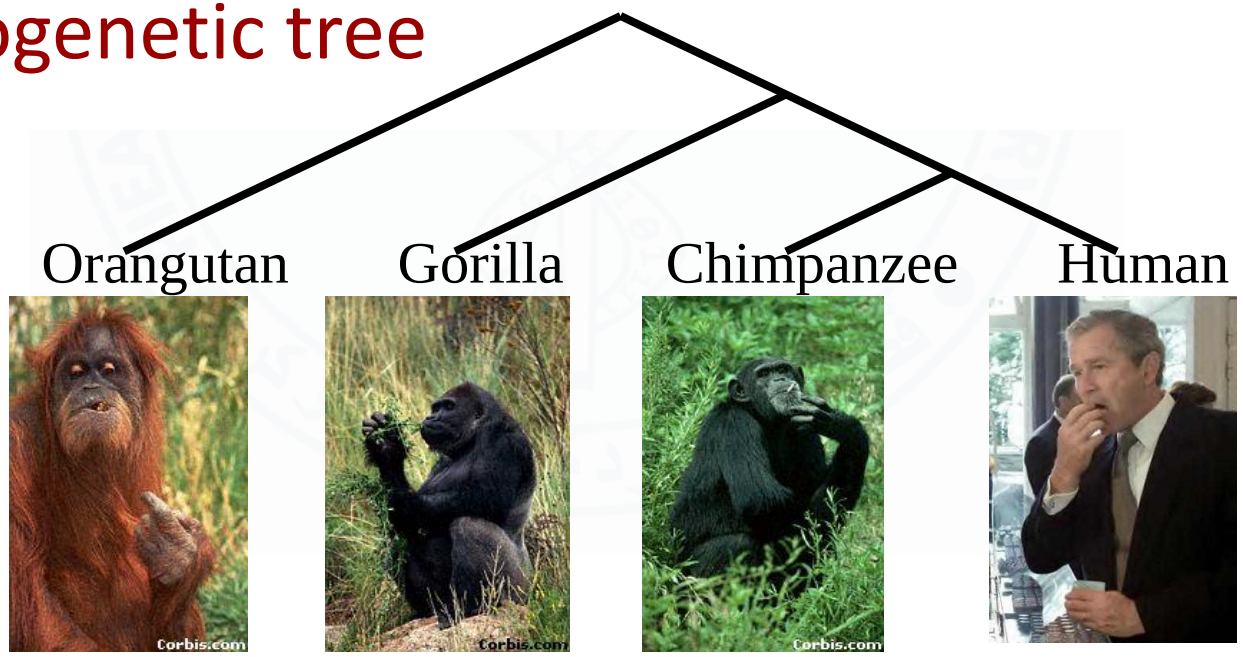
Phylogeny: building the tree of life

Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Acknowledgements: Most of these slides are taken from Lectures by Dr. Asa-Ben-Hur, CSU.

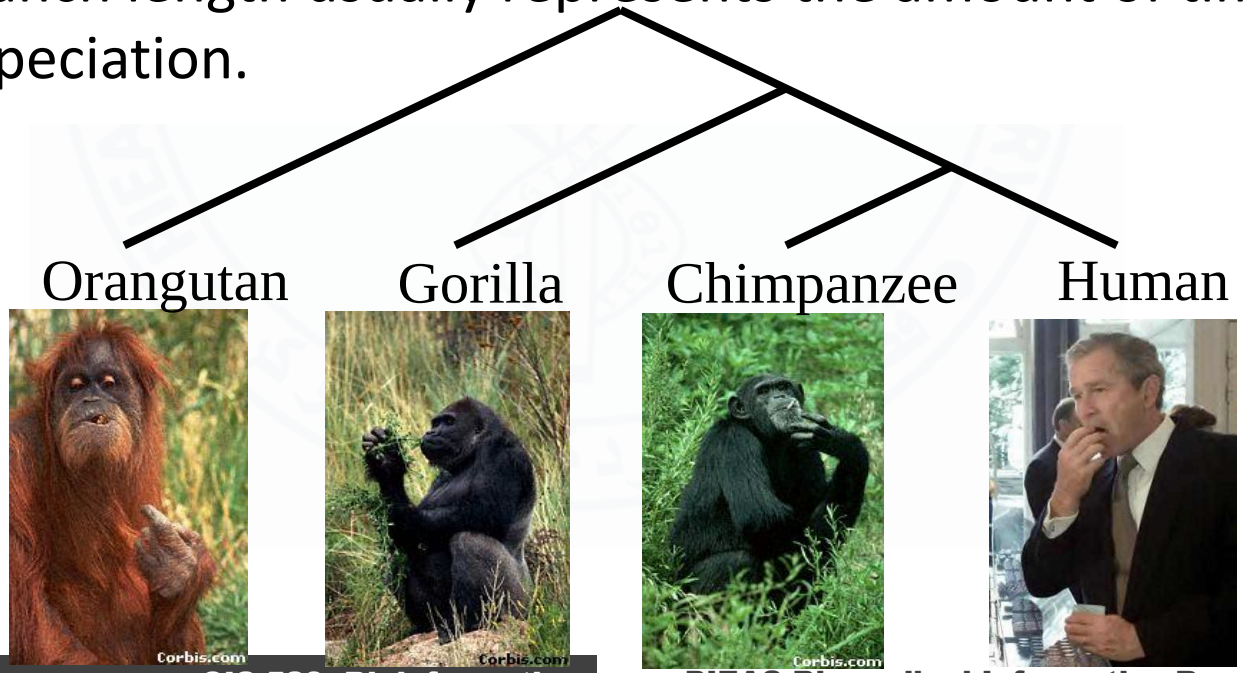
Phylogeny and Phylogenetic Trees

- The biological evidence suggests that all life evolved from a common ancestor
- Phylogeny is the study of the evolutionary histroy of a group of species (taxa)
- The relationship can be represented using a phylogenetic tree



Phylogeny and Phylogenetic Trees

- The leaves of the tree corresponds to extant taxa
- Internal nodes correspond to speciation events, and represent ancestral taxa.
- Branches define the relationships among the taxa (descent and ancestry)
- A **phylogenetic tree** is a binary tree.
- The branch length usually represents the amount of time since speciation.



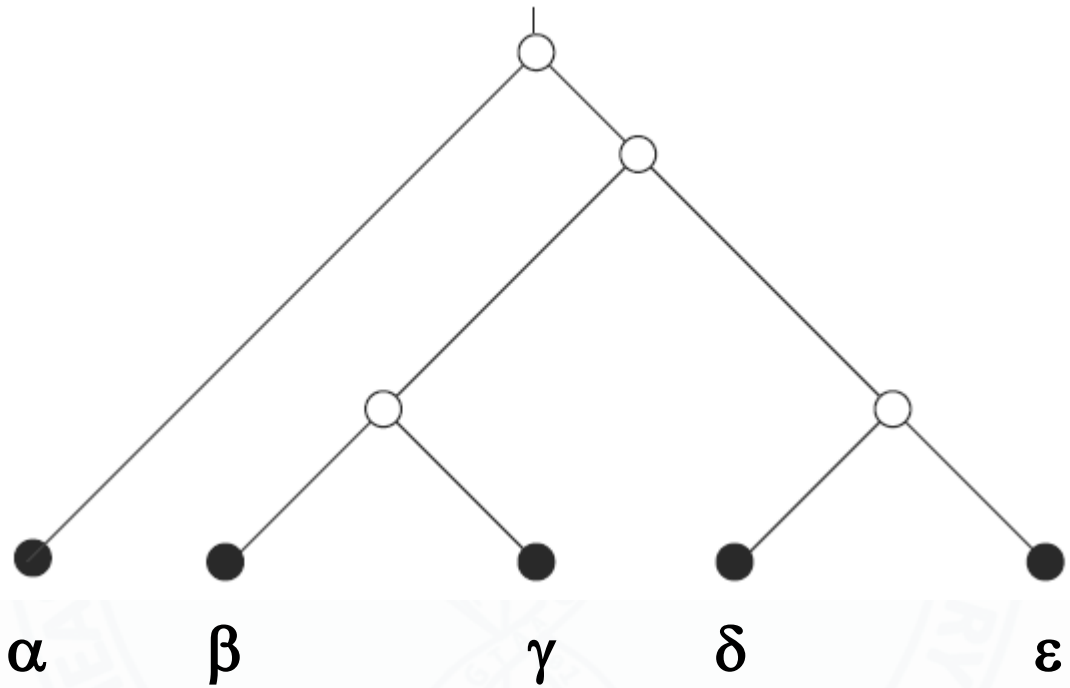
Objectives

- Reconstruct the ancestry of species, i.e. find the tree of life and
- To estimate the time of divergence between organisms since they last shared a common ancestor.

Cladograms and Phylogenetic trees from Phonotypical characteristics

- See video!

A Phylogenetic Tree



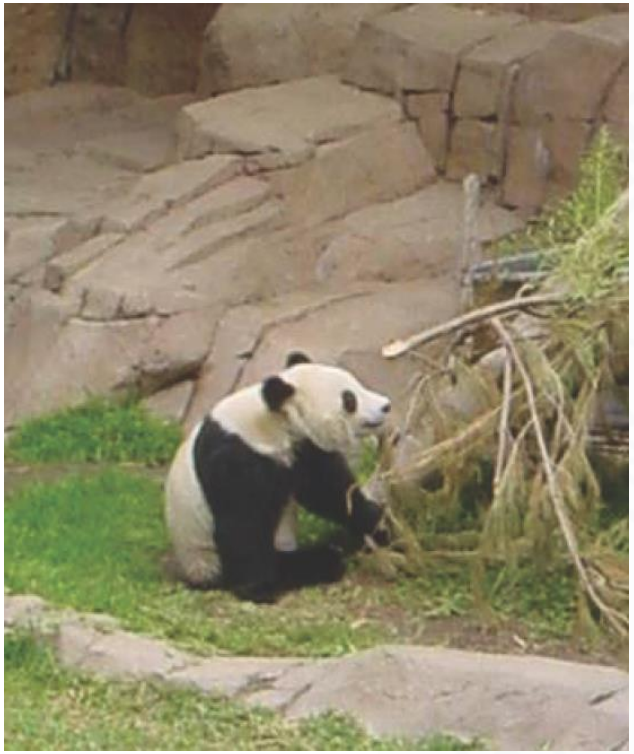
- A (rooted) tree for 5 species.
- Leaves α, β, γ, δ, and ε, correspond to contemporary species, for which data has been collected
- Internal nodes correspond to (inferred) ancestors

Using Morphological Characteristics

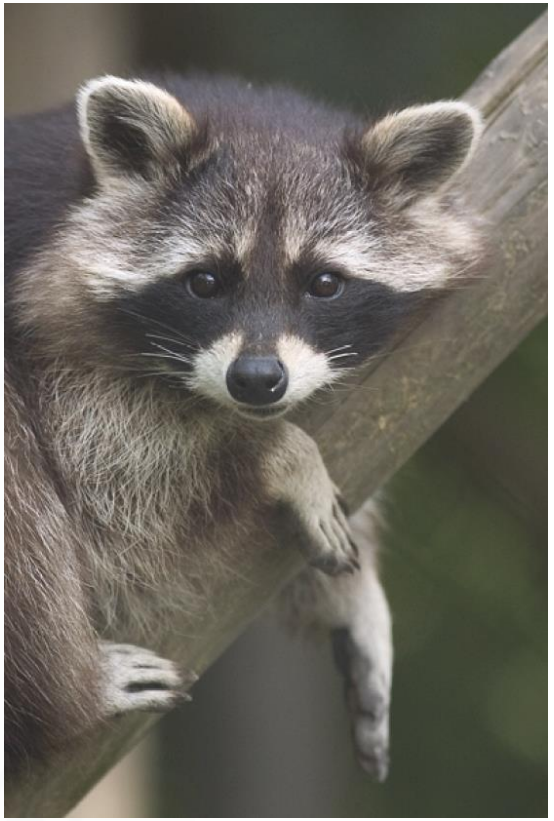
	Characters					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	1	0	0	1	1	0
β	0	0	1	0	0	0
γ	1	1	0	0	0	0
δ	1	1	0	1	1	1
ϵ	0	0	1	1	1	0

Here, the 0s and 1s are indicating the presence or absence of a character (has feathers?, lays eggs?, curved beak?, flies?, . . .).

