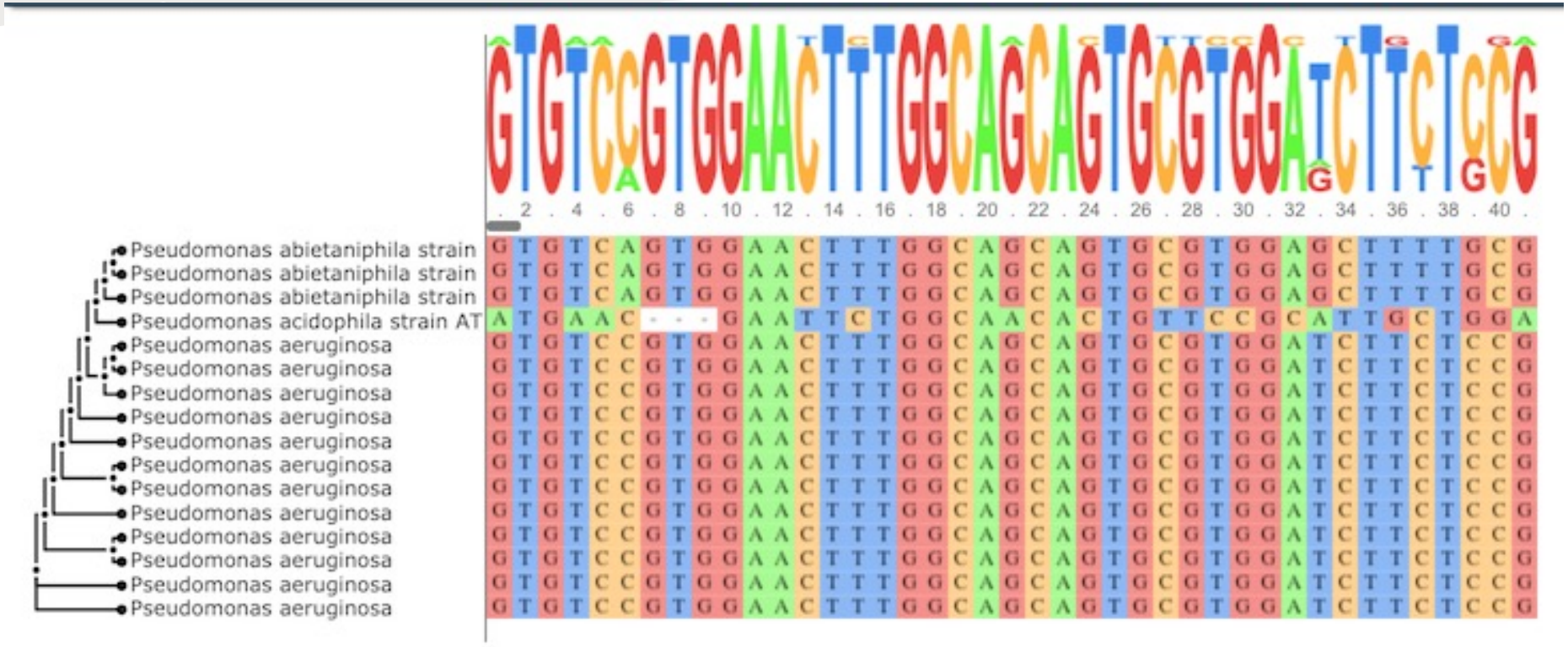
Parsimony analysis in practice…

TABLE 9.1 The Huge Number of Possible Tree Topologies

# of Taxa	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225

Even with computers, trees cannot be exhaustively searched for most analyses so these programs employ algorithms to efficiently search “tree space”

The modern era molecular phylogenetics

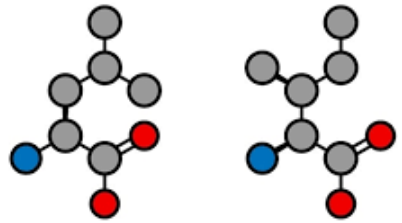


Multiple sequence alignment

- Multiple sequence alignment (MSA) is important for phylogenetic estimation or model-based inference of evolutionary processes
- The goal of MSA is to introduce gaps into sequences so that columns of an aligned matrix contain character states that are homologous
- Homology cannot be directly observed but can be inferred

