Molecular Phylogeny and Evolution

Alejandro Giorgetti

from Bioinformatics Pevsner
Five kingdom system (Haeckel, 1879)

- Mammals
- Vertebrates
- Invertebrates
- Protozoa
- Monera
- Plants
- Fungi
- Protists
- Animals
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
  [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.
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)
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.
<table>
<thead>
<tr>
<th></th>
<th>signal peptide</th>
<th>B chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>cow</td>
<td>MALWTRLRPLLALLALAWPPPPARA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>sheep</td>
<td>MALWTRLVPLLALLALAWPAPAH</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>pig</td>
<td>MALWTRLLPLLALLALAWAPAPAQA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>human</td>
<td>MALWMLRLLPLLALLALWGDPAAA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKT</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>MALWMLRLLPLLALLALWGDPASA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKT</td>
</tr>
<tr>
<td>dog</td>
<td>MALWMRLPLLALLALAWAPAPTRA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>rat</td>
<td>MALWMRFPLLALLALWEPKPAQA</td>
<td>FVKNQHLCGPHLVEALYLVCGERGFFYTPKS</td>
</tr>
<tr>
<td>mouse</td>
<td>MALLVHFLPLLALLALWEPKPTQA</td>
<td>FVKNQHLCGPHLVEALYLVCGERGFFYTPKS</td>
</tr>
<tr>
<td>rabbit</td>
<td>MASLAALLPLLALLVLCLRPAPAQA</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>sperm whale</td>
<td>-------------------------------------------------</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKA</td>
</tr>
<tr>
<td>elephant</td>
<td>-------------------------------------------------</td>
<td>FVNQHLCGSHLVEALYLVCGERGFFYTPKT</td>
</tr>
<tr>
<td>chicken</td>
<td>MALWIRSLPLLALLVFSPGPTSYA</td>
<td>AANQHLCGSHLVEALYLVCGERGFFYSPKA</td>
</tr>
</tbody>
</table>

|        | C peptide                             | A chain                             |
| cow    | RREVEGPQVGAELAGAGPG--AGGLEGPQKKRGIVEQCCASVCSLYQLENYCN |
| sheep  | RREVEGPQVGAELAGAGPG--AGGLEGPQKKRGIVEQCCAGVCSLYQLENYCN |
| pig    | RREAENPQAGAVELGGGLGG--LQALALEGPPQQKRGIVEQCCCTSICSLYQLENYCN |
| human  | RREAELQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCCTSICSLYQLENYCN |
| chimpanzee | RREAELQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCCTSICSLYQLENYCN |
| dog    | RREVEDLQVDVELAGAPEGGGLQPLALEGALQKRGIVEQCCCTSICSLYQLENYCN |
| rat    | RREVEDPQVPELLEGGPGPEAGDLQTLALEVARQKRGIVEQCCCTSICSLYQLENYCN |
| mouse  | RREVEDPQVQELEGSLGSP--GDLQTLALEVARQKRGIVEQCCCTSICSLYQLENYCN |
| rabbit | RREVEELQVGAELGGPGAGGLQPSALELALQKRGIVEQCCCTSICSLYQLENYCN |
| sperm whale | ------------------------------------------------- | GIVEQCCCTSICSLYQLENYCN             |
| elephant | ------------------------------------------------- | GIVEQCCGTGCSLYQLENYCN              |
| chicken | RRDVEQPLVSS--PLRGEAG--VLPFQQQEYEYKVRGIVEQCCCHINTCSLYQLENYCN |
Note the sequence divergence in the disulfide loop region of the A chain.
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.
<table>
<thead>
<tr>
<th>Species</th>
<th>Signal Peptide</th>
<th>B Chain</th>
<th>C Peptide</th>
<th>A Chain</th>
</tr>
</thead>
<tbody>
<tr>
<td>cow</td>
<td>MALWTRLRPLLALLALWPPPAPA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEGPQVGAELEGPPQKR</td>
<td>GIVEQCCASVCSLYQLENYCN</td>
</tr>
<tr>
<td>sheep</td>
<td>MALWTRLVPLLALLALWAPAPAHA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEGPQVGAELEGPPQKR</td>
<td>GIVEQCCAGVCSLYQLENYCN</td>
</tr>
<tr>
<td>pig</td>
<td>MALWTRLLLPLLALLALWAPAPAQA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREAEENLQVQVGAELEGPPQKR</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>human</td>
<td>MALWRMRLLLPLLALLALWGPDPAAA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREAEQLQVQVQVELEGPPQKR</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>MALWRMRLPLLALLALWAPAPTRA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEDLQVRDVEALGAGGLQPLALEGALQKR</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>dog</td>
<td>MALWMRLLPLLALLALWAPAPTRA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEELQVQSALEGPAQLQKRGIV</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>rat</td>
<td>MALWMRFLPLLALLVLWEPKPAQA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEDPQVPQLELGPGPEAGDLLQTLALEVARQKRGIV</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>mouse</td>
<td>MALLVHFLPLLALLALWEPKPTQA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEVQPEOLQLEGSP---</td>
<td>GDLQTLALEVARQKRGIVDQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>rabbit</td>
<td>MASLAALLPLLALLVLCLRPAPQA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYTPKA</td>
<td>RREVEELQVQSALEGPAQLQKRGIV</td>
<td>GIVEQCCCTSCIYQLENYCN</td>
</tr>
<tr>
<td>sperm whale</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>elephant</td>
<td>-------------------------</td>
<td>------------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>chicken</td>
<td>MAlWIRSLPLLALLLVPSGPGTSA</td>
<td>FVNQHLCGSHELVEALYLVCGERGFFYSPKA</td>
<td>RRDVEQPLVSS---PLRGEAG---</td>
<td>GIVEQCCCHINTCSLYQLENYCN</td>
</tr>
</tbody>
</table>

