Concept of molecular clock

Evolution Lectures 4

Dengue

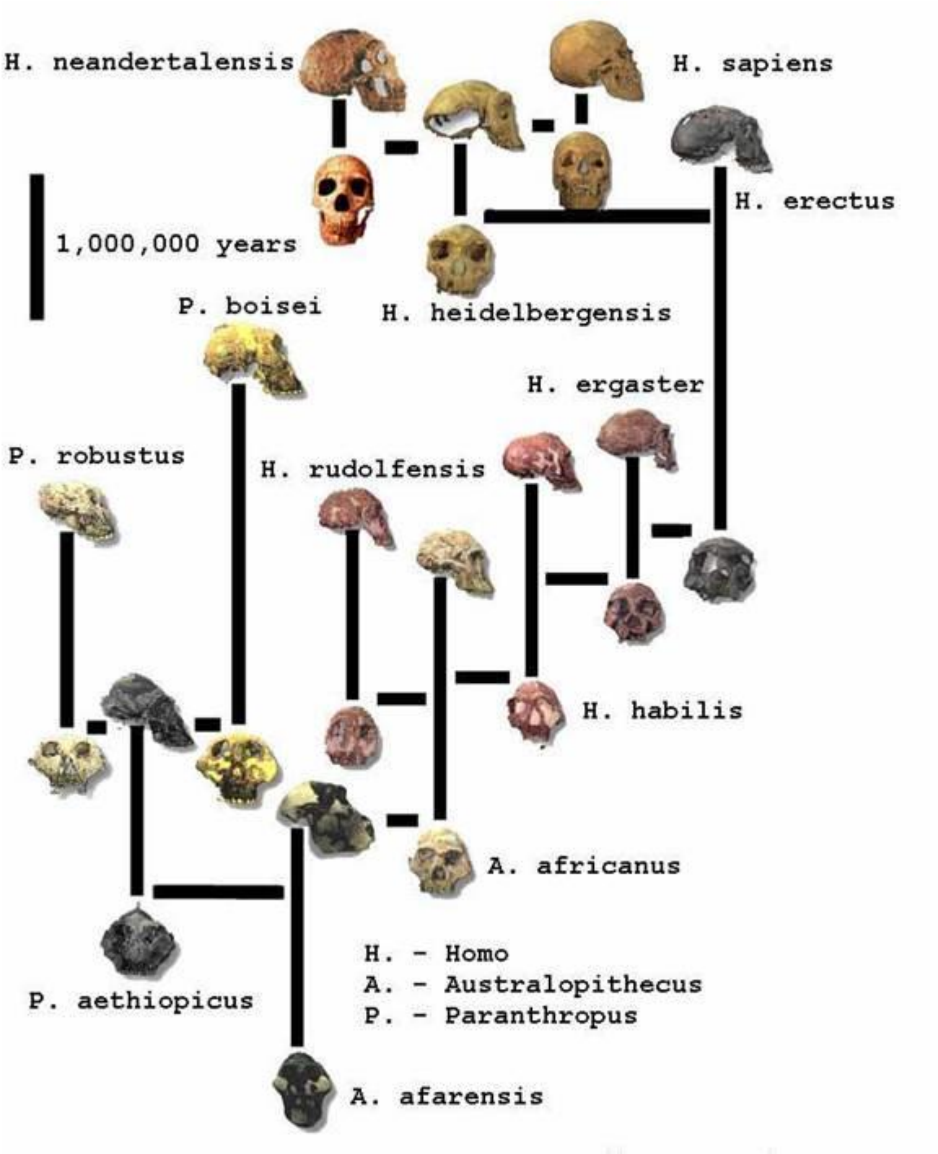
- More than 2.5 billion people – over 40% of the world’s population – are now at risk from dengue. WHO currently estimates there may be 50–100 million dengue infections worldwide every year (WHO 2014)
- The agent of these diseases, dengue virus, is an RNA virus with a positive-sense genome of approximately 11 kb, belonging to the genus Flavivirus and existing as four genetically distinct serotypes (DEN-1 to DEN-4)
- Our analysis reveals that dengue virus generally evolves according to a <u>molecular clock</u>, ………, with the virus appearing approximately 1000 years ago. Furthermore, we estimate that the zoonotic transfer of dengue from sylvatic (monkey) to sustained human transmission occurred between 125 and 320 years ago
  - Mol. Biol. Evol. 20(1):122–129. 2003

Science at work

- Zuckerkandl, E. & Pauling, L. in *Horizons in Biochemistry* (eds Kasha, M. & Pullman, B.) 189–225 (Academic Press, New York, 1962).
  - plotted the numbers of amino acid differences between different proteins (mostly hemoglobin) of a number of organisms against the estimated time back to the common ancestor of those organisms
  - An informal proposal of molecular clock
- Margoliash, E. Primary structure and evolution of cytochrome *c*. *Proc. Natl Acad. Sci. USA* **50**, 672–679 (1963).
  - A formal proposal of molecular clock

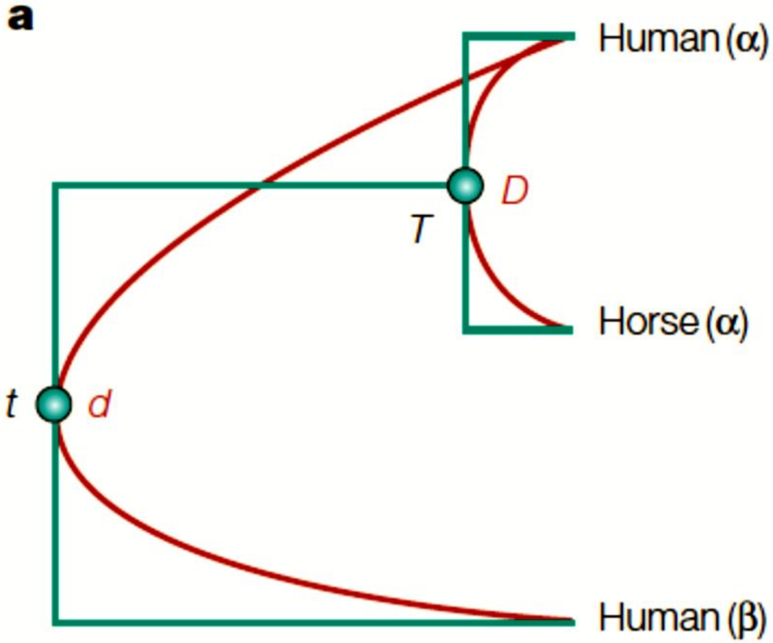
Nut shell

- In 60s
  - We get to know that the sequence of proteins are similar but not exact
  - But we can track how many changes have been made
  - We know (from the fossil evidence) who came from who
  - We also know (again fossil evidence and radiometric analysis) what is last know common ancestor
    - *H. heidelbergensis* for Neanderthal and Sapiens



The first proposal

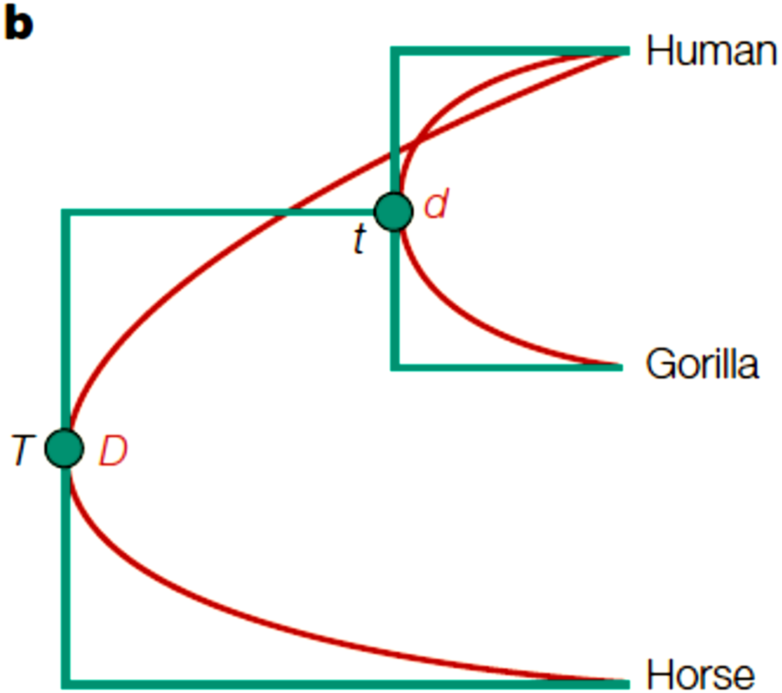
- Divergence of four members of the haemoglobin gene family (α, β, γ and δ)
- number of observed sequence differences (*D*) between the horse and human α-haemoglobin proteins and the divergence time between the two species (*T*)



