

Phylogeny

- PHYLOGENY, coined by Haeckel (1866)
 1. the line of descent or evolutionary development of any plant or animal species
 2. the origin and evolution of a division, group or race of animals or plants

phylogeny

Bioinformatics Fall 2011

1

Goals

- Understand evolutionary history
 - African eve example
 - Analysis of mitochondrial DNA from 182 individuals
 - Origin of europeans
- Assist in epidemiology
 - of infectious diseases
 - of genetic defects
- Aid in prediction of function of novel genes

phylogeny

Bioinformatics Fall 2011

2

Mitochondria and phylogeny

- **Mitochondrial DNA (mtDNA):** Extra-nuclear DNA, transmitted through maternal lineage.
 - Allows tracing of a single genetic line
 - No recombinations
- 16.5 Kb circular DNA contains genes: coding for 13 proteins, 22 tRNA genes, 2 rRNA genes.
- mtDNA has a mutation rate 10 times faster than nuclear DNA: provides a way to infer relationships between closely related individuals
- Phylogeny based on human mtDNA can give us molecular (hence accurate?) information about human evolution.

phylogeny

Bioinformatics Fall 2011

3

African eve

- Statistical analysis of mtDNA extracted from placental tissue of 182 women of different races and regions. (Cann, Stoneking, & Wilson, 87).
- Phylogenetic tree (assuming a constant molecular clock). Root closest to the modern African woman. ~200K years ago
 - Greatest diversity around Africa
 - Maximum time span in Africa
- **Conclusion:** Modern man emerged from Africa 200,000 years ago. Race differences arose 50,000 years ago: "Mitochondrial Eve Hypothesis"
- Also supported by analysis of ZFY gene on the Y-chromosome.
- Not the first human; other women would have co-existed (only maternal lineage).

phylogeny

Bioinformatics Fall 2011

4

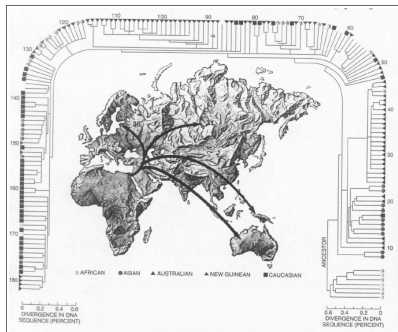


Figure 8.4: (Source: [24]) African origin for all modern human is indicated by the genetic evidence. The arrows on the maps (center) indicate the minimum number of unislated females who colonized major geographic areas, as inferred from the branching pattern.

