Computational Biology

January 2020

Disclaimer
Parts of the information provided within this document may be incomplete and/or incorrect.

1 Models of Molecular Evolution

1.1 Substitution Rate Matrices

1.1.1 JC69

- All substitutions have the same rate λ
- 1 parameter

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

1.1.2 K80

- This model differentiates between transitions (T-C/A-G) and transversions.
- 2 parameters

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

1.1.3 TN93

- Transitions between T/C happen with rate $\alpha_1 \times \pi$
- Transitions between A/G happen with rate $\alpha_2 \times \pi$
- Transversions happen with rate $\beta \times \pi$
- 3 + 3 (π_x) parameters

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

- If $\alpha_1 = \alpha_2$, the model is named **HKY**

1.1.4 GTR - Generalised Time Reversible

+ quite flexible
+ time-reversible
- not completely general

- 6 + 3 (π_x) parameters

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

1.1.5 UNREST

- Unrestricted model
- Each substitution has a different rate

+ Most general model
+ Other models are special cases of UNREST

- Mathematically complicated to handle
- Not time-reversible

- 12 parameters

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

1.2 Calculating Sequence Distance

1.2.1 Transition Probability Matrix

Using the substitution rate matrix Q we derive the transition probability matrix $P(t)$, which gives the probabilities of nucleotide i changing to nucleotide j in any time interval t.

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

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

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

1.2.2 Stationary Distribution

- For $t \to \infty$ we reach a stationary distribution; where the transition probabilities tend towards their equilibrium frequencies.
- Any long sequence will thus be composed of equal amounts of T,C,A and G at $t \to \infty$ under JC69.

1.3 Maximum Likelihood Estimators

Likelihood Function
Describes a hypersurface whose peak represents the combination of model parameter values that maximize the probability of drawing the obtained sample.

Maximum Likelihood Estimator
Is an estimator of a model parameter that maximises the probability to obtain the observed results.

Example: Estimate the probability that a die shows side 6. $\to$ The die is thrown $n = 100$ times and we obtained a 6 $x = 40$ times.

- Define probability of throwing a 6 x times out of n tries $\to$ Binomial-distribution, thus: $P = \binom{n}{x} p^x (1-p)^{n-x}$, where p is the probability of throwing a 6
- We use this probability as our likelihood function and plug in the given values: $L(p;x) = \binom{100}{40} p^{40}(1-p)^{60}$
- To find the maximum likelihood we calculate the first derivative of our likelihood function and set $L' = 0$ to find the maximum.
- Transformations are sometimes applied to the likelihood function. e.g.: $l(p;x) = \log(L(p;x))$
- We estimate the probability by solving for p and get $p = 0.4 \neq 1/6$

1.3.1 Confidence Intervals

Interval which tries to capture the uncertainty of a parameter estimate.

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

Confidence intervals may be calculated on the basis of likelihood intervals.
Let X be a random variable with a distribution parametrised in θ. Based on collected data x of a huge sample, the maximum likelihood estimation for the parameter is $\hat{\theta}$. Then, $2(l(\hat{\theta}) - l(\theta)) \sim \chi^2_k$

- Determine the value of the log likelihood function in $\hat{\theta}$: $l(\hat{\theta};x)$
- Calculate $l(\hat{\theta};x) - 0.5\chi^2_{k,5\%}$; subtract half of the 5% most extreme values according to the χ^2-distribution
- Determine those θ values for which the the following holds: $l(\theta;x) = l(\hat{\theta};x) - 0.5\chi^2_{k,5\%}$

1.3.2 MLE for Sequence Distance

The MLE framework can be used to derive a maximum likelihood estimation for the sequence distance under a JC69 model. We have the transition probability matrix as seen in subsection (1.2.1) with:

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

For two sequences of length n with x differences the probability that any one position is different is $p = 3p_1(t)$. We define $d = 3\lambda t$ as the expected distance in time t and get for the probability that x out of n positions are different:

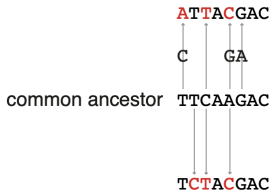
$$L(d;x) = \binom{n}{x} p^x (1-p)^{n-x} =$$
$$\binom{n}{x} \left(\frac{3}{4} - \frac{3}{4}e^{-\frac{4}{3}d}\right)^x \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{n-x}$$

- We compute $l(d;x) = \log(L(d;x))$
- And calculate the first derivative and set it to zero $l'(d;x) = 0$

Which gives us the MLE of the JC69 distance:

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

1.3.3 Example of JC69 MLE



- Length of gene: $n = 8$
- Differences between the two sequences: $x = 2$

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

Determine the 95% confidence interval for the given parameters.

1.4 Variable Substitution Rates

- Not all sites evolve at the same rate
- Mutation rates may vary across sites
- The acting selection in the phenotypic level exerts different evolutionary pressure on different sites
- Extend the existing models by replacing the constant rate by a Γ-distributed random variable

1.5 Codon Substitution Models

N_d : number of nonsynonymous differences
S_d : number of synonymous differences (taking into account all possible ways from seq1 to seq2)
N : number of nonsynonymous sites
S : number of synonymous sites

- Use $\frac{N_d}{N}$ and $\frac{S_d}{S}$ as p-distances
- Calculate d_N and d_S with their respective p-distance with respect to the chosen substitution model
- $\frac{d_N}{d_S} < 1$ implies that nonsynonymous mutations occur less frequently than synonymous mutations (purifying selection)
- $\frac{d_N}{d_S} > 1$ implies that nonsynonymous mutations occur more frequently than synonymous mutations (positive selection)

2 Sequence Distance

t

3 Phylogenetic Inference

There are three phylogenetic approaches

- Phenetic approach
  - Algoritmic approach (e.g. UPGMA algorithm)
  - Optimality approach (e.g. Least squares methods)
- Cladistic approach
  - Parsimony approach
- Mechanistic approach
  - Maximum Likelihood approach

While phenetic approach give statistically consistent result (true tree returned if infinite amount of data available) cladistic approach is not.

