

Molecular Phylogeny and Evolution

December 15, 2008

Bioinformatics
J. Pevsner
pevsner@kennedykrieger.org

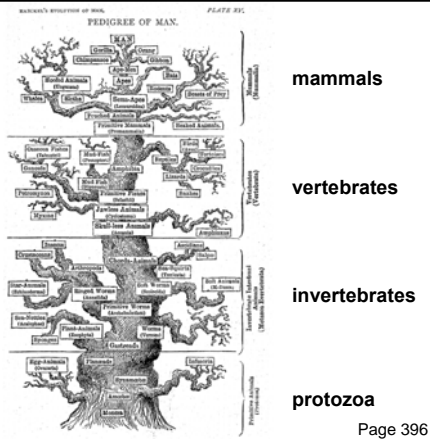
Copyright notice

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>

Five kingdom system
(Haeckel, 1879)



animals
plants
fungi
protists
monera

Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

- Five stages of molecular phylogeny:
- [1] selecting sequences
 - [2] multiple sequence alignment
 - [3] models of substitution
 - [4] tree-building
 - [5] tree evaluation

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.

Page 357

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.

Page 358

Historical background

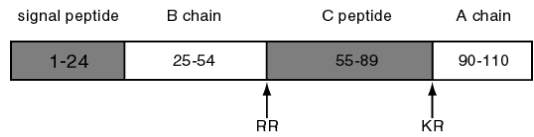
Studies of molecular evolution began with the first sequencing of proteins, beginning in the 1950s.

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

(The accession number of human insulin is NP_000198)

Page 358

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

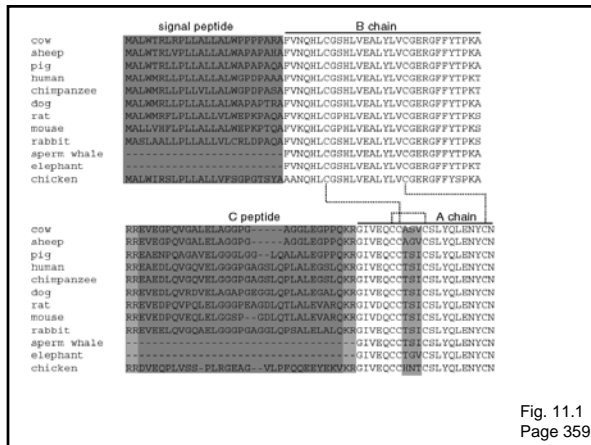


Fig. 11.1
Page 359

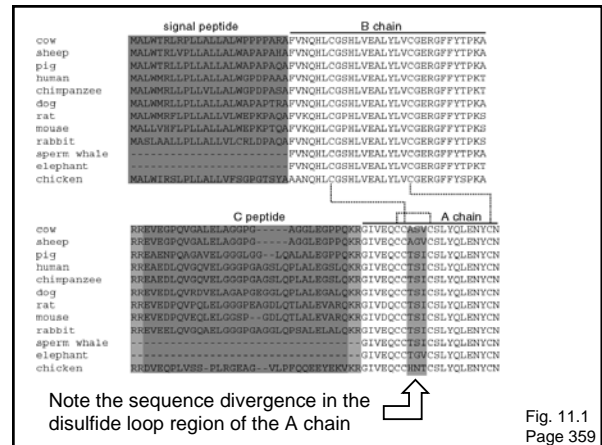


Fig. 11.1
Page 359

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.

Page 358

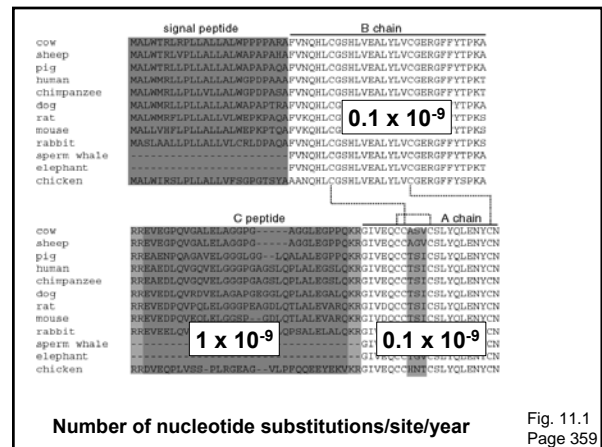


Fig. 11.1
Page 359

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.

Page 360

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

```

human      MALWMRLRLLPLLALLALWGPDPAAAFVWQHLCGSHLVEALYLVCGERGFFYPPTK
mouse      MALLVHFLPLLALLALWEPKPTQAFVKQHLCPHLVEALYLVCGERGFFYPPTKS
guinea pig MALWMHLLTVLALLALWGPNTQQA FVSRRLCGSNLVEVLYSVCQDDGFFYPK
human      RREAEDLQVQVELGGGPGAGSLQPLALEFSLQKRGIVDCCCTSI CSLYOLENYCN
mouse      RREVEDPQVEQELGGSP--GDLQTLALVARQKRIVDCCCTSI CSLYOLENYCN
guinea pig RRELEDPQVEQTELGMGLGAGLQPLALEMALQKRIVDCCCTGCTRHOLESYCN
    
```

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

Fig. 11.1
Page 359

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

Page 360

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.

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

Page 360

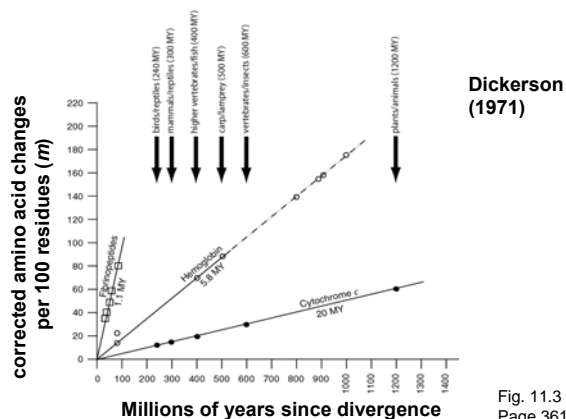


Fig. 11.3
Page 361

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.

Page 361

Molecular clock hypothesis: implications

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

Page 362

Positive and negative selection

Darwin's theory of evolution suggests that, at the phenotypic level, traits in a population that enhance survival are selected for, while traits that reduce fitness are selected against. For example, among a group of giraffes millions of years in the past, those giraffes that had longer necks were able to reach higher foliage and were more reproductively successful than their shorter-necked group members, that is, the taller giraffes were selected for.

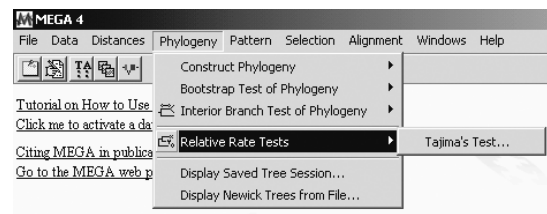
In the mid-20th century, a conventional view was that molecular sequences are routinely subject to positive (or negative) selection.

Positive and negative selection

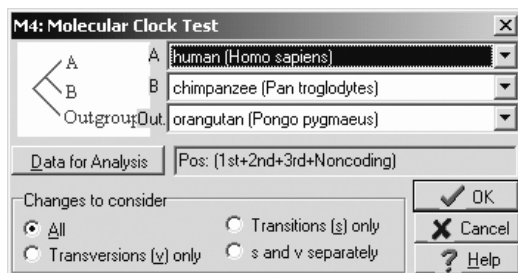
Darwin's theory of evolution suggests that, at the phenotypic level, traits in a population that enhance survival are selected for, while traits that reduce fitness are selected against. For example, among a group of giraffes millions of years in the past, those giraffes that had longer necks were able to reach higher foliage and were more reproductively successful than their shorter-necked group members, that is, the taller giraffes were selected for.

Positive selection occurs when a sequence undergoes significantly increased rates of substitution, while negative selection occurs when a sequence undergoes change slowly. Otherwise, selection is neutral.

