

ΑΝΩΤΑΤΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΊΔΡΥΜΑ ΠΕΙΡΑΙΑ Τ.Τ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΑΥΤΟΜΑΤΙΣΜΟΥ

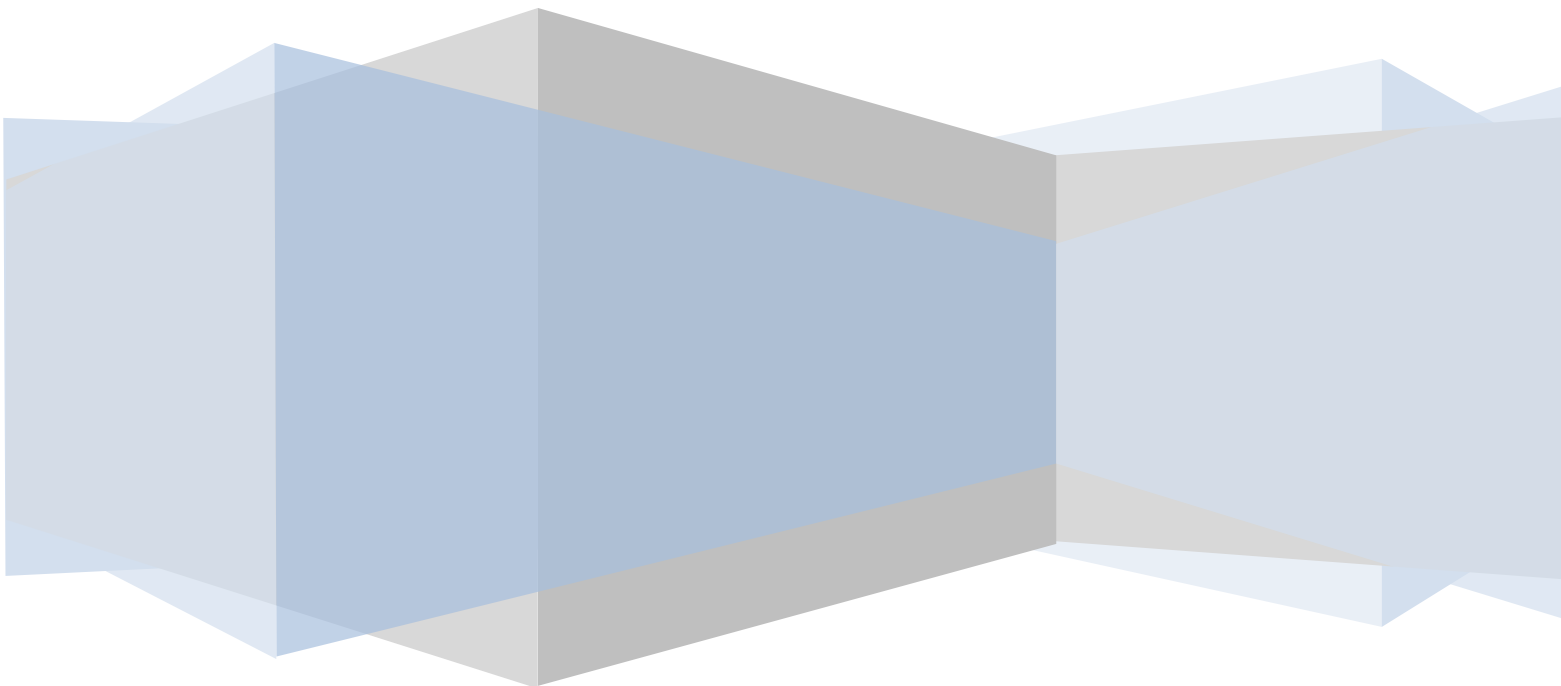


ΑΝΩΤΑΤΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΊΔΡΥΜΑ ΠΕΙΡΑΙΑ
ΤΕΧΝΟΛΟΓΙΚΟΥ ΤΟΜΕΑ

Πτυχιακή Εργασία: Αλγόριθμοι στη Μοριακή και Γενετική Βιολογία.

[Γουρνάρη Βασιλική]

Επιβλέπων καθηγητής: κ.Τσελές Δημήτριος



ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

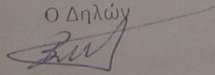
Ο / Η κάτωθι υπογεγραμμένος / η Γεωργίου Βασίλειος,
του Γεωργίου, με αριθμό μητρώου 44557 φοιτητής / τρια του
Τμήματος Μηχανικών Αυτοματισμού Τ.Ε. του Α.Ε.Ι. Πειραιά Τ.Τ. πριν αναλάβω την
εκπόνηση της Πτυχιακής Εργασίας μου, δηλώνω ότι ενημερώθηκα για τα παρακάτω:

«Η Πτυχιακή Εργασία (Π.Ε.) αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο του
συγγραφέα, όσο και του Ιδρύματος και θα πρέπει να έχει μοναδικό χαρακτήρα και
πρωτότυπο περιεχόμενο.

Απαγορεύεται αυστηρά οποιοδήποτε κομμάτι κειμένου της να εμφανίζεται
αυτούσιο ή μεταφρασμένο από κάποια άλλη δημοσιευμένη πηγή. Κάθε τέτοια πράξη
αποτελεί προϊόν λογοκλοπής και εγείρει θέμα Ηθικής Τάξης για τα πνευματικά δικαιώματα
του άλλου συγγραφέα. Αποκλειστικός υπεύθυνος είναι ο συγγραφέας της Π.Ε., ο οποίος
φέρει και την ευθύνη των συνεπειών, ποινικών και άλλων, αυτής της πράξης.

Πέραν των όποιων ποινικών ευθυνών του συγγραφέα σε περίπτωση που το Ίδρυμα
του έχει απονείμει Πτυχίο, αυτό ανακαλείται με απόφαση της Συνέλευσης του Τμήματος. Η
Συνέλευση του Τμήματος με νέα απόφαση της, μετά από αίτηση του ενδιαφερόμενου, του
αναθέτει εκ νέου την εκπόνηση της Π.Ε. με άλλο θέμα και διαφορετικό επιβλέποντα
καθηγητή. Η εκπόνηση της εν λόγω Π.Ε. πρέπει να ολοκληρωθεί εντός τουλάχιστον ενός
ημερολογιακού βμήνου από την ημερομηνία ανάθεσης της. Κατά τα λοιπά εφαρμόζονται τα
προβλεπόμενα στο άρθρο 18, παρ. 5 του ισχύοντος Εσωτερικού Κανονισμού.»

Ο Δηλών



Ημερομηνία

11/12/2016

Πίνακας περιεχομένων

Περίληψη.....	5
Κεφάλαιο 1	6
Βασικές έννοιες της Μοριακής Βιολογίας	6
1.1 Τι είναι η Μοριακή Βιολογία;	6
1.2 Πρωτεΐνες	6
1.3 Δομή των πρωτεϊνών	7
1.4 Ποία η σημασία των αμινοξέων για τον ανθρώπινο οργανισμό;	8
1.5 Πεπτιδικός δεσμός και οργάνωση των πρωτεϊνικών μορίων.	9
1.6 DNA(Δεοξυριβονουκλεϊκό οξύ)	9
1.7 Δομή DNA.....	10
1.8 Ριβονουκλεϊκό οξύ(RNA)	10
1.9 Χρωμοσώματα και γονίδιο;.....	11
1.10 Γενετικός κώδικας και βασικά χαρακτηριστικά.....	12
Λέξεις Κλειδιά.....	12
Κεφάλαιο 2ο.....	12
Μοριακή-Υπολογιστική Βιολογία.....	12
2.1Βασικοί ορισμοί μοριακής-υπολογιστικής βιολογίας.	13
2.2 Αλφάβητο	13
2.3 Ακολουθία	13
2.4 Αντιστοίχιση ακολουθίας(Sequence Alignment)	14
2.5 Βαθμολόγηση αντιστοίχισης(Score alignment).	14
2.6 Πίνακες βαθμολόγησης(Scoring Matrix)	15
2.7 Ποιος ο ρόλος της αντιστοίχισης-σύγκρισης των ακολουθιών;	18
Κεφάλαιο 3	19
Αλγόριθμοι Μοριακής Βιολογίας.....	19
3.1 Γενικός ορισμός αλγορίθμου.....	19
3.2 Είδη μοριακών αλγορίθμων.	19
3.2.1 Εισαγωγή στους αλγορίθμους δυναμικού προγραμματισμού, BLAST και FASTA..	19
3.3 Δυναμικός προγραμματισμός (Dynamic programming).....	19
3.3.1Κατηγορίες δυναμικού προγραμματισμού.	20

3.3.2 Πότε εφαρμόζεται ο δυναμικός προγραμματισμός;	20
3.3.3 Βήματα κατασκευής δυναμικού προγραμματισμού.	20
3.4 Αλγόριθμος BLAST(Basic Logical Alignment and Search Tool)	24
3.4.1 Εισαγωγή	24
3.4.2 Τρόπος εκτέλεσης του αλγορίθμου.	24
3.5 Αλγόριθμος Fasta	24
3.5.1 Εισαγωγή	24
3.5.2 Δομή και λειτουργία αλγορίθμου.	25
3.6 Αλγόριθμος Smith-Waterman.	25
Κεφάλαιο 4	26
Προγράμματα και γλώσσες προγραμματισμού μοριακών αλγορίθμων.	26
4.1 Γλώσσες προγραμματισμού.	26
4.2 Προγράμματα Μοριακής Βιολογίας.	26
4.2.1 Πρόγραμμα BLAST	26
4.2.2 Πρόγραμμα CLUSTALW	35
Κεφάλαιο 5	41
Γενετικοί Αλγόριθμοι	41
5.1 Εισαγωγή στους Γενετικούς Αλγορίθμους	41
5.2 Χαρακτηριστικά Γενετικού αλγορίθμου	41
5.3 Λειτουργίες βασικού Γενετικού αλγορίθμου.	41
5.4 Θεμελιώσεις των Γενετικών Αλγορίθμων.	48
5.5 Ασκήσεις Κατανόησης γενετικού αλγορίθμου.	55
Άσκηση 1	55
Άσκηση 2	57
Άσκηση 3	59
5.5 Πλεονεκτήματα των Γενετικών αλγορίθμων.	63
5.6 Προσδοκώμενα αποτελέσματα.	63
Βιβλιογραφία:	64

Περίληψη

Η εξέλιξη της επιστήμης σε πολλούς τομείς έχει προσφέρει στην ανθρωπότητα ένα μεγάλο πλήθος λύσεων που μέχρι τώρα φανταζόταν αδύνατον να επινοηθούν. Συγκεκριμένα η επιστήμη της Μοριακής και Γενετικής βιολογίας εξελίσσεται με ραγδαίους ρυθμούς στην προσπάθειά της να επιλύσει προβλήματα που θεωρούνται πολύπλοκα όχι μόνο για την υγεία του ανθρώπου αλλά και για κάθε ζωντανό οργανισμό. Τη λύση στα προβλήματα αυτά την δίνουν οι αλγόριθμοι που χρησιμοποιεί η επιστήμη της Μοριακής και Γενετικής Βιολογίας. Αρχικά στην παρούσα εργασία αναφέρουμε κάποιες γενικές έννοιες της βιολογίας που θα βοηθήσουν τον αναγνώστη στην κατανόηση της εργασίας. Στο δεύτερο κεφάλαιο θα ασχοληθούμε με τις έννοιες της υπολογιστικής βιολογίας οι οποίες αναφέρονται κυρίως στις πρωτεΐνες και στο γενετικό υλικό. Το τρίτο κεφάλαιο αποτελεί το κυρίως θέμα της εργασίας καθώς γίνεται αναφορά σε κάποιους από τους αλγόριθμους που χρησιμοποιεί η Μοριακή βιολογία με ιδιαίτερη έμφαση στον αλγόριθμο δυναμικού προγραμματισμού. Τέλος μελετάμε τους αλγόριθμους της Γενετικής Βιολογίας , γίνεται πλήρης αναφορά των βημάτων που χρησιμοποιεί ένας Γενετικός αλγόριθμος επιλύοντας κάποιες ασκήσεις που θα βοηθήσουν στην κατανόησή του. Στο σημείο αυτό θα ήθελα να αναφέρω ότι οι εν λόγω αλγόριθμοι αναφέρονται στα γενετικά χαρακτηριστικά ενός ατόμου στις πιθανότητες διασταύρωσής του και μετάλλαξής του. Στα πλαίσια εκπόνησης αυτής της πτυχιακής χρησιμοποιήσαμε πληροφορίες και στοιχεία από πολλές πηγές οι οποίες αναφέρονται και στο τέλος της εργασίας στην βιβλιογραφία. Όμως ιδιαίτερη χρήση των πληροφοριών έγινε κυρίως στο πρώτο και δεύτερο κεφάλαιο αντλώντας πληροφορίες από την διπλωματική εργασία του κ. Πέτρου Γαβρά (Αλγόριθμοι αναζήτησης σε βάσεις βιολογικών δεδομένων και υλοποίηση σε java) και της κ. Κωνσταντίνας Δημητρακοπούλου (Αναγνώριση λειτουργικών υποδομών στο πρωτεϊνικό δίκτυο του *Saccharomyces cerevisiae* συνδυάζοντας δεδομένα έκφρασης γονιδίων και αλληλεπίδρασης πρωτεϊνών) ενώ για το τρίτο, τέταρτο και πέμπτο κεφάλαιο χρησιμοποιήσαμε στοιχεία από τις διαλέξεις του κ. Δημητρίου Ψούνη στα πλαίσια της ενότητας (ΠΛ30,ΠΛ31).

Κεφάλαιο 1

Βασικές έννοιες της Μοριακής Βιολογίας

Στο κεφάλαιο αυτό αναλύουμε κάποιους βασικούς ορισμούς της μοριακής βιολογίας. Η έλλειψη των απαραίτητων γνώσεων σε θέματα της βιολογίας μας οδήγησε στη δημιουργία ενός μικρού βοηθήματος, έτσι ώστε ο αναγνώστης να καταλάβει το σκοπό και τη μελέτη της παρούσας εργασίας. Παρακάτω παρουσιάζουμε συνοπτικά τις γενικότερες έννοιες διευκρινίζοντας μόνο τα στοιχεία και τους ορισμούς που θα χρησιμοποιήσουμε για τη μελέτη της εργασίας.

1.1 Τι είναι η Μοριακή Βιολογία;

Μοριακή Βιολογία είναι ο κλάδος της βιολογίας που ασχολείται με την μελέτη των μορίων που συμμετέχουν στο θαύμα της ζωής. Πιο συγκεκριμένα, ασχολείται με την μελέτη της σύνθεσης, της δομής και της λειτουργίας της γενετικής πληροφορίας(DNA), και ειδικότερα με την αντιγραφή, μεταγραφή και μετάφραση του (RNA). Αντικείμενο της μοριακής βιολογίας είναι η μελέτη και η εξέλιξη της αλληλεπίδρασης της γενετικής πληροφορίας με τις πρωτεΐνες και των συστημάτων του κυττάρου.

1.2 Πρωτεΐνες

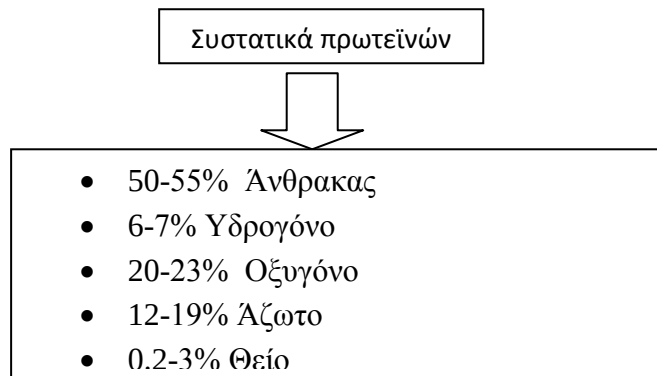
Ποιός ο ρόλος των πρωτεϊνών;

Είναι ουσίες που απαρτίζουν ένα σημαντικό μέρος του συνόλου των υλικών στοιχείων(όργανα, σκελετός, ιστοί κτλ.) κάθε ζωντανού οργανισμού. Αποτελούνται από στοιχεία αζώτου άνθρακα υδρογόνου και οξυγόνο. Το άζωτο είναι το στοιχείο που τις κάνει να διαφέρουν από τα υπόλοιπα συστατικά γι αυτό αναφέρονται ως αζωτούχες οργανικές ενώσεις. Συμβάλλουν σημαντικά στις διάφορες λειτουργίες κάθε ζωντανού οργανισμού όπως στην ανάπτυξη, εξέλιξη και την διατήρηση. Ανάλογα με τις λειτουργίες τους κατατάσσονται σε τρεις βασικές κατηγορίες : πρωτεΐνες με βάση τη δομική τους σύσταση, πρωτεΐνες με βάση τη βιολογική τους λειτουργία και πρωτεΐνες τροφίμων.

- Με κριτήριο τη δομική σύσταση των πρωτεϊνών αναφερόμαστε στις ινώδεις πρωτεΐνες οι οποίες περιέχουν ουσίες όπως το κολλαγόνο και την ελαστίνη που βρίσκονται στα οστά, στους μύς, στο δέρμα και στα όργανα κάθε ζωντανού οργανισμού.
- Με κριτήριο τη βιολογική τους δράση οι πρωτεΐνες αποτελούνται από τις ορμόνες (πχ η ινσουλίνη) που είναι υπεύθυνες για τη μεταφορά πληροφοριών και τη σωστή λειτουργία του οργανισμού. Τα χιλιάδες είδη ενζύμων γνωστοί και ως βιολογικοί καταλύτες και λειτουργούν με εξαιρετικά γρήγορες διαδικασίες στις διάφορες χημικές αντιδράσεις των κυττάρων. Επίσης οι μεταφορικές πρωτεΐνες που έχουν ως κύρια λειτουργία τη μεταφορά κάποιων ουσιών εσωτερικά του οργανισμού (αιμοσφαιρίνη, μυοσφαιρίνη, τρανσφερίνη), οι αποθηκευτικές ή αποταμιευτικές πρωτεΐνες εξασφαλίζοντας ειδικό χώρο για μελλοντική χρήση και οι προστατευτικές πρωτεΐνες (ανοσοσφαιρίνες) που αναγνωρίζουν άγνωστες πρωτεΐνες (ιούς, βακτήρια) και τις καταστρέφουν [1].

- Πρωτεΐνες τροφίμων ή διατροφικές πρωτεΐνες αν και δεν κατατάσσονται ως κάποια ξεχωριστή κατηγορία δεδομένου ότι και οι παραπάνω κατηγορίες ανήκουν στα είδη των διατροφικών πρωτεϊνών είναι σημαντικό να αναφέρουμε ότι είναι υπεύθυνες για τη λειτουργία του πεπτικού συστήματος του οργανισμού και που δεν έχουν τοξική επίδραση σε αυτόν και είναι διαθέσιμες σε μεγάλες ποσότητες .

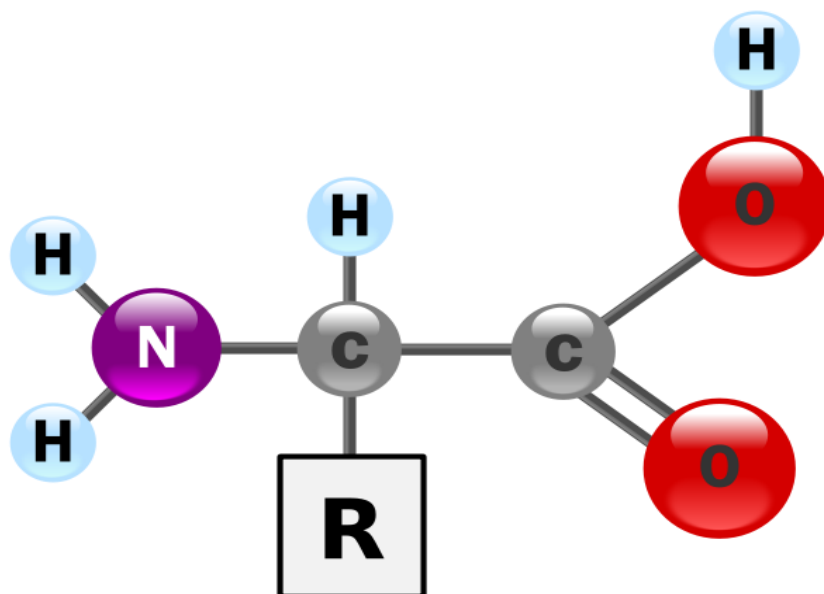
Η πολύτιμη σημασία των πρωτεϊνών διαπιστώνεται και από το γεγονός ότι στο γενετικό υλικό(DNA) κάθε οργανισμού αποταμιεύονται δεδομένα που αφορούν μόνο τη δομική σύσταση των πρωτεϊνών. Τα υπόλοιπα συστατικά και οι χημικές αντιδράσεις των κυττάρων έπονται των πρωτεϊνών και είναι εξαρτώμενα από αυτές.



Σχήμα 1(συστατικά πρωτεϊνών)

1.3 Δομή των πρωτεϊνών

Σημαντικό ρόλο στη δομή των πρωτεϊνών έχουν τα αμινοξέα. Τα αμινοξέα είναι απλές χημικές ενώσεις αποτελούμενες από ένα τουλάχιστον μονοκαρβονικό οξύ (RCOOH) και από μία τουλάχιστον ομάδα αζωτούχων οργανικών ενώσεων δηλαδή μιας αμινομάδας (-NH₂). Η ραγδαία εξέλιξη της χημείας έχει ως αποτέλεσμα την γνωστοποίηση πολλών αμινοξέων στους οργανισμούς αλλά με τη μόνη διαφορά ότι στους ζωντανούς οργανισμούς υπάρχουν μόνο 20 αμινοξέα τα οποία διαφέρουν μεταξύ τους. Αυτό που κάνει τα αμινοξέα να ξεχωρίζουν είναι η πλευρική αλυσίδα [1].



Εικόνα 1 Γενικός τύπος αμινοξέων[2]

Στο παραπάνω σχήμα το NH_2 είναι η αμινοομάδα, C άνθρακας, COOH καρβοξυλικό οξύ και το R αποτελεί την πλευρική αλυσίδα. Αν πχ το R είναι ένα άτομο υδρογόνου τότε το αμινοξύ θα ονομάζεται γλυκίνη, που θεωρείται ένα από τα αμινοξέα με την πιο απλή μορφή [3].

1.4 Ποία η σημασία των αμινοξέων για τον ανθρώπινο οργανισμό;

Ο ανθρώπινος οργανισμός απαιτεί την ύπαρξη των αμινοξέων για τη σύσταση των διάφορων πρωτεϊνών και πεπτιδίων. Όλα αυτά τα αμινοξέα τα προμηθεύεται ο οργανισμός από τις βρώσιμες πρωτεΐνες (αυτές που υπάρχουν στην τροφή) και με ειδική επεξεργασία της υδρόλυσης μας δίνουν τα αμινοξέα, καθώς και με τη σύνθεση άλλων πρώτων υλών. Τα αμινοξέα όπου ο οργανισμός έχει την ικανότητα να τα συνθέσει από μόνος του ονομάζονται μη απαραίτητα.

<u>Ονομασία αμινοξέος</u>	<u>Σύμβολο</u>	<u>Σύντμηση</u>
Αλανίνη	A	Ala
Αργινίνη	R	Arg
Ασπαργίνη	N	Asn
Ασπαρτικό οξύ	D	Asp
Βαλίνη	V	Val
Γλουταμίνη	Q	Gln
Γλουταμινικό οξύ	E	Glu
Γλυκίνη	G	Gly
Ιστιδίνη	H	His
Ισολευκίνη	I	Ile
Λευκίνη	L	Leu
Λυσίνη	K	Lys
Μεθειοθονίνη	M	Met
Προλίνη	P	Pro
Φαινυλανίνη	F	Phe
Σερίνη	S	Ser

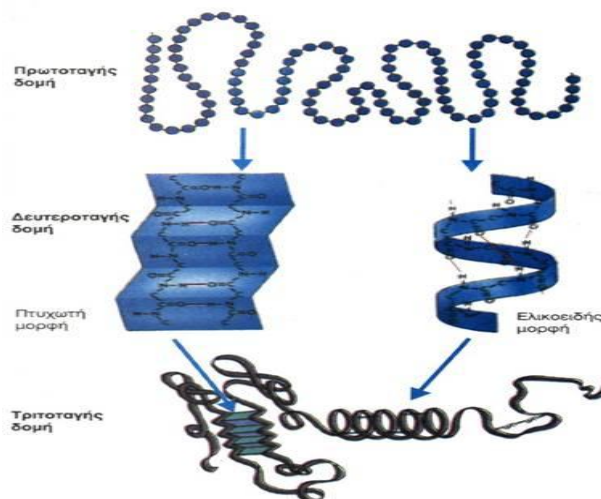
Θρεονίνη	T	Thr
Τρυπτοφάνη	W	Trp
Τυροσίνη	Y	Tyr
Κυστεΐνη	C	Cys

Πίνακας 1(Είκοσι βασικών αμινοξέων)

1.5 Πεπτιδικός δεσμός και οργάνωση των πρωτεϊνικών μορίων.

