Κεφάλαιο 6

Φυλογενετική Ανάλυση

Σύνοψη
Στο κεφάλαιο αυτό εξετάζονται οι υπολογιστικές όψεις της φυλογενετικής ανάλυσης, δηλαδή, της διαδικασίας εκτίμησης των εξελικτικών σχέσεων των οργανισμών, μέσα από τη μελέτη των αντίστοιχων βιολογικών αλληλουχιών τους. Θα δούμε στην αρχή τους βασικούς ορισμούς για τα φυλογενετικά δέντρα και τα βασικά πιθανοθεωρητικά μοντέλα της εξέλιξης αλληλουχιών. Κατόπιν, θα παρουσιάσουμε τις βασικές κατηγορίες μεθόδων κατασκευής φυλογενετικών δέντρων, και θα σχολιάσουμε τις ομοιότητες και τις διαφορές τους. Τέλος, θα παρουσιάσουμε τα αντίστοιχα πακέτα λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό, θα σχολιάσουμε τα σχετικά πλεονεκτήματα και μειονεκτήματα τους, και θα δώσουμε πρακτικές συμβουλές.

Προαπαιτούμενη γνώση
Βασικές γνώσεις εξελικτικής βιολογίας. Βασικές γνώσεις πιθανοτήτων. Το κεφάλαιο απαιτεί επίσης κατανόηση των μεθόδων του κεφαλαίου 3 και του κεφαλαίου 4.

6. Εισαγωγή

Το θέμα που θα μας απασχολήσει σε αυτό το κεφάλαιο είναι το πρόβλημα του προσδιορισμού των φυλογενετικών σχέσεων, δηλαδή, το πως θα μπορέσουμε από την αμινοξική αλληλουχία κάποιων πρωτεϊνών (ή τις περισσότερες φορές, από την αλληλουχία των αντιστοίχων γονιδίων), οι οποίες προέρχονται από διάφορους οργανισμούς, να προσδιορίσουμε τις εξελικτικές σχέσεις των οργανισμών αυτών.

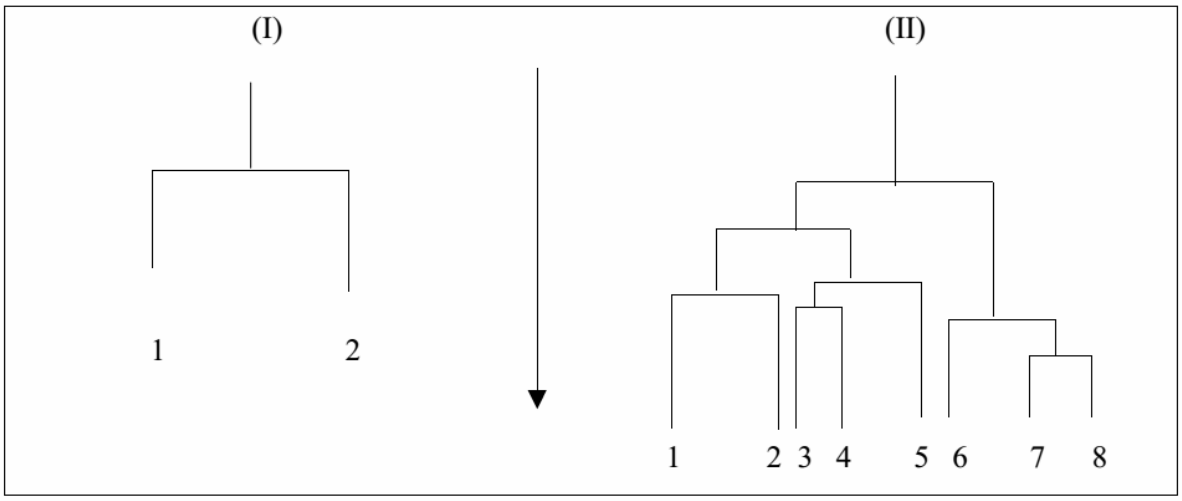
Το θέμα αυτό είναι τεράστιο με πολλές προεκτάσεις (φιλοσοφικές, αλλά και ιδεολογικές) και έχει γίνει από την εποχή του Darwin αντικείμενο για πολλές διαμάχες. Για τους πολέμιους της εξελικτικής θεωρίας (που στη βιολογία βέβαια είναι ελάχιστοι) το θέμα σταματά εδώ καθώς τίποτα από τα παρακάτω δεν έχει νόημα (και ίσως πολλά από όσα προηγήθηκαν), καθώς όπως είπε και ο Thedosius Dobzhansky: «*Nothing in Biology makes sense except in the light of evolution*». Για τους υπόλοιπους, το θέμα αποκτά έντονο ενδιαφέρον καθώς αν δεχθούμε ως βασική αλήθεια την ύπαρξη κοινών προγόνων για όλους τους ζώντες οργανισμούς και την εξέλιξη όλων των σημερινών ειδών από παλαιότερα μέσω της μετάλλαξης και της φυσικής επιλογής, το πρόβλημα της εκτίμησης αυτών των εξελικτικών σχέσεων είναι ένα κατ' εξοχήν μαθηματικό και υπολογιστικό πρόβλημα και κάποιες βασικές όψεις του θα προσπαθήσουμε να παρουσιάσουμε εδώ.

Οι μεθοδολογίες τις οποίες θα πραγματευθούμε σε αυτό το κεφάλαιο, έχουν μακρά ιστορία στο χώρο της βιολογίας. Οι επιστήμονες από τον καιρό του Darwin, προσπαθούσαν να κατασκευάσουν φυλογενετικά δέντρα που να αποδίδουν τις εξελικτικές σχέσεις των οργανισμών και χρησιμοποιούσαν αρχικά για το σκοπό αυτό, τα φαινοτυπικά χαρακτηριστικά. Με την ανάπτυξη όμως της μοριακής βιολογίας, τα μοριακά χαρακτηριστικά (δηλαδή, οι αλληλουχίες των γονιδίων και των πρωτεϊνών), είναι αυτά που κέρδισαν το ενδιαφέρον καθώς αυτά αποτελούν το βασικό υπόστρωμα πάνω στο οποίο δρουν οι εξελικτικές δυνάμεις (η μετάλλαξη και η φυσική επιλογή). Κατά συνέπεια, στο κεφάλαιο αυτό, θα παρουσιάσουμε τους βασικούς τρόπους φυλογενετικής μελέτης βιολογικών αλληλουχιών, θα αναδείξουμε τις ομοιότητες αλλά και τις διαφορές μεταξύ τους, θα εστιάσουμε στα σχετικά πλεονεκτήματα και μειονεκτήματα κάθε μιας, και θα παρουσιάσουμε τα βασικότερα εργαλεία λογισμικού που υπάρχουν διαθέσιμα για το σκοπό αυτό.

6.1. Βασικές Αρχές

Κατ' αρχήν πρέπει να είμαστε σίγουροι για το τι συγκρίνουμε. Αν θέλουμε να εκτιμήσουμε φυλογενετικές σχέσεις από τις αλληλουχίες κάποιων γονιδίων, πρέπει να συγκρίνουμε αντίστοιχα γονίδια, για να εντοπίσουμε την ομολογία τους. *Ομόλογες* πρωτεΐνες (ή γονίδια), λέγονται γενικά οι πρωτεΐνες που έχουν προκύψει μέσω της εξέλιξης από κάποιον κοινό πρόγονο. Συνήθως αυτές επιτελούν παρόμοια λειτουργία (κατ' αντιστοιχία με τα ομόλογα όργανα των οργανισμών), και κατά συνέπεια θα έχουν παρόμοια δομή και

αλληλουχία. Τα αντίστοιχα γονίδια σε διαφορετικούς οργανισμούς αναφέρονται και ως *ορθόλογα* (*orthologues*) στη σχετική βιβλιογραφία, και θεωρούμε ότι η όποια διαφοροποίηση τους έχει προκύψει λόγω της ειδογένεσης. Αντίθετα, ομόλογες πρωτεΐνες, ή γονίδια, μέσα στο ίδιο είδος, ονομάζονται *παράλογα* (*paralogues*), και θεωρούμε ότι έχουν προκύψει από γονιδιακό διπλασιασμό και ανεξάρτητη εξέλιξη μέσα στο είδος. Παράδειγμα της πρώτης περίπτωσης είναι οι α-αλυσίδες της αιμοσφαιρίνης των θηλαστικών (π.χ. του ανθρώπου, του χιμπαντζή, του σκύλου κ.α.), ενώ για τη δεύτερη περίπτωση θα μπορούσαμε να αναφέρουμε μέσα στο ίδιο είδος (π.χ. τον άνθρωπο), τις α, β, γ, δ, ε, ζ, θ αλυσίδες της αιμοσφαιρίνης αλλά και τη μυοσφαιρίνη. Τέλος, υπάρχουν και τα λεγόμενα *ξενόλογα* (*xenologues*) γονίδια, τα οποία είναι ομόλογα γονίδια τα οποία έχουν προκύψει από κάποια διαδικασία οριζόντιας γονιδιακής μεταφοράς (συνήθως από προκαρυωτικό οργανισμό).

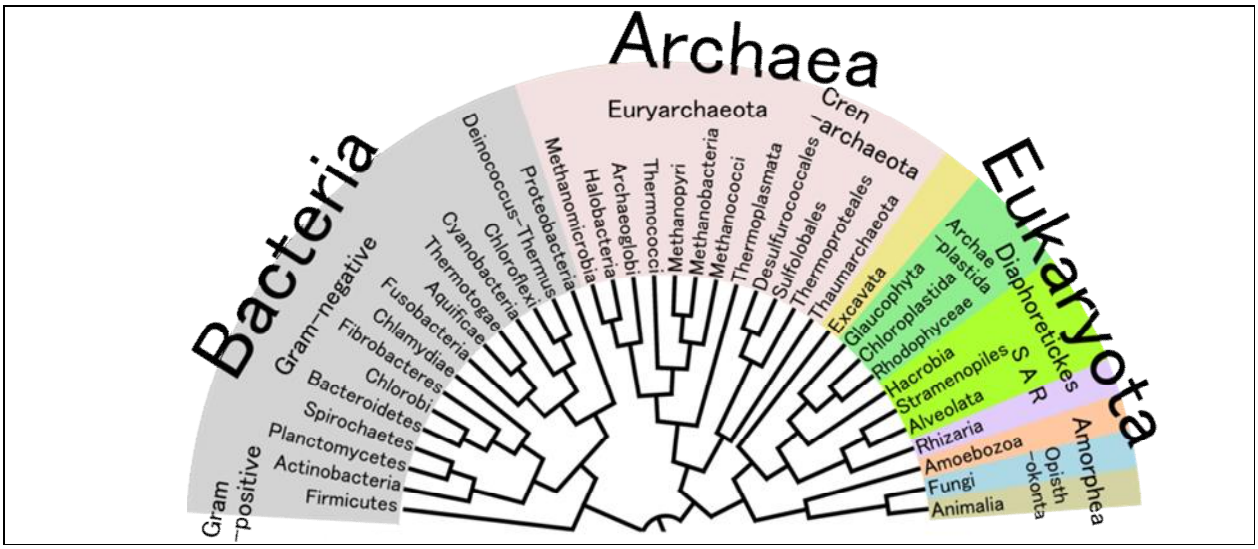


Εικόνα 6.1: *Παράδειγμα ενός δέντρου με δύο κλάδους (I), και ενός άλλου με 8 (II). Και τα δύο δέντρα είναι με ρίζα.*

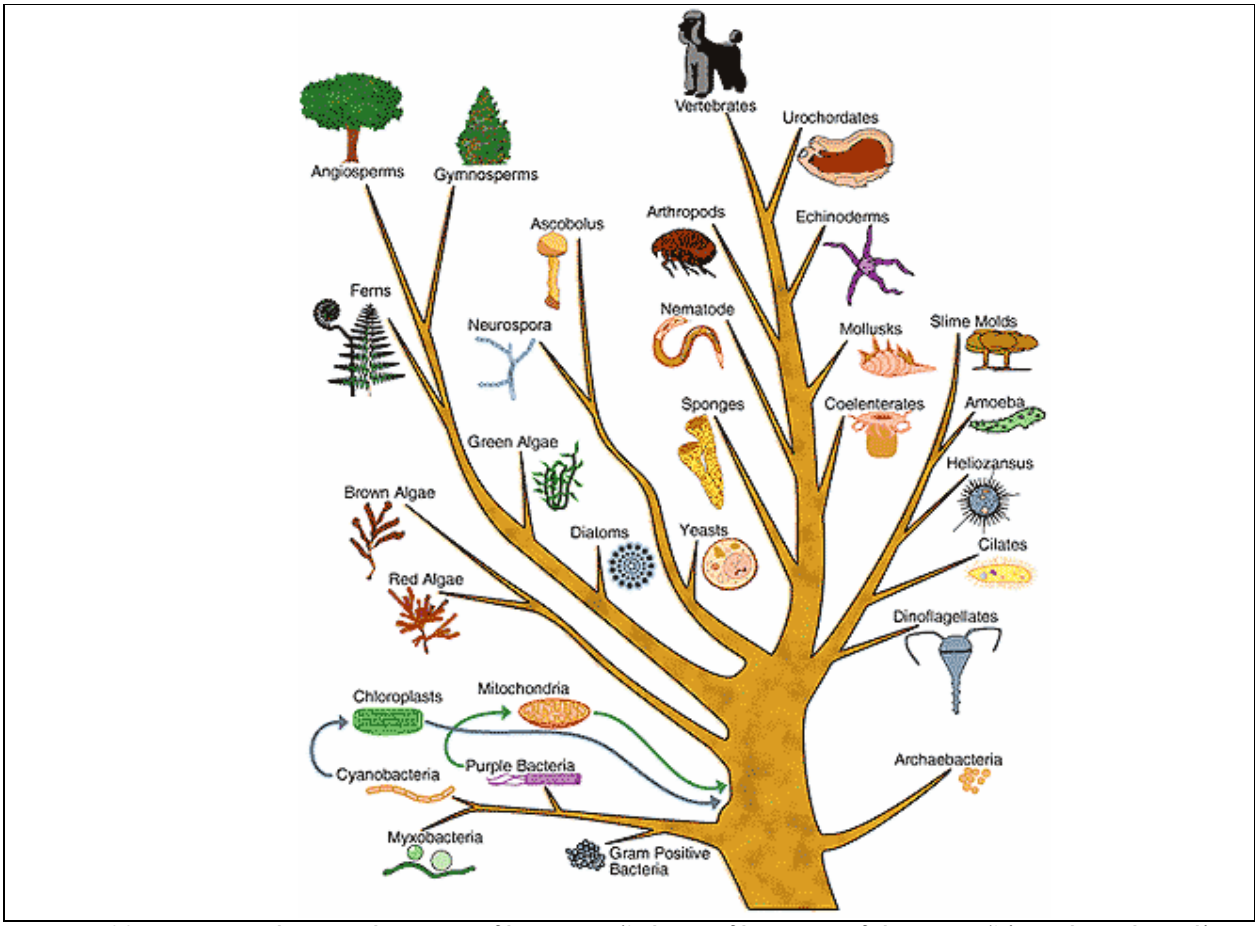
Προτού προχωρήσουμε, είναι απαραίτητο να δώσουμε κάποιους ορισμούς που αφορούν τα φυλογενετικά δέντρα. Ένα φυλογενετικό δέντρο είναι μια αναπαράσταση που συμβολίζει την εξελικτική διαδικασία. Όταν έχουμε κάποιες αλληλουχίες και θέλουμε να εκτιμήσουμε τις φυλογενετικές τους σχέσεις, μια αναπαράσταση σε μορφή δέντρου μας δείχνει πόσο κοντά βρίσκεται η μια αλληλουχία στην άλλη, δηλαδή με ποια σειρά οι αλληλουχίες εξελίχθηκαν η μια από την άλλη, έτσι ώστε, γυρνώντας πίσω στο χρόνο να εντοπίσουμε τελικά τον κοινό τους πρόγονο. Οι ακμές αυτού του δέντρου, είναι οι αλληλουχίες ή γενικότερα, οι ταξινομικές βαθμίδες (taxa) οι οποίες συγκρίνονται. Οι κόμβοι στο δέντρο, δείχνουν τα σημεία διακλάδωσης, δηλαδή το χρονικό σημείο ύπαρξης κοινού προγόνου. Τα μήκη των βραχιόνων, από έναν κόμβο σε μία ακμή, συμβολίζουν τον χρόνο που έχει περάσει. Έναν κλάδο, αποτελούν όλοι οι βραχίονες που ξεκινάνε από έναν κόμβο, και αυτός συμβολίζει μια μονοφυλετική ομάδα (μια ομάδα η οποία περιλαμβάνει όλους τους οργανισμούς που προέρχονται από τον κοινό πρόγονο, χωρίς όμως να περιέχει άλλους οργανισμούς που δεν κατάγονται από αυτόν). Στην Εικόνα 6.1 δίνονται δυο παραδείγματα δέντρων με 2 και 8 αλληλουχίες αντίστοιχα.

