ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
& ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΤΗΡΙΟ ΜΙΚΡΟΕΠΕΞΕΡΓΑΣΤΩΝ
ΚΑΙ ΥΛΙΚΟΥ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΕΤΗ, ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ ΦΥΛΟΓΕΝΕΤΙΚΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΣΕ ΑΝΑΔΙΑΤΑΣΣΟΜΕΝΗ ΛΟΓΙΚΗ

ΝΙΚΟΣ ΑΛΑΧΙΩΤΗΣ

Επιβλέπων : Καθηγητής Απόστολος Δόλλας

Εξεταστική Επιτροπή:

Καθηγητής	Απόστολος Δόλλας
Αναπληρωτής Καθηγητής	Διονύσιος Πνευματικάτος
Επίκουρος Καθηγητής	Γιάννης Παπαευσταθίου

Περίληψη

Από τις αρχές της δεκαετίας του εξήντα άρχισε να επικρατεί η υπόθεση ότι συγκεκριμένες περιοχές του γενετικού υλικού μπορεί να περιέχουν σημαντικές πληροφορίες για την εξέλιξη του είδους και έτσι ξεκίνησαν να χρησιμοποιούνται ειδικές περιοχές του DNA για την μελέτη της εξελικτικής διαδικασίας και την κατασκευή εξελικτικών δέντρων. Οι αλγόριθμοι που χρησιμοποιούνται σήμερα σε φυλογενετικές αναλύσεις είναι ιδιαίτερα ακριβοί υπολογιστικά και απαιτούν πολύ χρόνο για να εκτελεστούν σε συμβατικούς υπολογιστές. Το υπολογιστικό κόστος αυξάνεται ακόμη περισσότερο με την συνεχή αύξηση του μεγέθους των βάσεων μοριακών δεδομένων.

Στα πλαίσια της παρούσας διπλωματικής εργασίας μελετήθηκε το πρόγραμμα RAxML, το οποίο χρησιμοποιείται ευρέως για την διεξαγωγή μεγάλης κλίμακας φυλογενετικών αναλύσεων εφαρμόζωντας την μέθοδο της Μέγιστης Πιθανοφάνειας. Το μεγαλύτερο μέρος του χρόνου εκτέλεσης του συγκεκριμένου προγράμματος καταναλώνεται στον υπολογισμό του βαθμού πιθανοφάνειας μεγάλου αριθμού διαφορετικών δέντρων. Η συνάρτηση υπολογισμού του βαθμού πιθανοφάνειας καταναλώνει το 95% του χρόνου εκτέλεσης του RAxML. Το υψηλό αυτό ποσοστό χρόνου, οδήγησε στην ιδέα της σχεδίασης συστημάτος βασισμένου σε αναδιατασσόμενη λογική.

Με σκοπό την επιτάχυνση του συγκεκριμένου προγράμματος, αυτή η εργασία παρουσιάζει μια νέα αρχιτεκτονική η οποία υπολογίζει το βαθμό πιθανοφάνειας δοσμένης τοπολογίας δέντρου. Η συνάρτηση φυλογενετικής πιθανοφάνειας χρησιμοποιείται επίσης και από άλλα γνωστά προγράμματα μέγιστης πιθανοφάνειας όπως τα IQPNNI, PHYML, GARLI αλλά και από προγράμματα Μπεϊσιανής φυλογενετικής ανάλυσης όπως το MrBayes. Το σύστημα που σχεδιάστηκε αποτελεί μια γενική υλοποίηση της συγκεκριμένης συνάρτησης δίνοντας την δυνατότητα να χρησιμοποιηθεί από όλα τα προγράμματα που υπολογίζουν το βαθμό πιθανοφάνειας για δοσμένη τοπολογία δέντρου και δεν αποτελει εξειδικευμένη έκδοση του προγράμματος RAxML.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Απόστολο Δόλλα, ο οποίος ήταν και ο επιβλέπων καθηγητής της παρούσας διπλωματικής εργασίας, για την εποικοδομητική συνεργασία που είχαμε σε επίπεδο διπλωματικής αλλά και για τον χρόνο που αφιέρωσε για να με ενημερώσει και να με κατατοπίσει σχετικά με κρίσιμες αποφάσεις που χρειάστηκε να πάρω.

Επίσης, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή κ. Διονύσιο Πνευματικάτο και τον Επίκουρο Καθηγητή κ.Γιάννη Παπαευσταθίου , οι οποίοι δέχθηκαν να αξιολογήσουν την διπλωματική μου εργασία.

Ακόμη, θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα κ. Ε.Σωτηριάδη για την συνεργασία που είχαμε τα τελευταία δύο χρόνια και να αναγνωρίσω την σημαντική βοήθεια και καθοδήγηση που μου παρείχε το τελευταίο εξάμηνο σχετικά με την διπλωματική μου εργασία.

Όπωσδήποτε, δεν πρέπει να παραλείψω να ευχαριστήσω τον Dr. Αλέξανδρο Σταματάκη, Junior Research Group Leader στο Technical University του Μονάχου, για την συνεργασία που είχαμε το τελευταίο εξάμηνο καθώς και για τον χρόνο που αφιέρωσε τόσο κατά τις επισκέψεις του στα Χανιά όσο και κατά την επίσκεψη μου στο Μόναχο για να μου λύσει σημαντικές απορίες.

Επίσης, θα ήθελα να ευχαριστήσω τον θείο μου Σταμάτη Αλαχιώτη, Καθηγητή Γενετικής στο Πανεπιστήμιο Πατρών, για τα βιβλία που μου έστειλε και την βοήθεια που μου παρείχε σε βιολογικό επίπεδο.

Ακόμη, θα ήθελα να ευχαριστήσω τον κ. Μάρκο Κιμιωνή, ο οποίος με εφοδίασε με τον κατάλληλο τεχνολογικό εξοπλισμό που χρειάστηκε για την διεκπεραίωση της παρούσας διατριβής.

Οπωσδήποτε, θα ήθελα να ευχαριστήσω όλους τους φίλους μου για την ευχάριστη παρέα τόσα χρόνια καθώς και όλους τους συμφοιτητές μου και μέλη του εργαστηρίου Μικροεπεξεργαστών και Υλικού.

Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου και την αδερφή μου, που με στηρίζουν και με βοηθούν σε όποια απόφαση κι αν πάρω.

Περιεχόμενα

Λίστα Εικόνων

Λίστα Πινάκων

1

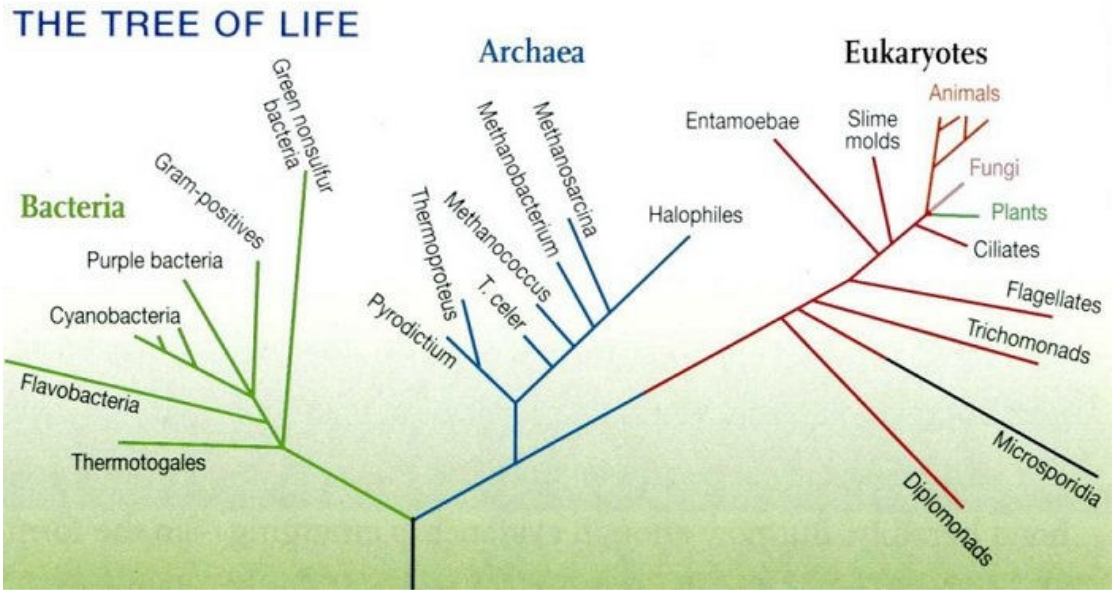
Εισαγωγή

Στο Πρώτο Κεφάλαιο της παρούσας διπλωματικής εργασίας γίνεται εισαγωγή στην εξέλικτική διαδικασία. Επίσης αναφέρεται η επιστημονική συνεισφορά της δουλειάς και περιγράφεται η δομή της εργασίας.

1.1 Εισαγωγή στην Εξέλιξη

Εξέλιξη σημαίνει αλλαγή στη μορφή και στην συμπεριφορά των ζωντανών οργανισμών από γενιά σε γενιά. Πρέπει να τονιστεί ότι ο όρος εξέλιξη δεν αναφέρεται στην αλλαγή κατά την ανάπτυξη ενός οργανισμού. Η εξελικτική θεωρία, η οποία ξεκίνησε από την εποχή του Δαρβίνου (Charles Darwin, 1809-1882) και συνεχίζεται μέχρι και σήμερα, «εξελίσεται» με πολύ γοργούς ρυθμούς. Ο όγκος των δεδομένων προς επεξεργασία και μελέτη αυξάνεται με την πάροδο των χρόνων και έτσι η διεξαγωγή μελετών με την χρήση των συμβατικών υπολογιστικών συστήματων αποτελεί σημαντικό εμπόδιο στην εξαγωγή αξιόπιστων αποτελεσμάτων σε εύλογο χρονικό διάστημα. Η βιοπληροφορική και η υπολογιστική βιολογία έχουν ως αντικείμενο μελέτης την ανάπτυξη και εφαρμογή υπολογιστικά απαιτητικών μεθόδων με σκοπό την κατανόηση βιολογικών διαδικασιών.

Οι εξελικτικές σχέσεις μεταξύ των ζωντανών οργανισμών αναπαρίστανται από ένα εξελικτικό δέντρο. Η κατασκευή του «δέντρου της ζωής», δηλαδή ένος εξελικτικού δέντρου που θα περιλαμβάνει όλους τους ζωντανούς και μη οργανισμούς αποτελεί μια ιδιαίτερα προκλητική ιδέα και έχει κεντρίσει το ερευνητικό ενδιαφέρον πολλών βιολόγων τόσο κατά το παρελθόν όσο και σήμερα. Παλαιότερα , οι οργανισμοί σε ένα εξελικτικό δέντρο κατατάσσονταν με κριτήριο την παρουσία ή απουσία παρατηρηθέντων χαρακτηριστικών. Το πρόβλημα όμως που παρουσιάστηκε ήταν το πως θα εντοπιστεί η σωστή θέση οργανισμών οι οποίοι δεν έχουν προφανή χαρακτηριστικά όπως για παράδειγμα τα Βακτήρια ή τα Αρχαιοβακτήρια. Μια γενική εικόνα του ταξινομημένου δέντρου της ζωής που ξεκινά 3.7 δισεκατομμύρια χρόνια πριν δείχνει τρείς κύριες υποδιαιρέσεις των έμβιων οργανισμών: τα Βακτήρια, τα Αρχαιοβακτήρια και τα Ευκαρυωτικά. (εικόνα 1-1). Για την γενικότερη αυτή ταξινόμηση όσο και για την ταξινόμηση των επιμέρους οργανισμών (taxa) ισχυρό εργαλείο σήμερα αποτελούν οι νουκλεοτιδικές ομοιότητες και διαφορές που αξιοποιούνται στην βάση ορισμένων παραδοχών και περιορισμών.



Εικόνα 1-1 : «Πανοραμική» άποψη του δέντρου της ζωής από απόσταση $3.7*10^9$ ετών (Πηγή: http://universe-review.ca/)

Από τις αρχές της δεκαετίας του εξήντα άρχισε να επικρατεί η υπόθεση ότι συγκεκριμένες περιοχές του γενετικού υλικού μπορεί να περιέχουν σημαντικές πληροφορίες για την εξελιξη του είδους και έτσι άρχισε να καλλιεργείται η ιδέα της χρησιμοποίησης ιδιαίτερα διατηρημένων περιοχών του DNA για την μελέτη της εξελικτικής διαδικασίας και κατασκευής εξελικτικών δέντρων με μεγάλο αριθμό οργανισμών. Αυτό οδήγησε στην ραγδαία αύξηση των δεδομένων προς επεξεργασία καθώς και την ανάπτυξη πολύπλοκων και απαιτητικών υπολογιστικά αλγορίθμων.

1.2 Επιστημονική Συνεισφορά

Σήμερα, υπάρχουν περισσότερα από 200 προγράμματα που χρησιμοποιούνται σε φυλογενετικές αναλύσεις και εφαρμόζουν περισσότερες από 25 διαφορετικές μεθόδους υπολογισμού φυλογενετικών δέντρων. Ενδεικτικά αναφέρεται ότι μέσα στον τελευταίο χρόνο δημιουργήθηκαν και άρχισαν να διανέμονται επίσημα στον παγκόσμιο ιστό 25 νέα προγράμματα σύμφωνα με στοιχεία που δημοσιεύθηκαν από το Department of Genome Sciences του τμήματος Ιατρικής του πανεπιστημίου της Ουάσιγκτον. (Πηγή: http://evolution.genetics.washington.edu/phylip/software.html)

Οι μέθοδοι της μέγιστης πιθανοφάνειας και μπεϊσιανής φυλογενετικής ανάλυσης, έχει αποδειχθεί ότι αποτελούν ένα επαρκές και ακριβές μοντέλο για τον υπολογισμό μεγάλων και πολύπλοκων δέντρων. Ο υπολογισμός όλων των δυνατών τοπολογιών για ένα δέντρο αυξάνεται πολύ γρήγορα καθώς προστίθενται νέοι οργανισμοί , γεγονός που κάνει το πρόβλημα ιδιαίτερα δύσκολο. Επίσης ο υπολογισμός του βαθμού πιθανοφάνειας (likelihood score) κάθε πιθανής τοπολογίας δέντρου είναι υπερβολικά απαιτητικός υπολογιστικά.

Με κίνητρο την μεγάλη χρονική διάρκεια εκτέλεσης φυλογενετικών αλγορίθμων που βασίζονται στην μέθοδο της μέγιστης πιθανοφάνειας, η παρούσα διπλωματική εργασία καλύπτει τον σχεδιασμό και την υλοποίηση της συνάρτησης φυλογενετικής πιθανοφάνειας σε αναδιατασσόμενη λογική (Field Programmable Gate Array-FPGA). Η κύρια συνεισφορά της εργασία εντοπίζεται στην σημαντική επιτάχυνση του υπολογισμού του βαθμού πιθανοφάνειας για μια δεδομένη τοπολογία δέντρου.

Αρχικά, μελετήθηκε το πρόγραμμα RAxML (Randomized Axelerated Maximum Likelihood), το οποίο κατά γενική ομολογία θεωρείται το πιο γρήγορο ανάμεσα σε άλλα για τον υπολογισμό του δέντρου με το maximum likelihood score. Σύμφωνα με υπόδειξη του κ. Σταματάκη, δημιουργού του προγράμματος RAxML, προτάθηκε η απεικόνιση συγκεκριμένων συναρτήσεων οι οποίες εκτελούνταν σε επίπεδο κλαδιού του δέντρου. Αφού εντοπίστηκαν και μελετήθηκαν οι συγκεκριμένες συναρτήσεις, μελετήθηκε η τεχνολογία αναδιατασσόμενης λογικής ώστε να εντοπιστεί η κατάλληλη FPGA για τις ανάγκες του προβλήματος. Το πρόβλημα εισόδου/εξόδου όμως που υπήρχε οδήγησε σε διαφορετική αντιμετώπιση του προβλήματος της επιτάχυνσης της συνάρτησης φυλογενετικής πιθανοφάνειας και προσανατόλισε την μελέτη στην δημιουργία σχεδίασης σε επίπεδο δέντρου η οποία δεν αντιμετώπιζε πρόβλημα εισόδου/εξόδου.

1.3 Δομή της Διπλωματικής Εργασίας

Το Κεφάλαιο 2 της παρούσας εργασίας αναφέρεται στις φυλογενετικές σχέσεις. Τι είναι φυλογένεια καθώς και τι περιλαμβάνει μια φυλογενετική ανάλυση εξηγούνται στο κεφάλαιο αυτό. Επίσης παρουσιάζονται οι τύποι των φυλογενετικών δέντρων και οι μέθοδοι κατασκευής τους. Στο Κεφάλαιο 3 εξηγείται η μέθοδος της μέγιστης πιθανοφάνειας όπως εφαρμόζεται σε μια φυλογενετική ανάλυση. Επίσης περιγράφονται τα βασικότερα μοντέλα υποκατάστασης των αλληλουχιών καθώς και ο τρόπος υπολογισμού του βαθμού πιθανοφάνειας ενός εξελικτικού δέντρου.

Το Κεφάλαιο 4 παρουσιάζει την αρχιτεκτονική που προτείνεται, αναλύωντας τα επιμέρους υποσυστήματα και εξηγώντας τις σχεδιαστικές επιλογές. Το Κεφάλαιο 5 περιλαμβάνει υλοποιηστικές λεπτομέρειες καθώς και πληροφορίες για την απόδοση της αρχιτεκτονικής υλοποιημένης σε FPGA σε σύγκριση με PC.

Τέλος, το Κεφάλαιο 6 περιέχει συμπεράσματα καθώς και ιδέες για μελλοντική επέκταση της αρχιτεκτονικής.

2

Φυλογενετικές Σχέσεις

Το Κεφάλαιο αυτό περιλαμβάνει το απαραίτητο βιολογικό υπόβαθρο. Εξηγούνται βασικές έννοιες σχετικά με τα φυλογενετικά δέντρα και ορολογία που θα χρησιμοποιηθεί στα επόμενα κεφάλαια.

2.1 Φυλογένεια και Φυλογενετική Ανάλυση

Ζωντανοί οργανισμοί υπάρχουν παντού πάνω στη γη, από τους πόλους μέχρι τον ισημερινό, από τα βάθη της θάλασσας μέχρι τον αέρα, από τα παγωμένα νερά μέχρι τις ερήμους. Τα τελευταία 3.7 δισεκατομύρια χρόνια, οι ζωντανοί οργανισμοί έχουν διαφοροποιηθεί και προσαρμοστεί σχεδόν σε κάθε περιβάλλον.

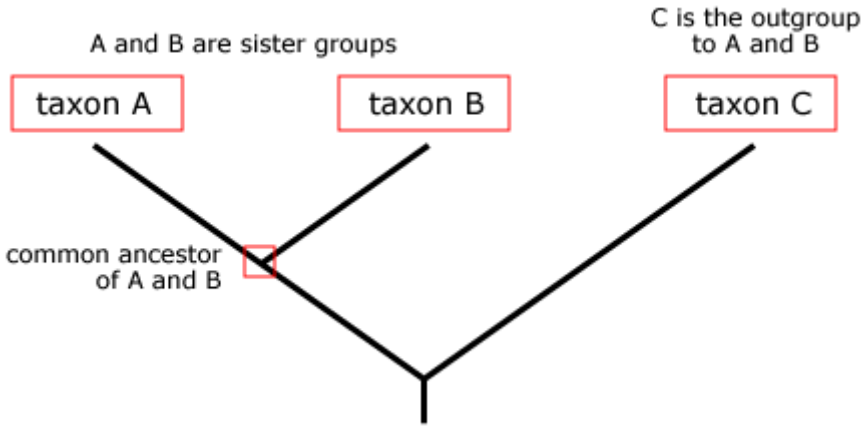
Όλοι οι γνωστοί οργανισμοί που υπάρχουν σήμερα καθώς και αυτοί που έχουν εξαλειφθεί «ενώνονται» μεταξύ τους μέσω της εξελικτικής ιστορίας. Όλοι οι οργανισμοί είναι εξελικτικά «ξαδέρφια» - κλαδιά στο δέντρο της ζωής. Πώς όμως θα μπορούσε κάποιος να συμπεράνει ότι ο άνθρωπος λόγου χάρη, και ο χιμπατζής έχουν έναν πιο πρόσφατο εξελικτικό κοινό πρόγονο απ' ότι έχουν ο ένας ή ο άλλος με την αμοιβάδα; Μια γρήγορη σκέψη οδηγεί στην άποψη ότι, επειδή ο άνθρωπος και ο χιμπατζής μοιάζουν περισσότερο μεταξύ τους παρά με την αμοιβάδα, θα είναι αυτό το ζευγάρι με τον πιο πρόσφατο κοινό πρόγονο. Η ομοιότητα αυτή βασίζεται στην σύγκριση των εξωτερικών τους χαρακτηριστικών.

Η ταξινομική αξιοποίηση των χαρακτηριστικών διαφόρων φυλογενετικών ομάδων βασίζεται στην αρχή της φειδωλότητας (parsimony), κατά την οποία μια φυλογενετική κατάταξη διαφόρων ομάδων ή ειδών (taxa) είναι περισσότερο πειστική αν οι εξελικτικές αλλαγές των εξετασθέντων χαρακτηριστικών είναι οι λιγότερες μεταξύ των ειδών που κατατάσσονται. Φυλογενετική επιστήμη ταξινόμησης είναι το όνομα του πεδίου της βιολογίας που ασχολείται με την επανακατασκευή της εξελικτικής ιστορίας και την μελέτη των σχέσεων μεταξύ των οργανισμών.

Σε μοριακό επίπεδο μια τέτοια προσέγγιση βασίζεται στην γενική αρχή ότι η αλληλουχία του DNA των έμβιων όντων κρύβει την εξελικτική τους ιστορία, καθώς ο βαθμός συγγένειας δύο οργανισμών είναι ανάλογος του βαθμού ομοιότητας των πληροφοριακών γενετικών στοιχείων τους.

2.1.1 Τι είναι ένα φυλογενετικό δέντρο

Μια φυλογένεια ή αλλιώς εξελικτικό δέντρο ή αλλιώς φυλογενετικό δέντρο ή αλλιώς κλαδόγραμμα είναι μια δενδρική δομή η οποία αναπαριστά τις εξελικτικές σχέσεις ανάμεσα σε ένα σύνολο από οργανισμούς ή ομάδες οργανισμών, τα taxa. Τα φύλλα (tips) του δέντρου αναπαριστούν τις νεότερες χρονολογικά ομάδες οργανισμών και συνήθως τα είδη(species), ενώ οι κόμβοι(nodes) του δέντρου αναπαριστούν κοινούς προγόνους(common ancestors) των ειδών. Δύο απόγονοι ενός κοινού προγόνου ονομάζονται αδερφικές ομάδες(sister groups). Επίσης είναι συχνό το φαινόμενο να υπάρχει ένα ή περισσότερα taxa που να έχει κοινό πρόγονο με κάποια sister groups αλλά να μην είναι το ίδιο sister group με τα υπόλοιπα(outgroup). Η εικόνα 2-1 δείχνει τις προαναφερθείσες έννοιες.



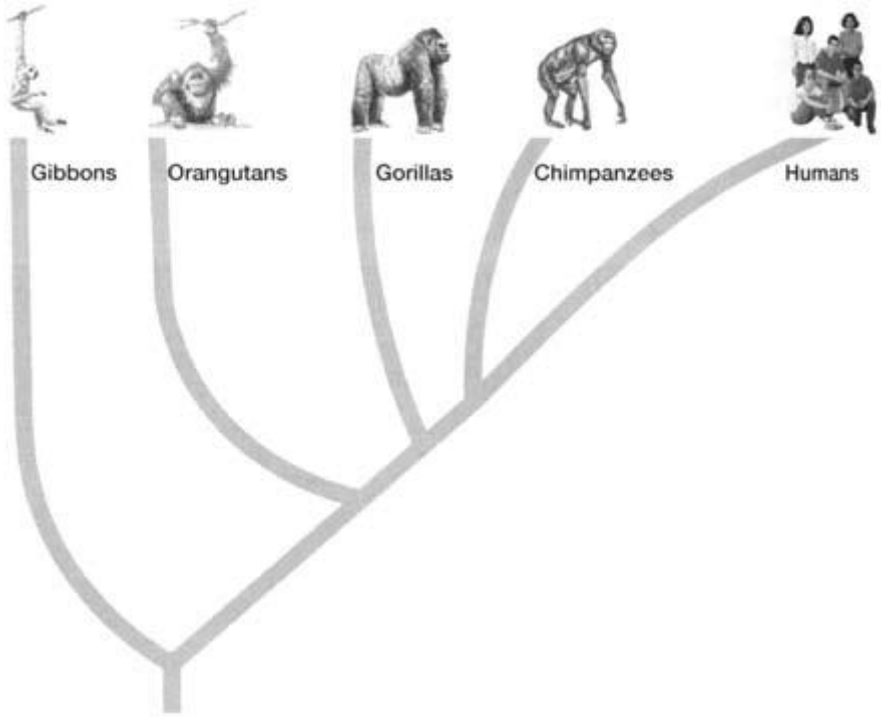
Εικόνα 2-1 : Ένα εξελικτικό δέντρο όπου φαίνονται οι έννοιες common ancestor, sister groups, και outgroup. (Πηγή: http://evolution.berkeley.edu/evolibrary/article/phylogenetics_02)

2.1.2 Τι είναι μια φυλογενετική ανάλυση

Μια φυλογενετική ανάλυση (phylogenetic inference) περιλαμβάνει την προσπάθεια υπολογισμού της εξελικτικής ιστορίας μια συλλογής οργανισμών. Αποτελείται από δύο κύρια συστατικά: τον υπολογισμό του εξελικτικού δέντρου(μιας προσέγγισης για την ακρίβεια) και την χρησιμοποίηση του προσεγγιστικού δέντρου για επιπλέον αναλυτική μελέτη. Η διαδικασία που ακολουθείται κατά την διεξαγωγή μια φυλογενετικής ανάλυσης με μοριακό επίπεδο, μπορεί να συνοψιστεί σε τρία βήματα. Αρχικά, γίνεται σύγκριση δύο ή περισσότερων ακολουθιών νουκλεοτιδίων που έχουν προέλθει από συγκεκριμένα σημεία του γενετικού υλικού των οργανισμών προς μελέτη, τα οποία πιστεύεται ότι δεν έχουν υποστεί σημαντικές μεταλλάξεις κατά την πάροδο του χρόνου. Στη συνέχεια γίνεται ανάλυση των οικογενειών των γονιδίων συμπεριλαμβάνοντας λειτουργικές προβλέψεις. Τέλος, γίνεται προσεγγιστικός υπολογισμός των εξελικτικών σχέσεων μεταξύ των οργανισμών.

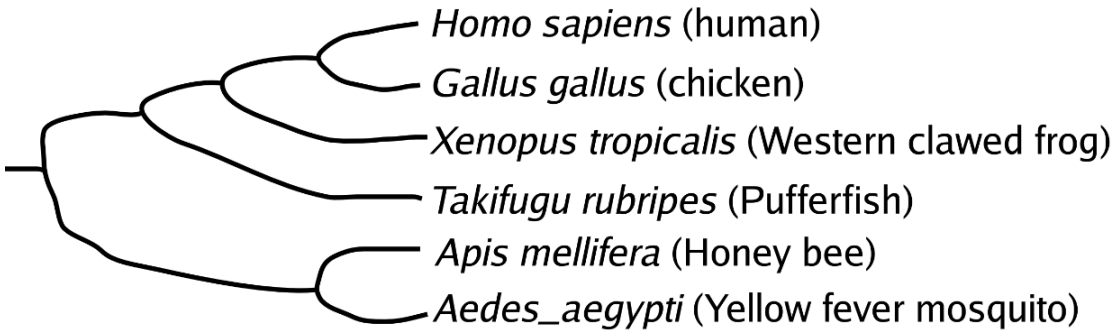
2.2 Τύποι Φυλογενετικών Δέντρων

Όπως έχει ήδη αναφερθεί, ένα φυλογενετικό δέντρο είναι μια γραφική αναπαράσταση της φυλογένειας μια ομάδας από taxa ή γονίδια και κατασκευάζεται με την αξιοποίηση των γενετικών πληροφοριών ενός ή λίγων γονιδίων. Οι αντικειμενικοί στόχοι της διεξαγωγής μιας φυλογενετικής μελέτης είναι δύο: η αναπαράσταση των πραγματικών γενεαλογικών σχέσεων των οργανισμών και η χρονολόγηση της διάσπασης των ειδών από τον τελευταίο τους πρόγονο.

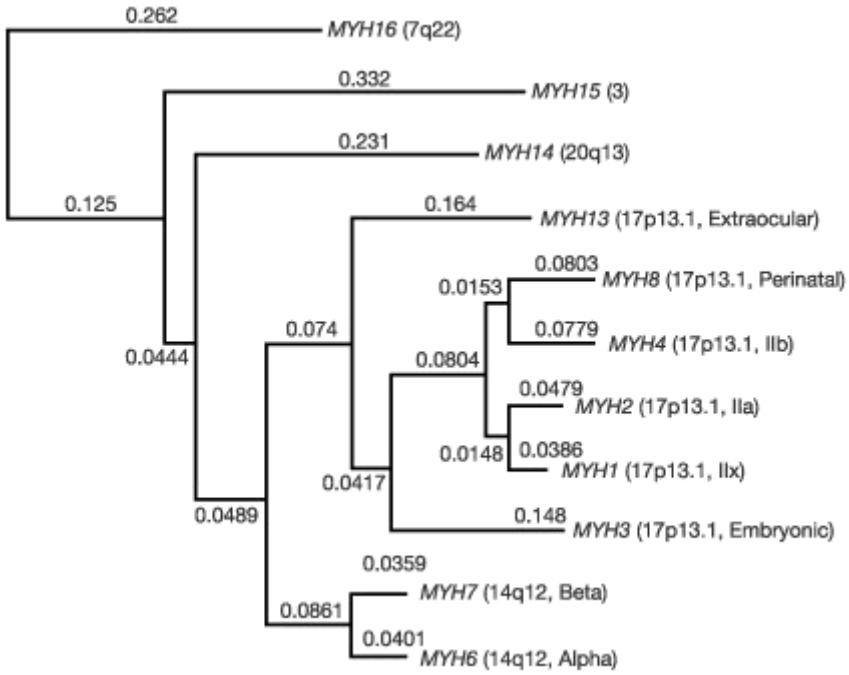


Εικόνα 2-2 : Ένα δέντρο ειδών (species tree) που δείχνει την εξελικτική σχέση των πιθήκων με τον άνθρωπο (Πηγή http://www.answersingenesis.org/)

Ένα φυλογενετικό δέντρο μπορεί να χρησιμοποιηθεί για την προσέγγιση των προαναφερθέντων στόχων. Όταν το δέντρο αντανακλά τις φυλογενετικές σχέσεις ομάδων πληθυσμών ή ειδών λέγεται φυλογενετικό δέντρο ειδών ή πληθυσμών ενώ όταν κατασκευάζεται με βάση τις νουκλεοτιδικές αλλαγές ενός γονιδίου ή λίγων γονιδίων από κάθε είδος τότε λέγεται γονιδιακό. Γενικότερα, το δέντρο που προκύπτει από μια φυλογενετική ανάλυση, είναι γονιδιακό δέντρο, και όχι μια φυλογένεια των ειδών (δέντρο ειδών) από τα οποία πάρθηκαν τα γονίδια, αν και στην ιδανική περίπτωση τα δύο αυτά δέντρα ταυτίζονται. Οι εικόνες 2-2 , 2-3 και 2-4 δείχνουν παραδείγματα δέντρων ειδών και γονιδιακών δέντρων.

