

How to read and make phylogenetic trees Zuzana Starostová

How to make phylogenetic trees? Workflow:



- quality check
- sequence alignment
- calculating genetic distances (optional)
- phylogeny estimation topology and branch length
- reliability test (bootstrap)
- tree visualization

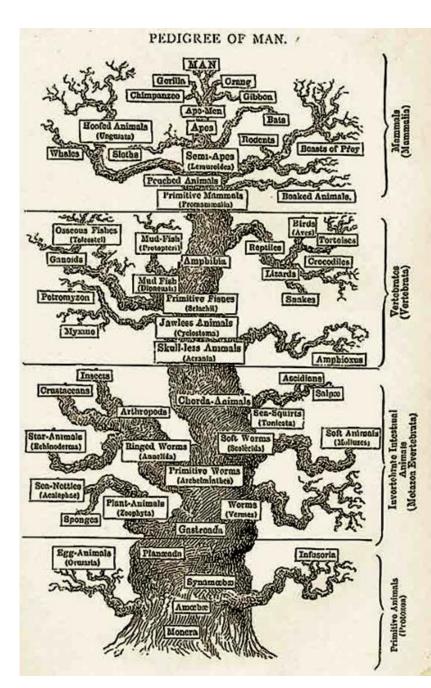
• the process of evolution produces a pattern of relationships between species - as lineages evolve and split and modifications are inherited, their evolutionary paths diverge

 this produces a branching pattern of evolutionary relationships –
 phylogenetic tree

• **phylogeny** – the evolutionary history of a species or group of related species

• phylogenies trace patterns of shared ancestry between lineages

• each lineage has a part of its history that is unique to it alone and parts that are shared with other lineages



Tree of life

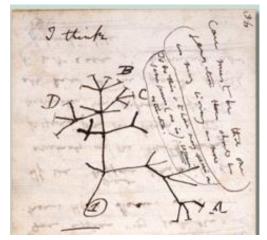


Charles Darwin species share common ancestor relationships among species compared to "the great Tree of Life"

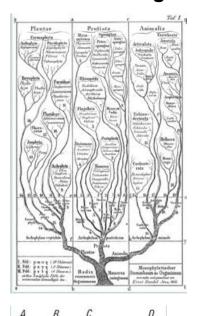
Ernst Haeckel – 1866

based on general similarity

diagram of relationships among species

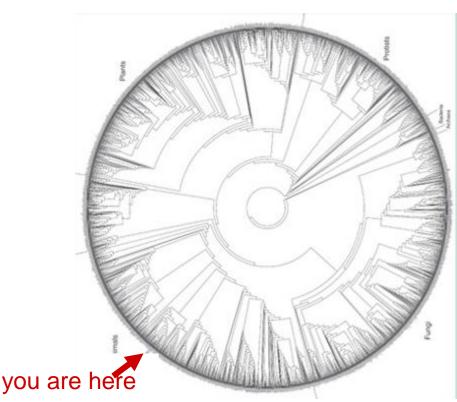


Darwin's diary, 1837



Willi Hennig – 1960s foundation of modern phylogenetics





http://www.peabody.yale.edu/exhibits/treeoflife/index.html

Why is phylogeny important?

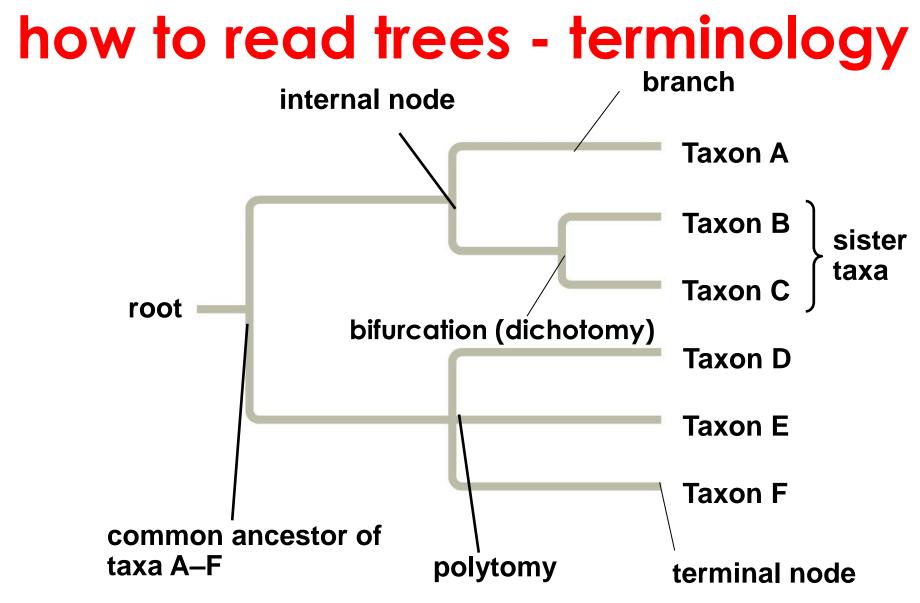
- understanding and classifying the diversity of life on Earth

We use phylogenetic trees for:

- comparative analysis and character evolution
- biogeography
- dating age of different taxa
- genetic ingeneering
- disease epidemiology
- conservation

•••

The Tree of Life Web Project (ToL) is a collaborative effort of biologists and nature enthusiasts from around the world. On more than 10,000 World Wide Web pages, the project provides information about biodiversity, the characteristics of different groups of organisms, and their evolutionary history. http://tolweb.org/tree/



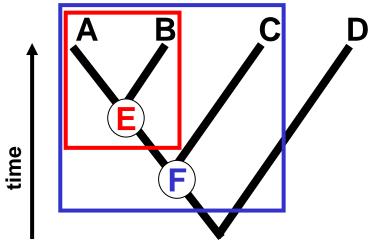
•root of the tree represents the ancestral lineage, and the tips of the branches represent the descendants of that ancestor

• as one moves from the root to the tips = moving forward in time

phylogenetic tree

branching diagram showing relationships between taxa based on their shared common ancestors

taxa (e.g.: species):



A and B are more closely related because they share a common ancestor (here "E") that C and D do not share

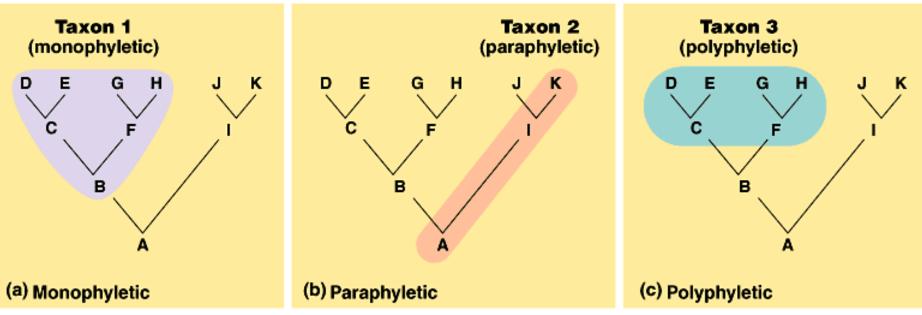
A+B+C are more closely related to each other than to D because they share a common ancestor ("F") that D does not share