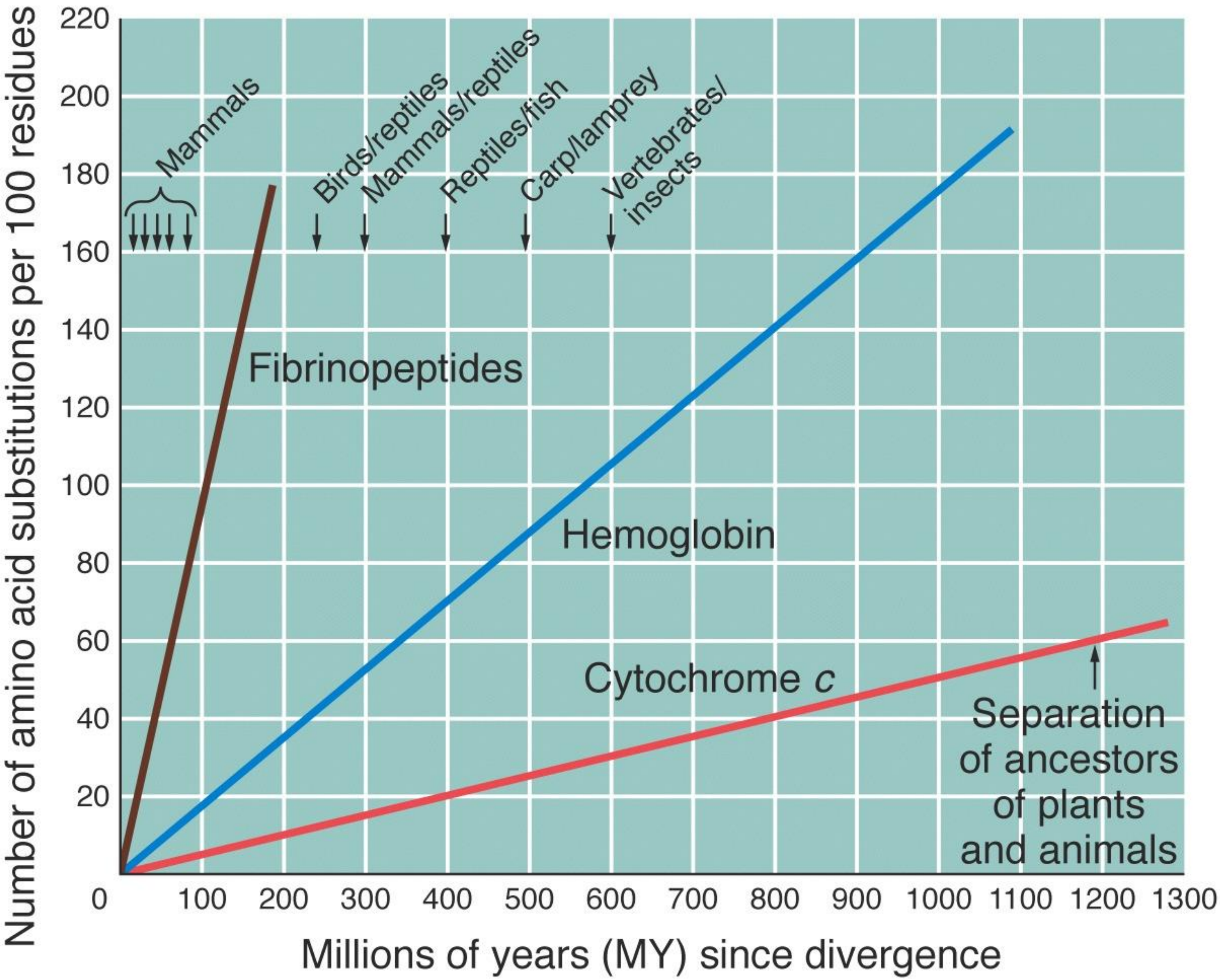
- The molecular-clock calibration was carried out by dividing twice the known divergence time by the amount of sequence divergence (2T/D)
- They calculated the molecular-clock calibration to be 11 to 18 million years per amino-acid substitution

The second proposal

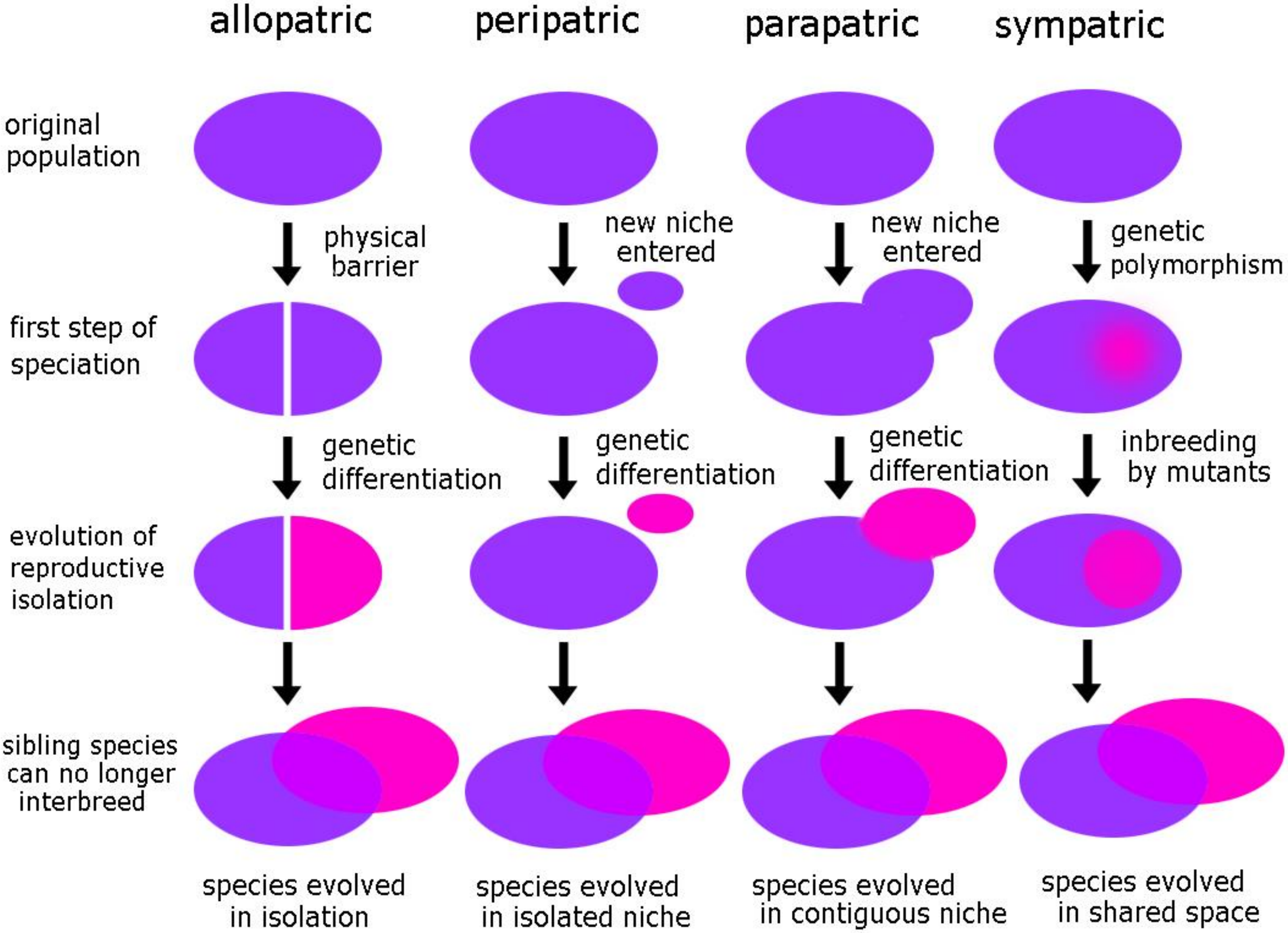
- Using an average calibration of 14.5 Myr per substitution, the human–gorilla divergence was dated to have occurred 14.5 and 7.25 Mya by α- and β-chains, because human and gorilla show two and one differences in these chains, respectively. Therefore, Zuckerkandl and Pauling reported a mean date of 11 Mya for the human–gorilla divergence from an analysis of the two proteins.



