BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html

July 2011

Darwin: Evolution & Natural Selection

- Charles Darwin's 1859 book (*On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*) introduced the Theory of Evolution.
- Struggle for existence induces a natural selection. Offspring are dissimilar from their parents (that is, variability exists), and individuals that are more fit for a given environment are selected for. In this way, over long periods of time, species evolve. Groups of organisms change over time so that descendants differ structurally and functionally from their ancestors.

Slide by Pevsner

Dominant View of Evolution

- All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.
- Organization: **Directed Rooted Tree**; Existing species: Leaves; Common ancestor species (divergence event): Internal node; Length of an edge: Time.

Five kingdom system (Haeckel, 1879)

Slide by Pevsner

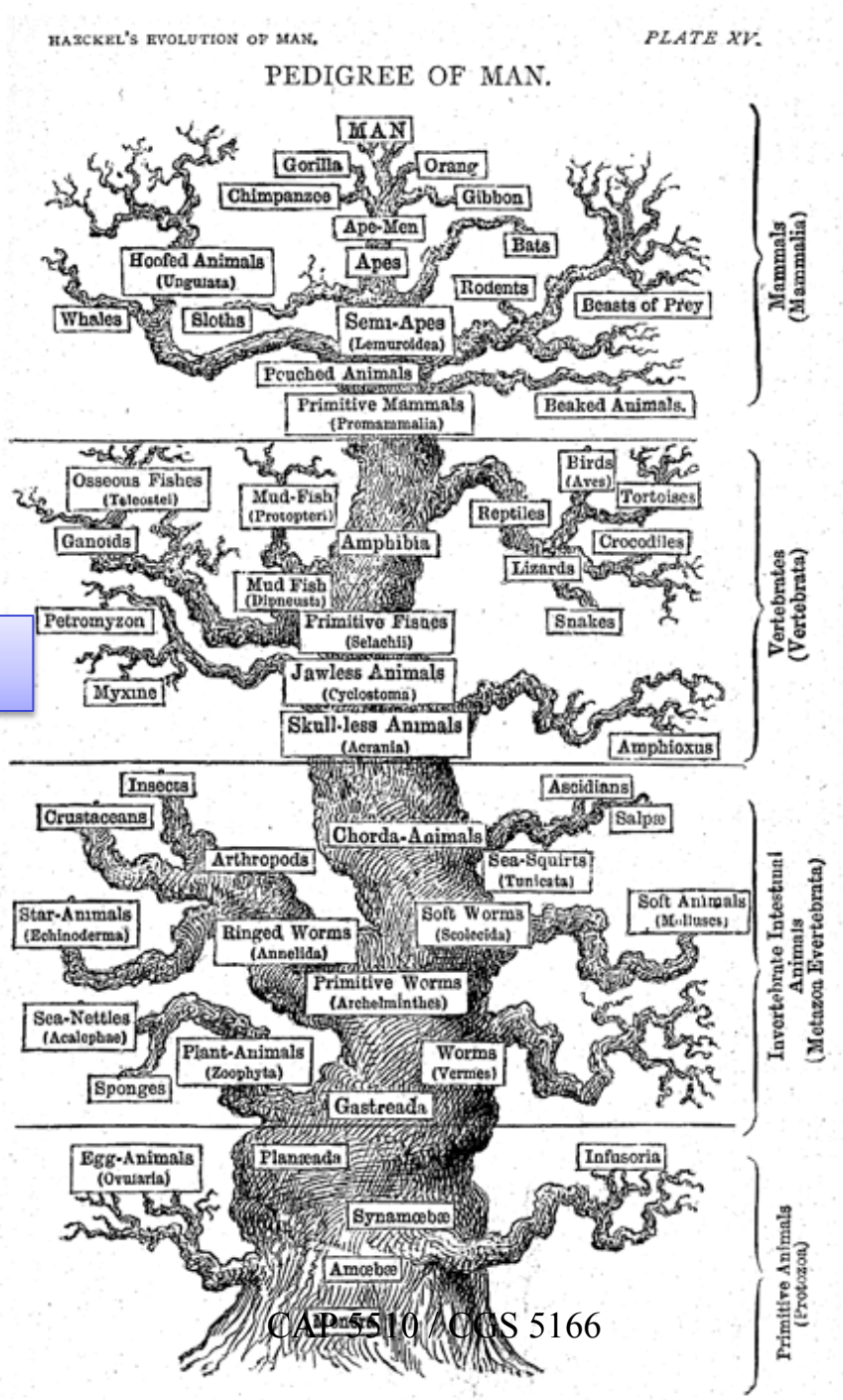
animals

plants

fungi

protists

monera



mammals

vertebrates

invertebrates

protozoa

Evolution & Phylogeny

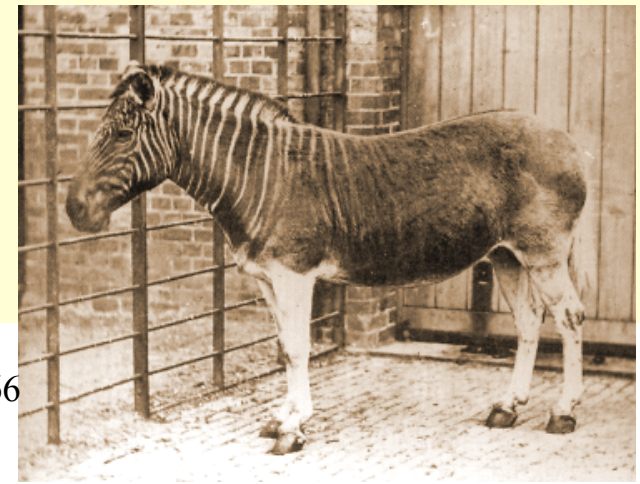
- At the molecular level, evolution is a process of mutation with selection.
- Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life.
- Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny relied on the comparison of morphological features between organisms. Today, molecular sequence data are also used for phylogenetic analyses.

Slide by Pevsner

Questions for Phylogenetic Analysis

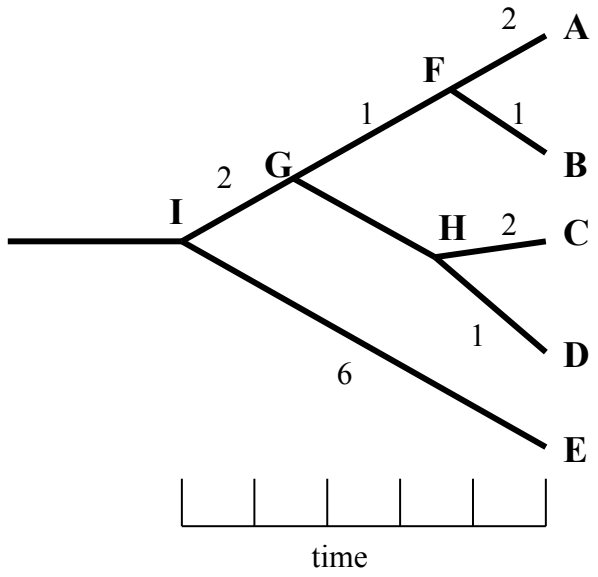
- How many genes are related to my favorite gene?
- How related are whales, dolphins & porpoises to cows?
- Where and when did HIV or other viruses originate?
- What is the history of life on earth?
- Was the extinct quagga more like a zebra or a horse?

Slide by Pevsner



Phylogenetic Trees

- Molecular phylogeny uses trees to depict evolutionary relationships among organisms. These trees are based upon DNA and protein sequence data.



Slide by Pevsner

Tree Nomenclature

Slide by Pevsner

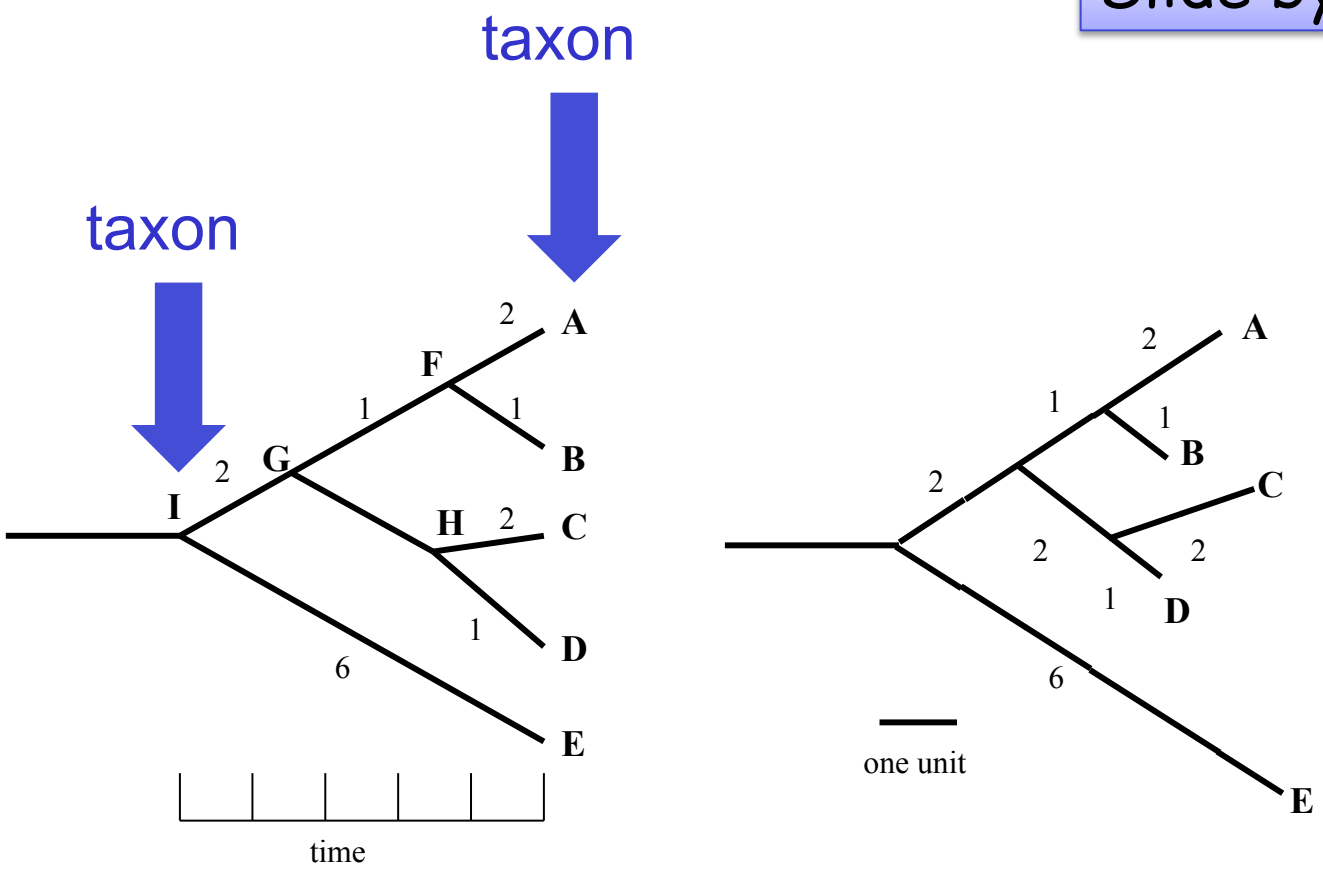


Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

operational taxonomic unit (OTU)

