Introduction to Bioinformatics

Bioinformatics

GUEST LECTURE:
Phylogenetic Analysis

11 December 2012, Université de Liège

Ronald Westra,
Biomathematics Group,
Maastricht University
Overview

1. Introduction
2. On trees and evolution
3. Inferring trees
4. Combining multiple trees
5. Case study: the phylogenetic analysis of SARS
6. References and recommended reading
On trees and evolution

* Traditionally, the evolutionary history connecting any group of (related) species has been represented by an evolutionary tree

* The analysis of the evolutionary history involving evolutionary trees is called Phylogenetic Analysis
Nothing in Biology makes sense except in the light of Evolution, and in the light of evolution everything in Biology makes perfectly sense. (Theodosius Dobzhansky)
The only *figure* in Darwin’s “On the origin of species” is a tree.
The biological basis of evolution

Mother DNA: tctgcctc

tctgcctc  gatgcctc  tctgcctc

Present species

phylogenetic trees
PHYLOGENETIC TREES

Phylogenetics

phylogenetics is the study of evolutionary relatedness among various groups of organisms (e.g., species, populations).
Visualizing phylogenetic relations

Multi-Dimensional Scaling (MDS map)

Dendrogram
Visualizing phylogenetic relations
On trees and evolution

* Normal procreation of individuals is via a tree

* In case of horizontal gene transfer a **phylogenetic network** is more appropriate → Presentation of Steven Kelk
From phylogenetic data to a phylogenetic tree

1. Homology vs homoplasy, and orthologous vs paralogous

2. Sequence alignment (weights)

3. Multiple substitutions: corrections

4. (In)dependence and uniformity of substitutions

5. Phylogenetic analysis: tree, timing, reconstruction of ancestors
Character and Distance

A phylogenetic tree can be based on

1. based on **qualitative aspects** like **common characters**, or

2. **quantitative measures** like the **distance** or **similarity** between species or number of acquired mutations from last common ancestor (LCA).
Character based comparison

- character 1
- character 2
- character 3

Non-numerical data: has/has not

Figure 1 Comparison of apheliscine and macroscelidean long bones.
Constructing Phylogenetic Trees

There are three main methods of constructing phylogenetic trees:

* **character-based methods** such as maximum likelihood or Bayesian inference,

* **distance-based methods** such as UPGMA and neighbour-joining, and

* **parsimony-based methods** such as maximum parsimony.

**Parsimony** is a 'less is better' concept of frugality, economy, stinginess or caution in arriving at a hypothesis or course of action. The word derives from Latin *parsimonia*, from *parcere*: to spare.
Cladistics

As treelike relationship-diagrams called "cladogram" is drawn up to show different hypotheses of relationships.

A cladistic analysis is typically based on morphological data.

This traditionally is character based.
Cladistics: tree of life
Phylogenetic Trees

A phylogenetic tree is a tree showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. A phylogenetic tree is a form of a cladogram. In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and edge lengths correspond to time estimates.

Each node in a phylogenetic tree is called a taxonomic unit. Internal nodes are generally referred to as Hypothetical Taxonomic Units (HTUs) as they cannot be directly observed.
Rooted and Unrooted Trees

A **rooted phylogenetic tree** is a directed tree with a unique node corresponding to the (usually imputed) **most recent common ancestor** of all the entities at the leaves of the tree.
Phylogenetic Tree of Life

**LCA** = last common ancestor
Rooted and **Unrooted** Trees

**Unrooted phylogenetic trees** can be generated from rooted trees by omitting the root from a rooted tree, a root cannot be inferred on an unrooted tree without either an outgroup or additional assumptions.
Unrooted Phylogenetic Tree

PHYLOGENETIC TREES
Trees and Branch Length

A tree can be a branching tree-graph where branches indicate close phylogenetic relations.

Alternatively, branches can have length that indicate the phylogenetic closeness.
Tree without Branch Length
Tree with Branch Length
On trees and evolution

* Relation between “taxa”

* Internal nodes and external nodes (leafs)

* Branches connects nodes

* Bifurcating tree: **internal** nodes have **degree**: 3, **external** nodes degree: 1, root **degree**: 2.