phylogeny

Bioinformatics Fall 2011

5

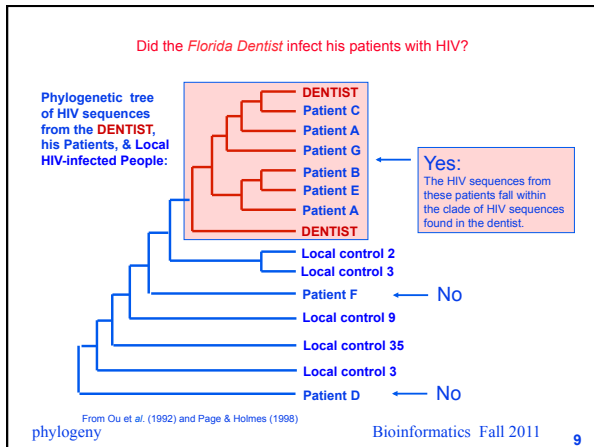
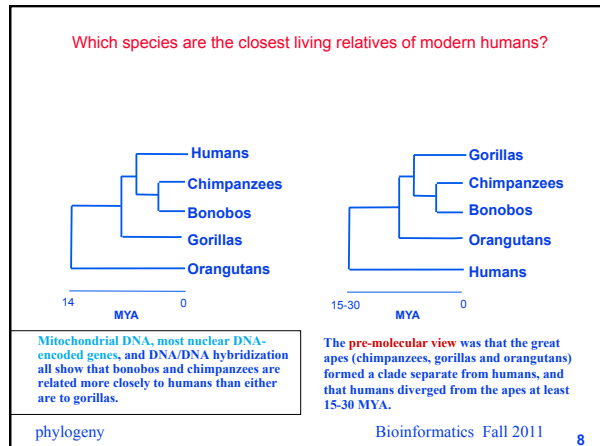
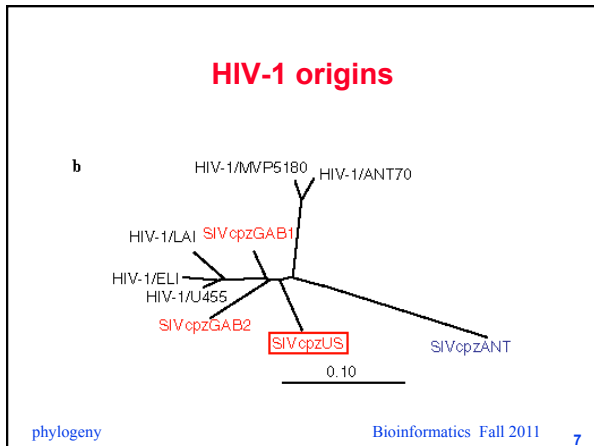
Mitochondrial eve's Africanness

- A simple reordering of the data could result in 100 distinct trees at most 2 steps away---all supporting non-African hypothesis. (Templeton)
- Not completely consistent with sparse fossil evidence

phylogeny

Bioinformatics Fall 2011

6



Building phylogenies: Phenotype information has problems

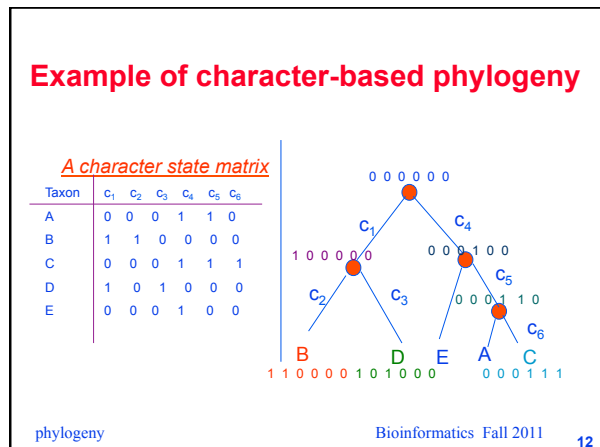
- Can be difficult to observe
 - Bacteria
- Difficult to compare diverse species
 - Plants, bacteria, animals
- Convergent evolution

phylogeny Bioinformatics Fall 2011 10

Data for building phylogenies

- Characteristics
 - Traits (continuous or discrete)
 - DNA/Protein sequence
 - *character state matrix*
- Numerical distance estimates
 - *distance matrix*

phylogeny Bioinformatics Fall 2011 11



Different kinds of trees

- Order of evolution
 - Rooted: indicates direction of evolution
 - Unrooted: only reflects the distance
- Rate of evolution
 - Edge lengths: distance (scaled trees)
 - Time?
 - Unscaled trees

phylogeny

Bioinformatics Fall 2011

13

Rooted and unrooted trees

- Most phylogenetic methods produce unrooted trees. This is because they detect differences between sequences, but have no means to orient residue changes relatively to time.
- Two means to root an unrooted tree :
 - The outgroup method: include in the analysis a group of sequences known *a priori* to be external to the group under study; the root is by necessity on the branch joining the outgroup to other sequences.
 - Make the molecular clock hypothesis: all lineages are supposed to have evolved with the same speed since divergence from their common ancestor. Root the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. The root is at the equidistant point from all tree leaves.

