

Molecular Phylogeny

Part 1 of 2

October 16, 2006

Introduction to Bioinformatics
J. Pevsner
pevsner@kennedykrieger.org

Copyright notice

Many of the images in this powerpoint presentation are from *Bioinformatics and Functional Genomics* by J Pevsner (ISBN 0-471-21004-8).
Copyright © 2003 by Wiley.

These images and materials may not be used without permission from the publisher.

Visit <http://www.bioinfbook.org>

Goal of the lectures today and Monday

Introduction to evolution and phylogeny

Nomenclature of trees

Four stages of molecular phylogeny:

[1] selecting sequences

[2] multiple sequence alignment

[3] tree-building

[4] tree evaluation

Practical approaches to making trees

Introduction

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.

To Darwin, the 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.

Introduction

Darwin did not understand the mechanisms by which hereditary changes occur. In the 1920s and 1930s, a synthesis occurred between Darwinism and Mendel's principles of inheritance.

The basic processes of evolution are

- [1] mutation, and also
- [2] genetic recombination as two sources of variability;
- [3] chromosomal organization (and its variation);
- [4] natural selection
- [5] reproductive isolation, which constrains the effects of selection on populations

(See Stebbins, 1966)

Introduction

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.

Historical background

Studies of molecular evolution began with the first sequencing of proteins, beginning in the 1950s.

In 1953 Frederick Sanger and colleagues determined the primary amino acid sequence of insulin.

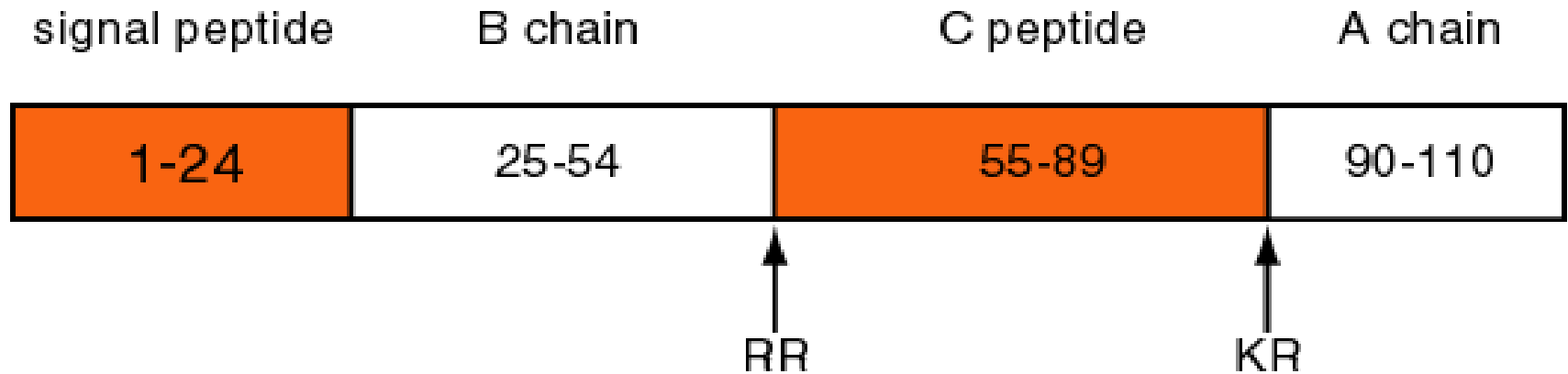
(The accession number of human insulin is NP_000198)

Sanger and colleagues sequenced insulin (1950s)

Human	CGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCT <u>S</u> ICSLYQLEN
chimpanzee	CGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCT <u>S</u> ICSLYQLEN
rabbit	CGERGFFYTPKSRREVEELQVGQAEELGGGPGAGGLQPSALELALQKRGIVEQCCT <u>S</u> ICSLYQLEN
dog	CGERGFFYTPKARREVEDLQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCT <u>S</u> ICSLYQLEN
horse	CGERGFFYTPKAXXEAEDPQVGEVELGGGPGGLGGLQPLALAGPQQXXGIVEQCCTG <u>I</u> C SLYQLEN
mouse	CGERGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVAQQKRGIVDQCCT <u>S</u> ICSLYQLEN
rat	CGERGFFYTPMSRREVEDPQVAQLELGGGPGAGDLQTLALEVARQKRGIVDQCCT <u>S</u> ICSLYQLEN
pig	CGERGFFYTPKARREAENPQAGAVELGG--GLGGLQALALEGPPQKRGIVEQCCT <u>S</u> ICSLYQLEN
chicken	CGERGFFYSPKARRDVEQPLVSSPLRG---EAGVLPFQQEYKVKRGIVEQCCH <u>N</u> TCSLYQLEN
sheep	CGERGFFYTPKARREVEGPQVGALELAGGPGAG-----GLEGPPQKRGIVEQCCAG <u>V</u> C SLYQLEN
bovine	CGERGFFYTPKARREVEGPQVGALELAGGPGAG-----GLEGPPQKRGIVEQCCAS <u>V</u> C SLYQLEN
whale	CGERGFFYTPKA-----GIVEQCCT <u>S</u> ICSLYQLEN
elephant	CGERGFFYTPKT-----GIVEQCCTG <u>V</u> C SLYQLEN

We can make a multiple sequence alignment of insulins from various species, and see conserved regions...

Mature insulin consists of an A chain and B chain heterodimer connected by disulphide bridges



The signal peptide and C peptide are cleaved, and their sequences display fewer functional constraints.

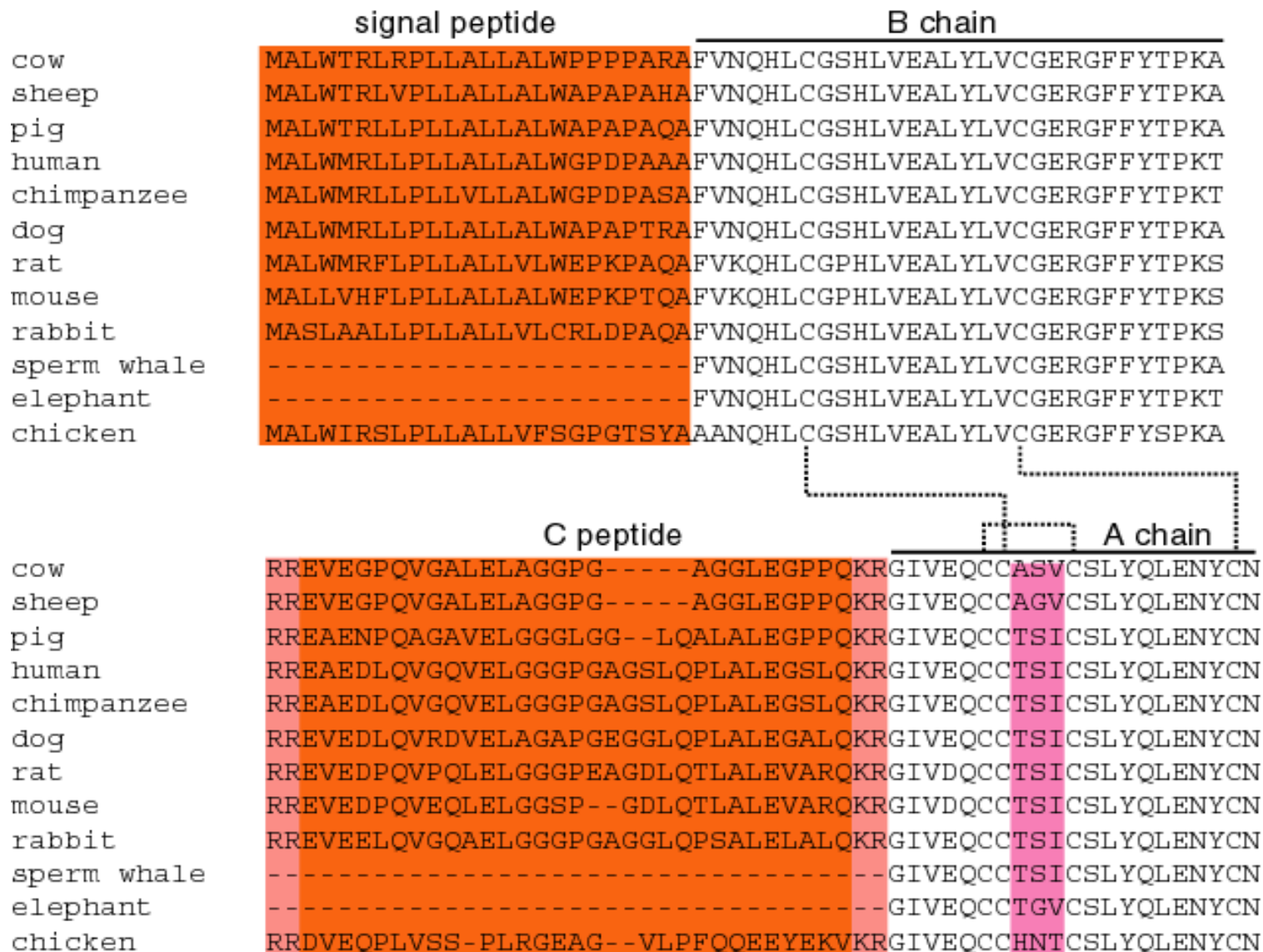
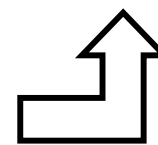


Fig. 11.1
Page 359

	signal peptide	B chain		C peptide	A chain		
cow	MALWTRLRPLLALLALWPPPPARA	FVNQHLCGSHLVEALYLVCGERGFFYTPKA		RREVEGPQVGALELAGGPG----	AGGLEGPPQKR	GIVEQCCASVCSLYQLENYCN	
sheep	MALWTRLVPLLALLALWAPAPAHAF	FVNQHLCGSHLVEALYLVCGERGFFYTPKA		RREVEGPQVGALELAGGPG----	AGGLEGPPQKR	GIVEQCCAGVCSLYQLENYCN	
pig	MALWTRLRPLLALLALWAPAPAQA	FVNQHLCGSHLVEALYLVCGERGFFYTPKA		RREAENPQAGAVELGGGLGG--	LQALALEGPPQKR	GIVEQCCTSI	CSLYQLENYCN
human	MALWMRLRPLLALLALWGPDPAAAF	FVNQHLCGSHLVEALYLVCGERGFFYTPKT		RREAEDLQVGQVELGGGPGAGSL	QPLALEGSLQKR	GIVEQCCTSI	CSLYQLENYCN
chimpanzee	MALWMRLRPLLALLALWGPDPASAF	FVNQHLCGSHLVEALYLVCGERGFFYTPKT		RREAEDLQVGQVELGGGPGAGSL	QPLALEGSLQKR	GIVEQCCTSI	CSLYQLENYCN
dog	MALWMRLRPLLALLALWAPAPTRA	FVNQHLCGSHLVEALYLVCGERGFFYTPKA		RREVEDLQVRDVELAGAPGEGGL	QPLALEGALQKR	GIVEQCCTSI	CSLYQLENYCN
rat	MALWMRFLPLLALLVLWEPKPAQA	FVKQHLCGPHLVEALYLVCGERGFFYTPKS		RREVEDPQVPQLELGGGPEAGDL	QTLALEVARQKR	GIVDQCCTSI	CSLYQLENYCN
mouse	MALLVHFLPLLALLALWEPKPTQA	FVKQHLCGPHLVEALYLVCGERGFFYTPKS		RREVEDPQVEQLELGGSP--	GDLQTLALEVARQKR	GIVDQCCTSI	CSLYQLENYCN
rabbit	MASLAALLPLLALLVLCRLDPAQA	FVNQHLCGSHLVEALYLVCGERGFFYTPKS		RREVEELQVGQAELGGGPGAGGL	QPSALELALQKR	GIVEQCCTSI	CSLYQLENYCN
sperm whale	-----	FVNQHLCGSHLVEALYLVCGERGFFYTPKA		-----	-----	GIVEQCCTSI	CSLYQLENYCN
elephant	-----	FVNQHLCGSHLVEALYLVCGERGFFYTPKT		-----	-----	GIVEQCCTGV	CSLYQLENYCN
chicken	MALWIRSLPLLALLVFSGPGTSYAA	ANQHLCGSHLVEALYLVCGERGFFYSPKA		RRDVEQPLVSS-PLRGEAG--	VLPFQOEEYEKVKR	GIVEQCCHNT	CSLYQLENYCN



