Το πρόγραμμα έχει ως στόχο την μετάφραση μιας συμπληρωματικής νουκλεοτιδικής αλληλουχίας, cDNA (complementary DNA), σε πρωτεΐνη.

Στη πραγματικότητα, in vivo,η RNA-polymerase σκανάρει το DNA, βρίσκει τον promoter και ξεκίνα τη μεταγραφή του σε mRNA. Έξω από το πύρινα, στα ριβοσώματα, μεταφράζεται το mRNA σε πρωτεΐνη. Το RNA που δημιουργείται έχει κατεύθυνση 5’->3’, αφού ο κλώνος του DNA που χρησιμοποιείτε είναι πάντα 3’->5’. Μια διαφορά ανάμεσα στα ευκαρυωτικά και τα προκαρυωτικά κύτταρα είναι ότι τα γονίδια στα ευκαρυωτικά κύτταρα μπορεί να περιέχουν εσώνια (ή αλλιώς ιντρόνια), τα οποία δεν μεταφράζονται σε αμινοξέα, και πρέπει να αφαιρεθούν από το mRNA πριν από την μετάφραση.

Πολλές φορές υπάρχει ανάγκη να εκφραστούν ευκαρυωτικά γονίδια από προκαρυωτικά κύτταρα. Μια απλή μέθοδος είναι να μεταφερθεί το DNA ενός ευκαρυωτικού κυττάρου σε ένα προκαρυωτικό κύτταρο - ξενιστή, το οποίο θα μεταγραφεί σε mRNA και θα μεταφραστεί σε πρωτεΐνη. Όμως, τα προκαρυωτικά κύτταρα δεν έχουν τη δυνατότητα να αφαιρέσουν τα εσώνια από το πρόδρομο mRNA, επομένως η αφαίρεσή τους πρέπει να γίνει πριν την μεταφορά. Αυτού του είδους το DNA, το οποίο έχει προέλθει από ώριμο mRNA, ώστε να μην περιέχει εσώνια, είναι το cDNA. Για να γίνει έκφραση απαιτείται επιπλέον η μεταφορά μεταγραφικών παραγόντων, όπως ο υποκινητής, στα προκαρυωτικά κύτταρα. Κάθε προκαρυωτικό κύτταρο που δέχεται το γονίδιο πολλαπλασιάζεται, παράγοντας έτσι ένα βακτηριακό κλώνο. Το σύνολο των βακτηριακών κλώνων που περιέχουν αντίγραφα όλων των ώριμων mRNA των γονιδίων που εκφράζονται σε έναν κυτταρικό τύπο ονομάζεται cDNA βιβλιοθήκη.

Όπως είπαμε, το RNA στο οποίο μεταγράφηκε η πληροφορία μπορεί να ταξιδέψει εκτός του πυρήνα και μόλις συναντήσει κάποιο ριβόσωμα αρχίζει η δημιουργία της πολυπεπτιδικής αλυσίδα. (Σε αυτό το στάδιο συμμετέχουν και άλλα δύο είδη RNA, το sRNA και το tRNA.) Χαρακτηριστικό αυτής της διαδικασίας είναι ότι η γενετική πληροφορία που είναι κωδικοποιημένη στο RNA σε τέσσερα διαφορετικά χαρακτηριστικά συμπλέγματα (τις τέσσερις αζωτούχες βάσεις) μετατρέπεται σε είκοσι διαφορετικά συμπλέγματα, τα είκοσι αμινοξέα.

Τα tRNA συνδέονται με αμινοξέα. Το κάθε tRNA φέρει ένα σημείο με το οποίο μπορεί να συνδεθεί στο mRNA, το οποίο ονομάζεται αντικωδικόνιο. Το αντικωδικόνιο αποτελείται από τρία διαδοχικά νουκλεοτίδια τα οποία είναι συμπληρωματικά τριών διαδοχικών νουκλεοτιδίων του mRNA, η οποία διαδοχική τριάδα νουκλεοτιδίων ονομάζεται κωδικόνιο, δηλαδή υπάρχει συμπληρωματικότητα ως προς τα κωδικόνια και τα αντικωδικόνια. Κάθε κωδικόνιο στο mRNA, αντιστοιχεί σε ένα αμινοξύ, αφού κάθε κωδικόνιο αντιστοιχεί ακριβώς σε ένα είδος tRNA που συνδέεται με ένα συγκεκριμένο αμινοξύ.

