Lecture Handout 24: MOLECULAR EVOLUTION AND PHYLOGENETICS 2

Adaptations from:
Page and Holmes "Molecular Evolution: Phylogenetic Approach" © Blackwell Science 1998
Terry Brown "Genomes 3" © Garland Science 2007
Phillip Benfey and Alexander Protopapas. "Genomics" © Pearson / Prentice Hall 2007
Thanks also to Taras Oleksyk

Gene trees vs. species trees:
Mutation vs. speciation

- Mutation can precede speciation
  - Both alleles present in the population before the split
  - Random drift or selection could fix frequencies in the two populations
  - Mutation can also follow the isolation event

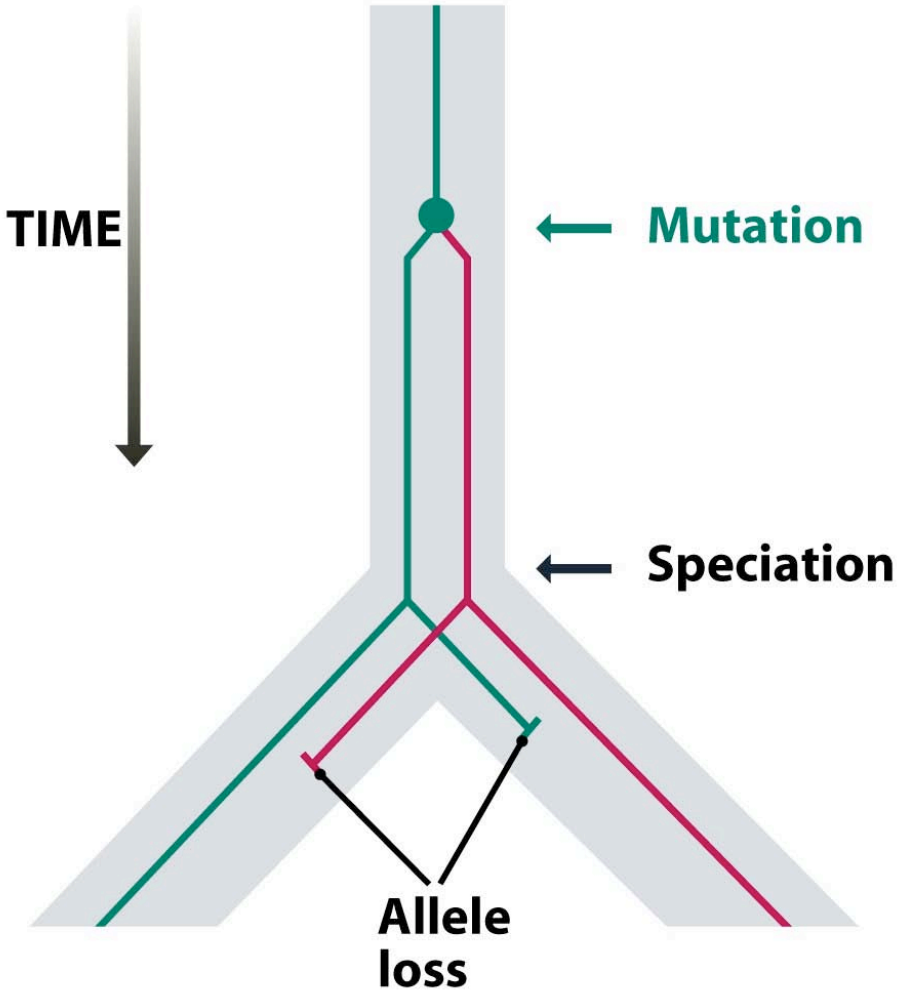


Figure 19.8 *Genomes 3* (© Garland Science 2007)

Gene trees vs. species trees:
Example: incongruent lineage sorting

- A gene tree can be different from a species tree
- If mutation precedes speciation it could also give an incorrect time for a speciation event if a **molecular clock** is used

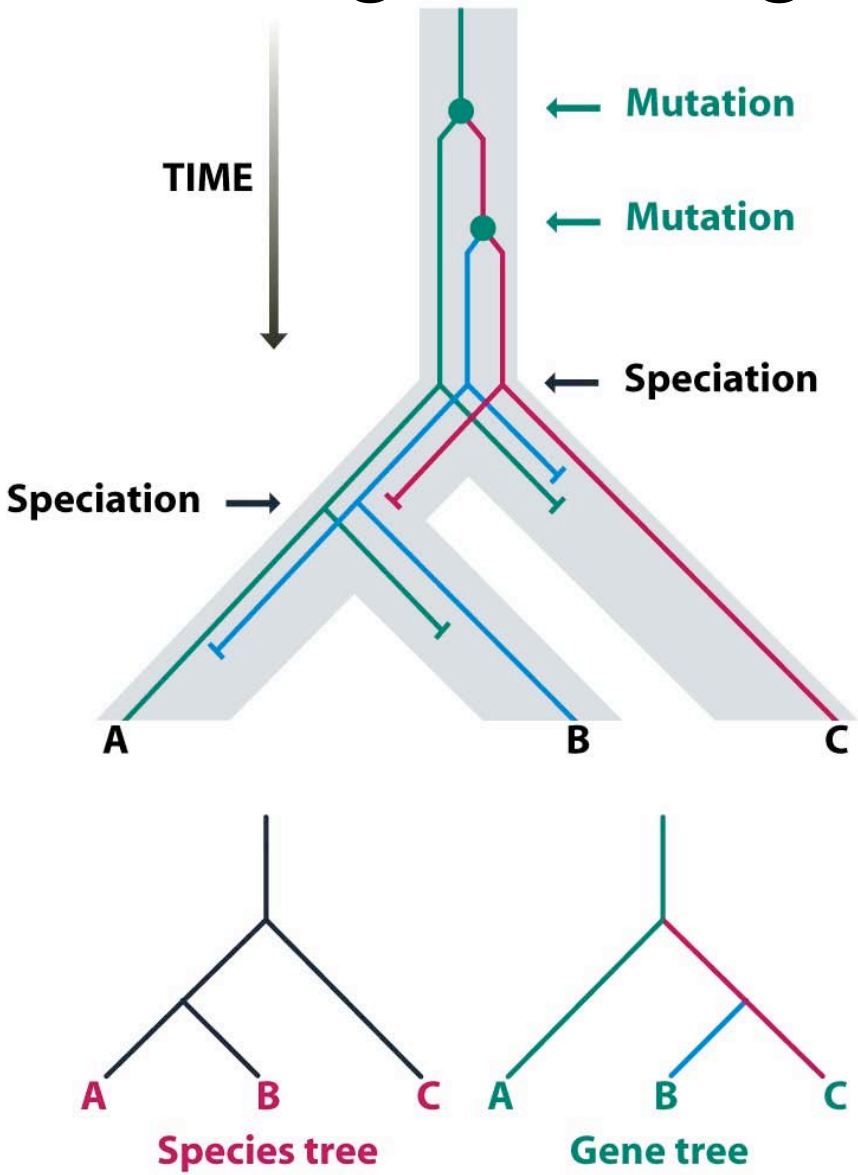
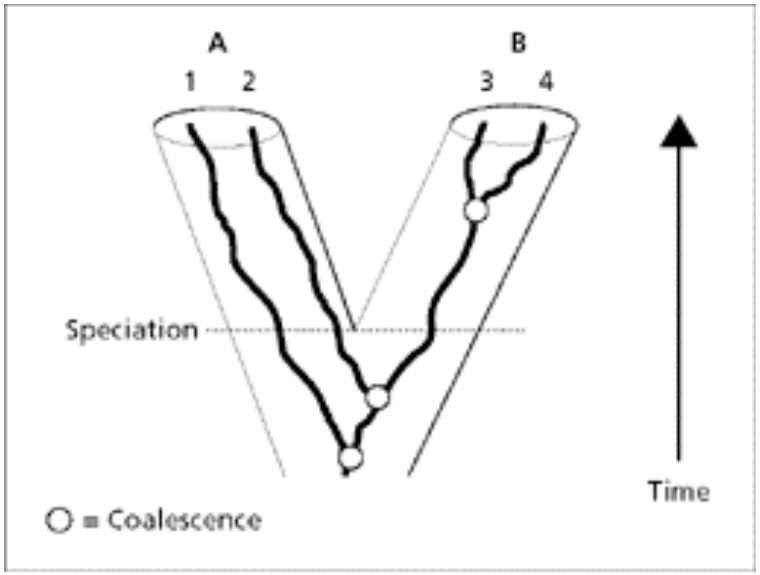


