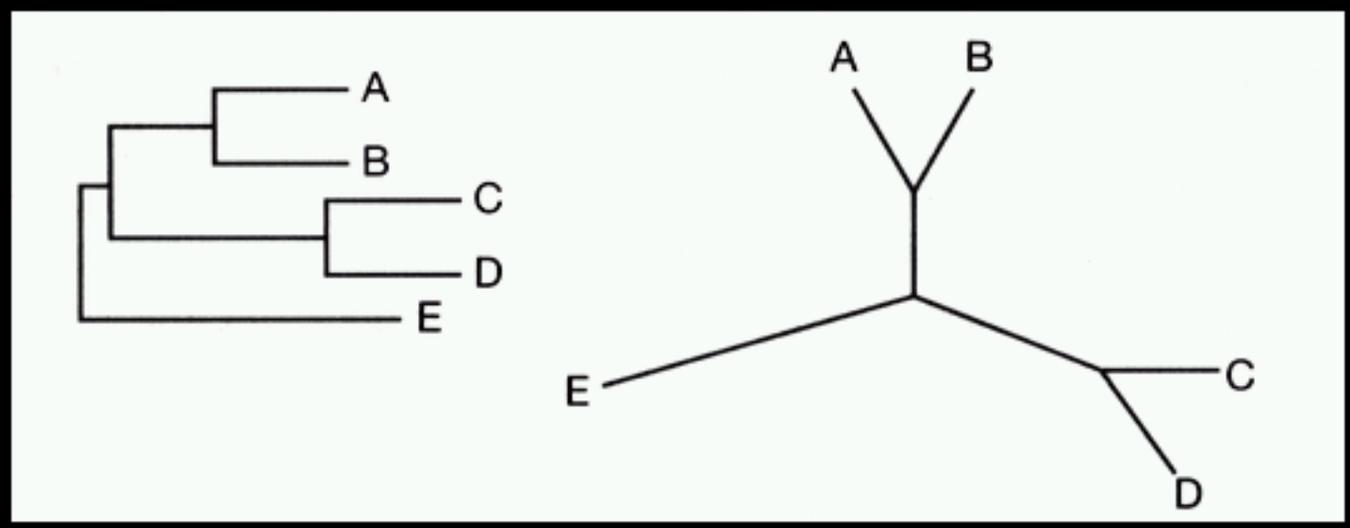
Βιοπληροφορική

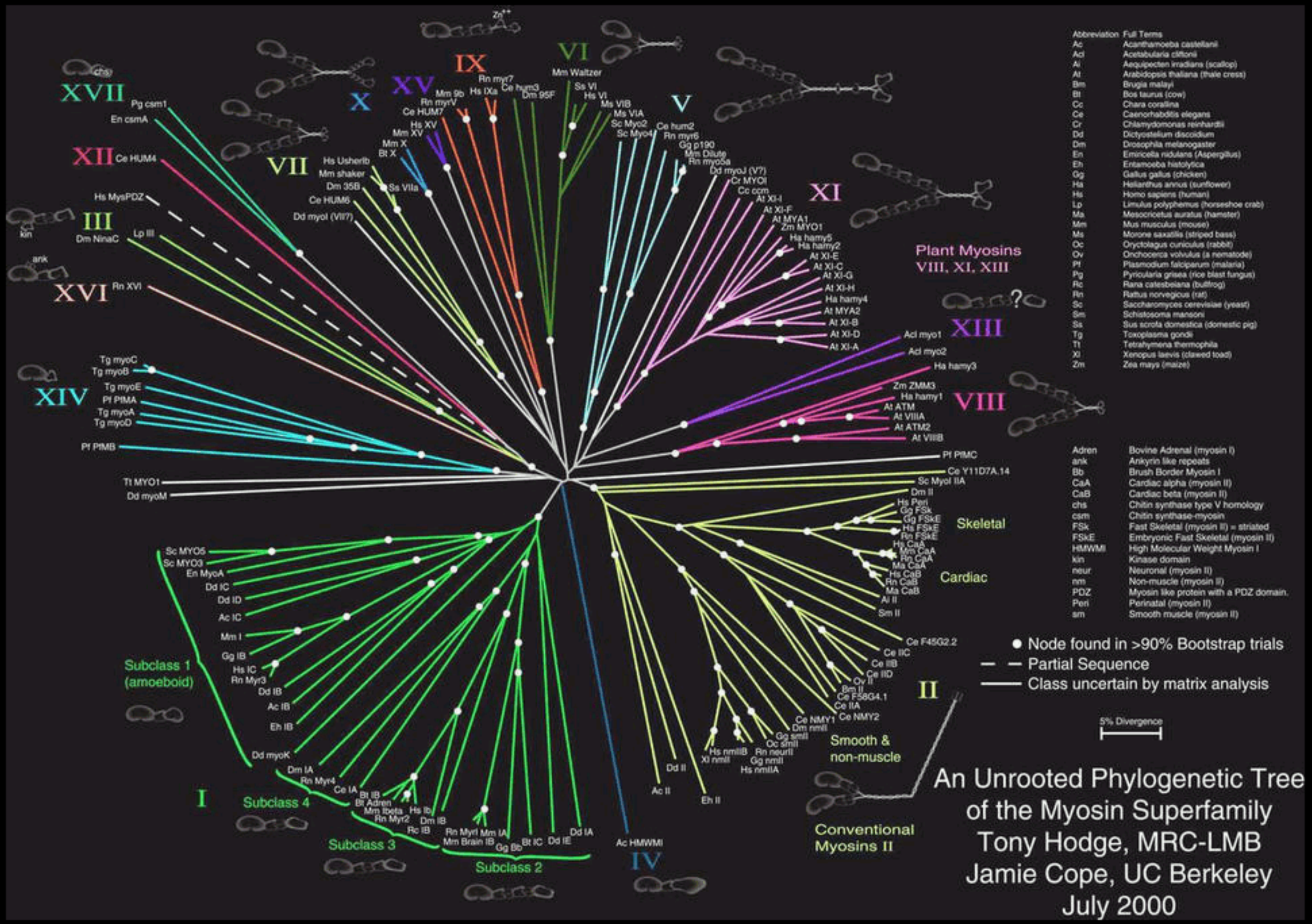
Διάλεξη 7η :

Φυλογενετική ανάλυση αλληλουχιών :
Εισαγωγή, Διαδικασία, Τύποι αλγορίθμων.
Αλγόριθμοι αποστάσεων : UPGMA.

Εισαγωγή

Φυλογενετική είναι η μελέτη εξελικτικών σχέσεων.
Η φυλογενετική ανάλυση χρησιμοποιείται ως το μέσο για τη συναγωγή (ή εκτίμηση) αυτών των σχέσεων.
Η συνηθέστερη μέθοδος για την αναπαράσταση της εξελικτικής ιστορίας (οργανισμών, αλληλουχιών, κοκ), είναι τα φυλογενετικά δένδρα ή δενδρογράμματα.




