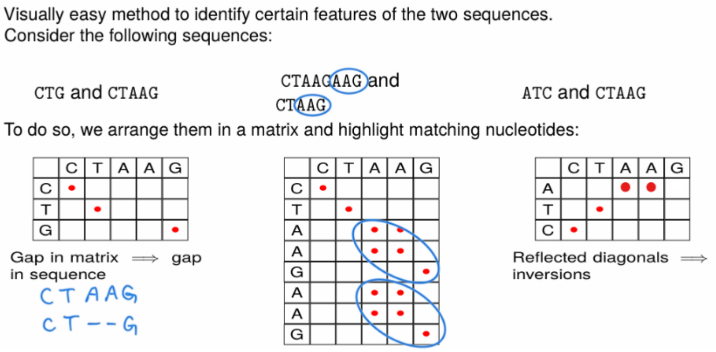
# Overview

# 1. Alignment Methods

## 1.1 Pairwise sequence alignment

### 1.1.1 Dot-matrix



Visually easy method to identify certain features of the two sequences.
Consider the following sequences:

CTG and CTAAG

CTAAGAAG and
CTAAG

ATC and CTAAG

To do so, we arrange them in a matrix and highlight matching nucleotides:

|   | C | T | A | A | G |
|---|---|---|---|---|---|
| C | ● |   |   |   |   |
| T |   | ● |   |   |   |
| G |   |   |   |   | ● |

Gap in matrix ⟹ gap in sequence

CTAAG
CT--G

|   | C | T | A | A | G |
|---|---|---|---|---|---|
| C | ● |   |   |   |   |
| T |   | ● |   |   |   |
| A |   |   | ● | ● |   |
| A |   |   | ● | ● |   |
| G |   |   |   |   | ● |
| A |   |   | ● | ● |   |
| A |   |   | ● | ● |   |
| G |   |   |   |   | ● |

|   | C | T | A | A | G |
|---|---|---|---|---|---|
| A |   |   | ● | ● |   |
| T |   | ● |   |   |   |
| C | ● |   |   |   |   |

Reflected diagonals ⟹ inversions

Advantages:

- Visually easy to identity sequence features such as indels, repeats, inversions, and inverted repeats.

Disadvantages

- Time consuming
- Does not give one optimal alignment

### 1.1.2 Quantitative brute force

Reward a match, e.g., score = 3; punish a mismatch, e.g., score = -1, punish a gap, e.g., score = -2. Calculate the final score by multiplying the corresponding number of matches, mismatches, and gaps. Brute force is to find the optimal alignment, meaning that it must find all alignments and score them. Suppose we have:

- Sequence $a = a_1 a_2 \dots a_m$ with length of $m$
- Sequence $b = b_1 b_2 \dots b_n$ with length of $n$
- $n \leq m$
- $k$ is the number of gaps to be introduced into sequence $a$
- There are $\binom{m+k}{k}$ possibilities to place these gaps between the $m$ letters of $a$
- There are $\binom{m}{m-n+k}$ possibilities to place the number of gaps in $b$ as the corresponding positions of $a$
- Total number of alignments is $\sum_{k=0}^{n} \binom{m+k}{k} \binom{m}{m-n+k}$

Advantages:

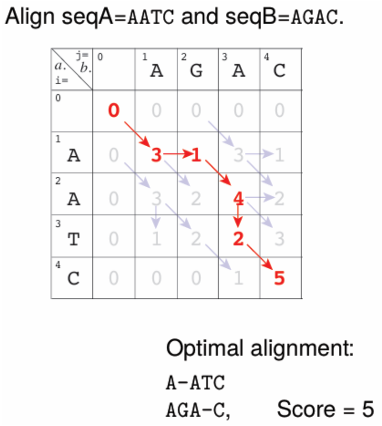
- Optimal guaranteed

Disadvantages:

- Way too time consuming

### 1.1.3 Smith-Waterman algorithm (local alignment)

$$H[i,j] = \max \begin{Bmatrix} 0 & \\ H[i-1,j-1] + 3 \text{ if } a_i = b_j & \text{match} \\ H[i-1,j-1] - 1 \text{ if } a_i \neq b_j & \text{mismatch} \\ H[i-1,j] - 2 & \text{gap in sequence seqB} \\ H[i,j-1] - 2 & \text{gap in sequence seqA} \end{Bmatrix}$$



Align seqA=AATC and seqB=AGAC.

| i= a. \ j= b. | 0 | 1 A | 2 G | 3 A | 4 C |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 A | 0 | 3 | 1 | 3 | 1 |
| 2 A | 0 | 3 | 2 | 4 | 2 |
| 3 T | 0 | 1 | 2 | 2 | 3 |
| 4 C | 0 | 0 | 0 | 1 | 5 |

Optimal alignment:

A-ATC
AGA-C,     Score = 5

Key points:

- 0th row and column set to 0, the other cells are for associated nucleotides
- Start from highest number (score) until 0 is reached

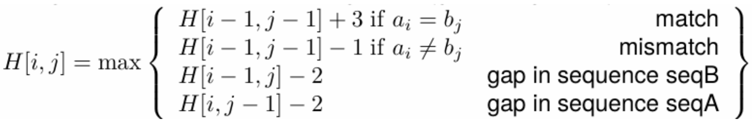
Advantages:

- Fast than brute force, only $m \times n$ steps
- Finds the optimal local alignment

Disadvantages:

- Only for local alignment
- Only pairwise alignment possible
- Multiple alignments possible
- Still slow for large libraries

### 1.1.4 Needleman-Wunsch algorithm (global optimal guaranteed)



$$H[i,j] = \max \left\{ \begin{array}{ll} H[i-1,j-1] + 3 \text{ if } a_i = b_j & \text{match} \\ H[i-1,j-1] - 1 \text{ if } a_i \neq b_j & \text{mismatch} \\ H[i-1,j] - 2 & \text{gap in sequence seqB} \\ H[i,j-1] - 2 & \text{gap in sequence seqA} \end{array} \right\}$$

Key points:

- Similar to Smith-Waterman algorithm
- Initiate the 0th row and column according to $H[0,j] = j \times w$ and $H[i,0] = i \times w$
- Starts from the bottom right (the score is also there), follow the path to $(0,0)$

## 1.2 Heuristic approach: BLAST

Process:

1. Split the query sequence into subsequences of length $k$
2. Search these $k$-letter words are allowed but scored, e.g., match +5, mismatch -3
3. Keep only the sequences with the highest scores
4. Expand the $k$-letter word to the right and left, note the scores
5. Stop if the score drops below a certain threshold
6. Keep only the pairwise alignments that are above the threshold
7. Report these database sequences

Advantages:

- Up to 50 – 100 times faster than SW and NW methods
- Allows search for exact matches but also for similarity up to a predefined degree

Disadvantages:

- Does not guarantee the optimal pairwise alignments of the query and database sequences
- Can only find a match when the gene/sequence is available in the database

## 1.3 Multiple sequence alignment

### 1.3.1 Pairwise alignment against a reference strain

Define a reference strain for the genome and pairwise align all sequences with the reference strain.

Advantages:

- Position numbering is the same for each sequence

Disadvantages:

- Only possible when one knows which species the sequences come from

### 1.3.2 Arrays

Generalization of Needleman-Wunsch algorithm. Scales terribly with the number of sequences, $k$ sequences of length $m$ requires approximately $m^k$ steps. Can be made practical by excluding sequences with too many gaps.

# 2. Evolutionary models

## 2.1 Genome-Wide Association Studies (GWAS)

### 2.1.1 Odds ratio

|  | case | control | total |
|---|---|---|---|
| minor | $D_S = 2104$ | $H_S = 2676$ | 4780 |
| major | $D_N = 1896$ | $H_N = 3324$ | 5520 |
| total | 4000 | 6000 | 10000 |

- $OR = \frac{D_S/H_S}{D_N/H_N} = \frac{2104/2676}{1896/3324} = 1.38$
- $D_S/H_S$: odds of having the disease among individuals with minor variant on SNP position
- $D_N/H_N$: odds of having the disease among individuals with major variant on SNP position
- SNP: Single nucleotide polymorphism
- Control group: healthy
- Case group: with certain disease
- > 1 minor variant increases the risk of disease, = 1 no association, < 1 minor variant decreases the risk of disease

### 2.1.2 Fisher's exact test

Based on the contingency table, we want to test whether A is lined to B.

|  | $B_1$ | $B_2$ |  |
|---|---|---|---|
| $A_1$ | a | b | a + b |
| $A_2$ | c | d | c + d |
|  | a + c | b + d | n |

- $H_0$: The random variable of the number of individuals expressing both $A_1$ and $B_1$ is distributed according to a hypergeometric distribution.
- $p - value = \sum_{i=a}^{a+b} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{n}{a+c}}$

Disadvantages:

- Fisher's exact test only works for small numbers, the choose operator cannot be calculated correctly for large numbers

### 2.1.3 Pearson's Chi-square test

One calculates the deviance between observed and expected numbers based on a hypergeometric distribution. For example, $E_{(A_1,B_1)} = \frac{a+c}{a+b+c+d} \cdot \frac{a+b}{a+b+c+d} \cdot n$.

Observed: $\left(O_{ij}\right)_{i,j\in\{1,2\}}$

|  | case | control | total |
|---|---|---|---|
| minor | $D_S = 2104$ | $H_S = 2676$ | 4780 |
| major | $D_N = 1896$ | $H_N = 3324$ | 5520 |
| total | 4000 | 6000 | 10000 |

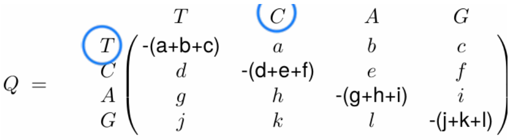
Expected: $\left(E_{ij}\right)_{i,j\in\{1,2\}}$

|  | case | control | total |
|---|---|---|---|
| minor | $D_S = 1912$ | $H_S = 2868$ | 4780 |
| major | $D_N = 2088$ | $H_N = 3132$ | 5520 |
| total | 4000 | 6000 | 10000 |

- Test statistics: $x = \sum_{i,j=1}^{2} \frac{\left(O_{ij}-E_{ij}\right)^2}{E_{i.j}}$
- It is Chi-squared distributed, p-value can be obtained by the Chi-square chart

## 2.2 Markov chain

**Stochastic process**, i.e., a series of random experiments through time. Lives on a **state space** and jumps to the different states. It is **memoryless**, the probability of jumping to a state only depends on the current state $P\left(X_{t_{n+1}} = x_{t_{n+1}} \middle| X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, \dots\right) = P\left(X_{t_{n+1}} = x_{t_{n+1}} \middle| X_{t_n} = x_{t_n}\right)$. If the transition probabilities on the state space do not change over time, the Markov chain is called **time homogenous**. State space of each nucleotide position $S = \{T, C, A, G\}$, the substitution matrix is



$$Q = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{cccc} T & C & A & G \\ \end{array}$$

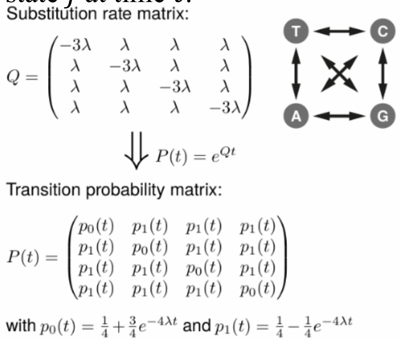
$$Q = \begin{array}{c|cccc} & T & C & A & G \\ \hline T & -(a+b+c) & a & b & c \\ C & d & -(d+e+f) & e & f \\ A & g & h & -(g+h+i) & i \\ G & j & k & l & -(j+k+l) \end{array}$$

Rate vs. Probabilities

- **Rate**: measures events per time unit
    - Deterministic, fixed quantity, describes averages
- **Probability**: measure of chance that a random event occurs
    - Stochastic, describes an exact event
- Suppose $\alpha$ is the rate of an event $E$ happening per unit of time. The probability that this happens in a very small time step $\Delta t$ is defined as $\alpha \Delta t$. We denote the time until the event happens as a random variable $X$. The probability that it does not happen in $\Delta t$ is $P(X > \Delta t) = 1 - \alpha \Delta t$. Let $\tau$ be a time with $\tau = k\Delta t$. We can divide the probability that $E$ does not happen in $\tau$ into $k$ time intervals in which the event does not happen: $P(X > \tau) = (1 - \alpha\Delta t)^k = (1 - \alpha\Delta t)^{\tau/\Delta t} \rightarrow^{\Delta t \rightarrow 0} e^{-\alpha\tau}$.
- $P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$ is the c.d.f.
- $f(x) = \frac{dP}{dt}(x) = \alpha e^{-\alpha x}$ is the p.d.f. of an exponential distribution
- An event occurring with rate $\alpha$ means that it occurs after an exponentially distributed wait time with parameter $\alpha$

Rate matrix to transition probabilities

- $P(t) = \left(p_{ij}(t)\right)_{i,j \in S}$ be the transition probability matrix with all probabilities that given the Markov chain is in state $i$ at time 0, the Markov chain will be in state $j$ at time $t$.
- $P(t + \Delta t) = P(t) + P(t)Q\Delta t \Rightarrow \frac{P(t+\Delta t) - P(t)}{\Delta t} = P(t)Q$
- Take limit to get $\lim_{\Delta t \rightarrow 0} \frac{P(t+\Delta t)-P(t)}{\Delta t} = \frac{dP}{dt}(t) = P(t)Q$
- $P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Q_t)^i}{i!}$



Substitution rate matrix:

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

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

Transition probability matrix:

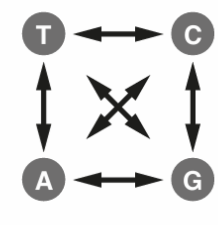
$$P(t) = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

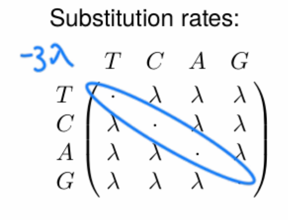
with $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$ and $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$

## 2.3 Substitution models

### 2.3.1 JC69

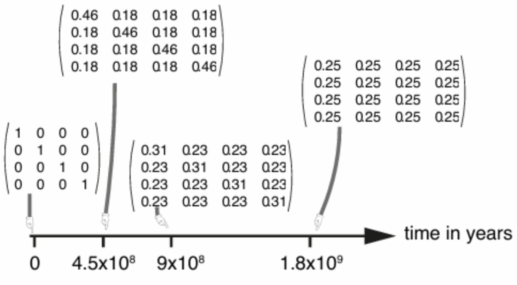
- Number of parameters: 1
- All substitutions have the same rate $\lambda$





Substitution rates:

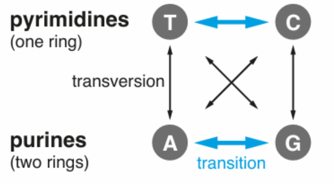
$$\begin{array}{c} -3\lambda \\ \\ \end{array} \quad \begin{array}{c|cccc} & T & C & A & G \\ \hline T & \cdot & \lambda & \lambda & \lambda \\ C & \lambda & \cdot & \lambda & \lambda \\ A & \lambda & \lambda & \cdot & \lambda \\ G & \lambda & \lambda & \lambda & \cdot \end{array}$$

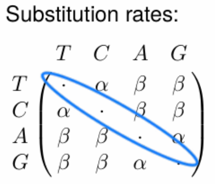


$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0.46 & 0.18 & 0.18 & 0.18 \\ 0.18 & 0.46 & 0.18 & 0.18 \\ 0.18 & 0.18 & 0.46 & 0.18 \\ 0.18 & 0.18 & 0.18 & 0.46 \end{pmatrix} \quad \begin{pmatrix} 0.31 & 0.23 & 0.23 & 0.23 \\ 0.23 & 0.31 & 0.23 & 0.23 \\ 0.23 & 0.23 & 0.31 & 0.23 \\ 0.23 & 0.23 & 0.23 & 0.31 \end{pmatrix} \quad \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

### 2.3.2 K80

- Number of parameters: 2
- Transitions happen at rate $\alpha$
- Transversions happen at rate $\beta$

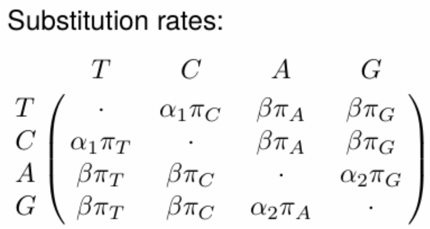



Substitution rates:

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

### 2.3.3 TN93 & HKY

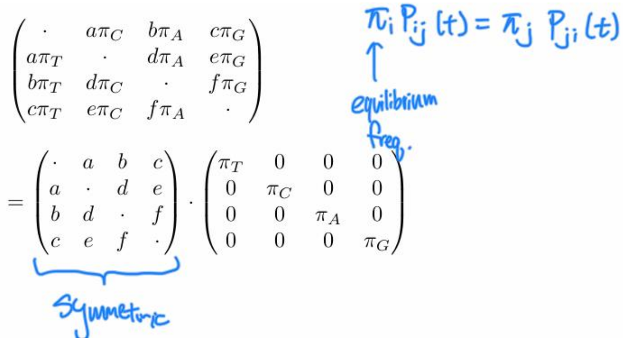
- Number of parameters TN93: 3 + 3
- Number of parameters HKY: 2 + 3
- Transitions between $T \leftrightarrow C$ happen at rate $\alpha_1 \times \pi_{\{T,C,A,G\}}$
- Transitions between $A \leftrightarrow G$ happen at rate $\alpha_2 \times \pi_{\{T,C,A,G\}}$
- Transversions happen at rate $\beta \times \pi_{\{T,C,A,G\}}$
- $\pi_{\{T,C,A,G\}}$ is the nucleotide equilibrium frequency
- When $\alpha_1 = \alpha_2$, then TN93 becomes HKY


Substitution rates:

$$\begin{array}{c|cccc} & T & C & A & G \\ \hline T & \cdot & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ C & \alpha_1\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ A & \beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_G \\ G & \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \cdot \end{array}$$

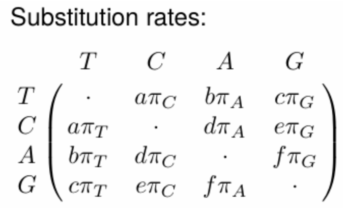
### 2.3.4 GTR

- Number of parameters: 6 + 3
- Generalized time-reversible model


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

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

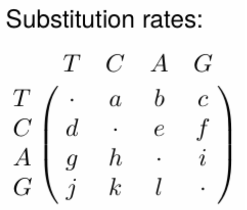
$$\underset{\uparrow\ \text{equilibrium freq.}}{\pi_i}\, P_{ij}(t) = \pi_j\, P_{ji}(t)$$


Substitution rates:

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

### 2.3.5 UNREST

- Number of parameters: 12
- Each substitution has a (different) rate
- All other models are special cases of UNREST
- Mathematically very complicated and not handy to use
- Not time-reversible


Substitution rates:

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

**Note**: Substitution models account for hidden evolutionary events - such as multiple substitutions at the same site - which, if ignored, would lead to inaccurate conclusions about species relationships, evolutionary rates, and molecular adaptation.

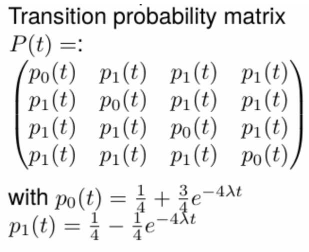
## 2.4 Maximum likelihood estimator (MLE)

An estimator that returns the value of a model parameter which maximizes the probability of the observed results.

- Looking for the maximum of the log-likelihood function: $l(p;x) = \log(L(p;x))$

- Multiplying $L$ with a constant

**JC69 MLE for sequence distance**



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

Suppose having 2 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 $t$).

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

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

The log-likelihood function is:

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

Set the derivative to 0 and solving the equation for $d$ yields $\hat{d}_{JC69} = -\frac{3}{4}\log\left(1 - \frac{4x}{3n}\right)$.

## 2.5 Variable substitution rates across genome

Substitution rates may differ across the genomes (not always evolve at the same rate) due to mutation rate difference across sites, and selective pressure difference on the phenotypic level.

Gamma distribution

- Probability distribution: $g(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}$
- Γ-function: $\Gamma(\alpha) = \int e^{-t} t^{\alpha-1} dt$; $\Gamma(n) = (n-1)!$
- A $\Gamma(\alpha,\beta)$-distributed random variable $X$ has mean $E[X] = \frac{\alpha}{\beta}$ and variance $Var[X] = \frac{1}{\alpha}$.

JC69+**Γ**

- $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda R t} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$
- $E[p] = \int \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}\right) g(r)dr = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{\alpha}\right)^{-\alpha}$
- $l(d;x) = \binom{n}{x}(E[p])^x(1 - E[p])^{n-x}$
- $\hat{d}_{JC69+\Gamma} = \frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-\frac{1}{\alpha}} - 1\right)$
- $\hat{d}_{JC69+\Gamma} \geq \hat{d}_{JC69}$, ignoring site variation tends to lead to underestimation of the sequence distance

## 2.6 Amino acid substitution models

In case of AA substitutions, $P(t)$ as well as the substitution rate matrix have dimension $20 \times 20$. To determine $P(t)$, we need the Q-matrix, however

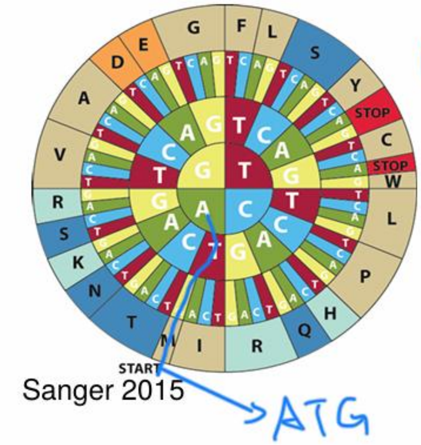
- More difficult than for nucleotide substitutions
- Two approaches: empiric, mechanistic
- Desired: time-reversibility

Suppose all substitutions happen at the same rate $\lambda$

- Substitution rate for any substitution is $19\lambda$
- The expected time until a substitution happens is $\frac{1}{19\lambda}$
- With $t = \frac{d}{19\lambda}$ we have $\hat{d}_{AA} = -\frac{19}{20}\log\left(1 - \frac{20x}{19n}\right)$
- We can also use MLE to estimate the Q-matrix

## 2.7 Codon substitution models

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

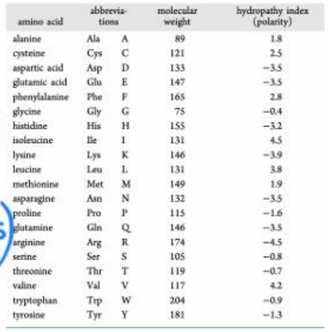


Sanger 2015

ATG

$64 = 4^3$ codons.

20 distict values (amino acids)



| amino acid | abbrevia-tions | | molecular weight | hydropathy index (polarity) |
|---|---|---|---|---|
| alanine | Ala | A | 89 | 1.8 |
| cysteine | Cys | C | 121 | 2.5 |
| aspartic acid | Asp | D | 133 | −3.5 |
| glutamic acid | Glu | E | 147 | −3.5 |
| phenylalanine | Phe | F | 165 | 2.8 |
| glycine | Gly | G | 75 | −0.4 |
| histidine | His | H | 155 | −3.2 |
| isoleucine | Ile | I | 131 | 4.5 |
| lysine | Lys | K | 146 | −3.9 |
| leucine | Leu | L | 131 | 3.8 |
| methionine | Met | M | 149 | 1.9 |
| asparagine | Asn | N | 132 | −3.5 |
| proline | Pro | P | 115 | −1.6 |
| glutamine | Gln | Q | 146 | −3.5 |
| arginine | Arg | R | 174 | −4.5 |
| serine | Ser | S | 105 | −0.8 |
| threonine | Thr | T | 119 | −0.7 |
| valine | Val | V | 117 | 4.2 |
| tryptophan | Trp | W | 204 | −0.9 |
| tyrosine | Tyr | Y | 181 | −1.3 |

Pal et al. 2022

- Synonymous substitutions: AA does not change / protein not changed -> neutral substitution
- Non-synonymous substitutions: AA does change / protein changes, selective processes can act on new protein

The basic model

- The substitution rate matrix has dimension 61 × 61
- We incorporate
  - $\kappa$: transition/transversion rate ratio
  - $\omega$: nonsynonymous/synonymous rate ratio
  - $\pi_I = \frac{1}{C}\pi^*_{i_1}\pi^*_{i_2}\pi^*_{i_3}$: equilibrium frequency of codon $I$ consisting of nucleotides $i_1, i_2, i_3$ with equilibrium frequencies $\pi^*_{i_1}\pi^*_{i_2}\pi^*_{i_3}$

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

-

Evidence of selection

Compare amounts of nonsynonymous and synonymous substitutions

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

Counting method for $d_N/d_S$ ratio

- Count the non-synonymous ($N_d$) and synonymous ($S_d$) differences (observed)
- Count the non-synonymous ($N$) and synonymous ($S$) sites (possible)
- Account for the possible evolutionary history by applying the pairwise distance formula to the ratio $\frac{N_d}{N}$ and $\frac{S_d}{S}$.

Under JC69, we have the distances as

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

- $d_S = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{S_d}{S}\right)$
- Ratio is $d_N/d_S$
- If ratio < 1, nonsynonymous mutations occur less frequently -> purifying selection
- If ratio > 1, nonsynonymous mutations occur more frequently -> positive selection

# 3. Phylogenetic methods & molecular clock models

## 3.1 Terminologies

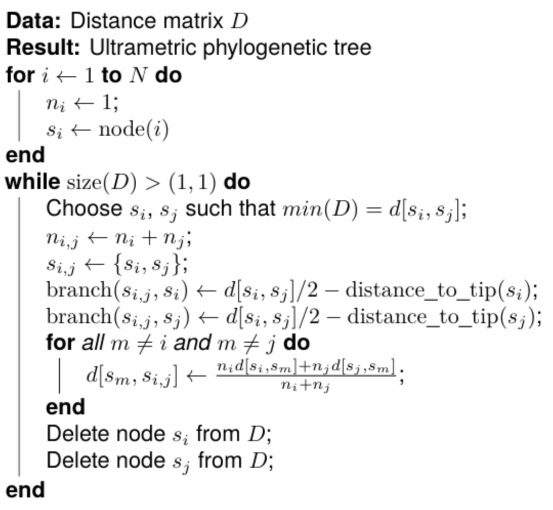
- A tree is a graph consisting of nodes and branches, without a loop
- An unrooted phylogenetic tree is a tree with two types of nodes:
    - Tip/leaf: node with 1 branch attached
    - Internal node: node with 3 branches attaches
- A rooted phylogenetic tree is an unrooted tree in which one branch is subdivided by a new node (root)
- Each branch may have a length >= 0 assigned
- **Pendant branch**: a branch attached to a tip
- **Cherry**: a pair of tips which are only separated by one internal node
- **Caterpillar tree**: a rooted tree with only one cherry
- **Monophyletic group / clade**: all descendants of a common ancestor
- **Ultrametric tree**: the sum of branch lengths from any tip to the root is the same
- **Polytomy**: the definition of a phylogenetic tree is extended such that internal nodes may have more than 3 branches attached. Such a node is a polytomy.

## 3.2 Phenetic approach (based on overall similarity, distance-based)

A phylogenetic reconstruction method is **statistically consistent** if the true tree is returned for an infinite amount of data (i.e., infinitely long sequences). Both UPGMA and least squares are consistent methods. Problems: (1) disregard information beyond pairwise distances; (2) large distances come with large variances, which are typically ignored.

