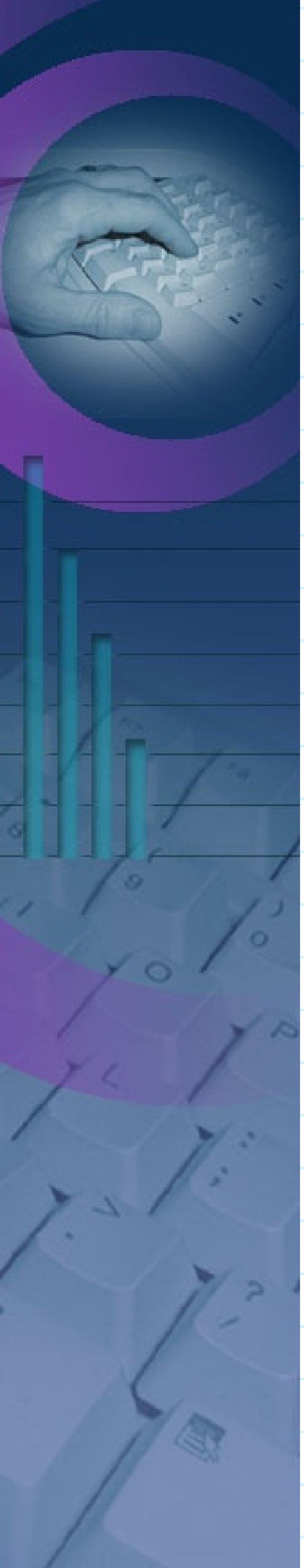
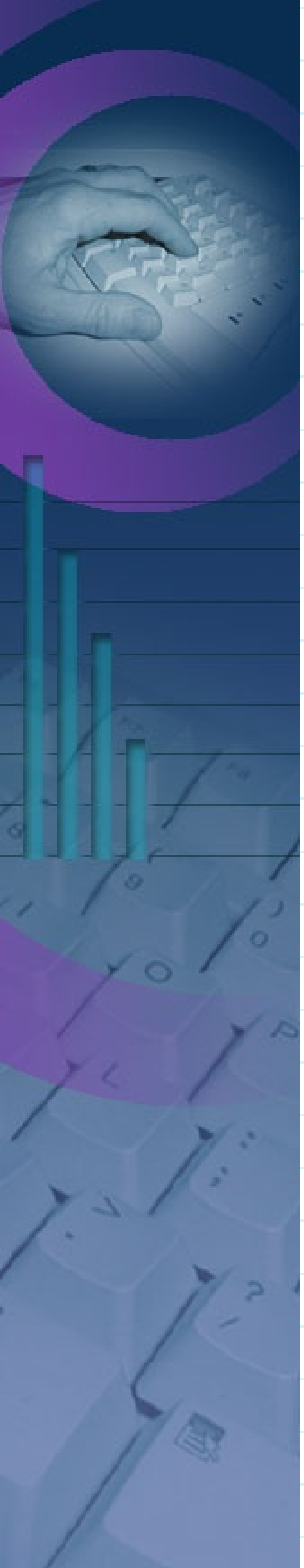