XVII
Pg csm1
En csmA
XII
Ce HUM4
III
Hs MysPDZ
Dm NinaC
Lp III
XVI
Rn XVI
XIV
Tg myoC
Tg myoB
Tg myoE
Pf PfMA
Tg myoA
Tg myoD
Pf PfMB
Tt MYO1
Dd myoM
X
IX
XV
Rn myr7
Mm 9b
Rn myrV
Ce HUM7
Hs XV
Mm XV
Mm X
Bt X
Hs IXa
Ce hum3
Dm 95F
VII
Hs Usherlb
Mm shaker
Dm 35B
Ce HUM6
Dd myoI (VII?)
Ss VIIa
VI
Mm Waltzer
Ss VI
Hs VI
Ms VIB
Ms VIA
Sc Myo2
Sc Myo4
V
Ce hum2
Rn myr6
Gg p190
Mm Dilute
Rn myo5a
Dd myoJ (V?)
Cr MYOI
Cc ccm
XI
At XI-I
At XI-F
At MYA1
Zm MYO1
Ha hamy5
Ha hamy2
At XI-E
At XI-C
At XI-G
At XI-H
Ha hamy4
At MYA2
At XI-B
At XI-D
At XI-A
Plant Myosins
VIII, XI, XIII
XIII
Acl myo1
Acl myo2
Ha hamy3
VIII
Zm ZMM3
Ha hamy1
At ATM
At VIIIA
At ATM2
At VIIIB
Pf PfMC
Ce Y11D7A.14
Sc Myo1 IIA
Dm II
Hs Peri
Gg FSk
Gg FSkE
Hs FSkE
Rn FSkE
Skeletal
Hs CaA
Mm CaA
Rn CaA
Ma CaA
Hs CaB
Rn CaB
Ma CaB
Cardiac
Al II
Sm II
Ce F45G2.2
Ce IIC
Ce IIB
Ce IID
Ov II
Bm II
Ce F58G4.1
Ce IIA
Ce NMY2
Ce NMY1
Dm nmII
Gg smII
Cc smII
Rn neurII
Gg nmII
Hs nmIIA
Hs nmIIB
Xl nmII
Dd II
Ac II
Eh II
II
Smooth &
non-muscle
Conventional
Myosins II
Sc MYO5
Sc MYO3
En MycA
Dd IC
Dd ID
Ac IC
Mm I
Gg IB
Hs IC
Rn Myr3
Dd IB
Ac IB
Eh IB
Dd myoK
Subclass 1
(amoeboid)
I
Dm IA
Rn Myr4
Ce IA
Bt IB
Bt Adren
Mm Ibeta
Rn Myr2
Rc IB
Hs Ib
Dm IB
Subclass 4
Subclass 3
Rn Myr1
Mm IA
Mm Brain IB
Gg Bb
Bt IC
Dd IE
Dd IA
Subclass 2
Ac HMWMI
IV
Abbreviation Full Terms
Ac Acanthamoeba castellanii
Acl Acetabularia cliftonii
Ai Aequipecten irradians (scallop)
At Arabidopsis thaliana (thale cress)
Bm Brugia malayi
Bt Bos taurus (cow)
Cc Chara corallina
Ce Caenorhabditis elegans
Cr Chlamydomonas reinhardtii
Dd Dictyostelium discoidium
Dm Drosophila melanogaster
En Emericella nidulans (Aspergillus)
Eh Entamoeba histolytica
Gg Gallus gallus (chicken)
Ha Helianthus annus (sunflower)
Hs Homo sapiens (human)
Lp Limulus polyphemus (horseshoe crab)
Ma Mesocricetus auratus (hamster)
Mm Mus musculus (mouse)
Ms Morone saxatilis (striped bass)
Oc Oryctolagus cuniculus (rabbit)
Ov Onchocerca volvulus (a nematode)
Pf Plasmodium falciparum (malaria)
Pg Pyricularia grisea (rice blast fungus)
Rc Rana catesbeiana (bullfrog)
Rn Rattus norvegicus (rat)
Sc Saccharomyces cerevisiae (yeast)
Sm Schistosoma mansoni
Ss Sus scrofa domestica (domestic pig)
Tg Toxoplasma gondii
Tt Tetrahymena thermophila
Xl Xenopus laevis (clawed toad)
Zm Zea mays (maize)
Adren Bovine Adrenal (myosin I)
ank Ankyrin like repeats
Bb Brush Border Myosin I
CaA Cardiac alpha (myosin II)
CaB Cardiac beta (myosin II)
chs Chitin synthase type V homology
csm Chitin synthase-myosin
FSk Fast Skeletal (myosin II) = striated
FSkE Embryonic Fast Skeletal (myosin II)
HMWMI High Molecular Weight Myosin I
kin Kinase domain
neur Neuronal (myosin II)
nm Non-muscle (myosin II)
PDZ Myosin like protein with a PDZ domain.
Peri Perinatal (myosin II)
sm Smooth muscle (myosin II)
Node found in >90% Bootstrap trials
Partial Sequence
Class uncertain by matrix analysis
5% Divergence
An Unrooted Phylogenetic Tree
of the Myosin Superfamily
Tony Hodge, MRC-LMB
Jamie Cope, UC Berkeley
July 2000

Εισαγωγή

Ένα φυλογενετικό δένδρο είναι ένα συνδεδεμένο μη κυκλικό δυαδικό γράφημα το οποίο αποτελείται από :

- Ένα σύνολο εσωτερικών και εξωτερικών κόμβων οι οποίοι αντιπροσωπεύουν τάξα.
- Ένα σύνολο από κλάδους οι οποίοι συνδέουν τα τάξα, έτσι ώστε (1) ο κάθε κόμβος να συνδέεται με κάθε άλλο κόμβο μέσω μίας και μόνο μίας διαδρομής, και, (2) τα μήκη των κόμβων να αντιπροσωπεύουν αποστάσεις (χρονικές, μοριακών διαφορών, κοκ.).

Εισαγωγή

Οι συνηθισμένες υποθέσεις στις οποίες βασίζεται η φυλογενετική ανάλυση (στα πλαίσια της εξελικτικής θεωρίας) είναι :

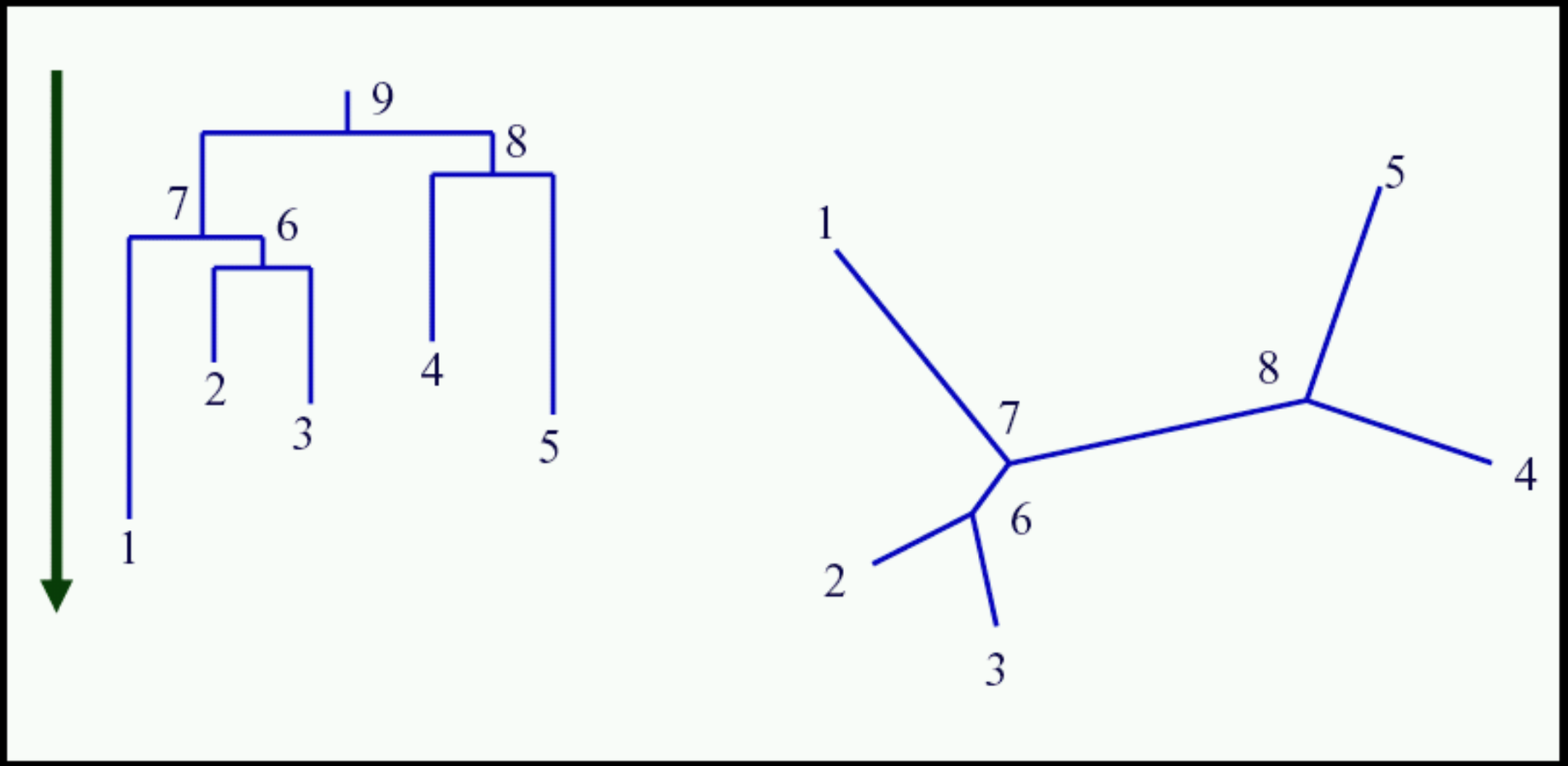
- Κάθε ομάδα οργανισμών προέρχεται από κάποιο κοινό πρόγονο.
- Η μεταβολή των χαρακτηριστικών γίνεται προοδευτικά με την πάροδο του χρόνου (δηλαδή η μεταβολή χαρακτηριστικών είναι μονότονη συνάρτηση του χρόνου).
- Οι φυλογενετικές σχέσεις (και τα αντίστοιχα δενδρογράμματα) είναι δυαδικά (κάθε κόμβος χωρίζεται σε δύο κλάδους).