* Root connects to ‘outgroup’

* Multifurcating trees
ON TREES AND EVOLUTION

- **root**
- **branch**
- **internal node**
- **external node**
ON TREES AND EVOLUTION

unrooted tree

*Where is the root?*
* Any rotation of the internal branches of a tree keeps the phylogenetic relations intact
rotation invariant
ON TREES AND EVOLUTION

Number of possible trees

* $n$ is number of taxa

* # unrooted trees for $n > 2$: $(2n - 5)!/(2^{n-3}(n-3)!)$

* # rooted trees for $n > 1$: $(2n - 3)!/(2^{n-2}(n-2)!)$

* $n = 5$: #rooted trees = 105

* $n = 10$: #rooted trees = 34,459,425
Representing trees

* Various possibilities

* Listing of nodes

* \( n \) taxa = \( n \) external nodes: \( (n - 1) \) internal nodes

* internal nodes with children: \( (n - 1) \times 3 \) matrix

* \(( \text{internal node}, \text{daughter}_1, \text{daughter}_2)\)

* Newick format: see next slide for example
ON TREES AND EVOLUTION

Newick format: (((1,2),3),((4,5),(6,7)))
Under **parsimony**, the preferred phylogenetic tree is the tree that requires the **least evolutionary change** to explain some observed data.

Given a family of trees $T(\theta)$ with minimum substitutions $n(i,j|\theta)$ between branches $i$ and $j$:

$$\theta^* = \min \Sigma n(i,j|\theta)$$

The obtained result is the **maximum parsimonious tree**.
The aim of maximum parsimony is to find the shortest tree, that is the tree with the smallest number of changes that explains the observed data.

Example:

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence1</td>
<td>T</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>Sequence2</td>
<td>T</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Sequence3</td>
<td>A</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Sequence4</td>
<td>A</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

1. draw all the possible trees
2. consider each position separately
3. find tree with fewest changes to explain data

(1,2): 4
(1,3): 5
(1,4): 6
So: shortest tree : ((1,2)(3,4))
INFERRING TREES

PARSIMONY

* Real evolution may have more substitutions!
* So maximum parsimonious tree is a lower bound on the evolution
Inferring distance based trees

* input: distance table

QUESTION: which distances ?!
Estimating genetic distance

* Substitutions are independent (?)
* Substitutions are random
* Multiple substitutions may occur
* Back-mutations mutate a nucleotide back to an earlier value
Multiple substitutions and Back-mutations conceal the real genetic distance

PHYLOGENETIC ANALYSIS

evolutionary time

observed : 2 (= d)
actual : 4 (= K)
The **actual** genetic distance $K$ for an **observed** gene-gene dissimilarity $d$ is the **Jukes-Cantor formula**:

$$K \approx -\frac{3}{4} \ln\left(1 - \frac{4}{3}d\right)$$
Jukes-Cantor

\[-\frac{3}{4} \log(1 - 4d/3)\]
Inferring trees

* $n$ taxa \( \{t_1, \ldots, t_n\} \)

* $D$ matrix of pairwise genetic distances + JC-correction

* **Additive** distances: distance over path from $i \rightarrow j$ is: $d(i,j)$

* (total) length of a tree: sum of all branch lengths.
Ultrametric trees:

If the distance from the root to all leafs is equal the tree is ultrametric

Ultrametricity must be valid for the real tree, but due to noise this condition will in practice generate erroneous trees.
INFERRING TREES

Ultrametric - Minimum length
INFERRING TREES

MINIMUM LENGTH TREE

Find phylogenetic tree with minimum total length of the branches

Given a family of trees $T(\theta)$ with branch length $\lambda(i,j|\theta)$ between nodes $i$ and $j$ – and genetic distance $d(i,j)$

$L^* = \min \sum \lambda(i,j|\theta)$ subject to $\lambda(i,j|\theta) \geq d(i,j|\theta) \geq 0$

The obtained result is the minimum length tree

This looks much like the maximum parsimonious tree
NEIGHBOR JOINING algorithm:

Popular distance-based clustering method

Iteratively combine closest nodes
Finding Branche lengths:

Three-point formula:

\[ L_x + L_y = d_{AB} \]
\[ L_x + L_z = d_{AC} \]
\[ L_y + L_z = d_{BC} \]

\[ L_x = \frac{(d_{AB} + d_{AC} - d_{BC})}{2} \]
\[ L_y = \frac{(d_{AB} + d_{BC} - d_{AC})}{2} \]
\[ L_z = \frac{(d_{AC} + d_{BC} - d_{AB})}{2} \]
INFERRING TREES

Four-point formula:

\[ d(1,2) + d(i,j) < d(i,1) + d(2,j) \]

\[ R_i = \sum_j d(t_i, t_j) \]

\[ M(i,j) = (n-2)d(i,j) - R_i - R_j \]

\[ M(i,j) < M(i,k) \text{ for all } k \text{ not equal to } j \]

When \((1,2)\) and \((i,j)\) are neighbor couples!

Minimize \(d(i,j)\) AND total distance in tree

If \(i\) and \(j\) are neighbours!
NEIGHBOR JOINING algorithm:

Input: \(nxn\) distance matrix \(D\) and an outgroup
Output: rooted phylogenetic tree \(T\)

**Step 1:** Compute new table \(M\) using \(D\) – select smallest value of \(M\) to select two taxa to join

**Step 2:** Join the two taxa \(t_i\) and \(t_j\) to a new vertex \(V\) - use 3-point formula to calculate the updates distance matrix \(D'\) where \(t_i\) and \(t_j\) are replaced by \(V\).

**Step 3:** Compute branch lengths from \(t_k\) to \(V\) using 3-point formula, \(T(V,1) = t_i\) and \(T(V,2) = t_j\) and \(TD(t_i) = L(t_i,V)\) and \(TD(t_j) = L(t_j,V)\).

**Step 4:** The distance matrix \(D'\) now contains \(n - 1\) taxa. If there are more than 2 taxa left go to step 1. If two taxa are left join them by an branch of length \(d(t_i,t_j)\).

**Step 5:** Define the root node as the branch connecting the outgroup to the rest of the tree. (Alternatively, determine the so-called “mid-point”)
INFERRING TREES

UPGMA and ultrametric trees:

For ultrametric trees use $D$ instead of $M$ and the algorithm
is called UPGMA (Unweighted Pair Group Method).

EVALUATING TREES

- (un)decidability
- Hypothesis testing: models of evolution
- Using numerical simulation
CONSENSUS TREES

Different genes/proteins can/will give different trees
OTHER APPLICATIONS

Language families
Fig. 2.6.2 The genetic tree comparing linguistic families and superfamilies published in Cavalli-Sforza et al. (1988). Populations pooled on the basis of linguistic classification belong to the following groups: Bantu, Niger-Kordofanian family; Nilotic, Nilo-Saharan family; Southeast Indian, Dravidian family; Samoyeds, Uralic family from Russia; North Turkic, branch of Altaic family; Northwest Amerind, Na-Dene family. The genetic tree was constructed by average linkage analysis of Nei’s genetic distances and is the same as that of figure 2.3.2A.
CASE STUDY:

Phylogenetic Analysis of the 2003 SARS epidemic
**SARS: the outbreak**

* February 28, 2003, Hanoi, the Vietnam French hospital called the WHO with a report of an influenza-like infection.

* Dr. Carlo Urbani (WHO) came and concluded that this was a new and unusual pathogen.

* Next few days Dr. Urbani collected samples, worked through the hospital documenting findings, and organized patient quarantine.

* Fever, dry cough, short breath, progressively worsening respiratory failure, death through respiratory failure.
SARS: the outbreak

* Dr. Carlo Urbani was the first to identify Severe Acute Respiratory Syndrome: SARS.

* In three weeks Dr. Urbani and five other healthcare professionals from the hospital died from the effects of SARS.

* By March 15, 2003, the WHO issued a global alert, calling SARS a worldwide health threat.
Dr. Carlo Urbani (1956-2003)
WHO

Hanoi, the Vietnam French hospital, March 2003
Origin of the SARS epidemic

* Earliest cases of what now is called SARS occurred in November 2002 in Guangong (P.R. of China)

* Guangzhou hospital spread 106 new cases

* A doctor from this hospital visited Hong Kong, on Feb 21, 2003, and stayed in the 9th floor of the Metropole Hotel

* The doctor became ill and died, diagnosed pneumonia

* Many of the visitors of the 9th floor of the Metropole Hotel now became disease carriers themselves
Origin of the SARS epidemic

* One of the visitors of the 9th floor of the Metropole Hotel was an American business man who went to Hanoi, and was the first patient to bring SARS to the Vietnam French hospital of Hanoi.

