Molecular evolution

Joe Felsenstein

GENOME 453, Autumn 2011
A data example for phylogeny inference

Five DNA sequences, for some gene in an imaginary group of species whose names are Alpha, Beta, Gamma, Delta, and Epsilon:

<table>
<thead>
<tr>
<th>species</th>
<th>site</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Alpha</td>
<td>A</td>
</tr>
<tr>
<td>Beta</td>
<td>C</td>
</tr>
<tr>
<td>Gamma</td>
<td>A</td>
</tr>
<tr>
<td>Delta</td>
<td>A</td>
</tr>
<tr>
<td>Epsilon</td>
<td>C</td>
</tr>
</tbody>
</table>
The tree we are evaluating

Alpha    Delta    Gamma    Beta    Epsilon
Steps in character 1

1 2 3 4 5 6

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T G A G C</td>
<td>C T C T A C</td>
<td>A G G T A C</td>
<td>A G G A G T</td>
<td>C T C A G C</td>
</tr>
</tbody>
</table>

or

A A A A C C

Alpha Delta Gamma Beta Epsilon
Steps in character 2

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>Beta</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Gamma</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Delta</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Epsilon</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

Molecular evolution – p.5/62
Steps in character 3

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>Beta</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Gamma</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Delta</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Epsilon</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

Molecular evolution – p.6/62
Steps in character 4

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
</tbody>
</table>

**or**

![Molecular evolution](attachment://image.png)
Steps in character 5

### Alpha Beta Gamma Delta Epsilon

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Beta</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>5</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

Or

A T G A
G
C
C T C A
G
C
A G G A
G
T
C T C A
G
C
A G G T
A
C
A G G A
G
T
C T C A
G
C
A G G T
A
C
A G G A
G
T
C T C A
G
C
Steps in character 6

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>Beta</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Gamma</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Delta</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Epsilon</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

![Diagram of molecular evolution with characters A, T, G, and C at different positions across species Alpha, Delta, Gamma, Beta, and Epsilon.](image-url)
Steps in all characters

showing one of their possible placements
The most parsimonious tree

with one possible placement of the changes
The same tree as an unrooted tree

shown as an unrooted tree
root can be anywhere
changes can occur in either direction
Direction of change depends on where root is

Placement of the root affects which way bases change but not how many changes there are

<table>
<thead>
<tr>
<th></th>
<th>Alpha</th>
<th>Gamma</th>
<th>Delta</th>
<th>Epsilon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

if root is here

Molecular evolution – p.13/62
Changing the root changes the direction

Placement of the root affects which way bases change but not how many changes there are.
All possible trees (15 in all)
Their best numbers of nucleotide substitutions
The most parsimonious tree
Distance matrix methods

Each possible tree (with branch lengths) predict pairwise distances

Find the tree which comes closest to predicting the observed pairwise distances

observed distances calculated from the data
An example


Xenopus  ?TACCTGGTTGATCCTGCCAGTAG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCACG
Sebastol  ???????????????????????????????AG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCAG
Latimeri  ?TACCTGGTTGATCCTGCCAGTAG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCATG
Squalus  ???????????????????????????????AG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCATG
Myxine  ??CCCTGGTTGATCCTGCCAGCGCAGTAG-CATATGCTTGCTCTCAAAAGACTAAGCCATGCATG
Petromyz  ???CTGGTTGATCCTGCCAGCTAG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCATG
Branch  ???CTGGTTGATCCTGCCAGCTAG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCATG
Styela  ??ATCTGGTTGATCCTGCCAGCTAG-CATATGCTTGCTCTCAAAAGATTAAGCCATGCATG
Herdman  ??TATCTGGTTGATCCTGCCAGCTAGTCAAA-GATTAAGCCATGCATG
Saccogl  ??ACCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Ophiophol  ??ACCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Strongyl  ??ACCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Placopec  CAACCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Limicol  ?TATCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Eurypelm  ?TACCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG
Tenebrio  ??TCCCTGGTTGATCCTGCCAGCTAGTCAAAAGATTAAGCCATGCATG

(and so on for 33 more pages)
The tree with images of the animals

- **Placopecten** (scallop)
- **Strongylocentrotus** (sea urchin)
- **Saccoglossus** (acorn worm)
The tree with images of the animals

*Branchiostoma* (amphioxus)

*Myxine* (hagfish)

*Tunicate*
The tree with images of the animals

*Petromyzon*
(lamprey)

dogfish

dogfish
The tree with images of the animals

- **Sebastoles** (lrockfish)
- **Latimeria** (coelacanth)
- **Xenopus** (clawed frog)
Blair and Hedges’ alternative tree

(in Molecular Biology and Evolution, 2005.)
Molecular evolution (1963 on)

Linus Pauling in 1963  Emile Zuckerkandl, more recently
The late Margaret Dayhoff

Responsible for the first computer-produced molecular phylogeny (1966), the start of protein sequence databases (1965), the recognition of gene families (1960s-1970s)
Morris Goodman