Εισαγωγή

- Κλάδος ενός δενδρογράμματος είναι ένα μονοφυλετικό τάξο, δηλ. ομάδες οργανισμών ή αλληλουχιών οι οποίες περιλαμβάνουν τον πλέον πρόσφατο εξελικτικό τους πρόγονο και όλους τους απογόνους αυτού του προγόνου.
- Συχνά, το μήκος των διακλαδώσεων ενός δενδρογράμματος είναι ανάλογο της εξελικτικής απόκλισης μεταξύ των κλάδων του.

Εισαγωγή

Δένδρα με ή χωρίς ρίζα (Rooted vs. unrooted trees)



Εξελικτικό μοντέλο

Η πράξη εύρεσης ενός φυλογενετικού δένδρου από ένα σύνολο στοιχισμένων αλληλουχιών προϋποθέτει ένα εξελικτικό μοντέλο του οποίου οι βασικές υποθέσεις είναι :

- Οι αλληλουχίες είναι ομόλογες.
- Κάθε θέση ή τμήμα των στοιχισμένων αλληλουχιών είναι ομόλογο με κάθε άλλη θέση ή τμήμα της στοίχισης.
- Οι αλληλουχίες έχουν κοινή φυλογενετική ιστορία (δεν ισχύει, για παράδειγμα, στη σύγκριση ανάμεσα σε πυρηνικές - μιτοχονδριακές αλληλουχίες).
- Το ταξινομικό εύρος των αλληλουχιών αρκεί για την προσδοκώμενη φυλογενετική ανάλυση.

Εξελικτικό μοντέλο

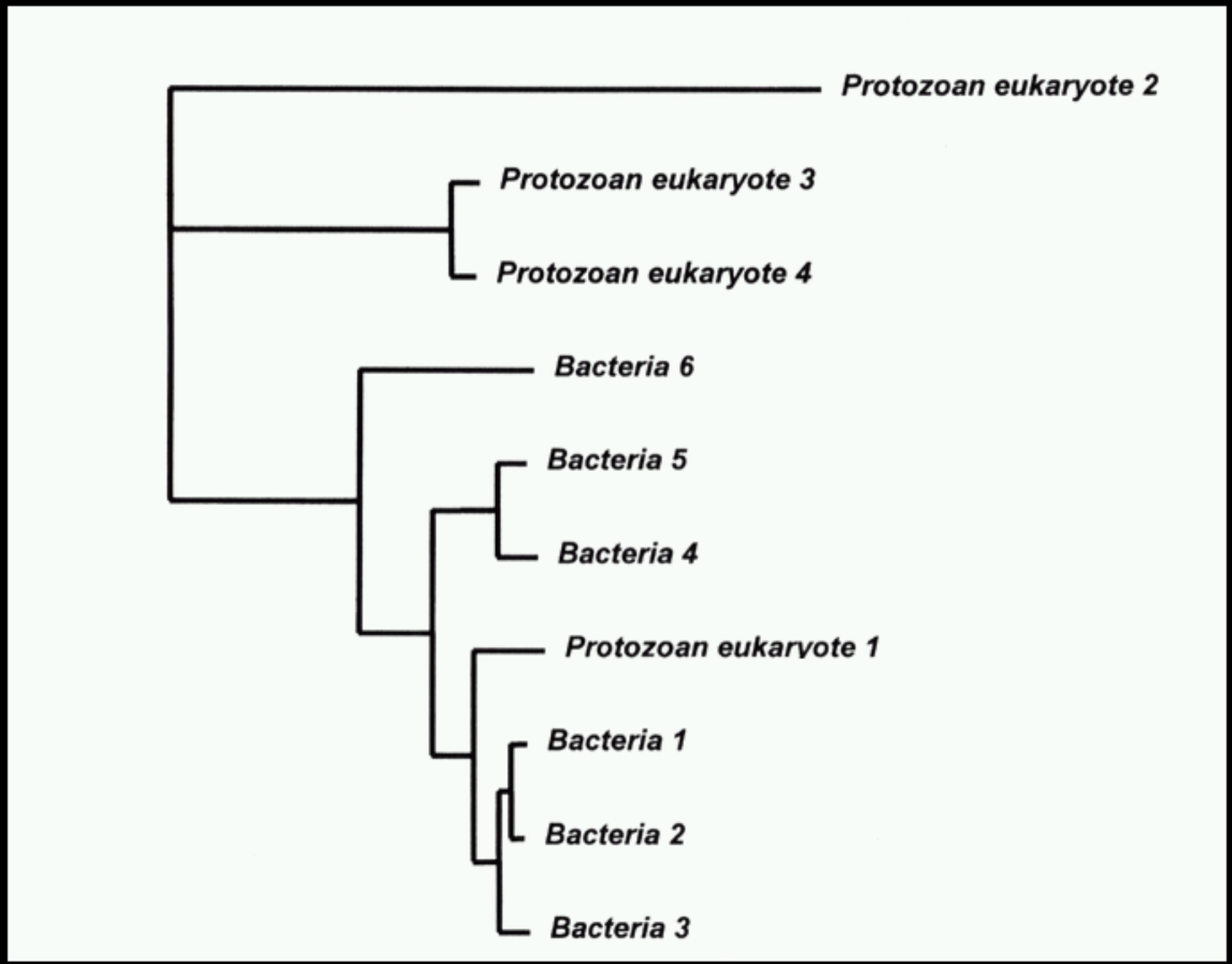
- Η ποικιλότητα των αλληλουχιών που χρησιμοποιούνται για την ανάλυση είναι αντιπροσωπευτική της ποικιλότητας (σε επίπεδο αλληλουχιών) που υπάρχει στα τάξα υπό έλεγχο.
- Η ποικιλότητα των αλληλουχιών του δείγματος εμπεριέχει αρκετή φυλογενετική πληροφορία ώστε να μπορεί να πραγματοποιηθεί η προσδοκώμενη φυλογενετική ανάλυση.

Εξελικτικό μοντέλο

Μερικές από τις μεθόδους που θα περιγραφούν θέτουν και επιπλέον περιορισμούς, όπως για παράδειγμα :

- Οι αλληλουχίες του δείγματος θεωρείται ότι εξελίχθηκαν μέσω μίας τυχαίας (στοχαστικής) διαδικασίας, η οποία ήταν η ίδια για όλες τις αλληλουχίες.
- Όλες οι θέσεις των αλληλουχιών θεωρείται ότι εξελίχθηκαν μέσω της ίδιας τυχαίας διαδικασίας.
- Κάθε θέση των αλληλουχιών θεωρείται ότι εξελίχθηκε ανεξάρτητα από τις υπόλοιπες θέσεις.

Εξελικτικό μοντέλο



Διαδικασία ανάλυσης

Η φυλογενετική ανάλυση αλληλουχιών μπορεί να χωριστεί σε πέντε στάδια, αν και μερικά από αυτά είναι αλληλοεξαρτώμενα και μπορεί να επαναλαμβάνονται κυκλικά μέχρις συγκλίσεως. Τα στάδια είναι :

- Στοίχιση αλληλουχιών.
- Μετατροπή της στοίχισης πολλών αλληλουχιών ώστε να γίνει κατάλληλη για φυλογενετική ανάλυση.
- Προσδιορισμός του μοντέλου υποκατάστασης.
- Κατασκευή δένδρου.
- Εκτίμηση δενδρογράμματος.

