**Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών "Πληροφορική Επιστημών Ζωής" (ΠΕΖ)**

**Μάθημα «Εισαγωγή στη Σχεδίαση, Υλοποίηση και Ανάκτηση Πληροφορίας από Βιοϊατρικές Βάσεις Δεδομένων»**

**Εργασία Ακ. Έτος 2016-2017**

**Περίληψη**

Ο σκοπός της εργασίας είναι να σχεδιαστεί και να υλοποιηθεί μια σχεσιακή βάση δεδομένων που θα αποθηκεύει βιολογικές πληροφορίες σχετικά με κάποια συγκεκριμένη ασθένεια. Στόχος της βάσης δεδομένων είναι να υποστηρίξει μια βιολογική έρευνα που αφορά τη συγκεκριμένη ασθένεια.

**Προδιαγραφές**

Η βάση δεδομένων θα αποθηκεύει δεδομένα σχετικά με τη βιολογική πληροφορία μιας συγκεκριμένης ασθένειας. Η βιολογική πληροφορία αφορά τα γονίδια που σχετίζονται με την ασθένεια, τις παραγόμενες πρωτεΐνες βάση αυτών των γονιδίων, τις λειτουργίες και τις ισομορφές των πρωτεϊνών.

Ειδικότερα, για κάθε γονίδιο θα αποθηκεύονται τα εξής:

* Σύμβολο Γονιδίου (Gene Symbol) σύμφωνα με την HUGO Gene Nomenclature Committee.
* Όνομα Γονιδίου.
* Το σκορ που δείχνει τη σχετικότητα με την ασθένεια σύμφωνα με τη DisGenNet.
* Πλήθος δημοσιεύσεων (# of PMIDs – PubMed IDs) που υποστηρίζουν τη σχετικότητα του γονιδίου με την ασθένεια.

Υποθέτουμε ότι κάθε γονίδιο σχετίζεται με τουλάχιστον μία πρωτεΐνη και για κάθε πρωτεΐνη αποθηκεύονται τα εξής:

* Κωδικός πρωτεΐνης σύμφωνα με την βάση Uniprot.
* Όνομα πρωτεΐνης.
* Πειραματική επιβεβαίωση της πρωτεΐνης.
* Σχετιζόμενο γονίδιο.

Επιπρόσθετα, κάθε πρωτεΐνη σχετίζεται με κάποιες λειτουργίες σύμφωνα με την Gene Ontology (GO). Για κάθε λειτουργία θα πρέπει να αποθηκεύονται να εξής:

* Όρος λειτουργίας σύμφωνα με την GO (GO Term). Ο όρος της λειτουργίας είναι μοναδικός για κάθε λειτουργία.
* Όνομα λειτουργίας.
* Είδος λειτουργίας.

Επίσης, κάθε πρωτεΐνη σχετίζεται με κάποιες ισομορφές. Μια ισομορφή ανήκει σε μια μόνο πρωτεΐνη. Για κάθε ισομορφή θα πρέπει να αποθηκεύονται να εξής:

* H ακολουθία της ισομορφής σε μορφοποίηση FASTA.

**Συγκέντρωση Δεδομένων**

Για τη συγκέντρωση των δεδομένων της άσκησης θα χρησιμοποιηθούν τρεις δικτυακές βάσεις δεδομένων βιολογικής πληροφορίας:

* Η **DisGenNet** (http://www.disgenet.org/) που περιέχει πληροφορίες για τη συσχέτιση ασθενειών και γονιδίων.
* Η **Hint-KB 2** (http://hintkb.ceid.upatras.gr/) που περιέχει πληροφορία για πρωτεΐνες και τις αλληλεπιδράσεις τους.
* Η **UniProt** (http://www.uniprot.org/) που περιέχει ολοκληρωμένη πληροφορία για πρωτεΐνες.

Η συγκέντρωση των δεδομένων θα ξεκινήσει από την DisGenNet η οποία περιέχει φόρμα αναζήτησης σύμφωνα με κάποια ασθένεια. Στη σελίδα του DisGenNet πάμε στην επιλογή “Search”, επιλέγετε το disease πάνω από το κουτί αναζήτησης και αναζητείτε βάση της ασθένειας. Στα αποτελέσματα πηγαίνετε στο πλαίσιο «Top 10 gene associations for this disease» και επιλέγεται το «Browse details….». **Ένας αναλυτικός πίνακας με όλη τη πληροφορία (και όχι μόνο) που χρειάζεστε εμφανίζεται στο πλαίσιο (tab) «Summary of Associated Genes»**.

