Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Φυλογενετικές σχέσεις

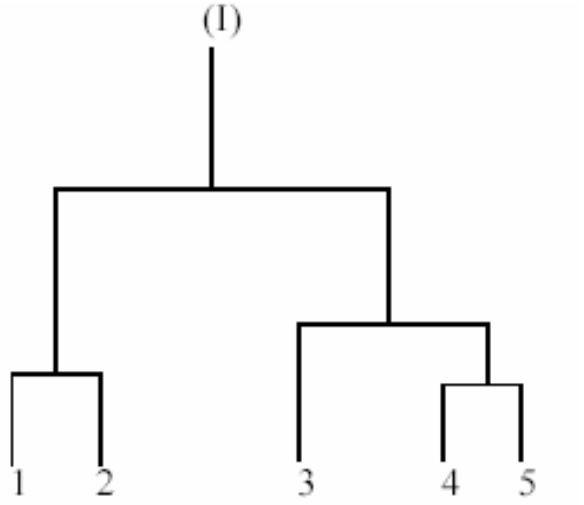
- Χρησιμοποιούνται οι ομοιότητες και οι διαφορές των μελετούμενων οργανισμών
- Τα χαρακτηριστικά που μελετώνται, ενώ παλιά ήταν μορφολογικά-ανατομικά, τώρα πλέον είναι όλο και περισσότερο μοριακά

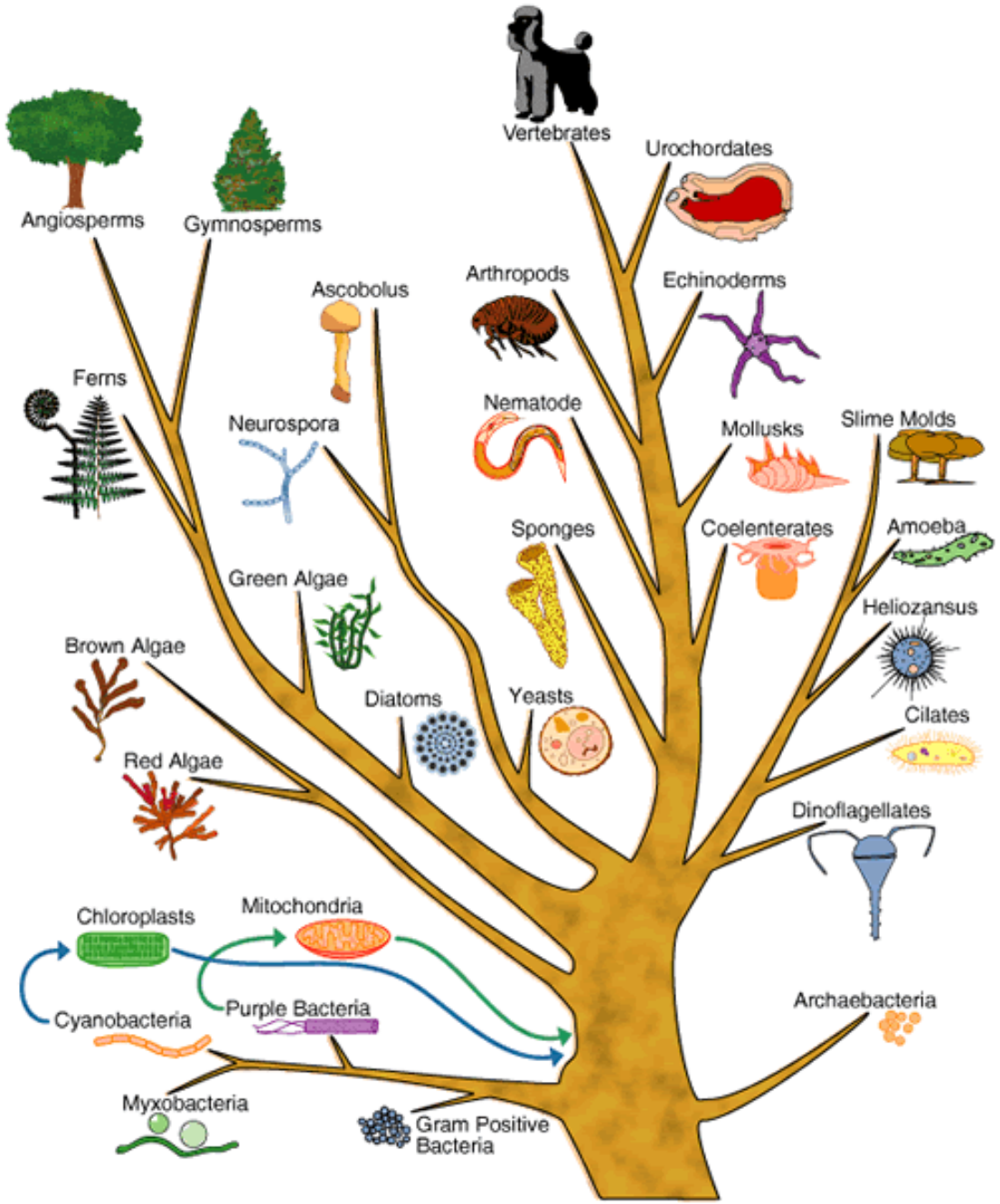
Βασικές αρχές

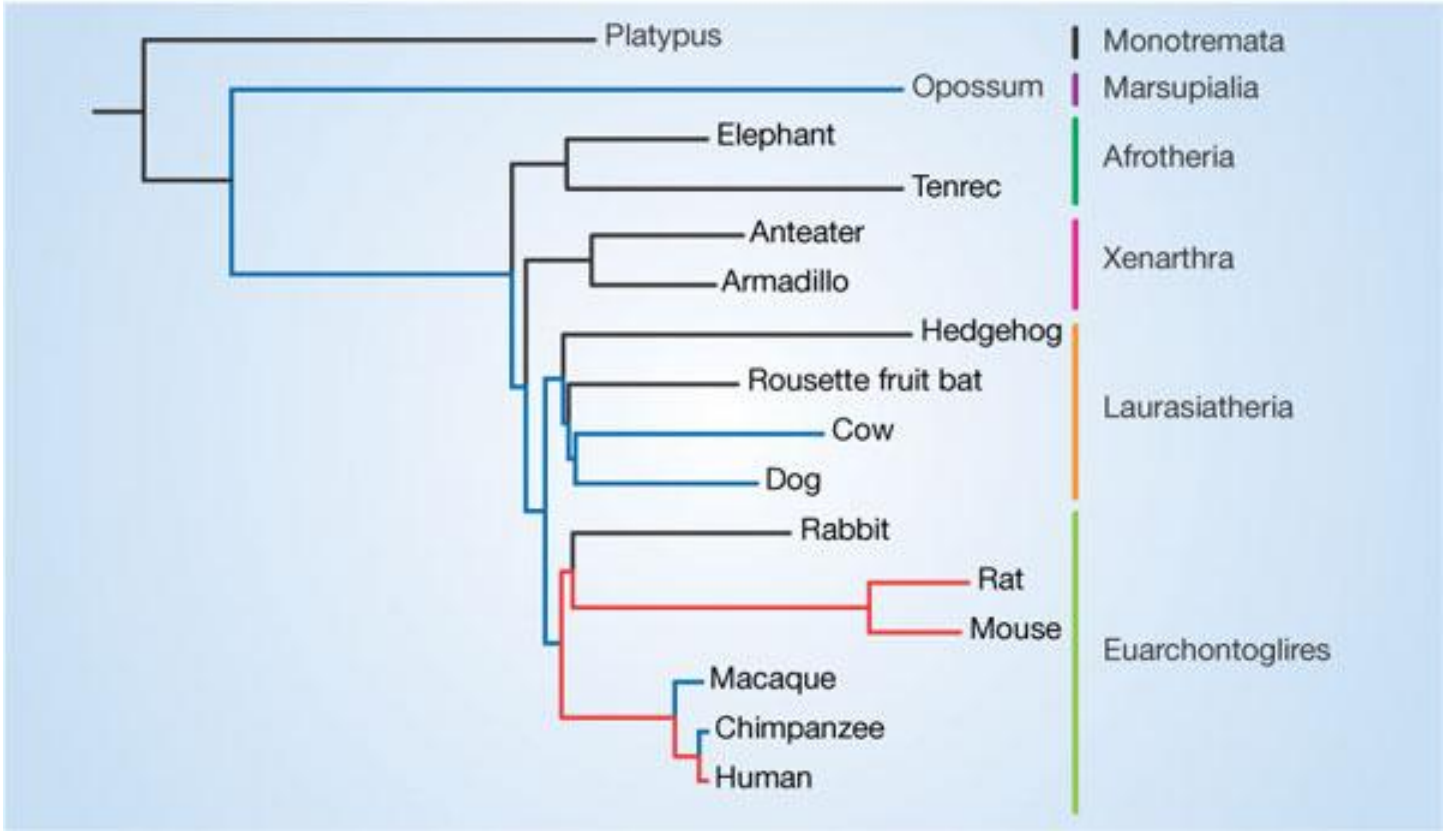
- Οποιαδήποτε ομάδα οργανισμών προέρχεται από έναν κοινό πρόγονο μέσω της εξέλιξης
- Υπάρχει διχαλωτό πρότυπο στην εξέλιξη
- Αλλαγή στα χαρακτηριστικά εμφανίζεται μετά το πέρασμα πολλών γενιών

