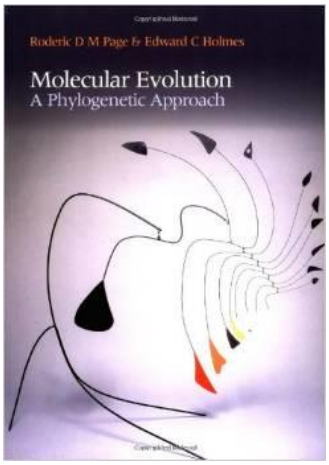
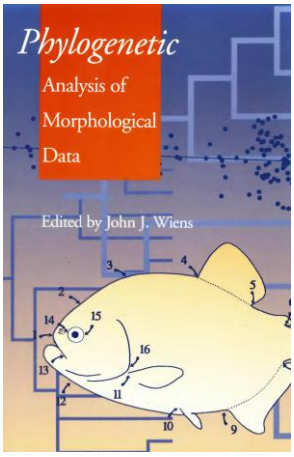
References





Useful reading:

Nei, M. & Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York, NY, 333 pp.
Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA. [**F**]

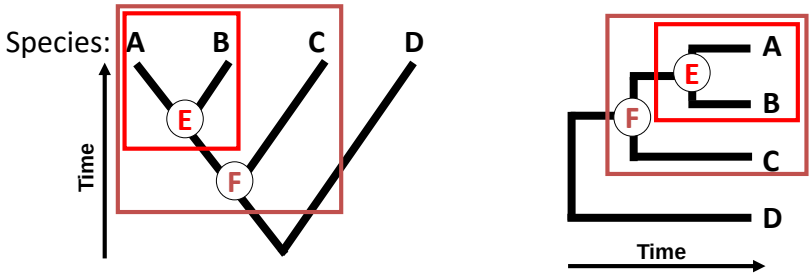
Assesment

- No mid-term exam
- Talks 7
- Final exam 13

Phylogeny

What is a phylogeny?

Branching diagram showing relationships between species (or higher taxa) based on their shared common ancestors



A and B are most closely related because they share a common ancestor (call the ancestor "E") that C and D do not share

A+B+C are more closely related to each other than to D because they share a common ancestor ("F") that D does not share

Tree of life

Why phylogeny is important?

Understanding and classifying the diversity of life on Earth

Testing evolutionary hypotheses:
- trait evolution
- coevolution
- mode and pattern of speciation
- correlated trait evolution
- biogeography
- geographic origins
- age of different taxa
- nature of molecular evolution
- disease epidemiology

…and many more applications!

Phylogenetic classification

- Based on known (inferred) evolutionary history.

- Advantage:
  - Classification reflects pattern of evolution
  - Classification not ambiguous

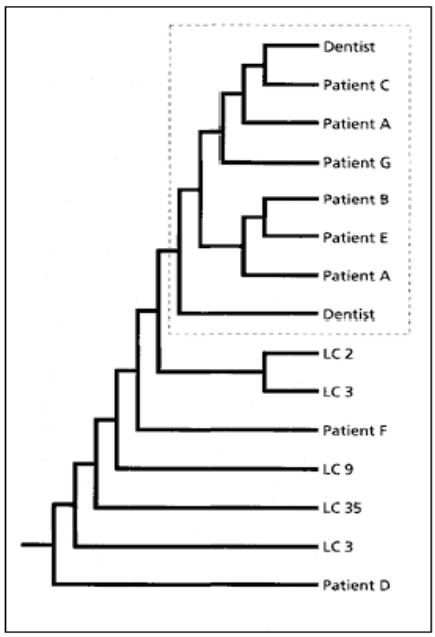


Fig. 1.2
The case of the Florida dentist. Each branch represents the sequence from part of the envelope (*env*) gene of HIV-1. Viral sequences were obtained from the dentist and seven of his former patients (labelled A to G), also infected with the virus. Five of these patients (A, B, C, E and G), have sequences very closely related to those of the dentist (boxed), suggesting that he infected them. Two of his other former patients (D and F) had other risk factors for HIV infection and their viruses are separated from the dentist by sequences taken from local controls (LC)HIV-infected individuals living within a 90-mile radius of the dentist's surgery. Because HIV-1 is so variable, two different sequences are included for the dentist and patient A. Data taken from Ou *et al.* (1992).

the 'Tree of Life', is one of the prime goals of evolutionary biology

- Our goal here is to give you some familiarity with trees so that interpreting them eventually becomes second nature

2.1 Introduction to Trees
2.1.1 Tree Terminology
2.1.2 A Shorthand for Trees
2.1.3 Cladograms, Additive Trees and Ultrametric Trees
2.1.4 Rooted and Unrooted Trees
2.1.5 Tree Shape
2.1.6 Splits
2.2 Reconstructing the History of Character Change
2.2.1 Ancestors
2.3 Trees and Distances
2.3.1 Metric Distances
2.3.2 Ultrametric Distances
2.3.3 Additive Distances
2.3.4 Tree Distances
2.4 Organismal Phylogeny
2.4.1 Clades and Classification
2.4.2 Gene Trees and Species Trees
2.4.3 Lineage Sorting and Coalescence
2.5 Consensus Trees
2.6 Networks
2.7 Summary

Tree Terminology

- Weighted tree and unweighted tree

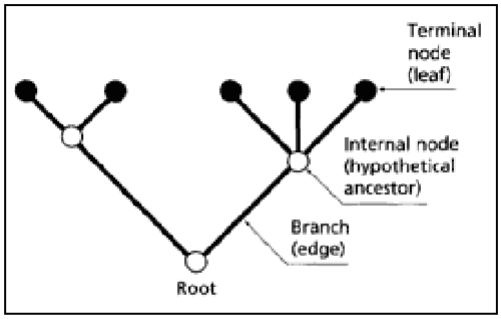
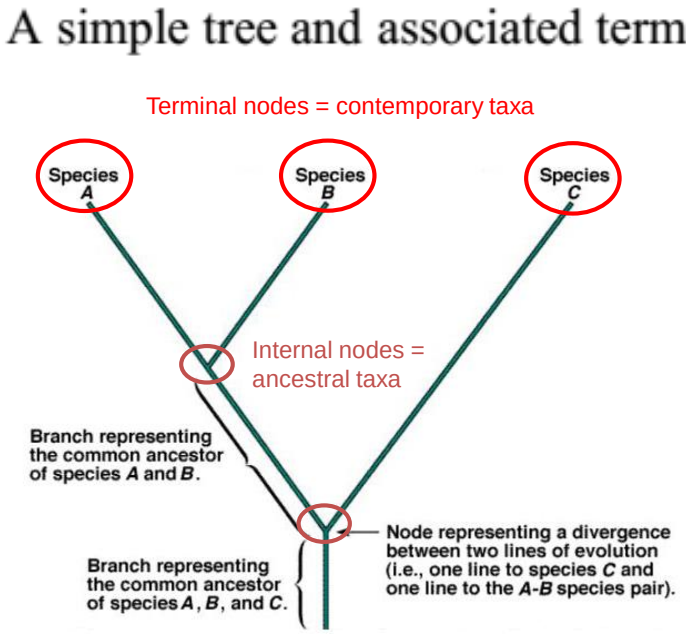
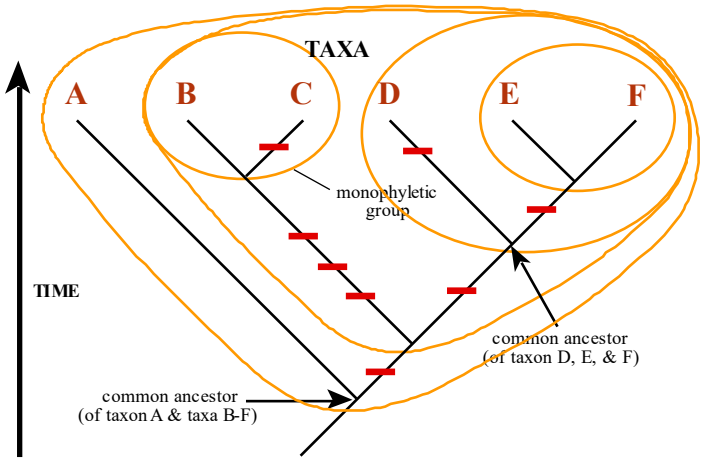


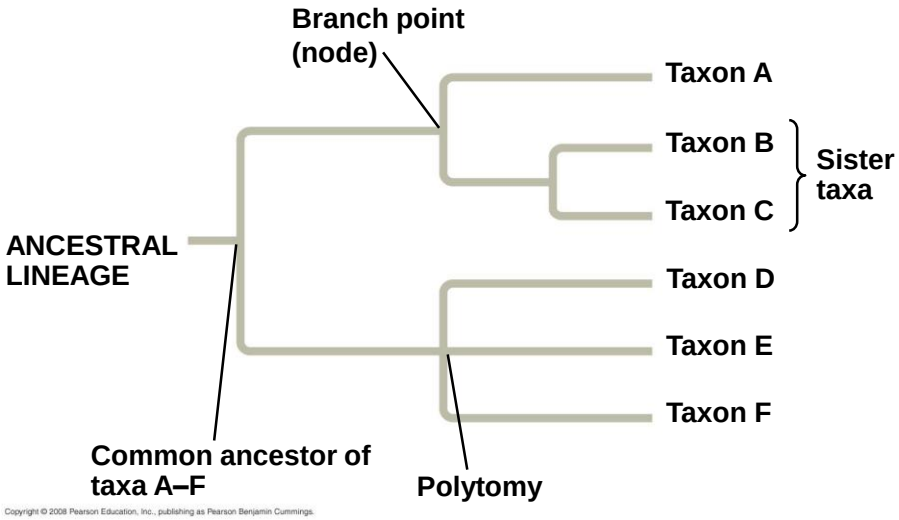
Fig. 2.1
A simple tree and associated terms.





Cladogram or Phylogenetic Tree

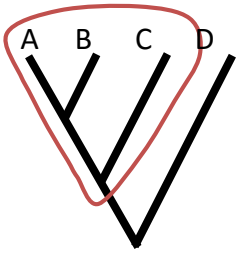
Fig. 26-5



Phylogeny and classification

Monophyletic group

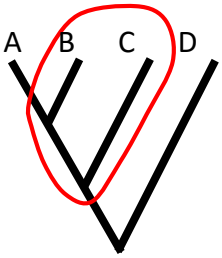
Includes an ancestor all of its descendants



How could this happen?