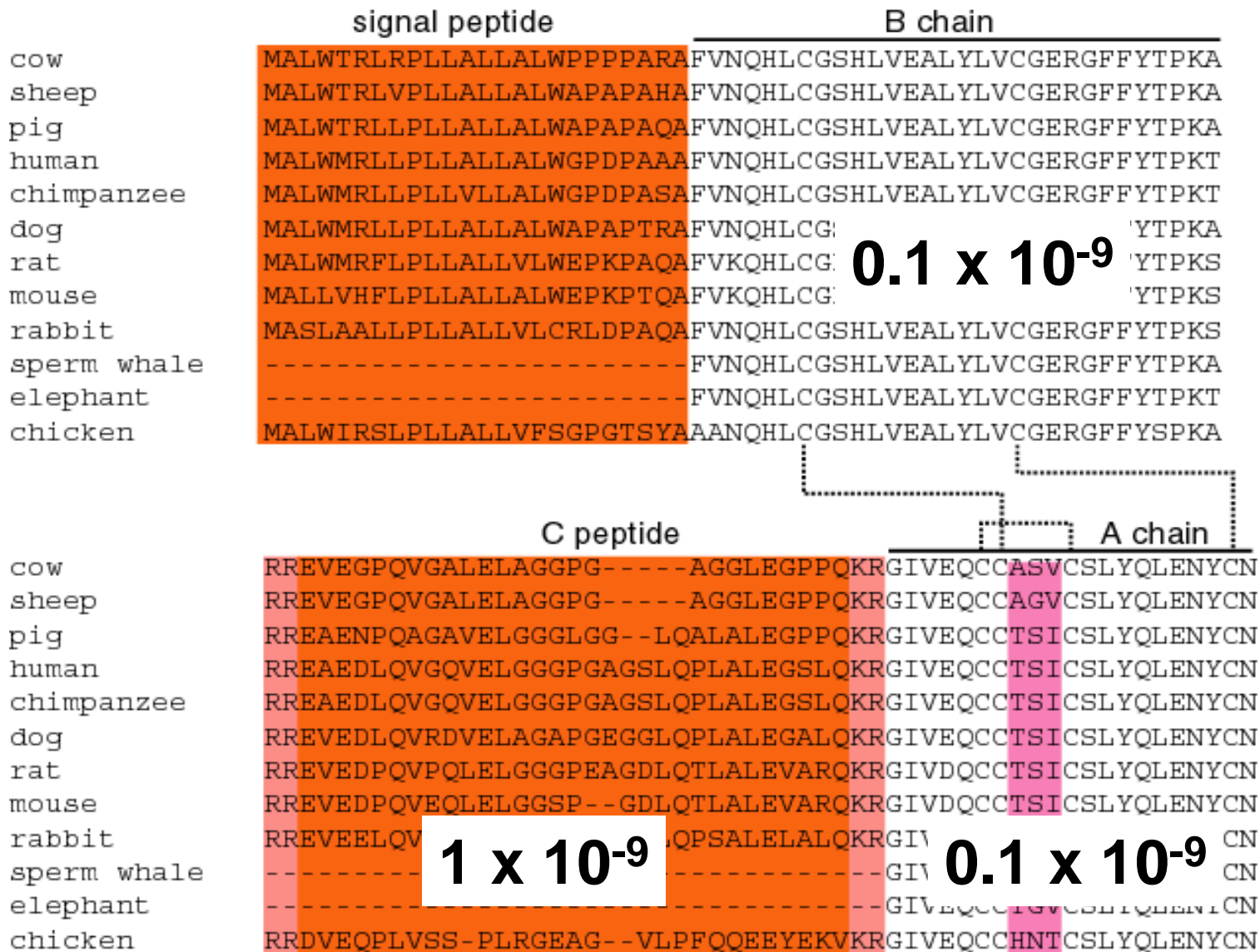
Note the sequence divergence in the disulfide loop region of the A chain



Historical background: insulin

By the 1950s, it became clear that amino acid substitutions occur nonrandomly. For example, Sanger and colleagues noted that most amino acid changes in the insulin A chain are restricted to a disulfide loop region. Such differences are called “neutral” changes (Kimura, 1968; Jukes and Cantor, 1969).

Subsequent studies at the DNA level showed that rate of nucleotide (and of amino acid) substitution is about six- to ten-fold higher in the C peptide, relative to the A and B chains.



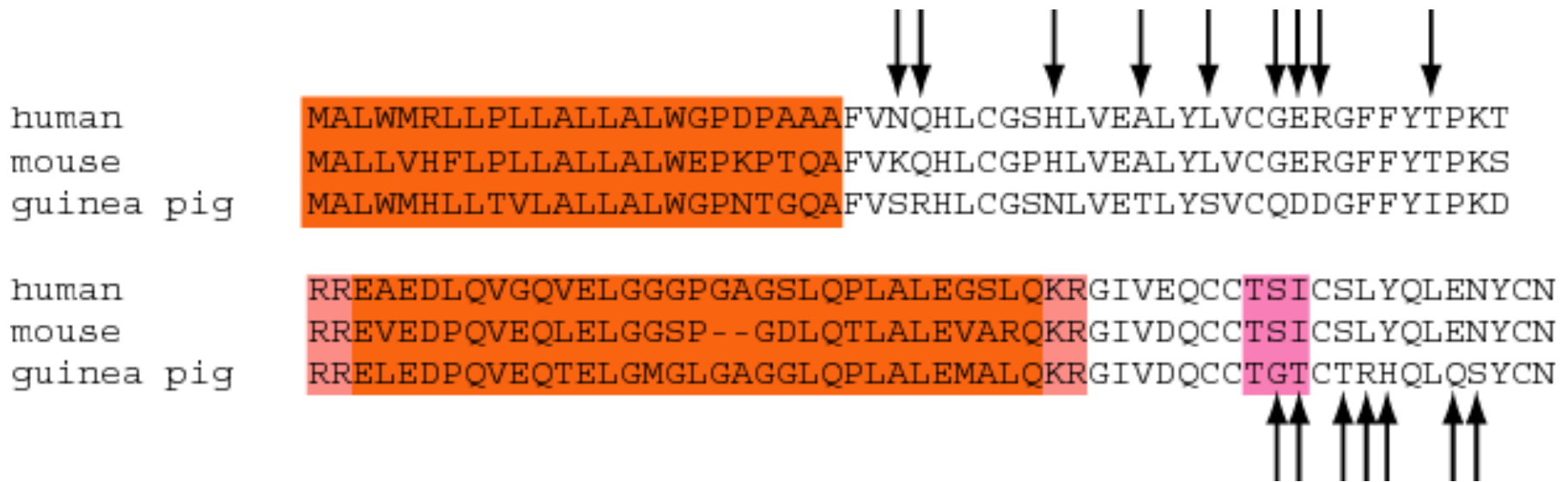
Number of nucleotide substitutions/site/year

Historical background: insulin

Surprisingly, insulin from the guinea pig (and from the related coypu) evolve seven times faster than insulin from other species. Why?

The answer is that guinea pig and coypu insulin do not bind two zinc ions, while insulin molecules from most other species do. There was a relaxation on the structural constraints of these molecules, and so the genes diverged rapidly.

Guinea pig and coypu insulin have undergone an extremely rapid rate of evolutionary change



Arrows indicate positions at which guinea pig insulin (A chain and B chain) differs from both human and mouse

Historical background

Oxytocin	CYIQNCPLG
Vasopressin	CYFQNCPRG

In the 1950s, other labs sequenced oxytocin and vasopressin. These peptides differ at only two amino acid residues, but they have distinctly different functions. It became clear that there are significant structural and functional consequences to changes in primary amino acid sequence.

Molecular clock hypothesis

In the 1960s, sequence data were accumulated for small, abundant proteins such as globins, cytochromes *c*, and fibrinopeptides. Some proteins appeared to evolve slowly, while others evolved rapidly.

Linus Pauling, Emanuel Margoliash and others proposed the hypothesis of a molecular clock:

For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages

Molecular clock hypothesis

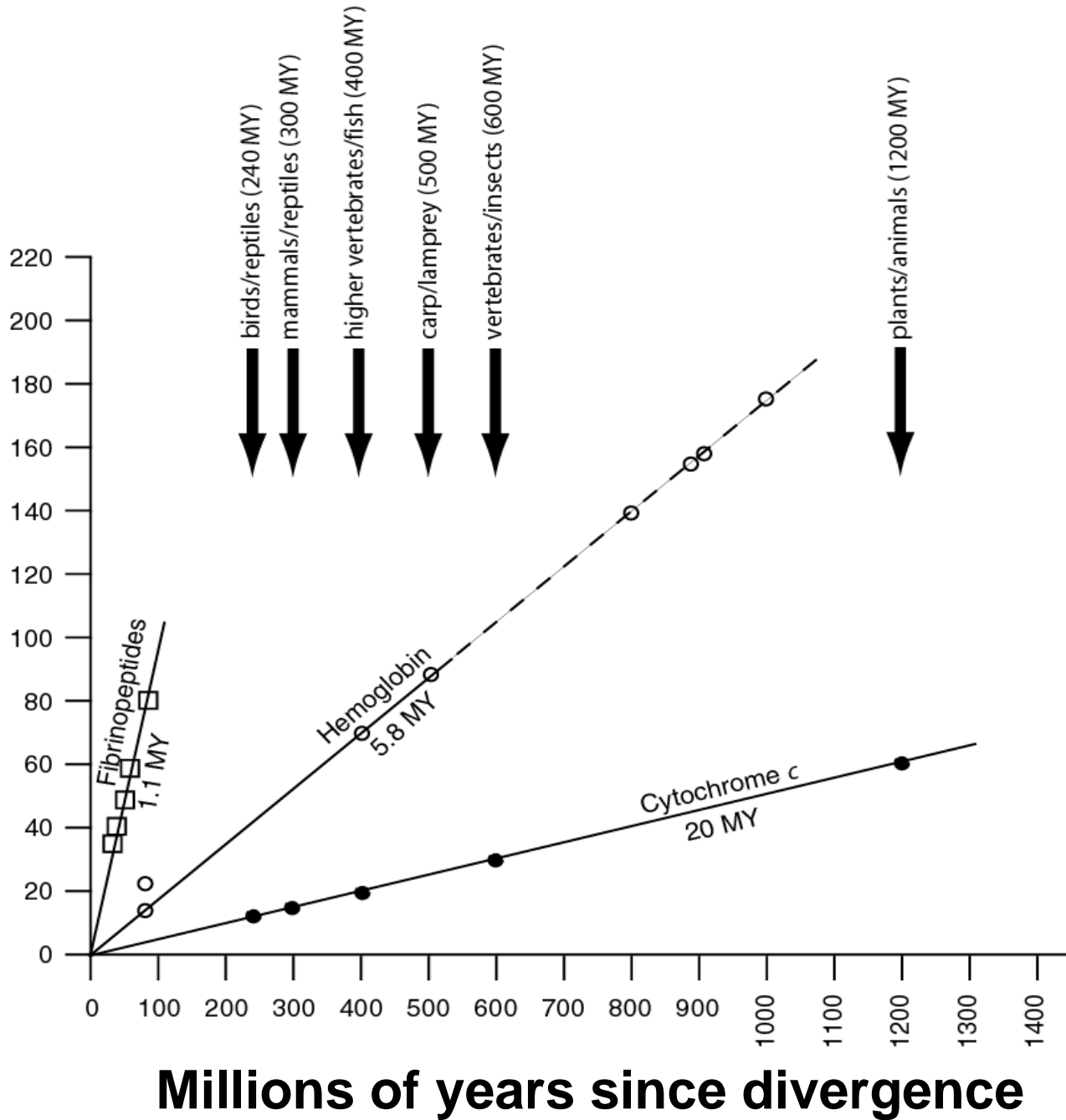
As an example, Richard Dickerson (1971) plotted data from three protein families: cytochrome *c*, hemoglobin, and fibrinopeptides.

The x-axis shows the divergence times of the species, estimated from paleontological data. The y-axis shows m , the corrected number of amino acid changes per 100 residues.

n is the observed number of amino acid changes per 100 residues, and it is corrected to m to account for changes that occur but are not observed.

$$\frac{N}{100} = 1 - e^{-(m/100)}$$

**corrected amino acid changes
per 100 residues (m)**



**Dickerson
(1971)**

Fig. 11.3
Page 361

Molecular clock hypothesis: conclusions

Dickerson drew the following conclusions:

- For each protein, the data lie on a straight line. Thus, the rate of amino acid substitution has remained constant for each protein.
- The average rate of change differs for each protein. The time for a 1% change to occur between two lines of evolution is 20 MY (cytochrome c), 5.8 MY (hemoglobin), and 1.1 MY (fibrinopeptides).
- The observed variations in rate of change reflect functional constraints imposed by natural selection.

Molecular clock hypothesis: λ and PAM

The rate of amino acid substitution is measured by λ , the number of substitutions per amino acid site per year.

Consider serum albumin:

$$\lambda = 1.9 \times 10^{-9}$$

$$\lambda \times 10^9 = 1.9$$

Dayhoff et al. (Box 3.3, page 50) reported the rate of mutation acceptance for serum albumin as 19 PAMs per amino acid residue per 100 million years.

(19 subst./1 aa/10⁸ years = 1.9 subst./100 aa/10⁹ years)

Molecular clock for proteins: rate of substitutions per aa site per 10^9 years

Fibrinopeptides	9.0
Kappa casein	3.3
Lactalbumin	2.7
Serum albumin	1.9
Lysozyme	0.98
Trypsin	0.59
Insulin	0.44
Cytochrome c	0.22
Histone H2B	0.09
Ubiquitin	0.010
Histone H4	0.010

Partial alignment of histones from PFAM ($\lambda = 0.05$)

H2A1_HUMAN/4-119	R.KGNYAERV	GAGAPVYLAA	VLEYLTAEIL	ELAGNAARDN	KKTRIIPR
H2A1_YEAST/3-120	R.RGNYAQRI	GSGAPVYLTA	VLEYLAAEIL	ELAGNAARDN	KKTRIIPR
H2A3_VOLCA/5-119	K.KGKYAERI	GAGAPVYLAA	VLEYLTAEVL	ELAGNAARDN	KKNRIVPR
H2A_PLAFA/5-120	K.KGKYAKRV	GAGAPVYLAA	VLEYLCAEIL	ELAGNAARDN	KKSRTIPR
H2A1_PEA/11-128	K.KGRYAQRV	GTGAPVYLAA	VLEYLAAEVL	ELAGNAARDN	KKNRISPR
H2A1_TETPY/7-123	K.HGRYSERI	GTGAPVYLAA	VLEYLAAEVL	ELAGNAAKDN	KKTRIVPR
H2AM_RAT/4-116	K.KGHPKYRI	GVGAPVYMAA	VLEYLTAEIL	ELAGNAARDN	KKGRVTPR
H2A_EUGGR/18-134	R.AGRYAKRV	GKGAPVYLAA	VLEYLSAELL	ELAGNASRDN	KKKRITPR
H2A2_XENLA/4-119	R.KGNYAERV	GAGAPVYLAA	VLEYLTAEIL	ELAWERLPEI	TKRPVLSP
H2AV_CHICK/6-121	KTRTTSHGRV	GATAAVYSAA	ILEYLTAEVL	ELAGNASKDL	KVKRITPR
H2AV_TETHH/6-131	KGRVSAKNRV	GATAAVYAAA	ILEYLTAEVL	ELAGNASKDF	KVRRITPR

Partial alignment of casein from PFAM ($\lambda = 3.3$)

CASK_BOVIN/2-190	VLSRYPSYGL	NYYQQKPVAL	.INNQFLPYP	YYAKPAAVRS	PAQILQWQVL
CASK_CERNI/2-190	ALSRYPYGL	NYYQHRPVAL	.INNQFLPYP	YYVKPGAVRS	PAQILQWQVL
CASK_CAMDR/1-182	VQSRYPYGI	NYYQHRLAVP	.INNQFIPYP	NYAKPVAIRL	HAQIPQCQAL
CASK_PIG/2-188	MLNRFPSYGF	.FYQHRSVAVS	.PNRQFIPYP	YYARPVVAGP	HAQKPQWQDQ
CASK_HUMAN/1-182	VPNSYPYYGT	NLYQRRPAIA	.INNPYVPRT	YYANPAVVRP	HAQIPQRQYL
CASK_RABIT/2-179	VMNRYPQYEP	SYYLRRQAVP	.TLNPFMLNP	YYVKPIVFKP	NVQVPHWQIL
CASK_CAVPO/2-181	VLNNYLRTAP	SYYQNRASVP	.INNPYLCHL	YYVPSEVLWA	QGQIPKGPVS
CASK_MOUSE/2-181	VLN.FNQYEP	NYYHYRPSLP	ATASPYMYYP	LVRLLLLLRS	PAPISKWQSM
CASK_RAT/2-178	VLN.RNHYEP	IYYHYRTSVP	.VSPYAYFP	VGLKLLLLRS	PAQILKWQPM

Most conserved proteins in worm, human, and yeast

Protein	worm/ human	worm/ yeast	yeast/ human
H4 histone	99% id	91% id	92 % id
H3.3 histone	99	89	90
Actin B	98	88	89
Ubiquitin	98	95	96
Calmodulin	96	59	58
Tubulin	94	75	76

See Copley et al. (1999), who performed reciprocal BLAST searches

Molecular clock hypothesis: implications

If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.

Molecular clock hypothesis: implications

If protein sequences evolve at constant rates, they can be used to estimate the times that sequences diverged. This is analogous to dating geological specimens by radioactive decay.

N = total number of substitutions

L = number of nucleotide sites compared
between two sequences

$K = \frac{N}{L}$ = number of substitutions
per nucleotide site

See Graur and Li (2000), p. 140

Rate of nucleotide substitution r and time of divergence T

r = rate of substitution

= 0.56×10^{-9} per site per year for hemoglobin alpha

$K = 0.093$ = number of substitutions

per nucleotide site (rat versus human)

$$r = K / 2T$$

$$T = .093 / (2)(0.56 \times 10^{-9}) = 80 \text{ million years}$$

Neutral theory of evolution

An often-held view of evolution is that just as organisms propagate through natural selection, so also DNA and protein molecules are selected for.

According to Motoo Kimura's 1968 neutral theory of molecular evolution, the vast majority of DNA changes are not selected for in a Darwinian sense. The main cause of evolutionary change is random drift of mutant alleles that are selectively neutral (or nearly neutral). Positive Darwinian selection does occur, but it has a limited role.

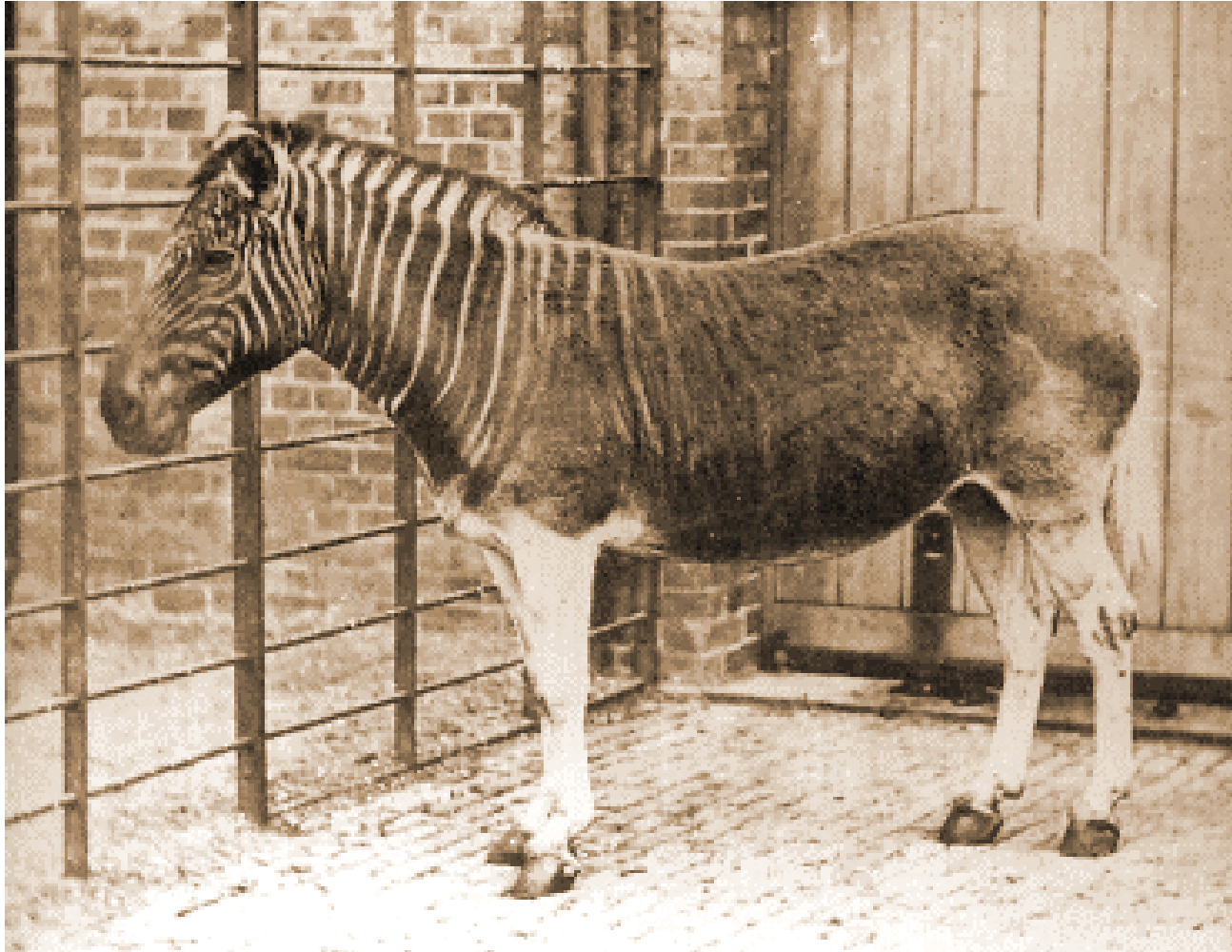
As an example, the divergent C peptide of insulin changes according to the neutral mutation rate.