Από τον πίνακα αποτελεσμάτων της DisGenNet (στο tab genes) παίρνουμε για κάθε γονίδιο τη σχετιζόμενη πρωτεΐνη (στήλη uniprot). Στη Hint-KB 2 επιλέγουμε από το μενού την επιλογή «By Protein» και αναζητούμε σύμφωνα με τον Uniprot κωδικό της πρωτεΐνης. Στα αποτελέσματα στο πλαίσιο (tab) “Protein’s Functions” υπάρχει η λίστα με τις λειτουργίες της πρωτεΐνης.

Επίσης με τον Uniprot κωδικό της πρωτεΐνης μπορούμε να κάνουμε αναζήτηση στη βάση δεδομένων Uniprot όπου θα βρούμε το όνομα της πρωτεΐνης (Protein), αν υπάρχει πειραματική επιβεβαίωση ύπαρξης της πρωτεΐνης (Status) και τις ισομορφές της (στην ενότητα Sequence). Για να υπάρχει πειραματική επιβεβαίωση πρέπει η τιμή «Annotation score» να είναι «Experimental evidence at protein level» αλλιώς δεν θεωρείτε επιβεβαιωμένη. Για κάθε ισομορφή πέρα από τον identifier μπορείτε να δείτε και τη FASTA ακολουθία πατώντας την επιλογή FASTA.

**Ερώτημα 1:**

Σχεδιάστε το διάγραμμα Οντοτήτων-Συσχετίσεων για την παραπάνω Βάση Δεδομένων. Κρίνεται μόνοι σας και δικαιολογήστε τους περιορισμούς και τους τύπους δεδομένων που θα χρησιμοποιήσετε.

Για το διάγραμμα ΟΣ μπορείτε να χρησιμοποιήσετε κάποια σχεδιαστική εφαρμογή όπως MS WORD, MS Powerpoint, MS Vidio, OpenOffice/LibreOffice, Dia, κτλ. ή να την σχεδιάσετε στο χαρτί και να την σκανάρετε ώστε να μπει στην αναφορά σας. Στη δεύτερη περίπτωση ο σχεδιασμός, τα σχήματα και τα γράμματα πρέπει να είναι ξεκάθαρα.

**Ερώτημα 2:**

Δημιουργήστε την βάση δεδομένων σε MySQL μέσω του εργαλείου phpMyAdmin. Ορίσετε τους πίνακες και τους περιορισμούς – συσχετίσεις.

Ανάλογα με το άθροισμα των ψηφίων του ΑΜ στο μεταπτυχιακό θα επιλέξετε να συγκεντρώσετε δεδομένα για μια από τις παρακάτω ασθένειες:

* Aν το **άθροισμα των ψηφίων του ΑΜ σας είναι μονός αριθμός** τότε αναζητείτε για «Alzheimer's Disease» και επιλέγεται τα γονίδια με score, όπως δίνεται από την DisGenNet, μεγαλύτερο του 0.3.
* Aν το **άθροισμα των ψηφίων του ΑΜ σας είναι ζυγός αριθμός** τότε αναζητείτε για «Parkinson's Disease» και επιλέγεται τα γονίδια με score όπως δίνεται από την DisGenNet, μεγαλύτερο του 0.25.

*Σημείωση 1: Σε περίπτωση που δεν έχετε AM τότε το άθροισμα είναι 0 (Ζυγός).*

*Σημείωση 2: Επειδή κάποιες πρωτεΐνες έχουν πολλές ισομορφές, στη ΒΔ θα αποθηκεύσετε το μέγιστο τρεις ισομορφές για κάθε πρωτεΐνη.*

Αφού συγκεντρώσετε τα δεδομένα, αποθηκεύστε τα στη ΒΔ που φτιάξατε. Η διαδικασία που θα ακολουθήσετε για την **συγκέντρωση και εισαγωγή των δεδομένων στη ΒΔ αφήνετε στην επιλογή σας**, αλλά θα πρέπει να αναφερθεί και να αναλυθεί στην αναφορά σας. Προτείνονται οι παρακάτω διαδικασίες:

* Συγκέντρωση χειροκίνητα της πληροφορίας σε excel αρχεία και αποθήκευση στη ΒΔ με python script παρόμοιο με αυτό που διδαχτήκατε.
* Συγκέντρωση χειροκίνητα της πληροφορίας σε excel αρχεία και χρήση της επιλογής Import του phpMyAdmin.
* Συγκέντρωση χειροκίνητα της πληροφορίας σε excel αρχεία και χειροκίνητη αποθήκευση στη ΒΔ δημιουργώντας SQL ερωτήματα.