* He infected 80 people before dying

* Other visitors of the 9th floor of the Metropole Hotel brought the disease to Canada, Singapore and the USA.

* By end April 2003, the disease was reported in 25 countries over the world, on 4300 cases and 250 deaths.
PHYLOGENETIC TREES

SARS panic & Mediahype, April-June 2003
SARS: the outbreak

**The SARS corona virus**

* Early March 2003, the WHO coordinated an international research.

* End March 2003, laboratories in Germany, Canada, United States, and Hong Kong independently identified a novel virus that caused SARS.

* The SARS corona virus (SARS-CoV) is an RNA virus (like HIV).

* Corona viruses are common in humans and animals, causing ~25% of all upper respiratory tract infections (e.g. common cold).
SARS: the outbreak

SARS STATISTICS

PEOPLE PROBABLY INFECTED

MORTALITY RATE

DATE

20-Mar 4-Apr 19-Apr 4-May 19-May 3-Jun 18-Jun 3-Jul

CHINA
HONG KONG
CANADA
TAIWAN
OTHER
TOTAL

MORTALITY RATE

AVERAGE MORTALITY RATE

Last Update July 7th, 2003
SARS: Cumulative Number of Reported Probable* Cases

Total number of cases: 2722 as of 9 Apr 2003, 15:00 GMT+2

*Due to differences in the case definitions being used at a national level, probable cases are reported by all countries except the United States of America, which is reporting suspect cases under investigation.

Cumulative number of Reported Cases
(From 1 November 02 to 9 April 03)

1 101 - 1000
2 - 10 11 - 100
>1000 no local transmission

Type of transmission

Data Source: World Health Organization
Map Production: Public Health Mapping Team
Communicable Diseases (CDS)
© World Health Organization, April 2003

The presentation of material on the maps contained herein does not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or areas or of its authorities, or concerning the delimitation of its frontiers or boundaries.
SARS: the outbreak

The SARS corona virus
The SARS coronavirus
SARS: the outbreak

The SARS corona virus

[Diagram showing the interaction between the SARS virus and the human cell membrane via the ACE-2 receptor]
The SARS corona virus

* April 2003, a laboratory in Canada announced the entire RNA genome sequence of the SARS CoV virus.

* Phylogenetic analysis of the SARS corona virus showed that the most closely related CoV is the *palm civet*.

* The palm civet is a popular food item in the Guangdong province of China.
SARS: the outbreak

Palm civet as Chinese food item

Palm civet alive
SARS: the outbreak

Phylogenetic analysis of SARS CoV

* May 2003, 2 papers in Science reported the full genome of SARS CoV.

* Genome of SARS CoV contains 29,751 bp.

* Substantially different from all human CoVs.

* Also different from bird CoVs – so no relation to bird flue.

* End 2003 SARS had spread over the entire world
Phylogenetic analysis of SARS CoV

Phylogenetic analysis helps to answer:

* What kind of virus caused the original infection?

* What is the source of the infection?

* When and where did the virus cross the species border?

* What are the key mutations that enabled this switch?

* What was the trajectory of the spread of the virus?
Case study: phylogenetic analysis of the SARS epidemic

* Genome of SARS-CoV: 6 genes
* Identify host: Himalayan Palm Civet
* The epidemic tree
* The date of origin
* Area of Origin
phylogenetic analysis of SARS: Identifying the Host

Neighbor-joining tree using Jukes-Cantor model
phylogenetic analysis of SARS: The epidemic tree
**phylogenetic analysis of SARS : Area of origin**

Multidimensional scaling - Genetic distance between SARS spike genes

Largest variation in Guangzhou provence
phylogenetic analysis of SARS: Date of origin

95% confidence interval: 16 September 2002.

The genetic distance of samples from the palm civet increases +/- linearly with time.
Newick format: (((1,2),3),((4,5),(6,7)))
REFERENCES AND RECOMMENDED READING

GENERAL:


Computational Genomics, a case study approach, Nello Christianini, Matthew Hahn, Cambridge University press, Cambridge UK, 2007

APPLY AND USE:


MATHEMATICAL BACKGROUND:

END of LECTURE