Ορολογία στα δέντρα

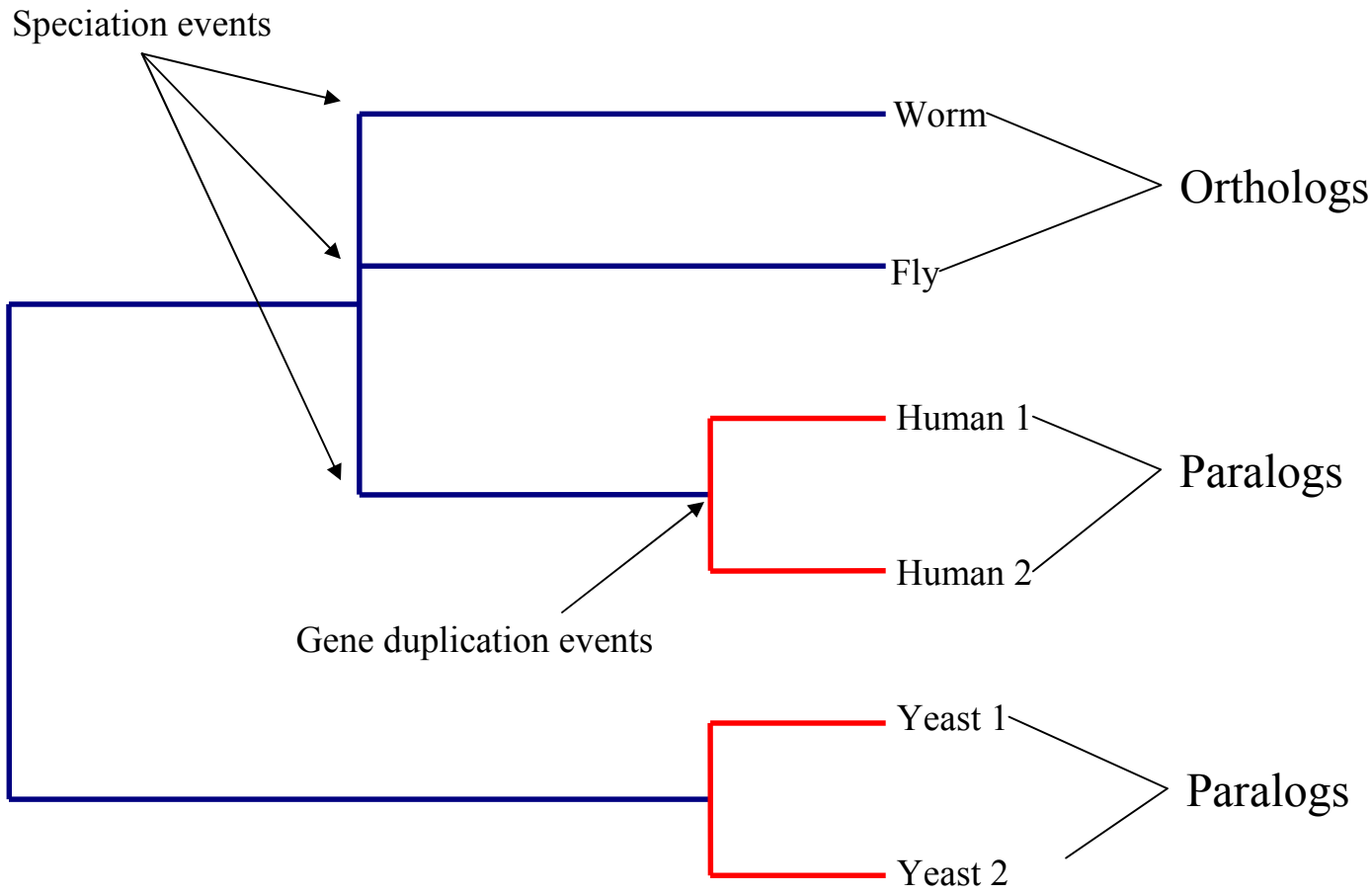
- Clade, Taxon, Node
- Στις πιο πολλές αναλύσεις, το μήκος των βραχιόνων ανταποκρίνεται στη φυλογενετική απόσταση-απόκλιση



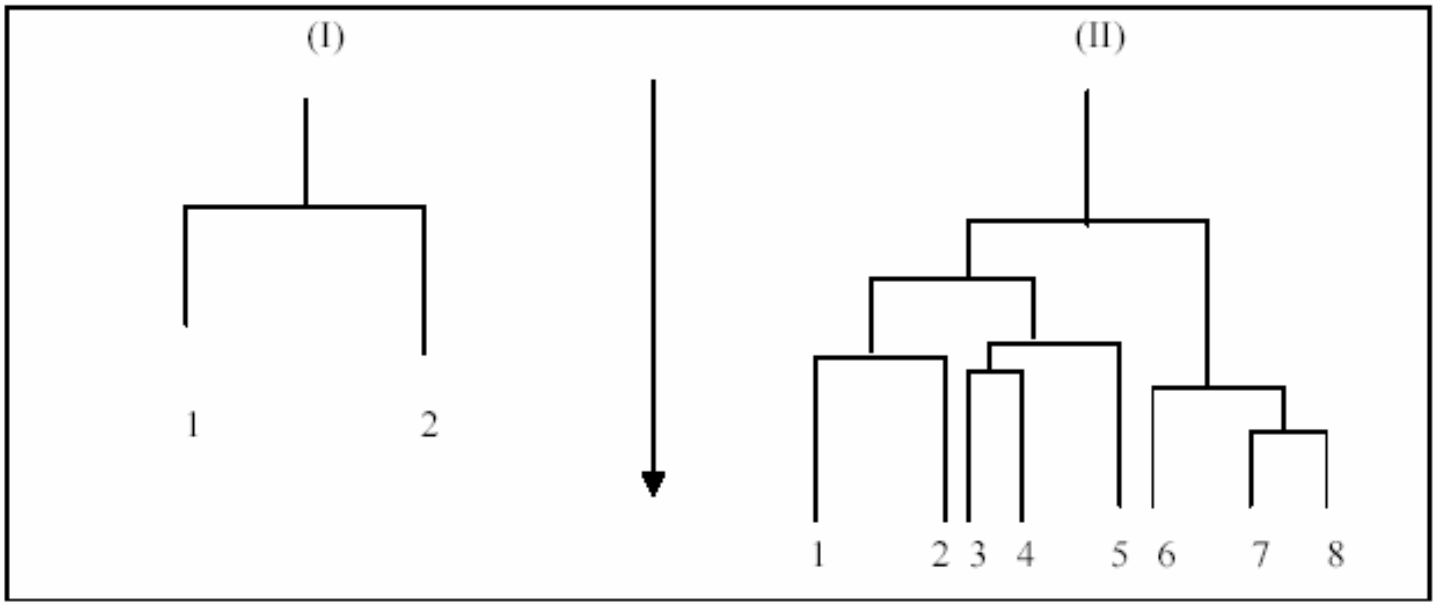

Vertebrates
Urochordates
Angiosperms
Gymnosperms
Arthropods
Echinoderms
Ascobolus
Ferns
Nematode
Neurospora
Mollusks
Slime Molds
Sponges
Coelenterates
Amoeba
Green Algae
Heliozansus
Brown Algae
Diatoms
Yeasts
Ciliates
Red Algae
Dinoflagellates
Chloroplasts
Mitochondria
Archaebacteria
Purple Bacteria
Cyanobacteria
Myxobacteria
Gram Positive Bacteria