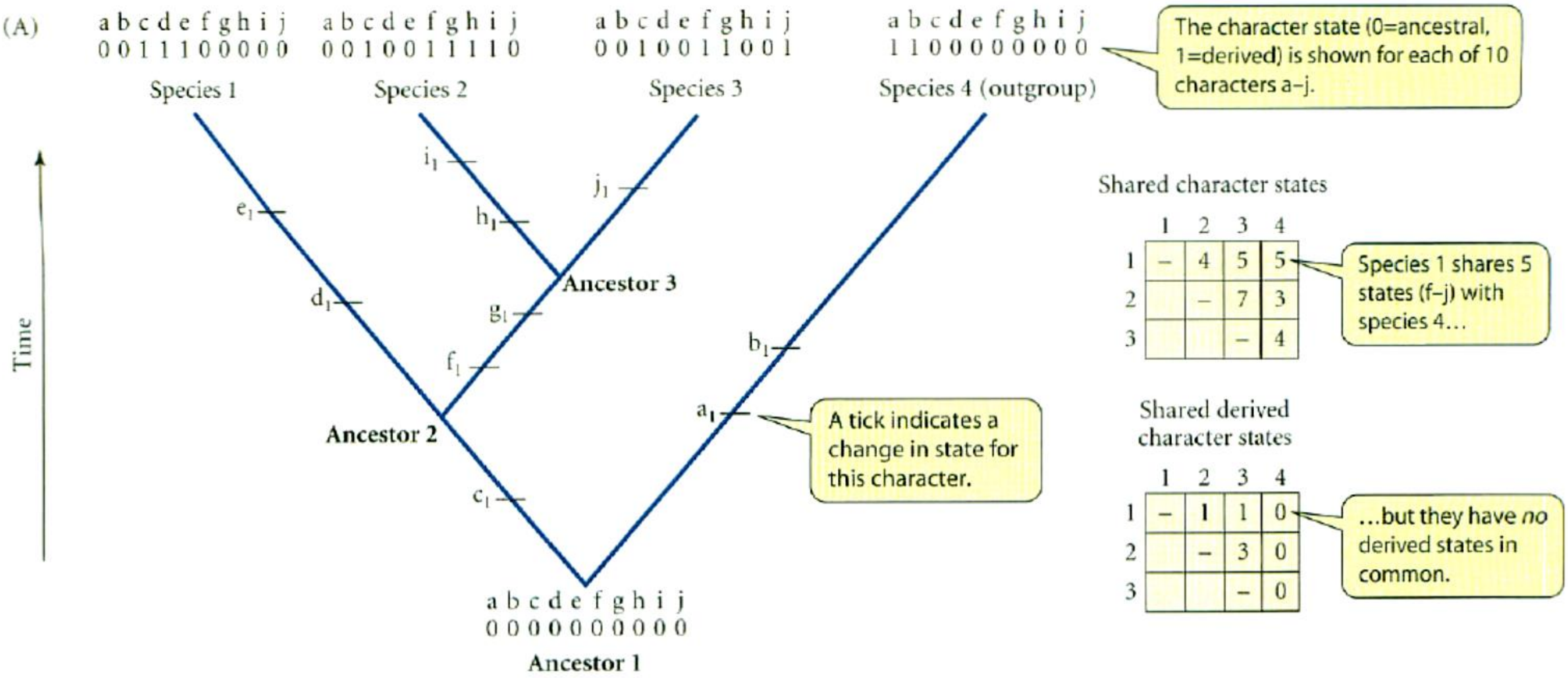
Combined (Margoliash)



Speciation


allopatric
peripatric
parapatric
sympatric
original population
physical barrier
new niche entered
new niche entered
genetic polymorphism
first step of speciation
genetic differentiation
genetic differentiation
genetic differentiation
inbreeding by mutants
evolution of reproductive isolation
sibling species can no longer interbreed
species evolved in isolation
species evolved in isolated niche
species evolved in contiguous niche
species evolved in shared space

First, basics



The character state (0=ancestral, 1=derived) is shown for each of 10 characters a–j.

A tick indicates a change in state for this character.

Shared character states

	1	2	3	4
1	–	4	5	5
2		–	7	3
3			–	4

Species 1 shares 5 states (f–j) with species 4…

Shared derived character states

	1	2	3	4
1	–	1	1	0
2		–	3	0
3			–	0

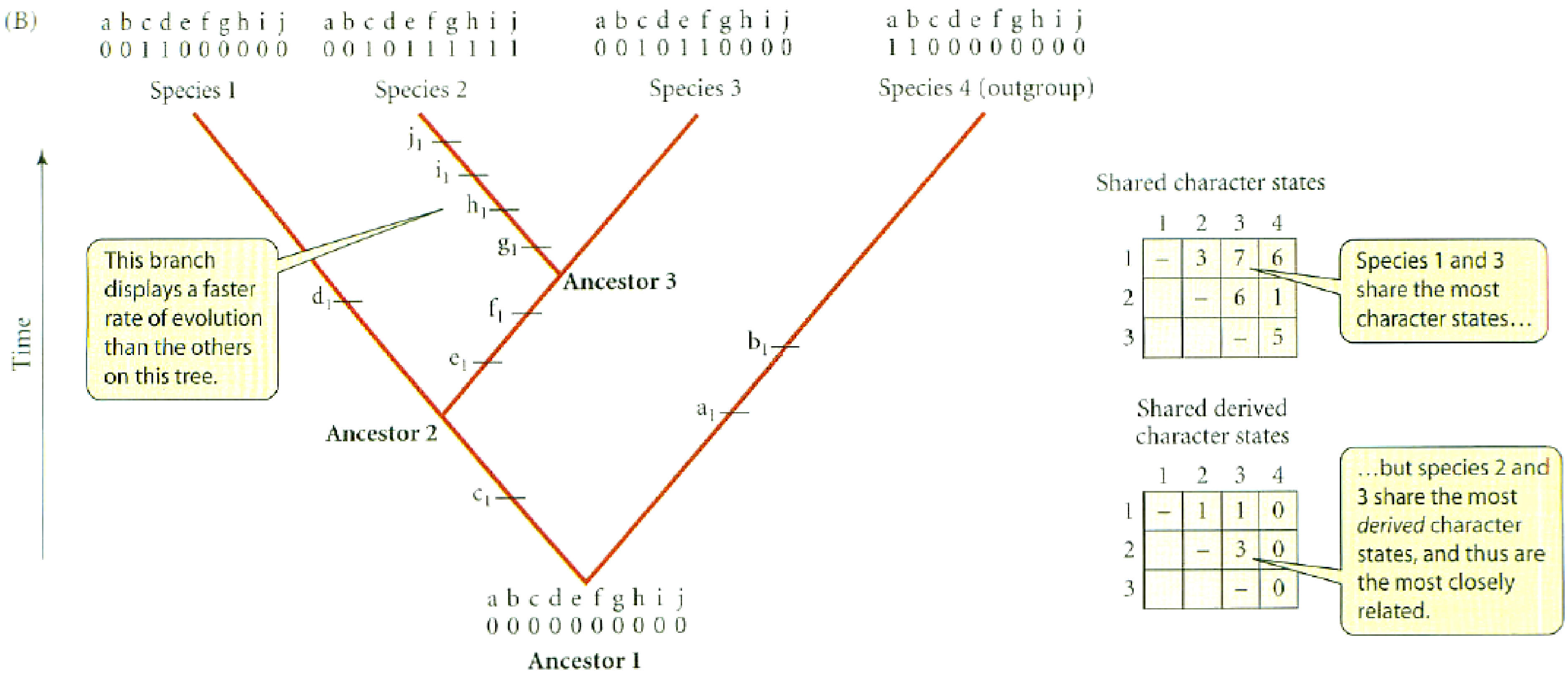
…but they have *no* derived states in common.