3.1 Phenetic approach

3.1.1 UPGMA algorithm

- Output an ultrametric tree (all sequences must be sampled at the same time)
- Assumes evolution according to a strict molecular clock
- The distances from the UPGMA tree might not be exactly the same as is the distance matrix

3.1.2 Least squares method

- Uses an optimality criterion
- Tries to lower the difference between the distance of two sequences from the tree and the distance matrix
- Needs a tree topology to be proposed and then give the optimal branch lengths given the topology

3.2 Cladistic approach

3.2.1 Parsimony method

- Tries to lower the number of mutations
- Considers every possible unrooted tree given sequences alignment and calculate the parsimony score for each of them
- Output the unrooted tree with minimal parsimony score
- Uses Fitch algorithm to speed up the process
- Statistically inconsistent (because of long branch attraction in the Felsenstein zone)

4 Cladistic and ML Inference

MISSING CONTENT PP5-6; Here starts PP7

4.1 Searching Tree Space

To search the tree space for the maximum likelihood tree we need to propose different trees for evaluation.

- We propose different unrooted trees using various defined moves to alter the tree
- We propose trees with different branch lengths; thus we multiply each branch length by some factor

→ We can then use "hill-climbing" strategies to find the ML tree

4.1.1 Modifying Unrooted Trees

- **Nearest-Neighbour Interchange (NNI)**: Swap two subtrees of opposing sides of one branch.
- **Suptree Pruning and Regrafting (SPR)**: Remove one random subtree and attach at a random position of the tree.
- **Branch swapping by tree bisection and reconnection (TBR)**: Cut the tree into two and reconnect at random by selecting branches on both trees and connecting the subtrees between them.

4.2 Model Testing

Here we introduce methods for model selection and assessing the confidence of our parameters.

4.2.1 Likelihood Ratio Testing

- Consider two models: H_0 as a general model parameterised in θ_0 and H_1 as a nested model parameterised in θ_1.
- Derive the likelihood function for both models and the maximum likelihood estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ for given dataset.
- Compute if $2(\log L(\hat{\theta}_1) - \log L(\hat{\theta}_0))$ is in the α tail of $\chi^2_{df'}$ then reject the null model H_0

Likelihood Ratio Test
Is used to assess the goodness of fit of two models based on the ratio of their likelihoods. One model is found by maximizing the likelihood over the whole parameter space while the other is evaluated under some constraints. If the constraint (hypothesis) is supported by the data the likelihoods should not differ significantly.

4.2.2 Testing Non-Nested Models

Akaike Information Criterion (AIC): Used for testing non-nested models:

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

where p_i is the number of parameters and L_i the likelihood function of model i.

- Calculate the AIC for each model
- Choose the model with the lowest AIC; thereby minimizing the Kullback-Leibler distance to the true model

Rules of thumb for multiple model comparisons:

- AIC $\leq$ 1-2 + minimum $\rightarrow$ substantial support, should receive consideration in inference
- AIC $\leq$ 4-7 + minimum $\rightarrow$ low support
- AIC $\geq$ 10 + minimum $\rightarrow$ essentially no support

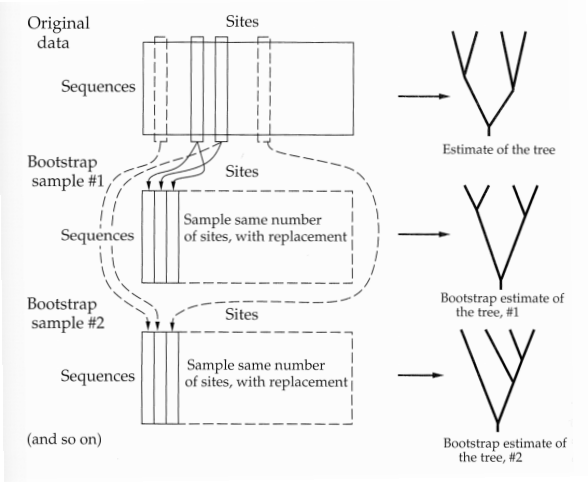
4.2.3 Confidence Intervals

Each parameter value which is not rejected based on the likelihood ratio test at the 0.05 level is within the 95% interval. $\rightarrow$ Use the strategy as introduced before.

- Determine the value of the log likelihood function in $\hat{\theta}$: $l(\hat{\theta}; x)$
- Calculate $l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$; subtract half of the 5% most extreme values according to the χ^2-distribution
- Determine those θ values for which the the following holds: $l(\theta; x) = l(\hat{\theta}; x) - 0.5\chi^2_{k,5\%}$

4.2.4 Bootstrapping

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



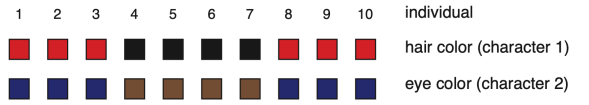
4.3 Overview of ML Inference

1. Infer a ML tree
   - Felsenstein's pruning algorithm for each tree and branch length
   - Choose the tree with branch lengths that optimize the likelihood
   - Do this for each substitution model and calculate its AIC
2. Determine the substitution model and tree with highest support using AIC
3. Determine the confidence interval for the substitution model parameters based on the likelihood ratios
4. Determine the confidence in your maximum likelihood tree using bootstrapping

5 Comparative Methods

5.1 Comparing Discrete Characters

Example: We want to know whether eye and hair color are correlated.



To test whether there is a true correlation we perform Fisher's exact test.
H_0: Having brown eyes is equally likely among red- and black-haired individuals.

hair/eyes	**brown**	**blue**
red	0	6
black	4	0

Evaluating the contingency table above yields the following.

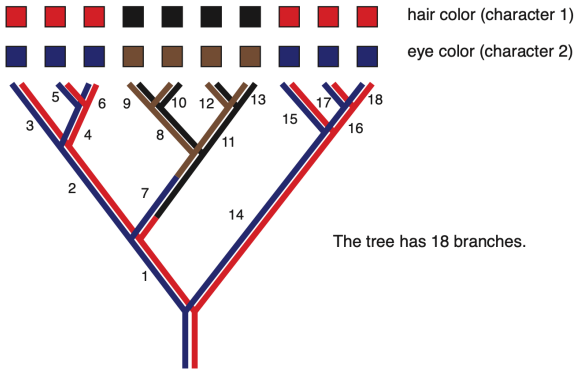
$$P(red/brown) = \frac{(RBr\ in\ R) \times (BlBr\ in\ Bl)}{Br\ in\ All}$$
$$= \frac{\binom{6}{0}\binom{4}{4}}{\binom{10}{4}} = 0.0048 < 0.05$$

Thus we reject the hypothesis of independent character evolution on the 0.05 significance level, indicating that there is a correlation.

Caveat: We should consider that there may be a bias due to the relatedness of the individuals. Fisher's exact test assumes independence which may not be given here.

