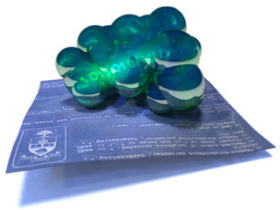
A
Bioinformatics
Course

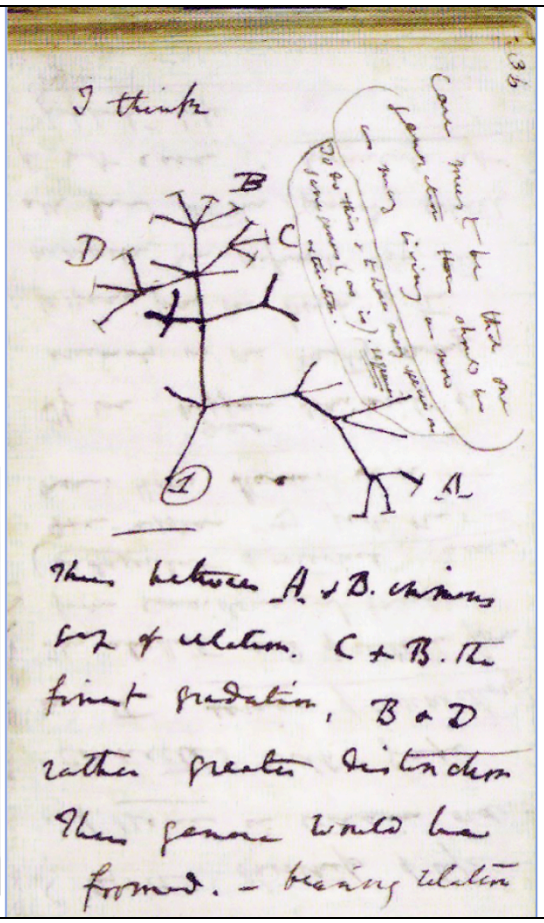
Concepts of Phylogenetic Analysis



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

1837: The Tree



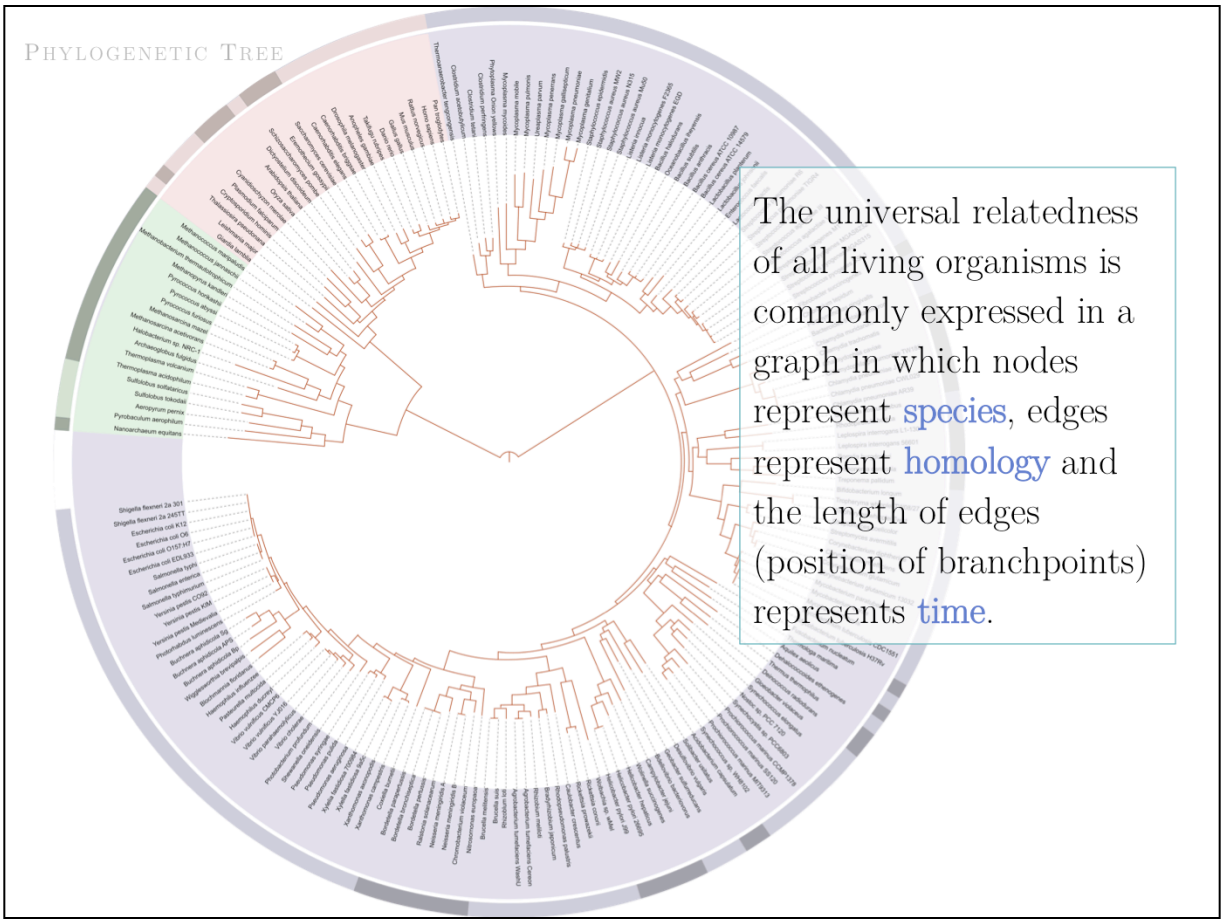
This page from Darwin's notebooks around July 1837 shows his first sketch of an evolutionary tree.