Tajima's relative rate test in MEGA



Tajima's relative rate test



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.

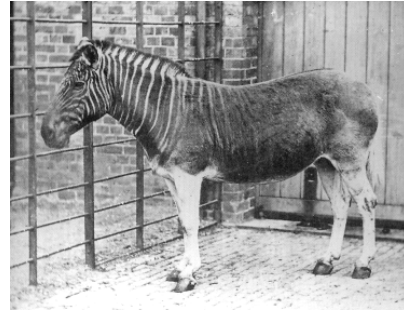
Page 363

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?



Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

Five stages of molecular phylogeny:

- [1] selecting sequences
- [2] multiple sequence alignment
- [3] models of substitution
- [4] tree-building
- [5] tree evaluation

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.

Page 366

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

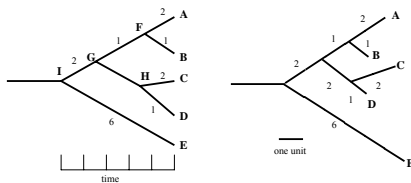


Fig. 11.4
Page 366

Tree nomenclature

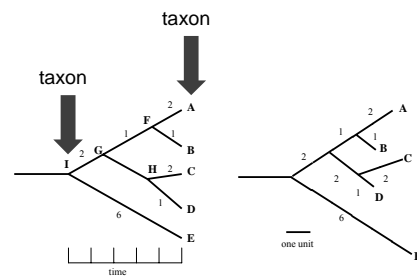
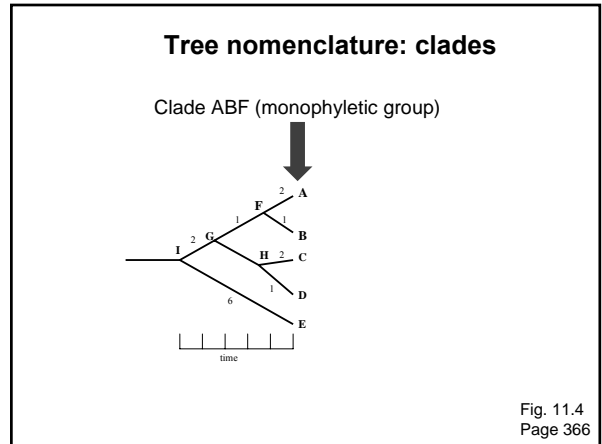
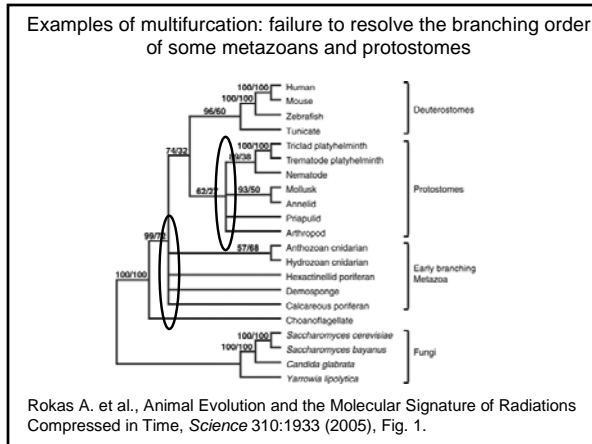
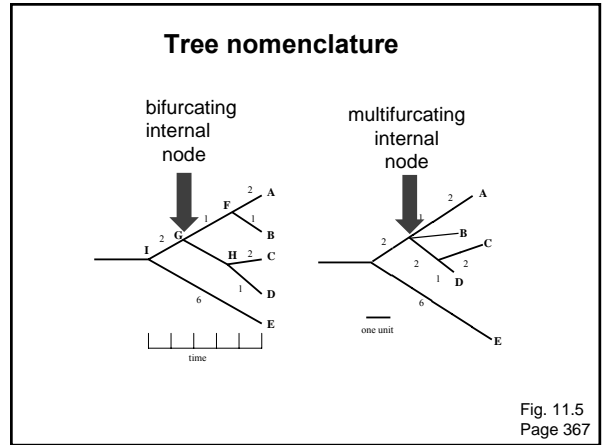
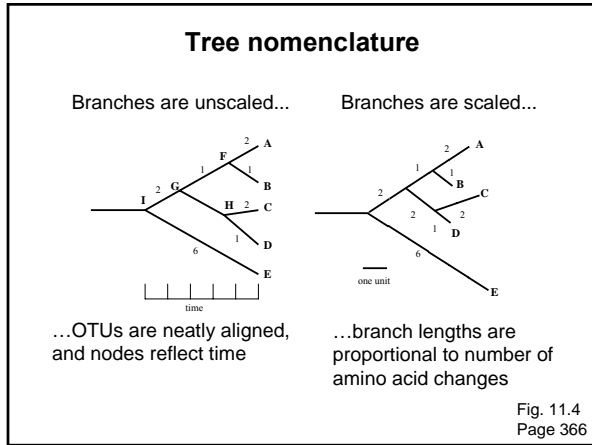
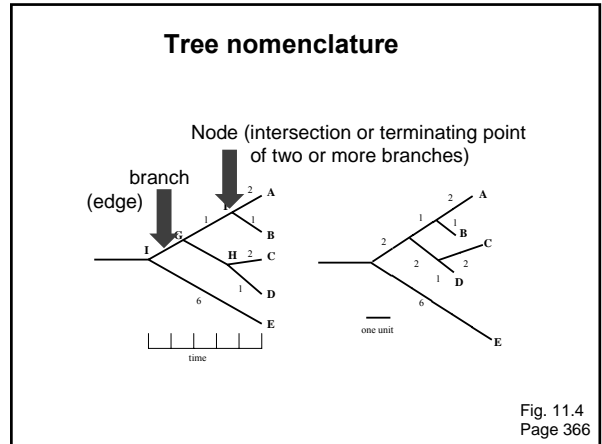
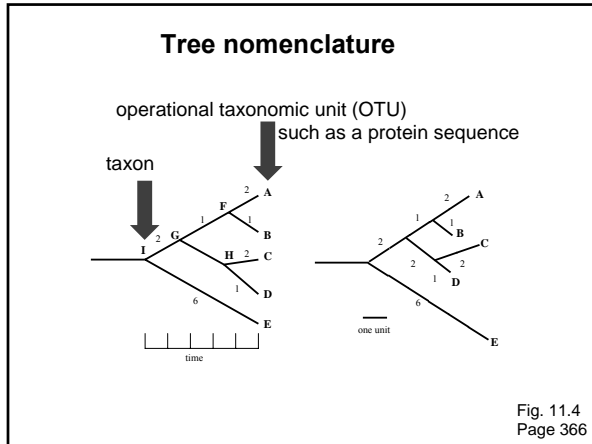
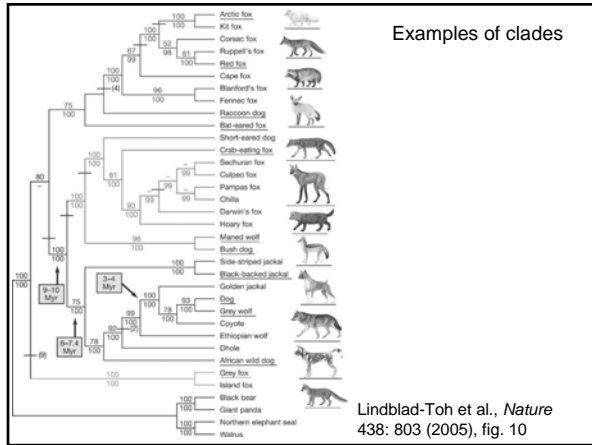
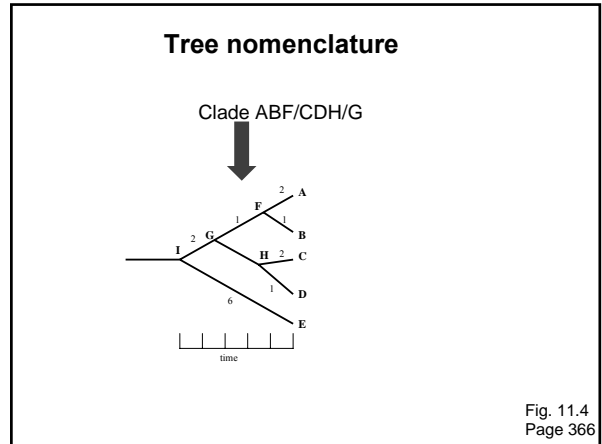
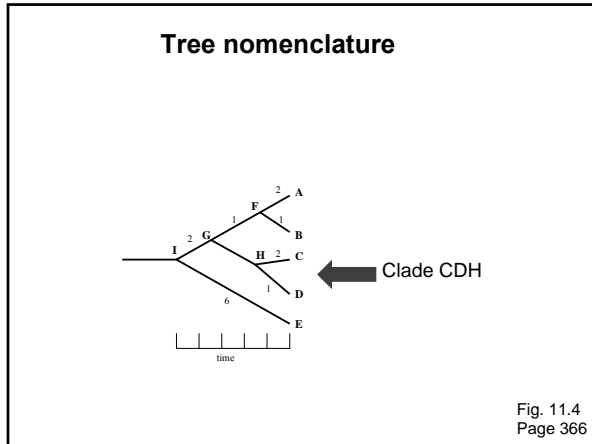


