BIOINFORMATICS

or why biologists need computers

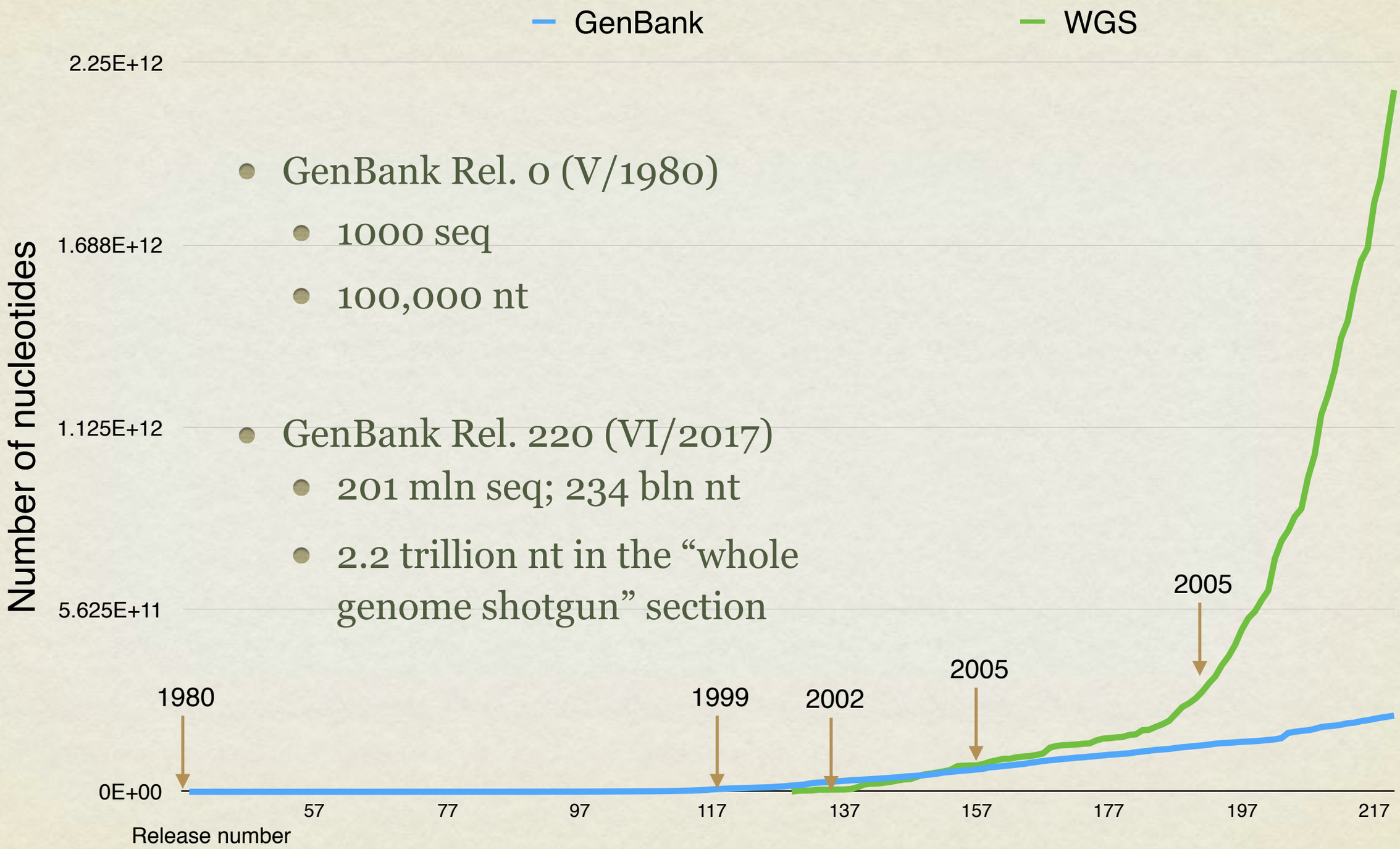
Wojciech Makałowski
Institute of Bioinformatics
http://bioinformatics.uni-muenster.de

It's sink or swim as a tidal wave of data approaches

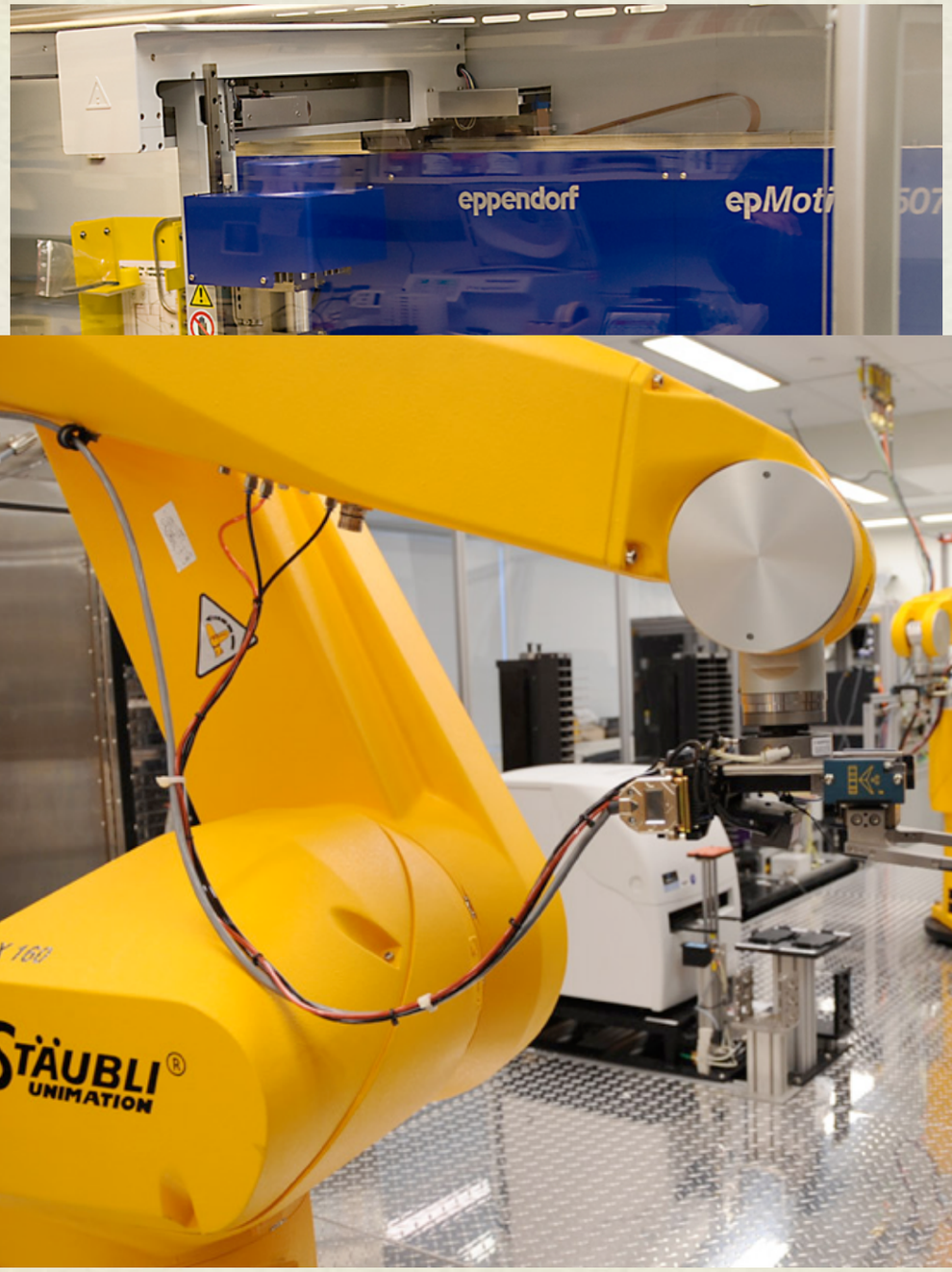
Nature 399:517 10 June 1999

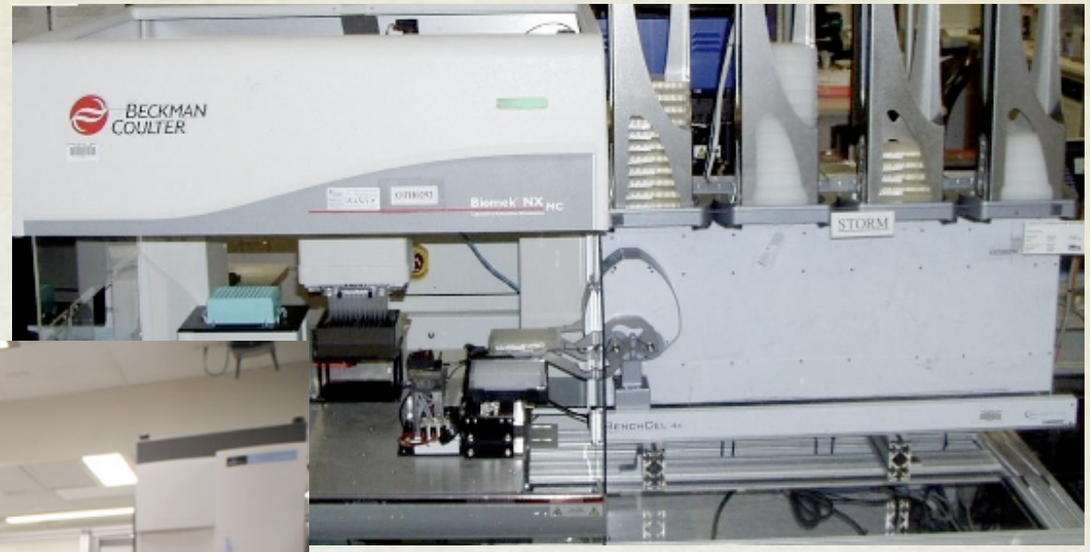
GROWTH OF BIOMEDICAL INFORMATION - GENBANK

- GenBank Rel. 0 (V/1980)
  - 1000 seq
  - 100,000 nt
- GenBank Rel. 220 (VI/2017)
  - 201 mln seq; 234 bln nt
  - 2.2 trillion nt in the “whole genome shotgun” section

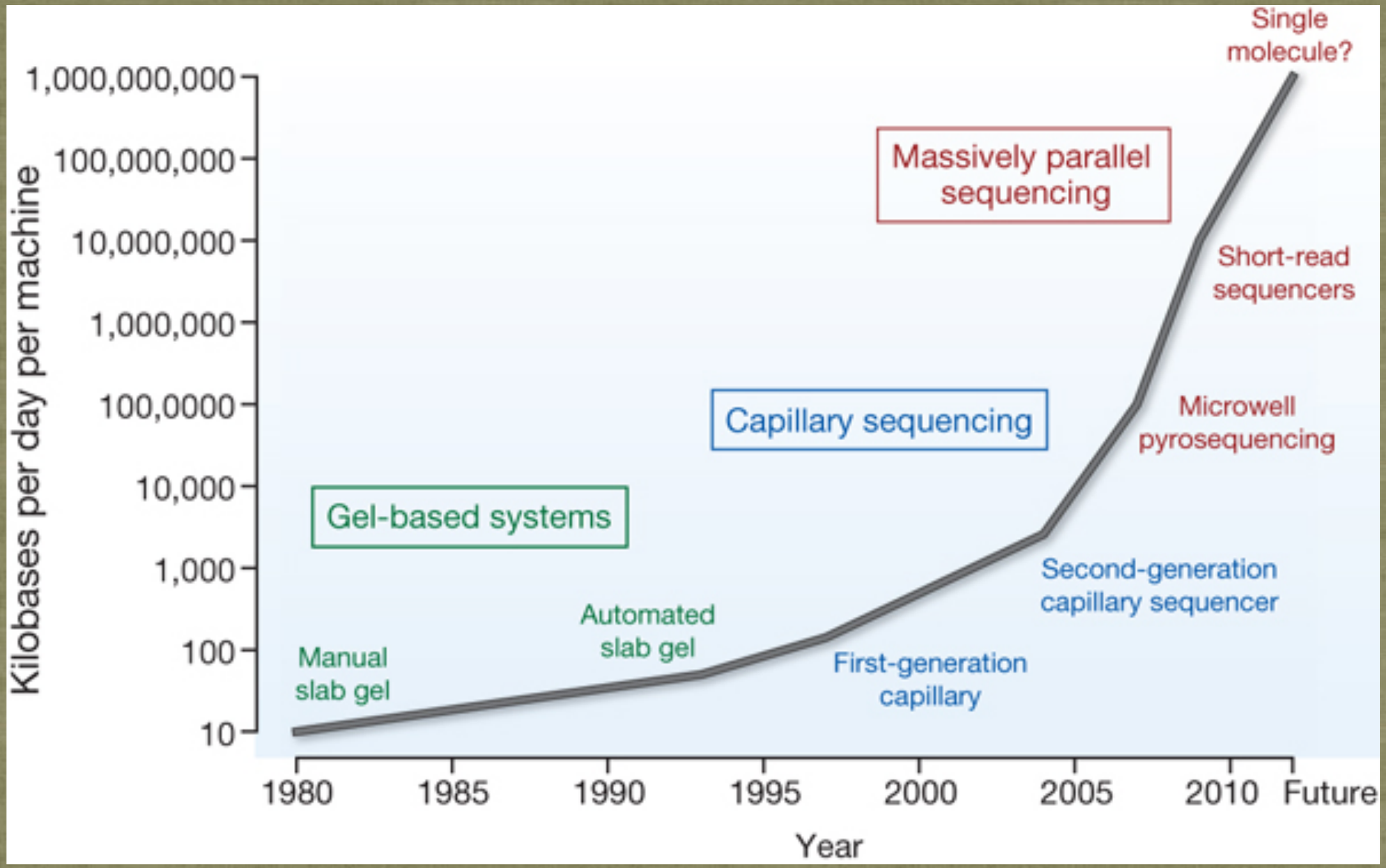


TECHNOLOGY MEETS BIOLOGY





IMPROVING TECHNOLOGY

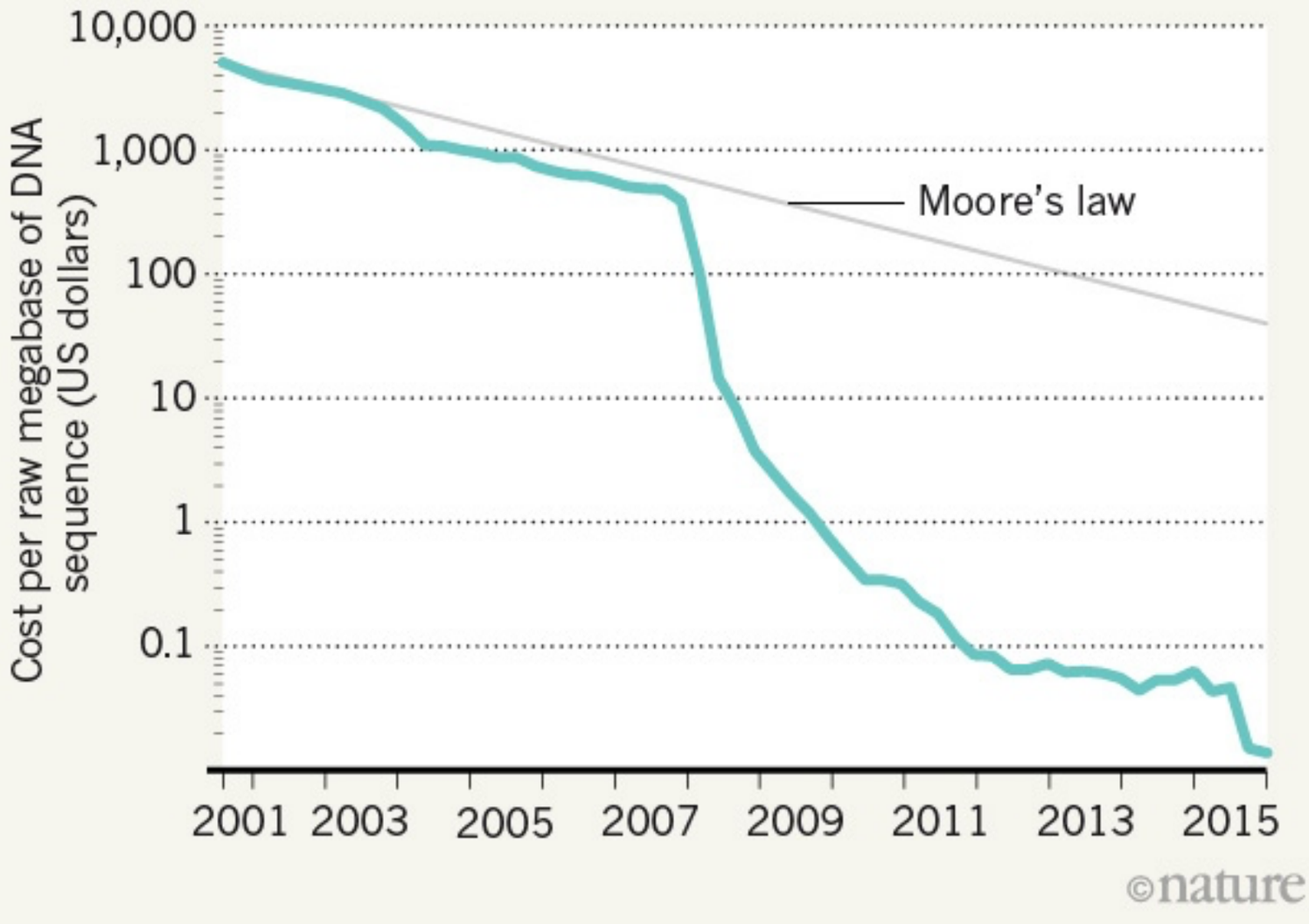


MR Stratton *et al. Nature* **458**, 719-724 (2009)

IMPROVING TECHNOLOGY

PLUNGING COSTS OF SEQUENCING

Since 2008, new sequencing technologies have driven the costs of DNA sequencing down faster than the rapid improvement in microprocessor power represented by Moore's Law.