such as a protein sequence

taxon

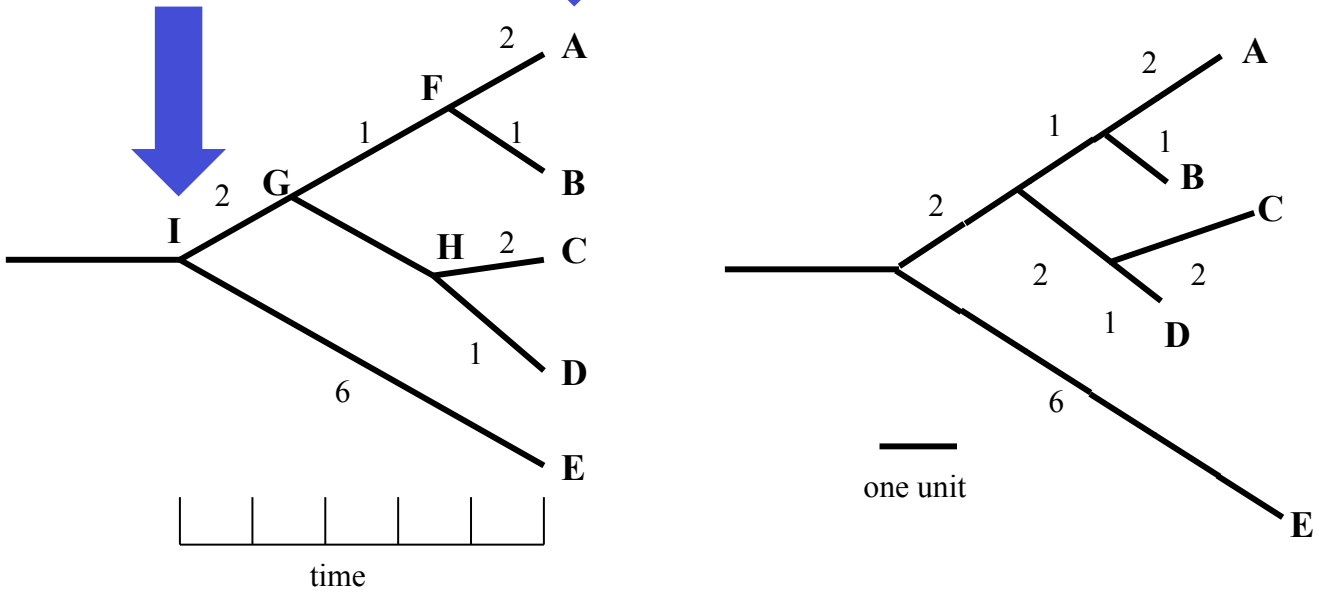


Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

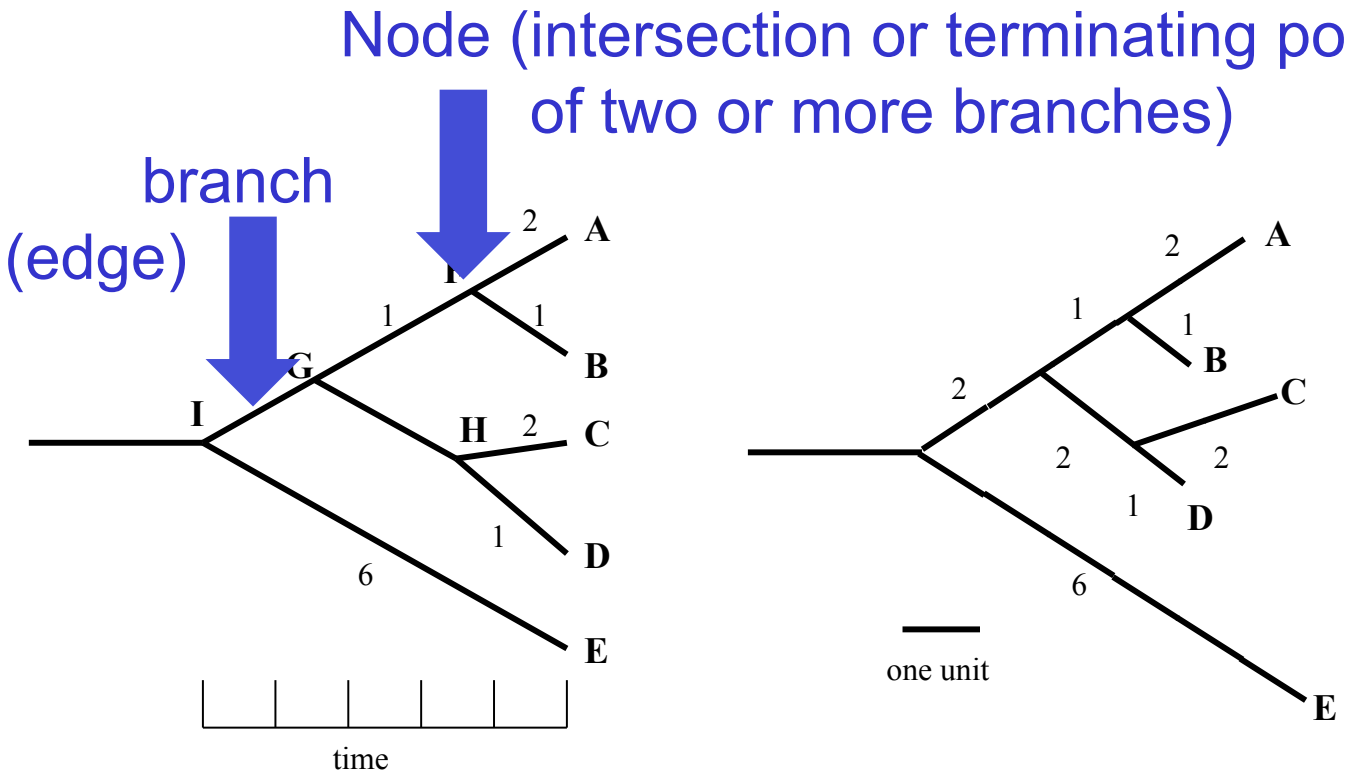
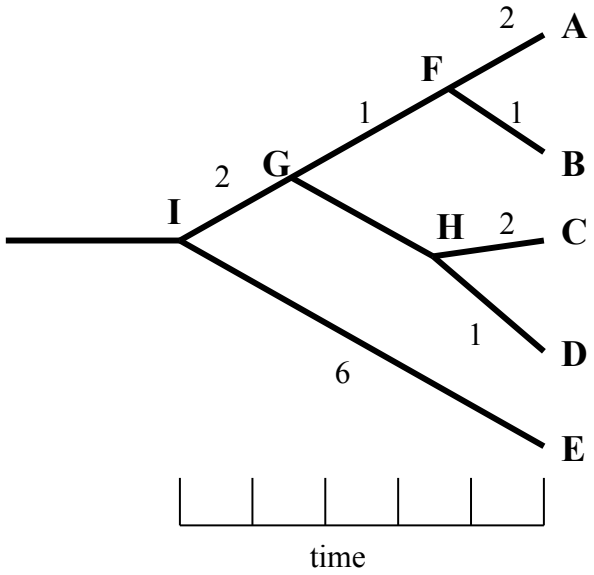


Fig. 7.8
Page 232

Tree nomenclature

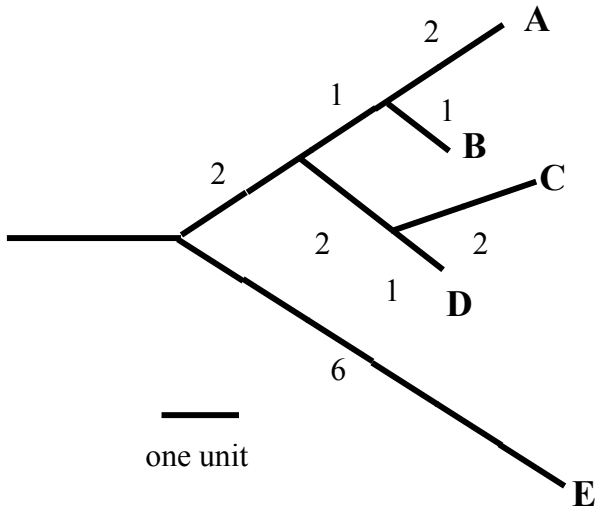
Slide by Pevsner

Branches are unscaled...



…OTUs are neatly aligned, and nodes reflect time

Branches are scaled...



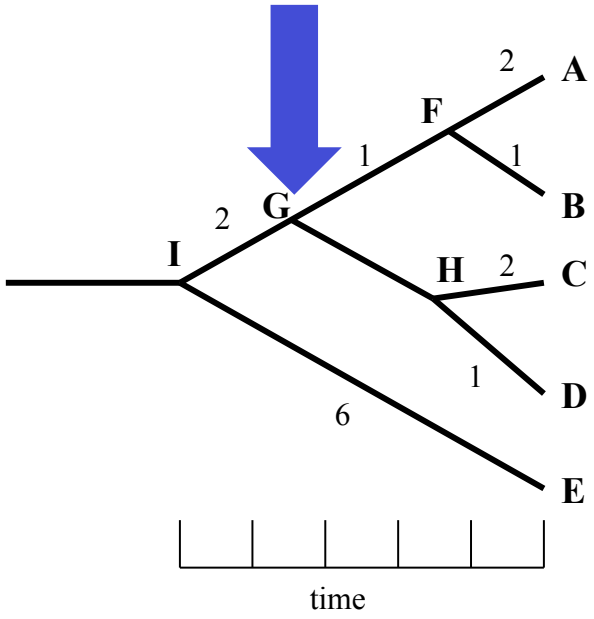
…branch lengths are proportional to number of amino acid changes

Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

bifurcating internal node



multifurcating internal node

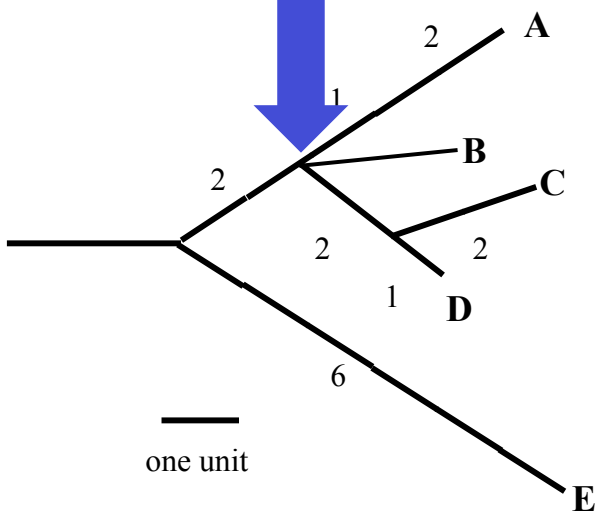
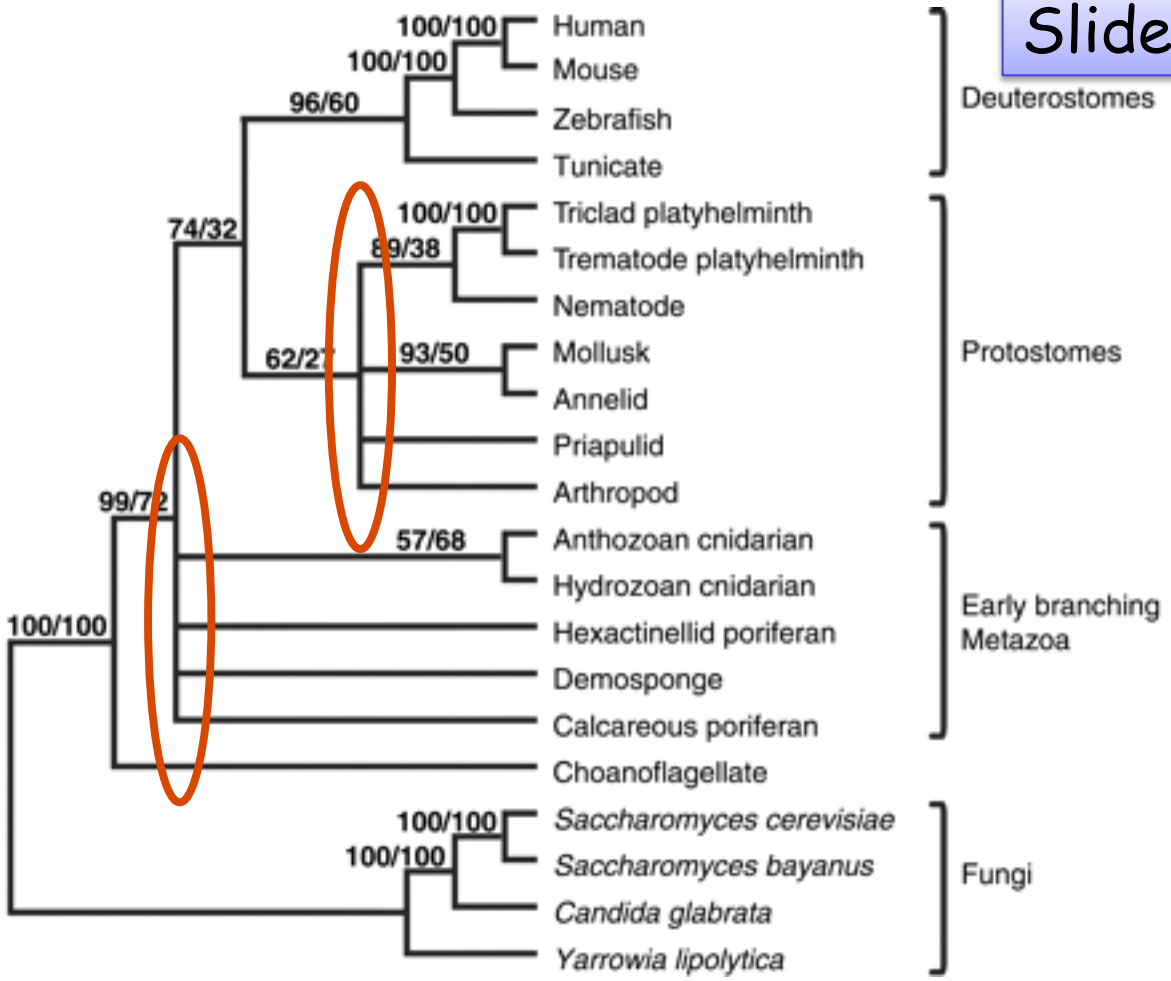


Fig. 7.9
Page 233

Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes

Slide by Pevsner



Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933 (2005), Fig. 1.

Tree nomenclature: clades

Slide by Pevsner

Clade ABF (monophyletic group)

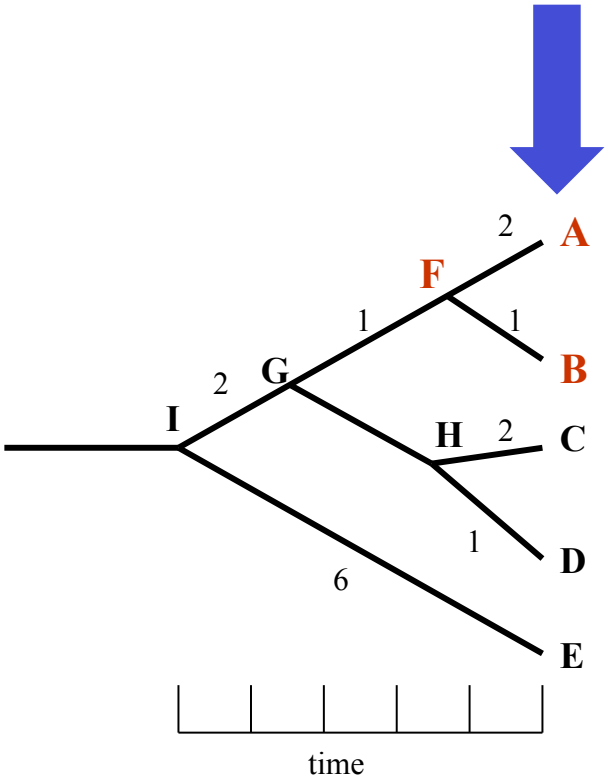
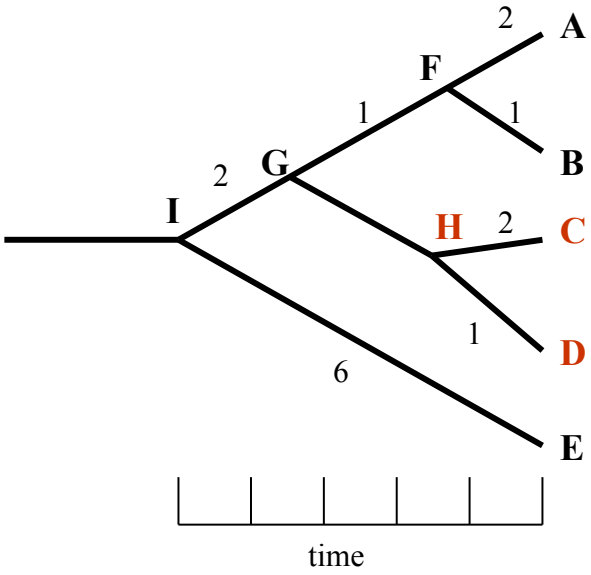


Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner



Clade CDH

Fig. 7.8
Page 232

Tree nomenclature

Slide by Pevsner

Clade ABF/CDH/G

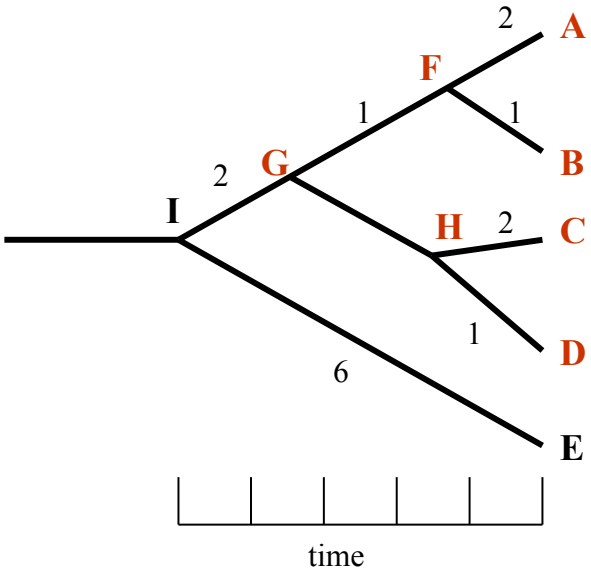
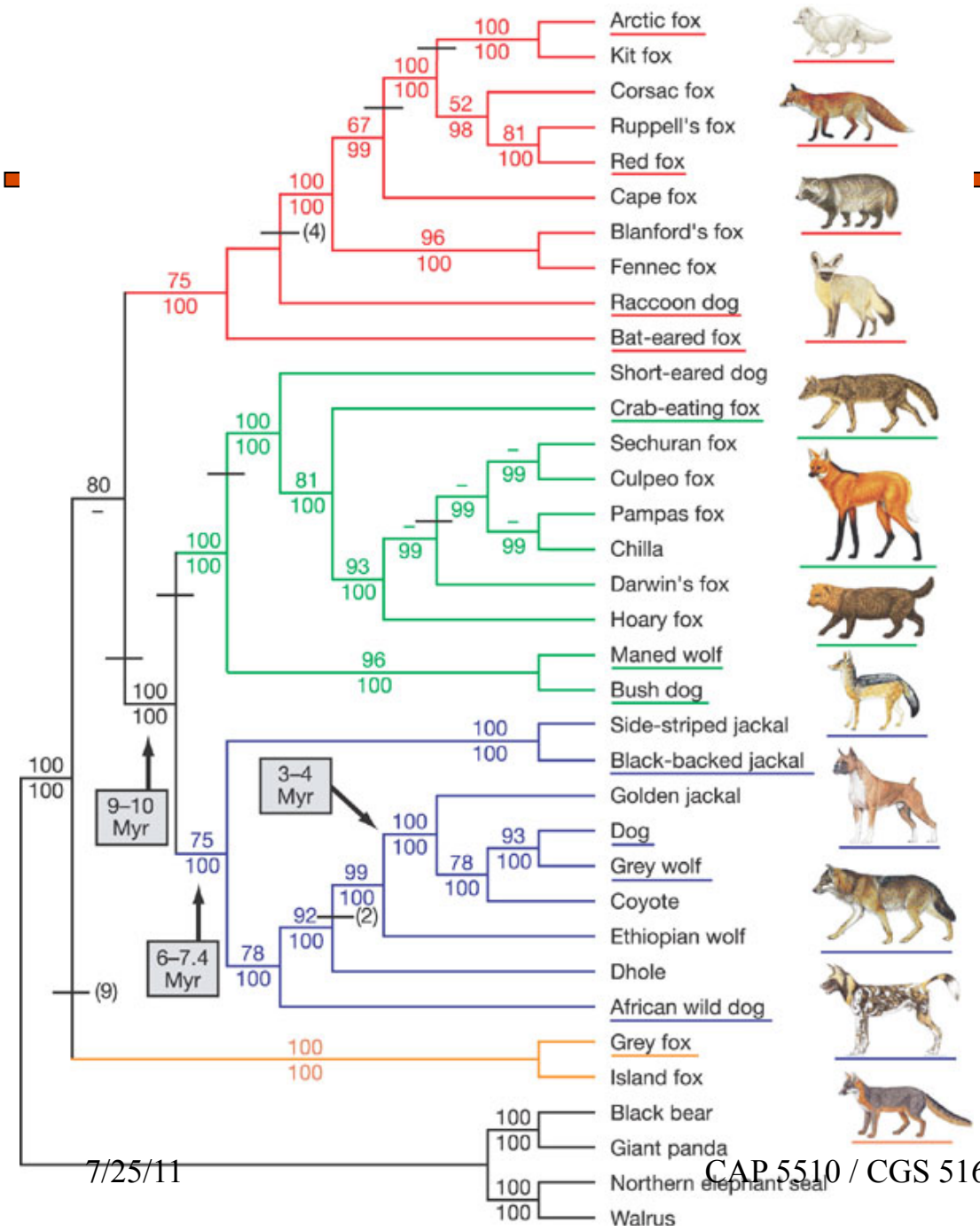