Raccoon or Bear ?



- Based on anatomical and behavioral characteristics, the panda was originally classified as a raccoon (1870)
- In 1985, the panda was re-classified as a bear when an analysis based on molecular data was done



Platypus

The Platypus (Ornithorhynchus anatinus) is a venomous, egg-laying, duck-billed mammal; the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)

commons.wikimedia.org/wiki/Image:Ornithorhynchidae-00.jpg

Need for Other Approaches

Hard to resolve relationships using morphology and behavior alone

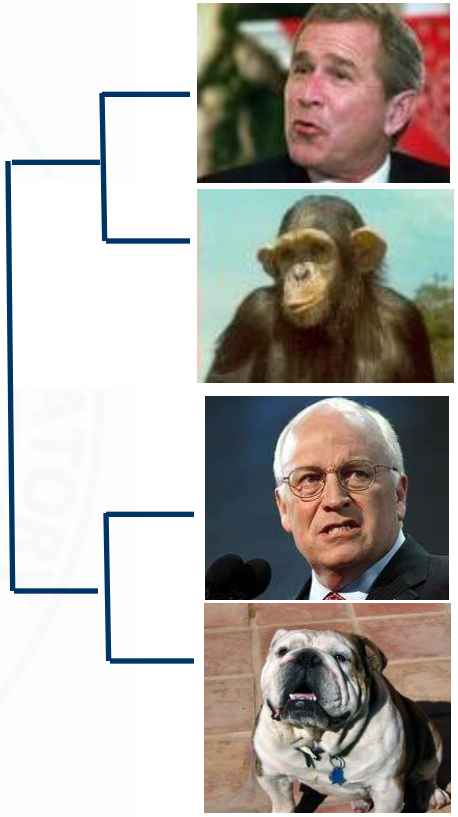
1. Similar characteristics can evolve independently in distantly related organisms — convergent-evolution

2. It is often difficult to find characteristics that are common to all the organisms under study

Inferring Phylogenies

Trees can be inferred by several criteria:

- Morphology of the organisms
  - *Can lead to mistakes!*
- **Sequence comparison**



Example:

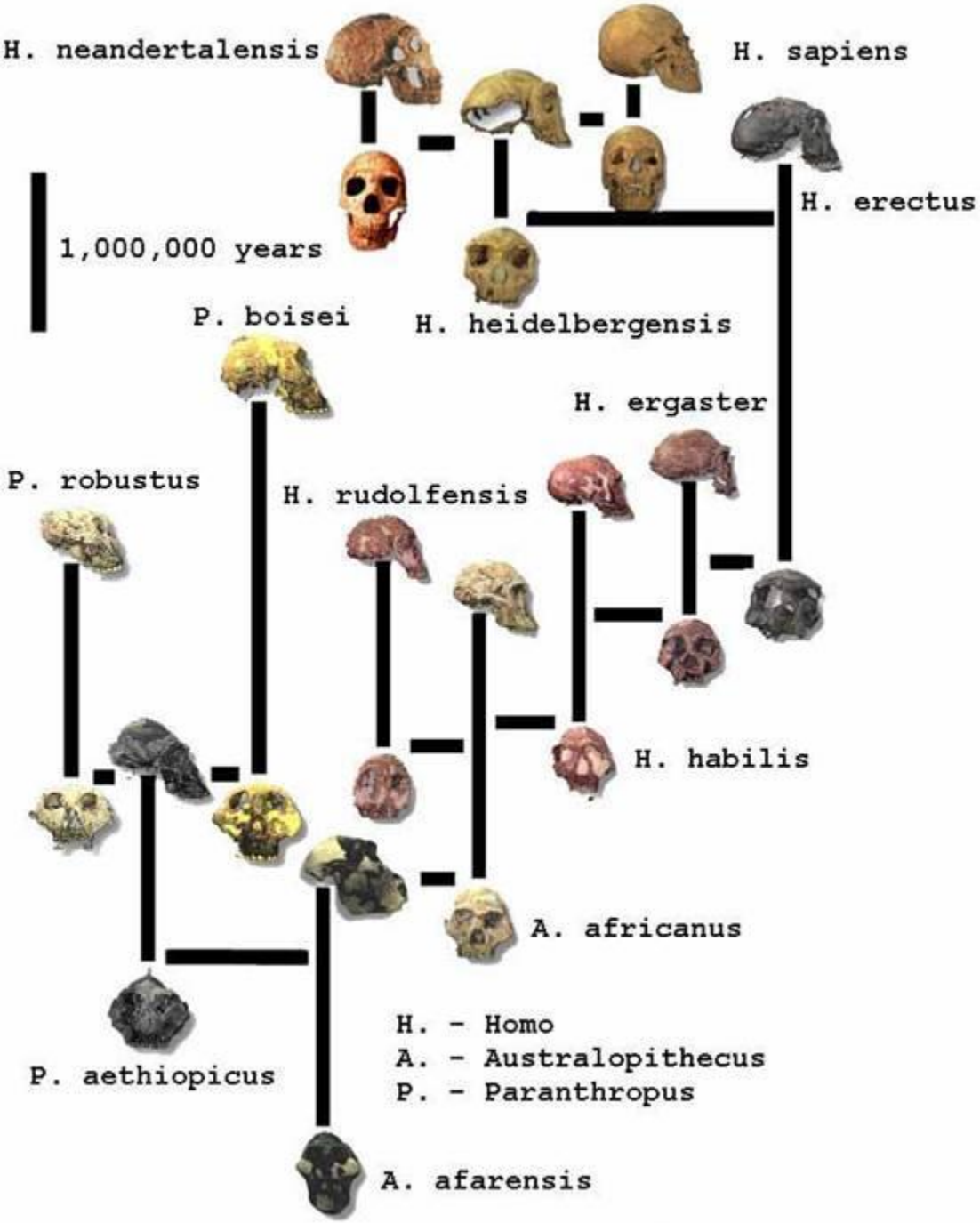
Orc: ACAGTGACGCCCCAAACGT

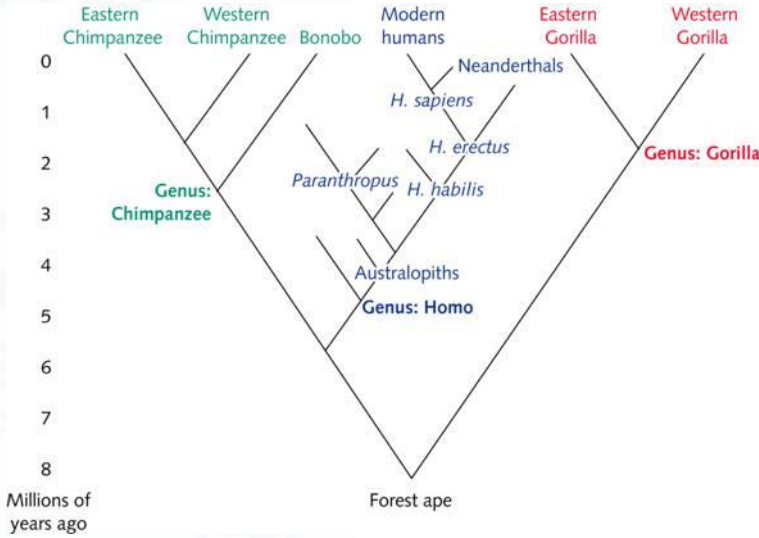
Elf: ACAGTGACGCTACAAACGT

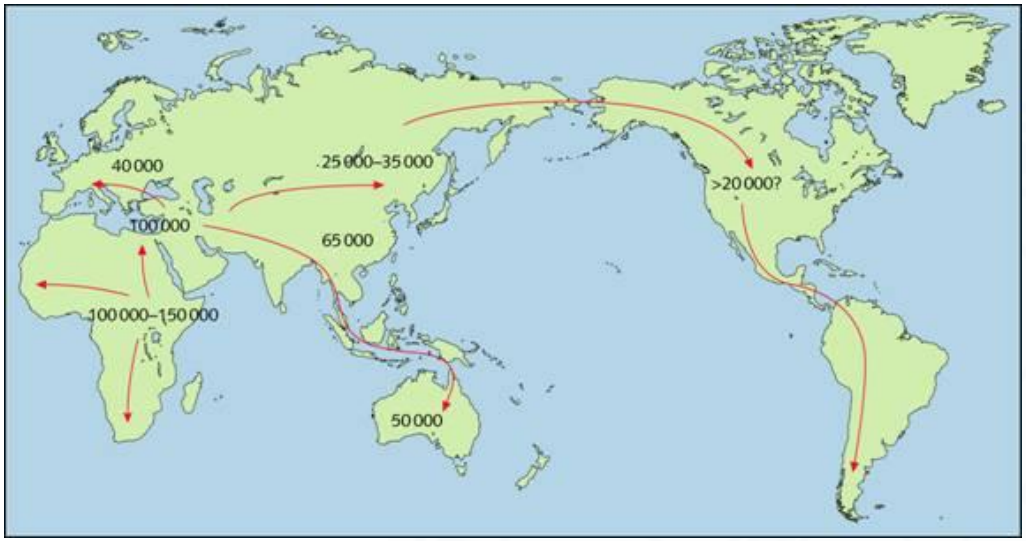
Dwarf: CCTGTGACGTAACAAACGA

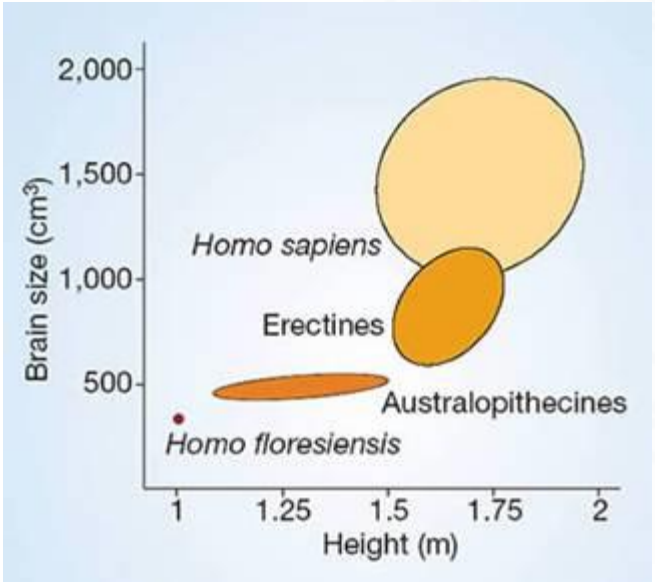
Hobbit: CCTGTGACGTAGCAAACGA

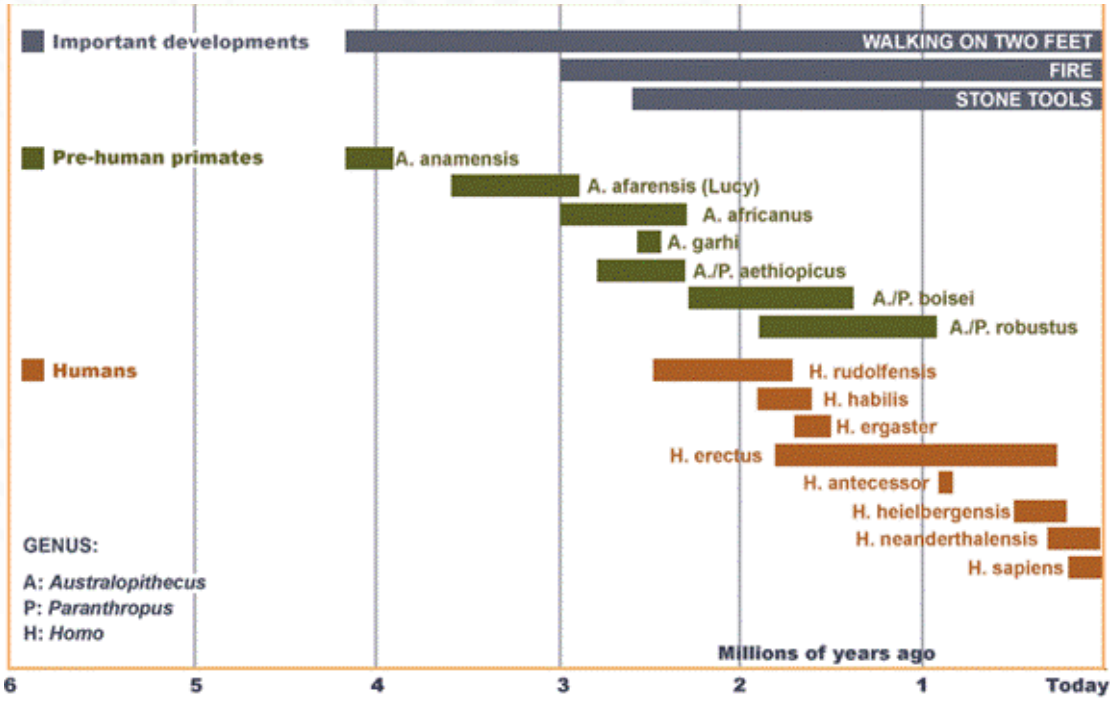
Human: CCTGTGACGTAGCAAACGA

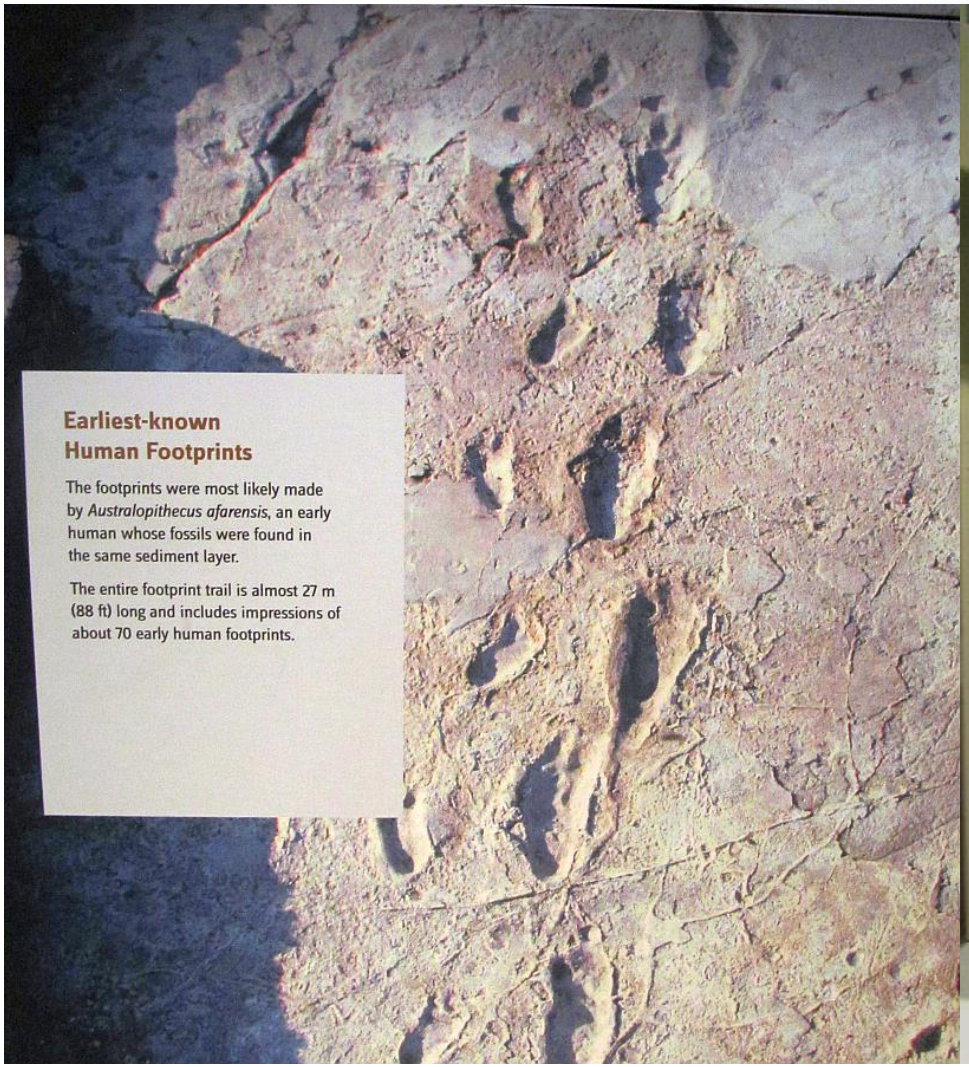

H. neandertalensis
H. sapiens
H. erectus
1,000,000 years
P. boisei
H. heidelbergensis
H. ergaster
P. robustus
H. rudolfensis
H. habilis
A. africanus
H. - Homo
A. - Australopithecus
P. - Paranthropus
P. aethiopicus
A. afarensis


Eastern Chimpanzee
Western Chimpanzee
Bonobo
Modern humans
Eastern Gorilla
Western Gorilla
Neanderthals
H. sapiens
H. erectus
Paranthropus
H. habilis
Genus: Chimpanzee
Genus: Gorilla
Australopiths
Genus: Homo
Forest ape
0
1
2
3
4
5
6
7
8
Millions of years ago


40 000
25 000–35 000
>20 000?
100 000
65 000
100 000–150 000
50 000


2,000
1,500
1,000
500
Brain size (cm³)
Homo sapiens
Erectines
Australopithecines
Homo floresiensis
1
1.25
1.5
1.75
2
Height (m)

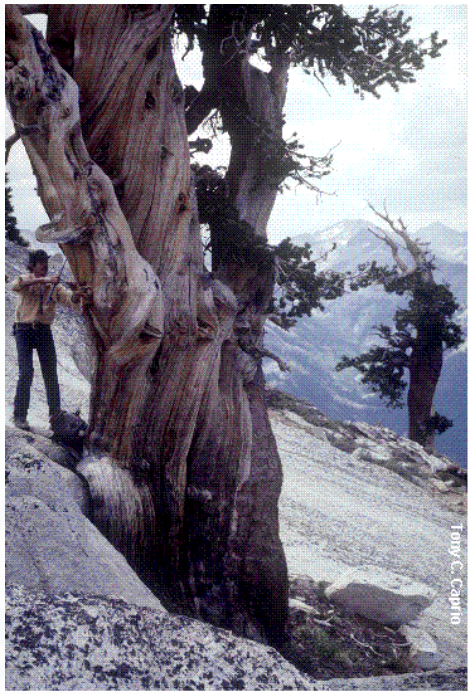

Important developments
WALKING ON TWO FEET
FIRE
STONE TOOLS
Pre-human primates
A. anamensis
A. afarensis (Lucy)
A. africanus
A. garhi
A./P. aethiopicus
A./P. boisei
A./P. robustus
Humans
H. rudolfensis
H. habilis
H. ergaster
H. erectus
H. antecessor
H. heielbergensis
H. neanderthalensis
H. sapiens
GENUS:
A: Australopithecus
P: Paranthropus
H: Homo
Millions of years ago
6
5
4
3
2
1
Today


Earliest-known
Human Footprints
The footprints were most likely made by *Australopithecus afarensis*, an early human whose fossils were found in the same sediment layer.
The entire footprint trail is almost 27 m (88 ft) long and includes impressions of about 70 early human footprints.


The First Step: Walking Upright

What is phylogeny good for ?

- Evolutionary history (“tree of life”)
- Population history
- Origins of diseases
- Prediction of sequence function
- MSA
- Can be applied to organisms, sequences, viruses, languages, etc.



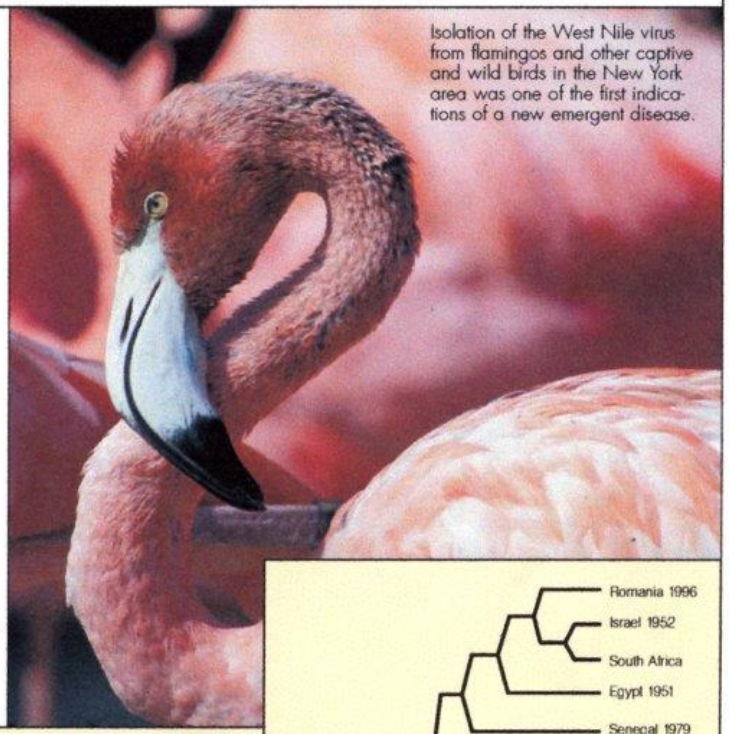
The Tree of Life and Human Health

Understanding how organisms, as well as their genes and gene products, are related to one another has become a powerful tool for identifying disease organisms, tracing the history of infections, and predicting disease outbreaks.