Phylogeny and classification

phylogenetic (cladistic) classification reflects evolutionary history

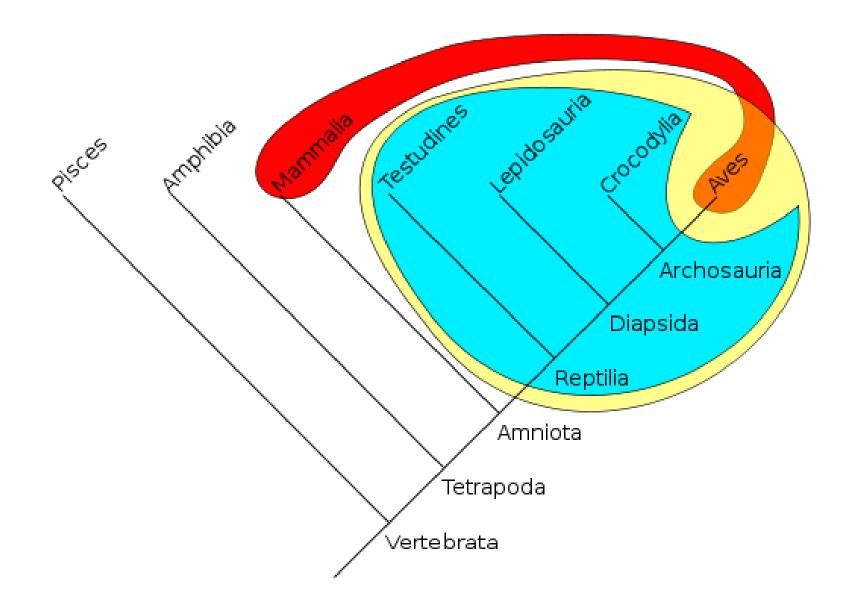
the only valid group for classification is **monophyletic** group = clade = group that includes a common ancestor and all the descendants (living and extinct) of that

ancestor.



Copyright @ Pearson Education, Inc., publishing as Benjamin Cummings.





-

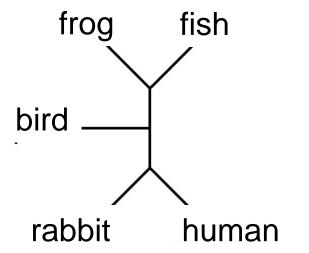
rooted vs. unrooted phylogenetic trees

rooted tree – root leads to the common ancestor of all studied taxa (e.g.: species) rooting the tree = indicating the direction of the evolutionary process helps to determine what is plesiomorphic and apomorphic

How to root a tree?

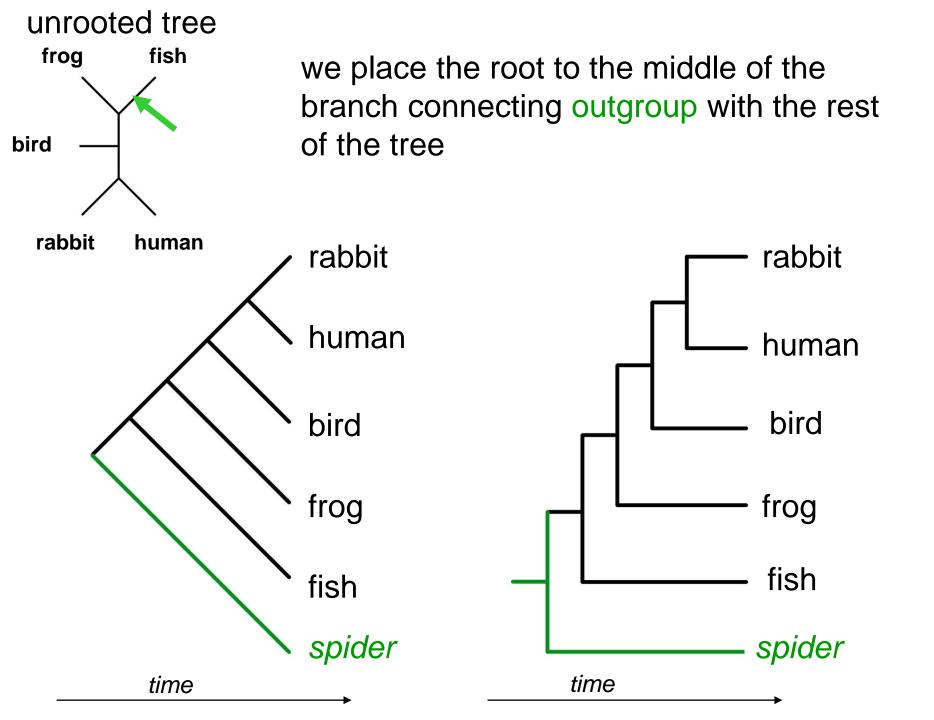
- introducing outgroup

outgroup – a species or group of species that is clearly distant from all the species of interest (ingroup), but still closely related

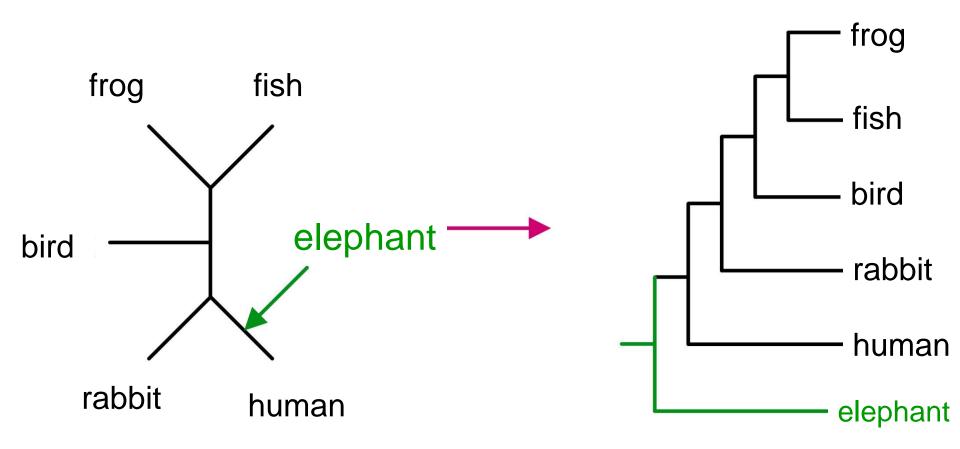


here we will use an invertebrate species to root the tree

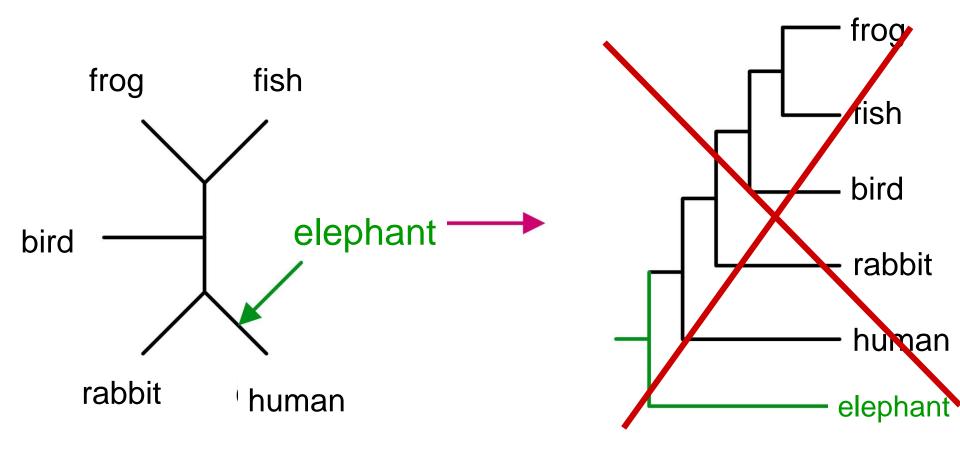
(since these species are all Craniata it would be better to use Urochordata (tunicates) to root the tree)



But what if our outgroup is wrong?