- Four species 1-4; ten characters a-j; each node is a character
- 0=Ancestral; 1= Derived

Explanation

- 3 monophylatic groups [1, 2, 3 & 4; 1, 2 & 3; 1 & 2]
- Calculating similarities of divergent species: counting the number of characters they share
  - 2 & 3 are most similar
  - 4 is the least similar – outgroup
- But if we count the derives character states: Synapomorphies

Another scenario



This branch displays a faster rate of evolution than the others on this tree.

Shared character states

	1	2	3	4
1	–	3	7	6
2		–	6	1
3			–	5

Species 1 and 3 share the most character states…

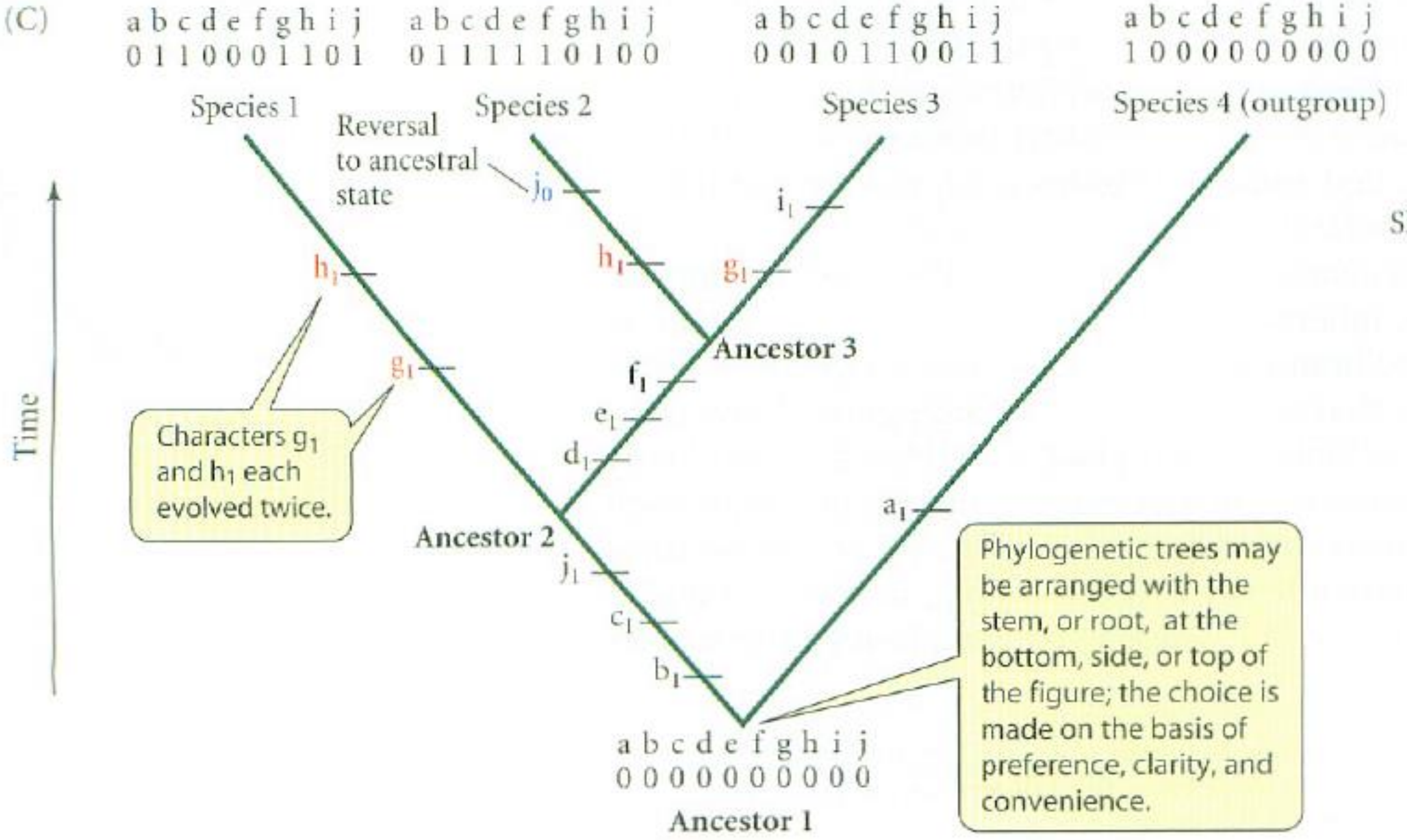
Shared derived character states

	1	2	3	4
1	–	1	1	0
2		–	3	0
3			–	0

…but species 2 and 3 share the most *derived* character states, and thus are the most closely related.

- In this case the divergence is variable
- In this case most familiar species are 1 & 3
- BUT 2 & 3 share most derived character states
  - – That makes them closest relative

One more scenario



Characters g_1 and h_1 each evolved twice.

Phylogenetic trees may be arranged with the stem, or root, at the bottom, side, or top of the figure; the choice is made on the basis of preference, clarity, and convenience.