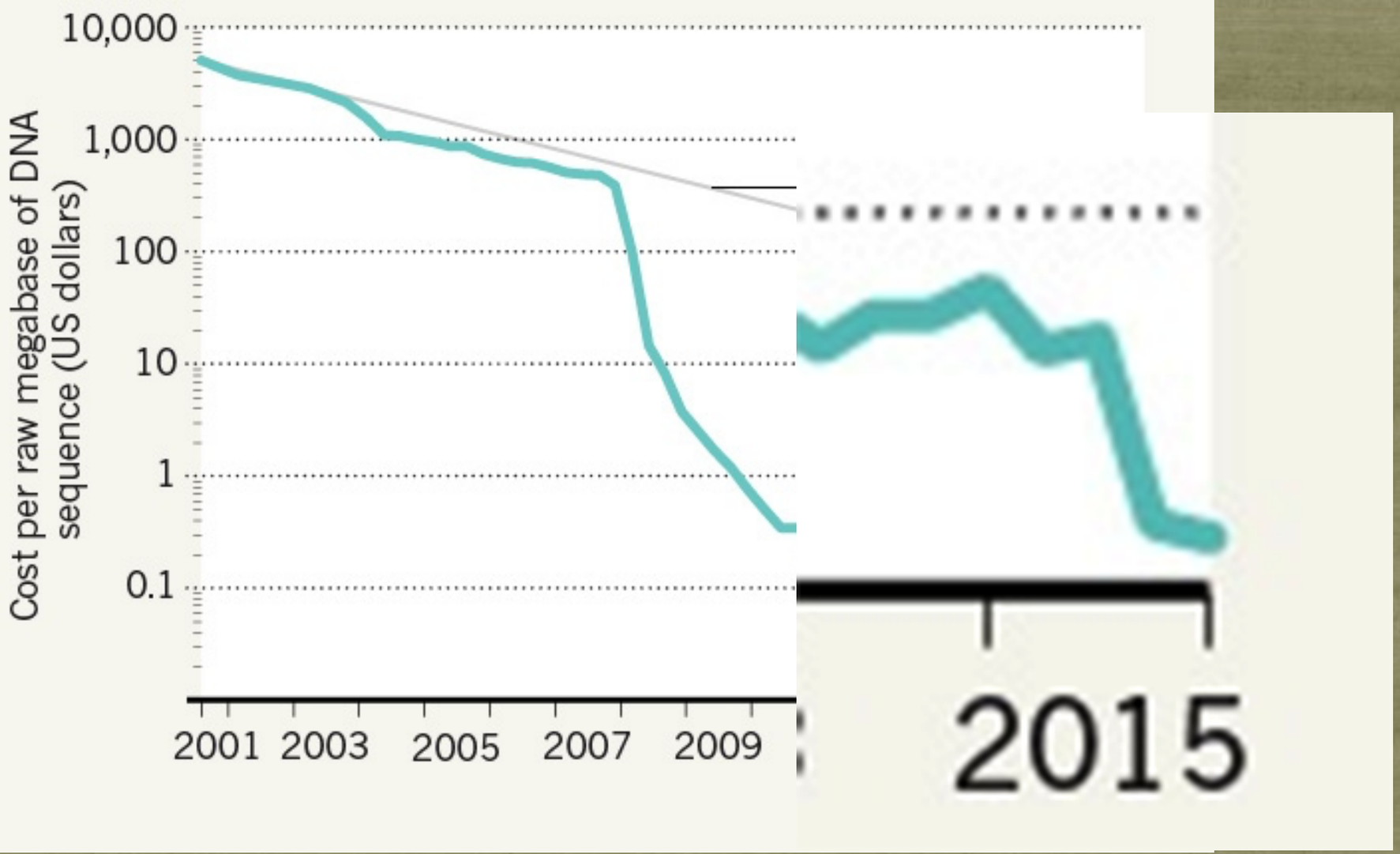


http://ark-invest.com/genomic-revolution/declining-costs-of-genome-sequencing

IMPROVING TECHNOLOGY

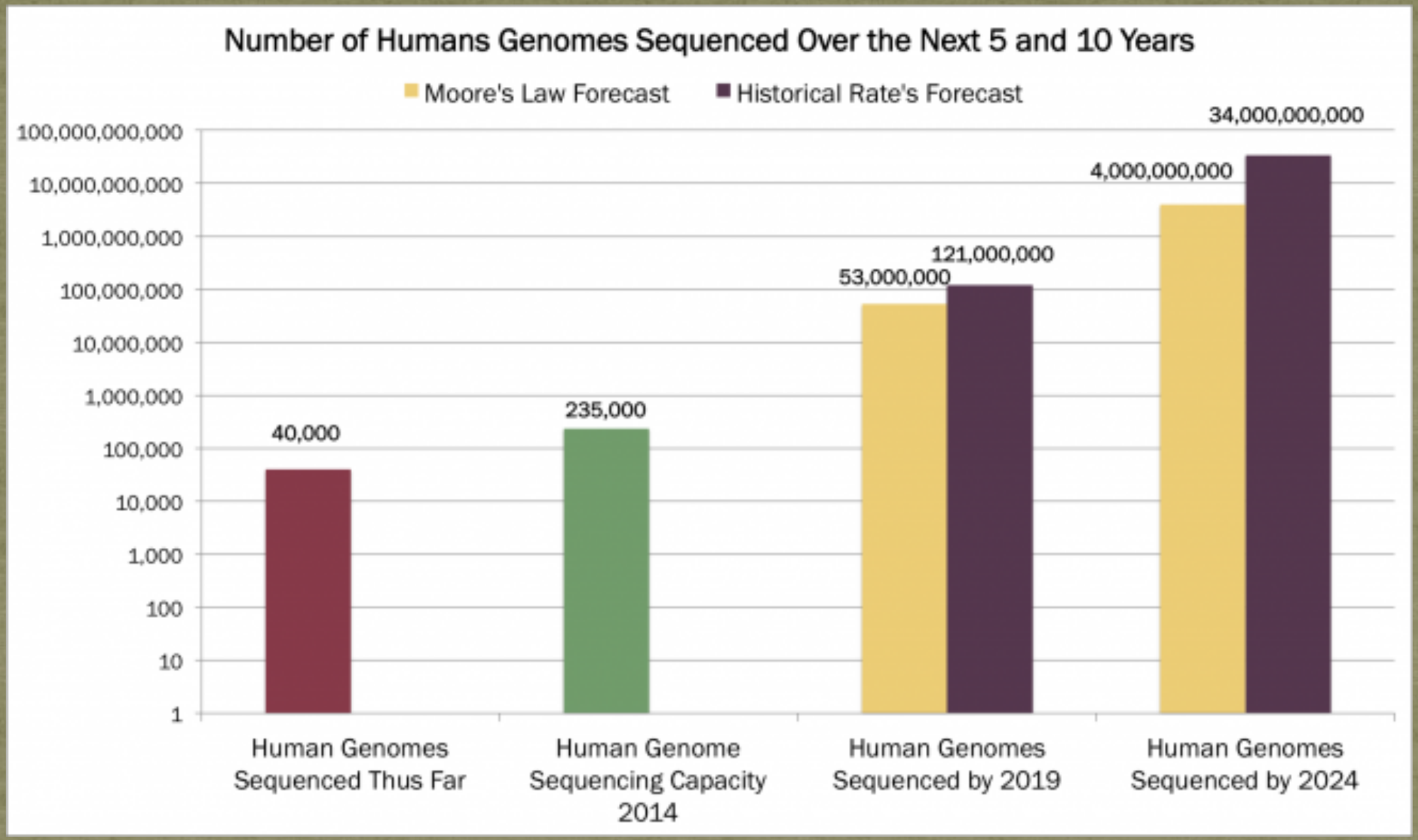
PLUNGING COSTS OF SEQUENCING

Since 2008, new sequencing technologies have driven the costs of DNA sequencing down faster than the rapid improvement in microprocessor power represented by Moore's Law.



http://ark-invest.com/genomic-revolution/declining-costs-of-genome-sequencing

IMPROVING TECHNOLOGY



http://ark-invest.com/genomic-revolution/declining-costs-of-genome-sequencing

GETTING SEQUENCES

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

READING ≠ UNDERSTANDING

Carmina qui quondam studio
florente peregi, flebilis heu maestos
cogor inire modos.

Ecce mihi lacerae dictant scribenda
Camenae et ueris elegi fletibus ora
rigant.

Boethius, *Consolatio Philosophiae*

READING ≠ UNDERSTANDING

We shall best understand the probable course of natural selection by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

READING ≠ UNDERSTANDING

We shall best understand the probable course of [redacted] by taking the case of a country undergoing some physical change. If the country were open were open on its borders, new forms would certainly immigrate, and this also would bla, bla bla become extinct inhabitants.

Charles Darwin - *The Origin of Species*

CHALLENGE: HOW FROM THIS...

TGCATCGATCGTAGCTAGCTAGCGCATGCTAGCTAGCTAGCTAGCTACGATGCATCG
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
CGCGCGCATTATGCCGCGGCATGCTGCGCACACACAGTACTATAGCATTAGTAAAAA
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT
AGCTAGTGTAGCTAGCTAGCATGCTGCTAGCATGCAGCATGCATCGGGCGCGATGCT
GCTAGCGCTGCTAGCTAGCTAGCTAGCTAGGCGCTAATTATTTATTTTGGGGGGTTA
AAAAAAAAAATTTCGCTGCTTATACCCCCCCCCACATGATGATCGTTAGTAGCTACT
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
AAAGACCCCATCTCTCTCTCTTTTCCCTTCTCTCGCTAGCGGGCGGTACGATTTACC
GGCCGCGTATATTTTACACGATAGTGCGGCGCGGCGCGTAGCTAGTGCTAGCTAGTC
AGCTCTCATCGCGCGGGGGGATGCTTAGCGTGGTGTGTGTGTGTGGTGTGTGTGGTC
TGCATCGATCGATGCATGCTAGCTAGCTAGCTAGCATGCTAGCTAGCTAGCTATTGG
CTATAATTAGTGCATCGGCGCATCGATGGCTAGTCGATCGATCGATTTTATATATCT
CGCTAGCTAGCATGCATGCATGCATCGATGCATCGATTATAAGCGCGATGACGTCAG
TCCGGTTACACAGGTAGCTAGCTAGCTGCTAGCTAGCTGCTGCATGCATGCATTAGT

Infer this

HOW TO SOLVE THE PROBLEM - A HUMAN OR A COMPUTER?

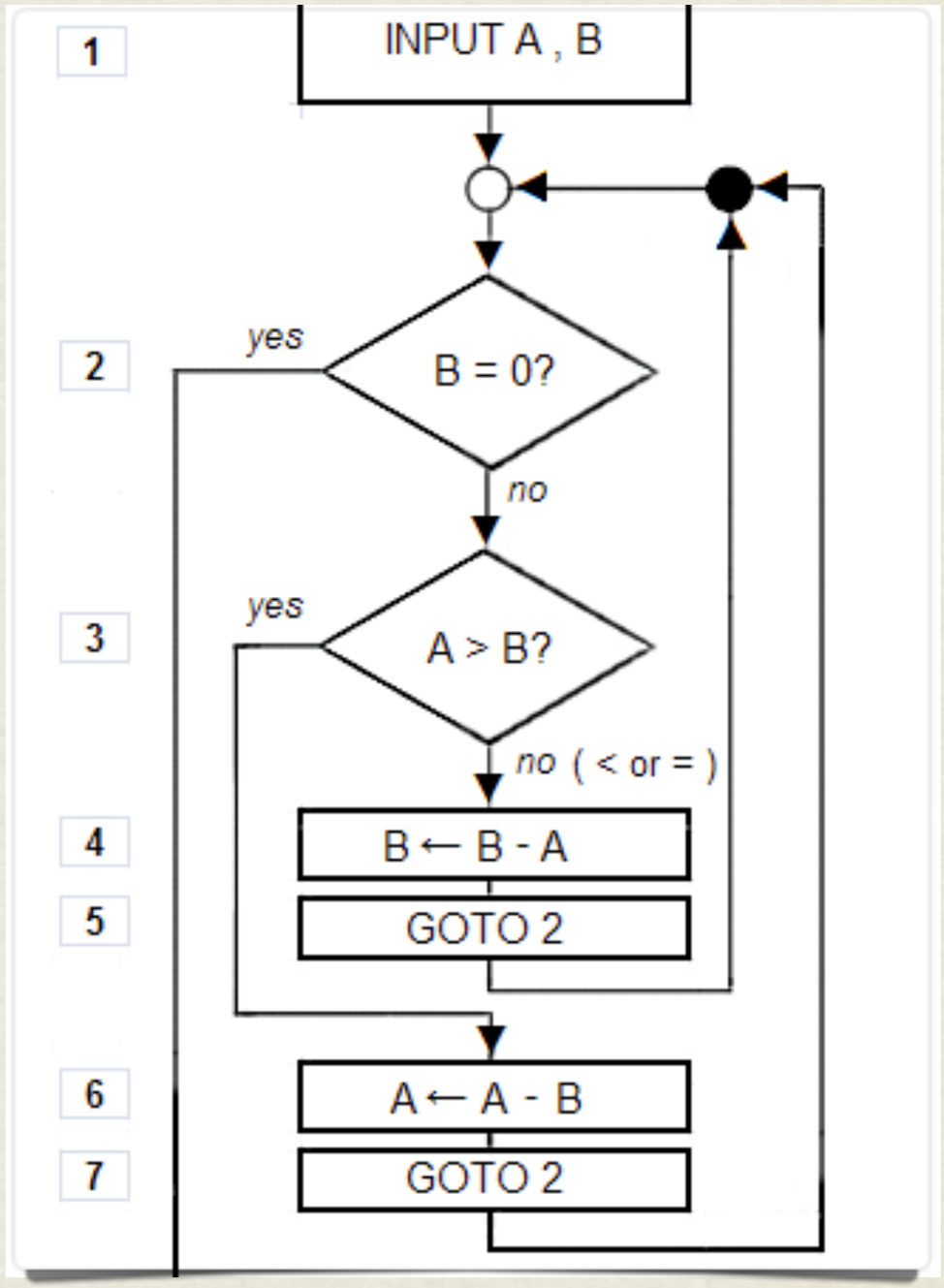
- very smart
- slow
- error prone
- doesn't like repetitive tasks

- not so smart (stupid)
- extremely fast
- very accurate
- doesn't understand human languages;

needs instruction provided in a special way

ALGORITHM

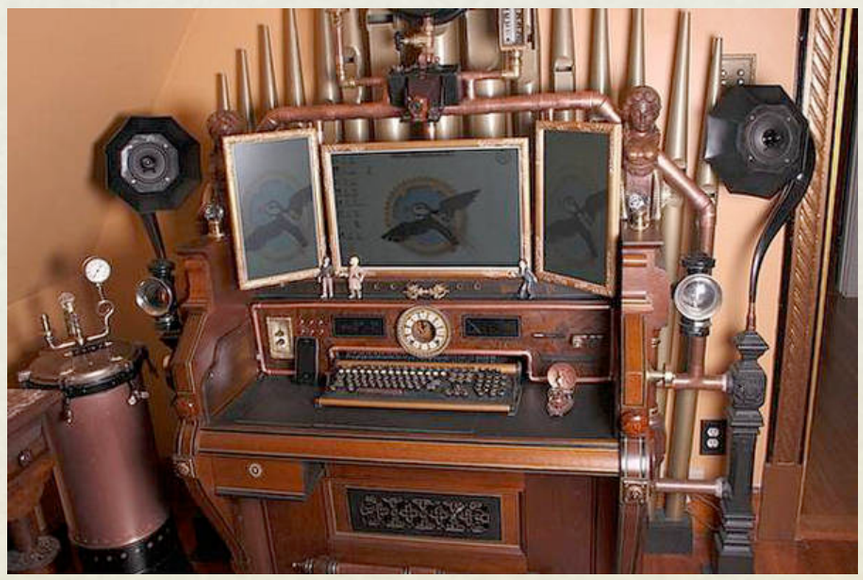
A step-by-step problem-solving procedure, especially an established, recursive computational procedure for solving a problem in a finite number of steps.



EXAMPLE TASK: PUT SHOES ON!

A human just understands an order and often executes it automatically even without thinking

A computer needs detailed instruction (an algorithm)



PUT SHOES ON!
INSTRUCTION FOR A COMPUTER

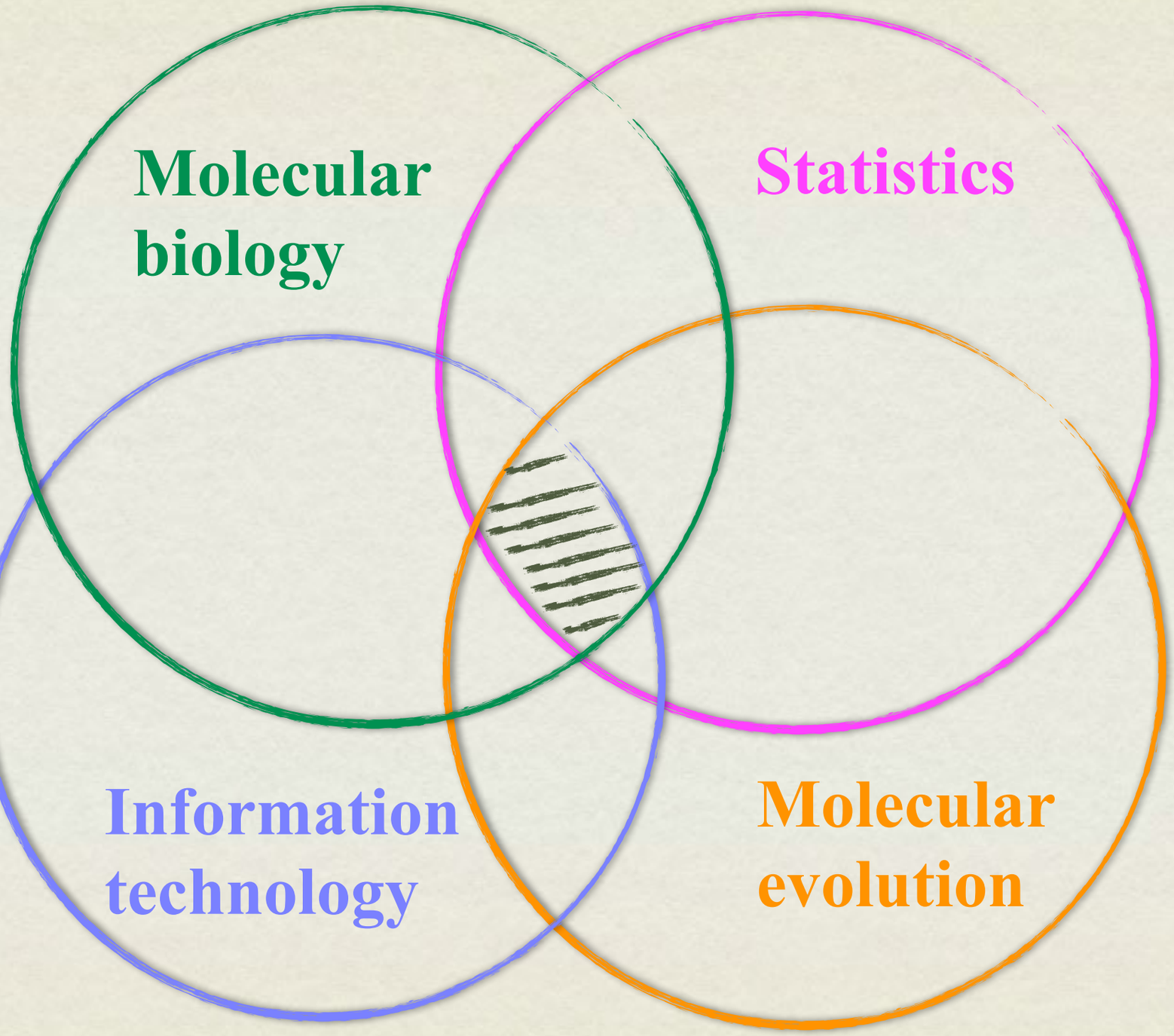
1. Find two the same shoes
2. Check if you have left and right shoe
3. Check if they are of the same size
4. Check if this is the right size
5. Put the left shoe on
6. Put the right shoe on
7. Tie the laces

THE ORIGIN OF THE FIELD

Paulien Hogeweg coined the term *bioinformatica* to define “the study of informatic processes in biotic systems”. Hesper B, Hogeweg P (1970) Bioinformatica: een werkconcept. Kameleon 1(6): 28–29. (In Dutch.) Leiden: Leidse Biologen Club.

