Models of Sequence Evolution with Selection

Julien Dutheil

dutheil@evolbio.mpg.de

Max Planck Institute for Evolutionary Biology

June 18th 2015

Markov model

- We model the evolution of each position (site) of a sequence independently

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\Pr(X(t) = A) \quad = \quad \Pr(X(t_0) = A) \times \Pr(A \to A)$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned}\Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\ &+ \Pr(X(t_0) = C) \times \Pr(C \to A)\end{aligned}$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned}
\Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\
&+ \Pr(X(t_0) = C) \times \Pr(C \to A) \\
&+ \Pr(X(t_0) = G) \times \Pr(G \to A)
\end{aligned}$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned}\Pr(X(t)=A) &= \Pr(X(t_0)=A)\times\Pr(A\to A)\\ &+ \Pr(X(t_0)=C)\times\Pr(C\to A)\\ &+ \Pr(X(t_0)=G)\times\Pr(G\to A)\\ &+ \Pr(X(t_0)=T)\times\Pr(T\to A)\end{aligned} \tag{1}$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned}
\Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\
&+ \Pr(X(t_0) = C) \times \Pr(C \to A) \\
&+ \Pr(X(t_0) = G) \times \Pr(G \to A) \\
&+ \Pr(X(t_0) = T) \times \Pr(T \to A) \qquad (1)
\end{aligned}$$

- Similar equations can be written for $\Pr(X(t) = C)$, $\Pr(X(t) = G)$ and $\Pr(X(t) = T)$.

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

- And we can write

$$x(t) = x(t_0) \times \underbrace{\begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}}_{P}$$

where $p_{ij} = \Pr(i \rightarrow j)$.

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

- And we can write

$$x(t) = x(t_0) \times \underbrace{\begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}}_{P}$$

  where $p_{ij} = \Pr(i \rightarrow j)$.

- 'P' defines the *substitution process.*

A few more considerations

- We have

$$\forall i, \sum_j p_{i,j} = 1$$

that is

$$\Pr(A \to A) + \Pr(A \to C) + \Pr(A \to G) + \Pr(A \to T) = 1$$

- If we assume that all types of mutations are equi-probable (Jukes and Cantor, 1969), we can simplify:

$$P_{(JC69)} = \begin{pmatrix} 1-3r & r & r & r \\ r & 1-3r & r & r \\ r & r & 1-3r & r \\ r & r & r & 1-3r \end{pmatrix}$$

Continuous time

We assume that the process does not change over time, so we can write the equations for any time t:

$$t = t_0 + dt_0, \quad r = \alpha \cdot dt_0$$

$$x(t_0 + dt_0) = x(t_0) \times \begin{pmatrix} 1 - 3\alpha dt_0 & \alpha dt_0 & \alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & 1 - 3\alpha dt_0 & \alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & \alpha dt_0 & 1 - 3\alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & \alpha dt_0 & \alpha dt & 1 - 3\alpha dt_0 \end{pmatrix}$$

$$x(t_0 + dt_0) = x(t_0) + x(t_0) \cdot Q dt_0$$

$$\frac{x(t_0 + dt_0) - x(t_0)}{dt_0} = x(t_0) \cdot Q$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

- Q is called the *generator* of the substitution process, and we have

$$Q_{(JC69)} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

with

$$\forall i, \sum_j q_{i,j} = 0$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

- Q is called the *generator* of the substitution process, and we have

$$Q_{(JC69)} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

with

$$\forall i, \sum_j q_{i,j} = 0$$

- This resolves into

$$x(t) = x(t_0) \cdot \exp(Q \cdot t)$$

Conclusion

We can compute the probability that a certain sequence ($x(t_0)$) transforms into another given sequence ($x(t)$) after a known time (t) and given a certain substitution process specified by its generator (Q).

