Friday, 12 February 2021

Computational Biology

Explanation of Colour Coding

- Black: Normal things to remember
- Dark Grey: Algorithms
- Blue: Things that I forget
- Green: Things I feel it's important / pros of something
- Red: Drawback of something
- Orange: Time or space complexity of one algorithm

Lecture 01

- 名词解释
  - Macroevolution: evolution above the species level, an individual in macroevolution is one species
  - Gene: entity encoding a phenotype
  - Allele: version of a Gene
  - Genotype: the collection of genes of one individual
  - Dominant Allele: The Allele that determines the phenotype
  - Recessive allele: The other allele is called recessive allele
  - DNA may change due to
    - Point mutations
    - Recombinations
    - Insertions
    - Deletions
  - Modern Darwin Evolution
    - Multiplication of DNA leads to offspring
    - Variations in the multiplication of DNA leads to variations in phenotype
    - Heredity of phenotype occurs due to DNA being passed on
    - Competition between phenotypes to survive and multiply occurs.
    - Neglect the environmental effect, which gives room to Epigenetics

Friday, 12 February 2021

- Q & A
  - Name one weakness and one strength for each of the different sequencing methods
    - **Answer:**

Sequencing Technology	Advantages	Disadvantages
Sanger/First Generation	Very low error rate	Very slow
Next Generation	Higher speed	Low maximum read length
Third Generation / Nanopore	No maximum read length & No amplification	High error rate

  - With which of the methods presented could you conceivably sequence the genome of a particular cell?
    - **Answer:** Nanopore Sequencing, because it does not require amplification.
  - Imagine you want to perform a paternity test. How would you go about testing whether the potential father is really the father? Could you think of reasons to find a false-negative relatedness?
    - Sequence both the father and the son and make a Sequence Comparison. One possible false negative relatedness might be mutations and sequence errors.

Lecture 02 — Sequence alignments and BLAST

- 名词解释
  - Codon: three nucleotides encode for one amino acid
  - Homologous: Sequences with shared ancestry
  - Global Alignment: aligns one sequence to the other from start to the end

- Local Alignment: finds the longest subsequences with highest similarity

- Dynamic Programming: an algorithmic technique for solving an optimisation problem by breaking it down into simpler subproblems and utilising the fact that the optimal solution to the overall problem depends upon the optimal solution to its subproblems.
- BLAST: Basic Local Alignment Search Tool

- General things to remember
  - Alignment are based on the idea that there is a common ancestor from which the genes have evolved.
  - Ways to find a pairwise sequence alignment
    - Qualitative method : Dot-matrix method
    - Dynamic programming
      - Needleman-Wunsch: global alignment
      - Smith-Waterman: local alignment
      - Heuristic and fast methods: BLAST
  - Number of possible alignments of two sequences.
    - Let $\alpha = \alpha_1\alpha_2 \dots \alpha_m$ be a sequence of length m and $b = b_1b_2 \dots b_n$ a sequence of length n, we assume $n \leq m$. Let k be the number of gaps to be introduced into sequence α
      - There are $\binom{m+k}{k}$ possibilities to place k gaps between m letters of α
    - Since gaps can not align to gaps, and there are only $m + k - n$ gaps to be inserted in b, we have $\binom{m}{m+k-n}$ possibilities to place the number of gaps in b at the non-gaped positions of α
      - In total, we have $\binom{m+k}{k}\binom{m}{m+k-n}$ possibilities of alignment with respect to k.
    - K can take a value for 0 to n, so in total, we have

$$\sum_{k=0}^{n} \binom{m+k}{k} \binom{m}{m-n+k}$$

- 算法
  - **The dot-matrix method**
    - The algorithm:
      - Arrange two sequences in matrixes
        - Gap in matrix -> gap in sequence
        - Repeated patterns / repeated blocks -> repeats
        - “Reflected diagonals” -> inversions
    - Pros and Cons
      - Pros: visually easy method to identify sequence features such as indels, repeats, inversions and inverted repeats
      - Cons: time-consuming, Does not give one optimal alignment
  - **Dynamic Programming Method**
    - **Smith-Waterman Algorithm for local alignment**
      - Initialisation
        - 0th row and 0th column: set to 0
        - Remaining rows and columns correspond to nucleotides
      - Iteratively calculate entry $H(i,j)$ of matrix H using the following formula, also denote the direction from where the optimal alignment comes.

$$H(i,j) = max \begin{cases} 0 & \\ H(i-1,j-1) + s(i,j) & \text{match or mismatch } \searrow \\ H(i-1,j) + w & \text{gap in seqB } \downarrow \\ H(i,j-1) + w & \text{gap in seqA } \rightarrow \end{cases}$$

      - Start from the highest score and walk backwards until a 0 is reached.

```
Initialize empty alignments ali_A and ali_B;
Initialize current position (p_A, p_B) such that
  score_matrix[p_A, p_B] = max(score_matrix);
while score_matrix[p_A, p_B] > 0 do
    if path_matrix[p_A, p_B] = diag then
        Prepend character seq_A[p_A − 1] to ali_A;
        Prepend character seq_B[p_B − 1] to ali_B;
        (p_A, p_B) ← (p_A − 1, p_B − 1);
    else if path_matrix[p_A, p_B] = left then
        Prepend a gap to ali_A;
        Prepend character seq_B[p_B − 1] to ali_B;
        (p_A, p_B) ← (p_A, p_B − 1);
    else if path_matrix[p_A, p_B] = up then
        Prepend character seq_A[p_A − 1] to ali_A;
        Prepend a gap to ali_B;
        (p_A, p_B) ← (p_A − 1, p_B);
end
```

- Pros and Cons
  - Pros: fast in comparison to brute force; finds the optimal local alignment
  - Cons: only pairwise alignment possible (not suitable for global alignment); still too slow for scanning against big libraries.

- **Needleman-Wunsch Algorithm for Global Alignment**
  - Initialise the 0th row and 0th column according to:

$$H(0,j) = j * w \text{ and } H(i,0) = i * w$$

  - Iteratively calculate $H(i, j)$ using the following formula

$$H(i,j) = max \begin{cases} H(i-1,j-1) + s(i,j) & \text{match or mismatch } \searrow \\ H(i-1,j) + w & \text{gap in seqB } \downarrow \\ H(i,j-1) + w & \text{gap in seqA } \rightarrow \end{cases}$$

  - Always Start from the bottom right field $(m\ .\ n)$ and follow the path up to the top left field $(0,0)$

- **BLAST: Basic Local Alignment Search Tool**
  - Algorithms
    - Split the query sequence into k-mers
    - Search these k-mers in the database sequences allowing mismatches but scored. Eg. Match +5, mismatch -3
    - Keep only the highest score k-mer alignment
    - Extend the alignment on both directions, keep track of the scores
    - Whenever score is below a threshold, delete the alignment
    - Keep only the alignments that are above the threshold
    - Report these database sequences
  - Pros and Cons
    - Pros: faster than Smith-Waterman & Needleman-Wunsch; allows exact match and also some similarity up to a pre-defined degree.
    - Cons: does not guarantee the optimal pairwise alignments; Can only find genes/sequences that are already available in the database.

- **Multiple sequence alignment (MSA)**
  - *Ad hoc* approach
    - Pairwise alignment against a reference strain
    - Define a reference strain for the genome and pairwise align app sequins with the reference strain.

    - Pros and Cons
        - Pros: position numbering is the same for each sequence
        - Cons: only possible when one knows which species the sequences come from
    - Extension to Smith-Waterman algorithm
        - Extend Smith-Waterman algorithm into more dimensions
        - Extremely SLOW: k sequences of length m required m^k steps.

- Q&A
    - What are the weaknesses and strengths of the different alignment methods (dot-matrix method, Smith-Waterman, Needleman-Wunsch, BLAST)?
        - **Answer:**

Alignment methods	Strengths	Weaknesses
Dot-Matrix	Easy to detect indels; repeats; inversions; inverted repeats	Time consuming; Does not give on optimal alignment
Smith-Waterman	Faster than brute-force; gives optimal local alignment;	Only pairwise alignment; still not fast enough for alignment to a large database
Needleman-Wunsch	Faster than brute-force; gives optimal Gobal alignment;	Only pairwise alignment; still not fast enough for alignment to a large database
BLAST	Fastest; exact match while allowing to a certain degree of similarity	Do not guarantee the optimal pairwise alignment; Can only find genes/sequences that are already in the database.

    - With Which alignment methods do you get an optimal alignment?
        - Smith-Waterman gives the optimal local alignment
        - Needleman-Wunsch gives the optimal Global alignment
    - Do you obtain the same alignments when using different scoring schemes in the Smith-Waterman and Needleman-Wunsch algorithms?
        - No. Different scoring schemes usually means weighting different kinds of mutation (gap or mismatches) differently. This leads to different alignment. However, if we just time all the penalty and match score by a factor, we would get the same alignment.

Lecture 03 - GWAS and molecular evolution

- 名词解释
    - GWAS: Genome Wide Association studies

- p-value: Goven a random variable X and a realisation x. Let up assume a null hypothesis H_0, which is a statement on the distribution of X. The p-value is the probability of observing x or a more extreme realisation under the assumption the null hypothesis was true
  - $p - value := P(X = x \text{ or more extreme value} \mid H_0)$
- Significance level: The significance level α, is a value that one defines before performing a statistical test. If p-value lies below this significance level, the observed result of the random experiment cannot support the null hypothesis which leads to rejection of the null hypothesis
- Hyper geometric distribution: Assume an urn with r red and s black balls, k balls are drawn without replacement. How is the number of red balls among the k drawn balls, R_k, distributed?
  - $$P(R_k = i) = \frac{\binom{r}{i}\binom{b}{k-i}}{\binom{r+b}{k}}$$
- Fisher's exact test: Statistical test to examine the significance of the association between two kinds of classifications

	Case	Control	Total
Minor	a	b	a+b
Major	c	d	c+d
	a+c	b+d	n=a+b+c+d

  - H_0: Class A is not linked to class B. (The number of individuals expressing both A_1 and B_1 is based on chance
  - $$p - value = \sum_{i=a}^{a+b} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{n}{a+c}}$$
    - Fisher's exact test is a fixed margin test, meaning that margins like $a+b;\ c+d;\ a+c;\ b+d$; are **FIXED**
- Pearsons χ^2-test
  - Fisher's exact test only works for small numbers because $\binom{n}{k}$ can not be calculated correctly for bigger numbers
  - Pearsons χ^2 -test
    - One calculates the deviance between observed and expected numbers on hypergeometric distribution

- To test whether this result is significant, we calculate the following test statistic: $\chi = \sum_{i,j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- Pearson proved that this statistic is χ^2-distributed
- The p-value is then the probability of a X^2-distributed random variable being $\geq \chi$
- Degree of freedom is calculated by $(\#\text{ of rows} - 1) * (\#\text{ of columns - 1})$
- E.g

Observed	Case	Control	Total
minor	2104	2676	4780
Major	1896	3324	5220
total	4000	6000	10000

Expected	Case	Control	Total
minor	A : 1912	B:2868	4780
Major	C:2088	D:3132	5220
total	4000	6000	10000

- $A = 10000 * \frac{4000}{10000} * \frac{4780}{10000} = 1912$
- $\chi = 61, p - value = 5.055 * 10^{-15}$

- Rate & probability
  - Rate: measures events per time unit
    - Deterministic, fixed quantity
    - Birth rate, substitution rate
    - Describes averages
  - Probability: measures the chance that a random event occurs
    - Stochastic
    - Obtaining 6 when throwing a die, time to substitution
    - Describes an exact event

- General things to remember
  - GWAS:
    - Case-control setup
      - Two large group of individuals: one healthy control group, and one group with a certain disease
      - All individuals are genotyped for the majority of known SNP locations

- Molecular evolution models

- 算法
  - GWAS:
    - Statistical analysis
      - Case control setup: two groups (i) patients with disease, (ii) healthy patients
      - For each SNP, count the number of healthy individuals without this specific mutation, H_N, and with the mutation H_S, as well as the number of diseased individuals without the mutation, D_N, and with the mutation, D_S

	Healthy	Diseased
Mutation	H_S	D_S
No Mutation	H_N	D_N

      - Calculate the odds ratio (OR) of diseased and healthy people with respect to the abundance of each SNP

$$OR = \frac{D_S/H_S}{D_N/H_N}$$

      - Example

Observed	Case	Control	Total
minor	2104	2676	4780
Major	1896	3324	5220
total	4000	6000	10000

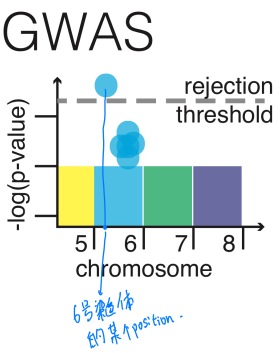
From the above table, we have $H_N = 3324$, $H_S = 2676$, $D_N = 1896$, $D_S = 2104$,

Then the odds ratio of this setup would be

$$\begin{aligned} OR &= \frac{D_S/H_S}{D_N/H_N} \\ &= \frac{2104/2676}{1896/3324} \\ &= 1.38 > 1 \end{aligned}$$

Which means there is an increase in odds ratio between this mutation and the disease.

- Statistical significance
  - The odds ratio can only inform us about a potential association between a SNP and a genetic disease
  - How can we test for statistical significance?
    - Contingency table tests to calculate p-value
      - Fisher's exact test
      - Pearson's χ^2 test
- Typical results
  - Graph with the SNP positions (sorted according to chromosome) on the x-axis and -log(p-value) on the y-axis



  - SNPs (as well as chromosomes) with extremely low p-values could be associated with the specific disease
- GWAS drawbacks
  - Missing quality control steps
  - Multiple testing: Correction for multiple testing needed (Bonferroni-correction : set threshold to $\frac{\alpha}{\text{\# of total tests}}$)
  - Correlations only between single SNPs but not between genes tested

- Quantifying variation between sequences
  - Measurements of sequence differences
    - Hamming distance
      - Number of segregating sites
      - Count the sites that vary

    - p-distance

      $$\frac{\text{number of segregating sites}}{\text{sequence length}}$$

- Molecular evolution models
  - Distance between two sequences
    - The distance between two sequences is the expected number of nucleotide substitutions per site
    - This definition includes all evolutionary steps in between two sequences
    - Employ a mathematical model for estimating the distance, as we do not know the evolution steps, this model must include
      - (Stochastic) process modelling the substitution through time
      - Substitution rates
  - Nucleotide substitutions as a Markov chain
    - Definition of a Markov chain
      - Stochastic process, a series of random experiments through time
        - A series of random variables $(X_t)_{t\in\tau}$. If τ is discrete, the Markov chain is called discrete, otherwise continuous
        - The process is called stationary if $(X_{t_1}, X_{t_2}, \dots, X_{t_n})$ and $(X_{t_1+r}, X_{t_2+r}, \dots, X_{t_n+r})$ have the same distribution for all $t_1, t_2, \dots, t_n, r \in T$
      - Lives on a state space and jumps to different states, state space S
      - **Memorylessness**: the probability of jumping to a state only depends on the current states
        $P(X_{t_{n+1}} = x_{t_{n+1}} \mid X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, \dots) = P(X_{t_{n+1}} = x_{t_{n+1}} \mid X_{t_n} = x_{t_n})$
      - If the transition probabilities on the state space do not change over time, the Markov chain is called time homogenous
    - Nucleotide substitution as a Markov chain
      - State space $S = \{T, C, A, G\}$
      - Substitution rate matrix

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

- From rate to probabilities
  - Let α be the rate of an event E happening per unit of time. The probability that this event happens in a very small time step Δt is defined as $\alpha \Delta t$. We denote the time until the event happens with the random variable X. The probability that the event does not happen in Δt is
    - $P(X > \Delta t) = 1 - \alpha \Delta t$
  - Let $\tau = k\Delta t$, we can divide the probability that event E does not happen in τ as
    - $P(X > \tau) = (1 - \Delta t)^k$
    - $= (1 - \Delta t)^{\tau/\Delta t} \rightarrow_{\Delta t \to 0} e^{-\alpha\tau}$
  - The probability that event E happen within time period τ is then
    - $P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$
  - $P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$ is the cumulative density function
  - $f(x) = \frac{dP}{dt}(x) = \alpha e^{-\alpha x}$ is the probability density function of an exponential distribution
    - **An event occurring with rate α means that it occurs after an exponential distributed waiting time with parameter α**
- **From rate matrix to transition probabilities**
  - Let $P(t) = (P_{ij}(t))_{i,j \in S}$ 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. When we look at an infinitesimally time step Δt, in which only one event can happen, we can calculate the transition probability at time $t + \Delta t$ as

    $$P(t + \Delta t) = P(t) + P(t)Q\Delta t \Leftrightarrow \frac{P(t + \Delta t) - P(t)}{\Delta t} = P(t)Q$$

    Let $\Delta t \to 0$, thus it follows:

$$lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t} = \frac{dP}{dt}(t) = P(t)Q$$

This is a differential equation with the solution

$$P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}$$

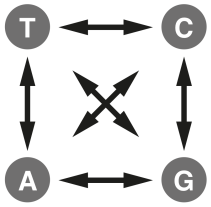
- Every possible series of states is visited in time t

- Q & A
  - In a GWAS, why can you not reject your null hypothesis if the p-value is $< \alpha$?
    - Multiple testing. When testing multiple SNPs(doing statistical tests multiple times), it could be that the p-value is $< \alpha$ only by chance. We need to perform corrections on that p-value before rejecting null hypothesis
    - Bonferroni correction: instead of using α, we use $\frac{\alpha}{\text{\# of tests}}$
  - Why is the Markov Chain model a good model for sequence evolution
    - Memorylessness: a nucleotide substitution happens independently from the substitution history at this site
    - Substitution rate matrix defines the transition probabilities
    - The transition probabilities take into account every possible substitution path
  - Why is it not advisable to reconstruct a phylogeny based on the Hamming distance?
    - # This question was not discussed in the Q&A section, the answers are given by the understanding of Chenxi Nie
    - It can not take all evolutionary trajectories into account.
    - It is more of a static model rather than stochastic process like Markov Chain which has the ability of describing a random process over time.

Lecture 4: Nucleotide, amino acid and codon substitution models

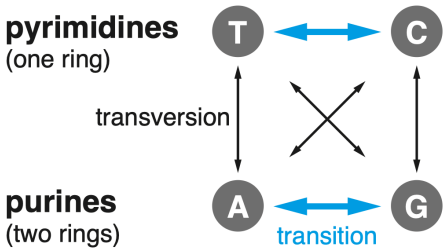
- 名词解释
  - Substitution rate matrices
    - JC69: all substitution have the same rate λ

- $$\begin{pmatrix} & T & C & A & G \\ T & -3\lambda & \lambda & \lambda & \lambda \\ C & \lambda & -3\lambda & \lambda & \lambda \\ A & \lambda & \lambda & -3\lambda & \lambda \\ G & \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$



- K80: transitions happen at α, transversions at rate β
  - Transition: A <-> G; C<->T
  - Transversions: A <-> T; C <-> G

$$\begin{pmatrix} & T & C & A & G \\ T & \alpha - 2\beta & \alpha & \beta & \beta \\ C & \alpha & \alpha - 2\beta & \beta & \beta \\ A & \beta & \beta & \alpha - 2\beta & \alpha \\ G & \beta & \beta & \alpha & \alpha - 2\beta \end{pmatrix}$$



- TN93:
  - Transitions between $T \leftrightarrow C$ happen at rate α_1 * (nucleotide equilibrium frequency)
  - Transitions between $A \leftrightarrow G$ happen at rate α_2 * (nucleotide equilibrium frequency)
  - Transversions happen at rate β * (nucleotide equilibrium frequency)
  - Substitution rate matrix
    - $$\begin{pmatrix} & T & C & A & G \\ T & . & \alpha_1 * \pi_C & \beta * \pi_A & \beta * \pi_G \\ C & \alpha_1 * \pi_T & . & \beta * \pi_A & \beta * \pi_G \\ A & \beta * \pi_T & \beta * \pi_C & . & \alpha_2 * \pi_G \\ G & \beta * \pi_T & \beta * \pi_C & \alpha_2 * \pi_A & . \end{pmatrix}$$
- GTR(REV)
  - Generalised time-reversible model
  - The most generalised time-reversible model

- $$\begin{pmatrix} & T & C & A & G \\ T & . & a*\pi_C & b*\pi_A & c*\pi_G \\ C & a*\pi_T & . & d*\pi_A & e*\pi_G \\ A & b*\pi_T & d*\pi_C & . & f*\pi_G \\ G & c*\pi_T & e*\pi_C & f*pi_A & . \end{pmatrix}$$
  - Pros: Time reversible & Flexibility
  - Cons: not completely general

- UNREST: The most general substitution model

- $$Q = \begin{pmatrix} & T & C & A & G \\ T & . & a & b & c \\ C & d & . & e & f \\ A & g & h & . & i \\ G & j & k & l & . \end{pmatrix}$$
  - Pros: most general case, all other models are special cases of UNREST
  - Cons: Mathmatically very complicated and not handy to use & not time-reversible

- **Time reversible:**
  - A *stationary Markov chain* with rate matrix Q is *time reversible* if and only if the rate matrix can be decomposed into a symmetric matrix $S = (s_{ij})_{i,j\in\{1,2,\ldots,n\}}$ and the diagonal matrix Π. The equilibrium frequencies, also referred to as stationary distribution, are on the diagonals of Π, i.e

$$Q = \begin{pmatrix} s_{1,1} & S_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & S_{2,2} & \cdots & s_{2,n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ s_{n,1} & S_{n,2} & \cdots & s_{n,n} \end{pmatrix} \begin{pmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & \pi_n \end{pmatrix}$$

  - Intuitively, time reversibility can be interpreted such that the probability flux from state i to j must equal the probability flux out of state j to i
  - If a stochastic process is time reversible, it shows the same statistical behaviour forward and backward in time
  - If time reversible rate matrix is adopted, the probability of sequences given a phylogenetic tree does not depend on where the root of the tree is positioned
  - Synonymous substitutions: codon changes but the output amino acid does not change.

- Algorithms
  - Calculating transition probabilities and sequence distance
    - JC69:
      - Given substitution rate matrix:

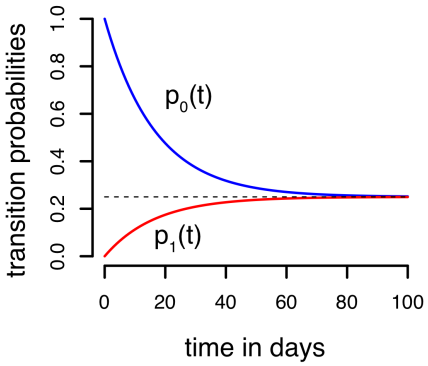
$$Q = \begin{pmatrix} & T & C & A & G \\ T & -3\lambda & \lambda & \lambda & \lambda \\ C & \lambda & -3\lambda & \lambda & \lambda \\ A & \lambda & \lambda & -3\lambda & \lambda \\ G & \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

      - We can now calculate the transition probability matrix P

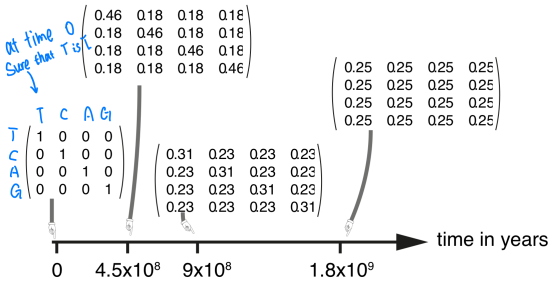
$$P(t) = e^{Qt} = \begin{pmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{pmatrix}$$

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

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



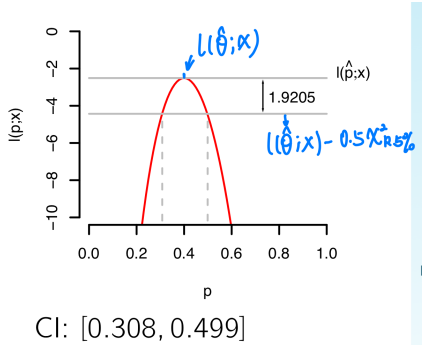
      - Stationary distribution
        - sequence TCAG evolves according to JC69
        - $\lambda = 2.2/3 * 10^{-9} \frac{\text{substitutions per site}}{\text{year}}$
        - What is the probability to observe a certain nucleotide $\alpha_i \in \{T, C, A, G\}$ at position $i \in \{1,2,3,4\}$ after time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



-When $t \rightarrow \infty$ stationary distribution is reached

-Any long sequence at time 0, will be composed of equal amounts of T,C,A,G after $t \rightarrow \infty$

- Maximum likelihood estimators of sequence distance
  - Maximum likelihood estimator: an estimator of a model parameter that maximises the probability to obtain the observed results
  - Confidence interval: if we repeatedly estimated the parameter form realisations of the random experiment and the interval estimate for each realisation of the random experiment, we could expect 95% of these intervals to contain the true parameter.
    - The interval is an estimate itself based on a realisation of a random experiment
    - Confidence interval for parameter estimate can be calculated based on likelihood intervals
      - Let X be a random variable with a distribution parameterised in θ. Based on collected data x of a huge sample, the MLE for the parameter is $\hat{\theta}$. Then, one can show that, $2(l(\hat{\theta}) - l(\theta))$ has a χ^2_k-distribution where k are the degrees of freedom (the vector length of θ)
      - To obtain a 95% confidence interval (CI):
        - Determine the value if the log-likelihood function in $\hat{\theta}$, $l(\hat{\theta}; x)$ ($l(\hat{\theta}; x)$ is a constant number!)
        - Subtract $0.5\chi^2_{k,5\%}$: $l(\hat{\theta}; x)$ - $0.5\chi^2_{k,5\%}$ (also a constant number)
        - Determine those θ values for which $l(\theta; x) = l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$



CI: [0.308, 0.499]

- The 95% Confidence Interval include all θs which are not in the 0.05 tail of the χ^2 distribution. This means that when performing the experiment e.g. 100 times, then 5 times the estimated parameter is expected not to be contained in the 95% Confidence Interval.