Figure 19.9 *Genomes 3* (© Garland Science 2007)

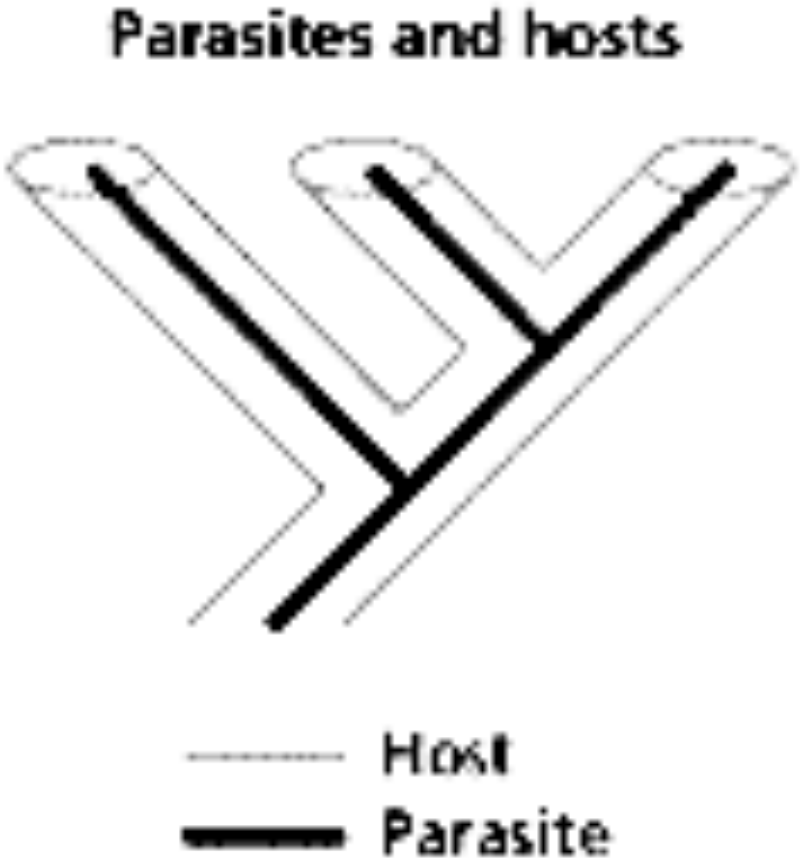
Lineage sorting:
Coalescence

- The most recent common ancestor:
  - **Coalescent point**
  - **Coalescent time** – the time at which the most recent common ancestor occurs
- Does not necessarily coincide with speciation
- Examples…



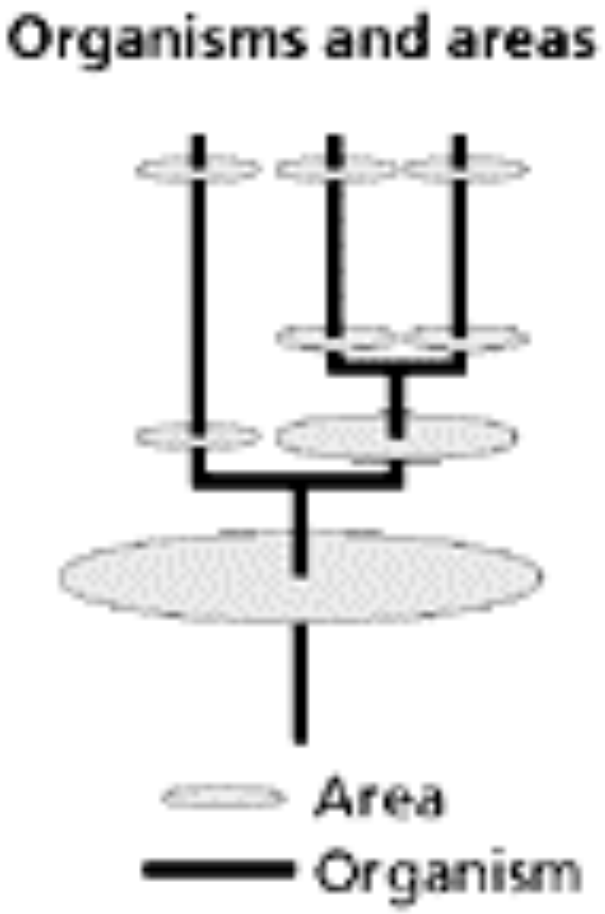
Page, Holmes
Molecular Evolution

Historical associations



- Some hosts & their parasites (including viruses) may have a long evolutionary history of close association reflected in their evolutionary trees