But what if our outgroup is wrong?



rooted vs. unrooted phylogenetic trees

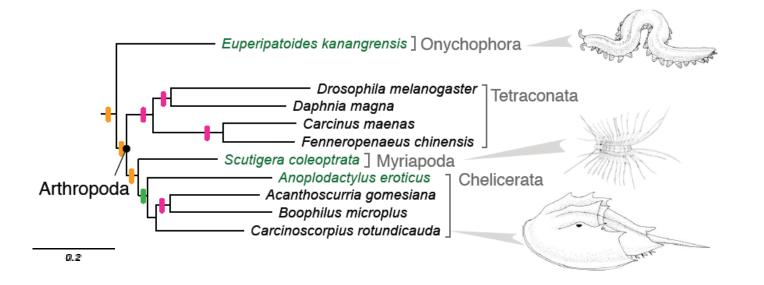
(2n – 3)!

 $2^{n-2}(n-2)!$ Number of bifurcating rooted trees for n taxa (OTUs) OTU = operational taxonomical unit

Number of bifurcating trees is increasing with number of taxa. The number of unrooted trees for n taxa is equal for the number of rooted trees for (n-1) taxa.

No. of taxa	No. of unrooted trees	No. of rooted trees
3	1	3
4	3	15
5	15	105
10	2 027 025	34 459 425
11	34 459 425	654 729 075
12	654 729 075	13 749 310 575
50	1.00986 x10 ⁵⁷	2.75292x10 ⁷⁶

From Zima a kol. Genetické metody v zoologii



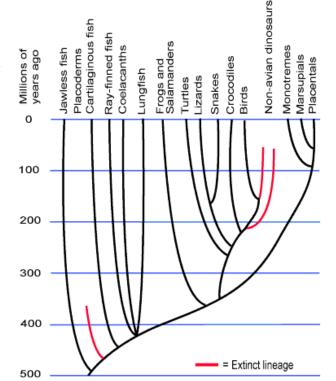
branch length = time of divergence from the common ancestor

molecular clock: sequence divergence increases over time linearly

when molecular clock holds (accumulation is linear over time) – all lineages in the tree have accumulated substitutions at the same rate

evolutionary rate dependent on metabolic rate, generation time, bottleneck events,...

we need calibration points (fossils, geological events)



How to make trees?

like family trees, phylogenetic trees represent patterns of ancestry

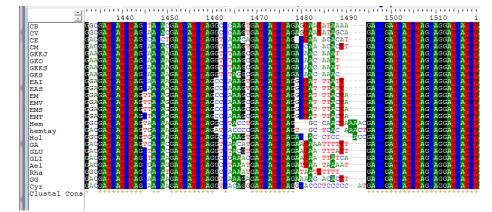
to reveal phylogenetic relationship, we have to compare characters which are **inherited from a common ancestor**

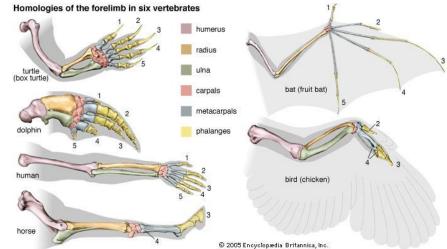
homologous characters (x analogous characters are result of convergent evolution)

character = heritable traits that can be compared across organisms

two or more forms = character state

types: physical characteristics (morphology), genetic sequences, and behavioral traits







Pros and cons of molecular characters (mainly sequence data)

- molecular data much more abundant (human genome 3,1 Gb, E. coli 4.6 Mb)
- independent (one position in the sequence of nucleotides x an eye the eye is missing, but it means that also cornea, retina, etc. are missing)
- easy to describe (A, C, T, G at the position 175 in the cytochrome b gene) x some structure on the bone more pointed
- can resolve relationships at all different levels of organization, from species and populations to phyla and kingdoms
- less subjective
- neutral number of shared characters mirrors phylogenetic relationship not just the same environmental selection pressures



Pros and cons of molecular characters (mainly sequence data)

- expensive
- technically demanding
- also non-experts on a given group can use them (lack of insight)
- no information about phenotype

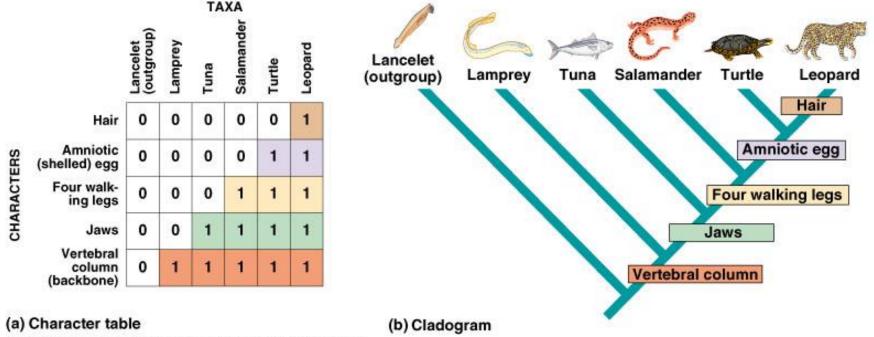
How to make a tree - morphological example

-selected species have shared primitive and derived characters -for reconstructing phylogeny derived shared characters = **synapomorphies** are important

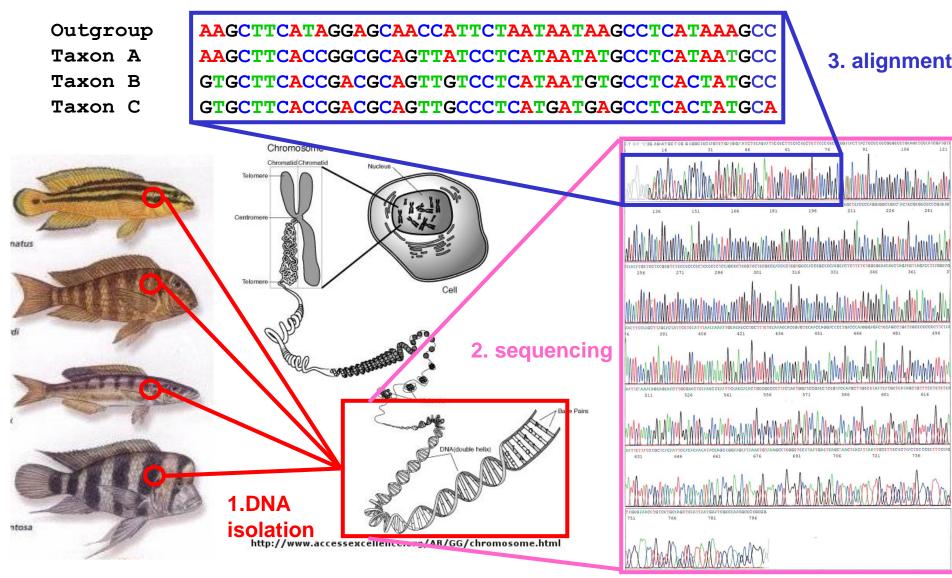
-create a character table (matrix) with variable characters -group the taxa based on synapomorphies - the more shared characters, the more closely related are the species