### 3.2.1 UPGMA

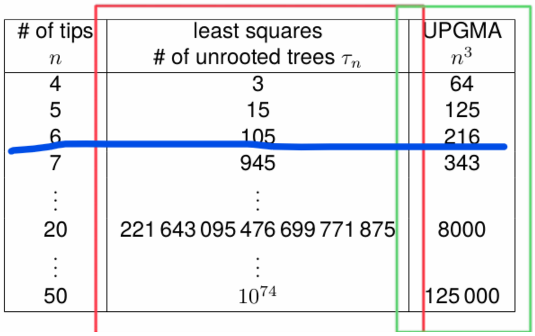
- **Input**: distance matrix (need to use substitution models such as JC69)
- **Output**: Ultrametric phylogenetic tree (rooted tree)
- Assumptions:
    - All sequences must come from the **same time point**
    - Assumes evolution according to a **strict molecular clock**, i.e., **constant evolution rate** over time
- Can use **Neighbor-Joining Algorithm** to relax these assumptions, i.e., branch lengths correspond to number of mutations, output tree will be **unrooted**



```
Data: Distance matrix $D$
Result: Ultrametric phylogenetic tree
for $i \leftarrow 1$ to $N$ do
    $n_i \leftarrow 1$;
    $s_i \leftarrow \text{node}(i)$
end
while $\text{size}(D) > (1,1)$ do
    Choose $s_i, s_j$ such that $min(D) = d[s_i, s_j]$;
    $n_{i,j} \leftarrow n_i + n_j$;
    $s_{i,j} \leftarrow \{s_i, s_j\}$;
    $\text{branch}(s_{i,j}, s_i) \leftarrow d[s_i, s_j]/2 - \text{distance\_to\_tip}(s_i)$;
    $\text{branch}(s_{i,j}, s_j) \leftarrow d[s_i, s_j]/2 - \text{distance\_to\_tip}(s_j)$;
    for all $m \neq i$ and $m \neq j$ do
        $d[s_m, s_{i,j}] \leftarrow \frac{n_i d[s_i,s_m] + n_j d[s_j,s_m]}{n_i + n_j}$;
    end
    Delete node $s_i$ from $D$;
    Delete node $s_j$ from $D$;
end
```

Running time:

- Prune nodes $n$ times
- $n^2$ calculations per pruning
- $O(n^3)$ running time for $n$ sequences



| # of tips<br>$n$ | least squares<br># of unrooted trees $\tau_n$ | UPGMA<br>$n^3$ |
|---|---|---|
| 4 | 3 | 64 |
| 5 | 15 | 125 |
| 6 | 105 | 216 |
| 7 | 945 | 343 |
| ⋮ | ⋮ | |
| 20 | 221 643 095 476 699 771 875 | 8000 |
| ⋮ | ⋮ | |
| 50 | $10^{74}$ | 125 000 |

### 3.2.2 Least square methods

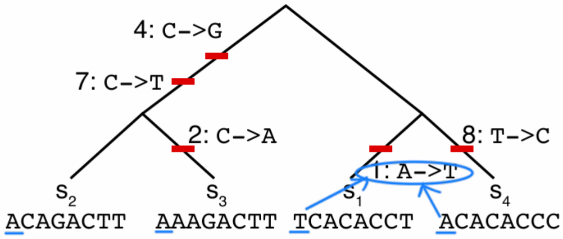
- **Input**: distance matrix
- Repeat until all tree topologies have been considered
  - Propose an unrooted tree topology (without branch lengths)
  - Minimizes square error $S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{i,j}\left(D_{i,j} - d_{i,j}\right)^2$, where $D$ is the sequence distances matrix, $d$ the tree distance matrix for a proposed tree, and $w$ weights (1 or inversely proportional to $D_{i,j}$, not using 1 because we might want to acknowledge that some data is "noisier" than others, so we let the model ignore the noise to focus on the signal)
- **Output**: tree with the smallest $S$
- Running time:
  - An unrooted tree with $n$ tips has $b_n = 2n - 3$ branches,
  - With $n$ tips, there are $\tau_n = (2n-5)!!$ unrooted trees
  - With $n$ tips, there are $\tau_n^r = (2n-5)!!\,(2n-3) = (2n-3)!!$ rooted trees
  - Runtime $O(e^{nlnn})$

## 3.3 Cladistic approach (character based, sequences with shared characters)

- **Statistically inconsistent**: no back substitutions or parallel substitutions are considered, which leads to **long-branch attraction**. If you have a tree with two very **long branches** (rapid evolution) separated by a short internal branch, Parsimony will **incorrectly group the long branches together** because they will share "homoplasies" (random convergent mutations) simply by chance.
- **Slow**: exponential time complexity – the whole tree space has to be visited.

### 3.3.1 Parsimony

- Find tree needing smallest number of mutations.
- **Parsimony score of a tree**: the lowest number of mutations required to explain the sequences at the tips of the tree
- **Parsimony tree**: the tree with lowest parsimony score

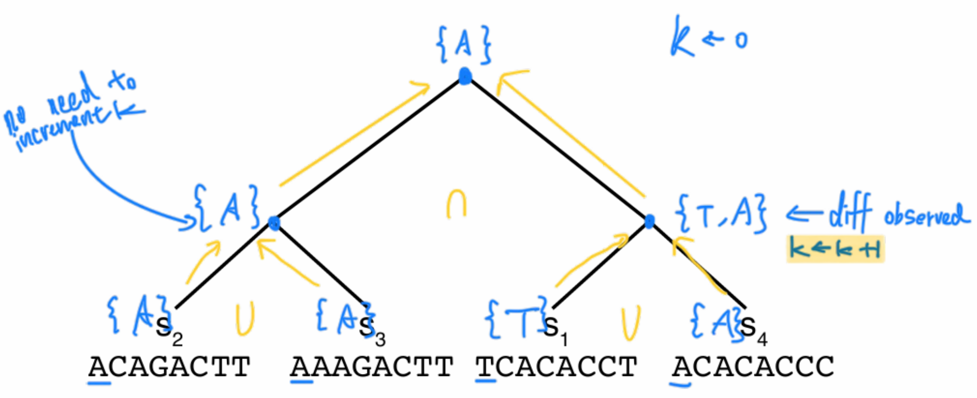


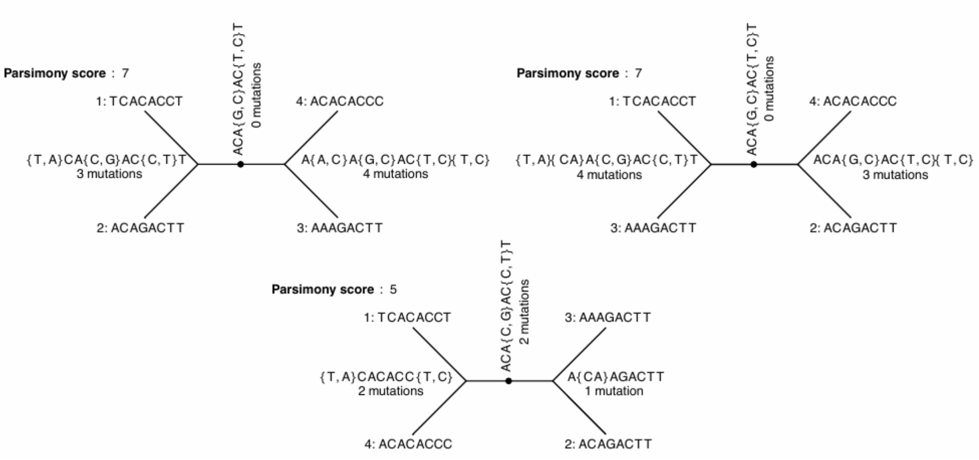
- Label **internal nodes** with each possible ancestral sequence and then determine the number of mutations required for each assignment. Minimal number of mutations required is the parsimony score.
- For $n$ tips with $m$ sequences, the cost of considering all ancestors is $4^{n-1} \cdot m$.
- **Rooted tree from the same unrooted tree have the same parsimony score.**

- **Input**: Sequence alignment of $n$ sequences, with sequence length $m$
- Iterate:
  - Consider each unrooted tree ($(2n-5)!!$ overall)
  - Calculate parsimony score for the considered unrooted tree (requires considering $4^{n-1}m$ internal character assignments)
- **Output**: unrooted tree with lowest parsimony score

### 3.3.2 Fitch algorithm for parsimony score

- Root the tree at an arbitrary edge
- $k \leftarrow 0$
- While the root has no sequence assigned, iterate
    - Choose a node in the tree where all descending nodes have sequences assigned
    - Assign a sequence to the chosen node:
        - For $i = 1, \dots, m$:
            - Let $C_l$ and $C_r$ be the sets of nucleotides being assigned to the two direct descents of the chosen node for site $i$
            - **If $\boldsymbol{C_l \cap C_r \neq \emptyset}$,** we assign $C_l \cap C_r$ to nucleotide $i$ of the chosen node
            - **If $\boldsymbol{C_l \cap C_r = \emptyset}$,** we assign $C_l \cup C_r$ to nucleotide $i$ of the chosen node and set $\boldsymbol{k \leftarrow k+1}$
- **Output**: parsimony score $k$ of the tree, i.e., minimal number of mutations required to explain the sequence at the tips.





Running time:

- Must visit each internal node of the rooted tree, for a tree with $n$ tips, there are $n-1$ internal nodes
- With $m$ sequences, the running time is $(n-1)m$
- With exhaustive search, the number of tree grows as $O(\exp(n))$, NP-complete.

## 3.4 Mechanistic approach (character based)

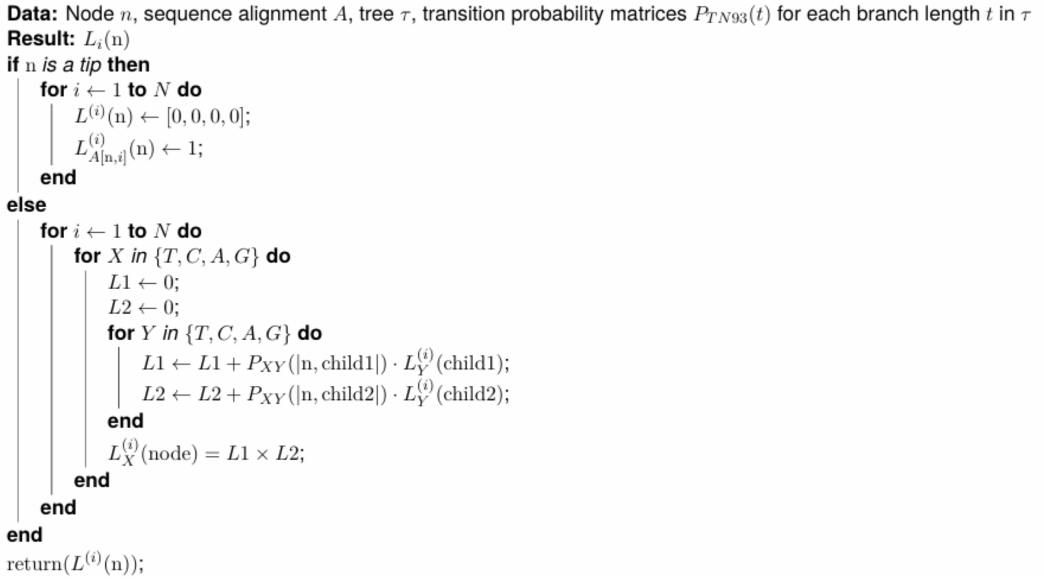
**Statistically consistent.**

### 3.4.1 Maximum likelihood

- **Input**: sequence alignment

- **Output**: Tree which maximizes the probability of the sequences given the tree and the sequence evolution parameters (**unrooted**, need outgroup to root the tree)
    - Requires an evolutionary model (JC69, HKY, etc.)
    - Parameters of the evolutionary model can be co-estimated with the tree
- Mechanistic model for evolution of the data
    - Each unrooted tree T with branch lengths is a parameter
    - Sequences evolve on the tree according to parameters provided in the rate matrix $Q$
- Mechanistic model description allows us to simulate sequence alignments (data $D$) for given parameters
- $L(T, Q; D) = P(D|T, Q)$ is called the likelihood function of the parameters $T$, $Q$ for given sequence data
- Inference: determine the $T$, $Q$ which best explain the alignment $\max_{T,Q} L(T, Q; D)$
- We determine the best tree by evaluating the likelihood for many different proposed trees.
    - Look at sites **independently** $P(s_1, \dots, s_n|T, Q) = \prod_{j=1}^{m} P(s_{1,j}, \dots, s_{n,j}|T, Q)$.
    - **Sum over internal node sequences** $P\left(s_{1,j}, \dots, s_{n,j}\middle|T, Q\right) = \sum_{s_{n+1,j} \in \{A,C,G,T\}} \cdots \sum_{s_{2n-1,j} \in \{A,C,G,T\}} P(s_{1,j}, \cdots, s_{2n-1,j}|T, Q)$
    - **Multiply all branches** in tree $P\left(s_{1,j}, \cdots, s_{2n-1,j}\middle|T, Q\right) = \pi\left(s_{2n-1,j}\right) \prod_{l=1}^{2n-2} P_{s_{l_1,j}, s_{l_2,j}}(t_l)$, $s_{l_1}$ is the starting sequence, $s_{l_2}$ is the ending sequence, $t_l$ is the branch length, a rooted tree with $n$ tips has $2n - 2$ branches
- Running time
    - We need to visit each single tree in the tree space
    - For each tree we need to calculate the likelihood
        - Multiply over all sites $O(m)$
        - Sum over internal nucleotides at $n - 1$ internal nodes $O(4^{n-1})$
        - Multiply over $2n - 2$ branches $O(2n - 2)$
    - Running time of likelihood calculation is $O(m4^n n)$
    - Felsenstein's pruning algorithm speeds up this calculation using DP

## Felsenstein's pruning algorithm



```
Data: Node $n$, sequence alignment $A$, tree $\tau$, transition probability matrices $P_{TN93}(t)$ for each branch length $t$ in $\tau$
Result: $L_i(\mathrm{n})$
if n is a tip then
    for $i \leftarrow 1$ to $N$ do
        $L^{(i)}(\mathrm{n}) \leftarrow [0, 0, 0, 0]$;
        $L^{(i)}_{A[\mathrm{n},i]}(\mathrm{n}) \leftarrow 1$;
    end
else
    for $i \leftarrow 1$ to $N$ do
        for $X$ in $\{T, C, A, G\}$ do
            $L1 \leftarrow 0$;
            $L2 \leftarrow 0$;
            for $Y$ in $\{T, C, A, G\}$ do
                $L1 \leftarrow L1 + P_{XY}(|\mathrm{n}, \mathrm{child1}|) \cdot L^{(i)}_Y(\mathrm{child1})$;
                $L2 \leftarrow L2 + P_{XY}(|\mathrm{n}, \mathrm{child2}|) \cdot L^{(i)}_Y(\mathrm{child2})$;
            end
            $L^{(i)}_X(\mathrm{node}) = L1 \times L2$;
        end
    end
end
return($L^{(i)}(\mathrm{n})$);
```

Running time:

- Each recursion step is a summation over four times four states, i.e., constant, we have $O(n)$ nodes and thus the recursion has running time $O(n)$
- The recursion has to be performed for each of the $m$ sites, $O(m)$
- Running time $O(mn)$

### 3.4.2 Bayesian

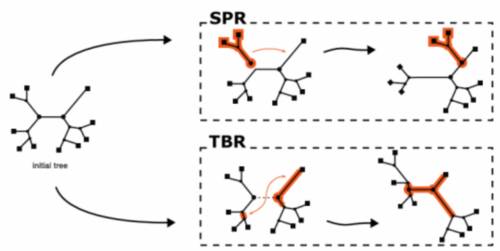
- The phylogenetic likelihood $P(A|T,Q)$
    - $A$ is a sequence alignment
    - $T$ is a phylogenetic tree
    - $Q$ is the substitution rate matrix
    - Use Felsenstein algorithm to calculate
- The phylogenetic posterior $P(T,Q,\eta|A) = \frac{1}{P(A)}P(A|T,Q)P(T|\eta)P(Q,\eta)$
    - $\eta$ are the phylodynamic model parameters
    - $P(T|\eta)$ is the tree prior or phylodynamic likelihood
    - $P(Q,\eta) = P(Q)P(\eta)$ are the parameter prior distributions
- **Assumption**: Separating the process of **tree generation** from that of **sequence evolution** implies the **sequence evolution is effectively neutral**.
- The MCMC algorithm proposes new states $T'$, $Q'$, $\eta'$ based on state $T$, $Q$, $\eta$ and evaluates the numerator of Bayes formula
    - New phylogenetic tree $T'$ is proposed using specialized tree-space proposal distribution
    - Other parameters are real scalar variables and new states can be proposed via random sampling, uniform random walks, etc.
- Require a set of proposal distributions $q_i(T|T)$, where $T$ is a point in the space of rooted time trees

## 3.5 Maximum likelihood and testing

### 3.5.1 Search tree space

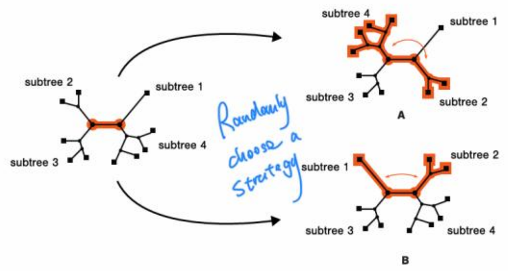
- Propose different **unrooted** trees: NNI, SPR, and TBR moves
- Propose different branch lengths: multiply each branch length by some factor
- Use "**hill-climbing**" strategies to find the optimum

Subtree Prune and Regraft (SPR) and
Tree move by Bisection and Reconnection (TBR)



In SPR, a subtree is detached and then reattached to a different location. For TBR, the tree is broken into two subtrees, which are then joined at a random pair of edges.