Για όσους θέλουν να χρησιμοποιήσουν τις υπηρεσίες ιστού που έχουν οι ανοικτές βάσεις δεδομένων μπορούν να διαβάσουν τους παρακάτω συνδέσμους:

1. http://www.uniprot.org/help/programmatic\_access

2. http://www.disgenet.org/web/DisGeNET/menu/rdf#releaseInfo

3. http://hintkb.ceid.upatras.gr/hintkb/page/services

Παραδοτέο είναι ο SQL κώδικας για την **δημιουργία πινάκων και περιορισμών ξένου κλειδιού** και η διαδικασία **εισαγωγής δεδομένων**.

**Ερώτημα 3α (για φοιτητές με πτυχίο σε επιστήμες της υγείας ή πτυχίο που δεν ανήκει στις επιστήμες πληροφορικής):**

Συντάξτε τα κατάλληλα SQL ερωτήματα που θα επιστρέφουν την κατάλληλη πληροφορία για κάθε μια από τα παρακάτω αιτήματα

**Αίτημα 1** Tο όνομα, το σύμβολό και το πλήθος δημοσιεύσεων των γονιδίων που είναι αποθηκευμένα στη ΒΔ σε αύξουσα σειρά σύμφωνα με τον κωδικό της πρωτεΐνης.

**Αίτημα 2** Τον κωδικό πρωτεΐνης και το πλήθος των σχετιζόμενων ισομορφών για κάθε πρωτεΐνη.

**Αίτημα 3** Τον κωδικό και όνομα των πρωτεΐνων που έχουν επιβεβαιωθεί πειραματικά.

**Αίτημα 4** Το όνομα, το σύμβολο και το σκορ σχετικότητας με την ασθένεια για τα γονίδια που το σκορ συσχέτισης με την ασθένεια είναι μεγαλύτερο του 0.35 σε κατάταξη σύμφωνα με το σκορ (το μεγαλύτερο σκορ πρώτο).

**Αίτημα 5** Τον κωδικό πρωτεΐνης και το όνομα της κάθε σχετιζόμενης λειτουργίας.

**Αίτημα 6** Τον κωδικό και το όνομα των πρωτεϊνών που έχουν λειτουργία που ονομάζεται «protein binding».

**Αίτημα 7** Τον κωδικό και το πλήθος των ισομορφών των πρωτεϊνών των οποίων το σχετιζόμενο γονίδιο έχει πάνω από 150 σχετικές με την ασθένεια δημοσιεύσεις.

**Αίτημα 8** Το όνομα του γονιδίου και το όνομα της αντίστοιχης πρωτεΐνης για την οποία υπάρχει κάποια ισομορφή που περιέχει την ακολουθία «EEG» (για τη Parkinson's Disease) ή «ALL» (για Alzheimer's Disease).

**Αίτημα 9** Τον όρο και το όνομά των λειτουργιών που σχετίζονται με πρωτεΐνες που ο κωδικός πρωτεΐνης τους ξεκινάει από «P» και η λειτουργία ανήκει στην κατηγορία «molecular\_function».

**Αίτημα 10** Το σύμβολο του γονιδίου και το πλήθος των ισομορφών της αντίστοιχης πρωτεΐνης που περιέχουν την ακολουθία «EEG» (για τη Parkinson's Disease) ή «ALL» (για Alzheimer's Disease).

Παραδοτέο είναι ο SQL κώδικας των ερωτημάτων.

**Ερώτημα 3β (για φοιτητές με πτυχίο σε επιστήμες πληροφορικής):**

Δημιουργήστε Python Script (GEO\_parse.py) που θα συνδέεται με την βάση δεδομένων Gene Expression Omnibus, θα κατεβάζει το σύνολο δεδομένων GDS2519.soft και θα γράφει τα περιεχόμενά του σε ένα tab delimited αρχείο. Για αυτό τον σκοπό θα χρησιμοποιήσετε το Module GEOparse της python όπως φαίνεται στο παρακάτω υπόδειγμα:

import GEOparse

gds = GEOparse.get\_GEO(geo="GDS2519")

gds.table.to\_csv('gds2619.tsv', '\t')

Στη συνέχεια το script σας θα πρέπει να δημιουργεί πίνακα gene\_expression στην βάση σας με τόσα πεδία όσες και οι στήλες του αρχείου που δημιουργήσατε. Τα ονόματα των πεδίων θα είναι αυτά που εμφανίζονται στην πρώτη γραμμή του αρχείου που δημιουργήσατε. Το όνομα του πρώτου πεδίου (το οποίο θα είναι και το πρωτεύων κλειδί του πίνακα) θα είναι το id.