Fig. 7.8
Page 232

Examples of clades

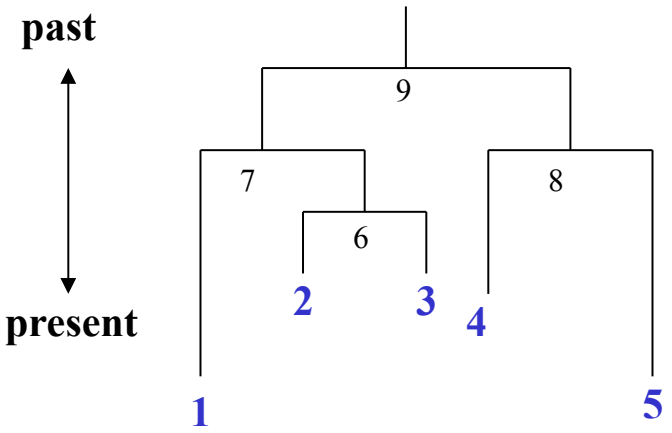
Slide by Pevsner

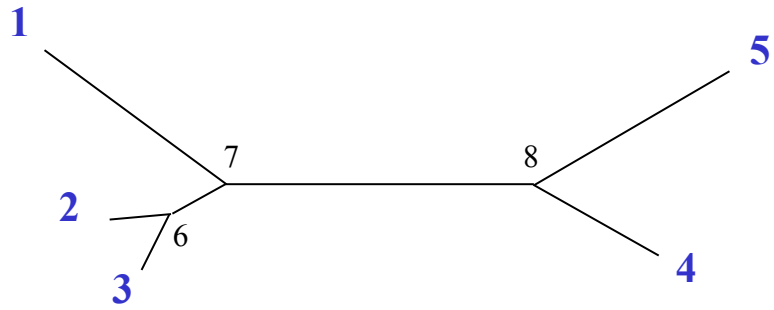


Lindblad-Toh et al., *Nature* 438: 803 (2005), fig. 10

Tree nomenclature: roots

Slide by Pevsner





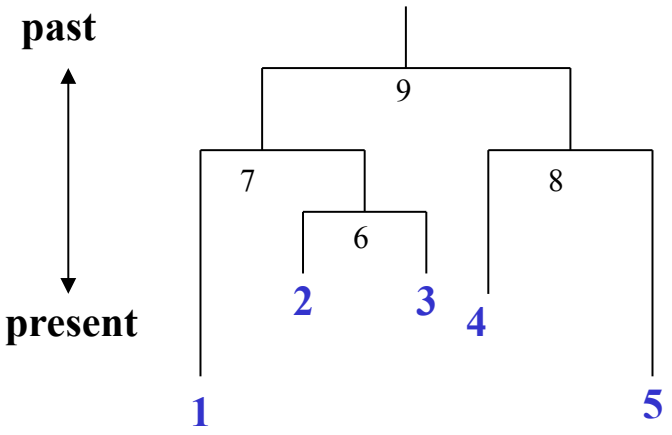
Rooted tree (specifies evolutionary path)

Unrooted tree

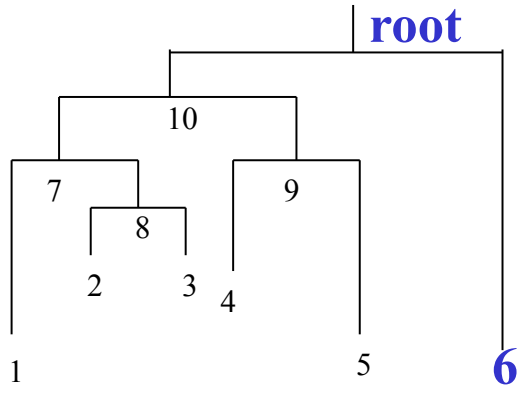
Fig. 7.10
Page 234

Tree nomenclature: outgroup rooting

Slide by Pevsner



Rooted tree



Outgroup
(used to place the root)

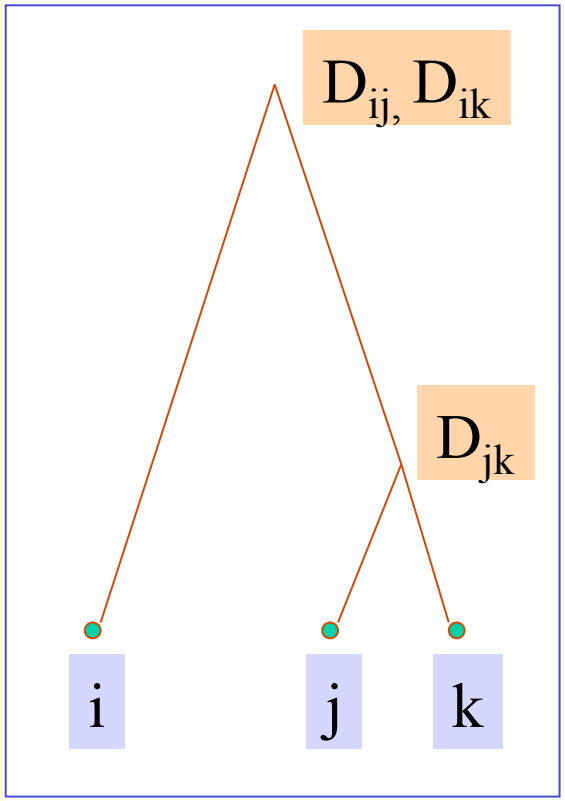
Fig. 7.10
Page 234

Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Ultrametric

- An ultrametric tree:
  - decreasing internal node labels
  - distance between two nodes is label of least common ancestor.
- An ultrametric distance matrix:
  - Symmetric matrix such that for every i, j, k, there is tie for maximum of D(i,j), D (j,k), D(i,k)



Ultrametric: Assumptions

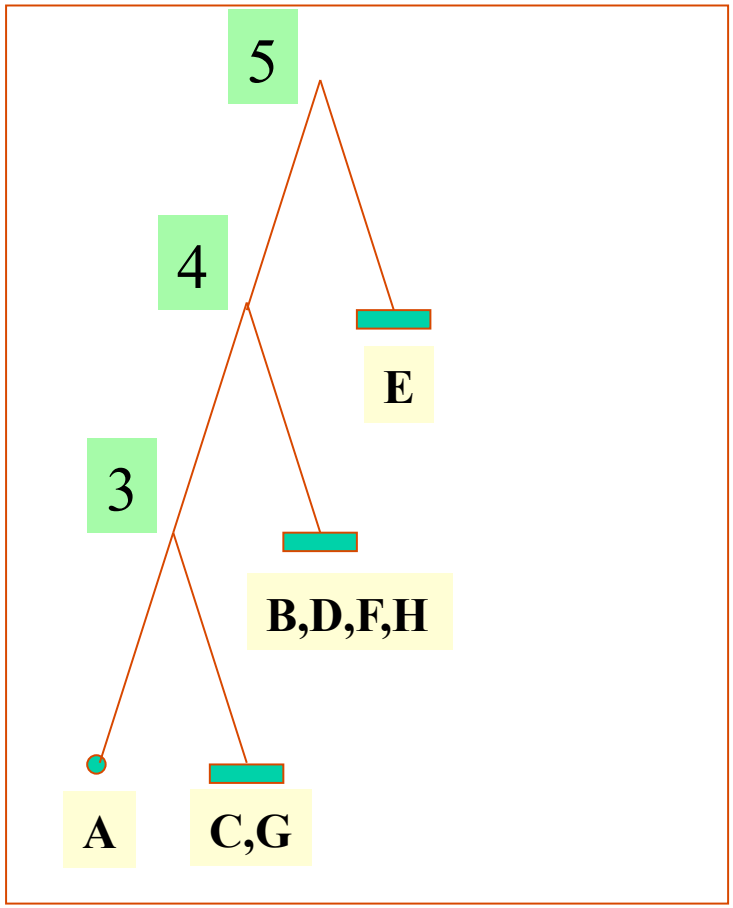
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

Ultrametric Data Sources

- Lab-based methods: hybridization
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: distance

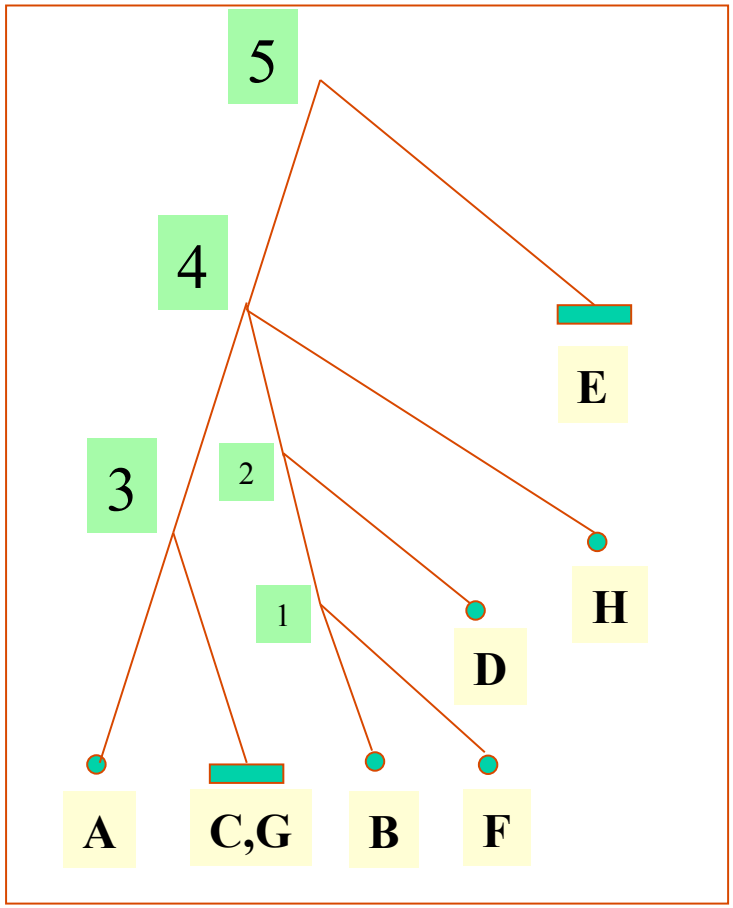
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B								
C								
D								
E								
F								
G								
H								



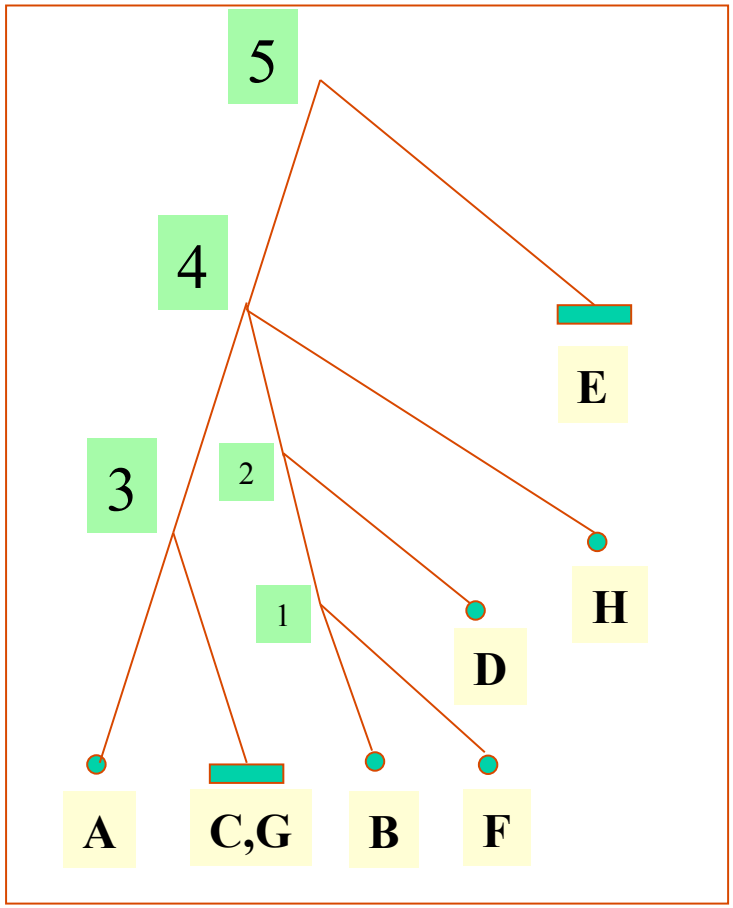
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C								
D								
E								
F								
G								
H								