… but its origin can be tracked back many decades earlier.

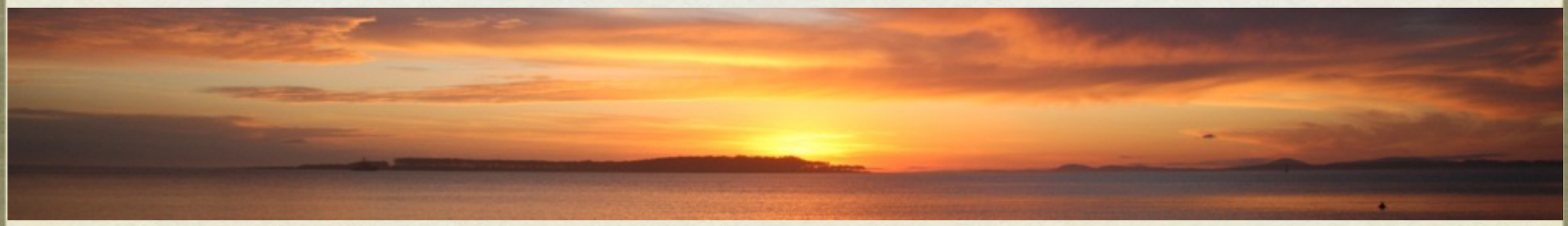
BIOINFORMATICS EMERGED AS AN INTERSECTION BETWEEN DIFFERENT DISCIPLINES



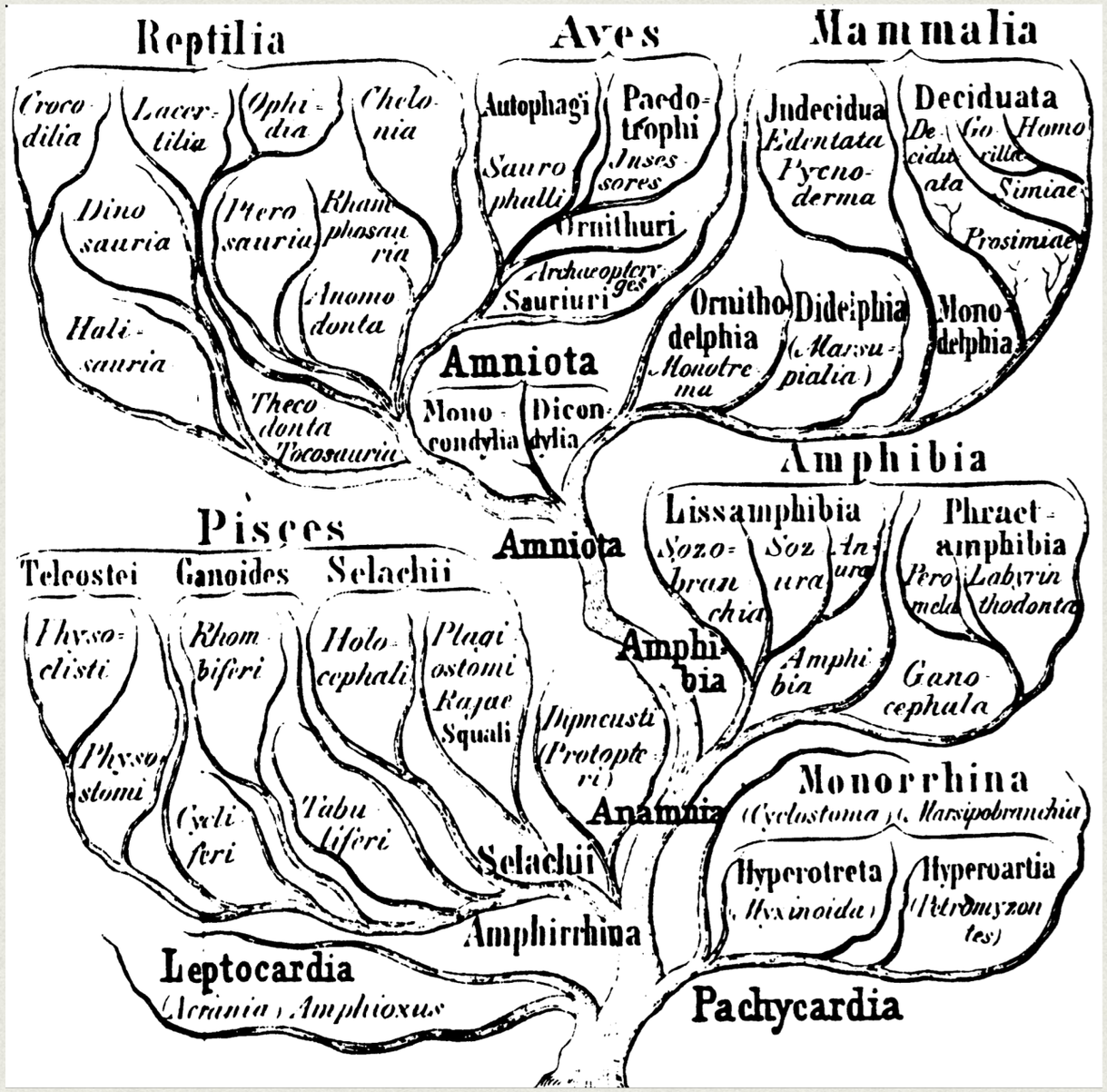
BIOINFORMATICS - DEFINITION

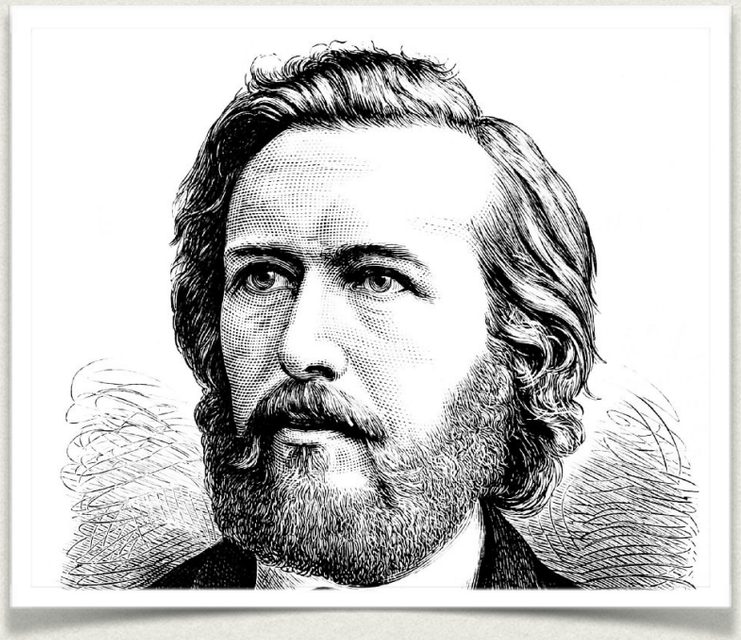
- Research, development, or application of computational tools and approaches for expanding the use of biological data, including those to acquire, store, organize, archive, analyze, or visualize such data.

- Its goal is to enable biological discovery based on existing information or in other words transform biological data into information and eventually into knowledge.



PHYLOGENETIC ANALYSIS





Haeckel (1866)
Generelle Morphologie der Organismen

ROLE OF BIOINFORMATICS IN MODERN BIOLOGY

- molecular biology
- molecular evolution
- genomics
- system biology
- protein engineering
- drug design
- personalized medicine
- biogeography

WHAT IS PHYLOGENETIC ANALYSIS?

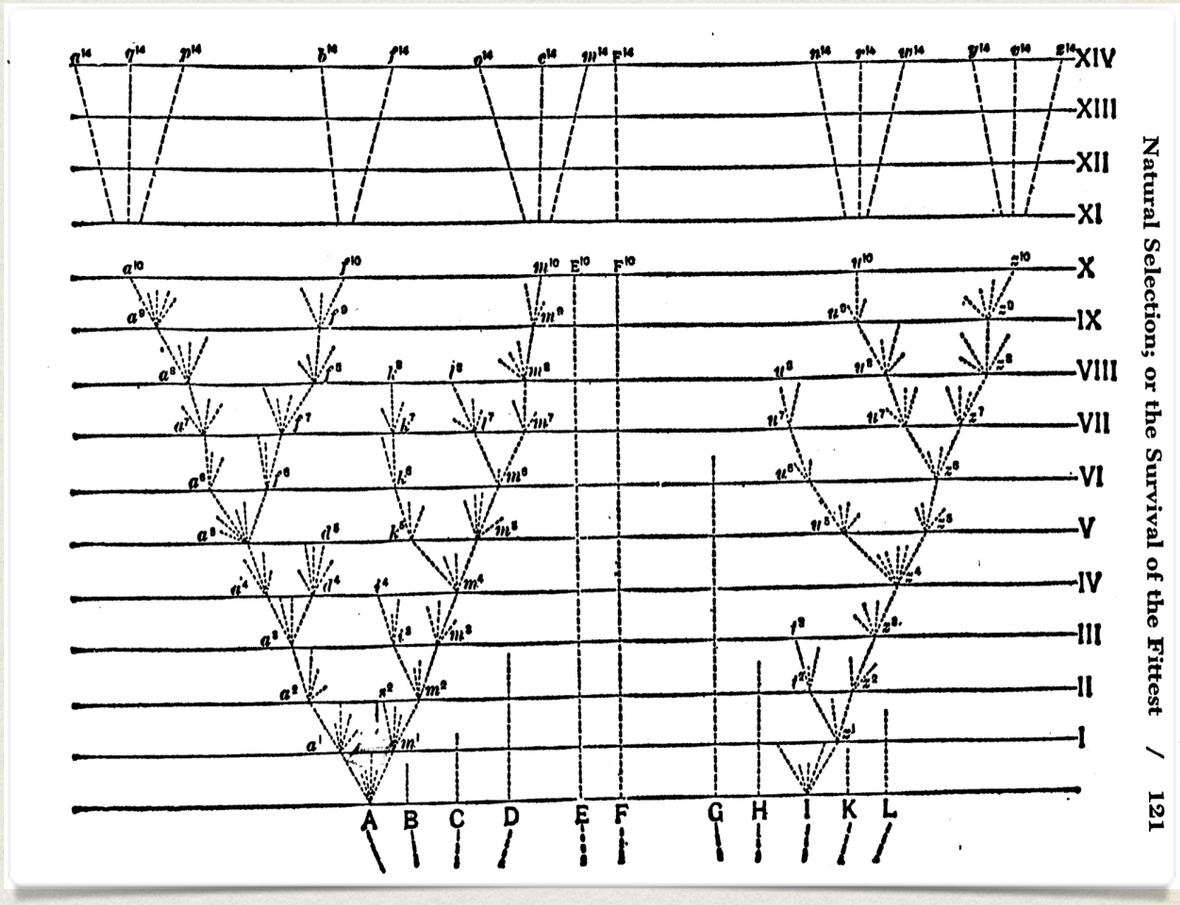
- Phylogenetics is the study of evolutionary relationships
- Phylogenetic analysis is the means used to estimate evolutionary relationships based on observable evidence
- Evidence can include morphology, physiology, and other properties of organisms. Paleontological and geological evidence is also used.

THE ONLY FIGURE IN "THE ORIGIN OF SPECIES"

The affinities of all the beings of the same class have sometimes be represented by a great tree. I believe this simile largely speaks the truth……

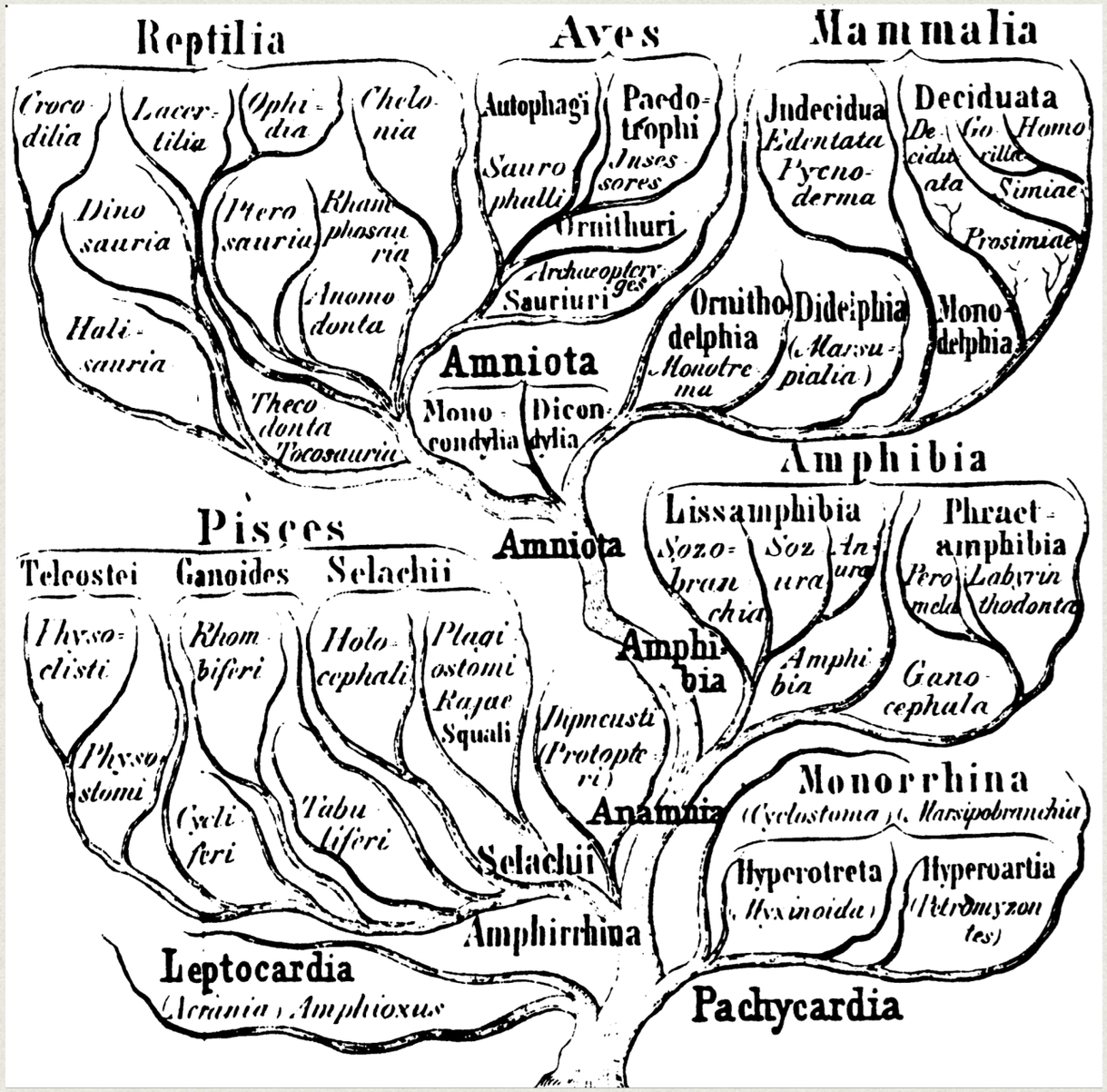
…The green and budding twigs may represent existing species; and those produced during former years may represent the long succession of extinct species…..

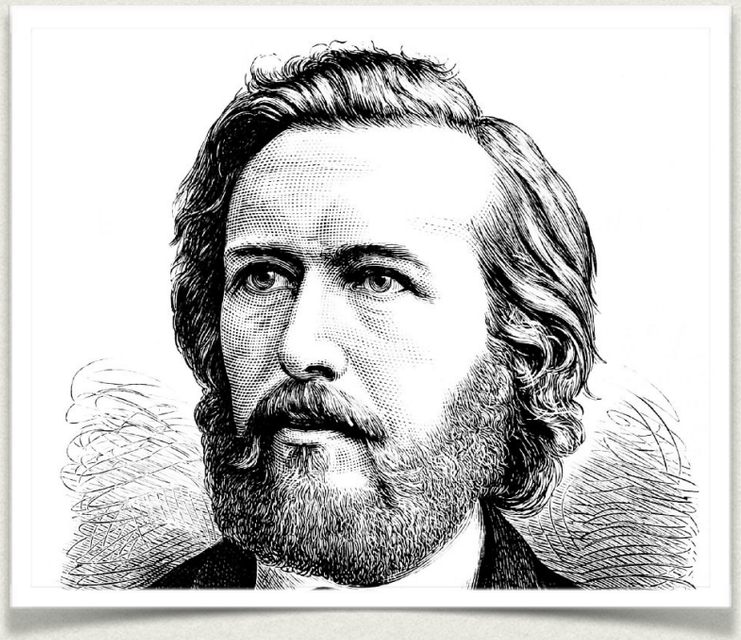
….the great Tree of Life….covers the earth with ever-branching and beautiful ramifications



