After this lecture, you can...

... recognize different types of nodes in phylogenetic trees
... interpret branch lengths in the light of evolutionary time
... collapse branches with low support values
... rotate branches and interpret radial and polar trees
... infer the phenotype of ancestors using parsimony
... distinguish orthologous and paralogous genes in a tree
... explain how gene functions evolve
... transfer functional annotations between genes
... reconcile gene trees and species trees
... identify gene losses and horizontal transfer events
... list the causes of phylogenetic inconsistencies
Phylogeny

• Term coined by Ernst Haeckel (1866)
  – Phylon (φύλον)
    • Tribe
    • Race
  – Genus
    • Birth
    • Origin

• At every node in the tree, a new lineage is born
• All lineages in a tree are related because they descend from the same root
• Tree topology shows how the lineages are related

Phylogenetic trees

• A phylogenetic tree represents the phylogeny of species or sequences

The horizontal lines are branches and represent evolutionary lineages changing over time.

The vertical lines represent nodes or evolutionary splits. Line length has no meaning; lines just show which branches are connected.

The branch length represents the evolutionary time between two nodes. Unit: substitutions per sequence site.

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny
Nodes in a tree

- **Tips** (sometimes called “leaves” or “terminal nodes”)
  - Present day species or sequences
  - The only things we can directly measure
  - Contain information used to build the tree

- **Ancestral nodes** (or “internal nodes”)
  - Last common ancestor of its daughter lineages

- **Root** (sometimes)
  - Last common ancestor of whole tree
  - If the tree is rooted, then the time axis is defined away from the root

---

The molecular clock

- The concept of a molecular clock states that the number of mutations between two sequences increases linearly with evolutionary time
  - Zuckerkandl and Pauling found that the number of amino acid differences in hemoglobin correlates with fossil dates

- Sequence space is so large, that sequences almost never converge in evolution

- The molecular clock holds in some cases
Evolutionary time

- The evolutionary divergence between lineages can be measured by using evolving characters
  - For molecular phylogenies, the unit is generally: mutations per sequence site

- Using the principle of the molecular clock in reverse, we can estimate the dates of evolutionary events (splits in the tree)
  - We need to have a few known dates for this to calibrate the tree
  - For example from the fossil record or known dates of infection

Introduction of cholera in Haiti in 2010

U.N. Admits Role in Cholera Epidemic in Haiti

By JONATHAN M. KATZ  AUG. 12, 2010

Protesters marching to the United Nations base housing Nepalese peacekeepers in Mirebalais, Haiti, on Oct. 21, 2010. Assistant Press
Evolutionary distance

- The evolutionary distance between two nodes can be calculated as the sum of all the horizontal branch lengths between them.
  - For example, the distance between virus3 and virus7 is:

\[
= 11 \times 0.77 \text{ mutations/site}
\]

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny

Topology and branch lengths

- **Topology** (branching order) quickly shows you the evolutionary relationships.

- **Branch lengths** show you how many mutations occurred in the evolutionary time between lineages.

- As a whole, the phylogenetic tree quickly shows you the most distinctive clusters.
  - ...it is not always this clear 😊
A case story

- In August 1994 a nurse in Lafayette, LA, tests negative for HIV
- A few weeks later, she breaks off a messy 10 year affair with a doctor
- Three weeks later, while suffering from chronic fatigue symptoms, the doctor gives his ex-mistress a vitamin B-12 shot, somewhat against her will
- In January 1995, the nurse tests positive for both HIV and hepatitis C. Investigation reveals no obvious means of infection (positive test for a sexual partner, accident with a patient, et cetera). The vitamin B-12 shot becomes suspicious
- The doctor’s office records from the day are conveniently missing but eventually found by police buried in the back of a closet. The records show that the doctor had withdrawn blood samples from a known HIV patient and a known hepatitis C patient the same day as the vitamin B-12 shot. The record keeping is not in line with standard office procedure and there is no information as to what happened to either blood sample
- The nurse never had contact with either patient
- Seemingly strong, but otherwise circumstantial, evidence that the doctor deliberately infected the nurse with HIV and hepatitis C