Το σύστημα αντιστοιχίας κωδικονίων-αντικωδικονίων-αμινοξέων είναι παγκόσμιο, ισχύει το ίδιο ακριβώς σε κάθε οργανισμό και ονομάζεται γενετικός κώδικας. Επειδή, το κάθε κωδικόνιο του mRNA αντιστοιχεί ακριβώς σε μια τριάδα νουκλεοτιδίων του DNA (κωδικοποίηση σε τριπλέτες), οι διαδοχικές ακολουθίες νουκλεοτιδίων στο DNA καθορίζουν ποιες πολυπεπτιδικές αλυσίδες θα παραχθούν στο κύτταρο και έμμεσα ποιες πρωτεΐνες θα παραχθούν σε αυτό, και μέσω αυτών ποιες λειτουργίες, ποια χαρακτηριστικά, ποιες αντιδράσεις θα έχει το κύτταρο στον κύκλο ζωής του.

Στα ευκαρυωτικά κύτταρα η μετάφραση δεν αρχίζει από την πρώτη τριάδα του mRNA, αλλά από το πρώτο AUG, γνωστό ως κωδικόνιο έναρξης. Αντίστοιχα, η μετάφραση τερματίζεται όταν εμφανιστεί ένα από τα κωδικόνια λήξης (UAA, UAG και UGA). Στη περίπτωση του cDNA βέβαια, η πυριμιδίνη ουρακίλη αντικαθίσταται με τη θυμίνη. Έτσι το αρχικό αμινοξύ των γονιδίων, είναι η μεθιονίνη με το κωδικόνιο ATG (βλέπε πίνακα). Για την τρυπτοφάνη σημάνει το κωδικόνιο TGG, και για τα υπόλοιπα 18 αμινοξέα υπάρχουν δυο, τρία ή και περισσότερα κωδικόνια τα οποία και φαίνονται στο παρακάτω πίνακα:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Amino acid** | **Codons** | **Compressed** |  | **Amino acid** | **Codons** | **Compressed** |
| **Ala/A** | GCT, GCC, GCA, GCG | GCN | **Leu/L** | TTA, TTG, CTT, CTC, CTA, CTG | YTR, CTN |
| **Arg/R** | CGT, CGC, CGA, CGG, AGA, AGG | CGN, MGR | **Lys/K** | AAA, AAG | AAR |
| **Asn/N** | AAT, AAC | AAY | **Met/M** | ATG | |
| **Asp/D** | GAT, GAC | GAY | **Phe/F** | TTT, TTC | TTY |
| **Cys/C** | TGT, TGC | TGY | **Pro/P** | CCT, CCC, CCA, CCG | CCN |
| **Gln/Q** | CAA, CAG | CAR | **Ser/S** | TCT, TCC, TCA, TCG, AGT, AGC | TCN, AGY |
| **Glu/E** | GAA, GAG | GAR | **Thr/T** | ACT, ACC, ACA, ACG | ACN |
| **Gly/G** | GGT, GGC, GGA, GGG | GGN | **Trp/W** | TGG | |
| **His/H** | CAT, CAC | CAY | **Tyr/Y** | TAT, TAC | TAY |
| **Ile/I** | ATT, ATC, ATA | ATH | **Val/V** | GTT, GTC, GTA, GTG | GTN |
| **START** | ATG | | **STOP** | TAA, TGA, TAG | TAR, TRA |

( <https://en.wikipedia.org/wiki/DNA_codon_table> )

Επειδή πολλά αμινοξέα μπορούν να παραχθούν με πολλές διαφορετικές τριπλέτες λέμε πως έχουμε εκφυλισμό του γενετικού κώδικα.   
(61 κωδίκια -> 20 αμινοξέα)

Το γενετικό υλικό μπορεί να υποστεί αλλαγές με πολλούς διαφορετικούς τρόπους. Οι αλλαγές στην αλληλουχία του DNA, που ονομάζονται μεταλλάξεις, δημιουργούν συνήθως ένα διαφορετικό φαινότυπο, χωρίς όμως αυτό να είναι πάντοτε απαραίτητο. Αυτό εξαρτάται από τον τρόπο με τον οποίο η αλλαγή επιδρά στο γονιδιακό προϊόν, δηλαδή στην πρωτεΐνη.

