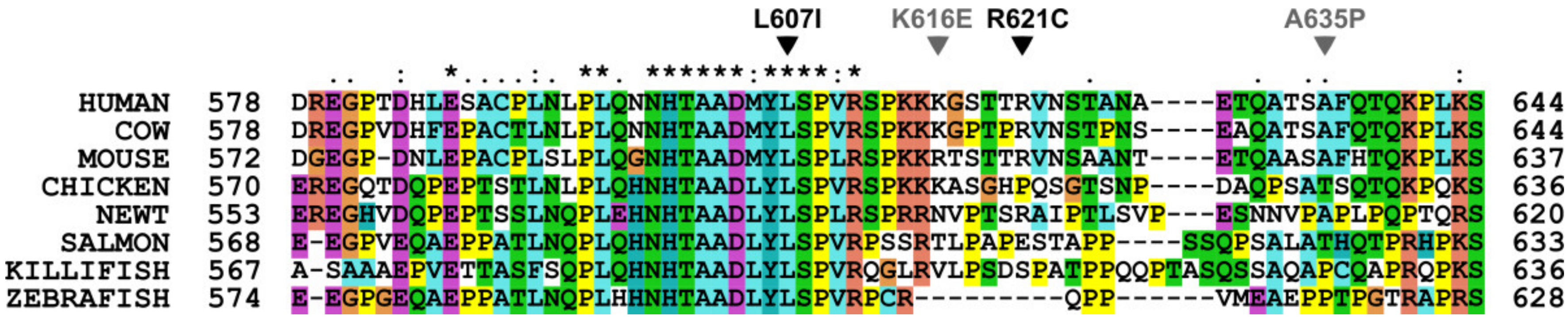
Πολλαπλή στοίχιση
Φυλογένεση

MSA: Τι είναι

- Στοίχιση για 3 ή περισσότερες ακολουθίες.
- Αποκαλύπτονται οι συντηρημένες περιοχές μεταξύ των ακολουθιών μιας οικογένειας.
- Χρειάζεται για:
  - Δημιουργία profiles/motifs που χαρακτηρίζουν μια επικράτεια (domain).
  - Ανίχνευση συντηρημένων DNA-binding sites σε προμότορες γονιδίων
  - Φυλογένεση.
  - Πρόβλεψη δευτεροταγούς και τριτοταγούς δομής πρωτεϊνών.
  - Σχεδιασμό εκφυλισμένων εκκινητών PCR

MSA


L607I
K616E
R621C
A635P
HUMAN 578 DREGPTDHLESACPLNLPLQNNHTAADMYLSPVRSPKKKGSTTRVNSTANA----ETQATSAFQTQKPLKS 644
COW 578 DREGPVDHFEPACTLNLPLQNNHTAADMYLSPVRSPKKKGPTPRVNSTPNS----EAQATSAFQTQKPLKS 644
MOUSE 572 DGEGP-DNLEPACPLSLPLQGNHTAADMYLSPLRSPKKRTSTTRVNSAANT----ETQAASAFHTQKPLKS 637
CHICKEN 570 EREGQTDQPEPTSTLNLPLQHNHTAADLYLSPVRSPKKKASGHPQSGTSNP----DAQPSATSQTQKPQKS 636
NEWT 553 EREGHVDQPEPTSSLNQPLEHNHTAADLYLSPLRSPRRNVPTSRAIPTLSVP---ESNNVPAPLPQPTQRS 620
SALMON 568 E-EGPVEQAEPPATLNQPLQHNHTAADLYLSPVRPSSRTLPAPESTAPP----SSQPSALATHQTPRHPKS 633
KILLIFISH 567 A-SAAAEPVETTASFSQPLQHNHTAADLYLSPVRQGLRVLPSDSPATPPQQPTASQSSAQAPCQAPRQPKS 636
ZEBRAFISH 574 E-EGPGEQAEPPATLNQPLHHNHTAADLYLSPVRPCR---------QPP------VMEAEPPTPGTRAPRS 628

MSA

- Sum of pairs
- Σκοπός: η μεγιστοποίηση αυτού του score

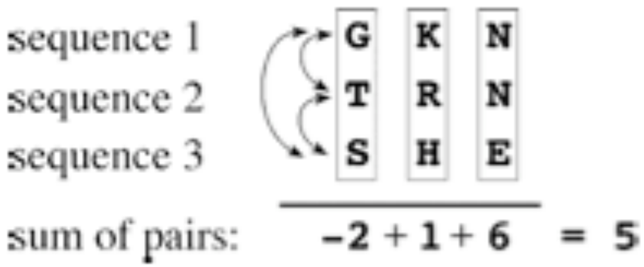


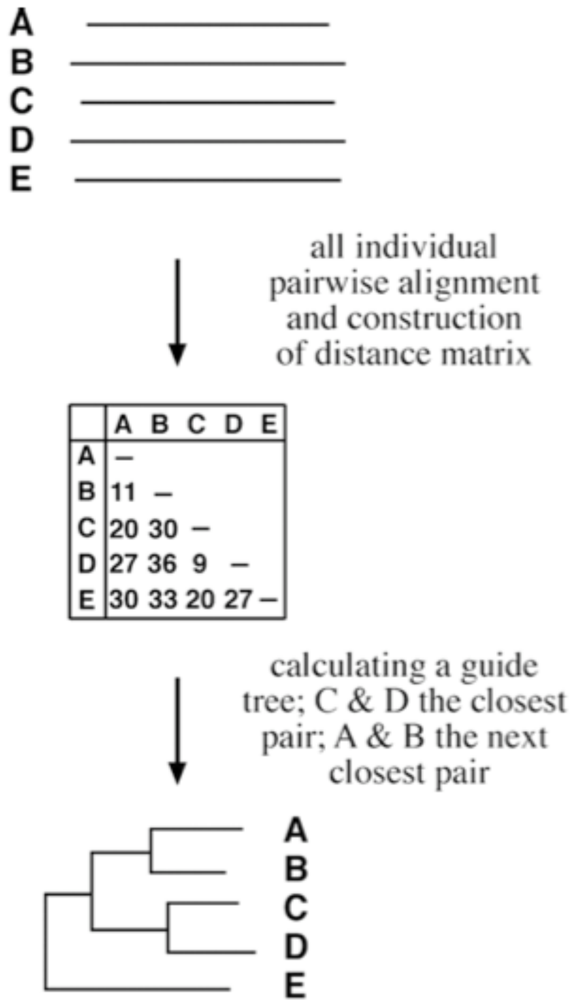
Figure 5.1: Given a multiple alignment of three sequences, the sum of scores is calculated as the sum of the similarity scores of every pair of sequences at each position. The scoring is based on the BLOSUM62 matrix (see Chapter 3). The total score for the alignment is 5, which means that the alignment is $2^5 = 32$ times more likely to occur among homologous sequences than by random chance.

MSA

- Πολλαπλή στοίχιση με:
  - Δυναμικό προγραμματισμό (dynamic programming).
  - Με ευρετικές μεθόδους (heuristics).
    - Προοδευτική στοίχιση (progressive alignment)
    - Στοίχιση με διαδοχικές βελτιώσεις (iterative alignment)
    - Στοίχιση βασισμένη σε blocks

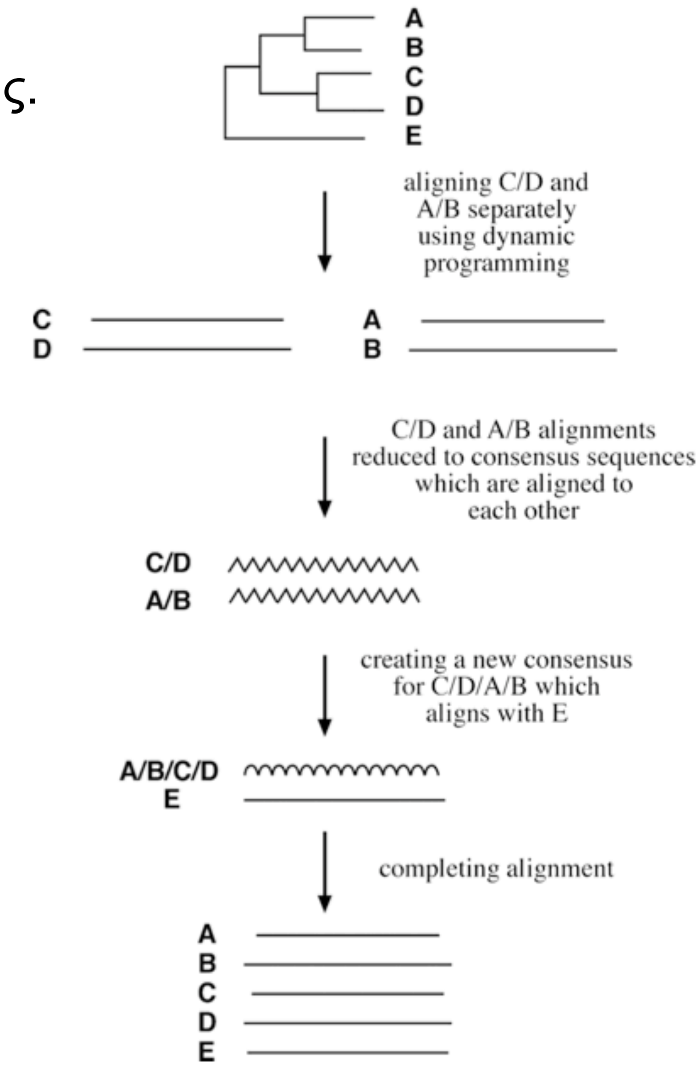
ClustalW (i)

- Ολική στοίχιση (Needlman-Wunsch) κάθε πιθανού ζεύγους
- Πίνακας αποστάσεων (identities ή πίνακες Blossum/PAM).
- Μετατροπή των αποστάσεων σε εξελικτικές αποστάσεις.
- Δημιουργία φυλογενετικού δένδρου - οδηγού (guide tree) (neighbor joining).
  - Χαμηλότερης εμπιστοσύνης από ένα κανονικό φυλογενετικό δένδρο, ωστόσο καταδεικνύει ικανοποιητικά τις βασικές σχέσεις



ClustalW (ii)

- Οι 2 κοντινότερες ακολουθίες στοιχίζονται και δημιουργείται μια ακολουθία συναίνεσης.
- Με βάση το δένδρο-οδηγό, η ακολουθία συναίνεσης στοιχίζεται (δυναμικός προγραμματισμός) με την επόμενη πιο κοντινή ακολουθία ή την επόμενη πιο κοντινή ακολουθία συναίνεσης.
- Η διαδικασία επαναλαμβάνεται έως ότου στοιχιθούν όλες οι ακολουθίες.



ClustalW (iii)

- Ανάλογα με την απόσταση 2 ακολουθιών στο δένδρο-οδηγό, χρησιμοποιείται και ο κατάλληλος πίνακας αντικατάστασης (Blossum62, Blossum 45) για την ολική στοίχιση κατά ζεύγη .
- Οι ποινές των κενών προσαρμόζονται ανάλογα με την παρατηρούμενη συντήρηση μιας περιοχής και ανάλογα με την δευτεροταγή δομή.
- Συντελεστής βαρύτητας ανάλογα με την εξελικτική απόσταση 2 ακολουθιών

Προβλήματα της προοδευτικής στοίχισης

- Δεν ενδύκνειται για ακολουθίες με πολύ διαφορετικά μήκη (λόγω ολικής στοίχισης).

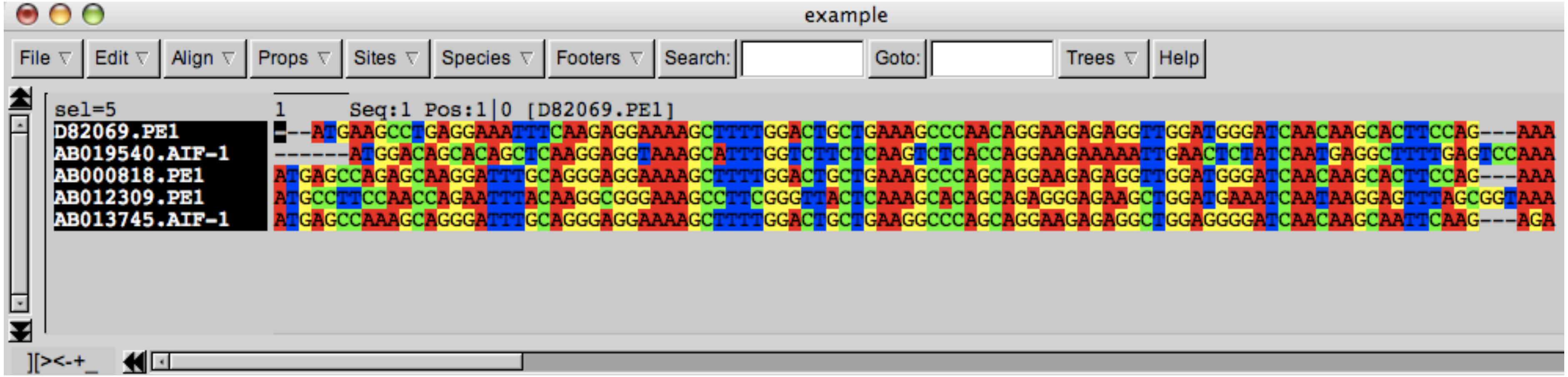
- Η τελική πολλαπλή στοίχιση εξαρτάται από τη σειρά με την οποία θα γίνουν οι επιμέρους στοιχίσεις κατά ζεύγη.

- Ένα αρχικό λάθος θα επηρεάσει τα υπόλοιπα στάδια της πολλαπλής στοίχισης.

