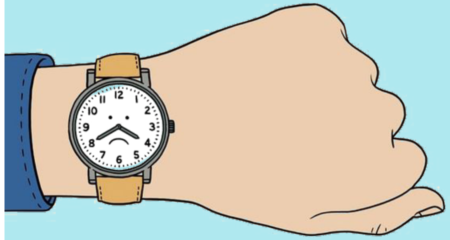


Phylogenetics: quantifying evolution

Episode in the series on phylogenetics

https://www.seh.ox.ac.uk/blog/mammals-during-the-extinction-of-dinosaurs

Mikael Bodén

Phylogenetics: quantifying evolution (Part 1)

Context: distances and models

Evolutionary distance and corrections (p-distance, Poisson and Gamma)

Molecular clock

Bear in mind...

most related sequences have *many* positions that have mutated, *some* of which have mutated several times

We need to effectively capture such dynamic changes

Metric of distance v. model of change

- The **evolutionary distance** between two sequences is an estimate of the number of mutations that has occurred since they diverged from their common ancestor

- While largely random, general rules may be governing which mutations lead to *changes* over time, imprinted in DNA, RNA and amino acid sequence

- **Evolutionary models** attempt to formalise tendencies of *change* in `<INSERT-ALPHABET-HERE>` sequences

`ALPHABETs`

- DNA: `A,C,G,T`
- RNA: `A,C,G,U`
- Protein: `A,R,N,D,`…
- more…

Sets of species/sequences: Distance matrices

- Define sequence distance
- Calculate all pairwise distances

Suppose we have three species i, j and k and a distance metric D

$$D = \begin{bmatrix} D_{i,i} & D_{i,j} & D_{i,k} \\ D_{j,i} & D_{j,j} & D_{j,k} \\ D_{k,i} & D_{k,j} & D_{k,k} \end{bmatrix}$$

p-distance (fractional alignment difference)

The simplest “evolutionary distance” between two sequences is the observed number of mutations since diverged.

$$p = \frac{D}{L}$$

Positions at which sequences differ

Total number of positions

$$\boxed{1 - p} = \frac{L - D}{L}$$

Positions at which sequences are the same

Total number of positions

The probability of “no change”

p-distance (fractional alignment difference)

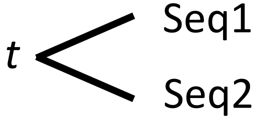
$$p = \frac{D}{L}$$

- Example
  - AAABBA
  - ABABAA
  - p = 2/6 = 1/3 = 0.333

- Two conditions for evolutionary time to be proportional to number of changes observed from an alignment
  1. all sequences mutate at a constant rate
  2. no position has mutated more than once

Poisson distance correction accounts for multiple mutations at site

- **Mutation rate** per site: r
- After time t, **expected** number of mutations at site: rt
- *No* mutation at t: e^{-rt} Euler's number (constant)
- Two sequences share ancestor at t, so $2rt$ away

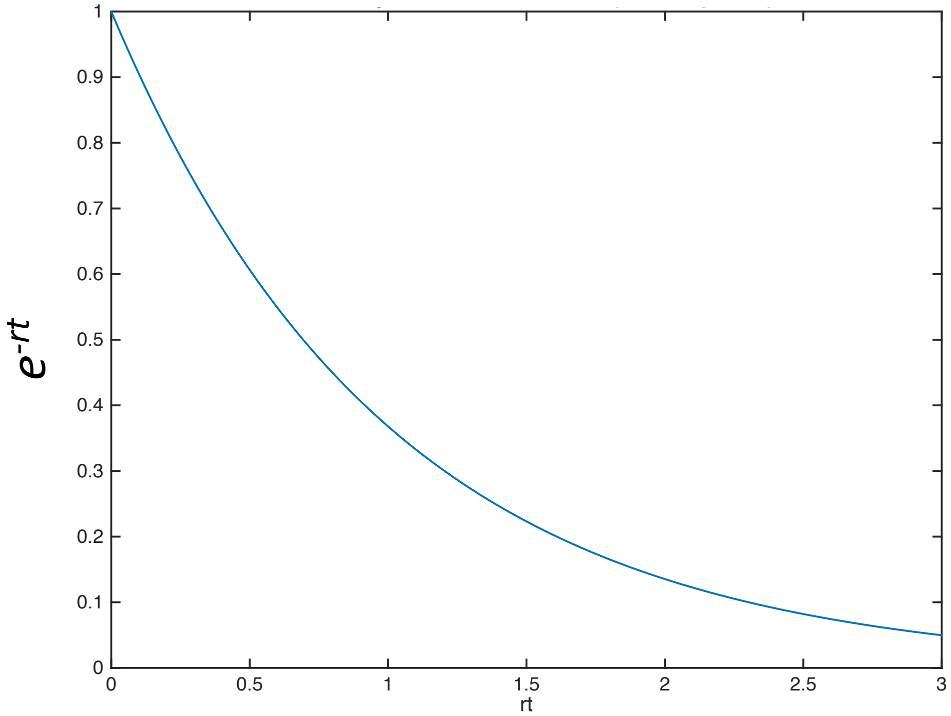


$2rt = d$

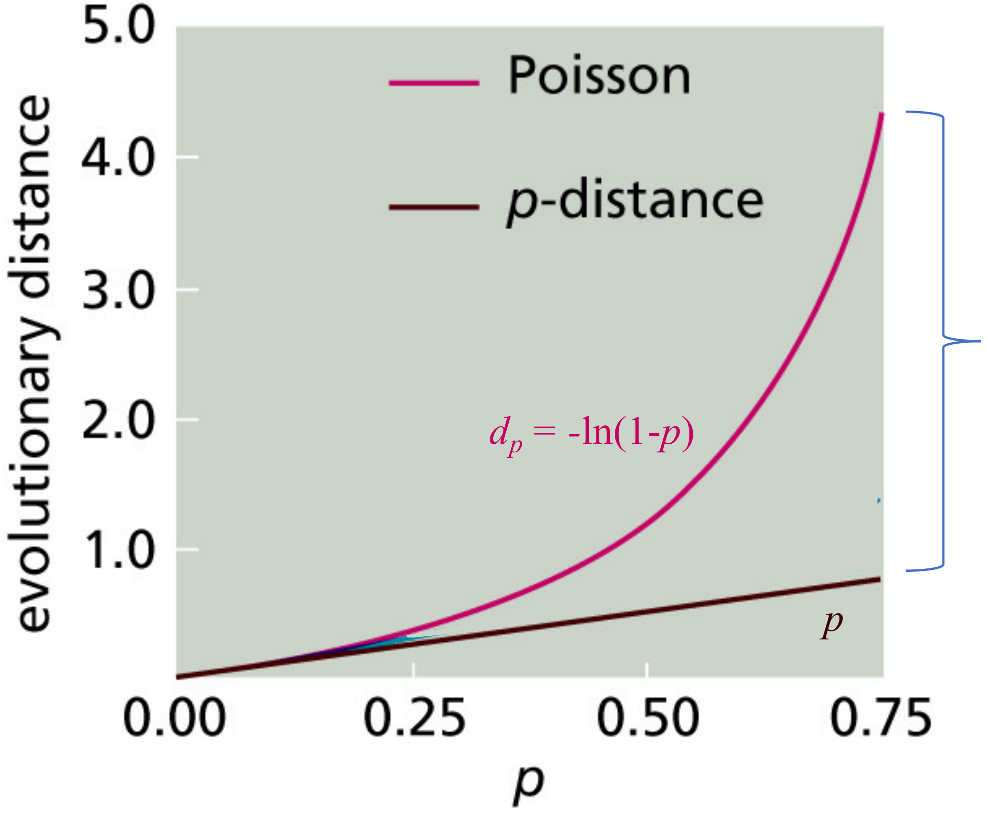
$1\text{-}p = e^{-2rt} = e^{-d}$ ← The probability of "no change"