Paraphyletic group

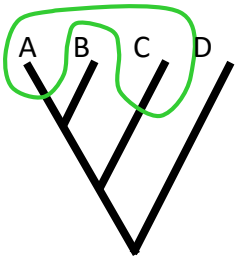
Includes ancestor and some, but not all of its descendants



Taxon A is highly derived and looks very different from B, C, and ancestor

Polyphyletic group

Includes two convergent descendants but not their common ancestor

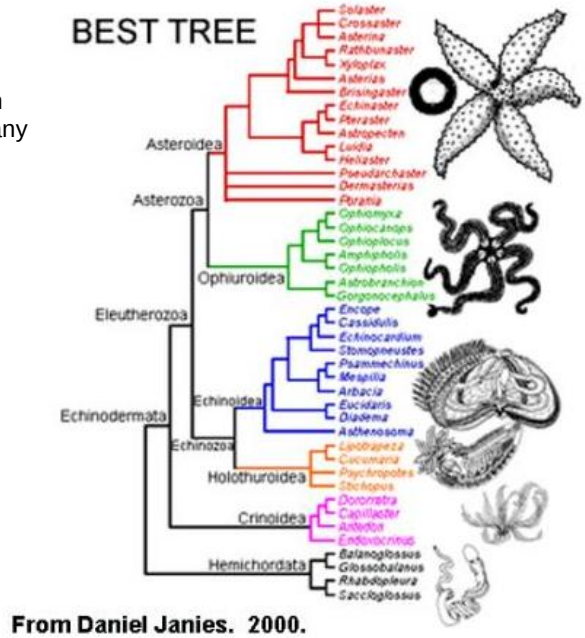


Taxon A and C share similar traits through convergent evolution

Only monophyletic groups (**clades**) are recognized in cladistic classification

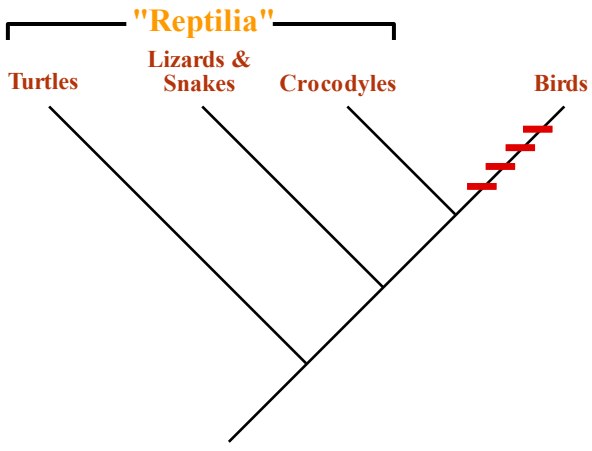
Monophyly

- Each group shares a common ancestor that is not shared by any members of another group

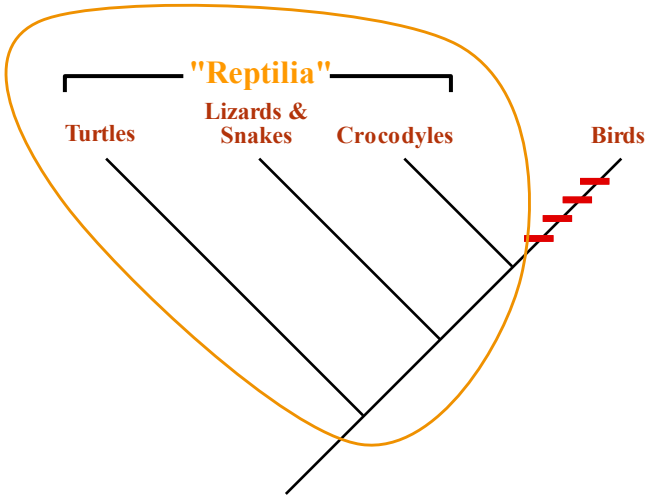


From Daniel Janies. 2000.

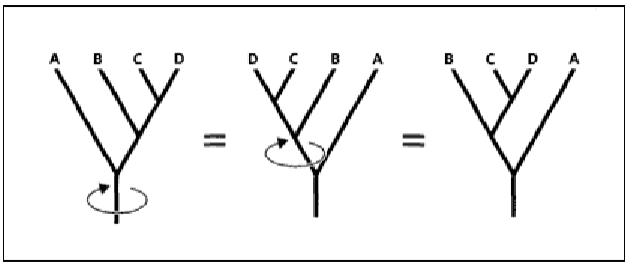
Paraphyletic group



"Reptilia" here paraphyletic



Trees are like mobiles



Degree of the resolution for trees

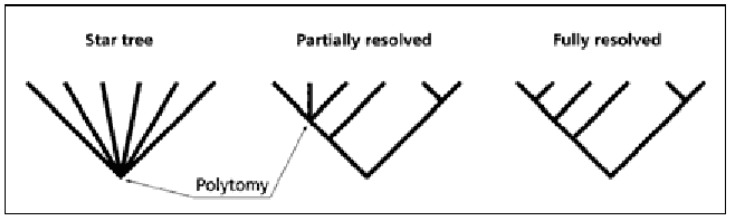


Fig. 2.2
Three trees showing various degrees of resolution, ranging from a complete lack of resolution (star tree) to a fully resolved tree. Any internal node with more than two immediate descendants is a polytomy.

What polytomies can represent?

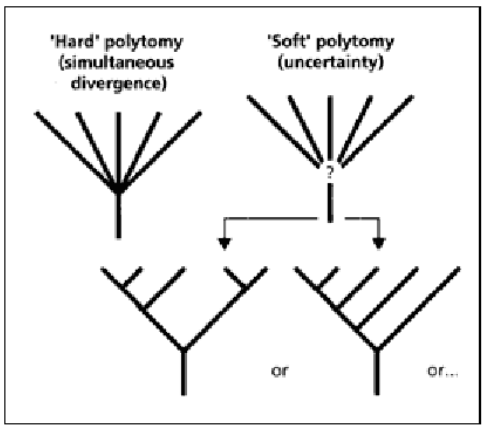
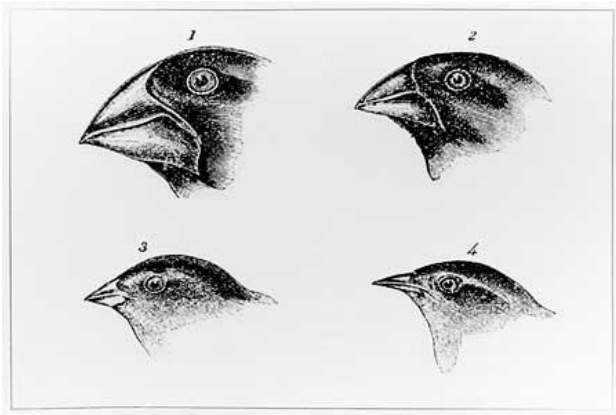


Fig. 2.3
Polytomies can represent either simultaneous divergence of multiple sequences ('hard'), or lack of resolution due to insufficient data or conflicting trees ('soft').

What kind of polytomy does the Galapagos finches show?



A shorthand for trees

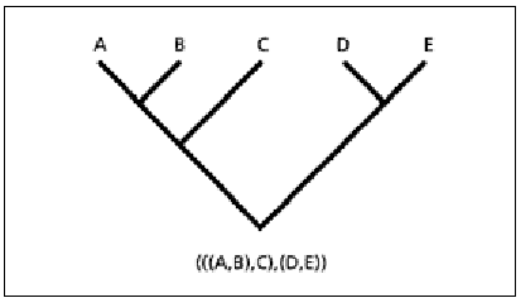
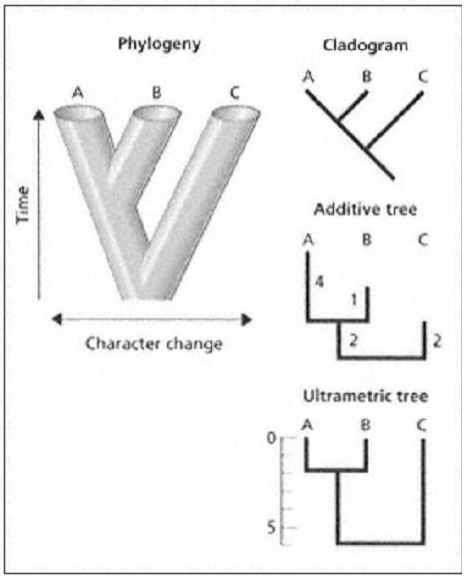


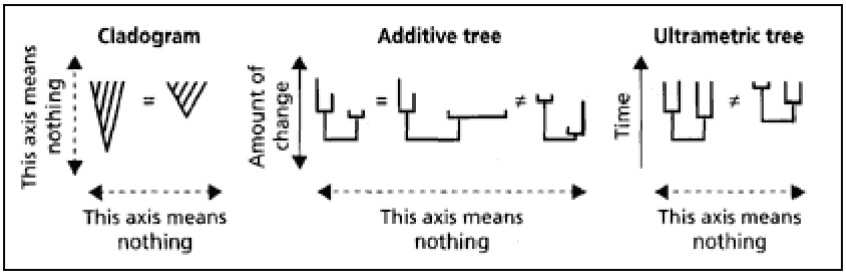
Fig. 2.4
A tree and its shorthand representation using nested parentheses.

Cladograms, Additive Trees and Ultrametric Trees

- A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.



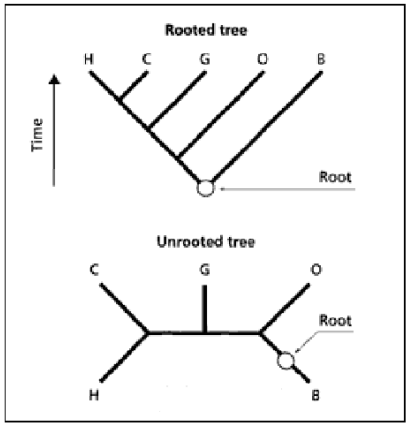
what do the horizontal and vertical axes represent?



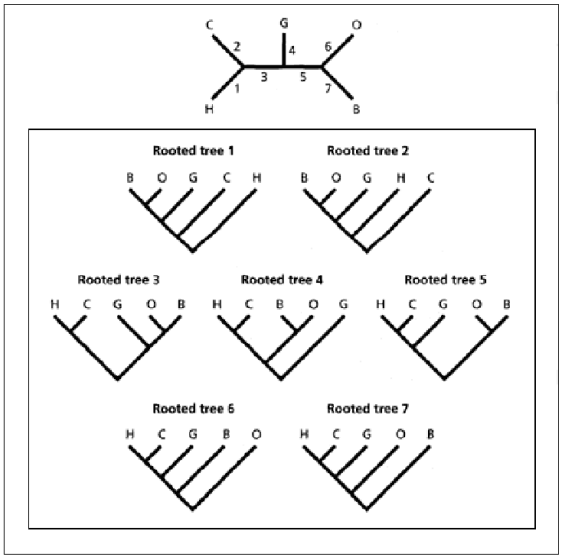
Rooted & Unrooted tree

-**rooted tree has direction.**
- **This direction corresponds to evolutionary time.**
-**It specify evolutionary relationship**

Terminology
Ingroup – studiedgroup
Outgroup – group not part of ingroup, used to "root" tree



Rooted and unrooted trees for human (H), chimp (C), gorilla (G), orangutan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.



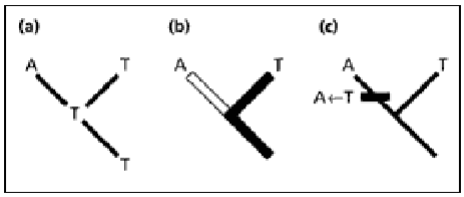
The seven rooted trees that can be derived from an unrooted tree for five sequences. Each rooted tree 17 corresponds to placing the root on the corresponding numbered branch of the unrooted tree.

2

Reconstructing the history of character change

Reconstructing the History of Character Change

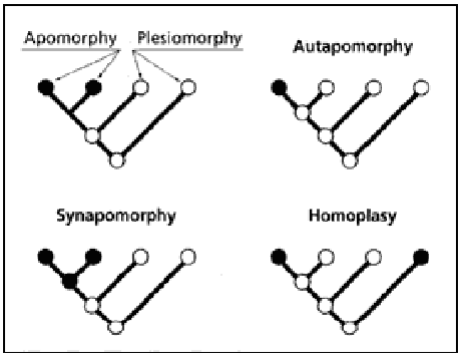
- The tree relating a set of sequences tells us only part of what we want to know as:
  - Inheritance from a common ancestor or
  - independent evolution



Three equivalent ways of representing the same evolutionary change on the same tree. (a) Each node is labelled by the corresponding nucleotide; (b) each branch is coloured corresponding to the nucleotide at the end of each branch; and (c) indicating on which branch the change took place.

Some basic terminology

- **Ancestral (primitive) and derived character states**
  - **Apomorphy**
  - **plesiomorphy**



Apomorphy (derived trait)

- = a new, derived feature. E.g., for this evolutionary transformation

  scales --------> feathers
  (ancestral feature) (derived feature)

- Presence of feathers is an **apomorphy** for birds.

Taxa are grouped by apomorphies

Apomorphies are the result of evolution.

