ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΔΙΚΤΥΩΝ

ΜΕΛΕΤΗ ΕΞΕΛΙΚΤΙΚΩΝ ΚΑΙ ΓΕΝΕΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ ΜΕ ΕΦΑΡΜΟΓΗ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ

Επιβλέπων καθηγητής: Μποζάνης Παναγιώτης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μπαλάση Παναγιώτα ΑΕΜ 568

Στην οικογένεια μου και τους φίλους μου

Ευχαριστίες

Με την περάτωση αυτής της διπλωματικής εργασίας θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της εργασίας κ. Μποζάνη Παναγιώτη για την εμπιστοσύνη που έδειξε στο πρόσωπό μου καθώς και τη στήριξη και τη βοήθεια που μου παρείχε αυτά τα έξι χρόνια των σπουδών μου.
Επιπλέον, θα ήθελα να ευχαριστήσω τους φίλους μου για την ευχάριστη παρέα τόσα χρόνια, τη βοήθεια και τη στήριξή τους.
Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους γονείς μου και στον αδερφό μου για την αμέριστη συμπαράσταση, αγάπη και υποστήριξή τους σε όποια απόφαση κι αν πάρω.

Μπαλάση Παναγιώτα
Βόλος, 2012

Περιεχόμενα

Πίνακας Εικόνων

1.Περίληψη

Τα τελευταία χρόνια οι υπολογιστές έχουν κατακτήσει σημαντική θέση σε κάθε τομέα της ζωής μας, αλλά πολύ ενδιαφέρουσα και προκλητική σε αρκετούς τομείς διαφόρων επιστημών.
Η Βιοπληροφορική αποτελεί ένα σύγχρονο τομέα έρευνας και ανάπτυξης για μοριακούς βιολόγους αλλά και επιστήμονες της πληροφορικής. Η συνεργασία αυτή έρχεται να ρίξει φως στην ερμηνεία και το ρόλο της γονιδιακής πληροφορίας και κατ`επέκταση σε αρκετές διαδικασίες της ζωής που ζητούν ερμηνεία.
Η πρόοδος της τεχνολογίας των υπολογιστών επιτρέπει την προσπάθεια ανάλυσης προβλημάτων που προκύπτουν στον τομέα της μοριακής βιολογίας. Λόγω της αναπτυγμένης τεχνολογίας των γραφικών είναι δυνατή η απεικόνιση των διαμορφώσεων της δομής των βιολογικών μορίων στη οθόνη του υπολογιστή. Ο μεγάλος αριθμός δεδομένων που μεταφράζονται στην επιστήμη της μοριακής βιολογίας και ειδικότερα στον τομέα της αλληλούχισης του γονιδιώματος (δηλαδή της αλληλουχίας του DNA), αποτελεί μεγάλη πρόκληση για τους επιστήμονες του σχεδιασμού της ανάλυσης αλγορίθμων.
Συγκεκριμένα η ερμηνεία αυτών των δεδομένων μπορεί να διευκολύνει την αναζήτηση λύσεων σε προβλήματα όπως είναι η αναγνώριση γονιδίων, ο καθορισμός της δομής των κωδικοποιημένων πρωτεϊνών, η ανακάλυψη των μηχανισμών με τους οποίους οι πρωτεΐνες εκτελούν τη βιολογική λειτουργία τους, η απόκτηση γνώσης για το ρόλο των μη κωδικοποιημένων περιοχών του DNA στη μορφολογία και έκφραση των γονιδίων.
Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η επισκόπηση διαφόρων εξελικτικών-γενετικών αλγορίθμων με βάση τους οποίους έχουν αναπτυχθεί τεχνικές για επίλυση προβλημάτων βιοπληροφορικής τα τελευταία 25 χρόνια.
Συγκεκριμένα, τα προβλήματα στα οποία θα αναφερθούμε στα επόμενα κεφάλαια είναι : η χαρτογράφηση περιορισμού, η εύρεση μοτίβων, οι αναδιατάξεις γονιδιώματος, η γονιδιακή πρόγνωση, η σύγκριση ακολουθιών και τέλος η μοριακή εξέλιξη

2.Βασικές Έννοιες

2.1 Τι είναι Αλγόριθμος;

Η λέξη *αλγόριθμος* (algorithm) προέρχεται από μια μελέτη του Πέρση μαθηματικού Abu Ja'far Mohammed ibn Musa al Khowarizmi, που έζησε περί το 825 μ. Πέντε αιώνες αργότερα η μελέτη αυτή μεταφράστηκε στα λατινικά και άρχιζε με τη φράση "Algoritmi dixit ..." (ο αλγόριθμος λέει ....). Η μελέτη του al Khowarizmi υπήρξε η πρώτη πλήρης πραγματεία άλγεβρας (όρος που και αυτός προέρχεται από το αραβικό aljabr=αποκατάσταση), γιατί ένας από τους σκοπούς της άλγεβρας είναι και η αποκατάσταση της ισότητας μέσα σε μια εξίσωση. Ο όρος αλγόριθμος επέζησε επί χίλια χρόνια ως σπάνιος όρος, που σήμαινε κάτι σαν "συστηματική διαδικασία αριθμητικών χειρισμών". Τη σημερινή του αξία απόκτησε από την αρχή του 20ού αιώνα με την ανάπτυξη της ομώνυμης θεωρίας και φυσικά με την επικαιρότητα των ηλεκτρονικών υπολογιστών.
Ο όρος αλγόριθμος, λοιπόν, χρησιμοποιείται για να δηλώσει μεθόδους που εφαρμόζονται για την επίλυση προβλημάτων. Ωστόσο, ένας πιο αυστηρός ορισμός της έννοιας αυτής είναι ο εξής.

Ορισμός: *Αλγόριθμος είναι μια πεπερασμένη σειρά ενεργειών, αυστηρά καθορισμένων και εκτελέσιμων σε πεπερασμένο χρόνο, που στοχεύουν στην επίλυση ενός προβλήματος. Κάθε αλγόριθμος απαραίτητα ικανοποιεί τα επόμενα κριτήρια.*

- ***Είσοδος*** (input). Καμία, μία ή περισσότερες τιμές δεδομένων πρέπει να δίνονται ως είσοδοι στον αλγόριθμο. Η περίπτωση που δεν δίνονται τιμές δεδομένων εμφανίζεται, όταν ο αλγόριθμος δημιουργεί και επεξεργάζεται κάποιες πρωτογενείς τιμές με τη βοήθεια συναρτήσεων παραγωγής τυχαίων αριθμών ή με τη βοήθεια άλλων απλών εντολών.
- ***Έξοδος*** (output). Ο αλγόριθμος πρέπει να δημιουργεί τουλάχιστον μία τιμή δεδομένων ως αποτέλεσμα προς το χρήστη ή προς έναν άλλο αλγόριθμο.
- ***Καθοριστικότητα*** (definiteness). Κάθε εντολή πρέπει να καθορίζεται χωρίς καμία αμφιβολία για τον τρόπο εκτέλεσής της. Λόγου χάριν, μία εντολή διαίρεσης πρέπει να θεωρεί και την περίπτωση, όπου ο διαιρέτης λαμβάνει μηδενική τιμή.
- ***Περατότητα*** (finiteness). Ο αλγόριθμος να τελειώνει μετά από πεπερασμένα βήματα εκτέλεσης των εντολών του. Μία διαδικασία που δεν τελειώνει μετά από ένα συγκεκριμένο αριθμό βημάτων δεν αποτελεί αλγόριθμο, αλλά λέγεται απλά *υπολογιστική διαδικασία* (computational procedure).
- ***Αποτελεσματικότητα*** (effectiveness). Κάθε μεμονωμένη εντολή του αλγορίθμου να είναι απλή. Αυτό σημαίνει ότι μία εντολή δεν αρκεί να έχει ορισθεί, αλλά πρέπει να είναι και εκτελέσιμη.

Η έννοια του αλγόριθμου είναι θεμελιώδης για την επιστήμη της Πληροφορικής. Η μελέτη των αλγορίθμων είναι πολύ ενδιαφέρουσα, γιατί είναι η πρώτη ύλη για τη μελέτη και εμβάθυνση, αν όχι σε όλες, τουλάχιστον σε πάρα πολλές γνωστικές περιοχές της επιστήμης αυτής.

Η Πληροφορική, λοιπόν, μπορεί να ορισθεί ως η επιστήμη που μελετά τους αλγορίθμους από τις ακόλουθες σκοπιές:

- ***Υλικού*** (hardware). Η ταχύτητα εκτέλεσης ενός αλγορίθμου επηρεάζεται από τις διάφορες τεχνολογίες υλικού, δηλαδή από τον τρόπο που είναι δομημένα σε μία ενιαία αρχιτεκτονική τα διάφορα συστατικά του υπολογιστή (δηλαδή ανάλογα με το αν ο υπολογιστής έχει κρυφή μνήμη και πόση, ανάλογα με την ταχύτητα της κύριας και δευτερεύουσας μνήμης κοκ.).
- ***Γλωσσών Προγραμματισμού*** (programming languages). Το είδος της γλώσσας προγραμματισμού που χρησιμοποιείται (δηλαδή, χαμηλότερου ή υψηλότερου επιπέδου) αλλάζει τη δομή και τον αριθμό των εντολών ενός αλγορίθμου. Γενικά μία γλώσσα που είναι χαμηλότερου επιπέδου (όπως η assembly ή η γλώσσα C) είναι ταχύτερη από μία άλλη γλώσσα που είναι υψηλοτέρου επιπέδου (όπως η Basic ή Pascal). Ακόμη, σημειώνεται ότι διαφορές συναντώνται μεταξύ των γλωσσών σε σχέση με το πότε εμφανίσθηκαν. Για παράδειγμα, παλαιότερα μερικές γλώσσες προγραμματισμού δεν υποστήριζαν την αναδρομή
- ***Θεωρητική*** (theoretical). Το ερώτημα που συχνά τίθεται είναι, αν πράγματι υπάρχει ή όχι κάποιος αποδοτικός αλγόριθμος για την επίλυση ενός προβλήματος. Η προσέγγιση αυτή είναι ιδιαίτερα σημαντική, γιατί προσδιορίζει τα όρια της λύσης που θα βρεθεί σε σχέση με ένα συγκεκριμένο πρόβλημα.
- **Αναλυτική** (analytical). Μελετώνται οι υπολογιστικοί πόροι (computer resources) που απαιτούνται από έναν αλγόριθμο, όπως για παράδειγμα το μέγεθος της κύριας και της δευτερεύουσας μνήμης, ο χρόνος για λειτουργίες CPU και για λειτουργίες εισόδου/εξόδου κ.λ.π

2.2 Τι είναι Βιοπληροφορική;

Στα μέσα του 19ου αιώνα, σε μία πόλη της Γαλλίας ο Louis Pasteur (1822-1895) μελετούσε το πως η ζύμωση της αλκοόλης συνδέεται με την ύπαρξη ενός μικροοργανισμού. Την ίδια εποχή στην Αγγλία ο Charles Babbage (1791-1871) κατασκεύαζε την Αναλυτική Μηχανή στην οποία η Ada Lovelace (1815-1852) - μία μαθηματικός που είχε κατανοήσει το όραμα του Babbage - προσπαθούσε να υπολογίσει τους αριθμούς Bernoulli . Είναι σχεδόν βέβαιο πως αν ποτέ συναντιόταν ο Babbage , που σήμερα θεωρείται ο πατέρας της επιστήμης των υπολογιστών, με τον Pasteur που θεωρείται ο πατέρας της Βιοτεχνολογίας, δεν θα φαντάζονταν πως οι δύο επιστήμες θα είχαν τόσα κοινά, ώστε η εξέλιξη της μίας να επιδρά ζωτικά στην εξέλιξη της άλλης.

Το 1978 οι Paulien Hogeweg και Ben Hesper μελετώντας τις διαδικασίες πληροφορικής σε βιοτικά συστήματα εισήγαγαν τους όρους Βιοπληροφορική και Υπολογιστική Βιολογία.

Βιοπληροφορική *είναι ο επιστημονικός χώρος όπου η σύμπραξη της Βιολογίας με την Πληροφορική, την Στατιστική και τα Μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της Βιολογίας. Πρόκειται για γνωστικό χώρο με συγκεκριμένο όσο και ευρύ πεδίο εφαρμογών και αλληλεπίδρασης με τη σύγχρονη δομική, μοριακή, πληθυσμιακή και περιβαλλοντική βιολογία.*

Ο κλάδος της Βιοπληροφορικής σήμερα θεωρείται, παγκόσμια, ένας από τους πλέον αναπτυσσόμενους, ενώ έχει ήδη επιδείξει σημαντικά επιτεύγματα και έχει συγκεντρώσει ιδιαίτερα σημαντικές επενδύσεις. Ουσιαστικά, κατέχει κεντρική θέση στις σύγχρονες εξελίξεις των Επιστημών της Ζωής, με πιο χαρακτηριστικό παράδειγμα τα προγράμματα "Αποκωδικοποίησης" των γονιδιωμάτων, περιλαμβανομένου και αυτού του Ανθρώπου.
Στόχος της έρευνας στη βιοπληροφορική είναι η κατανόηση της λειτουργίας των ζωντανών όντων , ο σχεδιασμός φαρμάκων, η αναγνώριση γενετικών παραγόντων κινδύνου, η γονιδιακή θεραπεία, η γενετική τροποποίηση φυτών και ζώων και η βελτίωση μέσων βιολογικού πολέμου
Οι σημαντικές αλλαγές που συντελέστηκαν τις τελευταίες δεκαετίες στο πεδίο της Μοριακής Βιολογίας (κλασσικής και δομικής), σε συνδυασμό με την εξέλιξη της τεχνολογίας της γονιδιωματικής, αλλά και στη μελέτη της βιοποικιλότητας και της διατήρησης της φύσης, οδήγησαν σε εκθετική αύξηση των πληροφοριών που παράγονται από τη βιολογική κοινότητα. Το γεγονός αυτό, κατέστησε απόλυτα αναγκαία τη διαχείριση, τον έλεγχο και την ανάλυση όλων αυτών των δεδομένων με τελικό σκοπό την αξιοποίησή τους για την εξαγωγή σημαντικών Βιολογικών συμπερασμάτων. Άμεση συνέπεια των ανωτέρω, ήταν η ανάγκη για την ανάπτυξη εξειδικευμένων υπολογιστικών εργαλείων (λογισμικού), αλλά και την προσαρμογή ήδη υπαρχόντων δοκιμασμένων συστημάτων, για την αποθήκευση, οπτικοποίηση και ανάλυση των δεδομένων, δίνοντας το έναυσμα για τη μεγάλη ανάπτυξη, που παρατηρείται στις μέρες μας, στο πεδίο της Βιοπληροφορικής.
Το ερευνητικό πεδίο της Βιοπληροφορικής προϋπήρχε της τεράστιας έκρηξης στη συλλογή των γονιδιωματικών, κυρίως, αλλά και των πληθυσμιακών και οικολογικών πληροφοριών και είχε αρχίσει να αναπτύσσεται από τις αρχές της δεκαετίας του 1970. Αρκετές από τις σημερινές κατευθύνσεις στον τομέα αυτό βασίζονται σε θεμέλια που είχαν ήδη τεθεί από την περίοδο εκείνη. Ο πλούτος και η ποικιλομορφία των πληροφοριών που διατίθενται στις μέρες μας και χρήζουν ανάλυσης και επεξεργασίας έδωσαν νέα ώθηση και προεκτάσεις στο πεδίο αυτό, το οποίο αποτελεί, σε αρκετές περιπτώσεις, την αιχμή του δόρατος στην βασική αλλά και εφαρμοσμένη έρευνα των Βιολογικών-Βιοϊατρικών επιστημών.
Χαρακτηριστικό φαινόμενο καθ' όλη την ιστορία του κλάδου της Βιοπληροφορικής έχει αποτελέσει η "στρατολόγηση" ειδικών από διάφορα γνωστικά αντικείμενα (Βιολογία, Πληροφορική, Μαθηματικά, Φυσική, Χημεία κλπ.) με κοινό παρονομαστή τη χρήση μαθηματικών μεθόδων και υπολογιστικών τεχνικών για την περιγραφή και ανάλυση Βιολογικών Συστημάτων. Η ολοένα αυξανόμενη πολυπλοκότητα των προς ανάλυση δεδομένων και η ποικιλία τους καθιστά επιτακτική τη συνεισφορά και συνεργασία ειδικών από όλα τα παραπάνω πεδία, εκπαιδευμένων κατάλληλα με βάση τις γενικότερες αρχές και τη μεθοδολογία της σύγχρονης Βιοπληροφορικής, ώστε να είναι δυνατόν να ανταπεξέλθουν στις αυξημένες απαιτήσεις του πεδίου στην έρευνα και την παραγωγή.

Υπάρχουν τρεις σημαντικές παιδικότητες στο πλαίσιο της βιοπληροφορικής.

1) Ανάπτυξη νέων αλγορίθμων και μοντέλων για την αξιολόγηση διαφορετικών σχέσεων μεταξύ των μελών ενός μεγάλου βιολογικού συνόλου δεδομένων με τρόπο που να επιτρέπει στους ερευνητές να έχουν πρόσβαση στις υπάρχουσες πληροφορίες, και να παρέχουν νέες όταν αυτές παράγονται.
2) Ανάλυση και ερμηνεία των διαφόρων τύπων δεδομένων, συμπεριλαμβανομένων νουκλεοτιδίων και ακολουθίες αμινοξέων, τομείς πρωτεΐνης και τις δομές των πρωτεϊνών.

3) Ανάπτυξη και εφαρμογή εργαλείων που επιτρέπουν την αποτελεσματική πρόσβαση και διαχείριση διαφορετικών τύπων πληροφοριών.

2.3 Εξελικτικοί-Γενετικοί αλγόριθμοι

2.3.1 Εξελικτικοί αλγόριθμοι

Πρόσφατα, εξελικτικοί αλγόριθμοι (evolutionary algorithms (EAs),), μια κατηγορία τυχαίων αναζητήσεων και βελτιστοποιημένων τεχνικών, καθοδηγούμενη από τις αρχές της εξέλιξης και της φυσικής γενετικής, έχει κερδίσει την προσοχή των ερευνητών για την επίλυση προβλημάτων της βιοπληροφορικής. Γενετικοί αλγόριθμοι (Genetic algorithms (GAs)), στρατηγικές εξέλιξης (evolutionary strategies (ES)), και γενετικός προγραμματισμός (and genetic programming (GP) είναι τα βασικά συστατικά των εξελικτικών αλγορίθμων Από αυτά οι γενετικοί αλγόριθμοι χρησιμοποιούνται περισσότερο κι αυτό γιατί είναι αποτελεσματικοί καθώς παράγουν κοντινές βέλτιστες λύσεις. Εργαλεία ανάλυσης δεδομένων που χρησιμοποιούνταν νωρίτερα στη βιοπληροφορική βασίζονταν κυρίως σε στατιστικές τεχνικές, όπως παλινδρόμηση και εκτίμηση. Ο ρόλος των γενετικών αλγορίθμων στην βιοπληροφορική απέκτησε σημασία με την ανάγκη της διαχείρισης μεγάλων συνόλων δεδομένων στον τομέα της βιολογίας με ένα ισχυρό και υπολογιστικά αποδοτικό τρόπο.

2.3.2 Γενετικοί αλγόριθμοι

Η πρώτη εμφάνιση των Γενετικών Αλγορίθμων χρονολογείται στις αρχές του 1950. Η συστηματική τους ανάπτυξη όμως που οδήγησε στην μορφή με την οποία τους γνωρίζουμε σήμερα πραγματοποιήθηκε στις αρχές του 1970 από τον Ben Lovelace και τους συνεργάτες του στο Πανεπιστήμιο του Michigan, ενώ αργότερα ο Goldberg παρουσίασε τον πρώτο Γενετικό Αλγόριθμο. Στη συνέχεια ο Fogel παρουσίασε τον Εξελικτικό Προγραμματισμό και ο Ada το Γενετικό Προγραμματισμό.
Η βασική ιδέα που κρύβεται πίσω από τους Γ.Α είναι η μίμηση των μηχανισμών της βιολογικής εξέλιξης που απαντώνται στη φύση και χρησιμοποιούν ορολογία δανεισμένη από τη Γενετική. Αναφέρονται σε άτομα (individuals) ή γενότυπους (genotypes) μέσα σε ένα πληθυσμό. Κάθε άτομο ή γενότυπος αποτελείται μόνο από ένα χρωμόσωμα και αναπαριστά μια πιθανή λύση σε ένα πρόβλημα. Μια διαδικασία εξέλιξης που εφαρμόζεται πάνω σε ένα πληθυσμό αντιστοιχεί σε ένα εκτενές ψάξιμο στο χώρο των πιθανών λύσεων.
Οι Γ.Α διατηρούν έναν πληθυσμό πιθανών λύσεων, του προβλήματος που μας ενδιαφέρει, πάνω στο οποίο δουλεύουν, σε αντίθεση με άλλες μεθόδους αναζήτησης που επεξεργάζονται ένα μόνο σημείο διαστήματος αναζήτησης. Έτσι ένας Γ.Α πραγματοποιεί αναζήτηση σε πολλές κατευθύνσεις και υποστηρίζει καταγραφή και ανταλλαγή πληροφοριών μεταξύ αυτών των κατευθύνσεων. Ο πληθυσμός υφίσταται μια προσομοιωμένη γενετική εξέλιξη. Σε κάθε γενιά, οι σχετικά «καλές» λύσεις αναπαράγονται, ενώ οι σχετικά «κακές» απομακρύνονται. Ο διαχωρισμός και η αποτίμηση των διαφόρων

λύσεων γίνεται με τη βοήθεια μιας αντικειμενικής συνάρτησης (objective or fitness function) η οποία παίζει το ρόλο του περιβάλλοντος μέσα στο οποίο εξελίσσεται ο πληθυσμός. Η αντικειμενική συνάρτηση υπολογίζει για κάθε άτομο (δηλαδή για κάθε πιθανή λύση) την ικανότητά του για επιβίωση.

Η δομή ενός απλού Γ.Α έχει σε γενικές γραμμές ως εξής:

- Κατά τη διάρκεια της γενιάς t, ο Γ.Α διατηρεί έναν πληθυσμό P(t) από n πιθανές λύσεις (individuals): $P(t)=x_1, x_2, \ldots, x_n$
- Κάθε λύση (individual) x_i αποτιμάται και δίνει ένα μέτρο της καταλληλότητας και ορθότητάς της βάσει της αντικειμενικής συνάρτησης.
- Αφού ολοκληρωθεί η αποτίμηση όλων των μελών του πληθυσμού, δημιουργείται ένας νέος πληθυσμός (γενιά t + 1) που προκύπτει από την επιλογή των πιο κατάλληλων στοιχείων του πληθυσμού της προηγούμενης γενιάς.
- Μερικά μέλη από τον καινούριο πληθυσμό υφίστανται αλλαγές με τη βοήθεια των γενετικών τελεστών (της διασταύρωσης και της μετάλλαξης) σχηματίζοντας νέες πιθανές λύσεις.

1: $t \leftarrow 0$.
2: Αρχικοποίησε το $P(t)$.
3: Αξιολόγησε το $P(t)$.
4: **while** not συνθήκη τερματισμού **do**
5: $t \leftarrow t + 1$
6: Επιλογή του $P(t)$ από το $P(t-1)$
7: Τροποποίηση του $P(t)$
8: Αξιολόγηση του $P(t)$
9: **end while**

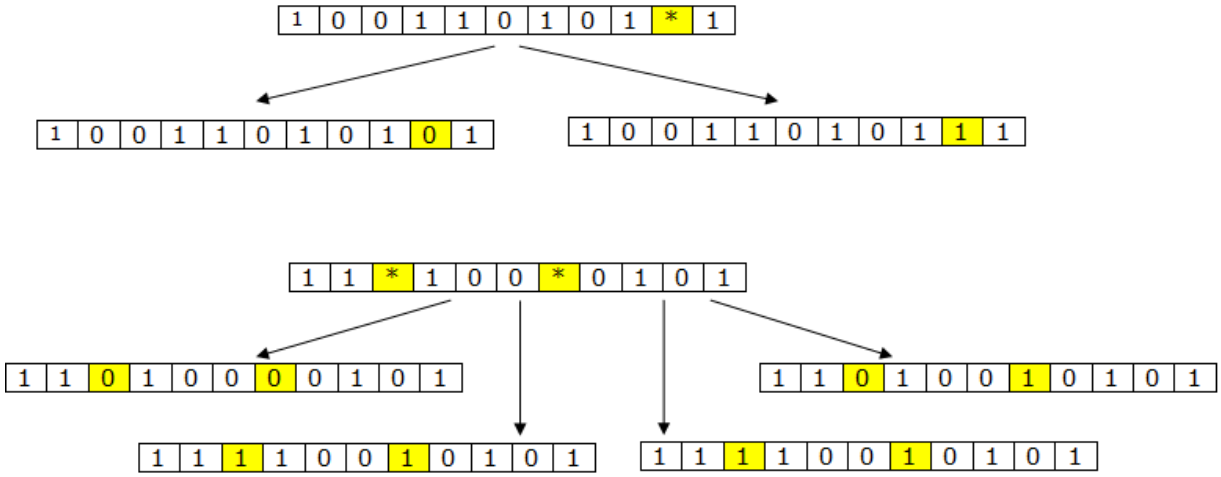
***Εικόνα 1** : Δομή Γενετικού Αλγορίθμου*

Η προσομοιωμένη γενετική εξέλιξη κάθε πληθυσμού πραγματοποιείται με τη βοήθεια τριών τελεστών

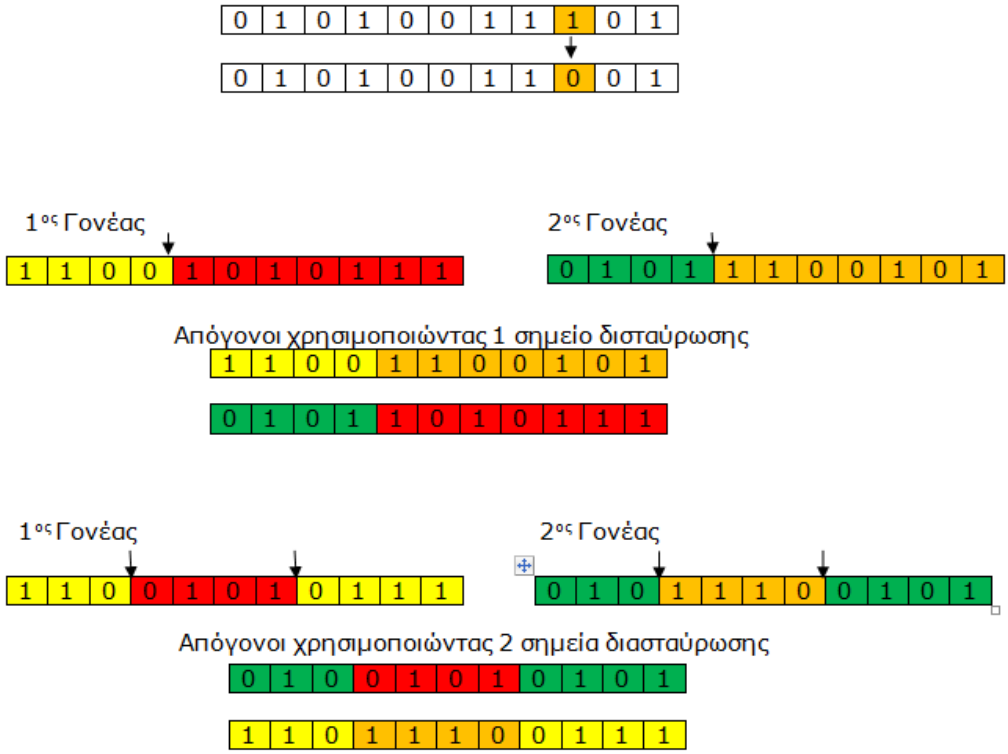
- ***Επιλογή:*** Καθορίζει ποια άτομα από τον υπάρχοντα πληθυσμό θα λάβουν μέρος στην αναπαραγωγή. Η τεχνική της επιλογής οφείλει να δίνει μεγαλύτερες πιθανότητες αναπαραγωγής σε άτομα που αξιολογούνται ως πιο ικανά. Γενετικός αλγόριθμος χωρίς επιλογή ισοδυναμεί σε τυχαίο ψάξιμο

- ***Διασταύρωση:*** κατά τη διασταύρωση, δυο άτομα του πληθυσμού παίρνουν μέρος σε μια απλή λειτουργία ανταλλαγής γενετικού υλικού. Στόχος είναι η νέα γενιά που θα προκύψει να περιλαμβάνει άτομα που θα διαφέρουν από τους γονείς τους και θα φέρουν συνδυασμό των καλύτερων χαρακτηριστικών τους. Η διασταύρωση λαμβάνει χώρα με πιθανότητα διασταύρωσης (crossover probability) p_c που καθορίζεται από το σχεδιαστή του Γενετικού Αλγορίθμου.
- ***Μετάλλαξη:*** μεταλλάσσει τη γενετική πληροφορία που μεταφέρει κάθε άτομο. Η επιλογή της πληροφορίας που αλλάζει από το γονέα στον απόγονο γίνεται τυχαία με μικρή πιθανότητα, την πιθανότητα μετάλλαξης (mutation probability) p_m.

Η διασταύρωση συνδυάζει τα στοιχεία των χρωμοσωμάτων δύο γονέων για να δημιουργήσει δύο νέους απογόνους απαλλάσσοντας κομμάτια από τους γονείς. Για παράδειγμα έστω ότι οι δύο γονείς αναπαριστώνται με χρωμοσώματα πέντε γονιδίων (a_1, b_1, c_1, d_1, e_1) και (a_2, b_2, c_2, d_2,e_2) αντίστοιχα τότε οι απόγονοι που θα προκύψουν από διασταύρωση με σημείο διασταύρωσης (crossover point) το σημείο 2 είναι οι (a_1, b_1, c_2, d_2, e_2) και (a_2, b_2, c_1, d_1, e_1). Διαισθητικά μπορούμε να πούμε ότι η διασταύρωση εξυπηρετεί την ανταλλαγή πληροφοριών μεταξύ διαφορετικών πιθανών λύσεων.
Αντίστοιχα μπορούμε να πούμε ότι η μετάλλαξη εξυπηρετεί την εισαγωγή νέων πιθανών λύσεων, διαφορετικών από τις υπάρχουσες. Η διαδικασία της μετάλλαξης αλλάζει αυθαίρετα ένα ή περισσότερα γονίδια ενός συγκεκριμένου χρωμοσώματος. Πραγματοποιείται με τυχαία αλλαγή γονιδίων με πιθανότητα ίση με το ρυθμό μετάλλαξης (mutation rate). Η μετάλλαξη λειτουργεί ως ασφαλιστική δικλείδα για τις περιπτώσεις που η επιλογή και η διασταύρωση χάσουν κάποιες πολύτιμες γενετικές πληροφορίες.
Η θεωρητική θεμελίωση των Γ.Α βασίζεται στην αναπαράσταση των δυνατών λύσεων σε δυαδικές συμβολοσειρές καθώς και στην έννοια του σχήματος (schema), μιας φόρμας (template) που επιτρέπει τον προσδιορισμό της ομοιότητας μεταξύ των χρωμοσωμάτων. Ένα σχήμα κατασκευάζεται εισάγοντας το λεγόμενο αδιάφορο σύμβολο (don`t care symbol) στο αλφάβητο των γονιδίων Σ=0,1. Ένα σχήμα αναπαριστά όλες τις συμβολοσειρές οι οποίες ταιριάζουν σε όλες τις θέσεις εκτός από αυτές με το αδιάφορο σύμβολο
Ας θεωρήσουμε για παράδειγμα τις συμβολοσειρές και τα σχήματα μήκους 10. Στο σχήμα (*111100100) ταιριάζουν οι δυο συμβολοσειρές: (0111100100, 1111100100) ενώ στο σχήμα (*1*1100100) ταιριάζουν τέσσερις συμβολοσειρές (0101100100, 0111100100, 1101100100, 1111100100) (Εικόνα 2).



Παράδειγμα εφαρμογής του τελεστή της μετάλλαξης ως προς 1



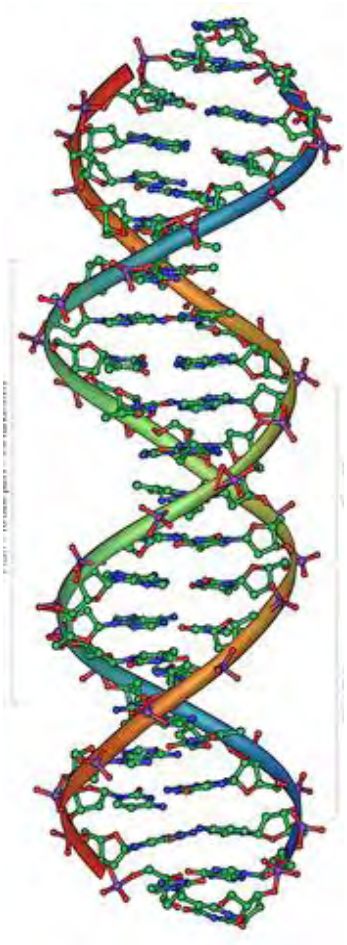
***Εικόνα 2** : Παραδείγματα χρήσης των τελεστών της μετάλλαξης και της διασταύρωσης*

2.4 Βασικές έννοιες της Βιοπληροφορικής

2.4.1 DNA

Το **δε(σ)οξυριβο(ζο)νουκλεϊ(νι)κό οξύ (Deoxyribonucleic acid -** DNA**)** είναι ένα νουκλεϊκό οξύ που περιέχει τις γενετικές πληροφορίες που καθορίζουν τη βιολογική ανάπτυξη όλων των κυτταρικών μορφών ζωής και των περισσοτέρων ιών και συνήθως έχει τη μορφή διπλής έλικας.
Η αποκωδικοποίηση του DNA, η αποσαφήνιση δηλαδή του τρόπου με τον οποίο η δομή του DNA καθορίζει συγκεκριμένες γενετικές επιλογές, επέτρεψε στους επιστήμονες να κατανοήσουν καλύτερα την γενετική της ζωής και την κληρονόμηση ορισμένων χαρακτηριστικών και νόσων. Η ανακάλυψη της δομής του DNA πραγματοποιήθηκε το 1953 από τους Τζέιμς Γουάτσον (James D. Watson) και Φράνσις Κρικ (Francis Crick). Από πολλούς η ανακάλυψη της διπλής έλικας του DNA θεωρείται ως η μεγαλύτερη βιολογική ανακάλυψη του 20ου αιώνα.
Το DNA Πρόκειται για μια μεγαλομοριακή ένωση που συγκροτείται από αζωτούχες-πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα (πεντόζη), την δε(σ)οξυριβόζη. Στα ευκαρυωτικά κύτταρα ανιχνεύεται κυρίως μέσα στον πυρήνα του κυττάρου αλλά και σε μερικά άλλα οργανίδια, όπως τα μιτοχόνδρια και τα πλαστίδια, επιτρέποντάς τους να αναπαράγονται αυτόνομα (ημιαυτόνομα οργανίδια).
Το σύνολο των μορίων DNA που υπάρχουν σε ένα κύτταρο αποτελούν το γενετικό υλικό του. Το DNA είναι ο φορέας των γενετικών πληροφοριών του κυττάρου, όχι μόνον με την έννοια της μεταβίβασης χαρακτηριστικών, αναλλοίωτων από γενεά σε γενεά, αλλά και της ρύθμισης της φυσιογνωμίας εξειδίκευσης κάθε κυττάρου για την επιτέλεση των ιδιαίτερων λειτουργιών του. Τέλος, το DNA επιτρέπει τη δημιουργία γενετικής ποικιλότητας, υφιστάμενο μεταλλάξεις.
Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμηκών αλυσίδων οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι αζωτούχες βάσεις στο DNA είναι τέσσερις: κυτοσίνη (C), γουανίνη (G), θυμίνη (T), αδενίνη (A).
Σύμφωνα με το μοντέλο Watson- Crick το μόριο του DNA παρουσιάζεται με τα ακόλουθα βασικά χαρακτηριστικά:

1. Αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες σε μορφή δύο αντιτακτών κλώνων που σχηματίζουν δεξιόστροφη διπλή έλικα.
2. Οι αζωτούχες βάσεις (ή πρωτεϊνικές) κάθε κλώνου είναι κάθετες ως προς τον άξονα του μορίου και προεξέχουν προς το εσωτερικό της συστροφής.
3. Οι δύο δημιουργούμενοι κλώνοι συγκρατούνται μεταξύ τους με δεσμούς υδρογόνου. Τα δε ζευγάρια των αζωτούχων βάσεων όπου αναπτύσσονται μεταξύ τους δεσμοί υδρογόνου είναι καθορισμένα: η αδενίνη με τη θυμίνη και η γουανίνη με την κυτοσίνη.
4. Μεταξύ της αδενίνης και της θυμίνης σχηματίζονται δύο δεσμοί υδρογόνου , ενώ μεταξύ της γουανίνης και της κυτοσίνης τρεις δεσμοί υδρογόνου



Εικόνα 3 *: Τρισδιάστατη απεικόνιση του μοντέλου ελικοειδούς δομής ενός τμήματος DNA*

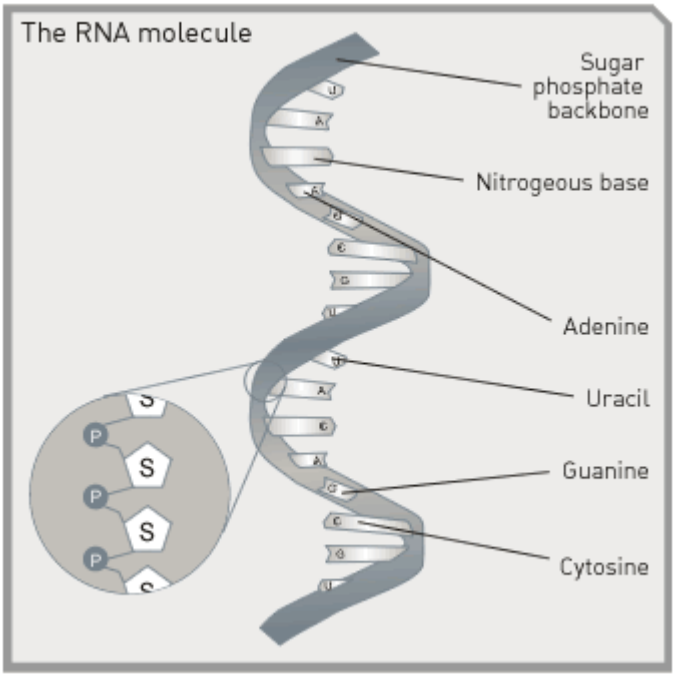
Πηγή: Google-εικόνες για DNA

2.4.2 RNA

Το **ριβο(ζο)νουκλεϊκό** οξύ **(RNA)** , είναι μία από τις δύο κατηγορίες των πολυμερών νουκλεϊκών οξέων στο κύτταρο. Αποτελείται από μονομερή νουκλεοτίδια που παίζουν σημαντικό ρόλο στη διαδικασία της μετάφρασης του γενετικού κώδικα από την έτερη κατηγορία νουκλεϊκού οξέος, το δεοξυριβονουκλεϊκό οξύ (συντομογραφικά DNA), σε πρωτεϊνικά προϊόντα. Το RNA χαρακτηρίζεται ως ο «αγγελιοφόρος» μεταξύ του DNA και των πρωτεϊνικών συμπλεγμάτων που είναι γνωστά σαν ριβοσώματα στο κυτταρόπλασμα του κυττάρου (αγγελιαφόρο RNA, mRNA). Έτσι το RNA μαζί με το DNA αποτελούν το γενετικό υλικό των οργανισμών.

Στα βακτηριακά κύτταρα το μεγαλύτερο μέρος του απαντώμενου RNA εντοπίζεται στο κυτταρόπλασμα, ενώ μια ακόμη ποσότητα (κατά το στάδιο της βιοσύνθεσης) εντοπίζεται να συνδέεται με μη ομοιοπολικούς χημικούς δεσμούς με το DNA. Επίσης το RNA εντοπίζεται σε όλα τα είδη των ευκαρυωτικών κυττάρων. Για παράδειγμα στα ηπατικά κύτταρα περίπου το 11% της συνολικής ποσότητας RNA απαντάται στον πυρήνα, το 15% στα μιτοχόνδρια, ένα 24% στο κυτοσόλυο και το υπόλοιπο 50% στα ριβοσώματα.

Από χημικής άποψης το RNA είναι όμοιο με το DNA. Και οι δύο αυτές κατηγορίες νουκλεϊκών οξέων είναι μακρομοριακές ενώσεις μεγάλου μοριακού βάρους. Το μακρομόριο του RNA αποτελείται από επαναλαμβανόμενες δομικές μονάδες τα νουκλεοτίδια. Το μόριο του RNA περιλαμβάνει (όπως και του DNA), τέσσερις τύπους νουκλεοτιδίων που συνδέονται μεταξύ τους με 3'-5' φωσφοδιεστερικούς δεσμούς. Ωστόσο κύρια διαφορά του RNA από το DNA είναι ότι το μόριό του είναι μονόκλωνο έναντι του δίκλωνου του DNA, αποτελείται δηλαδή από μια μόνο αλυσίδα, ανάλογη της μιας εκ των δύο εκείνων της διπλής έλικας του DNA. Βασική επίσης διαφορά είναι ότι το σάκχαρο στα νουκλεοτίδια του είναι η ριβόζη, εξ ου και η ονομασία τους ριβονουκλεοτίδια, αντί της δεοξυριβόζης στο DNA, και ότι περιέχει την πυριμιδίνη ουρακίλη αντί της θυμίνης (που υπάρχει στο μόριο του DNA), χωρίς να είναι γνωστός ο λόγος της τελευταίας αυτής διαφοράς. Η μακρομοριακή πολυνουκλεοτιδική αλυσίδα του RNA εμφανίζεται από ελικοειδής μέχρι ευθύγραμμη.