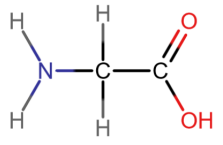
Inferring Homology

- MSA algorithm attempts to produce homologous alignments by scoring many plausible alignments and choosing one with the best score
- Aligning two positions that display the same nucleotide improves the score
- Aligning two positions that are not the same decrease the score

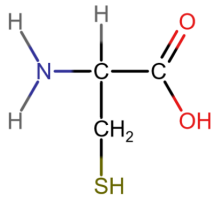


L-leucine (Leu, L)

L-isoleucine (Ile, I)



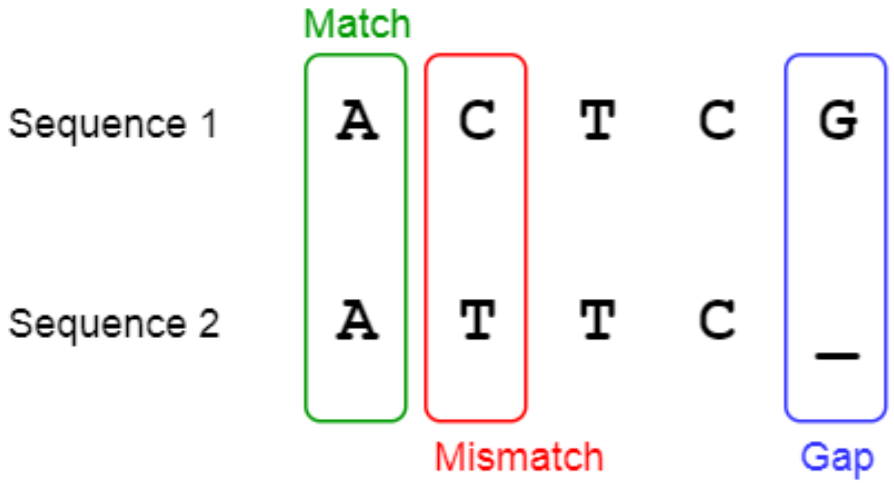
glycine

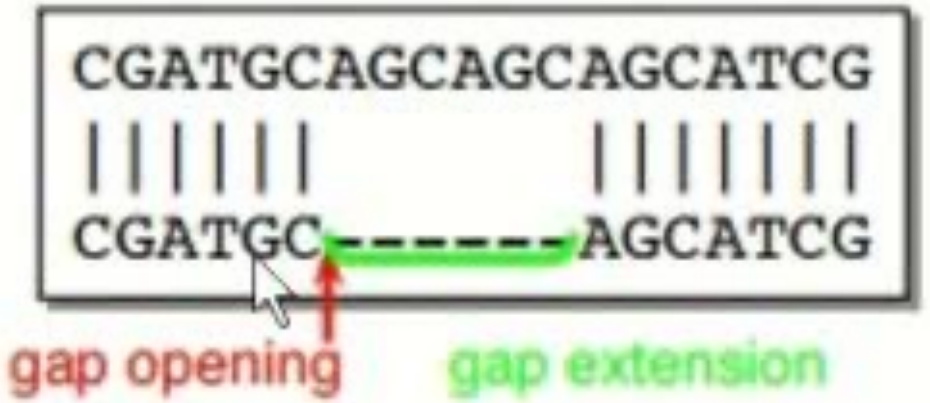


cysteine

Inferring Homology

- Placing gaps in a sequence is penalized too
- Introducing a new gap usually has a higher cost than extending an existing gap