Platypus
Opossum
Elephant
Tenrec
Anteater
Armadillo
Hedgehog
Rousette fruit bat
Cow
Dog
Rabbit
Rat
Mouse
Macaque
Chimpanzee
Human
Monotremata
Marsupialia
Afrotheria
Xenarthra
Laurasiatheria
Euarchontoglires

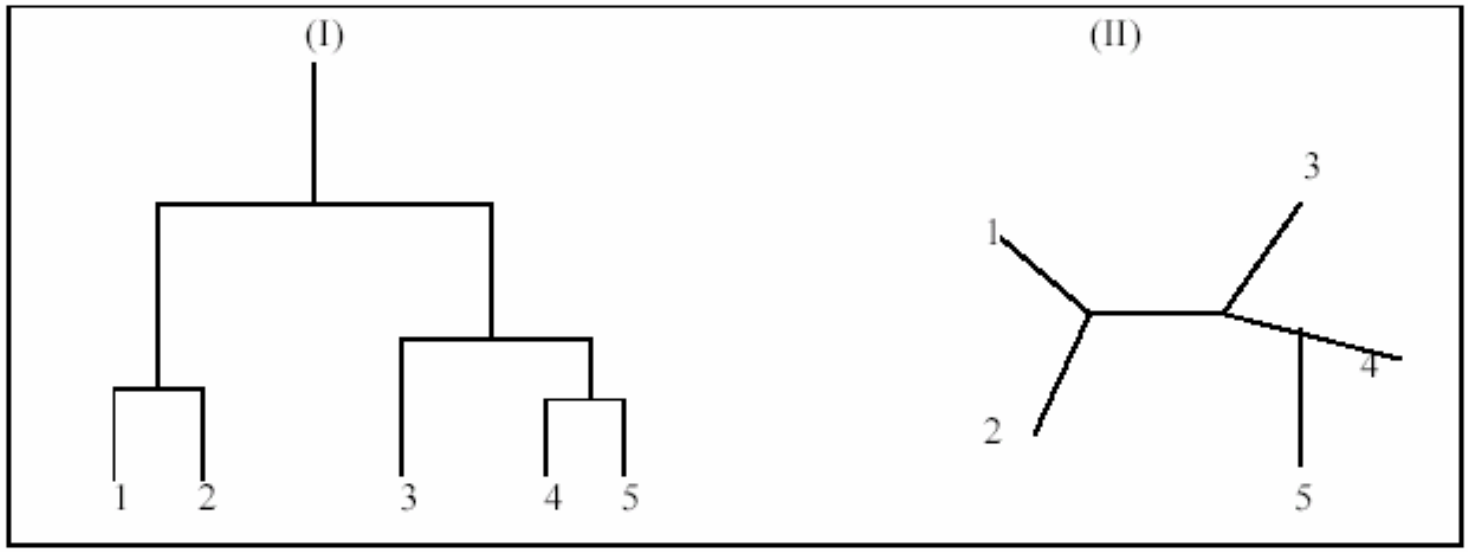
Orthologs & Paralogs



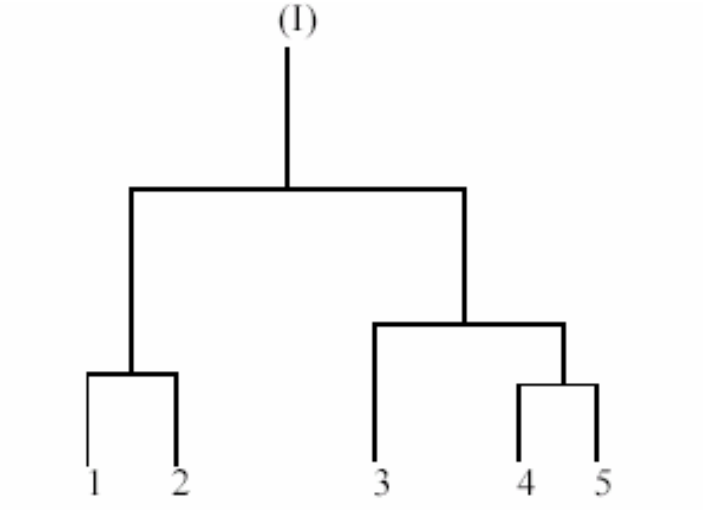
Όταν έχουμε κάποιες ακολουθίες και θέλουμε να εκτιμήσουμε τις φυλογενετικές σχέσεις, μια αναπαράσταση σε μορφή δέντρου μας δείχνει πόσο κοντά βρίσκεται η μια ακολουθία στην άλλη, δηλαδή με ποια σειρά οι ακολουθίες εξελίχθηκαν η μια από την άλλη, και γυρνώντας πίσω στο χρόνο να εντοπίσουμε τελικά τον κοινό τους πρόγονο. Παρακάτω δίνονται δυο παραδείγματα δέντρων με 2 και 8 ακολουθίες αντίστοιχα.



Τα δέντρα που είδαμε στο παραπάνω διάγραμμα είναι δέντρα με ρίζα (rooted). Σε αυτά έχουμε ξεκάθαρη κατεύθυνση του χρόνου, και έτσι μπορούμε να προσδιορίσουμε τον αρχαίο κοινό προγονό. Εναλλακτικά μπορούμε να έχουμε δέντρα χωρίς ρίζα (unrooted), στα οποία δεν μπορούμε να προσδιορίσουμε την κατεύθυνση κατά την οποία εχει συντελεστεί η εξελικτική διαδικασία. Παρακάτω βλέπουμε ένα παράδειγμα για δέντρο με ρίζα (I), και ένα χωρίς ρίζα (II), αμφότερα για 5 ακολουθίες.

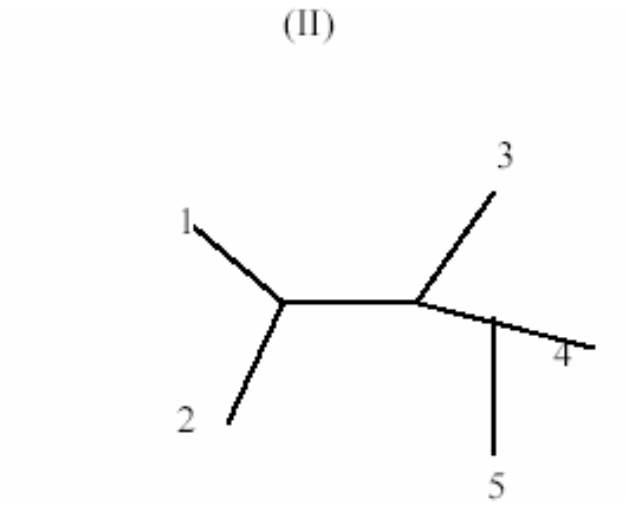


Δέντρα με ρίζα (rooted trees)



$$N_{rooted} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Δέντρα χωρίς ρίζα (unrooted trees)



$$N_{unrooted} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Για να προσθέσουμε ρίζα επιλέγουμε ένα Outgroup

Τα βήματα μιας φυλογενετικής ανάλυσης

- Στοίχιση (πολλές φορές χρειάζεται manual editing)
- Καθορισμός του μοντέλου αντικατάστασης
- Κατασκευή του δέντρου
- Αξιολόγηση του δέντρου