Fig. 11.4
Page 366



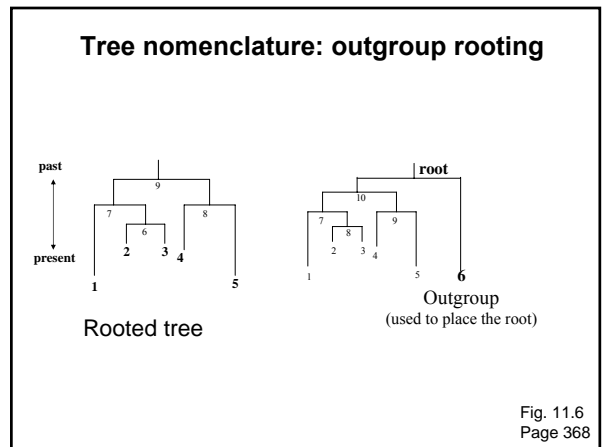
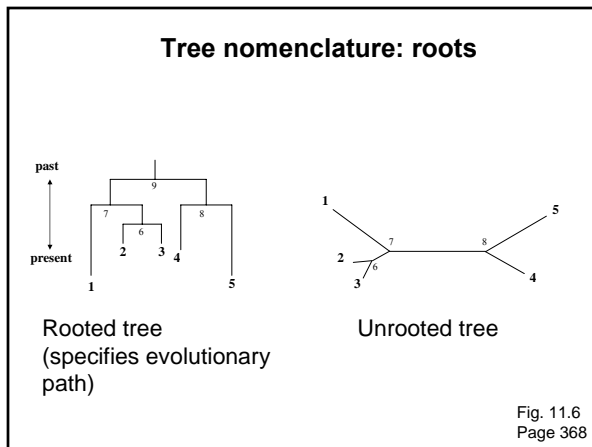


Tree roots

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).

Page 368



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.

Page 368

Numbers of trees

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

Box 11-2
Page 369

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.

Page 370

Species trees versus gene/protein trees

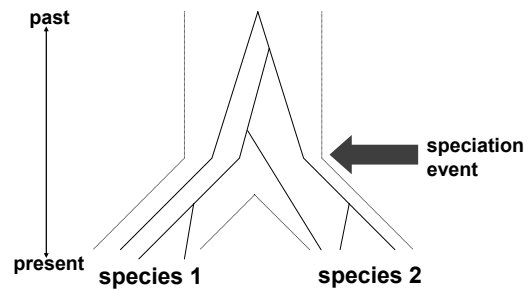


Fig. 11.9
Page 372

Species trees versus gene/protein trees

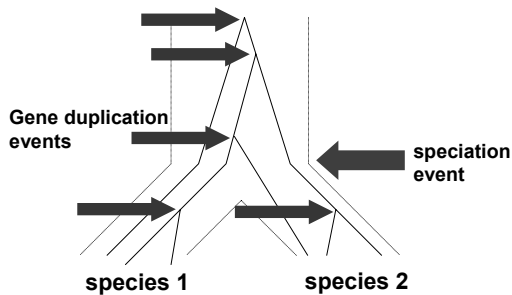


Fig. 11.9
Page 372

Species trees versus gene/protein trees

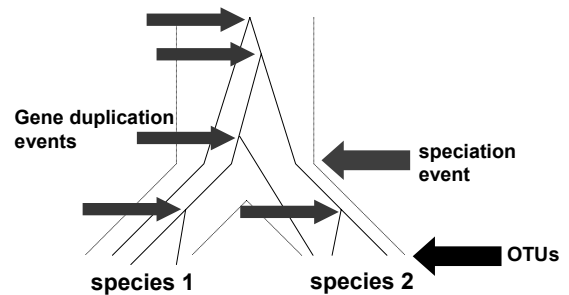


Fig. 11.9
Page 372

Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

Five stages of molecular phylogeny:

- [1] selecting sequences
- [2] multiple sequence alignment
- [3] models of substitution
- [4] tree-building
- [5] tree evaluation

Stage 1: Use of DNA, RNA, or protein

For some phylogenetic studies, it may be preferable to use protein instead of DNA sequences.

We saw that in pairwise alignment and in BLAST searching, protein is often more informative than DNA (Chapter 3). Proteins have 20 states (amino acids) instead of only four for DNA, so there is a stronger phylogenetic signal.

Page 371

Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

--The protein-coding portion of DNA has synonymous and nonsynonymous substitutions. Thus, some DNA changes do not have corresponding protein changes.

Page 371

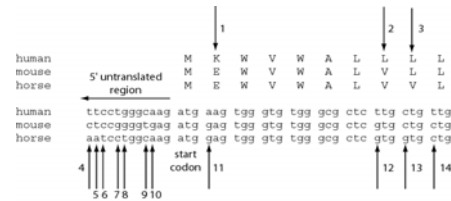


Fig. 11.10
Page 373

Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

--The protein-coding portion of DNA has synonymous and nonsynonymous substitutions. Thus, some DNA changes do not have corresponding protein changes.

If the synonymous substitution rate (\hat{d}_S) is greater than the nonsynonymous substitution rate (\hat{d}_N), the DNA sequence is under negative (purifying) selection. This limits change in the sequence (e.g. insulin A chain).

If $\hat{d}_S < \hat{d}_N$, positive selection occurs. For example, a duplicated gene may evolve rapidly to assume new functions.

Page 372

Stage 1: Use of DNA, RNA, or protein

You can measure the synonymous and nonsynonymous substitution rates by pasting your fasta-formatted sequences into the SNAP program at the Los Alamos National Labs HIV database (hiv-web.lanl.gov/).



Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

--Some substitutions in a DNA sequence alignment can be directly observed: single nucleotide substitutions, sequential substitutions, coincidental substitutions.

Page 372

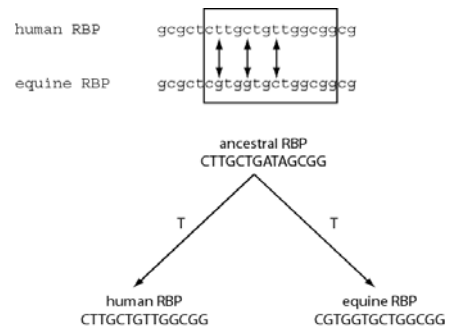


Fig. 11.11
Page 374

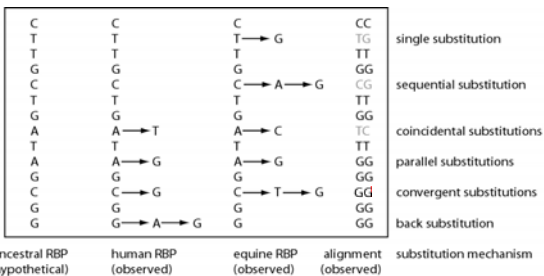


Fig. 11.11
Page 374

Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

--Some substitutions in a DNA sequence alignment can be directly observed: single nucleotide substitutions, sequential substitutions, coincidental substitutions.