Ultrametric: Distances Computed

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



Ultrametric: Assumptions

- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: **Accepted** point mutations in amino acid sequence of a protein occurs at a **constant** rate.
  - Varies from protein to protein
  - Varies from one part of a protein to another

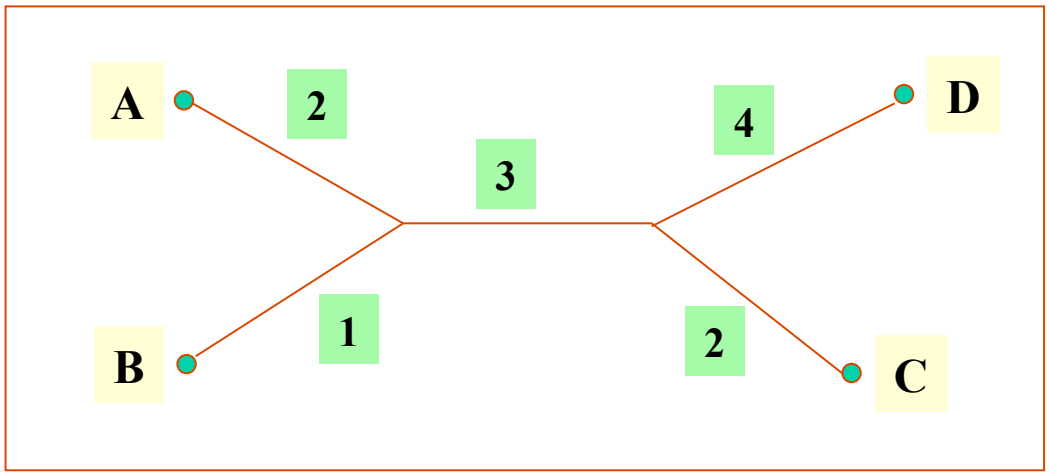
Ultrametric Data Sources

- Lab-based methods: hybridization
  - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: distance

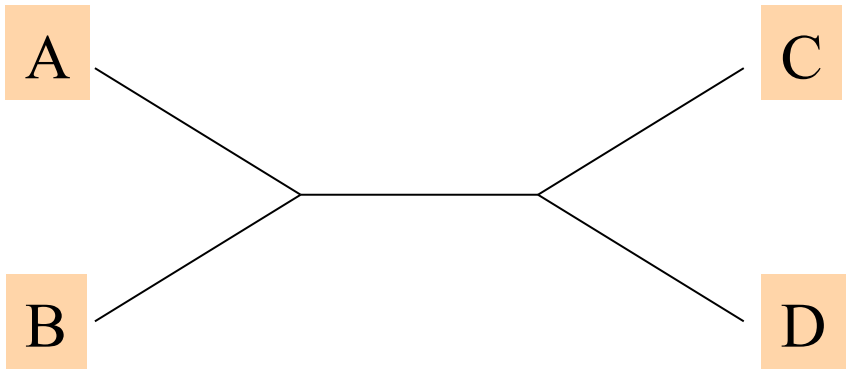
Additive-Distance Trees

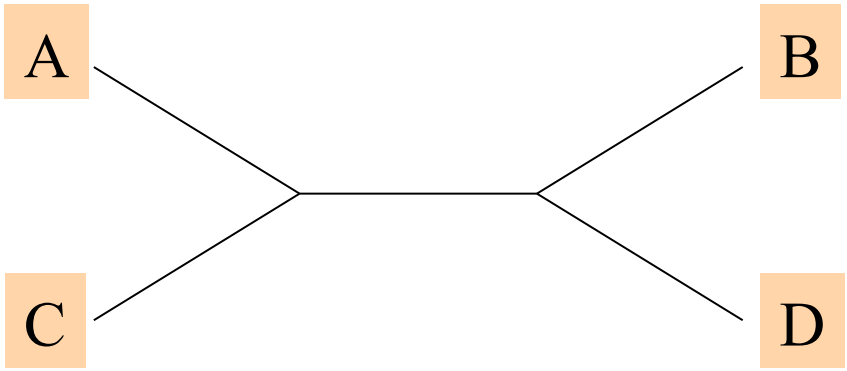
Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

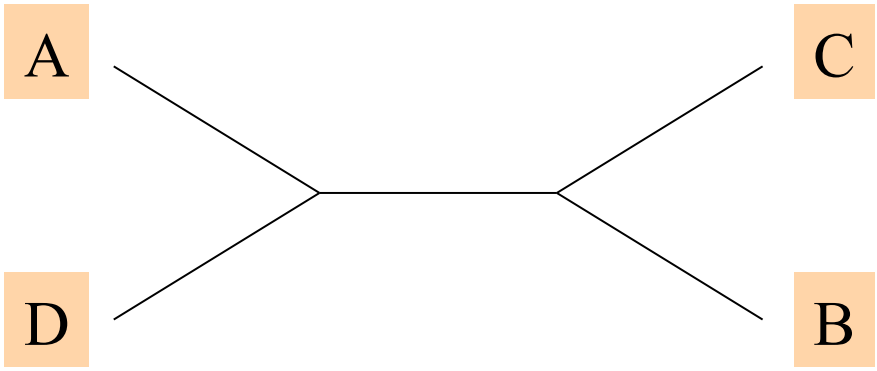
	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

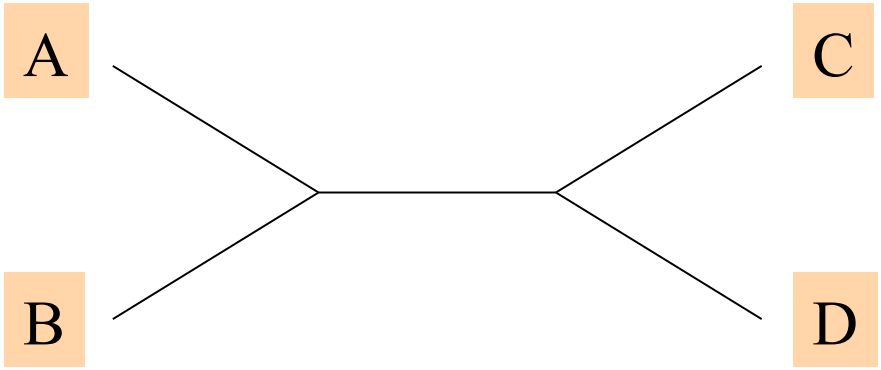






Four-Point Condition

- If the true tree is as shown below, then
  1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and
  2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

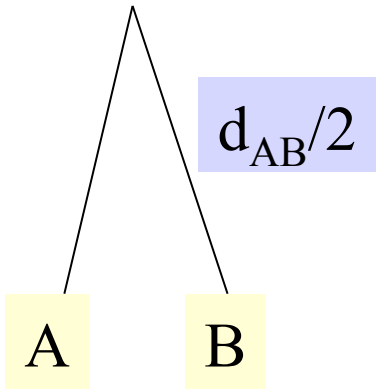


Unweighted pair-group method with arithmetic means (UPGMA)

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

	AB	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$$d_{(AB)C} = (d_{AC} + d_{BC}) /2$$



Transformed Distance Method

- UPGMA makes errors when rate constancy among lineages does not hold.
- Remedy: introduce an outgroup & make corrections
- Now apply UPGMA

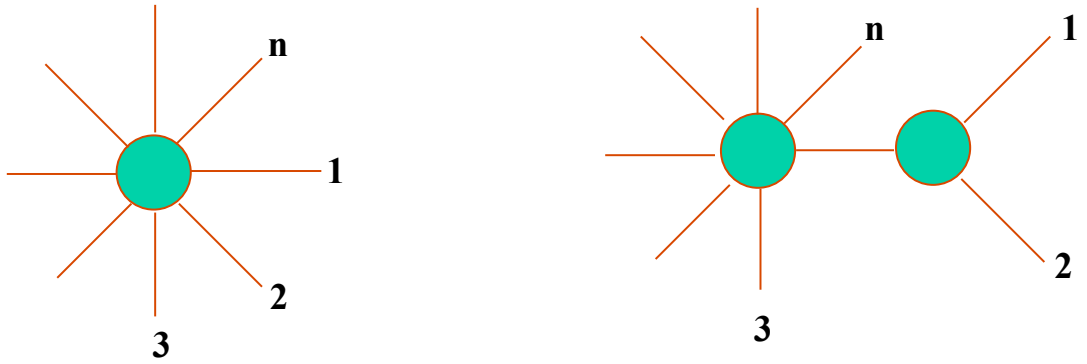
$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\sum_{k=1}^{n} D_{kO} \Big/ n \right)$$

Saitou & Nei: Neighbor-Joining Method

- Start with a star topology.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

Neighbor-Joining



$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \le i \le j \le n} D_{ij}$$

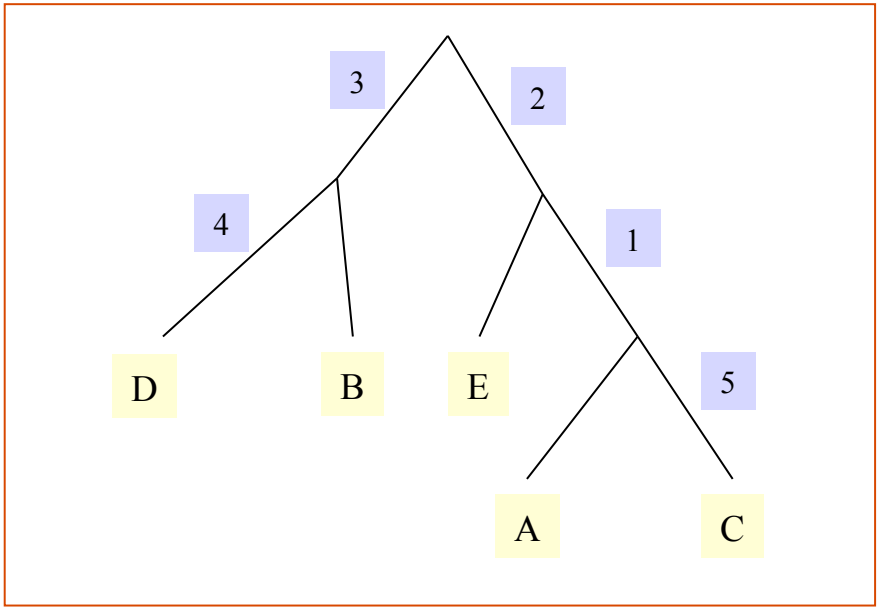
Constructing Evolutionary/Phylogenetic Trees