Identifying Emerging Diseases: West Nile Virus

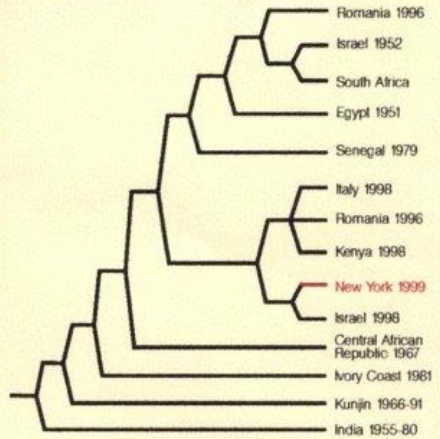
When an encephalitis-like viral infection emerged in people living in the New York region in 1999, it was first suspected to be the St. Louis encephalitis virus. Transmitted by mosquitos, the virus was simultaneously found to be associated with a high mortality in wild and domesticated birds. It has currently spread as far west as California and has resulted in numerous human deaths.

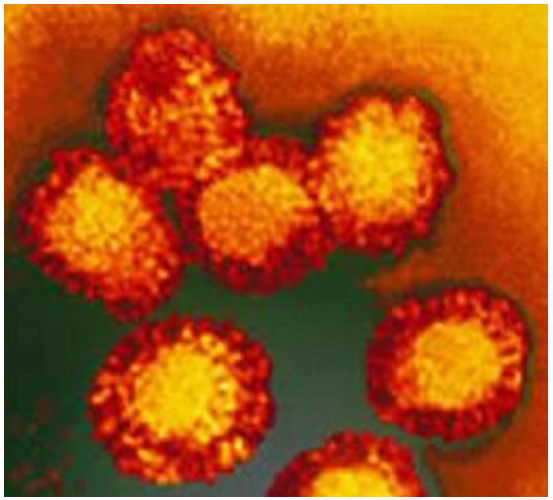
In two separate studies, health workers used phylogenetic analysis to identify viral isolates from mosquitos and birds as a new outbreak of the West Nile virus rather than St. Louis encephalitis. The viral tree (right) demonstrated that the New York isolate was most closely related to one found in dead birds in Israel, East Africa, and eastern Europe. This knowledge provided health officials with key information about the basic biology of the virus that was needed for diagnosis and predicting its spread. Such knowledge was critical in preventing human and animal infection.



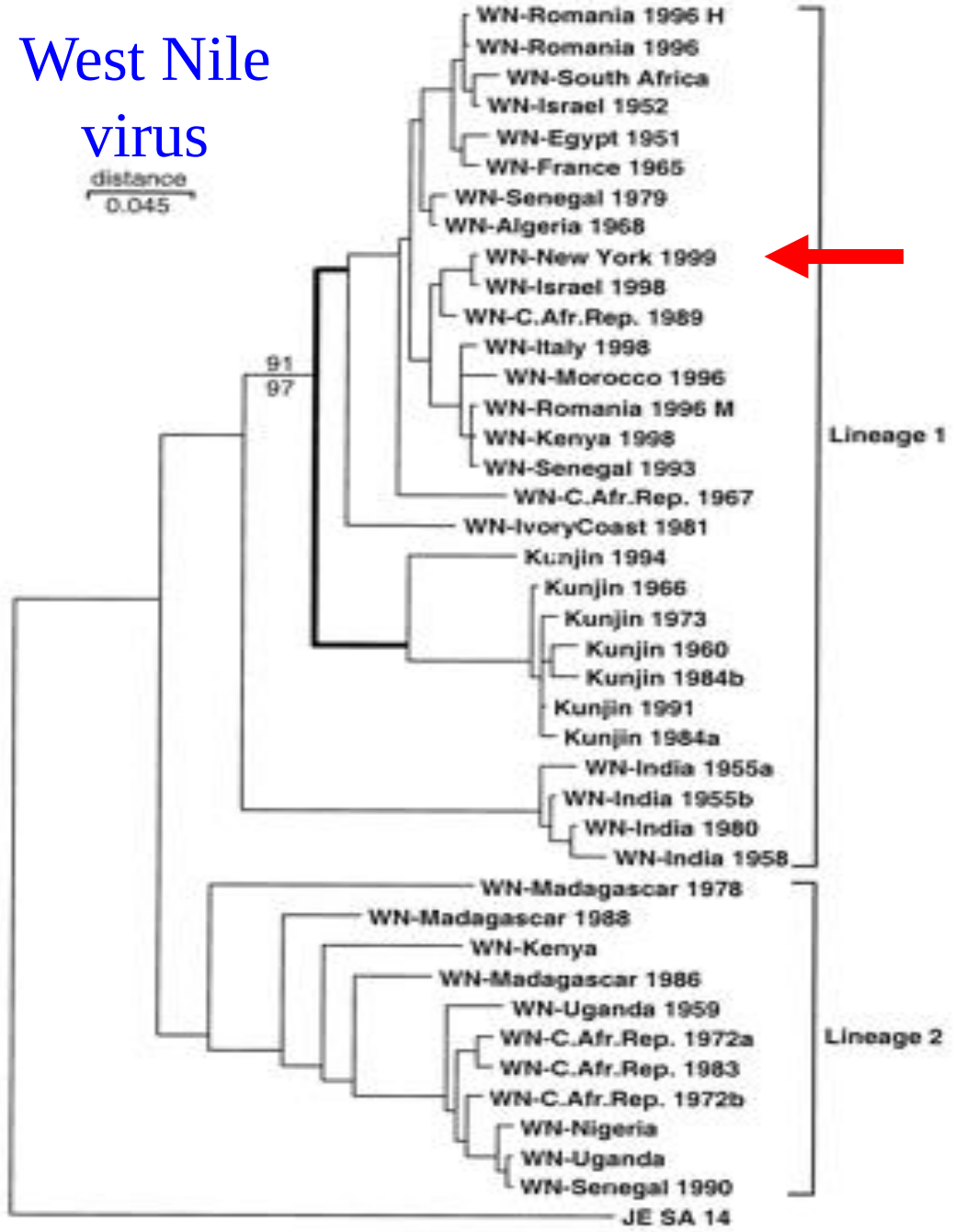
Isolation of the West Nile virus from flamingos and other captive and wild birds in the New York area was one of the first indications of a new emergent disease.

Several species of mosquitos, including the recently invasive tiger mosquito, transmit West Nile virus from bird and mammal hosts to human populations. Understanding the phylogenetic relationships of disease vectors such as mosquitos is important in tracing the origin and spread of diseases.