Taxa which sharing apomorphies underwent same evolutionary history and should be grouped together.

Sequentially group taxa by
shared derived character states (apomorphies)

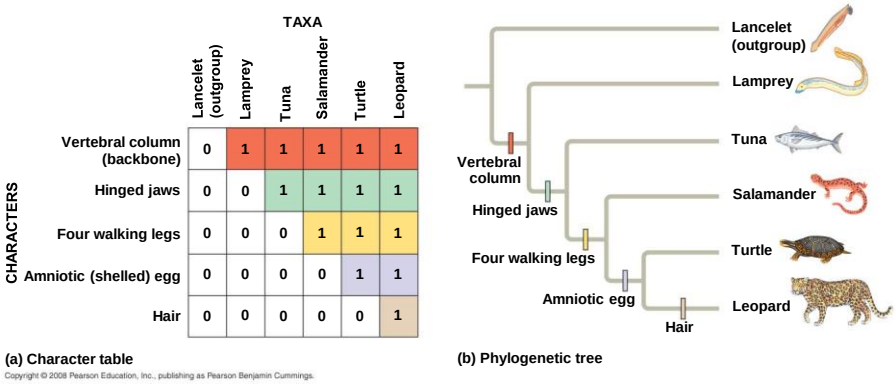
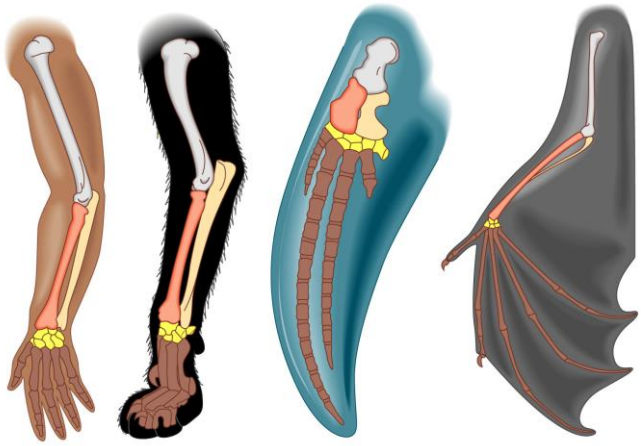


Fig. 26-11

Homology versus Homoplasy

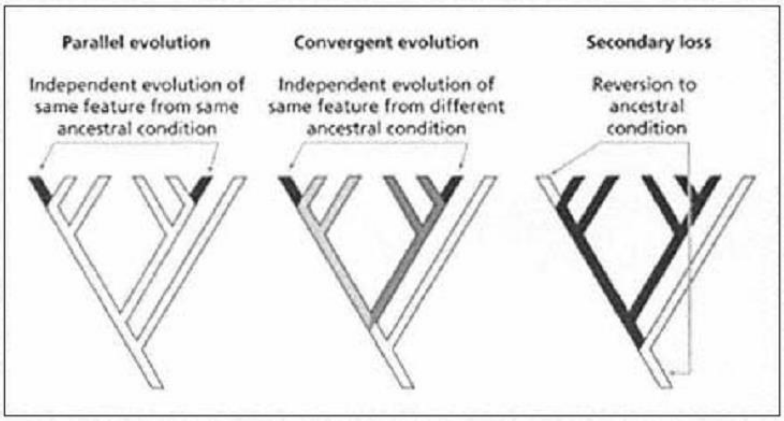


Homoplasy (analogy)

- Similarity <u>not</u> due to common ancestry
- **Parallel evolution** – gain of new, similar features from same ancestral condition.

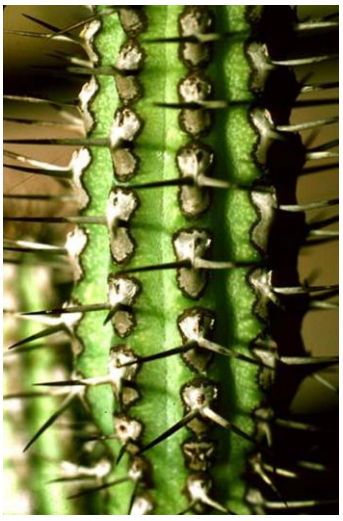
  Convergence– gain of new, similar features from different ancestral condition
- **Secondary loss –** revision to ancestral condition

Homoplasy



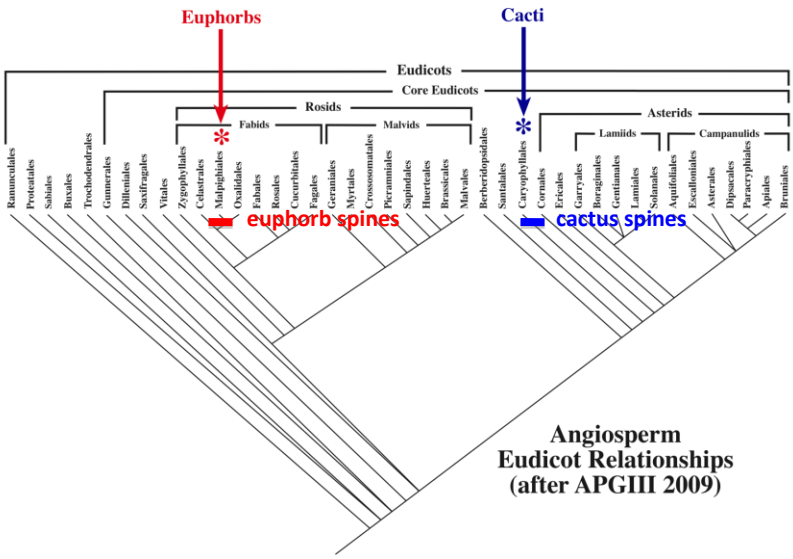
Convergent evolution:
spines of cacti & euphorbs

Cactus



Euphorb

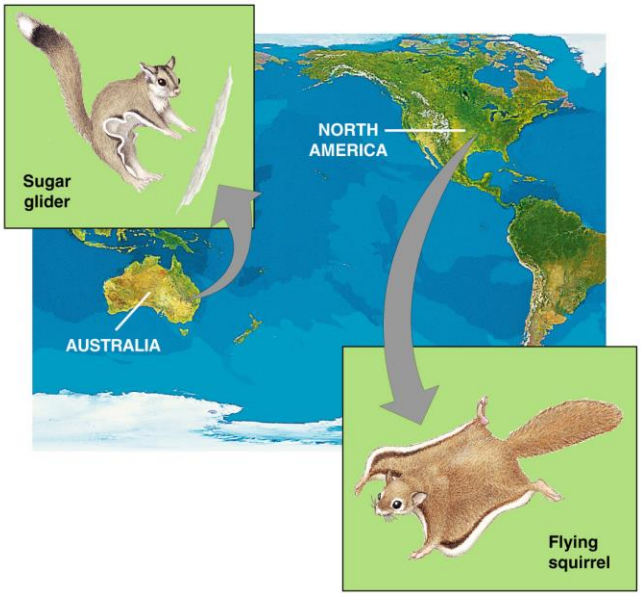
Convergent evolution: spines of cacti & euphorbs



بال درحشرات و دیگر جانوران

Old and New World Vultures

Convergent evolution:
wings of some animals evolved independently



Leg-less lizards

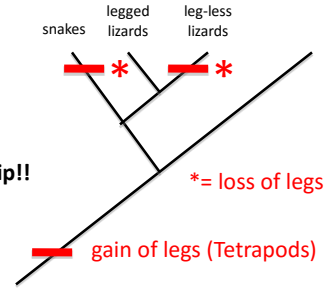
Snake

Both examples of **reversal** within Tetrapods:
loss of a derived feature – forelimbs.

Example of **Parallel evolution** relative to one another!
snakes and leg-less lizards

Homoplasy is a poor indicator for evolutionary relationship!!

Why?



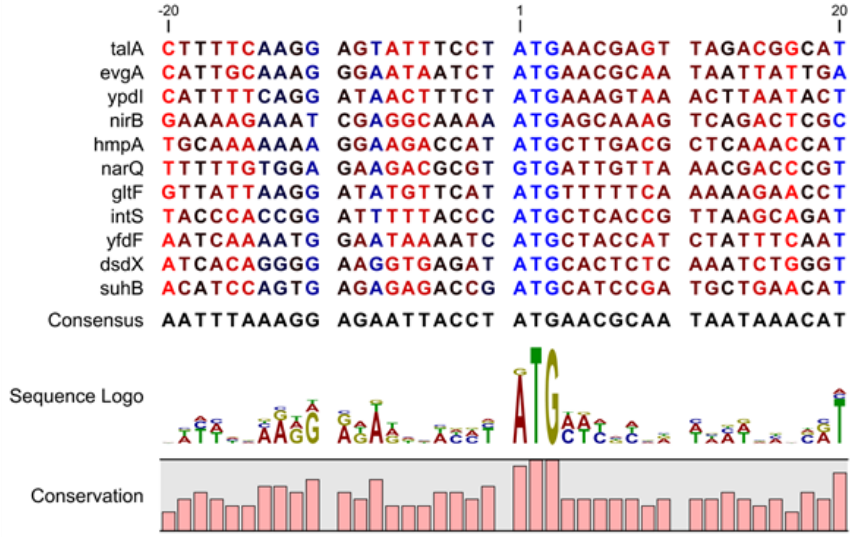
Ancestors

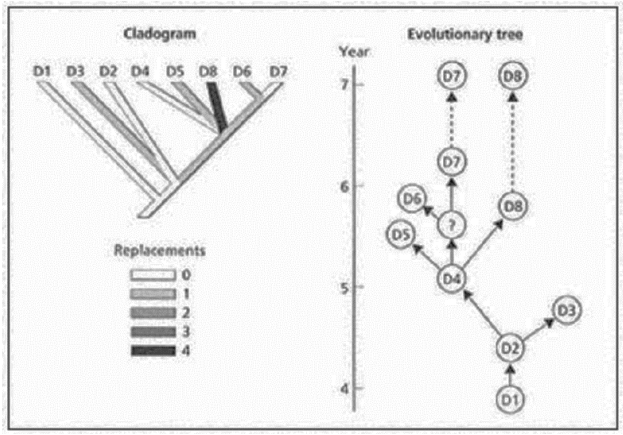
- They are now extinct but left descendants
- They are represented by the internal nodes of a tree
- These ancestors are hypothetical
- recognition of ancestors is possible if only:
  - Recovery of DNA from extinct taxa
  - Increasing number of sequences (works for fast evolving genes or taxa) (figure)

- How we can consider a DNA sequence belonging to an extinct taxa as an ancestral condition !!!!!!!

No autapomorphy (figure)

Recovering ancestral shape

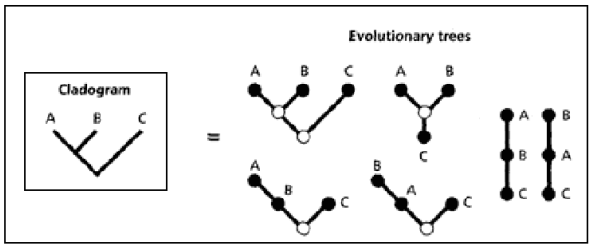


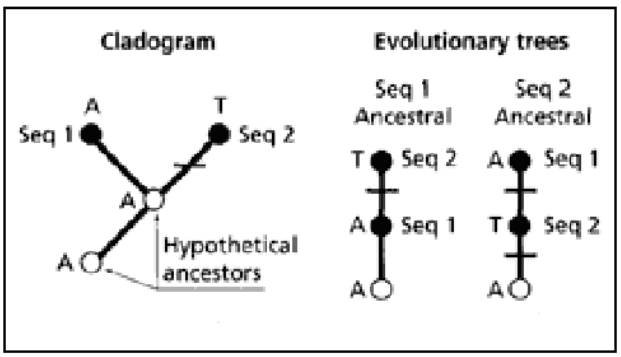


Cladogram and corresponding evolutionary tree for eight V3 loop amino acid sequences for HIV samples taken from a single patient over 3 years. In the cladogram on the left all eight sequences are depicted as terminal nodes; however, four sequences (D1, D2, D4 and D7) have no autapomorphies (i.e. there are no replacements along the branch leading to each sequence) and hence are possible ancestors. The evolutionary tree on the right depicts the same relationships as the cladogram, but the sequences lacking autapomorphies (except D7) are treated as ancestors which is consistent with the order of appearance of the sequences. Modified from Holmes et al. (1992).

