Διπλωματική εργασία του

Πριονίδη Κωνσταντίνου

προπτυχιακού φοιτητή του τμήματος ΜΗΥΤΔ

με τίτλο

"ΜικράRNA και Βιολογικά Μονοπάτια"

(micro-RNAs and biological pathways)

Βόλος – Αλεξάνδρεια – Αθήνα
2011

Περιεχόμενα

Λίστα Εικόνων

Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η μελέτη μεθόδων και η ανάπτυξη εφαρμογών, που εντάσσονται στο τομέα της βιοπληροφορικής και ως στόχο έχουν να προσδιορίσουν τον ρόλο που παίζουν τα microRNAs στην μεταβολή των βιολογικών μονοπατιών.

Ειδικότερα, για μια λίστα από microRNAs, γίνεται αναζήτηση σε όλα τα γνωστά βιολογικά μονοπάτια και υπολογίζεται ο βαθμός που τα microRNAs, πιθανόν, να τροποποιούν τα αντίστοιχα μονοπάτια. Η παρουσίαση των αποτελεσμάτων γίνεται με αριθμητικό τρόπο, από όπου γίνεται εμφανής ο βαθμός τροποποίησης των μονοπατιών, αλλά και με γραφικό τρόπο όπου φαίνεται η συμμετοχή των microRNA πάνω στο γράφο του μονοπατιού.

Στο πρώτο κεφάλαιο γίνεται η ανάπτυξη των σχετικών βιολογικών εννοιών ώστε να είναι κατανοητή η περαιτέρω ανάπτυξη των θεμάτων και ειδικότερα των κινήτρων και των μεθόδων της εφαρμογής. Στο δεύτερο κεφάλαιο αναλύονται τα κίνητρα που μας οδηγούν και παρουσιάζονται οι μέθοδοι-αλγόριθμοι της εφαρμογής ενώ στο τρίτο κεφάλαιο γίνεται η παρουσίαση της χρήσης της εφαρμογής. Στα κεφάλαια τέσσερα και πέντε αναφέρονται τεχνικές λεπτομέρειες σχετικά με την δομή, ανάπτυξη και υλοποίηση της εφαρμογής ενώ στο έκτο κεφάλαιο παρουσιάζονται συμπεράσματα σχετικά με τη χρησιμότητα, τη χρηστικότητα και τις επιδόσεις της εφαρμογής.

Το υπάρχον (σχετικό) πρόγραμμα, παρουσιάζει αρκετές ελλείψεις δομής και λειτουργικότητας, οπότε, ένας επιπλέον στόχος της εργασίας αυτής είναι ο σχεδιασμός, από την αρχή, μιας σωστά δομημένης εφαρμογής -πρόγραμμα- ανοιχτού κώδικα (open-source) που να ξεπερνά την προηγούμενη σε επιδόσεις και χρηστικότητα και να προωθεί τη γνώση και την έρευνα μέσω της κατανόησης και της τροποποίησης του κώδικα.

Η παρούσα εργασία εκπονείται στο τμήμα ΜΗΥΤΔ του πανεπιστημίου Θεσσαλίας σε συνεργασία με το ερευνητικό κέντρο Φλέμινγκ και αναμένεται να αποτελέσει μέρος των εφαρμογών του εργαστηρίου DIANA-LAB.

Κεφάλαιο 1

Βιολογικό υπόβαθρο

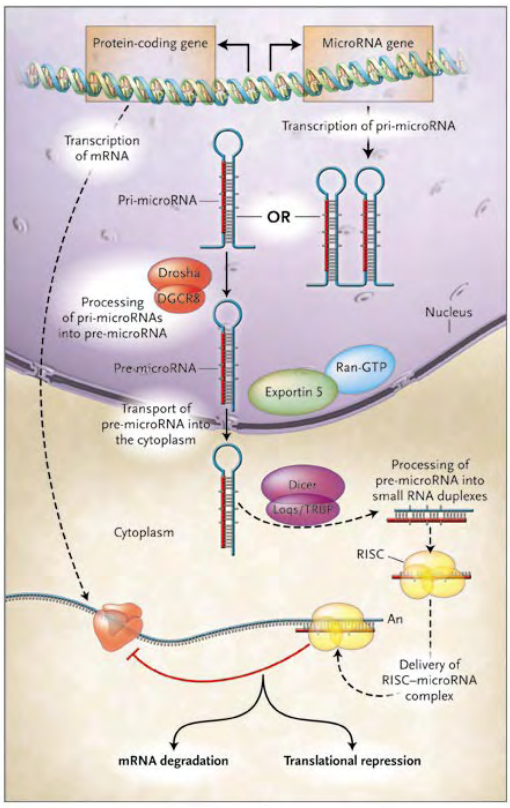
Το παρόν κεφάλαιο έχει ως στόχο την ανάπτυξη ενός βιολογικού υποβάθρου ώστε να είναι κατανοητή η περαιτέρω ανάπτυξη των θεμάτων και ειδικότερα των κινήτρων και των μεθόδων της εφαρμογής.

$ 1.1 ΜικράRNA (microRNA)

Τα μικράRNA (microRNA/miRNA) ορίζονται ως μικρά κομμάτια RNA, απλής έλικας, μήκους περίπου 22 νουκλεοτιδίων, παραγόμενα από ενδογενή αντίγραφα που μπορούν να σχηματίσουν τοπικές δομές φουρκέτας. Αποτελούν μια μεγάλη λειτουργική οικογένεια, μη-κωδικών RNAs (non-coding RNAs), που παίζουν τον ρόλο μετα-μεταγραφικών ρυθμιστών καταστέλλοντας τη μετάφραση των mRNA (messenger RNA).

Έχει αναφερθεί πώς τα microRNAs εμπλέκονται σε διαδικασίες σχετικές με τον καρκίνο λειτουργώντας ως ογκογονίδια ή ως ογκοκαταστολείς. Για το λόγο αυτό, τα microRNAs είναι σημαντικά τόσο για τη διερεύνηση της ρύθμισης της έκφρασης των γονιδίων όσο και για την κατανόηση της παθογένεσης διάφορων ασθενειών.

Τα γονίδια των οποίων η έκφραση φαίνεται ότι ρυθμίζεται από συγκεκριμένα microRNAs ονομάζονται **γονίδια-στόχοι** του microRNA (target genes) και η αναγνώριση τους είναι σημαντική στον προσδιορισμό των λειτουργιών του εκάστοτε microRNA.



Εικόνα 1.1 : MicroRNA

$ 1.2 Βιολογικά μονοπάτια

$ 1.2.1 Τι είναι τα βιολογικά μονοπάτια

Τα βιολογικά μονοπάτια είναι μια σειρά από ενέργειες μεταξύ των μορίων ενός κυττάρου που οδηγούν είτε σε ένα συγκεκριμένο προϊόν ή σε μια αλλαγή στο κύτταρο. Ένα τέτοιο μονοπάτι μπορεί να ενεργοποιήσει τη σύνθεση νέων μορίων όπως λίπος ή πρωτεΐνη. Επίσης, μπορεί να ενεργοποιήσει/απενεργοποιήσει ένα γονίδιο ή να ωθήσει ένα κύτταρο να κινηθεί.

Υπάρχουν αρκετά είδη βιολογικών μονοπατιών, με τα πιο συνηθισμένα να εμπλέκονται στο μεταβολισμό, στη ρύθμιση των γονιδίων και στην εκπομπή σημάτων.