Additional mutational events can be inferred by analysis of ancestral sequences. These changes include parallel substitutions, convergent substitutions, and back substitutions.

Page 372

Stage 1: Use of DNA, RNA, or protein

For phylogeny, DNA can be more informative.

--Noncoding regions (such as 5' and 3' untranslated regions) may be analyzed using molecular phylogeny.

--Pseudogenes (nonfunctional genes) are studied by molecular phylogeny

--Rates of transitions and transversions can be measured.

Transitions: purine (A ↔ G) or pyrimidine (C ↔ T) substitutions
 Transversion: purine ↔ pyrimidine

Page 372

Models of nucleotide substitution

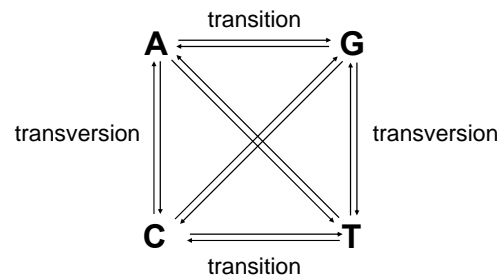
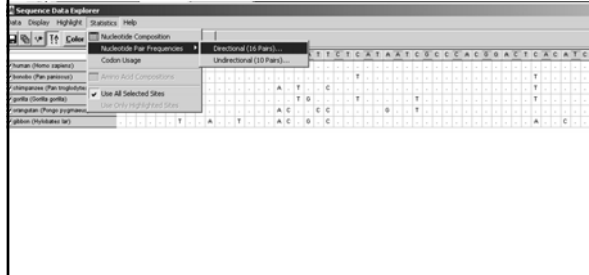
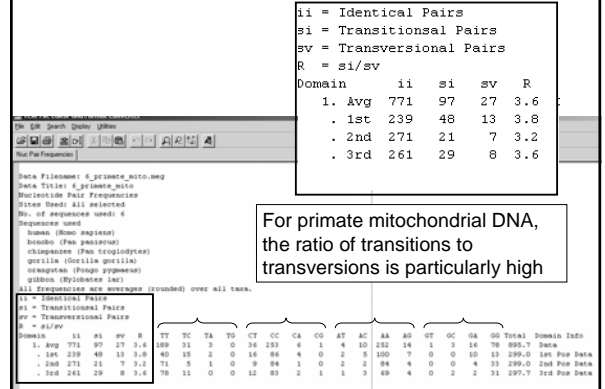


Fig. 11.14
Page 379

MEGA outputs transition and transversion frequencies



MEGA outputs transition and transversion frequencies



Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

Five stages of molecular phylogeny:

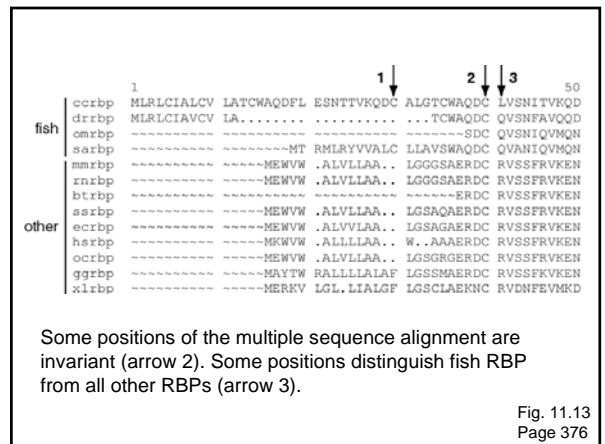
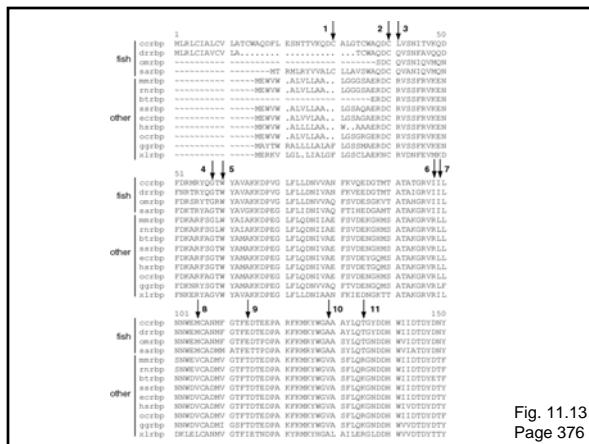
- [1] selecting sequences
- [2] multiple sequence alignment
- [3] models of substitution
- [4] tree-building
- [5] tree evaluation

Stage 2: Multiple sequence alignment

The fundamental basis of a phylogenetic tree is a multiple sequence alignment.

(If there is a misalignment, or if a nonhomologous sequence is included in the alignment, it will still be possible to generate a tree.)

Consider the following alignment of 13 orthologous retinol-binding proteins.



Stage 2: Multiple sequence alignment

- [1] Confirm that all sequences are homologous
- [2] Adjust gap creation and extension penalties as needed to optimize the alignment
- [3] Restrict phylogenetic analysis to regions of the multiple sequence alignment for which data are available for all taxa (delete columns having incomplete data).
- [4] Many experts recommend that you delete any column of an alignment that contains gaps (even if the gap occurs in only one taxon)

In this example, note that four RBPs are from fish, while the others are vertebrates that evolved more recently.

Page 375

Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

Five stages of molecular phylogeny:

- [1] selecting sequences
- [2] multiple sequence alignment
- [3] models of substitution
- [4] tree-building
- [5] tree evaluation

Stage 3: Tree-building models: distance

The simplest approach to measuring distances between sequences is to align pairs of sequences, and then to count the number of differences. The degree of divergence is called the Hamming distance. For an alignment of length N with n sites at which there are differences, the degree of divergence D is:

$$D = n / N$$

Page 378

Stage 3: Tree-building models: distance

The simplest approach to measuring distances between sequences is to align pairs of sequences, and then to count the number of differences. The degree of divergence is called the Hamming distance. For an alignment of length N with n sites at which there are differences, the degree of divergence D is:

$$D = n / N$$

But observed differences do not equal genetic distance! Genetic distance involves mutations that are not observed directly (see earlier figure).

Page 378

Stage 3: Tree-building models: distance

Jukes and Cantor (1969) proposed a corrective formula:

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

This model describes the probability that one nucleotide will change into another. It assumes that each residue is equally likely to change into any other (i.e. the rate of transversions equals the rate of transitions). In practice, the transition is typically greater than the transversion rate.