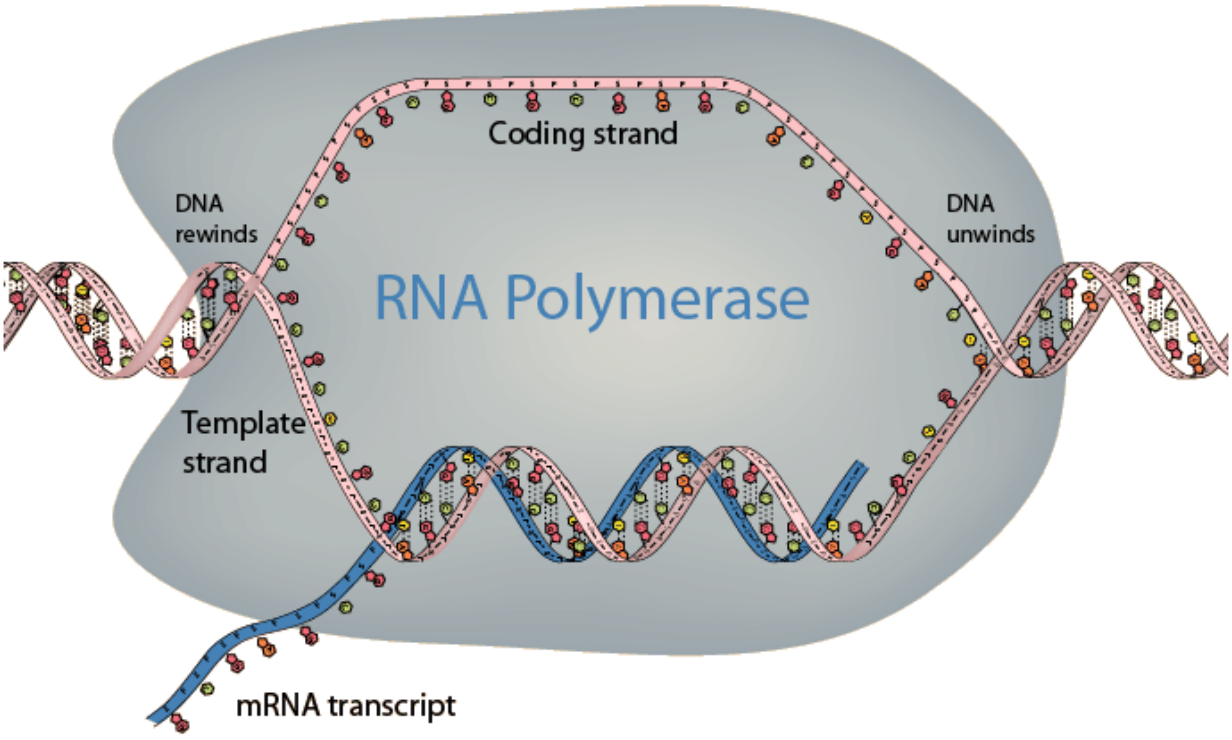


***Εικόνα 4**: Απεικόνιση του μορίου RNA*

Πηγή: Google-εικόνες για RNA

Έχουμε 5 είδη RNA:

- **Αγγελιοφόρο RNA mRNA**(Messenger RNA)
είναι το RNA που μεταφέρει τη γενετική πληροφορία από το DNA στα ριβοσώματα για την πρωτεϊνοσύνθεση των κυττάρων. Στα ευκαρυωτικά κύτταρα όταν το mRNA μεταγράφεται από το DNA υφίσταται επεξεργασία πριν εξαχθεί από τον πυρήνα στο κυτταρόπλασμα. Εκεί συνδέεται με τα ριβοσώματα και μεταφράζεται στην αντίστοιχη πρωτεΐνη με τη βοήθεια του μεταφορικού RNA tRNA. Στα προκαρυωτικά κύτταρα, όπου δεν υπάρχει σαφής διάκριση μεταξύ πυρήνα και κυτταροπλάσματος, το mRNA μπορεί να συνδεθεί με τα ριβοσώματα ενώ μεταγράφεται από το DNA. Μετά από κάποιο χρονικό διάστημα το γενετικό μήνυμα υποβαθμίζεται στα συστατικά του νουκλεοτίδια συνήθως με τη βοήθεια ειδικών ενζύμων.

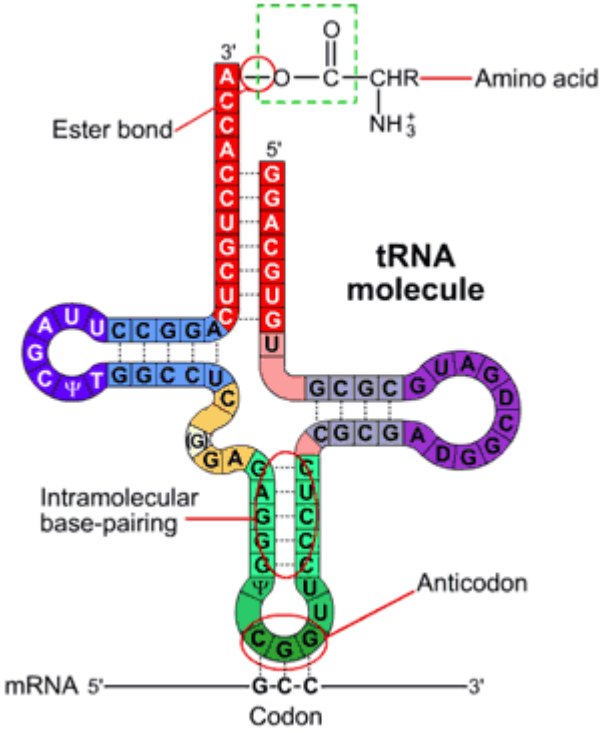


***Εικόνα 5**:Απεικόνιση του mRNA*

Πηγή: Google-εικόνες mRNA

- **Μεταφορικό RNA tRNA** (Transfer RNA)
είναι μικρή αλυσίδα RNA, με μήκος 74-95 νουκλεοτιδίων, που μεταφέρει ειδικά αμινοξέα σε μια επεκτεινόμενη πολυπεπτιδική αλυσίδα στα ριβοσώματα του κυττάρου, με βάση τις οδηγίες του αγγελιαφόρου RNA. Έτσι γίνεται η πρωτεϊνοσύνθεση κατά τη διάρκεια της μετάφρασης στο κύτταρο. Είναι ένας τύπος μη κωδικοποιητικού RNA. Το tRNA διαθέτει ειδικούς υποδοχείς για την

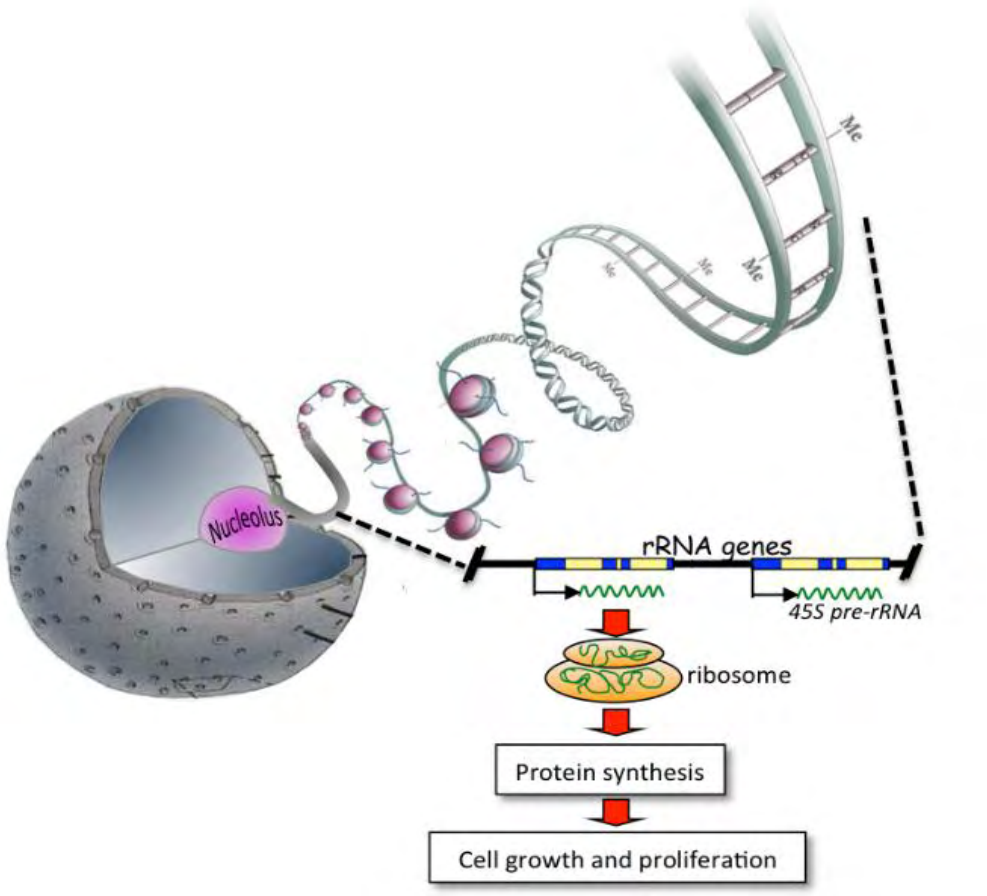
πρόσδεση αμινοξέων καθώς και μια περιοχή αντικωδικονίου, για την αναγνώριση του τρέχοντος κωδικονίου στο αγγελιαφόρο RNA. Η αναγνώριση αυτή γίνεται με βάση τη συμπληρωματικότητα των βάσεων. Ένας τύπος tRNA αντιστοιχεί σε πολλά κωδικόνια αλλά μόνο σε ένα αμινοξύ.



***Εικόνα 6**: Απεικόνιση tRNA*

Πηγή: Google-εικόνες tRNA

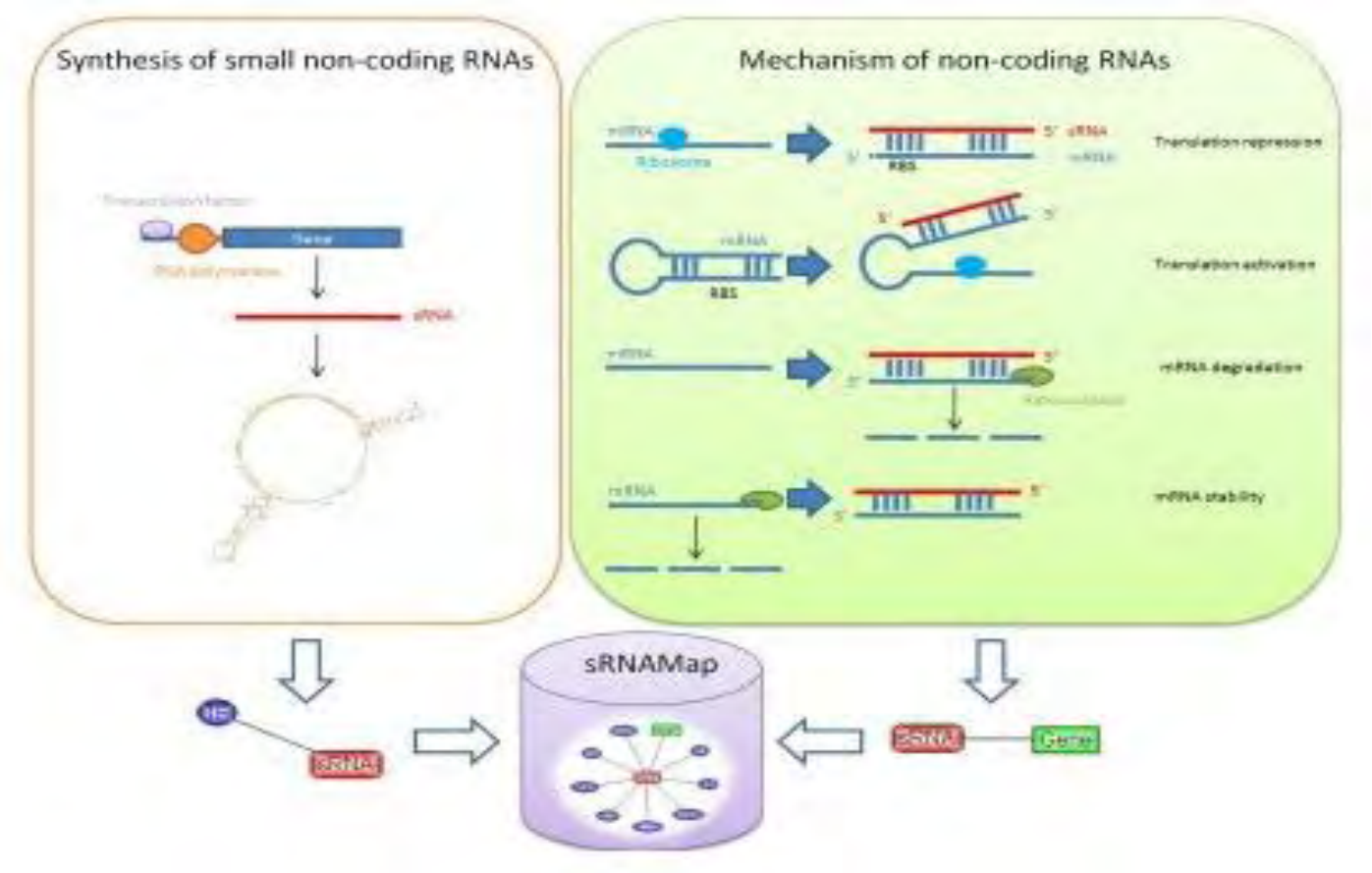
- **Ριβοσωμικό RNA rRNA** (Ribosomal RNA)
είναι ένας τύπος RNA των ριβοσωμάτων που καταλύει την πρωτεϊνοσύνθεση στο κύτταρο. Τα ευκαρυωτικά ριβοσώματα αποτελούν το εργοστάσιο πρωτεϊνοσύνθεσης του κυττάρου και περιέχουν τέσσερα διαφορετικά μόρια rRNA: τα 18S, 5.8S, 28S, και 5S rRNA. Τρία από αυτά τα μόρια rRNA συντίθενται στον πυρηνίσκο του κυττάρου. Τα διαφορετικά είδη rRNA είναι πολυάριθμα στο κύτταρο και αποτελούν το 80% των ολικών RNA σε ένα τυπικό ευκαρυωτικό κύτταρο. Στο κυτταρόπλασμα το ριβοσωμικό RNA συνδυάζεται με ειδικές πρωτεΐνες του πλάσματος και συνθέτουν ένα νουκλεοπρωτεϊνικό σύμπλεγμα που ονομάζεται ριβόσωμα. Το ριβόσωμα συνδέεται με το mRNA και είναι υπεύθυνο για την πρωτεϊνική σύνθεση, όπου και λαμβάνει χώρα. Ορισμένα ριβοσώματα μπορούν να είναι συνδεδεμένα με μία μονή αλυσίδα mRNA σε όλη τη διάρκεια της ζωής τους.



***Εικόνα 7**: Απεικόνιση rRNA*

Πηγή: Google-εικόνες rRNA

- **Μη κωδικοποιητικό RNA** (Non-coding RNA)
  είναι ένα μόριο RNA που δεν μεταφράζεται σε πρωτεΐνη (όπως το αγγελιαφόρο RNA), αλλά έχει κάποιον άλλο λειτουργικό ρόλο στο κύτταρο. Τα πιο αντιπροσωπευτικά είδη μη κωδικοποιητικού RNA είναι το μεταφορικό RNA (tRNA) και το ριβοσωμικό RNA (rRNA), που παίζουν σημαντικό υποστηρικτικό ρόλο στην μετάφραση. Η περιοχή του γενετικού κώδικα όπου κωδικοποιείται ένα μη κωδικοποιητικό RNA (αντί για μία πρωτεΐνη), ονομάζεται «γονίδιο RNA».



***Εικόνα 8**: Απεικόνιση μη κωδικοποιητικού RNA*

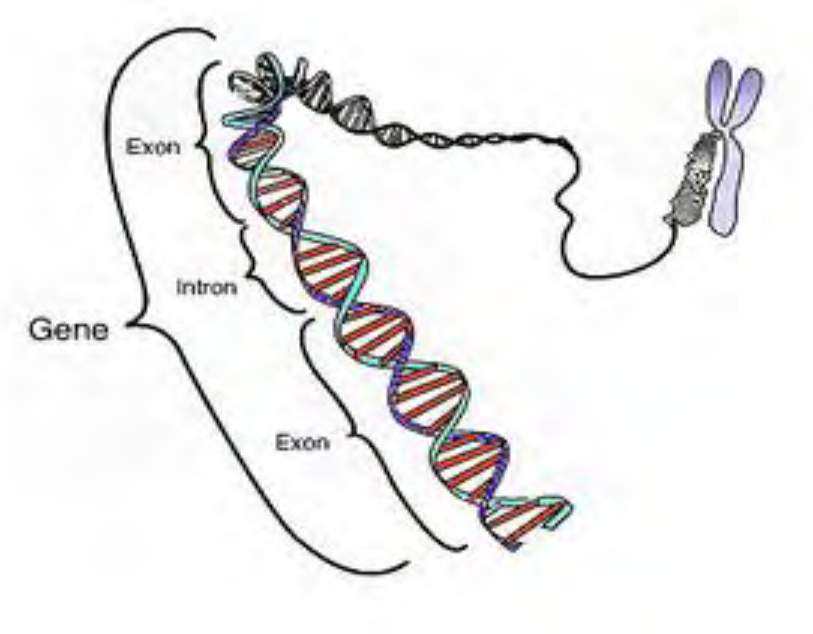
Πηγή: Google-εικόνες μη κωδικοποιητικού RNA

- **Καταλυτικό RNA** (Catalytic RNA)
  είναι ένα είδος RNA που καταλύει μια χημική αντίδραση. Παρότι το RNA περιέχει μόνο τέσσερις αζωτούχες βάσεις σε σχέση με τα 20 αμινοξέα που απαντώνται στις πρωτεΐνες, ορισμένα είναι ικανά να καταλύουν ειδικές χημικές αντιδράσεις που λαμβάνουν χώρα στο κύτταρο. Τέτοιες αντιδράσεις περιλαμβάνουν την διάσπαση και επανένωση άλλων μορίων RNA και επίσης η κατάλυση του πεπτιδικού δεσμού που λαμβάνει χώρα στα ριβοσώματα του κυτταροπλάσματος.

2.4.3 Γονίδιο

Το γονίδιο είναι η βασική φυσική μονάδα κληρονομικότητας στους ζωντανούς οργανισμούς η οποία και μεταβιβάζει πληροφορίες από το ένα κύτταρο σε άλλο και κατ᾽ επέκταση από τη μια γενιά στην άλλη. Τα γονίδια συνθέτουν το γονιδίωμα ενός οργανισμού, που αποτελείται από το DNA και το RNA, και κατευθύνουν τη φυσική ανάπτυξη και συμπεριφορά του οργανισμού. Τα περισσότερα γονίδια κωδικοποιούν πρωτεΐνες. Μερικά γονίδια δεν κωδικοποιούν πρωτεΐνες, αλλά τα μόρια του RNA διαδραματίζουν βασικούς ρόλους στην βιοσύνθεση πρωτεϊνών και στον έλεγχο της γονιδιακής έκφρασης. Τα μόρια που προκύπτουν από την γονιδιακή έκφραση, είτε RNA είτε πρωτεΐνη, είναι γνωστά ως γονιδιακά προϊόντα.

Τα περισσότερα γονίδια περιέχουν κάποιες περιοχές που δεν κωδικοποιούν γονιδιακά προϊόντα, αλλά συχνά ρυθμίζουν τη γονιδιακή έκφραση. Μια κρίσιμη περιοχή μη-κωδικοποίησης είναι το γονίδιο-υποκινητής, μια σύντομη ακολουθία DNA, απαραίτητη για την έναρξη της γονιδιακής έκφρασης. Στα γονίδια ευκαρυωτικών οργανισμών περιέχονται συχνά κάποιες περιοχές που αποκαλούνται ιντρόνια ή εσόνια και αφαιρούνται από το mRNA σε μια διαδικασία γνωστή ως συγκόλληση ή ωρίμανση του mRNA, που γίνεται στον πυρήνα, (splicing). Οι περιοχές που κωδικοποιούν πραγματικά το προϊόν γονιδίων, το οποίο μπορεί να είναι πολύ μικρότερο από τα ιντρόνια, είναι γνωστές ως εξόνια. Έχει καθιερωθεί διεθνώς οτιδήποτε αναφέρεται σε, ή έχει σχέση με γονίδια, να χαρακτηρίζεται γενετικός ή γενετικό (genetic).



***Εικόνα 9**: Απεικόνιση γονιδίου*

Πηγή: http://el.wikipedia.org/wiki/%CE%93%CE%BF%CE%BD%CE%AF%CE%B4%CE%B9%CE%B

2.4.4 Πρωτεΐνες

Οι πρωτεΐνες είναι μεγάλα σύνθετα βιομόρια, με μοριακό βάρος από 10.000 μέχρι πάνω από 1 εκατομμύριο, αποτελούμενα από αμινοξέα, τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας μια γραμμική αλυσίδα, καλούμενη αλυσίδα πολυπεπτιδίων. Όλες οι πρωτεΐνες περιέχουν άνθρακα, οξυγόνο και άζωτο και οι περισσότερες εξ αυτών και θείο. Η ακολουθία αμινοξέων σε μια πρωτεΐνη καθορίζεται από ένα γονίδιο και κωδικοποιείται κατά τον γενετικό κώδικα DNA. Παρόλο που ο γενετικός κώδικας κωδικοποιεί 20 αμινοξέα, τα αμινοξέα που συνιστούν την πρωτεΐνη συχνά υφίστανται χημικές αλλαγές κατά τη μετά-μεταγραφική τροποποίηση: είτε προτού να μπορέσει η πρωτεΐνη να λειτουργήσει στο κύτταρο, είτε ως τμήμα των μηχανισμών ελέγχου. Περισσότερες από μια πρωτεΐνες συχνά λειτουργούν μαζί για να επιτύχουν κάποια συγκεκριμένη λειτουργία, ή μπορεί ακόμα και να συσσωματωθούν για να διαμορφώσουν τα σταθερά σύμπλοκα.

Εικόνα 10 *: Αναπαράσταση της τρισδιάστατης δομής της μυογλοβίνης, που παρουσιάζεται με χρωματισμένες τις άλφα έλικες. Αυτή ήταν η πρώτη πρωτεΐνη, η δομή της οποίας προσδιορίστηκε με κρυσταλλογραφία ακτινών X*

Πηγή: http://el.wikipedia.org/wiki/%CE%A0%CF%81%CF%89%CF%84%CE%B5%CE%90%CE%BD%CE%B7

.

Διαφορετικά βιολογικά προβλήματα που εμπίπτουν στο πεδίο της βιοπληροφορικής περιλαμβάνουν τη μελέτη των γονιδίων, πρωτεϊνών, πρόβλεψη δομής νουκλεϊκών οξέων. Στην παρούσα διπλωματική εργασία θα ασχοληθούμε με τα προβλήματα:

1) Χαρτογράφηση περιορισμού
2) Εύρεση μοτίβων
3) Αναδιατάξεις γονιδιώματος
4) Γονιδιακή πρόγνωση
5) Σύγκριση ακολουθιών
6) Μοριακή εξέλιξη

χρησιμοποιώντας αλγορίθμους εξαντλητικής αναζήτησης και άπληστους, δυναμικό προγραμματισμό, αλγορίθμους ομαδοποίησης και δέντρα.

3.Αλγόριθμοι διεξοδικής αναζήτησης (brute-force)

Με βάση την τεχνική των αλγορίθμων διεξοδικής αναζήτησης (brute-force), το πρόβλημα επιλύεται με τον πιο απλό και προφανή τρόπο, ο οποίος όμως δεν είναι πάντα (σχεδόν σε όλες τις περιπτώσεις) και ο πιο καλός. Το γεγονός ότι επιλύει τα προβλήματα με τον πιο απλό και προφανή δυνατό τρόπο καθιστά τον αλγόριθμο κατανοητό και εύκολα υλοποιήσιμο . Μπορεί να είναι πολύ καλή και πρακτική μέθοδος για προβλήματα που η είσοδος είναι μικρού ή μεσαίου μεγέθους.
Χαρακτηριστικό παράδειγμα για τη διευκρίνιση της τεχνικής brute-force είναι το πως θα υπολογίζαμε μια δύναμη, για παράδειγμα τη δύναμη a^{16} :

Brute-Force:

$$\underbrace{a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a \cdot a}_{16}$$

Ενώ με κάποια άλλη τεχνική (πολύ καλύτερη) θα μπορούσε να υπολογισθεί σαν:

Καλύτερη τεχνική:

$$(((a^2)^2)^2)^2$$

δηλαδή 4 πολλαπλασιασμοί αντί για 16.

Θα αναφερθούμε σε δύο βιολογικά προβλήματα:

- τη χαρτογράφηση περιορισμού του DNA και
- την εύρεση μοτίβων

των οποίων οι λύσεις, με αλγόριθμο διεξοδικής αναζήτησης, δεν είναι πρακτικές. Έπειτα θα χρησιμοποιήσουμε την τεχνική branch-and-bound για να μετατρέψουμε έναν αναποτελεσματικό αλγόριθμο brute-force σε αποτελεσματικό.

3.1 Χαρτογράφηση περιορισμού

3.1.1 Ένζυμα περιορισμού

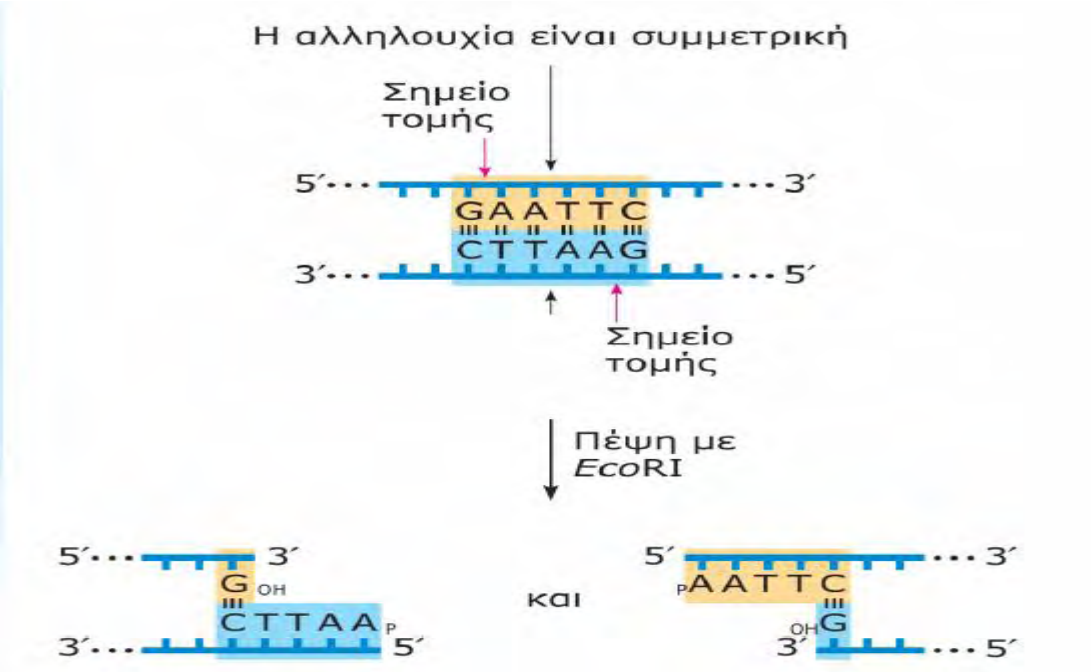
Πριν αναλύσουμε τη διαδικασία χαρτογράφησης του DNA θα πρέπει να αναφερθούμε στα ένζυμα περιορισμού που διαδραμάτισαν σημαντικό ρόλο σε αυτή.
Τα ένζυμα περιορισμού (restriction enzymes) είναι ενδονουκλεάσες που απομονώθηκαν από διάφορους προκαρυωτικούς οργανισμούς, κυρίως βακτήρια και παίρνουν το όνομα τους από το είδος του βακτηρίου από το οποίο απομονώθηκαν (π.χ. EcoRI από την Escherichia coli, Sau3A από τον Staphylococcus aureus). Προστατεύουν τους μικροοργανισμούς αυτούς από την εισβολή ξένου DNA, και έχουν ως ιδιότητα να πετούν το DNA σε συγκεκριμένες αλληλουχίες.
Τα ένζυμα περιορισμού χαρακτηρίζονται από τη θέση αναγνώρισης και το σημείο κοπής.

Θέσεις αναγνώρισης: είναι συνήθως 4 ή 6 bp σε μήκος και παλινδρομικές (π.χ. CTGCAG ή GGCC).

Σημείο κοπής: για ορισμένα ένζυμα περιορισμού είναι στο μέσο της θέσης αναγνώρισης, οπότε δημιουργούνται **λεία άκρα** (blunt ends). Άλλα ένζυμα κόβουν τις δύο αλυσίδες του DNA σε συμμετρικά έκκεντρα σημεία (staggered cuts) μία ή δύο βάσεις από το μέσο, οπότε δημιουργούνται προεξέχοντα άκρα 2 ή 4 βάσεων αντίστοιχα. Τα προεξέχοντα άκρα είναι προς την 5' ή την 3' κατεύθυνση, ανάλογα με το ένζυμο. Επειδή οι αλληλουχία της μονόκλωνης προέκτασης είναι παλινδρομική, μπορεί να συστοιχίζεται με τον εαυτό της και έτσι αυτά τα άκρα αποκαλούνται **κολλώδη άκρα** (sticky ends).

Μία ενδονουκλεάση περιορισμού κόβει την εξειδικευμένη θέση αναγνώρισής της (την **θέση περιορισμού** της) οπουδήποτε αυτή συναντάται μέσα σε ένα δείγμα DNA, εκτός αν η θέση αυτή είναι προστατευμένη με μεθυλίωση σε μία ή περισσότερες από τις βάσεις που την αποτελούν (η μεθυλίωση κατά τα άλλα δεν επηρεάζει τις λειτουργίες του DNA).

Στην παρακάτω εικόνα (εικόνα 11) φαίνεται η θέση αναγνώρισης της *Eco*RI. Η αλληλουχία είναι παλίνδρομη είναι ίδια δηλαδή και στις δύο αλυσίδες του DNA όταν διαβάζεται στην ίδια κατεύθυνση (στο παράδειγμα αυτό είναι 5′ GAATTC 3΄).
Στην εικόνα 12 παρουσιάζονται διάφορα ένζυμα περιορισμού με την αλληλουχία αναγνώρισης και τη θέση κοπής τους.



Εικόνα 11 *: Παράδειγμα πέψης DNA από το ένζυμο περιορισμού EcoRI*

Πηγή: Διάλεξη 1 («ενδονουκλεάσες περιορισμού») του μαθήματος «Μοριακή Βιολογία», Δρ. Χρήστος Παναγιωτίδης-Τμήμα φαρμακευτικής

	Ονομασία ενζύμου	Οργανισμός από τον οποίο προέρχεται το ένζυμο	Αλληλουχία αναγνώρισης και θέση κοπής*
Ένζυμα με αλληλουχία αναγνώρισης 6 bp	*Bam*HI	*Bacillus amyloliquefaciens H*	5′-G↓G A T C C-3′ 3′-C C T A G↑G-5′
	*Bgl*II	*Bacillus globigi*	A↓G A T C T T C T A G↑A
	*Eco*RI	*E. coli RY13*	G↓A A T T C C T T A A↑G
	*Hae*II	*Haemophilus aegypticus*	R G C G C↓Y Y↑C G C G R
	*Hin*dIII	*Haemophilus influenzae* R_d	A↓A G C T T T T C G A↑A
	*Pst*I	*Providencia stuartii*	C T G C A↓G G↑A C G T C
	*Sal*I	*Streptomyces albus*	G↓T C G A C C A G C T↑G
	*Sma*I	*Serratia macrescens*	C C C↓G G G G G G↑C C C
Ένζυμα με αλληλουχία αναγνώρισης 4 bp	*Hae*III	*Haemophilus aecypticus*	G G↓C C C C↑G G
	*Hha*I	*Haemophilus haemolyticus*	G C G↓C C↑G C G
	*Hpa*II	*Haemophilus parainfluenzae*	C↓C G C G G C↑C
	*Sau*3A	*Staphylococcus aureus 3A*	↓G A T C C T A G↑
Ένζυμο με αλληλουχία αναγνώρισης 8 bp	*Not*I	*Nocardia otitidis-caviarum*	G C↓G G C C G C C G C C G G↑C G
Ένζυμο με μη συμμετρική αλληλουχία αναγνώρισης	*Bst*XI	*Bacillus stearothermophilus*	C C A N N N N N↓N T G G G G T N↑N N N N N A C C

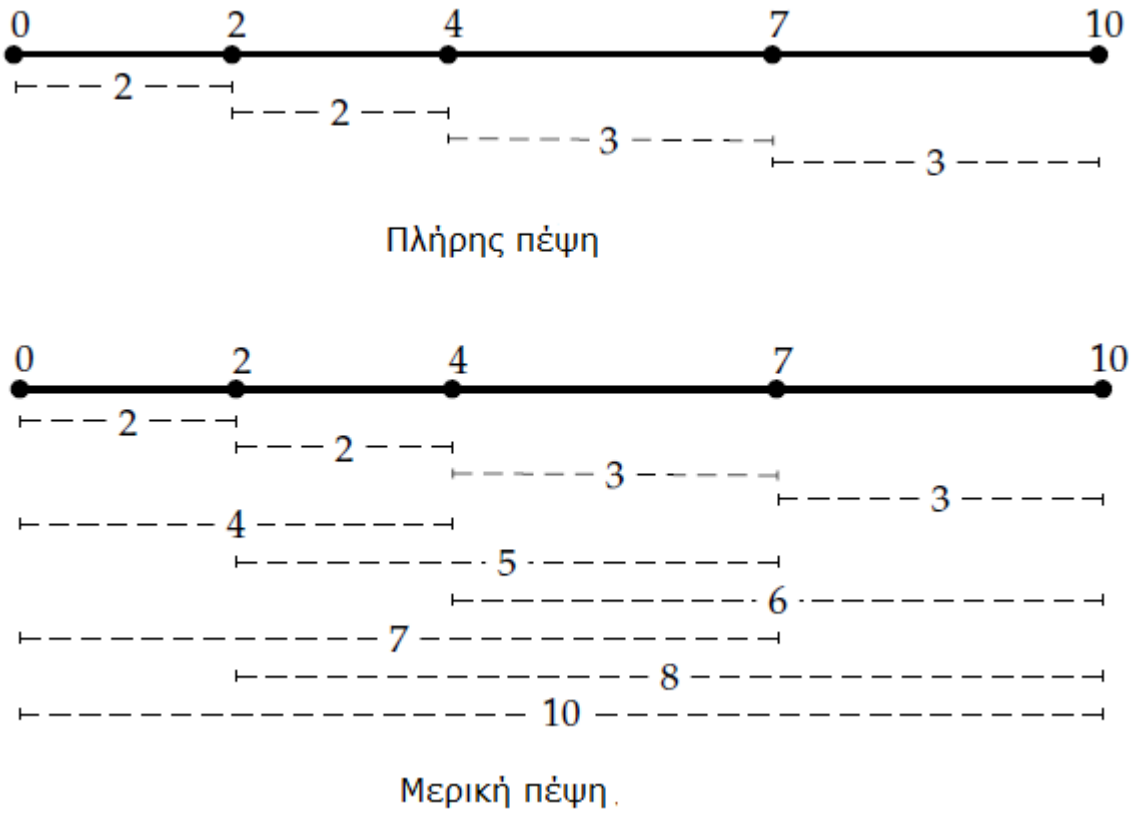
*Σε αυτή τη στήλη παρουσιάζονται οι δύο αλυσίδες του DNA και οι θέσεις κοπής υποδεικνύονται με βέλη.
R = πουρίνη, Y = πυριμιδίνη, N = οποιαδήποτε βάση.

Εικόνα 12 Χαρακτηριστικά μερικών ενζύμων περιορισμού

Πηγή: Διάλεξη 1 («ενδονουκλεάσες περιορισμού») του μαθήματος «Μοριακή Βιολογία», Δρ. Χρήστος Παναγιωτίδης-Τμήμα φαρμακευτικής

3.1.2 Χάρτες περιορισμού

Τα ένζυμα περιορισμού είναι ιδανικά αντιδραστήρια για χαρτογράφηση του DNA, αφού κάθε αναγνωριζόμενη αλληλουχία (**θέση περιορισμού**) παίζει τον ρόλο οροσήμου πάνω στο DNA. Στην πράξη, επωάζουμε ένα δείγμα DNA με κάποιο περιοριστικό ένζυμο και μετράμε το μέγεθος των θραυσμάτων μετά από ηλεκτροφόρηση σε πήκτωμα αγοράζης (η ταχύτητα μετακίνησης μέσα στο πήκτωμα είναι ανάλογη με το αρνητικό του λογαρίθμου του μήκους του θραύσματος). Με συνδυασμό απλών και διπλών (χρησιμοποιώντας συγχρόνως δύο περιοριστικά ένζυμα) πέψεων πετυχαίνουμε χαρτογράφηση των περιοριστικών θέσεων στο δείγμα μας.
Υπάρχουν 2 είδη πέψης του DNA, η πλήρης και η μερική. Η διαφορά τους είναι ότι στην πλήρη πέψη δημιουργούνται θραύσματα μόνο μεταξύ 2 διαδοχικών σημείων κοπής ενώ στην μερική μεταξύ οποιονδήποτε δύο σημείων κοπής



Εικόνα 13 : *Σχηματική απεικόνιση των διαφόρων τύπων πέψης του DNA*

Πηγή: *An introduction to Bioinformatics Algorithms, Neil C. Jones and Pavel A. Pevzner*

3.1.3 Διατύπωση προβλήματος χαρτογράφησης περιορισμού με μερική πέψη (partial digest problem):

Εάν X = {x_1 = 0, x_2,. . . , x_n} είναι ένα σύνολο n σημείων σε αύξουσα σειρά σε ένα ευθύγραμμο τμήμα,
τότε το ΔX το σύνολο όλων των αποστάσεων μεταξύ δύο σημείων του συνόλου X

$$X = \{xj - xi : 1 <= i < j <= n\} .$$

Για παράδειγμα, αν X = {0, 4, 14, 19, 22}, τότε ΔX = {3, 4, 5, 8, 10, 14, 15, 18, 19, 22}, που είναι τα δέκα ζεύγη αποστάσεων μεταξύ των σημείων του X.

	0	4	14	19	22
0		4	14	19	22
4			10	15	18
14				5	8
19					3
22					

***Εικόνα 14**: Αναπαράσταση του ΔX = {3, 4, 5, 8, 10, 14, 15, 18, 19, 22} σε ένα δισδιάστατο πίνακα, με τα στοιχεία του X = {0, 4, 14, 19, 22} στην πρώτη γραμμή και στην πρώτη στήλη του πίνακα.. Το στοιχείο στη θέση (i, j) του πίνακα έχει τιμή xj - xi όπου 1<= i < j<=n*

Στη χαρτογράφηση περιορισμού μας δίνεται το ΔX, το οποίο πρόκειται για τα πειραματικά δεδομένα σχετικά με τα μήκη των θραυσμάτων.
Το πρόβλημα είναι να ανακατασκευάσουμε X από το ΔX. Δηλαδή, για το παραπάνω παράδειγμα το πρόβλημα είναι πως μπορούμε να συμπεράνουμε ότι το σύνολο ΔX = {3, 4, 5, 8, 10, 14, 15, 18, 19, 22} προήλθε από το X={0, 4, 14 , 19, 22};
Αν και η διαδικασία ηλεκτροφόρησης μας επιτρέπει να καθορίσουμε εύκολα τα μήκη των θραυσμάτων του DNA, είναι δύσκολο να διακρίνουμε την πολλαπλότητα τους. Δηλαδή, είναι δύσκολο να προσδιοριστεί ο αριθμός των διαφόρων τμημάτων ενός συγκεκριμένου μήκους. Ωστόσο, πειραματικά και με πολλή δουλειά είναι δυνατόν να προσδιοριστεί. Χάριν απλότητας του προβλήματος, στα πλαίσια αυτής της διπλωματικής, θα θεωρήσουμε ότι η πληροφορία αυτή παρέχεται σαν είσοδος στο πρόβλημα.

3.1.4 Αλγόριθμος (Partial Digest algorithm)

Το 1990, ο Steven Skiena περιέγραψε έναν αλγόριθμο για το πρόβλημα της χαρτογράφησης περιορισμού με μερική πέψη (partial digest problem ή PDP). Ο αλγόριθμος αυτός παίρνει ως είσοδο τη λίστα L με όλες τις αποστάσεις των θραυσμάτων και επιστρέφει το σύνολο X τέτοιο ώστε ΔX=L.
Βήματα αλγορίθμου:

1) Δημιουργούμε το σύνολο X με αρχικό στοιχείο το 0 (X={0})
2) Βρίσκουμε το μέγιστο στοιχείο της λίστας L (έστω max)
3) Βρίσκουμε τις αποστάσεις του max από τα ακριανά σημεία του X (την πρώτη φορά μόνο από το 0)
4) Διαλέγουμε μία από τις παραπάνω αποστάσεις (έστω y)
5) Αν η απόσταση του y από τα σημεία του συνόλου X αποτελούν στοιχεία του L, τότε τοποθετούμε το y στο σύνολο X και διαγράφουμε από το L τόσο το y όσο και τα στοιχεία-αποστάσεις του y από το X.
6) Αν η παραπάνω συνθήκη δεν ισχύει προχωράμε τη διαδικασία με το επόμενο μέγιστο στοιχείο του L.
7) Όταν η λίστα L μείνει κενή τότε έχουμε τη λύση του προβλήματος, η οποία είναι το σύνολο X
8) Σε περίπτωση που αλγόριθμος δεν καταφέρνει να βρει λύση οπισθοδρομούμε και στο βήμα 4 παίρνουμε άλλη απόφαση για το y.
9) Με τη διαδικασία της οπισθοδρόμησης μπορούμε να βρούμε ακόμα να βρούμε και εναλλακτική λύση, δηλαδή άλλο σύνολο X.

Κώδικας

```
PARTIALDIGEST(L)
    width ← Maximum element in L
    DELETE(width, L)
    X ← {0, width}
    PLACE(L, X)

PLACE(L, X)
    if L is empty
        output X
        return
    y ← Maximum element in L
    if Δ(y, X) ⊆ L
        Add y to X and remove lengths Δ(y, X) from L
        PLACE(L, X)
        Remove y from X and add lengths Δ(y, X) to L
    if Δ(width − y, X) ⊆ L
        Add width − y to X and remove lengths Δ(width − y, X) from L
        PLACE(L, X)
        Remove width − y from X and add lengths Δ(width − y, X) to L
    return
```

Πηγή: *An introduction to Bioinformatics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Παράδειγμα

Για να κατανοήσουμε τον παραπάνω αλγόριθμο ας δούμε ένα παράδειγμα. Έστω:

L= {3, 4, 5, 8, 10, 14, 15, 18, 19, 22}

Ορίζουμε αρχικά το X:

X= {0}

Το μέγιστο στοιχείο του L είναι το 22. Η απόσταση του 22 από το στοιχείο 0 του X είναι 22, στοιχείο το οποίο υπάρχει στο L. Επομένως τώρα έχουμε:

L= {3, 4, 5, 8, 10, 14, 15, 18, 19}

X= {0, 22}

Προχωράμε στο επόμενο μέγιστο στοιχείο του L που είναι το 19. Ακολουθώντας το βήμα 3 έχουμε y=|0-19|=19 ή y=|22-19|=3. Διαλέγουμε το y=19 και βρίσκουμε τις αποστάσεις του από τα σημεία του X.

D(y, X) = {3, 19}. Σβήνουμε τα στοιχεία 3 και 19 από το L και προσθέτουμε το 19 στο σύνολο X. Έτσι έχουμε:

L= {4, 5, 8, 10, 14, 15, 18}

X= {0, 19, 22}.

Συνεχίζοντας τη διαδικασία με το 18, έχουμε ότι y=|0-18|=18 ή y=|22-18=4. Διαλέγουμε το y=4 του οποίου η απόσταση από τα σημεία του X είναι D(y, X)={4, 15, 18}. Με τον ίδιο τρόπο όπως και προηγουμένως τα σύνολα μας τώρα μετατρέπονται στα:

L= {5, 8, 10, 14}

X= {0, 4, 19, 22}.

Παίρνοντας το επόμενο μέγιστο στοιχείο (το 14) του L έχουμε y=14 ή y=8. Διαλέγοντας το y=14 έχουμε D(y, X) = {14, 10, 5, 8}. Επομένως τα σύνολά μας τώρα έχουν ως εξής:

L= {}

X= {0, 4, 14, 19, 22}

Παρατηρούμε ότι η λίστα L έμεινε άδεια επομένως ο αλγόριθμος έφτασε στο τέλος του και η λύση μας είναι το σύνολο

X= {0, 4, 14, 19, 22}

Πολυπλοκότητα

Μετά από κάθε αναδρομική κλήση της συνάρτησης PLACE , αναιρούμε τις αλλαγές στα σύνολα X και L, ώστε να τα χρησιμοποιήσουμε στην επόμενη αναδρομική κλήση. Είναι σημαντικό να τονίσουμε ότι αυτός ο αλγόριθμος θα εμφανίσει όλα τα σύνολα X για τα οποία ισχύει ΔX = L.