Phylogenetic Trees

What They Are
Why We Do It
&
How To Do It

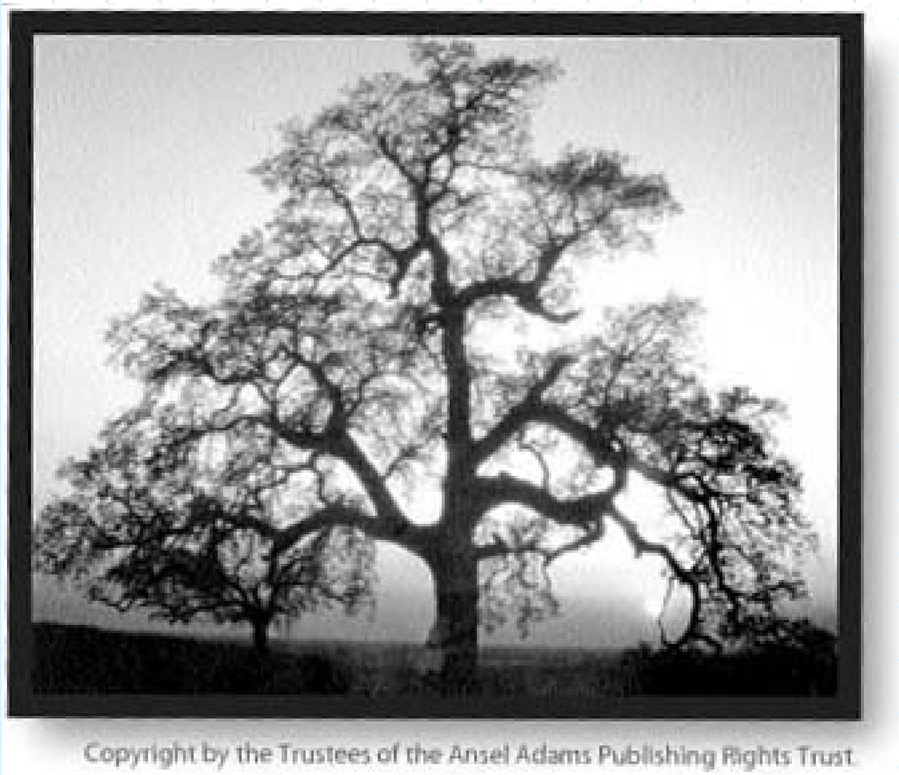
Presented by
Amy Harris
Dr Brad Morantz



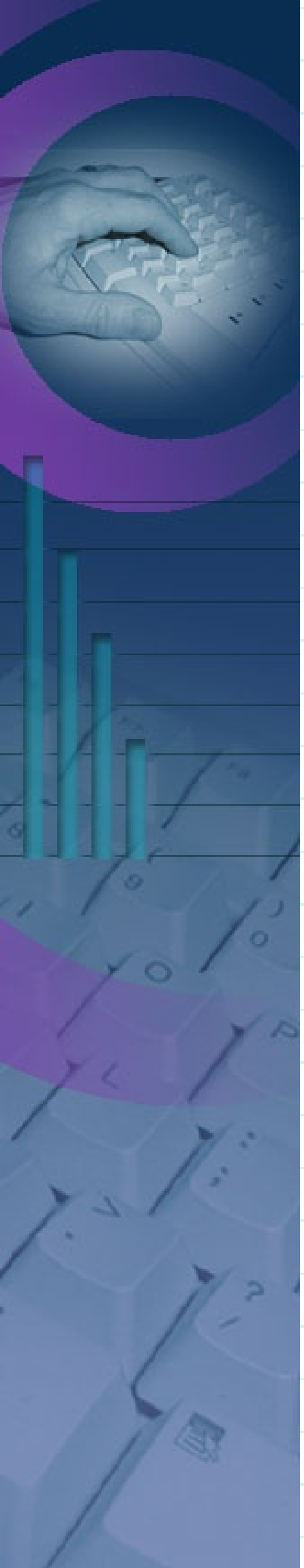
Overview

- What is a phylogenetic tree
- Why do we do it
- How do we do it
- Methods and programs
- Parallels with Genetic Algorithms (time permitting)

Definition: Phylogenetic Tree



A tree (graphical representation) that shows evolutionary relationships based upon common ancestry. Describes the relationship between a set of objects (species or taxa). [Israngkul]



Phylogenetic Tree

- Finding a tree like structure that defines certain ancestral relationships between a related set of objects. [Reijmers et al, 1999]
- Composed of branches/edges and nodes
- Can be gene families, single gene from many taxa, or combination of both. [Baldauf, 2003]

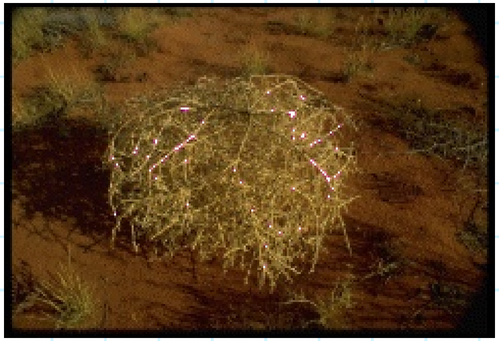
Terminology

- Branches – connections between nodes
- Evolutionary tree – patterns of historical relationships between the data
- Leaves – terminal node; taxa at the end of the tree
- Nodes – represent the sequence for the given data; Internal nodes correspond to the hypothetical last common ancestor of everything arising from it. [Baldauf, 2003]
- Taxa (a car for hire) – individual groups
- Tree (number after two) - mathematical structure consisting of nodes which are connected by branches