Πεπτιδικός δεσμός καλείται η ένωση δύο αμινοξέων συνήθως με ομοιοπολικούς δεσμούς. Ο πεπτιδικός δεσμός είναι πιο διαδεδομένος στις πρωτεΐνες ή αλλιώς πολυπεπτιδικές αλυσίδες όπου τα αμινοξέα ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς. Δηλαδή η τοποθέτηση και η ένωση όλων των αμινοξέων σε σειρά δημιουργώντας μια πρωτεϊνική αλυσίδα. Για να επιτελέσουν οι πρωτεΐνες μια διεργασία πρέπει να υποστούν κάποια επεξεργασία και αφού πάρουν την τελική τους μορφή τότε οργανώνονται σε τέσσερα επίπεδα.[4]

1. Πρωτοταγής δομή όπου τα αμινοξέα ενώνονται μεταξύ τους με ένα καθορισμένο τρόπο.
2. Δευτεροταγής δομή δημιουργείται όταν η πολυπεπτιδική αλυσίδα αναδιπλώνεται και αποκτά την εικόνα μιας ελικοειδούς μορφής.
3. Τριτοταγής δομή η μορφή που αποκτά στο χώρο η πολυπεπτιδική αλυσίδα. Η μορφή αυτή προκύπτει από την αναδίπλωση των ελικοειδών περιοχών και παραμένει αμετάβλητη λόγω των δεσμών που δημιουργούνται μεταξύ των διπλανών ομάδων των αμινοξέων.
4. Τεταρτοταγής δομή όπου διαφέρει από τις προηγούμενες δομές επειδή δημιουργείται σε πρωτεΐνες που αποτελούνται από περισσότερες από μια πολυπεπτιδικές αλυσίδες ενώ οι υπόλοιπες εμφανίζονται σε όλες τις



πρωτεΐνες. .

Εικόνα 2[5] Οργάνωση των πρωτεϊνών σε επίπεδα.

1.6 DNA(Δεοξυριβονουκλεϊκό οξύ)

Το DNA είναι το βασικό δομικό στοιχείο όλων των κυττάρων το οποίο περιέχει πληροφορίες για την ανάπτυξη και εξέλιξη όλων των οργανισμών υπό μορφών ζωής και την πλειονότητα των ιών. Βασικός ρόλος του είναι η αποθήκευση όλων των πληροφοριών με σκοπό τη σύνθεση όλων των πρωτεϊνών των οργανισμών υπό μορφή ζωής. Θεωρείται φορέας της

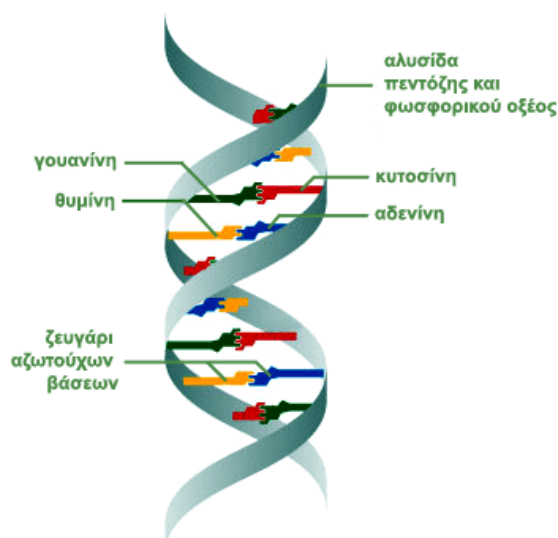
αναλλοίωτης γενετικής πληροφορίας του κυττάρου, μέσω των κυτταρικών διαιρέσεων με αποτέλεσμα τη δημιουργία αντιγράφων που απαιτούνται.

1.7 Δομή DNA

Η δομή του DNA ανακαλύφθηκε το 1953 από τους Τζέιμς Γουάτσον(*James D.Watson*) και Φράνσις Κρικ (*Francis Crick*) και θεωρείται ίσως και η μεγαλύτερη και πιο σπουδαία ανακάλυψη στο τομέα της βιολογίας του 20ού αιώνα. Το σύνολο των μορίων που περιέχει ένα κύτταρο αποτελεί το γενετικό υλικό του. Το DNA αποτελείται από ένα μόνο μόριο δημιουργεί μια αλυσίδα από απλούστερα μόρια όπως συμβαίνει και στις πρωτεΐνες. Το DNA αποτελείται από τέσσερις αζωτούχες βάσεις (θυμίνη, γουανίνη, κυτοσίνη και αδερίνη). Συνδέονται μεταξύ τους με ένα μοναδικό τρόπο γι αυτό και ονομάζονται συμπληρωματικές. Δηλαδή η αδερίνη είναι συμπληρωματική της θυμίνης δηλαδή συνδέεται μόνο με αυτή και η γουανίνη συμπληρωματική της κυτοσίνης. Οι αζωτούχες βάσεις που βρίσκονται στη μία αλυσίδα θα συνδέονται με δεσμούς υδρογόνου με αυτές που είναι στην απέναντι αλυσίδα . Υπάρχουν στο εσωτερικό του μορίου και είναι υδρόφοβες.(αδιάλυτες στο νερό). Η σύνδεση των αζωτούχων βάσεων αναγκάζουν το DNA να παίρνει μια ελικοειδή μορφή όπως φαίνεται στο σχήμα 1.4[6]

ΒΑΣΗ	ΣΥΜΒΟΛΟ
ΑΔΕΝΙΝΗ	A
ΘΥΜΙΝΗ	T
ΓΟΥΑΝΙΝΗ	G
ΚΥΤΟΣΙΝΗ	C

Πίνακας 2 Είδη βάσεων DNA και τα σύμβολά τους.



Εικόνα 3[6] Ελικοειδής μορφή DNA κατά τη σύνδεση αζωτούχων βάσεων.

1.8 Ριβονουκλεϊκό οξύ(RNA)

Το RNA ή αλλιώς το ριβοζονουκλεϊκό οξύ είναι μια κατηγορία νουκλεϊκού οξέος που μαζί με το DNA αποτελεί το γενετικό υλικό όλων των οργανισμών. Βρίσκεται σε όλα τα ευκαρυωτικά κύτταρα και στο κυτταρόπλασμα των βακτηριακών κυττάρων. Σε ένα σημαντικό βαθμό το RNA και το DNA παρουσιάζουν αρκετές ομοιότητες . Ωστόσο, υπάρχουν τέσσερις βασικές διαφορές που είναι οι εξής:

1. Για το σχηματισμό των μορίων του RNA χρησιμοποιείται το σάκχαρο ριβόζη ενώ στο DNA χρησιμοποιείται η 2'- δεοξυριβόζη.
2. Σε αντίθεση με το DNA σχηματίζουν έναν απλό μηχανισμό και όχι διπλή έλικα με αποτέλεσμα να έχουν πολύπλοκη τρισδιάστατη δομή
3. Βασική διαφορά είναι ότι δεν υπάρχει η αζωτούχα βάση θυμίνη(T) αλλά έχουμε στη θέση της την ουρακίλη(U) η οποία συνδέεται και αυτή με την αδενίνη .
4. Τα μόρια του DNA έχουν ως βασική λειτουργία την κωδικοποίηση των γενετικών πληροφοριών, ενώ το RNA πραγματοποιεί και άλλες λειτουργίες.

Τα μόρια του RNA χωρίζονται σύμφωνα με την λειτουργία που εκτελούν σε τρία είδη[7]:

1. tRNA (μεταφορικό RNA) είναι μια μικρού μήκους αλυσίδα RNA που μεταφέρει αμινοξέα στα ριβοσώματα. Κάθε τύπος του tRNA αντιστοιχεί σε ένα μόνο αμινοξύ.
2. rRNA (ριβισωμικό RNA) κύρια λειτουργία του, είναι η σύνθεση του ριβοσώματος μέσω κάποιων ειδικών πρωτεϊνών που υπάρχουν στο κυτταρόπλασμα.
3. mRNA (αγγελιοφόρο RNA) αντιγράφει και μεταφέρει τη γενετική πληροφορία από το DNA στα ριβοσώματα.

1.9 Χρωμοσώματα και γονίδιο;

Στους οργανισμούς κυρίως των ευκαρυωτικών κυττάρων το DNA βρίσκεται στον πυρήνα το οποίο αποτελείται από μεγάλα μόρια DNA αλλά λίγα σε αριθμό δημιουργώντας κάποιες διακριτές δομές που ονομάζονται **χρωμοσώματα**[8]. Σε κάποια στάδια ανάπτυξης του κυττάρου τα χρωμοσώματα μπορούμε να τα διακρίνουμε με οπτικό μικροσκόπιο. Υπάρχει συγκεκριμένος αριθμός χρωμοσωμάτων για κάθε οργανισμό. Για παράδειγμα κάθε σωματικό κύτταρο του ανθρώπου αποτελείται από 46 χρωμοσώματα τα οποία σχηματίζουν 23 ομόλογα ζεύγη είναι δηλαδή ανά δυο όμοια ή αλληλόμορφα. Ωστόσο μπορεί να διαφέρουν ως προς την πληροφορία που περιέχουν έστω και αν αποτελούνται από όμοια γονίδια που βρίσκονται στις ίδιες θέσεις. Τα χρωμοσώματα μπορούμε να τα διακρίνουμε σε φυλετικά και αυτοσωμικά. Φυλετικά είναι αυτά που καθορίζουν το φύλο ενός οργανισμού όπως στον άνθρωπο ενώ αυτοσωμικά είναι αυτά που δεν σχετίζονται με το φύλο. Η γυναίκα περιέχει δύο X χρωμοσώματα και συμβολίζεται (XX), ο άντρας περιέχει ένα X χρωμόσωμα και ένα Y και συμβολίζεται (XY). Κάθε παιδί αποκτά το 50% των γονιδίων και χρωμοσωμάτων δηλαδή τα κληρονομικά χαρακτηριστικά από τον πατέρα του και 50% από την μητέρα του. Κληρονομεί ένα χρωμόσωμα από κάθε ζεύγος αλληλόμορφων χρωμοσωμάτων από τον πατέρα του και ένα από την μητέρα του. Έστω ότι πάρει το χρωμόσωμα Y από τον πατέρα του (XY) τότε το παιδί θα είναι αγόρι αν πάρει από την μητέρα του (XX) το X χρωμόσωμα τότε θα είναι κορίτσι. Από την μητέρα του κληρονομεί πάντα ένα X χρωμόσωμα.

Τα τμήματα του DNA που περιέχουν πληροφορίες για την σύνθεση των πρωτεϊνών ονομάζονται **γονίδια**[8]. Υπάρχουν πρωτεΐνες που καθορίζουν τη λειτουργία των κυττάρων. Αυτές οι πρωτεΐνες των γονιδίων όταν δεν λειτουργούν σωστά έχουν ως αποτέλεσμα την ταχύτερη αύξηση του αριθμού των κυττάρων , με σκοπό την μετάλλαξη των γονιδίων υπό μορφή όγκων που οδηγεί στην εμφάνιση της αντίστοιχης μορφής της νόσου.

1.10 Γενετικός κώδικας και βασικά χαρακτηριστικά.

Ο γενετικός κώδικας είναι ο κώδικας αναλογίας των αζωτούχων βάσεων με τα αμινοξέα των πρωτεϊνών. Γνωρίζουμε ότι για τη δημιουργία των πρωτεϊνών χρειαζόμαστε είκοσι διαφορετικά αμινοξέα. Στο RNA υπάρχουν τέσσερις αζωτούχες βάσεις(A,U,G,C). Για να κωδικοποιηθούν αυτά τα 20 διαφορετικά αμινοξέα θα πρέπει οι 4 αζωτούχες βάσεις να συνδυάζονται ανά τρεις επειδή $4^3=64$ περισσότερα δηλαδή από 20 αμινοξέα. Άρα για να κωδικοποιήσουμε ένα αμινοξύ χρειαζόμαστε μια τριάδα νουκλεϊτιδίων δηλαδή ένα κωδικόνιο.

Βασικά χαρακτηριστικά γενετικού κώδικα.

- Υπάρχει στους περισσότερους οργανισμούς
- Είναι διαρκής , μπορεί να διαβιβάζει με συγκεκριμένη κατεύθυνση (5'→3')το mRNA με τη παρουσία όλων των νουκλεοτιδίων .
- Είναι μη επικαλυπτόμενος δηλαδή έχει ως σημείο έναρξης της μετάφρασης καθορισμένο σημείο(5'-AUG-3') συνεχίζοντας τρία νουκλεοτίδια κάθε φορά. Έτσι δεν υπάρχει περίπτωση ένα νουκλεοτίδιο να διαβιβαστεί δυο φορές.
- Για να ξεκινήσει τη μετάφραση έχει ένα κωδικόνιο έναρξης (AUG) και τρία κωδικόνια λήξης(UAA,UAG,UGA) υπεύθυνα για τον τερματισμό της όπου δεν κωδικοποιούν κανένα αμινοξύ.

Λέξεις Κλειδιά

- Πρωτεΐνη: κύριο στοιχείο που καθορίζει την ανάπτυξη, λειτουργία και δομή όλων των ζωντανών οργανισμών.
- Αμινοξέα: απαραίτητο συστατικό των πρωτεϊνών. Υπάρχουν είκοσι βασικά αμινοξέα που συμβάλλουν στην ανάπτυξη όλων των οργανισμών.
- Γονίδιο: κομμάτια DNA που περιέχουν πληροφορίες για τη λειτουργία των πρωτεϊνών.
- DNA: βασικό δομικό στοιχείο όλων των κυττάρων.
- RNA: χωρίζεται σε μεταφορικό(tRNA), ριβοσωμικό(rRNA) και αγγελιοφόρο (mRNA).
- χρωμόσωμα: μόρια DNA που καθορίζουν βασικά χαρακτηριστικά του ζωντανού οργανισμού όπως το φύλο του.
- γενετικός κώδικας: κώδικας αντιστοίχισης αζωτούχων βάσεων με τα αμινοξέα των πρωτεϊνών.
- κύτταρο: δομικό και λειτουργικό στοιχείο που καθορίζει την ύπαρξη ζωής.

Κεφάλαιο 2ο

Μοριακή-Υπολογιστική Βιολογία.

Στο κεφάλαιο αυτό θα μελετήσουμε βασικούς ορισμούς της μοριακής -υπολογιστικής βιολογίας όπου αποτελούν την βάση για την υλοποίηση πολλών αλγορίθμων και κυρίως των FASTA και BLAST που θα ασχοληθούμε στην ενότητα του 3ου κεφαλαίου.

2.1 Βασικοί ορισμοί μοριακής-υπολογιστικής βιολογίας.

Στην ενότητα αυτή αναφέρουμε κάποιους βασικούς ορισμούς που τους έχουμε δανειστεί από την εργασία του Γαβρά Πέτρου με θέμα "Αλγόριθμοι αναζήτησης σε βάσεις βιολογικών δεδομένων και υλοποίηση σε Java."

2.2 Αλφάβητο

Ορισμός: Αλφάβητο καλείται ένα πεπερασμένο σύνολο από σύμβολα.

Για τους αλγορίθμους που θα ασχοληθούμε στα παρακάτω κεφάλαια είναι απαραίτητο η δημιουργία δύο διαφορετικών αλφαβήτων αποτελούμενα από σύμβολα των πινάκων (1.1 και 1.2), ένα αλφάβητο δηλαδή για τις πρωτεΐνες και ένα για το DNA. Ως αποτέλεσμα έχουμε τη δημιουργία των παρακάτω αλφαβήτων.

$$\Sigma_{\text{DNA}} = \{A, T, G, C\}$$

και

$$\Sigma_{\text{PROTEIN}} = \{A, R, D, N, V, Q, E, G, H, I, L, K, M, P, F, S, T, W, Y, C\}$$

Υπάρχουν επίσης εκτός από τα κύρια σύμβολα και άλλα σύνολα από σύμβολα τα οποία τα ονομάζουμε αμφίσημα (ambiguous symbols). Αυτά τα σύμβολα τα χρησιμοποιούμε όταν δεν γνωρίζουμε το συμβολισμό και την σειρά του συμβόλου.

2.3 Ακολουθία

Ορισμός: Ως ακολουθία καλείται η καθορισμένη σειρά συμβόλων ενός αλφαβήτου.

Η ακολουθία θεωρείται από τις βασικότερες έννοιες της μοριακής βιολογίας καθώς αναπαριστά τα μόρια του DNA και των πρωτεϊνών. Το μήκος μιας ακολουθίας υπολογίζεται από το σύνολο των συμβόλων που την απαρτίζουν. Αν μια ακολουθία έχει μήκος s τότε συμβολίζεται με $|s|$. Αν το μήκος της ακολουθίας είναι ίσο με το μηδέν τότε η ακολουθία ονομάζεται κενή και συμβολίζεται με e . Το σύμβολο που υπάρχει στη θέση i μιας ακολουθίας συμβολίζεται με $s[i]$.

Παράδειγμα 1

Θεωρούμε την ακόλουθη συμβολοσειρά GATTACA ποιο το μήκος και η θέση της ακολουθίας;

Απάντηση

Το μήκος της ακολουθίας είναι 7 γιατί αποτελείται από 7 σύμβολα και η θέση της είναι: $s[1]=G$, $s[2]=A$, $s[3]=T$, $s[4]=T$, $s[5]=A$, $s[6]=C$, $s[7]=A$ και είναι η ακολουθία dna ή πρωτεΐνης.

Παράδειγμα 2

Τι μήκος και σε ποιο αλφάβητο ανήκει η παρακάτω συμβολοσειρά;

Απάντηση

Δεν μπορούμε να γνωρίζουμε το μήκος της ακολουθίας γιατί το σύμβολο U δεν υπάρχει σε κανένα από τα δύο αλφάβητα, συνεπώς δεν είναι έγκυρη ακολουθία.

2.4 Αντιστοίχιση ακολουθίας(Sequence Alignment)

Ορισμός: Θεωρούμε δυο ακολουθίες s_1 και s_2 , η αντιστοίχιση των ακολουθιών επιτυγχάνεται με την ομοιότητα των συμβόλων των ακολουθιών και την εισαγωγή κενών(gaps) σε τυχαία σημεία έτσι ώστε οι ακολουθίες να αποκτήσουν το ίδιο μήκος.

Τα κενά μπορούν να εισαχθούν σε όποιο σημείο της ακολουθίας εμείς επιλέξουμε, δηλαδή στην αρχή, μέση και τέλος των συμβόλων. Οι αντιστοιχίες χωρίζονται σε τρία είδη :

1. Ολική αντιστοίχιση(global alignment): Είναι η αντιστοίχιση ακολουθιών που έχουν μικρό συνήθως μήκος ακολουθίας και περισσότερα όμοια σύμβολα.
2. Τοπική αντιστοίχιση(local alignment): Οι ακολουθίες αυτής της αντιστοίχισης έχουν μεγάλο μήκος και όχι απαραίτητα ίσο επομένως είναι πιο δύσκολο να βρούμε το μήκος της ακολουθίας λόγω των διαφορετικών συμβόλων που έχουν.
3. Πολλαπλή αντιστοίχιση(Multiple sequence alignment): Εφαρμόζεται κυρίως μέσω υπολογιστών όταν έχουμε να αντιμετωπίσουμε πάρα πολλές ακολουθίες με διαφορετικά μήκη και σύμβολα. Ένα παράδειγμα χρήσης αυτής της αντιστοίχισης είναι το φυλογενετικό δέντρο(phylogenetic tree).

2.5 Βαθμολόγηση αντιστοίχισης(Score alignment).

Ορίζεται ως βαθμολόγηση αντιστοίχισης το άθροισμα των αντιστοιχίσεων των συμβόλων μαζί με την εισαγωγή των κενών που χρησιμοποιήσαμε. Η μέγιστη βαθμολόγηση της ακολουθίας ονομάζεται ομοιότητα(similarity). Η πλήρης ομοιότητα των δυο συμβόλων ονομάζεται (match) και η διαφορά(mismatch).

Παράδειγμα 3

Θεωρούμε τις εξής ακολουθίες $s_1 \rightarrow A T C G$ και $s_2 \rightarrow T C A$ ποια η βαθμολόγηση αυτών των ακολουθιών;

Απάντηση.

Όπως ήδη ξέρουμε για να βρούμε την βαθμολόγηση δύο ακολουθιών θα πρέπει να έχουν το ίδιο μήκος. Οι παραπάνω ακολουθίες δεν έχουν το ίδιο μήκος γι αυτό και θα προσπαθήσουμε να τις φέρουμε στο ίδιο μήκος για να βρούμε την βαθμολόγηση.

Αρχικά βρίσκουμε ποια σύμβολα είναι όμοια και ποια όχι.

$$1. \quad s_1 \rightarrow A T C G$$

$$s_2 \rightarrow \text{---} T C A$$

Παρατηρούμε ότι δεν υπάρχει σύμβολο κάτω από την αδερίνη(A) γι αυτό και εισάγουμε ένα κενό. Έχουμε δυο ομοιότητες της θυμίνης(T) και της κυτοσίνης(C) δηλαδή (match) και μια διαφορά συμβόλων μεταξύ γουανίνης(G) και αδερίνης(A)

δηλαδή (mismatch). Στη συνέχεια βρίσκουμε όλους τους πιθανούς συνδυασμούς των ακολουθιών όπως φαίνεται παρακάτω:

$$2. \quad s_1 \rightarrow A \ T \ C \ G$$

$$s_2 \rightarrow T \ C \ A \ -$$

Στο σημείο αυτό παρατηρούμε ότι δεν έχουμε κανένα όμοιο σύμβολο γεγονός που καθιστά τη βαθμολόγηση όχι και από τις ιδανικές λύσεις.

$$3. \quad s_1 \rightarrow A \ T \ C \ G$$

$$s_2 \rightarrow T \ - \ C \ A$$

Ενώ σ' αυτές τις ακολουθίες παρατηρούμε ότι έχουμε ένα τουλάχιστον κοινό σύμβολο.

Άρα στην πρώτη περίπτωση έχουμε: match +2, mismatch -1, gap-2 το οποίο παραμένει ίδιο και για τις άλλες περιπτώσεις το μόνο που αλλάζει είναι η βαθμολόγηση. Επομένως η βαθμολόγηση για την πρώτη περίπτωση είναι 1.

$$s_1 \rightarrow A \ T \ C \ G$$

$$s_2 \rightarrow - \ T \ C \ A$$

$$-2, +2, +2, -1 \rightarrow 1$$

Στη δεύτερη περίπτωση η βαθμολόγηση είναι -5.

$$s_1 \rightarrow A \ T \ C \ G$$

$$s_2 \rightarrow T \ C \ A \ -$$

$$-1 \ -1 \ -1 \ -2 \rightarrow -5$$

Και στην τρίτη περίπτωση η βαθμολόγηση είναι -2.

$$s_1 \rightarrow A \ T \ C \ G$$

$$s_2 \rightarrow T \ - \ C \ A$$

2.6 Πίνακες βαθμολόγησης(Scoring Matrix)

Ορισμός: Ως πίνακα βαθμολόγησης ορίζουμε έναν πίνακα N ο οποίος περιέχει m*m συμβόλων, όπου κάθε στοιχείο N[i,j] μας δίνει την βαθμολόγηση της αντιστοίχισης των συμβόλων.

Το είδος των πινάκων βαθμολόγησης θεωρείται μια από τις πιο βασικές παραμέτρους στην εφαρμογή προγραμμάτων τοπικής αντιστοίχισης. Είναι σημαντικό να αναφέρουμε ότι στην Μοριακή Βιολογία χρησιμοποιούμε τους πίνακες PAM και BLOSUM. Οι πίνακες PAM βασίζονται σε Μαρκοβιανές αλυσίδες δηλαδή τη χρήση ενός μαθηματικού συστήματος που έχει ως στόχο να μεταβάλλει μια κατάσταση σε μια άλλη προκαλώντας ένα είδος αμνησίας στην επόμενη κατάσταση. Οι πίνακες

BLOSUM σχεδιάστηκαν για να μπορούν να ανακαλύπτουν τα κύρια στοιχεία των πρωτεϊνών, μπορούν να βρουν την στοίχιση απομακρυσμένων ακολουθιών και τις τοπικές ομοιότητες. Κάθε πίνακας έχει σχεδιαστεί με τέτοιο τρόπο έτσι ώστε να δημιουργεί τις καλύτερες στοιχίσεις σε μια δεδομένη εξελικτική σχέση πρωτεϊνών. Για παράδειγμα, άλλος πίνακας θα ταιριάζει στην περίπτωση που έχουμε δύο πολύ ομόλογες πρωτεΐνες και άλλος για δύο πολύ απομακρυσμένες εξελικτικά πρωτεΐνες. Γι αυτό και είναι καλύτερο να χρησιμοποιούμε κυρίως τους πιο αποδεκτούς πίνακες όπως τον πίνακα BLOSUM62. Οι δύο πιο γνωστοί πίνακες υποκαταστάσεως είναι οι BLOSUM62 και PAM250. Η διαφορά που έχουν οι δύο πίνακες είναι στα μεγέθη Entropy(σχετική εντροπία) και Expected(αναμενόμενο score).[13]. Παρακάτω παραθέτουμε δύο από τους πιο αποδοτικούς πίνακες τον BLOSUM62 και τον PAM250.

