**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**

**ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ**
**ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

# «ΧΡΗΣΗ ΔΕΝΤΡΩΝ ΕΠΙΘΕΜΑΤΩΝ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ ΓΟΝΙΔΙΩΜΑΤΙΚΩΝ ΑΚΟΛΟΥΘΙΩΝ»

ΕΚΠΟΝΗΣΗ
**ΜΑΡΙΑ Κ. ΠΑΠΑΔΟΠΟΥΛΟΥ**

ΕΠΙΒΛΕΠΟΝΤΕΣ ΚΑΘΗΓΗΤΕΣ:

ΗΛΙΑΣ ΧΟΥΣΤΗΣ
ΤΡΙΑΣ ΘΗΡΑΙΟΥ
ΒΑΣΙΛΕΙΟΣ ΑΤΛΑΜΑΖΟΓΛΟΥ

ΒΟΛΟΣ , ΟΚΤΩΒΡΙΟΣ 2008

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ**

**ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ**

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ**
**ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

# «ΧΡΗΣΗ ΔΕΝΤΡΩΝ ΕΠΙΘΕΜΑΤΩΝ ΓΙΑ ΤΗΝ ΜΕΛΕΤΗ ΓΟΝΙΔΙΩΜΑΤΙΚΩΝ ΑΚΟΛΟΥΘΙΩΝ»

ΕΚΠΟΝΗΣΗ
**ΜΑΡΙΑ Κ. ΠΑΠΑΔΟΠΟΥΛΟΥ**

ΕΠΙΒΛΕΠΟΝΤΕΣ ΚΑΘΗΓΗΤΕΣ:

ΗΛΙΑΣ ΧΟΥΣΤΗΣ
ΤΡΙΑΣ ΘΗΡΑΙΟΥ
ΒΑΣΙΛΕΙΟΣ ΑΤΛΑΜΑΖΟΓΛΟΥ

ΒΟΛΟΣ , ΟΚΤΩΒΡΙΟΣ 2008

Η σελίδα αυτή είναι σκόπιμα λευκή.

# Περίληψη

Σκοπός της συγκεκριμένης διπλωματικής εργασίας είναι η δημιουργία ενός αλγορίθμου όπου με χρήση δένδρων επιθεμάτων θα γίνεται εύρεση επαναληπτικών ακολουθιών σε ένα αρχείο DNA. Σε πρώτη θεώρηση, επιδιώκεται να γίνει μια συστηματική παρουσίαση βιολογικών εννοιών που αφορούν την Βιοπληροφορική. Επιπλέον, παρουσιάζονται αναλυτικά οι διαφορετικές μέθοδοι της που έχουν χρησιμοποιηθεί έως τώρα, όσον αφορά την κατασκευή δένδρων επιθεμάτων, καθώς και οι απαιτήσεις συστήματος που χρειάζονται. Σε δεύτερη φάση, παρουσιάζεται η χρήση της εφαρμογής στις ακολουθίες του arabidopsis chrM και chrC, την ακολουθία Enterobacteria phage phiX174 genome, Human herpesvirus 4 type 2 genome, και UP_seq_antagomir.

Πιο συγκεκριμένα, το Κεφάλαιο 1 δίνει μια γενικότερη περιγραφή του υπό μελέτη προβλήματος που προσπαθήσαμε να αναλύσουμε. Περιλαμβάνει εισαγωγικές βιολογικές έννοιες και λειτουργίες απαραίτητες για την κατανόηση των προβλημάτων που ανακύπτουν. Το Κεφάλαιο 2 παρουσιάζονται οι βασικότερες τάσεις που υπάρχουν σήμερα στο χώρο της βιοπληροφορικής. Αυτές σε συνδυασμό με την ανάπτυξη των διαδικτυακών εργαλείων προσφέρονται για την δημιουργία νέων μορφών προγραμματισμού με απώτερο σκοπό την εξυπηρέτηση των ερευνητών στον τομέα της βιοπληροφορικής. Επίσης περιγράφεται το περιβάλλον εργασίας της Java και οι δυνατότητες που προσφέρει η χρήση της. Στο Κεφάλαιο 3 γίνεται μια αφορά στα δέντρα επιθεμάτων. Στην συνέχεια αναλύονται οι διάφοροι μέθοδοι και τα αποτελέσματά τους. Τέλος παρουσιάζονται τα λογικά διαγράμματα της υλοποιημένης εφαρμογής. Στο Κεφάλαιο 4 παρουσιάζεται η εκτέλεση της εφαρμογής για τις ακολουθίες του arabidopsis chrM και chrC, την ακολουθία Enterobacteria phage phiX174 genome, Human herpesvirus 4 type 2 genome. Παρατίθενται επίσης πειραματικά αποτελέσματα με τον απαραίτητο σχολιασμό και οι συγκρίσεις που προκύπτουν.

## Λέξεις Κλειδιά

Δέντρα επιθεμάτων, ολιγομερή, επαναληπτικές ακολουθίες ,ο αλγόριθμος TDD, βιοπληροφορική, ακολουθίες DNA

## Abstract

The purpose of the present diploma thesis is the creation of an algorithm which will use suffix trees to find repetitive sequences in a file DNA. During the primary stage, a systematic approach of biological definitions, which concern Bioinformatics, is presented. Moreover, the different methods, which have been used until now for the creation of suffix trees, are discussed. Furthermore system requirements are analysed. During the second phase, the use of the application in the sequences Arabidopsis chrM and chrC, the sequence Enterobacteria phage phiX174 genome, Human herpesvirus 4 type 2 genome, and UP_seq_antagomir is presented.

Specifically, Chapter 1 gives a general description over the problem we worked towards to analyze. Introductive biological significances and operations, essential for the comprehension of the problems that emerge, are included. In Chapter 2 the basic fields in the space of bioinformatics are introduced. These in combination with the growth of internet tools are offered for the creation of new forms of programming, which will aim researchers in the sector of bioinformatics. Moreover the environment of Java is described as well as the possibilities that are offered with its use. Chapter 3 focuses in suffix trees .Various methods and their results are analyzed. Finally logical diagrams are presented. In Chapter 4 the implementation of the application for the sequences Arabidopsis chrM and chrC, the sequence Enterobacteria phage phiX174 genome, Human herpesvirus 4 type 2 genome is presented. Experimental results and comparisons are also mentioned.



### *Key Words*

Suffix trees, oligomers, repetitive sequences, algorithm TDD, bioinformatics, Sequences DNA

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ηλία Χούστη επιβλέποντα της διπλωματικής μου, για τη στήριξη και τη βοήθεια που μου παρείχε αυτά τα έξι χρόνια των σπουδών μου. Ένα μεγάλο μέρος αυτής της εργασίας ανήκει στους επιβλέποντες καθηγητές μου, κ. Άντα Θηραίου και κ. Βασίλη Ατλαμαζόγλου, που έβρισκαν πάντα χρόνο για να με καθοδηγήσουν και να απαντήσουν στις δεκάδες ερωτήσεις μου. Τους ευχαριστώ θερμά για όλα.

Επιπλέον, ένα μεγάλο ευχαριστώ στον Λεωνίδα για τις πολύτιμες συμβουλές του και στους φίλους και συμφοιτητές μου για την αλληλοϋποστήριξη και τα αξέχαστα χρόνια που περάσαμε. Τέλος, ευχαριστώ μέσα από την καρδιά μου την οικογένειά μου για την αμέριστη συμπαράσταση και αγάπη τους, την οικονομική τους στήριξη όλα αυτά τα χρόνια των σπουδών μου και το σεβασμό και την κατανόηση που έχουν πάντα για τις επιλογές μου.

# Περιεχόμενα

# ΚΕΦΑΛΑΙΟ 1

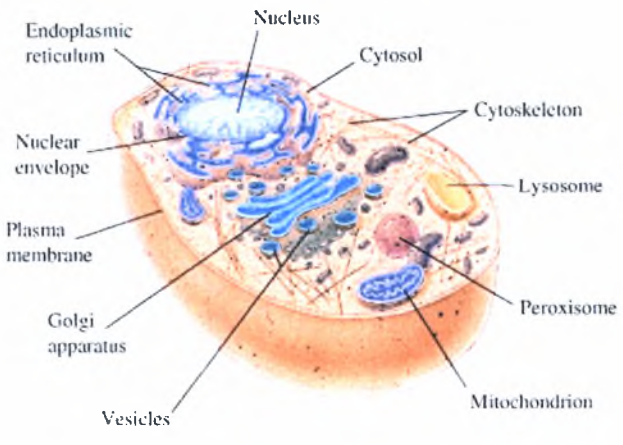
## Εισαγωγή στη βιολογία του προβλήματος

### *1.1 Εισαγωγή*

Οι απλούστερες δομικές μονάδες στις οποίες μπορεί να διαιρεθεί ένας πολύπλοκος οργανισμός και να εξακολουθούν αν διατηρούν τις χαρακτηριστικές λειτουργίες της ζωής , ονομάζονται κύτταρα. Κάθε ανθρώπινος οργανισμός ξεκινά ως ένα μοναδικό κύτταρο, που διαιρείται δημιουργώντας δύο, καθένα από τα οποία διαιρείται και ούτω κάθε εξής. Κατά τη διάρκεια της ανάπτυξης , κάθε κύτταρο εξειδικεύεται για την εκτέλεση συγκεκριμένων λειτουργιών. Η διαδικασία της μετατροπής ενός μη εξειδικευμένου κυττάρου σε εξειδικευμένο ονομάζεται κυτταρική διαφοροποίηση. Βάσει της πολυπλοκότητας της δομής τους τα κύτταρα διακρίνονται σε δύο κατηγορίες, τα ευκαρυωτικά και τα προκαρυωτικά.

Η κυτταρική θεωρία δηλώνει ότι:

1. Όλοι οι οργανισμοί αποτελούνται από ένα ή περισσότερα κύτταρα.
2. Όλα τα κύτταρα προέρχονται από τα προϋπάρχοντα κύτταρα.
3. Οι ζωτικής σημασίας λειτουργίες ενός οργανισμού εμφανίζονται μέσα στα κύτταρα.
4. Όλα τα κύτταρα περιέχουν όλες τις κληρονομικές πληροφορίες που είναι απαραίτητες για τη ρύθμιση των λειτουργιών των κυττάρων και για τη διαβίβαση των πληροφοριών στην επόμενη γενεά των κυττάρων.




**Σχήμα 1.1 Απεικόνιση του κυττάρου**

## *1.2 Βασικές Βιολογικές Έννοιες*

Η σύγχρονη βιολογία του κυττάρου περικλείει τρία διακριτά αλλά παράλληλα αλληλοσυμπληρούμενα επιστημονικά πεδία, την κυτταρολογία (cytology), τη βιοχημεία (biochemistry) και τη γενετική (genetics). Η κυτταρολογία ασχολείται με την δομή του κυττάρου και μελετά τις ιδιότητες , τη συμπεριφορά, την αλληλεπίδραση, και το περιβάλλον του, ενώ η βιοχημεία καλύπτει την χημεία μιας βιολογικής δομής και λειτουργίας. Η γενετική επικεντρώνει στην διάχυση των γενετικών πληροφοριών, που ενυπάρχουν μέσα στο κύτταρο, σε συγκεκριμένες μορφές και την αποκωδικοποίησή τους με επίσης συγκεκριμένο τρόπο. Βασικές έννοιες που καθορίζουν την ροή και την αποκωδικοποίηση των γενετικών πληροφοριών είναι αυτές του DNA, του γονιδίου, του χρωμοσώματος, του RNA και της πρωτεΐνης.

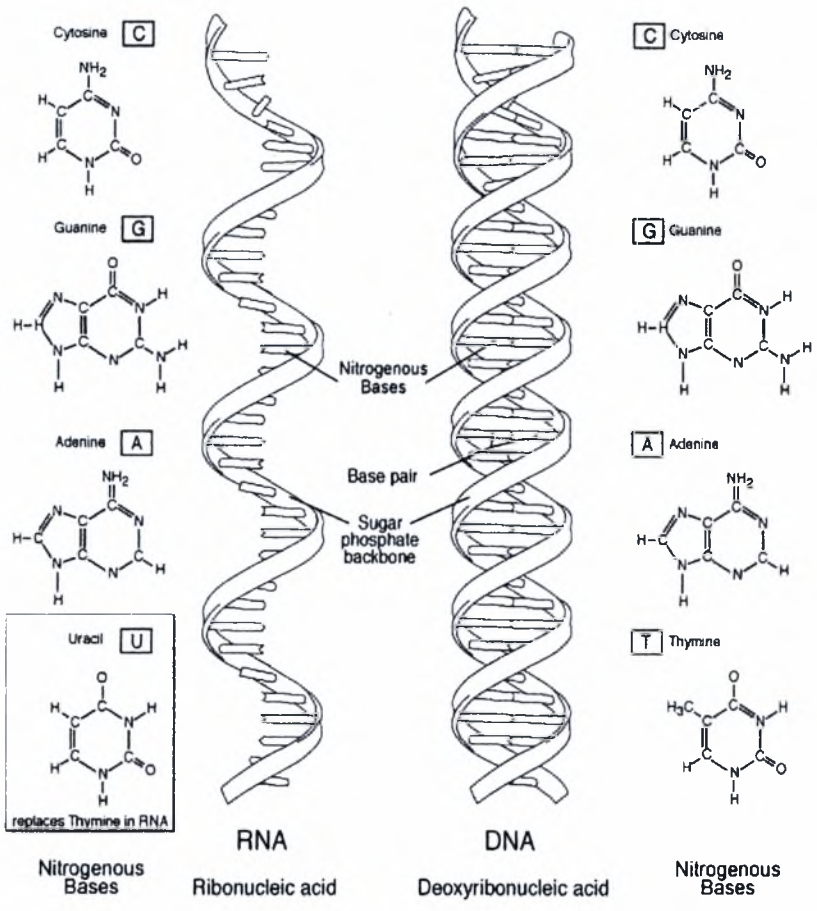
Μία από τις κυριότερες ιδιομορφίες της ζώσας ύλης είναι ότι περιέχει μακρομόρια. Ο όρος **μακρομόρια** αναφέρεται συνήθως σε μόρια υψηλού μοριακού βάρους. Γενικά ως "βιολογικά μακρομόρια" χαρακτηρίζονται σύνθετες οργανικές ενώσεις μεγάλου μοριακού βάρους ($10^3$ - $10^9$) όπως είναι οι πρωτεΐνες, τα νουκλεϊκά οξέα, οι πολυσακχαρίτες και τα λιπίδια. Τα μακρομόρια που εμπλέκονται στην λειτουργία του κυττάρου, είναι:

- Τα **νουκλεϊκά οξέα**
- Οι **πρωτεΐνες**

### *1.2.1 Τα Νουκλεϊκά Οξέα*

Τα νουκλεϊκά οξέα παίρνουν την ονομασία τους από τον Φριντριχ Μίσερ ο οποίος το 1869 ανακάλυψε στους πυρήνες των κυττάρων μια ουσία με συγκεκριμένη όξινη αντίδραση την οποία στα γερμανικά ονόμασε *Nuklein* , δηλαδή ουσία του πυρήνα. Αργότερα το 1889 ο μαθητής του Ρίτσαρντ Άλτμαν την μετονόμασε σε *Nukleinsäure* που μεταφράζεται στα ελληνικά ως νουκλεϊκό οξύ.

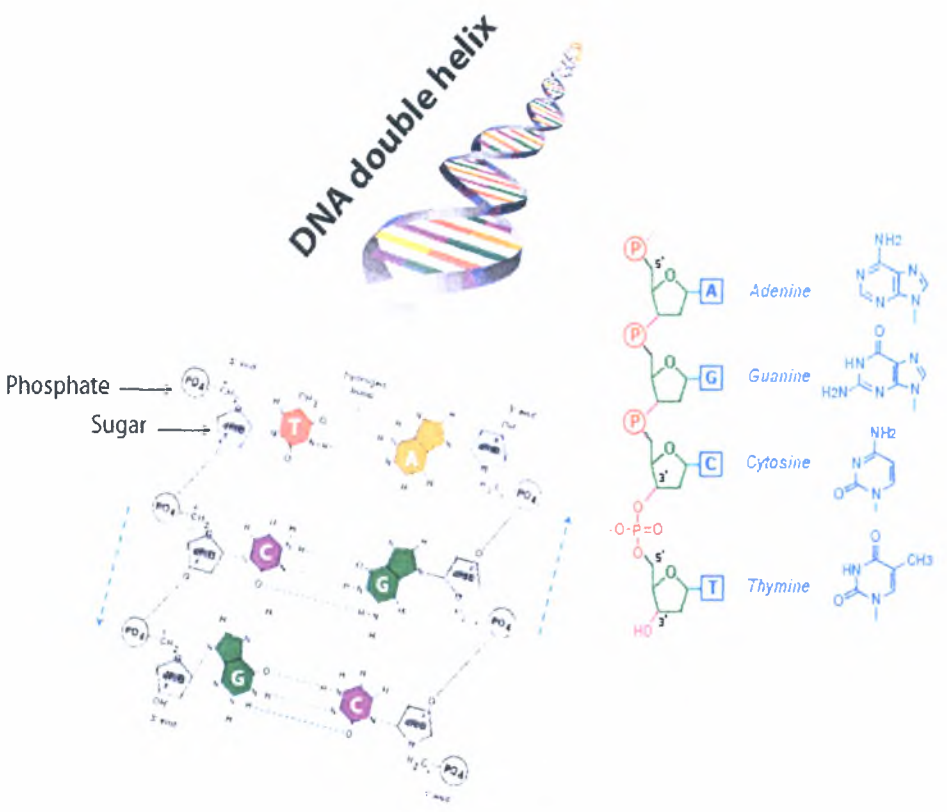
Τα νουκλεϊκά οξέα (nucleic acids) είναι εκείνα τα μακρομόρια τα οποία περιέχουν τη γενετική πληροφορία και εξασφαλίζουν τη μεταβίβασή της. Είναι πολυμερή μόρια τα οποία συνίστανται από μια αλληλουχία δομικών μονάδων, η συγκρότηση της οποίας πραγματοποιείται με βάση τέσσερις δομικές μονάδες, τέσσερα διαφορετικά νουκλεοτίδια. Το κάθε νουκλεοτίδιο αποτελείται από τρία απλούστερα μόρια, που συνδέονται μεταξύ τους με χημικούς δεσμούς: μια αζωτούχα βάση, ένα σάκχαρο και μια φωσφορική ομάδα. Το μόνο από τα τρία επιμέρους μόρια ενός νουκλεοτιδίου που μεταβάλλεται είναι η αζωτούχα βάση, ενώ το σάκχαρο και η φωσφορική ομάδα παραμένουν τα ίδια και για τα τέσσερα είδη δομικών νουκλεοτιδίων. Μπορούμε, δηλαδή, να πούμε ότι τα δυο τελευταία αποτελούν τη «σπονδυλική στήλη» της πολυνουκλεοτιδικής αλυσίδας, πάνω στην οποία έχουν προσαρτηθεί οι τέσσερις διαφορετικές αζωτούχες βάσεις. Κατά συνέπεια μπορούμε να θεωρήσουμε ότι η διαδοχή των νουκλεοτιδίων αντιστοιχεί πρακτικά στην διαδοχή των αζωτούχων βάσεων που ανήκουν στα διαφορετικά νουκλεοτίδια. Οι δυο τύποι νουκλεϊκών οξέων που υπάρχουν είναι το DNA και το RNA.




**Σχήμα 1.2 Απεικόνιση του DNA και του RNA**

## *1.2.1.1 Δεσοξυριβουνουκλεϊκό οξύ (DNA)*

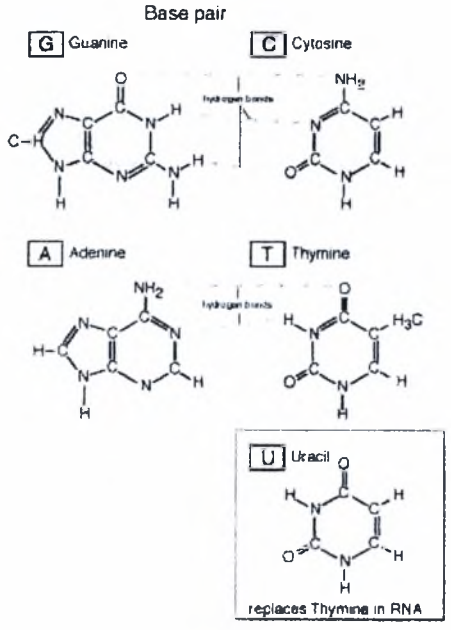
**Το DNA (δεσοξυριβονουκλεϊκό οξύ) είναι το υπόβαθρο και ο φορέας των γενετικών πληροφοριών.** Οι τέσσερις αζωτούχες βάσεις που μπορούν να περιγράψουν την γραμμική απεριοδική ακολουθία του DNA είναι η αδενίνη, η κυτοσίνη, η γουανίνη και η θυμίνη, οι οποίες συμβολίζονται με τα αρχικά των αντίστοιχων ονομάτων στο λατινικό αλφάβητο: A, C, G και T. Το DNA αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες όπου η μία περιελίσσεται στην άλλη και οι οποίες συνδέονται μεταξύ τους με δεσμούς υδρογόνου ανάμεσα στις συμπληρωματικές βάσεις. Οι συμπληρωματικές βάσεις στο DNA είναι η αδενίνη (A) με την θυμίνη (T) και η γουανίνη (G) με την κυτοσίνη (C). Με αυτόν τον τρόπο δημιουργείται ένα σπειροειδές μόριο, η γνωστή **διπλή έλικα** του DNA.




**Σχήμα 1.3 Απεικόνιση της διπλής έλικας του DNA**

Η κάθε έλικα του DNA έχει μια εγγενή κατευθυντικότητα που συμβολίζεται με βάση την αρίθμηση των ατόμων άνθρακα στο σάκχαρο, έχοντας ως σημείο εκκίνησης το άτομο του άνθρακα που συνδέεται με την αζωτούχα βάση. Με αυτόν τον τρόπο η βασική πολυνουκλεοτιδική αλυσίδα συνδέεται με το 3ο άτομο άνθρακα στο επόμενο νουκλεοτίδιο και με το 5ο άτομο άνθρακα με την φωσφορική ομάδα. Η κατεύθυνση αυτή συμβολίζεται σαν 5'→ 3'. Η συμπληρωματική πολυνουκλεοτιδική αλυσίδα έχει αντιπαράλληλη κατευθυντικότητα, και προφανώς συμβολίζεται σαν 3'→ 5'. Με την λειτουργία της μεταγραφής και της μετάφρασης που θα περιγραφεί μετέπειτα, το DNA καθορίζει ποια πρωτεΐνη θα οικοδομηθεί και ποιον ρόλο θα έχει αυτή μέσα σε έναν συγκεκριμένο οργανισμό.

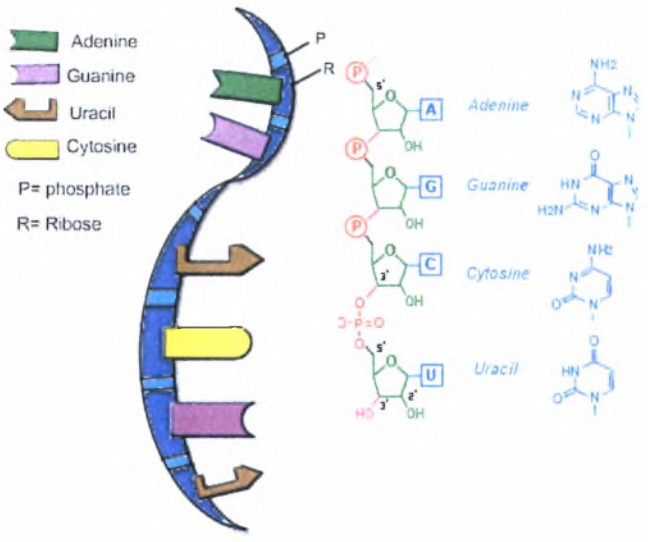
**Τα χρωμοσώματα είναι οργανωμένες δομές DNA και πρωτεϊνών και βρίσκονται μέσα στα κύτταρα.** Ένα χρωμόσωμα είναι μια μεμονωμένη αλυσίδα DNA, η οποία περιέχει γονίδια (genes), ρυθμιστικά στοιχεία και νουκλεοτιδικές ακολουθίες. **Το σύνολο των γενετικών πληροφοριών που κληρονομεί ένας έμβιος οργανισμός στους απογόνους του ονομάζεται γονιδίωμα (genome).** Για παράδειγμα, το γονιδίωμα του ανθρώπου είναι το σύνολο των γενετικών πληροφοριών που λαμβάνουμε από τα 23 χρωμοσώματα που έχει. Το μέγεθος του γονιδιώματος μετριέται συνήθως σε **ζεύγη βάσεων (base pairs – bp).**




**Σχήμα 1.4 Απεικόνιση των δεσμών βάσεων**

### *1.2.1.2 Ριβονουκλεϊκό οξύ (RNA)*

**Το RNA (ριβονουκλεϊκό οξύ) είναι υπεύθυνο για την ορθή μεταφορά της γενετικής πληροφορίας και την αποκωδικοποίησή της από το DNA.** Οι κύριες διαφορές του RNA από το DNA, όσον αφορά τη δομή του, είναι η ύπαρξη του σακχάρου της ριβόζης αντί της δεσοξυριβόζης στο DNA και η αντικατάσταση της βάσης της θυμίνης με αυτήν της ουρακίλης (T→U). Παράλληλα, το RNA στις περισσότερες περιπτώσεις δεν βρίσκεται σε μορφή διπλής έλικας, αλλά έχει μόνο μια πολυνουκλεοτιδική αλυσίδα. Οι λειτουργίες που επιτελεί το RNA εξαρτώνται από την ειδική μορφή που αυτό παίρνει: το **αγγελιοφόρο RNA (mRNA)** είναι υπεύθυνο για τη μεταφορά του γενετικού κώδικα που έχει αντιγραφεί από το DNA, το **ριβοσωμικό RNA (rRNA)** εμπλέκεται στην διαδικασία κατασκευής των ριβοσωμάτων, ενώ το **μεταφορικό RNA (tRNA)** εμπλέκεται στη σύνθεση των αμινοξέων σε μια πρωτεϊνική αλυσίδα.