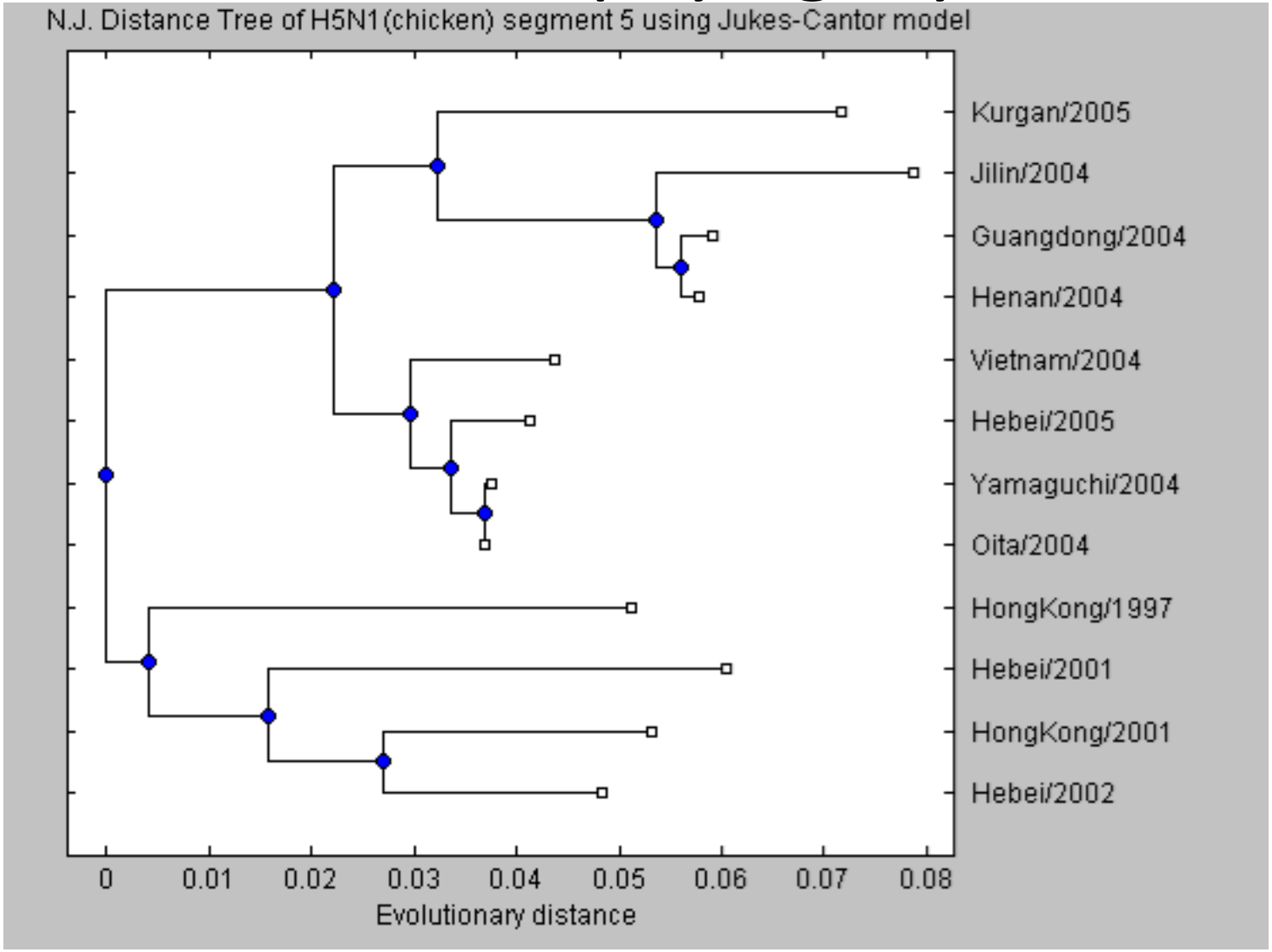


West Nile virus

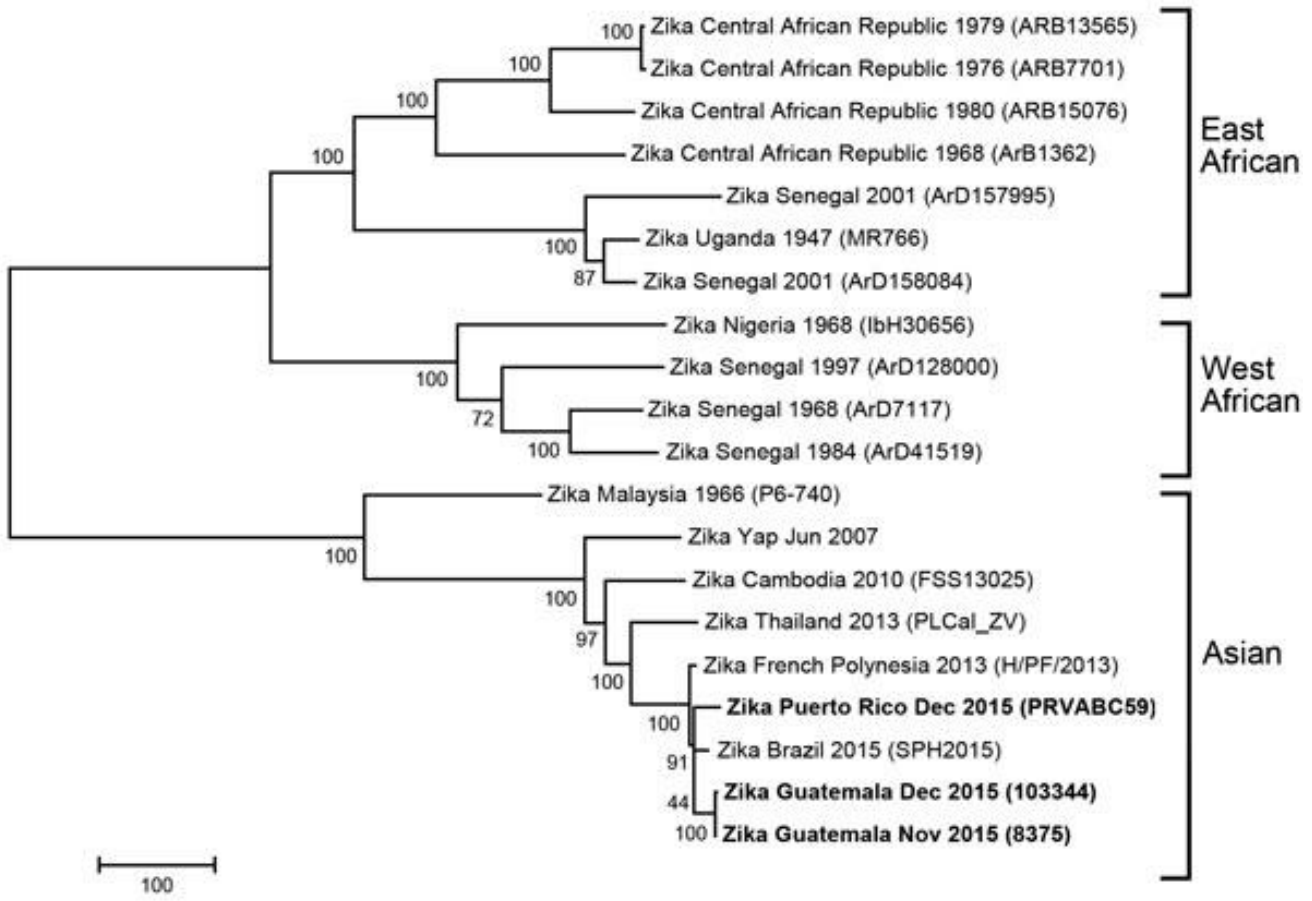


Lanciotti et al. (1999)

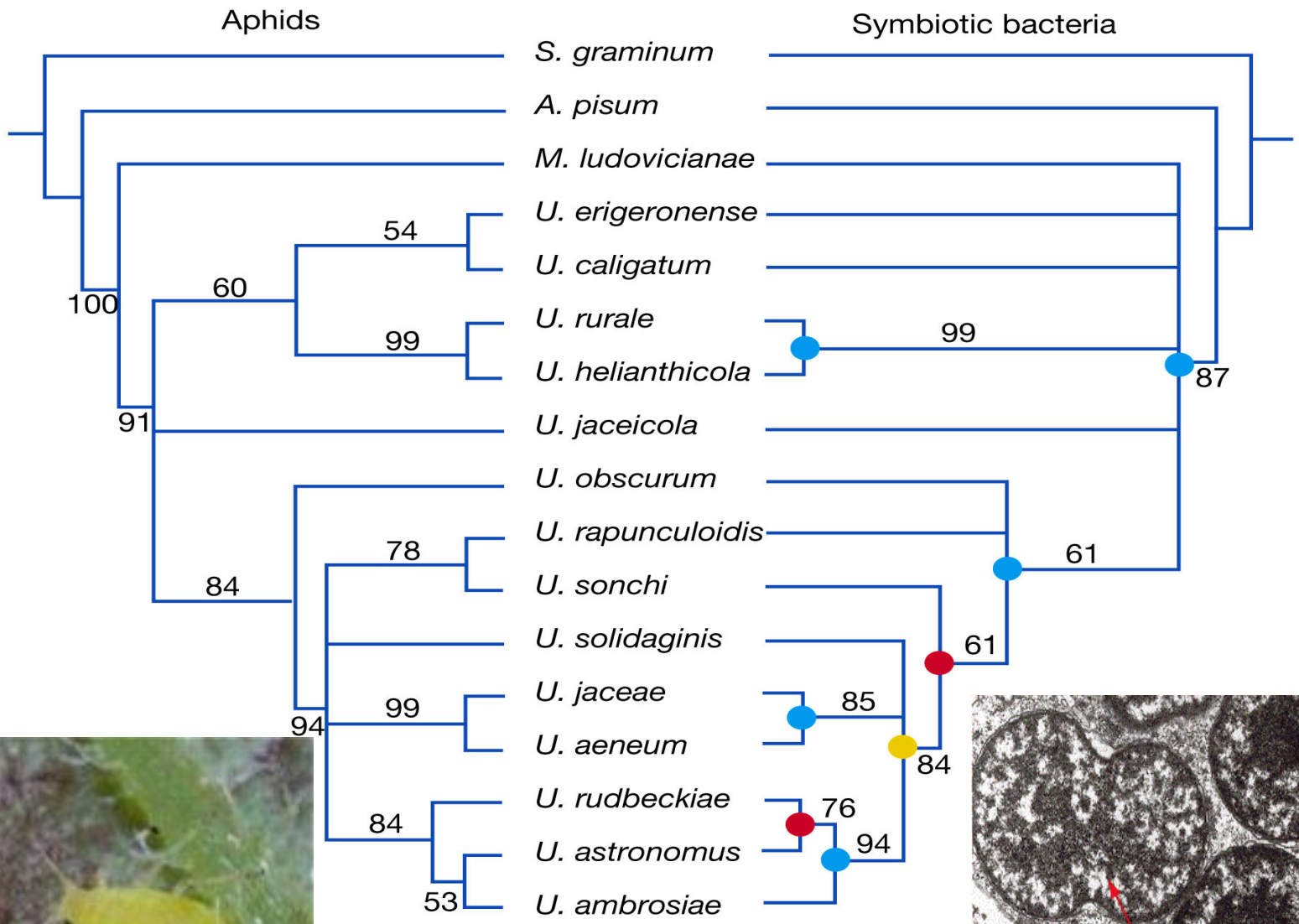
Bird flu phylogeny

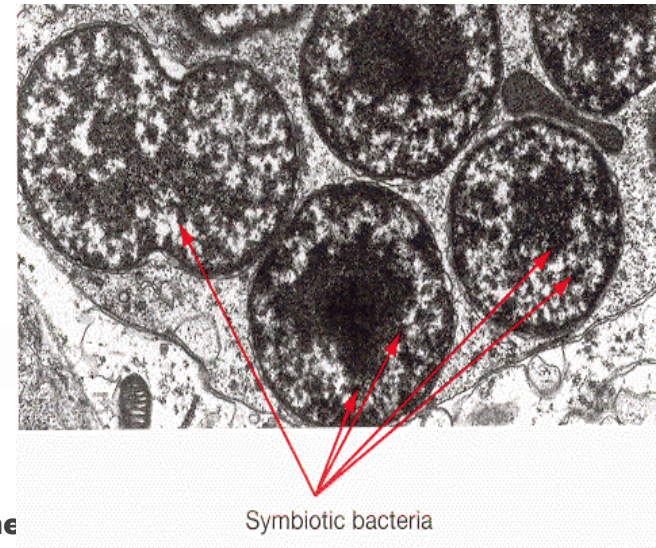


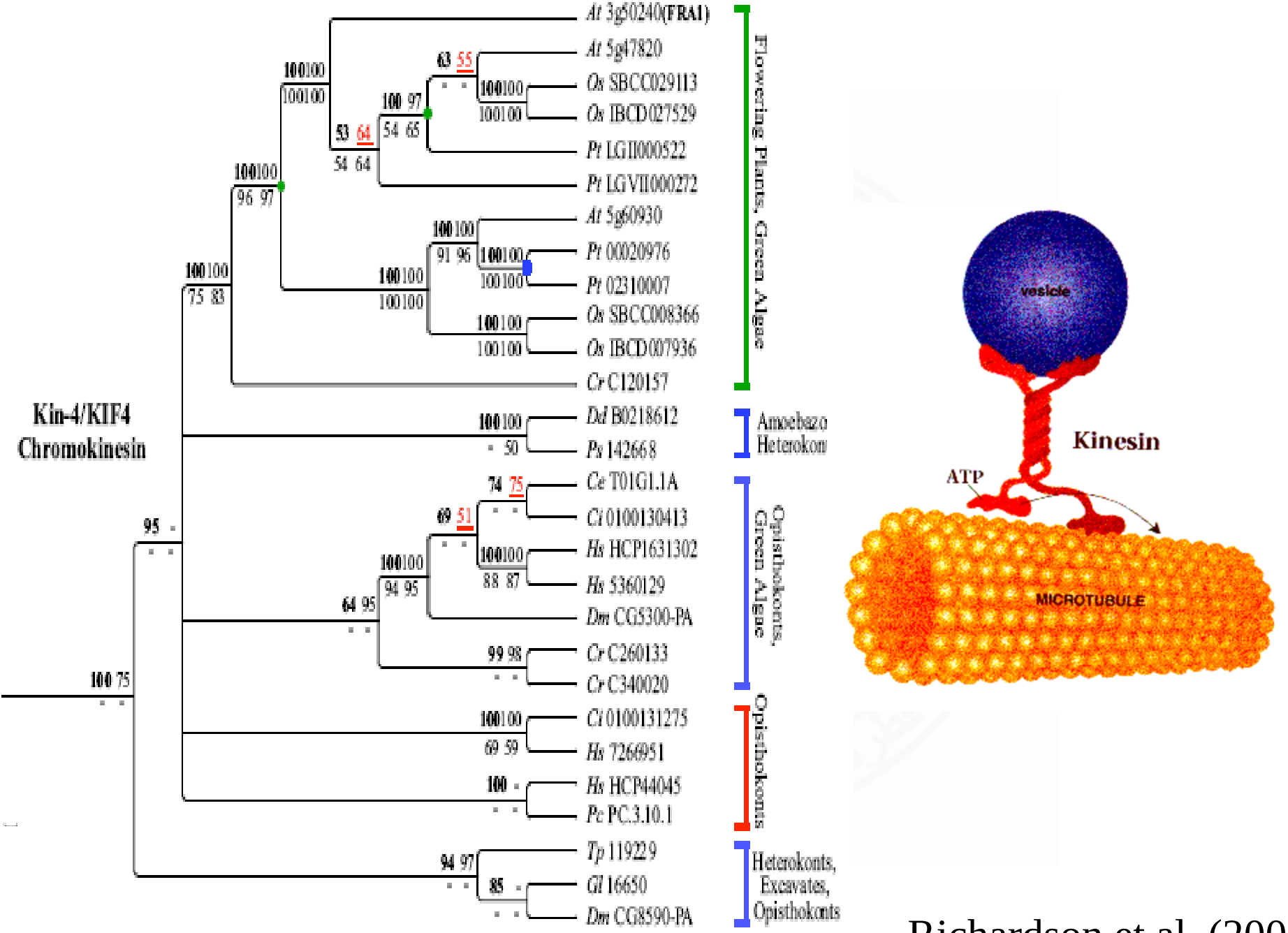
Zika Virus



Lanciotti, Robert S., Amy J. Lambert, Mark Holodniy, Sonia Saavedra, and Leticia del Carmen Castillo Signor. 2016. "Phylogeny of Zika Virus in Western Hemisphere, 2015." *Emerging Infectious Diseases* 22 (5). doi:10.3201/eid2205.160065.