Goals of molecular phylogeny

Phylogeny can answer questions such as:

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

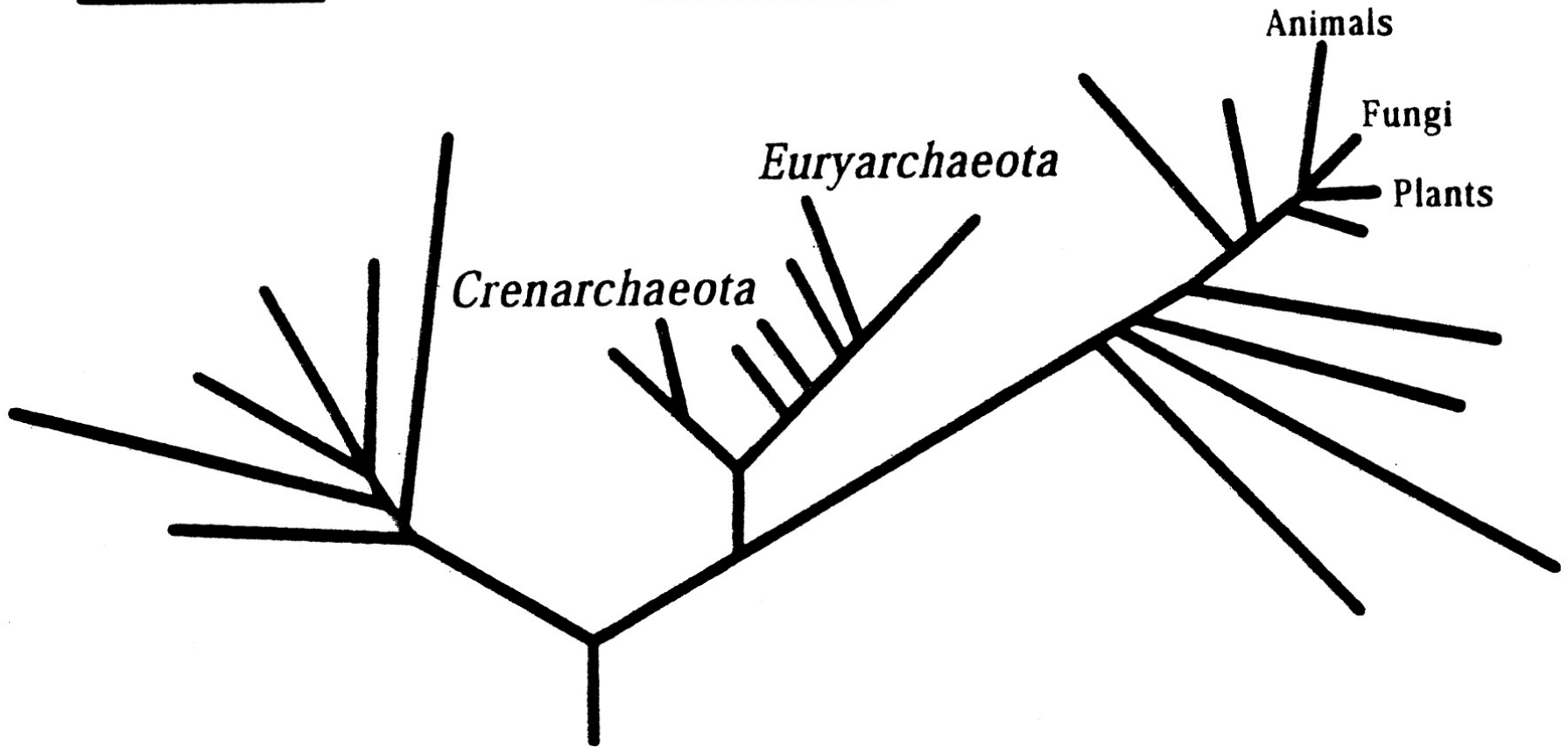
Was the quagga (now extinct) more like a zebra or a horse?



Bacteria

Archaea

Eucarya



Woese PNAS

Molecular phylogeny in bioinformatics

Many of the topics we have discussed so far involve explicit or implicit models of evolution.

Dayhoff et al. (1978) describe scoring matrices: “An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form.

Molecular phylogeny in bioinformatics

Many of the topics we have discussed so far involve explicit or implicit models of evolution.

Feng and Doolittle (1987, p. 351) use the Needleman-Wunsch algorithm “to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The sequences are assumed a priori to share a common ancestor, and the trees are constructed from different matrices derived directly from the multiple alignment.”

Molecular phylogeny: nomenclature of trees

There are two main kinds of information inherent to any tree: topology and branch lengths.

We will now describe the parts of a tree.

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

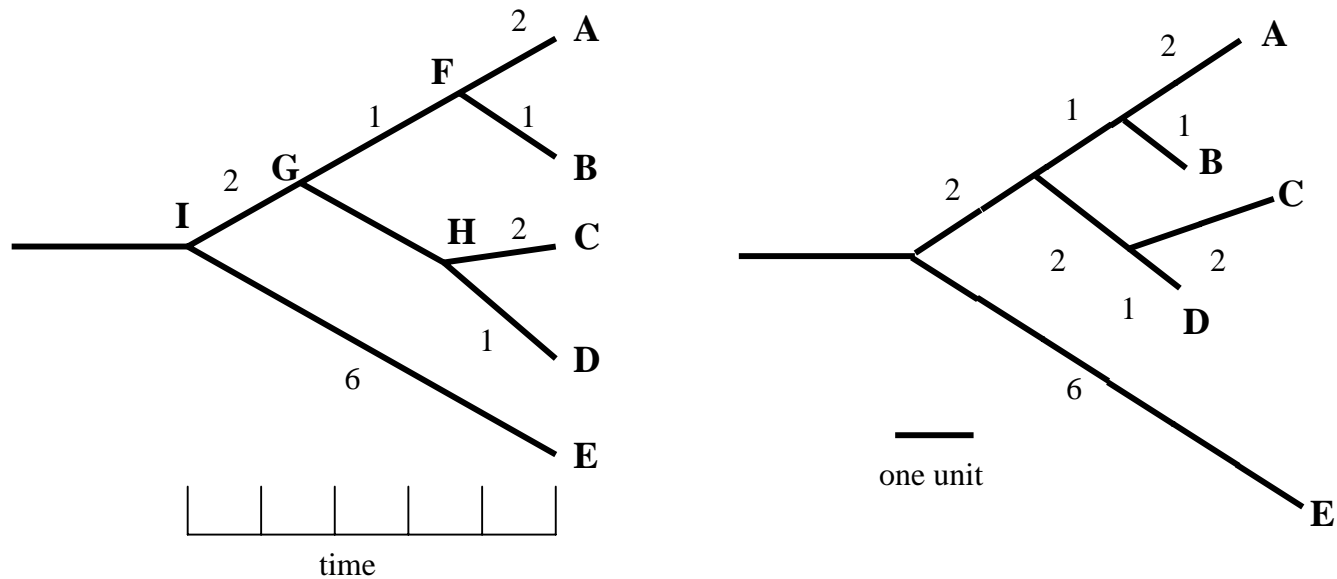


Fig. 11.4
Page 366

Tree nomenclature

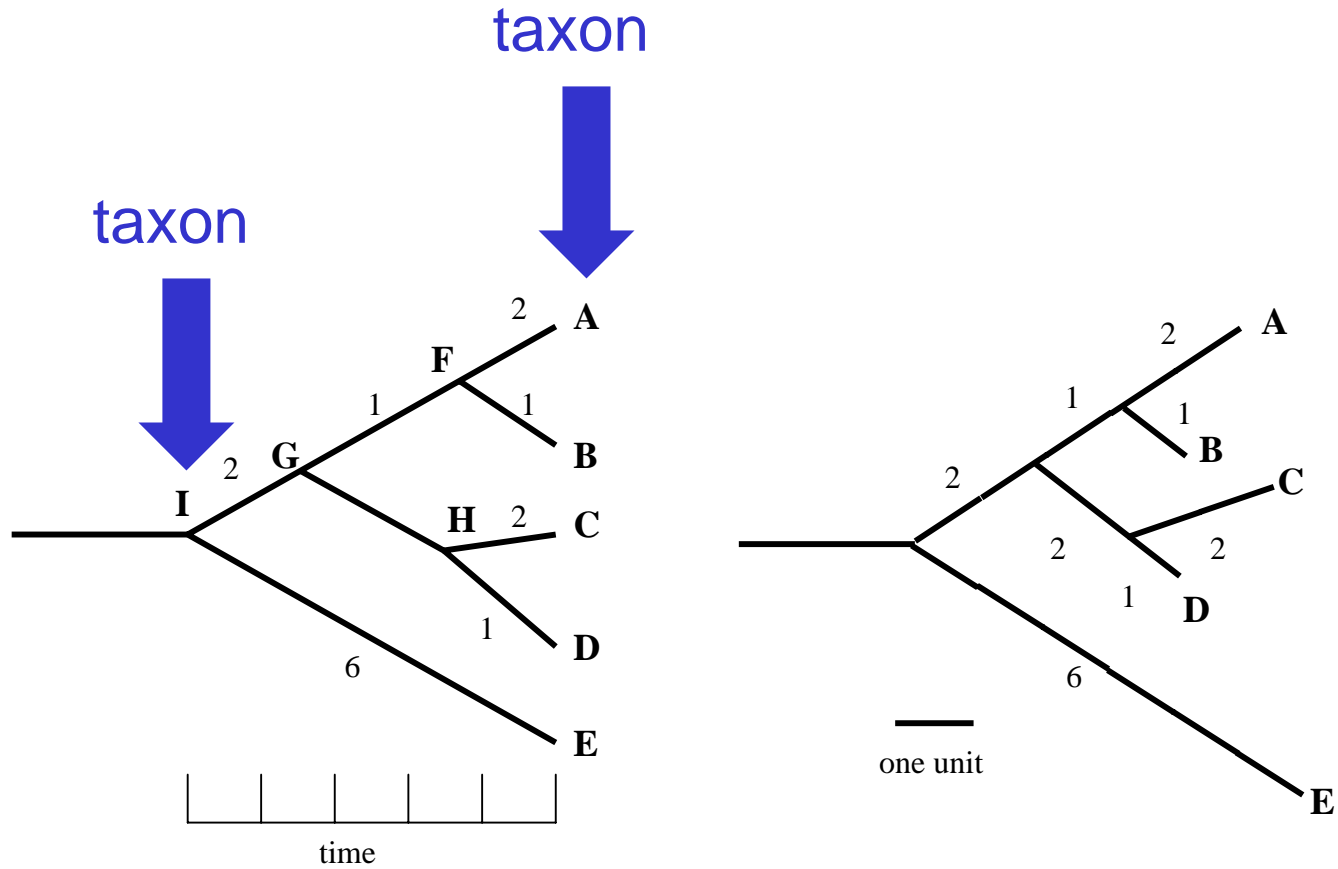


Fig. 11.4
Page 366

Tree nomenclature

operational taxonomic unit (OTU)