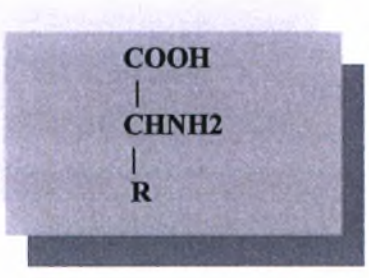



**Σχήμα 1.5 Απεικόνιση του RNA**

## *1.2.2 Οι Πρωτεΐνες*

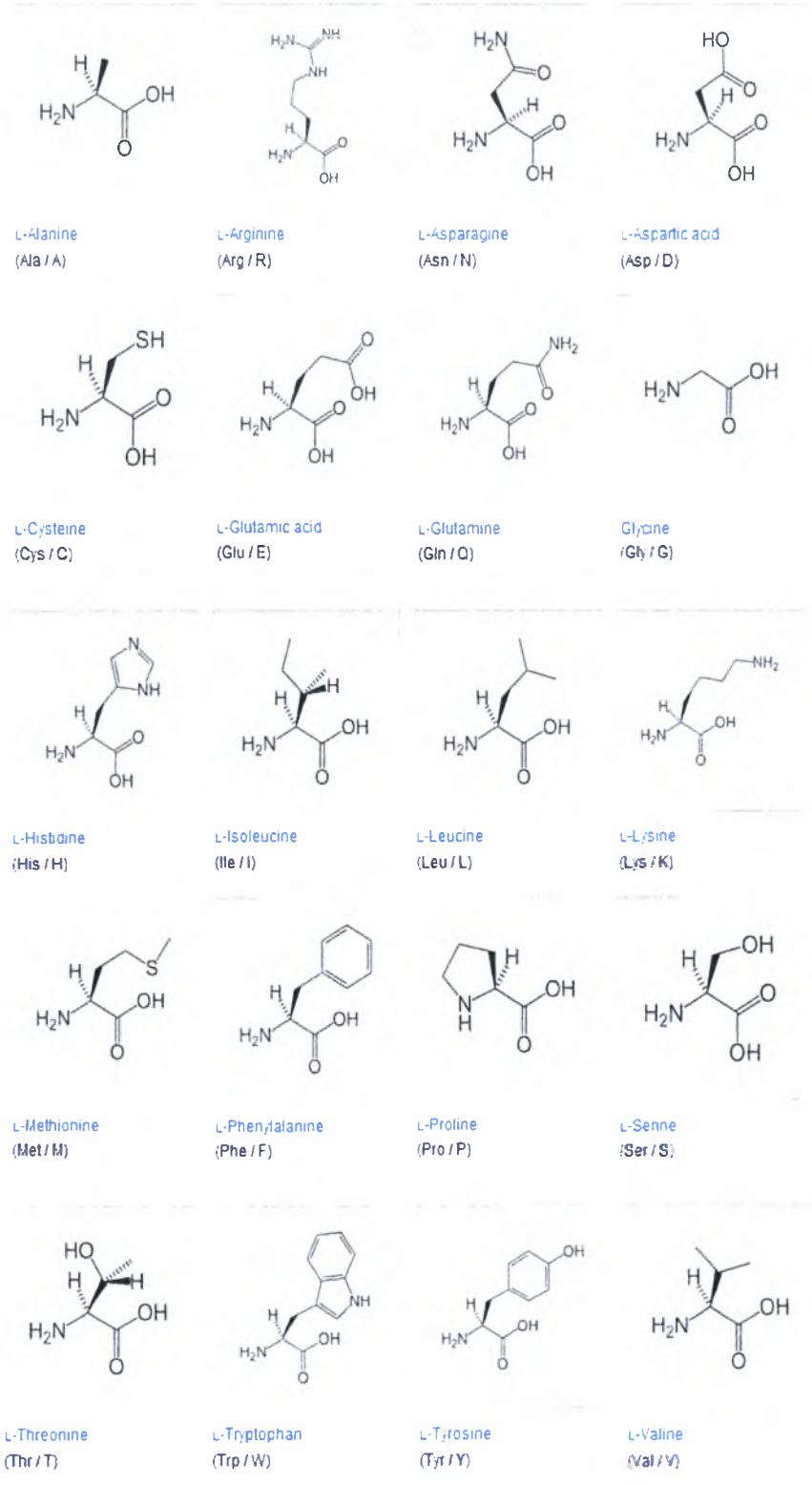
**Οι πρωτεΐνες αποτελούν τα πιο διαδεδομένα και πολυδιάστατα τόσο στη μορφή όσο και στη λειτουργικότητά τους μακρομόρια.** Ακόμη και σε ένα απλό κύτταρο των βακτηρίων εντοπίζονται εκατοντάδες διαφορετικές πρωτεΐνες που κάθε μια εξ αυτών έχει ιδιαίτερο ρόλο. Οι πρωτεΐνες αποτελούν είτε το δομικό συστατικό του κυττάρου είτε συνεργούν σε κάποια συγκεκριμένη λειτουργία του.

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.




**Σχήμα 1.6 Δομή της πρωτεΐνης**

Κάθε αμινοξύ περιγραφικά διακρίνεται από δύο τμήματα ένα "σταθερό", που αποτελείται από ένα άτομο υδρογόνου, μια αμινομάδα και μια καρβοξυλομάδα (που φέρονται ενωμένα σε κοινό άτομο άνθρακα), και ένα "μεταβλητό" που αποτελείται από την "πλευρική ομάδα R". Σημειώνεται πως η "πλευρική ομάδα R" έχει διαφορετική χημική δομή σε κάθε αμινοξύ.



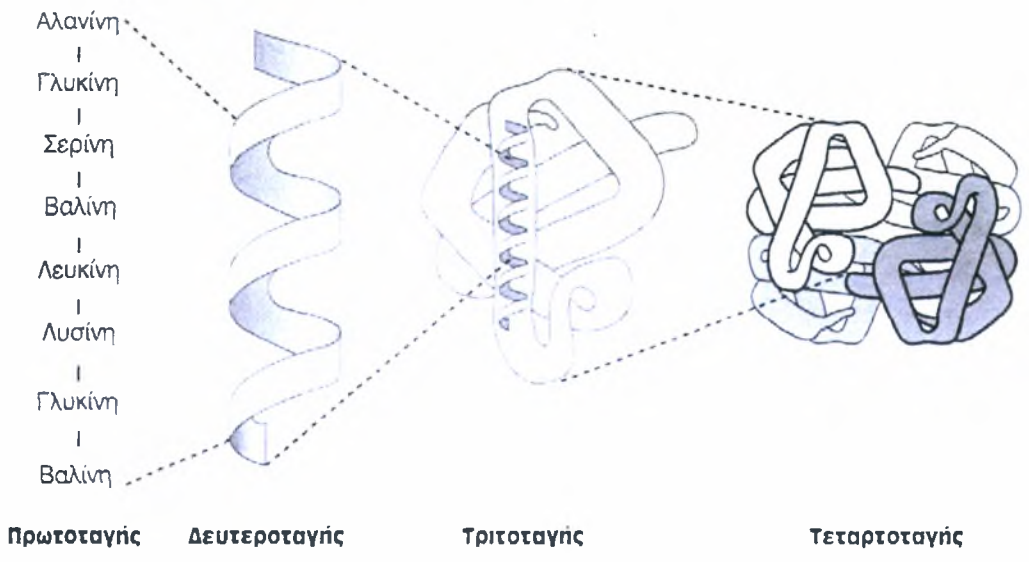

**Σχήμα 1.7 Απεικόνιση των 20 αμινοξέων**

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.

Οι πρωτεΐνες παράγονται στο κυτόπλασμα και συγκεκριμένα στα ριβοσώματα όπου ξεκινούν ως απλές μη διακλαδωμένες αλληλουχίες αμινοξέων, σχηματίζοντας την "πρωτοταγή δομή", για την οποία καθοριστικοί παράγοντες είναι τα νουκλεϊκά οξέα. Στη συνέχεια όλα τα πρωτεϊνικά μόρια υφίστανται μια φυσική αναδιάταξη προκειμένου να δώσουν μια "δευτεροταγή δομή" η οποία προκαλείται από δεσμούς υδρογόνου μεταξύ των καρβοξυλομάδων και των αμινομάδων των αμινοξέων. Κατά τη δευτεροταγή δομή δε λαμβάνονται υπ' όψιν οι αλληλεπιδράσεις μεταξύ των πλευρικών ομάδων των αμινοξέων. Ο πλέον διαδεδομένος τύπος τέτοιας μορφής είναι η λεγόμενη "α-έλικα", δεξιόστροφη, όπου οι σπείρες διατηρούνται στη θέση τους με δεσμούς υδρογόνου μεταξύ των καρβοξυλομάδων και αμινομάδων.

Μια άλλη δευτεροταγής δομή είναι η λεγόμενη "β-πτυχωτή επιφάνεια" όπου στη περίπτωση αυτή διασταυρώνονται παράλληλες αλυσίδες πολυπεπτιδίων που ενώνονται στις διασταυρώσεις με δεσμούς υδρογόνου σχηματίζοντας έτσι μια εξαιρετικά σφιχτή δομή. Οι πρωτεΐνες με τέτοιες σχετικά απλές δισδιάστατες δευτερογενείς δομές ονομάζονται γενικά **ινώδεις πρωτεΐνες**. Επιπλέον οι πρωτεΐνες υφίστανται ένα ακόμα ποιο περίπλοκο δίπλωμα το οποίο καλείται "τριτοταγής δομή". Με τον όρο τριτοταγή δομή, εννοούμε το τελικό και λειτουργικό σχήμα που αποκτά η πρωτεΐνη μετά κι από την αλληλεπίδραση των πλευρικών ομάδων των αμινοξέων. Τέλος, υπάρχουν και πρωτεΐνες που αποτελούνται από πολλές πολυπεπτιδικές αλυσίδες που είναι χαλαρά ενωμένες και αυτό αποτελεί τη λεγόμενη "τεταρτοταγή δομή".

**Σχήμα 1.8 Αναπαράσταση της τρισδιάστατης δομής της μυογλοβίνης, που παρουσιάζεται με χρωματισμένες τις άλφα έλικες.**



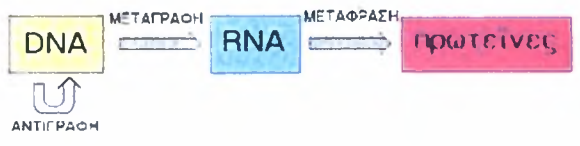

**Σχήμαα 1.9 Στάδια δημιουργίας τεταρτοταγούς δομής της πρωτεΐνης**

## *1.3 Βασικές Βιολογικές Λειτουργίες*

Το **κεντρικό δόγμα της μοριακής βιολογίας** διατυπώθηκε αρχικά από Francis Crick το 1958. Τα πρώτα χρόνια χρήσης του όρου, οι ερευνητές γνώριζαν τους εξής πιθανούς τρόπους ροής της γενετικής πληροφορίας:

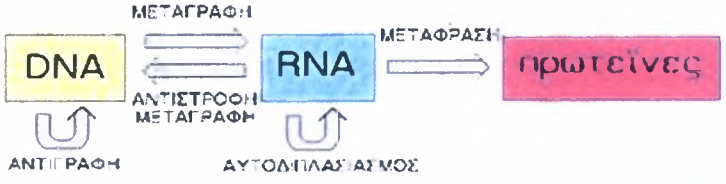
- Αντιγραφή DNA.
- Μεταγραφή, μεταφορά δηλαδή της γενετικής πληροφορίας από μορφή DNA σε μορφή αγγελιοφόρου RNA (mRNA).
- Μετάφραση, έκφραση δηλαδή της πληροφορίας στη γλώσσα των αμινοξέων, με βάση το γενετικό κώδικα.

Κατά συνέπεια, το κεντρικό δόγμα της Μοριακής Βιολογίας περιγραφόταν από το παρακάτω διάγραμμα.

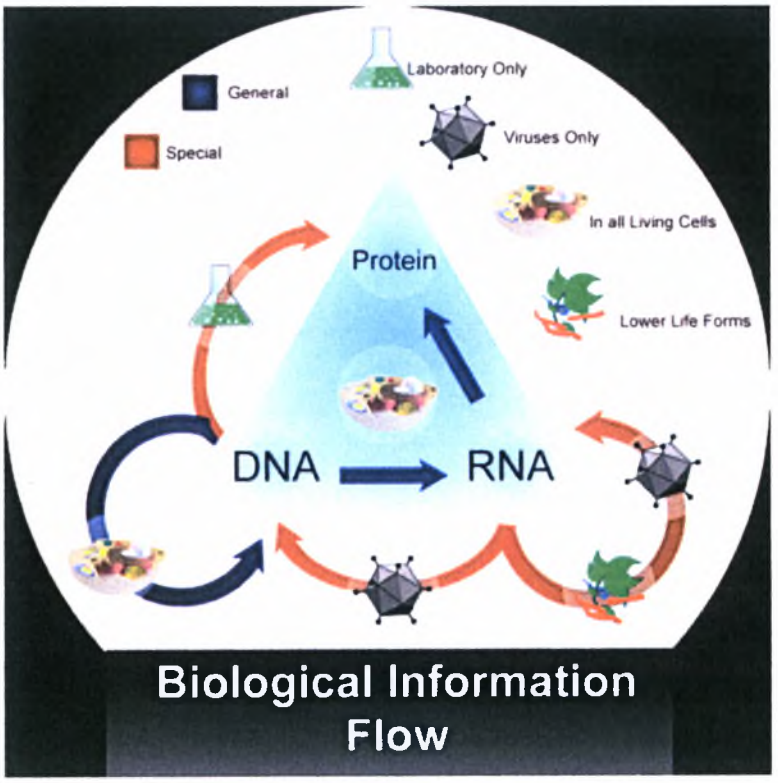



**Σχήμα 1.10 Το αρχικό διάγραμμα του κεντρικού δόγματος**

Παρατηρούμε ότι το βέλος στο διάγραμμα (δηλαδή η γενετική πληροφορία) πηγαίνει μόνο προς μία κατεύθυνση. Η κύρια εξαίρεση στο κεντρικό δόγμα της Βιολογίας όπως έχει προκύψει σήμερα είναι μια διαδικασία γνωστή σαν αντίστροφη μεταγραφή, κατά την οποία κωδικοποιημένη πληροφορία που υπάρχει στο RNA ορισμένων ιών μπορεί να μεταγραφεί σε DNA. Έτσι, σήμερα, το κεντρικό δόγμα της Μοριακής Βιολογίας περιγράφεται από το παρακάτω διάγραμμα.




**Σχήμα 1.11 Το τελικό διάγραμμα του κεντρικού δόγματος**




**Σχήμα 1.12 Απεικόνιση της ροής της βιολογικής πληροφορίας**

Σύμφωνα με το κεντρικό δόγμα της Βιολογίας η σύνθεση των πρωτεϊνών γίνεται σε δύο στάδια:

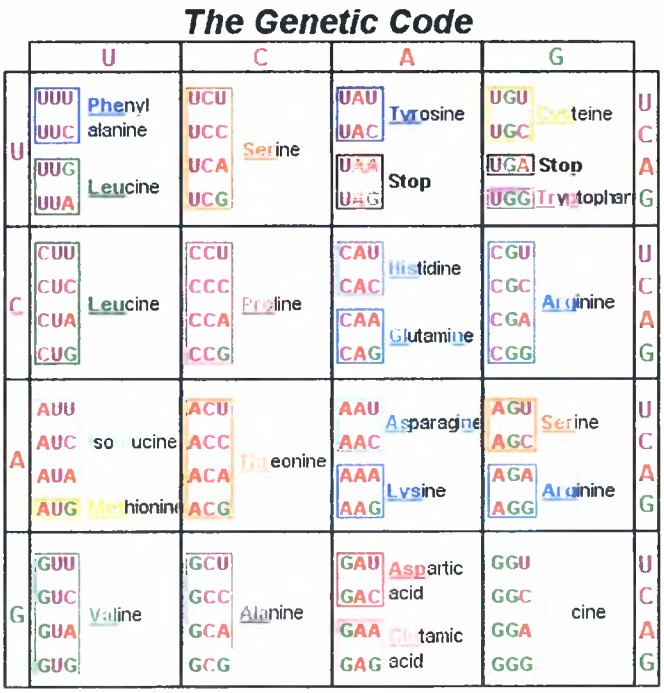
1) Δημιουργία του αγγελιοφόρου (mRNA) που είναι συμπληρωματικό ενός τμήματος του DNA με μια διαδικασία που ονομάζεται μεταγραφή (transcription).

2) Το mRNA κινείται προς το κυτταρόπλασμα (για τους ευκαρυωτικούς οργανισμούς) όπου μεταφράζεται σε μια συγκεκριμένη αλληλουχία αμινοξέων. Η διαδικασία αυτή ονομάζεται μετάφραση (translation).

Σε αντίθεση με την αντιγραφή του DNA, η οποία συμβαίνει μια μόνο φορά κατά τον κύκλο ζωής ενός κυττάρου, η μεταγραφή και η μετάφραση είναι φαινόμενα που επαναλαμβάνονται αδιάκοπα.

Το κεντρικό δόγμα καθορίζει ότι η αλληλουχία των νουκλεοτιδίων στο DNA καθώς και στο συμπληρωματικό αντίγραφο mRNA πρέπει με κάποιο τρόπο να κατευθύνει τη σωστή σειρά τοποθέτησης των αμινοξέων στην κατασκευή της πρωτεΐνης. Είναι γνωστό ότι υπάρχουν τέσσερις διαφορετικές βάσεις αλλά είκοσι διαφορετικά αμινοξέα

Γίνεται λοιπόν φανερό ότι η κωδική λέξη για την κωδικοποίηση κάθε αμινοξέος δεν θα μπορούσε να είναι ούτε μία μόνο ούτε δύο αζωτούχες βάσεις. Εάν ήταν **μία** βάση π.χ. Α ή Τ, μόνο τέσσερα διαφορετικά αμινοξέα θα μπορούσαν να κωδικοποιηθούν. Εάν ήταν **δύο** βάσεις π.χ. ΑΤ ή ΑC μόνο δεκαέξι διαφορετικά αμινοξέα θα μπορούσαν να κωδικοποιηθούν . Εάν όμως η κωδικοποίηση γινόταν με συνδυασμό **τριών** αζωτούχων ενώσεων όπως π.χ. ACT, τότε θα υπήρχαν 64 πιθανά αμινοξέα που θα μπορούσαν να κωδικοποιηθούν από το συνδυασμό των 4 αζωτούχων βάσεων ανά τρεις . Τα είκοσι, επομένως, αμινοξέα που χρησιμοποιούνται στην κατασκευή των πρωτεϊνών θα μπορούσαν εύκολα να κωδικοποιηθούν με τη χρήση ενός τέτοιου κώδικα βασισμένου σε τριπλέτες βάσεων (triplets).




**Σχήμα 1.13 Απεικόνιση του γενετικού κώδικα**

Μια προσεκτική εξέταση του πίνακα των τριπλετών (κωδικωνίων - codons) επιτρέπει να εξαγάγουμε τα ακόλουθα σημαντικά συμπεράσματα που σχετίζονται με το γενετικό κώδικα:

1. **Ο γενετικός κώδικας είναι εκφυλισμένος** , δηλαδή, πολλές τριπλέτες μπορούν να κωδικοποιήσουν για το ίδιο αμινοξύ. Η λευκίνη, λόγου χάρη, κωδικοποιείται από έξι τριπλέτες: (UUA, UUG, CUU, CUC, CUA, CUG).
2. **Ο γενετικός κώδικας έχει μία τριπλέτα έναρξης AUG και τρεις τριπλέτες λήξης** (UUA, UAG, UGA).
3. **Η τρίτη βάση είναι λιγότερο σημαντική απ' ό,τι οι δύο πρώτες στον καθορισμό του αμινοξέος που θα κωδικοποιηθεί.**
4. Ο γενετικός κώδικας που περιγράφηκε έχει σχεδόν παγκόσμια εφαρμογή στους ζωντανούς οργανισμούς.

### *1.3.1 Η Διαδικασία της Αντιγραφής του DNA (DNA replication)*

Η **αντιγραφή του DNA** είναι η διαδικασία κατά την οποία το DNA αυτοδιπλασιάζεται προκειμένου να διατηρήσει και να μεταβιβάσει τη γενετική πληροφορία από κύτταρο σε κύτταρο.

Για να αντιγραφεί το μόριο DNA πρέπει πρώτα να ξεχωρίσουν οι δύο αλυσίδες του. Το ένζυμο **ελικάση** χρησιμεύει για να ξετυλιχθεί η διπλή έλικα του μορίου σπάζοντας τους ασθενείς υδρογονικούς δεσμούς μεταξύ των δύο συμπληρωματικών αλυσίδων.

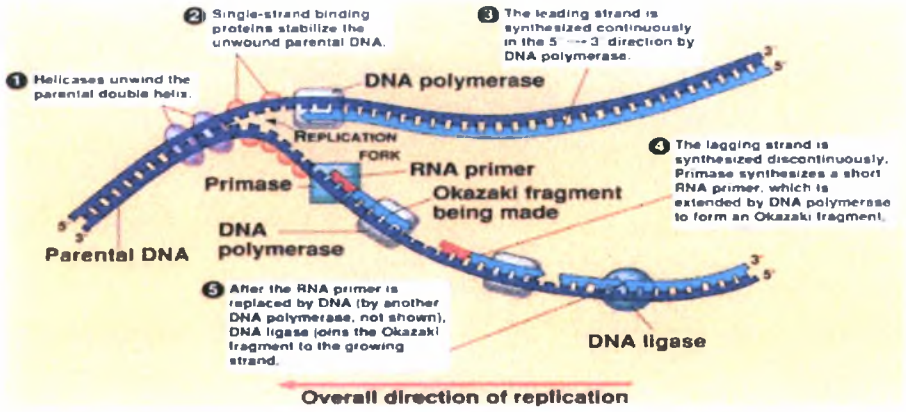
Με τον τρόπο αυτό σχηματίζονται δύο ξεχωριστές πολυνουκλεοτιδικές αλυσίδες που, στη συνέχεια, χρησιμεύουν η καθεμία για το σχηματισμό μίας συμπληρωματικής της.

Ελεύθερα νουκλεοτίδια του πυρήνα τοποθετούνται διαδοχικά απέναντι από κάθε νουκλεοτίδιο καθεμίας από τις δύο αλυσίδες, με βάση τον κανόνα της συμπληρωματικότητας των βάσεων , με τη βοήθεια του ενζύμου DNA- πολυμεράση.

Το ένζυμο αυτό χρησιμεύει επίσης και για τη συνένωση των διαδοχικών νουκλεοτιδίων με φωσφοδιεστερικούς δεσμούς με τρόπο ώστε να σχηματίσουν τελικά δύο δίκλωνα μόρια DNA, που είναι πανομοιότυπα , τόσο μεταξύ τους, όσο και με το αρχικό μόριο από το οποίο προήλθαν.

Επειδή στα μόρια αυτά μία μόνο αλυσίδα είναι καινούρια, ενώ η άλλη προέρχεται από το αρχικό μόριο DNA (καλούπι), ο συγκεκριμένος τρόπος αντιγραφής του DNA ονομάζεται **συντηρητικός.**

Η αντιγραφή του DNA γίνεται μια φορά κατά τον κύκλο της ζωής του κυττάρου. Στα ευκαρυωτικά κύτταρα γίνεται πριν από τη μίτωση ή τη μείωση, ώστε το κύτταρο να κληροδοτήσει στα θυγατρικά του το σωστό αριθμό χρωμοσωμάτων. Η όλη διαδικασία διπλασιασμού του γενετικού υλικού χαρακτηρίζεται από μεγάλη ταχύτητα, που στα μεν προκαρυωτικά κύτταρα φθάνει στις 500 βάσεις το δευτερόλεπτο, ενώ στα ευκαρυωτικά κύτταρα του ανθρώπου ο ρυθμός αντιγραφής δεν ξεπερνά τις 50 βάσεις το δευτερόλεπτο.