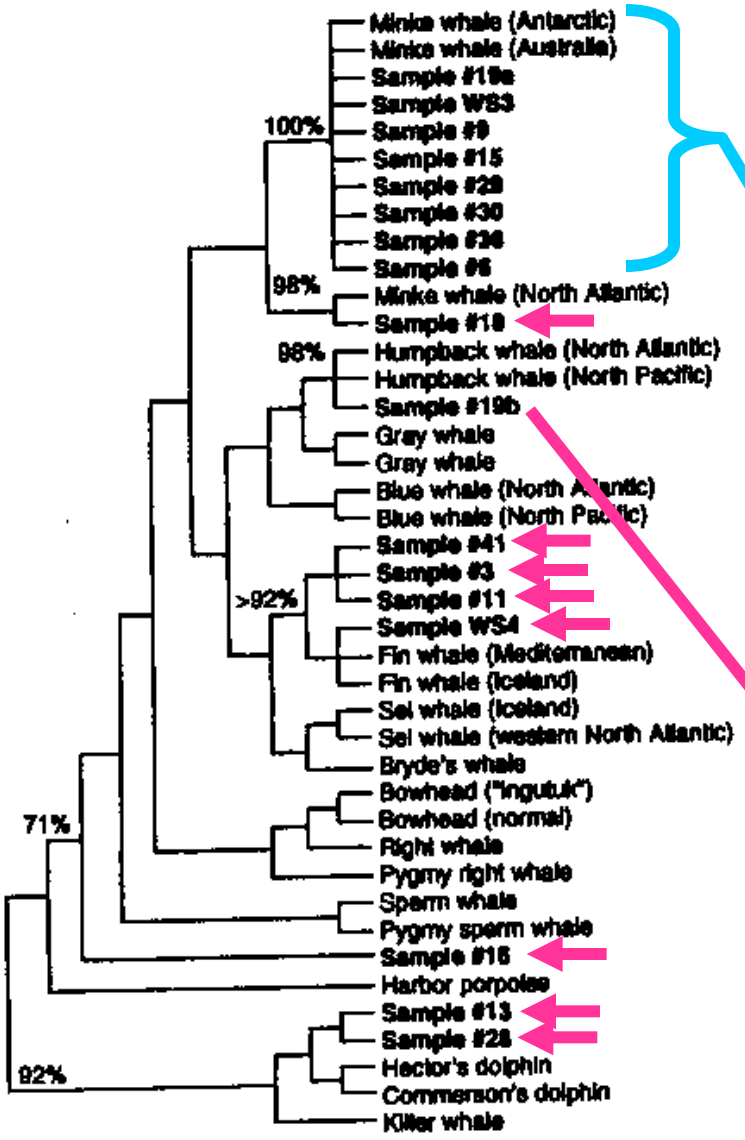

Aphids
Symbiotic bacteria
S. graminum
A. pisum
M. ludovicianae
U. erigeronense
U. caligatum
U. rurale
U. helianthicola
U. jaceicola
U. obscurum
U. rapunculoidis
U. sonchi
U. solidaginis
U. jaceae
U. aeneum
U. rudbeckiae
U. astronomus
U. ambrosiae
100
54
60
99
91
99
87
84
78
61
61
94
99
85
84
84
76
94
53


Symbiotic bacteria



Richardson et al. (2006)

Which Whales Are Hunted? A Molecular Genetic Approach to Monitoring Whaling



09-11-01

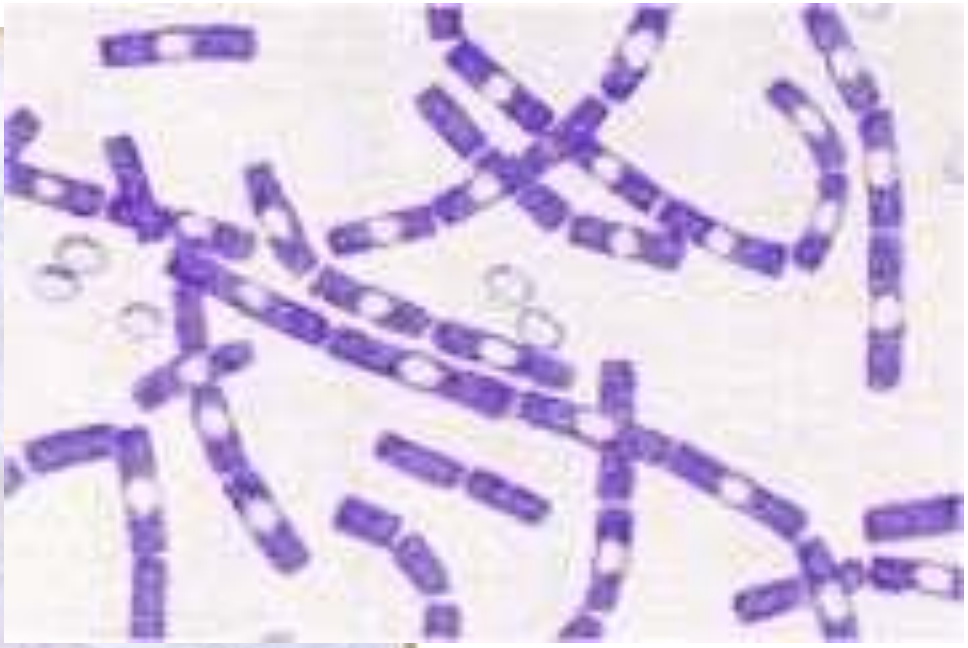
THIS IS NEXT

TAKE PENACILIN NOW

DEATH TO AMERICA

DEATH TO ISRAEL

ALLAH IS GREAT



TOM BROKAW
NBC TV
30 ROCKEFELLER PLAZA
NEW YORK NY 10112

Zuckerkandl & Pauling (1962)

- Zuckerkandl & Pauling (1962) demonstrated that molecular sequences provide large amounts of information and help understand evolutionary relationships

Sequence Data

	Characters					
Species	1	2	3	4	5	6
α	A	G	A	C	G	G
β	C	G	T	G	A	G
γ	A	C	A	G	A	G
δ	A	C	A	C	G	A
ϵ	C	G	T	C	G	G

Nowadays, biologists rely on molecular sequence data, in particular DNA or RNA sequences, which allows the comparison of a broader range of species.

E. coli, yeast, clam shell, human, …….

Species vs. Gene Trees

- The taxonomic units (nodes of the tree) can represent genes, species or populations.
- A gene tree represents the evolutionary history of a single gene; e.g. the evolution of the globin family.
- A species tree will generally be built using multiple genes.
- We will focus mainly on species trees.

Methods for Tree Reconstruction

Distance based

- Tree is constructed on the basis of a distance measure between species.

Maximum parsimony

- The tree that explains the data with the smallest number of evolutionary events.

Maximum likelihood

- Statistical method of phylogeny reconstruction
- Explicit model for how data set is generated with nucleotide or amino acid substitutions
- Find topology that maximizes the probability of the observed data given the model and the parameter values (estimated from data)

Distance Based Tree Reconstruction

Definition: A distance measure d_{ij} between *n* species is said to be a tree-derived distance if there exists a tree T with the species at the leaves such that d_{ij} is the sum of the branch lengths along a path between i and j.

- Tree-derived distances are typically called additive distances.
- In general we don't have available to us the "real" tree-distance. Given surrogate information about the tree-distance, we would like to obtain the best estimate of the phylogenetic tree.

Tree Reconstruction - Ultrametric Case

If:

Pairwise distances d_{ij} among *n* extant species satisfy the Ultrametric property

Then:

There exists a unique rooted tree (up to trivial relabeling) with these species as leaves, that gives these distances.

Thus the original tree can be recovered exactly.

Proof: See Ewens and Grant – Statistical Methods in Bioinformatics (2nd edition)

UPGMA (Unweighted Pair Group Method using Averages)

- Uses a hierarchical clustering method. Initially each species is in its own cluster. The algorithm sequentially combines two clusters and forms a new node in the tree.

- First define distance d_{ij} between two clusters C_i and C_j to be the average distance between pairs of species from each cluster:

$$d_{ij} = \frac{1}{|C_i||C_j|} \sum_{p \text{ in } C_i, q \text{ in } C_j} d_{pq},$$

where $|C_i|$ denotes the number of species in cluster C_i.

- Note that if C_k is the union of C_i and C_j and C_l is another cluster, then

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}.$$

This is also known as **average linkage** cluster analysis

Alternatives are “maximum” or “minimum” linkage

UPGMA

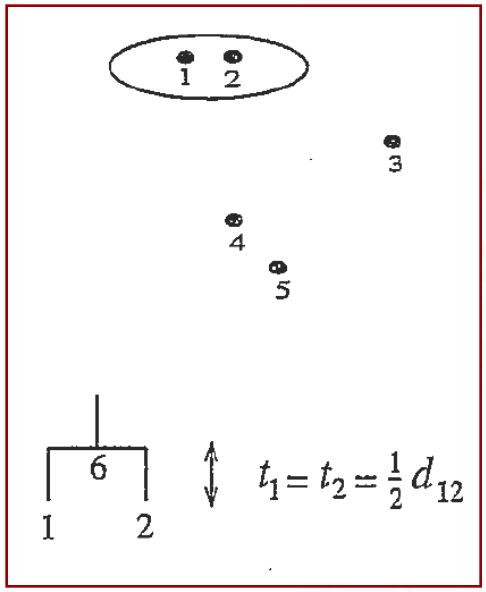
- Initialization:
  - Assign each species i to its own cluster C_i
  - Define a leaf of T for each species. Place it at height zero.
- Iteration:
  - Find i,j so that d_{ij} is minimal (in case of ties, choose randomly)
  - Define new cluster k by $C_k = C_i$ union C_j. Define d_{kl} for all l by

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}.$$

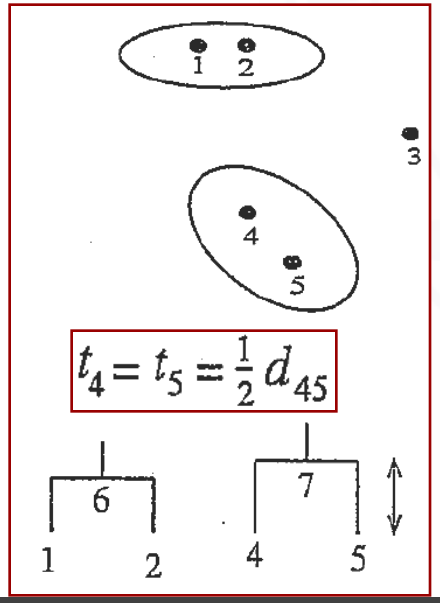
  - Define a node k with "daughter" nodes i and j and place it at height $d_{ij}/2$.
  - Add k to the current clusters and remove i and j.
- Termination:
  - When only two clusters i,j remain, place the root at height $d_{ij}/2$.