BLOSUM62

#	BLOSUM Clustered Scoring Matrix in 1/2 Bit Units																								
#	Cluster Percentage: >= 62																								
#	Entropy = 0.6979, Expected = -0.5209																								
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*	
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4	
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4	
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4	
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4	
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4	
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4	
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4	
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4	
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4	
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4	
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4	
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4	
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4	
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4	
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4	
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4	
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4	
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4	
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4	
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4	
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4	
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1	

Πίνακας BLOSUM62[13]

PAM250

```
# This matrix was produced by "pam" Version 1.0.6 [28-Jul-93]
# PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049
# Expected score = -0.844, Entropy = 0.354 bits
# Lowest score = -8, Highest score = 17
```

	A	R	N	D	C	Q	F	G	H	T	I	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	7	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	1	2	4	5	2	3	1	1	2	4	0	3	6	1	0	0	7	4	2	3	3	1	8
C	2	4	4	5	12	5	5	3	3	2	6	5	5	4	3	0	2	8	0	2	4	5	3	8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
T	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
I	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	1	3	0	1	0	0	2	3	1	2	5	6	1	0	6	5	1	1	0	1	8
S	1	0	1	0	0	1	0	1	1	1	3	0	2	3	1	2	1	2	3	1	0	0	0	8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

Πίνακας PAM250[13]

Παράδειγμα 3

Υπολογίστε τον πίνακα βαθμολόγησης των παρακάτω ακολουθιών:

$$s_1 \rightarrow A T C G \text{ και } s_2 \rightarrow T C A$$

Δίνεται ότι το match=+1, mismatch=-1, gap=-2

Απάντηση

Για να βρούμε την τιμή του κάθε στοιχείου βρίσκεται από τρεις θέσεις..

- κάτω κελί(bottom box)
- διπλανό κελί(besides box)
- διαγώνιο κελί(diagonal box)

Η τιμή εξαρτάται από το match/mismatch για το διαγώνιο κελί και από το gap για τα (bottom box, besides box).

	#	A	T	C	G
#	0	-2	-4	-6	-8
T	-2	-1	-1	-3	-5
C	-4	-3	-2	0	-2
A	-6	-3	-4	-2	-1

Πίνακας 4 βαθμολόγησης

Θα χρησιμοποιήσουμε συγκεκριμένα βήματα για να βρούμε τις τιμές των κελιών.

Στην πρώτη στήλη και γραμμή συμπληρώνουμε τις τιμές των κενών. Για να βρούμε την τιμή στο κελί $a[3,3]$ παρατηρούμε αν υπάρχει ομοιότητα συμβόλων. Δεν υπάρχει άρα έχουμε mismatch. Προσέχουμε την τιμή δίπλα από το κελί $a[3,3]$ είναι -2 άρα θα προσθέσουμε σε αυτό το -2 που είναι η τιμή του gap. Επομένως η τιμή είναι -4. Συνεχίζουμε τώρα με την τιμή που βρίσκεται πάνω από το κελί $a[3,3]$ και είναι -2 και σε αυτήν θα προσθέσουμε την τιμή του gap που είναι το -2 και το αποτέλεσμα είναι -4. Για να βρούμε την τιμή που προέρχεται από το διαγώνιο κελί προσθέτουμε στο 0 το mismatch δηλαδή το -1 και μας δίνει το αποτέλεσμα -1. Το τελικό αποτέλεσμα του κελιού καθορίζεται από την μέγιστη τιμή των τριών αυτών τιμών (-4,-4,-1) που είναι το -1.

Στη συνέχεια θα δείξουμε πως υπολογίζουμε τις τιμές των κελιών όταν υπάρχει match. Ένα τέτοιο παράδειγμα υπάρχει στο κελί $a[3,4]$. Κοιτάζουμε την τιμή δίπλα από το κελί $a[3,4]$ και προσθέτουμε σε αυτό το -2 που είναι το gap άρα το αποτέλεσμα είναι -3. Το ίδιο και για την τιμή που βρίσκεται πάνω από το κελί $a[3,4]$ και έχουμε το αποτέλεσμα -6. Για να βρούμε την τιμή που προέρχεται από την διαγώνιο προσθέτουμε στην τιμή του διαγώνιου κελιού την τιμή +1 που είναι το match άρα έχουμε ως αποτέλεσμα το -1. Τελικό αποτέλεσμα είναι το -1 γιατί είναι το μέγιστο από τις τιμές (-3,-6,-1). Έτσι συνεχίζουμε και για τα υπόλοιπα κελιά.

Παρατηρούμε στον πίνακα ποια είναι η μέγιστη τιμή. Στην συγκεκριμένη περίπτωση είναι το 0. Άρα οι ακολουθίες που σχηματίζονται με βάση τον πίνακα είναι:

$$s_1 \rightarrow A T C$$

$$s_2 \rightarrow _ T C$$

Όταν η μέγιστη τιμή προέρχονται από την διαγώνιο τότε οι ακολουθίες θα σχηματίζονται από τα στοιχεία της διαγωνίου. Αν η μέγιστη τιμή είναι από το διπλανό κελί τότε παίρνουμε τα στοιχεία του κατακόρυφου άξονα. Ενώ αν είναι από το κάτω κελί τότε παίρνουμε τον οριζόντιο άξονα.

2.7 Ποιος ο ρόλος της αντιστοίχισης-σύγκρισης των ακολουθιών;

Η σύγκριση των ακολουθιών αναφέρεται στις ομοιότητες και διαφορές που υπάρχουν μεταξύ των συμβόλων της ακολουθίας. Ειδικά στην επιστήμη της βιολογίας η σύγκριση ακολουθίας έχει καθοριστική σημασία καθώς οι ακολουθίες που έχουν χαρακτηριστικές ομοιότητες ίσως αναφέρονται σε μόρια πρωτεΐνης ή Dna. Τα αποτελέσματα αυτών των ομοιοτήτων δίνουν στους μοριακούς βιολόγους

λεπτομέρειες σχετικά με τις λειτουργίες και άλλων ακολουθιών. Συνοψίζοντας καταλήγουμε στο συμπέρασμα ότι οι αλγόριθμοι εξαρτώνται από την σύγκριση αυτών των ακολουθιών.

Κεφάλαιο 3

Αλγόριθμοι Μοριακής Βιολογίας.

Στο κεφάλαιο αυτό αναφέρουμε αναλυτικά τον ορισμό και τα είδη του αλγορίθμου καθώς και βασικές έννοιες που χρησιμοποιούμε στην υπολογιστική-μοριακή βιολογία. Οι αλγόριθμοι αποτελούν σημαντικό κομμάτι της ζωής όχι μόνου του ανθρώπου αλλά και όλων ανεξαιρέτως των οργανισμών. Με άλλα λόγια όλη μας η ζωή αποτελείται από αλγορίθμους που εν μέρει κάποιοι γίνονται κατανοητοί και κάποιοι άλλοι όχι. Ακόμα και ο τρόπος όλων των λειτουργιών που εκτελούμε καθημερινά αποτελεί έναν αλγόριθμο. Έτσι και στη μοριακή βιολογία χρησιμοποιούνται συγκεκριμένα είδη που συμβάλλουν εξαιρετικά στην εξέλιξή της.

3.1 Γενικός ορισμός αλγορίθμου.

Αλγόριθμος καλείται μια σειρά εντολών-λειτουργιών που είναι απόλυτα καθορισμένες και εκτελέσιμες σε συγκεκριμένο χρόνο με σκοπό την επίλυση του προβλήματος. Κύριο χαρακτηριστικό του αλγορίθμου είναι ότι όλες οι ενέργειές του είναι πεπερασμένες έχουν δηλαδή αρχή και τέλος.

3.2 Είδη μοριακών αλγορίθμων.

Στη παρούσα εργασία θα ασχοληθούμε με την μελέτη τριών ειδών μοριακών αλγορίθμων: τον αλγόριθμο δυναμικού προγραμματισμού, τον αλγόριθμο FASTA και τον αλγόριθμο BLAST. Καθένας από τους τρεις χρησιμοποιείται για συγκεκριμένο σκοπό όπως περιγράφεται στο επόμενο κεφάλαιο. Περισσότερη αναφορά γίνεται στον αλγόριθμο δυναμικού προγραμματισμού αφού αποτελεί το θεμέλιο για την δημιουργία των άλλων δυο αλγορίθμων.

3.2.1 Εισαγωγή στους αλγορίθμους δυναμικού προγραμματισμού, BLAST και FASTA.

Η χρήση των παραπάνω αλγορίθμων έχει συμβάλλει σημαντικά στον τομέα της Μοριακής Βιολογίας. Αρχικά ο αλγόριθμος δυναμικού προγραμματισμού έχει ως κύριο χαρακτηριστικό το συνδυασμό επαναλαμβανόμενων λύσεων των υποπροβλημάτων. Ο αλγόριθμος BLAST είναι εύκολος στη χρήση και γρήγορος. Η βασική λειτουργία του είναι ο εύρεση ενός ζεύγους δύο ακολουθιών που εμφανίζουν ομοιότητα. Ο FASTA αλγόριθμος είναι και αυτός ένας ευριστικός αλγόριθμος που έχει ως στόχο την υπολογιστική προσέγγιση της σύγκρισης-αντιστοίχισης δυο ακολουθιών. Στις επόμενες ενότητες αναφέρουμε αναλυτικά τη χρήση και την λειτουργία για κάθε έναν από τους πιο πάνω αλγορίθμους.

3.3 Δυναμικός προγραμματισμός (Dynamic programming)

Ο δυναμικός προγραμματισμός είναι μια μέθοδος υπολογισμού προβλημάτων που δεν μπορούν να λυθούν με άλλες μεθόδους. Έχει ως βάση την αρχή της βελτιστοποίησης αναζητά δηλαδή την βέλτιστη λύση του προβλήματος. Για πρώτη φορά το 1953 χρησιμοποιήθηκε ο όρος του δυναμικού προγραμματισμού από τον μαθηματικό *Richard Ernest Bellman* (August 26, 1920-March 19, 1984)[9] με σκοπό

να περιγράψει τον τρόπο επίλυσης προβλημάτων που αναφέρεται σε σειρά διαδοχικών ζητημάτων. Αποτελεί μια τεχνική σχεδιασμού που επιλύει δύσκολα προβλήματα με τον πιο αποτελεσματικό τρόπο χρησιμοποιώντας ένα πρόγραμμα για την εφαρμογή του αλγορίθμου σε υπολογιστή.

3.3.1 Κατηγορίες δυναμικού προγραμματισμού.

Ο δυναμικός προγραμματισμός, σύμφωνα με την μέθοδο που υπολογίζονται οι τιμές-λύσεις ενός προβλήματος, χωρίζεται σε δυο κατηγορίες τον <<σύνθετο>> και τον <<αναλυτικό>>. Κύρια λειτουργία σύνθετου δυναμικού προγραμματισμού είναι ο υπολογισμός και η αποθήκευση των τιμών μιας συνάρτησης με την σειρά ξεκινώντας από την μικρότερη τιμή. Ο αναλυτικός δυναμικός προγραμματισμός είναι πιο αποτελεσματικός αφού χρησιμοποιεί έναν μηχανικό τρόπο που σχηματίζει και ρυθμίζει αυτόματα τη σειρά που υπολογίζονται τα υποπροβλήματα αποφεύγοντας κάποια από αυτά που δεν είναι απαραίτητα για λύση.

3.3.2 Πότε εφαρμόζεται ο δυναμικός προγραμματισμός;

Ο δυναμικός προγραμματισμός εφαρμόζεται όταν ένα πρόβλημα έχει τα ακόλουθα χαρακτηριστικά:

- ο Όταν πρέπει να λύσουμε το πρόβλημα και να πάρουμε την βέλτιστη λύση.
- ο Όταν αποτελείται από μικρό αριθμό υποπροβλημάτων.
- ο Επικαλυπτόμενα υποπροβλήματα δηλαδή οδηγούμαστε στην επαναλαμβανόμενη λύση των προβλημάτων με αποτέλεσμα να χάνουμε χρόνο.

Στον δυναμικό προγραμματισμό ξεκινάμε με διαφορετικό τρόπο να λύνουμε τα προβλήματα. Αρχίζουμε από το μικρότερο πρόβλημα και πηγαίνουμε στο μεγαλύτερο. Ξεκινάμε δηλαδή όπως θα δούμε και παρακάτω από κάτω προς τα πάνω.

3.3.3 Βήματα κατασκευής δυναμικού προγραμματισμού.

1. Διατυπώνουμε τον αλγόριθμο που λύνει το πρόβλημα.
2. Παρατηρούμε αν ισχύουν οι τρεις παραπάνω συνθήκες για την δημιουργία του αλγορίθμου δυναμικού προγραμματισμού.
3. Επιλύουμε τα υποπροβλήματα ξεκινώντας από τα μικρότερα προς τα μεγαλύτερα (δηλαδή από κάτω προς τα πάνω).
4. Ορίζουμε τον επαναληπτικό αλγόριθμο που δίνει την λύση του προβλήματος.
5. Βρίσκουμε την πολυπλοκότητα του αλγορίθμου.

Για να γίνει κατανοητή η χρήση του αλγορίθμου δυναμικού προγραμματισμού θα δώσουμε στη συνέχεια ένα παράδειγμα όπου θα αναλύσουμε όλα τα παραπάνω βήματα που αναφέραμε με κύριο σκοπό να δείξουμε στον αναγνώστη της παρούσας εργασίας τον τρόπο με τον οποίο σκεφτόμαστε για να λύσουμε ένα πρόβλημα με την χρήση του δυναμικού προγραμματισμού.

Παράδειγμα 4[10]

Έστω ένας πίνακας ακεραίων αριθμών A που αποτελείται από m γραμμές και n στήλες. Θεωρούμε ένα παιχνίδι όπου ξεκινώντας από οποιαδήποτε κελί της κάτω γραμμής του πίνακα, προσπαθούμε να φτάσουμε σε κάποιο κελί της πάνω γραμμής του πίνακα περνώντας από κελιά ελαχίστου συνολικού κόστους. Σαν κόστος $A(i,j)$, ενός κελιού (i,j) θεωρούμε τον αριθμό που αναγράφεται στο συγκεκριμένο κελί. Από

ένα κελί του πίνακα μπορούμε να κινηθούμε είτε προς το κελί που βρίσκεται ακριβώς από πάνω του, είτε διαγώνια αριστερά, είτε διαγώνια δεξιά.

Για παράδειγμα στον πίνακα του επόμενου σχήματος, μπορούμε να πάμε από το κελί (2,2), προς κάποιο από τα κελιά (3,1), (3,2) ή (3,3). Σαν πρώτη γραμμή θεωρείται η κάτω γραμμή του πίνακα (με μωβ αναφέρουμε τον αριθμό των γραμμών και με γαλάζιο των αριθμό των στηλών)

A=

4	3	9	9	6	9
3	5	6	7	3	4
2	6	8	7	7	2
1	4	3	6	5	9
	1	2	3	4	5

α)Ζητείται να περιγράψετε μια λύση του πιο πάνω προβλήματος με τη βοήθεια του δυναμικού προγραμματισμού.

β)Να δώσετε τον αντίστοιχο αλγόριθμο υπολογισμού του ελάχιστου κόστους μετάβασης από την πρώτη γραμμή του πίνακα προς την τελευταία.

γ)Ποια η πολυπλοκότητα χώρου του αλγορίθμου αυτού.

Απάντηση

α)Σύμφωνα με τα πιο πάνω βήματα η λύση του πρώτου ερωτήματος έχει ως εξής:

Αρχικά υπολογίζουμε τις βέλτιστες λύσεις του νέου πίνακα [OPT]. Όμως πάντα οι τιμές των κελιών της πρώτης γραμμής του πίνακα παραμένουν ίδιες με τις τιμές του αρχικού μας πίνακα. Για να βρούμε τις τιμές του κελιού (2,1) θα χρησιμοποιήσουμε τα κελιά που βρίσκονται κάτω, διαγώνια αριστερά και διαγώνια δεξιά από το κελί του πίνακα A προσθέτοντας στην τιμή 6 την μικρότερη τιμή του κελιού. Το κελί που έχει την μικρότερη τιμή είναι το διαγώνιο κελί δηλαδή το(1,2) άρα η βέλτιστη λύση είναι $6+3=9$. Το ίδιο κάνω και για το κελί (2,2). Κοιτάζω τις τιμές που βρίσκονται κάτω , διαγώνια αριστερά και διαγώνια δεξιά , βρίσκω ποια είναι η ελάχιστη τιμή και τη προσθέτω στην τιμή του κελιού. Άρα $8+3=11$. Ομοίως συνεχίζω και με τα υπόλοιπα κελιά. Έτσι έχουμε τον ακόλουθο πίνακα:

OPT=

4	17	23	19	16	19
3	14	15	17	10	11
2	9	11	10	12	7
1	4	3	6	5	9
	1	2	3	4	5

Η βέλτιστη λύση είναι αυτή που βρίσκεται στην τελευταία γραμμή του πίνακα και είναι η μικρότερη τιμή. Δηλαδή το 16. Συνεπώς έχουμε ήδη φτιάξει τον αλγόριθμο του δυναμικού προγραμματισμού χρησιμοποιώντας την παραπάνω διαδικασία αλλά πρέπει να τον διατυπώσουμε με μαθηματικά. Αυτό γίνεται γράφοντας την

αναδρομική σχέση που υπολογίζει την βέλτιστη λύση για κάθε κελί του πίνακα όπως φαίνεται παρακάτω.

Για την πρώτη γραμμή του πίνακα και για οποιοδήποτε κελί βρίσκεται στην πρώτη γραμμή βάλαμε την τιμή που βρίσκεται στον αρχικό πίνακα A. Αυτή η διαδικασία περιγράφεται μαθηματικά ως εξής:

1) $OPT[i,j]=\min\{A[i,j], i=1, 1\leq j\leq n\}$ όπου n ο αριθμός των στηλών.

Στη συνέχεια θα γράψουμε μια άλλη σχέση που να υπολογίζει την τιμή ενός κελιού που βρίσκεται σε μια τυχαία θέση. Έστω ότι θέλουμε να βρούμε πώς υπολογίσαμε την τιμή 17 του πίνακα OPT. Εδώ υπολογίσαμε πιο ήταν το ελάχιστο από τα κελιά που βρίσκονται ακριβώς κάτω από το 17 ακριβώς κάτω αριστερά και δεξιά και προσθέσαμε την ελάχιστη τιμή στην τιμή του κελιού του πίνακα A. Αυτό όμως δεν το κάναμε για όλα τα κελιά το κάναμε για οποιαδήποτε γραμμή εκτός της πρώτης, και για όλα τα κελιά που είναι από την δεύτερη στήλη έως και μία στήλη πριν από το τέλος. Η σχέση που περιγράφει την διαδικασία αυτή είναι η εξής:

2) $OPT[i,j]=\{\min\{OPT[i-1,j-1], OPT[i-1,j], OPT[i-1,j+1]\} + A[i,j], 2\leq i\leq m, 2\leq j\leq n-1\}$

Συνεχίζουμε περιγράφοντας πώς υπολογίσαμε τις τιμές του πρώτης στήλης. Χρησιμοποιήσαμε το κελί που βρίσκεται ακριβώς από κάτω και δεξιά πήραμε την ελάχιστη τιμή και προσθέσαμε σ' αυτή την τιμή του κελιού του πίνακα A. Αυτό ισχύει για τα κελιά που βρίσκονται από την δεύτερη γραμμή πάνω και μόνο στην πρώτη στήλη. Η σχέση 3 περιγράφει την διαδικασία που αναφέραμε προηγουμένως.

3) $OPT[i,j]=\{\min\{OPT[i-1,j], OPT[i-1,j+1]\} + A[i,j], 2\leq i\leq m, j=1\}$

Όμως πρέπει να αναφέρουμε πώς υπολογίσαμε τις τιμές της τελευταίας στήλης. Πήραμε το κελί που βρίσκεται ακριβώς από κάτω και αυτό που βρίσκεται κάτω αριστερά και επιλέξαμε από αυτά το ελάχιστο. Προσθέσαμε σ' αυτό την τιμή του κελιού του πίνακα A, αλλά αυτό το κάναμε μόνο για τα κελιά που βρίσκονται στην τελευταία στήλη και σε οποιαδήποτε γραμμή του πίνακα μεγαλύτερη ή ίση της δεύτερης. Στην 4 σχέση περιγράφεται η διαδικασία αυτή.

4) $OPT[i,j]=\{\min\{OPT[i-1,j], OPT[i-1,j-1]\} + A[i,j], 2\leq i\leq m, j=n\}$

Για να βρούμε την βέλτιστη λύση πήραμε με την τιμή του ελαχίστου κελιού της τελευταίας γραμμής του πίνακα. Η πράξη αυτή περιγράφεται μαθηματικά ως εξής:

5) $OPT^* = \min\{A[i,j], j=1....n\}$

Τώρα που υπολογίσαμε την αναδρομική σχέση μπορούμε να απαντήσουμε στην ερώτηση του πρώτου ερωτήματος. Άρα η αναδρομική σχέση που περιγράφει την βέλτιστη λύση του δυναμικού προγραμματισμού είναι η 5. Ένα άλλο σημαντικό στοιχείο που πρέπει να καταγράψουμε εδώ είναι ότι ισχύουν οι τρεις συνθήκες του Δυναμικού Προγραμματισμού.

1) Ισχύει η ιδιότητα των βέλτιστων επιμέρους δομών αφού υπολογίζουμε τις βέλτιστες λύσεις συνδυάζοντας τις βέλτιστες λύσεις των υποπροβλημάτων.

2) Παρατηρώ ότι έχω μικρό αριθμό υποπροβλημάτων.

3) Ισχύει η ιδιότητα των επικαλυπτόμενων επιμέρους δομών δηλαδή ο υπολογισμός ενός κελιού πολλές φορές

β) Το ερώτημα β μας ζητάει να γράψουμε τον αλγόριθμο που βρήκαμε σε ψευδοκώδικα αυτό γίνεται ως εξής:

- Αρχικά ονομάζουμε τον κώδικα
- Υπολογίζουμε τις βέλτιστες λύσεις στην πρώτη γραμμή γράφοντάς τες σε ψευδοκώδικα χρησιμοποιώντας τη σχέση 1.
- Γράφω στη συνέχεια το κυρίως σώμα του κώδικα χρησιμοποιώντας τις σχέσεις 2,3,4.
- Στη συνέχεια υπολογίζω το ελάχιστο για όλα τα στοιχεία της τελευταίας γραμμής.
- Τέλος τερματίζει ο αλγόριθμος επιστρέφοντας την τιμή πίσω ως το κόστος της βέλτιστης λύσης.

Άρα ο αλγόριθμος σε ψευδογλώσσα είναι ο εξής:

Procedure Dynamic(A)

for (j=1 to n)

OPT[1,j]=A[1,j]

end for

for (i=2 to m)

OPT[i,1]=min{OPT[i-1,2]} +A[i,1]

for(j=2 to n-1)

OPT[i,j]=min{OPT[i-1,j-1]},OPT[i-1,j],OPT[i-1,j+1]+A[i,j]

end for

OPT[i,n]=min{OPT[i-1,n-1],OPT[i-1,n]} +A[i,n]

minimum=OPT[m,1]

for(i=2 to n)

if(OPT[m,j]<minimum

end if

end for

return minimum

end procedure

3.4 Αλγόριθμος BLAST(Basic Logical Alignment and Search Tool)

3.4.1 Εισαγωγή

Ο αλγόριθμος BLAST δημιουργήθηκε το 1990 από τον Stephan Altschul (*Stephan Frank Altschul, 1990*) και από τους συνεργάτες του σε μια εποχή όπου ήταν επιτακτική η δημιουργία ενός ευριστικού αλγορίθμου που είχε ως στόχο την γρήγορη αναζήτηση και σύγκριση ακολουθιών στις βάσεις δεδομένων. Ο αλγόριθμος BLAST χρησιμοποιείται ευρέως στον τομέα κυρίως της βιοπληροφορικής αφού η γρήγορη εφαρμογή του, επιτρέπει την ενσωμάτωσή του σε servers καθώς και την εκτέλεσή του όταν υπάρχει πρόσβαση στο διαδίκτυο.

3.4.2 Τρόπος εκτέλεσης του αλγορίθμου.

Ο αλγόριθμος εκτελείται μέσω τριών σταδίων.

- Στο πρώτο στάδιο εκτέλεσης του αλγορίθμου κατασκευάζουμε μία λίστα από λέξεις- ακολουθίες οι οποίες διαιρούνται σε υποακολουθίες που έχουν συγκεκριμένο μήκος(w).
- Στο δεύτερο στάδιο ενώ έχουμε κατασκευάσει την λίστα- κατάλογο λέξεων για την αλληλουχία που εξετάζουμε κάθε φορά, παίρνουμε με τη σειρά την κάθε αλληλουχία και εξετάζουμε αν υπάρχει κοινό ζεύγος λέξεων. Η διαδικασία είναι λίγο χρονοβόρα όμως το θετικό στοιχείο αυτής της διαδικασίας είναι ότι απορρίπτει αυτόματα τις αλληλουχίες που δεν έχουν κοινές λέξεις.
- Στο τρίτο στάδιο σημαντικό ρόλο παίζει η παράμετρος E δηλαδή το κριτήριο στατικής σημασίας. Αυτό αναφέρεται στο επιθυμητό ζεύγος λέξεων. Η παράμετρος αυτή επηρεάζεται από τη βαθμολόγηση της ακολουθίας δηλαδή αν η βαθμολόγηση είναι μικρή τότε θα έχουμε ως αποτέλεσμα την απόρριψη αρκετών ζευγών λέξεων.