- for morphological data we use usually maximum parsimony method

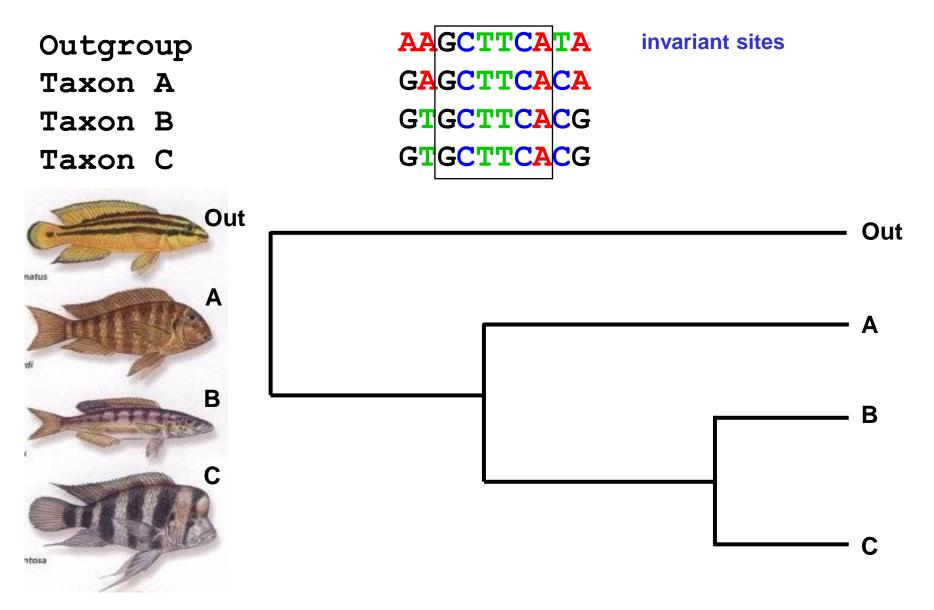
which prefers the simplest explanation of observed data



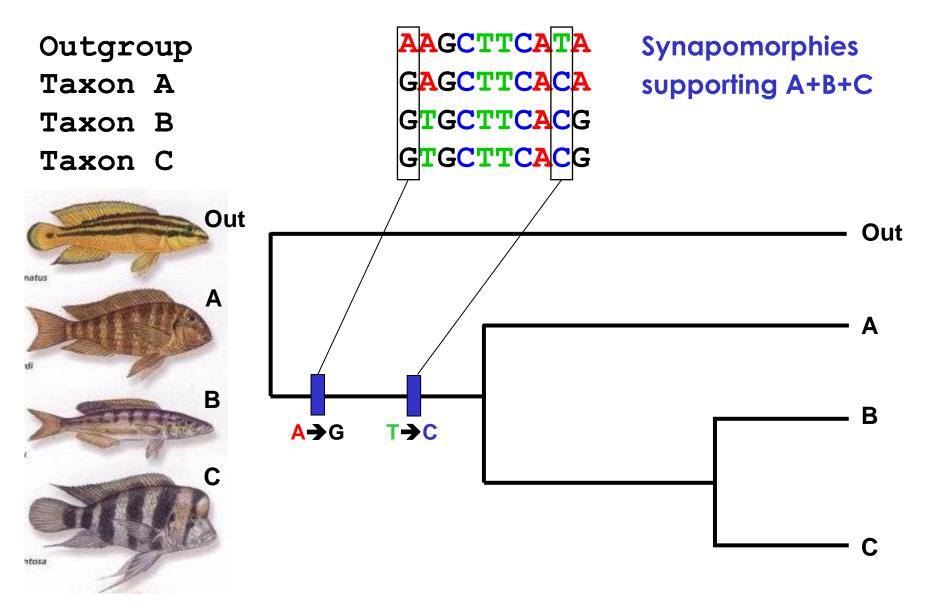
Convright @ Pearson Education, Inc., publishing as Reniamin Cummings.



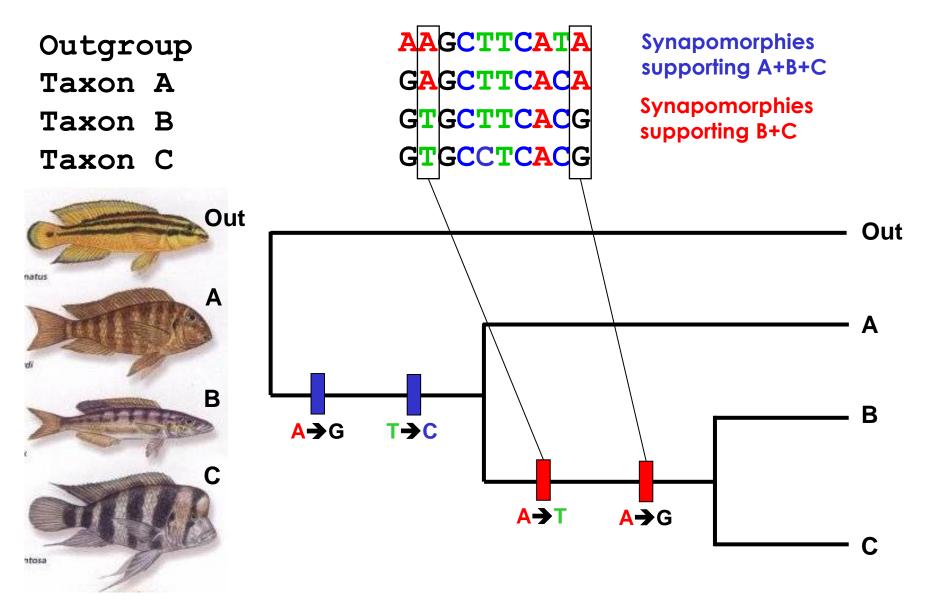
picture from R. Cox - Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html



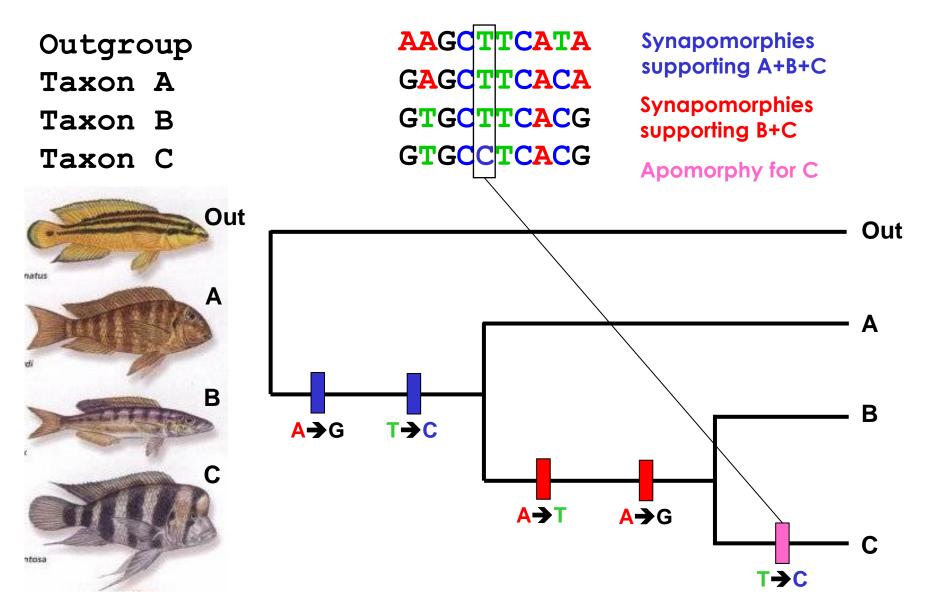
picture from R. Cox - Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html



picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html



picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html



picture from R. Cox – Phylogeny. http://www.dartmouth.edu/~robertcox/Teaching.html