Έπειτα το script σας με την εκτέλεση επαναληπτικά της κατάλληλης εντολής insert θα εισαγάγει στον πίνακα που δημιουργήσατε την πληροφορία από το αρχείο γραμμή-γραμμή.

Παρουσιάστε το καινούργιο διάγραμμα ΟΣ που περιλαμβάνει και τον καινούργιο πίνακα.

Δημιουργήστε τα παρακάτω Python script:

**Script 1. getGenesPublications.py** που θα τυπώνει στην οθόνη του χρήστη το όνομα, το σύμβολό και το πλήθος δημοσιεύσεων των γονιδίων που είναι αποθηκευμένα στη ΒΔ σε αύξουσα σειρά σύμφωνα με τον κωδικό της πρωτεΐνης.

**Script 2. getProteinsIsoformsCount.py** που θα τυπώνει στην οθόνη του χρήστη τον κωδικό πρωτεΐνης και το πλήθος των σχετιζόμενων ισομορφών για κάθε πρωτεΐνη.

**Script 3. getProteinExperimentalVerified.py** που θα τυπώνει στην οθόνη του χρήστη τον κωδικό και όνομα των πρωτεΐνων που έχουν επιβεβαιωθεί πειραματικά.

**Script 4. getGeneByScore.py –s «σκορ»** που θα τυπώνει στην οθόνη του χρήστη το όνομα, το σύμβολο και το σκορ σχετικότητας με την ασθένεια για τα γονίδια που το σκορ συσχέτισης με την ασθένεια είναι μεγαλύτερο του Χ (όπου Χ το σκορ που δίνεται ως όρισμα) σε κατάταξη σύμφωνα με το σκορ (το μεγαλύτερο σκορ πρώτο)

**Script 5. getProteinFunctions.py** που θα τυπώνει στην οθόνη του χρήστη τον κωδικό πρωτεΐνης και το όνομα της κάθε σχετιζόμενης λειτουργίας.

**Script 6. getProteinsByFunction.py –f «λειτουργία»** που θα τυπώνει στην οθόνη του χρήστη τον κωδικό και το όνομα των πρωτεϊνών που έχουν λειτουργία που ονομάζεται Χ (όπου Χ η λειτουργία που δίνεται ως όρισμα).

**Script 7. getProteinIsoformCountByPubsCount.py –p «αριθμός»** που θα τυπώνει στην οθόνη του χρήστη τον κωδικό και το πλήθος των ισομορφών των πρωτεϊνών των οποίο το σχετιζόμενο γονίδιο έχει πάνω από X σχετικές με την ασθένεια δημοσιεύσεις (όπου Χ ο αριθμός που δίνεται ως όρισμα).

**Script 8. getGeneByIsoformSequence.py –s «ακολουθία»** που θα τυπώνει στην οθόνη του χρήστη το όνομα του γονιδίου και το όνομα της αντίστοιχης πρωτεΐνης για την οποία υπάρχει κάποια ισομορφή που περιέχει την ακολουθία X (όπου Χ η ακολουθία που δίνεται ως όρισμα).

**Script 9. getProteinsFunctionByNamespace.py –n «κατηγορία λειτουργίας»** που θα τυπώνει στην οθόνη του χρήστη τον όρο και το όνομά των λειτουργιών που σχετίζονται με πρωτεΐνες που ο κωδικός πρωτεΐνης τους ξεκινάει από «P» και η λειτουργία ανήκει στην κατηγορία Χ (όπου Χ η κατηγορία λειτουργίας που δίνεται ως όρισμα).

**Script 10. getGenesIsoformsCountBySequence.py –s «ακολουθία»** που θα τυπώνει στην οθόνη του χρήστη το σύμβολο του γονιδίου και το πλήθος των ισομορφών της αντίστοιχης πρωτεΐνης που περιέχουν την ακολουθία X (όπου Χ η ακολουθία που δίνεται ως όρισμα).

**Script 11.** **getGenesMaximumExpression.py** που θα τυπώνει στην οθόνη του χρήστη την το gene\_symbol με την μεγαλύτερη γονιδιακή έκφραση στο πείραμα GSM153425.

**Script 12. GetGeneExpressionsByUniprot.py –s «uniprot\_id»** που θα τυπώνει στην οθόνη τα gene\_symbols των γονιδίων που αντιστοιχούν με το uniprot\_id που δίνει ο χρήστης και τις γονιδιακές εκφράσεις τους από τον πίνακα gene\_expression.