Case story continued

- HIV evolves very fast
  - This is partly why it has been so difficult to develop a cure
- Can we show that the HIV in the nurse is related to the HIV from the patient?
  1. Take samples of HIV from the nurse
  2. Take samples of HIV from the patient
  3. Take samples of HIV from other HIV positive people from the same town
  4. Sequence HIV gene sequences
  5. Construct a phylogeny of the HIV
Molecular evidence of HIV-1 transmission in a criminal case


*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; †Department of Ecology and Evolutionary Biology and Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720; ‡‡Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

Edited by Wouter M. Fith, University of California, Irvine, CA, and approved September 4, 2002 (received for review May 2, 2002)

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States. Phylogenetic analyses of HIV-1 reverse transcriptase and env DNA sequences isolated from the victim, the patient, and a local population sample of HIV-1-positive individuals showed the victim's HIV-1 sequences to be most closely related to and nested within a lineage comprised of the patient's HIV-1 sequences. This finding of paraphyly for the patient's sequences was consistent with the direction of transmission from the patient to the victim. Analysis of the victim's viral reverse transcriptase sequences revealed genotypes consistent with known mutations that confer resistance to AZT, similar to those genotypes found in the patient. A priori establishment of the patient and victim as a suspected transmission pair provided a clear hypothesis for phylogenetic testing. All phylogenetic models and both genes examined strongly supported the close relationship between the HIV-1 sequences of the patient and the victim. Resampling of blood from the suspected transmission pair and independent sequencing by different laboratories provided precaution against laboratory error.

(15) This case was the first time that phylogenetic analysis has been used as evidence in a United States criminal proceeding. Here we present the phylogenetic evidence that constituted part of the prosecution's case that resulted in the conviction of the Louisiana gastroenterologist on the charge of attempted second-degree murder.

Materials and Methods
Criminal Investigation. The prosecution's case was based on circumstantial evidence indicating that on August 4, 1994, a Lafayette, LA, gastroenterologist made a mixture of blood or blood products from two patients under the doctor's care, one infected with HIV-1 and the other with hepatitis C, and injected his former girlfriend by intramuscular injection. Our efforts for the criminal investigation involved only the molecular analysis of HIV-1 sequences, which represented only one part of the prosecution's case against the physician.

Risk factors associated with HIV-1 infection for the victim were determined through the course of the criminal investigation. From 1984 to 1995, the victim reported having sexual contacts with seven men, including the doctor, all of whom were interviewed by local law enforcement agents. The seven men were tested between the
**Branch support**

- Support values show you how reliable a branching split is.
- Mostly displayed as values or circles.
  - Often between 0 and 1, or 0-100%.
- Branches that are not well supported might be collapsed.
  - This means the topology is unclear.
  - Bifurcating → multifurcating branch.

*Note:* this is the same split (in an unrooted tree).

**Rotating branches**

- In the tree on the left, it looks like bat viruses are all grouped together, but they are not in one lineage!
- Because the vertical dimension has no meaning, branches can be freely rotated.
- The trees below are identical, one branch is rotated.
Phylogenetic mobile of mammals

Diameter: 3.5 m. Location: lobby of the Broad Institute in Cambridge, Massachusetts, USA. Design: Peter Agoos.

Different portraits of the same tree
• Because the vertical dimension has no meaning, trees can be displayed in different ways

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny
The time axis is always away from the root

Phylogenies allow us to look back in time

Origin of life
Origin of eukaryotes
Origin of animals
Origin of vertebrates
Origin of mammals
Earliest fossils
Ancestral states

- Can we figure out what animal the very first virus infected?
- We know that evolution tends to be conservative
- We can infer ancestral states by assuming the fewest possible changes in the tree
- This is called the parsimony principle

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny

- Multiple parsimonious solutions means that the ancestral state is ambiguous
Ancestral states

- Can we figure out what animal the very first virus infected?
- We know that evolution tends to be conservative
- We can infer ancestral states by assuming the fewest possible changes in the tree
- This is called the parsimony principle

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny

Orthology* and paralogy

*Not to be confused with Ornithology

Bas E. Dutilh
Systems Biology: Bioinformatic Data Analysis
Utrecht University, February 14th 2017
Reminder: nodes in a phylogenetic tree

- **Terminal nodes**: the present day sequences that were used to create the tree
- **Ancestral nodes**: ancestors of the present-day sequences
- **Root**: Last common ancestor of all the sequences in the tree

Time axis: away from root

http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny

Using phylogenetics to study the evolution of genes

- Speciation
- Gene Duplication
The origin of homologs

- Observed branch in phylogenetic tree:

- So what happened at this node?