- JC69: Maximum likelihood estimator for sequence distance

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

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

  - Given two sequences of length n with x differences. The probability that a position is different is $p = 3p_1(t)$. We define $d = 3\lambda t$ (the expected distance in time t)
  - Thus the probability that x positions out of n are different is

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

    - Note that here x and n are KNOWN variables

  - $l(d;x) = log(L(d;x)) = log\binom{n}{x} + xlog(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}) + (n-x)log(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d})$
  - Setting the first derivative to 0 and solving this equation with respect to d leads to the MLE of the JC69 distance:

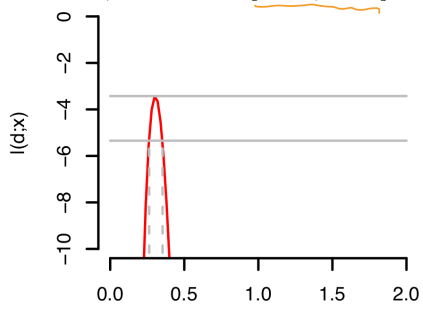
$$\hat{d} = -\frac{3}{4}log(1 - \frac{4x}{3n})$$

  - Example:
    - Distance between
      - Sequence 1: ATTACGAC
      - Sequence 2: TCTACGAC
      - Length of gene n = 8 and defences x = 2
    - Answer : $\hat{d} = -\frac{3}{4}log(1 - \frac{4*2}{3*8}) = 0.3$
    - Confidence interval [0.05, 1.17]

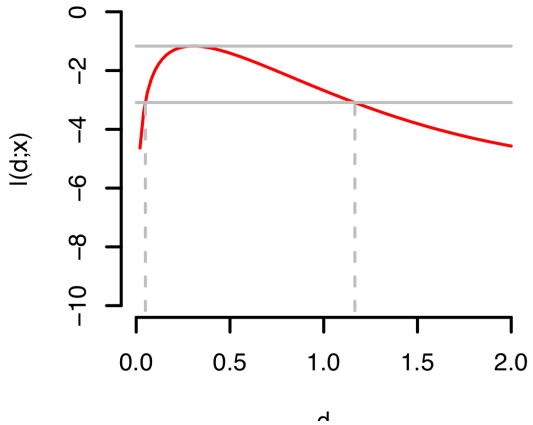
dence interval:

The more data you have, the more confident you can be about estimation (confidence interval ↓ 变小了)

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



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

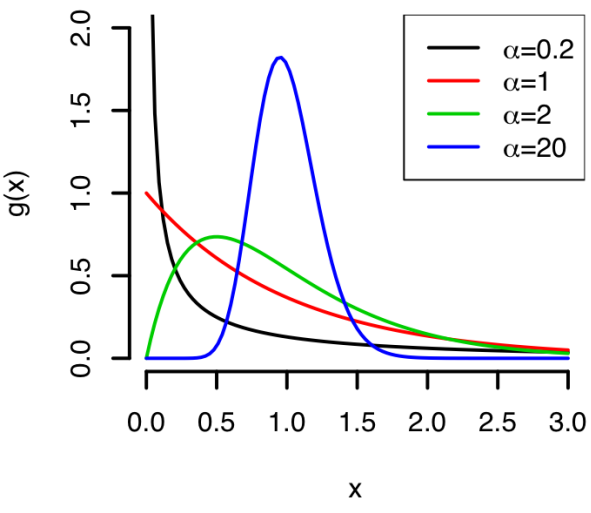


- Variable substitution rates across the genome
  - Substitution rates might differ across the genome
    - Mutation rates might differ across sites
    - Selective pressure might be different on the phenotypic level
  - Extend existing models by replacing the constant rates by Γ-distributed random variable
    - JC69 + Γ, K80 + Γ, ...
    - Sequence distance: expected number of substitutions per site, averaged over all sites
  - Γ-distribution

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

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

☞ Γ-distribution very flexible



- JC69 + Γ
  - Replace the substitution rate λ by λR, where R is a $\Gamma(\alpha, \alpha)$-distributed random variable (mean 1)
  - JC69: $p(d) = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}$
  - JC69 + Γ: $p(d, R) = \frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}dR}$
    - To obtain the probability of a substitution at one site, we calculate

$$\mathbb{E}[p] = \int_0^\infty (\frac{3}{4} - \frac{3}{4}e^{=\frac{4}{3}dr})g(r)dr = \frac{3}{4} - \frac{3}{4}(1 + \frac{4d}{\alpha})^{-\alpha}$$

  - To estimate the distance d, we use the MLE as described before with $L(d; x) = \binom{n}{x}(\mathbb{E}[p])^x(1 - \mathbb{E}[p])^{n-x}$
    - $\hat{d} = \frac{3}{4}\alpha((1 - \frac{4}{3}\hat{p})^{-\frac{1}{\alpha}} - 1)$ where $\hat{p} = \frac{x}{n}$
  - Example
    - Distance between
      - Sequence 1: ATTACGAC
      - Sequence 2: TCTACGAC
      - Site variation: $\Gamma(2,2)$-distributed, $\alpha = 2$
      - Length of gene n = 8 and defences x = 2
    - Answer : $\hat{d} = -\frac{3}{4} * 2((1 - \frac{4 * 2}{3 * 8})^{-\frac{1}{2}} - 1) = 0.34 > 0.3$
    - **Ignoring site variation leads to underestimation of the sequence distance**
- Distance based phylogenetic reconstruction
  - We can now replace the Hamming distance with the evolutionary distance for distance based phylogenetic reconstruction
- Amino acid substitution models
  - Generally the same Markov model is used for amino acid substitutions as for nucleotide substitutions. The transition probability matrix can be calculated according to the equation $P(t) = e^{Qt}$

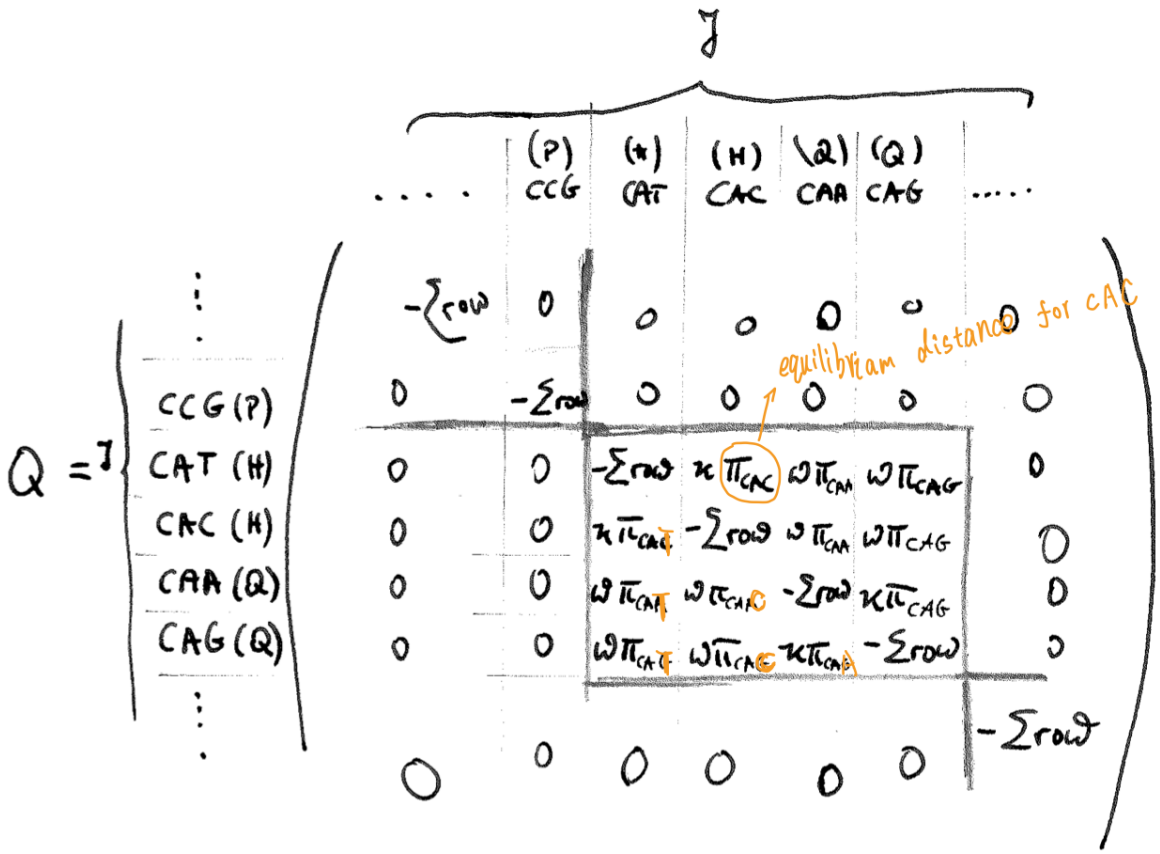
- In the case amino acid, $P(t)$ as well as the substitution rate matrix have dimension 20 * 20. To determine P(t), we need to derive the Q-matrix, which is a bit more complicated than nucleotide substitutions
  - Two approaches
    - Empiric
    - Mechanistic
  - Desired: time-reversibility
- Easiest amino acid substitution
  - All substitutions happen at the same rate λ
  - As we have 20 amino acid, the substitution rate for any substitution is then 19λ
  - With $d = 19\lambda t$ we can also derive the distance with the maximum likelihood approach

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

- Codon substitution models
  - Also model codon substitution with a Markov chain model, with 61 states (no stop codon) are allowed
  - The substitution rate matrix has dimension 61 * 61
  - Denote codons with capital letters and nucleotides with small letters
  - Incorporate
    - k: transition / transversion rate ratio
    - ω: nonsynonymous/synonymous rate ratio
    - π_I: equilibrium frequency of codon I consisting of nucleotides $i_1 i_2 i_3$, with equilibrium frequencies $\pi_{i_1}\pi_{i_2}\pi_{i_3}$, $\pi_I = \frac{1}{c}\pi^*_{i_1}\pi^*_{i_2}\pi^*_{i_3}$
  - Model:

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

- The substitution rate matrix



- Evidence for selection
  - Synonymous substitutions do not change the protein, these substitutions are seen as neutral
  - Nonsynonymous substitutions do change the protein and selective processes can act on the new protein
  - To discover selection, one compares amounts of nonsynonymous and synonymous substitutions
    - d_N: distances at non synonymous codon positions
    - d_S: distance at synonymous codon positions
  - d_N/d_Sratio: Counting method
    - Given two sequences
      - Sequence 1: TTTCCTCCTCCT
      - Sequence 2: TTCCAGCCTCCT
    - We have the following codon table

	Codon 1	Codon 2	Codon 3	Codon 4
Sequence 1	TTT	CCT	CCT	CCT
Sequence 2	TTC	CAG	CCT	CCT

- TTT & TTC -> F; CCT -> P & CAG -> Q; CCT&CCT -> P
- Algorithm
  - Count the (non-) synonymous differences
  - Count the (non-) synonymous sites
  - Account for the possible evolutionary history
- Example of the algorithm
  - **Number of (non-) synonymous differences**
    - N_d: number of non synonymous differences
    - S_d: number of synonymous differences

$$S_d$$

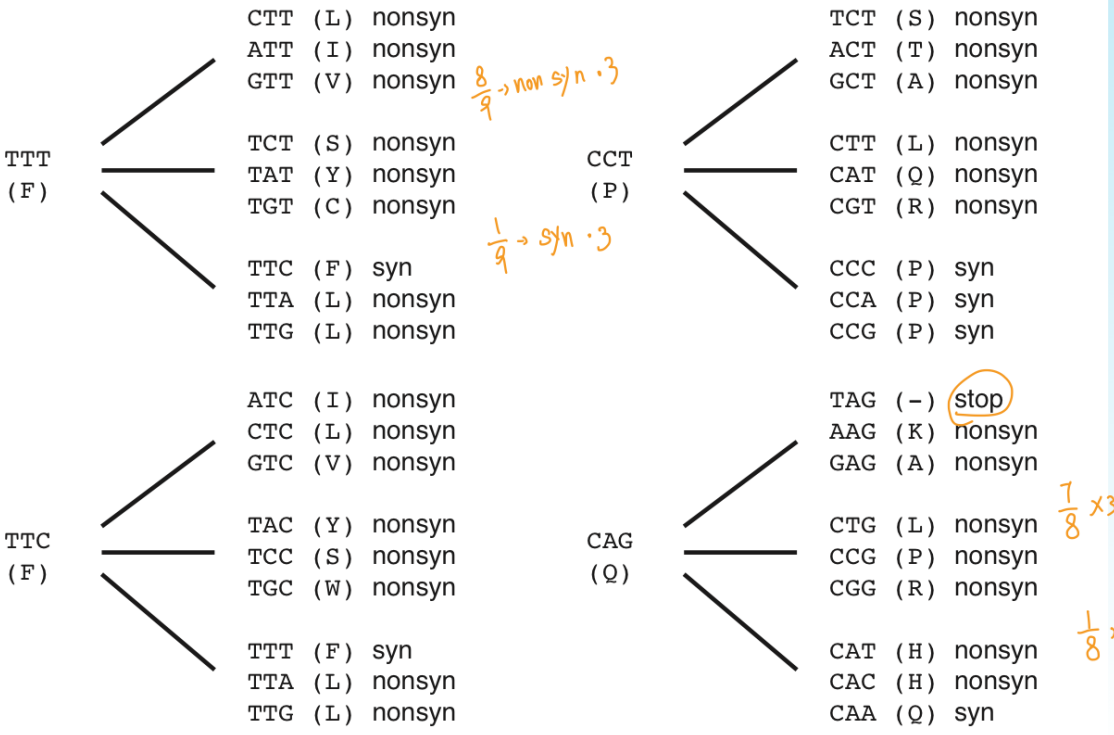
	Codon 1	Codon 2	Codon 3	Codon 4
Sequence 1	TTT	CCT	CCT	CCT
Sequence 2	TTC	CAG	CCT	CCT
N_d	0	1.5	0	0
S_d	1	0.5	0	0

Two possible ways from CCT to CAG	S_d	N_d
CCT(P) -> CAT(H) -> CAG(Q)	0	2
CCT(P) -> CCG(P) -> CAG(Q)	1	1
Average	(0+1) / 2 = 0.5	(1 + 2) /2 = 1.5

  - **Number of (non-) synonymous sites**
    - N ; number of non synonymous sites
    - S: number of synonymous sites
    - averaging over all potential one point mutations
    - The sum of (non-)synonymous mutations per codon must sum up to 3

	cdn1	cdn2	cdn3	cdn4	
seq1	TTT	CCT	CCT	CCT	
seq2	TTC	CAG	CCT	CCT	
nonsyn					
seq1		+	+	+	=
seq2		+	+	+	=
average					N =
syn					
seq1		+	+	+	=
seq2		+	+	+	=
average					S =

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



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

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

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

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

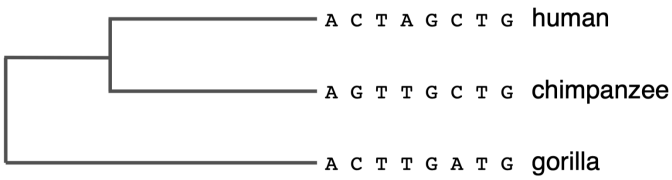
  - Using N_d, N, S_d, S calculated previously , we have $\frac{d_N}{d_S} = 0.23$

- Interpretation of $d_N/d_S - ratio$
  - d_N/d_S < 1: non synonymous mutations occur less frequently than synonymous mutations (purifying selection)
  - d_N/d_S > 1: non synonymous mutations occur more frequently than synonymous mutations (positive selection)

- Q & A
  - What are the differences between the JC69 and the TN93 models
    - parameters: JC69 only has one model parameter λ; TN93 counts transitions between $T \leftrightarrow C, A \leftrightarrow G$ and transversions happen rate different, resulting in 3 + 3* parameters, the 3* here indicates the nucleotide equilibrium frequency

  - Biological realism: JC69 model does not differentiate transition and transversions, while TN93 rates transitions between $T \leftrightarrow C, A \leftrightarrow G$ and transversions differently.
  - Time reversible: Both models are time reversible.
- Which of the two models would you chose if you were to perform a phylogenetic analysis based on sequence distances
  - GTR: most general one
  - TH93: more general than nJC69 but not too complex
  - General rule: few data -> simple model; large data -> complex model
- In lecture 3, we tried to naively reconstruct a phylogeny based on three sequences. The pairwise Hamming distances were all the same. Would you expect that any of the presented nucleotide sequence models would result in different trees?

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



  - Given the three sequences above
    - human ACTAGCTG
    - chimpanzee: AGTTGCTG
    - gorilla: ACTTGATG
  - We can calculate the distances between the sequences using JC69 and $JC69 + \Gamma$.
  - For JC69 model, we calculate the distances using

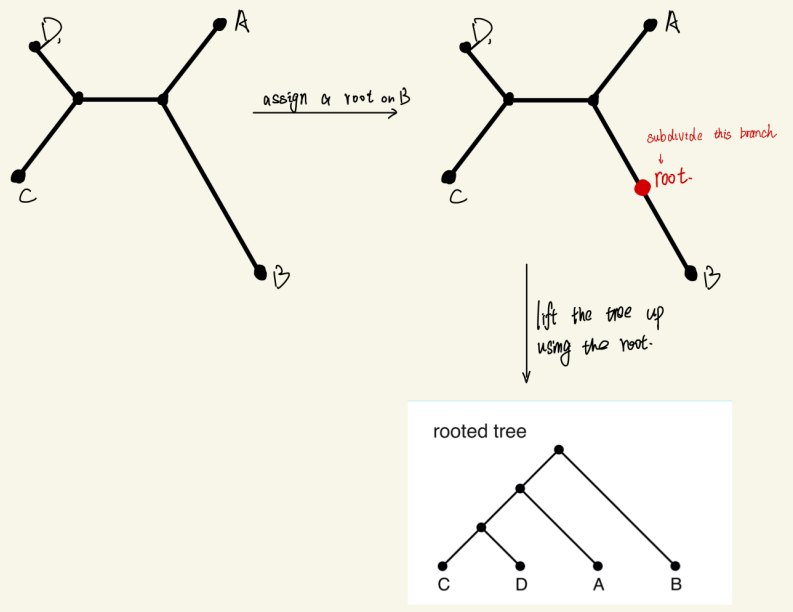
$$\hat{d} = -\frac{3}{4}\log(1 - \frac{4x}{3n})$$

    where x is the number of different nucleotides between two sequences.
  - For example: human and chimpanzee differ by x = 2, then the distance between the human and chimpanzee is $\hat{d} = -\frac{3}{4}\log_e^{(}1 - \frac{4*2}{3*8}) = 0.3$

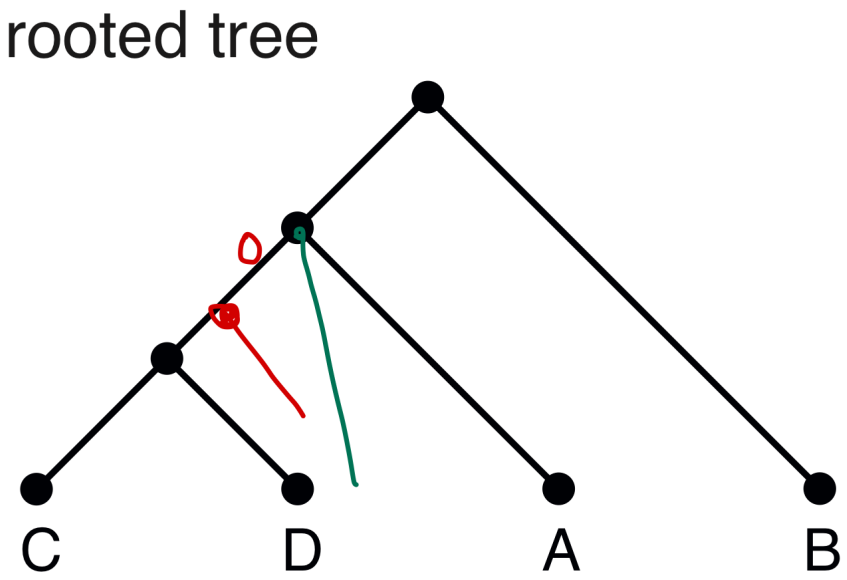
- Since all three pairs have the same number of differences (x = 2). So if we use the JC69 model, we would have the same distances of all three pairs. So JC69 and $JC69 + \Gamma$ model would not be of much help.
- Also, since all the differences here are all transversions, we will not be able to get different distances using more complex models like TN96 and K80.
- This is a illustration of problems we might run into if we calculate the distances only based on sequences.

Lecture 5: Phylogenetics

- Introduction to phylogenetics
  - Trees
    - 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 attached
    - **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
    - Linkage of unrooted and rooted trees
      - Unrooted trees can be rooted with an outgrip (here B). which means that the branch ending in B is subdivided by the root node. B is chosen by the user as a very distantly related organism to the remaining organisms in the tree.



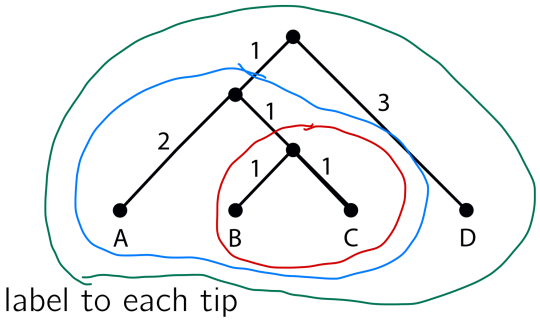
- Phylogenetic terms:
  - **pendant branch**: a branch attached to a tip
  - **cherry:** a pair of tips which are only separated by one internal node
  - **caterpillar tree:** a rooted tree with only one cherry
  - **monophyletic group / clade** : **all** descendants of a common ancestor
  - **ultrametric tree:** the sum of branch lengths from any tip to the root is the same
  - **polygamy:** the definition of a phylogenetic tree is extended such that internal nodes may have more than 3 branches attached. Such a node is a polygamy. It can be represented as a classic phylogenetic tree with branch lengths of 0.



    - The green line indicates the polygamy and the red dot means it can be represented as a classic phylogenetic tree with branch lengths to 0.

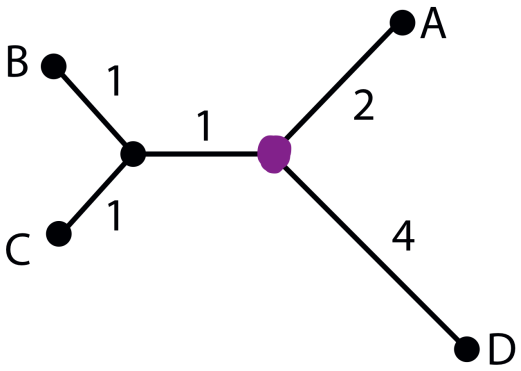
- Newick format
  - recursively: choose two tips (e.g. C and D) forming a cherry and replace them by the new tip $(C : t_C, D : t_D)$, where t_x is the length of the branch ancestral to node X; the length of the branch ancestral to new tip is the branch length ancestral to the cherry
  - E.g

((A:2, (B:1,C:1) : 1) : 1, D:3)



label to each tip

- Newick format for unrooted trees
  - Choose arbitrarily an internal node
  - proceed as in the rooted tree towards this arbitrary internal node
  - connect the three last tips X, Y, Z to $(X : t_X, Y : t_Y, Z : t_Z)$
  - E.g.



-((B:1,C:1):1, A:2, D:4)

-A tree can have different but equivalent Newick representations

- Examples of phylogenies
  - phylogeny of species
    - species phylogeny of simians
      - tips: simian species consisting of apes, old world monkeys, and new world monkeys
      - branching events: speciation events
      - branch lengths: time between speciation events
    - phylogenies contain information about species relationships
  - Phylogeny of pathogens
    - pathogen phylogeny of HIV epidemic

        - tips: different infected hosts
        - branching events: transmission events
        - branch length: time between transmission events
    - pathogen phylogeny is an approximation of part of the transmission chain, and thus contains information about transmission dynamics.
        - origin of epidemics
        - transmission group interaction, criminal cases etc.
- Phylogenetic inferences: How to construct phylogenetic trees
  - Overview of phylogenetic reconstruction methods
    - Similar individuals are clustered together. Similarity may be defined in different ways
    - **phenetic:**
      - based on overall similarity
      - pairwise distance-based
      - methods: UPGMA, least squares algorithm
    - **cladistic:**
      - based on shared characteristics
      - character-based
      - method: parsimony
    - **mechanistic:**
      - based on evolutionary model
      - character-based
      - methods: maximum-likelihood, Bayesian inference
  - Phenetic Approach
    - Distance-based methods
      - basic idea:
        - define how to measure distance between sequences
        - calculate distance between all pairs of sequences
        - find a tree where the distances, i.e. the branch lengths, between the pairs of tips “most closely” follow the sequence distance matrix.
      - two strategies:
        - **algorithmic approach:** sequences separated by the smallest distance are clustered iteratively in a tree

- **optimality approach:** minimise the difference of the sequence distance matrix to the inferred tree distance matrix

- only distances between pairs of sequences are used but not any higher order correlations between sequences.
- The distance matrix:

sequence 1 (s_1): TCACACCT
sequence 2 (s_2): ACAGACTT
sequence 3 (s_3): AAAGACTT
sequence 4 (s_4): ACACACCC

Hamming distance / len(S_i)

Hamming-distance

H	s_1	s_2	s_3	s_4
s_1	-	3	4	2
s_2		-	1	3
s_3			-	4
s_4				-

JC69-distance

$$\hat{d} = -\frac{3}{4}\log(1 - \frac{4}{3}\hat{p})$$

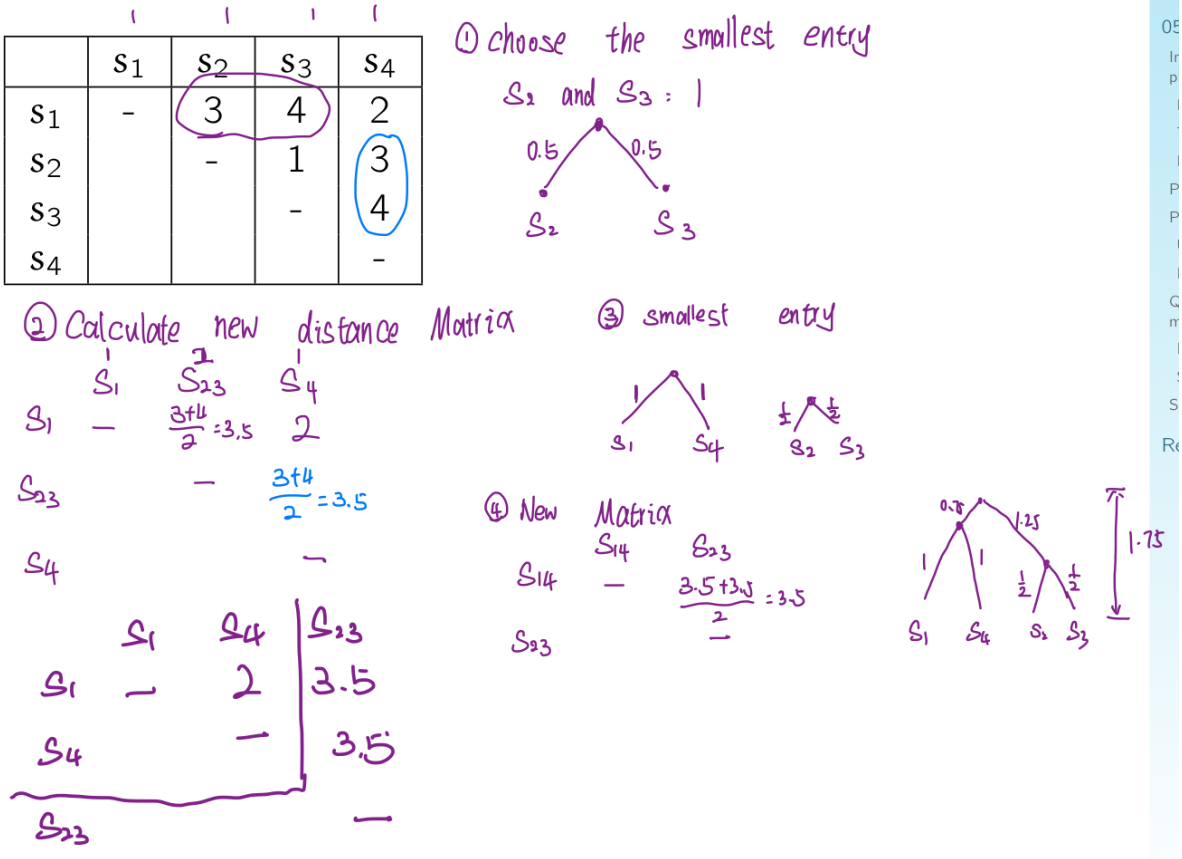
JC	s_1	s_2	s_3	s_4
s_1	-	0.52	0.82	0.30
s_2		-	0.14	0.52
s_3			-	0.82
s_4				-

- **Algorithmic approach: UPGMA**
  - UPGMA: Unweighted Pair-Group Method using Arithmetic Averages
  - Input : distance matrix
  - Output : ultrametric phylogenetic tree
  - Property of this algorithm
    - all sequences must come from the same time point
    - the algorithm assumes evolution according to a **strict molecular clock**
      - the rate of DNA/RNA/protein sequence evolution is constant over time
    - the output is an **ultrametric tree:**
      - rooted tree, the total branch lengths from all tips to the root are equal
  - Computational Steps
    - Initialize the size of each node s_i as $n_i = 1$
    - While the distance matrix is not empty, iterate:

1. Choose nodes s_i and s_j such that $d(s_i, s_j)$ is the smallest entry in the distance matrix (in case of several minima choose one uniformly at random).
2. Coalesce s_i and s_j to node $s_{i,j}$, with size $n_{i,j} = n_i + n_j$. The branch length between $s_i, s_{i,j}$ and between $s_j, s_{i,j}$ is chosen such that all tips descending from $s_{i,j}$ have the same distance $\frac{d(s_i, s_j)}{2}$ to $s_{i,j}$
3. If the distance matrix includes more than 2 nodes, include $s_{i,j}$ into the distance matrix with $d(s_m, s_{i,j}) = \frac{n_i d(s_i, s_m) + n_j d(s_j, s_m)}{n_i + n_j}$ where s_m is a node in the distance matrix
4. Delete nodes s_i and s_j from the distance matrix

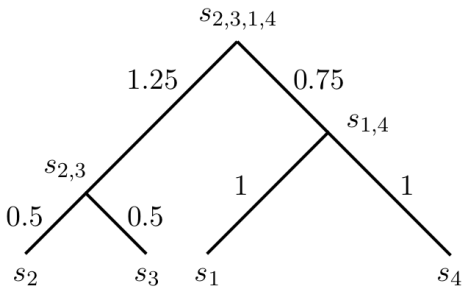
- Output: Ultrametric phylogenetic tree

- Example:



- Problem with UPGMA:

we obtain the tree:



original distance matrix:

	s_1	s_2	s_3	s_4
s_1	-	3	4	2
s_2		-	1	3
s_3			-	4
s_4				-

Any strange observation?

new distance matrix:

	s_1	s_2	s_3	s_4
s_1	-	3.5	3.5	2
s_2		-	1	3.5
s_3			-	3.5
s_4				-

New matrix differs!