UPGMA Example

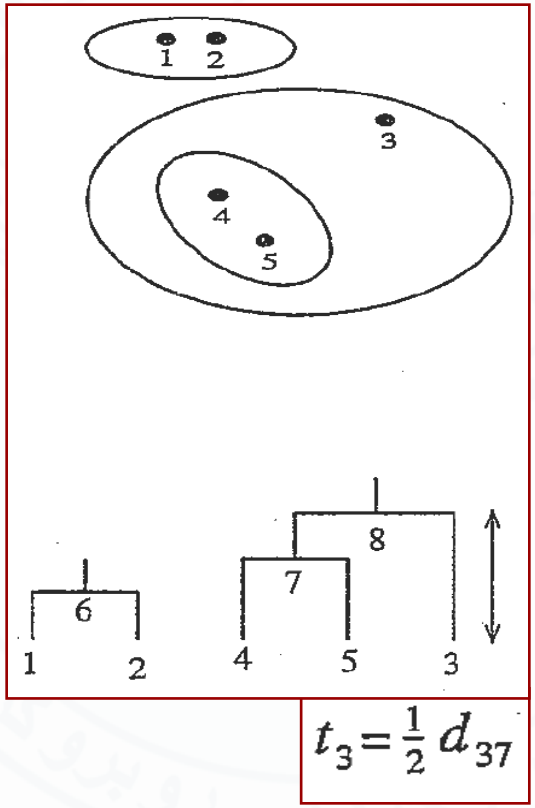
(1)



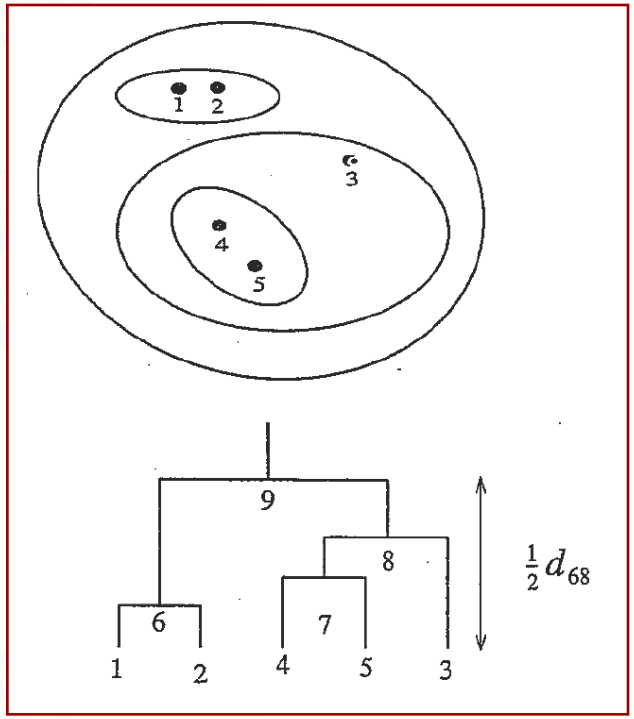
(2)



(3)



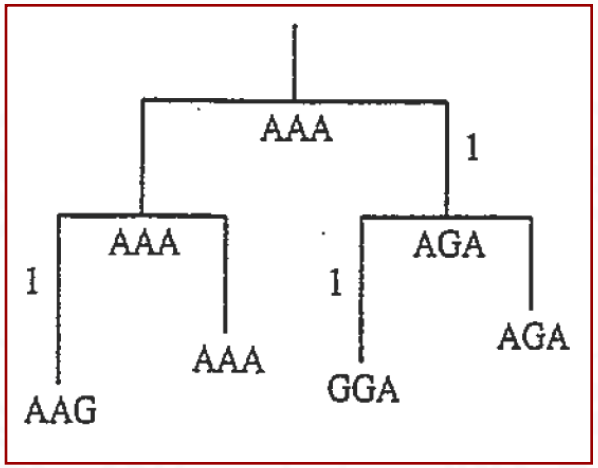
(4)



Other methods

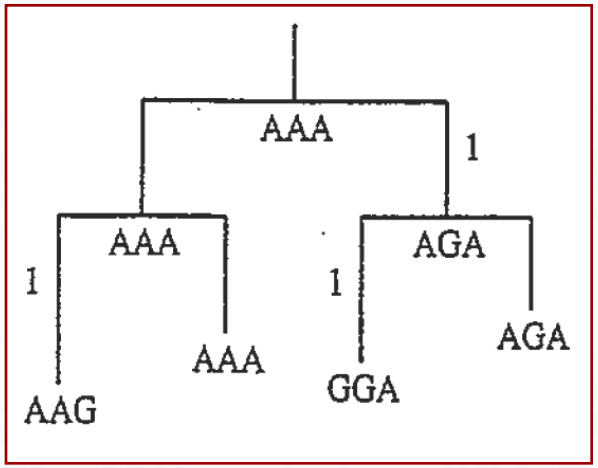
- Neighbor joining
  - constructs unrooted trees

Parsimony



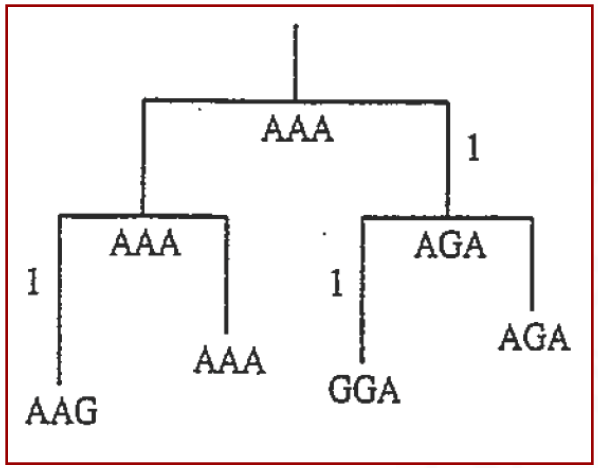
- One of the most commonly used methods for tree reconstruction
- Uses sequence information directly rather than distances that reflect sequence similarity.
- The idea: find a tree that explain the observed sequences with minimal number of substitutions (Occam's razor)

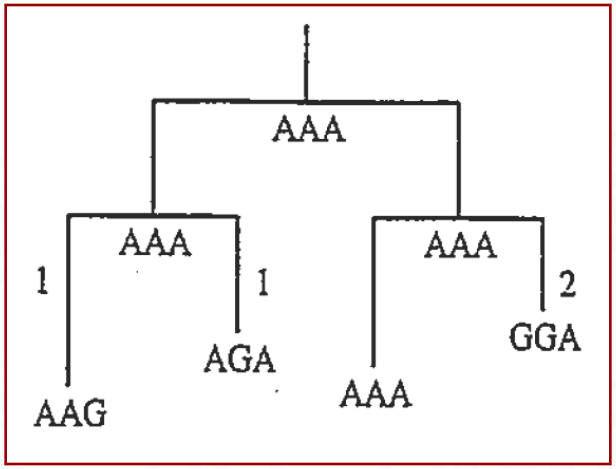
Parsimony

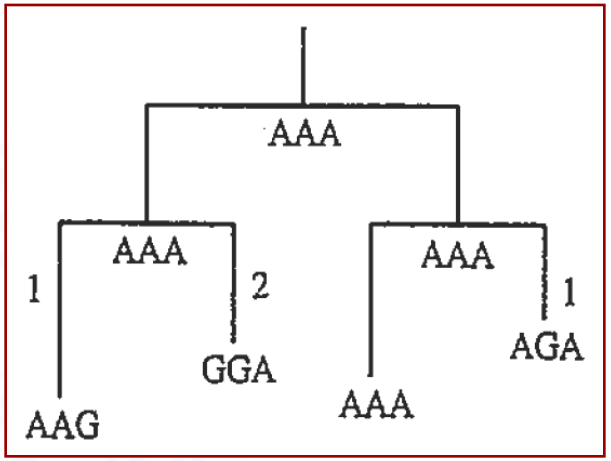


- Assign a score to a tree (with ancestral sequences included) as the total number of substitutions along all branches of the tree.
- The objective: find the tree that minimizes this score
- Minimization is over all choices of inferred ancestral sequences and tree topologies.

Parsimony







Suppose we are given the four sequences:

AAG, AAA, GGA, AGA

- For each possible topology for a rooted tree with 4 leaves, assign sequences to the ancestral nodes such that the total number of changes needed is minimized.
- Minimize this number over all trees

Parsimony

Two step process:

1. Given a tree topology and an assignment of residues to the leaf nodes, find an assignment of residues to internal nodes that minimizes the number of mutations (*note parsimony treats each position independently*)

   This is sometimes called the SMALL PARSIMONY PROBLEM.

   Solvable using dynamic programming

2. Search over all possible trees and find the "best" tree, that is, the tree having the minimal parsimony score.

   This is sometimes called the LARGE PARSIMONY PROBLEM.

Large Parsimony

# Species	# unrooted trees
4	3
5	15
6	105
7	945
8	10,395
9	1,35,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625

Exhaustive Search

For each **unrooted** tree with n leaves, calculate the minimum parsimony score and choose the tree with minimum score.

Feasible only for small n.

Parsimony score is independent of root position.

Large parsimony is NP complete

Maximum Likelihood phylogeny

- Given a tree T with *n* leaves, where sequence x^j is at leaf node j, and a set of edge lengths denoted by $t_1, t_2, \ldots, t_{2n-2}$ (written as **t**), we want to define

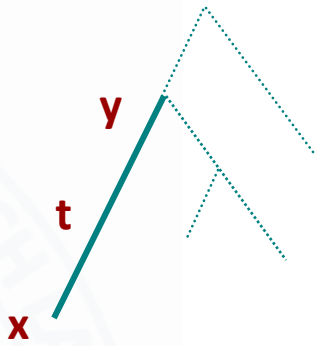
$$P(\mathbf{x} \mid T, \mathbf{t})$$

- This requires a model of evolution – probabilities of mutation events that change sequences along the edges of the evolutionary tree

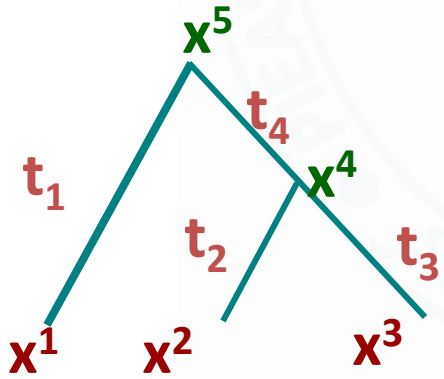
Probability Model

Define:

$P(x \mid y, t)$



the probability that a given ancestral sequence y will evolve into x along an edge of length t.



