Βιοπληροφορική ΙΙ

Ασκήσεις Εργαστηρίου

Άσκηση 1

Γράψτε ένα πρόγραμμα σε γλώσσα Perl το οποίο θα παίρνει σαν είσοδο ένα αρχείο με πολλαπλές εγγραφές από τη Uniprot (π.χ. το αρχείο transmem_proteins.swiss που βρίσκεται στο http://www.compgen.org/material/courses/bioinformatics2/transmembrane-proteins στο τέλος της σελίδας) και θα δίνει σαν αποτέλεσμα ένα αρχείο .txt με τις ακόλουθες πληροφορίες για κάθε πρωτεΐνη (εγγραφή):

1. Στην πρώτη γραμμή θα εμφανίζει το σύμβολο > και στη συνέχεια το ID, το AC (μόνο το ισχύον) και το μήκος της πρωτεΐνης σε αμινοξικά κατάλοιπα και ενδιάμεσα θα τυπώνεται ο χαρακτήρας «|».

2. Στην δεύτερη γραμμή θα τυπώνεται η αμινοξική ακολουθία της πρωτεΐνης χωρίς κενά.

3. Στην τρίτη γραμμή θα εμφανίζεται η θέση των διαμεμβρανικών τμημάτων της πρωτεΐνης σε μία συμβολοσειρά ίδιου μήκους με την αμινοξική ακολουθία όπου οι θέσεις των διαμεμβρανικών θα συμβολίζονται με «M» και όλες οι υπόλοιπες με «-».

4. Στην τέταρτη γραμμή θα εμφανίζονται τα σύμβολα «//» που υποδεικνύουν το τέλος της εγγραφής στη UniProt.

Το αρχείο αποτελεσμάτων θα είναι της μορφής (δείχνεται τμήμα μιας εγγραφής):

```
>ADBR2_HUMAN|P07550|413aa
MGQPGNGSAFLLAPNGSHAPDHDVTQERDEVWVVGMGIVMSLIVLAIVFGNVLVITAIAKFERLQTVTN
----------------------------------MMMMMMMMMMMMMMMMMMMMMMM----------
//
```

Άσκηση 2

Γράψτε ένα πρόγραμμα σε γλώσσα Perl το οποίο θα παίρνει σαν είσοδο ένα αρχείο μιας εγγραφής από τη Uniprot (μπορείτε να δουλέψετε με την πρωτεΐνη με AC P07550 αλλά το πρόγραμμα πρέπει να μπορεί να χρησιμοποιηθεί σε οποιαδήποτε εγγραφή) και χρησιμοποιώντας μια κλίμακα υδροφοβικότητας π.χ. την Kyte-Doolitle (http://www.compgen.org/material/courses/bioinformatics2/transmembrane-proteins) θα απομονώνει την αμινοξική ακολουθία και θα κάνει πρόβλεψη για τις θέσεις των διαμεμβρανικών τμημάτων με χρήση της μεθόδου που περιγράψαμε στο εργαστήριο για τη μέση υδροφοβικότητα σε κυλιόμενο παράθυρο. Το μέγεθος του παραθύρου να δίνεται στο πρόγραμμα σαν παράμετρος την οποία θα καθορίζει ο χρήστης. Επιπλέον το πρόγραμμα θα θεωρεί ως πραγματικά όρια διαμεμβρανικών αυτά που δίνονται στην UniProt και θα υπολογίζει την επιτυχία του αλγορίθμου πρόγνωσης με δύο μέτρα αξιοπιστίας, την **ακρίβεια** (accuracy = (TP+TN)/(TP+FP+TN+FN)) και τον **συντελεστή συσχέτισης Matthews**

($MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$).

Τέλος το πρόγραμμα θα το εκτελέσετε 3 φορές για διαφορετικά μεγέθη παραθύρου **9, 15 και 21** και θα σχολιάσετε την επιτυχία της μεθόδου σε κάθε περίπτωση.

Άσκηση 3

Γράψτε ένα πρόγραμμα σε Perl το οποίο θα παίρνει σαν είσοδο ένα αρχείο με πολλαπλές εγγραφές από τη UniProt, θα το μετατρέπει σε ένα one-line fasta αρχείο (η ακολουθία σε μία γραμμή) και στη συνέχεια στο fasta αρχείο θα ψάχνει να βρει αν οι ακολουθίες έχουν στην αρχή τους το πεπτίδιο οδηγητή των λιποπρωτεϊνών στα βακτήρια. Το pattern που πρέπει να αναζητήσετε στην αρχή κάθε ακολουθίας είναι:

```
X(οσεσδήποτε φορές)-{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-
[AGS]-C
```

Το τελικό αρχείο εξόδου του προγράμματος θα περιλαμβάνει τα πεπτίδια οδηγητές που εντοπίστηκαν (ένα σε κάθε γραμμή). Για να δουλέψετε χρησιμοποιήστε σαν αρχείο εισόδου το αρχείο με τις 63 ακολουθίες που σας δίνετε στον ακόλουθο σύνδεσμο http://www.compgen.org/material/courses/bioinformatics2/lipoprotein-signals