Shared character states

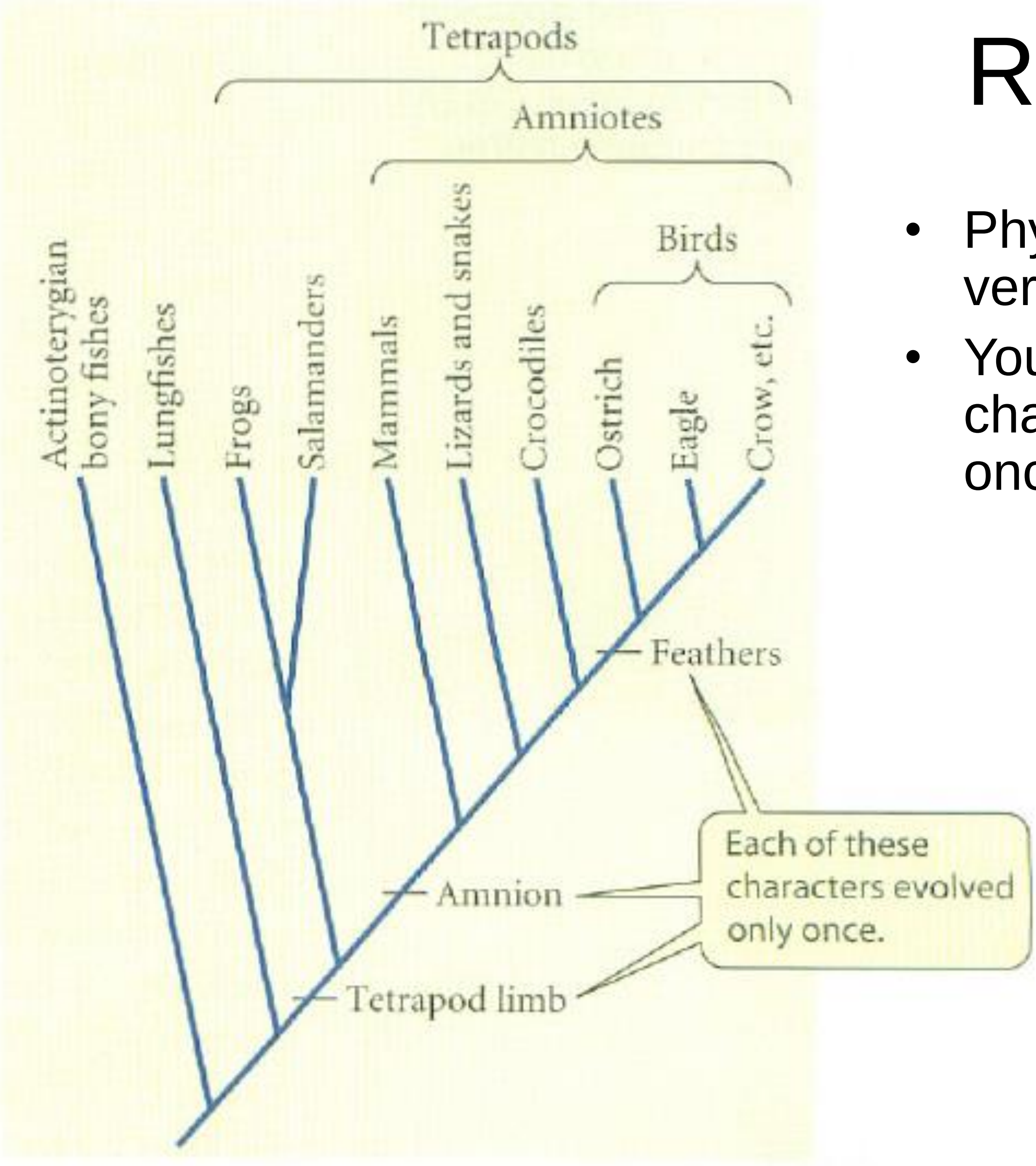
	1	2	3	4
1	–	5	5	4
2		–	6	3
3			–	1

Shared derived character states

	1	2	3	4
1	–	3	4	0
2		–	5	0
3			–	0

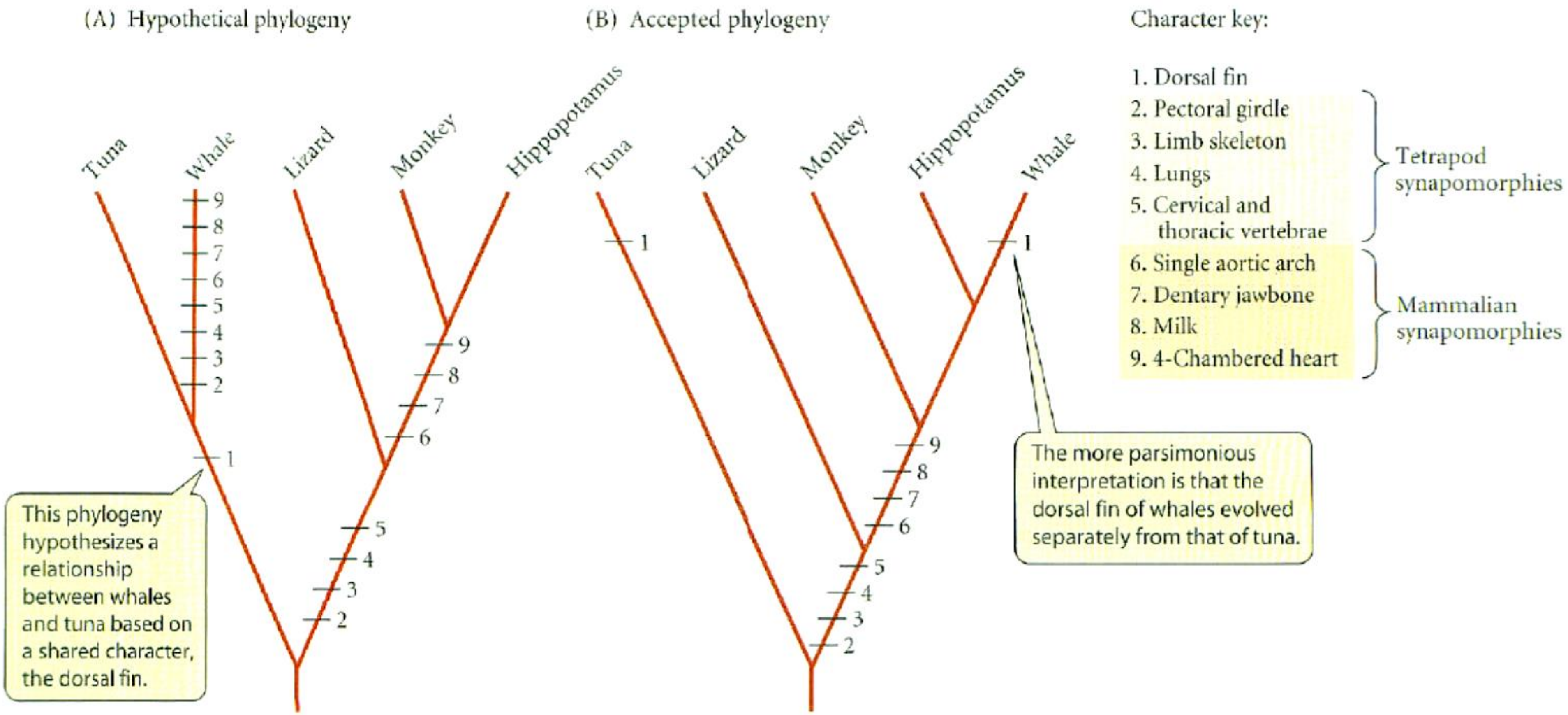
- Most reasonable scenario; some characters evolving more than once
- And in this case 2 & 3 is most similar and closest relative

Real example



- Phylogenetic map of vertebrate evolution
- You see three characters evolving once

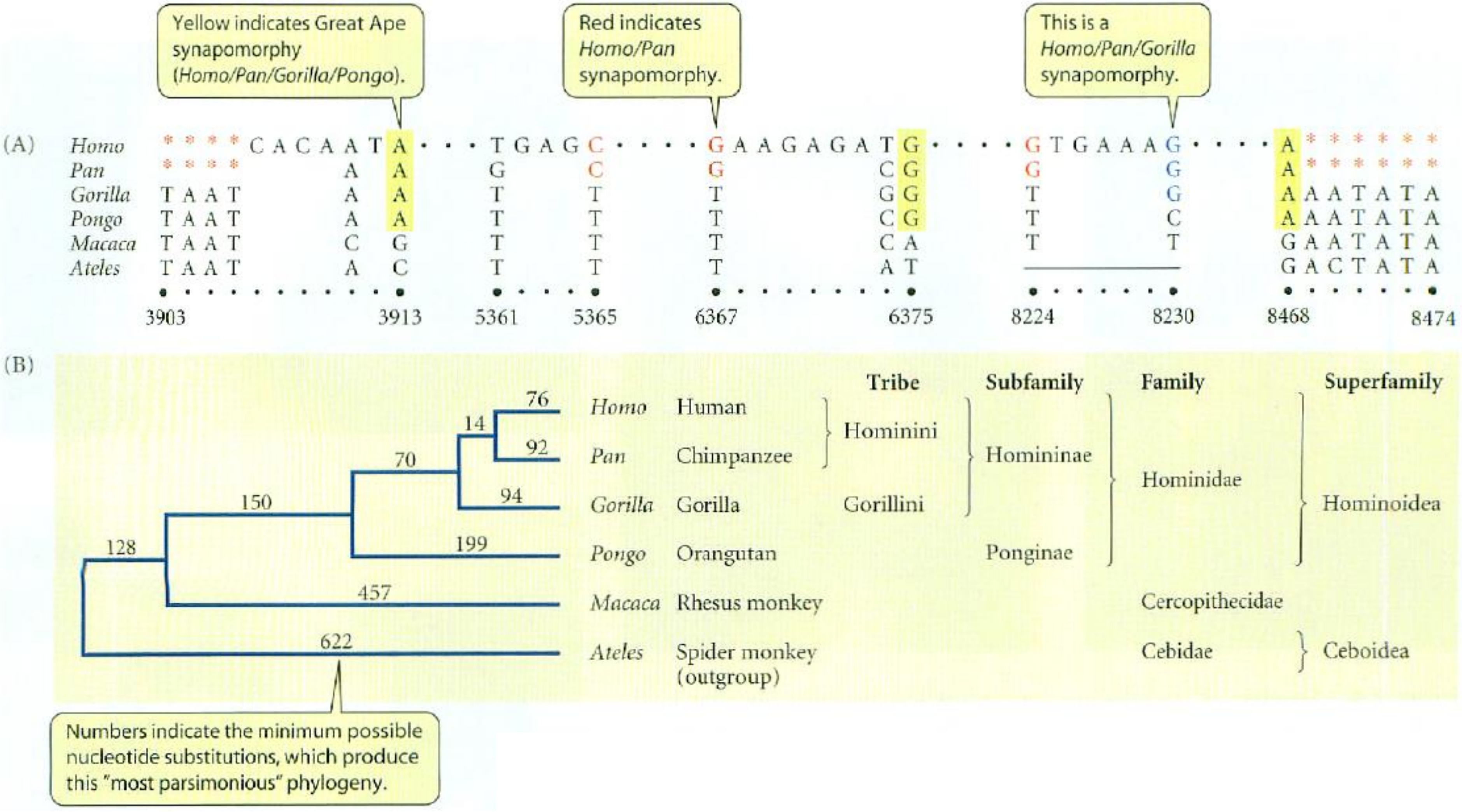
Maximum parsimony



- But in case of characters evolving more than once we follow the simplest path

A real example

ψη-Globin pseudogene



Yellow indicates Great Ape synapomorphy (*Homo/Pan/Gorilla/Pongo*).