$ 1.2.2 Πώς δουλεύουν τα βιολογικά μονοπάτια

Για να αναπτυχθεί σωστά το σώμα μας και να παραμένει υγιές πρέπει αρκετές διαδικασίες να λειτουργούν μαζί σε διάφορα επίπεδα - από τα όργανα έως τα κύτταρα και τα γονίδια. Τα κύτταρα, δέχονται συνεχώς ερεθίσματα τόσο από το εσωτερικό όσο και από το εξωτερικό του σώματος όπως χτυπήματα, μολύνσεις ή ακόμη και το φαγητό. Για να αντιδράσουν και να προσαρμοστούν σε αυτά τα ερεθίσματα, τα κύτταρα στέλνουν και δέχονται σήματα μέσω των βιολογικών μονοπατιών. Τα μόρια, από τα οποία αποτελούνται τα βιολογικά μονοπάτια, αλληλεπιδρούν με τα σήματα αλλά και μεταξύ τους ώστε να φέρουν εις πέρας συγκεκριμένες εργασίες. Ωστόσο, εάν κάτι δεν δουλέψει σωστά σε ένα βιολογικό μονοπάτι το αποτέλεσμα μπορεί να είναι μια ασθένεια όπως ο καρκίνος ή ο διαβήτης.

$ 1.2.3 Τι μπορούν να μας πουν τα βιολογικά μονοπάτια για τις ασθένειες

Οι ερευνητές μπορούν να γνωρίσουν πολλά από τη μελέτη των βιολογικών μονοπατιών. Γνωρίζοντας ποια γονίδια, πρωτεΐνες η άλλα μόρια εμπλέκονται σε ένα βιολογικό μονοπάτι μπορεί να δώσει στοιχεία για το τι πάει στραβά όταν χτυπάει μια ασθένεια. Για παράδειγμα, οι ερευνητές μπορούν να συγκρίνουν συγκεκριμένα μονοπάτια ενός υγιούς ατόμου με αυτά ενός που πάσχει από κάποια ασθένεια ώστε να ανακαλύψουν τις ρίζες της δυσλειτουργίας. Γνωρίζοντας ποιο μονοπάτι εμπλέκεται σε μια ασθένεια και ποιο κομμάτι του μονοπατιού έχει διαφοροποιηθεί από αυτή, μπορεί να οδηγήσει σε ειδικευμένες στρατηγικές διάγνωσης, θεραπείας και πρόληψης της ασθένειας.

Κεφάλαιο 2

Κίνητρα και μέθοδοι της εφαρμογής

$ 2.1 Κίνητρα και στόχοι της εφαρμογής

Τα microRNAs (miRNAs) είναι λειτουργικά συνδεδεμένα τόσο με τα signaling όσο και με τα μεταβολικά δίκτυα και αλληλεπιδρούν, εκτενώς, με μεταγραφικούς παράγοντες μέσω διακριτών τοπολογικών προτύπων (topological patterns), ενσωματώνοντας μεταγραφικούς και μετα-μεταγραφικούς μηχανισμούς σε ρυθμιστικά βιολογικά δίκτυα. Έχει δειχθεί πως τα miRNAs παίζουν ρόλο σε διάφορες ψυχολογικές και παθολογικές ανθρώπινες καταστάσεις όπως την ανάπτυξη όγκων, στη μετάσταση κ.α. Ακόμη, είναι γνωστό πως διαθέτουν πολλαπλά γονίδια-στόχους και υπάρχουν ισχυρές ενδείξεις πώς μπορούν να δράσουν σε συνεργασία ώστε να διαφοροποιήσουν ένα μοριακό μονοπάτι. Πάραυτα, η συστηματική ενσωμάτωση των miRNAs στα βιολογικά μονοπάτια παραμένει μάλλον ατελής.

Είναι τα παραπάνω που μας οδηγούν στην μελέτη μεθόδων και στην ανάπτυξη μιας εφαρμογής ιστού (web-based application) που εκτελεί μια εμπλουτισμένη ανάλυση των προβλεπόμενων γονιδίων-στόχων, ενός ή περισσοτέρων miRNAs, ανάμεσα στα βιολογικά μονοπάτια. Η συνδυαστική επίδραση των συν-εκφραζόμενων miRNAs στην διαμόρφωση ενός μονοπατιού διευθετείται από την εφαρμογή μέσω της ταυτόχρονης ανάλυσης πολλαπλών miRNAs. Τα γονίδια-στόχοι των miRNAs που εμπλέκονται σε ένα μονοπάτι, επισημαίνονται με γραφικό τρόπο πάνω στο χάρτη του μονοπατιού παρέχοντας μια επισκόπηση των (πιθανών) διαφοροποιημένων κομματιών και διευκολύνοντας την ερμηνεία της ρύθμισης των βιολογικών μονοπατιών.

$ 2.2 Οι μέθοδοι της εφαρμογής

Τα δεδομένα εισόδου της εφαρμογής είναι η λίστα από γονίδια-στόχους των miRNAs οριζόμενη από το όνομα ενός miRNA και του αντίστοιχου λογισμικού "πρόβλεψης-στόχων" που επιλέγονται μέσα από ένα, φιλικό προς τον χρήστη, web-interface. Η ανάκτηση των γονιδίων-στόχων των miRNAs είναι αυτοματοποιημένη για τρία λογισμικά πρόβλεψης (DIANA-microT, PicTar και TargetScan). Εναλλακτικά, σαν είσοδος μπορεί να χρησιμοποιηθεί και οποιαδήποτε λίστα αποτελούμενη από γονίδια, την οποία έχει συντάξει ο χρήστης.

Η εφαρμογή εκτελεί μια ανάλυση μεταξύ των δεδομένων εισόδου, συγκρίνοντας κάθε σετ γονιδίων ανάμεσα σε όλα τα διαθέσιμα βιολογικά μονοπάτια που παρέχονται από το Kyoto Encyclopedia of genes and genomes (KEGG). Το KEGG είναι μια βάση δεδομένων που παρέχει γνώση για διάφορα γονιδιώματα όπως και για τη σχέση αυτών στα βιολογικά συστήματα και χρησιμοποιείται συστηματικά ως βάση γνώσης για την μοριακή βιολογία. Ειδικότερα, η βάση δεδομένων του KEGG σχετικά με τα μονοπάτια (KEGG PATHWAY DATABASE) παρέχει γράφους αλληλεπιδραστικών και αντιδραστικών δικτύων μεταξύ των γονιδίων.

Η ανάλυση των δεδομένων εισόδου (input datasets) εκτελείται με τη βοήθεια του Pearson's chi-squared test $\{x^2 = \sum [(O-E)^2 / E]\}$ όπου O (Observed) είναι ο

αριθμός των γονιδίων ενός dataset που βρέθηκε να συμμετέχει σε ένα συγκεκριμένο μονοπάτι, E (Expected) είναι ο αριθμός των γονιδίων που αναμενόταν εκ τύχης να συμμετέχει στο μονοπάτι -δεδομένων των μεγεθών του μονοπατιού και της αντίστοιχης λίστας. Η ανάλυση, αναθέτει για κάθε dataset εισόδου και για κάθε μονοπάτι, μια τιμή (σκορ) που παρουσιάζεται ως ο αρνητικός φυσικός λογάριθμος του P-value που έχει υπολογιστεί από το chi-squared test.