Historical associations



- On the larger scale, organisms may track ecological events or geological history
  - Such as continental breakup & drift

Cladistics & classification

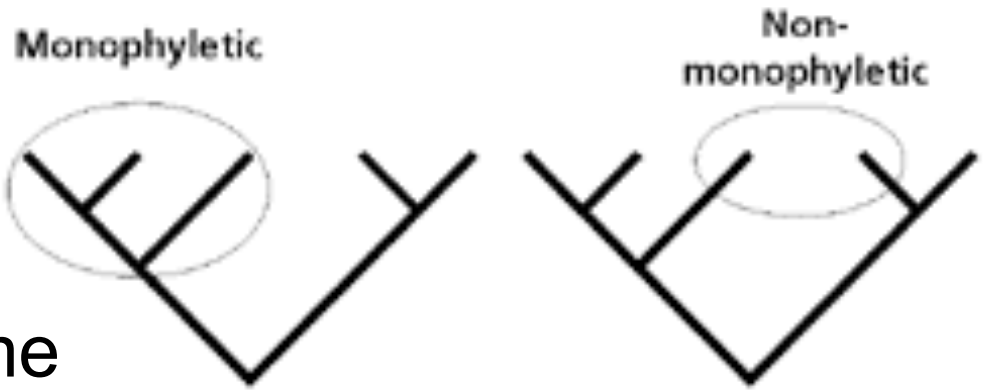
- Phylogenetics forms the basis of taxonomy
  - The formal naming of groups of organisms

- Cladistic classification recognize only monophyletic groups …

Monophyly

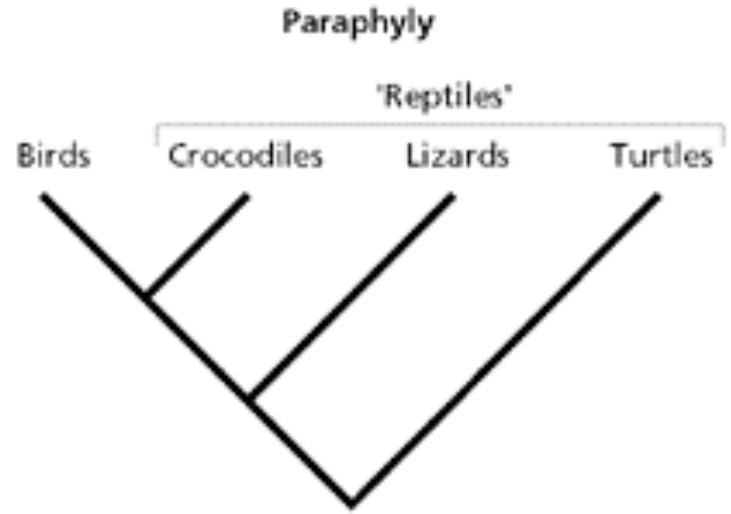
- **A clade** –all the sequences descended from a common ancestral sequence or node

- **Monophyletic clade** – when all members of a clade belong to the same taxon
  - In a non-monophyletic group, one or more descendants are not included

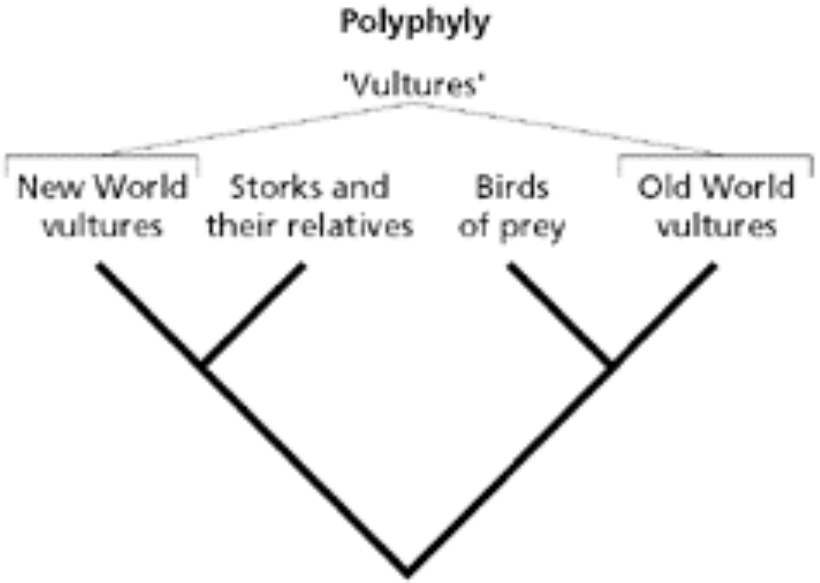


Non-monophyletic groups

- **Paraphyletic group** – a group of sequences or taxa that excludes some members of a clade (one ancestor)



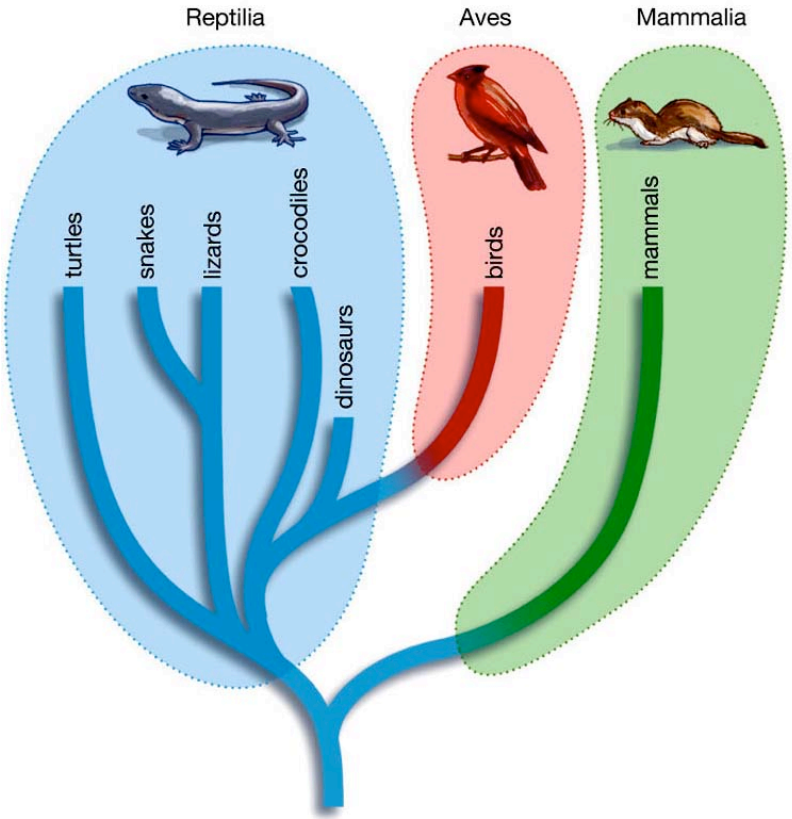
- **Polyphyletic group** – a group of sequences that derive from two or more distinct ancestral sequences