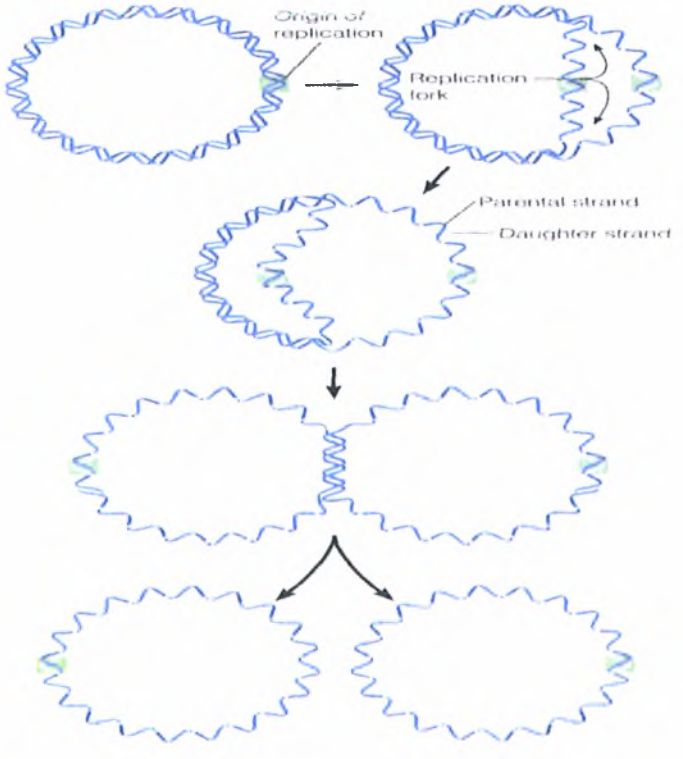

**Σχήμα 1.14 Σύνοψη της αντιγραφής του DNA**

**Σχήμα 1.15 Απεικόνιση των 2 πανομοιότυπων αλυσίδων που προκύπτουν**

Ειδικά για τα προκαρυωτικά κύτταρα, δεδομένου ότι το μόριο του DNA είναι κυκλικό, η αντιγραφή του DNA έχει ένα συγκεκριμένο σημείο εκκίνησης ή, θα λέγαμε, μια συγκεκριμένη ακολουθία DNA (**origin of replication**). Με βάση τα

πρώτα πειράματα που πραγματοποιήθηκαν για να μελετηθεί ο τρόπος αντιγραφής του DNA, παρατηρήθηκε ότι από συγκεκριμένα σημεία των μορίων DNA δημιουργούνταν **διχάλες αντιγραφής (replication forks)**, οι οποίες κινούνται κατά μήκος της ακολουθίας του DNA, «ξετυλίγοντας» και αντιγράφοντας την ακολουθία της αλυσίδας του DNA, σε αντίθετη κατεύθυνση η μία από την άλλη. Ειδικά για τους οργανισμούς με κυκλικό DNA, η διαδικασία της αντιγραφής του DNA ονομάζεται **«αντιγραφή θήτα» (theta replication)**, καθώς, όπως φαίνεται και στο σχήμα, κατά την διάρκεια της αντιγραφής δημιουργούνται μορφές παρόμοιες με το σχήμα του ελληνικού γράμματος θ. Η αντίστοιχη διαδικασία είναι περίπλοκη σε σχέση με αυτή των προκαρυωτικών κυττάρων. Σε αυτήν την περίπτωση το σημείο εκκίνησης της αντιγραφής του DNA δεν είναι συγκεκριμένο και δημιουργεί πολλές μονάδες αντιγραφής, τα **ρεπλικόνια (replicons)**. Τα ρεπλικόνια καθορίζουν τα σημεία στα οποία δημιουργούνται και σταδιακά αναπτύσσονται οι **φυσαλίδες αντιγραφής (replication bubbles)**, οποίες συγχωνεύονται, με αποτέλεσμα να δημιουργείται η νέα ακολουθία DNA.

Η παραπάνω περιγραφή συνοψίζει την συγκεκριμένη διαδικασία :




**Σχήμα 1.16 Απεικόνιση της αντιγραφής σε κυκλικό μόριο DNA**

### *1.3.2 Η Διαδικασία της Μεταγραφής του DNA (DNA transcription)*

Το πρώτο στάδιο για την έκφραση της γενετικής πληροφορίας υλοποιείται με τη σύνθεση του mRNA. Χρησιμοποιούμε για τη διαδικασία αυτή τον όρο «μεταγραφή», διότι η γενετική πληροφορία , που ήταν καταγραμμένη στη γλώσσα DNA, περνά στο RNA, που χρησιμοποιεί την ίδια γλώσσα, δηλαδή τη γλώσσα των νουκλεοτιδίων. Κάθε mRNA είναι αντίγραφο ενός γονιδίου, δηλαδή ενός τμήματος μιας από τις δύο αλυσίδες του DNA. Προκύπτει με μια διαδικασία αντίστοιχη της αντιγραφής, κατά ην οποία τα ριβονουκλεοτίδια που χρησιμοποιούνται υπακούουν στον ίδιο κανόνα της συμπληρωματικότητας των βάσεων. Η μόνη διαφορά είναι ότι η βάση ουρακίλη αντικαθιστά τη θυμίνη στο ζευγάρωμα με την αδενίνη. Το μόριο που φτιάχνεται με τον τρόπο αυτό περιέχει μια συγκεκριμένη αλληλουχία νουκλεοτιδίων , τα οποία χρησιμεύουν για να κατευθύνουν τη σειρά τοποθέτησης των αμινοξέων και το σχηματισμό της πρωτεΐνης. Συγκεκριμένα:

1. **Ξετυλίγεται** η διπλή έλικα του DNA στην περιοχή του γονιδίου που πρόκειται να γίνει η μεταγραφή και ανοίγει με το σπάσιμο των δεσμών υδρογόνου που συγκρατούν τις δύο αλυσίδες.

2. **Η RNA πολυμεράση** (ένζυμο της μεταγραφής) **συνδέεται με τον υποκινητή**, μια μικρή αλληλουχία νουκλεοτιδίων πάνω στο DNA που δρα ως σημείο εκκίνησης. Η σειρά των βάσεων της αλληλουχίας αυτής καθορίζει επίσης ποια από τις δύο αλυσίδες του DNA θα χρησιμεύσει σαν καλούπι για τη σύνθεση του RNA. Η αλυσίδα του DNA που θα μεταγραφεί ονομάζεται «μεταγραφόμενη», ενώ η άλλη «κωδική αλυσίδα».

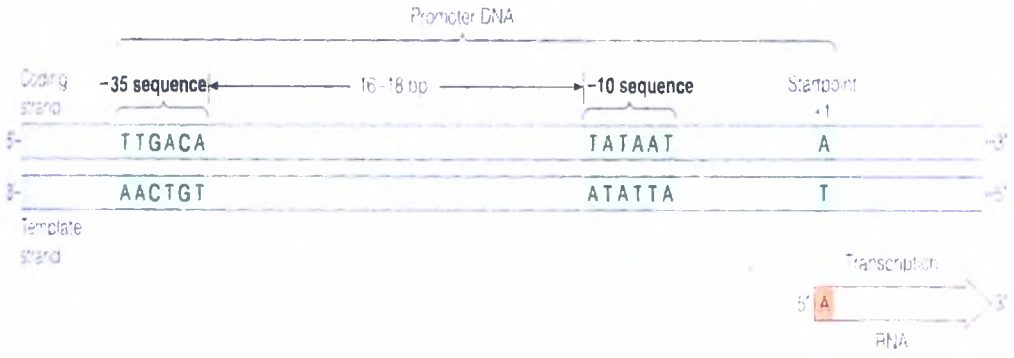
3. **Η έναρξη** της δημιουργίας της RNA αλυσίδας σηματοδοτείται από τη συνένωση των δύο πρώτων ριβονουκλεοτιδίων (η περιοχή του υποκινητή δεν μεταγράφεται).

4. **Η επιμήκυνση της αλυσίδας** γίνεται με την RNA πολυμεράση να τοποθετεί διαδοχικά τα συμπληρωματικά ριβονουκλεοτίδια απέναντι από κάθε νουκλεοτίδιο του DNA και να τα ενώνει μεταξύ τους με φωσφοδιεστερικό δεσμό. Καθώς η αλυσίδα του mRNA επιμηκύνεται, αποσυνδέεται από την αλυσίδα του DNA και το μόριο του DNA ξανατυλίγεται. Το μόριο RNA που δημιουργείται είναι μεν συμπληρωματικό με την αλυσίδα από την οποία μεταγράφηκε, αλλά ταυτόχρονα πανομοιότυπο με την αλυσίδα του DNA που δεν μεταγράφηκε (με εξαίρεση ότι η θυμίνη έχει αντικατασταθεί από την ουρακίλη).

5. **Λήξη.** Η RNA πολυμεράση αντιλαμβάνεται ένα μήνυμα τερματισμού από μια ειδική αλληλουχία βάσεων και με τον τρόπο αυτό το ολοκληρωμένο μόριο mRNA ελευθερώνεται πλήρως από το καλούπι του DNA.


**Σχήμα 1.17 Η διαδικασία της μεταγραφής του DNA για έναν προκαρυωτικό οργανισμό**

Μετά από μια σειρά ερευνών και καθορισμού των ακολουθιών των υποκινητών με κύρια μέθοδο αυτή του DNA footprinting, παρατηρήθηκε ότι οι υποκινητές διαφέρουν χαρακτηριστικά μεταξύ τους. Εύλογα, το ερώτημα που τέθηκε ήταν με ποιόν τρόπο η RNA-πολυμεράση καταφέρνει να αναγνωρίσει το σύνολο αυτών των διαφορετικών μεταξύ τους ακολουθιών. Η πιο πρόσφατη απάντηση στο ερώτημα αυτό είναι ότι η αναγνώριση των ακολουθιών αυτών γίνεται όχι σε όλο το μήκος του υποκινητή, αλλά σε πολύ μικρές στο μήκος υποακολουθίες του. Οι υποακολουθίες αυτές βρίσκονται σε συγκεκριμένες θέσεις σε κάθε περιοχή υποκινητή. Ένας τυπικός προκαρυωτικός υποκινητής παρουσιάζεται στο παρακάτω σχήμα:




**Σχήμα 1.18 Ένας τυπικός προκαρυωτικός υποκινητής**

## *1.3.3 Η Διαδικασία της Μετάφρασης του DNA (DNA translation)*

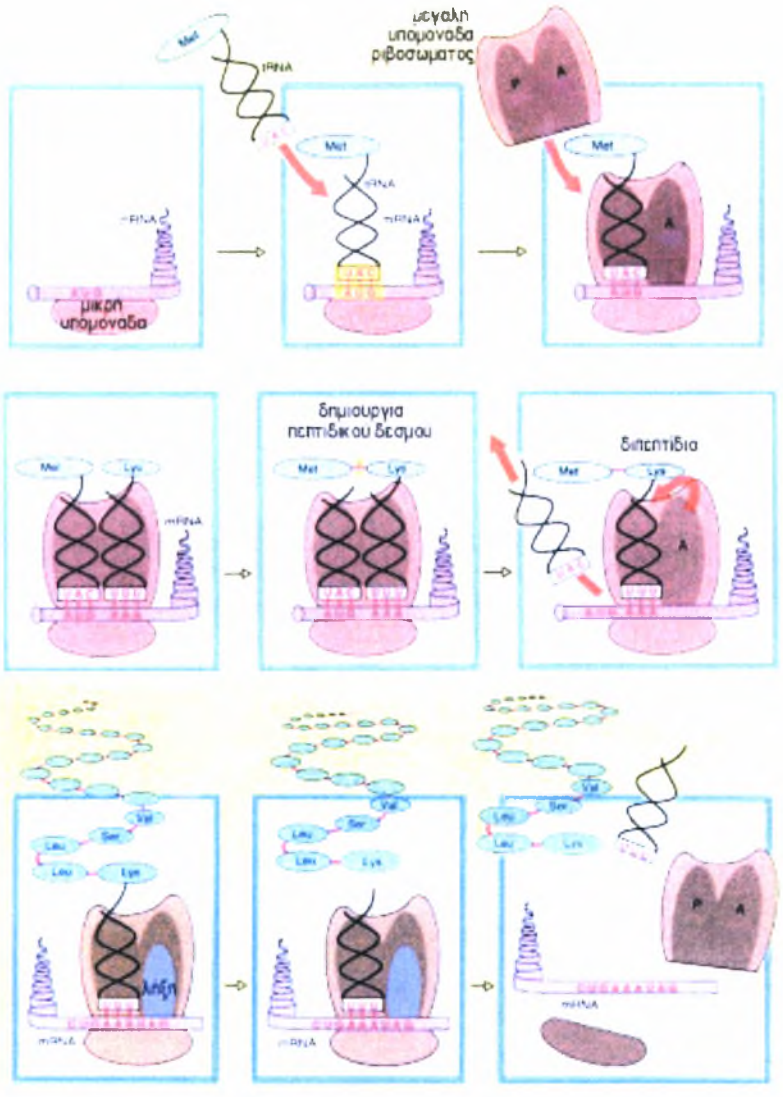
Το τελευταίο στάδιο για την έκφραση της γενετικής πληροφορίας υλοποιείται με τη μετάφραση της ,δηλαδή , με την παραγωγή του πρωτεϊνικού προϊόντος. Χρησιμοποιούμε για τη διαδικασία αυτή τον όρο «μετάφραση», διότι ενώ η γενετική πληροφορία ήταν μέχρι τώρα καταγραμμένη στη γλώσσα των νουκλεϊκών οξέων και χρησιμοποιούσε ένα αλφάβητο 4 γραμμάτων, τώρα πρέπει να μεταφραστεί στη γλώσσα των πρωτεϊνών που χρησιμοποιεί ένα αλφάβητο 20 γραμμάτων.

Η διαδικασία της μετάφρασης πρέπει να γίνει με ένα πολύ δομημένο και μεθοδικό τρόπο ώστε να διασφαλιστεί η αποφυγή λαθών στην τελική αλληλουχία αμινοξέων που θα παραχθεί. Τα σημαντικότερα στάδια αυτής της διαδικασίας είναι η έναρξη, η επιμήκυνση και ο τερματισμός.

1. Έναρξη της μετάφρασης γίνεται, όταν η μικρή υπομονάδα του ριβοσώματος ενώνεται με το mRNA στην περιοχή του κωδικονίου έναρξης (AUG) . Το πρώτο tRNA, που ονομάζεται **εναρκτής,** που έχει σαν αντικωδικόνιο την τριπλέτα UAC, αναγνωρίζει το κωδικόνιο AUG χάρη στη συμπληρωματικότητα των βάσεων και ζευγαρώνει μαζί του. Αυτό το αρχικό tRNA μεταφέρει μια τροποποιημένη μορφή του αμινοξέος μεθειονίνη. Όλες οι πολυπεπτιδικές αλυσίδες έχουν, δηλαδή, ως πρώτο αμινοξύ αυτή την τροποποιημένη μεθειονίνη. Το αμινοξύ αυτό αργότερα , κατά την επεξεργασία της πρωτεΐνης μπορεί να αποχωριστεί. Η μεγάλη ριβοσωμικής υπομονάδα συνδέεται με το πρώτο tRNA στη P θέση (πεπτιδική θέση).

2. Επιμήκυνση της αλυσίδας. Ένα δεύτερο σύμπλεγμα αμινοξέος tRNA με αντικωδικόνιο που είναι συμπληρωματικό με το δεύτερο κατά σειρά κωδικόνιο, συνδέεται στη δεύτερη θέση Α της μεγάλης υπομονάδας του ριβοσώματος με τον ίδιο τρόπο (συμπληρωματικότητα κωδικονίου - αντικωδικονίου). Αν το tRNA αυτό δεν είναι σωστό, τότε απομακρύνεται και μέσα από μια διαδικασία δοκιμής και σφάλματος επιλέγεται το σωστό. Όταν συμπληρωθούν και οι δύο θέσεις του ριβοσώματος (Α και P), δημιουργείται ένας πεπτιδικός δεσμός μεταξύ των δύο αμινοξέων. Ταυτόχρονα ο δεσμός που συνδέει το πρώτο αμινοξύ με το tRNA του σπάζει, προσφέροντας την απαιτούμενη ενέργεια για τη δημιουργία του πεπτιδικού δεσμού στο

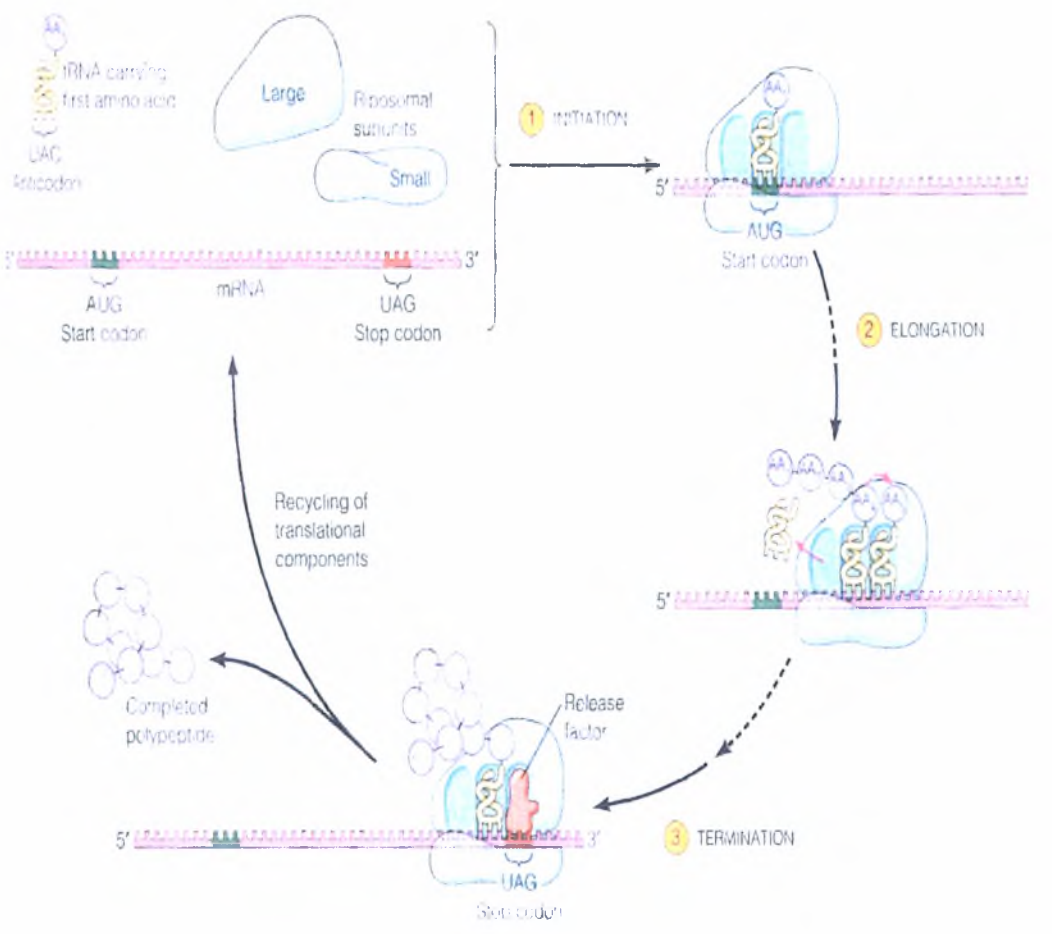
σχηματιζόμενο διπεπτίδιο. Στη συνέχεια, το πρώτο απελευθερώνεται στο κυτταρόπλασμα. Το διπεπτίδιο είναι τώρα ενωμένο με το δεύτερο tRNA.




**Σχήμα 1.19 Διαδικασία μετάφρασης του DNA**

Στη συνέχεια , το ριβόσωμα μετακινείται κατά μήκος του μορίου mRNA προς το επόμενο κωδικόνιο, έτσι ώστε, το δεύτερο tRNA να καταλάβει την P θέση, ελευθερώνοντας με τον τρόπο αυτό τη θέση Α. Ένα τρίτο σύμπλεγμα αμινοξύ - tRNA κινείται προς την ελεύθερη θέση Α. Με τη διαδικασία «δοκιμή και σφάλμα» επιλέγεται το σωστό και η διαδικασία της δημιουργίας πεπτιδικού δεσμού με το τρίτο αμινοξύ επαναλαμβάνεται. Καθώς το ριβόσωμα κινείται διαδοχικά σε όλο το μήκος της αλυσίδας του mRNA, κωδικόνιο προς κωδικόνιο, τοποθετείται με τον ίδιο τρόπο ένα καινούριο σύμπλεγμα αμινοξέος-tRNA κάθε φορά στη θέση Α απέναντι από το αντίστοιχο κωδικόνιο, και επομένως ένα επιπλέον αμινοξύ προστίθεται κάθε φορά στο σχηματιζόμενο πολυπεπτίδιο. Το πολυπεπτίδιο αυτό είναι πάντα ενωμένο στο τελευταίο tRNA .

3. Ο τερματισμός της αλυσίδας επέρχεται, όταν το ριβόσωμα αντιληφθεί ένα σήμα λήξης με τη μορφή ενός κωδικονίου λήξης (UAG, UAA και UGA). Κανένα tRNA δεν θα τοποθετηθεί στην περιοχή Α, μια που κανένα tRNA δεν φέρει αντικωδικόνιο συμπληρωματικό με αυτές τις αλληλουχίες. Το πολυπεπτίδιο βγαίνει από το ριβόσωμα και οι δύο ριβοσωμικές υπομονάδες διαχωρίζονται.




**Σχήμα 1.20 Τα τρία βήματα με τα οποία ολοκληρώνεται η μετάφραση σε ένα κύτταρο**

Η σύνθεση της πολυπεπτιδικής αλυσίδας δε σημαίνει ταυτόχρονα και δημιουργία ολοκληρωμένης πρωτεΐνης. Όπως έχει ήδη αναφερθεί πολλές από τις πολυπεπτιδικές αλυσίδες, προκειμένου να αποτελέσουν ή να συμμετάσχουν στη δημιουργία ενός λειτουργικού πρωτεϊνικού μορίου, πρέπει πρώτα να υποστούν κάποια ενζυμική επεξεργασία ή προσθήκες άλλων μορίων. Αυτές γίνονται στα κατάλληλα οργανίδια του κυττάρου. Παρακάτω αναφέρονται διάφορες λειτουργίες του οργανισμού όπου συμβάλλουν σημαντικά η πρωτεΐνες:

- **Ενζυμική κατάλυση.**
  Σχεδόν όλες οι χημικές αντιδράσεις στα βιολογικά συστήματα καταλύονται από ειδικά μακρομόρια που ονομάζονται ένζυμα. Μερικές από

τις αντιδράσεις αυτές είναι αρκετά απλές. Άλλες, όπως η αντιγραφή ενός ολοκλήρου χρωμοσώματος, είναι αρκετά πολύπλοκες. Σχεδόν όλα τα ένζυμα επιδεικνύουν τεράστια καταλυτική ικανότητα. Το αξιοσημείωτο είναι ότι όλα τα γνωστά ένζυμα είναι πρωτεΐνες. Έτσι, οι πρωτεΐνες διαδραματίζουν τον μοναδικό ρόλο του καθορισμού της πορείας των χημικών αντιδράσεων που πραγματοποιούνται στα βιολογικά συστήματα.

- **Μεταφορά και αποθήκευση.**

Πολλά μικρά μόρια και ιόντα μεταφέρονται από ειδικές πρωτεΐνες. Για παράδειγμα, η αιμοσφαιρίνη μεταφέρει οξυγόνο στα ερυθροκύτταρα, ενώ η μυοσφαιρίνη στους μυς. Σίδηρος μεταφέρεται στο πλάσμα του αίματος από την τρανσφερρίνη και αποθηκεύεται στο συκώτι ως σύμπλοκο μαζί με τη φερριτίνη, μία διαφορετική πρωτεΐνη.

- **Κίνηση.**

Οι πρωτεΐνες είναι τα κύρια συστατικά των μυών. Η συστολή των μυών επιτυγχάνεται με την ολισθητική κίνηση δύο ειδών πρωτεϊνικών νηματίων.

- **Μηχανική στήριξη.**

Η μεγάλη αντοχή του δέρματος και των οστών οφείλεται στην παρουσία του κολλαγόνου, μιας πρωτεΐνης που σχηματίζει ίνες.

- **Ανοσοπροστασία.**

Τα αντισώματα είναι πολύ ειδικές πρωτεΐνες που αναγνωρίζουν και συνενώνονται με ξένες ουσίες όπως οι ιοί, τα βακτήρια και κύτταρα από άλλους οργανισμούς. Ο ζωτικός ρόλος, επομένως, των πρωτεϊνών είναι προφανής.

- **Δημιουργία και μετάδοση νευρικών παλμών.**

Η απόκριση των νευρικών κυττάρων σε ειδικά ερεθίσματα υποβοηθείται από πρωτεΐνες-υποδοχείς. Μόρια υποδοχείς που μπορούν να «διεγερθούν» με ειδικά μόρια, όπως η ακετυλοχολίνη, είναι υπεύθυνα για τη μετάδοση των νευρικών παλμών στις επαφές μεταξύ νευρικών κυττάρων.

- **Ελεγχος της ανάπτυξης και διαφοροποίησης.**

Η ελεγχόμενη έκφραση των γενετικών πληροφοριών είναι ουσιώδης για την κανονική ανάπτυξη και διαφοροποίηση των κυττάρων. Μόνο ένα μικρό κλάσμα του γονιδιώματος ενός κυττάρου εκφράζεται κάθε φορά.

# ΚΕΦΑΛΑΙΟ 2

## *2.1 Βιοπληροφορική - Υπολογιστική Βιολογία*


**Σχήμα 2.1 Διάγραμμα συσχέτισης Βιοπληροφορικής με άλλες επιστήμες**

**Η βιοπληροφορική** και η **υπολογιστική βιολογία** περιλαμβάνουν τη χρήση ή την ανάπτυξη των τεχνικών, συμπεριλαμβανομένων των εφαρμοσμένων μαθηματικών, της πληροφορικής, των στατιστικών, της τεχνητής νοημοσύνης , της χημείας και της βιοχημείας για να λύσουν τα βιολογικά προβλήματα, συνήθως στο μοριακό επίπεδο. Ο αρχικός στόχος της βιοπληροφορικής είναι να αυξηθεί η κατανόηση των βιολογικών διαδικασιών. Αυτό που την θέτει εκτός από άλλες προσεγγίσεις, εντούτοις, είναι η εστίασή της να αναπτύξει και να εφαρμόσει υπολογιστικά τις εντατικές τεχνικές (π.χ., αλγόριθμους εξόρυξης δεδομένων ,εκμάθησης μηχανών) για να επιτευχθεί αυτός ο στόχος. Σημαντικές ερευνητικές προσπάθειες στον τομέα περιλαμβάνουν την ευθυγράμμιση ακολουθίας, την εύρεση γονιδίων, τη συγκέντρωση γονιδιώματος, την πρωτεϊνική ευθυγράμμιση δομών, την πρωτεϊνική πρόβλεψη δομών, την πρόβλεψη της έκφρασης γονιδίων και των πρωτεϊνικών αλληλεπιδράσεων, και το πρότυπο της εξέλιξης.