Page 379

Models of nucleotide substitution

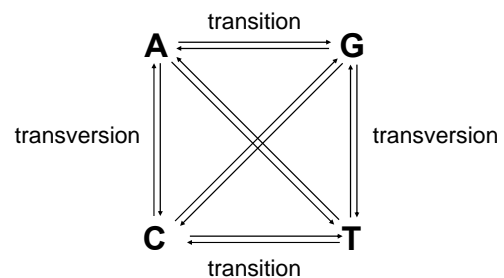


Fig. 11.14
Page 379

Jukes and Cantor one-parameter model of nucleotide substitution ($\alpha=\beta$)

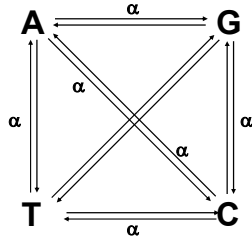


Fig. 11.14
Page 379

Kimura model of nucleotide substitution (assumes $\alpha \neq \beta$)

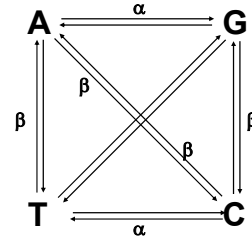


Fig. 11.14
Page 379

Stage 3: Tree-building models: distance

Jukes and Cantor (1969) proposed a corrective formula:

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

Stage 3: Tree-building models: distance

Jukes and Cantor (1969) proposed a corrective formula:

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

Consider an alignment where 3/60 aligned residues differ. The normalized Hamming distance is $3/60 = 0.05$. The Jukes-Cantor correction is

$$D = \left(-\frac{3}{4}\right) \ln \left(1 - \frac{4}{3} 0.05\right) = 0.052$$

When 30/60 aligned residues differ, the Jukes-Cantor correction is more substantial:

$$D = \left(-\frac{3}{4}\right) \ln \left(1 - \frac{4}{3} 0.5\right) = 0.82$$

Use MEGA to display a pairwise distance matrix of 13 globins

(a) number of differences

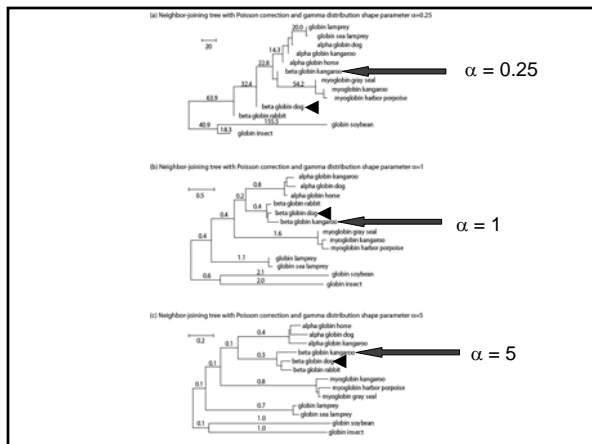
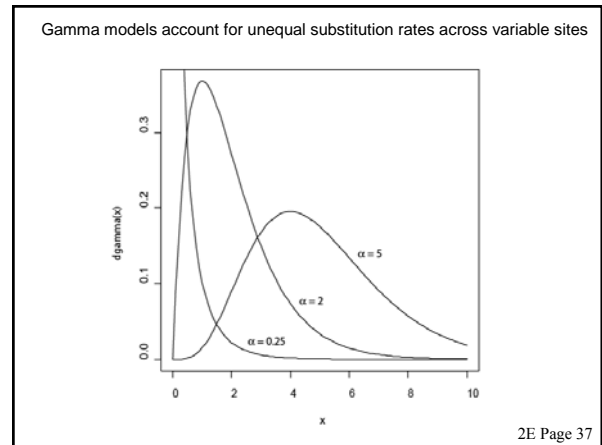
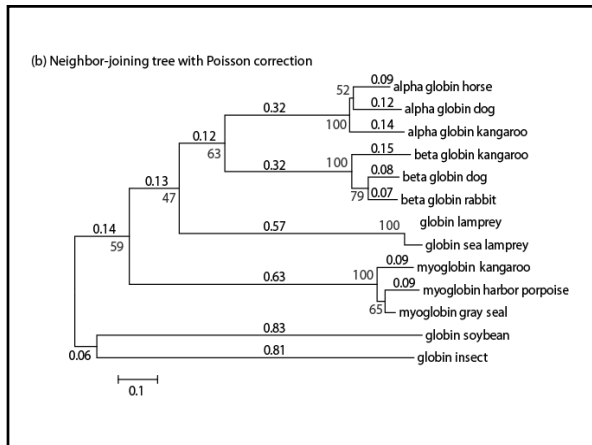
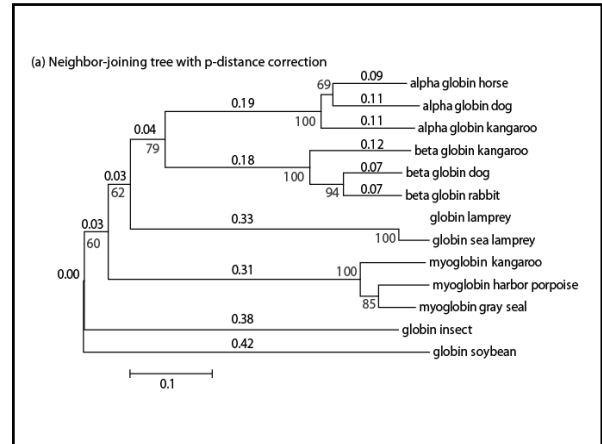
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|
| 1. mbk1angaco P02134 Macropus nullus (red) | | | | | | | | | | | | |
| 2. mbharboi poposa P00278 Phococera pho. | 19 | | | | | | | | | | | |
| 3. infogay isa P00581 Haliastur pringi | 16 | 17 | | | | | | | | | | |
| 4. alphahorse P01958 Equus caballus | 84 | 84 | 84 | | | | | | | | | |
| 5. alphalangaco P01975 Macropus gigante | 85 | 87 | 84 | 24 | | | | | | | | |
| 6. alphadog P00523 Canis lupus familiaris | 86 | 88 | 86 | 22 | 27 | | | | | | | |
| 7. betadog 2P 537902 Canis lupus familiaris | 80 | 79 | 78 | 66 | 69 | 67 | | | | | | |
| 8. betawater NP 001072723 Oryzias latipes | 80 | 81 | 78 | 64 | 67 | 65 | 16 | | | | | |
| 9. betokangaco P02106 Macropus gigante | 83 | 82 | 80 | 68 | 69 | 66 | 26 | 28 | | | | |
| 10. globinlansey E00951A Lampetta fluvia | 86 | 82 | 88 | 77 | 77 | 76 | 83 | 83 | 81 | | | |
| 11. globinsealangay P02208 Petrosyon ma | 83 | 81 | 83 | 76 | 77 | 76 | 83 | 85 | 81 | 8 | | |
| 12. globinsoybean T11674A Glycine max (soy) | 80 | 87 | 87 | 83 | 83 | 87 | 90 | 90 | 93 | 84 | | |
| 13. globinseal P02228 Chironomus fluvi | 87 | 88 | 86 | 82 | 83 | 87 | 90 | 94 | 88 | 89 | 8 | |

(b) p-distance

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|------|------|------|------|------|------|------|------|------|------|------|------|
| 1. mbk1angaco P02134 Macropus nullus (red) | | | | | | | | | | | | |
| 2. mbharboi poposa P00278 Phococera pho. | 0.17 | | | | | | | | | | | |
| 3. infogay isa P00581 Haliastur pringi | 0.14 | 0.11 | | | | | | | | | | |
| 4. alphahorse P01958 Equus caballus | 0.74 | 0.74 | 0.74 | | | | | | | | | |
| 5. alphalangaco P01975 Macropus gigante | 0.75 | 0.77 | 0.74 | 0.21 | | | | | | | | |
| 6. alphadog P00523 Canis lupus familiaris | 0.78 | 0.78 | 0.76 | 0.19 | 0.24 | | | | | | | |
| 7. betadog 2P 537902 Canis lupus familiaris | 0.71 | 0.70 | 0.69 | 0.58 | 0.61 | 0.59 | | | | | | |
| 8. betawater NP 001072723 Oryzias latipes | 0.71 | 0.72 | 0.69 | 0.57 | 0.58 | 0.56 | 0.14 | | | | | |
| 9. betokangaco P02106 Macropus gigante | 0.73 | 0.73 | 0.71 | 0.60 | 0.61 | 0.58 | 0.22 | 0.25 | | | | |
| 10. globinlansey E00951A Lampetta fluvia | 0.78 | 0.81 | 0.78 | 0.68 | 0.68 | 0.67 | 0.73 | 0.73 | 0.72 | | | |
| 11. globinsealangay P02208 Petrosyon ma | 0.79 | 0.81 | 0.79 | 0.67 | 0.68 | 0.67 | 0.73 | 0.75 | 0.72 | 0.07 | | |
| 12. globinsoybean T11674A Glycine max (soy) | 0.67 | 0.66 | 0.66 | 0.62 | 0.62 | 0.62 | 0.77 | 0.80 | 0.80 | 0.62 | 0.83 | |
| 13. globinseal P02228 Chironomus fluvi | 0.77 | 0.78 | 0.76 | 0.61 | 0.62 | 0.62 | 0.81 | 0.80 | 0.83 | 0.76 | 0.79 | 0.81 |

