Maximum Likelihood

Greg Ewing

CIBIV

Outline

Outline

1. Introduction
   - Markov Process
2. The Likelihood
   - The Rate Matrix
   - Rates and Probabilities

Outline

1. Introduction
   - Markov Process
2. The Likelihood
   - The Rate Matrix
   - Rates and Probabilities
3. Optimisation
   - Local Maxima

Outline

Outline

Outline

Stochastic Models

Mathematical Model

A mathematical description of the process of interest, usually describing how things change over time.

- Mathematically define how things change over time.

Stochastic Models

Mathematical Model

A mathematical description of the process of interest, usually describing how things change over time.

- Mathematically define how things change over time.
- So if we have a given state, we can predict what will happen next how the system will behave.

Stochastic Models

Mathematical Model

A mathematical description of the process of interest, usually describing how things change over time.

- Mathematically define how things change over time.
- So if we have a given state, we can predict what will happen next how the system will behave.
- Sometimes we can only predict the probability that something will happen at some time in the future.

Stochastic Models

Mathematical Model

A mathematical description of the process of interest, usually describing how things change over time.

- Mathematically define how things change over time.
- So if we have a given state, we can predict what will happen next how the system will behave.
- Sometimes we can only predict the probability that something will happen at some time in the future.
- This is a stochastic model.

Stochastic Models

Mathematical Model

A mathematical description of the process of interest, usually describing how things change over time.

- Mathematically define how things change over time.
- So if we have a given state, we can predict what will happen next how the system will behave.
- Sometimes we can only predict the probability that something will happen at some time in the future.
- This is a stochastic model.
- Allows a more rigorous mathematical treatment of the problem of tree reconstruction.

Introduction: ML on Coin Tossing

Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right)$

Introduction: ML on Coin Tossing

Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right)$

We take out one coin and toss it 20 times:

$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

Introduction: ML on Coin Tossing

Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right)$

We take out one coin and toss it 20 times:

$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

Probability

$p(k \text{ heads in } n \text{ tosses}|\theta)$

Introduction: ML on Coin Tossing

Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right)$

We take out one coin and toss it 20 times:

$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

Probability		**Likelihood**
$p(k$ heads in n tosses$\|\theta)$	$\equiv$	$L(\theta\|k$ heads in n tosses$)$

Introduction: ML on Coin Tossing

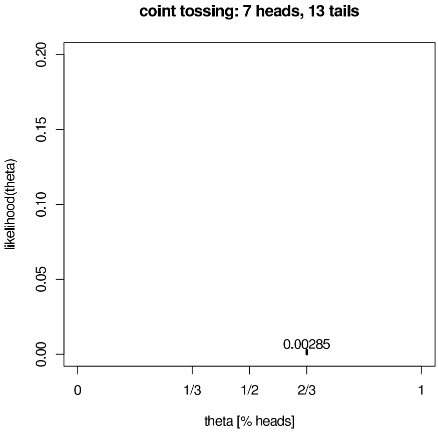
Given a box with 3 coins of different fairness $\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right)$

We take out one coin and toss it 20 times:

$$H, T, T, H, H, T, T, T, T, H, T, T, H, T, H, T, T, H, T, T$$

Probability		**Likelihood**
$p(k$ heads in n tosses$\vert\theta)$	$\equiv$	$L(\theta\vert k$ heads in n tosses$)$
	$=$	$\binom{n}{k}\theta^k(1-\theta)^{n-k}$
		(The binomial distribution)

Introduction: ML on Coin Tossing (Estimate)

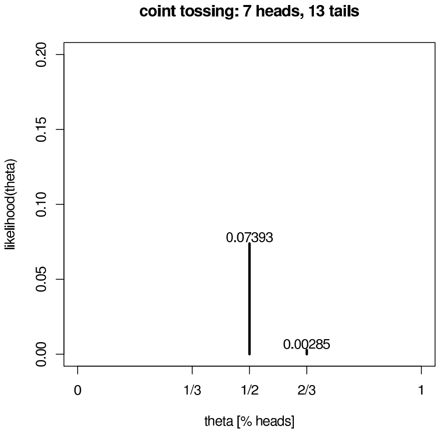


Three coin case

$$L(\theta|7 \text{ in } 20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

Introduction: ML on Coin Tossing (Estimate)

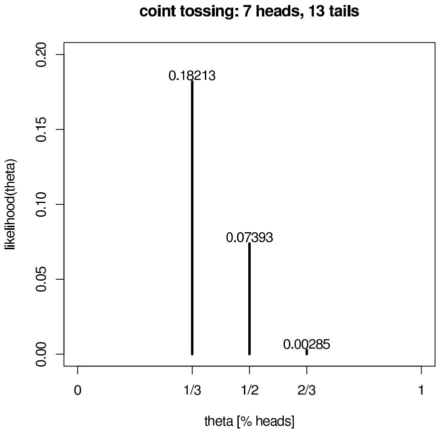


Three coin case

$$L(\theta|7 \text{ in } 20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

Introduction: ML on Coin Tossing (Estimate)

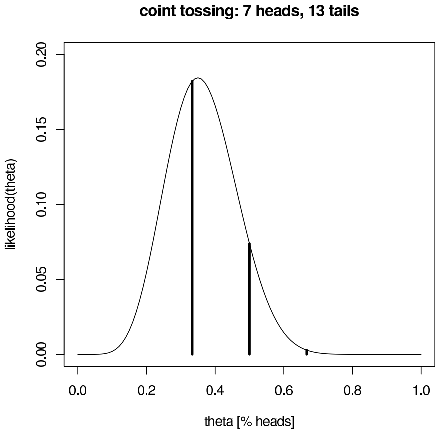


Three coin case

$$L(\theta|7\text{ in }20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

Introduction: ML on Coin Tossing (Estimate)



Three coin case

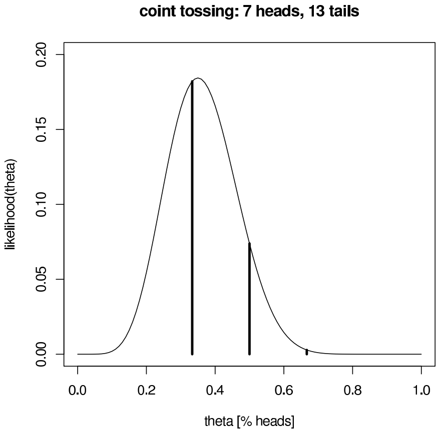
$$L(\theta|7 \text{ in } 20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

For infinitely many coins

$\theta = (0...1)$

Introduction: ML on Coin Tossing (Estimate)



Three coin case

$$L(\theta|7 \text{ in } 20) = \binom{20}{7}\theta^7(1-\theta)^{13}$$

for each coin $\theta \in \left\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\right\}$

For infinitely many coins

$\theta = (0...1)$

ML estimate: $L(\hat{\theta}) = 0.1844$ where coin shows $\hat{\theta} = 0.35$ heads

Coins and Mutations

- Consider 4 coins labelled A, G, T, C.

Coins and Mutations

- Consider 4 coins labelled A, G, T, C.
- At each time step we pick any coin at random and flip it.

Coins and Mutations

- Consider 4 coins labelled A, G, T, C.
- At each time step we pick any coin at random and flip it.
- If a coin comes up heads, we replace it from a random pick of the other coins.

Coins and Mutations

- Consider 4 coins labelled A, G, T, C.
- At each time step we pick any coin at random and flip it.
- If a coin comes up heads, we replace it from a random pick of the other coins.
- Note that the statistics of any column is independent of other columns.

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

- We can extend this to continuous time.

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

- We can extend this to continuous time.
- Each coin can be biased.

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

- We can extend this to continuous time.
- Each coin can be biased.
- Formally a Markov process.

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

- We can extend this to continuous time.
- Each coin can be biased.
- Formally a Markov process.
- Result is that we can calculate a probability of a sequence at some time in the future or past, given the sequence now.

Coins and Mutations

Flip coins

ACACTTTGTGGTGTGGTGGT
ACACATTGTGGTGTGGTGGT
ACACATTGTAGTGTGGTGGT
ACACATTGTAGTTTGGTGGT
ACACATTGTAGTTTGGAGGT

- We can extend this to continuous time.
- Each coin can be biased.
- Formally a Markov process.
- Result is that we can calculate a probability of a sequence at some time in the future or past, given the sequence now.
- Need to get mathematical.

Markov Property

The probability distribution of the next state is completely determined by the previous state.

Markov Property

The probability distribution of the next state is completely determined by the previous state.

As Maths

$$\Pr(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n)$$

Markov Property

The probability distribution of the next state is completely determined by the previous state.

As Maths

$$\Pr(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n)$$

- In the coin example above, the probability of the new sequence is completely determined by the previous state.

Markov Property

The probability distribution of the next state is completely determined by the previous state.

As Maths

$$\Pr(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n)$$

- In the coin example above, the probability of the new sequence is completely determined by the previous state.
- Consider Evolution. The probability of a DNA sequence of the next generation is completely determined by the current generation's DNA sequence.

Markov Property

The probability distribution of the next state is completely determined by the previous state.

As Maths

$$\Pr(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n)$$

- In the coin example above, the probability of the new sequence is completely determined by the previous state.
- Consider Evolution. The probability of a DNA sequence of the next generation is completely determined by the current generation's DNA sequence.
- In other words the process is memoryless.

Markov Property

The probability distribution of the next state is completely determined by the previous state.

As Maths

$$\Pr(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = \Pr(X_{n+1} = x | X_n = x_n)$$

- In the coin example above, the probability of the new sequence is completely determined by the previous state.
- Consider Evolution. The probability of a DNA sequence of the next generation is completely determined by the current generation's DNA sequence.
- In other words the process is memoryless.
- We can therefore use a Markov process to model evolution.

Assumptions

- Ergodic. That is, there is some equilibrium distribution.

Assumptions

- Ergodic. That is, there is some equilibrium distribution.
- Stationary. The base frequencies are in this equilibrium distribution.

Assumptions

- Ergodic. That is, there is some equilibrium distribution.
- Stationary. The base frequencies are in this equilibrium distribution.
- Reversible. The model is the same when time is reversed.

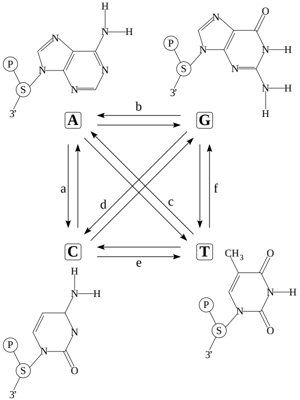
Assumptions

- Ergodic. That is, there is some equilibrium distribution.
- Stationary. The base frequencies are in this equilibrium distribution.
- Reversible. The model is the same when time is reversed.
- Each site in the alignment is independent and identically distributed.

Outline

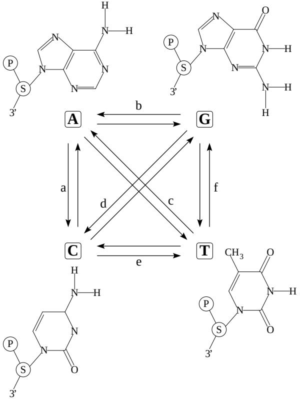
Substitution Models

Evolutionary models are often described using a substitution rate matrix R and character frequencies π. Here, 4×4 matrix for DNA models:



Substitution Models

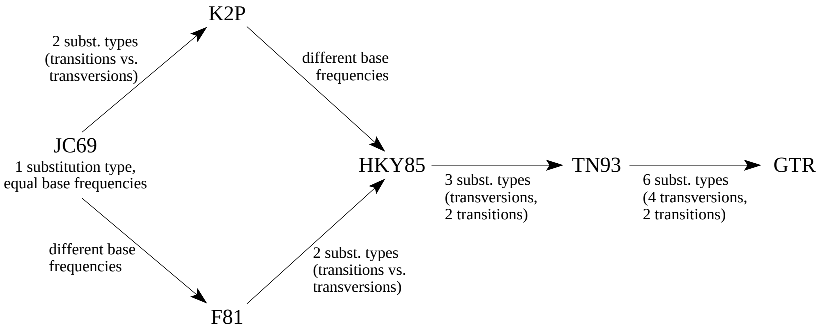
Evolutionary models are often described using a substitution rate matrix R and character frequencies π. Here, 4×4 matrix for DNA models:



$$R = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

$$\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$$

Relations between DNA models



Protein Models

Generally this is the same for protein sequences, but with 20×20 matrices. However unlike DNA the matrix is never optimised. Some protein models are:

- Poisson model ("JC69" for proteins)
- Dayhoff (Dayhoff *et al.*, 1978)
- JTT (Jones *et al.*, 1992)
- mtREV (Adachi & Hasegawa, 1996)
- cpREV (Adachi *et al.*, 2000)
- VT (Müller & Vingron, 2000)
- WAG (Whelan & Goldman, 2000)
- BLOSUM 62 (Henikoff & Henikoff, 1992)

From Substitution rates to probabilities

... R and π are combined into the instantaneous rate matrix Q

$$Q = \begin{pmatrix} \tilde{A} & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \tilde{C} & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \tilde{G} & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \tilde{T} \end{pmatrix} \qquad \begin{array}{l} \tilde{A} = -(a\pi_C + b\pi_G + c\pi_T) \\ \tilde{C} = -(a\pi_A + d\pi_G + e\pi_T) \\ \tilde{G} = -(b\pi_A + d\pi_C + f\pi_T) \\ \tilde{T} = -(c\pi_A + e\pi_C + f\pi_G) \end{array}$$

(where the row sums are zero).

From Substitution rates to probabilities

... R and π are combined into the instantaneous rate matrix Q

$$Q = \begin{pmatrix} \tilde{A} & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \tilde{C} & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \tilde{G} & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \tilde{T} \end{pmatrix} \qquad \begin{array}{l} \tilde{A} = -(a\pi_C + b\pi_G + c\pi_T) \\ \tilde{C} = -(a\pi_A + d\pi_G + e\pi_T) \\ \tilde{G} = -(b\pi_A + d\pi_C + f\pi_T) \\ \tilde{T} = -(c\pi_A + e\pi_C + f\pi_G) \end{array}$$

(where the row sums are zero).
Given now the instantaneous rate matrix Q, we can compute a substitution probability matrix P at time t as

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

. With this matrix P we can compute the probability $P_{ij}(t)$ of a change $i \rightarrow j$ over a time t.

From Substitution rates to probabilities

. . . R and $\boldsymbol{\pi}$ are combined into the instantaneous rate matrix Q

$$Q = \begin{pmatrix} \tilde{A} & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & \tilde{C} & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & \tilde{G} & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & \tilde{T} \end{pmatrix} \qquad \begin{array}{l} \tilde{A} = -(a\pi_C + b\pi_G + c\pi_T) \\ \tilde{C} = -(a\pi_A + d\pi_G + e\pi_T) \\ \tilde{G} = -(b\pi_A + d\pi_C + f\pi_T) \\ \tilde{T} = -(c\pi_A + e\pi_C + f\pi_G) \end{array}$$

(where the row sums are zero).
Given now the instantaneous rate matrix Q, we can compute a substitution probability matrix P at time t as

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

. With this matrix P we can compute the probability $P_{ij}(t)$ of a change $i \rightarrow j$ over a time t.
That is $\Pr(X_t = j | X_0 = j) = P_{ij}(t)$

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.
- First we calculate $P(t) = e^{Qt}$ usually using some eigenvalue decomposition of Qt.

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.
- First we calculate $P(t) = e^{Qt}$ usually using some eigenvalue decomposition of Qt.
- Let s_i be the character at the i'th position, ℓ be the number of characters in s and s'. $P_{ij}(t)$ is the probability that character i changed to character j.

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.
- First we calculate $P(t) = e^{Qt}$ usually using some eigenvalue decomposition of Qt.
- Let s_i be the character at the i'th position, ℓ be the number of characters in s and s'. $P_{ij}(t)$ is the probability that character i changed to character j.

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.
- First we calculate $P(t) = e^{Qt}$ usually using some eigenvalue decomposition of Qt.
- Let s_i be the character at the i'th position, ℓ be the number of characters in s and s'. $P_{ij}(t)$ is the probability that character i changed to character j.

$$P(s'|s,t) = \prod_{i=1}^{\ell} P_{s_i s'_i}(t)$$

Probability of the data

- Start with a sequence $s = \{AGGT\}$ at time 0.
- We can calculate the probability that the sequence changed to $s' = \{ACGA\}$ at t.
- First we calculate $P(t) = e^{Qt}$ usually using some eigenvalue decomposition of Qt.
- Let s_i be the character at the i'th position, ℓ be the number of characters in s and s'. $P_{ij}(t)$ is the probability that character i changed to character j.

$$P(s'|s,t) = \prod_{i=1}^{\ell} P_{s_i s'_i}(t)$$

- Consider finding the value of t where this is maximised.

Computing ML Distances Using $\mathrm{P_{ij}(t)}$

The Likelihood of sequence s evolving to s' in time t:

$$L(t|s \to s') = P(s'|s,t) = \prod_{i=1}^{\ell} P_{s_i s'_i}(t)$$

Computing ML Distances Using $\mathrm{P_{ij}(t)}$

The Likelihood of sequence s evolving to s' in time t:

$$L(t|s \rightarrow s') = P(s'|s,t) = \prod_{i=1}^{\ell} P_{s_i s_i'}(t)$$

Likelihood surface for two sequences under JC69:

```
GATCCTGAGAGAAATAAAC
GGTCCTGACAGAAATAAAC
```

Computing ML Distances Using $\mathrm{P_{ij}(t)}$

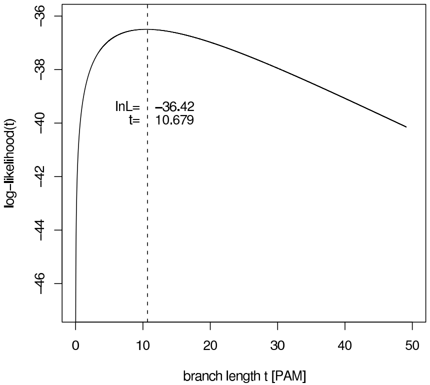
The Likelihood of sequence s evolving to s' in time t:

$$L(t|s \to s') = P(s'|s,t) = \prod_{i=1}^{\ell} P_{s_i s'_i}(t)$$

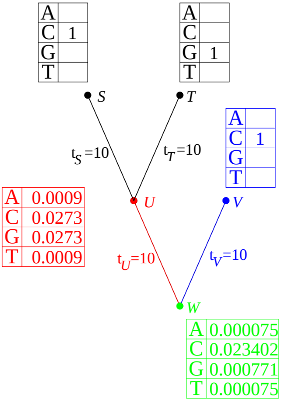
Likelihood surface for two sequences under JC69:

GATCCTGAGAGAAATAAAC
GGTCCTGACAGAAATAAAC

Note: we do not compute the probability of the distance t but that of the data $D = \{s, s'\}$.

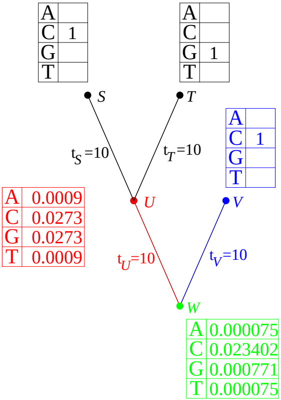


Likelihoods of a Single column tree



Likelihoods of nucleotides at inner nodes:

Likelihoods of a Single column tree

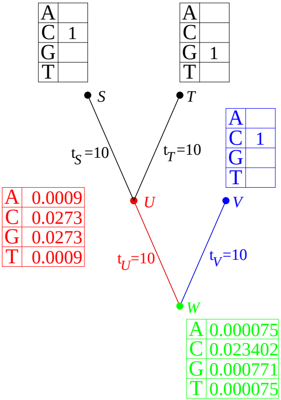


Likelihoods of nucleotides at inner nodes:

$$L_U(i) = [P_{iC}(10) \cdot L(C)] \cdot [P_{iG}(10) \cdot L(G)]$$

$$L_W(i) = \left[\sum_{u \in \Omega} P_{iu}(t_U) \cdot L_U(u)\right] \cdot \left[\sum_{v \in \Omega} P_{iv}(t_V) \cdot L_V(v)\right]$$

Likelihoods of a Single column tree



Likelihoods of nucleotides at inner nodes:

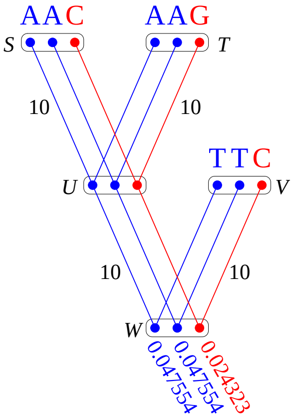
$$L_U(i) = [P_{iC}(10) \cdot L(C)] \cdot [P_{iG}(10) \cdot L(G)]$$

$$L_W(i) = \left[\sum_{u \in \Omega} P_{iu}(t_U) \cdot L_U(u)\right] \cdot \left[\sum_{v \in \Omega} P_{iv}(t_V) \cdot L_V(v)\right]$$

Site-Likelihood of an alignment column k:

$$L^{(k)} = \sum_{i \in \Omega} \boldsymbol{\pi}_i \cdot L_W(i) = 0.024323$$

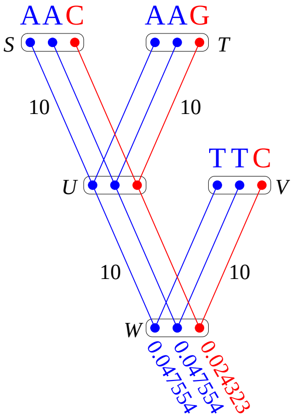
Likelihoods of Trees (multiple columns)



Considering this tree with $n = 3$ sequences of length $\ell = 3$ the tree likelihood of this tree is

$$\mathcal{L}(T) = \prod_{k=1}^{\ell} L^{(k)}$$

Likelihoods of Trees (multiple columns)

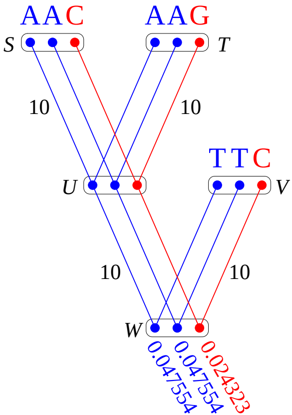


Considering this tree with $n = 3$ sequences of length $\ell = 3$ the tree likelihood of this tree is

$$\mathcal{L}(T) = \prod_{k=1}^{\ell} L^{(k)} = 0.047554^2 \cdot 0.024323$$

$$= 0.000055$$

Likelihoods of Trees (multiple columns)



Considering this tree with $n = 3$ sequences of length $\ell = 3$ the tree likelihood of this tree is

$$\mathcal{L}(T) = \prod_{k=1}^{\ell} L^{(k)} = 0.047554^2 \cdot 0.024323$$
$$= 0.000055$$

or the log-likelihood

$$\ln \mathcal{L}(T) = \sum_{k=1}^{\ell} \ln L^{(k)} = -9.80811$$

Outline

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths
- Starting with a branch, adjust the length calculating the log Likelihood until a maximum is found.

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths
- Starting with a branch, adjust the length calculating the log Likelihood until a maximum is found.
- Do the same to other branches and repeat until no further improvement can be made.

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths
- Starting with a branch, adjust the length calculating the log Likelihood until a maximum is found.
- Do the same to other branches and repeat until no further improvement can be made.
- Model parameters can also be optimised (ie π).

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths
- Starting with a branch, adjust the length calculating the log Likelihood until a maximum is found.
- Do the same to other branches and repeat until no further improvement can be made.
- Model parameters can also be optimised (ie π).
- Note traditional multivariate optimisation can apply.

Optimise branch lengths

To compute optimal branch lengths:

- Initialise the branch lengths
- Starting with a branch, adjust the length calculating the log Likelihood until a maximum is found.
- Do the same to other branches and repeat until no further improvement can be made.
- Model parameters can also be optimised (ie π).
- Note traditional multivariate optimisation can apply.
- Changing the topology is much harder.

Finding the ML Tree

Exhaustive Search

Guarantees to find the optimal tree, because all trees are evaluated, but not feasible for more than 10-12 taxa.

Finding the ML Tree

Exhaustive Search

Guarantees to find the optimal tree, because all trees are evaluated, but not feasible for more than 10-12 taxa.

Branch and Bound

Guarantees to find the optimal tree, without searching certain parts of the tree space – can run on more sequences, but often not for current-day datasets.

Finding the ML Tree

Exhaustive Search

Guarantees to find the optimal tree, because all trees are evaluated, but not feasible for more than 10-12 taxa.

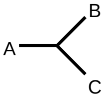
Branch and Bound

Guarantees to find the optimal tree, without searching certain parts of the tree space – can run on more sequences, but often not for current-day datasets.

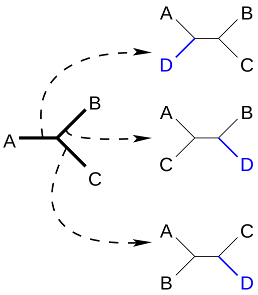
Heuristics

Cannot guarantee to find the optimal tree, but are at least able to analyse large datasets.

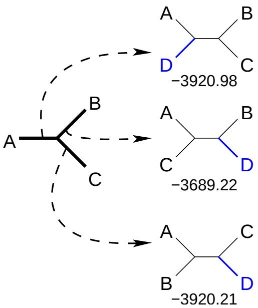
Build up a tree: Stepwise Insertion



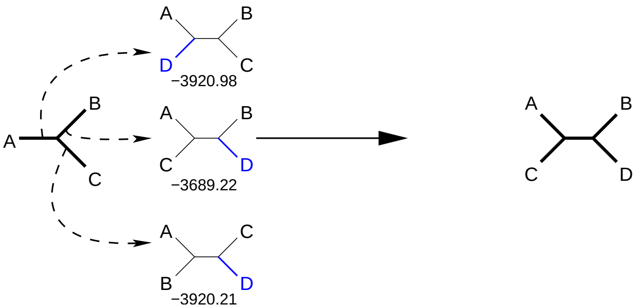
Build up a tree: Stepwise Insertion



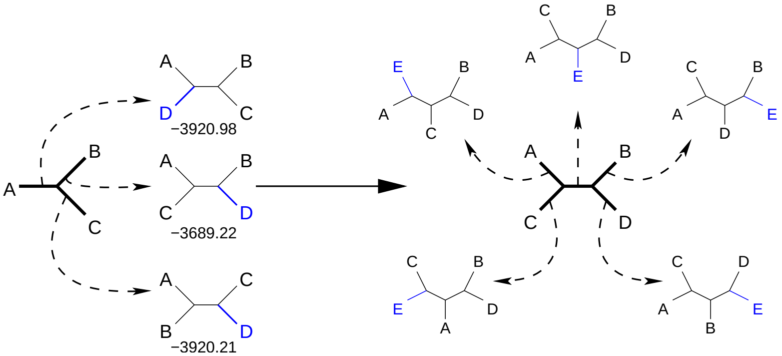
Build up a tree: Stepwise Insertion



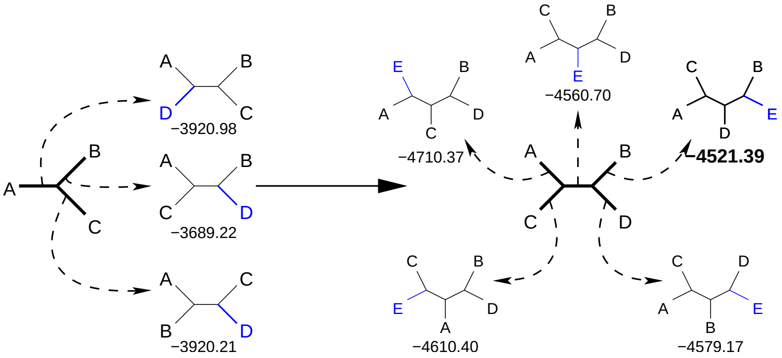
Build up a tree: Stepwise Insertion



Build up a tree: Stepwise Insertion

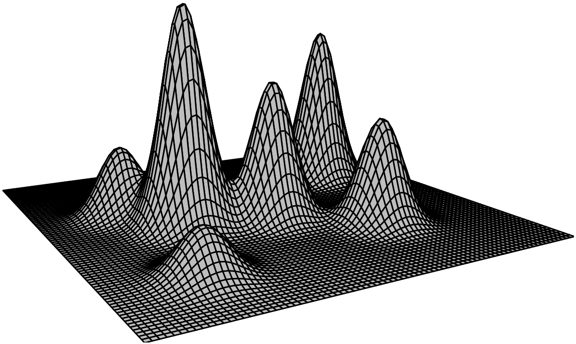


Build up a tree: Stepwise Insertion



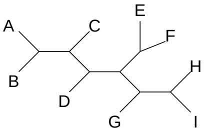
Local Maxima

What if we have multiple maxima in the likelihood surface?

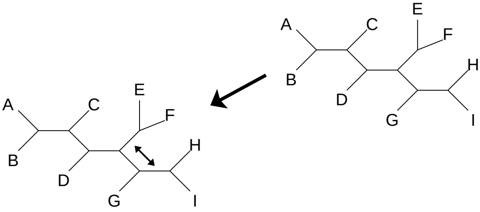


Use Tree rearrangements to escape local maxima.

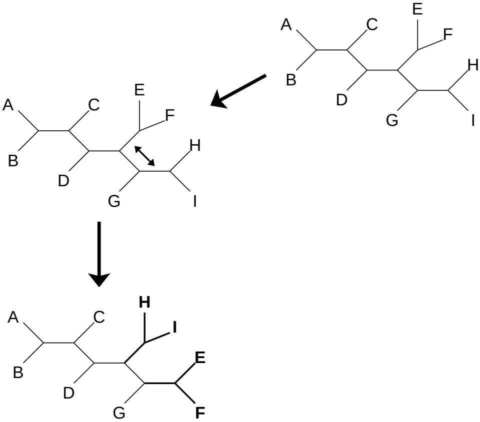
Tree Rearrangements



Tree Rearrangements



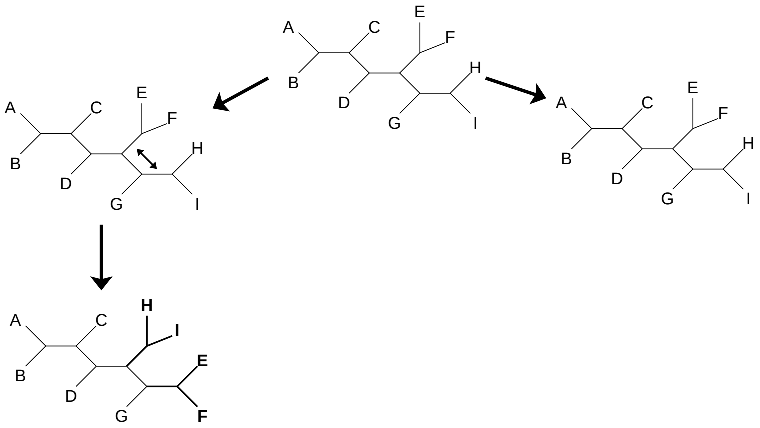
Tree Rearrangements



Nearest Neighbor Interchange

Possible NNI trees = O(n)

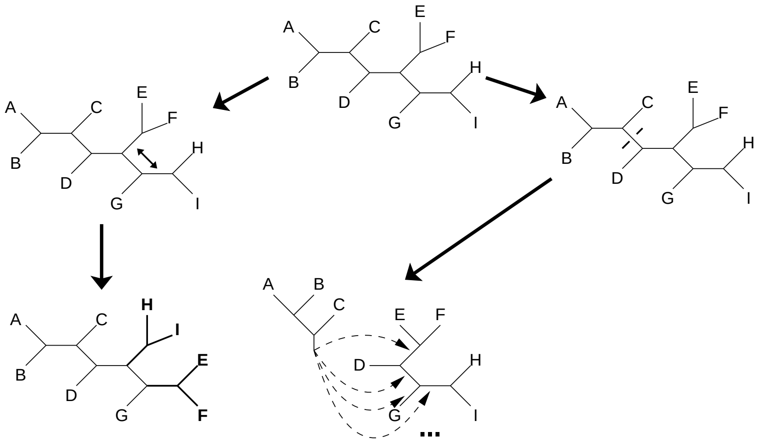
Tree Rearrangements



Nearest Neighbor Interchange

Possible NNI trees = O(n)

Tree Rearrangements



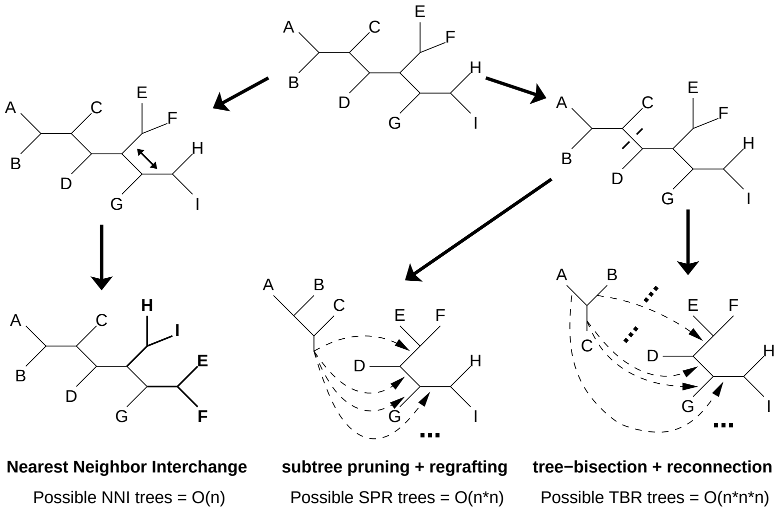
Nearest Neighbor Interchange

Possible NNI trees = O(n)

subtree pruning + regrafting

Possible SPR trees = O(n*n)

Tree Rearrangements



Outline

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.
- We want to be able to do the same with trees.

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.
- We want to be able to do the same with trees.
- The bootstrap is a general statistical method that can be used in this case.

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.
- We want to be able to do the same with trees.
- The bootstrap is a general statistical method that can be used in this case.
    - Nonparametric bootstrap, just re-samples the alignment.

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.
- We want to be able to do the same with trees.
- The bootstrap is a general statistical method that can be used in this case.
  - Nonparametric bootstrap, just re-samples the alignment.
  - Parametric bootstrap uses model parameters to generate replicate data.

Bootstraps

- Usually when we estimate some parameter from data, we have some measure of variability. ie Mean and standard deviation.
- We want to be able to do the same with trees.
- The bootstrap is a general statistical method that can be used in this case.
  - Nonparametric bootstrap, just re-samples the alignment.
  - Parametric bootstrap uses model parameters to generate replicate data.
- Bayesian methods usually get this for “free” because we already have a large set of trees that represent potions in the posterior density.

Pros and Cons

Pros

- Established statistical method.

Cons

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.

Cons

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.
- Studies indicate that it’s quite conservative.

Cons

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.
- Studies indicate that it's quite conservative.

Cons

- Results have no convenient interpretation. ie 50% support does not mean 50% probability.

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.
- Studies indicate that it's quite conservative.

Cons

- Results have no convenient interpretation. ie 50% support does not mean 50% probability.
- Some strong assumptions are imposed on the data. ie iid.

Pros and Cons

Pros

- Established statistical method.
- Simple to implement.
- Studies indicate that it's quite conservative.

Cons

- Results have no convenient interpretation. ie 50% support does not mean 50% probability.
- Some strong assumptions are imposed on the data. ie iid.
- Relies on the fact that the data sample we are using is representative of entire “population” of data.

Bootstrap flow

- Estimate a ML tree and the model parameters θ.

Bootstrap flow

- Estimate a ML tree and the model parameters θ.
- From the data/or estimateted parameters, generate replicate data sets.

Bootstrap flow

- Estimate a ML tree and the model parameters θ.
- From the data/or estimateted parameters, generate replicate data sets.
- For each replicate data set estimate a replicate ML tree.

Bootstrap flow

- Estimate a ML tree and the model parameters θ.
- From the data/or estimateted parameters, generate replicate data sets.
- For each replicate data set estimate a replicate ML tree.
- Combine the replicate ML trees into some kind of consensus tree.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
  - A site, or column in the alignment is picked at random.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
  - A site, or column in the alignment is picked at random.
  - This column of sequence data is placed into the replicate alignment.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
    - A site, or column in the alignment is picked at random.
    - This column of sequence data is placed into the replicate alignment.
    - Some columns will appear more than once in the replicate alignment.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
  - A site, or column in the alignment is picked at random.
  - This column of sequence data is placed into the replicate alignment.
  - Some columns will appear more than once in the replicate alignment.
  - Other columns will not appear at all.

Nonparametric Bootstrap

- Nonparametric bootstrap samples the alignment with replacement.
    - A site, or column in the alignment is picked at random.
    - This column of sequence data is placed into the replicate alignment.
    - Some columns will appear more than once in the replicate alignment.
    - Other columns will not appear at all.
- Requires that the data is IID across sites.

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C
G
C
T
T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C
G
C
T
T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A
G	A
C	A
T	T
T	A

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T
G	A	T
C	A	-
T	T	T
T	A	T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C
G	A	T	G
C	A	-	C
T	T	T	T
T	A	T	T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C
G	A	T	G	T
C	A	-	C	C
T	T	T	T	C
T	A	T	T	-

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T
G	A	T	G	T	T
C	A	-	C	C	G
T	T	T	T	C	T
T	A	T	T	-	T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T
G	A	T	G	T	T	A
C	A	-	C	C	G	T
T	T	T	T	C	T	T
T	A	T	T	-	T	T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T
G	A	T	G	T	T	A	T
C	A	-	C	C	G	T	G
T	T	T	T	C	T	T	T
T	A	T	T	-	T	T	T

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C
G	A	T	G	T	T	A	T	T
C	A	-	C	C	G	T	G	C
T	T	T	T	C	T	T	T	C
T	A	T	T	-	T	T	T	-

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C	G
G	A	T	G	T	T	A	T	T	G
C	A	-	C	C	G	T	G	C	C
T	T	T	T	C	T	T	T	C	C
T	A	T	T	-	T	T	T	-	C

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C	G
G	A	T	G	T	T	A	T	T	G
C	A	-	C	C	G	T	G	C	C
T	T	T	T	C	T	T	T	C	C
T	A	T	T	-	T	T	T	-	C

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C	G
G	A	T	G	T	T	A	T	T	G
C	A	-	C	C	G	T	G	C	C
T	T	T	T	C	T	T	T	C	C
T	A	T	T	-	T	T	T	-	C

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C	G
G	A	T	G	T	T	A	T	T	G
C	A	-	C	C	G	T	G	C	C
T	T	T	T	C	T	T	T	C	C
T	A	T	T	-	T	T	T	-	C

Original Data

A	C	A	C	G	C	T	T	T	A
A	G	A	T	G	C	T	T	A	A
A	C	C	C	C	-	-	G	T	A
A	T	A	C	C	C	T	T	T	T
A	T	-	-	C	C	T	T	T	A

Re-sampled Data

C	A	T	C	C	T	T	T	C	G
G	A	T	G	T	T	A	T	T	G
C	A	-	C	C	G	T	G	C	C
T	T	T	T	C	T	T	T	C	C
T	A	T	T	-	T	T	T	-	C

Jackknife is the same without replacement

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.
  - Using these estimated parameters and the estimated ML tree simulate a new replicate data set.

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.
  - Using these estimated parameters and the estimated ML tree simulate a new replicate data set.
  - Estimate a new ML tree and parameters θ'.

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.
  - Using these estimated parameters and the estimated ML tree simulate a new replicate data set.
  - Estimate a new ML tree and parameters θ'.
  - In some cases model parameters can be fixed.

Parametric Bootstrap

- Instead of re-sampling the data, we use estimated model parameters.
  - Start by estimating a ML tree and model parameters θ.
  - Using these estimated parameters and the estimated ML tree simulate a new replicate data set.
  - Estimate a new ML tree and parameters θ'.
  - In some cases model parameters can be fixed.
- Parametric bootstraps do not make any extra assumptions about the data over the model.

Combining the trees

- 50% Majority rule is conservative and all nodes cannot be conflicting.

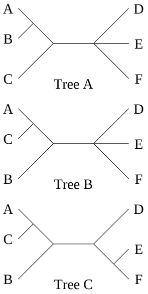
Combining the trees

- 50% Majority rule is conservative and all nodes cannot be conflicting.
- Extended consensus rules can vary slightly in implementation.

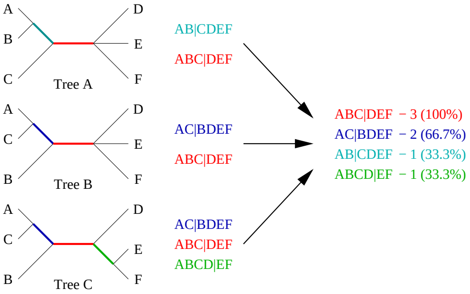
Combining the trees

- 50% Majority rule is conservative and all nodes cannot be conflicting.
- Extended consensus rules can vary slightly in implementation.
- In particular the extended majority rule (default in Consensus) can have nodes in the final tree that conflict with nodes that are more frequent.

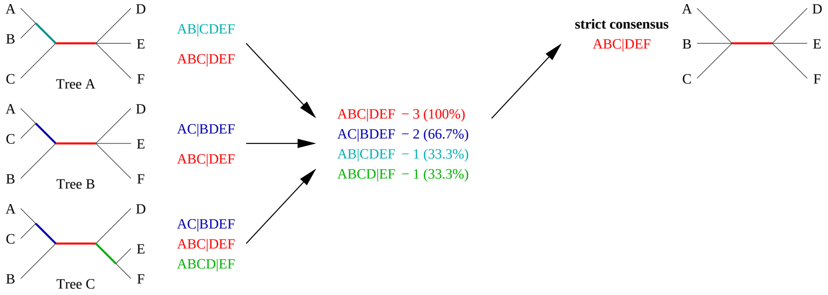
Summarising Trees: Consensus Methods



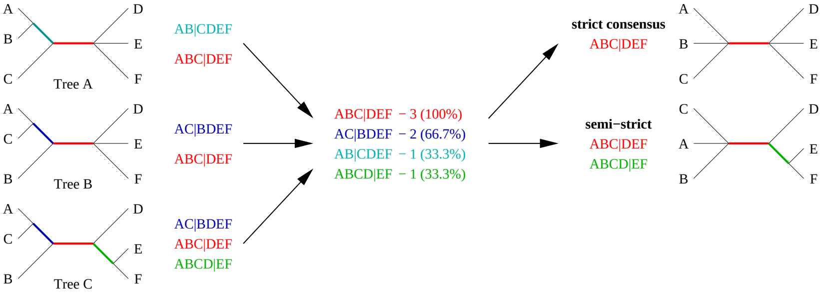
Summarising Trees: Consensus Methods



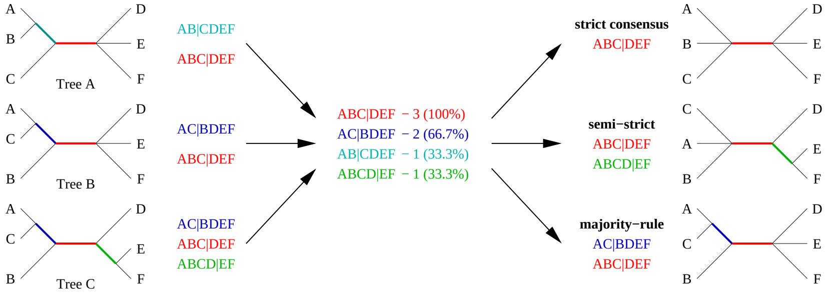
Summarising Trees: Consensus Methods



Summarising Trees: Consensus Methods



Summarising Trees: Consensus Methods



Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.

Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.
- It is not a probability.

Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.
- It is not a probability.
- Generally it appears to be somewhat conservative.

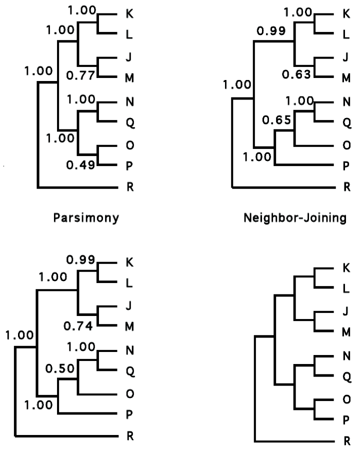
Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.
- It is not a probability.
- Generally it appears to be somewhat conservative.
- On the other hand it is not uncommon to see high bootstrap support for the wrong tree.

Interpretation

- Unfortunately in this setting interpreting bootstrap scores is not straight forward.
- It is not a probability.
- Generally it appears to be somewhat conservative.
- On the other hand it is not uncommon to see high bootstrap support for the wrong tree.
- One interpretation is that the bootstrap attempts to measure sampling variance. (Swofford, et al 1996)

Example Support of a known tree



Hills et al, 1992. Bacteriophage T7 DNA sequences with a known phylogeny.

Outline

Hypothesis testing

- What question do I want to answer?

Hypothesis testing

- What question do I want to answer?
- Say should I use the JC model or the GTR model?

Hypothesis testing

- What question do I want to answer?
- Say should I use the JC model or the GTR model?
- Or perhaps, Is tree A statistically significantly different from tree B?

Hypothesis testing

- What question do I want to answer?
- Say should I use the JC model or the GTR model?
- Or perhaps, Is tree A statistically significantly different from tree B?
- Answering these question is the advantage of using ML.

Hypothesis testing

- What question do I want to answer?
- Say should I use the JC model or the GTR model?
- Or perhaps, Is tree A statistically significantly different from tree B?
- Answering these question is the advantage of using ML.
- It's important to note that you should know the null hypothesis/hypotheses before you "collect" the data.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.
- In such a situation we can consider the likelihood of both models.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.
- In such a situation we can consider the likelihood of both models.
- The Hypothesis: Is the more complicated model better?

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.
- In such a situation we can consider the likelihood of both models.
- The Hypothesis: Is the more complicated model better?
- The Null Hypothesis: Both models are equally good.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.
- In such a situation we can consider the likelihood of both models.
- The Hypothesis: Is the more complicated model better?
- The Null Hypothesis: Both models are equally good.
- Note that the more complicated model always has an equal or higher likelihood.

Nested models

- A model is nested in another model, if it is a simplification of the complicated model.
- eg Star topology. GTR vrs JC.
- In such a situation we can consider the likelihood of both models.
- The Hypothesis: Is the more complicated model better?
- The Null Hypothesis: Both models are equally good.
- Note that the more complicated model always has an equal or higher likelihood.
- We can use a Log Likelihood ratio test.

LRT

Log Likelihood ratio test

$$\lambda = -2\log\frac{L_0}{L_1} = 2(\log L_1 - \log L_0)$$

- λ is asymptotically distributed to the χ^2 distribution with the appropriate degrees of freedom.

LRT

Log Likelihood ratio test

$$\lambda = -2\log\frac{L_0}{L_1} = 2(\log L_1 - \log L_0)$$

- λ is asymptotically distributed to the χ^2 distribution with the appropriate degrees of freedom.
- The degrees of freedom are the difference between the two models i.e. Star tree compared to a given tree, it's the number of internal branches.

LRT

Log Likelihood ratio test

$$\lambda = -2 \log \frac{L_0}{L_1} = 2(\log L_1 - \log L_0)$$

- λ is asymptotically distributed to the χ^2 distribution with the appropriate degrees of freedom.
- The degrees of freedom are the difference between the two models i.e. Star tree compared to a given tree, it's the number of internal branches.
- We calculate λ and check if it's outside our P-value range on the χ^2 distribution.

Tree Tests

- LRT cannot be used on different topologies.

Tree Tests

- LRT cannot be used on different topologies.
- So two tree test methods have been developed. KH and SH

Tree Tests

- LRT cannot be used on different topologies.
- So two tree test methods have been developed. KH and SH
- Note that the first test (KH) is often misapplied.

Tree Tests

- LRT cannot be used on different topologies.
- So two tree test methods have been developed. KH and SH
- Note that the first test (KH) is often misapplied.
- The idea is similar to the LRT that there is a statistic that is compared to a distribution. Only now we must estimate that distribution.