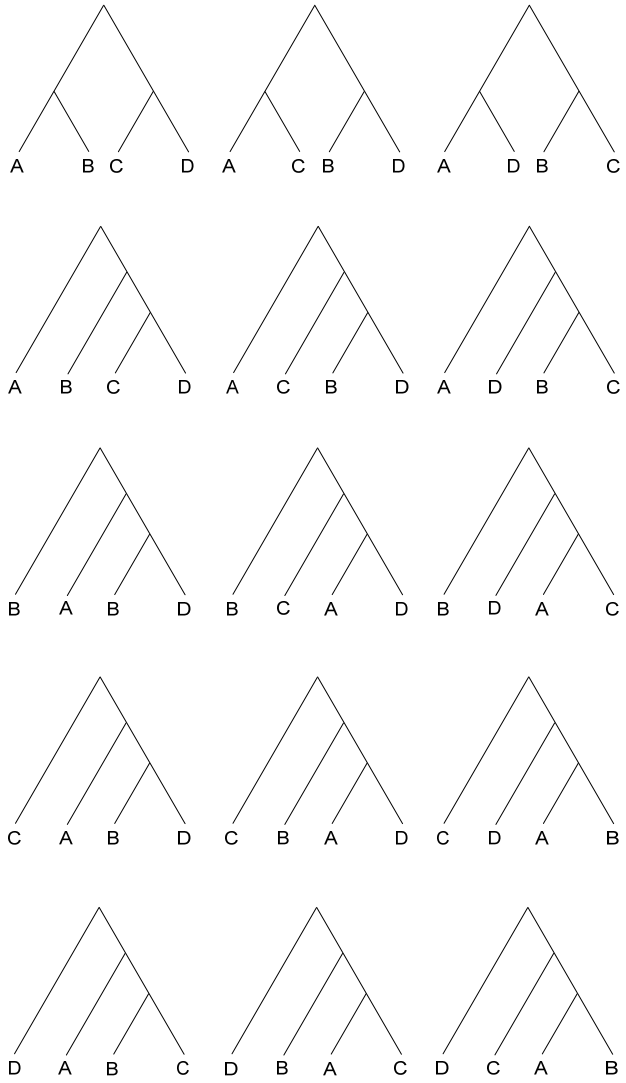


Εικόνα 2-3 : Ένα δέντρο ειδών (species tree) (Πηγή: http://bioinformatics.bio.uu.nl/)

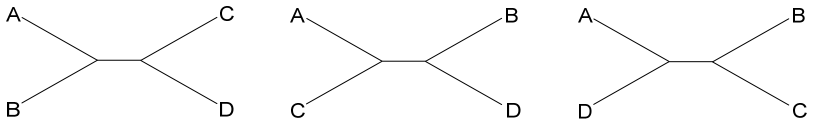


Εικόνα 2-4 : Ένα γονιδιακό δέντρο (gene tree) που έχει προκύψει από την σύγκριση του γονιδίου MYH (Πηγή http://genomic.unibe.ch/)

Τα φυλογενετικά δέντρα μπορεί να είναι με ρίζα (rooted), ως συνήθως, η οποία δείχνει τον κοινό πρόγονο καθώς και την εξελικτική κατεύθυνση, ή χωρίς ρίζα (unrooted), στα οποία δεν φαίνεται ούτε η ρίζα αλλά ούτε και η κατεύθυνση της εξελικτικής πορείας. Σε γενικές γραμμές, αξίζει να σημειωθεί ότι αν μελετώνται για παράδειγμα 3 είδη, τότε είναι δυνατά 3 δέντρα με ρίζα και 1 χωρίς ρίζα. Στις εικόνες 2-4 και 2-5, φαίνονται τα 15 δυνατά δέντρα με ρίζα και τα 3 χωρίς για 4 είδη.



Εικόνα 2-5 : Για 4 είδη (A,B,C,D) υπάρχουν 15 δυνατά δέντρα με ρίζα.



Εικόνα 2-6 : Για 4 είδη (A,B,C,D) υπάρχουν 3 δυνατά δέντρα χωρίς ρίζα.

Όπως είναι λογικό, όσο ο αριθμός των ειδών n αυξάνεται, τόσο αυξάνεται και ο αριθμός των πιθανών δέντρων. Ο αριθμός των δέντρων με ρίζα (Δ_ρ) για n είδη(OTUs : λειτουργικές ταξινομικές μονάδες, δλδ. οποιαδήποτε υπάρχουσα και αξιοποιούμενη στη μελέτη ταξινομική μονάδα όπως χαρακτήρας, είδος κ.α.) δίνεται από τον τύπο:

$$\Delta_{\rho} = \frac{(2n-3)!}{2^{n-2}(n-2)!} \quad \text{για } 2 \leq \text{n}$$

Ο αντίστοιχος αριθμός δέντρων χωρίς ρίζα ($\Delta_{\chi\rho}$) δίνεται από τον τύπο:

$$\Delta_{x\rho} = \frac{(2n-5)!}{2^{n-3}(n-3)!} \quad \text{για } 3 \leq \text{n}$$

Στον ακόλουθο πίνακα φαίνεται πως αυξάνεται ο αριθμός των δυνατών δέντρων με ρίζα και χωρίς ρίζα με την αύξηση του αριθμού των ειδών. Γίνεται εύκολα κατανοητό, τόσο από τους παραπάνω δύο τύπους όσο και από τον πίνακα, ότι είναι πολύ δύσκολο να πιστοποιηθεί το αληθινό φυλογενετικό δέντρο, ιδιαίτερα όταν ο αριθμός των ειδών μεγαλώνει.

Αριθμός Ειδών	Αριθμός Δέντρων με ρίζα	Αριθμός Δέντρων χωρίς ρίζα
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025
15	2,13458E+14	7,90585E+12
20	8,20079E+21	2,21643E+20
50	2,75292E+76	2,83806E+74

Πίνακας 2-1 : Πίνακας που δείχνει τον αριθμό διαφορετικών δέντρων, με ρίζα και χωρίς ρίζα, για αριθμό ειδών από 3 μέχρι 50.

Τέλος, ένα ακόμη στοιχείο που απασχολεί τους βιολόγους που διεξάγουν φυλογενετικές αναλύσεις, όσο αφορά την κατασκευή φυλογενετικών δέντρων είναι ο χρόνος που πέρασε από την διάσπαση κάθε ζευγαριού ειδών. Με βάση αυτό το κριτήριο δημιουργούνται δύο ειδών δέντρα: αναμενόμενα δέντρα απόστασης (expected distance tree) και ρεαλιστικά δέντρα απόστασης(realistic distance trees). Σε ένα αναμενόμενο

δέντρο απόστασης, πρέπει τα μήκη των δύο βραχιόνων που οδηγούν στο ζευγάρι των ειδών από τον κοινό τους πρόγονο να είναι ίσα, ακόμη και για γονιδιακά δέντρα. Αν ο ρυθμός της γονιδιακής αντικατάστασης είναι σταθερός, η αναμενόμενη εξελικτική απόσταση πρέπει να είναι η ίδια. Στα αναμενόμενα δέντρα απόστασης το μήκος των βραχιόνων είναι ανάλογο του εξελικτικού χρόνου. Μπορεί επίσης να συμβαίνει, ο πραγματικός αριθμός των γονιδιακών υποκαταστάσεων να μην είναι ο ίδιος στις δύο εξελικτικές γραμμές λόγω του στοχαστικού στοιχείου της αστάθειας του ρυθμού της γονιδιακής υποκατάστασης. Σε αντιδιαστολή λοιπόν με τα αναμενόμενα δέντρα απόστασης, χρησιμοποιούνται τα ρεαλιστικά δέντρα απόστασης.

Πρέπει να σημειωθεί ότι ένα δέντρο ειδών είναι πάντα αναμενόμενο δέντρο απόστασης, ενώ ένα γονιδιακό δέντρο μπορεί να είναι είτε αναμενόμενο είτε ρεαλιστικό.

2.3 Μέθοδοι Κατασκευής Φυλογενετικών Δέντρων

Οι μέθοδοι κατασκευής φυλογενετικών δέντρων κατατάσσονται σε δύο βασικές κατηγορίες: την κατηγορία μητρών απόστασης (distance matrix methods) και την κατηγορία που βασίζεται στην παρουσία ή απουσία πληροφοριακών χαρακτήρων (character-based methods). Στην πρώτη κατηγορία ανήκουν οι μέθοδοι: UPGMA, των μετασχηματισμένων αποστάσεων, των Fitch-Margoliash και των γειτονικών ζευγαριών (neighbor joining). Στην δεύτερη κατηγορία ανήκουν οι μέθοδοι: μέθοδος της μέγιστης φειδωλότητας (maximum parsimony), μέθοδος της μέγιστης πιθανοφάνειας (maximum likelihood) και μέθοδος της μπεϊσιανής ανάλυσης(bayesian analysis).

Οι προαναφερθέντες μέθοδοι θα αναπτυχθούν περιγραφικά στην συνέχεια , εκτός από την μέθοδο της μέγιστης πιθανοφάνειας η οποία θα αναλυθεί εκτενώς στο επόμενο κεφάλαιο.

2.3.1 Κατηγορία μητρών απόστασης

Στην κατηγορία αυτή, οι εξελικτικές αποστάσεις (στην συνηθισμένη περίπτωση με την μορφή νουκλεοτιδικών ή αμινοξεϊκών διαφορών) υπολογίζονται για όλα τα ζευγάρια και χρησιμοποιείται ένας αλγόριθμος για την κατασκευή του δέντρου.

2.3.1.1 Η μέθοδος UPGMA

Η μέθοδος αυτή (UPGMA-Unweighted pair-group method with arithmetic mean) προϋποθέτει σταθερούς ρυθμούς εξέλιξης μεταξύ των γενεαλογικών γραμμών, δλδ. υπάρχει γραμμική σχέση μεταξύ των εξελικτικών αποστάσεων και του χρόνου διάσπασης. Έστω 4 λειτουργικές ταξινομικές μονάδες (OTUs) των οποίων οι γενετικές αποστάσεις φαίνονται στην πίνακα 2-2.

OTU(i/j)	A	B	C
B	D_{AB}		
C	D_{AC}	D_{BC}	
D	D_{AD}	D_{BD}	D_{CD}

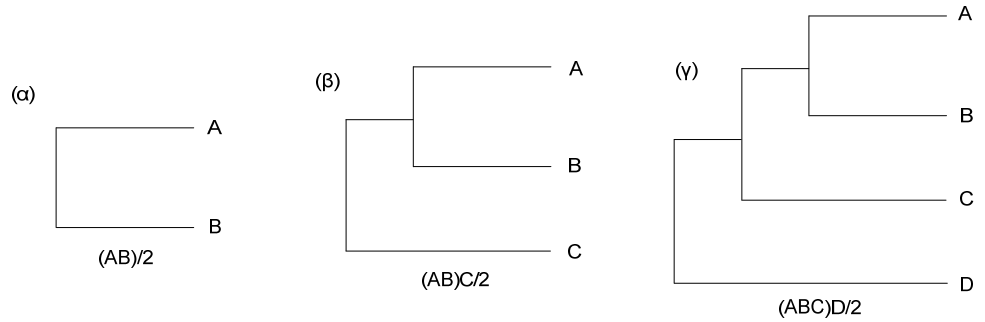
Πίνακας 2-2 : Κατάταξη γενετικών αποστάσεων D τεσσάρων ταξινομικών μονάδων OTUs με τη μορφή μήτρας

Αν γίνει η υπόθεση ότι η μικρότερη γενετική απόσταση είναι μεταξύ των μονάδων A και B τότε οι δύο αυτές μονάδες θα ομαδοποιηθούν. Το σημείο διακλάδωσης υπολογίζεται από την απόσταση $D_{AB}/2$(εικόνα 2-7α). Η ενοποιημένη μονάδα θεωρείται πλέον ως μια σύνθετη ταξινομική μονάδα και δημιουργείται η νέα μήτρα που φαίνεται στον πίνακα 2-3.

OUT(i/j)	AB	C
C	$D_{(AB)C}$	
D	$D_{(AB)D}$	D_{CD}

Πίνακας 2-3 : Κατάταξη γενετικών αποστάσεων D τριών ταξινομικών μονάδων OTUs ,μιας σύνθετης (AB) και δύο απλών (C,D), με τη μορφή μήτρας.

Η απόσταση μεταξύ μιας σύνθετης μονάδας και μιας απλής είναι η μέση τιμή τους, δηλαδή $D_{(AB)C} = (D_{AC}+D_{BC})/2$ και $D_{(AB)D} = (D_{AD}+D_{BD})/2$. Η μία από τις δύο αυτές αποστάσεις θα είναι μικρότερη και έτσι θα δημιουργηθεί το νέο ζευγάρι. Έστω ότι $D_{(AB)C} < D_{(AB)D}$. Προκύπτει λοιπόν η ένωση της σύνθετης λειτουργικής ταξινομικής μονάδας (AB) με την C με απόσταση διακλάδωσης ίση προς $D_{(AB)C}/2$(εικόνα 2-7β) . Το τελευταίο βήμα αφορά την ομαδοποίηση της τελευταίας μονάδας D με την σύνθετη ABC , ενώ η ρίζα ολόκληρου του δέντρου τοποθετείται σε απόσταση ίση με: $D_{(ABC)D}/2 = [(D_{AD}+D_{BD}+ D_{CD})/3]/2$(εικόνα 2-7γ) .



Εικόνα 2-7 : Βαθμιαία δόμηση ενός φυλογενετικού δέντρου με τέσσερις λειτουργικές ταξινομικές μονάδες με την χρησιμοποίηση της μεθόδουUPGMA.

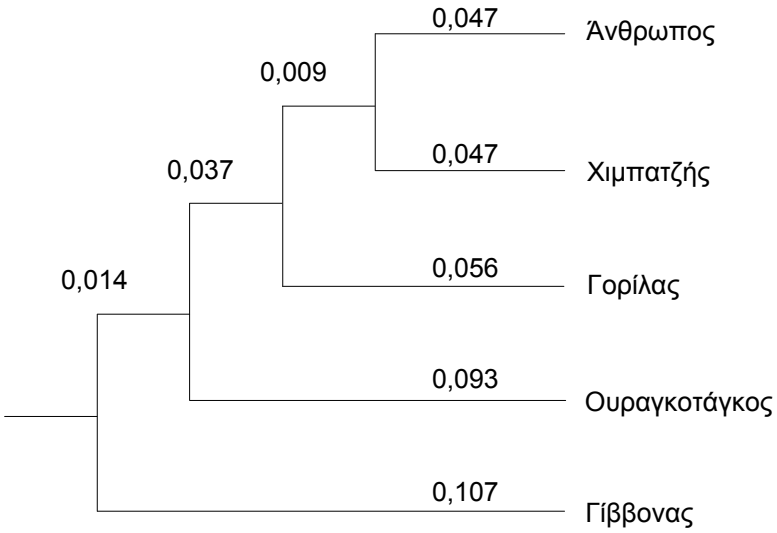
Παράδειγμα

Στον πίνακα 2-4 παρουσιάζονται οι υπολογισμένες νουκλεοτιδικές υποκαταστάσεις μιας περιοχής mtDNA πέντε ειδών Πρωτευόντων.

OTU(i/j)	Άνθρωπος (A)	Χιμπατζής(B)	Γορίλας(C)	Ουραγκοτάγκος(D)
Χιμπατζής(B)	0.094			
Γορίλας(C)	0.111	0.115		
Ουραγκοτάγκος(D)	0.180	0.194	0.188	
Γίββονας(E)	0.207	0.218	0.218	0.216

Πίνακας 2-4 : Υπολογισμένες νουκλεοτιδικές υποκαταστάσεις μιας περιοχής mtDNA σε πέντε είδη πρωτευόντων

Από τον πίνακα 2-4 , με εφαρμογή της μεθόδου UPGMA προκύπτει το φυλογενετικό δέντρο της εικόνας 2-8.



Εικόνα 2-8 : Φυλογενετικό δέντρο που κατασκευάστηκε με την χρησιμοποίηση της μεθόδου UPGMA.

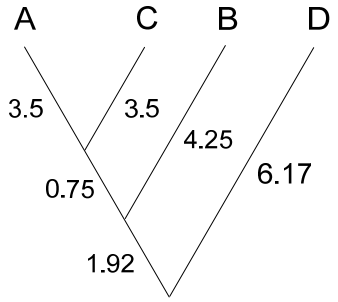
2.3.1.2 Η μέθοδος των μετασχηματισμένων αποστάσεων

Η μέθοδος UPGMA δεν μπορεί να εφαρμοστεί αν ο ρυθμός υποκατάστασης δεν είναι σταθερός σε όλες τις γενεαλογικές γραμμές, διότι προκύπτουν λαθεμένα δέντρα τόσο ως προς την τοπολογία όσο και ως προς το μήκος των βραχιόνων. Η μέθοδος των μετασχηματισμένων αποστάσεων χρησιμοποιεί ένα εξωτερικό είδος αναφοράς (πχ. για το είδος αναφοράς είναι γνωστό ότι έχει διασπαστεί πριν από τα άλλα είδη) για να κάνει διορθώσεις στις αποστάσεις, οι οποίες στην συνέχεια χρησιμοποιούνται από την μέθοδο UPGMA για την κατασκευή του δέντρου.

Έστω για παράδειγμα η ακόλουθη μήτρα θεωρητικών αποστάσεων, οι οποίες με την εφαρμογή της μεθόδου UPGMA δίνουν το δέντρο της εικόνας 2-9.

OTU(i/j)	A	B	C
B	8		
C	7	9	
D	12	14	11

Πίνακας 2-5 : Μήτρα θεωρητικών αποστάσεων (νουκλεοτιδικές διαφορές)



Εικόνα 2-9 : Φυλογενετικό δέντρο που κατασκευάστηκε με τη μέθοδο UPGMA χωρίς να ληφθεί υπόψη η πιθανότητα άνισων ρυθμών υποκατάστασης στους βραχίονες.

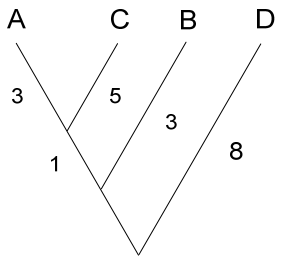
Το παραπάνω δέντρο είναι λαθεμένο εφόσον οι ρυθμοί υποκατάστασης ποικίλουν. Για να εφαρμοστεί η μέθοδος μετασχηματισμένων αποστάσεων έστω ότι θεωρείται ως είδος αναφοράς το D. Ακολουθεί μια διαδικασία διόρθωσης των αποστάσεων χρησιμοποιώντας την εξίσωση:

$d'_{ij} = [\frac{(d_{ij} - d_{iD} - d_{jD})}{2}] + d''_{D}$ όπου $d''_{D} = \frac{(d_{AD} - d_{BD} - d_{CD})}{3}$ και i=A,B ή C και d_{ij} η μετασχηματισμένη απόσταση.

Εφαρμόζωντας την εξίσωση της προηγούμενης σελίδας προκύπτει η νέα μήτρα με τις διορθωμένες τιμές που φαίνεται στον πίνακα 2-6.

OUT(i/j)	A	B
B	10/3	
C	10/3	13/3

Πίνακας 2-6 : Μήτρα υπολογισμένων αποστάσεων



Εικόνα 2-10 : Διορθωμένο Φυλογενετικό δέντρο με την μέθοδο των μετασχηματισμένων αποστάσεων.

2.3.1.3 Η μέθοδος των Fitch-Margoliash

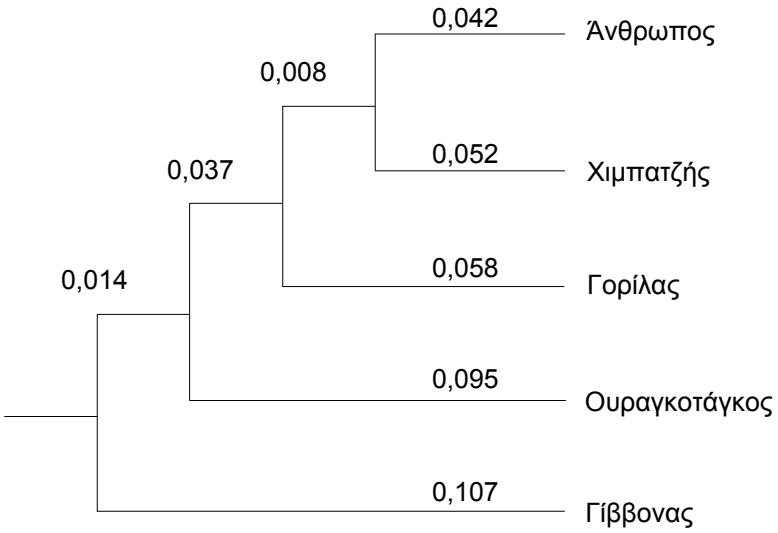
Η μέθοδος Fitch-Margoliash θεωρείται κατάλληλη για να πάρει κανείς πιο σωστά και πιο πραγματικά μήκη βραχιόνων σε ένα δέντρο του οποίου η τοπολογία έχει διορθωθεί με την εφαρμογή των μεθόδων UPGMA και μετασχηματισμένων αποστάσεων, όπως περιγράφηκε στην προηγούμενη ενότητα.
Έστω ο πίνακας 2-4 από τον οποίο προέκυψε με την εφαρμογή της μεθόδου UPGMA το δέντρο της εικόνας 2-8.Γενικά πρέπει να τονιστεί ότι αν υπάρχουν περισσότερες από τρείς ταξινομικές μονάδες, η γενική πορεία επεξεργασίας που ακολουθείται είναι να προβάλλονται κάθε φορά σε τρείς, με τη μια να είναι σύνθετη και να αποτελείται από όλες εκτός των δύο που δείχνουν την μικρότερη απόσταση.
Με βάση το γεγονός ότι η μικρότερη απόσταση είναι μεταξύ του ανθρώπου και του χιμπατζή, προκύπτουν οι ακόλουθες αποστάσεις:

$D_{AB} = 0.094$
$D_{AC} = (0.011+0.180+0.207)/3 = 0.166$
$D_{BC} = (0.115+0.194+0.218)/3 = 0.176$

Θεωρώντας την περίπτωση των τριών ταξινομικών μονάδων, οι αριθμοί των υποκαταστάσεων στον κάθε βραχίονα είναι :

$X=(D_{AB} + D_{AC} + D_{BC})/2 =0.042$
$Y=(D_{AB} - D_{AC} - D_{BC})/2 =0.052$
$Z=(-D_{AB} + D_{AC} + D_{BC})/2 =0.124$

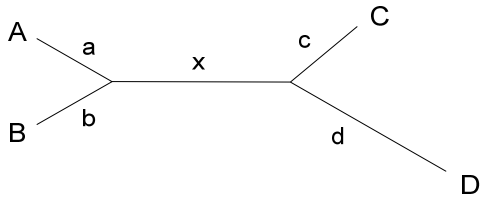
Στην συνέχεια, οι μονάδες Α και Β σχηματίζουν μια σύνθετη ταξινομική μονάδα την ΑΒ. Ξαναϋπολογίζονται οι αποστάσεις μεταξύ της νέας σύνθετης ταξινομικής μονάδας ΑΒ και των άλλων μονάδων και επιλέγονται και πάλι οι δύο που έχουν την μικρότερη απόσταση. Οι δύο αυτές μονάδες συμβολίζονται πάλι ως Α και Β, ενώ η C αναπαριστάνει την σύνθετη μονάδα, που αποτελείται από όλες τις υπόλοιπες. Στη συνέχεια ξαναϋπολογίζονται τα νέα X,Y,Z. Η πορεία αυτή συνεχίζεται μέχρι να συνδυαστούν όλες οι ταξινομικές μονάδες σε μια απλή οικογένεια. Τέλος, η μέθοδος καταλήγει στο ακόλουθο αναδομημένο φυλογενετικό δέντρο.



Εικόνα 2-11 : Αναδομημένο φυλογενετικό δέντρο με την μέθοδο Fitch-Margoliash.

2.3.1.4 Η μέθοδος των γειτονικών ζευγαριών

Η συγκεκριμένη μέθοδος (neighbor-joining method) δεν βασίζεται σε ίσους ρυθμούς εξέλιξης των αλληλουχιών DNA. Έστω το δέντρο χωρίς ρίζα της εικόνας 2-12, στο οποίο οι ταξινομικές μονάδες A και B είναι γειτονικές, όπως και οι C και D.



Εικόνα 2-12: Δέντρο χωρίς ρίζα για 4 OTUs.

Για το συγκεκριμένο δέντρο, υπο την προϋπόθεση ότι ισχύει η ιδιότητα της προσθετικότητας των αποστάσεων, ισχύει η σχέση:

$D_{AC} + D_{BD} + D_{AD} + D_{BC} = a + b + c + d + 2x = D_{AB} + D_{CD} + 2x$

Από την παραπάνω σχέση προκύπτουν δύο συνθήκες:

i) $D_{AB} + D_{CD} < D_{AC} + D_{BD}$ και ii) $D_{AB} + D_{CD} < D_{AD} + D_{BC}$

Οι παραπάνω συνθήκες εφαρμόζονται στην περίπτωση που έχουμε 4 ταξινομικές λειτουργικές μονάδες άγνωστης φυλογενετικής σχέσης, ώστε να εντοπιστούν τα γειτονικά ζευγάρια. Για 5 ταξινομικές λειτουργικές μονάδες, που θα μελετηθούν παρακάτω, υπάρχουν 5 δυνατές περιπτώσεις τετράδων, ενώ για x ταξινομικές λειτουγικές μονάδες, ο αριθμός των τετράδων στις οποίες θα επιχειρηθεί η εύρεση των γειτονικών ζευγαριών δίνεται από τον τύπο $\frac{x!}{[4!(x-4)!]}$.

Έστω λοιπόν η μήτρα που φαίνεται στον πίνακα 2-7 όπου παρουσιάζονται οι μέσες τιμές νουκλεοτιδικών υποκαταστάσεων ανα 100 θέσεις μιας διαγονιδιακής περιοχής γονιδίων της σφαιρίνης.

OTU(i/j)	Άνθρωπος (A)	Χιμπατζής(B)	Γορίλας(C)	Ουραγκοτάγκος(D)
Χιμπατζής(B)	1,45			
Γορίλας(C)	1,51	1,57		
Ουραγκοτάγκος(D)	2,98	2,94	3,04	
Γίββονας(E)	7,51	7,59	7,39	7,10

Πίνακας 2-7 : Μέσες τιμές νουκλεοτιδικών υποκαταστάσεων ανα 100 θέσεις μιας γονιδιακής περιοχής γονιδίων της σφαιρίνης.

Γενικά, για κάθε τετράδα ταξινομικών μονάδων, έστω i, j, k, και l, υπολογίζονται οι παραστάσεις: $D_{ij} + D_{kl}$, $D_{jk} + D_{jl}$ και $D_{il} + D_{jk}$. Έτσι λοιπόν, για τις ταξινομικές μονάδες του πίνακα 2-7 προκύπτει ο ακόλουθος πίνακας:

Ταξινομικές Μονάδες OTUs	Άθροισμα Ζευγαριών	Επιλεγόμενο γειτονικό ζευγάρι
A, B, C, D	$D_{AB} + D_{CD} = 4,49$ $D_{AC} + D_{BD} = 4,45$ $D_{AD} + D_{BC} = 4,49$	(AC), (BD)
A, B, C, E	$D_{AB} + D_{CE} = 8,84$ $D_{AC} + D_{BE} = 9,06$ $D_{AE} + D_{BC} = 9,08$	(AB), (CE)
A, B, D, E	$D_{AB} + D_{DE} = 8,55$ $D_{AD} + D_{BE} = 10,57$ $D_{AE} + D_{BD} = 10,45$	(AB), (DE)
A, C, D, E	$D_{AC} + D_{DE} = 8,61$ $D_{AD} + D_{CE} = 10,37$ $D_{AE} + D_{CD} = 10,55$	(AC), (DE)
B, C, D, E	$D_{BC} + D_{DE} = 8,67$ $D_{BD} + D_{CE} = 10,33$ $D_{BE} + D_{CD} = 11,59$	(BC), (DE)

Πίνακας 2-8 : Καταγραφή γειτονικών ζευγαριών με βάση τις αποστάσεις του πίνακα 2-7.

Για κάθε γραμμή του παραπάνω πίνακα, εξετάζονται τα αθροίσματα της δεύτερης στήλης και αφού επιλεγεί το μικρότερο, το ζευγάρι στο οποίο αντιστοιχεί αποθηκεύεται στην τρίτη στήλη. Έτσι λοιπόν, από τον συγκεκριμένο πίνακα προκύπτουν οι εξής τελικές καταμετρήσεις: (AB) = 2, (AC) = 2, (AD) = 0, (BC) = 1, (BD) = 1, (BE) = 0, (CD) = 0, (CE) = 1, (DE) = 3 . Την μεγαλύτερη συχνότητα έχει το ζευγάρι DE το οποίο αποτελεί πλέον το πρώτο γειτονικό ζευγάρι. Το ζευγάρι αυτό θεωρείται ως απλή ταξινομική μονάδα και προκύπτει η ακόλουθη μήτρα όπως και στην μέθοδο UPGMA:

OTU(i/j)	Άνθρωπος (A)	Χιμπατζής(B)	Γορίλας(C)
Χιμπατζής(B)	1,45		
Γορίλας(C)	1,51	1,57	
(DE)	5,25	5,25	5,22

Πίνακας 2-9: Νέα μήτρα με την συγχώνευση δύο ταξινομικών λειτουργικών μονάδων σε μια απλή με τη χρήση της μεθόδου γειτονικών ζευγαριών.