5.1.1 Reformulated Fisher's Test

Thus we reformulate our problem: **Is the change of characters on the branches correlated?**



$H_{0,new}$: The character changes are equally likely on every branch.

hair/eyes	yes	no
yes	1	0
no	0	17

This contingency table summarizes on how many branches we have a change in either the hair color, the eye color or both. Here we have **one** branch with a change in both and **17** branches with no change.

The probability for one branch on which both the hair and eye color changes (E) is under $H_{0,new}$:

$$P(E|H_{0,new}) = \frac{\binom{1}{1}\binom{17}{0}}{\binom{18}{1}} = 0.05555 > 0.05$$

Neglecting the phylogenetic background can lead to false conclusions on correlations between characters, because of non-independence of species data points as a result of shared ancestry.

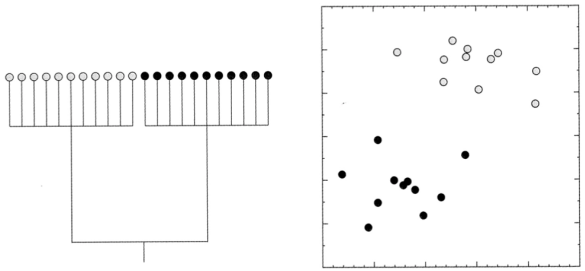
Here we do not consider differences in branch length, but these are important as changes are more likely to happen on longer branches.

5.2 Comparing Continuous Characters

We now switch our focus from discrete characters (e.g. color) to continuous phenotypic characters (e.g. height, weight, virulence).

5.2.1 Linear Regression on Phylogenies

We cannot use linear regression models two compare two characters which evolved on a phylogeny as we cannot distinguish between correlations and clade effects.



When characters evolve on a tree:

- ...they share common evolutionary history and are not independent realisations
- ...the variance/error added by Brownian motion is not equally distributed

Thus the prerequisites for linear regression (see: Linear Regression) are not given.

5.2.2 Brownian Motion

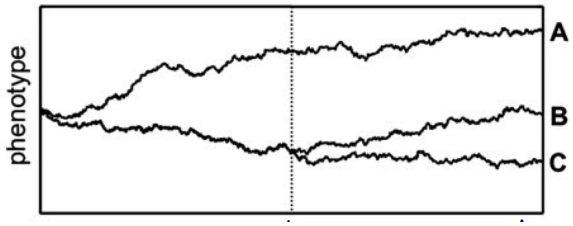
Brownian motion is a Wiener process and thus follows four conditions:

- $W_0 = 0$, the process start in 0
- W_t is almost surely continuous: $P(W_t \text{ continuous }) = 1$
- W_t has independent increments (memorylessness): For $0 \leq s_1 \leq t_1 < s_2 \leq t_2$, $(W_{t_1} - W_{s_1})$ and $(W_{t_2} - W_{s_2})$ are independent
- For $0 \leq s \leq t$, the $W_t - W_s \sim N(0, \sigma^2(t-s))$

There are analogies between models for evolution on discrete and continuous character space.

discrete	continuous
probability to visit any state	probability density on state space
memory-lessness due to Markov-chain model	memory-lessness due to Brownian motion
transition probabilities scale with time	variance scales with branch length

Given a phylogeny we can apply a Brownian motion model to evolve a continuous character.



Linear Regression
Determine dependency of a variable Y in another variable X. We measure Y and X for n independent realizations and fit a regression model to the data. The observations need to be:

- independent
- have the same (normally) distibuted errors

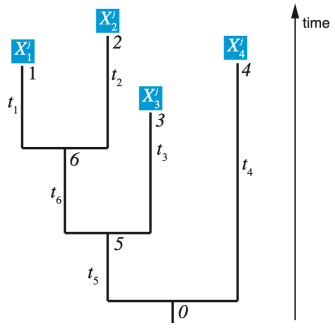
Then we have the model:

$$y_i = \beta x_i + b + \epsilon, \text{ where } \epsilon \sim N(0, \epsilon^2)$$

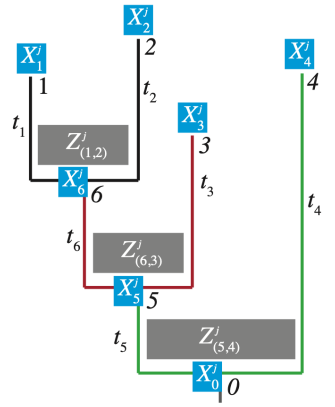
This is fit using a least squares method and the goodness of fit is estimated by R^2. An R^2 of 1 indicates a perfect fit.

5.2.3 Constructing Independent Variables

One method to overcome interdependencies of the evolutionary trait is the contrast method.
Suppose we have the following phylogeny with four species; traits X_1^j and X_2^j are not independent as they share the evolutionary lineages t_6 and t_7.



We now consider the contrasts of the characters instead of the characters. As these are independent.



We now calculate the values of the independent contrasts and their variances.

$$\begin{aligned} z^j_{(1,2)} &= x^j_1 x^j_2 \\ z^j_{(6,3)} &= x^j_6 x^j_3 \\ z^j_{(5,4)} &= x^j_5 x^j_4 \end{aligned}$$

We assume **character evolution according to Brownian motion**. And we consider that we observe tip values but have to estimate values of internal nodes.

In order to calculate the variance we apply the following formula.

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

The branch length at cherries can easily be calculated. The variance is proportional to the branch length between the two external nodes.

$$\begin{aligned} Var[X^j_1] &= \sigma^2(t_1 + t_6 + t_5) \\ Var[X^j_2] &= \sigma^2(t_2 + t_6 + t_5) \end{aligned}$$

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

Contrasts at Internal Nodes

We want to calculate $Z_{(i,l)} = X_i - X_l$ and $Var[Z_{(i,l)}]$ To calculate the values at the internal nodes we have:

$$X_i = \frac{t_n}{t_m + t_n} X_m + \frac{t_m}{t_m + t_n} X_n$$

and the corresponding variances:

$$Var[X_i] = Var[\frac{t_n}{t_m + t_n} X_m + \frac{t_m}{t_m + t_n} X_n] = \sigma^2 \left(\frac{t_m t_n}{t_m + t_n} + t_i + t_k + ... \right)$$

Normalisation of Contrasts To make contrasts comparable to each other they all need to have the same variance. Thus we normalise all contrasts in a last step.

