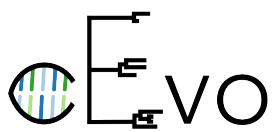
# Computational Biology

Lecturers:
Tanja Stadler, Carsten Magnus & Tim Vaughan

Teaching Assistants:
Jūlija Pečerska, Jérémie Sciré,
Sarah Nadeau & Marc Manceau

Computational Evolution
Department of Biosystems Science and Engineering

HS 2019



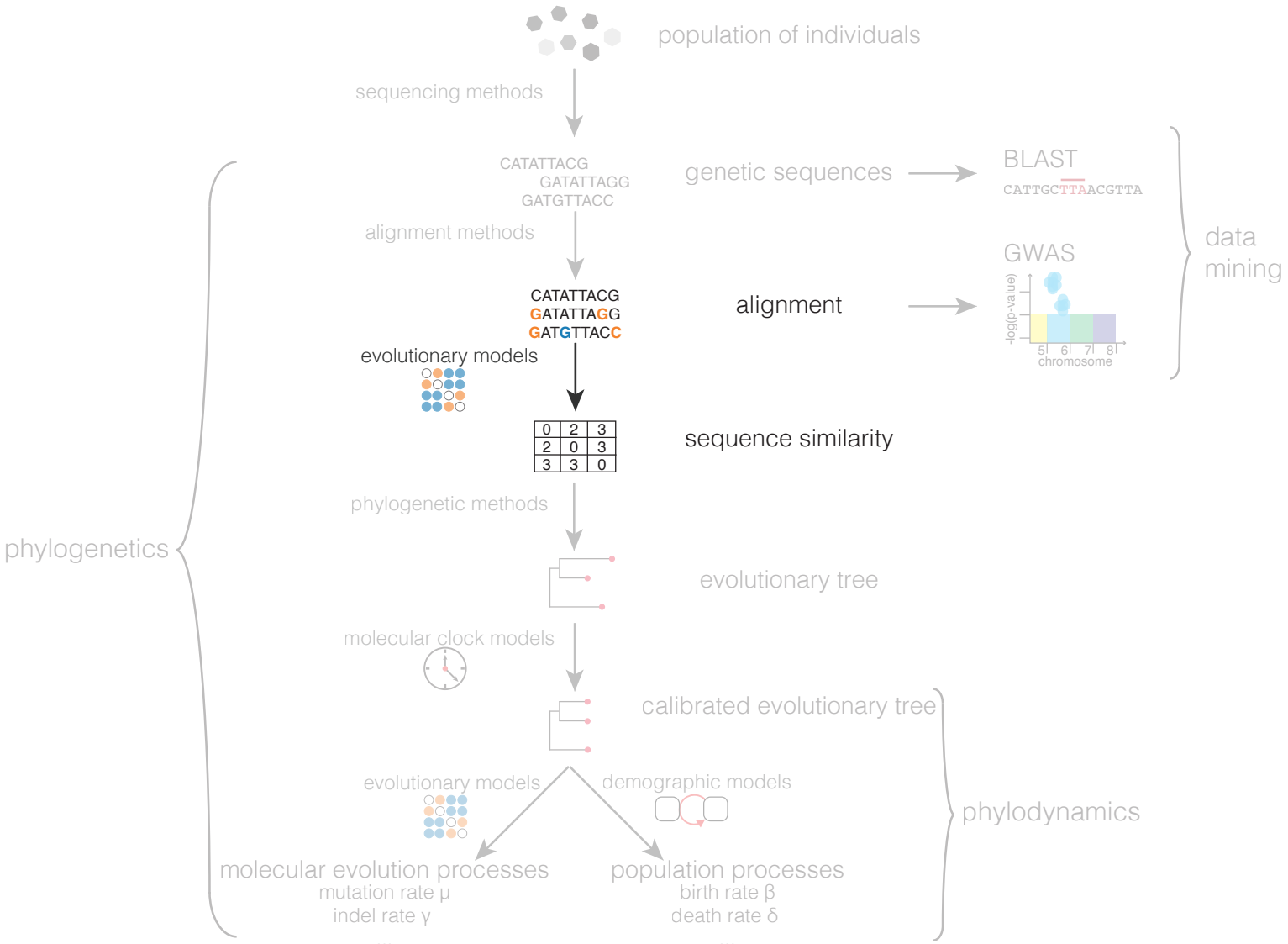
# GWAS, molecular evolution: Questions



- Does the odds ratio in a GWAS prove that a minor variant at a SNP position causes a genetic disease?
- In a GWAS, why can you not reject your null hypothesis if the $\mathrm{p}$-value is $< \alpha$?

# Overview





# What we will learn today



- nucleotide substitution models
- distance estimation under JC69
  - maximum likelihood approach
- modelling variable substitution rates across the genome
- amino acid substitution models
- codon substitution models

# Summary Markov chain model for nucleotide substitutions.

# Why Markov chains are a great model for nucleotide substitutions



- memorylessness: a nucleotide substitution happens independently from the substitution history at this site
- substitution rate matrix defines the transition probabilities
- the transition probabilities take into account every possible substitution path

# Why Markov chains are a great model for nucleotide substitutions

- memorylessness: a nucleotide substitution happens independently from the substitution history at this site
- substitution rate matrix defines the transition probabilities
- the transition probabilities take into account every possible substitution path

- applying theories of linear algebra we can calculate the transition probability matrix according to:
  $P(t) = e^{Qt} = U \, \mathrm{diag}(e^{\epsilon_1 t}, e^{\epsilon_2 t}, e^{\epsilon_3 t}, e^{\epsilon_4 t}) U^{-1}$
  - cookbook recipe provided on moodle
  - for further information on how to derive this formula please consult a textbook on linear algebra

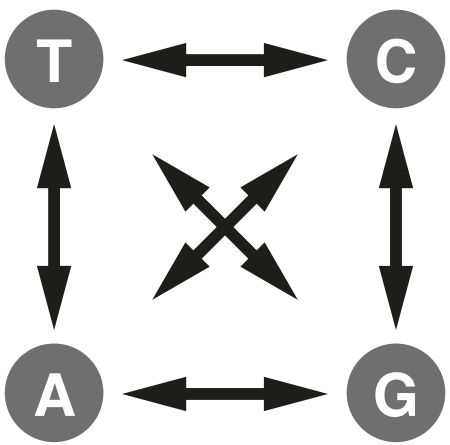
Substitution rate matrices and their properties.

# The easiest substitution model: JC69



**JC69:**

- named after TH Jukes, CR Cantor: Evolution of protein molecules. 1969 [Jukes and Cantor, 1969].
- all substitution have the same rate, $\lambda$




Substitution rate matrix:

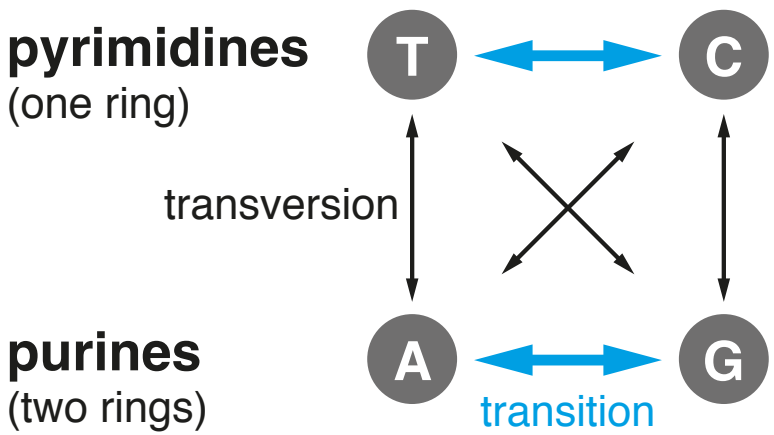
$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{pmatrix} \end{array}$$

# Accounting for transition/transversion: K80



**K80:**

- named after M Kimura: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. 1980. [Kimura, 1980]
- transitions happen at rate $\alpha$, transversions at rate $\beta$




Substitution rate matrix:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{pmatrix} \end{array}$$

# Including equilibrium frequencies: TN93 and HKY



**TN93:**

- named after [Tamura and Nei, 1993]
- transitions between $\mathrm{T} \leftrightarrow \mathrm{C}$ happen at rate $\alpha_1 \times$(nucleotide equilibrium frequency)
- transitions from $\mathrm{A} \leftrightarrow \mathrm{G}$ happen at rate $\alpha_2 \times$(nucleotide equilibrium frequency)
- transversions happen at rate $\beta \times$(nucleotide equilibrium frequency)

Substitution rate matrix:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha_1\pi_\mathrm{C} & \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} \\ \alpha_1\pi_\mathrm{T} & \cdot & \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} \\ \beta\pi_\mathrm{T} & \beta\pi_\mathrm{C} & \cdot & \alpha_2\pi_\mathrm{G} \\ \beta\pi_\mathrm{T} & \beta\pi_\mathrm{C} & \alpha_2\pi_\mathrm{A} & \cdot \end{pmatrix} \end{array}$$

# Including equilibrium frequencies: TN93 and HKY





**TN93:**

- named after [Tamura and Nei, 1993]
- transitions between $\mathrm{T} \leftrightarrow \mathrm{C}$ happen at rate $\alpha_1 \times$(nucleotide equilibrium frequency)
- transitions from $\mathrm{A} \leftrightarrow \mathrm{G}$ happen at rate $\alpha_2 \times$(nucleotide equilibrium frequency)
- transversions happen at rate $\beta \times$(nucleotide equilibrium frequency)

Substitution rate matrix:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & \alpha_1\pi_\mathrm{C} & \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} \\ \alpha_1\pi_\mathrm{T} & \cdot & \beta\pi_\mathrm{A} & \beta\pi_\mathrm{G} \\ \beta\pi_\mathrm{T} & \beta\pi_\mathrm{C} & \cdot & \alpha_2\pi_\mathrm{G} \\ \beta\pi_\mathrm{T} & \beta\pi_\mathrm{C} & \alpha_2\pi_\mathrm{A} & \cdot \end{pmatrix} \end{array}$$

- Note that the special case of $\alpha_1 = \alpha_2$ was described earlier [Hasegawa et al., 1984] and is named HKY

# A more general substitution model: GTR



**GTR (REV):**

- generalised time-reversible model
- based on three papers:
  [Tavaré, 1986, Yang, 1994, Zharkikh, 1994]

Substitution rate matrix:

$$\begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix} \end{array}$$

\+ quite flexible

\+ time-reversible

\- not completely general

# Time reversibility

- mathematical condition, making calculations and simulations easier

$$\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix}$$

$$= \begin{pmatrix} \cdot & a & b & c \\ a & \cdot & d & e \\ b & d & \cdot & f \\ c & e & f & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

# The most general substitution model



**UNREST:**

- unrestricted model first described in [Yang, 1994]
- each substitution has a (different) rate

Substitution rate matrix:

$$\begin{array}{c} \\ \mathtt{T} \\ \mathtt{C} \\ \mathtt{A} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{T} & \mathtt{C} & \mathtt{A} & \mathtt{G} \end{array} \\ \begin{pmatrix} \cdot & a & b & c \\ d & \cdot & e & f \\ g & h & \cdot & i \\ j & k & l & \cdot \end{pmatrix} \end{array}$$