Cladograms and evolutionary trees

- **'cladogram' refers to an evolutionary tree that has no information on branch lengths**
- In a **cladogram** the terminal taxa are always at the tips of the tree, no matter if the taxa are extant or extinct, or whether one or more of the taxa are ancestral to any of the others
- In an **evolutionary tree** some of the taxa may be ancestral to the others**.**





A cladogram for two sequences (Seq 1 and Seq 2) showing the nucleotide at a single site, and two of several possible evolutionary trees derived from that cladogram. We could postulate that either sequence is ancestral to the other. However, postulating Seq 2 to be an ancestor of Seq 1 requires the gain and subsequent loss of T, whereas if Seq 1 is an ancestor no additional substitutions need be postulated. Note that a third phylogeny would be identical to the cladogram

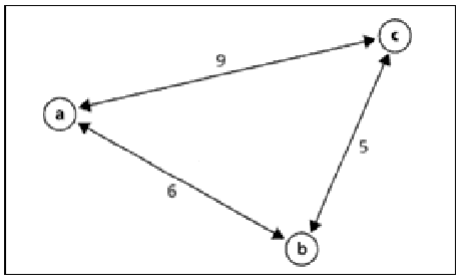
Trees and Distances

- **Measuring of sequence dissimilarity = estimation of the number of evolutionary changes**
  But since last shared common ancestor

➢**The distance measures are used for building evolutionary tree but it must meet two criteria:** metric and additive

Metric

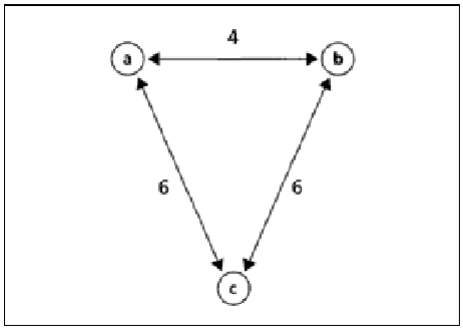
1 d(a, b) ³ 0	(non-negativity)
2 d(a, b) = d(b, a)	(symmetry)
3 d(a, c) £ d(a, b) + d(b, c)	(triangle inequality)
4 d(a, b) = 0 if and only if a = b	(distinctness)



- The triangle inequality. The distance between any pair of sequences must be no greater than that between those sequences and a third sequence.

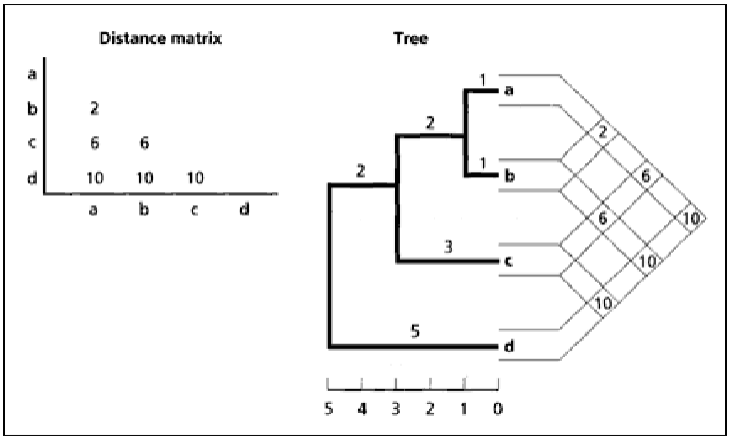
Ultrametric

- This criterion implies that the two largest distances are equal, so that they define an isosceles triangle



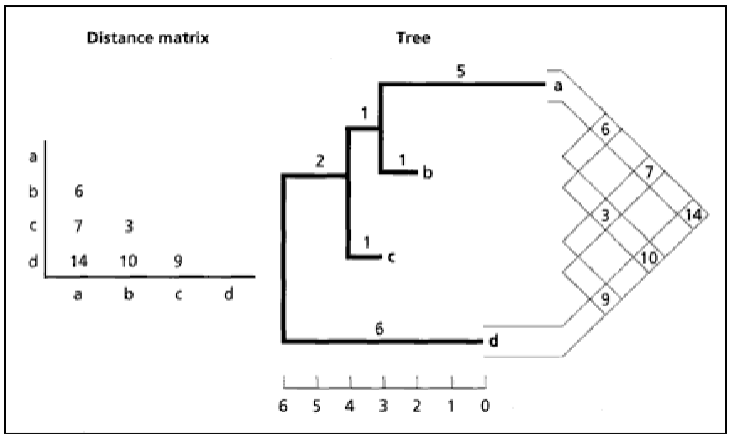
Tree Distances

- Ultrametric tree



- An ultrametric distance matrix between four sequences ad and the corresponding ultrametric tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

- Additive tree

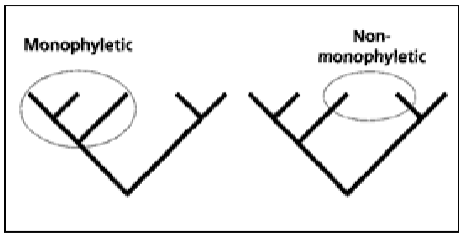


- An additive distance matrix between four sequences and the corresponding additive tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

Similarity versus Evolutionary relationship

Organismal phylogeny

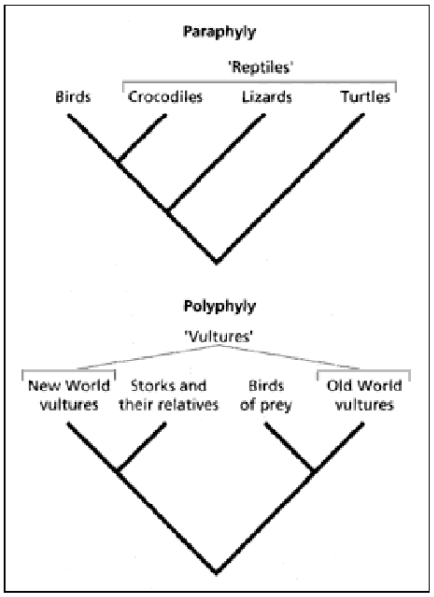
- Clade and classification
  - Monophyly



The difference between monophyly and non-monophyly. A monophyletic group includes all descendants of their common ancestor, whereas in a non-monophyletic group one or more descendant is not included.

Non-monophyly

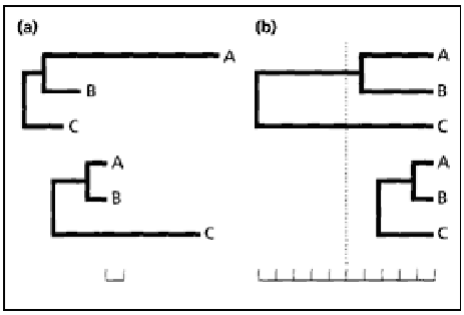
- Grouping paraphyletic group
- Grouping polyphyletic group



Cladistic classification

- Tell us little about the organisms themselves beyond who their nearest relatives are
- Cladistic classifications also have the great advantage of being immune to variation in rates of evolution (example)

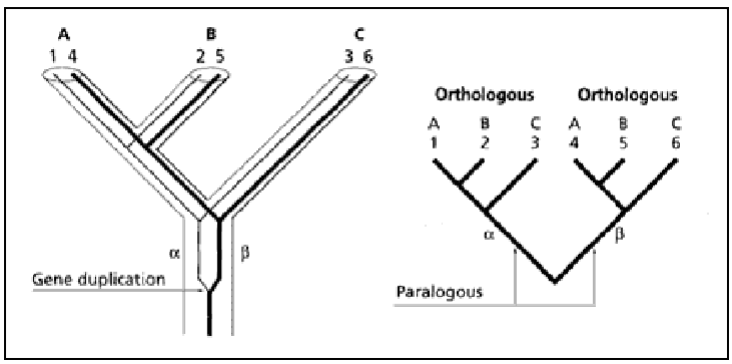
Examples of variation in rate of evolution among genes from the same organisms. For all four trees the cladistic group AB is preserved. Dashed line is an arbitrary threshold for placing species in different higher taxonomic groups.



Gene Trees and Species Trees?

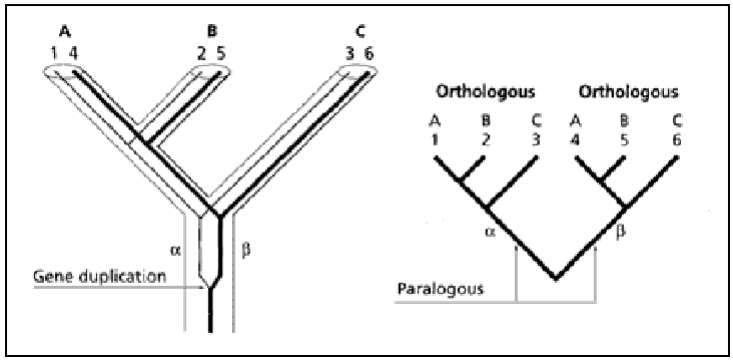
- Why phylogeny of genes do not match those of the organism.

  1- Gene duplication (orthologous and paralogous)



Why we need orthologous genes?

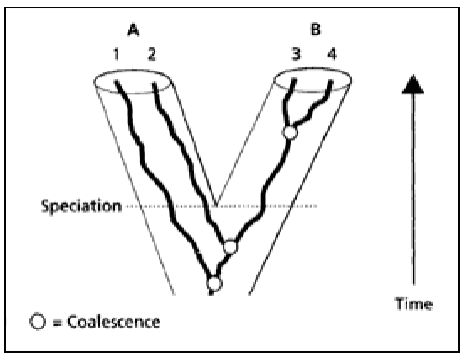
- Two homologous genes are orthologous if their most recent common ancestor did not undergo a gene duplication,
- otherwise they are termed paralogous, therefore give wrong signal



2- Lineage Sorting and Coalescence

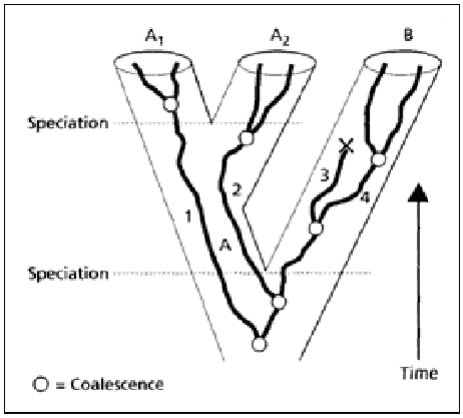
- Even if we restrict our attention to orthologous genes for the reason given above, the presence of ancestral polymorphism coupled with the differential survival of those alleles can result in allele phylogeny not matching organismal phylogeny.

- A gene tree for four alleles in two organismal lineages, A and B. The points at which pairs of allele lineages join (coalesce) are marked by open circles.
Alleles 3 and 4 coalesce within lineage B, but alleles 1 and 2 are older than lineage A. Note especially that alleles 1 and 2 do not form a monophyletic group 2 is more closely related to 3 and 4 than it is to the other allele (1) found in the same species.