Nearest Neighbour Interchange (NNI) move

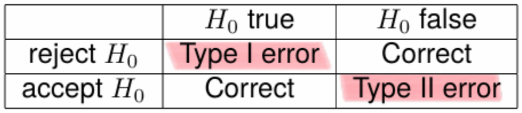


Each internal branch in the tree connects two pairs of nearest neighbours. In the NNI move each pair swaps a nearest neighbour with the other.

### 3.5.2 Likelihood ratio test (nested null hypothesis)

- Assume data evolved under a model $H_0$ with likelihood function $L_0$
- Consider a model $H_1$ (likelihood function $L_1$) which $H_0$ is nested (JC69 & K80, TN93 & HKY)
- MLE under $H_0$ is $\hat{\theta}_0$ and under $H_1$ is $\hat{\theta}_1$
- $\log\left[\left(\frac{L_1(\hat{\theta}_1)}{L_0(\hat{\theta}_0)}\right)^2\right] = 2\left(\log L_1(\hat{\theta}_1) - \log L_0(\hat{\theta}_0)\right) \sim \chi^2_{df}$
- The degree of freedom $df$ if the difference between the number of parameters in the general $H_1$ and in the nested $H_0$ model
- Example: test JC69 against GTR, we can only do the test on the same tree with fixed branch lengths. Because different trees have different parameters, thus the $H_0$ is not nested.
- Example: die rolling
    - $k$ is the event of rolling a die and get a 6
    - $n = 1000$ is the number of trials
    - $H_0: \theta_0 = \frac{1}{6}, H_1: \theta_1 \in (0,1)$
    - $\hat{\theta}_1 = \frac{k}{n}$
    - $2\left(\log L_1\left(\hat{\theta}_1 = \frac{k}{n}\right) - \log L_0\left(\hat{\theta}_0 = \frac{1}{6}\right)\right) > \chi^2_{1,5\%} = 3.84$, reject null at $\alpha = 0.05$

**Model testing error**

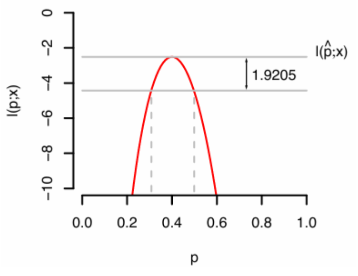


| | $H_0$ true | $H_0$ false |
|---|---|---|
| reject $H_0$ | Type I error | Correct |
| accept $H_0$ | Correct | Type II error |

- Accuracy = 1 – Type I error (**$H_0$ true but rejected**)
    - The Type I error is the significance level, and thus the accuracy is controlled by setting $\alpha$
- Power = 1 – Type II error (**$H_0$ false but accept**)
    - Power can be generally only be assessed via simulating under the general model $H_1$ and assessing the number of times that $H_0$ is accepted

### 3.5.3 Akaike information criterion (AIC) (non-nested null hypothesis)

- $AIC = -2\log L_i(\hat{\theta}_i) + p_i$, where $p_i$ is the number of parameters and $L_i$ the likelihood function of model $i$
- Workflow
    - Calculate the AIC for each model
    - Choose the model with the lowest AIC (smallest expected KL-divergence to the truth)
- $AIC \in [1,2]$: substantial support
- $AIC \in [4,7]$: considerably less support
- $AIC > 10$: essentially no support



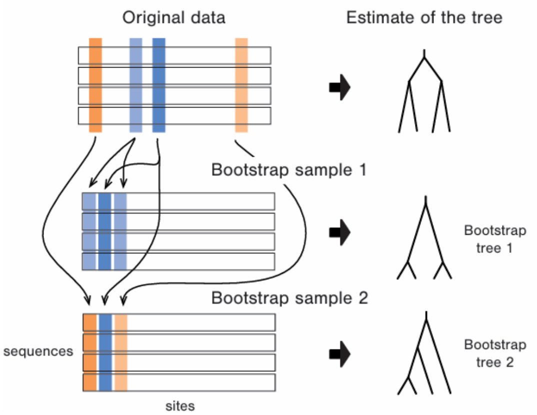
### 3.5.4 Confidence intervals

- Determine the value of the log-likelihood function at $\hat{\theta}$, $l(\hat{\theta}; x)$
- 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 $l(\theta; x) = l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$

### 3.5.5 Bootstrapping for phylogenies

Bootstrapping for phylogenies based on an alignment sequences with length $m$:

- Sample $m$ sites at random with replacement
- Infer a phylogeny based on the new data
- Repeat this procedure many times



# 4. Phenotypic evolution

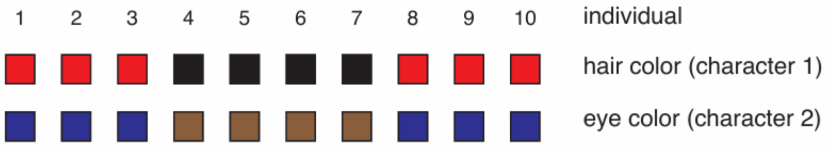
How to compare phenotypic traits / characters between individuals / species that evolved on a phylogeny? Discrete characters: hair color, eye color, etc. Continuous characters: height, surface to weight ratio, etc.

Analogies between models for evolution on discrete and continuous character space:

| discrete | continuous |
|---|---|
| probability to visit any state | probability density on state space |
| memorylessness due to Markov Chain model | memorylessness due to Brownian motion |
| transition probabilities scale with time | variance scales with ~~branch length~~ time |

## 4.1 Discrete characters: Fisher's exact test

Example: We want to know whether eye color is correlated with hair color. We examine 10 individuals:



To test whether there is a true correlation we need to perform a statistical test. In this situation we can apply **Fisher's exact test** with a significance level of 0.05:

$\mathcal{H}_0$: Having brown eyes is equally likely among red- and black-haired individuals.

With Fisher's exact test, we test whether the observed result happened due to chance alone:

contingency table:

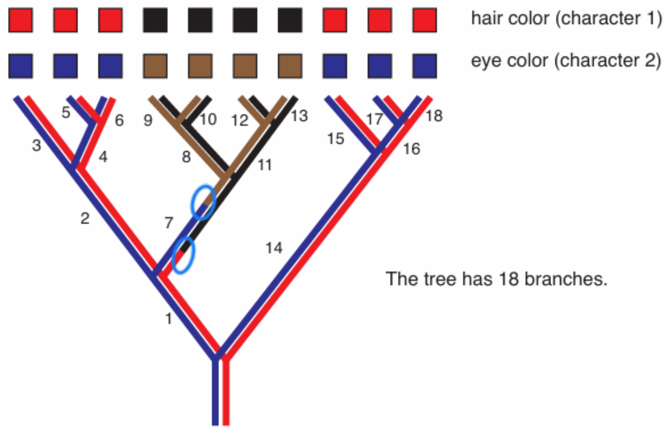
| hair \ eyes | 🟫 | 🟦 |
|---|---|---|
| 🟥 | 0 | 6 |
| ⬛ | 4 | 0 |

$$P(\text{red/brown}) = \frac{(\#\ \text{◩ in 🟥})\times(\#\ \text{◩ in ⬛})}{\#\ \text{comb 🟫 amongst all}}$$

$$= \frac{\binom{6}{0}\binom{4}{4}}{\binom{10}{4}} = 0.0048 < 0.05$$

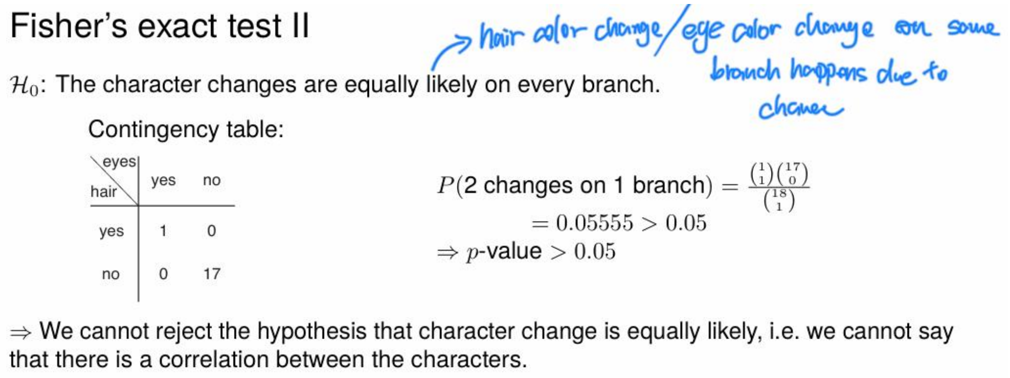
$\Longrightarrow$ We reject the hypothesis of independent evolution, i.e. we can see a correlation.

However, the analysis could be biased due to relatedness of the individuals:



The tree has 18 branches.

**Correct way to look at the problem:** Is the change of characters on the branches correlated?



Fisher's exact test II

$\mathcal{H}_0$: The character changes are equally likely on every branch.

→ hair color change/eye color change on some branch happens due to chance

Contingency table:

| hair \ eyes | yes | no |
|---|---|---|
| yes | 1 | 0 |
| no | 0 | 17 |

$$P(\text{2 changes on 1 branch}) = \frac{\binom{1}{1}\binom{17}{0}}{\binom{18}{1}}$$
$$= 0.05555 > 0.05$$
$$\Rightarrow p\text{-value} > 0.05$$

⇒ We cannot reject the hypothesis that character change is equally likely, i.e. we cannot say that there is a correlation between the characters.

**Important: Neglecting the phylogenetic background** can lead to **false conclusions on correlations between characters**. This is mainly the case because of the **non-independence of species data points as a result of shared ancestry**.

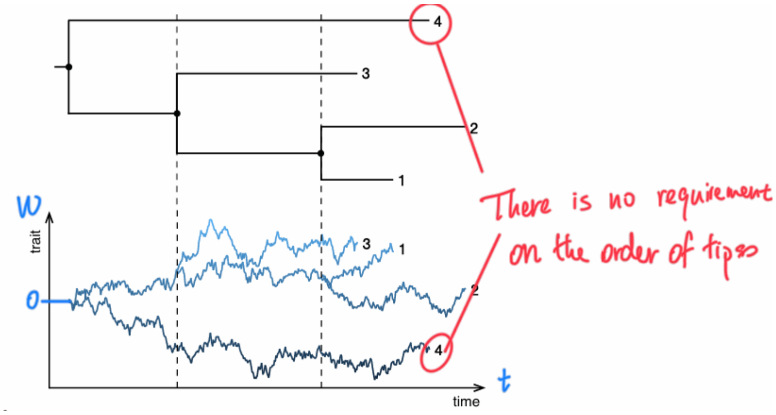
## 4.2 Continuous characters: Brownian motion

Linear regression cannot be used directly here because it will have clade effects:

- The characters share common evolutionary history (not independent realizations)
- The error (variance added by Brownian motion) is not equally distributed

Instead, Brownian motion model is commonly used to describe the evolution of continuous traits on a phylogeny.

- Brownian motion is a continuous time stochastic process on a continuous state space
- $(W_t)_{t \in T}$ fulfills the following conditions
  - $W_0 = 0$, i.e., the process starts at 0
  - $W_t$ is almost surely **continuous**
  - $W_t$ has independent increments (**memoryless** process)
  - $W_t$ has **normally distributed increments** with 0 mean and a variance which increases linearly with time

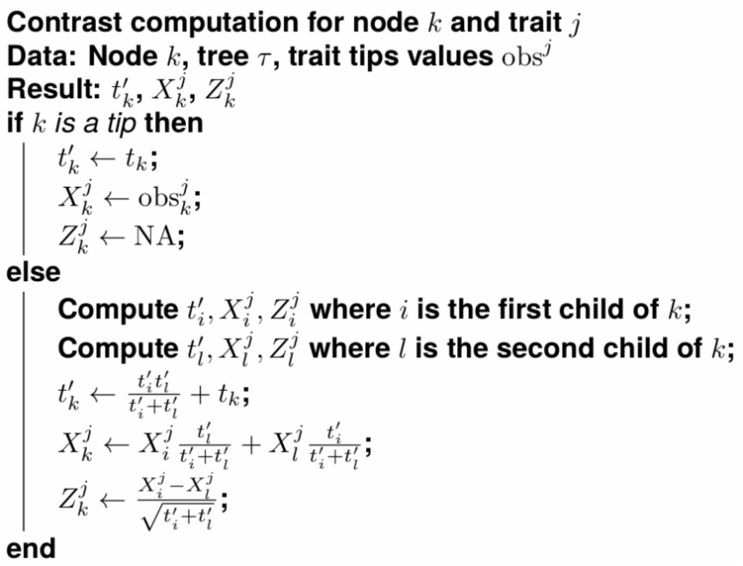


There is no requirement on the order of tips

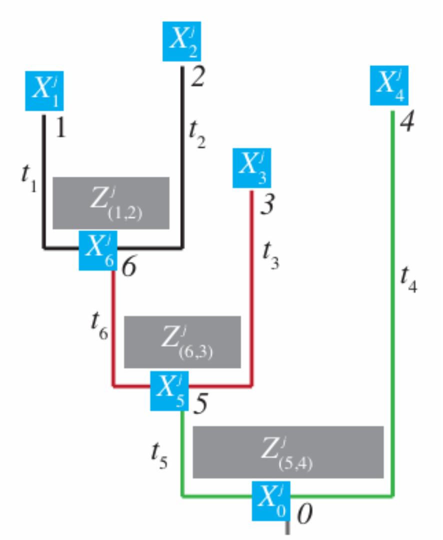
## 4.3 Independent contrasts (to allow us use linear regression)