Οι όροι βιοπληροφορική και υπολογιστική βιολογία συχνά χρησιμοποιούνται εναλλακτικά. Παρόλα αυτά ο όρος βιοπληροφορική αναφέρεται στην δημιουργία και στην ανάπτυξη αλγορίθμων, υπολογιστικών και στατιστικών δομών και στην θεωρία που χρειάζεται για την επίλυση πρακτικών προβλημάτων που προκύπτουν από την διαχείριση και την ανάλυση βιολογικών δεδομένων. Η υπολογιστική βιολογία, από την άλλη πλευρά, αναφέρεται στην υποθετική έρευνα ενός συγκεκριμένου βιολογικού προβλήματος με την χρήση υπολογιστών, που διεξάγονται με πειραματικά ή προσομοιωμένα δεδομένα, με πρωταρχικό στόχο την ανακάλυψη και την ανάπτυξη βιολογικής γνώσης. Με απλά λόγια, η βιοπληροφορική ασχολείται με τις πληροφορίες ενώ η υπολογιστική βιολογία με τις υποθέσεις.

Ένας παρόμοιος διαχωρισμός γίνεται από τα Εθνικά Ινστιτούτα Υγείας, με βάση τον δικό τους ορισμό για την βιοπληροφορική και την υπολογιστική βιολογία, δίνεται έμφαση στο γεγονός ότι υπάρχει ισχυρή συσχέτιση μεταξύ της γνώσης και των περισσότερο υποθετικών ερευνών. Η Βιοπληροφορική συχνά αναφέρεται και ως ένα εφαρμοσμένο υποσύνολο των γενικότερων αρχών της βιοϊατρικής πληροφορικής.

Ένας κοινός παράγοντας στα προγράμματα της βιοπληροφορικής και της υπολογιστικής βιολογίας είναι η χρήση μαθηματικών εργαλείων για την εξόρυξη χρήσιμων πληροφοριών από δεδομένα που παράγονται από τις βιολογικές τεχνικές όπως η γονιδιωματική αλληλουχία.

## *2.2 Κυριότεροι Τομείς Έρευνας στη Βιοπληροφορική*

Οι κυριότεροι τομείς έρευνας στη βιοπληροφορική και την υπολογιστική βιολογία είναι:

- Η ανάπτυξη εργαλείων που να επιτρέπουν την ανάλυση, σύγκριση και κατηγοριοποίηση ακολουθιών βιολογικών δεδομένων.
- Η ανάπτυξη εργαλείων που να επιτρέπουν την ερμηνεία αποτελεσμάτων βιολογικής σημασίας.
- Η αποδοτική οργάνωση των δεδομένων, ώστε να είναι δυνατή η αποθήκευση, ανάκτηση και ενημέρωσή τους.

### *2.2.1 Ανάλυση, Σύγκριση, Κατηγοριοποίηση και Ταξινόμηση ακολουθιών βιολογικών δεδομένων*

Η βασική υπόθεση για την ανάπτυξη και εφαρμογή τεχνικών διαχείρισης συμβολοσειρών βιολογικών δεδομένων είναι ότι **κάθε βιολογικό μόριο μπορεί να περιγραφεί ως μια ακολουθία συμβόλων από ένα ορισμένο αλφάβητο Σ**. Συγκεκριμένα, κάθε μόριο του DNA μπορεί να θεωρηθεί ως μια ακολουθία συμβόλων (συμβολοσειρά), από ένα αλφάβητο τεσσάρων χαρακτήρων / γραμμάτων: A,C,G,T, ενώ κάθε μόριο πρωτεΐνης μπορεί να θεωρηθεί ως μια ακολουθία συμβόλων (συμβολοσειρά) από ένα αλφάβητο είκοσι χαρακτήρων / γραμμάτων, των 20 αμινοξέων.

Κατά την ανάλυση ακολουθιών βιολογικών δεδομένων μας ενδιαφέρει είτε η ακριβής εύρεση προτύπου, είτε η προσεγγιστική εύρεση προτύπου. Με αυτόν τον τρόπο στις ακολουθίες DNA μπορούμε να προσδιορίσουμε τις περιοχές όπου βρίσκονται γονίδια, περιοχές όπου τερματίζει ή ξεκινάει η αντιγραφή του DNA κ.α. Σε πρωτεϊνικές ακολουθίες μπορούμε να καθορίσουμε εξελικτικές σχέσεις και να προβλέψουμε την δευτεροταγή ή τριτοταγή δομή τους.

Ένα από τα σημαντικότερα πεδία έρευνας αποτελεί η πολλαπλή στοίχιση ακολουθιών (multiple sequence alignment). Η μέθοδος είναι αναπόσπαστα δεμένη με την εξέλιξη (κληρονομούμενες αλλαγές πληροφορίας) ανεξάρτητα από το εάν η

εξέλιξη αποτελεί τμήμα του προβλήματος. Η στοίχιση ακολουθιών μπορεί να κατηγοριοποιηθεί σε:

1) τοπική ευθυγράμμιση (local alignment)

2) ολική ευθυγράμμιση (global alignment)

Στην τοπική ευθυγράμμιση αναζητούμε περιοχές τοπικής ομοιότητας. Ο πρώτος αλγόριθμος τοπικής ευθυγράμμισης δημιουργήθηκε από τους Smith-Waterman (SW) και αρκετές σύγχρονες τεχνικές βασίζονται σε αυτόν. Αντίστοιχα ο πρώτος αλγόριθμος ολικής ευθυγράμμισης δημιουργήθηκε από τους Needleman & Wunsch (NW). Και στις δυο περιπτώσεις υπάρχουν παραπάνω από μια δυνατές ευθυγραμμίσεις. Η βέλτιστη λύση πρέπει να ελαχιστοποιεί τις διαφορές ανάμεσα στις δυο ακολουθίες ή διαφορετικά να μεγιστοποιεί τη συνάρτηση ομοιότητας.

Το πρόβλημα εύρεσης τοπικών ευθυγραμμίσεων χρησιμοποιείται ευρέως στη σύγκριση μιας δοσμένης ακολουθίας μικρού μήκους (input query sequence) ως προς το σύνολο γνωστών ακολουθιών που αποθηκεύονται σε μια βάση δεδομένων. Παρά την μεγάλη αξία των αλγορίθμων SW και NW, εξαιτίας του ότι ανήκουν στην κατηγορία των λεγόμενων «σχολαστικών» (rigorous) αλγορίθμων, δεν ικανοποιούν από πλευράς ταχύτητας, ειδικά σε περίπτωση έρευνας βάσεων δεδομένων με εκατοντάδες χιλιάδες ή και εκατομμύρια ακολουθιών. Για αυτόν το λόγο αναπτύχθηκαν οι λεγόμενοι ευρεστικοί (heuristic) αλγόριθμοι.

Οι ευρεστικοί αλγόριθμοι χρησιμοποιούν προσεγγίσεις, οι οποίες επιτρέπουν η αναζήτηση ομόλογων ακολουθιών να γίνεται πολύ πιο γρήγορα (ομόλογες ονομάζονται οι ακολουθίες που έχουν αποκλίνει από μια κοινή προγονική ακολουθία). Τα προγράμματα που χρησιμοποιούνται πιο συχνά και βασίζονται στους ευρεστικούς αλγόριθμους, είναι τα BLAST και FASTA που αναζητούν περιοχές τοπικής ομοιότητας. Ουσιαστικά και τα δυο αυτά προγράμματα αποτελούν μια συλλογή εργαλείων ευθυγράμμισης ακολουθιών.

Το πρόγραμμα BLAST (Basic Local Alignment Search Tool) δημιουργήθηκε από τον Altschul το 1990 και βασίζεται στην κεντρική ιδέα της εύρεσης κοινών υποακολουθιών ίδιου μήκους (segment pairs) που εμφανίζονται και στη δοσμένη ακολουθία μικρού μήκους (input query sequence) και στο σύνολο των ακολουθιών μιας βάσης δεδομένων με βάση μια συγκεκριμένη συνάρτηση ομοιότητας (scoring threshold).

Ο αλγόριθμος FASTA δημιουργήθηκε από τους Lipman & Pearson το 1985 και βασίζεται στην κεντρική ιδέα της αναζήτησης μικρών λέξεων (words ή k-tuples) που εμφανίζονται και στις δυο ακολουθίες. Στην περίπτωση πρωτεϊνικών ακολουθιών το μήκος των λέξεων είναι 1-2 κατάλοιπα, ενώ για ακολουθίες DNA το μήκος μιας λέξης μπορεί να φθάνει τις 6 βάσεις. Ο αλγόριθμος χρησιμοποιεί ευρεστικές μεθόδους για να δημιουργήσει περιοχές που περιέχουν κοινές λέξεις. Η στοίχιση που προκύπτει περιλαμβάνει διαφορές ανάμεσα σε κοινές λέξεις.

Εκτός από τα δυο προγράμματα που αναφέραμε παραπάνω και τα οποία χρησιμοποιούνται ευρύτατα, μια σειρά από άλλες μεθόδους και αλγόριθμους βρίσκουν εφαρμογή στην ανάλυση και σύγκριση βιολογικών ακολουθιών. Πολλοί σχετικά πρόσφατοι αλγόριθμοι βρίσκουν εφαρμογή σε προβλήματα βιοπληροφορικής

(ενδεικτικά αναφέρουμε τους αλγόριθμους Boyer-Moore, Knuth-Morris-Prat για προβλήματα ακριβούς εύρεσης προτύπου). Μεγάλη εφαρμογή το τελευταίο διάστημα έχουν και τα δέντρα επιθεμάτων (suffix tree και generalized suffix tree) με τα οποία θα ασχοληθούμε στο επόμενο κεφάλαιο.

Πέρα από την ανάλυση και σύγκριση ακολουθιών, η κατηγοριοποίηση αποτελεί ένα ευρύ πεδίο έρευνας στην βιοπληροφορική και την υπολογιστική βιολογία. Οι τεχνικές ανάλυσης συστάδων (clustering) αποτελούν μια στατιστική διαδικασία πολλών μεταβλητών, η οποία, ξεκινώντας από ένα σύνολο δεδομένων, επιχειρεί να το οργανώσει σε ομάδες ομοειδών στοιχείων που ονομάζουμε συστάδες (clusters). Οι ομάδες αυτές δεν είναι εκ των προτέρων γνωστές, αλλά προκύπτουν δυναμικά. Μια σειρά από μεθόδους κατηγοριοποίησης χρησιμοποιούνται, οι οποίες θα μπορούσαν να καταταχθούν στις/στα :

1. **Ιεραρχικές μεθόδους (hierarchical methods)**

2. **Διαιρετικές μεθόδους (partitioning methods)**

3. **Γραφοθεωρητικές μεθόδους (graph-based methods)**

4. **Τεχνητά νευρωνικά δίκτυα (artificial neural networks)**

5. **Εξελικτικές μεθόδους (genetic algorithms)**

6. **Μηχανές υποστήριξης διανύσματος (support vector machines)**

Χωρίς να εισερχόμαστε σε μια ανάλυση της κάθε μεθόδου, θα πρέπει να σημειώσουμε ότι κάθε μέθοδος έχει σε κάποιες εφαρμογές καλή απόδοση, ενώ σε κάποιες άλλες δεν εμφανίζει σημαντικά ποσοστά επιτυχίας. Οι μεγαλύτερες διαφορές των μεθόδων κατηγοριοποίησης εντοπίζονται:

α) στη δομή των συστάδων,

β) στην ύπαρξη επικαλύψεων και στο αντίστοιχο ποσοστό

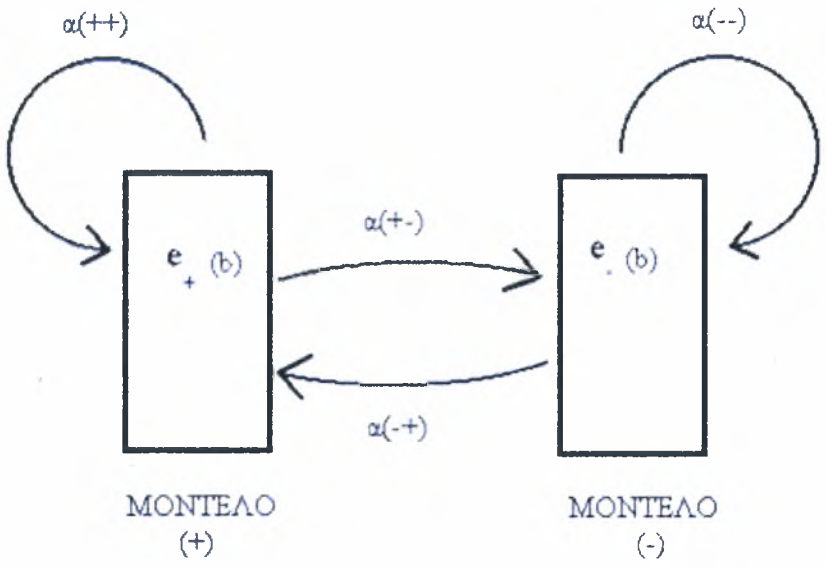
γ) στη μετρική ομοιότητας που χρησιμοποιείται.

Η δομή των συστάδων αναφέρεται τόσο στο πλήθος των συστάδων, και στο σχήμα τους, όσο και στη δυναμικότητά τους, δηλαδή το πλήθος των στοιχείων που περιλαμβάνουν σε απόλυτο ή σχετικό μέγεθος. Επίσης, η ύπαρξη επικάλυψης ανάμεσα σε συστάδες και το ποσοστό επικάλυψης που επιτρέπουμε επηρεάζουν τις παραμέτρους των μεθόδων κατηγοριοποίησης. Ανάλογα με το είδος των δεδομένων που επεξεργαζόμαστε επιλέγουμε τόσο τη μέθοδο ομαδοποίησης, όσο και τις αντίστοιχες παραμέτρους. Επομένως, η γνώση των χαρακτηριστικών και του είδους των δεδομένων λειτουργεί ως είσοδος στο πρόβλημα κατηγοριοποίησης .

Το ζήτημα της ταξινόμησης μπορεί να καταταχθεί στο ευρύτερο αντικείμενο της αναγνώρισης προτύπων (pattern recognition). Δυο από τα κυριότερα προβλήματα αναγνώρισης προτύπων στη βιοπληροφορική είναι η αναγνώριση γονιδίων και ο καθορισμός από τα συστατικά μιας ακολουθίας αμινοξέων της δευτεροταγούς δομής της πρωτεΐνης που θα προκύψει. Υπάρχουν αρκετοί τρόποι διεκπεραίωσης του

προβλήματος αναγνώρισης προτύπων σε βιολογικά μακρομόρια. Πολλά από αυτά βασίζονται στην μηχανική μάθηση (machine learning) και στα πιθανοθεωρητικά μοντέλα όπως επίσης και στα νευρωνικά δίκτυα.

Όσον αφορά τα πιθανοθεωρητικά μοντέλα η μορφή που χρησιμοποιείται ως επί το πλείστον είναι οι αλυσίδες Markov. Οι Αλυσίδες Markov (Markov Chains), είναι στοχαστικά μοντέλα, με τα οποία περιγράφουμε και αναλύουμε τις ακολουθίες βιολογικών πολυμερών όπως το DNA και τις πρωτεΐνες. Πρέπει εδώ να τονιστεί ότι το μοντέλο Markov θεωρείται από πολλούς ερευνητές ως το πιο φυσικό για να περιγράψει αλληλουχίες μακρομορίων όπως του DNA, αλλά και των πρωτεϊνών.




**Σχήμα 2.2 Η μορφή ενός βασικού Hidden Markov Model**

Ένα HMM, αφού υπολογιστούν οι παράμετροι του (πιθανότητες μεταβάσεως κλπ) από ένα γνωστό σύνολο δεδομένων (training set), χρησιμοποιείται για την πρόγνωση-αποκωδικοποίηση, σε ένα σύνολο δεδομένων με απροσδιόριστα χαρακτηριστικά (test set). Οι μέθοδοι αποκωδικοποίησης, δηλαδή εύρεσης της αλληλουχίας των καταστάσεων εάν είναι γνωστή αλληλουχία των συμβόλων, είναι βασικά 2, η αποκωδικοποίηση Viterbi, και η εκ των υστέρων αποκωδικοποίηση (posterior decoding). Συνήθως σε περιπτώσεις πολύπλοκων μοντέλων, είναι πιο χρήσιμη η εκ των υστέρων αποκωδικοποίηση.

Εκτός από την ευρεία εφαρμογή των HMM και των νευρωνικών δικτύων σε προβλήματα βιοπληροφορικής, μια σειρά από άλλες μεθόδους συναντιούνται στην αρθρογραφία όπως αυτή της υπολογιστικής γλωσσολογίας, η οποία βασίζεται στην γλωσσολογική θεωρία του Chomsky, θεωρία που βασίζεται στους γενικευμένους γραμματικούς κανόνες που υπάρχουν για να δημιουργηθεί μια πρόταση ή αλλιώς μια ακολουθία χαρακτήρων.

### 2.2.2 *Ανάπτυξη Μεθοδολογιών που επιτρέπουν την ερμηνεία αποτελεσμάτων βιολογικής σημασίας*

Η αύξηση των βιολογικών δεδομένων με εκθετικό ρυθμό τα τελευταία χρόνια δεν θα μπορούσε να προσφέρει στην επιστημονική κοινότητα αξιόλογες πληροφορίες χωρίς την ανάπτυξη μεθοδολογιών που θα εξασφάλιζαν την όσο το δυνατόν καλύτερη ερμηνεία των νέων δεδομένων.

Για αυτόν το λόγο τομείς όπως η φυλογενετική πρόβλεψη, ο καθορισμός τρισδιάστατων μορφών πρωτεϊνών από τις ακολουθίες μακρομορίων, η εξαγωγή συμπερασμάτων για την ρύθμιση της λειτουργίας ενός κυττάρου ή μιας πρωτεΐνης, επικεντρώνουν σε πολλές περιπτώσεις το ενδιαφέρον των ερευνητών. Παρακάτω παρουσιάζουμε συνοπτικά τις σημαντικότερες από αυτές τις μεθοδολογίες.

Η φυλογενετική πρόβλεψη υλοποιείται στις περισσότερες περιπτώσεις μέσα από τη δημιουργία εξελικτικών δέντρων . Τα δέντρα αυτά σχεδιάζονται μετά από σύγκριση βιολογικών ακολουθιών που ανήκουν σε διαφορετικούς οργανισμούς. Η ομαδοποίηση των βιολογικών ακολουθιών στο δέντρο γίνεται ανάλογα με τον βαθμό ομοιότητάς τους. Με αυτόν τον τρόπο έχουμε μια σαφή μορφή απεικόνισης πάνω στο πώς οι βιολογικές ακολουθίες, οι οποίες αντιπροσωπεύουν και διαφορετικούς βιολογικούς οργανισμούς, μετασχηματίστηκαν κατά τη διάρκεια της εξελικτικής διαδικασίας.

Οι μέθοδοι που χρησιμοποιούνται κυρίως για να δημιουργηθούν φυλογενετικά δέντρα, είναι:

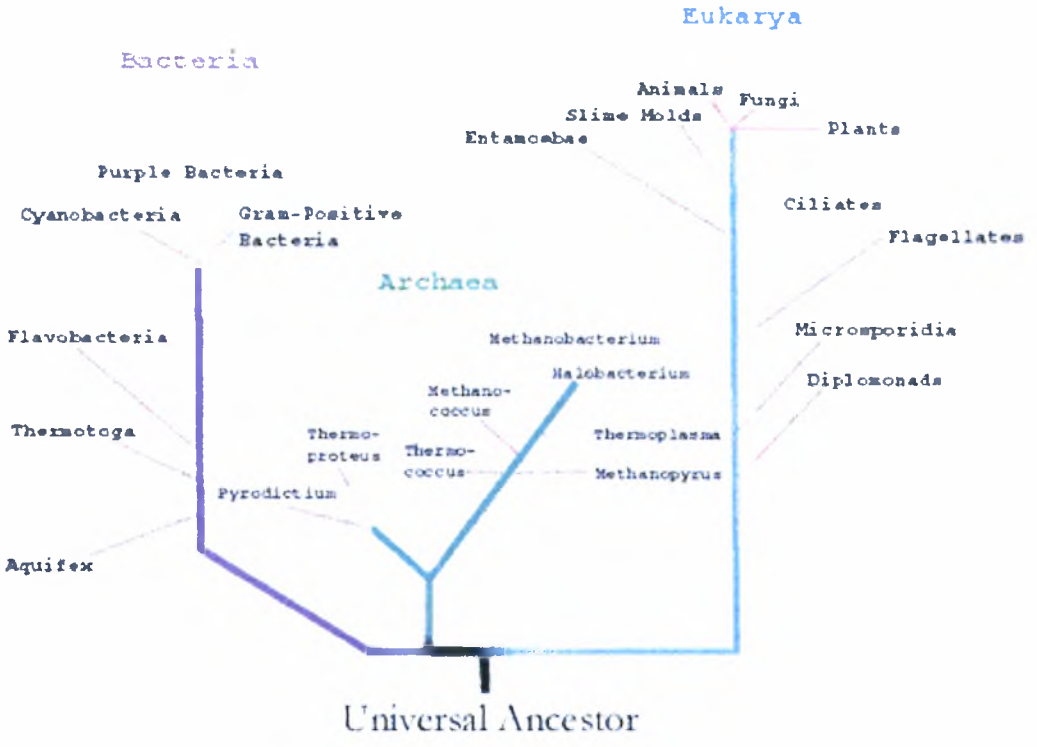
α) η μέθοδος μέγιστης συντήρησης (maximum parsimony),

β) η μέθοδος ιεραρχικής ομαδοποίησης (hierarchical clustering)

γ) η μέθοδος μέγιστης πιθανοφάνειας (maximum likelihood).

Ανάλογα με τον τρόπο δημιουργίας του φυλογενετικού δέντρου – είτε βάσει της απόστασης οπότε και επιλέγεται η ιεραρχική ομαδοποίηση, είτε βάσει εξελικτικού μοντέλου οπότε και επιλέγεται η μέθοδος μέγιστης συντήρησης ή η μέθοδος μέγιστης πιθανοφάνειας - επιλέγεται η μια από τις παραπάνω μεθόδους.

Μία από τις πιο γνωστές μορφές ιεραρχικής ομαδοποίησης είναι και ο αλγόριθμος UPGMA (Unweighted Pair Group Method using arithmetic Averages), όπως και το εργαλείο PHYLIP (Phylogenetic Inference Package) που μπορεί να βρεθεί και στο διαδίκτυο.

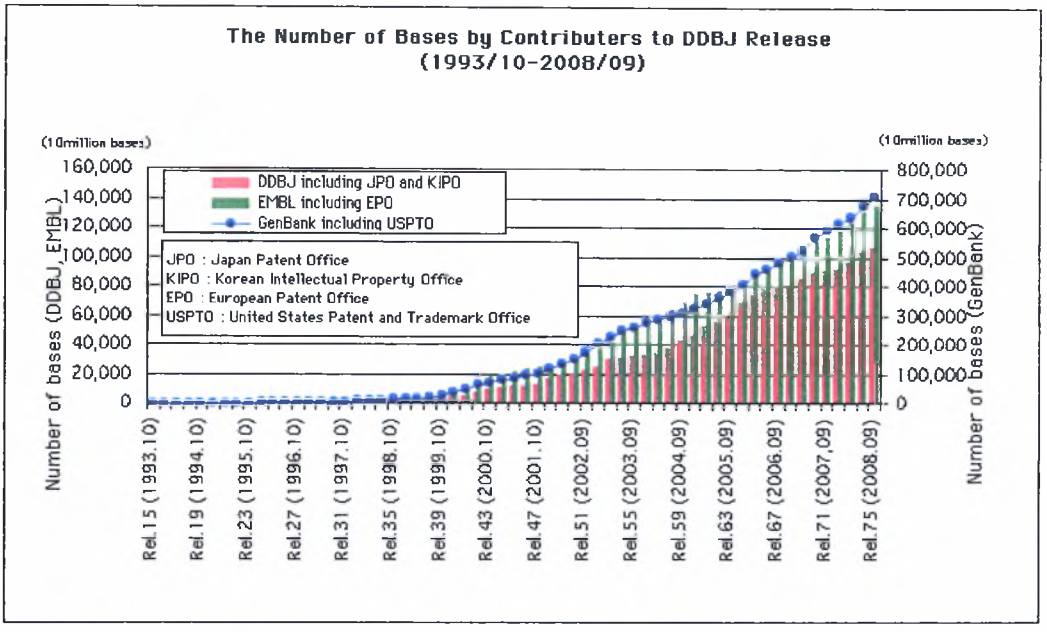



**Σχήμα 2.3 Παράδειγμα φυλογενετικού δέντρου**

Ο καθορισμός των τρισδιάστατων μορφών πρωτεϊνών από τις βιολογικές ακολουθίες έχει αποτελέσει ένα από τα σημαντικότερα σημεία έρευνας, για αυτόν άλλωστε το λόγο μια σειρά από τέτοιου είδους εφαρμογές μπορούν να βρεθούν στο διαδίκτυο. Υπολογιστικά, το πρόβλημα καθορισμού από μια ακολουθία RNA της τρισδιάστατης μορφής μιας πρωτεΐνης είναι δύσκολο, αφού απαιτούνται αλγόριθμοι πολυπλοκότητας τρίτου βαθμού. Παράλληλα, η εξαγωγή συμπεράσματος για το σχήμα που παίρνει η τρισδιάστατη μορφή μιας πρωτεΐνης από μια αμινοξική ακολουθία παραμένει ένα άλυτο πρόβλημα. Μεγάλη εφαρμογή και σε αυτήν την περίπτωση βρίσκουν τα HMM (Hidden Markov Models), ενώ εφαρμογή βρίσκει και η μέθοδος CFG (Context-Free Grammar), εφαρμογή που βασίζεται και πάλι στην υπολογιστική γλωσσολογία.