Alignment formats

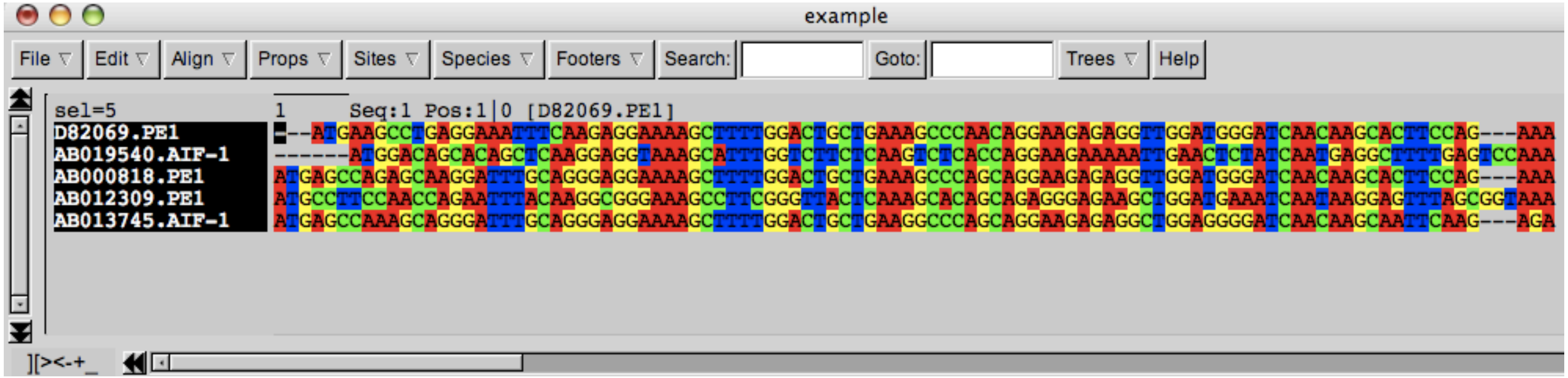
- FASTA (.fa ή .fasta ή .fst)
- Clustal (.aln)
- Phylip (.phy ή .phylip)
- MSF (.msf)
- Mase (.mase)
- Nexus (.nxs)
- Συνήθως, τα alignment editors μπορούν να μετατρέψουν το ένα format σε άλλο.
- Readseq
  - http://www.ebi.ac.uk/cgi-bin/readseq.cgi

Fasta format




```
>D82069.PE1 D82069.PE1 CDS /codon_start=1 /product="iba1, ionized calcium binding adapter mo
---atgaagcctgaggaaatttcaagaggaaaagcttttggactgctgaaagcccaacag
gaagagaggttggatgggatcaacaagcacttccag---aaa
>AB019540.AIF-1 AB019540.AIF-1 CDS /codon_start=1 /transl_table=1 /gene="AIF-1" /product="allogr
------atggacagcacagctcaaggaggtaaagcatttggtcttctcaagtctcaccag
gaagaaaaattgaactctatcaatgaggcttttgagtccaaa
>AB000818.PE1 AB000818.PE1 CDS /codon_start=1 /transl_table=1 /product="MRF-1" /db_xref="GOA:P
atgagccagagcaaggatttgcagggaggaaaagcttttggactgctgaaagcccagcag
gaagagaggttggatgggatcaacaagcacttccag---aaa
>AB012309.PE1 AB012309.PE1 CDS /codon_start=1 /transl_table=1 /product="allograft inflammatory
atgccttccaaccagaatttacaaggcgggaaagccttcgggttactcaaagcacagcag
agggagaagctggatgaaatcaataaggagtttagcggtaaa
>AB013745.AIF-1 AB013745.AIF-1 CDS /codon_start=1 /transl_table=1 /gene="AIF-1" /product="allogr
atgagccaaagcagggatttgcagggaggaaaagcttttggactgctgaaggcccagcag
gaagagaggctggaggggatcaacaagcaattcaag---aga
```

Clustal format



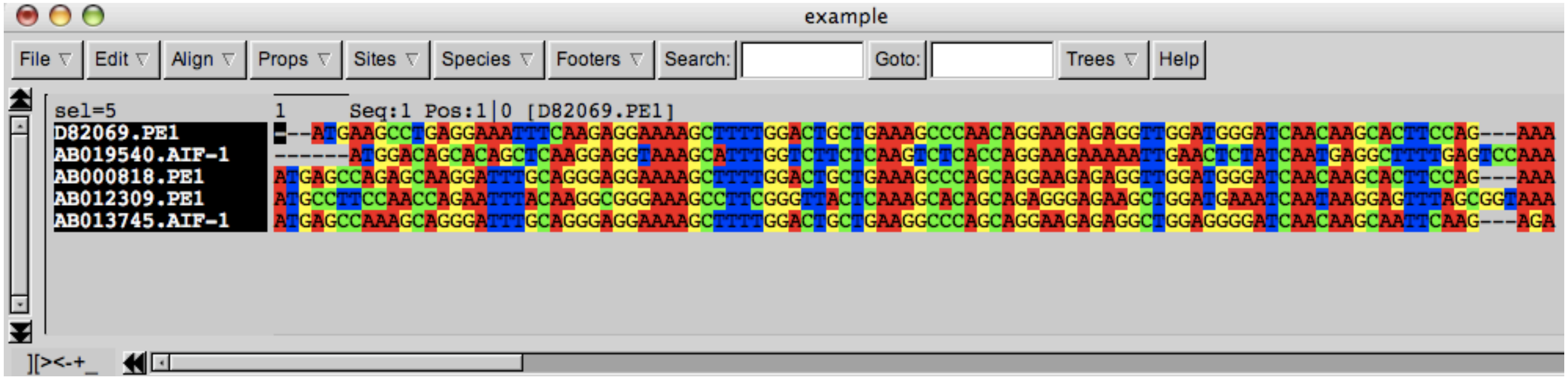

```
CLUSTAL W (1.7) multiple sequence alignment


D82069.PE1       ---atgaagcctgaggaaatttcaagaggaaaagcttttggactgctgaaagcccaacag
AB019540.AIF-1   ------atggacagcacagctcaaggaggtaaagcatttggtcttctcaagtctcaccag
AB000818.PE1     atgagccagagcaaggatttgcagggaggaaaagcttttggactgctgaaagcccagcag
AB012309.PE1     atgccttccaaccagaatttacaaggcgggaaagccttcgggttactcaaagcacagcag
AB013745.AIF-1   atgagccaaagcagggatttgcagggaggaaaagcttttggactgctgaaggcccagcag


D82069.PE1       gaagagaggttggatgggatcaacaagcacttccag---aaa
AB019540.AIF-1   gaagaaaaattgaactctatcaatgaggcttttgagtccaaa
AB000818.PE1     gaagagaggttggatgggatcaacaagcacttccag---aaa
AB012309.PE1     agggagaagctggatgaaatcaataaggagtttagcggtaaa
AB013745.AIF-1   gaagagaggctggaggggatcaacaagcaattcaag---aga
```

Phylip format

- Χρησιμοποιείται στο πρόγραμμα phylip για φυλογένεση




```
5   102
D82069.PE1       ---atgaagc ctgaggaaat ttcaagagga aaagcttttg gactgctgaa agcccaacag
AB019540.AIF-1   ------atgg acagcacagc tcaaggaggt aaagcatttg gtcttctcaa gtctcaccag
AB000818.PE1     atgagccaga gcaaggattt gcagggagga aaagcttttg gactgctgaa agcccagcag
AB012309.PE1     atgccttcca accagaattt acaaggcggg aaagccttcg ggttactcaa agcacagcag
AB013745.AIF-1   atgagccaaa gcagggattt gcagggagga aaagcttttg gactgctgaa ggcccagcag

                 gaagagaggt tggatgggat caacaagcac ttccag---a aa
                 gaagaaaaat tgaactctat caatgaggct tttgagtcca aa
                 gaagagaggt tggatgggat caacaagcac ttccag---a aa
                 agggagaagc tggatgaaat caataaggag tttagcggta aa
                 gaagagaggc tggaggggat caacaagcaa ttcaag---a ga
```

Seaview

- http://pbil.univ-lyon1.fr/software/seaview.html

- Online help
- http://pbil.univ-lyon1.fr/software/seaview_data/seaview.html

Φυλογένεση

Φυλογένεση

- Η εκτίμηση της εξελικτικής ιστορίας γονιδίων/πρωτεϊνών ή οργανισμών.
- Η απεικόνιση αυτής της ιστορίας γίνεται με φυλογράμματα/ κλαδογράμματα

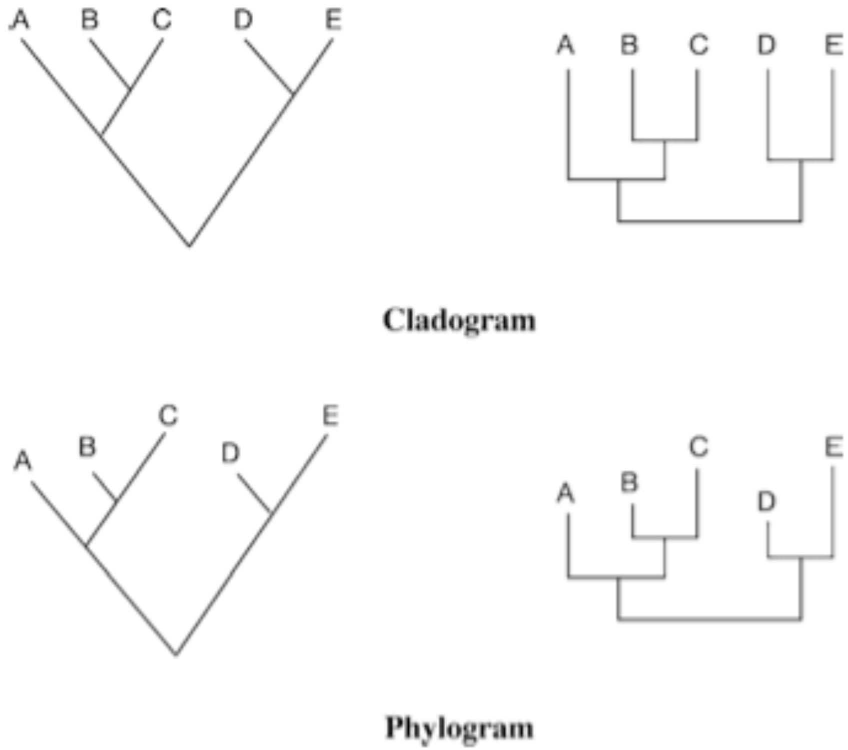


Figure 10.4: Phylogenetic trees drawn as cladograms (*top*) and phylograms (*bottom*). The branch lengths are unscaled in the cladograms and scaled in the phylograms. The trees can be drawn as angled form (*left*) or squared form (*right*).

Λίγη εξέλιξη: ομολογία

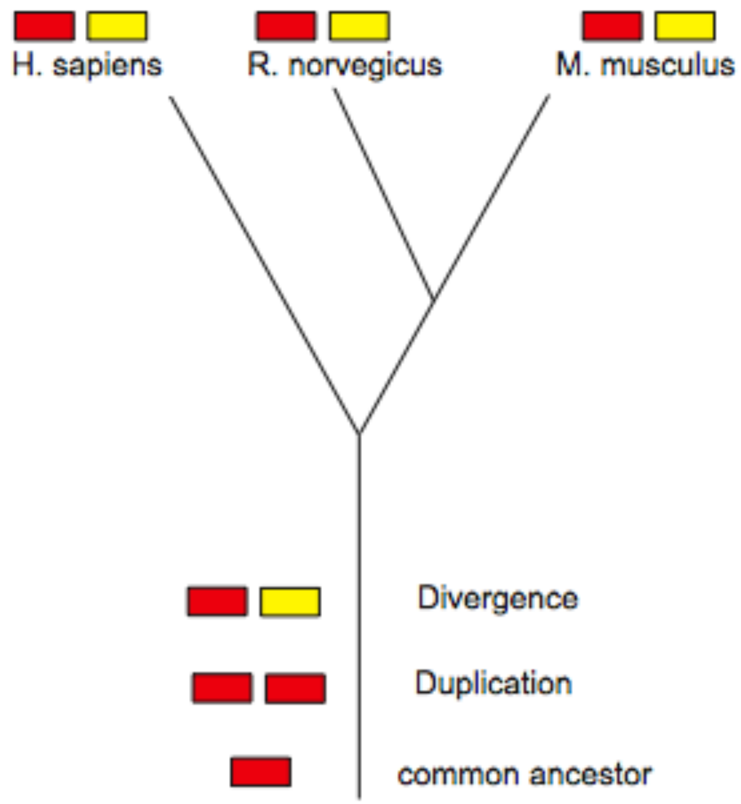
- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος.

- Ορθόλογα γονίδια: προέρχονται από ειδογένεση. Ουσιαστικά, ένα γονίδιο α (μεταλλαγμένο) σε δύο διαφορετικούς οργανισμούς. Συχνά έχουν την ίδια λειτουργία

- Παράλογα γονίδια: προέρχονται από γονιδιακό διπλασιασμό. Ανήκουν στην ίδια οικογένεια

- Ξενόλογα γονίδια: από οριζόντια μεταφορά

Λίγη εξέλιξη: ομολογία (ιι)

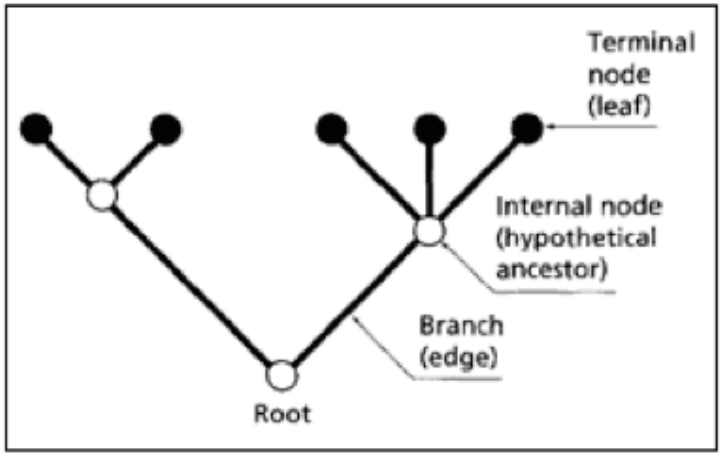


Στάδια φυλογενετικής ανάλυσης

- Επιλογή ακολουθιών:
  - Επιλογή μοριακών δεικτών
  - Εντοπισμός ομόλογων ακολουθιών
    - Π.χ. Blast, HMMs
- Πολλαπλή στοίχιση
  - Διορθώσεις στην στοίχιση
- Υπολογισμός φυλογενετικού δένδρου
  - Επιλογή εξελικτικού μοντέλου
  - Επιλογή μεθόδου δημιουργίας του δένδρου
  - Αξιολόγιση/αξιοπιστία του δένδρου

Στοιχεία ενός φυλογενετικού δένδρου

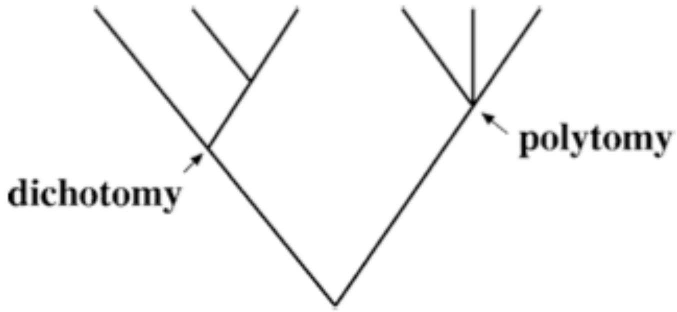
- Φύλλα (leafs)
  - Taxon
  - Operational taxonomic units (OTUs)
- Βραχίονες (branches)
- Κόμβοι (nodes)
- Κλάδοι (clades)
  - Μονοφυλετικά group
- Ρίζα (root)



Στοιχεία ενός φυλογενετικού δένδρου

- Τοπολογία του δένδρου
  - Διχοτόμιση (dichotomy)
  - Πολυτόμιση (polytomy)
    - Radiation
    - Unresolved phylogeny

Figure 10.2: A phylogenetic tree showing an example of bifurcation and multifurcation. Multifurcation is normally a result of insufficient evidence to fully resolve the tree or a result of an evolutionary process known as *radiation*.



Δένδρα με/χωρίς ρίζα

- Χωρίς ρίζα
  - Δεν γνωρίζουμε τον κοινό πρόγονο.
  - Απεικονίζονται μόνο οι σχετικές θέσεις των taxa.
  - Δεν φαίνεται η εξελικτική πορεία.
- Με ρίζα
  - Γνωρίζουμε τον κοινό πρόγονο.
  - Φαίνεται η εξελικτική πορεία.
  - Χρησιμοποιούνται:
    - Outgroup
    - Midpoint rooting approach (υποθέτει την ύπαρξη μοριακού ρολογιού - σταθερού ρυθμού εξέλιξης για όλες τις ακολουθίες).

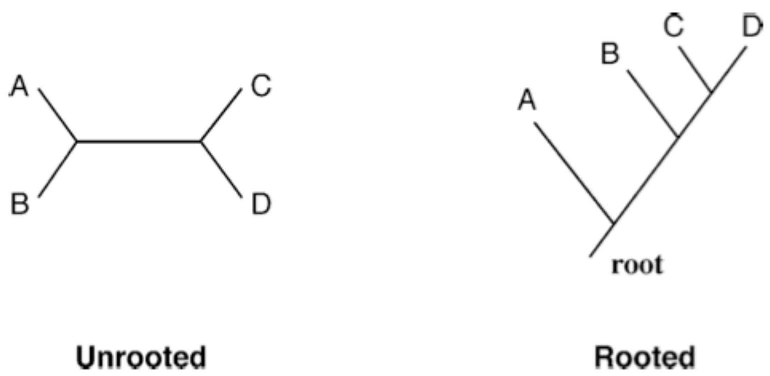


Figure 10.3: An illustration of rooted versus unrooted trees. A phylogenetic tree without definition of a root is unrooted (*left*). The tree with a root is rooted (*right*).