Example

Imagine that shortly after species A and B diverged, and while alleles 1 and 2 were still both extant, species A itself became two, species A1 and A2

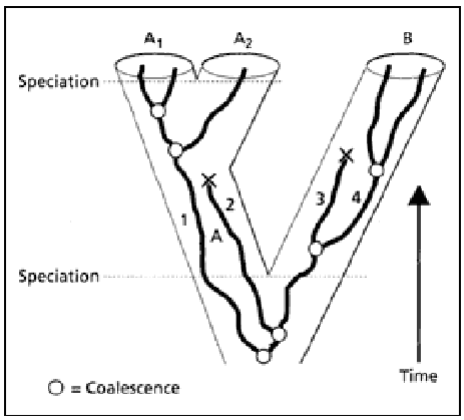


Lineage sorting is likely to be a problem for organismal phylogenetics if:

the time it takes for alleles within a lineage to coalesce is greater than the interval between successive speciation events.

When phylogeny faithfully reflects species phylogeny

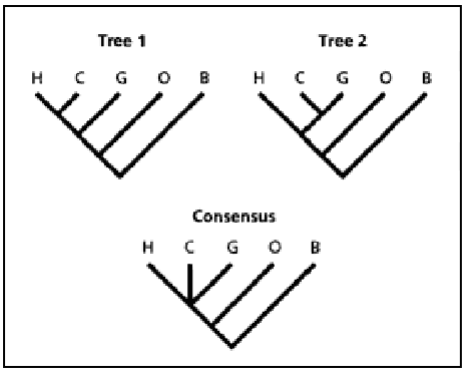
The same situation as last figure but lineage A speciating later in time, by which time allele 2 has gone extinct. Consequently species A1 and A2inherit a monophyletic set of alleles



- **The key difference between last two figures is the length of time between successive speciations of the same lineage. Due to a combination of chance and selection, allele lineages will either persist, radiate or go extinct. The longer the interval between speciation events the greater the chance that these processes will result in lineages with a monophyletic set of alleles.**

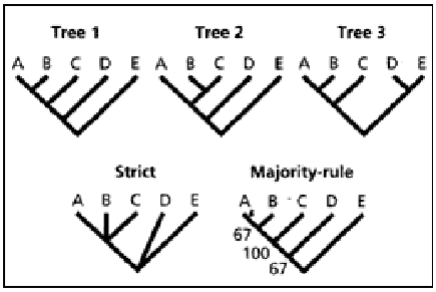
Consensus tree

- When we want to compare trees derived from different sequences, or from the same sequences using different methods



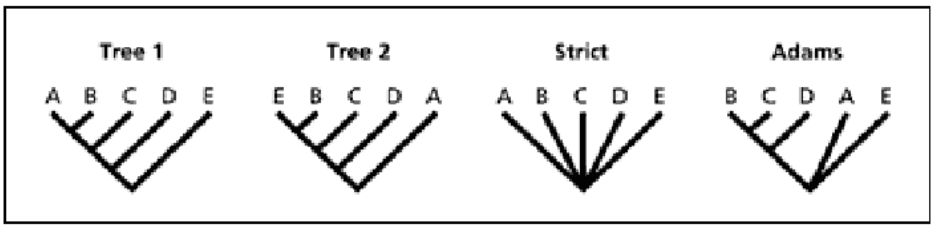
Types of consensus tree

- Strict consensus
- majority-rule consensus



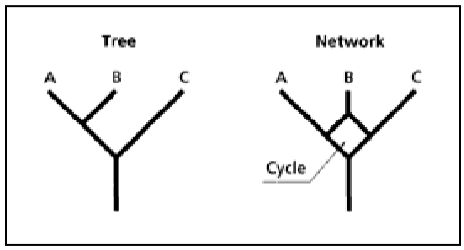
Types of consensus tree

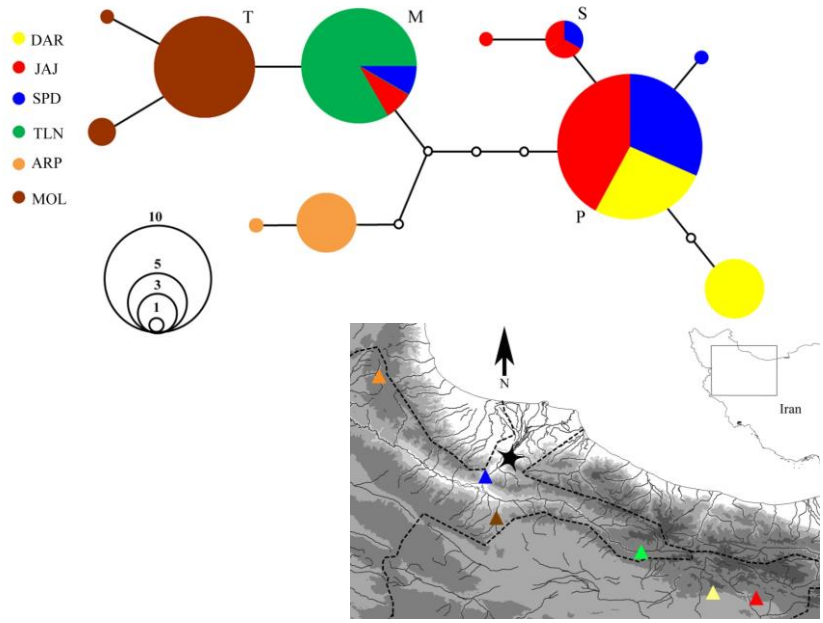
- Adams consensus tree



network

- A tree has single root and branches outwards such that the branches never meet, whereas in a family tree or pedigree every time a male and female organism mate their branches fuse. Generally the history of each individual gene can be adequately represented by a tree; however, in cases where a gene has undergone recombination a network may be more appropriate.





Summary

- 1- Evolutionary relationships can be represented by a variety of trees. Cladograms depict relative recency of common ancestry, additive trees incorporate branch lengths, ultrametric trees can be used to represent evolutionary time.
- 2- Trees may be either rooted or unrooted, but only rooted trees have an evolutionary direction.
- 3- The number of possible trees increases rapidly with increasing number of sequences.
- 4- Evolutionary trees can depict ancestordescendant relationships.
- 5- Distances satisfying the 'four-point condition' define a corresponding tree.
- 6- Gene trees may differ from species trees.

- **Molecular versus morphological in systematic**

Advantage of molecular data

- Large number of observable characters
  - How might we estimate the number of useful molecular characters
    - Characters independency
    - Inheritable

Advantage of molecular data

- Wide range of substitution rate
  - Help to recognize even distantly related lineage that is not possible using morphology

Advantage of molecular data

- Genetic base is known and non-independent but genetic base of morphological traits are not known.

Advantage of molecular data

- Characters can be selected and defined in an objective manners.
  - Although choice of the gene and alignment involve subjectivity.
  - But in morphological characters must be defined and delimited without explicit criteria (quite arbitrary).

Advantage of morphological data

- Easier, cheaper
- Extinct taxa
  - Understanding the relationship
  - Testing and rooting phylogenetic trees
- Evolution of genes may differ from the species evolution, because is based on one gene but in morphology

Advantage of morphological data

- Probably encoded by different genes
- Morphology plays crucial role in alpha taxonomy

Incongruence and conflict

- Reason is weak support for either or both of the estimates
- Using a single species to root a tree
- Phylogeny of genes differ from the phylogeny of species
  - Paralogy
  - Lineage sorting
  - Lateral transfer of genes between unrelated species

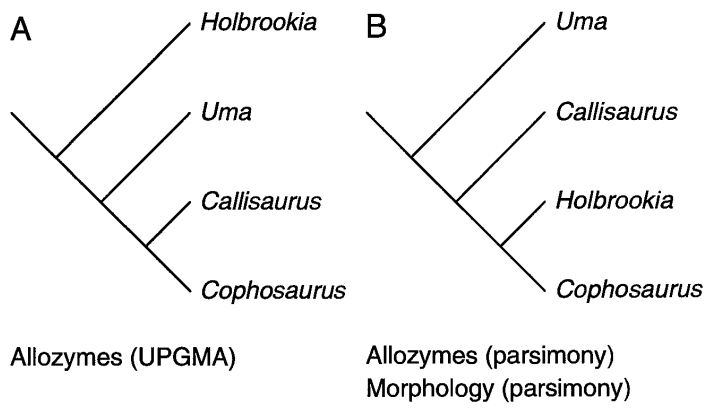


Figure 1.3. Apparent conflict between molecular and morphological data as an artifact of applying a different phylogenetic analysis method to each data set. **(A)** The phylogeny of the "sand lizard" clade of the family Phrynosomatidae based on allozyme data and UPGMA analysis (Adest 1978). **(B)** Sand lizard phylogeny based on morphology and parsimony analysis (Etheridge and de Queiroz 1988). The allozyme data used to construct tree A, when reanalyzed with the parsimony method, resulted in a tree identical to tree B (K. de Queiroz 1992).

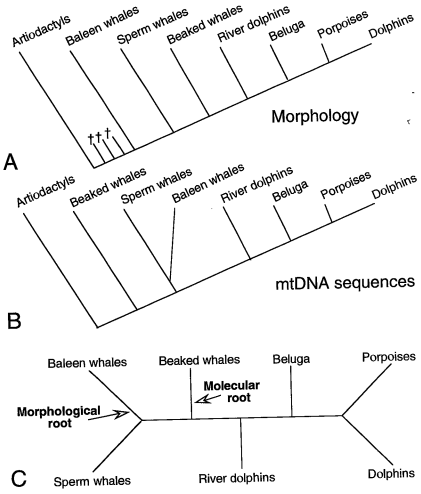


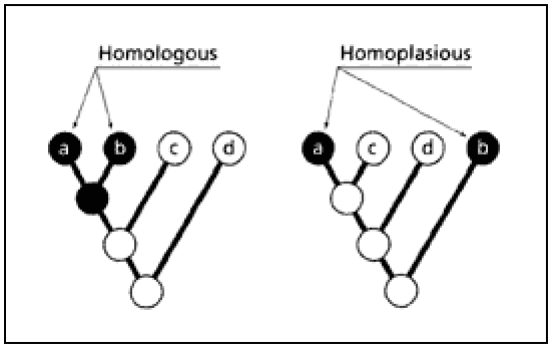
Figure 1.4. Apparent conflict between molecular and morphological data that is attributable to uncertain rooting of one of the phylogenetic trees (by use of a distant outgroup taxon). **(A)** The estimated phylogeny of the major whale taxa based on morphological data. The outgroups were three late Eocene fossil taxa (marked with daggers) that are closely related to modern whales but have some ancestral features such as pelvic limbs. Adapted from Messenger and McGuire (1998). **(B)** The estimated phylogeny of whales based on mitochondrial DNA sequences. Adapted from Milinkovitch et al. (1993, 1994). **(C)** Tree A and tree B superimposed. The two trees differ only in the position of the root. (Here, the "morphological root" is the root of tree A, and the "molecular root" is that of tree B.) Messenger and McGuire (1998) found that the position of the "molecular root" changed when the DNA data were reanalyzed with different combinations of extant artiodactyls used as the outgroup.