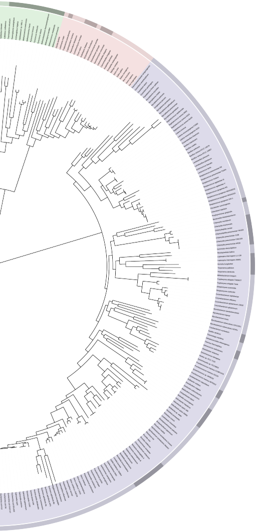
Conclusion

We can compute the probability that a certain sequence ($x(t_0)$) transforms into another given sequence ($x(t)$) after a known time (t) and given a certain substitution process specified by its generator (Q).

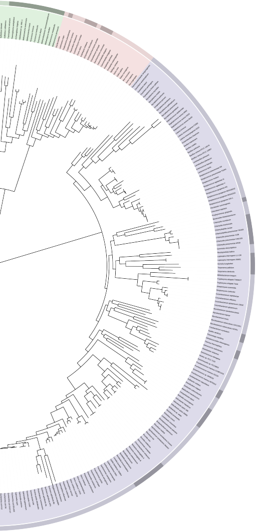
So what???

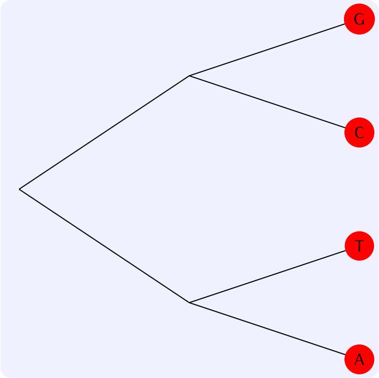
If we have two sequences and Q, we can compute t which maximizes this probability $\rightarrow$ unbiased estimate of the divergence between the two sequences!