Επιπλέον, ο αλγόριθμος εκτελεί την παραπάνω ανάλυση και για τα datasets της "Ένωσης" και "Τομής" (Union/Intersection) των γονιδίων των υπολοίπων σετ. Η ανάλυση και το σκορ για το dataset της Ένωσης σε ένα συγκεκριμένο μονοπάτι αντανακλά τη συντονιστική καταστολή του μονοπατιού από όλα τα συν-εκφραζόμενα miRNAs ενώ αυτή του dataset της Τομής δίνει μια επισκόπηση της συνεργατικής καταστολής από ένα μεμονωμένο γονίδιο μεταξύ όλων των εκφρασμένων miRNAs.

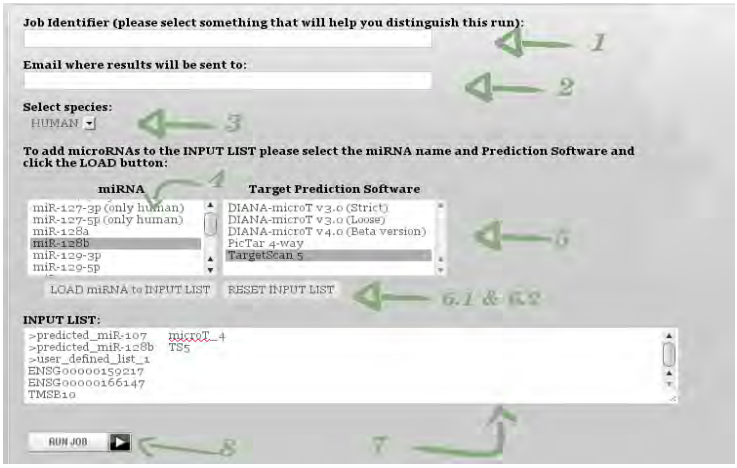
Η εφαρμογή, παρέχει σαν αποτέλεσμα ένα γράφημα με μπάρες από όπου μπορεί να γίνει η σύγκριση των διάφορων datasets (με βάση το σκορ) ανάμεσα στα μονοπάτια. Επίσης, στην σελίδα αποτελεσμάτων της εφαρμογής, εμφανίζεται για κάθε dataset το σκορ μαζί με τον αριθμό και τα ονόματα των γονιδίων που συμμετέχουν σε κάθε μονοπάτι διατεταγμένα σε φθίνουσα σειρά ως προς την τιμή του σκορ. Ακόμη, τα γονίδια-στόχοι των miRNAs που εμπλέκονται σε μονοπάτια, παρουσιάζονται με γραφικό τρόπο πάνω στους γράφους των μονοπατιών -με τρόπο που περιγράφεται αναλυτικά στο επόμενο κεφάλαιο- όπου υπάρχει η δυνατότητα να επισημανθούν από το χρήστη, μεμονωμένα γονίδια ή ολόκληρα datasets διευκολύνοντας έτσι την αναγνώριση αυτών πάνω στο χάρτη του μονοπατιού. Τέλος, παρέχεται ένα αρχείο (spreadsheet) με τα αποτελέσματα της ανάλυσης.

Κεφάλαιο 3

Παρουσίαση και χρήση της εφαρμογής

$ 3.1 Εισαγωγή δεδομένων

Ο χρήστης εισάγει τα δεδομένα προς ανάλυση από τη φόρμα που έχει τη μορφή της *Εικόνας 3.1*. Τα στοιχεία που εισάγει ο χρήστης είναι τα εξής:



Εικόνα 3.1 : Φόρμα εισαγωγής δεδομένων

- Ένα αναγνωριστικό για την συγκεκριμένη "προς-διεκπεραίωση" εργασία, ώστε ο ίδιος να μπορεί να ξεχωρίσει την συγκεκριμένη εργασία μεταξύ πιθανών υπολοίπων *(πλαίσιο 1)*
- Μια διεύθυνση ηλεκτρονικού ταχυδρομείου, στην οποία θα σταλεί ο σύνδεσμος προς τα αποτελέσματα *(πλαίσιο 2)*
- Επιλογή του είδους (άνθρωπος, ποντίκι ...) για το οποίο θα γίνει η ανάλυση *(πλαίσιο 3)*.
- Τη λίστα με τα microRNAS, σε συνδυασμό με το Target-Prediction software.

Για το τελευταίο βήμα ειδικότερα, ο χρήστης, μπορεί να "φορτώσει" στην είσοδο έναν τέτοιο συνδυασμό επιλέγοντας ένα microRNA από τη λίστα των microRNAs *(πλαίσιο 4)*, ένα prediction software από την αντίστοιχη λίστα *(πλαίσιο 5)* και πατώντας το κουμπί *"LOAD miRNA to INPUT LIST" (6.1)*. Ο συνδυασμός αυτός εμφανίζεται στο πεδίο *"INPUT LIST" (πλαίσιο 7)* που αποτελεί την λίστα-είσοδο προς ανάλυση. Ο χρήστης μπορεί να φορτώσει περισσότερους από έναν τέτοιος συνδυασμούς στη λίστα και επιπρόσθετα, μπορεί να συντάξει δικές του λίστες ορίζοντας το όνομα της κάθε λίστας με τα περιεχόμενα γονίδια.