\+ most general case

\+ all other models are special cases of UNREST

\- mathematically very complicated and not handy to use

\- not time-reversible

- Further models described in [Yang, 1994]

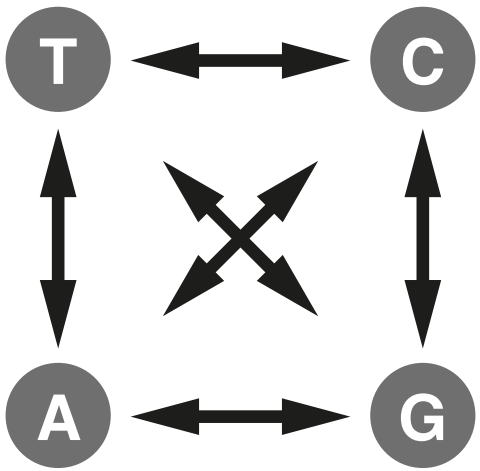
# Substitution models



| model | parameters | description |
|---|---|---|
| JC69 | 1 | all substitutions have the same rate |
| K80 | 2 | accounts for transition and transversions |
| HKY | 2+3* | distinction between transition and transversions, including equilibrium frequencies |
| TN93 | 3+3* | different rates for transitions |
| GTR | 6+3* | general, but still time-reversible |
| UNREST | 12 | most general, not time-reversible |

* equilibrium frequencies of nucleotides

# Calculating transition probabilities and sequence distance.




Substitution rate matrix:

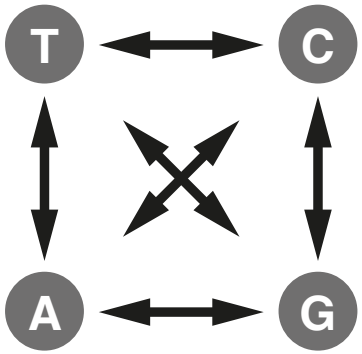
$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

# Example of transition probabilities: JC69



Substitution rate matrix:

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$



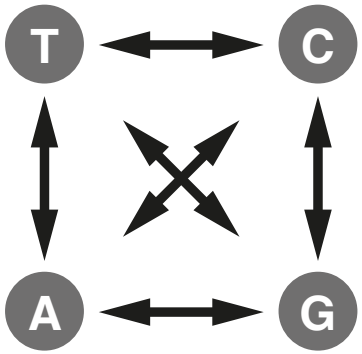

$$\Downarrow \quad P(t) = e^{Qt}$$

# Example of transition probabilities: JC69



Substitution rate matrix:

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$




$$\Downarrow \quad P(t) = e^{Qt}$$

transition probability matrix*:

$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$
and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

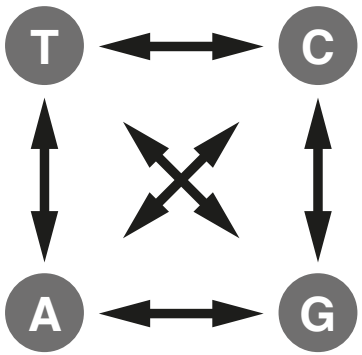
*cookbook recipe provided on moodle (for a proof of the matrix exponentiation please consult a textbook on linear algebra)

# Example of transition probabilities: JC69



Substitution rate matrix:

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$




$$\Downarrow \quad P(t) = e^{Qt}$$

transition probability matrix*:

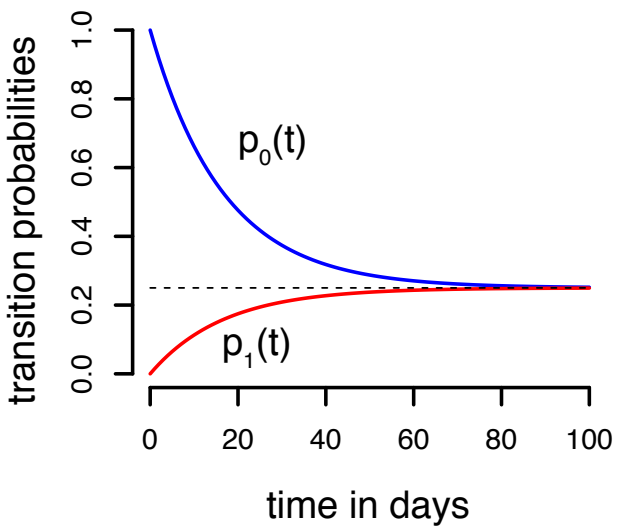
$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$
and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

*cookbook recipe provided on moodle (for a proof of the matrix exponentiation please consult a textbook on linear algebra)

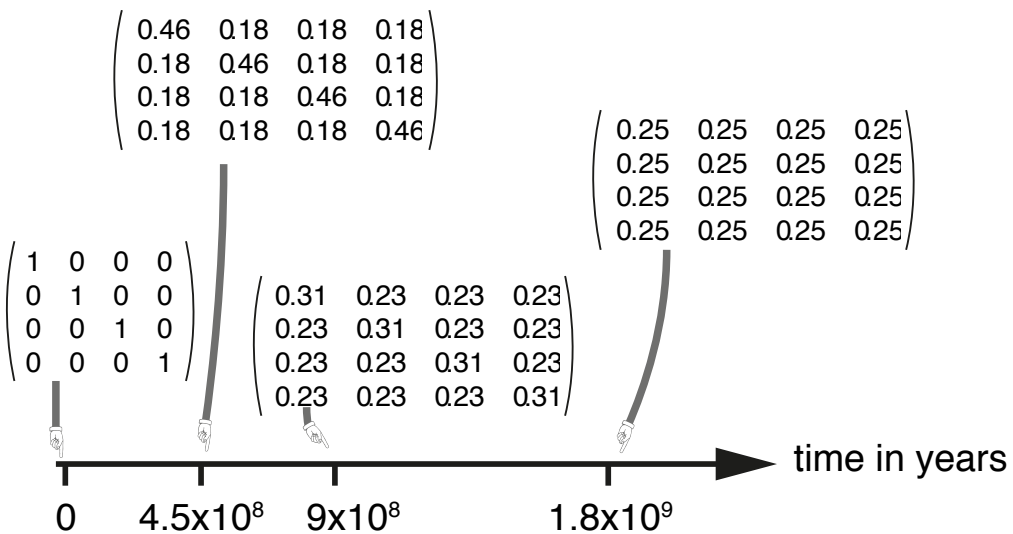
Example:

$\lambda = 0.015 \frac{\text{substitutions per site}}{\text{day}}$



# JC69: Stationary distribution

- sequence TCAG evolves according to JC69
- $\lambda = 2.2/3 \times 10^{-9} \frac{\text{substitutions per site}}{\text{year}}$
- What is the probability to observe a certain nucleotide $a_i \in \{T, C, A, G\}$ at position $i \in \{1, 2, 3, 4\}$ after time $t$?



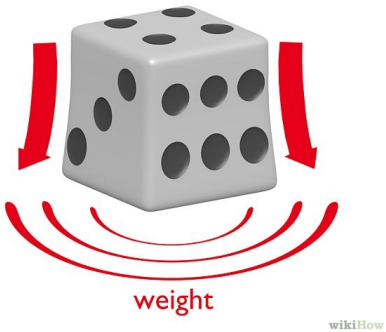

- when $t \to \infty$ stationary distribution is reached
- Any long sequence (e.g. TTTTTT...) at time 0, will be composed of equal amounts of T,C,A,G after time $t \to \infty$

# Maximum likelihood estimators of sequence distance.

# The idea behind maximum likelihood estimators



**Example:** We want to estimate the probability with which a six sided die shows 6. (Is the die loaded (unfair)?)



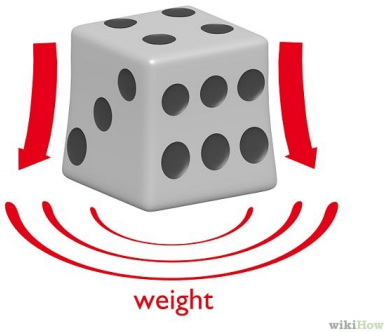

[loaded.dice, 2015]

- we throw the same die $\mathfrak{n} = 100$ times and note how many times we obtain a 6 (example: $\mathfrak{x} = 40$ times)

# The idea behind maximum likelihood estimators



**Example:** We want to estimate the probability with which a six sided die shows 6. (Is the die loaded (unfair)?)




[loaded.dice, 2015]

- we throw the same die $n = 100$ times and note how many times we obtain a 6 (example: $x = 40$ times)

**Strategy:**

- we use this data to obtain an estimate for the probability of throwing a 6
- we denote the probability of throwing a 6 with $p$; the probability of throwing $x$ times 6 out of $n$ tries is Binomial-distributed, i.e.
  $P(x \text{ times } 6) = \binom{n}{x} p^x (1-p)^{n-x}$

# The idea behind maximum likelihood estimators



**Strategy:**

- We use this data to obtain an estimate for the probability of throwing a 6
- We denote the probability of throwing a 6 with $p$. The probability of throwing $x$ times 6 out of $n$ tries is Binomial-distributed, i.e.
  $P(x \text{ times } 6) = \binom{n}{x} p^x (1-p)^{n-x}$
- As we know $x$ and $n$ (this is our data), we can interpret this probability as a function of the unknown $p$:
  $L(p; x) = P(x|p) = \binom{100}{40} p^{40} (1-p)^{60}$

# The idea behind maximum likelihood estimators



**Strategy:**

- We use this data to obtain an estimate for the probability of throwing a 6
- We denote the probability of throwing a 6 with $p$. The probability of throwing $x$ times 6 out of $n$ tries is Binomial-distributed, i.e.
  $P(x \text{ times } 6) = \binom{n}{x} p^x (1-p)^{n-x}$
- As we know $x$ and $n$ (this is our data), we can interpret this probability as a function of the unknown $p$:
  $L(p; x) = P(x|p) = \binom{100}{40} p^{40} (1-p)^{60}$
- **trick:** find the maximum of $L(p; x)$ in $p$ (necessary condition: first derivative set to 0)

# The idea behind maximum likelihood estimators



**Strategy:**

- We use this data to obtain an estimate for the probability of throwing a 6
- We denote the probability of throwing a 6 with $p$. The probability of throwing $x$ times 6 out of $n$ tries is Binomial-distributed, i.e.
  $P(x \text{ times } 6) = \binom{n}{x} p^x (1-p)^{n-x}$
- As we know $x$ and $n$ (this is our data), we can interpret this probability as a function of the unknown $p$:
  $L(p; x) = P(x|p) = \binom{100}{40} p^{40} (1-p)^{60}$
- **trick:** find the maximum of $L(p; x)$ in $p$ (necessary condition: first derivative set to 0)

