

Βιοπληροφορική II

Ασκήσεις Εργαστηρίου Μέρος 2^ο

Άσκηση 4. Hidden Markov Models και profile HMMs

Δίνεται η παρακάτω άγνωστη ακολουθία:

```
>OMPA_ECOLI
```

```
MKKTAIAIAVALAGFATVAQAAPKDNTWYTGAKLGWSQYHDTGFINNNGPTHENQLGAGA  
FGGYQVNPYVGFEMGYDWLGRMPYKGSVENGAYKAQGVQLTAKLGYPITDDLDIYTRLGG  
MVWRADTKSNVYGKNHDTGVSPVFAGGVEYAITPEIATRLEYQWTNNIGDAHTIGTRPDN  
GMLSLGVSYRFGQGEAAPVVAPOPAPAPEVQTKHF T LKSDVLFNFNKATLKPEGQAALDQ  
LYSQLSNLDPKDGSVVVLGYTDRI GSDAYNQGLSERRAQSVDYLI SKGIPADKISARGM  
GESNPVTGNTCDNVKQRAALIDCLAPDRRVEIEVKGIKDVVTQPQA
```

Να προσδιορίσετε τη δομή και λειτουργία της άγνωστης ακολουθίας χρησιμοποιώντας αλγορίθμους πρόγνωσης που στηρίζονται στην ακολουθία μιας πρωτεΐνης και μόνο, ακολουθώντας τα ακόλουθα βήματα:

1. Χρησιμοποιήστε το BLAST έναντι της nr και αξιολογήστε τα αποτελέσματα που πήρατε. Έχετε κάποια ένδειξη για την άγνωστη πρωτεΐνη;
2. Χρησιμοποιήστε το BLAST έναντι της SwissProt χωρίς να συμπεριλάβετε τον οργανισμό *E. coli* και συλλέξτε τα 10 πρώτα αποτελέσματα από διαφορετικούς οργανισμούς. Χρησιμοποιήστε τις 11 αυτές αλληλουχίες (μαζί με την άγνωστη) για να κατασκευάσετε μια πολλαπλή στοίχιση με το εργαλείο T-Coffee. Σχολιάστε την πολλαπλή στοίχιση.
3. Χρησιμοποιήστε αλγορίθμους πρόγνωσης 2ταγούς δομής (SecStr, PSI-PRED), διαμεμβρανικών τμημάτων (OCTOPUS, HMMTM, BOCTOPUS, PRED-TMBB) και πεπτιδίου οδηγητή (SIGNALP) και σχολιάστε τα αποτελέσματά τους.
4. Μέσω του <http://hmmer.org/>:
 1. Χρησιμοποιήστε την πολλαπλή στοίχιση και με τη βοήθεια του hmmsearch κατασκευάστε ένα pHMM και κάντε αναζήτηση έναντι της SwissProt. Στη συνέχεια, χρησιμοποιήστε το rhhmer και το BLAST για αναζήτηση στη SwissProt και συγκρίνετε τα αποτελέσματα των αναζητήσεων.
 2. Χρησιμοποιήστε το pHMM που κατασκευάσατε στο προηγούμενο βήμα και κάντε αναζήτηση έναντι της SwissProt μόνο για πρωτεΐνες του οργανισμού *E. coli* και συγκρίνετε τα αποτελέσματά σας με την αντίστοιχη αναζήτηση μέσω BLAST.
5. Προσπαθήστε με βάση όλα τα προηγούμενα βήματα αλλά και ό,τι άλλο θεωρείτε ότι μπορεί να σας βοηθήσει να προσδιορίσετε τη δομή και τη λειτουργία της άγνωστης πρωτεΐνης που σας δόθηκε.

Άσκηση 5. Artificial Neural Networks

Στόχος είναι η δημιουργία αλγορίθμου πρόγνωσης της υποκυτταρικής θέσης των υδατοδιαλυτών πρωτεϊνών των αρνητικών κατά Gram βακτηρίων με χρήση Τεχνητών Νευρωνικών Δικτύων όπως αυτός περιγράφηκε στο αντίστοιχο εργαστήριο. Στην εργασία πρέπει να εκπαιδεύσετε και να αξιολογήσετε τα NN με τις ακόλουθες αρχιτεκτονικές:

1. Χωρίς κρυφούς νευρώνες
2. Με ένα στρώμα 3 κρυφών νευρώνων
3. Με ένα στρώμα 7 κρυφών νευρώνων
4. Με ένα στρώμα 11 κρυφών νευρώνων
5. Με ένα στρώμα 15 κρυφών νευρώνων
6. Με δύο στρώματα 3 κρυφών νευρώνων το καθένα.

Αφού καταλήξετε μεταξύ ποιων αρχιτεκτονικών βρίσκεται η βέλτιστή, προσπαθήστε με επιπλέον δοκιμές να καταλήξετε στον ιδανικό αριθμό νευρώνων. Για την αξιολόγηση κάθε νευρωνικού δικτύου, χρησιμοποιήστε τον δείκτη C-index.

Η εντολή που τρέχετε είναι πάντα: `>np.exe prot.net 100 10`