Generic Alignment Scoring Parameters

match = +5
mismatch = -3
gap open = -5
gap extension = -2

Example 1

Alignment 1

```
AGTTCCCTG
AGTTA--TG
```

Alignment 2

```
AGTTCCCTG
AGTT-A-TG
```

Score

matches

6 x 5 = 30

− 3

− 5

− 2

20

= 17

Example 1

Alignment 1

```
AGTTCCCTG
AGTTA--TG
```

Alignment 2

```
AGTTCCCTG
AGTT-A-TG
```

Score

20

17

In example 1, the first alignment has a higher score for minimizing gap openings

Generic Alignment Scoring Parameters

match = +5
mismatch = -3
gap open = -5
gap extension = -2

Score

Generic Alignment Scoring Parameters

match = +5
mismatch = -3
gap open = -5
gap extension = -2

Example 2

Alignment 1

```
AGTTCCACTG
AGTTA---TG
```

= 18

Alignment 2

```
AGTTCCACTG
AGTT--A-TG
```

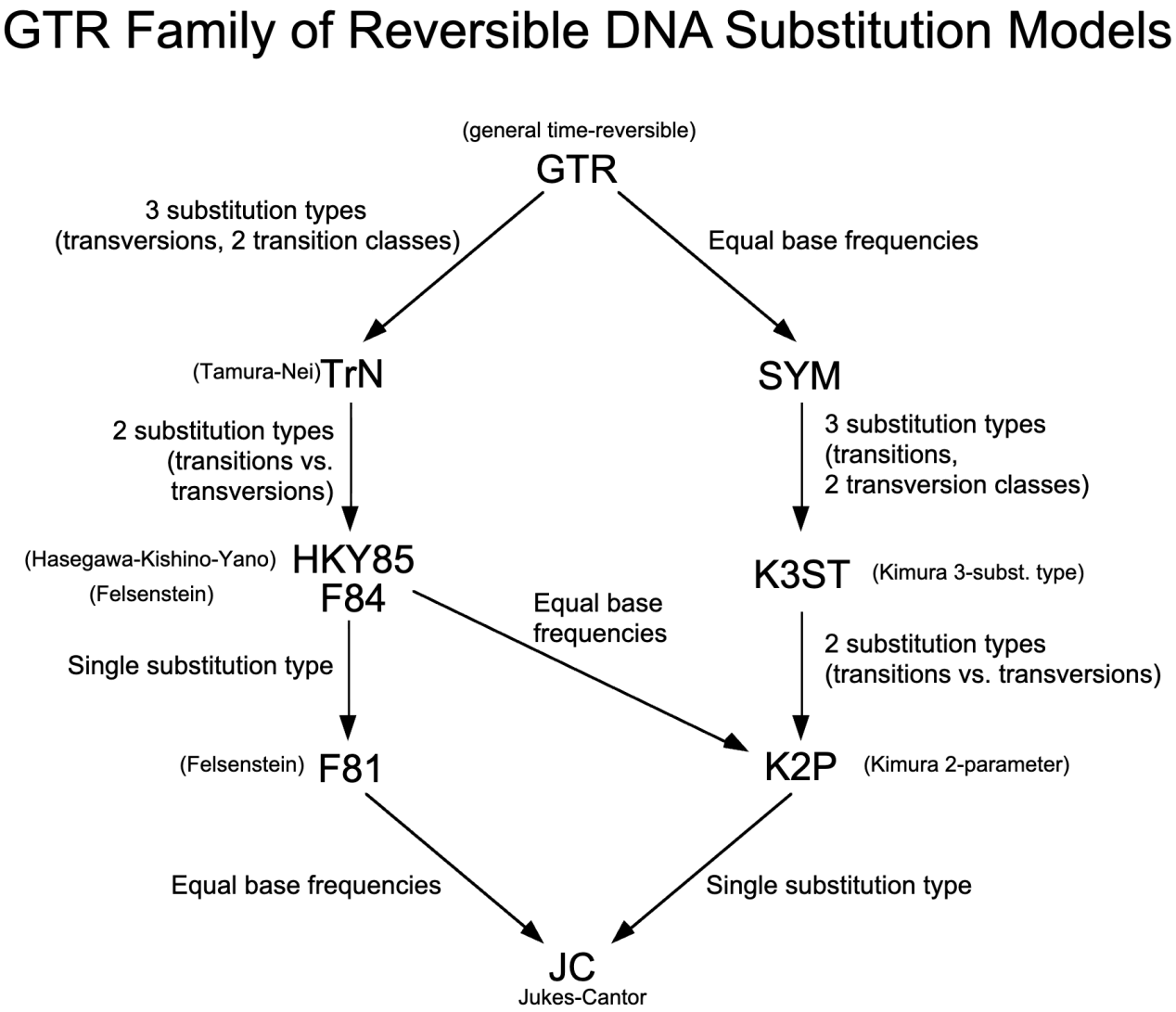
= 23

Alignment Software

After alignment: Inferring a phylogeny

1. Model for molecular evolution – How nucleotides change over time
2. Model for branch lengths – Time - branch lengths
   ↳ molecular clock
3. Search in the space of trees
   ↳ Different cladograms are going to have different "scores"

Models for molecular evolution



Continuous Time Markov ~~Models~~ Chains (CTMC)