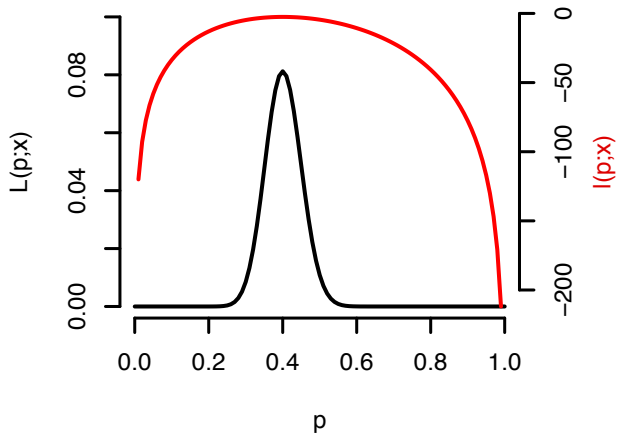
**Maximum likelihood estimator:** is an estimator of a model parameter that maximises the probability to obtain the observed results

# Mathematical tricks to calculate the MLE



sometimes it is handy to apply some transformation to $\mathrm{L}(\mathrm{p};\mathrm{x})$ that does not change the position of the maximum (but maybe the height):

- looking for the maximum of the log-likelihood function: $\mathrm{l}(\mathrm{p};\mathrm{x}) = \log(\mathrm{L}(\mathrm{p};\mathrm{x}))$
- multiplying $\mathrm{L}$ with a constant



# Throwing dice



coming back to our example:

- throw a die $\mathfrak{n} = 100$-times. $\mathfrak{x} = 40$ times the result is 6

# Throwing dice



coming back to our example:

- throw a die $n = 100$-times. $x = 40$ times the result is 6

The likelihood function is:
$L(p; x) = \binom{n}{x} p^x (1-p)^{n-x}$

# Throwing dice



coming back to our example:

- throw a die $n = 100$-times. $x = 40$ times the result is 6

The likelihood function is:
$L(p; x) = \binom{n}{x} p^x (1-p)^{n-x}$

Thus
$l(p; x) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p)$

Now we derive $l$ with respect to $p$ and set the derivative to 0:
$\frac{dl}{dp}(\hat{p}; x, n) = \frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} \stackrel{!}{=} 0 \quad \Leftrightarrow \hat{p} = \frac{x}{n}$

This means that we estimate the probability to throw a 6 to be $\frac{40}{100} = 0.4$ which is much higher than $\frac{1}{6} = 0.1667$.

# Confidence intervals



- interval that tries to capture the uncertainty of a parameter estimate

# Confidence intervals



- interval that tries to capture the uncertainty of a parameter estimate

**Definition: Confidence interval:** If we repeatedly estimated the parameter from realisations of the random experiment and the interval estimate for each realisation of the random experiment, we could expect 95% of these intervals to contain the true parameter.

definition adapted from [Sokal and Rohlf, 2012]

# Confidence intervals



- interval that tries to capture the uncertainty of a parameter estimate

**Definition: Confidence interval:** If we repeatedly estimated the parameter from realisations of the random experiment and the interval estimate for each realisation of the random experiment, we could expect 95% of these intervals to contain the true parameter.

definition adapted from [Sokal and Rohlf, 2012]

- note that this interval is an estimate itself based on a realisation of a random experiment

# Confidence intervals



Confidence intervals for the parameter estimate can be calculated based on **likelihood intervals**:

- technique based on likelihood ratio test

# Confidence intervals



Confidence intervals for the parameter estimate can be calculated based on **likelihood intervals**:

- technique based on likelihood ratio test
- Let $X$ be a random variable with a distribution parameterised in $\theta$. Based on collected data $x$ of a huge sample, the MLE for the parameter is $\hat{\theta}$. Then, one can show (under certain conditions) that, $2(l(\hat{\theta}) - l(\theta))$ has a $\chi^2_k$-distribution, where $k$ are the degrees of freedom (the vector length of $\theta$)

# Confidence intervals



Confidence intervals for the parameter estimate can be calculated based on **likelihood intervals**:

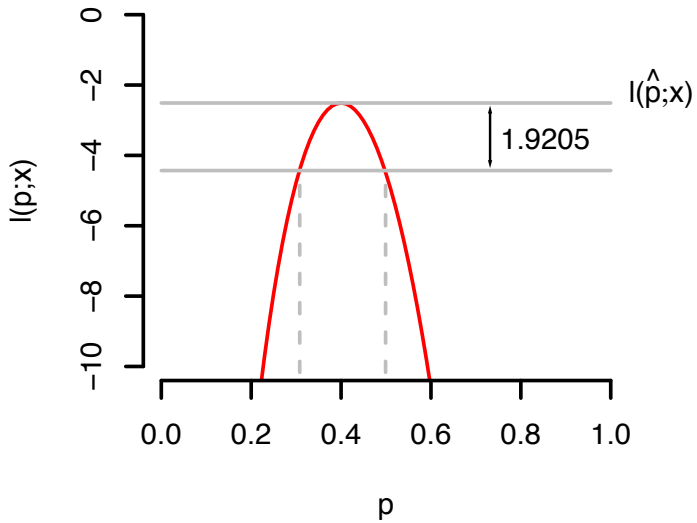
- technique based on likelihood ratio test
- Let $X$ be a random variable with a distribution parameterised in $\theta$. Based on collected data $x$ of a huge sample, the MLE for the parameter is $\hat{\theta}$. Then, one can show (under certain conditions) that, $2(l(\hat{\theta}) - l(\theta))$ has a $\chi^2_k$-distribution, where $k$ are the degrees of freedom (the vector length of $\theta$)
- Idea: the 95% CI includes all $\theta$s which are not in the 0.05 tail of the $\chi^2$ distribution, i.e. which would not be rejected in a statistical test with $\alpha = 0.05$. This means that when performing the experiment e.g. 100 times, then 5 times the estimated parameter is expected not to be contained in the 95% CI.

# Confidence intervals

- Let $X$ be a random variable with a distribution parameterised in $\theta$. Based on collected data $x$ of a huge sample, the MLE for the parameter is $\hat{\theta}$. Then, $2(l(\hat{\theta}) - l(\theta)) \sim \chi^2_k$.

**Strategy** to obtain a 95% confidence interval (CI):

- determine the value of the log-likelihood function in $\hat{\theta}$, $l(\hat{\theta}; x)$
- subtract $0.5\chi^2_{k,5\%}$*, i.e. calculate $l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$.
- determine those $\theta$ values for which $l(\theta; x) = l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$



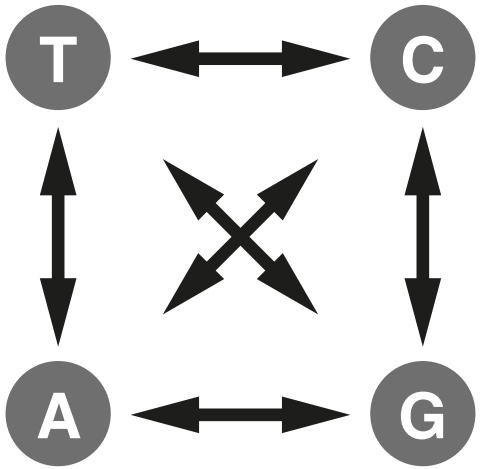

CI: $[0.308, 0.499]$

*A $\chi^2$ table can be found on the moodle page.

# JC69: MLE for sequence distance



We will apply the MLE framework to derive an estimator for sequence distance under JC69.




Transition probability matrix $\mathbf{P}(\mathbf{t}) =$:

$$\begin{pmatrix} \mathbf{p}_0(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) \\ \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_0(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) \\ \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_0(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) \\ \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_1(\mathbf{t}) & \mathbf{p}_0(\mathbf{t}) \end{pmatrix}$$

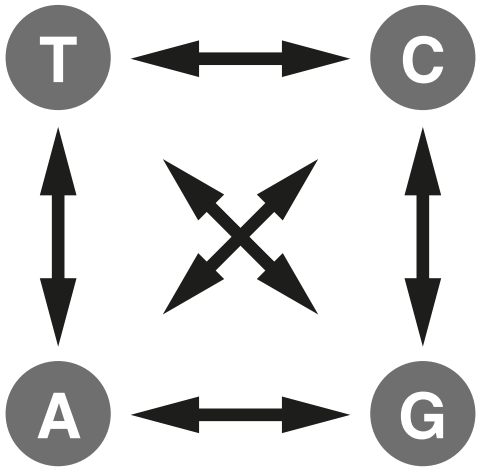
with $\mathbf{p}_0(\mathbf{t}) = \frac{1}{4} + \frac{3}{4}\mathbf{e}^{-4\lambda\mathbf{t}}$

$\mathbf{p}_1(\mathbf{t}) = \frac{1}{4} - \frac{1}{4}\mathbf{e}^{-4\lambda\mathbf{t}}$

Imagine we have two sequences of length $\mathbf{n}$ with $\mathbf{x}$ differences. The probability that a position is different is $\mathbf{p} = 3\mathbf{p}_1(\mathbf{t})$. We define $\mathbf{d} = 3\lambda\mathbf{t}$ (the expected distance in time $\mathbf{t}$).

# JC69: MLE for sequence distance

We will apply the MLE framework to derive an estimator for sequence distance under JC69.




Transition probability matrix $P(t) =$:

$$\begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$

$p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

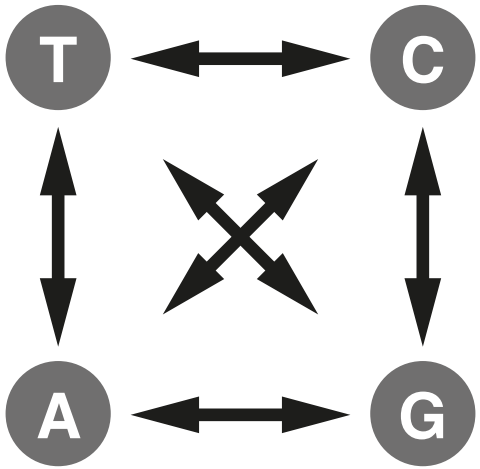
Imagine we have two sequences of length $n$ with $x$ differences. The probability that a position is different is $p = 3p_1(t)$. We define $d = 3\lambda t$ (the expected distance in time $t$).

Thus the probability that $x$ positions out of $n$ are different is:

$$\binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x}$$

# JC69: MLE for sequence distance

We will apply the MLE framework to derive an estimator for sequence distance under JC69.




Transition probability matrix $P(t) =$:

$$\begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$

$p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

Imagine we have two sequences of length $n$ with $x$ differences. The probability that a position is different is $p = 3p_1(t)$. We define $d = 3\lambda t$ (the expected distance in time $t$).