Red indicates *Homo/Pan* synapomorphy.

This is a *Homo/Pan/Gorilla* synapomorphy.

Numbers indicate the minimum possible nucleotide substitutions, which produce this "most parsimonious" phylogeny.

(A)
Homo * * * * C A C A A T A · · · T G A G C · · · · G A A G A G A T G · · · · G T G A A A G · · · · A * * * * * *
Pan * * * * A A G C G C G G G A * * * * * *
Gorilla T A A T A A T T T G G T G A A A T A T A
Pongo T A A T A A T T T C G T C A A A T A T A
Macaca T A A T C G T T T C A T T G A A T A T A
Ateles T A A T A C T T T A T G A C T A T A
3903
3913
5361
5365
6367
6375
8224
8230
8468
8474
(B)
76
14
92
70
94
150
199
128
457
622
Homo Human
Pan Chimpanzee
Gorilla Gorilla
Pongo Orangutan
Macaca Rhesus monkey
Ateles Spider monkey (outgroup)
Tribe
Subfamily
Family
Superfamily
Hominini
Gorillini
Homininae
Ponginae
Hominidae
Hominoidea
Cercopithecidae
Cebidae
Ceboidea

So……

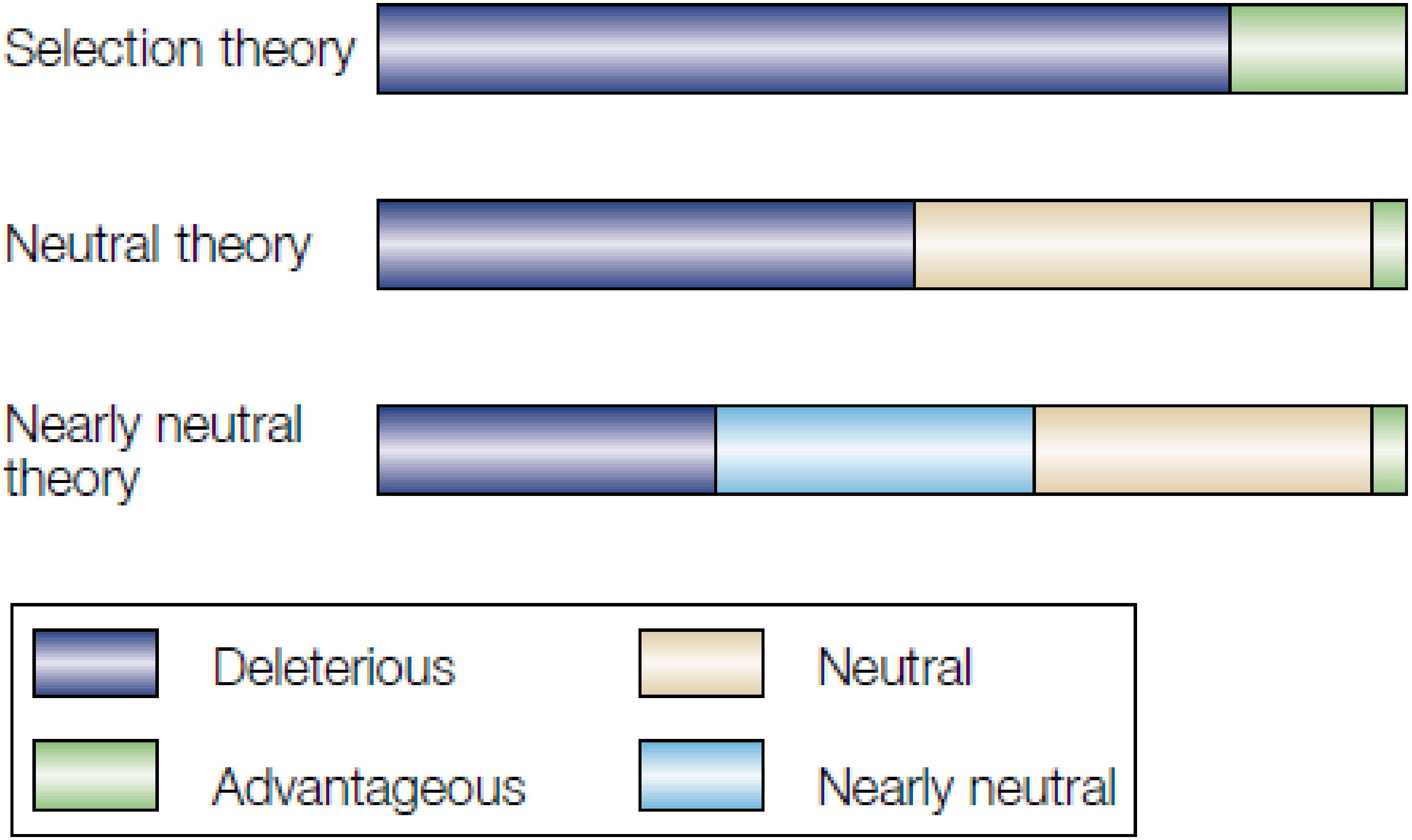
- Most amino-acid changes would be either favourable (positively selected) or deleterious (removed by selection)
- But you know that the proteins change with time
- And based on the changes one can define phylogenetic tree
- So if you know the changes and interval you can
  - Know who is our ancestors
  - When did we separate

- We all wish things are that easy

Ideas refined

- Motoo Kimura and Tomoko Ohta explained the constant characteristic rate for each protein by suggesting that most amino-acid changes in a protein were effectively neutral — that is, changing the amino-acid sequence had no influence on the fitness of an organism and, therefore, the rate of change was not affected by natural selection
  - Kimura, M. & Ohta, T.; *J. Mol. Evol.* **1**, 1–17 (1971).
- Tomoko Ohta extended the neutral theory by recognizing the critical role of effective population size. The fixation of nearly-neutral alleles of small selective effect is expected to be greatest in small populations
  - Ohta, T.; *Proc. Natl Acad. Sci. USA* **99**, 16134–16137 (2002).

In an image


Selection theory
Neutral theory
Nearly neutral theory
Deleterious
Neutral
Advantageous
Nearly neutral

Neutral theory

- Kimura and Ohta reasoned that advantageous mutations would be relatively rare, deleterious mutations would be rapidly removed by selection and that a large proportion of possible amino acid changes would have no practical effect on the functioning of the protein
  - overall mutation rate (μ) is the sum of deleterious (μ^-), neutral (μ^o), and positive (μ^+) mutations, $\mu = \mu^- + \mu^o + \mu^+$
- The theory focused on neutral mutations because advantages mutations (μ^+) were considered relatively rare, and in large populations deleterious mutations (μ^-) would be eliminated by negative selection. The mutation rate is expressed per individual per unit time — which might be per generation, per year or per cell replication.

Neutral theory, cont…

- The other parameter required is the EFFECTIVE POPULATION SIZE (N_e). For a haploid taxon, the number of mutations per time period = N_e. µº, the probability of fixation of a neutral mutation = 1/ N_e and, therefore, the number of neutral mutations fixed per time period = N_e.µº/ N_e = µº.
- In other words, although a larger population produces more mutations, the probability of a specific mutation being FIXED into the population declines proportionally with population size. So, according to a neutral model, population size cancels out to leave the molecular evolution rate determined by the mutation rate (µº). For diploids, the mutations double (2 N_e), but the probability of fixation halves (1/2 N_e) and so population size still cancels out.