Non-monophyletic groups: Paraphyly

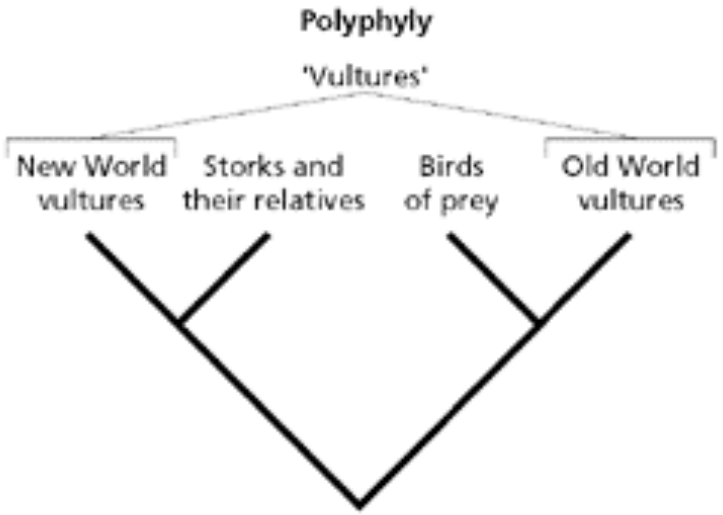
- Groupings are based on shared primitive characters (**plesiomorphies** and exclude members that have **autapomorphies**

- ‘Reptiles’ exclude birds because of their novel anatomy



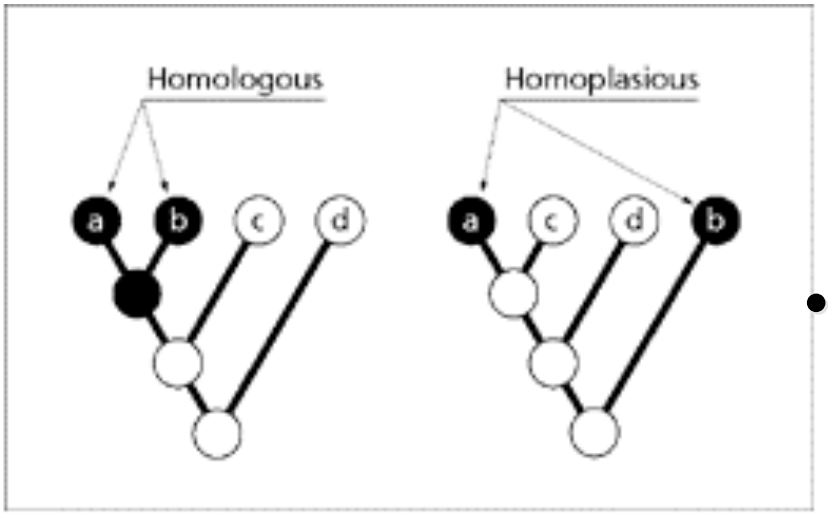
Non-monophyletic groups: Polyphyly

- Groupings are typically assemblages of taxa that have been erroneously grouped on the basis of convergent characters



- 'Vultures' are a polyphyletic grouping comprising two groups of birds that have independently evolved similar morphology and habits from different ancestors

Homology & Homoplasy
Examples



b Page, Holmes
Molecular Evolution

- **Homologous** (on the left): because inherited it directly from the ancestor that also had a “shaded” allele
- **Homoplasious** (on the right): where “shaded” allele evolved independently from two ancestors that had “unshaded” alleles

Birds & Bats

- **Homology** can depend on the aspect of the feature:
  - **Yes:** Both birds & bats inherited forearms from the last common ancestor (so forearms are homologous)
  - **No:** In both groups the forearms have been modified for flight independently (so wings are not homologous)

Homology

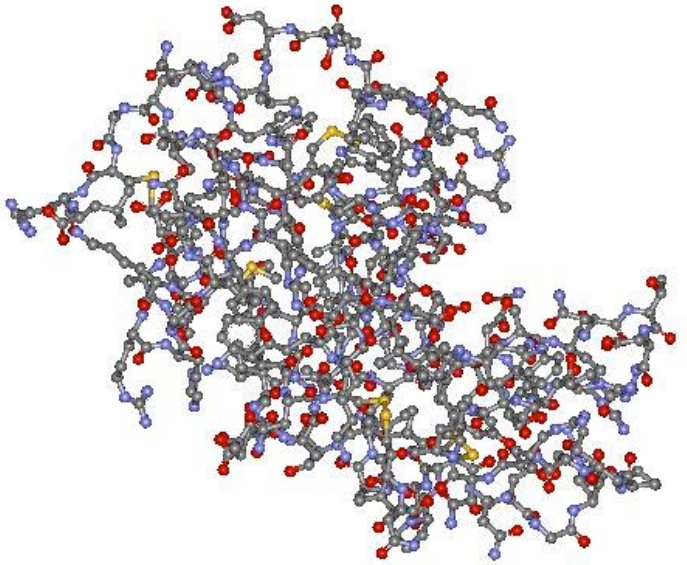
Structure vs. functionality

- **Yes:** Two proteins in two different organisms may be encoded by the same gene
- **Yes:** These two genes may share amino acids in common and even have similar function

- **No:** But if the functionality is acquired independently, then the functionality is ***homoplasious***