Charles Darwin, 1856

THE USE OF TREES AS METAPHORS WAS PROMOTED BY ERNST HAECKEL

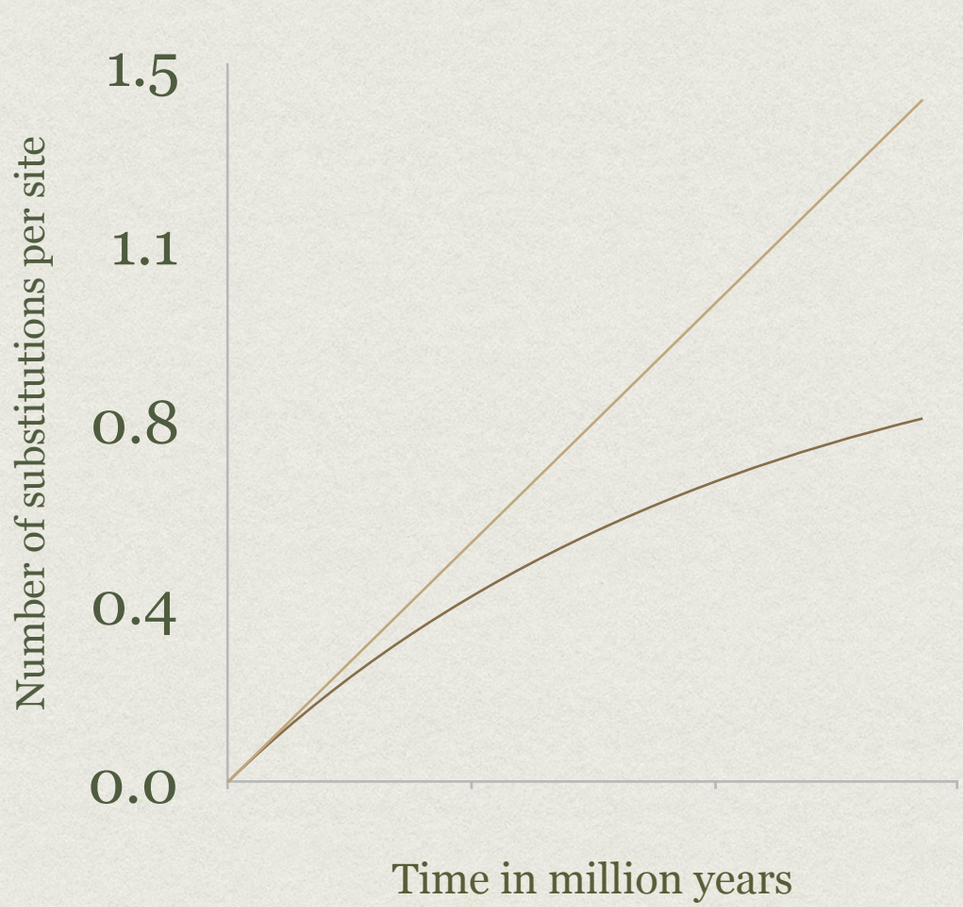




Haeckel (1866)
Generelle Morphologie der Organismen

MOLECULAR PHYLOGENETICS

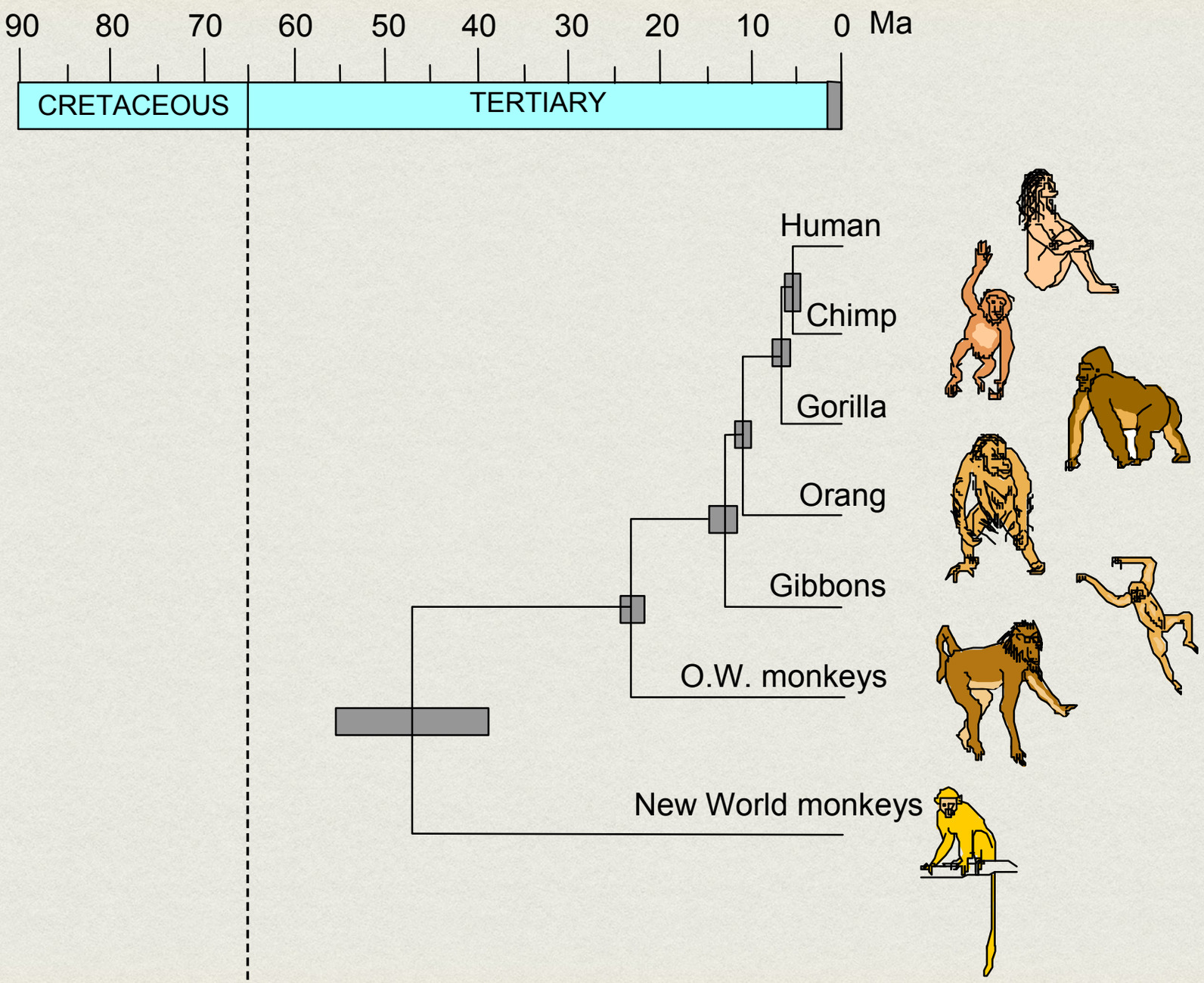
- The molecular biology of an organism can also provide evidence for phylogenetic analysis
- Accumulated mutational changes in DNA and protein sequence over time constitutes evidence
- Sequence-based phylogenetic analysis can be automated or semi-automated using computers



THINGS TO REMEMBER

- The events that determine a phylogeny happened in the past
- They cannot be known empirically, they can only be inferred from their "end products", whether these are morphological or molecular
- The tree is the model of evolutionary events that best explains the end product (diverged group of sequences)
- Phylogenetic analysis is modeling or estimation, and the quality or certainty of the analysis should be presented along with the result

EXAMPLES OF PHYLOGENETIC ANALYSIS: MOLECULAR TAXONOMY

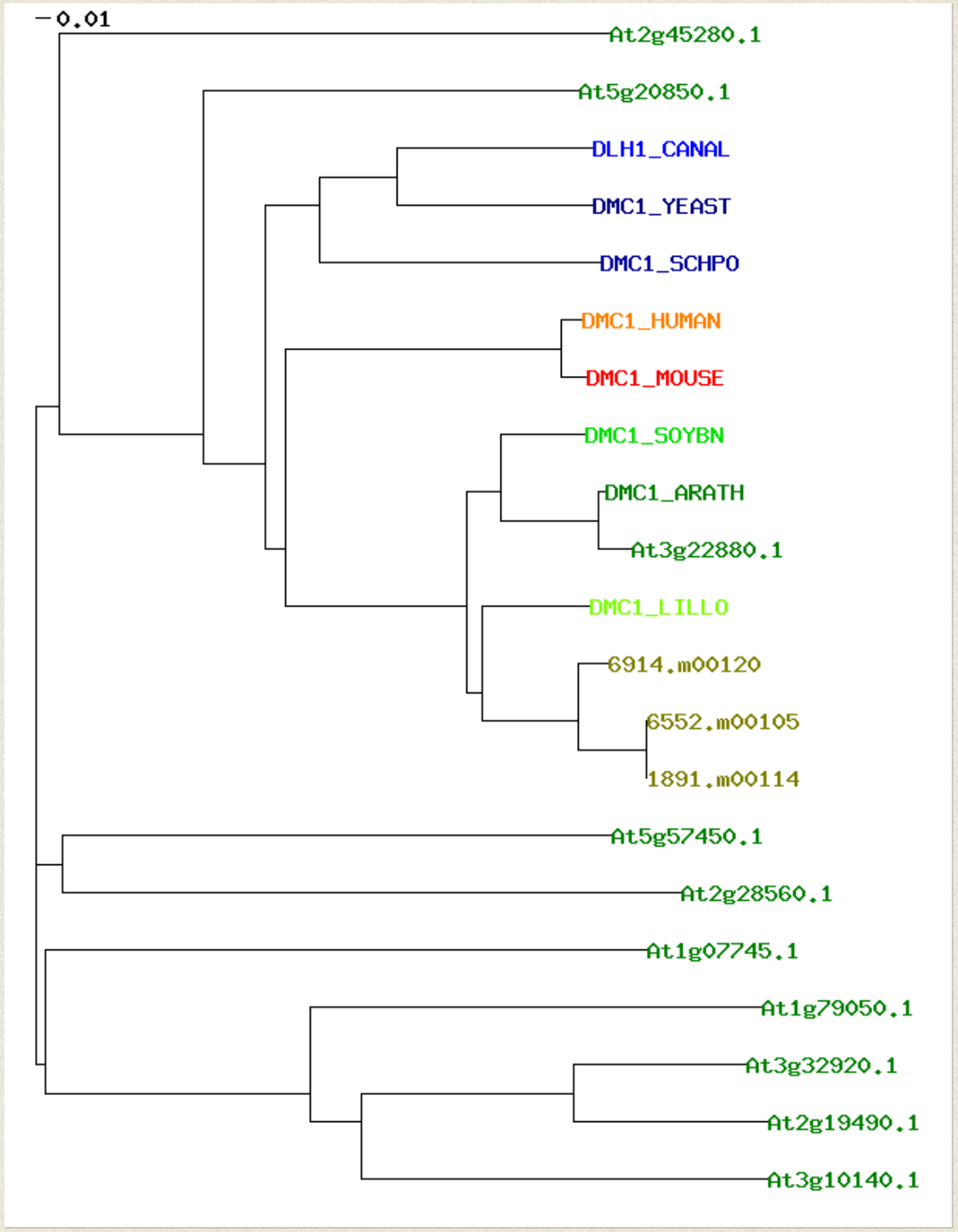


Stauffer et al. (2001);
Kumar & Hedges (1998)