- 2 broad categories:
  - Distance-based methods
    - Ultrametric
    - Additive:
      - UPGMA
      - Transformed Distance
      - Neighbor-Joining
  - Character-based
    - Maximum Parsimony
    - Maximum Likelihood
    - Bayesian Methods

Character-based Methods

- Input: characters, morphological features, sequences, etc.
- Output: phylogenetic tree that provides the history of what features changed. [Perfect Phylogeny Problem]
- one leaf/object, 1 edge per character, path ⇔changed traits

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

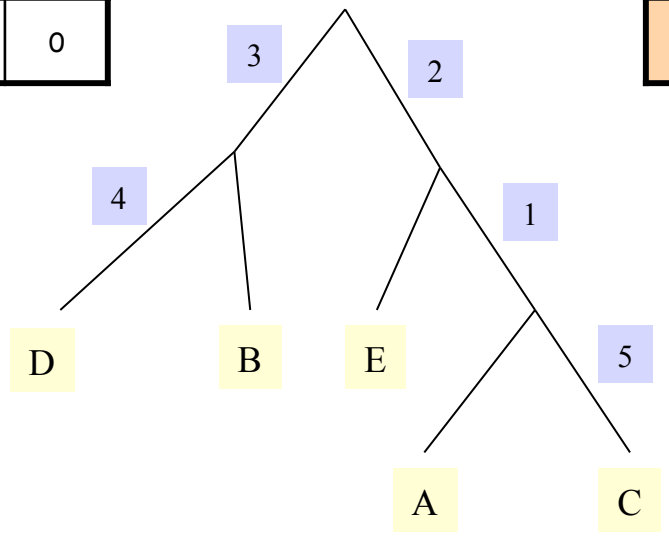


Example

- Perfect phylogeny does not always exist.

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1



Maximum Parsimony

- Minimize the total number of mutations implied by the evolutionary history

Examples of Character Data

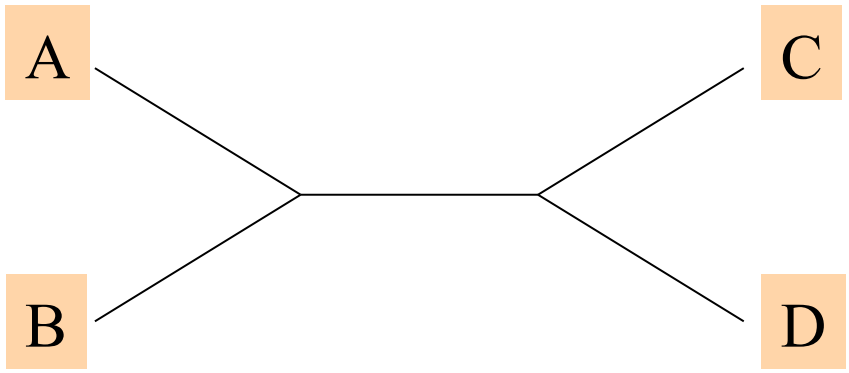
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

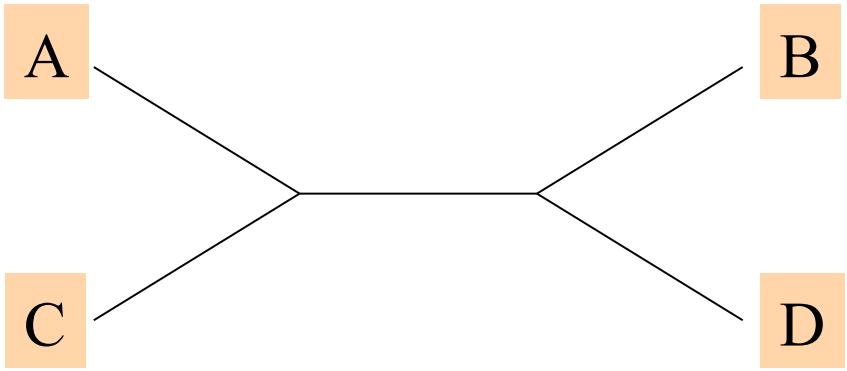
	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

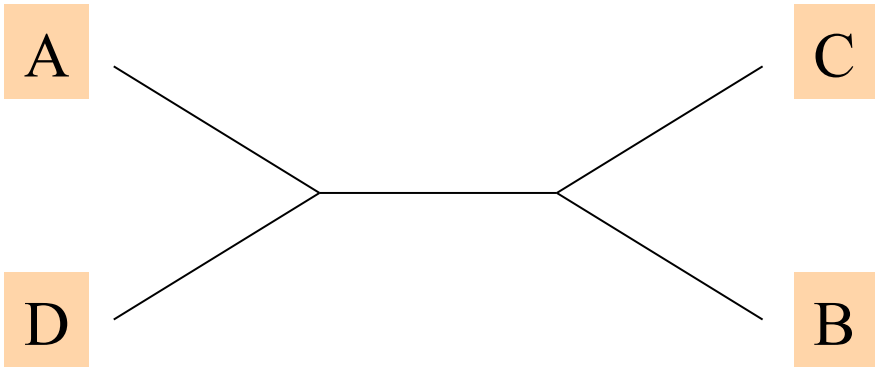
Maximum Parsimony Method: Example

	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Unrooted Trees on 4 Taxa







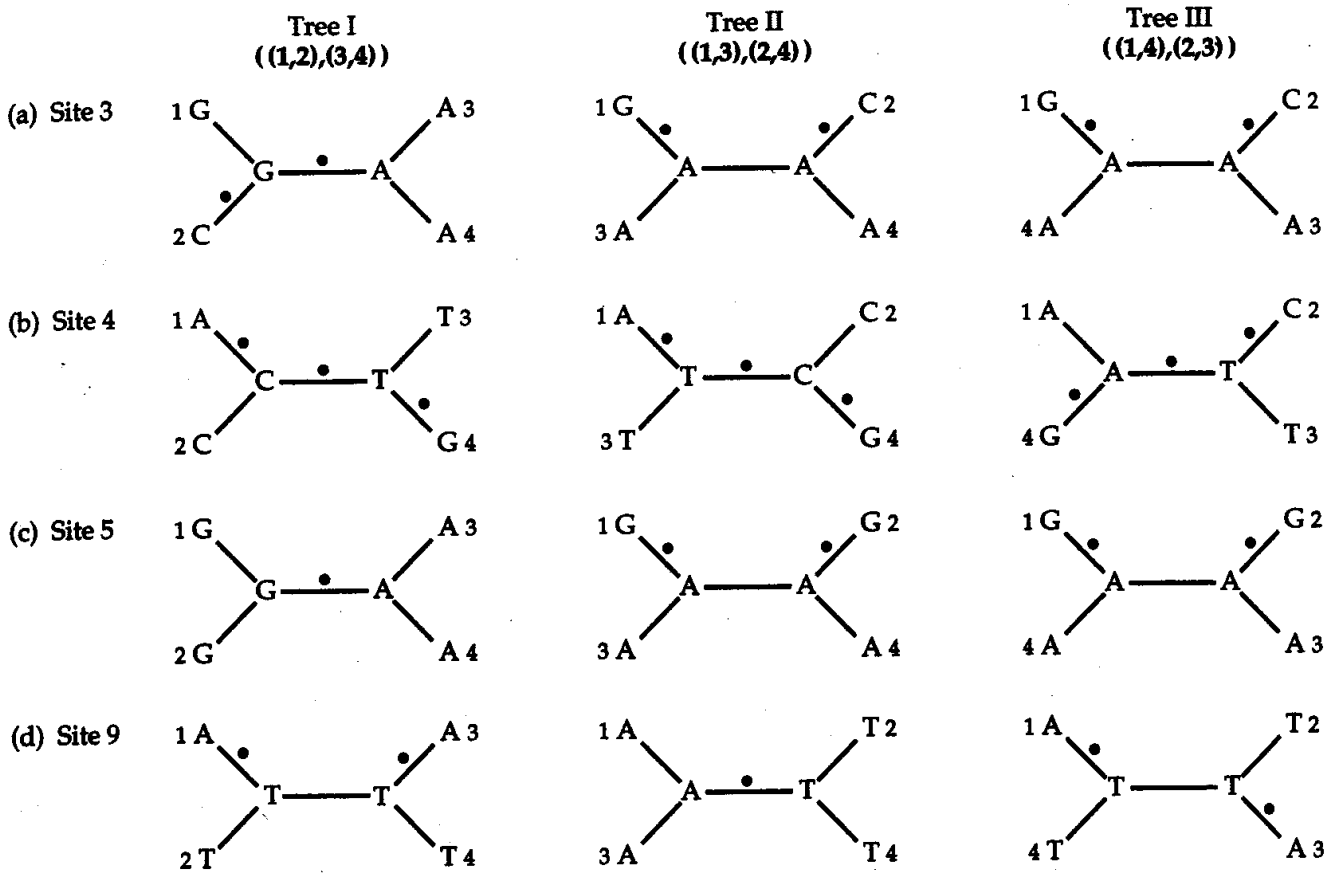


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III(d) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.

	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Inferring nucleotides on internal nodes

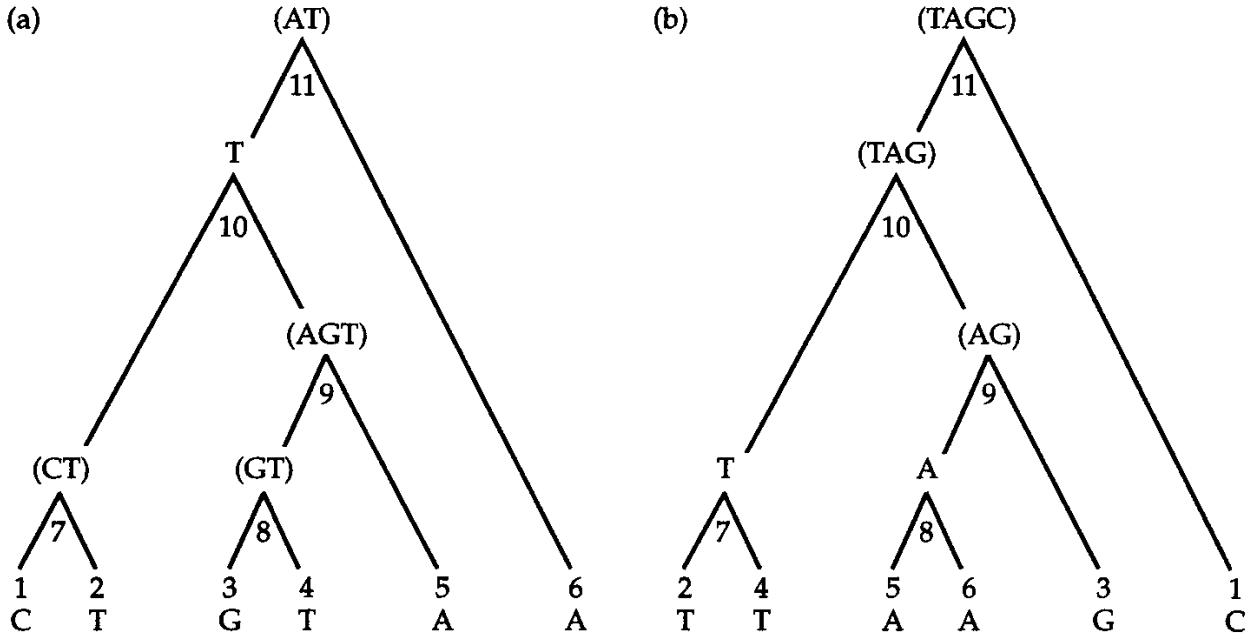


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).