Number of nucleotide substitutions/site/year

- 0.1 x 10^-9
- 1 x 10^-9
- 0.1 x 10^-9
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.

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Mouse</th>
<th>Guinea Pig</th>
</tr>
</thead>
<tbody>
<tr>
<td>A chain</td>
<td>MALWMRLLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT</td>
<td>MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKS</td>
<td>MALWHMLLTVLALLALWGPNTGQAFVSRLHLCGSNLVETLYSVCQDDGFFYIPKD</td>
</tr>
<tr>
<td>B chain</td>
<td>RREAEDLQVGQVELGGPGAGSLQPLALEGLQKRIGIVEQCCTSICSQLYQLENYCNR</td>
<td>RREVEDPQVEQLELGGS - GDLQTLALERVARQKRIGVDQCTSICSQLYQLENYCNC</td>
<td>RRELEDPQVEQTELGMGLGAGGLQPLALEMALQKRIGVDQCCTGTCTRQLQSYCN</td>
</tr>
</tbody>
</table>

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

Millions of years since divergence

Dickerson (1971)
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: 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.
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.
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
[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.
• Terminology

- **External nodes**: things under comparison; operational taxonomic units (OTUs)
- **Internal nodes**: ancestral units; hypothetical; goal is to group current day units
- **Root**: common ancestor of all OTUs under study. Path from root to node defines evolutionary path
- **Unrooted**: specify relationship but not evolutionary path
  – If have an **outgroup** (external reason to believe certain OTU branched off first), then can root
- **Topology**: branching pattern of a tree
- **Branch length**: amount of difference that occurred along a branch
Enumerating trees

Cavalii-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 $\approx 34$ million, and the number of unrooted trees is $\approx 2$ million. Many tree-making algorithms can exhaustively examine every possible tree for up to ten to twelve sequences.
### Numbers of trees

<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
<td>105</td>
</tr>
<tr>
<td>20</td>
<td>$8 \times 10^{21}$</td>
<td>$2 \times 10^{20}$</td>
</tr>
</tbody>
</table>
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 1

