EVOLUTIONARY INFERENCE:
Some basics of phylogenetic analyses.

Ana Rojas Mendoza
CNIO-Madrid-Spain.
Alfonso Valencia’s lab.
Aims of this talk:

1. To introduce relevant concepts of evolution to practice phylogenetic inference from molecular data.

2. To introduce some of the most useful methods and computer programmes to practice phylogenetic inference.

3. To show some examples I’ve worked in.
1-Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
Owen’s definition of homology

Richard Owen, 1843

- **Homologue**: the same organ under every variety of form and function (true or essential correspondence).

- **Analogy**: superficial or misleading similarity.
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- **Homology vs similarity.**
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- **Allele mutation vs allele substitution.**
- Rates of allele substitution.
- Neutral theory of evolution.
Similarity ≠ Homology

- Similarity: mathematical concept
- Homology: biological concept
  - Common Ancestry!!!
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Orthologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
Homologous genes

- **Orthologous genes**
  Derived from a process of new species formation (speciation)

- **Paralogous genes**
  Derived from an original gene duplication process in a single biological species
Homologous genes

Species1

A

A

B

Duplication

Species2

A

A

Orthologs

Orthologs

Paralogs

Homologs

SOME BASICS
HOMOLOGS/ORTHOLOGS/PARALOGS

Recent duplication

Ras (H. sapiens)

Ras2 (H. sapiens)

Ras (M. musculus)

Ras (C. elegans)

Rab (H. sapiens)

Rab (M. musculus)

Rab (C. elegans)

in-paralogs.

Group of orthologs and in-paralogs

Ras subfamily

Both families are paralogs between each other!

Rab subfamily

Group of ortholgs

SOME BASICS
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
Species trees vs Gene trees

Orthologous genes of Cytochrome
Each one is present in a biological species

• Paralogous genes of Globin
• a, b, d (Glob), Myo y Leg haemoglobin, each originated by duplication from an ancestral gene
Species trees and Gene trees

We often assume that gene trees give us species trees.
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Orthologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
Is there a molecular clock?

• The idea of a molecular clock was initially suggested by Zuckerkandl and Pauling in 1962.

• They noted that rates of amino acid replacements in animal haemoglobins were roughly proportional to time— as judged against the fossil record.
The molecular clock for alpha-globin:
Each point represents the number of substitutions separating each animal from humans

- Shark
- Carp
- Platypus
- Chicken
- Cow

SOME BASICS
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Orthologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
ALLELE MUTATION VS. FIXATION

Mutation = Individual

Fixation = Population
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Orthologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
Rates of amino acid replacement in different proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Rate (mean replacements per site per $10^9$ years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibrinopeptides</td>
<td>8.3</td>
</tr>
<tr>
<td>Insulin C</td>
<td>2.4</td>
</tr>
<tr>
<td>Ribonuclease</td>
<td>2.1</td>
</tr>
<tr>
<td>Haemoglobins</td>
<td>1.0</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>0.3</td>
</tr>
<tr>
<td>Histone H4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- *Evolutionary rates depends on functional constraints of proteins*
**Nucleotide substitution rate** $(r)$:

$\#$ substitutions per site per per year

$$r = \frac{K}{2T}$$

- **$K$**: number of substitutions per site between homologous sequences.
- **$T$**: Time of divergence.
1. Concepts of Molecular Evolution

- Homology vs Analogy.
- Homology vs similarity.
- Ortologous vs Paralogous genes.
- Species tree vs genes tree.
- Molecular clock.
- Allele mutation vs allele substitution.
- Rates of allele substitution.
- Neutral theory of evolution.
- Homoplasmy.
Neutral theory of evolution

- At molecular level, the most frequent change are those involving fixation in populations of neutral selective variants (Kimura, 1968).
  - Allelic variants are functionally equivalent.
  - Neutralism does not deny adaptive evolution.

- Fixation of new allelic variants occur at a constant rate, it is equal to mutation rate and independent of population parameters.

\[
\text{mutation in population} \quad \text{probability to fix} \\
\quad 2N \cdot m \times \frac{1}{2N} = m
\]
There is no universal clock

• The initial proposal saw the clock as a Poisson process with a constant rate
• Now known to be more complex - differences in rates occur for:
  – different sites in a molecule
  – different genes
  – different base position (synonymous-nonsynonymous)
  – different regions of genomes
  – different genomes in the same cell
  – different taxonomic groups for the same gene

• Molecular Clocks Not Exactly Swiss
2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram.
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.
Phylogenetic systematics

- Sees **homology as evidence** of common ancestry
- Uses tree diagrams to portray relationships based upon recency of common ancestry
- Monophyletic groups (clades) - contain species which are more closely related to each other than to any outside of the group
Phylogenetic Trees

- terminal branches
- node 1
- interior branches
- node 2
- LEAVES
- polytomy
- A CLADOGRAM

SOME BASICS
2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram.
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.
Cladograms and phylograms

Cladograms show branching order - branch lengths are meaningless

Phylograms show branch order and branch lengths

**SOME BASICS**

- Bacterium 1
- Bacterium 2
- Bacterium 3
- Eukaryote 1
- Eukaryote 2
- Eukaryote 3
- Eukaryote 4
2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram.
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.
Trees - Rooted and Unrooted

SOME BASICS
Rooting using an outgroup

Unrooted tree

Rooted by outgroup

Monophyletic Ingroup

Monophyletic Ingroup

SOME BASICS
2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.
Character:
A descriptor that can have different manifestations in different species. (character states)

Types of characters

• Morphological (characteristics of physical attributes).
• Behavioral.
• Ecological (nest type, host plant, prey type).
• Distributional (geographical).
• Physiological/chemical.
• Molecular.
Character evolution

- Heritable changes (in morphology, gene sequences, etc.) produce different character states.
- Similarities and differences in character states provide the basis for inferring phylogeny (i.e. provide evidence of relationships)
- The utility of this evidence depends on how often the evolutionary changes that produce the different character states occur independently.
Why to use molecular data?