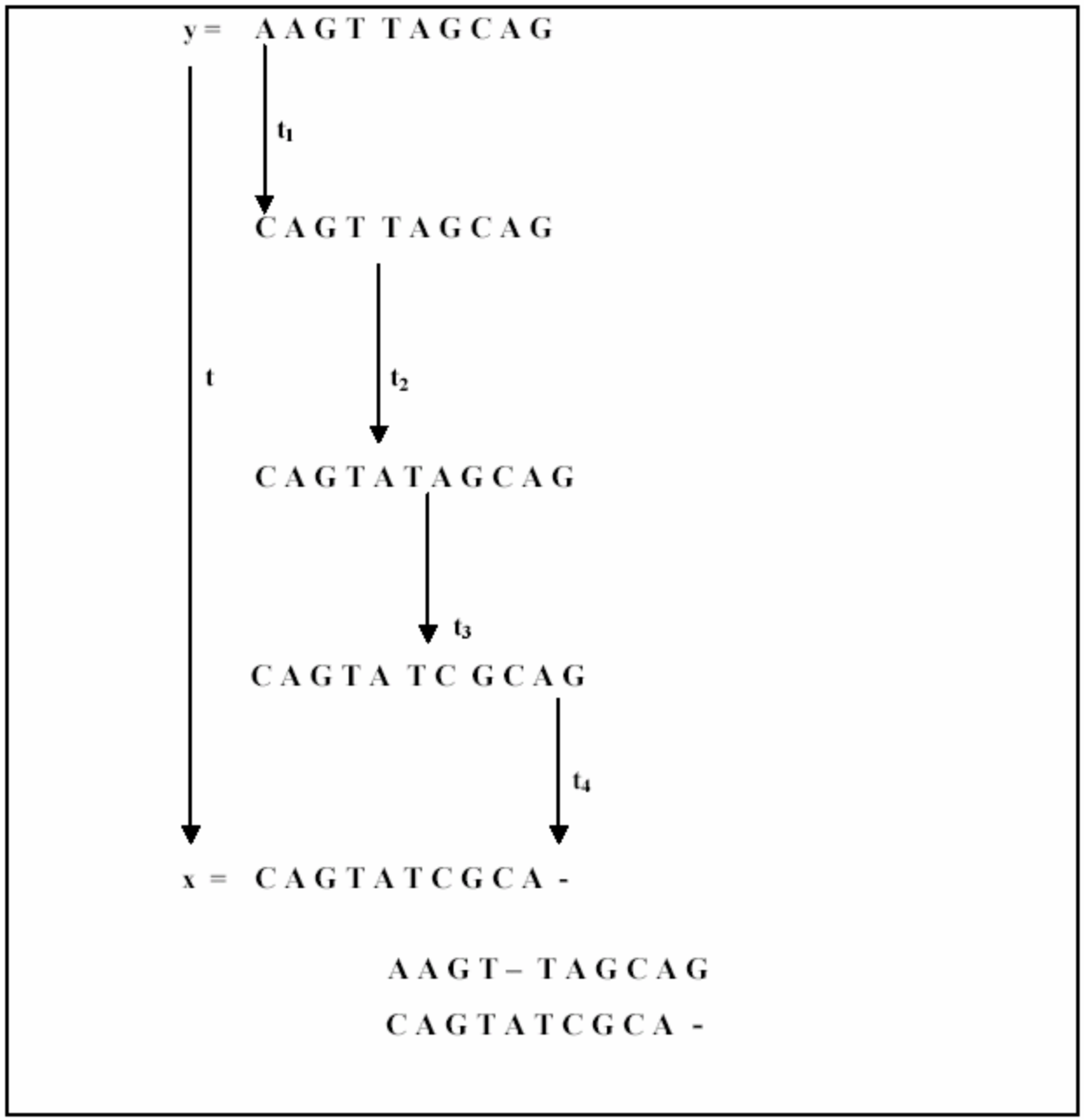
Πιθανοθεωρητικά μοντέλα της εξέλιξης

- Προυποθέτουν μια Μαρκοβιανή ανέλιξη για το ρυθμό αντικαταστάσεων των νουκλεοτιδίων

Έτσι οι πιθανότητες μεταβάσεως θα είναι :

$$p_{abt} = P(x_i = b \mid x_i = a, t)$$

δηλαδή η πιθανότητα το νουκλεοτίδιο b να αντικαταστήσει το α στην i θέση της ακολουθίας x έπειτα από χρόνο t. Όπως είναι φανερό εδώ έχουμε μια ανέλιξη Markov διακριτών καταστάσεων σε συνεχή χρόνο. Αν τώρα έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, \ldots, x_n$ και $\mathbf{y} = y_1, y_2, \ldots, y_n$ η πιθανότητα η $\mathbf{x}$ να έχει προκύψει από την $\mathbf{y}$ σε χρόνο t είναι $P(\mathbf{x} \mid \mathbf{y}, t) = \prod_{i=1}^{n} P(x_i \mid y_i, t)$


y = A A G T T A G C A G
t1
C A G T T A G C A G
t
t2
C A G T A T A G C A G
t3
C A G T A T C G C A G
t4
x = C A G T A T C G C A -
A A G T – T A G C A G
C A G T A T C G C A -

Ορίζουμε στη συνέχεια έναν 4X4 πίνακα πιθανοτήτων μεταβάσεως ή υποκαταστάσεως ο οποίος εξαρτάται από το t,

$$S(t) = \begin{bmatrix} P(A \mid A,t) & P(T \mid A,t) & P(G \mid A,t) & P(C \mid A,t) \\ P(A \mid T,t) & P(T \mid T,t) & P(G \mid T,t) & P(C \mid T,t) \\ P(A \mid G,t) & P(T \mid G,t) & P(G \mid G,t) & P(C \mid G,t) \\ P(A \mid C,t) & P(T \mid C,t) & P(G \mid C,t) & P(C \mid C,t) \end{bmatrix}$$

Στον πίνακα αυτό πρέπει να ισχύουν

$p_{i,j} \geq 0$ με $i, j = 1,2,3,4$ και

$$\sum_{j=1}^{4} p_{i,j} = 1 \qquad \text{για κάθε } i$$

δηλαδή ο πίνακας αυτός είναι στοχαστικός.

Πρέπει να τονίσουμε εδώ ότι μια ανέλιξη Markov, σαν αυτές που περιγράφουμε εδώ, μπορεί να έχει τρεις βασικές ιδιότητες (Κάκκουλος, 1995; Lio and Goldman, 1998). Πρώτον η αλυσίδα Markov μπορει να είναι ομογενής χρονικά (homogeneity), δηλαδή οι πιθανότητες μεταβάσεως να μην εξαρτώνται από το χρόνο. Σε μια ομογενή αλυσίδα οδηγούμαστε τελικά σε μια κατάσταση ισορροπίας (equilibrium). Δεύτερον, η αλυσίδα Markov να είναι στάσιμη (stationary), δηλαδή σε κάθε χρονική στιγμή η κατανομή των βάσεων είναι αυτή της κατάστασης ισορροπίας. Και τρίτον, είναι δυνατόν να ισχύει η αντιστρεπτοτητα (reversibility) των πιθανοτήτων μεταβάσεως, δηλαδή οι πιθανότητες να είναι ίδιες και για τις αντίστροφες μεταβάσεις. Στα μοντέλα φυλογενετικής εξέλιξης συνήθως υποθέτουμε ότι πληρούν και τις 3 παραπάνω προϋποθέσεις, για λόγους υπολογιστικής απλότητας

Έτσι αν ισχύουν τα παραπάνω τότε οι εξισώσεις Chapman-Kolmogorov γίνονται (Κάκκουλος, 1995):

$$S(t)S(s) = S(t+s)$$

Τέλος είναι αναγκαίο να ορίσουμε έναν πίνακα ρυθμού υποκαταστάσεως (Substitution Rate Matrix) R έτσι ώστε:

$$R = \begin{bmatrix} \delta & \alpha & \beta & \gamma \\ \alpha & \delta & \gamma & \beta \\ \beta & \gamma & \delta & \alpha \\ \gamma & \beta & \alpha & \delta \end{bmatrix}$$

στον οποίο για να πληρούνται και οι 3 παραπάνω προϋποθέσεις, πρέπει να ισχύει:

$$\delta = -(\alpha+\beta+\gamma)$$

δηλαδή οι γραμμές και οι στήλες του να αθροίζουν στο 0.

