Βιοπληροφορική

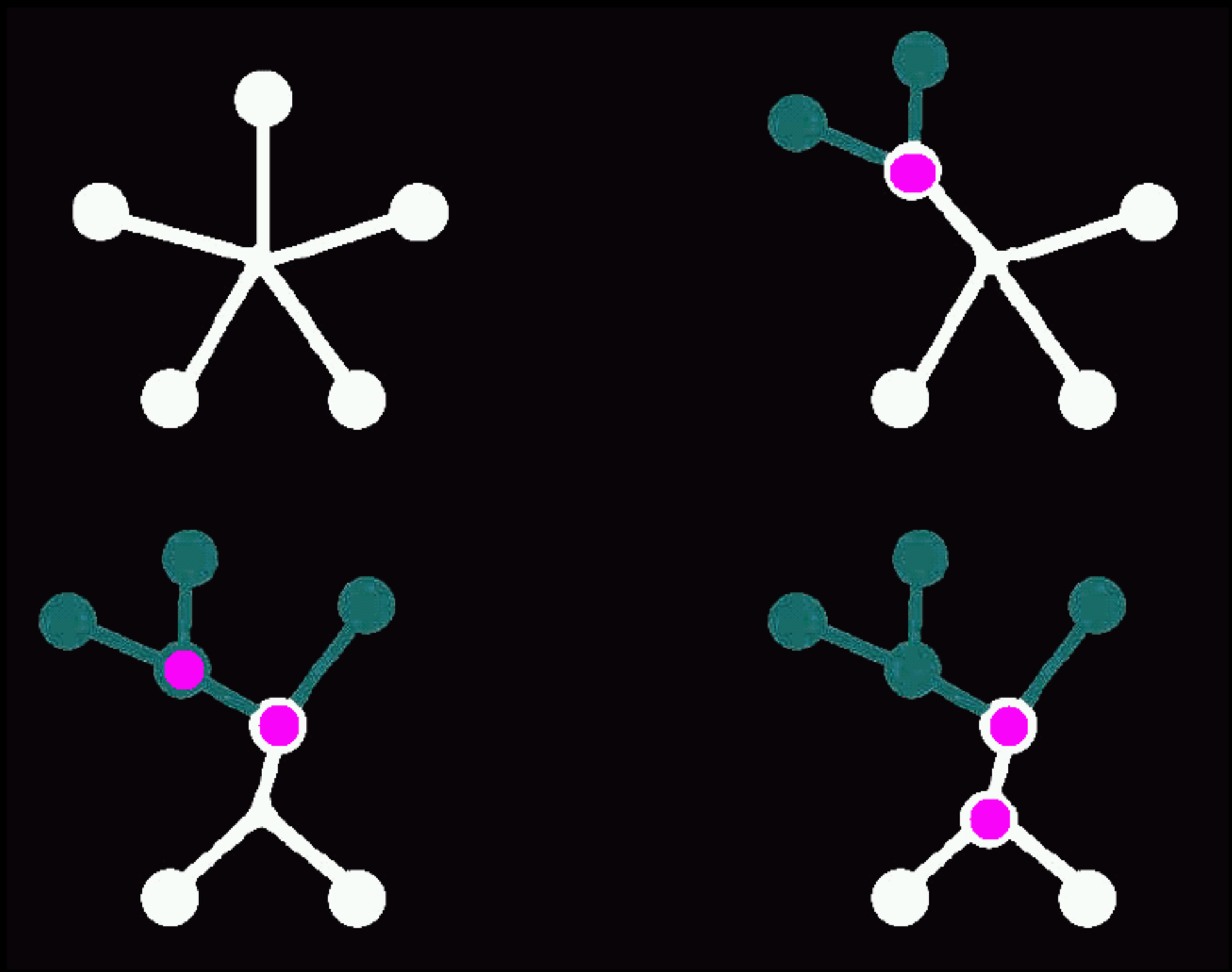
Διάλεξη 8η :

Φυλογενετική ανάλυση αλληλουχιών :
Αλγόριθμοι αποστάσεων : Neighbor Joining.
Εισαγωγή στους αλγόριθμους χαρακτήρων (maximum parsimony, maximum likelihood), Bootstrapping.

Neighbor Joining (NJ)

Ο αλγόριθμος UPGMA θα βρει το σωστό δένδρο όταν τα δεδομένα είναι υπερμετρικά. Το πρόβλημα είναι ότι τα υπερμετρικά δεδομένα είναι μάλλον σπάνια στη βιολογία : η πίεση της φυσικής επιλογής συνήθως διαφέρει για διαφορετικούς οργανισμούς, χρονικές περιόδους, γονίδια, περιοχές γονιδίων κοκ. Η μέθοδος NJ είναι μία διαδεδομένη και σχετικά γρήγορη μέθοδος δημιουργίας δένδρων η οποία δεν απαιτεί την ύπαρξη ενός μοριακού ρολογιού αλλά απαιτεί ότι τα δεδομένα έχουν την προσθετική ιδιότητα για τα μήκη των κλάδων. Η NJ είναι η μέθοδος που χρησιμοποιείται από το Clustal για τη δημιουργία του δενδρογράμματος-οδηγού.

Neighbor Joining (NJ)



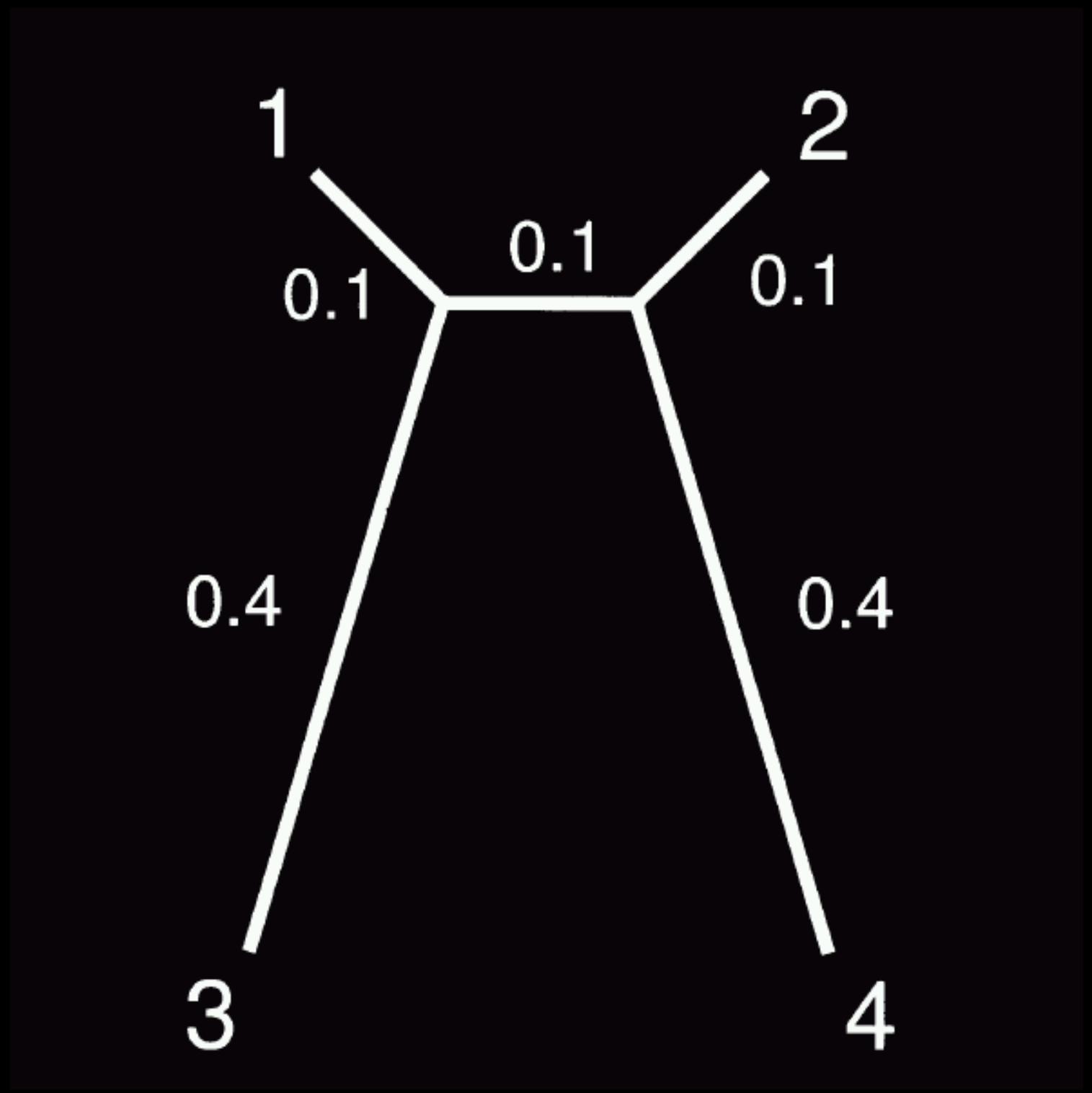
Προβλήματα της UPGMA

Έστω τέσσερις αλληλουχίες 1,2,3,4 με πίνακα αποστάσεων :

	1	2	3	4
1	-	.3	.5	.6
2		-	.6	.5
3			-	.9
4				-

Ποιο είναι το UPGMA δένδρο ; Υπάρχει άλλη προφανής λύση ;

Neighbor Joining (NJ)



Neighbor Joining (NJ)

- Όρισε μία νέου τύπου απόσταση D_{ij} ανάμεσα στις αλληλουχίες k και l η οποία να λαμβάνει υπόψη τη μέση απόσταση των αλληλουχιών ως προς τις υπόλοιπες αλληλουχίες :

$$D_{ij} = d_{ij} - (r_i + r_j)$$

όπου, d_{ij} είναι η απόσταση από τον αρχικό πίνακα αποστάσεων,

$$r_i = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik}$$

είναι ανάλογο της απόσταση της αλληλουχίας i από τις υπόλοιπες αλληλουχίες, και $|L|$ είναι το πλήθος των κόμβων του δένδρου.

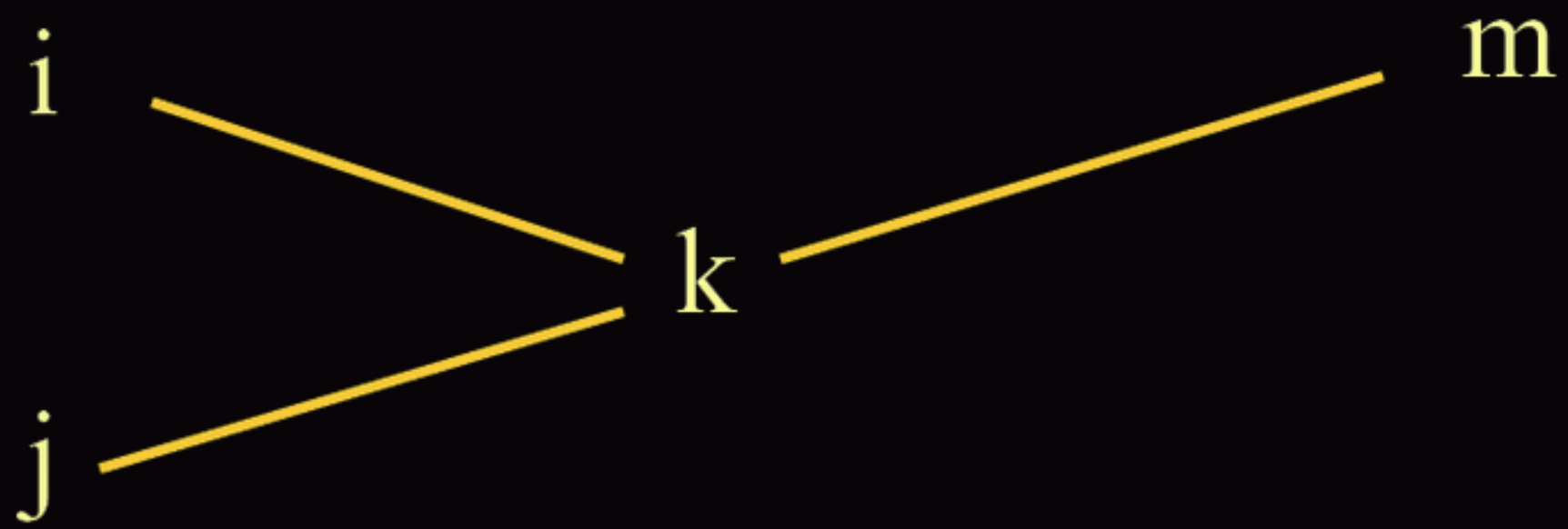
Neighbor Joining (NJ)

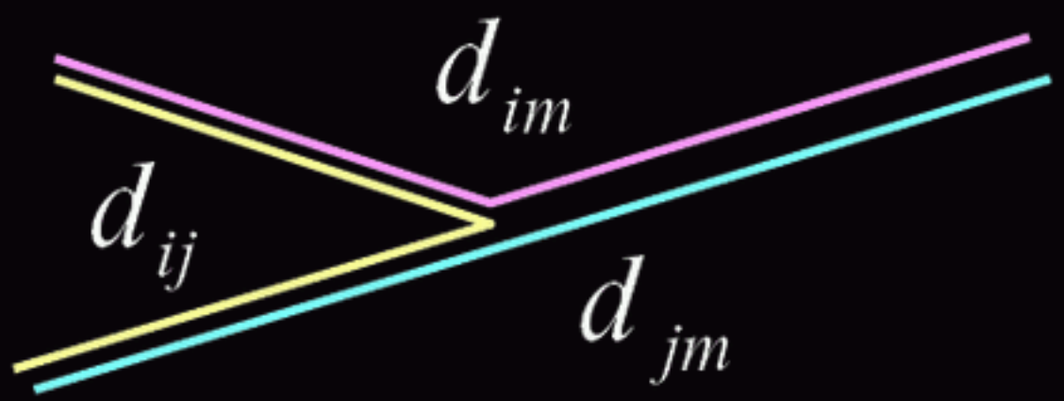
- Ένωσε τις αλληλουχίες i και j με το μικρότερο D_{ij}. Ας ονομάσουμε k τον καινούργιο κόμβο που εισαγάγαμε (τον πατρικό των i και j). Η απόσταση του k από κάποιον από τους αρχικούς κόμβους m είναι :

$$d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$$

Neighbor Joining (NJ)

$$d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$$





Neighbor Joining (NJ)

- Οι αποστάσεις των i και j από τον κόμβο k (που τα συνδέει) είναι :

$$t_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j)$$

και

$$t_{jk} = d_{ij} - t_{ik}$$

Στην έκφραση για το t_{ik} χρησιμοποιούνται οι μέσες αποστάσεις προς όλα τα άλλα φύλλα (τα r_i και r_j) αντί για τα d_{im} και d_{jm} ως διόρθωση για τις περιπτώσεις που τα δεδομένα δεν έχουν την προσθετική ιδιότητα.

- Αφαίρεσε τα i και j, και επανέλαβε τον κύκλο μέχρι να μείνουν δύο υποδένδρα ($|L| = 2$).

Neighbor Joining (NJ)

Παράδειγμα 1ο.

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

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

Ποιο είναι το NJ δένδρο ;

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπολογίζουμε τις τιμές των r = [Σ d(ij)] / (L-2)

	A	B	Γ	Δ	r
A	-	8	7	12	13.5
B		-	9	14	15.5
Γ			-	11	13.5
Δ				-	18.5

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπολογίζουμε τις τιμές των D(ij) = d(ij) - (ri + rj) και για ευκολία τις εισάγουμε στο κάτω μισό του πίνακα :

	Α	Β	Γ	Δ	r
Α	-	8	7	12	13.5
Β	-21	-	9	14	15.5
Γ	-20	-20	-	11	13.5
Δ	-20	-20	-21	-	18.5

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπάρχουν δύο ζεύγη αλληλουχιών με το ίδιο (μικρότερο) D(ij) : οι A-B και Γ-Δ. Όποιο και εάν διαλέξουμε οδηγεί στο ίδιο δένδρο. Διαλέγουμε να ενώσουμε τις A-B μέσω ενός νέου κόμβου που ας τον ονομάσουμε N1. Οι αποστάσεις του N1 από τις A,B είναι :

$$t(A\text{-}N1) = 0.5 * (d(A,B) + r(A) - r(B)) =$$
$$= 0.5 * (8 + 13.5 - 15.5) = 3$$

$$t(B\text{-}N1) = d(A,B) - t(A\text{-}N1) = 8 - 3 = 5$$

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Οι αποστάσεις του νέου κόμβου N1 από τις υπόλοιπες αλληλουχίες (Γ,Δ) είναι :

d(N1,Γ) = [d(A,Γ) + d(B,Γ) - d(A,B)] / 2 =
= [7 + 9 - 8] / 2 = 4

d(N1,Δ) = [d(A,Δ) + d(B,Δ) - d(A,B)] / 2 =
= [12 + 14 - 8] / 2 = 9

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Αφαιρούμε τις αλληλουχίες Α,Β, και κατασκευάζουμε το νέο πίνακα αποστάσεων :

	N1	Γ	Δ
N1	-	4	9
Γ		-	11
Δ			-

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπολογίζουμε τις τιμές των r = [Σ d(ij)] / (L-2)
Προσοχή : το L τώρα έχει τιμή 3.

	N1	Γ	Δ	r
N1	-	4	9	13
Γ		-	11	15
Δ			-	20

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπολογίζουμε τις τιμές των D(ij) = d(ij) - (ri + rj) και για ευκολία τις εισάγουμε στο κάτω μισό του πίνακα :

	N1	Γ	Δ	r
N1	-	4	9	13
Γ	-24	-	11	15
Δ	-24	-24	-	20

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Υπάρχουν τρία ζεύγη κόμβων και φύλλων με το ίδιο (μικρότερο) D(ij) : τα Ν1-Γ, Ν1-Δ και Γ-Δ.
Όποιο και εάν διαλέξουμε οδηγεί στο ίδιο δένδρο.
Διαλέγουμε να ενώσουμε τις Γ-Δ μέσω ενός νέου κόμβου που ας τον ονομάσουμε Ν2. Οι αποστάσεις του Ν2 από τις Γ,Δ είναι :

$$t(\Gamma\text{-}N2) = 0.5 * (d(\Gamma,\Delta) + r(\Gamma) - r(\Delta)) =$$
$$= 0.5 * (11 + 15 - 20) = 3$$

$$t(\Delta\text{-}N2) = d(\Gamma,\Delta) - t(\Gamma\text{-}N2) = 11 - 3 = 8$$

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Η απόσταση του νέου κόμβου N2 από τον N1 είναι :

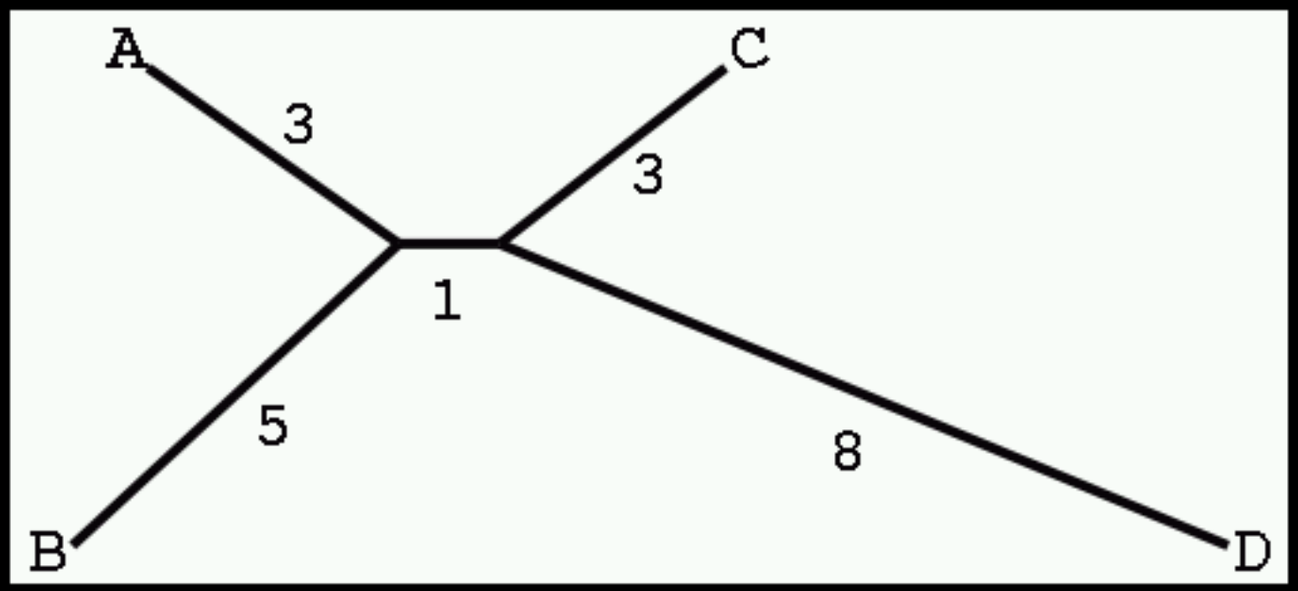
$$d(N1,N2) = [\, d(N1,Γ) + d(N1,Δ) - d(Γ,Δ) \,] / 2 =$$
$$= [\, 4 + 9 - 11 \,] / 2 = 1$$

Το δένδρο ολοκληρώνεται με τη σύνδεση των κόμβων N1 και N2 με μήκος κλάδου ίσο με την απόσταση τους.

Neighbor Joining (NJ)

Παράδειγμα 1ο.

Άρα, το δένδρο είναι :



Neighbor Joining (NJ)

Παράδειγμα 2ο.

Έστω πέντε αλληλουχίες Α,Β,Γ,Δ,Ε με πίνακα αποστάσεων :

	Α	Β	Γ	Δ	Ε
Α	-	12	13	7	6
Β		-	5	13	10
Γ			-	14	11
Δ				-	7
Ε					-

Ποιο είναι το NJ δένδρο ;

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	Α	Β	Γ	Δ	Ε	r
Α	-	12	13	7	6	38/3
Β		-	5	13	10	40/3
Γ			-	14	11	43/3
Δ				-	7	41/3
Ε					-	34/3

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	A	B	Γ	Δ	E	r
A	-	12	13	7	6	38/3
B	-14	-	5	13	10	40/3
Γ	-14	-22.	-	14	11	43/3
Δ	-19.	-14	-14	-	7	41/3
E	-18	-14.	-14.	-18	-	34/3

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Ενώνουμε Β-Γ μέσω κόμβου Ν1

t(Β-Ν1) = [d(Β,Γ) + r(Β) - r(Γ)] / 2 =
= (5 + 40/3 - 43/3) / 2 =
= 2

t(Γ-Ν1) = d(Β,Γ) - t(Β-Ν1) = 5 - 2 = 3

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Αποστάσεις N1 από Α,Δ,Ε :

d(N1,A) = [d(B,A) + d(Γ,A) - d(B,Γ)] / 2 =
= (12 + 13 - 5) / 2 =
= 10

d(N1,Δ) = [d(B,Δ) + d(Γ,Δ) - d(B,Γ)] / 2 =
= (13 + 14 - 5) / 2 =
= 11

d(N1,E) = [d(B,E) + d(Γ,E) - d(B,Γ)] / 2 =
= (10 + 11 - 5) / 2 =
= 8

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	N1	A	Δ	E
N1	-	10	11	8
A		-	7	6
Δ			-	7
E				-

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	N1	A	Δ	E	r
N1	-	10	11	8	29/2
A		-	7	6	23/2
Δ			-	7	25/2
E				-	21/2

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	N1	A	Δ	E	r
N1	-	10	11	8	29/2
A	-16	-	7	6	23/2
Δ	-16	-17	-	7	25/2
E	-17	-16	-16	-	21/2

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Ενώνουμε Α-Δ μέσω κόμβου Ν2

t(A-N2) = [d(A,Δ) + r(A) - r(Δ)] / 2 =

= (7 + 23/2 - 25/2) / 2 =

= 3

t(Δ-N2) = d(A,Δ) - t(A-N2) = 7 - 3 = 4

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Αποστάσεις N2 από E και N1 :

d(N2,N1) = [d(N1,A) + d(N1,Δ) - d(A,Δ)] / 2 =
= (10 + 11 -7) / 2 =
= 7

d(N2,E) = [d(E,A) + d(E,Δ) - d(A,Δ)] / 2 =
= (6 + 7 - 7) / 2 =
= 3

Neighbor Joining (NJ)

Παράδειγμα 2ο.

	N1	N2	E	r
N1	-	7	8	15/1
N2	-18	-	3	10/1
E	-18	-18	-	11/1

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Ενώνουμε N1-E μέσω κόμβου N3

t(N1-N3) = [d(N1,E) + r(N1) - r(E)] / 2 =
= (8 + 15 - 11) / 2 =
= 6

t(E-N3) = d(N1,E) - t(N1-N3) = 8 - 6 = 2

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Απόσταση του N3 από N2 :

d(N2,N3) = [d(N2,N1) + d(N2,E) - d(N1,E)] / 2 =

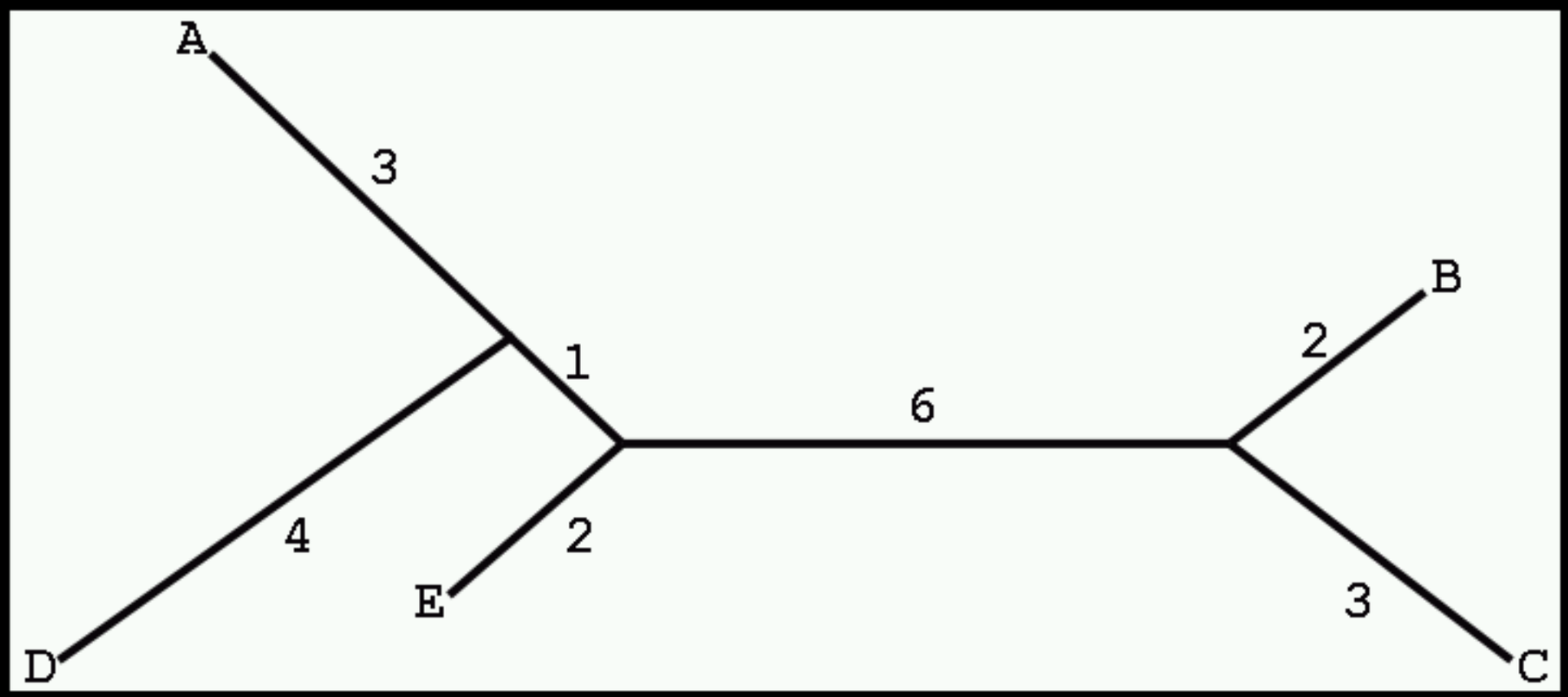
= (7 + 3 - 8) / 2 =

= 1

Neighbor Joining (NJ)

Παράδειγμα 2ο.

Άρα, το δένδρο είναι :



Neighbor Joining (NJ)

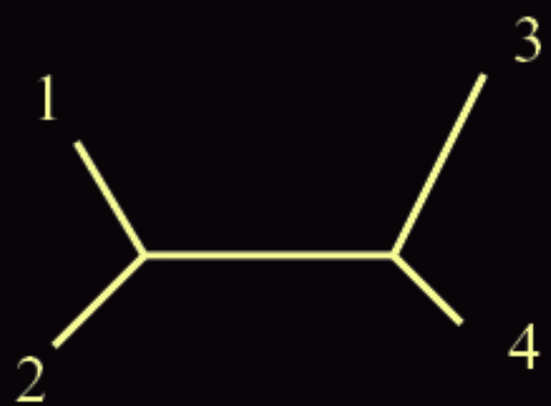
Προσθετικότητα

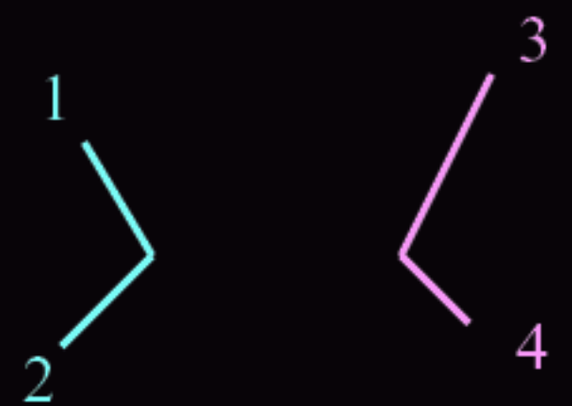
Ο NJ θα εντοπίσει το σωστό δένδρο εάν τα δεδομένα έχουν την προσθετική ιδιότητα. Η ύπαρξη της προσθετικής ιδιότητας ελέγχεται ως εξής :

Εάν για οποιαδήποτε σύνολο τεσσάρων φύλλων i, j, k και l οι αποστάσεις [d(ij)+d(kl)], [d(ik)+d(jl)] και [d(il)+d(jk)] είναι τέτοιες ώστε δύο από αυτές να είναι ίσες και μεγαλύτερες από την τρίτη, τότε τα δεδομένα έχουν την προσθετική ιδιότητα (και ο NJ θα επιστρέψει το σωστό δένδρο).

Neighbor Joining (NJ)

Προσθετικότητα





Παράδειγμα προγράμματος

ClustalW & NJplot

```
*************************************************************
******** CLUSTAL W (1.82) Multiple Sequence Alignments  ******
*************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)


Your choice:
```

Αλγόριθμοι χαρακτήρων

Οι αλγόριθμοι χαρακτήρων διαφέρουν από τους αλγόριθμους αποστάσεων στο ότι χρησιμοποιούν όλα τα δεδομένα (την πλήρη στοίχιση) για να συνάγουν το φυλογενετικό δένδρο. Οι πλέον γνωστές μέθοδοι που ανήκουν σε αυτή την κατηγορία είναι α. της μεγίστης φειδωλότητας (maximum parsimony), και, β. του βέλτιστου ενδεχομένου (maximum likelihood, ML).
Η ML έχει δειχθεί ότι δίνει τα ακριβέστερα αποτελέσματα από όλες τις μεθόδους και είναι η κυρίαρχη μέθοδος για φυλογενετικές μελέτες. Και οι δύο μέθοδοι έχουν ένα τόσο πλούσιο μαθηματικό υπόβαθρο που η πλήρης ανάπτυξη του ξεφεύγει από τα πλαίσια αυτού του μαθήματος.

Maximum parsimony

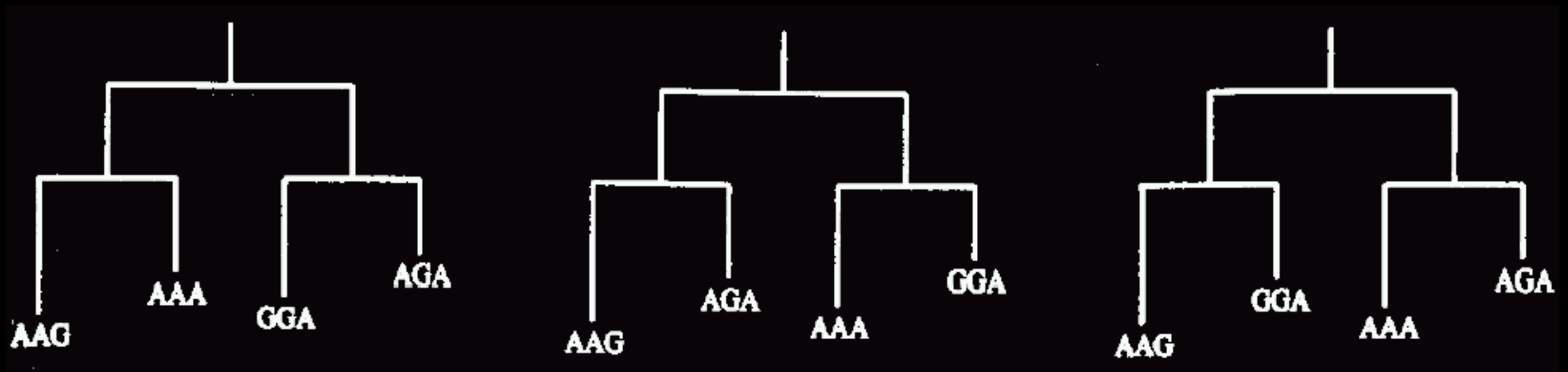
Η μέθοδος αναζητά το δένδρο για το οποίο ο αριθμός των υποκαταστάσεων των χαρακτήρων (βάσεων, αμινοξέων) ελαχιστοποιείται. Εάν ο αριθμός των πιθανών δένδρων είναι μικρός, τότε η εύρεση του δένδρου που ελαχιστοποιεί τον αριθμό υποκαταστά-σεων γίνεται μέσω μίας συστηματικής έρευνας. Εάν ο αριθμός των δένδρων είναι μεγάλος χρησιμοποιούνται στοχαστικοί αλγόριθμοι όπως simulated annealing (προσομοίωση ανόπτησης) ή ευρεστικοί αλγόριθμοι όπως η αντιμετάθεση κλάδων (branch-swapping).

Maximum parsimony

Παράδειγμα

Έστω τέσσερις αλληλουχίες AAG, AAA, GGA και AGA.

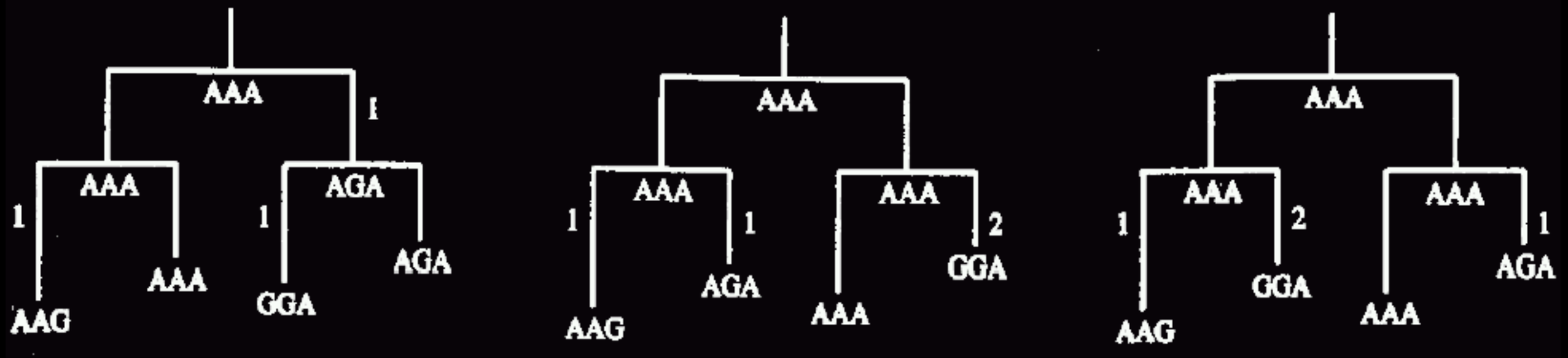
Υπάρχουν τρία πιθανά δένδρα (unrooted) :



Maximum parsimony

Παράδειγμα

Τα πλέον φειδωλά μοντέλα υποκατάστασης για κάθε ένα από αυτά τα δένδρα είναι :



και, συνεπώς, το maximum parsimony δένδρο είναι το πρώτο.

Maximum likelihood

Το θεώρημα του Bayes.

$$p(X, Y|I) = p(X|Y, I)p(Y|I)$$

$$p(Y, X|I) = p(Y|X, I)p(X|I)$$

$$p(X|Y, I) = \frac{p(Y|X, I)p(X|I)}{p(Y|I)}$$

Maximum likelihood

Το θεώρημα του Bayes.

$$p(X|Y,I) = \frac{p(Y|X,I)p(X|I)}{p(Y|I)}$$

$$p(hypothesis|data,I) = \frac{p(data|hypothesis,I)p(hypothesis|I)}{p(data|I)}$$

Maximum likelihood

Το θεώρημα του Bayes.

Ο όρος 'p(hypothesis|I)' είναι γνωστός ως prior (για prior probability) και αντιπροσωπεύει την εκ των προτέρων πιθανότητα να είναι η υπόθεση σωστή (με βάση τα όσα γνωρίζουμε για το πρόβλημα πριν έρθουν τα δεδομένα).

Ο όρος 'p(data|hypothesis,I)' είναι γνωστός ως likelihood και αντιπροσωπεύει την πιθανότητα να παρατηρηθούν τα συγκεκριμένα δεδομένα εάν η συγκεκριμένη υπόθεση είναι αληθής.

Ο όρος 'p(data|I)' αντιπροσωπεύει την πιθανότητα παρατήρησης των δεδομένων (κανονικοποιεί τις πιθανότητες στο εύρος 0-1).

Maximum likelihood

Υποθέτοντας ότι η εκ των προτέρων πιθανότητα είναι η ίδια για όλα τα ενδεχόμενα (ότι δηλ. όλες οι διαφορετικές υποθέσεις μας φαίνονται εκ των προτέρων ισοπίθανες), τότε η υπόθεση (μοντέλο) που έχει τη μέγιστη πιθανότητα να είναι σωστή (υπό το φως των δεδομένων) είναι αυτή που μεγιστοποιεί το likelihood. Άρα, το μοντέλο που αναζητούμε είναι αυτό για το οποίο ο όρος 'p(data|hypothesis,I)' μεγιστοποιείται, το οποίο σημαίνει :

Maximum likelihood

Το μοντέλο που θέλουμε είναι αυτό που κάνει προβλέψεις τέτοιες ώστε να μεγιστοποιείται η πιθανότητα να παρατηρήσουμε τα δεδομένα που παρατηρήθηκαν.

Αυτή είναι η αρχή του maximum likelihood, το οποίο έχει εφαρμογές σχεδόν σε όλα τα προβλήματα που ασχολούνται με εκτίμηση παραμέτρων και επιλογή μοντέλων.

Maximum likelihood

Εφαρμογή στην κατασκευή δένδρων

Από όλα τα πιθανά δένδρα και εξελικτικά μοντέλα, βρες εκείνο το δένδρο και εκείνο το μοντέλο για τα οποία η πιθανότητα να παρατηρηθούν τα δεδομένα που παρατηρήθηκαν (δηλ. η συγκεκριμένη στοίχιση) μεγιστοποιείται.

Εάν το εξελικτικό μοντέλο θεωρηθεί δεδομένο, τότε το πρόβλημα γράφεται :

Προσδιόρισε τις παραμέτρους (τοπολογία δένδρου, μήκη κλάδων) οι οποίες κάνουν τα δεδομένα που παρατηρήθηκαν να είναι τα πλέον πιθανά.

Στατιστικός έλεγχος

Αλγόριθμος : Bootstrapping (non-parametric)

- Κατασκεύασε πολλά νέα data sets μέσω τυχαίας επιλογής (με επαναφορά) θέσεων από την αρχική στοίχιση. Μια "θέση" είναι μία ολόκληρη στήλη της στοίχισης.
- Για κάθε ένα από αυτά τα data sets (στοιχίσεις) υπολόγισε ένα καινούργιο δένδρο με τον ίδιο αλγόριθμο και παραμέτρους που έδωσαν το αρχικό δένδρο.
- Βρες ποσοστό επανεμφάνισης του κάθε κόμβου του αρχικού δένδρου στα νέα δένδρα. Αυτό είναι η bootstrap τιμή του κόμβου. Τιμές μεγαλύτερες από 70% θεωρούνται στατιστικά σημαντικές.

Παράδειγμα προγράμματος

Bootstrapping in ClustalW

```
**************************************************************
******** CLUSTAL W (1.82) Multiple Sequence Alignments  ******
**************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)


Your choice:
```

Παράδειγμα προγράμματος