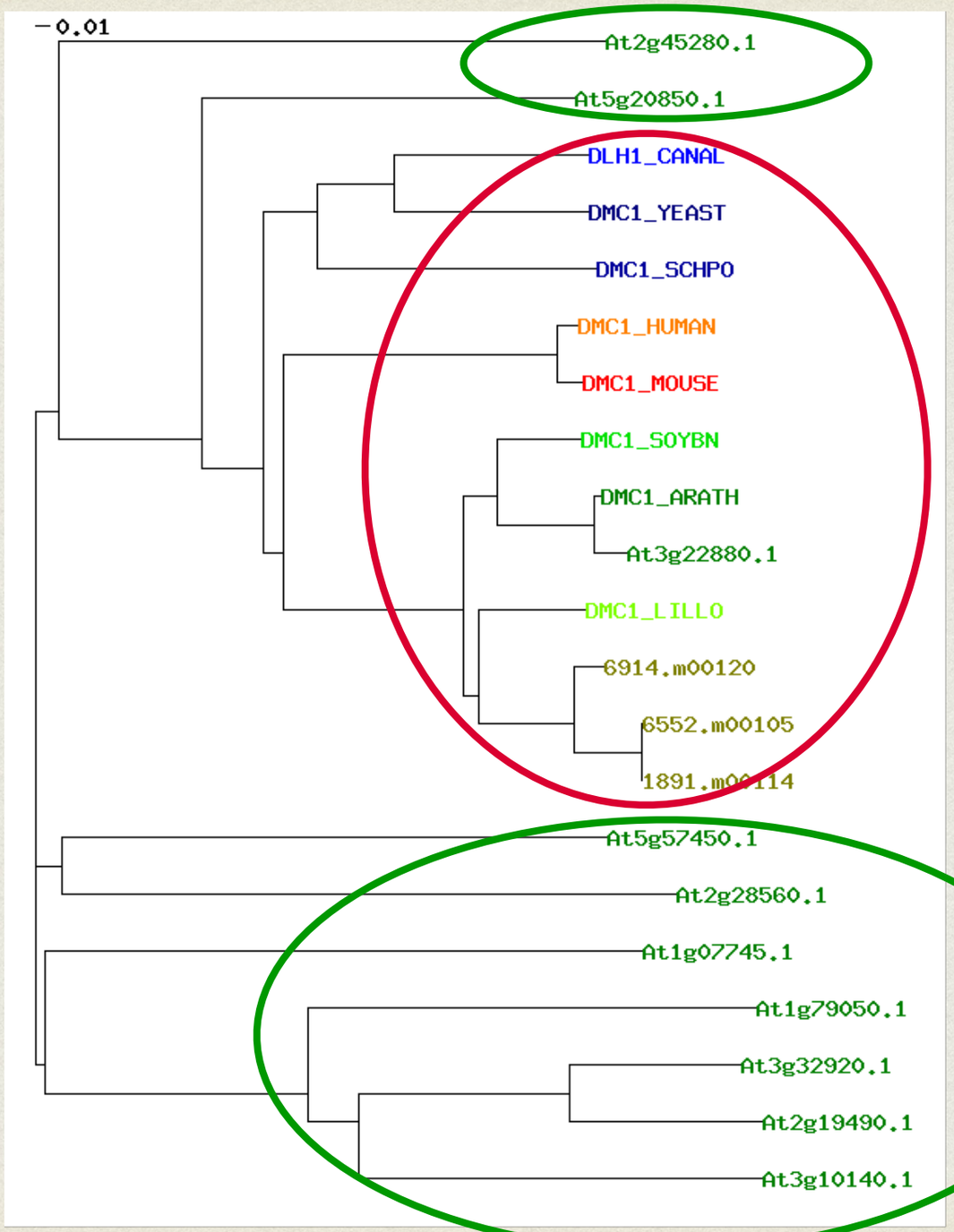
EXAMPLES OF PHYLOGENETIC ANALYSIS: EVOLUTIONARY HISTORY OF A SINGLE MOLECULE

DMC1

DNA meiotic recombinase 1



EXAMPLES OF PHYLOGENETIC ANALYSIS: EVOLUTIONARY HISTORY OF A SINGLE MOLECULE

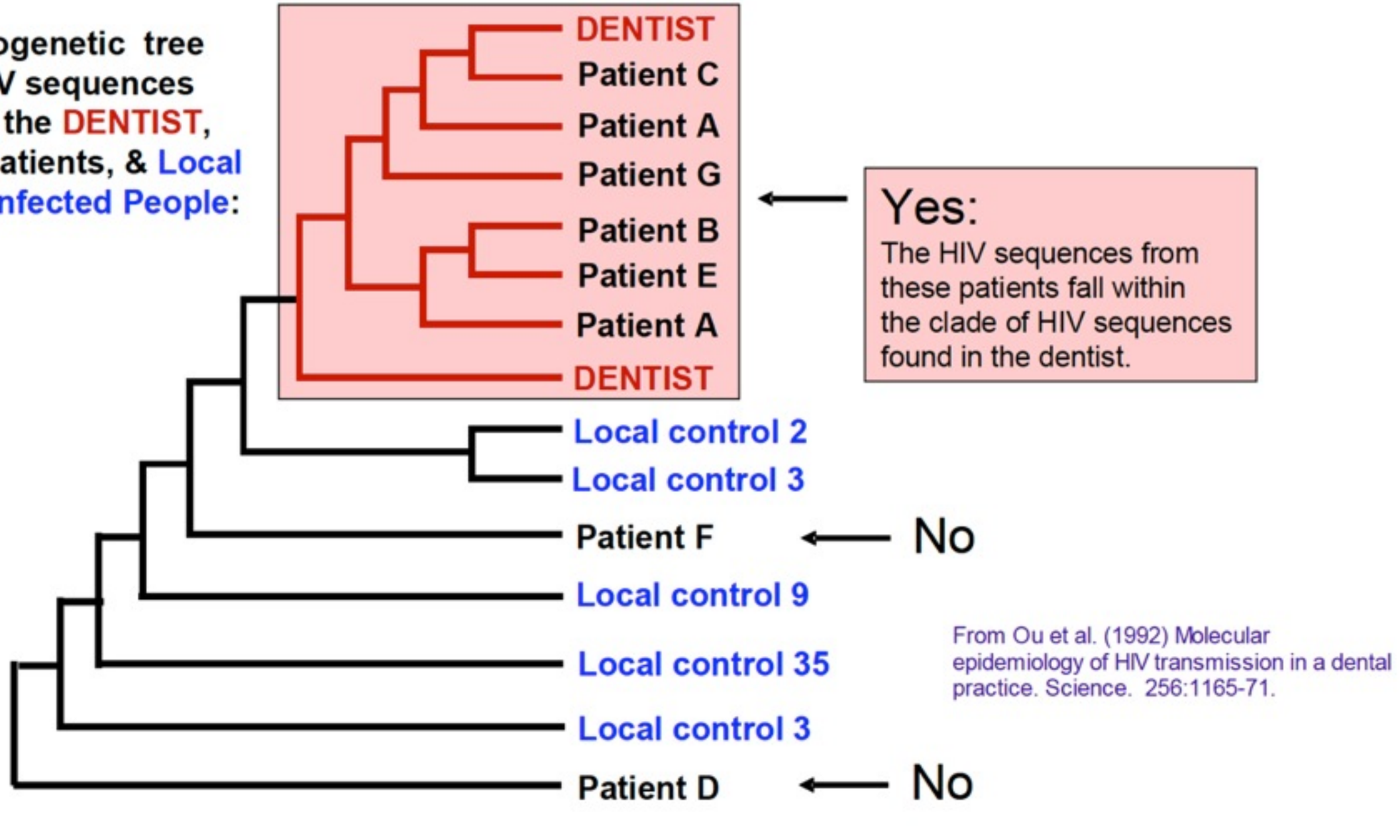


first cluster of paralogs
in Arabidopsis

DMC1 orthologs

second cluster of paralogs
in Arabidopsis

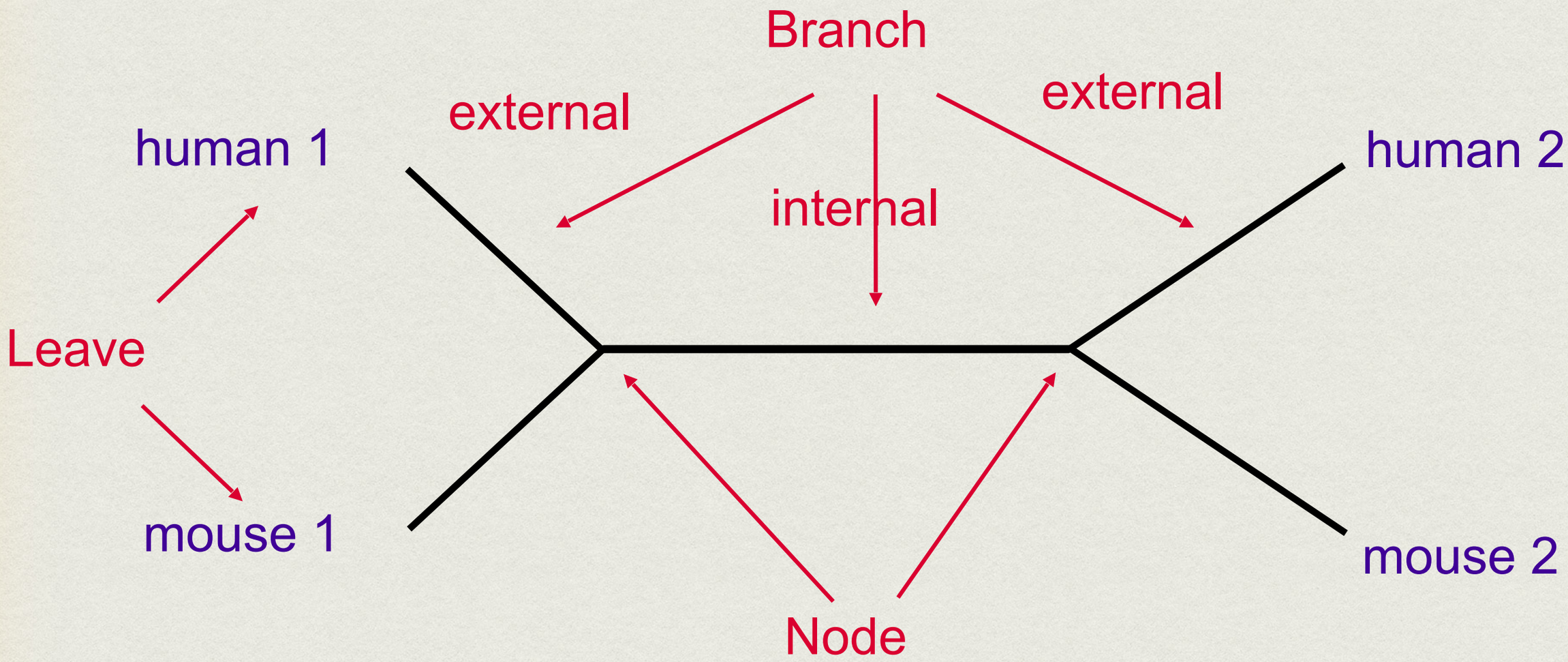
EXAMPLES OF PHYLOGENETIC ANALYSIS: MOLECULAR EPIDEMIOLOGY



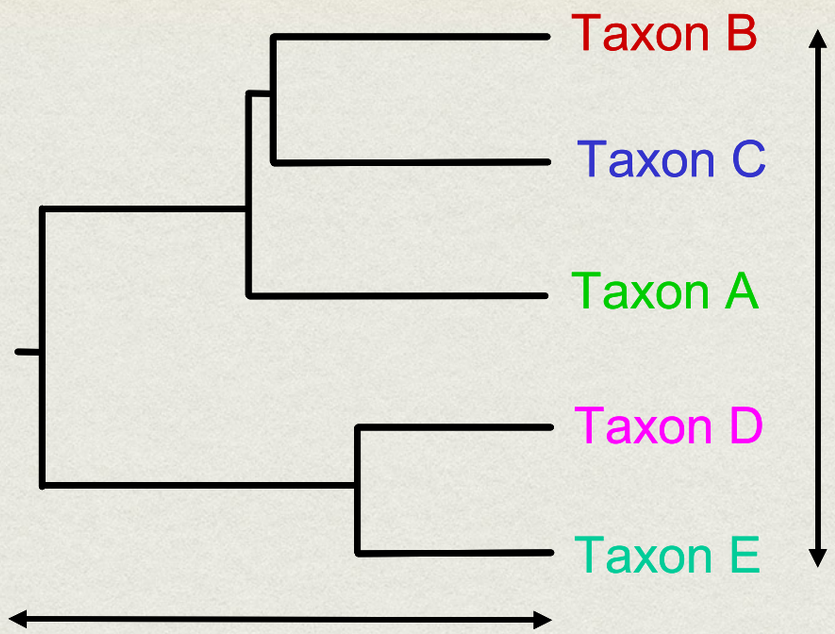
NOMENCLATURE

- A phylogenetic tree is characterized by ”leaves”, ”nodes” and ”branches.”
  - Leaves (vertices) represent species or sequences compared.
  - Nodes (vertices) are usually bifurcations and represent gene duplication or speciation events, hypothetical ancestor sequences.
  - Branches (edges) are always linear and represent sequence diversity but can also be of unit length.
  - The root (vertex) is optional and represents the hypothetical ancestor.

NOMENCLATURE



TREE INTERPRETATION



There's no meaning to the spacing between the taxa, or to the order in which they appear from top to bottom.

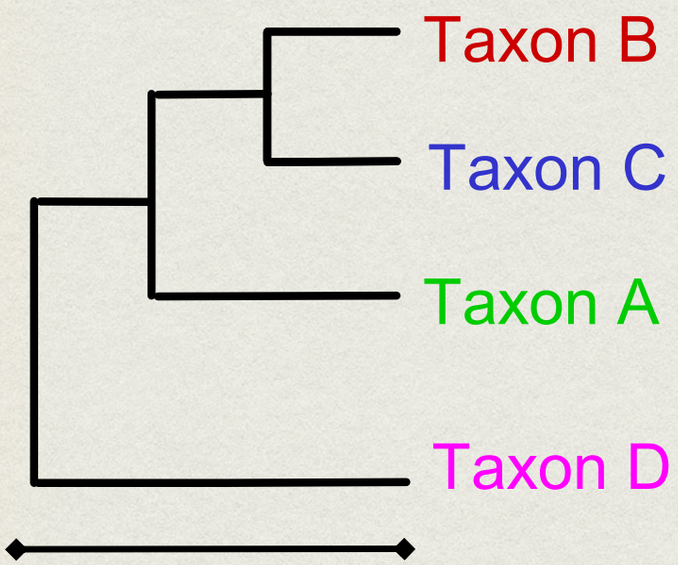
This dimension either can have no scale (for 'cladograms'), can be proportional to genetic distance or amount of change (for 'phylograms' or 'additive trees'), or can be proportional to time (for 'ultrametric trees' or true evolutionary trees).

((A,(B,C)),(D,E)) = The above phylogeny as nested parentheses, so called the Newick tree format

The above tree suugests that B and C are more closely related to each other than either is to A, and that A, B, and C form a clade that is a sister group to the clade composed of D and E. If the tree has a time scale, then D and E are the most closely related.

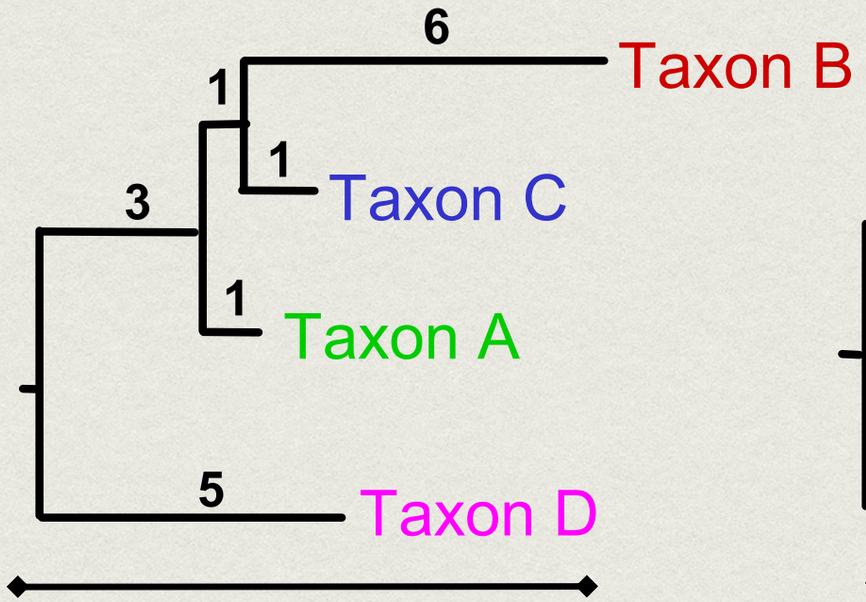
TYPES OF TREES

Cladogram



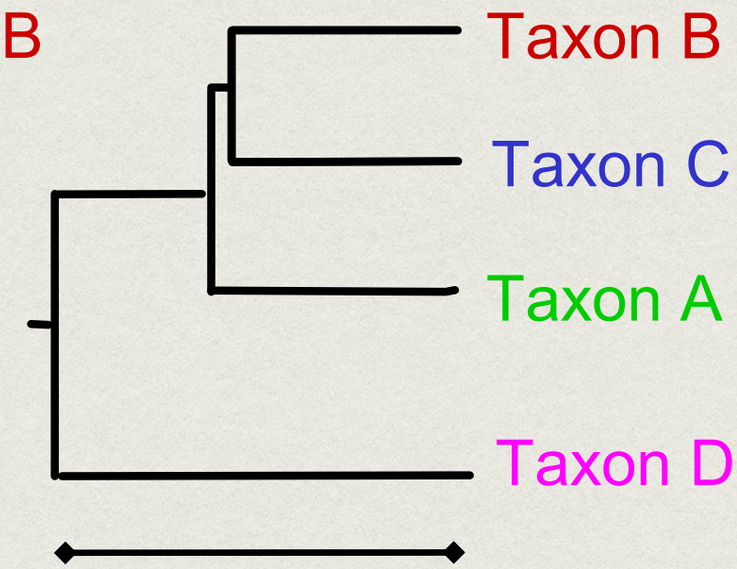
no meaning

Phylogram



genetic change

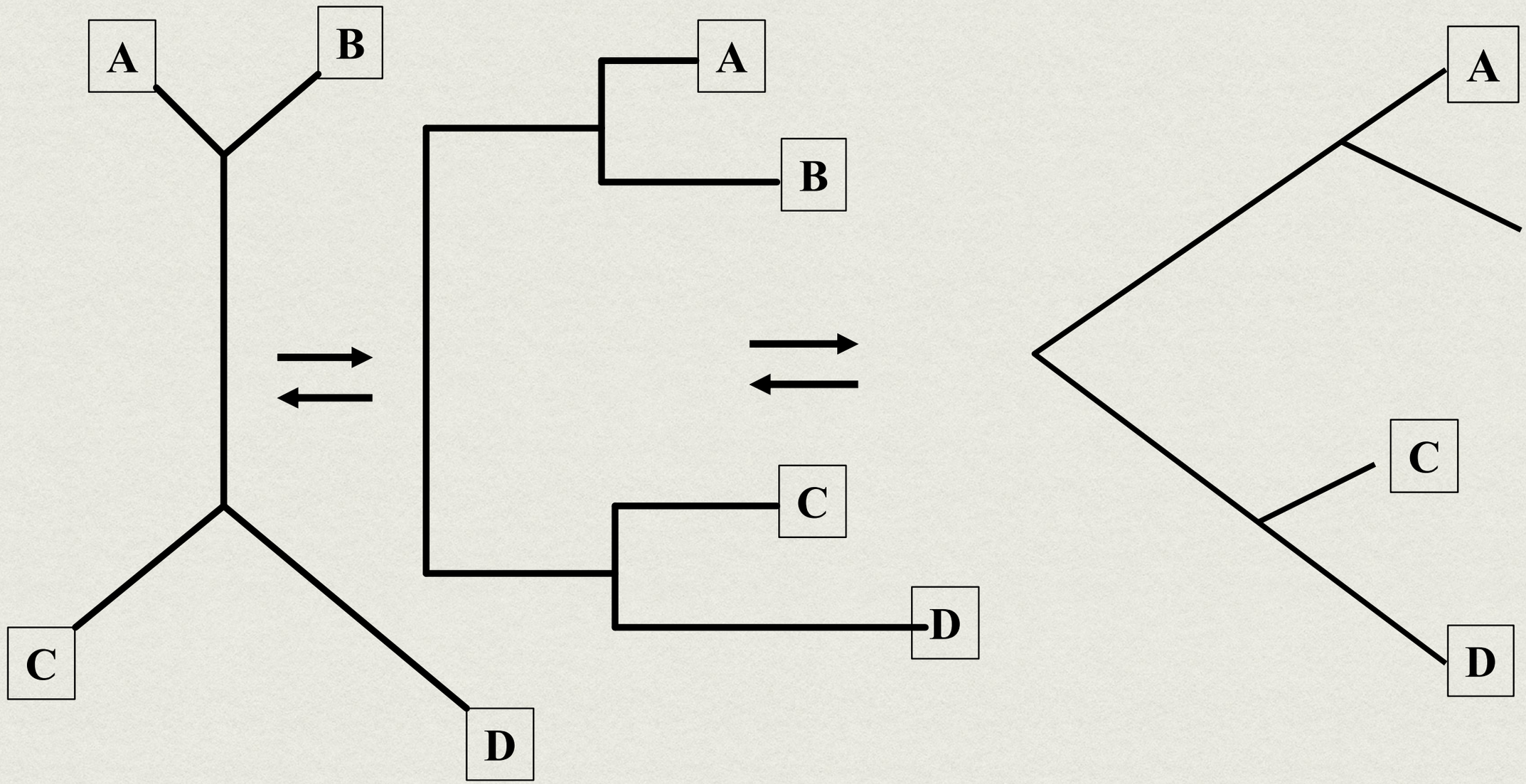
Ultrametric tree



time

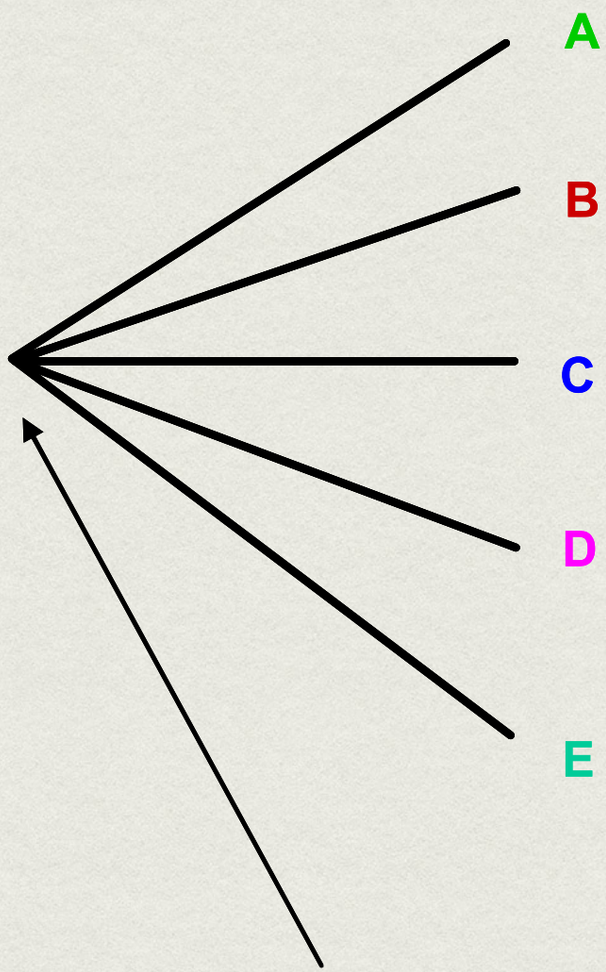
All show the same evolutionary relationships, or branching orders, between the taxa.

TREE PRESENTATION - DIFFERENT GRAPHS THE SAME MEANING



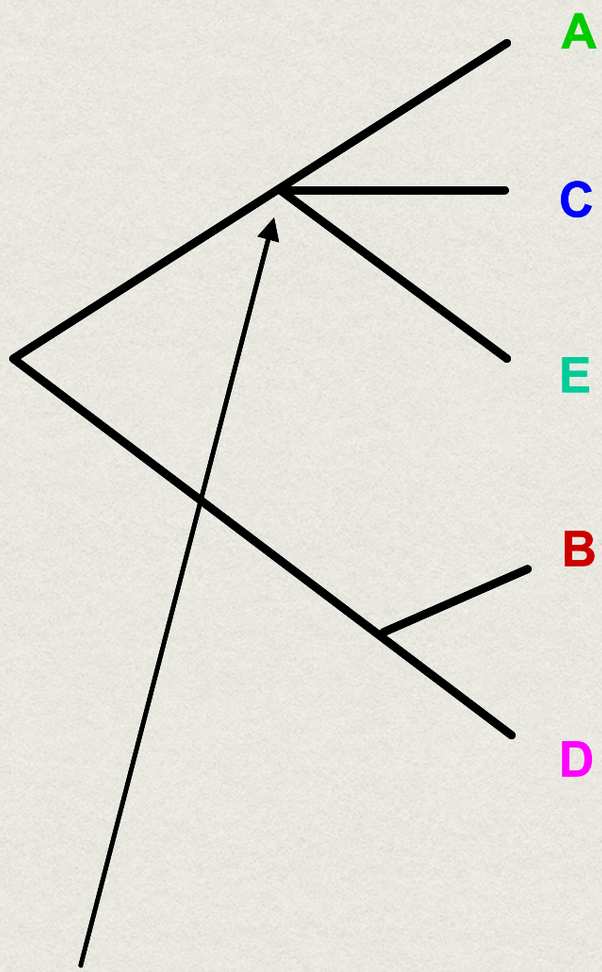
THE GOAL OF PHYLOGENY INFERENCE IS TO RESOLVE THE BRANCHING ORDERS OF LINEAGES IN EVOLUTIONARY TREES:

Completely unresolved or "star" phylogeny

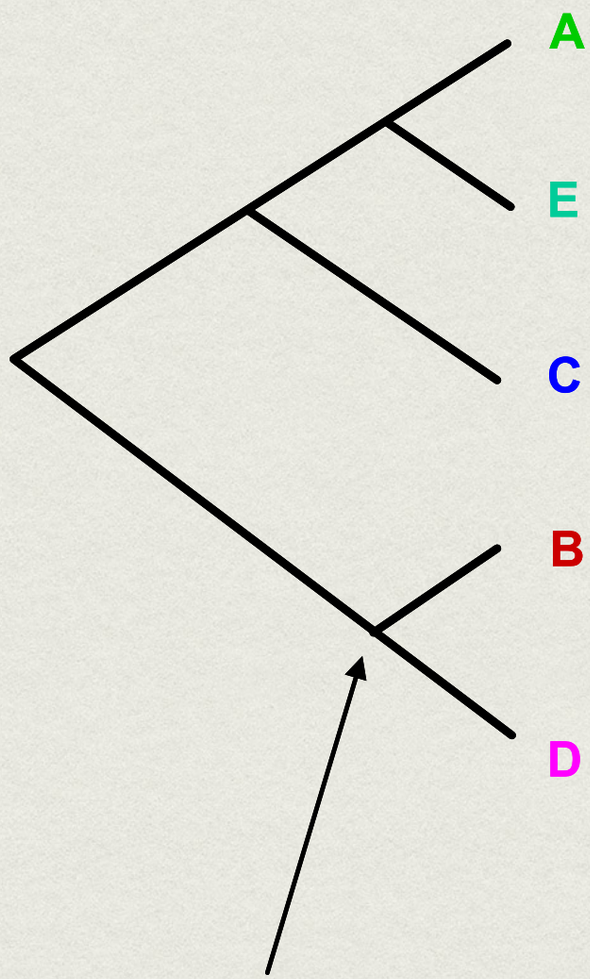


Polytomy or multifurcation

Partially resolved phylogeny

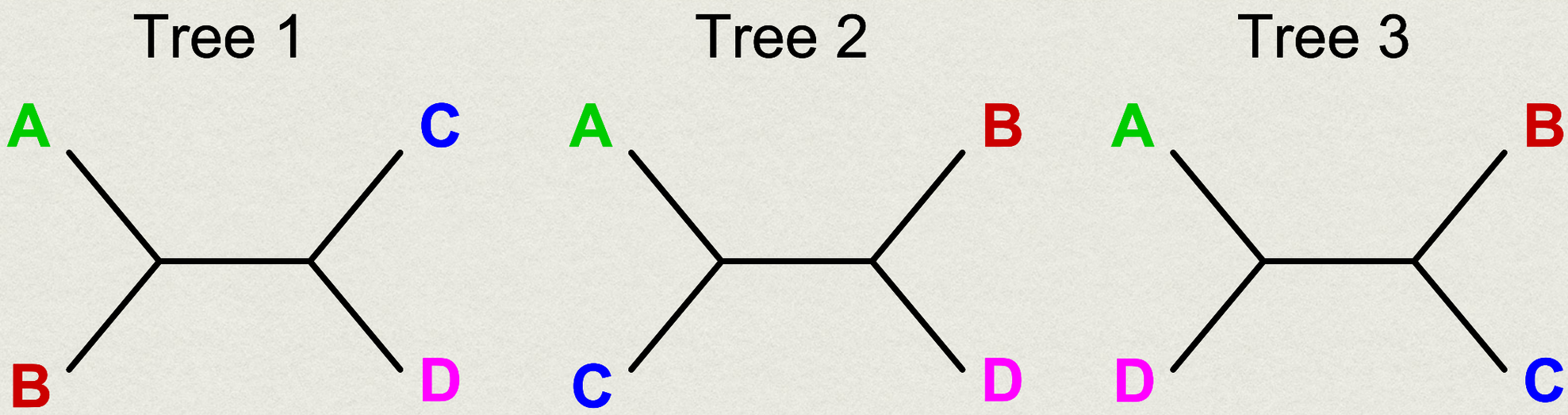


Fully resolved, bifurcating phylogeny



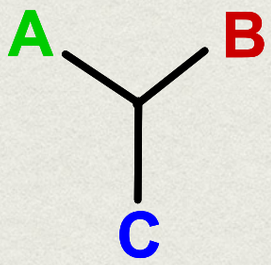
A bifurcation

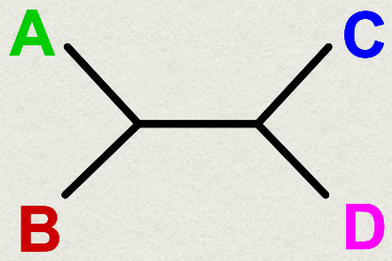
THERE ARE THREE POSSIBLE UNROOTED TREES FOR FOUR TAXA

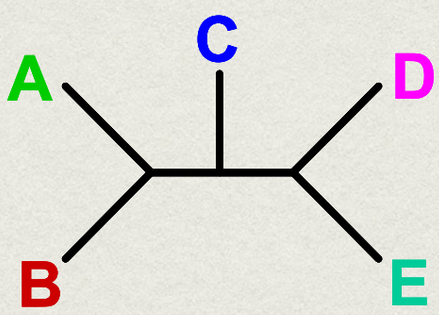


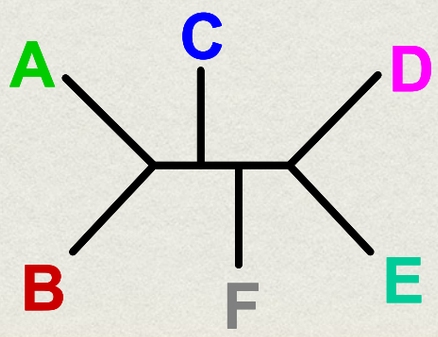
Phylogenetic tree building (or inference) methods are aimed at discovering which of the possible unrooted trees is "correct". We would like this to be the “true” biological tree — that is, one that accurately represents the evolutionary history of the taxa. However, we must settle for discovering the computationally correct or optimal tree for the phylogenetic method of choice.

THE NUMBER OF UNROOTED TREES INCREASES IN A GREATER THAN EXPONENTIAL MANNER WITH NUMBER OF TAXA



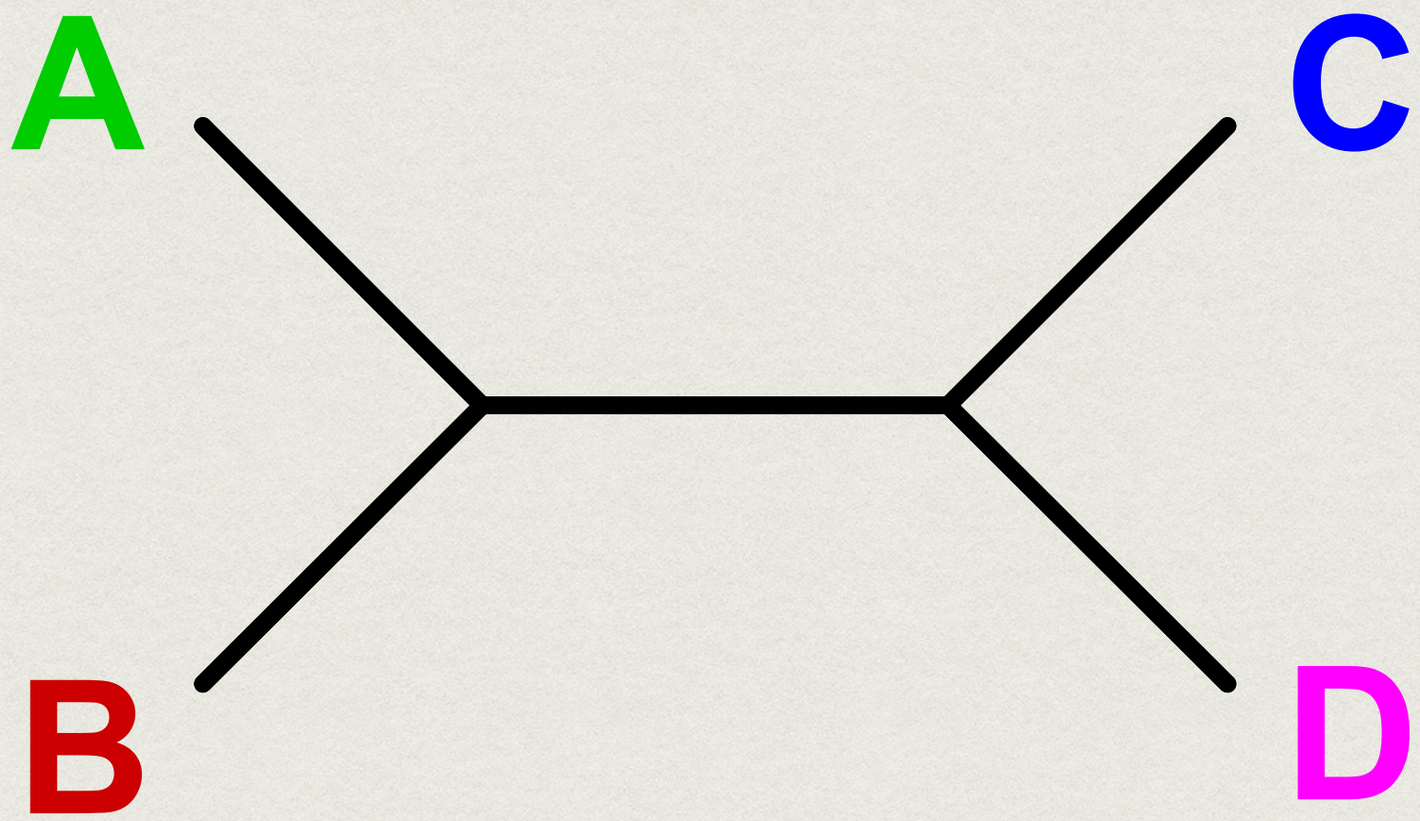






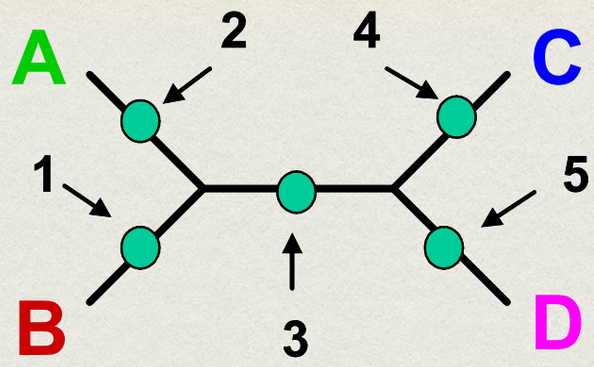
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
30	3.58×10^{36}

AN UNROOTED, FOUR-TAXON TREE CAN BE ROOTED IN FIVE DIFFERENT PLACES TO PRODUCE FIVE DIFFERENT ROOTED TREES

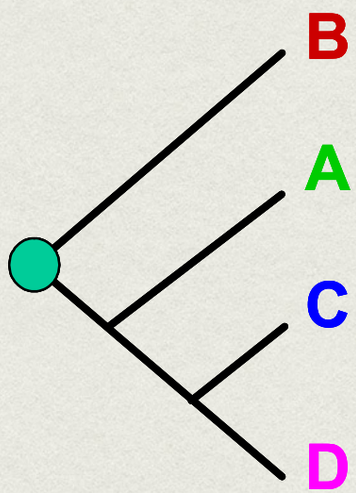


AN UNROOTED, FOUR-TAXON TREE CAN BE ROOTED IN FIVE DIFFERENT PLACES TO PRODUCE FIVE DIFFERENT ROOTED TREES

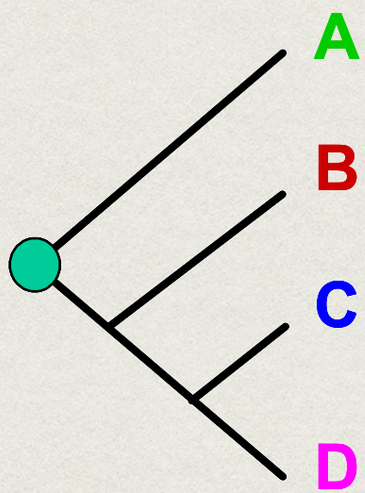
The unrooted tree:



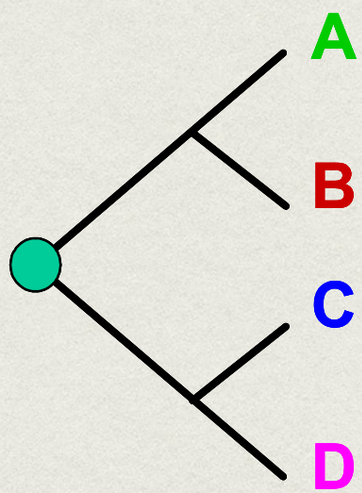
Rooted tree 1



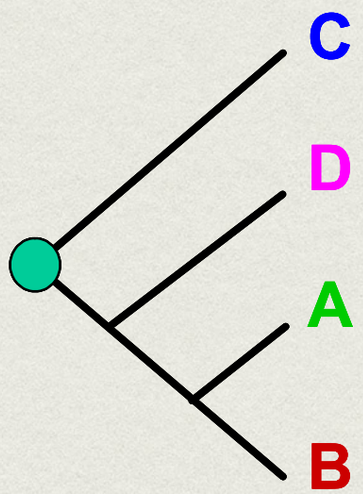
Rooted tree 2



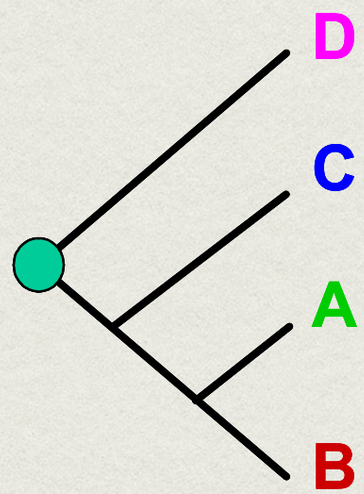
Rooted tree 3



Rooted tree 4



Rooted tree 5



These trees show five different evolutionary relationships among the taxa!

FIVE STEPS IN BUILDING A PHYLOGENETIC ANALYSIS

- Finding all homologs
- Multiple sequence alignment
- Building a tree
- Statistical assessment of a tree
- Viewing a tree and drawing conclusions

STEP 1: FINDING ALL HOMOLOGS

- Sequence homology search is the most popular approach:
  - use protein sequences
  - use PSI-BLAST or delta-BLAST not a simple BLASTp
- Text search in protein databases is often useful in finding distant, very diverged homologs
- Search protein domains database, e.g. Pfam

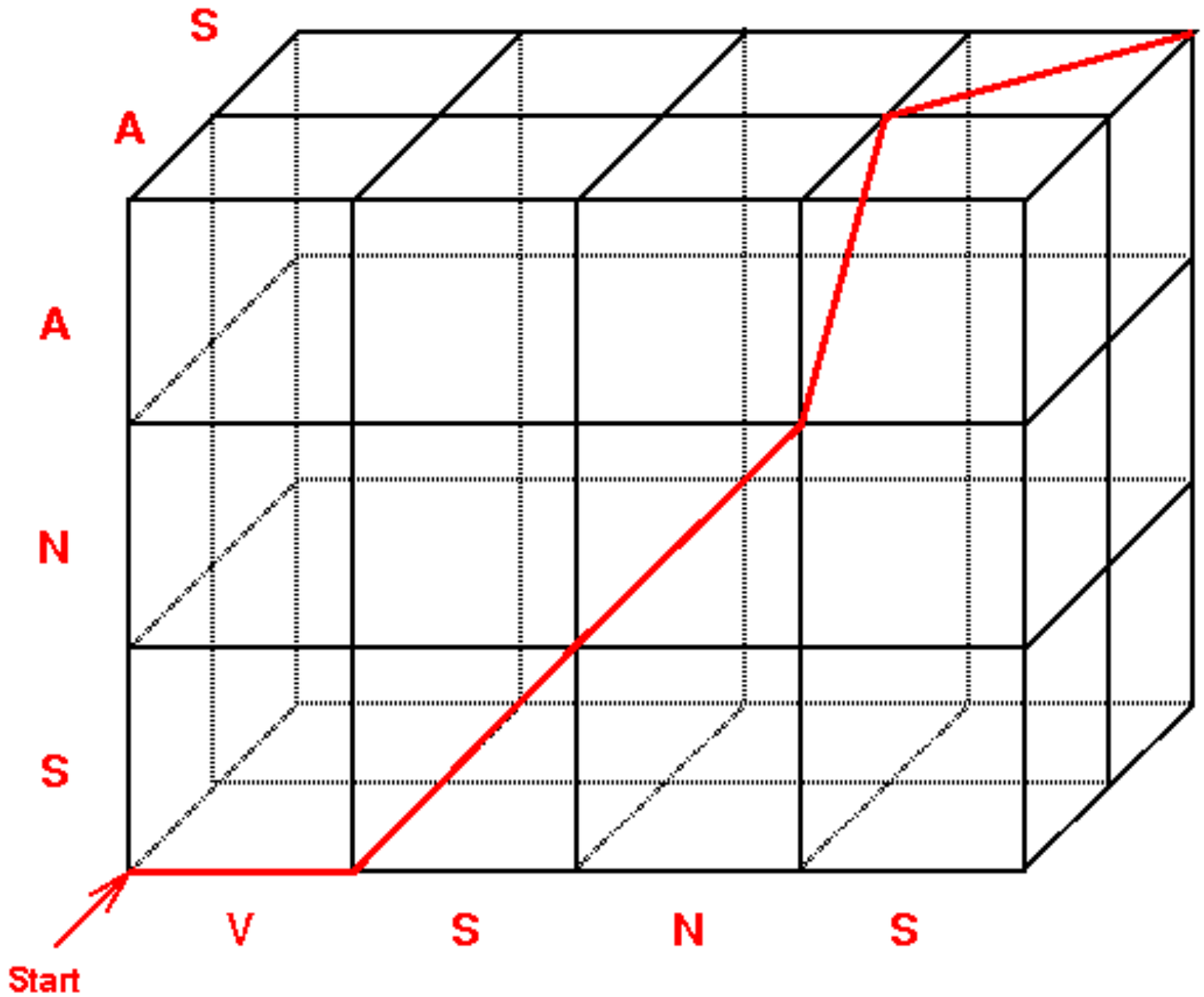
STEP 2: MULTIPLE SEQUENCE ALIGNMENT

- Approaches to Multiple Sequence Alignment
  - Dynamic Programming
  - Progressive Alignment
  - Iterative Alignment
  - Statistical Modeling

DYNAMIC PROGRAMMING APPROACH

- Dynamic programming with two sequences
  - Relatively easy to code
  - Guarantee to obtain optimal alignment
- Can this be extended to multiple sequences?

DYNAMIC PROGRAMMING WITH THREE SEQUENCES



V S N _ S
_ S N A _
_ _ _ A S

Figure source: http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/node2.html

MULTIPLE DYNAMIC PROGRAMMING COMPLEXITY

Memory requirements if each sequence has length of n

- 2 sequences: $O(n^2)$
- 3 sequences: $O(n^3)$
- k sequences: $O(n^k)$

Time problem:

$$O(2^k \prod_{i=1,\ldots,k} |s_i|)$$

If the calculation factor is one nanosecond, then for six sequences of length 100, we'll have a running time of $2^6 \times 100^6 \times 10^{-9}$, that's roughly 64000 seconds (almost 18 hours). Just add two sequences, and the running time increases to 2.56×10^9 seconds (over 81 years)!