- Molecular data are genetic data: $V_p = V_g + V_e$
- Molecular data led us to study a huge amount of characters.
- Any kind of homoplastic similarity vanishing at time more characters are considered.
- Indels, duplications and chromosomic rearrangements are rare events with strong weight of homology.
- Molecular data offers a common measure for evolutionary divergence.
Small subunit ribosomal RNA

18S or 16S rRNA

SOME BASICS
Molecular characters

1. Protein variation (1950s-present)
   Historically, first molecular characters
   a. Isozyme/allozyme variation
      · Used mostly at population level, sometimes Phylogenetic.
      · Misses lots of underlying variation
   b. Amino acid sequencing (1960s, Fitch, etc.)
      · Globin genes
      · Technically difficult
Molecular characters

2. DNA (1970s)
   - Has dominated molecular phylogenetics since.
   - Impact of polymerase chain reaction (PCR).

a. DNA-DNA hybridization (1970s-80s; rare now)
   - Famous studies in birds (Sibley and Ahlquist) made some big changes (birds infamous for lack of allozyme variation)
   - Not character-based; data are pairwise comparisons between OTUs (suitable only for distance analysis)
   - Advantage of looking at entire genome (single copy DNA anyway)
### Some Common Phylogenetic Methods

#### Table of Types of Data

<table>
<thead>
<tr>
<th>Tree building method</th>
<th>Cluster Algorithms</th>
<th>Optimal criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UPGMA</td>
<td>Minimum Evolution</td>
</tr>
<tr>
<td></td>
<td>NJ</td>
<td>Least Square</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Parsimony</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian Inference</td>
</tr>
</tbody>
</table>

**Types of Data**
- Distances
- Sites (nucleotides, aa)
2. Concepts of Phylogenetic Systematics

- What is Phylogenetic systematics?
- Cladogram and Phylogram
- Monophyletic, Paraphyletic and Polyphyletic groups.
- Rooted vs Unrooted trees.
- Ingroup and Outgroup.
- Character states and evolution.
- Homoplasy.
Homoplasy

• Convergent evolution: similarity due to adaptation, Not to common ancestry!

Involve bones that Ih human will make The hands

Both lineages had a hug evolutionary separation before They came fliers! They independently became fliers!

Involve bones that Ih human will make The arms

Human eye and squid (calamari!) eye...
**PHYLOGENETIC ANALYSIS**

**Trees:** cladograms - represents only the branching order of nodes  
phylograms - represents branching order and branch length  
(number of sequence changes between nodes)

**Distance:** number of substitutions that have taken place along a branch

**Tree construction:**

*algorithmic:* uses an algorithm to construct a tree from data  
(NJ, UPGM: distance methods) Fast, one tree ONLY.

*tree-searching:* builds many trees and then uses a criterion to decide which is the best tree. (Character based)

- **Parsimony:** several trees. The most likely scenario involves the fewest changes?
- **ML:** seeks for tree that maximizes the likelihood of observing data.
- **Bayesian:** seeks from several trees with the greatest likelihoods given the data.
PHYLOGENETICS
DANCING!
The five steps in phylogenetics dancing

1. Sequence data
2. Align Sequences
3. Phylogenetic signal? Patterns—>evolutionary processes?
4. Choose a method
   - Characters based methods
     - MB
     - ML
     - MP
   - Distance methods
     - Distance calculation (which model?)
       - Optimality criterion
         - LS
         - ME
       - Single tree
         - NJ
         - UPGMA
5. Calculate or estimate best fit tree
6. Test phylogenetic reliability

SEARCHING DB

QUERY SEQUENCE

Sequences retrieval

Related proteins (alignment)

Homologous sequences

Function?

ras (H. sapiens)
ras2 (H. sapiens)
ras (M. musculus)
ras (C. elegans)
rab (H. sapiens)
rab (M. musculus)
rab (C. elegans)
The five steps in phylogenetics dancing

1. **Sequence data**

2. **Align Sequences**

3. **Phylogenetic signal? Patterns—>evolutionary processes?**

   - **Characters based methods**
     - MB
     - ML
     - MP

   - **Distances methods**
     - Distance calculation (which model?)
     - Optimality criterion
     - Single tree
     - LS \rightarrow ME
     - NJ \rightarrow UPGMA

4. **Choose a method**

5. **Calculate or estimate best fit tree**

6. **Test phylogenetic reliability**

SEARCHING THE DATABASES

Pairwise

Searching : FASTA
(Lipman & Pearson, 1985, Pearson & Lipman 1988)

Basic Local Alignment Search Tool (BLAST)
Altschul, S.F., Gish W., Miller W., Myers E.W., and Lipman D.J.

Profile

PSI-BLAST:
Nucleic Acids Research (1997) v.25, n.17 3389-3402

Iterative search

USE of point position specific matrices.
Use the matrix to search again!
WHY SEARCHING THE DATABASES?

We want to obtain all the sequences related to our query!

OKAY, but which kind of sequences?

Am I looking for distant homologs? → PSI-BLAST

Am I looking for clear orthologs? → FASTA, BLAST

How many sequences should I use? Well depends:
Rule of the thumb:

Get a representative set of your sequences, remove redundancy at least say 80%!
The five steps in phylogenetics dancing

1. Sequence data

2. Align Sequences

3. Phylogenetic signal? Patterns—>evolutionary processes?

4. Characters based methods
   - MB
   - ML
   - MP

   - Model?
   - Model?
   - Weighting? (sites, changes)?

5. Distances methods
   - Calculate or estimate best fit tree
   - Test phylogenetic reliability

ALIGNING THE SEQUENCES

We want to align all the sequences obtained via searching the databases.

Alignment quality is CRUCIAL => bad alignment = bad tree!!!!