### *2.2.3 Αποδοτική Οργάνωση βιολογικών δεδομένων*

Η αποδοτική οργάνωση βιολογικών δεδομένων είναι κάτι παραπάνω από απαραίτητη εξαιτίας του τεράστια και ολοένα αυξανόμενου πλήθους βιολογικών ακολουθιών που πρέπει να αποθηκεύονται με έναν τρόπο που να διευκολύνει τόσο την ανάκτησή τους, όσο και την επεξεργασία τους. Όταν πρωτοξεκίνησε η δημιουργία των βιολογικών βάσεων δεδομένων, ο όγκος της πληροφορίας ήταν τόσο μικρός που ένας μικρός αριθμός ερευνητών αρκούσε για την συντήρηση και για την ανανέωση των βάσεων αυτών. Αν κάποιος ερευνητής ενδιαφερόταν να έχει πρόσβαση στις εγγραφές της βάσης, επικοινωνούσε με τους επιστημονικούς υπευθύνους και εκείνοι του έστελναν με συμβατικό ταχυδρομείο όλη τη βάση η οποία αρκούσε να αποθηκευτεί ακόμη και σε μερικές δισκέτες ή μια μαγνητοταινία.




**Σχήμα 2.4 Διάγραμμα αύξησης των νουκλεοτιδίων τα τελευταία 15 περίπου χρόνια στις κυριότερες νουκλεοτιδικές βάσεις.**

Την τελευταία δεκαπενταετία όμως η τεχνολογική εξέλιξη βοήθησε στη διεκπεραίωση μεγάλου όγκου πειραματικής εργασίας, η οποία σε συνάρτηση με τον διαρκή προσδιορισμό γονιδιωμάτων διαφόρων οργανισμών αύξησε τον όγκο της πληροφορίας στο επίπεδο της ακολουθίας, και όχι μόνο, σε υπέρογκα μεγέθη. Οι βάσεις πλέον δεν περιέχουν απλώς πολλά δεδομένα, αλλά και η διαδικασία ανανέωσης τους είναι απαραίτητη καθημερινή υπόθεση. Πλέον η συντήρηση μιας βάσης απαιτεί ένα πολυάριθμο επιτελείο επιστημόνων, οι οποίοι ασχολούνται αποκλειστικά με το σχολιασμό των νεοεισερχόμενων δεδομένων καθώς και με τη διόρθωση λαθών των ήδη υπαρχόντων.

Χαρακτηριστικά παραδείγματα αποτελούν η βάση πρωτεϊνικών ακολουθιών **SWISS-PROT** που περιέχει 398181 ακολουθίες (version 56.2 - 23-Sep-2008), ενώ η **EMBL Nucleotide Sequence Database** που περιέχει νουκλεοτιδικές αλληλουχίες έχει **233,090,586,920** νουκλεοτιδικές αλυσίδες σε **143,843,446** εγγραφές (release 96, *28 AUG 2008*).

Η πρόσβαση στις βάσεις αυτές είναι πλέον εύκολη μέσω της χρήσης του διαδικτύου. Ο χρήστης μπορεί να επισκεφτεί την ιστοσελίδα που διατηρείται από τους υπευθύνους της βάσης και να κάνει αναζητήσεις αποθηκεύοντας στον υπολογιστή του δεδομένα του άμεσου ενδιαφέροντός του. Παράλληλα έχουν δημιουργηθεί και μια σειρά από βάσεις που αποσκοπούν στην ταξινόμηση της πληροφορίας στο επίπεδο της ακολουθίας και της δομής, προκειμένου να οργανωθεί η πληροφορία και να εξαχθούν συμπεράσματα για την βιολογική τους σημασία.

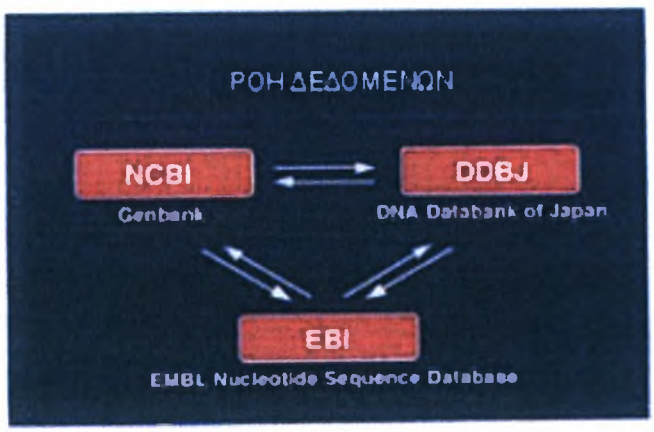
### *2.2.3.1 Βάσεις δεδομένων νουκλεοτιδικών ακολουθιών*

Οι βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών αποτελούν τις μεγαλύτερες βάσεις στο ευρύτερο πεδίο της Βιολογίας τόσο από άποψη του όγκου της πληροφορίας που περιέχουν, όσο και από την άποψη του εκθετικού ρυθμού συσσώρευσης δεδομένων που εμφανίζουν. Τα τελευταία χρόνια λόγω της εξέλιξης της τεχνολογίας στην εύρεση της αλληλουχίας (sequencing) πολυνουκλεοτιδίων έγινε εφικτός, σε μικρό χρονικό διάστημα, ο προσδιορισμός της αλληλουχίας ολόκληρων γονιδιωμάτων αρκετών οργανισμών, όπως ο άνθρωπος. Σε αρκετές περιπτώσεις μάλιστα υπάρχουν εξειδικευμένες βάσεις δεδομένων που περιέχουν τις αλληλουχίες για ένα και μόνο οργανισμό.

Εδώ πρέπει να σημειώσουμε τις τρεις μεγαλύτερες βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών που είναι ελεύθερα διαθέσιμες στην ακαδημαϊκή κοινότητα. Πρόκειται για τις :

- GENBANK (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
- EMBL-Bank: (http://www.ebi.ac.uk/embl/)
- DDJB: (http://www.ddbj.nig.ac.jp/)

Αυτές έχουν δημιουργήσει την International Nucleotide Sequence Database Collaboration. Η συνεργασία μεταξύ των βάσεων περιλαμβάνει την ανταλλαγή σε καθημερινή βάση εγγραφών που κατατίθενται ανεξάρτητα σε κάθε βάση δεδομένων έχοντας θέσει παράλληλα και κοινούς κανόνες για την ταξινόμηση και το σχολιασμό των δεδομένων. Στο παρακάτω σχήμα παρουσιάζεται η ροή της πληροφορίας ανάμεσα στις βάσεις.




**Σχήμα 2.5 Η επικοινωνία μεταξύ των τριών μεγαλύτερων βάσεων βιολογικών δεδομένων είναι συνεχής και αμφίδρομη.**

### *2.2.3.2 Βάσεις δεδομένων πρωτεϊνικών ακολουθιών και βάσεις για την ανάλυση ακολουθιών*

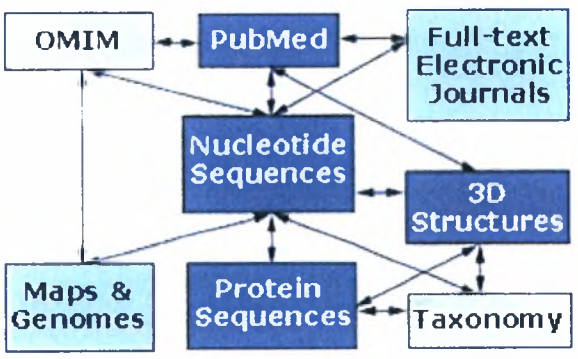
- SWISS-PROT (http://www.expasy.ch/sprot/)
- Protein Information Resource (PIR) (http://pir.georgetown.edu/)
- PROSITE (http://www.expasy.ch/prosite/)

### *2.2.3.3 Βάσεις Δεδομένων Δομικής Βιολογίας*

- Protein Data Bank (PDB): (www.rcsb.org)
- CATH: (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)
- SCOP: (http://scop.mrc-lmb.cam.ac.uk/scop/index.html)

### *2.2.3.4 Ολοκληρωμένα Συστήματα Ανάκλησης Πληροφοριών από Βάσεις Δεδομένων*

- SRS
- Entrez




**Σχήμα 2.6 Οι δυνατότητες ενός ολοκληρωμένου εργαλείου όπως το Entrez σε μορφή διαγράμματος**

## *2.3 Εργαλεία Διαχείρισης Βιολογικών Δεδομένων*

Η ανάπτυξη του διαδικτύου άλλαξε τα δεδομένα και στη βιοπληροφορική και τις εφαρμογές της, καθώς δεν χρειάζεται πλέον η εύρεση και εγκατάσταση ενός προγράμματος σε κάθε υπολογιστή, αλλά μοναδική προϋπόθεση είναι η ύπαρξη σύνδεσης στο διαδίκτυο. Με άλλα λόγια η ανάπτυξη του μοντέλου χρήστη-εξυπηρετητή , δίνει τη δυνατότητα να προσπελασθούν από το χρήστη βάσεις με τεράστιο αριθμό καταχωρίσεων βιολογικών δεδομένων, ενώ δίνεται και η δυνατότητα εκτέλεσης ενός προγράμματος στο υπολογιστικό σύστημα αυτού που παρέχει την υπηρεσία. Ο χρήστης ουσιαστικά βλέπει μόνο ένα user interface, και τα αποτελέσματα από την εκτέλεση του προγράμματος που επέλεξε. Το γεγονός αυτό έχει ιδιαίτερη σημασία για εφαρμογές στην βιολογία, καθώς τα δεδομένα έχουν μεγάλο μέγεθος και μπορούν να γίνουν με αυτόν τον τρόπο εκμεταλλεύσιμα με την απλή ύπαρξη λίγων αλλά ισχυρών εξυπηρετητών.

Άμεσο αποτέλεσμα των δυνατοτήτων που δόθηκαν από την ανάπτυξη του διαδικτύου ήταν η δημιουργία μιας σειράς εφαρμογών που διατίθενται μέσω αυτού. Η ανάπτυξη συλλογών εφαρμογών από τις ήδη γνωστές τράπεζες βιολογικών δεδομένων ήταν το επόμενο βήμα. Χαρακτηριστικό παράδειγμα είναι η συλλογή εργαλείων της τράπεζας EMBL, καθώς αντίστοιχα οργανωμένη είναι τόσο η τράπεζα NCBI όσο και η DDBJ με παρόμοιες εφαρμογές προς χρήση.

Εκτός από τα εργαλεία που προσφέρονται από τις συγκεκριμένες βάσεις δεδομένων μια σειρά άλλων υπάρχουν σε δικτυακούς τόπους ερευνητικών κέντρων και πανεπιστημιακών τμημάτων. Ειδικότερα όσον αφορά τα εργαλεία αναζήτησης γονιδίων, τα κυριότερα που προσφέρονται μέσα από δικτυακούς τόπους είναι τα εξής:

- GENSCAN http://genes.mit.edu/GENSCAN.html : Αποτελεί ένα από τα εργαλεία που χρησιμοποιείται κατά κόρον από τους ερευνητές. Βασίζεται σε πιθανοθεωρητικά μοντέλα για τη δομή γονιδίων.

- NetGene http://www.cbs.dtu.dk/services/NetGene2/ : Χρησιμοποιεί τεχνητά νευρωνικά δίκτυα για την ανάλυση splice περιοχών κυρίως στο ανθρώπινο γονιδίωμα.

- GeneID-3 http://www1.imim.es/geneid.html : Χρησιμοποιεί μοντέλα που βασίζονται σε κανόνες .

- GeneMark http://opal.biology.gatech.edu/GeneMark/ : Βασίζεται στη μέθοδο των Hidden Markov Models.

- GenLang http://www.cbil.upenn.edu/genlang/genlang.html : Χρησιμοποιεί γλωσσολογικές μεθόδους .

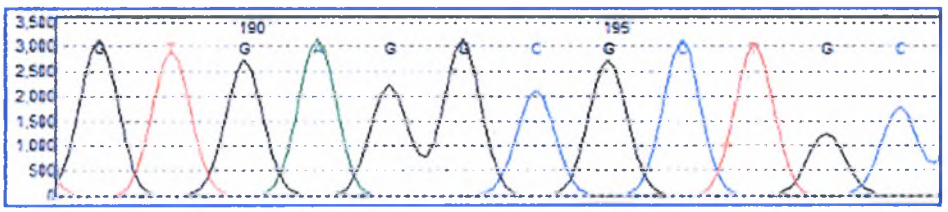
Να σημειώσουμε ότι τα εργαλεία αυτά απευθύνονται ως επί το πλείστον τόσο σε ευκαρυωτικούς, όσο και σε προκαρυωτικούς οργανισμούς.

## *2.4 Επαναληπτικές ακολουθίες*

Μια ακολουθία DNA ή γενετική ακολουθία είναι μια διαδοχή γραμμάτων που αντιπροσωπεύουν την αρχική δομή ενός πραγματικού ή υποθετικού μορίου DNA, με την ικανότητα να φέρουν πληροφορίες, τέτοιες όπως περιγράφηκαν από το κεντρικό δόγμα της μοριακής βιολογίας.

Τα πιθανά γράμματα είναι τα A,C,G,T και αντιπροσωπεύουν τις τέσσερις νουκλεοτιδικές βάσεις του μορίου DNA - αδενίνη, κυτοσίνη, γουανίνη, θυμίνη. Στη χαρακτηριστική περίπτωση, οι ακολουθίες εμφανίζονται η μία μετά την άλλη, χωρίς ενδιάμεσα κενά, όπως στην ακολουθία AAAGTCTGAC. Εάν διαβάσουμε την ακολουθία από τα αριστερά προς τα δεξιά έχουμε την 5′ 3′. Οι μικρές ακολουθίες νουκλεοτιδίων αναφέρονται ως ολιγονουκλεοτίδια και χρησιμοποιούνται σε μια σειρά εργαστηριακών εφαρμογών στη μοριακή βιολογία. Όσον αφορά τη βιολογική λειτουργία, μια ακολουθία DNA μπορεί να είτε να κωδικοποιείται είτε όχι. Οι ακολουθίες DNA μπορούν επίσης να περιέχουν και «junk DNA », δηλαδή κομμάτια των οποίων η βιολογική χρησιμότητα δεν είναι ακόμα γνωστή.

Οι ακολουθίες μπορούν να προέλθουν από τη βιολογική πρώτη ύλη μέσω μιας διαδικασίας αποκαλούμενης DNA sequencing.




**Σχήμα 2.7 DNA sequence trace**

Σε μερικές ειδικές περιπτώσεις, πρόσθετα γράμματα εκτός των A ,T ,C ,G είναι παρούσα σε μια ακολουθία. Αυτά τα γράμματα αντιπροσωπεύουν την ασάφεια. Από όλες τις δειγματοληψίες που έχουν γίνει, σε αυτά τα γράμματα αντιστοιχούν περισσότερα του ενός νουκλεοτίδια. Σύμφωνα με τους κανόνες της Διεθνούς Ένωσης καθαρής και Εφαρμοσμένης Χημείας (IUPAC) είναι τα ακόλουθα:

```
A= adenine, C= cytosine
G= guanine ,T= thymine
   R = G A (purine)
 Y = T C (pyrimidine)
    K = G T (keto)
   M = A C (amino)
S = G C (strong bonds)
 W = A T (weak bonds)
B = G T C (all but A)
D = G A T (all but C)
H = A C T (all but G)
V = G C A (all but T)
  N = A G C T (any)
```

**Σχήμα 2.8 Παρουσίαση των επιπλέον γραμμάτων σύμφωνα με την IUPAC**

Μια υπακολουθία, ένα πρόθεμα ή ένα επίθεμα μιας σειράς είναι ένα υποσύνολο των συμβόλων μιας ακολουθίας, όπου η σειρά των στοιχείων διατηρείται. Παραδείγματος χάριν, η σειρά νάνα είναι ίση με ένα επίθεμα της σειράς μπανάνα.

### *2.4.1 Διάφοροι τύποι*

Στην έρευνα των DNA ακολουθιών μπορούν να διαχωριστούν οι παρακάτω κύριοι τύποι επαναληπτικών ακολουθιών:

1. Tandem repeats:
    - Satellite DNA
    - Minisatellite
    - Microsatellites
2. Interspersed repeats:
    - SINEs(Short INterspersed Elements)
    - LINEs(Long INterspersed Elements)

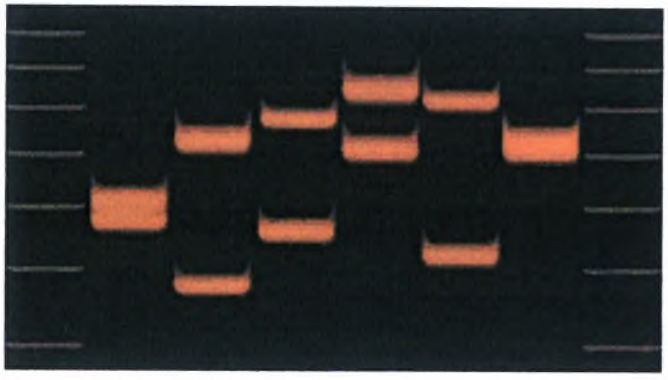
#### *2.4.1.1 Tandem repeats*

Τα ***Tandem repeats*** είναι διαδοχικές επαναλήψεις που εμφανίζονται στο DNA ως ένα σύνολο δύο η περισσοτέρων νουκλεοτιδίων που επαναλαμβάνονται και οι επαναλήψεις είναι η μία δίπλα στην άλλη. Παραδείγματος χάριν στην ακολουθία ATTCGATTCGATTCG το ATTCG επαναλαμβάνεται 3 φορές.

Μεταξύ επαναλήψεων 10 έως και 60 νουκλεοτιδίων , ονομάζονται minisatellite. Για λιγότερα νουκλεοτίδια ονομάζονται microsatellites ή short tandem repeats. Όταν ακριβώς δύο νουκλεοτίδια επαναλαμβάνονται , λέμε ότι έχουμε δινουκλεοτιδική επανάληψη, όταν τρία νουκλεοτίδια επαναλαμβάνονται λέμε ότι έχουμε τρινουκλεοτιδική επανάληψη κ.ο.κ. Όταν ο αριθμός δεν είναι γνωστός μερικές φορές ονομάζεται variable number tandem repeat.

Οι διαδοχικές επαναλήψεις περιγράφουν ένα μοτίβο που μπορεί να βοηθήσει στον καθορισμό των κληρονομικών γνωρισμάτων ενός ατόμου. Διότι, οι διαδοχικές επαναλήψεις είναι πολύ χρήσιμες στον καθορισμό της προέλευσης. Συγκεκριμένα τα short tandem repeats χρησιμοποιούνται για ορισμένα γενεαλογικά τεστ DNA.

Ένα variable tandem repeat είναι μια τοποθεσία σε ένα γονιδίωμα όπου μία μικρή νουκλεοτιδική ακολουθία οργανωμένη σαν διαδοχική επανάληψη. Αυτά μπορούν να βρεθούν στα χρωμοσώματα, και να παρουσιάσουν συχνά παραλλαγές στο μήκος μεταξύ των ατόμων . Κάθε παραλλαγή ενεργεί ως κληρονομημένο αλληλόμορφο γονίδιο, επιτρέποντας έτσι να χρησιμοποιηθεί για προσωπικό ή γονικό προσδιορισμό. Η ανάλυσή τους είναι χρήσιμη στην έρευνα της γενετικής και της βιολογίας, της ιατροδικαστικής και στο DNA fingerprinting.



**Σχήμα 2.9 Παραλλαγές των μηκών αλληλόμορφων γονιδίων VNTR (D1S80) σε 6 άτομα.**

Το Satellite DNA περιέχει ιδιαίτερα επαναλαμβανόμενο DNA, και αποκαλείται έτσι επειδή οι επαναλήψεις μιας σύντομης ακολουθίας DNA τείνουν να παραγάγουν μια διαφορετική συχνότητα νουκλεοτιδίων της αδενίνης, της κυτοσίνης, της γουανίνης και της θυμίνης. Κάποιοι τύποι Satellite DNA στους ανθρώπους είναι οι παρακάτω:

| Τύπος | Μέγεθος επαναληπτικής μονάδας (bp) | Τοποθεσία |
|---|---|---|
| α (alphoid DNA) | 171 | Όλα τα χρωμοσώματα |
| β | 68 | Κεντρομερή των χρωμοσωμάτων 1, 9, 13, 14, 15, 21, 22 και Y |
| Satellite 1 | 25-48 | Κεντρομερή και άλλοι χώροι των περισσοτέρων χρωμοσωμάτων |
| Satellite 2 | 5 | Στα περισσότερα χρωμοσώματα |
| Satellite 3 | 5 | Στα περισσότερα χρωμοσώματα |

**Πίνακας 2.1 Πίνακας που παρουσιάζει κάποιους τύπους Satellite DNA**

Το minisatellite είναι ένα τμήμα DNA το οποίο αποτελείται από μικρές ακολουθίες από βάσεις μεγέθους 10-60bp. Αυτό συμβαίνει για παραπάνω από 1000 τοποθεσίες στο ανθρώπινο γονιδίωμα. Κάποια minisatellite περιέχουν μία κεντρική (ή πυρήνα) ακολουθία γραμμάτων "GGGCAGGAXG" (όπου X μπορεί να είναι οποιοδήποτε γράμμα) ή γενικότερα ένα νήμα από πουρίνες (αδενίνη (A) και γουανίνη (G)) στη μία και ένα από την άλλη πλευρά από πυριμιδίνες (θυμίνη (T) κυτοσίνη (C)) .Θεωρείται ότι αυτές οι ακολουθίες ενθαρρύνουν τα χρωμοσώματα να αντιμεταθέτουν το DNA. Οι σωματικές αλλαγές σηματοδοτούνται από δυσκολίες στην αντιγραφή. Όταν συμβαίνουν τέτοια περιστατικά , τα minisatellite εμφανίζουν ελαφρώς διαφορετικό αριθμό επαναλήψεων σε περισσότερες από 1000 τοποθεσίες στο γονιδίωμα ενός ατόμου, πράγμα που τις κάνει μοναδικές.

Τα microsatellite είναι πολυμορφικές τοποθεσίες που παρουσιάζονται στο πυρηνικό και κυτταρικό DNA και αποτελούνται από επαναλαμβανόμενες μονάδες 1-6 bp σε μήκος. Είναι τυπικά, συν-κυρίαρχα Και χρησιμοποιούνται ως μοριακοί σηματοδότες οι οποίοι έχουν ευρεία χρήση σε εφαρμογές στον τομέα της γενετικής. Τα microsatellite μπορούν ακόμα να χρησιμοποιηθούν για την μελέτη του γονιδιώματος (αναζήτηση για αντιγραφές ή διαγραφές συγκεκριμένων γενετικών περιοχών).

Συγκεκριμένα τα microsatellites είναι τμήματα του DNA που αποτελούνται από ζευγαρωτές επαναλήψεις πολύ απλών μοτίβων όπως το (CT) στην παρακάτω ακολουθία :


**Σχήμα 2.10 Παράδειγμα microsatellite repeat**

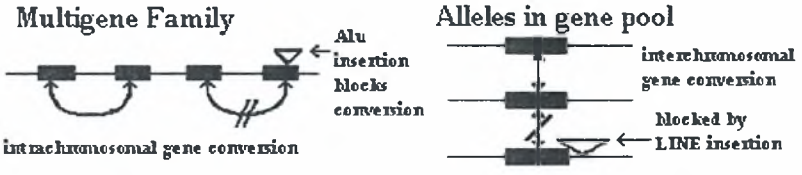
Το μήκος των ακολουθιών αυτών τείνει να κυμαίνεται μεταξύ ατόμων, εξαιτίας των υψηλών ρυθμών μεταλλάξεων. Όταν το DNA αντιγράφεται κατά την διάρκεια της μίτωσης, το ένζυμο της DNA πολυμεράσης μπορεί να γλιστράει μπρος πίσω στις επαναλαμβανόμενες μονάδες, διαγράφοντας ή προσθέτοντας επαναλαμβανόμενες μονάδες στον θυγατρικό κλώνο. Αυτό σημαίνει, ότι οι θυγατρικοί κλώνοι, μπορεί να έχουν ελαφρώς λιγότερες ή περισσότερες επαναλαμβανόμενες μονάδες από ότι ο πατρικός κλώνος. Οι γενετιστές μπορούν να μετρήσουν τον αριθμό των επαναλαμβανόμενων μονάδων για ένα δοθέν microsatellite. Εάν πολλά microsatellite μετρηθούν σε ένα άτομο, τα αποτελέσματα θα είναι ένα μοναδικό γενετικό αποτύπωμα.

Τα microsatellites έχουν πολλές χρήσεις :

1. Στην ταυτοποίηση μοναδικών ατόμων σε έναν πληθυσμό: ανθρώπων, ζώων, φυτών.
2. Στην εγκληματολογική επιστήμη, για να συγκρίνει τους υπόπτους με στοιχεία του εγκλήματος.
3. Για τον καθορισμό της πατρότητας ή της μητρότητας ατόμων.
4. Για την εκτίμηση πληθυσμών.
5. Για τον καθορισμό εάν κάποιος πληθυσμός υπέφερε από γενετικές επιβαρύνσεις.
6. Για φυλογενετικές μελέτες, σε εφαρμογές που περιλαμβάνουν πολύ κοντινά είδη.
7. Για γενετική αντιστοίχιση.
8. Εάν υπάρχει υποδιαίρεση πληθυσμού σε πολύ μεγάλους πληθυσμούς.