Thus the probability that $x$ positions out of $n$ are different is:

$$\binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x} = L(d; x)$$

# JC69: MLE for sequence distance



$$\binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3} d}\right)^x \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3} d}\right)^{n-x} = L(d; x)$$

From this equation we obtain the log-likelihood function:

$$l(d; x) = \log \binom{n}{x} + x \log \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3} d}\right) + (n - x) \log \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3} d}\right)$$

# JC69: MLE for sequence distance



$$\binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3} d}\right)^x \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3} d}\right)^{n-x} = L(d; x)$$

From this equation we obtain the log-likelihood function:

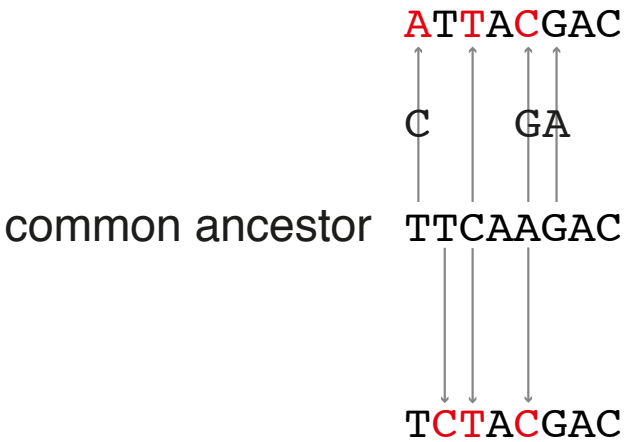
$$l(d; x) = \log \binom{n}{x} + x \log \left(\frac{3}{4} - \frac{3}{4} e^{-\frac{4}{3} d}\right) + (n - x) \log \left(\frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3} d}\right)$$

Setting the first derivative to 0 and solving this equation with respect to $d$ leads to the MLE of the JC69 distance:

$$\hat{d} = -\frac{3}{4} \log \left(1 - \frac{4x}{3n}\right)$$

# JC69: Example



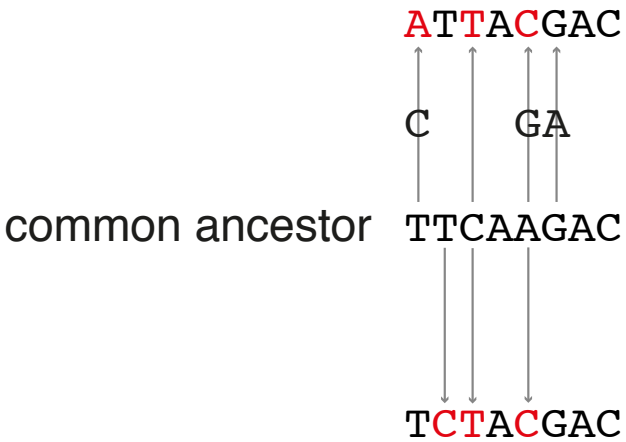



- length of gene: $n = 8$
- differences between the two sequences: $x = 2$

$$\Rightarrow \hat{d} = -\frac{3}{4}\log\left(1 - \frac{4\times 2}{3\times 8}\right) = 0.3$$

# JC69: Example



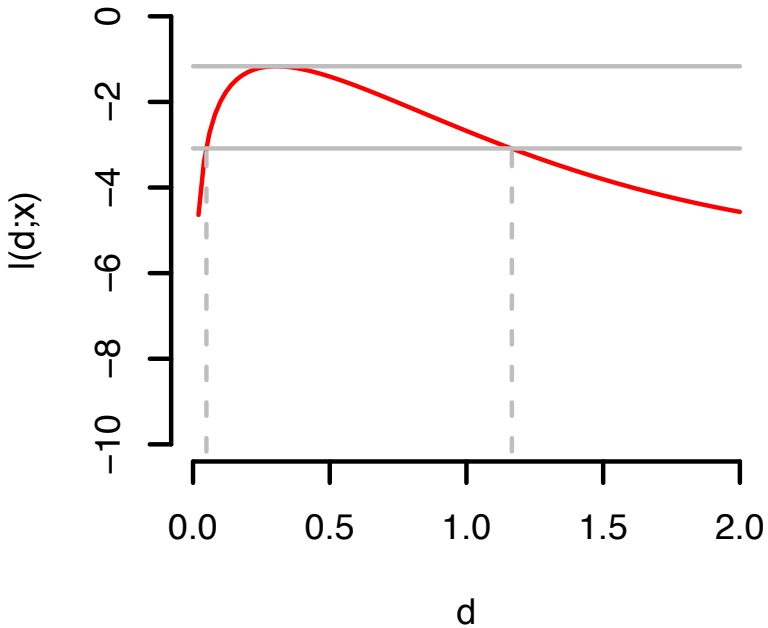



- length of gene: $n = 8$
- differences between the two sequences: $x = 2$

$$\Rightarrow \hat{d} = -\tfrac{3}{4}\log\left(1 - \tfrac{4\times 2}{3\times 8}\right) = 0.3$$

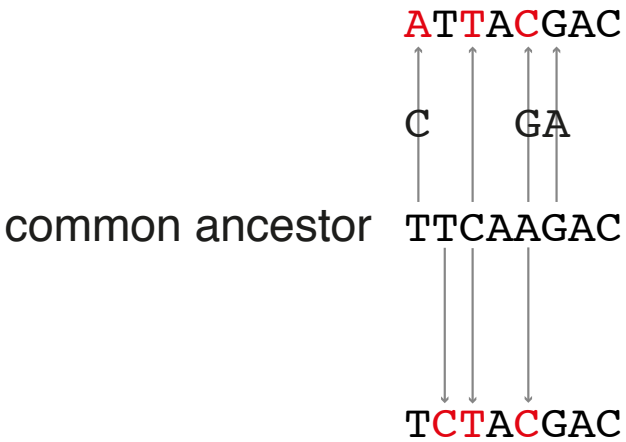
Determination of the 95% confidence interval:

$x = 2$, $n = 8$: $[0.05, 1.17]$



# JC69: Example



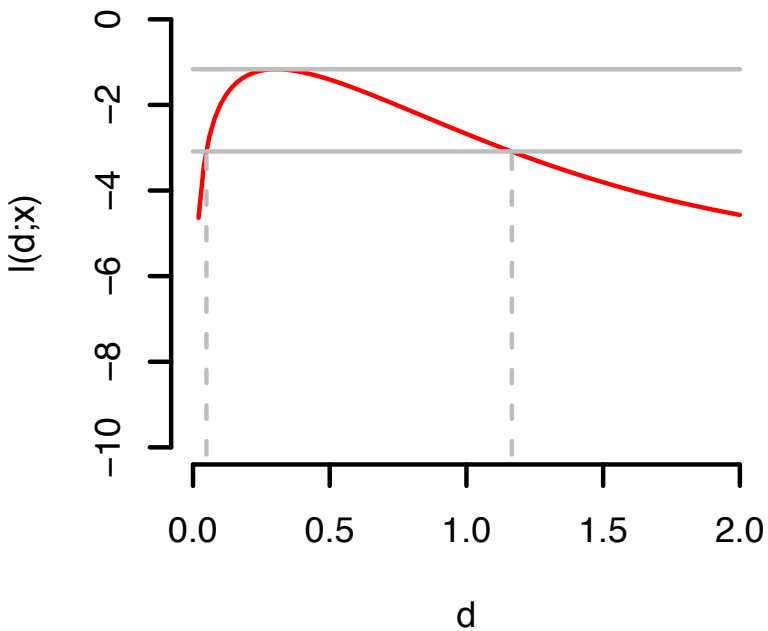



- length of gene: $\mathfrak{n} = 8$
- differences between the two sequences: $\mathfrak{x} = 2$

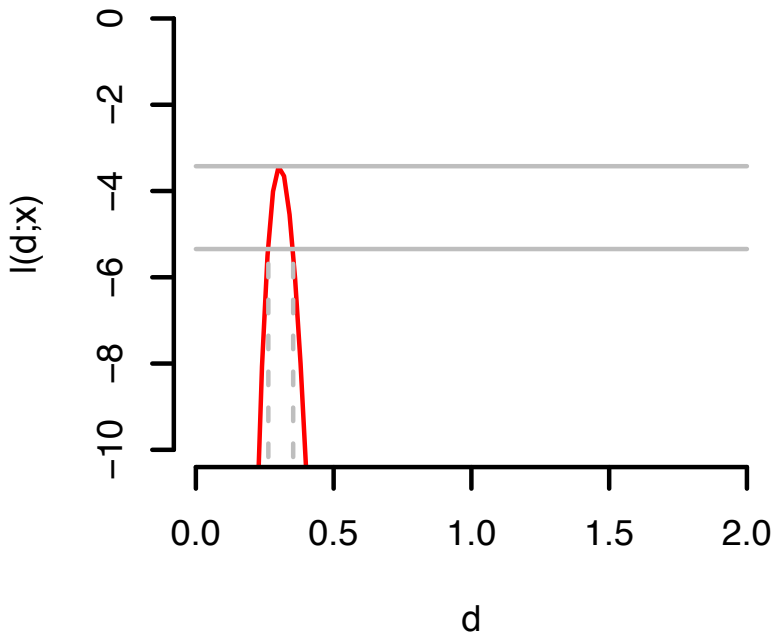
$$\Rightarrow \hat{\mathfrak{d}} = -\tfrac{3}{4}\log\left(1 - \tfrac{4\times 2}{3\times 8}\right) = 0.3$$

Determination of the 95% confidence interval:

$\mathfrak{x} = 2, \mathfrak{n} = 8$: $[0.05, 1.17]$




$\mathfrak{x} = 200, \mathfrak{n} = 800$: $[0.26, 0.35]$



# Variable substitution rates across the genome.

# Variable rates



- so far: all sites in the sequence evolve at the same rate
- but: substitution rates might differ across the genome
  - mutation rates might differ across sites
  - selective pressure might be different on the phenotypic level

# Variable rates



- so far: all sites in the sequence evolve at the same rate
- but: substitution rates might differ across the genome
  - mutation rates might differ across sites
  - selective pressure might be different on the phenotypic level

We extend the existing models, by replacing the constant rates by $\Gamma$-distributed random variables (notation: JC69+$\Gamma$, K80+$\Gamma$, ... )

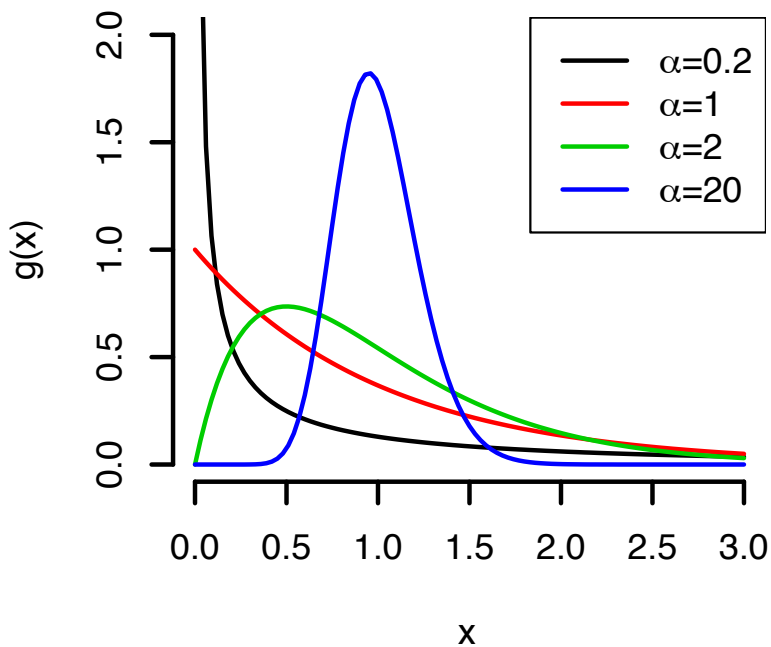
- sequence distance: expected number of substitutions per site, averaged over all sites