METHODS:

Greedy approaches: Progressive alignment (Feng & Doolittle, 1987-96)

PILEUP, ClustalW (improved the Progressive alignment) .. Too greedy! poor when %id < 30%

• Then the real improvements:

  T-COFFEE (Notredame et al, 2000): incorporates local and global information!
  ProbCons(Do, CB, et al, 2005): like T-Coffee with probabilistic estimations!
<table>
<thead>
<tr>
<th>Gene</th>
<th>ID</th>
<th>Length</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASH_HUMAN</td>
<td>5</td>
<td>164</td>
<td>KLVVGGAGGGVGSITIQVINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEYS</td>
</tr>
<tr>
<td>RRAS_HUMAN</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RTC1_HUMAN</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS2_HYDRA</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS2_DROME</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RASL_NEURO</td>
<td>1</td>
<td>159</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RASL_MUSE</td>
<td>1</td>
<td>159</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS1_YEAST</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS_SCHPO</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS_LINED</td>
<td>1</td>
<td>159</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAS2_RHIRA</td>
<td>1</td>
<td>160</td>
<td>KLVVGGGGGGGAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAL_A_HUMAN</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAL_A_RAT</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RALB_HUMAN</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RALB_RAT</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RALB_XENLA</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RAL_DISOM</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RALA_DROME</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>CE139344</td>
<td>1</td>
<td>157</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>CC42_DROME</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH04_YEAST</td>
<td>1</td>
<td>155</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH02_SCHPO</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH02_YEAST</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH08_HUMAN</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH06_HUMAN</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH03_YEAST</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH01_SCHPO</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH01_YEAST</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH01_ENTH</td>
<td>1</td>
<td>151</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
<tr>
<td>RH01_DROME</td>
<td>1</td>
<td>156</td>
<td>KVIMVGGGAAAGGGSITIQFINIFVDEMDFTIEEISYRKQVVIDGETCLLIDILDTAGGEEFG</td>
</tr>
</tbody>
</table>

**WHAT CAN I LEARN FROM MY ALIGNMENT?**

- **Correlated mutations**
- **Tree-determinant**

**ALIGNING DB**
Mutational behaviour
Pazos Valencia, 2001

Relative entropy cut,
del Sol, Valencia 2002

Casari, Sander, Valencia Nature Str. Biol. 95
Pazos, Valencia 2003
Romero, Valencia 04

Del Sol, Pazos, Valencia JMB 03
The five steps in phylogenetics dancing

1. Sequence data
2. Align Sequences
3. Phylogenetic signal? Patterns—>evolutionary processes?
4. Choose a method
   - MB
   - ML
   - MP
   - Distance methods
     - Optimal criteria
     - Single tree
     - Distance calculation
6. Calculate or estimate best fit tree
7. Test phylogenetic reliability

Distance Methods

• Distance Estimates attempt to estimate the mean number of changes per site since 2 species (sequences) split from each other.

• Simply counting the number of differences (p distance) may underestimate the amount of change - especially if the sequences are very dissimilar - because of multiple hits.

• We therefore use a model which includes parameters which reflect how we think sequences may have evolved.
Transitions: changes between Pyrs or purs.

Tranversions: changes between Pyrs AND purs (2X more frequent*)

Inversion: 180 rotation ds-DNA (more than 2 bases)
12 mutations acumulated

Only 3 detected!
METHODS

Distances: observed and real

<table>
<thead>
<tr>
<th>Obs</th>
<th>Real</th>
<th>Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>no</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>simple</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>coincidence</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>multiple</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>paralell</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>convergent</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>reverse</td>
</tr>
</tbody>
</table>

Obs might be <<<< Real changes!
## Distance calculations

### Similarity
- **SEQ1**  `ACGTACGTAA`
- **SEQ2**  `ACGTTTCGTAT`
- **SEQ3**  `TCCATTCGTAA`

### Distance

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>0.2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0.4</td>
<td>0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

- (1–2) 80%  
  \[1 - 0.8 = 0.2\]
- (1–3) 60%  
  \[1 - 0.6 = 0.4\]
- (2–3) 60%  
  \[1 - 0.6 = 0.4\]
Saturation in sequence data:

- Saturation is due to multiple changes at the same site subsequent to lineage splitting.

- Most data will contain some fast evolving sites which are potentially saturated (e.g. in proteins often position 3).

- In severe cases the data becomes essentially random and all information about relationships can be lost.
Multiple changes at a single site - hidden changes

Seq 1  AGCGAG
Seq 2  GCGGAC

Number of changes

1  2  3

Seq 1  C → G → T → A
Seq 2  C →  A
METHODS

Observed

Substitution numbers

time
The simplest model is that of Jukes & Cantor:
\[ d_{xy} = -(3/4) \ln (1-4/3 \ D) \]

- \( d_{xy} \) = distance between sequence \( x \) and sequence \( y \) expressed as the number of changes per site.

- (note \( d_{xy} = r/n \) where \( r \) is number of replacements and \( n \) is the total number of sites. This assumes all sites can vary and when unvaried sites are present in two sequences it will underestimate the amount of change which has occurred at variable sites).

- \( D \) = is the observed proportion of nucleotides which differ between two sequences (fractional dissimilarity).

- \( \ln \) = natural log function to correct for superimposed substitutions.

- The 3/4 and 4/3 terms reflect that there are four types of nucleotides and three ways in which a second nucleotide may not match a first - with all types of change being equally likely (i.e. unrelated sequences should be 25% identical by chance alone).
The natural logarithm \( \ln \) is used to correct for superimposed changes at the same site

- If two sequences are 95% identical, they are different at 5% or 0.05 (D) of sites thus:
  \[
  - d_{xy} = -3/4 \ln (1-4/3 \ 0.05) = 0.0517
  \]

- Note that the observed dissimilarity 0.05 increases only slightly to an estimated 0.0517 - this makes sense because in two very similar sequences one would expect very few changes to have been superimposed at the same site in the short time since the sequences diverged apart.