Ο παραπάνω ο αλγόριθμος με την πρώτη ματιά, φαίνεται αποτελεσματικός καθώς σε κάθε σημείο εξετάζουμε δύο εναλλακτικές λύσεις (την απόσταση από το αριστερό άκρο του X και την απόσταση από το δεξί άκρο), αποκλείοντας τις προφανώς εσφαλμένες αποφάσεις. Στις περισσότερες περιπτώσεις του Partial Digest Problem (PDP), ο αλγόριθμος είναι πολύ γρήγορος αφού συνήθως μόνο μία από τις δύο εναλλακτικές λύσεις είναι βιώσιμη σε οποιαδήποτε φάση. Για αρκετά χρόνια δεν ήταν σαφές αν ο αλγόριθμος αυτός είχε ή όχι πολυωνυμική πολυπλοκότητα στη χειρότερη περίπτωση. Αυτό, γιατί μερικές φορές και οι δύο εναλλακτικές λύσεις είναι βιώσιμες. Αν, λοιπόν, και οι δύο εναλλακτικές λύσεις είναι

βιώσιμες και αυτό συνεχίζει να συμβαίνει κατά τη διάρκεια των βημάτων του αλγορίθμου, η απόδοση του αλγορίθμου αυξάνεται σε $2^{κ}$, όπου κ είναι ο αριθμός αυτών των "διφορούμενων" βημάτων.
Έστω T (n) ο μέγιστος χρόνος που χρειάζεται ο PartialDigest για να βρει τη λύση για ένα PDP n σημείων. Αν υπάρχει μόνο μία βιώσιμη εναλλακτική λύση σε κάθε βήμα, τότε ο PartialDigest μειώνει σταθερά το μέγεθος του προβλήματος κατά ένα και αυτοκαλείται αναδρομικά. Έτσι έχουμε:

$$T(n) = T(n - 1) + O(n),$$

όπου O (n) είναι το έργο που δαπανάται για την προσαρμογή των συνόλων X και L.

Ωστόσο, αν υπάρχουν δύο εναλλακτικές λύσεις, τότε:

$$T(n) = 2T(n - 1) + O(n).$$

Ενώ οι εκφράσεις T (n) = T (n - 1) + O (n) και T (n) = 2T (n - 1) + O (n) φέρουν μια επιφανειακή ομοιότητα ως προς τη μορφή, ο καθένας οδηγεί σε πολύ διαφορετικές εκφράσεις για το χρόνο λειτουργίας του αλγορίθμου. Η μια πρόκειται για τετραγωνική πολυπλοκότητα ενώ η άλλη για εκθετική. Στην πραγματικότητα, αλγόριθμοι με πολυωνυμική πολυπλοκότητα για το PDP ήταν άγνωστοι μέχρι το 2002, όταν ο Maurice Nivat και οι συνεργάτες του σχεδίασαν τον πρώτο.

3.2 Εύρεση μοτίβων

Κάθε μόριο DNA αναπαρίσταται ως μια ακολουθία συμβόλων (συμβολοσειρά), από το αλφάβητο των τεσσάρων χαρακτήρων {A,T,C,G} όπου το A χρησιμοποιείται για την αδενίνη, το T για την θυμίνη, το C για την κυτοσίνη και το G για την γουανίνη. Ο προσδιορισμός αυτής της συμβολοσειράς για διαφορετικά μόρια ή ο προσδιορισμός της σειράς των συμβόλων βάσεων στα μόρια, είναι κρίσιμος για την κατανόηση των βιολογικών λειτουργιών των μορίων.
Τα ακολουθιακά μοτίβα είναι μικρά, επαναλαμβανόμενα μοτίβα στο DNA τα οποία εκτιμάται ότι έχουν βιολογική λειτουργικότητα. Κάποια δείχνουν ειδικές θέσεις πρόσδεσης στην ακολουθία για πρωτεΐνες, όπως οι νουκλεάσες και οι μεταγραφικοί παράγοντες (TF). Κάποια άλλα συμμετέχουν σε σημαντικές διεργασίες στο επίπεδο του RNA , συμπεριλαμβάνοντας τη σύνδεση με το ριβόσωμα, την επεξεργασία των mRNAs και τον τερματισμό της μεταγραφής.

3.2.1 Προφίλ

Στην παρακάτω εικόνα (εικόνα 15.α) παρουσιάζονται επτά τυχαίες αλληλουχίες DNA 32 νουκλεοτιδίων. Επίσης, φαίνονται (εικόνα 15.β) οι ίδιες ακολουθίες με εμφυτευμένο το «μυστικό μοτίβο» P = TCGACTAC μήκους L = 8 σε τυχαίες θέσεις.
Υποθέτοντας ότι δεν ξέρουμε ποιο είναι το μοτίβο P, ή σε ποιο σημείο κάθε αλληλουχίας έχει εμφυτευθεί (εικόνα 15.γ) τίθεται το ερώτημα αν μπορούμε να ανακατασκευάσουμε το P από την ανάλυση των αλληλουχιών του DNA.

AGTCCGATCGAATCGATCGATCGATTCGATCG
TTCGATCAATGCTAGCTACCGTACGTAGCTTA
CCTAGGCTTAGCTACGTCCTTTAAAGCTATCG
TTTGATCGATCCGATCGACGACGACCACTGAT
AACTAGTCGATCGATCCTTGATCGATCGATCG
TCTCTCATCAATCGATCGAACTGACTATCGAT
TACTACTCGAATCGATCGATCCTGACTGACTG

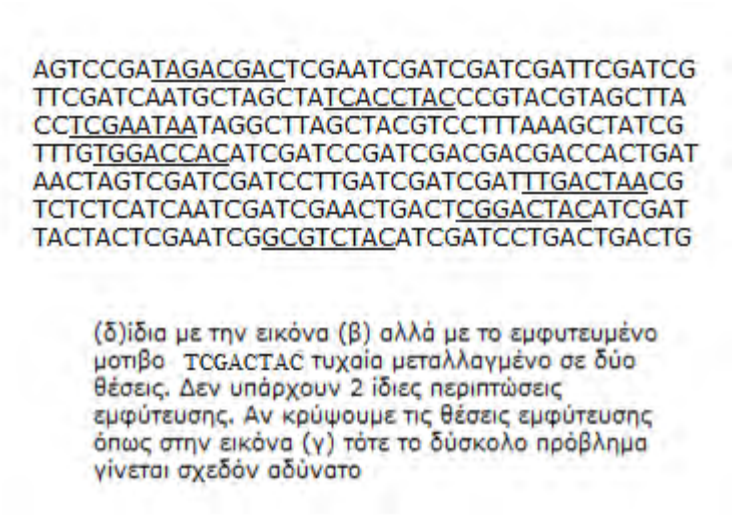
(a) 7 τυχαίες αλληλουχίες DNA

AGTCCGATCGACTACTCGAATCGATCGATCGATTCGATCG
TTCGATCAATGCTAGCTATCGACTACCCGTACGTAGCTTA
CCTCGACTACTAGGCTTAGCTACGTCCTTTAAAGCTATCG
TTTGTCGACTACATCGATCCGATCGACGACGACCACTGAT
AACTAGTCGATCGATCCTTGATCGATCGATTCGACTACCG
TCTCTCATCAATCGATCGAACTGACTTCGACTACATCGAT
TACTACTCGAATCGTCGACTACATCGATCCTGACTGACTG

(β) οι ίδιες αλληλουχίες με εμφυτευμένο το μοτίβο TCGACTAC

AGTCCGATCGACTACTCGAATCGATCGATCGATTCGATCG
TTCGATCAATGCTAGCTATCGACTACCCGTACGTAGCTTA
CCTCGACTACTAGGCTTAGCTACGTCCTTTAAAGCTATCG
TTTGTCGACTACATCGATCCGATCGACGACGACCACTGAT
AACTAGTCGATCGATCCTTGATCGATCGATTCGACTACCG
TCTCTCATCAATCGATCGAACTGACTTCGACTACATCGAT
TACTACTCGAATCGTCGACTACATCGATCCTGACTGACTG

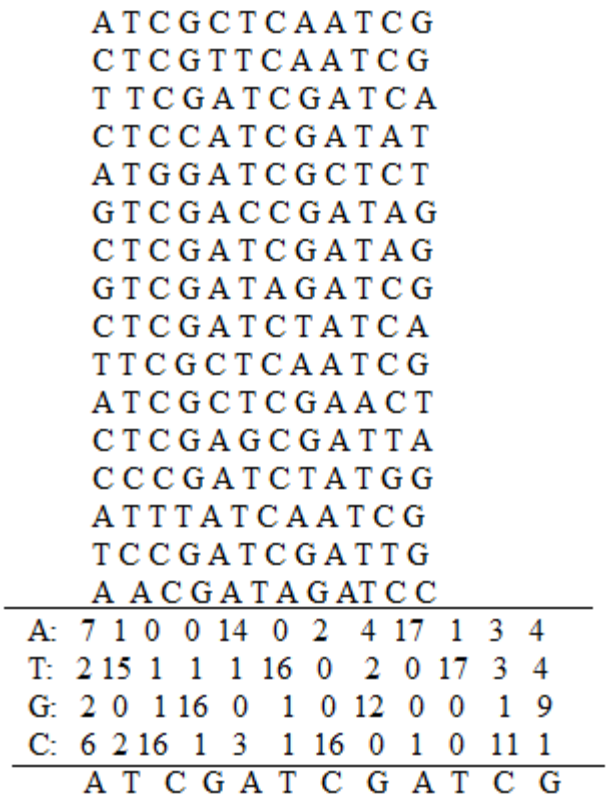
(γ)Ιδια με την εικόνα (β) με κρυμμένες τις θέσεις εμφύτευσης του μοτίβου



Εικόνα 15 *: DNA ακολουθίες με εμφυτευμένα μοτίβα*

Προκειμένου να απαντήσουμε στο ερώτημα αυτό θα μπορούσαμε απλά να μετρήσουμε τον αριθμό των φορών που κάθε συμβολοσειρά μήκους l, υπάρχει στο δείγμα. Δεδομένου ότι υπάρχουν μόνο $7 \cdot (32 + 8) = 280$ νουκλεοτίδια στο δείγμα, είναι απίθανο οποιοδήποτε άλλο 8-μερές εκτός από το εμφυτευμένο μοτίβο να εμφανίζεται περισσότερο από μία φορά (η πιθανότητα εμφάνισης οποιουδήποτε 8-μερές στο δείγμα είναι μικρότερη από 280 / 48= 0.004). Μετά την καταμέτρηση όλων των εμφανίσεων 8-μερών στο δείγμα παρατηρούμε ότι παρόλο που τα περισσότερα 8-μερή εμφανίζονται στο δείγμα μόνο μια φορά (με λίγα να εμφανίζονται δύο φορές), υπάρχει ένα 8-μερές που ύποπτα εμφανίζεται στο δείγμα πολλές φορές - επτά ή περισσότερες. Αυτό το πολυεμφανιζόμενο 8-μερές είναι το μοτίβο P που προσπαθούμε να βρούμε.

Σε αντίθεση με την παραπάνω απλή εμφύτευση μοτίβων, το DNA χρησιμοποιεί μια πιο εφευρετική έννοια των μοτίβων, η οποία επιτρέπει μετάλλαξη σε κάποιες θέσεις νουκλεοτιδίων (εικόνα 15.δ). Για παράδειγμα, στον πίνακα της εικόνας 16 παρουσιάζονται δεκαοχτώ διαφορετικά NF-κB μοτίβα. Παρατηρούμε ότι παρόλο που καμιά ακολουθία δεν είναι ίδια με την ακολουθία ATCGATCGATCG, δεν είναι ωστόσο και ουσιαστικά διαφορετική. Όταν επιτρέπεται μετάλλαξη στο εμφυτευμένο μοτίβο P, η ανασύνθεση του P γίνεται πιο περίπλοκη, δεδομένου ότι το 8-μερές δεν αποκαλύπτει το μοτίβο. Συγκεκριμένα, η συμβολοσειρά TCGACTAC δεν φαίνεται στην εικόνα 15.δ αλλά οι επτά μεταλλαγμένες εκδόσεις της εμφανίζονται στη θέση 8 στην πρώτη αλληλουχία, στη θέση 19 στη δεύτερη ακολουθία, 3 στην τρίτη, 5 στην τέταρτη, 31 στην πέμπτη, 27 στην έκτη, και 15 στην έβδομη.



***Εικόνα 16** : NF-κB μοτίβα*

3.2.2 Διατύπωση προβλήματος εύρεσης μοτίβων

Προκειμένου να διατυπώσουμε το πρόβλημα εύρεσης μοτίβων, χρειάζεται να προσδιορίσουμε με ακρίβεια τι εννοούμε με τον όρο «μοτίβο». Βασιζόμενοι σε μία μόνο συμβολοσειρά που αντιπροσωπεύει ένα μοτίβο συχνά αποτυγχάνουμε να εκπροσωπήσουμε την παραλλαγή του προτύπου σε πραγματικές βιολογικές ακολουθίες (όπως στην εικόνα 15.δ). Μια πιο ευέλικτη αναπαράσταση ενός μοτίβου χρησιμοποιεί έναν πίνακα προφίλ.
Υποθέτουμε ένα σύνολο από t αλληλουχίες DNA, καθεμία από τις οποίες έχει n νουκλεοτίδια και στη συνέχεια επιλέγουμε μία θέση σε καθεμία από αυτές τις t αλληλουχίες, σχηματίζοντας έτσι έναν πίνακα $s = (s_1, s_2, \ldots, s_t)$, με $1 <= s_i <= n - l + 1$. Τα l-μερή που ξεκινούν από αυτές τις θέσεις μπορούν να τοποθετούνται σε έναν $t \times l$ πίνακα στοίχησης ακολουθιών, του οποίου το (i, j) στοιχείο είναι το νουκλεοτίδιο στο $s_i + j - 1$ στοιχείο στην i-οστή ακολουθία (εικόνα 17).
Με βάση αυτόν τον πίνακα στοίχησης, μπορούμε να υπολογίσουμε τον $4 \times l$ πίνακα προφίλ του οποίου το (i, j) στοιχείο αποτελεί τον αριθμό των φορών που το νουκλεοτίδιο i εμφανίζεται στη στήλη j του πίνακα στοίχησης, όπου το i κυμαίνεται από 1 έως 4. Ο πίνακας προφίλ ή απλά προφίλ, δείχνει την μεταβλητότητα της νουκλεοτιδικής σύνθεσης σε κάθε θέση για μια συγκεκριμένη επιλογή ενός l-μερούς. Για παράδειγμα, οι θέσεις 3, 7, και 8 διατηρούνται, ενώ η θέση 4 όχι. Μπορούμε ακόμα να σχηματίσουμε μια «ομόφωνη»

συμβολοσειρά (consensus string) από τα πιο δημοφιλή στοιχεία σε κάθε στήλη του πίνακα στοίχησης. Με τον όρο πιο δημοφιλές εννοούμε το στοιχείο με τις περισσότερες εμφανίσεις (δηλαδή με τον μεγαλύτερο αριθμό στον πίνακα προφίλ). Η εικόνα 17 δείχνει τον πίνακα στοίχησης για s = (8, 19, 3, 5, 31, 27, 15) καθώς και το αποτέλεσμα της «ομόφωνης» συμβολοσειράς που είναι η ATGCAACT.
Μεταβάλλοντας τις αρχικές θέσεις στο s, μπορούμε να κατασκευάσουμε ένα μεγάλο αριθμό διαφορετικών πινάκων προφίλ για δεδομένο δείγμα. Χρειαζόμαστε, λοιπόν, κάποιον τρόπο της αξιολόγησης αυτών των προφίλ καθώς μερικά προφίλ αντιπροσωπεύουν υψηλή διατήρηση ενός μοτίβου ενώ άλλα όχι. Μια ασαφής διατύπωση του προβλήματος εύρεσης μοτίβων είναι να βρεθούν οι αρχικές θέσεις s που αντιστοιχούν στο προφίλ που διατηρεί περισσότερο το μοτίβο.

(α)Εισαγωγή των επτά 8-μερών της εικόνας 14.δ

Alignment

	T	A	G	A	C	G	A	C
	T	C	A	C	C	T	A	C
	T	C	G	A	A	T	A	A
	T	G	G	A	C	C	A	C
	T	T	G	A	C	T	A	A
	C	G	G	A	C	T	A	C
	G	C	G	T	C	T	A	C

Profile

A	0	1	1	5	1	0	7	2
T	5	1	0	1	0	5	0	0
C	1	3	0	1	6	1	0	5
G	1	2	6	0	0	1	0	0

Consensus T C G A C T A C

(β) Ο πίνακας στοίχησης, ο πίνακας προφίλ και η "ομόφωνη" συμβολοσειρά που σχηματίζονται από τα 8-μερή, τα οποία ξεκινούν από τις θέσεις s=(8, 19, 3, 5, 31, 27, 15) της εικόνας 15.δ

***Εικόνα 17** : Από το δείγμα DNA στον πίνακα στοίχησης, το προφίλ και τέλος, την «ομόφωνη» συμβολοσειρά. Αν s = (8, 19, 3, 5, 31, 27, 15) είναι ο πίνακας των αρχικών θέσεων των 8-μερών της εικόνας 15.δ, τότε Score (s) = 5 + 3 + 6 + 5 + 6 + 5 + 7 + 5 = 42*

Πρόβλημα Εύρεσης Μοτίβων (The Motif Finding Problem):
Δοσμένου ενός συνόλου αλληλουχιών DNA, βρίσκουμε ένα σύνολο l-μερών, ένα για κάθε αλληλουχία, τέτοιο ώστε να μεγιστοποιείται στο σκορ της «ομόφωνης» συμβολοσειράς (consensus string).

Είσοδος (Input): Ένας πίνακας DNA αλληλουχιών t × n και το μήκος l του μοτίβου

Έξοδος (Output): Τον πίνακα s με τις t θέσεις έναρξης του μοτίβου σε κάθε αλληλουχία s = (s1, s2, . . . , st) που μεγιστοποιεί το σκορ της «ομόφωνης» συμβολοσειρά (consensus string).

Μια άλλη οπτική αυτού του προβλήματος είναι να θεωρήσουμε το πρόβλημα εύρεσης μοτίβων ως ένα πρόβλημα εύρεσης μέσης συμβολοσειράς (median string). Δοσμένων δύο l-μερών v και w, μπορούμε να υπολογίσουμε τον αριθμό των θέσεων που οι δύο συμβολοσειρές διαφέρουν χρησιμοποιώντας την απόσταση Hamming (Hamming distance) d_H (v, w) μεταξύ τους.
Για παράδειγμα, d_H (TAACGT, ACAGGT) = 2:

```
T A A C G T
: X : X : :
A C A G G T
```

Τώρα, ας υποθέσουμε ότι s = (s_1, s_2, . . . , s_t) είναι ένας πίνακας με τις θέσεις έναρξης, και ότι είναι v είναι κάποιο l-μερές. Η συνολική απόσταση Hamming (total Hamming distance) μεταξύ του v και των l-μερών που ξεκινούν από τις θέσεις του πίνακα s είναι:

$$d_H(v, \mathbf{s}) = \sum_{i=1}^{t} d_H(v, s_i),$$

όπου d_H (v,s_i) είναι η απόσταση Hamming μεταξύ του v και του l-μερούς που ξεκινά από τη θέση s_i της i-οστής ακολουθίας DNA.
Χρησιμοποιούμε την ελάχιστη συνολική απόσταση Hamming της συμβολοσειράς από οποιοδήποτε σύνολο s που περιλαμβάνει τις θέσεις έναρξης στις αλληλουχίες του DNA. Η απόσταση αυτή δίνεται από τη σχέση:

$$\text{TotalDistance}(v, \text{DNA}) = \min_s (d_H(v, s))$$

Η εύρεση του TotalDistance είναι ένα απλό πρόβλημα. Πρώτα πρέπει να βρούμε το καλύτερο ταίριασμα του v με την πρώτη αλληλουχία του DNA (δηλαδή, μια θέση που να ελαχιστοποιεί το d_H (v, s_1) όπου 1 <= s_1 <= n − l + 1), έπειτα με τη δεύτερη και ούτω καθεξής. Έτσι, παίρνουμε την ελάχιστη από όλες τις πιθανές θέσεις έναρξης. Τέλος, ορίζουμε τη μεσαία συμβολοσειρά (median string) ως τη συμβολοσειρά v που ελαχιστοποιεί το TotalDistance (v, DNA).

Μπορούμε, λοιπόν, τώρα να διατυπώνουμε το πρόβλημα της εύρεσης μέσης συμβολοσειράς DNA ακολουθιών ως εξής:

Πρόβλημα μέσης συμβολοσειράς (Median String Problem)
Δοσμένου ενός συνόλου DNA αλληλουχιών βρίσκουμε τη μέση συμβολοσειρά

Είσοδος (Input): Ένας πίνακας DNA αλληλουχιών t × n και το μήκος l του μοτίβου

Έξοδος (Output): Η συμβολοσειρά v, l νουκλεοτιδίων που ελαχιστοποιεί το TotalDistance(v,DNA)

Στο σημείο αυτό πρέπει να τονίσουμε ότι έχουμε διπλή ελαχιστοποίηση. Βρίσκουμε, δηλαδή, μια συμβολοσειρά v που ελαχιστοποιεί την TotalDistance (v, DNA), που με τη σειρά της είναι η μικρότερη απόσταση μεταξύ όλων των επιλογών των σημείων εκκίνησης s στις ακολουθίες DNA.
Παρά το γεγονός ότι το πρόβλημα εύρεσης μέσης συμβολοσειράς (Median String Problem) είναι πρόβλημα ελαχιστοποίησης ενώ το πρόβλημα εύρεσης μοτίβων είναι πρόβλημα μεγιστοποίησης, τα δύο προβλήματα είναι υπολογιστικά ισοδύναμα.
Αν υποθέσουμε s το σύνολο με τις θέσεις έναρξης με σκορ «ομόφωνης» συμβολοσειράς ίσο με Score(s, DNA) και w την «ομόφωνη» συμβολοσειρά του αντίστοιχου προφίλ έχουμε:

$$d_H(w, s) = l * t - \text{Score}(s, DNA).$$

T **A** G A C **G** A C
T C **A** **C** C T A C
T C G A **A** T A **A**
T **G** G A C **C** A C
T **T** G A C T A **A**
C **G** G A C T A C
G C G **T** C T A C

Εικόνα 18 *: Υπολογισμός της συνολικής απόστασης Hamming*

Στην εικόνα 18, η απόσταση Hamming μεταξύ της συμβολοσειράς w και καθένα από τα επτά εμφυτευμένα μοτίβα είναι 2. Αυτό γιατί:

$$d_H(w, s) = 7 \cdot 8 - 42 = 14$$

ή αλλιώς

2 * 7 (2 τα νουκλεοτίδια που διαφέρουν οι αλληλουχίες DNA από την w και 7 οι αλληλουχίες DNA).

Όσο τα t και l είναι σταθερά, η μικρότερη τιμή του d_H μπορεί να επιτευχθεί μεγιστοποιώντας το Score (s, DNA), όλων των επιλογών του s:

$$\min_{\text{όλες τις επιλογές του s}} \min_{\text{όλες τις επιλογές του v}} dH(v, s) = l * t - \max_{\text{όλες τις επιλογές του s}} Score(s, DNA).$$

Το πρόβλημα στα αριστερά είναι το πρόβλημα εύρεσης μέσης συμβολοσειράς ενώ το πρόβλημα στα δεξιά είναι το πρόβλημα εύρεσης μοτίβων.
Με άλλα λόγια, δηλαδή, η «ομόφωνη» συμβολοσειρά (consensus), η λύση του προβλήματος εύρεσης μοτίβων αποτελεί τη μέση συμβολοσειρά (median string). Η τελευταία μπορεί να χρησιμοποιηθεί για να δημιουργήσει ένα προφίλ που να λύνει το πρόβλημα εύρεσης μοτίβων, αναζητώντας σε καθεμία από τις t ακολουθίες τη συμβολοσειρά με τη μικρότερη απόσταση Hamming από αυτή.

3.2.3 Δέντρα αναζήτησης

Τόσο στο πρόβλημα εύρεσης μοτίβων όσο και σε αυτό της εύρεσης μέσης συμβολοσειράς προκειμένου να βρούμε την καλύτερη λύση πρέπει να ελέγξουμε έναν μεγάλο αριθμό εναλλακτικών λύσεων. Για παράδειγμα, στο πρόβλημα εύρεσης μοτίβων πρέπει να λάβουμε υπόψη μας όλες τις $(n - l + 1)^t$ θέσεις έναρξης s (εικόνα 19) και στο πρόβλημα εύρεσης μέσης συμβολοσειράς όλα τα 4^l πιθανά l-μερή (εικόνα 20)

$$
\begin{array}{cccccc}
(& 1, & 1, & \ldots, & 1, & 1 &) \\
(& 1, & 1, & \ldots, & 1, & 2 &) \\
(& 1, & 1, & \ldots, & 1, & 3 &) \\
 & & & \vdots & & & \\
(& 1, & 1, & \ldots, & 1, & n-l+1 &) \\
(& 1, & 1, & \ldots, & 2, & 1 &) \\
(& 1, & 1, & \ldots, & 2, & 2 &) \\
(& 1, & 1, & \ldots, & 2, & 3 &) \\
 & & & \vdots & & & \\
(& 1, & 1, & \ldots, & 2, & n-l+1 &) \\
 & & & \vdots & & & \\
(& n-l+1, & n-l+1, & \ldots, & n-l+1, & 1 &) \\
(& n-l+1, & n-l+1, & \ldots, & n-l+1, & 2 &) \\
(& n-l+1, & n-l+1, & \ldots, & n-l+1, & 3 &) \\
 & & & \vdots & & & \\
(& n-l+1, & n-l+1, & \ldots, & n-l+1, & n-l+1 &)
\end{array}
$$

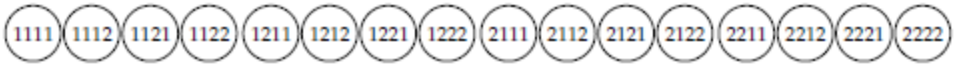
***Εικόνα 19** : Όλα τα πιθανά s στο πρόβλημα εύρεσης μοτίβων*

Πηγή: An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner

AA··· AA
AA··· AT
AA··· AG
AA··· AC
AA··· TA
AA··· TT
AA··· TG
AA··· TC
⋮
CC··· GG
CC··· GC
CC··· CA
CC··· CT
CC··· CG
CC··· CC

***Εικόνα 20** : Τα πιθανά l-μερή στο πρόβλημα εύρεσης μέσης συμβολοσειράς*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*



***Εικόνα 21** : Όλα τα 4-μερή με αλφάβητο {1, 2}*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η αναπαράσταση των 4-μερών της εικόνας 20 είναι ισοδύναμη με αυτή της εικόνας 21. Ας θεωρήσουμε, λοιπόν, ότι χρησιμοποιούμε το 1 στη θέση της αδενίνης (A), το 2 στη θέση της θυμίνης (T), το 3 στη θέση της γουανίνης (G) και 4 στη θέση της κυτοσίνης (C):

$$
\begin{array}{c}
(1,1,\ldots,1,1)\\
(1,1,\ldots,1,2)\\
(1,1,\ldots,1,3)\\
(1,1,\ldots,1,4)\\
(1,1,\ldots,2,1)\\
(1,1,\ldots,2,2)\\
(1,1,\ldots,2,3)\\
(1,1,\ldots,2,4)\\
\vdots\\
(4,4,\ldots,3,3)\\
(4,4,\ldots,3,4)\\
(4,4,\ldots,4,1)\\
(4,4,\ldots,4,2)\\
(4,4,\ldots,4,3)\\
(4,4,\ldots,4,4)
\end{array}
$$

***Εικόνα 22** : Όλα τα 4-μερή με αλφάβητο {1, 2, 3, 4}*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Αυτό που θέλουμε, λοιπόν, είναι να εξετάσουμε όλα τα $κ^L$ L-μερή σε ένα αλφάβητο κ γραμμάτων. Για το πρόβλημα εύρεσης μοτίβων έχουμε κ= n − l + 1 , ενώ για αυτό της εύρεσης μέσης συμβολοσειράς έχουμε κ=4.
Η παρακάτω συνάρτηση NEXTLEAF δείχνει πως μπορούμε να μεταβούμε από το ένα L-μερές a = (a_1 a_2 · · · a_L) στο επόμενο.

```
NEXTLEAF(a, L, k)
1  for i ← L to 1
2      if a_i < k
3          a_i ← a_i + 1
4          return a
5      a_i ← 1
6  return a
```

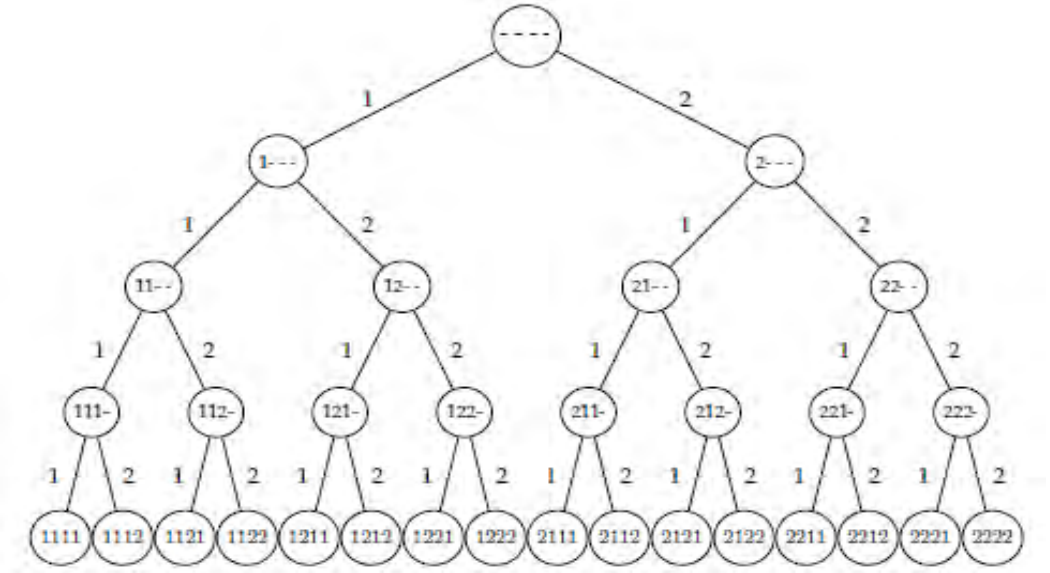
***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η συνάρτηση NEXTLEAF λειτουργεί με έναν παρόμοιο τρόπο με τη φυσική διαδικασία μέτρησης, δηλαδή, το $(a_1\ a_2 \cdot \cdot \cdot a_L)$ ακολουθείται από το $(a_1\ a_2 \cdot \cdot \cdot a_L + 1)$. Όταν $a_L = k$, η επόμενη επίκληση της NEXTLEAF επαναφέρει το a_L σε 1 και προσθέτει 1 στο a_{L-1} όπως ακριβώς συμβαίνει, καθώς μετράμε, στη μετάβαση από το 1019 στο 1020. Αν όμως υπάρχει μια ολόκληρη συμβολοσειρά με τιμή κ στη δεξιά πλευρά, ο αλγόριθμος πρέπει να επαναφέρει όλα τα στοιχεία σε 1, όπως και στη μετάβαση από το 139999 στο 140000.
Ο αλγόριθμος τερματίζει, ολοκληρώνει δηλαδή την εξέταση όλων των L-μερών, όταν επαναφέρει όλα τα στοιχεία k και τους δώσει τις τιμές (1, 1,..., 1).
Στην περίπτωση που η L =10, η συνάρτηση NEXTLEAF λειτουργεί όπως ακριβώς το μέτρημα των δεκαδικών αριθμών με τη διαφορά ότι χρησιμοποιεί «ψηφία» από το 1 έως το 10, και όχι από το 0 έως το 9.
Ο παρακάτω αλγόριθμος ALLLEAVES, χρησιμοποιεί την NEXTLEAF για την έξοδο όλων των 4-μερών με τη σειρά που φαίνονται στην εικόνα 21.

```
ALLLEAVES(L, k)
1  a ← (1, ..., 1)
2  while forever
3      output a
4      a ← NEXTLEAF(a, L, k)
5      if a = (1, 1, ..., 1)
6          return
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Οι επιστήμονες υπολογιστών αναπαριστούν όλα τα L-μερή, ως φύλλα ενός δέντρου, όπως ακριβώς φαίνεται στην εικόνα 23. L- μερή δέντρα έχουν L επίπεδα (με εξαίρεση το ανώτερο επίπεδο - ρίζα) και κάθε φύλλο έχει κ παιδιά.



Εικόνα 23 *: Όλα τα 4-μερή με αλφάβητο {1, 2} μπορούν να αναπαρασταθούν ως φύλλα ενός δέντρου*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Για να αναπαραστήσουμε όλες τις πιθανές θέσεις έναρξης στο πρόβλημα εύρεσης μοτίβων, μπορούμε να κατασκευάσουμε το δέντρο με L = t επίπεδα και k = n - l + 1 παιδιά ανά κορυφή.
Για το πρόβλημα εύρεσης μέσης συμβολοσειράς έχουμε L = l και k = 4. Η καταγραφή των φύλλων ενός δέντρου είναι απλή, αλλά η καταγραφή όλων των κόμβων του (δηλαδή, όλων των φύλλων και όλων των εσωτερικών κόμβων) είναι πιο περίπλοκη. Ξεκινάμε από επίπεδο 0 (τη ρίζα) και εξετάζουμε καθένα από τα κ παιδιά του με σειρά. Για κάθε παιδί, εξετάζουμε πάλι καθένα από τα κ παιδιά του και ούτω καθεξής.
Με τον αλγόριθμο NEXTVERTEX που ακολουθεί διασχίζουμε επαναληπτικά ολόκληρο το δέντρο. Ο αλγόριθμος παίρνει ως είσοδο την κορυφή a = (a_1, ... , a_L) του επιπέδου i και επιστρέφει την επόμενη κορυφή του δέντρου. Στην πραγματικότητα, σε επίπεδο i ο NEXTVERTEX χρησιμοποιεί μόνο τις τιμές a_1, ... , a_i και αγνοεί τις a_1, ... , a_L. Ο NEXTVERTEX παίρνει εισόδους παρόμοιες με τον NEXTLEAF, με την διαφορά ότι το «τρέχον φύλλο» είναι πλέον η «τρέχουσα κορυφή». Χρησιμοποιεί, λοιπόν, την παράμετρο i για το επίπεδο στο οποίο βρίσκεται η κορυφή. Δεδομένων των, L, i, και k, ο NEXTVERTEX επιστρέφει τον επόμενο κόμβο του δέντρου ως ζεύγος πίνακα και επιπέδου. Ο αλγόριθμος θα ολοκληρώσει την αναζήτηση και θα τερματίσει όταν επιστρέψει τον αριθμό επιπέδου 0.

```
NEXTVERTEX(a, i, L, k)
1  if i < L
2      a_{i+1} ← 1
3      return (a, i + 1)
4  else
5      for j ← L to 1
6          if a_j < k
7              a_j ← a_j + 1
8              return (a, j)
9  return (a, 0)
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Όμως, ένα δέντρο έχει και αδιάφορους κόμβους. Στην εικόνα 23 δίπλα από κάθε φύλλο έχουμε σημειώσει το σκορ των L-μερών. Οι βαθμολογίες στους εσωτερικούς κόμβους αντιπροσωπεύουν τη μέγιστη βαθμολογία των υποδέντρων εκείνου του κόμβου. Προκειμένου να βελτιώσουμε τους αλγορίθμους διεξοδικής αναζήτησης (brute-force) αγνοούμε τα υποδέντρα δεν περιέχουν φύλλα υψηλής βαθμολογίας. Για παράδειγμα, αφού το σκορ του πρώτου φύλλου είναι 24, δεν έχει νόημα να αναλυθούν το 4ο, 5ο, ή 6ο φύλλο των οποίων η βαθμολογία είναι 20, 4, και 5, αντίστοιχα. Επομένως, το υποδέντρο που περιέχει αυτούς τους κόμβους μπορεί να αγνοηθεί. Η προσέγγιση αυτή (branch and bound) υλοποιείται με τον αλγόριθμο BYPASS που ακολουθεί.

```
BYPASS(a, i, L, k)
1  for j ← i to 1
2      if a_j < k
3          a_j ← a_j + 1
4          return (a, j)
5  return (a, 0)
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

3.2.4 Αλγόριθμοι για το πρόβλημα εύρεσης μοτίβων

Η επίλυση του προβλήματος εύρεσης μοτίβων με διεξοδική αναζήτηση αναζητά όλες τις πιθανές θέσεις έναρξης. Υπάρχουν n − l + 1 επιλογές για το πρώτο δείκτη (s_1) και για καθεμία από αυτές, υπάρχουν n − l + 1 επιλογές για το δεύτερο δείκτη (s_2). Για κάθε μία από αυτές τις επιλογές, υπάρχουν n - l + 1 επιλογές για το τρίτο δείκτη και ούτω καθεξής. Επομένως, ο συνολικός αριθμός των θέσεων είναι $(n - l + 1)^t$, όπου t ο αριθμός των ακολουθιών. Για κάθε s, ο αλγόριθμος υπολογίζει το Score (s, DNA), το οποίο απαιτεί O (l) πράξεις. Έτσι, η συνολική πολυπλοκότητα του αλγορίθμου υπολογίζεται σε O (ln^t).

```
BRUTEFORCEMOTIFSEARCH(DNA, t, n, l)
1  bestScore ← 0
2  for each (s_1, ..., s_t) from (1, ..., 1) to (n − l + 1, ..., n − l + 1)
3      if Score(s, DNA) > bestScore
4          bestScore ← Score(s, DNA)
5          bestMotif ← (s_1, s_2, ..., s_t)
6  return bestMotif
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Απομένει, όμως, το ερώτημα του πως θα γράψουμε τη γραμμή 2 ψευδοκώδικα. Αυτό, είναι ιδιαίτερα εύκολο αν χρησιμοποιήσουμε τη συνάρτηση NEXTLEAF που αναλύσαμε παραπάνω Σε αυτή την περίπτωση, L=n-l+1 και k = t. Ξαναγράφοντας τον BRUTEFORCEMOTIFSEARCH με αυτόν τον τρόπο καταλήγουμε στον BRUTEFORCEMOTIFSEARCHAGAIN.

```
BRUTEFORCEMOTIFSEARCHAGAIN(DNA, t, n, l)
1  s ← (1, 1, ..., 1)
2  bestScore ← Score(s, DNA)
3  while forever
4      s ← NEXTLEAF(s, t, n − l + 1)
5      if Score(s, DNA) > bestScore
6          bestScore ← Score(s, DNA)
7          bestMotif ← (s_1, s_2, ..., s_t)
8      if s = (1, 1, ..., 1)
9          return bestMotif
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Τέλος, για την προσέγγιση branch and bound χρειαζόμαστε τον αντίστοιχο αλγόριθμο SIMPLEMOTIFSEARCH, ο οποίος χρησιμοποιεί τη συνάρτηση NEXTVERTEX για να εξερευνήσει κάθε φύλλο.

```
SIMPLEMOTIFSEARCH(DNA, t, n, l)
 1  s ← (1, ..., 1)
 2  bestScore ← 0
 3  i ← 1
 4  while i > 0
 5      if i < t
 6          (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
 7      else
 8          if Score(s, DNA) > bestScore
 9              bestScore ← Score(s, DNA)
10              bestMotif ← (s_1, s_2, ..., s_t)
11          (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
12  return bestMotif
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Παρατηρούμε ότι ορισμένα σύνολα θέσεων έναρξης μπορούν να αποκλειστούν αμέσως με βάση την εκτίμηση της βαθμολογίας τους.
Δοσμένου ενός συνόλου θέσεων έναρξης s = (s_1, s_2, . . . , s_t), ορίζουμε το partial consensus Score (Score (s, i, DNA)), ως το σκορ του i × l πίνακα στοίχησης που αφορά μόνο τις πρώτες γραμμές i του DNA που αντιστοιχούν στις θέσεις έναρξης (s_1, s_2, . . . s_i, - , -, . . . , -). Αν έχουμε το partial consensus Score s_1, s_2, . . . s_i , ακόμη και στην καλύτερη περίπτωση οι υπόλοιπες t - i σειρές μπορούν μόνο να βελτιώσουν το σκορ κατά (t - i) · l. Επομένως, το σκορ του κάθε πίνακα στοίχησης με τις i πρώτες θέσεις έναρξης (s_1, s_2, . . . s_i) μπορεί να είναι το πολύ Score (s, i, DNA) + (t - i) · l. Αυτό σημαίνει ότι αν το Score (s, i, του DNA) + (t-i) · l είναι μικρότερο από το τρέχον καλύτερο σκορ (bestScore), τότε δεν έχει νόημα να διερευνήσουμε οποιαδήποτε από τις υπόλοιπες t - i ακολουθίες του δείγματος. Επομένως το δεσμευμένο σκορ (bound Score (s, i, DNA) + (t - i) · l) μας αποδεσμεύει από το ψάξιμο $(n - l + 1)^{t-i}$ φύλλων.
Παρόλο που η στρατηγική branch-and-bound βελτιώνει τον αλγόριθμο μας για αρκετές περιπτώσεις του προβλήματος, δεν έχουμε βελτιώσει την απόδοση της χειρότερης περίπτωσης καθώς μπορούμε να υλοποιήσουμε ένα δείγμα με ένα εμφυτευμένο μοτίβο που απαιτεί εκθετικό χρόνο για να βρεθεί.

```
BRANCHANDBOUNDMOTIFSEARCH(DNA, t, n, l)
 1  s ← (1, ..., 1)
 2  bestScore ← 0
 3  i ← 1
 4  while i > 0
 5      if i < t
 6          optimisticScore ← Score(s, i, DNA) + (t − i) · l
 7          if optimisticScore < bestScore
 8              (s, i) ← BYPASS(s, i, t, n − l + 1)
 9          else
10              (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
11      else
12          if Score(s, DNA) > bestScore
13              bestScore ← Score(s)
14              bestMotif ← (s_1, s_2, ..., s_t)
15          (s, i) ← NEXTVERTEX(s, i, t, n − l + 1)
16  return bestMotif
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

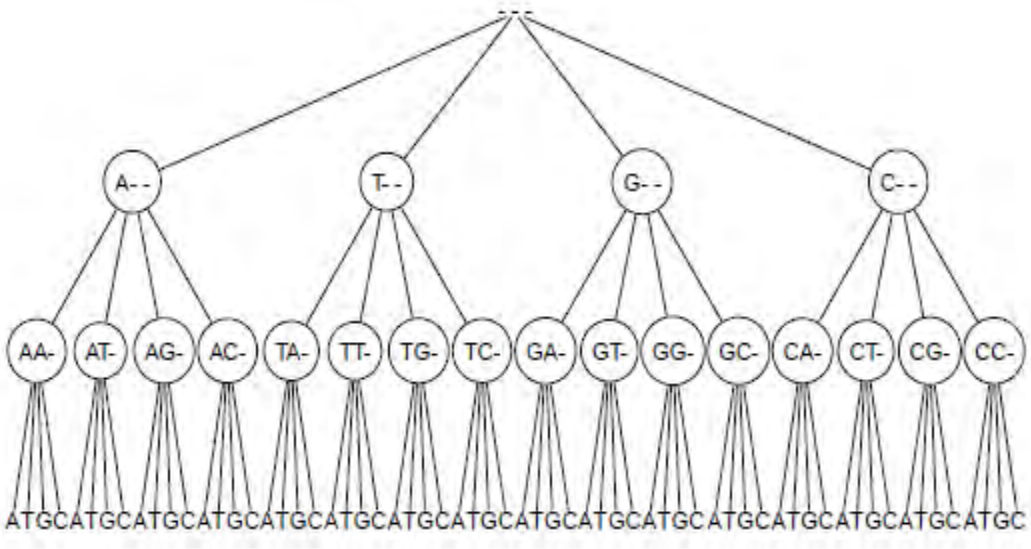
3.2.5 Αλγόριθμοι για το πρόβλημα εύρεσης μέσης συμβολοσειράς (Median String Problem)

Όπως ήδη έχουμε αναφέρει το πρόβλημα εύρεσης μέσης συμβολοσειράς (Median String Problem) αποτελεί μια εναλλακτική προσέγγιση της εύρεσης μοτίβων. Αν εφαρμόσουμε την τεχνική διεξοδικής αναζήτησης για την επίλυση αυτού του προβλήματος, καταλήγουμε στον αλγόριθμο BRUTEFORCEMEDIANSEARCH που ακολουθεί.