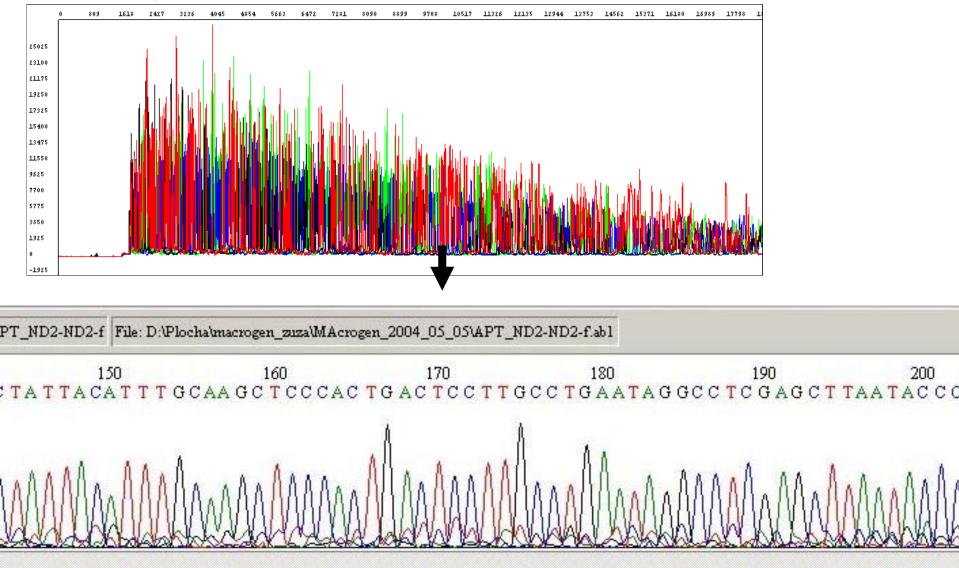
How to make phylogenetic trees? Workflow:



- quality check
- sequence alignment
- calculating genetic distances
- phylogeny estimation topology and branch length
- reliability test (bootstrap)
- tree visualization

Chromatogram files:

Can be opened in different software- Chromas, BioEdit, DNASTAR, ...





Checking sequences and chromatograms programs: Chromas, BioEdit, Geneious

Quality control:

Protein coding sequences: translation into amino acids

Checking for stop codons within the sequence of the gene (UAG, UAA, UGA)

Is start codon AUG (codes methionine (Met)) present?

Be aware of using correct translation table:

– nuclear vs. mitochondrial DNA, mitochondrial DNA (vertebrates, invertebrates, ...)

http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG2

LOCUS	AF248660 1036 bp DNA linear INV 14-NOV-2000	LOCUS	AFO
DEFINITION	Gryllus bimaculatus haplotype 2 cytochrome b gene, partial cds;	DEFINITION	Сус
	mitochondrial gene for mitochondrial product.		enc
ACCESSION	AF248660	ACCESSION	AFC
VERSION	AF248660.1 GI:11141460	VERSION	AFC
KEYWORDS	•	KEYWORDS	
SOURCE	mitochondrion Gryllus bimaculatus (two-spotted cricket)	SOURCE	mit
ORGANISM	Gryllus bimaculatus	ORGANISM	Сус
	Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;		Euk
	Neoptera; Orthopteroidea; Orthoptera; Ensifera; Grylloidea;		Lep
	Gryllidae; Gryllinae; Gryllus.	REFERENCE	1
REFERENCE	1 (bases 1 to 1036)	AUTHORS	Pet
AUTHORS	Huang,Y., Orti,G., Sutherlin,M., Duhachek,A. and Zera,A.	TITLE	AB
TITLE	Phylogenetic relationships of north American field crickets		chu
	inferred from mitochondrial DNA data	JOURNAL	Evo
JOURNAL	Mol. Phylogenet. Evol. 17 (1), 48-57 (2000)	REFERENCE	2
PUBMED	<u>11020304</u>	AUTHORS	Pet
REFERENCE	2 (bases 1 to 1036)	TITLE	Dir
AUTHORS	Huang,Y., Sutherlin,M. and Duhachek,A.	JOURNAL	Sub
TITLE	Direct Submission	OCONTRAL	Was
JOURNAL	Submitted (24-MAR-2000) School of Biological Sciences, University	FEATURES	was
	of Nebraska, 314 Manter Hall, Lincoln, NE 68588, USA		
FEATURES	Location/Qualifiers	source	
source	11036		
	/organism="Gryllus bimaculatus"		
	/organelle="mitochondrion"		
	/mol_type="genomic DNA"		
	/db_xref="taxon: <u>6999</u> "		
	/haplotype="2"	<u>e ne</u>	
	/country="Cyprus"		
CDS	<1>1036	CDS	
	/codon_start=2		
	/transl_table=5_		
	/product="cytochrome b"		
	/protein id="AAG31952.1"		
	/db_xref="GI:11141461"		
	/translation="FGSLLGLCLVIQIMTGIFLAMHYTADINMAFNSVSHICRDVNFG		
	WLLRTMHANGASFFFICLYTHVGRGMYYGSYNLMFTWMTGTLILFLVMAAAFMGYVLP		
	MCOMSENCE TUTTINE SET DYLOTDE MOMBILC FRUDWETLTD FFT FURNT DETURES F		

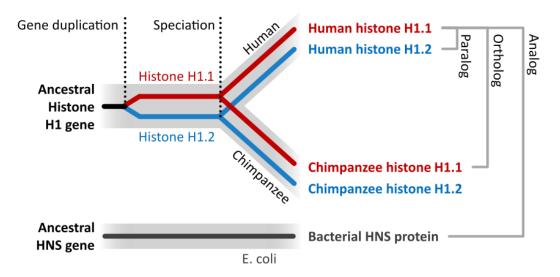
LOCUS	AF020255 902 bp DNA linear VRT 30-SEP-1997
DEFINITION	Cyclura nubila cytochrome b (cytb) gene, mitochondrial gene
	encoding mitochondrial protein, partial cds.
ACCESSION	AF020255
VERSION	AF020255.1 GI:2444366
KEYWORDS	•
SOURCE	mitochondrion Cyclura nubila
ORGANISM	Cyclura nubila
	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
	Lepidosauria; Squamata; Iguania; Iguanidae; Iguaninae; Cyclura.
REFERENCE	1 (bases 1 to 902)
AUTHORS	Petren, K. and Case, T.J.
TITLE	A Phylogenetic analysis of body size evolution and biogeography in
	chuckwallas (Sauromalus) and other iguanines
JOURNAL	
REFERENCE	2 (bases 1 to 902)
AUTHORS TITLE	Petren,K. and Case,T.J. Direct Submission
JOURNAL	Submitted (20-AUG-1997) EEB, Princeton University, Guyot Hall,
OOOKNAL	Washington Rd., Princeton, NJ 08544-1003, USA
FEATURES	Location/Qualifiers
source	
000200	/organism="Cvclura nubila"
	/organelle="mitochondrion"
	/mol type="genomic DNA"
	/db xref="taxon:51214"
	/note="specimen from Cuba, (pop. #27)"
<u>e</u> ene	<1>902
	/gene="cytb"
CDS	<1>902
	/gene="cytb"
	/codon_start=2
	/transl_table=2
	/product="cytochrome b"
	/protein_id="AAB71582.1"
	/db_xref="GI:2444367"
	/translation="FIDLPTPSNISSWWNFGSLLGLCLIIQILTGLFLAMHYTANISH

homology of molecular characters

it is necessary to compare sequences (of e.g. genes) which are **orthologous =** inferred to be descended from the same ancestral sequence separated by a speciation event X

xenolog – sequence of gene incorporated from other species by horizontal transfer

pseudogene – sequence copied from the mitochondrial genome to the nuclear DNA



Gene phylogeny as red and blue branches within grey species phylogeny. Top: An ancestral gene duplication produces two **paralogs** (histone H1.1 and 1.2). A speciation event produces **orthologs** in the two daughter species (human and chimpanzee). Bottom: in a separate species (E. coli), a gene has a similar function (histone-like nucleoid-structuring protein) but has a separate evolutionary origin and so is an **analog**.

