Phylogenetic reconstruction

Julie Thompson
Laboratory of Integrative Bioinformatics and Genomics
IGBMC, Strasbourg, France

julie@igbmc.fr
Phylogenetic reconstruction

- Introduction: basic concepts
- Principal methods for tree reconstruction
  - Distance based methods
  - Character based methods
- Evaluation of the reliability of a tree
- The limits of phylogeny
- Some programs
- Methods for phylogenomics
Introduction

- Phylogenetics: the study of **evolutionary** relatedness among **organisms**

  - Relationships represented by a hierarchical, tree-like structure
  - Trees are constructed based on the shared characters
  - Characters are heritable traits that can be compared across organisms
    - physical characteristics (morphology), behavioral traits or genetic sequences

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky, 1973

Julie Thompson -IGBMC
Introduction

- Phylogenetics: the study of evolutionary relatedness among organisms

- What for?
  - Reconstruction of the evolutionary relationships between species: living (extant) and dead (extinct). eg: tree of life
  - Classification of new species. eg: viral strain
  - Reconstruction of evolutionary history of a gene family
Basic concepts: phylogeny

A phylogenetic tree is characterised by:

- its topology (branching patterns)
- the lengths of the branches (possibly)

**Node**: represents a taxonomic unit (species, population, gene,…), either existing or ancestor

**Branch**: defines the relationship between the TUs (descent, ancestry)

**Root**: common ancestor of all taxonomic units on tree
Cladogram

(brand lengths are not meaningful)

Phylogram

(branch lengths are proportional to number of changes)

Unrooted tree
For an unrooted tree, there are several possible rooted trees

In most tree reconstruction programs, the position of the root is chosen arbitrarily:
- « midpoint rooting » (root placed in the centre of the longest branch)
- « outgroup rooting »

The user can define the sequence(s) that can be used as an outgroup to root the tree.
The outgroup sequence should be distantly related to all other sequences.

Julie Thompson - IGBMC
Basic concepts: branch order

The order of the branches belonging to the same node is not important. Rotating branches at a node does not change the topology of the tree.

Tree 1 = Tree 2
**Homology**: 2 genes are homologous if they have a common ancestor

**Orthology**: 2 genes are orthologous if they diverged after a speciation event

**Paralogy**: 2 genes are paralogous if they diverged after a duplication event

---

**Ancestral gene of insulin**

*Homology, orthology, paralogy*
Phylogenetic reconstruction

- Introduction: basic concepts
- Principal methods for tree reconstruction
  - Distance based methods
  - Character based methods
- Evaluation of the reliability of a tree
- The limits of phylogeny
- Some programs

Julie Thompson - IGBMC
Molecular phylogenetics

- DNA, RNA, and protein sequences can be considered as phenotypic traits
- Molecular phylogenetics attempts to determine the rates and patterns of change occurring in the sequences and to reconstruct the evolutionary history of genes and organisms
- How?
  - Choose the set of sequences to study
  - Build a multiple alignment
  - Validate, edit or mask the resulting alignment
  - Reconstruct the tree
  - Evaluate statistically the reliability of the tree
Multiple sequence alignment

- Errors in the initial alignment will lead to inaccurate trees

Alignment of 18s rRNA sequences (Morrison and Ellis, J Mol Evol, 1997)
- alignment algorithms Pileup, ClustalW, TreeAlign, MALIGN, SAM. Trees construction, neighbor-joining, weighted-parsimony, and maximum-likelihood.
- different alignments produced trees that were more dissimilar than did the different tree-building methods

Using genomic data from seven yeast species (Wong et al, Science 2008)
- build alignments using 7 different alignment programs and then estimate phylogeny (using maximum parsimony)
- 46.2% of the 1502 genes had one or more differing trees depending on the alignment procedure used
Alignment masking

- Remove or mask columns with gaps and unreliable regions
  - E.g. Gblocks (Castresana, 2000)

Full length alignment

Gblocks only

Julie Thompson - IGBMC
Removing columns with gaps and variable regions might reduce phylogenetic signal (Dessimoz and Gil *Genome Biology* 2010)
The quality of the multiple alignment is crucial for the quality of the tree!!