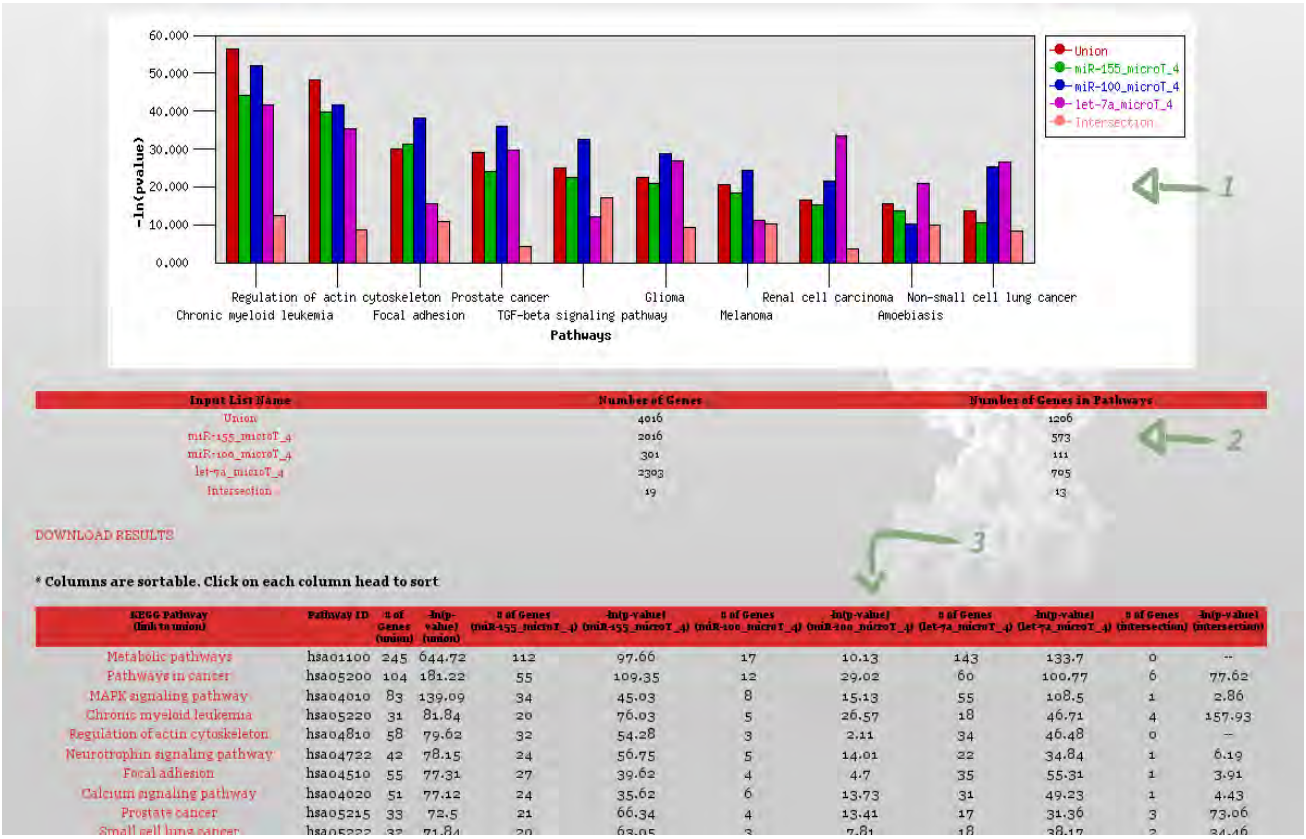
Σε αυτή την περίπτωση, η λίστα ορίζεται με το σύμβολο ">" ακολουθούμενο από το όνομα της λίστας και τα περιεχόμενα γονίδια, καθένα σε μια σειρά, περιγραφόμενα είτε με το ensembl αναγνωριστικό τους ή με το όνομα του γονιδίου. Για παράδειγμα, μια τέτοια λίστα με όνομα "user_defined_list_1" έχει τη μορφή :

>user_defined_list_1
ENSG00000159217
ENSG00000166147
TMSB10

Σε κάθε περίπτωση, ο χρήστης μπορεί να καθαρίσει την είσοδο που έχει εισάγει μέχρι στιγμής, πατώντας το κουμπί *"Reset List" (6.2).*

Όταν ο χρήστης είναι ευχαριστημένος με τα δεδομένα εισαγωγής, τότε πατώντας το κουμπί *"RUN JOB" (πλαίσιο 8)* στέλνει τα δεδομένα προς ανάλυση.

$ 3.2 Η αρχική σελίδα των αποτελεσμάτων



Εικόνα 3.2: Αρχική σελίδα αποτελεσμάτων

Η αρχική σελίδα των αποτελεσμάτων, που φαίνεται στην *εικόνα 3.2*, χωρίζεται σε τρία μέρη.

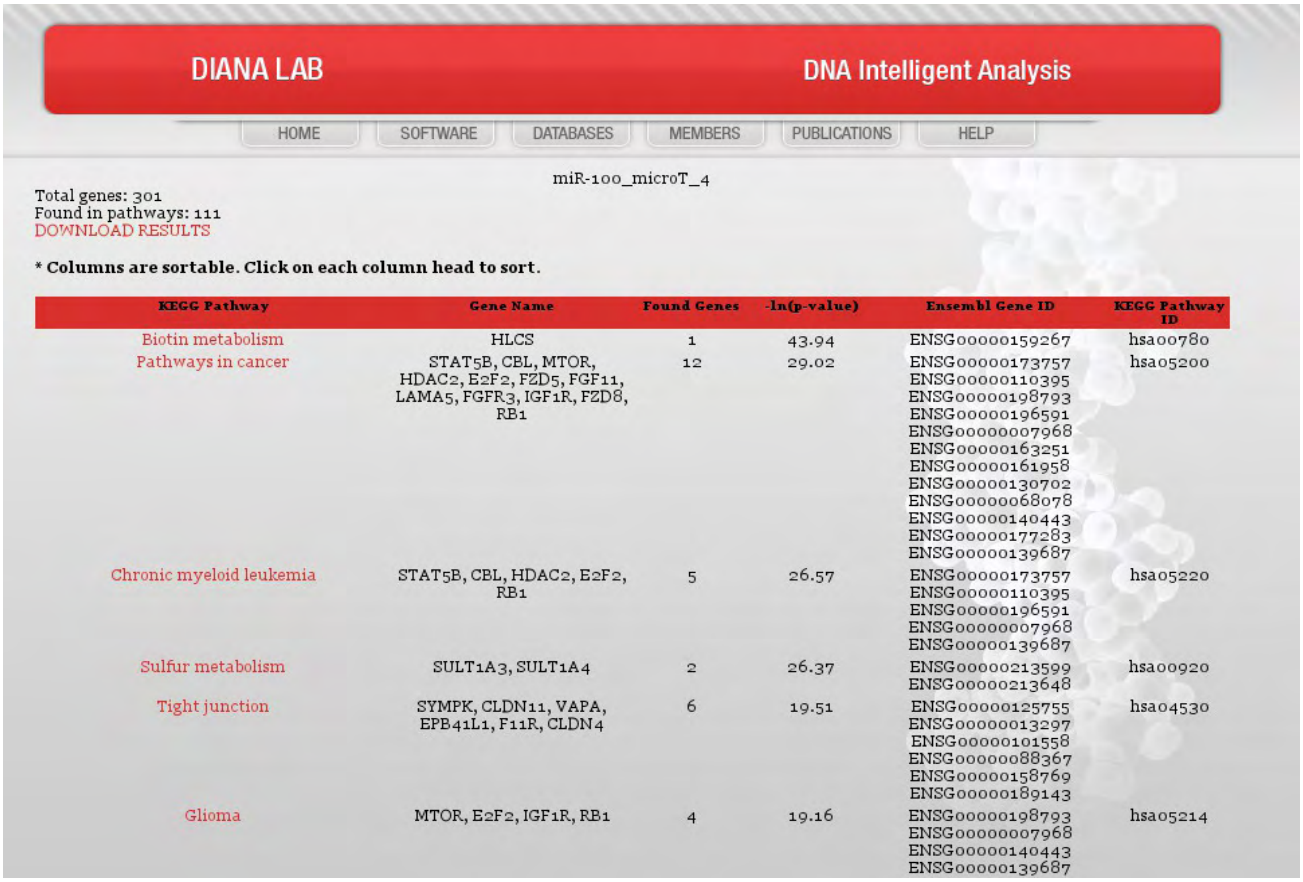
Στο πρώτο *(πλαίσιο 1)* περιέχεται το γράφημα με μπάρες, όπου για κάθε μονοπάτι εμφανίζεται το σκορ κάθε λίστας.

Στο δεύτερο μέρος *(πλαίσιο 2)* παρατίθενται τα ονόματα των λιστών της εισόδου, μαζί με τον αριθμό των γονιδίων-στόχων που ανήκουν στη λίστα όπως και τον αριθμό αυτών που συμμετέχουν σε μονοπάτια. Το όνομα κάθε λίστας περιέχει τον σύνδεσμο που οδηγεί στη σελίδα με τα αναλυτικότερα αποτελέσματα κάθε λίστας και που περιγράφεται στη επόμενη ενότητα. Υπάρχει επίσης και ένας σύνδεσμος, *"DOWNLOAD RESULTS"*, ο οποίος οδηγεί στο xls αρχείο με τα αποτελέσματα το οποίο ο χρήστης μπορεί να "κατεβάσει" στον υπολογιστή του.