```
BRUTEFORCEMEDIANSEARCH(DNA, t, n, l)
1  bestWord ← AAA···AA
2  bestDistance ← ∞
3  for each l-mer word from AAA...A to TTT...T
4      if TOTALDISTANCE(word, DNA) < bestDistance
5          bestDistance ← TOTALDISTANCE(word, DNA)
6          bestWord ← word
7  return bestWord
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ο αλγόριθμος BRUTEFORCEMEDIANSEARCH λαμβάνει υπόψη καθεμία από τις 4^L συμβολοσειρές μήκους l νουκλεοτιδίων και υπολογίζει το TOTALDISTANCE σε κάθε βήμα. Δοσμένης μιας συμβολοσειράς (word), μπορούμε να υπολογίσουμε το TotalDistance (word, DNA) με ένα μόνο πέρασμα (δηλαδή, σε χρόνο O(nt)), και όχι εξετάζοντας όλα τα πιθανά σημεία έναρξης στο δείγμα DNA. Επομένως, ο αλγόριθμος BRUTEFORCEMEDIANSEARCH απαιτεί χρόνο εκτέλεσης O ($4^l \cdot n \cdot t$), ο οποίος είναι συγκριτικά καλύτερος από το O (ln^t) του SIMPLEMOTIFSEARCH. Ένα τυπικό μοτίβο έχει μήκος (l) που κυμαίνεται από 8 έως 15 νουκλεοτίδια, ενώ το τυπικό μέγεθος των περιοχών που αναλύονται έχουν μήκος (n) που κυμαίνεται από 500 έως 1000 νουκλεοτίδια. Ο αλγόριθμος BRUTEFORCEMEDIANSEARCH είναι ένα πρακτικός αλγόριθμος για την εύρεση μικρών μοτίβων ενώ ο SIMPLEMOTIFSEARCH δεν είναι.
Μπορούμε να μετατρέψουμε τον αλγόριθμο εύρεσης μέσης συμβολοσειράς έτσι ώστε να εξερευνούμε ολόκληρο το δέντρο L συμβολοσειρών (εικόνα 24) και όχι μόνο τα φύλλα του, όπως στον αλγόριθμο BRUTEFORCEMEDIANSEARCH.
Μια κορυφή στο επίπεδο i σε αυτό το δέντρο αντιπροσωπεύει μια νουκλεοτιδική συμβολοσειρά
μήκους i. Ο αλγόριθμος SIMPLEMEDIANSEARCH υποθέτει ότι τα νουκλεοτίδια A, C, G, T κωδικοποιούνται σαν αριθμοί (1, 2, 3, 4). Το l-μερές (1,1, ... 1) αντιστοιχεί στην νουκλεοτιδική αλληλουχία AA. . . . . A



Εικόνα 24 *: Δέντρο για την αναζήτηση μέσης συμβολοσειράς. Κάθε κόμβος μπορεί να έχει μόνο τέσσερα παιδιά, σε αντίθεση με τα n-l+l παιδιά στο πρόβλημα εύρεσης μοτίβων*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

```
SIMPLEMEDIANSEARCH(DNA, t, n, l)
 1  s ← (1, 1, ..., 1)
 2  bestDistance ← ∞
 3  i ← 1
 4  while i > 0
 5      if i < l
 6          (s, i) ← NEXTVERTEX(s, i, l, 4)
 7      else
 8          word ← nucleotide string corresponding to (s_1, s_2, ... s_l)
 9          if TOTALDISTANCE(word, DNA) < bestDistance
10              bestDistance ← TOTALDISTANCE(word, DNA)
11              bestWord ← word
12          (s, i) ← NEXTVERTEX(s, i, l, 4)
13  return bestWord
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Σύμφωνα με την στρατηγική branch-and-bound βρίσκουμε την TotalDistance (word, DNA), σε κάθε κορυφή. Εάν για μια κορυφή ι, αυτή η συνολική απόσταση είναι μεγαλύτερη από τη μικρότερη που έχουμε συναντήσει μέχρι εκείνη τη στιγμή για ένα από τα φύλλα τότε δεν υπάρχει κανένας λόγος διερεύνησης των υποδέντρων της κορυφής αυτής. Αυτό γιατί όλες οι επεκτάσεις αυτής της κορυφής σε ένα l-μερές θα έχουν τουλάχιστον την ίδια συνολική απόσταση και ίσως μεγαλύτερη.

```
BRANCHANDBOUNDMEDIANSEARCH(DNA, t, n, l)
 1  s ← (1, 1, ..., 1)
 2  bestDistance ← ∞
 3  i ← 1
 4  while i > 0
 5      if i < l
 6          prefix ← nucleotide string corresponding to (s_1, s_2, ..., s_i)
 7          optimisticDistance ← TOTALDISTANCE(prefix, DNA)
 8          if optimisticDistance > bestDistance
 9              (s, i) ← BYPASS(s, i, l, 4)
10          else
11              (s, i) ← NEXTVERTEX(s, i, l, 4)
12      else
13          word ← nucleotide string corresponding to (s_1, s_2, ... s_l)
14          if TOTALDISTANCE(word, DNA) < bestDistance
15              bestDistance ← TOTALDISTANCE(word, DNA)
16              bestWord ← word
17          (s, i) ← NEXTVERTEX(s, i, l, 4)
18  return bestWord
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Όπως συνήθως με τους αλγορίθμους που χρησιμοποιούν την τεχνική branch and bound, ο αλγόριθμος BRANCHANDBOUNDMEDIANSEARCH δεν παρέχει καμία βελτίωση στο χρόνο εκτέλεσης της χειρότερης περίπτωσης

4 Άπληστοι αλγόριθμοι

Οι άπληστοι αλγόριθμοι εφαρμόζονται συνήθως για την **επίλυση προβλημάτων βελτιστοποίησης**, όπως π.χ. η εύρεση του συντομότερου μονοπατιού μεταξύ δύο κορυφών γράφου ή η εύρεση της βέλτιστης σειράς για την εκτέλεση από έναν υπολογιστή ενός συνόλου έργων.
Ένας άπληστος αλγόριθμος διαθέτει γενικά μία απλή δομή που βασικά αποτελείται από τα εξής στοιχεία:

- ένα σύνολο υποψηφίων επιλογών (π.χ. οι κορυφές ενός γράφου)
- ένα σύνολο επιλογών που έχουν ήδη χρησιμοποιηθεί
- μία **συνάρτηση ελέγχου**, που απαντά στο ερώτημα αν ένα συγκεκριμένο σύνολο υποψηφίων αποδίδει μία λύση, όχι απαραίτητα τη βέλτιστη για τη στιγμή που εξετάζεται
- μία **συνάρτηση που ελέγχει αν ένα σύνολο υποψηφίων επιλογών είναι εφικτό**, με την έννοια ότι μπορεί αυτό να συμπληρωθεί με τέτοιο τρόπο, ώστε να μας δώσει μία λύση στο πρόβλημα
- μία **συνάρτηση επιλογής**, που ανά πάσα στιγμή δείχνει ποια επιλογή έχει την καλύτερη προοπτική για να είναι μέρος της λύσης του προβλήματος
- μία **αντικειμενική συνάρτηση**, που δίνει την τιμή της λύσης: είναι αυτή που επιθυμούμε να βελτιστοποιήσουμε

Ένας άπληστος αλγόριθμος προχωράει στο επόμενο βήμα με την απόφαση που εκείνη τη στιγμή φαίνεται να είναι η καλύτερη για την επίλυση του προβλήματος. Αυτό όμως δε σημαίνει ότι αποδίδει πάντα τη βέλτιστη λύση, αν και μερικές φορές τη βρίσκει.

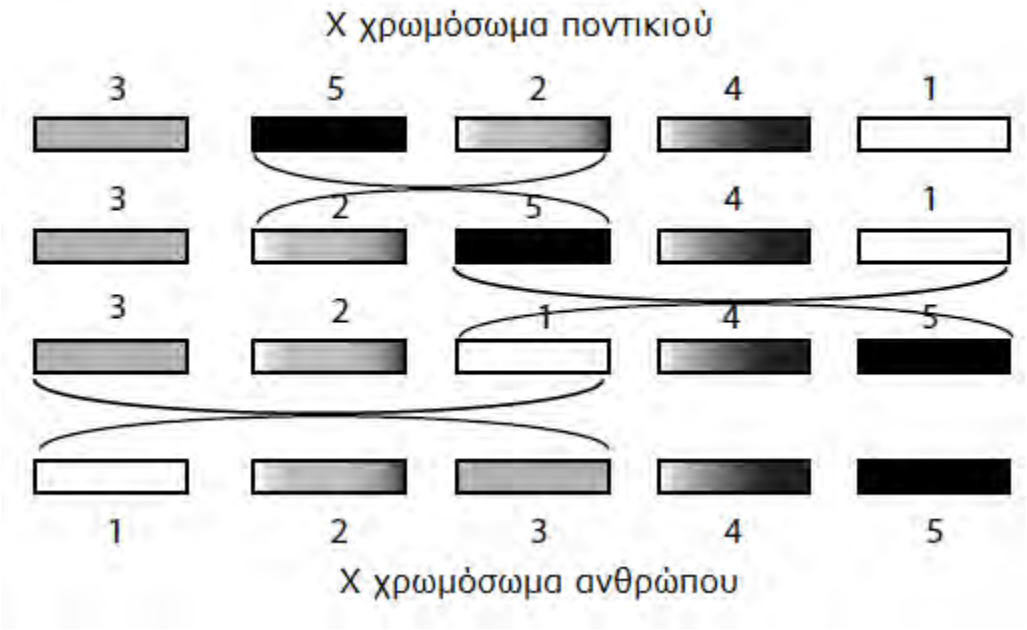
4.1 Ανακατατάξεις Γονιδιώματος

Το σύνδρομο Waardenburg είναι μια γενετική διαταραχή που οδηγεί σε απώλεια ακοής και χρωστικές ανωμαλίες, όπως δύο διαφορετικού χρώματος μάτια. Η ασθένεια πήρε το όνομά της από τον Ολλανδό οφθαλμίατρο που παρατήρησε πρώτος ότι οι άνθρωποι με δύο διαφορετικού χρώματος μάτια συχνά είχαν και προβλήματα ακοής.

***Εικόνα 25** : Βρέφος με σύνδρομο Waardenburg*

Πηγή:«Κλινικά και ακοολογικά ευρήματα στο σύνδρομο Waardenburg», Γ. Ψύλλας, Α. Ψηφίδης, Μ. Χίτογλου-Αντωνιάδου, Β. Νικολαΐδης, Α. Κουλούλας - τεύχος 25, Ιούλιος-Αύγουστος-Σεπτέμβριος 2006, σελίδες 33-36

Στις αρχές της δεκαετίας του 1990, οι βιολόγοι περιορίστηκαν στην αναζήτηση του γονιδίου του 2^{ου} χρωμοσώματος που ευθύνεται για το σύνδρομο Waardenburg, αλλά η ακριβής θέση του παρέμεινε άγνωστη για αρκετό χρονικό διάστημα. Παράλληλα, έλεγχαν ποντίκια για μεταλλάξεις και ένα από αυτά είχε χρωστικές ανωμαλίες όπως «μπαλώματα» λευκών κηλίδων, παρόμοια με εκείνα των ανθρώπων με σύνδρομο Waardenburg. Μέσω της αναπαραγωγής, το γονίδιο που ευθύνεται για τις κηλίδες αυτές χαρτογραφήθηκε σε ένα από τα χρωμοσώματα του ποντικιού. Όσο η χαρτογράφηση των γονιδίων προχωρούσε, κατέστη σαφές ότι υπάρχουν ομάδες γονιδίων σε ποντίκια που εμφανίζονται με την ίδια σειρά στον άνθρωπο. Κατά κάποιο τρόπο, δηλαδή, το ανθρώπινο γονιδίωμα είναι ακριβώς το γονιδίωμα του ποντικού κομμένο σε περίπου 300 μεγάλα γονιδιωματικά κομμάτια, που ονομάζονται synteny blocks τα οποία έχουν συνδεθεί με διαφορετική σειρά.
Το χρωμόσωμα 2 για παράδειγμα σε ανθρώπους είναι «χτισμένο» από θραύσματα παρόμοια με το DNA του ποντικιού που βρίσκονται στα χρωμοσώματα 1, 2, 3, 5, 6, 7, 10, 11, 12, 14, και 17. Γι αυτό το λόγο δεν αποτελεί έκπληξη, λοιπόν, ότι η ανακάλυψη ενός γονιδίου στα ποντίκια συχνά οδηγεί σε ενδείξεις σχετικά με τη θέση των σχετικών γονιδίων στον άνθρωπο.
Η παρακάτω εικόνα (εικόνα 26) παρουσιάζει ένα σενάριο αναδιοργάνωσης του χρωμοσώματος Χ του ποντικιού και μετατροπής του σε ανθρώπινο Χ χρωμόσωμα



***Εικόνα 26** : Παρουσίαση της μετατροπής της σειράς του γονιδίου του ποντικιού σε αυτή του ανθρώπινου γονιδίου για το χρωμόσωμα Χ (εδώ φαίνονται μόνο τα πέντε μεγαλύτερα synteny blocks).*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Οι βιολόγοι ενδιαφέρονται για το πιο φειδωλό εξελικτικό σενάριο, δηλαδή το σενάριο που περιλαμβάνει το μικρότερο αριθμό ανακατατάξεων. Παρόλο που δεν υπάρχει καμία εγγύηση ότι το παραπάνω σενάριο αποτελεί μια πραγματική εξελικτική ακολουθία, μας δίνει το κάτω όριο του αριθμού των ανακατατάξεων που έχουν συμβεί και δείχνει την ομοιότητα των δύο ειδών.
Ακόμη και για τον μικρό αριθμό των synteny blocks που χρησιμοποιούμε στην εικόνα 26, δεν είναι τόσο εύκολο να εξακριβωθεί ότι τα τρία εξελικτικά γεγονότα που παρουσιάζονται αντιπροσωπεύουν τη συντομότερη σειρά των αναστροφών για τη μετατροπή του X χρωμοσώματος του ποντικιού στο αντίστοιχο του ανθρώπου. Η τεχνική της διεξοδικής αναζήτησης, που παρουσιάσαμε στο προηγούμενο κεφάλαιο, δύσκολα θα μπορούσε να ασχοληθεί με τις ανακατατάξεις γονιδιώματος, καθώς ο αριθμός των παραλλαγών που πρέπει να διερευνηθούν γίνεται τεράστιος για περισσότερα από δέκα synteny blocks. Παρακάτω, θα αναλύσουμε δύο άπληστες προσεγγίσεις που έχουν διαφορετικό βαθμό επιτυχίας.

4.1.1 Ταξινόμηση με ανατροπές

Στην απλούστερη μορφή τους, τα γεγονότα αναδιάταξης μπορούν να μοντελοποιηθούν από μια σειρά ανακατατάξεων που μετατρέπουν ένα γονιδίωμα σε ένα άλλο. Η σειρά των γονιδίων (η καλύτερα των synteny blocks) σε ένα γονιδίωμα μπορεί να αναπαρασταθεί με μετάθεση της θέσης των στοιχείων του π, όπου $\pi = \pi_1\pi_2 \ldots \pi_n$
Για παράδειγμα η σειρά των synteny blocks στο X χρωμόσωμα του ανθρώπου αναπαρίσταται στην εικόνα 26 από το σύνολο (1, 2, 3, 4, 5), ενώ η αντίστοιχη σειρά στο ποντίκι από το (3, 5, 2, 4, 1) .
Μια ανακατάταξη ρ(i, j) έχει ως αποτέλεσμα την ανακατάταξη της σειράς των synteny blocks

$$\pi_i\pi_{i+1} \ldots \pi_{j-1}\pi_j$$

Στην πραγματικότητα, αυτό μετατρέπει το

$$\pi = \pi_1 \ldots \pi_{i-1}\ \pi_i\pi_{i+1} \ldots \pi_{j-1}\pi_j\pi_{j+1} \ldots \pi_n$$

στο

$$\pi * \rho(i, j) = \pi_1 \ldots \pi_{i-1}\ \pi_j\pi_{j-1} \ldots \pi_{i+1}\pi_i\pi_{j+1} \ldots \pi_n$$

Για παράδειγμα, εάν π= 1 2 4 3 7 5 6, τότε π * ρ (3, 6) = 1 2 **5 7 3 4** 6. Έτσι, είμαστε σε θέση να διατυπώσουμε το υπολογιστικό πρόβλημα που μιμείται τη βιολογική διαδικασία αναδιάταξης.

Πρόβλημα απόστασης αναστροφής (Reversal Distance Problem)
Δοσμένων δύο παραλλαγών, βρίσκουμε τη μικρότερη σειρά αναστροφών που μετατρέπουν τον ένα συνδυασμό στον άλλο.

Είσοδος (Input): Παραλλαγές π και σ.
Έξοδος (Output): Τη σειρά των λιγότερων αναστροφών $\rho_1, \rho_2, \ldots, \rho_t$ που μετατρέπουν τον συνδυασμό π στον σ (δηλαδή $\pi * \rho_1 * \rho_2 \ldots * \rho_t = \sigma$)

Καλούμε t την απόσταση αναστροφής ανάμεσα σε π και σ την ορίζουμε ως d (π, σ).
Το πρόβλημα ταξινόμησης αναστροφών (Sorting by Reversals Problem) είναι παρόμοιο με το πρόβλημα απόστασης αναστροφής με τη διαφορά ότι απαιτεί ως είσοδο μόνο μια παραλλαγή
Σε αυτή την περίπτωση, καλούμε t την απόσταση αναστροφής του π και την ορίζουμε ως d(π). Όταν, για παράδειγμα, ταξινομούμε μια παραλλαγή π = 1 2 3 6 4 5, δεν έχει νόημα να μετακινήσουμε τα τρία πρώτα στοιχεία που είναι ήδη ταξινομημένα. Εάν ορίσουμε ως prefix(π) τον αριθμό των ήδη ταξινομημένων στοιχείων, η λογική στρατηγική της ταξινόμησης είναι να αυξάνουμε το prefix(π) σε κάθε βήμα. Η προσέγγιση αυτή ταξινομεί το π σε 2 βήματα:

1 2 3 6 4 5 → 1 2 3 4 6 5 → 1 2 3 4 5 6.

Έτσι, γενικεύοντας όλα τα παραπάνω οδηγούμαστε σε έναν αλγόριθμο που ταξινομεί μια παραλλαγή με επανειλημμένες κινήσεις έτσι ώστε το i-στο στοιχείο να βρίσκεται στην i-στη θέση

```
SIMPLEREVERSALSORT(π)
1  for i ← 1 to n − 1
2      j ← position of element i in π (i.e., π_j = i)
3      if j ≠ i
4          π ← π · ρ(i, j)
5          output π
6      if π is the identity permutation
7          return
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ο αλγόριθμος SIMPLEREVERSALSORT είναι ένα παράδειγμα άπληστου αλγόριθμου που επιλέγει την "καλύτερη" αναστροφή σε κάθε βήμα. Ωστόσο, η έννοια "καλύτερη" εδώ είναι μάλλον κοντόφθαλμη καθώς αυξάνοντας απλά το prefix(π) δεν έχουμε εγγυημένα τον μικρότερο αριθμό αναστροφών.
Για παράδειγμα, ο SIMPLEREVERSALSORT ταξινομεί το 8 1 3 5 6 7 σε πέντε βήματα.

$\underline{8\,1}\,3\,5\,6\,7 \rightarrow 1\,\underline{8\,3}\,5\,6\,7 \rightarrow 1\,3\,\underline{8\,5}\,6\,7 \rightarrow 1\,3\,5\,\underline{8\,6}\,7 \rightarrow 1\,3\,5\,6\,\underline{8\,7} \rightarrow 1\,3\,5\,6\,7\,8$

Όμως η ίδια παραλλαγή μπορεί να ταξινομηθεί με 2 βήματα:

$\underline{8\,1\,3\,5\,6\,7} \rightarrow \underline{7\,6\,5\,3\,1}\,8 \rightarrow 1\,3\,5\,6\,7\,6$

Επομένως μπορούμε να πούμε με βεβαιότητα ότι ο αλγόριθμος SIMPLEREVERSALSORT δεν είναι σωστός. Στην πραγματικότητα απαιτεί n-1 βήματα για να ταξινομήσει την παραλλαγή π= n 1 2. . . (n-1) παρόλο που d(π) = 2.;
Ένας ανάλογος αλγόριθμος με τον SIMPLEREVERSALSORT ταξινομεί κάθε παραλλαγή με το πολύ 2* (n-1) ανακατατάξεις. Για παράδειγμα, μπορούμε να ταξινομήσουμε το 1 2 3 6 4 5 με 4 αναστροφές

$\underline{1\,2\,3\,6}\,4\,5 \rightarrow \underline{6\,3\,2\,1\,4\,5} \rightarrow \underline{5\,4\,1\,2\,3}\,6 \rightarrow \underline{3\,2\,1}\,4\,5\,6 \rightarrow 1\,2\,3\,4\,5\,6$

αλλά δεν είναι σαφές το κατά πόσο υπάρχει μια ακόμα μικρότερη σειρά αναστροφών για την ταξινόμηση αυτή.
Ο William Gates, προπτυχιακός φοιτητής του Harvard στα μέσα της δεκαετίας του 1970 και ο Χρήστος Παπαδημητρίου, καθηγητής του Harvard στα μέσα της ίδιας δεκαετίας, έκαναν την πρώτη απόπειρα να λύσουν το πρόβλημα και απέδειξαν ότι κάθε παραλλαγή μπορεί να ταξινομηθεί με το πολύ 5/3 * (n + 1) αναστροφές.

4.2 Μια άπληστη προσέγγιση για την εύρεση μοτίβων

Στο προηγούμενο κεφάλαιο είδαμε έναν αλγόριθμο διεξοδικής αναζήτησης (brute force) για την επίλυση του προβλήματος εύρεσης μοτίβων. Όμως, λόγω του απογοητευτικού χρόνου λειτουργίας του O ($l \cdot n^t$), δεν μπορούμε να τον χρησιμοποιήσουμε στην πράξη σε βιολογικά δείγματα. Επιλέγουμε, λοιπόν, να βασιστούμε σε μια πιο γρήγορη άπληστη λογική, ακόμη και αν δεν είναι σωστή και δεν οδηγεί σε έναν αλγόριθμο με εγγύηση καλής εκτέλεσης. Ο αλγόριθμος αυτός είναι προσεγγιστικός με άγνωστο λόγο προσέγγισης. Με βάση την προσέγγιση αυτή, αναπτύχθηκε το 1989 ένα δημοφιλές εργαλείο (CONSENSUS), από τον Gary Stormo και τον Gerald Hertz, το οποίο παράγει αποτελέσματα εξίσου καλά ή και καλύτερα από ότι πιο περίπλοκοι αλγόριθμοι.
Ο αλγόριθμος GREEDYMOTIFSEARCH ελέγχει κάθε αλληλουχία DNA μόνο μία φορά. Μόλις ελέγξουμε μια συγκεκριμένη ακολουθία i, αποφασίζουμε ποιο L-μερές έχει την καλύτερη συμβολή στο σκορ στοίχησης (Score(s, I, DNA)) για τις πρώτες i ακολουθίες και αμέσως ισχυριζόμαστε ότι αυτό το I-μερές αποτελεί μέρος της ευθυγράμμισης.

```
GREEDYMOTIFSEARCH(DNA, t, n, l)
 1  bestMotif ← (1, 1, ..., 1)
 2  s ← (1, 1, ..., 1)
 3  for s_1 ← 1 to n − l + 1
 4      for s_2 ← 1 to n − l + 1
 5          if Score(s, 2, DNA) > Score(bestMotif, 2, DNA)
 6              BestMotif_1 ← s_1
 7              BestMotif_2 ← s_2
 8  s_1 ← BestMotif_1
 9  s_2 ← BestMotif_2
10  for i ← 3 to t
11      for s_i ← 1 to n − l + 1
12          if Score(s, i, DNA) > Score(bestMotif, i, DNA)
13              bestMotif_i ← s_i
14      s_i ← bestMotif_i
15  return bestMotif
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Αρχικά, ο GREEDYMOTIFSEARCH βρίσκει τα δύο πιο «κοντινά» l-μερή (με τη βοήθεια της απόστασης Hamming) στις ακολουθίες 1 και 2 και σχηματίζει έναν 2 × l πίνακα. Αυτό το στάδιο απαιτεί $l*(n - l + 1)^2$ πράξεις. Σε καθεμία από τις εναπομείναντες t - 2 επαναλήψεις ο GREEDYMOTIFSEARCH μεγαλώνει τον πίνακα, προσθέτοντάς του μια γραμμή, ελέγχοντας την i-στη αλληλουχία (όπου 3<= i <= t) και επιλέγοντας το l-μερές με το μέγιστο Score(s, i). Αυτό ισοδυναμεί περίπου σε $l*(n - l + 1)$ πράξεις σε κάθε επανάληψη. Επομένως, ο χρόνος εκτέλεσης αυτού του αλγορίθμου είναι $O(ln^2 + lnt)$, ο οποίος είναι κατά πολύ καλύτερος από τον $O(ln^t)$ του SIMPLEMOTIFSEARCH ή ακόμη και από τον $O(4^l nt)$ του BRUTEFORCEMEDIANSTRING.

Όταν t είναι μικρό σε σύγκριση με το n, ο GREEDYMOTIFSEARCH απαιτεί χρόνο $O(ln^2)$, και το μεγαλύτερο μέρος του χρόνου στην πραγματικότητα δαπανάται στον εντοπισμό των l-μερών από τις δύο πρώτες ακολουθίες.

Λόγω του γεγονότος ότι οι ακολουθίες εξετάζονται διαδοχικά, είναι δυνατό να κατασκευάσουμε περιπτώσεις εισόδου στις οποίες ο GREEDYMOTIFSEARCH δε θα βρει το βέλτιστο μοτίβο. Μια σημαντική διαφορά μεταξύ της δημοφιλούς CONSENSUS εύρεσης μοτίβων και του αλγόριθμου που παρουσιάσαμε είναι ότι η CONSENSUS μπορεί να ελέγξει τις ακολουθίες με μια τυχαία σειρά, καθιστώντας έτσι πιο δύσκολη την κατασκευή εισόδων που προκαλούν συμπεριφορά χειρότερης περίπτωσης. Μια άλλη σημαντική διαφορά είναι ότι η CONSENSUS αποθηκεύει ένα μεγάλο αριθμό (συνήθως τουλάχιστον 1000) πινάκων σε κάθε επανάληψη και όχι μόνο έναν που αποθηκεύει ο GREEDYMOTIFSEARCH, καθιστώντας έτσι λιγότερο πιθανό να χάσει τη βέλτιστη λύση. Ωστόσο, παρά αυτά τα πλεονεκτήματα δεν είναι εγγυημένο ότι αυτή η άπληστη προσέγγιση θα βρει ένα βέλτιστο μοτίβο.

5 Δυναμικός Προγραμματισμός

Ο Δυναμικός Προγραμματισμός είναι μια υπολογιστική μέθοδος, η οποία εφαρμόζεται όταν πρόκειται να ληφθεί μια σύνθετη απόφαση που προκύπτει από τη σύνθεση επιμέρους αποφάσεων που αλληλοεξαρτώνται. Η μέθοδος επίλυσης τέτοιων προβλημάτων βασίζεται στη διασύνδεση των επιμέρους αποφάσεων με κατάλληλη αναδρομική σχέση ώστε η σύνθεση των επιμέρους αποφάσεων να δίνει την τελικά ζητούμενη απόφαση. Το αρχικό πρόβλημα διασπάται σε επιμέρους υποπροβλήματα τα οποία συνδέονται με τη βοήθεια κατάλληλων αναδρομικών σχέσεων.
Ο Δυναμικός προγραμματισμός παρέχει αλγορίθμους για την κατανόηση και σύγκριση DNA αλληλουχιών, πολλοί από τους οποίους έχουν χρησιμοποιηθεί από τους βιολόγους για να βγάλουν σημαντικά συμπεράσματα για τη λειτουργία των γονιδίων και την εξελικτική ιστορία. Χρησιμοποιείται, επίσης για την εύρεση γονιδίων και άλλα προβλήματα βιοπληροφορικής

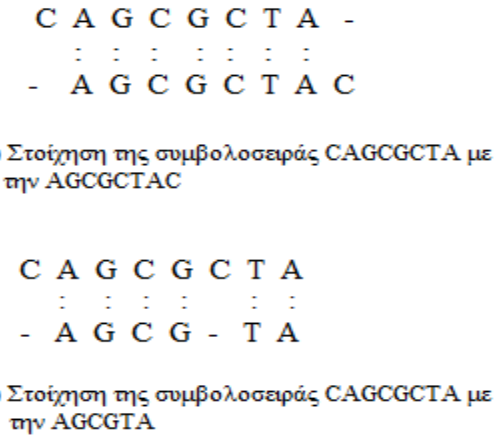
5.1 Η σημασία της σύγκρισης DNA ακολουθιών

Με την εύρεση ενός νέου γονιδίου, οι βιολόγοι δεν γνωρίζουν απολύτως τίποτα για τη λειτουργία του. Μια συνηθισμένη προσέγγιση για να αναφερθούν στη λειτουργία της αλληλουχίας ενός νέου γονιδίου είναι να βρουν ομοιότητες με γονίδια των οποίων η λειτουργία είναι γνωστή. Ένα παράδειγμα τέτοιας βιολογικής ανακάλυψης, που έγινε μέσω έρευνας ομοιοτήτων, συνέβη το 1984 όταν οι επιστήμονες χρησιμοποίησαν μια απλή υπολογιστική τεχνική για να συγκρίνουν τα νεο-ανακαλυφθείσα v-sis ογκογονίδια τα οποία προκαλούν καρκίνο, με όλα τα γνωστά (μέχρι εκείνη τη στιγμή) γονίδια. Προς έκπληξή τους, διαπίστωσαν ότι το καρκινογόνο γονίδιο ταίριαζε με ένα κανονικό γονίδιο που εμπλέκεται στην ανάπτυξη και ονομάζεται platelet-derived growth factor (PDGF). Μετά την ανακάλυψη αυτή της ομοιότητας, οι επιστήμονες άρχισαν να υποψιάζονται ότι ο καρκίνος μπορεί να προκληθεί από ένα φυσιολογικό γονίδιο ανάπτυξης το οποίο ενεργοποιείται τη λάθος στιγμή. Στην ουσία, ένα καλό γονίδιο κάνει το σωστό πράγμα τη λάθος στιγμή.
Ένα άλλο παράδειγμα επιτυχούς αναζήτησης ομοιοτήτων ήταν η ανακάλυψη του γονιδίου της κυστικής ίνωσης. Το γονίδιο αυτό κωδικοποιεί μια ρυθμιστική πρωτεΐνη (Cystic Fibrosis Transmembrane Conductance Regulator), η οποία ελέγχει την διέλευση χλωρίου διαμέσου των μεμβρανών των επιθηλιακών κυττάρων διαφόρων οργάνων του σώματος όπως των πνευμόνων, του παγκρέατος, των ιδρωτοποιών αδένων και του εντέρου. Μεταλλάξεις στο γονίδιο αυτό προκαλούν μειωμένη παραγωγή ή λειτουργικότητα της πρωτεΐνης με αποτέλεσμα να παράγεται παχύρρευστη κολλώδης βλέννα στο επιθήλιο των προσβαλλόμενων οργάνων. Η βλέννα αυτή φράσσει τους πόρους των αδένων με συνέπεια την προοδευτική καταστροφή του ιστού των οργάνων (ίνωση) και την τελική ανεπάρκεια τους.
Το 1989 η έρευνα για το γονίδιο της κυστικής ίνωσης περιορίστηκε σε μια περιοχή ενός εκατομμυρίου νουκλεοτιδίων στο χρωμόσωμα 7, αλλά η ακριβής θέση του γονιδίου παρέμενε άγνωστη. Όταν η αλληλουχία γύρω από το γονίδιο της κυστικής ίνωσης προσδιορίστηκε, οι βιολόγοι την σύγκριναν με όλα τα γνωστά γονίδια και ανακάλυψαν ομοιότητες μεταξύ ορισμένων τμημάτων της περιοχής αυτής καθώς και ένα γονίδιο που είχαν ήδη ανακαλύψει και ήταν γνωστό για την κωδικοποίηση δεσμευτικών πρωτεϊνών

τριφοσφορικής αδενοσίνης (ATP binding proteins) Αυτές οι πρωτεΐνες καλύπτουν την κυτταρική μεμβράνη, γεγονός που δίνει μια εύλογη εξήγηση για τη λειτουργία του γονιδίου της κυστικής ίνωσης, με δεδομένο ότι η νόσος περιλαμβάνει εκκρίσεις ιδρώτα με ασυνήθιστα υψηλή περιεκτικότητα σε νάτριο.
Ο καθορισμός μιας σχέσης μεταξύ καρκινικών γονιδίων και φυσιολογικών γονιδίων ανάπτυξης και η αποσαφήνιση της φύσης της κυστικής ίνωσης ήταν μόνο οι πρώτες επιτυχίες σύγκρισης ακολουθιών. Ακολούθησαν πολλές εφαρμογές αλγορίθμων σύγκρισης ακολουθιών και οι προσεγγίσεις της βιοπληροφορικής είναι μεταξύ των κυρίαρχων τεχνικών για την ανακάλυψη της λειτουργίας των γονιδίων.
Στο κεφάλαιο αυτό περιγράφουμε αλγορίθμους που επιτρέπουν στους βιολόγους να αποκαλύψουν την ομοιότητα μεταξύ διαφορετικών αλληλουχιών DNA.

5.2 Απόσταση σύνταξης και ευθυγράμμιση αλληλουχιών

Μέχρι στιγμής, έχουμε αναφερθεί στον όρο «ομοιότητα αλληλουχιών" ή "απόσταση" μεταξύ των αλληλουχιών του DNA. Η απόσταση Hamming (που αναφέραμε στο κεφάλαιο 4) ενώ είναι σημαντική στην επιστήμη των υπολογιστών, ουσιαστικά δεν χρησιμοποιείται για τη σύγκριση αλληλουχιών DNA ή πρωτεϊνών. Αυτό, γιατί ο υπολογισμός της απόστασης Hamming προϋποθέτει ότι το i-στο σύμβολο μιας ακολουθίας είναι ευθυγραμμισμένο με το υπ 'αριθμόν i σύμβολο της άλλης. Υπάρχει όμως η περίπτωση να μην ισχύει κάτι τέτοιο. Δηλαδή, το i-στο σύμβολο της μιας αλληλουχίας να αντιστοιχεί με ένα σύμβολο σε μια διαφορετική και άγνωστη θέση της άλλης αλληλουχίας.
Στην εικόνα 27 (α) φαίνεται ότι ενώ οι συμβολοσειρές CAGCGCTA και AGCGCTAC από την άποψη της απόστασης Hamming είναι πολύ διαφορετικές, γίνονται παρόμοιες αν κάποιος απλά μετακινήσει τη δεύτερη συμβολοσειρά ώστε να ευθυγραμμιστεί το (i-στο +1) γράμμα της AGCGCTAC με το i-στο γράμμα της CAGCGCTA όπου 1<= i <= 7. Οι συμβολοσειρές CAGCGCTA και AGCGTA αποτελούν παράδειγμα με περισσότερες ομοιότητες. Στην εικόνα 27 (β) αποκαλύπτονται αυτές οι ομοιότητες ευθυγραμμίζοντας τη θέση 2 της CAGCGCTA στη θέση 1 της AGCGTA. Έτσι έχουμε στοίχιση και στις θέσεις 3-2, 4-3, 5-4, 7-5, 8-6 ενώ στις θέσεις 1 και 6 της συμβολοσειράς CAGCGCTA όχι.

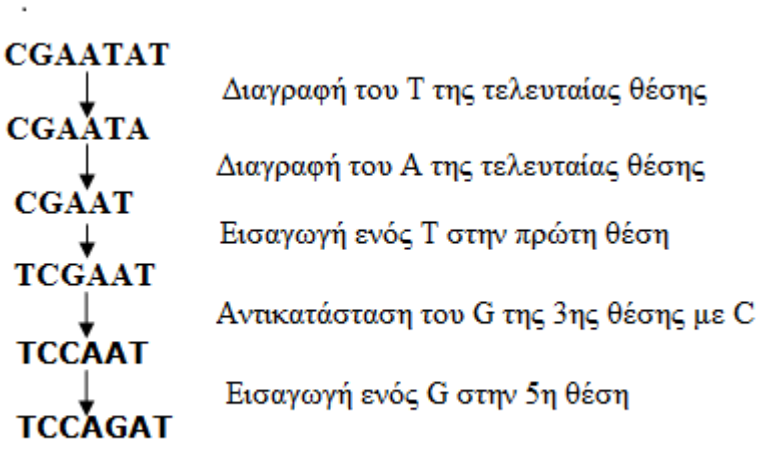


(α) Στοίχηση της συμβολοσειράς CAGCGCTA με την AGCGCTAC

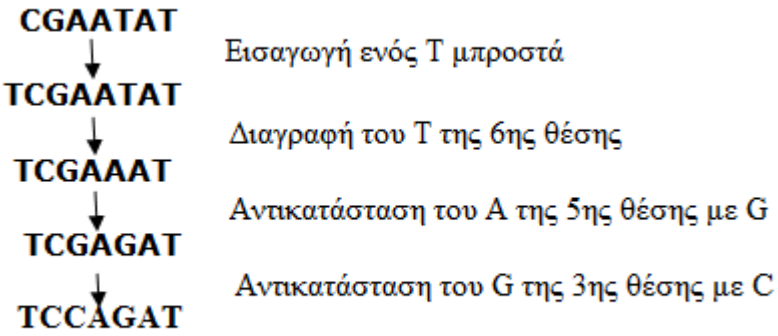
(β) Στοίχηση της συμβολοσειράς CAGCGCTA με την AGCGTA

Εικόνα 27 : Ευθυγράμμιση των CAGCGCTA – AGCGCTAC και των CAGCGCTA – AGCGTA

Το 1966, ο Vladimir Levenshtein διατύπωσε έναν αλγόριθμο ο οποίος εξετάζει την ομοιότητα μεταξύ δύο αλφαριθμητικών, και είναι ευρύτερα γνωστός με τον όρο Edit Distance, ο οποίος σε ελεύθερη μετάφραση σημαίνει απόσταση σύνταξης. Ο αλγόριθμος εξετάζει την ομοιότητα μεταξύ δύο αλφαριθμητικών, υπολογίζοντας τον ελάχιστο αριθμό πράξεων που χρειάζονται προκειμένου να επιτευχθεί η μετατροπή του ενός αλφαριθμητικού στο άλλο. Ο όρος πράξη αντιστοιχεί στην εισαγωγή, διαγραφή ή αντικατάσταση ενός μοναδικού χαρακτήρα στο αλφαριθμητικό. Για παράδειγμα, η αλληλουχία CGAATAT μπορεί να μετατραπεί στην TCCAGAT με πέντε πράξεις όπως φαίνεται στην εικόνα παρακάτω (εικόνα 28). Αυτό σημαίνει ότι η απόσταση σύνταξης μεταξύ τους είναι το πολύ 5. Στην πραγματικότητα, η απόσταση σύνταξης μεταξύ τους είναι 4, επειδή μπορούμε να μετασχηματίσουμε τη μια στην άλλη με μια πράξη λιγότερη, όπως φαίνεται στην εικόνα 29.



***Εικόνα 28** : Πέντε πράξεις για την μετατροπή της CGAATAT στην TCCAGAT*



***Εικόνα 29** : Τέσσερις πράξεις μπορούν επίσης να μετατρέψουν την CGAATAT στην TCCAGAT*

Σε αντίθεση με την απόσταση Hamming, η απόσταση σύνταξης (edit distance) μπορεί να επεξεργαστεί συμβολοσειρές διαφορετικού μήκους.
Η ευθυγράμμιση των συμβολοσειρών v (n χαρακτήρων) και w (m χαρακτήρων, με m όχι απαραίτητα ίδιο με το n), γίνεται σε έναν πίνακα δύο γραμμών όπου η πρώτη γραμμή περιέχει τους χαρακτήρες της v, ενώ η δεύτερη γραμμή περιέχει τους χαρακτήρες της w. Μπορούν να υπάρχουν σε διαφορετικά σημεία ανάμεσα στους χαρακτήρες των συμβολοσειρών κενά διαστήματα, επομένως ο πίνακας ευθυγράμμισης μπορεί να έχει το πολύ n + m στήλες.

