ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΛΟΓΙΑ
3η εργασία

Οδηγίες

Να παραδώσετε μέσω eclass, έναν συμπιεσμένο φάκελο zip με όνομα τον αριθμό μητρώου σας, που να περιέχει:

i. ένα **pdf ή doc** που θα περιέχει το ονοματεπώνυμό σας, τον αριθμό μητρώου σας και τις απαντήσεις σας στις ερωτήσεις του Μέρους Α και Γ
ii. ένα Perl script «**find_PFM.pl**» με τον κώδικά σας για το Μέρος Β
iii. το text αρχείο «**PFM.txt**» που δημιουργήσατε στο Μέρος Β
iv. το text αρχείο «**ten_dna_sequences.txt**», όπως αυτό σας δίνεται στο eClass.
v. το text αρχείο «**PFM2.txt**» που δημιουργήσατε στο Μέρος Γ

Για απορίες μπορείτε να χρησιμοποιήσετε τις Συζητήσεις, στο eClass του μαθήματος.

Μέρος Α (1,5 μονάδες)

Ύστερα από αναζήτηση στο internet και το υλικό του μαθήματος να απαντήσετε στις εξής ερωτήσεις:

Κατά τη μετατροπή ενός Position Probability Matrix (PPM) σε έναν Position Weight Matrix (PWM) οι τιμές αρχικά διαιρούνται με μία μεταβλητή b.

1. Στην πιο απλή περίπτωση η μεταβλητή b είναι ίση με 0.25 και για τις τέσσερις νουκλεοτιδικές βάσεις (A, T, C, G). Ποια παραδοχή κάνουμε όταν χρησιμοποιούμε την τιμή 0.25;
2. Αν οι υπό μελέτη αλληλουχίες βρίσκονται σε γονιδιωματική περιοχή όπου το “G-C content” είναι ίσο με 40% ποια/ες τιμή/ές για τη μεταβλητή b θα χρησιμοποιούσατε;

Μέρος Β (7,5 μονάδες)

Σας ζητείται να δημιουργήσετε κώδικα στη γλώσσα Perl (να τον αποθηκεύσετε ως “**find_PFM.pl**”) ο οποίος θα διαβάζει το text αρχείο “**ten_dna_sequences.txt**” που σάς δίνεται στο eClass. Στη συνέχεια, θα ελέγχει αν το text αρχείο περιέχει έγκυρες αλληλουχίες DNA (μόνο τα γράμματα A,T,C,G). Στην περίπτωση που τα περιεχόμενα του text αρχείου δεν είναι έγκυρα θα εκτυπώνει στη γραμμή εντολών μήνυμα λάθους και θα σταματάει, ενώ αν είναι έγκυρα θα υπολογίζει τον πίνακα Position Frequency Matrix (PFM) των αλληλουχιών αυτών.

Ο κώδικάς σας θα πρέπει να δημιουργεί ένα νέο text αρχείο με το όνομα “**PFM.txt**” με τα εξής περιεχόμενα:

1. τον πίνακα Position Frequency Matrix (PFM)
2. τη συναινετική αλληλουχία (consensus sequence), όπου σε κάθε θέση θα υπάρχει η πιο συχνή βάση, ενώ σε περίπτωση ισοψηφίας θα επιλέγετε μία από τις βάσεις.
3. Τη συναινετική αλληλουχία (consensus sequence), όπως στο προηγούμενο βήμα μόνο που τώρα σε περίπτωση ισοψηφίας θα εμφανίζετε όλες τις υποψήφιες βάσεις.

Παράδειγμα: η συνεναιτική αλληλουχία “AA|TCGC|A|G” υποδηλώνει ότι στη 2η θέση μπορεί να τοποθετηθεί είτε A είτε T. Στην τελευταία θέση μπορεί να τοποθετηθεί είτε C είτε A είτε T.

4. Το σκορ της συναινετικής αλληλουχίας.

Σημειώσεις:

1. Ο κώδικάς σας θα πρέπει να λειτουργεί δεχόμενος σαν input και άλλα text αρχεία που είναι παρόμοια με το “**ten_dna_sequences.txt”**. Δηλαδή, θα λειτουργεί με οποιοδήποτε text αρχείο περιέχει DNA αλληλουχίες, οι οποίες είναι μία σε κάθε γραμμή και έχουν το ίδιο μήκος.
2. Προτείνεται η χρήση δυσδιάστατου hash (two-dimensional hash) για την υλοποίηση της άσκησης, βλ. διάλεξη 7, διαφάνεια 36.
3. Στο eClass (φάκελος «Υλικό Διάλεξης 7») έχει αναρτηθεί η λύση ενός μέρους της άσκησης και μπορείτε να τη χρησιμοποιήσετε. Δώστε προσοχή στη χρήση της μεταβλητής $i.
4. Στον πίνακα Position Frequency Matrix (PFM) που θα εκτυπώσετε θα πρέπει να εκτυπώνετε το μηδέν, όπου η αντίστοιχη θέση είναι κενή.
5. Η σειρά με την οποία παρουσιάζονται οι 4 γραμμές του πίνακα PFM οι οποίες αντιστοιχούν στις βάσεις A, C, G, T μπορεί να είναι τυχαία.
6. Παρακάτω μπορείτε να δείτε ένα παράδειγμα του αρχείου “PFM.txt” που σας ζητείται, το οποίο υπολογίστηκε από άλλες input αλληλουχίες:

```
A    0   4   2   2   1   2   2   2   3   2   2   0   3   1   3
C    0   2   3   2   5   2   4   3   3   2   4   3   3   2   3
T    0   2   2   3   1   5   2   2   1   5   3   1   2   4   2
G   10   2   3   3   3   1   2   3   3   1   1   6   2   3   2
GACTCTCCATCGATA
GAC|GT|GCTCC|GA|C|GTCGA|CTA|C
Score=  65
```

Μέρος Γ (1 μονάδα)

Στη βάση δεδομένων JASPAR (https://jaspar.genereg.net/), και συγκεκριμένα στη συλλογή JASPAR CORE υπάρχουν αποθηκευμένα μοτίβα που αναγνωρίζονται από μεταγραφικούς παράγοντες και τα οποία έχουν αποδειχθεί με πειραματικές τεχνικές, όπως ChIP-seq. Για κάθε μοτίβο είναι αποθηκευμένες πληροφορίες όπως ο οργανισμός, το όνομα του μεταγραφικού παράγοντα, links προς άλλες βάσεις δεδομένων, ο πίνακας PFM “Frequency matrix” και διάφορα αρχεία που μπορεί να κατεβάσει ο χρήστης.

Το πιο πρόσφατο μοτίβο που έχει προταθεί για τον μεταγραφικό παράγοντα GATA1 έχει κωδικό MA0035.4 και μπορείτε να το δείτε εδώ: https://jaspar.genereg.net/matrix/MA0035.4/. Σας δίνεται ένα text αρχείο με το όνομα “**MA0035.4.sites.txt**” το οποίο περιέχει τις αλληλουχίες από το αντίστοιχο πείραμα ChIP-seq, τροποποιημένο για τους σκοπούς της παρούσας εργασίας.

1. Να εκτελέστε τον κώδικα που δημιουργήσατε στο μέρος Β, χρησιμοποιώντας ως αρχείο εισόδου το αρχείο “**MA0035.4.sites.txt**”, και επιβεβαιώστε ότι ο πίνακας PFM που υπολογίζετε είναι ίδιος με αυτόν στο site της JASPAR (βλ. παρακάτω εικόνα). Αποθηκεύστε τον πίνακα PFM σε ένα text αρχείο με το όνομα “**PFM2.txt**”.

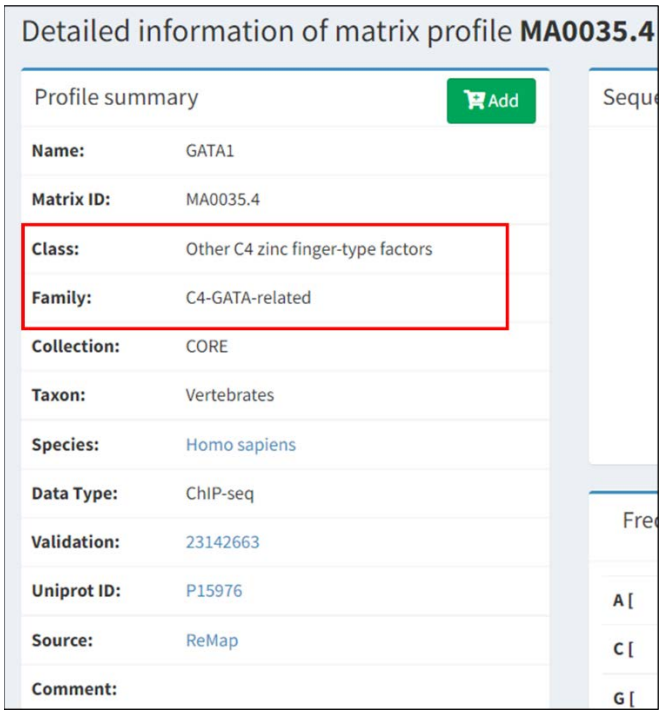
Frequency matrix

JASPAR | TRANSFAC | MEME | RAW PFM | Reverse comp.

A [	22209	17328	3953	1314	49692	67550	2206	2567	7397	26545	22656	]
C [	12209	14489	62419	595	710	1292	1238	65937	3025	11186	15261	]
G [	13955	11088	2712	652	856	988	618	1471	765	14358	9325	]
T [	23455	28923	2744	69267	20570	1998	67766	1853	60641	19739	24586	]

Bonus ερωτήσεις (1,5 μονάδες)

Για αυτό τον μεταγραφικό παράγοντα, GATA1, στη βάση JASPAR δίνεται η πληροφορία ότι ανήκει στην Κλάση (Class) μεταγραφικών παραγόντων "**Other C4 zinc finger-type factors**" και στην οικογένεια (Family) "**C4-GATA-related**" (βλ. εικόνα παρακάτω).



1. Ποιο σύστημα ταξινόμησης χρησιμοποιεί η JASPAR για να ταξινομήσει τους μεταγραφικούς παράγοντες; Να αναφέρετε με ποιο κριτήριο ταξινομούνται οι μεταγραφικοί παράγοντες στις διάφορες ομάδες, σε ποια δημοσίευση (link/authors) παρουσιάζεται αυτό το σύστημα ταξινόμησης και σε ποιο site μπορείτε να αναζητήσετε τις διάφορες ταξινομικές ομάδες και τα μέλη τους.
2. Τι σημαίνει η ονομασία "**Other C4 zinc finger-type factors**", που είναι το όνομα της Κλάσης του GATA1;
3. Σε ποια Υπερκλάση (Superclass) ανήκει ο μεταγραφικός παράγοντας GATA1;