Στο τρίτο μέρος *(πλαίσιο 3)*, υπάρχει ένας πίνακας που παραθέτει, με συνοπτικό τρόπο, τα αποτελέσματα κάθε λίστας για κάθε μονοπάτι. Ειδικότερα, για κάθε λίστα υπάρχουν δύο στήλες όπου στη μία αναφέρεται ο αριθμός των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι ενώ στη δεύτερη αναφέρεται το σκορ της λίστας για το μονοπάτι. Το κάθε όνομα μονοπατιού είναι ένας σύνδεσμος προς τη γραφική αναπαράσταση του μονοπατιού για την λίστα της ένωσης, το οποίο περιγράφεται στη συνέχεια. Ο πίνακας είναι διατεταγμένος με φθίνουσα σειρά ως προς το πεδίο του σκορ της ένωσης και ο χρήστης μπορεί να αλλάξει σε οποιαδήποτε στιγμή την διάταξη αυτή, με βάση οποιοδήποτε πεδίο, κάνοντας κλικ στην κεφαλή του πίνακα -δηλαδή κλικ

στην πρώτη γραμμή οποιουδήποτε πεδίου θα διατάξει τον πίνακα με βάση τις τιμές αυτού.

$ 3.3 Σελίδα αποτελεσμάτων λίστας



KEGG Pathway	Gene Name	Found Genes	-ln(p-value)	Ensembl Gene ID	KEGG Pathway ID
Biotin metabolism	HLCS	1	43.94	ENSG00000159267	hsa00780
Pathways in cancer	STAT5B, CBL, MTOR, HDAC2, E2F2, FZD5, FGF11, LAMA5, FGFR3, IGF1R, FZD8, RB1	12	29.02	ENSG00000173757 ENSG00000110395 ENSG00000198793 ENSG00000196591 ENSG00000007968 ENSG00000163251 ENSG00000161958 ENSG00000130702 ENSG00000068078 ENSG00000140443 ENSG00000177283 ENSG00000139687	hsa05200
Chronic myeloid leukemia	STAT5B, CBL, HDAC2, E2F2, RB1	5	26.57	ENSG00000173757 ENSG00000110395 ENSG00000196591 ENSG00000007968 ENSG00000139687	hsa05220
Sulfur metabolism	SULT1A3, SULT1A4	2	26.37	ENSG00000213599 ENSG00000213648	hsa00920
Tight junction	SYMPK, CLDN11, VAPA, EPB41L1, F11R, CLDN4	6	19.51	ENSG00000125755 ENSG00000013297 ENSG00000101558 ENSG00000088367 ENSG00000158769 ENSG00000189143	hsa04530
Glioma	MTOR, E2F2, IGF1R, RB1	4	19.16	ENSG00000198793 ENSG00000007968 ENSG00000140443 ENSG00000139687	hsa05214

Εικόνα 3.3: Παρουσίαση αποτελεσμάτων συγκεκριμένης λίστας

Στην σελίδα αυτή *(εικόνα 3.3)* παρουσιάζονται αναλυτικά τα αποτελέσματα της ανάλυσης για μια συγκεκριμένη λίστα. Αναφέρεται το όνομα της λίστας, ο αριθμός των γονιδίων-στόχων της λίστας και ο αριθμός αυτών που συμμετέχουν σε μονοπάτια. Επίσης, υπάρχει και σύνδεσμος προς το αρχείο xls των αποτελεσμάτων *"DOWNLOAD RESULTS"* ενώ ακολουθεί ο πίνακας με τα αποτελέσματα.

Εκεί, υπάρχει μια εγγραφή (σειρά) για κάθε μονοπάτι στο οποίο συμμετέχει τουλάχιστον ένα γονίδιο από τη λίστα. Τα πεδία του πίνακα είναι:

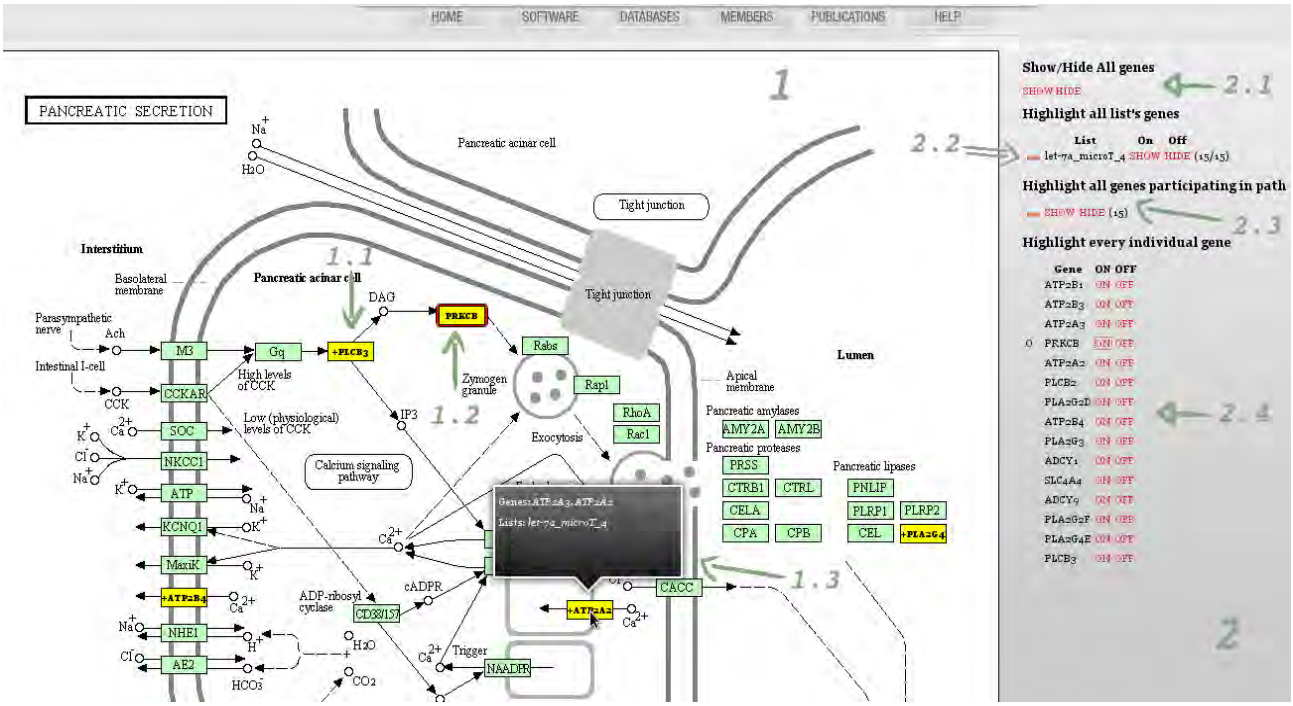
- Το όνομα του μονοπατιού *(KEGG PATHWAY)*, περιέχει και σύνδεσμο προς την γραφική αναπαράσταση της λίστας για το συγκεκριμένο μονοπάτι (αναλύεται στην επόμενη ενότητα)
- Τα ονόματα των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι *(Gene Name)*
- Ο αριθμός των γονιδίων της λίστας που βρέθηκαν να συμμετέχουν στο μονοπάτι *(Found Genes)*
- Το σκορ της λίστας για το συγκεκριμένο μονοπάτι *(-ln(p-value))*
- Τα ensembl αναγνωριστικά των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι

(Ensembl Gene ID)

- Το KEGG αναγνωριστικό του μονοπατιού *(KEGG Pathway ID)*

Ο πίνακας είναι διατεταγμένος σε φθίνουσα σειρά ως προς το σκορ, ενώ ο χρήστης μπορεί να διατάξει τον πίνακα με βάσει οποιοδήποτε πεδίο του πίνακα κάνοντας κλικ στην κεφαλή κάθε πεδίου. Όπως αναφέρθηκε και νωρίτερα, ο χρήστης κάνοντας κλικ πάνω στο όνομα του μονοπατιού οδηγείται στη γραφική αναπαράσταση του μονοπατιού.

