**Project Python Report - Thodoris Danis & Sofia Kampaki**

1. Προεργασία, αρχείο GENCODE

Αρχικά, διαβάζουμε μέσω pandas το αρχείο τύπου gff3 (GENCODE) και ορίζουμε τα ονόματα των στηλών σύμφωνα με την εκφώνηση του προτζεκτ. Αφαιρούμε όλες τις καταχωρήσεις που έχουν για source το HAVANA και επίσης αφαιρούμε τις στήλες που δεν μας χρειάζονται για τους σκοπούς των ερωτήσεων. Αυτές είναι οι source, score, frame και strand.  
  
Η στήλη attribute όμως περιέχει διάφορες πληροφορίες που μας είναι χρήσιμες σαν μορφή νέων στηλών, συνεπώς δημιουργούμε συναρτήσεις οι οποίες φτιάχνουν νέες στήλες στο αρχείο gencodeexin από την στήλη attribute. Αυτές είναι οι εξής:

* get\_transcript : για το transcript id
* get\_tag : για να πάρουμε το tag το οποίο χαρακτηρίζει την προτεραιότητα για τα μετάγραφα συμφωνα με την appris
* get\_genename: για το όνομα του γονιδίου
* get\_geneid: για το gene id

Καλώντας αυτές τις συναρτήσεις με apply στο dataframe genecodexin δημιουργούμε τις νέες στήλες.  
  
Για τις ερωτήσεις που ακολουθούν στο notebook κάνουμε το εξής:  
Για κάθε ερώτηση ψάχνουμε με regex λέξεις κλειδιά για να ξέρουμε τι ακριβώς πρέπει να απαντηθεί και “απαιτούμε” το όνομα του γονιδίου, μετάγραφου κλπ να είναι μέσα σε ‘...’

**Ερώτηση 1η: "Which gene is upstream to gene <Γονίδιο> ?" Which gene is downstream?", "Which gene is closest?"**

**Χρησιμοποιήστε το GENCODE.**

Με την βοήθεια του νέου dataset είναι πολύ εύκολο να βρουμε λόγω των θέσεων που βρίσκονται στις στήλες start, end το γονίδιο που είναι upstream/downstream/closest και από την στήλη genename που φτιάξαμε να πάρουμε το όνομα του.

**Ερώτηση 2.How many transcripts does the gene <Γονίδιο> have?.**

Ψάχνουμε τις λέξεις many και transcripts για να πάρουμε την πληροφορία με βάση το γονίδιο που δόθηκε στην ερώτηση και μετά μετράμε τα μετάγραφα που έχει το συγκεκριμένο γονίδιο.

**Ερώτηση 3."What is in <Θέση> ? (GENCODE)**

Αρχικά χωρίζουμε με ένα flag για το αν η θέση που ζήτησε ο χρήστης για να μάθει πληροφορίες είναι intergenic η όχι.  
Aν δεν είναι τότε συλλέγει όλες τις απαραίτητες πληροφορίες για την θέση του νουκλεοτιδίου σε ποιο εξόνιο ποιου γονιδίου είναι και τι υπάρχει downstream & upstream.  
Έπειτα μέσω του sequence\_region της Ensembl τυπώνει την ακολουθία του γονιδιώματος για +50 και -50 θέσεις από την θέση που έβαλε ο χρήστης.  
Αν είναι intergenic τυπώνει ότι είναι intergenic και επίσης τα upstream/downstream genes.

**Ερώτηση 4. "What is the name of the gene <Γονίδιο> ?"**

Συμπληρώνοντας το πεδίο ‘fields’ με τι κατάλληλες παραμέτρους (name, other\_names) αλλά και χρησιμοποιώντας το json αρχείο που λαμβάνουμε, μπουρούμε και εξάγουμε την πληροφορία που μας ζητείται.

**Ερώτηση 5. "Where is the location of <Γονίδιο, μετάγραφο ή μετάλλαξη> ?"**

Στην ερώτηση αυτή αφού πρώτα καταλήξουμε αν χρήστης αναφέρεται σε γονίδιο, μετάγραφο ή μετάλλαξη , τότε η υλοποίηση μας ανατρέχει στην κατάλληλη βάση δεδομένων για να επιστρέψει την πληροφορία που ζητάει ο χρήστης.

**Ερώτηση 6. "Which are the refseq transcripts of the gene <Γονίδιο> ?"**

Με το πεδίο ‘refseq’ αποκτάς την κατάληλλη πληροφορία για το ερωτηθέν γονίδιο ωστόσο για να χαρακτηριστεί primary χρησιμοποιήθηκε η appris, βασιζόμενοι στο σύστημα κατάταξης της και αφού παραλείφθηκαν γονίδια τα οποία δεν είχαν cds.

**Ερώτηση 7. "How many exons does <Γονίδιο ή Μετάγραφο> have ?"**

Η πληροφορία αυτή αποκτήθηκε αφού διαχωρίζεται το γονίδιο από το μετάγραφο της ερώτησης του χρήστη και της ανάθεσης του πεδίου ‘exons’ στις παραμέτρους του ‘fields’ στο σύνολο των παραμέτρων του request.

**Ερώτηση 8. "Where is the location of the CDS of the <Γονίδιο ή Μετάγραφο>?"**

**Ερώτηση 9. "What is the primary transcript of the gene <Γονίδιο> ?"**

Το primary trascript αποκτάται μέσω της πληροφορίας που παρέχει η appris.

**Ερώτηση 10. "In which gene does <Μετάγραφο, Μετάλλαξη ή θέση> belong"?**

Μέσω της snpeff της myvariant.info και genomic\_pos, symbol\_name για την κάθε περίπτωση, μπορεί να αποκτηθεί το επίσημο όνομα γονιδίου που ζητάται από τον χρήστη.

**Ερώτηση 11. "In which pathways is the gene <Γονίδιο> involved ?"**

Αφού κάνουμε το κατάλληλο parsing στο πεδίο pathway που θα ορίσουμε στα fields τότε επιστρέφουμε όλες τις βάσεις δεδομένων που σχετίζονται με βιοχημικά μονοπάτια.

**Ερώτηση 12. "Which proteins are transcribed from the <Γονίδιο/Μετάγραφο> ?"**

Μέσω του πεδίου refseq της mygene.info έγινε η αντιστοίχηση μεταγράφων και πρωτεινών και έπειτα ο χαρακτηρισμός σε primary transcript μέσω της appris.

**Ερώτηση 13. "What is the function of the gene <Γονίδιο> ?"**

Το πεδίο summary (‘fields’: ‘summary’) είναι αυτό που περιέχει την πληροφορία για την λειτουργία του γονιδίου.

**Ερώτηση 14. "Which organisms have a homologous gene to <Γονίδιο> ?"**

Αρχικά δημιουργήθηκε ένα νέο αρχείο με τους ήδη χαρακτηρισμένους οργανισμούς. Έπειτα γίνεται σύγκριση των ids που αντιστοιχούν στο υπό μελέτη γονίδιο με τα ήδη καταχωρημένα στο νέο αρχείο που κατασκευάστηκε ώστε να αποκτηθούν τα επιστημονικά ονόματα των οργανισμών.

**Ερώτηση 15. "What type is the gene <Γονίδιο> ?"**

Το πεδίο ‘type\_of\_gene’ στο fields των παραμέτρων του request είναι αυτό που χρησιμοποιήσαμε ώστα να επιστρέψουμε την επιθυμητή πληροφορία στον χρήστη.

**Ερώτηση 16. "What is the sequence in chromosome X between 1000000 and 1000100;"**

Μέσω της ensembl επιστρέφουμε την νουκλεοτιδική αλληλουχία για το χρωμόσωμα αλλά και το εύρος των βάσεων που ζητάει ο χρήστης.

**Ερώτηση 17. "What is the allele frequency of <Μετάλλαξη> ?"**

Αφού γίνει το κατάλληλο parsing στα πεδία 1000g και esp επιστρέφονται οι αλληλικές συχνότητες, εφόσον υπάρχουν στον χρήστη.

**Ερώτηση 18. "With which conditions is the mutation <Μετάλλαξη> associated ? "**

Οι κλινικές συνθήκες που σχετίζεται κάθε γονίδιο μπορούν να βρεθούν απο την βάση clinvar και το πεδίο rcv.

**Ερώτηση 19. "What is the rs-id of <Μετάλλαξη> ?"**

Υπάρχει ειδικό πεδίο για αυτό το είδος της ερώτησης ‘dbsnp.rsid’ .

**Ερώτηση 20. "What is the effect of <Μετάλλαξη> ?"**

Οι βάσεις που χρησιμοποιήθηκαν στο ερώτημα αυτό είναι h snpeff και τα διάφορα πεδία της cadd.

Αξίζει να σημειωθεί ότι κάθε βάση που χρησιμοποιήθηκε για το κάθε ένα από τα παραπάνω ερωτήματα δεν είχε κάποιο ομογενή και ολοκληρωμένο τρόπο παρουσίασης της πληροφορίας δυσκολεύοντας αρκετά χρησιμότητα της. Επίσης, η απουσία αρκετών πληροφοριών για αρκετά πεδία γονιδίων, μεταγράφων και μεταλλάξεων δημιουργεί αρκετά προβλήματα στην επικύρωση και την απόκτηση της επιθυμητης πληροφορίας.