Κλαδόγραμμα/φυλόγραμμα

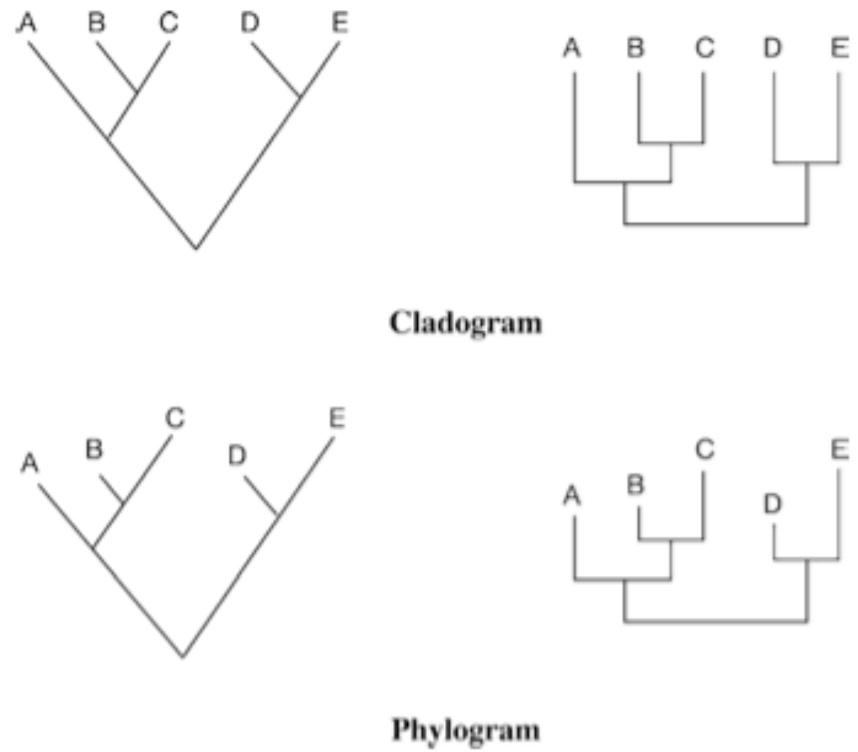
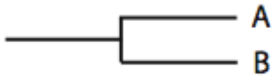


Figure 10.4: Phylogenetic trees drawn as cladograms (*top*) and phylograms (*bottom*). The branch lengths are unscaled in the cladograms and scaled in the phylograms. The trees can be drawn as angled form (*left*) or squared form (*right*).

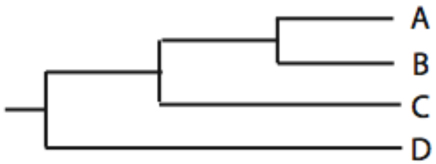
Newick format

- Δένδρα αποθηκεύονται σε μορφή Newick ή Nexus (παραλλαγή του Newick).

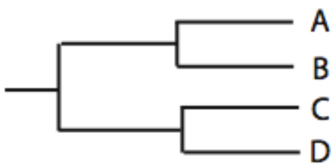


(A,B);

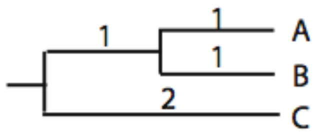
((A,B),C);



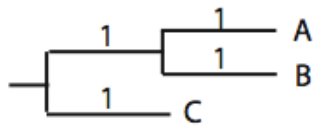
(((A,B),C),D);



((A,B),(C,D));



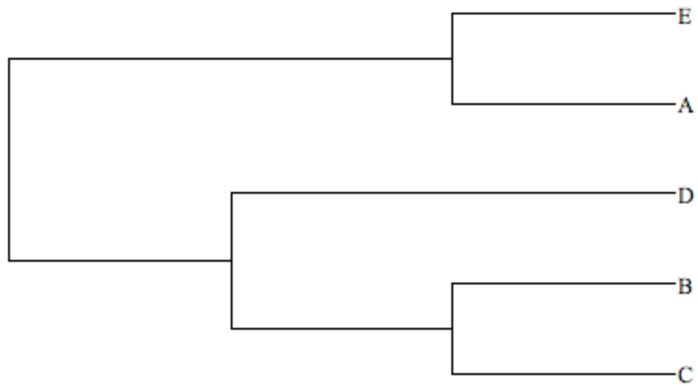
((A:1,B:1):1,C:2);



((A:1,B:1):1,C:1);

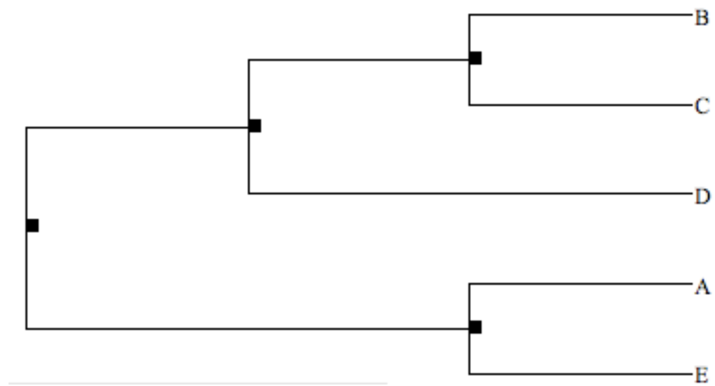
- Ποιό είναι το δένδρο: (((C,B),D),(A,E));

Newick format



(((C,B),D),(A,E));

Είναι το ίδιο δένδρο;



Newick format

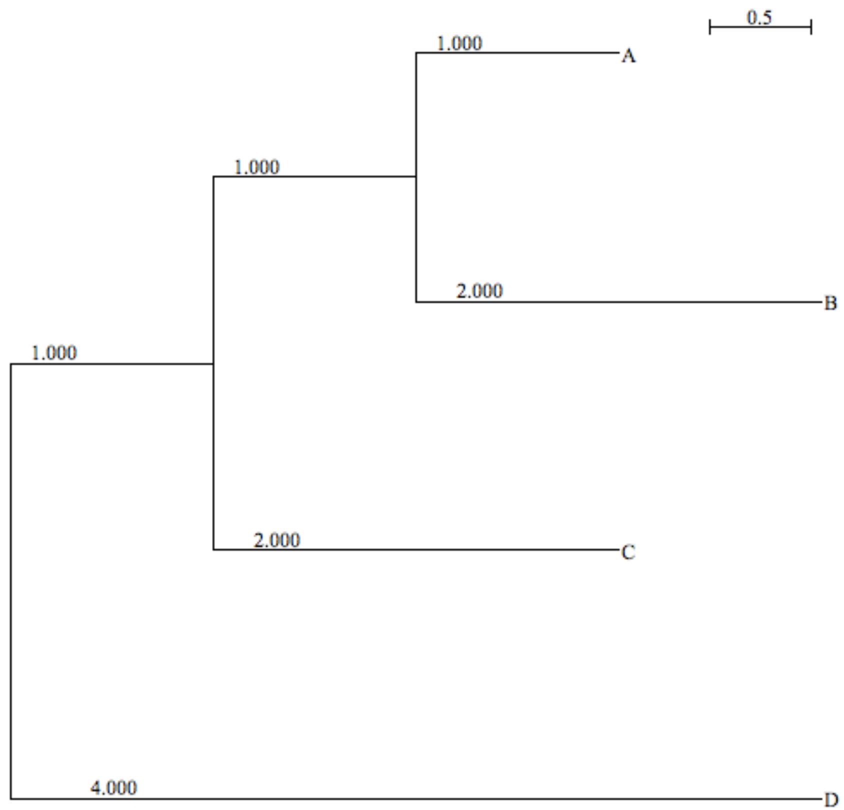
Ποιό είναι το δένδρο;

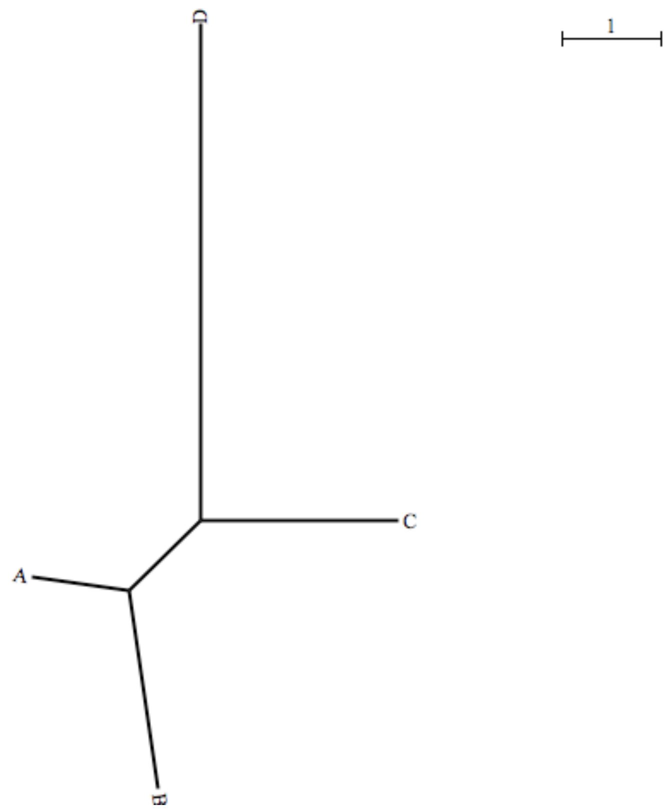
(((A:1,B:2):1,C:2):1,D:4);

Newick format

(((A:1,B:2):1,C:2):1,D:4);

distanceAC=1+1+2





Φυλογένεση γονιδίων/πρωτεϊνών

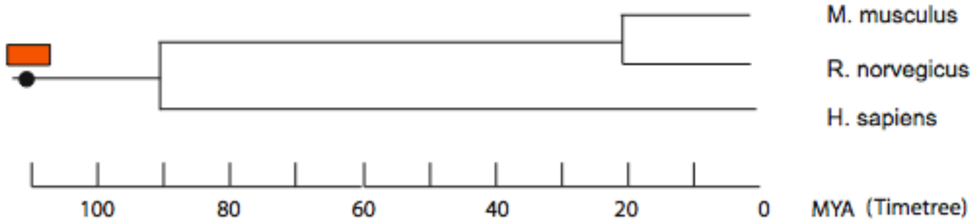
- Φυλογένεση γονιδίων ή πρωτεΐνών.
  - Δείχνει την εξελικτική πορεία μιας οικογένειας γονιδίων.
  - Κάθε κόμβος (node) στο δένδρο είναι ένας γονιδιακός διπλασιασμός ή ειδογένεση.
  - Το κάθε γονίδιο/πρωτεΐνη μπορεί να έχει διαφορετική εξελικτική πορεία (π.χ. Οριζόντια μεταφορά) ή ρυθμό εξέλιξης από τα υπόλοιπα γονίδια ενός οργανισμού.
  - Άρα, η εξελικτική πορεία ενός μόνο γονιδίου/πρωτεΐνης ενδέχεται να μην αντανακλά την εξελικτική πορεία ενός οργανισμού

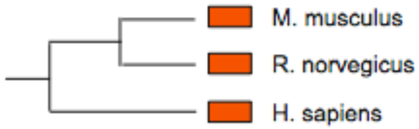
Φυλογένεση οργανισμών

- Δείχνει την εξελικτική πορεία μιας ομάδας οργανισμών.
- Οι κόμβοι (nodes) στο δένδρο απεικονίζουν γεγονότα ειδογένεσης.
- Η φυλογένεση μπορεί να γίνει από:
  - μια σειρά φαινοτυπικών χαρακτήρων
  - Ένα γονίδιο μοριακό δείκτη (π.χ. 16S rRNA)
  - Μια σειρά γονιδίων
  - Από την πλειψηφία των γονιδίων του κάθε γενώματος

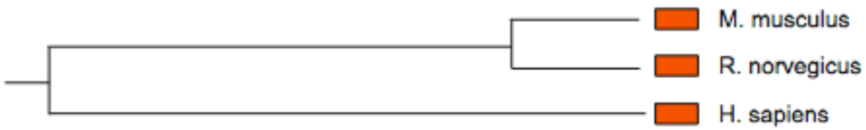
Φυλογένεση οργανισμών

- Επιλέγουμε/βρίσκουμε το ορθόλογο γονίδιο-δείκτη στους οργανισμούς που μελετάμε και ακολουθεί φυλογένεση





cladogram



phylogram

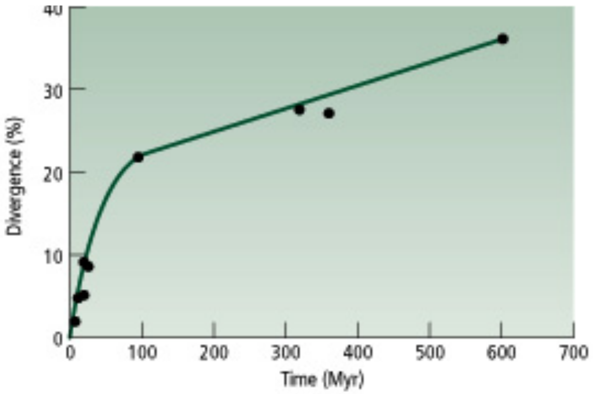
- Το ποντίκι και ο αρουραίος είχαν λιγότερο χρόνο να εξελιχθούν ξεχωριστά, από ότι ο άνθρωπος σε σχέση με το ποντίκι ή σε σχέση με τον αρουραίο. Οι μεταλλάξεις που συσσωρεύτηκαν σε κάθε ορθόλογη ακολουθία πρέπει να είναι ανάλογες του χρόνου απόκλισης των οργανισμών.
- Αν υποθέσουμε ότι ο ρυθμός μετάλλαξης είναι 1/1.000.000 χρόνια, πόσες μεταλλάξεις έχουν συσσωρευθεί σε κάθε ακολουθία, σε σχέση με τον κοινό πρόγονο;

Μοριακοί δείκτες για φυλογένεση οργανισμών

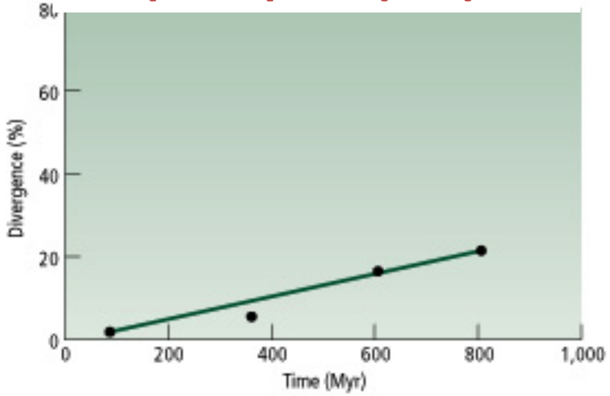
- DNA ή πρωτεΐνη, ανάλογα με την εξελικτική απόσταση των οργανισμών.
- Για πολύ ‘κοντινούς’ οργανισμούς:
  - Περιοχές του DNA που εξελίσσονται γρήγορα.
  - Π.χ. Για άτομα ενός ή περισσότερων πληθυσμών του ίδιου είδους, χρησιμοποιείται mtDNA που δεν κωδικοποιεί πρωτεΐνες.
- Για μέτρια αποκλίνοντες οργανισμούς:
  - rRNA ή πρωτεΐνες.
    - Mt-rRNA 10-100 MY
    - Nuc-rRNA 100-800 MY
- Για βαθιά αποκλείνοντες οργανισμούς:
  - Βαθιά συντηρημμένες πρωτεΐνες.