source: Wikipedia

Additional sources of DNA sequences:

- public databases 🛯



NCBI Resources 🕑 How To 🕑			
NCBI Nucleotide	•		Search
ional Center for technology Information	Search		
NCBI Home	Welcome to NCBI	Popular Resources	
Site Map (A-Z)		BLAST	
All Resources	The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.	Bookshelf	
Chemicals & Bioassays		Gene	
Data & Software	About the NCBI Mission Organization Research RSS Feeds	Genome	
DNA & RNA		Nucleotide	
Domains & Structures	Get Started	OMIM	
Genes & Expression	 Tools: Analyze data using NCBI software Downloads: Get NCBI data or software 	Protein	
Genetics & Medicine	 How-To's: Learn how to accomplish specific tasks at NCBI 	PubChem	
Genomes & Maps	Submissions: Submit data to GenBank or other NCBI databases	PubMed	
		PubMed Central	
Homology		SNP	
Literature	Education Resources		
Proteins	Education Resources		
Sequence Analysis	Central point of access for help	NCBI News	
Taxonomy	documents, teaching materials, news	New NCBI Newsletter 01 Dec 2011	
Training & Tutorials	outlets, and other educational resources.	Information on the new Genome Site, a new 16S BLAST database, updates to Sequin,	
Variation	II 1 2 3 4 5 6 7	NCEI will continue to operate SEA	

More

Subsequent to an announcement in February 2011 that NCBI was planning to phase out the

http://www.ncbi.nlm.nih.gov/

BLAST - Basic Local Alignment Search Tool

- algorithm for searching in databases for similar sequences

BLAST [®] Home Recent Results	Basic Local Alignment Search Tool	My NCBI 2 [Sign In] [Register]
NCBI/ BLAST Home BLAST finds regions of si	milarity between biological sequences. <u>more</u>	Your Recent Results New!
	New DELTA-BLAST, a more sensitive protein-protein search	🖹 All Recent results
BLAST Assembled R	tefSeq Genomes	News

How to make phylogenetic trees? Workflow:



- obtain DNA sequence
- oquality check
- sequence alignment
- calculating genetic distances
- phylogeny estimation topology and branch length
- reliability test (bootstrap)
- tree visualization

Where sequences differ and where are the same?

Mode: Select / Slide	 Selection: 0 Position: 7: Enotaeus 104 	Sequence Mask: None Numbering Mask: None	Start ruler at: 1
e i d I D	: 😚 c/0 -l- 🗠 🎇 🏬 👫 📗 👬 🗰 💼	🖬 💱 💱 👬 💱 👬	speed slow 🗛 🚽 fast
-	10 20	30 40 50	60 70
	CTTCGGATCAATGCTGCTCGCCTGCTCA	GCCCTACAAGTACTAACCGGGGTTCTTTCT	AGCCGTACACTACACAGCA
CM2M5MAURUS	CTTCGGATCAATGCTGTTCGCCTGCTCA	.G <mark>CCCTACAA</mark> GT <mark>ACTAACC</mark> GGGTTCTTTCT	AGCCGTACACTACACAGCA
zv3m5venezue	CTTCGGATCAATGCTGCTCGCCTGCTCA	GCCCTACAAGTACTAACCGGGTTCTTTCT	GGCCG <mark>H</mark> ACAC <mark>H</mark> ACACAGCA
		GCCCTACAAGTACTAACCGGGTTCTTCCT	
CA4M5ALVARE2		.ACCC <mark>T</mark> ACAAG <mark>T</mark> AC <mark>T</mark> AAC <mark>A</mark> GG <mark>CTTCTT</mark> CCT	
Emurinus	CTTCGGATCTATGCTGCTTACCTGCTCA	ACCCTACAAGTACTAACTGGGTTTTTCCT	AGCAGTTCACTACACAGC
Enotaeus		ACTCTACAAGTACTAACAGGCTTTTTTC1	
inornatus		.G <mark>CCCTACAAGTAC</mark> TAACAGG <mark>CTTCTTCC</mark> T	
inorn-MUJ		GCCCTACAAGTACTAACAGGCTTCTTCCT	
nonensis	TTTCGGATCTATACTGACCTGCTCA	GTCCTACAAGTACTAACAGGCTTCTTTCT	AGCCGTACATTACACAGC
subflavens	TTTTGGATCTATGCTACTCGCCTGCTCA	ACCCTACAAGTACTAACAGGATTCTTCTT	AGCCGTACATTACACAGC
JH1m5		ATACTACAAGTATTAACAGGCTTTTTCCCT	
713M5	CTTCGGATCCATACTACTCACTTGCTCA	ATACTACAAGTATTAACAGGCTTTTTCCT	GGCTATACACTACACAGC
93m5	CTTCGGATCCATACTACTCACTTGCTCA	ATACTACAAGTATTAACAGGCTTTTTCCT	GGCTATACACTACACAGC
34m5		ATACTACAAGTATTAACAGGCTTTTTCCT	GGCTATACACTACACAGC
42M5	CTTCGGATCCATACTACTCACTTGCTCA	ATTCTACAAGTATTAACAGGCTTTTTCCT	GGCCATACACTATACAGC
41m5	CTTCGGATCCATACTACTCACTTGCTCA	ATTCTACAAGTATTAACAGGCTTTTTCCT	GGCCATACACTATACAGC
0M5	CTTCGGATCCATACTACTCACTTGCTCA	ATACTACAAGTATTAACAGGCTTTTTCCT	GGCCATACACTACACAGC
3M5	CTTCGGATCCATACTACTCACTTGCTCA	ATACTACAAGTATTAACAGGCTTTTTCCT	GGCCATACACTACACAGC
Hm 5	CTTCGGATCCATACTACTCACTTGCTCA	ATACTACAAGTATTAACAGGCTTTTTCCT	GGCCATACACTACACAG
14M5	CTTCCCATCCATACTACTACTTCCTCA	ATCCTACAAGTATTAACAGGCTTCTTCCT	AGCTATACACTACACAGC
411M5	CTTCGGATCCATACTACTTACTTGCTCA	ATGCTACAAGTATTAACAGGCTTCTTCCT	AGCTATACACTACACAGC
46m5		ATGCTACAAGTATTAACAGGCTTCTTCCT	
842M5	CTTCCCATACTACTACTACTACTCC	ATCCTACAAGTATTAACAGGCTTCTTCCT	ACCTATACACTACACACC
ex1m5	CTTTCCCATACTACTTACTTACTTCCCCCA	ATGCTACAAGTATTAACAGGCTTCTTCCT	AGGMAMACAGMACAGAG
712m5		ATGCTACAAGTATTAACAGGCTTCTTCCT	
73m5		ATCCTACAAGTATTAACAGGCTTCTTCCT	
1.9m5	CTTTCCCATACTACTTACTTACTTCCCCCA	ATGCTACAAGTATTAACAGGCTTCTTCCT	AGGMAMACAGMACAGAG
77m5	CTTCCATCCATACTACTTACTTCCTCA	ATGCTACAAGTATTAACAGGCTTCTTCCT	ACCUATACACO
/m5 8M5		ATGCTACAAGTATTAACAGGCTTCTTCCT	
.0m5	C T T C C A T C C A T A C T A C T T A C T T C C T C A	ATCCTACAACTATTAACACCCTTCTTCCT	A GC MAMAGAC MAGACAC
L7m5 543m5		ATGCTGCAAGTATTAACAGGCTTCTTCCT	
		ATACTACAACTA TTAACAGCC TTC TTCCT	
UH0m5		A A CARCARO A A CAGO TICTICO	