Species 2

Past

Present

Speciation event
Species trees versus gene/protein trees

Gene duplication events

speciation event

species 1

species 2
Species trees versus gene/protein trees

Gene duplication events

species 1

species 2

speciation event

OTUs
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
[5] tree evaluation
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.
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.
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.
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.
<table>
<thead>
<tr>
<th>ancestral RBP (hypothetical)</th>
<th>human RBP (observed)</th>
<th>equine RBP (observed)</th>
<th>alignment (observed)</th>
<th>substitution mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>CC</td>
<td>single substitution</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>TG</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>CA→G</td>
<td>sequential substitution</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>TT</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A→T</td>
<td>A→C</td>
<td>TC</td>
<td>coincidental substitutions</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A→G</td>
<td>A→G</td>
<td>GG</td>
<td>parallel substitutions</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>C→G</td>
<td>C→T→G</td>
<td>GG</td>
<td>convergent substitutions</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>G</td>
<td>GG</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>G→A→G</td>
<td>G</td>
<td>GG</td>
<td>back substitution</td>
</tr>
</tbody>
</table>
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.
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
MEGA outputs transition and transversion frequencies

|       | T | C | G | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C |
| human (Homo sapiens) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| bonobo (Pan paniscus) | T | C | G | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C | A | T | C |
| chimpanzee (Pan troglodytes) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| gorilla (Gorilla gorilla) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| orangutan (Pongo pygmaeus) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| gibbon (Hylobates lar) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
MEGA outputs transition and transversion frequencies

For primate mitochondrial DNA, the ratio of transitions to transversions is particularly high.

For primate mitochondrial DNA, the ratio of transitions to transversions is particularly high.
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
[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.
Some positions of the multiple sequence alignment are invariant (arrow 2). Some positions distinguish fish RBP from all other RBPs (arrow 3).
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.
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
[5] tree evaluation
Computing Distances Between Sequences

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

Sequence 2: A A T G A A A G A A T C G C

Could compute fraction of mismatches between two sequences; however, this is an underestimate of actual distance
Computing Distances Between Sequences

E.g., many underlying substitutions possible

Use models of substitution to correct these values
Computing Distances Between Sequences

Jukes & Cantor model
- Each position in DNA sequence is independent
- Each position can mutate with same probability to any another base

Correction to observed substitution rate (see notes):

$$-0.75 \left( \ln \left(1 - \frac{4}{3} \left( \frac{\text{observed \# \ differences}}{\text{length}} \right) \right) \right)$$
Ex: Computing Distances Between Sequences

- Alignment of two DNA sequences
  - Length of alignment (non gapped positions): 100
  - Number of differences: 25

Naïve distance calculation = $25/100 = \frac{1}{4}$

Correction

$$-0.75 \ln \left(1 - \frac{4}{3}(1/4)\right) = -0.75 \ln \left(\frac{2}{3}\right) = .304$$

- Other models for DNA, also protein (e.g., PAM)
Use MEGA to display a pairwise distance matrix of 13 globins

<table>
<thead>
<tr>
<th>(a) number of differences</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. mbkangaroo P02194 Macropus rufus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. mbhabor porpoise P68278 Phocoena pho...</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. mbgray seal P68081 Halichoerus grypus</td>
<td>16</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. alphahorse P01958 Equus caballus</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. alphakangaroo P01975 Macropus gigante...</td>
<td>85</td>
<td>87</td>
<td>84</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. alphadog P60529 Canis lupus familiar...</td>
<td>88</td>
<td>88</td>
<td>86</td>
<td>22</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. betadog XP 537902 Canis lupus familia...</td>
<td>80</td>
<td>79</td>
<td>78</td>
<td>66</td>
<td>69</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. betarabbit NP 001075729 Oryctolagus c...</td>
<td>80</td>
<td>81</td>
<td>78</td>
<td>64</td>
<td>67</td>
<td>65</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. betakangaroo P02106 Macropus giganteu...</td>
<td>83</td>
<td>82</td>
<td>80</td>
<td>68</td>
<td>69</td>
<td>66</td>
<td>25</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. globinlamprey 690951A Lampetra fluvia...</td>
<td>88</td>
<td>92</td>
<td>88</td>
<td>77</td>
<td>77</td>
<td>76</td>
<td>83</td>
<td>83</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. globinsealamprey P02208 Petromyzon ma...</td>
<td>89</td>
<td>91</td>
<td>89</td>
<td>76</td>
<td>77</td>
<td>76</td>
<td>83</td>
<td>85</td>
<td>81</td>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. globinsoybean 711674A Glycine max [so...</td>
<td>98</td>
<td>97</td>
<td>97</td>
<td>93</td>
<td>93</td>
<td>93</td>
<td>87</td>
<td>90</td>
<td>90</td>
<td>93</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>13. globininsect P02229 Chironomus thummi...</td>
<td>87</td>
<td>88</td>
<td>86</td>
<td>92</td>
<td>93</td>
<td>97</td>
<td>92</td>
<td>90</td>
<td>94</td>
<td>88</td>
<td>89</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>mbkangaroo P02194 Macropus rufus (red...)</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mbharbor porpoise P68278 Phocoena pho...</td>
<td>0.14</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mbgray seal P68081 Halichoerus grypus</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alphahorse P01958 Equus caballus</td>
<td>0.75</td>
<td>0.77</td>
<td>0.74</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alphakangaroo P01975 Macropus gigante...</td>
<td>0.78</td>
<td>0.78</td>
<td>0.76</td>
<td>0.19</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alphadog P60529 Canis lupus familiar...</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.58</td>
<td>0.61</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>betadog XP 537902 Canis lupus familia...</td>
<td>0.71</td>
<td>0.72</td>
<td>0.69</td>
<td>0.57</td>
<td>0.59</td>
<td>0.58</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>betarabbit NP 001075729 Oryctolagus c...</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
<td>0.60</td>
<td>0.61</td>
<td>0.58</td>
<td>0.22</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>betakangaroo P02106 Macropus giganteu...</td>
<td>0.78</td>
<td>0.81</td>
<td>0.78</td>
<td>0.68</td>
<td>0.68</td>
<td>0.67</td>
<td>0.73</td>
<td>0.73</td>
<td>0.72</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>globinlamprey 690951A Lampetra fluvia...</td>
<td>0.97</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.77</td>
<td>0.80</td>
<td>0.80</td>
<td>0.82</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>globinsealamprey P02208 Petromyzon ma...</td>
<td>0.77</td>
<td>0.78</td>
<td>0.76</td>
<td>0.81</td>
<td>0.82</td>
<td>0.86</td>
<td>0.81</td>
<td>0.80</td>
<td>0.83</td>
<td>0.78</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>(c) poisson correction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>mbkangaroo P02194 Macropus rufus (red...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>mbbharbor porpoise P68278 Phocoena pho...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>mbgray seal P68081 Halichoerus grypus</td>
<td>0.15 0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>alphahorse P01958 Equus caballus</td>
<td>1.36 1.36 1.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>alphakangaroo P01975 Macropus gigante...</td>
<td>1.40 1.47 1.36 0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>alphadog P60529 Canis lupus familiar...</td>
<td>1.51 1.51 1.43 0.22 0.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>betadog XP 537902 Canis lupus familiar...</td>
<td>1.23 1.20 1.17 0.88 0.94 0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>betarabbit NP 001075729 Oryctolagus c...</td>
<td>1.23 1.26 1.17 0.84 0.90 0.86 0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>betakangaroo P02106 Macropus giganteu...</td>
<td>1.33 1.29 1.23 0.92 0.94 0.88 0.25 0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>globinlamprey 690951A Lampetra fluvia...</td>
<td>1.51 1.68 1.51 1.14 1.14 1.12 1.33 1.33 1.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>globinsalamprey P02208 Petromyzon ma...</td>
<td>1.55 1.64 1.55 1.12 1.14 1.12 1.33 1.40 1.26 0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>globinsoybean 711674A Glycine max [so...</td>
<td>2.02 1.95 1.95 1.73 1.73 1.73 1.47 1.59 1.59 1.73 1.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>globininsect P02229 Chironomus thummi...</td>
<td>1.47 1.51 1.43 1.68 1.73 1.95 1.68 1.59 1.78 1.51 1.55 1.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(a) Neighbor-joining tree with p-distance correction