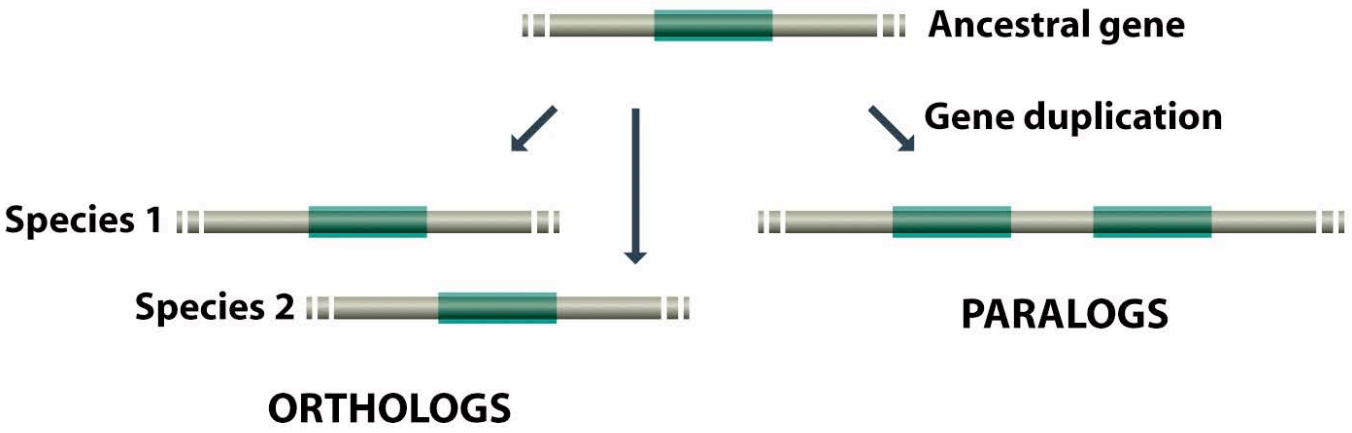
Homology
At the molecular level

- Parallel evolution of amino acids in lysozyme enzyme sequence of **langur** monkeys & cows



(http://lysozyme.co.uk

Basic types of homology



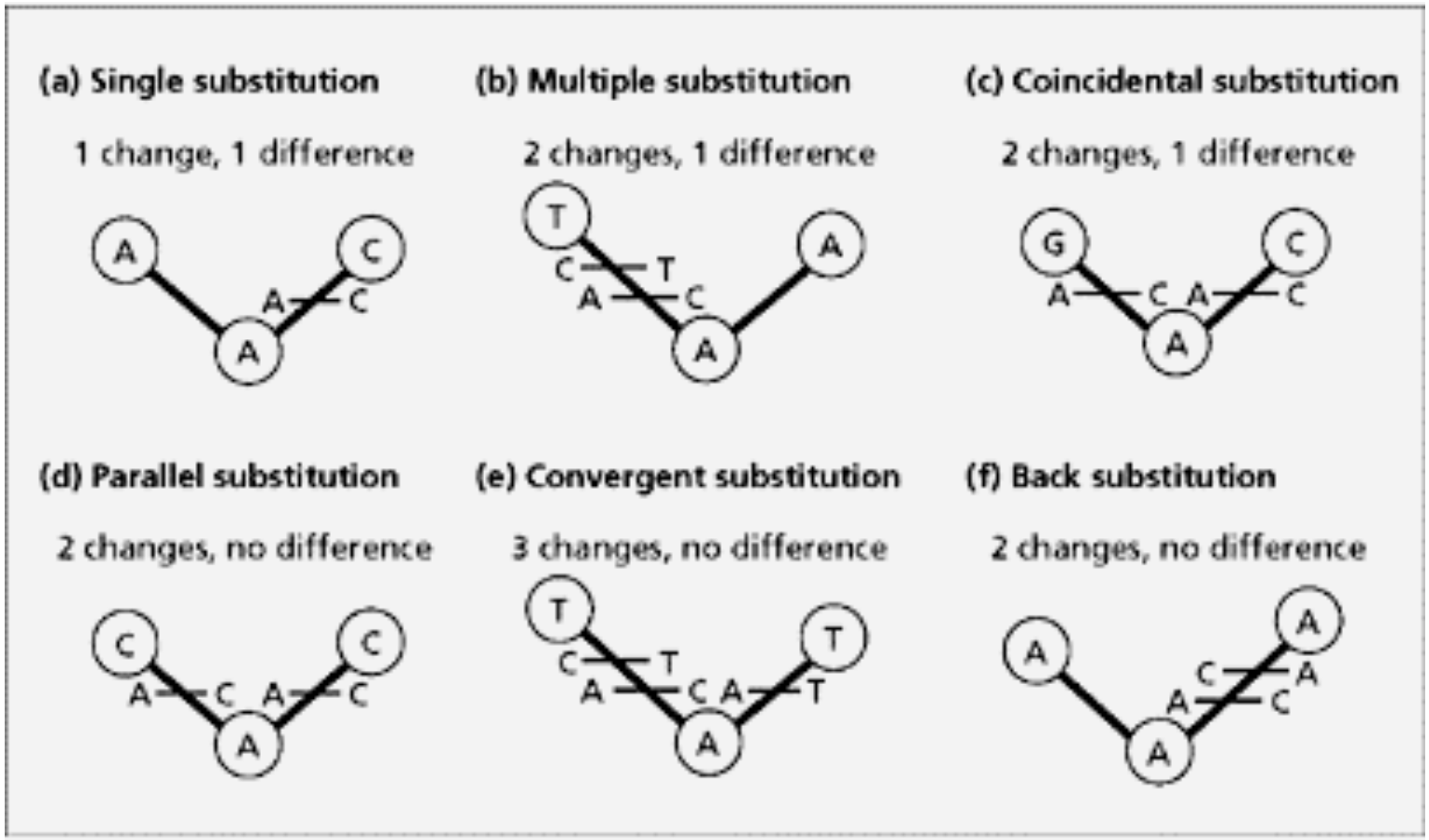
- **Orthology** – homology that arises via speciation
- **Paralogy** – homology that arises via gene duplication