such as a protein sequence

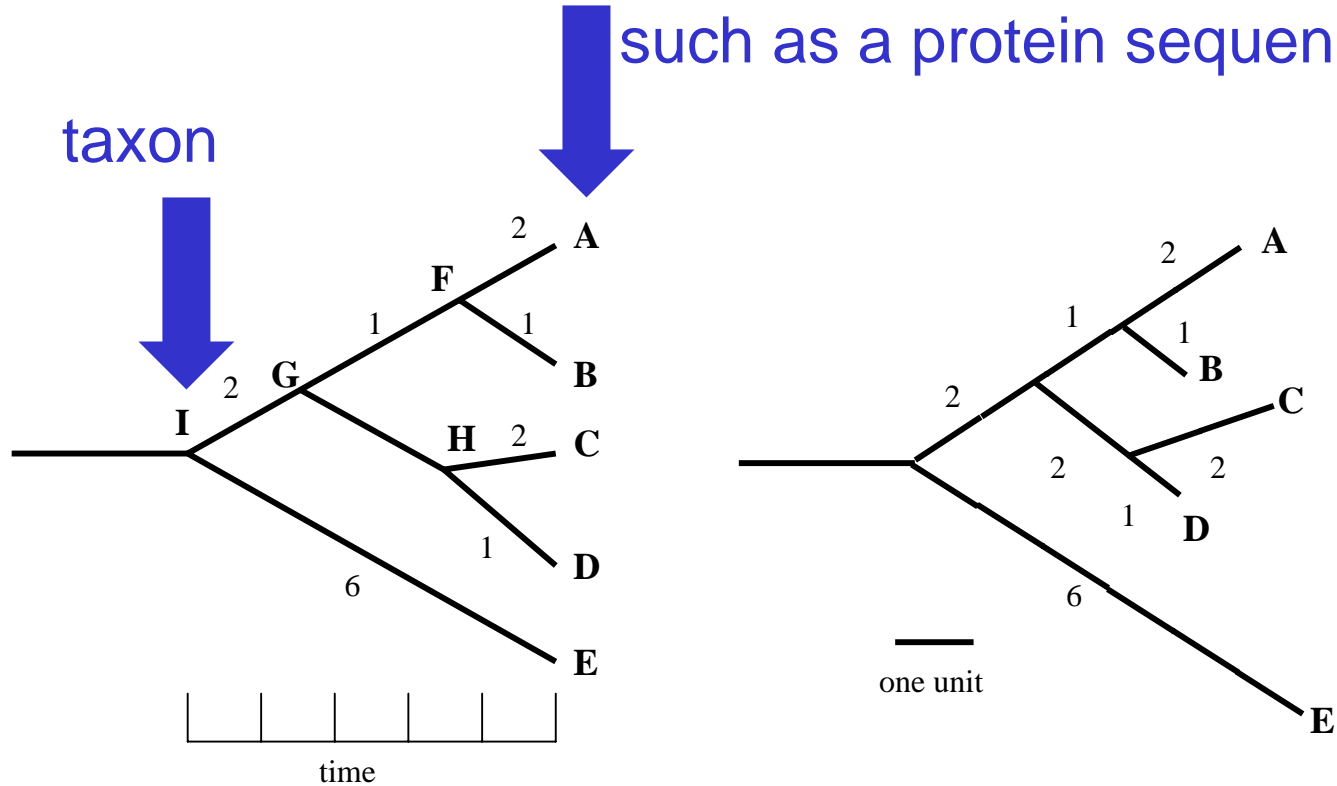


Fig. 11.4
Page 366

Tree nomenclature

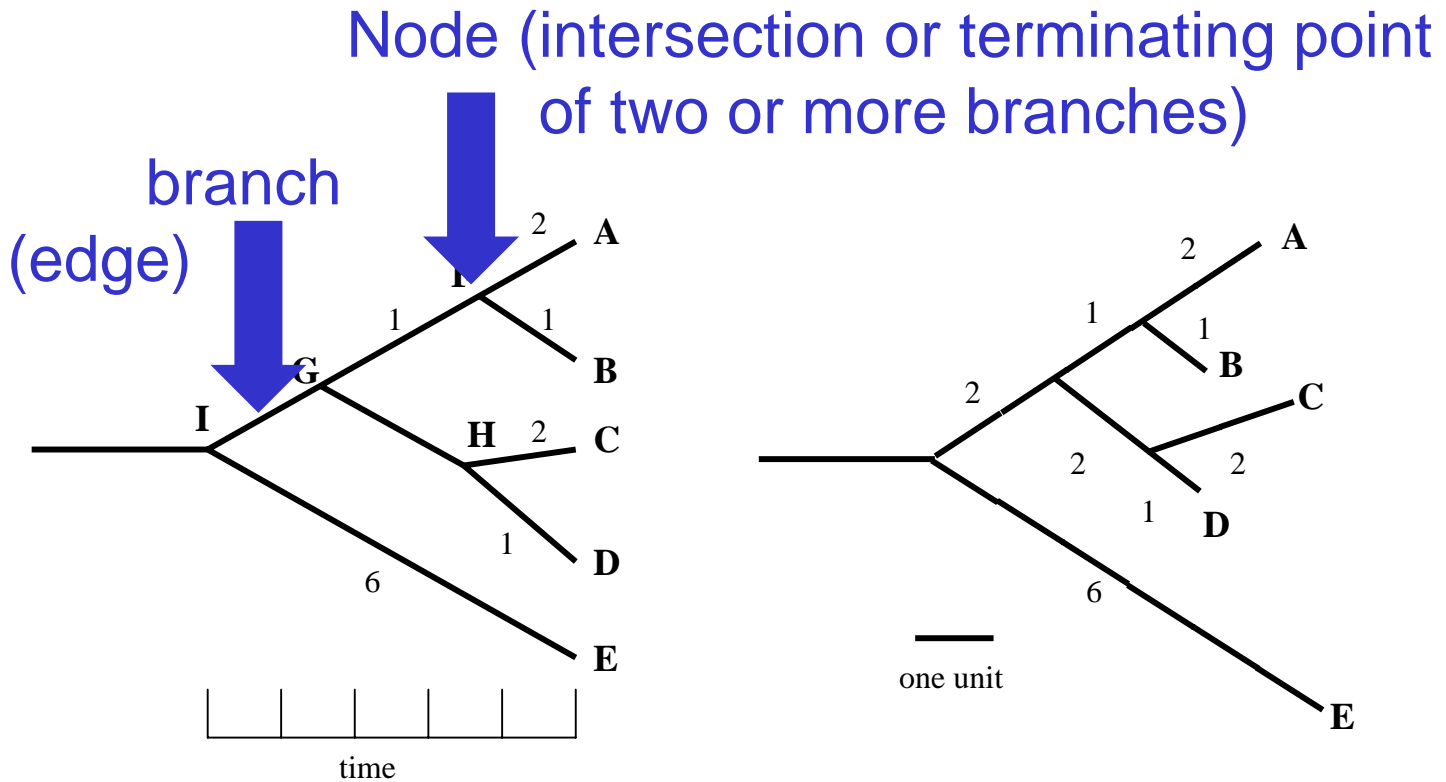
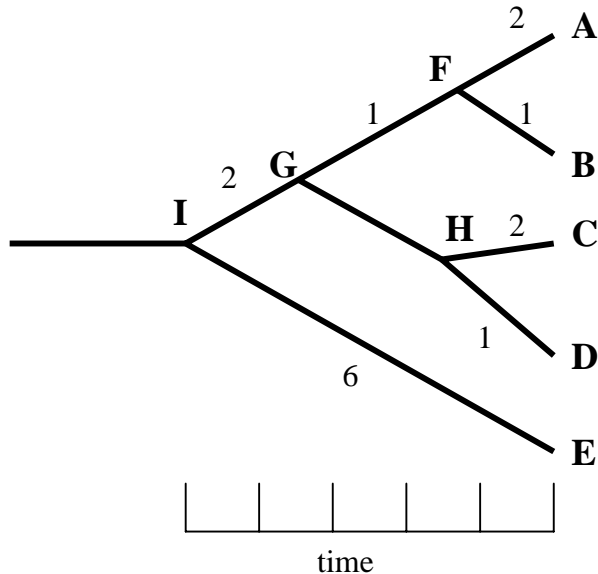


Fig. 11.4
Page 366

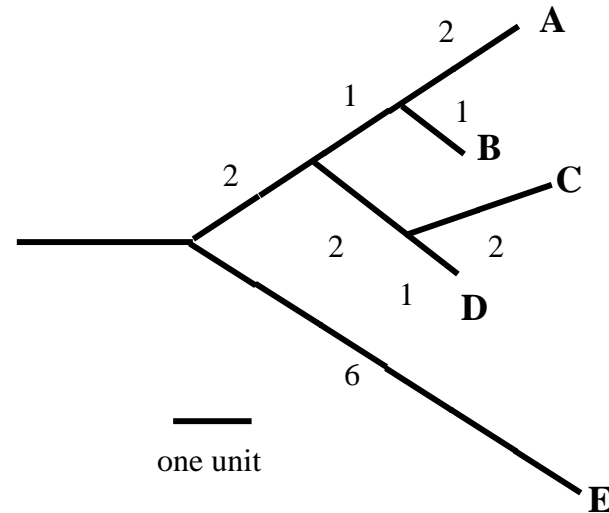
Tree nomenclature

Branches are unscaled...