Οι μεταλλάξεις ευθύνονται για τη δημιουργία γενετικής ποικιλότητας στον πληθυσμό, για πολλές κληρονομικές ασθένειες, καθώς και για πολλές περιπτώσεις καρκίνου. Μεταλλάξεις μπορεί να συμβούν σε οποιοδήποτε γεννητικό ή σωματικό κύτταρο. Μόνο οι μεταλλάξεις των γεννητικών κυττάρων, εν τούτοις, μπορεί να μεταβιβαστούν από τη μια γενιά στην επόμενη και γι' αυτό είναι υπεύθυνες για τη δημιουργία κληρονομικών ασθενειών. Αυτό όμως δε σημαίνει ότι οι σωματικές μεταλλάξεις είναι λιγότερο σημαντικές για την υγεία. Στην πραγματικότητα αποτελούν την πλειονότητα των μεταλλάξεων.

Μία ασθένεια μπορεί να είναι το αποτέλεσμα αντικατάστασης μίας μόνο από τα δισεκατομμύρια βάσεων DNA. Αλλαγές αυτού του τύπου ονομάζονται σημειακές μεταλλάξεις (SNPs), και μπορεί να έχουν ποικίλα αποτελέσματα στην πρωτεΐνη που παράγεται από το αντίστοιχο γονίδιο. Στην περίπτωση που η διαφορετική τριπλέτα που προέκυψε κωδικοποιεί το ίδιο αμινοξύ (συνώνυμο κωδικόνιο) δεν αλλάζει η ακολουθία αμινοξέων στην παραγόμενη πρωτεΐνη.

Στις περισσότερες όμως περιπτώσεις μια σημειακή μετάλλαξη δημιουργεί μια τριπλέτα που κωδικοποιεί ένα διαφορετικό αμινοξύ και κατά συνέπεια μια αλλαγμένη πρωτεΐνη. Εάν το διαφορετικό αμινοξύ βρίσκεται στο ενεργό κέντρο ενός ενζύμου ή κοντά σε αυτό, τότε η ενεργότητά του, δηλαδή η ικανότητα κατάλυσης αντιδράσεων, μπορεί να ελαττωθεί ή και να μηδενισθεί. Επίσης, σε άλλα είδη πρωτεϊνών η μετάλλαξη μπορεί να οδηγήσει σε αλλαγή της δομής τους και συνεπώς και της λειτουργίας τους. Σε ελάχιστες περιπτώσεις μια σημειακή μετάλλαξη μπορεί να μετατρέψει ένα κωδικόνιο, που κωδικοποιεί κάποιο αμινοξύ, σε ένα κωδικόνιο λήξης, με αποτέλεσμα τον τερματισμό της πολυπεπτιδικής αλυσίδας. Στις περισσότερες από αυτές τις περιπτώσεις καταστρέφεται η λειτουργικότητα της πρωτεΐνης.

Ένας άλλος σημαντικός τύπος γονιδιακών μεταλλάξεων περιλαμβάνει προσθήκη ή έλλειψη βάσεων. Αλλαγές στον αριθμό των βάσεων έχουν ως αποτέλεσμα την εμφάνιση μεταλλαγμένων φαινοτύπων. Η προσθήκη ή η έλλειψη διαδοχικών βάσεων σε οποιοδήποτε αριθμό πολλαπλάσιο του τρία δημιουργεί, αντίστοιχα, προσθήκη ή έλλειψη ενός ή περισσότερων αμινοξέων στην πολυπεπτιδική αλυσίδα, που μπορεί να αλλάζει τη λειτουργικότητά της. Αν όμως ο αριθμός των βάσεων είναι διαφορετικός του τρία ή πολλαπλασίων του, τότε διαταράσσεται το πλαίσιο ανάγνωσης των τριπλετών. Συνεπώς η αλληλουχία των αμινοξέων δεν εμφανίζει πλέον πολλές ομοιότητες με την αρχική. Μια τέτοιου τύπου μετάλλαξη σε ένα γονίδιο που κωδικοποιεί ένα ένζυμο συνήθως οδηγεί σε απώλεια της ενεργότητάς του.\*