Αφού υπάρχουν πλέον μόνο 4 ταξινομικές μονάδες υπολογίζονται οι 3 παραστάσεις που αναφέρθηκαν παραπάνω και από αυτές επιλέγεται αυτή που δίνει το μικρότερο

άθροισμα. Έτσι προκύπτει: $D_{AB} + D_{C(DE)} = 6,67 < D_{AC} + D_{B(DE)} = 6,76 < D_{A(DE)} + D_{BC)} = 6,82$ και άρα επιλέγεται το (AB) ως το ένα γειτονικό ζευγάρι και το C(DE) ως το άλλο γειτονικό ζευγάρι.
Τέλος, πρέπει να αναφερθεί ότι οι περισσότερες μέθοδοι κατασκευής φυλογενετικών δέντρων δίνουν δέντρα χωρίς ρίζα. Για να μετατραπεί ένα άριζο δέντρο σε ανάλογο με ρίζα χρησιμοποιείται ένα εξωτερικό είδος αναφοράς, γνωστό από άλλες πληροφορίες όπως π.χ. παλαιοντολογικές, και τοποθετείται η ρίζα μεταξύ αυτού του είδους και του σημείου που το συνδέει με τις άλλες μονάδες. Το είδος αναφοράς πρέπει να έχει διασπαστεί πριν από τα υπόλοιπα αλλά δεν πρέπει αν έχει πολύ μεγάλη απόσταση από τα άλλα, αλλά ούτε και πολύ μικρή για να μπορεί να διακρίνεται. Ωστόσο, υπάρχει περίπτωση να μην μπορεί να βρεθεί είδος αναφοράς. Σε αυτήν την περίπτωση, και υπο την προϋπόθεση ότι ο ρυθμός εξέλιξης είναι χοντρικά ο ίδιος για όλους τους βραχίονες, η ρίζα μπορεί να τοποθετηθεί στο μέσο σημείο της μεγαλύτερης απόστασης μεταξύ δύο μονάδων.

2.3.2 Κατηγορία μεθόδων που βασίζονται στην παρουσία-απουσία πληροφοριακών χαρακτήρων

2.3.2.1 Η μέθοδος της μέγιστης φειδωλότητας

Στις μεθόδους που περιγράφηκαν μέχρι τώρα χρησιμοποιούνταν όλες οι πολυμορφικές θέσεις, είτε αυτές είναι αμινοξικές είτε νουκλεοτιδικές υποκαταστάσεις, για την εύρεση της σωστής τοπολογίας του δέντρου. Η μέθοδος της μέγιστης φειδωλότητας (maximum parsimony) βασίζεται στην αξιοποίηση των μικρότερων εξελικτικών αλλαγών που απαιτούνται για να εξηγηθούν οι διαφορές οι οποίες παρατηρούνται μεταξύ των ταξινομικών λειτουργικών μονάδων. Στην μέθοδο αυτή γίνεται διάκριση μεταξύ πληροφοριακών και μη πληροφοριακών θέσεων.

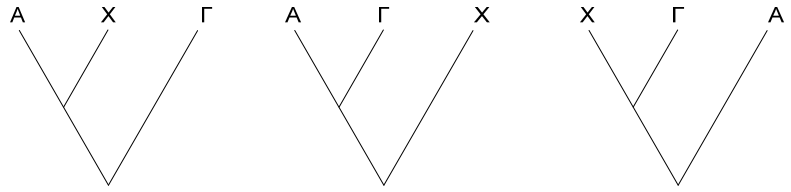
Στον πίνακα 2-10 φαίνονται 4 πληροφοριακές θέσεις από μια αλληλουχία DNA του γονιδιώματος του ανθρώπου, του χιμπατζή, του γορίλα και του ουρακοτάγκου.

Θέση	Άνθρωπος	Χιμπατζής	Γορίλας	Ουρακοτάγκος
1	A	G	A	G
2	C	C	A	A
3	-	-	T	T
4	G	G	A	A

Πίνακας 2-10: Τέσσερις πληροφοριακές θέσεις από μια αλληλουχία DNA του γονιδιώματος του ανθρώπου, του χιμπατζή, του γορίλα και του ουρακοτάγκου.

Μια νουκλεοτιδική θέση είναι πληροφοριακή μόνο αν αφορά τουλάχιστον δύο διαφορετικά νουκλεοτίδια που το καθένα απαντάται τουλάχιστον δύο φορές. Η συγκεκριμένη αυτή θέση αξιοποιείται για να βρεθεί το πιο φειδωλό δέντρο. Στην δεύτερη γραμμή του παραπάνω πίνακα λόγου χάρη, που αφορά τα νουκλεοτίδια C και A τα οποία βρίσκονται στην $34^{η}$ θέση των συγκρινόμενων αλληλουχιών DNA, τα C είναι κοινά στον άνθρωπο και στο χιμπατζή και διαφορετικά από τα αντίστοιχα νουκλεοτίδια των απόμακρων προγόνων τους. Οι πιθανές σχέσεις του ανθρώπου, του χιμπατζή και του γορίλα φαίνονται στην εικόνα 2-13, στην οποία αποτυπώνονται

τρείς υποθέσεις: κατά την πρώτη υπόθεση ο άνθρωπος και ο χιμπατζής ανήκουν σε έναν κλάδο, κατά την δεύτερη υπόθεση ο χιμπατζής και ο γορίλας ανήκουν σε έναν κλάδο, ενώ κατά την τρίτη ο άνθρωπος και ο γορίλας ανήκουν σε έναν κλάδο.



Εικόνα 2-13: Τρία πιθανά φυλογενετικά δέντρα με ρίζα, για τον άνθρωπο, το χιμπατζή και τον γορίλα.

Όλα τα παραδείγματα που παρουσιάστηκαν μέχρι τώρα υποστηρίζουν την άποψη ότι ο άνθρωπος έχει μεγαλύτερη εξελικτική συγγένεια με το χιμπατζή παρά με τον γορίλα. Στο ίδιο αποτέλεσμα καταλήγει και το παράδειγμα της μέγιστης φειδωλότητας.

Καταλήγωντας, το πιο σωστό φυλογενετικό δέντρο θεωρείται μεταξύ όλων των πιθανών εκείνο που κατασκευάζεται με τον μικρότερο αριθμό νουκλεοτιδικών αλλαγών. Στην μέθοδο αυτή, ο βαθμός κάθε δέντρου υπολογίζεται χρησιμοποιώντας έναν απλό αλγόριθμο ο οποίος καθορίζει πόσες εξελικτικές μεταλλάξεις απαιτούνται για να εξηγήσουν την κατανομή κάθε οργανισμού.

Παρόλο που η μέθοδος της μέγιστης φειδωλότητας είναι μια απλή προσέγγιση, δεν είναι στατιστικά συνεπής. Δοθείσης επαρκής ποσότητας πληροφορίας, δεν είναι εγγυημένο ότι θα παραχθεί το σωστό δέντρο με την υψηλότερη πιθανότητα. Ο όρος συνέπεια στο συγκεκριμένο ζήτημα σημαίνει την μονοτονική σύγκλιση σε σωστή απάντηση με την προσθήκη όλο και περισσότερων δεδομένων, ιδιότητα που είναι επιθυμητή σε κάθε στατιστική μέθοδο. Το 1978, ο Joe Felsenstein απέδειξε ότι η μέθοδος της μέγιστης φειδωλότητας μπορεί να γίνει ασυνεπής υπό συγκεκριμένες συνθήκες. Η κατηγορία των περιπτώσεων στις οποίες είναι γνωστό ότι θα παρατηρηθεί ασυνέπεια λέγεται long branch attraction και μπορεί να παρατηρηθεί για παράδειγμα όταν από κοινό πρόγονο ξεκινάνε δύο μακρυά κλαδία για δύο είδη, δλδ. υψηλό ποσοστό μεταλλάξεων, ενώ ξεκινάνε επίσης κοντά κλαδία για άλλα δύο είδη.

2.3.2.2 Η μέθοδος της μέγιστης πιθανοφάνειας

Ανάμεσα στις πιο διάσημες εναλλακτικές φυλογενετικές μεθόδους είναι η μέθοδος της μέγιστης πιθανοφάνειας (maximum likelihood method). Η συγκεκριμένη μέθοδος χρησιμοποιεί εξελικτικά μοντέλα αλληλουχιών DNA που είτε βασίζονται σε ίσους ρυθμούς υποκατάστασης για όλα τα νουκλεοτίδια, είτε συμβαίνουν με διαφορετικούς ρυθμούς. Με δεδομένο το μοντέλο που επιλέγεται, υπολογίζεται η πιθανοφάνεια των παρατηρούμενων δεδομένων με τη χρησιμοποίηση της μεθόδου της μέγιστης πιθανοφάνειας και η καλύτερη εκτίμηση της φυλογένειας εξάγεται από το δέντρο που μεγιστοποιεί την εν λόγω ποσότητα. Στο επόμενο κεφάλαιο θα αναλυθεί εκτενώς η συγκεκριμένη μέθοδος.

2.3.2.3 Η μέθοδος της μπεϊσιανής ανάλυσης

Στην μπεϊσιανή στατιστική ανάλυση(Bayesian phylogenetic inference) χρησιμοποιείται η συνάρτηση πιθανοφάνειας και συνήθως τα ίδια μοντέλα εξελικτικών μεταλλάξεων όπως και στην μέθοδο μέγιστης πιθανοφάνειας. Διαφέρει ωστόσο, τόσο σε θεωρητικό επίπεδο όσο και σε επίπεδο εφαρμογών. Η μπεϊσιανή ανάλυση χρησιμοποιεί το θεώρημα του Bayes, το οποίο σχετίζει την εκ των υστέρων (posterior) πιθανότητα ενός δέντρου με την εκ των προτέρων (prior) πιθανότητα του δέντρου και του μοντέλου αντικατάστασης. Σε αντίθεση όμως με τις μεθόδους μέγιστης φειδωλότητας και μέγιστης πιθανοφάνειας, η μπεϊσιανή ανάλυση δεν παράγει ένα δέντρο ή ένα σύνολο από δέντρα. Η συγκεκριμένη ανάλυση χρησιμοποιεί την πιθανοφάνεια των δέντρων σε μια Markov Chain Monte Carlo προσομοίωση ώστε να δειγματοληπτηθούν δέντρα ανάλογα με τον βαθμό πιθανοφάνειας τους, παράγωντας έτσι ένα αξιόπιστο δείγμα από δέντρα. Markov Chain Monte Carlo μέθοδοι είναι μια κατηγορία αλγορίθμων που χρησιμοποιούνται για δειγματολήπτηση από κατανομές πιθανότητας βασισμένοι στην κατασκευή μιας Μαρκοβιανής αλυσίδας (Markov chain) που έχει μια επιθυμητή κατανομή.

2.4 Το πρόγραμμα RAxML και ο τρόπος λειτουργίας του

Το πρόγραμμα RAxML (Randomized Axelerated Maximum Likelihood) χρησιμοποιείται ευρέως για την διεξαγωγή μεγάλης κλίμακας φυλογενετικών αναλύσεων εκτελώντας την μέθοδο μέγιστης πιθανοφάνειας.
Το συγκεκριμένο πρόγραμμα ξεκινάει την εκτέλεση του δημιουργώντας ένα αρχικό φειδωλό δέντρο χρησιμοποιώντας το πρόγραμμα dnapars του πακέτου PHYLIP. Έπειτα, ξεκινάει μια ακολουθία αναδιατάξεων στα υποδέντρα. Αν κάποια από τις αναδιατάξεις δημιουργήσει δέντρο με μεγαλύτερο βαθμό πιθανοφάνειας τότε το δέντρο αυτό χρησιμοποιείται ως δέντρο αναφοράς για τις επόμενες αναδιατάξεις ξεκινώντας πάλι την διαδικασία αναδιατάξεων από την αρχή. Επίσης, το πρόγραμμα RAxML βελτιστοποιεί τα μήκη των κλαδιών του δέντρου και έπειτα υπολογίζει το βαθμό πιθανοφάνειας.
Σε κάθε αναδιάταξη, κρατιέται μια λίστα με τα 20 καλύτερα δέντρα. Μετά από κάθε βήμα αναδιάταξης πραγματοποιείται η διαδικασία βελτιστοποίησης των μηκών των κλαδιών για τις 20 αυτές τοπολογίες. Οι αναδιατάξεις και οι βελτιστοποιήσεις των μηκών των κλαδιών επαναλαμβάνονται μέχρι να καλυφθούν ορισμένα κριτήρια σύγκλισης, οπότε και σταματάει η διαδικασία. Τότε, τα 20 καλύτερα δέντρα που έχουν προκύψει συνδυάζονται μεταξύ τους χρησιμοποιώντας το πρόγραμμα Consense του πακέτου PHYLIP και προκύπτει το τελικό δέντρο με την μέγιστη πιθανοφάνεια.
Όπως κάθε άλλο πρόγραμμα που χρησιμοποιείται σε φυλογενετικές αναλύσεις και εκτελεί την ίδια μέθοδο, έτσι και στο RAxML, το μεγαλύτερο ποσοστό του συνολικού χρόνου εκτέλεσης καταναλώνεται στον υπολογισμό του βαθμού πιθανοφάνειας μιας δοσμένης τοπολογίας δέντρου. Το συγκεκριμένο πρόγραμμα καταναλώνει περίπου το 95% του χρόνου εκτέλεσης για τον υπολογισμό του βαθμού πιθανοφάνειας για αρκετές χιλιάδες εναλλακτικές τοπολογίες δέντρων. Το ίδιο ισχύει και για άλλα προγράμματα όπως τα IQPNNI, PHYML, GARLI. Επιπλέον, υλοποιήσεις μπεϊσιανής φυλογενετικής ανάλυσης, όπως το πρόγραμμα MrBayes, αντιμετωπίζουν τις ίδιες υπολογιστικές προκλίσεις καθώς βασίζονται επίσης στον υπολογισμό της συνάρτησης φυλογενετικής πιθανοφάνειας.

Εξαιτίας, της σημαντικής προόδου που έχει πραγματοποιηθεί σε πειραματικές τεχνικές τα τελευταία χρόνια, η ανάπτυξη μοριακών δεδομένων στην GeneBank πραγματοποιείται με επιταχυνόμενους ρυθμούς. Ήδη υπάρχουν σετ δεδομένων που χρειάζονται πάνω από 2 000 000 ώρες CPU time σε supercomputers όπως το IBM BlueGene/L για να αναλυθούν.
Το πρόβλημα που παρουσιάζεται σχετικά με τον υπολογισμό της συνάρτησης φυλογενετικής πιθανοφάνειας, η οποία είναι μια από τις σημαντικότερες συναρτήσεις στον τομέα των Bioinformatics, είναι στην ουσία το ίδιο, για όλες τις προαναφερθείσες υλοποιήσεις, είτε προσανατολίζονται στην υλοποίηση της συνάρτηση μέγιστης πιθανοφάνειας είτε στην πραγματοποίηση φυλογενετικής μπεϊσιανης ανάλυσης, καθιστώντας επιτακτική την ανάγκη για δημιουργία υλοποιήσεων σε επίπεδο hardware προκειμένου να επιταχυνθεί η συγκεκριμένη συνάρτηση.

3

Η μέθοδος της Μέγιστης Πιθανοφάνειας

Το Κεφάλαιο αυτό περιγράφει την μέθοδο της μέγιστης πιθανοφάνειας. Επίσης παρουσιάζονται τα σημαντικότερα μοντέλα εξέλιξης των αλληλουχιών DNA καθώς και η συνάρτηση φυλογενετικής πιθανοφάνειας.

3.1 Μέγιστη Πιθανοφάνεια

Η μέθοδος μέγιστης πιθανοφάνειας όταν εφαρμόζεται σε μια φυλογενετική ανάλυση υπολογίζει μια υπόθεση σχετικά με την εξελικτική ιστορία των ειδών υπο μελέτη, δοθείσης της πιθανότητας ότι ένα προτεινόμενο μοντέλο σχετικά με την εξελικτική διαδικασία σε συνδυασμό με μια υποτιθέμενη ιστορική πορεία θα δημιουργούσαν τα παρατηρημένα δεδομένα. Εικάζεται ότι η ιστορική πορεία με την υψηλότερη πιθανότητα να δημιουργήσει τα πραγματικά δεδομένα είναι προτιμότερη από οποιαδήποτε άλλη με χαμηλότερη πιθανότητα. Η πρώτη εφαρμογή μεθόδων μέγιστης πιθανοφάνειας σε φυλογενετικές αναλύσεις πραγματοποιήθηκε από τους Cavalli-Sforza και Edwards (1967). Ο Felsenstein(1981, 1993) ανέπτυξε περαιτέρω την συγκεκριμένη μέθοδο επεκτείνωντας την σε νουκλεοτιδικό επίπεδο. Μερικά χρόνια αργότερα η μέθοδος της μέγιστης πιθανοφάνειας εφαρμόστηκε και σε αμινοξικές ακολουθίες από τους Kishino et. al (1990) και Adachi και Hasegawa(1992) .
Εκτός από το ότι η συγκεκριμένη μέθοδος πληρεί τις ιδιότητες της συνέπειας, επίσης παράγει προσεγγίσεις οι οποίες έχουν μικρότερη διακύμανση από άλλες μεθόδους. Γενικότερα, η μέθοδος της μέγιστης πιθανοφάνειας θεωρείται ότι επηρεάζεται λιγότερο από κάθε άλλη μέθοδο από λάθη τόσο σε επίπεδο δειγματοληψίας των ακολουθιών DNA που χρησιμοποιούνται όσο και από παραβιάσεις των υποθέσεων που χρησιμοποιούνται στα εξελικτικά μοντέλα. Η ιδιότητα αυτή προκύπτει κυρίως από το γεγονός ότι πολλά από τα μοντέλα εξέλιξης που χρησιμοποιούνται υποθέτουν ταυτόσημες κατανομές κατά μήκος κάθε θέσης των ακολουθιών DNA και έτσι μπορεί να γίνει η υπόθεση ότι οι πραγματικές μεταλλακτικές διαδικασίες που πραγματοποιήθηκαν σε κάθε θέση έχουν πολλά κοινά αν και δεν είναι απόλυτα όμοιες.
Πολλοί τομείς της βιολογίας σε ερευνητικό επίπεδο, όπως η γονιδιακή χαρτογράφηση (genetic mapping) , χρησιμοποιούν μεθόδους μέγιστης πιθανοφάνειας σε συστηματική βάση για έλεγχο υποθέσεων. Παρά ταύτα, η πραγματική πολυπλοκότητα εκτέλεσης των συγκεκριμένων μεθόδων, όταν υπάρχουν πολλές διαφορετικές υποθέσεις προς εξέταση έχει αναστείλει την γενική χρήση τους.

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, σκοπός μιας φυλογενετικής ανάλυσης είναι να υπολογιστεί μια πιθανή ιστορική πορεία ή ένα σύνολο από ιστορικές πορείες οι οποίες είναι οι πιο συνεπής με ένα σύνολο πραγματικών δεδομένων. Τα πραγματικά δεδομένα στην συγκεκριμένη περίπτωση είναι νουκλεοτιδικές ή αμινοξικές ακολουθίες ενώ τα ζητούμενα είναι η διάταξη των κλαδιών καθώς και το μήκος τους.
Για να εφαρμοστεί η μέθοδος της μέγιστης πιθανοφάνειας πρέπει να προσδιοριστεί κάποιο εξελικτικό μοντέλο το οποίο θα εξηγεί την μετάλλαξη από μια ακολουθία σε μια άλλη. Η μέθοδος, στην συνέχεια, υπολογίζει την πιθανότητα το επιλεγμένο εξελικτικό μοντέλο να έχει δημιουργήσει τα πραγματικά δεδομένα. Μια φυλογένεια δημιουργείται επιλέγωντας τα φυλογενετικά δέντρα που έχουν τον υψηλότερο βαθμό πιθανοφάνειας.

3.2 Μοντέλα Εξέλιξης των Αλληλουχιών

Αναμφίβολα, το DNA αποτελεί ένα καλό πληροφοριακό μόριο για τη μελέτη της εξέλιξης για τον λόγο ότι στο επίπεδο αυτό μπορούν να μελετηθούν οι γενετικές διαφορές όχι μόνο των περιοχών κωδικοποίησης αλλά και των περιοχών μη κωδικοποίησης της πληροφορίας όπως λόγου χάρη οι διαγωνιδιακές περιοχές ή οι σιωπηρές νουκλεοτιδικές υποκαταστάσεις. Η μελέτη της νουκλεοτιδικής αλληλουχίας μπορεί να δώσει λεπτομερείς πληροφορίες για τους μηχανισμούς των απαλοιφών, των ενθέσεων, του άνισου διασκελισμού, της μεταθεσιμότητας των γονιδίων, της γονιδιακής μετατροπής και της οριζόντιας μεταβίβασης γονιδίων. Η διαδικασία της νουκλεοτιδικής υποκατάστασης σε έναν πληθυσμό είναι γενικά τόσο αργή ώστε να μην μπορεί να παρατηρηθεί στη διάρκεια ζωής του ερευνητή. Για αυτόν τον λόγο, οι εξελικτικές αλλαγές στις αλληλουχίες DNA αποκαλύπτονται με σύγκριση μεταξύ ομάδων με κοινό πρόγονο στο εξελικτικό παρελθόν. Τέτοιες συγκρίσεις απαιτούν τη χρησιμοποίηση στατιστικών μεθόδων πολλές από τις οποίες αναπτύχθηκαν πρόσφατα και έδωσαν νέες αντιλήψεις για την εξελικτική διαδικασία.
Για τη μελέτη της δυναμικής της νουκλεοτιδικής υποκατάστασης πρέπει να γίνουν αρκετές παραδοχές σχετικά με την πιθανότητα υποκατάστασης ενός νουκλεοτιδίου από ένα άλλο. Η μελέτη περιορίζεται υποθέτωντας Μαρκοβιανά μοντέλα (Markov Models), σύμφωνα με τα οποία η αλλαγή από την κατάσταση i στην κατάσταση j σε ένα συγκεκριμένο κλαδί του δέντρου δεν εξαρτάται από προηγούμενες καταστάσεις πριν από την κατάσταση i. Για παράδειγμα, αν σε κάποια θέση i μιας ακολουθίας υπάρχει η πρωτεϊνική βάση αδενίνη (A) κάποια χρονική στιγμή t_0, η πιθανότητα ότι θα υπάρχει η βάση θυμίνη (T) στην συγκεκριμένη θέση κάποια επακόλουθη χρονική στιγμή t_1 εξαρτάται μόνο από το γεγονός ότι υπάρχει A την χρονική στιγμη t_0. Μια ακόμη υπόθεση που ισχύει είναι ότι οι πιθανότητες υποκατάστασης είναι οι ίδιες σε όλο το δέντρο, δηλαδή θεωρείται ότι οι εξελικτικοί μηχανισμοί που είναι υπεύθυνοι για τις υποκαταστάσεις αλλάζουν σύμφωνα με μια ομογενή Μαρκοβιανή διαδικασία (homogeneous Markov Process).

Η μαθηματική έκφραση ενός μοντέλου υποκατάστασης είναι ένας πίνακας απο ρυθμούς σύμφωνα με τους οποίους κάθε νουκλεοτίδιο αντικαθίσταται από κάθε άλλο. Στην περίπτωση μελέτης ακολουθιών DNA ο συγκεκριμένος πίνακας είναι μεγέθους 4 x 4 και συνηθίζεται να συμβολίζεται με το γράμμα Q. Στον πίνακα Q κάθε στοιχείο Q_{ij} αναπαριστά τον ρυθμό υποκατάστασης μιας πρωτεινικής βάσης i σε μια άλλη πρωτεινική βάση j κατά τη διάρκεια ενός απειροελάχιστου χρονικού διαστήματος dt. Η πιο γενική μορφή του πίνακα των ρυθμών υποκατάστασης Q είναι :

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_C + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_G + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

Στον παραπάνω πίνακα οι γραμμές και οι στήλες αντιστοιχούν στις πρωτεινικές βάσεις αδενίνη (A) , κυτοσίνη (C) , γουανίνη (G) και θυμίνη (T) αντίστοιχα. Ο παράγοντας μ ισούται με τη μέση στιγμιαία τιμή των ρυθμών υποκατάστασης και εξαρτάται από τις παραμέτρους a, b, c, d, e, f, g, h, i, j, k, l οι οποίες αντιστοιχούν σε κάθε δυνατόν μετασχηματισμό από μια βάση σε μια άλλη. Το γινόμενο κάθε μιας από τις παραμέτρους a, b, c, .. l με τη μέση στιγμιαία τιμή μ δίνουν τις παραμέτρους των ρυθμών υποκατάστασης(rate parameters). Οι παράμετροι π_A, π_C, π_G και π_T είναι οι συχνότητες των βάσεων A,C,G και T αντίστοιχα. Όσο αφορά τις συγκεκριμένες παραμέτρους ισχύει η υπόθεση ότι παραμένουν σταθερές με την πάροδο του χρόνου. Τα στοιχεία της διαγωνίου του πίνακα επιλέγονται έτσι ώστε το άθροισμα των στοιχείων κάθε γραμμής να ισούται με 0. Ανάλογοι πίνακες ρυθμών υποκατάστασης μπορούν να οριστούν και για αμινοξικές ακολουθίες και είναι μεγέθους 20 x 20.
Κάθε μοντέλο υποκατάστασης που χρησιμοποιείται σήμερα αποτελεί ειδική περίπτωση του παραπάνω πίνακα Q. Γενικά ισχύει η υπόθεση ότι ο συνολικός ρυθμός αλλαγών από μια βάση i σε μια βάση j σε ένα δεδομένο χρονικό διάστημα είναι ο ίδιος με τον ρυθμό αλλαγών από μια βάση j σε μια βάση i. Εξελικτικά μοντέλα που «υπακούουν» στην συγκεκριμένη υπόθεση ονομάζονται χρονικά αντιστρεπτά(time-reversible). Σε μαθηματικό επίπεδο η συγκεκριμένη υπόθεση μεταφράζεται στις ακόλουθες ισότητες που πρέπει να ισχύουν για τις παραμέτρους a, b, c, .. l : g=a, h=b, i=c, j=d, k=e και l=f.
Η πιο γενική μορφή του πίνακα των ρυθμών υποκατάστασης Q για το γενικό χρονικά αντιστρεπτο μοντέλο (GTR) είναι :

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_C + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_G + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{pmatrix}$$

Ένα σημαντικό επακόλουθο της υπόθεσης για χρονική αντιστρεψιμότητα είναι το γεγονός ότι ο βαθμός πιθανοφάνειας ένος δέντρου δεν εξαρτάται από την ρίζα του δέντρου. Ως συνέπεια, η μέθοδος μέγιστης πιθανοφάνειας συνηθίζεται να χρησιμοποιείται για υπολογισμό δέντρων χωρίς ρίζα και χρειάζονται κάποιες επιπλέον υποθέσεις ώστε να μετατραπεί το δέντρο χωρίς ρίζα σε δέντρο με ρίζα. Παρόλο που είναι δυνατόν να χαλαρωθούν οι υποθέσεις σχετικά με την χρονική αντιστρεψιμότητα, μια τέτοια χαλάρωση θα εισήγαγε σημαντική υπολογιστική πολυπλοκότητα συμπεριλαμβάνοντας την ανάγκη για υπολογισμό και δέντρων με ρίζα.
Τα περισσότερα από τα μοντέλα που χρησιμοποιούνται σήμερα για τον εντοπισμό του δέντρου με την μέγιστη πιθανοφάνεια προκύπτουν από περιορισμούς στις παραμέτρους του πίνακα Q.
Αν για παράδειγμα οι τύποι υποκατάστασης διαιρεθούν σε μεταστροφές(transversions), όταν η πουρίνη (A ή G) αντικαθίσταται από πυριμιδίνη (C ή T) ή το αντίστροφο, σε μεταπτώσεις(transitions) μεταξύ πουρινών και σε μεταπτώσεις μεταξύ πυριμιδινών τότε προκύπτει το μοντέλο Tamura και Nei (1993;TrN) θέτωντας a=c=d=f.
Το μοντέλο Kimura(1981;K3ST) θεωρεί ίσες συχνότητες για κάθε πρωτεινική βάση, δλδ. $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$ και διαιρεί τους τύπους υποκατάστασης σε μεταπτώσεις με b=e, μεταστροφές A $\leftrightarrow$T ή C $\leftrightarrow$G με c=d και μεταστροφές A $\leftrightarrow$C ή G $\leftrightarrow$T με a=f.
Άλλη μια παραλαγή του μοντέλου GTR είναι το Zharkish (1994;SYM) το οποίο είναι ισοδύναμο με το μοντέλο GTR αλλά θεωρεί επιπλέον ίσες συχνότητες βάσεων.
Το μοντέλο Jukes και Cantor (1969;JC) θεωρεί ίσες συχνότητες βάσεων ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$) σε συνδυασμό με ίσους ρυθμούς υποκαταστάσεων θέτωντας a=b=c=d=e=f=1 και έτσι ο πίνακας Q διαμορφώνεται ως εξής:

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}$$

Επειδή οι μεταπτώσεις είναι πιο συχνές από ότι οι μεταστροφές το μοντέλο των Jukes και Cantor δεν θεωρείται ρεαλιστικό. Ο Kimura(1980;K2P) πρότεινε ένα μοντέλο με δύο ρυθμούς υποκατάστασης ανά μονάδα χρόνου, έναν ρυθμό α για τις μεταπτώσεις και έναν ρυθμό β για τις μεταστροφές διατηρώντας και πάλι ίσες συχνότητες για τις βάσεις. Θέτωντας λοιπόν a=c=d=f=1 και b=e=κ προκύπτει:

$$Q=\begin{pmatrix} -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) \end{pmatrix}$$

Αν θεωρηθεί ο ρυθμός μεταπτώσεων α=μκ/4 και ο ρυθμός μεταστροφών β=μ/4 τότε ο παραπάνω πίνακας απλοποιείται ως εξής:

$$Q=\begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \alpha & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix}$$