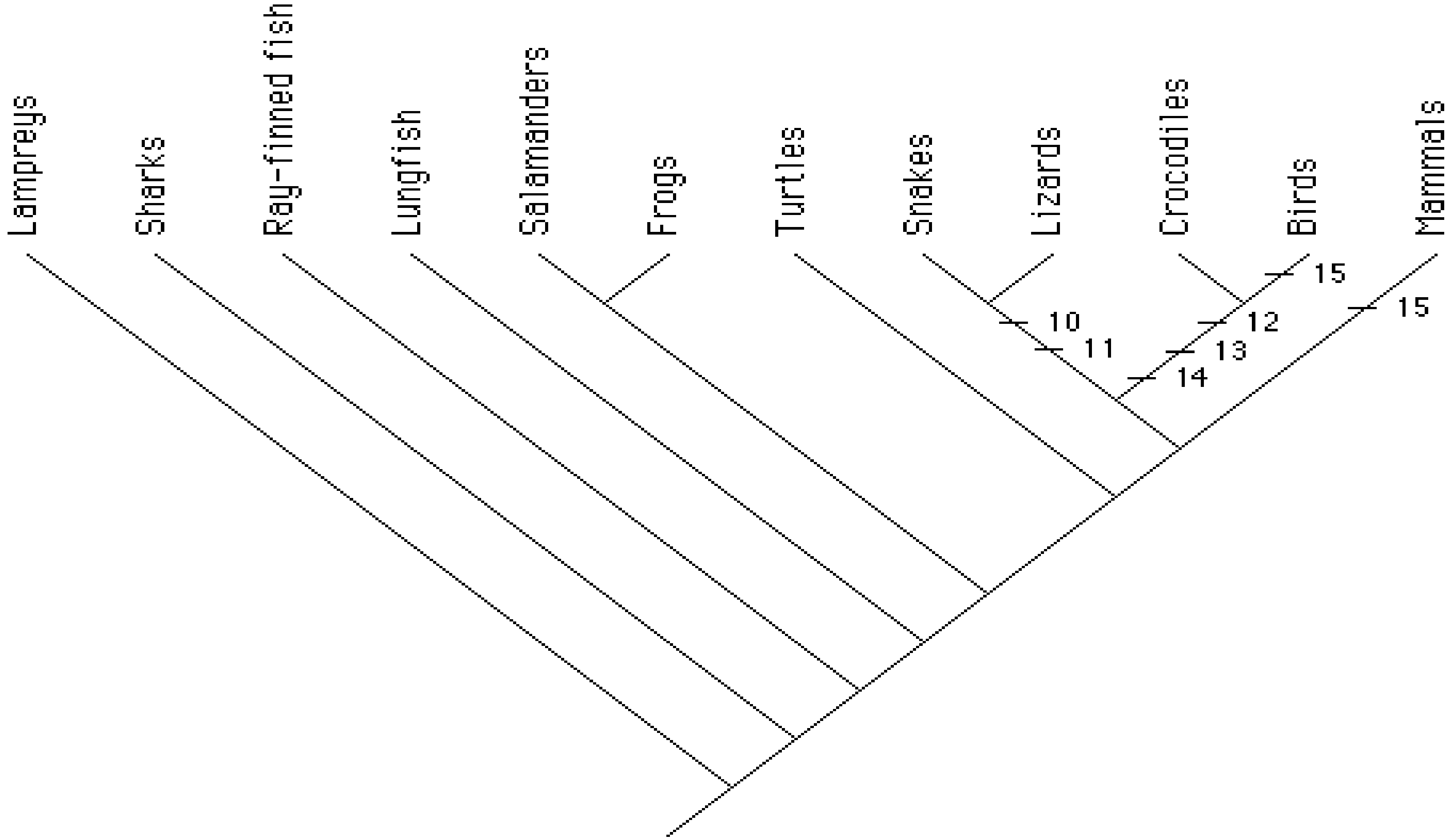
Rooted vs Unrooted

- Rooted tree
- Directed tree
  - Has a path
- Accepted common ancestor
- Doesn't blow over in the wind