The trees can vary depending on the region of the alignment selected.
Figure 7 : Représentation schématique des conservations présentes dans les Asp tRNA synthétases après analyse OrdAli. Colons noirs et grises, résidus strictement conservés ou présents dans au moins 80% des séquences. Colons rouges, bleus, jaunes, régions strictement conservées chez les Eucaryotes, Archaea et Bactéries respectivement. Colons vertes et violets, régions communes aux Eucaryotes et Archaea et aux Archaea et bactéries respectivement.
N terminal region

Bacteria
Archaea
Mitoch.
Tree reconstruction methods

- **Distance based methods**
  - Calculate distance between each pair of sequences => distance matrix
    - UPGMA, Neighbor-Joining
  - => relatively fast and simple
  - => but, the sequences themselves are not taken into account, so we lose information

- **Character based methods**
  - Each position in the sequence is considered as a trait or character
    - maximum parsimony
    - maximum likelihood
    - Bayesian inference
  - => calculation time can be very long
**Observed distance:** mean number of substitutions per site

\[
\text{Observed Dist} = \frac{\text{No. substitutions}}{\text{No. sites considered}}
\]

- **Sites considered:**
  - For DNA, the 3rd base of each codon can be excluded from analysis.
  - Alignment positions with gaps are generally eliminated

2 possibilities:

**Global gap removal:** 18 sites considered
- \(\text{Dist(Seq1,Seq2)} = \frac{3}{18} = 0.1667\)
- \(\text{Dist(Seq1,Seq3)} = \frac{4}{18} = 0.2222\)
- \(\text{Dist(Seq2,Seq3)} = \frac{1}{18} = 0.0556\)

**Pairwise gap removal:**
- \(\text{Dist(Seq1,Seq2)} = \frac{3}{19} = 0.1579\)
- \(\text{Dist(Seq1,Seq3)} = \frac{4}{19} = 0.2105\)
- \(\text{Dist(Seq2,Seq3)} = \frac{1}{18} = 0.0556\)

Total sites = 20

- No. substitutions (Seq1,seq2) = 3
- No. substitutions (Seq1,seq3) = 4
- No. substitutions (Seq2,seq3) = 1

*Julie Thompson - IGBMC*
Distance correction

For more distantly related sequences, the probability of several substitutions occurring at the same site increases.

=> The number of observed substitutions under-estimates the true number of substitutions between distantly related sequences.

<table>
<thead>
<tr>
<th></th>
<th>Sequence1</th>
<th>Sequence2</th>
<th>Observed no. substitutions</th>
<th>True no. substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single substitution</td>
<td>C</td>
<td>C=&gt;A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Multiple substitutions</td>
<td>C</td>
<td>C=&gt;A=&gt;T</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Coincident substitutions</td>
<td>C=&gt;G</td>
<td>C=&gt;A</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Parallel substitutions</td>
<td>C=&gt;A</td>
<td>C=&gt;A</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Convergent substitutions</td>
<td>C=&gt;A</td>
<td>C=&gt;T=&gt;A</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Reverse substitutions</td>
<td>C</td>
<td>C=&gt;T=&gt;C</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

=> Numerous methods available that estimate the true distance between sequences.
Distance correction: DNA substitution models

Jukes-Cantor (JC) model (1969)
- The 4 bases have the same frequency
- All substitutions occur with equal probabilities

\[ d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} D\right) \]

where \( D \) is observed distance

Kimura 2 parameter (K2P/K80) model (1980)
- The 4 bases have the same frequency
- Transitions (substitutions A-G or C-T) are more common than transversions

\[ d = -\frac{1}{2} \ln(1-2P-Q) - \frac{1}{4} \ln(1-2Q) \]

where \( P \) is mean no. of transitions
\( Q \) is mean no. of transversions

Tamura-Nei (TrN) model (1993)
- 4 bases have different frequencies
- Different transition frequencies (\( \alpha_1 \) between purines; \( \alpha_2 \) between pyrimidines), equal transversion frequencies (\( \beta \))
Distance calculation: amino acid substitution models