Given the contrast $Z^j_{(i,l)}$ with variance $Var[Z^j_{i,l}] = \sigma^2 c_{z^j_{(i,l)}}$ we know that $Var(\alpha X) = \alpha^2 Var(X)$. Thus we can replace the contrasts by the following.

$$Z^j_{(i,l),norm} = \frac{Z^j_{(i,l)}}{\sqrt{c_{z^j_{(i,l)}}}}$$

With $Z^j_{(i,l),norm} \sim N(0, \sigma^2)$, the contrasts may now be used for linear regression analysis.

6 Phylodynamic Inference and Birth Death Models

Phylodynamic trees encode past macroevolutionary dynamics.
We define the following terms.

- **Molecular Evolution**: Genetic makeup of species changes through time
- **Phylogenetics**: Phylogeny displays the relationship between species
- **Phylodynamics**: Dynamics of population; speciation and extinction processes

In epidemiology we these terms as follows.

- **Evolution**: Pathogen is evolving trough time
- **Phylogenetics**: The phylogeny display the transmission history
- **Phylodynamics**: Transmission and becoming non-infectious

In this context we look at the basic reproductive number R_0, which is the average number of secondary infections caused by a single infected individual.

Applications: Phylogenetic modelling is used outside its traditional realm.

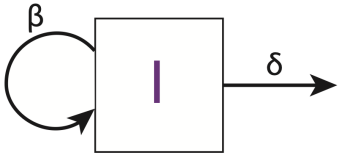
- **Immunology**: B cells are the unit of evolution
  - Phylogeny displays B cell differentiation trough somatic hypermutation
  - Population dynamics are represented by the B cell generation and loss
- **Cancer**: Cells are the unit of evolution
  - Phylogeny displays relationship of different cancer cells and healthy cells
  - Population dynamics is represented by the spread an loss of cell types
- **Languages**: Languages are the unit of evolution
  - Phylogeny displays language evolution
  - Population dynamics is the gain and loss of languages

6.1 Phylodynamics

Population dynamics models the birth and death of individuals (species, infected hosts, B cells, cancer cells and languages). Phylodynamics aims to understand and quantify the population dynamics based on a phylogenetic tree.

6.1.1 Population Dynamic Models

We look at a linear birth-death process which models the reproduction and death of individuals with the following simple model.



- β: The rate of birth of new individuals per individual in I

- δ: The rate of death per individual in I

Thus if we consider the fate of one individual we see that.

- The probability of giving birth to another individual in a very small time step Δt is $\beta \Delta t$.

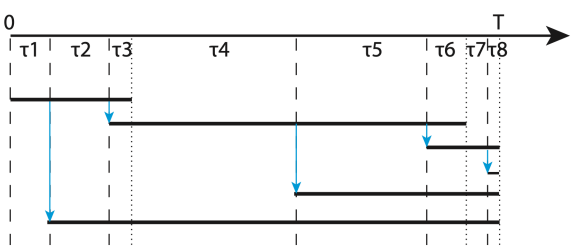
- The probability of dying in a very small time step Δt is $\delta \Delta t$

- The waiting time to the first birth or death event is exponentially distributed with parameter $\beta + \delta$, as the minimum of two exponentially distributed random variables with rates r_1, r_2 is exponentially distributed with rate $(r_1 + r_2)$.

Thus if we consider the waiting time of N individuals we find that.

- The waiting time of the first event is exponentially distributed with parameter $N(\beta + \delta)$

The following diagram illustrates the full population dynamics of a birth-death process which starts with one individual and is stopped after time T.



6.2 Phylodynamic Models

A phylodynamic model adds a sampling process of individuals to the population dynamics. In a simple model we have the following.

- Birth rate β

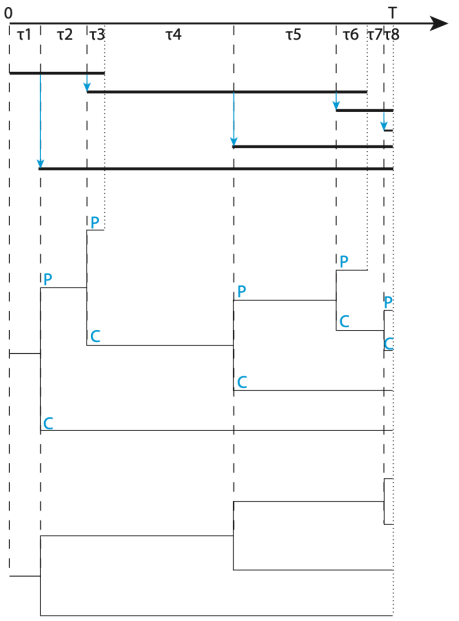
- Death rate δ

- Process duration T

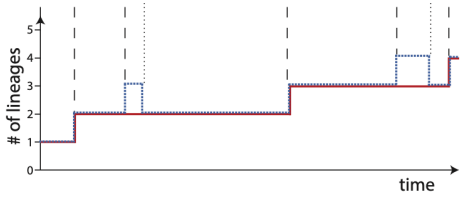
- Extant tip sampling probability ρ

- Extinct tip sampling probability ϕ

If we assume $\rho = 1$, $\phi = 0$ in the context of macro-evolution that means that we do not sample from fossils but only from extant (species still living today) species. The subtree of the complete population tree connecting the sampled individuals and ignoring parent-children labels, is called the phylogenetic tree. This is displayed in the lower part of the following figure, where the parent-child information is removed and only extant species are shown.

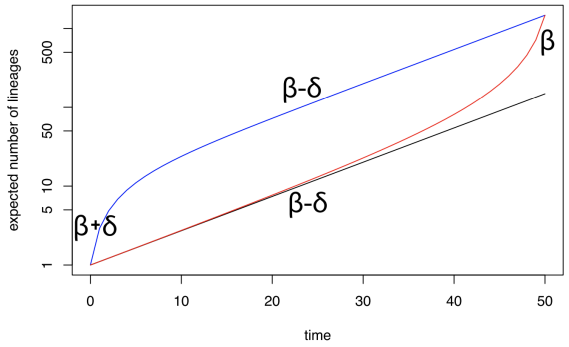


We introduce the lineage through time (LTT) plot here, it indicates the number of surviving lineages trough time. The dashed blue line is the LTT plot of the complete tree and includes population size through time while the red line only shows the number of surviving lineages.



LTT plots provide a method of estimating parameters of birt death models. The following plot

shows a large number of realization of a birth-death model simulation. The simulation was run with $T = 50, \beta > \delta$. If $\beta < \delta$ the population would decrease on average and most trees would rapidly go extinct. We plot the average over all LTT for the phylogenetic trees (red), the average over all LTT plots for the complete trees (blue) and the average total population size over all realizations (complete and extinct trees) (black).

