Comparative Genomics

trees/phylogenies/genomes

Berend Snel
Theoretical Biology & Bioinformatics, Department of Biology, Faculty of Science

Some of the slides in this lecture are courtesy of Jaap Heringa, Anders Gorm Pedersen, Can Kesmir, Bas Dutilh, Geert Kops and Michael Rosenberg

General idea of today

- Having all the genomes …
- What is the relation of my gene to homologous genes in the same and other organisms

Today

- What is a phylogenetic tree?
  - History
  - What can you do with a phylogenetic tree?
- How to “read” simple phylogenetic trees
- How can you make a phylogeny?
- How can you root a phylogeny?
- How to interpret a phylogenetic tree
  - Duplications
- (Genome Duplications -> evolutionary Genomics)

Tree: depiction (formalization) of hierarchical classification

NB no information in skunk left / otter right
Theory of evolution

Charles Darwin
1809-1882

The only figure in Darwin’s “On the Origin of Species” is...

Phylogenetic tree: historical pattern of relationships among organisms: interpretation of a tree

(molecular) Phylogenetic insights changed taxonomy

Divergence (= speciation) between skunk and otter
Ancestor of all extant carnivores
NB still no information in skunk left / otter right

cladogram
Today

- What is a phylogenetic tree?
- How to “read” simple phylogenetic trees
  - Types of trees
  - Unrooted vs rooted
  - Molecular clock vs no molecular clock
- How to make a phylogeny
- How to root a phylogeny
- Afternoon Lecture

Rooted trees with molecular clock

![Rooted Trees Diagram]

Evolutionary clock speeds

- Uniform clock: leads to identical distances from root to leaves (ultrametric tree)
- Non-uniform evolutionary clock: leaves have different distances … (additive tree)
- Root unknown: unrooted trees

**Figure 25.13 Campbell & Reece**
Unequal rates between species are a very real phenomenon

- Phylogram

No molecular clock means that a phylogenetic reconstruction method will infer only relations and no direction

- We “loose” flow of time
- Unambiguous representation: Unrooted tree
- NB most methods infer unrooted trees (because distances / process are modelled non-directional)

Radial tree (always explicitly unrooted)
Phylogram (explicitly rooted)
Rooted = directed tree

Introduce a root to go from unrooted to rooted (or vice versa)
One unrooted tree can be turned into multiple rooted trees

Exercise 1

Today

- What is a phylogenetic tree?
- How to “read” simple phylogenetic trees
- How to make a phylogeny
  - Distance methods
  - Parsimony
- How to root a phylogeny
- Afternoon Lecture

Trees vs blast, phylogeny vs homology

- Blast/hmm/psi-blast tell you
  - How likely it is that two (parts) of a sequence are homologous or not (and how high the similarity between a profile and a sequence of between two sequences is)
  - Which portions of the sequences are significantly similar; which section of which sequence is homologous to which section of which other sequence.
  - Homologous is a yes/no thing
- Trees/phylogeny tell you
  - How the sequences are related, i.e. In which order they diverged
How to make a molecular phylogenetic tree

1) Alignment
   - 2a) Distances
   - 3a) Clustering

2) Explit model of sequence evolution plus best fitting tree, choice between Parsimony and Maximum likelihood

Clustering algorithm: UPGMA (assumes ultrametric trees)

Initialisation:
- Fill distance matrix with pairwise distances
- Start with N clusters of 1 element (gene) each

Iteration:
- Merge cluster C_i and C_j for which d_ij is minimal
- Place internal node connecting C_i and C_j at d_ij/2
- Delete C_i and C_j; replace by new C with group average distances

Termination:
- When only two clusters i, j remain, put root at d_ij/2

Phylogenetic tree by Distance methods (Clustering)

- Multiple alignment
- Evolutionary Distance matrix
- Phylogenetic tree

UPGMA

Iteration:
- Merge cluster C_i and C_j for which d_ij is minimal
- Place internal node connecting C_i and C_j at d_ij/2
- Delete C_i and C_j; replace by new C with group average distances

Termination:
- When only two clusters i, j remain, put root at d_ij/2
How to make a molecular phylogenetic tree

1) Alignment
   - 2a) Distances
   - 3a) Clustering

2b) Explicit model of sequence evolution plus best fitting tree, choice between Parsimony and Maximum likelihood

Multiple sequence alignment

Model based approaches

Find best fit to evolutionary model

Maximum parsimony (MP) and likelihood (ML)

- Maximum parsimony (MP): the tree that requires the fewest evolutionary events to explain the alignment
  - Occam’s razor: the simplest explanation of the observations
- Maximum likelihood (ML): the tree most likely to have led to the alignment given a certain model of evolution (most applied strategy)

Maximum parsimony (MP)

- MP example for a single position “alignment” in 5 species:
  - Chimp
  - Gibbon
  - Gorilla
  - Human
  - Orangutan

- Draw all possible trees for the sequences/species present in your multiple alignment
- For each tree, identify where the mutations have taken place
- Make a parsimony assumption; minimum number of required mutations
Maximum parsimony (MP)

- How many trees are there?
  - # unrooted trees $N_U = (2n - 5)!!$
  - # rooted trees $N_R = (2n - 3)!!$

- E.g.
  - For 5 species 15 unrooted trees possible,
  - for 50 species, $2.84 \times 10^{74}$ unrooted trees possible

- (do you know how many species exist?)

- For parsimony & maximum-likelihood phylogeny: “heuristic searches”

Most parsimonious tree

The MP tree has the minimum number of required mutations

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>c</td>
<td>c</td>
<td>t</td>
<td>c</td>
<td>t</td>
<td>g</td>
<td>a</td>
</tr>
<tr>
<td>Chimp</td>
<td>c</td>
<td>c</td>
<td>t</td>
<td>g</td>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Gorilla</td>
<td>c</td>
<td>c</td>
<td>t</td>
<td>a</td>
<td>g</td>
<td>t</td>
<td>a</td>
</tr>
<tr>
<td>Gibbon</td>
<td>t</td>
<td>c</td>
<td>a</td>
<td>a</td>
<td>g</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>Orangutan</td>
<td>t</td>
<td>c</td>
<td>a</td>
<td>g</td>
<td>a</td>
<td>t</td>
<td></td>
</tr>
</tbody>
</table>

Chimp

Gibbon

Orangutan

NB unrooted tree! = Mutation modelled in two directions

Maximum likelihood

- If data = alignment, hypothesis = tree, and under a given evolutionary model:
  - compute “likelihood” that the hypothesis (=tree), given a model (e.g. substitution matrix), results in the observed data (= multiple sequence alignment).
  - maximum likelihood selects the hypothesis (tree) that maximises the observed data

- CPU intensive method

- Best approach to find the “true” tree

Parsimony, Maximum Likelihood or Neighbor-joining?

1. ML (PhyML, RaxML, IQ-tree) and bayesian methods (MrBayes and PhyloBase) are thought to be most accurate
2. Data is of greater importance than method
3. one must remember that a phylogenetic tree is a hypothesis of the true evolutionary history.
4. As a hypothesis it could be right or wrong or a bit of both.
Today

- What is a phylogenetic tree?
- How to “read” simple phylogenetic trees
- How to make a phylogeny
- How to root a phylogeny
- Afternoon Lecture

Unrooted trees

- Problem for interpretation in what order did my species diverge a tree without flow of time
- Representation: Unrooted tree
- NB most methods infer unrooted trees (because distances / process are not modeled directional)

Introduce a root

- root
- edge
- internal node
- leaf

How to root a tree: outgroup

- Human ccttgaa
- Frog ccttgat
- Lizard ccttgac

Human ccttgaa
Frog ccttgat
Lizard ccttgac
Insect aattgat
How to root a tree

- Outgroup – place root between distant (still homolog) sequence and rest group
- Midpoint – place root at midpoint of longest path (sum of branches between any two leafs)
- Gene duplication – place root between paralogous gene copies

Implication of the fact that rooting is (like) a display choice

- If a tree is not rooted by the method (e.g. ML, NJ, MP), you are free to root it yourself … as long as you explain where you rooted it (and why)
This lecture was about reading trees and how they are constructed …

After practicing with this we will have a lecture on recognizing gene duplications and gene loss in a gene tree (important for comparative genomic of eukaryotes)

Thinking in trees also very relevant for recognizing horizontal gene transfer, host-pathogen co-evolution and many more things …

Exercises 2 & 3

Further reading