**PAM matrices (Dayhoff, 1978)**

- A 1-PAM mutation matrix describes an amount of evolution which will change, on the average, 1% of the amino acids. In mathematical terms:
  \[
  \sum_i f_i (1 - M_{ii}) = 0.01
  \]
- estimated from the observation of accepted mutations between 34 superfamilies of closely related sequences
- extrapolated to other evolutionary distances (eg. PAM250)

**Blosum matrices (1992)**

- local, ungapped alignments of distantly related sequences to derive the BLOSUM series of matrices. Matrices of this series are identified by a number after the matrix (e.g. BLOSUM50), which refers to the minimum percentage identity of the blocks of multiple aligned amino acids used to construct the matrix.

**Gonnet matrices (1992)**

- similar to Dayhoff, but with more modern sequence databases
**Neighbor-Joining (NJ)**

**Method:** Find the tree that best fits the distance matrix

1. Start with a star topology

2. For each pair of sequences, create a new node and estimate sum of branch lengths.
   Choose pair that minimises sum of branch lengths

3. Recalculate the distance matrix by treating the joined sequences as one.
4. Repeat steps 2,3 until all taxa are joined and the tree is resolved.
Tree reconstruction methods

- **Distance based methods**
  - Allow different mutation rates between branches
  - Produce unrooted trees
  - Should find correct tree if distances are well estimated (closely related sequences)

- **Character based methods**
  - Each position in the sequence is considered as a trait or character.
    - maximum parsimony
    - maximum likelihood
    - Bayesian inference

  ⇒ calculation time can be very long
Maximum Parsimony

Used historically to study morphological characters

- prefer evolutionary scenario that involves the smallest number of events
(Okham’s razor => find the simplest explanation that works)

Application to sequences:

one column in the alignment = one character

Search for most parcimonious trees, i.e. those for which the topology involves minimum number of substitutions

Steps:

1) Search for all possible topologies
2) find smallest no. of substitutions for each topology (only informative sites)
3) Find topology with minimum substitutions (maximum parsimony)
First step: find all possible topologies

Seq1 AAGAGTGCA
Seq2 AGCCGTGCG
Seq3 AGATATCCA
Seq4 AGAGATCCG

No. of possible topologies (unrooted trees): 3
Second step: find smallest no. of substitutions for a given topology (Fitch method)

1) Root the tree (anywhere)

- Choose one nucleotide \( x \) from set \( N \) of root \( n \)
- For child node \( u \), choose one nucleotide:
  - \( x \) if \( x \in U \)
  - any nucleotide from set \( U \) otherwise

2) From leaves to root:

- Nodes \( u \) and \( v \) are children of \( n \)
- \( U, V \) and \( N \) are sets of nucleotides associated with these nodes

3) From root to leaves:
- Choose one nucleotide \( x \) from set \( N \) of root \( n \)
- For child node \( u \), choose one nucleotide:
  - \( x \) if \( x \in U \)
  - any nucleotide from set \( U \) otherwise

\[ N = U \cap V \text{ if } U \cap V \neq \emptyset \]
\[ N = U \cup V \text{ otherwise} \]
Maximum parsimony

- only include **informative sites**
  
  => sites that support one topology

Seq1: A A G A G T G C A
Seq2: A G C C G T G C G
Seq3: A G A T A T C C A
Seq4: A G A G A T C C G

1 1 1 3
Maximum parsimony

Find tree topology with minimum number of substitutions

**Branch-and-bound Method (Hendy & Penny, 1982)**

- **Exact algorithm, guaranteeing the optimal solution without exhaustive searching**

  1) add sequences one by one (following order of alignment)
     - calculate no. of substitutions for each topology