- Alpha globin horse
- Alpha globin dog
- Alpha globin kangaroo
- Beta globin kangaroo
- Beta globin dog
- Beta globin rabbit
- Globin lamprey
- Globin sea lamprey
- Myoglobin kangaroo
- Myoglobin harbor porpoise
- Myoglobin gray seal
- Globin insect
- Globin soybean
(b) Neighbor-joining tree with Poisson correction

- alpha globin horse
- alpha globin dog
- alpha globin kangaroo
- beta globin kangaroo
- beta globin dog
- beta globin rabbit
- globin lamprey
- globin sea lamprey
- myoglobin kangaroo
- myoglobin harbor porpoise
- myoglobin gray seal
- globin soybean
- globin insect
Gamma models account for unequal substitution rates across variable sites.
(a) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter $\alpha=0.25$

$\alpha = 0.25$

(b) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter $\alpha=1$

$\alpha = 1$

(c) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter $\alpha=5$

$\alpha = 5$
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.
Orthologs: members of a gene (protein) family in various organisms. This tree shows RBP orthologs.
Distance-based tree
Calculate the pairwise alignments; if two sequences are related, put them next to each other on the tree
Character-based tree: identify positions that best describe how characters (amino acids) are derived from common ancestors.
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.
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.
Stage 4: Tree-building methods

[1] distance-based


Stage 4: Tree-building methods: distance

Many software packages are available for making phylogenetic trees.
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/


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

Click green boxes to obtain options
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
Use of MEGA for a distance-based tree: UPGMA

Flipping branches around a node creates an equivalent topology
Tree-building methods: UPGMA

UPGMA is an unweighted pair group method using arithmetic mean.
Tree-building methods: UPGMA

Step 1: compute the pairwise distances of all the proteins. Get ready to put the numbers 1-5 at the bottom of your new tree.
Tree-building methods: UPGMA

Step 2: Find the two proteins with the smallest pairwise distance. Cluster them.
Tree-building methods: UPGMA

Step 3: Do it again. Find the next two proteins with the smallest pairwise distance. Cluster them.
Tree-building methods: UPGMA

Step 4: Keep going. Cluster.
Tree-building methods: UPGMA

Step 4: Last cluster! This is your tree.
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).
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.
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.
Tree-building methods: Neighbor joining

Define the distance from X to Y by

\[ d_{XY} = \frac{1}{2}(d_{1Y} + d_{2Y} - d_{12}) \]
Use of MEGA for a distance-based tree: NJ

Neighbor Joining produces a reasonably similar tree as UPGMA
Example of a neighbor-joining tree: phylogenetic analysis of 13 RBPs
Stage 4: Tree-building methods

We will discuss four tree-building methods:

[1] distance-based


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.
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).
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?
Tree-building methods: Maximum parsimony

In maximum parsimony, choose the tree(s) with the lowest cost (shortest branch lengths).
MEGA for maximum parsimony (MP) trees

Options include heuristic approaches, and bootstrapping
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)
Rectangular phylogram

(values are proportional to branch lengths)
Cladogram
(values are not proportional to branch lengths)
Rectangular cladogram

(values are not proportional to branch lengths)

These four trees display the same data in different formats.
Stage 4: Tree-building methods

We will discuss four tree-building methods:

[1] distance-based


Maximum Likelihood

- Given a probabilistic model for nucleotide (or protein) substitution (e.g., Jukes & Cantor), pick the tree that has highest probability of generating observed data
  - i.e., Given data $D$ and model $M$, find tree $T$ such that $Pr(D/T, M)$ is maximized
- Models gives values $p_{ij}(t)$, the probability of going from nucleotide $i$ to $j$ in time $t$
Maximum Likelihood

- Makes 2 independence assumptions
  - Different sites evolve independently
  - Diverged sequences (or species) evolve independently after diverging
- If $D_i$ is data for $i$th site
  \[
  Pr(D|T, M) = \prod_i Pr(D_i|T, M)
  \]
Maximum Likelihood

How to calculate $Pr(D_i/T,M)$?

$p_{xy}(t) \sim \text{prob of going from } x \text{ to } y \text{ in time } t$

$$Pr(i, j, k, l|T, M) = \sum_x \sum_y \sum_z pr(x)(p_{xl} \cdot (t_1 + t_2 + t_3) \cdot p_{xy}(t_1) \\
\cdot p_{yk}(t_2 + t_3) \cdot p_{yz}(t_2) \cdot p_{zi}(t_3) \cdot p_{zj}(t_3))$$
Maximum Likelihood

• Given tree topology and branch lengths, can efficiently calculate \( Pr(D/T, M) \) using dynamic programming
  – I.e., don’t have to enumerate over all internal states

• Finding best maximum likelihood tree is expensive
  – Must consider all topologies
  – Find best edge lengths for each topology
    • Idea: use some search procedure, e.g., EM, to optimize these lengths
We will discuss four tree-building methods:

[1] distance-based


Bayesian inference of phylogeny with MrBayes

Calculate:

\[
Pr [ \text{Tree} | \text{Data} ] = \frac{Pr [ \text{Data} | \text{Tree} ] \times Pr [ \text{Tree} ]}{Pr [ \text{Data} ]}
\]

\( Pr [ \text{Tree} | \text{Data} ] \) is the posterior probability distribution of trees. Ideally this involves a summation over all possible trees. In practice, Monte Carlo Markov Chains (MCMC) are run to estimate the posterior probability distribution.

Notably, Bayesian approaches require you to specify prior assumptions about the model of evolution.
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 5: Evaluating trees

The main criteria by which the accuracy of a phylogenetic tree is assessed are consistency, efficiency, and robustness. Evaluation of accuracy can refer to an approach (e.g. UPGMA) or to a particular tree.
Stage 5: Evaluating trees: bootstrapping

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?
Stage 5: Evaluating trees: bootstrapping

Bootstrapping is a commonly used approach to measuring the robustness of a tree topology. Given a branching order, how consistently does an algorithm find that branching order in a randomly permuted version of the original data set?

To bootstrap, make an artificial dataset obtained by randomly sampling columns from your multiple sequence alignment. Make the dataset the same size as the original. Do 100 (to 1,000) bootstrap replicates. Observe the percent of cases in which the assignment of clades in the original tree is supported by the bootstrap replicates. >70% is considered significant.
MEGA for maximum parsimony (MP) trees

Bootstrap values show the percent of times each clade is supported after a large number (n=500) of replicate samplings of the data.
In 61% of the bootstrap resamplings, ssrbp and btrbp (pig and cow RBP) formed a distinct clade. In 39% of the cases, another protein joined the clade (e.g. ecrbp), or one of these two sequences joined another clade.
Difference in Methods

• Maximum-likelihood and parsimony methods have models of evolution
• Distance methods do not necessarily
  – Useful aspect in some circumstances
    • E.g., trees built based on whole genomes, presence or absence of genes
• Religious wars over which methods to use
  – Most people now believe ML based methods are best: most sensitive at large evolutionary distances – but also most time-consuming & depend on specific model of evolution used
• Most commonly used packages contain software for all three methods: may want to use more than 1 to have confidence in built tree