### 2.4.1.2 Interspersed repeats (διασπαρμένες επαναλήψεις)

Το διασπαρμένο επαναληπτικό DNA εμφανίζεται σε όλους τα ευκαρυωτικά γονιδιώματα. Αυτές οι ακολουθίες πολλαπλασιάζονται με τη διαμεσολάβηση του tRNA και ονομάζονται και αλλιώς retroposons. Τα στοιχεία του διασπαρμένου επαναληπτικού DNA επιτρέπουν την εξέλιξη των γονιδίων. Το πετυχαίνουν με την αποσύνδεση όμοιων DNA ακολουθιών από την μετατροπή του γονιδιώματος κατά την διάρκεια της μείωσης. Το SINE επαναληπτικό DNA εξειδικεύεται σε intrachromosomal γονιδίωμα ενώ το LINE σε interchromosomal. Και στις δύο περιπτώσεις , οι διασπαρμένες επαναλήψεις μπλοκάρουν την μετατροπή των γονιδίων εισάγοντας περιοχές μη-ομόλογες μεταξύ κατά τα άλλα όμοιων DNA ακολουθιών. Οι δυνάμεις ομογενοποίησης που συνδέουν τις ακολουθίες DNA είναι με αυτόν τον τρόπο σπασμένες και οι ακολουθίες DNA είναι ελεύθερες να εξελιχθούν ανεξάρτητα. Αυτό οδηγεί στη δημιουργία των νέων γονιδίων και των νέων ειδών κατά τη διάρκεια της εξέλιξης. Με το σπάσιμο των συνδέσεων που ειδάλλως θα επικάλυπταν τις νέες ακολουθίες DNA, οι διασπαρμένες επαναλήψεις καταλύουν την εξέλιξη, που επιτρέπει στα νέα γονίδια και τα νέα είδη να αναπτυχθούν.




**Σχήμα 2.11 Απεικόνιση intrachromosomal και interchromosomal**

Τυπικό παράδειγμα SINE, αποτελεί η Alu family, η οποία επαναλαμβάνεται 300.000 φορές μέσα στο ανθρώπινο γονιδίωμα και καλύπτει σε μήκος το 5% περίπου του ανθρώπινου DNA και άλλων γονιδιωμάτων θηλαστικών.

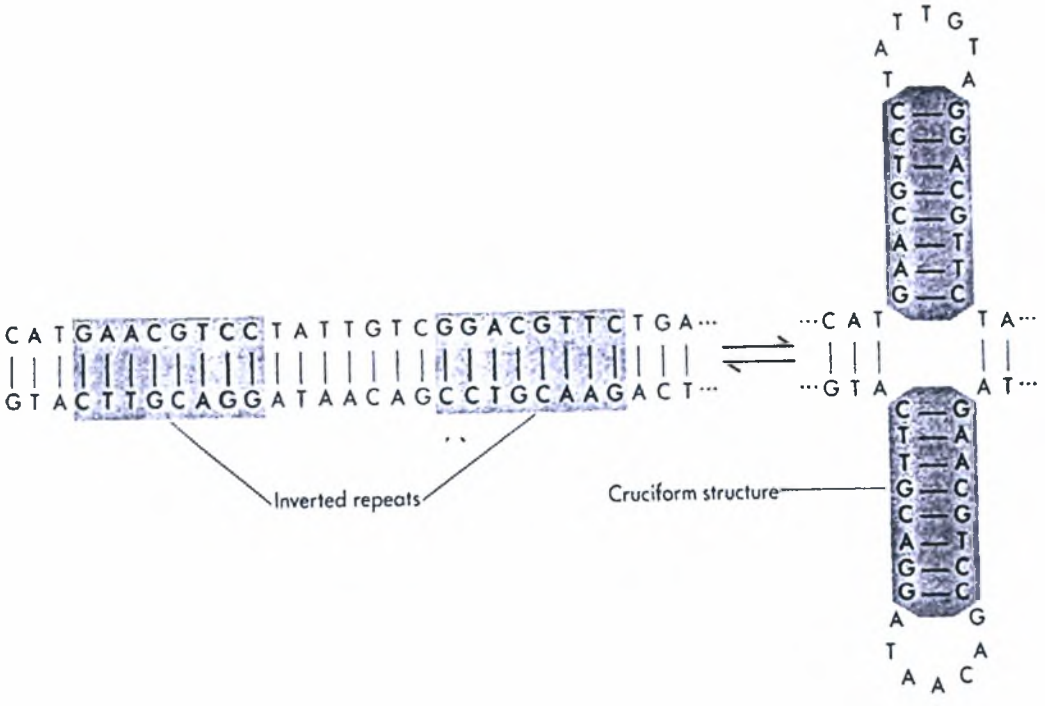
Η αναζήτηση επαναλαμβανόμενων μοτίβων, αποτελεί σημαντικό υπολογιστικό πρόβλημα στη Βιοπληροφορική, ειδικά μετά τη χαρτογράφηση του ανθρώπινου γονιδιώματος, αφού στοχεύει στην αναγνώριση δεικτών- markers, που υποδεικνύουν σημαντικές θέσεις ή λειτουργικά τμήματα στις βιολογικές ακολουθίες. Επίσης η αναζήτηση επαναλαμβανόμενων μοτίβων, μπορεί να στηρίζεται είτε στην ακριβή είτε στην προσεγγιστική προσέγγιση.

### 2.4.1.3 Παλινδρομική ακολουθία

Παλίνδρομες ακολουθίες είναι ακολουθίες από γράμματα/ λέξεις που είναι όμοια από μπροστά προς τα πίσω όπως και ανάποδα. Μια παλινδρομική ακολουθία DNA είναι μια ακολουθία από νουκλεοτιδικές βάσεις η οποία διαβάζεται με τον ίδιο τρόπο που διαβάζεται και η συμπληρωματική της ακολουθία. Παραδείγματος χάριν η ΄5-AATCGCGATT- ΄3 είναι παλινδρομική ακολουθία μας και η συμπληρωματική της ΄3-TTAGCGCTAA- ΄5 διαβάζεται το ίδιο, εάν διαβαστεί αντίστροφα. Οι παλινδρομικές ακολουθίες DNA είναι κρίσιμες για τον γενετικό έλεγχο.

| Ένζυμο | Πηγή | Ακολουθία αναγνώρισης | Περικοπή |
|---|---|---|---|
| EcoR1 | *Escherichia coli* | 5 " GAATTC<br>3 " CTTAAG | 5 "---Γ AATTC---3 "<br>3 "---CTTAA Γ---5 " |
| BamH1 | *Βάκιλος amyloliquefaciens* | 5 " GGATCC<br>3 " CCTAGG | 5 "---Γ GATCC---3 "<br>3 "---CCTAG Γ---5 " |
| Taq1 | *Aquaticus Thermus* | 5 " TCGA<br>3 " AGCT | 5 "---Τ CGA---3 "<br>3 "---AGC Τ---5 " |
| Alu1* | *Arthrobacter luteus* | 5 " AGCT<br>3 " TCGA | 5 "---CT ΑΡΤΥΡΟΥ---3 "<br>3 "---TC GA---5 " |

**Σχήμα 2.12 Παλινδρομικές περιοχές και τα ένζυμα περιορισμού που αναγνωρίζουν ακολουθούν**




**Σχήμα 2.13 Απεικόνιση παλινδρομικής ακολουθίας**

## *2.5 Η Γλώσσα Προγραμματισμού Java*

Η αλματώδης ανάπτυξη του κλάδου της Βιοπληροφορικής ανάγκασε μια σειρά από προγραμματιστές να εξετάσουν και να αναπτύξουν σε ήδη υπάρχουσες γλώσσες προγραμματισμού τη δυνατότητα διαχείρισης βιολογικών ακολουθιών από αυτές, όπως επίσης και τη δυνατότητα αποθήκευσής τους σε μορφή που να είναι πιο εύκολα διαχειρίσιμη και προσπελάσιμη.

Μια συνοπτική παρουσίαση της γλώσσας προγραμματισμού Java θα μπορούσε να κωδικοποιηθεί στα παρακάτω σημεία:

1)Ιστορικά, η γλώσσα προγραμματισμού Java, αναπτύχθηκε με γοργούς ρυθμούς από το 1991 με αρχική επιδίωξη την ανάπτυξη λογισμικού για έλεγχο ηλεκτρονικών συσκευών ευρείας κατανάλωσης. Ενώ αρχικά το όλο εγχείρημα θα υλοποιούνταν στη γλώσσα προγραμματισμού C++, η όχι καλή λειτουργικότητα της γλώσσας, οδήγησε στην ανάπτυξη μιας νέας γλώσσας προγραμματισμού, που αρχικά ονομάστηκε Oak και στη συνέχεια Java. Εξαιτίας της μεταφερσιμότητας, της ευελιξίας και της αποδοτικότητας που θα έπρεπε να τη διακρίνει ούτως ώστε να είναι χρήσιμη σε μια σειρά διαφορετικών συσκευών, αλλά και λόγω της μεγάλης αξιοπιστίας που θα έπρεπε να έχει, η Java είχε χαρακτηριστικά τα οποία ανταποκρίνονταν στις απαιτήσεις για νέους τρόπους ανάπτυξης και διανομής του λογισμικού με την ραγδαία εξάπλωση του Διαδικτύου (Internet) και του Παγκόσμιου Ιστού (World-Wide-Web). Η Java σχεδιάστηκε με σκοπό την ανάπτυξη εφαρμογών που τρέχουν σε ετερογενή δικτυακά περιβάλλοντα. Η γλώσσα προγραμματισμού Java υποστηρίζει τα ακόλουθα χαρακτηριστικά:

2) Η γλώσσα Java είναι αντικειμενοστραφής (object-oriented language). Η διαχείριση όμως των αντικειμένων γίνεται με διαφορετικό σε σχέση με τη γλώσσα προγραμματισμού C++ τρόπο. Πιο συγκεκριμένα, η γλώσσα προγραμματισμού Java δεσμεύει και αποδεσμεύει μνήμη αυτόματα καθώς το πρόγραμμα δημιουργεί και καταστρέφει αντικείμενα. Όπου οι προγραμματιστές σε Java δημιουργούν και χρησιμοποιούν, προβλέψιμες και ασφαλείς, άμεσες αναφορές (references) σε αντικείμενα, οι προγραμματιστές σε C++ χρησιμοποιούν δείκτες (pointers), μια πρακτική, η οποία εγκυμονεί κινδύνους από διάφορες απόψεις.

3) Μία ακόμα θεμελιώδης διαφορά ανάμεσα στις δύο γλώσσες βρίσκεται στους μηχανισμούς με τους οποίους δίνουν σε ένα αντικείμενο πολλαπλούς τύπους. Η C++ υποστηρίζει πολλαπλή κληρονομικότητα (multiple inheritance), το οποίο σημαίνει ότι οι κλάσεις μπορούν να έχουν πολλαπλές βασικές κλάσεις. Αυτός ο μηχανισμός οδηγεί σε ασάφεια και σύγχυση. Η γλώσσα Java αντικαθιστά αυτόν το μηχανισμό με ένα πιο ξεκάθαρο είδος αφαιρετικής δομής, τις διεπαφές (interfaces). Μία κλάση Java μπορεί να έχει μία μόνο βασική κλάση (parent class) αλλά μπορεί να υλοποιεί πολλαπλές διεπαφές.

4) Η Java χρησιμοποιεί αρχιτεκτονική εικονικής μηχανής (virtual machine), ή (όπως αλλιώς είναι γνωστή) *διερμηνέα* Java (**Java Interpreter ή Java runtime**). Με αυτόν τον τρόπο, η εικονική μηχανή μπορεί να υλοποιηθεί, ώστε να λειτουργεί σε διάφορα λειτουργικά συστήματα και πλατφόρμες με σταθερότητα, ενώ παράλληλα παρέχει αυστηρό έλεγχο σε εκτελούμενα προγράμματα, δίνοντας έτσι την δυνατότητα για ασφαλή εκτέλεση αναξιόπιστου κώδικα.

5) Η πλατφόρμα Java περιλαμβάνει μια εκτεταμένη συλλογή από συγκεκριμένες βιβλιοθήκες κλάσεων που ονομάζονται Διεπαφές Προγραμματισμού Εφαρμογών (**Application Programming Interfaces** σύντομα **APIs**). Αυτά υποστηρίζουν οτιδήποτε μπορεί να χρειάζεται μια εφαρμογή.

6) Υποστηρίζεται δημιουργία ανεξάρτητων εφαρμογών και applets.

7) Η Java είναι κατανεμημένη (distributed), δηλαδή ένα πρόγραμμα Java είναι δυνατό να εγκατασταθεί και να λειτουργήσει διαμέσου του Διαδικτύου ή ακόμα και από διαφορετικές συσκευές στο Διαδίκτυο.

8) Η Java είναι ασφαλής (secure), καθώς στο Διαδίκτυο ελλοχεύουν πολλοί κίνδυνοι για τον χρήστη – παραλήπτη μιας δικτυακής εφαρμογής, ενώ η Java έχει σχεδιαστεί έτσι ώστε να ελαχιστοποιείται η πιθανότητα προσβολής του συστήματος του χρήστη από κάποιο πρόγραμμα γραμμένο για αυτό το σκοπό.

9) Υποστηρίζει ταυτοχρονισμό (multithreading). Η Java υποστηρίζει εγγενώς την χρήση νημάτων (Threads). Προκειμένου να το πετύχει αυτό σε συστήματα με έναν επεξεργαστή, το **Java runtime system (interpreter)** υλοποιεί ένα δικό του χρονοδρομολογητή (scheduler), ενώ σε συστήματα που υποστηρίζουν πολυεπεξεργασία η δημιουργία νημάτων ανατίθεται στο λειτουργικό σύστημα. Η όλη διαδικασία είναι αόρατη τόσο στον προγραμματιστή όσο και στον χρήστη.

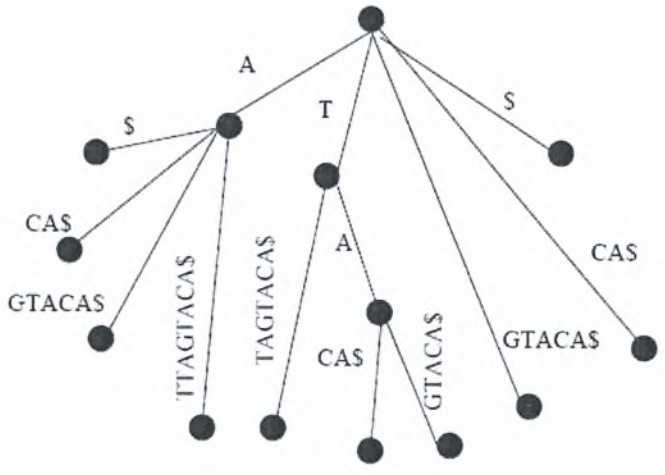
10) Τέλος η Java υποστηρίζει εφαρμογές πολυμέσων τόσο με την ευελιξία της ως γλώσσα προγραμματισμού, όσο και με τις πλούσιες και συνεχώς εμπλουτιζόμενες βιβλιοθήκες.

# ΚΕΦΑΛΑΙΟ 3

## Χρήση δέντρων επιθεμάτων για τη μελέτη γονιδιωματικών ακολουθιών

### *3.1 Δέντρα επιθεμάτων και αλγόριθμοι*

Μια συμβολοσειρά S, μήκους |S| =m, έχει m δυνατά μη κενά επιθέματα που είναι τα ακόλουθα: S[1…m], S[2…m], …. S[m-1…m] και S[m]. Για παράδειγμα για τη συμβολοσειρά "sequence", τα δυνατά επιθέματα είναι: sequence, equence, quence, uence, ence, nce, ce, e. Το Δέντρο Επιθεμάτων (Suffix Tree), αποθηκεύει όλα τα δυνατά επιθέματα της συμβολοσειράς S, όπως φαίνεται και στο ακόλουθο σχήμα.




**Σχήμα 3.1 Το Δέντρο Επιθεμάτων για τη συμβολοσειρά S = ATTAGTACA$**

Το Δέντρο Επιθεμάτων (Suffix Tree), T, μιας συμβολοσειρά S μεγέθους m (|S| =m) ορίζεται ως η κατευθυνόμενη δενδρική δομή με ακριβώς m φύλλα τα οποία είναι αριθμημένα από το 1 μέχρι το m. Κάθε εσωτερικός κόμβος ,ο οποίος δεν είναι η ρίζα, έχει τουλάχιστον δύο παιδιά και κάθε πλευρά αντιστοιχίζεται σε μία μη-μηδενική υπακολουθία του S.

Οι υπακολουθίες των πλευρών που εξέρχονται από τον ίδιο κόμβο δεν επιτρέπεται να έχουν κοινό τον πρώτο τους χαρακτήρα. Τέλος κύριο χαρακτηριστικό του δένδρου επιθεμάτων είναι το γεγονός ότι αν ενώσουμε τις ετικέτες μονοπατιών (path labels) που συναντάμε σε μια διαδρομή από τη ρίζα προς κάποιο από τα φύλλα, (έστω το φύλλο με αριθμό i), σχηματίζουμε το επίθεμα της συμβολοσειράς S που ξεκινά από την θέση i, δηλαδή το S[i..m].

Από τον παραπάνω ορισμό εξασφαλίζεται ότι υπάρχει Δέντρο Επιθεμάτων για κάθε συμβολοσειρά S. Εάν όμως ,για παράδειγμα , είχαμε την συμβολοσειρά S΄= CATTAGTACA και αφαιρέσουμε τους χαρακτήρες CA θα προέκύπτε η συμβολοσειρά S΄΄ = CATTAGTACA για την οποία το επίθεμα S[9…10]=CA δεν καταλήγει σε κάποιο φύλλο αλλά σε εσωτερικό κόμβο, αφού αποτελεί ταυτόχρονα και πρόθεμα της συμβολοσειράς. Για να αποφύγουμε αυτό το πρόβλημα σε κάθε συμβολοσειρά, S, προσθέτουμε έναν επιπλέον τελικό χαρακτήρα (τερματικό χαρακτήρα), ο οποίος δεν ανήκει στο αλφάβητο της συμβολοσειράς, άρα δεν εμφανίζεται πουθενά αλλού στην συμβολοσειρά. Συνήθως προστίθεται ως τερματικός χαρακτήρας (termination symbol) ο χαρακτήρας "$".

Ορίζουμε ως Ετικέτα Μονοπατιού (Path Label), από τη ρίζα του δέντρου σε κάποιο κόμβο, τη συμβολοσειρά που προκύπτει από τη συνένωση των υπακολουθιών που συναντάμε από τη ρίζα στον αντίστοιχο κόμβο. Μια απλοϊκή θεώρηση για την κατασκευή του Δέντρου Επιθεμάτων, για μια συμβολοσειρά S, περιλαμβάνει τα ακόλουθα βήματα:

1. Ένθεση μιας πλευράς στο δέντρο για το επίθεμα S[1…m]$,

2. Διαδοχική ένθεση των επιθεμάτων S[i…m]$, για i=2→m.

Στο πρώτο βήμα ο αλγόριθμος θεωρεί ότι το δέντρο αποτελείται μόνο από τη ρίζα και εισάγει σε αυτό το επίθεμα S[1..m]$, (ολόκληρη δηλαδή τη συμβολοσειρά και τον τερματικό χαρακτήρα), με αποτέλεσμα το δέντρο $N_1$ να αποτελείται από μια πλευρά με ετικέτα "S$" και ένα φύλλο αριθμημένο με τον αριθμό "1". Σε κάθε επόμενο βήμα δημιουργούμε το δέντρο $N_{i+1}$, από το δέντρο $N_i$, ως εξής: ξεκινώντας από τη ρίζα του δέντρου $N_i$, βρίσκουμε το μέγιστο σε μήκος μονοπάτι από τη ρίζα, για το οποίο η ετικέτα μονοπατιού ταιριάζει με κάποιο πρόθεμα του S[i+1..m]$, (συγκρίνοντας διαδοχικά τους χαρακτήρες). Έστω ότι στο χαρακτήρα S[k], με k≥i, έχουμε μη-ταίριασμα. Σε αυτή τη θέση υπάρχουν δύο δυνατές καταστάσεις: είτε βρισκόμαστε σε κάποιο κόμβο w του δέντρου $N_i$ είτε στο μέσο κάποιας πλευράς, μεταξύ των κόμβων (u,v). Στη δεύτερη περίπτωση χωρίζουμε την πλευρά στη μέση εισάγοντας ένα νέο εσωτερικό κόμβο, έστω w, αμέσως μετά τον τελευταίο χαρακτήρα του δέντρου που ταίριαζε σε κάποιον χαρακτήρα στο S[i+1…m]. Η νέα πλευρά (u, w), έχει ως ετικέτα μονοπατιού το τμήμα της πλευράς (u,v), που ταιριάζει στην υπακολουθία S[i+1…m], ενώ η πλευρά (w, v), αποκτά ως ετικέτα μονοπατιού το υπόλοιπο της πλευράς (u,v). Στη συνέχεια (το βήμα αυτό είναι κοινό και στην $1^η$ και στη $2^η$ περίπτωση), ο αλγόριθμος δημιουργεί μια νέα πλευρά (w, i+1), η οποία εκτείνεται από τον κόμβο w, σε ένα νέο φύλλο με αριθμό "i+1". Η νέα αυτή πλευρά έχει ως ετικέτα μονοπατιού από τη ρίζα στο φύλλο "i+1", το επίθεμα S[i+1..m]$.

Η απλοϊκή θεώρηση κατασκευής του Δέντρου Επιθεμάτων στοιχίζει $O(m^2)$ χρόνο, για ένα αλφάβητο πεπερασμένου μεγέθους. Πιο αποδοτικοί αλγόριθμοι για την κατασκευή του Δέντρου Επιθεμάτων, έχουν προταθεί στη σχετική βιβλιογραφία, ξεκινώντας με τον αλγόριθμο που παρουσίασε ο Weiner το 1973, ο McCreight το 1976 και τέλος το 1995 ο Ukkonen, ο οποίος απαιτεί γραμμικό χρόνο O(n).

Ειδικότερα, ο Weiner , που εισήγαγε τη δομή δεδομένων, έδωσε βέλτιστα 0(n)-χρόνο αλγόριθμο για την κατασκευή δέντρου επιθεμάτων μιας ακολουθίας n χαρακτήρων από ένα σταθερού μεγέθους αλφάβητο. Στο συγκριτικό μοντέλο, υπάρχει ένα τετριμμένο O(n log n)-χρόνου κατώτερο όριο που βασίζεται στην ταξινόμηση και ο αλγόριθμος του Weiner ταιριάζει σε αυτό το όριο.

Για τα αλφάβητα ακεραίων, ο γρηγορότερος γνωστός αλγόριθμος είναι O(n log n) χρόνου. Ο αλγόριθμος Ukkonen αρχίζει με ένα υποθετικό δέντρο επιθεμάτων που περιέχει μόνο τον πρώτο χαρακτήρα της σειράς. Κατόπιν διαπερνάει την σειρά προσθέτοντας τους διαδοχικούς χαρακτήρες μέχρι το δέντρο να γίνει πλήρες. Η απλοϊκή εφαρμογή αυτού του αλγορίθμου απαιτεί $O(n^2)$ ή ακόμα και χρόνο $O(n^3)$, όπου το n είναι ο αριθμός χαρακτήρων στη σειρά, αλλά με την εκμετάλλευση διάφορων αλγοριθμικών τεχνικών αυτό μπορεί να μειωθεί στο O (n) (γραμμικός) χρόνος. Τέλος ο αλγόριθμος του McCreight απαιτεί 28n bytes στην χειρότερη περίπτωση, όπου n είναι το μέγεθος του εισαγόμενου αλφαριθμητικού.

Στην πράξη οι απαιτήσεις χώρου είναι μικρότερες:

- Οι Manber και Myers δηλώνουν ότι η δική τους εφαρμογή δέντρων επιθεμάτων απασχολεί 18,8n και 24,4n bytes χώρου για πραγματικά αλφαριθμητικά εισόδου (μορφής text, code, DNA)

- Ο Karkkainen υποστηρίζει ότι ένα δέντρο επιθεμάτων μπορεί να υλοποιηθεί σε 15n-18n bytes. Δυστυχώς δεν έχει δείξει πως κάτι τέτοιο είναι εφικτό.

- Οι Chrochemore και Verin υποστηρίζουν πως τα δέντρα επιθεμάτων απαιτούν 32,7n bytes για τις DNA ακολουθίες.

- Το strmat λογισμικό από τους Knight, Gusfield και Stoye υλοποιεί τα δέντρα επιθεμάτων σε 24n-28n bytes για αλφαριθμητικά μεγέθους $2^{23}$=8,388,608. Παρόλα αυτά , το strmat μπορεί να διαχειρίζεται σύνολα από αλφαριθμητικά και δεν είναι ξεκάθαρο εάν οι απαιτήσεις χώρου είναι εξαιτίας αυτού του ιδιαίτερου χαρακτηριστικού του.

- Munro πρόσφατα περιέγραψε μια αναπαράσταση δέντρου επιθεμάτων η οποία απαιτεί $n[\log_2 n]+o(n)$ bits. Παρόλα αυτά περιορίζεται στην αναζήτηση αλφαριθμητικών κομματιών και δεν είναι ξεκάθαρο εάν υπάρχει αλγόριθμος γραμμικού χρόνου για άμεση αναπαράσταση. Ως αποτέλεσμα, κάποιος πρέπει πρώτα να κατασκευάσει ένα δέντρο επιθεμάτων με την συνήθη, λιγότερο αποδοτική σε απαιτήσεις χώρου αναπαράσταση. Έτσι η προσέγγιση Munro θυσιάζει την μεταβλητότητα και δεν δίνει ουσιαστικά πλεονεκτήματα στις απαιτήσεις χώρου.

## *3.2 Εφαρμογή για τη μελέτη γονιδιωματικών ακολουθιών με την χρήση δέντρων επιθεμάτων*

Οι περισσότεροι κατασκευαστικοί αλγόριθμοι δέντρων επιθεμάτων δεν κλιμακώνουν καλά για μεγάλο όγκο αρχείων. Το υψηλό σύνολο ανά χαρακτήρα αναγκάζει τις δομές δεδομένων να ξεπεράσουν την κύρια μνήμη και η φτωχή τοπικότητα της αναφοράς καθιστά την αποδοτική διαχείριση buffer δύσκολη. Έτσι καταλήγουμε σε μία κατασκευαστική τεχνική disk-based, που αποκαλείται «Top-Down Disk-based» ή αλλιώς TDD.