$d_p = \text{-}\ln(1\text{-}p)$

Probability of no mutation at site (ONE sequence) (Poisson distribution)



Distance varies when *p* is corrected

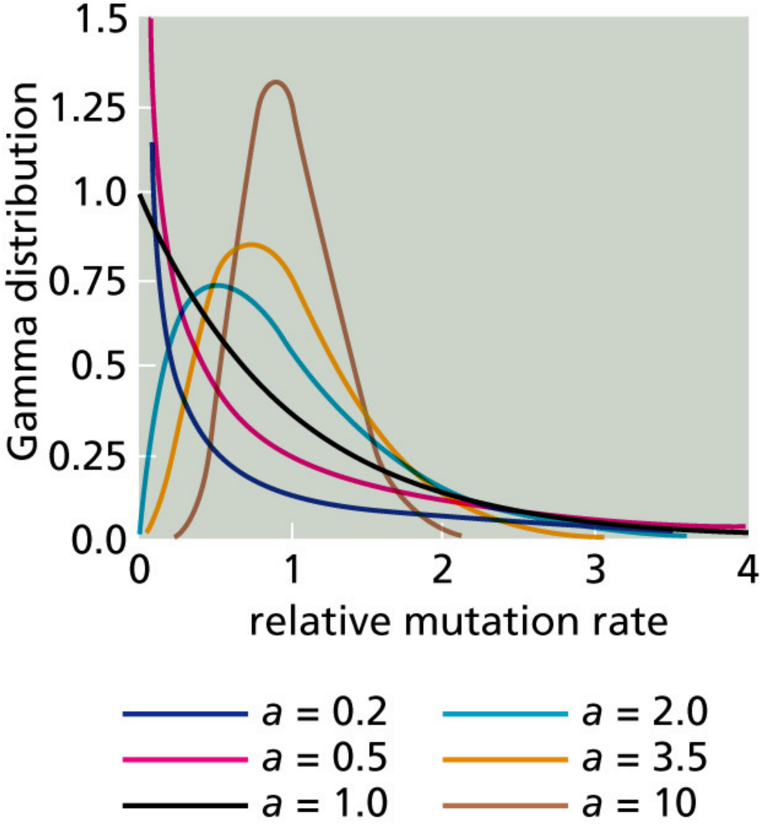


Mutations
missed
by *p*-distance

Gamma distance correction
Accounts for site-specific rates

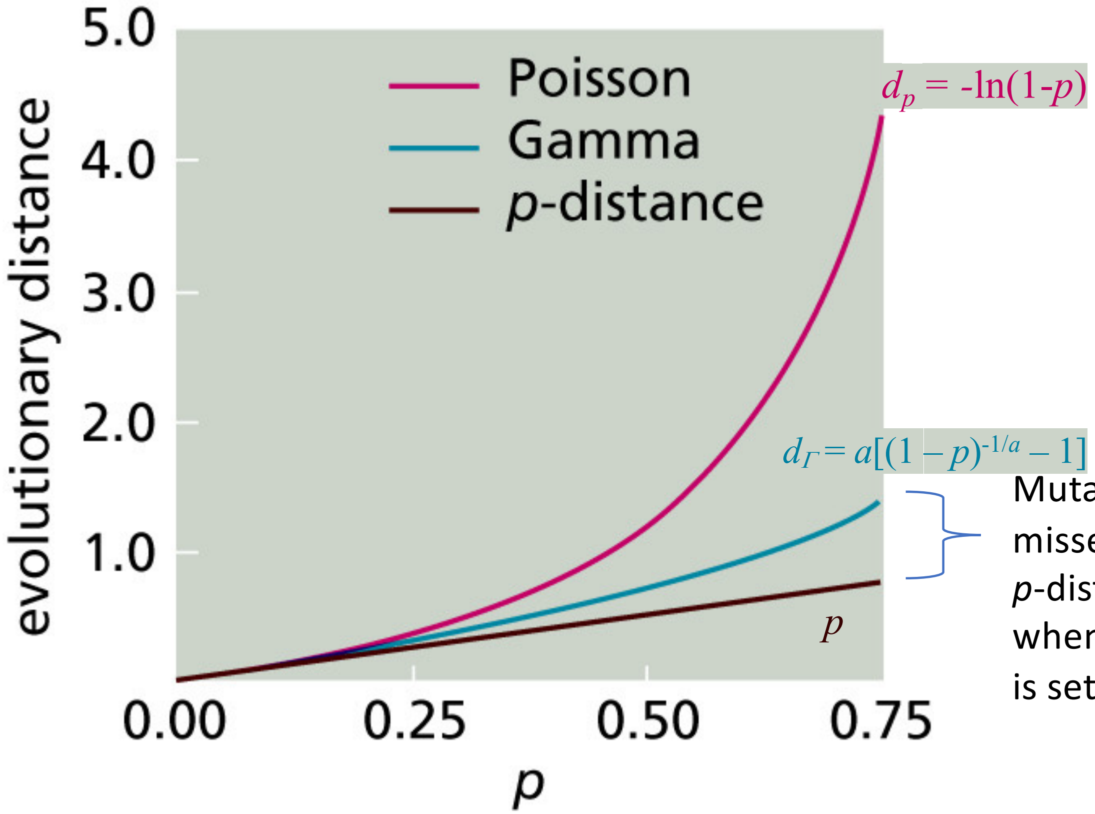
- Poisson only had one r for whole sequence
- The Gamma distribution can model how r varies across sites using a parameter a

$$d_{\Gamma} = a[(1 - p)^{-1/a} - 1]$$



Gamma corrects distance estimate for changes that can be explained by a variable rate
(a can be found by inspecting relevant data)