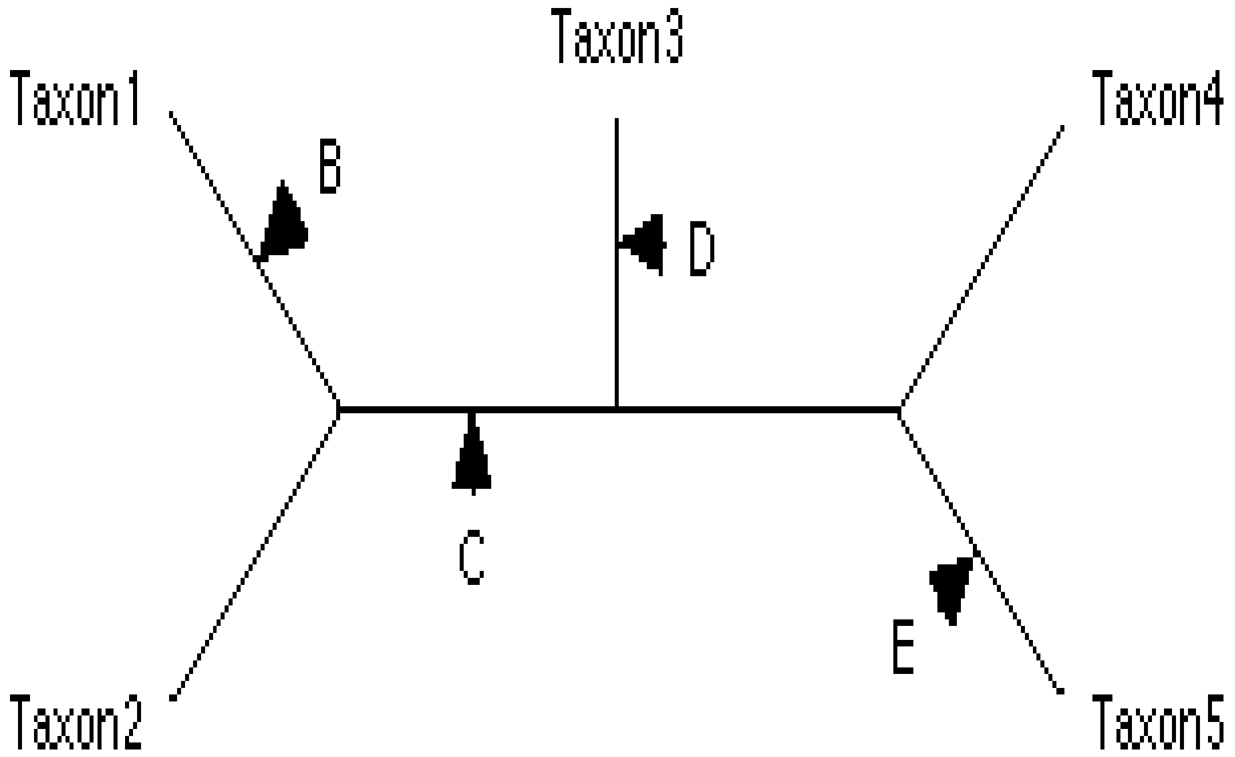
- Unrooted Tree
- Typical results
- Unknown common ancestor
- Common in Arizona, blowing around plains