C	G	-	T	A	C	T	G	-
C	G	C	T	A	-	T	-	A

Λέμε ότι έχουμε ταίριασμα (match) στις στήλες που έχουμε το ίδιο γράμμα και στις δύο σειρές, ενώ αναντιστοιχία (mismatch) σε αυτές που περιέχουν διαφορετικά λέμε. Οι στήλες που περιέχουν ένα κενό διάστημα καλούνται indels, με τις στήλες που περιέχουν κενό στην πρώτη γραμμή να ονομάζεται προσθήκες (insertions) και τις στήλες με κενό διάστημα στη δεύτερη γραμμή διαγραφές (deletions). Η ευθυγράμμιση της εικόνας 29 έχει 5 ταιριάσματα (matches), 0 αναντιστοιχίες (mismatches) και 4 indels. Ο αριθμός των ταιριασμάτων συν των αριθμό των αναντιστοιχιών συν τον αριθμό των indels είναι ίσος με το μήκος του πίνακα ευθυγράμμισης και πρέπει να είναι μικρότερο από n + m.
Καθεμία από τις δύο σειρές στον πίνακα ευθυγράμμισης περιέχει ένα αλφαριθμητικό με διάσπαρτα σύμβολα κενού διαστήματος « - ». Για παράδειγμα: CG-TACTG είναι η αναπαράσταση της γραμμής που αντιστοιχεί στη συμβολοσειρά v = CGTACTG, ενώ CGCTA – A - C είναι η αναπαράσταση της γραμμής που αντιστοιχεί στη συμβολοσειρά w = CGCTATA. Ένας άλλος τρόπος να αναπαραστήσουμε τη σειρά CG -TACTG - είναι 1 2 2 3 4 5 6 7 7, η οποία δείχνει τον αριθμό των συμβόλων της που έχουν παρουσιαστεί μέχρι μια δεδομένη θέση. Ομοίως, η CGCTA – T - A αναπαρίσταται ως 1 2 3 4 5 5 6 6 7. Όταν και οι δύο σειρές αναπαρίστανται με αυτόν τον τρόπο, τότε ο πίνακας ευθυγράμμισης έχει ως εξής:

0	1	2	2	3	4	5	6	7	7
0	1	2	3	4	5	5	6	6	7

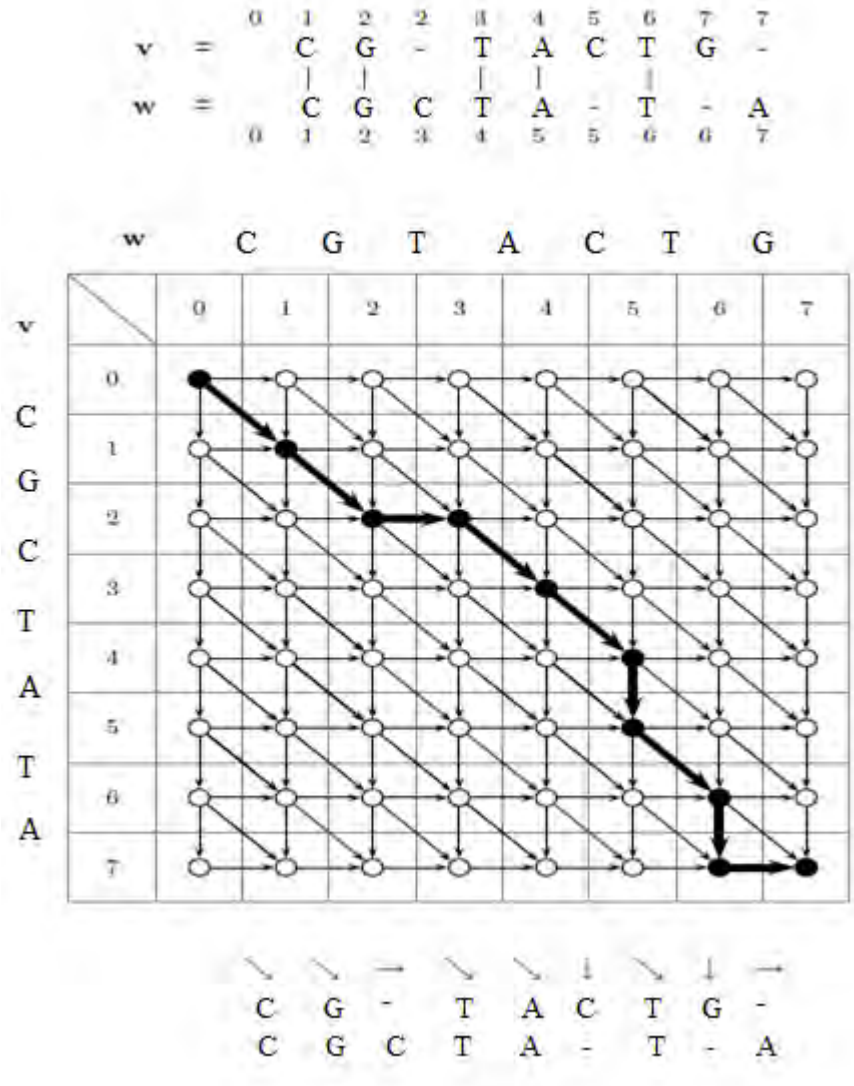
Το αποτέλεσμα της ευθυγράμμισης είναι το μονοπάτι

$$(0,0) \rightarrow (1,1) \rightarrow (2,2) \rightarrow (2,3) \rightarrow (3,4) \rightarrow (4,5) \rightarrow (5,5) \rightarrow (6,6) \rightarrow (7,6) \rightarrow (7,7)$$

Για να υπολογίσουμε το παραπάνω μονοπάτι και τις πράξεις που απαιτούνται για την ευθυγράμμιση 2 συμβολοσειρών, κατασκευάζουμε μια γραφική παράσταση, που ονομάζεται γράφημα σύνταξης. Συγκεκριμένα, κατασκευάζουμε έναν πίνακα (n * m) του οποίου οι γραμμές αντιστοιχούν στα σύμβολα της μίας ακολουθίας και οι στήλες στα σύμβολα της

άλλης. Κάθε κελί αυτού του πίνακα αντιστοιχεί σε ένα ταίριασμα γραμμάτων των δύο ακολουθιών. Σε κάθε κελί δημιουργούμε μία κορυφή από την οποία θα ξεκινούν οι ακμές που αντιπροσωπεύουν κάθε πιθανή πράξη. Προκειμένου να απεικονίσουμε στο γράφημα μια πράξη χρησιμοποιούμε τους δείκτες

που αντιπροσωπεύουν το ταίριασμα (match), την προσθήκη (insertion) και τη διαγραφή (deletion) αντίστοιχα. Έτσι, σε κάθε κελί στην περιεχόμενη κορυφή πρόσκεινται 3 ακμές (δείκτες) από τις οποίες έχουμε να επιλέξουμε τη μια. Το σύνολο τελικά των επιλεγμένων ακμών συνθέτει ένα μονοπάτι το οποίο αποτελεί τη λύση μας. Πρόκειται, δηλαδή, για το σύνολο των πράξεων που απαιτούνται για την ευθυγράμμιση των αλληλουχιών.



Εικόνα 30 *: Απεικόνιση ενός μονοπατιού ευθυγράμμισης των αλληλουχιών CGCTATA και CGTACTG*

Δεδομένων 2 συμβολοσειρών υπάρχει ένας μεγάλος αριθμός διαφορετικών πινάκων στοίχησης και αντίστοιχων μονοπατιών στο γράφο σύνταξης. Για να προσδιορίσουμε ποια πράξη (μονοπάτι ή δείκτης) υπερτερεί σε σχέση με τις άλλες, βασιζόμαστε στην έννοια μιας συνάρτησης βαθμολόγησης, η οποία λαμβάνει ως είσοδο έναν πίνακα στοίχησης (ή ισοδύναμα το τελικό μονοπάτι του γραφήματος σύνταξης) και παράγει ένα σκορ που καθορίζει πόσο καλή είναι η ευθυγράμμιση. Υπάρχει ποικιλία συναρτήσεων βαθμολόγησης που μπορούμε να χρησιμοποιήσουμε, αλλά θέλουμε αυτή που δίνει υψηλότερα σκορ στις ευθυγραμμίσεις με τα περισσότερα ταιριάσματα. Η απλούστερη συνάρτηση βαθμολόγησης βαθμολογεί μια στήλη με θετικό αριθμό αν και τα δύο γράμματα είναι τα ίδια, και με αρνητικό αριθμό εάν τα δύο γράμματα είναι διαφορετικά. Το σκορ για το σύνολο της ευθυγράμμισης είναι το άθροισμα των επιμέρους βαθμολογιών των στηλών. Αυτό το σύστημα βαθμολόγησης χρησιμοποιείται για τον καθορισμό βάρους στις ακμές στο γράφο σύνταξης (edit graph).
Επιλέγοντας διαφορετικές συναρτήσεις βαθμολόγησης, μπορούμε να λύσουμε διαφορετικά προβλήματα σύγκρισης συμβολοσειρών. Αν επιλέξουμε την πολύ απλή συνάρτηση βαθμολόγησης με +1 για ταίριασμα και 0 αλλιώς, τότε το πρόβλημα μετατρέπεται στο πρόβλημα αναζήτησης της μεγαλύτερης κοινής ακολουθίας ανάμεσα στις δύο συμβολοσειρές. Πριν, λοιπόν, περιγράψουμε τον τρόπο υπολογισμού της απόστασης Levenshtein θα αναφερθούμε στο πρόβλημα αναζήτησης της μεγαλύτερης κοινής υποακολουθίας.

5.3 Πρόβλημα εύρεσης της μεγαλύτερης κοινής υποακολουθίας 2 συμβολοσειρών

Η απλούστερη μορφή ανάλυσης της ομοιότητας ακολουθιών είναι το πρόβλημα της μεγαλύτερης κοινής υποακολουθίας (Longest Common Subsequence LCS), όπου εξαλείφουμε την πράξη της αντικατάστασης και επιτρέπονται εισαγωγές και διαγραφές.
Υποακολουθία μιας συμβολοσειράς v είναι μια ακολουθία όχι απαραίτητα συνεχόμενων χαρακτήρων της v. Για παράδειγμα, αν v = ATTGCTA, τότε η AGCA και ATTA είναι υποακολουθίες της v ενώ οι TGTT και TCG δεν είναι.
Κοινή υποακολουθία δύο συμβολοσειρών ονομάζεται η συμβολοσειρά που αποτελεί υποακολουθία και των δυο.
Ορίζουμε την κοινή υποακολουθία των συμβολοσειρών $v = v_1 \ldots v_n$ και $w = w_1 \ldots w_m$ ως μια ακολουθία από θέσεις της v,

$$1 \leq i_1 < i_2 < \cdots < i_k \leq n$$

και μια ακολουθία από θέσεις της w,

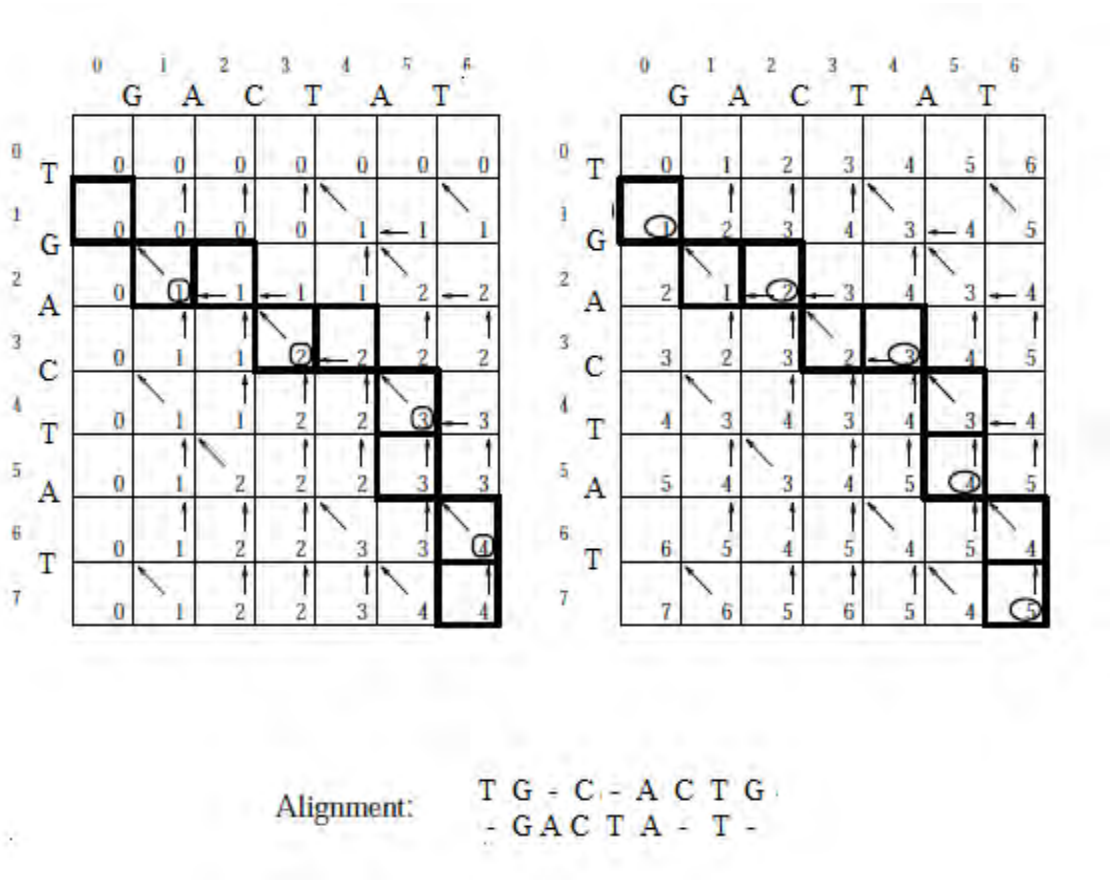
$$1 \leq j_1 < j_2 < \cdots < j_k \leq m$$

τέτοια ώστε τα σύμβολα στις αντίστοιχες θέσεις των v και w να συμπίπτουν:

$$v_{i_t} = w_{j_t} \text{ for } 1 \leq t \leq k.$$

Για παράδειγμα, η υποακολουθία TCTA είναι κοινή για τις συμβολοσειρές ATCTGAT και TGCATA.

Αν και υπάρχουν πολλές κοινές υποακολουθίες μεταξύ δύο συμβολοσειρών v και w, κάποια από αυτές είναι μεγαλύτερη από τις άλλες, δεν είναι αμέσως προφανές το πώς θα βρούμε αυτή τη μεγαλύτερη. Αν θεωρήσουμε ως s (v, w) το μήκος της μεγαλύτερης κοινής υποακολουθίας των v και w, τότε η απόσταση σύνταξης μεταξύ v και w (με την παραδοχή ότι επιτρέπονται μόνο εισαγωγές και διαγραφές) είναι d (v, w) = n + m - 2s (v, w), και αντιστοιχεί στον ελάχιστο αριθμό των εισαγωγών και διαγραφών που απαιτούνται για τη μετατροπή της v σε w. Στην εικόνα 31 φαίνεται η μέγιστη κοινή υποακολουθία τεσσάρων γραμμάτων και η μικρότερη ακολουθία δυο εισαγωγών και τριών διαγραφών για τη μετατροπή της συμβολοσειρών v = ATCTGAT στην w = TGCATA

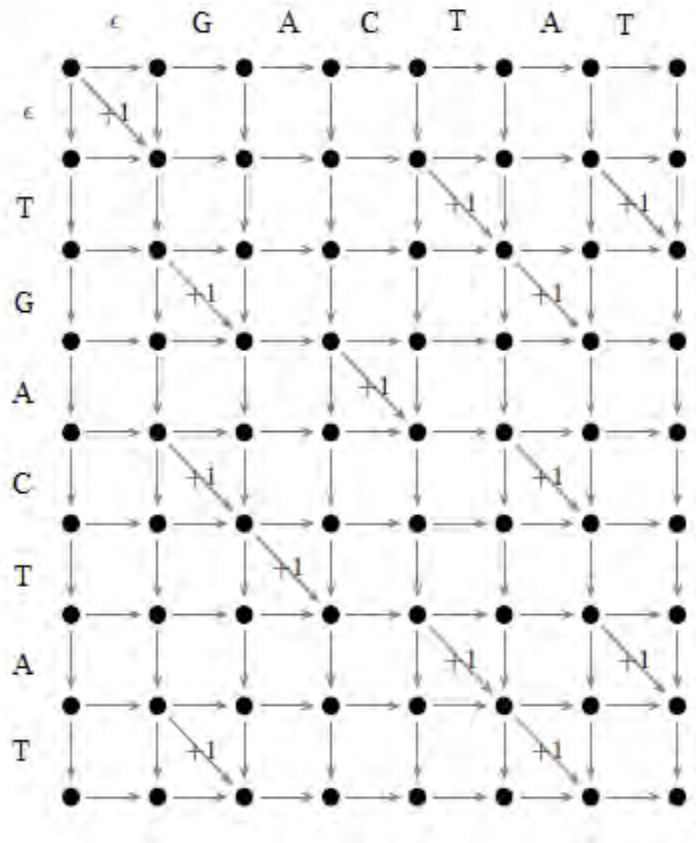


Εικόνα 31 *: Αλγόριθμος δυναμικού προγραμματισμού για την εύρεση μέγιστης κοινής υποακολουθίας*

Πρόβλημα μέγιστης κοινής υποακολουθίας (Longest Subsequence Problem)
Βρίσκουμε τη μέγιστη κοινή υποακολουθία δυο συμβολοσειρών

Είσοδος (Input): Δυο συμβολοσειρές v και w
Έξοδος (Output): Τη μέγιστη κοινή υποακολουθία των συμβολοσειρών v και w

Κάθε κοινή υποακολουθία αντιστοιχεί σε ευθυγράμμιση με καμία αναντιστοιχία. Αυτό μπορεί να επιτευχθεί μόνο με την άρση όλων διαγώνιων ακμών του γράφου στα κελιά που οι χαρακτήρες τους δεν ταιριάζουν όπως ακριβώς φαίνεται και στην εικόνα 30



Εικόνα 32 *: Γράφος σύνταξης ενός LCS προβλήματος*

Ορίζουμε $s_{i,j}$ το μήκος μιας LCS μεταξύ $v_1 \ldots v_i$ και $w_1 \ldots w_j$.

$$s_{i,0} = s_{0,j} = 0 \quad \text{για κάθε i όπου } 1 \leq i \leq n$$

Για $s_{i,j}$ ισχύει:

$$s_{i,j} = \max \begin{cases} s_{i-1,j} \\ s_{i,j-1} \\ s_{i-1,j-1} + 1 \text{ εάν } v_i = w_j \end{cases}$$

Ο πρώτος όρος αντιστοιχεί στην περίπτωση που δεν εμφανίζεται v_i στην LCS (εδώ πρόκειται για διαγραφή του v_i). Ο δεύτερος αντιστοιχεί στην περίπτωση που δεν εμφανίζεται w_j στην LCS (εδώ πρόκειται για εισαγωγή του w_j) και ο τρίτος όρος αντιστοιχεί στην περίπτωση όπου εμφανίζεται και v_i και w_j στην LCS (το v_i ταιριάζει με w_j).
Στη συνέχεια, χρησιμοποιούμε το s για να αναπαραστήσουμε τον πίνακα δυναμικού προγραμματισμού. Το μήκος μιας LCS μεταξύ v και w μπορεί να διαβαστεί από τον πίνακα δυναμικού προγραμματισμού (n, m), αλλά για να ανακατασκευάσουμε την LCS από αυτόν πρέπει να κρατήσουμε την επιπρόσθετη πληροφορία σχετικά με το ποια από τις τρεις ποσότητες, $s_{i-1,j}$, $s_{i,j-1}$ ή $s_{i-1,j-1} + 1$ είναι η μέγιστη για το $s_{i,j}$.
Ο ακόλουθος αλγόριθμος επιτυγχάνει το στόχο αυτό με την εισαγωγή δεικτών υπαναχώρησης (backtracking pointers) που λαμβάνουν μία από τις τρεις τιμές ← , ↑ ή ↖ . Οι δείκτες αυτοί καθορίζουν ποια από τις παραπάνω τρεις περιπτώσεις αντιπροσωπεύουν (δηλαδή , $s_{i-1,j}$, $s_{i,j-1}$ ή $s_{i-1,j-1} + 1$) και αποθηκεύονται σε ένα δισδιάστατο πίνακα B (Εικόνα 31)

```
LCS(v, w)
1  for i ← 0 to n
2      s_{i,0} ← 0
3  for j ← 1 to m
4      s_{0,j} ← 0
5  for i ← 1 to n
6      for j ← 1 to m
7          s_{i,j} ← max { s_{i-1,j}
                         { s_{i,j-1}
                         { s_{i-1,j-1} + 1,   if v_i = w_j
8          b_{i,j} ← { "↑"   if s_{i,j} = s_{i-1,j}
                     { "←"   if s_{i,j} = s_{i,j-1}
                     { "↖",  if s_{i,j} = s_{i-1,j-1} + 1
9  return (s_{n,m}, b)
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Το παρακάτω πρόγραμμα χρησιμοποιώντας αναδρομή εκτυπώνει τη μεγαλύτερη κοινή υποακολουθία χρησιμοποιώντας τις πληροφορίες που είναι αποθηκευμένες στον πίνακα B. Η αρχική κλήση που εκτυπώνει τη λύση του προβλήματος είναι PRINTLCS (b, v, n, m).
Ο πίνακας δυναμικού προγραμματισμού στην εικόνα 31 (αριστερά) παρουσιάζει τον υπολογισμό του σκορ s(v, w) μεταξύ v και w, ενώ ο πίνακας δεξιά παρουσιάζει τον υπολογισμό της απόστασης σύνταξης (edit distance) μεταξύ v και w με την παραδοχή ότι οι μόνες πράξεις που επιτρέπονται είναι οι εισαγωγές και οι διαγραφές.
Η απόσταση σύνταξης d (v, w) υπολογίζεται σύμφωνα με τις αρχικές συνθήκες $d_{i,0} = i$, $d_{0,j} = j$ για $1 <= i <= n$ και $1 <= j <= m$ και την παρακάτω σχέση

$$d_{i,j} = \min \begin{cases} d_{i-1,j} + 1 \\ d_{i,j-1} + 1 \\ d_{i-1,j-1}, \quad \text{εάν } v_i = w_j \end{cases}$$

```
PRINTLCS(b, v, i, j)
1  if i = 0 or j = 0
2      return
3  if b_{i,j} = "↖"
4      PRINTLCS(b, v, i − 1, j − 1)
5      print v_i
6  else
7      if b_{i,j} = "↑"
8          PRINTLCS(b, v, i − 1, j)
9      else
10         PRINTLCS(b, v, i, j − 1)
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

5.4 Γενίκευση προβλήματος ευθυγράμμισης

Το πρόβλημα LCS αντιστοιχεί σε μια μάλλον περιοριστική βαθμολόγηση καθώς προσδίδει 1 για τα ταιριάσματα και δεν υπάρχει καμία ποινή για τα indels.
Προκειμένου να γενικεύσουμε τη βαθμολόγηση, επεκτείνουμε το αλφάβητο A κ γραμμάτων ώστε να περιλαμβάνει τον κενό χαρακτήρα «-» και εξετάζουμε έναν τυχαίο πίνακα βαθμολόγησης δ με διαστάσεις (κ+1)*(κ+1) όπου k είναι συνήθως 4 ή 20, ανάλογα με τον τύπο των ακολουθιών (DNA ή πρωτεΐνη) που αναλύουμε. Το σκορ κάθε στήλης είναι το δ(x,y) και το σκορ ευθυγράμμισης ορίζεται ως το άθροισμα των βαθμολογιών των

στηλών. Με αυτόν τον τρόπο μπορούμε να λάβουμε υπόψη στη βαθμολόγηση τις αναντιστοιχίες και τα indels. Αντί να επιλέγουμε έναν συγκεκριμένο πίνακα βαθμολόγησης και στη συνέχεια να επιλύουμε εκ νέου το πρόβλημα ευθυγράμμισης, θέτουμε ένα γενικό πρόβλημα ευθυγράμμισης (global sequence alignment) που παίρνει ως είσοδο τον πίνακα βαθμολόγησης.

Οι αντίστοιχες σχέσεις για το σκορ $s_{i,j}$ της βέλτιστης ευθυγράμμισης μεταξύ της θέσης i της συμβολοσειράς v και j της συμβολοσειράς w έχουν ως εξής:

$$s_{i,j} = \max \begin{cases} s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j) \end{cases}$$

όπου οι αναντιστοιχίες «τιμωρούνται» με μία σταθερά –μ, τα indels με μια σταθερά –σ, και ταιριάσματα ανταμείβονται με +1. Έτσι, το αποτέλεσμα της βαθμολογίας είναι:

ταιριάσματα – μ * αναντιστοιχίες – σ * indels

Έτσι, η παραπάνω σχέση μπορεί να γραφεί ως εξής

$$s_{i,j} = \max \begin{cases} s_{i-1,j} - \sigma \\ s_{i,j-1} - \sigma \\ s_{i-1,j-1} - \mu, \text{ if } v_i \neq w_j \\ s_{i-1,j-1} + 1, \text{ if } v_i = w_j \end{cases}$$

Μπορούμε να αποθηκεύσουμε ξανά πληροφορίες με παρόμοιους δείκτες οπισθοδρόμησης ώστε να ανακατασκευάσουμε την ευθυγράμμιση από τον πίνακα δυναμικού προγραμματισμού. Παρατηρούμε ότι το πρόβλημα LCS είναι το γενικευμένο πρόβλημα ευθυγράμμισης με παραμέτρους μ = 0, σ= 0 (ή ισοδύναμα, μ= ∞ , σ = 0).
Το γενικευμένο πρόβλημα ευθυγράμμισης αναζητά ομοιότητες μεταξύ δύο ολόκληρων συμβολοσειρών. Αυτό είναι χρήσιμο όταν η ομοιότητα μεταξύ των συμβολοσειρών εκτείνεται σε ολόκληρο το μήκος τους, για παράδειγμα, σε ακολουθίες πρωτεϊνών από την ίδια οικογένεια. Αυτές οι ακολουθίες πρωτεϊνών συχνά διατηρούνται και έχουν σχεδόν το ίδιο μήκος σε όλους τους οργανισμούς από τα φρούτα και τις μύγες μέχρι τον άνθρωπο. Ωστόσο, σε πολλές βιολογικές εφαρμογές, το σκορ ευθυγράμμισης μεταξύ δύο υποακολουθιών των v και w μπορεί στην πραγματικότητα να είναι μεγαλύτερο από το σκορ ευθυγράμμισης μεταξύ ολόκληρων των v και w.
Για παράδειγμα, τα γονίδια homeobox, που ρυθμίζουν την εμβρυϊκή ανάπτυξη, παρουσιάζονται σε έναν μεγάλο αριθμό ειδών. Αν και τα γονίδια homeobox είναι πολύ διαφορετικά στα διάφορα είδη, μια περιοχή σε κάθε γονίδιο, που ονομάζεται homeodomain, διατηρείται σε μεγάλο βαθμό. Έτσι, το ερώτημα που προκύπτει είναι πώς θα βρούμε αυτό το τμήμα που διατηρείται αγνοώντας τις περιοχές που παρουσιάζουν μικρή ομοιότητα. Το

1981 οι Temple Smith and Michael Waterman πρότειναν μια έξυπνη τροποποίηση του γενικευμένου αλγορίθμου ευθυγράμμισης δυναμικού προγραμματισμού που λύνει το πρόβλημα τοπικής στοίχησης.
Στην εικόνα 33 παρουσιάζεται η σύγκριση δύο υποθετικών γονιδίων v και w ίδιου μήκους με το τμήμα που διατηρείται να βρίσκεται στις αρχές του v και στο τέλος του w. Για λόγους απλούστευσης, θα υποθέσουμε ότι το τμήμα που διατηρείται σε αυτά τα δύο γονίδια είναι πανομοιότυπο και καλύπτει το ένα τρίτο του συνολικού μήκους n των γονιδίων. Στην περίπτωση αυτή το μονοπάτι στον πίνακα δυναμικού προγραμματισμού περιλαμβάνει (2/3)*n οριζόντιες ακμές, (1/3) *n διαγώνιες ακμές δηλαδή ταιριάσματα και (2/3)*n κάθετες ακμές. Έτσι, η βαθμολογία του μονοπατιού αυτού είναι:

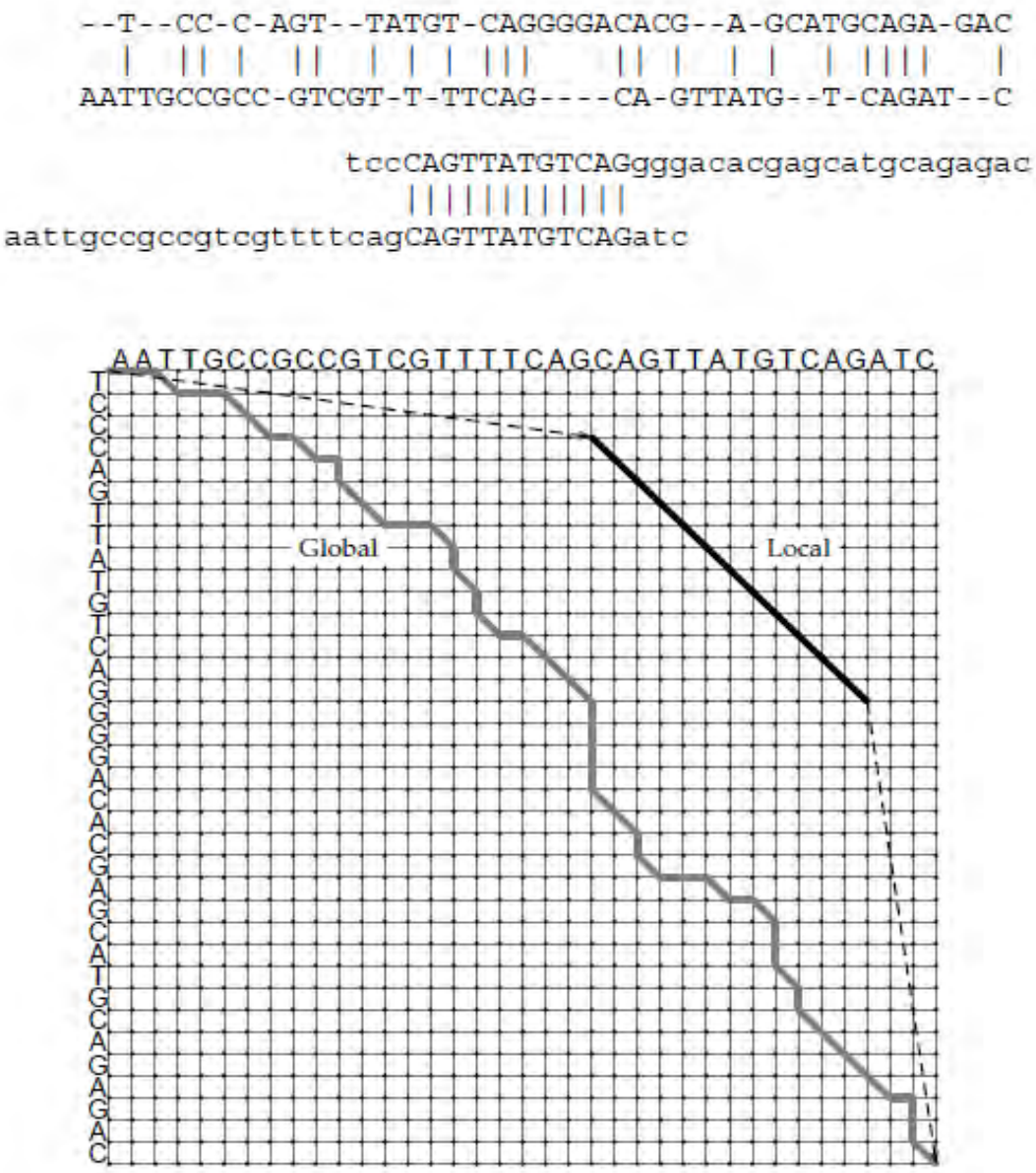
$$-\frac{2}{3}n\sigma + \frac{1}{3}n - \frac{2}{3}n\sigma = n\left(\frac{1}{3} - \frac{4}{3}\sigma\right)$$

Όμως, το μονοπάτι αυτό περιέχει τόσα πολλά indels που είναι απίθανο να είναι η ευθυγράμμιση με τη μεγαλύτερη βαθμολογία. Στην πραγματικότητα, άσχετα διαγώνια βιολογικά μονοπάτια από την αρχή μέχρι το τέλος είναι πιθανόν να έχουν μεγαλύτερο σκορ από αυτό της ευθυγράμμισης, καθώς οι αναντιστοιχίες «τιμωρούνται» λιγότερο από τα indels.
Η αναμενόμενη βαθμολογία ενός τέτοιου διαγώνιου μονοπατιού είναι :

$$n(\tfrac{1}{4} - \tfrac{3}{4}\mu)$$

αφού κάθε διαγώνια ακμή αντιστοιχεί σε έναν ταίριασμα με πιθανότητα 1/4 και σε αναντιστοιχία με πιθανότητα ¾.

Εφόσον $(\frac{1}{3} - \frac{4}{3}\sigma) < (\frac{1}{4} - \frac{3}{4}\mu)$ για πολλές κυρώσεις indel και αναντιστοιχίας, ο γενικευμένος αλγόριθμος ευθυγράμμισης θα χάσει τη σωστή λύση του πραγματικού βιολογικού προβλήματος, και είναι πιθανό να δώσει ως αποτέλεσμα ένα βιολογικά άσχετο μονοπάτι με σχεδόν διαγώνια πορεία. Η εικόνα 33 δείχνει ακριβώς αυτή την παρατήρηση.
Όταν βιολογικά σημαντικές ομοιότητες υπάρχουν σε ορισμένα μέρη τμημάτων DNA και δεν είναι παρούσες σε άλλα, οι βιολόγοι προσπαθούν να μεγαλώσουν το σκορ ευθυγράμμισης s (v_i... $v_{i`}$, $w_j$...$w_{j`}$), σε σχέση με όλες τις υποσυμβολοσειρές v_i... $v_{i`}$ της και $w_j$...$w_{j`}$ της w. Αυτό ονομάζεται τοπικό πρόβλημα στοίχισης (Local Alignment problem) εφόσον η ευθυγράμμιση δεν σημαίνει απαραίτητα ότι επεκτείνεται σε όλο το μήκος συμβολοσειράς όπως συμβαίνει με το γενικευμένο πρόβλημα ευθυγράμμισης

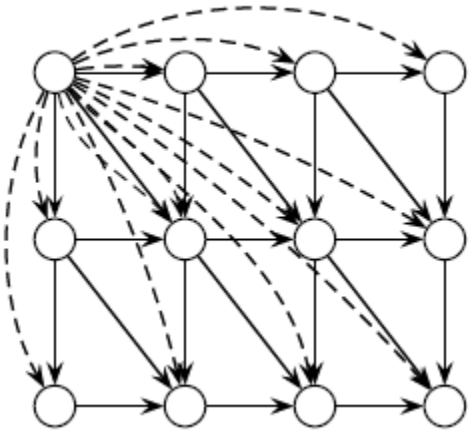


Εικόνα 33 *: (α) Γενικευμένη και (β) τις τοπική ευθυγράμμιση των δύο υποθετικών γονιδίων που το καθένα έχει μια διατηρημένη περιοχή. Η τοπική προσέγγιση έχει πολύ χειρότερη βαθμολογία σύμφωνα με το σύστημα βαθμολόγησης, αλλά εντοπίζει σωστά την περιοχή που διατηρείται*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η λύση σε αυτό το φαινομενικά πιο δύσκολο πρόβλημα είναι η συνειδητοποίηση ότι το γενικευμένο πρόβλημα ευθυγράμμισης αντιστοιχεί στην εύρεση της μεγαλύτερη τοπικής διαδρομής μεταξύ των κορυφών (0, 0) και (n, m) στο γράφο επεξεργασίας, ενώ το πρόβλημα τοπικής ευθυγράμμισης αντιστοιχεί στην εύρεση του μεγαλύτερου μεταξύ των μονοπατιών ανάμεσα στις αυθαίρετες κορυφές (i, j) και (i', j') στο γράφο επεξεργασίας. Μια απλή και αναποτελεσματική προσέγγιση αυτού του προβλήματος είναι να βρούμε το

μεγαλύτερο μονοπάτι για κάθε ζευγάρι κορυφών (i, j) και (i', j') και στη συνέχεια να επιλέξουμε το μεγαλύτερο από αυτά. Αντί να βρούμε το μεγαλύτερο μονοπάτι από κάθε κορυφή (i, j) σε κάθε άλλη κορυφή (i', j'), το πρόβλημα τοπικής ευθυγράμμισης μπορεί να περιοριστεί στην εύρεση των μεγαλύτερων μονοπατιών από την αρχή (source(0,0)) σε κάθε άλλη κορυφή, προσθέτοντας ακμές βάρους 0 στο γράφο επεξεργασίας. Αυτές οι ακμές κάνουν την κορυφή (0,0) πρόγονο κάθε κόμβου στο γράφο και προσφέρουν μια «ελεύθερη πρόσβαση" από αυτήν σε οποιαδήποτε άλλη κορυφή (i, j) (εικόνα 34).



Εικόνα 34 *Ο αλγόριθμος Smith-Waterman τοπικής ευθυγράμμισης εισάγει άκρες βάρους 0 (εδώ φαίνονται με διακεκομμένες γραμμές) από την πηγή κορυφή (0, 0) σε κάθε άλλη κορυφή στο γράφημα επεξεργασίας*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

$$s_{i,j} = \max \begin{cases} 0 \\ s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j) \end{cases}$$

Η μεγαλύτερη τιμή του $s_{i,j}$ σε ολόκληρο το γράφο επεξεργασίας αντιπροσωπεύει το σκορ της καλύτερης ευθυγράμμισης των v και w. Στο γενικευμένο πρόβλημα ευθυγράμμισης κοιτάμε το $s_{n,m}$. Η διαφορά μεταξύ τοπικής και γενικευμένης ευθυγράμμισης φαίνεται στην αρχή της εικόνας 33.
Η βέλτιστη τοπική ευθυγράμμιση είναι το μεγαλύτερο μονοπάτι στο γράφημα επεξεργασίας. Την ίδια στιγμή, διάφορες τοπικές ευθυγραμμίσεις μπορεί να έχουν βιολογική σημασία και έχουν αναπτυχθεί μέθοδοι για να βρίσκουμε τις k καλύτερες μη επικαλυπτόμενες ευθυγραμμίσεις. Οι μέθοδοι αυτές είναι ιδιαίτερα σημαντικές για τη σύγκριση πρωτεϊνών πολλών τομέων που μοιράζονται παρόμοια τμήματα (blocks). Στην περίπτωση αυτή, δεν υπάρχει μία τοπική ευθυγράμμιση που να αντιπροσωπεύει όλες τις σημαντικές ομοιότητες.

5.5 Γονιδιακή πρόγνωση

Το 1961 οι Sydney Brenner και Francis Crick απέδειξαν ότι κάθε τριπλέτα νουκλεοτιδίων (κωδικόνιο) σε ένα γονίδιο κωδικοποιεί ένα αμινοξύ της αντίστοιχης πρωτεΐνης. Ήταν σε θέση να εισάγουν διαγραφές στο DNA και παρατήρησαν ότι διαγραφή ενός μόνο νουκλεοτιδίου ή δύο συνεχόμενων νουκλεοτιδίων σε ένα γονίδιο, μεταβάλλει δραματικά το προϊόν πρωτεΐνης. Παραδόξως, η διαγραφή τριών διαδοχικών νουκλεοτιδίων είχε ως αποτέλεσμα μικρές αλλαγές στην πρωτεΐνη. Για παράδειγμα, η φράση : THE SLY FOX AND THE SHY DOG (γραμμένο σε τριπλέτες) μετατρέπεται σε ασυναρτησίες μετά τη διαγραφή ενός γράμματος (THE SYF OXA NDT HES HYD OG) ή δύο γραμμάτων (THE SFO XAN DTH ESH YDO G), αλλά έχει κάποιο νόημα μετά τη διαγραφή τριών νουκλεοτιδίων (THE SOX AND THE SHY DOG).

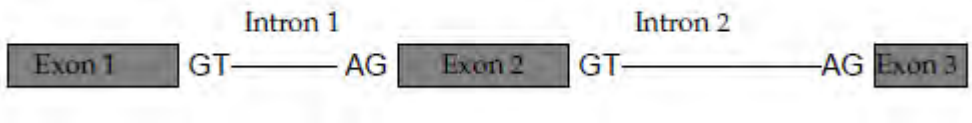
Εμπνευσμένος, λοιπόν, από αυτό το πείραμα ο Charles Yanofsky απέδειξε ότι ένα γονίδιο και το πρωτεϊνικό προϊόν του είναι γραμμικά, δηλαδή, το πρώτο κωδικόνιο στο γονίδιο κωδικοποιεί το πρώτο αμινοξύ της πρωτεΐνης, το δεύτερο κωδικόνιο κωδικοποιεί το δεύτερο αμινοξύ (και όχι, ας πούμε, το δέκατο έβδομο) και ούτω καθεξής. Το έξυπνο πείραμα Yanofsky είχε τόσο επιρροή που κανείς δεν αναρωτιόταν αν τα κωδικόνια αντιπροσωπεύονται από συνεχή τμήματα DNA και για τα επόμενα δεκαπέντε χρόνια οι βιολόγοι πίστευαν ότι μια πρωτεΐνη κωδικοποιείται από μια μεγάλη συμβολοσειρά συνεχόμενων τριπλέτων. Ωστόσο, το 1977, η ανακάλυψη διασκορπισμένων ανθρώπινων γονιδίων, απέδειξε ότι τα γονίδια εκπροσωπούνται από μια συλλογή υποσυμβολοσειρών και έθεσε το υπολογιστικό πρόβλημα της πρόβλεψης των θέσεων των γονιδίων σε ένα γονιδίωμα δοσμένης μόνο μιας γονιδιωματική ακολουθίας DNA.

Το ανθρώπινο γονιδίωμα είναι μεγαλύτερο και πιο σύνθετο από το γονιδίωμα των βακτηρίων. Αυτό δεν προκαλεί ιδιαίτερη έκπληξη δεδομένου ότι θα περίμενε κανείς να βρει περισσότερα γονίδια στους ανθρώπους από ότι σε βακτήρια. Ωστόσο, το μέγεθος του γονιδιώματος πολλών ευκαρυωτικών κυττάρων δεν φαίνεται να σχετίζεται με τη γενετική πολυπλοκότητα ενός οργανισμού. Για παράδειγμα, το γονιδίωμα της σαλαμάνδρας είναι δέκα φορές μεγαλύτερο από αυτό του ανθρώπου. Αυτό το παράδοξο επιλύθηκε με την ανακάλυψη ότι πολλοί οργανισμοί περιέχουν όχι μόνο τα γονίδια, αλλά και μεγάλες ποσότητες του λεγόμενου «άχρηστου» (junk) DNA που δεν κωδικοποιούν καθόλου πρωτεΐνες. Ειδικότερα, τα περισσότερα ανθρώπινα γονίδια είναι σπασμένα σε κομμάτια που ονομάζεται εξόνια, που χωρίζονται από αυτό «άχρηστο» DNA. Έτσι, η διαφορά στα μεγέθη της σαλαμάνδρας και του ανθρώπινου γονιδιώματος πιθανώς αντανακλά σε μεγαλύτερες ποσότητες «άχρηστου» DNA και επαναλήψεων στο γονιδίωμα της σαλαμάνδρας.

Τα διαχωρισμένα γονίδια είναι αντίστοιχα με ένα άρθρο περιοδικού που ξεκινά στη σελίδα 1, συνεχίζεται στη σελίδα 13, στη συνέχεια εμφανίζεται και πάλι στις σελίδες 43, 51, 74, 80, και 91, με σελίδες διαφημίσεων να εμφανίζονται ενδιάμεσα. Δεν γνωρίζουμε γιατί συμβαίνουν αυτά τα άλματα και ένα σημαντικό μέρος του ανθρώπινου γονιδιώματος είναι αυτή η «άχρηστη διαφήμιση» που χωρίζει τα εξώνια.

Περισσότερο μπέρδεμα προκαλεί το γεγονός ότι τα άλματα μεταξύ των γονιδίων δεν είναι συμβατά από είδος σε είδος. Ένα γονίδιο σε ένα γονιδίωμα εντόμου οργανώνεται με διαφορετικό τρόπο από το σχετικό γονίδιο σε αυτό του σκουληκιού. Ο αριθμός των τμημάτων (εξώνια) μπορεί να είναι διαφορετικός. Οι πληροφορίες, δηλαδή, που εμφανίζονται στο ανθρώπινο γονιδίωμα μπορεί να χωριστούν στα δύο στο ποντίκι ή το αντίστροφο. Παρόλο που τα γονίδιά τους σχετίζονται, μπορεί να είναι αρκετά διαφορετικά από άποψη δομής.

Η γονιδιακή πρόβλεψη είναι το πρόβλημα εντοπισμού γονιδίων σε μια γονιδιωματική αλληλουχία. Τα ανθρώπινα γονίδια αποτελούν μόνο το 3% του ανθρώπινου γονιδιώματος, και δεν υπάρχουν αλγόριθμοι αναγνώρισης που παρέχουν εντελώς αξιόπιστη αναγνώριση των γονιδίων.
Το μοντέλο introns-exons (εσώνια-εξώνια) για ένα γονίδιο φαίνεται να επικρατεί σε ευκαρυωτικούς οργανισμούς. Οι προκαρυωτικοί οργανισμοί (όπως τα βακτήρια) δεν έχουν σπασμένα γονίδια. Κατά συνέπεια, αλγόριθμοι γονιδιακής πρόβλεψης για τους προκαρυωτικούς οργανισμούς τείνουν να είναι απλούστεροι από ότι για τους ευκαρυωτικούς.
Υπάρχουν δύο κατηγορίες προσεγγίσεων που οι ερευνητές έχουν χρησιμοποιήσει για την πρόβλεψη της θέσης των γονιδίων. Η στατιστική προσέγγιση γονιδιακής πρόβλεψης είναι να ψάχνουμε για χαρακτηριστικά που εμφανίζονται συχνά στα γονίδια και σπάνια αλλού. Πολλοί ερευνητές προσπάθησαν να αναγνωρίσουν τις θέσεις των σημάτων διαχωρισμού (splicing signals) σε τμήματα εξωνίων-εσωνίων. Για παράδειγμα, τα δινουκλεοτίδια AG και GT στην αριστερή και δεξιά πλευρά ενός εξωνίου διατηρούνται σε υψηλό βαθμό (εικόνα 35). Υπάρχουν επίσης και άλλες θέσεις που διατηρούνται σε μικρότερο βαθμό και στις δύο πλευρές των εξωνίων. Ο απλούστερος τρόπος για να αναπαραστήσουμε τέτοιες θέσεις δέσμευσης είναι με ένα προφίλ που περιγράφει τις τάσεις των διαφόρων νουκλεοτίδιων να εμφανίζονται σε διαφορετικές θέσεις. Δυστυχώς, ο εντοπισμός σημείων διαχωρισμού με τη χρήση προφίλ έχει περιορισμένη επιτυχία καθώς τα προφίλ είναι αρκετά αδύναμα και έχουν την τάση να ταιριάζουν το γονιδίωμα με αδιάσπαστες περιοχές. Οι προσπάθειες για να βελτιώσουμε την ακρίβεια της γονιδιακής πρόβλεψης οδήγησε στη δεύτερη κατηγορία προσεγγίσεων : εκείνων που βασίζονται στην ομοιότητα.



