

Maximum Likelihood in Phylogenetics

23 January 2013

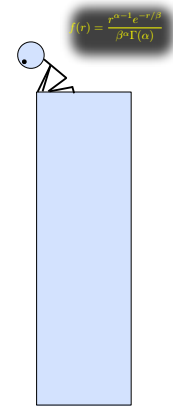
Workshop on Molecular Evolution
Český Krumlov

Paul O. Lewis
Department of Ecology & Evolutionary Biology
University of Connecticut, Storrs, CT

1

Goals

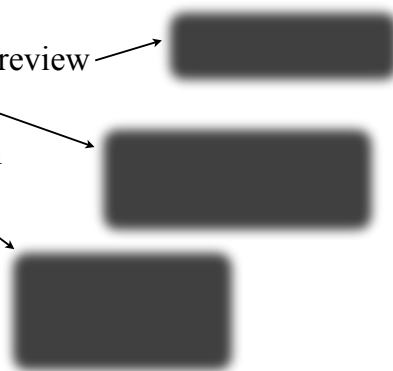
Explain jargon
Increase comfort level
Provide background
In other words...give a hand up



2

The Plan

- Probability review
- Likelihood
- Substitution models



3

Combining probabilities

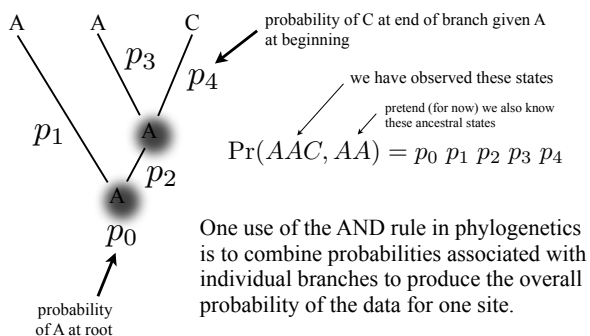
- *Multiply* probabilities if the component events must happen **simultaneously** (i.e. where you would naturally use the word AND when describing the problem)

Using 2 dice, what is the probability of

• AND • ?
 $(1/6) \times (1/6) = 1/36$

4

AND rule in phylogenetics



5

Combining probabilities

- *Add* probabilities if the component events are **mutually exclusive** (i.e. where you would naturally use the word OR in describing the problem)

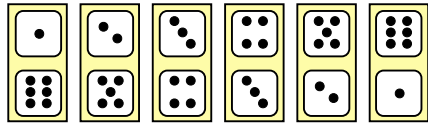
Using one die, what is the probability of

• OR • ?
 $(1/6) + (1/6) = 1/3$

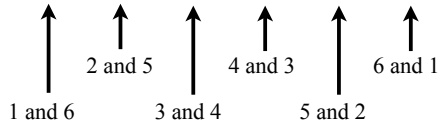
6

Combining AND and OR

What is the probability that the sum of two dice is 7?

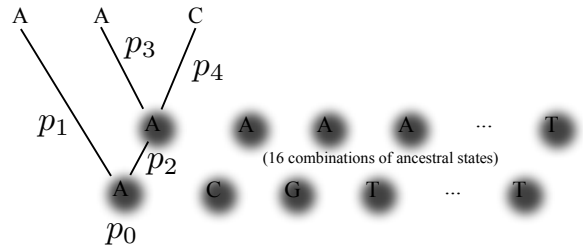


$$(1/36) + (1/36) + (1/36) + (1/36) + (1/36) + (1/36) = 1/6$$



7

Using both AND and OR in phylogenetics



AND rule used to compute probability of the observed data for *each combination* of ancestral states.

OR rule used to combine different combinations of ancestral states.