- However, if two sequences are only 50% identical they are different at 50% or 0.50 (D) of sites thus:
  \[
  - d_{xy} = -3/4 \ln (1-4/3 \ 0.5) = 0.824
  \]

- For dissimilar sequences, which may diverged apart a long time ago, the use of \( \ln \) infers that a much larger number of superimposed changes have occurred at the same site.
**METHODS**

Distance models can be made more parameter rich to increase their realism.

- It is better to use a model which fits the data than to blindly impose a model on data.

- The most common additional parameters are:
  - A correction for the proportion of sites which are unable to change.
  - A correction for variable site rates at those sites which can change.
  - A correction to allow different substitution rates for each type of nucleotide change.
A gamma distribution can be used to model site rate heterogeneity.
Distances: advantages:

• Fast - suitable for analysing data sets which are too large for ML.

• A large number of models are available with many parameters - improves estimation of distances.

• Use ML to test the fit of model to data.
Obtaining a tree using pairwise distances

Additive distances:

- If we could determine exactly the true evolutionary distance implied by a given amount of observed sequence change, between each pair of taxa under study, these distances would have the useful property of tree additivity
The branch lengths in the matrix and the tree path lengths match perfectly - there is a single unique additive tree
Distance estimates may not make an additive tree

Some path lengths are longer and others shorter than appear in the matrix

Jukes-Cantor distance matrix
Proportion of sites assumed to be invariable = 0.56; identical sites removed proportionally to base frequencies estimated from constant sites only

```
  1  2  4  5  6
1 ruber     -
2 Aquifex   0.38745 -
4 Deinococc 0.22455 0.47540 -
5 Thermus   0.13415 0.27313 0.23615 -
6 Bacillus  0.27111 0.33595 0.28017 0.28846 -
```
Obtaining a tree using pairwise distances

- Stochastic errors will cause deviation of the estimated distances from perfect tree additivity even when evolution proceeds exactly according to the distance model used.

- Poor estimates obtained using an inappropriate model will compound the problem.

- How can we identify the tree which best fits the experimental data from the many possible trees.
Obtaining a tree using pairwise distances

- We have uncertain data that we want to fit to a tree and find the optimal value for the adjustable parameters (branching pattern and branch lengths).

- Use statistics to evaluate the fit of tree to the data (goodness of fit measures)
  - Fitch Margoliash method - a least squares method
  - Minimum evolution method - minimises length of tree

- Note that neighbor joining while fast does not evaluate the fit of the data to the tree.
Fitch Margoliash Method 1968:

- Minimises the weighted squared deviation of the tree path length distances from the distance estimates.
Fitch Margoliash Method 1968:

Optimality criterion = distance (weighted least squares with power=2)
Score of best tree(s) found = 0.12243 (average %SD = 11.663)

Tree #       1       2
Wtd. S.S.   0.13817  0.12243
APSD        12.391  11.663

Tree 1

Tree 2 - best

METHODS
Minimum Evolution Method:

• For each possible alternative tree one can estimate the length of each branch from the estimated pairwise distances between taxa and then compute the sum (S) of all branch length estimates. The minimum evolution criterion is to choose the tree with the smallest S value.
METHODS

Minimum Evolution

Optimality criterion = distance (minimum evolution)
Score of best tree(s) found = 0.68998

Tree #           1       2
ME-score   0.68998 0.69163

Tree 1 - best
Aquifex 0.217
Bacillus 0.119
Deinococcus ruber 0.217
Thermus 0.056

Tree 2
Aquifex
Bacillus 0.057
Deinococcus ruber 0.017
Thermus 0.055
Thermus ruber 0.145
Neighbor joining method

- The neighbor joining method is a greedy heuristic which joins at each step, the two closest sub-trees that are not already joined.
- It is based on the minimum evolution principle.
- One of the important concepts in the NJ method is *neighbors*, which are defined as two taxa that are connected by a single node in an unrooted tree.
## Distance Matrix

<table>
<thead>
<tr>
<th>PAM</th>
<th>Spinach</th>
<th>Rice</th>
<th>Mosquito</th>
<th>Monkey</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinach</td>
<td>0.0</td>
<td>84.9</td>
<td>105.6</td>
<td>90.8</td>
<td>86.3</td>
</tr>
<tr>
<td>Rice</td>
<td>84.9</td>
<td>0.0</td>
<td>117.8</td>
<td>122.4</td>
<td>122.6</td>
</tr>
<tr>
<td>Mosquito</td>
<td>105.6</td>
<td>117.8</td>
<td>0.0</td>
<td>84.7</td>
<td>80.8</td>
</tr>
<tr>
<td>Monkey</td>
<td>90.8</td>
<td>122.4</td>
<td>84.7</td>
<td>0.0</td>
<td><strong>3.3</strong></td>
</tr>
<tr>
<td>Human</td>
<td>86.3</td>
<td>122.6</td>
<td>80.8</td>
<td><strong>3.3</strong></td>
<td>0.0</td>
</tr>
</tbody>
</table>
PAM distance 3.3 (Human - Monkey) is the minimum. So we'll join Human and Monkey to MonHum and we'll calculate the new distances.
After we have joined two species in a subtree we have to compute the distances from every other node to the new subtree. We do this with a simple average of distances:

\[
\text{Dist}[\text{Spinach, MonHum}] = \frac{\text{Dist}[\text{Spinach, Monkey}] + \text{Dist}[\text{Spinach, Human}]}{2} = \frac{90.8 + 86.3}{2} = 88.55
\]
METHODS

Last Joining

<table>
<thead>
<tr>
<th>PAM</th>
<th>SpinRice</th>
<th>MosMonHum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spinach</td>
<td>0.0</td>
<td>108.7</td>
</tr>
<tr>
<td>MosMonHum</td>
<td>108.7</td>
<td>0.0</td>
</tr>
</tbody>
</table>

(Spin-Rice)-(Mos-(Mon-Hum))
Unrooted Neighbor-Joining Tree
The five steps in phylogenetics dancing

1. Sequence data
2. Align Sequences
3. Phylogenetic signal? Patterns—>evolutionary processes?
4. Choose a method
   - Characters based methods
   - Distances methods
     - Model?
     - Model?
     - Weighting?
     - Distance calculation (which model?)
     - Optimality criterion
     - Single tree
   - Calculate or estimate best fit tree
5. Test phylogenetic reliability

ML: comparison with other methods.