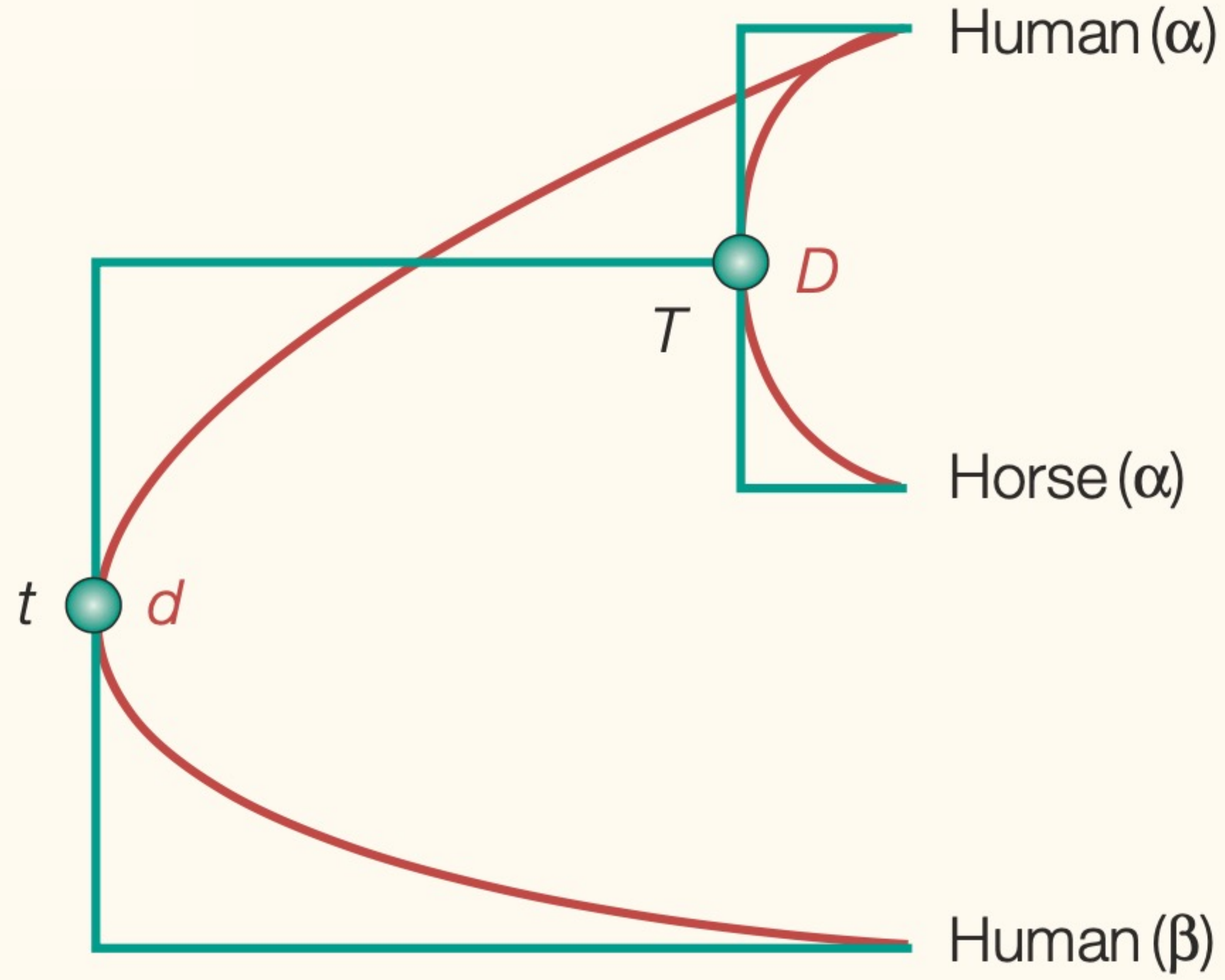
Distance varies when *p* is corrected



Mutations missed by *p*-distance when Gamma is set at *a*=0.70

Distance = Time? Molecular clock and rate variation

- Zuckerkandl and Pauling noticed that the number of amino acid differences between different lineages changes roughly linearly with time; the rate of evolutionary change was approx. constant over time and over different lineages—this is known as the molecular clock hypothesis
- This is challenged by
  - Changing generation times, population size, species-specific differences (metabolism, ecology, etc), change in function and in the intensity of natural selection



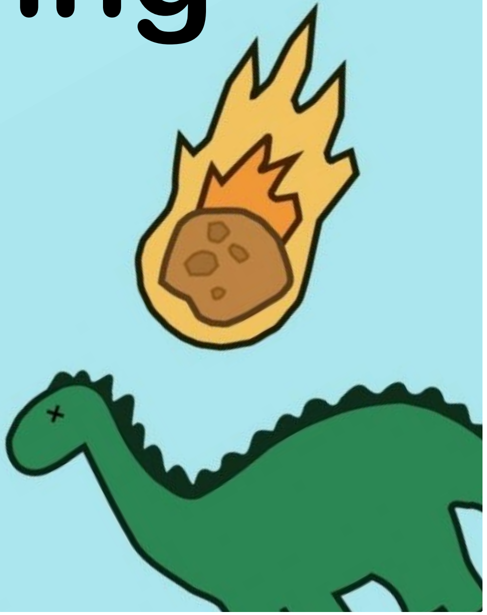
Kumar, S. Nature Reviews Genetics 6:654–662(2005)

Phylogenetics: quantifying evolution (Part 2)