  - This problem indicates that not all distance matrix can have a UPGMA tree that represents exactly the same distance in the distance matrix, instead, we can only get a “tree distance”.
  - This means that we just can not fit the same flexibility as the original distance matrix into an UPGMA tree.
  - If there is a tree that induces the sequence distance matrix, this tree will be the UPGMA tree of the sequence distance matrix.

- **Optimality Approach: Least squares methods**
  - Least squares methods use an optimality criterion instead of an algorithmic approach (UPGMA):
  - squared difference between the sequence distance matrix and the tree distance matrix is,

$$S := \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{imj}(D_{i,j} - d_{i,j})^2$$

  - where D is the sequence distances matrix, d the tree distance matrix for a proposed tree, and w weights ($w_{i,j}$ may be 1 or e.g inverse proportional to $D_{i,j}$)
  - Algorithm
    - Input : distance matrix

- Repeat until all tree topologies were proposed
    1. propose an unrooted tree topology (without branch lengths)
    2. minimise S, i.e. optimise the branch lengths
- Output: tree with the smallest S

- Quality assessment of phylogeny reconstruction methods
  - Runtime
    - Runtime of UPGMA
      - To get an estimate of the number of computation steps in UPGMA on n tips, we need to answer:
        - How many times do we prune nodes?
          - n
        - How many calculations do we perform per pruning?
          - n^2
      - UPGMA has runtime $O(n^3)$ for n sequences
    - Runtime of least squares methods
      - For each possible tree we need to optimise

$$S := \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{imj}(D_{i,j} - d_{i,j})^2$$

      - Thus we need to visit each tree in the space of trees
      - **How many trees on n tips exist?**
        - count number of branches in an unrooted tree with n tips, b_n
        - use b_n to calculate the number of unrooted trees with n tips, τ_n
        - use τ_n to calculate the number of rooted trees with n tips, τ_n^r
      - **Step 1: Counting branches**
        - A tree with n = 2 tips: $b_2 = 1$ branch
        - Consider an unrooted tree on n tips, every time we add an additional tip, we add another 2 branches : $b_{n+1} = b_n + 2$
        - In general : $b_n = b_2 + 2 * (n - 2) = 2n - 3$ branches
        - Proof by Induction : 数学归纳法

- **Step 2: Counting unrooted trees**
  - We now know that a tree with n tips has $b_n = 2n - 3$ branches
  - With this we can derive how many unrooted trees with n tips exist:
    - Unrooted trees on 2 tips : $\tau_2 = 1$
    - Given a tree on n tips, in how many ways can we add the (n + 1)th tip?
      - Every tree has b_n branches we can add on
      - There are τ_n trees in total
      - $\tau_{n+1} = \tau_n * b_n$
    - For $n \geq 3$ holds:

$$\begin{aligned}\tau_n &= \tau_{n-1} * b_{n-1} \\ &= \tau_{n-2} * b_{n-2} * b_{n-1} \\ &= \dots \\ &= \tau_2 * b_2 * b_3 * \dots * b_{n-1} \\ &= 1 * 1 * 3 * 5 * 7 * \dots (2n-5) \\ &= (2n-5)!!\end{aligned}$$

- **Counting rooted trees**
  - We obtain a rooted tree on n tips by choosing an unrooted tree on n tips and pick one branch which is subdivided by a root node.

$$\tau_n^r = \tau_n * b_n = (2n-5)!!(2n-3) = (2n-3)!!$$

- The least square decision problem is an **NP-complete problem! T**hus there is no polynomial-time algorithm unless $P = NP$. As a consequence, we essentially have to check all trees with n tips

- NP completeness of a problem
  - P = polynomial time, i.e. the runtime until a solution for a problem with input size n is found is n^k with k some fixed number (input-independent; e.g. n^3 for UPGMA)
  - NP = nondeterministic polynomial time
    - Consider a decision problem X (e.g. Is there a tree with a least squares difference of less than x?)
    - A decision problem is in NP if it can be verified in polynomial time. (e.g it is easy to determine the least squares difference for a given tree).
    - A decision problem in NP is NP complete if the travelling salesman problem can be solved using al algorithm to solve the decision problem X together with potentially a polynomial time transformation algorithm.

- Statistical consistency
  - A phylogenetic reconstruction method is statistically consistent if the true tree is returned for an infinite amount of data (i.e. infinitely long sequences)
  - Formally:
    - A phylogenetic reconstruction method is **statistically consistent** if for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(||\hat{T} - T|| < \epsilon) = 1$$

    - Where n is the sequence length, T the true tree, and $\hat{T}$ the inferred tree.
  - Let sequences evolve on a fixed tree with some model M, the distances in the tree in units of substitution . Then the maximum likelihood distance matrix based on the simulated sequences approaches the distances in the tree with increasing sequence length, since maximum likelihood estimators are statistically consistent.
  - UPGMA is consistent
  - Least squares method is consistent.
- Problem of phenetic approaches:
  - Disregard information beyond pairwise distances
  - Large distances come with large variances, which are typically ignored.

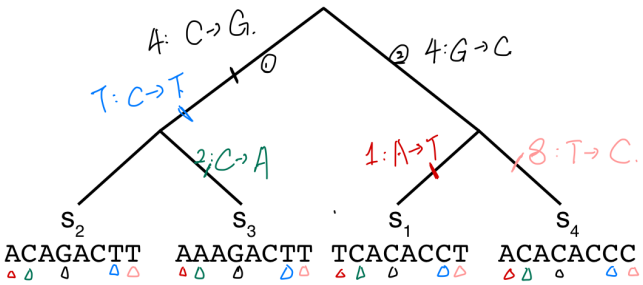
- Q & A:
  - What is the minimal number of cherries in a phylogenetic tree of 99 tips? What is the maximum number?
    - 1; $\lfloor \frac{99}{2} \rfloor = 49$
  - In how many ways can you write the Newick string for a rooted tree with species A,B,C? In how many ways can you write the Newick string for a rooted tree with n species
    - For a rooted tree with only 2 species, we can have 2 Newick strings which are either $(A : d_A, B : d_B)$ or $(B : d_B, A : d_A)$
    - For a rooted tree with k species, imagine we have already known the number of Newick string for this tree, then we consider k + 1 species,
      - There are two ways we can add this species to the Newick string which are $(k+1 : d_{k+1}, Newick_k)$ or $(Newick_k, k+1 : d_{k+1})$
      - This means when we add a new species to the tree, we double the number of newick string for this tree

  - So, the number of Newick string for a rooted tree with n species is 2^{n-1}

- Consider the least square method. Why would we use weights $w_{i,j}$ which are not equal to 1?
  - Since the distances are estimated and the larger the estimated distance is, the larger of the variance of that distance is. We can embed these kind of information by setting the weight to something inverse proportional to $D_{i,j}$.
  - It is a way to express the uncertainty of an estimated distance. The more uncertain we are about one distance, the less of the weight that distance may have.

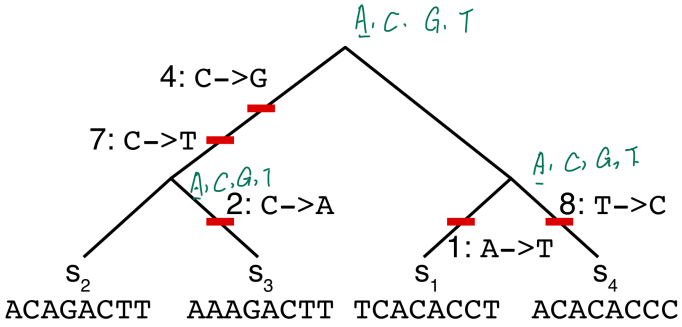
Lecture 6: Phylogenetics II

- Cladistic tree inference
  - Phenetic vs cladistic approach
    - phenetic approach: overall similar sequences cluster
    - cladistic approach: sequences with many shared characters cluster
      - **parsimony: find tree with minimal number of mutations**
  - Parsimony method:
    - 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
    - Example: For the given four sequences below, the parsimony tree would have at least 5 mutations
      - sequence 1: TCACACCT
      - sequence 2: ACAGACTT
      - sequence 3: AAAGACTT
      - sequence 4: ACACACCC



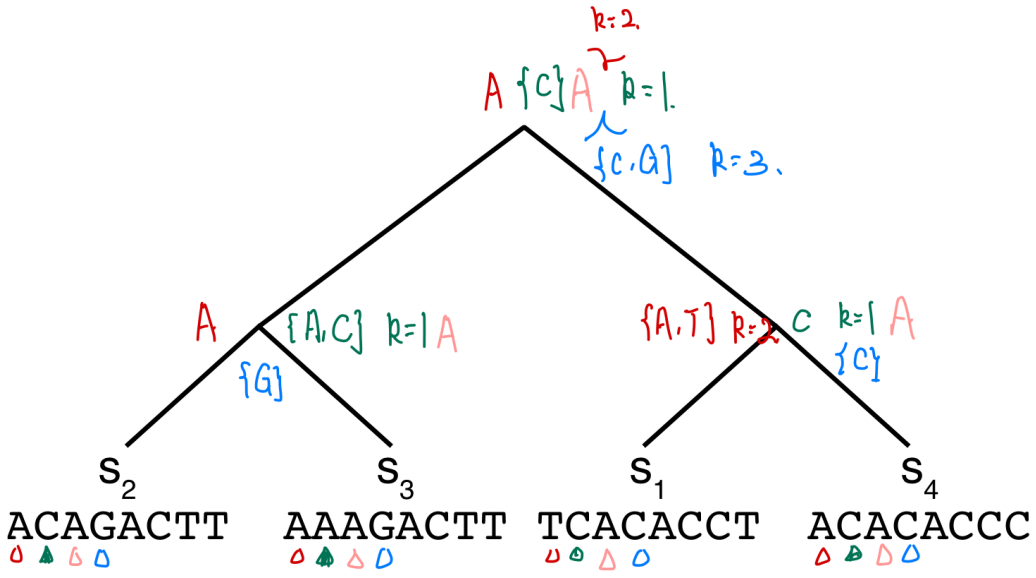
3个A，1个T，所以假设S1上从A突变到T.
3个C，1个A，所以假设 S3上从C突变到A
2个C. 2个A：两种解释：
① branch ①上从C到G
② branch ②上从G到C

- Parsimony score of UPGMA tree
  - **Rigorous:** Label internal nodes with each possible ancestral sequence and then determine the number of mutations required for each assignment. Minimal number of mutations requires is the parsimony score
    - For example: for position 1 of the above 4 sequences, there are 3 internal nodes which means there are $4 \times 4 \times 4 = 64$ different labels to try (from AAA to TTT). We try every single one of them and extract the minimum mutations required to explain the sequences at the tips of the tree.

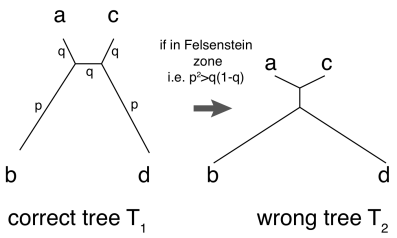


    - How many possible internal sequence assignments exist?
      - For every tree with n tips, there are n - 1 internal nodes, which means we would have to try out 4^{n-1} combinations and for each combination. We denote the size of the alignment to be m, then we have the overall time complexity of this algorithm $O(4^{n-1} \times m)$. -> EXPENSIVE
- Parsimony score of different tree structures
  - Rooted trees obtained from the same unrooted tree have the same parsimony score!
- Parsimony method
  - Input: Sequence alignment of n sequences, with sequence length m
  - Iterate:
    - Consider each unrooted tree (overall there are $(2n-5)!!$ unrooted trees). This step can not be improved much unless $P = NP$
    - Calculate parsimony score for the considered unrooted tree (requires to consider $4^{n-1} \times m$ internal sequence assignments). → This step CAN be improved considerably by the Fitch algorithm.
  - Output: Unrooted tree with the lowest parsimony score

- Fitch algorithm to quickly determine parsimony score
  - Input: Unrooted phylogenetic tree and an alignment of n sequences of length m, corresponding to the n tips of the tree.
  - Computational steps:
    - root the tree at an arbitrary edge
    - $k \leftarrow 0$
    - while the root has no sequence assigned iterate
      - choose a mode in the tree where all descending nodes have sequences assigned
      - assign a sequence to the chosen node:
        - For $i = 1,2,...,m$, do the following: Let C_l and C_r be the sets of nucleotides being assigned to the two direct descendants of the chosen node for site i. If $C_l \cap C_r \neq \emptyset$, we assign $C_l \cap C_r$ to nucleotide i of the chosen node. If $C_l \cap C_r = \emptyset$, we assign $C_l \cup C_r$ to nucleotide i of the chosen node and set $k \leftarrow k + 1$
  - Output: Parsimony score k of the tree. I.e. minimal number of mutations required to explain the sequences at the tip.
  - Example:



- Time complexity of Fitch algorithm
  - We have to visit each internal node of the rooted tree and for a rooted tree on n tips, we have $n-1$ internal nodes.
  - Total number of steps: $(n-1) \times m$ with m sequence length. Thus the Fitch algorithm improved the runtime from $O(4^{n-1} \times m)$ to $(n-1) \times m$ by using dynamic programming
  - Parsimony tree is found by calculating parsimony score for each unrooted tree
    - The parsimony decision problem is NP-complete
- Statistical inconsistency of parsimony method
  - Parsimony method does not consider back-substitutions.
    - A branch with apparently no substitution may have had two substitutions e.g. from $A \rightarrow G \rightarrow A$, but parsimony assumes no substitution at all in such a case.



• If the chance of a change is observed on both the long branches which is p^2 is bigger than actually observing a change on the small separating branch, then we would construct b and d being clustered

-

-

-

- Summary cladistic tree inference
  - **statistically inconsistent**: no back substitutions or parallel substitutions are considered, which leads to long-branch attraction
  - **Slow**: NP-complete decision problem - the whole tree space has to be visited

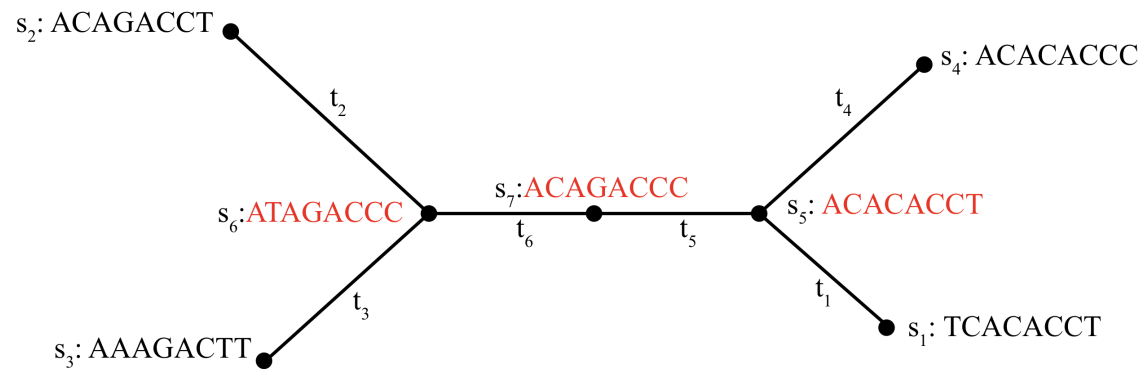
- UPGMA and Parsimony tree may be different from each other.

- Mechanistic tree inference
  - Maximum likelihood tree inference
    - Input: Sequence alignment
    - Output: Tree which maximises the probability of the sequences given the tree & the sequence evolution parameters.
      - Requires an evolution model (JC69,HKY,GTR etc)
      - Parameters of the evolution model can be co-estimated with the tree.
    - The Maximum likelihood framework
      - Likelihood of a parameter is the probability of the observed data given the parameter.
      - Maximum likelihood estimate is the parameter maximising the likelihood function given the data
        - The fraction of dice throws where we observe a six, assuming a binomial distribution, is the maximum likelihood estimate for the probability p of observing a six (p is the parameter).
        - The average of measurements, assuming a normal distribution, is the maximum likelihood estimate for the mean m of the normal distribution (m is the parameter)
      - What is the maximum likelihood phylogeny given the sequence data?
    - Maximum likelihood in phylogenetics
      - Mechanistic model for evolution of the data (sequences):
        - Each unrooted tree $\mathcal{T}$ 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
      - $L(\mathcal{T}, Q; D) := P(D \mid \mathcal{T}, Q)$ is called the likelihood function of the parameters $\mathcal{T}$,Q for the given sequence data
      - Inference: Determine the $\mathcal{J}, Q$ which best explain the alignment

$$\max_{\mathcal{T},Q} L(\mathcal{T}, Q; D)$$

      - We determine the best tree by evaluating the likelihood for “many” different trees.

- Likelihood calculation for a given tree



- Sites in the alignment evolve independent from each other, we can consider each site separately. Let the alignment consist of m sites, then,

$$P(s_1, \ldots s_n \,|\, \mathcal{T}, Q) = \prod_{j=1}^{m} P(s_{1,j}, \ldots, s_{n,j} \,|\, \mathcal{T}, Q)$$

(with $s_{k,j}$ being site j of sequence s_k)

- Typically, the substitution process is time-reversible. Thus for a proposed unrooted tree, we subdivide an arbitrary edge to obtain a root, with (unknown) sequence s_{2n-1}. For n sequences, the rooted tree has $n-1$ internal nodes with (unknown) sequences $s_{n+1}, \ldots, s_{2n-1}$.

- The probability of the nucleotides at the tips at site j is the sum over the probabilities of nucleotide states at the internal nodes and tips.

$$P(s_{1,j}, \ldots, s_{n,j} \,|\, \mathcal{T}, Q) = \sum_{s_{n+1,j} \in \{A,C,G,T\}} \sum_{s_{n+2,j} \in \{A,C,G,T\}} \cdots \sum_{s_{2n-1,j} \in \{A,C,G,T\}} P(s_{1,j}, \ldots, s_{2n-1,j} \,|\, \mathcal{T}, Q)$$

- Finally, $P(s_{1,j}, \ldots, s_{2n-1,j} \,|\, \mathcal{T}, Q)$ can be evaluated by calculating for each branch l (with starting sequences s_{l_1}, ending sequence s_{l_2}, and branch length t_l) the transition probability from the ancestral nucleotide to the descendant nucleotide.

- The root nucleotides (here s_7) are weighted by their equilibrium probabilities π, so overall,

$$P(s_{1,j}, \ldots, s_{2n-1,j} \,|\, \mathcal{T}, Q) =$$
$$\pi(s_{2n-1,j}) \prod_{l=1}^{2n-2} P_{s_{l_1,j}, s_{l_2,j}}(t_l)$$

(A rooted tree on n tips has $2n-2$ branches)

- Example:
  - For site j = 2, we have,

$$P(s_{1,2}, \ldots, s_{2n-1,2} \mid \mathcal{T}, Q) = \pi(s_{2n-1,2}) \prod_{l=1}^{2n-2} P_{s_{l_1,2}, s_{l_2,2}}(t_l)$$

$$= \pi_c P_{C,C}(t_5) P_{C,C}(t_4) P_{C,C}(t_1) P_{C,T}(t_6) P_{T,C}(t_2) P_{T,A}(t_3)$$

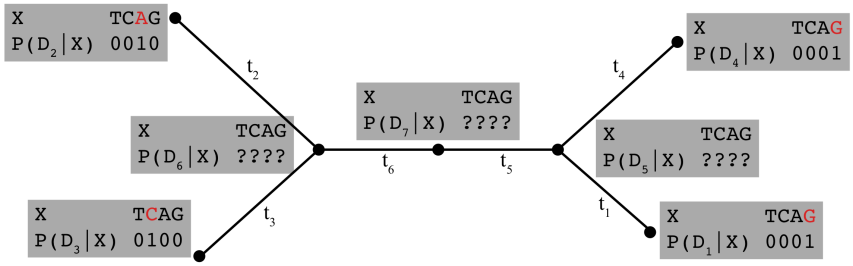
  - We have $4 \times 4 \times 4 = 64$ possibilities for nucleotides at internal nodes of site j. Thus the sum in our tree with three internal nodes consists of 64 summands.
- Taken together
  - The probability to observe the data given a specific tree, $\mathcal{T}$ and a substitution matrix, Q, is

$$P(s_1, \ldots s_n \mid \mathcal{T}, Q) = \prod_{j=1}^{m} [\sum_{s_{n+1}, j \in \{T,C,A,G\}} \cdots \sum_{s_{2n-1}, j \in \{T,C,A,G\}} (\pi(s_{2n-1,j}) \prod_{l=1}^{2n-2} P_{s_{l_1,j}, s_{l_2,j}}(t_l))]$$

Where $P_{s_{l_1,j}, s_{l_2,j}}(t_l)$ is the transition probability from the nucleotide at the start of branch l, $s_{l_1,j}$, to the nucleotide at the end of branch l, $s_{l_2,j}$, at site j, with branch length t_l. As derived in lecture 3, the substitution rate matrix Q defines the transition probabilities.

- Run 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 ($O(2n-2)$)
    - runtime of likelihood calculation is $O(m4^n n)$ -> Very slow
    - Felsenstein's pruning algorithm speeds up this calculation using a clever way to store intermediate steps that are reused in the summation.

- Improving the runtime of ML tree inference: Felsenstein's pruning algorithm



- Consider site j (and later multiply all sites as before). Let the probability of the tip nucleotides descending the node k be $P(D_k|X)$, given the nucleotide at node k is X. The tree is traversed as in the Fitch algorithm for $X \in [A, C, G, T]$ determining $P(D_k|X)$ for all nodes.
- At a tip k, $P(D_k|X) = 1$ iff X is the observed nucleotide; $P(D_k|X) = 0$ otherwise $(X \in \{A, C, G, T\})$
- "Cherries" are pruned recursively towards the root, let k be a node with the descendants l, m :

$$P(D_k|X) = \left(\sum_{Y \in \{A,C,G,T\}} P_{X,Y}(t_l) P(D_l|Y) \right) \times \left(\sum_{Z \in \{A,C,G,T\}} P_{X,Z}(t_m) P(D_m|Z) \right)$$

Where $P_{X,Y}(t_l)$ is the probability of the nucleotide change from X to Y in time t_l and can be calculated using rate matrix Q, $P(D_l|Y)$ is stored in the descendants l.

- Thus for the root r, we calculated $P(D_r|X)$ where $X \in \{A, C, G, T\}$
- Finally, the probability of the sequences at site j is

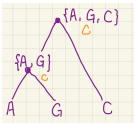
$$P(s_{1,j}, \ldots, s_{n,j}|\mathcal{T}, Q) = \sum_{X \in \{A,C,G,T\}} P(D_r|X)\pi_X$$

- Runtime of Felsenstein's pruning algorithm
  - each recursion step is a summation over two times four states, we have $O(n)$ nodes and thus the recursion has runtime $O(n)$
  - the recursion has to be performed for each of the m sites, $O(m)$
  - thus, in total, the runtime is $O(nm)$

- However, the problem of finding a tree and branch lengths with likelihood value $\leq L$ is NP-complete.

- Q&A:

  - Consider the Fitch algorithm. Do you obtain all most parsimonious ancestral sequences when choosing the different nucleotides in the curly brackets?



-No, for the parsimony trees given below, if we put Cs instead, the parsimony score is still 2

  - Does the maximum likelihood tree reconstruction methods return estimates for the internal sequences? Give a reason for your answer.

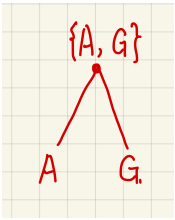
    - No, the probability that we focus on in the maximum likelihood tree reconstruction is the the probability of the tip nucleotide descending the node k, given the nucleotide at k is X, denoted as $P(D \mid X)$.

    - In this conditional probability, we assume that $X \in \{A, C, G, T\}$ is known, however, if we want to estimate the probability of the node at k is X, we would need to be estimating $P(X \mid D)$

    - This is can not be done using the Bayesian Rule since the tree structure are different.

  - Does the Fitch algorithm return the parsimony score for any phylogenetic tree and any sequence alignment? Or are there situations when the Fitch algorithm does not return the smallest number of mutations required?

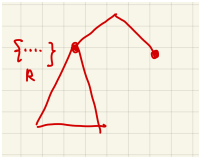
    - Yes, it returns the parsimony score for any phylogenetic tree and for any sequence alignment. Proof by induction



- For the simplest tree — a cherry, it follows that

  - If the two leaf nodes are different, we have $k = 1$, which is correct

  - If the two leaf nodes are the same, we have $k = 0$, which is correct.



-Assume we have proven that the algorithm works on the red rectangle tree, we now add a new node to it.

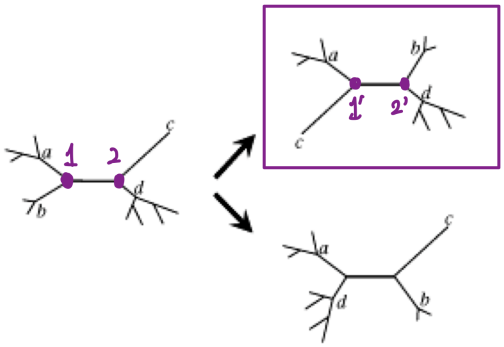
  - If the character in the new node is contained in the $\{ \ldots \}$, we know that the algorithm does not update k, which is correct

  - Else if the character in the new node is not contained in the $\{ \ldots \}$, we would update k to $k + 1$, which is then correct.

    - Then, we finish the proof by induction.

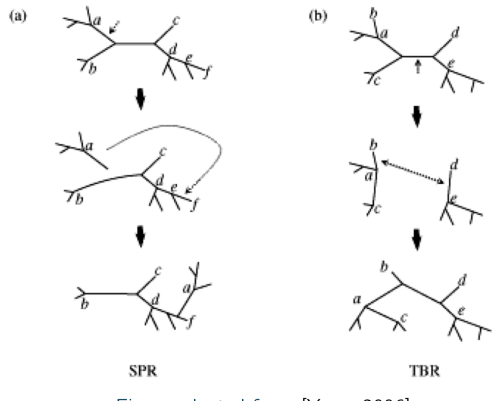
Lecture 7: Maximum likelihood method & testing

- Searching tree space
  - How to search tree space for maximum likelihood tree?
    - We need to propose different unrooted trees: NNI, SPR, and TBR moves (next slides)
    - We need to propose different branch lengths: multiply each branch length by some factor
    - We can use “hill-climbing” strategies to find the optimum
  - NNI move



-The nearest-neighbour interchange (NNI) algorithm. Each internal branch in the tree connects four subtrees or nearest neighbours (alb, cod). Interchanging a subtree on one side of the branch with another on the other side constitutes an NNI. Two such rearrangements are possible for each internal branch

  - SPR and TBR move



-Branch swapping by subtree pruning and regrafting(SPR). A subtree (for example, the one represented by node a) is pruned, and then reattached to a different location on the tree.

-Branch swapping by tree bisection and reconnection (TBR). The tree is broken into two subtrees by cutting an internal branch. Two branches, one from each subtree, are then chosen and rejoined to form a new tree.