$ 3.4 Γραφική αναπαράσταση μονοπατιού απλής λίστας



Εικόνα 3.4: Σελίδα γραφικής αναπαράστασης μονοπατιού

Σε αυτήν τη σελίδα, ο χρήστης έχει πρόσβαση στο γράφο του μονοπατιού με επισημασμένα τα γονίδια της λίστας πάνω σ' αυτόν. Τα γονίδια που συμμετέχουν στο μονοπάτι αποτελούν τις κορυφές του γράφου. Κάθε γονίδιο πάνω στο γράφο είναι ένας σύνδεσμος προς την ιστοσελίδα του KEGG όπου αναφέρονται λεπτομέρειες σχετικές με αυτό. Σε μερικά μονοπάτια τα γονίδια μπορεί να εμφανίζονται και ως ακμές του γράφου που συνδέουν δύο κορυφές.

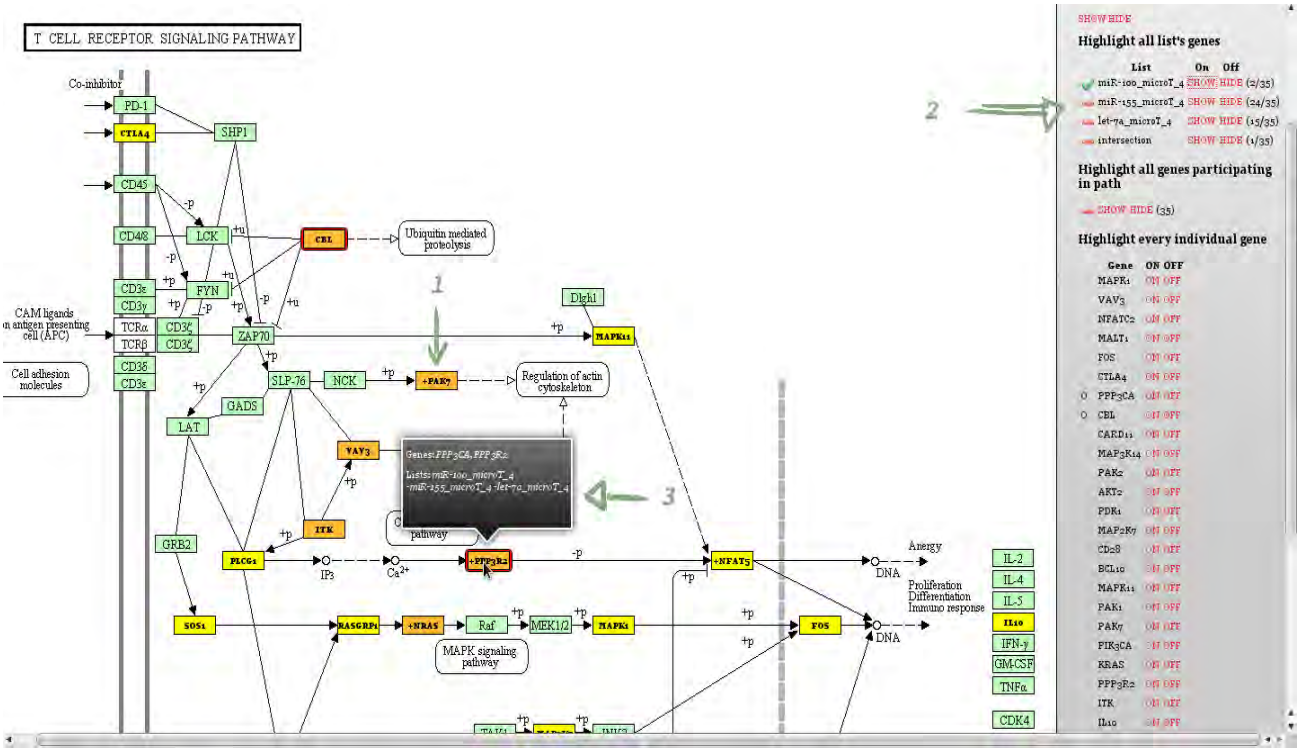
Η σελίδα, ένα στιγμιότυπο της οποίας εμφανίζεται στη εικόνα 3.4, αποτελείται από δύο μέρη. Το πρώτο *(πλαίσιο 1)* περιέχει το γράφο του μονοπατιού με τα περιεχόμενα γονίδια-στόχους της λίστας ενώ το δεύτερο *(πλαίσιο 2)* είναι το πάνελ από το οποίο ο χρήστης χειρίζεται την εμφάνιση των γονιδίων πάνω στο γράφο.

Στο γράφο, τα γονίδια της λίστας εμφανίζονται με κίτρινο φόντο ενώ αν σε ένα συγκεκριμένο σημείο του γράφου συμμετέχουν περισσότερα από ένα γονίδια-στόχοι της λίστας τότε αυτό επισημαίνεται με το σύμβολο "+" μπροστά από το όνομα του γονιδίου *(πλαίσιο 1.1)*. Ο χρήστης μπορεί να απενεργοποιήσει οποιαδήποτε στιγμή την επισήμανση των γονιδίων πάνω στο γράφο από την επιλογή *"Show/Hide All Genes" (πλαίσιο 2.1)*.

Υπάρχει και ένας δεύτερος τρόπος επισήμανσης των γονιδίων πάνω στο γράφο που ονομάζεται highlight και στόχο έχει να ξεχωρίζονται συγκεκριμένα γονίδια της λίστας πάνω στο γράφο. Στην περίπτωση που ένα γονίδιο είναι highlighted τότε εμφανίζεται με κόκκινο φόντο πάνω στον γράφο ή στην περίπτωση που είναι ενεργοποιημένη και η προηγούμενη επιλογή τότε δημιουργείται ένα κόκκινο περίγραμμα γύρω από το γονίδιο όπως φαίνεται στο *πλαίσιο 1.2*. Ο χρήστης μπορεί να κάνει highlight όλα τα γονίδια συγκεκριμένης λίστας από την επιλογή *"Highlight all list's genes" (πλαίσιο 2.2)* ή και να κάνει highlight όλα τα γονίδια της λίστας που εμφανίζονται στο μονοπάτι από την επιλογή *"Highlight all genes participating in path" (πλαίσιο 2.3)*. Στην περίπτωση απλής λίστας - δηλαδή διαφορετικής της ένωσης - δεν υπάρχει διάκριση μεταξύ των παραπάνω αλλά η σημαντικότητα τους θα φανεί στη γραφική αναπαράσταση της ένωσης που παρουσιάζεται στην επόμενη ενότητα. Στο κατώτερο κομμάτι του πάνελ *(πλαίσιο 2.4)* υπάρχει η επιλογή *"Highlight every individual gene"* από την οποία ο χρήστης μπορεί να κάνει highlight κάθε ξεχωριστό γονίδιο της λίστας πάνω στο γράφο. Αν ένα γονίδιο είναι highlighted τότε εμφανίζεται το σύμβολο "O" δίπλα από το όνομα του σε αυτό το σημείο του πάνελ.