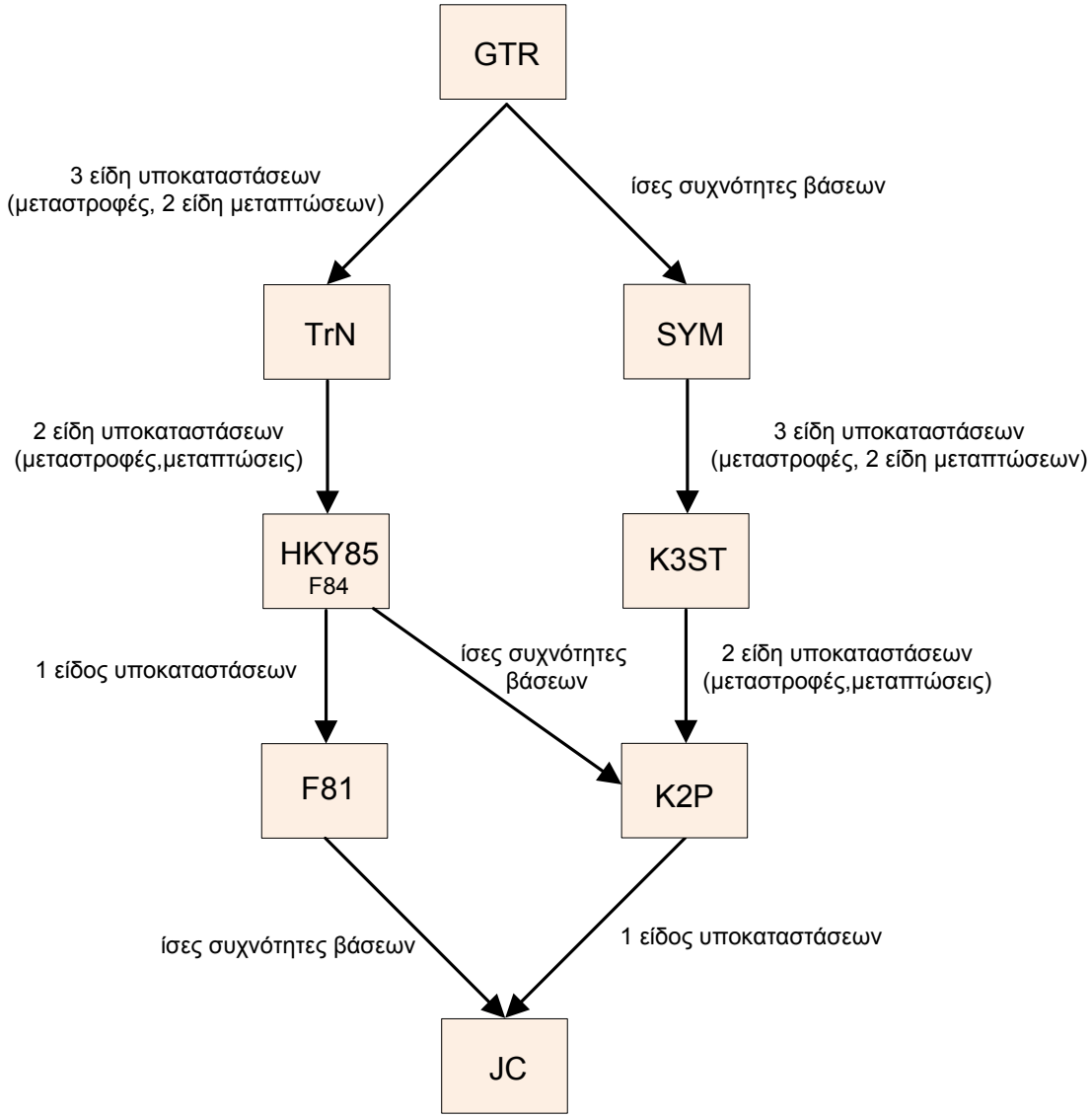
Ο λόγος κ=α/β εκφράζει την τάση μεταπτώσεων (transition bias). Όταν κ=1 προφανώς α=β και το μοντέλο Kimura ταυτίζεται με το μοντέλο Jukes Cantor.
Το μοντέλο K2P μπορεί να γενικευτεί επιτρέποντας άνισες συχνότητες για τις βάσεις (Hasegawa et. al, 1985). Στην περίπτωση αυτή ο πίνακας Q για το μοντέλο που προκύπτει (HKY85) είναι:

$$Q=\begin{pmatrix} -\mu(\kappa\pi_G+\pi_Y) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\kappa\pi_T+\pi_R) & \mu\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A+\pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\kappa\pi_C+\pi_R) \end{pmatrix}$$

όπου α=μ, β=μκ, $\pi_R = \pi_A + \pi_G$ και $\pi_Y = \pi_C + \pi_T$. Ο παραπάνω πίνακας ανταποκρίνεται στο μοντέλο GTR με περιορισμούς a=c=d=f=1 και b=e=κ.
Ακόμη, το μοντέλο JC μπορεί να γενικευτεί και να υποστηρίζει και αυτό διαφορετικές συχνότητες βάσεων (Falsenstein 1981;F81) θέτωντας κ=1 στον παραπάνω πίνακα δημιουργώντας έτσι τον ακόλουθο πίνακα:

$$Q=\begin{pmatrix} -\mu(\pi_G+\pi_Y) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_T+\pi_R) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_A+\pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_C+\pi_R) \end{pmatrix}$$

Στην εικόνα 3-1 φαίνονται οι σχέσεις μεταξύ των μοντέλων της οικογένειας των χρονικά αντιστρεπτών μοντέλων υποκατάστασης(time-reversible family of substitution models). Τα βέλη δείχνουν την μετατροπή από ένα πιο γενικό σε ένα πιο περιορισμένο μοντέλο.



Εικόνα 3-1 : Η ιεραρχία των μοντέλων υποκατάστασης των αλληλουχιών DNA.

3.3 Υπολογισμός των πιθανοτήτων υποκατάστασης

Ο πίνακας των ρυθμών υποκατάστασης Q προσδιορίζει τους ρυθμούς αλλαγής μεταξύ νουκλεοτιδίων για μια απειροελάχιστη χρονική στιγμή dt. Για να υπολογιστεί η πιθανοφάνεια ένος δέντρου χρειάζονται οι πιθανότητες αλλαγής από κάθε δυνατή κατάσταση σε οποιαδήποτε άλλη κατά μήκος ενός κλαδιού μήκους t. Ο πίνακας πιθανοτήτων υποκατάστασης δίνεται από τον τύπο: $P(t) = e^{Qt}$. Ο υπολογισμός του εκθετικού μπορεί να πραγματοποιηθεί αποσυνθέτωντας τον πίνακα ρυθμών υποκατάστασης Q στις ιδιοτιμές(eigenvalues) και τα ιδιοδιανύσματα(eigenvectors) του. Για αρκετά μοντέλα υπάρχουν απλές εκφράσεις για τις ιδιοτιμές τους, επιτρέπωντας άμεσο αναλυτικό υπολογισμό των στοιχείων του πίνακα P. Για παράδειγμα, για το μοντέλο K2P όσο αφορά τις υποκαταστάσεις πρωτεϊνικών βάσεων, υπάρχουν μόνο τρείς διαφορετικές τιμές πιθανοτήτων: η πιθανότητα υποκατάστασης τύπου μεταστροφής, η πιθανότητα υποκατάστασης τύπου μετάπτωσης και η πιθανότητα καμμίας υποκατάστασης.

Οι συγκεκριμένες πιθανότητες είναι:

$$\text{K2P} \quad : \quad \mathrm{P_{ij}(t)} = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-\mu t} + \frac{1}{2}e^{-\mu t\frac{(\kappa+1)}{2}} & \text{για i = j(όχι υποκατάσταση)} \\ \frac{1}{4} + \frac{1}{4}e^{-\mu t} - \frac{1}{2}e^{-\mu t\frac{(\kappa+1)}{2}} & \text{για i} \neq \text{j(μετάπτωση)} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \text{για i} \neq \text{j(μεταστροφή)} \end{cases}$$

Θέτωντας τις παραπάνω τιμές πιθανοτήτων A,B,C αντίστοιχα προκύπτει ο συνολικός πίνακας πιθανοτήτων P:

$$P(t) = \begin{pmatrix} A & C & B & C \\ C & A & C & B \\ B & C & A & C \\ C & B & C & A \end{pmatrix}$$

Στη συνέχεια φαίνονται οι πιθανότητες υποκατάστασης για μερικά DNA μοντέλα υποκατάστασης.

$$\text{JC} \quad : \quad \mathrm{P_{ij}(t)} = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \text{για i = j} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \text{για i} \neq \text{j} \end{cases}$$

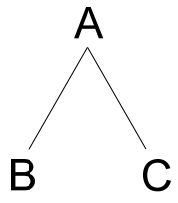
$$\text{F81} \quad : \quad P_{ij}(t) = \begin{cases} \pi_j + (1-\pi_j)e^{-\mu t} & \text{για i = j} \\ \pi_j(1-e^{-\mu t}) & \text{για i} \neq \text{j} \end{cases}$$

$$\text{HKY85,F84} \quad : \quad P_{ij}(t) = \begin{cases} \pi_j + \pi_j(\frac{1}{\Pi_j} - 1)\, e^{-\mu t} + (\frac{\Pi_j - \pi_j}{\Pi_j})\, e^{-\mu t A} & \text{για i = j} \\ \pi_j + \pi_j(\frac{1}{\Pi_j} - 1)\, e^{-\mu t} - (\frac{\pi_j}{\Pi_j})\, e^{-\mu t A} & \text{για i} \neq \text{j(μετάπτωση)} \\ \pi_j(1-e^{-\mu t}) & \text{για i} \neq \text{j(μεταστροφή)} \end{cases}$$

όπου $A=1+\Pi_j\,(\kappa-1)$ για το μοντέλο HKY85 και $A=K+1$ για το μοντέλο F84, και $\Pi_j = \pi_A + \pi_G$ αν η βάση j είναι πουρίνη(A ή G) και $\Pi_j = \pi_C + \pi_T$ αν η βάση j είναι πυριμιδίνη(C ή T).

3.3 Υπολογισμός της πιθανοφάνειας δέντρου

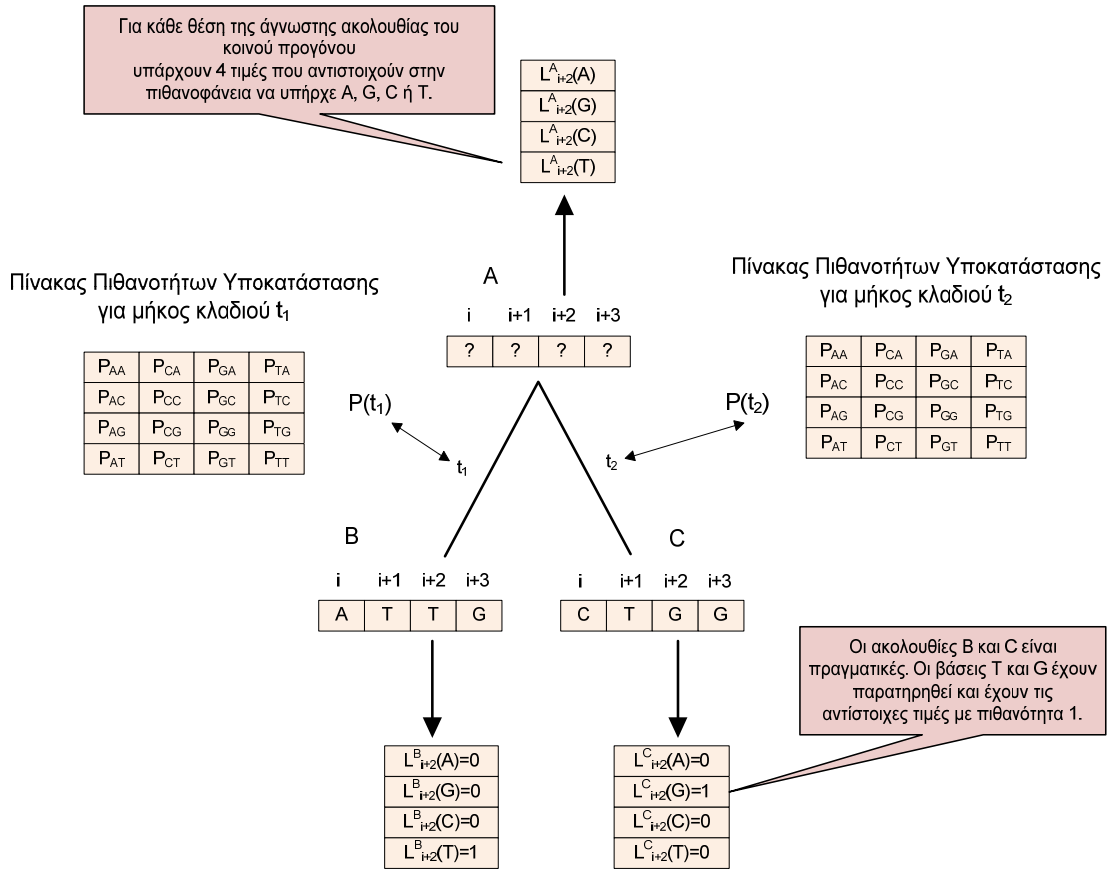
Για να υπολογιστεί η πιθανοφάνεια ενός δέντρου είναι απαραίτητο να εκφραστούν οι πιθανοφάνειες εμφάνισης της κάθε κατάστασης σε κάθε κόμβο του δέντρου ως συνάρτηση της τοπολογίας και του μήκους των κλαδιών. Όπως γίνεται και με άλλες μεθόδους οι οποίες προσδιορίζουν το βέλτιστο δέντρο βάση ενός κριτηρίου, τώρα θα υποτεθεί ότι το δέντρο είναι δοσμένο και σκοπός είναι να αξιολογηθεί πόσο καλό είναι.
Η μέθοδος για τον υπολογισμό του βαθμού πιθανοφάνειας ενός δέντρου ξεκινάει από μια υποθετική ρίζα η οποία τοποθετείται σε κάποιο σημείο του δέντρου και υπολογίζονται οι πιθανοφάνειες όλων των υποδέντρων-παιδιών της ρίζας. Για χρονικά αντιστρεπτά μοντέλα υποκατάστασης η επιλογή της θέσης της ρίζας δεν επηρεάζει την πιθανοφάνεια του δέντρου. Έστω A, B, C τρία είδη με A τον κοινό πρόγονο των B και C. Η σχέση των A,B,C φαίνεται στην επόμενη εικόνα:



Η υπο συνθήκη πιθανοφάνεια της κατάστασης i (όπου i οι πρωτεϊνικές βάσεις A,G,C και T) σε μια θέση j της ακολουθίας A είναι:

$$L(x_{Aj} = i) = [\sum_k P_{ik}(u_{AB})\, L(x_{Bj} = k)][\sum_l P_{il}(u_{AC})\, L(x_{Cj} = l)]$$ όπου u_{XY} είναι το μήκος του κλαδιού που συνδέει τις ακολουθίες X και Y.

Ο όρος υπο συνθήκη χρησιμοποιείται καθώς η παραπάνω ποσότητα εκφράζει την πιθανοφάνεια του υποδέντρου που αποτελεί απόγονο του κόμβου A δοθέντος μιας κατάστασης i για τον κόμβο A. Με άλλα λόγια, η υπο συνθήκη πιθανοφάνεια ότι ο κόμβος A έχει μια κατάσταση i είναι το γινόμενο των πιθανοφανειών η κατάσταση i να δημιούργησε τις ακολουθίες B και C. Ο πρώτος όρος του γινομένου είναι το άθροισμα της πιθανότητας η κατάσταση i να αλλάξει σε κατάσταση k στο χρονικό διάστημα u_{AB} , P_{ik} (u_{AB}), επί την πιθανοφάνεια ότι η ακολουθία B έχει την κατάσταση k στην θέση υπο μελέτη, για όλες τις τιμές του k. Αν η ακολουθία B είναι γνωστή, τότε η πιθανότητα ότι η θέση j έχει την κατάσταση k είναι 1 αν η k ισούται με την παρατηρούμενη κατάσταση στην ακολουθία αλλιώς 0. Η επόμενη εικόνα εξηγεί τα παραπάνω.



$$L^A_{i+2}(A) = [P_{AA}(t_1)L^B_{i+2}(A)+P_{AC}(t_1)L^B_{i+2}(C)+P_{AG}(t_1)L^B_{i+2}(G)+P_{AT}(t_1)L^B_{i+2}(T)] \times [P_{AA}(t_2)L^C_{i+2}(A)+P_{AC}(t_2)L^C_{i+2}(C)+P_{AG}(t_2)L^C_{i+2}(G)+P_{AT}(t_2)L^C_{i+2}(T)]$$

$$L^A_{i+2}(C) = [P_{CA}(t_1)L^B_{i+2}(A)+P_{CC}(t_1)L^B_{i+2}(C)+P_{CG}(t_1)L^B_{i+2}(G)+P_{CT}(t_1)L^B_{i+2}(T)] \times [P_{CA}(t_2)L^C_{i+2}(A)+P_{CC}(t_2)L^C_{i+2}(C)+P_{CG}(t_2)L^C_{i+2}(G)+P_{CT}(t_2)L^C_{i+2}(T)]$$

$$L^A_{i+2}(G) = [P_{GA}(t_1)L^B_{i+2}(A)+P_{GC}(t_1)L^B_{i+2}(C)+P_{GG}(t_1)L^B_{i+2}(G)+P_{GT}(t_1)L^B_{i+2}(T)] \times [P_{GA}(t_2)L^C_{i+2}(A)+P_{GC}(t_2)L^C_{i+2}(C)+P_{GG}(t_2)L^C_{i+2}(G)+P_{GT}(t_2)L^C_{i+2}(T)]$$

$$L^A_{i+2}(T) = [P_{TA}(t_1)L^B_{i+2}(A)+P_{TC}(t_1)L^B_{i+2}(C)+P_{TG}(t_1)L^B_{i+2}(G)+P_{TT}(t_1)L^B_{i+2}(T)] \times [P_{TA}(t_2)L^C_{i+2}(A)+P_{TC}(t_2)L^C_{i+2}(C)+P_{TG}(t_2)L^C_{i+2}(G)+P_{TT}(t_2)L^C_{i+2}(T)]$$

Εικόνα 3-2 : Τρόπος υπολογισμού του διανύσματος πιθανοφάνειας για μια τυχαία θέση προγόνου.

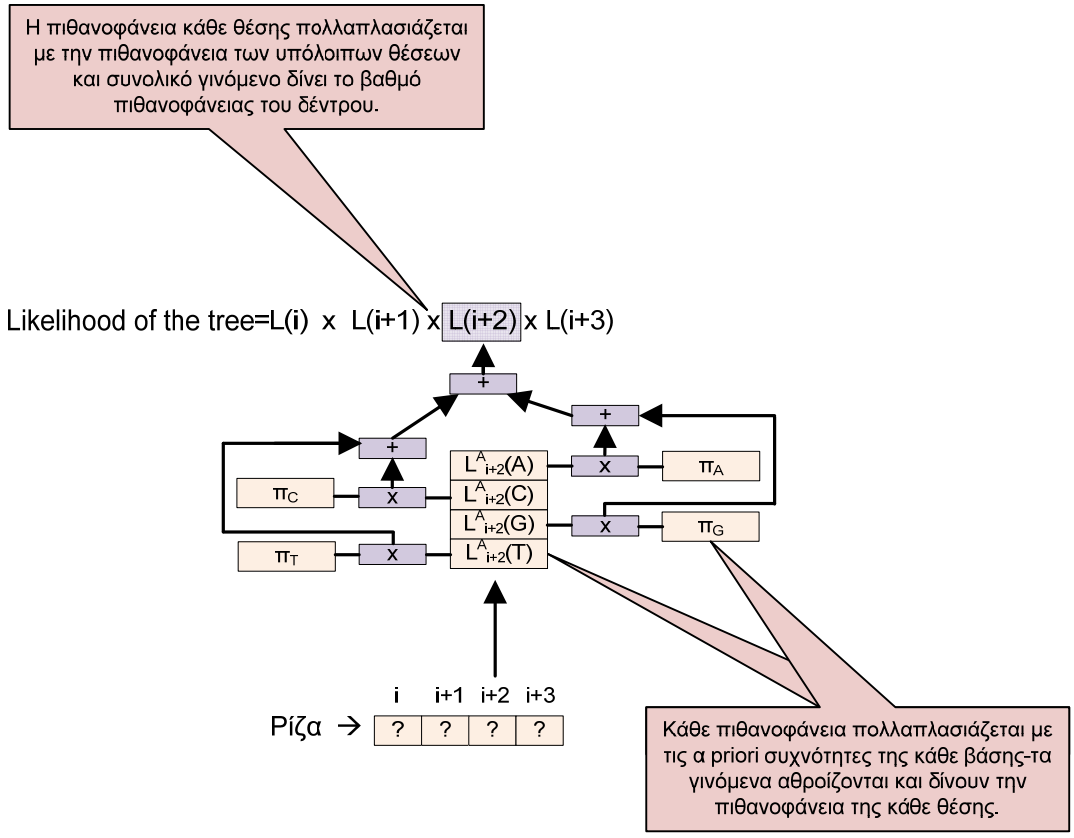
Αν ο κόμβος Β για παράδειγμα είναι πρόγονος άλλων ειδών τότε υπολογίζονται αναδρομικά οι πιθανοφάνειες για κάθε θέση της άγνωστης ακολουθίας του Β. Με αυτόν τον τρόπο, ξεκινώντας από την ρίζα και εκτελώντας postorder αναδρομή υπολογίζονται οι τιμές πιθανοφάνειας για κάθε θέση των άγνωστων ακολουθιών που υπάρχουν στους εσωτερικούς κόμβους, μέχρι να υπολογιστούν οι τιμές της ρίζας του δέντρου.
Μόλις τελειώσει η παραπάνω αναδρομική ακολουθία υπολογισμών υπολογίζονται οι πιθανοφάνειες για κάθε θέση j της ακολουθίας της ρίζας από τον τύπο:

$$L(j) = \sum_{k} \pi_k \, L(x^{root}_{j} = k) \quad \text{με } k=A,C,G,T$$

Για κάθε θέση δηλαδή υπολογίζεται το άθροισμα των συχνοτήτων εμφάνισης κάθε βάσης επί την υπο συνθήκη πιθανοφάνεια της αντίστοιχης βάσης για όλες τις περιπτώσεις βάσεων. Με αυτόν τον τρόπο προκύπτει μια συνολική πιθανοφάνεια για την κάθε θέση της ακολουθίας της ρίζας. Οι τιμές αυτές στην συνέχεια πολλαπλασιάζονται μεταξύ τους και προκύπτει η ζητούμενη συνολική πιθανοφάνεια του δέντρου:

$$L = \prod_{j} L(j) \quad \{ \text{ ή } L = \sum_{j} \log(L(j)) \}$$

Η επόμενη εικόνα εξηγεί τα παραπάνω.



Εικόνα 3-3 : Τρόπος υπολογισμού του βαθμού πιθανοφάνειας κάθε θέσης της ρίζας και του δέντρου.

Για την ακρίβεια, συνηθίζεται να υπολογίζεται η λογαριθμική πιθανοφάνεια του δέντρου(log-likelihood), ως το άθροισμα των επιμέρους λογαριθμικών πιθανοφανειών κάθε θέσης. Ο λόγος για τον οποίο γίνεται αυτό είναι η αποφυγή αριθμητικών υποχειλίσεων.

4

Αρχιτεκτονική για την Συνάρτηση Φυλογενετικής Πιθανοφάνειας

Το Κεφάλαιο αυτό παρουσιάζει μια νέα αρχιτεκτονική για τον υπολογισμό του βαθμού πιθανοφάνειας ενός φυλογενετικού δέντρου. Περιγράφονται λεπτομερώς τα επιμέρους υποσυστήματα της αρχιτεκτονικής και εξηγούνται σχεδιαστικές επιλογές.

4.1 Ανάλυση της Συνάρτησης Φυλογενετικής Πιθανοφάνειας

Η συνάρτηση φυλογενετικής πιθανοφάνειας εφαρμόζεται για κάθε θέση των νουκλεοτιδικών ακολουθιών δύο αδερφικών ταξινομικών λειτουργικών μονάδων σε ένα φυλογενετικό δέντρο και επαναλαμβάνεται αναδρομικά μέχρι να υπολογιστούν τα διανύσματα πιθανοφανειών για κάθε θέση των άγνωστων ακολουθιών νουκλεοτιδίων μέχρι και της ρίζας. Έπειτα ακολουθείται μια διαδικασία πράξεων για την εξαγωγή του βαθμού πιθανοφάνειας του δέντρου. Συνοψίζωντας, οι μαθηματικοί τύποι για την εφαρμογή της συνάρτησης είναι οι εξής:

$$L(x_{Aj} = i) = [\sum_{k} P_{ik}(u_{AB})\, L(x_{Bj} = k)][\sum_{l} P_{il}(u_{AC})\, L(x_{Cj} = l)] \qquad (A)$$

$$L(j) = \sum_{k} \pi_k \, L(x^{root}_{j} = k) \quad \mu\varepsilon\ k=A,C,G,T \qquad (B)$$

$$L = \prod_{j} L(j) \quad \{ \ \acute{\eta}\ L = \sum_{j} \log(L(j)) \ \} \qquad (\Gamma)$$

Ο τύπος Α εφαρμόζεται αναδρομικά σε όλο το δέντρο ενώ οι τύποι Β και Γ εφαρμόζονται μόνο στην ρίζα για την εξαγωγή του ζητούμενου βαθμού πιθανοφάνειας. Όσον αφορά τον τύπο Γ, χρησιμοποιείται συνήθως η λογαριθμική μορφή προς αποφυγή αριθμητικών υποχειλίσεων.

4.2 Η Βασική Υπολογιστική Μονάδα (basic cell)

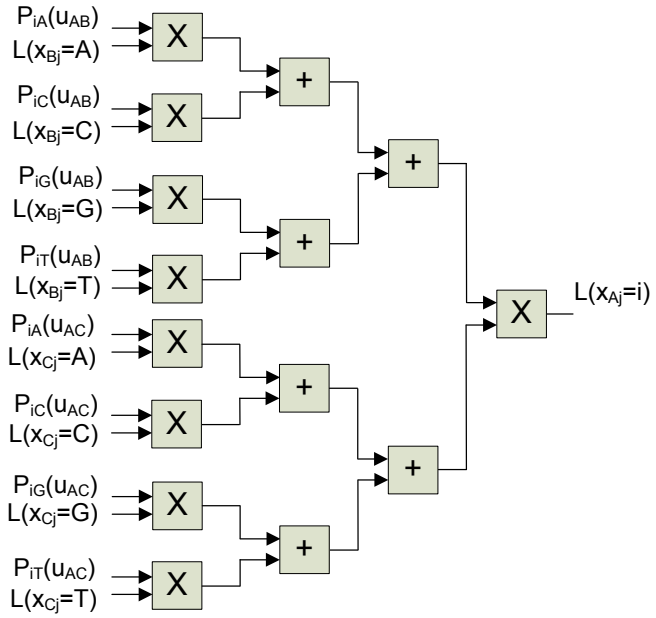
Η βασική υπολογιστική μονάδα(BC) που θα παρουσιαστεί στη συνέχεια υλοποιεί τον τύπο Α. Αν η μελέτη περιοριστεί σε πρωτεινικές ακολουθίες (DNA) τότε κάθε άθροισμα της συγκεκριμένης αριθμητικής παράστασης έχει 4 όρους οι οποιοί αντιστοιχούν στις 4 πρωτεϊνικές βάσεις A, C, G και T.

Αναπτύσσοντας τα αθροίσματα στον παραπάνω τύπο προκύπτει:

$$
\begin{aligned}
L(x_{Aj} = i) = [\, & P_{iA}(u_{AB})L(x_{Bj} = A) + P_{iC}(u_{AB})L(x_{Bj} = C) + \\
& P_{iG}(u_{AB})L(x_{Bj} = G) + P_{iT}(u_{AB})L(x_{Bj} = T)] \\
\text{x } [\, & P_{iA}(u_{AC})L(x_{Cj} = A) + P_{iC}(u_{AC})L(x_{Cj} = C) + \\
& P_{iG}(u_{AC})L(x_{Cj} = G) + P_{iT}(u_{AC})L(x_{Cj} = T)]
\end{aligned}
$$

Παρατηρείται ότι, για να υπολογιστεί κάθε μια πιθανοφάνεια της αντίστοιχης θέσης του κοινού προγόνου των δύο λειτουργικών ταξινομικών μονάδων υπο μελέτη χρειάζονται 8 πολλαπλασιασμοί (4 για κάθε μονάδα), 6 προσθέσεις (3 για κάθε μονάδα) και 1 πολλαπλασιασμός, για να συνδυαστούν οι τιμές που προέκυψαν από τις δύο μονάδες.

Στην εικόνα 4-1 φαίνεται η διάταξη της βασικής υπολογιστικής μονάδας.

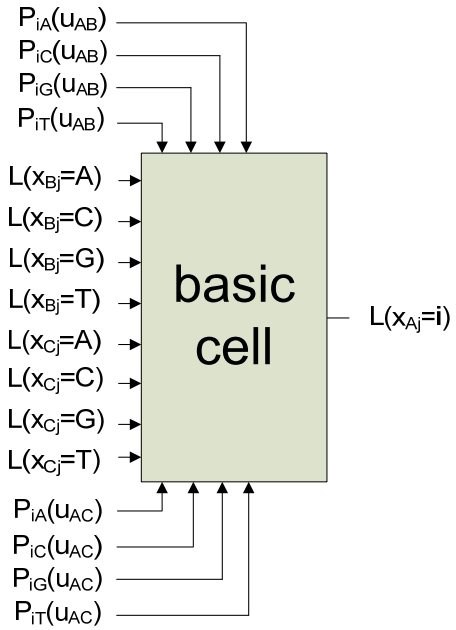


Εικόνα 4-1 : Η διάταξη πολλαπλασιαστών και αθροιστών της βασικής υπολογιστικής μονάδας.

Προφανώς, θα μπορούσε κάποιος να αναπτύξει περαιτέρω την παραπάνω αριθμητική παράσταση εκτελώντας τον πολλαπλασιασμό μεταξύ των δύο ζευγαριών αγκυλών.