Ας περάσουμε τώρα στην υλοποίηση του προγράμματος.

Αρχικά ζητείται από το χρήστη το αρχείο fasta από το οποίο θα διαβάσουμε την ακολουθία (input file), και ένα αρχείο οπού θα τυπωθούν τα αποτελέσματα (output file). Για να ξεπεράσουμε το πρόβλημα που πρόεκυπτε στη μνήμη από μεγάλες πρωτεΐνες(>5000bp), υπολογίζουμε πρώτα τις γραμμές του αρχείου, και μέσω αυτών ορίζεται ένα string (cDNA) αρκετά μεγάλο ώστε να χωρά όλα τα νουκλεοτίδια.

Μετά, αποθηκεύουμε όλη την νουκλεοτιδική ακολουθία σε αυτό το string.

Έτσι μπορούμε να κλείσουμε το αρχείο αυτό, καθώς έχουμε αποθηκεύσει όλη τη πληροφορία που χρειαζόμαστε στο sting με όνομα cdna. Εν συνεχεία, μιμούμενοι το τρόπο οπού γίνεται η μετάφραση στα ριβοσώματα, διαβάζουμε ανά τρία τα νουκλεοτίδια του cdna και τα μεταφράζουμε σε πρωτεΐνη. Αυτό γίνεται με τη συνάρτηση check( ), η οποία αναγνωρίζει όλες τις τριπλέτες του παραπάνω πίνακα συμπεριλαμβανόμενων και των Compressed Codons. Στο κωδικόνιο λήξης αντιστοιχεί το σύμβολο της δίεσης ‘#’ .

Το επόμενο βήμα είναι ο καθορισμός των SNPs από το χρήστη, και ο υπολογισμός της νέας πρωτεΐνης. Για την ευκολότερη ανάγνωση των αποτελεσμάτων, και για να μπορεί η σύγκριση μεταξύ των δυο αλληλουχιών να γίνει άμεσα, ακολουθήθηκε μια διαδικασία εμπνευσμένη από άλλα εργαλεία στοίχισης. Έτσι τοποθετείται σε μια γραμμή η πρώτη αλληλουχία, δυο γραμμές πιο κάτω η αλληλουχία της δεύτερης πρωτεΐνης και ανάμεσα τους με το σύμβολο ‘^’ δείχνεται οποιο αμινοξύ διαφέρει. Έτσι μια σιωπηλή μετάλλαξη δε θα έχει καθόλου ‘^’ και μια μετάλλαξη που επηρεάζει το τρίτο, π.χ. αμινοξύ μιας ακολουθίας θα φαινόταν κάπως έτσι:

MAEKLMN#

^

MAGKLMN#

Καθ’ όλη τη διάρκεια ο χρήστης καθοδηγείται με σαφήνεια από το πρόγραμμα προς αποφύγει λαθών.

Τα αποτελέσματα του προγράμματος έχουν ελεγχθεί με το διεθνώς αναγνωρισμένο εργαλείο EMBOSS Transeq του EMBL. <http://www.ebi.ac.uk/Tools/st/emboss_transeq/>

\*Η παράγραφος αυτή, που είναι σημαδεμένη με αστερίσκο, περιγράφει κάτι που άρχισα να κάνω σαν επέκταση του προγράμματος αλλά λόγω εξεταστικής προς το παρόν αναβάλλεται. Δε παύει να είναι όμως, μια από τις προοπτικές του συγκεκριμένου προγράμματος.