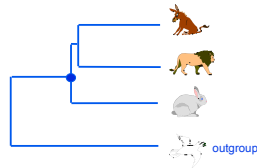
phylogeny

Bioinformatics Fall 2011

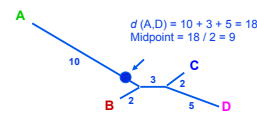
14

Rooting unrooted trees

By outgroup:



By midpoint or distance:

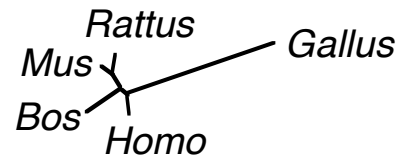


phylogeny

Bioinformatics Fall 2011

15

Unrooted tree



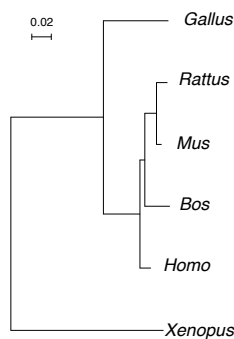
0.02
|

phylogeny

Bioinformatics Fall 2011

16

Rooted tree



phylogeny

Bioinformatics Fall 2011

17

Tree building methods

- Character-based methods
 - Maximum parsimony
 - Maximum likelihood
- Distance-based methods
 - UPGMA
 - NJ

phylogeny

Bioinformatics Fall 2011

18

Character-based models

- Given m characters, n species, build the best tree that maximizes some target function
- Assumptions
 - Independence of characters
- Number of non-isomorphic, binary, unrooted trees with n leaves is $1 \cdot 3 \cdot \dots \cdot (2n-5)$, $n \geq 3$ (1 for $n=2$)
 - Show by induction
 - Number of edges in an unrooted tree with n leaves = $2n-3$
 - Can extend any edge
 - Rooted trees?
- Approx. 2 million unrooted trees for 10 leaves
 - Need better ways of exploring the search space

phylogeny

Bioinformatics Fall 2011

19

Parsimony

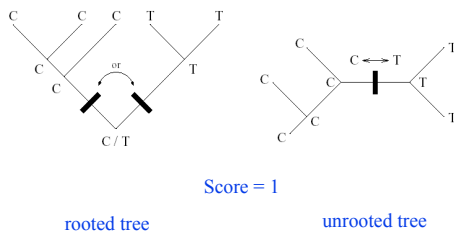
- Minimize Hamming distance summed over all edges of the tree
- Justification: minimum possible number of evolutionary (rare) events
- Weighted parsimony
- Insensitive to branch lengths

phylogeny

Bioinformatics Fall 2011

20

Example



phylogeny

Bioinformatics Fall 2011

21

Another example

- Trees that could explain the phylogeny of the following sequences: AAG, AAA, GGA, AGA



- Parsimony prefers the second tree because it requires the fewest substitution events

phylogeny

Bioinformatics Fall 2011

22

Parsimony based approaches

These approaches involve two separate components:

- A search through the space of trees (Large parsimony problem)
- A procedure to find the minimum number of changes needed to explain the data for a given tree topology (Small parsimony problem)

phylogeny

Bioinformatics Fall 2011

23

Fitch's Algorithm [1971] for Small Parsimony

1. Traverse tree from leaves to root determining set of possible *states* (e.g. nucleotides) for each internal node
2. traverse tree from root to leaves picking ancestral states for internal nodes

phylogeny

Bioinformatics Fall 2011

24

Fitch's Algorithm – Step 1

- Perform a post-order (from leaves to root) traversal of tree
- Determine possible states R_i of internal node i with children j and k

$$R_i = \begin{cases} R_j \cup R_k & \text{if } R_j \cap R_k = \phi \\ R_j \cap R_k & \text{otherwise} \end{cases}$$