Πρέπει να σημειωθεί ότι, στην περίπτωση αυτή, οι συνέπειες σε σχεδιαστικό επίπεδο θα ήταν ιδιαίτερα αρνητικές όσο αφορά την απόδοση αν αναλογιστεί κανείς ότι αντί για 9 πολλαπλασιασμούς και 6 προσθέσεις, θα χρειαζόντουσαν συνολικά 24 πολλαπλασιασμοί (8 για πολλαπλασιασμό με τις πιθανότητες P κ' 16 για συνδυασμό των προηγούμενων 8 αποτελεσμάτων μεταξύ τους) και 15 προσθέσεις. Αξίζει να τονιστεί λοιπόν ότι η σωστή επιλογή της μορφής της μαθηματικής συνάρτησης προς υλοποίηση μπορεί να συνεισφέρει σημαντικά στην τελική απόδοση της σχεδίασης. Στην συγκεκριμένη περίπτωση, η επιλογή της παραπάνω μορφής έναντι της ανεπτυγμένης, προσφέρει μείωση κατά 62% στον αριθμό των πολλαπλασιαστών που θα χρειαζόντουσαν και κατά 60% στον αριθμό των αθροιστών.
Στο πρώτο επίπεδο πράξεων (από αριστερά προς τα δεξιά) εκτελούνται παράλληλα 8 πολλαπλασιασμοί. Στο δεύτερο επίπεδο εκτελούνται παράλληλα 4 προσθέσεις ενώ στο τρίτο 2 προσθέσεις. Ο τελικός πολλαπλασιασμός χρησιμοποιείται για τον συνδυασμό των αποτελεσμάτων που προέκυψαν, ένα για κάθε μια από τις δύο λειτουργικές ταξινομικές μονάδες. Τόσο οι πολλαπλασιαστές που χρησιμοποιήθηκαν όσο και οι προσθετές είναι ομόχειροι και η αριθμητική που χρησιμοποιείται είναι floating point double precision(IEEE-754). Προσπάθειες για χρησιμοποιήση single precision αριθμητικής πραγματοποιήθηκαν από τον κ. Σταματάκη στο πρόγραμμα RAxML χωρίς επιθυμητά αποτελέσματα λόγω αριθμητικών υποχειλίσεων.

Η διεπαφή της βασικής υπολογιστικής μονάδας φαίνεται στην εικόνα 4-2.

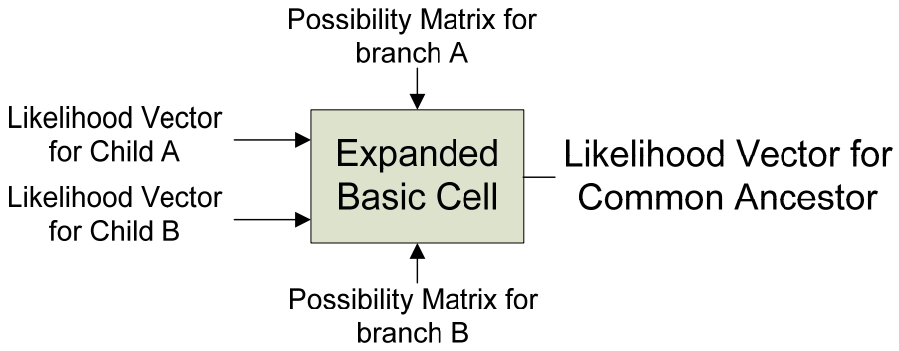


Εικόνα 4-2 : Η διεπαφή της βασικής υπολογιστικής μονάδας.

Η παραπάνω υπολογιστική μονάδα υπολογίζει την μια από τις 4 πιθανοφάνειες του ζητούμενου διανύσματος πιθανοφανειών για κάθε θέση της ακολουθίας προς υπολογισμό. Προφανώς, για τον υπολογισμό ολόκληρου του διανύσματος θα

μπορούσαν να χρησιμοποιηθούν 4 basic cells παράλληλα. Στην συγκεκριμένη αρχιτεκτονική επιλέχθηκε να χρησιμοποιηθεί ένα basic cell το οποίο να υπολογίζει το ένα μετά το άλλο τις 4 ζητούμενες πιθανοφάνειες. Η συγκεκριμένη σχεδιαστική επιλογή κρίθηκε καλύτερη από την παράλληλη τοποθέτηση $4^{ων}$ basic cells για τον λόγο ότι η χρονική καθυστέρηση που δημιουργείται από την χρησιμοποίηση του ίδιου basic cell είναι ελάχιστη σε σύγκριση με την οικονομία σε πόρους που πραγματοποιείται(DSP48E slices και block ram blocks).
Οι πολλαπλασιαστές που χρησιμοποιήθηκαν έχουν latency 15 κύκλους και throughput 1 αποτέλεσμα/κύκλο ενώ οι αθροιστές έχουν latency 14 κύκλους και επίσης throughput 1 αποτέλεσμα/κύκλο. Λόγω ομοχειρίας, θα υπάρξει μια καθυστέρηση 3 κύκλων για να υπολογιστεί το διάνυσμα πιθανοφανειών από ένα basic cell. Η συγκεκριμένη σχεδιαστική επιλογή είναι μόλις 0.052 φορές πιο αργή από την χρησιμοποίηση $4^{ων}$ basic cells αλλά 4 φορές πιο οικονομική σε επίπεδο πόρων. Η συγκεκριμένη επιλογή είναι καθοριστικής σημασίας για την ανάπτυξη της συγκεκριμένης αρχιτεκτονικής. Ο λόγος για τον οποίο συμβαίνει αυτό θα εξηγηθεί στη συνέχεια.
Ως το βασικό υποσύστημα της αρχιτεκτονικής μπορεί να οριστεί το basic cell μαζί με την κατάλληλη λογική ώστε να δημιουργείται ολόκληρο το διάνυσμα πιθανοφανειών. Η συγκεκριμένη σχεδιαστική επιλογή λοιπόν δίνει μια επεκταμένη βασική υπολογιστική(expanded basic cell-EBC) μονάδα με latency 58 κύκλων και throughput 1 πιθανοφάνεια/κύκλο ή 1 διάνυσμα πιθανοφανειών ανά 4 κύκλους.

Η διεπαφή της επεκταμένης βασικής υπολογιστικής μονάδας φαίνεται στην εικόνα 4-3.



Εικόνα 4-3 : Η διεπαφή της επεκταμένης βασικής υπολογιστικής μονάδας.

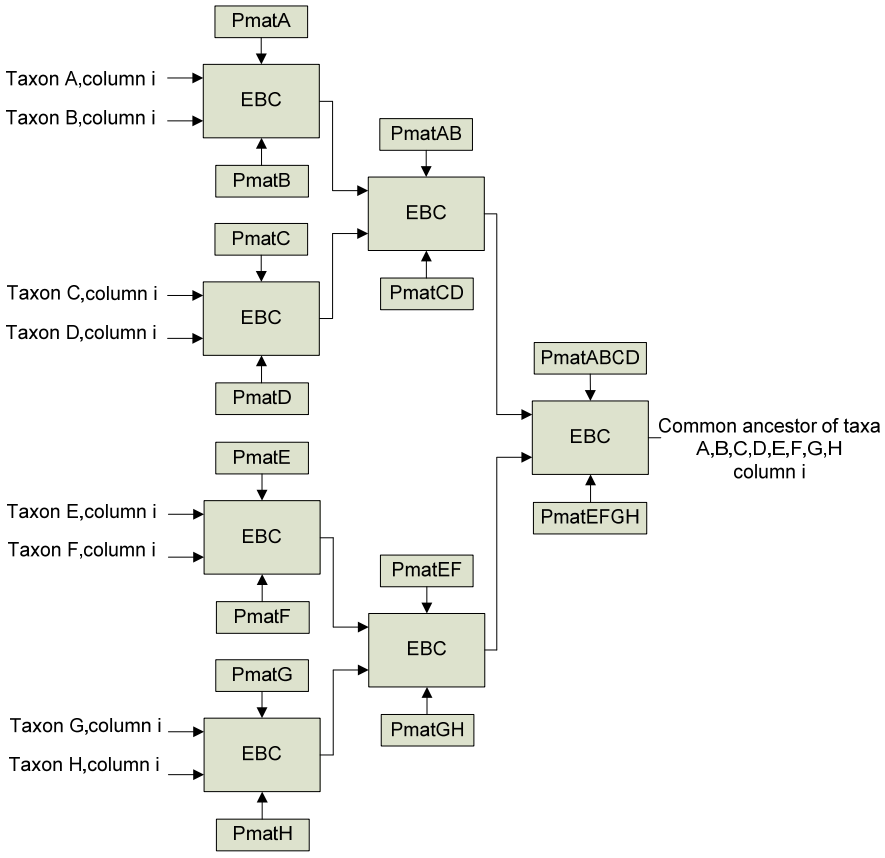
Το EBC αποτελείται από το basic cell καθώς και τους κατάλληλους καταχωρητές για να αποθηκεύονται προσωρινά οι τιμές των πιθανοφανειών μέχρι να δημιουργηθεί ολόκληρο το ζητούμενο διάνυσμα.

4.3 Το Μονοπάτι Δεδομένων (datapath)

Δεδομένης της δενδρικής μορφής των φυλογενετικών δένδρων θεωρείται ότι μια δενδρική διάταξη των basic cells θα ήταν η βέλτιστη. Παρ'όλα αυτά, περαίτερω έρευνα πρέπει να διεξαχθεί σχετικά με αυτό το θέμα καθώς η μορφή του δέντρου, δλδ. το πόσο ισοζυγισμένο είναι, παίζει σημαντικό ρόλο στην τελική απόδοση της σχεδίασης. Στη συνέχεια προτείνονται δύο εναλλακτικές τοπολογίες των βασικών υπολογιστικών μονάδων και εξηγούνται τα θετικά και αρνητικά της κάθε μιας.

4.3.1 Το Μονοπάτι Δεδομένων – Εναλλακτική 1

Σύμφωνα με την πρώτη εναλλακτική σχεδιαστική προσέγγιση που προτείνεται, τα expanded basic cells τοποθετούνται σε δενδρική δομή όπως φαίνεται στην εικόνα 4-4.



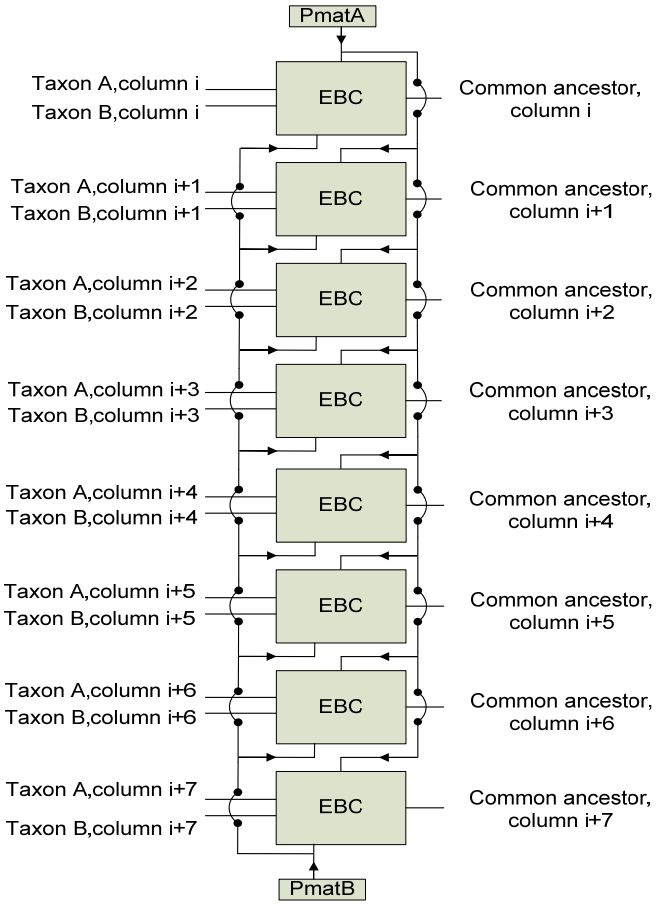
Εικόνα 4-4 : Δενδρική Τοπολογία της επεκταμένης βασικής υπολογιστική μονάδας

Η παραπάνω σχεδιαστική επιλογή παρουσιάζει 7 expanded basic cells σε δενδρική διάταξη. Τα 4 EBCs του πρώτου επιπέδου λειτουργούν παράλληλα και επεξεργάζονται δεδομένα από 8 διαφορετικές λειτουργικές ταξινομικές μονάδες. Η πληροφορία που προκύπτει αποτελεί είσοδο στα EBCs του δεύτερου επιπέδου και τα αποτελέσματα του δεύτερου επιπέδου αποτελούν είσοδο για το τρίτο. Με αυτόν τον τρόπο, στο πρώτο επίπεδο υπολογίζονται παράλληλα 4 διανύσματα πιθανοφανειών για μια θέση i

των ακολουθιών των κοινών προγόνων των 8 αντίστοιχων λειτουργικών μονάδων. Ακολούθως, στο δεύτερο επίπεδο υπολογίζονται παράλληλα 2 διανύσματα πιθανοφανειών για την ίδια θέση i των ακολουθιών των κοινών προγόνων των 4 λειτουργικών μονάδων που υπολογίστηκαν στο προηγούμενο επίπεδο και τέλος υπολογίζεται το διάνυσμα πιθανοφανειών για την θέση i του κοινού προγόνου των δύο λειτουργικών ταξινομικών μονάδων.
Αν αναλογιστεί κανείς την περιορισμένη μνήμη BRAM που υπάρχει σε μια FPGA, η παραπάνω σχεδίαση είναι ιδιαίτερα οικονομική. Για κάθε 2048 bit πληροφορίας (8(είδη)x256bits(1 διάνυσμα πιθανοφανειών/είδος))χρειάζεται να αποθηκεύονται στη μνήμη 256 bits γεγονός το οποίο εξοικονομεί μνήμη η οποία μπορεί να χρησιμοποιηθεί όπως θα εξηγηθεί στη συνέχεια για να βελτιώσει σημαντικά τόσο την απόδοση της σχεδίασης όσο και το μέγιστο μέγεθος του φυλογενετικού δέντρου που μπορεί να υπολογιστεί από την συγκεκριμένη αρχιτεκτονική.

4.3.2 Το Μονοπάτι Δεδομένων – Εναλλακτική 2

Υπάρχει η δυνατότητα τοποθέτησης των EBCs στο ίδιο επίπεδο και όχι σε δενδρική διάταξη, όπως φαίνεται στην εικόνα 4-5.



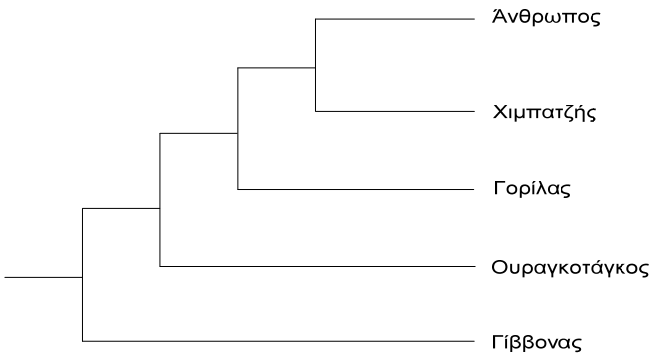
Εικόνα 4-5 : Διανυσματική τοπολογία της επεκταμένης βασικής υπολογιστικής μονάδας.

Η παραπάνω σχεδιαστική επιλογή είναι ανεξάρτητη από την μορφή του δέντρου και δέχεται είσοδο από τις ακολουθίες 2 λειτουργικών ταξινομικών μονάδων σε αντίθεση με την εναλλακτική 1 η οποία δέχεται είσοδο από 8 λειτουργικές ταξινομικές

μονάδες. Η εναλλακτική 2 υπολογίζει παράλληλα τα διανύσματα πιθανοφανειών για 8 θέσεις της άγνωστης ακολουθίας του κοινού προγόνου των δύο λειτουργικών ταξινομικών μονάδων. Δεδομένου του τρόπου επεξεργασίας των πληροφορίας, δλδ. περισσότερες από μια θέση των ακολουθιών υπολογίζονται παράλληλα για δύο OTUs(εναλλακτική 2) έναντι μιας θέσης για περισσότερα από ένα ζευγάρια OTUs (εναλλακτική 1), είναι πιο πιθανό η εναλλακτική 2 να έχει καλύτερη απόδοση από την εναλλακτική 1 καθώς τα basic cells θα υπολογίζουν πάντα χρήσιμη πληροφορία. Στην πρώτη σχεδιαστική επλογή, είναι πολύ πιθανό αρκετά από τα 7 basic cells να μην υπολογίζουν χρήσιμη πληροφορία κατά τη διάρκεια υπολογισμού των διανυσμάτων πιθανοφανειών και αυτό γιατί, όπως προαναφέρθηκε, η πρώτη εναλλακτική σχεδιαστική επιλογή εξαρτάται άμεσα και επηρεάζεται από την τοπολογία του δέντρου. Παρ'όλα αυτά, προτιμήθηκε η πρώτη σχεδιαστική επιλογή εξαιτίας της περιορισμένης μνήμης που υπάρχει στην FPGA. Η συγκεκριμένη επιλογή προσφέρει την δυνατότητα υπολογισμού δέντρου με αρκετές εκατοντάδες είδη καθώς κάθε φορά η μνήμη καταναλώνεται για συγκεκριμένο αριθμό στηλών των νουκλεοτιδικών ακολουθιών σε αντίθεση με την δεύτερη εναλλακτική προσέγγιση η οποία θα χρειαζόταν να αποθηκεύσει τα διανύσματα πιθανοφανειών για κάποιες θέσεις των ακολουθιών για όλες τις ταξινομικές λειτουργικές μονάδες.

4.3.3 Το Μονοπάτι Δεδομένων – Συνολική Εικόνα

Αφού επιλέχτηκε η σχεδιαστική επιλογή που τοποθετεί τα EBCs σε δενδρική διάταξη, η σχεδίαση επεκτάθηκε ώστε να μπορούν να υπολογιστούν δέντρα με περισσότερα από 8 OTUs. Τοποθετήθηκαν μνήμες BRAM ώστε να αποθηκεύονται προσωρινά τα διανύσματα πιθανοφανειών για τις x θέσεις των ακολουθιών των ταξινομικών λειτουργικών μονάδων μέχρι να υπολογιστούν ανα 8άδες όλες οι ταξινομικές λειτουργικές μονάδες. Επίσης προστέθηκε η κατάλληλη λογική ώστε να προωθούνται όπου και όποτε χρειάζεται τα διανύσματα πιθανοφανειών. Προώθηση των διανυσμάτων πιθανοφανειών των θέσεων μπορεί να χρειαστεί να πραγματοποιηθεί σε περιπτώσεις δέντρων που έχουν μορφή παρεμφερή με του δέντρου που φαίνεται στην εικόνα 4-6.

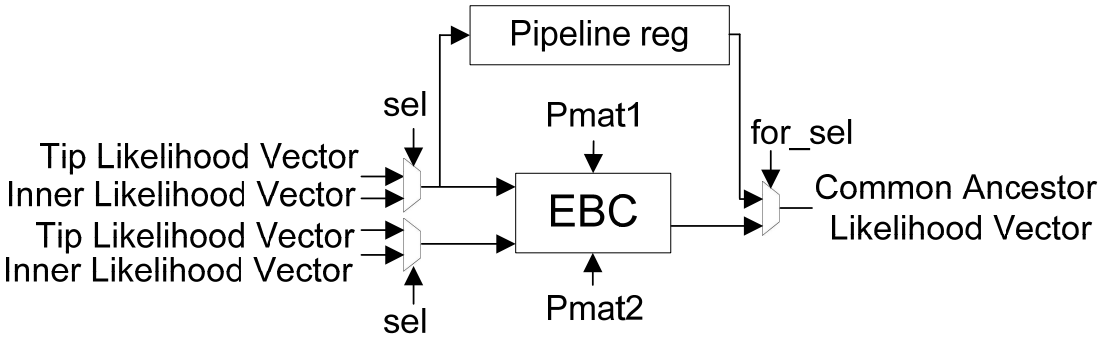


Εικόνα 4-6 : Ενδεικτική μορφή μη ισοζυγισμένου φυλογενετικού δέντρου.

Στο δέντρο που φαίνεται στην παραπάνω εικόνα τα διανύσματα πιθανοφανειών των ανθρώπου και του χιμπατζή θα συνδυαστούν μεταξύ τους μέσω ενός EBC αλλά ο κοινός απόγονος που θα προκύψει, πρέπει να συνδυαστεί με τον γορίλα. Η προώθηση

των διανυσμάτων πιθανοφανειών του γορίλα θα φέρει το συγκεκριμένο είδος ένα επίπεδο λιγότερο βαθιά στο δέντρο, στο ίδιο επίπεδο με τον κοινό πρόγονο του ανθρώπου και του χιμπατζή, με τον οποίο πρέπει να συνδυαστεί.

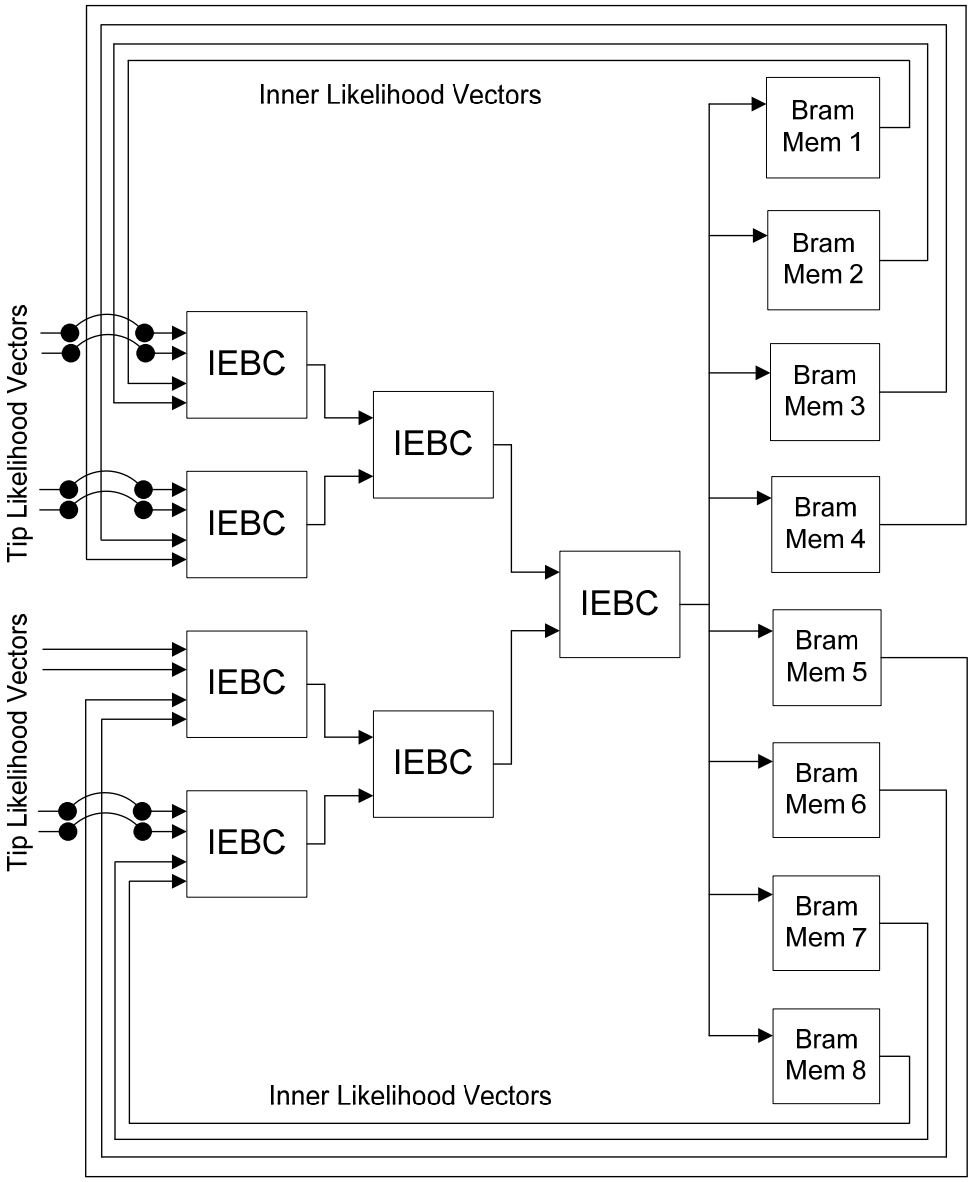
Στην επόμενη εικόνα φαίνεται το βελτιωμένο EBC(IEBC-Improved Expanded Basic Cell) με τους κατάλληλους πολυπλέκτες και ένα διάνυσμα από καταχωρητές με σκοπό την προώθηση του διανύσματος πιθανοφανειών της εισόδου στην έξοδο.



Εικόνα 4-7 : Η βελτιωμένη και επεκταμένη βασική υπολογιστική μονάδα.

Τα Tip Likelihood Vectors προέρχονται από τον μεταφραστή χαρακτήρων που θα παρουσιαστεί στην επόμενη παράγραφο, και κατά συνέπεια ανήκουν σε κάποια ταξινομική λειτουργική μονάδα που βρίσκεται σε κάποιο από τα φύλλα του δέντρου. Τα Inner Likelihood Vectors έχουν υπολογιστεί από κάποιο IEBC και βρίσκονταν αποθηκευμένα σε κάποια από τις 8 μνήμες που υπάρχουν για τον σκοπό αυτό. Τα σήματα sel και for_sel δημιουργούνται από τον control unit που θα παρουσιαστεί στη συνέχεια.

Έτσι λοιπόν, η συνολική εικόνα του μονοπατιού δεδομένων μέχρι το σημείο που έχει παρουσιαστεί φαίνεται στην εικόνα 4-8. Στο επόμενο σχήμα φαίνεται μόνο η διαδρομή δεδομένων και παραλείπονται όλα τα υπόλοιπα σήματα όπως τα σήματα ελέγχου και οι πίνακες πιθανοτήτων.



Εικόνα 4-8 : Το μονοπάτι δεδομένων- Δενδρική Τοπολογία με χρήση των βελτιωμένων και επεκταμένων βασικών υπολογιστικών μονάδων.

Η λογική επεξεργασίας των δεδομένων είναι να επεξεργάζονται ανα 8άδες οι ταξινομικές λειτουργικές μονάδες και να αποθήκεύονται στις μνήμες τα διανύσματα πιθανοφανειών των κοινών προγόνων των ταξινομικών λετουργικών μονάδων που βρίσκονται 3 επίπεδα λιγότερο βαθιά στο δέντρο. Αυτός είναι ο λόγος για τον οποιό η συγκεκριμένη αρχιτεκτονική μπορεί να υπολογίσει δέντρα με μεγάλο αριθμό από OTUs χρησιμοποιώντας μόνο την διαθέσιμη μνήμη της FPGA(Block RAM). Γίνεται πλέον κατανοητό γιατί επιλέχθηκε η πρώτη σχεδιαστική επιλογή, δλδ. η δενδρική τοπολογία των basic cells.

4.4 Κωδικοποίηση και Μετάφραση Χαρακτήρων

Όλες οι ακολουθίες νουκλεοτιδίων που χρησιμοποιούνται ως είσοδος για τον υπολογισμό του βαθμού πιθανοφάνειας ενός φυλογενετικού δέντρου προέρχονται από αλγόριθμους πολλαπλής ταύτισης. Ένα από τα πιο γνωστά προγράμματα που χρησιμοποιούνται σήμερα για πολλαπλές ταυτίσεις είναι το ClustalW. Πρέπει να τονιστεί το γεγονός ότι το τελικό αποτέλεσμα όσο αφορά τον βαθμό πιθανοφάνειας του δέντρου προς υπολογισμό εξαρτάται άμεσα από το πόσο επιτυχημένη είναι η πολλαπλή ταύτιση των ακολουθιών που αντιστοιχούν στα φύλλα του δέντρου.

4.4.1 Κωδικοποίηση

Λόγω του γεγονότος ότι οι ακολουθίες που χρησιμοποιούνται προέρχονται από προγράμματα πολλαπλής ταύτισης το αλφάβητο χαρακτήρων διευρύνεται καθώς εισάγονται χαρακτήρες αμφιβολίας κατά την διαδικασία ταύτισης. Το σύνολο όλων των χαρακτήρων που μπορεί να εμφανιστούν σε ένα αρχείο PHYLIP καθώς και η κωδικοποίηση που ακολουθείται από την συγκεκριμένη αρχιτεκτονική φαίνονται στον ακόλουθό πίνακα.

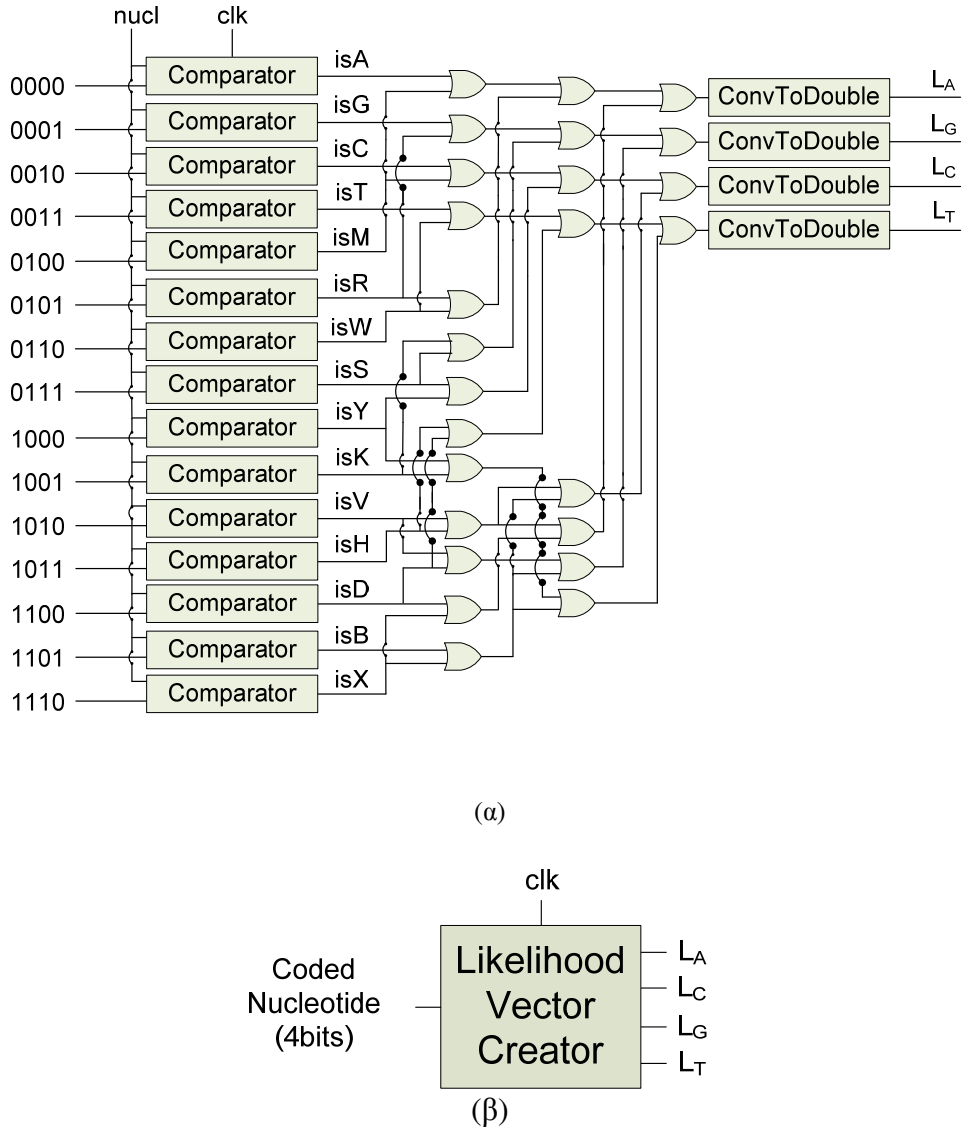
IUPAC-IUB/GCG Code	Meaning	Coding
A	A	0000
G	G	0001
C	C	0010
T/U	T	0011
M	A or C	0100
R	A or G	0101
W	A or T	0110
S	C or G	0111
Y	C or T	1000
K	G or T	1001
V	A or C or G	1010
H	A or C or T	1011
D	A or G or T	1100
B	C or G or T	1101
X/N	A or C or G or T	1110
.	not (A or C or G or T)	1111

Πίνακας 4-1 : Χαρακτήρες Αμφιβολίας και η κωδικοποίηση τους

Οι παραπάνω χαρακτήρες αμφιβολίας έχουν οριστεί από το MRC HGU(Medical Research Council Human Genetics Unit).