For any set of ancestral sequences assigned to internal nodes, we can calculate the probability of observing such a fully specified tree.

Maximum Likelihood Tree

- The tree with **topology T** and **edge lengths t** that **maximizes** $P(\mathbf{x} \mid T, \mathbf{t})$

- **Finding the maximum likelihood tree involves a search over:**
  - Tree topologies
  - Assignment to ancestor sequences
  - Edge lengths **t**

Phylogenetic Trees using MEGA

- MEGA is a commonly used tool for this purpose

Building Phylogenetic Trees from Molecular Data with MEGA

Barry G. Hall*
Bellingham Research Institute, Bellingham, Washington
*Corresponding author: E-mail: barryghall@gmail.com.
Associate editor: Joel Dudley

Abstract

Phylogenetic analysis is sometimes regarded as being an intimidating, complex process that requires expertise and years of experience. In fact, it is a fairly straightforward process that can be learned quickly and applied effectively. This Protocol describes the several steps required to produce a phylogenetic tree from molecular data for novices. In the example illustrated here, the program MEGA is used to implement all those steps, thereby eliminating the need to learn several programs, and to deal with multiple file formats from one step to another (Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 28:2731–2739). The first step, identification of a set of homologous sequences and downloading those sequences, is implemented by MEGA's own browser built on top of the Google Chrome toolkit. For the second step, alignment of those sequences, MEGA offers two different algorithms: ClustalW and MUSCLE. For the third step, construction of a phylogenetic tree from the aligned sequences, MEGA offers many different methods. Here we illustrate the maximum likelihood method, beginning with MEGA's Models feature, which permits selecting the most suitable substitution model. Finally, MEGA provides a powerful and flexible interface for the final step, actually drawing the tree for publication. Here a step-by-step protocol is presented in sufficient detail to allow a novice to start with a sequence of interest and to build a publication-quality tree illustrating the evolution of an appropriate set of homologs of that sequence. MEGA is available for use on PCs and Macs from www.megasoftware.net.

Protocol

A phylogenetic tree is an estimate of the relationships among taxa (or sequences) and their hypothetical common ancestors (Nei and Kumar 2000; Felsenstein 2004; Hall 2011). Today most phylogenetic trees are built from molecular data: DNA or protein sequences. Originally, the purpose of most molecular phylogenetic trees was to estimate the relationships among the species represented by those sequences, but today the purposes have expanded to include understanding the relationships among the sequences themselves without regard to the host species, inferring the functions of genes that have not been studied experimentally (Hall et al. 2009), and elucidating mechanisms that lead to microbial outbreaks (Hall and Barlow 2006) among many others. Building a phylogenetic tree requires four distinct steps: (Step 1) identify and acquire a set of homologous DNA or protein sequences, (Step 2) align those sequences, (Step 3) estimate a tree from the aligned sequences, and (Step 4) present that tree in such a way as to clearly convey the relevant information to others.

Typically you would use your favorite web browser to identify and download the homologous sequences from a national database such as GenBank, then one of several alignment programs to align the sequences, followed by one of many possible phylogenetic programs to estimate the tree, and finally, a program to draw the tree for exploration and publication. Each program would have its own interface and its own required file format, forcing you to interconvert files as you moved information from one program to another. It is no wonder that phylogenetic analysis is sometimes considered intimidating!

MEGA5 (Tamura et al. 2011) is an integrated program that carries out all four steps in a single environment, with a single user interface eliminating the need for interconverting file formats. At the same time, MEGA5 is sufficiently flexible to permit using other programs for particular steps if that is desired. MEGA5 is, thus, particularly well suited for those who are less familiar with estimating phylogenetic trees.

Step 1: Acquiring the Sequences

Ironically, the first step is the most intellectually demanding, but it often receives the least attention. If not done well, the tree will be invalid or impossible to interpret or both. If done wisely, the remaining steps are easy, essentially mechanical, operations that will result in a robust meaningful tree.

Often, the investigator is interested in a particular gene or protein that has been the subject of investigation and wishes to determine the relationship of that gene or protein to its homologs. The word "homologs" is key here. The most basic assumption of phylogenetic analysis is that all the sequences on a tree are homologous, that is, descended from a common ancestor. Alignment programs will align sequences, homologous or not. All tree-building programs will make a tree from that alignment. However, if the sequences are not actually descended from a common ancestor, the tree will be

PROTOCOL

BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 299–306 doi:10.1093/bib/bbn017
Advance Access publication April 16, 2008

MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences

Sudhir Kumar, Masatoshi Nei, Joel Dudley and Koichiro Tamura

Submitted: 18th January 2008; Received (in revised form): 8th March 2008

Abstract

The Molecular Evolutionary Genetics Analysis (MEGA) software is a desktop application designed for comparative analysis of homologous gene sequences either from multigene families or from different species with a special emphasis on inferring evolutionary relationships and patterns of DNA and protein evolution. In addition to the tools for statistical analysis of data, MEGA provides many convenient facilities for the assembly of sequence data sets from files or web-based repositories, and it includes tools for visual presentation of the results obtained in the form of interactive phylogenetic trees and evolutionary distance matrices. Here we discuss the motivation, design principles and priorities that have shaped the development of MEGA. We also discuss how MEGA might evolve in the future to assist researchers in their growing need to analyze large data set using new computational methods.

Keywords: *phylogenetics; genome; evolution; software*

Biologist-friendly software tools are crucial in this age of burgeoning sequence databases. These tools not only make it possible to use computational and statistical methods but also allow scientists to select methods and algorithms best suited to understand the function, evolution and adaptation of genes and species. The Molecular Evolutionary Genetics Analysis (MEGA) software aims to serve both of these purposes in inferring evolutionary relationships of homologous sequences, exploring basic statistical properties of genes and estimating neutral and selective evolutionary divergence among sequences (Figure 1).

The MEGA software project grew out of our own need for employing statistical methods in the phylogenetic analysis of DNA and protein sequences in the early 1990s. At this time, most computer programs available did not allow us to explore the primary data visually and lacked a user-friendly interface. There were two primary general-purpose computer programs for inferring phylogenetic trees. One was PAUP for constructing parsimony trees, and the other was PHYLIP for inferring phylogenetic trees using various character and statistical methods such as maximum likelihood, parsimony and distance methods [1, 2]. These programs were (and continue to be) very useful, but the former lacked statistical methods at that time and the latter did not provide a point-and-click user interface [3–6].

In order to make statistical methods available for phylogenetic analysis in a user-friendly manner,

Corresponding author. Sudhir Kumar, Biodesign Institute, A240, Arizona State University, Tempe, AZ 85287-5301, USA. E-mail: s.kumar@asu.edu
Sudhir Kumar is conducting large-scale analysis of genome sequences and spatial patterns of gene expression, and developing statistical methods and bioinformatics tools. He is the Director of the Center for Evolutionary Functional Genomics at Arizona State University, AZ, USA.
Masatoshi Nei is one of the founders of molecular evolutionary genetics and pursues statistical analysis of molecular and genome evolution. He is the Director of the Institute of Molecular Evolutionary Genetics at Pennsylvania State University, PA, USA.
Joel Dudley is interested in translational bioinformatics, genomic medicine and the application of molecular evolution in translational genomic research. He is a bioinformatics specialist at the Stanford Center for Biomedical Informatics Research at Stanford University, USA.
Koichiro Tamura's research interests are in the area of molecular evolution with emphasis on the pattern of gene and genome evolution, and on development of statistical methods and computer programs. He is an associate professor at the Tokyo Metropolitan University, Japan.

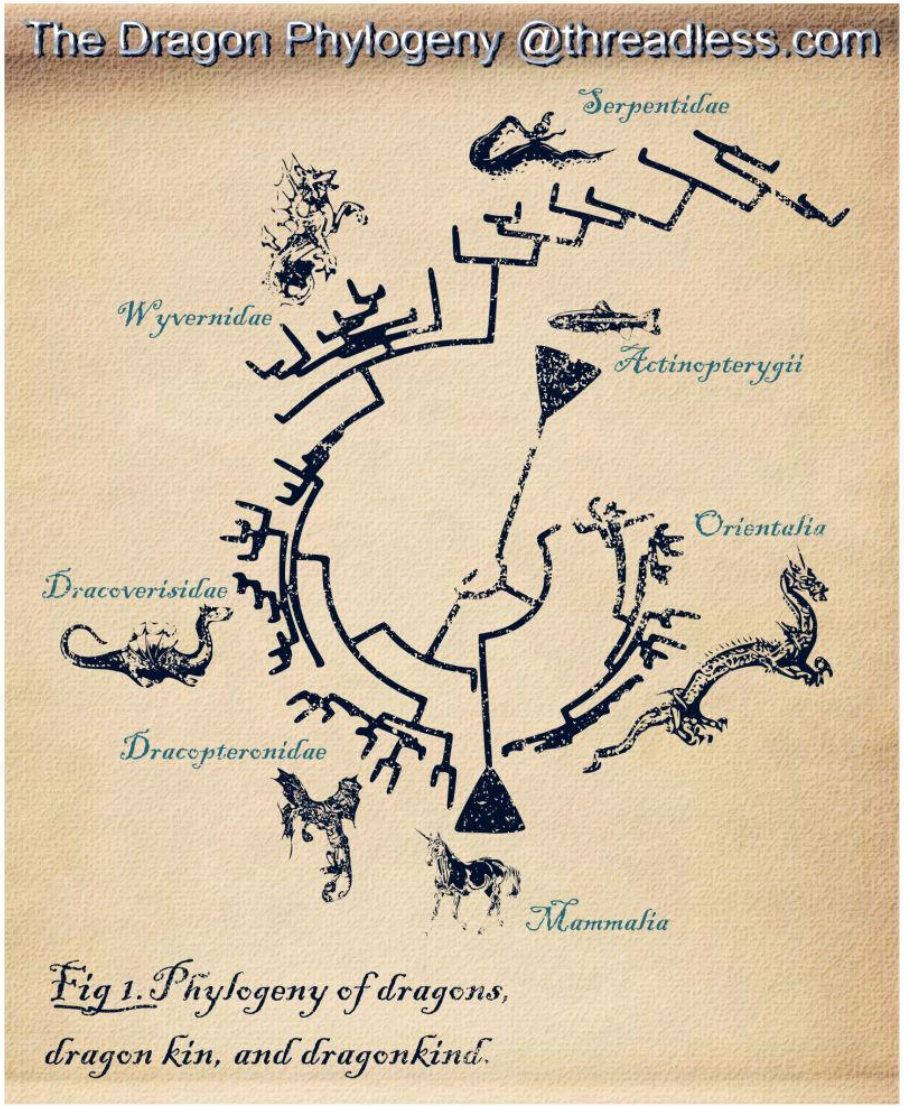
MEGA steps

- BLAST the given sequence to get more sequences – must be done carefully
- Construct MSA using MUSCLE or CLUSTALW
  - MUSCLE is better
- Construct tree
  - Distance based (distance matrix …)
  - Parsimony based
  - Maximum likelihood based
- Generate confidence values (using bootstrapping)
  - Optional but recommended

For Dengue viruses

- http://faculty.pieas.edu.pk/fayyaz/ippy/html_demos/Bioinformatics.html

End of Lecture



http://io9.gizmodo.com/5936427/the-evolutionary-history-of-dragons-illustrated-by-a-scientist