- The contrasts $Z^j_{(1,2)} = X^j_1 - X^j_2$, $Z^j_{(6,3)} = X^j_6 - X^j_3$, etc. are mutually independent
- Assume character evolution according to Brownian motion
- Observe tip values but have to estimate the values at the internal nodes
- $Var[aX - bY] = a^2 Var[X] + b^2 Var[Y] - 2abCov[X,Y]$

- $Var\left[Z^j_{(1,2)}\right] = Var\left[X^j_1 - X^j_2\right] = Var\left[X^j_1\right] + Var\left[X^j_2\right] - 2Cov\left[X^j_1, X^j_2\right] = \sigma^2(t_1 + t_6 + t_5) + \sigma^2(t_2 + t_6 + t_5) - 2\sigma^2(t_6 + t_5) = \sigma^2(t_1 + t_2)$



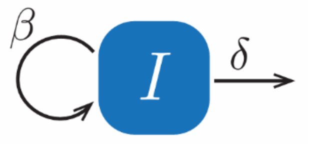
```
Contrast computation for node $k$ and trait $j$
Data: Node $k$, tree $\tau$, trait tips values $\mathrm{obs}^j$
Result: $t'_k$, $X^j_k$, $Z^j_k$
if $k$ is a tip then
    $t'_k \leftarrow t_k$;
    $X^j_k \leftarrow \mathrm{obs}^j_k$;
    $Z^j_k \leftarrow \mathrm{NA}$;
else
    Compute $t'_i, X^j_i, Z^j_i$ where $i$ is the first child of $k$;
    Compute $t'_l, X^j_l, Z^j_l$ where $l$ is the second child of $k$;
    $t'_k \leftarrow \frac{t'_i t'_l}{t'_i + t'_l} + t_k$;
    $X^j_k \leftarrow X^j_i \frac{t'_l}{t'_i + t'_l} + X^j_l \frac{t'_i}{t'_i + t'_l}$;
    $Z^j_k \leftarrow \frac{X^j_i - X^j_l}{\sqrt{t'_i + t'_l}}$;
end
```



# 5. Population processes

- **Population dynamics** models the birth and death of individuals (species, infected hosts, cells, and languages). **The birth and death process gives rise to a phylogenetic tree**.
- **Phylodynamic** aims to understand and quantify the population dynamics based on a phylogenetic tree. Today, we quantify birth and death dynamics given the phylogenetic tree and also the $R_0$.

## 5.1 Birth-death phylodynamic



- $\beta$: Rate of birth of new individuals per individual in $I$
- $\delta$: Rate of death per individual in $I$
- $\beta\Delta t$: The probability of giving birth to another individual in a very small time step $\Delta t$
- $\delta\Delta t$: The probability of dying in a very small time step $\Delta t$
- The waiting time to a birth event is **exponentially distributed** with parameter $\beta$
- The waiting time to the first event (birth or death) is **exponentially distributed** with $\beta + \delta$
- Consider the fact of $N$ individuals
  - The waiting time to the first event (birth or death) is exponentially distributed with parameter $N(\beta + \delta)$

**A simple phylodynamic model**

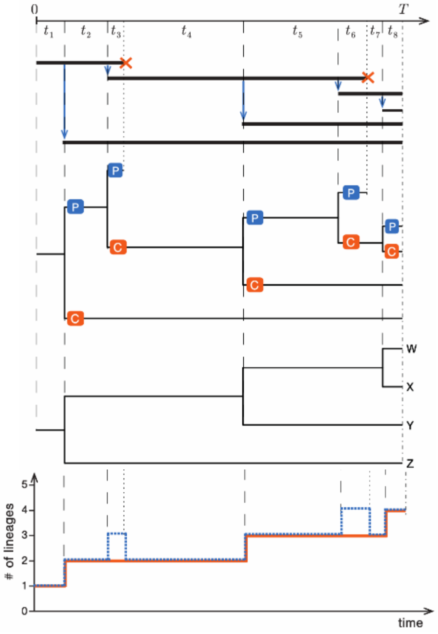
A phylodynamic model adds a sampling process of individuals to the population dynamics. The simplest phylodynamic model is:

- Birth rate $\beta$
- Death rate $\delta$
- Process duration $T$
- Extant tip sampling probability $\rho$
- Extinct tip sampling probability $\phi$

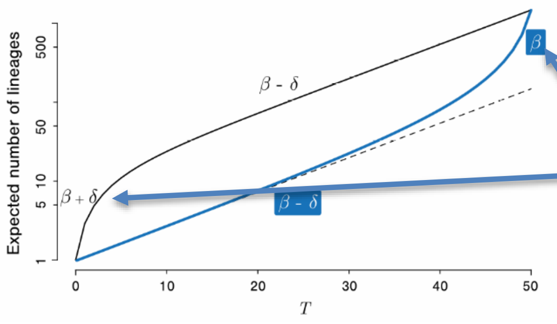
Assume $\rho = 1, \phi = 0$, meaning no fossil sampling and complete extant species sampling. The subtree of complete population tree connecting the sampled individuals, and ignoring the parent-children labels, is called the phylogenetic tree, which is inferred from the data.

We can also plot the number of lieages (y-axis) and time (x-axis) is called the **lineages-through-time (LTT)** plot.

- Blue line: shows the population size through time
- Orange line: shows the number of surviving lineages through time.



**Estimate the birth and death rate from trees (average LTT for a population of age $T = 50$)**



- **Blue**: average number of lineages in the phylogenetic tree.
- **Solid black**: average of surviving population trajectories.
- **Push of the past**: only individuals with a quick replication early on will produce surviving population.
- **Pull of present**: very recent lineages have not yet had time to go extinct.

- **Dashed line**: represents the expected number of individuals.
- Fit a regression line to the early branching events, the slope is about $\beta - \delta$
- Fit a regression line to the late branch events, the slope is about $\beta$
- **Problem**: it is not clear how to incorporate the variances into the regression, and how to choose the time interval for the two regression lines.

**Birth-death phylodynamic likelihood**

**For $\boldsymbol{p(0|t)}$, i.e., no offspring:**

- $p(0|t+\Delta t) = \delta\Delta t + \beta\Delta t\, p(0|t)^2 + (1-(\beta+\delta)\Delta t)p(0|t) + \Delta t^2$
    - $\delta\Delta t$ is the **death** event probability, immediately dies after $\Delta t$
    - $\beta\Delta t$ is the **birth** even probability, new split after $\Delta t$ but then both branches die
    - $1-(\beta+\delta)\Delta t$ is the **no event** happening probability, still no offspring
- $\lim_{\Delta t\to\infty} \frac{p(0|t+\Delta t)-p(0|t)}{\Delta t} = \lim_{\Delta t\to\infty} \delta + \beta p(0|t)^2 - (\delta+\beta)p(0|t) + \Delta t = \frac{dp}{dt}$
- $\frac{dp}{dt} = \delta + \beta p(0|t)^2 - (\delta+\beta)p(0|t)$
- $p(0|t) = \frac{\delta\left(1-e^{-(\beta-\delta)t}\right)}{\beta-\delta e^{-(\beta-\delta)t}}$
- Assume $p(0|0) = 0$

**For $\boldsymbol{p(1|t)}$, i.e., 1 offspring:**

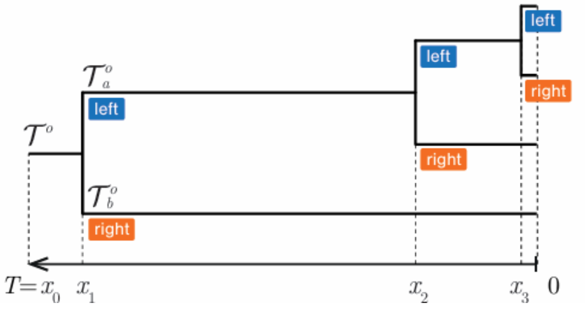
- $p(1|t+\Delta t) = \beta\Delta t\, p(1|t)p(0|t) + \beta\Delta t\, p(0|t)p(1|t) + (1-(\beta+\delta)\Delta t)p(1|t) + \Delta t^2$
    - $\beta\Delta t\, p(1|t)p(0|t) + \beta\Delta t\, p(0|t)p(1|t)$ is **either** one of the descendants of the **birth** event leading to the **surviving individual after time $\boldsymbol{t}$**
    - $(1-(\beta+\delta)\Delta t)p(1|t)$ is **no event** happening probability
- $\lim_{\Delta t\to\infty} \frac{p(1|t+\Delta t)-p(1|t)}{\Delta t} = \lim_{\Delta t\to\infty} 2\beta\, p(1|t)p(0|t) - (\beta+\delta)p(1|t) + \Delta t = \frac{dp}{dt}$
- $\frac{dp}{dt} = 2\beta\, p(1|t)p(0|t) - (\beta+\delta)p(1|t)$
- $p(1|t) = e^{-(\beta-\delta)t}\left(1-p(0|t)\right)^2$

**For $\boldsymbol{p(n|t)}$, $\boldsymbol{n} \geq$ 2:**

- $p(n|t) = p(1|t)\left(\frac{\beta}{\delta}p(0|t)\right)^{n-1}$

**Probability density**

- Label each child of a branching event with "left" or "right" – oriented trees
- Easy to convert densities over oriented trees to densities over trees labeled only at the tips (labeled trees)
- $f_l(T) = f_o(T)\frac{2^{n-1}}{n!}$



- Let $p(x_0, x_1)$ be the probability density for a branch length $x_0 - x_1$ extending from an individual at time $x_0$ in the past
- The **probability density of a tree $\boldsymbol{T^o}$** with age $x_0$ is $p(T^o|x_0) = p(x_0,x_1)\beta p(T^o_a|x_1)p(T^o_b|x_1) = \beta^{n-1}\prod_{i=0}^{n-1} p(1|x_i)$
- The **probability of the branch** between $t$ and $x_1$, $p(t,x_1)$ is $p(t+\Delta t, x_1) = (1-(\beta+\delta)\Delta t)p(t,x_1) + 2\beta\Delta t\, p(t,x_1)p(0|t)$

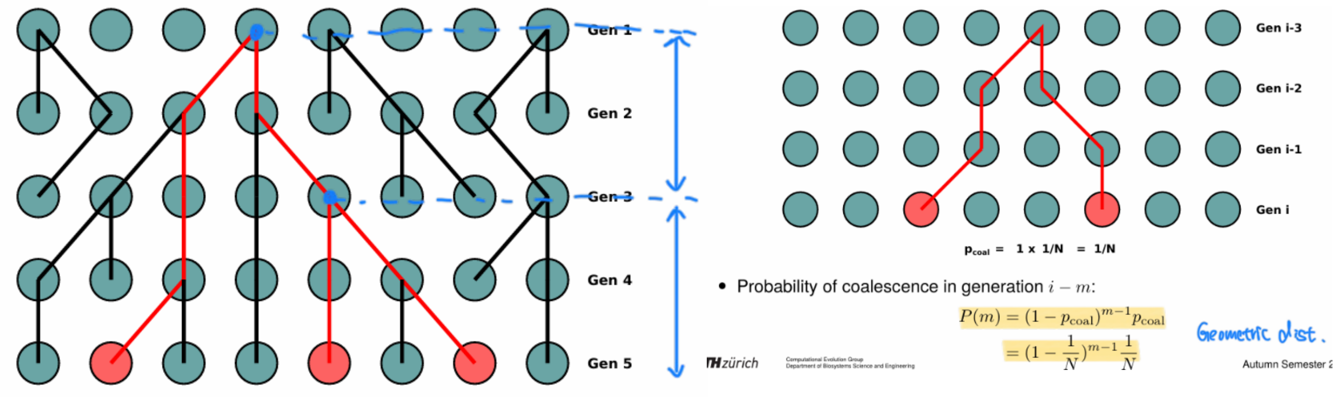
- Assume $p(x_1, x_1) = 1$
- $\frac{d}{dt} p(t, x_1) = -(\beta + \delta) p(t, x_1) + 2\beta p(t, x_1) p(0|t)$
- $p(x_0, x_1) = \frac{p(1|x_0)}{p(1|x_1)}$

## 5.2 Coalescent phylodynamic

- Can be derived as a **limiting distribution** from several population genetic models
- Common assumption is that the underlying population dynamics are **deterministic**
- Often used as **the basis for phylodynamic inference of population size and dynamics**

### 5.2.1 Wright-Fisher process

- **Discrete generation**
- Each generation consists of $N$ individuals
- Each individual in the offspring population choses its **parent uniformly at random** from the $N$ parents
    - A given **parent has a binomially distributed number of offspring**
- For phylogenies of a particular gene, ploidy can be taken into account by multiplying $N$ by a factor which accounts for the number of copies of a gene present in each individual



$p_{coal} = 1 \times 1/N = 1/N$

- Probability of coalescence in generation $i - m$:

$$P(m) = (1 - p_{\text{coal}})^{m-1} p_{\text{coal}}$$
$$= (1 - \frac{1}{N})^{m-1} \frac{1}{N}$$

- $P(m|N) = \left(1 - \frac{1}{N}\right)^{m-1} \frac{1}{N}$ (geometric distribution)
- For large $N$, $P(m|N) \to \exp\left[-\frac{m-1}{N}\right] \frac{1}{N}$ (exponential distribution)
- Note that $\exp\left[-\frac{m-1}{N}\right] = \left(\exp\left[-\frac{1}{N}\right]\right)^{m-1} = \left(1 - \frac{1}{N} + O\left(\frac{1}{N^2}\right)\right)^{m-1}$
- The coalescent in calendar time
    - ***m* is the number of generations**, let ***g* be the calendar time of a generation**. Thus $\Delta t = gm$ is the calendar time span of $m$ generations
    - In calendar time the probability density function for the coalescence time of two lineages is $\frac{1}{gN} e^{-\frac{\Delta t}{gN}}$.
    - In the large $N$ limit, the time to coalescence is exponentially distributed with mean $gN$.
- For sampled k-individual phylogeny: $p_{coal} \approx \binom{k}{2} \frac{1}{N}$

### 5.2.2 Kingman's coalescent process

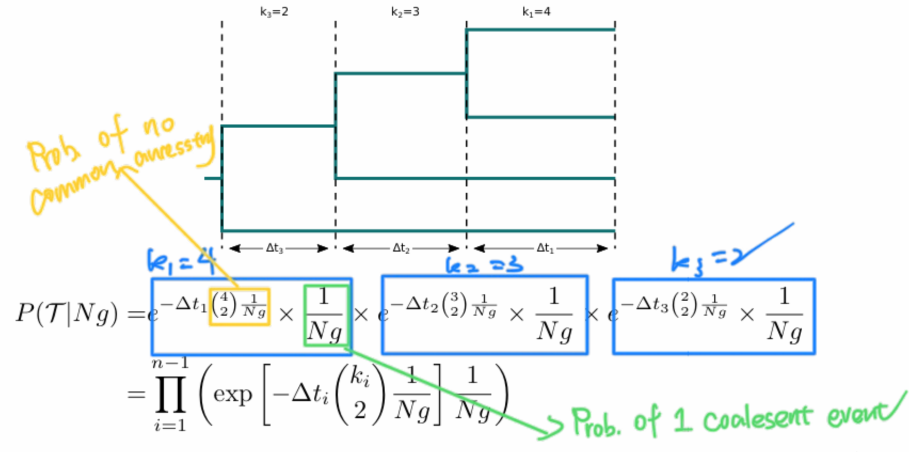
- **Continuous-time Markov process** which produces sampled time trees

- Process occurs **backwards** in time
- Equivalent to sampled trees produced by WF model when $N$ is much larger than the number of samples
- These between coalescence events are drawn from exponential distributions with rate parameters $\binom{k}{2}\frac{1}{N_g}$, i.e., $p(\Delta t|N, g, k) = \exp\left[-\Delta t\binom{k}{2}\frac{1}{N_g}\right]\binom{k}{2}\frac{1}{Ng}$.