4.4.2 Μετάφραση Χαρακτήρα σε Διάνυσμα Πιθανοφανειών

Για την «μετάφραση» των χαρακτήρων του παραπάνω πίνακα σε διανύσματα πιθανοφανειών δημιουργήθηκε ένα σύστημα μετάφρασης. Πρέπει να σημειωθεί ότι η μετάφραση των νουκλεοτιδίων σε διανύσματα πιθανοφανειών έχει νοήμα μόνο για τις ακολουθίες που βρίσκονται στα φύλλα ενός φυλογενετικού δέντρου. Για τους εσωτερικούς κόμβους οι ακολουθίες είναι άγνωστες και μόνο τα διανύσματα πιθανοφανειών μπορούν να υπολογιστούν. Στη επόμενη εικόνα φαίνεται το σύστημα μετάφρασης το οποίο ματαφράζει τις 4 πρωτεϊνικες βάσεις και τους 12 χαρακτήρες αμφιβολίας.



Εικόνα 4-9 : (α) Μεταφραστής Νουκλεοτιδίου Εισόδου σε Διάνυσμα Πιθανοφανειών, (β) Η διεπαφή του Μεταφραστή

Το διάνυσμα συγκριτών χρησιμοποιείται για να εντοπιστεί ποιος είναι ο χαρακτήρας στην είσοδο. Έτσι λοιπόν, μόνο ένα από τα σήματα isA, isG, isC, isT, … , isX που φαίνονται στην παραπάνω εικόνα μπορεί να είναι 1 κάθε φορά. Ανάλογα με το ποιό σήμα γίνεται 1 προκύπτει διαφορετικό διάνυσμα πιθανοφανειών το οποίο μπορεί να περιέχει από κανέναν μέχρι και τέσσερις άσσους στις τέσσερις θέσεις του. Ποιές θέσεις θα πάρουν την τιμή 1 υπολογίζεται από το σύμπλεγμα πυλών OR που φαίνονται στη εικόνα 4-9. Αφού υπολογιστούν τέσσερα σήματα(οι έξοδοι των $4^{ων}$ πυλών OR που φαίνονται στο τρίτο επίπεδο πυλών-από αριστερά προς τα δεξιά), ένα για κάθε θέση του ζητούμενου διανύσματος, η τιμή αυτή μετρατρέπεται στην αντίστοιχη τιμή σε αριθμητική κινητής υποδιαστολής διπλής ακρίβειας.
Όπως μπορεί κανείς να παρατηρήσει, η κωδικοποιημένη τιμή "1111" που αντιστοιχεί στον χαρακτήρα . (τελεία) δεν λαμβάνεται υπόψιν κατά την αποκωδικοποίηση καθώς καλύπτεται από τους υπόλοιπους συγκριτές. Αν υπάρχει είσοδος "1111" κανέναν συγκριτής δεν θα δώσει θετική τιμή και κανένα σήμα δεν θα πάρει ποτέ την τιμή 1 πριν την τελική μετατροπή σε αναπαράσταση κινητής υποδιαστολής διπλής ακρίβειας.

4.5 Οι Φάσεις Λειτουργίας

Η εκτέλεση περνάει από τρείς φάσεις μέχρι να υπολογιστει ο βαθμός πιθανοφάνειας του φυλογενετικού δέντρου. Η πρώτη φάση είναι η φάση υπολογισμού των διανυσμάτων για τους πιο πρόσφατους εξελικτικά προγόνους των ειδών που βρίσκονται στα φύλλα του δέντρου. Η δεύτερη φάση είναι η φάση υπολογισμού των εσωτερικών κόμβων του δέντρου ενώ η τρίτη και τελευταία φάση είναι ο υπολογισμός του βαθμού πιθανοφάνειας από τα διανύσματα πιθανοφανειών των θέσεων της ρίζας.

4.5.1 Η πρώτη φάση λειτουργίας

Κατά τη φάση αυτή, «διαβάζονται» οι ακολουθίες νουκλεοτιδίων των φύλλων του δέντρου από την είσοδο του συστήματος. Στην συγκεκριμένη σχεδίαση θεωρήθηκε ότι η είσοδος προέρχεται από μια εξωτερική μνήμη. Η σχεδιαστική επιλογή που ακολουθήθηκε στην αρχιτεκτονική επεξεργάζεται παράλληλα 8 νουκλεοτίδια, καθένα από την ίδια θέση 8 διαφορετικών νουκλεοτιδικών ακολουθιών. Αφού δημιουργούνται τα διανύσματα πιθανοφανειών για τις 8 αυτές πρωτεϊνικες βάσεις, ξεκινάει η λειτουργία των βασικών υπολογιστικών μονάδων. Οι μονάδες επεξεργάζονται την είσοδο ανα 8άδες νουκλεοτιδίων και το αποτελέσμα, που είναι το διάνυσμα πιθανοφανειών της θέσης υπο μελέτη του κοινού προγόνου των 8 ειδών, αποθηκεύεται σε μια μνήμη BRAM. Οι 8 μνήμες που έχουν δεσμευτεί για τον σκοπό αυτό «γεμίζουν» διαδοχικά, και ανάλογα με το μέγεθος του δέντρου εκτελείται συγκεκριμένος αριθμός εγγραφών σε κάθε μνήμη. Η πρώτη φάση λειτουργίας τελειώνει έχωντας γεμίσει όλες οι μνήμες με διανύσματα πιθανοφανειών, μόνο για τους πιο πρόσφατους εξελικτικά προγόνους των φύλλων οι οποίοι βρίσκονται τρία επίπεδα λιγότερο βαθιά στο δέντρο, όπως εξηγήθηκε νωρίτερα.

4.5.2 Η δεύτερη φάση λειτουργίας

Στην παρούσα φάση, υπολογίζονται τα διανύσματα πιθανοφανειών για όλους τους εσωτερικούς κόμβους του δέντρου μέχρι και της ρίζας. Κατά τη διάρκεια της φάσης αυτή η είσοδος στις βασικές υπολογιστικές μονάδες προέρχεται από τις μνήμες BRAM και όχι από την εξωτερική μνήμη και η έξοδος αποθηκεύεται και πάλι στις ίδιες μνήμες, σε θέσεις που έχουν διαβαστεί αρκετούς κύκλους νωρίτερα και έτσι δεν περιέχουν πλέον χρήσιμη πληροφορία. Η δεύτερη φάση λειτουργίας τελειώνει έχωντας υπολογίσει και αποθηκεύσει στη μνήμη τα διανύσματα για την ρίζα του δέντρου.

4.5.3 Η τρίτη φάση λειτουργίας

Στην τελευταία φάση διαβάζονται τα διανύσματα πιθανοφανειών που αντιστοιχούν στις θέσεις της άγνωστης ακολουθίας της ρίζας και υπολογίζεται ο βαθμός πιθανοφάνειας του δέντρου. Διαβάζεται ένα μέρος της μιας από τις 8 μνήμες και γίνεται ένα σύνολο μαθηματικών πράξεων, προσθέσεων και πολλαπλασιασμών. Αν επαρκεί ο αριθμός των περισσευούμενων DSPs για επιπλέον πολλαπλασιαστές και αθροιστές η παρούσα φάση μπορεί να επιταχυνθεί σημαντικά, αν αναλογιστεί κανείς ότι η προηγούμενη φάση τελειώνει έχωντας αποθηκεύσει χρήσιμη πληροφορία σε ένα μέρος της μιας από τις 8 μνήμες. Προτείνεται, η ίδια πληροφορία να αποθηκεύεται σε όλες τις μνήμες έτσι ώστε κατά τη διάρκεια της τελευταίας φάσης να διαβάζεται και να επεξεργάζεται παράλληλα πληροφορία και από τις 8 μνήμες. Το πόσο παραλληλισμού που μπορεί κανείς να επιτύχει κάνωντας αυτό το τέχνασμα φράσσεται τόσο από το αριθμό των μνημών (8) όσο και από τον αριθμό των διαθέσιμων DSPs.

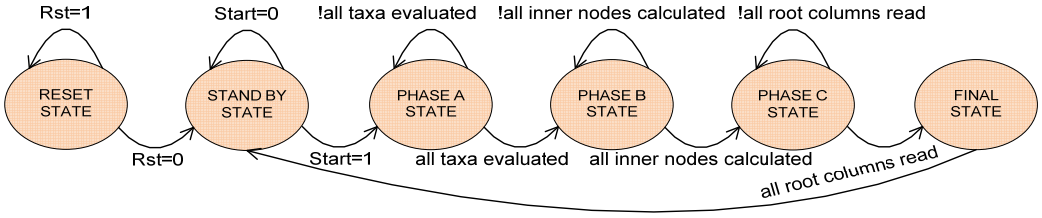
4.6 Η Μονάδα Ελέγχου

Για να λειτουργήσει σωστά το μονοπάτι δεδομένων είναι απαραίτητη η παρουσία μιας μονάδας ελέγχου. Η μονάδα αυτή, αποτελείται από ένα σύνολο μηχανών πεπερασμένων καταστάσεων (Finite State Machines-FSMs) οι οποίες είναι οργανωμένες σε ιεραρχία. Το σύνολο των μηχανών αυτών δημιουργεί τα σωστά σήματα για τους πολυπλέκτες στις εισόδους και εξόδους των βασικών υπολογιστικών μονάδων σε κάθε φάση λειτουργίας της αρχιτεκτονικής καθώς και για εγγραφή και ανάγνωση από τις μνήμες BRAM. Στη συνέχεια θα περιγραφούν οι μηχανές πεπερασμένων καταστάσεων που συντονίζουν την λειτουργία της αρχιτεκτονικής.

4.6.1 FSM 1

Η πρώτη FSM βρίσκεται στο υψηλότερο επίπεδο της ιεραρχίας και θέτει σε λειτουργία δύο άλλες FSMs. Το πρώτο, το δεύτερο και το τελευταίο στάδιο είναι κοινά σε όλες τις μηχανές και ονομάζονται RESET_STATE , STAND_BY_STATE και FINAL_STATE αντίστοιχα. Η FSM 1 αποτελείται από 3 βασικά στάδια, καθένα από τα οποία αντίστοιχεί σε κάθε μια από τις τρείς φάσεις λειτουργίας.

Η συγκεκριμένη FSM φαίνεται στην επόμενη εικόνα.



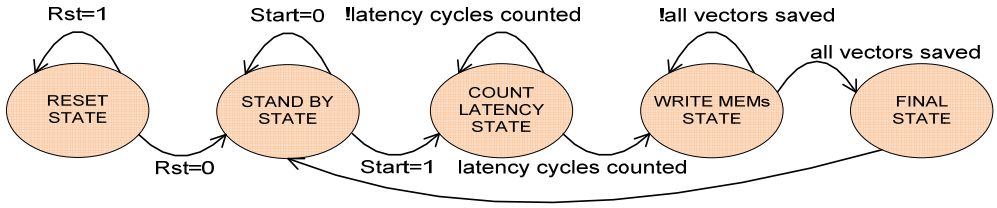
Εικόνα 4-10 : Μηχανή Πεπερασμένων Καταστάσεων του πρώτου επιπέδου της ιεραρχίας για συντονισμό των υπόλοιπων μηχανών.

Η PHASE A STATE θέτει σε λειτουργία την FSM 2 και η μηχανή παραμένει σε αυτήν την κατάσταση μέχρι να διαβαστεί όλη η εξωτερική μνήμη και να γεμίσουν οι 8 μνήμες BRAM της FPGA. Η PHASE B STATE θέτει σε λειτουργία την FSM 2 και η μηχανή παραμένει στην ίδια κατάσταση μέχρι να διαβαστούν και να επανεγραφούν οι 8 μνήμες όσες φορές χρειαστεί μέχρι να υπολογιστούν όλοι οι εσωτερικοί κόμβοι του δέντρου έως την ρίζα.
Η PHASE C STATE θέτει σε λειτουργία την FSM 3 και διαρκεί όσο χρειάζεται για να υπολογιστεί ο βαθμός πιθανοφάνειας του δέντρου από τα διανύσματα πιθανοφανειών της ρίζας.

4.6.2 FSM 2

Η δεύτερη FSM βρίσκεται στο δεύτερο επίπεδο της ιεραρχίας και θέτει σε λειτουργία δύο άλλες FSMs. Η συγκεκριμένη FSM αποτελείται από δύο βασικά στάδια. Το πρώτο στάδιο αντιστοιχεί στον χρόνο που χρειάζεται η ομοχειρία να δώσει το πρώτο αποτέλεσμα (latency) ενώ το δεύτερο στάδιο αντιπροσωπεύει την φάση αποθήκευσης των αποτελεσμάτων στις 8 μνήμες.

Η συγκεκριμένη FSM φαίνεται στην επόμενη εικόνα.



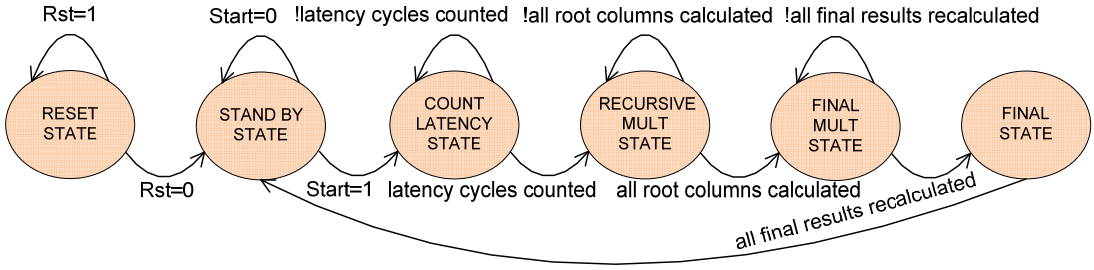
Εικόνα 4-11 : Μηχανή Πεπερασμένων Καταστάσεων του δεύτερου επιπέδου της ιεραρχίας για εγγραφή των μνημών.

Η COUNT LATENCY STATE ενεργοποιεί έναν counter ο οποίος μετράει τους κύκλους ρολογιού που χρειάζεται το μονοπάτι δεδομένων για να δώσει το πρώτο αποτέλεσμα. Έπειτα η μηχανή μεταβαίνει στην κατάσταση WRITE MEMs και

ενεργοποιεί την FSM 4 η οποία ελέγχει ποια μνήμη γράφεται κάθε φορά καθώς και την διεύθυνση εγγραφής.

4.6.3 FSM 3

Η τρίτη FSM βρίσκεται επίσης στο δεύτερο επίπεδο της ιεραρχίας. Λειτουργεί μόνο κατά τη διάρκεια της φάσης υπολογισμού του βαθμού πιθανοφάνειας του δέντρου. Ελέγχει την ανάγνωση των μνημών καθώς και τις εισόδους στους τελικούς πολλαπλασιαστές. Αποτελείται από τρείς βασικές καταστάσεις λειτουργίας όπως φαίνεται στην επόμενη εικόνα.

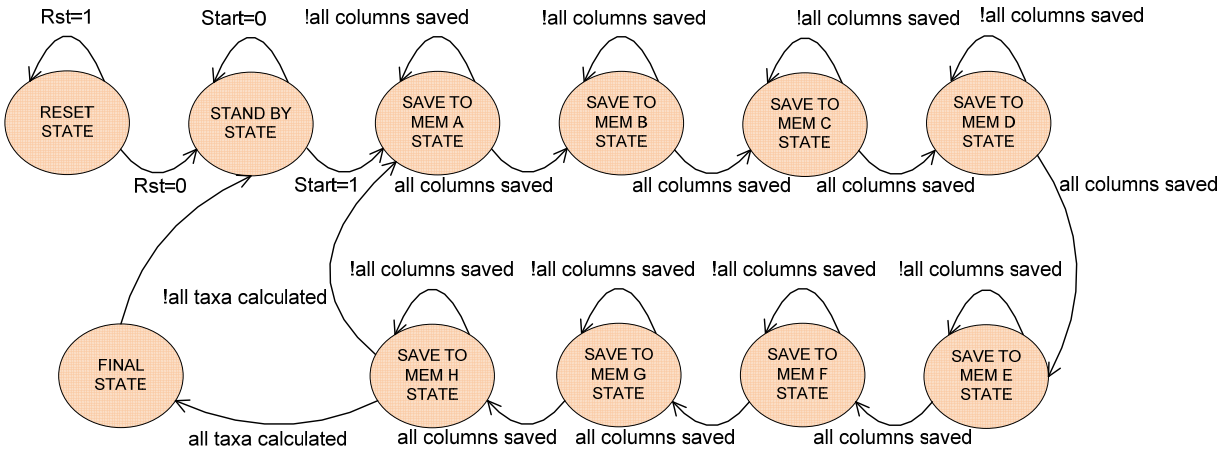


Εικόνα 4-12 : Μηχανή Πεπερασμένων Καταστάσεων του δεύτερου επιπέδου της ιεραρχίας για υπολογισμό του τελικού βαθμού πιθανοφάνειας του δέντρου.

Η COUNT LATENCY STATE ενεργοποιεί, όπως και στην FSM 2, έναν counter ο οποίος μετράει τους κύκλους ρολογιού που χρειάζεται το τελικό σύστημα πολλαπλασιαστών-αθροιστών να δώσει το πρώτο αποτέλεσμα για την πρώτη θέση της μνήμης. Κατά την διάρκεια της RECURSIVE MULT STATE ένας πολλαπλασιαστής χρησιμοποιείται για να παράγει το γινόμενο όλων των αποτελεσμάτων που προκύπτουν από το σύστημα πολλαπλασιαστών-αθροιστών. Λόγω του latency των 16 κύκλων που έχει ο συγκεκριμένος πολλαπλασιαστής προκύπτουν 16 αποτελέσματα τα οποία συνδυάζονται σε ένα τελικό σύμπλεγμα πολλαπλασιαστών κατά τη διάρκεια του FINAL MULT STATE προκειμένου να επιταχυνθεί η συγκεκριμένη διαδικασία. Η FINAL MULT STATE τελειώνει με την δημιουργία του βαθμού πιθανοφάνειας του δέντρου και την αποθήκευση του στην πρώτη θέση της πρώτης μνήμης.

4.6.4 FSM 4

Η τέταρτη FSM βρίσκεται στο τρίτο επίπεδο της ιεραρχίας. Ενεργοποιείται κατά τις φάσεις λειτουργίας A και B και δημιουργεί τα σήματα εγγραφής (wren) των μνημών καθώς και τα σήματα που ελέγχουν τις διευθύνσεις εγγραφής. Η συγκεκριμένη FSM αποτελείται από 8 βασικές καταστάσεις, μια για κάθε μια από τις 8 μνήμες, και φαίνεται στην επόμενη εικόνα.



Εικόνα 4-13 : Μηχανή Πεπερασμένων Καταστάσεων του τρίτου επιπέδου της ιεραρχίας για συντονισμό της εγγραφής των μνημών.

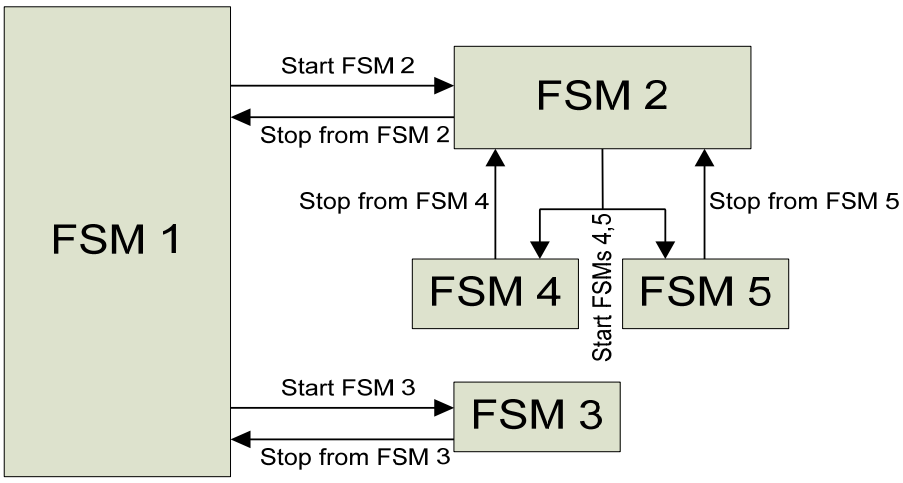
Κάθε μια από τις 8 βασικές καταστάσεις ενεργοποιεί εγγραφή σε μια από τις 8 μνήμες και παράλληλα αυξάνει τον μετρητή που χρησιμοποιείται για διευθυνσιοδότηση της αντίστοιχης μνήμης προς εγγραφή.

4.6.5 FSM 5

Η πέμπτη FSM βρίσκεται στο τρίτο επίπεδο της ιεραρχίας και είναι ένα απλός μετρητής ο οποίος χρησιμοποιείται για την διευθυνσιοδότηση των μνημών για ανάγνωση.

4.6.6 Η Ιεραρχία των Μηχανών Πεπερασμένων Καταστάσεων

Το ακόλουθο σχήμα δείχνει την διάταξη των πέντε μηχανών πεπερασμένων καταστάσεων που συνθέτουν την μονάδα λειτουργίας της αρχιτεκτονικής.



Εικόνα 4-14 : Η ιεραρχία των μηχανών πεπερασμένων καταστάσεων που συνθέτουν την μονάδα ελέγχου και ο τρόπος επικοινωνίας.

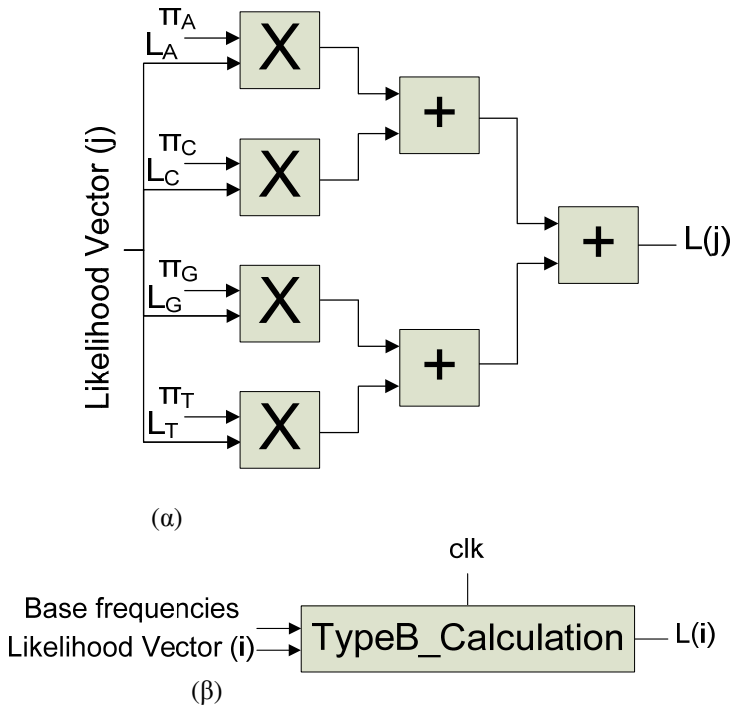
4.7 Σύστημα Εξαγωγής του Βαθμού Πιθανοφάνειας

Κατά την τρίτη φάση λειτουργίας υπολογίζεται ο βαθμός πιθανοφάνειας του δέντρου, λαμβάνοντας υπόψιν τις εκ των προτέρων(a priori) πιθανότητες εμφάνισης των πρωτεϊνικών βάσεων καθώς και τις πιθανοφάνειες που υπολογίστηκαν για κάθε θέση της ακολουθίας της ρίζας. Το σύστημα που σχεδιάστηκε για τον σκοπό αυτό υλοποιεί τους τύπους Β και Γ της συνάρτησης φυλογενετικής πιθανοφάνειας:

$$L(j) = \sum_{k} \pi_k \; L(x^{root}_{\;j} = k) \quad \text{με } k=A,C,G,T \qquad (B)$$

$$L = \prod_{j} L(j) \quad \{ \text{ ή } L = \sum_{j} \log(L(j)) \; \} \qquad (Γ)$$

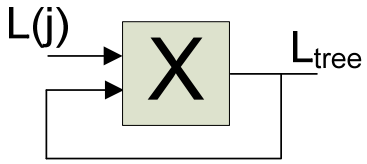
Η βασική μονάδα που υλοποιεί τον τύπο Β αποτελείται από 4 πολλαπλασιαστές και τρείς αθροιστές και φαίνεται στην εικόνα 4-15.



Εικόνα 4-15 : (α) Η βασική μονάδα για τον υπολογισμό της συνολικής πιθανοφάνειας για κάθε θέση της ακολουθίας της ρίζας (β) Η διεπαφή της συγκεκριμένης μονάδας

Η συγκεκριμένη μονάδα επεξεργάζεται το διάνυσμα πιθανοφανειών για κάθε θέση j της ρίζας, και παράγει την τιμή L(j). Πιο συγκεκριμένα, οι τέσσερις πολλαπλασιαστές παράγουν τα γινόμενα π_k x L($root_j$=k) για k=A,C,G και T , ενώ οι τρείς αθροιστές δίνουν το ζητούμενο άθροισμα.

Η μονάδα που υπολογίζει τον τύπο Γ είναι ένας απλός πολλαπλασιαστής ο οποίος ανατροφοδοτεί στην είσοδο το αποτέλεσμα του, όπως φαίνεται στην εικόνα 4-16.



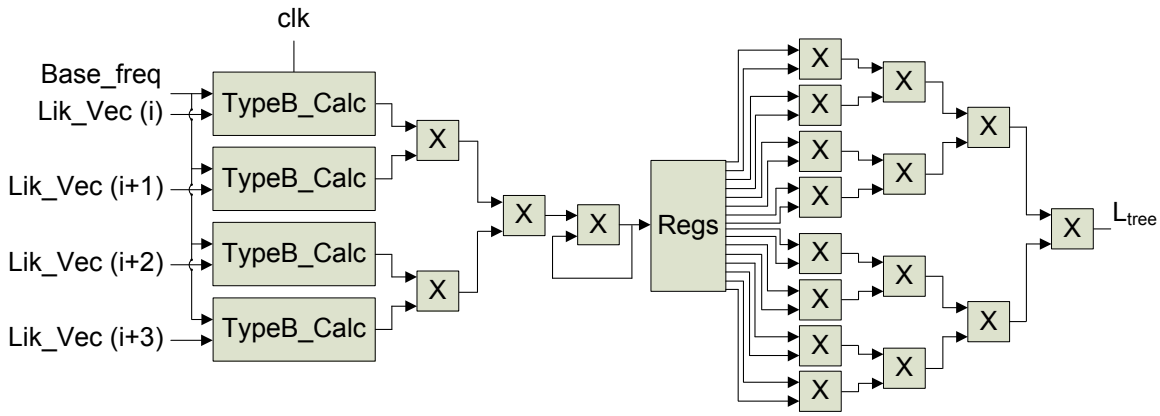
Εικόνα 4-16 : Πολλαπλασιαστής που ανατροφοδοτεί την έξοδο στην είσοδο, για υπολογισμό του γινομένου αγνώστου πλήθους πιθανοφανειών

Η πρώτη είσοδος του πολλαπλασιαστή προέρχεται από την έξοδο του συστήματος που υλοποιεί τον τύπο Β και παρουσιάστηκε προηγουμένως ενώ η δεύτερη είσοδος προέρχεται από την έξοδο του ίδιου του πολλαπλασιαστή. Βέβαια, η συγκεκριμένη είσοδος είναι αρχικοποιημένη στην τιμή 1 μέχρι να δημιουργηθεί το πρώτο αποτέλεσμα από τον πολλαπλασιαστή ο οποίος έχει latency 16 κύκλων. Ο αριθμός των εισόδων στον συγκεκριμένο πολλαπλασιαστή είναι ίδιος με το μήκος των νουκλεοτιδικών ακολουθιών των φύλλων του δέντρου. Προφανώς, λόγω του latency των 16 κύκλων του πολλαπλασιαστή, τα 16 τελευταία αποτελέσματα που προκύπτουν, δεν έχει νόημα να ανατροφοδοτηθούν στην είσοδο και πρέπει να πολλαπλασιαστούν μεταξύ τους. Για τον λόγο αυτό χρειάζεται ένας πολυπλέκτης στην πρώτη είσοδο του πολλαπλασιαστή ώστε να επιλέγεται αν η συγκεκριμένη είσοδος θα προέρχεται από τους προηγούμενους υπολογισμούς (από την μονάδα που υπολογίζει τον τύπο Β) ή από την έξοδο του πολλαπλασιαστή.