Misconception in the molecules versus morphology debate

- Sensitivity to convergent evolution

Measuring Genetic Change

- homologous or homoplasious
  - Wings of bats and birds



parallel evolution of amino acid sequences in the lysozyme enzyme in leaf-eating langur monkeys and in cows

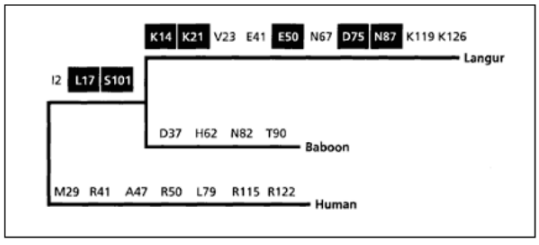
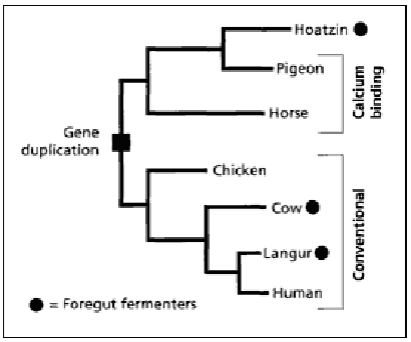


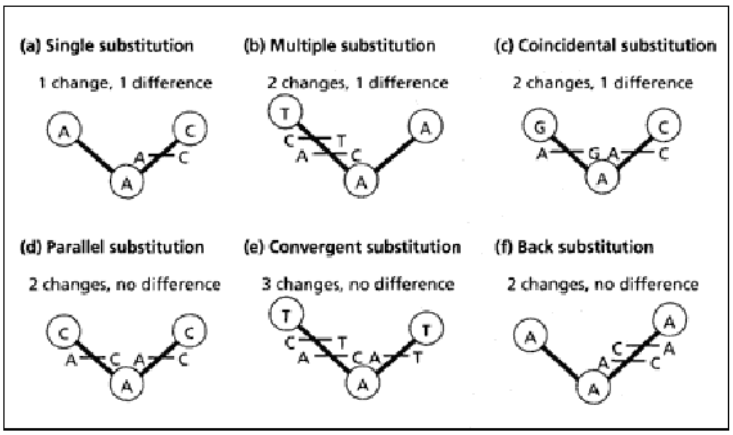
Fig. 5.2
Independent evolution of amino acid replacements in cows and langur monkeys. Although langur monkey lysozyme is phylogenetically closely related to other primate lysozymes it has independently acquired several amino acid substitutions in common with cow lysozyme (these are indicated by the black squares). Redrawn from Li and Graur (1991).

Homology can sometimes be difficult to distinguish from homoplasy, especially at the molecular level



Phylogeny of some bird and mammal lysozymes. The lysozyme in cows, langur monkeys and the hoatzin bird have all independently evolved similar digestive properties. The two mammalian genes are orthologous, and are paralogous with respect to the hoatzin gene, which is related to calcium binding lysozymes found in birds and mammals. After Kornegay et al. (1994).

Kind of substitution



Homology among Sequences

Sequence 1 A T G C G T C G T T

Sequence 2 A T G C G T C G T

```
A T G C G T C G T T
| | | | | | | | |
A T G C G T C G T
```

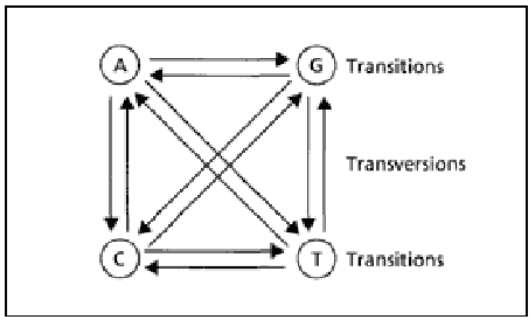
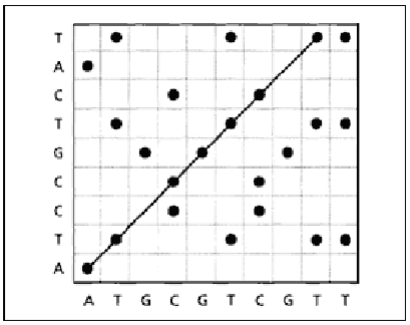


Fig. 5.10
The possible substitutions among the four nucleotides.

cost of alignment

$$D = s + wg$$



➢If indels are thought to be rare then w should be large; conversely, if indels are frequent then low values of ware more appropriate

Mutation rate

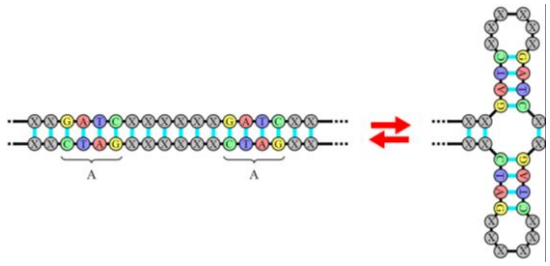
- **Nucleus**
- **Mitochondria**
  - relative ease with which they can be amplified (stable and numerous copies per cell)
  - they are only maternally inherited and lack introns and recombination
  - higher mutation rates and thus greater variability than nuclear DNA

shortcomings of mtDNA

- discovery of selective sweeps,
- mitochondrial introgressions,
- nuclear copies of mtDNA (numts)

Mutation rate

- Chloroplast
- Hotspots
  - 5` CG 3`
  - 5` TT3`
  - Repetition - palindrome repeat



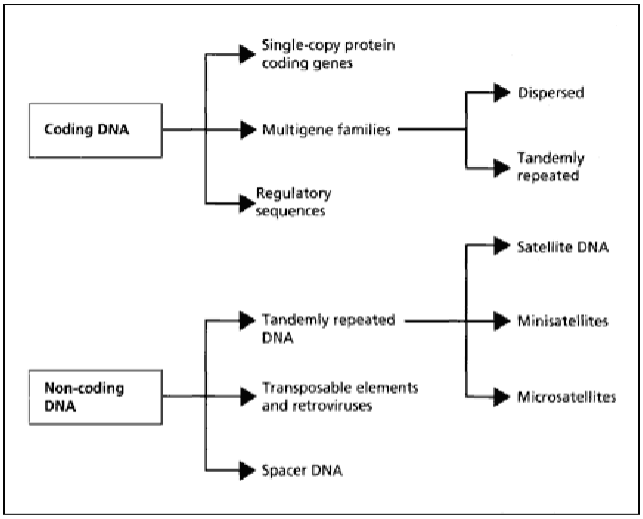


Fig. 3.17
A roadmap of the eukaryote genome. Non-coding regions within genes, such as introns, are not considered separately here.

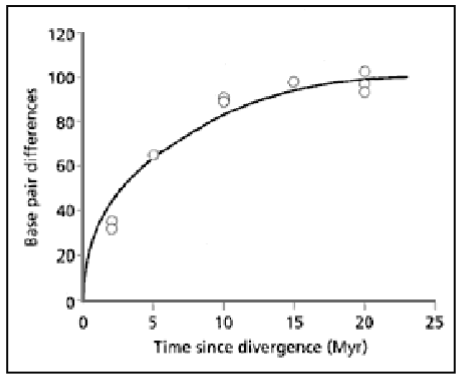


Fig. 5.11
Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al.* (1996).

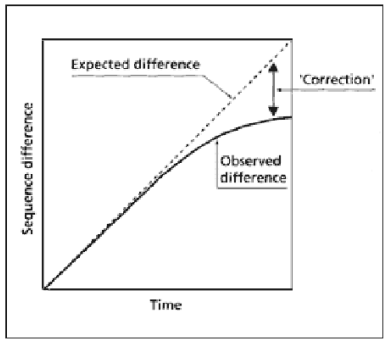


Fig. 5.12
The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

Evolutionary models

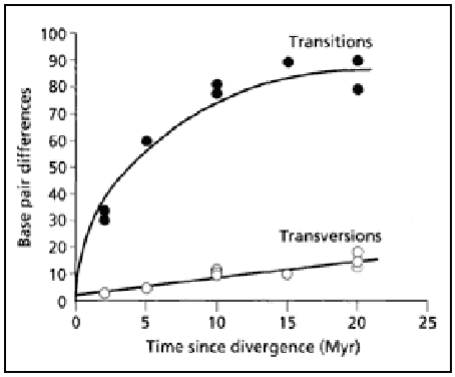
- JukesCantor (JC)

$$\mathbf{P}_t = \begin{bmatrix} . & \alpha & \alpha & \alpha \\ \alpha & . & \alpha & \alpha \\ \alpha & \alpha & . & \alpha \\ \alpha & \alpha & \alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}]$$

Evolutionary models

- Kimura's 2 Parameter Model (K2P)

Fig. 5.13
The number of transitions and transversions between the same bovid mammal sequences used in Fig. 5.11. Transitions accumulate much more rapidly than transversions and become saturated, whereas transversions accumulate more slowly and show no evidence of saturation.



Kimura's 2 Parameter Model (K2P)

$$\mathbf{P}_t = \begin{bmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{bmatrix}, \qquad \mathbf{f} = [\tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}\ \tfrac{1}{4}].$$

Evolutionary models

- Felsenstein (1981)
- where pi is the frequency of the ith base averaged over the sequences being compared. Note that if pA= pC=pG= pT= then the F81 model is the same as the JC model.

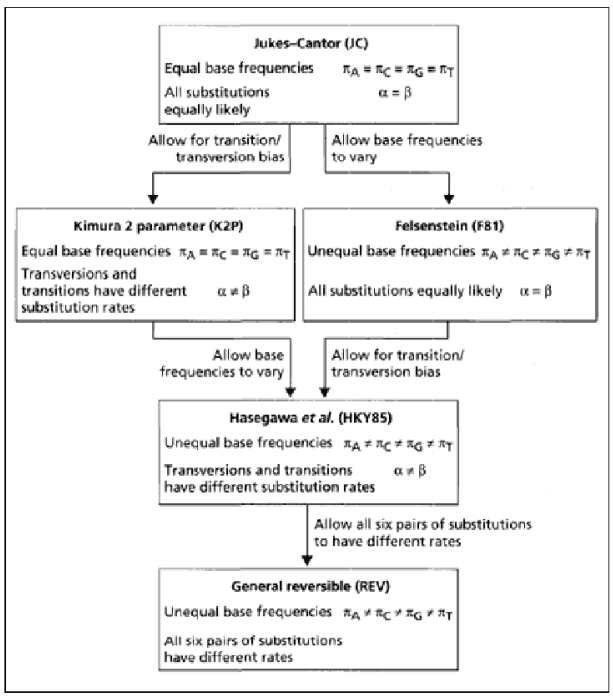
The F81 model has the following form:

$$\mathbf{P}_t = \begin{bmatrix} . & \pi_C\alpha & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & . & \pi_G\alpha & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & . & \pi_T\alpha \\ \pi_A\alpha & \pi_C\alpha & \pi_G\alpha & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A\ \pi_C\ \pi_G\ \pi_T]$$

Evolutionary models

- Hasegawa, Kishino and Yano (1985)

$$\mathbf{P}_i = \begin{bmatrix} . & \pi_C\beta & \pi_G\alpha & \pi_T\beta \\ \pi_A\beta & . & \pi_G\beta & \pi_T\alpha \\ \pi_A\alpha & \pi_C\beta & . & \pi_T\beta \\ \pi_A\beta & \pi_C\alpha & \pi_G\beta & . \end{bmatrix}, \qquad \mathbf{f} = [\pi_A\ \pi_C\ \pi_G\ \pi_T]$$



How can we decide whether our hypothesis is an adequate explanation of the data?

- flipping a coin **Likelihood** $L = \Pr(D|H)$

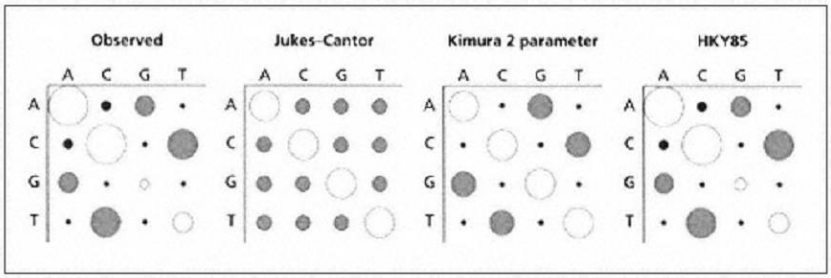


Fig. 5.15
Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

All the methods we have discussed share these assumptions:

- All nucleotide sites change independently.
- The substitution rate is constant over time and in different lineages.
- The base composition is at equilibrium.
- The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.

While these assumptions make the methods tractable they are in many cases unrealistic. We consider some of these assumptions below.