If that is true, then…

- Dickerson compared what was then known about the protein structures of histones, cytochrome *c*, haemoglobins and fibrinopeptides, and concluded that their different rates of change could be explained by the proportion of neutral sites that each protein contained — the greater the proportion of neutral sites, the faster the rate of molecular evolution
  - Dickerson, R. E.; *J. Mol. Evol.* **1**, 26–45 (1971)
- The molecular clock is a ‘sloppy’ clock
  - nucleotide distance between sister species on Hawaiian islands, plotted against geological estimates of island age, gave impressively linear relationships for both birds and fruit flies

Sloppy clock

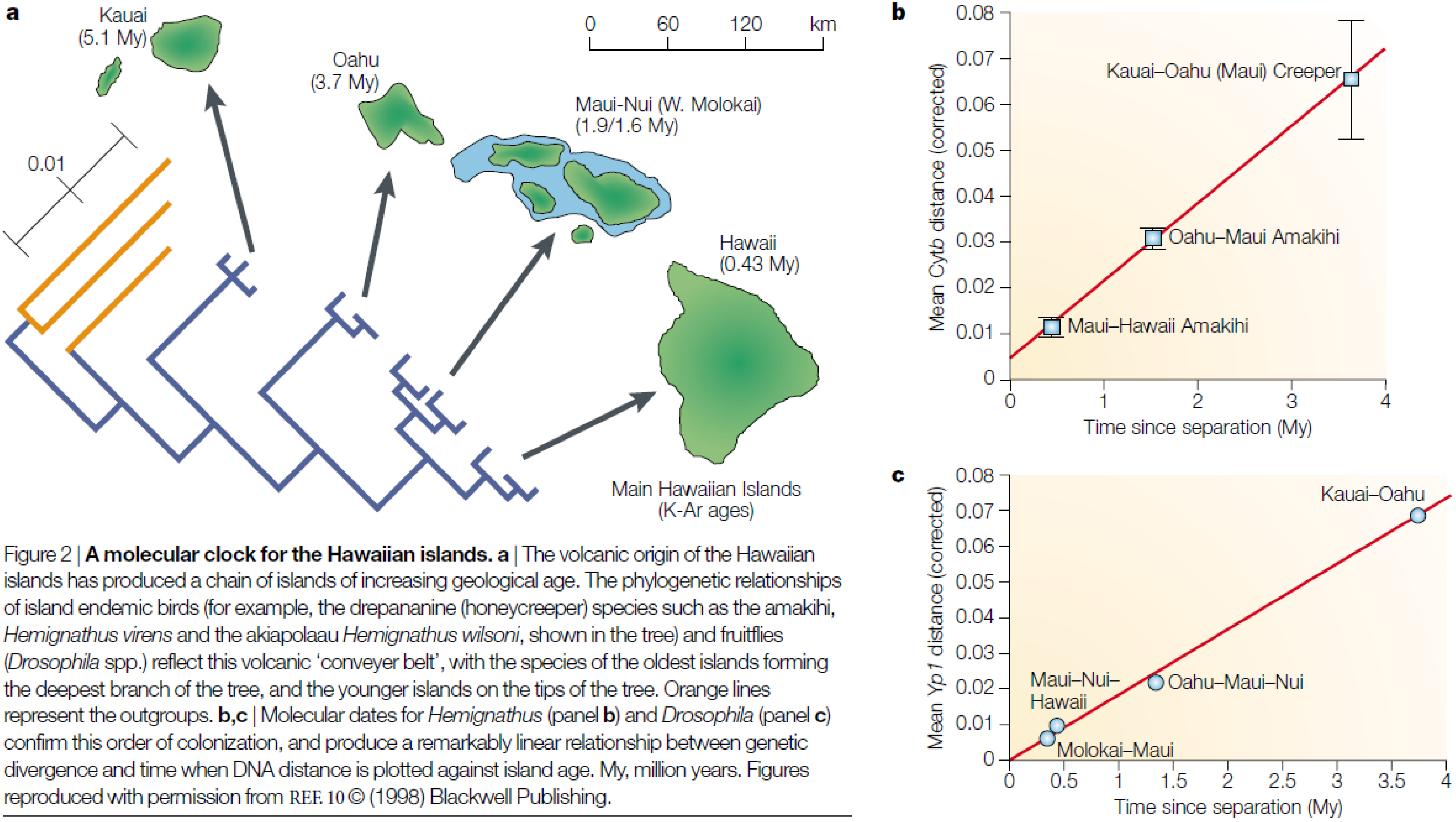


Figure 2 | **A molecular clock for the Hawaiian islands. a** | The volcanic origin of the Hawaiian islands has produced a chain of islands of increasing geological age. The phylogenetic relationships of island endemic birds (for example, the drepananine (honeycreeper) species such as the amakihi, *Hemignathus virens* and the akiapolaau *Hemignathus wilsoni*, shown in the tree) and fruitflies (*Drosophila* spp.) reflect this volcanic 'conveyer belt', with the species of the oldest islands forming the deepest branch of the tree, and the younger islands on the tips of the tree. Orange lines represent the outgroups. **b,c** | Molecular dates for *Hemignathus* (panel **b**) and *Drosophila* (panel **c**) confirm this order of colonization, and produce a remarkably linear relationship between genetic divergence and time when DNA distance is plotted against island age. My, million years. Figures reproduced with permission from REF. 10 © (1998) Blackwell Publishing.

However

- Empirical studies have also shown a great deal of variation in the rate of molecular evolution. The neutral theory allows for two sources of rate variation in the molecular clock:
  - the ‘sloppiness’ of the ‘tick rate’ and
  - changes in the mutation rate.
- These types of rate variation do not necessarily arise from different mechanisms. However, they do give rise to two types of error in molecular date estimates that contribute to ‘residual effects’ (unevenness of substitution rate in a lineage) and ‘lineage effects’ (variation in substitution rate between lineages) on the rate of molecular evolution.

Further characterization

- The molecular clock is probabilistic, not deterministic
- The three-dimensional structure assumed by proteins has an important effect on the patterns of substitutions
  - Some substitutions can happen without changing 3D structure
- The strength of selection on particular sites can change, producing bursts of substitutions as a molecule is adapted to a new role or responds to changes in another part of the genome
  - Ruminants and leaf eating langurs have co-opted lysozyme and ribonuclease for their herbivory (evolved twice)
- The selection coefficient of an entire gene can change
  - Pseudogenes: copies of a gene that have been rendered non-functional by mutation. Now all mutations to that gene is neutral

The nearly neutral theory

- Slightly deleterious mutations will tend to be removed by selection in very large populations; however, they can be fixed by chance in smaller populations in which selection can be overpowered by random sampling events (GENETIC DRIFT)
- Ohta showed that the fixation of these nearly neutral mutations with small selection coefficients (*s*), whether positive or negative, would be governed by chance events in small populations, just as if they were neutral. So, a mutation would be effectively neutral if $|s| < 1/4N_e$, where N_e is the effective population size.
- In other words, whether a mutation behaves according to the neutral expectation is determined not simply by the selection properties of the mutation (*s*), but also the size of the population in which it arises (N_e).

Continued

- Ohta then considered the fate of a range of slightly deleterious or slightly advantageous mutations with a VARIANCE on their selection coefficients (σ_s) — some will be slightly deleterious and some slightly advantageous.
- Mutations are divided into three categories: mutations for which selection is the predominant force ($4N_e\sigma_s > 3$); nearly neutral mutations, which are governed by both selection and drift ($3 \geq 4N_e\sigma_s \leq 0.2$); and effectively neutral mutations, the fate of which is determined only by drift ($4N_e\sigma_s < 0.2$).
- So, the nearly neutral theory describes how the rate of molecular evolution can vary not only with changes in the mutation rate, but also through the changing balance between selection and drift.

Furthermore

- Then the fate of a range of slightly deleterious or slightly advantageous mutations with a VARIANCE on their selection coefficients (σ_s). Mutations are divided into three categories:
  - – mutations for which selection is the predominant force ($4N_e\sigma_s > 3$);
  - – nearly neutral mutations, which are governed by both selection and drift ($3 \geq 4N_e\sigma_s \leq 0.2$); and
  - – effectively neutral mutations, the fate of which is determined only by drift ($4N_e\sigma_s < 0.2$).
- So, the nearly neutral theory describes how the rate of molecular evolution can vary not only with changes in the mutation rate, but also through the changing balance between selection and drift.