  2) explore different topologies

     - If no. of substitutions during addition > no. found for best topology ⇒ stop

\[ A \rightarrow B \rightarrow C \]

---

*Julie Thompson - IGBMC*
Maximum parsimony

Find tree topology with minimum number of substitutions

Heuristic methods can handle more sequences, but do not guarantee optimal solution

1) construct initial tree by progressively adding TUs
2) then, rearrange initial tree to reduce length (branch swapping)

result depends on initial order of sequences

perform several tests with different initial ordering: option « jumble »
Problem: long branch attraction

If the sequences are evolving at very different rates, the probability of convergent substitutions is significant in long branches => The most parsimonious clustering can lead to false topology

- Implicit evolutionary model: all changes occur at equal rates, no correction for multiple substitutions
- Can lead to several equally parsimonious trees
- Relatively slow, not suitable for a large number of sequences
- No information about branch lengths (generally)
Maximum likelihood

⇒ maximizes the probability that a given tree could have produced the observed data (that is, the likelihood)

- As for parsimony method:
  - Each column is considered to be a character
  - All possible trees are considered
  - For trees requiring many mutations, probability is low
    => trees requiring few mutations are preferred

- Differences:
  - Use of an explicit evolutionary model
  - Allows variable substitution rates for each branch

- Can be used to estimate reliability of tree
Alignment of 4 sequences

| Seq a | TTGC... |
| Seq b | TTGC... |
| Seq c | ATAC... |
| Seq d | GTAC... |

Maximum likelihood

3 possible unrooted trees

For position 1, possible combinations of bases at each node:
- 3 internal nodes
- 4 possible bases at each node

No. of possible combinations: $4 \times 4 \times 4 = 64$
**Maximum likelihood**

*Example of evolutionary model:*

L0 = frequency of T \( \sim 0.25 \)

L2 = probability of transversion of T\( \Rightarrow \)G

L5 = probability of transition of G\( \Rightarrow \)A

L1, L3, L4, L6 = \( \sim 1 \)

Probability of this combination for position 1:

\[ L = L_0 \times L_1 \times L_2 \times L_3 \times L_4 \times L_5 \times L_6 \]

---

*Estimation of likelihood of tree 1 for position 1:
sum of probabilities for each of 64 combinations*

*Estimation de likelihood of tree 1:
Sum of probabilities obtained for each position*

*The calculation is performed for all possible tree (3 here).
The tree with the maximum likelihood is selected.*

---

<table>
<thead>
<tr>
<th>No of</th>
<th>No of possible unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
</tr>
<tr>
<td>50</td>
<td>( &gt;3 \times 10^{74} )</td>
</tr>
</tbody>
</table>

*Julie Thompson - IGBMC*
Bayesian inference

Use Bayes's theorem to combine the prior probability of a phylogeny with the likelihood to produce a posterior probability distribution on trees:

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

- Prior probability of a tree before observations are made (generally, all trees are equally probable)

- Likelihood is proportional to probability of the observations, conditional on the tree (makes assumptions about processes generating observations)

- Posterior probability of a tree is the probability of the tree conditional on the observations, obtained by combining prior and likelihood for each tree

From Huelsenbeck et al, 2001

Bayesian inference

- Computation of probabilities is generally not possible

- Some numerical methods are available that allow the posterior probability of a tree to be approximated

- E.g. Markov chain Monte Carlo (MCMC) in MrBayes is used to generate a statistical sample from the posterior distribution of trees (Huelsenbeck et al, 2001)
  - Credible sets of trees: include trees in order of decreasing probability to obtain, e.g. 95% credible set
  - Choose tree with the highest posterior probability as the best estimate of phylogeny
Phylogenetic reconstruction

- Introduction: basic concepts
- Principal methods for tree reconstruction
  - Distance based methods
  - Character based methods
- Evaluation of the reliability of a tree
- The limits of phylogeny
- Some programs
Evaluation of the reliability of a tree