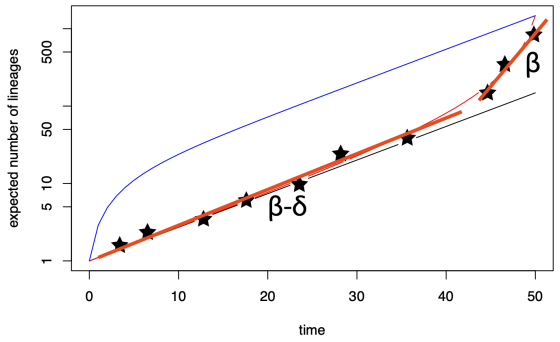


At the start of the complete LTT plot we see an increased slope, this effect is called "push-of-the-past"; the increased slope at the end of the phylogenetic LTT plot is called "pull-of-the-present".

- **Average Total Population Size**: The average total population size through time has the constant slope $\beta - \delta$ on the log scale, corresponding to the average total population size of $e^{(\beta-\delta)t}$ at time t.
- **Complete LTT Plot**: The complete LTT plot goes through a period of accelerated growth at the beginning of the process, before growing exponentially at rate $\beta - \delta$. This may be explained by the fact that the complete LTT plot only includes populations that survive to the present. Thus we expect populations that grow slowly at the start to be less likely to survive to the end of the process and are thus not included in the complete LTT plot (push-to-the-past).
- **Phylogenetic LTT Plot** The phylogenetic LTT plot grows exponentially with rate $\beta - \delta$ until the present when the growth accelerates to exponential growth with rate β. This may be explained by the fact that lineages appearing close to the end of the process have not enough time to go extinct and are thus more likely to be sampled.

6.2.1 Parameter Estimation with LTT

The following figure indicates how the paramters β and δ can be determined from the phylogenetic LTT plot. The stars indicate a branching event at time t with the number of lineages after the event on the y-axis. We fit a line to the initial slope and find $\beta - \delta$ and to the recent slope to find β.



However estimating parameters of the birth-death model in this manner is problematic.

- The variance in the timing of the next branching event (next star) decreases with increasing population size. Thus a classic linear regression assuming the same variance for each data point (homoscedasticity) is not valid.
- The time transition between the two phases of the curve is unclear. This poses the difficulty of deciding where to place the cutoff between the first and second regression line.

6.2.2 Probability Density of a Tree

We recall that in phylogenetics we calculate the phylogenetic tree by evaluating the following likelihood function.

$$L(T, Q; D) = P(D|T, Q)$$

Where T is the phylogenetic tree, Q is the substitution rate matrix and D is the sequence

alignement.

In phylodynamics we want to compute the phylodynamic likelihood which is defined as follows.

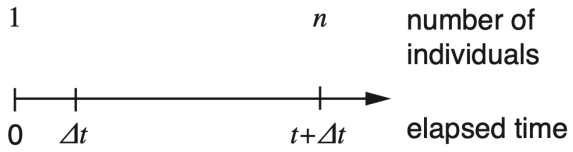
$$L(\eta = (\beta, \delta, T, \rho, \psi, r); T) = P(T|\eta)$$

Where T is the phylogenetic tree and η summarizes the birth-death parameters.

We use a maximum likelihood approach with assumption of complete extant sampling (no fossil sampling), thus we have $\rho = 1$, $\psi = 0$. We start by deriving the likelihood of a single individual to leave 0 or 1 offspring after time t.

The probability that no surviving individuals remain after time t if we start with one individual is abbreviated as $p(0|t)$. We consider very small timesteps (Δt) during which only one event occurs. During this timestep for a single individual a death event happens with probability $\delta \Delta t$ and a birth event happens with probability $\beta \Delta t$. With probability $1 - (\beta + \delta)\Delta t$ no event occurs. We derive the differential equation for $p(0|t)$ below.

$$p(0|t + \Delta t) = \underbrace{(1 - (\beta + \delta)\Delta t)p(0|t)}_{\text{No event}} + \underbrace{\delta \Delta t}_{\text{Death event}} + \underbrace{\beta \Delta t p(0|t)^2}_{\text{Birth event}}$$



The events are derived as follows.

- **No event**: We have probability $1 - P(\text{Any event happens in time } \Delta t)$ that no event occurs in the time step $[0, \Delta t]$, to derive the probability that we have extinction we multiply with the probability of one individual dying out ($p(0|t)$).
- **Birth**: We have the probability $\beta \Delta t$ that a birth happens times the probability that both these individuals die out ($p(0|t)^2$).
- **Death** We have the probability $\delta \Delta t$ for a death event, in this case the probability for extinction is one as no individual remains.

We rearrange the equation and get.

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

When we take the limit $\Delta t \to 0$, the following remains.

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

With the initial condition of $p(0|0) = 0$ we solve the differential equation to get this.

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

We extend the equation to the probability of n surviving lineages after time t and get the following after [Kendall et al., 1948].

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

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

Proof

We prove the equation for $p(1|t)$ and write the following.

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

The factor of two accounts for either one of the descendants of the birth event leading to a surviving individual at time t. We evaluate both sides of the equation using $p(1|t) = e^{-(\beta - \delta)t}(1 - p(0|t))^2$, thus showing that this function is a solution to the differential equation.

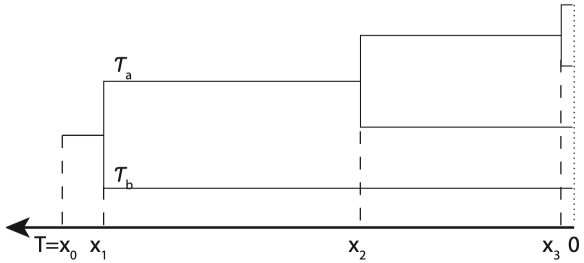
Solving the equation then leads to the following expression.

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

6.2.3 Expansion of the Model to a Tree

We assume that the time in this model is measured as age relative to the present, thus $t = 0$ is the present. We spilt the calculation into multiple sub-problems using dynamic programming. So 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 T with age x_0 is the following.

$$p(T|x_0) = \underbrace{p(x_0, x_1)}_{P(\text{Initial branch})} \underbrace{\beta}_{\text{Branching rate}} \underbrace{p(T_a|x_1)p(T_b|x_1)}_{P(\text{Subtrees})}$$



With $p(T|x) = p(T|\rho = (\beta, \delta, T = x))$

If we calculate the probability $p(t, x_1)$ density of the branch between t and x_1 we get.

$$p(t + \Delta t, x_1) = (1 - (\beta + \delta)\Delta t)p(t, x_1) + 2\beta\Delta t p(t, x_1)p(0|t)$$

This leads to the following differential equation.