**Age of a coalescent tree**

- The average time required for $n$ lineages to coalesce into one is $E[t_{root}] = \sum_{k=2}^{n}\frac{Ng}{\binom{k}{2}} = Ng\sum_{k=2}^{n}\frac{1}{\binom{k}{2}}$
- As the number of lineages $n$ becomes large, we find that $E[t_{root}] \rightarrow 2Ng$
- This is an **upper bound** on the expectation, individual coalescent trees can be older than this

**Probability of a coalescent tree**



$$P(T|Ng) = e^{-\Delta t_1\binom{4}{2}\frac{1}{Ng}} \times \frac{1}{Ng} \times e^{-\Delta t_2\binom{3}{2}\frac{1}{Ng}} \times \frac{1}{Ng} \times e^{-\Delta t_3\binom{2}{2}\frac{1}{Ng}} \times \frac{1}{Ng}$$

$$= \prod_{i=1}^{n-1}\left(\exp\left[-\Delta t_i\binom{k_i}{2}\frac{1}{Ng}\right]\frac{1}{Ng}\right)$$

- The exponentials give the probability of nothing happening in interval $\Delta t_i$
- $1/gN$ factors are the probability densities of the particular coalescent events.

**Inference and effective population size**

- For a given coalescent tree $T$, $L(Ng; T) = P(T|Ng)$
- **Real populations are structured**, but WF population is assumed to be completely **homogeneous**
- The inferred population size is referred to as the **effective population size**
- This is the size of a WF population which **shares some statistical similarity with the real population**

**Robustness of the coalescent**

- The coalescent distribution/process is often derived as a **limit of the Wright-Fisher process**
- It also appears as the limit of many other population processes
- The fact that the coalescent distribution persists in the face of many departures from the WF model is sometimes termed the "**robustness of the coalescent**"

**General assumptions of the coalescent**

- Samples are members of a population that is at **demographic equilibrium** (use of fixed or slowly varying population size)
- **Number of samples is small** compared to the total population size (neglect of > 2 lineages coalescing in the same generation)
- Population is "**well mixed**", samples are drawn **uniformly at random** (the coalescent rate between any pair of sampled lineages being equal, population structure violates this assumption)

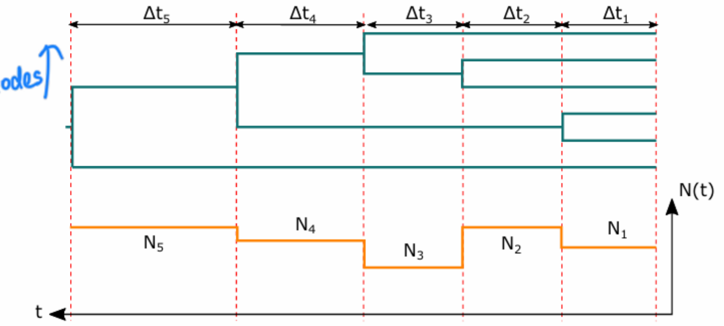
### 5.2.3 Population dynamics

**Parametric population dynamic inference**

- Under a coalescent model with a deterministically varying population size $N(t)$, the probability of a sampled tree becomes $P(T|N(t)g) = \prod_{i=1}^{n-1}\left(\exp\left[-\int_{t_i}^{t_{i+1}}\binom{k_i}{2}\frac{dt}{N(t)g}\right]\frac{1}{N(t_{i+1})g}\right)$, where $t_i$ is the time at the beginning of interval $i$
- For a given parametric form, e.g., $N(t) = N_0 e^{-\gamma t}$, yields $P(T|N_0,\gamma) = L(N_0,\gamma;T)$, i.e., the likelihood for the demographic model parameters
- Thus, we can directly compare and test different demographic scenarios for a given tree

**Non-parametric population dynamics inference**

# params ↑
with tips & nodes ↑

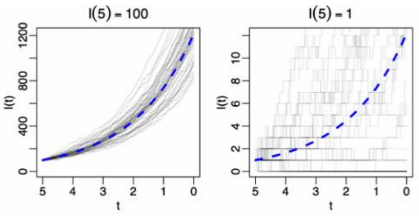


- Assume population has distinct values in each interval between coalescent events.
- Can obtain a separate ML estimate for each population size.
- Resulting population function estimate is the "skyline plot" (Pybus, A. Rambaut, and Harvey 2000).

Can be noisy

**Coalescent approximation of birth-death models**

- Assume the approximate ODE solution $I(t) = I(T)e^{(\beta-\delta)(T-t)}$ for the linear birth-death process is correct
- Birth events occur at time $t$ with overall rate of $\beta I(t)$
- Every birth is a potential coalescence between sampled lineages
- Probability of choosing a sampled lineage pair is $\binom{k}{2}/\binom{I(t)}{2}$
- Approximate coalescence rate is $\beta I(t)\frac{k(k-1)}{I(t)(I(t)-1)} \approx \binom{k}{2}\frac{2\beta}{I(t)}$



- Quality of approximation depends heavily on **how well** birth-death population dynamics are approximated by **deterministic ODE solution**
- This approximation can perform very **poorly** when **population size is small**, as it always is at the start of an epidemic

### 5.2.4 Birth-death vs. coalescent models

| Feature | Birth-death | Coalescent |
|---|---|---|
| Concept | - Parameters: transmission rates, removal rates, sampling rates/proportions<br>- Model sampling process (sample times/locations are data) | - Parameters: effective population size (NOT actual population size)<br>- Assumes number of sampled lineages are small ($k \ll N$) |
| Advantages | - Accounts for stochastic variability in population dynamics | - Generally fast likelihood calculations |

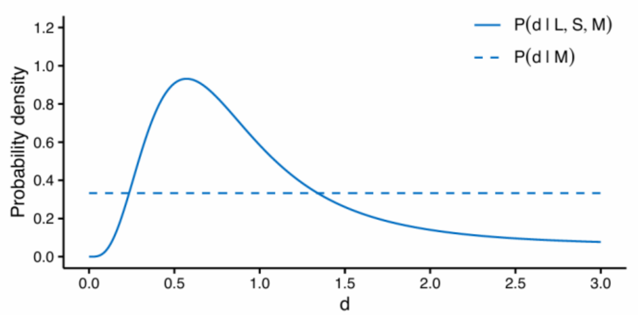
| | | |
|---|---|---|
| | - Generally easier interpretation of parameters<br>- Uses information about sampling | - Easy to extend to complex population dynamics<br>- Naturally account for incomplete sampling |
| Disadvantages | - Sensitive to **unmodeled** changes in sampling fractions<br>- Difficult to extend to complex population models | - Sensitive to **uncertainty** in population dynamics at **high sampling**<br>- Sensitive to **hidden population structure** and **nonrandom sampling** |

## 5.3 Bayesian

### 5.3.1 Bayesian inference

Suppose we have two sequences and the number of substitutions is $S = 4$ and the total number of sites is $L = 10$, with JC69 transition probabilities $p_{ij}(d) = \frac{1}{4} + \frac{3}{4}\exp\left(-\frac{4}{3}d\right)$ if $i = j$, $p_{ij} = \frac{3}{4} - \frac{3}{4}\exp\left(-\frac{4}{3}d\right)$ if $i \neq j$. So the probability for the pairwise alignment is $P[S|d, L] = \left[\frac{1}{4} + \frac{3}{4}\exp\left(-\frac{4}{3}d\right)\right]^{L-S} \times \left[\frac{3}{4} - \frac{3}{4}\exp\left(-\frac{4}{3}d\right)\right]^{S}$. The probability that a position is different is $p = 3p_1(t)$. We define $d = 3\lambda t$ (the expected distance in $t$).

- Note that our model $M$ has provided us the probability of the number of segregating sites $S$ given the genetic distance $d$ and the length $L$ of the sequences $P(S|d, L, M)$
- The Bayesian interpretation of probabilities means that it is sensible to talk about the probability of $d$ given $S$ and $L$: $P(d|S, L, M)$
- $P(d|S, L, M) = \frac{P(S,d|L,M)}{P(S|L,M)} = \frac{P(S|d,L,M)P(d|L,M)}{P(S|L,M)}$
- $P(d|L, M) = P(d|M)$ quantifies the knowledge of $d$ in the absence of the observation
- $P(S|L, M)$ is the distribution over possible numbers of segregating sites given the JC69 model and any independent knowledge of $d$
- Assume our prior information here is $P(d|M) = \frac{1}{3}$ for $0 \leq d \leq 3$ and 0 otherwise.



### 5.3.2 Bayes theorem

$$P(\theta_M|D, M) = \frac{P(D|\theta_M, M)P(\theta_M|M)}{P(D|M)}$$

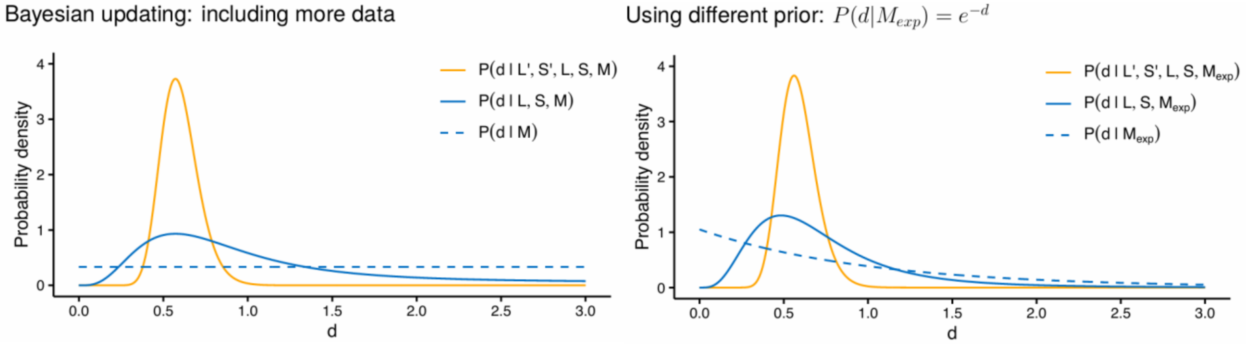
- $\theta_M$ are parameters of some model $M$, while $D$ are data assumed to be generated by the same model.
- $P(\theta_M|M)$ is the prior for the model parameters
- $P(D|\theta_M, M)$ is the likelihood of the parameters given the data
- $P(D|M)$ is the marginal likelihood of the model, DIFFICULT to calculate due to $P(D|M) = \int P(D|\theta_M, M)P(\theta_M|M)d\theta_M$ the integration

- $P(\theta_M|D,M)$ is the posterior of the model parameters given the model and the data

**Bayesian updating**

Suppose we have a new alignment with $L' = 90$ and $S' = 48$. We can apply the Bayes theorem with the posterior from the previous analysis as the prior:

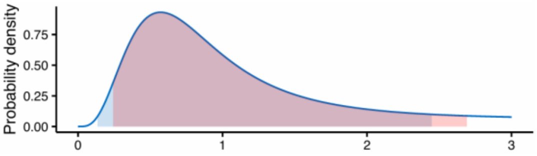
$$P(d|S',S,L',L,M) = \frac{P(S'|d,L',M)P(d|S,L,M)}{P(S'|S,L,L',M)} = \frac{P(S'|d,L',M)P(S|d,L,M)P(d|M)}{P(S'|S,L,L',M)P(S|L,M)} = \frac{P(S',S|d,L',L,M)P(d|M)}{P(S',S|L',L,M)}$$



**Credible interval**

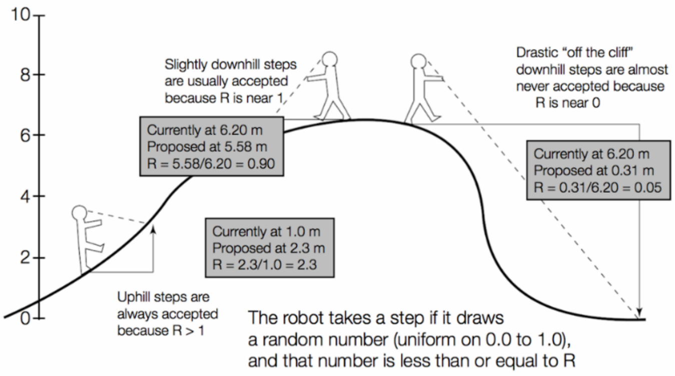
A 95% credible interval is an interval of the posterior distribution containing 95% of the probability. Popular choice include:

- Neglect 2.5% of the samples on both ends of the posterior distribution. Sometimes called the central credible interval.
- Select the smallest interval spanned by 95% of the probability mass. This is the 95% highest posterior density (HPD) interval.
    - Can be found by lowering a threshold density until the area under the curve where the density exceeds the threshold 0.95
    - **Meaning**: the probability of the unknown value falling in this region 95% given the observed data
    - **This is DIFFERENT from a 95% confidence interval**, which is instead an interval produced by a method which generates truth-containing intervals 95% of the time when averaging over all possible datasets.