Searching for the Maximum Parsimony Tree: Exhaustive Search

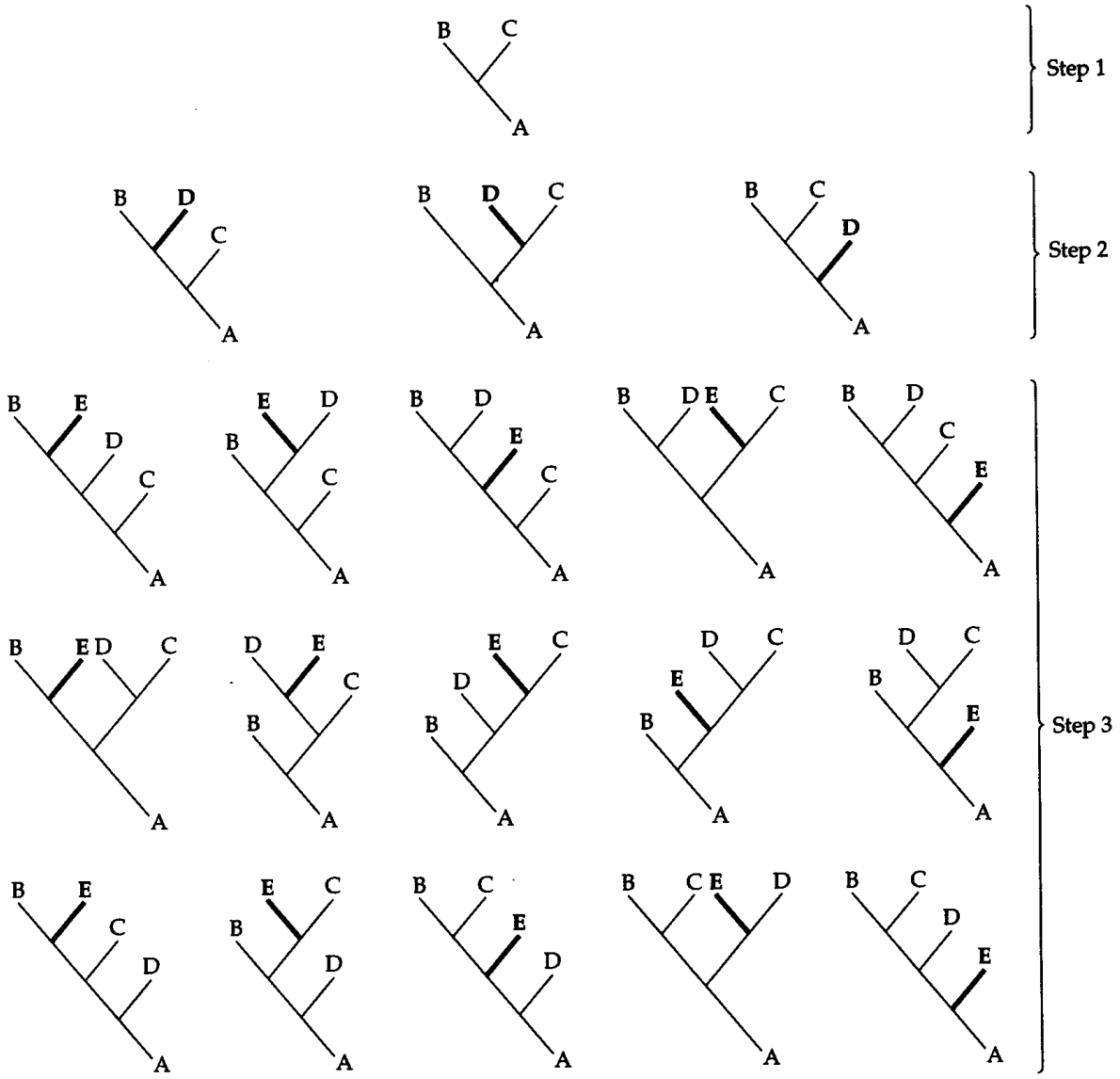
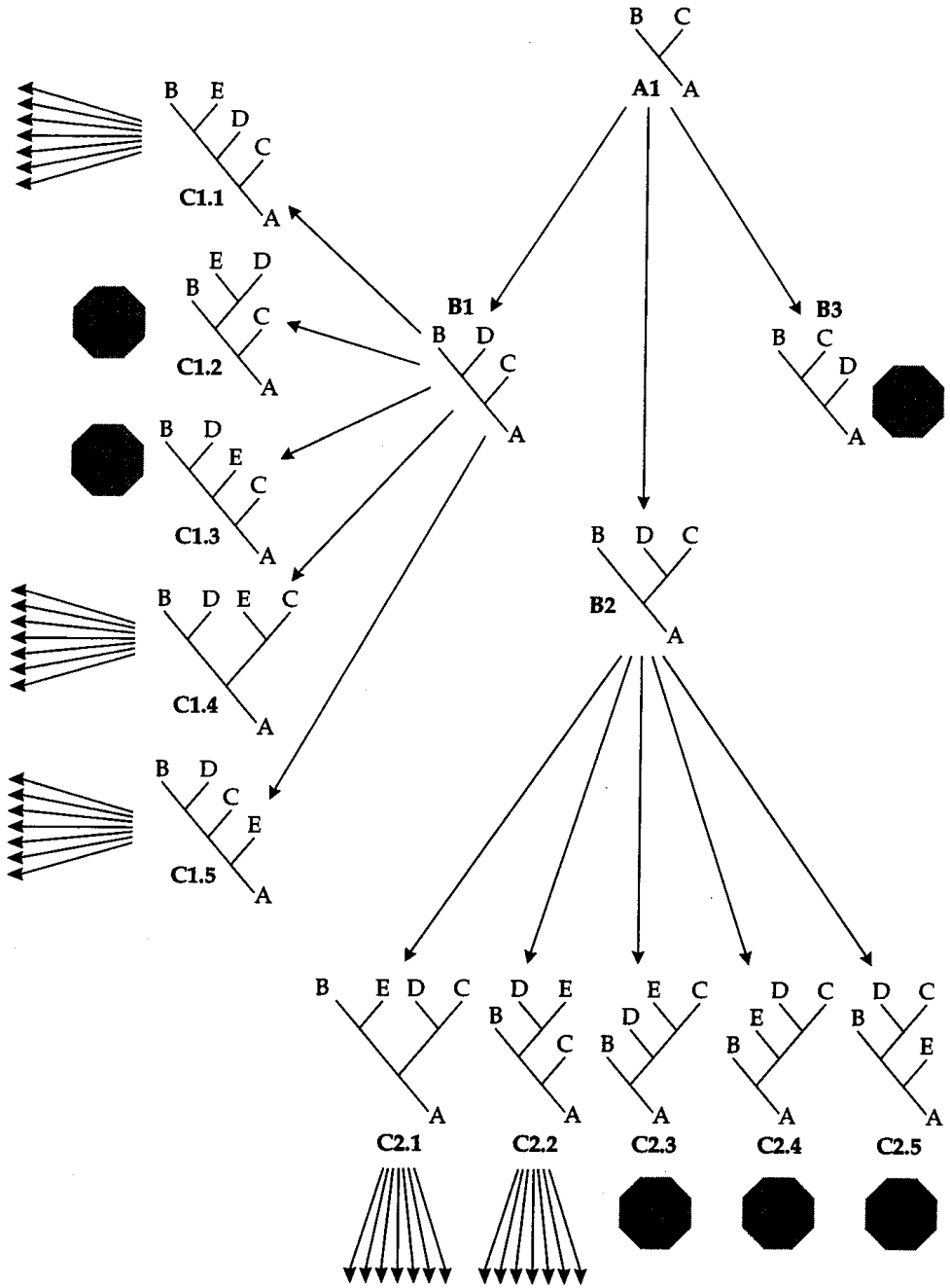


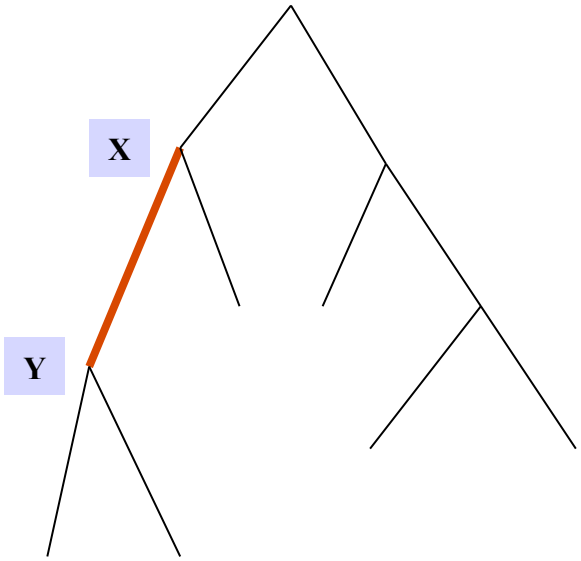
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four OTUs. In step 3, we add OTU E to each of the five branches of the three trees in step 2, thereby generating 15 unrooted trees. Additions of OTUs are shown as heavier lines. Modifed from Swofford et al. (1996).



Searching for the Maximum Parsimony Tree: Branch-&-Bound

Probabilistic Models of Evolution

- Assuming a model of substitution,
  - $Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$,
- Using this formula it is possible to compute the likelihood that data D is generated by a given phylogenetic tree T under a model of substitution. Now find the tree with the maximum likelihood.