8

Independence

This is always true...

$$\text{Pr(A and B)} = \text{Pr(A)} \text{Pr(B|A)}$$

joint probability conditional probability

If we can say this...

$$\text{Pr(B|A)} = \text{Pr(B)}$$

...then events A and B are **independent** and we can express the joint probability as the product of Pr(A) and Pr(B)

$$\text{Pr(A and B)} = \text{Pr(A)} \text{Pr(B)}$$

9

Likelihood

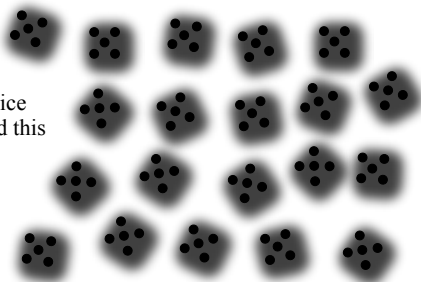
10

The Likelihood Criterion

The probability of the observations computed using a model tells us how surprised we should be.

The preferred model is the one that surprises us least.

Suppose I threw 20 dice down on the table and this was the result...

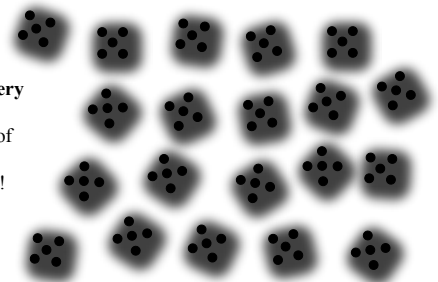


11

The Fair Dice model

$$\text{Pr}(\text{obs.} | \text{fair dice model}) = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$

You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!



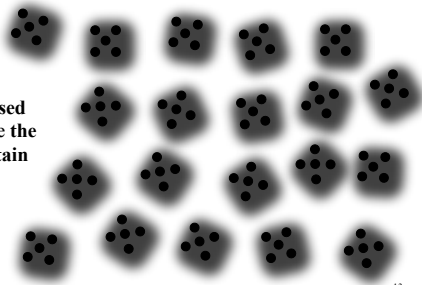
12

The Trick Dice model

(assumes dice each have 5 on every side)

$$\Pr(\text{obs.} | \text{trick dice model}) = 1^{20} = 1$$

You should **not be surprised at all** at this result because the **observed outcome is certain** under this model



13

Results

Model	Likelihood	Surprise level
Fair Dice	$\frac{1}{3,656,158,440,062,976}$	Very, very, very surprised
Trick Dice	1.0	Not surprised at all

winning model maximizes likelihood (and thus minimizes surprise)

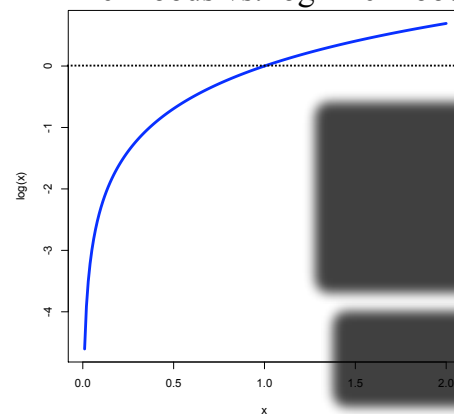
14

Likelihood and model comparison

- Analyses using likelihoods ultimately involve **model comparison**
- The models compared can be **discrete** (as in the fair vs. trick dice example)
- More often the models compared differ **continuously**:
 - Model 1: branch length is 0.05
 - Model 2: branch length is 0.06

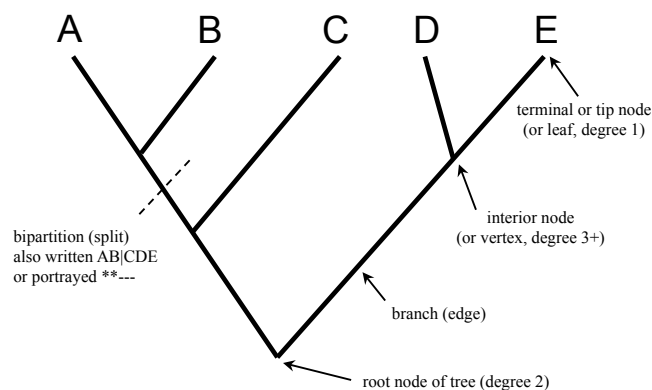
15

Likelihoods vs. log-likelihoods



16

Tree jargon



17

Likelihood of a single tip node

First 32 nucleotides of the $\psi\eta$ -globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