- ML is similar to many other methods in many ways.
- In many ways it is fundamentally different.
- ML assumes a model of sequence evolution (so does Maximum Parsimony and so do distance matrix methods).

- ML attempts to answer the question: What is the probability that I would observe these data (a multiple sequence alignment), given a particular model of evolution (a tree and a process).
Maximum Likelihood - goal

- To estimate the probability that we would observe a particular dataset, given a phylogenetic tree and some notion of how the evolutionary process worked over time.

\[ P(D/H) \]

Probability of \( \pi = [a,c,g,t] \) given
The model

- The two parts of the model are the tree and the process (the model).

- The model is composed of the composition and the substitution process - rate of change from one character state to another character state.

\[
\pi = [a,c,g,t]
\]
Does changing a model affect the outcome?

There are different models

**Jukes and Cantor (JC69):**
All base compositions equal (0.25 each), rate of change from one base to another is the same

**Kimura 2-Parameter (K2P):**
All base compositions equal (0.25 each), different substitution rate for transitions and transversions.

**Hasegawa-Kishino-Yano (HKY):**
Like the K2P, but with base composition free to vary.

**General Time Reversible (GTR):**
Base composition free to vary, all possible substitutions can differ.

All these models can be extended to accommodate invariable sites and site-to-site rate variation.
Strengths of ML

• Does not try to make an observation of sequence change and then a correction for superimposed substitutions. **There is no need to ‘correct’ for anything**, the models take care of superimposed substitutions.

• Accurate branch lengths.

• Each site has a likelihood.

• If the model is correct, we should retrieve the correct tree*.

• You can use a model that fits the data.

• ML uses all the data (no selection of sites based on informativeness, all sites are informative).

• ML can not only tell you about the phylogeny of the sequences, but also the process of evolution that led to the observations of today’s sequences.

*If we have long-enough sequences and a sophisticated-enough model.
Weaknesses of ML

- Can be inconsistent if we use models that are not accurate.

- Model might not be sophisticated enough (you can ‘max-out’ on models).

- Very computationally-intensive. Might not be possible to examine all models (substitution matrices, tree topologies, etc.).
Parsimony Analysis

• Given a set of characters, such as aligned sequences, parsimony analysis works by determining the fit (number of steps) of each character on a given tree

• The sum over all characters is called Tree Length
  – Most parsimonious trees (MPTs) have the minimum tree length needed to explain the observed distributions of all the characters
RESULTS

Results of parsimony analysis

• One or more most parsimonious trees.

• Hypotheses of character evolution associated with each tree (where and how changes have occurred).

• Branch lengths (amounts of change associated with branches).

• Various tree and character statistics describing the fit between tree and data.

• Suboptimal trees – optional.
Parsimony - advantages

- is a simple method - easily understood operation.
- does not seem to depend on an explicit model of evolution.
- gives both trees and associated hypotheses of character Evolution.
- should give reliable results if the data is well structured and homoplasy is either rare or widely (randomly) distributed on the tree.
Parsimony - disadvantages

- May give misleading results if homoplasy is common or concentrated in particular parts of the tree, e.g:
  - thermophilic convergence
  - base composition biases
  - long branch attraction

- Underestimates branch lengths.

- Model of evolution is implicit - behaviour of method not well understood.

- Parsimony often justified on purely philosophical grounds – we must prefer simplest hypotheses - particularly by morphologists.

- For most molecular systematists this is unconvincing.
Parsimony can be inconsistent

- Felsenstein (1978) developed a simple model phylogeny including four taxa and a mixture of short and long branches.
- Under this model parsimony will give the wrong tree

\[
\begin{array}{c}
A \\
p \\
C \quad q \quad q \\
q \\
D
\end{array} \quad \begin{array}{c}
B \\
p \\
q \\
q \\
\text{Rates or branch lengths} \\
p \gg \gg q
\end{array}
\]

- With more data the certainty that parsimony will give the wrong tree increases - so that parsimony is statistically inconsistent.

- Advocates of parsimony initially responded by claiming that Felsenstein’s result showed only that his model was unrealistic.

- It is now recognised that the long-branch attraction in the Felsenstein Zone is one of the most serious problems in phylogenetic inference.
Bayesian Inference of Phylogeny

- Clustering methods: UPGMA, NJ
- Parsimony: minimization of cost
- Statistical approaches
  - Maximum Likelihood
  - Bayesian Inference
Statistical methods

- **Maximum likelihood**
  - Standard statistical approach
  - Philosophy widely accepted
  - Computationally difficult, especially for confidence intervals

- **Bayesian inference**
  - Old but marginal statistical approach until recently
  - Philosophy controversial (subjective probability)
  - Computationally efficient numerical solutions to difficult, high-dimensional problems
METHODS

Infer relationships among three species:

Outgroup:
Three possible trees (topologies):
Bayes’ rule

Joint probabilities

- $\Pr(B) = 0.6$
- $\Pr(S) = 0.5$
- $\Pr(W) = 0.4$
- $\Pr(D) = 0.5$

- $\Pr(\bigcirc) = \Pr(B, D) = 0.2$
- $\Pr(\bullet) = \Pr(B, S) = 0.4$
- $\Pr(\bigcirc) = \Pr(W, D) = 0.3$
- $\Pr(\bigcirc) = \Pr(W, S) = 0.1$

Conditional probabilities

- $\Pr(B|D) = \frac{2}{5} = 0.2$

- Hide all solid marbles (leaving 5 with dot)
- Of those left, 2 are black

Bayes’ rule

\[
\Pr(B|D) = \frac{\Pr(B) \Pr(D|B)}{\Pr(D)}
\]

\[
\Pr(D) = \Pr(B, D) + \Pr(W, D)
\]

\[
\Pr(D) = \Pr(B) \Pr(D|B) + \Pr(W) \Pr(D|W)
\]