- Homologous genes can result from:
  - **Speciation**: when separate species lineages diverge from a common ancestor
    - This means that the two homologs are present in the genome sequences of different species
    - These homologs are called orthologs
  - **Gene duplication**: when a gene is duplicated within a genome and the duplication becomes fixed in the population
    - This means that two the two homologs are present in the same genome sequence
    - These homologs are called paralogs

The evolution of a gene

- Orthologs are derived from a speciation event
  - Orthologs are directly related to same ancestral gene

- Paralogs are derived from a gene duplication event
  - Paralogs have evolved together in the same genome for a while
Question

• Observe the two simplified gene trees above of two homologs from mouse and two homologs from human.

1. Which are the speciation nodes and which are the gene duplication nodes?

2. What kind of homologs are Mouse A and Human A in Tree 1?

3. What kind of homologs are Human C and Human D in Tree 2?

4. What kind of homologs are Mouse D and Human D in Tree 2?

5. Which genes do you think may have the same function in Tree 1?

6. Which genes do you think may have the same function in Tree 2?

How to tell the difference?

• Rule of thumb:
  - If the two daughter branches of a node contain the same species, it may be a gene duplication node
  - Otherwise it may be a speciation node

• There are exceptions where the rule of thumb does not hold
  - Horizontal gene transfer (HGT)
  - Unrecognized paralogy
Evolution of function

• Orthologs are directly related to same ancestral gene
  → Likely perform same ancestral function (evolution is conservative)

• Paralogs evolved together in the same genome for a while
  → Function might diverge (unlikely that two genes have the exact same function in one genome)

Gene function prediction

• Transfer of functional annotations between genes is only reliable between orthologs

• Function is more likely to diverge between paralogs
Orthology & paralogy are evolutionary concepts

- Researchers are often trying to identify orthologs in model organisms, with the goal of transferring functional annotation
- However, note that orthology and paralogy are originally evolutionary definitions that say nothing about function

Species trees and gene trees

- The phylogenetic tree of a gene family can be much more complex than the species tree
- It can be challenging to reconcile the gene tree with the species tree
- It helps if you have prior knowledge of the species tree
Species tree reconciliation

- Using species tree reconciliation, we can deduce where gene losses occurred.

(b) Gene Tree

Gene losses

- Species tree reconciliation allows us to answer questions like:
  - How many copies of this gene did the last common ancestor of all fishes have?
  - How many copies of this gene did the last common ancestor of all mammals have?
Horizontal gene transfer (HGT)

• Sometimes, you need to assume many losses to explain a gene tree
• It may be more parsimonious to assume just one HGT event
• ... or just one contaminated sample 😞

Unrecognized paralogy

• Another mechanism that causes of conflict between the phylogenetic tree of a gene and the species tree is unrecognized paralogy

- Gene invention
- Speciation node (orthologs)
- Gene duplication node (paralogs)
- Gene loss

...with all these processes going on in thousands of genes evolution can get very complex!
**Phylogenetic inconsistencies**

- The phylogenies of different genes from the same genomes can be inconsistent

![Phylogenetic tree]

- This can be the result of:
  - Evolution of the gene is different than the evolution of the genome
    - Horizontal gene transfer
    - Unrecognized paralogy
  - Technical issues
    - Bad model of evolution
    - Bad alignments
    - Bad phylogenies
  - Biological noise
    - Mutational saturation: multiple mutations at the same sequence site
    - Different rates of evolution in different lineages (inconsistent molecular clock)

**The “Tree” of Life?**

- Due to evolutionary processes like HGT, gene loss, endosymbiosis (mitochondrion, chloroplast) relationships might better be represented as a network than as a tree of life

![Diagram of the "Tree" of Life]