- **Xenology** – gene owes its presence from the transfer from another organism

Genetic distance

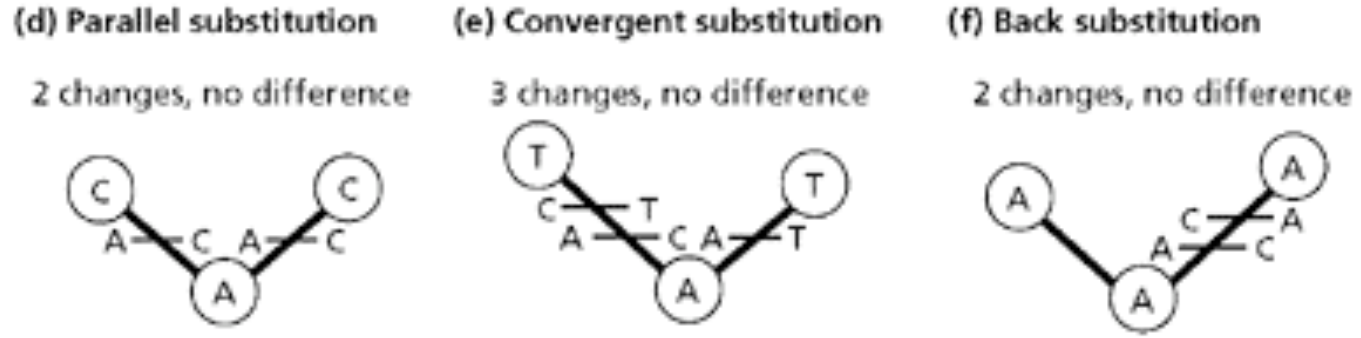
- DNA segments are not very informative about their evolutionary history
  - Sequences are either similar or not
  - For any given site, maximum number of differences detectable is one
  - There are only four states: A,G,C and T
- However there are other complications:
  - What if there were **more than one event** at the same site?
  - What if there are **different frequencies** for different events?

Types of substitutions



Page, Holmes
Molecular Evolution

Types of substitutions
Homoplasious similarity



- In these three cases nucleotides are identical in both descendant sequences, but the resulting alleles are not inherited from the ancestral sequence
- Potentially, these have much more serious implications:
  - Homoplasy can obscure the actual number of evolutionary events

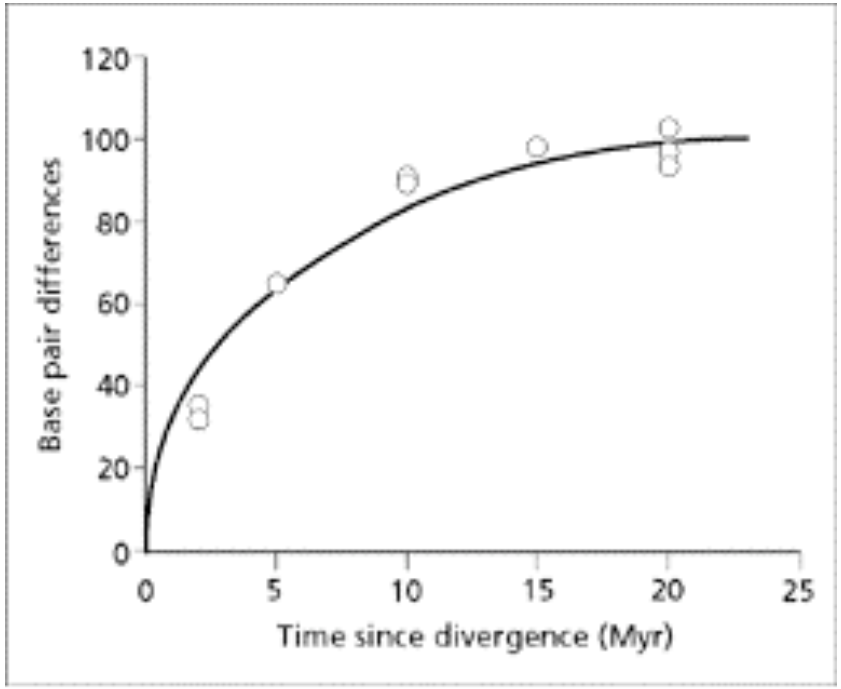
Distance measures

Observed distances

- The simplest measure of distance between two nucleotide sequences
- However, for all but very similar sequences this is a poor measure:
  - Since the same site can undergo multiple substitutions
  - As time goes by, the number of differences becomes less and less of an accurate estimator

Observed distances

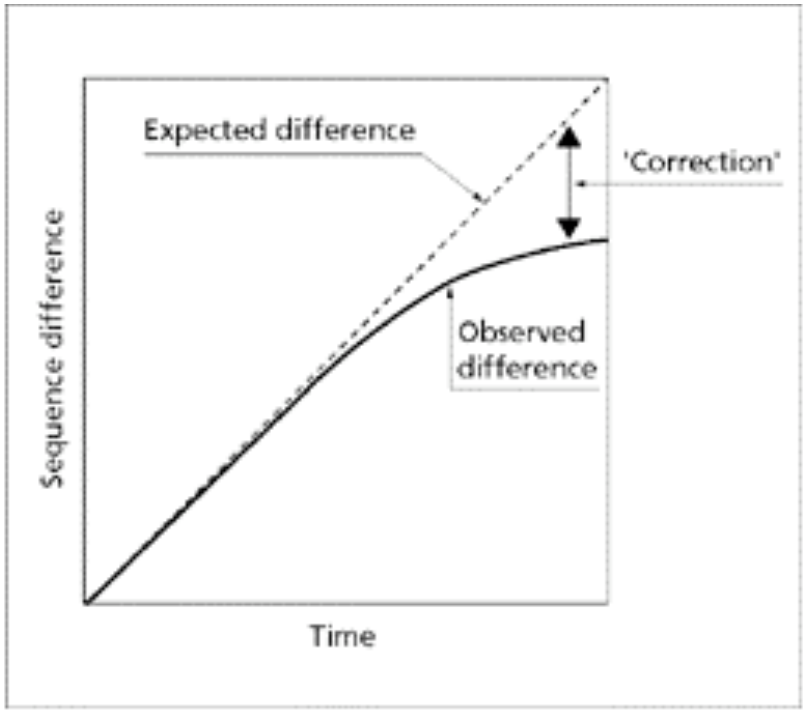
Example



Page, Holmes
Molecular Evolution

- Number of nucleotide differences between mammalian (bovid) mitochondrial sequences vs. time of divergence
- **Note:** the observed number of substitutions is not linear

Distance correction



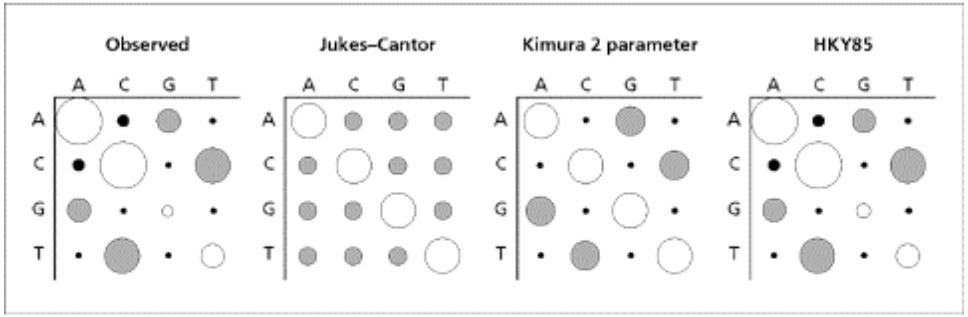
Page, Holmes
Molecular Evolution

- As more substitutions accumulate they become **saturated**
  - Most of the sites that are changing have changed before
- The goal of **distance correction methods** is to recover the “overprinted” amount of evolutionary changes