(c) poisson correction

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|--|------|------|------|------|------|------|------|------|------|------|------|------|
| 1. mbl kangaroo P02194 Macropus rufus fwd. | | | | | | | | | | | | |
| 2. mbl harbor porpoise P66278 Phocoena pho. | 0.10 | | | | | | | | | | | |
| 3. mbl gray seal P65901 Halobius glabris | 0.15 | 0.11 | | | | | | | | | | |
| 4. mbl horse P01199 Equus caballus | 1.36 | 1.36 | 1.36 | | | | | | | | | |
| 5. alpha kangaroo P01575 Macropus gigante. | 1.40 | 1.47 | 1.36 | 0.24 | | | | | | | | |
| 6. alpha dog P02529 Canis lupus familiaris | 1.51 | 1.51 | 1.43 | 0.22 | 0.27 | | | | | | | |
| 7. beta dog 2P 517302 Canis lupus familiaris | 1.23 | 1.20 | 1.17 | 0.80 | 0.94 | 0.90 | | | | | | |
| 8. beta sea lamp P01072720 Dyplosidius sp. | 1.23 | 1.26 | 1.17 | 0.64 | 0.90 | 0.96 | 0.15 | | | | | |
| 9. beta kangaroo P02196 Macropus gigante. | 1.33 | 1.29 | 1.23 | 0.92 | 0.94 | 0.89 | 0.26 | 0.28 | | | | |
| 10. globin lamprey S30951A Lampetra fluvia. | 1.51 | 1.68 | 1.51 | 1.14 | 1.14 | 1.12 | 1.33 | 1.33 | 1.26 | | | |
| 11. globin sea lamprey P02208 Petromyzon ma. | 1.55 | 1.64 | 1.55 | 1.12 | 1.14 | 1.12 | 1.33 | 1.40 | 1.26 | 0.07 | | |
| 12. globin soybean 7116748 Glycine max iso | 2.02 | 1.95 | 1.95 | 1.73 | 1.73 | 1.47 | 1.59 | 1.59 | 1.73 | 1.78 | | |
| 13. globin insect P02229 Chironomus thummi. | 1.47 | 1.51 | 1.43 | 1.69 | 1.73 | 1.95 | 1.60 | 1.59 | 1.79 | 1.51 | 1.95 | 1.84 |



Goals of the lecture

Introduction to evolution and phylogeny

Nomenclature of trees

Five stages of molecular phylogeny:

- [1] selecting sequences
- [2] multiple sequence alignment
- [3] models of substitution
- [4] tree-building
- [5] tree evaluation

Stage 4: Tree-building methods

We will discuss two tree-building methods: distance-based and character-based.

Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score. Examples of distance-based algorithms are UPGMA and neighbor-joining.

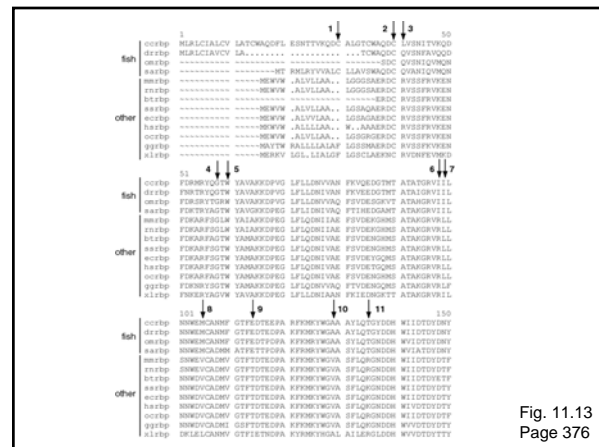
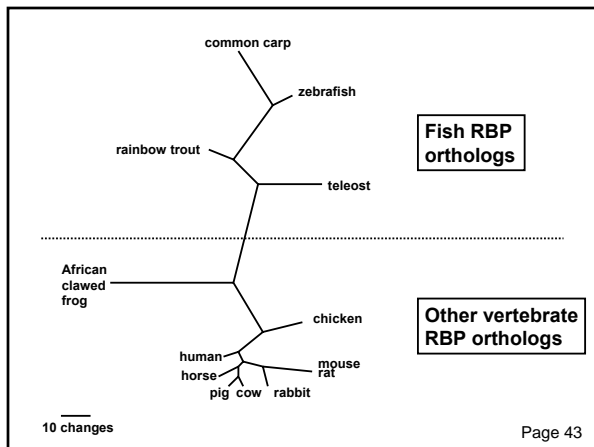
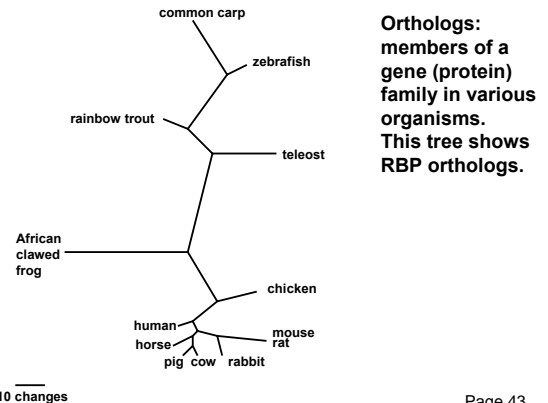
Stage 4: Tree-building methods

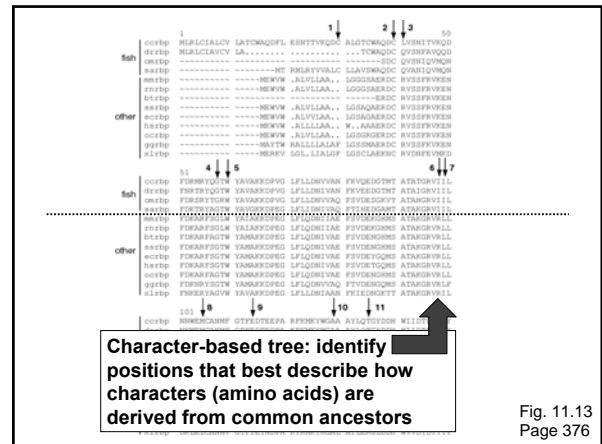
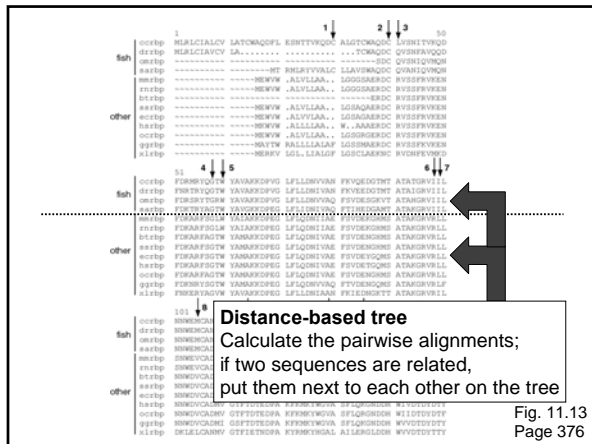
Distance-based methods involve a distance metric, such as the number of amino acid changes between the sequences, or a distance score. Examples of distance-based algorithms are UPGMA and neighbor-joining.

Character-based methods include maximum parsimony and maximum likelihood. Parsimony analysis involves the search for the tree with the fewest amino acid (or nucleotide) changes that account for the observed differences between taxa.

Stage 4: Tree-building methods

We can introduce distance-based and character-based tree-building methods by referring to a tree of 13 orthologous retinol-binding proteins, and the multiple sequence alignment from which the tree was generated.