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

Which is the same differential equation as for $p(1|t)$, but the initial condition differs. Here we have $p(x_1, x_1) = 1$. Thus for the solution we get.

$$p(x_0, x_1) = p(1|x_0)/p(1|x_1)$$

For a tree of n present day tips, age of the process X_0 and branching times $x_1, x_2, ..., x_{n-1}$ we have the following probability density.

$$p(T|x_0) = p(x_0, x_1)\beta p(T_a|x_1)p(T_b|x_1) = \beta^{n-1}\prod_{i=0}^{n-1} p(1|x_i)$$

An analogous strategy provides us with a tree probability density when $\rho < 1$ (indicating incomplete extant sampling) and $\phi > 0$ (sampling through time).

7 Coalescent Models

While birth-death models allow the population size to vary stochastically, coalescent models instead treat the population size as a given. Thus the population size itself becomes a target of the inference.
A common assumption is that the underlying population dynamics are deterministic.

7.1 Wright-Fisher Process

Is a model for the propagation of traits in a population of fixed size. It includes discrete generations where each generation consists of N individuals.

- Each individual in the offspring population chooses its parent uniformly at random from the N parents.
- Thus a given parent has a binomially-distributed number of offspring.
- For phylogenies of a particular gene, ploidy can be taken into account by multiplying N by a factor which accounts for the number of copies of gene present in each individual. (e.g. for a diploid organism the number of gene copies is $2N$)
- The model assumes distinct non-overlapping generations and each member of a given generation has exactly one parent in the previous generation.
- What these elements represent depends on the system of study; they might be genes or asexual organisms.
- The selection of the parent is completely random; the Wright-Fisher process is therefore neutral as fitness values are not considered.

We want to determine the probability that the most recent common ancestor (MRCA) of two samples occurred at m generations before the present.

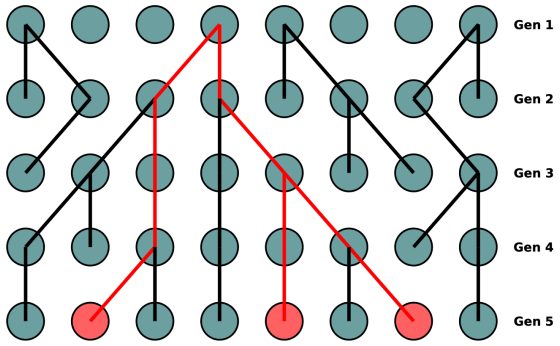
We consider the following.

- Since the parent of each individual is randomly picked, the probability that two individuals in the same generation have the same parent is $\frac{1}{N}$.
- Thus the probability that two individuals in the same generation do not have a common ancestor in the previous generation is $\left(1-\frac{1}{N}\right)$.

The probability that two sampled individuals first share a common ancestor in the m^{th} generation before the present is the following.

$$P_{\text{MRCA}(m)} = \underbrace{\left(1-\frac{1}{N}\right)^{m-1}}_{\substack{\text{No common ancestor in} \\ m-1 \text{ generations}}} \underbrace{\frac{1}{N}}_{\substack{\text{Common ancestor in the} \\ m^{\text{th}} \text{ generation}}}$$

This results in a success probability of $1/N$. Since the mean of such a distribution is the inverse of the success probability we must wait on average N generations to see a common ancestor of two samples from a population of size N.



7.1.1 Coalescent in Calendar Time

If m is the number of generations let g be the calendar time of a generation (e.g. 5 days). Therefore we we have the following for the calendar time span of m generations.

$$\Delta t = gm$$

In calendar time the probability density function for the coalescence of two lineages is the following

$$pdf(N, \Delta t) = \frac{1}{gN} \mathrm{e}^{-\frac{\Delta t}{gN}}$$

For large N the time limit of the coalescence is exponentially distributed with mean gN.

If the number of samples k is much smaller than the population size N we have the following probability for a coalesence between any of the $\binom{k}{2}$ pairs.

$$p_{\text{coal}} \approx \binom{k}{2} \frac{1}{N}$$

7.2 Kingman's Coalescent

Is a continuous-time Markov chain which produces time-trees. The process runs backwards in time by successive merging of events known as "coalesence".

If $N \gg k$ then it will produce the same tree as the Wright-Fisher model.

The times between coalesence events are drawn from an exponential distribution with rate parameters $\binom{k}{2}\frac{1}{Ng}$.

$$P(\Delta t|N, g, k) = \exp\left(-\Delta t \binom{k}{2} \frac{1}{Ng}\right) \binom{k}{2} \frac{1}{Ng}$$

We derive the average time for n lineages to coalesce into one.

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

Thus the mean time until all coalescent events happen is the sum over all probabilities in the tree.

Coming from the previous equation we write the following.

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

We can expand the term $\frac{2}{k(k-1)}$ to $\frac{2}{k-1} - \frac{2}{k}$ and get.

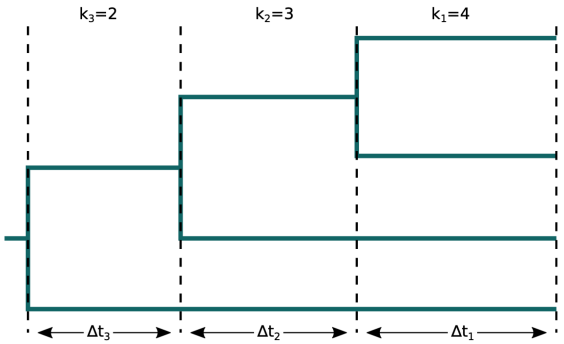
$$\sum_{k=2}^{n} \frac{1}{\binom{k}{2}} = \sum_{k=1}^{n-1} \frac{2}{k} - \sum_{k=2}^{n} \frac{2}{k} = 2(1 - \frac{1}{n})$$

Therefore we find that $E[t_{\text{root}}] \to 2Ng$ as the number of lineages n. (i.e. number of leaves in the coalescent tree) becomes large. This is an upper bound on the expectation, while individual coalescent trees can be older than this.