cf. http://en.wikipedia.org/wiki/Tree_of_life_(science)

Phylogenetic Tree

http://wallpoper.com/images/00/23/94/22/trees-grass_00239422.jpg

The tree – originating from a single stem and spreading out into an uncountable number of branches, terminating in single leaves – is the most common conceptual metaphor for the process of evolution.



All life originates from a common ancestor species, and has diversified in a process of speciation into the complexity we observe today.

Phylogenetic Analysis

Phylogenetic analysis uses observed states to infer the *evolutionary distance* between related species. Once the relative distances are computed, a tree can be constructed. The distances are important for:

- Evolutionary Trace method of discovering functional residues
- Quantifying conservation
- Inferring histories of descendance
- Distinguishing orthologues and paralogues

Fundamental to such analysis is the quantitative interpretation of evolutionary distance under a branching model of stochastic variation and selection.

Stochastic variation (i.e. random, undirected variation) changes gene sequences. Sequence changes lead to changes in function and global fitness. *Selection* will cause changes to become fixed in a population. However, this process is not the same in different populations if the populations do not constantly share and mix genetic material. Once populations become separated in reproduction, their genome sequences diverge. This divergence is described as branching from a common ancestor.

For sets of genes that have diverged from the same ancestor (a *cenancestor*, or *LCA* – Last Common Ancestor), the amount of observed divergence allows us to order branching events on a tree. This requires (i) quantifying divergence, and (ii) building a tree that best explains the observed divergence between contemporary genes (leaves of the tree, or *OTU* – Operational Taxonomic Units).

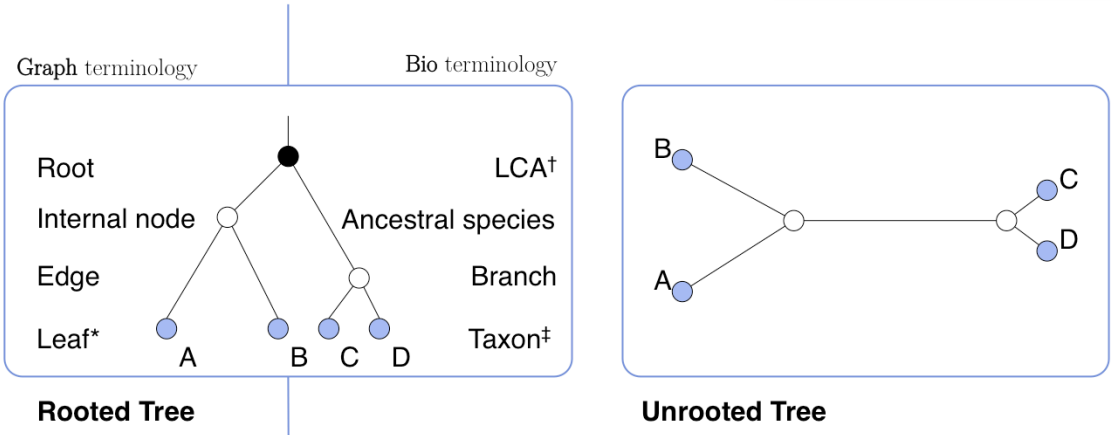
Phylogenetic Tree

A phylogenetic tree is a formal graph-abstraction for computable representations of evolutionary relationships. **Leafs** (terminal nodes) represent contemporary species, genes etc., **internal nodes** represent hypothetical states of ancestral species. The **topology** represents the evolutionary relationship (ancestry and descent) and the **branch-length** represents similarity.