Παράδειγμα 5

Έστω ότι έχουμε την ακολουθία Q=AGTCGATCGA το μήκος της ακολουθίας είναι l=10. Αν το μήκος της κάθε λέξης είναι w=6 πόσες ακολουθίες λέξεων θα δημιουργηθούν;

Απάντηση

Για να βρούμε τον αριθμό των λέξεων που θα δημιουργηθούν κάνουμε την εξής πράξη: $l-w+1=10-6+1=5$

AGTCGA
GTCGAT
TCGATC
CGATCG
GATCGA

3.5 Αλγόριθμος Fasta

3.5.1 Εισαγωγή

Αν και ο αλγόριθμος αυτός δεν έχει πολλές διαφορές σε σχέση με τον αλγόριθμο BLAST, θεωρείται και αυτός έναν ευριστικό αλγόριθμο στην γρήγορη αναζήτηση πρωτεϊνικών ακολουθιών σε βάσεις δεδομένων. Με τον όρο ευριστικό εννοούμε τον αλγόριθμο που επιτυγχάνει μια εύκολη και ταχεία λύση αλλά όχι βέλτιστη.

Δημιουργήθηκε το 1985 από τους Lipman και Pearson(*David J.Lipman and William R.Pearson,1985*) και η εφαρμογή αυτού του αλγορίθμου είχε ως στόχο την προσέγγιση των ομοιοτήτων των ακολουθιών σε βάσεις δεδομένων.

3.5.2 Δομή και λειτουργία αλγορίθμου.

Σημαντικό ρόλο στη δομή και στη λειτουργία του αλγορίθμου αυτού παίζει η παράμετρος κτυρ όπου καθορίζεται κάθε φορά από τον χρήστη και αναφέρεται στην αναζήτηση μικρών λέξεων των ακολουθιών. Αρχικά αλγόριθμος αναζητά αυτές τις λέξεις στον πίνακα δυναμικού προγραμματισμού που δημιουργείται μέσω των ακολουθιών. Οι λέξεις που ανακαλύπτει ο αλγόριθμος ονομάζονται 'hot spots'. Ο εντοπισμός αυτών των λέξεων προέρχεται από τα ζεύγη(i,j) του πίνακα. Θα πρέπει δηλαδή το στοιχείο (i) του πίνακα της μιας ακολουθίας να ταυτίζεται με το στοιχείο (j) του πίνακα της άλλης ακολουθίας. Αφού ο αλγόριθμος εντοπίσει όλα αυτά τα ζεύγη επιλέγει εκείνα που σχηματίζουν διαγώνιο. Επιλέγει κυρίως τα δέκα καλύτερα ζεύγη. Τα στοιχεία που δεν δημιουργούν ζεύγη τα συμβολίζουμε με μηδενικά στον πίνακα. Τέλος παίρνουμε τα ζεύγη των διαγωνίων και τα ενώνουμε μεταξύ τους δημιουργώντας έτσι το τελικό αποτέλεσμα της ακολουθίας.

3.6 Αλγόριθμος Smith-Waterman.

Ο αλγόριθμος Smith-Waterman είναι ένας αλγόριθμος που πραγματοποιεί τοπική ευθυγράμμιση ακολουθιών. Δημιουργήθηκε για πρώτη φορά το 1981 από τους Smith-Waterman(*F.Smith, Michael S.Waterman,1981*) χρησιμοποιώντας την τεχνική του δυναμικού προγραμματισμού. Στόχος του είναι η σύγκριση συμβολοσειρών μεγάλου μήκους παράγοντας το βέλτιστο αποτέλεσμα ομοιότητας της ακολουθίας. Πραγματοποιώντας κάποιους υπολογισμούς εμφανίζει έναν πίνακα H ο οποίος περιέχει κάποια κόστη από τη σύγκριση των ακολουθιών. Τα κόστη αυτά καταγράφονται, αθροίζονται και επιλέγεται το μεγαλύτερο ως αποτέλεσμα[18]. Για κάθε θέση του πίνακα έχουμε την ακόλουθη τιμή:

$$H(i,k)=\max \left\{ \begin{array}{l} 0 \\ H(i-1,k-1) + S(a_i,b_j) \\ H(i-1,k) -G \\ H(i,k-1)-G \end{array} \right\}$$

Όπου $S(a_i,b_j)$ μια συνάρτηση αλφαβητικής αντιστοίχισης για μεταβολές του κόστους και G η τιμή του κόστους σε κάθε μεταβολή.

Κεφάλαιο 4

Προγράμματα και γλώσσες προγραμματισμού μοριακών αλγορίθμων.

4.1 Γλώσσες προγραμματισμού.

Για την επίλυση των μοριακών αλγορίθμων οι επιστήμονες χρησιμοποιούν ειδικά προγράμματα που ο κώδικάς τους είναι γραμμένος σε κάποια γλώσσα προγραμματισμού όπως JavaScript, Perl και R.

Perl: Είναι μια πολύ δημοφιλής αντικειμενοστραφής γλώσσα προγραμματισμού η οποία σχεδιάστηκε από τον Larry Wall (Larry Wall, 1987), και είναι μια διερμηνεύσιμη γλώσσα που κατατάσσεται στο ελεύθερο λογισμικό βάσει της άδειας του κώδικα της GPL [14]. Βασικό χαρακτηριστικό της είναι ότι μπορεί να επεξεργάζεται αριθμούς (δεκαδικούς και ακέραιους) καθώς και συμβολοσειρές. Η Perl μπορεί να διαχειρίζεται και πίνακες απλούς ή συσχετικούς. Στους συσχετικούς πίνακες έχουμε το προνόμιο της δημιουργίας δυναμικών δομών όπως δέντρα, στοίβες, ουρές και άλλα [14]. Η υλοποίησή της βασίζεται στη γλώσσα προγραμματισμού C. Έχει πολλά πλεονεκτήματα καθώς εφαρμόζεται σε πολλές πλατφόρμες όπως Unix, Windows. Χρησιμοποιείται κυρίως στον τομέα της Βιοπληροφορικής δεδομένου του γεγονότος ότι μπορεί να υποστηρίξει εκφράσεις που περιγράφουν την δομή μιας συμβολοσειράς. Γι αυτό και στην Βιοπληροφορική χρησιμοποιείται η Bioperl ένα πρόγραμμα βασισμένο στη χαρτογράφηση του ανθρώπινου γονιδιώματος [15].

JavaScript: Είναι μια διερμηνεύσιμη γλώσσα προγραμματισμού δημιουργήθηκε το 1995 από τον προγραμματιστή Brendan Eich (Brendan Eich, 1995), της εταιρείας Netscape με την επωνυμία Mocha [16]. Το 1995 μετονομάστηκε σε JavaScript επειδή η ανάπτυξή της επηρεάστηκε περισσότερο από τη Java [16]. Διαφέρει από τις άλλες γλώσσες προγραμματισμού διότι η επεξεργασία του κώδικα περιεχομένου δεν εκτελούνται στον διακομιστή αλλά στο πρόγραμμα περιήγησης επισκεπτών. Έγινε ιδιαίτερα γνωστή σε προγράμματα δυναμικού περιεχομένου για την εκτέλεση κώδικα σε ιστοσελίδες.

R: Είναι μια γλώσσα προγραμματισμού βασισμένη στη C αλλά και στη Fortran. Η πρώτη της έκδοση κυκλοφόρησε το 1994 από τους Ross Ihaka και Robert Gentleman (Ross Ihaka, Robert Gentleman, 1994) στο Πανεπιστήμιο του Auckland στη Νέα Ζηλανδία [17]. Ονομάστηκε R από τα αρχικά γράμματα των ονομάτων των δημιουργών της. Ένα από τα πλεονεκτήματά της είναι ότι περιέχει πολλά έτοιμα πακέτα αναλύσεων, η γρήγορη αποθήκευση μεγάλου όγκου δεδομένων καθώς και η σχεδίαση ποιοτικών διαγραμμάτων [15].

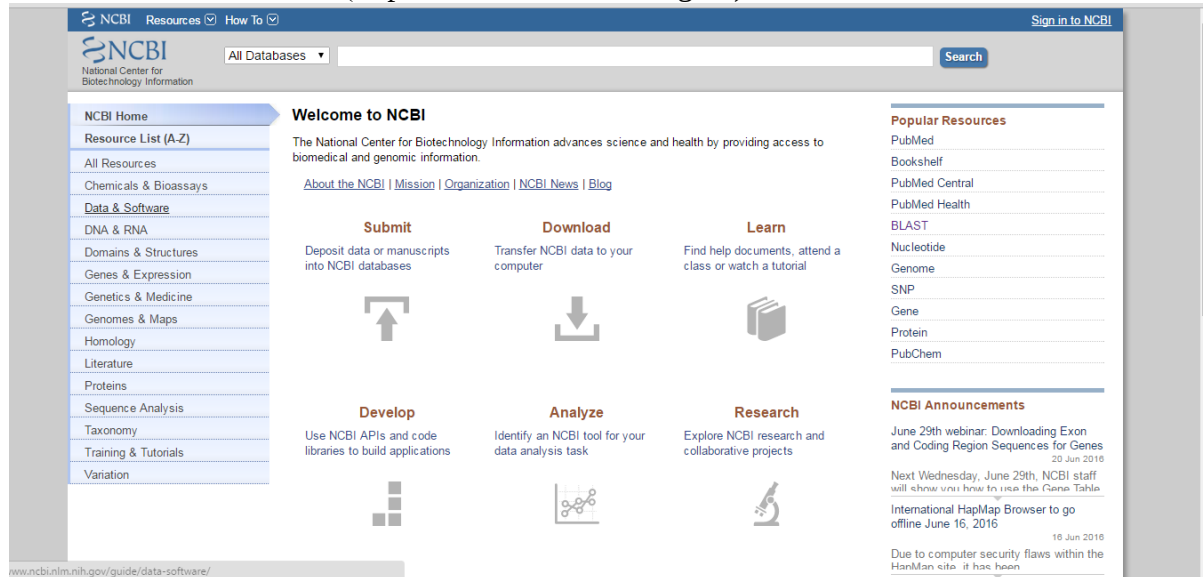
4.2 Προγράμματα Μοριακής Βιολογίας.

4.2.1 Πρόγραμμα BLAST

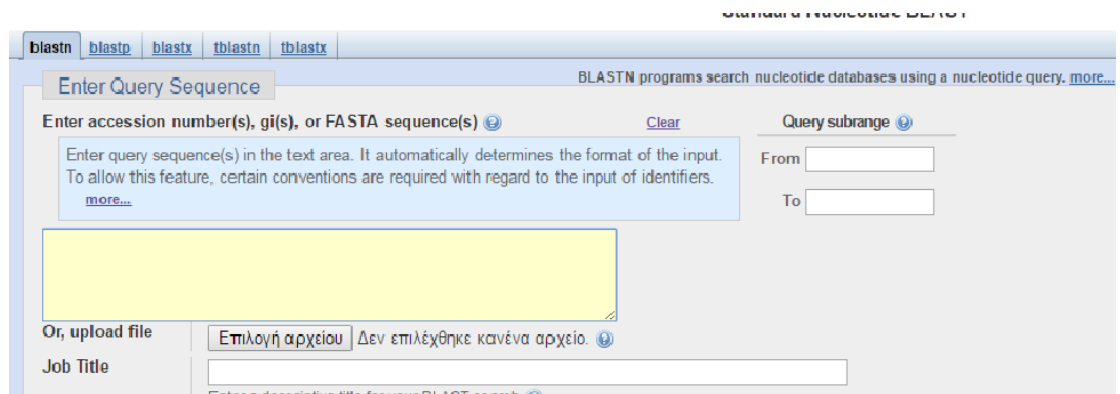
Υπάρχουν αρκετά προγράμματα που χρησιμοποιεί η Μοριακή βιολογία κυρίως για τη σύγκριση ακολουθιών όπως το πρόγραμμα το Blast που βρίσκεται στην ιστοσελίδα του Εθνικού Κέντρου Βιοτεχνολογίας (<http://www.ncbi.nlm.nih.gov/BLAST/download.shtml>). Το πρόγραμμα αυτό αποτελείται από βάσεις δεδομένων που περιέχουν γενετικές ακολουθίες που έχουν ως στόχο τη σύγκριση βάσης προς βάση ακολουθιών. Γίνεται δηλαδή σύγκριση δυο ακολουθιών που τις εισχωρεί ο χρήστης με κατάλληλο τρόπο. Το πρόγραμμα αυτό

βρίσκει τρομερή απήχηση κυρίως στη σύγκριση ακολουθιών διαφόρων ιών. Παρακάτω αναφέρουμε κάποιες από τις λειτουργίες του προγράμματος.

- Αρχικά για να έχουμε πρόσβαση στο Blast εισερχόμαστε στη σελίδα του ncbi που είναι (<http://www.ncbi.nlm.nih.gov/>).



- Από αυτήν την σελίδα μπορούμε να εισάγουμε τις ακολουθίες που θέλουμε να συγκρίνουμε. Επιλέγουμε την επιλογή Blast που είναι στην δεξιά στήλη και εμφανίζεται το ακόλουθο παράθυρο:



Εμφανίζεται ένα παράθυρο με τις ακόλουθες επιλογές:

- **blastn**: Κάνει έλεγχο μιας νουκλεοτιδικής ακολουθίας ή σύγκριση μεταξύ δυο νουκλεοτιδικών ακολουθιών σε μια βάση δεδομένων.
- **blastp**: Συγκρίνει και ελέγχει πρωτεϊνικές ακολουθίες σε μια βάση δεδομένων.
- **blastx**: Συγκρίνει μια νουκλεοτιδική ακολουθία με ακολουθίες πρωτεϊνών. Πραγματοποιεί έλεγχο της ακολουθίας που έχει εισχωρήσει ο χρήστης και προσπαθεί να βρει παρόμοιες ακολουθίες στη πρωτεϊνική βάση δεδομένων.
- **tblastn**: Συγκρίνει πρωτεϊνικές ακολουθίες με νουκλεοτιδικές στη νουκλεοτιδική βάση δεδομένων.

Εμείς στη συγκεκριμένη εργασία θα αναφέρουμε κάποια παραδείγματα σύγκρισης ακολουθιών για κάποιους ιούς.

Παράδειγμα σύγκρισης ιού H1V1 που θεωρείται υπεύθυνος για το AIDS(Σύνδρομο Επίκτητης Ανοσολογικής Ανεπάρκειας) με τον αντίστοιχο ιό που προσβάλλει τα πιθηκοειδή SIV1.

- Γράφουμε τις ακολουθίες των ιών στα πλαίσια της βάσης δεδομένων όπως φαίνεται παρακάτω έχοντας χρησιμοποιήσει την ακολουθία για τον ιό SIV1 (L0604.2) και για τον HIV1(NC_001802.1) και επιλέγουμε την επιλογή BLAST κάτω αριστερά :

BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

NC_001802.1

[Clear](#) [Query subrange](#) [?](#)

From

To

Or, upload file Δεν επιλέχθηκε κανένα αρχείο. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☒ Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

L06042.1

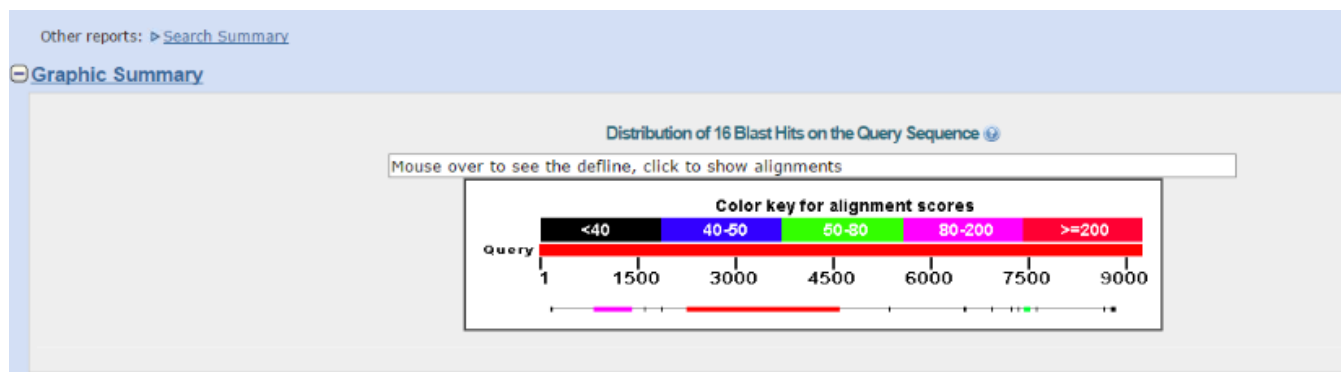
[Clear](#) [Subject subrange](#) [?](#)

From

To

Or, upload file Δεν επιλέχθηκε κανένα αρχείο. [?](#)

- Εμφανίζεται το παρακάτω παράθυρο με την γραφική απεικόνιση των αποτελεσμάτων χρησιμοποιώντας 5 διαφορετικά χρώματα.



- Παρακάτω φαίνονται τα αποτελέσματα σύγκρισης ακολουθιών:

Human immunodeficiency virus 1, complete genome

Sequence ID: [ref|NC_001802.1|](#) Length: 9181 Number of Matches: 16

Range 1: 2255 to 4594 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
417 bits(462)	1e-118	1560/2397(65%)	114/2397(4%)	Plus/Plus
Query 2906	AATCCTTATAATACCCCACTATTTGTAATTAGGAAAAAGGACAAAAAGAAATGGAGGATG			2965
Sbjct 2255	AATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAGACAGTACTAAATGGAGAAAA			2314
Query 2966	CTCATTGATTTTAGAAAATTAAATGAGGCCACACAAGATTTTTTCGAGGTACAATTGGGA			3025
Sbjct 2315	TTAGTAGATTTTACAGAACTTAATAAGAGAACTCAAGACTTCTGGGAAGTTCAATTAGGA			2374
Query 3026	ATACCCCAACCCAGGAGGACTAAAACGGCAAAAAC---TAACAATTATAGATTTAAAAGAT			3082
Sbjct 2375	ATACCACATCCCGCAGGGTTAAAAAGAAAAAATCAGTAACAGTACTGGATGTGGGTGAT			2434
Query 3083	GCATATTACTCAGTCCCTCCGCAAAAGAGTTTAGGCCATACACGGCATTACAGTGCCA			3142
Sbjct 2435	GCATATTTTTCAGTCCCTTATGATGAAGACTTCAGGAAGTACTGCATTTACCATACCT			2494
Query 3143	AGCATAAATAATGCCTCCCGGGAGAAAGATATCAATTTACAGTGTACCACAGGGATGG			3202
Sbjct 2495	AGTATAAACAATGAGACACAGGGATTAGATATCAGTACAATGTCTTCCACAGGGATGG			2554
Query 3203	AAAGGATCTCCCACTATCTTCCAAAGCACAATAAATCAGATACTACAGCCATTTAGAAAG			3262
Sbjct 2555	AAAGGATCACCAGCAATATTCCAAAGTAGCATGACAAAAATCTTAGAGCCTTTAGAAAA			2614
Query 3263	AAGTATTCAGATCTCACTCTAATACAGTACATGGATGACTTATTGATTGGCACTGATAGA			3322
Sbjct 2615	CAAAATCCAGACATAGTTATCTATCAATACATGGATGATTTGTATGTAGGATCTGAC---			2671
Query 3323	TCAGAGAAGGCACATCAGGAGATAGTACAACAAATAGTG-----ACAGCA-CTATTA			3373
Sbjct 2672	TTAGAAATAGGGCAGCA-----TAGAACAA-AAATAGAGGAGCTGAGACAACATCTGTTG			2725
Query 3374	AAAGTGGGATTTAAGGTCCCAAAAGAAAAATGGCAGGACCAATATCCAATGCAGTGGTTA			3433
Sbjct 2726	AGGTGGGGACTTACCACACCAGACAAAAAACATCAGAAAGAACCTCCATTCTTTGGATG			2785
Query 3434	GGTTACACTCTTCATCCAGACAAGTGG-CAGTTGCAGAAAATAGAACTCCCTAATATAGA			3492
Sbjct 2786	GGTTATGAATCCATCCTGATAAATGGACAGTA-CAGCCTATAGTGCTGCC-----AGA			2838
Query 3493	TGACGAAATC---ACAGTGAATCAGTTACAGAAATTGATAGGAGTACTTAATTGGGCAAG			3549
Sbjct 2839	AAAAGACAGCTGGACTGTCAATGACATACAGAAAGTTAGTGGGGAAATTGAATTGGGCAAG			2898
Query 3550	CCAGATTTATTTCAGGAATTAAGACAAAAGAGCTCTGTAAGTGTATCAGAGGAACTAAACA			3609
Sbjct 2899	TCAGATTTACCCAGGGATTAAAGTAAGGCAATTATGTAACTCCTTAGAGGAACCAAAGC			2958
Query 3610	ATTAACAGAAAGTACTCACACTTACAGAAAGCAGCTGAGGCAGAAATTAGAAGAGAATAGACA			3669
Sbjct 2959	ACTAACAGAAAGTAAATACCACTAACAGAAAGAGCAGAGCTAGAATGGCAGAAAACAGAGA			3018
Query 3670	AATCTTAAAGGAA---GAACAGGCAGGGTCTTATTATGACCTTAAAAAACCTTTAGAAAGC			3726
Sbjct 3019	GATTCTAAAGAACCAAGTACATGGAGTGT---ATTATGACCCATCAAAAGACTTAATAGC			3075
Query 3727	ACACATCACTAACTAGGGTCCAGC-AATGGGGATACATGATTAAGCAAGAGC-----A			3780
Sbjct 3076	AGAAAT-ACAGAAAGCAGGGGCAAGGCCAATGGACATATCAAATTTATCAAGAGCCATTTA			3134
Query 3781	AAAAGGCCACCCCTAATAACAGGAAAGACAGCAAAAACCTTTGCAGCACATAGTAATGA			3840
Sbjct 3135	AAAAT-----CTGAAAACAGGAAATATGCAAGAATGAGGGGTGCCACACTAATGA			3186
Query 3841	CTATCAATCTCTTG-CTCAATTGTTAAATAAAATAGGGATCCAATCACTATGGTATT--G			3897
Sbjct 3187	-TGTAACAATTAACAGAGGCAGTGCAAAAAATAACACAGAA--AGCATAGTAATATG			3243
Query 3898	GGGAAAGGTACCAACATTCCATCTCCAGTAAAAAGAGAGGAATGGGAAAAATGGTGGAC			3957
Sbjct 3244	GGGAAAGACTCCTAAATTTAACTGCCCATACAAAAGGAAACATGGGAAACATGGTGGAC			3303
Query 3958	AGACTATTGGCAAGCCACATGGGTACCAGAAGTCAAATTCATAAGTACCCCGCCATTGGT			4017
Sbjct 3304	AGAGTATTGGCAAGCCACCTGGATTCTGAGTGGGAGTTGTATATACCCCTCCCTTAGT			3363

Query 4018 TAGGTGGTACTATAACTTTGGTACCAGAACCCATACCAGAAGCAGTAACCTTCTACGTAGA 4077
Sbjct 3364 GAAATTATGGTACCAGTTAGAGAAAGAACCCATAGTAGGAGCAGAAACCTTCTATGTAGA 3423

Query 4078 TGGAGCAGCTAATAGAGACAGCAAGACAGGAAATGCAGGATATGTAGCAAGTGATGGGAC 4137
Sbjct 3424 TGGGGCAGCTAACAGGGAGACTAAATTAGGAAAAGCAGGATATGTTACTAATAGAGGAAG 3483

Query 4138 ACAGAGAGT----ACAGTATCTGGAGCAAAACAACCAACAGGCAGAAATTAGAGGGC- 4192
Sbjct 3484 ACAAAAAGTTGTCAACCCTAACTG----ACACAACAAATCAGAAGACTGAGTTACAAGCAA 3539

Query 4193 TTGCTCATGGCATTGCAGGACTCAAAAGACAAAGTAAATATAGTAGTAGATTACAATAC 4252
Sbjct 3540 TTTATC-TAGCTTTGCAGGATTCGGGATTAGAAAGTAAACATAGTAACAGACTCACAATAT 3598

Query 4253 TCTTATGGCATTCTTATGACATGCCCCACTAA--CACAGAACACCCAA--TAGTGGAGCA 4308
Sbjct 3599 GCATTAGGAATCATT----CAAGCACAAACAGATCAAAGTGAATCAGAGTTAGTCAATCA 3654

Query 4309 GATAATACAAGAGGCCATAAAAAAGGAAGCCATATATGTGACATGGGTCCCAGGCCATAA 4368
Sbjct 3655 AATAATAGAGCAGTTAATAAAAAAGGAAAAGGTCTATCTGGCATGGGTACCAGCACACAA 3714

Query 4369 AGGAATAGGAGGAAATGAAGCTGTTGATAAACTAGTCAGTAAAGGAATAAGAAAAATCCT 4428
Sbjct 3715 AGGAATTGGAGGAAATGAACAAGTAGATAAATTAGTCAGTGCTGGAATCAGGAAAGTACT 3774