Rooted Phylogenetic Tree



Unrooted Phylogenetic Tree



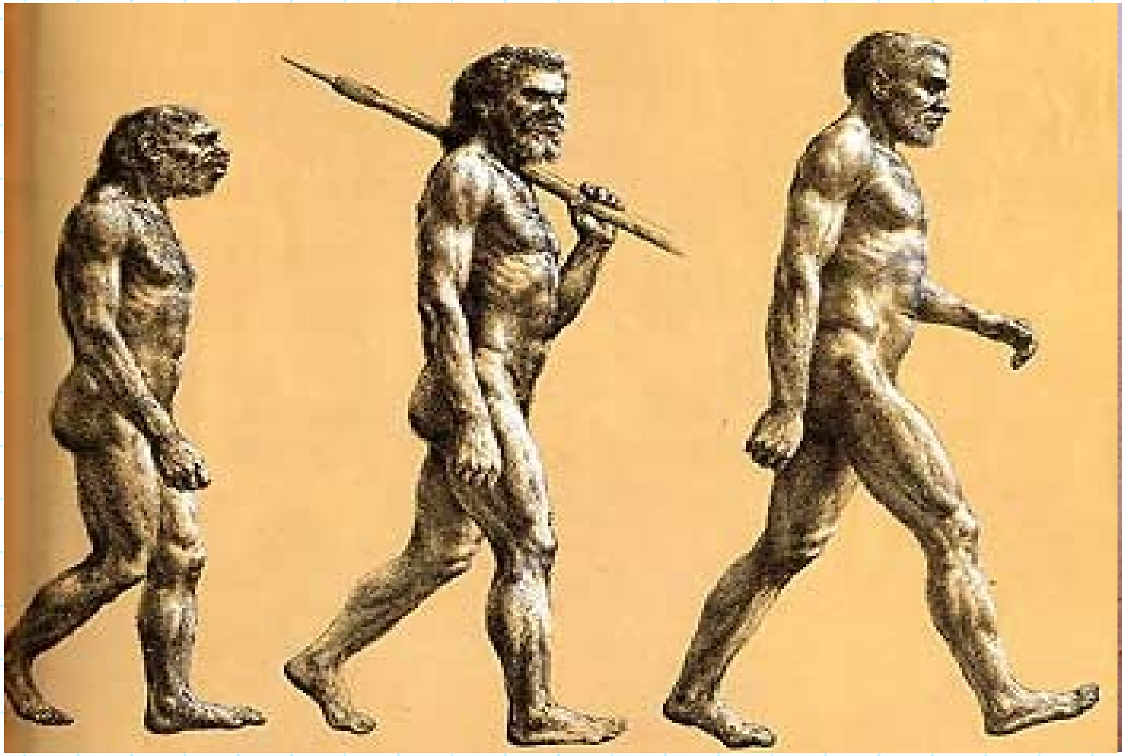
Combinatorial Explosion

- Number of topologies =
  - Product i =3 to x (2i – 5)
  - Ten objects yields 2,027,025 possible trees
  - 25 objects yields about 2.5 x 10^{28} trees
- Branches
  - 2x -3 branches
    - X are peripheral
    - (X - 3) are interior

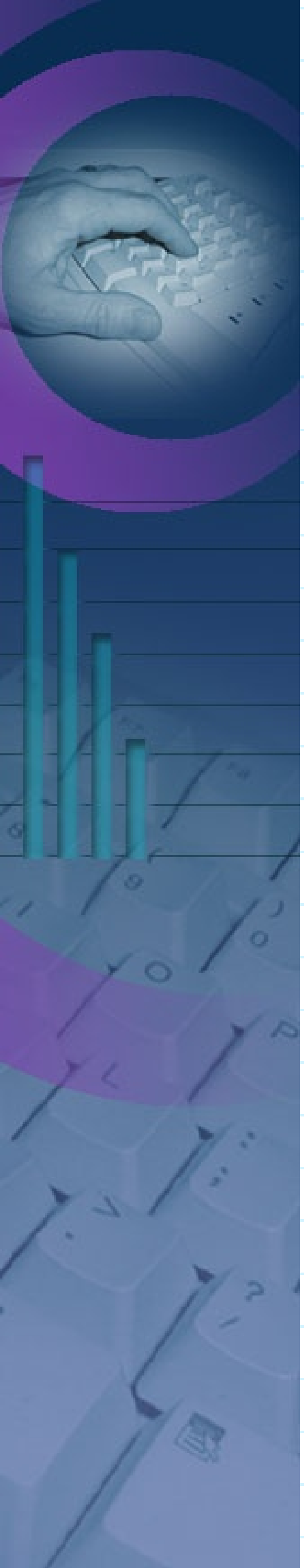
Why Do We Do It?

- Understand evolutionary history
  - Show visual representation of relationships and origins
  - Healthcare
    - Origins of diseases
    - Show how they are changing
    - Help produce cures and vaccines
- Model allows calculations to determine distances
  - Forensics
  - Our history