Εικόνα 35 *: Τα εξώνια συνήθως πλαισιώνονται από τα δινουκλεοτίδια AG και GT.*

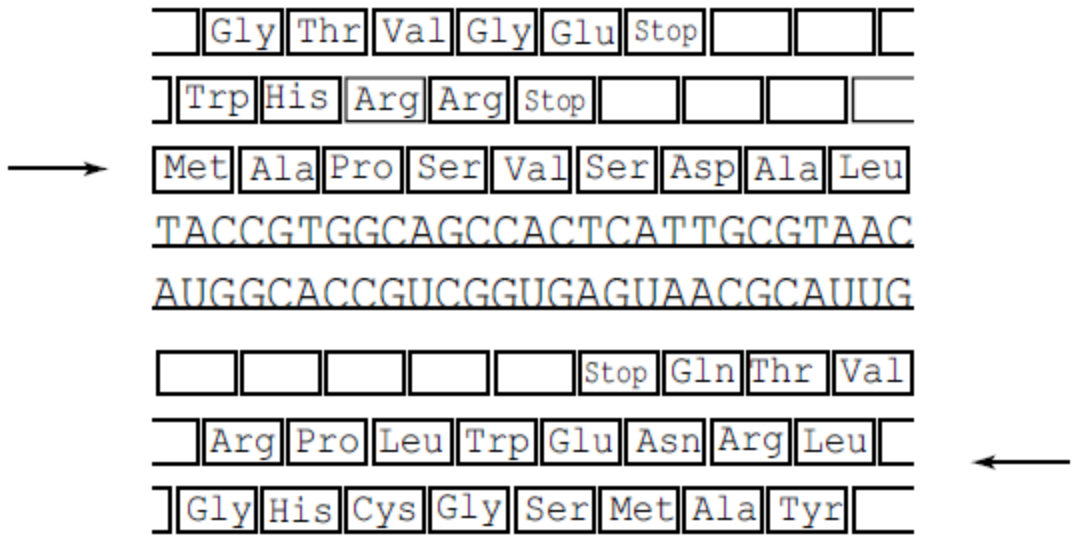
Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η βασισμένη στην ομοιότητα (similarity-based) προσέγγιση γονιδιακής πρόβλεψης βασίζεται στην παρατήρηση ότι μια νέα αλληλουχία γονιδίου έχει μεγάλη πιθανότητα να συσχετιστεί με κάποια που είναι ήδη γνωστή. Για παράδειγμα, το 99% των γονιδίων του ποντικιού έχουν ανθρώπινες αναλογίες. Ωστόσο, δεν μπορούμε απλά να κοιτάξουμε για μια παρόμοια ακολουθία στο γονιδίωμα ενός οργανισμού βασιζόμενοι στα γονίδια που γνωρίζουμε στο άλλο, για τους λόγους που προαναφέρθηκαν: τόσο η ακολουθία όσο και η δομή του εξωνίου του σχετικού γονιδίου σε διαφορετικά είδη είναι διαφορετικές. Το κοινό μεταξύ των σχετικών γονιδίων είναι ότι και στους δύο οργανισμούς παράγουν παρόμοιες πρωτεΐνες. Κατά συνέπεια, αντί να ασχολούμαστε με τη στατιστική ανάλυση των εξωνίων, μέθοδοι βασισμένες στην ομοιότητα προσπαθούν να λύσουν το συνδυαστικό γρίφο: βρείτε ένα σύνολο συμβολοσειρών (υποτιθέμενα εξώνια) σε μια γονιδιωματική ακολουθίας (ας πούμε του ποντικιού), των οποίων η σύνδεση ταιριάζει με μια γνωστή ανθρώπινη πρωτεΐνη. Σε αυτό το σενάριο, υποθέτουμε ότι γνωρίζουμε μια ανθρώπινη πρωτεΐνη, και θέλουμε να

ανακαλύψουν τη δομή των εξωνίων του σχετικού γονιδίου στο γονιδίωμα του ποντικιού. Όσο περισσότερα δεδομένα συλλέγουμε, τόσο περισσότερο ακριβείς και αξιόπιστες γίνονται οι μέθοδοι που βασίζονται στην ομοιότητα. Επομένως, η τάση στη γονιδιακή πρόβλεψη έχει πρόσφατα μετατοπιστεί από τις στατιστικές προσεγγίσεις σε αλγόριθμους. βασισμένους στην ομοιότητα.

5.5.1 Στατιστικές προσεγγίσεις

Όπως αναφέραμε παραπάνω, οι στατιστικές προσεγγίσεις στην εύρεση γονίδιων βασίζονται στην ανίχνευση λεπτών στατιστικών διαφορών μεταξύ κωδικής (εξώνια) και μη κωδικής περιοχής. Ο απλούστερος τρόπος για την ανίχνευση πιθανών περιοχών κωδικοποίησης είναι να ψάξουμε σε ανοιχτά πλαίσια ανάγνωσης (open reading frames) ή ORFs. Μπορούμε να αναπαραστήσουμε ένα γονιδίωμα μήκους n ως μια ακολουθία από n/3 κωδικόνια. Τα τρία κωδικόνια τερματισμού, (TAA, TAG, και TGA) σπάνε αυτή την ακολουθία σε τμήματα, το καθένα ανάμεσα σε δύο διαδοχικά κωδικόνια τερματισμού. Τα υποτμήματα αυτών που ξεκινούν από ένα κωδικόνιο έναρξης, ATG, είναι ORFs. ORFs μέσα σε μία ενιαία ακολουθία γονιδιώματος μπορεί να επικαλύπτονται, δεδομένου ότι υπάρχουν έξι πιθανά "πλαίσια ανάγνωσης": τρία για το ένα από τα σκέλη που αρχίζει στις θέσεις 1, 2, και 3, και τρία για το αντίθετο σκέλος, όπως φαίνεται στην εικόνα 36.
Θα περίμενε κανείς να βρει συχνά κωδικόνια τερματισμού στο μη κωδικοποιήσιμο DNA, αφού ο μέσος αριθμός των κωδικόνιων μεταξύ δύο διαδοχικών κωδικόνιων τερματισμού σε τυχαίο DNA πρέπει να είναι 64/3= 21. Αυτό είναι πολύ μικρότερο από τον αριθμό των κωδικονίων σε μια μέση πρωτεΐνη, η οποία έχει περίπου 300. Επομένως, ORFs μετά από κάποιο μήκος δείχνουν ενδεχόμενα γονίδια. Ωστόσο, οι αλγόριθμοι γονιδιακής πρόβλεψης που βασίζονται στην επιλογή σημαντικά μεγάλων ORFs μπορεί να αποτύχουν να ανιχνεύσουν μικρά γονίδια ή γονίδια με μικρά εξώνια. Πολλοί στατιστικοί αλγόριθμοι γονιδιακής πρόβλεψης βασίζονται σε στατιστικά χαρακτηριστικά περιοχών που κωδικοποιούν πρωτεΐνη. Μπορούμε να εισάγουμε τη συχνότητα εμφάνισης κάθε κωδικονίου σε μια δεδομένη ακολουθία, στον πίνακα (όπως φαίνεται στην εικόνα) με όλα τα κωδικόνια που κωδικοποιούν κάποιο αμινοξύ (64 κωδικόνια).



Εικόνα 36 *: Τα έξι πλαίσια ανάγνωσης για την ακολουθία ATGCTTAGTCTG. Η συμβολοσειρά μπορεί να διαβάσει προς τα εμπρός ή προς τα πίσω, και υπάρχουν τρία πλαίσια για κάθε κατεύθυνση*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Για παράδειγμα, στα ανθρώπινα γονίδια τα κωδικόνια CGC και AGG κωδικοποιούν το ίδιο αμινοξύ (Arg), αλλά έχουν πολύ διαφορετικές συχνότητες: το CGC είναι 12 φορές πιο πιθανό να χρησιμοποιηθεί στα γονίδια από ότι το AGG. Επομένως, ένα ORF που «προτιμά» το CGC από το AGG και κωδικοποιεί το αμινοξύ Arg είναι ένα πιθανό γονίδιο. Μπορούμε να χρησιμοποιήσουμε την προσέγγιση του λόγου πιθανότητας για να υπολογίσουμε τις υπό συνθήκη πιθανότητες της ακολουθίας του DNA ενός πλαισίου, με την υπόθεση ότι το πλαίσιο περιέχει κωδικοποιήσιμη αλληλουχία, και με την υπόθεση ότι το πλαίσιο περιέχει μη κωδικοποιήσιμη ακολουθία. Αν ολισθήσουμε αυτό το πλαίσιο κατά μήκος της γονιδιωματικής ακολουθίας DNA (και υπολογίσουμε το λόγο πιθανότητας σε κάθε σημείο), τα γονίδια αποτελούν τις κορυφές στη γραφική αναπαράσταση του λόγου πιθανότητας.
Ενώ οι παραπάνω μέθοδοι είναι επιτυχείς σε προκαρυωτικούς οργανισμούς, η εφαρμογή τους στους ευκαρυωτικούς περιπλέκεται από τη δομή εξωνίου-εσωνίου. Το μέσο μήκος των εξωνίων σε σπονδυλωτά είναι 130 νουκλεοτίδια. Εξώνια αυτού του μήκους είναι πολύ μικρά για να παράγουν αξιόπιστες κορυφές στη γραφική αναπαράσταση του λόγου πιθανότητας κατά την ανάλυση των ORFs, επειδή δεν διαφέρουν αρκετά από τυχαίες διακυμάνσεις που θα ανιχνευθούν. Πολλοί ερευνητές έχουν χρησιμοποιήσει μια πιο βιολογικά προσανατολισμένη προσέγγιση και έχουν προσπαθήσει να αναγνωρίσουν τις θέσεις των σημάτων διαχωρισμού σε τμήματα εξωνίων-εσωνίων . Υπάρχει μια (ασθενώς) διατηρούμενη ακολουθία των οκτώ νουκλεοτιδίων, στα όρια ενός εξωνίου και ενός εσωνίου (donor splice site) και μια ακολουθία των τεσσάρων νουκλεοτιδίων στα όρια ενός εσωνίου και ενός εξωνίου ((acceptor splice site). Ενώ τα προφίλ για τις περιοχές διαχωρισμού είναι ανεπιτυχή, οι προσεγγίσεις αυτές είχαν περιορισμένη επιτυχία και έχουν αντικατασταθεί από το μοντέλο των αλυσίδων Markov (HMM).

	U		C		A		G	
U	UUU Phe	57	UCU Ser	16	UAU Tyr	58	UGU Cys	45
	UUC Phe	43	UCC Ser	15	UAC Tyr	42	UGC Cys	55
	UUA Leu	13	UCA Ser	13	UAA Stp	62	UGA Stp	30
	UUG Leu	13	UCG Ser	15	UAG Stp	8	UGG Trp	100
C	CUU Leu	11	CCU Pro	17	CAU His	57	CGU Arg	37
	CUC Leu	10	CCC Pro	17	CAC His	43	CGC Arg	38
	CUA Leu	4	CCA Pro	20	CAA Gln	45	CGA Arg	7
	CUG Leu	49	CCG Pro	51	CAG Gln	66	CGG Arg	10
A	AUU Ile	50	ACU Thr	18	AAU Asn	46	AGU Ser	15
	AUC Ile	41	ACC Thr	42	AAC Asn	54	AGC Ser	26
	AUA Ile	9	ACA Thr	15	AAA Lys	75	AGA Arg	5
	AUG Met	100	ACG Thr	26	AAG Lys	25	AGG Arg	3
G	GUU Val	27	GCU Ala	17	GAU Asp	63	GGU Gly	34
	GUC Val	21	GCC Ala	27	GAC Asp	37	GGC Gly	39
	GUA Val	16	GCA Ala	22	GAA Glu	68	GGA Gly	12
	GUG Val	36	GCG Ala	34	GAG Glu	32	GGG Gly	15

Εικόνα 37*: Ο γενετικός κώδικας και η χρήση κωδικόνιων στο Homo sapiens.Το κωδικόνιο για μεθειονίνη, ή AUG δρα επίσης και ως κωδικόνιο έναρξης. Όλες οι πρωτεΐνες αρχίζουν με Met. Οι αριθμοί δίπλα σε κάθε κωδικόνιο δείχνουν τη συχνότητα εμφάνισης του εν λόγω κωδικονίου στην κωδικοποίηση αμινοξέων. Για παράδειγμα, για τη λυσίνη (Lys), το κωδικόνιο AAG παράγει το 25% αυτής, ενώ το κωδικόνιο AAG παράγει το 75%. Οι συχνότητες αυτές διαφέρουν μεταξύ των ειδών.*

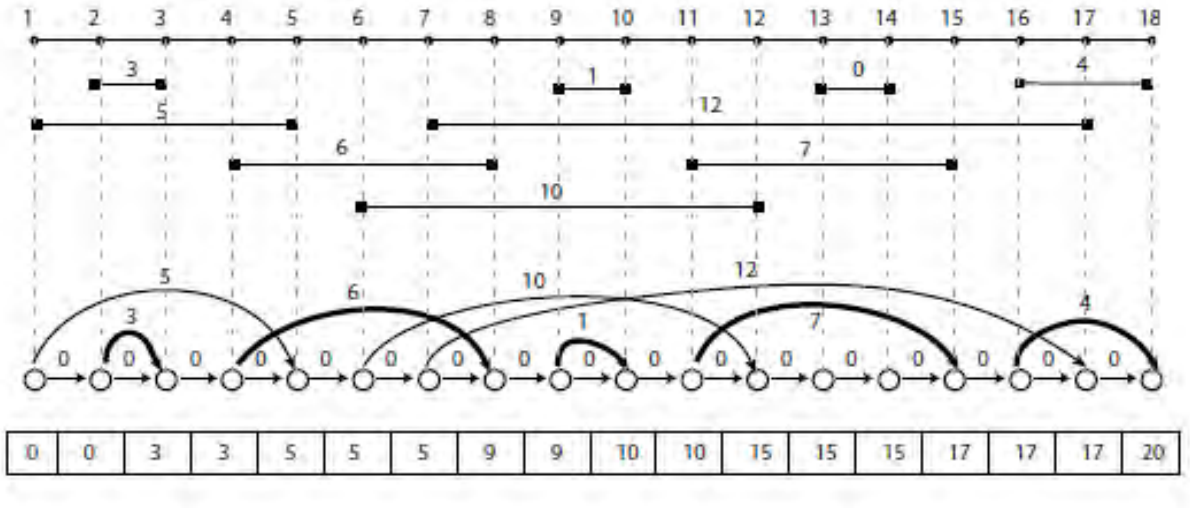
Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

5.5.2 Προσεγγίσεις βασισμένες στην ομοιότητα

Μια προσέγγιση που βασίζεται ομοιότητα, για την πρόβλεψη γονιδίων χρησιμοποιεί προηγούμενες αλληλουχίες γονιδίων και τα πρωτεϊνικά προϊόντα τους ως πρότυπα για την αναγνώριση άγνωστων γονίδιων. Αυτή η μέθοδος προσπαθεί να λύσει τον εξής συνδυαστικό γρίφο: με δεδομένη μια γνωστή πρωτεΐνη (στόχος) και μια γονιδιωματική ακολουθία, να βρούμε ένα σύνολο συμβολοσειρών (υποψήφια εξώνια) της γονιδιωματικής ακολουθίας των οποίων η συνένωσης (splicing) ταιριάζει καλύτερα στο στόχο.
Μια απλοϊκή προσέγγιση διεξοδικής αναζήτησης είναι να βρούμε όλες τις τοπικές ομοιότητες μεταξύ της γονιδιωματικής αλληλουχίας και της πρωτεΐνης-στόχου. Κάθε υποσυμβολοσειρά από την αλληλουχία του γονιδιώματος που παρουσιάζει επαρκή ομοιότητα με την πρωτεΐνη-στόχο θα μπορούσε να θεωρηθεί ως εξώνιο (putative exon). Έτσι τα εξώνια που επιλέγονται μπορεί να στερούνται των κανονικών εξώνιο-συνοδευτικών δινουκλεοτιδίων AG και GT αλλά μπορούμε να τα επεκτείνουμε ή να μειώσουμε ελαφρά ώστε να τα πλαισιώσουμε από τα AG και GT. Το αποτέλεσμα μπορεί να περιέχει επικαλυπτόμενες υποσυμβολοσειρές, και το πρόβλημα είναι να επιλέξουμε το καλύτερο υποσύνολο μη επικαλυπτόμενων υποσυμβολοσειρών.

Θα μοντελοποιήσουμε ένα υποθετικό εξώνιο (putative model), στη γονιδιωματική αλληλουχία, με ένα διάνυσμα βάρους, το οποίο περιγράφεται από τρεις παραμέτρους (l, r, w), όπως στην εικόνα 38.
Εδώ, το l είναι η αριστερή θέση, το r είναι η δεξιά θέση, και το w είναι το βάρος του εξωνίου. Το βάρος w μπορεί να αποτελεί είτε το σκορ της τοπικής ευθυγράμμισης του γονιδιακού διανύσματος με την ακολουθία πρωτεΐνης-στόχου είτε την ισχύ των συνοδευτικών acceptor και donor sites, ή οποιοδήποτε συνδυασμό αυτών. Μια αλυσίδα (chain) είναι κάθε σύνολο μη επικαλυπτόμενων βεβαρημένων διαστημάτων. Το συνολικό βάρος της αλυσίδας είναι το άθροισμα των βαρών των διαστημάτων στην αλυσίδα. Μεγαλύτερη αλυσίδα είναι η αλυσίδα με το μέγιστο συνολικό βάρος μεταξύ όλων των πιθανών αλυσίδων. Παρακάτω υποθέτουμε ότι τα βάρη όλων των διαστημάτων είναι θετικά (w> 0).
Το πρόβλημα εύρεσης της αλυσίδας εξωνίων για n διαστήματα μπορεί να λυθεί με έναν γράφο G δυναμικού προγραμματισμού 2n κορυφών, με n να εκπροσωπούν τις θέσεις έναρξης των διαστημάτων (l) και n τις θέσεις τερματισμού (r), όπως στην εικόνα 38. Υποθέτοντας ότι το σύνολο των διαστημάτων είναι ταξινομημένο σε αύξουσα σειρά και ότι όλες οι θέσεις είναι διαφορετικές, έχουμε διαταγμένες τις κορυφές (v_1, v_2, . . . v_{2n}) στο G.
Υπάρχουν 3n - 1 ακμές:

- μια ακμή ανάμεσα σε κάθε l_i και r_i βάρους w_i για i από 1 έως n
- και 2n - 1 επιπρόσθετες ακμές βάρους 0 που απλά συνδέουν γειτονικές κορυφές (v_i, v_{i+1}) σχηματίζοντας ένα μονοπάτι στο γράφημα από τη v_1 έως τη v_{2n}.



3

Εικόνα 38 *: Μια σύντομη «γονιδιωματική» αλληλουχία, ένα σύνολο από εννέα βεβαρημένα διαστήματα, και το διάγραμμα που χρησιμοποιείται για τη λύση δυναμικού προγραμματισμού του προβλήματος Exon Chaining. Φαίνονται με έντονες ακμές τα πέντε βεβαρημένα διαστήματα, (2, 3, 3), (4, 8, 6), (9, 10, 1), (11, 15, 7) και (16, 18, 4) που αποτελούν τη βέλτιστη λύση στο πρόβλημα. Ο πίνακας στο κάτω μέρος δείχνει τις τιμές S_1, S_2,. . . , S_{2n} που παράγονται από τον αλγόριθμο EXONCHANING.*

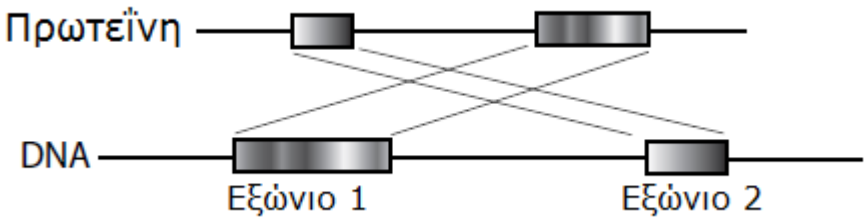
Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Στον παρακάτω αλγόριθμο το s_i αντιπροσωπεύει το μήκος του μεγαλύτερου μονοπατιού στο γράφημα που καταλήγει στην κορυφή v_i. Επομένως το s_{2n} αποτελεί τη λύση στο πρόβλημα εύρεσης αλυσίδας εξωνίων.

```
EXONCHAINING(G, n)
1  for i ← 1 to 2n
2      s_i ← 0
3  for i ← 1 to 2n
4      if vertex v_i in G corresponds to the right end of an interval I
5          j ← index of vertex for left end of the interval I
6          w ← weight of the interval I
7          s_i ← max{s_j + w, s_{i-1}}
8      else
9          s_i ← s_{i-1}
10 return s_{2n}
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Μια αδυναμία αυτής της προσέγγισης είναι ότι τα τελικά σημεία των υποθετικών εξωνίων (putative exons) δεν είναι πολύ καλά καθορισμένα, και αυτή η μέθοδος συναρμολόγησης δεν επιτρέπει οποιαδήποτε ευελιξία σε αυτά τα σημεία. Επιπλέον, η βέλτιστη αλυσίδα των διαστημάτων δεν μπορεί να αντιστοιχηθεί σε οποιαδήποτε έγκυρη ευθυγράμμιση. Για παράδειγμα, το πρώτο διάστημα στη βέλτιστη αλυσίδα μπορεί να είναι παρόμοιο με την κατάληξη της πρωτεΐνης, ενώ το δεύτερο διάστημα της βέλτιστης αλυσίδας μπορεί να είναι παρόμοιο με το πρόθεμα (την αρχή της πρωτεΐνης). Στην περίπτωση αυτή, τα putative exons των δύο διαστημάτων που αντιστοιχούν στην έγκυρη αλυσίδα δεν μπορούν να συνδυαστούν σε μια έγκυρη αλληλουχία ώστε να παράγουν την πρωτεΐνη-στόχο εικόνα 39.



Εικόνα 39 *: Μια ανέφικτη αλυσίδα που ενδέχεται να έχει τη μέγιστη βαθμολογία. Το πρώτο εξώνιο αντιστοιχεί σε μία περιοχή στο τέλος της πρωτεΐνης-στόχου, ενώ το δεύτερο εξώνιο αντιστοιχεί σε μια περιοχή στην αρχή της πρωτεΐνης-στόχου. Αυτά τα εξώνια δεν μπορεί να συνδυάζονται σε μια έγκυρη ευθυγράμμιση DNA-πρωτεΐνης.*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

6. Ομαδοποίηση και δένδρα

Ένα συνηθισμένο βιολογικό πρόβλημα είναι ο διαμελισμός ενός συνόλου πειραματικών δεδομένων σε ομάδες (clusters), με τέτοιο τρόπο ώστε τα δεδομένα εντός της ίδιας ομάδας να είναι παρόμοια σε μεγάλο βαθμό, ενώ τα δεδομένα διαφορετικών ομάδων να είναι τελείως διαφορετικά. Αυτό το πρόβλημα δεν είναι καθόλου απλό. Στο κεφάλαιο αυτό αναλύουμε διάφορους αλγόριθμους που εκτελούν διαφορετικούς τύπους ομαδοποίησης. Δεν υπάρχει απλή συνταγή για την επιλογή μιας συγκεκριμένης προσέγγισης αντί μιας άλλης, για ένα συγκεκριμένο πρόβλημα ομαδοποίησης, όπως ακριβώς δεν υπάρχει καθορισμένη έννοια του τι συνιστά την «καλή ομάδα». Παρ 'όλα αυτά, αυτοί οι αλγόριθμοι προσφέρουν σημαντική εικόνα για τα δεδομένα και επιτρέπουν σε κάποιον, για παράδειγμα, να εντοπίσει ομάδες γονιδίων με παρόμοιες λειτουργίες ακόμα και όταν δεν είναι σαφής ο ρόλος που διαδραματίζουν τα γονίδια αυτά. Επιπλέον, στο κεφάλαιο αυτό αναφέρουμε μελέτες ανοικοδόμησης εξελικτικών δέντρων που σχετίζονται με την ομαδοποίηση.

6.1 Ανάλυση της γονιδιακής έκφρασης

Η σύγκριση ακολουθιών βοηθά να ανακαλύψουμε την λειτουργία μιας νέας αλληλουχίας γονιδίου με την εύρεση ομοιοτήτων μεταξύ του νέου γονιδίου και αλληλουχιών γονιδίων με γνωστές λειτουργίες. Ωστόσο, για πολλά γονίδια, η ακολουθία τους δεν έχει μεγάλη ομοιότητα με μια οικογένεια λειτουργιών με αποτέλεσμα να μην μπορούμε αξιόπιστα να βρούμε τη λειτουργία της νέας αλληλουχίας γονιδίου βασιζόμενοι στην ακολουθία και μόνο. Επιπλέον, γονίδια που έχουν ίδια λειτουργία μερικές φορές δεν έχουν καμία ομοιότητα ακολουθιών. Επομένως, οι λειτουργίες πάνω από το 40% των γονιδίων είναι ακόμη άγνωστες.
Κατά την τελευταία δεκαετία, έχει προκύψει μια νέα προσέγγιση για την ανάλυση των λειτουργιών των γονιδίων. Συστοιχίες DNA μας επιτρέπουν να αναλύσουμε τα επίπεδα έκφρασης (ποσότητα του mRNA που παράγεται στο κύτταρο) πολλών γονιδίων σε διαφορετικές χρονικές στιγμές κάτω από διαφορετικές συνθήκες και να ανακαλύψουμε ποια γονίδια ενεργοποιούνται και απενεργοποιούνται στο κύτταρο. Το αποτέλεσμα αυτής της μελέτης είναι ένας πίνακας έκφρασης (expression matrix) n × m, με n σειρές που αντιστοιχούν σε γονίδια, και m στήλες που αντιστοιχούν σε διαφορετικές χρονικές στιγμές και διαφορετικές συνθήκες. Ο πίνακας έκφρασης I αναπαριστά τις εντάσεις των σημάτων υβριδισμού, που δημιουργούνται από μια συστοιχία DNA.
Το στοιχείο $I_{i,j}$ του πίνακα έκφρασης αντιπροσωπεύει το επίπεδο έκφρασης του γονιδίου i στο πείραμα j. Ολόκληρη η i-οστή γραμμή του πίνακα έκφρασης καλείται πρότυπο έκφρασης (expression pattern) του γονιδίου i. Μπορεί κανείς να αναζητήσει στον πίνακα έκφρασης ζεύγη γονιδίων με παρόμοιο πρότυπο έκφρασης, που θα εμφανιστούν ως δύο παρόμοιες σειρές. Επομένως, εάν τα πρότυπα έκφρασης δύο γονιδίων είναι παρόμοια, υπάρχει μεγάλη πιθανότητα αυτά τα γονίδια να σχετίζονται με κάποιο τρόπο, δηλαδή, είτε να επιτελούν παρόμοιες λειτουργίες, είτε να συμμετέχουν στην ίδια βιολογική διαδικασία. Κατά συνέπεια, αν το πρότυπο έκφρασης μιας νέας αλληλουχίας γονιδίου είναι παρόμοιο με το πρότυπο έκφρασης ενός γονιδίου με γνωστή λειτουργία, ένας βιολόγος μπορεί έχει

βάσιμες υποψίες ότι αυτά τα γονίδια εκτελούν όμοιες ή παρεμφερείς λειτουργίες. Μια άλλη σημαντική εφαρμογή της ανάλυσης της έκφρασης είναι η αποκρυπτογράφηση ρυθμιστικών μονοπατιών. Ωστόσο, η ανάλυση της έκφρασης θα πρέπει να γίνεται με προσοχή εφόσον οι συστοιχίες DNA παράγουν συνήθως δεδομένα με θόρυβο, δηλαδή, με υψηλά ποσοστά σφαλμάτων.

Οι αλγόριθμοι ομαδοποίησης κατηγοριοποιούν τα γονίδια με παρόμοια πρότυπα έκφρασης σε ομάδες, με την ελπίδα ότι αυτές οι ομάδες αντιστοιχούν σε ομάδες λειτουργικά σχετιζόμενων γονιδίων. Για να συλλέξουμε τα δεδομένα έκφρασης, ο n × m πίνακας έκφρασης μετατρέπεται σε ένα n × n πίνακα αποστάσεων **$d = (d_{i,j})$** όπου $d_{i,j}$ δείχνει πόσο όμοια είναι τα πρότυπα έκφρασης των γονιδίων i και j (εικόνα 40)

Ο στόχος της ομαδοποίησης είναι η κατηγοριοποίηση των γονιδίων σε ομάδες που πληρούν τις εξής δύο προϋποθέσεις:

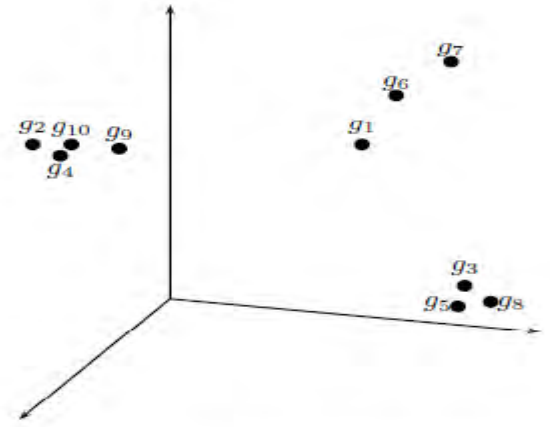
- **Ομοιογένεια**: Τα γονίδια (ή καλύτερα τα πρότυπα έκφρασής τους) μέσα σε μια ομάδα θα πρέπει να είναι παρόμοια μεταξύ τους. Δηλαδή, η $d_{i,j}$ θα πρέπει να είναι μικρή, αν τα i και j ανήκουν στην ίδια ομάδα
- **Διαχωρισμός**: Τα γονίδια από διαφορετικές ομάδες θα πρέπει να είναι πολύ διαφορετικά. Δηλαδή, η $d_{i,j}$ θα πρέπει να είναι μεγάλη, αν τα i και j ανήκουν σε διαφορετικές ομάδες.

Time	1 hr	2 hr	3 hr
g_1	10.0	8.0	10.0
g_2	10.0	0.0	9.0
g_3	4.0	8.5	3.0
g_4	9.5	0.5	8.5
g_5	4.5	8.5	2.5
g_6	10.5	9.0	12.0
g_7	5.0	8.5	11.0
g_8	2.7	8.7	2.0
g_9	9.7	2.0	9.0
g_{10}	10.2	1.0	9.2

(α) πίνακας I

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

(β) πίνακας αποστάσεων d

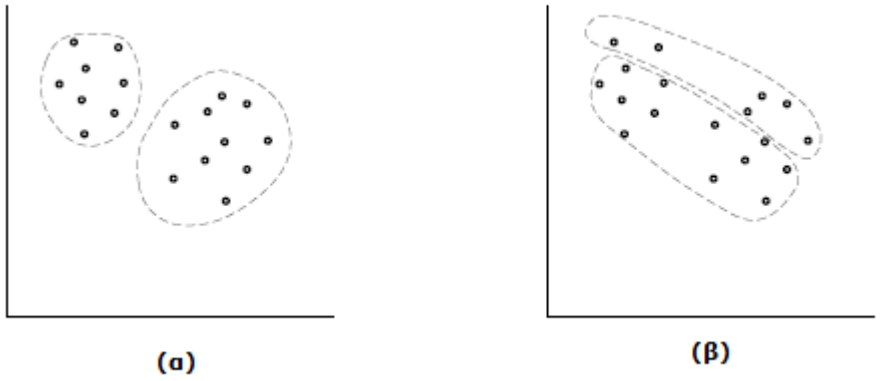


(γ) πρότυπα έκφρασης ως σημεία στον τρισδιάστατο χώρο

***Εικόνα 40** : Ένας πίνακας "έκφρασης" δέκα γονιδίων σε τρία χρονικά σημεία, και ο αντίστοιχος πίνακας αποστάσεων. Οι αποστάσεις υπολογίζονται ως η Ευκλείδεια απόσταση στον τρισδιάστατο χώρο.*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ένα παράδειγμα ομαδοποίησης φαίνεται στην εικόνα 41. Το (α) μέρος της εικόνας δείχνει μια καλή κατηγοριοποίηση σύμφωνα με τις παραπάνω δύο ιδιότητες, ενώ το (β) μια κακή. Οι αλγόριθμοι ομαδοποίησης προσπαθούν να βρουν έναν καλό διαχωρισμό. "Καλή" ομαδοποίηση δεδομένων είναι αυτή που πληροί τις παραπάνω προϋποθέσεις στόχους. Ενώ εμείς ελπίζουμε ότι μια καλύτερη ομαδοποίηση των γονιδίων οδηγεί σε μια καλύτερη ομαδοποίηση των γονιδίων σε επίπεδο λειτουργιών, η τελική ανάλυση των ομάδων που προκύπτουν αφήνεται στους βιολόγους.
Διαφορετικοί ιστοί εκφράζουν διαφορετικά γονίδια, και υπάρχουν συνήθως περισσότερα από 10.000 γονίδια που εκφράζονται σε έναν οποιοδήποτε ιστό. Δεδομένου ότι υπάρχουν περίπου 100 διαφορετικοί τύποι ιστών, και δεδομένου ότι τα επίπεδα έκφρασης μετριούνται σε πολλές χρονικές στιγμές, τα πειράματα γονιδιακής έκφρασης παράγουν τεράστιες ποσότητες δεδομένων που είναι δύσκολο να ερμηνευθούν. Επιδεινώνοντας αυτές τις δυσκολίες, τα επίπεδα έκφρασης των γονιδίων που σχετίζονται ποικίλουν κατά αρκετές τάξεις μεγέθους, δημιουργώντας έτσι το πρόβλημα επίτευξης ακρίβειας μετά από ένα μεγάλο εύρος μετρήσεων των επιπέδων έκφρασης (γονίδια με χαμηλά επίπεδα έκφρασης μπορεί να σχετίζονται με γονίδια με υψηλά επίπεδα έκφρασης).



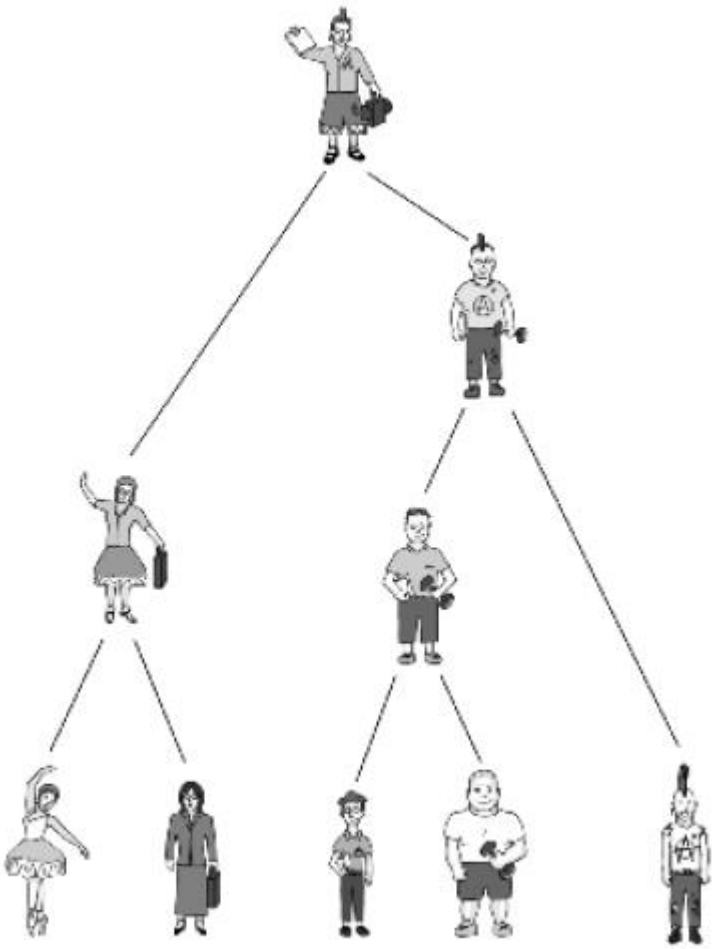
Εικόνα 41 *: Τα δεδομένα μπορούν να κατηγοριοποιηθούν σε ομάδες. Μερικές ομάδες είναι καλύτερες από άλλες: οι δύο ομάδες στο α) παρουσιάζουν καλή ομοιογένεια και διαχωρισμό, ενώ οι ομάδες στο β) όχι*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

6.2 Ιεραρχική ομαδοποίηση

Σε πολλές περιπτώσεις, οι ομάδες έχουν υποομάδες, αυτές έχουν άλλες υποομάδες και ούτω καθεξής. Για παράδειγμα, τα θηλαστικά μπορεί να χωριστούν σε πρωτεύοντα θηλαστικά, σαρκοφάγα ζώα, νυχτερίδες, μαρσιποφόρα, και πολλά άλλα. Τα σαρκοφάγα ζώα μπορούν να διαχωριστούν περαιτέρω σε γάτες, ύαινες, αρκούδες, φώκιες, και πολλά άλλα. Τέλος, οι γάτες μπορούν να διαχωριστούν σε τριάντα επτά είδη: λιοντάρια, τίγρεις, λεοπαρδάλεις, τζάγκουαρ, λύγκας, τσιτάχ, πούμα και πολλά άλλα.

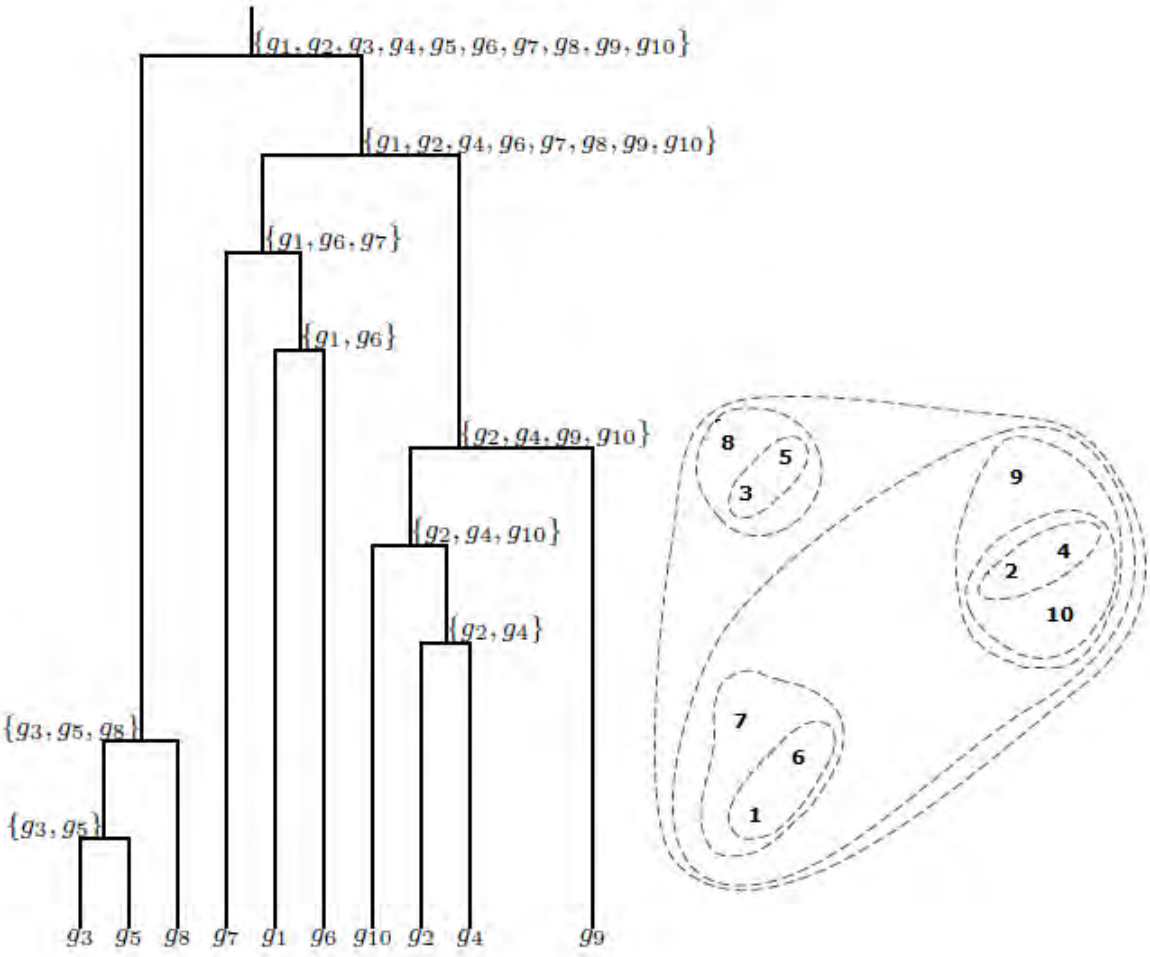
Η Ιεραρχική ομαδοποίηση (εικόνα 42) είναι μια τεχνική που οργανώνει τα στοιχεία σε ένα δέντρο, αντί να διαμορφώνει μια ρητή στεγανοποίηση των στοιχείων σε ομάδες. Στην περίπτωση αυτή, τα γονίδια αναπαρίστανται ως τα φύλλα ενός δέντρου. Στις ακμές των δέντρων έχουν ανατεθεί μήκη και οι αποστάσεις μεταξύ των φύλλων, δηλαδή το μήκος της διαδρομής στο δέντρο που συνδέει δύο φύλλα, συσχετίζονται με τις καταχωρήσεις στον πίνακα αποστάσεων. Τέτοια δέντρα χρησιμοποιούνται τόσο στην ανάλυση της έκφρασης όσο και σε μελέτες της μοριακής εξέλιξης.



***Εικόνα 42** : Σχηματική αναπαράσταση της ιεραρχικής ταξινόμησης*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η εικόνα 43 δείχνει ένα δέντρο που αντιπροσωπεύει την ομαδοποίηση των δεδομένων της εικόνας 40. Αυτό το δέντρο στην πραγματικότητα περιγράφει μια οικογένεια διαφορετικών ομάδων που η καθεμία έχει διαφορετικό αριθμό υποομάδων. Μπορούμε να δούμε ότι αυτές οι κατατμήσεις σχεδιάζονται με μια οριζόντια γραμμή στο δέντρο. Κάθε τέτοια γραμμή διασταυρώνεται με το δέντρο στο σημείο i (1<= i<= k) και αντιστοιχεί στην ομάδα i.



Εικόνα 43 : *Ιεραρχική ομαδοποίηση των δεδομένων της εικόνας 40*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ο αλγόριθμος HIERARCHICALCLUSTERING παίρνει ως είσοδο έναν πίνακα αποστάσεων d n×n, και σταδιακά παράγει ως έξοδο n διαφορετικές κατατμήσεις των δεδομένων όπως το δέντρο. Η μεγαλύτερη ομάδα έχει n υποομάδες του ενός στοιχείου. Η δεύτερη μεγαλύτερη ομάδα συνδυάζει τις δύο πιο κοντινές υποομάδες της μεγαλύτερης, και έτσι έχει n - 1 υποομάδες. Σε γενικές γραμμές, η i-οστή ομάδα συνδυάζει τις δύο πιο κοντινές υποομάδες από την (i - 1) υποομάδα και έχει n - i + 1 υποομάδες.