Επειδή

$$S(t) = \exp(Rt) \cong I + Rt + \frac{(Rt)^2}{2!} + \frac{(Rt)^3}{3!} + \ldots$$

αν προχωρήσουμε σε φασματική διάσπαση (spectral decomposition) επιπλέον ισχύει (Κάκκουλος, 1995):

$$S(t) = U.diag\left\{e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\right\}.U^{-1}$$

όπου λ_i οι ιδιοτιμές (eigenvalues) του R,και U ο πίνακας που τις περιέχει.

Ο πίνακας υποκαταστάσεως για ένα «μικρό» χρονικό διάστημα ε γίνεται:

$$S(\varepsilon) = I + R\varepsilon \Rightarrow S(t+\varepsilon) = S(t)S(\varepsilon) = S(t)(I + R\varepsilon)$$

$$\Rightarrow \frac{S(t+\varepsilon) - S(t)}{\varepsilon} \approx S(t)(I + R\varepsilon)$$

και παίρνοντας το όριο καθώς $\varepsilon \to 0$ έχουμε

$$S'(t) = S(t)R.$$

Λύνοντας αυτές τις εξισώσεις μπορούμε να πάρουμε τις τιμές για τις πιθανότητες μεταβάσεως η υποκαταστάσεως.

Jukes and Cantor (1969)

- Αν α=β=γ, τότε:

$$R=\begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} \qquad S(t)=\begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}$$

$$r_t = \tfrac{1}{4}\left(1+3e^{-4\alpha t}\right) \qquad s_t = \tfrac{1}{4}\left(1-e^{-4\alpha t}\right)$$

- Το ποιο απλό μοντέλο αλλά έχει σημαντικές αδυναμίες
- Οι μεταπτώσεις (πουρίνη σε πουρίνη) δεν έχουν ίδιο ρυθμό εμφάνισης με τις μεταστροφές (πουρίνη σε πυριμιδίνη, και αντίστροφα)

Kimura (1980)

$$R=\begin{bmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{bmatrix} \qquad S(t)=\begin{bmatrix} r_t & s_t & u_t & s_t \\ s_t & r_t & s_t & u_t \\ u_t & s_t & r_t & s_t \\ s_t & u_t & s_t & r_t \end{bmatrix}$$

με λύσεις $s_t = \frac{1}{4}\left(1-e^{-4\beta t}\right)$

$$u_t = \frac{1}{4}\left(1+e^{-4\beta t}-2e^{-2(\alpha+\beta)t}\right)$$

και $r_t = 1-2s_t-u_t$

Όπως είναι φανερό το μοντέλο αυτό είναι δι-παραμετρικό καθώς προβλέπει άλλες πιθανότητες υποκαταστάσεως για μεταπτώσεις (π.χ. $A \leftrightarrow G, T \leftrightarrow C$) και άλλες για μεταστροφές (π.χ. $A \leftrightarrow T, G \leftrightarrow C$)

F81 (Felsenstein, 1981)

- χρησιμοποιεί μόνο μία παράμετρο (μ) για τις μεταπτώσεις και τις μεταστροφές, αλλά μοντελοποιεί τα τέσσερα νουκλεοτίδια με διαφορετικές πιθανότητες εμφάνισης (πA, πG, πC, πT):

$$R = \begin{bmatrix} -\mu(\pi_C + \pi_G + \pi_T) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_C + \pi_A + \pi_T) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_C + \pi_G + \pi_A) \end{bmatrix}$$

HKY85

- οι μεταπτώσεις έχουν διαφορετικό ρυθμό από τις μεταστροφές, χρησιμοποιώντας δύο παραμέτρους (κ και μ):

$$R = \begin{bmatrix} -\mu\left(\pi_C + \kappa\pi_G + \pi_T\right) & \mu\pi_C & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\pi_A & -\mu\left(\pi_A + \pi_G + \kappa\pi_T\right) & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu\left(\pi_C + \kappa\pi_A + \pi_T\right) & \mu\kappa\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu\left(\kappa\pi_C + \pi_G + \pi_A\right) \end{bmatrix}$$

GTR

- Τέλος, το πιο γενικό μοντέλο αυτής της κατηγορίας, είναι το λεγόμενο γενικό χρονικά αντιστρεπτό μοντέλο (general time reversible model), το οποίο συμβολίζεται ως GTR (Tavare, 1986) και περιγράφεται από τη σχέση:

$$R = \begin{bmatrix} -\mu\left(a\pi_C + b\pi_G + c\pi_T\right) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu\left(a\pi_A + d\pi_G + e\pi_T\right) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu\left(d\pi_C + b\pi_A + f\pi_T\right) & \mu\kappa\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu\left(e\pi_C + f\pi_G + c\pi_A\right) \end{bmatrix}$$

Συνέπεια των απλών μοντέλων…

- Σε κατάσταση ισορροπίας ($t \to \infty$), έχουμε ισοκατανομή των νουκλεοτιδίων
- Αυτό όμως ξέρουμε ότι δεν ισχύει καθώς οι οργανισμοί διαφέρουν στο ποσοστό GC%
- Άλλα μοντέλα, πιο περίπλοκα έχουν προταθεί (Lio and Goldman, Felsenstein, Penny and Hendy) όπου ο πίνακας δεν έχει την συμμετρία που είδαμε πρίν

Κατασκευή των δέντρων

- Μέθοδοι βασισμένες στην απόσταση (distance-based methods)
- Μέθοδοι βασισμένες στους χαρακτήρες (character-based methods)

Μέθοδοι βασισμένες στην απόσταση

- UPGMA (unweighted pair group method using arithmetic mean) (Clustering)
- Ενώνει τους βραχίονες με τη μεγαλύτερη ομοιότητα, και παίρνει το μέσο όρο για να ορίσει το νέο cluster.
- Αναμένεται να αποδώσει τα μέγιστα, αν ισχύει το «μοριακό ρολόι»

$d_{ii}=0$

$d_{ij}= d_{ji} >0$ για $i \neq j$

$d_{ij} \leq d_{ik}+ d_{kj}$ (τριγωνική ανισότητα)

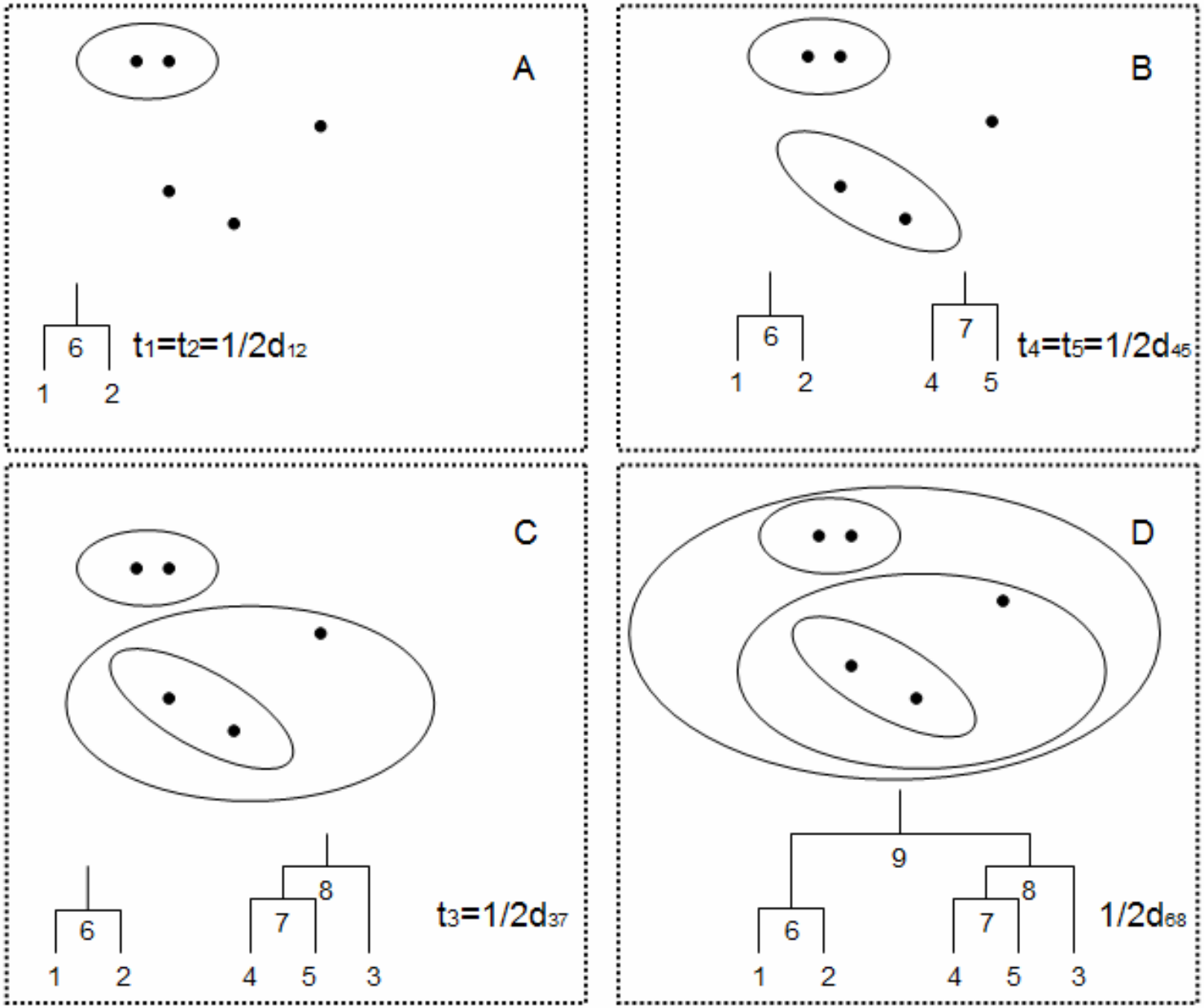
$$d_{ij} = \frac{1}{|C_i||C_j|}\sum d_{pq}$$