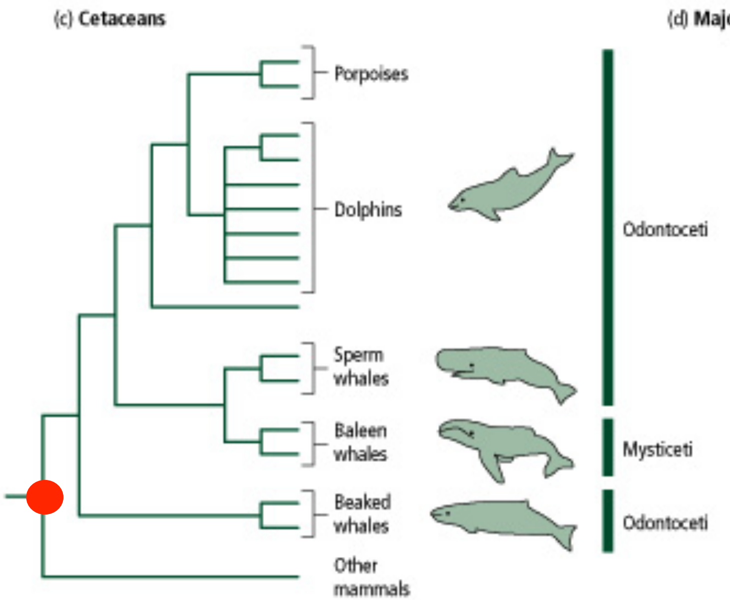
Διαφορετικά γονίδια για διαφορετικά ερωτήματα

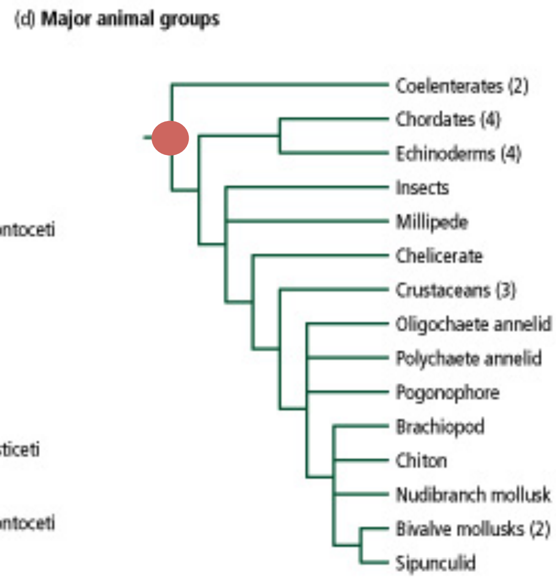
Μοριακό χρονόμετρο



Μοριακή κλεψύδρα



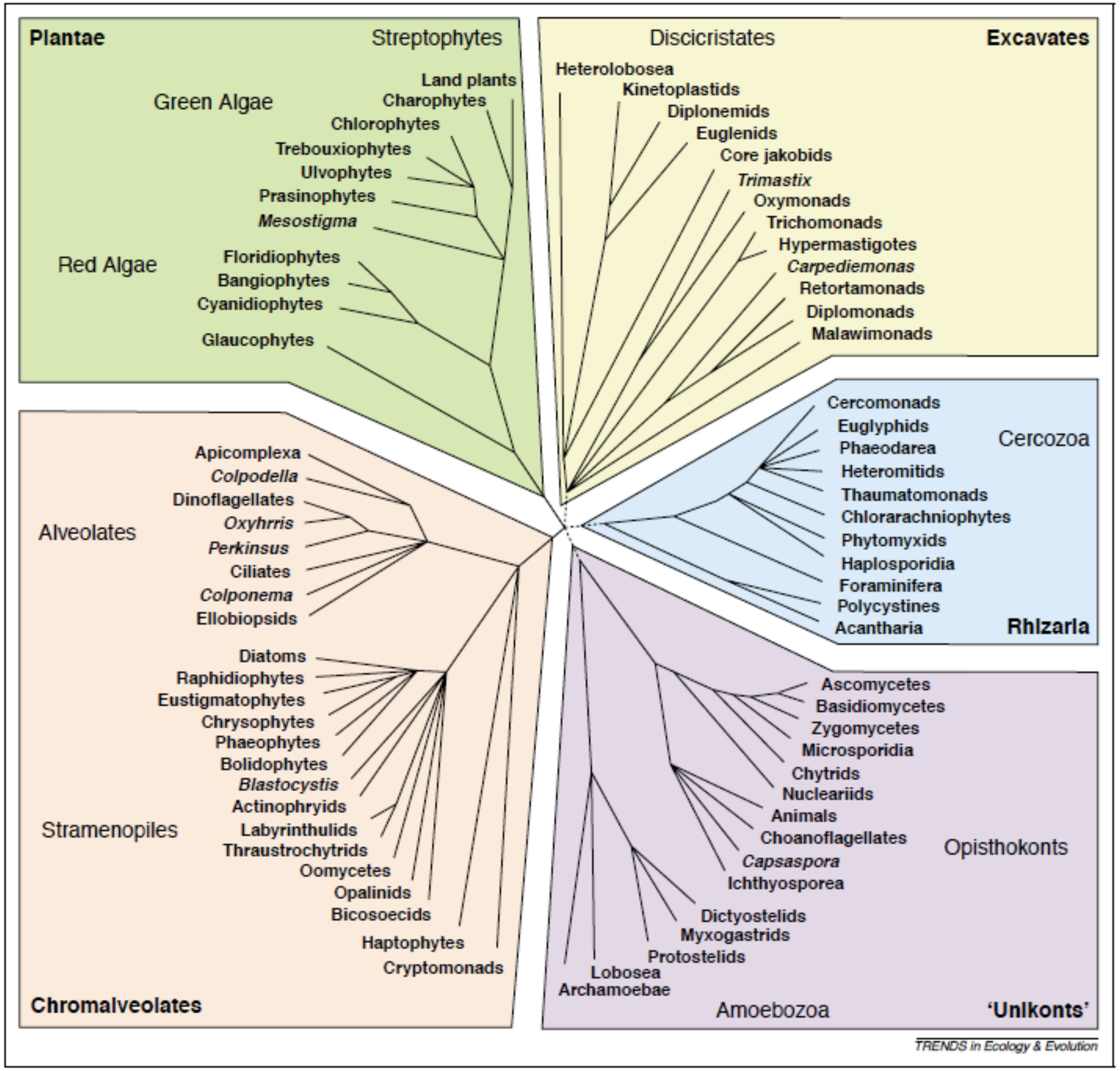




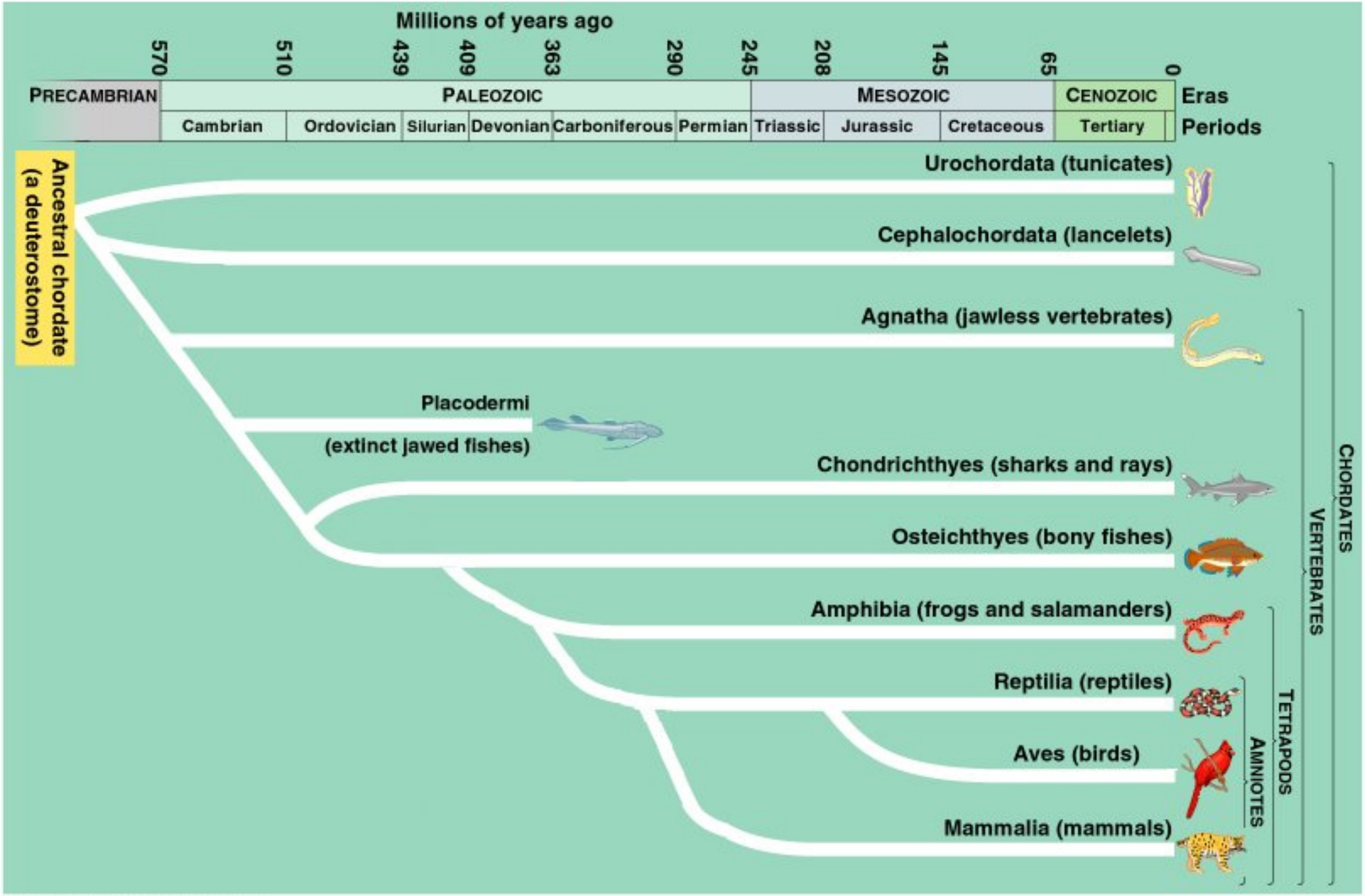
Βαθύτερη ρίζα: **35 mya (με mtRNA)** **600 mya (με πυρηνικό rRNA)**

Φυλογένεση οργανισμών

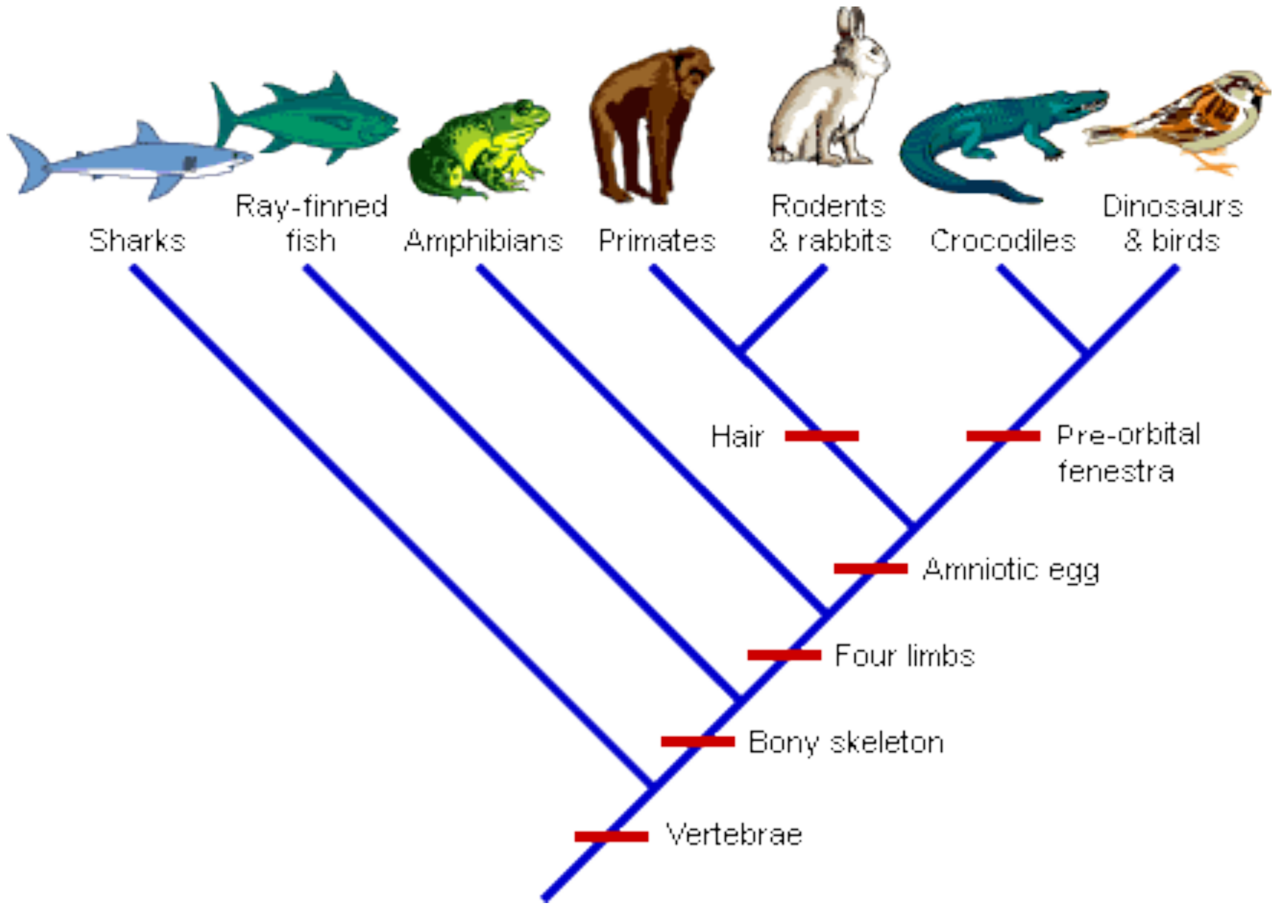
672 Review TRENDS in Ecology and Evolution Vol.20 No.12 December 2005



Φυλογένεση χορδωτών

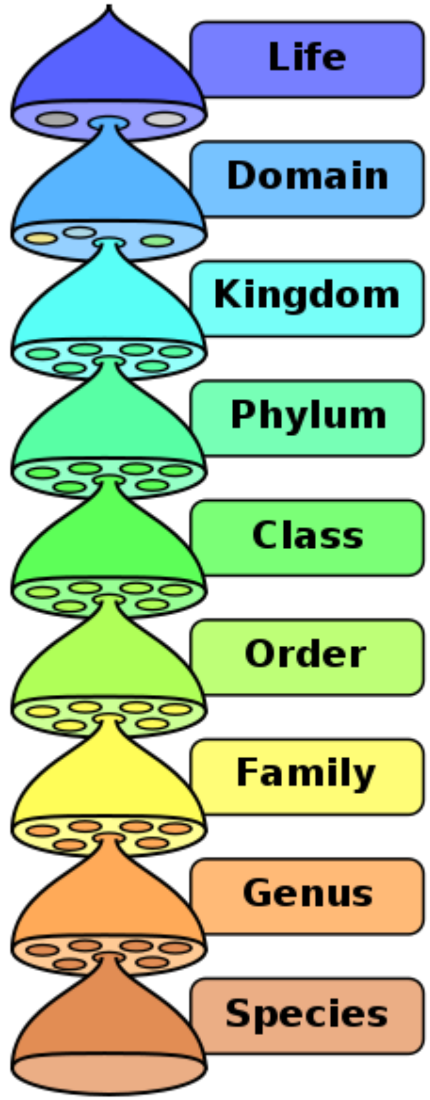


Φυλογένεση σπονδυλωτών

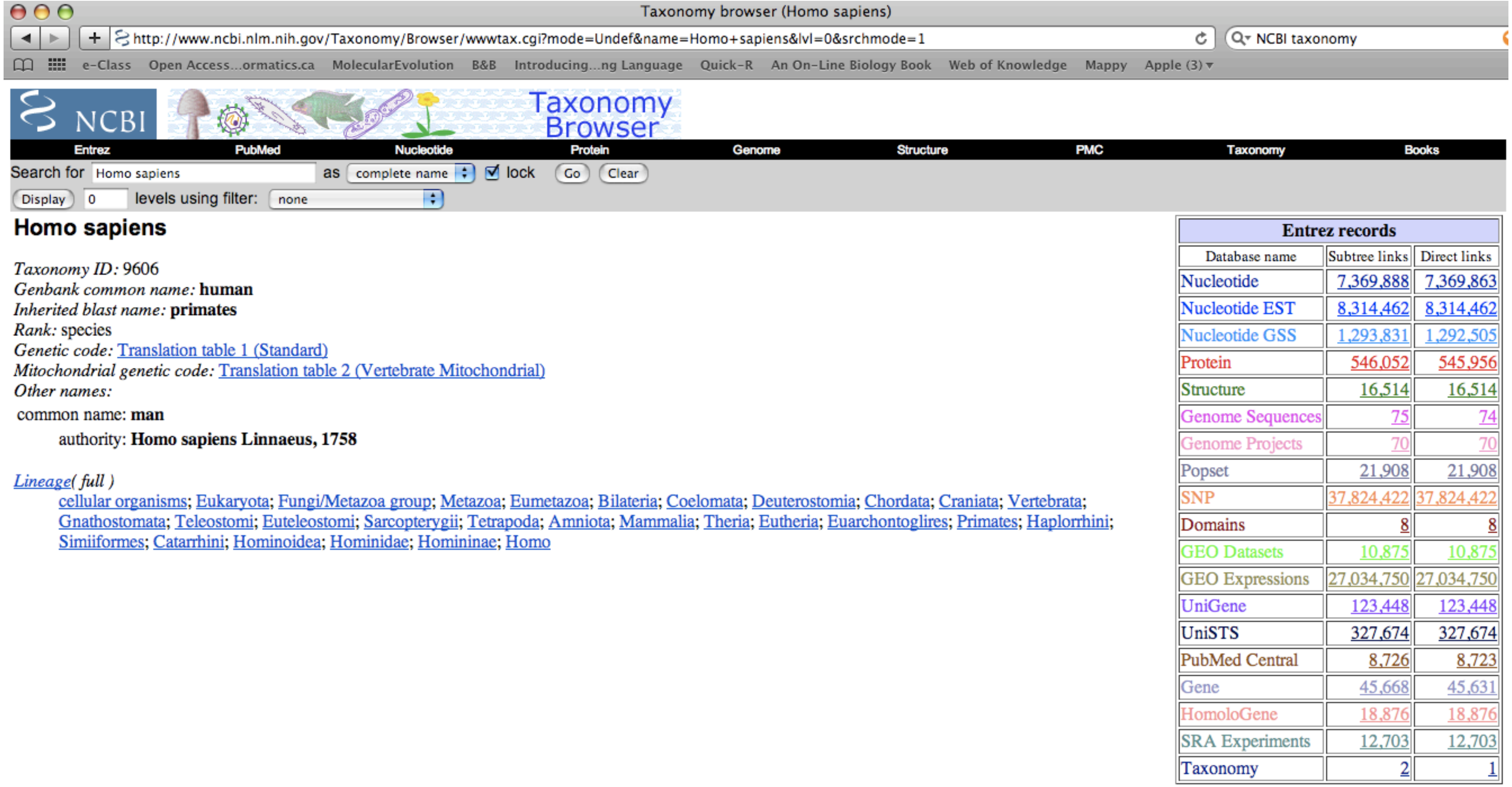


Ταξινόμιση οργανισμών

- Ιεραρχική κατηγοριοποίηση/ ομαδοποίηση οργανισμών.
- Linnaeus (1707-1778) ομαδοποίησε οργανισμούς με βάση κοινούς χαρακτήρες.
- Αργότερα, η ταξινόμιση προσαρμόστηκε στην εξελικτική θεωρία του Δαρβίνου, ώστε να ομαδοποιούνται οι οργανισμοί με βάση την κοινή τους προέλευση.



NCBI taxonomy



Taxonomy browser (Homo sapiens)

http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Homo+sapiens&lvl=0&srchmode=1

NCBI taxonomy

e-Class Open Access...ormatics.ca MolecularEvolution B&B Introducing...ng Language Quick-R An On-Line Biology Book Web of Knowledge Mappy Apple (3)

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search for Homo sapiens as complete name lock Go Clear

Display 0 levels using filter: none

Homo sapiens

Taxonomy ID: 9606
Genbank common name: **human**
Inherited blast name: **primates**
Rank: species
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)
Other names:
common name: **man**
authority: **Homo sapiens Linnaeus, 1758**

Lineage(*full*)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Entrez records

Database name	Subtree links	Direct links
Nucleotide	7,369,888	7,369,863
Nucleotide EST	8,314,462	8,314,462
Nucleotide GSS	1,293,831	1,292,505
Protein	546,052	545,956
Structure	16,514	16,514
Genome Sequences	75	74
Genome Projects	70	70
Popset	21,908	21,908
SNP	37,824,422	37,824,422
Domains	8	8
GEO Datasets	10,875	10,875
GEO Expressions	27,034,750	27,034,750
UniGene	123,448	123,448
UniSTS	327,674	327,674
PubMed Central	8,726	8,723
Gene	45,668	45,631
HomoloGene	18,876	18,876
SRA Experiments	12,703	12,703
Taxonomy	2	1

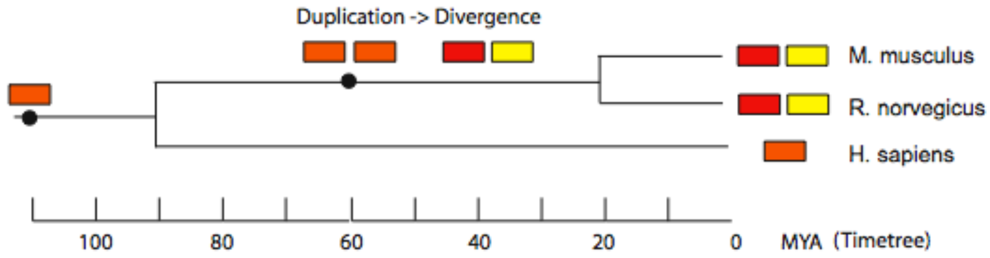
Timetree

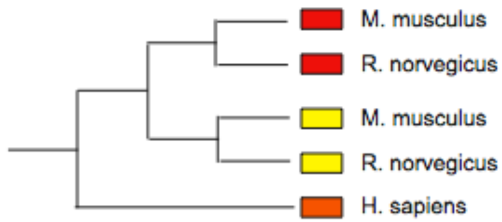

Time Tree :: The Timescale of Life
http://www.timetree.org/
e-Class
Open Access...ormatics.ca
MolecularEvolution
B&B
Introducing...ng Language
Quick-R
An On-Line Biology Book
TIME TREE
THE TIMESCALE OF LIFE
TIMETREE is a public resource for knowledge on the timescale and evolutionary history of life.
Search the database below or go to the TIMETREE OF LIFE for other resources
ABOUT
SEARCH
BOOK
RESOURCES
NEWS
FAQs
CONTACT
TIMETREE OF LIFE BOOK
CONSORTIUM
TimeTree Search
Find time of divergence
(Example: Homo sapiens, Lagomorpha, dog, horses)
Taxon A:
Taxon B:
Clear
Search
Search by Author
Last Name:
Clear
Search
How It Works
TIME
Feliformia
Taxon A
(e.g., cat)
Caniformia
Taxon B
(e.g., dog)
1. Two species or higher taxa are queried (e.g., cat and dog.)
2. TimeTree compares all taxa in one inclusive group (e.g., Feliformia) with those in the other group (e.g., Caniformia) to find all published times of divergence for the evolutionary split.
Citing TimeTree:
Hedges SB, Dudley J & Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms.
Bioinformatics 22:2971-2972 [Download PDF]
THE TIMETREE OF LIFE book
Eubacteria
Vertebrates
Plants
Invertebrates
Search "TimeTree" in App Store
You Tube

Φυλογένεση γονιδίων

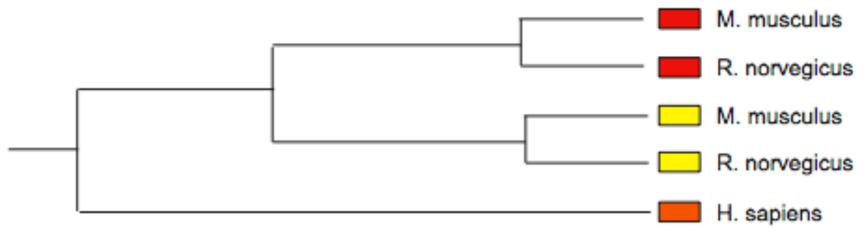
Βρίσκουμε τις ομόλογες ακολουθίες στους οργανισμούς που μας ενδιαφέρουν και ακολουθεί φυλογένεση, για να καταλάβουμε πότε συνέβησαν οι γονιδιακοί διπλασιασμοί, και ποιά ομόλογα είναι πιο κοντινά μεταξύ τους.

Πρέπει να γνωρίζουμε τις εξελικτικές σχέσεις των οργανισμών





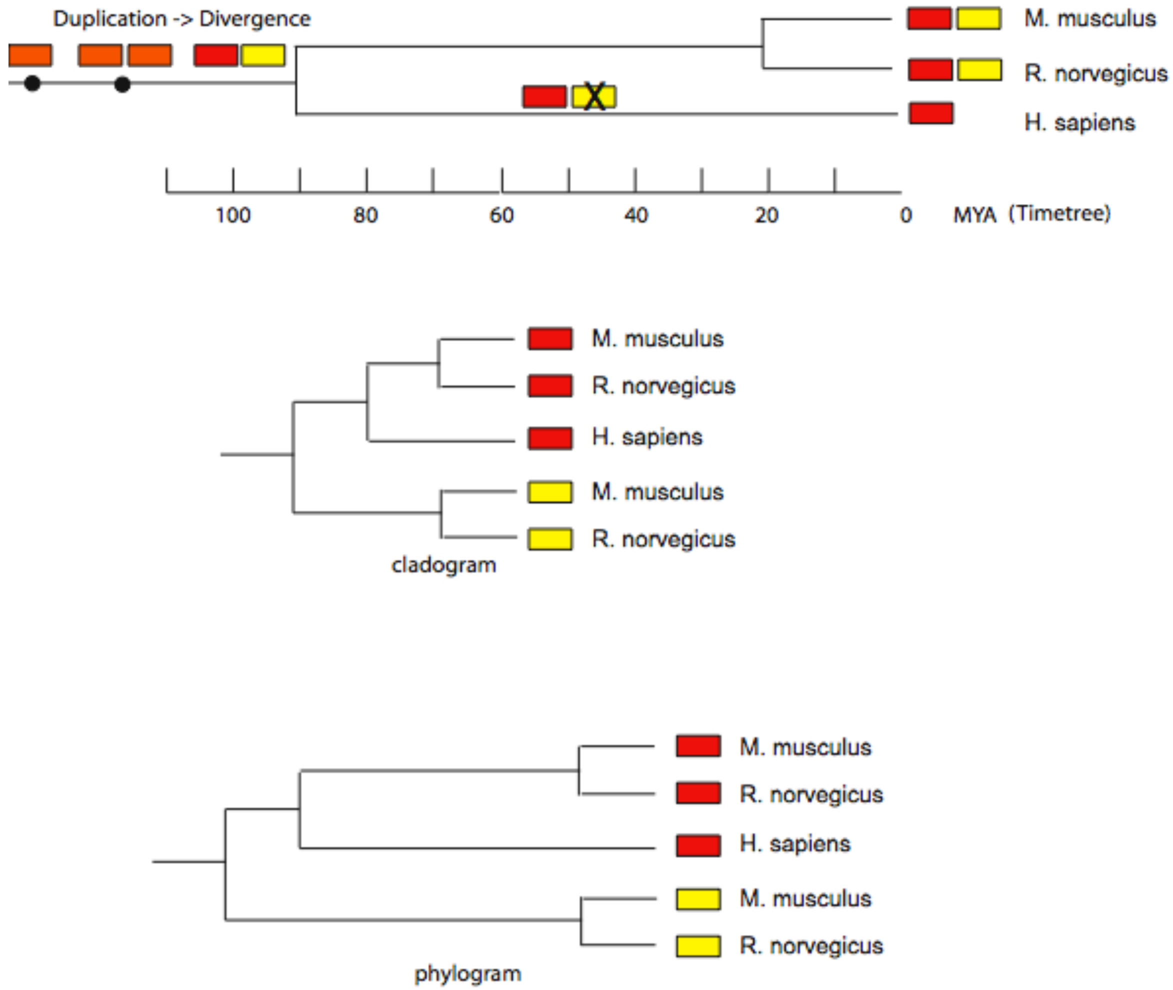
cladogram



phylogram

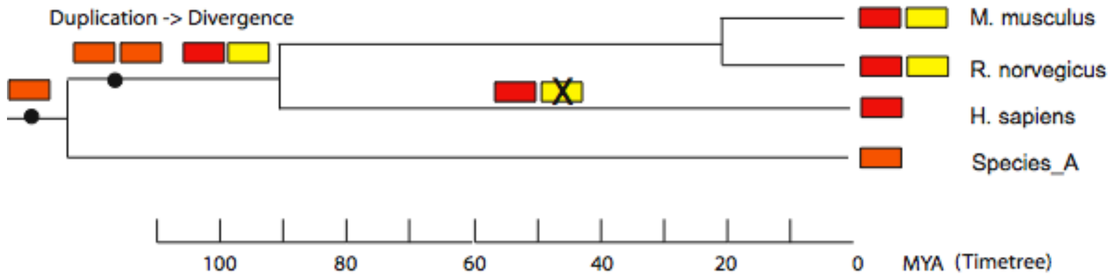
Φυλογένεση γονιδίων

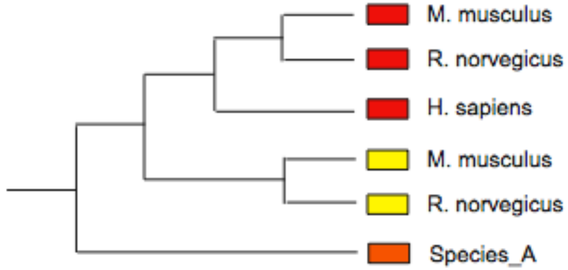
- Απώλεια αντίγραφου



Φυλογένεση γονιδίων

Δειγματοληψία ορθόλογων από πιο απομακρυσμένους οργανισμούς, μέχρι να εντοπιστεί ο χρόνος που συνέβη ο διπλασιασμός. Απολιθώματα βοηθούν στην χρονολόγιση





cladogram

phylogram

Δένδρα συναίνεσης

- Μια μέθοδος μπορεί να οδηγήσει σε περισσότερα από ένα εξίσου καλά δένδρα.
- Ή, από τα ίδια δεδομένα, δημιουργούνται δένδρα με διαφορετικές μεθόδους.
- Το δένδρο συναίνεσης δείχνει ποιοί κόμβοι είναι κοινοί μεταξύ των διαφόρων δένδρων.
- Για κόμβους που δεν παρατηρείται συμφωνία, εμφανίζονται ως πολυτομημένοι.
- Μέθοδοι δημιουργίας δένδρου συναίνεσης:
  - απόλυτη συναίνεση (strict consensus) (100%)
  - Μέθοδος πλειοψηφίας (majority rule) (>50%)

Δένδρα συναίνεσης

- Το παράδειγμα της φυλογενετικής σχέσης ανθρώπου-χιμπατζή-γορίλα

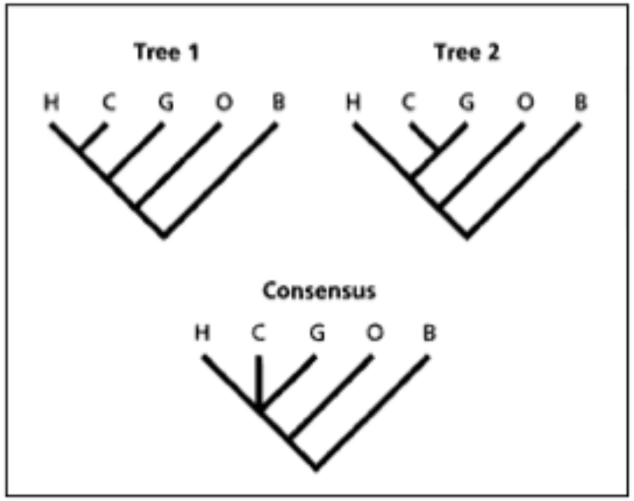


Fig. 2.26
Two different trees for humans (H), chimps (C), gorillas (G), orang-utans (O) and gibbons (B), and their consensus tree.

Πόσα πιθανά δένδρα;

- Το σύνολο των πιθανών διαφορετικών δένδρων για ένα αριθμό taxa αυξάνει εκθετικά

$$N_R = (2n-3)!/2^{n-2}(n-2)! \quad \text{(Eq. 10.1)}$$

In this formula, $(2n-3)!$ is a mathematical expression of factorial, which is the product of positive integers from 1 to $2n-3$. For example, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$.

For unrooted trees, the number of unrooted tree topologies (N_U) is:

$$N_U = (2n-5)!/2^{n-3}(n-3)! \quad \text{(Eq. 10.2)}$$

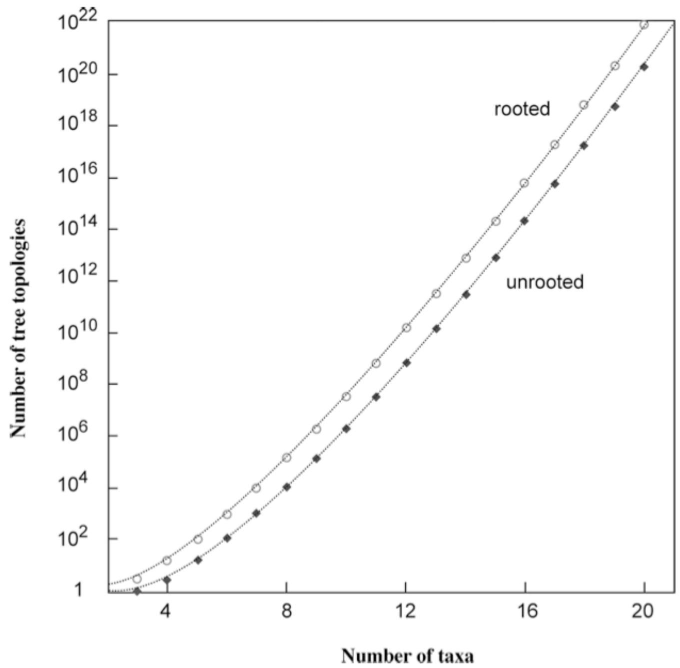
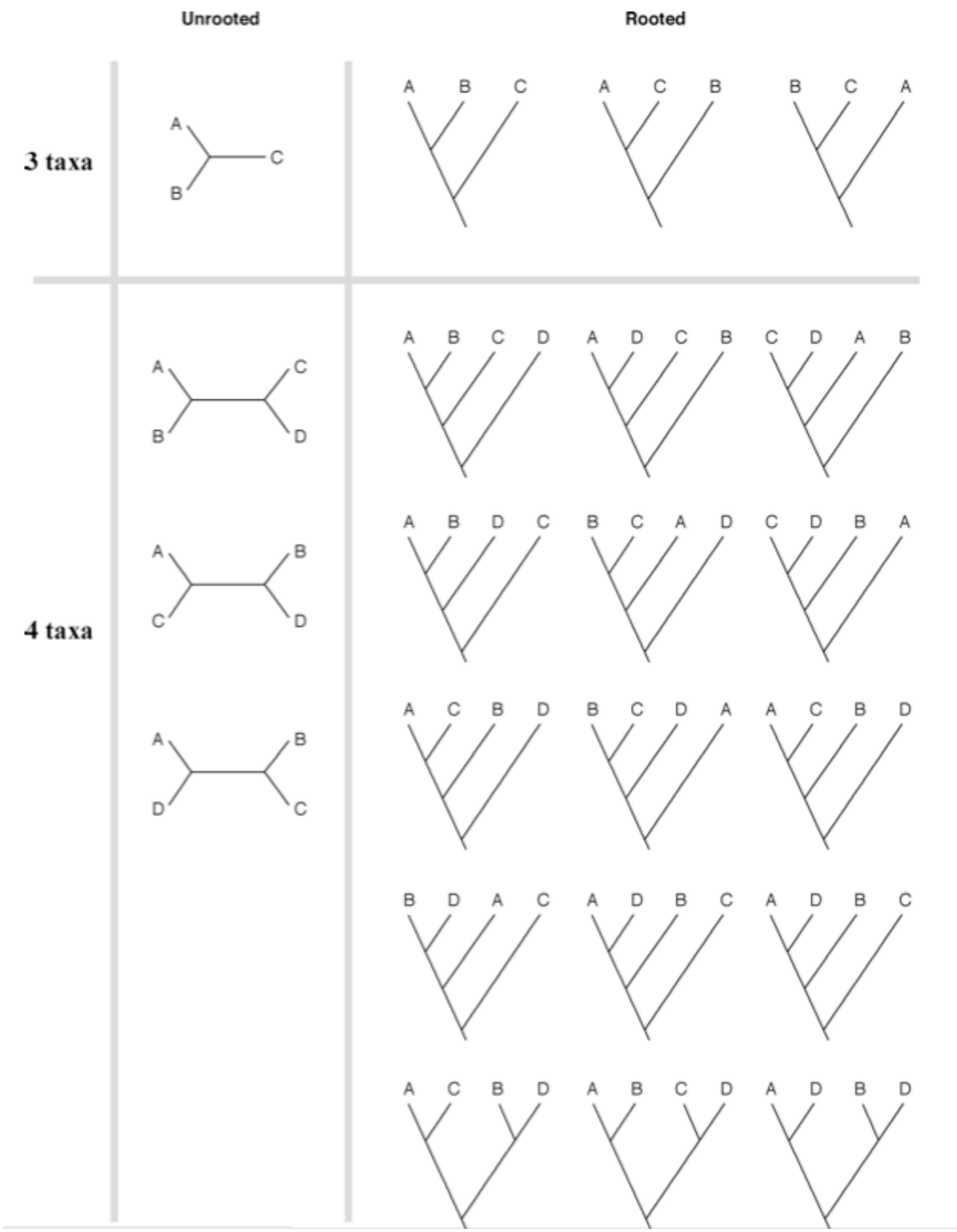


Figure 10.8: Total number of rooted (○) and unrooted (◆) tree topologies as a function of the number of taxa. The values in the *y*-axis are plotted in the log scale.

Πόσα πιθανά δένδρα;



Μέθοδοι Φυλογένεσης

- Μέθοδοι που βασίζονται σε αποστάσεις
  - UPGMA
  - Κοντινότερης γειτονίας (Neighbor joining)
  - Fitch-Margoliash
  - Ελάχιστης εξέλιξης

- Μέθοδοι που βασίζονται σε χαρακτήρες
  - Μέγιστη φειδωλότητα (Maximum Parsimony)
  - Μέγιστη πιθανοφάνεια (Maximum Likelihood)

Μέθοδοι αποστάσεων

- Αρχικά υπολογίζονται οι αποστάσεις ανάμεσα σε όλα τα πιθανά ζεύγη ακολουθιών.
- Δημιουργείται ένας πίνακας αποστάσεων.
- Με βάση τον πίνακα αυτό, δημιουργούνται δένδρα με μεθόδους που βασίζονται:
  - Στην ομαδοποίηση. Η ομαδοποίηση ξεκινάει από τις πιο κοντινές ακολουθίες και σταδιακά ενσωματώνει όλο και πιο απομακρυσμένες:
    - UPGMA
    - Neighbor joining
  - Στην βελτιστοποίηση. Ο αλγόριθμος συγκρίνει τις πιθανές τοπολογίες και επιλέγει αυτή που οι αποστάσεις πάνω στο δένδρο ταιριάζουν καλύτερα με τις αποστάσεις στον αρχικό πίνακα αποστάσεων:
    - Fitch-Margoliash
    - Ελάχιστη εξέλιξη

Υπολογισμός της απόστασης μεταξύ δύο ακολουθιών

- Παρατηρούμενη απόσταση: από την στοίχιση, μπορούμε να δούμε σε ποιές θέσεις δεν ταιριάζουν οι χαρακτήρες.
- Η παρατηρούμενη απόσταση δεν συμπίπτει με την πραγματική (εξελικτική) απόσταση, λόγω πολλαπλών αντικαταστάσεων στην ίδια θέση. Όσο μεγαλύτερη η απόσταση, τόσο πιο πολλές αντικαταστάσεις συνέβησαν στην ίδια θέση.

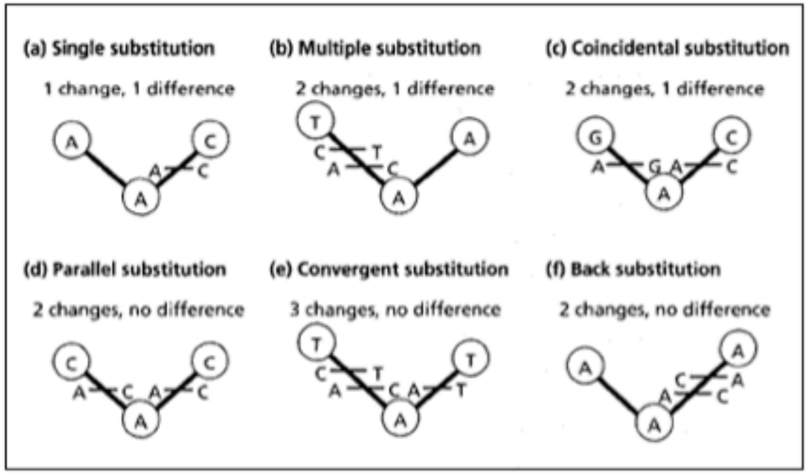


Fig. 5.9
Six kinds of nucleotide substitution. In each case the ancestral nucleotide was A. In all except the case of a single substitution, the number of substitutions that actually occurred is greater than would be counted if we just compared the two descendant sequences. In the lower three cases the nucleotides are identical in both descendant sequences, but this similarity has not been directly inherited from the ancestral sequence. Such similarity is termed 'homoplasious'.

Υπολογισμός της απόστασης μεταξύ δύο ακολουθιών

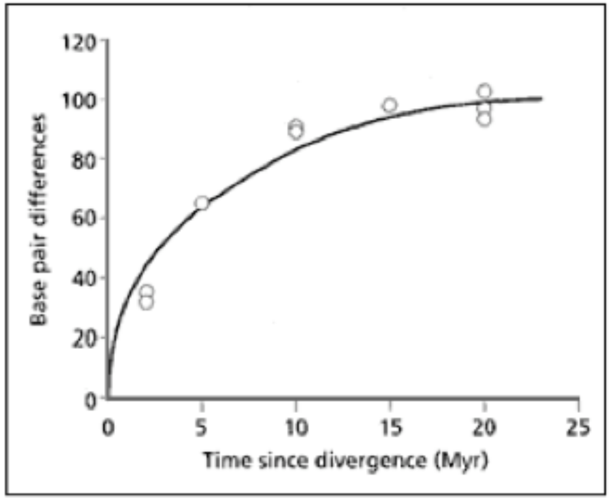


Fig. 5.11
Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al*. (1996).

Διόρθωση της απόστασης μεταξύ 2 ακολουθιών

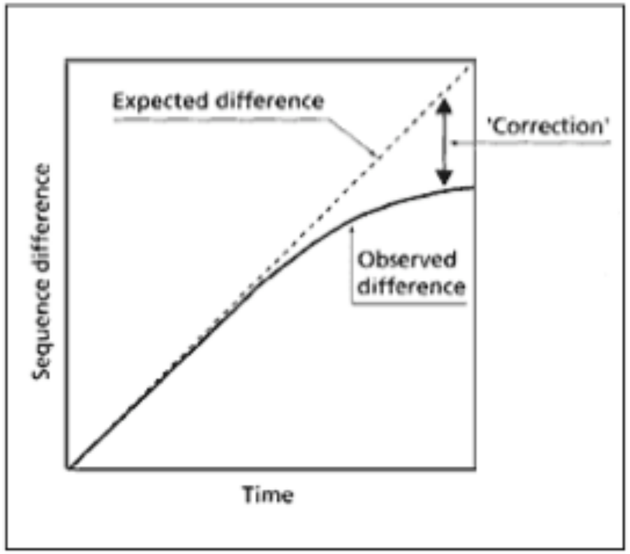
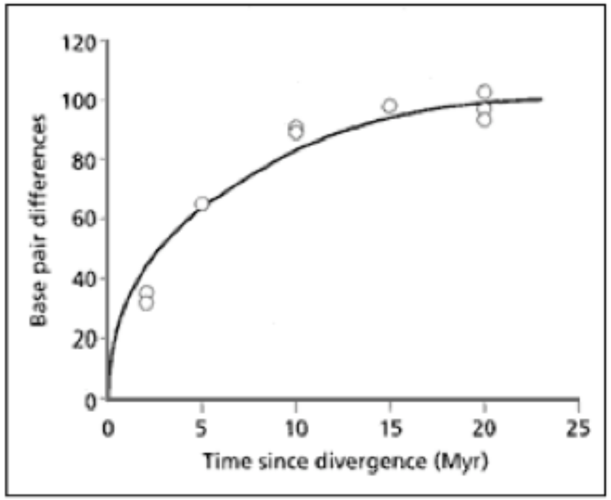


Fig. 5.12
The need to correct observed sequence differences. The extent of observed differences between two sequences is not linear with time (as we would expect if the rate of molecular evolution is approximately constant) but curvilinear due to multiple hits. The goal of distance correction methods is to recover the amount of evolutionary change that the multiple hits have overprinted and to 'correct' the distances for unobserved hits. In effect, the methods seek to 'straighten out' the line representing observed differences.

Μοντέλα αντικατάστασης

- Στατιστικά μοντέλα που λαμβάνουν υπόψην τις πολλαπλές αντικαταστάσεις (για την ίδια θέση) και διορθώνουν την παρατηρούμενη απόσταση, μετατρέποντας την σε εξελικτική.
- Αν η απόσταση είναι πολύ μεγάλη, τότε έχει επέλθει κορεσμός και δεν είναι δυνατόν να γίνει σωστή διόρθωση.



Μοντέλο αντικατάστασης Jukes - Cantor

- Είναι το απλούστερο μοντέλο για ακολουθίες DNA.
- κάθε νουκλεοτίδιο εμφανίζεται με την ίδια συχνότητα
- έχει την ίδια πιθανότητα να μεταλλαχθεί σε ένα από τα υπόλοιπα 3 νουκλεοτίδια

$$d_{AB} = -(3/4)\ \ln[1 - (4/3)\,p_{AB}] \qquad \text{(Eq. 10.3)}$$

where d_{AB} is the evolutionary distance between sequences A and B and p_{AB} is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

For example, if an alignment of sequences A and B is twenty nucleotides long and six pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3. To correct for multiple substitutions using the Jukes–Cantor model, the corrected evolutionary distance based on Equation 10.3 is:

$$d_{AB} = -3/4\ln[1 - (4/3 \times 0.3)] = 0.38$$

Μοντέλο αντικατάστασης Kimura

- Πιο εξελιγμένο μοντέλο.
- κάθε νουκλεοτίδιο εμφανίζεται με την ίδια συχνότητα
- Θεωρεί ότι οι μεταπτώσεις έχουν άλλη πιθανότητα να συμβούν, από ότι οι μεταστροφές.

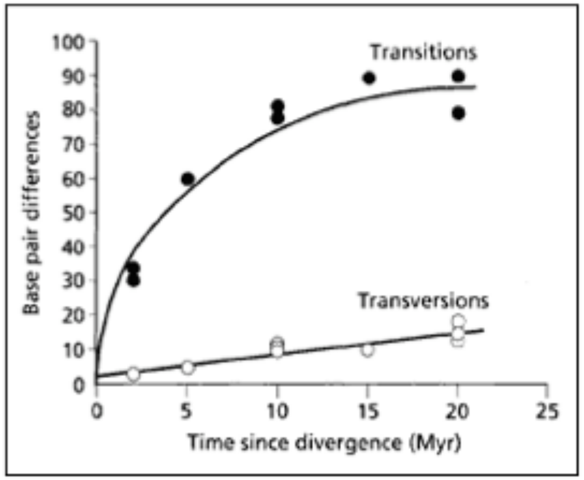


Fig. 5.13
The number of transitions and transversions between the same bovid mammal sequences used in Fig. 5.11. Transitions accumulate much more rapidly than transversions and become saturated, whereas transversions accumulate more slowly and show no evidence of saturation.

Μοντέλο αντικατάστασης Kimura

$$d_{AB} = -(1/2)\ \ln(1 - 2p_{ti} - p_{tv}) - (1/4)\ \ln(1 - 2p_{tv}) \qquad \text{(Eq. 10.4)}$$

An example of using the Kimura model can be illustrated by the comparison of sequences A and B that differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions, the evolutionary distance can be calculated using Equation 10.4:

$$d_{AB} = -1/2\ \ln(1 - 2 \times 0.2 - 0.1) - 1/4\ \ln(1 - 2 \times 0.1) = 0.40$$

Μοντέλα αντικατάστασης για DNA

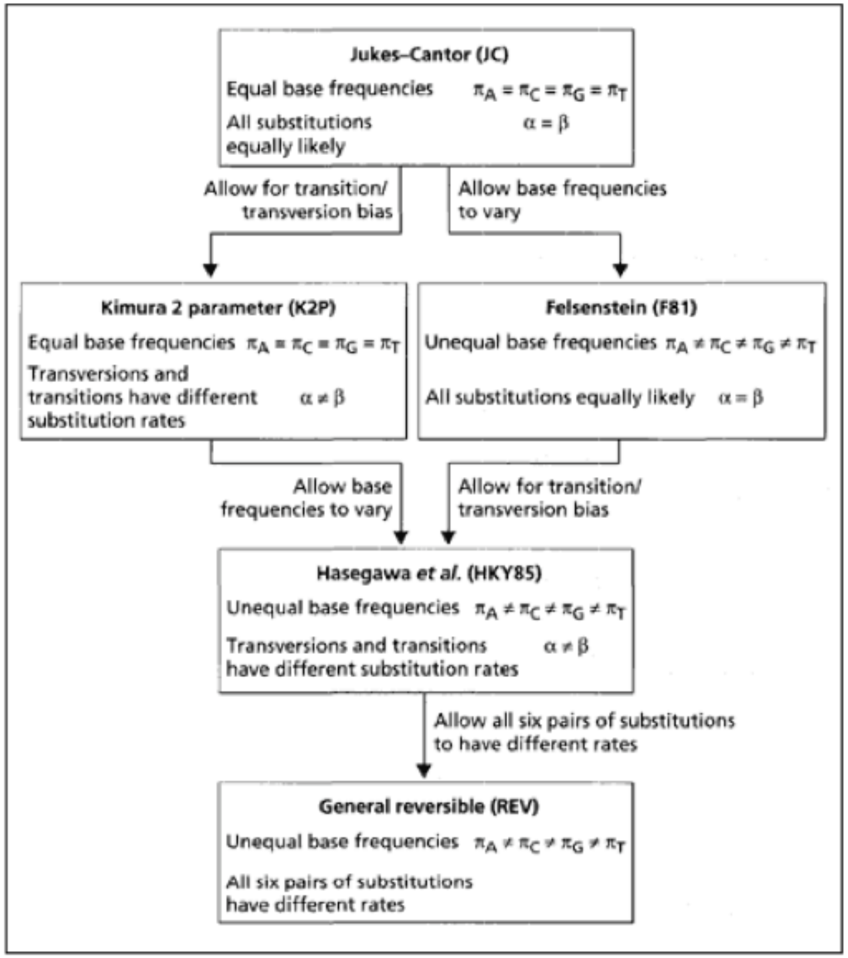


Fig. 5.14
Interrelationships among five models for estimating the number of nucleotide substitutions among a pair of DNA sequences. The JC, K2P, F81 and HKY85 models can all be generated by constrainingvarious parameters of the REV model.

Διόρθωση των παρατηρούμενων αποστάσεων για πρωτεΐνες

2.3 *Estimation of distance between two protein sequences* • **47**

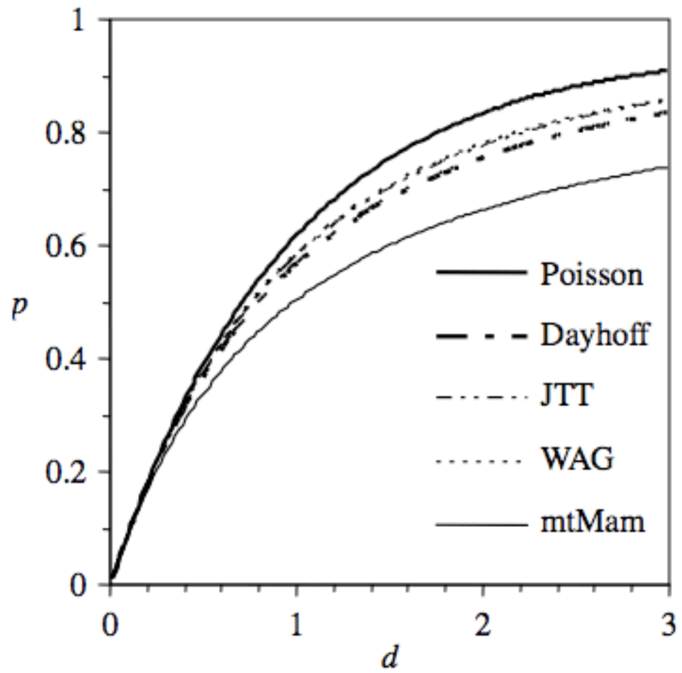


Fig. 2.2 The expected proportion of different sites (p) between two sequences separated by time or distance d under different models. The models are, from top to bottom, Poisson, WAG (Whelan and Goldman 2001), JTT (Jones *et al.* 1992), DAYHOFF (Dayhoff *et al.* 1978), and MTMAM (Yang *et al.* 1998). Note that the results for WAG, JTT, and DAYHOFF are almost identical.

Διόρθωση των παρατηρούμενων αποστάσεων για πρωτεΐνες

- Διόρθωση με πίνακες αντικατάστασης:
  - PAM
  - JTT (Jones-Taylor-Thornton)

- Διόρθωση με αντίστοιχες μεθόδους Jukes-Cantor ή Kimura, προσαρμοσμένες για πρωτεΐνες.

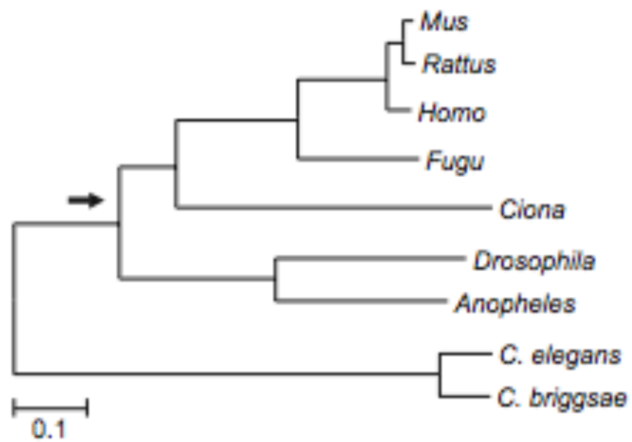
distances. For example, the Kimura model for correcting multiple substitutions in protein distances is:

$$d = -\ln(1 - p - 0.2p^2) \qquad \text{(Eq. 10.5)}$$

whereas p is the observed pairwise distance between two sequences.

Μέθοδος σύνδεσης γειτονίας neighbor joining

- Είναι παρόμοια μέθοδος με το UPGMA.
- Ωστόσο, δεν θεωρεί ότι όλες οι ακολουθίες εξελίσσονται με τον ίδιο ρυθμό.
- Το δένδρο που παράγεται είναι άρριζο και πρέπει εμείς να επιλέξουμε που είναι η ρίζα.



Μέθοδοι βελτιστοποίησης

- Οι μέθοδοι που βασίζονται σε ομαδοποίηση παράγουν ένα δένδρο.
- Δεν γνωρίζουμε πόσο καλύτερο είναι αυτό το δένδρο από άλλα εναλλακτικά δένδρα.
- Οι μέθοδοι βελτιστοποίησης ελέγχουν τα διάφορα πιθανά δένδρα και βρίσκουν αυτό που ταιριάζει καλύτερα στον αρχικό πίνακα αποστάσεων.

Υπέρ και κατά μεθόδων βασισμένων σε αποστάσεις

- Οι μέθοδοι βελτιστοποίησης δίνουν καλύτερα αποτελέσματα από τις μεθόδους ομαδοποίησης, αλλά είναι πιο αργές.
- Αν τα δεδομένα είναι πολλά, τότε προτιμάται μια μέθοδος ομαδοποίησης.
- Οι μέθοδοι αποστάσεων διορθώνουν τις παρατηρούμενες αποστάσεις. Όταν οι ακολουθίες είναι απομακρυσμένες, αυτή η διόρθωση έχει μεγάλες επιπτώσεις και πρέπει να γίνεται.
- Με τις μεθόδους αποστάσεων χάνεται πληροφορία και δεν είναι δυνατόν να ανακατασκευαστεί μια προγονική ακολουθία.

Μέθοδοι που βασίζονται σε χαρακτήρες

Μέγιστη φειδωλότητα (Maximum Parsimony)

Μέγιστη πιθανοφάνεια (Maximum Likelihood)

Βασίζονται στους χαρακτήρες των ακολουθιών και όχι στις αποστάσεις μεταξύ των ακολουθιών.

Είναι δυνατή η ανακατασκευή των προγονικών ακολουθιών.

Έλξη μεταξύ μακρινών βραχιόνων (long branch attraction).

- Τάξα που εξελίσονται με γρήγορους ρυθμούς και επομένως έχουν μακρείς βραχίονες, έλκονται μεταξύ τους.

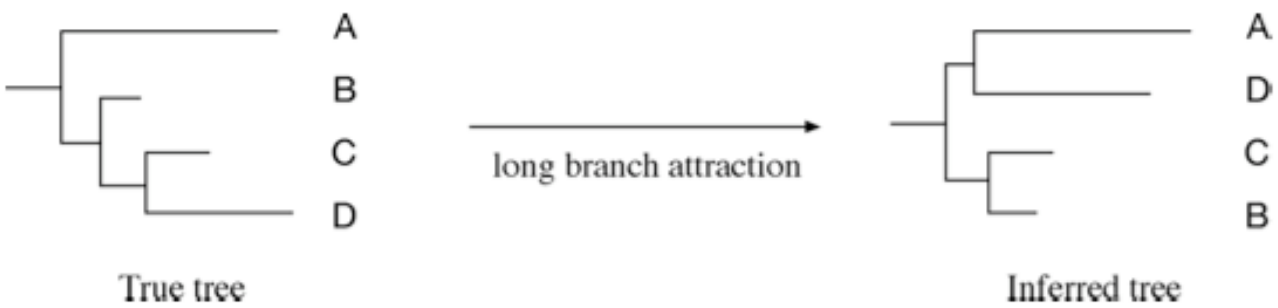


Figure 11.7: The LBA artifact showing taxa A and D are artifactually clustered during phylogenetic construction.

Μέγιστη πιθανοφάνεια

- Βασίζεται σε χαρακτήρες.
- Χρησιμοποιεί όλες τις θέσεις μια πολλαπλής στοίχισης.
- Χρησιμοποιεί πιθανότητες και μοντέλα αντικατάστασης.
- Υπολογίζονται οι χαρακτήρες σε κάθε προγονική ακολουθία.
- Υπολογίζει για το κάθε πιθανό εξελικτικό μονοπάτι (προγονικές ακολουθίες και δένδρο) την πιθανότητα του, με βάση τα παρατηρούμενα σημερινά δεδομένα και ένα συγκεκριμμένο μοντέλο εξέλιξης (μοντέλο αντικατάστασης).
- Οι πιθανότητες μετατρέπονται σε log-likelihood scores.
- Δένδρο με το μεγαλύτερο log-likelihood score επιλέγεται.

Μέγιστη πιθανοφάνεια

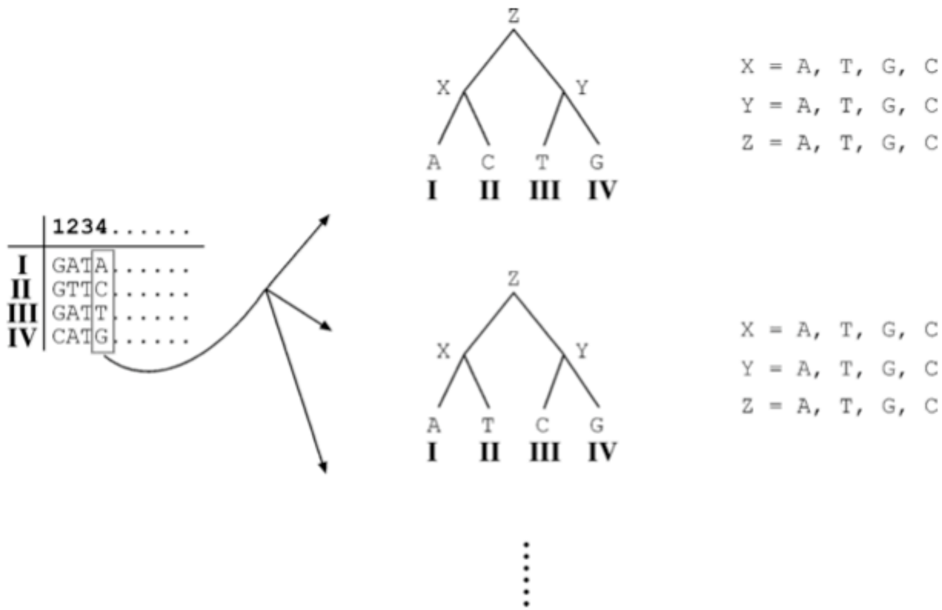


Figure 11.8: Schematic representation of the ML approach to build phylogenetic trees for four taxa, I, II, III, and IV. The ancestral character states at the internal nodes and root node are assigned X, Y, and Z, respectively. The example only shows some of the topologies derived from one of the sites in the original alignment. The method actually uses all the sites in probability calculation for all possible trees with all combinations of possible ancestral sequences at internal nodes according to a predefined substitution model.

$$\mathrm{L}_{(4)} = \Pr(Z \to X) * \Pr(Z \to Y) * \Pr(X \to A) * \Pr(X \to C) * \Pr(Y \to T) * \Pr(Y \to G)$$

$$\ln \mathrm{L}_{(4)} = \ln \Pr(Z \to X) + \ln \Pr(Z \to Y) + \ln \Pr(X \to A) + \ln \Pr(X \to C) + \ln \Pr(Y \to T) + \ln \Pr(Y \to G)$$

Αξιολόγηση του δένδρου

- Bootstrap:
  - Τυχαία δειγματοληψία θέσεων της πολλαπλής στοίχισης.
  - Μια θέση μπορεί να επιλεγεί περισσότερες από μια φορές ή και καμία.
  - Δημιουργία μιας νέας αλλαγμένης πολλαπλής στοίχισης
  - Η διαδικασία επαναλαμβάνεται 100-1000 φορές.
  - Για κάθε νέα πολλαπλή στοίχιση, υπολογίζεται το δένδρο.
  - Τα νέα δένδρα συγχωνεύονται σε ένα νέο δένδρο (consensus tree).
  - Boostrap -> συχνότητα εμφάνισης ενός κόμβου.
  - Bootstrap 70% -> 95% εμπιστοσύνη.
  - Αν η μεθοδολογία δημιουργίας του δένδρου είναι λάθος, μπορεί να πάρουμε υψηλές τιμές bootstrap για το λάθος δένδρο.

bootstrap

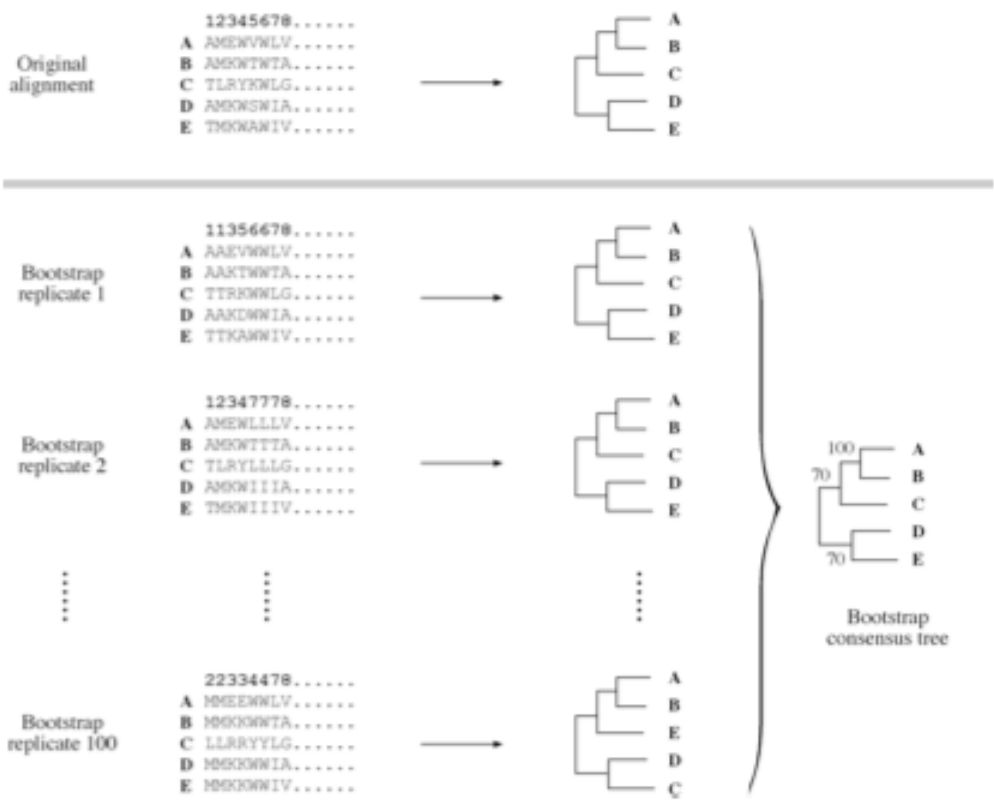


Figure 11.10: Schematic representation of a bootstrap analysis showing the original alignment and modified replicates in which certain sites are randomly replaced with other existing sites. The resulting altered replicates are used to building trees for statistical analysis at each node.

Jacknife

- Το Jacknife είναι παρόμοιο με το bootstrap.
- Επιλέγονται τυχαία (δίχως αντικατάσταση) οι μισές στήλες της πολλαπλής στοίχισης.
- Πρόβλημα: τα νέα δένδρα δημιουργούνται από λιγότερα δεδομένα.

Άσκηση

Άσκηση (2)

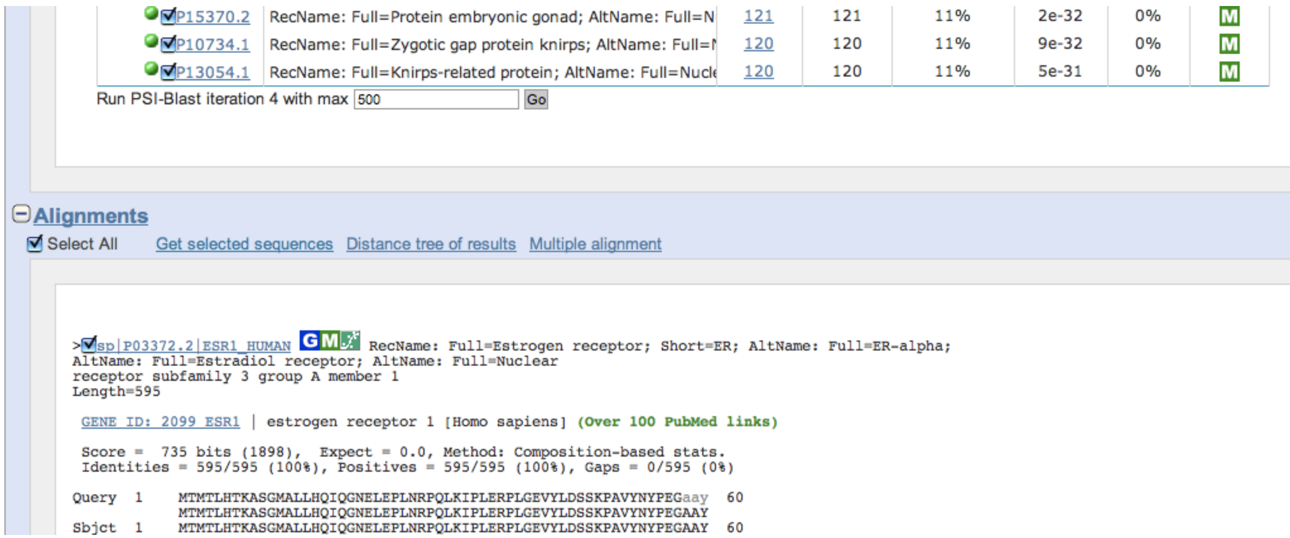
- Βρείτε την πρωτεϊνική ακολουθία του human estrogen receptor alpha (Uniprot id: P03372) σε μορφή FASTA.

- Με την ακολουθία αυτή (P03372), βρείτε τις ομόλογες πρωτεϊνικές ακολουθίες της, στη Drosophila melanogaster και στον άνθρωπο ταυτόχρονα, με τη βοήθεια του PSI-BLAST. Κάνετε το PSI-Blast στην ιστοσελίδα του NCBI, χρησιμοποιώντας την Swissprot, expectation value 1e-10 και low-complexity filtering. Επαναλάβετε τους κύκλους του PSI-blast μέχρι να συγκλίνει ο αλγόριθμος.

- Αποθηκεύεστε σε ένα αρχείο (με όνομα sequences.fasta) με μορφή FASTA τις ακολουθίες από την παραπάνω αναζήτηση.

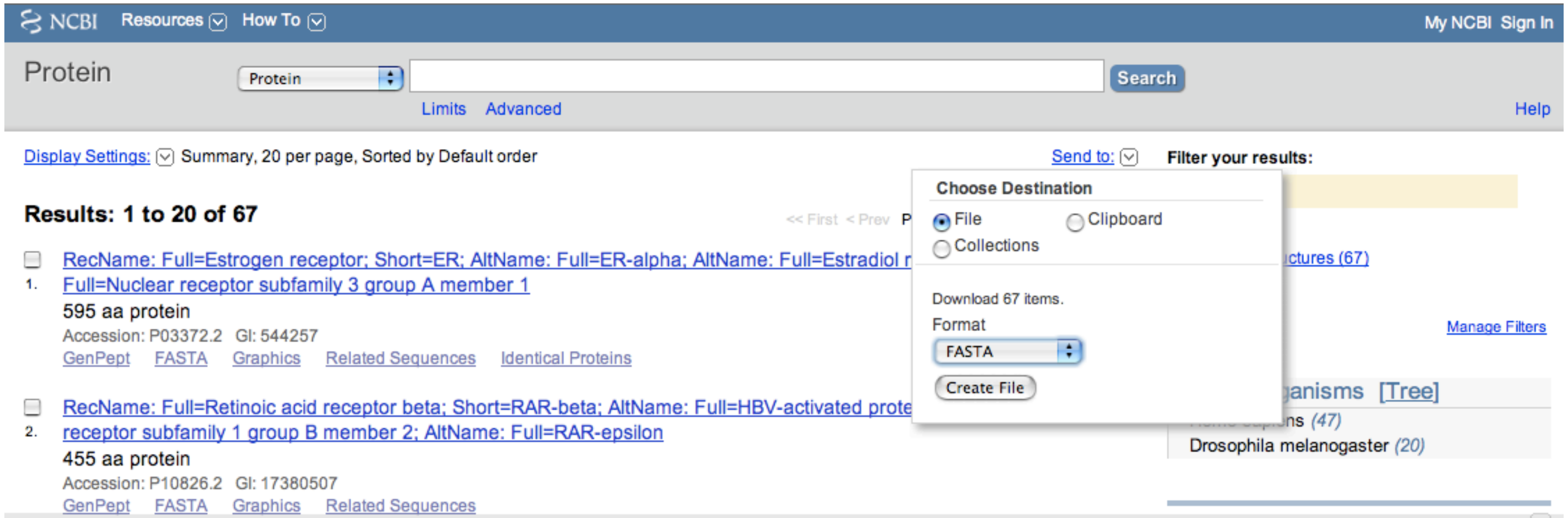
Αποθήκευση ακολουθιών από το Blast

- Select all
- Get selected sequences



Αποθήκευση ακολουθιών από το Blast

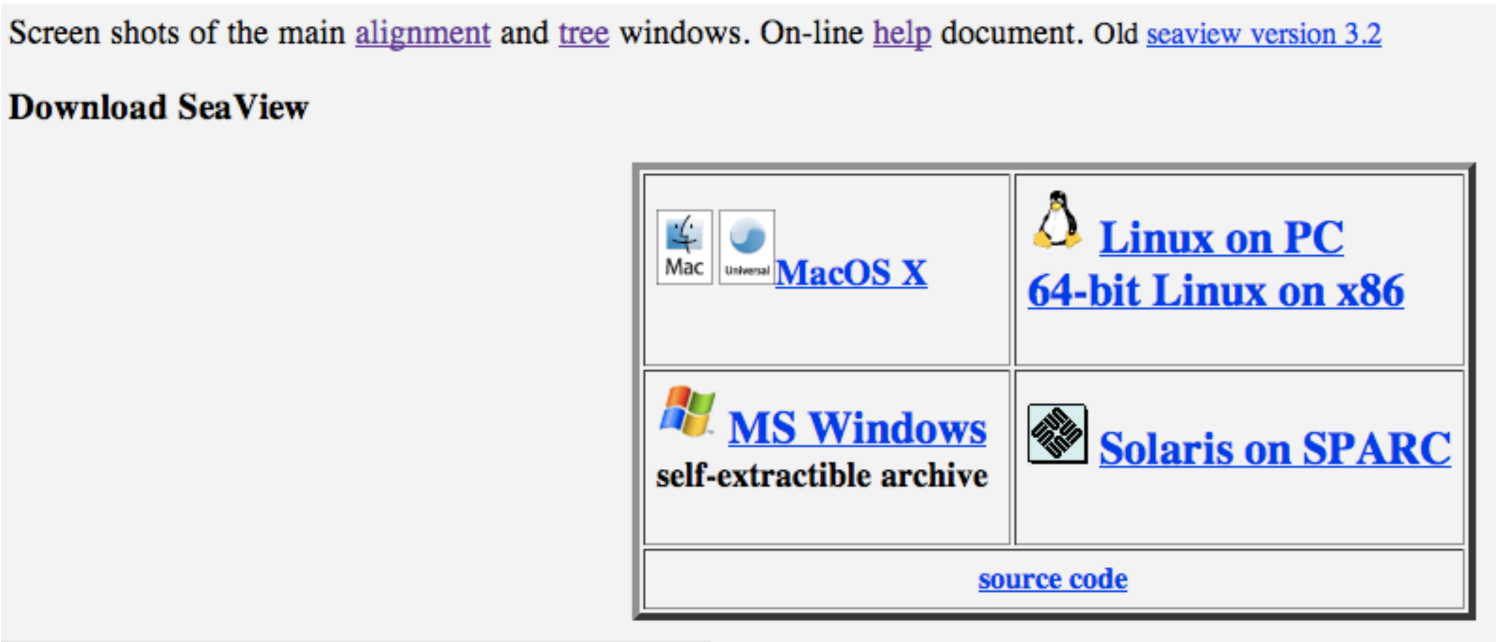
- Send to ->
- File ->
- Format: FASTA ->
- Creat file



Seaview

- ‘Κατεβάστε’ το seaview (MS Windows self-extractible archive) από την διεύθυνση

http://pbil.univ-lyon1.fr/software/seaview.html



- Online help για το πρόγραμμα θα βρείτε στην διεύθυνση

http://pbil.univ-lyon1.fr/software/seaview_data/seaview.html

Άσκηση (2ο)

- Από το Psi-Blast δημιουργήθηκε ένα αρχείο (sequences.fasta) με τις ομόλογες ακολουθίες που βρήκατε.
- Φορτώστε το αρχείο (sequences.fasta) στο πρόγραμμα Seaview.
  - File -> Open -> Fasta
  - Η απλά τραβήξτε το αρχείο μέσα στο seaview.
- Αλλάξτε το όνομα των ακολουθιών.
  - Επιλέξτε την ακολουθία -> Edit -> Rename sequence.

- Κάνετε πολλαπλή στοίχιση των ακολουθιών με το πρόγραμμα muscle.
  - Align -> alignment options -> muscle
  - Align -> Align all

Άσκηση (3)

Απομακρύνετε τις περιοχές που δεν είναι συντηρημένες

- Για να κάνετε Editing την πολλαπλή στοίχιση:
  - Props-> allow seq. editing
  - Επιλέξτε τις ακολουθίες που θέλετε να τροποποιήσετε (σε αυτό το παράδειγμα επιλέξτε όλες τις ακολουθίες).
  - Τοποθετήστε τον κέρσορα μέσα στην πολλαπλή στοίχιση (σε περιοχή που θέλετε να διαγράψετε) και χρησιμοποιήστε το πλήκτρο delete.

Δημιουργήστε το φυλογενετικό δένδρο με τη μέθοδο Neighbor joining

- Trees -> Distance Methods -> NJ (Poisson, ignore all gap sites).
- Στην προηγούμενη εργαστηριακή άσκηση το human estrogen receptor alpha & το Seven-up από τη Drosophila δεν ήταν τα καλύτερα ανταποδοτικά χτυπήματα του Blast. Μπορείτε να καταλάβετε από το φυλογενετικό δένδρο γιατί συνέβη αυτό;