Η τεχνική TDD αποτελείται από ένα κατασκευαστικό αλγόριθμο , αποκαλούμενο PWOTD , και μια σχετική στρατηγική διαχείρισης buffer. Ο προαναφερθείσας αλγόριθμος είναι βασισμένος στον wotdeager αλγόριθμο που προτείνεται από τον Kurtz. Ουσιαστικά βελτιώνεται αυτός ο αλγόριθμος με την χρήση μιας φάσης χωρισμού που επιτρέπει την κατασκευή μεγαλύτερων και ανεξάρτητων sub-tree στη μνήμη.

Το δέντρο επιθεμάτων αντιπροσωπεύεται σε ένα γραμμικό πίνακα. Στο παρακάτω σχήμα επεξηγείται η αναπαράσταση στη μνήμη ενός δέντρου επιθεμάτων, της ακολουθίας ATTAGTACA$. Οι σκιασμένες καταχωρήσεις στον πίνακα αντιπροσωπεύουν τους κόμβους φύλλων, και όλες οι υπόλοιπες καταχωρήσεις αντιπροσωπεύουν τους κόμβους μη-φύλλων. Το R στη χαμηλότερη δεξιά γωνία μιας καταχώρησης δηλώνει ένα δεξιό παιδί.

Ένας διακλαδωμένος κόμβος αντιπροσωπεύεται από δύο ακέραιους αριθμούς. Ο πρώτος ακέραιος είναι ένας δείκτης στην αρχική ακολουθία, ο χαρακτήρας σε αυτόν τον δείκτη είναι χαρακτήρας- εναρκτής της συγκεκριμένης υπακολουθίας. Το πλήθος επαναλήψεων αυτής της υπακολουθίας μπορεί να βρεθεί από το πλήθος των παιδιών της. Ο δεύτερος ακέραιος δείχνει στο πρώτο παιδί. Οι κόμβοι φύλλων δεν έχουν δεύτερη καταχώρηση. Ο κόμβος φύλλων απαιτεί μόνο τον αρχικό δείκτη στην ακολουθία. Το τέλος της ακολουθίας σηματοδοτείται από τον τερματικό χαρακτήρα.

### Ο αλγόριθμος PWOTD

Αποτελείται από δύο φάσεις. Στην πρώτη φάση χωρίζουμε τις υπακολουθίες της αρχικής ακολουθίας σε $|A|^{prefixlen}$ ,όπου |A| είναι το μέγεθος του αλφαβήτου της ακολουθίας και prefixlen είναι το μήκος του επιθέματος . Το βήμα διαχωρισμού εκτελείται όπως παρακάτω:

Η αρχική ακολουθία διαβάζεται από αριστερά στα δεξιά. Σε κάθε θέση δείκτη i το prefixlen χρησιμοποιείται για τον καθορισμό των $|A|^{prefixlen}$ διαχωρισμών. Αυτός ο δείκτης έπειτα γράφεται στο υπολογισμένο buffer διαχωρισμού. Στο τέλος του

διαβάσματος κάθε διαχωρισμένο κομμάτι θα περιέχει δείκτες επιθεμάτων που έχουν το ίδιο επίθεμα μεγέθους prefixlen.

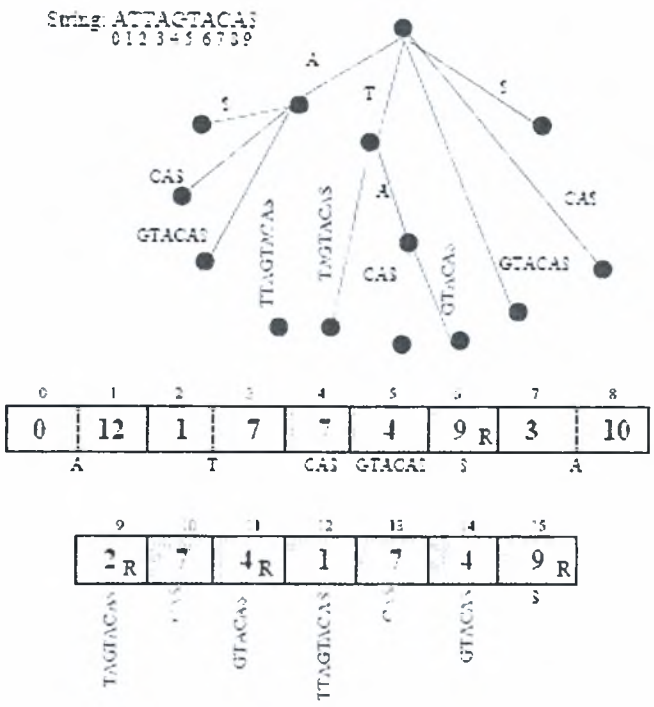
Για να γίνει περισσότερο κατανοητό παρατίθεται το παρακάτω παράδειγμα. Διαχωρίζουμε την ακολουθία ATTAGTACA$ χρησιμοποιώντας prefixlen 1 θα δημιουργούνταν 4 υπακολουθίες, μία για κάθε σύμβολο του αλφαβήτου. Ο διαχωρισμός επιθεμάτων για το γράμμα A θα ήταν {0, 3, 6} αντιπροσωπεύοντας τις υπακολουθίες {ATTAGTACA$, AGTACA$, ACA$,A$}. Ο διαχωρισμός επιθεμάτων για το γράμμα T θα ήταν {1, 2, 5} αντιπροσωπεύοντας τις υπακολουθίες {TTAGTACA$, TAGTACA$, TACA$}.

Στη δεύτερη φάση χρησιμοποιούμε τον αλγόριθμο wotdeager για να κτίσουμε το δέντρο επιθεμάτων χρησιμοποιώντας top down κατασκευή. Ο αλγόριθμος PWOTD απαιτεί 4 δομές για τη κατασκευή του δέντρου επιθεμάτων: έναν πίνακα αλφαριθμητικών(για συντομία Π.Α.), έναν πίνακα Επιθεμάτων(για συντομία Π.Ε.), έναν Προσωρινό πίνακα (για συντομία Π.Π.), και το Δέντρο Επιθεμάτων(για συντομία Δ.Ε.).

Ο Π.Ε. γεμίζει αρχικά από επιθέματα το διαχωρισμό των πρώτων prefixlen χαρακτήρων. Χρησιμοποιώντας το ίδιο παράδειγμα με πριν, ATTAGTACA$, έχουμε τον διαχωρισμό T , όπου τα επιθέματα βρίσκονται στις θέσεις 1,2,5. Καθώς όλα αυτά τα επιθέματα μοιράζονται το ίδιο πρόθεμα ,T, προσθέτουμε 1 σε κάθε καταχώρηση έτσι ώστε να δημιουργήσουμε τον νέο πίνακα επιθεμάτων {2, 3, 6}. Το επόμενο βήμα περιλαμβάνει την ταξινόμηση αυτού του πίνακα βάσει του πρώτου χαρακτήρα. Οι πρώτοι χαρακτήρες αυτών των επιθεμάτων είναι οι {T, A, A}. Η ταξινόμηση γίνεται με την χρήση ενός αποτελεσματικού αλγορίθμου καλούμενου count-sort σε γραμμικό χρόνο. Σε μία διαπέραση, για κάθε χαρακτήρα του αλφαβήτου, μετράμε την συχνότητα εμφάνισης του ως πρώτο χαρακτήρα για κάθε επίθεμα, και αντιγράφουμε τους δείκτες αυτούς σε έναν προσωρινό πίνακα. Παρατηρούμε ότι ο υπολογισμός για το A είναι 2 ,για το T είναι 1 και για το C ,G και $ είναι 0.

Χρησιμοποιούμε αυτές τις μετρήσεις για να καθορίσουμε τα όρια κάθε ομάδας . Η ομάδα A ξεκινάει από την θέση 0 με δύο καταχωρήσεις, η ομάδα T αρχίζει στην θέση 2 με μία καταχώρηση. Έπειτα διαπερνάμε τον Π.Π. και παράγουμε νέο πίνακα επιθεμάτων βάση τους πρώτους χαρακτήρες. Ο Π.Ε. τώρα είναι {3,6,2}.Η ομάδα A έχει 2 μέλη άρα είναι διασυνδεδεμένος κόμβος. Αυτά τα δύο επιθέματα καθορίζουν το υπο-δέντρο κάτω από τον κόμβο. Δεσμεύεται χώρος στο Δ.Ε. έτσι ώστε να γραφεί ο διασυνδεδεμένος (μη-φύλλο) κόμβος όταν θα επεκταθεί, έπειτα ο κόμβος αποθηκεύεται σε μία στοίβα.

Καθώς η ομάδα του T έχει μόνο ένα μέλος, είναι φύλλο και θα γραφεί αυτόματα στο Δ.Ε. Επειδή δεν υπάρχουν άλλα παιδιά, δεν κρατιούνται επιπλέον θέσεις στη στοίβα και ο κόμβος που περιέχει θα χρησιμοποιηθεί πρώτος. Όταν είναι να χρησιμοποιηθεί, και κατά συνέπεια αφαιρεθεί από τη στοίβα, υπολογίζεται το Μέγιστο Κοινό Πρόθεμα (LCP) για κάθε κόμβο της ομάδας. Εξετάζοντας τις θέσεις 4 (G) και 7(C) καθορίζουμε το LCP ίσο με 1. Κάθε δείκτης επιθέματος αυξάνεται κατά LCP και το αποτέλεσμα εκτελείται όπως και πριν. Ο υπολογισμός συνεχίζει μέχρι να επεκταθούν όλοι οι κόμβοι και η στοίβα μείνει άδεια.




**Εικόνα 3.2 Απεικόνιση του πίνακα επιθεμάτων (Π.Ε.) καθώς και του δέντρου επιθεμάτων (Δ.Ε.) για την ακολουθία ATTAGTACA$**

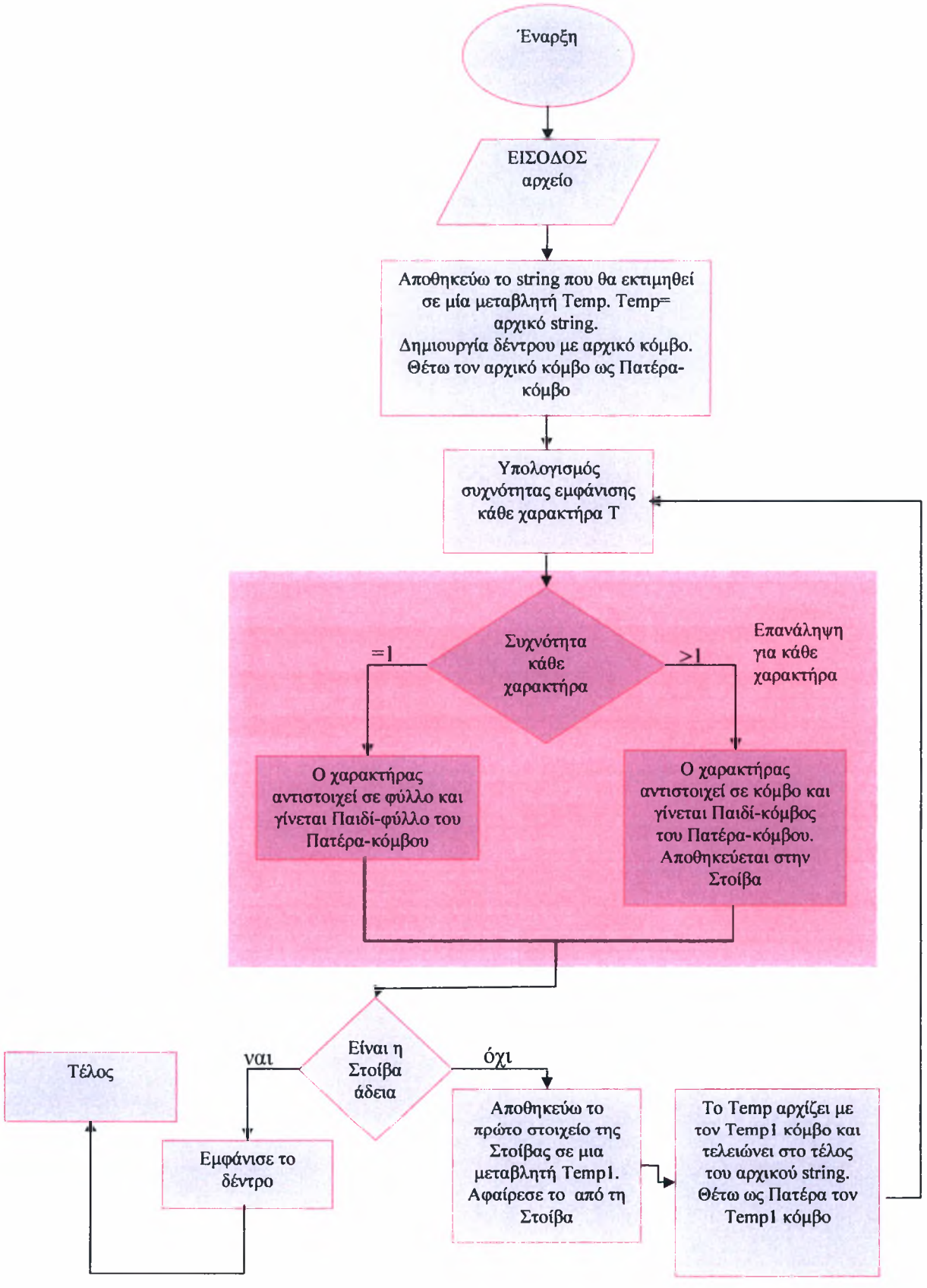
**Algorithm PWOTD(*String,prefixlen*)**
**Phase1**:
Scan the *String* and partition *Suffixes* based on the first *prefixlen* symbols of each suffix
**Phase2**: Do for each partition:
1. START BuildSuffixTree
2. Populate *Suffixes* from current partition
3. Sort *Suffixes* on first symbol using *Temp*
4. Output branching and leaf nodes to the *Tree*
5. Push the nodes pointing to an unevaluated range onto the *Stack*

While *Stack* is not empty
6. Pop a node
7. Find the Longest Common Prefix (LCP) of all the suffixes in this range by checking the *String*
8. Sort the range in *Suffixes* on the first symbol using *Temp*
9. Write out branching nodes or leaf nodes to *Tree*
10. Push the nodes pointing to an unevaluated range onto the *Stack*

11. END

**Εικόνα 3.3 Απεικόνιση του ψευτοκώδικα υλοποίησης του PWOTD**

Παρακάτω παρουσιάζεται και το λογικό διάγραμμα , που αναπαριστά την υλοποίηση της κατασκευής του δέντρου επιθεμάτων.




**Σχήμα 3.1 Λογικό διάγραμμα για την εμφάνιση του δέντρου (suffix tree)**

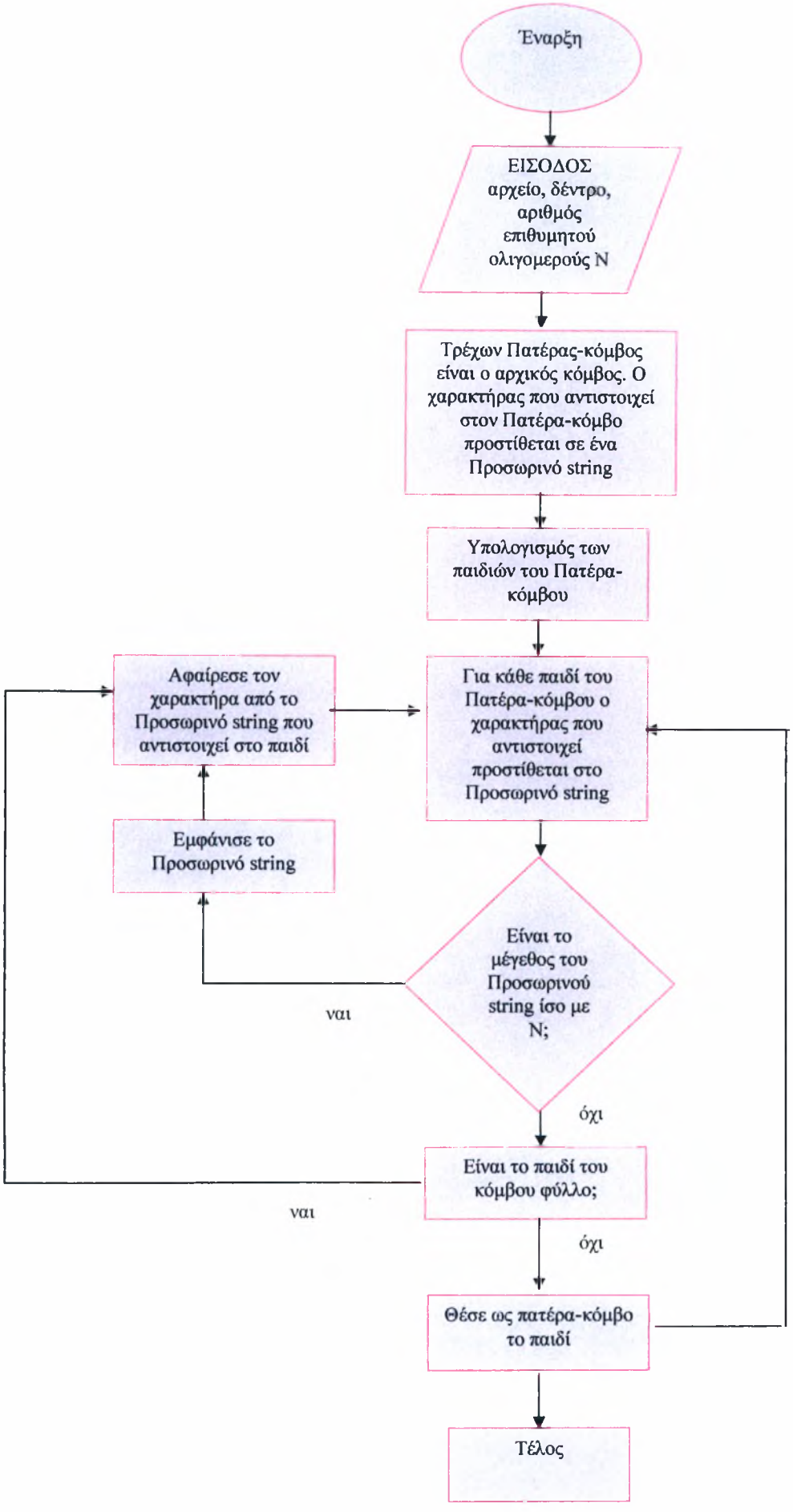
Αφού κατασκευαστεί το δέντρο επιθεμάτων, επόμενος στόχος είναι η εξαγωγή των ολιγομερών και της συχνότητα εμφάνισης στην αρχική ακολουθία και των θέσεων στο αρχικό αρχείο που εμφανίζονται αυτές οι επαναλήψεις..

Ο αλγόριθμος απαιτεί 3 δομές για τον υπολογισμό των επαναλήψεων. Το αλφαριθμητικό εισόδου (input string), το δέντρο επιθεμάτων (Δ.Ε.), και τον επιθυμητό αριθμό ολιγομερούς (N), εάν επιθυμούμε να εμφανίσουμε όλα τα 6μερή που εμφανίζονται στην αρχική ακολουθία τότε N=6.

Διαλέγουμε τον πρώτο κόμβο από το Δ.Ε. έχουμε ένα προσωρινό αλφαριθμητικό όπου προσθέτουμε τον χαρακτήρα που αντιστοιχεί σε αυτόν τον κόμβο. Στην συνέχεια υπολογίζουμε τα παιδιά αυτού του κόμβου. Για κάθε παιδί του κόμβου προσθέτουμε τον χαρακτήρα που αντιστοιχεί στο προσωρινό αλφαριθμητικό. Ελέγχουμε εάν το μέγεθος του προσωρινού αλφαριθμητικού έχει φτάσει στο ζητούμενο μέγεθος N. Εάν έχει φτάσει τότε έχουμε ένα ταίριασμα. Έπειτα υπολογίζουμε το πλήθος των παιδιών του τελευταίου χαρακτήρα, αυτό το πλήθος μας δείχνει το πλήθος επαναλήψεων και οι θέσεις των παιδιών του κόμβου αυτού μας δίνουν τις θέσεις που εμφανίζεται το ολιγομερές στο αρχικό αρχείο. Έπειτα το αλφαριθμητικό αποθηκεύεται σε 2 αρχεία, στο πρώτο αρχείο εμφανίζεται το κάθε ολιγομερές σε αντιστοιχία με τον αριθμό επαναλήψεων, και στο δεύτερο ομοίως αλλά και με πρόσθετες τις θέσεις. Έπειτα αφαιρούμε από το προσωρινό αλφαριθμητικό τον τελευταίο χαρακτήρα και συνεχίζουμε την εκτέλεση για το επόμενο παιδί του κόμβου.

Εάν όμως το προσωρινό αλφαριθμητικό δεν έχει φτάσει στο ζητούμενο μέγεθος, ελέγχουμε εάν ο κόμβος είναι κόμβος φύλλο ή διασυνδεδεμένος κόμβος. Εάν πρόκειται για κόμβο φύλλο έχουμε μη ταίριασμα. Αφαιρείται ο τελευταίος χαρακτήρας από το προσωρινό αλφαριθμητικό και συνεχίζεται η αναζήτηση με το επόμενο παιδί. Εάν όμως έχουμε διασυνδεδεμένο κόμβο, τότε συνεχίζουμε την αναζήτηση διαλέγοντας ως κόμβο έναρξης τον κόμβο αυτό.

Η αναζήτηση συνεχίζεται για όλα τα παιδιά του αρχικού κόμβου. Παρακάτω εμφανίζεται το λογικό διάγραμμα για τον αλγόριθμο εμφάνισης των ολιγομερών.

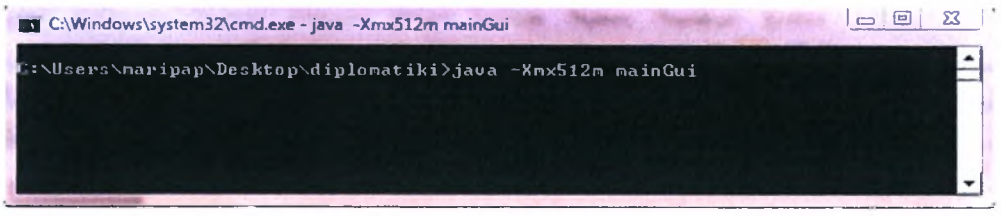



**Σχήμα 3.2 Λογικό διάγραμμα για την εμφάνιση των επαναλήψεων των ολιγομερών**

# ΚΕΦΑΛΑΙΟ 4

## *4.1 Εισαγωγή*

Η εφαρμογή έχει γραφτεί στην γλώσσα προγραμματισμού JAVA. Η εκτέλεση, γίνεται από την κονσόλα με την εντολή >java –Xmx512m mainGui, με αυτή την εντολή δεσμεύουμε επιπλέον RAM 512 MB από αυτή που μας παρέχει ως default η Java, όπως φαίνεται και παρακάτω:




Η εφαρμογή δέχεται αρχεία .dna, .txt . Χρειάζεται μια τροποποίηση όμως σε κάθε αρχείο, καθώς χρειάζεται να προστεθεί ο τερματικός χαρακτήρας στο τέλος κάθε ακολουθίας '\'.

## *4.2 Ανώτατο όριο μεγέθους εισαγόμενου αλφαριθμητικού και απαιτήσεις συστήματος*

Συνήθως η διαθέσιμη η μνήμη οριοθετεί το μέγιστο μήκος της σειράς εισαγωγής που μπορεί να υποβληθεί σε μια επεξεργασία. Εντούτοις, σε μερικούς υπολογιστές με πάρα πολύ μεγάλη μνήμη, κάποιος πρέπει να λάβει υπόψη του και το διαθέσιμο χώρο διευθύνσεων. Η παρούσα εφαρμογή διαχειρίζεται αρχεία μεγέθους μέχρι 406.260 (bp). Η εφαρμογή μεταγλωττίστηκε με τον JCreator 4.00 Pro version 4.00.028. έπειτα το πρόγραμμα εκτελέστηκε στην κονσόλα cmd των Windows Vista, όπως περιγράφεται παραπάνω.

Ο υπολογιστής που εκτελέστηκε η εφαρμογή έχει CPU Intel Core 2 Duo στα 1,66 GHz και μνήμη RAM 2 GB. Σε περίπτωση που είχαμε σύστημα με μεγαλύτερη RAM θα μπορούσαμε να δεσμεύσουμε περισσότερη μνήμη με την εντολή:

>java –Xmx4G mainGui

Σε μια τέτοια περίπτωση, λογικά θα μπορούσαμε να υποστηρίξουμε μεγαλύτερου μήκους ακολουθίες.

## *4.3 Εκτέλεση της εφαρμογής*

Η εφαρμογή αποτελείται από πέντε αρχεία java. Παίρνει σαν είσοδο μια ακολουθία DNA, την διαβάζει και κατασκευάζει ένα δέντρο επιθεμάτων. Με άλλα λόγια, κατασκευάζει ένα δέντρο το οποίο περιέχει όλες τις δυνατές υπακολουθίες της αρχικής ακολουθίας DNA. Στην συνέχεια υπολογίζει τα βάρη στους κόμβους του δέντρου. Υπολογίζονται δηλαδή τα παιδιά κάθε κόμβου και αυτός ο αριθμός αποθηκεύεται σε κάθε κόμβο.