SOLUTION: PROGRESSIVE ALIGNMENTS

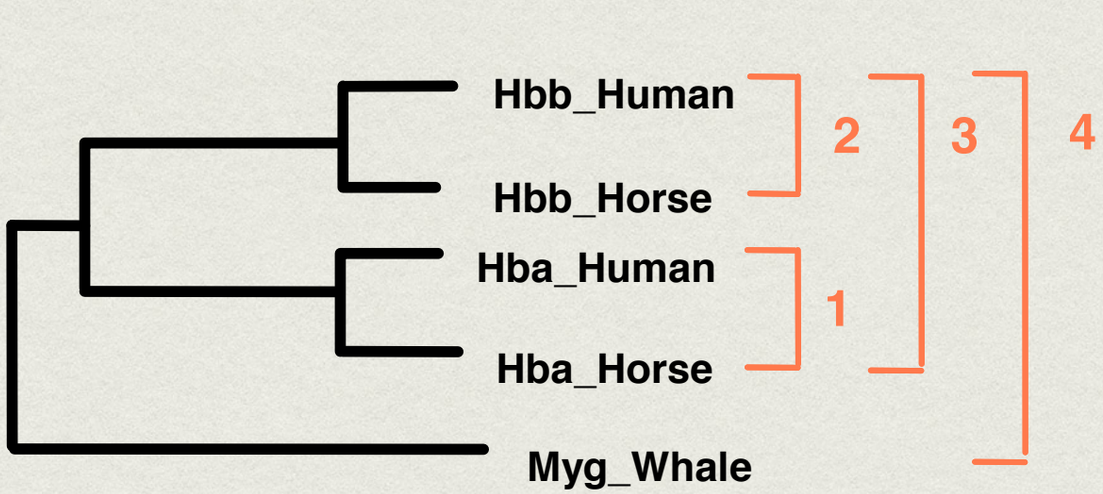
- Align most related sequences
- Add on less related sequences to initial alignment
- Software Examples:
  - ClustalW
  - MultAlin

PROGRESSIVE ALIGNMENT

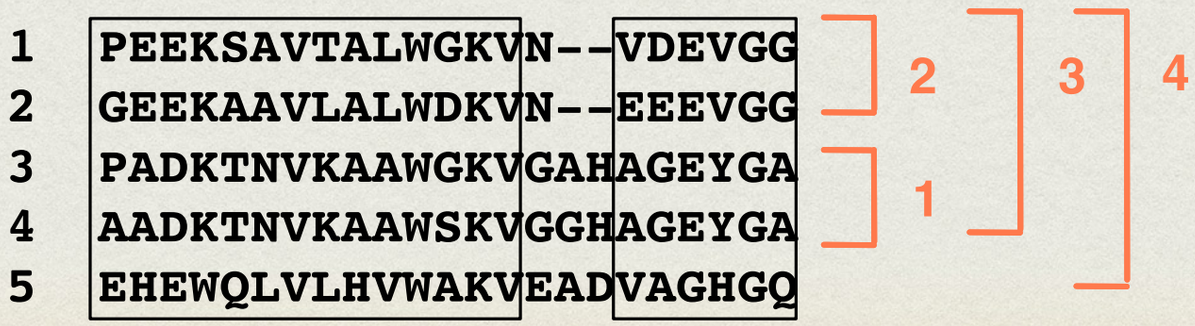
- Devised by Feng and Doolittle in 1987
  - Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 25(4):351-60
- Essentially a heuristic method and as such is not guaranteed to find the ‘optimal’ alignment
- Requires n-1+n-2+n-3…n-n+1 pairwise alignments as a starting point
- Most successful implementation is Clustal
  - Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.
  - Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882.

CLUSTALW - AN OVERVIEW

Hbb_Human	1	-				
Hbb_Horse	2	.17	-			
Hba_Human	3	.59	.60	-		
Hba_Horse	4	.59	.59	.13	-	
Myg_Whale	5	.77	.77	.75	.75	-



alpha-helices



CLUSTAL W

↓

Quick pairwise alignment:
calculate distance matrix

↓

Neighbor-joining tree
(guide tree)

↓

Progressive alignment
following guide tree

CLUSTALW- PAIRWISE ALIGNMENTS

- First perform all possible pairwise alignments between each pair of sequences. There are (n-1)+(n-2)...(n-n+1) possibilities.
- Calculate the ‘distance’ between each pair of sequences based on these isolated pairwise alignments.
- Generate a distance matrix.

CLUSTALW- GUIDE TREE

- Generate a Neighbor-Joining ‘guide tree’ from these pairwise distances

- This guide tree gives the order in which the progressive alignment will be carried out

CLUSTALW - FIRST PAIR

- Align the two most closely-related sequences first.
- This alignment is then ‘fixed’ and will never change. If a gap is to be introduced subsequently, then it will be introduced in the same place in both sequences, but their relative alignment remains unchanged

CLUSTALW- DECISION TIME

Consult the guide tree to see what alignment is performed next.

Align a third sequence to the first two

or

align two entirely different sequences to each other

Option 1 **Option 2**

CLUSTALW- PROGRESSION

The alignment is progressively built up in this way, with each step being treated as a pairwise alignment, sometimes with each member of a ‘pair’ having more than one sequence

CLUSTALW - GOOD POINTS/BAD POINTS

- Advantages
  - Speed
- Disadvantages
  - No objective function
  - No way of quantifying whether or not the alignment is good
  - No way of knowing if the alignment is ‘correct’
  - Potential problems:
  - Local minimum problem. If an error is introduced early in the alignment process, it is impossible to correct this later in the procedure
  - Arbitrary alignment

CLUSTALW - INCREASING THE SOPHISTICATION OF THE ALIGNMENT PROCESS

- realignment of selected sequences
- realignment of selected regions
- limited iteration of the alignment process
- pairwise alignment guided by protein secondary structure
- no penalty for terminal gaps

CLUSTALW- CAVEATS

- Sequence weighting
- Varying substitution matrices
- Residue-specific gap penalties and reduced penalties in hydrophilic regions (external regions of protein sequences), encourage gaps in loops rather than in core regions
- Positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage openings in subsequent alignments

ADVICE ON PROGRESSIVE ALIGNMENT

- Progressive alignment is a mathematical process that is completely independent of biological reality.
- Can be a very good estimate
- Can be an impossibly poor estimate
- Requires user input and skill
- Treat cautiously
- Can be improved by eye (usually)
- Often helps to have colour-coding
- Depending on the use, the user should be able to make a judgement on those regions that are reliable or not
- For phylogeny reconstruction, only use those positions whose hypothesis of positional homology is certain

FIVE STEPS IN BUILDING A PHYLOGENETIC ANALYSIS

- Finding all homologs
- Multiple-sequence alignment
- Building a tree
- Statistical assessment of a tree
- Viewing a tree and drawing conclusions

MOLECULAR PHYLOGENETIC TREE BUILDING METHODS

DATA TYPE	COMPUTATIONAL METHOD: Optimality criterion	COMPUTATIONAL METHOD: Clustering algorithm
Characters	PARSIMONY MAXIMUM LIKELIHOOD BAYESIAN INFERENCE	
Distances	MINIMUM EVOLUTION LEAST SQUARES	UPGMA NEIGHBOR-JOINING

TYPES OF DATA USED IN PHYLOGENETIC INFERENCE

Character-based methods: Use the aligned characters, such as DNA or protein sequences, directly during tree inference.

Taxa	Characters
Species A	**ATGGCTATTCTTATAGTACG**
Species B	**ATCGCTAGTCTTATATTACA**
Species C	**TTCACTAGACCTGTGGTCCA**
Species D	**TTGACCAGACCTGTGGTCCG**
Species E	**TTGACCAGTTCTCTAGTTCG**

Distance-based methods: Transform the sequence data into pairwise distances (dissimilarities), and then use the matrix during tree building.

	A	B	C	D	E
Taxon A	X	0.20	0.50	0.45	0.40
Taxon B	0.23	X	0.40	0.55	0.50
Taxon C	0.87	0.59	X	0.15	0.40
Taxon D	0.73	1.12	0.17	X	0.25
Taxon E	0.59	0.89	0.61	0.31	X

p-distances - the average difference per site (observed sequence difference)

Kimura 2-parameter distance (estimate of the true number of substitutions between taxa)

TYPES OF COMPUTATIONAL METHODS

- Clustering algorithms: Use pairwise distances.
  - These are purely algorithmic methods, in which the algorithm itself defines the tree selection criterion. Tend to be very fast programs that produce singular trees rooted by distance. No objective function to compare to other trees, even if numerous other trees could explain the data equally well.
  - Warning: finding a singular tree is not necessarily the same as finding the "true" evolutionary tree.
- Optimality approaches:
  - Use either character or distance data. First define an optimality criterion (minimum branch lengths, fewest number of events, highest likelihood), and then use a specific algorithm for finding trees with the best value for the objective function. Can identify many equally optimal trees, if such exist.
  - Warning: Finding an optimal tree is not necessarily the same as finding the "true" tree.

COMPUTATIONAL METHODS FOR FINDING OPTIMAL TREES

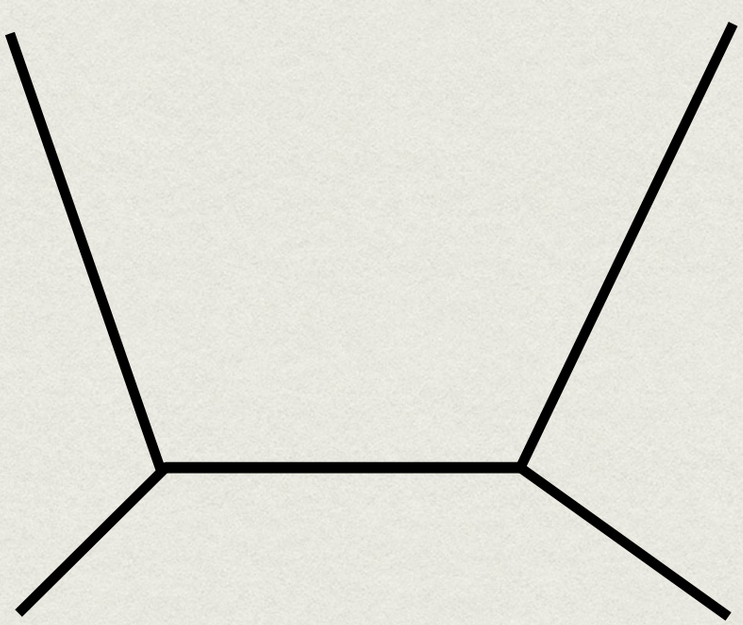
- Exact algorithms
  - Guarantee to find the optimal or "best" tree for the method of choice. Two types used in tree building:
  - Exhaustive search: Evaluates all possible unrooted trees, choosing the one with the best score for the method.
  - Branch-and-bound search: Eliminates the parts of the search tree that only contain suboptimal solutions
- Heuristic algorithms
  - Approximate or “quick-and-dirty” methods that attempt to find the optimal tree for the method of choice, but cannot guarantee to do so. Heuristic searches often operate by “hill-climbing” methods.

PARSIMONY METHODS

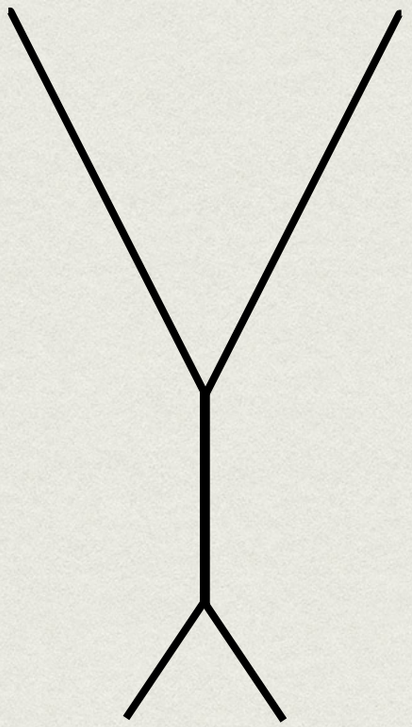
- Optimality criterion:
  - The ‘most-parsimonious’ tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences
- Advantages:
  - Are simple, intuitive, and logical (many possible by ‘pencil-and-paper’).
  - Can be used on molecular and non-molecular (e.g., morphological) data.
  - Can be used for character (can infer the exact substitutions) and rate analysis.
  - Can be used to infer the sequences of the extinct (hypothetical) ancestors.
- Disadvantages:
  - Can be fooled by high levels of homoplasy (‘same’ events).
  - Can become positively misleading in the “Felsenstein Zone” (long branch attraction)

PARSIMONY METHODS LONG BRANCH ATTRACTION

First time described by J. Felsenstein in 1978 (Syst. Zool. 27:401-410)



True tree



Inferred tree

MAXIMUM LIKELIHOOD (ML) METHODS

- Optimality criterion:
  - ML methods evaluate phylogenetic hypotheses in terms of the probability that a proposed model of the evolutionary process and the proposed unrooted tree would give rise to the observed data. The tree found to have the highest ML value is considered to be the preferred tree.

MAXIMUM LIKELIHOOD (ML) METHODS

- Advantages:
  - Are inherently statistical and evolutionary model-based.
  - Usually the most consistent of the methods available.
  - Can be used for character (can infer the exact substitutions) and rate analysis.
  - Can be used to infer the sequences of the extinct (hypothetical) ancestors.
  - Can help account for branch-length effects in unbalanced trees.
  - Can be applied to nucleotide or amino acid sequences, and other types of data.

MAXIMUM LIKELIHOOD (ML) METHODS

- Disadvantages:
  - Are not as simple and intuitive as many other methods.
  - Are computationally very intense.
  - Like parsimony, can be fooled by high levels of homoplasy.
  - Violations of the assumed model can lead to incorrect trees.
  - If model is wrong the inferred tree will be likely incorrect

BAYSIAN INFERENCE OF PHYLOGENY

- Start with best guess of a tree (prior probability)
- Simulation of trees (MCMC, Markov Chain Monte Carlo)
- Keep all the best trees
- Posterior tree with probabilities

MINIMUM EVOLUTION (ME) METHODS

- Optimality criterion:
  - The tree(s) with the shortest sum of the branch lengths (or overall tree length) is chosen as the best tree.
- Advantages:
  - Can be used on indirectly-measured distances (immunological, hybridization).
  - Distances can be ‘corrected’ for unseen events.
  - Usually faster than character-based methods.
  - Can be used for some rate analyses.
  - Has an objective function (as compared to clustering methods).
- Disadvantages:
  - Information lost when characters transformed to distances.
  - Cannot be used for character analysis.
  - Slower than clustering methods.

CLUSTERING METHODS (UPGMA & N-J)

- Optimality criterion:
  - NONE.
- Advantages:
  - Can be used on indirectly-measured distances (immunological, hybridization).
  - Distances can be ‘corrected’ for unseen events.
  - The fastest of the methods available.
  - Can therefore analyze very large datasets quickly.
- Disadvantages:
  - Similarity and relationship are not necessarily the same thing, so clustering by similarity does not necessarily give an evolutionary tree.
  - Cannot be used for character analysis!
  - Have no explicit optimization criteria, so one cannot even know if the program worked properly to find the correct tree for the method.

DISTANCE METHODS

- Based on precomputed pairwise distances between sequences according to the scoring scheme; the actual sequence is discarded once a distance matrix is computed
- Distance score is based on number of observed differences between two aligned sequences
- Pairwise alignment identity scores can be converted directly to distance scores; more sophisticated models contain heuristics to adjust for predicted number of multiple events at each site

DISTANCE METHODS

- Simplest distance measure = Hamming distance, number of changes (n) per unit sequence (N) = n/N; gaps can be ignored or treated as substitutions
- Assumes every change occurs only once, there are no duplicate changes at each site
- Can result in a zero or even negative branch length if that assumption is incorrect
- Alternate distance models -- e.g. probabilistic models like Jukes-Cantor, Kimura -- can be used to estimate probabilities that multiple changes have occurred at a site

MAJOR DISTANCE-BASED METHODS

- UPGMA (Unweighted pair group method with arithmetic mean) is a hierarchical clustering method that assumes a constant molecular clock (rate of evolution) along all branches of the tree.
- Two closest sequences are clustered first, then next two closest, etc. A rooted tree is produced.
- UPGMA assumes a molecular clock and results in a fixed (and error-prone) rooted tree topology. UPGMA methods are not recommended unless evolutionary rates can be assumed to be consistent in all branches in an entire protein group.

UPGMA - ALGORITHM

- Given a matrix of pairwise distances, find the clusters (taxa) i and j such that dij is the minimum value in the table
- Define the depth of the branching between i and j (lij) to be dij/2
- If i and j were the last two clusters, the tree is complete. Otherwise, create a new cluster called u.
- Define the distance from u to each other cluster (k, with k g i or j) to bean average of the distances dki and dkj.
- Go back to step 1 with one less cluster; cluster i and j have been eliminated, and cluster u has benn added.

CLUSTER ANALYSIS (UPGMA) OF 5S rRNA EVOLUTIONARY DISTANCES ESTIMATES

	Bsu	Bst	Lvi	Amo	Mlu
Bacillus subtilis	x	0.1715	0.2147	0.3091	0.2326
Bacillus stearothermophilus		x	0.2991	0.3399	0.2058
Lactobacillus viridescens			x	0.2795	0.3943
Acholeplasma modicum				x	0.4289
Micrococcus luteus					x

Create a cluster between two taxa with the minimum distance - Bsu and Bst in the example above. Recalculate distances with Bsu-Bst cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

CLUSTER ANALYSIS (UPGMA) OF 5S rRNA EVOLUTIONARY DISTANCES ESTIMATES

	Bsu-Bst	Lvi	Amo	Mlu
Bsu-Bst	x	0.2569	0.3245	0.2192
Lactobacillus viridescens		x	0.2795	0.3943
Acholeplasma modicum			x	0.4289
Micrococcus luteus				x

Create a cluster between two taxa with the minimum distance - Bsu-Bst and Mlu in the example above. Recalculate distances with Bsu-Bst-Mlu cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

CLUSTER ANALYSIS (UPGMA) OF 5S rRNA EVOLUTIONARY DISTANCES ESTIMATES

	Bsu-Bst-Mlu	Lvi	Amo
Bsu-Bst-Mlu	x	0.3027	0.3593
Lactobacillus viridescens		x	0.2795
Acholeplasma modicum			x

Create a cluster between two taxa with the minimum distance - Lvi and Amo in the example above. Recalculate distances with Lvi-Amo cluster as a new operational unit.