- Model testing
  - Which evolutionary models are appropriate for our data?
    - likelihood ratio test
    - AIC

- How confident can we be in the inferred parameters, i.e. how confident can we be in the phylogeny and the substitution rates?
  - confidence assessed via likelihood ratios
  - confidence assessed via bootstrapping
- All methods above, with the exception of bootstrapping, require a maximum likelihood method
- Reminder from Lecture 03
  - We used the hypergeometric distribution as a null model H_0 and asked if the p-value lies below the significance level α. In that case, the null model H_0 was rejected
  - We did not have an alternative model. However, we could approximately evaluate $P(X \mid \theta)$ for all possible data X and fixed parameter θ under the null model.
  - Now we want to ask if we reject H_0 in favour of model H_1. This is advantageous in the phylogenetic setting, as it is not straightforward to evaluate $P(X \mid \theta)$ where X is now an alignment) for all possible alignments
- Likelihood ratios
  - Assume the data evolve under a model H_0
  - Assume a model H_1 within which H_0 is nested
    - H_0 is a special case of H_1
  - For the given data, let the maximum likelihood parameter estimate under model H_0 be $\hat{\theta}_0$ and under H_1 be $\hat{\theta}_1$
  - Now under some “mild” conditions (in particular large amount of data),

    $$2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0)) \sim \chi^2_{df}$$

    Where df is the degree of freedom
  - The degree of freedom in the χ^2_{df} distribution is the difference between the number of parameters in the general and in the nested model. Be careful though when the special model parameter is at the parameter boundary (e.g. 0 or ∞), then the degree of freedom loss is typically only 0.5
  - $2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0)) = 2 \log \left(\frac{L(\hat{\theta}_1)}{L(\hat{\theta}_0)} \right)$ explains the name “likelihood ratio”

- Likelihood ratio test
  - Consider two models : H_1 general model parameterised in θ_1, H_0 nested model parameterised in θ_0
  - derive the likelihood function for both models and the maximum likelihood estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ for a given data set
  - calculate $2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0))$
  - reject the null model if $2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0))$ is in the α tail of the χ^2_{df}
    - most often $\alpha = 0.05$
    - α is called the significance level
    - if the null model was the true model, we would falsely reject it in a proportion α of tests
    - if the null model was the false model, then we expect the null model to have a much lower likelihood than the tree model, and thus we would accept the null model only in a very low proportion of tests.
- Connections to Lecture 3
  - In lecture 3, we considered $P(X \mid \theta)$ under the null model, and asked if θ is an extreme outcome. Thus, for the die experiment, we checked if the number of times a 6 was thrown is in the 5% tail of the binomial distribution
  - Here we compare nested models, i.e. when a simple model (H_0) is obtained by restricting parameters in the general model (H_1). For our die, the simple model had no free parameter, but in general $\hat{\theta}_0$ may be the maximum likelihood estimate of a simple model with free parameters
  - While in lecture 3 we assessed the overall fit of a. model, here we only assess the fit of H_0 relative to H_1. In other words, even though H_1 might be a very bad model, we may reject H_0 in favour of H_1 since H_0 is even much worse.
- Model testing erros

	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)
    - The type I error is the significance level, and thus the accuracy is controlled by setting α
  - Power = 1 - (Type II error)

  - Power can generally only be assessed via simulating under the general model H_1 and assessing the number of times that H_0 is accepted.

- Example : Die rolling
  - simulation of 10 000 die rolling experiments with different θ_1 s where we report the fraction of 6 among 1000 die rolling
  - H_0: the probability to obtain a 6 is $\theta_0 = 1/6$
  - We test with significance level $\alpha = 0.05$
  - simulations with different $\theta_1 s$, tested under the same H_0:
    - $\theta_1 = 1/6$
      - H_0 is rejected in 5.1% of the 10000 experiments. This simply highlights again that we chose an accuracy of 0.95
    - $\theta_1 = 1/5$
      - H_0 is rejected in 77.66% of the 10000 experiments. Thus the power is estimated to be 0.78
    - $\theta_1 = 1/2$
      - H_0 is rejected in all of the 10000 experiments. Thus the power is estimated to be 1.
  - The power increases with an increasing difference of the true model and the null model H_0

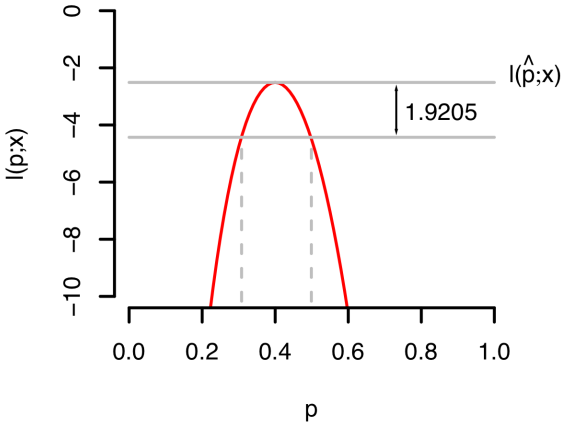
- Testing non-nested models
  - The likelihood ratio test can only test between two models that H_0 is nested in H_1
  - Akaike Information Criterion (AIC) is designed for testing non-nested models:

$$AIC = -2 \log L_i(\hat{\theta}_i) + 2p_i$$

    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
    - rationale: AIC aims to pick the model with the smallest *expected Kullback-Leibler* distance to the true model

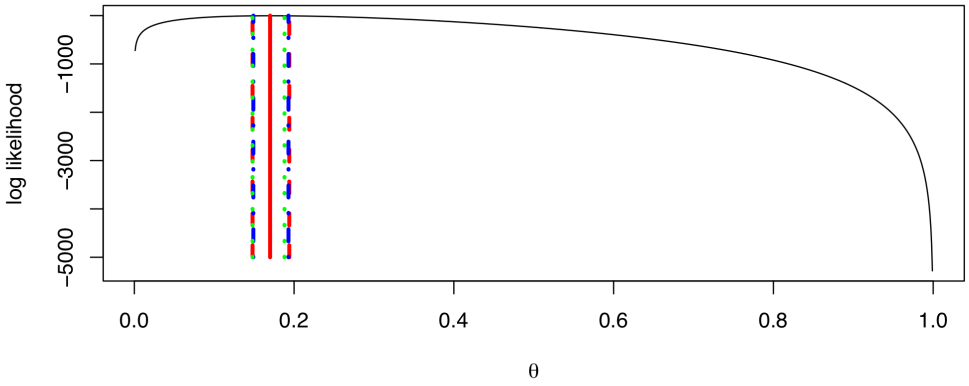
- Rule of thumb for multiple model comparison
  - models having AIC within 1-2 of the minimum: substantial support, should receive consideration in inference.
  - models having AIC within 4-7 of the minimum: considerably less support
  - models having AIC > 10 above the minimum: essentially no support
- Given you want to test Jukes-Cantor against the more general GTR model. Can you do a likelihood ratio test?
  - Yes, given you perform the test on the same tree with fixed branch lengths
  - No, if you need to perform the test on different trees (e.g. on the maximum likelihood tree for Jukes-Cantor and GTR, which may be different). Each tree is a different parameter, thus the full models are not nested. You need to employ the AIC.

- Confidence intervals
  - Each parameter value which is not rejected based on the likelihood ratio test at the 0.05 level is within the 95% confidence interval
  - In lecture 4, we learned that
    - determine the value of the log-likelihood function in $\hat{\theta}, l(\hat{\theta}; x)$
    - subtract $0.5\chi^2_{k,5\%}$ i.e. calculate $l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$
    - determine those θ values for which $l(\theta; x) = l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$



  - Same strategy can be used to calculate confidence intervals for the evolutionary parameters given a fixed tree.
  - How confident are we in the maximum likelihood estimate for rolling a 6 in our die experiment?
    - We calculate the confidence interval as explained above. (However, for complex objects like tree topologies, this is not possible)
    - We can do more experiments. The 95% interval of the experiment outcomes is obtained by ignoring the smallest 2.5% and largest 2.5% outcomes, and then considering the minimum and maximum outcome. However, for many questions, it is not possible to do more experiments (e.g. we can not repeat plant speciation)
    - We can mimic more experiments by “bootstrapping”. Bootstrapping refers to tests relying on random sampling with replacement. For an additional

die experiment, we sample 1000 results of the die roll from our initial 1000 die rolls (**with replacement**; otherwise we get the original result back). If we had enough die rolls initially, then the bootstrap results are the same as rolling the die again, and we obtain the 95% interval as in case 2 (redoing experiments).



- We show the log likelihood function of one experiment in solid black, the maximum likelihood estimate $\hat{\theta}_1 = k/n$ in a solid read line, and the 95%. confidence interval with dashed lines (case1). the 95% interval obtained from another 100 experiments are in green (case2). the 95% bootstrap confidence intervals are in blue (case3)

- Bootstrapping for phylogenies based on an alignment sequences with length m
  - sample m sites at random with replacement
    - allowing one particular sites to be picked up multiple times
  - infer a phylogeny based on the new data
  - repeat this procedure many times

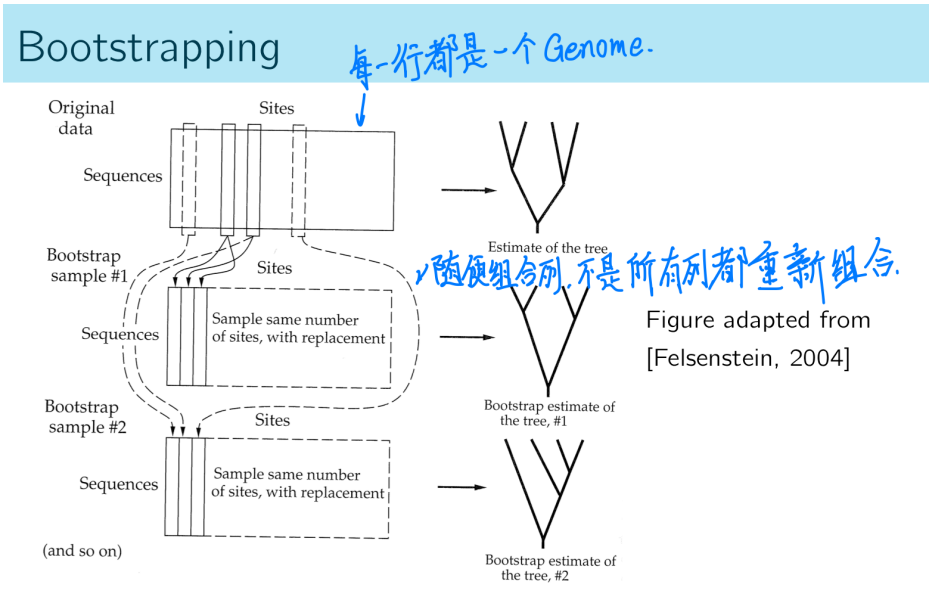
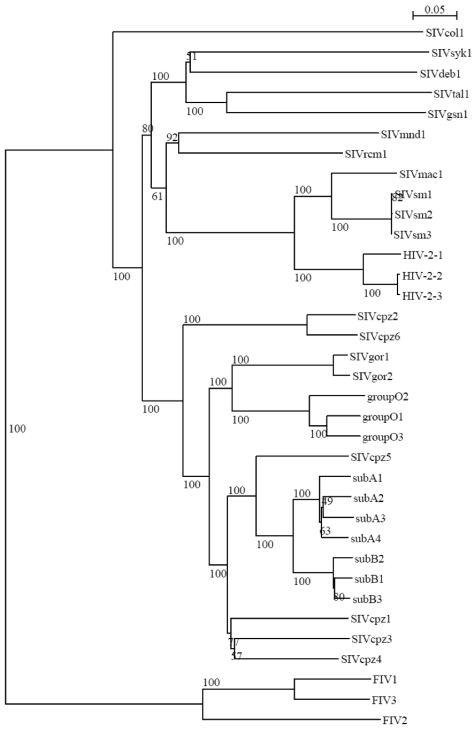


Figure adapted from [Felsenstein, 2004]

- Bootstrapping for an HIV/SIV tree



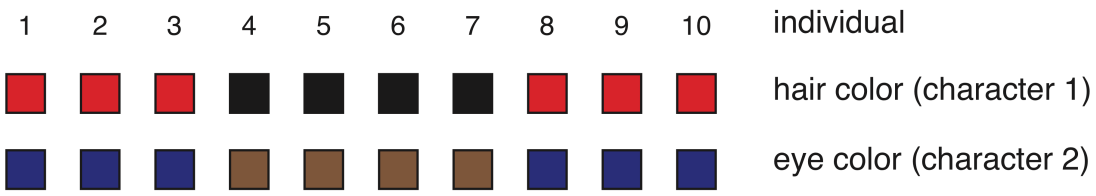
-Each node in the tree is labelled with the number of boot strap trees containing a node with the same descendant tips

- Overview of maximum likelihood inference
  - Infer a maximum likelihood tree
    - employ Felsenstein's pruning algorithm for each tree & branch lengths
    - choose the tree with branch lengths maximising the likelihood
    - do this for each substitution model and calculate its AIC
  - determine the substitution model & tree with highest support using AIC
  - determine the confidence interval for the substitution model parameters based on the likelihood ratios
  - determine the confidence in your maximum likelihood tree using bootstrap
- Phylogenetics in HIV research
  - See the lecture videos

- Questions:
  - Is there a way to test how to best root a maximum likelihood tree?
    - No, because there are no best way to root a maximum likelihood tree, since as long as the chosen nucleotide substitution model is time reversible, it does not matter how the tree is rooted.
  - Can you use the bootstrapping ideas for assessing confidence in a UPGMA tree?
    - Yes, we can use the idea of the bootstrapping for assessing confidence in the UPGMA tree. Essentially, we sample with replacement on the given sequences, then calculate the UPGMA tree. We repeat this processor for a lot of times and we can assess the confidence of the UPGMA tree.
  - What is required to infer the direction of transmission from a phylogeny?
    - In the phylogeny tree, if the sequence of patient A is nested within the sequence of patient B, it is not that likely that patient A transmitted the virus to patient B
    - But we can not just assume that patient B transmitted the virus to patient A because it might be that patient B transmitted the virus to patient C who we did not sequence and then C transmitted the virus to patient A.

Lecture 8: Continuous trains and comparative methods

- Outline of lecture 8
  - How can we compare **phenotypic traits/characters** between individuals/ species that **evolved on a phylogeny**?
  - Such characters/traits could be
    - Discrete
      - Spike numbers of HIV irons
      - number of legs in arthropods
      - fur patterns in rodents
    - Continuous
      - height
      - surface to weight ratio
      - virulence of influenza
      - shape of dinosaur jaws
- Comparing discrete characters
  - The problem by example
    - We want to know whether eye colour is correlated with hair colour. 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.
    - 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 happen due to chance alone.
    -

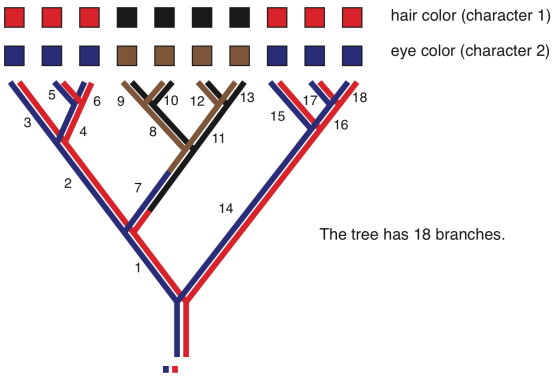
	brown eye	blue eye
red hair	0	6
back hair	4	0

$$P(\text{red hair/brown eye}) = \frac{(\#\text{ of red hair \& brown eye in red hair}) \times (\#\text{ of black hair \& brown eye in black hair})}{\#\text{ of comb brown eyes amongst all}}$$

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

  - We reject the hypothesis of independent character evolution, i.e. we can see a correlation

- The problem: a phylogenetic approach
  - However the analysis could be biased due to relatedness of the individuals
  - For example, if the above 10 individuals have the underlying evolution tree,



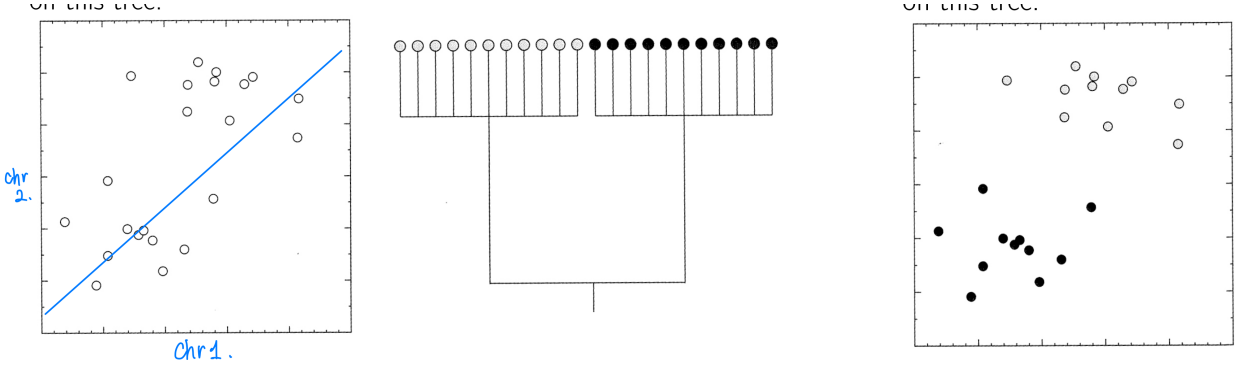
  - We can see that even through the hair colour and the eye colour clearly does not change simultaneously, (e.g. on branch 7, the hair and the eye colour do not change at the same time) We would be tricked into thinking that there is some kind of correlation by just looking at the individuals without the phylogenetic tree.
  - Correct way to look at the problem: Is the change of characters on the branches correlated?
  - $\mathscr{H}_0$: The character changes are equally likely on every branch

number of branches	change in eye colour	no change in eye colour
change in hair colour	1	0
no change in hair colour	0	17

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

- We cannot reject the hypothesis that character change is equally likely, i.e. we cannot say that there is a correlation between the characters
- To summarise: Neglecting the phylogenetic background can lead to false conclusions on correlations between characters. This is mainly the case because of non-independence of species data points as a result of shared ancestry.

- Comparing continuous characters
  - Continuous characters
    - So, far, we have mostly looked at evolution on a discrete space
      - nucleotide substitution models in lectures 3 and 4
      - codon and amino acid substitution models in lecture 4
      - space of tree topologies (not considering branch lengths) in phylogenetic reconstruction based on nucleotide sequences in lectures 5-7
      - correlation between discrete phenotypic characters in this lecture
    - For the rest of the lecture, we want to learn how evolution of **continuous phenotypic characters** (e.g. height, weight, virulence) can be modelled and how we can test for correlations amongst continuous traits.
  - Why linear regression can not be used to compare two characters evolved on a phylogeny?
    - Given a bunch of species with continuous phenotypic characters chr1 and chr2 on the left below, we can use linear regression to find a regression line for chr1 and chr2.

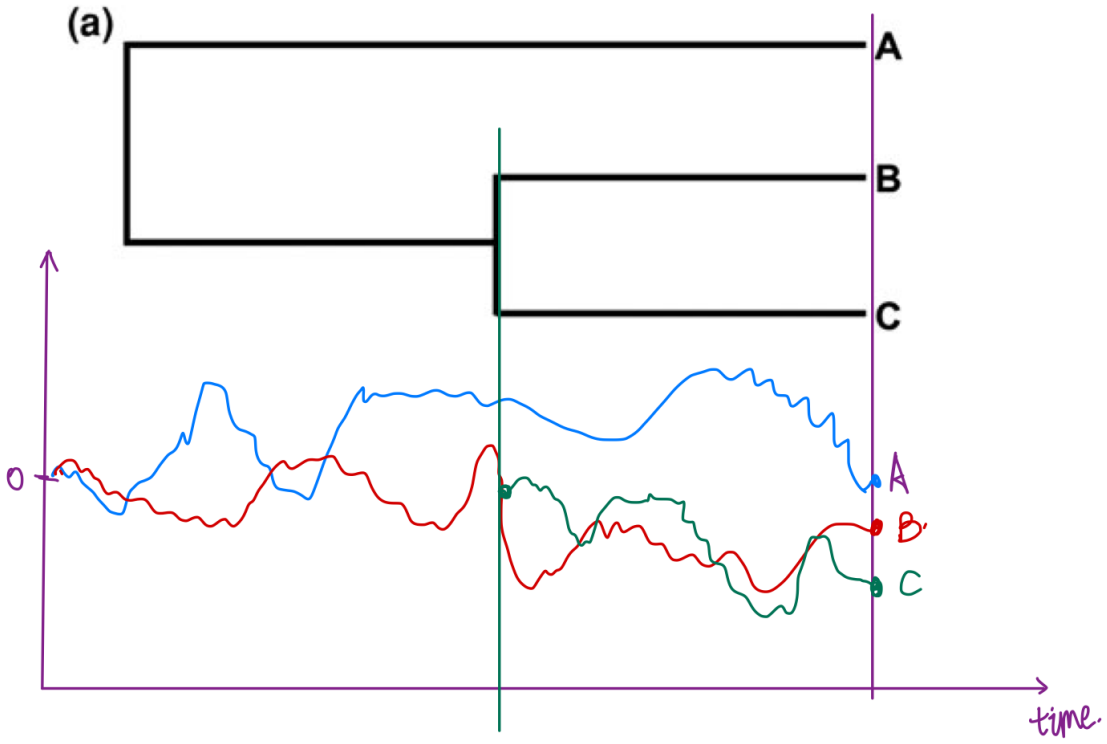


    - However, if we consider the phylogeny in the middle and the labeled graph on the right, we can see that there is no correlation between this two characters. There's only clade effects.
  - Toolbox: Brownian motion
    - One commonly used model to describe evolution of continuous traits on a phylogeny is the Brownian motion model

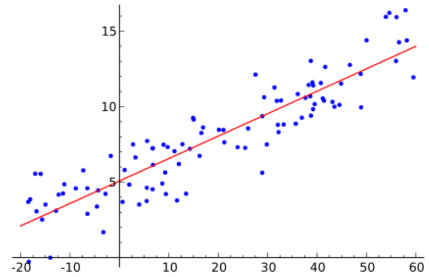
- In short: Brownian motion is described as a Wiener process, $(W_t)_{t\in T}$, which fulfils the following four conditions
  - $W_0 = 0$
  - W_t is almost surely continuous
  - W_t has independent increments (implies memorylessness)
    - for $0 \leq s_1 \leq t_1 < s_2 \leq t_2, (W_{t_1} - W_{s_1})$ and $(W_{s_1} - W_{s_2})$ are indipendent.
  - for $0 \leq s \leq t$, the $W_t - W_s \sim \mathcal{N}(0,\sigma^2(t-s))$
- Analogies between models for evolution indiscreet 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

- Given a phylogeny, we can apply a Brownian motion model on this phylogeny to evolve a continuous character



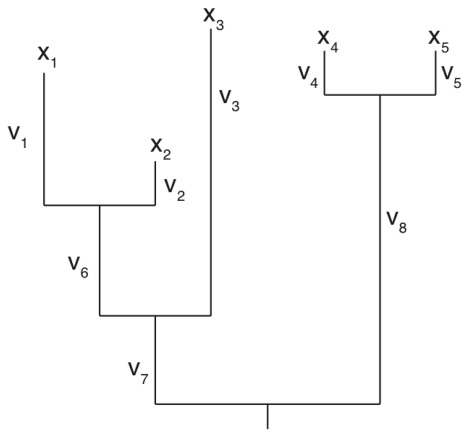
- Toolbox: Linear regression
  - Mathematical method to determine the dependency of a variable Y on another variable X. We measure X and Y for n independent realisations and fit a regression model to the data. The Observations $(x_1, y_1), \ldots, (x_n, y_n)$ need to be
    - independent
    - with the same (normally) distributed errors
  - Model
    - $y_i = \beta x_i + b + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\sigma^2)$
  - Fitting
    - Least squares method
    - Goodness of fit:
      - R^2: Perfect fit if close to 1; no dependency if close to 0



en.wikipedia.org/wiki/Linear_regression

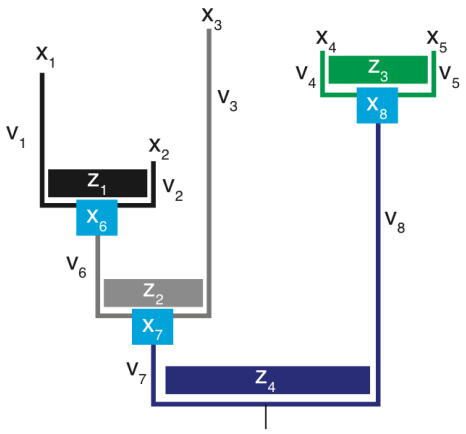
  - When two characters evolve on a tree
    - they share common evolutionary history (not independent realisations!)
    - the “error” variance added by Brownian motion is not equally distributed
- Constructing independent variables
  - One method to overcome interdependencies of the evolutionary trait is the contrast method

Suppose a phylogeny of 5 species:

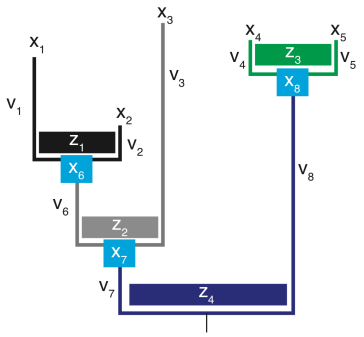


x_1 and x_2 are not independent as they share the evolutionary lineages v_6, v_7

Instead of characters, we look at their contrasts:



- Independent contrasts
  - We need to calculate/estimate the values of the contrasts and their variances in order to perform a linear regression on the contrasts.



$$z_3 = x_4 - x_5$$

$$z_1 = x_1 - x_2$$

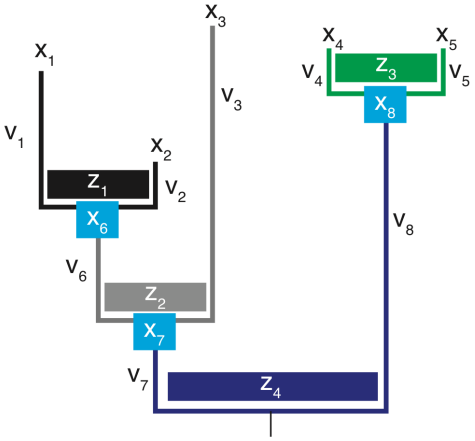
$$z_2 = x_6 - x_3$$

$$z_4 = x_7 - x_3$$

  - We assume character evolution according to Brownian motion
  - We observed the tip values, but we have to estimate the values at internal nodes
  - To calculate the variance, we just apply

$$Var[\alpha X + \beta Y] = \alpha^2 Var[X] + \beta^2 Var[Y] + 2\alpha\beta Cov[X, Y]$$

- Constrasts at cherries



$$z_1 = x_1 - x_2$$

$$Var[x_1] = \sigma^2(v_1 + v_6 + v_7) \text{ ..... from tip } x_1 \text{ to root}$$

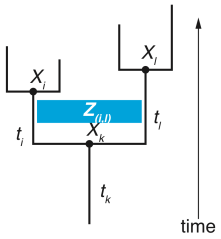
$$Var[x_2] = \sigma^2(v_2 + v_6 + v_7) \text{ ..... from tip } x_2 \text{ to root}$$

$$Cov[x_1, x_2] = \sigma^2(v_6 + v_7) \text{..... co-ancestor to the root}$$

$$\begin{aligned} Var[z_1] &= var[x_1 - x_2] \\ &= Var[x_1] + Var[x_2] - 2Cov[x_1, x_2] \\ &= \sigma^2(v_1 + v_6 + v_7 + v_2 + v_6 + v_7 - 2(v_6 + v_7)) \\ &= \sigma^2(v_1 + v_2) \end{aligned}$$

  - The value for contrasts at cherries can easily be calculated. The variance is proportional to the branch lengths between the two external nodes.