Marginal probability

It is obtained by “marginalizing over” color
Bayes’ theorem

\[ f(\theta \mid X) = \frac{p(\theta)l(X \mid \theta)}{\int p(\theta)l(X \mid \theta)d\theta} \]

Posterior distribution
Prior distribution
Likelihood function

Unconditional probab.

\[ \Pr \left[ \text{Tree/Data} \right] = \frac{(\Pr \left[ \text{Tree} \right] \times \Pr \left[ \text{Data/Tree} \right])}{\Pr \left[ \text{Data} \right]} \]
METHODS

Prior probability distribution

Data (observations)

Posterior probability distribution
Model: tree and branch lengths

\[ \theta \text{ Parameters} \]

- topology \((\tau)\) (branching order)
- branch lengths \((v_i)\) (expected amount of change per site or character)

\[ \theta = (\tau, v) \]
METHODS

Data

X The data

Taxon Characters

<table>
<thead>
<tr>
<th>ACG</th>
<th>TTA</th>
<th>TTA</th>
<th>AAT</th>
<th>TGT</th>
<th>CCT</th>
<th>CTT</th>
<th>TTC</th>
<th>AGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACG</td>
<td>TGT</td>
<td>TTC</td>
<td>GAT</td>
<td>CGT</td>
<td>CCT</td>
<td>CTT</td>
<td>TTC</td>
<td>AGA</td>
</tr>
<tr>
<td>ACG</td>
<td>TGT</td>
<td>TTA</td>
<td>GAC</td>
<td>CGA</td>
<td>CCT</td>
<td>CGG</td>
<td>TTA</td>
<td>AGG</td>
</tr>
<tr>
<td>ACA</td>
<td>GGA</td>
<td>TTA</td>
<td>GAT</td>
<td>CGT</td>
<td>CCG</td>
<td>CTT</td>
<td>TTC</td>
<td>AGA</td>
</tr>
</tbody>
</table>
Markov Chain Monte Carlo (MCMC)

\[ p(\theta) l(X \mid \theta) \]

METHODS
Model parameters 1

METHODS

General Time Reversible substitution model

$$Q = \begin{pmatrix}
\pi_A r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\
\pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\
\pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\
\pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & -
\end{pmatrix}$$

topology ($\tau$) branch lengths ($v_i$)
MODEL PARAMETERS 2

Gamma-shaped rate variation across sites

Gamma-shaped rate variation across sites
METHODS

stationary phase sampled with thinning (rapid mixing essential)

\[ \ln L \]

\[ \text{Generation} \]
Majority rule consensus tree from an MCMC run (insect 18S data, GTR + G)

Frequencies represent the posterior probability of the clades

Probability of clade being true given data and model
Bayesian inference pitfalls

- To what extent is the posterior distribution influenced by the prior?
- How do we know that the chains have converged onto the stationary distribution?
- Most common approach is to compare independent runs starting from different points in parameter space
PROGRAMS:

ML:
PAUP: http://paup.csit.fsu.edu/about.html  David Swofford. (U-L,M,W) UNIX,MAC,Windows
PAML: http://abacus.gene.ucl.ac.uk/software/paml.html  Ziheng Yang (U-L,M)
MOLPHY: Jun Adachi and Massami Hasegawa (Pascal)
PASSML: Pietro Lio (Hidden Markov) (U)

MB:
BAMBE: http://www.mathcs.duq.edu/larget/bambe.html  Donald Simon & B. Larget UNIX, Windows
Mac5: http://www.agapow.net/software/mac5/  Paul-Michael Agapow
       UNIX,Windows,MAC

OTHERS!
MEGA2: http://www.megasoftware.net/  Kumar et al. DOS/Windows
      Check out the list of Joe Felsenstein!
Here are some 194 of the phylogeny packages, and 16 free servers, that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or importance. Updates to these pages are made about twice a year (however, almost no dates have been made since the start of 2001, and this will continue until at least the end of 2002 when I hope to complete a major writing project).

Some of these programs are available over Internet from ftp server machines, or by World Wide Web.

The programs listed below include both free and non-free ones; in some cases I do not know whether a program is free. I have listed as free those that I knew were free, for the others you have to ask their distributor.

If you discover any inaccuracies, or feel that I have left any important programs or facts out, or if links do not work properly, please e-mail me at: joe@genetics.washington.edu.

List of packages arranged...
DNA
DNAPARS. Estimates phylogenies by the parsimony method using nucleic acid sequences.
DNAMOVE. Interactive construction of phylogenies from nucleic acid sequences, with their evaluation by parsimony and compatibility.
DNAPENNY. Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search.
DNACOMP. Estimates phylogenies from nucleic acid sequence data using the compatibility criterion.
DNAINVAR. For nucleic acid sequence data on four species, computes Lake's and Cavender's phylogenetic invariants.
DNAML. Estimates phylogenies from nucleotide sequences by maximum likelihood.
DNAMLK. Same as DNAML but assumes a molecular clock.
DNADIST. Computes four different distances between species from nucleic acid sequences.

Proteins
PROTPARS. Estimates phylogenies from protein sequences using the parsimony method.
PROTDIST. Computes a distance measure for protein sequences.

Restriction
RESTML. Estimation of phylogenies by maximum likelihood using restriction sites data.

Continuous
CONTML. Estimates phylogenies from gene frequency data by maximum likelihood.
GENDIST. Computes one of three different genetic distance formulas from gene frequency data.

Discrete characters
MIX. Wagner parsimony method and Camin-Sokal parsimony method.
MOVE. Interactive construction of phylogenies from discrete character data.
Evaluates parsimony and compatibility criteria.
PENNY. Finds all most parsimonious phylogenies.
DOLLOP. Estimates phylogenies by the Dollo or polymorphism parsimony criteria.
DOLMOVE. Interactive DOLLOP.
DOLPENNY. branch-and-bound method.
CLIQUE. Finds the largest clique of mutually compatible characters.