We usually draw phylogenetic trees with the root at the top or at the left. This reflects our intution about the tree representing a process, a sequence of events, and aligning this with our reading conventions: top to bottom, left to right.

Phylogenetic Tree

A phylogenetic tree is a formal graph-abstraction for computable representations of evolutionary relationships. **Leafs** (terminal nodes) represent contemporary species, genes etc., **internal nodes** represent hypothetical states of ancestral species. The **topology** represents the evolutionary relationship (ancestry and descent) and the **branch-length** represents similarity.



(*) also: Terminal Node

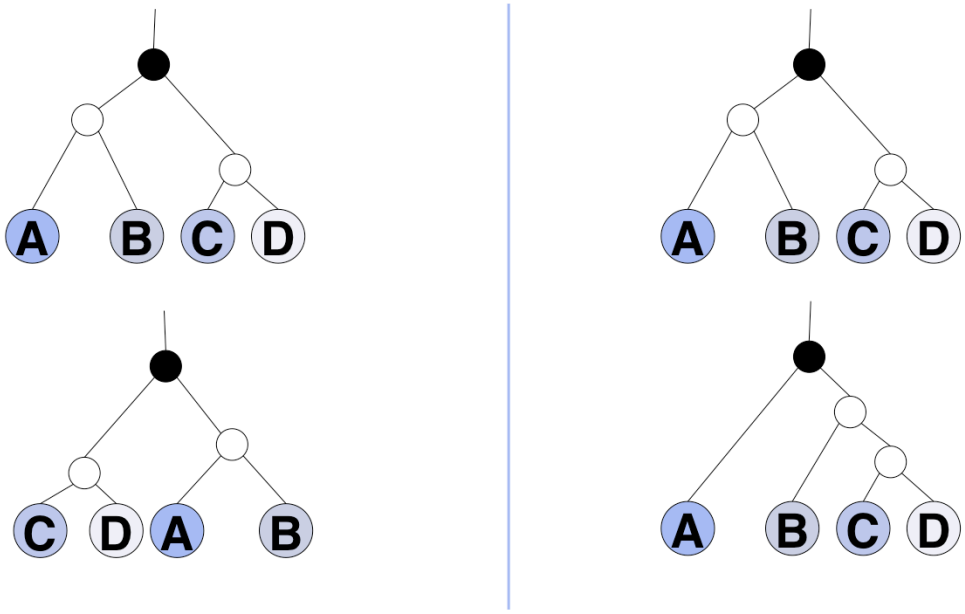
(‡) also: OTU (Operational Taxonomic Unit), gene, population, species ...

(†) (Last Common Ancestor) also Cenancestor, also LUCA (Universal)

.

Phylogenetic Tree

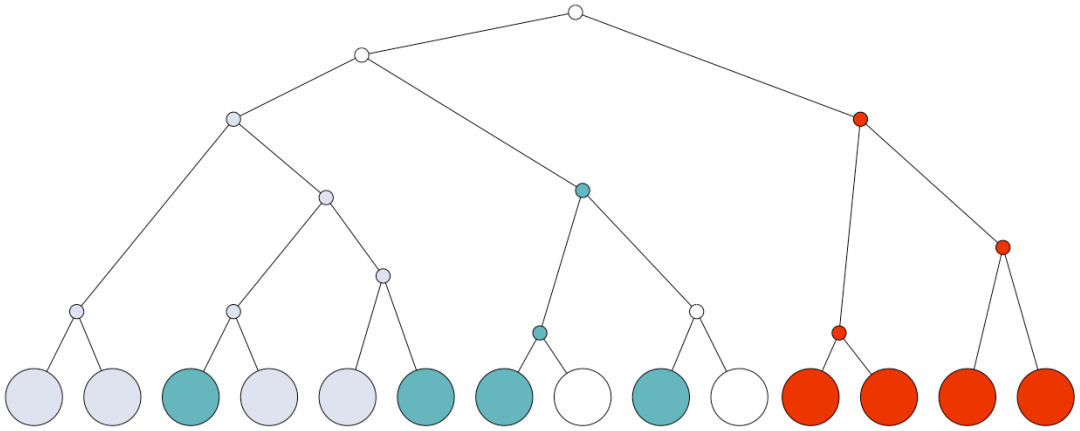
Tree graphs differ only in their topology or "branching pattern", not the order in which neighbouring OTUs are drawn. The trees on the left are identical. The trees on the right are not.