Evolution along a tree

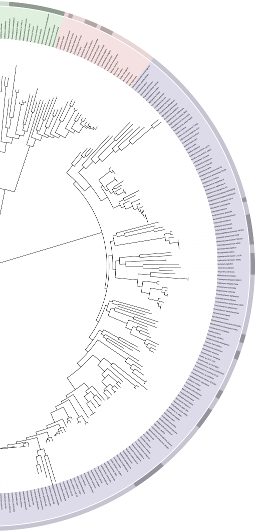


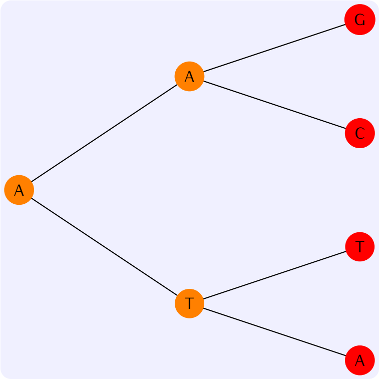
Evolution along a tree



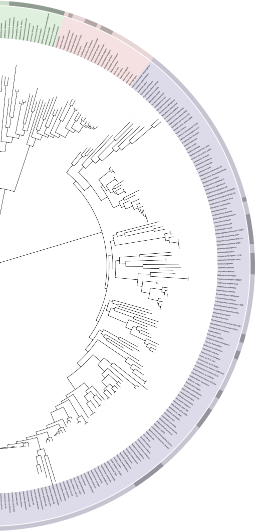


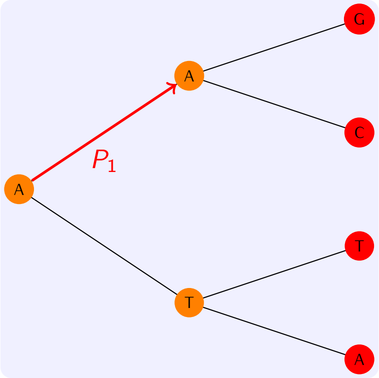
Evolution along a tree



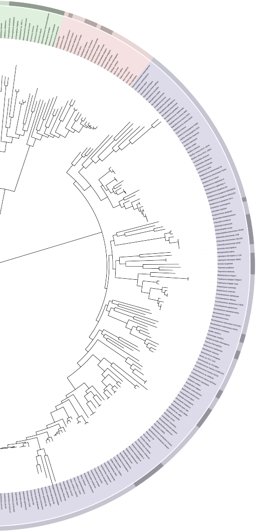


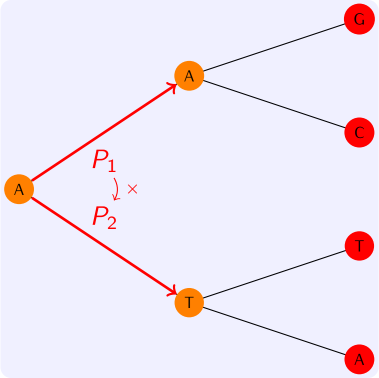
Evolution along a tree



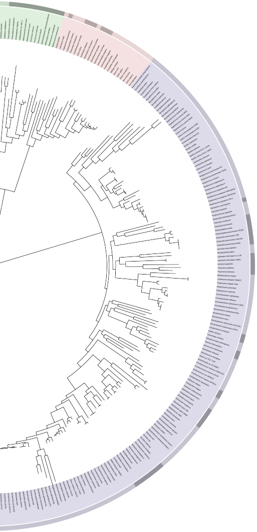


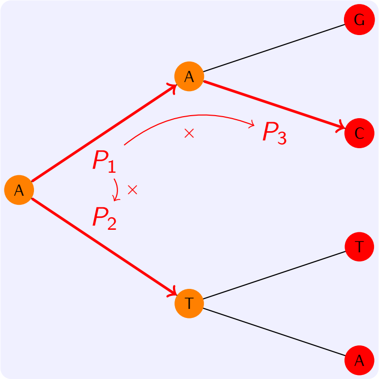
Evolution along a tree



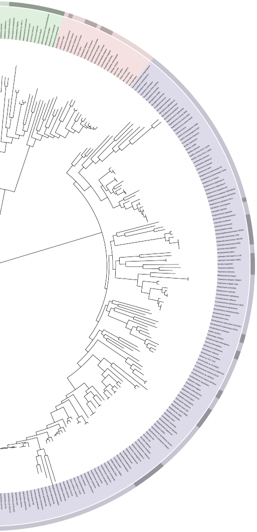


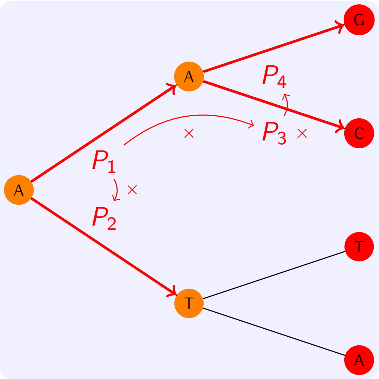
Evolution along a tree



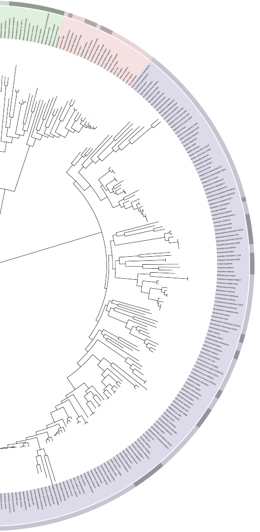


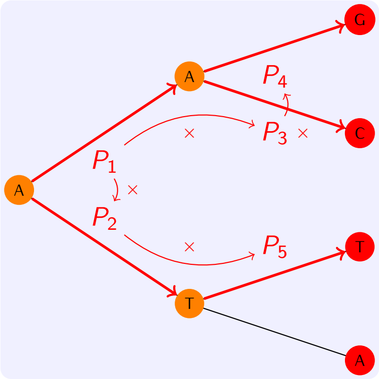
Evolution along a tree



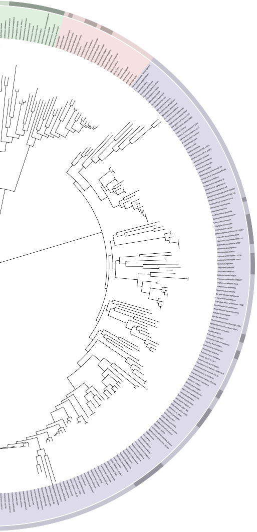


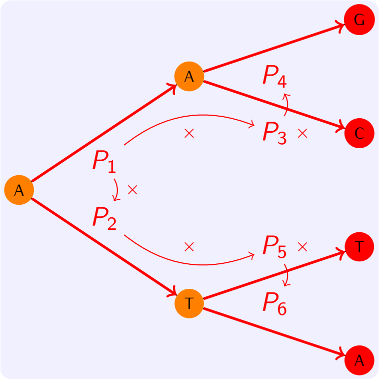
Evolution along a tree



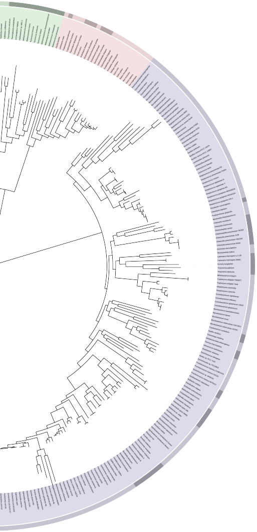


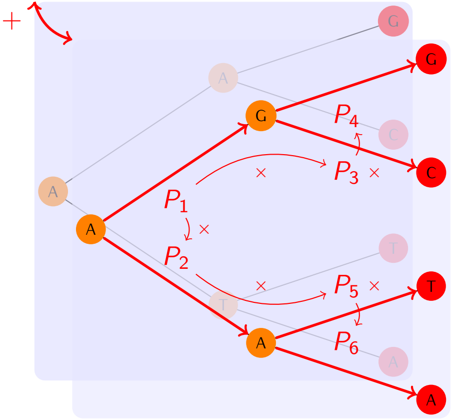
Evolution along a tree



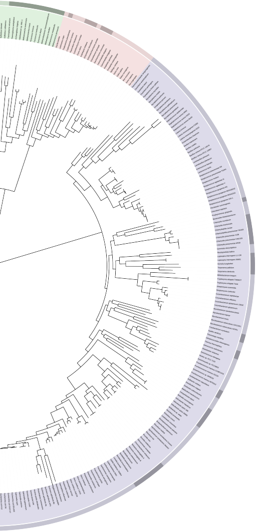


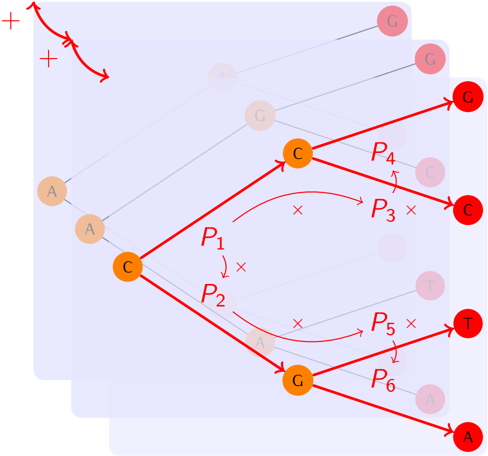
Evolution along a tree



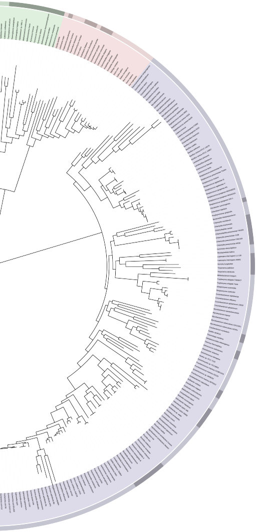


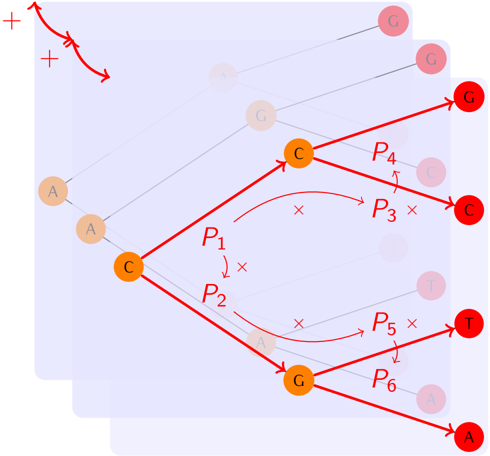
Evolution along a tree





Evolution along a tree





$$L_i = \sum_{\text{Ancestors}} P_1 \times P_2 \times P_3 \times P_4 \times P_5 \times P_6$$

Common nucleotide substitution models

Model	Authors	Parameters