και η απόσταση του Cluster των i,j (k) με μια άλλη ακολουθία l είναι

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}$$

- Ο πίνακας των αποστάσεων δυο ακολουθιών και, θα μπορούσε γενικά να προκύψει με πολλούς τρόπους. Ένας εύκολος τρόπος θα ήταν μετρώντας απλά το ποσοστό από τις θέσεις, στις οποίες τα κατάλοιπα διαφέρουν. Αυτό είναι ένα λογικό μέτρο, αλλά δεν αποδίδει καλά για ασυσχέτιστες ακολουθίες, καθώς θέλουμε σε αυτή την περίπτωση η απόσταση να αυξάνει. Μια καλύτερη λύση, προκύπτει από την αρχική πολλαπλή στοίχιση με χρήση κάποιου από τα πιθανοθεωρητικά μοντέλα της εξέλιξης που παρουσιάσαμε στην προηγούμενη παράγραφο. Για παράδειγμα, το μοντέλο JC69 δίνει την απόσταση:

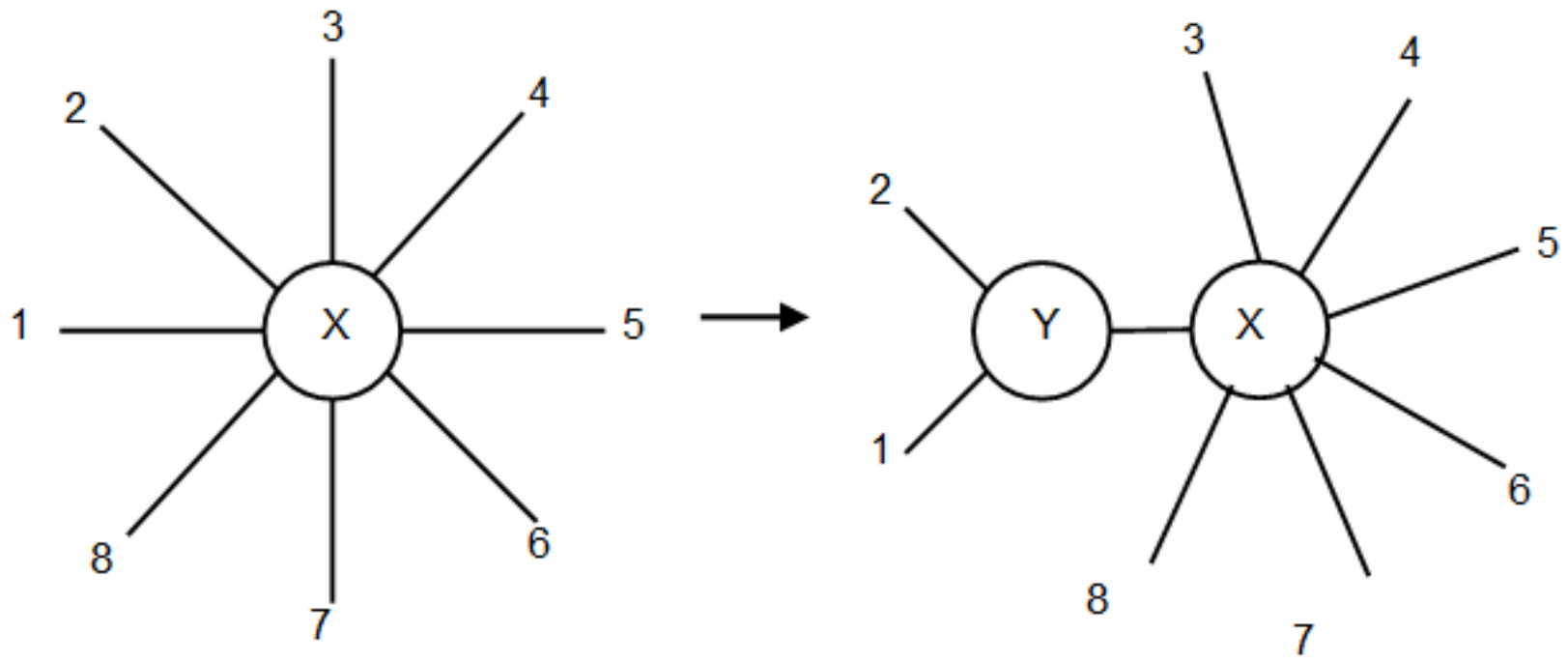
$$d_{ij} = -\frac{3}{4}\log\left(1-\frac{4}{3}f\right) \qquad d_{ij} = -\frac{1}{2}\log\left(1-2f-g\right)-\frac{1}{4}\log\left(1-2g\right)$$


A
6
1
2
t1=t2=1/2d12
B
6
7
1
2
4
5
t4=t5=1/2d45
C
6
8
7
1
2
4
5
3
t3=1/2d37
D
9
6
8
7
1
2
4
5
3
1/2d68

Μέθοδοι βασισμένες στην απόσταση

- Neighbor-Joining Method (Ένωση Γειτόνων)
- Ενώνει διαδοχικά τα πιο κοντινά ζευγάρια, παγίωνοντας την θέση τους σαν ένα νεό κλάδο
- Προχωράει προοδευτικά μέχρι τέλους
- Είναι αρκετά γρήγορη

$$D_{ij} = d_{ij} - \frac{1}{L-2}\sum\left(d_{ik} + d_{jk}\right)$$


3
2
4
1
X
5
8
6
7
3
4
2
5
Y
X
1
6
8
7

Μέθοδοι βασισμένες στην απόσταση

- Fitch and Margoliash (FM)
- Μεγιστοποιεί την προσαρμογή (fit) των ανά δυο αποστάσεων σε ένα δέντρο
- Ελαχιστοποιεί την τετραγωνική απόκλιση των αποστάσεων, σε σχέση με όλα τα πιθανά μήκη μονοπατιών στο δέντρο
- Διάφορες παραλλαγές για το στάθμισμα του σφάλματος

Μέθοδοι βασισμένες στους χαρακτήρες (character-based methods)

- Μέγιστη Πιθανοφάνεια (Maximum Likelihood)
- Μέγιστη Φειδωλότητα (Maximum Parsimony)

Μέγιστη Πιθανοφάνεια (Maximum Likelihood)

- Η πιθανοφάνεια είναι δεσμευμένη στην αρχική στοίχιση
- Πρέπει να αθροιστεί για όλες τις πιθανές τοπολογίες του δέντρου
- Δεν μπορεί να υπολογιστεί αναλυτικά

$$\mathbf{X_1} = x_{11}, x_{12}, ..., x_{1n}$$
$$\mathbf{X_2} = x_{21}, x_{22}, ..., x_{2n}$$

η πιθανότητα το *i* νουκλεοτίδιο στις δυο ακολουθίες να έχει προκύψει από κάποιο *α* αρχικό είναι :

$$P(x_{1i}, x_{2i}, a \mid T, t_1, t_2) = q_a P(x_{i1} \mid a, t_1) P(x_{i2} \mid a, t_2)$$

όπου *α* το άγνωστο αρχικό νουκλεοτίδιο (τ.μ), *T* το υποτιθέμενο δέντρο, και t_1, t_2 τα μήκη των βραχιόνων του δέντρου (χρόνος κατά τον οποίο έχουν αποκλίνει εξελικτικά).