...OTUs are neatly aligned,
and nodes reflect time

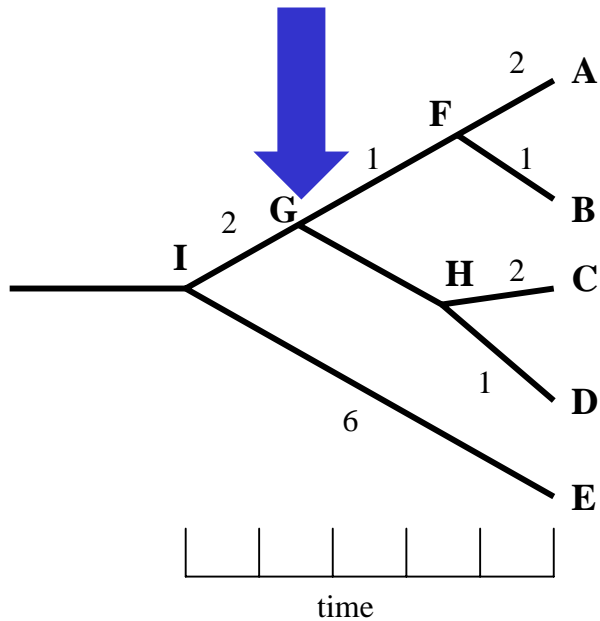
Branches are scaled...



...branch lengths are
proportional to number of
amino acid changes

Tree nomenclature

bifurcating
internal
node



multifurcating
internal
node

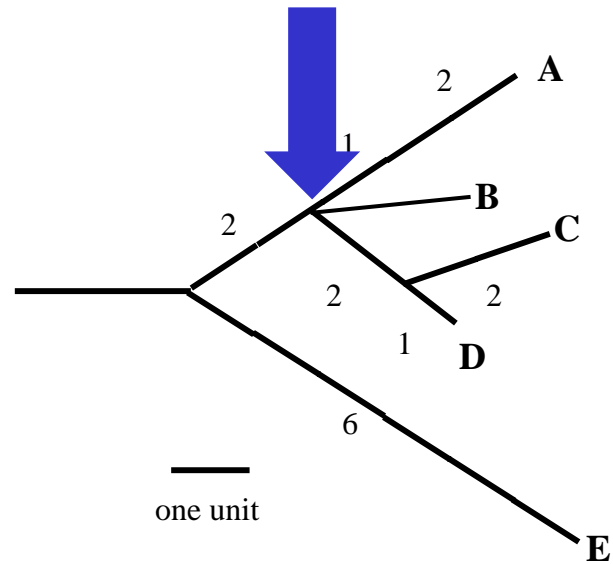
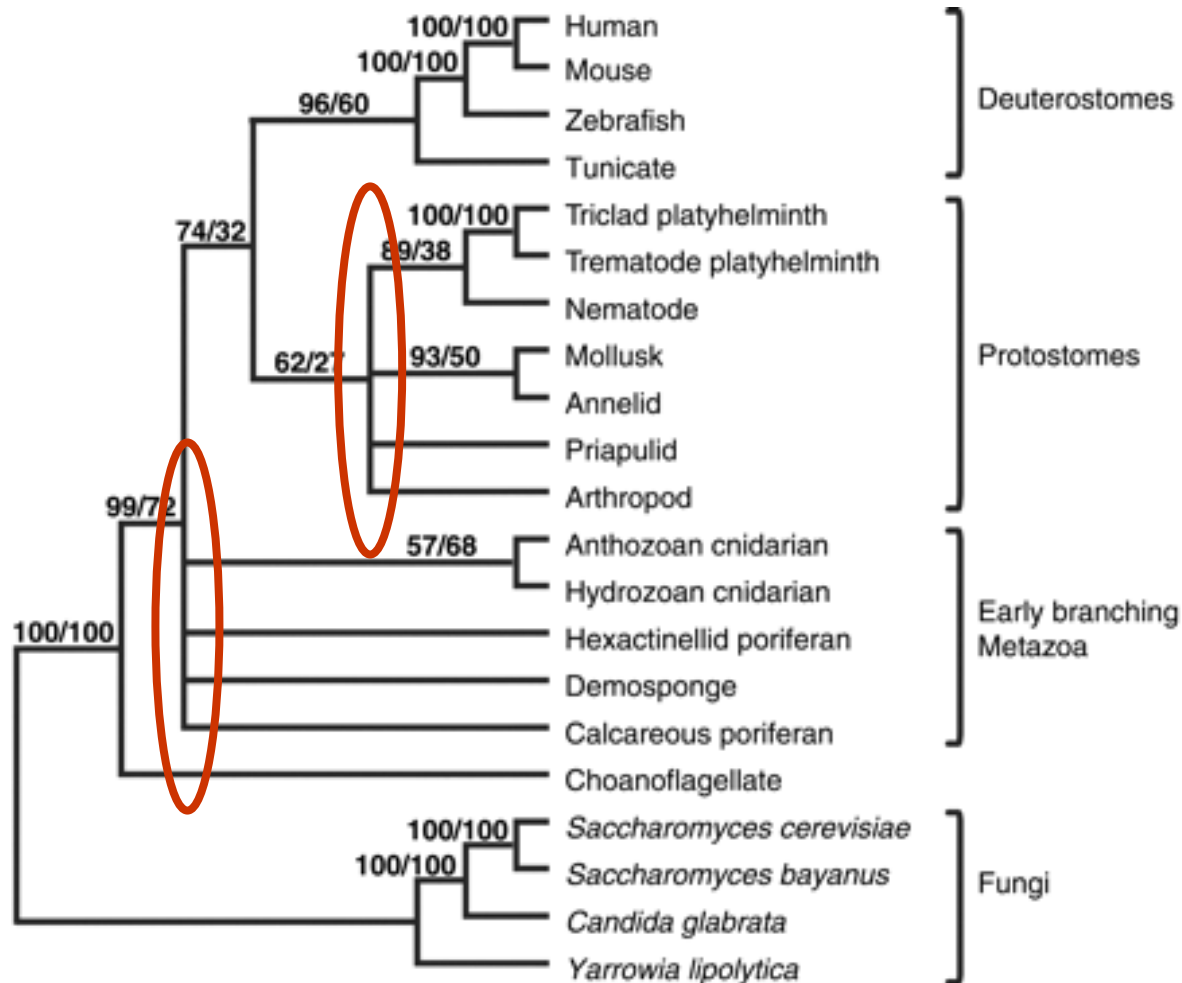


Fig. 11.5
Page 367

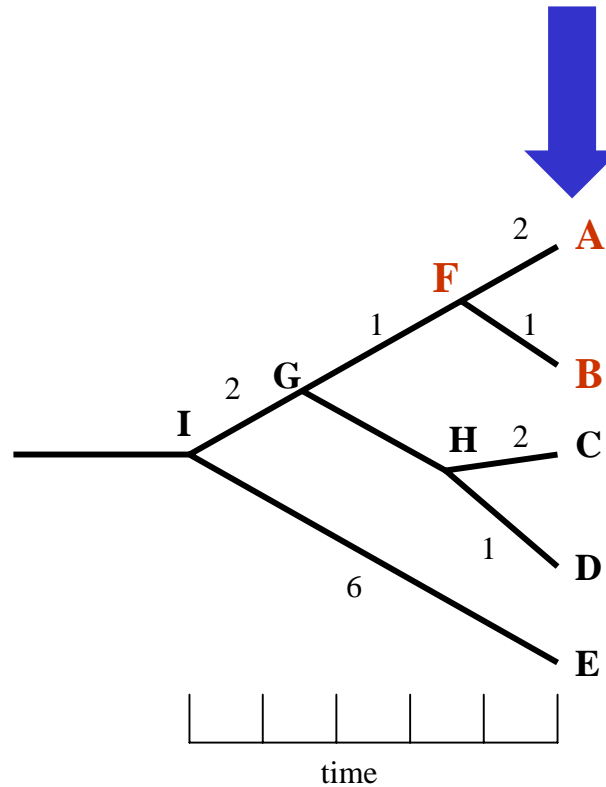
Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes



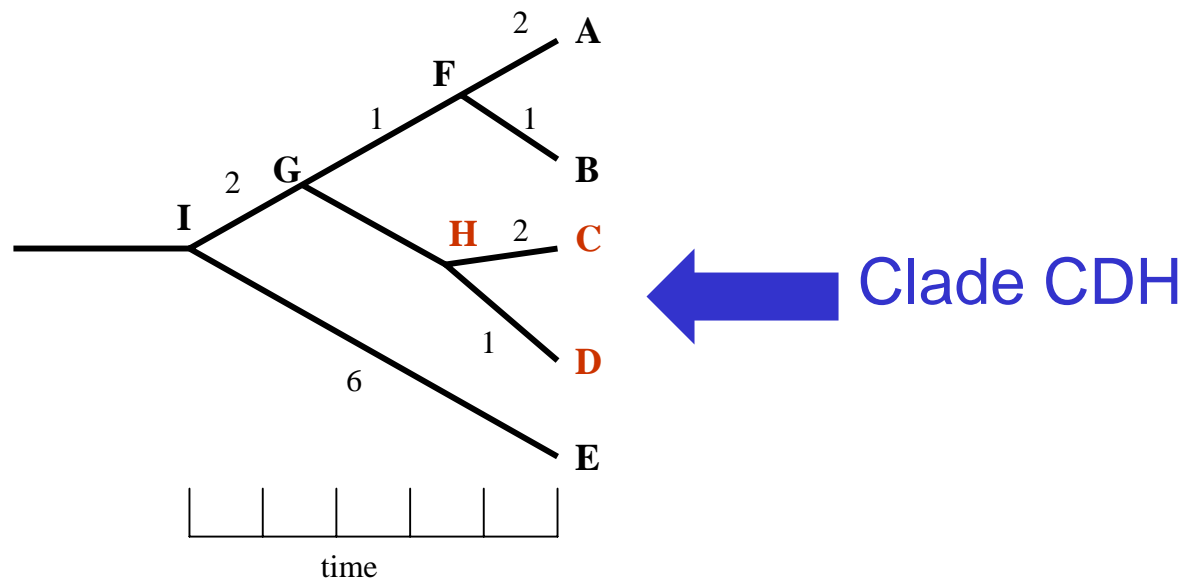
Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933, 23 December 2005, Fig. 1.