- Time elapsed? Δ
- Prob of change along edge? $Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$
- Prob of data? Product of prob for all edges

(a)

	1	2	3	4	5	6	7	8	9	...*n*
OTU1	A	A	G	A	C	T	T	C	A	...N
OTU2	A	G	C	C	C	T	T	C	T	...N
OTU3	A	G	A	T	A	T	C	C	A	...N
OTU4	A	G	A	G	G	T	C	C	T	...N

(b)

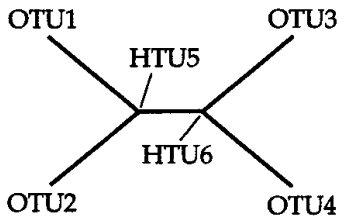
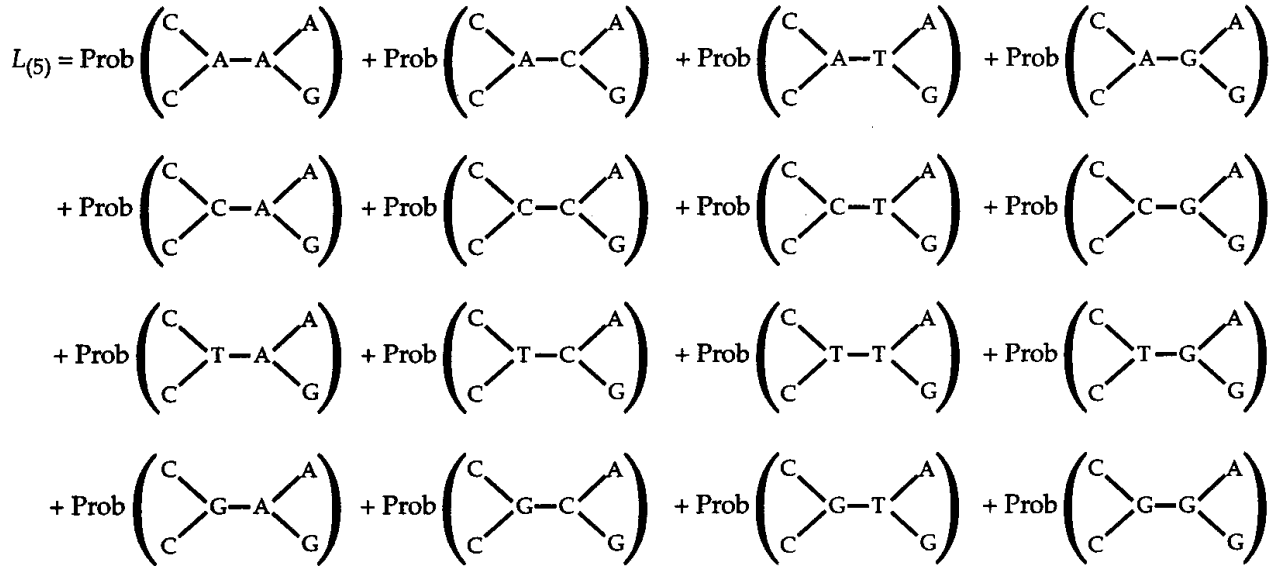


FIGURE 5.19 **Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length *n*. (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all *n* sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).**

(c)



(d) $$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times ... \times L_{(n)} = \prod_{i=1}^{n} L_{(i)}$$

(e) $$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + ... + L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)}$$

Computing Maximum Likelihood Tree

Basic Population Genetics

- Allele: one of two or more forms of DNA sequence of a particular gene
  - The word "allele" is a short form of allelomorph ('other form')
- Diploid: organisms with two sets of chromosomes
  - Homozygous alleles: if both copies of the allele are the same
  - Heterozygous alleles
- Alleles may be
  - Dominant: allele that is more often expressed in heterozygous individuals
  - Recessive
- Genotype: set of alleles in an individual, i.e., genetic composition

Genetic Characters

- Characters can be
  - Mendelian, i.e., single-gene effects, OR
  - Polygenic, i.e., caused by combined effect of multiple genetic factors, OR
  - Environmental
- Characters can be:
  - discrete (e.g., disease) or
  - continuous (e.g., height)
- Gene loci involved in continuous characters are called Quantitative Trait Loci

Hardy-Weinberg Principle

- G.H. Hardy & Wilhelm Weinberg (1908)
  - Allele and genotype frequencies in a population remain constant.

		Females	
		A (p)	a (q)
Males	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

  - Assumptions:
    - Diploid; sexual reproduction; non-overlapping generations
    - Biallelic loci; Allele frequencies independent of gender
    - Mating is random
    - Population size is infinite
    - Mutations can be ignored
    - Migration is negligible
    - Natural selection does not affect allele in question
    - Equilibrium attained in one generation

Genetic Linkage

- Meiosis: Cell division necessary for sexual reproduction
  - Produces gametes like sperm and egg cells.
- Meiosis: Starts with one diploid cell with 2 copies of each chromosome and produces four haploid cells, each with one copy of each chromosome. Each chromosome is recombined from the 2 copies.
  - At start of meiosis, chromosome pair recombine and exchange sections. Then they separate into two chromosomes.
  - Recombination: alleles on same chromosome may end up in different daughter cells
  - If two alleles are far apart, then there is a higher probability of a cross-over event between them putting them on different chromosomes.
  - Genetically linked traits are caused by alleles sufficiently close to each other. Used to produce genetic maps or linkage maps.

Linkage Disequilibrium (D)

- D = Difference between observed and expected allelic frequencies
- Given 2 bi-allelic loci A and B

AB	x_{11}
Ab	x_{12}
aB	x_{21}
ab	x_{22}

Allele	Frequency
A	$P_1 = x_{11} + x_{12}$
a	$P_2 = x_{21} + x_{22}$
B	$q_1 = x_{11} + x_{21}$
b	$q_2 = x_{12} + x_{22}$

- $D = x_{11} - p_1q_1$

	A	a	Total
B	$x_{11} = p_1q_1 + D$	$x_{11} = p_2q_1 - D$	q_1
b	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	P_1	P_2	1

Linkage Disequilibrium

- Linkage (dis)equilibrium: when genotype at loci are (not) independent
- Assumptions of basic population genetics
  - Transmission of alleles (across generations) at two loci are independent
  - Fitness of genotypes at different loci are independent
- Both assumptions are not true in general
- There exists non-random associations of alleles at different loci
- The extent of these associations are measured by Linkage Disequilibrium

SNPs

- SNP: single nucleotide polymorphism
  - Mutations in single nucleotide position
  - Occurred once in human history
  - Passed on through heredity
  - ~10M SNPs in human genome
  - 1 SNP every 300 bp, most with a frequency of 10-50%
- Most variations within a population characterized by SNPs
- Want to correlate SNPs to human disease
- Genotype
  - Gives bases at each SNP for both copies of chromosome, but loses information as to the chromosome on which it appears. NO LABEL!
- Haplotype
  - Gives bases at each SNP for each chromosome. LABELED!

Genotype vs Haplotype

- If the first locus is bi-allelic with two possible alleles (say, A & G)
  - Genotypes: AA, GG, AG
- If a second bi-allelic locus has alleles G & C
  - Genotypes: GG, CC, GC
- Genotypes & Haplotypes for the two loci are:

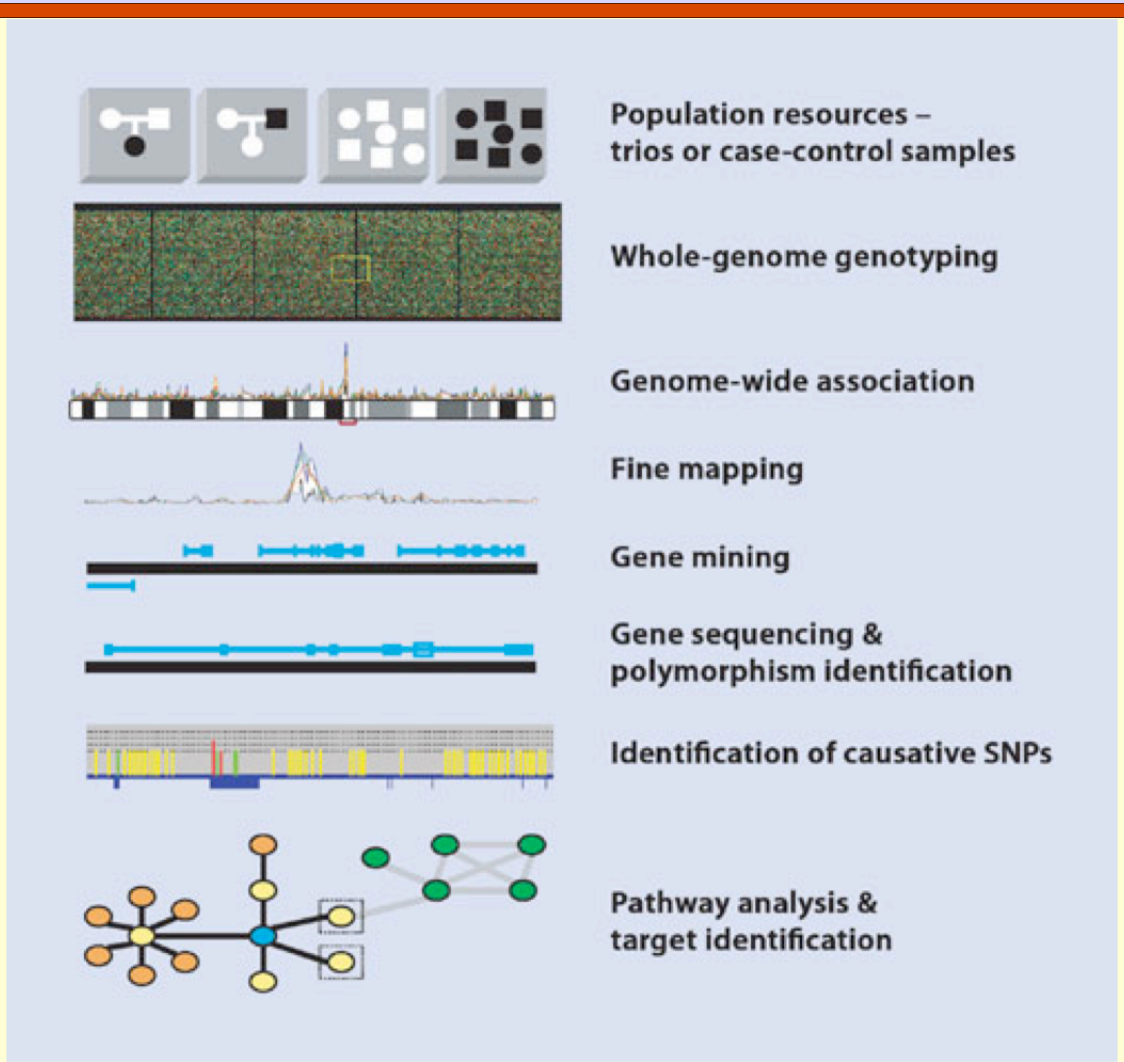
Haplotypes		Second Locus		
		GG	GC	CC
First Locus	AA	AG AG	AG AC	AC AC
	AG	AG GG	AG GC or AC GG	AC GC
	GG	GG GG	GG GC	GC GC

- Interesting problem:
  - Given genotypes, resolve the haplotypes

Genome-wide Association Studies (GWAS)

- To identify patterns of polymorphisms that vary systematically between individuals with different disease states
  - To identify risk-enhancing or risk-decreasing alleles
- Examples of GWAS (900 studies; 3500 associations)
  - Prostate Cancer: Nature Genetics, 1 Apr 2007
  - Type 2 Diabetes: Science Express, 26 Apr 2007
  - Heart Diseases: Science Express, 3 May 2007
  - Breast Cancer, Nature & Nature Genetics, 27 May 2007
  - …
  - See: http://www.genome.gov/Pages/About/OD/ReportsPublications/GWASUpdateSlides-9-19-07.pdf
- Since variation is inherited in blocks / groups, it is enough to study a sample of the population, instead of looking at the whole population.
- GWA databases at NIH: dbGaP, caBIG, and CGEMS

GWAS Process



Analysis

- Summary statistics for quality control
  - Allele, genotypes frequencies, missing genotype rates, inbreeding stats, non-Mendelian transmission in family data, Sex checks based on X chromosome SNPs
- Population stratification detection
  - Complete linkage hierarchical clustering
  - Multidimensional scaling analysis to visualise substructure
  - Significance test for whether two individuals belong to the same population
- Association Testing:
  - Case vs Control
    - Standard allelic test, Fisher's exact test, Cochran-Armitage trend test, Mantel-Haenszel and Breslow-Day tests for stratified samples, Dominant/recessive and general models, Model comparison tests
  - Family-based associations
  - QTLs
- ...

Software

- PLINK: for analysis of genotype, phenotype data
- EIGENSOFT: for population structure analysis
- IMPUTE, SNPTEST, MACH, ProbABEL, BimBam, QUICKTEST