# Toolbox: The Γ-distribution

- probability distribution: $g(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}$, $x \geqslant 0$
- Γ-function: $\Gamma(\alpha) = \int_0^{\infty} e^{-t} t^{\alpha-1} dt$; $\Gamma(n) = (n-1)!$
- a $\Gamma(\alpha, \beta)$-distributed random variable $X$ has mean $E[X] = \frac{\alpha}{\beta}$ and variance $var[X] = \frac{1}{\alpha}$

Here, we fix the mean of the distribution to be 1, this is the case if and only if $\alpha = \beta$

☞ Γ-distribution very flexible



# JC69+Γ



- we replace the substitution rate $\lambda$ by $\lambda R$, where $R$ is a $\Gamma(\alpha, \alpha)$-distributed random variable (! mean 1)
- JC 69: $p = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$
  JC69+Γ: $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$

# JC69+Γ



- we replace the substitution rate $\lambda$ by $\lambda R$, where $R$ is a $\Gamma(\alpha, \alpha)$-distributed random variable (! mean 1)
- JC 69: $p = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$
  JC69+Γ: $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$

  To obtain the probability of a substitution at one site, we calculate $\mathbb{E}[p]$:
  $\mathbb{E}[p] = \int_0^\infty (\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dr})g(r)dr = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{\alpha}\right)^{-\alpha}$

# JC69+Γ

- we replace the substitution rate $\lambda$ by $\lambda R$, where $R$ is a $\Gamma(\alpha, \alpha)$-distributed random variable (! mean 1)
- JC 69: $p = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$
  JC69+Γ: $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$

  To obtain the probability of a substitution at one site, we calculate $\mathbb{E}[p]$:
  $\mathbb{E}[p] = \int_0^\infty (\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dr})g(r)dr = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{\alpha}\right)^{-\alpha}$
- to estimate the distance $d$, we use the MLE as described before with $l(d; x) = \binom{n}{x}(\mathbb{E}[p])^x(1 - \mathbb{E}[p])^{n-x}$

# JC69+Γ



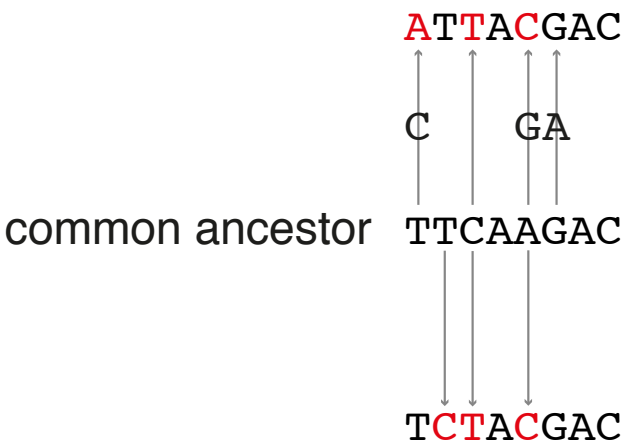
- we replace the substitution rate $\lambda$ by $\lambda R$, where $R$ is a $\Gamma(\alpha, \alpha)$-distributed random variable (! mean 1)
- JC 69: $p = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$
  JC69+Γ: $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$

  To obtain the probability of a substitution at one site, we calculate $\mathbb{E}[p]$:
  $\mathbb{E}[p] = \int_0^\infty (\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dr})g(r)dr = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{\alpha}\right)^{-\alpha}$
- to estimate the distance $d$, we use the MLE as described before with $l(d; x) = \binom{n}{x}(\mathbb{E}[p])^x(1 - \mathbb{E}[p])^{n-x}$

$$\hat{d} = \frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right)$$

# JC69+Γ: Example





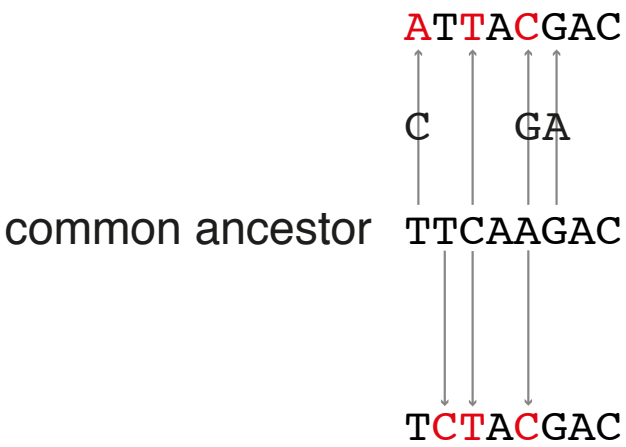

- length of gene: $n = 8$
- differences between the two sequences: $x = 2$

$\Rightarrow \hat{p} = 2/8 = 0.25$

- site variation: $\Gamma(2, 2)$-distributed, i.e. $\alpha = 2$

# JC69+Γ: Example






- length of gene: $n = 8$
- differences between the two sequences: $x = 2$

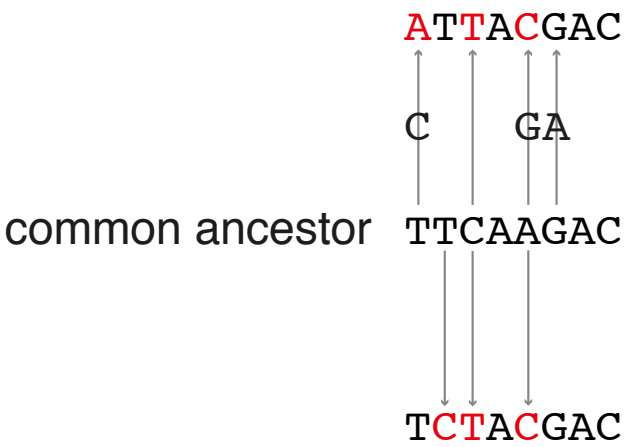
$\Rightarrow \hat{p} = 2/8 = 0.25$

- site variation: $\Gamma(2, 2)$-distributed, i.e. $\alpha = 2$

Thus: $\frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right) = 0.34$

# JC69+Γ: Example






- length of gene: $\mathfrak{n} = 8$
- differences between the two sequences: $\mathfrak{x} = 2$

$\Rightarrow \hat{\mathfrak{p}} = 2/8 = 0.25$

- site variation: $\Gamma(2, 2)$-distributed, i.e. $\alpha = 2$

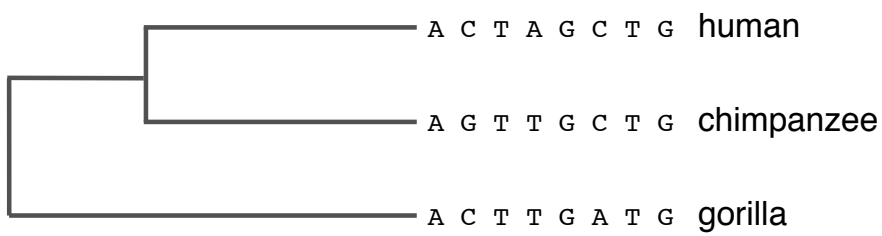
Thus: $\frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{\mathfrak{p}}\right)^{-1/\alpha} - 1\right) = 0.34$

This is bigger than $\hat{\mathfrak{d}} = 0.3$ in JC69 $\Rightarrow$

Ignoring site variation leads to underestimation of the sequence distance

# Distance usage: distance based phylogenetic reconstruction

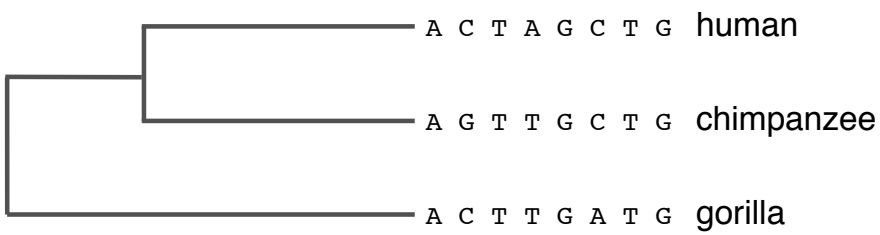
We can now replace the Hamming-distance with the evolutionary distance for distance based phylogenetic reconstruction:



# Distance usage: distance based phylogenetic reconstruction



We can now replace the Hamming-distance with the evolutionary distance for distance based phylogenetic reconstruction:




- here we only derived distance estimators for the most simple nucleotide substitution model (JC69), other substitution rate models would lead to other distances estimated
- not all phylogenetic reconstruction methods are based on distances (other methods will be discussed from lecture 5 on)

# Amino acid substitution models.

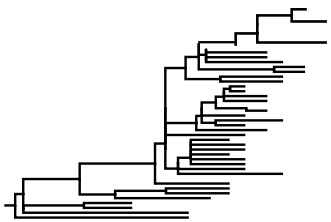
# Levels of evolution



**genotype**

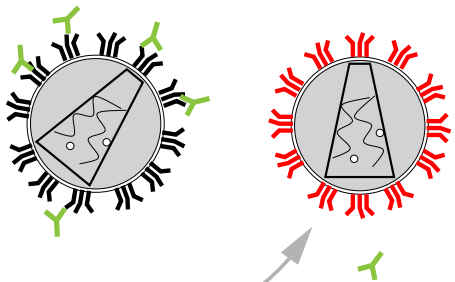
sequence level

ACUGAACGUGACUACUG
ACUGAACGUAACUACUG



**phenotype**

e.g. antigenic level: Antibody binding to HIV



**codon**: three nucleotides encode for one amino acid

one nucleotide change can already change the phenotype

**alphabet:**

4 nucleotides: DNA: TCAG
RNA: UCAG

$64=4^3$ codons

20 amino acids

# Method for estimating distance between two strains

In general, the same Markov model is used for amino acid substitutions as for nucleotide substitutions. The transition probability matrix can be calculated according to the equation

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Of course, in case of AA substitutions, $\mathbf{P}(t)$ as well as the substitution rate matrix have dimension $20 \times 20$. To determine $\mathbf{P}(t)$, we need to derive the $\mathbf{Q}$-matrix:

# Method for estimating distance between two strains

In general, the same Markov model is used for amino acid substitutions as for nucleotide substitutions. The transition probability matrix can be calculated according to the equation

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Of course, in case of AA substitutions, $\mathbf{P}(t)$ as well as the substitution rate matrix have dimension $20 \times 20$. To determine $\mathbf{P}(t)$, we need to derive the $\mathbf{Q}$-matrix:

- more difficult than for nucleotide substitutions
- two approaches
  - empiric
  - mechanistic
- desired: time-reversibility

# Distance estimation of AA sequences



Let us consider the easiest case of AA substitutions: All substitutions happen at the same rate $\lambda$.

- as we have 20 AA, the substitution rate for any substitution is then $19\lambda$
- the expected time until a substitution happens is then $\frac{1}{19\lambda}$
- with $\mathbf{t} = \frac{\mathbf{d}}{19\lambda}$ we can also derive the distance with the maximum likelihood approach

# Distance estimation of AA sequences



Let us consider the easiest case of AA substitutions: All substitutions happen at the same rate $\lambda$.