Independence

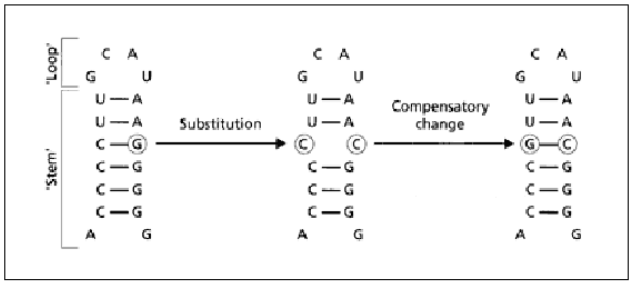


Fig. 5.16
RNA molecules have a secondary structure comprising 'stems' of WatsonCrick paired nucleotides and 'loops' of unpaired nucleotides. A substitution in a stem can destroy the WatsonCrick bond and reduce the stability of the molecule. A substitution that restores the stem is a compensatory change. After Hickson *et al.* (1996).

Base composition

- over the collection sequences being studied the base composition is roughly the same. Deviations from this assumption do occur and can lead to problems inferring the correct evolutionary tree

Log Det to solve the problem

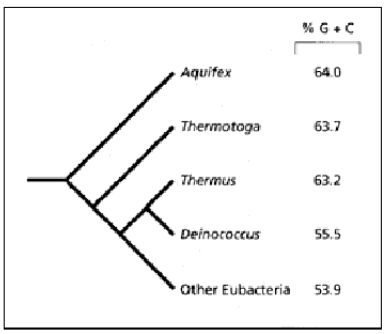


Fig. 5.17
art of the phylogeny of bacteria showing the variation in percentage G + C content in small subunit rRNA. After Galtier and Gouy (1995: Fig. 4).

Variation in Rates of Substitution among Sites

- one assumption that they all models mentioned is that each nucleotide site in a sequence is equally likely to undergo a substitution.
  - Pseudogenes, which have lost all functionality evolve most rapidly, with fourfold degenerate sites close behind. As we might expect, non-degenerate sites (at which any nucleotide substitution results in an amino acid replacement) evolve relatively slowly.

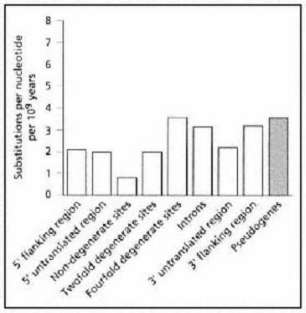


Fig. 5.18
Average rates of substitution in different parts of mammalian genes and pseudogenes. From Li and Graur (1991).

If some sites are not free to vary then sequences that evolve at a fast rate can, over evolutionary time, paradoxically show less divergence than more slowly evolving sequences that have fewer constraints

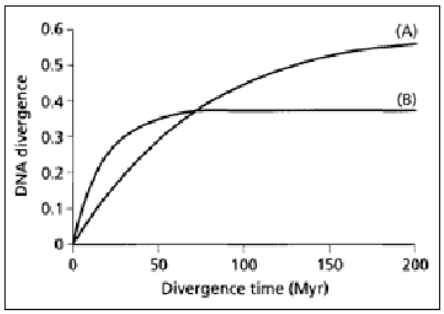


Fig. 5.19
DNA sequence divergence plotted against time since divergence. For the upper curve (A) the rate of substitution is 0.5%/Myr and 80% of sites are free to vary, whereas for the lower curve (B) the rate of substitution is higher (2%/Myr) but only half the sites are free to vary. After Palumbi (1989).

Distribution of Rates

- sites show a range of probabilities of substitution, rather than simply the two categories of zero and non zero

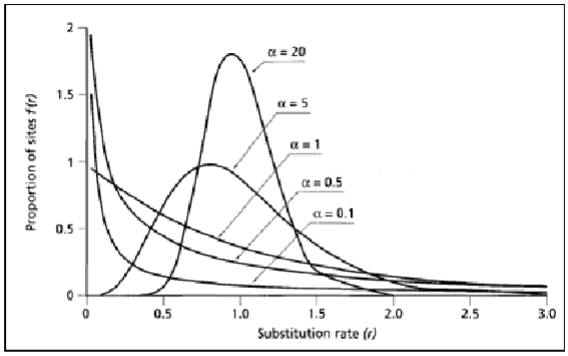


Fig. 5.20
The distribution of relative substitution rate *r* corresponding to different values of the gamma shape parameter a. Low a corresponds to large rate variation. As a gets larger the range of variation diminishes, until as a approaches ¥ all sites have the same substitution rate. After Yang (1996: Fig. 1).

Chapter 3
Genes:
Organisation, Function and Evolution

3.1 Levels of Genetic Organisation

3.1.1 DNA, Proteins and Chromosomes

3.1.2 The Genetic Code

3.1.3 Mitochondria and Chloroplasts

3.1.4 The Structure of Genes

3.1.5 Multigene Families

3.2 How Genes Function

3.2.1 DNA Replication

3.2.2 Protein Synthesis

3.2.3 Mutation

3.2.4 Recombination

3.3 Genome Organisation and Evolution

3.3.1 Species Differ in Genome Size and Gene Number

3.3.2 The Evolution of Multigene Families

3.3.3 Non-Coding Repetitive DNA Sequences

3.4 Summary

Levels of Genetic Organization

- **DNA has two extremely important properties:**
  - it contains the instructions for how organisms should be put together in the enormous variety of ways that characterizes life on earth, and
  - second it can be copied or replicated so that these instructions are passed on to success

Errors, or mutations, continually arise which provide the raw material upon which evolution works.

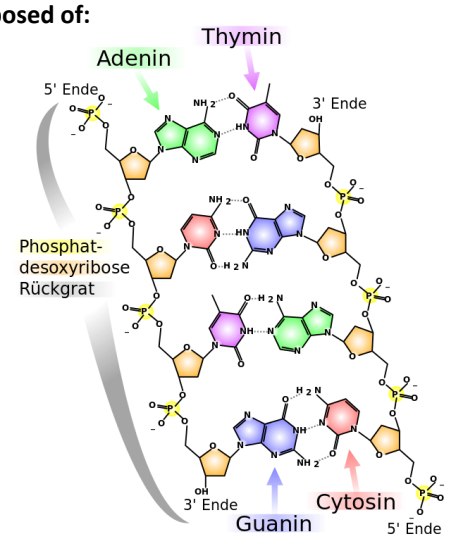
This chapter:

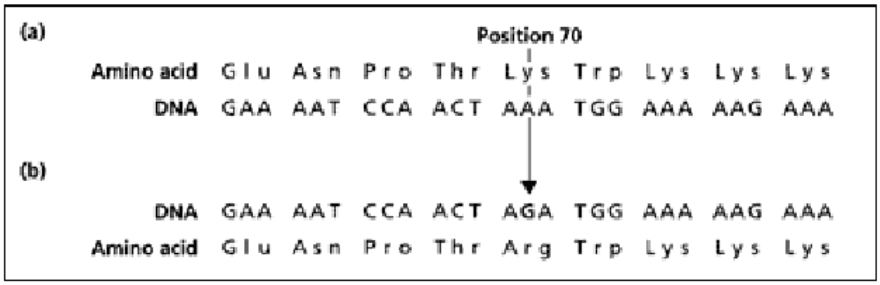
- How genetic information is organised ?

- How this organization evolved, as well ?

DNA, Proteins and Chromosomes

- **DNA is known as a nucleic acid composed of:**





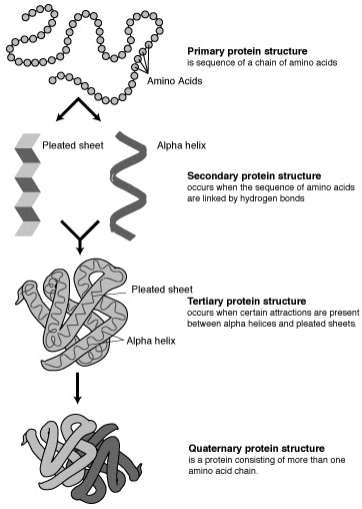
A DNA sequence. Part (a) shows the DNA sequence from part of the pol gene of HIV-1 and the amino acids this sequence encodes. Part (b) shows an A to G mutation at amino acid 70 (lysine changes to arginine) which confers resistance to the drug AZT.

What is RNA world?

- **The first living systems on earth may have been composed of RNA rather than DNA**

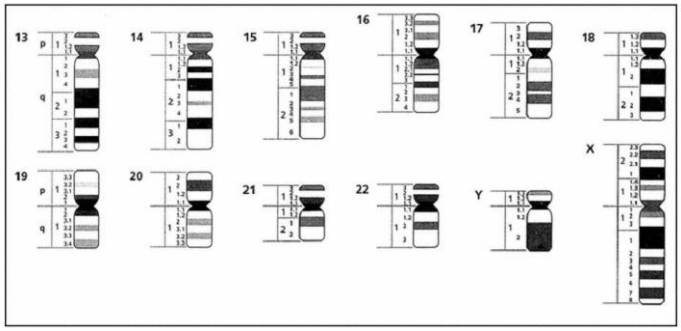
Definitions

- Non coding DNA
- Coding DNA
  - Structural proteins
  - Regulatory proteins



Chromosome

- Chromatin
  - Euchromatin
  - Heterochromatin



Genetic code

- 20 amino acid versus 64 genetic code
  - N fold degenerated site

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Mutation

- Substitutional mutation
- recombinations
- Deletions
- Insertions
- inversions

Evolution is just this

Substitutional mutation

- Transition
- Transversion

- Synonymous (silent)
- Asynonymous
  - Probability of amino acid change according to the changed nucleotide position

- recombinations
- Deletions
- Insertions
- Inversions

Recombinations

- Reciprocal
- Gene conversion

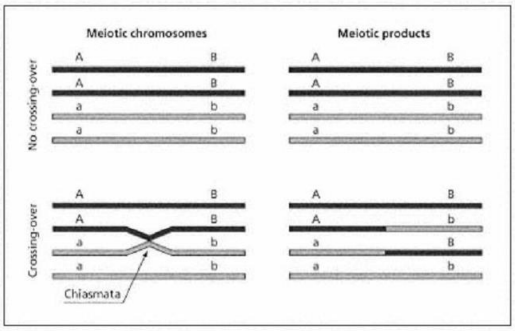


Fig. 3.14
Crossing-over between non-sister chromatids on duplicated homologous chromosomes (one shaded black, the other grey) during meiosis. Two loci (A and B) with two alleles at each are shown. In the top example, no crossing-over takes place but in the bottom one crossing-over leads to a new combination of alleles in the meiotic products. Adapted from Griffiths *et al.* (1993).

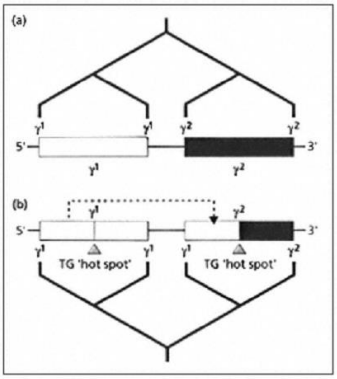


Fig. 3.16
Gene conversion in the duplicated g-globin genes of primates. Part (a) depicts the expected relationships between the duplicated g1 and g2 genes from different species: each gene evolves independently so that g1 sequences are more closely related to other g1 sequences than they are to g2 sequences. However, in part (b) a gene conversion event occurs such that the 5' part of g1 is superimposed on the 5' part of g2. This means that the 5' region of g2 is more closely related to g1 than it is to the g2 sequence found in other species. The boundary for this conversion event is marked by the repeated TG 'hotspot' sequence.

problem

- **Fewer parameters might give an inaccurate estimate; more parameters may decrease the precision of our estimate.**

Distance Measures for Protein Sequences

1st letter	Second Letter: U	Second Letter: C	Second Letter: A	Second Letter: G	3rd letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

$$d = -\ln(1 - p - 0.2p^2)$$

Measuring Evolutionary Change on a Tree

- Parsimony explicitly seeks to reconstruct the ancestral sequences themselves, rather than just the edge lengths.