Κατά την διάρκεια της εκτέλεσης εμφανίζεται ένα παράθυρο. Αρχικά ζητείται η εισαγωγή του ονόματος του αρχείου προς εκτίμηση. Όταν έχει ήδη διαβαστεί το αρχείο εμφανίζεται ένα μήνυμα «File Reading....ok». Μετά την κατασκευή του δέντρου εμφανίζεται το μήνυμα «Create Tree....ok». Τέλος όταν υπολογιστούν και τα βάρη εμφανίζεται το μήνυμα «Calculate Weights....ok». Στο κάτω μέρος του παραθύρου εμφανίζονται ο χρόνος και η μνήμη που δεσμεύτηκε κατά τη διάρκεια της εκτέλεσης. Με το πέρας αυτής της διαδικασίας ενεργοποιείται η επιλογή για τον καθορισμό του μεγέθους των ολιγομερών που αναζητούμε. Εάν χρειαζόμαστε 6μερη δίνεται ο αριθμός 6, εάν αναζητούμε 21μερή , ο αριθμός 21 κτλ.

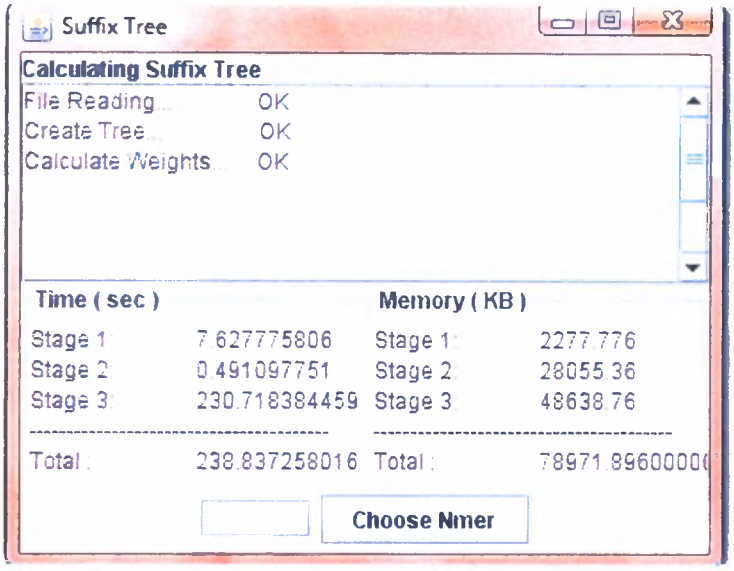
Μετά την επιλογή αυτή δημιουργούνται 2 αρχεία τα ονόματα των οποίων εμφανίζονται στο παράθυρο ,τα File_Name_Nmers.tree και File_Name_index_Nmers.tree . Το πρώτο αρχείο περιέχει όλα τα ολιγομερή που προκύπτουν, με αλφαβητική σειρά. Δίπλα σε κάθε ολιγομερές εμφανίζεται επίσης και το πλήθος των επαναλήψεων του μέσα στην αρχική ακολουθία. Στο δεύτερο αρχείο εμφανίζεται κάθε ολιγομερές, το πλήθος επαναλήψεων αλλά και οι ακριβείς θέσεις όπου εμφανίζεται στην αρχική ακολουθία.

Το πρόγραμμα χρησιμοποιήθηκε για την εκτίμηση 5 αρχείων FASTA. Τα αρχεία FASTA είναι μία από τις πιο διαδεδομένες μορφές με τις οποίες είναι αποθηκευμένα βιολογικά δεδομένα στις βάσεις δεδομένων. Ειδικότερα , ένα αρχείο FASTA -μορφής μπορεί να περιέχει περισσότερες τις μιας ακολουθίας. Μια ακολουθία σε FASTA -μορφή ξεκινάει με μια περιγραφή για την ακολουθία στης πρώτη γραμμή και συνεχίζει με τη συγκεκριμένη ακολουθία. Η γραμμή περιγραφής ξεκινάει με το σύμβολο (“>”). Παρακάτω εμφανίζεται μια μορφή ακολουθίας σε FASTA –μορφή .

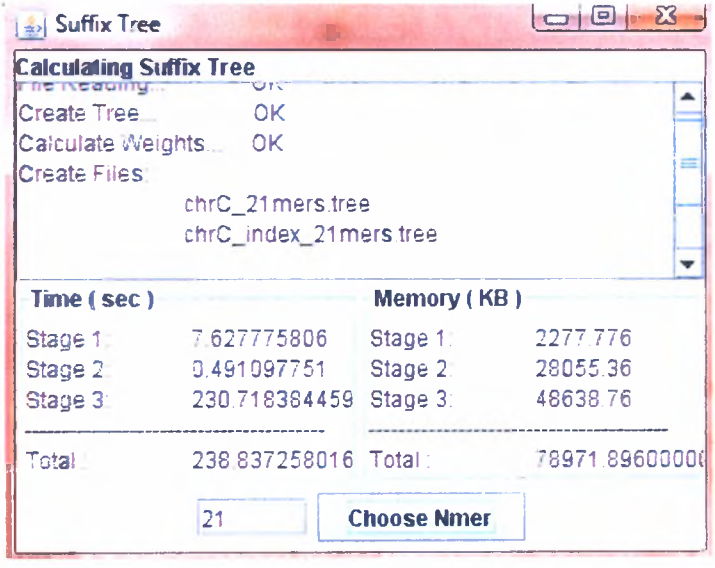
>10:ENSMUSG00000000290:protein_coding:1
AGCATGAGTTATCATAATCAAGCAGATGTGACCCCCTCAGACCACGCCTC
CTCCCCTCTGCAAACACAACGTGGCTTACAGCTCACCCCAGTGCTGCCAA
GGATCCAAAAGCCTGCTCGGTTTCTTTCCGCCATTATATCAAG

Συγκεκριμένα η εφαρμογή χρησιμοποιήθηκε για τις παρακάτω ακολουθίες:

Από το Arabidopsis genome η ακολουθία chrC kai chrM, την ακολουθία Enterobacteria phage phiX174 genome, Human herpesvirus 4 type 2 genome, και UP_seq_antagomir. Τα αποτελέσματα όσον αφορά το χρόνο και την μνήμη φαίνονται παρακάτω:



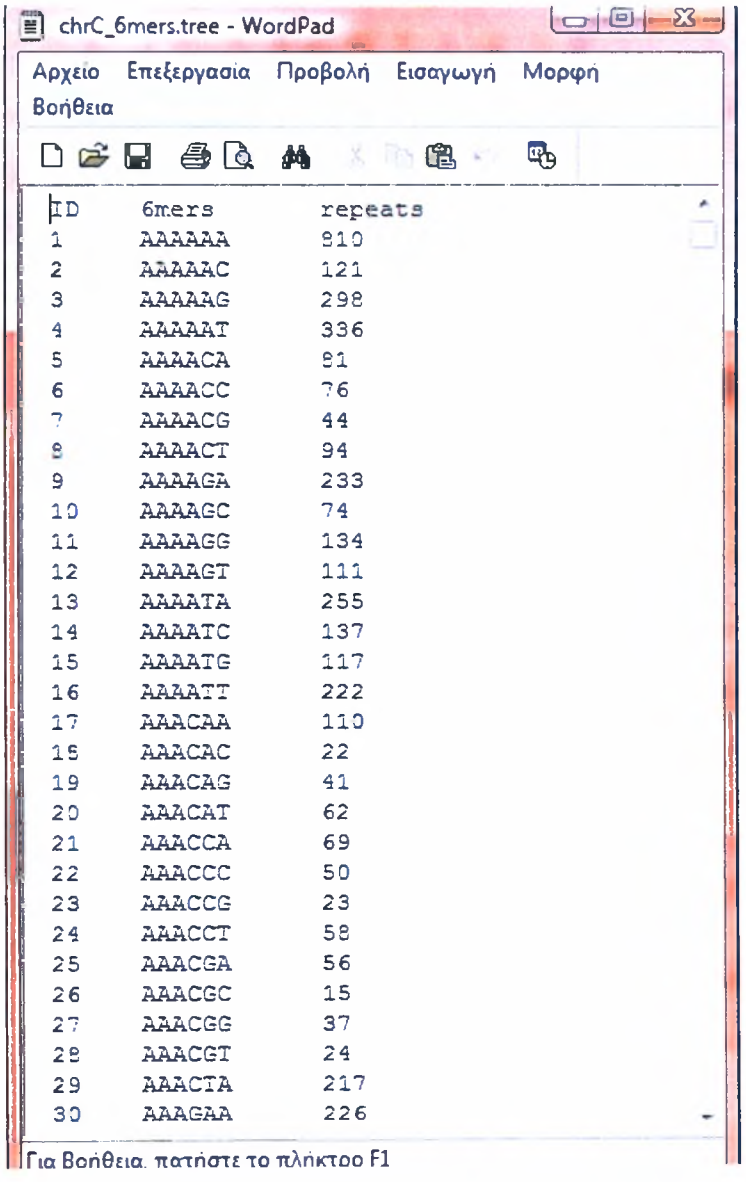

Παρακάτω φαίνονται τα διάφορα στάδια της εκτέλεσης, καθώς επίσης και ο χρόνος και η μνήμη που χρειάστηκε να δεσμευτεί σε κάθε στάδιο.




Εδώ φαίνεται η επιλογή του μεγέθους του ολιγομερούς και η εμφάνιση των αρχείων που δημιουργούνται.

Ένα παράδειγμα των παραγόμενων αρχείων , είναι της ακολουθίας Arabidopsis το chrC για 6μερείς ακολουθίες.

**1° ΑΡΧΕΙΟ (chrC_6mers.tree)**

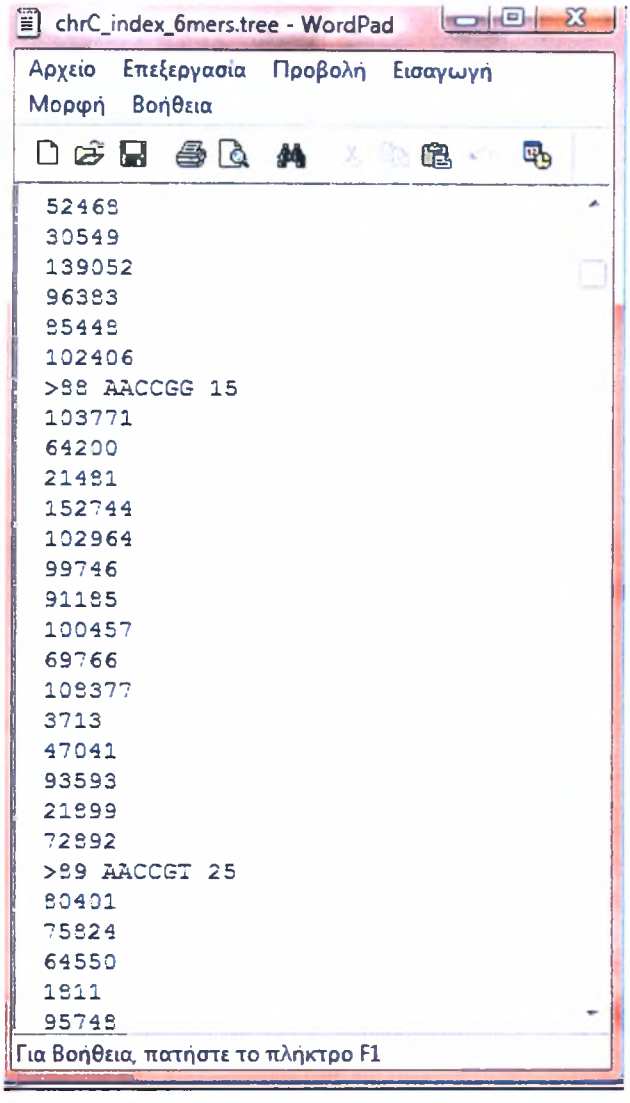

chrC_6mers.tree - WordPad

Αρχείο Επεξεργασία Προβολή Εισαγωγή Μορφή Βοήθεια

| ID | 6mers | repeats |
|---|---|---|
| 1 | AAAAAA | 810 |
| 2 | AAAAAC | 121 |
| 3 | AAAAAG | 298 |
| 4 | AAAAAT | 336 |
| 5 | AAAACA | 81 |
| 6 | AAAACC | 76 |
| 7 | AAAACG | 44 |
| 8 | AAAACT | 94 |
| 9 | AAAAGA | 233 |
| 10 | AAAAGC | 74 |
| 11 | AAAAGG | 134 |
| 12 | AAAAGT | 111 |
| 13 | AAAATA | 255 |
| 14 | AAAATC | 137 |
| 15 | AAAATG | 117 |
| 16 | AAAATT | 222 |
| 17 | AAACAA | 110 |
| 18 | AAACAC | 22 |
| 19 | AAACAG | 41 |
| 20 | AAACAT | 62 |
| 21 | AAACCA | 69 |
| 22 | AAACCC | 50 |
| 23 | AAACCG | 23 |
| 24 | AAACCT | 58 |
| 25 | AAACGA | 56 |
| 26 | AAACGC | 15 |
| 27 | AAACGG | 37 |
| 28 | AAACGT | 24 |
| 29 | AAACTA | 217 |
| 30 | AAAGAA | 226 |

Για Βοήθεια, πατήστε το πλήκτρο F1

## 2ο ΑΡΧΕΙΟ (chrC_index_6mers.tree)

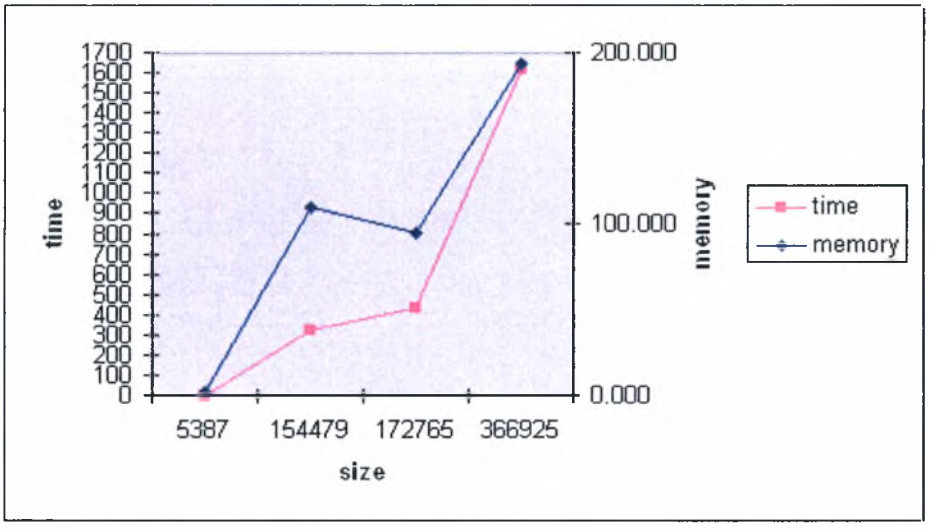


## *4.4 Πειραματικά αποτελέσματα και Συγκρίσεις*

Τα συνολικά αποτελέσματα φαίνονται στους 2 παρακάτω πίνακες και απεικονίζονται στο ακόλουθο διάγραμμα:

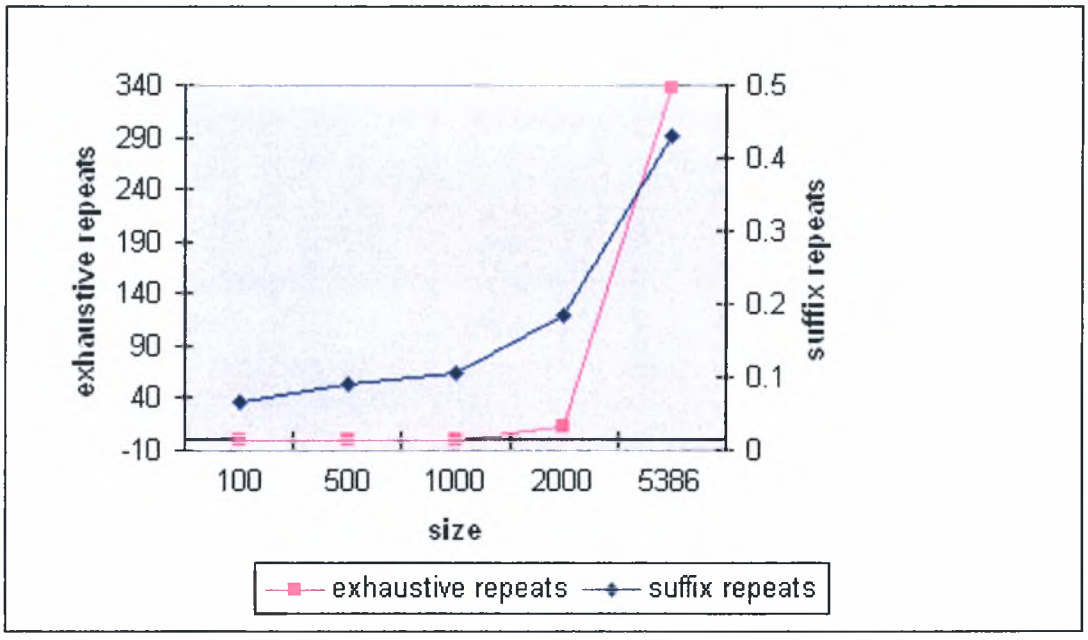
| Size (bp) | | Time (sec) | Memory (MB) |
|---|---|---|---|
| ***chrC*** | | | |
| *154479* | *File Reading* | 7.628 | 2.224 |
| | *Create Tree* | 0.491 | 27.398 |
| | *Calculate Weights* | 230.718 | 47.499 |
| ***chrM*** | | | |
| *366925* | *File Reading* | 66.147 | 3.467 |
| | *Create Tree* | 1.165 | 65.619 |
| | *Calculate Weights* | 1280.411 | 170.246 |
| ***Enterobacteria*** | | | |
| *5387* | *File Reading* | 0.046 | 0.686 |
| | *Create Tree* | 0.0709 | 0.851 |
| | *Calculate Weights* | 0.395 | 1.701 |
| ***Human_herpesvirus*** | | | |
| *172765* | *File Reading* | 17.436 | 2.824 |
| | *Create Tree* | 2.818 | 29.690 |
| | *Calculate Weights* | 377.9120 | 45.754 |

| sequences | *Size (bp)* | *time ( sec )* | *Memory (MB)* |
|---|---|---|---|
| ***chrC*** | 154479 | 321.17 | 109.599 |
| ***chrM*** | 366925 | 1615.37 | 193.331 |
| ***Enterobacteria*** | 5387 | 0.309 | 2.394 |
| ***Human_herpesvirus*** | 172765 | 438.08 | 95.193 |




Παρακάτω δίνονται αποτελέσματα από την σύγκριση της μεθόδου με δέντρα επιθεμάτων και τις μεθόδου εξαντλητικών επαναλήψεων. Η δεύτερη μέθοδος διαπερνάει σειριακά όλο το αρχείο και ψάχνει ένα προς ένα κάθε ολιγομερή και το πλήθος των επαναλήψεών του. Το παράδειγμα έγινε στο Enterobacteria phage phiX174, complete genome„με 5386 bp.

| Size (bp) | Exhaustive repeats ( sec ) | Suffix ( sec ) |
|---|---|---|
| **100** | 0,008 | 0,066 |
| **500** | 0,107 | 0,090 |
| **1000** | 1,208 | 0,105 |
| **2000** | 14,251 | 0,185 |
| **5386 (max)** | 337,977 | 0,430 |




Παρατηρώντας τα παραπάνω δεδομένα βλέπουμε ότι ο αλγόριθμος TDD (suffix tree) έχει αρκετά μικρότερες απαιτήσεις . από άποψη χρόνου, σε σχέση με την μέθοδο exhaustive repeats. Αυτό ήταν αναμενόμενο καθώς ξέρουμε ότι η χρονική διάρκεια εκτέλεσης της μεθόδου exhaustive repeats είναι της τάξης $O(n!)$ σε αντίθεση με τον TDD ο οποίος είναι της τάξης $O(n^2)$ , όπου n είναι το πλήθος γραμμάτων της ακολουθίας που βάζουμε σαν είσοδο στις παραπάνω μεθόδους.
Ως αποτέλεσμα η μέθοδος exhaustive repeats δεν είναι αποδοτική για πολύ μεγάλα αρχεία. Παρακάτω παρουσιάζεται ένας συγκριτικός πίνακας μεταξύ **Exhaustive repeats** και επαναλήψεων με χρήση **Suffix tree.**

| sequences | Exhaustive repeats ( sec ) | Suffix ( sec ) |
|---|---|---|
| **chrC** | ∞ | 321,517 |
| **chrM** | ∞ | 1615,737 |
| **Enterobacteria** | 337,977067197 | 0,4309 |
| **Human_herpesvirus** | ∞ | 438,808 |
| **UP_seq_antagomir** | ∞ | 2014,17458 |

# Συμπεράσματα

Η Βιολογία στον 21ο αιώνα μετασχηματίζεται από καθαρά εργαστηριακή επιστήμη και σε επιστήμη πληροφοριών. Οι πληροφορίες αυτές περιλαμβάνουν αναλυτικές απόψεις της ακολουθίας του DNA, της έκφρασης του RNA, των πρωτεϊνικών αλληλεπιδράσεων ή των μοριακών στερεοδιατάξεων. Οι βιολογικές μελέτες ολοένα και περισσότερο ξεκινούν με την μελέτη τεράστιων βάσεων δεδομένων για να διατυπωθούν συγκεκριμένες υποθέσεις ή να σχεδιαστούν πειράματα μεγάλης κλίμακας.

Η παρούσα διπλωματική εργασία είχε παράλληλα δύο στόχους: ο πρώτος ήταν να συστηματοποιήσει όσο το δυνατόν καλύτερα τις τάσεις που υπάρχουν σήμερα στο χώρο της βιοπληροφορικής, ιδιαίτερα δε αυτές που μελετούνε επαναληπτικές ακολουθίες και δέντρα επιθεμάτων. Αυτό δεν θα μπορούσε να γίνει αν προηγουμένως δεν γινόταν μια συνοπτική παρουσίαση της σύγχρονης βιολογίας του κυττάρου και της γενετικής των οργανισμών.

Ο δεύτερος στόχος ήταν να προσπαθήσουμε να δημιουργήσουμε μια εφαρμογή που να χρησιμοποιεί δέντρα επιθεμάτων και να δέχεται βιολογικές ακολουθίες σαν δεδομένα. Να γίνεται εύρεση των επαναληπτικών ακολουθιών καθώς και μια προσέγγιση στην χρονική πολυπλοκότητα της υλοποίησης. Μετά την ολοκλήρωση της εργασίας θεωρούμε ότι αυτό μπορεί να αποτελέσει τη βάση για μια ακόμα πιο πετυχημένη προσέγγιση του αντικειμένου της βιοπληροφορικής.

# Βιβλιογραφία


1. Βιβλίο Βιοπληροφορικής- Andreas D. Baxevanis, B.F. Francis Quellette- Δεύτερη Έκδοση
2. W. Becker, L. Kleinsmith, J. Hardin – The world of the cell – Benjamin Cummings, 2002
3. Φίλιπ Κουρίλσκι – DNA, Το νήμα της ζωής και οι γενετικές επεμβάσεις – Εκδόσεις Ράππα, 1993
4. K. P. Talaro, Arthur Talaro – Foundations in microbiology – WCB/McGraw-Hill, 1999
5. Schulz G.E. and Schirmer R.H. - Principles of Protein Structure – 1979
6. Α.Θ. Καλοφούτης, Κ.Ε. Σέκερης - Ιατρική Βιοχημεία, Κλινική Προσέγγιση – 2000
7. Dickerson, R.E. and Geis, I. - The Structure and Action of Proteins – 1969
8. K.Arms and P.S. Camp – Biology, a journey into life – Sounders College Publishing, 1991
9. http://genome.tugraz.at/Theses/Rader2005.pdf
10. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
11. https://www.cbs.dtu.dk/services/GenomeAtlas/
12. http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/RNA.htm
13. http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/DNA.htm
14. http://nayfahsai.blogspot.com/2007/11/intro-about-cell.html
15. http://www.genome.gov/search.cfm
16. http://dnadir.blogcu.com/dna-ne-rna_4928501.html
17. http://en.wikipedia.org/wiki/DNA_replication
18. βιολογία γενικής παιδείας γ λυκείου, Από τον Γιώργο Ντράνο: Καθηγητή Βιολογίας του 3ου ΕΛ Αργυρούπολης
19. J. Cohen - Bioinformatics-An introduction for Computer Scientists – ACM Computing Surveys, Vol.36,No.2, June 2004
20. Introduction in Computational Biology – http://www.lisha.ufsc.br/~guto
21. Gusfield D. – Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology – Cambridge University Press, 1997
22. Rabiner, L. R. – A tutorial on hidden Markov models and selected applications in speech recognition – Proc. IEEE 77, 2, 257-286
23. Ο δικτυακός τόπος http://www.BioinformaticsOnline.org διαθέτει μια πλούσια συλλογή εργαλείων που χρησιμοποιούνται τόσο για εύρεση γονιδίων, όσο και για πρόβλεψη υποκινητών
24. http://pir.georgetown.edu/
25. http://www.ebi.ac.uk/
26. http://www.ncbi.nlm.nih.gov/
27. http://en.wikipedia.org/wiki/Category:Repetitive_DNA_sequences
28. http://en.wikipedia.org/wiki/Tandem_repeats
29. http://en.wikipedia.org/wiki/Interspersed_repeat
30. http://en.wikipedia.org/wiki/Palindromic_sequence
31. http://en.wikipedia.org/wiki/Trinucleotide_repeat_disordershttp://en.wikipedia.org/wiki/Suffix_tree
32. http://www.csd.uoc.gr/~hy463/2006/download/lectures/indexing_6.pdf

33. http://homepage.usask.ca/~ctl271/857/suffix_tree.shtml
34. Reducing the Space Requirement of Suffix Trees-STEFAN KURTZ
35. http://www.cs.ucdavis.edu/~martel/122a/suffix.pdf
36. Suffix Tree Construction-Arthur Dardia Suzanne Matthews
37. Practical Suffix Tree Construction- Sandeep Tata ,Richard A. Hankins ,Jignesh M. Patel
38. A rapid method for detection of putative RNAi target genes in genomic data
39. Yair Horesh, Amihood Amir, Shulamit Michaeli and Ron Unger