- **Goal:**
  Statistically estimate the reliability of a given tree topology

- **Example: bootstrapping**
  
  Build $n$ pseudo-alignments by random sampling of the columns in the initial alignment
  
  - each column can be used 0, 1 or more times
  - the pseudo-alignments have the same length as the initial alignment
  - the number of pseudo-alignments should allow for significant statistical testing ($n \geq$ number of columns)
  - for each pseudo-alignment, build a tree
  - for each branch in the initial tree, count the number of times this branch is found in the $n$ trees
Evaluation of the reliability of a tree

Initial alignment

| Seq A | A G G C T C C A A A |
| Seq B | A G G T T C G A A A |
| Seq C | A G C C C C G A A A |
| Seq D | A T T T C C G A A C |

| Seq A | G G G T T T C A A A |
| Seq B | G G G T T T G A A A |
| Seq C | G C C C C C G A A A |
| Seq D | T T T C C C G A A C |

Random sampling

1

| Seq A | A T T C C C C A A A |
| Seq B | A T T C C C G A A A |
| Seq C | A C C C C C G A A A |
| Seq D | A C C C C C G C C C |

| Seq A | A G T T C C C A A A |
| Seq B | A G T T C C C G A A A |
| Seq C | A G C C C C C G A A A |
| Seq D | A T C C C C G A C C |

2

| Seq A | A T T C C C C A A A |
| Seq B | A T T C C C G A A A |
| Seq C | A C C C C C G A A A |
| Seq D | A C C C C C G C C C |

| Seq A | A G T T C C C A A A |
| Seq B | A G T T C C C G A A A |
| Seq C | A G C C C C C G A A A |
| Seq D | A T C C C C G A C C |

3

| Seq A | A T T C C C C A A A |
| Seq B | A T T C C C G A A A |
| Seq C | A C C C C C G A A A |
| Seq D | A C C C C C G C C C |

| Seq A | A G T T C C C A A A |
| Seq B | A G T T C C C G A A A |
| Seq C | A G C C C C C G A A A |
| Seq D | A T C C C C G A C C |

Julie Thompson -IGBMC
Phylogenetic tree of the ribosomal proteins L2p

bootstrap value of a node is the percentage of times that node is present in the trees built from the random samples

Archaea

Eucarya

Bacteria
Phylogenetic reconstruction

- Introduction: basic concepts
- Principal methods for tree reconstruction
  - Distance based methods
  - Character based methods
- Evaluation of the reliability of a tree
- The limits of phylogeny
- Some programs
Phylogenetic tree-building models make certain assumptions:

- The sequences are homologous (descended from a shared ancestral sequence)
- Each of the sequences has a common phylogenetic history with the other sequences
- At each position in the alignment, the characters are homologous with each other
- The sequence variability in the sample contains phylogenetic signal adequate to resolve the problem under study.

Careful selection of sequences and evolutionary model
Some programs…

Software suites:

- **Phylip** (very comprehensive)
- **PAUP** (Phylogenetic Analysis Using Parsimony)
  http://www.lms.si.edu/PAUP/about.html
- **TREE-PUZZLE**
  http://www.tree-puzzle.de/
- **phylowin** (graphical interface)
  http://pbil.univ-lyon1.fr/software/phylowin.html

Integrated workbenches

- **MEGA** http://www.megasoftware.net/
- **Mesquite** http://mesquiteproject.org/

Visualisation, manipulation

- **Njplot, baobab, treeedit, phylodendron**…
- **Treeview**
  http://taxonomy.zoology.gla.ac.uk/rod/treev
- **DensiTree** (sets of trees, e.g. Ba)
  http://compevol.auckland.ac.nz/software

Interactive Tree Of Life (iTOL)
an online tool for phylogenetic tree display and annotation
http://itol.embl.de/

No algorithm is perfect

- It is never certain that the reconstructed tree is the real one!
- The same data can result in different trees depending on the algorithm used
Phylogenomics

