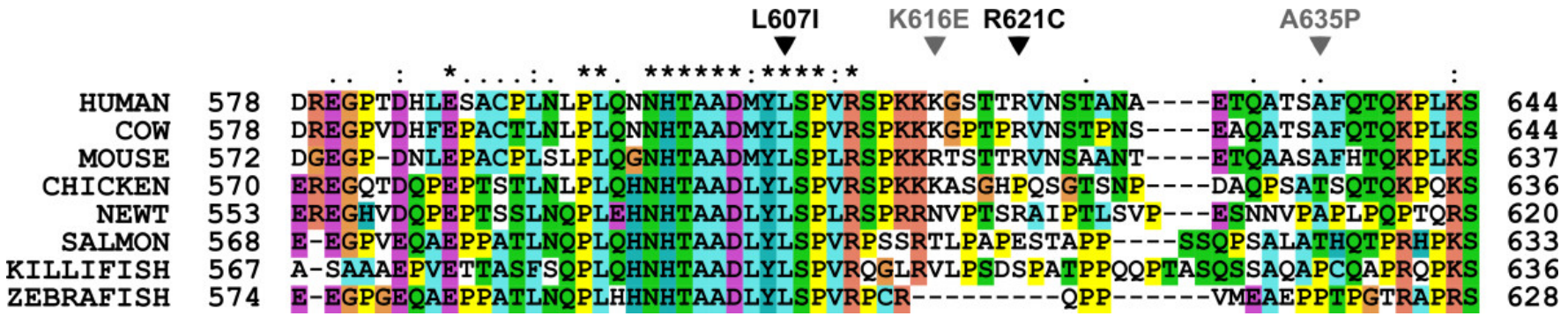
Πολλαπλή στοίχιση - Φυλογένεση

4ο εργαστήριο

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

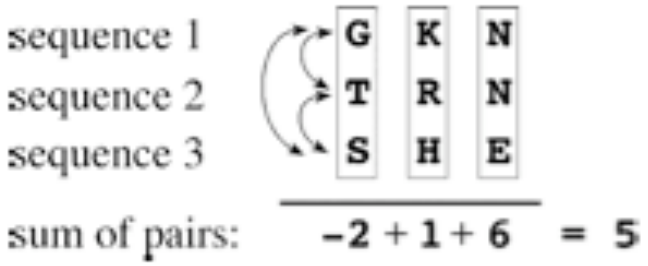


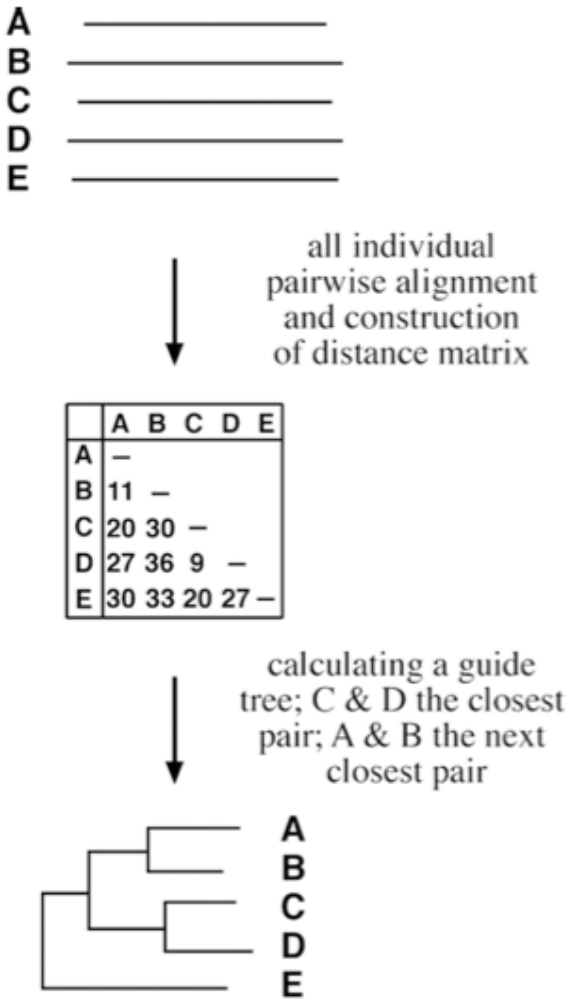
Figure 5.1: Given a multiple alignment of three sequences, the sum of scores is calculated as the sum of the similarity scores of every pair of sequences at each position. The scoring is based on the BLOSUM62 matrix (see Chapter 3). The total score for the alignment is 5, which means that the alignment 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

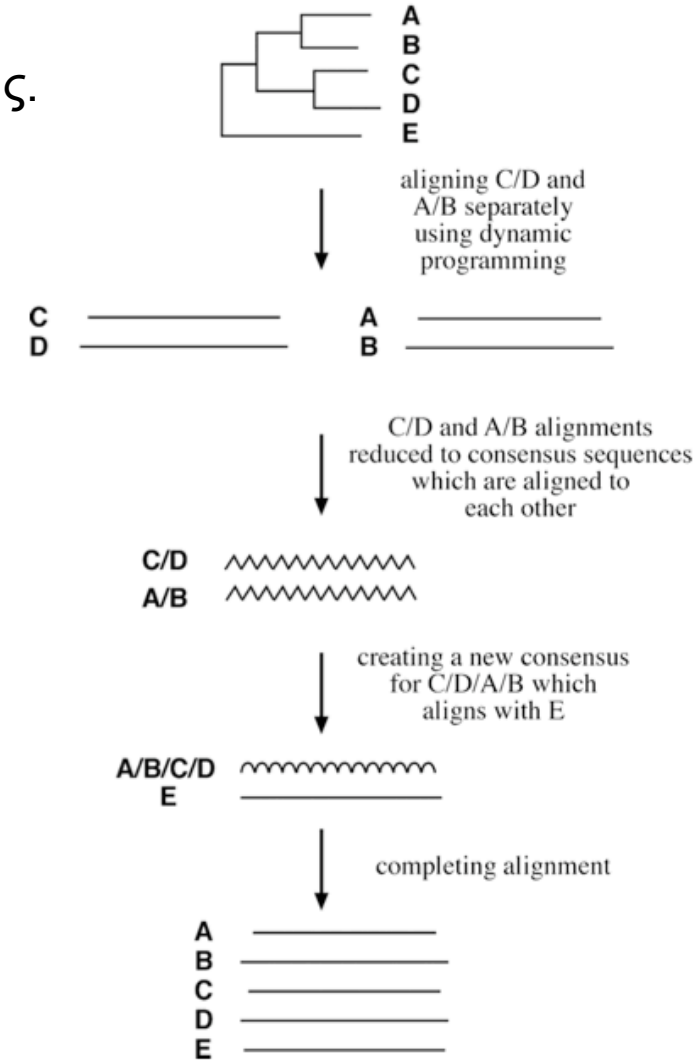
ClustalW (i)

- 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

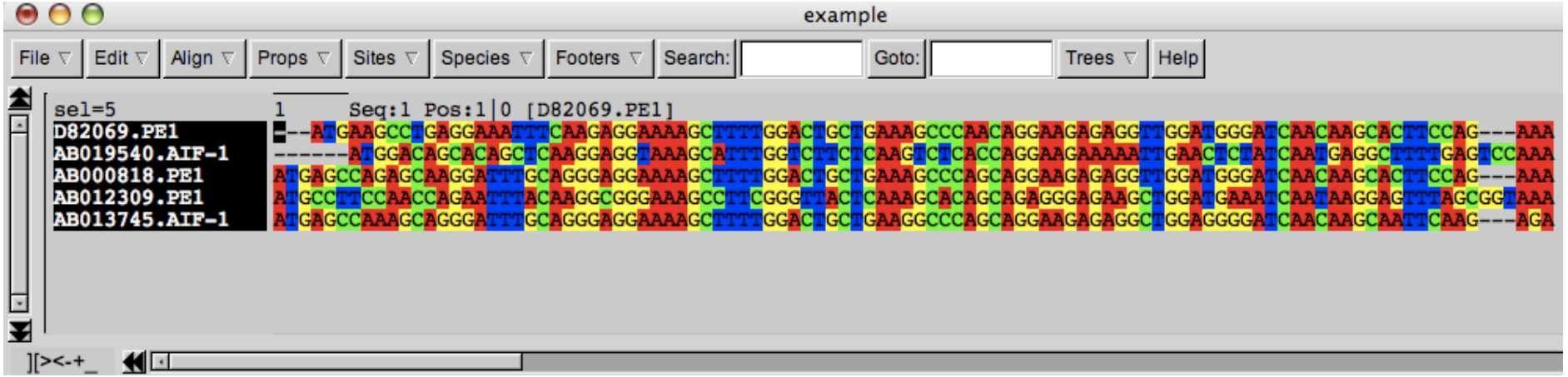
Προβλήματα της προοδευτικής στοίχισης

- 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

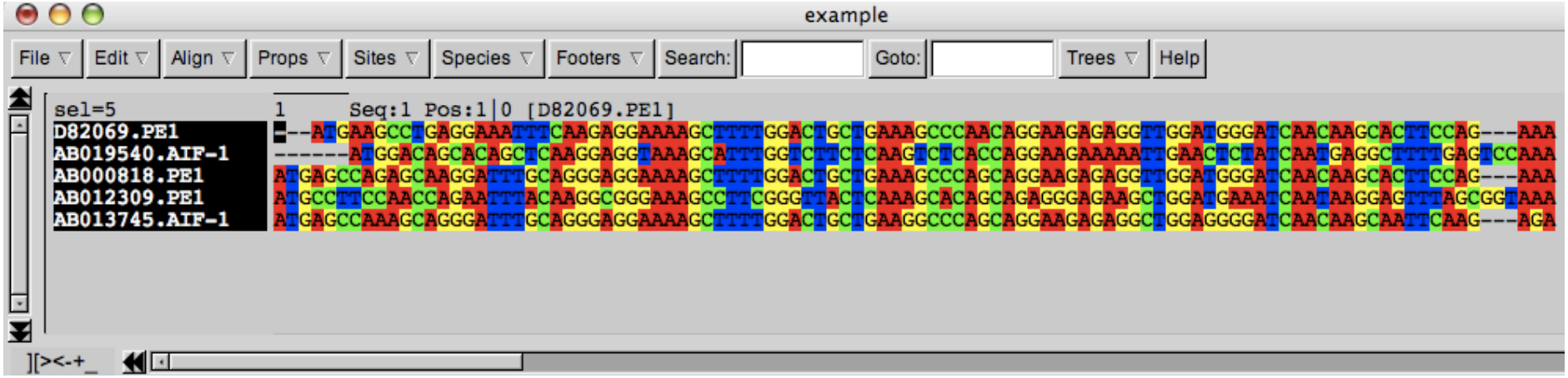
- FASTA (.fa ή .fasta ή .fst)
- Clustal (.aln)
- Phylip (.phy ή .phylip)
- MSF (.msf)
- Mase (.mase)
- Nexus (.nxs)
- Συνήθως, τα alignment editors μπορούν να μετατρέψουν το ένα format σε άλλο.
- Readseq
  - http://www.ebi.ac.uk/cgi-bin/readseq.cgi

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



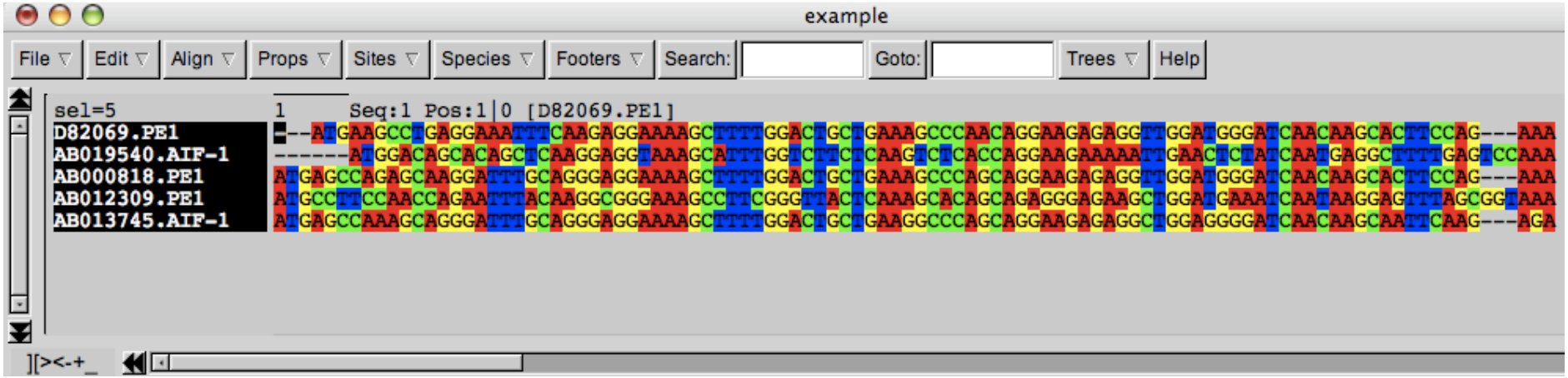

```
CLUSTAL W (1.7) multiple sequence alignment


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

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

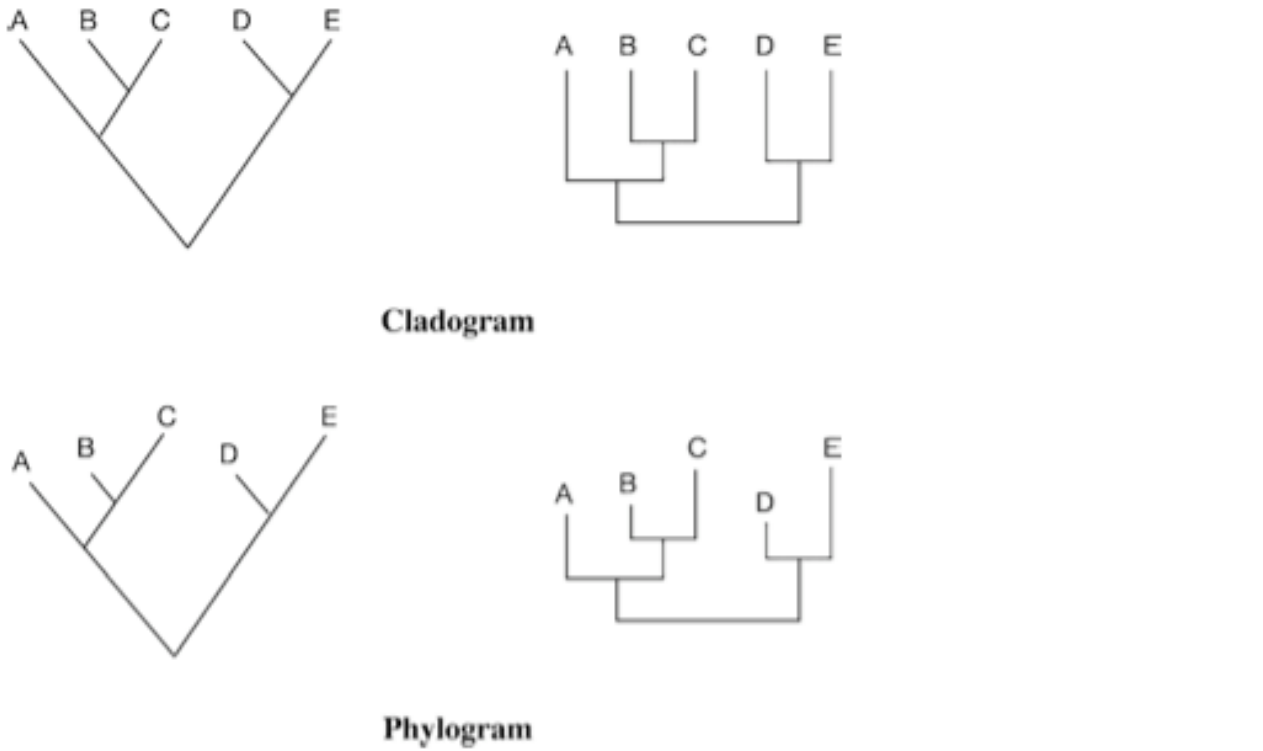


Figure 10.4: Phylogenetic trees drawn as cladograms (*top*) and phylograms (*bottom*). The branch lengths are unscaled in the cladograms and scaled in the phylograms. The trees can be drawn as angled form (*left*) or squared form (*right*).

Λίγη εξέλιξη: ομολογία

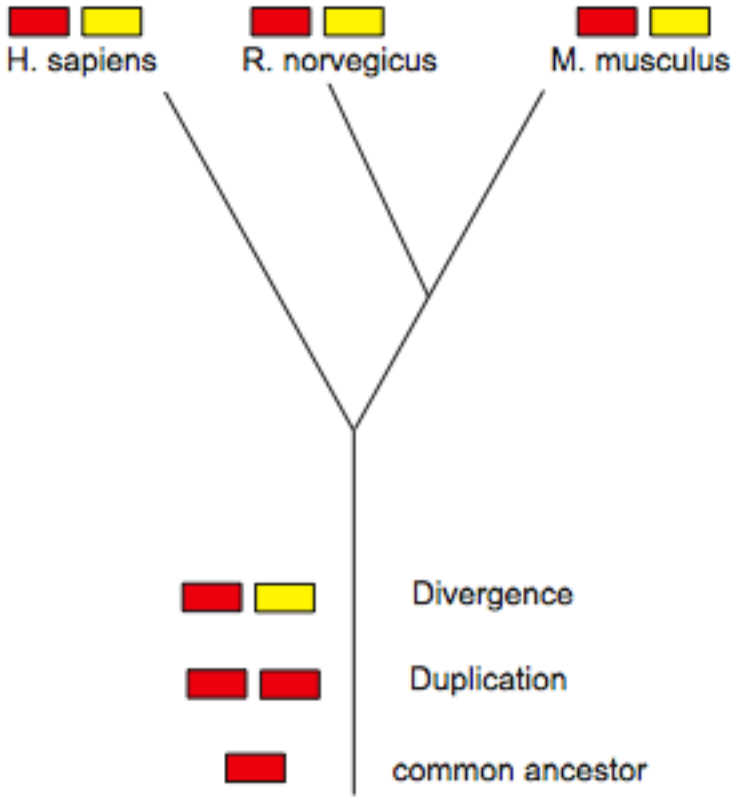
- Ομόλογα γονίδια: κοινός εξελικτικός πρόγονος. Χιμαιρικές πρωτεΐνες;

- 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 (ii)

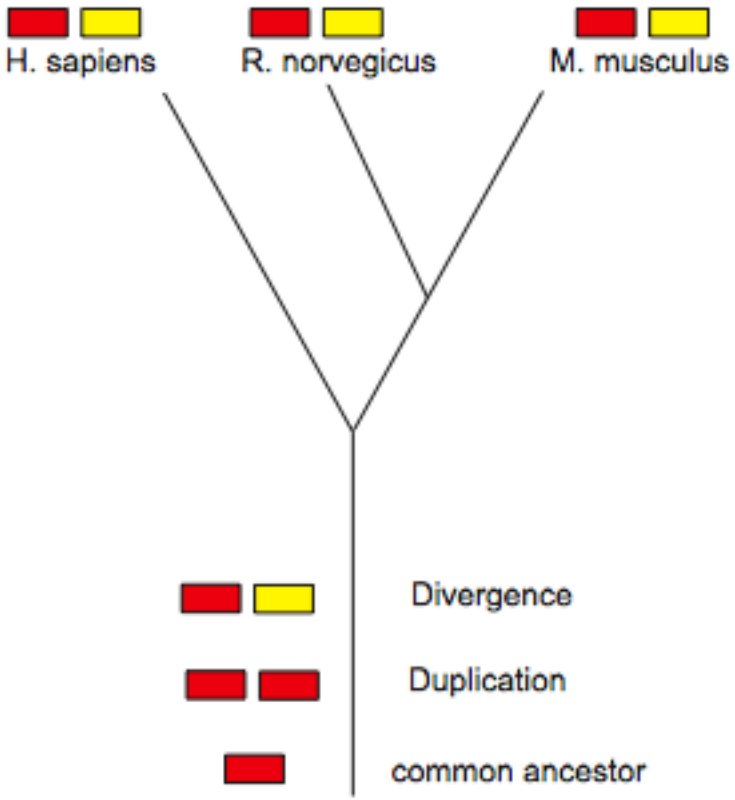


Στάδια φυλογενετικής ανάλυσης

- Εντοπισμός ομόλογων ακολουθιών
  - Π.χ. Blast, HMMs
- Πολλαπλή στοίχιση
  - Διορθώσεις στην στοίχιση
- Υπολογισμός φυλογενετικού δένδρου

Στοιχεία ενός φυλογενετικού δένδρου

- Φύλλα (leafs)
- Βραχίονες (branches)
- Κόμβοι (nodes)
- Κλάδοι (clades)



Δένδρα με/χωρίς ρίζα

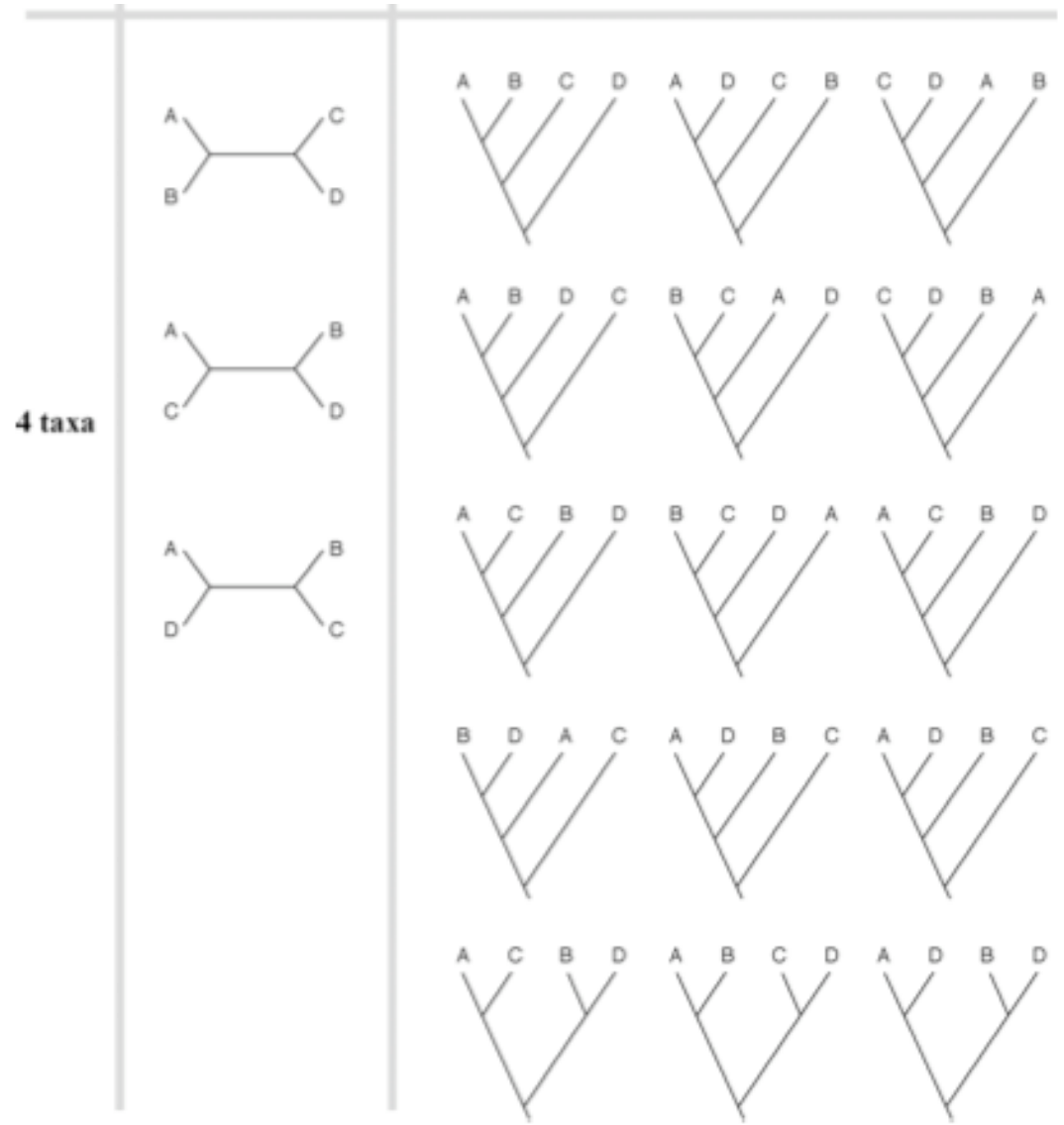


Figure 10.7: All possible tree topologies for three and four taxa. For three taxa, there are one unrooted and three rooted trees. For four taxa, there are three unrooted and fifteen rooted trees.

Μέθοδοι κατασκευής δένδρων

- Μέθοδοι αποστάσεων
  - Ένωση γειτόνων (neighbor joining)
  - UPGMA (unweighted pair group method using arithmetic averages)
  - Λιγότερων τετραγώνων (least squares)
  - Ελάχιστης εξέλιξης (minimum evolution)

Μέθοδοι κατασκευής δένδρων

- Μέθοδοι βασισμένες σε χαρακτήρες (discrete methods).
  - Maximum parsimony:Απαιτεί τον ελάχιστο αριθμό αντικαταστάσεων για την ερμηνεία των ακολουθιών
  - Maximum likelihood: Αναζητά το εξελικτικό μονοπάτι με την μέγιστη πιθανότητα για τα υπάρχοντα δεδομένα

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

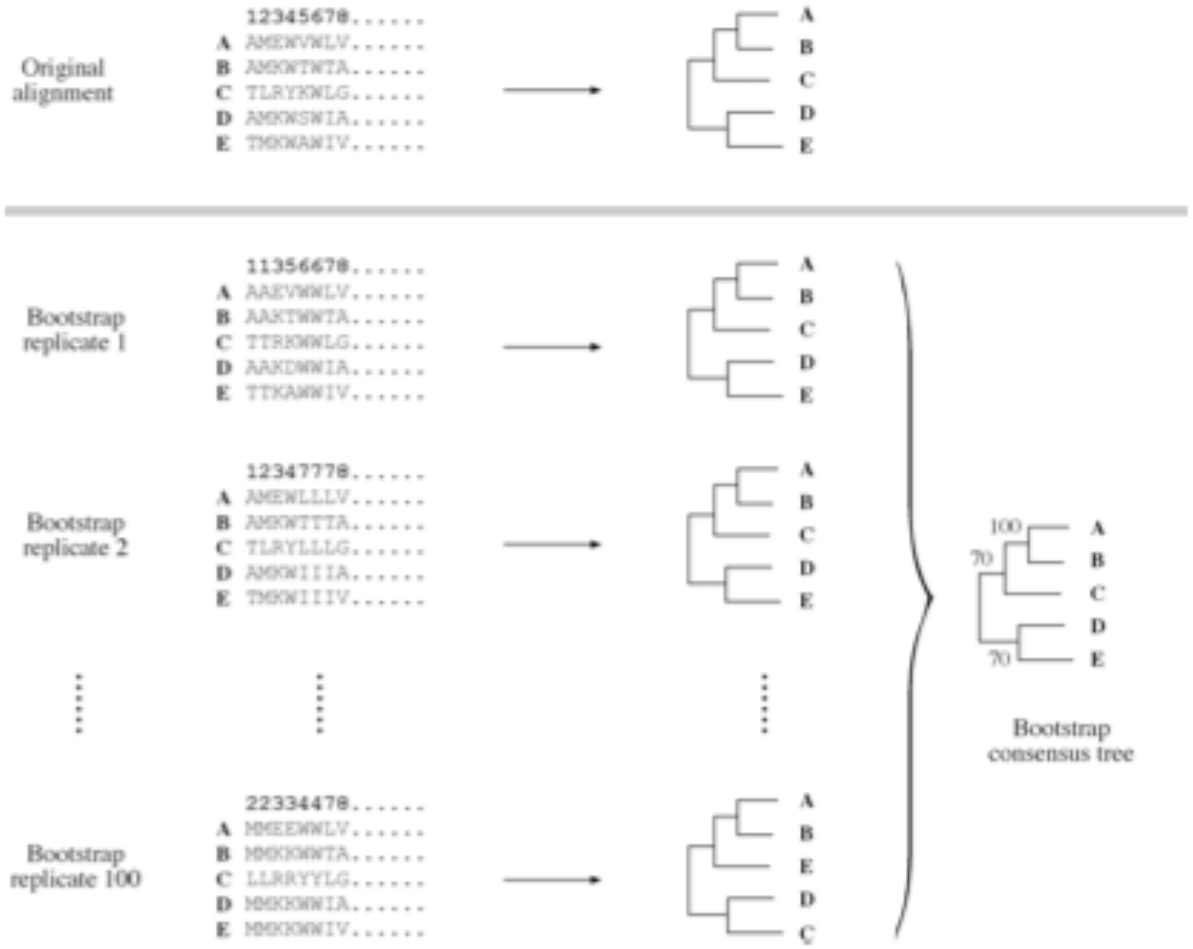


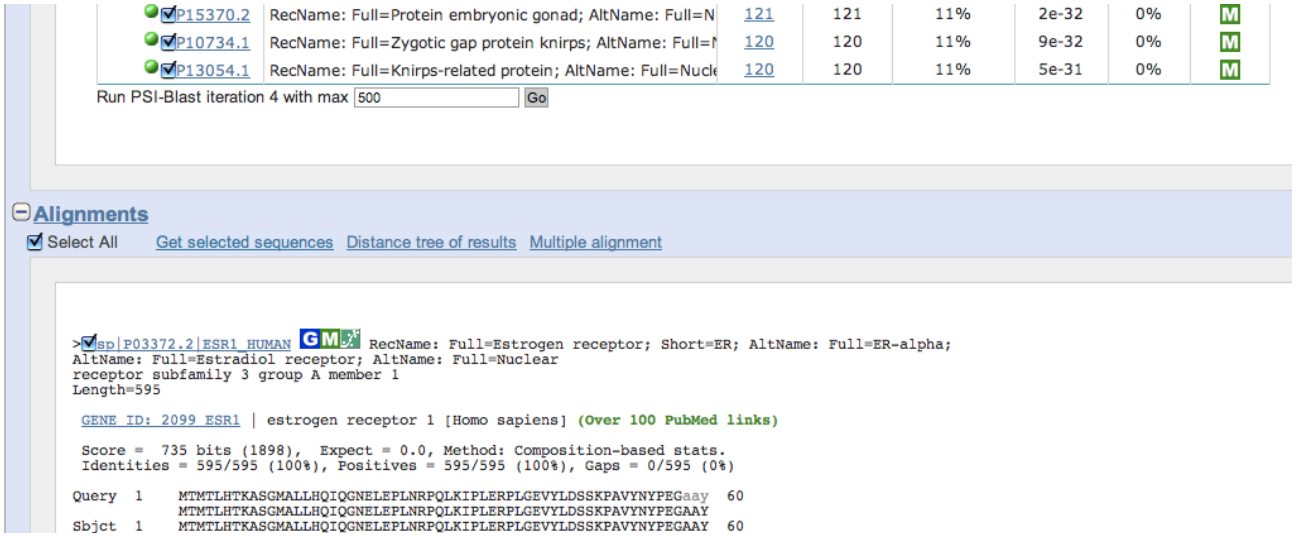
Figure 11.10: Schematic representation of a bootstrap analysis showing the original alignment and modified replicates in which certain sites are randomly replaced with other existing sites. The resulting altered replicates are used to building trees for statistical analysis at each node.

Άσκηση (1)

- 1) Βρείτε την πρωτεϊνική ακολουθία του human estrogen receptor alpha (Uniprot id: P03372) σε μορφή FASTA.
- 2) Με την ακολουθία αυτή (P03372), βρείτε τις ομόλογες πρωτεϊνικές ακολουθίες της, στη Drosophila melanogaster και στον άνθρωπο, 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.
- 3) Αποθηκεύεστε σε ένα αρχείο (με όνομα sequences.fasta) με μορφή FASTA τις ακολουθίες από την παραπάνω αναζήτηση.

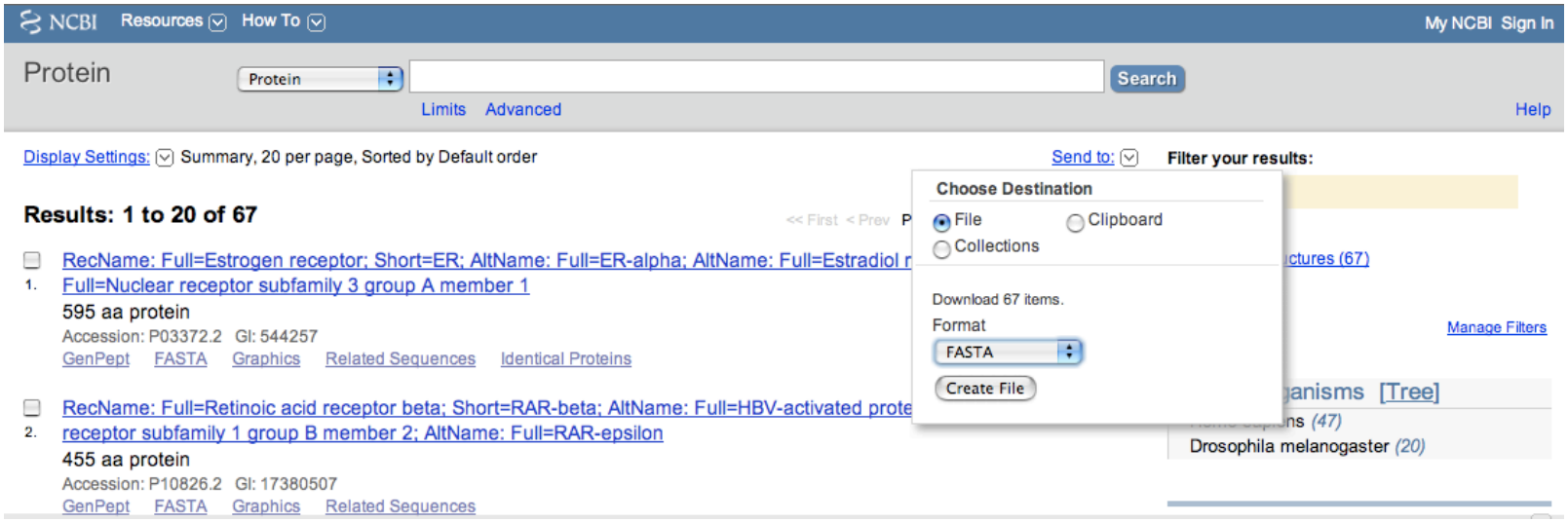
Αποθήκευση ακολουθιών από το Blast

- Select all
- Get selected sequences



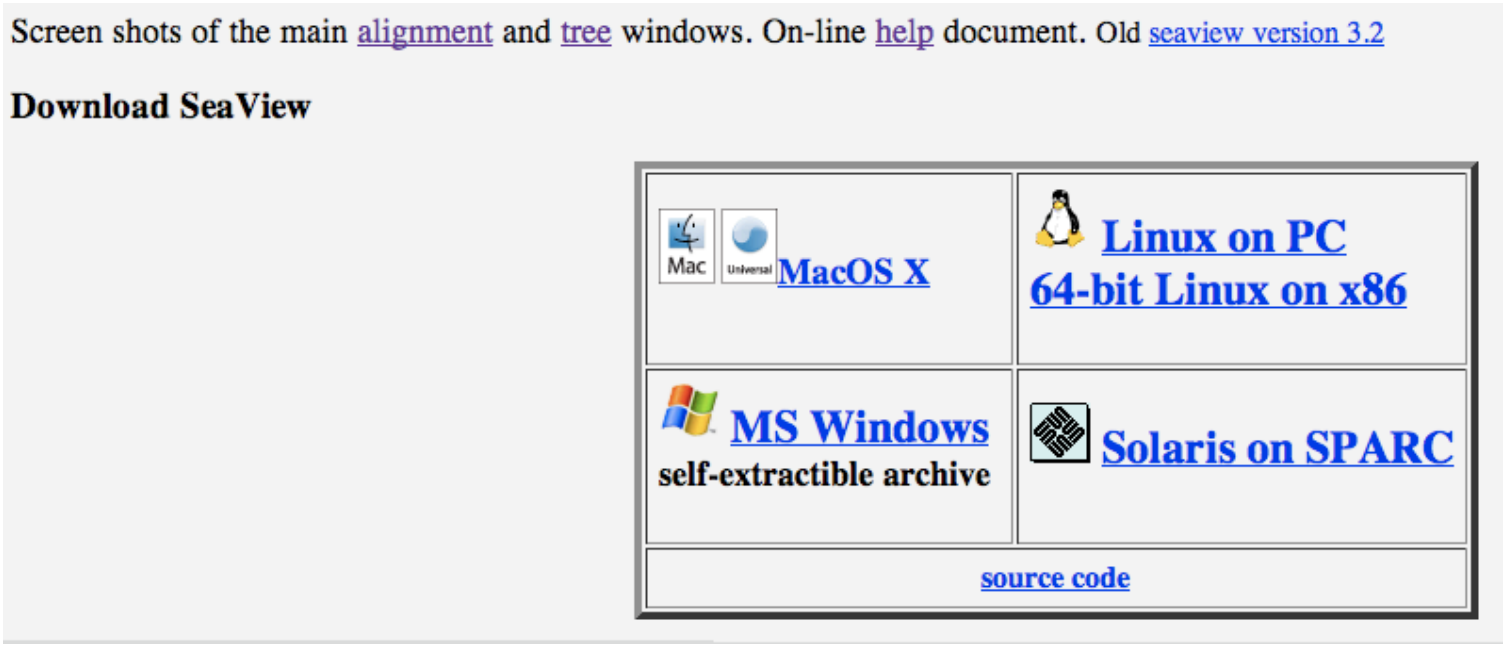
Αποθήκευση ακολουθιών από το Blast

- 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 (2)

- 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 & 100 Boostraps.

- Trees -> Distance Methods -> NJ (Poisson, ignore all gap sites, bootstrap 100).
- 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;