- as we have 20 AA, the substitution rate for any substitution is then $19\lambda$
- the expected time until a substitution happens is then $\frac{1}{19\lambda}$
- with $\mathrm{t} = \frac{\mathrm{d}}{19\lambda}$ we can also derive the distance with the maximum likelihood approach

$$\hat{\mathrm{d}} = -\frac{19}{20} \log\left(1 - \frac{20\mathrm{x}}{19\mathrm{n}}\right)$$

# Distance estimation of AA sequences



Let us consider the easiest case of AA substitutions: All substitutions happen at the same rate $\lambda$.

- as we have 20 AA, the substitution rate for any substitution is then $19\lambda$
- the expected time until a substitution happens is then $\frac{1}{19\lambda}$
- with $\mathrm{t} = \frac{\mathrm{d}}{19\lambda}$ we can also derive the distance with the maximum likelihood approach

$$\hat{\mathrm{d}} = -\frac{19}{20} \log\left(1 - \frac{20\mathrm{x}}{19\mathrm{n}}\right)$$

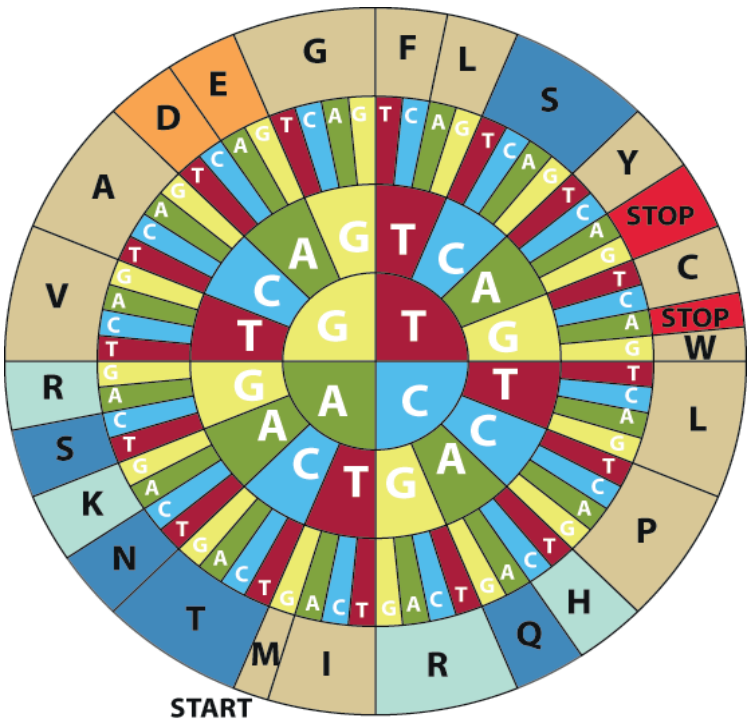
- to estimate the distance with an empirical or mechanistic Q-matrix, we also use a MLE

# Codon substitution models.

# The codon sun



A codon consists of three nucleotides, translating to one of the 20 amino acids:



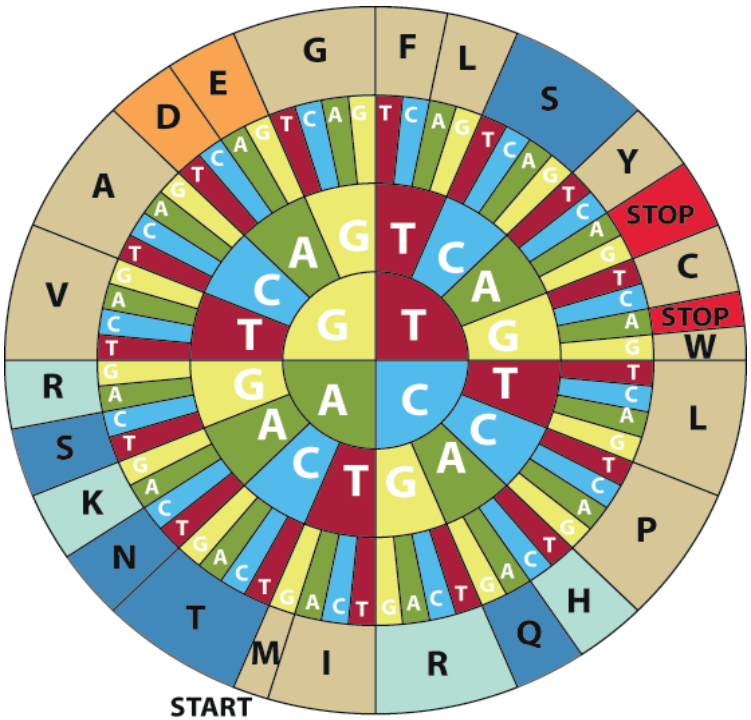

[Sanger, 2015]

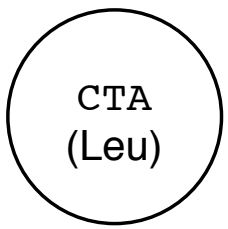
| Amino Acid | Three-Letter Abbreviation | One-Letter Symbol | Molecular Weight |
|---|---|---|---|
| Alanine | Ala | A | 89Da |
| Arginine | Arg | R | 174Da |
| Asparagine | Asn | N | 132Da |
| Aspartic acid | Asp | D | 133Da |
| Asparagine or aspartic acid | Asx | B | 133Da |
| Cysteine | Cys | C | 121Da |
| Glutamine | Gln | Q | 146Da |
| Glutamic acid | Glu | E | 147Da |
| Glutamine or glutamic acid | Glx | Z | 147Da |
| Glycine | Gly | G | 75Da |
| Histidine | His | H | 155Da |
| Isoleucine | Ile | I | 131Da |
| Leucine | Leu | L | 131Da |
| Lysine | Lys | K | 146Da |
| Methionine | Met | M | 149Da |
| Phenylalanine | Phe | F | 165Da |
| Proline | Pro | P | 115Da |
| Serine | Ser | S | 105Da |
| Threonine | Thr | T | 119Da |
| Tryptophan | Trp | W | 204Da |
| Tyrosine | Tyr | Y | 181Da |
| Valine | Val | V | 117Da |

[Promega, 2015]

# Example: Codon CTA

Which one-point mutations lead to the codon CTA for leucine?

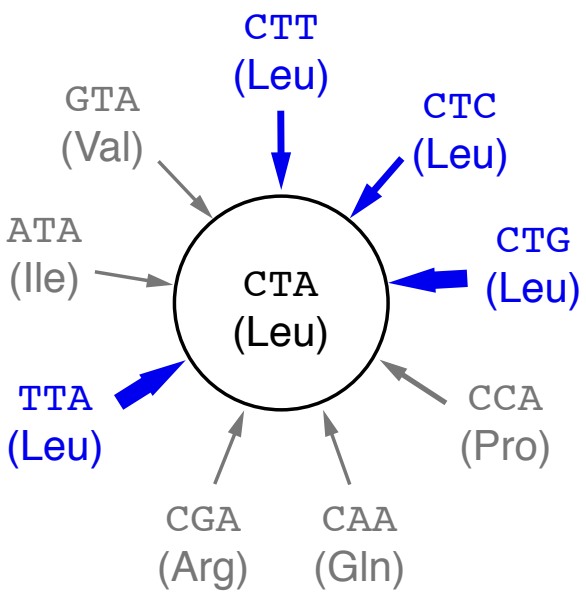



[Sanger, 2015]

# Example: Codon CTA

Overview over substitution rates to the same codon CTA, the thickness of arrows represent different rates:




- **synonymous substitutions:** AA does not change
- **nonsynonymous substitutions:** AA does change
- bigger arrows: transition
- smaller arrows: transversion

adapted from [Yang, 2014]

# The basic model



We also model codon substitution with a Markov chain model, where 61 states (no stop codons) are allowed:

- The substitution rate matrix has dimension $61 \times 61$

# The basic model



We also model codon substitution with a Markov chain model, where 61 states (no stop codons) are allowed:

- The substitution rate matrix has dimension $61 \times 61$
- We denote codons with capital letters, e.g. $\mathtt{I}, \mathtt{J}$, but nucleotides with small letters, e.g. $\mathtt{i}, \mathtt{j}$

# The basic model

We also model codon substitution with a Markov chain model, where 61 states (no stop codons) are allowed:

- The substitution rate matrix has dimension $61 \times 61$
- We denote codons with capital letters, e.g. $\mathrm{I}, \mathrm{J}$, but nucleotides with small letters, e.g. $\mathrm{i}, \mathrm{j}$
- We incorporate:
  - $\kappa$: transition/transversion rate ratio
  - $\omega$: nonsynonymous/synonymous rate ratio
  - $\pi_{\mathrm{I}}$: equilibrium frequency of codon $\mathrm{I}$ consisting of nucleotides $\mathrm{i}_1\mathrm{i}_2\mathrm{i}_3$, with equilibrium frequencies $\pi_{\mathrm{i}_1}, \pi_{\mathrm{i}_2}, \pi_{\mathrm{i}_3}$ $\pi_{\mathrm{I}} = \frac{1}{\mathrm{C}}\pi^*_{\mathrm{i}_1}\pi^*_{\mathrm{i}_2}\pi^*_{\mathrm{i}_3}$

# The basic model

We also model codon substitution with a Markov chain model, where 61 states (no stop codons) are allowed:

- The substitution rate matrix has dimension $61 \times 61$
- We denote codons with capital letters, e.g. $\mathrm{I}, \mathrm{J}$, but nucleotides with small letters, e.g. $\mathrm{i}, \mathrm{j}$
- We incorporate:
  - $\kappa$: transition/transversion rate ratio
  - $\omega$: nonsynonymous/synonymous rate ratio
  - $\pi_{\mathrm{I}}$: equilibrium frequency of codon $\mathrm{I}$ consisting of nucleotides $\mathrm{i}_1\mathrm{i}_2\mathrm{i}_3$, with equilibrium frequencies $\pi_{\mathrm{i}_1}, \pi_{\mathrm{i}_2}, \pi_{\mathrm{i}_3}$ $\pi_{\mathrm{I}} = \frac{1}{\mathrm{C}}\pi^*_{\mathrm{i}_1}\pi^*_{\mathrm{i}_2}\pi^*_{\mathrm{i}_3}$