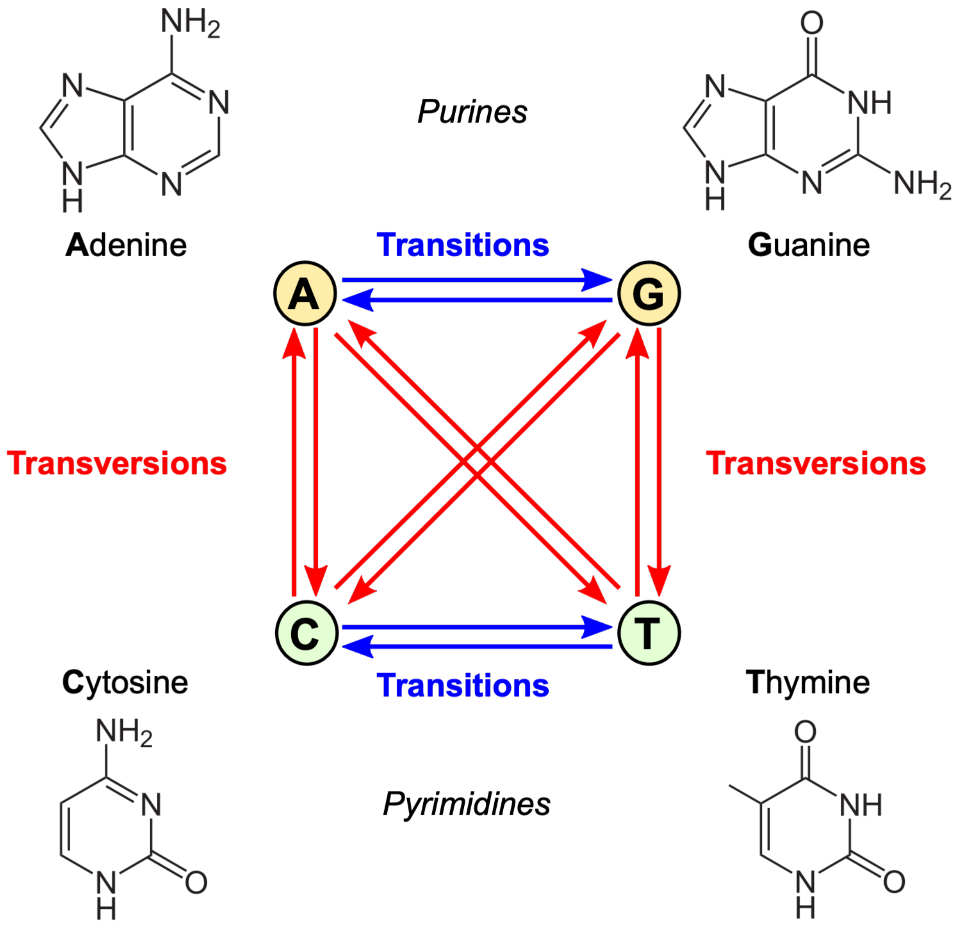
DNA models

Evolutionary models and rate matrices

Probabilistic meaning

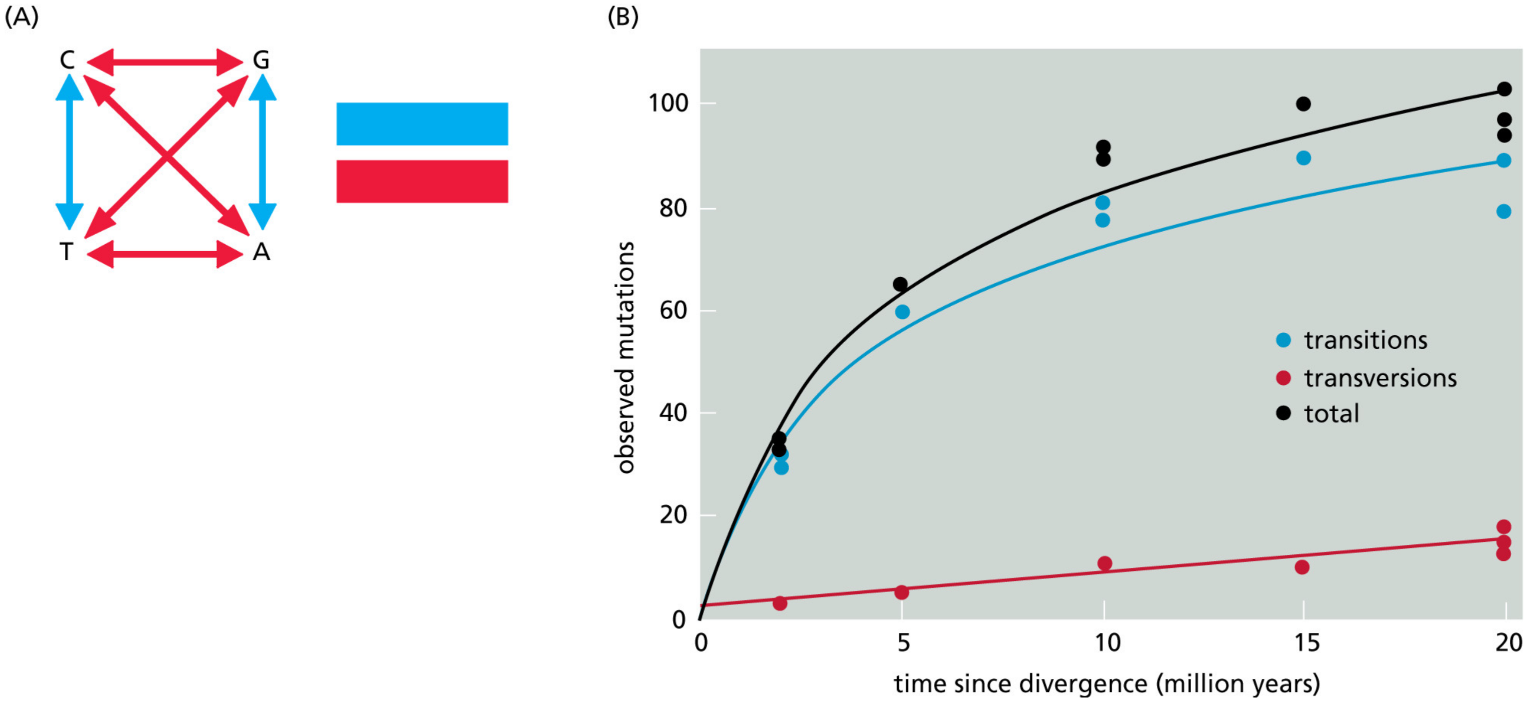


Transition vs. transversion



https://en.wikipedia.org/wiki/Transversion

Transition vs. transversion



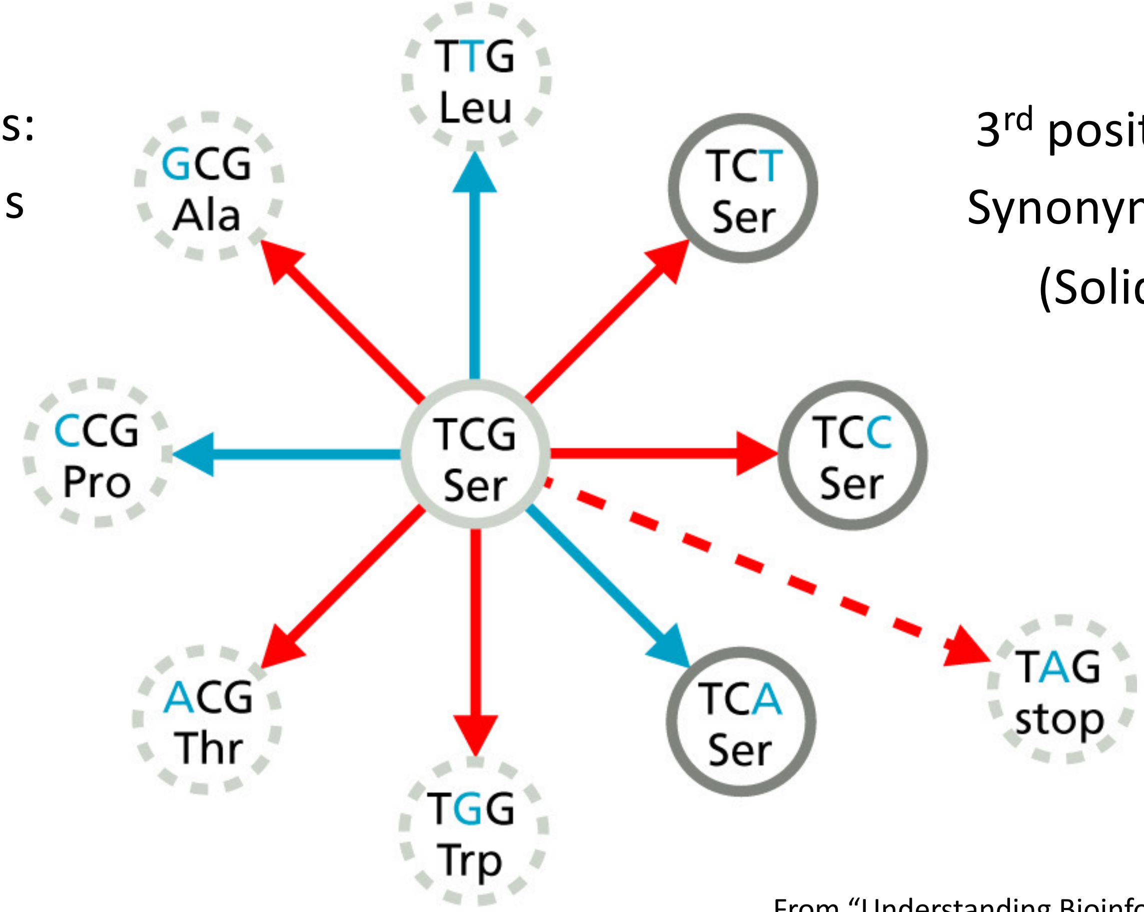
From “Understanding Bioinformatics”, Zvelebil & Baum, p238.

Different codon positions have different mutation rates

1st & 2nd positions:

Non-synonymous

(Dashed)



3rd position:

Synonymous

(Solid)

From “Understanding Bioinformatics”, Zvelebil & Baum, p241.

Evolutionary models

Model name	Base Composition	Different transition and transversion rates	All transition rates identical	All transversion rates identical	Reference
Jukes-Cantor (JC69)	1:1:1:1	No	Yes	Yes	Jukes and Cantor (1969)
Felsenstein 81 (F81)	Variable	No	Yes	Yes	Felsenstein (1981)
Kimura 2 Param (K80)	1:1:1:1	Yes	Yes	Yes	Kimura (1980)
HKY85	Variable	Yes	No	No	Hasegawa et al. (1985)
Tamura-Nei (TN)	Variable	Yes	No	Yes	Tamura and Nei (1993)
K3P (K81)	Variable	Yes	No	Yes	Kimura (1981)
SYM	1:1:1:1	Yes	No	No	Zharkikh (1994)
REV (GTR)	Variable	Yes	No	No	Rodriguez et al. (1990)

From “Understanding Bioinformatics”, Zvelebil & Baum, p253.