- Contrasts further down in the tree



- Wanted: $z_k = x_i - x_j$ and $Var[z_k]$

- We have to calculate the values at internal nodes

$$x_i = \frac{v_n}{v_m + v_n} x_m + \frac{v_m}{v_m + v_n} x_n$$

$$Var[x_i] = Var[\frac{v_n}{v_m + v_n} x_m + \frac{v_m}{v_m + v_n} x_n]$$

$$= \sigma^2 \left(\frac{v_m v_n}{v_m + v_n} + v_i + v_k + \cdots \right)$$ ..... the path from x_k to root

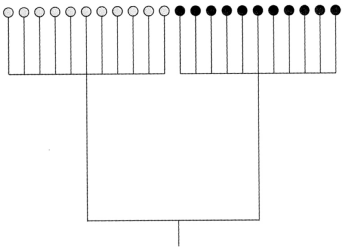
- We calculate the corrected branch length using:

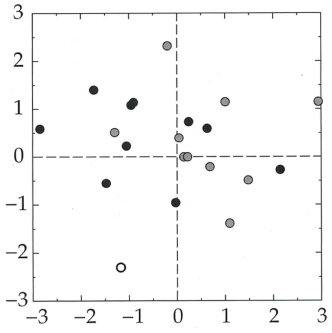
$$v_i' = \frac{v_m v_n}{v_m + v_n} + v_i$$

- then the contrasts and the variance can be calculated using

$$z_k = x_i - x_j \text{ and } Var[z_k] = \sigma^2 (v_i' + v_j')$$

- Normalisation of contrasts
  - To be able to compare the independent contrast, all contrasts need to have the same variance. Thus, we need to normalise the contrasts in a last step.
  - Given the contrast z_k with variance $Var[z_k] = \sigma^2 c_{z_k}$ where $c_{z_k} = v_i' + v_j'$. We know that $Var(\alpha X) = \alpha^2 Var(X)$. Thus we can replace the contrasts by $Z_k = z_k / \sqrt{c_{z_k}}$
  - Therefore all $Z_k \sim \mathcal{N}(0, \sigma^2)$ and are ready for a linear regression. Of course, one needs to calculate and normalise the values of the second measured trait following the same formulae.
- Example
  - We use the calculation of independent contrasts on the linear regression example above,





    - The independent contrast method supports our early suspicion that no correlation between the two characters evolved on this particular phylogeny can be found.

- Q & A
  - In a Fisher's exact test, how would you calculate which values for one of the cells in the contingency table would lead to a rejection of the null hypothesis, given that row and column sums remain the same?
    - Fisher's exact test: Statistical test to examine the significance of the association between two kinds of classifications

	Case	Control	Total
Minor	a	b	a+b
Major	c	d	c+d
	a+c	b+d	n=a+b+c+d

- H_0: Class A is not linked to class B. (The number of individuals expressing both A_1 and B_1 is based on chance

$$p - value = \sum_{i=a}^{a+b} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{n}{a+c}}$$

    - If we see the p-value as a function of a, we can see that the function is monotone, so we can just increase from a = 0The var, till a = a + b and then we choose one a that makes the p value < 0.05

  - Is the Brownian motion model a good model for all continuous traits? Could you imagine situations where this is not the case and which assumption in this model could be violated?
    - Brownian Motion model
      - the good:
        - Memorylessness ⇒ independent contract method
        - Simple and intuitive
      - the bad
        - No selective pressure
        - No direction
        - The variance increases with time. The more you wait, the more "extreme" result can occur, like a height = 30cm.
  - Do you think it is a good strategy to first determine the species tree and then look at character evolution, or would a co-estimation of characters and phylogeny make more sense?
    - Co-estimation of characters
      - Good:

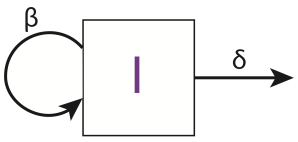
        - Generally it's good to assume that character and the species do not evolve independently.
    - Bad
        - Trait and sequences are intertwined

Lecture 09 - Phylodynamics

- Introduction
    - So far, we have learned how to infer a tree-like phylogeny from a bunch of given sequences. Now we would like to know what is the process that generated the phylogenetic tree?
    - Phylogenetic trees encode past macroevolutionary dynamics
        - Macroevolution: individuals = species
            - (Molecular). Evolution
                - (Genetic) makeup of species changes through time
            - Phylogenetics
                - Phylogeny displays species relationships
            - Phylodynamics
                - Population dynamics is the speciation and extinction process
                - Quantify the speciation and extinction rate
        - Epidemiology: individuals = infected hosts
            - Evolution
                - Pathogen is evolving through time
            - Phylogenetics
                - Phylogeny displays transmission history
            - Phylodynamics
                - Population dynamics is the transmission and becoming non-infectious process.
    - Phylodynamics:
        - Population dynamics models the birth and death of individuals
            - Immunology: individuals = B cells
                - Phylogeny displays B cell differentiation through somatic hypermutation
                - Population dynamics is the B cell generation and loss process
            - Cancer: Individuals = cells

    - Phylogeny displays relationship of different cancer cells and healthy cells
    - Population dynamics is the spread and loss of cell types
  - Languages: individuals = languages
    - Phylogeny displays language evolution
    - Population dynamics is the gain and loss of languages
- The birth and death process gives rise to a phylogenetic tree
- Phylodynamics aim to understand and quantify the population dynamics based on a phylogenetic tree. Today we quantify birth and death dynamics given the phylogenetic tree and then also R_0
  - R_0 is the average number of secondary infections caused by a single infected individual at the start of an epidemic

- Population dynamic models
  - Birth-Death process



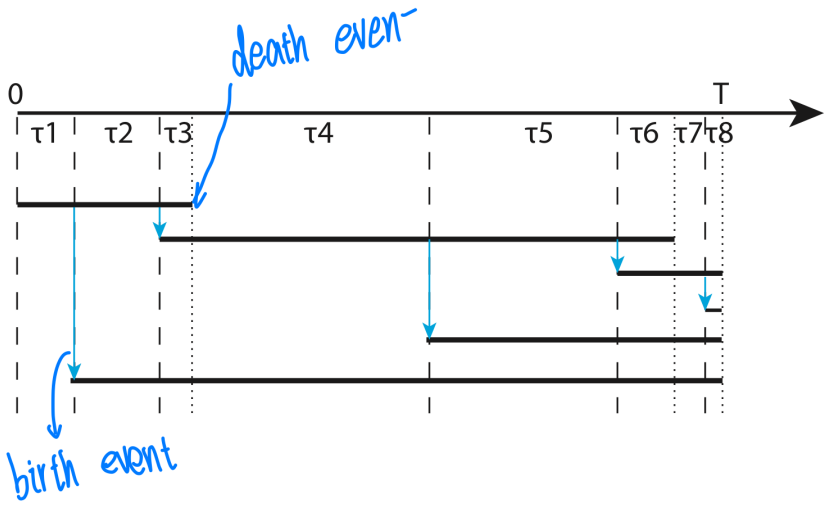
-Rate of birth of new individuals per individual in I: β

-Rate of death per individual in I: δ

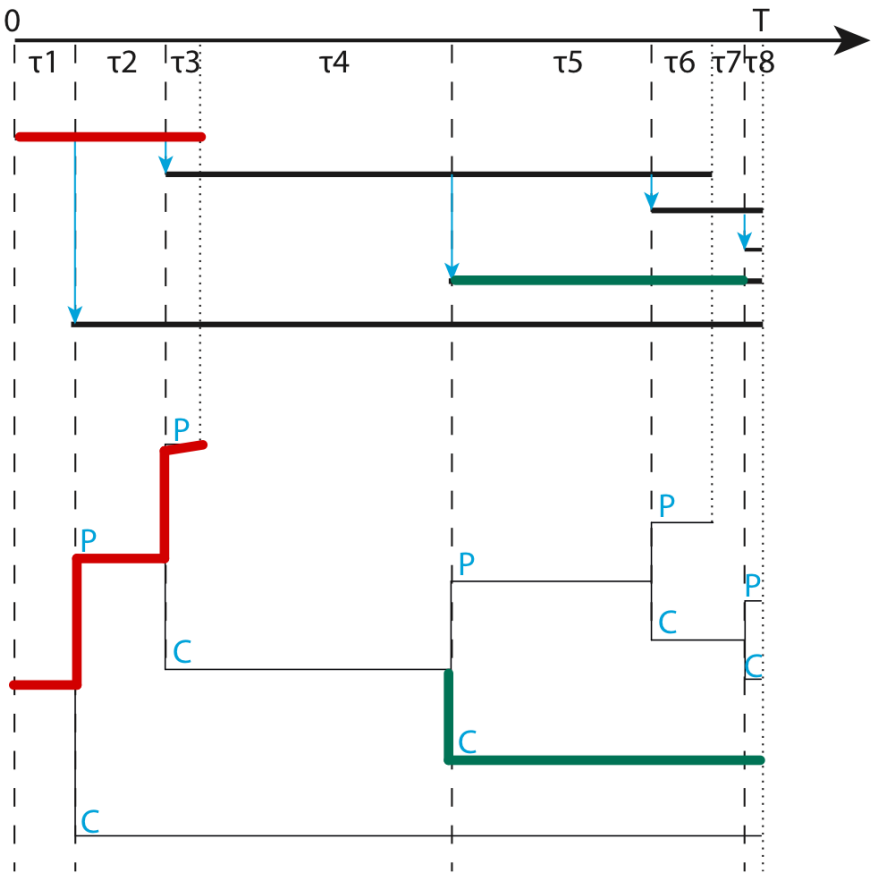
-Such a process is called a (linear) birth-death process

  - Stochastic population dynamics
    - Consider the fate of one individual
      - The probability of giving birth to another individual in a very small time step Δt is $\beta \Delta t$
      - The probability of dying in a very small time step Δt is $\delta \Delta t$
      - The waiting time to a birth event is exponentially distributed with parameter β
        - see section “from rate to probability” on page 12
      - The waiting time to the first event (birth or death) is exponentially distributed with parameter $\beta + \delta$ (minimum of two exponentially distributed random variables with rates r_1, r_2 is exponentially distributed with the rate $r_1 + r_2$)
    - Consider the fate of N individuals
      - The waiting time to the first event (birth or death) is exponentially distributed with parameter $N(\beta + \delta)$

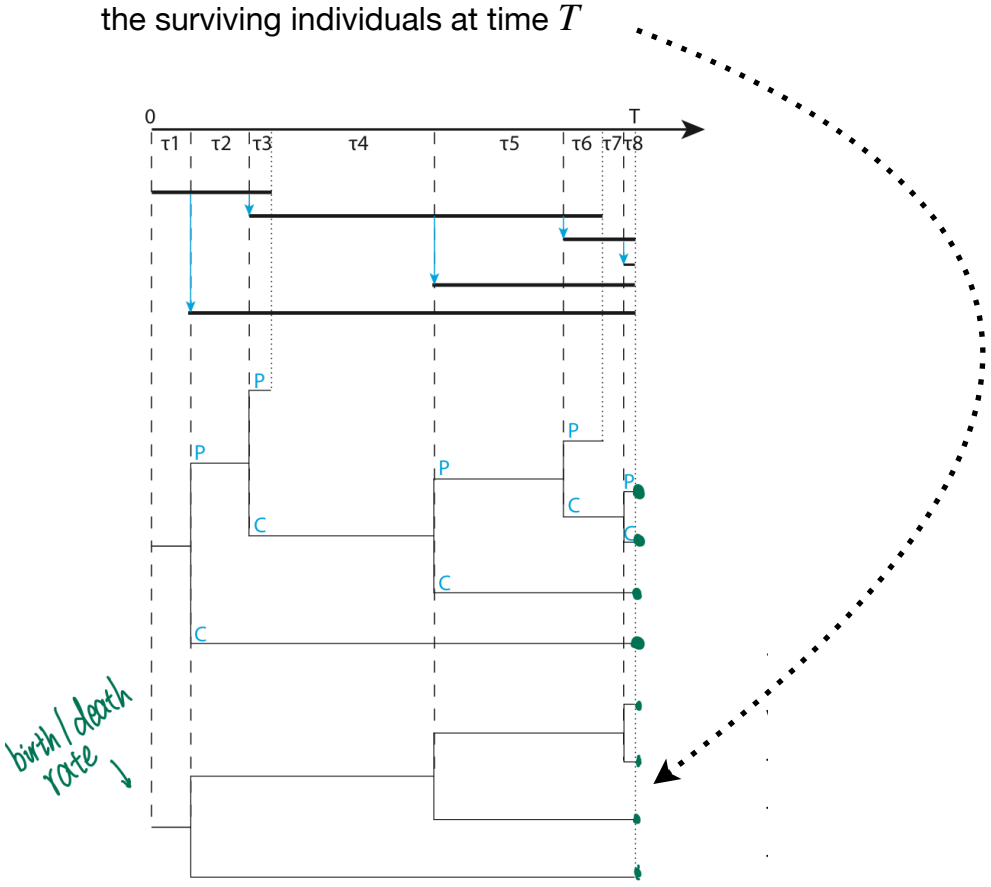
- From population dynamics to phylogenetic trees
  - The diagram illustrates the full population dynamics of a birth-death process which starts with one individual and is stopped after time T. Each solid black line is the lifetime of an individual. Blue arrows are birth events



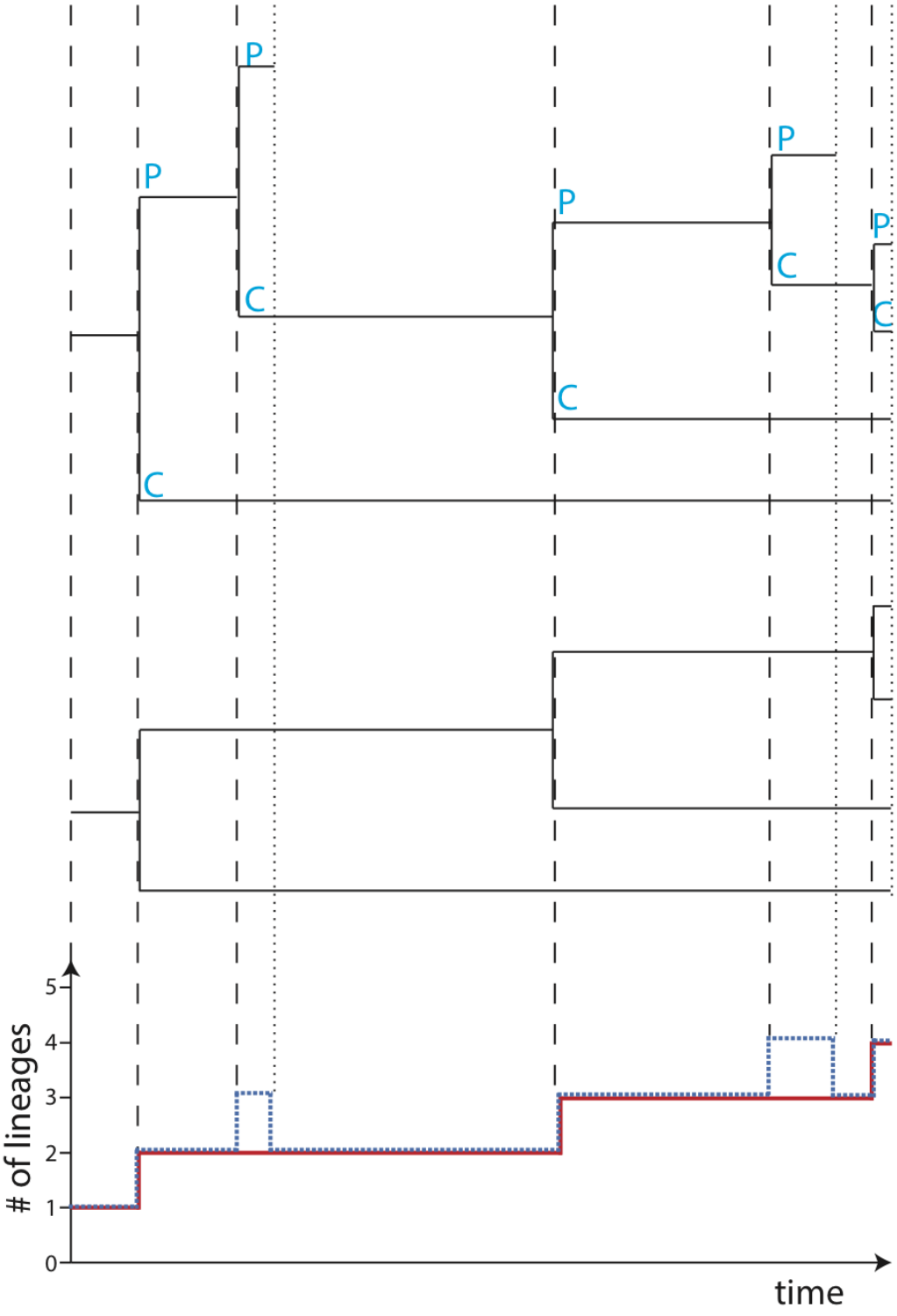
  - A complete population tree displays the full population dynamics. The labels P and C illustrate the parent-child associations



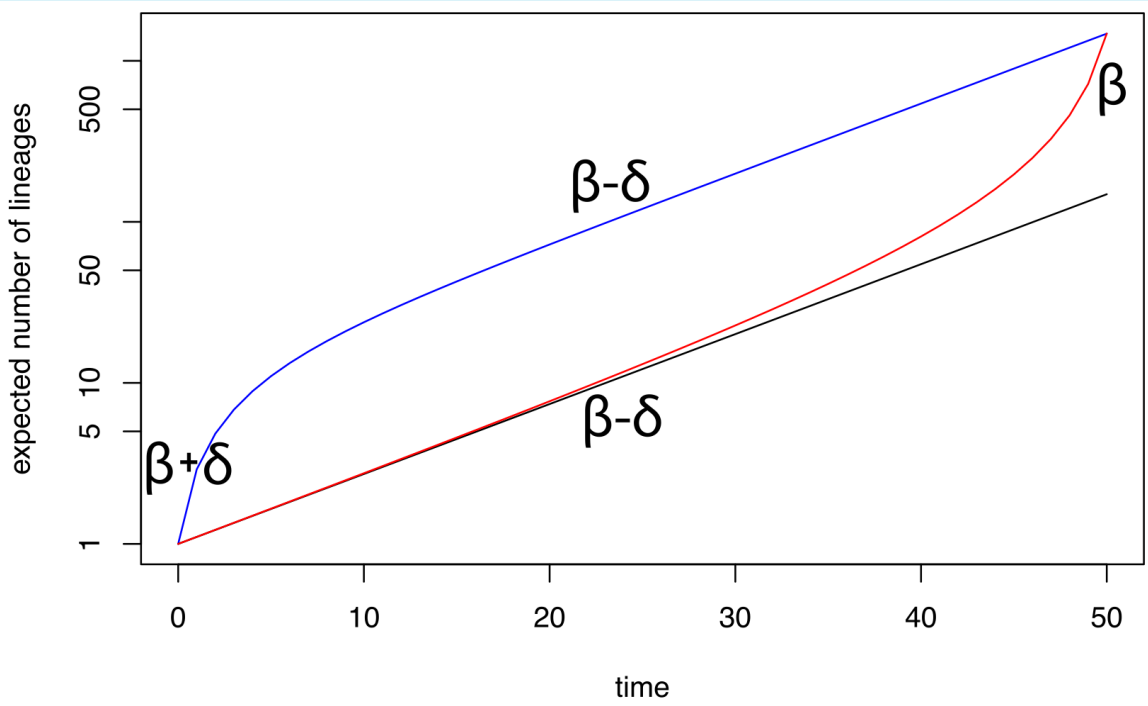
- Phylodynamic models
  - A simple phylodynamic model
    - A phylodynamic model adds a sampling process of individuals to the population dynamics (sampling is equivalent to sequencing the individual and adding it to the phylogenetic tree). The simplest phylodynamic model is:
      - Birth rate β
      - Death rate δ
      - Process duration T
      - Extant(尚存的) tip sampling probability ρ
      - Extinct tip sampling probability Φ
    - We will now assume $\rho = 1, \Phi = 0$. For macroevolution, that means no fossil sampling and complete extant species sampling. The subtree of the complete population tree connecting the sampled individuals, and ignoring the parent-children labels, is called the *phylogenetic tree*. The phylogenetic tree is the tree we infer from data.
  - From population dynamics to phylogenetic trees
    - The phylogenetic tree with $\rho = 1, \Phi = 0$ displays the dynamics giving rise to the surviving individuals at time T



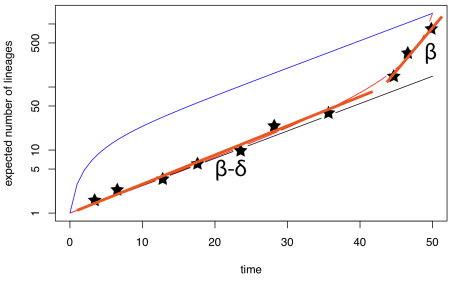
- Lineages-through-time plot
  - Plotting the number of lineages (y-axis) vs time(x-axis) is called lineage-through-time(LTT) plot. The LTT plot of the complete tree (blue; dashed) shows the population size through time. The LTT plot of the phylogenetic tree (red) shows the number of surviving lineages through time.



- Can we estimate the birth and death rates from reconstructed phylogenies?
  - Average LTTs for a population of age $T = 50$:
    - Red: Average number of lineages in the phylogenetic tree.
      - much like the LTT above, we can simulate for phylogenetic trees and draw more red lines, then for each time point calculate the average number of lineages. Then, you plot the log transformed number of lineages and you get the red line below.
    - Blue: Average of surviving population trajectories.
      - Just like mentioned above, but this time, we simulate on the full dynamic population tree model. (The first subplot in the plot on page 66).
    - Black: Exponential growth curve (linear with slope $\beta - \delta$ on the log scale.
  - The early blue part is called push-of-the-past, the late red part is called pull-of-the-present
    - We observe a push-of-the-past as only individuals with a quick replication early on will produce surviving populations
    - We observe the pull-of-the-present (i.e. an apparent acceleration in diversification towards the present) as the very recent lineages did not yet have time to go extinction.



- Estimating β and α
  - In the plot from the right, each black star is a branching event and the y axis is the log-ed number of lineages.
  - We could fit a regression line to the early branching events and estimate its slope. This is an estimation for $\beta - \delta$
  - We could fit a regression line to the late branching events and estimate its slope. This is an estimation for β



  - 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.
    - where to cut the line.

- Probabilities
  - Probability density of a tree
    - Phylogenetic likelihood : $L(\mathcal{T}, Q) = P(A \mid \mathcal{T}, Q)$
    - Phylodynamic likelihood: $L(\eta = (\beta, \delta, T)) = P(\mathcal{T} \mid \eta)$
    - We aim to determine the maximum likelihood estimate for the parameters (β, δ, T), given a phylogenetic tree. Note that in this lecture, the age of the process T is assumed to be fixed to a known value. (This will be relaxed in lecture 11)
    - In order to do maximum likelihood estimation, we now derive $P(\mathcal{T} \mid \eta)$. This requires us to first derive the probability of a single individual after time t leaving 0 or 1 offspring, we denote this by $p(0 \mid t, \beta, \delta)$ and $p(1 \mid t, \beta, \delta)$
  - Probability of extinction, $p(0 \mid t)$
    - Suppose you start the birth-death process with one individual. What is the probability that no surviving individuals remains after time t ($p(0 \mid t, \beta, \delta)$)?
    - On the following pages we will use the abbreviated notation $p(0 \mid t) = p(0 \mid t, \beta, \delta)$
    - Consider a small tilmestep Δt during which only one event occurs
    - During that time step, for a single individual, a death event happens with probability $\delta \Delta t$ and a birth event happens with probability $\beta \Delta t$. No event happens with probability $1 - (\beta + \delta)\Delta t$

- The resulting individual after time Δt have probability $p(0 \mid t)$ to go extinct within time interval t

- Thus

$$p(0 \mid t + \Delta t) = (1 - (\beta + \delta)\Delta t)p(0 \mid t) + \delta \Delta t + \beta \Delta t p(0 \mid t)^2$$

- Rearranging leads to:

$$\frac{p(0 \mid t + \Delta t) - p(0 \mid t)}{\Delta t} = -(\beta + \delta)p(0 \mid t) + \delta + \beta p(0 \mid t)^2$$

- Taking the limit $\Delta t \to 0$ leads to

$$\frac{d}{dt}p(0 \mid t) = -(\beta + \delta)p(0 \mid t) + \delta + \beta p(0 \mid t)^2$$

- The initial condition is $p(0 \mid 0) = 0$ meaning that at time point 0, there must be at least one individual.

- Solution for $p(0 \mid t)$

  • From the above calculation, in order to obtain $p(0 \mid t)$ we have to solve the following differential equation

$$\begin{cases} \frac{d}{dt}p(0 \mid t) = -(\beta + \delta)p(0 \mid t) + \delta + \beta p(0 \mid t)^2 \\ p(0 \mid 0) \quad = 0 \end{cases}$$

  • The solution to this equation is

$$p(0 \mid t) = \frac{\delta(1 - e^{-(\beta - \delta t)})}{\beta - \delta e^{-(\beta - \delta)t}}$$

- Probability of n descendants, $p(n \mid t)$

  - In general the probability to obtain n surviving lineages after time t, $p(n \mid t; \beta, \delta)$ (for which we again write short $p(n \mid t)$) is

$$p(0 \mid t) = \frac{\delta(1 - e^{-(\beta - \delta t)})}{\beta - \delta e^{-(\beta - \delta)t}}$$

$$p(1 \mid t) = e^{-(\beta - \delta)t}(1 - p(0 \mid t))^2$$

$$p(n \mid t) = p(1 \mid t)\left(\frac{\beta}{\delta}p(0 \mid t)\right)^{n-1} \qquad \text{for } n \geq 2$$

  - We now provide a proof for $p(1 \mid t)$. A proof for $p(n \mid t)$ can be obtained using an induction.

- Probability of 1 descendant, $p(1 \mid t)$

$$p(1 \mid t + \Delta t) = (1 - (\beta + \delta)\Delta t)p(1 \mid t) + \beta \Delta t \binom{2}{1} p(1 \mid t)p(0 \mid t)$$

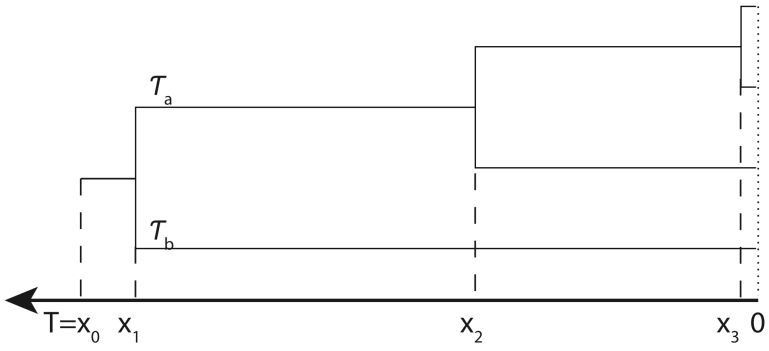
So that we have

$$\frac{d}{dt} p(1 \mid t) = -(\beta + \delta)p(1 \mid t) + 2\beta p(1 \mid t)p(0 \mid t)$$

with initial condition $p(1 \mid 0) = 1$

- The fraction of 2 in the differential equation for $p(1 \mid t)$ accounts for either one of the descendants of the birth event leading to the surviving individual after time t
- Evaluating the left- and right-hand side of the differential equation using $p(1 \mid t) = e^{-(\beta-\delta)t}(1 - p(0 \mid t))^2$ shows that this function is a solution to the differential equation

- Probability density of a tree, $P(\mathcal{T} \mid x_0)$

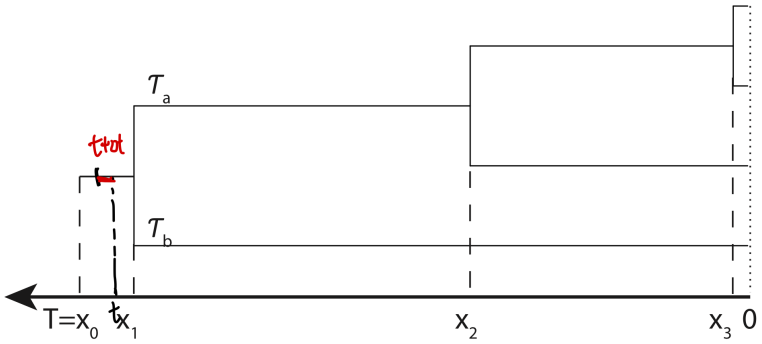


- Time is measured as age relative to the present (i.e. where $t = 0$)
- We take a dynamic programming approach
- Let $p(x_0, x_1)$ be the probability density for a branch of length $x_0 - x_1$ extending from an individual at time x_0 in the past
- Then the probability density of a tree $\mathcal{T}$ with age x_0 is

$$p(\mathcal{T} \mid x_0) = p(x_0, x_1)\beta p(\mathcal{T}_a \mid x_1)p(\mathcal{T}_b \mid x_1)$$

- with $p(\mathcal{T} \mid x) := p(\mathcal{T} \mid \eta = (\beta, \delta, T = x))$.