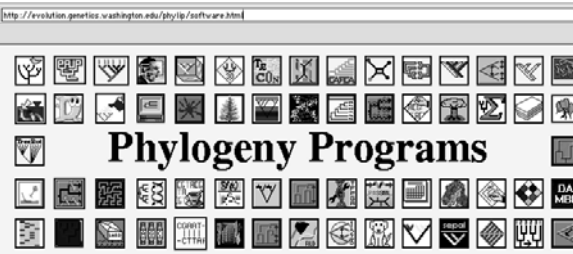
Stage 4: Tree-building methods

Regardless of whether you use distance- or character-based methods for building a tree, the starting point is a multiple sequence alignment.

ReadSeq is a convenient web-based program that translates multiple sequence alignments into formats compatible with most commonly used phylogeny programs such as PAUP and PHYLIP.

Page 378

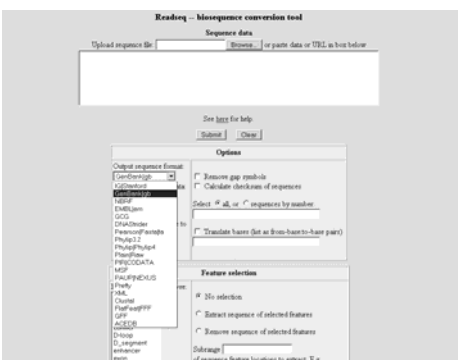
<http://evolution.genetics.washington.edu/phylip/software.html>



Phylogeny Programs

This site lists 200 phylogeny packages. Perhaps the best-known programs are PAUP (David Swofford and colleagues) and PHYLIP (Joe Felsenstein).

ReadSeq is widely available; try the "tools" menu at the LANL HIV database



ReadSeq - biosequence conversion tool

Sequence data

Upload sequence data: or paste data or URL in box below

See help for help

Options

Output sequence format: fasta fasta2 fasta3 fasta4 fasta5 fasta6 fasta7 fasta8 fasta9 fasta10 fasta11 fasta12 fasta13 fasta14 fasta15 fasta16 fasta17 fasta18 fasta19 fasta20 fasta21 fasta22 fasta23 fasta24 fasta25 fasta26 fasta27 fasta28 fasta29 fasta30 fasta31 fasta32 fasta33 fasta34 fasta35 fasta36 fasta37 fasta38 fasta39 fasta40 fasta41 fasta42 fasta43 fasta44 fasta45 fasta46 fasta47 fasta48 fasta49 fasta50 fasta51 fasta52 fasta53 fasta54 fasta55 fasta56 fasta57 fasta58 fasta59 fasta60 fasta61 fasta62 fasta63 fasta64 fasta65 fasta66 fasta67 fasta68 fasta69 fasta70 fasta71 fasta72 fasta73 fasta74 fasta75 fasta76 fasta77 fasta78 fasta79 fasta80 fasta81 fasta82 fasta83 fasta84 fasta85 fasta86 fasta87 fasta88 fasta89 fasta90 fasta91 fasta92 fasta93 fasta94 fasta95 fasta96 fasta97 fasta98 fasta99 fasta100

Feature selection

No selection

Extract sequence of selected features

Extract sequence of selected features

Subrange: of sequence feature locations to extract. E.g., 100-1000

- ### Stage 4: Tree-building methods
- [1] distance-based
 - [2] character-based: maximum parsimony
 - [3] character- and model-based: maximum likelihood
 - [4] character- and model-based: Bayesian

Stage 4: Tree-building methods: distance

Many software packages are available for making phylogenetic trees.

Page 379

Stage 4: Tree-building methods: distance

Many software packages are available for making phylogenetic trees. We will describe two programs.

[1] MEGA (Molecular Evolutionary Genetics Analysis) by Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. Download it from <http://www.megasoftware.net/>

[2] Phylogeny Analysis Using Parsimony (PAUP), written by David Swofford. See <http://paup.csit.fsu.edu/>.

We will next use MEGA and PAUP to generate trees by the distance-based method UPGMA.

Page 379

How to use MEGA to make a tree

- [1] Enter a multiple sequence alignment (.meg) file
- [2] Under the phylogeny menu, select one of these four methods...

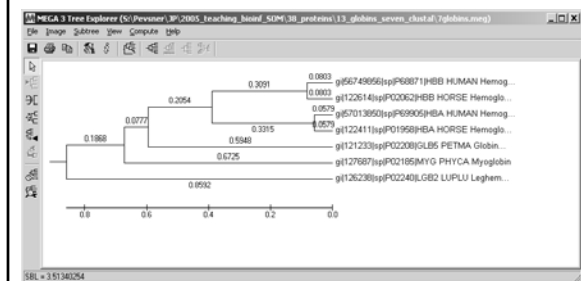
Neighbor-Joining (NJ)
Minimum Evolution (ME)
Maximum Parsimony (MP)
UPGMA

Use of MEGA for a distance-based tree: UPGMA

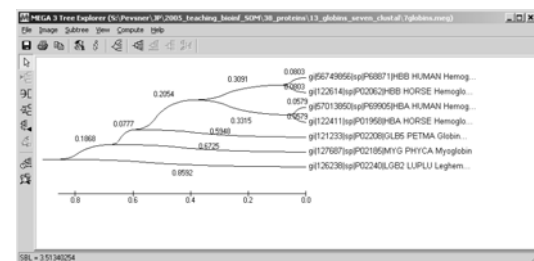
Click green boxes to obtain options

Click compute to obtain tree

Use of MEGA for a distance-based tree: UPGMA



Use of MEGA for a distance-based tree: UPGMA



A variety of styles are available for tree display

Tree-building methods: UPGMA

Step 4: Last cluster! This is your tree.

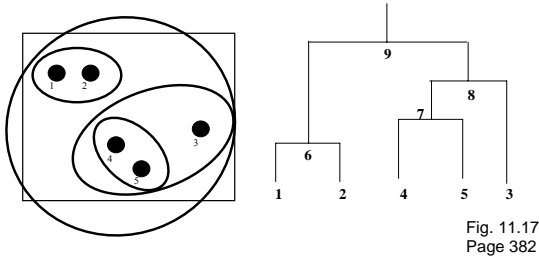


Fig. 11.17
Page 382

Distance-based methods: UPGMA trees

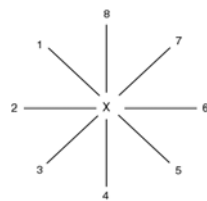
UPGMA is a simple approach for making trees.

- An UPGMA tree is always rooted.
- An assumption of the algorithm is that the molecular clock is constant for sequences in the tree. If there are unequal substitution rates, the tree may be wrong.
- While UPGMA is simple, it is less accurate than the neighbor-joining approach (described next).

Page 383

Making trees using neighbor-joining

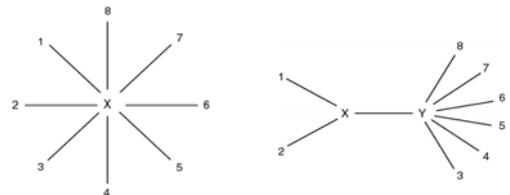
The neighbor-joining method of Saitou and Nei (1987) is especially useful for making a tree having a large number of taxa.



Begin by placing all the taxa in a star-like structure.

Page 383

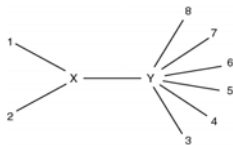
Tree-building methods: Neighbor joining



Next, identify neighbors (e.g. 1 and 2) that are most closely related. Connect these neighbors to other OTUs via an internal branch, XY. At each successive stage, minimize the sum of the branch lengths.