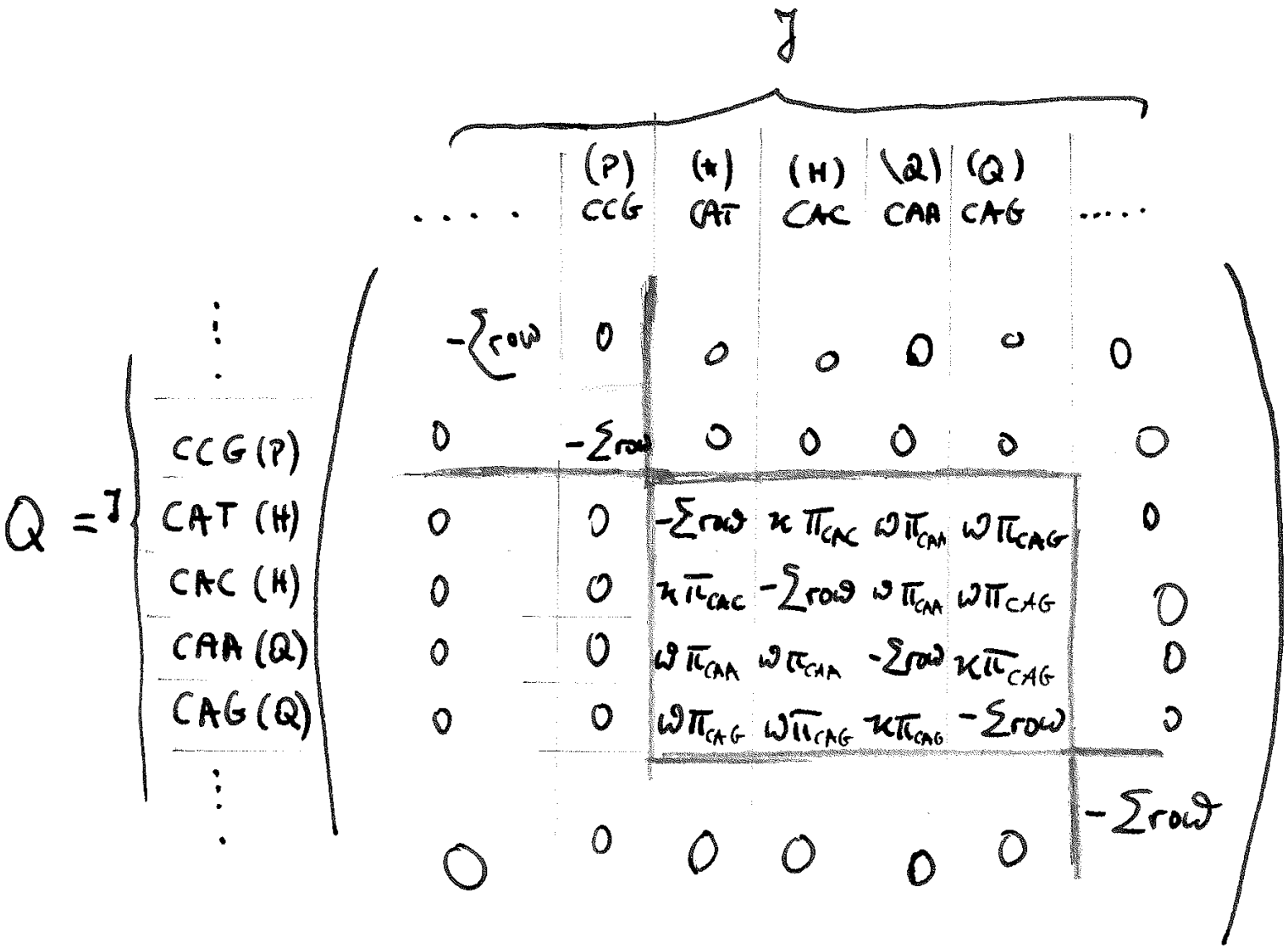
Model:

$$q_{\mathrm{IJ}} = \begin{cases} 0 & \text{if I and J differ at more than 1 positions} \\ \pi_{\mathrm{J}} & \text{if I and J differ by a synonymous transversion} \\ \kappa\pi_{\mathrm{J}} & \text{if I and J differ by a synonymous transition} \\ \omega\pi_{\mathrm{J}} & \text{if I and J differ by a nonsynonymous transversion} \\ \omega\kappa\pi_{\mathrm{J}} & \text{if I and J differ by a nonsynonymous transition} \end{cases}$$

# The substitution rate matrix

# The substitution rate matrix





# Evidence for selection



- synonymous substitutions do not change the protein, thus, these substitutions are seen as neutral
- nonsynonymous substitutions do change the protein and selective processes can act on the new protein

# Evidence for selection



- synonymous substitutions do not change the protein, thus, these substitutions are seen as neutral
- nonsynonymous substitutions do change the protein and selective processes can act on the new protein
- to discover selection, one compares amounts of nonsynonymous and synonymous substitutions

# Evidence for selection



- synonymous substitutions do not change the protein, thus, these substitutions are seen as neutral
- nonsynonymous substitutions do change the protein and selective processes can act on the new protein
- to discover selection, one compares amounts of nonsynonymous and synonymous substitutions

$d_N$: distance at nonsynonymous codon positions
$d_S$: distance at synonymous codon positions

# $d_N/d_S$ ratio: Counting method



In the following we will look at a method based on counting synonymous and non-synonymous substitutions. We will illustrate this method with the two sequences `TTTCCTCCTCCT` and `TTCCAGCCTCCT`:

# $d_N/d_S$ ratio: Counting method



In the following we will look at a method based on counting synonymous and non-synonymous substitutions. We will illustrate this method with the two sequences `TTTCCTCCTCCT` and `TTCCAGCCTCCT`:

| | codon 1 | codon 2 | codon 3 | codon 4 |
|---|---|---|---|---|
| sequence 1 | `TTT` | `CCT` | `CCT` | `CCT` |
| sequence 2 | `TTC` | `CAG` | `CCT` | `CCT` |

# $d_N/d_S$ ratio: Counting method



In the following we will look at a method based on counting synonymous and non-synonymous substitutions. We will illustrate this method with the two sequences `TTTCCTCCTCCT` and `TTCCAGCCTCCT`:

| | codon 1 | codon 2 | codon 3 | codon 4 |
|---|---|---|---|---|
| sequence 1 | TTT | CCT | CCT | CCT |
| | F | P | P | P |
| | F | Q | P | P |
| sequence 2 | TTC | CAG | CCT | CCT |

# $d_N/d_S$ ratio: Counting method



In the following we will look at a method based on counting synonymous and non-synonymous substitutions. We will illustrate this method with the two sequences `TTTCCTCCTCCT` and `TTCCAGCCTCCT`:

| | codon 1 | codon 2 | codon 3 | codon 4 |
|---|---|---|---|---|
| sequence 1 | TTT | CCT | CCT | CCT |
| | F | P | P | P |
| | F | Q | P | P |
| sequence 2 | TTC | CAG | CCT | CCT |

**Algorithm:**

1. count the (non-) synonymous differences
2. count the (non-) synonymous sites
3. account for the possible evolutionary history

# Step 1: number of (non-) synonymous differences



$N_d$ : number of nonsynonymous differences

$S_d$ : number of synonymous differences

. . . taking into account all possible ways from seq1 to seq2

| | codon 1 | | codon 2 | | codon 3 | | codon 4 | |
|---|---|---|---|---|---|---|---|---|
| seq1 | TTT | | CCT | | CCT | | CCT | |
| seq2 | TTC | | CAG | | CCT | | CCT | |
| $N_d$ = | 0 | + | | + | 0 | + | 0 | = |
| $S_d$ = | 1 | + | | + | 0 | + | 0 | = |

# Step 1: number of (non-) synonymous differences

$N_d$ : number of nonsynonymous differences

$S_d$ : number of synonymous differences

. . . taking into account all possible ways from seq1 to seq2

| | codon 1 | | codon 2 | | codon 3 | | codon 4 | |
|---|---|---|---|---|---|---|---|---|
| seq1 | TTT | | CCT | | CCT | | CCT | |
| seq2 | TTC | | CAG | | CCT | | CCT | |
| $N_d$ = | 0 | + | | + | 0 | + | 0 | = |
| $S_d$ = | 1 | + | | + | 0 | + | 0 | = |

| pathway | $S_d$ at codon 2 | $N_d$ at codon 2 |
|---|---|---|
| CCT (P) → CAT (H) → CAG (Q) | | |
| CCT (P) → CCG (P) → CAG (Q) | | |
| average | | |

# Step 1: number of (non-) synonymous differences

$N_d$ : number of nonsynonymous differences

$S_d$ : number of synonymous differences

. . . taking into account all possible ways from seq1 to seq2

| | codon 1 | | codon 2 | | codon 3 | | codon 4 | |
|---|---|---|---|---|---|---|---|---|
| seq1 | TTT | | CCT | | CCT | | CCT | |
| seq2 | TTC | | CAG | | CCT | | CCT | |
| $N_d$ = | 0 | + | | + | 0 | + | 0 | = |
| $S_d$ = | 1 | + | | + | 0 | + | 0 | = |

| pathway | $S_d$ at codon 2 | $N_d$ at codon 2 |
|---|---|---|
| CCT (P) → CAT (H) → CAG (Q) | 0 | 2 |
| CCT (P) → CCG (P) → CAG (Q) | 1 | 1 |
| average | 0.5 | 1.5 |

# Step 1: number of (non-) synonymous differences

$N_d$ : number of nonsynonymous differences

$S_d$ : number of synonymous differences

. . . taking into account all possible ways from seq1 to seq2

| | codon 1 | | codon 2 | | codon 3 | | codon 4 | |
|---|---|---|---|---|---|---|---|---|
| seq1 | TTT | | CCT | | CCT | | CCT | |
| seq2 | TTC | | CAG | | CCT | | CCT | |
| $N_d$ = | 0 | + | 1.5 | + | 0 | + | 0 | = 1.5 |
| $S_d$ = | 1 | + | 0.5 | + | 0 | + | 0 | = 1.5 |

| pathway | $S_d$ at codon 2 | $N_d$ at codon 2 |
|---|---|---|
| CCT (P) → CAT (H) → CAG (Q) | 0 | 2 |
| CCT (P) → CCG (P) → CAG (Q) | 1 | 1 |
| average | 0.5 | 1.5 |

# Step 2: number of (non-) synonymous sites



$\mathcal{N}$ : number of nonsynonymous sites

$\mathcal{S}$ : number of synonymous sites

. . . averaging over all potential one point mutations

**important:** the sum of (non-) synonymous mutations per codon must sum up to 3

| | cdn1 | cdn2 | cdn3 | cdn4 | |
|---|---|---|---|---|---|
| seq1 | TTT | CCT | CCT | CCT | |
| seq2 | TTC | CAG | CCT | CCT | |
| nonsyn | | | | | |
| seq1 | | + | + | + | = |
| seq2 | | + | + | + | = |
| average | | | | | $\mathcal{N}$ = |
| syn | | | | | |
| seq1 | | + | + | + | = |
| seq2 | | + | + | + | = |
| average | | | | | $\mathcal{S}$ = |