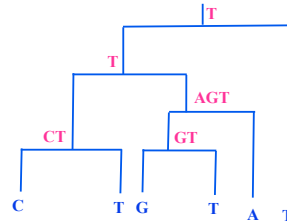
phylogeny

Bioinformatics Fall 2011

25

Fitch's Algorithm – Step 1

of changes = # union operations



phylogeny

Bioinformatics Fall 2011

26

Fitch's Algorithm – Step 2

- Perform a pre-order (from root to leaves) traversal
- Select state r_j of internal node j with parent i

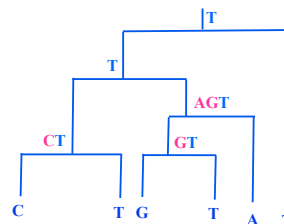
$$r_j = \begin{cases} r_i & \text{if } r_i \in R_j \\ \text{arbitrary state} \in R_j & \text{otherwise} \end{cases}$$

phylogeny

Bioinformatics Fall 2011

27

Fitch's Algorithm – Step 2



phylogeny

Bioinformatics Fall 2011

28

Complexity

- $O(mnk)$
 - m characters, n leaves, k possible values for a character
- Weighted parsimony
 - Assign weights to changes [Sankoff 75]

phylogeny

Bioinformatics Fall 2011

29

Large Parsimony Problem

- NP-hard under many different conditions
- Informative and uninformative sites
 - Example: consider sequences GGGGGG, GGGAGT, GGATAG, GATCAT
 - Which sites are informative?
- Speeding up the search
 - Branch & bound
 - Hill climbing methods

phylogeny

Bioinformatics Fall 2011

30

Maximum Likelihood

- Hypotheses
 - The substitution process follows a probabilistic model whose mathematical expression, but not parameter values, is known *a priori*.
 - Sites evolve independently from each other.
 - All sites follow the same substitution process.
 - Substitution probabilities do not change with time on any tree branch. They may vary between branches.

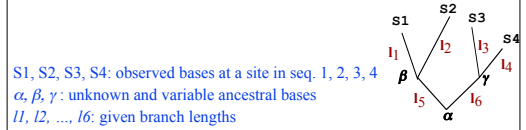
phylogeny

Bioinformatics Fall 2011

31

Maximum Likelihood

- Step 1: Let us consider a given rooted tree, a given site, and a given set of branch lengths. Let us compute the probability that the observed pattern of nucleotides at that site has evolved along this tree.



$$P(S1, S2, S3, S4 | \text{Tree}, l_1, l_2, l_3, l_4, l_5, l_6) = \sum_{\alpha} \sum_{\beta} \sum_{\gamma} P(\alpha) P_{l_5}(\alpha, \beta) P_{l_6}(\alpha, \gamma) P_{l_1}(\beta, S1) P_{l_2}(\beta, S2) P_{l_3}(\gamma, S3) P_{l_4}(\gamma, S4)$$

phylogeny

Bioinformatics Fall 2011

32

Substitution rates

- DNA sequences
 - Jukes and Cantor model
 - Kimura model
- Protein sequences
 - PAM and BLOSUM matrices

phylogeny

Bioinformatics Fall 2011

33

Maximum Likelihood

- Step 2: compute the probability that entire sequences have evolved along the tree:

$$P(Sq1, Sq2, Sq3, Sq4 | \text{Tree}, l_1, l_2, l_3, l_4, l_5, l_6) =$$

$$\prod_{\text{all sites}} P(S1, S2, S3, S4 | \text{Tree}, l_1, l_2, l_3, l_4, l_5, l_6)$$

- Step 3: compute branch lengths l_1, l_2, \dots, l_6 that give the highest value for the above probability. This is the *likelihood* of the tree.
- Step 4: compute the likelihood of all possible trees. The tree predicted by the method is that having the highest likelihood.