Using immunological methods with proteins, defined the human-chimp-gorilla clade (1962), later pioneered work on evolution of gene families, especially the globin family.
The late Allan Wilson

With Vincent Sarich, supported the human-chimp-gorilla clade. Advocated use of the “molecular clock” by which they dated the divergence of these species to 5 million years ago. Later (as we shall see) found the tree of human mitochondria, whose ancestor was “mitochondrial Eve”.
An example: who is most closely related to whales?

from Amrine-Madsen, H. et al., 2003, Molecular Phylogenetics and Evolution
The tree of human ancestry

Old world monkeys
Orangutan
Chimp
Bonobo
Humans
Gorilla
Gibbon
New world monkeys
Just who are you calling an ape?

- Old world monkeys
- Orangutan
- Chimp
- Bonobo
- Humans
- Gorilla
- Gibbon
- New world monkeys
Just who are you calling an ape?

Old world monkeys
- Orangutan
- Chimp
- Bonobo
- Humans
- Gorilla
- Gibbon

New world monkeys
32 mammals from Homeobox-containing protein 1

- Ornithorhyncus
  - Canis
  - Echinops
  - Procavia
  - Callithrix
  - Pteropus
  - Erinaceus
  - Microcebus
  - Sorex
  - Tarsius
  - Myotis
  - Gorilla
  - Otolemur
  - Pan
  - Equus
    - Macaca
      - Mus
        - Rattus
          - Sus
            - Dipodomys
              - Cavia
                - Ochotona
      - Pongo
      - Tursiops
        - Macropus
          - Vicugna
            - Dasypus
              - Choloepus
                - Homo
          - Monodelphi
    - Bos
... coloring in branches that are or aren’t true
The same 32 using E3 ubiquitin-protein ligase
... does considerably better
But using both of these loci ...
... is not bad at all!
Using 19 loci, the consensus tree is
Using 19 loci concatenated, the tree is

- Sorex
  - Myotis
  - Bos
    - Tursiops
  - Equus
    - Canis
    - Cavia
      - Cavia
      - Mus
        - Homo
          - Macaca
            - Choloepus
            - Loxodonta
              - Monodelphi
              - Ornithorhy
Do these trees agree with each other? Well, ...
The “expert tree” from Timetree.org
Some named groups widely agreed upon

- Sorex
- Myotis
- Tursiops
- Equus
- Canis
- Cavia
- Mus
- Homo
- Macaca
- Choloepus
- Loxodonta
- Monodelphis
- Ornithorhynchus

- Laurasiatheria
- Boroeutheria
- Eutherians (Placental Mammals)
  - Atlantotheria
  - Rodents
  - Primates
  - Xenarthra
  - Afrotheria

- Cetartiodactyla
How does the consensus tree agree?
How does the concatenated tree agree?

- Sorex
- Myotis
- Bos
- Tursiops
- Equus
- Canis
- Cavia
- Mus
- Homo
- Macaca
- Choloepus
- Loxodonta
- Monodelphis
- Ornithorhynchus

Colors indicate agreement:
- Red: disagrees
- Orange: weakly disagrees
- Green: agrees
Molecular phylogenies

Some examples of other important conclusions from molecular phylogenies:

- Using immunological distances, Morris Goodman (1962 on) and later Wilson and Sarich (1966) show that humans, gorilla, and chimps were a clade.
- Wilson and Sarich (in that work, 1967) date the divergence of humans to 5 million years.
- Carl Woese (1978) uses rRNA trees to introduce evolution into microbiology, argue for the domain Archaea.
- Much progress on early radiation of angiosperms
- Protostome-deuterostome tree of metazoans (more or less) replaced by deuterostome-lophotrichozoa-ecdysozoa tree.
- Fungi closer to animals than either is to plants.
- Symbiotic origin of mitochondria and of chloroplasts verified.
- Amphioxus diverged before split of tunicates from craniate chordates.
- Lots of horizontal gene transfer in prokaryotes, almost not a tree.
Different parts of hemoglobin genes

Alignment of hemoglobin ε loci of Human, Tarsier

<table>
<thead>
<tr>
<th>region</th>
<th>bases</th>
<th>differences</th>
<th>% different</th>
</tr>
</thead>
<tbody>
<tr>
<td>upstream</td>
<td>100</td>
<td>12</td>
<td>12.0</td>
</tr>
<tr>
<td>exon 1</td>
<td>92</td>
<td>9</td>
<td>9.8</td>
</tr>
<tr>
<td>intron 1</td>
<td>126</td>
<td>26</td>
<td>20.6</td>
</tr>
<tr>
<td>exon 2</td>
<td>223</td>
<td>26</td>
<td>11.7</td>
</tr>
<tr>
<td>intron 2</td>
<td>820</td>
<td>239</td>
<td>29.1</td>
</tr>
<tr>
<td>exon 3</td>
<td>129</td>
<td>13</td>
<td>10.1</td>
</tr>
<tr>
<td>downstream</td>
<td>100</td>
<td>13</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Differences in exons

position 1  10
position 2  5
position 3  33
Higher Rates of Substitution for ...

1. ... some proteins than others
2. ... some sites within proteins than others (less in active sites, interior sites)
3. ... some amino acid replacements than others (less changes to chemically more similar amino acids)
4. ... silent changes than nonsilent ones
5. ...“in-between" DNA than introns, introns than coding sequences
6. ... transitions than transversions
Morris Goodman tabulation for $\beta$ hemoglobin

<table>
<thead>
<tr>
<th>where</th>
<th>sites</th>
<th>change/100myr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heme contacts</td>
<td>21</td>
<td>0.02</td>
</tr>
<tr>
<td>Nonheme contacts</td>
<td>10</td>
<td>0.02</td>
</tr>
<tr>
<td>Salt bridges $\beta$-$\beta$</td>
<td>4</td>
<td>0.00</td>
</tr>
<tr>
<td>2,3-DPG binding</td>
<td>4</td>
<td>0.10</td>
</tr>
<tr>
<td>Nonsalt bridge $\alpha$,$\beta$ contacts</td>
<td>16</td>
<td>0.16</td>
</tr>
<tr>
<td>Remaining interior</td>
<td>21</td>
<td>0.09</td>
</tr>
<tr>
<td>Remaining exterior</td>
<td>70</td>
<td>0.20</td>
</tr>
<tr>
<td>All</td>
<td>146</td>
<td>0.13</td>
</tr>
</tbody>
</table>
PhyloHMM analysis of multiple genomes

Fig. 5. A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the MET gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities ($\times 1000$) of state $s_1$ in the phylo-HMM (dark gray). Plots similar to this one, showing phylo-HMM-based conservation scores across the whole human genome, can be viewed online at http://genome.ucsc.edu.

From a paper by Siepel and Haussler (*Journal of Computational Biology*, 2004) describing the machinery for finding conserved regions of multiple genomes.
Rates of change from neutral and selective mechanisms

**Neutral mutations**

A fraction $\mu$ of all copies of a gene mutate. Of these $\frac{1}{2N}$ (equal to the initial frequency of the mutant) succeed in drifting to fixation for the mutant.

There are in all $2N$ copies of the gene available to mutate.

The resulting rate of substitution is

$$\mu \times \frac{1}{2N} \times 2N = \mu$$

So the rate of substitution of neutral mutations is equal to the mutation rate (the mutation rate of neutral mutants, not the total mutation rate).
Change by neutral mutation

end of a lineage which is the ancestry of a single copy of the gene:
change is expected to be $\mu$ per generation

Interestingly, the expected rate of change is the same no matter what the population size is (small populations make all copies more likely to become the same, but all are still expected to differ from their ancestors by this amount)
Rates from neutral and selective mechanisms

Selectively advantageous mutations

A fraction $\mu$ of copies of the gene mutate. There are in all $2N$ copies available. A fraction $2s$ succeed in fixing.

The resulting rate of substitution is

$$\mu \times 2N \times 2s = 4Ns\mu$$

Note that this is $4Ns$ times as high as for neutral mutants, if the mutation rate in both categories were equal (which it isn’t).
Substitution of neutral and advantageous mutations

Suppose that the population size is \( N = 1,000,000 \), and mutation rates are:

\[
\text{Advantageous mutations} \quad u_a = 10^{-7} \\
\text{Neutral mutations} \quad u_n = 10^{-6}
\]

If the selection coefficient in favor of advantageous mutations is \( s = 0.0001 \), the rates of substitution expected are:

\[
\text{Advantageous mutations} \quad (4Ns)u_a = 4 \times 10^{-5} \\
\text{Neutral mutations} \quad u_n = 10^{-6}
\]
The molecular clock (from Wilson, 1976)
A not-quite-clocklike tree of *Plethodon* salamanders

(Palmer et al., 2005, in *Molecular Biology and Evolution*)
Phylogeny and gene tree with a gene duplication

A phylogeny with a gene duplication event:

- Frog
- Human
- Monkey
- Squirrel

gene duplication

species boundary
gene duplication
tree of genes
Phylogenies, gene trees, and gene duplication

So when genes are all aligned with each other, their "gene tree" is:

These two identical subtrees should be identical.
A gene family

MUTATIONAL LINK LENGTHS
(O: ORIGINAL, A: AUGMENTED)

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>A</th>
<th>O</th>
<th>A</th>
<th>O</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>43</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>11</td>
<td>16</td>
<td>21</td>
<td>44</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td>24</td>
<td>47</td>
<td>63</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>14</td>
<td>20</td>
<td>27</td>
<td>50</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>51</td>
<td>68</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>16</td>
<td>22</td>
<td>29</td>
<td>52</td>
<td>85</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>18</td>
<td>29</td>
<td>31</td>
<td>54</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>32</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>34</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>36</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>