Πανεπιστημιο Κυπρου
Τμημα Βιολογικων Επιστημων
Ερευνητικο Εργαστηριο Βιοπληροφορικης
Βασιλειος Ι. Προμπονας, PhD
P.O. Box 20537, 1678 Λευκωσια, Κυπρος
email: vprobon@ucy.ac.cy, web: http://troodos.biol.ucy.ac.cy

BIO 331 - Αρχές και Μέθοδοι Βιοπληροφορικής I

Θέματα εργασιών και αρχικές οδηγίες, 16 Νοεμβρίου 2011.

Σύνθεση ομάδων εργασίας και εργασίες ανά κατηγορία

Ομάδα	Φοιτητής 1	Φοιτητής 2	Φοιτητής 3	A	B
1	Παναγιώτα Δημοσθένους	Κωνσταντίνα Ιωάννου	•	4	3
2	Thekla Christodoulou	Agata Combi	•	2	1
3	Maria Eleftheriou	Kyriaki Nikolaou	•	3	1
4	Κατερίνα Όθωνος	Τιμοθέα Κωνσταντίνου	Μιχαλένα Ηλία	4	3
5	Χριστόφορος Παύλου	Γιώργος Μιχαήλ	•	7	2
6	Αθηνά Βλασίου	Ελένη Κουμενή	Χρυσοστόμη Περιστιάνη	1	3
7	Κωνσταντίνα Κουτσοφτή	Ελένη Χρυσάνθου	Μαρίνα Ρωτσίδου	2	1
8	Κωσταρή Μαριλένα	Μασούρα Βασιλική	Μοδέστου Χριστίνα	1	1
9	katerina constantinou	eleni ioannidou	kyriaki antoniou	6	2
10	Μελπομένη Στεφάνη	Μαργαρίτα Καλυβά	Χριστίνα Θρασίου	7	2
11	Σταυρούλα Αρτεμίου	Ραφαέλα Πολυκάρπου	Μαρία Λαζάρου	3	3

Αρχικές οδηγίες

Ομάδα Α:

Εργασία 1. Μελέτη χαρακτηριστικών σε πολλαπλές στοιχίσεις νουκλεοτιδικών ή αμινοξικών αλληλουχιών.

Να ξεκινήσετε μελετώντας τους τύπους μορφοποίησης της εξόδου του λογισμικού CLUSTALW (http://www.clustal.org/clustal2/). Από τις δυνατές μορφοποιήσεις να επικεντρωθείτε στο CLUSTAL format. Κατεβάστε την τελευταία έκδοση στον υπολογιστή που έχετε πρόσβαση και εξοικειωθείτε με την εκτέλεση της εφαρμογής.

- Θα χρησιμοποιήσετε αρχείο με τις κωδικές αλληλουχίες (CDSs) ομόλογων του γονιδίου SufI από οργανισμούς του γένους *Escherichia* and *Shigella*
- Πραγματοποιήστε πολλαπλή στοίχιση με το λογισμικό CLUSTALW.
- Να γράψετε πρόγραμμα το οποίο διαβάζει τη στοίχιση και αναγνωρίζει τα μεταλλαγμένα κωδικόνια.
- Στην έξοδο να τυπώνει αρχείο που να περιέχει για τις θέσεις με μεταλλάξεις το original codon - mutated codon.

Σημείωση:

- Να χρησιμοποιήσετε ως αλληλουχία αναφοράς αυτήν του γονιδίου από το στέλεχος *E. coli* K12 substr. MG1655.
- Να χρησιμοποιήσετε την κλίμακα ταχύτητας μετάφρασης των κωδικωνίων (αρχείο sortscale) για την E. coli και να σχολιάσετε τα αποτελέσματά σας.

Εργασία 2. Μελέτη χαρακτηριστικών των πρωτεϊνών που αποτελούν γνωστούς στόχους αντικαρκινικών φαρμάκων.

Να ξεκινήσετε μελετώντας τους τύπους δεδομένων που περιλαμβάνονται στη διαδικτυακή βάση δεδομένων DrugBank (http://drugbank.ca/). Συγκεκριμένα, εντοπίστε τι δεδομένα που αφορούν αλληλουχίες πρωτεϊνών μπορείτε να ανακτήσετε από τη DrugBank άμεσα ή έμμεσα (δηλαδή από συνδεδεμένες βάσεις δεδομένων), και τους τρόπους με τους οποίους μπορείτε να τα ανακτήσετε (π.χ. κάνοντας download).

- Να ανακτήσετε το αρχείο της DrugBank (Full database) σε XML μορφή.
- Να γράψετε ένα πρόγραμμα που θα διαβάζει το XML αρχείο και θα τυπώνει σε ένα νέο αρχείο μια λίστα με τους κωδικούς της DrugBank που αντιστοιχούν σε αντικαρκινικά φάρμακα. Αναζητήστε στα πεδία "Categories" (όρος Antineoplastic) και "Indication" (όροι cancer, tumor, leukemia, melanoma).
- Να ανακτήσετε τις αμινοξικές αλληλουχίες όλων των πρωτεϊνών που είναι καταγεγραμμένοι ως στόχοι στη DrugBank (σε FASTA format).
- Να γράψετε ένα πρόγραμμα που θα διαβάζει το FASTA αρχείο και θα αντιγράφει σε ένα νέο αρχείο τις αλληλουχίες που αντιστοιχούν σε στόχους αντικαρκινικών φαρμάκων.
- Να χρησιμοποιήσετε το λογισμικό CAST (θα σας δοθεί από το διδάσκοντα) για να εντοπίσετε τις αλληλουχίες με ακραία αμινοξική σύσταση.
- Να γράψετε πρόγραμμα που υπολογίζει:

- Το ποσοστό των πρωτεϊνών που έχουν τουλάχιστον μια περιοχή ακραίας αμινοξικής σύστασης.
- Τη συχνότητα εμφάνισης περιοχών ακραίας αμινοξικής σύστασης ανά τύπο αμινοξικού καταλοίπου.
- Να συγκρίνετε τα παραπάνω με τα αντίστοιχα αποτελέσματα για τη UniRef50 τα οποία θα σας δώσει η κα Στέλλα Ταμανά.

Εργασία 3. Μελέτη των αλληλουχιών περιοχών διαμεμβρανικών πρωτεϊνών οι οποίες μερικώς διαπερνούν τη μεμβράνη.

Να ξεκινήσετε μελετώντας τους τύπους δεδομένων που περιλαμβάνονται στη διαδικτυακή βάση δεδομένων PDB_TM (http://pdbtm.enzim.hu/). Επικεντρωθείτε στην κατηγορία πρωτεϊνών που διαπερνούν τη μεμβράνη με α-έλικες. Η μορφοποίηση (format) των εγγραφών περιγράφονται στο http://pdbtm.enzim.hu/?m=manual.

- Να ανακτήσετε το αρχείο XML με όλες τις εγγραφές της PDB_TM που διαπερνούν τη μεμβράνη με α-έλικες
- Να ανακτήσετε το αρχείο FASTA χωρίς πλεονασμό (non-redundant) με τις αλυσίδες των εγγραφών της PDB_TM που διαπερνούν τη μεμβράνη με α-έλικες (προσοχή στα κενά που έχουν οι αλληλουχίες)
- Να γράψετε ένα πρόγραμμα το οποίο:
  - Θα διαβάζει τα αρχεία XML και FASTA.
  - Θα δημιουργεί δύο νέα FASTA αρχεία στα οποία θα αποθηκεύει (με μοναδικό κωδικό) αντίστοιχα την αλληλουχία κάθε περιοχής η οποία (α) μερικώς διαπερνά τη μεμβράνη (L: membrane embedded region not crossing the membrane) και (β) είναι διαμεμβρανική (H: alpha helix).

– Ο μοναδικός κωδικός να είναι της μορφής:

```
>PDBID:CHAINID:SEGMENTcode:Hydro:PropHelix
```

όπου PDBID και CHAINID ο κωδικός της εγγραφής PDB και της αντίστοιχης αλυσίδας, SEGMENTcode ο αύξων αριθμός του τμήματος και Hydro, PropHelix η μέση υδροφοβικότητα της περιοχής και η μέση τάση σχηματισμού α-ελίκων αντίστοιχα.

Σημείωση: Για τον υπολογισμό της μέσης υδροφοβικότητας να χρησιμοποιήσετε μια από τις κλίμακες που δίνοται στο http://blanco.biomol.uci.edu/hydrophobicity_scales.html, ενώ για τη μέση τάση σχηματισμού α-ελίκων την κλίμακα που περιγράφεται στο άρθρο http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1299714/?tool=pmcentrez.

Εργασία 4. Μελέτη των νουκλεοτιδικών περιοχών που κωδικοποιούν ομοπολυμερή πεπτίδια σε μονοκύτταρους οργανισμούς.

Να ξεκινήσετε μελετώντας τους τύπους δεδομένων που περιλαμβάνονται στις διαδικτυακές βάσεις δεδομένων:

- Ecogene (http://www.ecogene.org/), και
- SGD (http://www.yeastgenome.org/),

που αφορούν τους πρότυπους οργανισμούς *Escherichia coli* και *Saccharomyces cerevisiae* αντίστοιχα. Να επικεντρωθείτε στα δεδομένα που αφορούν νουκλεοτιδικές αλληλουχίες.

- Να ανακτήσετε το αρχείο με τις νουκλεοτιδικές αλληλλουχίες για το γονιδίωμα των δύο οργανισμών ή/και τις αλληλουχίες των γονιδίων που κωδικοποιούν πρωτεϊνικά μόρια σε μορφή FASTA.

- Να γράψετε ένα πρόγραμμα το οποίο:
  - Θα διαβάζει τον κατάλληλο γενετικό κώδικα για τον οργανισμό για τον οποίο γίνεται η ανάλυση. Χριιμοποιήστε τα δεδομένα από την ιστοσελίδα http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi
  - Θα διαβάζει τα αρχεία FASTA και θα εντοπίζει περιοχές οι οποίες κωδικοποιούν ομοπολυμερή (διαδοχικά συνώνυμα in-frame κωδικόνια) με μήκος μεγαλύτερο από k, όπου k παράμετρος που θα δίνεται από το χρήστη.
  - Θα υπολογίζει (α) το πλήθος των ομοπολυμερών ανά τύπο αμινοξικού καταλοίπου, και (β) την κατανομή των μηκών ανά τύπο και ανά οργανισμό.

Να σχολιάσετε τα αποτελέσματα, συγκρίνοντας τη μέση κατανομή του μήκους των γονιδίων από τους 2 οργανισμούς.

Εργασία 5. Μελέτη της (χαμηλής ευκρίνειας) 3D δομής γονιδιωμάτων μονοκύτταρων οργανισμών.

—

ή

Εργασία 6. Έλεγχος επίδοσεων παραλληλοποιήσιμου κώδικα ανάλυσης αμινοξικών αλληλουχιών.

Να ξεκινήσετε μελετώντας τις ιστοσελίδες:

- http://www.bioperf.org/
- http://www.ece.umd.edu/biobench/
- http://hpc.cs.tsinghua.edu.cn/research/cluster/pbb/index.html

που αναφέρονται στον έλεγχο επιδόσεων ποικιλίας εφαρμογών βιοπληροφορικής. Να

εστιάσετε στις εφαρμογές που αφορούν σύγκριση αλληλουχιών βιολογικών μακρομορίων και να εντοπίσετε τους ελέγχους οι οποίοι συνήθως πραγματοποιούνται.

- Θα πρέπει να συλλέξετε σύνολα δεδομένων αλληλουχιών σε μορφή FASTA, τα οποία έχουν χρησιμοποιηθεί σε έλεγχο επιδόσεων παράλληλου κώδικα σε εφαρμογές σύγκρισης αλληλουχιών/αναζήτησης ομοιότητας σε βάσεις δεδομένων.

- Θα συγγράψετε κώδικα ο οποίος θα μπορεί να διαβάζει αρχεία δεδομένων αλληλουχιών σε μορφή FASTA και θα δημιουργεί νέα αρχεία με shuffled αλληλουχίες ("τυχαίο ανακάτεμα").

- Θα πραγματοποιήσετε συγκριτικό έλεγχο επιδόσεων για λογισμικά μελέτης περιοχών χαμηλής πολυπλοκότητας για:

  – Δύο εκδόσεις του λογισμικού CAST (σειριακή και multithreaded).

  – Το λογισμικό SEG.

Να σχολιάσετε τα αποτελέσματά σας σε συνάρτηση με διαφορετικές υπολογιστικές πλατφόρμες.

Σημείωση: Τα σχετικά εκτελέσιμα αρχεία θα σας δοθούν από το διδάσκοντα.

Εργασία 7. Μελέτη της σύστασης του "παν-γονιδιώματος" (pan-genome) σε οργανισμούς βιοιατρικής ή/και βιοτεχνολογικής σημασίας.

Να ξεκινήσετε μελετώντας τους τύπους δεδομένων που περιλαμβάνονται στη διαδικτυακή βάση δεδομένων PlasmoDB (http://plasmodb.org/). Επικεντρωθείτε στα δεδομένα που αφορούν ορθόλογα γονίδια και δεδομένα αλληλουχιών γονιδίων και πρωτεϊνών. Εντοπίσετε τους τρόπους με τους οποίους μπορείτε να ανακτήσετε δεδομένα από αυτή τη βάση δεδομένων.

Από την PlasmoDB

- Να επιλέξετε "Identify genes by" - > "Evolution" -> "Orthology Phylogenetic Profile"
- Στη συνέχεια, στο "Select ortholog group profile" επιλέξτε τη συνομοταξία (order): "Haemosporida (HAEM)" .
- Να συλλέξετε το πλήρες σύνολο δεδομένων (μέσω της λειτουργίας Select columns) σε τοπικό αρχείο.

Να συγγράψετε πρόγραμμα το οποίο:

- Να υπολογίζει από κάθε νουκλεοτιδική ή αμινοξική αλληλουχία ένα διάνυσμα σύστασης.
- Να υπολογίζει για το pan-genome τη μέση νουκλεοτιδική ή αμινοξική σύσταση.
- Να υπολογίζει την ευκλείδια απόσταση του διανύσματος σύστασης κάθε αλληλουχίας από τη μέση σύσταση του pan-genome.

Να υπολογίζει τα αντίστοιχα μεγέθη για τα γονίδια του *Plasmodium falciparum* τα οποία δεν ανήκουν στο pan-genome.

Ομάδα Β:

Ελεύθερες εργασίες (επιλέγετε μόνοι σας εάν χρειάζεται συγγραφή κώδικα, χρήση διαδικτυακών ή άλλων εργαλείων):

Εργασία 1. Μελέτη χαρακτηριστικών πρωτεϊνών από υπερθερμόφιλους οργανισμούς.

Εργασία 2. Μελέτη χαρακτηριστικών μεταλλοπρωτεϊνών από βακτήρια.

Εργασία 3. Μελέτη χαρακτηριστικών πρωτεϊνών που διαπερνούν την εξωτερική μεμβράνη σε αρνητικά κατά Gram βακτήρια.