Οι βασικές αρχές της φυλογενετικής ανάλυσης, μπορούμε να πούμε ότι στηρίζονται σε μερικές απλές παραδοχές (Brinkman & Leipe, 2001):

- Οποιαδήποτε ομάδα οργανισμών (ή αλληλουχιών) προέρχεται από κάποιον κοινό πρόγονο μέσω της εξέλιξης. Αν οι οργανισμοί (ή οι αλληλουχίες) είναι πολύ διαφορετικοί, ο κοινός πρόγονος υπάρχει αλλά βρίσκεται πολύ πίσω στον εξελικτικό χρόνο.
- Υπάρχει διχαλωτό πρότυπο στην εξέλιξη. Η διαδικασία της εξέλιξης οδηγεί πάντα σε διχοτόμηση ενός taxon ή μιας αλληλουχίας, έτσι ώστε να δημιουργούνται δύο βραχίονες κάτω από έναν κόμβο.
- Αλλαγή στα παρατηρήσιμα χαρακτηριστικά των οργανισμών εμφανίζεται μετά το πέρασμα πολλών γενιών.

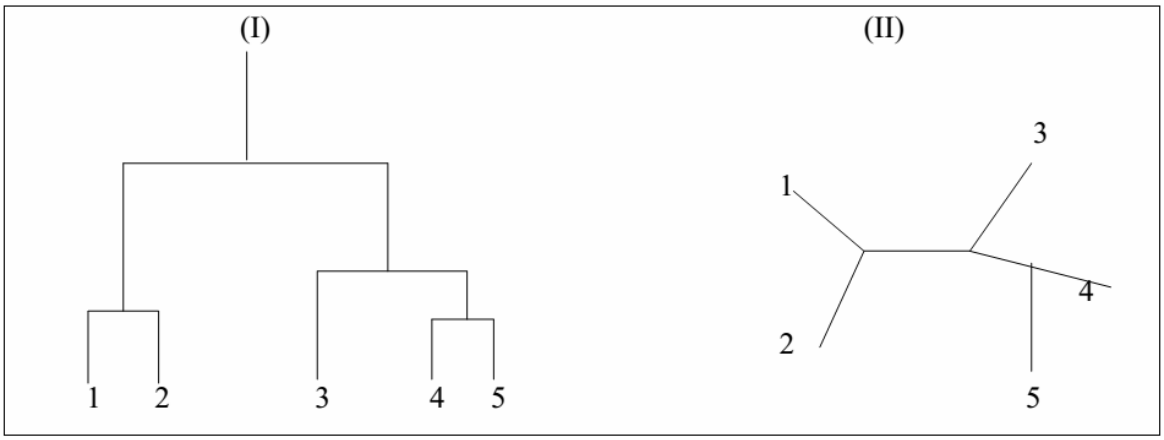


Εικόνα 6.2: *Ένα φυλογενετικό δέντρο όλων των σύγχρονων ομάδων οργανισμών (πηγή: http://commons.wikimedia.org/wiki/File:Phylogenetic_Tree_of_Life.png)*



Εικόνα 6.3: *Μια γραφική αναπαράσταση του δέντρου της ζωής, του δέντρου που δείχνει την εξελικτική συγγένεια όλων των σύγχρονων ομάδων οργανισμών. Η πληροφορία είναι βασικά η ίδια με αυτή της εικόνας 6.2 με τη διαφορά ότι τα μήκη των κλαδιών δεν αντικατοπτρίζουν τους εξελικτικούς χρόνους. Για παράδειγμα, τα βακτήρια και τα αρχαία, φαίνονται σαν δυο μικροί κλάδο, παρόλο που περιέχουν πολλές και διακριτές μεταξύ τους ομάδες οργανισμών (πηγή: http://creationwiki.org/Macroevolution)*

Τα δέντρα που είδαμε στις προηγούμενες εικόνες είναι δέντρα με ρίζα (rooted). Σε αυτά έχουμε ξεκάθαρη κατεύθυνση του χρόνου, και έτσι μπορούμε να προσδιορίσουμε τον αρχαίο κοινό προγονό (ο οποίος όπως είπαμε, είναι σίγουρο ότι υπάρχει). Εναλλακτικά μπορούμε να έχουμε δέντρα χωρίς ρίζα (unrooted), στα οποία δεν μπορούμε να προσδιορίσουμε την κατεύθυνση κατά την οποία έχει συντελεστεί η εξελικτική διαδικασία. Τέτοια δέντρα δημιουργούνται συνήθως από κάποιους αλγορίθμους (είναι περιορισμός των αλγορίθμων αυτών). Παρακάτω (Εικόνα 6.4) βλέπουμε ένα παράδειγμα για δέντρο με ρίζα (I), και ένα χωρίς ρίζα (II), αμφότερα για 5 αλληλουχίες.



Εικόνα 6.4: *Τυπικά παραδείγματα πιθανών δέντρων για 5 αλληλουχίες. (I) δέντρο με ρίζα, (II) δέντρο χωρίς ρίζα*

Το αν θα έχουμε τελικά δέντρο με η χωρίς ρίζα είναι αποτέλεσμα της μεθόδου που χρησιμοποιείται καθώς άλλες μέθοδοι παράγουν δέντρα με ρίζα και άλλες χωρίς. Πάντως ακόμα και αν έχουμε δέντρο χωρίς ρίζα είναι δυνατόν να προστεθεί εκ των υστέρων (και μάλιστα, το επιδιώκουμε αυτό), αν συγκρίνουμε όλους τους οργανισμούς με ένα άλλο είδος για το οποίο ξέρουμε ότι απέχει «πολύ» εξελικτικά, από τα υπό μελέτη είδη του δέντρου (το είδος αυτό ονομάζεται outgroup). Όσον αφορά τον αριθμό των κλάδων που έχουν τα παραπάνω δέντρα για L ακολουθίες, ξέρουμε από την συνδυαστική ότι για τα δέντρα με ρίζα θα είναι $2L$-1 (1,2,…L για τα τελικά κλαδιά που αντιστοιχούν στις L ακολουθίες και L+1, L+2,…, $2L$-1 για τα εσωτερικά κλαδιά), και για τα δέντρα χωρίς ρίζα $2L$-3. Ο αριθμός N των πιθανών δέντρων που αντιστοιχούν σε L ακολουθίες θα είναι, για τα δέντρα με ρίζα

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

ενώ για τα δέντρα χωρίς ρίζα αντίστοιχα, θα έχουμε:

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

Έτσι για παράδειγμα, αν έχουμε L=10 ακολουθίες, τότε μπορεί να κατασκευαστούν $N \approx 35$ εκ. δέντρα με ρίζα και $N \approx 2$εκ.δεντρα χωρίς ρίζα. Γενικά τα πιθανά δέντρα με ρίζες θα είναι $2L$-3 φορές περισσότερα από τα αντίστοιχα χωρίς ρίζα.

Γενικά η διαδικασία φυλογενετικής ανάλυσης και η κατασκευή φυλογενετικών δέντρων, αποτελείται από τέσσερα διακριτά σημεία:

- Μία πολλαπλή στοίχιση. Από αυτήν ξεκινάνε όλα, και όλα βασίζονται σε αυτή. Αν η αρχική στοίχιση είναι λάθος, όλες οι παρακάτω αναλύσεις θα είναι επισφαλείς. Γιαυτό, πολλές φορές χρειάζεται εμπειρία και χειροκίνητη επεξεργασία.
- Καθορισμός του μοντέλου αντικατάστασης, δηλαδή του μαθηματικού μοντέλου της εξελικτικής αλλαγής. Αυτή είναι μια απαίτηση των περισσότερων μεθόδων (με εξαίρεση αυτή της μέγιστης φειδωλότητας), και χρειάζεται ιδιαίτερη προσοχή, καθώς ένα απλό μοντέλο μπορεί να κάνει εύκολους τους υπολογισμούς αλλά μπορεί να μην είναι ρεαλιστικό.
- Κατασκευή του δέντρου. Σε αυτό το σημείο, υπάρχουν οι βασικότερες διαφοροποιήσεις των αλγορίθμων. Κάποιες μέθοδοι είναι γρήγορες, άλλες πιο χρονοβόρες, άλλες κάνουν περισσότερες υποθέσεις κ.ο.κ. Γενικά, οι μέθοδοι χωρίζονται σε δύο μεγάλες κατηγορίες,

στις μεθόδους που χρησιμοποιούν απόσταση και στις μεθόδους που χρησιμοποιούν χαρακτήρες.

- Αξιολόγηση του δέντρου. Αφού το δέντρο κατασκευαστεί, πρέπει να υπάρχει και ένας τρόπος να υπολογιστεί η αξιοπιστία του. Ανάλογα με τη μέθοδο κατασκευής, και με το χρησιμοποιούμενο λογισμικό, μπορεί να υπάρχουν και διαφορετικοί τρόποι ελέγχου της αξιοπιστίας του δέντρου.

Θα προσπαθήσουμε να αναλύσουμε αυτά τα θέματα (με την εξαίρεση της πολλαπλής στοίχισης, την οποία μελετήσαμε σε προηγούμενο κεφάλαιο), με τη σειρά που τέθηκαν παραπάνω. Στο τέλος, θα παρουσιάσουμε το διαθέσιμο λογισμικό και θα δώσουμε μερικές πρακτικές συμβουλές.

6.2. Πιθανοθεωρητικά Μοντέλα της Εξέλιξης των Νουκλεοτιδικών Αλληλουχιών

Το πρώτο θέμα που χρειάζεται σχεδόν σε όλες τις φυλογενετικές αναλύσεις, με την εξαίρεση της φειδωλότητας, και αφού θεωρήσουμε δεδομένη την πολλαπλή στοίχιση, είναι ο καθορισμός του μοντέλου με το οποίο θεωρούμε ότι έχει συντελεστεί η εξελικτική διαδικασία. Προχωρώντας στην κατασκευή πιθανοθεωρητικών μοντέλων που να περιγράφουν την εξέλιξη των νουκλεοτιδικών αλληλουχιών μέσω της μετάλλαξης θα συναντήσουμε την έννοια της ανέλιξης Markov. Η διαφορά εδώ σε σχέση με τα μοντέλα που θα παρουσιαστούν στο κεφάλαιο 8, είναι ότι δεν ενδιαφερόμαστε για το ποιο νουκλεοτίδιο ακολουθεί κάποιο άλλο στην ίδια αλυσίδα αλλά ποιο νουκλεοτίδιο θα αντικαταστήσει ένα συγκεκριμένο, σε κάποια μελλοντική χρονική στιγμή. Έτσι, οι πιθανότητες μεταβάσεως θα είναι (Durbin, Eddy, Krogh, & Mitchison, 1998):

$$p_{abt} = P(x_i = b \mid x_i = a, t) \tag{6.1}$$

δηλαδή η πιθανότητα το νουκλεοτίδιο *b* να αντικαταστήσει το *α* στην θέση *i* της αλληλουχίας *x* έπειτα από χρόνο *t*. Όπως είναι φανερό, εδώ έχουμε μια ανέλιξη Markov διακριτών καταστάσεων σε συνεχή χρόνο. Αν τώρα έχουμε δυο αλληλουχίες DNA $\mathbf{x} = x_1, x_2, ..., x_n$ και $\mathbf{y} = y_1, y_2, ..., y_n$, η πιθανότητα η **x** να έχει προκύψει από την **y** σε χρόνο t είναι

$$P(\mathbf{x} \mid \mathbf{y}, t) = \prod_{i=1}^{n} P(x_i \mid y_i, t) \tag{6.2}$$

Ορίζουμε στη συνέχεια, έναν 4x4 πίνακα πιθανοτήτων μεταβάσεως ή υποκαταστάσεως ο οποίος εξαρτάται από το t,

$$S(t) = \begin{bmatrix} P(A \mid A,t) & P(T \mid A,t) & P(G \mid A,t) & P(C \mid A,t) \\ P(A \mid T,t) & P(T \mid T,t) & P(G \mid T,t) & P(C \mid T,t) \\ P(A \mid G,t) & P(T \mid G,t) & P(G \mid G,t) & P(C \mid G,t) \\ P(A \mid C,t) & P(T \mid C,t) & P(G \mid C,t) & P(C \mid C,t) \end{bmatrix} \tag{6.3}$$

Στον πίνακα αυτό πρέπει να ισχύουν

$$\begin{aligned} &p_{i,j} \geq 0 \text{ με } i, j = 1,2,3,4 \text{ και} \\ &\sum_{j=1}^{4} p_{i,j} = 1 \text{ για κάθε } i \end{aligned} \tag{6.4}$$

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 (Lio & Goldman, 1998). Πρώτον, η αλυσίδα Markov μπορεί 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 αντιστρεπτότητα (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:

$$S(t)S(s)=S(t+s) \tag{6.5}$$

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

$$R=\begin{bmatrix}\delta & \alpha & \beta & \gamma\\ \alpha & \delta & \gamma & \beta\\ \beta & \gamma & \delta & \alpha\\ \gamma & \beta & \alpha & \delta\end{bmatrix} \tag{6.6}$$

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

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

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 αποικοδόμηση (spectral decomposition), ισχύει επιπλέον:

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

όπου λ_i οι ιδιοτιμές (eigenvalues) του R, και U το αντίστοιχο ιδιοδιάνυσμα. Ο πίνακας υποκαταστάσεως για ένα «μικρό» χρονικό διάστημα ε γίνεται:

$$S(\varepsilon)=I+R\varepsilon \overset{(5.20)}{\Rightarrow} S(t+\varepsilon)=S(t)S(\varepsilon)=S(t)(I+R\varepsilon)$$

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

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

$$S'(t)=S(t)R \tag{6.7}$$

Λύνοντας αυτές τις εξισώσεις μπορούμε να πάρουμε τις τιμές για τις πιθανότητες μεταβάσεως. Στην περίπτωση που στη σχέση (6.6) έχουμε $a=\beta=\gamma$ τότε:

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

και προκύπτουν οι λύσεις

$$\begin{aligned} r_t &= \tfrac{1}{4}\left(1+3e^{-4at}\right)\\ s_t &= \tfrac{1}{4}\left(1-e^{-4at}\right)\end{aligned} \tag{6.8}$$

Το μοντέλο αυτό ήταν το πρώτο σχετικό μοντέλο που προτάθηκε από τους Jukes και Cantor (Jukes & Cantor, 1969) και χαρακτηρίζεται ως ένα απλό μοντέλο ανέλιξης Poisson (για συντομία, ονομάζεται JC69). Είναι το πιο απλό ανάμεσα στα σχετικά μοντέλα, αλλά χάνει με αυτόν τον τρόπο κάποια σημαντικά χαρακτηριστικά της εξελικτικής διαδικασίας. Για παράδειγμα δεν αποδίδεται σωστά το γεγονός ότι οι μεταπτώσεις (πουρίνη σε πουρίνη) δεν έχουν τον ίδιο ρυθμό με τις μεταστροφές (πουρίνη σε πυριμιδίνη και αντίστροφα). Η επόμενη παραλλαγή είναι το μοντέλο του διάσημου εξελικτικού βιολόγου Kimura (Kimura, 1980) το οποίο συμβολίζεται ως K2P και προβλέπει:

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

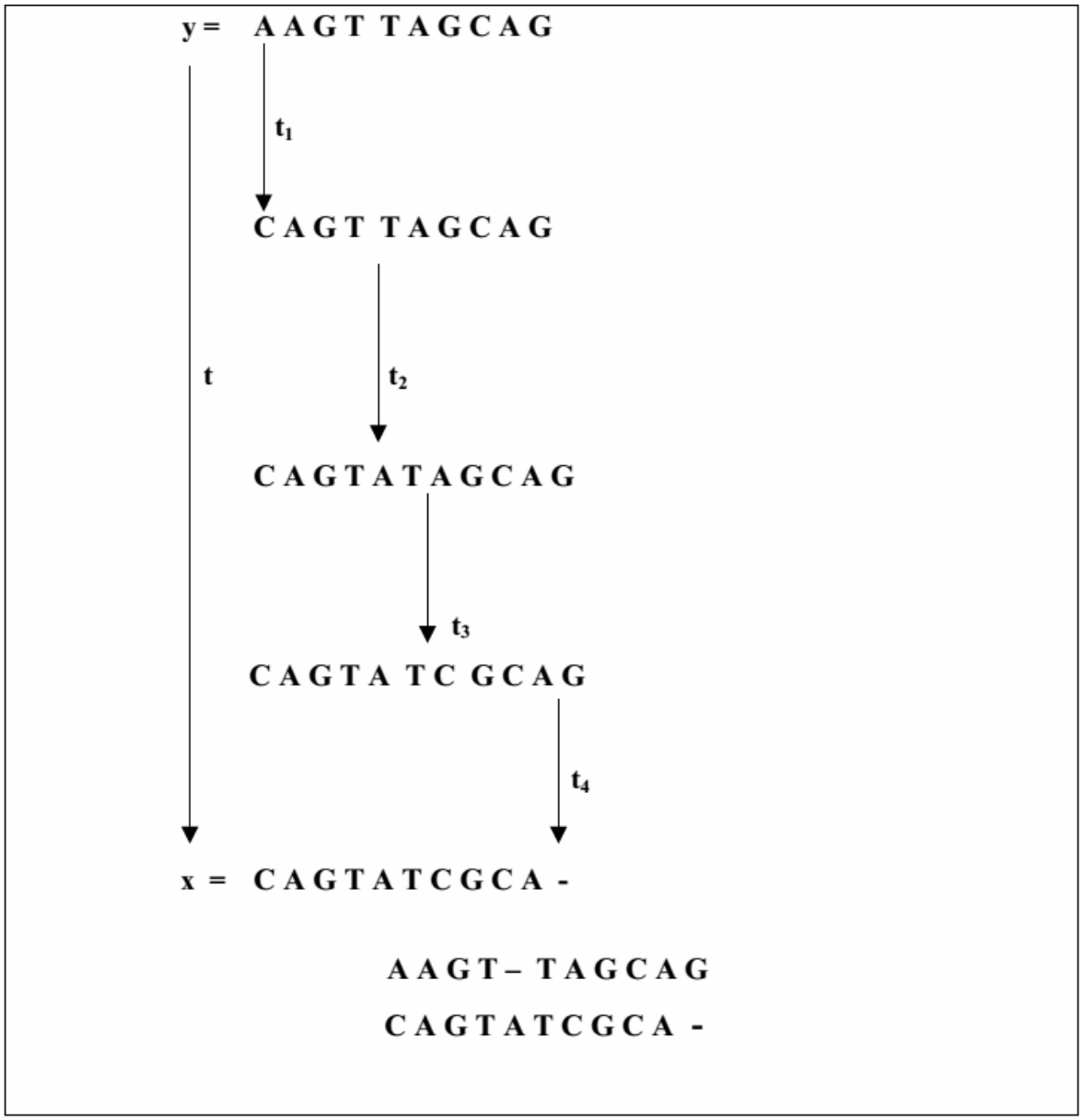
με λύσεις

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

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

και

$$r_t = 1 - 2s_t - u_t \tag{6.9}$$



Εικόνα 6.5: *Σχηματική αναπαράσταση της εξελικτικής διαδικασίας μέσα από την οποία μια αλληλουχία y, μέσα από διαδοχικές μεταλλάξεις (αντικαταστάσεις, απαλοιφές, εισαγωγές), μετά από την πάροδο μεγάλου χρονικού διαστήματος t, οδηγεί σε μια αλληλουχία x.*

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. $A \leftrightarrow T, G \leftrightarrow C$). Πρέπει να υπενθυμίσουμε εδώ ότι βασική ιδιότητα των δυο αυτών μοντέλων (JC69, K2P) είναι ότι καθώς $t \to \infty$ ισχύει $q_A = q_T = q_G = q_C = \frac{1}{4}$ δηλαδή στην κατάσταση ισορροπίας μετά την παρέλευση «άπειρου» χρόνου θα έχουμε μια ισοκατανομή των βάσεων του DNA. Αυτό όμως ξέρουμε ότι δεν είναι τόσο ρεαλιστικό καθώς οι οργανισμοί παρουσιάζουν μεγάλη μεταβλητότητα στο λόγο των βάσεων (A+T)/(G+C) και για να αντιμετωπισθεί αυτό έχουν προταθεί άλλα πιο σύνθετα μοντέλα (Felsenstein, 1981; Lio & Goldman, 1998;

Penny & Hendy, 2001). Σ 'αυτά, ο πίνακας των ρυθμών υποκαταστάσεως δεν έχει την ιδιότητα (6.6) αλλά στηρίζεται σε παρατηρηθησες συχνότητες νουκλεοτιδίων στις υπό μελέτη ακολουθίες (π_A, π_G, π_C, π_T), και κατά συνέπεια επιτρέπει στην κατάσταση ισορροπίας να έχουμε διαφορετικές κατανομές των νουκλεοτιδίων, μια προσέγγιση η οποία είναι πιο ρεαλιστική. Για παράδειγμα, το μοντέλο F81του Fenselstein (Felsenstein, 1981), μοιάζει αρκετά με το κλασικό μοντέλο JC69, στο ότι χρησιμοποιεί μόνο μία παράμετρο (μ) για τις μεταπτώσεις και τις μεταστροφές, αλλά μοντελοποιεί τα τέσσερα νουκλεοτίδια με διαφορετικές πιθανότητες εμφάνισης (π_A, π_G, π_C, π_T):

$$R = \begin{bmatrix} -\mu(\pi_C + \pi_G + \pi_T) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_C + \pi_A + \pi_T) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_C + \pi_G + \pi_A) \end{bmatrix} \quad (6.10)$$

Μια λογική επέκταση αυτού του μοντέλου, είναι εφικτή αν θεωρήσουμε ότι οι μεταπτώσεις έχουν διαφορετικό ρυθμό από τις μεταστροφές, χρησιμοποιώντας δύο παραμέτρους (κ και μ):

$$R = \begin{bmatrix} -\mu(\pi_C + \kappa\pi_G + \pi_T) & \mu\pi_C & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \kappa\pi_T) & \mu\kappa\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\pi_C + \kappa\pi_A + \pi_T) & \mu\kappa\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C + \pi_G + \pi_A) \end{bmatrix} \quad (6.11)$$

Αυτό είναι το μοντέλο των Hasegawa-Kishono-Yano (Hasegawa, Kishino, & Yano, 1985) (HKY85), ενώ παρόμοιο είναι και το μεταγενέστερο μοντέλο του Fenselstein, το λεγόμενο F84, το οποίο μοντελοποιεί το ίδιο φαινόμενο αλλά διαφέρει στην παραμετροποίηση.

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:

$$R = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(d\pi_C + b\pi_A + f\pi_T) & \mu\kappa\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(e\pi_C + f\pi_G + c\pi_A) \end{bmatrix} \quad (6.12)$$

Όπως είναι φανερό, το μοντέλο αυτό είναι το πιο γενικό και μπορεί να συμπεριλάβει όλα τα προηγούμενα ως ειδικές περιπτώσεις. Αντίστοιχα, το HKY85 περιλαμβάνει σαν ειδικές περιπτώσεις τα F81, K2P και JC69, ενώ το F81 και το K2P περιλαμβάνουν το καθένα σαν (διαφορετική) ειδική περίπτωση το JC69. Γενικά, όπως και σε κάθε διαδικασία μοντελοποίησης το πιο σύνθετο μοντέλο είναι και το καλύτερο, αλλά, από την άλλη απαιτεί περισσότερα δεδομένα και αλλά και υπολογιστική ισχύ, καθώς χρειάζεται η εκτίμηση μεγαλύτερου αριθμού παραμέτρων. Πάντως, με τους σύγχρονους υπολογιστές και τις μεθόδους εκτίμησης, ο αριθμός των παραμέτρων και η πολυπλοκότητα του μοντέλου δεν αποτελεί πρακτικό πρόβλημα, γι' αυτό και στις περισσότερες πρακτικές αναλύσεις τα σύγχρονα λογισμικά χρησιμοποιούν πλέον κατά βάση το πιο γενικό μοντέλο, το GTR.

Άλλα μοντέλα ακόμα πιο σύνθετα έχουν προταθεί κατά καιρούς αυτά όμως δεν ασχολούνται με τα ίδια τα νουκλεοτίδια αλλά με τα κωδικόνια επιτρέποντας έτσι απευθείας υπολογισμούς που ανάγονται στο πρωτεϊνικό επίπεδο (Yang, 1994). Έχουν προταθεί και άλλα μοντέλα τα οποία επιτρέπουν διαφορετικούς ρυθμούς αντικατάστασης θεωρώντας ότι αυτοί προέρχονται από έναν πληθυσμό που ακολουθεί την κατανομή Γάμμα (το λεγόμενο random effects model) (Yang, 1993). Πολλές φορές, το μοντέλο της ετερογένειας των ρυθμών εξέλιξης για τις διάφορες θέσεις, μπορεί να συνδυαστεί με κάποιο από τα μοντέλα που παρουσιάσαμε πριν, οπότε μιλάμε για το μοντέλο JC69+Γ, ή GTR+Γ, κ.ο.κ. Η ύπαρξη σταθερών ρυθμών αντικατάστασης είναι γνωστή ως η υπόθεση του «μοριακού ρολογιού» (molecular clock).

Παρόλο που δεν είναι τόσο συνηθισμένο, υπάρχουν και μοντέλα τα οποία δουλεύουν κατευθείαν πάνω σε αμινοξικές αλληλουχίες πρωτεϊνών. Η πιο φυσική επιλογή σε μια τέτοια περίπτωση, είναι οι πίνακες PAM (Dayhoff, Schwartz, & Orcutt, 1978) οι οποίοι έχουν προκύψει από ένα ξεκάθαρο (και μαρκοβιανό) μοντέλο εξελικτικής αλλαγής, ανάλογο με αυτό της σχέσης (6.3). Οι πίνακες αυτής της οικογένειας έχουν ακριβώς τις ίδιες ιδιότητες (χρονική ομογένεια και αντιστρεπτότητα) και σε κατάσταση ισορροπίας οδηγούν

μια κατανομή των αμινοξέων ίδια με αυτή που είχε η βάση δεδομένων από την οποία προήλθαν. Πρέπει να τονιστεί εδώ η ολοφάνερη εξελικτική έννοια που παίρνουν οι πίνακες αντικατάστασης αμινοξέων και νουκλεοτιδίων που συζητήσαμε στις τεχνικές στοίχισης αλληλουχιών στο 3ο κεφαλαίο, καθώς μπορούν να ιδωθούν (Lio & Goldman, 1998) ως πίνακες μεταβάσεως της στοχαστικής ανελίξεως της μετάλλαξης και για την ακρίβεια σαν το όριο τους καθώς $t \to \infty$ (δεχόμενοι δηλαδή ότι έχει περάσει «άπειρος» χρόνος και έχουμε φτάσει σε μια κατάσταση ισορροπίας, όχι όμως με ισοκατανομή των νουκλεοτιδίων ή των αμινοξέων). Μια τελική παρατήρηση αφορά στη χρονική συμμετρία όλων των σχετικών μοντέλων Markov που αναφέραμε, δηλαδή στην αδυναμία τους να διαχωρίσουν ποια από τις ακολουθίες προέκυψε από την άλλη, με άλλα λόγια δεν μπορούν αν χρησιμοποιηθούν σε μια ανάλυση μέγιστης πιθανοφάνειας να παράγουν φυλογενετικό δέντρο με ρίζα. Παρ' όλα αυτά έχουν προταθεί και μοντέλα που δεν χαρακτηρίζονται από χρονική συμμετρία (Lio & Goldman, 1998).

6.3. Μέθοδοι βασισμένες στην απόσταση

Η μία μεγάλη κατηγορία μεθόδων, είναι οι λεγόμενες *μέθοδοι που χρησιμοποιούν τις αποστάσεις* (*distance-based methods*). Οι μέθοδοι αυτές, ξεκινούν από μία πολλαπλή στοίχιση, υπολογίζουν με κάποιον τρόπο έναν πίνακα αποστάσεων για όλα τα ζευγάρια αλληλουχιών και μετά με βάση αυτόν τον πίνακα κατασκευάζουν το φυλογενετικό δέντρο (Durbin, et al., 1998). Η απόσταση (d_{ij}) μεταξύ των αλληλουχιών i,j , έχει συνήθως τις εξής ιδιότητες:

$$\begin{aligned} & d_{ii} = 0 \\ & d_{ij} = d_{ji} > 0,\ i \neq j \\ & d_{ij} \leq d_{ik} + d_{kj} \end{aligned} \tag{6.13}$$

Η μεγαλύτερη ομάδα από τις προαναφερόμενες μεθόδους, είναι στην ουσία κλασικές τεχνικές της *στατιστικής ομαδοποίησης* (*clustering*), και συγκεκριμένα, αυτές οι μέθοδοι που ονομάζονται «ιεραρχικές», οι οποίες παράγουν ένα είδους δέντρο το οποίο δηλώνει την ομαδοποίηση των δεδομένων. Τέτοια παραδείγματα, είδαμε ήδη στο κεφάλαιο 4 με εφαρμογές στην προοδευτική πολλαπλή στοίχιση, ενώ θα τις ξανασυναντήσουμε στο κεφάλαιο 13 (ανάλυση βιολογικών δικτύων) και στο κεφάλαιο 12 (ανάλυση δεδομένων γονιδιακής έκφρασης).

Ο πίνακας των αποστάσεων, d_{ij}, δυο αλληλουχιών i και j, θα μπορούσε γενικά να προκύψει με πολλούς τρόπους. Ένας εύκολος τρόπος θα ήταν μετρώντας απλά το ποσοστό f από τις θέσεις u, στις οποίες τα κατάλοιπα x_u^i και x_u^j, 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:

$$d_{ij} = -\frac{3}{4}\log\left(1 - \frac{4}{3}f\right) \tag{6.14}$$

Για το K2P, αντίστοιχα θα έχουμε:

$$d_{ij} = -\frac{1}{2}\log\left(1 - 2f - g\right) - \frac{1}{4}\log\left(1 - 2g\right) \tag{6.15}$$

όπου f είναι το ποσοστό των αλλαγών που οφείλονται σε μεταπτώσεις και g το ποσοστό των αλλαγών που οφείλονται σε μεταστροφές. Παρόμοιοι υπολογισμοί, μπορούν να γίνουν και για τα υπόλοιπα μοντέλα, μόνο που οι εκφράσεις είναι πιο σύνθετες καθώς περιέχουν διαφορετικές συχνότητες για τις τέσσερις βάσεις.

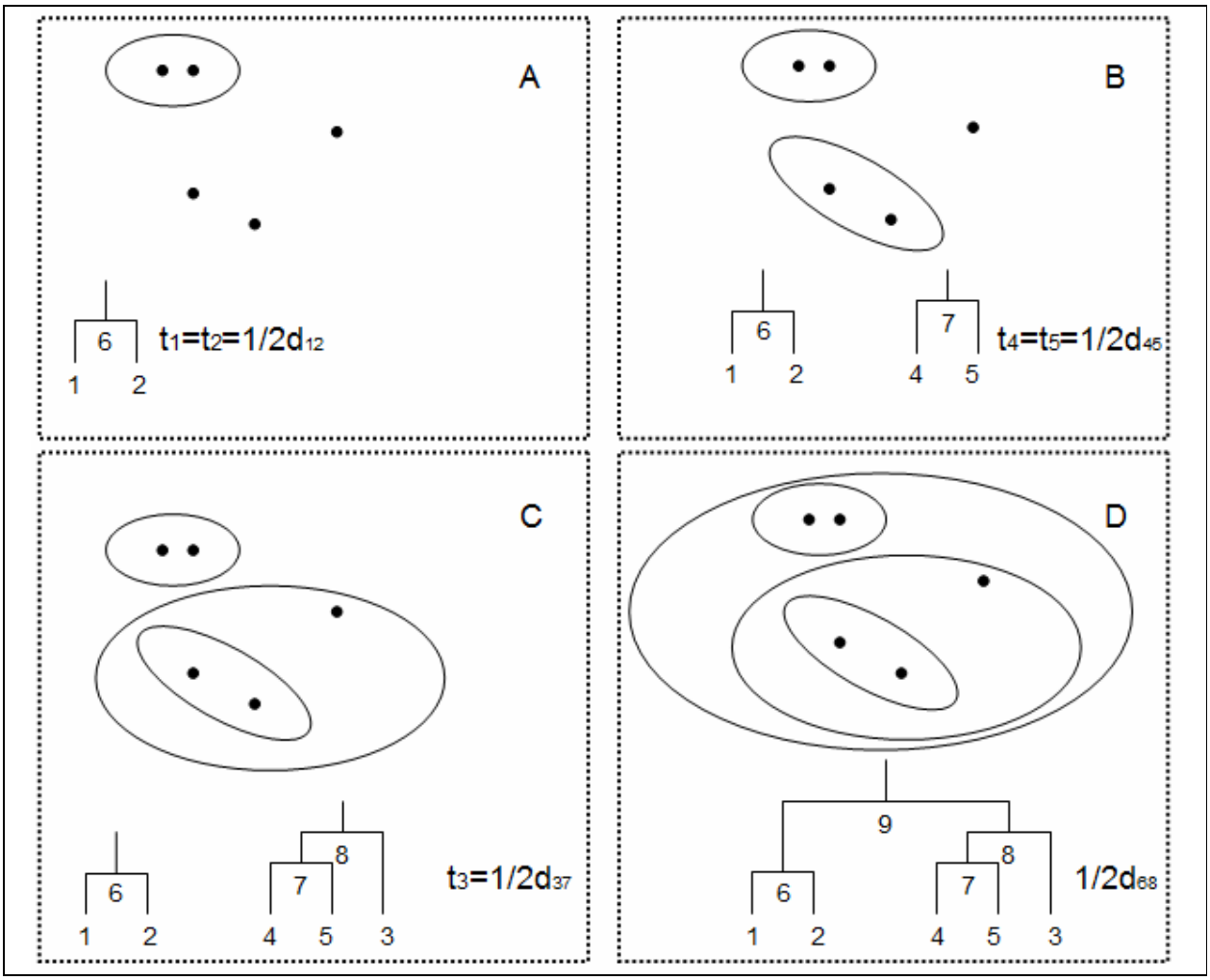
Μόλις οι αποστάσεις υπολογιστούν, η πολλαπλή στοίχιση δεν χρησιμοποιείται ξανά, και όλοι οι υπολογισμοί γίνονται με χρήση του πίνακα. Η πρώτη τέτοια μέθοδος, προτάθηκε από τους Socal και Michener, (Sokal & Michener, 1958) και είναι η γνωστή *UPGMA* (*Unweighted Pair Group using Arithmetic Mean*), η οποία ορίζει την απόσταση μεταξύ δυο ομάδων (Clusters) να είναι η μέση απόσταση μεταξύ ζευγών αλληλουχιών από τις δύο ομάδες:

$$d_{ij} = \frac{1}{|C_i||C_j|}\sum_{p \in C_i, q \in C_j} d_{pq} \tag{6.16}$$

Σε αυτή τη σχέση, $|C_i|$ και $|C_j|$ είναι οι αριθμοί των αλληλουχιών στις ομάδες i και j. Η απόσταση της ομάδας k η οποία αποτελεί την ένωση των ομάδων i και j με μια άλλη ομάδα l θα δίνεται από τη σχέση:

$$d_{kl} = \frac{d_{il}\left|C_i\right| + d_{jl}\left|C_j\right|}{\left|C_i\right| + \left|C_j\right|} \tag{6.17}$$

Ο αλγόριθμος, δουλεύει όπως και οι αλγόριθμοι ιεραρχικής ομαδοποίησης που περιγράψαμε στην πολλαπλή στοίχιση: στην αρχή ξεκινάει τοποθετώντας κάθε μια ακολουθία στη δική της ομάδα, και προχωράει ιεραρχικά, τοποθετώντας στην ίδια ομάδα τις ακολουθίες με τη μικρότερη απόσταση. Ο κόμβος σε κάθε βραχίονα τοποθετείται στο ύψος $d_{ij}/2$ (Εικόνα 6.6). Ο αλγόριθμος ανακατασκευάζει το δέντρο σε χρόνο της τάξης του $O(n^2)$. Η μέθοδος αυτή είναι απλή, διαισθητικά σωστή, εύκολα ερμηνεύσιμη και παράγει φυλογενετικά δέντρα με ρίζα, παρ' όλα αυτά πολλές φορές μπορεί να δώσει λάθος αποτελέσματα, όταν δεν ικανοποιούνται κάποιες προϋποθέσεις, η βασικότερη από τις οποίες είναι ο σταθερός ρυθμός εξελικτικής διαδικασίας σε όλες τις αλληλουχίες, δηλαδή, το «μοριακό ρολόι». Αξίζει να αναφερθεί, ότι στη στατιστική ορολογία, η μέθοδος ονομάζεται «average linkage». Μέθοδοι που βασίζονται και στις υπόλοιπες κατηγορίες linkage (complete linkage, simple linkage κλπ), έχουν επίσης προταθεί για φυλογενετική ανάλυση, αλλά εμπειρικές αναλύσεις έδειξαν ότι δεν αποδίδουν τόσο καλά όσο η μέθοδος UPGMA.



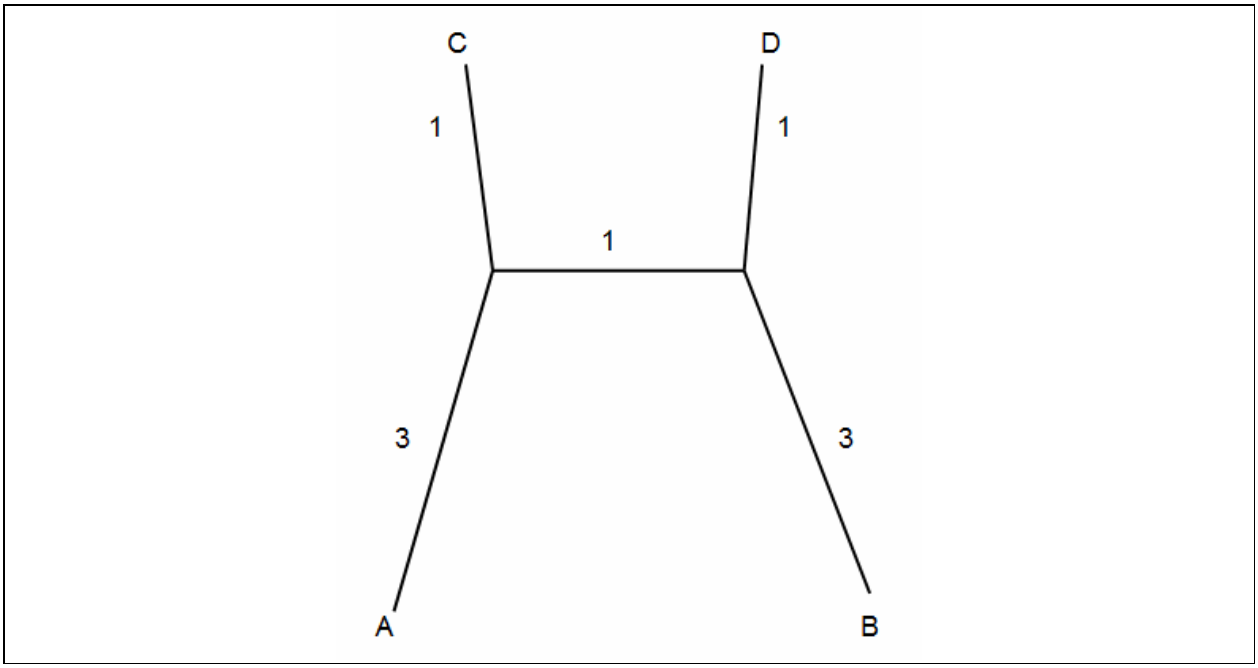
Εικόνα 6.6: *Υποθετικό παράδειγμα της μεθόδου UPGMA.Καταχρηστικά, οι πέντε αλληλουχίες αναπαρίστανται σαν να αντιστοιχούν σε σημεία στο επίπεδο (δεν ισχύει πάντα λόγω των διαφορετικών τρόπων ορισμού της απόστασης). Ο αλγόριθμος προχωράει σε βήματα (A,B,C,D) κατά τα οποία ανακατασκευάζει το δέντρο, ομαδοποιώντας σταδιακά τις πιο όμοιες αλληλουχίες.*

Η μέθοδος UPGMA, εκτός από την υπόθεση του μοριακού ρολογιού, έχει και μια άλλη ιδιότητα, παράγει εξ'ορισμού δέντρα στα οποία ισχύει η προσθετική ιδιότητα. Με αυτό, εννοούμε δέντρα στα οποία η απόσταση δύο οποιονδήποτε άκρων, είναι ίση με το άθροισμα των μηκών των ακμών που τα συνδέουν.

Παρόλα αυτά, είναι δυνατόν να υπάρχουν περιπτώσεις στις οποίες δεν ισχύει η υπόθεση του μοριακού ρολογιού, αλλά η προσθετική ιδιότητα να εξακολουθεί να ισχύει. Σε αυτή την περίπτωση, πρέπει να αναζητηθεί κάποιος άλλος κατάλληλος αλγόριθμος. Η βασική ιδέα, είναι να έχουμε ένα δέντρο με αθροιστικό μήκως κλαδιών. Τότε, για δύο γειτονικά κλαδιά *i*, *j* τα οποία έχουν κοινό κόμβο τον *k*, θα πρέπει να αφαιρέσουμε τα κλαδιά αυτά από το δέντρο, να προσθέσουμε το *k* σαν κλαδί του δέντρου και να ορίσουμε την απόστασή του από ένα άλλο κλαδί, *m*, ως:

$$d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right) \tag{6.18}$$

Επειδή ισχύει η προσθετική ιδιότητα, οι αρχικές αποστάσεις του δέντρου, διατηρούνται. Άρα, μπορούμε να προχωρήσουμε βήμα-βήμα και να αφαιρούμε κάθε φορά από ένα κλαδί του δέντρου, μέχρι να ομαδοποιήσουμε όλες τις παρατηρήσεις. Το βασικό πρόβλημα, είναι να μπορέσουμε να εντοπίσουμε τα γειτονικά κλαδιά, χρησιμοποιώντας τις αποστάσεις και μόνο. Αυτό δεν είναι πάντα απλό, όπως φαίνεται στην Εικόνα 6.8.



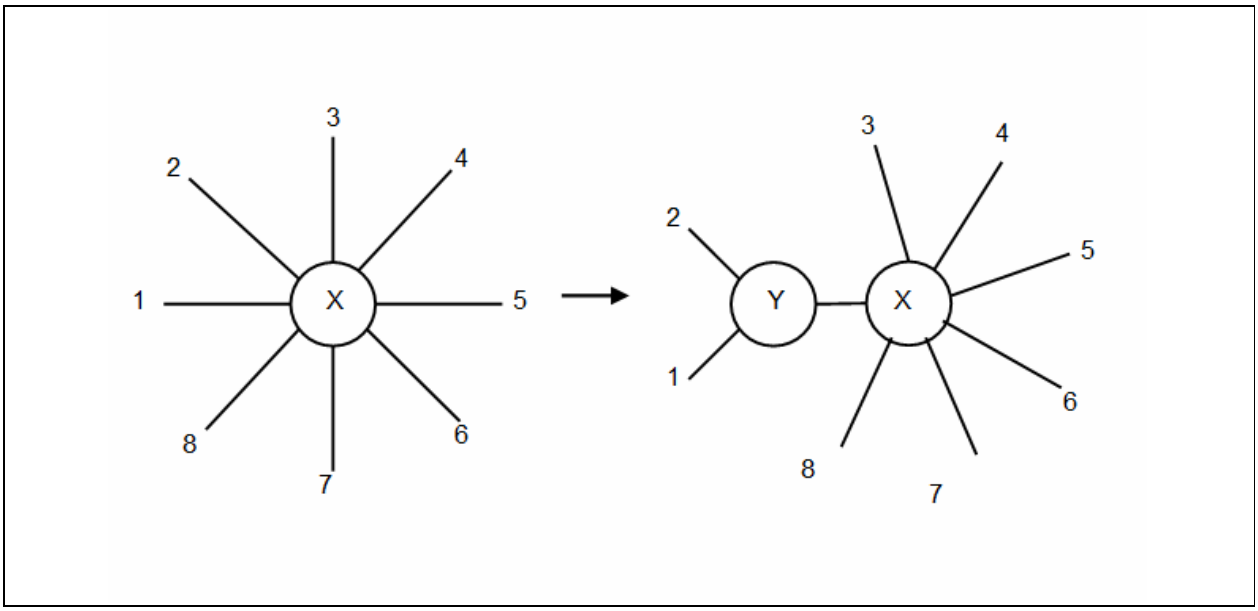
Εικόνα 6.7: *Ένα παράδειγμα υποθετικού δέντρου στο οποίο φαίνεται ότι δύο γειτονικοί βραχίονες, είναι δυνατόν να μην είναι οι πιο κοντινοί. Οι αποστάσεις των γειτόνων (A-C και B-D) είναι ίσες με 4, αλλά οι αποστάσεις των μη γειτόνων (A-B και C-D) είναι μικρότερες (ίσες με 3). Το πρόβλημα αυτό, το λύνει η μετασχηματισμένη απόσταση που χρησιμοποιεί η μέθοδος Neighbour-Joining.*

Η μέθοδος της *ένωσης γειτόνων* (*neighbour joining*, NJ) (Saitou & Nei, 1987), είναι ίσως μια από τις πιο συζητημένες και ευρέως χρησιμοποιούμενες μεθόδους αποστάσεων, η οποία ορίζει μια μετασχηματισμένη απόσταση:

$$D_{ij} = d_{ij} - \frac{1}{L-2}\sum_{\forall k}\left(d_{ik} + d_{jk}\right) \tag{6.19}$$

όπου *L* ο αριθμός των κλαδιών του δέντρου, και κατατάσσει τις αλληλουχίες σε ζευγάρια για να βρει τελικά τις πιο «γειτονικές». Με αυτόν τον τρόπο, λύνει το πρόβλημα των αποστάσεων που αναφέραμε στην Εικόνα 6.8 και εντοπίζει τα γειτονικά κλαδιά του δέντρου (δηλαδή, αυτά στα οποία το D_{ij} είναι ελάχιστο). Στο επόμενο βήμα, η μέθοδος θα ενώσει τα δύο επόμενα κλαδιά, κ.ο.κ. Ο αλγόριθμος αυτός, για ένα δέντρο με *L* κλαδιά, απαιτεί *L*-3 επαναλήψεις και σε κάθε μία από αυτές, υπολογίζεται ο πίνακας D_{ij} ο οποίος έχει διαστάσεις *L*x*L*. Κατά συνέπεια, ο αλγόριθμος έχει πολυπλοκότητα της τάξης $O(L^3)$, αν και υπάρχουν τροποποιήσεις που επιτυγχάνουν καλύτερους χρόνους.

Η μέθοδος NJ, όπως είπαμε, έχει το σημαντικό πλεονέκτημα ότι δεν κάνει την απαίτηση του μοριακού ρολογιού. Επίσης, αν οι αποστάσεις ακολουθούν ή προσεγγίζουν την προσθετική ιδιότητα, τότε το δέντρο που ανακατασκευάζεται θα είναι το σωστό. Τέλος, είναι πολύ γρήγορη, και αυτό την κάνει ελκυστική, ειδικά για αναλύσεις μεγάλων συνόλων δεδομένων ή για εφαρμογή στατιστικών τεχνικών όπως το bootsrap. Σε σύγκριση με την UPGMA, είναι πιο αργή, αλλά αυτό αντισταθμίζεται από την χαλάρωση της απαίτησης του μοριακού ρολογιού. Από την άλλη, παράγει φυλογενετικά δέντρα χωρίς ρίζα και για την εύρεση αυτής μπορούμε να χρησιμοποιήσουμε σαν σημείο αναφοράς ένα «μακρινό» είδος (outgroup) το οποίο ξέρουμε ότι βρίσκεται πολύ μακριά εξελικτικά από όσα εξετάσαμε. Η μέθοδος, προτάθηκε ειδικά για φυλογενετικές αναλύσεις, αλλά είναι στην ουσία μια μέθοδος ομαδοποίησης, η οποία μπορεί να βρει εφαρμογή και σε άλλα πεδία, φτάνει να οριστεί κατάλληλα η απόσταση. Ένα τέτοιο παράδειγμα, είδαμε στην προοδευτική πολλαπλή στοίχιση, καθώς το γνωστό πρόγραμμα CLUSTAL, χρησιμοποιεί αυτόν τον αλγόριθμο για να κατασκευάσει το δέντρο-οδηγό.



Εικόνα 6.8: *Ένα υποθετικό παράδειγμα της λειτουργίας της μεθόδου Neighbour-Joining. Οι αλληλουχίες 1 και 2 έχουν τη μικρότερη (μετασχηματισμένη) απόσταση, και κατά συνέπεια ενώνονται για να δώσουν ένα νέο κόμβο (Υ). Στη συνέχεια, ο αλγόριθμος θα προχωρήσει ενώνοντας διαδοχικά κάθε φορά τους βραχίονες που οδηγούν στο επόμενο πιο κοντινό ζευγάρι γειτόνων.*

Τέλος, πρέπει να αναφέρουμε και μια άλλη μέθοδο που χρησιμοποιεί αποστάσεις. Αυτή, είναι η μέθοδος των *Fitch-Margoliash* (Fitch & Margoliash, 1967), η οποία βασίζεται στη στατιστική τεχνική της γραμμικής παλινδρόμησης, δηλαδή, της ευθείας ελαχίστων τετραγώνων. Η βασική ιδέα της μεθόδου, είναι να ελαχιστοποιήσει το άθροισμα των τετραγώνων των αποκλίσεων που έχουν οι παρατηρηθείσες αποστάσεις σε ένα δέντρο (d_{ij}), από τις θεωρητικές ($\hat{d}_{ij}$). Συνεπώς, η ποσότητα που ελαχιστοποιείται, θα δίνεται από τον τύπο:

$$Q = \sum_{i=1}^{L}\sum_{j=1}^{L} w_{ij}\left(\hat{d}_{ij} - d_{ij}\right)^2 \qquad (6.20)$$

Αυτός είναι ακριβώς το κριτήριο που ελαχιστοποιείται και στην κλασική ευθεία ελαχίστων τετραγώνων ($y=a+bx$). Τα βάρη w_{ij} παίζουν εδώ ακριβώς τον ίδιο ρόλο που παίζουν και στην γραμμική παλινδρόμηση. Η πρώτη εμφάνιση της μεθόδου, όπως προτάθηκε από τους (Cavalli-Sforza & Edwards, 1967) είχε βάρη ίσα με $w_{ij}=1$ (απλή γραμμική παλινδρόμηση), ενώ η πιο προχωρημένη μέθοδος των (Fitch & Margoliash, 1967), χρησιμοποίησε βάρη αντίστροφα του τετραγώνου της απόστασης ($wi_j = 1/\hat{d}_{ij}^2$, σταθμισμένη γραμμική παλινδρόμηση). Σε κάθε περίπτωση, με τη μέθοδο αυτή αναζητούμε το δέντρο, τα μήκη των βραχιόνων του οποίου θα έχουν τη μικρότερη τετραγωνική απόκλιση των αποστάσεων, σε σχέση με όλα τα πιθανά μήκη μονοπατιών. Η μέθοδος είναι απλή, κατανοητή και βασίζεται σε ένα στέρεο στατιστικό υπόβαθρο, αλλά έχει

το βασικό μειονέκτημα ότι δεν έχει μηχανισμό για να εντοπίζει τα δέντρα. Αντίθετα, πρέπει η τετραγωνική απόκλιση (*Q*) να μετρηθεί για κάθε πιθανό δέντρο. Για αυτό το σκοπό, έχουν αναπτυχθεί μια σειρά από αλγόριθμοι οι οποίοι υπολογίζουν σε εύλογο χρονικό διάστημα τα πιθανά δέντρα τα οποία θα ελέγξει η μέθοδος. Παρόλα αυτά, σήμερα δεν χρησιμοποιείται πολύ, γιατί δεν εμφανίζει πολλά πλεονεκτήματα και έχει ξεπεραστεί από την πιο σύγχρονη και πιο ευέλικτη μέθοδο της μέγιστης πιθανοφάνειας.

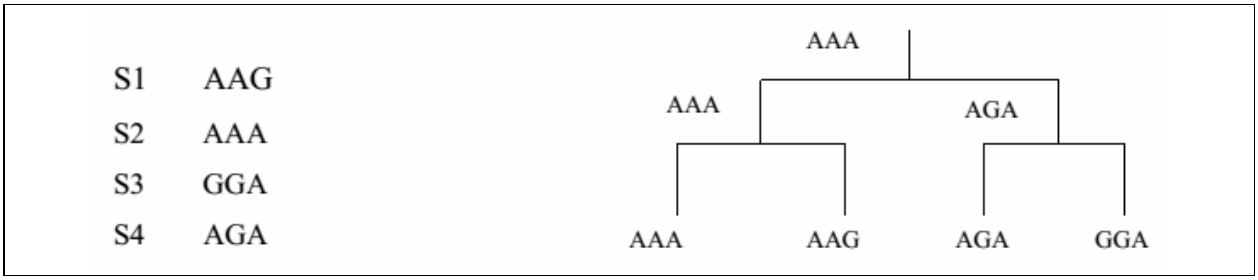
6.4. Μέθοδοι βασισμένες στους χαρακτήρες

Οι *μέθοδοι που βασίζονται στους χαρακτήρες* (*character-based methods*), σε αντίθεση με τις μεθόδους αποστάσεων, δεν μετασχηματίζουν τις αλληλουχίες, αλλά τις χρησιμοποιούν σε όλη τη διαδικασία της εκτίμησης, αντιμετωπίζοντας αυτές, όπως ακριβώς είναι: μια ακολουθία διακριτών συμβόλων από ένα πεπερασμένο αλφάβητο, Ω, (Durbin, et al., 1998). Διακρίνονται σε δύο μεγάλες ομάδες: στις μεθόδους φειδωλότητας, οι οποίες δεν κάνουν καμιά υπόθεση για τον τρόπο που συντελέστηκε η εξελικτική διαδικασία, και στις μεθόδους μέγιστης πιθανοφάνειας, οι οποίες χρησιμοποιούν (ή καλύτερα, απαιτούν) ένα ξεκάθαρο μαθηματικό μοντέλο για την εξέλιξη. Αυτές οι δύο μέθοδοι θα παρουσιαστούν στις επόμενες υπο-ενότητες.

6.4.1 Μέθοδος Φειδωλότητας

Οι μέθοδοι που στηρίζονται στη *μέγιστη φειδωλότητα* (*maximum parsimony*) διαφέρουν ριζικά από τις προηγούμενες μεθόδους των αποστάσεων, στο ότι κάνουν διάκριση μεταξύ πληροφοριακών και μη-πληροφοριακών θέσεων στις αλληλουχίες, με τις πληροφοριακές θέσεις να είναι αυτές που παρουσιάζουν πολυμορφισμό (ύπαρξη πάνω από δυο ειδών νουκλεοτιδίων) τουλάχιστον δυο φορές. Η μέθοδος αυτή εφαρμόζεται στην εξελικτική βιολογία προτού να εμφανιστεί η μοριακή φυλογένεση (εφαρμοζόταν για παράδειγμα σε διάφορα φαινοτυπικά χαρακτηριστικά) και έχει σκοπό να εξηγήσει τις εξελικτικές διάφορες με το μικρότερο δυνατό αριθμό αλλαγών. Είναι δηλαδή, κατά μία έννοια, το φυλογενετικό ανάλογο της φιλοσοφικής μεθόδου του «ξυραφιού του Οκάμ» (Okham Razor), η οποία με απλά λόγια δηλώνει ότι η απλούστερη εξήγηση είναι και η προτιμότερη. Συνήθως η έκφραση αποδίδεται στα Λατινικά ως «*Pluralitas non est ponenda sine necessitate*», το οποία σε ελεύθερη απόδοση σημαίνει «*Όταν δύο θεωρίες παρέχουν εξίσου ακριβείς προβλέψεις, πάντα επιλέγουμε την απλούστερη*».

Ένα σημαντικό χαρακτηριστικό της μεθόδου, είναι ότι απλά αποδίδει κόστος σε μια δεδομένη τοπολογία ενός δέντρου, οπότε πρέπει να έχουμε στο μυαλό μας ότι απαιτείται ειδικός αλγόριθμος για να εντοπίσει το δέντρο του οποίου το κόστος θα υπολογίσουμε. Η πλέον χρησιμοποιούμενη μέθοδος είναι αυτή που χρησιμοποιεί τον αλγόριθμο του Fitch (Fitch, 1971), στην οποία κάθε διαφορά σε μία θέση «σκοράρει», δηλαδή αποδίδει κόστος ίσο με +1 σε όλες τις αλλαγές, αλλά έχουν προταθεί και παραλλαγές οι οποίες σταθμίζουν με διαφορετικό τρόπο τις διαφορές (weighted parsimony), οπότε ο σκοπός της μεθόδου είναι να ελαχιστοποιηθεί το κόστος αυτό. Στην Εικόνα 6.9, απεικονίζονται 4 αλληλουχίες οι οποίες είναι ήδη στοιχισμένες, και θέλουμε να βρούμε ένα φυλογενετικό δέντρο με τη χρήση της μεθόδου της φειδωλότητας. Το δέντρο που δίνεται είναι αυτό το οποίο εξηγεί τις νουκλεοτιδικές αλλαγές με τον μικρότερο αριθμό αντικαταστάσεων (3 συνολικά) από όλα τα άλλα δέντρα με ρίζα (συνολικά υπάρχουν 15 τέτοια δέντρα).



Εικόνα 6.9: *Ένα παράδειγμα δέντρου που εκτιμήηκε με τη μέθοδο της φειδωλότητας. Το δέντρο που δίνεται είναι αυτό το οποίο εξηγεί τις διαφορές των αλληλουχιών με τον μικρότερο αριθμό αντικαταστάσεων (3 συνολικά) από όλα τα άλλα πιθανά δέντρα με ρίζα (συνολικά υπάρχουν 15 τέτοια δέντρα).*

Γενικά, παρόλο που η μέθοδος είναι ιδιαίτερα γρήγορη, η αναζήτηση ανάμεσα σε όλα τα πιθανά δέντρα, γίνεται απαγορευτική όταν οι υπό σύγκριση αλληλουχίες είναι υπερβολικά πολλές. Έχουν προταθεί

για αυτό το σκοπό διάφοροι αλγόριθμοι, εκ των οποίων ο λεγόμενος *branch and bound*, είναι αυτός που δίνει εγγυήσει ότι θα βρει το καλύτερο δέντρο χωρίς να ανατρέξει σε όλες τις πιθανές τοπολογίες. Η βασική του ιδέα είναι να ξεκινάει από τυχαία δέντρα και να προσθέτει βραχίονες στην τύχη, και εκμεταλλεύεται το γεγονός (το οποίο είναι χαρακτηριστικό της μεθόδου της φειδωλότητας), ότι οι αλλαγές στο δέντρο μπορούν να συμβούν μόνο όταν προστεθεί ένας βραχίονας. Αν τώρα, ένας δεδομένος βραχίονας αυξήσει το συνολικό ελάχιστο αριθμό αντικαταστάσεων που έχει παρατηρηθεί μέχρι εκείνη τη στιγμή, τότε η αναζήτηση σε αυτή την κατεύθυνση εγκαταλείπεται και ο βραχίονας διαγράφεται (όπως επίσης και όλες οι πιθανές αλλαγές από εκείνο το σημείο και κάτω).

Τα βασικά πλεονεκτήματα της μεθόδου είναι αφενός μεν η ταχύτητά της, που την καθιστά ικανή για αναλύσεις πολλών αλληλουχιών, αφετέρου δε η απλότητα της, καθώς δεν προϋποθέτει κανένα μοντέλο για την εξέλιξη των αλληλουχιών. Πρέπει να τονιστεί βέβαια, ότι η μέθοδος της φειδωλότητας είναι αντικείμενο πολλών αντιπαραθέσεων στην εξελικτική βιολογία, καθώς πολλοί ερευνητές δε δέχονται ότι διαθέτει στατιστική τεκμηρίωση ενώ μερικοί αμφισβητούν ακόμα και τη σχέση της με την μόλις παραπάνω αναφερθείσα φιλοσοφική μέθοδο της φειδωλότητας (Yang, 1996). Αξίζει να σημειωθεί τέλος ότι η φειδωλότητα προτάθηκε αρχικά (Edwards & Cavalli-Sforza, 1963) ως μέθοδος υπολογιστικής προσέγγισης στη μέγιστη πιθανοφάνεια την οποία θα αναλύσουμε στην επόμενη ενότητα.

6.4.2 Η Μέθοδος της Μέγιστης Πιθανοφάνειας

Η πιο προφανής από άποψη στατιστικής, μέθοδος εκτίμησης που θα μπορούσε να χρησιμοποιηθεί είναι αυτή της *Μέγιστης Πιθανοφάνειας* (*Maximum Likelihood*). Σε γενικές γραμμές η μέθοδος αυτή αντιμετωπίζει τις ακολουθίες ως ένα σετ *L* μεταβλητών με *n* παρατηρήσεις η κάθε μια. Έτσι στις ακολουθίες:

$$\mathbf{x_1} = x_{11}x_{12}...x_{1n}$$

$$\mathbf{x_2} = x_{21}x_{22}...x_{2n}$$

.....

$$\mathbf{x_L} = x_{L1}x_{L2}...x_{Ln}$$

τα αντίστοιχα διανύσματα των παρατηρήσεων τα οποία αντιστοιχούν στις θέσεις της πολλαπλής στοίχισης, θα είναι:

$$X_1 = (x_{11}, x_{21}, ..., x_{L1}), X_2 = (x_{12}, x_{22}, ..., x_{L2}), ..., X_L = (x_{1n}, x_{2n}, ..., x_{Ln})$$

Πρέπει εδώ να τονιστούν κάποιες θεμελιώδεις διαφορές ανάμεσα στις γνωστές μεθόδους Μέγιστης Πιθανοφάνειας (Maximum Likelihood) που χρησιμοποιούνται για την εκτίμηση π.χ. παραμέτρων σε ένα Γενικευμένο Γραμμικό Μοντέλο και στην εκδοχή της μεθόδου που χρησιμοποιείται για την εκτίμηση των φυλογενετικών σχέσεων.

- Για να προχωρήσουμε στην ανάλυση, είναι αναγκαίο να έχει γίνει πρώτα μια πολλαπλή στοίχιση των αλληλουχιών (και άρα το αποτέλεσμα μας θα είναι δεσμευμένο στη στοίχιση αυτή). Παρ' όλα αυτά έχουν προταθεί και κάποιες μέθοδοι ταυτόχρονης στοίχισης και φυλογενετικής ανάλυσης.
- Για να εφαρμοστούν οι μέθοδοι αυτές πρέπει να ορίσουμε εξ' αρχής ένα πιθανοθεωρητικό μοντέλο το οποίο και να περιγράφει την εξέλιξη των αλληλουχιών (π.χ μοντέλο JC69 ή κάποιο άλλο από αυτά που μελετήσαμε στην αντίστοιχη παράγραφο).
- Η πιθανοφάνεια κάθε φορά υπολογίζεται ως συνάρτηση της τοπολογίας του δέντρου και του μήκους των βραχιόνων του.

Αποτέλεσμα των παραπάνω είναι το γεγονός ότι η συνολική πιθανοφάνεια δεν μπορεί να υπολογιστεί αναλυτικά με κάποιον απλό τρόπο, αλλά απαιτεί υπολογισμούς που ανάγονται στο άθροισμα όλων των πιθανοφανειών για όλα τα πιθανά δέντρα. Στην περίπτωση που έχουμε δύο ακολουθίες:

$$\mathbf{x_1} = x_{11}, x_{12}, ..., x_{1n}$$

$$\mathbf{x_2} = x_{21}, x_{22}, ..., x_{2n}$$

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

$$P\left(x_{1i}, x_{2i}, a \mid T, t_1, t_2\right) = q_\alpha P\left(x_{i1} \mid a, t_1\right) P\left(x_{i2} \mid a, t_2\right) \tag{6.21}$$

όπου α το άγνωστο αρχικό νουκλεοτίδιο, T το υποτιθέμενο δέντρο, και t_1,t_2 τα μήκη των βραχιόνων του δέντρου (χρόνος κατά τον οποίο έχουν αποκλίνει εξελικτικά). Επειδή δεν γνωρίζουμε το α πρέπει να αθροίσουμε όλες τις εναλλακτικές άρα:

$$P\left(x_{1i}, x_{2i} \mid T, t_1, t_2\right) = \sum_a q_\alpha P\left(x_{i1} \mid a, t_1\right) P\left(x_{i2} \mid a, t_2\right) \tag{6.22}$$

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

$$P\left(\mathbf{x}_1, \mathbf{x}_2 \mid T, t_1, t_2\right) = \prod_{i=1}^{n} P\left(x_{1i}, x_{2i} \mid T, t_1, t_2\right) \tag{6.23}$$

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

$$\log P\left(\mathbf{x}_1, \mathbf{x}_2 \mid T, t_1, t_2\right) = \sum_{i=1}^{n} \log P\left(x_{1i}, x_{2i} \mid T, t_1, t_2\right) \tag{6.24}$$

Στο δεξί μέρος της σχέσης (6.24), θα πρέπει να χρησιμοποιηθούν οι αντίστοιχες εκφράσεις που απορρέουν από το εκάστοτε μοντέλο της εξέλιξης, όπως για παράδειγμα το JC69 ή το GTR, για να καταλήξουμε τελικά σε μια αναλυτική έκφραση για τη συνάρτηση πιθανοφάνειας. Στη γενικότερη περίπτωση που χρησιμοποιούμε L ακολουθίες, θα έχουμε:

$$P\left(x_{1i}, x_{2i}, ..., x_{Li} \mid T, t^\bullet\right) = \sum_{a^{L+1}, ..., a^{2L-1}} q_{a^{2L-1}} \prod_{k=L+1}^{2L-2} P\left(a^k \mid a^{a(k)}, t_k\right) \prod_{k=1}^{L} P\left(x_{ki} \mid a^{a(k)}, t_k\right) \tag{6.25}$$

Οι επιπλέον συμβολισμοί που εισάγουμε εδώ είναι το $\alpha(k)$ που συμβολίζει το αρχικό νουκλεοτίδιο για κάθε ένα από τα παρακλάδια του δέντρου. Από την συνδυαστική βρίσκουμε ότι για L ακολουθίες έχουμε $2L$-1 παρακλάδια και $2L$-2 σημεία διασταύρωσης των κλαδιών, όταν το δέντρο έχει ρίζα. Από αυτά τα πρώτα L αντιστοιχούν στα παρακλάδια (βραχίονες) που οδηγούν σε μια ακολουθία ενώ τα υπόλοιπα, από L+1εως $2L$-2, αντιστοιχούν στους κλάδους 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 L+1 έως το $2L$-1 (είναι ένα παραπάνω γιατί πρέπει να μετρήσουμε και το α στη ρίζα του δέντρου). Τελικά η συνολική πιθανοφάνεια για τις r ακολουθίες θα προκύψει αφού αθροίσουμε τη συνεισφορά όλων των n θέσεων:

$$P\left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L \mid T, t_\circ\right) = \prod_{i=1}^{n} P\left(x_{1i}, x_{2i}, ..., x_{Li} \mid T, t_\circ\right) \tag{6.26}$$

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

$$\log P\left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_L \mid T, t_\circ\right) = \sum_{i=1}^{n} \log P\left(x_{1i}, x_{2i}, ..., x_{Li} \mid T, t_\circ\right) \tag{6.27}$$

Το παραπάνω μοντέλο, θεωρεί όλες τις θέσεις ανεξάρτητες και υποθέτει ότι ο ρυθμός της εξέλιξης είναι σταθερός για όλα τα παρακλάδια του δέντρου. Όπως είδαμε, έχουν προταθεί και άλλα μοντέλα τα οποία επιτρέπουν διαφορετικούς ρυθμούς αντικατάστασης (Yang, 1993). Η ύπαρξη σταθερών ρυθμών αντικατάστασης είναι γνωστή ως η υπόθεση του «*μοριακού ρολογιού*» (*molecular clock*) και είναι απαραίτητη προϋπόθεση και για την εφαρμογή της μεθόδου UPGMA, την οποία είδαμε σε προηγούμενη παράγραφο. Μια άλλη παρατήρηση που αφορά τη μέγιστη πιθανοφάνεια, η οποία δεν είναι αμέσως εμφανής, είναι ότι η μέθοδος δεν είναι ικανή να δώσει τη θέση της ρίζας του δέντρου, καθώς τα περισσότερα από τα στοχαστικά μοντέλα της εξέλιξης, έχουν την ιδιότητα να είναι χρονικώς αντιστρεπτά (time reversibility). Κατά συνέπεια, επειδή ο πίνακας των αντικαταστάσεων δεν μπορεί να διακρίνει μια αλλαγή A σε T, από μια αλλαγή T σε A, η πιθανοφάνεια του δέντρου είναι ίδια και για τις δύο εναλλακτικές υποθέσεις, οπότε τελικά, η πιθανοφάνεια δεν θα εξαρτάται από τη θέση της ρίζας.

Όπως ήδη είπαμε, η συνολική πιθανοφάνεια δεν μπορεί να υπολογιστεί αναλυτικά, αλλά πρέπει να αθροιστούν οι συνεισφορές όλων των πιθανών δέντρων. Ακόμα και όταν αναφερόμαστε σε ένα δεδομένο δέντρο (ανάμεσα στα πολλά πιθανά), η αναλυτική έκφραση για τη σχέση (6.27) είναι ιδιαίτερα πολύπλοκη, και απαιτεί για την επίλυση της κάποιου είδους επαναληπτική διαδικασία, κλασική σε προβλήματα μέγιστης

πιθανοφάνειας, όπως για παράδειγμα κάποια παραλλαγή της μεθόδου Gradient Descent, Newton-Raphson (Ypma, 1995), ή του αλγορίθμου EM (Dempster, Laird, & Rubin, 1977). Κατά συνέπεια, η μέθοδος έχει το ίδιο πρόβλημα όπως και η μέγιστη φειδωλότητα: χρειάζεται κάποιος τρόπος για να υπολογιστούν γρήγορα όλα τα πιθανά δέντρα, ή τουλάχιστον, τα πιο πιθανά. Ο υπολογισμός της πιθανοφάνειας γίνεται με κάποιον αλγόριθμο ο οποίος αθροίζει τις συνεισφορές για όλα τα πιθανά δέντρα, και ο πιο γνωστός αλγόριθμος που έχει προταθεί για αυτό το σκοπό, είναι αυτός του Felsenstein (Felsenstein, 1981). Αν και έχουν προταθεί πολλές παραλλαγές, σε γενικές γραμμές η διαδικασία που ακολουθείται για να υπολογιστεί η πιθανοφάνεια είναι η εξής: Ο αλγόριθμος εντοπίζει ένα πιθανό δέντρο και οι παράμετροι για αυτό το δέντρο αλλάζουν λίγο-λίγο με κάποια από τις επαναληπτικές διαδικασίες που αναφέραμε παραπάνω έως ότου να βρεθεί το βέλτιστο δέντρο. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα πιθανά δέντρα και αυτό που δίνει τη μέγιστη πιθανοφάνεια, από όλα τα δέντρα, επιλέγεται τελικά ως το δέντρο μέγιστης πιθανοφάνειας.

Η μέθοδος της μέγιστης πιθανοφάνειας έχει μερικά ξεκάθαρα πλεονεκτήματα σε σχέση με τις άλλες μεθόδους (Yang & Rannala, 2012). Καταρχάς, όλες οι προϋποθέσεις των μοντέλων δηλώνονται ξεκάθαρα και μπορούν να αξιολογηθούν εκ των υστέρων. Διαθέτει επίσης μια πληθώρα πιθανοθεωρητικών μοντέλων για την εξέλιξη των αλληλουχιών, τα οποία μπορούν να υλοποιηθούν, να εφαρμοστούν και να ελεγχθούν. Το στέρεο μαθηματικό της υπόβαθρο, το οποίο χρησιμοποιείται και σε πολλές άλλες εφαρμογές στη βιοπληροφορική, επιτρέπει τη χρήση εργαλείων όπως ο έλεγχος του πηλίκου πιθανοφάνειας (likelihood ratio test), με σκοπό τον έλεγχο καλής προσαρμογής και τη σύγκριση ανταγωνιστικών μοντέλων. Συνέπεια όλων αυτών, είναι να αποτελεί πανίσχυρο εργαλείο όχι μόνο στην απλή ανακατασκευή φυλογενετικών δέντρων, αλλά και στη διερεύνηση των ίδιων των μηχανισμών της εξελικτικής διαδικασίας, όπως για παράδειγμα στον έλεγχο της υπόθεση του μοριακού ρολογιού, ή του τρόπου με τον οποίο επηρεάζει η δαρβινική επιλογή την εξέλιξη των πρωτεϊνών. Ένα βασικό μειονέκτημα της μεθόδου, είναι ότι είναι υπολογιστικά εντατική, με το πιο δύσκολο κομμάτι να αποτελεί η αναζήτηση των δέντρων κάτω από το κριτήριο της μέγιστης πιθανοφάνειας. Παρόλα αυτά, η εξέλιξη των υπολογιστών, η αύξηση της υπολογιστικής ισχύος αλλά και μια σειρά βελτιώσεις σε αλγοριθμικό επίπεδο, έχουν κάνει τη μέθοδο να είναι η πλέον αποδεκτή και η περισσότερο χρησιμοποιούμενη τα τελευταία χρόνια, καθώς συνεχώς παρουσιάζονται παράλληλες υλοποιήσεις αλγορίθμων, υλοποιήσεις σε GPU αλλά και υλοποιήσεις σε FPGA.

Μια διαφορετική οπτική στη μέθοδο μέγιστης πιθανοφάνειας, χρησιμοποιούν οι λεγόμενες *Μπεϋζιανές μέθοδοι* (*Bayesian methods*) (Huelsenbeck, Ronquist, Nielsen, & Bollback, 2001). Με την Μπευζιανή στατιστική ανάλυση, ενσωματώνεται στο μοντέλο με «φυσικό» τρόπο η αβεβαιότητα στην εκτίμηση των παραμέτρων και απαντάται με άμεσο τρόπο το βιολογικό ερώτημα. Η κλασική μέθοδος της μέγιστης πιθανοφάνειας, χρησιμοποιεί την πιθανότητα των παρατηρήσεων, δεδομένου του δέντρου και των παραμέτρων, $P(\mathbf{x}|T,\theta)$, δηλαδή τη σχέση (6.26). Σε αντίθεση (Bland & Altman, 1998), οι μπεϋζιανές μέθοδοι, βασιζόμενες στο θεώρημα του Bayes, αντιστρέφουν το πρόβλημα και αντιμετωπίζουν τις παραμέτρους σαν τυχαίες μεταβλητές χρησιμοποιώντας την εκ των υστέρων κατανομή (posterior distribution):

$$P(T,\theta \mid \mathbf{x}) = \frac{P(T,\theta)\,P(\mathbf{x} \mid T,\theta)}{P(\mathbf{x})} \tag{6.28}$$

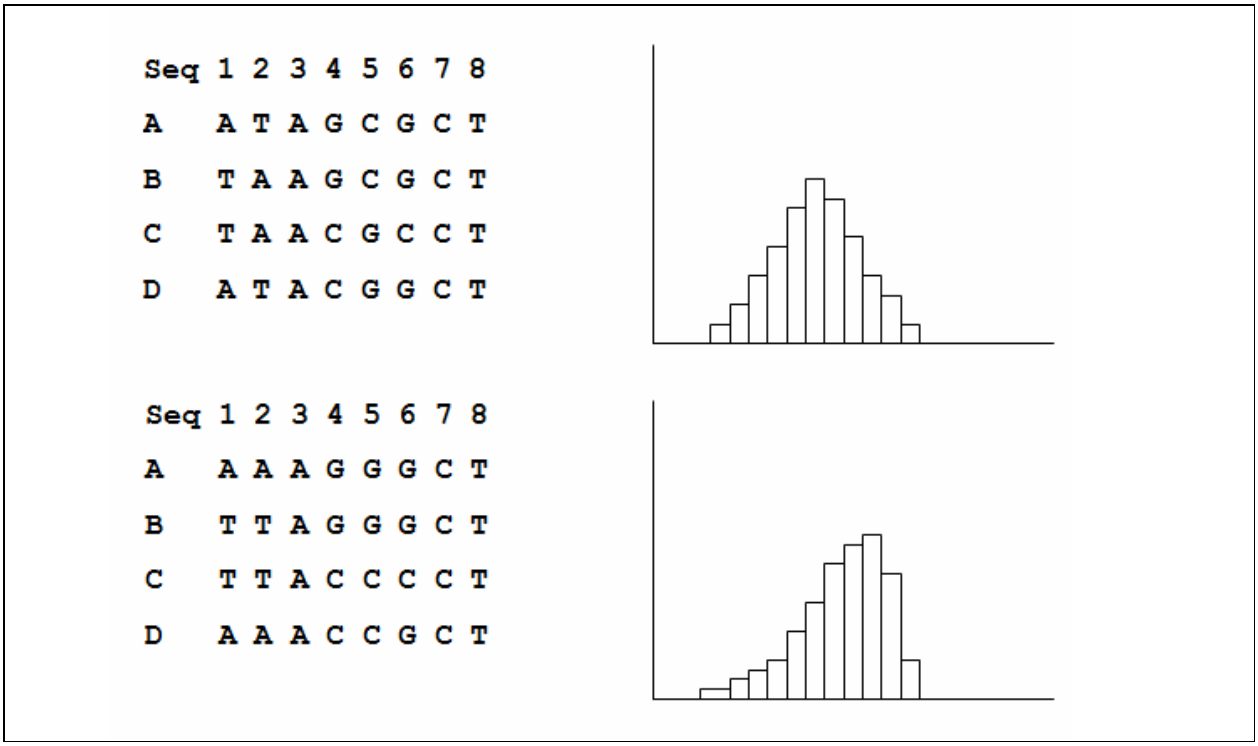
Στη σχέση (6.28), το $P(T,\theta\,|\mathbf{x})$ είναι η εκ των υστέρων κατανομή, $P(\mathbf{x}|T,\theta)$ η πιθανοφάνεια, ενώ $P(T,\theta)$ είναι η εκ των προτέρων κατανομή του δέντρου και των παραμέτρων. Ο παρονομαστής, $P(\mathbf{x})$, είναι μια σταθερά κανονικοποίησης η οποία χρησιμοποιείται έτσι ώστε η εκ των υστέρων κατανομή να είναι όντως πιθανότητα (δηλαδή, να αθροίζει στο ένα). Σε γενικές γραμμές, ακόμα και για τα απλά προβλήματα κλασικής στατιστικής, η εκ των υστέρων κατανομή δεν μπορεί να υπολογιστεί αναλυτικά, καθώς περιέχει δύσκολα ολοκληρώματα με πολλές διαστάσεις. Παρόλα αυτά, τέτοιες τεχνικές, χρησιμοποιούνται ευρέως τα τελευταία χρόνια στη βιοστατιστική και τη βιοπληροφορική και πραγματοποιούν τις εκτιμήσεις των παραμέτρων κάνοντας δειγματοληψία από την εκ των υστέρων κατανομή που προκύπτει από μια προσομοίωση με τη χρήση του MCMC (Markov Chain Monte Carlo) (Gilks, Richardson, & Spiegelhalter, 1996). Για το λόγο αυτό, είναι και πιο απαιτητικές από πλευρά υπολογιστικής ισχύος.

Το βασικό πλεονέκτημα της μεθόδου αυτής είναι ότι απαντάει με άμεσο και φυσικό τρόπο, μέσω της εκ των υστέρων κατανομής, στο βιολογικό ερώτημα («ποια είναι η πιθανότητα το δέντρο T να είναι σωστό, δεδομένων των παρατηρήσεων μου και του μοντέλου;»). Σε αντίθεση, η χρήση των τεχνικών της πιθανοφάνειας και των ελέγχων υποθέσεων γενικά, έχει μια δυσκολία στην κατανόηση από τους μη ειδικούς (Bland & Altman, 1998). Ένα άλλο πλεονέκτημα της μεθόδου είναι ότι επιτρέπει την ενσωμάτωση της εκ των προτέρων πληροφορίας η οποία μπορεί να προέρχεται από εξειδικευμένη γνώση. Πρακτικά όμως, τέτοιες περιπτώσεις είναι σπάνιες και μη-πληροφοριακές εκ των προτέρων κατανομές χρησιμοποιούνται στις

περισσότερες περιπτώσεις. Το βασικό μειονέκτημα της μεθόδου, είναι ότι είναι ιδιαίτερα απαιτητική υπολογιστικά (περισσότερο και από τη μέθοδο μέγιστης πιθανοφάνειας), ενώ η εκ των υστέρων κατανομή είναι ίσως περισσότερο ευαίσθητη σε περιπτώσεις λάθος ορισμού του μοντέλου. Παρόλα αυτά, είναι μια υποσχόμενη μέθοδος, η οποία κερδίζει συνεχώς έδαφος τα τελευταία χρόνια σε πολλούς τομείς της βιοπληροφορικής και της βιοστατιστικής.

6.5. Αξιολόγηση των δέντρων

Τελευταία, και ίσως πιο δύσκολη διαδικασία σε μια φυλογενετική ανάλυση, είναι το να εκτιμήσουμε πόσο καλό είναι το δέντρο που κατασκευάσαμε και αν ικανοποιούνται οι αρχικές προϋποθέσεις της ανάλυσης. Συνήθως δύο κατηγορίες μεθόδων είναι αυτές που χρησιμοποιούνται, οι εμπειρικές, δηλαδή αυτές που βασίζονται σε κάποια προσομοίωση ή μέθοδο τυχαίας δειγματοληψίας, και οι μέθοδοι που βασίζονται στις μαθηματικές ιδιότητες του ίδιου του μοντέλου.

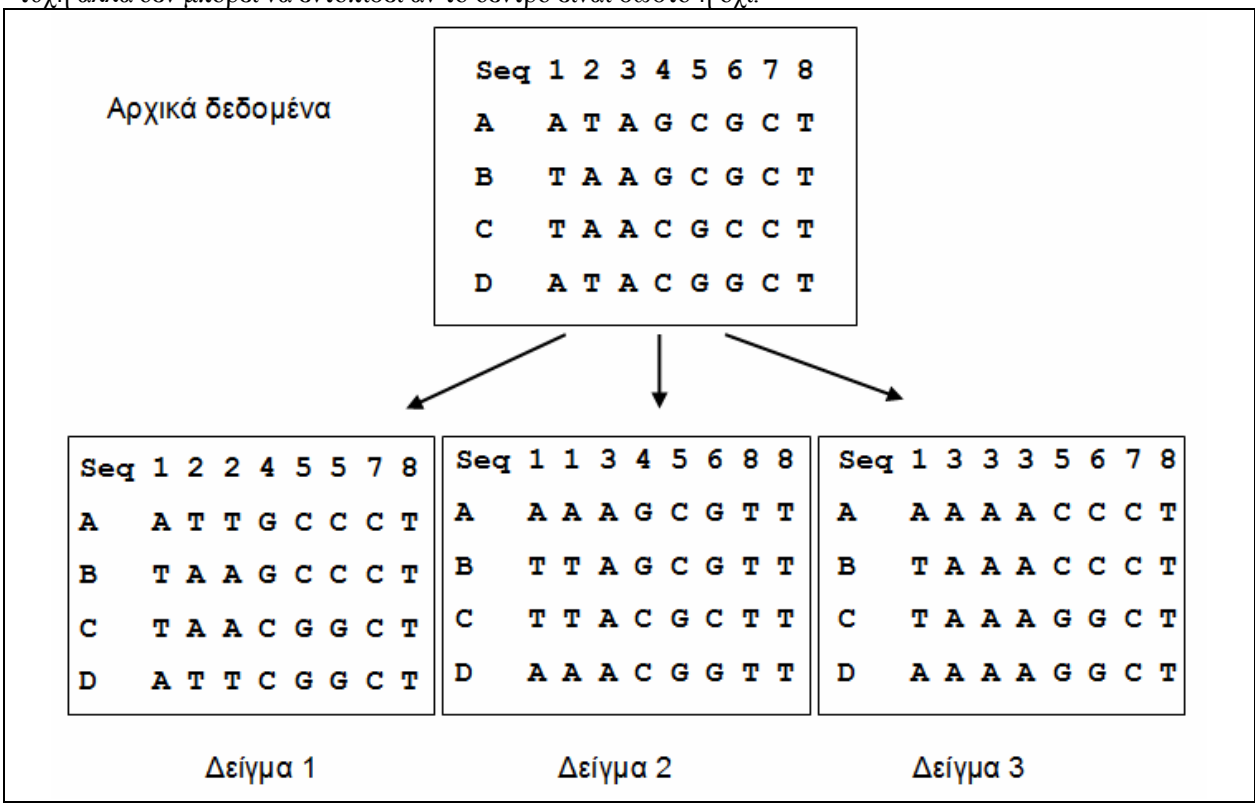


Εικόνα 6.10: *Αριστερά: υποθετικό παράδειγμα της μεθόδου της αντιμετάθεσης (permutation). Πάνω, φαίνεται μια αρχική πολλαπλή στοίχιση στην οποία υπάρχει συσχέτιση μεταξύ των στηλών (στήλες 1 με 2 και5 με 6), ενώ από κάτω μια τυχαία αντιμετάθεση στην οποία η συσχέτιση έχει εν πολλοίς αφαιρεθεί. Η ανάλυση των μηκών των βραχιόνων τέτοιων δέντρων που προκύπτουν από διαδοχικές αντιμεταθέσεις, θα πρέπει να δείξει μια κατανομή στην οποία, το παρατηρηθέν δέντρο θα βρίσκεται στο αριστερό άκρο (στο ελάχιστο). Δεξιά: η κατανομή των μηκών των βραχιόνων από πολλά τελείως τυχαιοποημένα (randomised) δέντρα. Πάνω, φαίνεται η συμμετρική κατανομή των μηκών από τυχαιοποιημένα δέντρα στα οποία δεν υπάρχει φυλογενετικό σήμα. Κάτω, φαίνεται η λοξή κατανομή που προκύπτει από ένα σύνολο δεδομένων στο οποίο υπάρχει τέτοιο σήμα.*

Στην τελευταία κατηγορία, ανήκει ο γνωστός έλεγχος του πηλίκου πιθανοφάνειας, ο οποίος προφανώς εφαρμόζεται μόνο στην περίπτωση ανάλυσης μέγιστης πιθανοφάνειας. Σε αυτή την περίπτωση, δύο ανταγωνιστικά «μοντέλα» (συνήθως το ένα να αποτελεί ειδική περίπτωση του άλλου) ελέγχονται μέσω του πηλίκου της πιθανοφάνειάς τους και η διαφορά συγκρίνεται με μια θεωρητική κατανομή χ^2. Γενικά η μέθοδος δεν έχει καλές ιδιότητες όταν πρόκειται να συγκριθούν ανταγωνιστικά φυλογενετικά δέντρα για να βρεθεί η σωστή τοπολογία, αλλά δουλεύει σωστά όταν πρόκειται να συγκρίνει εξελικτικά μοντέλα αντικατάστασης, ειδικά σε συνδυασμό με την παραμετρική bootsrap που θα δούμε παρακάτω (Goldman, 1993; Posada & Crandall, 1998). Ένα πλεονέκτημα των μπεϋζιανών μεθόδων, είναι όπως είπαμε, το ότι η εκ

των υστέρων κατανομή απάντά άμεσα και απλά στο πρόβλημα της πιθανότητας του δέντρου (αν και η μέθοδος έχει κατηγορηθεί ότι παράγει κάπως υπερβολικά αισιόδοξα ποσοστά).

Μια μέθοδος που χρησιμοποιείται συνήθως με την ανάλυση μέγιστης φειδωλότητας, είναι να ελέγχεται η κατανομή που δίνουν τα μήκη των τυχαιοποιημένων δέντρων και να συγκρίνεται, κυρίως με βάση τη λοξότητα της κατανομής, με αυτή που έχουν δέντρα τελείως τυχαία (χωρίς καμία εξελικτική πληροφορία). Στα λεγόμενα permutation tests, στήλες από την πολλαπλή στοίχιση αντιμετατίθενται, δηλαδή ανακατεύεται η σειρά εμφάνισης των χαρακτήρων τους, έτσι ώστε να πάρουμε ένα νέο σύνολο δεδομενων, δηλαδή μια στοίχιση, στην οποία οι παρατηρηθήσες συχνότητες σε κάθε στήλη να είναι ίδιες αλλά να έχει αφαιρεθεί η επίδραση της συσχέτισης μεταξύ θέσεων της ίδιας ακολουθίας. Η μέθοδος αυτή εφαρμόζεται συνήθως σε αναλύσεις φειδωλότητας για να δείξει κατά πόσο είναι πιθανό ένα δεδομένο δέντρο να έχει προέλθει κατά τύχη αλλά δεν μπορεί να εντοπίσει αν το δέντρο είναι σωστό ή όχι.



Εικόνα 6.11: *Υποθετικό Παράδειγμα της μεθόδου bootstrap. Πάνω φαίνεται μια αρχική πολλαπλή στοίχιση, ενώ από κάτω μια σειρά από τυχαίες δειγματοληψίες με επανάθεση ανάμεσα στις στήλες. Κάθε δείγμα θα αναλυθεί με την ίδια μέθοδο για να ελεγχθεί η σταθερότητα της. Προσέξτε ότι λόγω της δειγματοληψίας κάποιες στήλες δεν θα επιλεγούν σε κάποιο δείγμα, ενώ άλλες μπορεί να επιλεγούν περισσότερες από μία φορά.*

Ίσως η πιο γενική και ισχυρή μέθοδος, είναι αυτή του bootstrap. Είναι μια γνωστή μέθοδος στη στατιστική, και χρησιμοποιείται για ελέγχους σημαντικότητας σε δύσκολες περιπτώσεις (Efron & Tibshirani, 1993). Θα την ξανασυναντήσουμε στο κεφάλαιο 12 όπου θα δούμε τη χρήση της στον έλεγχο υποθέσεων για δεδομένα διαφορικής έκφρασης γονιδίων. Εν συντομία, η μέθοδος λειτουργεί σε δύο βήματα: 1) δημιουργεί πολλά «τεχνητά» σύνολα δεδομένων κάνοντας *δειγματοληψία με επανάθεση* από τις παρατηρήσεις (δηλαδή τις στήλες της στοίχισης) του αρχικού συνόλου δεδομένων, και 2) επαναλαμβάνει την ανάλυση για κάθε ένα από τα νέα αυτά σύνολα δεδομένων. Αν και φαίνεται εκ πρώτης όψεος παράδοξο, η μέθοδος έχει άριστες μαθηματικές ιδιότητες και οδηγεί σε καλό υπολογισμό των p-values και διαστημάτων εμπιστοσύνης. Στην περίπτωση των φυλογενετικών δέντρων (Hillis & Bull, 1993; Solitis & Solitis, 2003), η ερμηνεία είναι λίγο περίεργη, καθώς αν και αρχικά η μέθοδος προτάθηκε για να μετρήσει την επαναληψιμότητα μιας ανάλυσης, στην πράξη χρησιμοποιήθηκε από πολλούς για να δώσει μια εκτίμηση της πιθανότητας ότι το δέντρο είναι σωστό. Σε ανάλυση μέγιστης πιθανοφάνειας κάτω από προϋποθέσεις (να ισχύουν συγκεκριμένων μοντέλα της εξέλιξης, και τα δεδομένα να είναι πολλά), έχει δειχθεί ότι όντως η μέθοδος δίνει όντως μια κατανομή που προσεγγίζει την εκ των υστέρων πιθανότητα του δέντρου (Durbin, et al., 1998). Στην πραγματικότητα όμως,

και κάτω από συνθήκες στις οποίες το μοντέλο που χρησιμοποιείται δεν είναι το σωστό, η σωστή ερμηνεία είναι ότι αυτό που μετράει η bootstrap είναι η αξιοπιστία ενός συγκεκριμένου κλάδου του δέντρου (δηλαδή, το πόσο συχνά αυτός ο κλάδος ανακατασκευάζεται σωστά από να «τεχνητά» δεδομένα). Μια άλλη, αλλά όχι τόσο σωστή χρήση της μεθόδου είναι να κατασκευάζει ένα συναινετικό δέντρο (consensus tree) από τα αποτελέσματα των επαναλήψεων και να περιλαμβάνει σε αυτό κλάδους που εμφανίζονται στην πλειοψηφία των επαναλήψεων. Η γενική μέθοδος που περιγράψαμε, ονομάζεται μη-παραμετρική bootstrap και ειναι δυνατό να εφαρμοστεί με κάθε μέθοδο κατασκευής δέντρου.

Μια παραλλαγή της μεθόδου, η λεγόμενη παραμετρική bootstrap, παράγει τεχνητά δεδομένα του ίδιου μεγέθους με το αρχικό προσομοιώνονται κάτω από τις προϋποθέσεις του ίδιου εξελικτικού μοντέλου με το οποίο έγινε η ανάλυση (Goldman, 1993; Wollenberg & Atchley, 2000). Κάθε σύνολο δεδομένων αναλύεται με τον ίδιο τρόπο, και τα αποτελέσματα ερμηνεύονται περίπου με τον ίδιο τρόπο όπως και για την μη-παραμετρική bootstrap. Όπως είναι προφανές, η μέθοδος αυτή έχει ιδιαίτερη αξία αν χρησιμοποιηθεί με κάποια μέθοδο κατασκευής δέντρων η οποία υποθέτει ένα συγκεκριμένο μοντέλο της εξελικτικής διαδικασίας γιατί με τον τρόπο αυτό μπορούν να ελεγχθούν καλύτερα τόσο η ικανότητα ανακατασκευής του σωστού μοντέλου, όσο και η σταθερότητα της μεθόδου έναντι σε λάθος καθορισμό του μοντέλου (model misspecification). Οι μέθοδοι αυτές, λόγω της επαναληπτικής τους φύσης, είναι υπολογιστικά εντατικές καθώς απαιτούνται εκατοντάδες επαναλήψεις, στις οποίες ο αλγόριθμος κατασκευής δέντρων θα πρέπει επίσης να εφαρμοστεί επίσης επαναλληπτικά.

Γενικά, το πρόβλημα της εκτίμησης της πιθανότητας του δέντρου και της σύγκρισης ανταγωνιστικών δέντρων, είναι σύνθετο και με μεγάλη βιβλιογραφία. Όπως είπαμε παραπάνω, ούτε οι έλεγχοι πηλίκου πιθανοφάνειας, ούτε οι τιμές bootstrap από μόνες τους, προσφέρουν καλή εκτίμηση της πιθανότητα του δέντρου να έχει κατασκευαστεί σωστά. Έχουν αναπτυχθεί παρόλα αυτά, μια σειρά σύνθετες μέθοδοι οι οποίες χρησιμοποιούν τα αποτελέσματα της πιθανοφάνειας από τις διαδοχικές επαναλήψεις της bootstrap και κατασκευάζουν έναν έλεγχο για την ορθότητα του δέντρου, με αρκετά καλές στατιστικές ιδιότητες. Οι πιο γνωστές από αυτές τις τεχνικές, είναι ο έλεγχος των Kishino-Hasegawa (KH), ο έλεγχος των Shimodaira-Hasegawa (SH), ο σταθμισμένος (weighted) έλεγχος των Shimodaira-Hasegawa (WSH), και ο λεγόμενος Approximately Unbiased (AU) έλεγχος του Shimodaira, ο οποίος θεωρείται και ο καλύτερος. Το λογισμικό CONSEL, υλοποιεί τους παραπάνω ελέγχους, δεχόμενο σαν είσοδο τα αποτελέσματα από τις πιθανοφάνειες των ανταγωνιστικών δέντρων και τις αντίστοιχες τιμές από τις διαδοχικές επαναλήψεις bootsrap και έχει το πλεονέκτημα ότι μπορεί να χρησιμοποιηθεί με δεδομένα που προέρχονται από διαφορετικούς αλγόριθμους και διαφορετικά λογισμικά (Shimodaira & Hasegawa, 2001).

6.6. Η διαμάχη για την Εγκυρότητα των Μεθόδων-Πρακτικές Συμβουλές

Όπως ήδη αναφέραμε, η διαμάχη για το ποια μέθοδος (Μέγιστη Πιθανοφάνεια ή Μέγιστη Φειδωλότητα) είναι προτιμότερη για την εκτίμηση των φυλογενετικών σχέσεων, είναι παλιά και συνεχίζεται ακόμα, αν και μπορούμε να πούμε ότι τα τελευταία χρόνια, έχει κοπάσει κάπως. Πολλές μελέτες έχουν γίνει με χρήση προσομοιώσεων, οι οποίες έδειξαν αντικρουόμενα αποτελέσματα ως προς την προτίμηση στις δυο ανταγωνιστικές μεθόδους όσον αφορά τη ορθή ανακατασκευή των δέντρων, ανάλογα από το ποιες ήταν οι συνθήκες (το αρχικό μοντέλο) που παρήγαγαν τα δεδομένα.

Αξίζει να αναφερθεί, ότι ενώ αρχικά η μέθοδος της φειδωλότητας είχε προταθεί ως υπολογιστική απλοποίηση για την εύρεση της συνάρτησης πιθανοφάνειας, αμφισβητήθηκε έντονα στη συνέχεια, με κύριο επιχείρημα το ότι δεν έχει στατιστική ερμηνεία και δεν κάνει καθόλου υποθέσεις για τον τρόπο με τον οποίο έγινε η εξέλιξη. Παρ' όλα αυτά ο Felsenstein (Felsenstein, 1973, 1996) ισχυρίστηκε ότι όταν έχει συντελεστεί «λίγη» εξελικτική διαδικασία και ο ρυθμός της είναι περίπου σταθερός, τότε η παραπάνω προσέγγιση δίνει έγκυρα αποτελέσματα. Και αυτό όμως έχει αμφισβητηθεί από πολλούς που στηρίχθηκαν σε προσομοιώσεις χρησιμοποιώντας κάποιες ακραίες συνθήκες. Όταν οι ρυθμοί της εξέλιξης δεν είναι ίδιοι σε όλες τις εξελικτικές γραμμές (κάτι που συμβαίνει σχετικά συχνά), η μέθοδος της φειδωλότητας θα ανακατασκευάζει λανθασμένα δέντρα με μεγάλη πιθανότητα, η οποία θα μεγαλώνει καθώς το μέγεθος και ο αριθμός των ακολουθιών θα μεγαλώνει (Yang, 1996).

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 (Penny & Hendy, 2001; Steel & Penny, 2000), 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 (Felsenstein, 1996). 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 (Ota & Li, 2000).