Tree nomenclature: clades

Clade ABF (monophyletic group)



Tree nomenclature



Tree nomenclature

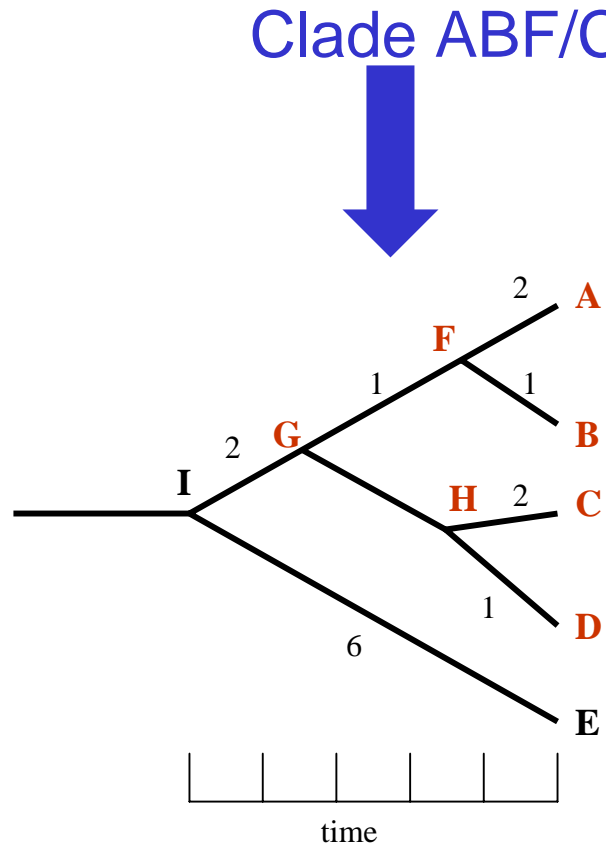
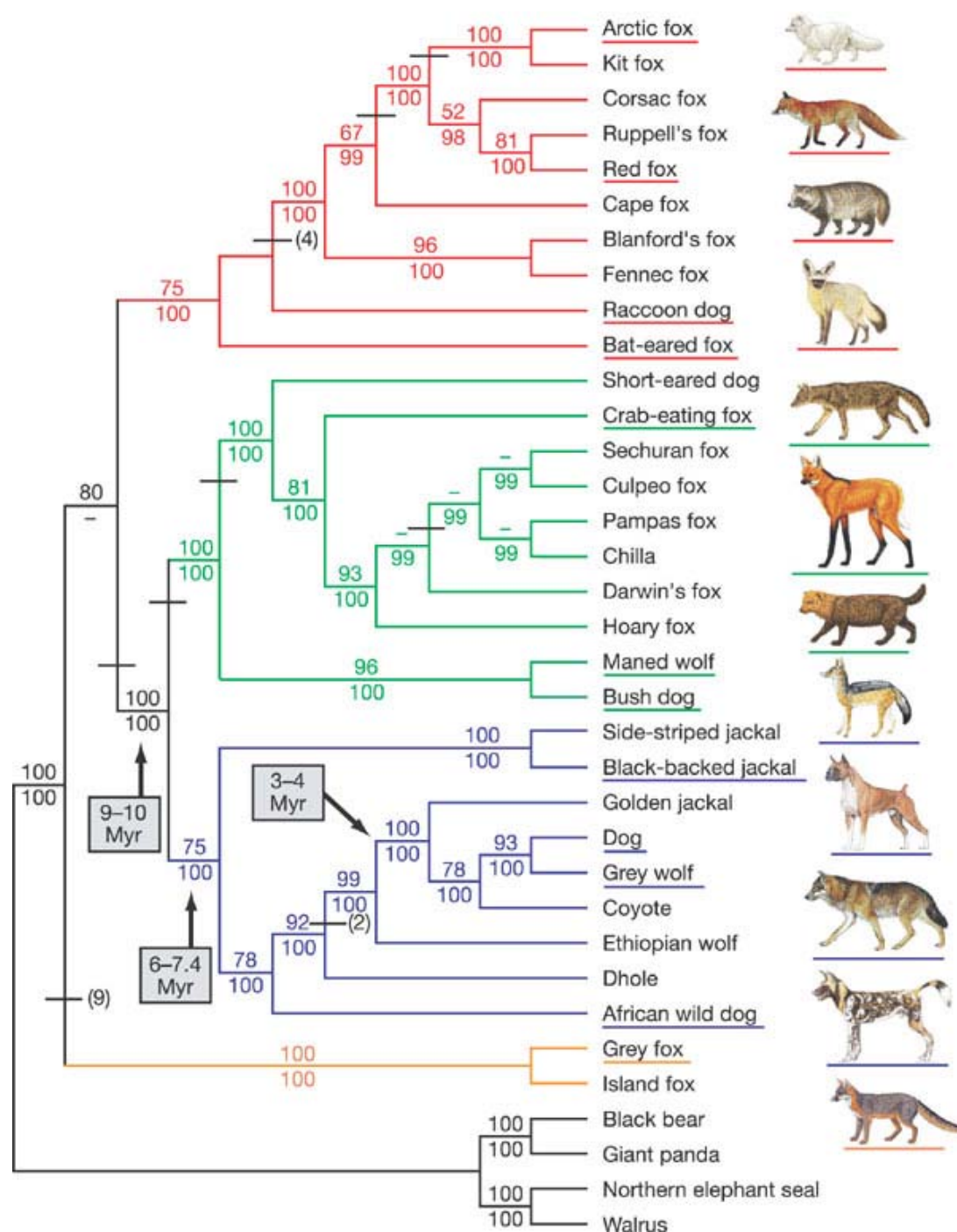


Fig. 11.4
Page 366

Examples of clades



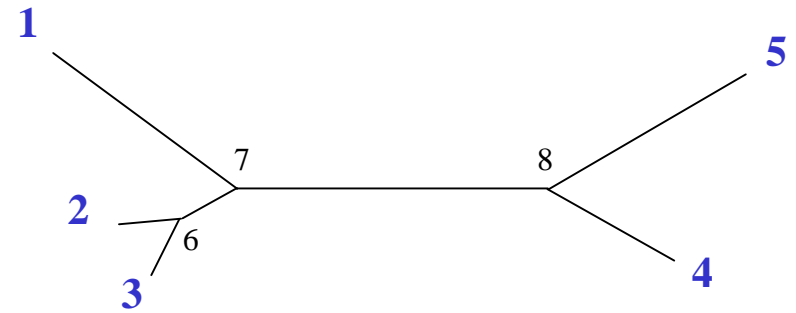
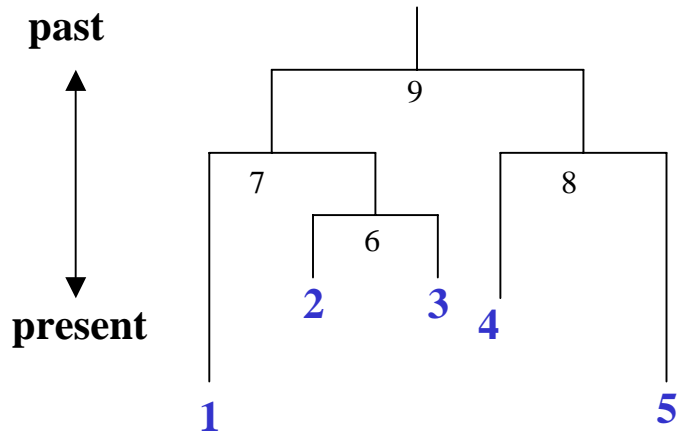
Lindblad-Toh et al., *Nature* 438: 803, 8 Dec. 2005, fig. 10

Tree roots

The root of a phylogenetic tree represents the common ancestor of the sequences. Some trees are unrooted, and thus do not specify the common ancestor.

A tree can be rooted using an outgroup (that is, a taxon known to be distantly related from all other OTUs).

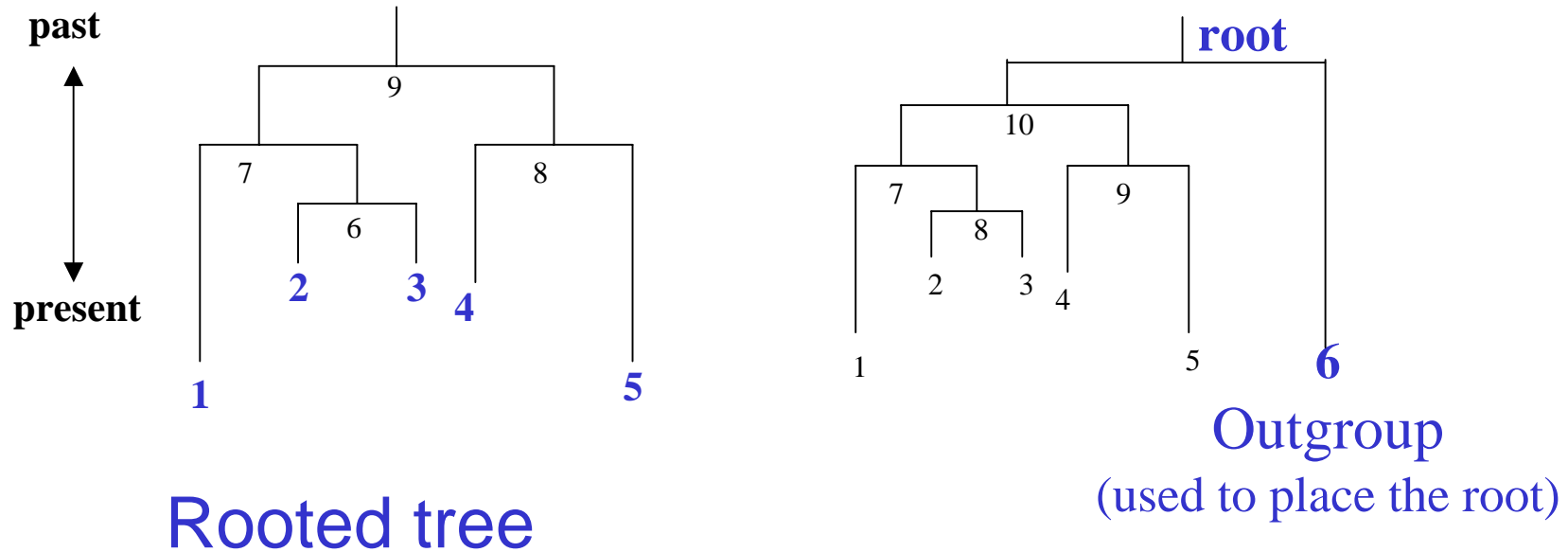
Tree nomenclature: roots



Rooted tree
(specifies evolutionary
path)

Unrooted tree

Tree nomenclature: outgroup rooting



Enumerating trees

Cavalli-Sforza and Edwards (1967) derived the number of possible unrooted trees (N_U) for n OTUs ($n \geq 3$):

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

The number of bifurcating rooted trees (N_R)

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

For 10 OTUs (e.g. 10 DNA or protein sequences), the number of possible rooted trees is ≈ 34 million, and the number of unrooted trees is ≈ 2 million. Many tree-making algorithms can exhaustively examine every possible tree for up to ten to twelve sequences.