$$L = \pi_G \pi_A \pi_A \pi_G \pi_G \pi_C \pi_C \pi_T \pi_T \pi_A \pi_C \pi_T \pi_T \pi_C \pi_C \pi_A \pi_A \pi_A \pi_T \pi_T \pi_A \pi_A \pi_C \pi_T \pi_G \pi_C \pi_A \pi_C \pi_C \pi_A \pi_C \pi_T \pi_G \pi_G$$

$$= \pi_A^{12} \pi_C^7 \pi_G^7 \pi_T^6$$

$$\log L = 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T)$$

18

Model ranking using AIC

The Akaike Information Criterion (AIC) can be used to evaluate whether an **unconstrained** model ("free") fits the data significantly better than a **constrained** version ("equal") of the same model.

Find *maximum* logL under the *unconstrained* model:

$$\begin{aligned}\log L_{\text{free}} &= 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T) \\ &= 12 \log(0.375) + 7 \log(0.219) + 7 \log(0.219) + 6 \log(0.188) \\ &= -43.1\end{aligned}$$

Find *maximum* logL under the *constrained* model:

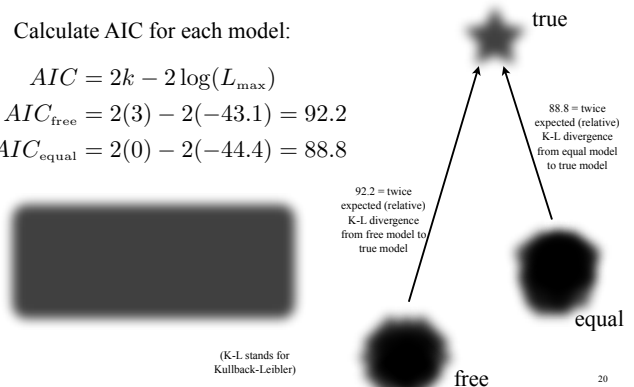
$$\begin{aligned}\log L_{\text{equal}} &= 12 \log(\pi_A) + 7 \log(\pi_C) + 7 \log(\pi_G) + 6 \log(\pi_T) \\ &= 12 \log(0.25) + 7 \log(0.25) + 7 \log(0.25) + 6 \log(0.25) \\ &= -44.4\end{aligned}$$

19

Model ranking using AIC

Calculate AIC for each model:

$$\begin{aligned}AIC &= 2k - 2 \log(L_{\text{max}}) \\ AIC_{\text{free}} &= 2(3) - 2(-43.1) = 92.2 \\ AIC_{\text{equal}} &= 2(0) - 2(-44.4) = 88.8\end{aligned}$$



20

Likelihood of the simplest tree

sequence 1 ————— sequence 2

To keep things simple, assume that the sequences are only 2 nucleotides long:

$$\begin{aligned}L &= L_1 L_2 \\ &= \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right) \right] \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \right]\end{aligned}$$

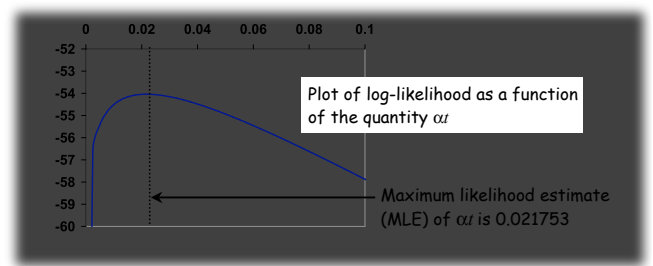
21

Maximum likelihood estimation

First 32 nucleotides of the $\psi\eta$ -globin gene of gorilla and orangutan:

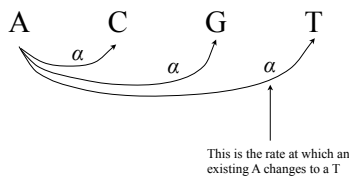
gorilla **G**A**A**GTCCTTGAGAAATAAACTGCACACACTGG
orangutan **G**G**A**CTCCTTGAGAAATAAACTGCACACACTGG

$$L = \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} + \frac{3}{4} e^{-4\alpha t} \right) \right]^{30} \left[\left(\frac{1}{4} \right) \left(\frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \right) \right]^2$$