```
HIERARCHICALCLUSTERING(d, n)
1  Form n clusters, each with 1 element
2  Construct a graph T by assigning an isolated vertex to each cluster
3  while there is more than 1 cluster
4      Find the two closest clusters C1 and C2
5      Merge C1 and C2 into new cluster C with |C1| + |C2| elements
6      Compute distance from C to all other clusters
7      Add a new vertex C to T and connect to vertices C1 and C2
8      Remove rows and columns of d corresponding to C1 and C2
9      Add a row and column to d for the new cluster C
10 return T
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η γραμμή 6 στον αλγόριθμο είναι διφορούμενη καθώς οι αλγόριθμοι ομαδοποίησης ποικίλουν ως προς τον τρόπο υπολογισμού της απόστασης μεταξύ της νεοσχηματισμένης ομάδας με κάθε άλλη ομάδα. Διαφορετικοί τύποι για τον υπολογισμό των αποστάσεων δίνουν διαφορετικές απαντήσεις για τον ίδιο αλγόριθμο ιεραρχικής ομαδοποίησης. Για παράδειγμα, μπορούμε να ορίσουμε την απόσταση ανάμεσα σε δύο ομάδες ως τη μικρότερη απόσταση μεταξύ οποιουδήποτε ζεύγους στοιχείων τους:

$$d_{min}(C^*, C) = \min_{x \in C^*, y \in C} d(x, y)$$

ή ως τη μέση απόσταση μεταξύ των στοιχείων τους

$$d_{avg}(C^*, C) = \frac{1}{|C^*||C|} \sum_{x \in C^*, y \in C} d(x, y).$$

Μια άλλη συνάρτηση εκτιμά την απόσταση βασιζόμενη στο διαχωρισμό του C_1 και C_2 στον αλγόριθμο HIERARCHICALCLUSTERING:

$$d(C^*, C) = \frac{d(C^*, C_1) + d(C^*, C_2) - d(C_1, C_2)}{2}$$

Σε μια από τις πρώτες μελέτες για την ανάλυση της έκφρασης, ο Michael Eisen και οι συνεργάτες του χρησιμοποίησαν την ιεραρχική ομαδοποίηση για την ανάλυση των προφίλ έκφρασης 8600 γονιδίων σε δεκατρείς χρονικές στιγμές για να βρουν τα γονίδια που είναι υπεύθυνα για την αίσθηση της πείνας ενός ανθρώπου. Το αποτέλεσμα του HIERARCHICALCLUSTERING ήταν ένα δέντρο που αποτελούταν από πέντε βασικά υποδέντρα και πολλά μικρότερα υποδέντρα. Τα γονίδια σε αυτές τις πέντε ομάδες είχαν παρόμοιες λειτουργίες, επιβεβαιώνοντας έτσι ότι οι προκύπτουσες ομάδες είναι βιολογικά λογικές.

6.3 Ομαδοποίηση κ-μέσων

Μια από τις πιο δημοφιλείς μεθόδους ομαδοποίησης των σημείων σε πολυδιάστατους χώρους ονομάζεται ομαδοποίηση κ-μέσων (k-means clustering). Δεδομένου ενός συνόλου n σημείων σε έναν m-διάστατο χώρο και έναν ακέραιο k, το πρόβλημα είναι να καθορίσουμε ένα σύνολο k σημείων , ή κέντρων, στο m-διάστατο χώρο που ελαχιστοποιούν τη συνάρτηση τετραγωνικού σφάλματος που ορίζεται παρακάτω. Δεδομένου ενός σημείου v και ενός συνόλου k κέντρων το $X = \{x_1, \ldots x_k\}$, ορίζει την απόσταση του v από τα κέντρα X, ως την απόσταση του v από το πλησιέστερο σημείο του X, δηλαδή ,

$$d(v, \mathcal{X}) = \min_{1 \le i \le k} d(v, x_i)$$

Υποθέτουμε ότι d (v, x_i) είναι η Ευκλείδεια απόσταση σε m διαστάσεις. Το τετραγωνικό σφάλμα ενός συνόλου n σημείων $V = \{v_1, \ldots v_n\}$, και ενός συνόλου k κέντρων $X = \{x_1, \ldots x_k\}$, ορίζεται ως η μέση τιμή των τετραγώνων των αποστάσεων κάθε σημείου από το πλησιέστερο κέντρο του

$$d(\mathcal{V}, \mathcal{X}) = \frac{\sum_{i=1}^{n} d(v_i, \mathcal{X})^2}{n}$$

Αν και η παραπάνω διατύπωση δεν ορίζει μια ρητή ομαδοποίηση n σημείων σε k συστάδες, μια ομαδοποίηση μπορεί να επιτευχθεί με απλή ανάθεση κάθε σημείου στο πλησιέστερο κέντρο του. Αν και το πρόβλημα ομαδοποίησης k-μέσων φαίνεται σχετικά απλό, δεν υπάρχουν αποτελεσματικοί (πολυωνυμικοί) αλγόριθμοι γνωστοί για αυτό. Ο αλγόριθμος ομαδοποίησης k-μέσων Lloyd k-means είναι ένας από τους πιο δημοφιλής που δημιουργεί συχνά καλές λύσεις όσον αφορά την ανάλυση της γονιδιακής έκφρασης. Ο αλγόριθμος Lloyd επιλέγει τυχαία μια αυθαίρετη κατάτμηση σημείων σε k συστάδες και προσπαθεί να βελτιώσει αυτόν το διαχωρισμό μετακινώντας κάποια σημεία μεταξύ των ομάδων. Στην αρχή μπορούμε να επιλέξουμε αυθαίρετα σημεία k σημεία ως «εκπροσώπους των ομάδων» Ο αλγόριθμος εκτελεί επαναληπτικά τις ακόλουθες δύο ενέργειες έως ότου είτε να συγκλίνει ή μέχρι οι διακυμάνσεις να γίνουν πολύ μικρές:

- Αναθέτει κάθε σημείο στην ομάδα C_i αντιστοιχίζοντας το με τον κοντινότερο εκπρόσωπο ομάδας x_i (όπου 1<= i<= k)

- Μετά την ανάθεση όλων των σημείων n, υπολογίζει τους νέους εκπροσώπους των ομάδων σύμφωνα με το κέντρο βάρους κάθε ομάδας, δηλαδή, ο νέος εκπρόσωπος για κάθε ομάδα C είναι ο

$$\frac{\sum_{v \in C} v}{|C|}$$

Ο αλγόριθμος Lloyd συχνά συγκλίνει σε ένα τοπικό όχι σε ένα ολικό ελάχιστο της συνάρτησης τετραγωνικού σφάλματος. Δυστυχώς, ενδιαφέρουσες συναρτήσεις, εκτός από αυτή του τετραγωνικού σφάλματος οδηγούν σε παρόμοια δύσκολα προβλήματα. Για παράδειγμα, η εύρεση μιας καλής ομαδοποίησης μπορεί να αρκετά δύσκολή, αν, αντί για το τετραγωνικό σφάλμα

$(\sum_{i=1}^{n} d(v_i, \mathcal{X})^2)$

προσπαθήσουμε να ελαχιστοποιήσουμε το k-median πρόβλημα

$\sum_{i=1}^{n} d(v_i, \mathcal{X})$

ή να μεγιστοποιήσουμε το k-center πρόβλημα

$\max_{1 \le i \le n} d(v_i, \mathcal{X})$

Παρατηρούμε ότι όλοι αυτοί οι τύποι κόστους της ομαδοποίησης τονίζουν τη συνθήκη της ομοιογένειας και λίγο ή πολύ αγνοούν τον άλλο σημαντικό στόχο της ομαδοποίησης, το διαχωρισμό. Επιπλέον, σε ορισμένες άτυχες περιπτώσεις του προβλήματος ομαδοποίησης k-μέσων, ο αλγόριθμος μπορεί να συγκλίνει σε ένα τοπικό ελάχιστο που είναι κακό σε σύγκριση με μια βέλτιστη λύση.

Αν και ο αλγόριθμος Lloyd είναι πολύ γρήγορος, μπορεί να αναδιατάσσει κάθε ομάδα σε κάθε επανάληψη. Μια πιο συντηρητική προσέγγιση είναι να μετακινεί μόνο ένα στοιχείο μεταξύ των ομάδων σε κάθε επανάληψη. Υποθέτουμε ότι κάθε κατάτμηση P n-στοιχείων σε ένα σύνολο k ομάδων έχει ένα κόστος ομαδοποίησης, που συμβολίζεται ως cost(P) και μετρά την ποιότητα της κατάτμησης P: όσο μικρότερο είναι το κόστος ομαδοποίησης της κατάτμησης, τόσο καλύτερη είναι η ομαδοποίηση. Το τετραγωνικό σφάλμα είναι μια συγκεκριμένη επιλογή του cost(P) και υποθέτει ότι κάθε κέντρο αποτελεί το κέντρο βάρους της ομάδας του. Ο παρακάτω ψευδοκώδικας υποθέτει ότι το cost(P) μπορεί αποτελεσματικά να υπολογιστεί είτε με βάση τον πίνακα αποστάσεων είτε τον πίνακα έκφρασης. Δεδομένης μιας κατάτμησης P, μίας ομάδα C μέσα σε αυτή την κατάτμηση και ενός στοιχείου i έξω από τη C, το $P_{i \to C}$ ορίζει τον διαχωρισμό που προκύπτει από το P μετακινώντας το στοιχείο i από αυτό στην ομάδα C. Αυτή η κίνηση βελτιώνει το κόστος της ομαδοποίησης μόνο αν

$$\Delta(i \to C) = cost(P) - cost(P_{i \to C}) > 0,$$

και ο αλγόριθμος PROGRESSIVEGREEDYK-MEANS αναζητά την "καλύτερη" κίνηση σε κάθε βήμα (δηλαδή, μια κίνηση που μεγιστοποιεί το $\Delta(i \to C)$ για όλα τα C και για όλα τα i που ανήκουν στο C)

```
PROGRESSIVEGREEDYK-MEANS(k)
 1  Select an arbitrary partition P into k clusters.
 2  while forever
 3      bestChange ← 0
 4      for every cluster C
 5          for every element i ∉ C
 6              if moving i to cluster C reduces the clustering cost
 7                  if Δ(i → C) > bestChange
 8                      bestChange ← Δ(i → C)
 9                      i* ← i
10                      C* ← C
11      if bestChange > 0
12          change partition P by moving i* to C*
13      else
14          return P
```

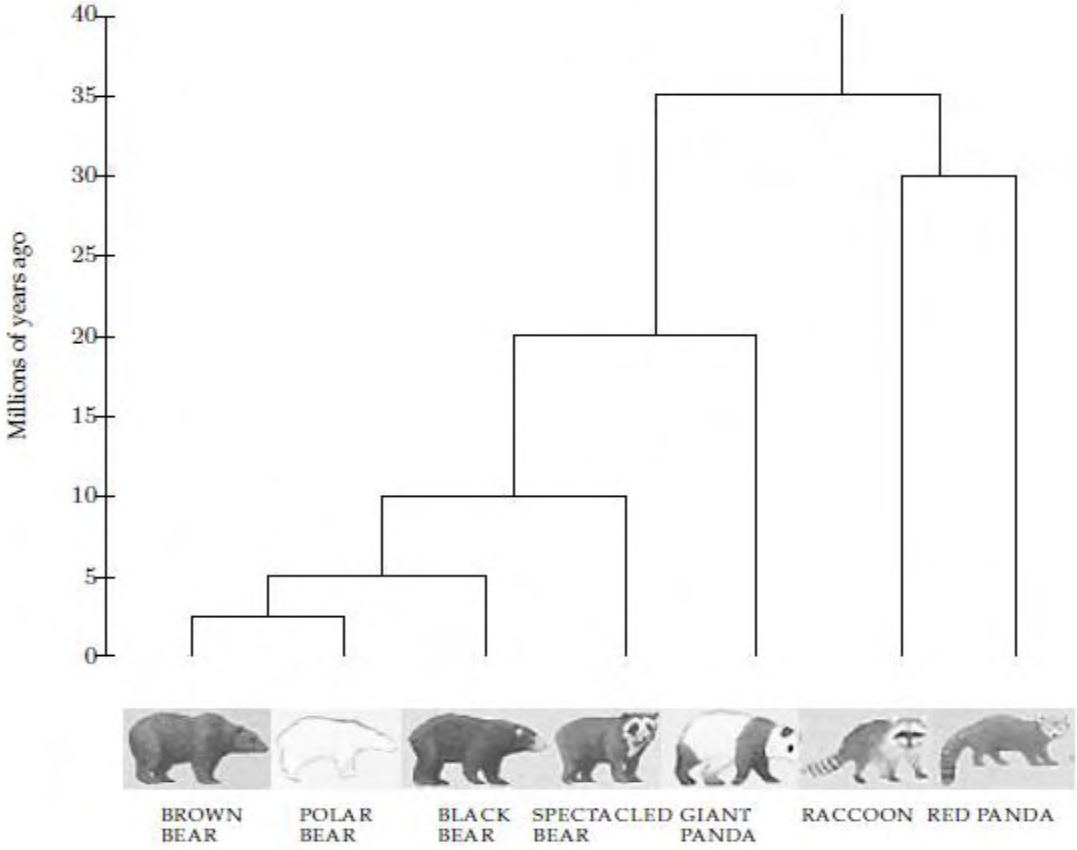
Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Παρόλο που η $2^{η}$ γραμμή δημιουργεί την εντύπωση ότι αυτός ο αλγόριθμος πέφτει σε ατέρμονα βρόχο η εντολή return της γραμμής 14 μας σώζει από μια απείρως μεγάλη αναμονή. Σταματάμε την επανάληψη όταν καμία μετακίνηση δεν έχει βελτίωση στη βαθμολογία.

6.4 Εξελικτικά δένδρα

Στο παρελθόν, οι βιολόγοι επικαλούνταν μορφολογικά χαρακτηριστικά, όπως το σχήμα του ράμφους ή την παρουσία ή την απουσία πτερυγίων για την κατασκευή εξελικτικών δέντρων. Σήμερα, βασίζονται στις ακολουθίες DNA για την ανασυγκρότηση των εξελικτικών δέντρων. Η εικόνα 44 παρουσιάζει το εξελικτικό δέντρο με βάση το DNA της αρκούδας και του ρακούν που βοήθησε τους βιολόγους να αποφασίσουν αν το γιγάντιο πάντα ανήκει στην οικογένεια της αρκούδας ή στην οικογένεια του ρακούν. Αυτή η ερώτηση δεν είναι τόσο προφανής όσο μπορεί να ακούγεται στην αρχή. Παρόλο που οι αρκούδες εμφανίστηκαν 35 εκατομμύρια χρόνια πριν τα ρακούν έχουν πολλά κοινά μορφολογικά χαρακτηριστικά.
Για πάνω από εκατό χρόνια οι βιολόγοι δεν μπορούσαν να συμφωνήσουν αν το γιγάντιο πάντα θα πρέπει να ταξινομηθεί στην οικογένεια της αρκούδας ή στην οικογένεια του ρακούν. Το 1870 ο ερασιτέχνης Père Armand David, επέστρεψε στο Παρίσι από την Κίνα με τα οστά ενός μυστήριου πλάσματος το οποίο ονόμασε απλά «μαύρη και άσπρη αρκούδα». Βιολόγοι εξέτασαν τα οστά και κατέληξαν στο συμπέρασμα ότι περισσότερο έμοιαζαν με τα οστά του κόκκινου πάντα παρά με εκείνα της αρκούδας. Εφόσον τα κόκκινα πάντα ήταν,

πέρα από κάθε αμφιβολία, μέρος της οικογένειας των ρακούν, το γιγάντιο πάντα κατηγοριοποιήθηκε κι αυτό ως ρακούν.
Παρόλο που το γιγάντιο πάντα μοιάζει με τις αρκούδες, έχει χαρακτηριστικά που είναι ασυνήθιστα για τις αρκούδες και τυπικά για τα ρακούν: δεν πέφτει σε χειμερία νάρκη όπως κάνουν οι άλλες αρκούδες, τα γεννητικά όργανα των αρσενικών είναι μικροσκοπικά και έχουν κατεύθυνση προς τα πίσω όπως των ρακούν και δεν βρυχάται σαν τις αρκούδες αλλά βελάζει όπως τα ρακούν.
Η κατηγοριοποίηση τελικά του γιγαντιαίου πάντα επιλύθηκε τελικά το 1985 από τον Steven O'Brien και τους συνεργάτες του οι οποίοι χρησιμοποίησαν, για την επίλυση της διαμάχης, ακολουθίες DNA και αλγορίθμους και όχι συμπεριφορά και ανατομικά χαρακτηριστικά εικόνα 44. Η τελική ανάλυση έδειξε ότι οι αλληλουχίες του DNA παρέχουν σημαντική πηγή πληροφοριών για τη δοκιμή εξελικτικών υποθέσεων. O'Brien στη μελέτη χρησιμοποίησε περίπου 500.000 νουκλεοτίδια, προκειμένου να κατασκευάσει το εξελικτικό δέντρο των αρκούδων και των ρακούν.
Περίπου την ίδια στιγμή που ο Steven O'Brien έλυσε τη διαμάχη για το γιγάντιο πάντα, οι Rebecca Cann, Mark Stoneking και Allan Wilson κατασκεύασαν ένα εξελικτικό δέντρο των ανθρώπων και αμέσως δημιουργήθηκε νέα διαφωνία. Αυτό το δέντρο οδήγησε στην Out of Africa υπόθεση, η οποία ισχυρίζεται ότι οι άνθρωποι έχουν ένα κοινό πρόγονο που έζησε στην Αφρική πριν από 200.000 χρόνια. Η μελέτη αυτή μετέτρεψε το θέμα της ανθρώπινης προέλευσης σε ένα αλγοριθμικό πάζλ.
Το δέντρο κατασκευάστηκε από ακολουθίες μιτοχονδριακού DNA (mtDNA) ανθρώπων από διαφορετικές φυλές και εθνικότητες. Ο Wilson και οι συνεργάτες του σύγκριναν ακολουθίες μιτοχονδριακού DNA από άτομα που προέρχονταν από τις χώρες της Αφρικής, της Ασίας, της Αυστραλίας, της Νέας Γουινέας και από τα Καυκάσια όρη και βρήκαν 133 παραλλαγές του mtDNA. Στη συνέχεια, κατασκεύασαν το εξελικτικό δέντρο για αυτές τις αλληλουχίες DNA που έδειξε το διαχωρισμό του κορμού σε δύο κύρια κλαδιά. Το ένα κλαδί αποτελούνταν μόνο από Αφρικανούς και το άλλο περιελάμβανε μερικούς σύγχρονους Αφρικανούς και μερικούς ανθρώπους από οπουδήποτε αλλού. Κατέληξαν στο συμπέρασμα ότι ο πληθυσμός της Αφρικής, οι πρώτοι σύγχρονοι άνθρωποι, αποτελεί τον κορμό και το πρώτο κλαδί του δέντρου, ενώ το δεύτερο κλαδί αντιπροσωπεύει μια υποομάδα που άφησε την Αφρική και αργότερα απλώθηκε στο υπόλοιπο κόσμο. Όλα τα μιτοχονδριακά DNA, ακόμη και τα δείγματα από τις περιοχές του κόσμου μακριά από την Αφρική, ήταν εντυπωσιακά παρόμοια. Αυτό υποδηλώνει ότι το είδος μας είναι σχετικά νέο. Όμως τα αφρικανικά δείγματα είχαν τις περισσότερες μεταλλάξεις, υπονοώντας έτσι ότι η Αφρικανική καταγωγή είναι η παλαιότερη και ότι όλοι οι σύγχρονοι άνθρωποι εντοπίζουν τις ρίζες τους πίσω στην Αφρική. Επιπλέον, εκτίμησαν ότι ο σύγχρονος άνθρωπος προέρχεται από την Αφρική, 200.000 χρόνια πριν με τις φυλετικές διαφορές να προκύπτουν μόλις 50.000 χρόνια πριν.

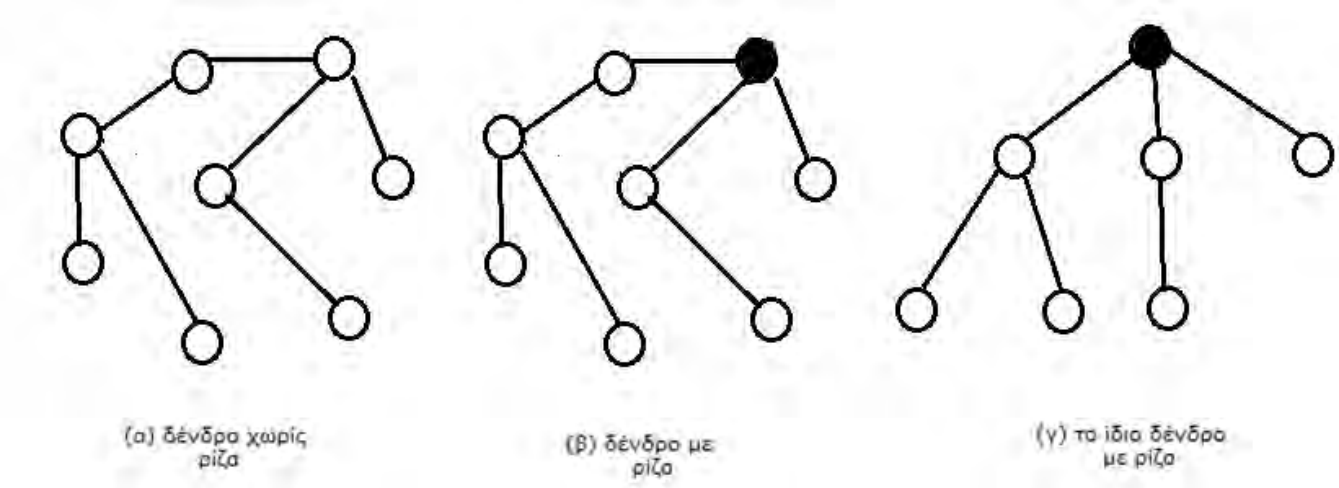


Εικόνα 44 : *Ένα εξελικτικό δέντρο που δείχνει τον διαχωρισμό των ρακούν και των αρκούδων. Παρά τη διαφορά τους στο μέγεθος και το σχήμα, αυτές οι δύο οικογένειες σχετίζονται στενά.*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Λίγο μετά τον AllanWilson και τους συνεργάτες του που κατασκεύασαν το ανθρώπινο εξελικτικό δέντρο με βάση το mtDNA υποστηρίζοντας την υπόθεση Out of Africa, ο Alan Templeton κατασκεύασε 100 διαφορετικά δέντρα που ήταν επίσης συνεπή με τα δεδομένα και παρείχε αποδεικτικά στοιχεία κατά της υπόθεσης της αφρικανικής προέλευσης.
Οι βιολόγοι χρησιμοποιούν εξελικτικά δέντρα είτε με ρίζα είτε χωρίς. Η διαφορά μεταξύ τους φαίνεται στην εικόνα 45. Σε ένα εξελικτικό δέντρο με ρίζα, η ρίζα αντιστοιχεί στο αρχαιότερο πρόγονο στο δέντρο, και το μονοπάτι από τη ρίζα σε ένα φύλλο ονομάζεται εξελικτικό μονοπάτι. Τα φύλλα των εξελικτικών δέντρων αντιστοιχούν στα υπάρχοντα είδη ενώ οι εσωτερικές κορυφές αντιστοιχούν σε υποθετικά προγονικά είδη. Στην περίπτωση που δεν υπάρχει ρίζα, δεν κάνουμε καμία υπόθεση σχετικά με τη θέση του εξελικτικού

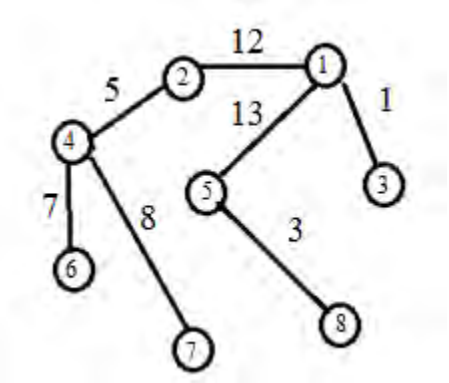
προγόνου στο δέντρο. Έχουμε, επίσης, παρατηρήσει ότι τα δένδρα με ρίζες (που επίσημα ορίζονται ως μη κατευθυνόμενα γραφήματα) μπορούν να αντιμετωπισθούν ως κατευθυνόμενα γραφήματα αν κατευθύνουμε τις ακμές από τη ρίζα προς τα φύλλα.
Οι βιολόγοι συχνά λειτουργούν με δυαδικά βεβαρημένα δένδρα όπου κάθε εσωτερική κορυφή έχει βαθμό ίσο με 3 και σε κάθε ακμή έχει ανατεθεί ένα θετικό βάρος (μερικές φορές αναφέρεται ως το μήκος). Το βάρος μιας ακμής (v, w) αποτελεί το αριθμό των μεταλλάξεων στην εξελικτική πορεία από το v στο w ή την εκτίμηση του χρόνου για την εξέλιξη των ειδών από το v στο w. Μερικές φορές υποθέτουμε την ύπαρξη ενός μοριακού ρολογιού που αποδίδει ένα χρόνο t(v) σε κάθε εσωτερική κορυφή v του δέντρου και μήκος σε μια ακμή (v, w) t(w) - t(v). Εδώ, ο χρόνος αντιστοιχεί στην «στιγμή», που το είδος v παράγει τους απογόνους του. Κάθε είδος-φύλλο αντιστοιχεί στη στιγμή 0 και κάθε εσωτερική κορυφή αντιστοιχεί προφανώς σε αρνητικές στιγμές.



***Εικόνα 45** : Η διαφορά μεταξύ (α) δέντρου χωρίς ρίζα και (β) δέντρου με ρίζα. Και οι δύο απεικονίσεις περιγράφουν το ίδιο δέντρο, αλλά το δέντρο α δεν κάνει καμία υπόθεση για την προέλευση των ειδών. Τα δένδρα με ρίζες συχνά αναπαρίσταται με τη ρίζα κορυφή στην κορυφή (γ), τονίζοντας ότι η ρίζα αντιστοιχεί στα προγονικά είδη.*

6.5 Ανασυγκρότηση δέντρου με βάση την απόσταση

Αν μας δίνεται ένα βεβαρημένο δέντρο Τ με n φύλλα, μπορούμε να υπολογίσουμε το μήκος του μονοπατιού $d_{i,j}$ (Τ) μεταξύ δύο φύλλων i και j (εικόνα 46). Οι βιολόγοι συχνά αντιμετωπίζουν το αντίθετο πρόβλημα: υπολογίζουν τον n × n πίνακα αποστάσεων ($D_{i,j}$) και, στη συνέχεια πρέπει να αναζητήσουν ένα δέντρο Τ που έχει n φύλλα και ταιριάζει με τα δεδομένα, τέτοιο ώστε να είναι $d_{i,j}$ (Τ)= $D_{i,j}$ για κάθε δύο φύλλα i και j.



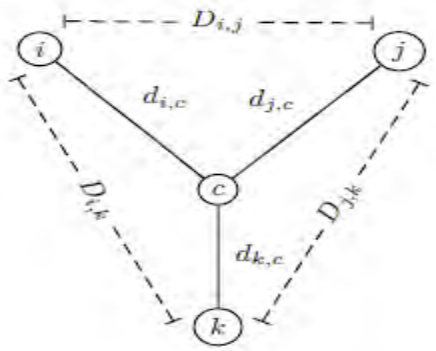
Εικόνα 46 : *Βεβαρημένο δέντρο χωρίς ρίζα. Το μήκος της διαδρομής μεταξύ δύο κορυφών μπορεί να υπολογιστεί ως το άθροισμα των βαρών των ακμών στο μεταξύ τους μονοπάτι. Για παράδειγμα, δ (1, 6) = 12 + 5 + 7 =24*

Δεν είναι δύσκολο να κατασκευάσουμε ένα δέντρο που ταιριάζει σε οποιοδήποτε δεδομένο πίνακα D 3 × 3. Αυτό το δυαδικό χωρίς ρίζα δέντρο έχει τέσσερις κορυφές - φύλλα i, j, k και την κορυφή c ως το κέντρο. Το μήκος της κάθε ακμής στο δέντρο ορίζεται από τις ακόλουθες τρεις εξισώσεις χρησιμοποιώντας τρεις μεταβλητές $d_{i,c}$, $d_{o,k}$ και $d_{o,c}$ (εικόνα 47):

$$d_{i,c} + d_{j,c} = D_{i,j} \qquad d_{i,c} + d_{k,c} = D_{i,k} \qquad d_{j,c} + d_{k,c} = D_{j,k}.$$

Η λύση δίνεται από τη σχέση

$$d_{i,c} = \frac{D_{i,j} + D_{i,k} - D_{j,k}}{2} d_{j,c} = \frac{D_{j,i} + D_{j,k} - D_{i,k}}{2} d_{k,c} = \frac{D_{k,i} + D_{k,j} - D_{i,j}}{2}.$$

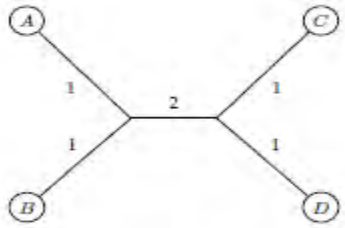


***Εικόνα 47** : Δέντρο με 3 φύλλα*

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ένα δυαδικό δέντρο χωρίς ρίζα με n φύλλα έχει 2n - 3 ακμές. Έτσι το ταίριασμα ενός δοσμένου δέντρου με έναν πίνακα αποστάσεων D n × n οδηγεί στην επίλυση ενός συστήματος $\binom{n}{2}$ εξισώσεων με 2n − 3 μεταβλητές. Για n = 4 έχουμε την επίλυση έξι εξισώσεων με μόνο πέντε μεταβλητές. Φυσικά, δεν είναι πάντα δυνατό να λυθεί αυτό το σύστημα καθιστώντας δύσκολο ή αδύνατο να κατασκευάσουμε ένα δέντρο από το D. Ένας πίνακας ($D_{i,j}$) καλείται πρόσθετος (additive) εάν υπάρχει δέντρο T με $d_{i,j}$ (T)= $D_{i,j}$, και μη πρόσθετος αλλιώς (εικόνα 48).
Το πρόβλημα της φυλογένεσης με βάση την απόσταση μπορεί να μην έχει λύση, αλλά αν έχει (δηλαδή αν ο D είναι πρόσθετος) υπάρχει ένας απλός αλγόριθμο για την επίλυσή του.

	A	B	C	D
A	0	2	4	4
B	2	0	4	4
C	4	4	0	2
D	4	4	2	0



(α) Πρόσθετος πίνακας και το αντίστοιχο δέντρο

	A	B	C	D
A	0	2	2	2
B	2	0	3	2
C	2	3	0	2
D	2	2	2	0

?

(β) μη πρόσθετος πίνακας

Εικόνα 48 : Πρόσθετοι και μη πρόσθετοι πίνακες

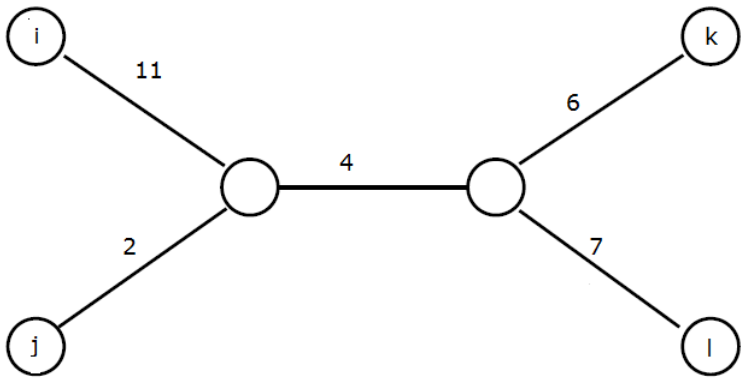
Πηγή: An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner

6.6 Ανοικοδόμηση δένδρων από πρόσθετους πίνακες

Ένας "απλός" τρόπος για να λύσουμε το πρόβλημα της φυλογένεσης με βάση την απόσταση από πρόσθετους πίνακες, είναι να βρούμε ένα ζευγάρι γειτονικών φύλλων, δηλαδή, φύλλων που έχουν γονέα την ίδια κορυφή. Η εικόνα 49 δείχνει ότι για ένα ζευγάρι γειτονικών φύλλων i και j και γονέα την κορυφή k, ισχύει η παρακάτω ισότητα για οποιοδήποτε άλλο φύλλο m του δέντρου:

$$D_{k,m} = \frac{D_{i,m} + D_{j,m} - D_{i,j}}{2} \quad (1)$$

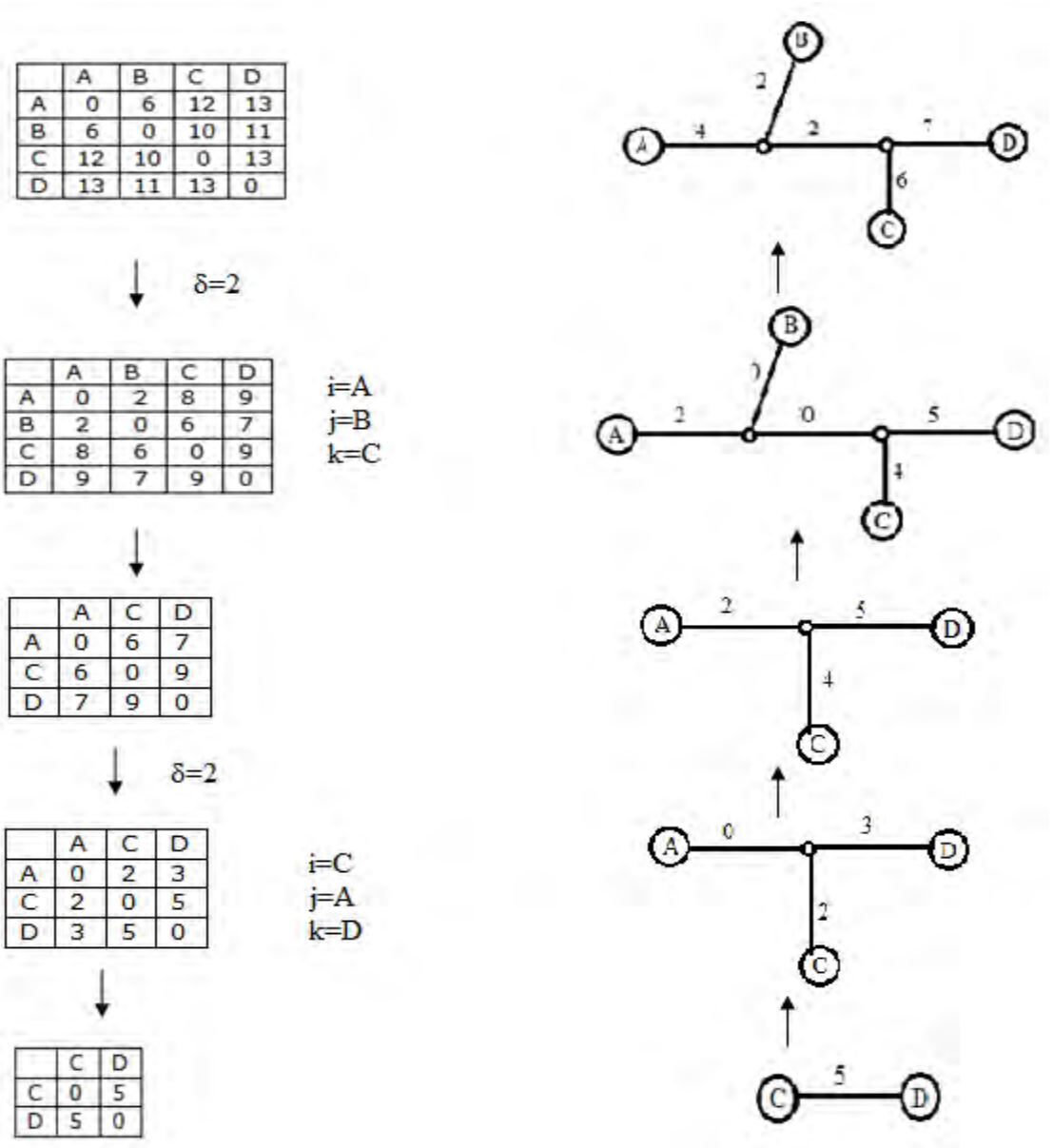
Επομένως, όταν βρούμε ένα ζευγάρι γειτονικών φύλλων i και j, μπορούμε να αφαιρέσουμε τις αντίστοιχες σειρές και στήλες i και j από τον πίνακα αποστάσεων και να προσθέσουμε νέα γραμμή και νέα στήλη που αντιστοιχεί στον κόμβο - γονέα k. Εφόσον ο πίνακας αποστάσεων είναι πρόσθετος, οι αποστάσεις του k από τα άλλα φύλλα υπολογίζονται από τον παραπάνω τύπο (1). Ο μετασχηματισμός αυτός οδηγεί σε έναν απλό αλγόριθμο που βρίσκει ένα ζευγάρι γειτονικών φύλλων και μειώνει το μέγεθος του δέντρου σε κάθε βήμα.
Το πρόβλημα με αυτή την προσέγγιση είναι ότι δεν είναι πολύ εύκολο να βρούμε γειτονικά φύλλα. Θα μπορούσε κανείς να μπει στον πειρασμό να σκεφτεί ότι το ζευγάρι φύλλων που βρίσκεται πιο κοντά (δηλαδή, τα φύλλα i και j, με την ελάχιστη απόσταση $D_{i,j}$) αποτελούν και ζευγάρι γειτονικών φύλλων, αλλά μια ματιά στην εικόνα 49 θα δείξει ότι αυτό δεν είναι αλήθεια.



***Εικόνα 49** : Αν i και j είναι γειτονικά φύλλα και k είναι ο γονέας τους τότε η απόσταση οποιασδήποτε άλλης κορυφής m στο δέντρο δίνεται από τον τύπο (1).*

***Πηγή:** An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Η εικόνα 50 δείχνει την διαδικασία της συντόμευσης όλων των ακμών του δέντρου Τ που "κρέμονται", δηλαδή, των ακμών που καταλήγουν σε φύλλα. Αν μειώσουμε το μήκος κάθε τέτοιας ακμής κατά την ίδια μικρή ποσότητα δ, τότε ο πίνακας αποστάσεων του δέντρου που προκύπτει ($d_{i,j}$ – 2δ) δεδομένου ότι η απόσταση μεταξύ δύο φύλλων μειώνεται κατά 2δ. Αργά ή γρήγορα αυτή η διαδικασία θα οδηγήσει σε «πτώση» κάποιου από τα φύλλα, όταν το μήκος της αντίστοιχης ακμής από την οποία «κρέμεται» γίνει ίσο με 0 (οπότε θέτουμε το δ ίσο με το μήκος της συντομότερης «κρεμάμενης» ακμής). Σε αυτό το σημείο, το αρχικό δέντρο Τ = T_n με n φύλλα θα μετατραπεί σε ένα δέντρο T_{n-1} με n - 1 ή λιγότερα φύλλα.



Εικόνα 50 *: Η επαναληπτική διαδικασία μείωσης των άκρων που «κρέμονται» σε ένα δέντρο.*

Παρόλο που ο πίνακας αποστάσεων D δεν περιέχει καμία ρητή ενημέρωση για το δ, είναι εύκολο να αντλήσουμε πληροφορία και για το δ και για τη θέση του φύλλου που «έπεσε». Επομένως, μπορούμε παραστήσουμε μια σειρά από μετασχηματισμούς δένδρων

$$T_n \to T_{n-1} \to \ldots \to T_3 \to T_2$$

να κατασκευάσουμε έτσι το δέντρο Τ2 (το οποίο είναι εύκολο, αφού αποτελείται από μία μόνο ακμή), και στη συνέχεια να εκτελέσουμε μια σειρά αντίστροφων μετασχηματισμών

$$T_2 \to T_3 \to \ldots \to T_{n-1} \to T_n$$

ανακτώντας την πληροφορία σχετικά με την ακμή που έπεσε σε κάθε βήμα (εικόνα 50)

Μια τριάδα διακριτών στοιχείων για τα οποία ισχύει:

$$1 \leq i, j, k \leq n$$

ονομάζεται «εκφυλισμένη» αν Di, j + Dj, k = Di, k , που ουσιαστικά είναι μόνο μια ένδειξη ότι η κορυφή j βρίσκεται στο μονοπάτι μεταξύ της i και της k στο δέντρο. Εάν ο D είναι πρόσθετος, ισχύει:

$$D_{i,j} + D_{j,k} \geq D_{i,k}$$

για κάθε τριάδα i, j, k στο δέντρο. Καλούμε τον πίνακα D εκφυλισμένο αν έχει τουλάχιστον μια εκφυλισμένη τριάδα. Σε περίπτωση που (i, j, k) είναι μια εκφυλισμένη τριάδα, και κάποιο δέντρο Τ εκπροσωπείται από τον πίνακα D, τότε η κορυφή j βρίσκεται στο μονοπάτι μεταξύ i και k του Τ. Ένας άλλος τρόπος να το πούμε αυτό είναι ότι η j συνδέεται με αυτό το μονοπάτι με μια ακμή βάρους 0 και το σημείο σύνδεσης βρίσκεται σε απόσταση $D_{i,j}$ από την κορυφή i. Ως εκ τούτου, εάν ένας n×n πρόσθετος πίνακας D έχει μια εκφυλισμένη τριάδα, μπορεί να μετατραπεί σε έναν (n - 1) × (n - 1) πρόσθετο πίνακα αφαιρώντας απλά την υπό εξέταση κορυφή j. Η θέση της j θα ανακτηθεί κατά τη διάρκεια των αντίστροφων μετασχηματισμών. Αν ο πίνακας D δεν έχει εκφυλισμένη τριάδα, μειώνουμε τις τιμές όλων των στοιχείων του D κατά την ίδια ποσότητα 2δ μέχρι το σημείο στο οποίο ο πίνακας αποστάσεων εκφυλιστεί για πρώτη φορά (δηλαδή, δ είναι η ελάχιστη τιμή για την οποία ο ($D_{i,j}$ -2δ) έχει μια εκφυλισμένη τριάδα για κάποια i και j). Η ενέργεια αυτή αντιστοιχεί σε μείωση κατά δ όλων των ακμών του Τ που «κρέμονται», έως ότου ένα από τα φύλλα καταλήξει στο εξελικτικό μονοπάτι μεταξύ δύο άλλων φύλλων για πρώτη φορά. Ο παρακάτω αναδρομικός αλγόριθμος βρίσκει το δέντρο που αντιστοιχεί στα δεδομένα.

```
ADDITIVEPHYLOGENY(D)
 1  if D is a 2 × 2 matrix
 2      T ← the tree consisting of a single edge of length D_{1,2}
 3      return T
 4  if D is non-degenerate
 5      δ ← trimming parameter of matrix D
 6      for all 1 ≤ i ≠ j ≤ n
 7          D_{i,j} ← D_{i,j} − 2δ
 8  else
 9      δ ← 0
10  Find a triple i, j, k in D such that D_{ij} + D_{jk} = D_{ik}
11  x ← D_{i,j}
12  Remove jth row and jth column from D.
13  T ← ADDITIVEPHYLOGENY(D)
14  Add a new vertex v to T at distance x from i to k
15  Add j back to T by creating an edge (v, j) of length 0
16  for every leaf l in T
17      if distance from l to v in the tree T does not equal D_{l,j}
18          output "Matrix D is not additive"
19          return
20  Extend hanging edges leading to all leaves by δ
21  return T
```

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

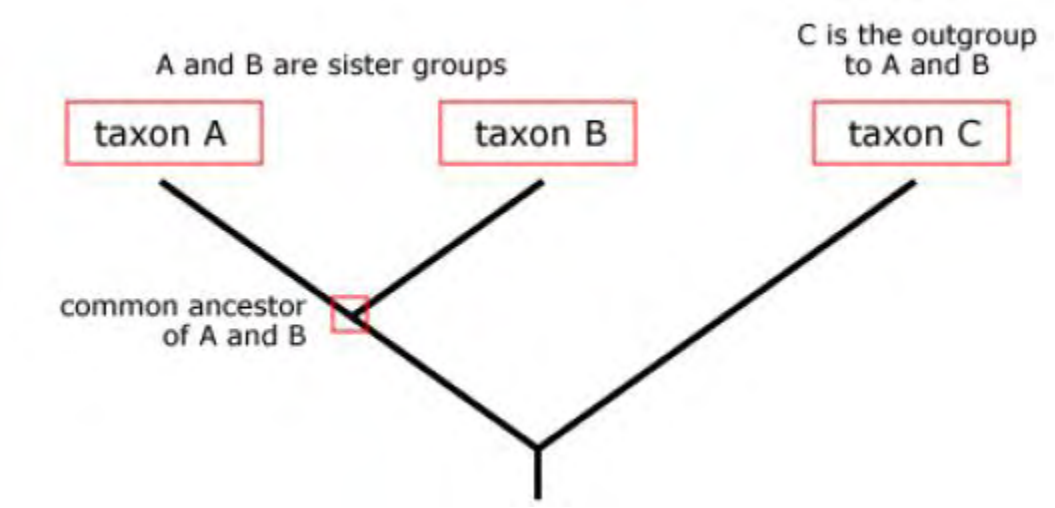
Ο παραπάνω αλγόριθμος ADDITIVEPHYLOGENY παρέχει έναν τρόπο για να ελέγξουμε αν ο πίνακας D είναι πρόσθετος. Αν και αυτός ο αλγόριθμος είναι απλός, δεν είναι ο πιο αποτελεσματικός τρόπος για να κατασκευάσουμε τα δέντρα. Ένας άλλος τρόπος είναι η χρήση της «συνθήκης των 4 σημείων» (four point condition). Έστω τέσσερα διαφορετικά i, j, k, l για τα οποία ισχύει 1<=i, j, k, l<= n. Υπολογίζουμε τα 3 αθροίσματα $D_{i,j} + D_{k,l}$, $D_{i,k} + D_{j,l}$ και $D_{i,l} + D_{j,k}$. Εάν ο D είναι ένας πρόσθετος πίνακας τότε αυτά τα τρία αθροίσματα μπορούν να παρουσιαστούν σε ένα δέντρο με τέσσερα φύλλα. Επιπλέον, δύο από αυτά τα αθροίσματα αντιπροσωπεύουν τον ίδιο αριθμό (το άθροισμα του μήκους όλων των ακμών στο δέντρο συν το μήκος της μεσαίας ακμής), ενώ το τρίτο άθροισμα αντιπροσωπεύει έναν άλλο μικρότερο αριθμό (το άθροισμα των μηκών όλων των ακμών στο δέντρο μείον το μήκος της μεσαίας ακμής). Λέμε ότι τα στοιχεία 1<=i, j, k, l<= n ικανοποιούν την four point condition αν τα δύο από τα αθροίσματα $D_{i,j} + D_{k,l}$, $D_{i,k} + D_{j,l}$ και $D_{i,l} + D_{j,k}$ είναι ίδια, και το τρίτο είναι μικρότερο από αυτά τα δύο.