Η συγκεκριμένη σχεδιαστική επιλογή είναι ιδιαίτερα αργή αν αναλογιστεί κανείς ότι για τον πολλαπλασιασμό x στοιχείων με $x \leq$ MultLatency χρειάζονται (x-1)*MultLatency κύκλοι, που στην συγκεκριμένη περίπτωση είναι 15*16=240 κύκλοι. Παρ'όλα αυτά, λόγω του νόμου του Amdhal, η προσπάθεια επιτάχυνσης των συγκεκριμένων 16 πολλαπλασιασμών θα έχει ελάχιστη επίδραση στον συνολικό χρόνο εκτέλεσης καθώς οι 240 αυτοί κύκλοι αποτελούν λιγότερο από το 7% του συνολικού χρόνου εκτέλεσης για μικρά φυλογενετικά δέντρα, και το συγκεκριμένο ποσοστό μειώνεται με την αύξηση του μεγέθους των δέντρων.
Όπως αναφέρθηκε κατά την περιγραφή της τρίτης φάσης λειτουργίας, η συγκεκριμένη φάση μπορεί να επιταχυνθεί. Αν κατά την τελευταία εγγραφή των μνημών στην φάση υπολογισμού των εσωτερικών κόμβων του δέντρου, οι πιθανοφάνειες των θέσεων της ρίζας αποθηκευτούν σε περισσότερες από μια μνήμες, τότε μπορούν να επεξεργαστούν παράλληλα περισσότερες από μια θέσεις της ρίζας, με την τοποθέτηση της μονάδας που υπολογίζει τον τύπο Β σε παράλληλη διάταξη, περισσότερες από μια φορές. Φυσικά, δεν υπάρχει κανέναν όφελος αν η συγκεκριμένη μονάδα τοποθετηθεί περισσότερες από 8 φορές καθώς δεν θα υπάρχει μνήμη για να τροφοδοτήσει τις επιπλέον μονάδες. Η υλοποίηση των 7 basic cells που συνθέτουν το μονοπάτι δεδομένων χρησιμοποιεί 567 DSP48E slices για πολλαπλασιαστές και 126 DSP48E slices για αθροιστές, γεγονός που αφήνει διάθεσιμα 363 DSP48E slices. Από αυτά, τα 9 πρέπει να χρησιμοποιηθούν για την υλοποίηση του παραπάνω πολλαπλασιαστή και έτσι μένουν 354 διαθέσιμα τα οποία μπορούν να χρησιμοποιηθούν για την επιτάχυνση της φάσης υπολογισμού του βαθμού πιθανοφάνειας.

Η υλοποίηση που προτείνεται προϋποθέτει την αποθήκευση των πιθανοφανειών της ρίζας σε 4 μνήμες. Η τοποθέτηση $4^{\omega v}$ βασικών μονάδων που υπολογίζουν τον τύπο Β για 4 θέσεις της ακολουθίας παράλληλα επιταχύνει τον υπολογισμό του τύπου Β για τις θέσεις τις ρίζας 4 φορές και επίσης παραμένουν διαθέσιμα αρκετά DSP48E slices για την επιτάχυνση και του υπολογισμού του τύπου Γ, παρ'όλο που είναι γνωστό εξαρχής ότι η συγκεκριμένη βελτίωση δεν θα έχει αισθητή επίδραση στον συνολικό χρόνο εκτέλεσης του συστήματος.

Το συνολικό σύστημα που δημιουργήθηκε φαίνεται στην εικόνα 4-17.



Εικόνα 4-17 : Υποσύστημα που υπολογίζει τον βαθμό πιθανοφάνειας του δέντρου από τα διανύσματα πιθανοφανειών των θέσεων της ρίζας

Επιταχύνωντας περίπου 4 φορές τους υπολογισμούς που επιδέχονταν βελτίωση επιτεύχθηκε συνολικό speed up του συγκεκριμένου υπολογισμού 1,6x για μήκη ακολουθιών μεγαλύτερα από 1000 νουκλεοτίδια, που είναι και η συνήθης περίπτωση σε πραγματικές φυλογενετικές αναλύσεις.

4.8 Αναδιάταξη της πληροφορίας του αρχείου εισόδου

Για τα πλαίσια της παρούσας διατριβής θεωρήθηκε ότι η πληροφορία εισόδου «διαβάζεται» από μια εξωτερική μνήμη DRAM. Η παρούσα αρχιτεκτονική δέχεται ως είσοδο ένα αρχείο PHYLIP μορφής για πρωτεϊνικές βάσεις. Το συγκεκριμένο αρχείο αναδιατάσσεται κατάλληλα σύμφωνα με τις δυνατότητες της μνήμης να παρέχει δεδομένα έτσι ώστε να μην υπάρχει πρόβλημα εισόδου στη σχεδίαση και να μην σταματάει ποτέ η ομοχειρία να τροφοδοτείται με νέα δεδομένα. Ένα παράδειγμα αρχείου PHYLIP φαίνεται στην ακόλουθη εικόνα.

```
10 705
Cow         ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACATCACCAATCATAGAAGAACTA
Carp        ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGCCATACCCGTTATAGAGGAACTT
Chicken     ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCATCCCCCATCATAGAAGAGCTC
Human       ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATCATAGAAGAGCTT
Loach       ATGGCACATCCCACACAATTAGGATTCCAAGACGCGGCCTCACCCGTAATAGAAGAACTT
Mouse       ATGGCCTACCCATTCCAACTTGGTCTACAAGACGCCACATCCCCTATTATAGAAGAGCTA
Rat         ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACATCACCTATCATAGAAGAACTT
Seal        ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAACCTCTCCCATTATAGAGGAGTTA
Whale       ATGGCATATCCATTCCAACTAGGTTTCCAAGATGCAGCATCACCCATCATAGAAGAGCTC
Frog        ATGGCACACCCATCACAATTAGGTTTTCAAGACGCAGCCTCTCCAATTATAGAAGAATTA

CTTCACTTTCATGACCACACGCTAATAATTGTCTTCTTAATTAGCTCATTAGTACTTTAC
CTTCACTTCCACGACCACGCATTAATAATTGTGCTCCTAATTAGCACTTTAGTTTTATAT
GTTGAATTCCACGACCACGCCCTGATAGTCGCACTAGCAATTTGCAGCTTAGTACTCTAC
ATCACCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTCCTAGTCCTGTAT
CTTCACTTCCATGACCATGCCCTAATAATTGTATTTTTGATTAGCGCCCTAGTACTTTAT
ATAAATTTCCATGATCACACACTAATAATTGTTTTCCTAATTAGCTCCTTAGTCCTCTAT
ACAAACTTTCATGACCACACCCTAATAATTGTATTCCTCATCAGCTCCCTAGTACTTTAT
CTACACTTCCATGACCACACATTAATAATTGTGTTCCTAATTAGCTCATTAGTACTCTAC
CTACACTTTCACGATCATACACTAATAATCGTTTTTCTAATTAGCTCTTTAGTTCTCTAC
CTTCACTTCCACGACCATACCCTCATAGCCGTTTTTCTTATTAGTACGCTAGTTCTTTAC

.....
.....
.....
```

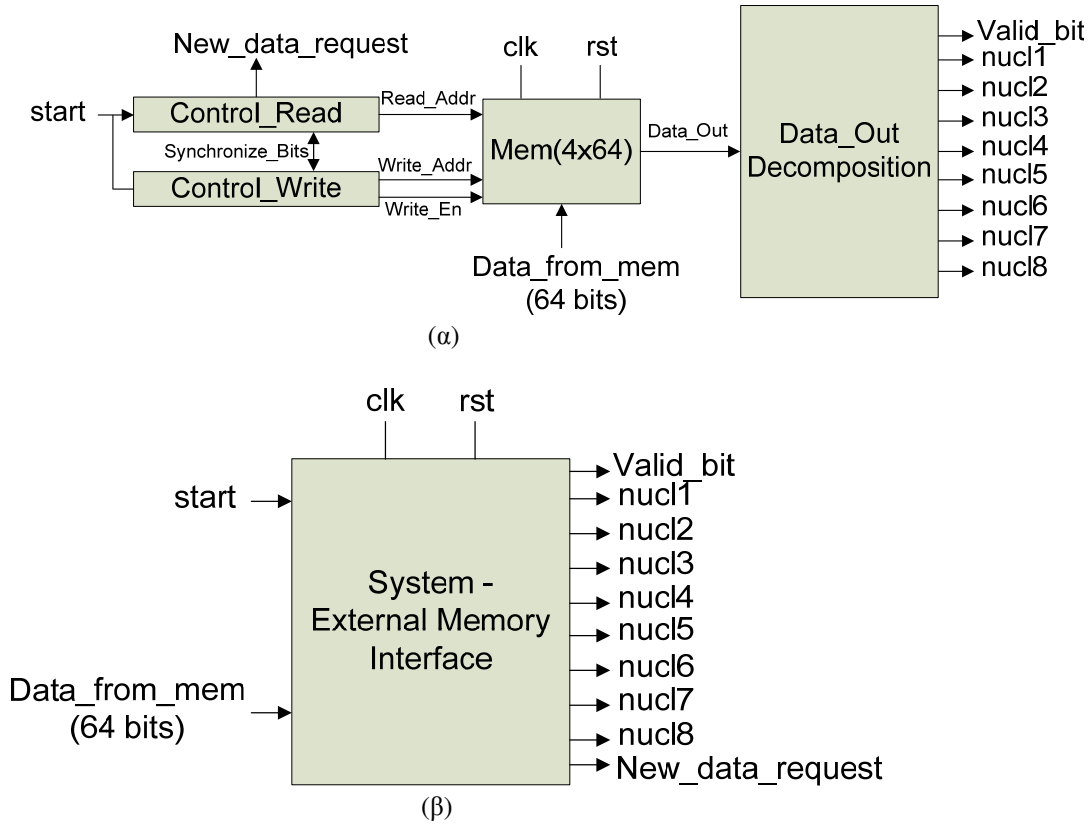
Εικόνα 4-18 : Ενδεικτική μορφή αρχείου PHYLIP που αποτελεί είσοδο στο σύστημα.

Για την αναδιάταξη των δεδομένων ενός αρχείου PHYLIP δημιουργήθηκε ένα πρόγραμμα σε γλώσσα C το οποίο δέχεται ως εισόδο το αρχείο PHYLIP προς αναδιάταξη καθώς και επιπλέον παραμέτρους σχετικά με την FPGA στην οποία θα απεικονιστεί η αρχιτεκτονική. Ανάλογα με το μέγεθος του δέντρου τόσο σε αριθμό ειδών(το πρώτο νούμερο του αρχείου PHYLIP) όσο και σε μήκος ακολουθίας(το δεύτερο νούμερο του αρχείου) καθώς και σε συνδυασμό με την πληροφορία για το μέγεθος της μνήμης BRAM που υπάρχει στην FPGA υπολογίζεται ο βέλτιστος αριθμός στηλών από κάθε ακολουθία νουκλεοτιδίων που θα τοποθετηθούν διαδοχικά στην εξωτερική DRAM. Ο λόγος για τον οποίο είναι απαραίτητη μια τέτοιου είδους αναδιάταξη είναι για να τροφοδοτείται συνεχώς η ομοχειρία αν η μνήμη λειτουργεί σε burst mode.

4.9 Η διεπαφή συστήματος-εξωτερικής μνήμης

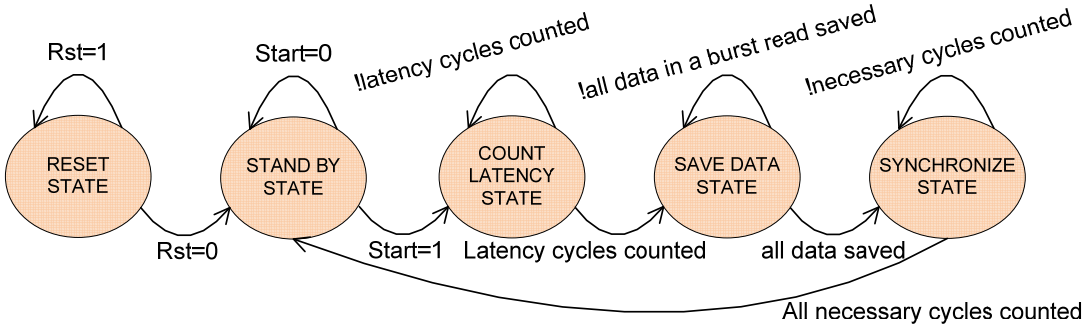
Για να υπάρχουν συνεχώς δεδομένα στην είσοδο του συστήματος και να μην δημιουργούνται καθυστερήσεις αφήνωντας το σύστημα σε αδράνεια περιμένωντας για νέα είσοδο από την εξωτερική μνήμη, δημιουργήθηκε λογική για ανάγνωση δεδομένων από την εξωτερική μνήμη και τροφοδότηση του συστήματος.
Το συγκεκριμένο υποσύστημα, που λειτουργεί ως διεπαφή μεταξύ της σχεδίασης και της εξωτερικής μνήμης, διαβάζει δεδομένα από την DRAM με ρυθμό με τον οποίο μπορεί να τα παρέχει η συγκεκριμένη μνήμη και τα διοχετεύει στην είσοδο του συστήματος με ρυθμό με τον οποίο μπορούν να καταναλωθούν. Για τον λόγο αυτό δημιουργήθηκε μια ουρά $4^{\omega v}$ θέσεων. Τα δεδομένα που έρχονται από την εξωτερική μνήμη αποθηκεύονται στην ουρά με διαφορετικό ρυθμό από ότι διαβάζονται για να τροφοδοτήσουν το σύστημα. Λόγω της χρονικής αυτής ασυνέπειας είναι επιτακτική η ανάγκη συγχρονισμού, διαφορετικά το μέγεθος της ουράς θα έπρεπε να ήταν υπερβολικά μεγάλο για να μην παραληφθούν δεδομένα κατά την μετάβαση από την εξωτερική μνήμη στην είσοδο του συστήματος.

Το υποσύστημα που δημιουργήθηκε για τον σκοπό αυτό φαίνεται στην εικόνα 4-19.

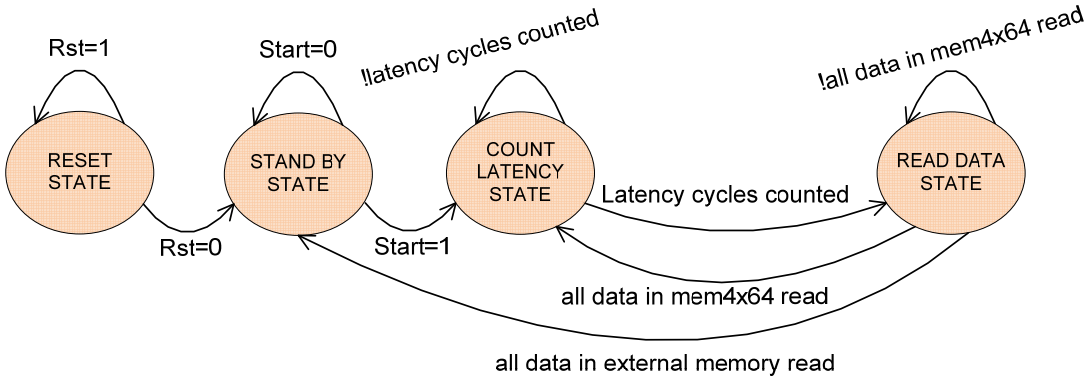


Εικόνα 4-19 : (α) Ουρά Τροφοδοσίας της σχεδίασης με δεδομένα από την εξωτερική μνήμη (β) Η διεπαφή του συγκεκριμένου υποσυστήματος

Οι μονάδες Control_Read και Control_Write είναι μηχανές πεπερασμένων καταστάσεων και χρησιμοποιούνται για τον έλεγχο του ρυθμού ανάγνωσης και εγγραφής δεδομένων στην MEM_4x64 (Control_Write) και τον έλεγχο του ρυθμού ανάγνωσης από την MEM_4x64(Control_Read). Υπάρχει ένα 2bit σήμα επικοινωνίας (Synchronize_Bits) το οποίο χρησιμοποιείται για να ενημερώνεται η μονάδα Control_Read ότι η μονάδα Control_Write έχει αποθηκεύσει τα δεδομένα στην μνήμη και ότι μπορεί να γίνει αίτηση για νέα δεδομένα. Όποτε γίνει νέα αίτηση για δεδομένα από την μονάδα Control_Read προς την εξωτερική μνήμη, η μονάδα Control_Write ενημερώνεται ώστε να είναι σε αναμονή για νέα δεδομένα προς εγγραφή, μέσω του δεύτερου bit του σήματος Synchronize_Bits. Οι μηχανές πεπερασμένων καταστάσεων Control_Read και Control_Write φαίνονται στην εικόνα 4-20.



(α)



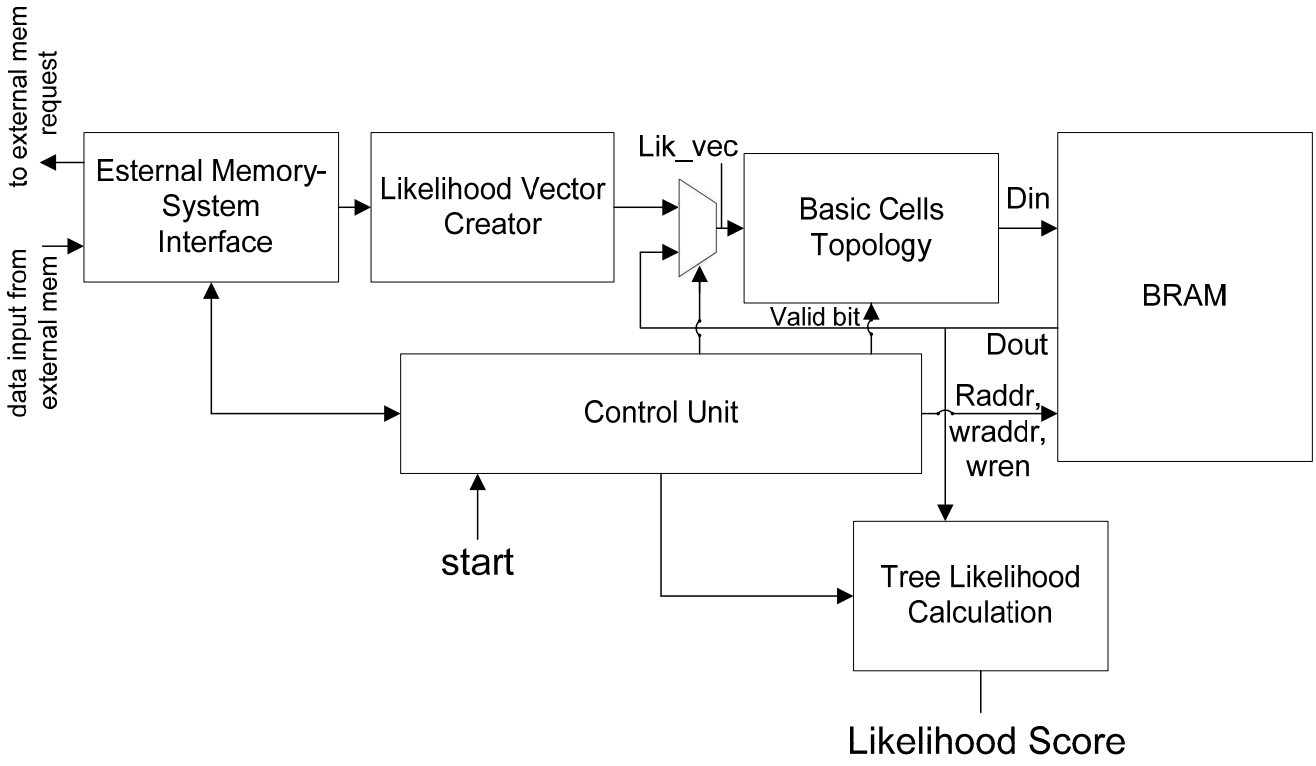
(β)

Εικόνα 4-20 : (α) Η μηχανή πεπερασμένων καταστάσεων της μονάδας Control Write (β) Η μηχανή πεπερασμένων καταστάσεων της μονάδας Control Read

Όπως φάινεται από τις παραπάνω μηχανές πεπερασμένων καταστάσεων, πραγματοποιείται συνεχώς ανάγνωση δεδομένων από την ουρά καθώς αυτός είναι και ο σκοπός παρουσίας της, η συνεχής τροφοδότηση της ομοχειρίας, ενώ η μονάδα Control Write είναι αυτή που συγχρονίζει τις αιτήσεις για νέα δεδομένα από την εξωτερική μνήμη.
Τέλος, η μονάδα Data_Out_Decomposition διασπάει την ακολουθία των 64 bits στα 16 νουκλεοτίδια τα οποία αναπαριστούν. Τα 16 αυτά νουκλεοτίδια συνθέτουν δύο 8άδες οι οποίες θα τροφοδοτήσουν την είσοδο του datapath η μια μετά την άλλη με διαφορά $4^{ων}$ κύκλων.

4.10 Η Συνολική Εικόνα της Αρχιτεκτονικής

Αφού παρουσιάστηκαν και περιγράφηκαν λεπτομερώς τα επιμέρους υποσυστήματα της αρχιτεκτονικής, φαίνεται στην επόμενη εικόνα το συνολικό block diagram.



Εικόνα 4-21 : Block Diagram της αρχιτεκτονικής για τον υπολογισμό της συνάρτησης φυλογενετικής πιθανοφάνειας

5

Υλοποίηση,Ταυτοποίηση και Αποτίμηση Απόδοσης

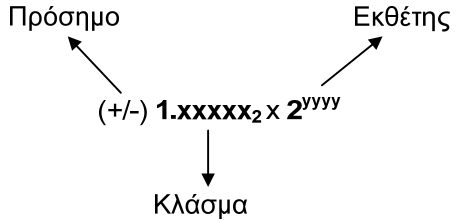
Το παρών Κεφάλαιο περιγράφει τον τρόπο δημιουργίας των βασικών υποσυστημάτων που συνθέτουν την αρχιτεκτονική και παρουσιάζει αποτελέσματα απόδοσης σε σύγκριση με PC.

5.1 Εισαγωγή στην Υλοποίηση της Αρχιτεκτονικής

Η αρχιτεκτονική που παρουσιάστηκε στο Κεφάλαιο 4 είναι ιδιαίτερα απαιτητική σε μνήμη και πράξεις κινητής υποδιαστολής. Εξαιτίας του γεγονότος αυτού επιλέχθηκε η FPGA V5SX240T της οικογένειας Virtex 5 η οποία έχει τον μεγαλύτερο αριθμό από DSP48E slices(1056) τα οποία εκτλελούν πράξεις κινητής υποδιατολής και τον μεγαλύτερο αριθμό από block ram blocks(516 blocks of 36Kb) από όλες τις FPGA που υπάρχουν στην αγορά σήμερα. Χρησιμοποιήθηκε η γλώσσα VHDL και το εργαλείο Xilinx Ise Design Suite 10.1 για την ανάπτυξη του συστήματος.

5.2 Αναπαράσταση Αριθμών Κινητής Υποδιαστολής

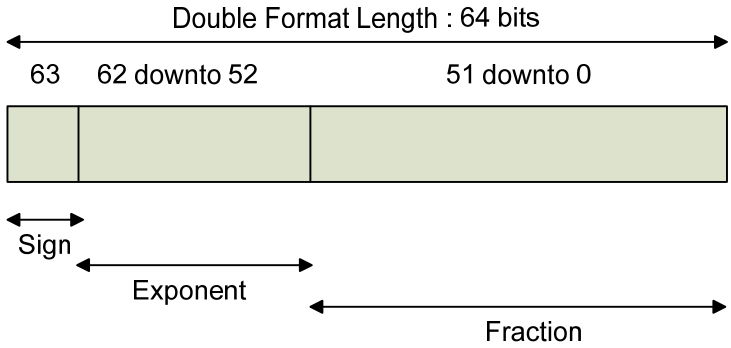
Όλες οι αριθμητικές πράξεις που εκτελούνται στην συγκεκριμένη αρχιτεκτονική είναι πράξεις κινητής υποδιαστολής διπλής ακρίβειας. Οι αριθμοί κινητής υποδιαστολής αποτελούνται από τρία μέρη: το πρόσημο(Sign), το κλάσμα(Fraction) και τον εκθέτη(Exponent) όπως φαίνεται στην εικόνα 5-1.



Εικόνα 5-1 : Περιγραφή δυαδικών αριθμών με κινητή υποδιαστολή.

Η αναπαράσταση των αριθμών αυτών ακολουθεί το πρότυπο IEEE-754. Το συγκεκριμένο πρότυπο χρησιμοποιείται σχεδόν από όλους τους υπολογιστές που σχεδιάστηκαν από το 1980 και μετά. Η αρχική μονάδα ενός κανονικοποιημένου αριθμού

παραλείπεται κατά την αναπαράσταση σε bits. Το πεδίο του κλάσματος περιέχει μόνο το κλασματικό μέρος. Κατά το πρότυπο IEEE-754, οι αριθμοί διπλής ακρίβειας χρησιμοποιούν 11 bits για τον εκθέτη και 52 bits για το κλάσμα. (Εικόνα 5-2).



Εικόνα 5-2 : Αναπαράσταση αριθμών κινητής υποδιαστολής διπλης ακρίβειας (IEEE-754 Standard)

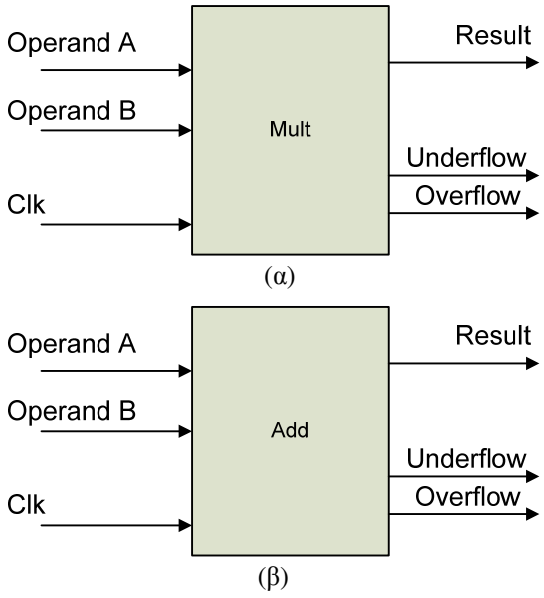
Αν S το πρόσημο, E ο εκθέτης και F το κλάσμα, τότε η δεκαδική τιμή ενός αριθμού κινητής διαστολής δίνεται από τον τύπο:

$$V= (-1)^S \times (1+ F) \times 2^E$$

Το μηδέν αναπαρίσταται ως $0\ldots00_2$ και όχι σε μορφή κανονικοποιημένου επιστημονικού συμβολισμού.

5.3 Πολλαπλασιαστής και Αθροιστής Κινητής Υποδιαστολής Διπλής Ακρίβειας

Οι πολλαπλασιαστές και οι αθροιστές που χρησιμοποιήθηκαν στην συγκεκριμένη σχεδίαση δημιουργήθηκαν με το εργαλείο Xilinx Core Generator 10.1. Εκτελούν πράξεις κινητής υποδιαστολής διπλής ακρίβειας. Η διεπαφές του πολλαπλασιαστή και του αθροιστή φαίνονται στην εικόνα 5-3.



Εικόνα 5-3 : (α) Διεπαφή Πολλαπλασιαστή, (β) Διεπαφή Αθροιστή

Στους πίνακες 5-1 και 5-2 φαίνονται τα σήματα εισόδου/εξόδου για τον πολλαπλασιαστή και τον αθροιστή αντίστοιχα, καθώς και περιγραφή της λειτουργίας τους.

Signal	Width(bits)	Type	Description
Operand A	64	Input	InputMult Operand A
Operand B	64	Input	Input Mult Operand B
Clk	1	Input	Clock Signal
Result	64	Output	Mult Result (AxB)
Underflow	1	Output	Signal to indicate arithmetic underflow
Overflow	1	Output	Signal to indicate arithmetic overflow

Πίνακας 5-1 : Περιγραφή σημάτων εισόδου εξόδου του πολλαπλασιαστή.

Signal	Width(bits)	Type	Description
Operand A	64	Input	Input Add Operand A
Operand B	64	Input	Input Add Operand B
Clk	1	Input	Clock Signal
Result	64	Output	Add Result (AxB)
Underflow	1	Output	Signal to indicate arithmetic underflow
Overflow	1	Output	Signal to indicate arithmetic overflow

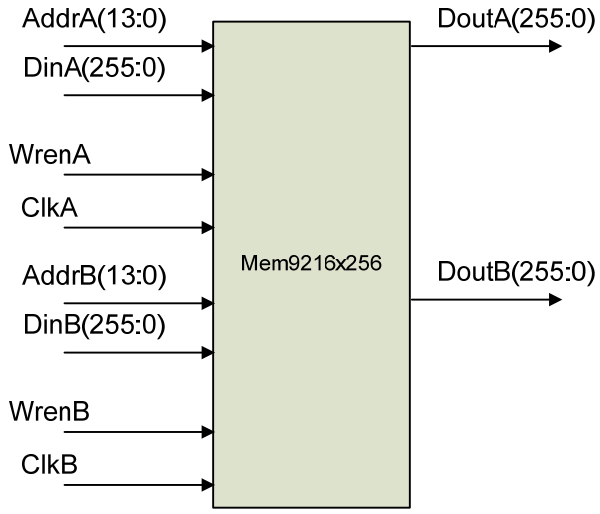
Πίνακας 5-2 : Περιγραφή σημάτων εισόδου εξόδου του αθροιστή.

Κάθε πολλαπλασιαστής χρησιμοποιεί 9 DSP48E slices με σκοπό να αποφευχθεί το ενδεχόμενο να επηρεάσει το κρίσιμο μονοπάτι(critical path). Έχει latency 15 κύκλων και λειτουργεί ομόχειρα δίνωντας ένα γινόμενο σε κάθε κύκλο(throughput=1αποτέλεσμα/κύκλο). Κάθε αθροιστής χρησιμοποιεί 3 DSP48E slices. Έχει latency 14 κύκλων και λειτουργεί επίσης ομόχειρα δίνωντας ένα άθροισμα σε κάθε κύκλο(throughput=1αποτέλεσμα/κύκλο).
Συνολικά, χρησιμοποιήθηκαν 91 πολλαπλασιαστές και 54 αθροιστές για την υλοποίηση της αρχιτεκτονικής. Έτσι λοιπόν, το 77.5% των διαθέσιμων DPS48E slices καταναλώθηκαν σε πολλαπλασιασμούς, ενώ το 15,3% σε αθροίσεις.