Details

- Rate of molecular evolution can vary in three ways: through changes to mutation rate, population size or selective coefficients
- Mutation rate clearly varies between taxa and much of this variation is due to differences in repair equipment (biochemical factors).
  - During DNA replication, and
    - RNA viruses provide an extreme example: they copy their genomes using highly error-prone RNA polymerases or reverse transcriptases that lack proofreading function. This contributes to the high rates of molecular evolution in the retrovirus HIV — around a million times faster than the rate of evolution of mammalian genomes, which use a battery of replication and repair enzymes to reduce the mutation rate

Still mutation rate

- Even within mammalian cells, different replication and repair enzymes vary in mutation rate: mitochondrial DNA is copied by DNA polymerase-γ, which has a higher error rate than other mammalian DNA polymerases, and this contributes to the higher mutation rate of mitochondrial genes over nuclear genes.
- Smaller-bodied species of vertebrates with faster metabolic rates have higher substitution rates than larger-bodied species
- The mutation rate might be best measured as a per-generation rate, particularly in animals that have separate germline and somatic cells

– Damage that is not repaired

Population size

- The nearly neutral theory also predicts that any consistent effect on effective population size could influence substitution rate. These effects include ecological factors (such as a reduced population size on different islands), correlates of life-history evolution (such as inbreeding or EUSOCIALITY) and aspects of niche or lifestyle (such as endoparasitism).

Why molecular clock

- If the rate of molecular evolution is relatively constant, then the amount of genetic difference between two species gives a measure of the time since their evolutionary separation
- Can provide insights into the history of all organisms from which we can obtain genetic sequences but no fossil record: viruses

Is it perfect?

- A study that produced a molecular date estimate for the split between kingdoms that is markedly younger than the earliest fossils
  - Controversial
- Two main types of error in molecular clock estimates.
  - First, the sloppy nature of the substitution rate results in large variance around the amount of genetic difference expected for any given time period
  - Second, the nearly neutral theory predicts that the rate of molecular evolution is influenced by mutation rate, population size and the relative proportions of sites with different selective coefficients; these factors might differ between genes, between species and over time, potentially resulting in consistent bias in date estimates

So

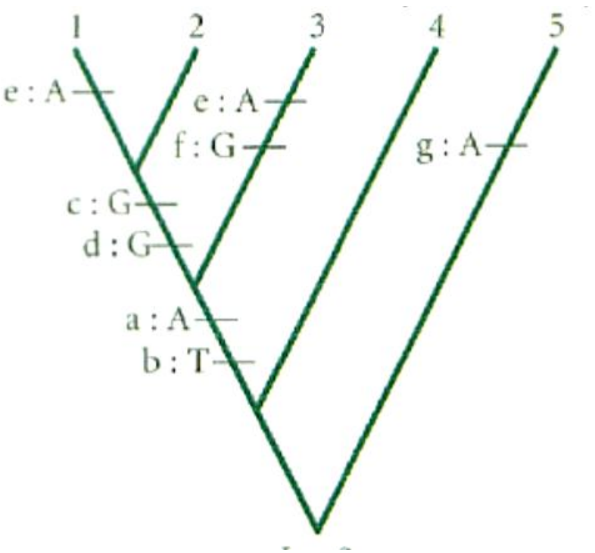
- Few studies present molecular date estimates with confidence intervals that accurately portray the variance in the clock due to the sloppiness of the tick rate, or the lineage variation in rates. Such confidence intervals allow molecular dates to be used to test evolutionary hypotheses within the bounds of the accuracy and precision of molecular clocks, by asking whether the range of possible date estimates is consistent with a specific evolutionary hypotheses
  - Eg: animal phyla arose in the early Cambrian
- Both theory and observation show that the molecular clock is much more complex than was initially supposed

The future

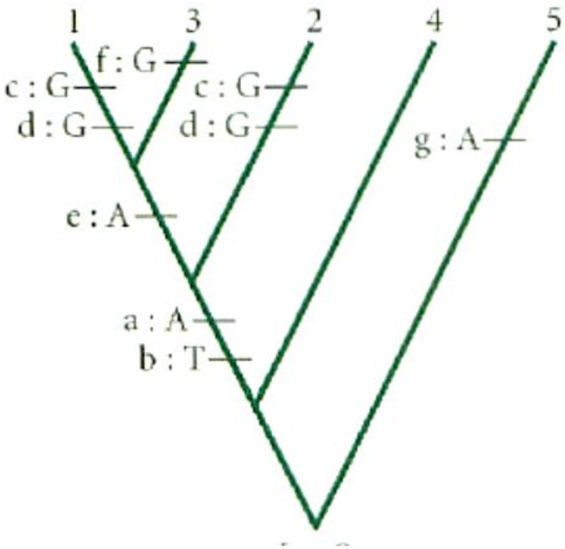
- There are three approaches to allowing for the variation in the rate of molecular evolution.
  - First, molecular date estimates should be presented with confidence intervals that accurately portray the variance in the rate of molecular evolution, both in and between lineages.
  - Second, new molecular clock methods that incorporate variation in the rate of molecular evolution should be developed.
  - Third, an understanding of the mechanisms that generate the rate variation will inform judgment of the reliability of molecular date estimates (including the identification of cases to which molecular clocks cannot be reasonably applied).

Test

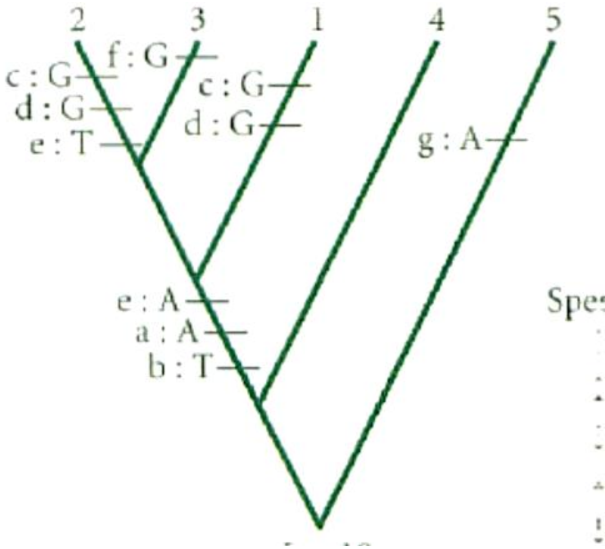
Tree 1



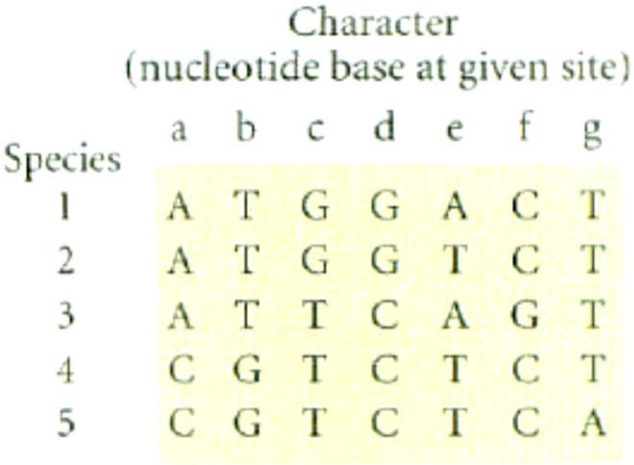
Tree 2



Tree 3



Q: which tree is most probable?



Character
(nucleotide base at given site)

Species	a	b	c	d	e	f	g
1	A	T	G	G	A	C	T
2	A	T	G	G	T	C	T
3	A	T	T	C	A	G	T
4	C	G	T	C	T	C	T
5	C	G	T	C	T	C	A

Answer is in futuyama but I want you to solve it yourself

Futuyama

Nature Reviews Genetics 4, 216-224 (March 2003) | doi:10.1038/nrg1020

Nature Reviews Genetics 6, 654-662 (August 2005) | doi:10.1038/nrg1659

& some google images