SEQBOOT. Reads in a data set, and produces multiple data sets from it by bootstrap resampling.
FITCH. Estimates phylogenies from distance matrix data under the "additive tree model".
KITSCH. Estimates phylogenies from distance matrix data under the "ultrametric" model.
NEIGHBOR. An implementation of Saitou and Nei's "Neighbor Joining Method," and of the UPGMA (Average Linkage clustering) method.
CONSENSE. Computes consensus trees by the majority-rule consensus tree method.
TOOL: MR. BAYES

Based on concept of posterior probabilities: probabilities that are estimated, based on some models (prior expectations), after learning something about the data (Mau et al., 1999). The user postulates a model of evolution, and the program searches for the best trees consistent with both the model, and the data (aln)
Method: Metropolis-coupled Markov Chain Monte Carlo: is a set of independent searches that occasionally exchanges information.

Model for aa replacement: Jones.
Number of markov chains: 4
Number of generations: >900,000
Number of trees generated: 1 tree each 100 generations.
Only trees generated after likelihood convergence are sampled (usually I discard 20% of the initial trees).
MrBayes v3.0B4
(Bayesian Analysis of Phylogeny)

by

John P. Huelsenbeck and Fredrik Ronquist

Section of Ecology, Behavior and Evolution
Division of Biological Sciences
University of California, San Diego
johnn@biomail.ucsd.edu

Department of Systematic Zoology
Evolutionary Biology Centre
Uppsala University
fredrik.ronquist@ebc.uu.se

Type "help" or "help <command>" for information
on the commands that are available.

MrBayes > execute <filename>
Expecting <name>

MrBayes > sumt filename=<filename.t> contype=allcompat burnin=300

http://morphbank.ebc.uu.se/mrbayes3/info.php

Linusx, windows, mac...
### PROGRAMS

**Shortcut to MrBayes3_0b4.exe**

- **Reading MrBayes block**
- **Setting Nst to 6**
- **Setting Rates to Gamma**
- **Successfully set likelihood model**
- **Setting autoclose to yes**
- **Setting number of generations to 1000**
- **Setting number of chains to 4**
- **Setting print frequency to 100**
- **Setting sample frequency to 10**
- **Setting program to save branch lengths**
- **Setting chain output file name to 22all.t**
- **Running Markov chain**
- **MCMC stamp = 1713738510**

**Model settings:**

- **Datatype = DNA**
- **Nucmodel = 4by4**
- **Nst = 6**
- **Substitution rates of the rate subclasses : 0.01.0.01.0.01.0.01**

- **Covariation = No**
- **W States = 4**
- **State frequencies = Gamma**
- **Rates = Gamma**
- **Gamma shape parameter distributed on the Gamma distribution**

**Active parameters:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revmat</td>
<td>1</td>
</tr>
<tr>
<td>Statefreq</td>
<td>2</td>
</tr>
<tr>
<td>Shape</td>
<td>3</td>
</tr>
<tr>
<td>Topology</td>
<td>4</td>
</tr>
<tr>
<td>Brlens</td>
<td>5</td>
</tr>
</tbody>
</table>

**Chain results:**

1 -- [6040.383] [6095.569] [6124.074] [6362.007] [6636.007] [6814.007] [7014.007] [7224.007] [7444.007] [7674.007] [7914.007] [8164.007] [8424.007] [8694.007] [8974.007] [9264.007] [9564.007] [9874.007] [10194.007] [10524.007] [10864.007] [11214.007] [11574.007] [11944.007] [12324.007] [12714.007] [13114.007] [13524.007] [13954.007] [14394.007] [14854.007] [15334.007] [15834.007] [16364.007] [16924.007] [17514.007] [18134.007] [18784.007] [19474.007] [20204.007] [20974.007] [21784.007] [22644.007] [23554.007] [24514.007] [25524.007] [26594.007] [27714.007] [28894.007] [29164.007] [30344.007] [31624.007] [32944.007] [34324.007] [35834.007] [37454.007] [39204.007] [41094.007] [43134.007] [45324.007] [47684.007] [50214.007] [53004.007] [56044.007] [59344.007] [62944.007] [66844.007] [71114.007] [75804.007] [81004.007] [86814.007] [93144.007] [100004.007] [107944.007] [117014.007] [127584.007]

- **Overwrite information in this file (yes/no): yes**
- **Overwriting file "22all.t"**

**File "22all.p" already exists**

- **Overwrite information in this file (yes/no): yes**
- **Overwriting file "22all.p"**

**Chain results:**

<table>
<thead>
<tr>
<th>Generation</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-6040.883</td>
</tr>
<tr>
<td>100</td>
<td>-5553.538</td>
</tr>
<tr>
<td>200</td>
<td>-5349.851</td>
</tr>
<tr>
<td>300</td>
<td>-5196.156</td>
</tr>
<tr>
<td>400</td>
<td>-4987.319</td>
</tr>
<tr>
<td>500</td>
<td>-4863.180</td>
</tr>
<tr>
<td>600</td>
<td>-4811.581</td>
</tr>
<tr>
<td>700</td>
<td>-480.029</td>
</tr>
<tr>
<td>800</td>
<td>-4751.444</td>
</tr>
<tr>
<td>900</td>
<td>-4721.009</td>
</tr>
<tr>
<td>1000</td>
<td>-4688.981</td>
</tr>
<tr>
<td>1100</td>
<td>-4686.537</td>
</tr>
<tr>
<td>1200</td>
<td>-4660.167</td>
</tr>
<tr>
<td>1300</td>
<td>-4636.410</td>
</tr>
<tr>
<td>1400</td>
<td>-4630.269</td>
</tr>
<tr>
<td>1500</td>
<td>-4616.658</td>
</tr>
<tr>
<td>1600</td>
<td>-4584.277</td>
</tr>
<tr>
<td>1700</td>
<td>-4549.489</td>
</tr>
<tr>
<td>1800</td>
<td>-4532.018</td>
</tr>
<tr>
<td>1900</td>
<td>-4519.365</td>
</tr>
<tr>
<td>2000</td>
<td>-4490.951</td>
</tr>
<tr>
<td>2100</td>
<td>-4479.080</td>
</tr>
<tr>
<td>2200</td>
<td>-4448.220</td>
</tr>
<tr>
<td>2300</td>
<td>-4415.675</td>
</tr>
<tr>
<td>2400</td>
<td>-4406.824</td>
</tr>
<tr>
<td>2500</td>
<td>-4401.989</td>
</tr>
</tbody>
</table>