Models of (DNA) evolution (based on rate)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

JC69 model (Jukes and Cantor 1969)

$$Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$$

K81 model (Kimura 1981)

$$Q = \begin{array}{c c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{C} & \mathtt{T} \end{array} & \\ \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix} & \begin{array}{c} \mathtt{A} \\ \mathtt{G} \\ \mathtt{C} \\ \mathtt{T} \end{array} \end{array}$$

F81 model (Felsenstein 1981)

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

GTR model (Tavaré 1986)

Models of (DNA) evolution: as rate matrix

https://en.wikipedia.org/wiki/Models_of_DNA_evolution

Imaginary (DNA) lineage over (discrete) time

A ← Ancient time
A
A
A
G
G
C
C
G
G
G ← Present time

To

From

	A	C	G	T
A	-0.3	0.1	0.1	0.1
C	0.1	-0.3	0.1	0.1
G	0.1	0.1	-0.3	0.1
T	0.1	0.1	0.1	-0.3

Note: Theory of continuous time Markov chain (CTMC)
- models change of state of a *single* discrete random variable
- defines probabilities of state changes, satisfying the *Markov property* (i.e. decision of future state depends only on current state)

The Jukes-Cantor (JC) model

- Treats substitutions uniformly
- Sites have identical rates, but depend on nucleotide identity

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \scriptstyle A & \scriptstyle C & \scriptstyle G & \scriptstyle T \end{array} \\ \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} \end{array}$$

Re-running evolution (forward)

P(A) = ?

← Ancient time

Evolutionary model

Predict

← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time

Evolutionary model

Predict

← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time

Evolutionary model

Predict

P(C|A) = ? ← T = 0.4

← Present time

Re-running evolution (forward)

P(A) = 0.25 ← Ancient time

P(C|A) = ? ← T = 1.0

← Present time

Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

$$T\text{=1} \quad \begin{array}{c|cccc} & \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ \hline \mathbf{A} & 0.75 & 0.08 & 0.08 & 0.08 \\ \mathbf{C} & 0.08 & 0.75 & 0.08 & 0.08 \\ \mathbf{G} & 0.08 & 0.08 & 0.75 & 0.08 \\ \mathbf{T} & 0.08 & 0.08 & 0.08 & 0.75 \end{array}$$

$$T\text{=0} \quad \begin{array}{c|cccc} & \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ \hline \mathbf{A} & 1 & 0 & 0 & 0 \\ \mathbf{C} & 0 & 1 & 0 & 0 \\ \mathbf{G} & 0 & 0 & 1 & 0 \\ \mathbf{T} & 0 & 0 & 0 & 1 \end{array}$$

⬆

$$Q\text{=} \quad \begin{array}{c|cccc} & \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ \hline \mathbf{A} & -0.3 & 0.1 & 0.1 & 0.1 \\ \mathbf{C} & 0.1 & -0.3 & 0.1 & 0.1 \\ \mathbf{G} & 0.1 & 0.1 & -0.3 & 0.1 \\ \mathbf{T} & 0.1 & 0.1 & 0.1 & -0.3 \end{array}$$

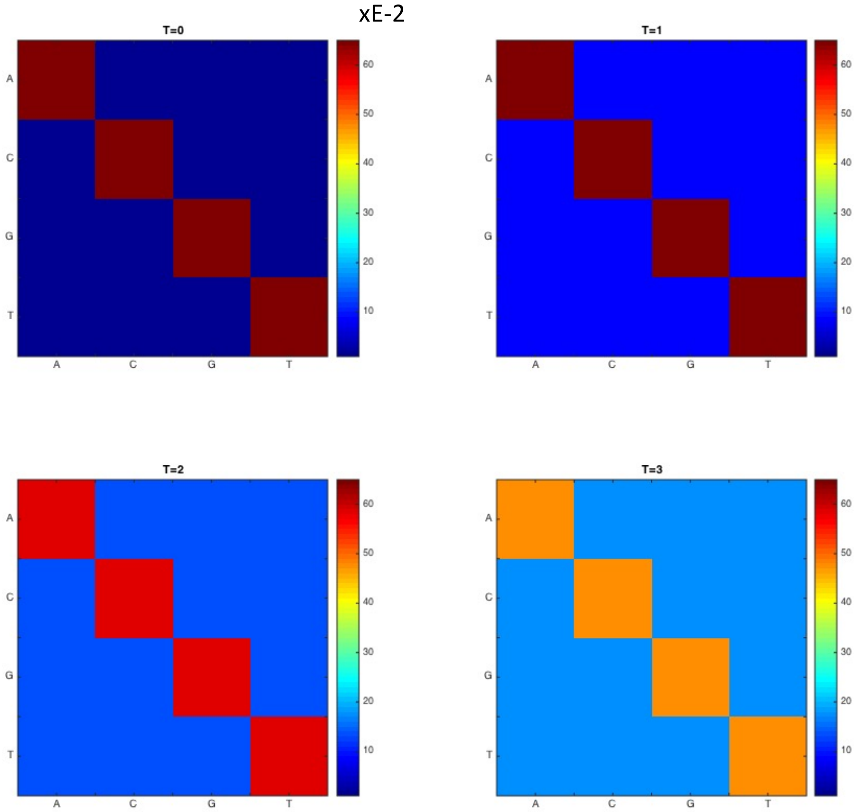
Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$\mathbf{P}(T + dT) =$

$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$



Rate matrix for proteins

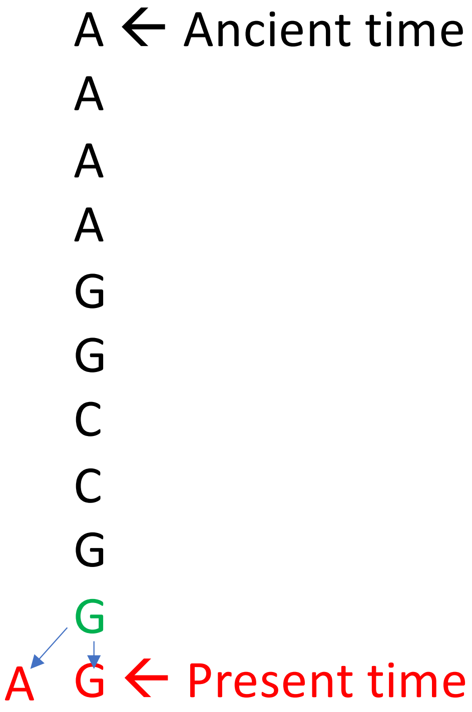
Dayhoff (remember PAM)