Our Family



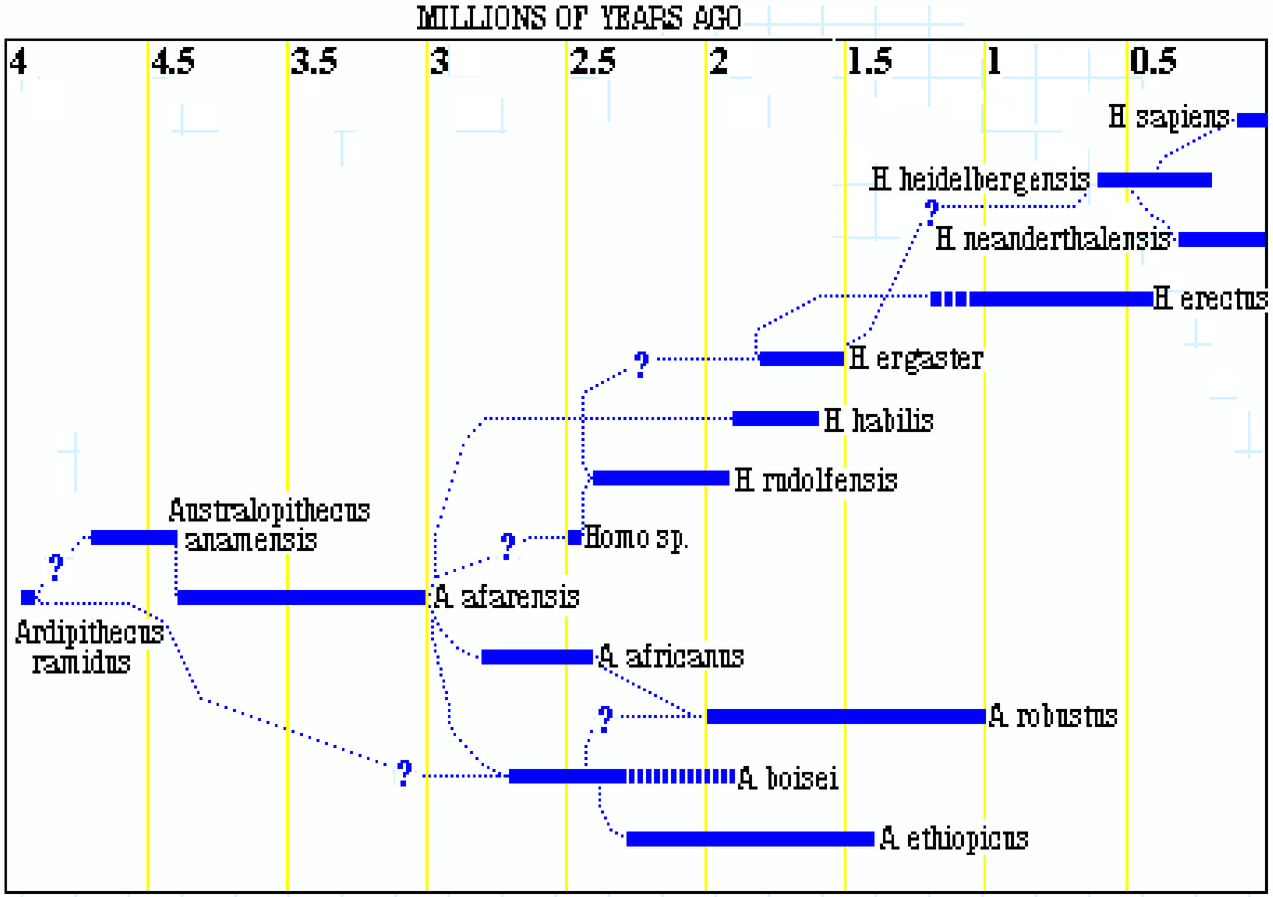
Terminal Node



Evolution

- Theory that groups of organisms change over time so that descendants differ structurally and/or functionally from their ancestors. [Pevsner, 2003]
- Biological process by which organisms inherit morphological and physiological features than define a species. [Pevsner, 2003]
- Biological theory that postulating that the various types of animals and plants have their origin in other preexisting types and that distinguishable differences are due to modifications in successive generations. [Encyclopaedia Britannica, 2004]

Example


MILLIONS OF YEARS AGO
4
4.5
3.5
3
2.5
2
1.5
1
0.5
H sapiens
H heidelbergensis
H neanderthalensis
H erectus
H ergaster
H habilis
H rudolfensis
Australopithecus anamensis
Homo sp.
A afarensis
Ardipithecus ramidus
A africanus
A robustus
A boisei
A ethiopicus
?

Ways to do it

- Parsimony
- Maximum Likelihood Estimator
- Distance based methods
- Clustering
- Genetic algorithm