Construction of species trees: problem

- The evolutionary history of a single gene family is not always transposable to the species
  - Not all genes evolve at the same rate (different selection pressure)
  - Gene duplication/loss
  - Horizontal transfer
  - Convergent or parallel evolution

Solution

- Integrate the phylogenetic information from different gene families to form a single species phylogeny
Construction of species trees

- Define groups of orthologous sequences
- Then use:
  - Whole genome features (complete genome alignment, gene content)
  - Supermatrix (simultaneous-analysis, combined-analysis)
  - Supertree (separate analysis)

Delsuc et al, Nature reviews, 2005
Supermatrix (superalignment)

- multiple alignments for each gene are concatenated to form a superalignment
- Use conventional phylogenetic reconstruction methods (e.g. distance or MP)  
- Example: RibAlign
- analysis of 16S ribosomal RNA (rRNA) sequences has been the de-facto gold standard for the assessment of phylogenetic relationships among prokaryotes
- concatenation of ribosomal protein sequences (MAFFT, Phylip: ProML, MrBayes)
Supertree

- Reconstruct phylogenetic trees for each gene family separately

- Combine the multiple gene family trees to form a single phylogenomic tree (Gene Tree Reconciliation)

(Bininda-Emonds, 2004; Daubin et al., 2002)
Gene tree reconciliation methods

- Consensus tree methods are used to combine fully overlapping source trees (strict, majority consensus rules, …)
  - (eg. Mincut Semple and Steele 2000)

de Queiroz and Gatesy, Trends Ecol Evol, 2007

Julie Thompson - IGBMC
Gene tree reconciliation methods

- Indirect supertree construction represents individual source trees as matrices, then combines them using an optimization criterion:
  - Matrix representation using parsimony (MRP)
  - “flip” supertrees
  - Average consensus procedure
  - Most Similar Supertree (MSSA)
  - Maximum Quartet Fit (QFIT)
  - Maximum Splits Fit (SFIT).

From Bininda-Emonds et al, 2002
Software Clann, http://bioinf.may.ie/software/clann/
Problems

- Large amounts of data: need automatic pipelines
- Need a reliable method to identify genuine orthologues
- Missing data: some genes missing from some species (incomplete sequencing)
- Factors leading to an incorrect tree, even with use of genome-scale data:
  - nucleotide or amino acid compositional bias
  - long-branch attraction caused by unequal evolutionary rates among lineages
  - sparse taxon sampling
  - heterotachy (the shift of position specific evolutionary rates)
Comparison supermatrix/supertree

**Supermatrix methods**

- Include all sequence information (reduces noise)
- Can yield relationships that are not present in the set of source trees
- Ignore differences in rates or modes of evolution
- More sensitive to missing data
- Computationally expensive

**Supertree methods**

- Relatively efficient => allow construction of large trees
- Estimate an independent set of parameters for every gene
- Allow incorporation of diverse kinds of data, e.g. characters from fossils, morphobank
- Less sensitive to missing data
- Use heuristic algorithms that cannot be justified rigorously on a statistical basis.
- Ignore uncertainties in the subtrees (bootstrap values, Bayesian posterior probabilities, …) but some recent algorithms may solve this problem (Burleigh, 2006; Moore, 2006)
- May over-fit the data and cause large variances in the estimates

*Julie Thompson - IGBMC*
Statistical modelling approach

- statistical likelihood provides a framework for combining information from different experiments
- combine data from multiple genes while accommodating differences in the evolutionary process
- define a model that estimates the probability of obtaining a series of subtree topologies, given a hypothesized supertree

select supertree that maximizes the likelihood (product of likelihoods of all subtrees)

Applications: tree of life

- Mammalian tree topology
  - 70 mammalian species, plus Marsupialia and Monotremata as outgroups
  - Supermatrix approach using 1st, 2nd codon positions of mitochondrial protein-coding genes and MrBayes

Applications: tree of life

- Current status and future challenges

Nature Reviews | Genetics

Julie Thompson - IGBMC