**Ερώτημα 4:**

..Ετοιμάστε μία παρουσίαση των 5 λεπτών πάνω στην δημοσίευση που έχετε αναλάβει από την λίστα που ανακοίνωσε ο κύριος Κλεφτογιάννης. Για αυτό το ερώτημα επιτρέπεται συνεργασία μεταξύ των 2 ατόμων που θα αναλάβουν την ίδια δημοσίευση Οι φοιτητές με πτυχίο σε επιστήμες πληροφορικής θα πρέπει να παρουσιάσουν την υπολογιστική μέθοδο και την τεχνολογία βάσης δεδομένων που παρουσιάζεται στην δημοσίευση που έχουν αναλάβει. Οι υπόλοιποι θα πρέπει να παρουσιάσουν το βιολογικό πρόβλημα που επιλύεται στην βάση δεδομένων/εργαλείο που παρουσιάζεται στην δημοσίευση καθώς καιτην βιολογική πληροφορία που μπορεί να εξαχθεί (αλλά και τον τρόπο για να γίνει αυτό).

Παραδοτέο είναι ο SQL κώδικας των ερωτημάτων και ο κώδικας Python.

**Παραδοτέα**

Το παραδοτέο της εργασίας θα είναι μια αναφορά σε ηλεκτρονική επεξεργάσιμη μορφή (doc, docx, odt) με τα παρακάτω:

* Το σχεδιάγραμμα οντοτήτων-συσχετίσεων (Ερώτημα 1).
* Τα SQL ερωτήματα για την δημιουργία της βάσης, των πινάκων και των περιορισμών (Ερώτημα 2).
* Τα SQL ερωτήματα του ερωτήματος 3α συμπεριλαμβανομένου τα ερωτήματα εισαγωγής και τα ερωτήματα ανάκτησης.
* Τα SQL ερωτήματα και τον κώδικα Python για κάθε script του ερωτήματος 3β
* Παρουσίαση σε Powerpoint της εργασίας σας για την προφορική εξέταση (**Η παρουσίαση δεν θα πρέπει να ξεπερνά τα 10 λεπτά-Το χρονικό όριο είναι αυστηρό και δεν θα ξεπεραστεί**)

Για την εργασία σας θα δουλέψετε στον MySQL Server στην διεύθυνση <https://pez.insybio.com>. Θα φτιάξετε μια νέα βάση δεδομένων και θα εκτελέσετε τα απαραίτητα SQL ερωτήματα μέσα από το phpMyAdmin περιβάλλον στη <https://pez.insybio.com>/phpmyadmin/ με τους κωδικούς που έχουν δοθεί στον καθένα. Η βάση που θα φτιάξετε θα είναι και αυτή μέρος των παραδοτέων σας.

Για όσους χρειάζεται να δουλέψετε σε python, μπορείτε να χρησιμοποιήσετε τους λογαριασμούς σας στον server python.insybio.com με τον τρόπο που δουλέψατε στην εργαστηριακή άσκηση ΙΙ (χρήση λογισμικού notepad++ και Putty)

Σε περίπτωση που έχετε κάποιο πρόβλημα με την σύνδεση με τη ΒΔ παρακαλώ επικοινωνήστε με την κυρία Αίγλη Κορφιάτη (a.korfiati@insybio.com).

Για οποιαδήποτε απορία μπορείτε να επικοινωνήστε μέχρι και τέσσερεις μέρες πριν την ημερομηνία παράδοσης με τους διδάσκοντες: Σ. Μαυρουδή (mavroudi@teiwest.gr), Χ. Αλεξάκος ([c.alexakos@insybio.com](mailto:c.alexakos@insybio.com)), K. Θεοφιλάτο ([k.theofilatos@insybio.com](mailto:k.theofilatos@insybio.com)), Δ. Κλεφτογιάννη ([dimitrios.kleftogiannis@kaust.edu.sa](mailto:dimitrios.kleftogiannis@kaust.edu.sa)).

Τα παραδοτέα της εργασίας θα κατατεθούν μέσω e-class (ενότητα εργασίες).

**Η παράδοση της εργασίας σας είναι απαραίτητη προϋπόθεση για την προφορική εξέταση στην οποία θα παρουσιάσετε την εργασία και θα κριθείτε για τις γνώσεις σας στο μάθημα ώστε να διαμορφωθεί ο τελικός βαθμό σας. Η παράδοση της εργασίας μόνο, δεν εξασφαλίζει προβιβάσιμο βαθμό. Η προφορική εξέταση θα είναι ατομική. Η ημερομηνία κατάθεσης καθώς και οι ημερομηνίες εξέτασης θα σας ανακοινωθούν σύντομα.**