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

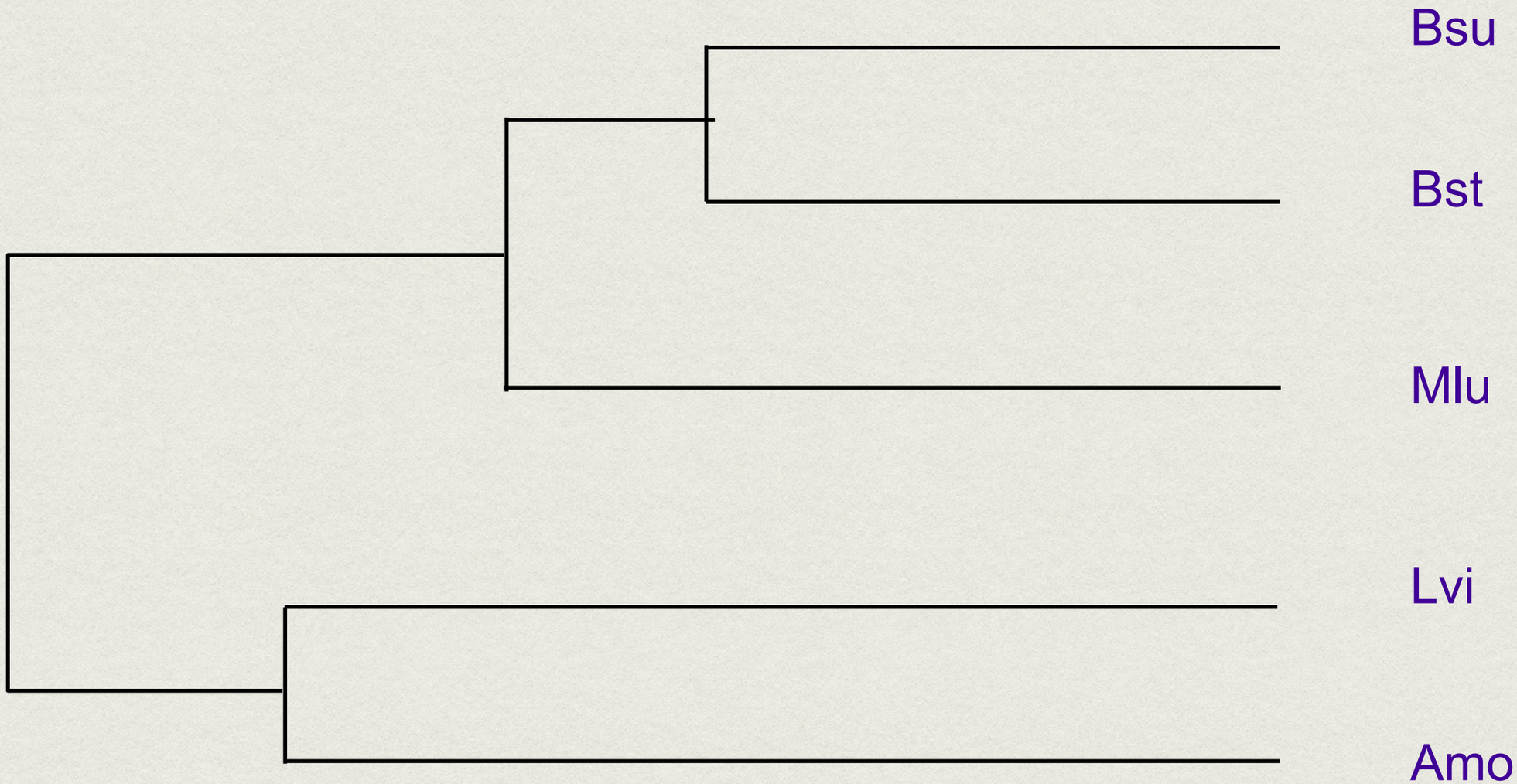
CLUSTER ANALYSIS (UPGMA) OF 5S rRNA EVOLUTIONARY DISTANCES ESTIMATES

	Bsu-Bst-Mlu	Lvi
Bsu-Bst-Mlu	x	0.3310
Lvi-Amo		x

Create the last cluster. Draw the tree

Data from Olsen (1988) Phylogenetic analysis using ribosomal RNA. Meth. Enzymol. 164: 793-838.

CLUSTER ANALYSIS (UPGMA) OF 5S rRNA EVOLUTIONARY DISTANCES ESTIMATES - INFERRED TREE



MAJOR DISTANCE-BASED METHOD

- Neighbor-joining (NJ) is in some sense the opposite of the UPGMA process. Rather than starting with closest sequence pairs and allowing early selections to bias the tree topology, NJ begins with an unresolved star-like cluster topology and selectively decomposes the alignment from this topology.

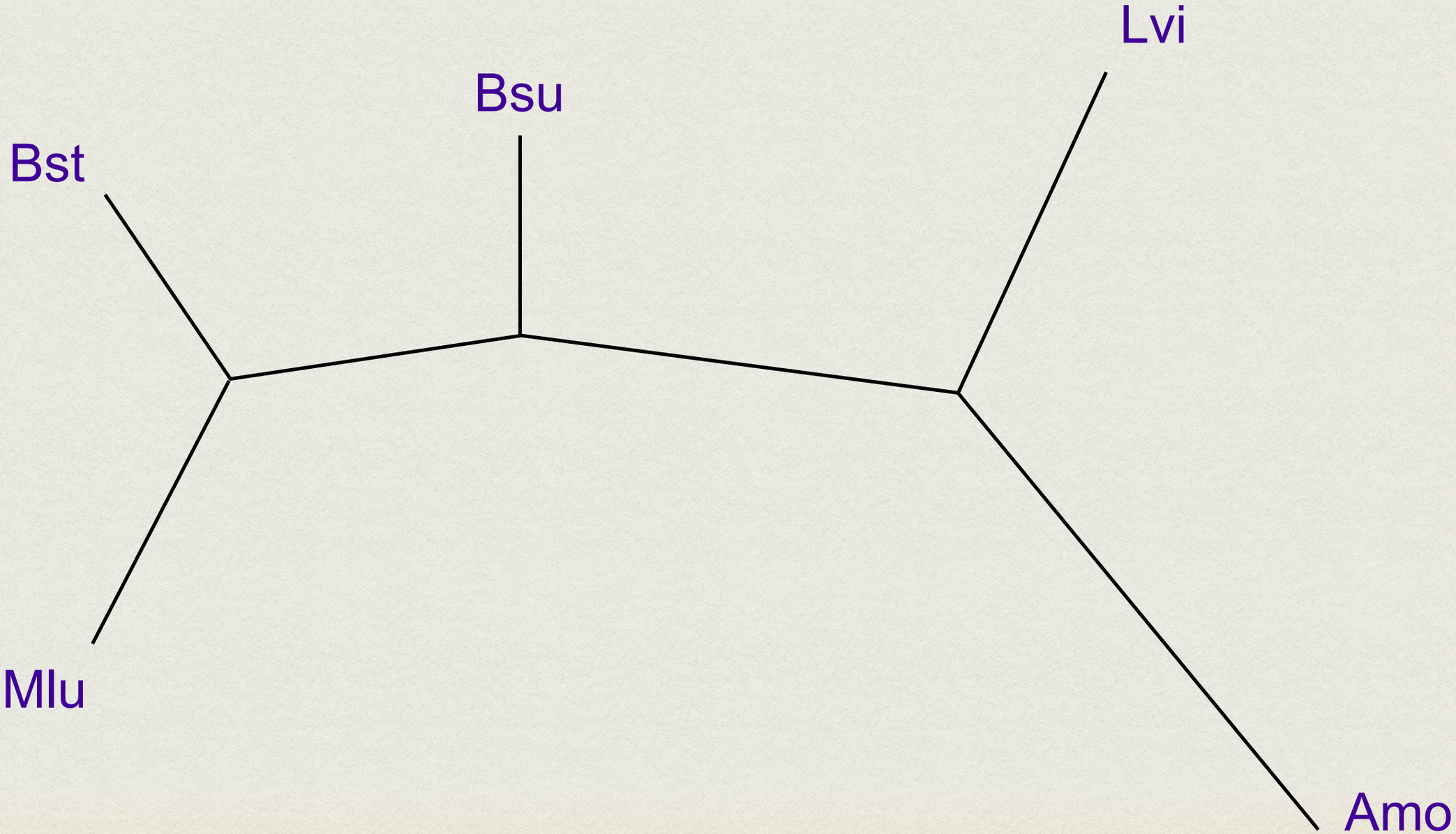
- Advantages: fast, yields one tree, usually reproduces trees close to those produced by more computationally intensive methods, does not assume consistent rates of evolution in each branch of the tree

DISTANCE METHODS - CONCLUSIONS

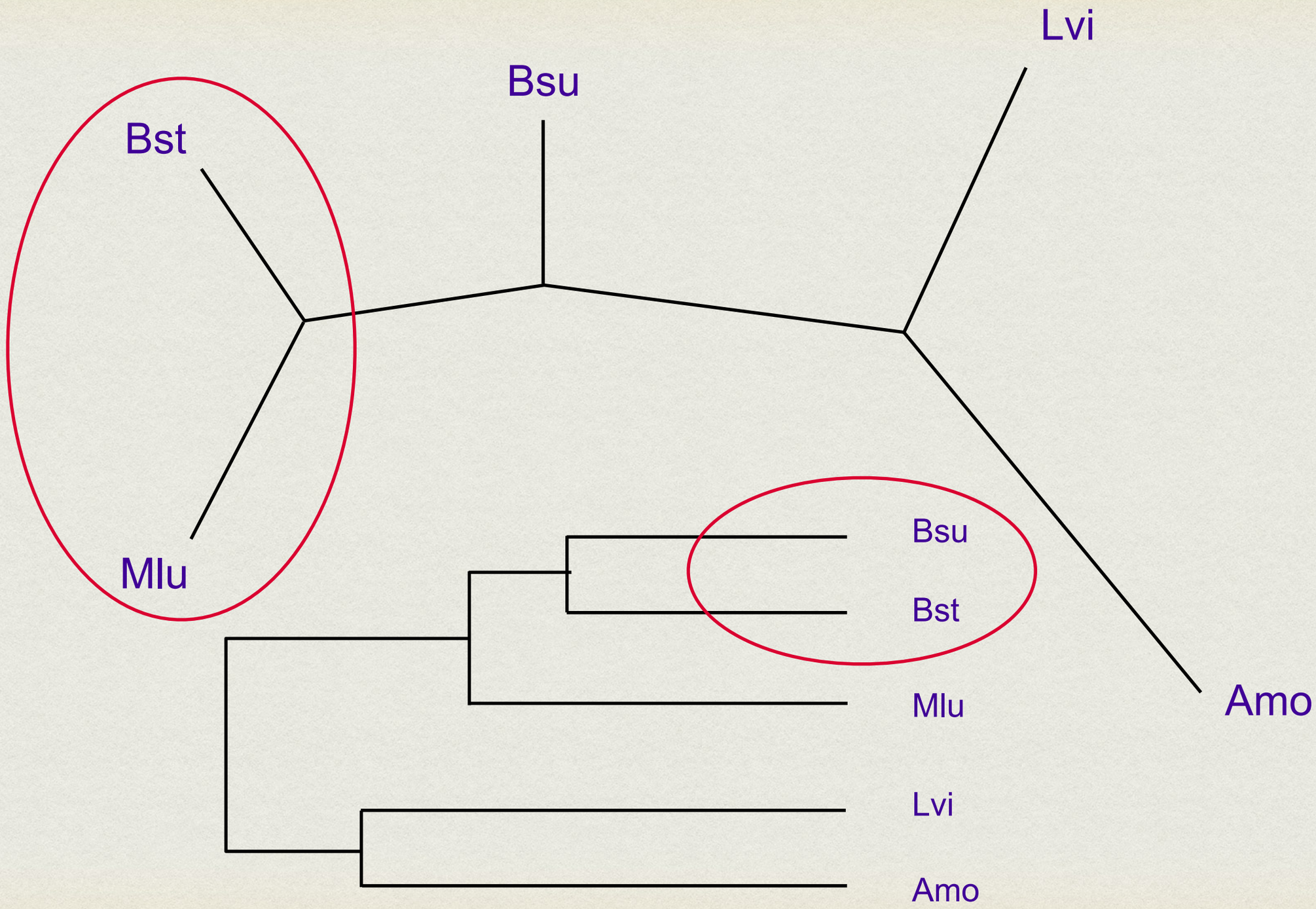
- Distance methods boil sequence data down to a single distance score
- By correcting that scored for multiple hits one tries to satisfy the additivity criterion
- For additive data NJ will work
- Otherwise ME or least-squares (FM) can be used to find the best tree for the distances

DIFFERENT METHODS - DIFFERENT RESULTS

Neighbor-joining (NJ) on 5S rRNA data



CLUSTER ANALYSIS OF 5S RRNA EVOLUTIONARY DISTANCES ESTIMATES - INFERRED UPGMA AND N-J TREES



FIVE STEPS IN BUILDING A PHYLOGENETIC ANALYSIS

- Finding all homologs
- Multiple-sequence alignment
- Building a tree
- Statistical assessment of a tree
- Viewing a tree and drawing conclusions

STATISTICAL ASSESMENT OF A TREE

- Tests of one overall hypothesis (tree) against other hypotheses
  - Wilson's "winning sites" test
  - Templeton's test
  - Kishino-Hasegawa ML test
- Tests of strength of support for lineages within trees
  - Bootstrap
  - Jack-knife
  - Decay index

BOOTSTRAPING - THE MOST FREQUENTLY USED STATISTICAL TEST FOR A TREE ASSESSMENT

1. Random sampling of columns in the original alignment to create a new alignment
2. Building a tree based on the new alignment
3. Repeat step 1 and 2 many times (usually 1000 times)
4. Calculate how many times a given topology appears in all replicas

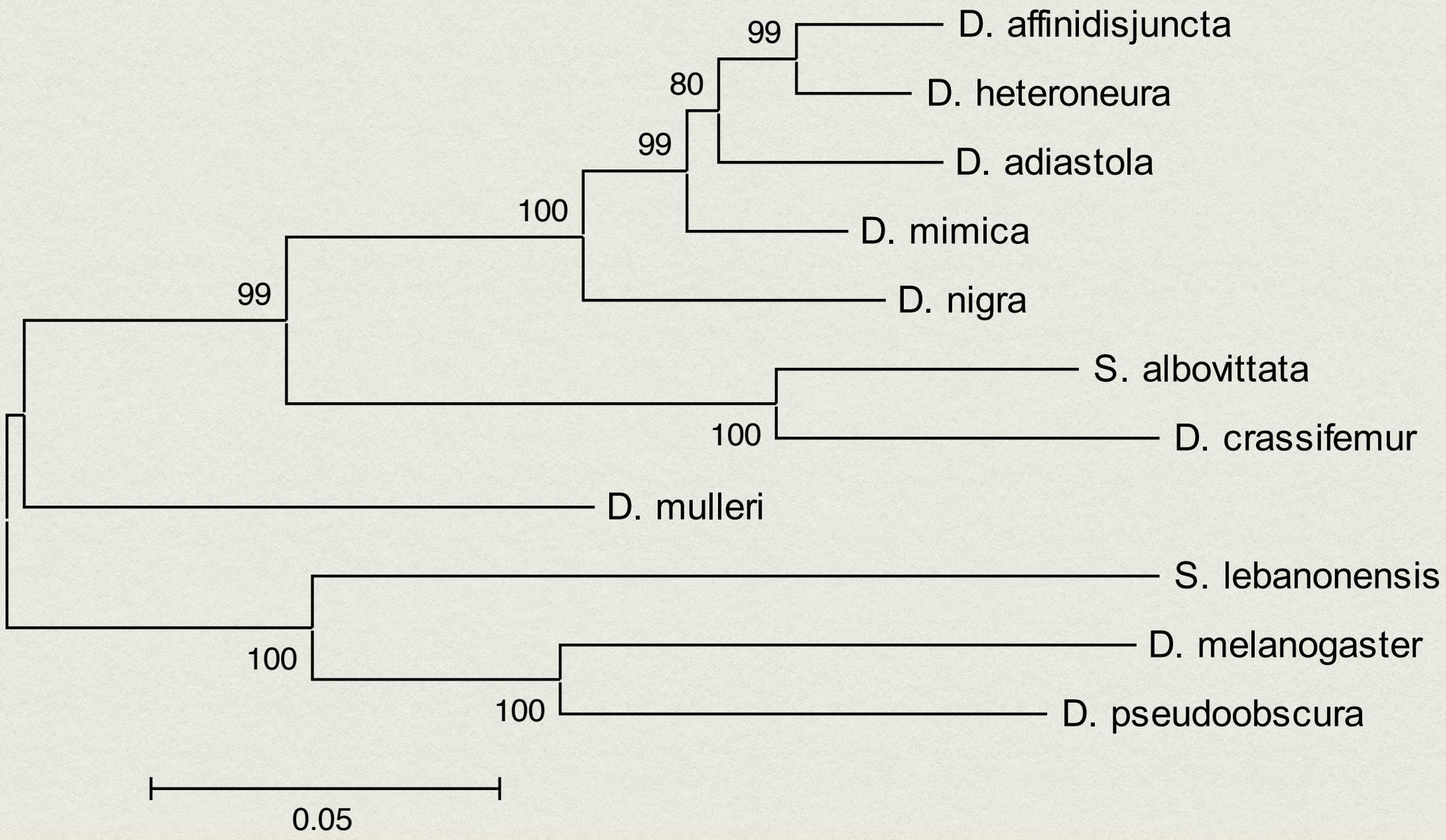
```
ATGGCTATTCTTATAGTACG
ATCGCTAGTCTTATATTACA
TTCACTAGACCTGTGGTCCA
TTGACCAGACCTGTGGTCCG
TTGACCAGTTCTCTAGTTCG
```

original alignment

```
AGGGGCTAATTCTATAGTAC
ACGGGCTAAGTCTATATTAC
TCAAACTAAGACCGTGGTCC
TGAAACCAAGACCGTGGTCC
TGAAACCAAGTTCCTAGTTC
```

resampled alignment

BOOTSTRAPING - THE MOST FREQUENTLY USED STATISTICAL TEST FOR A TREE ASSESSMENT



COMPARISON OF TREE BUILDING METHODS

Distance based	Maximum parsimony	Maximum likelihood
Uses only pairwise distances	Uses only shared derived characters	Uses all data
Minimizes distance between nearest neighbors	Minimizes total distance	Maximizes tree likelihood given specific parameter values
Very fast	Slow	Very slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree	Best option when tractable (<30 taxa, homoplasy rare)	Good for very small data sets and for testing trees built

DIFFICULTIES WITH PHYLOGENETIC ANALYSIS

- Horizontal or lateral transfer of genetic material (for instance through viruses) makes it difficult to determine phylogenetic origin of some evolutionary events
- Genes selective pressure can be rapidly evolving, masking earlier changes that had occurred phylogenetically two sites within comparative sequences may be evolving at different rates
- Rearrangements of genetic material can lead to false conclusions
- Duplicated genes can evolve along separate pathways, leading to different functions

WHICH PROCEDURE SHOULD WE USE?

- All that we can
- Each method has its own strengths
- Use multiple methods for cross-validation
- In some cases, none of the method gives the correct phylogeny

MORE ADVISE

- Selecting a high-quality input data set is the most critical step in developing a phylogeny
- The order of the input set can affect results. Good phylogenetics software provides tools for randomizing input sets
- Check for consistency by applying more than one method (NJ, MP, ML) to the same data set
- If you obtain an unreliable tree
- GET MORE DATA

SELECTED SOFTWARE

- Kumar S, Stecher G, and Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets *Molecular Biology and Evolution* **33**:1870-1874
  - http://www.megasoftware.net/
- Yang, Z. (1998) PAML: Phylogenetic Analysis using Maximum Likelihood.
  - http://abacus.gene.ucl.ac.uk/software/paml.html
- PHYLIP (the PHYLogeny Inference Package)
  - http://evolution.genetics.washington.edu/phylip/phylipweb.html

BIOINFORMATICS CREED

- Do not trust the data
- Use statistics
- Know the limits
- Remember about biology!!!