Θεώρημα: Ένας n × n πίνακας D είναι πρόσθετος αν και μόνο αν η four point condition ισχύει για κάθε 4 διακριτά στοιχεία 1<=i, j, k, l<= n.

Εάν ο πίνακας αποστάσεων D δεν είναι πρόσθετος, μπορούμε να βρούμε το δέντρο που προσεγγίζει τον D χρησιμοποιώντας το άθροισμα των τετραγώνων των λαθών $\Sigma_{i,j}(d_{i,j}(T) - D_{i,j})^2$ ως μέτρο για την ποιότητα της προσέγγισης. Αυτό οδηγεί στο πρόβλημα φυλογένεσης ελαχίστων τετραγώνων με βάση την απόσταση

6.7 Φυλογενετικά δέντρα και ιεραρχική ομαδοποίηση

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 του δέντρου (tips) 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 συχνό είναι το φαινόμενο να υπάρχει ένα ή περισσότερα taxa που να έχει κοινο πρόγονο με κάποια sister groups αλλά να μην είναι το ίδιο sister group με τα υπόλοιπα (outgroup). Η παρακάτω εικόνα (εικόνα 51) δείχνει τις προαναφερθείσες έννοιες

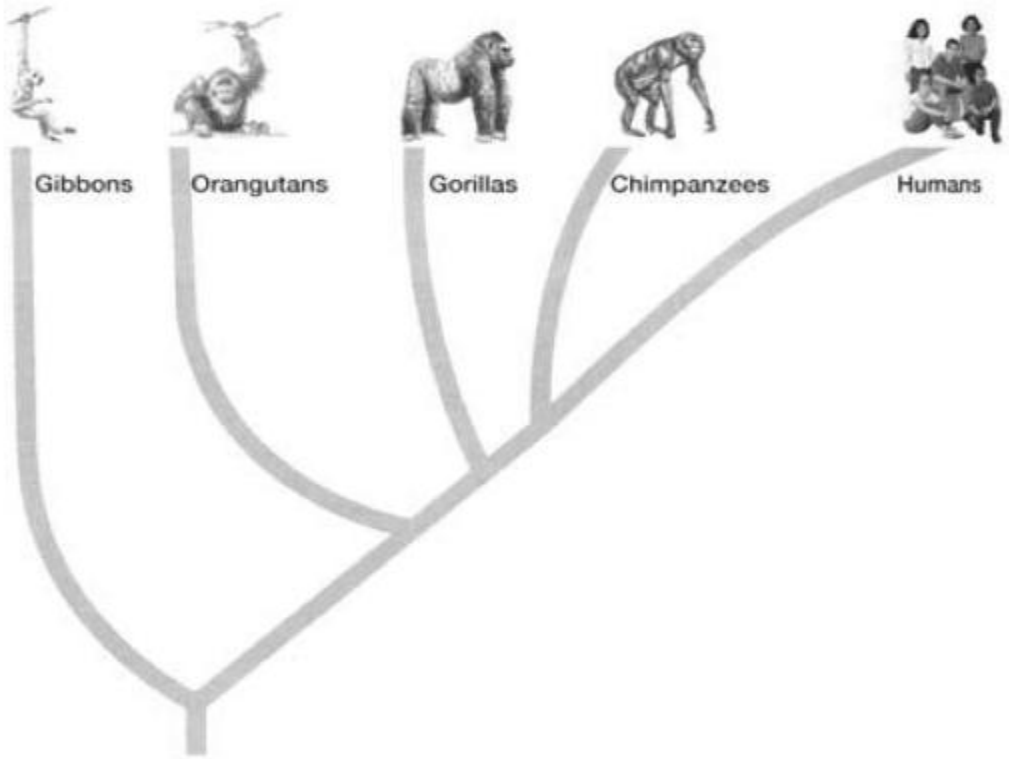


Εικόνα 51*: Ένα εξελικτικό δέντρο οπου φαίνονται οι έννοιες common ancestor, sister groups και outgroup*

Πηγή: http://evolution.berkeley.edu/evolibrary/article/phylogenetics_02

6.7.1 Τυποι Φυλογενετικών δέντρων

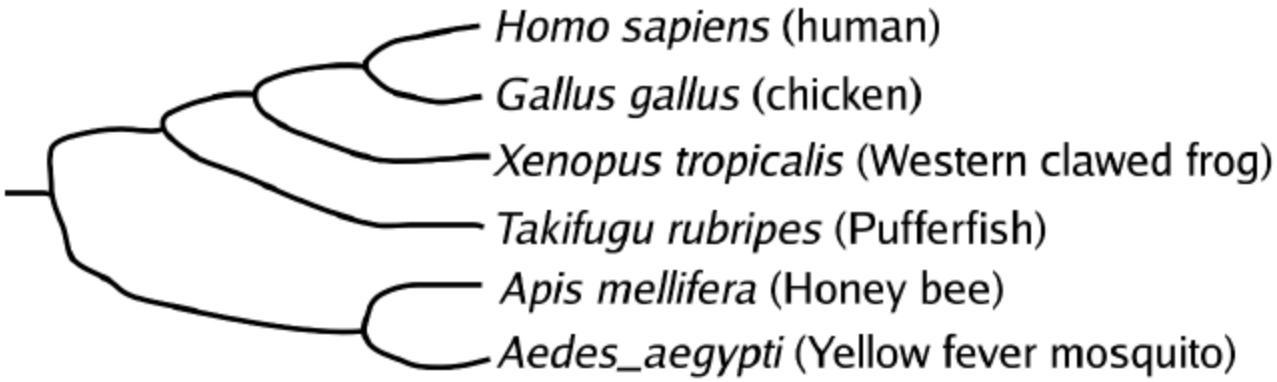
Όπως αναφέραμε, ένα φυλογενετικό δέντρο έναι μια γραφική αναπαράσταση της φυλογένειας μιας ομάδας από taxa ή γονίδια και κατασκευάζεται με την αξιοποίηση γενετικών πληροφοριών ενός ή λίγων γονιδίων. Οι αντικειμενικοί στόχοι της διεξαγωγής μιας φυλογενετικής μελέτης είναι δύο: η αναπαράσταση των πραγματικών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



Εικόνα 52*: Ένα δέντρο ειδών που δείχνει την εξελικτική σχέση των πιθήκων με τον άνθρωπο*

Πηγή: http://www.answersingenesis.org/

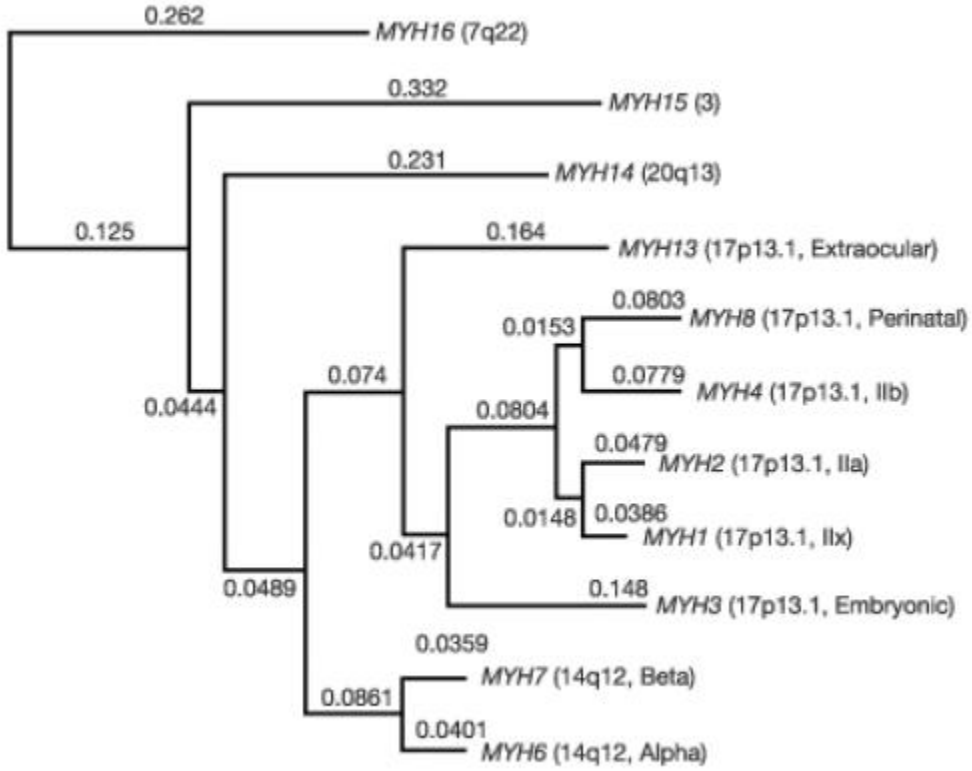
Ένα φυλογενετικό δέντρο μπορεί να χρησιμοποιηθεί για την προσέγγιση των προαναφερθέντων στόχων. Όταν το δέντρο αντανακλά τις φυλογενετικές σχέσεις ομάδων πληθυσμών ή ειδών λέγεται φυλογενετικό δέντρο ειδών η πληθυσμών ενώ όταν κατασκευάζεται με βάση τις νουκλεοτιδικές αλλαγές ενός γονιδίου ή λίγων γονιδίων από κάθε είδος τότε λέγεται γονιδιακό. Γενικότερα, το δέντρο που προκύπτει από μια φυλογενετική ανάλυση, είναι γονιδιακό δέντρο, και όχι μια φυλογένεια των ειδών από τα οποία πάρθηκαν τα γονίδια, αν και στην ιδανική περίπτωση τα δύο αυτά δέντρα ταυτίζονται. Οι εικόνες 52, 53, 54 δείχνουν παραδείγματα δέντρων ειδών και γονιδιακών δέντρων.



\

Εικόνα 53*: Ένα δέντρο ειδών (species tree)*

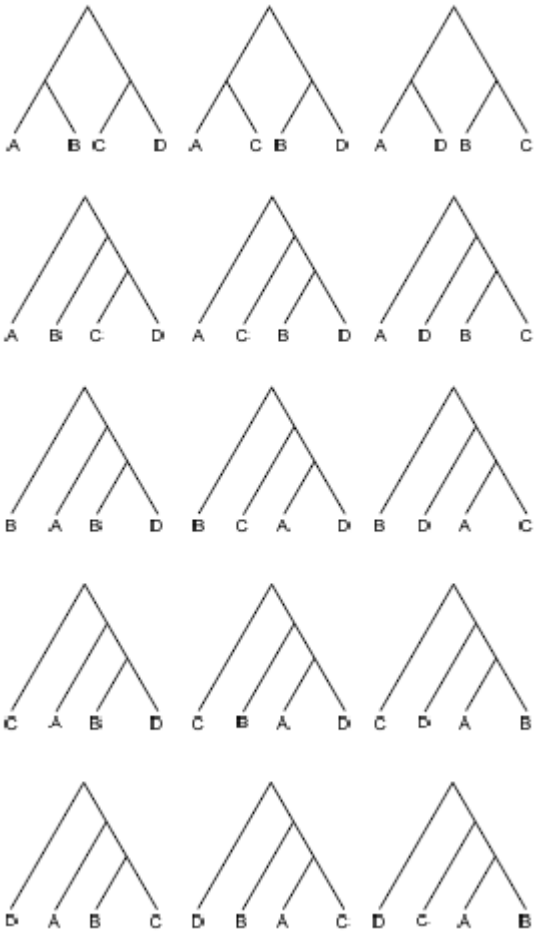
Πηγή: http://bioinformatics.bio.uu.nl/



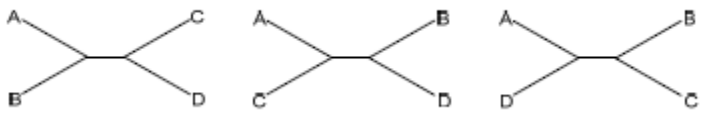
Εικόνα 54*: Ένα γονιδιακό δέντρο (gene tree) που έχει προκύψει από τη σύγκριση του γονιδίου*

Πηγή: http://genomic.unibe.ch/

Τα φυλογενετικά δέντρα μπορεί να είναι με ρίζα, ως συνήθως, η οποία δείχνει τον κοινό πρόγονο καθώς και την εξελικτική κατεύθυνση, ή χωρίς ρίζα, 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 παάραδειγμα 3 είδη, τότε είναι δυνατά 3 δέντρα με ρίζα και 1 χωρίς ρίζα. Στις εικόνες 55, 56 φαίνονται 15 δυνατά δέντρα με ρίζα και τα 3 χωρίς για 4 είδη.



Εικόνα 55*: Για 4 είδη (A, B, C, D) υπάρχουν 15 δυνατά δέντρα με ρίζα*



Εικόνα 56: *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 κα αξιοποιούμενη στη μελέτη ταξινομική μονάδα όπως χαρακτήρας, είδος κ.α) δίνεται από τον τύπο:

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

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

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

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

Τέλος, ένα ακόμη στοιχείο που απασχολεί τους βιολόγους που διεξάγουν φυλογενετικές αναλύσεις, όσο αφορά την κατασκευή φυλογενετικών δέντρων είναι ο χρόνος που πέρασε από τη 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 τα γονιδιακά δέντρα. Αν ο ρυθμός της γονιδιακής αντικατάστασης είναι σταθερός, η αναμενόμενη εξελικτική απόσταση πρέπει να είναι 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.

Ένα δέντρο ειδών είναι πάντα αναμενόμενο δέντρο απόστασης, ενώ ένα γονιδιακό δέντρο μπορεί να είναι είτε αναμενόμενο είτε ρεαλιστικό.

6.7.2 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 την 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). Στη δεύτερη κατηγορία ανήκει η μέθοδος της μέγιστης φειδωλότητας (maximum parsimony).

Κατηγορία μητρών απόστασης

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

Η μέθοδος UPGMA

Ο UPGMA (Unweighted Pair Group Method with Arithmetic Mean) είναι ένας ιδιαίτερα απλός αλγόριθμος ομαδοποίησης. Ο αλγόριθμος UPGMA είναι μια παραλλαγή του HIERARCHICALCLUSTERING που χρησιμοποιεί μια διαφορετική προσέγγιση για τον υπολογισμό της απόστασης μεταξύ των ομάδων, και αναθέτει ύψη στις κορυφές του κατασκευασμένου δέντρου. Το μήκος μιας ακμής (u, v) ορίζεται ως η διαφορά των υψών των κορυφών u και v. Το ύψος παίζει το ρόλο του μοριακού ρολογιού, και μας επιτρέπει να χρονολογήσουμε το σημείο απόκλισης για κάθε κορυφή του εξελικτικού δέντρου.
Δοθέντων 2 ομάδων C1 και C2, ο UPGMA ορίζει την απόσταση μεταξύ τους ως το μέσο όρο των αποστάσεων ανά ζεύγη:

$$D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} D(i, j).$$

UPGMA(D, n)

1 Form n clusters, each with a single element
2 Construct a graph T by assigning an isolated vertex to each cluster
3 Assign height $h(v) = 0$ to every vertex v in this graph
4 **while** there is more than one cluster
5 Find the two closest clusters C_1 and C_2
6 Merge C_1 and C_2 into a new cluster C with $|C_1| + |C_2|$ elements
7 **for** every cluster $C^* \neq C$
8 $D(C, C^*) = \frac{1}{|C| \cdot |C^*|} \sum_{i \in C} \sum_{j \in C^*} D(i, j)$
9 Add a new vertex C to T and connect to vertices C_1 and C_2
10 $h(C) \leftarrow \frac{D(C_1, C_2)}{2}$
11 Assign length $h(C) - h(C_1)$ to the edge (C_1, C)
12 Assign length $h(C) - h(C_2)$ to the edge (C_2, C)
13 Remove rows and columns of D corresponding to C_1 and C_2
14 Add a row and column to D for the new cluster C
15 **return** T

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Ο αλγόριθμος UPGMA παράγει έναν ειδικό τύπο δέντρου με ρίζα που είναι γνωστός ως ultrametric.
Στα δένδρα ultrametric η απόσταση της ρίζας από οποιοδήποτε φύλλο είναι η ίδια.
Έστω 4 λειτουργικές ταξινομικές μονάδες (OTUs) των οποίων οι γενετικές αποστάσεις φαίνονται στον παρακάτω πίνακα

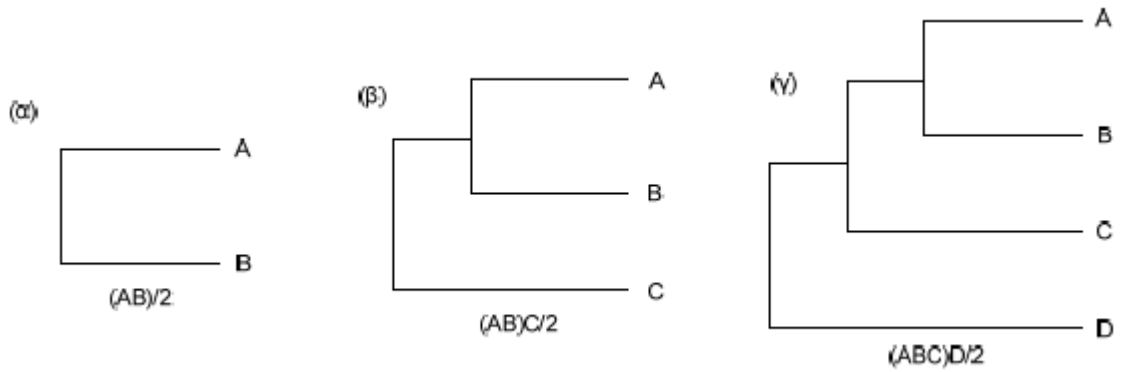
OTU(i/j)	A	B	C
B	D_{AB}		
C	D_{AC}	D_{BC}	
D	D_{AD}	D_{BD}	D_{CD}

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 $D_{AB}/2$. Η ενοποιημένη μονάδα θεωρείται πλέον ως μια σύνθετη ταξινομική μονάδα και δημιουργείται η νέα μήτρα που φαίνεται στον επόμενο πίνακα

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

Η απόσταση 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 μια 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 $D_{(AB)C}/2$. 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 57*:Βαθμιαία δόμηση ενός φυλογενετικού δέντρου με 4 λειτουργικές ταξινομικές μονάδες με τη χρήση της μεθόδου UPGMA*

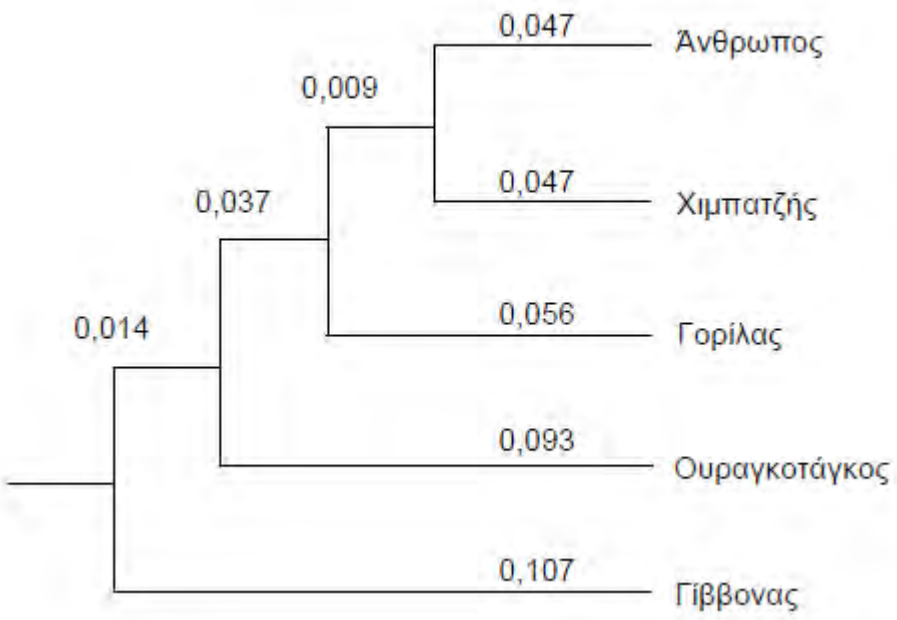
Παράδειγμα

Στον πίνακα που ακολουθεί 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 1

Με εφαρμογή της μεθόδου UPGMA προκύπτει το παρακάτω φυλογενετικό δέντρο

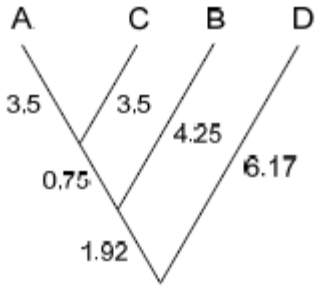


***Εικόνα 58**: Φυλογενετικό δέντρο που κατασκευάστηκε με τη χρήση της μεθόδου UPGMA*

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 στη συνέχεια χρησιμοποιούνται από τη μέθοδο UPGMA για την κατασκευή δέντρου.
Έστω για παράδειγμα η ακόλουθη μήτρα θεωρητικών απόστασεων, οι οποίες με την εφαρμογη της μεθόδου UPGMA δίνουν το δέντρο της εικόνας 59

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



Εικόνα 59: Φυλογενετικό δέντρο που κατασκευάστηκε με τη μέθοδο UPGMA χωρίς να ληφθεί υπόψη η πιθανότητα άνισων ρυθμών υποκατάστασης στους βραχίονες.

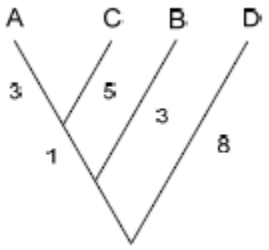
Το παραπάνω δέντρο είναι λαθεμένο εφόσον οι ρυθμοί υποκατάστασης ποικίλουν. Για να εφαρμοστεί η μέθοδος μετασχηματισμένων αποστάσεων έστω ότι θεωρείται ως είδος αναφοράς το D. Ακολουθεί μια διαδικασία διόρθωσης των αποστάσεων χρησιμοποιώντας την εξίσωση:

$$d'_{ij} = [\frac{(d_{ij} - d_{iD} - d_{jD})}{2}] + d''_{D} \quad \text{όπου} \quad d''_{D} = \frac{(d_{AD} - d_{BD} - d_{CD})}{3}$$

και i=A, B ή C και d_{ij} η μετασχηματισμένη απόσταση.

Εφαρμόζοντας την εξίσωση της προκύπτει η νέα μήτρα με τις διορθωμένες τιμές που ακολουθεί.

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



***Εικόνα 60**: Διορθωμένο φυλογενετικό δέντρο με τη μέθοδο των μετασχηματισμένων αποστάσεων*

Η μέθοδος των Fitch-Margoliash

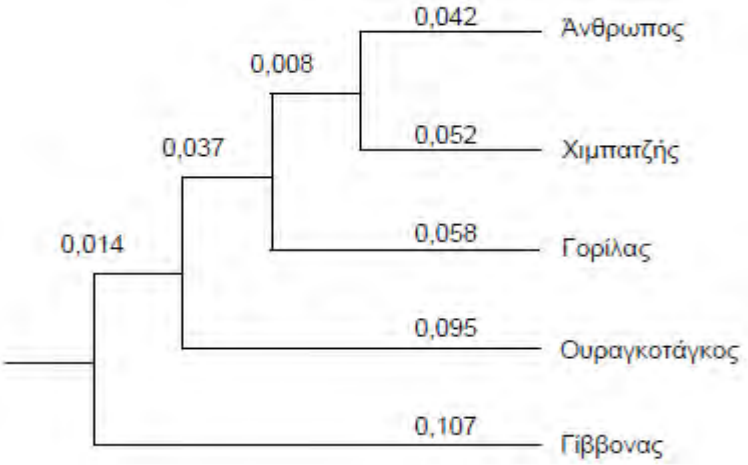
Η μέθοδος Fitch-Margoliash θεωρείται κατάλληλη για να πάρει κανείς πιο σωστά και πραγματικά μέρη βραχιόνων σε ένα δέντρου του οποίου η τοπολογία έχει διορθωθεί με την εφαρμογή των μεθόδων UPGMA και μετασχηματισμένων αποστάσεων.
Έστω ο πίνακας 1 από τον οποίο προέκυψε με την εφαρμογή της μεθόδου UPGMA το δέντρο της εικόνας 58. Γενικά πρέπει να τονιστεί ότι αν υπάρχουν περισσότερες από τρεις ταξινομικές μονάδες, η γενική πορεία επεξεργασίας που ακολουθείται είναι να προβάλλονται κάθε φορά σε τρεις, 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 σε κάθε βραχίονα είναι:

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

Στη συνέχεια οι μονάδες Α και Β σχηματίζουν μια σύνθετη ταξινομική μονάδα την ΑΒ. Ξαναυπολογίζονται 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 τη σύνθετη μονάδα που αποτελείται από όλες τις υπόλοιπες. Στη συνέχεια ξαναυπολογίζονται τα νέα X, Y, Z. Η πορεία αυτή συνεχίζεται μέχρι να συνδυαστούν όλες οι ταξινομικές μονάδες σε απλή οικογένεια. Τέλος, η μέθοδος καταλήγει στο ακόλουθο αναδομημένο φυλογενετικό δέντρο.



***Εικόνα 61**: Αναδομημένο φυλογενετικό δέντρο με τη μέθοδο Fitch-Margoliash*

Η μέθοδος των γειτονικών ζευγαριών

Μπορούμε τώρα να επιστρέψουμε στα "γειτονικά φύλλα" ιδέα που αναπτύξαμε και στη συνέχεια εγκαταλείψαμε στην προηγούμενη ενότητα. Το 1987 Naruya Saitou και Masatoshi Nei ανέπτυξαν έναν έξυπνο αλγόριθμο ένωσης γειτόνων για ανοικοδόμηση φυλογενετικού δέντρου. Στην περίπτωση των additive δένδρων, ο αλγόριθμος NEIGHBORJOINING βρίσκει κάπως μαγικά ζεύγη γειτονικών φύλλων και προχωράει αντικαθιστώντας τα εν λόγω ζεύγη με τον «πατέρα» των φύλλων. Ωστόσο, αλγόριθμος NEIGHBORJOINING λειτουργεί καλά όχι μόνο για τους πρόσθετους πίνακες απόστασης αλλά για πολλούς άλλους: δεν **αναλαμβάνει** την ύπαρξη ενός μοριακού ρολογιού και εξασφαλίζει ότι οι ομάδες που συγχωνεύονται κατά τη διάρκεια της ανοικοδόμησης του δέντρου δεν είναι μόνο κοντά η μια στην άλλη (όπως στον UPGMA), αλλά και πολύ μακριά από τις υπόλοιπες.
Για μια ομάδα C, ορίζουμε το

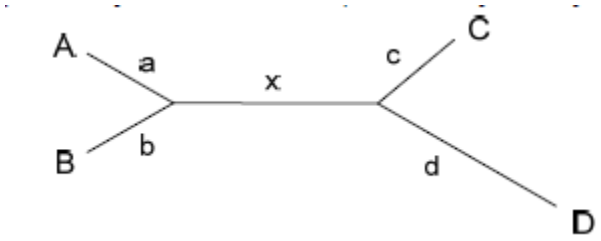
$$u(C) = \frac{1}{\text{number of clusters} - 2} \sum_{\text{all clusters } C'} D(C, C')$$

ως μέτρο του διαχωρισμού της C από τις άλλες ομάδες. Για να επιλέξουμε ποιες δύο ομάδες θα συγχωνεύσουμε, ψάχνουμε για τις ομάδες C_1 και C_2, που είναι ταυτόχρονα κοντά η μία στην άλλη και μακριά από τις άλλες. Κάποιος μπορεί να προσπαθήσει να συγχωνεύσει ομάδες που ταυτόχρονα ελαχιστοποιούν το D (C_1, C_2) και μεγιστοποιούν το u (C_1) + u (C_2). Ωστόσο, είναι απίθανο να υπάρχει τέτοιο ζευγάρι. Ως εναλλακτική λύση, επιλέγουμε να ελαχιστοποιούμε D (C_1, C_2) - u (C_1) - u (C_2). Η προσέγγιση αυτή χρησιμοποιείται στον αλγόριθμο NEIGHBORJOINING που ακολουθεί.

NEIGHBORJOINING(D, n)
1 Form n clusters, each with a single element
2 Construct a graph T by assigning an isolated vertex to each cluster
3 **while** there is more than one cluster
4 Find clusters C_1 and C_2 minimizing $D(C_1, C_2) - u(C_1) - u(C_2)$
5 Merge C_1 and C_2 into a new cluster C with $|C_1| + |C_2|$ elements
6 Compute $D(C, C^*) = \frac{D(C_1, C) + D(C_2, C)}{2}$ to every other cluster C^*
7 Add a new vertex C to T and connect it to vertices C_1 and C_2
8 Assign length $\frac{1}{2}D(C_1, C_2) + \frac{1}{2}(u(C_1) - u(C_2))$ to the edge (C_1, C)
9 Assign length $\frac{1}{2}D(C_1, C_2) + \frac{1}{2}(u(C_2) - u(C_1))$ to the edge (C_2, C)
10 Remove rows and columns of D corresponding to C_1 and C_2
11 Add a row and column to D for the new cluster C
12 **return** T

Πηγή: *An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner*

Έστω το δέντρο χώρίς ρίζα της εικόνας 62 στο οποίο οι ταξινομικές μονάδες Α και Β είναι γειτονικές, όπως και οι C, D



Εικόνα 62: Δέντρο χωρίς ρίζα για 4 OTUs

Για το συγκεκριμένο δέντρο, υπό την προυπόθεση 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 περιπτωση που έχουμε 4 ταξινομικές λειτουργικές μονάδες άγνωστης φυλογενετικής σχέσης, ώστε να εντοπιστούν τα γειτονικά ζευγάρια. Για 5 ταξινομικές λειτουργικές μονάδες, που θα μελετήσουμε, υπάρχουν 5 δυνατές περιπτώσεις 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 παρακάτω πίνκα όπου παρουσιάζονται οι μέσες τιμές νουκλεοτιδικών υποκαταστάσεων ανά 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

Γενικά, για κάθε τετράδα ταξινομικών μονάδων, έστω i, j, k και l, υπολογίζονται οι παραστάσεις:

$$D_{ij} + D_{kl}\ ,\ D_{jk} + D_{jl} \text{ και } D_{il} + D_{jk}$$

Έτσι, λοιπόν, για τις ταξινομικές μονάδες του πίνακα 2 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)

Για κάθε γραμμή του παραπάνω πίνακα, εξετάζονται τα αθροίσματα της δεύτερης στήλης και αφού επιλεχτεί το μικρότερο, το ζευγάρι στο οποίο αντιστοιχεί αποθηκεύεται στην τρίτη στήλη. Έτσι, λοιπόν, από το 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. Τη 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 στη 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

Αφού πλέον υπάρχουν 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 απιλέγεται το (AB) ως το ένα γειτονικό ζευγάρι και το C(DE) ως το άλλο γειτονικό ζευγάρι.

Τέλος, πρέπει να ανεφερθεί 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 το 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 να έχει πολύ μεγάλη απόσταση από τα άλλα, αλλά ούτε και πολύ μικρή για να μπορέο να διακρίνεται. Ωστόσο υπάρχει περίπτωση να μην μπορεί να βρεθεί είδος αναφοράς. Σε αυτή την περίπτωση και υπό την προυπόθεση ότι ο ρυθμός εξέλιξης είναι χοντρικά ο ίδιος γιο όλους τους βραχίονες, η ρίζα μπορεί να τοποθετηθεί στο μέσο σημείο της μεγαλύτερης απόστασης μεταξύ δύο μονάδων.

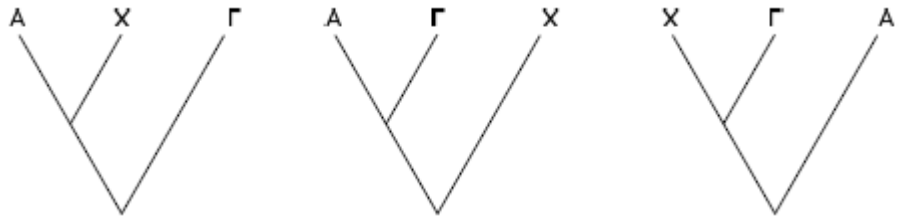
Κατηγορία μεθόδων που βασίζονται στην παρουσία-απουσία πληροφοριακών χαρακτήρων

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 και λειτουργικών μονάδων. Στη μέθοδο αυτή γίνεται διάκριση πληροφοριακών και μη πληροφοριακών θέσεων.
Στον παρακάτω πίνακα 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

Μια νουκλεοτιδική θέση είναι πληροφοριακή μόνο αν αφορά τουλάχιστον δύο διαφορετικά νουκλεοτίδια που το καθένα απαντάται τουλάχιστον 2 φορές. Η συγκεκριμένη αυτή θέση αξιοποιείται για να βρεθεί το φειδωλό δέντρο Στη δεύτερη γραμμή του παραπάνω πίνακα για παράδειγμα, που αφορά τα νουκλεοτίδια C και A τα οποία βρίσκονται στην 34η 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 63, στην οποία αποτυπώνονται τρεις υποθέσεις: κατά την πρώτη υπόθεση ο άνθρωπος και ο χιμπατζής ανήκουν σε έναν κλάδο, κατά τη δεύτερη υπόθεση ο χιμπατζής και ο γορίλας ανήκουν σε έναν κλάδο, ενω κατά την τρίτη ο άνθρωπος και ο γορίλας ανήκουν σε έναν κλάδο.



Εικόνα 63*: 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.
Καταλήγοντας, το πιο σωστό φυλογενετικό δέντρο θεωρείται μεταξύ όλων των πιθανών εκείνο που κατασκευάζεται με τον μικρότερο αριθμό νουκλεοτιδικών αλλαγών. Στη μέθοδο αυτή ο βαθμός κάθε δένδρου 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 με 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 κλαδιά για δύο είδη, δλδ υψηλό ποσοστό μεταλλάξεων, ενώ ξεκινάνε επίσης κοντά κλαδιά για άλλα δύο είδη.

7. Επίλογος

Μέσα από αυτή τη διπλωματική εργασία παρατηρήσαμε τη ραγδαία ανάπτυξη της Βιοπληροφορικής, ακολουθώντας την περιγραφή των αλγορίθμων που παρουσιάσαμε. Η ανάπτυξη αυτή την καθιστά όχι μόνο ένα νέο επιστημονικό πεδίο με ποικίλες εφαρμογές αλλά ταυτόχρονα δίνει ώθηση στην ανάπτυξη νέων πεδίων και στην επιστήμη της πληροφορικής για τη διαχείριση και την ανάλυση βιολογικών ακολουθιών.

Μπορούμε να πούμε πως οι εφαρμογές της Βιοπληροφορικής συνοψίζονται στην ανεύρεση της λειτουργίας των πρωτεϊνών, ομαδοποίηση πρωτεϊνών σε λειτουργικές ομάδες, ανεύρεση αλληλεπιδράσεων των πρωτεϊνών μεταξύ τους και κατανόηση της πολυπλοκότητας των βιολογικών συστημάτων, στη σύγκριση του γονιδιώματος διαφόρων ειδών, και στην εύρεση των εξελικτικών σχέσεων των οργανισμών μεταξύ τους, στην απόκτηση γνώσης για το ρόλο των μη κωδικοποιημένων περιοχών του DNA στη μορφολογία και έκφραση των γονιδίων, στην προσπάθεια αντιμετώπισης διαφόρων ασθενειών με την ανάπτυξη νέων διαγνωστικών μέτρων και θεραπευτικών μεθόδων, στην παραγωγή πιο αποτελεσματικών φαρμακευτικών προϊόντων με όλο και περισσότερες εφαρμογές τα επόμενα έτη στην Βιολογία και στην Ιατρική

Οι μελλοντικές επιδιώξεις εστιάζονται στο υπάρχον έλλειμμα υπολογιστικής ισχύος που δεν καλύπτει τη βιολογική πολυπλοκότητα. Γι’ αυτό «επιστρατεύεται» η Βιοχημεία για το σχεδιασμό βιοχημικών δικτύων και η Φυσικοχημεία για την καλύτερη κατανόηση των μικρών διαστάσεων του κυττάρου και των οργανιδίων του.

Ένας νέος λοιπόν συντονισμός, βιολόγων, φυσικών, χημικών και πληροφορικών ίσως οδηγήσει σε μια νέα επιστημονική σύνθεση.

Το μέλλον θα προχωρήσει με τη δημιουργία λογισμικών που θα απεικονίζουν τη λειτουργία των πρωτεϊνών και την αποκαλούμενη «έκφραση των γονιδίων» (ασθένειες), κατασκευή υπολογιστών στους οποίους θα έχει αντικατασταθεί η μνήμη με ζωντανά νευρικά κύτταρα, κάτι που θα πλαισιωθεί με τη δημιουργία νέων ειδικοτήτων, όπως αυτή του βιολόγου-λογισμικού, του ιατρού-προγραμματιστή, του δικτυακού-οικονομολόγου (Αλαχιώτης, 2003; Krane et al., 2003; de Silva et al., 2006; Perez-Iratxeta et al., 2006)

8. Βιβλιογραφία

- Wikipedia

- Evolutionary Computation in Bioinformatics: A Review
  Sankar K. Pal, Fellow, IEEE, Sanghamitra Bandyopadhyay, Senior Member, IEEE, and Shubhra Sankar Ray

- An introduction to Bionformtics Algorithms, Neil C. Jones and Pavel A. Pevzner

- «Αλγόριθμοι διαχείρισης και ανάλυσης ακολουθιών βιολογικών δεδομένων με εφαρμογή σε προβλήματα βιοπληροφορικής», διδακτορική διατριβή: Περδικούρη Αικατερίνη

- «Χρήση δέντρων επιθεμάτων για την μελέτη γονιδιωματικών ακολουθιών», διπλωματική εργασία: Μαρία Κ. Παπαδοπούλου-Βόλος 2008

- Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον. Τάξη Γ΄ Ενιαίου Λυκείου , Βακάλη Α., Γιαννόπουλος Η., Ιωαννίδης Ν., Κοίλιας Χ. Μάλαμας Κ., Μανωλόπουλος Ι., Πολίτης Π.

- Διάλεξη 1 («ενδονουκλεάσες περιορισμού») του μαθήματος «Μοριακή Βιολογία», Δρ. Χρήστος Παναγιωτίδης-Τμήμα φαρμακευτικής

- Διαλέξεις του μαθήματος «Εισαγωγή στη πληροφορική», Άρτεμις Χατζηγεωργίου – Τμήμα Μηχανικών Η/Υ Τηλεπικοινωνιών και Δικτύων-Βόλος 2010

- Διάλεξη 6(«Αναζήτηση μοτίβων σε αλληλουχίες. Ομοιότητα αλληλουχιών») του μαθήματος Υπολογιστική Βιολογία, Χριστόφορος Νικολάου-2011

- «Κλινικά και ακοολογικά ευρήματα στο σύνδρομο Waardenburg», Γ. Ψύλλας, Α. Ψηφίδης, Μ. Χίτογλου-Αντωνιάδου, Β. Νικολαΐδης, Α. Κουλούλας - τεύχος 25, Ιούλιος-Αύγουστος-Σεπτέμβριος 2006, σελίδες 33-36

- http://www.cs.vu.nl/~gusz/ecbook/Eiben-Smith-Intro2EC-Ch2.pdf

- http://www.solver.com/gabasics.htm

- http://www.cs.cmu.edu/Groups/AI/html/faqs/ai/genetic/part2/faq-doc-1.html

- http://www.enthesis.net/index.php?option=com_content&view=article&id=611:enthesis8055&catid=27:e-health&Itemid=6

- http://bioinformatics.biol.uoa.gr/msc/gr/general.html
- http://www.biology.uoc.gr/courses/BIO105_Genetiki/Lectures/notes%202.pdf
- http://mmlab.ceid.upatras.gr/bioinfo/Part-A-Kef1.pdf
- http://mmlab.ceid.upatras.gr/bioinfo/Part-A-Kef3.pdf