# Step 2: number of (non-) synonymous sites

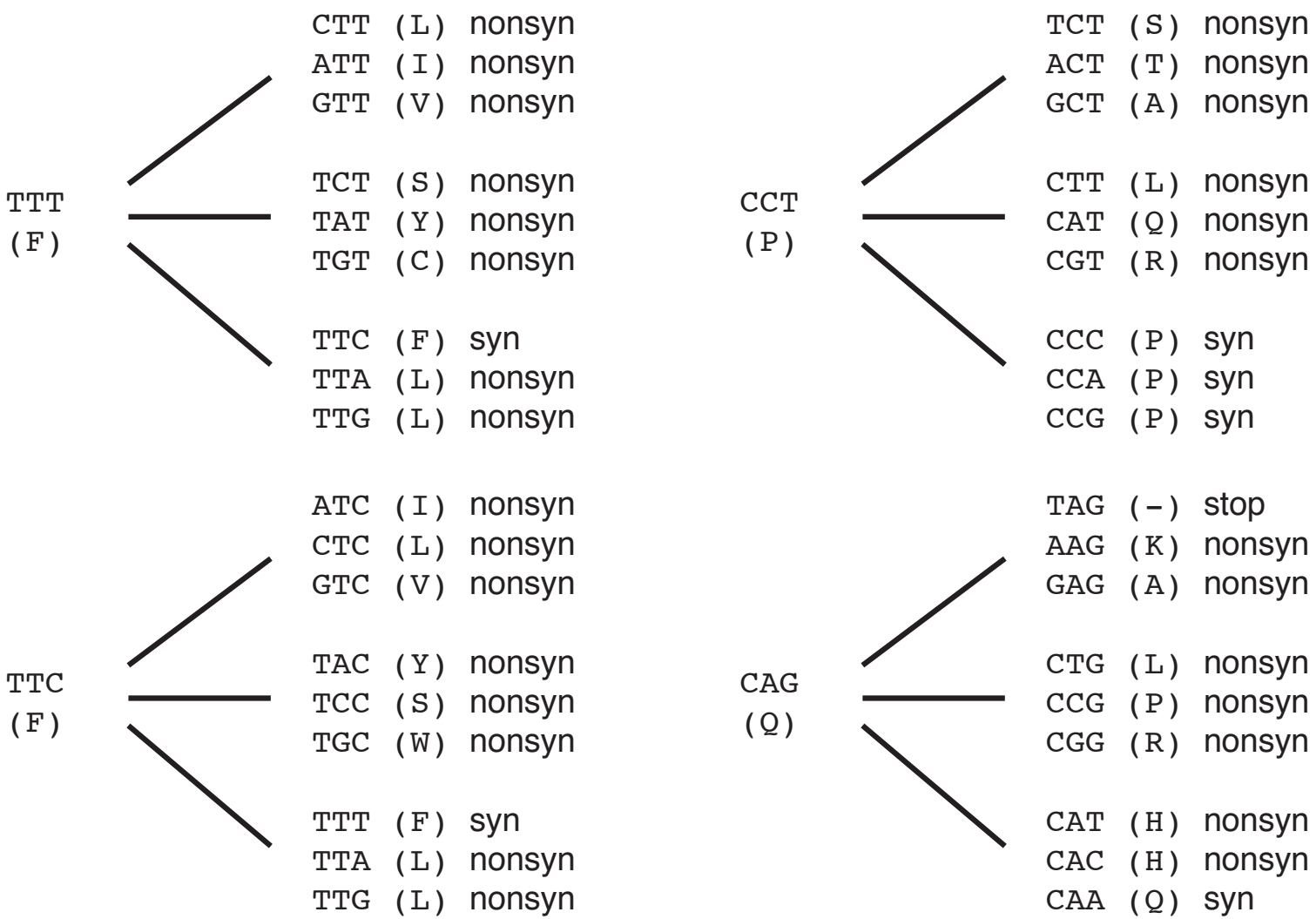


| | cdn1 | cdn2 | cdn3 | cdn4 |
|---|---|---|---|---|
| seq1 | TTT | CCT | CCT | CCT |
| seq2 | TTC | CAG | CCT | CCT |

# Step 2: number of (non-) synonymous sites



| | cdn1 | cdn2 | cdn3 | cdn4 |
|---|---|---|---|---|
| seq1 | TTT | CCT | CCT | CCT |
| seq2 | TTC | CAG | CCT | CCT |



# Step 2: number of (non-) synonymous sites



| | cdn1 | | cdn2 | | cdn3 | | cdn4 | |
|---|---|---|---|---|---|---|---|---|
| seq1 | TTT | | CCT | | CCT | | CCT | |
| seq2 | TTC | | CAG | | CCT | | CCT | |
| nonsyn | | | | | | | | |
| seq1 | 8/3 | + | 2 | + | 2 | + | 2 | = 8.67 |
| seq2 | 8/3 | + | 21/8 | + | 2 | + | 2 | = 9.29 |
| average | | | | | | | | $\mathbf{N} = 8.98$ |
| syn | | | | | | | | |
| seq1 | 1/3 | + | 1 | + | 1 | + | 1 | = 3.33 |
| seq2 | 1/3 | + | 3/8 | + | 1 | + | 1 | = 2.71 |
| average | | | | | | | | $S = 3.02$ |

# Step 3: accounting for evolution



Under the JC69 substitution model, we can now calculate the distances at (non-)synonymous codon positions, $\mathrm{d_N}$ and $\mathrm{d_S}$:

$$\begin{aligned} \mathrm{d_N} &= -\frac{3}{4}\log\left(1-\frac{4}{3}\frac{\mathrm{N_d}}{\mathrm{N}}\right) \\ \mathrm{d_S} &= -\frac{3}{4}\log\left(1-\frac{4}{3}\frac{\mathrm{S_d}}{\mathrm{S}}\right) \end{aligned}$$

# Step 3: accounting for evolution

Under the JC69 substitution model, we can now calculate the distances at (non-)synonymous codon positions, $d_N$ and $d_S$:

$$d_N = -\frac{3}{4}\log\left(1-\frac{4}{3}\frac{N_d}{N}\right)$$

$$d_S = -\frac{3}{4}\log\left(1-\frac{4}{3}\frac{S_d}{S}\right)$$

In our example, the $d_N/d_S$-ratio is:

$$\frac{d_N}{d_S} = \frac{-\frac{3}{4}\log\left(1-\frac{4}{3}\frac{N_d}{N}\right)}{-\frac{3}{4}\log\left(1-\frac{4}{3}\frac{S_d}{S}\right)} = \frac{\log\left(1-\frac{4}{3}\frac{1.5}{8.98}\right)}{\log\left(1-\frac{4}{3}\frac{1.5}{3.02}\right)} = 0.23$$

# The interpretation of the $d_N/d_S$-ratio



$d_N/d_S < 1$: nonsynonymous mutations occur less frequently than synonymous mutations (purifying selection)

$d_N/d_S > 1$: nonsynonymous mutations occur more frequently than synonymous mutations (positive selection)

# The interpretation of the $d_N/d_S$-ratio



$d_N/d_S < 1$: nonsynonymous mutations occur less frequently than synonymous mutations (purifying selection)

$d_N/d_S > 1$: nonsynonymous mutations occur more frequently than synonymous mutations (positive selection)

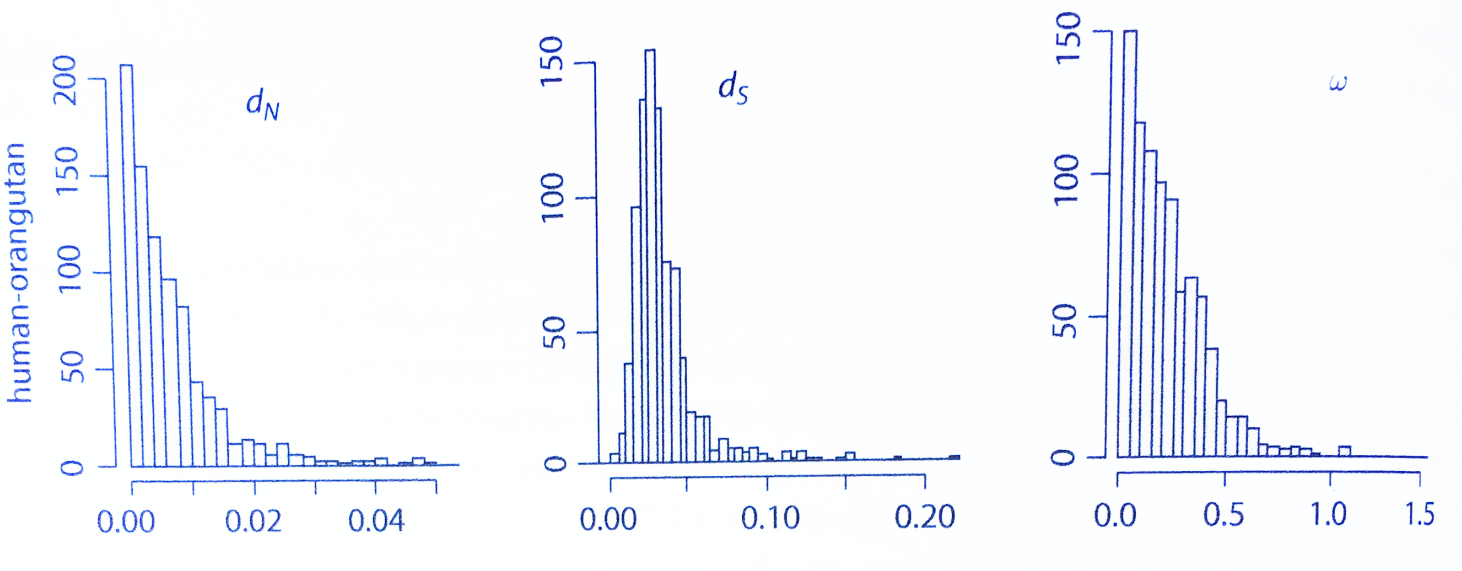
☞ In addition, ML methods have been developed. However:

> *"It is worth noting that the estimation of $d_S$ and $d_N$ is a complicated exercise, as manifested by the fact that even the simple transition/transversion rate difference is nontrivial to deal with. Unfortunately, different methods can often produce very different estimates." [Yang, 2014]*

# Example: $d_N$, $d_S$ between humans and orangutans



Histograms of $d_N$, $d_S$ and $\omega = \frac{d_N}{d_S}$ for 857 orthologous genes between humans and orangutans:




adapted from Figure 2.6 [Yang, 2014]

# What we learnt today



- nucleotide substitution models
- maximum likelihood estimators for distances
- modeling site variation with $\Gamma$-distribution

- amino acid substitution models

- codon substitution models
  - $\mathrm{d_N}$ and $\mathrm{d_S}$ as a tool to study selection

# Molecular evolution, AA and codon substitution models: Questions

- What are the differences between the JC69 and the TN93 models?
- Which of the two models would you chose if you were to perform a phylogenetic analysis based on sequence distances?
- In lecture 3 we tried to naively reconstruct a phylogeny based on three sequences (the same tree and sequences appear on slide 33 of this lecture). The pairwise Hamming distances were all the same. Would you expect that any of the presented nucleotide sequence models would result in different trees?

# References I




- Hasegawa, M., Yano, T., and Klishino, H. (1984). A New Molecular Clock of Mitochondrial-Dna and the Evolution of Hominoids. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences*, 60(4):95–98.
- Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism.*, pages 21–123.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.
- loaded.dice (2015). How to load dice: http://rss2.com/feeds/How-to-of-the-Day/page51.
- Promega (2015). The amino acids: https://www.promega.com/ /media/files/resources/technical references/amino acid abbreviations and molecular weights.pdf.
- Sanger (2015). The codon sun: ftp://ftp.sanger.ac.uk/pub/yourgenome/downloads/activities/kras-cancer-mutation/krascodonwheel.pdf.
- Sokal, R. and Rohlf, F. (2012). *Biometry. Fourth Edition.* W.H. Freeman and Company.
- Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology—DNA sequence analysis (New York, 1984)*, pages 57–86. Amer. Math. Soc., Providence, RI.
- Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111.
- Yang, Z. (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.
- Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–329.