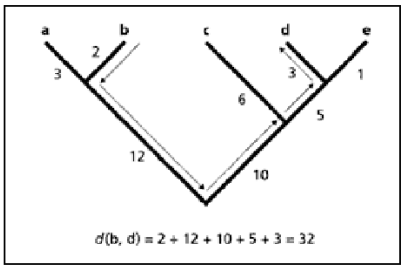
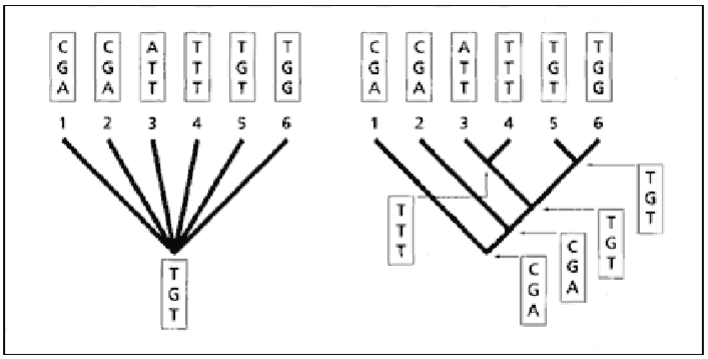
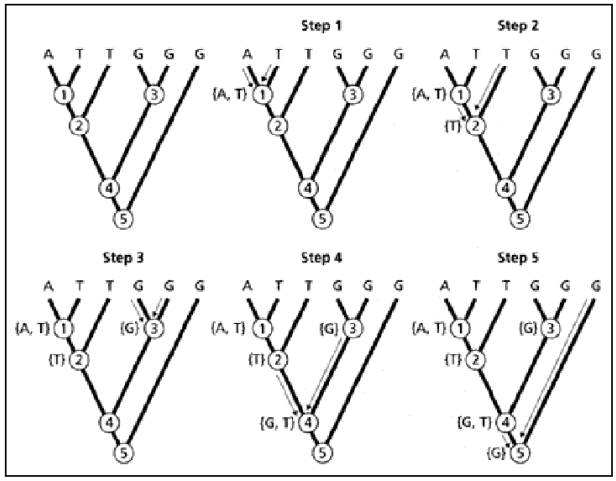


Fig. 5.21
The evolutionary distance between b and d is the sum of the edge lengths along the path in the tree between the two sequences.

- Such an approach is often used to construct a consensus sequence which summarises the common features of a set of sequences
- However, a consensus sequence is not necessarily the same as an ancestral sequence
- But
  - it should not be confused with an ancestral sequence. This is because sequence consensus methods implicitly assume that sequences are equally related that is by a star tree

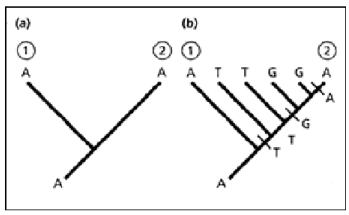


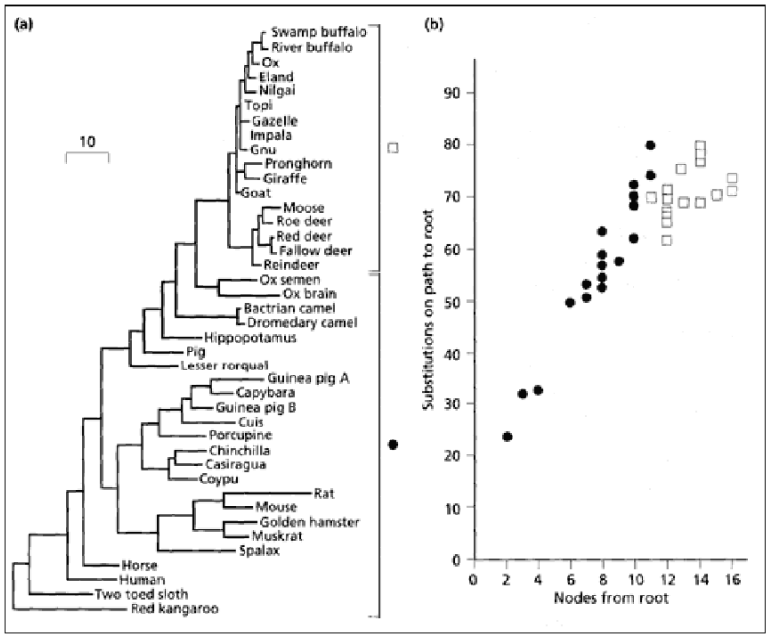
example



Estimating Branch Lengths

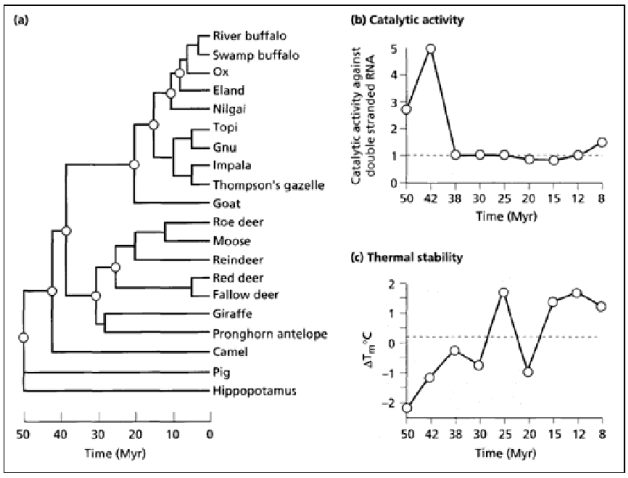
- Just because two sequences may have the same nucleotide at the same position, this does not mean that there has been no evolutionary change between those two sequences





Testing Ancestral Reconstructions

- Result: Specifically, the earliest artiodactyls had much higher levels of activity against double-stranded RNA and were slightly less thermally stable



Inferring Molecular Phylogeny

The methods for constructing phylogenetic trees from molecular data can be grouped according to the kind of data they use,

•Kinds of Data:

Distances Versus Discrete Characters
Clustering Methods Versus Search Methods

Distance Methods

Goodness of Fit Measures
Minimum Evolution

Discrete Methods

Maximum Parsimony
Maximum Likelihood
Parsimony and Likelihood

1- Kinds of Data: Distances Versus Discrete Characters state

- Character-state methods retain the original character status of the taxa and therefore, can be used to **reconstruct the character state of ancestral nodes.**
- In contrast, distance-matrix methods start by calculating some measure of the dissimilarity of each pair of OTUs to produce a pairwise distance matrix, and then infer the phylogenetic relationships of the OTUs from that matrix.

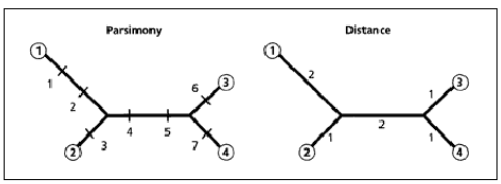


Fig. 6.1
A parsimony tree and a distance tree for the same sequence data. Note that both trees have the same topology and branch lengths, but that the parsimony tree identifies which site contributes to the length of each branch.

sequences

sites

sequences	1	2	3	4	5	6	7
1	T	T	A	T	T	A	A
2	A	A	T	T	T	A	A
3	A	A	A	A	A	T	A
4	A	A	A	A	A	A	T

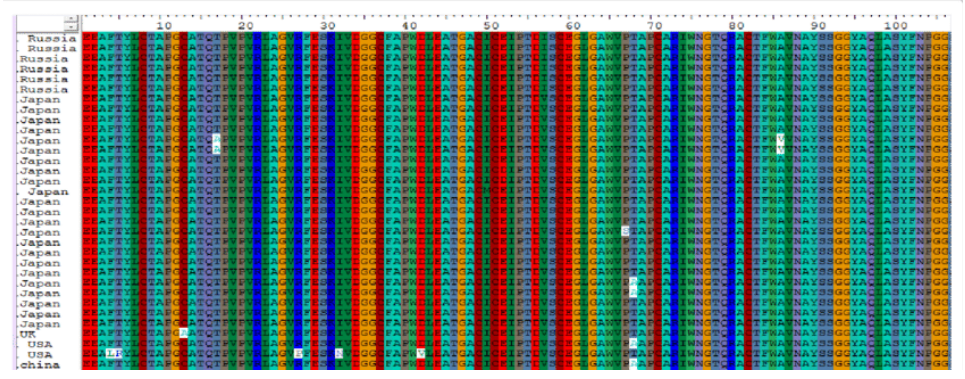
distances

sequences	**1**	**2**	**3**
2	3		
3	5	4	
4	5	4	2

sequences

- **The major advantage of distance methods** is that they are generally computationally inexpensive, which is important when many taxa have to be analyzed.

The major advantage of distance methods computationally inexpensive, which is important when many taxa have to be analyzed.



2- Clustering Methods Versus Search Methods

- What is the **clustering method**?
- Clustering methods have the advantage of being easy to implement, resulting in very **fast computer programs**. Furthermore they almost always produce **a single tree**.

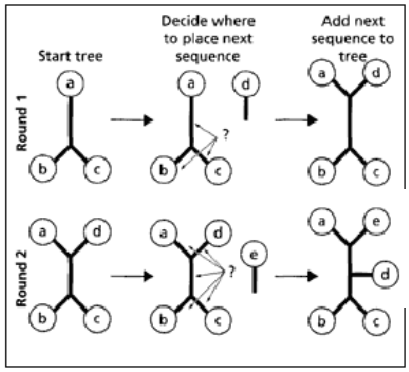


Fig. 6.2
An example of how a clustering method builds a tree. The tree is constructed by starting with the tree for three sequences, then adding each remaining sequence in turn until finally all sequences have been added.

Shortcomings

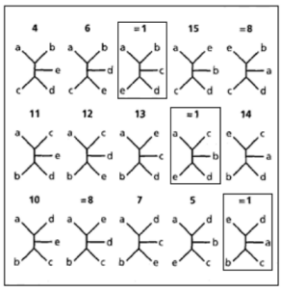
- simple clustering algorithms often **depends on the order** in which we add the sequences to the growing tree.

- But the biggest limitation is that cluster methods do not allow us to evaluate competing hypotheses since it results in one tree

Optimality criteria

- This criterion is used to assign to each tree a 'score' or rank which is a function of the relationship between tree and data (examples include maximum parsimony and maximum likelihood)

- The Achilles' heel of optimality very expensive.
  - firstly, for a given data set and a given tree, what is the value of the optimality criterion for that tree? For example, what is the minimum number of evolutionary events required to explain the observed data?
  - Secondly, which of all the possible trees has the maximum value of this criterion?



Comparison

- The majority of distance-matrix methods use clustering algorithms to compute the “best” tree, whereas most character-state methods employ an optimality criterion

Table 1.5 Classification of phylogenetic analysis methods and their strategies

	Optimality search criterion	Clustering
Character state	Maximum parsimony (MP)	
	Maximum likelihood (ML)	
	Bayesian inference	
Distance matrix	Fitch–Margoliash	UPGMA
		Neighbor-joining (NJ)

- The majority of distance-matrix methods use clustering algorithms to compute the “best” tree, whereas most character-state methods employ an optimality criterion

how can we decide which tree is better than others?

- Given the range of tree-building methods available, David Penny and colleagues have suggested five desirable properties a tree-building method should have:
- efficiency (how fast is the method)?
- power (how much data does the method need to produce a reasonable result)?
- consistency (will it converge on the right answer given enough data)?
- robustness (will minor violations of the method's assumptions result in poor estimates of phylogeny)?
- falsifiability (will the method tell us when its assumptions are violated, i.e. that we should not be using the method at all)

Niebuhrjoining tree