Numbers of trees

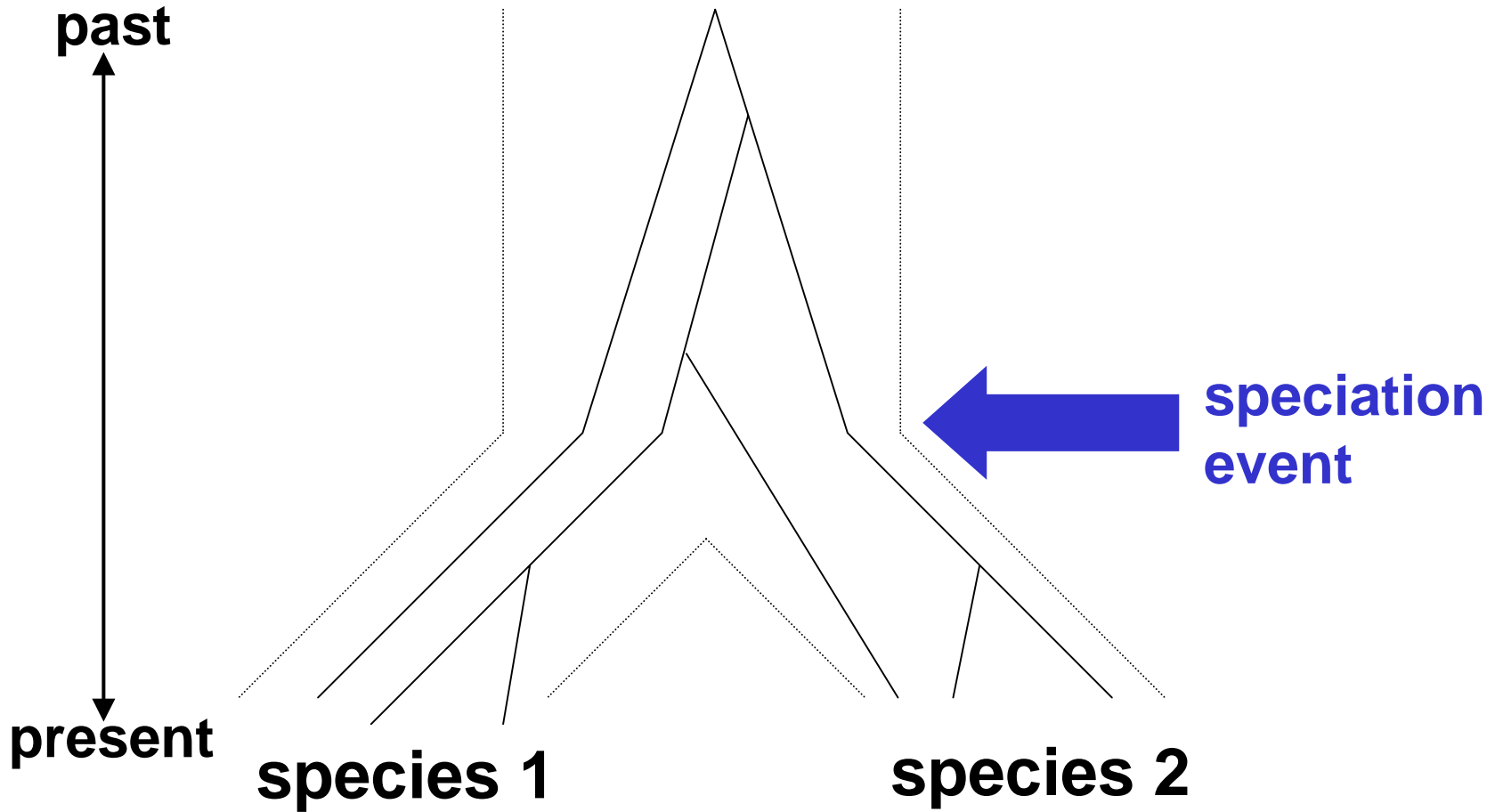
<u>Number of OTUs</u>	<u>Number of rooted trees</u>	<u>Number of unrooted trees</u>
2	1	1
3	3	1
4	15	3
5	105	15
10	34,459,425	105
20	8×10^{21}	2×10^{20}

Species trees versus gene/protein trees

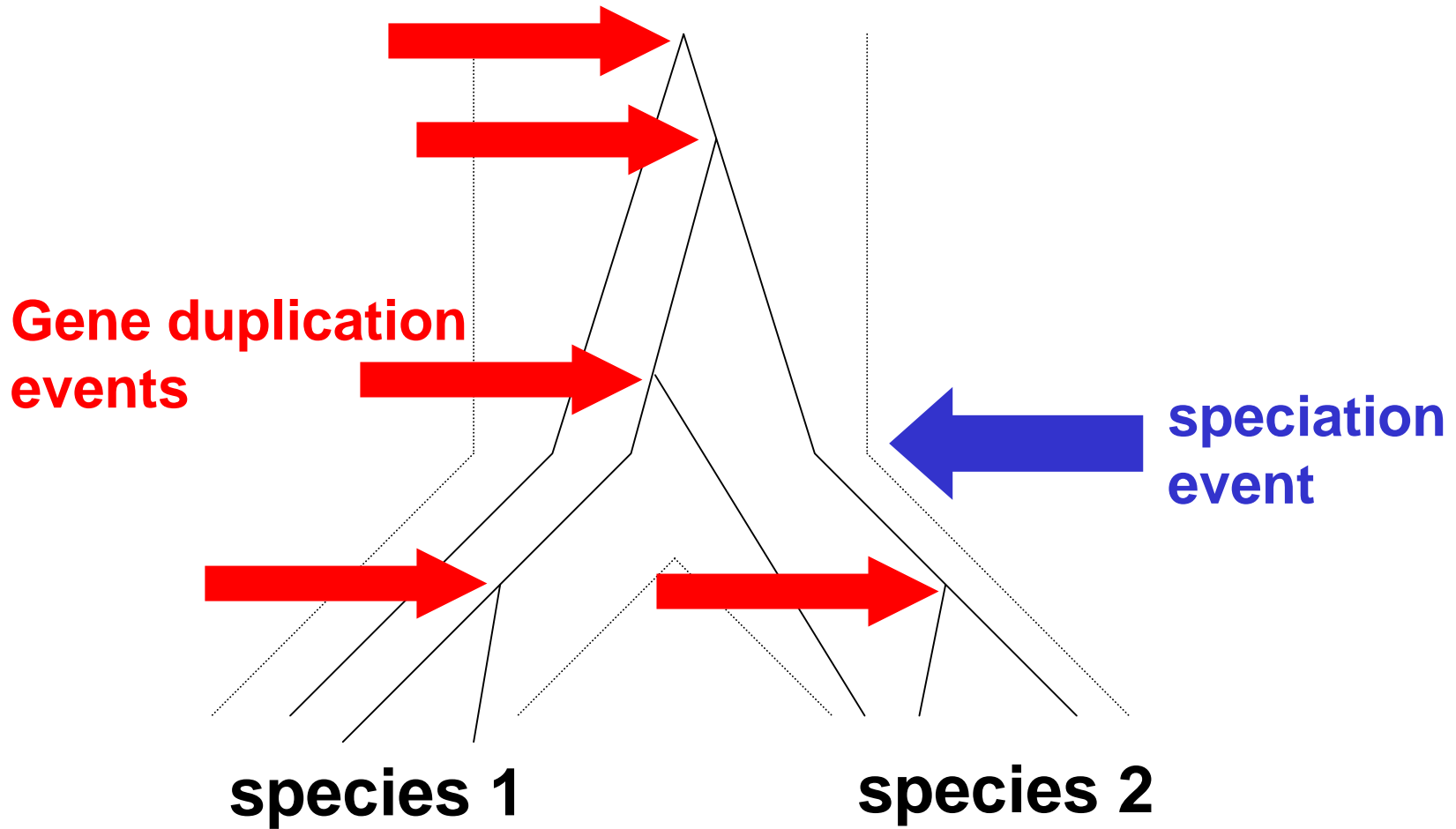
Molecular evolutionary studies can be complicated by the fact that both species and genes evolve. speciation usually occurs when a species becomes reproductively isolated. In a species tree, each internal node represents a speciation event.

Genes (and proteins) may duplicate or otherwise evolve before or after any given speciation event. The topology of a gene (or protein) based tree may differ from the topology of a species tree.

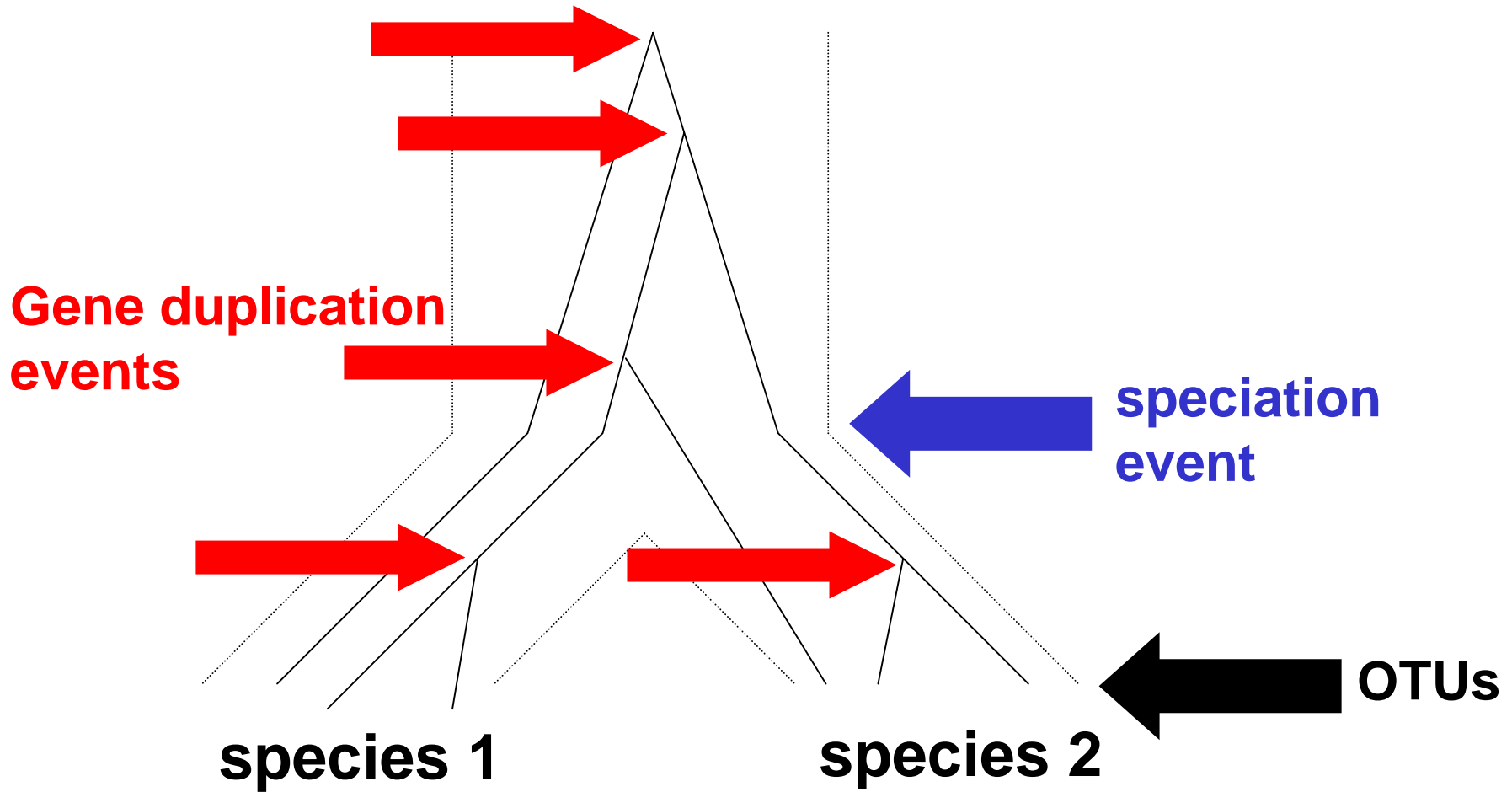
Species trees versus gene/protein trees



Species trees versus gene/protein trees



Species trees versus gene/protein trees



Computer lab

We have loaded MEGA3 onto PCs in the computer lab. Try to use this program before Friday's lab. Sample datasets are available on the course website.

For the find-a-gene project, try putting your novel protein into a phylogenetic tree using PAUP.

The phylogeny lecture continues with part II on Wednesday.