Επειδή τώρα δεν γνωρίζουμε το α πρέπει να αθροίσουμε όλες τις εναλλακτικές άρα:

$$P(x_{1i}, x_{2i} \mid T, t_1, t_2) = \sum_a q_a P(x_{i1} \mid a, t_1) P(x_{i2} \mid a, t_2)$$

και κατόπιν μπορούμε να υπολογίσουμε τη συνολική πιθανότητα για τις n θέσεις των δυο ακολουθιών ως εξής:

$$P(\mathbf{x_1}, \mathbf{x_2} \mid T, t_1, t_2) = \prod_{i=1}^{n} P(x_{1i}, x_{2i} \mid T, t_1, t_2)$$

Αυτή είναι η πιθανοφάνεια των δυο ακολουθιών (likelihood). Για αριθμητική ευκολία εργαζόμαστε συνήθως με το λογάριθμό της (log-likelihood) που είναι:

$$\log P(\mathbf{x_1}, \mathbf{x_2} \mid T, t_1, t_2) = \sum_{i=1}^{n} \log P(x_{1i}, x_{2i} \mid T, t_1, t_2)$$

Στη γενικότερη περίπτωση που χρησιμοποιούμε *r* ακολουθίες έχουμε:

$$P(x_{1i}, x_{2i}, ..., x_{ri} \mid T, t^{\bullet}) =$$

$$= \sum_{a^{r+1}, ..., a^{2r-1}} q_{a^{2r-1}} \prod_{k=r+1}^{2r-2} P(a^k \mid a^{a(k)}, t_k) \prod_{k=1}^{r} P(x_{ki} \mid a^{a(k)}, t_k)$$

Οι επιπλέον συμβολισμοί που εισάγουμε εδώ είναι το *α(k)* που συμβολίζει το αρχικό νουκλεοτίδιο για κάθε ένα από τα παρακλάδια του δέντρου. Από την συνδυαστική βρίσκουμε ότι για *r* ακολουθίες έχουμε *2r-1* παρακλάδια και *2r-2* σημεία διασταύρωσης των κλαδιών, για ένα δέντρο με ρίζα. Από αυτά τα πρώτα *r* αντιστοιχούν στα παρακλάδια (βραχίονες) που οδηγούν σε μια ακολουθία ενώ τα υπόλοιπα από *r*+1εως *2r-2* αντιστοιχούν στα κλαδιά (υπέρ-βραχίονες) που ομαδοποιούν τις ακολουθίες. Έτσι δικαιολογούνται όλοι οι δυνατοί συνδυασμοί και τα γινόμενα στην παραπάνω εξίσωση, και το τελικό άθροισμα είναι για τα κλαδιά του δέντρου από το *r+1* έως το *2r-1* (είναι ένα παραπάνω γιατί πρέπει να μετρήσουμε και το *α* στη ρίζα του δέντρου). Τελικά η πιθανοφάνεια για τις *r* ακολουθίες θα είναι αντίστοιχα :

$$P(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_r} \mid T, t_\circ) = \prod_{i=1}^{n} P(x_{1i}, x_{2i}, ..., x_{ri} \mid T, t_\circ)$$

και δουλεύοντας ως συνήθως με το λογάριθμο της (log-likelihood) θα έχουμε:

$$\log P(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x} \mid T, t_\circ) = \sum_{i=1}^{n} \log P(x_{1i}, x_{2i}, ..., x_{ri} \mid T, t_\circ)$$

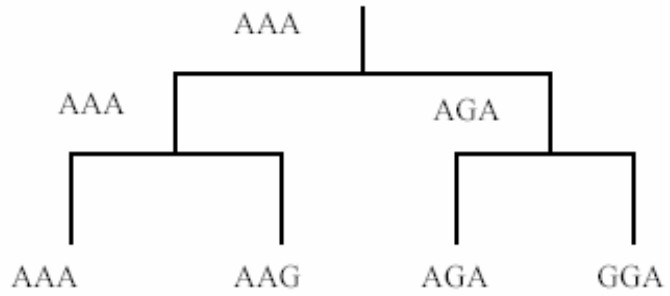
Μέγιστη Φειδωλότητα (Maximum Parsimony)

- Διαφέρει ριζικά από τις προηγούμενες, στο ότι κάνει διάκριση μεταξύ πληροφοριακών και μη-πληροφοριακών θέσεων στις ακολουθίες, με τις πληροφοριακές θέσεις να είναι αυτές που παρουσιάζουν πολυμορφισμό (ύπαρξη πάνω από δυο ειδών νουκλεοτιδίων) τουλάχιστον δυο φορές.
- Εφαρμόζεται στην εξελικτική βιολογία προτού να εμφανιστεί η μοριακή φυλογένεια (π.χ. εφαρμοζόταν σε διάφορα φαινοτυπικά χαρακτηριστικά) και έχει σκοπό να εξηγήσει τις εξελικτικές διάφορες με το μικρότερο αριθμό αλλαγών.
- Προτάθηκε αρχικά, σαν υπολογιστική προσέγγιση στη Μέγιστη Πιθανοφάνεια
- Έχει βρεθεί στο επίκεντρο πολλών αντιπαραθέσεων

Παράδειγμα

Έστω ότι έχουμε τις παρακάτω 4 ακολουθίες που είναι ήδη στοιχισμένες, και θέλουμε να βρούμε ένα φυλογενετικό δέντρο με τη χρήση της μεθόδου της φειδωλότητας.

S1	AAG
S2	AAA
S3	GGA
S4	AGA

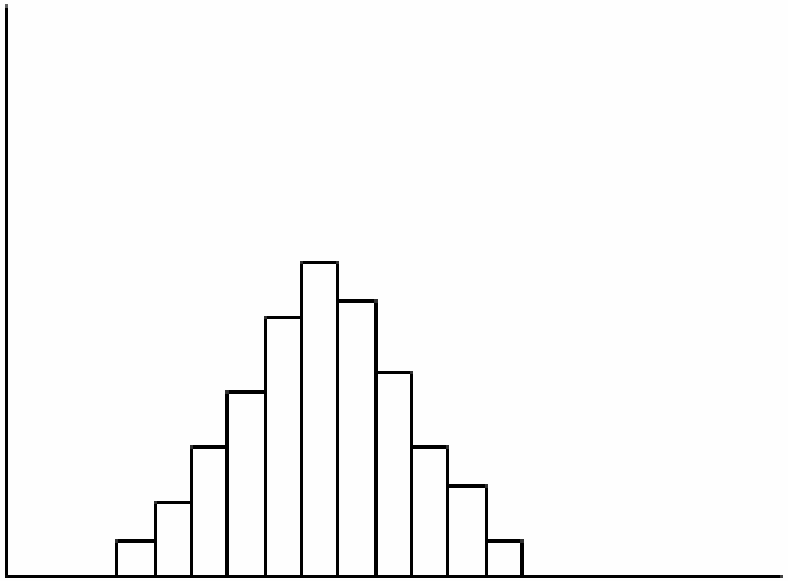


Το δέντρο που δίνεται παραπάνω είναι αυτό το οποίο εξηγεί τις νουκλεοτιδικές αλλαγές με τον μικρότερο αριθμό αντικαταστάσεων (3 συνολικά) από όλα τα άλλα δέντρα με ρίζα (συνολικά 15 τέτοια δέντρα)