Number of taxa = 22
Number of characters = 1463
Compressing data matrix for division 1
Division 1 has 264 unique sites per taxon
The chain will use the following parameters:
- Model of sequence evolution: DNA, 6 substitutions, Gamma distribution
- Markov chain: No, 4 states
- State frequencies: Gamma
- Rates: Gamma
- Gamma shape parameter: distributed on the Gamma distribution

With each chain, some chains will show slower convergence.
THE PROBLEM OF THE EUKARYA LINEAGE

DOMAIN SHUFFLING
WHAT TO DO THEN?

DOMAIN ANALYSES

CHECK CONSISTENCY BETWEEN DOMAIN DISTRIBUTION AND PHYLOGENETIC DISTRIBUTION

CHECK IF SHUFFLING IS RECENT OR OLD...
REMARKS

DOMAIN ARCHITECTURES

NALP2

MATER

CARD4

NOD2

NAIP

COS1.5

CLAN

NAC

PAAD  NACHT  LRR’S

?  NACHT  LRR’S

CARD  NACHT  LRR’S

CARD  CARD  NACHT  LRR’S

BIR  BIR  BIR  NACHT  LRR’S

?  NACHT  LRR’S

CARD  NACHT  LRR’S

PAAD  ?  NACHT  LRR’S  ?  CARD
NACHT DISTRIBUTION: POSSIBLE SCENARIO

![Diagram showing the distribution and interaction of different species with NACHT-like proteins](image)

**Remarks**

- Bacteria (anabaena)
- Plants
- Fungi
- Invertebrates

**Key Elements**

- CARD
- NACHT
- LRR
- PAAD

**Species Indications**

- Hs
- Mm
- Fugu
- Card

**Diagram Notes**

- NACHT LIKE*
- NB-arc
- TIR
- WD
SOME PRACTICAL EXAMPLES

• DESCRIPTION OF NEW SPECIES
  *Erwinia toletana* sp. nov.

• PLACEMENT OF NEW ISOLATED GENES
  Occurrence of serin proteases in sponge and jellyfish
• DESCRIPTION OF NEW SPECIES

   **Goal:** to obtain a natural antagonist of *P. savastanoi*.

   **Data:** Bacterial species isolated from wild trees’ knots (Olives, oleander...)

   total of 81 bacterial strains!

*(Rojas et al, 2004 IJESM)*
The problem: Resemble phenotypically to several..

What to do?:

• Choose an universal conserved marker: i.e. 16SRNA, Extract similar sequences Build phylogenetic trees

Gene sequencing:
16SRNA, 23SRNA, gnd, mdh

WHY THESE GENES? ?????
METHOD FOR 16SRNA

From 81 sequences only the longest retained (61) stand-alone blasted against a filtered EMBL DB. A total of 19,184 sequences retained (from 80,807 initial sequences).

The 2 most similar are retained to phylogenetic tree reconstruction using Parsimony, Maximum likelihood, and BioNJ with 1000 bootstrap.

CONSENSUS!

(Some examples: Rojas et al, 2004 IJESM)
(Rojas et al, 2004 IJESM)
SOME EXAMPLES

(Rojas et al, 2004 IJESM)
PLACEMENT OF NEW ISOLATED GENES
Ocurrence of serin proteases in sponge and jellyfish

Goal: Confirm the existence of serin proteases in early-divergent phyla, *cnidaria* and *porifera*. Where they come from?

Data: SP are absent in plants, and protists and in fungi are restricted to *Streptomyces*. However, there are hundreds in animals!

(Rojas & Doolittle, 2002, JME)
SOME EXAMPLES

(Rojas & Doolittle, 2002, JME)
SOME EXAMPLES

DIGESTIVE (Rojas & Doolittle, 2002, JME)

NON DIGESTIVE
SOME EXAMPLES

DIGESTIVE (Rojas & Doolittle, 2002, JME)

NON DIGESTIVE
WHICH ONE IS THE REAL ONE?
WHAT IS THE ORIGIN OF THE CHYMOTRYPSIN FAMILY?

ADDITIONAL INFORMATION:

- Sponge has a D189 diagnostic for trypsin (Hannenshalli & Russell, 2000)
  Jelly has N189.

- Codon for Serine at the active site:
  sponge signature for trypsin: TCT
  jelly: AGT, AGC

- When blasted against NR:
  sponge 48% with arthropod trypsin
  jelly 36% with RAT elastase

Disulfide bonds:
- sponge 5 disulfide bonds and cys match with chymotrypsin-elastase (first tree)
Jelly has digestive system with organs, sponge are loose cells.

(Rojas & Doolittle, 2002, JME)
SOME EXAMPLES

(Digestive) (Rojas & Doolittle, 2002, JME)
WHY THE FUNGAL ONES CLADE WITH ANIMALS?

H.G.T!

SCENARIO1

Plants and all fungi-except *Streptomyces* lost it!
Fungi should be more similar to jelly and sponge

then Plants and all fungi never had it.
They appeared when digestion was invented.
Fungi have them because HGT in both directions.

(SOme Examples)
(Rojas & Doolittle, 2002, JME)
Acknowledgements:

Frederik Ronquist for slides I borrowed

THANK YOU!!