Query 4429 ATTCCTAGAAAAGATTCCCCAAGCACAGGAAGATCATGAAAGGTACCACTCAATATGGA 4488
Sbjct 3775 ATTTTATAGTAGGAATAGATAAGGCCCAAGATGAACATGAGAAATATCACAGTAAT-TGGA 3833

Query 4489 GTATCTTAGGC-AG-GAGTTTCATCTACCCCGACAGGTAGCAAAAGCAATTATA--CAAC 4544
Sbjct 3834 G-AGCAATGGCTAGTGATTTTAACTGCCACCTGTAGTAGCAAAAGAAATAGTAGCCAGC 3892

Query 4545 AATGTCCAAAATGTCAAAATAGAGGAGAACCAAAACATGGGCAAGTAGA-TGTAGATATA 4603
Sbjct 3893 --TGTGATAAATGTCAGCTAAAAGGAGAAGCCATGCATGGACAAGTAGACTGTAG-TCCA 3949

Query 4604 TATAATTGGCAGATGGACTGTACTCATGAGGAAGGCAAAAGTAATCTGTGTAGCAGTAAAC 4663
Sbjct 3950 GGAATATGGCACTAGATTGTACACATTTAGAAAGGAAAAGTTATCCTGGTAGCAGTTCAT 4009

Query 4664 ACAGCTAGTGGATATATAGAAAC-CAAGATCTTAAAAAGGGAACGGGGGATGAAACAGC 4722
Sbjct 4010 GTAGCCAGTGGATATATAGAAGCAGAAGTTATT-CCAGCAGAAACAGGGCAGGAAACAGC 4068

Query 4723 ATTGTTTCTCATGCAAAATAGCCAGTAGATGGCCAATAAAACAGATACACACTGATAATGG 4782
Sbjct 4069 ATATTTTCTTTTAAAAATAGCAGGAAGATGGCCAGTAAAAACAATACATACTGACAATGG 4128

Query 4783 ACCAAACTTTGTCACTGACAAAGTTCAAAGCAGCATGTTGGTGGTGTGGCATAGAACACAC 4842
Sbjct 4129 CAGCAATTTCAACGGTGTACGGTTAGGGCCGCCTGTTGGTGGGCGGGAATCAAGCAGGA 4188

Query 4843 TACAGGTATTCCATACAACCCACAAAGCCAGGGAATAGTAGAGTCCAAAAATAGGTATTT 4902
Sbjct 4189 ATTTGGAAATCCCTACAATCCCAAGTCAAGGAGTAGTAGAATCTATGAATAAAGAATT 4248

Query 4903 AAAAGAAGCTATCTCACAATATAGAGACGACGTACACATTTGCAAAACAGCAGTGGCAAT 4962
Sbjct 4249 AAAGAAAATTATAGGACAGGTAAGAGATCAGGCTGAACATCTTAAGACAGCAGTACAAAT 4308

Query 4963 GGCAACCTTCATCTTAAATTTTAAAAGAAAAGGGAGGAATAGGGGGGATATCA--CCAGGA 5020
Sbjct 4309 GGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGG-TA-CAGTGCAGGG 4366

Query 5021 GAAAG---GTACA-TTAATATGCTATATACA---GAACTACAATTACAACAAAATACAA 5072
Sbjct 4367 GAAAGAATAGTAGACATAATA-GCAACAGACATACAACTA-AAGAATTACAAAA-ACAA 4423

Query 5073 CATCACCAAAATTTTCGAATTTTAGGGTTTACTAC--CGCCAGGGGAAGAACGA--GTGG 5128
Sbjct 4424 -ATTA-CAAAAATCAAAAATTTTCGGGTTTATTACAGGGACAGCAGAA-ATCCACTTTGG 4480

Query 5129 AAGGGACCAGCTCGTCTACTTTGGAAAGGTGAAGGTGCAGTAGTGGTGCAACAGAGGAG 5188
Sbjct 4481 AAAGGACCAGCAAGCTCCTCTGGAAAGGTGAAGGGGCAGTAGTAATACA--AGATAAT 4537

Query 5189 GGAGACATTTTTCAGTACCAAGAAGAAAGGCAAAAATAATTACAGACCATGGAGAA 5245
Sbjct 4538 AGTGACATAAAAGTAGTGCCAAGAAGAAAAGCAAGATCATTAGGGATTATGGAAAA 4594

Στο παραπάνω παράδειγμα έγινε σύγκριση μεταξύ των δυο ιών και τα αποτελέσματα έδειξαν ότι οι δυο ιοί εμφανίζουν ομοιότητα 65%. Το αποτέλεσμα αυτό το παρατηρούμε από το παράθυρο description στην περιγραφή Ident.

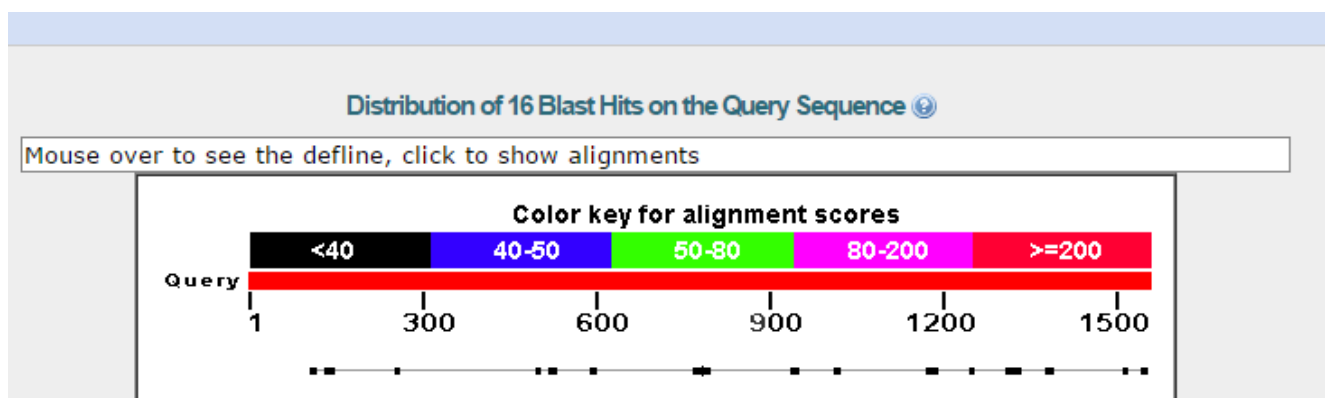
	Max score	Total score	Query cover	E value	Ident	Accession
	417	978	34%	1e-118	65%	NC_001802.1

Στο σημείο αυτό θα πρέπει να αναφέρουμε ότι ο ιός δεν μεταδίδεται σε καμία περίπτωση από τον πίθηκο στον άνθρωπο . Γνωρίζοντας ότι ο ιός SIV1 υπήρχε πριν από τον ιό HIV1, καταλήγουμε στο συμπέρασμα ότι ο ιός HIV1 ίσως δημιουργήθηκε από τον ιό SIV1 χωρίς ωστόσο να υπάρχουν αποδείξεις για το συμπέρασμα αυτό[18]. Η μετάλλαξη του ιού είχε ως αποτέλεσμα την γρήγορη εξάπλωσή του όχι μόνο στην Αφρική, που τα αποτελέσματα είναι πραγματικά απογοητευτικά, αλλά και σε άλλες χώρες διότι είναι ένας από τους πιο επικίνδυνους ιούς παγκοσμίως.

Σύγκριση ιού της γρίπης τύπου A με ιο γρίπης τύπου B.

Στη συνέχεια θα συγκρίνουμε δυο ιούς που εμφανίστηκαν αρχικά στο Πεκίνο το 1987 και στην Ταϊλάνδη το 2004. Ο ιός που εμφανίστηκε στο Πεκίνο ήταν ο ιός της γρίπης τύπου B ενώ στην Ταϊλάνδη ο ιός της γρίπης τύπου A. Θα πρέπει πάντα να συγκρίνουμε ιούς που εμφανίζουν κάποια ομοιότητα.

Συγκρίνουμε τα δύο στελέχη μεταξύ τους . Το στέλεχος για τον ιό της γρίπης τύπου A είναι AY555151.3 και του ιού της γρίπης τύπου B είναι M54967.2 Εμφανίζονται τα παρακάτω παράθυρα με την γραφική απεικόνιση των αποτελεσμάτων και της σύγκρισης των ακολουθιών.



Download [GenBank](#) [Graphics](#) Sort by: **E value**

Influenza A virus (A/Thailand/1(KAN-1)/2004(H5N1)) neuraminidase gene, complete cds

Sequence ID: [gb|AY555151.3|](#) Length: 1377 Number of Matches: 16

Range 1: 660 to 675 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
24.7 bits(26)	0.075	15/16(94%)	0/16(0%)	Plus/Minus

Query 1373 AGCAACAGCCATTAC 1388
 Sbjct 675 AGCAAGAGCCATTAC 660

Range 2: 260 to 271 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
22.9 bits(24)	0.26	12/12(100%)	0/12(0%)	Plus/Minus

Query 1506 TTGTCCTTACTG 1517
 Sbjct 271 TTGTCCTTACTG 260

Range 3: 506 to 520 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
22.9 bits(24)	0.26	14/15(93%)	0/15(0%)	Plus/Plus

Query 590 TGCATGCCATGATGG 604
 Sbjct 506 TGCTTGCCATGATGG 520

Range 4: 203 to 216 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
21.1 bits(22)	0.92	13/14(93%)	0/14(0%)	Plus/Plus

Query 521 TTCAGTCAAATTAG 534
 Sbjct 203 TTCAGTAAAATTAG 216

Range 5: 257 to 275 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
21.1 bits(22)	0.92	16/19(84%)	0/19(0%)	Plus/Minus

Query 134 ACTGTCATACTTACTGTAT 152
 Sbjct 275 ACTGTTGTCCTTACTGTAT 257

Range 6: 778 to 788 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
21.1 bits(22)	0.92	11/11(100%)	0/11(0%)	Plus/Plus

Query 110 ATTATCACTAT 120
 Sbjct 778 ATTATCACTAT 788

Range 7: 1146 to 1161 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
21.1 bits(22)	0.92	14/16(88%)	0/16(0%)	Plus/Plus

Query 768 ATAAGTATGGCTCAG 783
 Sbjct 1146 ATAAGTATGGCTCAG 1161

Range 8: 1290 to 1305 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
21.1 bits(22)	0.92	14/16(88%)	0/16(0%)	Plus/Minus

Query 935 TTACACAGCAAAAAGA 950
 Sbjct 1305 TTACACCAAAAAGA 1290

Range 9: 26 to 35 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	10/10(100%)	0/10(0%)	Plus/Minus

Query 1012 CTTATTTGGA 1021
 Sbjct 35 CTTATTTGGA 26

Range 10: 225 to 234 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	10/10(100%)	0/10(0%)	Plus/Minus

Query 255 AAAGAGATGA 264
 Sbjct 234 AAAGAGATGA 225

Range 11: 554 to 576 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	18/23(78%)	0/23(0%)	Plus/Plus

Query 1168 AGAGAATGGGGATGGAAGTGTAT 1190
 Sbjct 554 AGACAATGGGGCTGTGGCTGTAT 576

Range 12: 699 to 708 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	10/10(100%)	0/10(0%)	Plus/Plus

Query 1244 AGTAATGGTC 1253
 Sbjct 699 AGTAATGGTC 708

Range 13: 793 to 805 [GenBank](#) [Graphics](#) [▼](#) Next Match [▲](#) Previous Match [▲](#) First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	12/13(92%)	0/13(0%)	Plus/Plus

Query 1537 AATGCTCTTGTTA 1549
 Sbjct 793 AATGCTCCTGTTA 805

▼

Next Match

▲

Previous Match

▲

First Match

Range 14: 804 to 828

[GenBank](#)
[Graphics](#)

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	19/25(76%)	0/25(0%)	Plus/Minus

Query

1307

ATGTGATGTCCCTGTATTGGGATA

1331

Sbjct

828

ATGTGATTTCGCCGGCATCAGGATA

804

Range 15: 1053 to 1067

[GenBank](#)
[Graphics](#)

▼

Next Match

▲

Previous Match

▲

First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	13/15(87%)	0/15(0%)	Plus/Minus

Query

783

GCTTCAGGAATTAGT

797

Sbjct

1067

GCTCCTGGAATTAGT

1053

Range 16: 1219 to 1228

[GenBank](#)
[Graphics](#)

▼

Next Match

▲

Previous Match

▲

First Match

Score	Expect	Identities	Gaps	Strand
19.3 bits(20)	3.2	10/10(100%)	0/10(0%)	Plus/Minus

Query

497

CAGAAACAAG

506

Sbjct

1228

CAGAAACAAG

1219

Η ομοιότητα των δυο ιών είναι 94% όπως φαίνεται στο ακόλουθο παράθυρο:

	Max score	Total score	Query cover	E value	Ident	Accession
	24.7	330	15%	0.075	94%	AY555151.3

4.2.2 Πρόγραμμα CLUSTALW

Το CLUSTALW είναι ένα πρόγραμμα σύγκρισης μοριακών ακολουθιών το οποίο περιέχει έναν πολύ δύσκολο αλγόριθμο που επιτρέπει την στοίχιση πολλών πρωτεϊνικών ή νουκλεοτιδικών ακολουθιών. Το πρόγραμμα CLUSTALW υπάρχει στη σελίδα του EBI(European Bioinformatics Institute) που είναι η ακόλουθη (<http://www.ebi.ac.uk/clustalw/>).

Πώς χρησιμοποιούμε το πρόγραμμα;

1. Πηγαίνουμε στη σελίδα του προγράμματος ncbi και βρίσκουμε τις ακολουθίες που θέλουμε να συγκρίνουμε. Εμφανίζεται ένα παράθυρο που μας επιτρέπει να κατεβάσουμε τις ακολουθίες που θέλουμε να συγκρίνουμε αποθηκεύοντας το αρχείο που περιέχει τις ακολουθίες σε ένα φάκελο :

NCBI Resources ▾ How To ▾

HomoloGene

[Save search](#) [Limits](#) [Advanced](#)

[Display Settings:](#) ▾ Summary

[Send to:](#) ▾ **Filter your results:**

All (2)

Fungi (0)

Mammals (0)

Results: 2

☐ [glycerol-3-phosphate dehydrogenase 1 \(soluble\)](#)

1. GPD1, Gpd1, gpd1b, gpd1a, **Gpdh**, **gpdh**-2, GPD2, gpd1, gpd2 - conserved in Opisthokonta
hgid: 5593

☐ [glycerol-3-phosphate dehydrogenase 2 \(mitochondrial\)](#)

2. GPD2, Gpd2, gpd2, Gpo-1, T25G3.4, SDP6 - conserved in Eukaryota
hgid: 352

Find related data

Database:

2. Στη συνέχεια χρησιμοποιούμε το πρόγραμμα CLUSTALW για να συγκρίνουμε τις ακολουθίες. Το πρόγραμμα υλοποιείται σε τρία βήματα:

STEP 1 - Enter your input sequences

[Enter or paste](#) a set of sequences in any supported format:

Or, [upload](#) a file: homologene (1).txt

STEP 2 - Set your parameters

OUTPUT FORMAT

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

Στο STEP1 εισάγουμε τις πρωτεϊνικές ακολουθίες που επιθυμούμε. Μπορούμε να επιλέξουμε το αρχείο που αποθηκεύσαμε από το πρόγραμμα ncbi επιλέγοντας το εικονίδιο Επιλογή αρχείου και εισάγουμε το φάκελο με τις ακολουθίες.

Στο STEP2 και STEP3 επιλέγουμε τις κατάλληλες παραμέτρους των ακολουθιών.

Τέλος πατάμε το κουμπί Submit.

3. Εμφανίζεται το παράθυρο με τα αποτελέσματα της σύγκρισης.

```
gi|115474601|ref|NP_001060897.1|
gi|50311981|ref|XP_456022.1|
gi|398364523|ref|NP_012483.3|
gi|398366083|ref|NP_0011708.3|
gi|6322468|ref|NP_012542.1|
gi|389640441|ref|XP_003717853.1|
gi|85082731|ref|XP_956977.1|
gi|17534677|ref|NP_496237.1|
gi|17534679|ref|NP_496192.1|
gi|32566163|ref|NP_508534.3|
gi|17568413|ref|NP_508535.1|
gi|19112946|ref|NP_596154.1|
gi|19112028|ref|NP_595236.1|
gi|169403947|ref|NP_001108586.1|
gi|52345806|ref|NP_001004949.1|
gi|46048961|ref|NP_989636.1|
gi|7669492|ref|NP_002037.2|
gi|55637711|ref|XP_508955.1|
gi|109095230|ref|XP_001105471.1|
gi|6679937|ref|NP_032110.1|
gi|149273202|ref|XP_001476757.1|
gi|8393418|ref|NP_058704.1|
gi|77404273|ref|NP_001029206.1|
gi|345807932|ref|XP_003435697.1|
gi|50978862|ref|NP_001003142.1|
gi|345798394|ref|XP_003434435.1|
gi|58391523|ref|XP_318655.2|
gi|442616493|ref|NP_001259584.1|
gi|22023983|ref|NP_525108.2|
```

```

IKIKDSKTLLLG-----EKPVTVFGRINPDEIPWAEAGAEYVVESTGVFTDKKEAAHKL
TT--SGMDLVID-----GHKIAVFQEKDPANLPMWGLGVDIVDSTGVFKELDSAQKHLD
SH--DDKHHIID-----GVKIATYQERDPANLPMWGLSKIDVADSTGVFKELDTAQKHID
SH--DDKHHIVD-----GHKIATFQERDPANLPMWASLNIDIAIDSTGVFKELDTAQKHID
EV--SGSDLVVN-----GKKVKFYTERDPANIPWSETGAEYVVESTGVFTTIDOKASAHKL
EV--DGADFLVN-----GKKVKFYTERDPAAPWSETGADYIVESTGVFTTTEKASAHKL
TY--DGDFLTVQDGGKSSHKIKVFNKSDPAAIAGWSVKADFVVESTGVFTTEKASAHLQ
TY--DGDFLTVQDGGKSSHKIKVFNKSDPAAIAGWSVKADFVVESTGVFTTEKASAHLQ
AH--EGDYLLVAKEGKSQHKIKVYNSRDPAEIQWAGSAGADYVVESTGVFTTEKANAHLK
AH--EGDYLLVAKEGKSQHKIKVYNSRDPAEIQWAGSAGADYVVESTGVFTTEKANAHLK
EI--KDGKLVID-----GNAIDVHNERDPADIKWSTSGADYIVESTGVFTTQETASAHKL
ET--KDGKLVID-----GHSIDVHNERDPANIKWASGAEYVVESTGVFTTETASAHKL
KA--EGGKLVID-----GHAITVYSERDPANIKWGDAGATYVVESTGVFTTEKASAHIK
CV--ENGKLVIN-----GKAVTVFQERDPSNIKWGDAGAYVVESTGVFTTEKAGLHLK
KA--ENGKLVIN-----GHAITIFQERDPSNIKWADAGAEYVVESTGVFTTEKAGAHKL
KA--ENGKLVIN-----GNPITIFQERDPSIKWGDAGAEYVVESTGVFTTMEKAGAHLQ
KA--ENGKLVIN-----GNPITIFQERDPSIKWGDAGAEYVVESTGVFTTMEKAGAHLQ
KA--ENGKLVIN-----GSPITIFQERDPSIKWGDAGAEYVVESTGVFTTTEKAGAHLQ
KA--ENGKLVIN-----GKPITIFQERDPTNIKWGEAGAEYVVESTGVFTTEKAGAHKL
KA--ENGKLVIN-----GKPITIFQERDPTNIKWGEAGAEYVVESTGVFTTMEKAGAHKL
KA--ENGKLVIN-----GKPITIFQERDPANIKWGDAGAEYVVESTGVFTTMEKAGAHKL
KA--ENGKLVIN-----GKAITIFQERDPANIKWGDAGAEYVVESTGVFTTMEKAGAHKL
KA--ENGKLVIN-----GKSISIFQERDPANIKWGDAGAEYVVESTGVFTTMEKAGAHKL
KA--ENGKLVIN-----GKSISIFQERDPANIKWGDAGAEYVVESTGVFTTMEKAGAHKL
SA--QDGCFLVN-----GQKIAVFQERDPKAIPMWKGAGAEYVVESTGVFTTEKASAHLE
AA--EGGFLVNN-----GQKITVFSERDPANINWASAGAEYIVESTGVFTTIDKASTHLK
AA--EGGFLVNN-----GQKITVFSERDPANINWASAGAEYVVESTGVFTTIDKASTHLK

```

```
gi|115474601|ref|NP_001060897.1|
gi|50311981|ref|XP_456022.1|
gi|398364523|ref|NP_012483.3|
gi|398366083|ref|NP_0011708.3|
gi|6322468|ref|NP_012542.1|
gi|389640441|ref|XP_003717853.1|
gi|85082731|ref|XP_596977.1|
gi|17534677|ref|NP_496237.1|
gi|17534679|ref|NP_496192.1|
gi|32566163|ref|NP_508534.3|
gi|17568413|ref|NP_508535.1|
gi|19112946|ref|NP_596154.1|
gi|19112028|ref|NP_595236.1|
gi|169403947|ref|NP_001108586.1|
gi|52345806|ref|NP_001004949.1|
gi|46048961|ref|NP_989636.1|
gi|7669492|ref|NP_002037.2|
gi|55637711|ref|XP_508955.1|
gi|109095230|ref|XP_001105471.1|
gi|6679937|ref|NP_032110.1|
gi|149273202|ref|XP_001476757.1|
gi|8393418|ref|NP_058704.1|
gi|77404273|ref|NP_001029206.1|
gi|345807932|ref|XP_003435697.1|
gi|50978862|ref|NP_001003142.1|
gi|345798394|ref|XP_003434435.1|
gi|58391523|ref|XP_318655.2|
gi|442616493|ref|NP_001259584.1|
gi|22023983|ref|NP_525108.1|
```



```
gi|115474601|ref|NP_001060897.1|
gi|50311981|ref|XP_456022.1|
gi|398364523|ref|NP_012483.3|
gi|398366083|ref|NP_011708.3|
gi|6322468|ref|NP_012542.1|
gi|389640441|ref|XP_003717853.1|
gi|85082731|ref|XP_956977.1|
gi|17534677|ref|NP_496237.1|
gi|17534679|ref|NP_496192.1|
gi|32566163|ref|NP_508534.3|
gi|17568413|ref|NP_508535.1|
gi|19112946|ref|NP_596154.1|
gi|19112028|ref|NP_595236.1|
gi|169403947|ref|NP_001108586.1|
gi|52345806|ref|NP_001004949.1|
gi|46048961|ref|NP_989636.1|
gi|7669492|ref|NP_002037.2|
gi|55637711|ref|XP_508955.1|
gi|109095230|ref|XP_001105471.1|
gi|6679937|ref|NP_032110.1|
gi|149273202|ref|XP_001476757.1|
gi|8939418|ref|NP_058704.1|
gi|77404273|ref|NP_001029206.1|
gi|345807932|ref|XP_003435697.1|
gi|50978862|ref|NP_001003142.1|
gi|345798394|ref|XP_003434435.1|
gi|58391523|ref|XP_318655.2|
gi|442616493|ref|NP_001259584.1|
gi|22023983|ref|NP_525108.2|
```

```
| SIFDAKAGIALNDNFVKLVWYDNEWGYSNRVIDLIRHMAKTQ-----
| TIFDASAGIQLSPKFVKVVAWYDNEYGYSERVVDLVEHVA-----
| SIFDASAGIQLSPKFVKLISWYDNEYGY SARVVDLIEYVAKA-----
| SIFDASAGIQLSPKFVKLVSWYDNEYGYSTRVVDLVEHVAKA-----
| SIFDAAAGIQLSPKFVKLVSWYDNEYGYSTRVVDLVEHVAKA-----
| SIFDAQAGIALNDKFVKLVSWYDNEWGYSSRRVIDLVTYISKVDGKG-----
| SIFDAKAGISLNKNFVKLVSWYDNEWGYSSRRVLDLISYISKVDAKKA-----
| SIFDAGACISLNPNFVKLVSWYDNEYGYSNRVDLIGYIATRG-----
| SIFDTGACISLNPNFVKLVSWYDNEYGYSNRVDLIGYIATRG-----
| SIFDAGASISLNPHFVKLVSWYDNEFGYSNRVDLISYIATKA-----
| SIFDAGASISLNPHFVKLVSWYDNEFGYSNRVDLISYIATKA-----
| SIFDASAGIQLSPQFVKLVSWYDNEWGYSSRRVVDLVAYTAADKN-----
| SIFDASAGIQLSPQFVKLVSWYDNEWGYSHRVVDLVAYTASKD-----
| SIFDAGAGIALNDHFVKLVWYDNEFGYSNRVCDLMAHMASKE-----
| SIFDAGAGIALNDNFVKLVSWYDNECGYSHRVVDLMCHMASKE-----
| STFDAGAGIALNDHFVKLVSWYDNEFGYSNRVDLMVHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMAHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMAHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMAHMASKE-----
| STFDAGAGIALNDNFVKLISWYDNEYGYSNRVDLMAYMASKE-----
| STFDAGAGIALNDNFVKLISWYDNEYGYSNRVDLMAYMASKE-----
| STFDAGAGIALNDNFVKLISWYDNEYGYSNRVDLMAYMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMVHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMVHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNRVDLMVHMASKE-----
| STFDAGAGIALNDHFVKLISWYDNEFGYSNWVDLMVHMASKE-----
| SIFDAKAGIQLSDTFVKLISWYDNEYGYSNRVDLIKYMQTKD-----
| SVFDAKAGISLNDKFVKLISWYDNEFGYSNRVIDLIKYMQSKD-----
| SVFDAKAGISLNDKFVKLISWYDNEFGYSNRVIDLIKYMQSKD-----
|
| * * * * *
| * * * * *
```

4. Μπορούμε να αλλάξουμε το χρώμα των ακολουθιών με την επιλογή Show Colors :

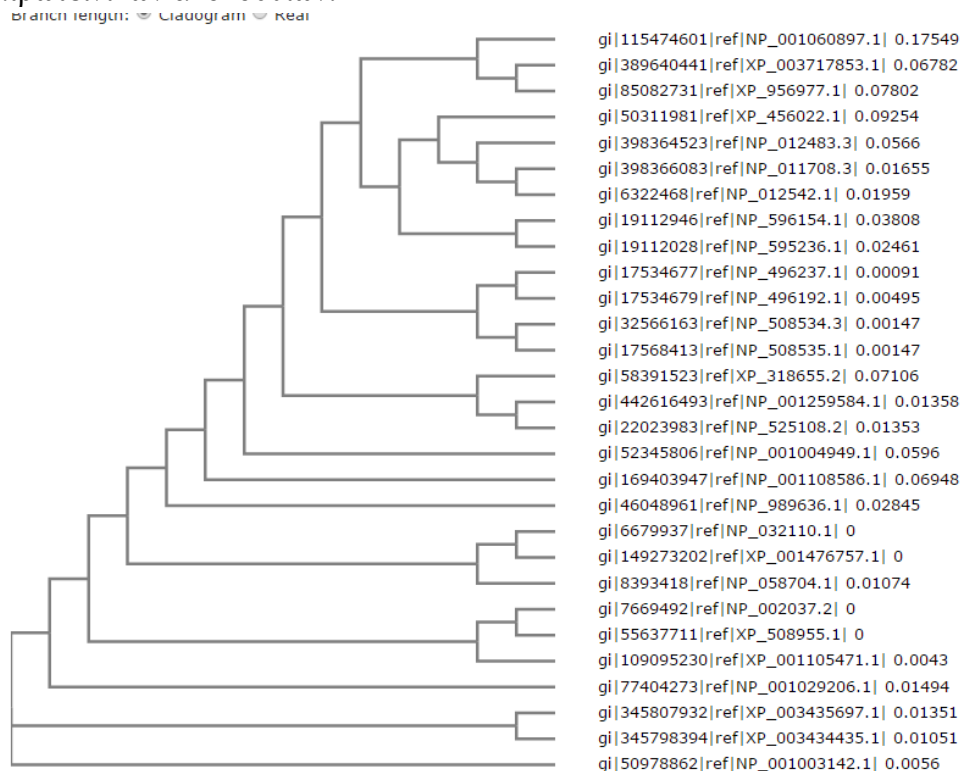
CLUSTAL O(1.2.1) multiple sequence alignment

```

gi|115474601|ref|NP_001060897.1|      MGKIKIGINGFGRIGRLVARVALQSEDEVLVAVNDPFITTDYMTYMFKYDTHVGQNKHSD
gi|50311981|ref|XP_456022.1|          --MKVVAINGFGRIGRLVRLIALQRKALEVVAVNDPFISVDYAAVMFYDSTHGRYKGEV
gi|398364523|ref|NP_012483.3|          --MIRIAINGFGRIGRLVRLIALQRKALIEVVAVNDPFISNDYAAVMFYDSTHGRYKGTV
gi|398366083|ref|NP_011708.3|          --MVRVAINGFGRIGRLVMRIALSRPNVEVVALNDPFITNDYAAVMFYDSTHGRYAGEV
gi|6322468|ref|NP_012542.1|           --MVRVAINGFGRIGRLVMRIALQRKNVEVVALNDPFISNDYSAYMFKYDSTHGRYAGEV
gi|389640441|ref|XP_003717853.1|        --MVKCGINGFGRIGRIVFRNAIEHPDCEIVAVNDPFIEPKYAKYMLEYDSTHGRFKGTV
gi|85082731|ref|XP_956977.1|           --MVVKVINGFGRIGRIVFRNAIEHDDIHIVAVNDPFIEPKYAAVMFYDSTHGRFKGTI
gi|17534677|ref|NP_496237.1|           MSKANVGINGFGRIGRLVRAAEKDTVQVAVNDPFITIDYMVYLFKYDSTHGRFKGTV
gi|17534679|ref|NP_496192.1|           MSKANVGINGFGRIGRLVRAAEKDTVQVAVNDPFITIDYMVYLFKYDSTHGRFKGTV
gi|32566163|ref|NP_508534.3|           MTKPSVINGFGRIGRLVRAAEKDSVNVAVNDPFISIDYMVYLFQYDSTHGRFKGTV
gi|17568413|ref|NP_508535.1|           MPKPSVINGFGRIGRLVRAAEKDSVNVAVNDPFISIDYMVYLFQYDSTHGRFKGTV
gi|19112946|ref|NP_596154.1|           MAIPKVINGFGRIGRIVLRNALVAKTIQVAVNDPFIDLEYMAYMFKYDSTHGRFDGSV
gi|19112028|ref|NP_595236.1|           MAIPKVINGFGRIGRIVLRNALITGKIQVAVNDPFIDLYMAYMFKYDSTHGRFEQSV
gi|169403947|ref|NP_001108586.1|        --MVVKVINGFGRIGRLVTRAAFLTKKVEIVAVNDPFIDLYMVMFYQYDSTHGRYKGEV
gi|52345806|ref|NP_001004949.1|         -----MAYMFKYDSTHGRFKGTV
gi|46048961|ref|NP_989636.1|           --MVKVGVNGFGRIGRLVTRAAVLGKQVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|7669492|ref|NP_002037.2|           MGKVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|55637711|ref|XP_508955.1|           MGKVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|109095230|ref|XP_001105471.1|        MGKVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|6679937|ref|NP_032110.1|           --MVKVGVNGFGRIGRLVTRAAICSGKVEIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|149273202|ref|XP_001476757.1|        --MVKVGVNGFGRIGRLVTRAAICSGKVEIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|8393418|ref|NP_058704.1|           --MVKVGVNGFGRIGRLVTRAAFCSDKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|77404273|ref|NP_001029206.1|        --MVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|345807932|ref|XP_003435697.1|        --MVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|50978862|ref|NP_001003142.1|        --MVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|345798394|ref|XP_003434435.1|        --MVKVGVNGFGRIGRLVTRAAFNSSGKVDIVAVNDPFIDLYMVMFYQYDSTHGRFKGTV
gi|58391523|ref|XP_318655.2|           --MSKIGINGFGRIGRLVRAAIDK--GASVAVNDPFIVDYMVYLFKYDSTHGRFKGEV
gi|442616493|ref|NP_001259584.1|        --MSKIGINGFGRIGRLVRAAIDK--GANVAVNDPFIDVNMVYLFKFDSTHGRFKGTV
gi|22023983|ref|NP_525108.2|           --MSKIGINGFGRIGRLVRAAIDK--GASVAVNDPFIDVNMVYLFKFDSTHGRFKGTV

```

5. Στην επιλογή Result Summary εμφανίζεται ένας πίνακας με τα SCORE που προκύπτουν από τη σύγκριση ανά δυο των πρωτεϊνικών ακολουθιών που εισάγαμε.
6. Στην επιλογή Phylogenetic Tree εμφανίζεται το κλαδικό δέντρο των πρωτεϊνικών ακολουθιών.



Κεφάλαιο 5

Γενετικοί Αλγόριθμοι

5.1 Εισαγωγή στους Γενετικούς Αλγορίθμους.

Οι Γενετικοί αλγόριθμοι αποτελούν μια μέθοδο εύρεσης και επίλυσης πολλών προβλημάτων τόσο στον τομέα της γενετικής επιστήμης όσο και σε προβλήματα της καθημερινής μας ζωής. Για πρώτη φορά οι γενετικοί αλγόριθμοι κάνουν την εμφάνισή τους στις αρχές του 1950 όταν μια ομάδα βιολόγων θέλησαν να λύσουν μέσω της χρήσης των υπολογιστών δύσκολες εφαρμογές βιολογικών συστημάτων. Η ανάπτυξη τους ήρθε τελικά το 1970 από τον Αμερικανό επιστήμονα John Henry Holland(*John Henry Holland,1970*) και τους συνεργάτες του[11].

5.2 Χαρακτηριστικά Γενετικού αλγορίθμου

- Εκτελούνται με κωδικοποίηση των τιμών του.
- Αναζητούν πληροφορίες σε πολλά σημεία και όχι μόνο σε ένα
- Χρήση της αντικειμενικής συνάρτησης
- Δουλεύουν κυρίως με πιθανότητες.

5.3 Λειτουργίες βασικού Γενετικού αλγορίθμου.

Σύμφωνα με τη θεωρία του Δαρβίνου ένας γενετικός αλγόριθμος εκτελεί 5 βασικές λειτουργίες καθεμία για τις οποίες θα μιλήσουμε αναλυτικά στην συγκεκριμένη ενότητα. Αυτές είναι:

- Κατασκευή των πιθανών λύσεων ή αλλιώς αρχικοποίηση του πληθυσμού στο χώρο αναζήτησης.
- Επανάληψη, όπου πραγματοποιεί τα τέσσερα επόμενα βήματα που είναι και τα πιο σημαντικά για τη δομή του αλγορίθμου.
- Εκτίμηση κάθε στοιχείου του πληθυσμού
- Χρήση του τελεστή επιλογής ενός νέου πληθυσμού.
- Χρήση του τελεστή διασταύρωσης για τη δημιουργία ενός νέου στοιχείου.
- Μετάλλαξη των στοιχείων του πληθυσμού μέσω του τελεστή μετάλλαξης.

Τέλος τερματίζουμε τον γενετικό αλγόριθμο. Συνεπώς και τα πέντε βήματα μας δείχνουν ότι εξερευνούν τον χώρο αναζήτησης με στόχο να δώσουν το καλύτερο αποτέλεσμα. Στη συνέχεια θα αναφέρουμε την έννοια των πιο πάνω βημάτων αναλυτικά με τη χρήση ενός παραδείγματος.

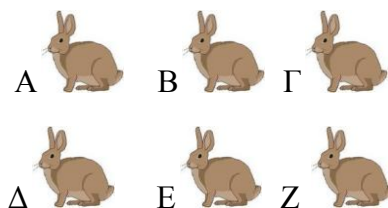
Παράδειγμα 6.

Έστω ότι θέλουμε να δημιουργήσουμε τον τέλειο λαγό. Αναφέρεται τα βήματα που θα καθορίσουν τη δημιουργία του σε έναν πληθυσμό που αποτελείται από έξι λαγούς. Σημειώνεται ότι κάθε λαγός έχει θετικά και αρνητικά χαρακτηριστικά. Η κωδικοποίηση και η συνάρτηση αξιολόγησης των λαγών φαίνεται παρακάτω:

A=010111, B=110011, Γ=000101, Δ=111011, E=011100, Z=001100

A=8, B=4, Γ=9, Δ=2, E=6, Z=1.

Επίσης για το βήμα της επιλογής δίνονται οι τυχαίοι αριθμοί $r=0.55$, $r=0.89$, $r=0.74$, $r=0.09$, $r=0.42$, $r=0.09$.



Εικόνα 4 Πληθυσμός 6 λαγών

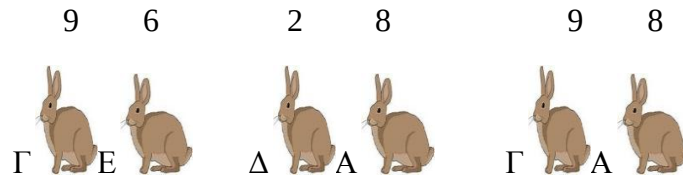
Αρχικοποίηση: Επιλέγει τα στοιχεία του πληθυσμού με τυχαίο τρόπο εφόσον προηγουμένως έχουν καθοριστεί οι παράμετροι που αναφέρουν πόσα είναι τα στοιχεία του πληθυσμού και συνήθως συμβολίζεται με `pop_size`(population size) και είναι σταθερό, καθώς και πως θα συμβολίσουμε τα στοιχεία του πληθυσμού στον υπολογιστή. Άρα στο συγκεκριμένο παράδειγμα το `pop_size` είναι 6. Γνωρίζοντας ότι κάθε λαγός έχει θετικά και αρνητικά χαρακτηριστικά, εμείς θα ασχοληθούμε με τη μελέτη δυο χαρακτηριστικών την ταχύτητά του και το μέγεθος των αυτιών του. Είναι σημαντικό να αναφέρουμε ότι στη θεωρία των γενετικών αλγορίθμων χρησιμοποιούνται κάποιες λέξεις κλειδιά που καθορίζουν τη λύση του αυτές είναι:

- φαινότυπος, ο χαρακτηρισμός ενός είδους χωρίς κωδικοποίηση της λύσης στον υπολογιστή.
- γονότυπος, η κωδικοποίηση της λύσης στον υπολογιστή συνηθίζεται να είναι δυαδική.
- γονίδιο, κάθε bit της κωδικοποίησης της λύσης.

Στη συνέχεια θα ασχοληθούμε με τα βήματα της επανάληψης.

Αξιολόγηση: Ο βαθμός που καθορίζει την καταλληλότητα του κάθε είδους. Δίνεται δηλαδή ένας αριθμός ή μια συνάρτηση που αναφέρει πόσο καλό είναι το κάθε είδος. Στο παράδειγμα με τους λαγούς μας δίνεται η συνάρτηση αξιολόγησης από 1-10. Αυτός που έχει την χειρότερη αξιολόγηση είναι ο Z γιατί έχει τον μικρότερο βαθμό $Z=1$. Αντίθετα ο Γ έχει την καλύτερη αξιολόγηση γιατί έχει τον μεγαλύτερο βαθμό $\Gamma=9$. Όμως ο Z μπορεί να έχει κάποιο χαρακτηριστικό που να μην το έχει ο Γ προκειμένου να γίνει ο τέλειος λαγός γι' αυτό και δεν απορρίπτουμε καμία λύση.

Επιλογή: Στο βήμα της επιλογής κατασκευάζουμε τον προσωρινό πληθυσμό έναν πληθυσμό που έχει κάποια χαρακτηριστικά. Επιλέγονται τα στοιχεία με μεγαλύτερη αξιολόγηση περισσότερες φορές από εκείνα που έχουν την μικρότερη. Δημιουργούμε δηλαδή έναν νέο πληθυσμό που έχει ίσο αριθμό στοιχείων με τον προηγούμενο αλλά έχουμε το δικαίωμα να επαναλάβουμε μερικά από τα στοιχεία στο νέο πληθυσμό καθώς και να χρησιμοποιήσουμε και κάποιο στοιχείο με χαμηλή βαθμολογία. Στο τρέχων παράδειγμα ο νέος πληθυσμός είναι:



Εικόνα 5 (Νέος πληθυσμός)

Οι αριθμοί (9,8,6,2) αναφέρονται στο βαθμό αξιολόγησης του κάθε στοιχείου. Η μέθοδος επιλογής των ίδιων στοιχείων πολλές φορές ονομάζεται εξαναγκασμένη ρουλέτα. Τα βήματα εκτέλεσης της εξαναγκασμένης ρουλέτας είναι τα ακόλουθα:

- Αρχικά αθροίζονται οι αξιολογήσεις όλων των στοιχείων του πληθυσμού, χρησιμοποιώντας τον παρακάτω τύπο:

$$F = \sum_{i=1}^{pop\ size} (1)$$

Τύπος (1) [Αθροιση αξιολογήσεων στοιχείων πληθυσμού]

- Βρίσκουμε πόσες φορές μπορούν να επιλεγθούν τα στοιχεία δηλαδή την πιθανότητα επιλογής κάθε στοιχείου η οποία μας δίνεται από τον τύπο:

$$p_i = \frac{eval(v_i)}{F} (2)$$

Τύπος (2) [Πιθανότητα κάθε στοιχείου]

όπου $eval(v_i)$ είναι η αξιολόγηση του κάθε στοιχείου και F η συνολική αξιολόγηση των στοιχείων.

- Τέλος υπολογίζουμε την αθροιστική πιθανότητα:

$$q_i = \sum_{j=1}^i p_j (3)$$

Τύπος (3) [Αθροιστική πιθανότητα]

Άρα τα βήματα της εξαναγκασμένης ρουλέτας στο παράδειγμα θα έχουν ως εξής:

1) Αθροισμα αξιολογήσεων των στοιχείων:

$$F = 8 + 4 + 9 + 2 + 6 + 1 = 30$$

2) Πιθανότητα επιλογής κάθε στοιχείου:

$$p_1 = \frac{eval(v_1)}{F} = \frac{8}{30} = 0.27$$

$$p_2 = \frac{eval(v_2)}{F} = \frac{4}{30} = 0.13$$

$$p_3 = \frac{eval(v_3)}{F} = \frac{9}{30} = 0.30$$

$$p_4 = \frac{\text{eval}(v_4)}{F} = \frac{2}{30} = 0.07$$

$$p_5 = \frac{\text{eval}(v_5)}{F} = \frac{6}{30} = 0.20$$

$$p_6 = \frac{\text{eval}(v_6)}{F} = \frac{1}{30} = 0.03$$

3) Αθροιστική πιθανότητα των στοιχείων

$$q_1 = p_1 = 0.27$$

$$q_2 = p_1 + p_2 = 0.27 + 0.13 = 0.40$$

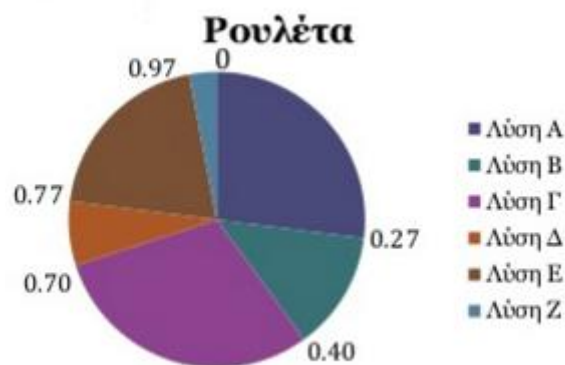
$$q_3 = p_1 + p_2 + p_3 = 0.27 + 0.13 + 0.30 = 0.70$$

$$q_4 = p_1 + p_2 + p_3 + p_4 = 0.27 + 0.13 + 0.30 + 0.07 = 0.77$$

$$q_5 = p_1 + p_2 + p_3 + p_4 + p_5 = 0.27 + 0.13 + 0.30 + 0.07 + 0.20 = 0.97$$

$$q_6 = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 0.27 + 0.13 + 0.30 + 0.07 + 0.20 + 0.03 = 1$$

Το άθροισμα στην τελευταία αθροιστική πιθανότητα θα πρέπει πάντα να είναι ίσο με 1. Στη συνέχεια κατασκευάζουμε τη ρουλέτα με τα αποτελέσματα των αθροιστικών πιθανοτήτων.



Εικόνα 5(Σχηματική απεικόνιση για τη ρουλέτα των αθροιστικών πιθανοτήτων)[12]

Χρησιμοποιώντας τους τυχαίους αριθμούς που μας δίνονται στην εκφώνηση παρατηρούμε ότι:

Για $r=0.55$ επιλέγεται το στοιχείο Γ

Για $r=0.89$ επιλέγεται το στοιχείο Ε

Για $r=0.74$ επιλέγεται το στοιχείο Δ

Για $r=0.09$ επιλέγεται το στοιχείο Α

Για $r=0.42$ επιλέγεται το στοιχείο Γ

Για $r=0.09$ επιλέγεται το στοιχείο Α

Άρα ο προσωρινός πληθυσμός θα είναι ο(Γ,Ε,Δ,Α,Γ,Α) ο οποίος θα διασταυρωθεί και θα μεταλλαχθεί όπως θα δούμε στη συνέχεια.

Διασταύρωση: Στο βήμα της διασταύρωσης επιλέγονται τυχαία τα στοιχεία του προσωρινού πληθυσμού και διασταυρώνονται με στόχο την παραγωγή νέων στοιχείων. Για να γίνει αυτό πρέπει να ληφθούν κάποιοι παράμετροι.

1. Ο αριθμός των διασταυρώσεων που θα γίνουν
2. Ποια στοιχεία του πληθυσμού θα λάβουν μέρος σε κάθε διασταύρωση.
3. Ποια μέθοδο θα επιλέξουμε για να γίνει η διασταύρωση(συνήθως επιλέγουμε τη μέθοδο ενός σημείου.)

Γενικά χρησιμοποιούνται αρκετές μέθοδοι διασταύρωσης, η πιο συχνή είναι η μέθοδος μονού σημείου . Στη μέθοδο αυτή χωρίζουμε τις δυο συμβολοσειρές σε ένα κοινό σημείο και παράγουμε τα νέα στοιχεία ως εξής:



Το στοιχείο A' που παράγεται είναι από το αριστερό τμήμα του A στοιχείου με το δεξί τμήμα του B στοιχείου. Ομοίως και το στοιχείο B' παράγεται από το αριστερό τμήμα του B στοιχείου με το δεξί τμήμα του A στοιχείου. Η προεργασία αυτή φαίνεται πιο αναλυτικά παρακάτω.

Προεργασία Διασταύρωσης Μονού Σημείου

- Αρχικά καθορίζεται το μήκος της συμβολοσειράς. Αν η συμβολοσειρά έχει μήκος n bits τότε τα πιθανά σημεία διασταύρωσης θα είναι $n-1$.
- Βρίσκω την πιθανότητα επιλογής των σημείων διασταύρωσης που είναι $1/(n-1)$.
- Επιλέγω το σημείο διασταύρωσης χρησιμοποιώντας κάποιους τυχαίους αριθμούς.

Στο παράδειγμα με τους λαγούς έχουμε ότι ο πληθυσμός τους είναι 6. Άρα το μήκος της συμβολοσειράς θα είναι 6 και τα πιθανά σημεία διασταύρωσης θα είναι $n-1=6-1=5$. Η πιθανότητα επιλογής σημείων διασταύρωσης είναι $1/5=0.2$. Το σημείο διασταύρωσης επιλέγεται τυχαία με βάση τους τυχαίους αριθμούς ανάμεσα στα παρακάτω διαστήματα.

Πρώτη επιλογή είναι οι θέσεις 1 έως 2 αν ο τυχαίος αριθμός βρίσκεται ανάμεσα στο διάστημα $[0.00,0.20]$.

Δεύτερη επιλογή είναι οι θέσεις 2 έως 3 αν ο τυχαίος αριθμός βρίσκεται ανάμεσα στο διάστημα $(0.20,0.40]$.

Τρίτη επιλογή είναι οι θέσεις 3 έως 4 αν ο τυχαίος αριθμός βρίσκεται ανάμεσα στο διάστημα $(0.40,0.60]$.

Τέταρτη επιλογή είναι οι θέσεις 4 έως 5 αν ο τυχαίος αριθμός βρίσκεται ανάμεσα στο διάστημα $(0.60,0.80]$.

Πέμπτη επιλογή είναι οι θέσεις 5 έως 6 αν ο τυχαίος αριθμός βρίσκεται ανάμεσα στο διάστημα(0.80,1.00].

Εφαρμογή Διασταύρωσης

Χωρίζω τον προσωρινό πληθυσμό σε ζεύγη (ΓΕ,ΔΑ,ΓΑ) . Επιλέγουμε έναν τυχαίο αριθμό(r) που μας δίνεται στην εκφώνηση. $A \leq p_c$ τότε το ζεύγος διασταυρώνεται. Αν είναι μεγαλύτερος τότε περνούν τα στοιχεία στον επόμενο πληθυσμό χωρίς διασταύρωση. Στο παράδειγμά μας το $p_c=0.60$. Η διασταύρωση γίνεται με τους τυχαίους αριθμούς 0.42, 0.65, 0.89, 0.21, 0.29.

Για το πρώτο ζεύγος (ΓΕ) έχουμε:

Συγκρίνουμε τον πρώτο τυχαίο αριθμό με την πιθανότητα διασταύρωσης δηλαδή $0.42 \leq 0.60$ άρα τα στοιχεία διασταυρώνονται. Η διασταύρωση επηρεάζεται από τον πρώτο τυχαίο αριθμό. Για τον αριθμό 0.65 παρατηρούμε ότι βρίσκεται μεταξύ των θέσεων 4-5 σύμφωνα πάντα με την προεργασία που κάναμε πριν. Επομένως έχουμε:

$\Gamma=000101$ $A'=000100$

$E=011100$ $B'=011101$

Για το δεύτερο ζεύγος (ΔΑ) έχουμε:

Ο τυχαίος αριθμός $0.89 > p_c$ άρα τα στοιχεία δεν διασταυρώνονται και παράγονται στον επόμενο πληθυσμό χωρίς καμία αλλαγή.