Τέλος, εάν ο χρήστης αφήσει το ποντίκι πάνω από ένα συγκεκριμένο σημείο με επισήμανση πάνω στο γράφο, τότε εμφανίζεται ένα tooltip *(πλαίσιο 1.3)* στο οποίο αναφέρονται τα γονίδια που συμμετέχουν σ' αυτό το σημείο.

$ 3.5 Γραφική αναπαράσταση μονοπατιού της ένωσης



Εικόνα 3.5: Σελίδα γραφικής αναπαράστασης λίστας/ένωσης

Η γραφική αναπαράσταση μονοπατιού της λίστας της ένωσης (ένα στιγμιότυπο της οποίας φαίνεται στην *εικόνα 3.5*) ακολουθεί τους κανόνες και τη λειτουργία της απλής

λίστας. Επειδή όμως η ένωση είναι η λίστα που περιέχει όλα τα στοιχεία των υπολοίπων (συμπεριλαμβανομένης και της τομής), η γραφική αναπαράσταση αυτής εμφανίζει κάποιες εξτρά λειτουργίες. Ειδικότερα, εάν σε ένα σημείο του γράφου συμμετέχουν γονίδια που ανήκουν σε περισσότερες από μία λίστες εισόδου τότε το φόντο του γονιδίου γίνεται πορτοκαλί από κίτρινο που ήταν προηγουμένως *(πλαίσιο 1)*. Επίσης, στην επιλογή *"Highlight all list's genes" (πλαίσιο 2)* στο πάνελ χειρισμού, εμφανίζονται πλέον όλες οι λίστες της εισόδου -με τον αριθμό των γονιδίων τους που συμμετέχουν στο μονοπάτι- και έτσι ο χρήστης μπορεί να κάνει highlight πάνω στο μονοπάτι γονίδια συγκεκριμένων λιστών. Αν από αυτή την επιλογή είναι ενεργοποιημένες περισσότερες από μια λίστες, τότε γίνονται highlight τα γονίδια της ένωσης αυτών. Τέλος, το tooltip που εμφανίζεται, περιέχει και τα ονόματα των λιστών που συμμετέχουν στο συγκεκριμένο σημείο *(πλαίσιο 3)*.

Κεφάλαιο 4

Δομή και υλοποίηση του πυρήνα της εφαρμογής

$ 4.1 Εισαγωγή

Ο πυρήνας της εφαρμογής, είναι το κομμάτι όπου γίνεται η βασική ανάλυση και ο υπολογισμός των τιμών (scores) για τις λίστες των microRNAs που έχουν εισαχθεί από τον χρήστη. Για την υλοποίηση αυτού του κομματιού έχει επιλεγεί το λογισμικό (γλώσσα προγραμματισμού) Perl ώστε ο κώδικας να είναι εύκολα αντιληπτός και κατανοητός από κάποιον που θέλει να τον διαβάσει ή τον να αλλάξει, μιας και η Perl είναι τη κοινώς αποδεκτή γλώσσα που χρησιμοποιείται στον τομέα της βιοπληροφορικής.

Τα δεδομένα σχετικά με τα γονίδια-στόχοι των microRNAs παρέχονται από το ινστιτούτο Φλέμινγκ. Πληροφορίες και δεδομένα σχετικά με τα βιολογικά μονοπάτια παρέχονται από το KEGG. Δεδομένα σχετικά με τα γονίδια (γενικότερα) παρέχονται και από τις δύο παραπάνω πηγές.

Το παρόν κεφάλαιο, εκτός από την παρουσίαση και την ανάλυση της δουλείας που έχει γίνει σχετικά με αυτό το κομμάτι έχει και ως στόχο να λειτουργήσει και σαν εγχειρίδιο για κάποιον που θέλει να χειριστεί (κατανοήσει/παραμετροποιήσει) τον κώδικα της εφαρμογής.

$ 4.2 Ορίσματα εισόδου στην εφαρμογή (Input Arguments)

Η εφαρμογή δέχεται (υποχρεωτικά) σαν είσοδο ένα όρισμα. Αυτό το όρισμα είναι η διαδρομή (path) ενός φακέλου στον οποίο θα αποθηκευτούν τα δεδομένα (έξοδος) που παράγει η εφαρμογή. Επίσης, δέχεται και μια επιλογή (option) "-d" ή "--database" το οποίο δηλώνει εάν θέλουμε τα δεδομένα να εξάγονται από τη βάση δεδομένων SQL. Ως προεπιλογή, η εφαρμογή εξάγει ότι δεδομένα χρειάζεται από αρχεία που περιέχουν τις ίδιες πληροφορίες με την βάση δεδομένων.

Σε περίπτωση που η εφαρμογή καλεστεί με όχι ορθό τρόπο, τότε η εκτέλεση της αναβάλλεται και εμφανίζεται ένα μήνυμα που περιγράφει τον ορθό τρόπο με τον οποίο αυτή πρέπει να καλεστεί. Το ίδιο μήνυμα εμφανίζεται και στην περίπτωση που καλεστεί με το option "-h" ή "--help".

Σε αυτό το σημείο να σημειώσουμε ότι μόλις ο χρήστης δηλώσει (submit) την δουλειά του από το web-interface τότε εκτελείται πρώτα ένα άλλο πρόγραμμα (script) το οποίο διαβάζει την είσοδο του χρήστη, προσδίδει ένα μοναδικό αριθμό στην συγκεκριμένη δουλειά και δημιουργεί έναν φάκελο στο σύστημα, με όνομα αυτόν τον μοναδικό αριθμό. Επίσης, παράγει και δύο αρχεία, τα "lists.dat" και "params.dat" που περιέχουν αντίστοιχα την είσοδο που έχει δώσει ο χρήστης - δηλαδή την λίστα με τα microRNAs ή άλλες λίστες που έχει συντάξει ο ίδιος - και πληροφορίες σχετικά με την συγκεκριμένη δουλειά - όπως το είδος (species) που ενδιαφέρει τον χρήστη, το e-mail του κ.α- τα οποία και τοποθετεί στον συγκεκριμένο φάκελο. Εν συνεχεία, δηλώνει στη βάση δεδομένων, πως υπάρχει μια διαθέσιμη δουλειά προς διεκπεραίωση. Ένα άλλο script, ελέγχει περιοδικά τη βάση δεδομένων για διαθέσιμες δουλειές και μόλις εντοπίσει μια

τέτοια, τότε αυτό με τη σειρά του καλεί την εφαρμογή μας με τα αντίστοιχα ορίσματα.