The Probability of a Coalescent Tree

We calculate the probability of the following coalescent tree.

$$P(T|Ng) = \overbrace{\exp(-\Delta t_1 \binom{4}{2}\frac{1}{Ng})}^{\text{Nothing happens in } \Delta t_i} \times \overbrace{\frac{1}{Ng}}^{\text{Probability of particular coalescent event}} \times$$
$$\exp(-\Delta t_2 \binom{3}{2}\frac{1}{Ng}) \times \frac{1}{Ng} \times$$
$$\exp(-\Delta t_3 \binom{2}{2}\frac{1}{Ng}) \times \frac{1}{Ng} =$$
$$\prod_{i=1}^{n-1} \exp\left(-\Delta t_i \binom{k_i}{2}\frac{1}{Ng}\right)\frac{1}{Ng}$$



- If we take a real tree (inferred from genetic data sampled from a real biological population) and infer population size the result might be biased sine the real dynamics differ from Wright-Fisher dynamics
- Real populations are structured while Wright-Fisher populations are completely homogeneous.
- The coalescent process is often derived as a limit of the Wright-Fisher process as done here and appears as the limit of many other population processes. → Thus the coalescent is believed to be fairly robust.

General Assumption of the Coalescent

1. Samples are members of a population that is at **demographic equilibrium**, which justifies the use of fixed or slowly varying population sizes.
2. **Small sample number** compared to the total population size, which justifies the neglection of more than two lineages coalescing at the same time.
3. Populations are **"well-mixed"**, thus samples are drawn uniformly at random, which justifies the coalescent rate between any pair of sampled lineages being equal. Population structure violates this assumption.

Extension of Population Size Changes

$P(T|N(t))$ is calculated via the rate of coalescence $\frac{1}{N(t)}$, where $N(t)$ is the population size as a function of time t. For large population sizes we have a slower coalescence rate.

Under a Wright-Fisher model with varying population size the probability of a sampled tree becomes.

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

For a given parametric form e.g. $N(t) = N_0 \exp(-\gamma t)$ the model yields the likelihood for the given demographic model parameters. This allows us to directly test different demographic scenarios for a given tree.

Non-parametric Population Dynamics
We assume that the population has distinct constant sizes in each interval between coalescent events. We can obtain a separate maximum likelihood estimate for each population size.

7.3 Coalescent Approximation of Birth-Death Models

We can develop coalescent distributions that approximate the probability density of sampled phylogenies generated by birth-death processes.

We assume that the ODE approximation $I(t) = I(T)\exp((\beta - \delta)(T - t))$ holds. Birth events occur at time t with the rate $\beta I(t)$ where I is the number of individuals. Every birth is a potential coalesence between sampled lineages and the probability of choosing a sampled lineage pair is $\binom{k}{2}/\binom{I(t)}{2}$ and the approximate coalesence rate is $\beta I(t)\frac{k(k-1)}{I(t)(I(t)-1)} \approx \binom{k}{2}\frac{2\beta}{I(t)}$.

The quality of the approximation heavily depends on how well birth-death population dynamics are approximated by the deterministic ODE solution. The approximation can perform very poorly if the population size is small, as it is always the case when an epidemic starts.

7.4 Comparing Birth-death models to Coalescent Models

These models are used to probabilistically relate a populations demography to its phylogenetic history. Both allow for inference of demographic and epidemiological parameters but may differ in their parametrization.

Birth-death Models

Advantages	Disadvantages
Accounts for stochastic variability in population dynamics	Sensitive to unmodeled changes in sampling fractions
Easier interpretation of parameters	Difficult to extend to complex population models
Uses sampling information	

Coalescent Models

Advantages	Disadvantages
Fast likelihood calculations	Sensitive to uncertainty in population dynamics at high sampling
Easy to extend to complex population dynamics	Sensitive to hidden population structure and nonrandom sampling
Accounts naturally for incomplete sampling	

8 Bayesian Inference

There are two main approaches to statistics, the frequentist approach and the Bayesian approach. In a frequentist view we see probabilities as relative frequencies of outcomes of repeatable random experiments (e.g. dice roll). Thus probabilities are only assignable to repeatable experiments and they are treated as an intrinsic property of the system. Furthermore, under this view the inference of model parameters is treated as fundamentally different from the prediction of outcomes.

The Bayesian approach on the other hand sees probabilities as plausibilities of propositions conditional on available information. Under this view probabilities are assignable to any unambiguous proposition and they represent a lack of information to predict the outcome with certainty. Here the inference of model parameters is treated in the same way as the prediction of outcomes.

8.1 Inference of Genetic Distances

We have a Jukes-Cantor model for the likelihood of the alignment with $S = 4$ substitutions given the total number of sites $L = 10$ and the distance d.

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

Thus under our model we can say that the number of segregating sites follows the likelihood $P(S|d,L,M)$. Under a Bayesian viewpoint we can talk about the probability of a certain distance d given the substitutions, the length and the model, thus we have $P(d|S,L,M)$. To find $P(d|S,L,M)$ we use Bayes rule.

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

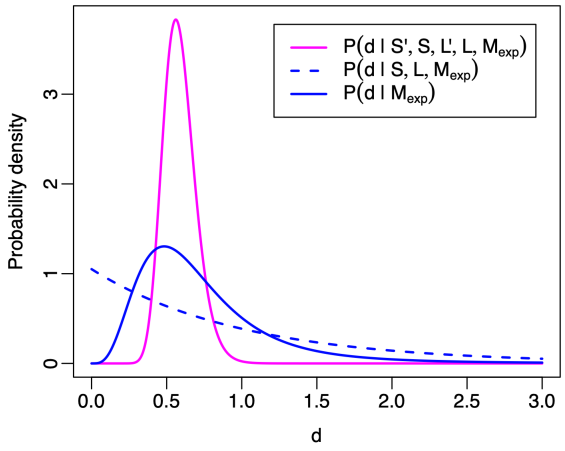
Where $P(d|L,M)$ quantifies knowledge of d in the absence of the observation S. $P(S|L,M)$ is the distribution over possible number of segregating sites S given the JC69 model and any independent d.
Here we assume that our prior information is $0 \leq d \leq 3$.

$$\begin{cases} \frac{1}{3}, & \text{if } 0 \leq d \leq 3 \\ 0, & \text{otherwise} \end{cases}$$

This yields a probability distribution with $L_{max} = 0.7$.

If new data is acquired we can update this estimate using Bayes. We have a new alignment with $(L' = 90, S' = 48)$. Thus we update our estimate by using the previously calculated posterior as our new prior.

This is equivalent to inferring d from both data sets simultaneously.



Credible Intervals
The 95 % credible interval is an interval of the posterior distribution containing 95 % of the probability. We ignore the 2.5 % of the sample on both sides. The interval is often chosen, such that it has the smallest size, this is called the highest posterior density (HPD).

- The 95% HPD can also be found by lowering a threshold density under the curve where the density exceeds the 95 % threshold.
- The interval can be interpreted as **the probability of an unknown value falling into this region is 95 % given the data.**
- This is different from a 95 % **confidence interval** which is the truth-containing interval 95% of the time when averaging over all possible data sets.

8.2 Difficulties of Bayesian Inference

Bayes' Law is given by the following formula.

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

In practice it is often difficult to determine the denominator $P(D|M)$ of this equation, this term can be seen as a normalizing constant for the posterior distribution.

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta$$

This integral is often not numerically solvable if θ has many dimensions, which is true for most phylogenetic and phylodynamic problems.

Thus we use Monte Carlo methods.

- Algorithms which produce random samples of values in order to characterize a probability distribution.
- Markov Chain Monte Carlo is an example of such an approach which is used very often for phylogenetic and phylodynamic problems.

→ See the excellent article on Wikipedia.
`https://en.wikipedia.org/wiki/Metropolis-Hastings_algorithm`

In Bayesian phylogenetics, we take samples from the posterior distribution in order to characterize the probability distribution of the tree T.

The MCMC algorithm relies on the use of appropriate proposal distributions that allow the Markov chain to probe the parameter space efficiently.

Here we want to calculate the probability of a given tree τ, the substitution model parameters Q and η which are the parameters of our phylodynamic model.

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

$P(\tau|\eta)$ is the prior of our tree and $P(Q, \eta) = P(Q)P(\eta)$ are the parameter prior distributions.

This Bayesian approach has the following characteristics.

- Joint inference of the phylogenetic tree, the substitution model parameters and the phylodynamic model parameters.
- Accounts for uncertainty in the tree and in the model parameters.
- It allows us to include additional sources of information such as constraints on the tree topology.
- The resulting posterior distribution naturally includes the uncertainty.

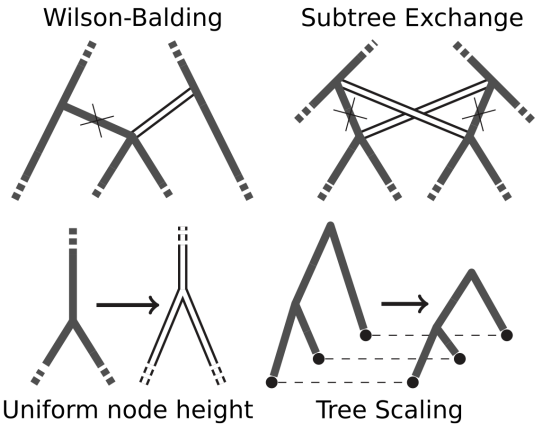
Here we assume that sequence evolution is neutral due to way we factorized the joint probability for the tree and the model parameters (Tree generation is separate from sequence evolution).

The MCMC algorithm proposes new state τ', θ', η' based on state τ', θ', η' and evaluates the numerator of Bayes formula. Here the new tree is proposed using tree-space proposal distributions. The other parameters are scalars and can be proposed via random scaling, random walks, etc.

$\rightarrow$ Acceptance/Rejection of the new state leads to a set of the accepted states which is s sample from the posterior distribution $P(\tau, Q, \eta|D)$.

8.2.1 Tree Space Proposal Distributions

We want get a proposal distribution from the whole space of rooted time trees. Thus to get the proposal distribution $q_i(\tau'|\tau)$ we generate new random trees using the following methods.



- **Wilson-Balding**: A branch swapping move proposed by WILSON and BALDING 1998 which involves removing a subtree and reattaching it on a new parent branch
- **Subtree Exchange**: Two subtrees are randomly swapped.
- **Uniform Node Height**: Randomly selects true internal tree node (i.e. not the root) and move node height uniformly in interval restricted by the nodes parent and children.
- **Tree Scaling**: Scale the branch length of a tree proportionately. CONFIRM

9 Phylodynamic Applications

9.1 Structured Populations

Populations often have some internal structure such as geographical separation between parts of the population.

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

Spatial/Geographic Structuring

Gene flow between subpopulations can be limited due to geographic separation. How strongly this impacts phylodynamic parameters depends on the rate of migration (relative to the local birth rate) between subpopulations.

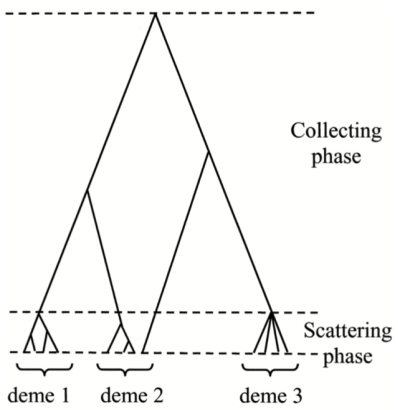
Non-spatial Structuring

Spatially mixed populations can be mixed, this is often the case in pathogen populations which are composed of within-host sub-populations. Some sub-populations may show traits such as drug resistance which have a strong influence on the reproductive success.

Individuals of the population may be in different epidemiological states (e.g. exposed vs infectious).

Sampled animals may even be members of different species, between which there has been horizontal gene transfer, although this is very rare.

Population structure can be an important factor in shaping the phylogenetic relationships between samples.



→ Here the coalesence rate within the demes (a local group of individuals, i.e. from the same taxon, that interbreed with each other and share a gene pool) is much higher than in between the demes.

If one fails to account for such structure there is a risk that the results will be biased. We can use structure aware phylodynamic models to account for such biases.

Incorporating structure into phylodynamic models also allows us to directly address questions relating to population structure. We can consider questions regarding the migration rate between islands or find information regarding the sizes of sub-populations. Even questions about when a disease entered a geographic location can be answered.

Example

Consider a population of individuals of which some are sensitive to some drug and others are not. When sampling from the population we do not know about the history of the lineages but only about their current resistance status. Thus if we reconstruct the tree, the history of how the resistance was transmitted will be missing. However if we find many clustered leaves then it is very likely that the drug resistance was transmitted; if the resistance occurred de-novo many times then drug-resistant and drug-sensitive tips will be mixed.

9.2 Structured Birth-Death Models

We extend the birth-death model to include individuals of different types and transitions between them.

10 Phylogenetic Networks