$\Delta=111011$ $\Gamma'=111011$

$A=010111$ $\Delta'=010111$

Για το τρίτο ζεύγος (ΓΑ) έχουμε:

Το $0.21 < p_c$ άρα τα στοιχεία διασταυρώνονται. Παρατηρούμε ότι ο τυχαίος αριθμός 0.29 βρίσκεται μεταξύ των θέσεων 2-3 έτσι η διασταύρωση έχει ως εξής:

$\Gamma=000101$ $E'=000111$

$A=010111$ $Z'=010101$

Στο παράδειγμα αυτό κάναμε εφαρμογή της διασταύρωσης μονού σημείου. Υπάρχουν και άλλες μέθοδοι διασταύρωσης όπως η διασταύρωση διπλού σημείου. Περιληπτικά αναφέρουμε τον τρόπο εφαρμογής της παρακάτω:

Αν έχω δυο στοιχεία A,B χρησιμοποιώντας τη μέθοδο διπλού σημείου τα χωρίζω σε τρία τμήματα και παίρνω για το στοιχείο A' το πρώτο αριστερό τμήμα του A , το δεύτερο τμήμα του στοιχείου B και το τρίτο δεξιό τμήμα του στοιχείου A. Το ίδιο θα κάνω και για το στοιχείο B'.



Μετάλλαξη: Η τροποποίηση κάποιων χαρακτηριστικών των στοιχείων του πληθυσμού. Η εφαρμογή του τελεστή μετάλλαξης καθορίζεται μέσω κάποιων

πιθανοκρατικών παραμέτρων , δηλαδή ο τρόπος που επιλέγονται τα στοιχεία του πληθυσμού και η πιθανότητα αλλαγής κάποιων χαρακτηριστικών. Ο τρόπος που γίνεται η μετάλλαξη για κάθε μέλος του πληθυσμού είναι ο εξής:

Για κάθε μέλος του πληθυσμού και για κάθε bit κυρίως από αριστερά προς τα δεξιά επιλέγω έναν τυχαίο αριθμό από 0 έως 1 και αν αυτός ο αριθμός είναι μικρότερος ή ίσος από την πιθανότητα μετάλλαξης(p_m) τότε το bit αντιστρέφεται γίνεται δηλαδή από 0→1 ή από 1→0. Το πλήθος των στοιχείων που επιλέγονται για μετάλλαξη είναι:

$$pop_size * m * p_m$$

όπου pop_size το μέγεθος του πληθυσμού, m το μήκος της συμβολοσειράς και p_m η πιθανότητα μετάλλαξης.

Αν στο παράδειγμα με τους λαγούς για $p_m=0.20$ χρησιμοποιήσω την εξής ακολουθία τυχαίων αριθμών: 0.77, 0.23, 0.12, 0.93, 0.28, 0.22, 0.15, 0.82, 0.34, 0.32, 0.44 τότε ο πίνακας μετάλλαξης θα είναι ο εξής:

	1 ^ο bit	2 ^ο bit	3 ^ο bit	4 ^ο bit	5 ^ο bit	6 ^ο bit	
A'=000100	0.77 0	0.23 0	0.12 <u>1</u>	0.93 1	0.28 0	0.22 0	A"=001100
B'=011101	0.15 <u>1</u>	0.82 1	0.34 1	0.32 1	0.44 0	0.77 1	B"=111101
Γ'=111011	0.23 1	0.12 <u>0</u>	0.93 1	0.28 0	0.22 1	0.15 <u>0</u>	Γ"=101010
Δ'=010111	0.82 0	0.34 1	0.32 0	0.44 1	0.77 1	0.23 1	Δ"=010111
E'=000111	0.12 <u>1</u>	0.93 0	0.28 0	0.22 1	0.15 <u>0</u>	0.82 1	Δ"=100101
Z'=010101	0.34 0	0.32 1	0.44 0	0.77 1	0.23 0	0.12 <u>0</u>	Z"=010100

Πίνακας 5(Πίνακας με τα στοιχεία της μετάλλαξης)

Τα A',B',Γ',Δ',E',Z' είναι στοιχεία που προκύπτουν από το βήμα της διασταύρωσης ενώ τα A",B",Γ",Δ",E",Z" από το βήμα της μετάλλαξης.

Συμπέρασμα:

Καταλήγουμε στο συμπέρασμα ότι με τη χρήση των πιο πάνω τελεστών δηλαδή του τελεστή επιλογής , διασταύρωσης και μετάλλαξης καταφέραμε από τη γενιά 0 των λαγών να παραχθεί η γενιά 1 όπως φαίνεται στον επόμενο πίνακα. Το αποτέλεσμα αυτού του αλγορίθμου είναι ότι η πιο πάνω διαδικασία θα επαναλαμβάνεται συνεχώς μέχρι τελικά να παραχθεί ο τέλειος λαγός.

ΓΕΝΙΑ 0	ΕΠΙΛΟΓΗ	ΔΙΑΣΤΑΥΡΩΣΗ	ΜΕΤΑΛΛΑΞΗ	ΓΕΝΙΑ 0
A=010111	Γ=000101	A'=000100	A''=001100	A=001100
B=110011	E=011100	B'=011101	B''=111101	B=111101
Γ=000101	Δ=111011	Γ'=111011	Γ''=101010	Γ=101010
Δ=111011	A=010111	Δ'=010111	Δ''=010111	Δ=010111
E=011100	Γ=000101	E'=000111	E''=100101	E=100101
Z=001100	A=010111	Z'=010101	Z''=010100	Z=01100

Πίνακας 6(Πίνακας με τα αποτελέσματα της χρήσης όλων των τελεστών)

Ολοκληρώνουμε την παραπάνω διαδικασία με το κριτήριο του τερματισμού του γενετικού αλγορίθμου. Προγραμματιστικά αυτό σημαίνει ότι ο γενετικός αλγόριθμος τερματίζει όταν έχει συμπληρωθεί ο μέγιστος αριθμός γενεών. Αν κατά τη διάρκεια της εκτέλεσης δεν βελτιωθεί η απόδοση των μελών του πληθυσμού τότε επιστρέφεται σε όλες τις γενιές το καλύτερο μέλος που πέρασε κατά την εκτέλεση του αλγορίθμου.

5.4 Θεμελιώσεις των Γενετικών Αλγορίθμων.

Τα θεμέλια των Γενετικών Αλγορίθμων αναφέρονται στη χρήση ενός τύπου ο οποίος κάνει μια πιθανή πρόβλεψη σχετικά με το πόσες συμβολοσειρές έχουν κάποια ομοιότητα θα έχουμε στο επόμενο βήμα εφόσον έχουν ολοκληρωθεί οι επαναληπτικές διαδικασίες της επιλογής, μετάλλαξης και διασταύρωσης. Στη διαδικασία της μετάλλαξης αναπαριστούσαμε τις συμβολοσειρές οι οποίες παρουσίαζαν μια αλλαγή με ένα σχήμα. Ο ορισμός του σχήματος δίνεται παρακάτω:

Ορισμός σχήματος :Έστω S το αλφάβητο των συμβόλων κυρίως των δυαδικών(0,1) που χρησιμοποιεί ο γενετικός αλγόριθμος για την κωδικοποίηση των χρωμοσωμάτων ενός πληθυσμού. Ένα σχήμα S θα είναι ένα χρωμόσωμα που θα περιέχει το σύμβολο (*) το οποίο μας επιτρέπει την αντικατάστασή του από άλλο σύμβολο του αλφαβήτου[20].

Για παράδειγμα αν έχω ένα σχήμα S=11*01 το αστεράκι που βρίσκεται στη θέση 3 μας δηλώνει ότι μπορεί να ταιριάζουν δύο συμβολοσειρές (11001) ή (11101). Ενώ αν έχω το σχήμα S=*1*1 θα ταιριάζουν τέσσερις συμβολοσειρές(0101,1111,0111,1101).

Παράμετροι σχήματος:

Αν έχω ένα χρωμόσωμα μήκους n που αναπαρίσταται με δυαδική συμβολοσειρά τότε θα ισχύει:

1. Το σχήμα χωρίς * θα αναπαριστά μια συμβολοσειρά.
2. Το σχήμα που έχει k * θα έχει 2^k συμβολοσειρές.
3. Αν το σχήμα έχει μόνο * τότε θα έχει 2^n συμβολοσειρές.

Για παράδειγμα στο σχήμα $S=01*10*01*$ ταιριάζουν $2^3=8$ συμβολοσειρές, ενώ στο σχήμα $S=*****$ ταιριάζουν $2^7=128$ συμβολοσειρές.

Τάξη και οριστικό μήκος σχήματος.

Τάξη ενός σχήματος $[o(S)]$ είναι ο αριθμός των θέσεων με 0 και 1 δηλαδή το σύνολο των θέσεων που δεν έχουν *. Αν έχω $S=01***11$ η τάξη αυτού του σχήματος είναι $o(S)=4$.

Οριστικό μήκος $\delta(S)$ ενός σχήματος είναι η απόσταση μεταξύ της τελευταίας και της πρώτης θέσης των συμβολοσειρών 0 και 1. Αν έχω $S=10*000*1$ τότε το σχήμα έχει οριστικό μήκος $\delta(S)=8-1=7$.

Η αρίθμηση των θέσεων ξεκινάει από αριστερά προς τα δεξιά.

Συνολική και μέση απόδοση του πληθυσμού.

Για μια χρονική στιγμή t της εκτέλεσης ενός αλγορίθμου η συνολική απόδοση ορίζεται ως το άθροισμα όλων των αξιολογήσεων του πληθυσμού ενώ η μέση απόδοση ορίζεται η συνολική απόδοση προς το μέγεθος του πληθυσμού.

Τύπος συνολικής απόδοσης:

$$F(t) = \sum_{i=1}^N f(v_i)$$

Τύπος μέσης απόδοσης-

$$\overline{F(t)} = \frac{F(t)}{N}$$

Όπου N το μέγεθος του πληθυσμού και $f(v)$ η αντικειμενική συνάρτηση.

Εξέλιξη σχήματος μέσω των βημάτων του γενετικού αλγορίθμου.

Έστω ένα σχήμα S και p το πλήθος των χρωμοσωμάτων την χρονική στιγμή t τότε ορίζουμε μέση απόδοση του σχήματος τον μέσο όρο των αποδόσεων των χρωμοσωμάτων που ταιριάζουν με το σχήμα την τρέχουσα χρονική στιγμή.

$$\text{eval}(S,t) = \frac{1}{p} \sum_{j=1}^p \text{eval}(v_j) \quad [20]$$

Η πιθανότητα επιλογής μιας συμβολοσειράς που ταιριάζει στο σχήμα S είναι:

$$\frac{\text{eval}(S,t)}{F(t)}$$

Πρέπει να αποδείξουμε ότι το πλήθος των χρωμοσωμάτων που ταιριάζουν την χρονική στιγμή (t+1) με το σχήμα S δίνεται από τον τύπο:

$$\xi(S,t+1) \geq \xi(S,t) * \frac{eval(S,t)}{F(t)} * [1 - p_c \frac{\delta(S)}{m-1} - o(S) * p_m]$$

Η απόδειξη του παραπάνω τύπου θα πραγματοποιηθεί με τη βοήθεια των τριών γενετικών τελεστών (επιλογής, διασταύρωσης και μετάλλαξης).

Διαδικασία επιλογής

Με τη διαδικασία της επιλογής μας δίνεται ο τύπος:

$$\xi(S,t+1) = \xi(S,t) * N * \frac{eval(S,t)}{F(t)}$$

Το $\xi(S,t+1)$ είναι ο εκτιμώμενος αριθμός συμβολοσειρών

Το $eval(S,t)$ δηλώνει τη μέση απόδοση των συμβολοσειρών που έχει το σχήμα

Το $\xi(S,t)$ είναι ο αριθμός των συμβολοσειρών του σχήματος S

Το $F(t)$ είναι η συνολική απόδοση των συμβολοσειρών.

Το N αναφέρεται στο πόσες φορές θα γυριστεί η ρουλέτα. Επομένως ο τύπος καθορίζει πόσα χρωμοσώματα θα επιλεγθούν από τη ρουλέτα τα οποία θα ταιριάζουν στο σχήμα S.

Στον παραπάνω τύπο μπορώ να γράψω ότι το:

$$N * \frac{eval(S,t)}{F(t)} = \frac{eval(S,t)}{F(t)/N} = \overline{F(t)}$$

γιατί το $\frac{F(t)}{N}$ είναι ίσο με την μέση απόδοση όλου του πληθυσμού.

Συνεπώς αν $\frac{eval(S,t)}{\overline{F(t)}} > 1$ τότε το πλήθος των συμβολοσειρών στην επόμενη γενιά θα

αυξάνεται ενώ αν $\frac{eval(S,t)}{\overline{F(t)}} < 1$ τότε το πλήθος των συμβολοσειρών στην επόμενη

γενιά θα μειώνεται.

Διαδικασία διασταύρωσης.

Θα ορίσουμε τις πιθανότητες επιβίωσης ή καταστροφής ενός σχήματος κατά τη διασταύρωση μονού σημείου για συμβολοσειρές που έχουν m χρωμοσώματα συνήθως συμβολίζονται με bits.

Η πιθανότητα καταστροφής ενός σχήματος δίνεται από τον τύπο:

$$p_D(S) = p_C \frac{\delta(S)}{m-1}$$

όπου m-1 τα σημεία διαχωρισμού

p_C η πιθανότητα διασταύρωσης

$\delta(S)$ το οριστικό μήκος σχήματος.

Πιθανότητα επιβίωσης ενός σχήματος

Η πιθανότητα επιβίωσης ενός σχήματος εξαρτάται από την πιθανότητα καταστροφής και ορίζεται από τον παρακάτω τύπο:

$$p_s(S) = 1 - p_D(S) = 1 - p_C \frac{\delta(S)}{m-1}$$

Με τη χρήση των παραπάνω τύπων αποδεικνύεται ότι ο αριθμός των συμβολοσειρών μετά την διαδικασία της διασταύρωσης θα είναι:

$$\xi(S, t+1) \geq \xi(S, t) * \frac{eval(S, t)}{F(t)} * p_s(S)$$

Αν τον παραπάνω τύπο τον πολλαπλασιάσω με την πιθανότητα επιβίωσης $p_s(S)$ τότε θα έχω:

$$\xi(S, t+1) \geq \xi(S, t) * \frac{eval(S, t)}{F(t)} * [1 - p_C \frac{\delta(S)}{m-1}]$$

Το σύμβολο $\geq$ δηλώνει ότι κατά την διαδικασία της διασταύρωσης

1. έχω δυο συμβολοσειρές
2. Αν έχω μια συμβολοσειρά που ταιριάζει στο σχήμα και τη διασταυρώσουμε σε κάποιο σημείο στη μέση τότε υπάρχει η περίπτωση να επιβιώσει το σχήμα.

Πιθανότητα μετάλλαξης

Αν έχω μια συμβολοσειρά π.χ. $S = 1^{**}01^{*}$ και κάνω μετάλλαξη σε ένα αδιάφορο σύμβολο(*) τότε παρατηρώ ότι δεν καταστρέφεται η συμβολοσειρά. Η περίπτωση που θα καταστραφεί η συμβολοσειρά είναι αν κάνω μετάλλαξη σε ένα σταθερό σύμβολο(0 ή 1).

Η πιθανότητα να επιβιώσει ένα σχήμα στη διαδικασία της μετάλλαξης καθορίζεται από τον παρακάτω τύπο:

$$p_s(S) = 1 - [p_m * o(S)]$$

όπου $o(S)$ η τάξη του σχήματος.

Άρα ο αριθμός των συμβολοσειρών μετά την διαδικασία της μετάλλαξης θα είναι:

$$\xi(S, t+1) \geq \xi(S, t) * \frac{eval(S, t)}{F(t)} * [1 - p_C \frac{\delta(S)}{m-1} - o(S) * p_m]$$

Ο παραπάνω τύπος είναι γνωστός και ως θεώρημα των σχημάτων.

Επικράτηση-Εξαφάνιση σχημάτων ενός πληθυσμού

Με τον όρο επικράτηση- εξαφάνιση αναφερόμαστε σε ποια σχήματα θα επικρατήσουν ή θα εξαφανιστούν από τον πληθυσμό.

Για παράδειγμα θέλουμε να μελετήσουμε την εξέλιξη ενός σχήματος $S=01*****$ μετά από κάποιες γενιές. Υπάρχουν δυο περιπτώσεις. Η πρώτη είναι να εξαφανιστούν οι συμβολοσειρές που αρχίζουν με (0,1) και η δεύτερη να επικρατήσουν αν υπάρχει καλή απόδοση.

Θα αναφέρουμε την επικράτηση - εξαφάνιση σχημάτων για ένα πλήθος συμβολοσειρών που ταιριάζουν σε ένα σχήμα S μετά από K γενιές, μόνο στη διαδικασία της επιλογής με τη χρήση του παρακάτω τύπου.

$$\xi(S,t+K)=\xi(S,t)*(1+\varepsilon)^K$$

όπου $\xi(S,t+K)$ είναι το αναμενόμενο πλήθος συμβολοσειρών που ταιριάζουν στο σχήμα S μετά από K γενιές.

ε η ει τοις εκατό απόκλιση της μέσης απόδοσης του πληθυσμού σε σχέση με την μέση απόδοση του σχήματος η οποία δίνεται από τον τύπο:

$$\varepsilon = \frac{\overline{eval(S)} - \overline{F(t)}}{\overline{F(t)}}$$

$\overline{eval(S)}$ η μέση απόδοση του σχήματος και $\overline{F}$ η μέση απόδοση του πληθυσμού.

Παράδειγμα

Έστω η αντικειμενική συνάρτηση f , που όταν δέχεται εόσοδο τη δυαδική συμβολοσειρά x , με μήκος $\lambda=4$, να δίνει αποτέλεσμα που ισούται με τον ακέραιο αριθμό που αναπαριστάται με τον δυαδικό x (π.χ $f(0011)=3$, $f(1111)=15$)

1. Ποια είναι η μέση απόδοση του σχήματος $S_1=1***$, σε σχέση με την αντικειμενική συνάρτηση f ; Ποια είναι η αντικειμενική συνάρτηση του σχήματος $S_2=0***$ σε σχέση με τη συνάρτηση f ; Να υποθέσετε ότι ο πληθυσμός αποτελείται 16 χρωμοσώματα, ένα για κάθε διαφορετικό δυαδικό αριθμό από το 0000 έως το 1111.
2. Ποιο από τα δύο σχήματα θα επικρατήσει σε ολόκληρο τον πληθυσμό μετά από κάποιες γενιές και ποιο θα εξαφανιστεί από τον πληθυσμό; Να υπολογίσετε μετά από πόσες γενιές θα συμβεί αυτό, για κάθε περίπτωση, αν θεωρήσουμε την επίδραση της διασταύρωσης και της μετάλλαξης είναι αμελητέα.

Απάντηση

1) Στο σχήμα $S_1=1^{***}$ ταιριάζουν οι παρακάτω συμβολοσειρές:

$$1000 \rightarrow \text{eval}(1000)=8$$

$$1001 \rightarrow \text{eval}(1001)=9$$

$$1011 \rightarrow \text{eval}(1011)=11$$

$$1010 \rightarrow \text{eval}(1010)=10$$

$$1100 \rightarrow \text{eval}(1100)=12$$

$$1101 \rightarrow \text{eval}(1101)=13$$

$$1110 \rightarrow \text{eval}(1110)=14$$

$$1111 \rightarrow \text{eval}(1111)=15$$

Τα αποτελέσματα 8,9,11 κτλ προκύπτουν μετρώντας τη θέση που βρίσκονται μόνο οι δυαδικοί αριθμοί (1).

π.χ. για την συμβολοσειρά 1000 έχουμε :

$$\begin{array}{cccc} 1 & 0 & 0 & 0 \\ | & | & | & | \\ 2^3 & 2^2 & 2^1 & 2^0 \end{array}$$

$$\text{δηλαδή } \text{eval}(1000)=2^3=8$$

Άρα η μέση απόδοση του σχήματος S_1 θα είναι ίση με το άθροισμα των καταλληλοτήτων προς το πλήθος των συμβολοσειρών:

$$\bar{F}=(8+9+11+10+12+13+14+18)/8=92/8=11.5$$

Για το σχήμα $S_2=0^{***}$ ταιριάζουν οι παρακάτω συμβολοσειρές:

$$0000 \rightarrow \text{eval}(0000)=0$$

$$0100 \rightarrow \text{eval}(0100)=4$$

$$0110 \rightarrow \text{eval}(0110)=6$$

$$0111 \rightarrow \text{eval}(0111)=7$$

$$0011 \rightarrow \text{eval}(0011)=3$$

$$0001 \rightarrow \text{eval}(0001) = 1$$

$$0101 \rightarrow \text{eval}(0101) = 5$$

$$0010 \rightarrow \text{eval}(0010) = 2$$

Άρα η μέση απόδοση του σχήματος S_2 θα είναι ίση με :

$$\bar{F} = (0+4+6+7+3+1+5+2)/8 = 28/8 = 3,5.$$

2) Παρατηρούμε ότι η μέση απόδοση του σχήματος S_1 είναι καλύτερη από τη μέση απόδοση του σχήματος S_2 . Άρα οι συμβολοσειρές του σχήματος S_1 θα επικρατήσουν στον πληθυσμό.

Η συνολική απόδοση του πληθυσμού είναι το συνολικό άθροισμα των καταλληλοτήτων για τα σχήματα S_1 και S_2 δηλαδή 120. Επομένως η συνολική μέση απόδοση και για τα δυο σχήματα είναι $\bar{F} = 120/16 = 7,5$.

Η % απόκλιση της μέσης απόδοσης του πληθυσμού είναι :

$$\varepsilon = (\text{eval}(S_1)/\bar{F}) - 1 = (11,5/7,5) - 1 = 1,53 - 1 = 0,53$$

Χρησιμοποιώντας την εξίσωση $\xi(S, t+k) = \xi^*(1+\varepsilon)^k$ έχουμε:

$$\text{Για } t=0 \text{ έχουμε } \xi(S, 0) = 8, \text{ άρα πρέπει να υπολογίσουμε το } \xi(S_1, t+k) = 8^*(1+\varepsilon)^k$$

Γνωρίζοντας ότι όλος ο πληθυσμός έχει 16 άτομα και τα 8 από αυτά είναι συμβολοσειρές που ταιριάζουν στο σχήμα S_1 επιδιώκουμε και οι 16 συμβολοσειρές να ταιριάζουν στο σχήμα S_1 .

Άρα για $\varepsilon = 0,53$ και $\xi(S, t+k) = 16$ έχουμε:

$$16 = 8^*(1+0,53)^k = 8*1,53^k \rightarrow 2 = 1,53^k$$

Εισάγοντας λογάριθμο και στα δυο μέρη έχω:

$\log 2 = k * \log(1,53) \rightarrow k = 0,301/0,185 = 1,62$ που στρογγυλοποιώντας το αποτέλεσμα διαπιστώνουμε ότι το σχήμα S_1 θα επικρατήσει στον πληθυσμό μετά από 2 γενιές.

Για το σχήμα S_2 το $\varepsilon = (\text{eval}(S_2)/\bar{F}) - 1 = (3,5/7,5) - 1 = -0,53$ άρα θα τείνει να μικρύνει.

$$\xi(S, t+k) < 1 \rightarrow 1 > 8^*(1-0,53)^k \rightarrow 0 > 8*0,47^k \rightarrow 0 > \log 8 + k * \log(0,47) \rightarrow 0 > 0,903 + k(-0,327) \rightarrow -0,327k < -0,903 \rightarrow k > 2,73 \text{ άρα 3 γενιές.}$$

Μετά από 3 γενιές το σχήμα S_2 θα εξαφανιστεί από τον πληθυσμό.

5.5 Ασκήσεις Κατανόησης γενετικού αλγορίθμου.

Άσκηση 1

Δίνεται ο ακόλουθος πληθυσμός συμβολοσειρών με τις αντίστοιχες καταλληλότητες, στη γενιά 0.

	Χρωμόσωμα	Καταλληλότητα
A	011101	5
B	101001	1
C	111001	4
D	010000	2
E	011001	4
F	111101	4

Θεωρήστε ότι:

1. Ο γενετικός αλγόριθμος χρησιμοποιεί τελεστή επιλογής που βασίζεται στην τιμή της αντικειμενικής συνάρτησης για κάθε άτομο(κάθε άτομο επιλέγεται ανάλογα με την απόδοσή του) και εκτελεί πλήρη αντικατάσταση των ατόμων κάθε γενιάς(κανένας γονέας δεν αντιγράφεται απευθείας στην επόμενη γενιά)
2. Η επιλογή των ατόμων που θα συμμετέχουν στη διασταύρωση γίνεται με βάση τον τελεστή επιλογής και τους τυχαίους αριθμούς που προέκυψαν από τη γεννήτρια τυχαίων αριθμών. Ο τελεστής διασταύρωσης είναι μονού σημείου με πιθανότητα διασταύρωσης 1.0 με το σημείο διασταύρωσης να επιλέγεται τυχαία.
3. Η πιθανότητα μετάλλαξης είναι ίση με 0
4. Τα δυο παιδιά που προκύπτουν από μια διασταύρωση προστίθενται στον πληθυσμό της επόμενης γενιάς. Δίνεται και η ακόλουθη λίστα τυχαίων αριθμών που έχει παραχθεί με χρήση μιας γεννήτριας τυχαίων αριθμών: 0.12, 0.63, 0.20, 0.35, 0.48, 0.93, 0.51, 0.46, 0.28, 0.15, 0.52, 0.81, 0.65, 0.25, 0.73.