phylogeny

Bioinformatics Fall 2011

34

Maximum Likelihood: properties

- This is the best justified method from a theoretical viewpoint.
- Sequence simulation experiments have shown that this method works better than all others in most cases.
- Position of root does not matter for *reversible* and *multiplicative* substitution matrices.
 - Detailed balance
 - $S(t_1)S(t_2) = S(t_1+t_2)$, S is the substitution matrix
- But it is a very compute-intensive method. Impossible to evaluate all possible trees.
 - MCMC to explore the tree space

phylogeny

Bioinformatics Fall 2011

35

Distance Matrix Methods

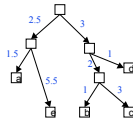
- Given a pairwise distance matrix D
- Produce a tree such that the *tree distance* between leaves i and j (sum of edge weights in the path between i and j) equals d_{ij}
- Optimize the error between d and D
 - Least square error metric: LSQ
 - $LSQ(d, D) = \sum \sum (d_{ij} - D_{ij})^2$
 - NP-complete
- Heuristics (usually based on agglomerative clustering)
 - UPGMA
 - NJ
 - Assume distance d is a metric
 - symmetry
 - triangle inequality
 - $d(x, y) = 0$ iff $x = y$
 - $d(x, y) \geq 0$

phylogeny

Bioinformatics Fall 2011

36

Additive Matrix



	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

There exists a tree with zero error

phylogeny

Bioinformatics Fall 2011

37

Four point condition

An additive matrix is characterized by the **four point condition**:
For any 4 points x,y,u,v, one of the following 3 conditions hold:

$$d(x,y) + d(u,v) \leq d(x,u) + d(y,v) = d(x,v) + d(y,u) \text{ or ... or ...}$$

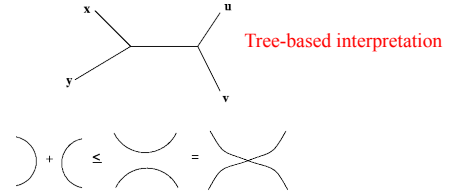


Figure: The four point condition

phylogeny

Bioinformatics Fall 2011

38

Trees from Additive Matrices

- Verify that the distance matrix is additive
- Choose a pair of objects, which results in the first path in the tree.
- Choose a pair of leaves in the tree constructed so far and compute the point at which a newly chosen object is inserted.
 1. The new path branches off an existing node in the tree: Do the insertion step once more in the branching path.
 2. The new path branches off an edge in the tree: This insertion is finished.

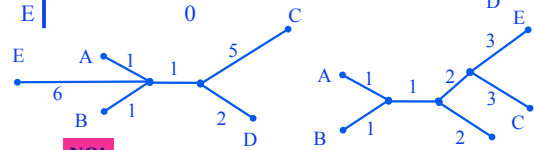
phylogeny

Bioinformatics Fall 2011

39

Example

	A	B	C	D	E
A	0	2	7	4	7
B		0	7	4	7
C			0	7	6
D				0	7
E					0



NO!

phylogeny

Bioinformatics Fall 2011

40

Approximating Additive Matrices

In practice, the distance matrix between molecular sequences will not be additive.

An additive tree T whose distance matrix approximates the given one is used.

The methods for exact tree reconstruction provide an inventory for heuristics for tree construction based on approximating additive metrics.

Heuristics give exact results when operating on additive metrics.

phylogeny

Bioinformatics Fall 2011

41

UPGMA

- Unweighted Pair-Group Method with Arithmetic Mean
 - Sokal and Michener 1958
- Agglomerative clustering
- Ultrametric tree
 - distances from root to all leaves are equal
- Cluster distances defined as

$$d_{AB} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d_{ab}$$

phylogeny

Bioinformatics Fall 2011

42

UPGMA Step 1 combine B and C

Choose two clusters with minimum distance and combine them



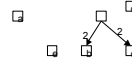
	A	B	C	D	E
A	0	10	12	8	7
B		0	4	4	14
C			0	6	16
D				0	12
E					0

phylogeny

Bioinformatics Fall 2011

43

Updating distance matrices



	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

Distance of new cluster to nodes in original clusters is half of original distance