22

number of substitutions = rate $\times$ time

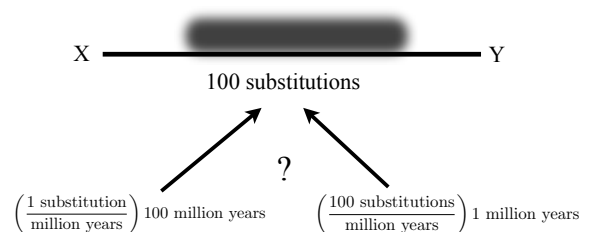


Overall substitution rate is 3α , so the expected number of substitutions (v) is

$$v = 3\alpha t$$

23

Rate and time are confounded



Later this week you will be introduced to models in which constraints on times can be used to infer rates (and vice versa), but without some extra information or constraints, sequence data allow only estimation of the **number** of substitutions.

24

A convenient convention

Because rate and time are confounded, it is convenient to arbitrarily standardize things by setting the rate to a value such that **one substitution** is expected to occur in **one unit of time** for each site.

This results in “time” (the length of a branch) being measured in units of **evolutionary distance (expected number of substitutions per site)** rather than years (or some other calendar unit).

evolutionary distance $v = 3\alpha t$

$$v = 3 \left(\frac{1}{3} \right) t \quad \text{Setting } \alpha=1/3 \text{ results in } v \text{ equalling } t$$

25

Evolutionary distances for several common models

Model	Expected no. substitutions: $v = \{r\}t$
JC69	$v = \{3\alpha\}t$
F81	$v = \{2\mu(\pi_R\pi_Y + \pi_A\pi_G + \pi_C\pi_T)\}t$
K80	$v = \{\beta(\kappa + 2)\}t$
HKY85	$v = \{2\mu[\pi_R\pi_Y + \kappa(\pi_A\pi_G + \pi_C\pi_T)]\}t$

In the formulas above, the overall rate r (in curly brackets) is a function of all parameters in the substitution model.

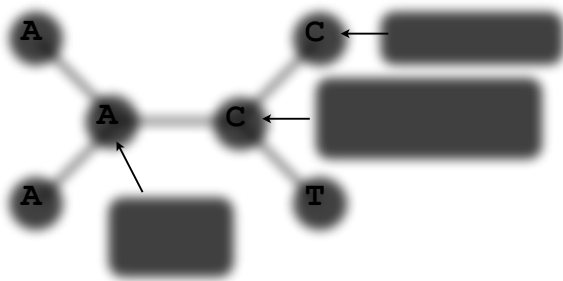
Note that one of the parameters of the substitution model can always be determined from the branch length (using our convention that $v = t$).

Typically, all other model parameters are estimated for the *entire tree* (for example, each branch uses the same value of κ)

26

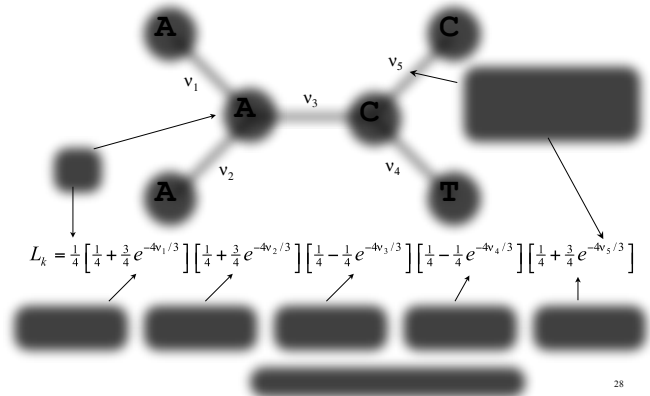
Likelihood of an unrooted tree

(data shown for only one site)



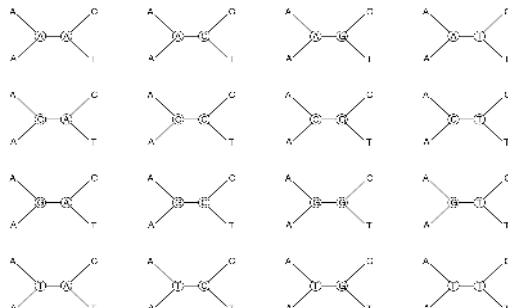
27

Likelihood for site k



28

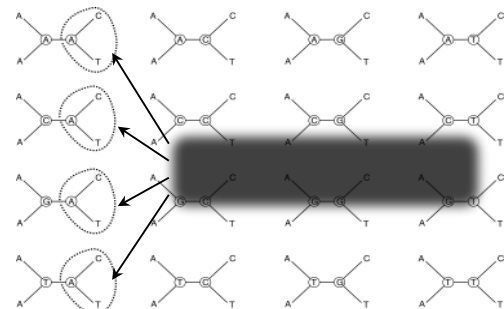
Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum them



29

Pruning algorithm

(same result, less time)



Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368-376

30

Substitution Models

31

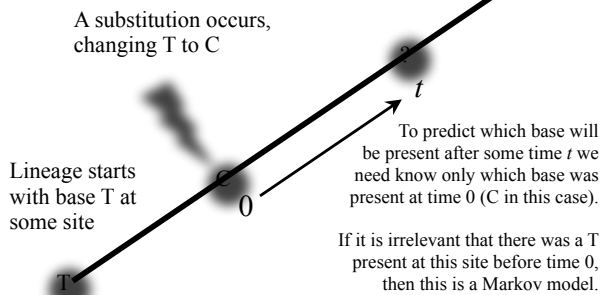
Jukes-Cantor (JC69) model

- The four bases (A, C, G, T) are expected to be **equally frequent** in sequences ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$)
- Assumes **same rate** for all types of substitution ($r_{A \rightarrow C} = r_{A \rightarrow G} = r_{A \rightarrow T} = r_{C \rightarrow G} = r_{C \rightarrow T} = r_{G \rightarrow T} = \alpha$)
- Usually described as a **1-parameter** model (the parameter being the branch length)
 - Remember, however, that each branch in a tree can have its own length, so there are really as many parameters in the model as there are edges in the tree!
- Assumes substitution is a **Markov** process...

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 in H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

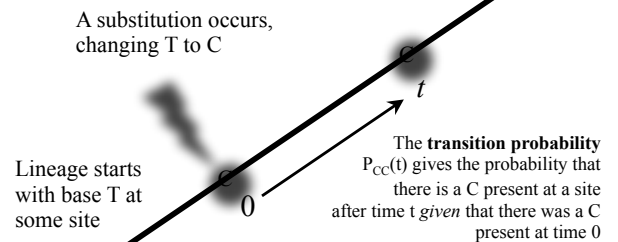
32

What is a Markov model?



33

Transition Probabilities



Note: the term *transition* here comes from the terminology of stochastic processes and refers to any change of state (and even non-changes!). If this kind of transition represents a change from one nucleotide state to a different nucleotide state, it could thus be either a transition-type or a transversion-type substitution.

34

Jukes-Cantor transition probabilities

Here is the probability that a site starting in state T will end up in state G after time t when the individual substitution rates are all α :

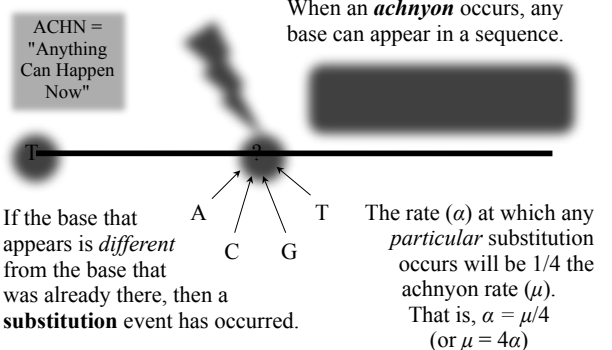
$$P_{TG}(t) = \frac{1}{4} (1 - e^{-4\alpha t})$$

The JC69 model has only one unknown quantity: αt

(The symbol e represents the base of the natural logarithms: its value is 2.718281828459045...)

35

"ACHNyons" vs. substitutions



36

Deriving a transition probability

Calculate the probability that a site currently T will change to G over time t when the rate of this particular substitution is α :

$$\Pr(\text{zero achnyons}) = e^{-\mu t} \quad (\text{Poisson probability of zero events})$$

$$\Pr(\text{at least 1 achnyon}) = 1 - e^{-\mu t}$$

$$\Pr(\text{last achnyon results in base G}) = \frac{1}{4}$$

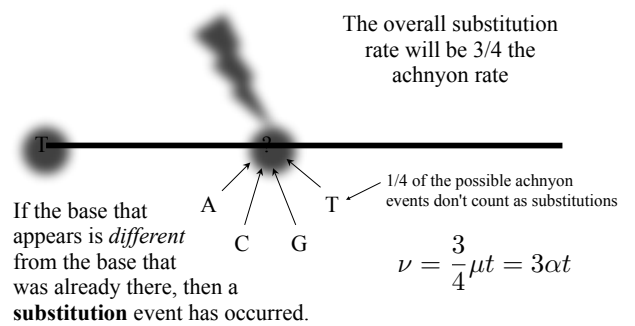
$$\Pr(\text{end in G} \mid \text{start in T}) = \frac{1}{4} (1 - e^{-\mu t})$$

Remember that the rate (α) of any particular substitution is one fourth the achnyon rate (μ):

$$P_{GT}(t) = \frac{1}{4} (1 - e^{-4\alpha t})$$

37

Expected number of substitutions



38

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TC}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TG}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TT}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$1 - e^{-4\alpha t}$$

39

Transition Probabilities: Remarks

$$P_{TA}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TC}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TG}(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

$$P_{TT}(t) = \frac{1}{4}(1 - e^{-4\alpha t}) + e^{-4\alpha t}$$

40

Equilibrium frequencies

- The JC69 model assumes that the frequencies of the four bases (A, C, G, T) are equal
- The equilibrium relative frequency of each base is thus 0.25
- Why are they called *equilibrium* frequencies?

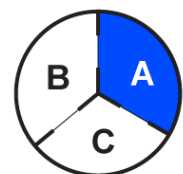
41

Equilibrium Frequencies

Imagine a bottle of perfume has been spilled in room A.

The doors to the other rooms are closed, so the perfume has, thus far, not been able to spread.

What would happen if we opened all the doors?

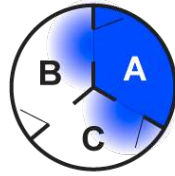


42

Equilibrium Frequencies

If the doors are suddenly opened, the perfume would begin diffusing from the area of highest concentration to lowest.

Molecules of perfume go both ways through open doors, but more pass one way than another, leading to a net flow from room A to rooms B and C.



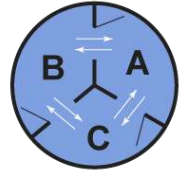
In the instant that the doors are opened, A is losing perfume molecules at *twice the rate* each of the other rooms is gaining molecules. As diffusion progresses, however, the rate of loss from A drops, approaching an equilibrium.

43

Equilibrium Frequencies

Eventually, all four rooms have essentially the same concentration of perfume.

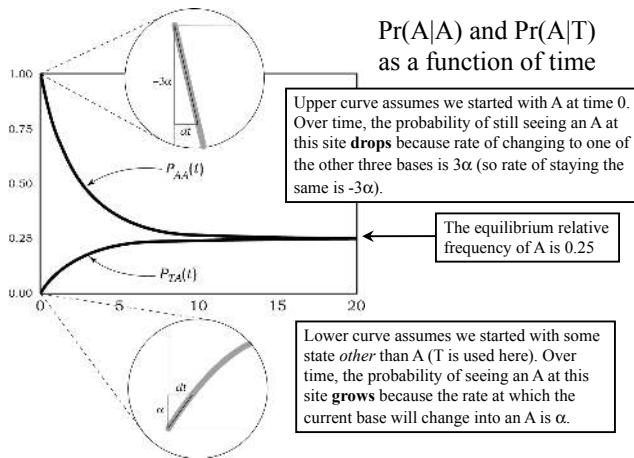
Molecules still move through doors, but now the rates are the same in all directions.



Back to sequence evolution: assume a sequence began with only A nucleotides (a poly-A sequence). Over time, substitution would begin converting some of these As to Cs, Gs, and Ts, just as the perfume diffused into adjacent rooms.

44

Pr(A|A) and Pr(A|T) as a function of time



45

JC69 rate matrix

		To			
		A	C	G	T
From	A	-3α	α	α	α
	C	α	-3α	α	α
	G	α	α	-3α	α
	T	α	α	α	-3α

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21-132 in H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York.

46

K80 (or K2P) rate matrix

		To			
		A	C	G	T
From	A	$-\alpha - 2\beta$	β	α	β
	C	β	$-\alpha - 2\beta$	β	α
	G	α	β	$-\alpha - 2\beta$	β
	T	β	α	β	$-\alpha - 2\beta$

transition rate transversion rate

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

47

K80 rate matrix (looks different, but actually the same)

		A	C	G	T
From	A	$-\beta(\kappa + 2)$	β	$\kappa\beta$	β
	C	β	$-\beta(\kappa + 2)$	β	$\kappa\beta$
	G	$\kappa\beta$	β	$-\beta(\kappa + 2)$	β
	T	β	$\kappa\beta$	β	$-\beta(\kappa + 2)$

All I've done is re-parameterize the rate matrix, letting κ equal the *transition/transversion rate ratio* $\rightarrow \kappa = \frac{\alpha}{\beta}$

48

49

50

51

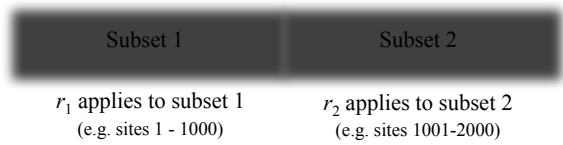
■

52

—

Site-specific rates

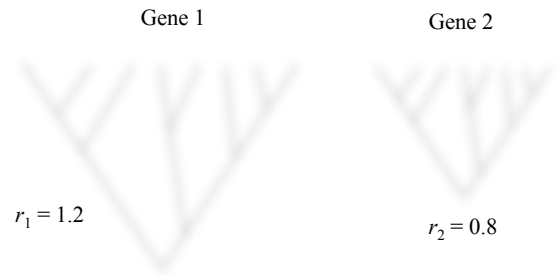
Each defined subset (e.g. gene, codon position) has its own relative rate



55

Site-specific rates

$$L = \underbrace{\Pr(D_1|r_1) \cdots \Pr(D_{1000}|r_1)}_{\text{Gene 1}} \underbrace{\Pr(D_{1001}|r_2) \cdots \Pr(D_{2000}|r_2)}_{\text{Gene 2}}$$



56

Site-specific rates

JC69 transition probabilities that would be used for every site if rate *homogeneity* were assumed:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

57

Site specific rates

JC69 transition probabilities that would be used for sites in **gene 1**:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4r_1\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_1\alpha t}$$

JC69 transition probabilities that would be used for sites in **gene 2**:

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4r_2\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4r_2\alpha t}$$

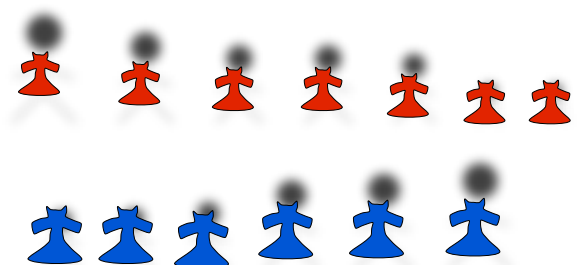
58

Site-specific Approach



59

Site-specific Approach

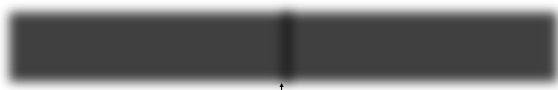


Pro: costs less: need to buy just one coat for every person
Con: every person in a group has to wear the same size coat, so the fit will be poor for some people if they are much bigger or smaller than the average size for the group in which they have been placed