Algorithms that draw trees need to decide how to order the terminal nodes. Is the top, or the bottom tree on the left “better”? In this example it really makes no difference, and the arrangement could be randomly chosen. But if the distances to the root node would be different, as they often are, we could arrange the tree so that the difference in distance to the root between adjacent genes is minimized. This would place more closely related leaves closer to each other in the drawing.

However, such *layout* decisions have nothing to do with the topology of the tree that represents the evolutionary relationships, nor with the objective function under which the tree is constructed.

Phylogenetic Tree



Monophyletic group:	a *Clade* – a node and all of its descendants.
Paraphyletic group:	a clade minus some of its members.
Polyphyletic group:	members of several clades, perhaps grouped by a convergent feature or other superficial similarity.

Definitions ...

The *molecular clock* hypothesis relates distance to time:

The concept of a molecular clock stems from the early 1960s, when Pauling and Zuckerkandl noted a correlation between hemoglobin diversity and species divergence time.

Strictly speaking such a relation is expected only for species with similar evolutionary landscapes, generation times and mutation rates. There are a number of reasons why a molecular clock might *not* apply:

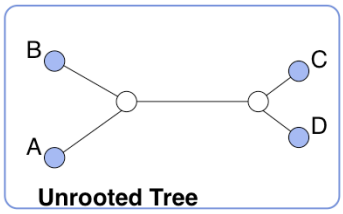
- Shorter generation times can fix more mutations in a length of clock-time.
- Very large populations may make the effects of individual mutations too small to confer a selective advantage, thus slowing the mutation rate.
- Species have different replication error rates.
- Evolution markers (individual genes) may have significantly different rates of acceptance of mutations.
- Environmental conditions may be very different, thus placing one species under much larger adaptive pressure than another.

Nevertheless assuming a constant clock is a useful first-order approach.

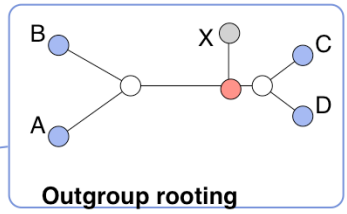
Apparently some of the reasons given above why a molecular clock should be inaccurate, cancel each other. As a result, the "molecular clock" actually works.

Root

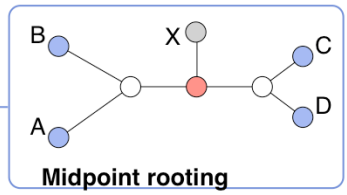
A tree does not necessarily imply a particular *direction* of time among its branches. Without a concept of time, it merely represents *distance* i.e. **relationship**, not **descendance**.



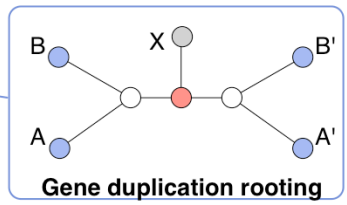
Under the assumption of a "molecular clock" the tree can be rooted and its branches acquire a *direction* in time.



Outgroup rooting places the root at the LCA of a distant relative and the ingroup.



Midpoint rooting places the root at the midpoint of the longest branch (essentially a molecular clock model).



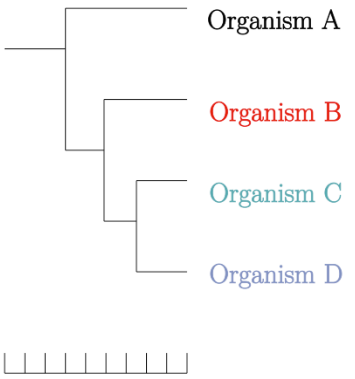
Gene duplication rooting places the root at the midpoint of duplicated paralogues (correction for unequal rates is needed).

Why does midpoint rooting place the root into the longest branch?

Because the root branch is twice as long as it should be – since it is missing a branching node: the root.