Alignment

- a way of arranging the sequences of DNA (also e.g. amino acid in the protein sequence) to identify regions of similarity
- start of every phylogenetic analysis
- assessing of position homology of each base in the sequence
- each position (column in the alignment) in the sequence represents character potentially useful for the phylogenetic analysis
- different programs for calculating and editing alignments
 - manual: BioEdit, Macaw
 - automatic different algorithms

Clustal X, PileUp, Multalin, Mafft – often online

Mode: Select / Slide	-	Selection: 0 Position: 7: Enotaeus 104		equence Mask: umbering Mask:			Start ruler at	1	
e i d I d	<u>D</u> 😚 G/D - -		👬 💼 👬 🖬 🗛		🚯 мі 🎹	😵 sp	eed slow 🕁 .	🚽 fast	
-		io 20	30	40	ŚO	6	0	70	
cenchria CM2M5MAUBUS		CAATGCTGCTCGCCTG CAATGCTGTTCGCCTG							
cv3m5venezue		CAATGCTGCTCGCCTG							
CE1M5cenchri		CAATACTGCTCGCCTG							
CA4M5ALVARE2		CAATACTACTACCCTC							
Emurinus		CTATGCTGCTTACCTG							
Enotaeus		CCATACTAATTGCCTG							
inornatus	CTTTCCAT	CAATACTACTCACTTG	CTCAGCCCTACA	GTACTAAC	GGCTTCTT	CTACC	TGTACACT	ATACAGC	A A
inorn-MUJ		CAATACTACTCACCTG							
monensis		CTATACTACTGACCTG							
subflavens	mmmm CCA	CTATGCTACTCGCCTG	CTCAACCCTACA	CTACTA AC	CCATTOT	mmAcc	COMACATT	ACACACC	A A
UH1m5	Cmmc ccAm	CCATACTACTCACTTC	CHCAAHACHACA	CmAmmAAC2	CCOmmmm	Cmac	TATACACT	ACACACC	CA.
F13M5		CCATACTACTCACTTC							
93m5		CCATACTACTCACTTG							
84m5		CCATACTACTCACTTC							
M2M5		CCATACTACTCACTTG							
M1m5	CTTCGGAT	CCATACTACTCACTTG	CTCAATTCTACA	GTATTAAC	GGCTTTTT	CTGGC	CATACACT	ATACAGO	CA.
DM5		CCATACTACTCACTTG							
GM5		CCATACTACTCACTTG							
Hm5		CCATACTACTCACTTG							
U4M5		CCATACTACTTACTT							
M11M5		CCATACTACTTACTT							
M6m5	CTTCGGAT	CCATACTACTTACTTG	CTCAATGCTACAA	GTATTAACA	GGCTTCTT	CTAGC	TATACAC	ACACAGC	TA T
E42M5									
ex1m5	CTTCGGAT	CCATACTACTTACTTG CCATACTACTTACTTG	CTCAATGCTACAA	GTATTAACA	GGCTTCTT	CTAGO	TATACAC	ACACAGC	
F12m5	CTTCGGAT	CCATACTACTTACTTG	CTCAATGCTACA	GTATTAAC	GGCTTCTT	CTAGC	TATACACI	ACACAGC	TA.
F3m5	CTTCGGAT	CCATACTACTTACTTG	CTCAATGCTACA/	GTATTAAC	GGCTTCTTC	CTAGC	TATACACI	ACACAGC	TA
1 9m5	CTTCGGAT	CCATACTACTTACTTG CCATACTACTTACTTG	CTCAATGCTACAA	GTATTAACA	GGCTTCTT	CTAGC	TATACACI	ACACAGC	TA.
77m5	CTTCGGAT	CCATACTACTTACTTC	CTCAATGCTACAA	GTATTAACA	GGCTTCTTC	CTAGC	TATACACI	ACACAGC	
18M5	CTTCGGAT	CCATACTACTTACTTG	CTCAATGCTACAA	GTATTAACA	GGCTTCTT	CTAGC	TATACACI	ACACAGC	πА.
17m5	CTTCGGAT	CCATACTACTTACTTC	CTCAATGCTACAA	GTATTAACA	GGCTTCTTC	CTAGC	TATACACI	ACACAGC	ΠA
E43m5		CCATACTACTTACTTG							
UHOm5		CCATACTACTTACTTG							
UH0m5		<u>CCATACTACTTACTT</u> G	CTCAATACTACA	GTATTAACA		CTAGC	TATACACI	ACACAGO	

Alignment - pairwise alignment (two sequences)

- multiple alignment (more sequences)

AATGCCCTAAA AATGCGGCTAAA AACGCGCTAAA ATGCTAA AATGCC-CTAAA AATGCGGGCTAAA AACGCG-CTAAA -ATG---CTAA-

gaps are inserted between the bases so that identical or similar characters are aligned in successive columns

during alignment: algorithm is searching for the best position for each base and **receives points**:

Plus points for every match, Minus points for inserting or extension of the gap,

The goal is to obtain the best score

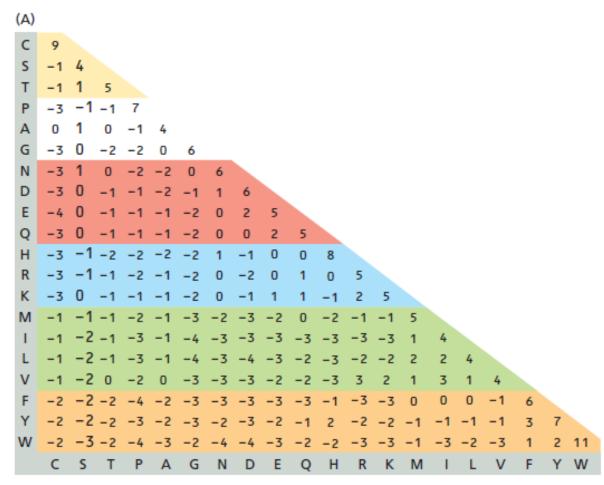


Figure 4.4

Amino acid substitution scoring matrices. (A) The BLOSUM-62 matrix and (B) the PAM120 substitution matrix. Each cell represents the score given to a residue paired with another residue (row × column). The values are given in half-bits, as discussed in Section 5.1. The colored shading indicates different physicochemical properties of the residues (see Figure 2.3): small and polar, yellow; small and nonpolar, white; polar or acidic, red; basic, blue; large and hydrophobic, green; aromatic, orange.