Distance of new cluster to other clusters is weighted mean of individual distances

phylogeny

Bioinformatics Fall 2011

44

UPGMA step 2 combine BC and D



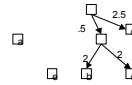
	A	BC	D	E
A	0	11	8	7
BC		0	5	15
D			0	12
E				0

phylogeny

Bioinformatics Fall 2011

45

Updating distance matrices



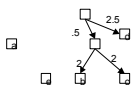
	A	BCD	E
A	0	10	7
BCD		0	14
E			0

phylogeny

Bioinformatics Fall 2011

46

UPGMA step 3 combine A and E



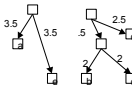
	A	BCD	E
A	0	10	7
BCD		0	14
E			0

phylogeny

Bioinformatics Fall 2011

47

Updating distance matrices



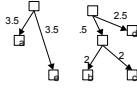
	AE	BCD
AE	0	12
BCD		0

phylogeny

Bioinformatics Fall 2011

48

UPGMA step 4 combine AE and BCD



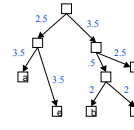
	AE	BCD
AE	0	12
BCD		0

phylogeny

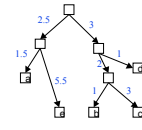
Bioinformatics Fall 2011

49

UPGMA Result



produced tree



actual tree

phylogeny

Bioinformatics Fall 2011

50

Limitations of UPGMA

- Ultrametric tree
 - Distance from the root to each leaf is the same
 - Molecular clock assumption - distance is proportional to evolutionary time
 - Falling out of favor - “mounting evidence that different lineages can evolve at different rates”
- Ultrametric distance
 - Usual metric conditions Why does this imply additive distance matrix?
 - $d(x,y) \leq \max[d(x,z), d(y,z)]$
 - 2 largest distances in any group of 3 are equal
 - meaning in a tree setting?
- UPGMA works correctly for ultrametric distances

phylogeny

Bioinformatics Fall 2011

51

Neighbor Joining (NJ)

- Saitou and Nei, 1987
 - Join clusters that are close to each other and also far from the rest
- Produces unrooted tree
- NJ is a fast method, even for hundreds of sequences.
- NJ always finds the correct tree if distances are additive (tree-like).
- Also performs well when substitution rates vary among lineages.
- Quality close to Maximum Likelihood at much less the cost.

phylogeny

Bioinformatics Fall 2011

52

Algorithm

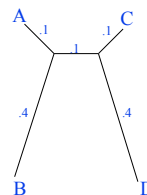
- Define $u_i = \sum_k D_{ik} / (n-2)$
 - measure of average distance from other nodes
- Iterate until 2 nodes are left
 - choose pair (i,j) with smallest $D_{ij} - u_i - u_j$
 - close to each other and far from others
 - merge to a new node (ij) and update distance matrix
 - $D_{k,(ij)} = (D_{ik} + D_{jk} - D_{ij})/2$ -- consider the tree paths
 - $D_{i,(ij)} = (D_{ij} + u_i - u_j)/2$ -- average over all leaves m of $(D_{ij} + D_{im} - D_{jm})/2$
 - $D_{j,(ij)} = D_{ij} - D_{i,(ij)}$ -- similarly
 - delete nodes i and j
- For the final group (i,j), use D_{ij} as the edge weight.

phylogeny

Bioinformatics Fall 2011

53

NJ picks out neighboring leaves



- What would NJ do?
- How about UPGMA?