Cladogram

Depending on which relationship is represented in a tree, different tree-types can be created:

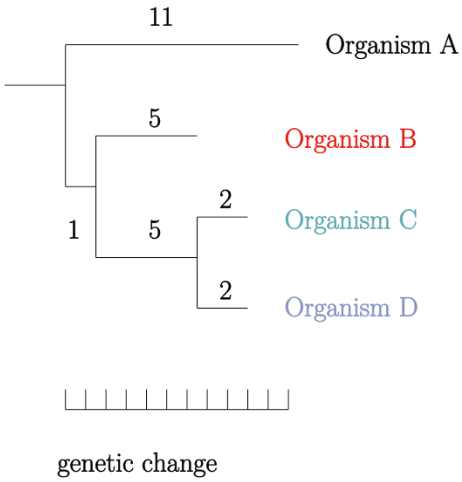


In a *Cladogram*, branch length has no meaning. It shows only the branching pattern.

.

PHYLOGRAM

Depending on which relationship is represented in a tree, different tree-types can be created:

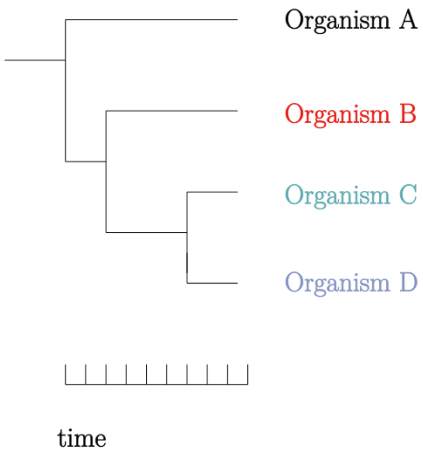


Phylograms constrain the tree so that the same branch length corresponds to the same amount of genetic change.

Phylograms are the most frequently used diagrams for phylogenetic relationships.