JC69	Jukes Cantor	1 substitution rate
K80	Kimura	1 transition rate, 1 transversion rate
K81	Kimura	1 transition rate, 2 transversion rates
F81 = TN84	Felsenstein, Tajima et Nei	1 substitution rate and 3 frequencies
HKY85	Hasegawa, Kishino et Yano	1 transition rate, 1 transversion rate and 3 frequencies
TN93	Tamura et Nei	1 transition rate, 2 transversion rates and 3 frequencies
Z94	Zharkikh	6 substitution rates
T92	Tamura	1 transition rate, 1 transversion rate and 1 GC rate
GTR	"General time reversible"	6 substitution rate and 3 frequencies

Common nucleotide substitution models

$$\begin{tikzcd}
 & \begin{array}{c} v_1 = v_2 = v_3 = v_4 \\ s_1 = s_2 \end{array} & \begin{array}{c} v_1 = v_4 \\ v_2 = v_3 \\ s_1 = s_2 \end{array} & \begin{array}{c} v_1, v_2, v_3, v_4 \\ s_1, s_2 \end{array} & \\
\begin{array}{c} \text{JC69} \\ \text{1 parameter} \end{array} \arrow[r] \arrow[d] & \begin{array}{c} \text{K80} \\ \text{2 parameters} \end{array} \arrow[r] \arrow[d] & \begin{array}{c} \text{K81} \\ \text{3 parameters} \end{array} \arrow[r] \arrow[d] & \begin{array}{c} \text{Z94} \\ \text{6 parameters} \end{array} \arrow[d] & \begin{array}{l} \pi_{\mathtt{A}} = \pi_{\mathtt{U}} \\ = \pi_{\mathtt{G}} = \pi_{\mathtt{C}} \end{array} \\
\begin{array}{c} \text{TN84} \\ \text{4 parameters} \end{array} \arrow[r] & \begin{array}{c} \text{HKY85} \\ \text{5 parameters} \end{array} \arrow[r] \arrow[d] & \begin{array}{c} \text{TN93} \\ \text{6 parameters} \end{array} \arrow[r] & \begin{array}{c} \text{GTR} \\ \text{9 parameters} \end{array} & \pi_{\mathtt{A}}, \pi_{\mathtt{U}}, \pi_{\mathtt{G}}, \pi_{\mathtt{C}} \\
 & \begin{array}{c} \text{T92} \\ \text{3 parameters} \end{array} & \begin{array}{l} \pi_{\mathtt{A}} = \pi_{\mathtt{U}} \\ \pi_{\mathtt{G}} = \pi_{\mathtt{C}} \end{array} & &
\end{tikzcd}$$

Probability of an alignment

Site independence

If sites evolve independently:

$$\begin{aligned}\Pr(D|\Theta) &= \prod_i \Pr(D_i|\Theta) \\ L &= \prod_i L_i\end{aligned}$$

Probability of an alignment

Site independence

If sites evolve independently:

$$\begin{aligned} \Pr(D|\Theta) &= \prod_i \Pr(D_i|\Theta) \\ L &= \prod_i L_i \end{aligned}$$

Parameters

- Branch lengths
- Entries in the substitution matrix
- Tree topology

Maximum likelihood

Maximum-likelihood estimation (MLE)

MLE is a method of estimating the parameters of a statistical model. For a given dataset and underlying statistical model, the maximum likelihood estimator corresponds to the set of values of the model parameters that maximizes the likelihood function. (The method was initially proposed by statistician Ronald Aylmer Fisher in 1922.)

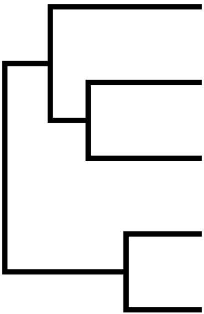
Maximum likelihood

Maximum-likelihood estimation (MLE)

MLE is a method of estimating the parameters of a statistical model. For a given dataset and underlying statistical model, the maximum likelihood estimator corresponds to the set of values of the model parameters that maximizes the likelihood function. (The method was initially proposed by statistician Ronald Aylmer Fisher in 1922.)

- General statistical framework
- Allows to perform model comparisons
- Allows to get confidence intervals of estimates

Modelling the evolution of a codon sequence alignment



ATG TGG CGT CGC TAT CCT ...

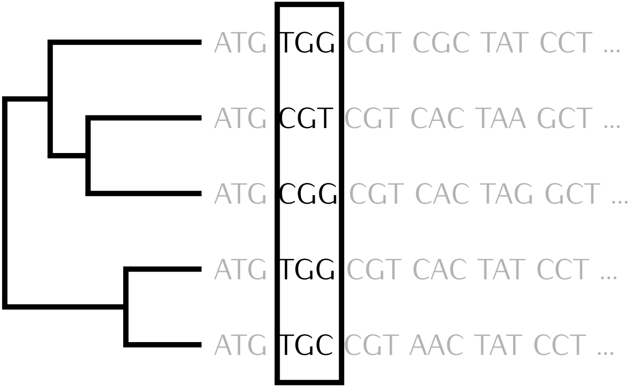
ATG CGT CGT CAC TAA GCT ...

ATG CGG CGT CAC TAG GCT ...

ATG TGG CGT CAC TAT CCT ...

ATG TGC CGT AAC TAT CCT ...

Modelling the evolution of a codon sequence alignment


ATG TGG CGT CGC TAT CCT ...
ATG CGT CGT CAC TAA GCT ...
ATG CGG CGT CAC TAG GCT ...
ATG TGG CGT CAC TAT CCT ...
ATG TGC CGT AAC TAT CCT ...

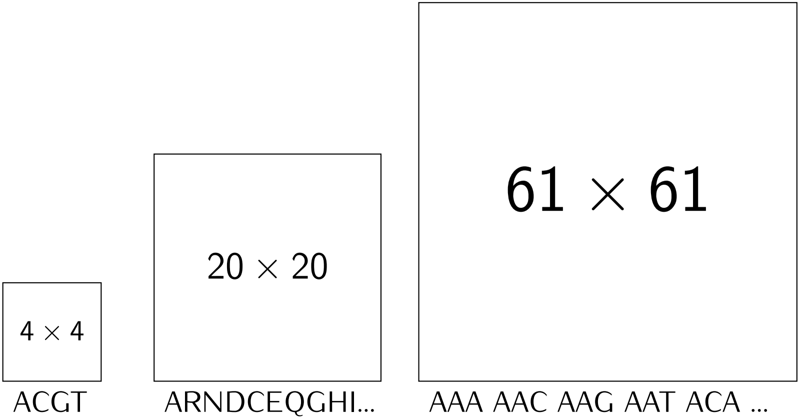
- We assume that all columns (sites) are independent

Modelling the evolution on a branch

- Mutations can occur at any time, with a given rate

Modelling the evolution on a branch

- Mutations can occur at any time, with a given rate
- The probability of each type of mutation is given by a matrix:



Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications…
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications...
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications…
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*
  - We allow nucleotide transitions and transversions to occur at a distinct rate. The ratio of the two is noted *kappa*

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications…
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*
  - We allow nucleotide transitions and transversions to occur at a distinct rate. The ratio of the two is noted *kappa*
- We can therefore express all mutation probabilities with only two parameters

Models for codon evolution

Muse and Gaut (1994), Goldman and Yang (1994)

Instantaneous substitution rate

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j & \text{if } i \text{ and } j \text{ differ by one synonymous transversion} \\ \kappa \cdot \pi_j & \text{if } i \text{ and } j \text{ differ by one synonymous transition} \\ \omega \cdot \pi_j & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transversion} \\ \omega \cdot \kappa \cdot \pi_j & \text{if } i \text{ and } j \text{ differ by one nonsynonymous transition} \end{cases}$$

Models for codon evolution

On the π_j...

- 61 codon frequencies: F61 model, 61 parameters)

Models for codon evolution

On the π_j…

- 61 codon frequencies: F61 model, 61 parameters)
- Consider only the frequencies of A, T, C and G, and deduce the frequencies of all codons: F1X4 model, 3 parameters

Models for codon evolution

On the π_j...

- 61 codon frequencies: F61 model, 61 parameters)
- Consider only the frequencies of A, T, C and G, and deduce the frequencies of all codons: F1X4 model, 3 parameters
- Consider the frequencies of A, T, C and G, independently at the three codon positions: F3X4 model, 9 parameters

Models for codon evolution

On the π_j…

- 61 codon frequencies: F61 model, 61 parameters)
- Consider only the frequencies of A, T, C and G, and deduce the frequencies of all codons: F1X4 model, 3 parameters
- Consider the frequencies of A, T, C and G, independently at the three codon positions: F3X4 model, 9 parameters
- Consider all codon equally frequent: F0 model, 0 parameter

Models for codon evolution

On the π_j…

- 61 codon frequencies: F61 model, 61 parameters)
- Consider only the frequencies of A, T, C and G, and deduce the frequencies of all codons: F1X4 model, 3 parameters
- Consider the frequencies of A, T, C and G, independently at the three codon positions: F3X4 model, 9 parameters
- Consider all codon equally frequent: F0 model, 0 parameter

✍ Frequencies can be estimated, or fixed to their observed values

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site
3. Summing over all possible ancestral states

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site
3. Summing over all possible ancestral states
4. Multiplying for all sites in the alignment

Site heterogeneity

☞ Model assumes homogeneous selective pressure along the alignment. How to account for heterogeneity?

Site heterogeneity

☞ Model assumes homogeneous selective pressure along the alignment. How to account for heterogeneity?

- We consider several possible scenarios for each site:
  - $\omega_1 = 1$ neutral evolution
  - $\omega_0 < 1$ negative selection
  - $\omega_2 > 1$ positive selection

Site heterogeneity

☞ Model assumes homogeneous selective pressure along the alignment. How to account for heterogeneity?

- We consider several possible scenarios for each site:
  - $\omega_1 = 1$ neutral evolution
  - $\omega_0 < 1$ negative selection
  - $\omega_2 > 1$ positive selection
- Each site can therefore "chose" between several omegas

Site heterogeneity

☞ Model assumes homogeneous selective pressure along the alignment. How to account for heterogeneity?

- We consider several possible scenarios for each site:
  - $\omega_1 = 1$ neutral evolution
  - $\omega_0 < 1$ negative selection
  - $\omega_2 > 1$ positive selection
- Each site can therefore "chose" between several omegas
- The likelihood of site i becomes

$$L_i = \sum_{\omega} L_i(\omega) \times \Pr(\omega)$$

where $L_i(\omega)$ is the likelihood for site i for a given value of ω, and $\Pr(\omega)$ is the probability of this given ω (the frequency of sites in the alignment which evolve with this particular ω).

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

M1a (variable selective pressure) $1 - p_0$ positions with $\omega = 1$, p_0 positions with $\omega < 1$

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

M1a (variable selective pressure) $1 - p_0$ positions with $\omega = 1$, p_0 positions with $\omega < 1$

M2a (variable selective pressure with positive selection) some positions with $\omega = 1$, others with $\omega < 1$ and some with $\omega > 1$

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

M1a (variable selective pressure) $1 - p_0$ positions with $\omega = 1$, p_0 positions with $\omega < 1$

M2a (variable selective pressure with positive selection) some positions with $\omega = 1$, others with $\omega < 1$ and some with $\omega > 1$

M7 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$, with $\beta(p, q)$ being the beta distribution between 0 and 1

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

M1a (variable selective pressure) $1 - p_0$ positions with $\omega = 1$, p_0 positions with $\omega < 1$

M2a (variable selective pressure with positive selection) some positions with $\omega = 1$, others with $\omega < 1$ and some with $\omega > 1$

M7 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$, with $\beta(p, q)$ being the beta distribution between 0 and 1

M8 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$ and some with $\omega > 1$

Some site-heterogeneous codon models

Yang, Nielsen, Goldman and Pedersen (2000)

M0 (one ratio) all positions identical, one parameter ω

M1a (variable selective pressure) $1 - p_0$ positions with $\omega = 1$, p_0 positions with $\omega < 1$

M2a (variable selective pressure with positive selection) some positions with $\omega = 1$, others with $\omega < 1$ and some with $\omega > 1$

M7 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$, with $\beta(p, q)$ being the beta distribution between 0 and 1

M8 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$ and some with $\omega > 1$

M9 (variable selective pressure) some positions with $\omega \circlearrowleft \beta(p, q)$ and some with $\omega \circlearrowleft \Gamma(a, b) + 1$, where $\Gamma(a, b) + 1$ is the gamma distribution between 1 and $+\inf$.

Model comparison

☞ Compare a model with selection to a neutral model

1. Fit both models (*e.g.* M1a and M2a, M7 and M8 or M9)

Model comparison

☞ Compare a model with selection to a neutral model

1. Fit both models (*e.g.* M1a and M2a, M7 and M8 or M9)
2. Perform a likelihood ration test (LRT): compute

$$S = 2 \times \ln(L1/L0) = 2 \times (\ln(L1) - \ln(L0))$$

where $L1$ is the likelihood of the model with selection, and $L0$ the likelihood of the neutral model.

Model comparison

☞ Compare a model with selection to a neutral model

1. Fit both models (*e.g.* M1a and M2a, M7 and M8 or M9)
2. Perform a likelihood ration test (LRT): compute

   $$S = 2 \times \ln(L1/L0) = 2 \times (\ln(L1) - \ln(L0))$$

   where $L1$ is the likelihood of the model with selection, and $L0$ the likelihood of the neutral model.
3. $S \circlearrowleft \chi(n1 - n0)$, where $n1$ and $n2$ are the number of parameters of the model with and without selection, respectively. For M2a-M1a and M8-M7, $n1 - n0 = 2$, for $M9 - M7$, $n1 - n0 = 3$.

Model comparison

☞ Compare a model with selection to a neutral model

1. Fit both models (*e.g.* M1a and M2a, M7 and M8 or M9)
2. Perform a likelihood ration test (LRT): compute

$$S = 2 \times \ln(L1/L0) = 2 \times (\ln(L1) - \ln(L0))$$

   where $L1$ is the likelihood of the model with selection, and $L0$ the likelihood of the neutral model.
3. $S \circlearrowleft \chi(n1 - n0)$, where $n1$ and $n2$ are the number of parameters of the model with and without selection, respectively. For M2a-M1a and M8-M7, $n1 - n0 = 2$, for $M9 - M7$, $n1 - n0 = 3$.
4. If significant, use a Bayesian approach to identify positions where M2a/M8/M9 has a higher posterior probability than M1a/M7