### 5.3.3 Markov chain Monte Carlo

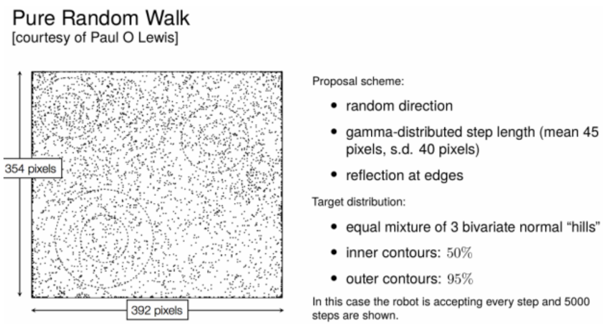
- Monte Carlo methods are algorithms which produce random samples of values in order to characterize a probability distribution


10
8
6
4
2
0
Slightly downhill steps are usually accepted because R is near 1
Drastic "off the cliff" downhill steps are almost never accepted because R is near 0
Currently at 6.20 m
Proposed at 5.58 m
R = 5.58/6.20 = 0.90
Currently at 6.20 m
Proposed at 0.31 m
R = 0.31/6.20 = 0.05
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Uphill steps are always accepted because R > 1
The robot takes a step if it draws a random number (uniform on 0.0 to 1.0), and that number is less than or equal to R

- Usually, the method produces an arbitrary number of independent samples of possible parameter values $\theta_M$ drawn from the posterior distribution $P(\theta_M|D,M)$
- Let $\theta_M$ be the current state, let $\theta'_M$ be a proposed parameter set for the new state
- We calculate $R = \frac{P(\theta'_M|D,M)}{P(\theta_M|D,M)} = \frac{P(D|\theta'_M,M)P(\theta'_M|M)}{P(D|\theta_M,M)P(\theta_M|M)}$
- Drawing a uniform number $u$ on $(0,1)$. We accept the proposed step if $u < R$

**Metropolis-Hastings algorithm**

- The introduced MCMC robot implements the "Metropolis algorithm"
- The robot will produce a sample from the posterior distribution
- The Metropolis algorithm requires that the proposal probability $q$ for a new state $\theta'_M$ given $\theta_M$ satisfies $q(\theta'_M|\theta_M) = q(\theta_M|\theta'_M)$
- The Metropolis-Hastings algorithm allows for non-symmetric proposals by using $R = \frac{P(\theta'_M|D,M)q(\theta_M|\theta'_M)}{P(\theta_M|D,M)q(\theta'_M|\theta_M)}$
- Some other methods:



Pure Random Walk
[courtesy of Paul O Lewis]

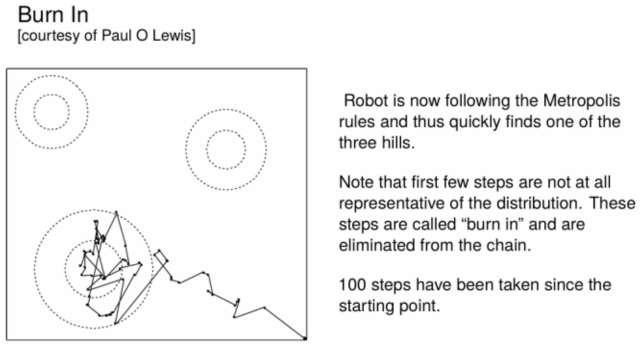
Proposal scheme:

- random direction
- gamma-distributed step length (mean 45 pixels, s.d. 40 pixels)
- reflection at edges

Target distribution:

- equal mixture of 3 bivariate normal "hills"
- inner contours: 50%
- outer contours: 95%

In this case the robot is accepting every step and 5000 steps are shown.

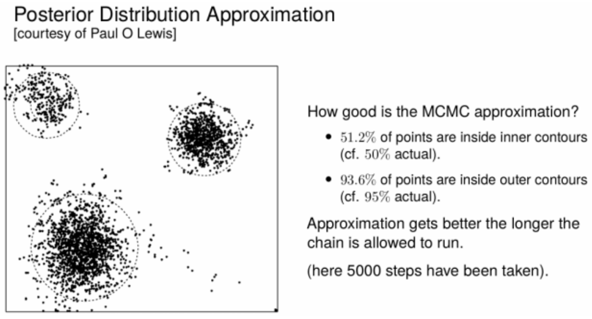


Burn In
[courtesy of Paul O Lewis]

Robot is now following the Metropolis rules and thus quickly finds one of the three hills.

Note that first few steps are not at all representative of the distribution. These steps are called "burn in" and are eliminated from the chain.

100 steps have been taken since the starting point.



Posterior Distribution Approximation
[courtesy of Paul O Lewis]

How good is the MCMC approximation?

- 51.2% of points are inside inner contours (cf. 50% actual).
- 93.6% of points are inside outer contours (cf. 95% actual).

Approximation gets better the longer the chain is allowed to run.

(here 5000 steps have been taken).

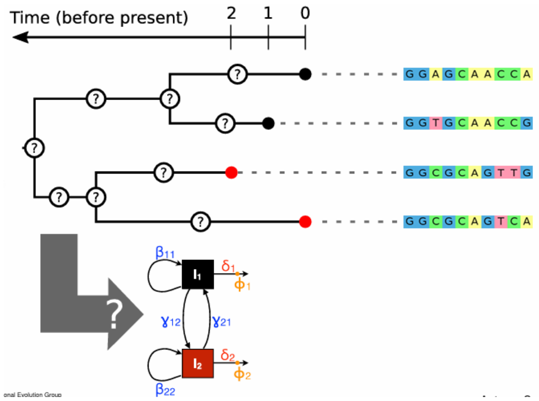
# 5.4 Structured models

## 5.4.1 Population structure

A population is structured if its members possess one or more traits (e.g. location, group membership, …) that affects their phylodynamic parameters (e.g., birth rate, death rate, sampling rate, coalescence rate).

- Plays an important role in shaping the phylogenetic relationships between samples
- Failing to account for existing structure in phylodynamic analyses can bias results
- We can also learn about parameters of structured models (e.g., local birth / death rates and sub-population sizes) using structure-aware phylodynamic models.

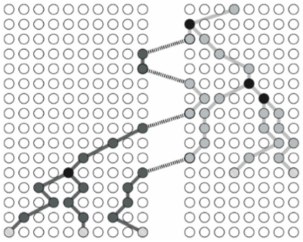
## 5.4.2 Structured birth-death models



- $\beta_{11}$ is the transmission (birth) rate of drug sensitive strains
- $\beta_{22}$ is the transmission (birth) rate of dug resistance strains
- $\gamma_{12}$ is the rate of resistance evolution
- The $\beta_{ij}$ represent the rate at which individuals of type $i$ **produce** individuals of type $j$, while $\gamma_{ij}$ is the rate at which individuals of type $i$ **becomes** type $j$
- Different compartments may represent pathogen strains, geographic locations, host risk groups, etc.

**Multi-type birth-death phylodynamic likelihood**

- Let $p_i(t)$ be the probability that an individual of type $i$, gives rise to no sampled descendants after time $t$.
- $P_i(t+\Delta t) = \left(1 - \Delta t\left(\delta_i + \sum_j(\beta_{ij} + \gamma_{ij})\right)\right) p_i(t) + \Delta t\delta_i + \sum_j \beta_{ij} p_i(t) p_j(t)\,\Delta t + \sum_j \gamma_{ij} p_j(t)\Delta t$
    - $\left(1 - \Delta t\left(\delta_i + \sum_j(\beta_{ij} + \delta_{ij})\right)\right) p_i(t)$ stands for no event happens
    - $\Delta t\delta_i$ stands for death event happens
    - $\sum_j \beta_{ij} p_i(t) p_j(t)\,\Delta t$ stands for birth event happens but dies after $t$, $p_j$ for the other individual, both of them have to be extinct later on
    - $\sum_j \gamma_{ij} p_j(t)\Delta t$ stands for migration to $j$.
- Satisfies the ODE: $\frac{d}{dt} p_i(t) = -\left(\sum_{j=1}^{d}(\beta_{ij} + \gamma_{ij}) + \delta_i\right) p_i(t) + \sum_{j=1}^{d} \beta_{ij} p_i(t) p_j(t) + \sum_{j=1}^{d} \gamma_{ij} p_j(t) + \delta_i$
- Bayesian inference for multi-type birth-death models
- $P(T,Q,\eta|A,L) = \frac{P(A|Q,T)P(T,L|\eta)P(Q)P(\eta)}{P(A,L)}$
    - $T$ is a phylogenetic tree
    - $L$ are the locations / types associated with the sequences
    - $\eta$ are the parameters of the multi-type birth-death model
    - $P(T,L|\eta)$ is the structured phylodynamic likelihood
    - $A$ represents the sequence alignment
    - $Q$ the substitution rate matrix



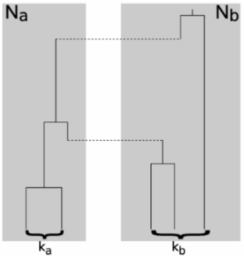
### 5.4.3 Structured coalescent models

**Structured Wright-Fisher model**

- Assume a **single population is divided into sub-populations** (demes) of size $N_i$ for $i \in [1,d]$
- Allows for migration between demes at rate $q_{ij}$ (**forward**)
- Assume a fixed time interval $g$ between successive generations

- Assume the **subpopulation sizes are unaffected by migration** in the long term

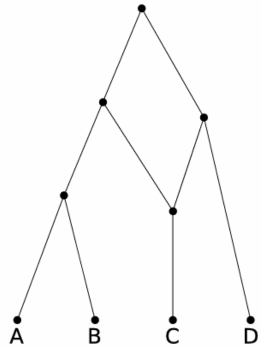
**Structured calescent**



- Coalescence rate in deme $i$: $\binom{k_i}{2}\frac{1}{gN_i}$
- Migration rate (backward) $i \to j$: $k_i m_{ij}$
- The backward-time migration rate $m_{ij}$ (immigration rate) is related to the forward time rate $q_{ij}$ from the structured WF model by $m_{ij} = q_{ji}\frac{N_j}{N_i}$
- For a symmetric 2 deme model ($N_1 = N_2 = N$ and $m_{12} = m_{21} = m$) we can have the expected time for two lineages to coalesce
    - Define $T_d$ and $T_s$ as the expected coalescence times when lineages are in different / same demes respectively
    - Lineages in distinct demes cannot coalescence, so we have $T_d = \frac{1}{2m} + T_s$
    - Lineages in the same deme wait for average time $1/(2m + 1/Ng)$ before either coalescing or migrating: $T_s = \frac{1}{2m+1/Ng} + \frac{1/Ng}{2m+1/Ng}0 + \frac{2m}{2m+1/Ng}T_d$
    - Solving this pair of simultaneous equations yields: $T_s = 2Ng$ and $T_d = \frac{1}{2m} + 2Ng$
- Bayesian inference for structured coalescent models
- $P(T_{col}, Q, \theta | A, L) = \frac{P(A|Q,T_{col})P(T_{col}|\theta,L)P(Q)P(\theta)}{P(A|L)}$
    - $T_{col}$ is a phylogenetic tree with ancestral location marked
    - $L$ are the locations / types associated with the sequences
    - $\theta$ are the parameters of the structured coalescent model
    - $P(T_{col}|\theta, L)$ is structured coalescent likelihood

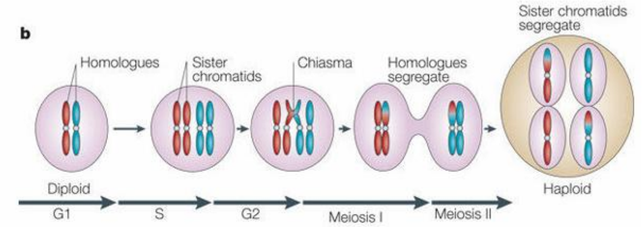
## 5.5 Phylogenetic networks

Just for a tree, phylogenetic networks represent a wide range of evolutionary relationships:



- For **species**, the network represents species ancestry and nodes with multiple parents represent hybridization or horizontal gene transfer (HGT) events
- For **individuals**, the network represents ancestry of individual lineages and nodes with multiple parents represent either hybridization, HGT or simply a node in a pedigree (family tree) of a sexually reproducing organism
- For **genes or chromosomes**, the network represents ancestry of sequence data and nodes with multiple parents represent recombination events
- There are **infinite** number of possible ancestral network topologies

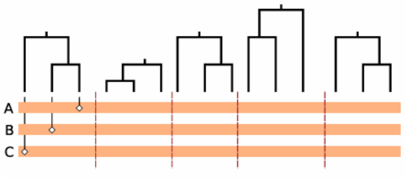
**Sexual reproduction and genetic linkage**



- Genetic linkage is the tendency for nearby sites to be inherited together
- For sexually reproducing organisms, sites on different chromosomes are completely unlinked

- **Sites on the same chromosome are inherited together** unless a homologous recombination event divides them

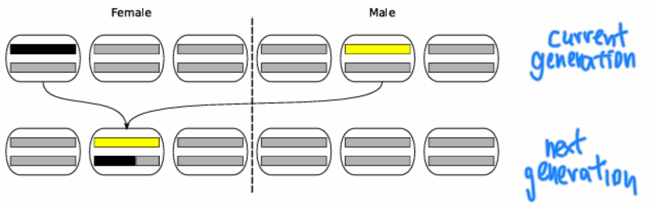
**Effect of recombination on phylogeny**



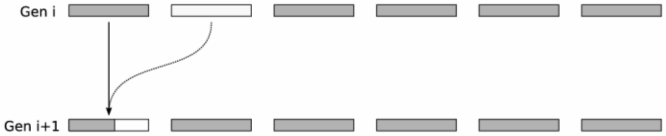
- Different sites correspond to different trees
- The further away sites are on the alignment, the more likely they are to possess different ancestry
- Single nucleotide polymorphisms (SNPs) are usually widely separated and are thus assumed to be completely unlinked (independent) – necessary for GWAS
- Short gene sequences often assumed to be completely linked (one tree for all sites)
- Even for asexual entities (viruses, bacteria, etc.) reality is usually somewhere between these extremes

### 5.5.1 Wright-Fisher with recombination

- Focus on a small segment of a single autosome
    - An autosome is a chromosome which is a member of a homologous pair, i.e., not a sex chromosome

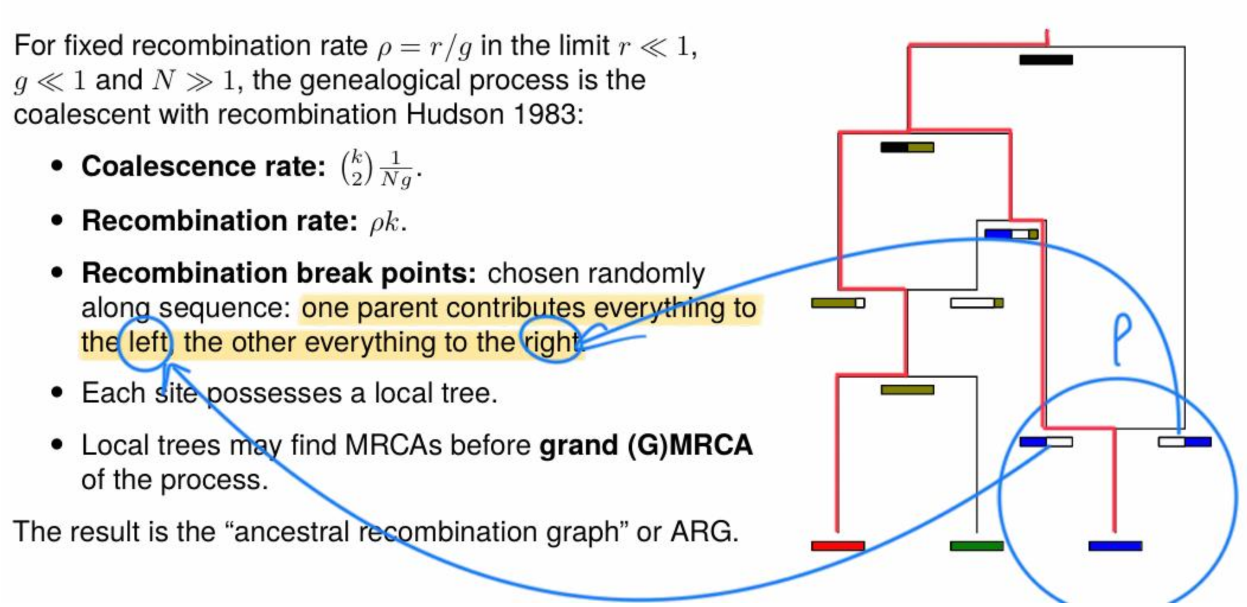


- Each child selects 1 male and 1 female parent randomly from the previous generation
- With probability $r$ (which depends on the segment length) the homologous pair from one of the parents is recombined
- Since the specific pairing of chromosomes only matters over a single generation, in the long term the haploid approximation is very good:



- Each child in $i + 1$ selects a parent at random from generation $i$
- With probability $r$, an additional parent is selected
- In this case, a break-point is chosen randomly on the chromosome, and everything to the right is replaced by the homologous section of the second parent's chromosome

### 5.5.2 The coalescent with recombination



For fixed recombination rate $\rho = r/g$ in the limit $r \ll 1$, $g \ll 1$ and $N \gg 1$, the genealogical process is the coalescent with recombination Hudson 1983:

- **Coalescence rate:** $\binom{k}{2}\frac{1}{Ng}$.
- **Recombination rate:** $\rho k$.
- **Recombination break points:** chosen randomly along sequence: one parent contributes everything to the left, the other everything to the right.
- Each site possesses a local tree.
- Local trees may find MRCAs before **grand (G)MRCA** of the process.