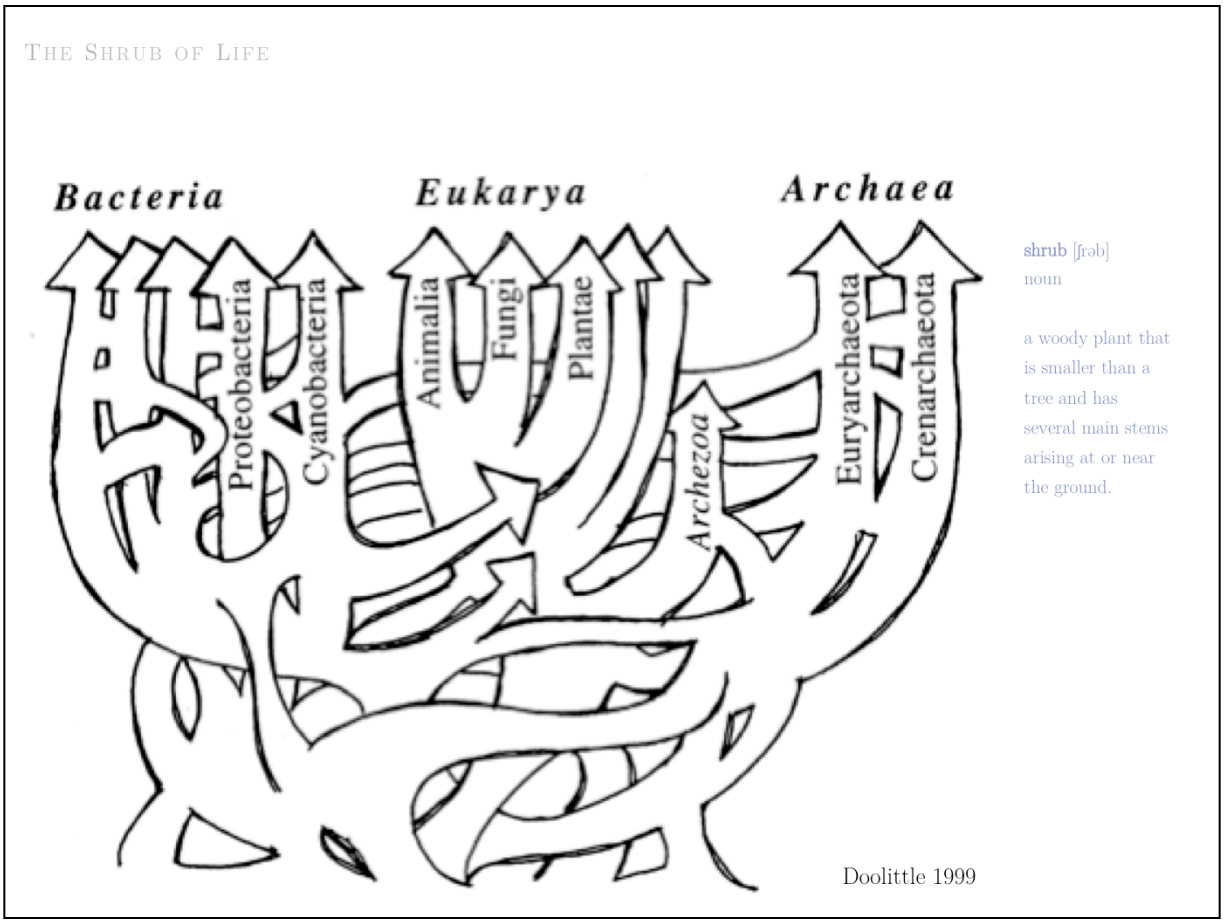
Ultrametric Tree

Depending on which relationship is represented in a tree, different tree-types can be created:



Ultrametric trees constrain the tree so that the same amount of time has passed from the LCA to all OTUs

Ultrametric trees may at first glance look like cladograms, but the branching point heights (distance from the root) are drawn proportional to the amount of time that has passed. In contrast, in a cladogram, branch lengths have no meaning and only the topology carries information.



But is a tree even the correct way to describe the evolutionary relationship of genes?

Remember what it describes: an individual gene, evolving over time and its relationship to other genes, independently evolving after speciation or duplication events.

Thus, in principle the relationship between genes is constrained by the universal Tree of Life.

However, there is a third type of event that needs to be considered, albeit it is less frequent than the other two: that of genetic material passing from one species to another in *horizontal gene transfer*.

Horizontal Gene Transfer is widespread

In a recent study, Crisp *et al.* (2015) identified 145 genes in the human genome whose evolutionary relationship clearly are incompatible with the Common Tree. That is 0.7% !

Phylogenetic tree for the human gene HAS1. For each branch the species name and UniProt accession is shown. The human gene under analysis is shown in orange, proteins from chordates are in red, other metazoa in black, fungi in pink, plants in green, protists in grey, archaea in light blue and bacteria in dark blue.

Numbers indicate aLRT support values for each branch where higher than 0.75 (on short terminal branches the support values are not shown).

(From: Crisp *et al.* 2015)

But is this the *major* mechanism of evolution ?

W. Ford Doolittle, Halifax

Requirements:

- Proximity to donor DNA
- Physical proximity to organism
- Stability of DNA in environment
- Vector transmission
- Uptake and insertion
- Maintenance
- Stabilization
- Selection

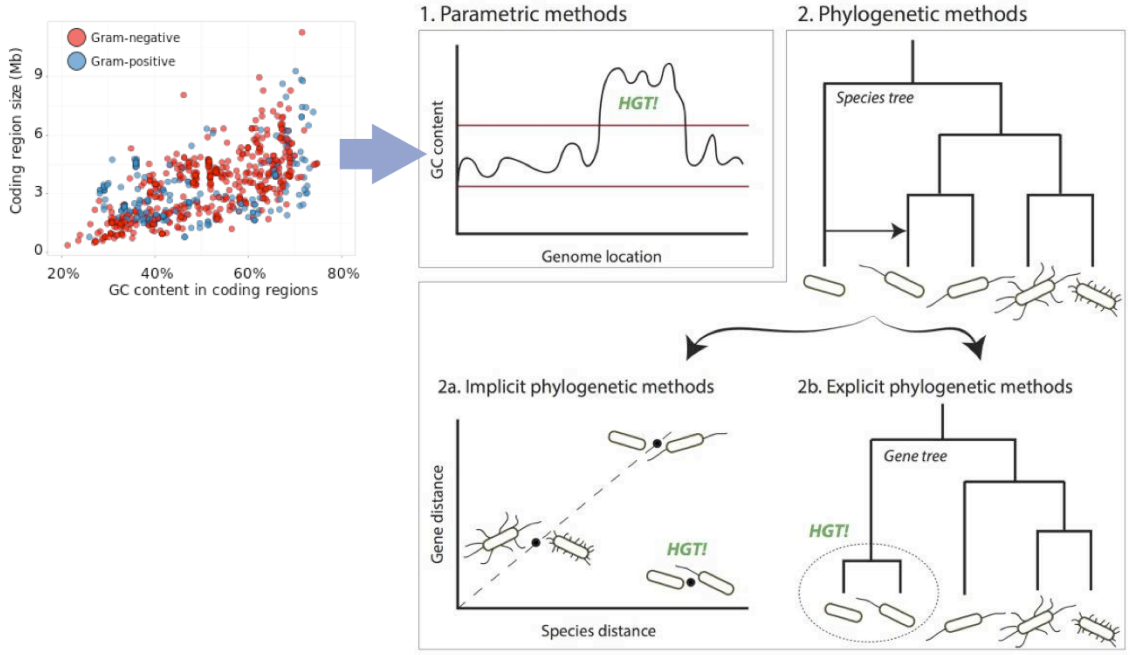
After Exchange ...