Zvelebil and Baum, sec 5.1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	**-1.33**	0.01	0.04	0.06	0.01	0.03	0.10	0.21	0.01	0.02	0.04	0.02	0.01	0.01	0.13	0.28	0.22	0.00	0.01	0.13
R	0.02	**-0.87**	0.01	0.00	0.01	0.09	0.00	0.01	0.08	0.02	0.01	0.37	0.01	0.01	0.05	0.11	0.02	0.02	0.00	0.02
N	0.09	0.01	**-1.78**	0.42	0.00	0.04	0.07	0.12	0.18	0.03	0.03	0.25	0.00	0.01	0.02	0.34	0.13	0.00	0.03	0.01
D	0.11	0.00	0.36	**-1.41**	0.00	0.05	0.56	0.11	0.03	0.01	0.00	0.06	0.00	0.00	0.01	0.07	0.04	0.00	0.00	0.01
C	0.03	0.01	0.00	0.00	**-0.27**	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00	0.01	0.11	0.01	0.00	0.03	0.03
Q	0.08	0.10	0.04	0.06	0.00	**-1.24**	0.35	0.03	0.20	0.01	0.06	0.12	0.02	0.00	0.08	0.04	0.03	0.00	0.00	0.02
E	0.17	0.00	0.06	0.53	0.00	0.27	**-1.36**	0.07	0.02	0.02	0.01	0.07	0.00	0.00	0.03	0.06	0.02	0.00	0.01	0.02
G	0.21	0.00	0.06	0.06	0.00	0.01	0.04	**-0.65**	0.00	0.00	0.01	0.02	0.00	0.01	0.02	0.16	0.02	0.00	0.00	0.04
H	0.02	0.10	0.21	0.04	0.01	0.23	0.02	0.01	**-0.88**	0.00	0.04	0.02	0.00	0.02	0.05	0.03	0.01	0.00	0.04	0.03
I	0.06	0.03	0.03	0.01	0.02	0.01	0.03	0.00	0.00	**-1.28**	0.22	0.04	0.05	0.08	0.01	0.02	0.11	0.00	0.01	0.57
L	0.04	0.01	0.01	0.00	0.00	0.03	0.01	0.01	0.02	0.09	**-0.53**	0.02	0.08	0.06	0.02	0.01	0.02	0.01	0.01	0.11
K	0.02	0.19	0.13	0.03	0.00	0.06	0.04	0.02	0.01	0.02	0.02	**-0.75**	0.04	0.00	0.02	0.07	0.08	0.00	0.00	0.01
M	0.06	0.04	0.00	0.00	0.00	0.04	0.02	0.02	0.00	0.12	0.45	0.19	**-1.25**	0.04	0.01	0.04	0.06	0.00	0.00	0.17
F	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.07	0.13	0.00	0.01	**-0.55**	0.01	0.03	0.01	0.01	0.21	0.01
P	0.22	0.04	0.02	0.01	0.01	0.06	0.03	0.03	0.03	0.00	0.03	0.03	0.00	0.00	**-0.75**	0.17	0.05	0.00	0.00	0.03
S	0.35	0.06	0.20	0.05	0.05	0.02	0.04	0.21	0.01	0.01	0.02	0.08	0.01	0.02	0.12	**-1.60**	0.32	0.01	0.01	0.02
T	0.32	0.01	0.09	0.03	0.01	0.02	0.02	0.03	0.01	0.07	0.03	0.11	0.02	0.01	0.04	0.38	**-1.29**	0.00	0.01	0.10
W	0.00	0.08	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.04	0.00	0.00	0.03	0.00	0.05	0.00	**-0.24**	0.02	0.00
Y	0.02	0.00	0.04	0.00	0.03	0.00	0.01	0.00	0.04	0.01	0.02	0.01	0.00	0.28	0.00	0.02	0.02	0.01	**-0.55**	0.02
W	0.18	0.01	0.01	0.01	0.02	0.01	0.02	0.05	0.02	0.33	0.15	0.01	0.04	0.01	0.03	0.02	0.09	0.00	0.01	**-0.99**

On *what* are protein models based?

```
DYISWWQQQ
DYISSWQEQ
DYISLWQEQ
DYISLWQDD
```



- Answer: Counts of character pairs from alignments of closely related sequences
- When sequences are *really* close (in time) the rates and probabilities of change are approximately linear
- Counts scaled based on sequence divergence, and averaging across many alignments

Probabilities come from the model

$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

$$\mathbf{P}_{T=1} = \left(\begin{matrix}
0.37 & 0.02 & 0.02 & 0.03 & 0.01 & 0.02 & 0.05 & 0.07 & 0.01 & 0.02 & 0.03 & 0.04 & 0.01 & 0.01 & 0.04 & 0.1 \\
0.04 & 0.40 & 0.03 & 0.02 & 0.01 & 0.05 & 0.03 & 0.04 & 0.03 & 0.01 & 0.03 & 0.14 & 0.01 & 0.01 & 0.02 & 0.0 \\
0.05 & 0.03 & 0.27 & 0.12 & 0.01 & 0.03 & 0.05 & 0.06 & 0.04 & 0.02 & 0.02 & 0.08 & 0.01 & 0.01 & 0.02 & 0.1 \\
0.05 & 0.01 & 0.08 & 0.42 & 0.00 & 0.03 & 0.14 & 0.05 & 0.02 & 0.01 & 0.01 & 0.04 & 0.00 & 0.00 & 0.02 & 0.0 \\
0.06 & 0.02 & 0.01 & 0.01 & 0.62 & 0.01 & 0.01 & 0.03 & 0.01 & 0.02 & 0.03 & 0.01 & 0.01 & 0.01 & 0.01 & 0.0 \\
0.06 & 0.07 & 0.04 & 0.05 & 0.00 & 0.28 & 0.11 & 0.03 & 0.04 & 0.01 & 0.04 & 0.10 & 0.01 & 0.01 & 0.03 & 0.0 \\
0.07 & 0.03 & 0.04 & 0.14 & 0.00 & 0.07 & 0.34 & 0.04 & 0.02 & 0.01 & 0.02 & 0.08 & 0.01 & 0.01 & 0.02 & 0.0 \\
0.07 & 0.02 & 0.03 & 0.04 & 0.01 & 0.01 & 0.03 & 0.62 & 0.01 & 0.01 & 0.01 & 0.02 & 0.00 & 0.00 & 0.01 & 0.0 \\
0.03 & 0.05 & 0.06 & 0.04 & 0.01 & 0.06 & 0.04 & 0.03 & 0.38 & 0.01 & 0.03 & 0.05 & 0.01 & 0.02 & 0.02 & 0.0 \\
0.04 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.35 & 0.15 & 0.02 & 0.04 & 0.03 & 0.01 & 0.0 \\
0.03 & 0.02 & 0.01 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.01 & 0.08 & 0.52 & 0.02 & 0.04 & 0.05 & 0.02 & 0.0 \\
0.05 & 0.10 & 0.05 & 0.04 & 0.00 & 0.06 & 0.07 & 0.03 & 0.02 & 0.01 & 0.03 & 0.36 & 0.01 & 0.01 & 0.02 & 0.0 \\
0.05 & 0.02 & 0.01 & 0.01 & 0.01 & 0.03 & 0.02 & 0.02 & 0.01 & 0.09 & 0.19 & 0.03 & 0.28 & 0.03 & 0.01 & 0.0 \\
0.02 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.02 & 0.04 & 0.11 & 0.01 & 0.02 & 0.51 & 0.01 & 0.0 \\
0.08 & 0.02 & 0.01 & 0.02 & 0.00 & 0.02 & 0.03 & 0.03 & 0.01 & 0.01 & 0.03 & 0.03 & 0.01 & 0.01 & 0.55 & 0.0 \\
0.12 & 0.03 & 0.06 & 0.04 & 0.02 & 0.03 & 0.04 & 0.07 & 0.02 & 0.02 & 0.03 & 0.04 & 0.01 & 0.01 & 0.04 & 0.2 \\
0.10 & 0.02 & 0.04 & 0.03 & 0.01 & 0.02 & 0.04 & 0.03 & 0.01 & 0.04 & 0.03 & 0.05 & 0.02 & 0.01 & 0.03 & 0.1 \\
0.02 & 0.03 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0.05 & 0.01 & 0.01 & 0.04 & 0.01 & 0.0 \\
0.02 & 0.02 & 0.03 & 0.02 & 0.01 & 0.01 & 0.01 & 0.01 & 0.05 & 0.02 & 0.04 & 0.01 & 0.01 & 0.13 & 0.01 & 0.0 \\
0.08 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.02 & 0.02 & 0.01 & 0.14 & 0.10 & 0.02 & 0.03 & 0.02 & 0.01 & 0.0
\end{matrix}\right.$$