- Probability density of a branch, $p(x_0, x_1)$
  - We calculate the probability density of the branch between t and x_1, $p(t, x_1)$



  - $p(t + \Delta t, x_1) = (1 - (\beta + \delta)\Delta t)p(t, x_1) + 2\beta\Delta t p(t, x_1)p(0 \mid t)$
  - $p(x_1, x_1) = 1$
  - This leads to the differential equation:
    - $\frac{d}{dt}p(1 \mid t) = -(\beta + \delta)p(1 \mid t) + 2\beta p(1 \mid t)p(0 \mid t)$
  - This is the same differential equation as for $p(1 \mid t)$
  - As the initial condition is different, we have

$$p(x_0, x_1) = p(1 \mid x_0)/p(1 \mid x_1)$$

- Probability density of a tree, $p(\mathcal{T} \mid x_0)$
  - For a tree on n present day tips, age of the process x_0, and branching times $x_1, x_2, \ldots, x_{n-1}$, we have the probability density

$$p(\mathcal{T} \mid x_0) = p(x_0, x_1)\beta p(\mathcal{T}_a \mid x_1)p(\mathcal{T}_b \mid x_1) = \beta^{n-1}\prod_{i=0}^{n-1} p(1 \mid x_i)$$

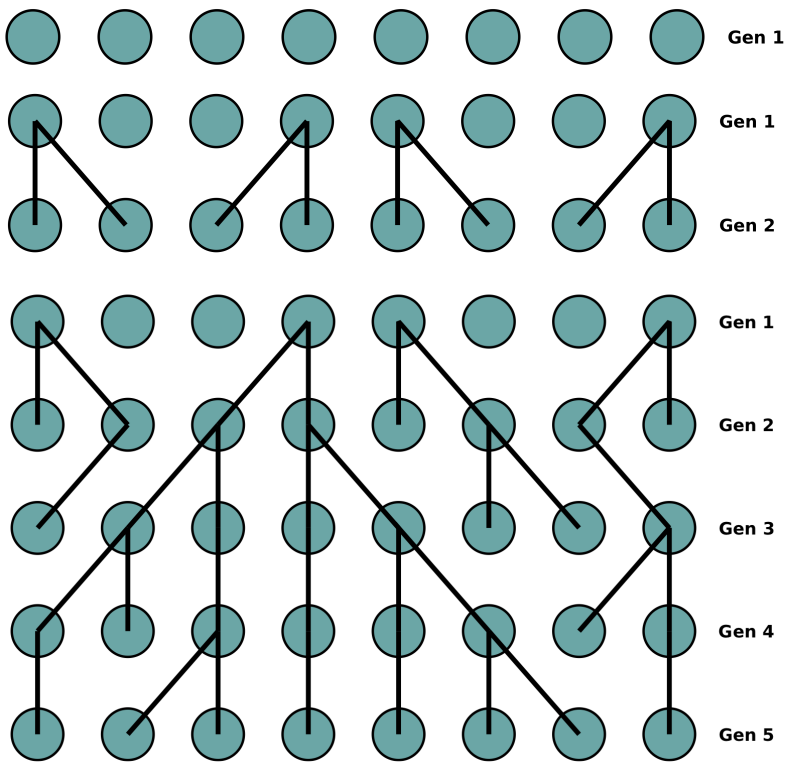
  - Analogous strategy provides us with a tree probability density when $\rho < 1$ (incomplete extant sampling) and $\Phi > 0$ (sampling through time)

- Applications
  - Quantifying the spread of Ebola in 2014
    - In each epidemic, it is crucial to know the basic reproductive number R_0. This is the number of secondary infectious caused by a single infected individual in a susceptible population. The number indicates the amount of public health effort for containing the epidemic, i.e. pushing the number of the second infectious below 1

  - By August 2014, 72 Ebola genomes from different patients in a Sierra Leone outbreak were available. The publication also included the phylogenetic tree
  - We want to calculate the basic reproductive number based on these trees
  - As it was early in the epidemic, they assumed a constant β (transmission rate) and δ (becoming-uninfectious rate, ending with recovery or death)
  - The data were sampled through time. i.e. $\rho = 0$ but $\Phi = 0.7$
  - $R_0 = \hat{\beta}/\hat{\delta}$, They obtained the maximum likelihood estimates $\hat{\beta}$ and $\hat{\delta}$ using the phylodynamic likelihood, and estimated $R_0 = 1.34$
  - Bayesian methods improved these estimates
- Questions
  - How does the approximate number of steps required to calculate the phylodynamic likelihood depend on the number of leaves in a phylogenetic tree? (I.e. what is the time complexity of this calculation?)
    - The number of steps that are required for this calculation equals to the number of internal nodes. For a phylodynamic tree with n leaves, there are $n - 1$ internal nodes. Thus the time complexity of this calculation is $O(n)$
  - What kind of population dynamic process could a decrease in slope in the LTT plot reflect?
    - The slope in the LTT plot is $\beta - \delta$, a decrease in slope would mean a decrease in the birth rate or an increase in the death rate
  - Assume a birth-death process where each individual at present is sampled with probability ρ. How is the derivation of $p(0 \mid t, \rho)$, the probability of sampling no individual at present, different compared to the derivation of $p(0 \mid t)$?
    - Since through out time, we do not do sampling on time interval $(\Delta t, t + \Delta t)$, so everything with respect to the differential equation should be the same
    - The difference is in the initial condition. If each individual at present is sampled with probability ρ, we can say that $p(0 \mid 0) = 1 - \rho$

Lecture 10: Coalescent Theory

- Introduction: Alternative tree-generating process: The Coalescent
  - Introduced by Kingman coalescent.
  - 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
- The Wright-Fisher process
  - Discrete generations
  - Each generation consists of N individuals
  - Each individual in the offspring population chooses 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 with accounts for the number of copies of a gene present in each individual.
    - “ploidy” means the number of complete sets of chromosomes in a cell
    - E.g. for a diploid organism, the number of copies of a gene in the population is 2N.
  - Example



- The sampled Wright-Fisher Phylogeny
  - imagine we randomly chose two individuals out of a total of N at generation i, and we want to know the probability of coalescence in generation $i - m$
    - The probability of coalescence at generation $i - 1$ is $p_{coal} = \binom{N}{1}\left(\frac{1}{N}\right)^2$

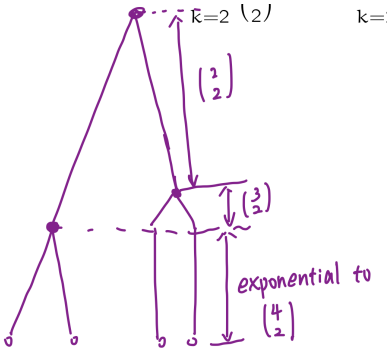
$$\begin{aligned} p(m) &= p(\text{no coalescence in generation m-1}) \cdot p_{coal} \\ &= (1 - p_{coal})^{m-1} \cdot p_{coal} \\ &= (1 - \frac{1}{N})^{m-1} \cdot \frac{1}{N} \end{aligned}$$

    - For large N, $p(m \mid N) \to \exp\left[-\frac{m-1}{N}\right]\frac{1}{N}$
  - The coalescent in calendar time
    - m is the number of generations. Let g be the calendar time of a generation (e.g. 5 days). 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
- Sampling on k-individual phylogeny
  - Question: How can this be generalised to k samples
  - Answer: Assuming $k \ll N$, we can model p_{coal} using $p_{coal} \simeq \binom{k}{2}\frac{1}{N}$

- Coalescent trees
  - Kingman's coalescent
    - Developed by John Kingman in a series of papers in 1982
    - Continuous-time Markov process which produces sampled time trees.
    - Process occurs *backwards* in time
    - Equivalent to sampled trees produced by Wright-Fisher model when N is much larger than the number of samples
    - Times between coalescence events are drawn from exponential distributions with rate parameters $\binom{k}{2}\frac{1}{Ng}$

$$p(\Delta t \mid N, g, k) = \exp\left[-\Delta t \binom{k}{2}\frac{1}{Ng}\right]\binom{k}{2}\frac{1}{Ng}$$

- The age of a coalescent tree
  - Under the coalescent model, the average time required for n lineages to coalesce into one is

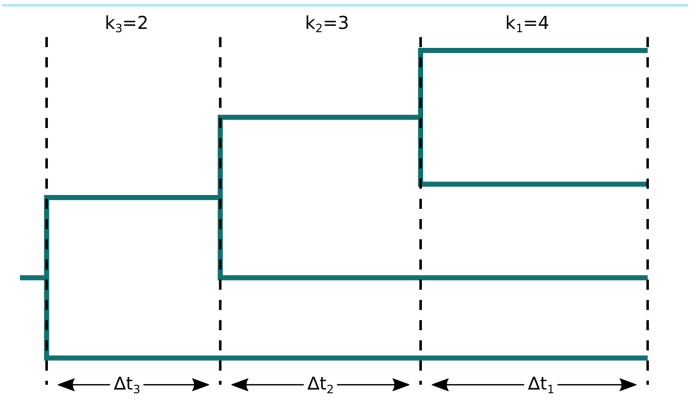


$$\mathbb{E}[t_{root}] = \sum_{k=2}^{n} \frac{Ng}{\binom{k}{2}} = Ng \sum_{k=2}^{n} \frac{1}{\binom{k}{2}}$$

- Can write $\sum_{k=2}^{n} \frac{1}{\binom{k}{2}} = \sum_{k=2}^{n} \frac{2}{k(k-1)}$
- Can expand $\frac{2}{k(k-1)} = \frac{2}{k-1} - \frac{2}{k}$
- Then $\sum_{k=2}^{n} \frac{1}{\binom{k}{2}} = \sum_{k=1}^{n-1} \frac{2}{k} - \sum_{k=2}^{n} \frac{2}{k} = 2(1 - \frac{1}{n})$

  - We therefore find that $\mathbb{E}(t_{root}) \to 2Ng$ as the number of n (i.e. number of leaves in the coalescent tree) becomes large.
  - This is an upper bound on the *expectation*: individual coalescent trees can be older than this.

- The probability of a coalescent tree



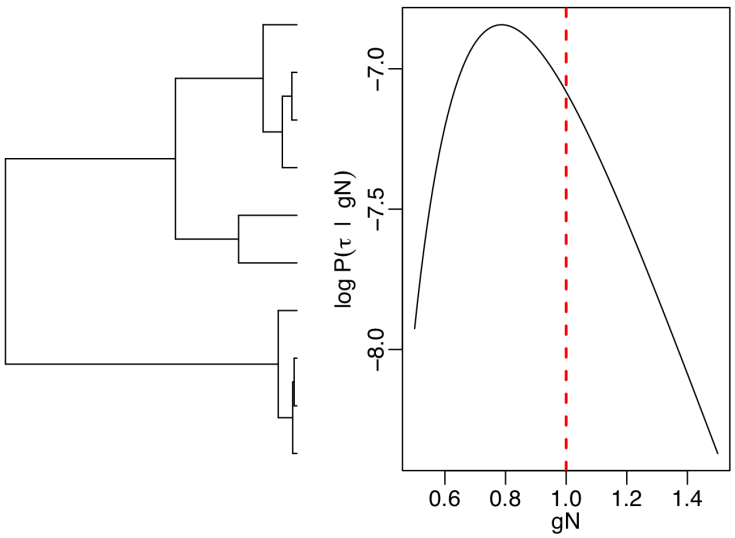
  - For a given tree above we can have the following probability

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

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

  - The exponentials give the probability of nothing happening in interval Δt_i, and the $1/gN$ factors are the probability densities of the particular coalescent events. (Note the units!)

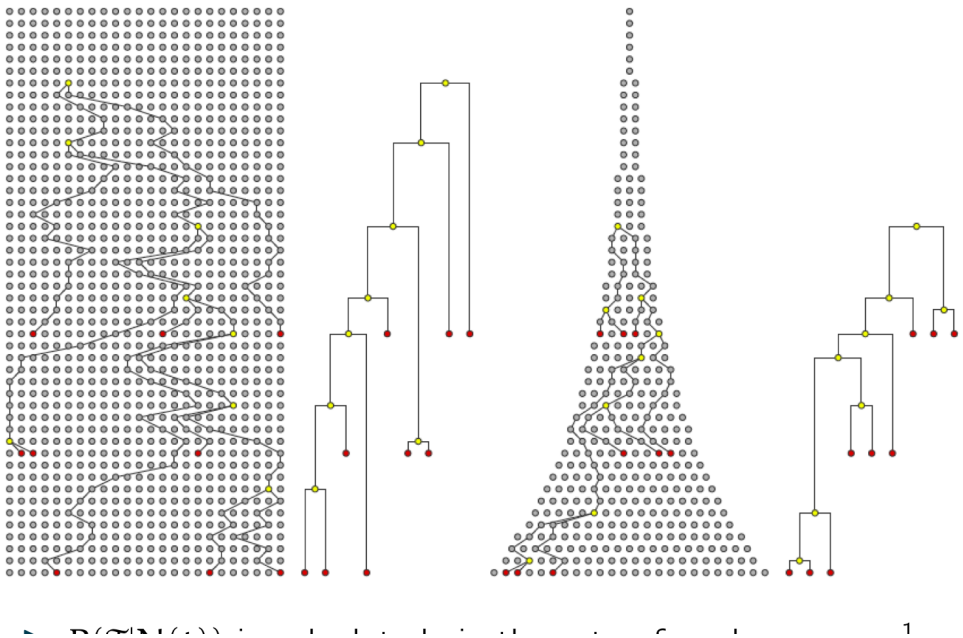
- Population size inference
  - For the plot given below, we generated the coalescent tree $\mathcal{T}$ using parameter $gN = 1$, if we plot the likelihood of the coalescent tree on the y-axis and gN on the x-axis, and do a maximum likelihood estimation, we can see that there is a bias.
  - For a given coalescent tree $\mathcal{T}$, $L(gN; \mathcal{T}) = P(\mathcal{T} \mid gN)$



- Inference and effective population size
  - If we take a real tree (e.g. one inferred from genetic data sampled from a real biological population) and infer the population size using the coalescent distribution, we can expect our result to be biased since the real dynamics differ from the WF dynamics
  - One of the important ways that reality differs is that real populations are *structured*, where the WF population is assumed to be completely homogeneous
  - In any case, the inferred population size is referred to as the *effective* population size, sometimes written as N_e
  - This is the size of a WF population which shares some statistical similarity with the real population
  - Care must be taken when drawing conclusions from effective population sizes
- Robustness of the coalescent
  - The coalescent distribution/process is often derived as a limit of the Wright-Fisher process, as we have done here.
  - It also appears as the limit of many other population processes, for example

  - The Canning model (generalisation of the Wright-Fisher model)
  - The Moran model (overlapping generations, fixed population size)
  - some stochastic logistic models (continuous time, population fluctuations)
  - …
- 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
    - Justifies use of fixed or slowly varying population size
  - Number of samples is small compared to the total population size
    - Justifies neglect of > 2 lineages coalescing in the same generation
  - Population is “well mixed”, samples are drawn uniformly at random
    - Justifies the coalescent rate between any pair of sampled lineages being equal
    - Population structure violates this assumption
- Extension: population size changes



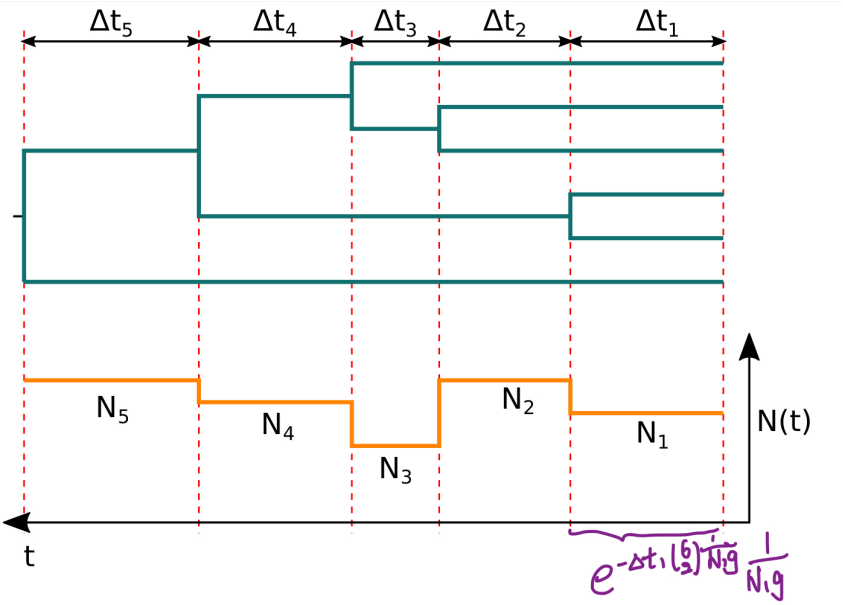
  - $P(\mathcal{T} \mid N(t))$ is calculated via the rate of coalescence $\frac{1}{N(t)}$ (where $N(t)$ is the population size as a function of time). Where $N(t)$ is large we have slower coalescence rate and thus longer branches

- Population dynamics
  - Parametric population dynamics inference
    - Under a Wright-Fisher model with a deterministically varying population size $N(t)$, the probability of a sampled tree becomes:

$$P(\mathcal{T} \mid N(t)) = \prod_{i=1}^{n-1} \left(\exp \left[-\int_{t_i}^{t_{i+1}} \binom{k_i}{2} \frac{dt}{N(t)g} \right] \frac{1}{N(t_{i+1})g} \right)$$

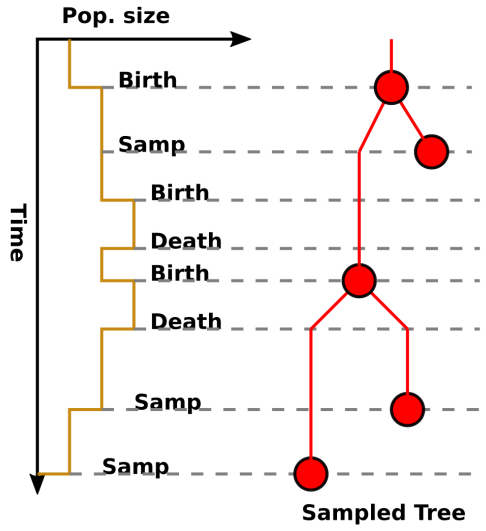
      where t_i is the time at the beginning of interval i

    - For a given parametric form, for example $N(t) = N_0 e^{-\gamma t}$ yields $P(\mathcal{T} \mid N_0, \gamma) = L(N_0, \gamma; \mathcal{T})$ 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



    - Assume population has distinct constant values in each interval between coalescent events
    - Can obtain a separate maximum likelihood estimation for each population size
      - Likelihood of Δt_1 would be $\exp \left[-\Delta t_i \binom{k_i}{2} \frac{1}{N_1 g} \right] \frac{1}{N_1 g}$
    - Resulting population function estimate is the “skyline plot”

- A connection to birth-death models



-Each birth event, there is a possibility of a coalescent event

-It is possible to develop coalescent distributions that approximate the probability density of sampled phylogenies generated by birth-death process

-Assume the approximate ODE (ordinary differential equation) $I(t) = I(T)e^{(\beta-\delta)(T-t)}$ for the linear birth-death process is correct

-Birth events occur at time t with the overall rate of $\beta I(t)$

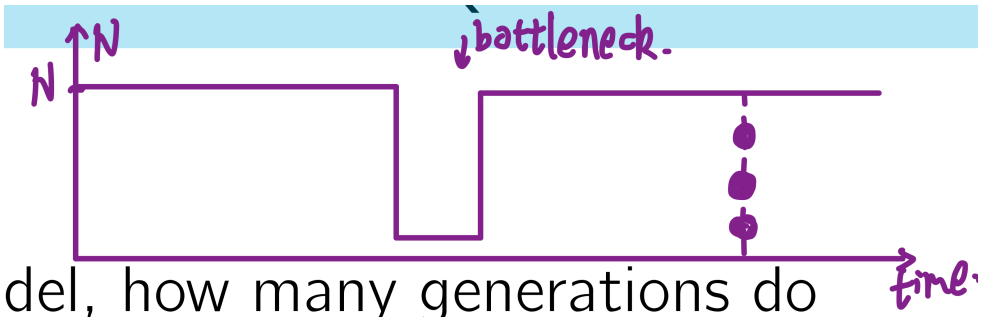
-Every birth is a potential coalescence between sampled lineage

- Probability of choosing a sampled lineage pairs is $\binom{k}{2} / \binom{I(t)}{2}$
- Approximate coalescence rate is $\beta I(t) \frac{k(k-1)}{I(t)(I(t)-1)} \simeq \binom{k}{2} \frac{2\beta}{I(t)}$
- One can use this coalescence rate to compute an approximation to $P(\mathcal{T} \mid \beta, \delta, T)$.
- 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

- Birth-death vs Coalescent models
  - Birth-death
    - Parameters: transmission rate, removal rate, sampling rates/proportions
    - Models sampling process (sample times/locations are data)
    - Advantages
      - Accounts for stochastic variability in population dynamics
      - Generally easier interpretation of parameters
      - Uses information about sampling
    - Disadvantages
      - Sensitive to unmodeled changes in sampling fractions
      - Difficult to extend to complex population models

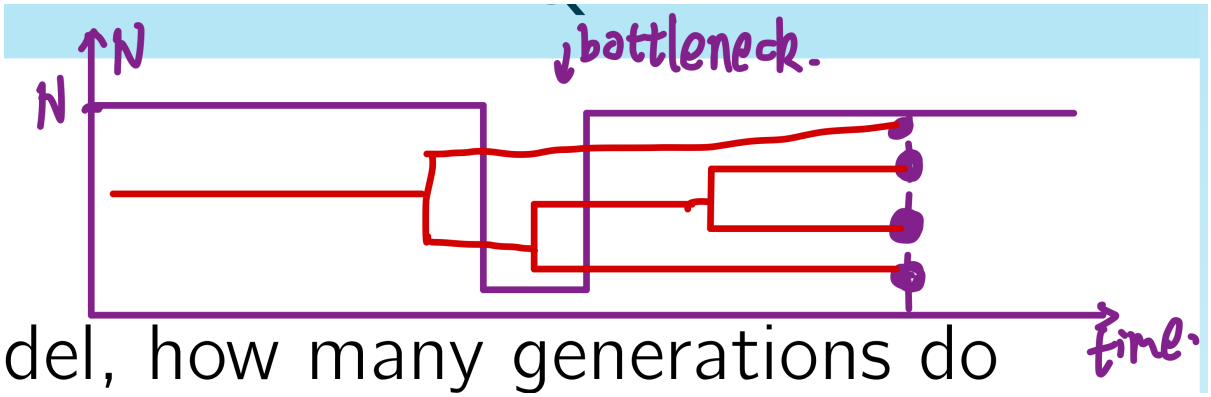
- Coalescent models
  - Parameters: effective population size (NOT actual population size)
  - Assumes number of sampled lineages are small ($k \ll N$)
  - Advantages
    - Generally fast likelihood calculations
    - Easy to extend to complex population dynamics
    - Naturally account for incomplete sampling
  - Disadvantages
    - Sensitive to uncertainty in population dynamics at high sampling
    - Sensitive to hidden population structure and nonrandom sampling

- Questions
  - Under the Wright-Fisher model, how many generations do we have to go back before we find the common ancestor of a pair of genes sampled from a haploid population of size N?
    - A pair of individuals finding the common ancestor in the previous generation is $\frac{1}{N}$
    - A pair of individuals not finding the common ancestor in the previous generation $(1 - \frac{1}{N})$
    - $P(m) = (1 - \frac{1}{N})^{m-1} \cdot \frac{1}{N}$ this is a geometric distribution, so the expectation of m would be $\mathbb{E}(m) = \frac{1}{p} = \frac{1}{\frac{1}{N}} = N$
  - Suppose you had a tree inferred using present-day samples from a population that experienced a severe bottleneck in its recent past. How and why would this bottleneck likely affect our ability to infer ancestral population dynamics?



del, how many generations do

- From the lecture, we know that the probability of a coalescence is in a reverse proportion relationship with the overall population size.
- So, in the bottleneck, it is highly likely that the samples would coalesce somewhere in the bottleneck, leaving no information for the population dynamics earlier than the bottleneck event.



del, how many generations do

- If we can sample through time, it would likely be helpful. Particularly, if we can collect sample before the bottleneck, we can get information about the population size earlier than the bottleneck.

- Imaging spreading a Wright-Fisher population across islands in an archipelago, so that movement between the islands is restricted but within each island the population is “well-mixed”. Qualitatively, how would you expect this population structure to influence estimates of the effective population size?
  - Since the coalescence rate is in a reverse proportion relationship with the population size and separating the population into islands makes the probability of coalescence into lineages within the island pretty easily and coalescence into lineages across island pretty hard, we are expected to end the coalescence much quicker compared to the non-island model. Then this leads to underestimate the population size.

Lecture 11: Bayesian Inference

- General idea of Bayesian Inference in Phylogenetics
  - We could assume a model for evolution of sequences (e.g. Jukes-Cantor) and for population dynamics (e.g. birth-death model)
  - We could assume some starting knowledge of the parameters (substitution rate, birth rate etc.) BEFORE looking at the data
  - We would then obtain and analyse sequencing data leading to a posterior distribution of trees & model parameters
  - If we received more data, we could use the knowledge obtained from the first analysis as the starting point for the analysis of the new data.
  - How can we make this kind of procedure quantitative?
- Probability
  - What do we mean by “probability”?
    - Frequency Interpretation
      - For this interpretation, probabilities are relative frequencies of outcomes of repeatable random experiments
      - For example: consider a dice rolling experiment, let N be the total number of rolls and n_5 be the total number of 5s rolled.
        - The probability of rolling a 5 is $P(d = 5) = \lim_{N \to \infty} \frac{n_5}{N}$
      - Characteristics of this view
        - Probabilities only assignable to outcomes of repeatable experiments (i.e. data)
        - Probabilities treated as an intrinsic property of the system
        - inference of model parameters is treated as a fundamentally distinct problem to the prediction of outcomes.
    - Bayesian Interpretation
      - Bayesian probabilities are the plausibilities of propositions conditional on available information
        - plausibilities of propositions → state of knowledge
      - The probability of a given proposition (e.g. the next roll of a dice will yield 5) depends on the information available.
      - Characteristics of this view
        - Probabilities assignable to any unambiguous proposition

    - Probabilities represent lack of information to predict with complete certainty.
    - Inference of model parameters is treated in the same way as prediction of outcomes.

- Aside 1: A word about notation
    - One way to think about Bayesian probabilities is that they assigned to propositions, i.e. statements that can either be true or false
        - Tim is a cat
        - N = 5 (where N represents some unknown quantity)
    - A statement such as $P(N)$ is therefore as meaningless as $P(Tim)$
    - However, where propositions concern the value of a variable like N, we often use $P(N)$ as shorthand for $P(N = n)$
    - In general, this shorthand is okay, but take care that it does not lead to confusion
- Aside 2: Continuous Variables
    - Propositions regarding continuous variables require special treatment
    - Suppose X may take any real value between 0 and 10
    - The probability $P(X = x)$ will usually be zero!
    - Instead, define

$$f(x) := \lim_{\delta \to 0} \frac{P(X \in [x + x + \delta])}{\delta}$$

    - The function $f(x)$ is a probability density function (PDF) and satisfies the following rules
        - It is normalised: $\int_0^{10} f(x)dx = 1$
        - It is positive: $f(x) \geq 0$
    - However, note that $f(x)$ may exceed 1

- Bayesian inference
  - Example: Inference of genetic distance
    - These two sequences are separated by an unknown genetic distance, d:

```
Sequence 1    A A T C T G T G T G
Sequence 2    A G C C T G G G T A
```

    - The Jukes-Cantor transition probabilities for each site are a function of the random variable d:

$$p_{ij}(d) = \begin{cases} \frac{1}{4} + \frac{3}{4}\exp(-\frac{4}{3}d) & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}\exp(-\frac{4}{3}d) & \text{if } i \neq j \end{cases}$$

    - The number of substitutions is $S = 4$ and the total number of sites is $L = 10$, so the likelihood for the pairwise alignment is

$$P[S \mid d, L] = \left[\frac{1}{4} + \frac{3}{4}\exp(-\frac{4}{3}d)\right]^{L-S} \times \left[\frac{1}{4} - \frac{1}{4}\exp(-\frac{4}{3}d)\right]^{S}$$

    - Our model M has provided us with the probability of the number of segregating sites S given the genetic distance d and the length L of the sequence: $P(S \mid d, L, M)$
    - Our Bayesian interpretation of probabilities means that it is sensible to talk about the probability of d given S and L: $P(d \mid S, L, M)$
    - This distribution quantifies our state of knowledge regarding d once the observed S is taken into account.

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

    - $p(d \mid L, M) = P(d \mid M)$ quantifies knowledge of d in the absence of the observation, while $P(S \mid L, M)$ is the distribution over possible numbers of segregating sites given the JC69 model and any independent knowledge of d.
    - Here we assume that our prior information is only that $0 \leq d \leq 3$, so we take

$$P(d \mid M) = \begin{cases} \frac{1}{3} & \text{for } 0 \leq d \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

      (i.e. a uniform distribution between 0 and 3).

- Bayes theorem
  - In answering this question we have discovered Bayes theorem:

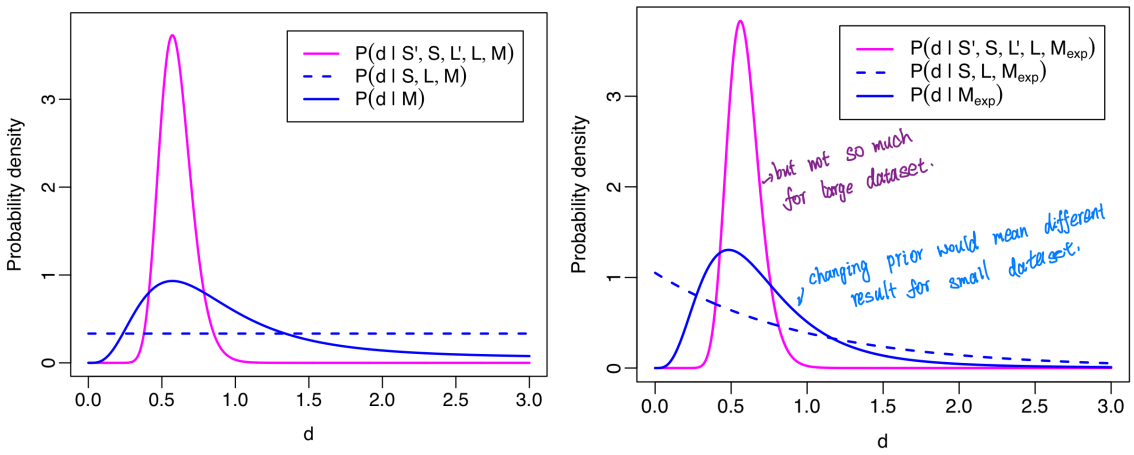
$$P(\theta_M \mid D, M) = \frac{P(D \mid \theta_M, M)P(\theta_M \mid M)}{P(D \mid M)}$$

  - Here θ_M are parameters of some model M, while D are data assumed to be generated by that same model.
  - The components of the theorem have the following names
    - $P(\theta_M \mid M)$ is the prior for the model parameters
    - $P(D \mid \theta_M, M)$ is the likelihood of the parameters given the data
    - $P(D \mid M)$ is the marginal likelihood of (or evidence for) the model
    - $P(\theta_M \mid D, M)$ is the posterior of the model parameters given the model and the data
- Bayesian updating: including more data
  - Suppose we acquired sequence data for an additional 90 sites from the same pair of genomes as the original 10 sites. This new alignments has $L' = 90$ and differs at $S' = 48$ sites.
  - Question: How can we update our estimate for d?
  - Answer: Simply apply Bayes theorem with the posterior of the previous analysis as the prior for the next analysis

$$\begin{aligned} P(d \mid S', S, L', L, M) &= \frac{P(S' \mid d, S, L, L', M) \cdot P(d \mid S, L, L', M)}{P(S' \mid S, L, L\, M)} \\ &= \frac{P(S' \mid d, L', M) \cdot P(d \mid S, L, M)}{P(S' \mid S, L, L\, M)} \\ &= \frac{P(S' \mid d, L', M) \cdot P(S \mid d, L, M) \cdot P(d \mid L, M)}{P(S' \mid S, L, L\, M)P(S \mid L, M)} \\ &= \frac{P(S', S \mid D, L', L, M)P(d \mid M)}{P(S', S \mid L, L', M)} \end{aligned}$$

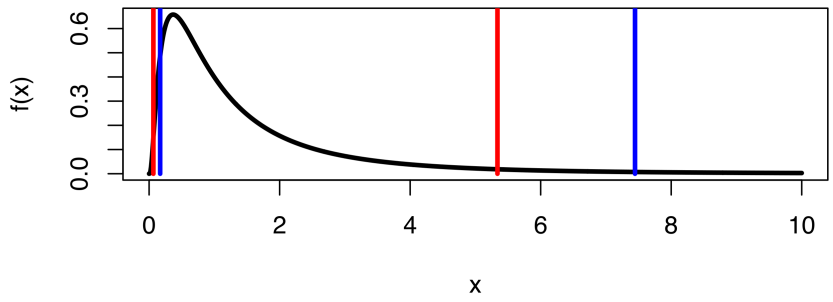
  - Equivalent to inferring d from both datasets simultaneously

- Effect of changing priors



  - Changing prior would mean different result for small dataset
  - But not so much for large dataset

- Prior probabilities
  - What is a prior probability distribution?
    - A prior probability distribution is
      - a probability distribution
        - i.e. a quantification of knowledge
      - Your knowledge about some variable in the absence of the data included in the likelihood term
        - You prior for the analysis of this data may be informed by other data
    - In principle, any two (rational) people/computers with access to precisely the same background information should specify exactly the same prior
    - In practice, truly objective prior selection is difficult to achieve
- Credible Intervals
  - The 95% credible interval is an interval of the posterior distribution containing 95% of the probability
    - One can neglect 2.5% of the samples on both ends of the posterior distribution (blue below)
    - Often, the smallest interval spanned by 95% of the samples is chosen (also called highest posterior density (HPD). red below)



- Highest posterior density (HPD intervals)
  - The 95% HPD interval can also be found by lowering a threshold density until the area under the curve where the density exceeds the threshold is 0.95
  - The meaning of this interval is simply: the probability of the unknown value falling in this region is 95% **given the observed data**
  - This is different to 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 data sets**

- The normalising constant
  - What is so difficult about Bayesian inference?
    - INTEGRATION
    - Bayes' theorem has a troublesome denominator:

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

    - The marginal likelihood $P(D \mid M)$ can be considered a normalising constant for the posterior distribution, and can be expanded as follows

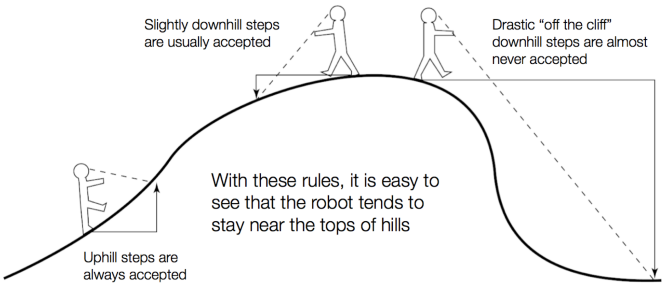
$$P(D \mid M) = \int P(D \mid \theta_M, M)P(\theta_M \mid M)d\theta_M$$

    - Unless you are very lucky, this integral can not be solved with pen and paper
    - If θ_M has many dimensions, you won't even be able to directly integrate this using a computer.
    - This is true for most phylogenetic and phylodynamic problems

- MCMC
  - Monte Carlo methods
    - In our context, Monte Carlo methods are algorithms which produce random samples of values in order to characterise a probability distribution
    - Usually, the algorithms we deal with seek to produce an arbitrary number of independent samples of possible parameter values θ_M drawn from the posterior distribution $P(\theta_M \mid D, M)$
    - Markov Chain Monte Carlo is an example of such an algorithm which is extremely popular in Bayesian phylogenetics and phylodynamics
  - Markov chain Monte Carlo (MCMC) robot
    - The MCMC robot produces a carefully constructed random walk on the domain of the target distribution

- In Bayesian MCMC the target distribution is the posterior distribution $P(\theta_M | D, M)$

$$P(MCMC) = \begin{cases} P(\text{uphill}) = 1 \\ P(\text{sharp downhill}) = \text{low} \\ P(\text{low downhill}) = \text{relatively high} \end{cases}$$



- Let θ_M be the current state. Let θ'_M be a proposed parameter set for the new state
  - For the proposed parameter set, we calculate

$$\alpha = \frac{P(\theta'_M | D, M)}{P(\theta_M | D, M)} = \frac{P(D | \theta'_M, M) P(\theta'_M | M)}{P(D | \theta_M, M) P(\theta_M | M)}$$

  - We draw a uniform number u on $(0,1)$. We accept the proposed step if $u \leq \alpha$
    - if $\alpha > 1$, it means posterior probability $P(\theta'_M | D, M) > P(\theta_M | D, M)$, this means we are climbing uphill, and we would always accept.
      - also $0 \leq u \leq 1$, we can see that $u \leq \alpha$ is always true in this case
    - if $\alpha < 1$, it means $P(\theta'_M | D, M) < P(\theta_M | D, M)$ and that we are going downhill. The smaller α is, the sharper we are going, the less likely this will be accepted.

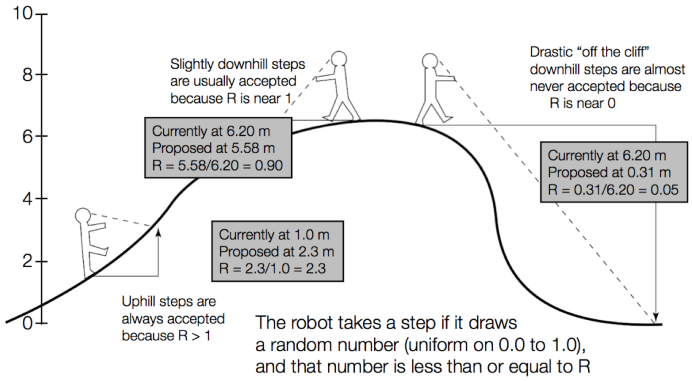


Figure adapted from PO Lewis.

- Metropolis-Hasting 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 θ'_M given θ_M satisfies $q(\theta'_M \mid \theta_M) = q(\theta_M \mid \theta'_M)$
    - symmetric proposal
    - $q(\theta'_M \mid \theta_M) = q(\theta_M \mid \theta'_M)$: the probability that θ'_M is generated given θ_M should be the same as the probability that θ_M is generated given θ'_M
  - The metropolis-Hastings algorithm allows for non-symmetric proposals by using

$$\alpha = \frac{P(\theta'_M \mid D, M)q(\theta_M \mid \theta'_M)}{P(\theta_M \mid D, M)q(\theta'_M \mid \theta_M)}$$

- Pure random walk (P100 on lecture 11)

- Bayesian phylogenetics: The Phylogenetic Likelihood

$$P(A \mid \mathcal{T}, Q)$$

- A is a sequence alignment
- $\mathcal{T}$ is a phylogenetic tree.
- Q is the substitution matrix (and possible other substitution model parameters)
- We are doing Bayesian inference though: we need a probability distribution for $\mathcal{T}$!
- The phylogenetic posterior

$$P(\mathcal{T}, Q, \eta \mid A) = \frac{1}{P(A)} P(A \mid \mathcal{T}, Q)P(\mathcal{T} \mid \eta)P(Q, \eta)$$

  - Here η are the phylodynamic model parameters.
    - $P(\mathcal{T} \mid \eta)$ is the “tree prior” or “phylodynamic likelihood”
    - $P(Q, \eta) = P(Q)P(\eta)$ are the parameter prior distributions
- Questions
  - Is the tree prior really a prior? (Does it depend on data?)
    - It partly depends on the data since it would require a sampling process on the data

  - What is $P(A)$?
    - Marginal likelihood
  - Is $P(A)$ feasible to calculate directly?
    - No. It requires integrate through all models and all tree structures
- Features of Bayesian Phylogenetic Inference
  - Some of the practical characteristics of the Bayesian approach include:
    - The approach jointly infers the phylogenetic tree, the substitution model parameters and the phylodynamic model parameters
    - The approach correctly accounts for uncertainty both in phylogenetic tree itself (due to our stochastic models of sequence evolution) and in the model parameters
    - Additional sources of information are straight-forward to include. (e.g. prior information about parameter values, constraints on tree topology, etc.)
    - Resulting posterior distributions naturally include the uncertainty in the inference results.
- Aside: neutrality assumption
  - Because of the way we have factorised the joint probability for the tree and model parameters, we are implicitly assuming our alignment could have been produced in the following fashion:
    - Sample parameters from prior
    - Sample tree from prior
    - Simulate sequence alignment
  - Separating the process of tree generation from that of sequence evolution implies the sequence evolution is effectively neutral.
- MCMC with Metropolis-Hasting algorithm

$$P(\mathcal{T}, Q, \eta \,|\, A) = \frac{1}{P(A)} P(A \,|\, \mathcal{T}, Q) P(\mathcal{T} \,|\, \eta) P(Q, \eta)$$

  - The MCMC algorithm proposes new state $\mathcal{T}', Q', \eta'$ based on state $\mathcal{T}, Q, \eta$ and evaluates the numerator of Bayes formula.
    - New phylogenetic tree $\mathcal{T}$ is proposed using specialised tree-space proposal distributions.
    - Other parameters are real scalar variables and new states can be proposed via random scaling, uniform random walks, etc
  - Acceptance/rejection of the new state leads eventually to a set of accepted states which is a sample from the posterior distribution $P(\mathcal{T}, Q, \eta \,|\, A)$

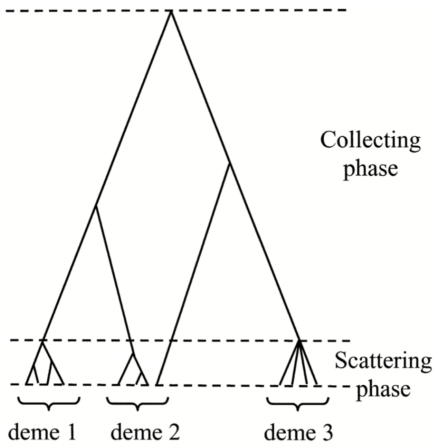
  - Proposal distributions for tree space
    - Require a set of proposal distributions $q_i(\mathcal{T} \mid \mathcal{T})$ where $\mathcal{T}$ is a point in the space of rooted time trees.
    - Some of the popular proposal distributions
      - Wilson-Balding: delete some edge, reattach the subtree of that edge to another branch.
      - Subtree Exchange: delete two edges, and reattach the subtree to the other edges' original branch
      - Uniform node height: modify the height of the some node to its neighbouring node
      - Tree Scaling: Scale the tree up or down.
- Bayesian phylogenetics in practice
  - Based on sequencing data (and potentially fossils), dated trees together with the evolutionary and population dynamic parameters are inferred.
  - We obtain a set of trees and parameters - a sample from the posterior distribution - which naturally allows us to assess the uncertainty (i.e. variance in the parameter estimates).
  - The samples visited by the MCMC algorithm (in principle draws from the posterior distribution) are recorded
    - A chain of length N will result in N trees and N values for each parameter.
- Application: quantifying the spread of Ebola
  - obtained the maximum likelihood estimates $\hat{\beta}$ and $\hat{\delta}$ using phylodynamic likelihood, and estimated $R_0 = \frac{\hat{\beta}}{\hat{\delta}}$
  - However, uncertainty in phylogenetic tree should contribute to uncertainty in this ML estimate of the parameters and R_0
  - Also, ML result does not incorporate prior knowledge of these parameters.
- Summary
  - Bayesian probability distributions quantify states of knowledge
  - Such probabilities apply equally well to data and parameters
  - Bayes theorem provides a natural way to quantitatively combine new data with prior knowledge of model parameters. The result (the posterior distribution for the parameters) provides a comprehensive representation of the final state of knowledge

- If priors do not exclude the truth, then two practitioners with the same likelihood functions will converge on the same inference with increasing amounts of data
- Markov chain Monte Carlo is a simple algorithm that can be employed to study large problems

- Questions:
  - Does Bayesian phylogenetic analysis of the kind described here allow one to directly infer ancestral sequences? Why/Why not?
    - No, Because, $P(\mathcal{T}, Q, \eta \mid A) = \frac{1}{P(A)} P(A \mid \mathcal{T}, Q) P(\mathcal{T} \mid \eta) P(Q, \eta)$, in this equation, the part that has connection with sequence alignment is $P(A \mid \mathcal{T}, Q)$ and when we are estimating the value of this probability using the Felsenstein's pruning algorithm, we have essentially summed over all the possible ancestral sequences and thus we do not mark some specific ancestral sequence.
  - How might we test to see whether a Bayesian MCMC analysis has explored the full state space supported by the posterior?
    - Just looking at the chain alone, in some way, it is impossible to know whether we have converged yet
    - Run multiple chains, 10, 100, distinct MCMC chains and each of them is starting from a random initial state. If we run all of them together and all of the sample distributions start to look the same, then we might have converged.
      - Quantitive way to compare these different chains
    - Or within a single chain, the position that is visited by the robot is close to the position in the previous step and this is called "auto-correlation". This means that if I have 1000 steps in my MCMC chain, the true number of effectively independent samples in the posterior is going to be much smaller. And we check if we have essentially enough effective points
  - Suppose you have conducted a Bayesian phylodynamic analysis and recovered a 95% HPD interval for the birth rate parameter. If you take this result and use it to construct a new prior for this parameter and use this prior to analyse the same data, would the resulting second posterior be valid?
    - No, the problem with this approach is that we are using the same dataset twice. The processor we used in the question is exactly the same as if we get a new dataset whose data is identical to the original dataset. This would lead to the second posterior narrower than the first posterior which means we are more confident about the result. BUT since we do not have another dataset, this is clearly not the case. (We've got no reason to improve the data and we shouldn't improve the data if we do not have more data)

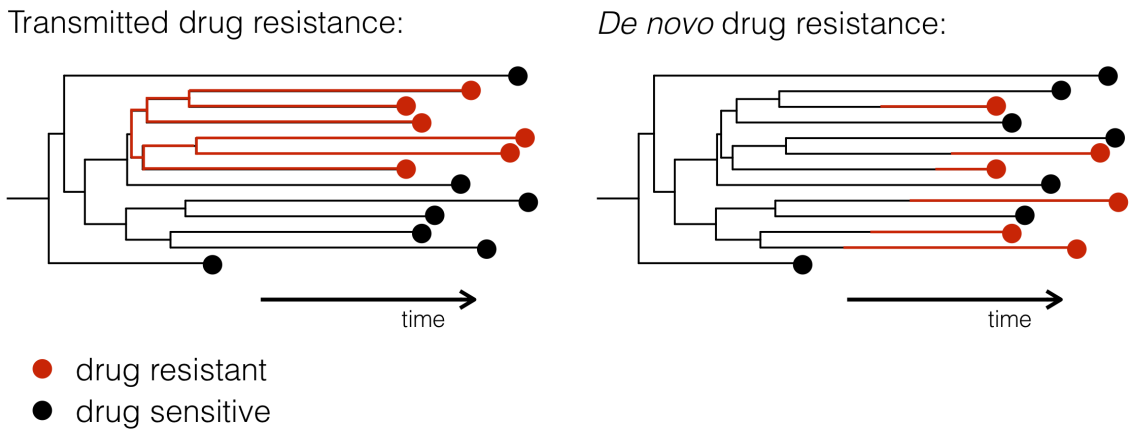
Lecture 12: Structured Population

- Population structures
  - Structured Populations
    - Biological populations often have some internal “structure”
      - fences between cow population etc.
    - What do we really mean by a “structured population”
      - 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).
- Geographic/Spatial structure
  - Gene flow limited by spatial/geographic segregation of subpopulations
  - Impact of structure depends on the rate of migration across boundaries relative to the local birth-rate
  - A population spanning an archipelago(群岛) is a classic example of a spatially structured population
- Non-spatial structuring
  - Even populations that are spatially mixed can be structured in other ways
    - pathogen populations are generally composed of many within-host sub-populations.
    - pathogen sub-populations may possess traits (e.g. drug resistance/ susceptibility) affecting reproductive success
    - infected individuals may be indifferent epidemiological(流行病学) states. (e.g. exposed vs infectious)
    - sampled animals may be members of different species, between which there may be (extremely rare) horizontal gene transfer.
- Structure in phylodynamic analysis
  - Population Structure can play 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 lear about parameters of structured models (e.g. local birth/death rates and sub-population sizes) using structure-aware phylodynamic models.

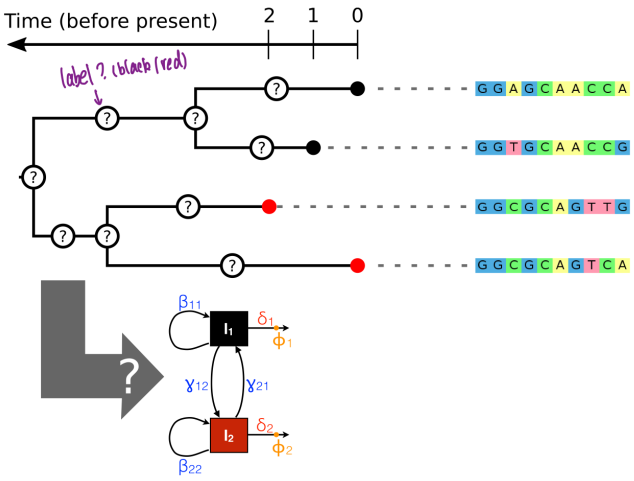


[Pannell, 2003]

- Phylogenetic trees contain information about population structure
  - There are two scenarios that a pathogen population can be formed, one is the transmitted drug resistance and one is the de novo drug resistance

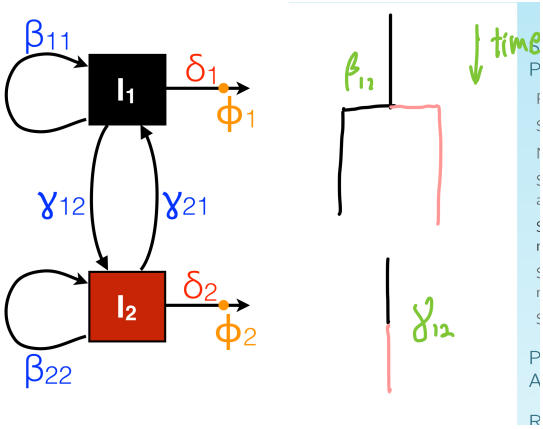


  - Each tip corresponds to one patient. The tree corresponds to the transmission chain.
  - Counting number of sensitive and number of resistant samples through time does not provide insight into which scenario happened (transmitted or de novo drug resistance)
  - However, the phylogeny with tip labels contains information about the scenarios
    - Clustered → Transmitted ; separated → De novo
    - Note that this information is hidden from traditional epidemiological data such as hospital case records.
- Basic structured population. Inference problem



- Tools for birth-death models to include individuals o different types, including
  - Binary State Speciation and Extinction (BiSSE)
  - Multi-State Speciation and Extinction (MuSSE)
  - Multi-type Birth Death (MTBD)

- Structured birth-death models :
  - Simple structured birth-death model



- In this case, β_{11} is transmission (birth) rate of drug sensitive strains, β_{22} is transmission (birth) rate of drug resistant strains, γ_{12} is the rate of resistance evolution.

-In general:

•Different compartments may represent pathogen strains, geographic locations, host risk groups, etc.

•The β_{ij} represent the rate at which individuals of type i **produce** individuals of type j, while γ_{ij} is the rate at which individuals of type i **become** type j

  - Multi-type birth-death phylodynamic likelihood
    - Very similar to the unstructured case
    - Let $p_i(t)$ be the probability that an individual of type i, alive at time t in the past, gives rise to no sampled descendants(used to be $p(0 \mid t)$)

$$P_i(t+\Delta t) = (1 - \Delta t(\sum_j (\beta_{ij} + \gamma_{ij}) + \delta_i)P_i(t) + \Delta t\delta_i + \Delta t(\sum_j \beta_{ij})P_i(t)P_j(t) + \Delta t(\sum_j \gamma_{ij})P_j(t)$$

$$\lim_{\Delta t \to 0} \frac{P_i(t+\Delta t) - P_i(t)}{\Delta t} = \frac{d}{dt}(P_i(t))$$

    - Satisfies the following ODE:

$$\frac{d}{dt}(P_i(t)) = -\left(\sum_{j=1}^{d} (\beta_{ij} + \gamma_{ij}) + \delta_i\right) P_i(t) + \sum_{j=1}^{d} \beta_{ij} P_i(t) P_j(t) + \sum_{j=1}^{d} \gamma_{ij} P_j(t) + \delta_i$$

    - Let $g_i^e(t)$ be the probability that an individual of type i, alive at time t in the past and belonging to sampled tree edge e, gives rise to the sampled phylogeny below that edge. (This obeys a similar ODE, used to be $P(1 \mid t)$)
    - Unlike unstructured case, no known analytical solutions to these equations exist: must be solved numerically to compute structured phylodynamic likelihood.

- Bayesian inference for multi-type birth-death models.
  - The phylogenetic posterior for a multi-type birth-death analysis becomes

$$P(\mathcal{T}, Q, \eta \,|\, A, L) = \frac{P(A \,|\, Q, \mathcal{T}) P(\mathcal{T}, L \,|\, \eta) P(Q) P(\eta)}{P(A, L)}$$

    $\mathcal{T}$ is a phylogenetic tree

    L are the locations/types associated with the sequences

    η are the parameters of the multi-type birth-death model

    $P(\mathcal{T}, L \,|\, \eta)$ is the structured phylodynamic likelihood

    As before, A represents the sequence alignment, Q the substitution matrix
  - Likelihood computed using equations on previous page integrates out (i.e. average over) ancestral states.
  - Can also derive probability of "coloured" tree with ancestral states marked. In that case, MCMC must be performed on this expanded state space of coloured trees.
    - Pro: Posterior distribution for ancestral types directly available
    - Con: Analyses much more computationally demanding
- Unknown Leaf Types
  - Newer methods are available that address the problem of accounting for unknown leaf types/locations
  - This problem is particularly challenging when the **number** of types/locations is also unknown.