* While there are numerous methods, these are among the most popular

Parsimony

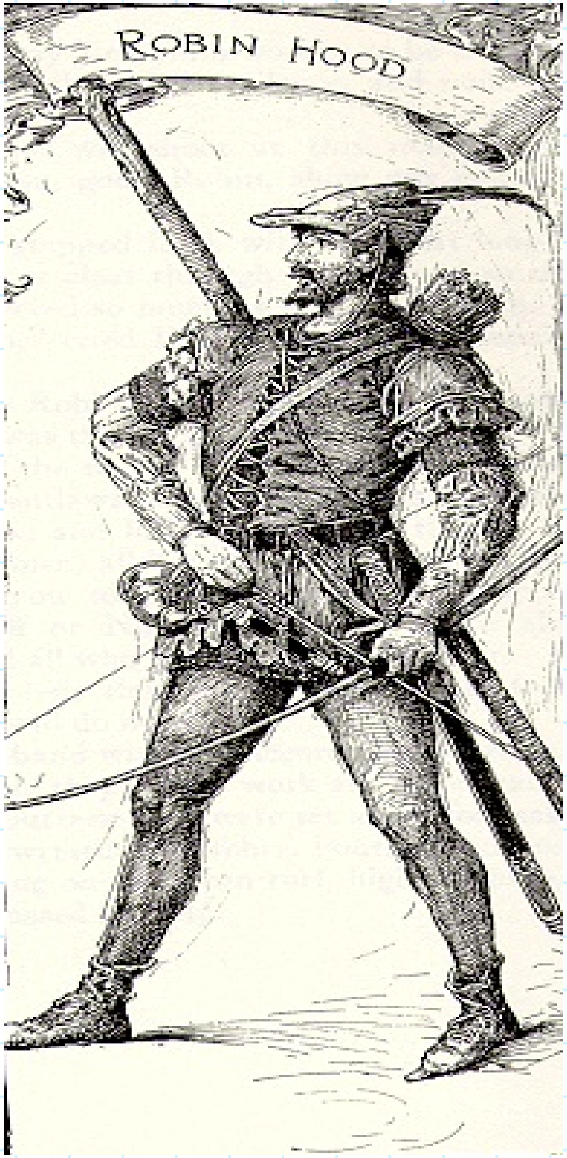
Parsimony

- Character based
- Search for tree with fewest number character changes that account for observed differences
- Best one has the least amount of evolutionary events required to obtain the specific tree
- Advantages:
  - Simple, intuitive, logical, and applicable to most models
  - Can be used on a wide variety of data
  - More powerful approach than distance to describe hierarchical relationship of genes & proteins (Pevsner, 2003)
- Disadvantages;
  - No mathematical origins
  - Fooled by same multiple or circuitous changes

Parsimony Methods

- The best tree is the one that minimizes the total number of mutations at all sites [*Israngkul*]
- The assumption of physical systematics is that genes exist in a nested hierarchy of relatedness and this is reflected in a hierarchical distribution of shared characters in the sequence. (Pevsner, 2003)

Maximum Likely Hood



Maximum Likelihood Estimator

- Tree with highest probability of evolving from given data
- Mathematical Process
  - Complex Math – many have problems with this
- Advantages:
  - Can be used for various types of data including nucleotides and amino acids
  - Usually the most consistent
- Disadvantages:
  - Computationally intense
  - Can be fooled by multiple or circuitous changes

Distance Based Methods

- Uses distances between leaves
  - Upper triangular matrix of distances between taxa
- Percent similarity
- Metric
  - Number of changes
  - Distance score
- Produces edge weighted tree
- Least squares error
  - Can use Matlab

Distances

- Hamming distance
  - n = # sites different
  - N = alignment length
  - $D = 100\% \times (n/N)$
  - ignore information of evolutionary relationship
- Jukes-Cantor
  - $D = -3/4 \ln (1-4P/3)$
- Kimura
  - Transitions more likely than transversions
  - Transitions given more weight

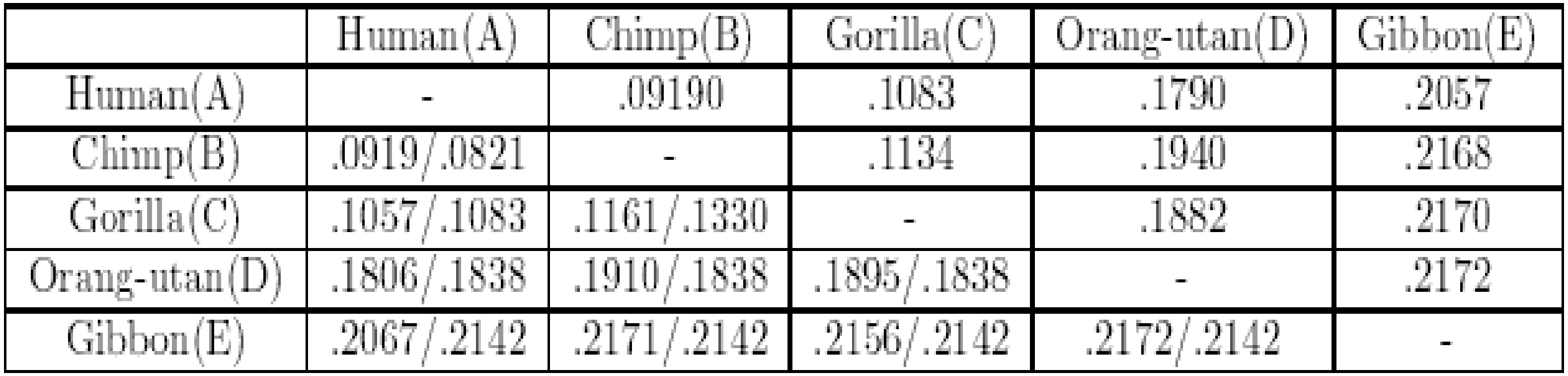
Distances

- The walk from the parking lot
- Now that's far!