Distance correction

Best model

- Observed and expected numbers of nucleotide pairs between human and chimp mtDNA sequences for three different models
  - As the models add parameters they start resembling the observed pattern closer and closer

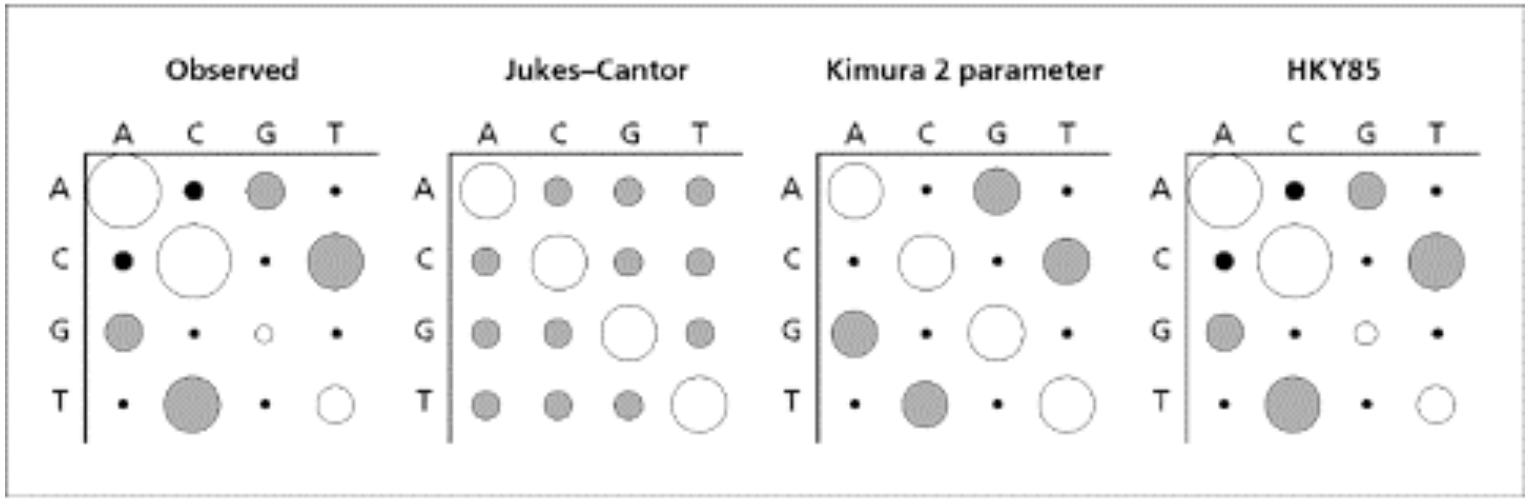


b Page, Holmes
Molecular Evolution

- How can we evaluate the best fit between each model and the data?

Choosing the model

Models of evolution:
Which model fits the data best?
With the least number of parameters?



Page, Holmes
Molecular Evolution

Distance correction
General framework

$$P_t = \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{AA} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

- In the general framework:
  - Probability of a given nucleotide substitution remains constant over time
  - Base composition of sequences is at equilibrium
- p_{AC} the probability that **A** changed to **C** at a site during time **t**

Distance correction

Jukes-Cantor (JC)

$$P_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}$$

$$f = [¼, ¼, ¼, ¼]$$

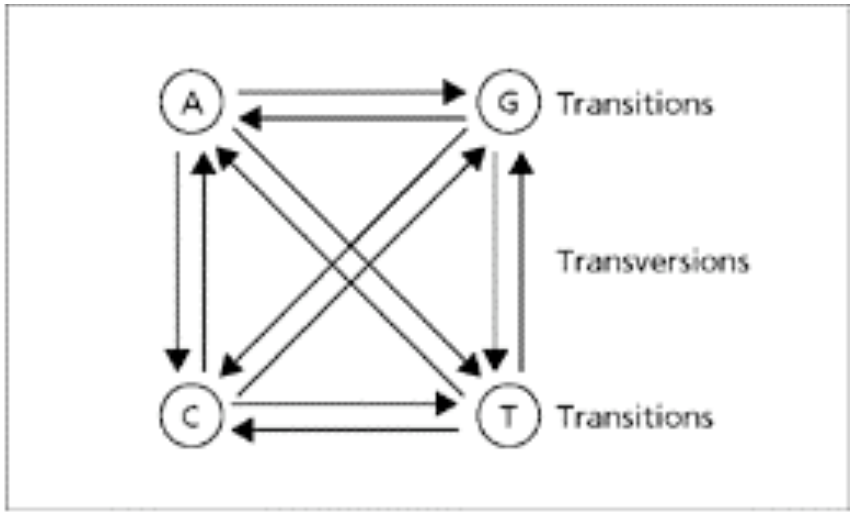
- **JC** model assumes that the four bases have equal frequencies & all substitutions are equally likely
- Distance between two sequences is given as:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

- p is proportion of nucleotides that are different in two sequences
- Very simple model, few parameters

Distance correction

Unequal substitution rates

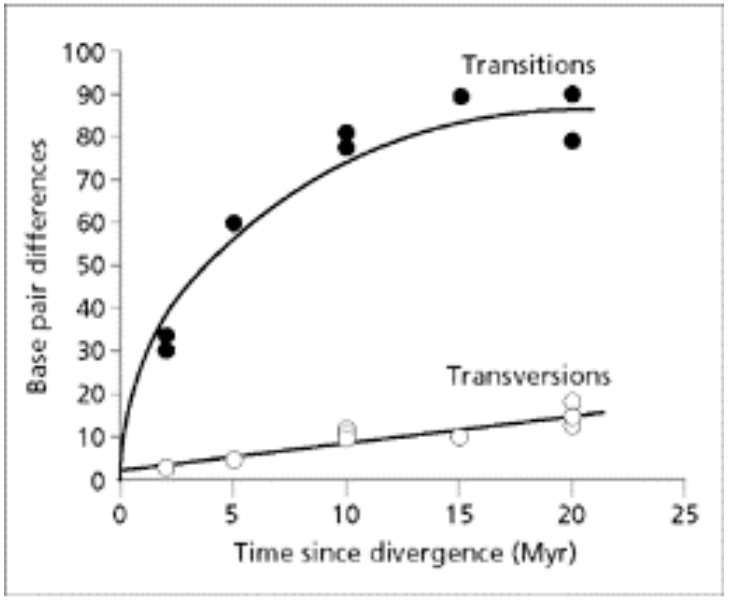


$S=(2\beta+\alpha)$

- One may expect transversions to be more common than transitions
  - There is one possible transition per two possible transversions
- However, transitions are much more common, especially for mtDNA