Ζητείται:

α) Η ρουλέτα για την επιλογή ατόμων από τη γενιά 0. Να φαίνεται το ποσοστό που αντιστοιχεί σε κάθε άτομο του αρχικού πληθυσμού και ο προσωρινός πληθυσμός.

β) Να εφαρμόσετε την διασταύρωση των ζευγών του πληθυσμού που προέκυψαν κατά τη γενιά 1.

Απάντηση:

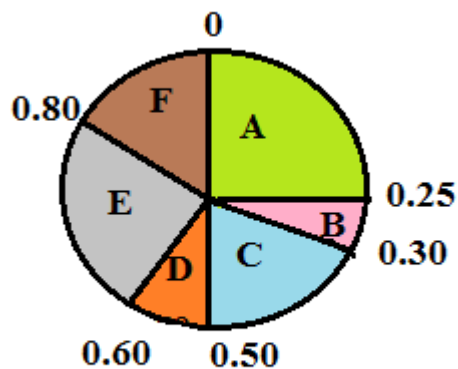
α) Για το ερώτημα α υπολογίζω τα εξής:

- Υπολογίζω την συνολική απόδοση του πληθυσμού που είναι το άθροισμα των καταλληλοτήτων των ατόμων. $F=5+1+4+2+4+4=20$
- Υπολογίζω τις πιθανότητες επιλογής των μελών του πληθυσμού.
 $p(A)=5/20=0.25$, $p(B)=1/20=0.05$, $p(C)=4/20=0.20$,

$$p(D)=2/20=0.1, p(E)=4/20=0.20, p(F)=4/20=0.20$$

- Βρίσκω τις αθροιστικές πιθανότητες.
 $q(A)=p(A)=0.25,$
 $q(B)=p(A)+p(B)=0.25+0.05=0.30$
 $q(C)=q(B)+0.20=0.30+0.20=0.50$
 $q(D)=q(C)+0.1=0.50+0.1=0.60$
 $q(E)=q(D)+0.2=0.60+0.2=0.80$
 $q(D)=q(E)+0.2=0.80+0.2=1.00$

Φτιάχνω τη ρουλέτα όπως φαίνεται παρακάτω:



- Επιλογή με βάση τους 6 πρώτους τυχαίους αριθμούς.
 Οι έξι πρώτοι τυχαίοι αριθμοί είναι: 0.12, 0.63, 0.20, 0.35, 0.48, 0.93) και σύμφωνα με αυτούς επιλέγεται ο παρακάτω προσωρινός πληθυσμός.
 0.12 → A
 0.63 → E
 0.20 → A
 0.35 → C
 0.48 → C
 0.93 → F

Άρα ο προσωρινός πληθυσμός είναι ο (A,E,A,C,C,F)

β) Δεδομένου ότι η πιθανότητα διασταύρωσης είναι ίση με 1 $p_c=1$ παρατηρούμε ότι θα διασταυρωθούν όλα τα ζεύγη χωρίς επιλογή τυχαίου αριθμού. Η προεργασία της διασταύρωσης είναι η εξής:

Βρίσκω τα ισοπίθανα σημεία διαχωρισμού. Το μήκος της συμβολοσειράς μου είναι $n=6$ άρα τα σημεία διαχωρισμού θα είναι $n-1=5$. Το $1/5=0.20$ άρα ο τυχαίος αριθμός θα είναι μεταξύ των παρακάτω θέσεων:

Για το διάστημα $[0.0, 0.20]$ επιλέγεται μεταξύ των θέσεων 1-2

Για το διάστημα $(0.20, 0.40]$ επιλέγεται μεταξύ των θέσεων 2-3

Για το διάστημα $(0.40, 0.60]$ επιλέγεται μεταξύ των θέσεων 3-4

Για το διάστημα $(0.60, 0.80]$ επιλέγεται μεταξύ των θέσεων 4-5

Για το διάστημα $(0.80, 1.00]$ επιλέγεται μεταξύ των θέσεων 5-6

Πρώτο ζεύγος(A,E):

Ο τυχαίος αριθμός είναι ο 0.51 και βρίσκεται μεταξύ των θέσεων 3-4 δηλαδή ο επόμενος από τους τυχαίους αριθμούς που χρησιμοποίησα πιο πάνω.

$A=011101 \rightarrow A'=011001$

$B=011001 \rightarrow B'=011101$

Δεύτερο ζεύγος(A,C):

Ο τυχαίος αριθμός είναι το 0.46 άρα πάλι μεταξύ των θέσεων 3-4

$A=011101 \rightarrow C'=011001$

$C=111001 \rightarrow D'=111101$

Τρίτο ζεύγος (C,F):

Ο τυχαίος αριθμός είναι το 0.28 άρα μεταξύ των θέσεων 2-3

$C=111001 \rightarrow E'=111101$

$F=111101 \rightarrow F'=111001$

Άρα τα μέλη της διασταύρωσης είναι:

$A'=011001, B'=011101, C'=011001, D'=111101, E'=111101, F'=111001.$

Μετάλλαξη δεν γίνεται γιατί η πιθανότητα μετάλλαξης είναι 0.

Άσκηση 2

Θεωρήστε τρεις δυαδικές συμβολοσειρές $A_1=11101111$, $A_2=00010100$, $A_3=01000011$ και έξι σχήματα:

$H1=1*****$, $H2=0*****$, $H3=*****11$, $H4=***0*00*$,

$H5=1*****1*$, $H6=1110**1*$.

1. Ποιές συμβολοσειρές ταιριάζουν σε κάθε σχήμα;
2. Βρείτε την τάξη και το οριστικό μήκος κάθε σχήματος,
3. Υπολογίστε την πιθανότητα επιβίωσης κάθε σχήματος αν γίνεται διασταύρωση με πιθανότητα διασταύρωσης $p_c=0.85$
4. Υπολογίστε την πιθανότητα επιβίωσης κάθε σχήματος αν γίνεται μετάλλαξη με πιθανότητα μετάλλαξης $p_m=0.001$.

Απάντηση

1) Στο σχήμα 1 ταιριάζει η συμβολοσειρά A_1 γιατί αρχίζει με 1.

Στο σχήμα 2 ταιριάζουν οι A_2 και A_3 .

Στο σχήμα 3 ταιριάζουν οι A_1 και A_3 .

Στο σχήμα 4 δεν ταιριάζει καμία συμβολοσειρά.

Στο σχήμα 5 ταιριάζει η A_1

Στο σχήμα 6 ταιριάζει η A_1

2) Η τάξη του σχήματος υπολογίζεται από τα σταθερά σύμβολα έχω σε κάθε σχήμα.

Το οριστικό μήκος υπολογίζεται από τη απόσταση του δεξιότερου και αριστερότερου συμβόλου.

Στο πρώτο σχήμα έχω 1 σταθερό σύμβολο και οριστικό μήκος $\delta(H1)=1-1=0$.

Στο δεύτερο σχήμα έχω 1 σταθερό σύμβολο και οριστικό μήκος $\delta(H2)=1-1=0$.

Στο τρίτο σχήμα έχω 2 σταθερά σύμβολα και οριστικό μήκος $\delta(H3)=8-7=1$.

Στο τέταρτο σχήμα έχω 3 σταθερά σύμβολα και οριστικό μήκος $\delta(H4)=7-4=3$.

Στο πέμπτο σχήμα έχω 2 σταθερά σύμβολα και οριστικό μήκος $\delta(H5)=7-1=6$.

Στο έκτο σχήμα έχω 5 σταθερά σύμβολα και οριστικό μήκος $\delta(H6)=7-1=6$.

3) Η πιθανότητα επιβίωσης κάθε σχήματος μας δίνεται από τον τύπο:

$p_s(S) = 1 - p_c \frac{\delta(S)}{m-1}$ όπου $\delta(S)$ το οριστικό μήκος, p_c η πιθανότητα διασταύρωσης και m το μήκος συμβολοσειράς.

$$p_s(H1) = 1 - p_c \frac{\delta(H1)}{m-1} = 1 - 0.85 \frac{0}{8-1} = 1$$

$$p_s(H2) = 1 - p_c \frac{\delta(H2)}{m-1} = 1 - 0.85 \frac{0}{8-1} = 1$$

$$p_s(H3) = 1 - p_c \frac{\delta(H3)}{m-1} = 1 - 0.85 \frac{1}{8-1} = 1 - 0.85 \frac{1}{7} = 0.879$$

$$p_s(H4) = 1 - p_c \frac{\delta(H4)}{m-1} = 1 - 0.85 \frac{3}{8-1} = 1 - 0.85 \frac{3}{7} = 0.6345$$

$$p_s(H5) = 1 - p_c \frac{\delta(H5)}{m-1} = 1 - 0.85 \frac{6}{8-1} = 1 - 0.85 \frac{6}{7} = 0.272$$

$$p_s(H6) = 1 - p_c \frac{\delta(H6)}{m-1} = 1 - 0.85 \frac{6}{8-1} = 1 - 0.85 \frac{6}{7} = 0.272$$

4)) Η πιθανότητα επιβίωσης κάθε σχήματος αν γίνει μετάλλαξη μας δίνεται από τον τύπο:

$$p_s(S) = 1 - [p_m * o(S)]$$

όπου p_m η πιθανότητα μετάλλαξης και $o(S)$ τάξη σχήματος.

$$p_s(H1)=1-[p_m*o(S)]=1-(0,001*1)=1-0,001=0,999$$

$$p_s(H2)=1-[p_m*o(S)]=1-(0,001*1)=1-0,001=0,999$$

$$p_s(H3)=1-[p_m*o(S)]=1-(0,001*2)=1-0,002=0,998$$

$$p_s(H4)=1-[p_m*o(S)]=1-(0,001*3)=1-0,003=0,997$$

$$p_s(H5)=1-[p_m*o(S)]=1-(0,001*2)=1-0,002=0,998$$

$$p_s(H6)=1-[p_m*o(S)]=1-(0,001*5)=1-0,005=0,995$$

Σχήματα	Συμβολοσειρές που ταιριάζουν	Τάξη ο(s)	Οριστικό μήκος	Πιθαν. Επιβίωσης διασταύρωσης	Πιθαν. Επιβίωσης μετάλλαξης
H1=1*****	A ₁	1	0	1	0,999
H2=0*****	A ₂ , A ₃	1	0	1	0,999
H3=*****11	A ₁ , A ₃	2	1	0,879	0,998
H4=***0*00*	Κανένα	3	3	0,6345	0,997
H5=1*****1*	A ₁	2	6	0,272	0,998
H6=1110**1*	A ₁	5	6	0,272	0,995

Πίνακας αποτελεσμάτων

Άσκηση 3

Έστω ότι πληθυσμός που δίνεται στον παρακάτω πίνακα είναι ο αρχικός πληθυσμός ενός Γενετικού Αλγορίθμου που εφαρμόζεται για τη μεγιστοποίηση της συνάρτησης $f(x_1, x_2)=x_2^2 - x_1^2$ όπου τα x_1, x_2 είναι ακέραιοι αριθμοί στο διάστημα $[0, 7]$. Για την αναπαράσταση κάθε ανεξάρτητης μεταβλητής χρησιμοποιούνται 3 bits οπότε για την αναπαράσταση κάθε χρωμοσώματος του πληθυσμού χρησιμοποιούνται 6 bits με τα τρία αριστερά να αναπαριστούν τη μεταβλητή x_1 και τα τρία δεξιά τη μεταβλητή x_2 . Για τα ζητήματα που ακολουθούν κάντε τις πράξεις με ακρίβεια 4 δεκαδικών ψηφίων.

	Χρωμόσωμα	X ₁	X ₂	Απόδοση Ικανότητα	Νέα ικανότητα
A	101011				
B	111000				
Γ	111011				
Δ	101000				
E	111001				

1) Να υπολογίσετε την απόδοση κάθε μέλους του πληθυσμού. Επειδή η καταλληλότητα- ικανότητα έχει αρνητικές τιμές, δεν μπορούν να υπολογιστούν οι

πιθανότητες επιλογής. Πως μπορεί να αντιμετωπιστεί αυτό το πρόβλημα. Αιτιολογήστε την απάντησή σας και συμπληρώστε την κενή στήλη του πίνακα, με την νέα ικανότητα. Ποιο πρόβλημα υπάρχει με τον υπολογισμό της πιθανότητας επιλογής του 2^{ου} ατόμου; Πως μπορεί να αντιμετωπιστεί;

2) Να υπολογίσετε την πιθανότητα επιλογής κάθε χρωμοσώματος χρησιμοποιώντας επιλογή εξαναγκασμένης ρουλέτας.

3) Να υπολογίσετε τον αναμενόμενο αριθμό αντιγράφων κάθε χρωμοσώματος στην επόμενη γενιά.

4) Αν η πρώτη γενιά που προκύπτει μετά την εφαρμογή των γενετικών τελεστών της επιλογής, διασταύρωσης και μετάλλαξης είναι αυτή που απεικονίζεται στον επόμενο πίνακα έχει βελτιωθεί η μέση απόδοση του πληθυσμού στη γενιά 1 σε σχέση με τη γενιά 0; Πόσο έχει μεταβληθεί η απόδοση του καλύτερου μέλους του πληθυσμού στη γενιά 1 σε σχέση με τη γενιά 0;

	Άτομο-Χρωμόσωμα
A	111 100
B	110 110
Γ	111 000
Δ	110 011
E	111 011

5) Ποιος θα είναι ο αναμενόμενος αριθμός ατόμων σύμφωνα με το θεώρημα σχημάτων, που θα αντιστοιχούν στο σχήμα $S=111***$ στη γενιά 2 αν θεωρήσουμε ότι εφαρμόζεται διασταύρωση με πιθανότητα $p_c=0,5$ και μετάλλαξη με πιθανότητα $p_m=0,1$;

Απάντηση

1) Οι αριθμοί που χρησιμοποιούνται στο διάστημα $[0, 7]$ είναι ακέραιοι αριθμοί. Εμείς θα χρησιμοποιήσουμε τη δυαδική τους συμβολοσειρά.

0→000

1→001

2→010

Σημείωση

3→011

Με τη βοήθεια των αριθμών αυτών θα συμπληρώσουμε τον

4→100

πίνακα. Στις στήλες X_1 X_2 θα γράψουμε τους αριθμούς

5→101

που αντιστοιχούν.

6→110

7→111

	Χρωμόσωμα	X ₁	X ₂	Απόδοση Ικανότητα	Νέα ικανότητα
A	101011	5	3	-16	
B	111000	7	0	-49	
Γ	111011	7	3	-49	
Δ	101000	5	0	-25	
E	111001	7	1	-48	

Χρησιμοποιούμε τη συνάρτηση $f(x_1, x_2) = x_2^2 - x_1^2$ για να υπολογίσουμε την ικανότητα.
 $f(x_1, x_2) = x_2^2 - x_1^2 = 3^2 - 5^2 = 9 - 25 = -16$

$$f(x_1, x_2) = x_2^2 - x_1^2 = 0^2 - 7^2 = 0 - 49 = -49$$

$$f(x_1, x_2) = x_2^2 - x_1^2 = 3^2 - 7^2 = 9 - 49 = -49$$

$$f(x_1, x_2) = x_2^2 - x_1^2 = 0^2 - 5^2 = 0 - 25 = -25$$

$$f(x_1, x_2) = x_2^2 - x_1^2 = 1^2 - 7^2 = 1 - 49 = -48$$

Επειδή όλες οι ικανότητες - καταλληλότητες μας δίνουν αρνητικό αποτέλεσμα πρέπει να προσθέσουμε μια κατάλληλη σταθερά για να έχουμε αποτέλεσμα θετικό. Η σταθερά αυτή είναι ο αριθμός 50. Προσθέτουμε το 50 γιατί αν προσθέσω το 49 το δεύτερο άτομο θα είναι 0. Το 0 είναι ένα αποτέλεσμα που μας καθορίζει ότι δεν έχει το άτομο την πιθανότητα να επιλεχθεί. Η αρχή του γενετικού αλγορίθμου μας αναφέρει ότι ακόμα και το χειρότερο άτομο πρέπει να έχει την πιθανότητα να επιλεχθεί. Επομένως η απάντηση για το αν υπάρχει πρόβλημα επιλογής δευτέρου ατόμου είναι ότι δεν υπάρχει πρόβλημα.

Η νέα απόδοση κάθε πληθυσμού προσθέτοντας τον αριθμό 50 είναι:

	Χρωμόσωμα	X ₁	X ₂	Απόδοση Ικανότητα	Νέα ικανότητα
A	101011	5	3	-16	34
B	111000	7	0	-49	1
Γ	111011	7	3	-49	10
Δ	101000	5	0	-25	25
E	111001	7	1	-48	2

2) Η πιθανότητα επιλογής χρησιμοποιώντας επιλογή εξαναγκασμένης ρουλέτας είναι:

$$p = \frac{\text{Νέα ικανότητα}}{\text{Σύνολο ικανοτήτων}}$$

$$\text{Σύνολο νέων ικανοτήτων} = 34 + 1 + 10 + 25 + 2 = 72$$

$$\text{Μέση απόδοση} \rightarrow \bar{F} = 72/5 = 14,4$$

$$p(A) = 34/72 = 0,4722$$

$$p(B)=1/72=0,0138$$

$$p(\Gamma)=10/72=0,1388$$

$$p(\Delta)=25/72=0,3472$$

$$p(E)=2/72=0,0277$$

3) Αριθμός αντιγράφων= αριθμός ατόμων * p

	Χρωμόσωμα	X ₁	X ₂	Απόδοση Ικανότητα	Νέα ικανότητα	Αριθμός αντιγράφων
A	101011	5	3	-16	34	2,3610
B	111000	7	0	-49	1	0,0694
Γ	111011	7	3	-49	10	0.6945
Δ	101000	5	0	-25	25	1,7360
E	111001	7	1	-48	2	0,1388

Ο παραπάνω πίνακας αναφέρεται στη γενιά 0.

4)

Πίνακας στη γενιά 1

	Άτομο- Χρωμόσωμα	X1	X2	Ικανότητα	Νέα ικανότητα
A	111 100	7	4	-33	17
B	110 110	6	6	0	50
Γ	111 000	7	0	-49	1
Δ	110 011	6	3	-27	23
E	111 011	7	3	-40	10

Η συνολική απόδοση είναι F=101

Η μέση απόδοση $\bar{F}=101/5=20,2$

Συγκρίνοντας την μέση απόδοση στη γενιά 0 με την μέση απόδοση στη γενιά 1 παρατηρούμε ότι έχει βελτιωθεί.

5) Χρησιμοποιώντας τον τύπο του θεωρήματος σχημάτων έχουμε:

$$\xi(S,t+1) \geq \xi(S,t) * \frac{eval(S,t)}{F(t)} * [1 - p_{\frac{\delta(S)}{m-1}} - o(S) * pm]$$

Οι συμβολοσειρές που ταιριάζουν με το σχήμα S=111*** είναι οι A, Γ, E που ανήκουν στην γενιά 1.

Άρα το $\xi(S,t)=3$

$$\text{eval}(S,t)=\frac{17+1+10}{3}=9,3333$$

$$\bar{F}(t)=20,2$$

$$pc=0,5, pm=0,1$$

$$\delta(S)=3-1=2, m=6$$

$$o(S)=3$$

Με αντικατάσταση έχουμε:

$$\xi(S,t+1) \geq 3 * \frac{9,3333}{20,2} * [1 - 0,5 \frac{2}{5} - 3 * 0,1] \geq 0,69$$

5.5 Πλεονεκτήματα των Γενετικών αλγορίθμων.

Η χρήση των γενετικών αλγορίθμων εμφανίζει τα ακόλουθα πλεονεκτήματα όπως:

1. Έχουν την δυνατότητα να επιλύουν πολύπλοκα προβλήματα γρήγορα και με ακρίβεια. Ο κύριος λόγος που χρησιμοποιούνται οι γενετικοί αλγόριθμοι είναι λόγω της μεγάλης τους αποδοτικότητας. Γι αυτό και τα περισσότερα πολύπλοκα προβλήματα λύνονται μέσω των γενετικών αλγορίθμων.
2. Επεξεργάζεται και εξελίσσει ένα πληθυσμό λύσεων σε σχέση με άλλες μεθόδους που επεξεργάζονται μόνο μια λύση στο χώρο αναζήτησης.
3. Έχουν πολλά πεδία εφαρμογής, μπορούν να εφαρμοστούν για τη λύση μαθηματικών εξισώσεων όπως και σε εφαρμογές νευρωνικών δικτύων.
4. Δέχεται δεδομένα σε κάθε βήμα της εκτέλεσής του με σκοπό την παραγωγή της καλύτερης λύσης.
5. Είναι η μόνη μέθοδος που έχει πετύχει συνδυασμό εξερεύνησης στο χώρο αναζήτησης και εκμετάλλευση της πληροφορίας. Δηλαδή χρησιμοποιεί τη συνάρτηση που του δίνεται και την εφαρμόζει τμηματικά κάνοντας τυχαία αναζήτηση επιτυγχάνοντας έτσι καλύτερη εξερεύνηση του χώρου.
6. Βασίζεται σε πιθανότητες δηλαδή χρησιμοποιεί τυχαίους αριθμούς σε περιοχές που θα δώσουν τα καλύτερα αποτελέσματα.

5.6 Προσδοκώμενα αποτελέσματα

Από τη μελέτη της πιο πάνω εργασίας καταλήγουμε στο συμπέρασμα ότι οι αλγόριθμοι επιλύουν ένα μεγάλο πλήθος υπολογιστικών βιολογικών προβλημάτων που έχουν καθοριστικό ρόλο στην εξέλιξη της Μοριακής και Γενετικής βιολογίας. Ευελπιστούμε ότι τόσο στον τομέα της Γενετικής όσο και στην επιστήμη της Ιατρικής οι επιστήμονες θα επινοήσουν αλγορίθμους οι οποίοι θα είναι βασισμένοι στους ήδη υπάρχοντες, έτσι ώστε να μπορέσουν να δοθούν λύσεις για την αντιμετώπιση ανίατων ασθενειών που μαστίζουν εκατομμύρια ανθρώπων.

Βιβλιογραφία:

- [1] (Klaus-Dieter Paul:Biologie, fur Schule und Beruf, Verlag Europa Lehrmittel Nourney,1993) Για την ελληνική έκδοση, μετάφραση Κάρολος Βούλγαρης, Βιολογία: Τεχνολογία και Περιβάλλον, Ευρωπαϊκές Τεχνικές Εκδόσεις,1997.
- [2] <https://el.wikipedia.org/wiki/Αμινοξέα>
- [3] Ψηφιακό σχολείο (Κεφάλαιο 2 Αμινοξέα-πεπτίδια)
<http://ebooks.edu.gr/modules/ebook/show.php/DSGL-C120/480/3166,12747/>
- [4] <http://ebooks.edu.gr/modules/ebook/show.php/DSGL-B106/85/678,2573/>
- [5] <http://ebooks.edu.gr/modules/ebook/show.php/DSGL-B106/85/678,2573/>
- [6] <http://kpe-kastor.kas.sch.gr/leaf/lexicon/d/dna.htm>
- [7] <https://el.wikipedia.org/wiki/RNA>
- [8] <http://nf1.bioacademy.gr/index.php/nf-mainmenu-30/mainmenu-32/mainmenu-33>
- [9] https://el.wikipedia.org/wiki/Δυναμικός_προγραμματισμός
- [10] [http://www.psounis.gr/πλη30/Ενότητα 2/σχεδίαση αλγορίθμων/μάθημα 2.2 δυναμικός προγραμματισμός/](http://www.psounis.gr/πλη30/Ενότητα_2/σχεδίαση_αλγορίθμων/μάθημα_2.2_δυναμικός_προγραμματισμός/)
- [11] [http://www.psounis.gr/πλη31/Ενότητα 4/γενετικοί αλγόριθμοι/μάθημα 4.1/Εισαγωγή στους Γενετικούς Αλγορίθμους/](http://www.psounis.gr/πλη31/Ενότητα_4/γενετικοί_αλγόριθμοι/μάθημα_4.1/Εισαγωγή_στους_Γενετικούς_Αλγορίθμους/)
- [12] <http://www.psounis.gr/πλη31/>
- [13] http://biophysics.biol.uoa.gr/courses/bioinformatics/BLAST_CLUSTAL/blast.html
- [14] <https://el.wikipedia.org/wiki/Perl>
- [15] <http://docplayer.gr/6142621-Ena-programma-entopismoy-gonidiakon-diplasiasmon.html>
- [16] <https://el.wikipedia.org/wiki/JavaScript>
- [17] [https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))
- [18] <http://docplayer.gr/4881059-Panepistimio-peiraios-tmima-pliroforikis-programma-metaptyhiakon-spydon-pliroforiki.html>
- [19] https://en.wikipedia.org/wiki/Smith–Waterman_algorithm
- [20] [http://www.psounis.gr/πλη31/Ενότητα 4/γενετικοί αλγόριθμοι/μάθημα 4.3 θέμelia γενετικών αλγορίθμων/](http://www.psounis.gr/πλη31/Ενότητα_4/γενετικοί_αλγόριθμοι/μάθημα_4.3_θεμέλια_γενετικών_αλγορίθμων/)