60

Mixture Models

All relative rates applied to every site



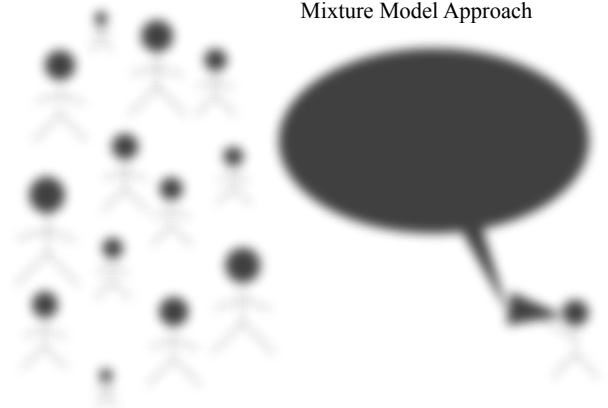
↑
site i

$$L_i = \Pr(D_i|r_1) \Pr(r_1) + \Pr(D_i|r_2) \Pr(r_2)$$

Common examples { Invariable sites (I) model
Discrete Gamma (G) model

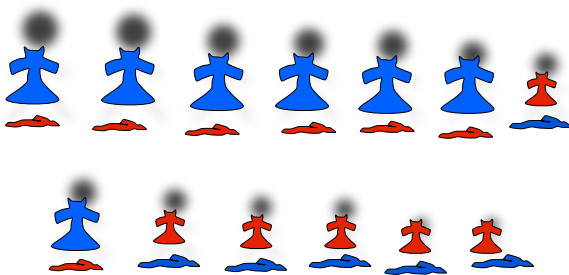
61

Mixture Model Approach



62

Mixture Model Approach

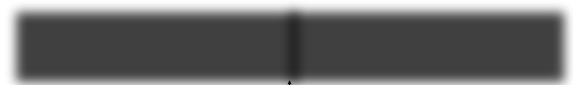


Pro: every person experiences better fit because they can choose the size coat that fits best
Con: costs more because two coats must be provided for each person

63

Invariable Sites Model

A fraction p_{invar} of sites are assumed to be invariable (i.e. rate = 0.0)



↑
site i

$$L_i = \Pr(D_i|r_1)p_{invar} + \Pr(D_i|r_2)(1 - p_{invar})$$

$$r_1 = 0.0$$

$$r_2 = \frac{1}{1 - p_{invar}}$$

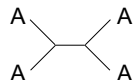
Allows for the possibility that any given site could be variable or invariable

Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution* 35:17-31.

64

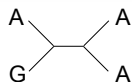
Invariable sites model

If site i is a *constant* site, both terms will contribute to the site likelihood:



$$L_i = \Pr(D_i|0.0)p_{invar} + \Pr(D_i|r_2)(1 - p_{invar})$$

If site i is a *variable* site, there is no way to explain the data with a zero rate, so the first term is zero:

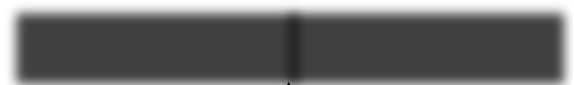


$$L_i = \cancel{\Pr(D_i|0.0)p_{invar}} + \Pr(D_i|r_2)(1 - p_{invar})$$

65

Discrete Gamma Model

No relative rates are exactly 0.0, and all are equally probable



↑
site i

$$L = \left(\frac{1}{4}\right) \Pr(D_i|r_1) + \left(\frac{1}{4}\right) \Pr(D_i|r_2) + \left(\frac{1}{4}\right) \Pr(D_i|r_3) + \left(\frac{1}{4}\right) \Pr(D_i|r_4)$$

Relative rates are constrained to a discrete gamma distribution
Number of rate categories can vary (4 used here)

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10:1396-1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39:306-314.

66

Interpreting codon model results

$\omega = \beta/\alpha$ is the nonsynonymous/synonymous rate ratio

omega	mode of selection	example(s)
$\omega < 1$	stabilizing selection (nucleotide substitutions rarely change the amino acid)	functional protein coding genes
$\omega = 1$	neutral evolution (synonymous and nonsynonymous substitutions occur at the same rate)	pseudogenes
$\omega > 1$	positive selection (nucleotide substitutions often change the amino acid)	envelope proteins in viruses under active positive selection