$$\mathbf{Q} = \left(\begin{matrix}
-1.12 & 0.03 & 0.02 & 0.04 & 0.02 & 0.04 & 0.10 & 0.12 & 0.01 & 0.01 & 0.04 & 0.06 & 0.0 \\
0.05 & -0.97 & 0.03 & 0.01 & 0.01 & 0.12 & 0.03 & 0.05 & 0.06 & 0.01 & 0.05 & 0.35 & 0.0 \\
0.05 & 0.03 & -1.45 & 0.32 & 0.01 & 0.06 & 0.06 & 0.10 & 0.10 & 0.03 & 0.01 & 0.20 & 0.0 \\
0.07 & 0.01 & 0.22 & -0.99 & 0.00 & 0.02 & 0.38 & 0.08 & 0.02 & 0.00 & 0.01 & 0.03 & 0.0 \\
0.09 & 0.02 & 0.01 & 0.00 & -0.49 & 0.00 & 0.00 & 0.03 & 0.01 & 0.01 & 0.04 & 0.01 & 0.0 \\
0.08 & 0.14 & 0.06 & 0.04 & 0.00 & -1.38 & 0.33 & 0.03 & 0.11 & 0.01 & 0.08 & 0.25 & 0.0 \\
0.14 & 0.02 & 0.04 & 0.37 & 0.00 & 0.21 & -1.24 & 0.05 & 0.02 & 0.01 & 0.01 & 0.17 & 0.0 \\
0.13 & 0.03 & 0.05 & 0.05 & 0.01 & 0.01 & 0.04 & -0.50 & 0.01 & 0.00 & 0.01 & 0.02 & 0.0 \\
0.03 & 0.10 & 0.16 & 0.06 & 0.01 & 0.17 & 0.04 & 0.02 & -0.99 & 0.01 & 0.05 & 0.06 & 0.0 \\
0.02 & 0.01 & 0.02 & 0.00 & 0.00 & 0.01 & 0.01 & 0.00 & 0.00 & -1.23 & 0.29 & 0.02 & 0.0 \\
0.04 & 0.02 & 0.01 & 0.01 & 0.01 & 0.03 & 0.01 & 0.01 & 0.01 & 0.16 & -0.73 & 0.02 & 0.1 \\
0.08 & 0.25 & 0.12 & 0.03 & 0.00 & 0.15 & 0.16 & 0.03 & 0.02 & 0.02 & 0.02 & -1.12 & 0.0 \\
0.08 & 0.03 & 0.01 & 0.01 & 0.01 & 0.06 & 0.02 & 0.02 & 0.01 & 0.22 & 0.44 & 0.06 & -1.3 \\
0.02 & 0.01 & 0.00 & 0.00 & 0.01 & 0.00 & 0.01 & 0.00 & 0.02 & 0.05 & 0.19 & 0.01 & 0.0 \\
0.13 & 0.03 & 0.01 & 0.03 & 0.00 & 0.04 & 0.04 & 0.02 & 0.02 & 0.01 & 0.04 & 0.04 & 0.0 \\
0.31 & 0.06 & 0.16 & 0.06 & 0.03 & 0.04 & 0.04 & 0.12 & 0.02 & 0.02 & 0.03 & 0.06 & 0.0 \\
0.19 & 0.03 & 0.08 & 0.02 & 0.01 & 0.03 & 0.05 & 0.02 & 0.01 & 0.07 & 0.03 & 0.09 & 0.0 \\
0.01 & 0.05 & 0.00 & 0.01 & 0.02 & 0.01 & 0.01 & 0.03 & 0.01 & 0.01 & 0.06 & 0.01 & 0.0 \\
0.02 & 0.02 & 0.05 & 0.02 & 0.01 & 0.01 & 0.01 & 0.01 & 0.10 & 0.02 & 0.04 & 0.01 & 0.0 \\
0.18 & 0.01 & 0.01 & 0.01 & 0.02 & 0.01 & 0.04 & 0.02 & 0.00 & 0.40 & 0.16 & 0.02 & 0.0
\end{matrix}\right.$$

Probabilities come from the model

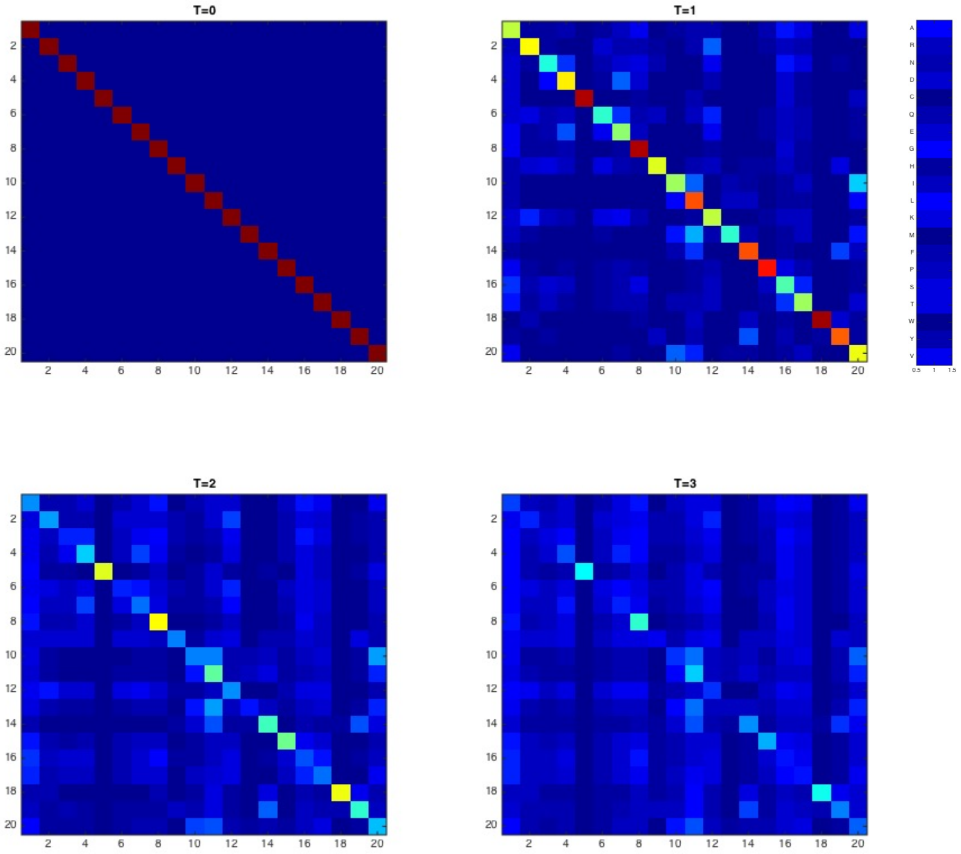
$P_{ij}(T)$ can be written as a matrix $\mathbf{P}(T)$

In discrete time:

$$\mathbf{P}(T + dT) =$$

$$\mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

So: there is a probability matrix for all possible time lapses.