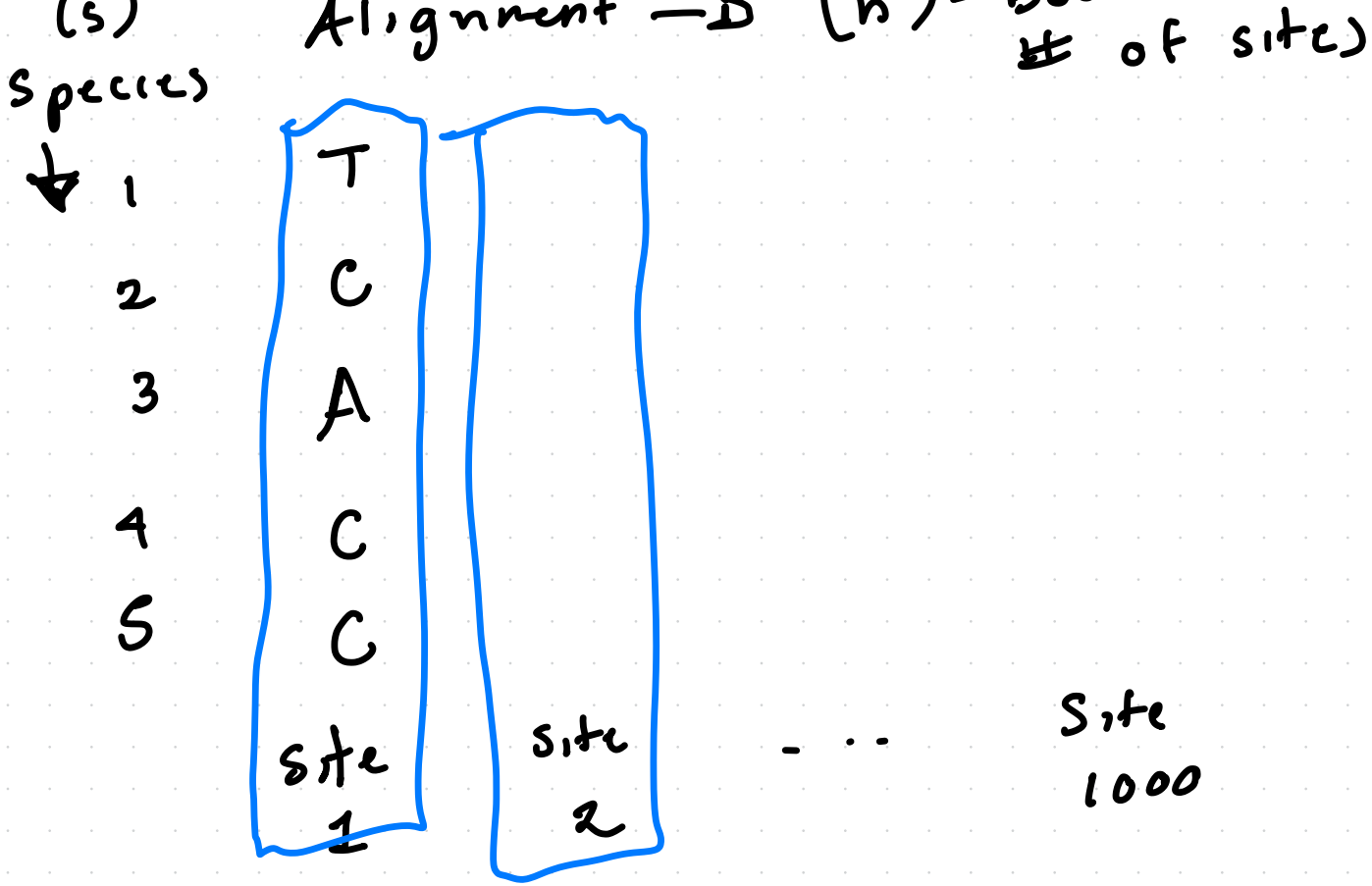
Models based on probability that follow a nucleotide over time

1. Notation
   - X -> a nucleotide A, C, G, T
   - $X(t)$ -> the value of a nucleotide at time t
   - t -> branch lenght

   E.g. $X(0.5) = A$

2. $X(t)$ is a random variable, this means it has a probability $X(t) \sim P$ (P: probability)