scoring matrices

Alignment:



Clustal W (Clustal X) www.clustal.org – frequently used software (Thompson et al. 1994 a 1997)

ClustalX (1.81)	
ile Edit Alignment Trees Colors Quality Help	
Multiple Alignment Parameters	
 CLOSE Multiple Parameters Gap Opening [0-100] : 10.00 Gap Extention [0-100] : 0.20 Delay Divergent Sequences (%) : 30 DNA Transition Weight [0-1] : 0.50 Use Negative Matrix OFF • Protein Weight Matrix BLOSUM series ○ PAM series Gonnet series ○ Identity matrix User defined Load protein matrix: NA Weight Matrix © IUB ○ CLUSTALW(1.6) ○ User defined 	

What can help:

• first use default parameters, later increase penalty for gap opening and decrease penalty for gap extension

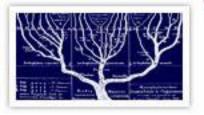
• coding genes align as amino acids

• use information about sequence secondary structure (genes for 12S and 16S RNA) – database with alignments with secondary structure in mind (http://www.arb-silva.de/)

• delete ambiguously aligned positions

	MOLECULAR PHYLOGENETICS AND EVOLUTION Vol. 2, No. 2, June, pp. 152–157, 1993
MOLECULAR PHYLOGENETICS AND EVOLUTION Vol. 4, No. 1, March, pp. 1–9, 1995	
	Alignment-Ambiguous Nucleotide Sites and the Exclusion of Systematic Data
Elision: A Method for Accommodating Multiple Molecular Sequence Alignments with Alignment-Ambiguous Sites	JOHN GATESY,* ROB DESALLE,† AND WARD WHEELER*
WARD C. WHEELER.* JOHN GATESY, T AND ROB DESALLE	

🌾	lioEdit S	Sequence A	lignment Ec	litor	-				_						-	-	
File	Edit	Sequence	Alignmer	nt View	Accesso	ry Applicati	on RNA	World W	/ide Web	Options	Window	Help					
	Ľ																
1	C:\Use	ers\Zuz\syno	cho-stary no	otas\synch	ro1\work				10-t=0,5	aln							
	3 📇	Courier Ne	w 🔻	11 💌	B	25 to	tal sequen	ces		shade	threshold 60	% 🔹 📩					
Мо	ode: Sel	ect / Slide 💽	•	Select Positio						lask: None lask: None			Start ruler at:	1			
	лт	DID	🔒 сл -		 I & ET		E 	GAT GAT	TAG A	•	MI 📅 🤇	Scrol		•			
	1		~					CATCAT :	n pre e re	· · · · ·			i slow 🕁 🖣				
		<u> </u>		1440		450	1460		1470	14		1490		500	151		1
CI CI				ATCTAG ATCTAG	CAACA	GATCAT GATCAT		CAAGT	GAT <mark>C</mark> A GAT <mark>C</mark> AT	CTAGAG			-G <mark>ACC</mark> G -G <mark>ACC</mark> G		TAGCAG		
CH	2					IGAT <mark>C</mark> AT GATCAT	TAGGT TAGGT		GAT <mark>C</mark> AT GATCAT				-G <mark>ACC</mark> G -G <mark>ACC</mark> G				CT.
GI	4 KKJ			A <mark>TC</mark> TAG ATCTAG	CAACA CAACA	GATCAT AGATCAT	TAGGT		GAT <mark>C</mark> AT GAT <mark>C</mark> AT		CAA-AC AAC-AA		GACCG GACCG		TAGCAG		CT
GI	KO 🛛		GAAG <mark>AT</mark> C	A <mark>TC</mark> TAG	CAACA	AGATCAT	TAGGT		GATCAT		AAC-AA		-G <mark>ACC</mark> G				ICT.
	KKS KS			A <mark>TCT</mark> AG ATCTAG		AGA <mark>TC</mark> AT AGATCAT	TAGGT TAGGT		GAT <mark>C</mark> AT GATCAT	C AGAC	AAC-AA AAC-AA		GACCG GACCG				
EZ			GGAGA <mark>TIC</mark>	A <mark>TC</mark> TAG	CAACA	AGA <mark>TCA</mark> T	TAGCC		GA <mark>TC</mark> AT		TAT-TT2	_	-G <mark>ACC</mark> G	A <mark>TCA</mark> T(TAGCAG	GA <mark>TC</mark> AT	ICT.
EZ				ATCTAG ATCTAG		AGA <mark>TCAT</mark> EGATCAT		CAAAGT CAAAGC	GAT <mark>C</mark> AT GATCAT		TAT-TT TAT-TT						S I
EN				ATCTAG		GATCAT GATCAT	TAGCC		GATCA GATCAT			ACTA ACTA	GACCG				CT
EN				A <mark>TC</mark> TAG	TT <mark>ACA</mark> (G <mark>atca</mark> t	TAGCC	CAAAGC	GA <mark>TC</mark> AT	CTAGGC	TAT-TT2	ACTA	G <mark>ACC</mark> G		TAGCAG	GA <mark>TC</mark> AT	ICT.
EN				ATCTAG ATCTAG	CTACA/ TAACA/	AGATCAT AGATCAT			GAT <mark>C</mark> AT GATCAT		-GC-CA		GACCG GACCG				
	em emtav	,		ATCTAG		AGATCAT	TAGAT		GATCAT	CTAGT-							T o
Ho	ol ¹			A <mark>TC</mark> TAG		AGA <mark>TC</mark> AT	TAGGC		GA <mark>TC</mark> A				CGA <mark>CC</mark> G		TAGCAG		CT
GI				ATCTAG ATCTAG		GATCAT GATCAT	TAGGT TAGGT		GAT <mark>C</mark> AT GATCAT	C AGAA	TAAATT: TAAATT:		-G <mark>ACC</mark> G -G <mark>ACC</mark> G		TAGCAG		
GI				A <mark>TC</mark> TAG	CTACAZ	AGATCAT	TAGGC	CGAAAC	GA <mark>TC</mark> AT	CTAGAA	TAA-TT	TCA	-G <mark>ACC</mark> G	A <mark>TCAT</mark> C	TAGCAG	GA <mark>TC</mark> AT	CT
Ae				ATCTAG		GATCAT			GAT <mark>C</mark> AT			AAT	-GACC G				CT
Rł				ATCTAG ATCTAG		IGATCAT Agatcat		CAAAAT CAAACA			AAT <mark>T</mark> TT AAC-AC		-G <mark>ACC</mark> G -G <mark>ACC</mark> G		TAGCAG		
C	/r		GACGA <mark>IIC</mark>	A <mark>TC</mark> TAG		AGA <mark>TC</mark> AT	TAGGT	ACAGG			ACCCTC	ccccz			TAGCAG		CT
C]	lusta	l Cons	* ****	******	**	*****	****	*	****	e de de de de			*****	*****	*******	******	r de de l
-		►	<u> </u>														



Phylogeny estimation

character based (maximum parsimony, maximum likelihood, Bayesian analysis)

two types of methods

distance based (Neighbour-joining, UPGMA)

Two different approaches:

algorithm – number of specific steps resulting in one best tree methods: UPGMA, Neighbour-joining

optimality criterion – consider and compare all theoretically possible trees based on selected criteria- number of evolutionary steps, likelihood value