- Amelioration (adaptation to host genome features)
- Functional changes
- Spread within new species
- Stabilization
- Divergence from donor species

Limitations:

- Instability in new host (e.g. repeats)
- Restriction systems
- GC/Codon usage incompatability
- Splicing and other signals incorrect
- RNA editing
- Lack of appropriate interacting genes (e.g. multi-subunit enzymes)

HGT can be detected with "parametric" and "phylogenetic" methods:



https://en.wikipedia.org/wiki/Inferring_horizontal_gene_transfer

Parametric methods are based on the analysis of sequence composition. GC contents is frequently used but virtually all organisms have characteristic sequence signatures regarding codon preferences, amino-acid profiles and GC contents.

Phylogenetic methods look at differences between trees for species (e.g. based on 16S rRNA trees) and trees for genes. In the example above, note the large distance that separate the two bacteria in 2., the small distance between the genes in 2b. However to infer the *direction* of the transfer requires evaluating more than two species, or using parametric evidence.

Ravenhall M, Škunca N, Lassalle F and Dessimoz C. (2015) Inferring horizontal gene transfer. PLoS Comput Biol. 11(5):e1004095. (Paper also availbale on Wikipedia: https://en.wikipedia.org/wiki/Inferring_horizontal_gene_transfer)

Split Networks

http://www.splitstree.org/

A split is a partition of the alignment, according to the observed states in a single column. Compatible splits can be represented in a tree.

A split network is a general type of phylogenetic graph that can represent any collection of splits, whether incompatible or not. For a compatible set of splits, it is always possible to represent each split by a single branch, and thus the resulting graph is a tree. In general, however, this will not be possible and in a split network usually a whole band of parallel branches (also called parallel edges) is required to represent a single split.

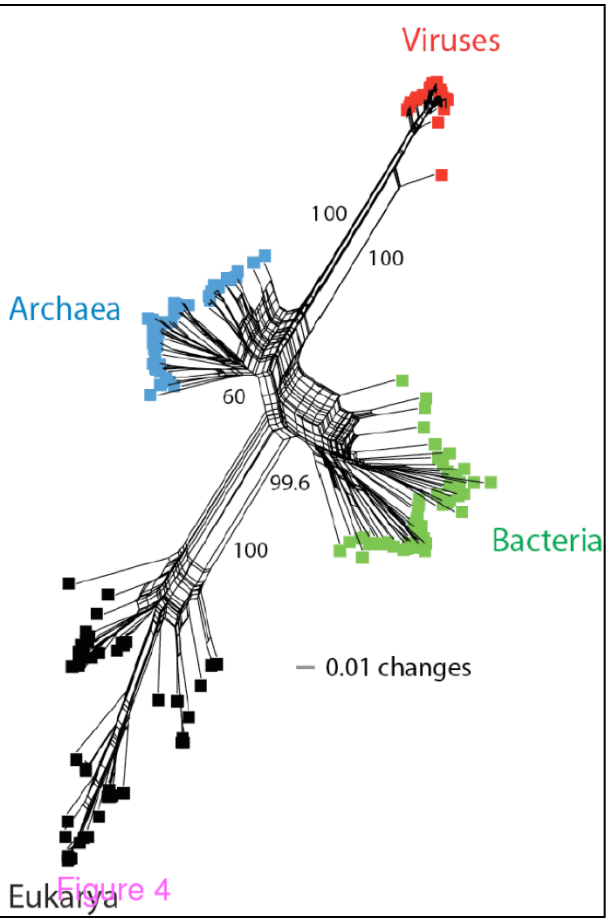


Figure 4

Especially in early evolution, horizontal gene transfer and mixing of evolutionary material may have been much more common than it is today, as current organisms have developed **very** sophisticated mechanisms to protect their genetic identity.