Rate matrix for proteins

Whelan & Goldman

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	-1.12	0.03	0.02	0.04	0.02	0.04	0.10	0.12	0.01	0.01	0.04	0.06	0.02	0.01	0.07	0.25	0.14	0.00	0.01	0.15
R	0.05	-0.97	0.03	0.01	0.01	0.12	0.03	0.05	0.06	0.01	0.05	0.35	0.01	0.00	0.03	0.09	0.04	0.02	0.01	0.02
N	0.05	0.03	-1.45	0.32	0.01	0.06	0.06	0.10	0.10	0.03	0.01	0.20	0.00	0.00	0.01	0.29	0.13	0.00	0.04	0.02
D	0.07	0.01	0.22	-0.99	0.00	0.02	0.38	0.08	0.02	0.00	0.01	0.03	0.00	0.00	0.02	0.08	0.02	0.00	0.01	0.01
C	0.09	0.02	0.01	0.00	-0.49	0.00	0.00	0.03	0.01	0.01	0.04	0.01	0.01	0.02	0.01	0.10	0.03	0.01	0.02	0.08
Q	0.08	0.14	0.06	0.04	0.00	-1.38	0.33	0.03	0.11	0.01	0.08	0.25	0.03	0.00	0.05	0.08	0.06	0.00	0.01	0.02
E	0.14	0.02	0.04	0.37	0.00	0.21	-1.24	0.05	0.02	0.01	0.01	0.17	0.01	0.00	0.03	0.05	0.05	0.00	0.01	0.04
G	0.13	0.03	0.05	0.05	0.01	0.01	0.04	-0.50	0.01	0.00	0.01	0.02	0.00	0.00	0.01	0.10	0.02	0.01	0.00	0.01
H	0.03	0.10	0.16	0.06	0.01	0.17	0.04	0.02	-0.99	0.01	0.05	0.06	0.01	0.03	0.03	0.05	0.03	0.00	0.14	0.01
I	0.02	0.01	0.02	0.00	0.00	0.01	0.01	0.00	0.00	-1.23	0.29	0.02	0.09	0.04	0.01	0.02	0.09	0.00	0.02	0.58
L	0.04	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.16	-0.73	0.02	0.10	0.09	0.02	0.03	0.02	0.01	0.02	0.13
K	0.08	0.25	0.12	0.03	0.00	0.15	0.16	0.03	0.02	0.02	0.02	-1.12	0.02	0.00	0.03	0.07	0.09	0.00	0.01	0.02
M	0.08	0.03	0.01	0.01	0.01	0.06	0.02	0.02	0.01	0.22	0.44	0.06	-1.32	0.05	0.01	0.04	0.10	0.01	0.02	0.15
F	0.02	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.05	0.19	0.01	0.02	-0.72	0.01	0.04	0.01	0.02	0.24	0.05
P	0.13	0.03	0.01	0.03	0.00	0.04	0.04	0.02	0.02	0.01	0.04	0.04	0.00	0.01	-0.61	0.12	0.05	0.00	0.01	0.02
S	0.31	0.06	0.16	0.06	0.03	0.04	0.04	0.12	0.02	0.02	0.03	0.06	0.01	0.02	0.08	-1.39	0.28	0.01	0.03	0.02
T	0.19	0.03	0.08	0.02	0.01	0.03	0.05	0.02	0.01	0.07	0.03	0.09	0.03	0.01	0.04	0.32	-1.16	0.00	0.01	0.10
W	0.01	0.05	0.00	0.01	0.02	0.01	0.01	0.03	0.01	0.01	0.06	0.01	0.01	0.06	0.01	0.04	0.01	-0.47	0.09	0.03
Y	0.02	0.02	0.05	0.02	0.01	0.01	0.01	0.01	0.10	0.02	0.04	0.01	0.01	0.26	0.01	0.06	0.02	0.04	-0.73	0.02
W	0.18	0.01	0.01	0.01	0.02	0.01	0.04	0.02	0.00	0.40	0.16	0.02	0.04	0.03	0.02	0.02	0.09	0.01	0.01	-1.09

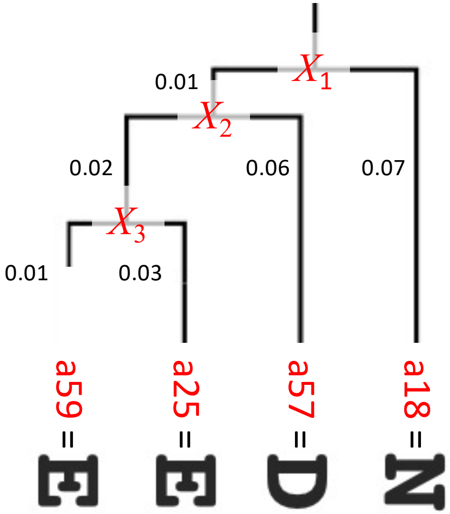
Transition probability matrix (CTMC)

Whelan & Goldman

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	W
A	**0.37**	0.02	0.02	0.03	0.01	0.02	0.05	0.07	0.01	0.02	0.03	0.04	0.01	0.01	0.04	0.10	0.07	0.00	0.01	0.07
R	0.04	**0.40**	0.03	0.02	0.01	0.05	0.03	0.04	0.03	0.01	0.03	0.14	0.01	0.01	0.02	0.05	0.03	0.01	0.01	0.02
N	0.05	0.03	**0.27**	0.12	0.01	0.03	0.05	0.06	0.04	0.02	0.02	0.08	0.01	0.01	0.02	0.10	0.06	0.00	0.02	0.02
D	0.05	0.01	0.08	**0.42**	0.00	0.03	0.14	0.05	0.02	0.01	0.01	0.04	0.00	0.00	0.02	0.05	0.03	0.00	0.01	0.02
C	0.06	0.02	0.01	0.01	**0.62**	0.01	0.01	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.01	0.06	0.03	0.01	0.02	0.05
Q	0.06	0.07	0.04	0.05	0.00	**0.28**	0.11	0.03	0.04	0.01	0.04	0.10	0.01	0.01	0.03	0.05	0.04	0.00	0.01	0.02
E	0.07	0.03	0.04	0.14	0.00	0.07	**0.34**	0.04	0.02	0.01	0.02	0.08	0.01	0.01	0.02	0.04	0.04	0.00	0.01	0.03
G	0.07	0.02	0.03	0.04	0.01	0.01	0.03	**0.62**	0.01	0.01	0.01	0.02	0.00	0.00	0.01	0.06	0.02	0.00	0.01	0.02
H	0.03	0.05	0.06	0.04	0.01	0.06	0.04	0.03	**0.38**	0.01	0.03	0.05	0.01	0.02	0.02	0.04	0.03	0.01	0.07	0.02
I	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.35**	0.15	0.02	0.04	0.03	0.01	0.02	0.05	0.00	0.01	0.21
L	0.03	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.08	**0.52**	0.02	0.04	0.05	0.02	0.02	0.02	0.01	0.02	0.09
K	0.05	0.10	0.05	0.04	0.00	0.06	0.07	0.03	0.02	0.01	0.03	**0.36**	0.01	0.01	0.02	0.05	0.05	0.00	0.01	0.02
M	0.05	0.02	0.01	0.01	0.01	0.03	0.02	0.02	0.01	0.09	0.19	0.03	**0.28**	0.03	0.01	0.03	0.05	0.01	0.01	0.09
F	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.04	0.11	0.01	0.02	**0.51**	0.01	0.03	0.02	0.02	0.12	0.04
P	0.08	0.02	0.01	0.02	0.00	0.02	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	**0.55**	0.06	0.04	0.00	0.01	0.02
S	0.12	0.03	0.06	0.04	0.02	0.03	0.04	0.07	0.02	0.02	0.03	0.04	0.01	0.01	0.04	**0.29**	0.10	0.01	0.02	0.03
T	0.10	0.02	0.04	0.03	0.01	0.02	0.04	0.03	0.01	0.04	0.03	0.05	0.02	0.01	0.03	0.12	**0.35**	0.00	0.01	0.06
W	0.02	0.03	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.05	0.01	0.01	0.04	0.01	0.03	0.01	**0.63**	0.06	0.02
Y	0.02	0.02	0.03	0.02	0.01	0.01	0.01	0.01	0.05	0.02	0.04	0.01	0.01	0.13	0.01	0.03	0.02	0.02	**0.51**	0.02
W	0.08	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.14	0.10	0.02	0.03	0.02	0.01	0.03	0.05	0.01	0.01	**0.40**

$t = 1$

Maximum likelihood can be used to determine the tree and ancestors



- Consider a single site (independent of all others)

Summary

- Various distance metrics available to quantify sequence similarity
  - Fractional (*p*-distance)
  - Poisson corrected
  - Gamma corrected
- Also need to account for chemical nature of sequence
  - Transitions/transversion
  - Codon dynamics
  - “Fixed” DNA models
- Evolutionary models based on real data capture similar trends
- Transition rate matrices help model evolution probabilistically

Reading

- Zvelebil & Baum (2008) *Understanding Bioinformatics*
  - ***Chapter 7*** *(7.1-3)*
  - ***Chapter 8*** *(8.1)*
- Kelley and Didulo, *Computational Biology: A Hypertextbook*
  - ***Chapter 6***