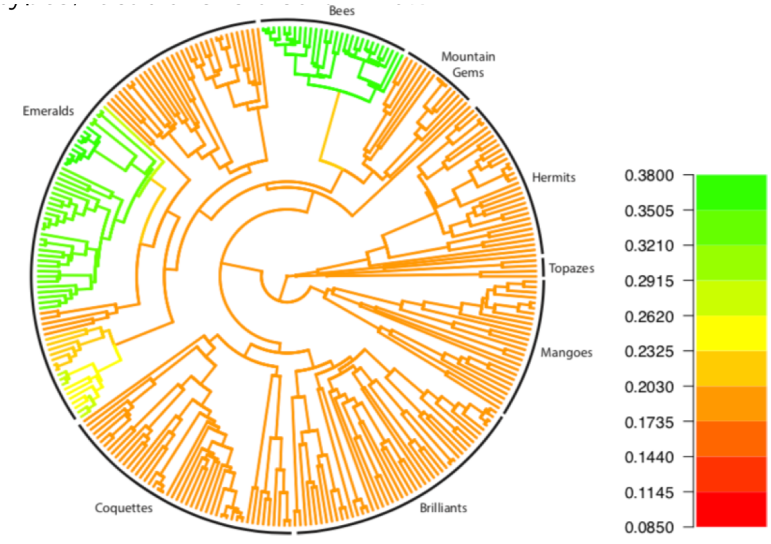
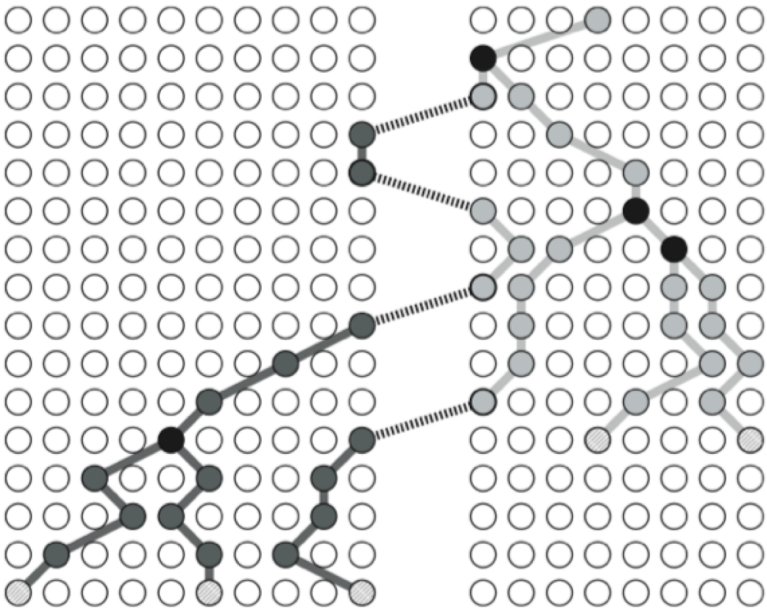
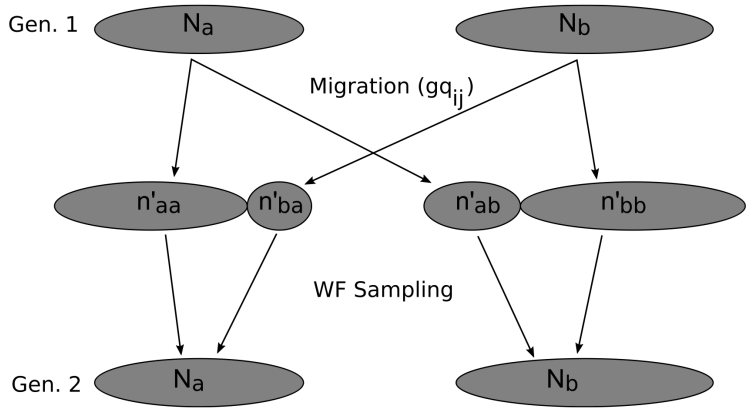


Figure adapted from [Barido-Sottani et al., 2018]
Speciation rate classes across phylogeny of hummingbird species.

- Structured coalescent models
  - Just as for the birth-death models, we can augment coalescent models to include multiple types.
  - Instead of type-specific birth, death and sampling rates however, this yields models with type-specific population sizes.
  - In a similar way to the unstructured coalescent process, the structured coalescent process can be derived from a variety of distinct forward-time models of population dynamics
  - We focus here on an extension to the Wright-Fisher model,
  - Structured Wright-Fisher model

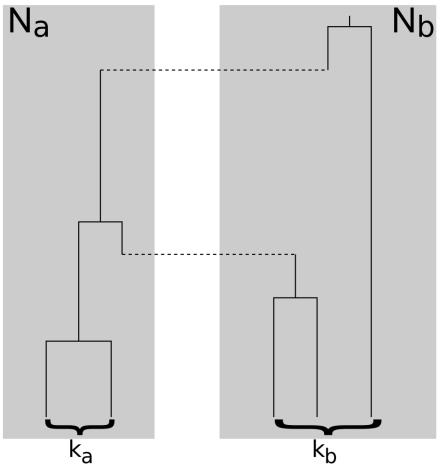


  - Assumes. a single population is divided into sub-populations(demes) of size N_i for $i \in [1,d]$.
  - Allows for migration between demes at rate q_{ij}
  - As in unstructured case, assume a fixed time interval g between successive generations.
  - Assumes that sub-population sizes are unaffected by migration in the long term



- Each generation in the structured WF model includes a migration phase and a sampling phase
  - Migration: Individuals migrate according to probabilities gq_{ij} to form intermediate generation
  - Sampling: Children in next generation sample parents uniformly from intermediate population
- Ensures subpopulation sizes remain fixed, even with asymmetric migration rates between demes
- (Assumption: migration rate is much slower than time necessary for sub-populations to return to equilibrium population size)

- The structured coalescent



-This is backward in time tree generation process

-Corresponding to the coalescent limit of the structured WF model.

-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} (also called the immigration rate) is related to the forward time rate q_{ij} from the structured WF model by

$$m_{ij} = q_{ji}\frac{N_j}{N_i}$$

- Expected coalescent time
  - For a symmetric 2 deme model ($N_1 = N_2 = N$ and $m_{12} = m_{21} = m$) we can derive the *expected* time for two lineages to coalesce:
  - Define T_d and T_s as expected coalescence times when lineages are in different/same demes respectively
  - Lineages in distinct demes cannot coalesce, so we have:

$$T_d = \frac{1}{2m} + T_s$$

    $\frac{1}{2m}$ is the average time for the two lineages (originally from different demes) to be in the same deme. (either deme 1 moved to deme 2 or deme 2 moved to deme 1) and since the migration rate is m, the expected time for either of the process to happen is $\frac{1}{2m}$
  - 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 fields: $T_s = 2Ng$ and $T_d = \frac{1}{2m} + 2Ng$
  - Interesting: Expected time to coalesce from the same deme is independent of migration rate!
- Bayesian inference for structured coalescent models
  - The phylogenetic posterior for a structured coalescent analysis becomes

$$P(\mathcal{T}_{\text{col}}, Q, \theta \mid A, L) = \frac{P(A \mid Q, \mathcal{T}_{\text{col}})P(\mathcal{T}_{\text{col}} \mid L, \theta)P(Q)P(\theta)}{P(A, L)}$$

    $\mathcal{T}_{\text{col}}$ is a phylogenetic tree with ancestral locations marked

    L are the locations/types associated with the sequences

    θ are the parameters of the structured coalescent model

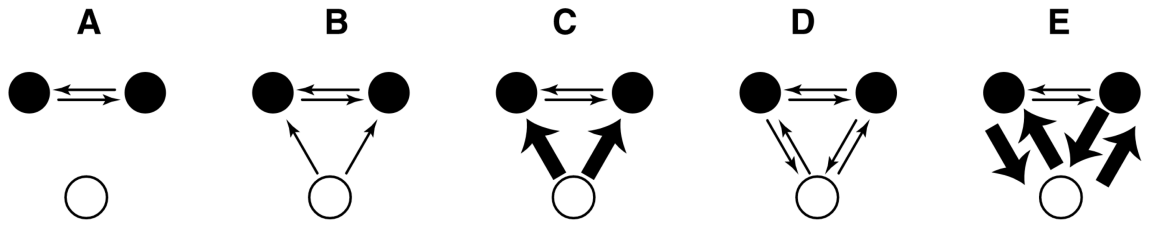
    $P(\mathcal{T}_{\text{col}}, Q, \theta \mid A, L)$ is the structured coalescent likelihood

    As before, A represents the sequence alignment, Q the substitution matrix
  - $P(\mathcal{T}_{\text{col}} \mid L, \theta)$ is a simple extension of the unstructured coalescent expression.

- Unlike the multi-type birth-death case, it is difficult to integrate over ancestral locations.
- Has been done approximately

• Unsampled Types (Ghost Demes)

- It is often the case (particularly for spatial structure) that not all types/ locations are sampled
- Unsampled demes are known as “ghost demes” and ignoring them can lead to overestimates of sub-population size



- Example models with an explicit ghost deme.

- Populations with dynamic structure

• Coalescent models have also been extended to account for **structure** with changes through time.

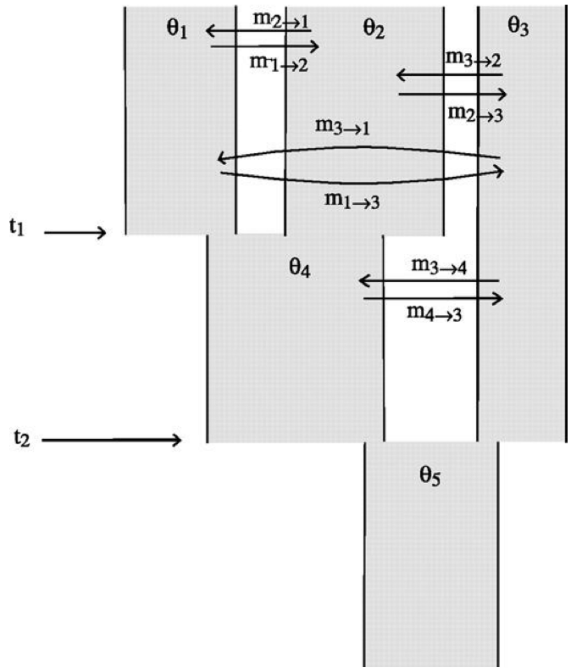


Figure adapted from [Hey, 2009]

Isolation-with-migration model.

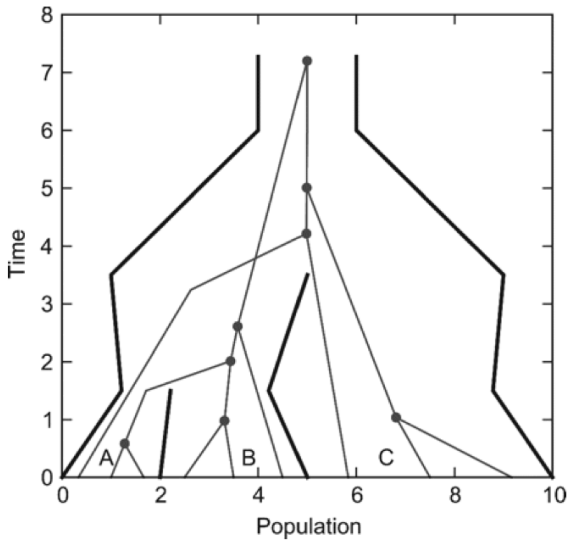


Figure adapted from [Heled and Drummond, 2010]

Multi-species coalescent model.

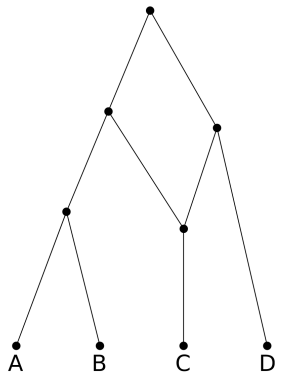
- Phylodynamics in Action
  - Macroevolution
    - individuals = species
    - (Molecular) Evolution
      - Genetic information and morphology of species changes through time
    - Phylogenetics
      - Phylogeny displays species relationship
    - Phylodynamics
      - Population dynamics is the speciation and extinction process
    - Examples: Dinosaurs; penguins
  - Epidemiology
    - individuals = infected hosts
    - Molecular Evolution
      - Genetic information of pathogens changes through time
    - Phylogenetics
      - Phylogeny displays transmission history
    - Phylodynamics
      - Population dynamics is the transmission and recovery process
    - Examples: Ebola, HCV, HIV, Zika
  - Immunology: antibody response
    - individuals = B cells
    - Molecular Evolution
      - B cells change through time due to recombination and somatic hypermutation, as response to pathogen exposure
    - Phylogenetics
      - Phylogeny displays B cell evolution
    - Phylodynamics
      - Population dynamics is the B cell generation and loss process
  - Developmental Biology
    - individuals = cells of a multicellular organism
    - Evolution: cell types change from stem cells to highly specialised cells
    - Phylogenetics: Phylogeny displays differentiation of cells through time.

    - Phylodynamics: Population dynamics is the gain and loss of cell types
  - Human migration
    - individual = human populations
    - Evolution: Human genomes evolve slowly and recombination makes analysis very hard!
      - Solution: Study portions of the genome that do not recombine (Y chromosome + mtDNA)
    - Phylogenetics
      - Phylogeny displays genetic relationships between human populations
    - Phylodynamics
      - Population dynamics is the migration process out of Africa
  - Language Evlution
    - individuals = languages
    - Evolution
      - words and letters change through time
    - Phylogenetics
      - Phylogeny displays language history
    - Phylodynamics
      - Population dynamics is the gain and loss of languages
  - Cultural evolution
    - individuals = human populations
    - Political systems: How does the complexity of systems change over time?
    - Religion: Are social structures correlated with certain religious practices such as ritual human sacrifice?
- Questions
  - Under a structured birth-death model, how do the sub-population size vary (if at all) through time?
    - Because it is an extension from the basic birth-death model, and in the basic birth death model, the population size varies exponentially with time, it is natural that the sub-population size varies exponentially through time
    - This is not the same with the structured coalescent model, it has a built in assumption that the size of the sub-population would not change.

- Suppose you perform a structured coalescent analysis on sequences collected from a relatively *unstructured population*. Would you expect the posterior migration rate to be very low or very high? Why?
  - It depends on how we sampled our data.
    - If we sampled from all the different demes, the migration rate would be very high. Because, if the migration rate is relatively small, we would see a very quick coalescent within each sub-population and then we would have to wait a long time before we can see a coalescent between the populations. If the migration rate is large, it would be meaningless to differentiate different sub-population since it migrates so fast, resulting in a rather unstructured dataset
    - If we sampled the data from only one deme, we would not be so sure about whether the migration rate would be large or small.
- How might the evolution of languages violate the assumptions of a substitution + birth/death phylodynamic model?
  - languages mix and influence each other

Lecture 13: Phylogenetic Networks

- Phylogenetic Networks
  - Just as for trees, phylogenetic networks represent a wide range of evolutionary relationships:
    - For species, the network represents species ancestry and nodes with multiple parents represent hybridisation or horizontal gene transfer (HGT) events
    - For individuals, the network represents ancestry of individual lineages and nodes with multiple parents represent either hybridisation, HGT or simply a node in a pedigree (family tree) if sexually reproducing organism
    - For gene or chromosomes, the network represents ancestry of sequence data and nodes with multiple parents represent recombination events.



  - How many networks are there?
    - Consider the ancestry of 4 species. How many distinct network topologies are there?
      - Recall that there are $(2n-3)!!$ rooted trees with n leaves
      - There are an infinite number of possible ancestral network topologies

- Linkage and phylogeny

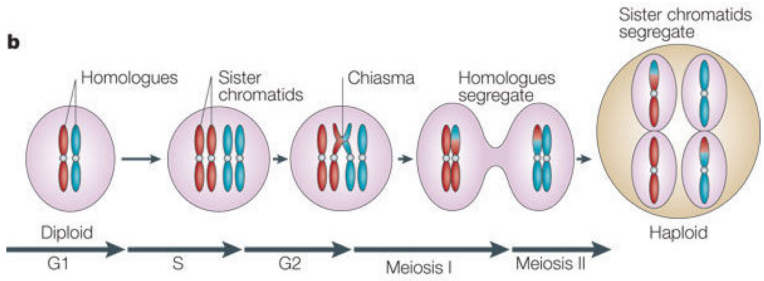
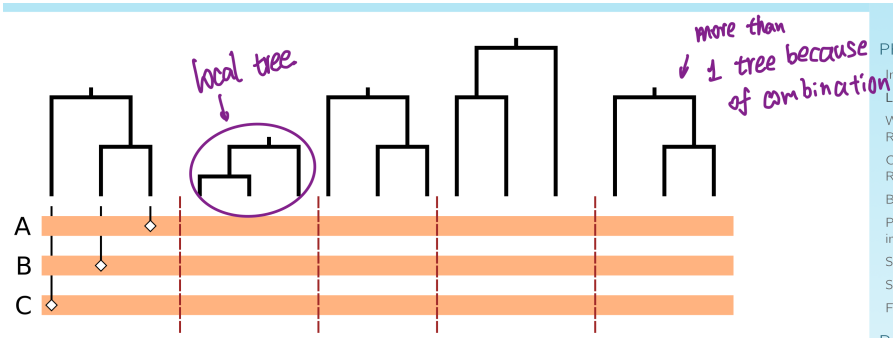


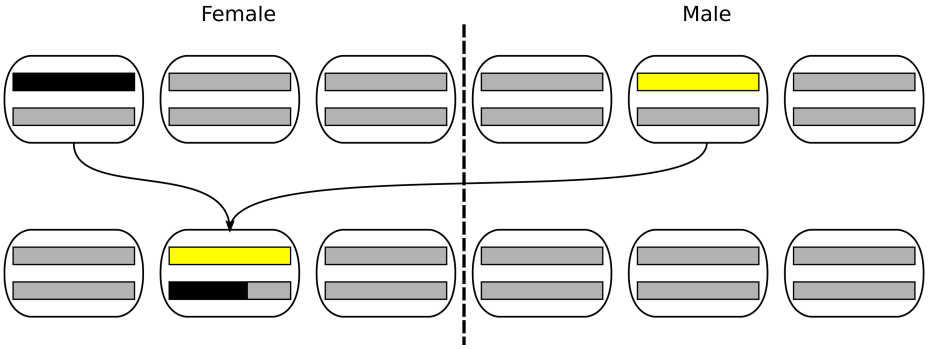
Figure adapted from [Marston and Amon, 2004]

- Sexual reproduction and genetic linkage
  - Recall: 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

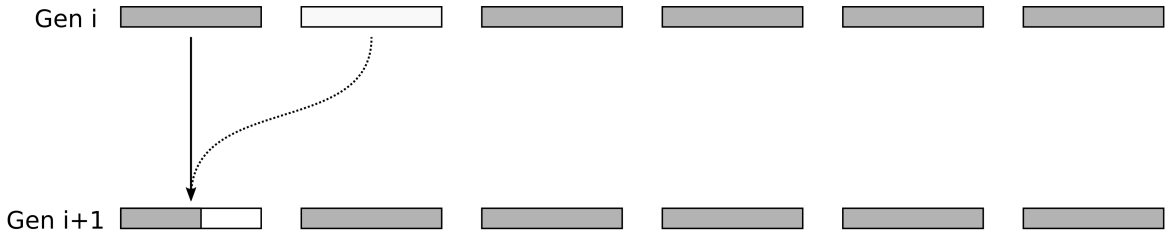


  - Different sites correspond to different trees
  - The further away sites are on the alignment, the more likely they are to possess different ancestry
  - Single nucleotide polymorphisms (SNPs) are usually widely separated and are thus assumed to be completely unlinked - necessary for the validity of GWAS analysis.
  - 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.

- Wright-Fisher with Recombination
  - Consider a Wright-Fisher population with female and male diploid individuals
  - 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

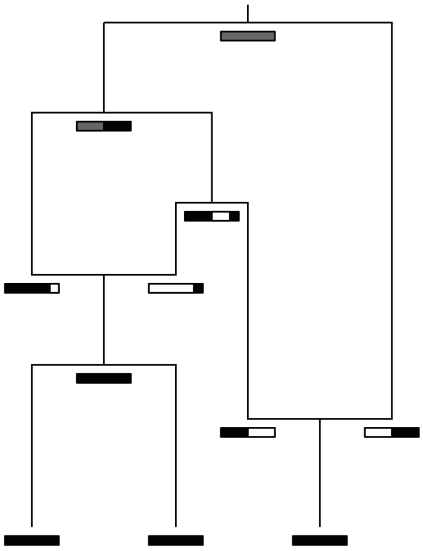


  - Each child selects 1 male and 1 female parent randomly from the previous generation
  - With probability r (which depends on the segment length) the homologous pair from one parents is recombined.
  - Since the specific of chromosomes only matters over a single generation, in the long term the haploid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

- 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
    - Coalescent rate: $\binom{k}{2}\frac{1}{Ng}$
    - Recombination rate: ρk
      - the probability per unit time that a lineage split into two
    - 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 (grey sites) before **grand (G)MRCA** of the process.
    - The result is the “ancestral recombination graph” or ARG.

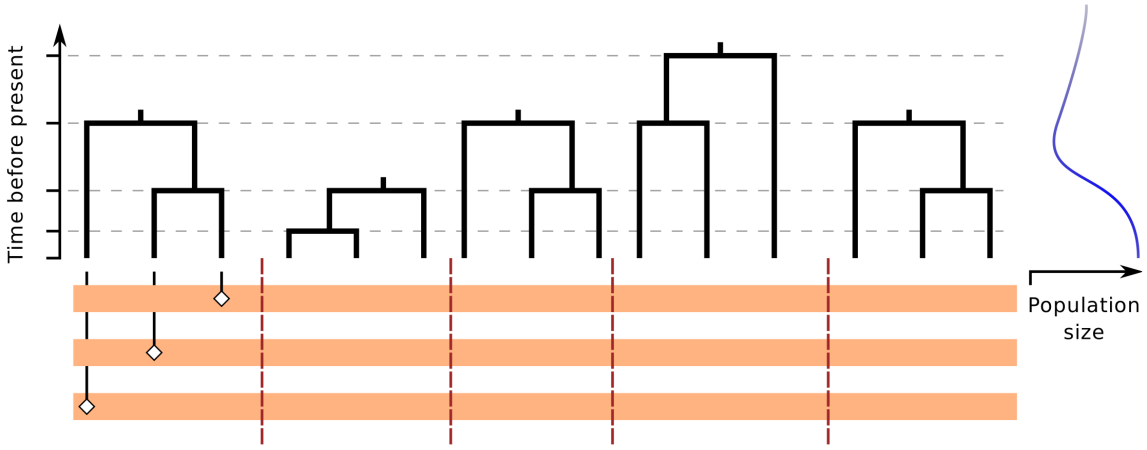


- Bayesian phylogenetic network inference
  - Can easily write down an an expression for posterior distribution for a phylogenetic network given a sequence alignment

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

  - G is the recombination graph/network
  - Q is the substitution rate matrix
  - ρ 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 (these features are “unidentifiable”)
    - the likelihood surface contains many distinct peaks, and
    - the volume of the space of phylogenetic networks with significant posterior probability is usually extremely large.
  - Despite this, many approximate algorithms exist.

- Inference of population dynamics



  - Each local tree contribute additional information to the inference of population size
  - The longer the sequence, assuming the local trees can be accurately inferred, the more powerful the population dynamics inference
- The sequentially Markovian Coalescent
  - produces the local trees along the sequence
  - skip network and go directly to producing local trees along the sequence
  - can be generally a good model for by omitting certain events from the coalescent with three parameters
  - Used greatly on human population dynamics inference
- Inference of human population dynamics
  - Li and Durbin developed an SMC-based hidden Markov model on pairs of alignments
  - Hidden states of HMM are local tree height at each site
  - Used to jointly infer heights and population size dynamics

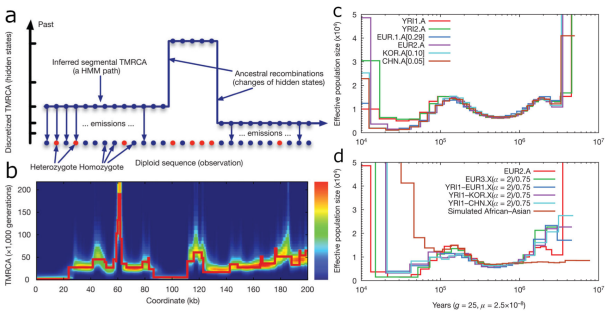
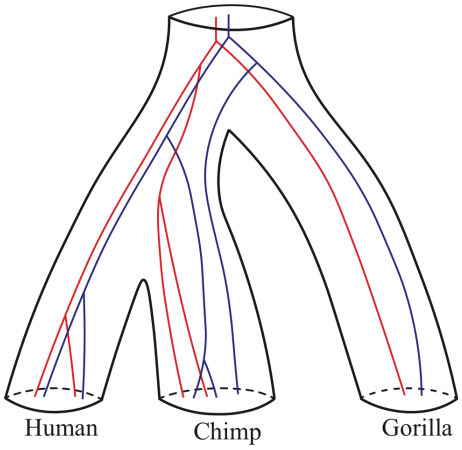


Figure adapted from [Li and Durbin, 2011]

- Species tree models: Multi-species coalescent



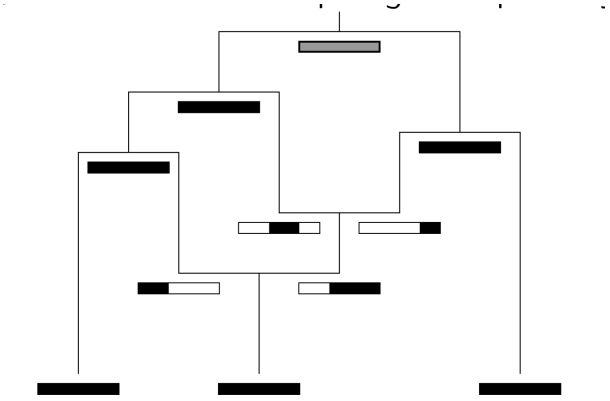
• Each gene has its own gene tree which is embedded within a species tree (or transmission tree)

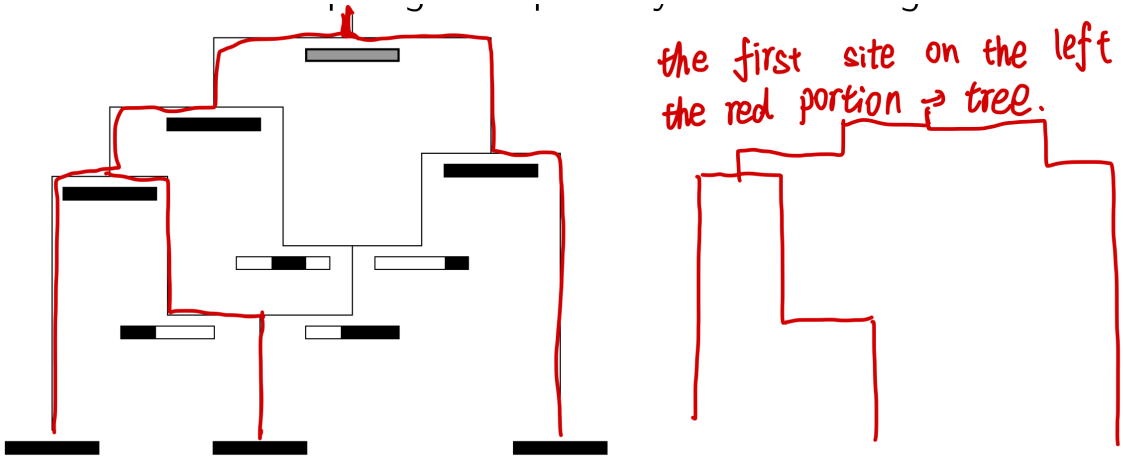
• Gene trees may be different to the species tree as the genes may coalesce any time prior to being in the same population.

• Failure of two homologous genes within the same species to coalesce during the lifetime of the species is called incomplete lineage sorting

•Software: *BEAST, *BEAST2

- Question
  - What is the maximum number of local trees that can correspond to a sequence alignment?
    - the length of the sequence alignment
  - Draw the local topologies implied by the following ARG:





- We saw how recombination can improve our ability to infer ancestral population dynamics. Would you expect higher recombination rates to always improve this? Why/ Why not?
  - If we have an extremely high recombination trees, we would have a lot of local trees. However, we have only one character associated with all these trees, meaning one-side problem, meaning your whole inference is going to be based on one single character and this would be extremely noisy and no phylogeny at all.