NJ will find the correct tree since the distances are additive

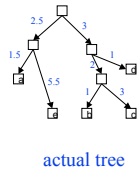
UPGMA will incorrectly Group A and C

phylogeny

Bioinformatics Fall 2011

54

Comparison of results



produced tree

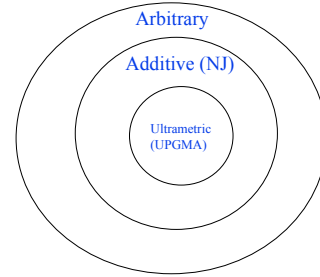
actual tree

phylogeny

Bioinformatics Fall 2011

55

Constraints on trees



phylogeny

Bioinformatics Fall 2011

56

Limitations of distance-based methods

- Loss of information
- Difficult to evaluate the quality of a tree or judge among competing hypotheses

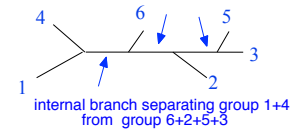
phylogeny

Bioinformatics Fall 2011

57

Reliability of phylogenetic trees: bootstrapping

- The phylogenetic information expressed by an unrooted tree resides entirely in its internal branches.



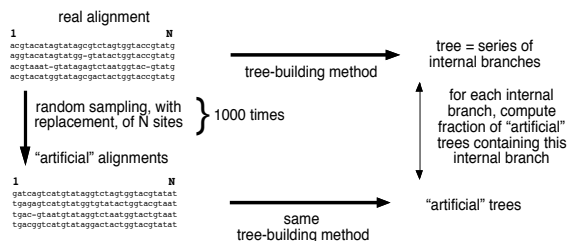
- The tree shape can be deduced from the list of its internal branches.
- Testing the reliability of a tree = testing the reliability of each internal branch.

phylogeny

Bioinformatics Fall 2011

58

Bootstrap procedure



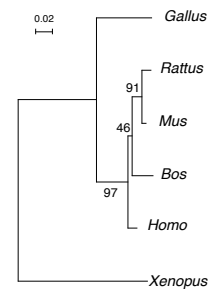
The support of each internal branch is expressed as percent of replicates.

phylogeny

Bioinformatics Fall 2011

59

"bootstrapped" tree



phylogeny

Bioinformatics Fall 2011

60

Bootstrap procedure : properties

- Internal branches supported by $\geq 90\%$ of replicates are considered as statistically significant.
- The bootstrap procedure only detects if sequence length is enough to support a particular branch.
- The bootstrap procedure does not help in determining if the tree-building method is good. A wrong tree can have 100 % bootstrap support for all its branches!

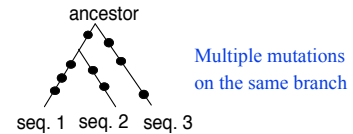
phylogeny

Bioinformatics Fall 2011

61

Saturation: loss of phylogenetic signal

- Closely related taxa: choose highly varying features
- Distantly related taxa: choose slowly varying features
- When compared homologous sequences have experienced too many residue substitutions since divergence, it is impossible to determine the phylogenetic tree, whatever the tree-building method used.



phylogeny

Bioinformatics Fall 2011

62

WWW Resources

- ⇒ PHYLIP : an extensive package of programs for all platforms <http://evolution.genetics.washington.edu/phylip.html>
- ⇒ CLUSTALX : beyond alignment, it also performs NJ
- ⇒ PAUP* : commercial package <http://paup.csit.fsu.edu/index.html>
- ⇒ PHYLO_WIN : a graphical interface, for unix only <http://pbil.univ-lyon1.fr/software/phylowin.html>
- ⇒ MrBayes : Bayesian phylogenetic analysis <http://morphbank.ebc.uu.se/mrbayes/>
- ⇒ PHYML : fast maximum likelihood tree building <http://www.lirmm.fr/~guindon/phyml.html>
- ⇒ WWW-interface at Institut Pasteur, Paris <http://bioweb.pasteur.fr/seqanal/phylogeny>
- ⇒ Tree drawing NJPLOT (for all platforms) <http://pbil.univ-lyon1.fr/software/njplot.html>
- ⇒ [Lecture notes of molecular systematics](#)

phylogeny

Bioinformatics Fall 2011

63