3. Markovian property
   "The future only depends on the present and not the past"

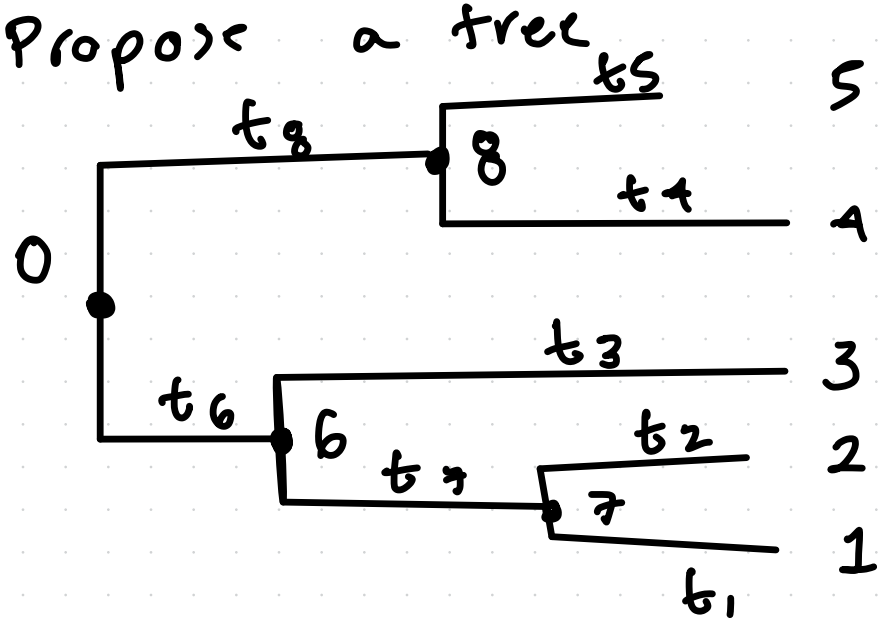


site 1 is independent from site 2

Watch out we are assuming each site is independent of each other

[because we aligned the sequences we are "ok" to proceed with caution]

1. Propose a tree



C
C
A
C
T
site 1

Probability of observing T C A C C given the tree and branches and assuming kimura 81

likelihood function

Likelihood function

Probability of observing sites (data) given
a model of evolution (nucleotide model +
a cladogram + branches (molecular clock)

$$P(D \mid M) = \mathcal{L}$$

data — given — model — likelihood

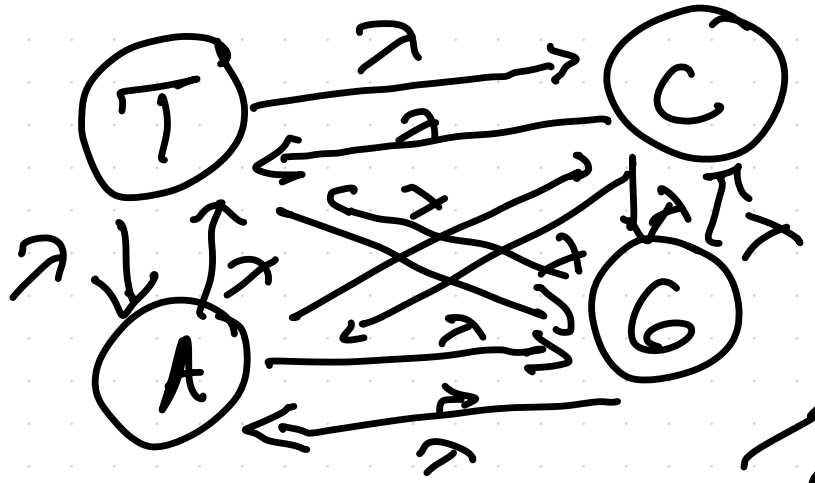
The pruning algorithm (example in your reading)

The pruning algorithm (example in your reading)

How do you mutate (point)
changing one letter A, C, G, T

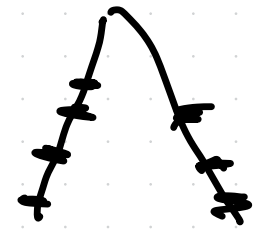
substitution

How quickly these substitutions happen?



λ (lambda) = rate = $\frac{\text{change DNA}}{\text{time}}$

Math representation is a matrix



↓ →	T	C	A	G
T	-	λ	λ	λ
C	λ	-	λ	λ
A	λ	λ	-	λ
G	λ	λ	λ	-

= Q-matrix

JC (69)