The result is the "ancestral recombination graph" or ARG.

### 5.5.3 Bayesian – Network phylodynamic posterior

$$P(Q,\rho,N,Q|A) = \frac{1}{P(A)}P(A|G,Q)P(G|\rho,N)P(\rho,N,Q)$$

- $G$ is the recombination graph/matrix
- $Q$ is the substitution rate matrix
- $\rho$ is the recombination rate
- $N$ is the effective population size

Sampling from this distribution is difficult since:

- Some features of $G$ do not contribute to the likelihood (unidentifiable)
- The likelihood surface contains many distinct peaks
- The volume of the space of phylogenetic networks is extremely large

## Appendix 1: Summary of substitution models

To convert substitution rate matrix $Q$ to a transition probability matrix $P$, we apply $P(t) = e^{Qt}$.

| Name | Params | Substitution rate matrix Q | Note |
|---|---|---|---|
| JC69 | 1 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & \lambda & \lambda & \lambda \\ \lambda & \cdot & \lambda & \lambda \\ \lambda & \lambda & \cdot & \lambda \\ \lambda & \lambda & \lambda & \cdot \end{bmatrix} \end{array}$ | - All substitutions have the same rate $\lambda$<br>- $P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}$<br>- $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$<br>- $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$<br>- Stationary, when $t \rightarrow \infty$, $p_0 = p_1 = 0.25$ |
| K80 | 2 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & \alpha & \beta & \beta \\ \alpha & \cdot & \beta & \beta \\ \beta & \beta & \cdot & \alpha \\ \beta & \beta & \alpha & \cdot \end{bmatrix} \end{array}$ | - Transitions (A <-> G / T <-> C) happen at rate $\alpha$<br>- Transversions happen at rate $\beta$ |
| TN93 | 3+3 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \cdot \end{bmatrix} \end{array}$ | - Transitions of T <-> C happen at $\alpha_1\pi_{\{TCAG\}}$<br>- Transitions of A <-> G happen at $\alpha_2\pi_{\{TCAG\}}$<br>- Transversions happen at $\beta\pi_{\{TCAG\}}$<br>- $\pi_{\{TCAG\}}$ is the nucleotide equilibrium frequency |
| HKY | 3+2 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha\pi_A & \cdot \end{bmatrix} \end{array}$ | - Variant of TN93<br>- When $\alpha_1 = \alpha_2$, TN93 becomes HKY |
| GTR | 6+3 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{bmatrix} \end{array}$ | - Generalized time-reversible model<br>- $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$ |
| UNREST | 12 | $\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} \cdot & a & b & c \\ d & \cdot & e & f \\ g & h & \cdot & i \\ j & k & l & \cdot \end{bmatrix} \end{array}$ | - Each substitution has a (different) rate<br>- Complicated<br>- Not time-reversible |

## Appendix 2: Phylogenetic tree reconstruction methods

| Name | Input/Output | Note | Runtime |
|---|---|---|---|
| UPGMA (Phenetic - distance based/overall similarity) | - **Input**: distance matrix (substitution model needed)<br>- **Output**: ultrametric tree (rooted) | Assumptions:<br>- All sequences must come from the **same time point**<br>- Evolution according to **strict molecular clock, constant evolution rate** over time<br>- Can use Neighbor-joining algorithm to relax these | - Prune nodes $n$ times<br>- $n^2$ calculation per pruning<br>- $O(n^3)$ |
| Least square (Phenetic - distance based/ overall similarity) | - **Input**: distance matrix (substitution model needed)<br>- **Output**: unrooted tree with smallest square error (outgroup required to form a rooted tree) | - Propose all unrooted trees<br>- Calculate the square error<br>- $n$ tips unrooted tree has $2n-3$ branches<br>- $n$ tips rooted tree has $2n-2$ branches<br>- $n$ tips, $(2n-5)!!$ unrooted trees<br>- $n$ tips, $(2n-3)!!$ rooted trees | - $O(e^{nlnm})$ |
| Parsimony (Cladistic - character based) | - **Input**: sequence alignment<br>- **Output**: unrooted tree with lowest parsimony score (outgroup required to form a rooted tree) | - Consider all unrooted trees<br>- Calculate parsimony scores, $4^{n-1}m$ internal character assignments | - $4^{n-1}m$, $n$ tips and $m$ sequences<br>- $O(e^n)$ |
| Fitch parsimony (Cladistic - character based) | Same as parsimony | - Parsimony score $k=0$<br>- For each site, if $C_l \cap C_r \neq \emptyset$, assign $C_l \cap C_r$ to the site, if $C_l \cap C_r = \emptyset$, assign $C_l \cup C_r$ to the site and $k += 1$ | - $n-1$ internal nodes<br>- $m$ sequences<br>- $(n-1)m$ |
| MLE (Mechanistic - character based/evolutionary model) | - **Input**: sequence alignment<br>- **Output**: unrooted tree (outgroup required to form a rooted tree) | - Phylogenetic likelihood calculation requires substitution models<br>- Multiply over all sites $O(m)$<br>- Sum over internal nucleotides at $n-1$ internal nodes $O(4^{n-1})$<br>- Multiply over branches $O(2n-2)$ | - $O(m4^n n)$ |
| Felsenstein MLE (Mechanistic - character based/evolutionary model) | Same as MLE | - Use dynamic programming to speed up<br>- Recursion $O(n)$, m sites $O(m)$ | - $O(mn)$ |
| Bayesian (Mechanistic - character based/evolutionary model) | - **Input**: sequence alignment, phylogenetic tree, substitution rate matrix<br>- **Output**: rooted tree | - Assume separating the process of tree generation from that of sequence evolution implies the sequence evolution is effectively neutral | Very slow due to the marginal probability – the integration |

## Appendix 3: Tree counting

| Name | # Internal nodes | # Branches | # Topologies |
|---|---|---|---|
| Unrooted tree | $n-2$ | $2n-3$ | $(2n-5)!!$ |
| Rooted tree | $n-1$ | $2n-2$ | $(2n-3)!!$ |
| Caterpillar tree | $n-1$ | $2n-2$ | $n!/2$<br>There are $n!$ ways to order the tips, but because the single cheery at the bottom can be flipped without changing the topology, so divided by 2 |
| Ultrametric tree | $n-1$ | $2n-2$ | $(2n-3)!!$ |

**Newick string:**

- Rooted tree: $2^{n-1}$ ways
- Unrooted tree: $(2n-3)\cdot 2^{n-1}$

**Note**: !! double factorial means multiplying every odd number, for example, $5!! = 1 \times 3 \times 5 = 15$.

## Appendix 4: JC69 properties

We use the Binomial distribution because we treat every nucleotide site as an **independent Bernoulli** trial (Match vs. Mismatch). This assumption allows us to calculate how “certain” or “uncertain” our estimated distance is (i.e., the confidence interval).

**MLE for JC69**

Given:

- $P(t) = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}$
- Prob. of the nucleotide remains the same after time $t$: $p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}$
- Prob. of the nucleotide changes after time $t$: $p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}$
- Prob. of a position is different $p = 3p_1(t)$
- Expected distance in time $t$ is $d = 3\lambda t$
- Suppose have 2 sequences with length $n$ and $x$ differences
- $L(d; x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x}$
- $l(d; x) = \log\binom{n}{x} + x \log\left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right) + (n-x) \log\left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)$
- $\boxed{\hat{d}_{JC69} = -\frac{3}{4}\log\left(1 - \frac{4x}{3n}\right)}$

**JC69+Γ**

- To solve the problem of rate heterogeneity across sites. In reality, it is not possible that every single nucleotide site evolves at the exact same speed.
- Prob. of a position is different $p = \frac{3}{4} - \frac{3}{4}e^{-4\lambda Rt} = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$, $R \sim \Gamma(\alpha, \alpha)$
- Prob. of a substitution at one site $\mathbb{E}[p] = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{\alpha}\right)^{-\alpha}$
- $l(d; x) = \binom{n}{x}(\mathbb{E}[p])^x (1 - \mathbb{E}[p])^{n-x}$
- $\boxed{\hat{d}_{JC69+\Gamma} = \frac{3}{4}\alpha\left(\left(1 - \frac{4}{3}\hat{p}\right)^{-\frac{1}{\alpha}} - 1\right)}$
- $\hat{d}_{JC69+\Gamma} \geq \hat{d}_{JC69}$, ignoring site variation tends to lead to underestimation of the sequence distance

**Amino Acids (AA)**

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

**Codon sun**

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

## Appendix 5: Wright-Fisher vs. Kingman

| Feature | Wright-Fisher | Kingman |
|---|---|---|
| Core concept | Sampling with replacement from the previous generation | A stochastic process describing time-to-common-ancestor |
| Time scale | Discrete | Continuous (Markov process) |
| Direction | Usually forward-in-time (simulate population then trace back) | Strictly backward-in-time (start with samples, find root) |
| Population limit | Finite, fixed $N$, small | Assume $N \to \infty$ |
| Merger events | Multiple & simultaneous mergers possible | Strictly binary pair mergers |
| Math basis | Geometric dist: $(1 - p_{coal})^{m-1} p_{coal}$ | Exponential dist: $\lambda e^{-\lambda t}$ |
| Key parameter | Probability ($p$) per generation | Rate ($\lambda$) per unit time |
| Runtime | $O(N)$ depends on the population size | $O(k)$ depends only on sample size |
| Equation | $p_{coal} = \binom{k}{2} \frac{1}{N}$ | $\lambda = \binom{k}{2} \frac{1}{Ng}$ |
| Primary use | Modeling small populations, forward simulations with selection | Phylogenetics (BEAST), inferring population history from DNA, large-scale simulations |

# Appendix 6: Model assumption summary

**Phylogenetic**:

- UPGMA:
    - **Strict molecular clock**: rate of evolution is consistent across all lineages
    - **Ultrametricity**: the output tree is ultrametric
- Least square
    - **Additivity**: evolutionary distances are "additive", the distance between two tips should equal to the sum of the branch lengths connecting them
    - **Distance reliability**: the input distance matrix accurately reflects the true number of substitutions
- Parsimony
    - Occam's Razor: the simplest explanation (**fewest mutations**) is the best
    - **Rare changes**: mutations are rare events
    - **Site independence**: assume each site evolves independently of the others
- Maximum likelihood (Felsenstein)
    - Model correctness: the substitution model correctly describes the underlying evolutionary process
    - **Site independence**: every site in the alignment evolves independently
    - **Markov property**: evolution is a "memoryless" process
    - **Stationary/reversible**: the base frequencies are constant over time and the process is time-reversible
- Bayesian inference
    - All ML assumptions
    - Valid prior distribution
    - **MCMC convergence**: the MCMC chain must run long enough to converge to the true posterior

**Phylodynamic & population dynamic:**

- Birth-death model
    - **Forward-in-time**: evolves forward with distinct speciation / transmission events
    - Sampling strategy: the $\rho$ and $\phi$
    - **Rate homogeneity**: assume birth and death rates are constant over time (basic model)
- Coalescent model
    - Wright-Fisher:
        - **Discrete generations**: non-overlapping, integer time steps
        - Constant size: standard WF assumes population size $N$ is constant
        - Panmixia: assumes random mating (any individual can pick any parent)
    - Kingman's coalescent:
        - **Continuous generations**
        - Large population: assumes $N \to \infty$
        - Small sample size: $k \ll N$
        - **Binary mergers**: assumes only two lineages coalesce at a time
        - **Neutrality**: assumes no natural selection is acting on the locus

**Structured birth-death & structured coalescent**

- **Distinct subpopulation**: the population is divided into defined groups (demes)

- **Restricted gene flow**: lineages can only coalesce if they are in the same deme at the same time
- **Migration**: assumes movement between demes occurs at specific rates $m$
- **Panmixia within demes**: assumes that within a deme, the population is well-mixed (Kingman)

**Phylogenetic networks**

- **Non-tree-like evolution**: assumes that the evolutionary history cannot be represented by a single bifurcating tree due to reticulate events
- **Recombination**: assumes a child can **have genetic material from two different parents** (not single parent assumption for standard trees)
- **Breakpoints**: assumes specific points along the sequence where the **ancestry switches from one parent to another**

## Appendix 7: Evolution, phylogenetics, and phylodynamics

| Case | Molecular evolution | Phylogenetics | Phylodynamics |
|---|---|---|---|
| General | How does the genetic information of individuals change through time? | How are the individuals related? | What are the population dynamics giving rise to the individuals in the phylogeny |
| Marcoevolution | Genetic information and morphology of species changes through time | Phylogeny displays species relationship | Population dynamics is the speciation and extinction process |
| Epidemiology | Genetic information of pathogens changes through time | Phylogeny displays transmission history | Population dynamics is the transmission and recovery process |