Στοίχιση αλληλουχιών

Ενώ στη χρήση των στοιχίσεων για την εύρεση ομολόγων πρωτεϊνών η έμφαση ήταν στην ομοιότητα (π.χ. % ταυτότητα) των αλληλουχιών, για την φυλογενετική ανάλυση η κυρίως πληροφορία βρίσκεται όχι τόσο στις ταυτόσημες υπακολουθίες, αλλά στις θέσεις της στοίχισης που οι αλληλουχίες διαφέρουν (χωρίς να παύουν, βέβαια, να είναι ομόλογες). Τυπικά, η αρχική στοίχιση των αλληλουχιών γίνεται με προγράμματα στοίχισης πολλών αλληλουχιών όπως το CLUSTALW. Ειδικά για την περίπτωση στοιχίσεων που θα χρησιμοποιηθούν για φυλογενετική ανάλυση, οι ιεραρχικές μέθοδοι στοίχισης είναι προτιμώμενες.

Στοίχιση αλληλουχιών

Οι σημαντικότεροι παράμετροι της στοίχισης φαίνεται να είναι τα gap penalties. Ιδιαίτερα για στοιχίσεις που θα χρησιμοποιηθούν για τη δημιουργία φυλογενετικών δένδρων τα μεταβαλλόμενα gap penalties (αυτά που εξαρτώνται την απόκλιση των αλληλουχιών) φαίνεται να είναι ιδιαίτερα χρήσιμα. Ευρείας αποδοχής είναι επίσης μέθοδοι που μειώνουν τη συνεισφορά στην ολική βαθμολογία της στοίχισης εκείνων των αλληλουχιών που είναι στενά συσχετιζόμενες και σε πολλά αντίγραφα (με αυτόν τον τρόπο μειώνεται η πιθανότητα να καθορίσουν οι συγγενείς αλληλουχίες τη στοίχιση των πλέον απομακρυσμένων).

Στοίχιση & δενδρογράμματα

Όπως αναφέρθηκε σε προηγούμενη διάλεξη, οι ιεραρχικές μέθοδοι στοίχισης χρησιμοποιούν ένα δενδρόγραμμα-οδηγό (βασισμένο στις ανά ζεύγη στοιχίσεις των αλληλουχιών) προκειμένου να πραγματοποιήσουν τη στοίχιση. Έχει δειχτεί θεωρητικά ότι το διπλό πρόβλημα στοίχιση/φυλογενετική ανάλυση μπορεί να λυθεί μέσω της ταυτόχρονης (επαναληπτικής) βελτιστοποίησης και των δύο. Αλγόριθμοι που προσπαθούν ταυτόχρονα να βελτιστοποιήσουν και τη στοίχιση και το δενδρόγραμμα έχουν περιγραφεί (π.χ. προγράμματα Malign & TreeAlign) αλλά δεν τυγχάνουν ευρείας χρήσης λόγω της πιθανότητας σύγκλισης τους σε λάθος ή ελλιπείς λύσεις.

Alignment surgery

Αυτό είναι το τμήμα της διαδικασίας φυλογενετικής ανάλυσης αλληλουχιών που συνήθως απαιτείται τόσο η ανθρώπινη παρέμβαση, όσο και η κατανόηση της βιολογικής σημασίας των δεδομένων και του εξελικτικού προβλήματος. Σε αυτό το στάδιο γίνεται η επεξεργασία της στοίχισης σκοπός της οποίας είναι : (1) να απαλείφουν από τη στοίχιση τμήματα αμφίβολης αξίας (βιολογικής σημασίας), και, (2) να μετατραπούν τα gaps (indels) ανάλογα με το ποιά μέθοδος θα χρησιμοποιηθεί για την κατασκευή του δένδρου.

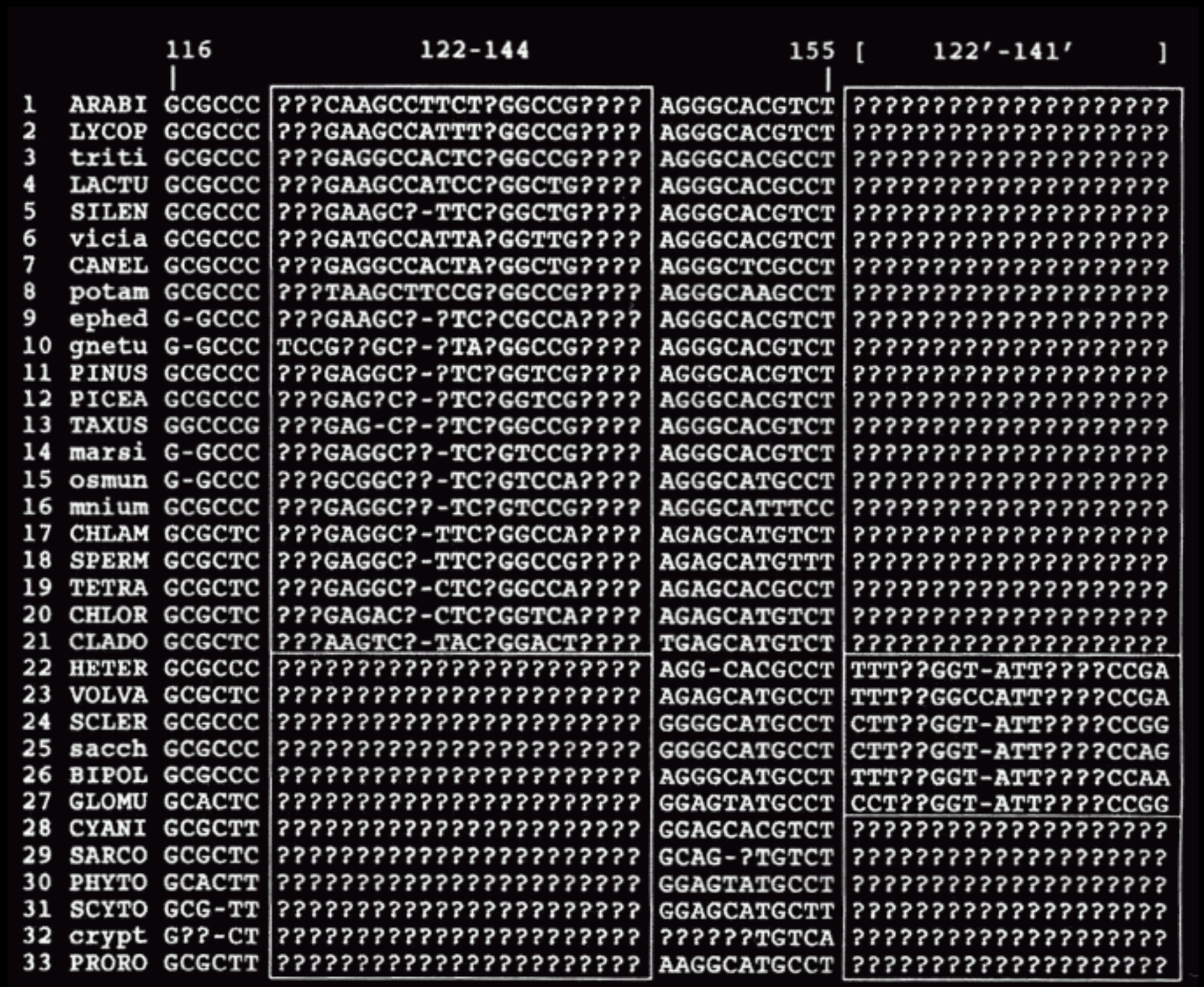
Χειρισμός κενών

Η πλέον ακραίας μορφής επέμβαση είναι η αφαίρεση από τη στοίχιση όλων των θέσεων που περιλαμβάνουν κενά. Αυτός ο χειρισμός αποφεύγει το πρόβλημα της εκτίμησης των σχέσεων μεταξύ των αλληλουχιών χωρίς τις επιπλοκές από την παρουσία indels. Το εμφανές πρόβλημα είναι ότι όλη η φυλογενετική πληροφορία που περιέχεται στις αντίστοιχες περιοχές δεν χρησιμοποιείται. Μερικές από τις μεθόδους κατασκευής δένδρων μπορούν να χρησιμοποιήσουν τα κενά ως ξεχωριστούς χαρακτήρες ('5ος' τύπος νουκλεοτιδίων, ή '21ο' αμινοξύ). Σε αυτές τις περιπτώσεις χρειάζεται προσοχή στην αφαίρεση διαδοχικών κενών από τη στοίχιση.

Alignment surgery

```
                 116          122-144              155
                 |                                   |
1  ARABI  GCGCCC ---CAAGCCTTCT-GGCCG----  AGGGCACGTCT
2  LYCOP  .....C ---GAAGCCATTT-GGCCG----  A..........
3  triti  .....C ---GAGGCCACTC-GGCCG----  A.......C..
4  LACTU  .....C ---GAAGCCATCC-GGCTG----  A.......C..
5  SILEN  .....C ---GAAGC--TTC-GGCTG----  A..........
6  vicia  .....C ---GATGCCATTA-GGTTG----  A..........
7  CANEL  .....C ---GAGGCCACTA-GGCTG----  A....T..C..
8  potam  .....C ---TAAGCTTCCG-GGCCG----  A.....A.C..
9  ephed  .-...C ---GAAGCC--TC-CGCCA----  A..........
10 gnetu  .-...C TCCG-AGCC--TA-GGCCG----  A..........
11 PINUS  .....C ---GAGGCC--TC-GGTCG----  A..........
12 PICEA  .....C ---GAGNCC--TC-GGTCG----  A..........
13 TAXUS  .GC..G ---GAG-C---TC-GGCCG----  A..........
14 marsi  .-...C ---GAGGC---TC-GTCCG----  A..........
15 osmun  .-...C ---GCGGC---TC-GTCCA----  A.....T.C..
16 mnium  .....C ---GAGGC---TC-GTCCG----  A.....TT..C
17 CHLAM  ....TC ---GAGGC--TTC-GGCCA----  A.A...T....
18 SPERM  ....TC ---GAGGC--TTC-GGCCG----  A.A...T..T.
19 TETRA  ....TC ---GAGGC--CTC-GGCCA----  A.A.....C..
20 CHLOR  ....TC ---GAGAC--CTC-GGTCA----  A.A...T....
21 CLADO  ....TC ---AAGTC--TAC-GGACT----  T.A...T....
22 HETER  .....C TTT--GGT-ATT----CCGA---  A..-....C..
23 VOLVA  ....TC TTT--GGCCATT----CCGA---  A.A...T.C..
24 SCLER  .....C CTT--GGT-ATT----CCGG---  G.....T.C..
25 sacch  .....C CTT--GGT-ATT----CCAG---  G.....T.C..
26 BIPOL  .....C TTT--GGT-ATT----CCAA---  A.....T.C..
27 GLOMU  ....TC CCT--GGT-ATT----CCGG---  G.A.T.T.C..
28 CYANI  ....TT TC--AGGAGAATTTTATTTTCCT  G.A........
29 SARCO  ....TC GC---GGTAA-TC--------CT  GCA.--T....
30 PHYTO  ..A.TT CCG--GGTTAGTC---CTG----  G.A.T.T.C..
31 SCYTO  ...-TT CCG--GGATATGC---CTG----  G.A...T.CT.
32 crypt  .---CT CC---AGC--TGA---CT-----  ------T...A
33 PRORO  ....TT TCG--GGATATCC---CTG----  AA....T.C..
```

Alignment surgery



Μοντέλο υποκατάστασης

Για πρωτεϊνικές αλληλουχίες, οι συχνότητες (πιθανότητες) υποκατάστασης δίδονται από τους γνωστούς πίνακες PAM, BLOSUM, ... Ο υπολογισμός της απόστασης μεταξύ δύο αλληλουχιών μπορεί να μετατραπεί σε άλλες μονάδες όπως η αναμενόμενη επί τοις εκατό αλλαγή αμινοξέων μεταξύ δύο αλληλουχιών. Για τα νουκλεϊκά οξέα, οι συχνότητες υποκατάστασης δίδονται είτε ως ένας 4x4 πίνακας (για κάθε βάση) ή και ως ένας πίνακας 61x61 (για τον γενετικό κώδικα). Στα μοντέλα υποκατάστασης μπορούν να ενσωματωθούν επίσης διαφορές στο ρυθμό υποκατάστασης διαφορετικών τμημάτων της αλληλουχίας (π.χ. 3η βάση του κώδικα).

Αλγόριθμοι δημιουργίας δένδρων

Το πρόβλημα

Το βασικό πρόβλημα της δημιουργίας δένδρων είναι ότι το πλήθος των πιθανών δένδρων αυξάνει πολύ γρήγορα με αυξανόμενο πλήθος αλληλουχιών : εάν το πλήθος των αλληλουχιών είναι n , τότε το πλήθος των διαφορετικών δένδρων με ή χωρίς ρίζα είναι :

$$T_{rooted} = (2n-3)\prod_{i=3}^{n}(2i-5)$$

$$T_{unrooted} = \prod_{i=3}^{n}(2i-5)$$

Αλγόριθμοι δημιουργίας δένδρων

Το πρόβλημα

Αριθμός Αλληλουχιών	Δένδρα χωρίς ρίζα	Δένδρα με ρίζα
4	3	15
5	15	105
6	105	945
8	10395	135135
10	2027025	34459425

Αλγόριθμοι

Οι αλγόριθμοι δημιουργίας δένδρων χωρίζονται σε δύο κύριες κατηγορίες : τους αλγόριθμους που στηρίζονται στις αποστάσεις μεταξύ των αλληλουχιών (distance-based), και τους αλγόριθμους που βασίζονται στους χαρακτήρες (κατάλοιπα-βάσεις) των αλληλουχιών (character-based methods). Και οι δύο κατηγορίες αλγορίθμων χρησιμοποιούνται ευρέως, όπως ευρεία είναι επίσης η βιβλιογραφική συζήτηση για τα πλεονεκτήματα και μειονεκτήματα τους.

Αλγόριθμοι αποστάσεων

Οι αλγόριθμοι αυτοί χρησιμοποιούν τη στοίχιση μόνο ως μέσο για τον υπολογισμό ενός πίνακα αποστάσεων μεταξύ των αλληλουχιών. Είναι μόνο αυτές οι αποστάσεις που χρησιμοποιούνται για την εύρεση του δένδρου (και όχι αυτή καθ'αυτή η στοίχιση). Οι αλγόριθμοι αυτοί μπορούν να υπολογίσουν το πραγματικό φυλογενετικό δένδρο μόνο εάν η αρχική στοίχιση περιλαμβάνει πληροφορία για το σύνολο των εξελικτικών γεγονότων που έχουν οδηγήσει τις αλληλουχίες σε απόκλιση. Το πρόβλημα είναι ότι άπαξ και μία θέση μεταλλαχθεί, εκ νέου μεταλλάξεις στην ίδια θέση δεν μπορούν να διαγνωστούν.

Αλγόριθμοι αποστάσεων

Για το λόγο αυτό, πολλοί από αυτούς τους αλγόριθμους διορθώνουν τον πίνακα αποστάσεων σε μια απόπειρα να λάβουν υπόψη τους τέτοιες 'αόρατες' υποκαταστάσεις. Οι ανά ζεύγη αποστάσεις μεταξύ των αλληλουχιών υπολογίζονται από τις συχνότητες υποκατάστασης και αντιπροσωπεύουν ένα μέτρο της απόκλισης τους (πόσο ανόμοιες είναι). Οι πλέον γνωστοί αλγόριθμοι αποστάσεων είναι οι

- UPGMA
- Neighbor Joining (NJ)
- Fitch-Margoliash (ελαχίστων τετραγώνων)
- Της ελαχίστης εξέλιξης (minimum evolution)

UPGMA

Είναι τα αρχικά του Unweighted Pair Group Method with Arithmetic mean. Είναι μία μέθοδος ομαδοποίησης η οποία υποθέτει την ύπαρξη ενός μοριακού ρολογιού (ρυθμού εξέλιξης) το οποίο να είναι κοινό και σταθερό για ολόκληρο το δένδρο. Παρά τις ισχυρές ενδείξεις ότι η μέθοδος δίδει μάλλον φτωχά αποτελέσματα για τη δημιουργία φυλογενετικών δένδρων, η βιβλιογραφική της παρουσία εξακολουθεί να ανθεί κυρίως λόγω της ταχύτητας του αλγόριθμου.

UPGMA

Αλγόριθμος

- Όρισε n ομάδες $C_1, C_2, ..., C_n$ μία για κάθε αλληλουχία.
- Αρχικό δένδρο : όλοι οι κόμβοι στο μηδέν.
- Όρισε την απόσταση ανάμεσα στις ομάδες C_k και C_l ως τη μέση απόσταση ανάμεσα σε όλα τα ζεύγη αλληλουχιών από κάθε ομάδα :

$$\delta_{kl} = \frac{1}{|C_k||C_l|} \sum_{i \in C_k, j \in C_l} d_{ij}$$

- Βρες τις ομάδες C_k και C_l με το μικρότερο δ_{kl}

UPGMA

Αλγόριθμος

- Όρισε μία καινούργια ομάδα $C_m = C_k \cup C_l$ και αφαίρεσε τις ομάδες C_k και C_l.
- Πρόσθεσε ένα κόμβο m με θυγατρικούς τους k και l. Ο κόμβος αυτός θα μπει σε ύψος $\delta_{kl}/2$.
- Αν υπάρχουν ακόμα περισσότεροι από δύο ομάδες, επανέλαβε τον κύκλο.
- Αλλιώς, τοποθέτησε τη ρίζα του δένδρου σε ύψος $\delta_{kl}/2$.

UPGMA

Παράδειγμα 1ο.

Έστω τέσσερις αλληλουχίες Α,Β,Γ,Δ με πίνακα αποστάσεων :

	Α	Β	Γ	Δ
Α	-	8	7	12
Β	8	-	9	14
Γ	7	9	-	11
Δ	12	14	11	-

Ποιο είναι το UPGMA δένδρο ;

UPGMA

Παράδειγμα 1ο.

Οι ομάδες με τη μικρότερη απόσταση είναι οι Α και Γ.
Άρα, φτιάχνουμε μία καινούργια ομάδα, την ομάδα Α-Γ
η οποία θα έχει ένα κόμβο σε ύψος 7 / 2 = 3.5

Η απόσταση της Α-Γ από την Β είναι :
δ(Α-Γ,Β) = [δ(Α,Β) + δ(Γ,Β)] / 2 = [8 + 9] / 2 = 8.5

και παρόμοια :
δ(Α-Γ,Δ) = [δ(Α,Δ) + δ(Γ,Δ)] / 2 = [12 + 11] / 2 = 11.5

UPGMA

Παράδειγμα 1ο.

Άρα, ο νέος πίνακας αποστάσεων είναι :

	Α-Γ	Β	Δ
Α-Γ	-	8.5	11.5
Β	8.5	-	14
Δ	11.5	14	-

ενώ το παρόν δένδρο είναι :

UPGMA

Παράδειγμα 1ο.

Οι ομάδες με τη μικρότερη απόσταση είναι οι Α-Γ και Β.
Άρα, φτιάχνουμε μία καινούργια ομάδα, την ομάδα Α-Γ-Β η οποία θα έχει ένα κόμβο σε ύψος 8.5 / 2 = 4.25

Η απόσταση της Α-Γ-Β από τη Δ είναι :

δ(Α-Γ-Β,Δ) = [δ(Α,Δ) + δ(Γ,Δ) + δ(Β,Δ)] / 3 =

[12 + 11 + 14] / 3 = 12.33

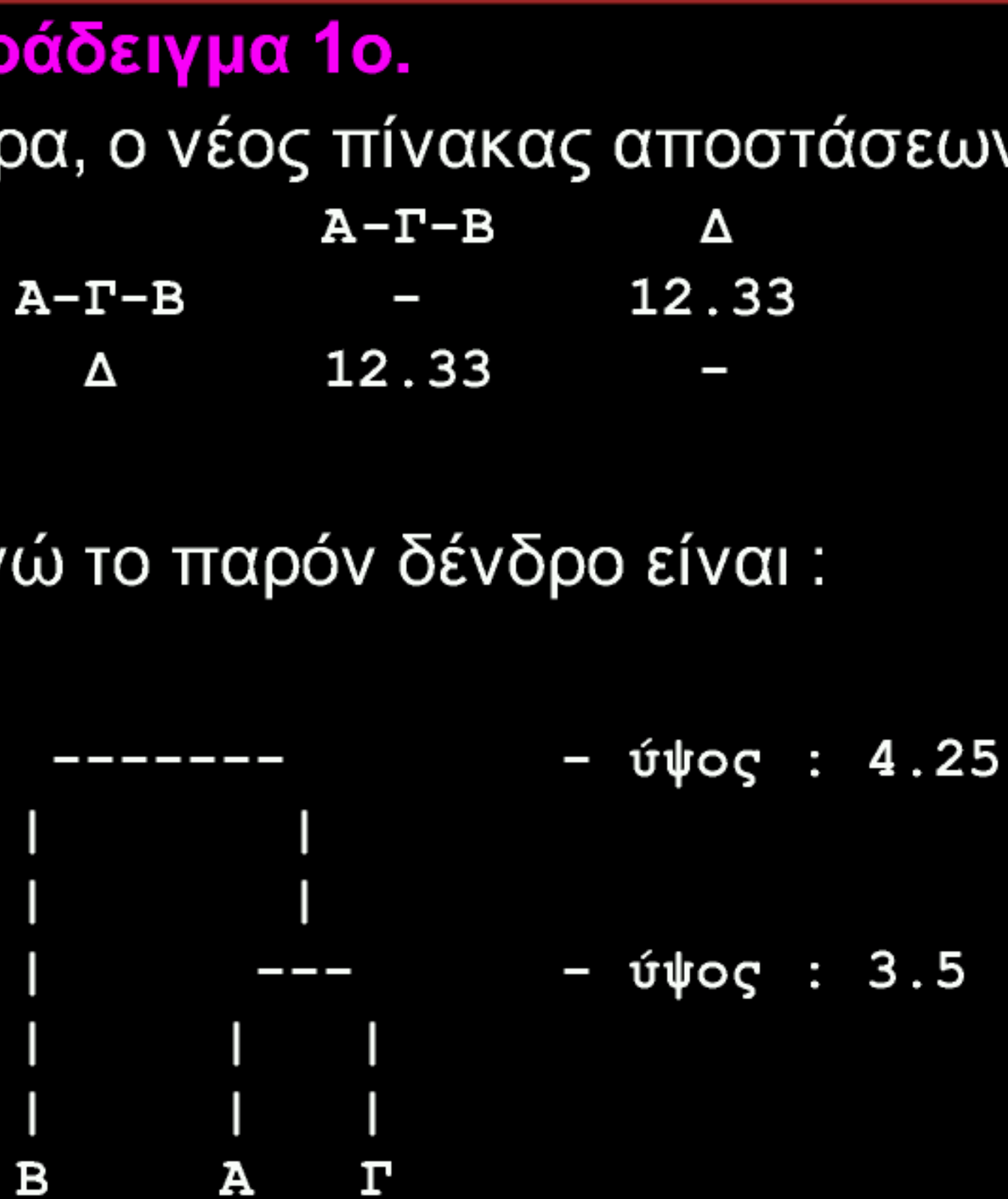
UPGMA

Παράδειγμα 1ο.

Άρα, ο νέος πίνακας αποστάσεων είναι :

	Α-Γ-Β	Δ
Α-Γ-Β	-	12.33
Δ	12.33	-

ενώ το παρόν δένδρο είναι :

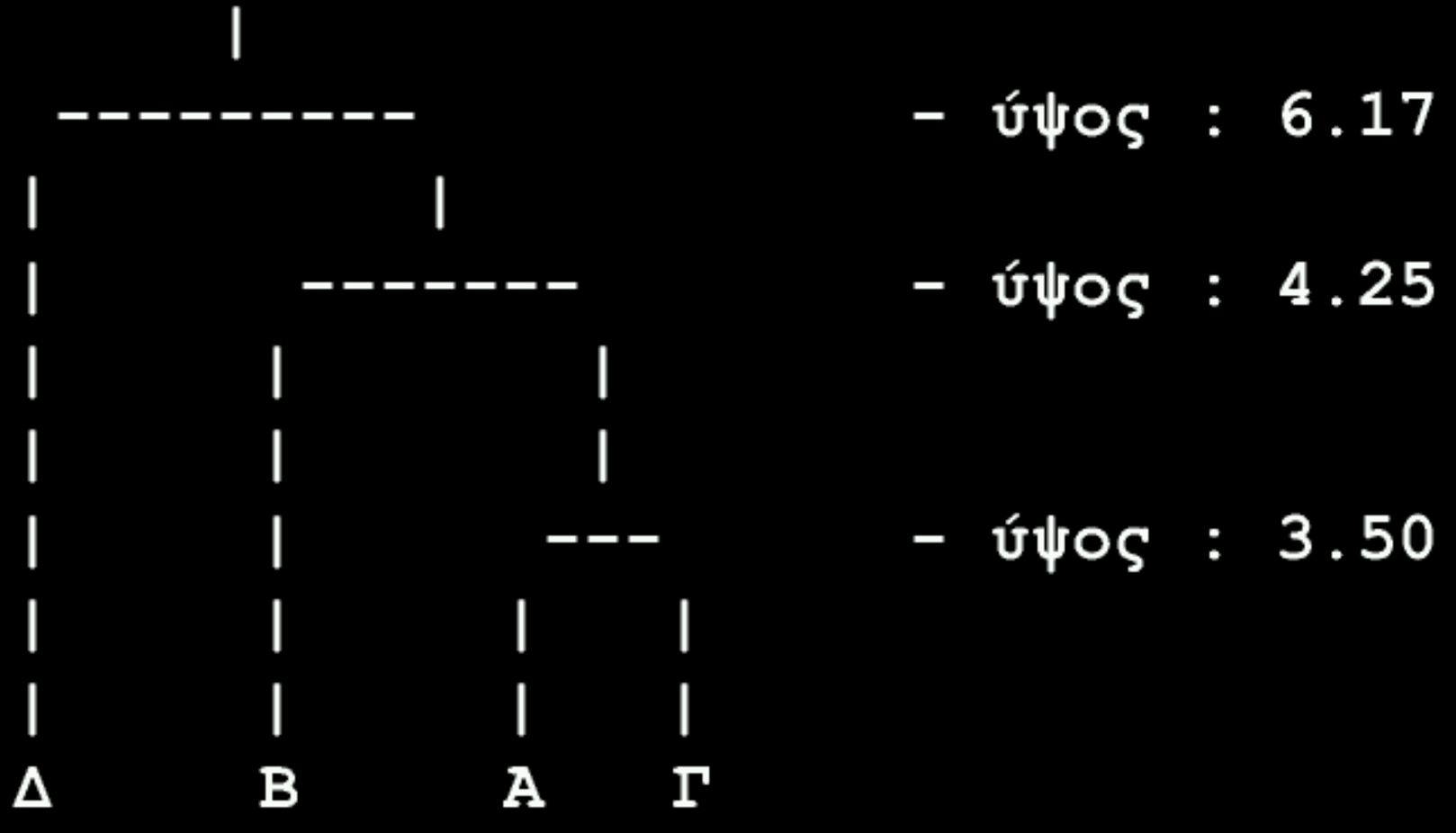


UPGMA

Παράδειγμα 1ο.

Οι δύο τελευταίες ομάδες ενώνονται μέσω ενός κόμβου σε ύψος 12.33 / 2 = 6.17

Άρα, το UPGMA δένδρο είναι :



UPGMA

Παράδειγμα 2ο.

Ο πίνακας αποστάσεων είναι :

	Α	Β	Γ	Δ	Ε
Α	-	8	8	5	3
Β		-	3	8	8
Γ			-	8	8
Δ				-	5
Ε					-

Ποιο είναι το UPGMA δένδρο ;

UPGMA

Παράδειγμα 2ο.

Κόμβος Β-Γ σε ύψος 1.5

	Β-Γ	Α	Δ	Ε
Β-Γ	-	8	8	8
Α		-	5	3
Δ			-	5
Ε				-

UPGMA

Παράδειγμα 2ο.

Κόμβος Α-Ε σε ύψος 1.5

	Β-Γ	Α-Ε	Δ
Β-Γ	-	8	8
Α-Ε		-	5
Δ			-

UPGMA

Παράδειγμα 2ο.

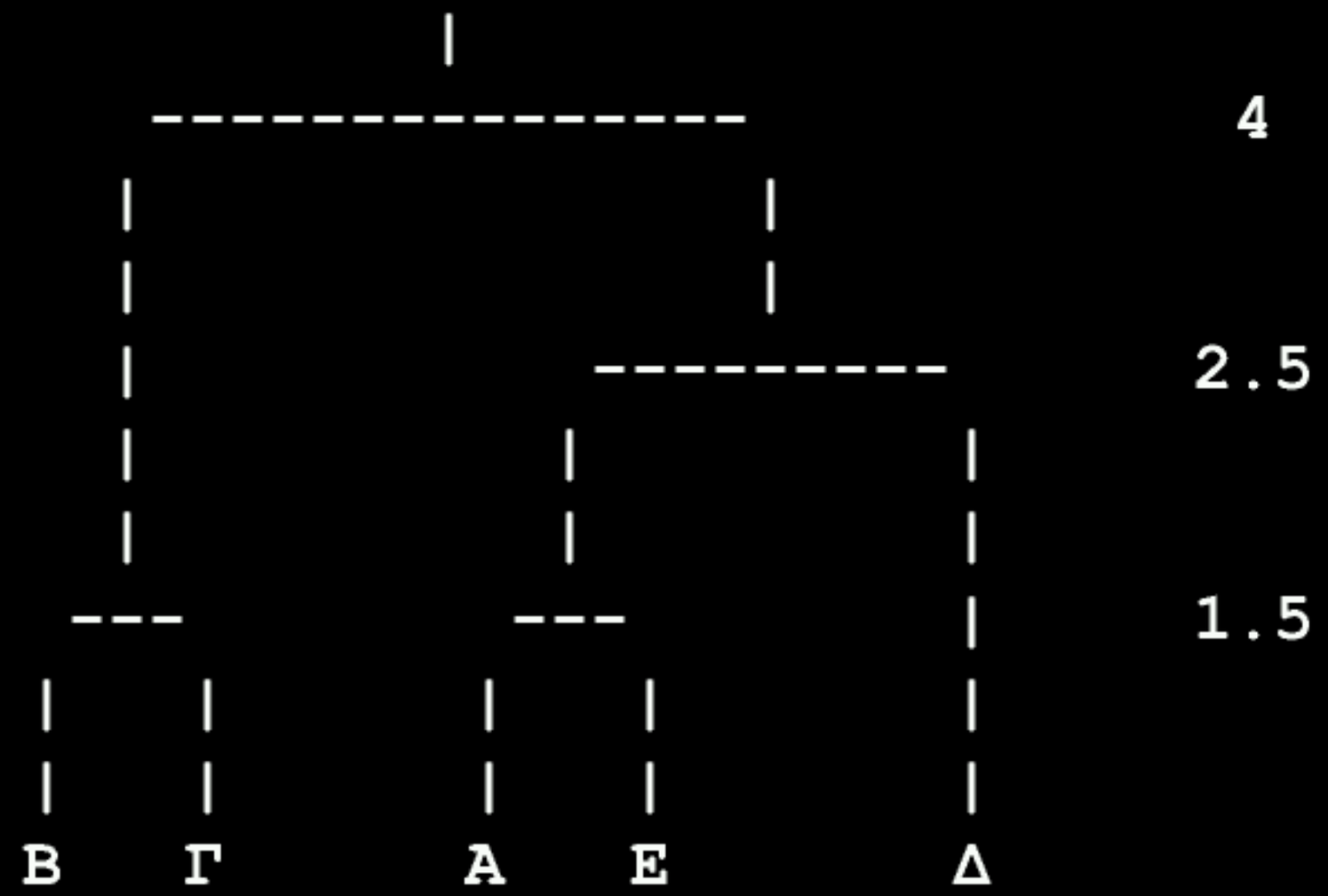
Κόμβος Α-Ε-Δ σε ύψος 2.5

	Β-Γ	Α-Ε-Δ
Β-Γ	-	8
Α-Ε-Δ		-

UPGMA

Παράδειγμα 2ο.

Ρίζα δένδρου σε ύψος 4



UPGMA

Παράδειγμα 3ο.

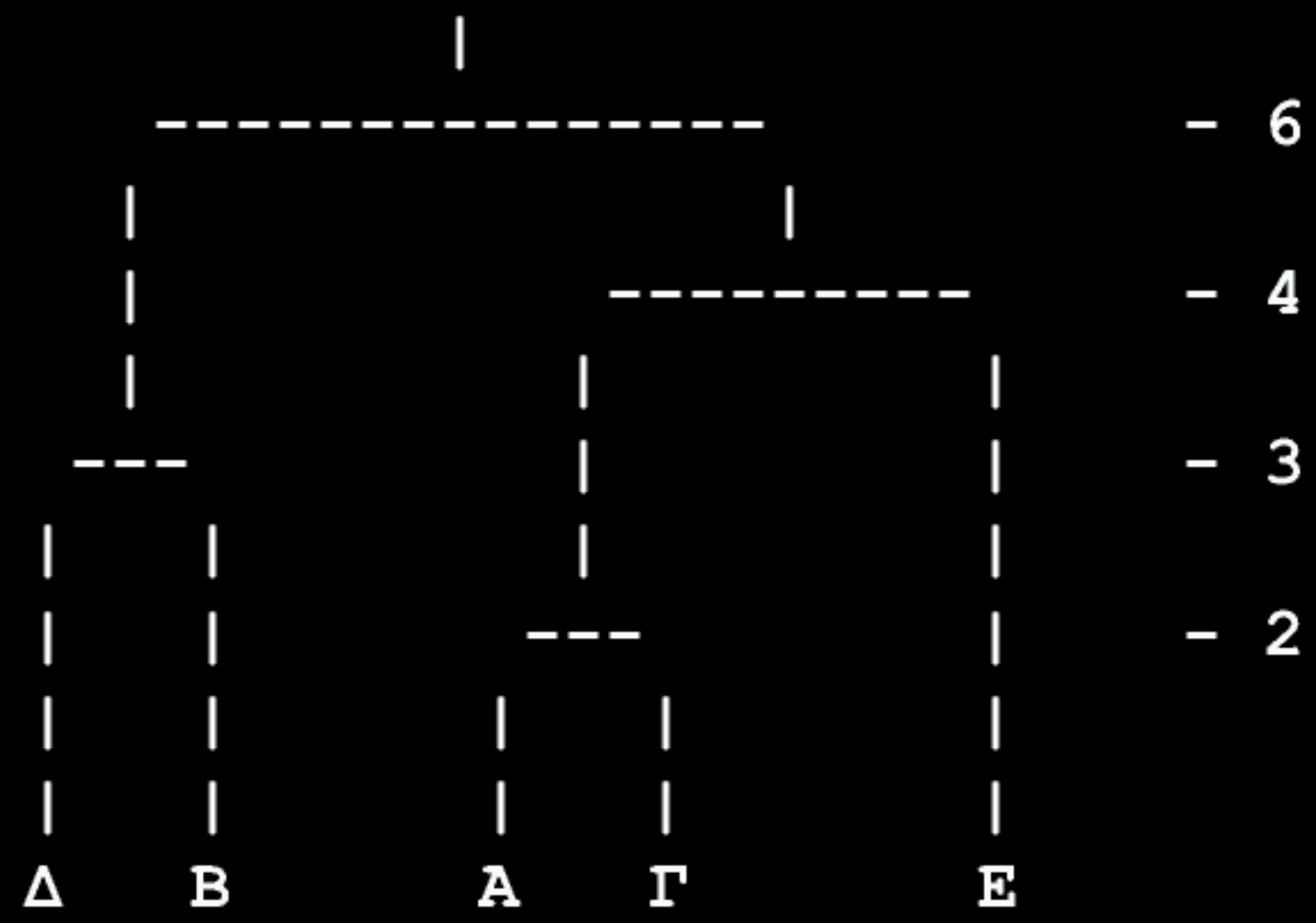
Ο πίνακας αποστάσεων είναι :

	Α	Β	Γ	Δ	Ε
Α	-	12	4	12	8
Β		-	12	6	12
Γ			-	12	8
Δ				-	12
Ε					-

Ποιο είναι το UPGMA δένδρο ;

UPGMA

Παράδειγμα 3ο.



UPGMA

Ενώ στο 2ο και 3ο παράδειγμα, η μέθοδος έδωσε ένα δένδρο το οποίο εξηγούσε πλήρως τα δεδομένα (δηλ. το άθροισμα των αποστάσεων των κλάδων ήταν ίσο με τις αποστάσεις που είχαν δοθεί στο αλγόριθμο), στο 1ο παράδειγμα αυτό δεν ίσχυε : Για παράδειγμα οι αποστάσεις Β-Α και Β-Γ ήταν από το δένδρο ίσες με το 8.5, ενώ τα δεδομένα ήταν : Β-Α = 8 και Β-Γ = 9.
Ο λόγος για αυτό είναι ότι τα δεδομένα του 1ου παραδείγματος δεν είναι συμβατά με την ύπαρξη ενός σταθερού και καθολικού (για όλο το δένδρο) μοριακού ρολογιού. Δεδομένα συμβατά με την ύπαρξη μοριακού ρολογιού ονομάζονται υπερμετρικά (ultrametric).

Υπερμετρικά δεδομένα

Η πλέον χαρακτηριστική ιδιότητα υπερμετρικών δεδομένων είναι ότι για οποιαδήποτε τριάδα αλληλουχιών, οι μεταξύ τους αποστάσεις θα είναι είτε όλες ίσες ή δύο θα είναι ίσες και η τρίτη μικρότερη. Συμβολικά : έστω τρεις αλληλουχίες i, j, k, και οι μεταξύ τους αποστάσεις d(ij), d(ik) και d(jk). Τότε, για να είναι τα δεδομένα του πίνακα αποστάσεων υπερμετρικά, θα πρέπει να ισχύει :

$$d(ij) = d(ik) = d(jk) \text{ ή}$$

$$d(ij) = d(ik) \text{ και } d(jk) < d(ij, ik) \text{ ή}$$

$$d(ij) = d(jk) \text{ και } d(ik) < d(ij, jk) \text{ ή}$$

$$d(ik) = d(jk) \text{ και } d(ij) < d(ik, jk)$$

Εξεταστικό παράδειγμα

Έστω πέντε αλληλουχίες Α, Β, Γ, Δ και Ε, και ο κάτωθι πίνακας μεταξύ τους αποστάσεων :

	Α	Β	Γ	Δ	Ε
Α	-	8	8	3	8
Β		-	5	8	3
Γ			-	8	5
Δ				-	8
Ε					-

Ποιο από τα παρακάτω είναι το σωστό ; (9/100)

Εξεταστικό παράδειγμα

1. Τα δεδομένα είναι υπερμετρικά. Χρησιμοποιώντας τον αλγόριθμο UPGMA βρίσκουμε το εξής δένδρο : B-E ενώνονται μέσω κόμβου N1 σε ύψος 1.5, οι A-Δ ενώνονται μέσω κόμβου N2 σε ύψος 1.5, οι N2-Γ ενώνονται μέσω κόμβου N3 σε ύψος 2.5, και, τέλος, οι N1-N3 ενώνονται σε ύψος 4.

2. Τα δεδομένα έχουν την προσθετική ιδιότητα. Χρησιμοποιώντας τον αλγόριθμο NJ βρίσκουμε το εξής δένδρο : A-N1 : 1.5, Δ-N1 : 1.5, N1-N2 : 4, N2-Γ : 2.5, N2-N3 : 1, N3-E : 1.5, N3-B : 1.5, όπου N1, N2 και N3 είναι ενδιάμεσοι κόμβοι του δένδρου, και οι αριθμοί είναι οι αποστάσεις (τα μήκη) των αντίστοιχων κλάδων.

Εξεταστικό παράδειγμα

3. Τα δεδομένα έχουν την προσθετική ιδιότητα. Χρησιμοποιώντας τον αλγόριθμο NJ βρίσκουμε το εξής δένδρο : A-N1 : 1.5, Δ-N1 : 1.5, N1-N2 : 5, N2-Γ : 2.5, N2-E : 1.5, N2-B : 1.5, όπου N1 και N2 είναι ενδιάμεσοι κόμβοι του δένδρου, και οι αριθμοί είναι οι αποστάσεις (τα μήκη) των αντίστοιχων κλάδων.

4. Τα δεδομένα είναι υπερμετρικά. Χρησιμοποιώντας τον αλγόριθμο UPGMA βρίσκουμε το εξής δένδρο : οι A-Δ ενώνονται μέσω κόμβου N1 σε ύψος 1.5, οι B-E ενώνονται μέσω κόμβου N2 σε ύψος 1.5, οι N2-Γ ενώνονται μέσω κόμβου N3 σε ύψος 2.5, και, τέλος, οι N1-N3 ενώνονται σε ύψος 4.

Εξεταστικό παράδειγμα

5. Τα δεδομένα είναι υπερμετρικά. Χρησιμοποιώντας τον αλγόριθμο UPGMA βρίσκουμε το εξής δένδρο : B-E ενώνονται μέσω κόμβου N1 σε ύψος 1.5, οι A-Γ ενώνονται μέσω κόμβου N2 σε ύψος 1.5, οι N2-Δ ενώνονται μέσω κόμβου N3 σε ύψος 2.5, και, τέλος, οι N1-N3 ενώνονται σε ύψος 4.