Thus an explicit treatment of non-dichotomous relationships is important, especially for **deep** evolutionary trees, such as the one that was recently proposed to establish *giant viruses* as a fourth superkingdom of life.

A fourth Superkingdom

Network tree generated from the presence/absence matrix of 1,739 Fold Superfamilies in 200 proteomes sampled equally from the four supergroups. The number of non-constant sites was 1,581. Nodes in the network tree are proteomes and are represented by rectangles labelled red, blue, green, and black for viruses, Archaea, Bacteria and Eukarya, respectively. Numbers on the major splits indicate bootstrap values.

Nasir A, Kim KM, Caetano-Anolles G. (2012) *BMC Evol Biol.* **12**:156

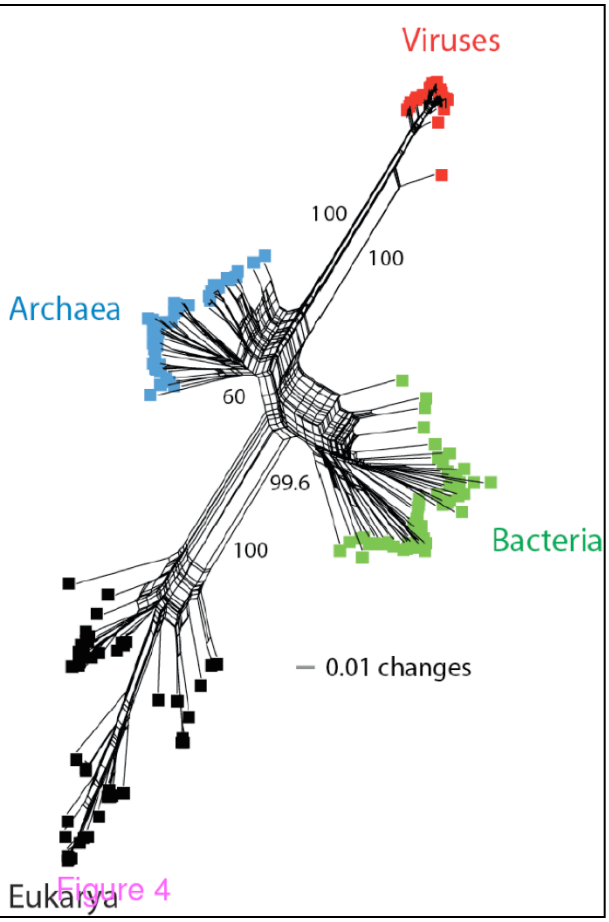
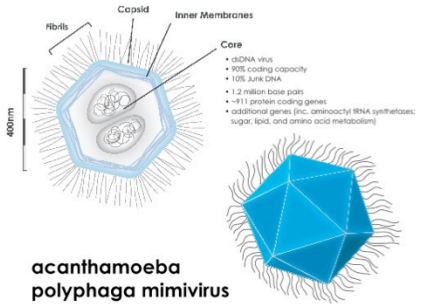


Figure 4

A Fourth Superkingdom



Nasir A, Kim KM, Caetano-Anolles G. (2012) Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol Biol.* **12**:156

ABSTRACT:BACKGROUND: The discovery of giant viruses with genome and physical size comparable to cellular organisms, remnants of protein translation machinery and virus-specific parasites (virophages) have raised intriguing questions about their origin. Evidence advocates for their inclusion into global phylogenomic studies and their consideration as a distinct and ancient form of life.

RESULTS: Here we reconstruct phylogenies describing the evolution of proteomes and protein domain structures of cellular organisms and double-stranded DNA viruses with medium-to-very-large proteomes (giant viruses). Trees of proteomes define viruses as a 'fourth supergroup' along with superkingdoms Archaea, Bacteria, and Eukarya. Trees of domains indicate they have evolved via massive and primordial reductive evolutionary processes. The distribution of domain structures suggests giant viruses harbor a significant number of protein domains including those with no cellular representation. The genomic and structural diversity embedded in the viral proteomes is comparable to the cellular proteomes of organisms with parasitic lifestyles. Since viral domains are widespread among cellular species, we propose that viruses mediate gene transfer between cells and crucially enhance biodiversity.

CONCLUSIONS: Results call for a change in the way viruses are perceived. They likely represent a distinct form of life that either predated or coexisted with the last universal common ancestor (LUCA) and constitute a very crucial part of our planet's biosphere.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA