

MODEL-BASED PHYLOGENETICS AND INFERENCE.

MOTIVATION

- Use a probabilistic model to describe the process of change from ancestral to derived states along a tree branch
- Model describe our mechanistic understanding of the evolutionary process.
- Using formal models, we can estimate biologically relevant parameters such as branch lengths and divergence times, substitution rates, measures of selection.
- Each model can be assigned a goodness of fit (usually derived from its Log L score and complexity)
- Models can be compared to decide which parameters are important, or to test what values biological quantities may take (hypothesis testing)
- Sequence data can be 'mined' for pattern discovery using collections of substitution models

JUKES-CANTOR 1969

- The idea is to model substitutions at a site using a **Markov Process**.
- Very much like a Markov chain, except time is now continuous, instead of being measured in discrete steps.
- $X(t)$ defines the probability distribution that the observed quantity follows at time $t \geq 0$.
- Markov property (memoryless process):

$$\Pr\{X(t) = x_0 | X(t_1) = x_1, \dots, X(t_n) = x_n, t > t_1 > \dots > t_n \geq 0\} = \Pr\{X(t) = x_0 | X(t_1) = x_1\}$$

MARKOV PROCESSES

- To completely define a Markov process, we need to specify the **transition probability function**: given that the process is in state A at time \mathbf{u} , what is the probability that it will be in state B at a later time, $\mathbf{u} + \mathbf{t}$?
- Often written as a matrix, $\mathbf{T}(\mathbf{u}, \mathbf{t})$:

$$T(u, t)_{AB} = \Pr\{X(u + t) = B | X(u) = A\}$$

- If one further assumes that the process is homogeneous, i.e. $\mathbf{T}(\mathbf{u}, \mathbf{t})$ does not depend on \mathbf{u} , then

$$T(t)_{AB} = \Pr\{X(t) = B | X(0) = A\}$$

MARKOV PROCESSES (CONT'D)

- For homogeneous processes, it is easier to define the process in terms of its **rate matrix Q**:

$$Q = \lim_{t \downarrow 0} \frac{T(t) - I}{t}$$

- Given **Q**, it can be shown that for $t \geq 0$,

$$T(t) = \exp Qt$$

- where the matrix exponential is defined by the standard Taylor series

$$\exp Qt = I + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \dots$$

- There are abundant numerical algorithms that compute the matrix exponential in $O(C^3)$ time, where C is the dimension of the rate matrix.

JUKES CANTOR ('69) DISTANCE: JC69

- The Markov process assumes that all four bases are equally probable and that nucleotides mutate to other nucleotides with equal rates.
- Diagonal rates are defined by the requirement that the *transition* matrix forms a valid probability distribution in each row: for this to hold, each row in the *rate* matrix must sum to 0.

Rate matrix Q

From↓ To →	A	C	G	T
A	-0.75	0.25	0.25	0.25
C	0.25	-0.75	0.25	0.25
G	0.25	0.25	-0.75	0.25
T	0.25	0.25	0.25	-0.75

Transition matrix T (t)

From↓ To →	A	C	G	T
A	$\frac{1}{4}(1+3e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$
C	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1+3e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$
G	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1+3e^{-t})$	$\frac{1}{4}(1-e^{-t})$
T	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1-e^{-t})$	$\frac{1}{4}(1+3e^{-t})$

T(0)

From↓ To →	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

T(0.5)

From↓ To →	A	C	G	T
A	0.352	0.216	0.216	0.216
C	0.216	0.352	0.216	0.216
G	0.216	0.216	0.352	0.216
T	0.216	0.216	0.216	0.352

T(0.1)

From↓ To →	A	C	G	T
A	0.753	0.082	0.082	0.082
C	0.082	0.753	0.082	0.082
G	0.082	0.082	0.753	0.082
T	0.082	0.082	0.082	0.753

T(∞)

From↓ To →	A	C	G	T
A	0.25	0.25	0.25	0.25
C	0.25	0.25	0.25	0.25
G	0.25	0.25	0.25	0.25
T	0.25	0.25	0.25	0.25

ML FITTING JC69

- The objective is to find the optimal **t**, given the data
- Use the principle of maximal likelihood to select **t**, which maximizes the probability of observing the alignment given the model

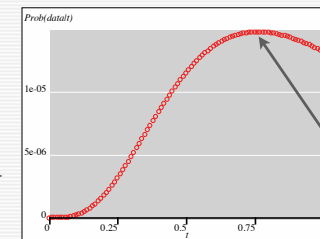
A T G A A A G C G A
A G T A G A G T G A

Simplify...

$$Pr\{data|t\} = \left(\frac{1}{4}[1+3e^{-t}]\right)^6 \left(\frac{1}{4}[1-e^{-t}]\right)^4$$

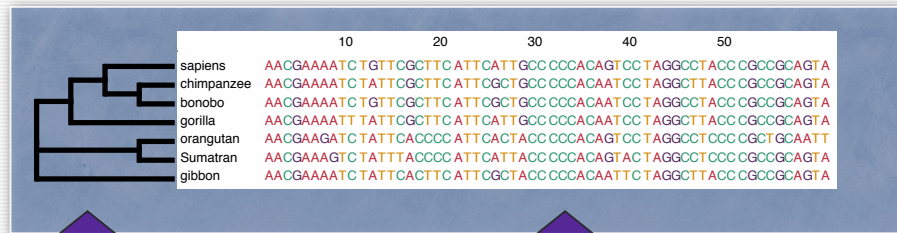
Independent sites

$$Pr\{data|t\} = Pr\{A \rightarrow A|t\}^4 \times Pr\{T \rightarrow G|t\} \times Pr\{G \rightarrow T|t\} \times Pr\{G \rightarrow G|t\}^2 \times Pr\{C \rightarrow T|t\}$$



t=0.76214

MAXIMUM LIKELIHOOD SEQUENCE ANALYSIS



Phylogeny

Alignment of homologous sequences (column = site)

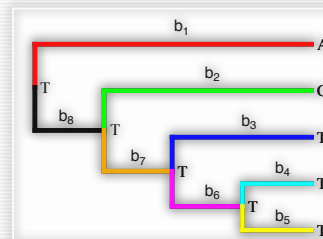
Substitution Models

“All models are wrong, but some models are useful” Box (1976)

MAXIMUM LIKELIHOOD MODELS

(PLEASE SEE [HTTP://WWW.HYPHY.ORG/DOCS/MAXIMUMLIKELIHOOD.PDF](http://www.hypHY.org/docs/maximumlikelihood.pdf) FOR DETAILS)

- Define the probability of a point substitution along a branch at a given site:
- $Q_{x,y}^i(t; \theta) = Pr_{\theta} \{x \text{ is replaced with } y \text{ in time } t : x, y \in C\}$
- Continuous time Markov chains.
- Typically, the models are stationary and time-reversible.

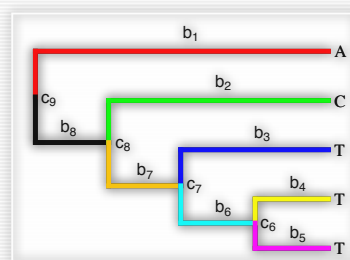


$$L(\mathcal{D}_s; \mathcal{T}, \theta) = Q_{T,A}^1(t_1; \theta) Q_{T,T}^8(t_8; \theta) Q_{T,C}^2(t_2; \theta) Q_{T,T}^4(t_4; \theta) Q_{T,T}^3(t_3; \theta) Q_{T,T}^5(t_5; \theta) Q_{T,T}^6(t_6; \theta) Q_{T,T}^7(t_7; \theta)$$

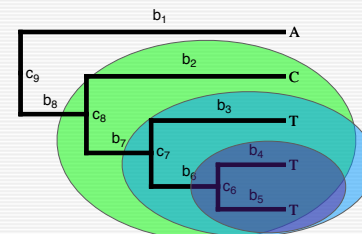
If ancestral states were known.

COMPUTING LIKELIHOOD

- Ancestral states are almost always unknown - must sum over all possible internal node character assignments.
- Computations can be done efficiently, in $O(C^2N)$ time, as opposed to $O(C^{2N})$ “brute force” time, using Felsenstein’s pruning algorithm (1981) that takes advantage of conditional independence of evolution along tree branches



$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in C} \sum_{c_8 \in C} \sum_{c_7 \in C} \sum_{c_6 \in C} \pi(c_9) Q_{c_9,A}^1(t_1; \theta) Q_{c_9,c_8}^8(t_8; \theta) Q_{c_8,C}^2(t_2; \theta) Q_{c_8,c_7}^7(t_7; \theta) Q_{c_7,T}^3(t_3; \theta) Q_{c_7,c_6}^6(t_6; \theta) Q_{c_6,T}^4(t_4; \theta) Q_{c_6,T}^5(t_5; \theta)$$



$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in C} \sum_{c_8 \in C} \sum_{c_7 \in C} \sum_{c_6 \in C} \pi(c_9) Q_{c_9,A}^1(t_1; \theta) Q_{c_9,c_8}^8(t_8; \theta) Q_{c_8,C}^2(t_2; \theta) Q_{c_8,c_7}^7(t_7; \theta) Q_{c_7,T}^3(t_3; \theta) Q_{c_7,c_6}^6(t_6; \theta) Q_{c_6,T}^4(t_4; \theta) Q_{c_6,T}^5(t_5; \theta)$$

Only depends on c_9

$$L(\mathcal{D}_s; \mathcal{T}, \theta) = \sum_{c_9 \in C} \pi(c_9) Q_{c_9,A}^1(t_1; \theta) \sum_{c_8 \in C} Q_{c_9,c_8}^8(t_8; \theta) Q_{c_8,C}^2(t_2; \theta) \times \sum_{c_7 \in C} \left(Q_{c_8,c_7}^7(t_7; \theta) Q_{c_7,T}^3(t_3; \theta) \sum_{c_6 \in C} (Q_{c_7,c_6}^6(t_6; \theta) Q_{c_6,T}^4(t_4; \theta) Q_{c_6,T}^5(t_5; \theta)) \right)$$

Only depends on c_8

Only depends on c_7

FELSENSTEIN'S PRUNING ALGORITHM

- Idea: for each node **n** in the tree, maintain a **C** (number of characters) - dimensional vector **L_n**, whose *i*-th element records the probability of the subtree rooted at **n**, given that the character at node **n** is **i**.
- For leaves, **L_n** is easy to compute. **L_n[i] = 1** if **n** is labeled with character **i**, and **L_n[i] = 0**, otherwise
- For interior nodes, **L_n[i]** is computed by iterating of all children of **n**, and computing the cumulative probability of changing from **i** to any other state at child **m** (this uses **L_m**), and then taking the product over all children
- At the root node, **r**, compute the likelihood of the site, by summing over all characters **L_r[i] x π(i)**, where **π(i)** is the (supplied) distribution of characters at the root.

LIKELIHOOD OF ALIGNMENTS

- We are forced make the assumption of independently evolving sites (can be relaxed a bit) to make the computation tractable. Hence the likelihood of an alignment is the product of site likelihoods
- Models are fitted by adjusting all the free parameters using numerical optimization techniques to maximize the likelihood (the probability of observing given sequence data under a model).
- Optimized values of model parameters are called MLEs - maximum likelihood estimates.
- ML estimation is a venerable approach in statistics. Many fundamental results exist to show that MLEs have desirable statistical properties, such as consistency, asymptotic efficiency, optimality of likelihood based hypothesis tests, etc.

A VERY BASIC EXAMPLE: HYPHY.

- Given a nucleotide alignment and a tree, fit a simple substitution model and interpret its output.

```

READ THE FOLLOWING DATA
8 species:
(B_FR_83_HXB2_ACC_K03455,B_US_83_RF_ACC_M17451,B_US_86_JRFL_ACC_U63632,B_US_90_WEAU160_ACC_U21135,D_CD_83_ELT_ACC_K03454,D_CD_83_NDK_ACC_M27323,D_CD_84_84ZR085_ACC_U88822,D_UG_94_94UG114_ACC_U88824);
Total Sites:1320;
Distinct Sites:118

A tree was found in the data file:
(C((D_CD_83_ELT_ACC_K03454,D_CD_83_NDK_ACC_M27323),D_UG_94_94UG114_ACC_U88824),D_CD_84_84ZR085_ACC_U88822),B_US_83_RF_ACC_M17451,((B_FR_83_HXB2_ACC_K03455,B_US_86_JRFL_ACC_U63632),B_US_90_WEAU160_ACC_U21135));

Would you like to use it:(Y/N)?y
    
```

```

RESULTS
Time taken = 0.23 seconds
AIC-Score = -6682.51
Log Likelihood = -3327.25252976199;
Shared Parameters:
R=0.111274

Tree givenTree=(B_US_83_RF_ACC_M17451:0.0262012,
((B_FR_83_HXB2_ACC_K03455:0.0116675,B_US_86_JRFL_ACC_U63632:0.0178118)
Node4:0.0022271,B_US_90_WEAU160_ACC_U21135:0.0209889)Node3:0.00507481,
(((D_CD_83_ELT_ACC_K03454:0.0188158,D_CD_83_NDK_ACC_M27323:0.0100127)
Node10:0.0105018,D_UG_94_94UG114_ACC_U88824:0.053071)Node9:0.00391151,D_CD_84_84ZR085_ACC_U88822:0.0283077)
Node8:0.0234111);
    
```

The screenshot shows the HyPhy interface. On the left is a phylogenetic tree with a scale bar from 0.01 to 0.06889. The tree has several branches labeled with species IDs like B_US_83_RF_ACC_M17451, B_FR_83_HXB2_ACC_K03455, B_US_86_JRFL_ACC_U63632, B_US_90_WEAU160_ACC_U21135, D_CD_83_ELT_ACC_K03454, D_CD_83_NDK_ACC_M27323, D_UG_94_94UG114_ACC_U88824, and D_CD_84_84ZR085_ACC_U88822. On the right is a 'PARAMETERS' table with columns for Parameter ID, Value, and Constraint. The table lists parameters for the givenTree, R, and various substitution rates (R%, a) for different branches and nodes.

TREE

Parameter	Value	Constraint
givenTree	0.111274	
R	0.0395069	
givenTree.B_US_83_RF_ACC_M17451.a	0.0887193	
givenTree.B_US_86_JRFL_ACC_U63632.a	0.0603122	
givenTree.B_US_90_WEAU160_ACC_U21135.a	0.0710701	
givenTree.D_CD_83_ELT_ACC_K03454.a	0.0637117	
givenTree.D_CD_83_NDK_ACC_M27323.a	0.0339038	
givenTree.D_CD_84_84ZR085_ACC_U88822.a	0.095852	
givenTree.D_UG_94_94UG114_ACC_U88824.a	0.179703	
givenTree.Node10.a	0.0355598	
givenTree.Node3.a	0.0171837	
givenTree.Node4.a	0.00754114	
givenTree.Node8.a	0.079272	
givenTree.Node9.a	0.0132447	

Equilibrium Freqs.	A	C	G	T
A	0.404451	*	R%	R%
C	0.166288	R%	*	R%
G	0.209564	a	R%	*
T	0.219697	R%	a	*

MODEL

TESTING HYPOTHESES

- Model (+constraints on parameters) = hypothesis
- To test hypotheses, we fit several models of varying complexity to the data and compare their goodness-of-fit. The model that fits the data significantly better than all others is chosen as the best explanation to how the data have arisen.
- The simplest case has two models
 - Null (simple)
 - Alternative (more complex)

EXAMPLE

- Test for presence of transition/transversion biases (κ) in p_{51}
- Fit a simpler model (F81) which can be obtained from HKY85 by constraining $\kappa = 1$
- Repeat the analysis of $p_{51.nex}$ using the F81 and record the logL value

```
----- READ THE FOLLOWING DATA -----
8 SPECIES:
(B_FR_83_HXB2_ACC_K03455,B_US_83_RF_ACC_M17451,B_US_86_JRFL_ACC_U63632,B_US_90_WEAU160_ACC_U21135,D_CD_83_EL1_ACC_K03454,D_CD_83_NDK_ACC_M27323,D_CD_84_84ZR085_ACC_U88822,D_UG_94_94UG114_ACC_U88824);
Total: Sites:1200;
Distinct Sites:118

A TREE WAS FOUND IN THE DATA FILE:
(((D_CD_83_EL1_ACC_K03454,D_CD_83_NDK_ACC_M27323),D_UG_94_94UG114_ACC_U88824),D_CD_84_84ZR085_ACC_U88822),B_US_83_RF_ACC_M17451,
((B_FR_83_HXB2_ACC_K03455,B_US_86_JRFL_ACC_U63632),B_US_90_WEAU160_ACC_U21135));

WOULD YOU LIKE TO USE IT: (Y/N)?Y

----- RESULTS -----
Time taken = 0.29 seconds
AIC Score = 6993.99
Log Likelihood = -3470.69378663355;
Tree givenTree=(B_US_83_RF_ACC_M17451:0.0257672,((B_FR_83_HXB2_ACC_K03455:0.0114795,B_US_86_JRFL_ACC_U63632:0.0177305)Node4:0.00237498,B_US_90_WEAU160_ACC_U21135:0.0208211)
Node3:0.00518708,(((D_CD_83_EL1_ACC_K03454:0.0104005,D_CD_83_NDK_ACC_M27323:0.0100241)Node10:0.0103215,D_UG_94_94UG114_ACC_U88824:0.0517432)
Node9:0.00366331,D_CD_84_84ZR085_ACC_U88822:0.0280537)Node8:0.0231355);
```

LIKELIHOOD RATIO TEST

- Alternative model (HKY85) yielded log L = **-3327.3**, while the simpler null (F81) model returned logL = **-3470.7**
- Smaller likelihood = worse fit
- However because HKY85 has one more parameter than F81, it should always beat (or at least match) the score of F81, even if F81 were the correct model.
- Is the improvement in fit (log(LR)- likelihood ratio = 143.4) large enough to be significant?
- Perform a Likelihood Ratio Test (LRT), comparing the distribution of $2*LR$ with the tail of the chi-squared distribution with as many degrees of freedom as there are additional parameters in the alternative model
- In this case, p-value - the probability that LR \geq observed value if the null model is correct is effectively 0, hence there is very strong evidence that transitions happen at higher rates than transversions in our data set.

DO THE SAME IN THE GUI.

- Fit the alternative model (HKY85)
- Save it in the likelihood panel
- Constrain the transition/transversion parameter to be one; re-optimize and save as null
- Compute the LRT
- Perform parametric bootstrap

Parameter ID	Value	Constr
p51_tree		
p51_part_Shared_TYTS	0.111273	
p51_tree.B_FR_83_H0B2_ACC_K034551	0.0395067	
p51_tree.B_US_83_RF_ACC_H174511	0.0887199	
p51_tree.B_US_86_JRF1_ACC_U636321	0.0605121	
p51_tree.B_US_90_WEAU160_ACC_U0211351	0.0710705	
p51_tree.D_CD_83_ELI_ACC_K034541	0.0637121	
p51_tree.D_CD_83_NDK_ACC_H273231	0.0399051	
p51_tree.D_CD_84_B4ZRO085_ACC_U888221	0.0958519	
p51_tree.D_US_94_9406114_ACC_U888221	0.1797903	
p51_tree.Node1.1	0.0792915	
p51_tree.Node10.1	0.00754072	
p51_tree.Node2.1	0.013245	
p51_tree.Node3.1	0.0355593	
p51_tree.Node9.1	0.0171841	

Log Likelihood = -3327.23, parameter count = 14, AIC = 6492.31.

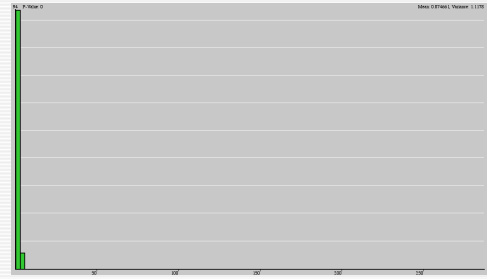
ALTERNATIVE

Parameter ID	Value	Constraint
p51_tree		
p51_part_Shared_TYTS	1	1
p51_tree.B_FR_83_H0B2_ACC_K034551	0.01602	
p51_tree.B_US_83_RF_ACC_H174511	0.0359591	
p51_tree.B_US_86_JRF1_ACC_U636321	0.0247435	
p51_tree.B_US_90_WEAU160_ACC_U0211351	0.0290956	
p51_tree.D_CD_83_ELI_ACC_K034541	0.0257889	
p51_tree.D_CD_83_NDK_ACC_H273231	0.0159882	
p51_tree.D_CD_84_B4ZRO085_ACC_U888221	0.0391496	
p51_tree.D_US_94_9406114_ACC_U888221	0.0722095	
p51_tree.Node1.1	0.0322862	
p51_tree.Node10.1	0.0033138	
p51_tree.Node2.1	0.00511234	
p51_tree.Node3.1	0.0144056	
p51_tree.Node9.1	0.00712652	

Log Likelihood = -3470.63, parameter count = 13, AIC = 6367.33.

NULL

PARAMETRIC BOOTSTRAP

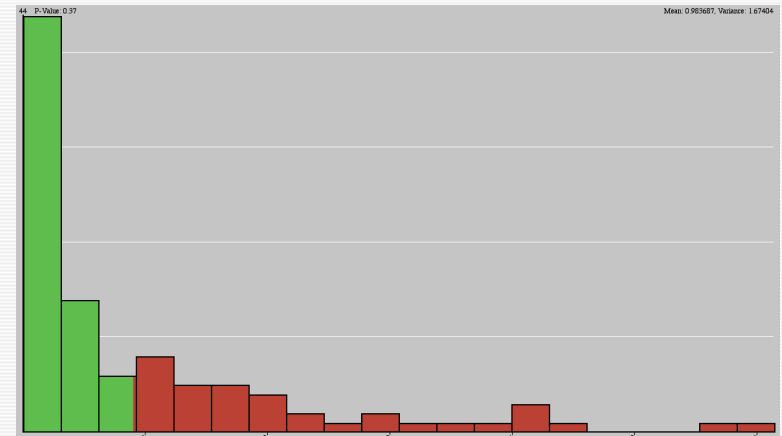


CHI²

LIKELIHOOD RATIO TEST
 2*LR = 286.883
 DF = 1
 P-VALUE = 0

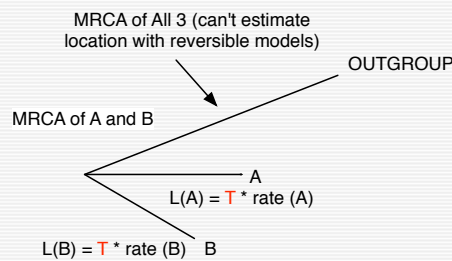
COMPARE TO WHEN THE NULL MODEL IS TRUE

USE ON SIMULATED F81.NEX, TEST F81 (NULL) VS HKY85 (ALTERNATIVE)



RELATIVE RATE TESTS

- For many models, we can't decouple substitution rates and evolutionary times; they are globbed together as 'expected substitutions/site'
- However, sometimes it is possible to factor out the time to directly compare evolutionary rates
- One of the first tests to do that was the relative ratio test: using an outgroup to 'polarize' substitutions.
- It is then possible to directly compare branch lengths and make statements about evolutionary rates.



If L(A) != L(B), this means that rate (A) != rate (B)

TESTING FOR EQUALITY OF EVOLUTIONARY RATES
 S. V. MUSE AND B. S. WEIR
 Genetics, (132), 269-276

RR EXAMPLES (USE PAIRWISE RELATIVE RATE.BF)

PRIMATE_MT DNA.NEX
 ND4/ND5 GENES
 HKY85 (GLOBAL)
 GIBBON OUTGROUP

---- RUNNING PAIRWISE RELATIVE RATE ANALYSIS ----

IN THE SUMMARY TABLE BELOW, BRANCH LENGTHS DENOTE THE EXPECTED NUMBER OF SUBSTITUTIONS PER THAT BRANCH.

(*) CORRESPONDS TO THE .05 SIGNIFICANCE LEVEL
 (**) CORRESPONDS TO THE .01 SIGNIFICANCE LEVEL
 (***) CORRESPONDS TO THE .001 SIGNIFICANCE LEVEL.

TAXA TRIPLET	LRT	P-VALUE
(GIBBON:0.186834,(HUMAN:0.0407094,(CHIMPANZEE:0.0546993)))		0.7338 0.391654
(GIBBON:0.176751,(HUMAN:0.0494128,(GORILLA:0.063398)))		0.6833 0.408466
(GIBBON:0.132351,(HUMAN:0.0860481,(ORANGUTAN:0.102599)))		0.7304 0.392766
(GIBBON:0.183243,(CHIMPANZEE:0.0579861,(GORILLA:0.0590667)))		0.0036 0.951884
(GIBBON:0.132985,(CHIMPANZEE:0.0961306,(ORANGUTAN:0.105782)))		0.1382 0.71012
(GIBBON:0.135942,(GORILLA:0.0959235,(ORANGUTAN:0.102316)))		0.0999 0.751965

SEQUENCE DATING

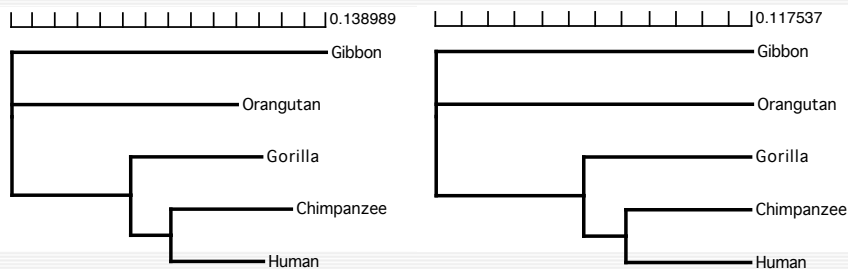
- Estimate 'dates' of ancestral nodes in a phylogenetic tree based on extant sequences
 - Timing of HIV-1 M MRCA
 - Dating speciation events (e.g. human/chimp split)
- Based on 'molecular clock'

RATES AND TIME

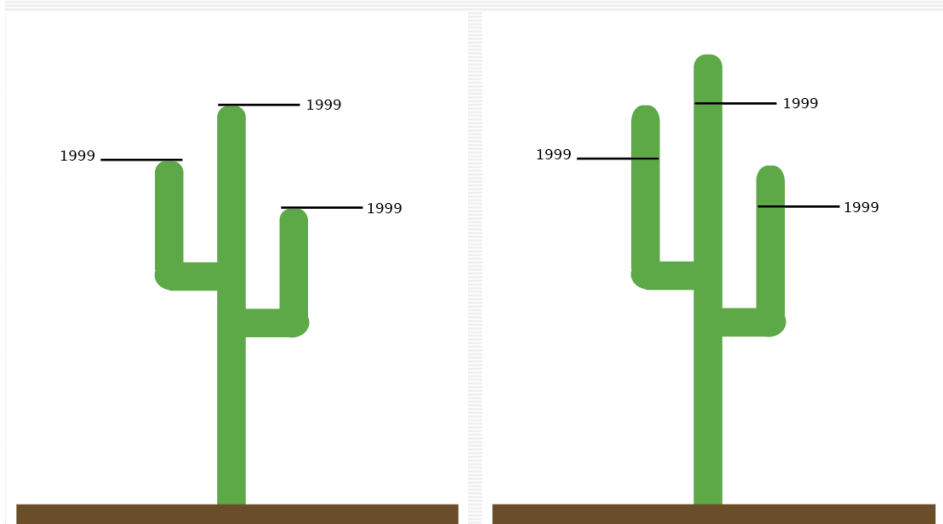
- Branch length = substitution rate (r) x time (t)
- In most evolutionary models only $r \cdot t$ can be estimated
- Branch Length = 0.1
 - $r = 0.1, t = 1$
 - $r = 0.001, t = 100$
- In order to decouple $r \cdot t$ we need
 - Calibration points (e.g. fossil record) to determine time scale or,
 - Sequences sampled at different times

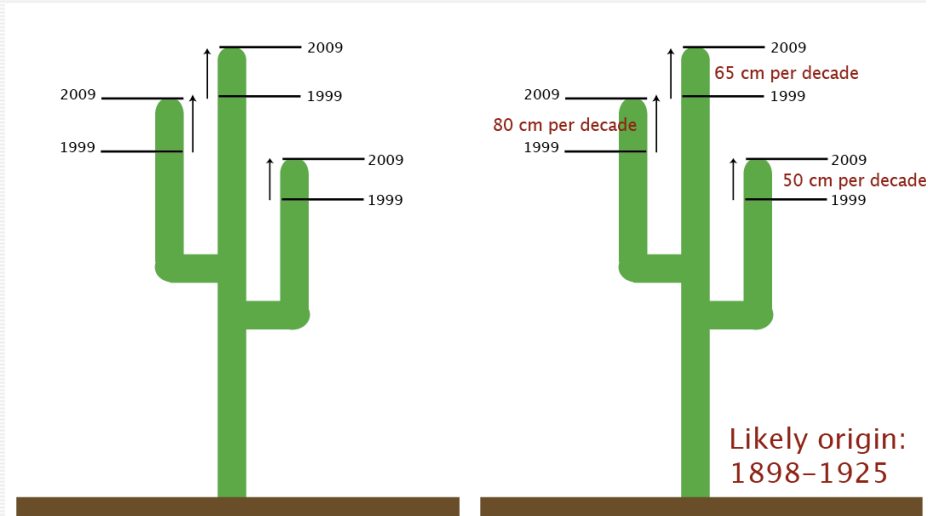
MOLECULAR CLOCK

- Tests whether that the rates (r) are constant along the tree, assuming all the tips are contemporaneous



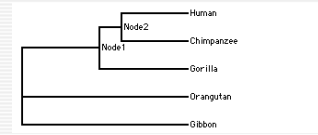
MOLECULAR CLOCK DATING



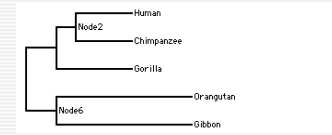


A NOTE ON ROOTING

- In the context of time-reversible models, we cannot place the root of the tree



```
CHIMPANZEE .A := HUMAN .A ;
GIBBON .A := HUMAN .A+NODE2 .A+NODE1 .A ;
ORANGUTAN .A := HUMAN .A+NODE2 .A+NODE1 .A ;
GORILLA .A := HUMAN .A+NODE2 .A ;
```



```
CHIMPANZEE .A := HUMAN .A ;
GIBBON .A := ORANGUTAN .A ;
GORILLA .A := HUMAN .A+NODE2 .A ;
```

WHEN A TREE WITH MORE THAN TWO ROOT CHILDREN IS INPUT, THE CLOCK IS ENFORCED ON ALL BRANCHES, I.E. THE DISTANCE FROM EVERY TIP TO THE SPECIFIED ROOT IS THE SAME

FOR BINARY ROOTED TREES, THE CLOCK MEANS THAT THERE IS A WAY TO SLIDE THE PLACEMENT OF THE ROOT ALONG THE TWO TOP-LEVEL BRANCHES WHICH WILL ENFORCE THE CLOCK (IN THE EXAMPLE ABOVE, THE BRANCH ABOVE NODE6 CAN BE DIVIDED IN A WAY TO ENFORCE THE CLOCK

EXAMPLE: MOLECULARCLOCK.BF

USE ON PRIMATE_MTDNA.NEX, USE HKY85 (GLOBAL), THEN SELECT ANALYSIS>RESULTS>PARAMETRIC BOOTSTRAP TO VERIFY P-VALUE

```
.....
RESULTS WITHOUT THE CLOCK:
Log Likelihood = -2666.7546945248;
Shared Parameters:
R=0.106532
Tree givenTree=((GORILLA:0.0575822,((CHIMPANZEE:0.0537616,HUMAN:0.0413744)Node4:0.0174717)
Node2:0.0530902,ORANGUTAN:0.10003,GIBBON:0.13881));
.....
RESULTS WITH THE CLOCK:
Log Likelihood = -2667.66014503937;
Shared Parameters:
R=0.106207
Tree clockTree=((GORILLA:0.0607272,((CHIMPANZEE:0.0457448,HUMAN:0.0457448)Node4:0.0149823)
Node2:0.0481899,ORANGUTAN:0.108917,GIBBON:0.139311));
.....

          86  2.8170332  0.6395349
          87  0.4380091  0.6321839
          88  12.0805096  0.6363636
          89  4.5595442  0.6404094
          90  4.9519407  0.6444444
          91  5.9528147  0.6483516
          92  1.0695719  0.6412043
          93  0.8584754  0.6344086
          94  5.5049459  0.6382979
          95  3.4711193  0.6421053
          96  1.2890321  0.6354167
          97  3.2898158  0.6391753
          98  3.8249806  0.6428571
          99  0.2620951  0.6464646
         100  5.7759134  0.6500000

          BOOTSTRAPPING SUMMARY

Likelihood Ratio Statistics:
MEAN      = 3.3995457
VARIANCE  = 7.3664000
Proportion larger that the original likelihood ratio=0.65

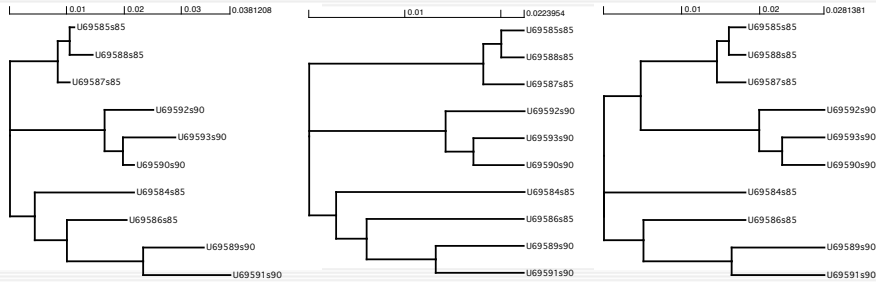
-2*ln Likelihood Ratio=1.8289
Constrained parameters:3
P-Value:0.608667
CPU time taken: 0.15 seconds.
```

OTHER MOLECULAR CLOCK TESTS IMPLEMENTED IN HYPHY

- Local clocks, i.e. only some clades are assumed to be clock-like
- Clocks with every possible root
- Multi-rate clock

DATED TIPS CLOCK

- Assumes that the rates (r) are constant along the tree, but corrects for sampling at different times.



NO CLOCK

LOG L = -5426.1, 26 PARAMETERS

CLOCK ($P < 0.01$)

LOG L = -5455.7, 16 PARAMETERS

DATED CLOCK ($P < 0.01$)

LOG L = -5437.7, 18 PARAMETERS

EXAMPLE: DATEDTIPSMOLECULARCLOCK.BF

USE ON ENV4B.NEX, SELECT SINGLE PARTITION, TIPDATE FORMAT,
YEARS, GRM (GLOBAL)

```

READ THE FOLLOWING DATES:
U69585585: 85 YEARS
U69588585: 85 YEARS
U69587585: 85 YEARS
U69590590: 90 YEARS
U69593590: 90 YEARS
U69592590: 90 YEARS
U69586585: 85 YEARS
U69589590: 90 YEARS
U69591590: 90 YEARS
U69584585: 85 YEARS

CPU TIME TAKEN: 0.66 SECONDS.

-----*
RESULTS WITHOUT THE CLOCK:
LOG LIKELIHOOD = -5426.1248759668;
SAVED PARAMETERS:
CG=0.143345
CI=1.06999
GT=0.164726
AT=0.229323
AC=0.639059

TREE GIVEN TREE=(((U69585585:0.000804528,U69588585:0.00405322)Node3:0.00202025,U69587585:0.00202579)Node2:0.00829657,((U69590590:0.0019399,U69593590:0.00904865)
Node8:0.00311844,U69592590:0.00828366)Node7:0.0165112)Node1:0.00440842,(U69586585:0.0102645,(U69589590:0.010505,U69591590:0.0149915)Node14:0.013145)Node12:0.00557041,U69584585:0.0171745);
-----*

RESULTS WITH DATED TIPS CLOCK:
LOG LIKELIHOOD: -5436.38056471177
CPU TIME TAKEN: 52.21 SECONDS.

-----*
-2*(LN LIKELIHOOD RATIO)=-20.3614
CONSTRAINED PARAMETERS: 8
ASYMPTOTIC P-VALUE: 0.00985197
TREE WITH BRANCH LENGTHS SCALED IN YEARS:
((((U69585585:0.734654,U69588585:0.734654):0.518432,U69587585:1.253089):3.83368,((U69590590:1.99489,U69593590:1.99489):1.12466,U69592590:3.11954):6.16723):1.6739,(U69586585:4.06101,
(U69589590:4.51879,U69591590:4.51879):4.54222):1.89966,U69584585:5.96867)
ROOT OF THE TREE PLACED AT 79.0393(95% CI: 78.6996,79.3181) YEARS
SUBSTITUTION RATE ESTIMATED AT 0.0027608(95% CI: 0.00247022,0.00307574) SUBS PER YEAR PER SITE
    
```