Least Squares

- Start with distance matrix
- Pick tree type to start
- Calculate distances to minimize SSE
- Try other trees
- Time consuming
- Exhaustive search will yield optimal tree, but also may take l-o-n-g time

Example of Distance Matrix



	Human(A)	Chimp(B)	Gorilla(C)	Orang-utan(D)	Gibbon(E)
Human(A)	-	.09190	.1083	.1790	.2057
Chimp(B)	.0919/.0821	-	.1134	.1940	.2168
Gorilla(C)	.1057/.1083	.1161/.1330	-	.1882	.2170
Orang-utan(D)	.1806/.1838	.1910/.1838	.1895/.1838	-	.2172
Gibbon(E)	.2067/.2142	.2171/.2142	.2156/.2142	.2172/.2142	-

Clustering

- Genetic algorithm
- Neighbor joining
- UPGMA
- PAUP contains the last two

Clustering

- Neighbor joining
  - Uses distances between pairs of taxa
    - i.e. Number of nucleotide differences
    - Not individual characters
  - Builds shortest tree by complex methods
- UPGMA
  - Unweighted Pair Group Method w/ Arithmetic Mean
  - Starts with first two most similar nodes
  - Compares this average/composite to the next
  - Never uses original nodes again

Summary

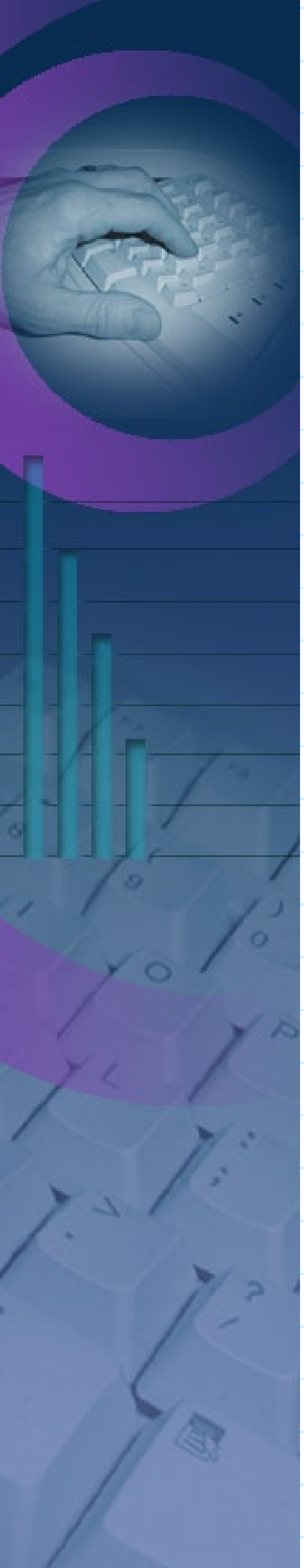
- Real life is usually not the optimum tree
- The best model is one that is obtained by several methods

Comments

- Sometimes difficult because
  - Do not have complete fossil record
  - Parallel evolution
  - Character reversals
  - Circuitous changes
- Bifurcating vs polytomy split
- No animals, plants, or robots were hurt in the making of this presentation

References

- *Baldauf, S.: 2003, Phylogeny for the faint of heart: a tutorial, Trends in Genetics, 19(6) pp. 345-351*
- *Encyclopaedia Brittanica; 2004, The Internet http://www.brittannica.com*
- *Futuyma, D; 1998, Evolutionary Biology, Third Edition, Sinauer Associates, Sunderland MA*
- *Israngkul, W; 2002, Algorithms for Phylogenetic Tree Reconstruction, the Internet: biohpc.learn.in.th/files/contents2003/Worawit/ Algorithms*
- *Opperdoes, F., 1997 Construction of a Distance Tree Using Clustering with the UPGMA, the Internet : http://www.icp.ucl.ac.be/~opperd/private/upgma.html*
- *Pevsner, J.; 2003, Bioinformatics & Functional Genomics, Wiley, Hoboken NJ*
- *Reijmers, T., Wehrens, R., Daeyaert, F., Lewi, P., & Buydens, L.; 1999, Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences, BioSystmes (49) pp. 31-43*
- *Renaut, R., 2004 Computational BioSciences Class, CBS-520, Arizona State University*



Genetic Algorithms

- Optimizing distance clustering (Reijmers et al)
- Optimal Distance method
- Not guaranteed the most optimal, only near optimal
- Exhaustive exploration not guaranteed
- Same solution may be checked multiple times
- Simulated time evolution (Brad's project for winter break)