Distance correction

Unequal substitution rates



b Page, Holmes
Molecular Evolution

- Transitions accumulate faster than transversions & become saturated, whereas transversions accumulate more slowly & show no evidence of saturation

- Figure: The number of transitions and transversions in bovid mtDNA

Distance correction
Kimura 2 Parameter (K2P)

$$P_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}$$

$$f = [¼, ¼, ¼, ¼]$$

- **K2P** model corrects for the unequal rate of substitution
- Distance between two sequences is given as:

$$d = \tfrac{1}{2}\ln\left(\frac{1}{(1-2P-Q)}\right) + \tfrac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

- ***P & Q*** *are the proportional differences between the two sequences due to transitions and transversions*

Distance correction

General Reversible Model (REW)

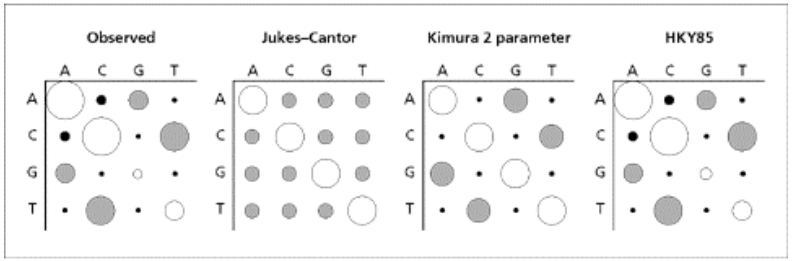
$$P_t = \begin{bmatrix} . & \pi_C a & \pi_G b & \pi_T c \\ \pi_A a & . & \pi_G d & \pi_T e \\ \pi_A b & \pi_C d & . & \pi_T f \\ \pi_A c & \pi_C e & \pi_G f & . \end{bmatrix}$$

$$f = [\pi_A \; \pi_C \; \pi_G \; \pi_T]$$

- Allows all six pairs of substitutions to have different rates; nucleotide frequencies to be different
- More general than any model, and therefore can be reduced to any model
- Very complex model; many parameters

Distance correction
Best model

- Observed and expected numbers of nucleotide pairs between human and chimp mtDNA sequences for three different models
  - As the models add parameters they start resembling the observed pattern closer and closer



b Page, Holmes
Molecular Evolution

- How can we evaluate the best fit between each model and the data?

Problem

Precision vs. accuracy

- **Precision:** how many alternatives are excluded:
  - A method that excludes all but one tree is very precise
- **Accuracy:** how close your tree is to the true tree
  - If the one tree found earlier is not the true tree, your method is inaccurate

More parameters: more accurate but less precise

Best model

Likelihood (e.g., *Modeltest* program)

- The probability of observing the data given a particular model
  - If you are flipping a coin and you get 1/100 as heads, that is a very unlikely outcome, since you expect about 50/50
- Given a model that specifies probabilities of observing various events, likelihood is

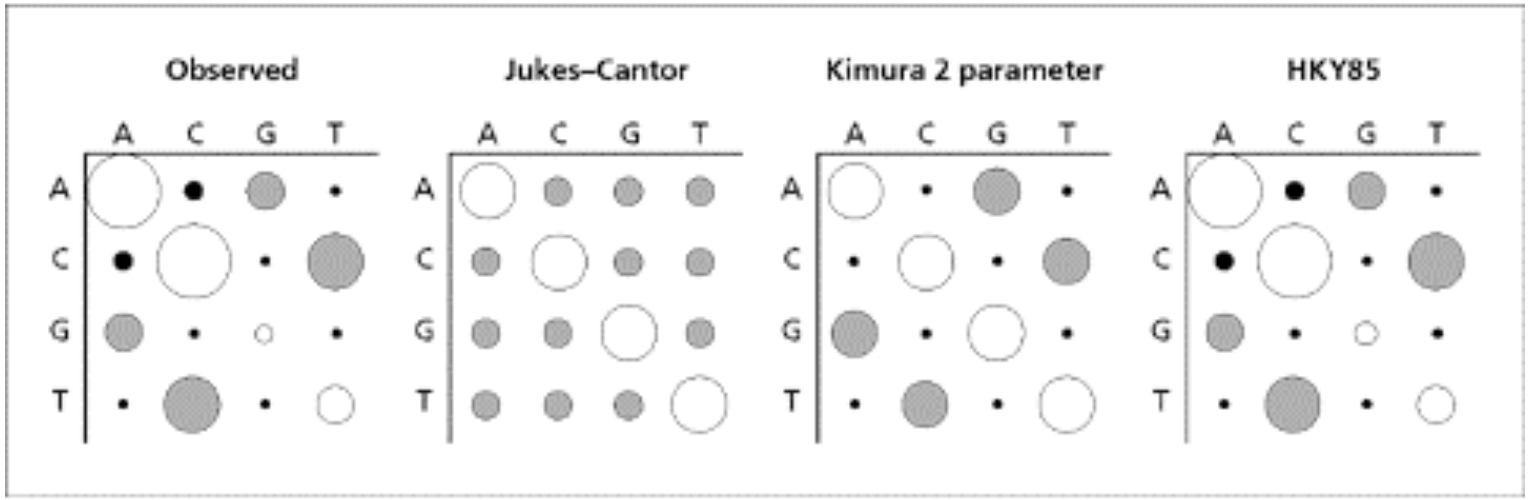
$$L = \Pr(D|H)$$

- **D** = data & **H**=hypothesis
- **Log likelihood:** Usually since L is very small, likelihoods are expressed as ln(L)

Choosing the model

Likelihood

- ***Theoretical best = 2,064.80***
- **JC** ln L = -2,691.76
- **K2P** ln L = 2,424.79
- **HKY85 ln L = -2,075.41***



Page, Holmes
Molecular Evolution

**** Best model resembles theoretical best lnL***