Συμπερασματικά, ο αναγνώστης θα πρέπει να έχει στο μυαλό του τις ακόλουθες συμβουλές πριν από κάθε ανάλυση (Brinkman & Leipe, 2001):

- Να γίνεται προσεκτικός έλεγχος των δεδομένων εισόδου. Αυτό είναι κάτι που οι περισσότεροι το ξεχνάνε, αλλά όπως είπαμε, όλες οι αναλύσεις στηρίζονται στην αρχική πολλαπλή στοίχιση και αν αυτή είναι λάθος, όλες οι μετέπειτα αναλύσεις είναι επισφαλείς (είναι αυτό που λένε: «*garbage in, garbage out*»)
- Ίσως το πιο σωστό είναι να χρησιμοποιούμε όσο το δυνατόν περισσότερες μεθόδους και να συγκρίνουμε τα αποτελέσματα, προσπαθώντας παράλληλα να ελέγξουμε τις προϋποθέσεις κάτω από τις οποίες ισχύει η κάθε μια (κάτι που δεν είναι και τόσο εύκολο). Γενικά, αν υπάρχει κάτι σημαντικό στα δεδομένα, τις περισσότερες φορές αν οι μέθοδοι εφαρμοστούν σωστά, θα δείξουν το ίδιο ή περίπου το ίδιο.
- Να γίνει έλέγχος της σειράς των αλληλουχιών. Όσο και αν φαίνεται παράξενο, κάποιες μέθοδοι (ειδικά οι παλιές) παράγουν διαφορετικά αποτελέσματα ανάλογα με τη σειρά εισόδου των αλληλουχιών. Αν δεν είμαστε τελείως σίγουροι, καλό είναι να τοποθετούμε τις «περίεργες» αλληλουχίες προς το τέλος, ή αν είναι δυνατόν, να επαναλαμβάνουμε τις αναλύσεις αλλάζοντας με το χέρι τη σειρά των αλληλουχιών.
- Να γίνει προσεκτική επιλογή Outgroup στις περιπτώσεις που η μέθοδος που θα χρησιμοποιήσουμε παράγει δέντρο χωρίς ρίζα. Εκτός από τις κλασικές παραμέτρους που πρέπει να προσέξουμε (να ανήκει σε οργανισμό με μεγάλη εξελικτική απόσταση από τους υπό μελέτη οργανισμούς), θα πρέπει να έχουμε υπόψη μας ότι μπορεί η επιλεγμένη ακολουθία να διαθέτει κάποια «ειδικά» χαρακτηριστικά που να την κάνουν να μοιάζει περισσότερο από το αναμενόμενο σε κάποιες από τις αλληλουχίες υπό σύγκριση. Τέτοια χαρακτηριστικά, είναι η σύσταση σε GC% και ο ρυθμοί εξελικτικής αλλαγής, οπότε πρέπει να είμαστε έτοιμοι για εναλλακτικές στρατηγικές (πχ να υπάρχει και δεύτερο outgroup διαθέσιμο).

6.7. Λογισμικό

Οι περισσότερες από τις μεθόδους που αναφέραμε στις προηγούμενες παραγράφους, υπάρχουν διαθέσιμες σε υλοποιήσεις λογισμικού, το οποίο διατίθεται ελεύθερα στον τελικό χρήστη. Τα πιο παλιά από τα προγράμματα αυτά, είναι το PAUP και το PHYLIP τα οποία εξελίσσονται συνεχώς, αλλά τα τελευταία χρόνια υπάρχουν νέες προσθήκες με πακέτα λογισμικού που προσφέρουν μεγάλη ευκολία στο χρήστη αλλά και

μεγάλες ικανότητες ανάλυσης κάτω από διαφορετικά μοντέλα και προϋποθέσεις (πχ MEGA, RAxML κλπ). Στην ενότητα αυτή, παρουσιάζεται μια μικρή, αλλά ελπίζουμε κατατοπιστική περιγραφή των βασικότερων αλγοριθμικών υλοποιήσεων και πακέτων λογισμικού για φυλογενετική ανάλυση.

Το **PAUP** (Phylogenetic analysis using parsimony* and other methods), είναι ένα από τα πιο παλιά και γνωστά πακέτα φυλογενετικής ανάλυσης (Wilgenbusch & Swofford, 2003). Όπως φανερώνει και το όνομα του, αρχικά ξεκίνησε υλοποιώντας μεθόδους φειδολώτητας αλλά σταδιακά εμπλουτίστηκε και πλέον προσφέρει και μεθόδους αποστάσεων αλλά και μέγιστης πιθανοφάνειας με πολλές επιλογές. Το μειονέκτημα του είναι ότι διατίθεται με εμπορική άδεια χρήσης (http://www.sinauer.com/detail.php?id=8060). Το **PHYLIP** (PHYLogeny Inference Package) είναι επίσης ένα από τα πιο παλιά και αξιόπιστα πακέτα το οποία αναπτύχθηκε αρχικά από τον Joe Felsenstein (Retief, 2000). Στις σύγχρονες εκδόσεις υλοποιεί πληθώρα μεθόδων τόσο για μεθόδους αποστάσεων όσο και για μέγιστη πιθανοφάνεια και φειδωλότητα, ενώ διανέμεται πλέον κάτω από άδεια ανοικτού κώδικα (http://evolution.gs.washington.edu/phylip.html). Το **MEGA** (Molecular evolutionary genetic analysis) είναι ίσως το πιο χρησιμοποιημένο τα τελευταία χρόνια πακέτο φυλογενετικής ανάλυσης (Kumar, Nei, Dudley, & Tamura, 2008). Ενσωματώνει μεθόδους αποστάσεων, μέγιστης πιθανοφάνειας και φειδωλότητας, και το δυνατό του σημείο είναι ότι είναι ιδιαίτερα εύχρηστο και κατάλληλο για τον απλό χρήστη καθώς τρέχει σε περιβάλλον Windows με παραθυρική διεπαφή (http://www.megasoftware.net).

Η αύξηση της υπολογιστικής ισχύος, έχει δώσει όπως είναι εμφανές, ένα μεγάλο προβάδισμα στις μεθόδους μέγιστης πιθανοφάνειας, καθώς αναλύσεις οι οποίες μέχρι πριν μία-δύο δεκαετίες δεν μπορούσαν να γίνουν παρά μόνο από υπερ-υπολογιστές, πλέον μπορούν να πραγματοποιηθούν από τον μέσο χρήστη στον προσωπικό του Η/Υ. Τέτοιες μέθοδοι, οι οποίες εστιάζονται αποκλειστικά στην χρήση πιθανοφάνειας, είναι το HYPHY, το PAML, το PhyML και το RAxML. Το **HYPHY** (Hypothesis testing using phylogenies), είναι ένα πρόγραμμα το οποίο χρησιμοποιεί μέγιστη πιθανοφάνεια για φυλογενετικές αναλύσεις. Η ιδιαιτερότητα του είναι ότι υλοποιεί μια υψηλού επιπέδου γλώσσα στην οποία ο χρήστης μπορεί να ορίσει το μοντέλο και να πραγματοποιήσει εύκολα ελέγχους πηλίκου πιθανοφάνειας για τη σύγκριση των ανταγωνιστικών μοντέλων (http://www.hyphy.org). Το **PAML** (Phylogenetic analysis by maximum likelihood), ήταν από τα πρώτα πακέτα που εστίαζαν αποκλειστικά στη μέγιστη πιθανοφάνεια. Αναπτύχθηκε από τον Ziheng Yang (Yang, 2007) και η μεγάλη του δύναμη βρίσκεται στους ελέγχους για θετική επιλογή, στην ανακατασκευή προγονικών αλληλουχιών, στη χρονολόγηση μέσω του μοριακού ρολογιού και στην υλοποίηση πολλών διαφορετικών πιθανοθεωρητικών μοντέλων, παρά στις αναζητήσεις και στις συγκρίσεις φυλογενετικών δέντρων (http://abacus.gene.ucl.ac.uk/software/paml.html). Το **PhyML** είναι ένα γρήγορο πρόγραμμα για αναζητήσεις δέντρων κάτω από τη μέγιστη πιθανοφάνεια, το οποίο μπορεί να χρησιμοποιήσει τόσο αλληλουχίες DNA όσο και πρωτεϊνών http://www.atgc-montpellier.fr/phyml/binaries.php, (Bazinet, Zwickl, & Cummings, 2014). Τέλος, το **RAxML**, το οποίο έχει αναπτυχθεί από τον Έλληνα επιστήμονα Αλέξανδρο Σταματάκη (Stamatakis, 2014), είναι ένα γρήγορο και αποτελεσματικό πρόγραμμα για αναλύσεις μέγιστης πιθανοφάνειας με το γενικό μοντέλο (GTR), κάνοντας χρήση τόσο αμινοξικών όσο και νουκλεοτιδικών αλληλουχιών. Το δυνατό του σημείο, είναι οι παράλληλες υλοποιήσεις των αλγορίθμων που επιτρέπουν την ανακατασκευή τεράστιων φυλογενετικών δέντρων http://scoh-its.org/exelixis/software.html.

Μεθόδους μέγιστης πιθανοφάνειας, χρησιμοποιούν και άλλα πακέτα, αλλά με ελαφρώς διαφορετικό τρόπο. Με την Μπευζιανή στατιστική ανάλυση, ενσωματώνεται στο μοντέλο με «φυσικό» τρόπο η αβεβαιότητα στην εκτίμηση των παραμέτρων. Τέτοιες τεχνικές, χρησιμοποιούνται ευρέως τα τελευταία χρόνια στη βιοστατιστική και τη βιοπληροφορική και πραγματοποιούν τις εκτιμήσεις των παραμέτρων κάνοντας δειγματοληψία από την εκ των υστέρων κατανομή που προκύπτει από μια προσομοίωση με τη χρήση του MCMC (Markov Chain Monte Carlo). Για το λόγο αυτό, είναι και πιο απαιτητικές από πλευρά υπολογιστικής ισχύος. Το **MrBayes** είναι ένα από τα πιο γνωστά και παλιά τέτοια εργαλεία, και υλοποιεί φυλογενετική ανάλυση με χρήση MCMC (Huelsenbeck & Ronquist, 2001). Περιέχει επιλογές για όλα τα γνωστά πιθανοθεωρητικά μοντέλα αντικατάστασης των νουκλεοτιδίων αλλά και μοντέλα για αμινοξέα και κωδικόνια (http://mrbayes.net). Το **BEAST** (Bayesian evolutionary analysis sampling tree), είναι ένα άλλο πρόγραμμα Μπεϋζιανής ανάλυσης με χρήση MCMC (Drummond, Suchard, Xie, & Rambaut, 2012). Παράγει δέντρα με ρίζα κάτω από τις προϋποθέσεις του μοριακού ρολογιού, αλλά υποστηρίζει και μια σειρά από μοντέλα που χαλαρώνουν αυτές τις προϋποθέσεις. Μπορεί να χρησιμοποιηθεί με αμινοξικές ή νουκλεοτιδικές αλληλουχίες, αλλά και με άλλου είδους δεδομένα (πχ μορφολογικά). Περιέχει επίσης ρουτίνες όπως το Tracer και το FigTree, οι οποίες χρησιμεύουν στα διαγνωστικά και στην οπτικοποίηση των αποτελεσμάτων (http://beast.bio.ed.ac.uk).

Το **GARLI** (Genetic Algorithm for Rapid Likelihood Inference), ακολουθεί μια διαφορετική προσέγγιση στη φυλογενετική ανάλυση με χρήση πιθανοφάνειας (Bazinet, et al., 2014). Χρησιμοποιεί Γενετικούς Αλγόριθμους (μια τεχνική της τεχνητής νοημοσύνης) στην αναζήτηση του δέντρου μέγιστης πιθανοφάνειας. Περιέχει τόσο το γενικό μοντέλο GTR και τις ειδικές περιπτώσεις του, όσο και το μοντέλο της κατανομής Γάμμα, ενώ μπορεί να αναλύσει νουκλεοτιδικές και αμινοξικές αλληλουχίες αλλά και κωδικόνια. Ένα μεγάλο πλεονέκτημα είναι ότι διαθέτει και παράλληλες εκδόσεις των αλγορίθμων http://code.google.com/p/garli. Το **TNT** (Tree analysis using new technology) (Goloboff, Farris, & Nixon, 2008) είναι ένα πολύ γρήγορο πρόγραμμα για φυλογενετική ανάλυση με χρήση της φειδωλότητας το οποίο είναι ιδιαίτερα ικανό για αναλύσεις μεγάλων δέντρων http://www.lillo.org.ar/phylogeny/tnt/. Τέλος, αξίζει να αναφερθούμε και στο **TreeView** (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) το οποίο είναι ένα ελεύθερα διαθέσιμο πρόγραμμα οπτικοποίησης φυλογενετικών δέντρων. Το λογισμικό διαβάζει τους περισσότερους τύπους αρχείων που χρησιμοποιούν τα σχετικά προγράμματα (NEXUS, PHYLIP, Hennig86, NONA, MEGA, και ClustalW/X) και μπορεί να υποστηρίξει γραμματοσειρές TrueType and Postscript αλλά και γραφικά PICT (Macintosh) και Windows metafile (Windows) τα οποία επιτρέπουν εύκολη μεταφορά και επεξεργασία. Είναι διαθέσιμο για όλες τις γνωστές πλατφόρμες (Windows, Unix/Linux, και Macintosh), και διαθέτει και editor για επεξεργασία των δέντρων.

Βιβλιογραφία

Bazinet, A. L., Zwickl, D. J., & Cummings, M. P. (2014). A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. *Syst Biol, 63*(5), 812-818.

Bland, J. M., & Altman, D. G. (1998). Bayesians and frequentists. *BMJ, 317*(7166), 1151-1160.

Brinkman, F. S., & Leipe, D. D. (2001). Phylogenetic Analysis. In A. D. Baxevanis & B. F. Ouellette (Eds.), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (pp. 323-358): John Wiley & Sons, Inc.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet, 19*(3 Pt 1), 233-257.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in Proteins. In M. Dayhoff (Ed.), *In Atlas of protein sequence and structure* (Vol. 5, Suppl. 3, pp. 345-352): National biomedical research foundation, Silver Spring, MD.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc B, 39*, 1-38.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution, 29*(8), 1969-1973.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic*

Models of Proteins and Nucleic Acids.

Edwards, A. W., & Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. . *Annals of Human Genetics, 27*, 105.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics, 25*(5), 471.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution, 17*(6), 368-376.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol, 266*, 418-427.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology, 20*(4), 406-416.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *science, 155*(3760), 279-284.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (Eds.). (1996). *Markov Chain Monte Carlo in Practice* Chapman & Hall/CRC.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution, 36*(2), 182-198.

Goloboff, P., A.,, Farris, J. S., & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics, 24*(5), 774–786.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution, 22*(2), 160-174.

Hillis, D. M., & Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic biology, 42*(2), 182-192.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17*(8), 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science, 294*(5550), 2310-2314.

Jukes, T., & Cantor, C. (1969). Evolution of protein molecules Pp. 21–132 in HN Munro, ed. Mammalian protein metabolism: Academic Press, New York.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution, 16*(2), 111-120.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform, 9*(4), 299-306.

Lio, P., & Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research, 8*(12), 1233-1244.

Ota, S., & Li, W.-H. (2000). NJML: a hybrid algorithm for the neighbor-joining and maximum-likelihood methods. *Molecular Biology and Evolution, 17*(9), 1401-1409.

Penny, D., & Hendy, M. (2001). Phylogenetics: parsimony and distance methods. In D. J. Balding, M. Bishop & C. Cannings (Eds.), *Handbook of Statistical Genetics* (pp. 445-484): John Wiley and Sons, Ltd.

Posada, D., & Crandall, K. A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics, 14*(9), 817-818.

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol Biol, 132*, 243-258.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution, 4*(4), 406-425.

Shimodaira, H., & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics, 17*(12), 1246-1247.

Sokal, R. R., & Michener, C. D. (1958). A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin, 38*, 1409-1438.

Solitis, P. S., & Solitis, D. E. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Stat Sci, 18*(2), 256-267.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313.

Steel, M., & Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and evolution, 17*(6), 839-850.

Tavare, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17*, 57–86.

Wilgenbusch, J. C., & Swofford, D. (2003). Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics, Chapter 6*, Unit 6 4.

Wollenberg, K. R., & Atchley, W. R. (2000). Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences, 97*(7), 3288-3291.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution, 10*(6), 1396-1401.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution, 39*(1), 105-111.

Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution, 42*(2), 294-307.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution, 24*(8), 1586-1591.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet, 13*(5), 303-314.

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *SIAM Review, 37*(4), 531-551.