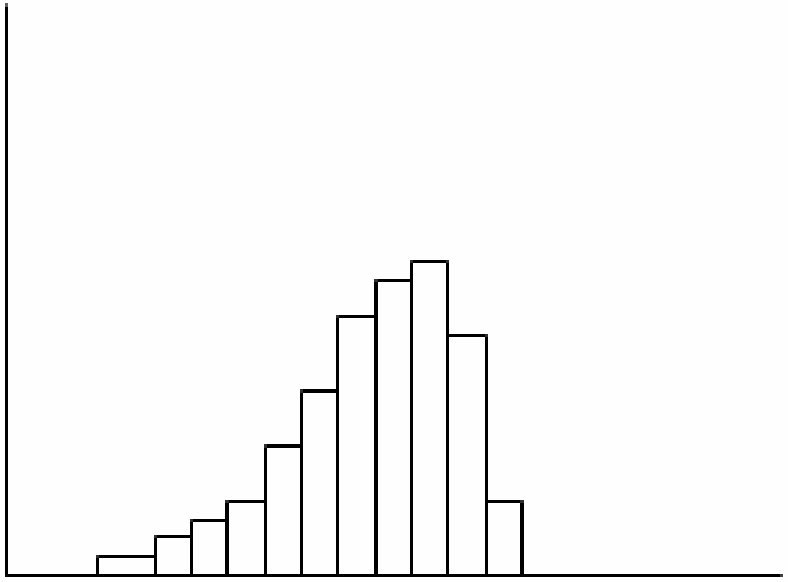
Αξιολόγηση των Δέντρων

- Τυχαιοποιημένα Δέντρα
- Permutation Tail Probability Test
- Bootstrap (parametric, non-parametric)
- Likelihood Ratio Test

Seq	1	2	3	4	5	6	7	8
A	A	T	A	G	C	G	C	T
B	T	A	A	G	C	G	C	T
C	T	A	A	C	G	C	C	T
D	A	T	A	C	G	G	C	T



Seq	1	2	3	4	5	6	7	8
A	A	A	A	G	G	G	C	T
B	T	T	A	G	G	G	C	T
C	T	T	A	C	C	C	C	T
D	A	A	A	C	C	G	C	T



Αρχικά δεδομένα

Seq	1	2	3	4	5	6	7	8
A	A	T	A	G	C	G	C	T
B	T	A	A	G	C	G	C	T
C	T	A	A	C	G	C	C	T
D	A	T	A	C	G	G	C	T

Seq	1	2	2	4	5	5	7	8
A	A	T	T	G	C	C	C	T
B	T	A	A	G	C	C	C	T
C	T	A	A	C	G	G	C	T
D	A	T	T	C	G	G	C	T

Δείγμα 1

Seq	1	1	3	4	5	6	8	8
A	A	A	A	G	C	G	T	T
B	T	T	A	G	C	G	T	T
C	T	T	A	C	G	C	T	T
D	A	A	A	C	G	G	T	T

Δείγμα 2

Seq	1	3	3	3	5	6	7	8
A	A	A	A	A	C	C	C	T
B	T	A	A	A	C	C	C	T
C	T	A	A	A	G	G	C	T
D	A	A	A	A	G	G	C	T

Δείγμα 3

Συμπεράσματα

- Συμπερασματικά, δεν μπορούμε να αποφανθούμε με 100% σιγουριά για το ποια μέθοδος είναι καλύτερη κάτω από όλες τις περιστάσεις, και έτσι χρειάζεται προσοχή όταν έχουμε να εκτιμήσουμε ένα φυλογενετικό δέντρο.
- Σε γενικές γραμμές, η μέγιστη πιθανοφάνεια, φαίνεται να έχει κερδίσει στη σχετική διαμάχη, κυρίως λόγω του στέρεου μαθηματικού υποβάθρου, της δυνατότητας χρήσης πολλών εξελικτικών μοντέλων αλλά και της ευκολίας την οποία προσδίδουν οι σύγχρονοι υπολογιστές και η αυξημένη υπολογιστική ισχύς. Επιπλέον δε, φαίνεται να αποδίδει καλύτερα την ανακατασκευή δέντρων κάτω από τα περισσότερα σενάρια προσομοιώσεων.
- Παρόλα αυτά, η μέθοδος NJ και η φειδωλότητα εξακολουθούν να είναι δημοφιλείς ειδικά για γρήγορες αναλύσεις μεγάλου όγκου δεδομένων.
- Γενικά επειδή η διαδικασία κατασκευής ενός δέντρου περιλαμβάνει 3 διακριτές λειτουργίες, δηλαδή: 1) το κριτήριο καταλληλότητας για το πόσο καλά «προσαρμόζονται» τα δεδομένα στο δέντρο, 2) τη στρατηγική αναζήτησης για να βρούμε το καλύτερο δέντρο, και τέλος 3) τον έλεγχο των προϋποθέσεων κάτω από τις οποίες έχει συντελεστεί η εξέλιξη, είναι δυνατόν να έχουμε συνδυασμό πολλών μεθόδων, πράγμα που εκμεταλλεύονται αρκετά από τα σύγχρονα λογισιμικά τα οποία παρουσιάζουμε στην επόμενη παράγραφο. Για παράδειγμα, η μη παραμετρική bootstrap μπορεί να χρησιμοποιηθεί σαν μέθοδος αξιολόγησης με κάθε μέθοδο κατασκευής δέντρων, ενώ τα κλασικά μοντέλα της εξέλιξης (πχ JC69, K2P κλπ), μπορούν να χρησιμοποιηθούν τόσο με τη NJ (και την UPGMA), όσο και με τη μέγιστη πιθανοφάνεια (αλλά προσοχή, όχι με τη φειδωλότητα!).
- Ο Fenselstein έδειξε επιπλέον, ότι χρησιμοποιόντας οποιοδήποτε από τα γνωστά μοντέλα της εξελικτικής διαδικασίας, μπορούν να οριστούν «αποστάσεις μέγιστης πιθανοφάνειας» (maximum likelihood distance), οι οποίες έχουν την προσθετική ιδιότητα . Αυτές οι αποστάσεις, μπορούν να χρησιμοποιηθούν με οποιαδήποτε μέθοδο αποστάσεων (NJ, UPGMA) για να δώσουν μια μέθοδο η οποία θα δίνει καλύτερα αποτελέσματα.
- Μια άλλη υβριδική μέθοδος είναι η NJML, η οποία αποτελεί συνδυασμό των Neighbour Joining και Maximum Likelihood. Στο πρώτο βήμα κατασκευάζει ένα δέντρο με NJ και η αναζήτηση των πιθανών δέντρων με τη μέθοδο μέγιστης πιθανοφάνειας γίνεται μόνο στα κλαδιά με μεγάλη τιμή bootstrap. Η NJML έδειξε ότι πετυχαίνει καλύτερα αποτελέσματα από την κλασική NJ αλλά σε χρόνο που είναι πολύ καλύτερος σε σχέση με τις ιδιαίτερα απαιτητικές μεθόδους πιθανοφάνειας

Πρακτικές Συμβουλές

- Έλεγχος των δεδομένων εισόδου (garbage in, garbage out)
- Χρήση διαφορετικών μεθόδων (αν υπάρχει κάτι σημαντικό, όλες θα δείξουν το ίδιο)
- Έλέγχος της σειράς των ακολουθιών (κάποιες μέθοδοι παράγουν άλλα αποτελέσματα ανάλογα με τη σειρά εισόδου)
- Επιλογή Outgroup

Διαθέσιμο Software

- PAUP (http://paup.csit.fsu.edu/)
- PHYLIP (http://evolution.genetics.washington.edu/phylip.html)
- PAML (http://abacus.gene.ucl.ac.uk/software/paml.html)
- HYPHY (http://www.hyphy.org)
- MrBayes (http://mrbayes.csit.fsu.edu/)
- MEGA (http://www.megasoftware.net/)
- PhyML (http://www.atgc-montpellier.fr/phyml/binaries.php)
- RAxML (http://scoh-its.org/exelixis/software.html)
- BEAST (http://beast.bio.ed.ac.uk)
- GARLI (http://code.google.com/p/garli)
- TNT (http://www.lillo.org.ar/phylogeny/tnt/)
- Treeview (http://taxonomy.zoology.gla.ac.uk/rod/rod.html)