5.4 Μνήμη BRAM

Ο συνολικός αριθμός των block ram blocks που υπάρχουν στην FPGA που επιλέχθηκε είναι 1032 blocks των 18Kb. Το σύνολο των blocks οργανώθηκε σε 8 μικρότερες μνήμες βάθους 9216 θέσεων των 256bits/γραμμή μνήμης καταναλώνοντας 129blocks. Χρησιμοποιήθηκε και πάλι το εργαλείο Xilinx Core Generator 10.1 για την οργάνωση της μνήμης στα 8 αυτά κομμάτια. Οι μνήμες που δημιουργήθηκαν είναι true dual port και η διεπαφή τους φαίνεται στην εικόνα 5-4. Τα

256 bits σε κάθε γραμμή της μνήμης αναπαριστούν ένα διάνυσμα πιθανοφανειών(4x64bits).



Εικόνα 5-4 : Διεπαφή True Dual Port Μνήμης 9216x256

Η χρήση μνημών dual port είναι επιτακτική καθώς επηρεάζει άμεσα την απόδοση. Κατά την πρώτη φάση λειτουργίας χρησιμοποιείται μόνο η μια θύρα και πραγματοποιούνται οι πρώτες εγγραφές στις μνήμες. Κατά την δεύτερη φάση λειτουργίας χρησιμοποιούνται και οι δύο θύρες. Η μια θύρα χρησιμοποιείται για ανάγνωση δεδομένων από τις μνήμες ενώ η δεύτερη για εγγραφή δεδομένων στις μνήμες. Κατά την εγγραφή, υπερεγγράφονται θέσεις μνήμης που έχουν διαβαστεί k*4 κύκλους νωρίτερα, όπου k είναι το latency του μονοπατιού δεδομένων το οποίο είναι 176 κύκλοι και η παρουσία του αριθμού 4 δικαιολογείται καθώς το μονοπάτι δεδομένων καταναλώνει καινούργια δεδομένα κάθε 4 κύκλους. Κατά την τρίτη φάση λειτουργίας και οι δύο θύρες χρησιμοποιούνται για ανάγνωση, όπως περιγράφηκε στο Κεφάλαιο 4, κατά τον υπολογισμό του τελικού βαθμού πιθανοφάνειας του δέντρου από τα διανύσματα πιθανοφανειών της ρίζας.

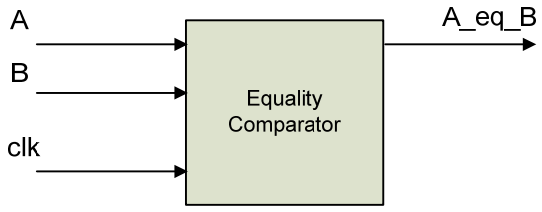
Στον πίνακα 5-3 φαίνονται τα σήματα εισόδου/εξόδου για τις μνήμες που χρησιμοποιήθηκαν, καθώς και περιγραφή της λειτουργίας τους.

Signal	Width(bits)	Type	Description
AddrA	14	Input	Input Address for Port A
AddrB	14	Input	Input Address for Port B
DinA	256	Input	Input Data for Port A
DinB	256	Input	Input Data for Port B
ClkA	1	Input	Clock Signal for Port A
ClkB	1	Input	Clock Signal for Port B
WrEnA	1	Input	Write Enable Signal for Port A
WrEnB	1	Input	Write Enable Signal for Port B
DoutA	256	Output	Ouput Data Read from Port A
DoutB	256	Output	Output Data Read from Port B

Πίνακας 5-3 : Περιγραφή σημάτων εισόδου εξόδου της μνήμης.

5.5 Συγκριτής Ισότητας

Με την χρήση του εργαλείου Xilinx Core Generator 10.1 δημιουργήθηκαν οι συγκριτές που χρησιμοποιήθηκαν κατά την υλοποίηση. Χρησιμοποιήθηκαν κυρίως για να εντοπίζονται οι κύκλοι που η κάθε FSM πρέπει να αλλάξει κατάσταση ελέγχωντας αν ισχύει ισότητα μεταξύ των εισόδων. Σε όλες τις περιπτώσεις που χρησιμοποιήθηκαν συγκριτές, η μια είσοδος παραμένει σταθερή και η άλλη αλλάζει. Μόλις παρατηρηθεί ισότητα ενημερώνεται κάποια FSM η οποία μπορεί να αλλάξει πλέον την είσοδο, η οποία παραμένει σταθερή για την επόμενη φορά που θα χρησιμοποιηθέι ο συγκριτής. Η διεπαφή του συγκριτή φαίνεται στην εικόνα 5-5 ενώ στον πίνακα 5-4 φαίνονται τα σήματα εισόδου εξόδου και περιγραφή της λειτουργίας τους.



Εικόνα 5-5 : Διεπαφή Συγκριτή Ισότητας

Signal	Width(bits)	Type	Description
A	32	Input	Input Operand A
B	32	Input	Input Operand B
Clk	1	Input	Clock Signal
A_eq_B	1	Output	Output Signal to Indicate Equality

Πίνακας 5-4 : Περιγραφή σημάτων εισόδου εξόδου του συγκριτή ισότητας.

Ο παραπάνω συγκριτής, συγκρίνει 32bit ποσότητες ελέγχωντας αν ταυτίζονται. Σε περίπτωση ισότητας ενεργοποιεί το σήμα A_eq_B με καθυστέρηση δύο κύκλων. Ο λόγος για τον οποίο χρησιμοποιήθηκε συγκριτής με latency κύκλων είναι για να αποφευχθεί το ενδεχόμενο ο συγκριτής να επηρεάσει το critical path. Όπως είναι όμως αναμενόμενο, η χρήση του συγκεκριμένου συγκριτή οδηγεί σε καθυστέρηση δύο κύκλων κάθε φορά που αναμένεται ένα σήμα ταύτισης από τον συγκεκριμένο συγκριτή. Ενώ δλδ. τα σήματα στην είσοδο του συγκριτή πληρούν το κριτήριο ισότητας , το σύστημα θα ενημερωθεί δύο κύκλους αργότερα ώστε να ξεκινήσει την προγραμματισμένη λειτουργία του. Φυσικά κάτι τέτοιο δεν είναι αποδεχτό και το συγκεκριμένο πρόβλημα υπερκεράστηκε δίνωντας ως σταθερή είσοδο στους συγκριτές μια τιμή μικρότερη κατά δύο μονάδες από την ζητούμενη. Αυτό γενικά δεν είναι εφικτό σε οποιαδήποτε σχεδίαση, αλλά στην συγκεκριμένη αρχιτεκτονική, όλοι σχεδόν οι συγκριτές εντοπίζουν κύκλους ρολογιού, που σημαίνει ότι το μεταβλητό σήμα στην είσοδο του συγκριτή αλλάζει σε κάθε κύκλο ρολογιού. Δεν χρησιμοποιούνται όμως όλοι οι συγκριτές για εντοπισμό των κύκλων. Παρ'όλα αυτά δεν δημιουργείται καμμία επιπλοκή στην σωστή λειτουργία του συστήματος

καθώς τα αποτελέσματα των συγκριτών πάντα αξιοποιούνται σε συγκεκριμένες καταστάσεις μηχανών πεπερασμένων καταστάσεων ενώ διαφορετικά αγνοούνται.

5.6 Ταυτοποίηση λειτουργίας

Η λειτουργία της αρχιτεκτονικής προσομοιώθηκε με το πρόγραμμα ModelSimSE 6.3f. Αρχικά προσομοιώθηκε η λειτουργία των επιμέρους υποσυστημάτων της αρχιτεκτονικής και διαπιστώθηκε η σωστή λειτουργία τους. Έπειτα ολοκληρώθηκε το σύστημα, συνδέωντας τα επιμέρους βασικά υποσυστήματα και προσομοιώθηκε η λειτουργία του.

Για την διαπίστωση της σωστής λειτουργίας χρησιμοποιήθηκαν τρία αρχεία PHYLIP και συγκρίθηκαν τα αποτελέσματα της προσομοίωσης με τα αποτελέσματα του πρόγραμματος RAxML Light.Το συγκεκριμένο πρόγραμμα δημιουργήθηκε από τον κ. Σταματάκη με σκοπό την ταυτοποίηση της συγκεκριμένης αρχιτεκτονικής και αποτελεί ένα μικρό και ελαφρώς διαμορφωμένο κομμάτι του προγράμματος RAxML 7.0.4. Το RAxML Light υπολογίζει το likelihood score ενός δέντρου για σταθερό μήκος κλαδιών.
Το πρώτο αρχείο PHYLIP που υπολογίστηκε περιέχει δέντρο με 8 taxa και μήκος ακολουθίας 705 νουκλεοτίδια. Το δεύτερο αρχείο PHYLIP περιέχει δέντρο με 64 taxa και μήκος ακολουθίας 1781 νουκλεοτίδια. Το τρίτο αρχείο PHYLIP περιέχει δέντρο με 512 taxa και μήκος ακολουθίας 1152 νουκλεοτίδια. Τα συγκεκριμένα αρχεία δώθηκαν από τον κ.Σταματάκη, δημιουργό του προγράμματος RAxML.

Τα πειράματα που διεξήχθησαν σε προσομοίωση με σκοπό την διαπίστωση της σωστής λειτουργίας της αρχιτεκτονικής.

Πείραμα	Αριθμός Ειδών	Μήκος Ακολουθίας
Α	8	705
Β	64	1781
Γ	512	1152

Πίνακας 5-5 : Τα πειράματα που διεξήχθησαν για την ταυτοποιήση της αρχιτεκτονικής

5.7 Απόδοση Συστήματος

Το τελικό σύστημα που δημιουργήθηκε αποτελείται από 7 βασικές υπολογιστικές μονάδες σε δενδρική τοπολογία όπως παρουσιάστηκε στο Κεφάλαιο 4. Πραγματοποιήθηκε Synthesis με το εργαλείο Xilinx Ise 10.1 και προέκυψαν τα ακόλουθα αποτελέσματα, όσο αφορά την συχνότητα ρολογιού και τους πόρους που καταναλώνει η συγκεκριμένη αρχιτεκτονική.

Clock Frequency : 284,152 MHz.

Device Utilization Summary (xc5vsx240t-2ff1738)			
Logic Utilization	Used	Available	Utilization
Number of Slice Registers	113269	149760	75%
Number of Slice LUTs	90643	149760	60%
Number of fully used LUT-FF pairs	76872	127040	60%
Number of bonded IOBs	132	960	13%
Number of Block RAM/FIFO	516	516	100%
Number of BUFG/BUFGCTRLs	2	32	6%
Number of DSP48Es	981	1056	92%

Πίνακας 5-6 : Χρησιμοποίηση πόρων της FPGA για την υλοποίηση της αρχιτεκτονικής

Επίσης, έγινε υλοποίηση μόνο της βασικής υπολογιστικής μονάδας, σε μια αρκετά μικρότερου μεγέθους FPGA όπως είναι η V4vfx60. Έγινε Synthesis με το εργαλείο Xilinx Ise 10.1 και προέκυψαν τα ακόλουθα αποτελέσματα, όσο αφορά τη συχνότητα ρολογιού και τους πόρους που καταναλώνει μόνο η βασική υπολογιστική μονάδα.

Clock Frequency : 267,386 MHz.

Device Utilization Summary (xc4vfx60-12ff1152)			
Logic Utilization	Used	Available	Utilization
Number of Slice Registers	13970	25280	55%
Number of Slice LUTs	22572	50560	44%
Number of fully used LUT-FF pairs	20107	50560	39%
Number of bonded IOBs	73	576	12%
Number of GCLKs	1	32	3%
Number of DSP48Es	99	128	77%

Πίνακας 5-7 : Χρησιμοποίηση πόρων μικρού μεγέθους FPGA για την υλοποίηση της βασικής μονάδας

5.8 Αποτίμηση Απόδοσης

Το πρόγραμμα RAxML Light εκτέλεστηκε σε PC με επεξεργαστή Intel Pentium 4 στα 2,66 GHz, με 1GB RAM και λειτουργικό σύστημα Ubuntu 8.04. Το συγκεκριμένο πρόγραμμα εκτελέστηκε περίπου 300 φορές (10 φορές για 30 διαφορετικά σετ δεδομένων)και χρησιμοποιήθηκε το πρόγραμμα Vtune για να μετρηθεί ο χρόνος CPU. Το 1/3 των σετ δεδομένων είναι πραγματικά δεδομένα που δώθηκαν από τον κ.Σταματάκη και αντιστοιχούν στην μέση περίπτωση εκτέλεσης του προγράμματος RAxML Light. Επίσης, δημιουργήθηκαν μη πραγματικά δεδομένα ώστε να προσσεγγιστεί τόσο η χειρότερη περίπτωση εκτέλεσης όσο και η βέλτιστη.
Για την χειρότερη περίπτωση εκτέλεσης του προγράμματος RAxML Light η οποία αντιστοιχεί σε εντελώς αποτυχημένη πολλαπλή ταύτιση του αρχείου εισόδου

εκτελέστηκαν τα πειράματα 1-10 των οποίων τα μεγέθη των σετ δεδομένων φαίνονται στον πίνακα 5-8. Ακριβώς ίδιου μεγέθους δεδομένα χρησιμοποιήθηκαν για να μετρηθεί και η μέση περίπτωση και η βέλτιστη περίπτωση, με διαφορετικές πολλαπλές ταυτίσεις.

Πείραμα	Αριθμός Ειδών	Μήκος Ακολουθίας
1	4	1000
2	8	1000
3	9	1000
4	16	1000
5	32	1000
6	64	1000
7	65	1000
8	128	1000
9	256	1000
10	512	1000

Πίνακας 5-8 : Πειράματα που δοκιμάστηκαν για αποτίμηση της απόδοσης για την χειρότερη, μέση και βέλτιστη περίπτωση εκτέλεσης του προγράμματος RAxML Light.

Ο πίνακας 5-9 που φαίνεται στην επόμενη σελίδα παρουσιάζει το σύνολο των μετρήσεων που έγιναν. Τα αποτελέσματα που παρουσιάζονται αποτελούν τον μέσο όρο 10 μετρήσεων για κάθε περίπτωση.

Χρόνος Εκτέλεσης CPU σε P4 (secs)

Πείραμα	Χειρότερη Περίπτωση	Μέση Περίπτωση	Βέλτιση Περίπτωση
1	0,006015	0,0050113	0,004259
2	0,007016	0,0060135	0,004886
3	0,007016	0,0060135	0,004886
4	0,021047	0,0130293	0,011025
5	0,029065	0,0240541	0,012027
6	0,052117	0,0380857	0,013029
7	0,052117	0,0380857	0,013029
8	0,092208	0,0721624	0,018041
9	0,171386	0,1513406	0,033074
10	0,347783	0,3269842	0,067151

Πίνακας 5-9 : Χρόνοι εκτέλεσης των πειραμάτων σε Pentium 4

Έπειτα, μετρήθηκαν οι απαιτούμενοι κύκλοι ρολογιού που χρειάστηκε το σύστημα που υλοποιήθηκε στην FPGA για να επεξεργαστεί τον όγκο δεδομένων του κάθε πειράματος και από το σύνολο των κύκλων υπολογίστηκε ο εκτιμώμενος χρόνος εκτέλεσης.

Πείραμα	Αριθμός Κύκλων	Υπολογισμένος Χρόνος Εκτέλεσης(secs)
1,2	18376	0,000064704
3,4,5,6	147156	0,000518155
7,8,9,10	1171936	0,004126535

Πίνακας 5-10 : Κύκλοι προσομοιωμένης εκτέλεσης των πειραμάτων σε FPGA και εκτιμώμενοι χρόνοι εκτέλεσης

Πρέπει να σημειωθεί ότι το σύστημα που σχεδιάστηκε και υλοποιήθηκε στην FPGA δεν αντιμετωπίζει πρόβλημα εισόδου/εξόδου. Οι απαιτήσεις της αρχιτεκτονικής για είσοδο είναι 8bits/cycle δλδ. 2,273Gbps. Η FPGA που χρησιμοποιήθηκε για την υλοποίηση του συστήματος έχει 24 RocketIO GTP Tranceivers σχεδιασμένους για μέγιστο ρυθμό μεταφοράς δεδομένων 3,75Gbps, που σημαίνει ότι οι ανάγκες για είσοδο καλύπτονται. Ακόμη, μπορεί να χρησιμοποιηθεί η έτοιμη λύση που παρέχεται από την FPGA των 4^{ων} Ethernet Macs. Αν χρησιμοποιηθούν 3 από τα 4 links με 1Gbps το καθένα, επίσης οι ανάγκες του συστήματος για είσοδο καλύπτονται. Αυτό συμβαίνει καθώς έχουμε point to point σύνδεση και η μεταφορά δεδομένων είναι μονόδρομη, μόνο προς την FPGA, και άρα δεν παρουσιάζονται «συγκρούσεις» δεδομένων.

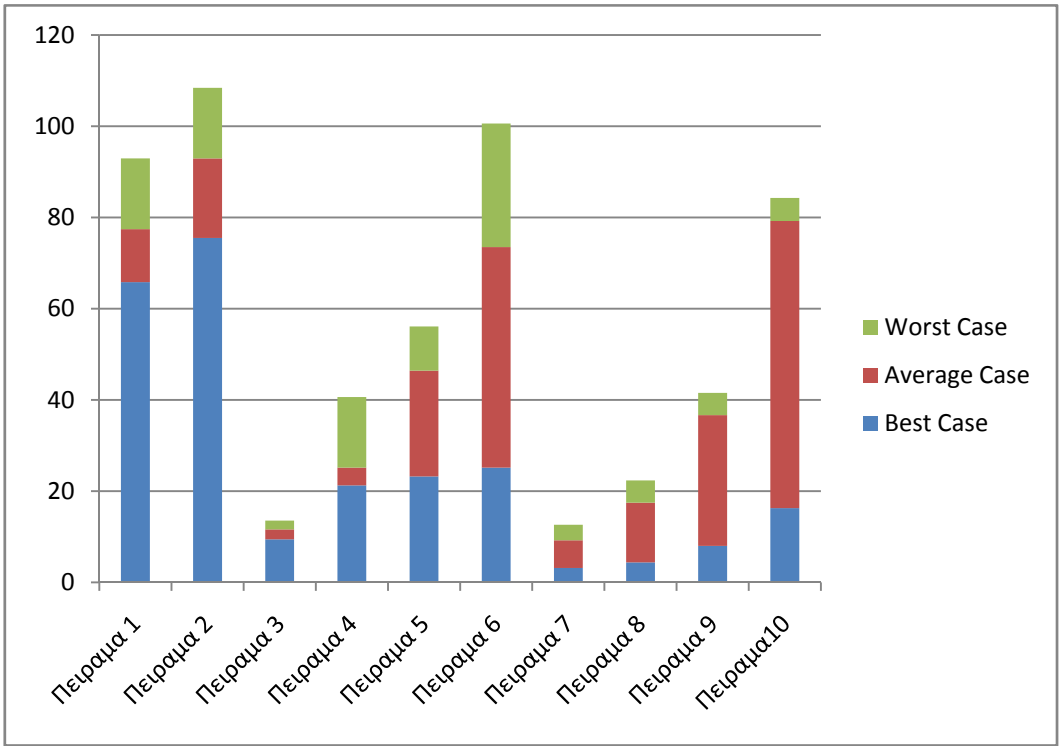
Τέλος, συνδυάζοντας τις πληροφορίες των παραπάνω πινάκων προκύπτει ο πίνακας που φαίνεται στην επόμενη σελίδα για την επιτάχυνση (speed up) που πραγματοποιείται από την χρήση FPGA αντί για Pentium 4 για το συγκεκριμένο πρόβλημα.

Πείραμα	Χειρότερη Περίπτωση	Μέση Περίπτωση	Βέλτιση Περίπτωση
1	92,96	77,45	65,83
2	108,43	92,94	75,51
3	13,54	11,61	9,43
4	40,62	25,15	21,28
5	56,09	46,42	23,21
6	100,58	73,50	25,15
7	12,63	9,23	3,16
8	22,35	17,49	4,37
9	41,53	36,67	8,02
10	84,30	79,24	16,27

Πίνακας 5-11 : Επιτάχυνση με τη χρήση FPGA vs P4 για τα παραπάνω πειράματα.

Στην επόμενη εικόνα φαίνεται το διάγραμμα που δείχνει τις επιταχύνσεις για τα 10 πειράματα που προσομοιώθηκαν για τις τρείς περιπτώσεις εκτέλεσης του προγράμματος RAxML Light .

Επιτάχυνση (x) FPGA V5 vs Pentium 4 για τα πειράματα



Εικόνα 5-6 : Διάγραμμα με τις επιταχύνσεις που μετρήθηκαν για τα 10 πειράματα και τις τρείς περιπτώσεις εκτέλεσης.

Αρχικά, πρέπει να σημειωθεί ότι, η καλύτερη περίπτωση δεν είναι η ίδια στην αρχιτεκτονική σε FPGA και στο RAxML Light. Το μέγεθος των δεδομένων προς επεξεργασία μειώνεται σημαντικά στην εκτέλεση του RAxML Light λόγω βελτιστοποιήσεων. Η εκτέλεση του συκγεκριμένου προγράμματος εξαρτάται άμεσα από τις στήλες νουκλεοτιδίων στο αρχείο εισόδου. Βέλτιστη περίπτωση εκτέλεσης για το σύστημα σε FPGA παρατηρείται όταν ο αριθμός των ταξινομικών λειτουργικών μονάδων του δέντρου που υπολογίζεται είναι δύναμη του 8. Αυτό συμβαίνει εξαιτίας της δενδρικής διάταξης των βασικών υπολογιστικών μονάδων.
Αντίστοιχα, η χειρότερη περίπτωση εκτέλεσης επίσης δεν είναι η ίδια στο σύστημα στην FPGA και στο RAxML Light. Χειρότερη περίπτωση για το σύστημα σε FPGA είναι ο υπολογισμός δέντρων με αριθμό ταξινομικών λειτουργικών μονάδων k*8+1, όπου k =0,1,2…
Τα πειράματα 2, 6 και 10 , των οποίων τα σετ δεδομένων είχαν 8, 64 και 512 ταξινομικές λειτουργικές μονάδες αντίστοιχα, αποτελούν την βέλτιστη περίπτωση για το σύστημα στην FPGA, γι' αυτό και το speed up είναι μεγαλύτερο απ' ότι στα υπόλοιπα πειράματα για όλες τις περιπτώσεις εκτέλεση του προγράμματος RAxML Light.

Τα πειράματα 3 και 7, των οποίων τα σετ δεδομένων είχαν 9 και 65 ταξινομικές λειτουργικές μονάδες αντίστοιχα αποτελούν την χείριστη περίπτωση εκτέλεσης για την σχεδίαση σε FPGA, γι'αυτό παρατηρούνται τόσο χαμηλές επιταχύνσεις. Τα υπόλοιπα πειράματα πλησιάζουν στην εκτέλεση μέσης περίπτωσης τόσο για το

σύστημα σε FPGA και για κάθε περίπτωση εκτέλεση του RAxML Light φαίνονται οι επιταχύνσεις. Αξίζει να σημειωθεί ότι ακόμη και στην βέλτιση περίπτωση για το RAxML Light η οποία όμως αποτελούσε την χειρότερη περίπτωση για το σύστημα στην FPGA μετρήθηκε speed up 3,16x.

6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Το τελευταίο Κεφάλαιο αναφέρει συμπεράσματα και μελλοντικές επεκτάσεις της αρχιτεκτονικής.

6.1 Συμπεράσματα

Η παρούσα διπλωματική διατριβή παρουσίασε την δημιουργία συστήματος σε αναδιατασσόμενη λογική το οποίο υπολογίζει τον βαθμό πιθανοφάνειας φυλογενετικών δέντρων πλήρως ισορροπημένων και για σταθερό μήκος κλαδιών, με τρόπο πιο αποδοτικό απ'ότι ένα PC με επεξεργαστή Pentium 4 στα 2,66GHz.

6.2 Μελλοντικές Επεκτάσεις

Η αρχιτεκτονική που παρουσιάστηκε δημιουργήθηκε με τέτοιο τρόπο ώστε να είναι δυνατόν να επεκταθεί εύκολα προς συγκεκριμένες κατευθύνσεις. Σημαντικές επεκτάσεις που μπορούν να πραγματοποιηθούν είναι :

i) Επέκταση της μονάδας ελέγχου με σκοπό την υποστήριξη δέντρων μη ισορροπημένων καθώς και λογικής για την μετάφραση της κωδικοποίησης των δέντρων.

ii) Δημιουργία υπολογιστικής μονάδας που να παράγει τους πίνακες P έτσι ώστε να υποστηρίζονται διαφορετικά μήκη κλαδιών. Η δημιουργία της συγκεκριμένης μονάδας έχει επιπλέον βιολογική σημασία πέραν της προφανούς καθώς η συγκεκριμένη αρχιτεκτονική θα μπορούσε να χρησιμοποιηθεί για αξιολόγηση αλλαγών σε παραμέτρους του μοντέλου.

iii) Προσαρμογή της προαναφερθείσης υπολογιστικής μονάδας με σκοπό να υποστηρίζονται περισσότερα από ένα μοντέλα υποκατάστασης.

iv) Δημιουργία κατάλληλου interface ώστε να μπορεί η αρχιτεκτονική να «επικοινωνήσει» τόσο με το πρόγραμμα RAxML αλλά και γενικότερα με όλα τα παραρεμφερή προγράμματα που χρησιμοποιούν την συνάρτηση φυλογενετικής πιθανοφάνειας.

Βιβλιογραφία

1. Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Maximum Likelihood Method. Alexandros Stamatakis . Phd Thesis.

2. Stamatakis, T. Ludwig, H. Meier. Parallel Inference of a 10.000-taxon Phylogeny with Maximum Likelihood. In *Proceedings of the 10th International Euro Par Conference (Euro-Par 2004)*, Springer Verlag, 997-1004, Pisa, Italy, September 2004.

3. Stamatakis, T. Ludwig, H. Meier. New Fast and Accurate Heuristics for Inference of Large Phylogenetic Trees. In *Proceedings of 18th International Parallel and Distributed Processing Symposium (IPDPS2004)*, Proceedings on CD, Abstract on page 193, Santa Fe, NewMexico, April 2004.

4. Stamatakis, T. Ludwig, H. Meier. A Fast Program for Phylogenetic Tree Inference with Maximum Likelihood. In *Arndt Bode, Franz Durst, Werner Hanke, and Siegfried Wagner, editors, High Performance Computing in Science and Engineering*, Springer Verlag, 273-284, 2004.

5. Stamatakis, T. Ludwig, H. Meier. RAxML: A Parallel Program for Phylogenetic Tree Inference. Poster abstract in *Proceedings of 2nd European Conference on Computational Biology (ECCB2003)*, 325–326, Paris, France, September 2003.

6. L. Jermiin et al. Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. *Mol. Biol. Evol.*,1296-1302, 1997.

7. Filip Blagojevic, Dimitrios S. Nikolopoulos, Alexandros Stamatakis, Christos D. Antonopoulos. *"RAxML-Cell: Parallel Phylogenetic Tree Inference on the Cell Broadband Engine".* In Proceedings of 21st IEEE International Parallel & Distributed Processing Symposium (IPDPS2007), Proceedings on CD, 327-346, Long Beach, California, USA, March 2007.

8. Filip Blagojevic, Dimitrios S. Nikolopoulos, Alexandros Stamatakis, and Christos D. Antonopoulos: "*Dynamic Multigrain Parallelization on the Cell Broadband Engine"*. In Proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'07), 90-100, San Jose, California, USA, March 2007.

9. Alexandros Stamatakis, Thomas Ludwig, and Harald Meier: ”*RAxML-III: A Fast Program for Maximum Likelihood-based Inference of Large Phylogenetic Trees.*” In Bioinformatics 21(4):456-463, 2005.

10. Craig A. Stewart1, David Hart1, Donald K. Berry1, Gary J. Olsen2, Eric A. Wernert1, William Fischer: *”Parallel implementation and performance of fastDNAml - a program for maximum likelihood phylogenetic inference”,20-20,* Denver,November 2001.

11. Introduction to Character and Parsimony Analysis.(pdf), Embnet, Διαθέσιμο στο: www.ch.embnet.org/CoursEMBnet/PHYL03/Slides/characters_mwilkinson.pdf , visited Mar 2008

12. Gokul Govindu, Ling Zhuo, Seonil Choi and Viktor Prasanna: “*Analysis of High-performance Floating-point Arithmetic on FPGAs”,149,* 2004

13. In Jae Myung : “*Tutorial on maximum likelihood estimation*.”, Journal of Mathematical Psychology*, 47*, 90-100, Ohio, 2002.

14. Dennis Pearl : “Maximum Likelihood Tree Construction” (lecture) September,2005. Διαθέσιμο στο: http://mbi.osu.edu/2005/tutorialmaterials/MBImle.pdf, visited Mar 2008

15. Alexandros Stamatakis, Michael Ott, and Thomas Ludwig: *"RAxML-OMP: An Efficient Program for Phylogenetic Inference on SMPs"* Proceedings of 8th International Conference on Parallel Computing Technologies (PaCT 2005), Lecture Notes in Computer Science, 3606:288-302, Sep 2005

16. A. Stamatakis, T. Ludwig, H. Meier, M.J. Wolf. Accelerating Parallel Maximum Likelihood-based Phylogenetic Tree Calculations using Subtree Equality Vectors. In *Proceedings of 15th Supercomputing Conference(SC2002)*,1-16, Proceedings on CD, Baltimore, Maryland, November 2002.

17. Inferring Phylogenies. Joseph Felsenstein. University of Washigton. Sinauer Associates, Inc.USA.2004

18. Molecular Systematics-Second Edition . David M.Hills, Craig Moritz, Barbara K.Bable. Sinauner Associates,Inc USA.1996

19. Εισαγωγή στην Εξέλιξη. Σταμάτης Ν.Αλαχιώτης. Α.Α.Λιβάνη, Αθήνα, 2007

20. Joseph Felsenstein:**“**Evolutionary Trees from DNA Sequences:A Maximum Likelihood Approach.”, Journal of Molecular Evolution, 368-376, USA,2005.

21. Museum of Paleontology-University of California, site: http://www.ucmp.berkeley.edu/exhibits/historyoflife.php , visited: Aug 2008

22. Understading Evolution- Berkeley, site: http://evolution.berkeley.edu/evolibrary/home.php , visited Jul 2008

23. Division of Biology and Medicine, Brown University, site: http://biomed.brown.edu/Courses/BIO48/25.Inference.HTML , visited Apr 2008

24. College of Chemical and Life Sciences, The Delwiche Lab(Molecular Systematics),University of Maryland, site: http://www.life.umd.edu/labs/delwiche/bsci348s/lec/Phylogenetics1.html , visited Apr 2008

25. ClustalW project site: http://www.ebi.ac.uk/Tools/clustalw2/index.html , visited Jun 2008