Fig. 11.18
Page 384

Tree-building methods: Neighbor joining

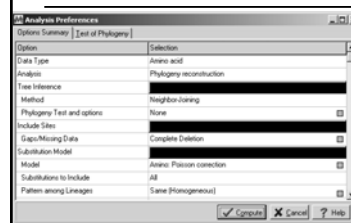


Define the distance from X to Y by

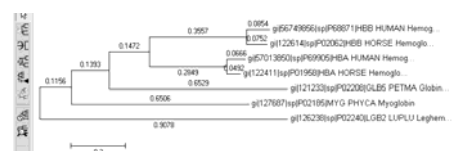
$$d_{XY} = 1/2(d_{1Y} + d_{2Y} - d_{12})$$

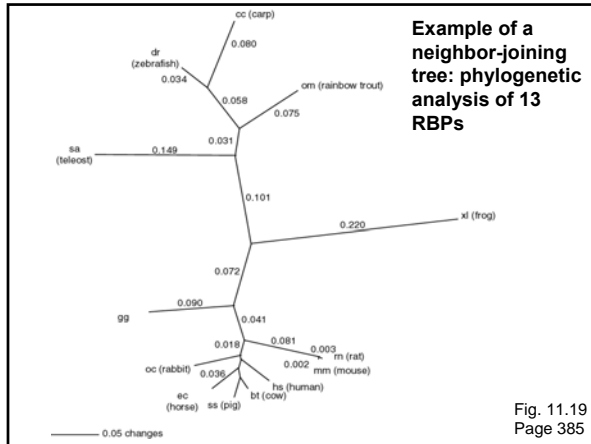
Fig. 11.18
Page 384

Use of MEGA for a distance-based tree: NJ



Neighbor Joining produces a reasonably similar tree as UPGMA





Stage 4: Tree-building methods

We will discuss four tree-building methods:

[1] distance-based

[2] character-based: maximum parsimony

[3] character- and model-based: maximum likelihood

[4] character- and model-based: Bayesian

Tree-building methods: character based

Rather than pairwise distances between proteins, evaluate the aligned columns of amino acid residues (characters).

Tree-building methods based on characters include maximum parsimony and maximum likelihood.

Page 383

Making trees using character-based methods

The main idea of character-based methods is to find the tree with the shortest branch lengths possible. Thus we seek the most parsimonious ("simple") tree.

- Identify informative sites. For example, constant characters are not parsimony-informative.
- Construct trees, counting the number of changes required to create each tree. For about 12 taxa or fewer, evaluate all possible trees exhaustively; for >12 taxa perform a heuristic search.
- Select the shortest tree (or trees).

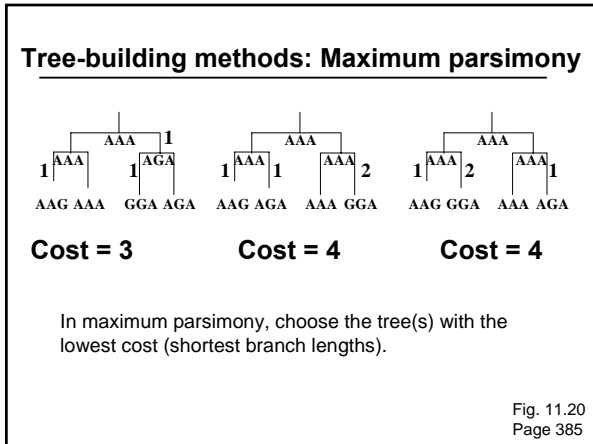
Page 383

As an example of tree-building using maximum parsimony, consider these four taxa:

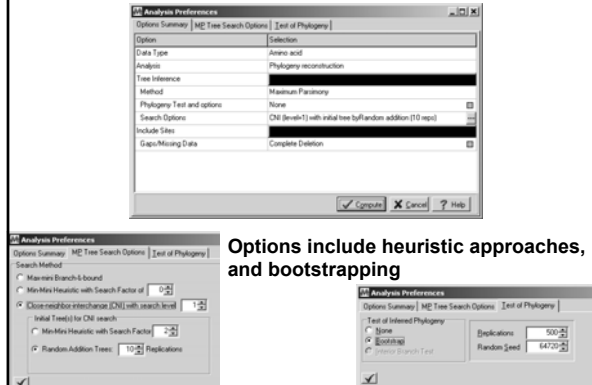
AAG
AAA
GGA
AGA

How might they have evolved from a common ancestor such as AAA?

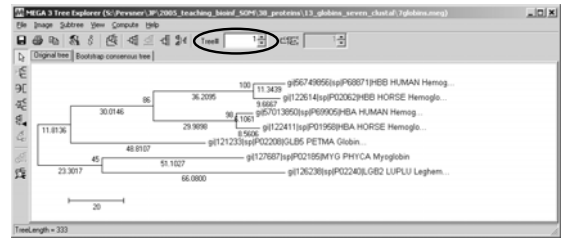
Fig. 11.20
Page 385



MEGA for maximum parsimony (MP) trees



MEGA for maximum parsimony (MP) trees



In maximum parsimony, there may be more than one tree having the lowest total branch length. You may compute the consensus best tree.

Phylogram

(values are proportional to branch lengths)

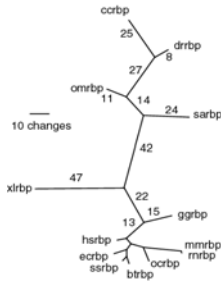


Fig. 11.22
Page 387

Rectangular phylogram

(values are proportional to branch lengths)

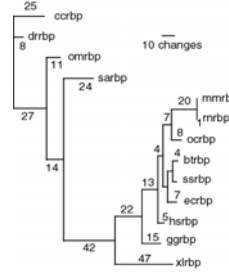


Fig. 11.22
Page 387

Cladogram

(values are not proportional to branch lengths)

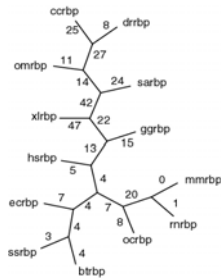
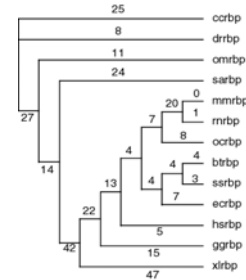


Fig. 11.22
Page 387

Rectangular cladogram

(values are not proportional to branch lengths)



These four trees display the same data in different formats.

Fig. 11.22
Page 387

Stage 4: Tree-building methods

We will discuss four tree-building methods:

- [1] distance-based
- [2] character-based: maximum parsimony
- [3] character- and model-based: maximum likelihood
- [4] character- and model-based: Bayesian

Making trees using maximum likelihood

Maximum likelihood is an alternative to maximum parsimony. It is computationally intensive. A likelihood is calculated for the probability of each residue in an alignment, based upon some model of the substitution process.

What are the tree topology and branch lengths that have the greatest likelihood of producing the observed data set?

ML is implemented in the TREE-PUZZLE program, as well as PAUP and PHYLIP.

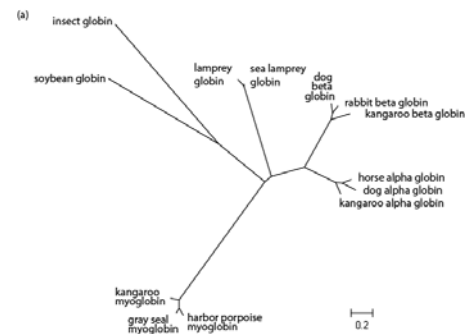
Page 386

Maximum likelihood: Tree-Puzzle

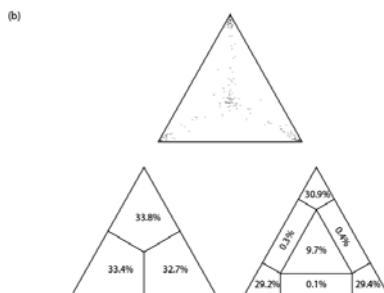
(1) Reconstruct all possible quartets A, B, C, D.
For 12 myoglobins there are 495 possible quartets.

(2) Puzzling step: begin with one quartet tree. N-4 sequences remain. Add them to the branches systematically, estimating the support for each internal branch. Report a consensus tree.

Maximum likelihood tree



Quartet puzzling



Stage 4: Tree-building methods

We will discuss four tree-building methods:

- [1] distance-based
- [2] character-based: maximum parsimony
- [3] character- and model-based: maximum likelihood
- [4] character- and model-based: Bayesian