$ 4.3 Ο κορμός της εφαρμογής

Αρχικά, κρατάμε στη μεταβλητή *$store_path* τη διαδρομή του μονοπατιού στο οποίο θα αποθηκευτεί η έξοδος της εφαρμογής αλλά και βρίσκονται τα σχετικά αρχεία με την είσοδο και τις λοιπές πληροφορίες του χρήστη. Στη συνέχεια, διαβάζουμε αυτές τις πληροφορίες μέσω της συνάρτησης *"get_user_info"* και αποθηκεύουμε τα δεδομένα σε μεταβλητές. Αυτές που θα μας απασχολήσουν ιδιαίτερα είναι οι μεταβλητές *"$species"* και *"$unique_id"* στις οποίες κρατάμε το είδος στο οποίο θα αναφέρεται η ανάλυση (άνθρωπος, ποντίκι, κλπ.) και τον μοναδικό αριθμό που έχει προσδοθεί στην συγκεκριμένη δουλειά, αντίστοιχα και προχωράμε στη "μετάφραση" του είδους στην επιστημονική ορολογία μέσω της συνάρτησης *"translate_species"* όπως για παράδειγμα HUMAN->hsa, MOUSE->mmu κλπ. Ακολουθεί το κομμάτι όπου εξάγουμε τα δεδομένα σχετικά με όλα τα γνωστά γονίδια.

$ 4.4 Εξαγωγή όλων των γνωστών γονιδίων

Η εφαρμογή προχωράει με την εξαγωγή όλων των γνωστών γονιδίων μέσω της συνάρτησης *"retrieve_gene_names"*, για το είδος για το οποίο ενδιαφερόμαστε και όπως αυτά είναι αποθηκευμένα στη βάση δεδομένων του ινστιτούτου Φλέμινγκ. Εδώ γίνεται η χρήση του option "-d" που τυχόν έχει δοθεί κατά την εκτέλεση της εφαρμογής. Αν το option είναι ενεργό τότε τα γονίδια εξάγονται από τη βάση δεδομένων και πιο συγκεκριμένα από τον πίνακα *"diana_protein_genes"*. Σε διαφορετική περίπτωση τα γονίδια διαβάζονται από το αρχείο *"Database.dat"*. Σκοπός είναι να κάνουμε γνωστά στην εφαρμογή όλα τα γνωστά γονίδια ώστε να είναι δυνατή η περαιτέρω ανάλυση και ο υπολογισμός των δεδομένων.

Στο σημείο αυτό να αναφέρουμε πώς γενικά παρουσιάζονται ασυμφωνίες μεταξύ των διάφορων βάσεων δεδομένων που έχουν να κάνουν με βιολογικές και γενετικές πληροφορίες, παγκοσμίως. Τέτοιες ασυμφωνίες εμφανίζονται και μεταξύ του ινστιτούτου Φλέμινγκ και του KEGG. Συγκεκριμένα κάποια γονίδια που περιέχονται στη βάση του Φλέμινγκ μπορεί να αντιστοιχούν σε ένα ή περισσότερα γονίδια στην πλευρά του KEGG και άλλα να μην έχουν καμιά αντιστοίχιση. Εφόσον η περαιτέρω ανάλυση στην εφαρμογή θα έχει να κάνει με δεδομένα και από τις δύο πλευρές θα κρατήσουμε μια αντιστοίχιση μεταξύ αυτών. Κάτι τέτοιο γίνεται δυνατό μέσω μιας αντιστοίχισης που μας δίνει το KEGG μεταξύ αναγνωριστικών γονιδίων με βάση το KEGG *(kegg_gene_id)* και ensembl αναγνωριστικών *(ensg_id)* μέσω του αρχείου *"kegg_to_ensembl.list"*

Η συνάρτηση επιστρέφει ένα hash το οποίο για κάθε γνωστό γονίδιο περιέχει μια εγγραφή που σαν key: έχει το ensembl_id του γονιδίου και σαν value: έχει μια αναφορά *(reference)* σε πίνακα *(array)*. Αυτός ο πίνακας, στο πρώτο κελί περιλαμβάνει το όνομα του γονιδίου και στο δεύτερο την αντιστοίχηση με τα δεδομένα του KEGG όπως φαίνεται στο παρακάτω γράφημα. Κρατάμε αυτή τη δομή σε μια μεταβλητή με όνομα *"%genes"*.

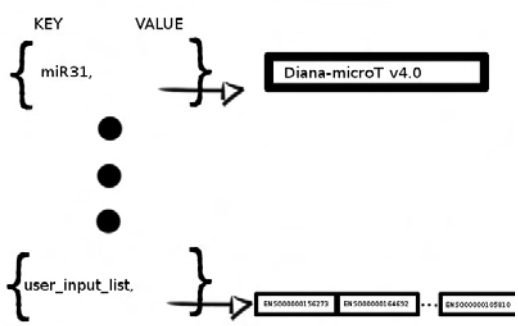
Κρατάμε, ακόμη, και μια αντιστοιχία μεταξύ των KEGG αναγνωριστικών και του ονόματος του γονιδίου μέσω της συνάρτησης *"kegg_to_name_translation"* και με τη βοήθεια του αρχείου *"list"* που μας παρέχεται από το KEGG. Αποθηκεύουμε την μετάφραση αυτή σε ένα hash όπου για κάθε kegg_id υπάρχει μια εγγραφή με key: το kegg_id και value:

το αντίστοιχο όνομα του γονιδίου και το αποθηκεύουμε στη μεταβλητή *"$keggid_to_name"* το οποίο είναι ένα reference σε hash.

$ 4.5 Διαβάζοντας την είσοδο του χρήστη και τα γονίδια-στόχους

Προχωράμε στην εφαρμογή διαβάζοντας την είσοδο - δηλαδή τη λίστα με τα microRNAs - που έχει δηλώσει ο χρήστης και με την εξαγωγή των γονιδίων στόχων των micro-RNAs για συγκεκριμένο prediction software. Πέρα από τα παραπάνω, γίνεται και έλεγχος ορθότητας της εισόδου. Κάτι τέτοιο είναι απαραίτητο καθώς ο χρήστης έχει τη δυνατότητα να συντάξει και δικές του λίστες οδηγώντας σε τυχόν παρατυπίες.

Μέσω της συνάρτησης *"get_user_input"* και του αρχείου *"lists.dat"*, που έχει δημιουργηθεί εκ των προτέρων όπως εξηγήθηκε νωρίτερα, διαβάζεται η είσοδος του χρήστη και όπως αυτός εισήγαγε τα δεδομένα από το web-interface της εφαρμογής. Κάθε ξεχωριστό micro-RNA και λίστα που έχει συντάξει ο χρήστης θα αναφέρεται από εδώ και στο εξής ως "λίστα".Η είσοδος του χρήστη αναλύεται και αποθηκεύεται σε ένα hash που για κάθε λίστα έχει μια εγγραφή με key: το όνομα της λίστας και value: μια αναφορά σε πίνακα όπου στην περίπτωση των micro-RNAs έχει μόνο ένα στοιχείο το οποίο περιέχει το όνομα του prediction software που θα χρησιμοποιηθεί -για το συγκεκριμένο micro-RNA- ενώ στην περίπτωση λίστας που έχει συντάξει ο χρήστης περιέχει σε κάθε του κελί κάθε γονίδιο από τη συγκεκριμένη λίστα. Σαν όνομα της λίστας μπαίνει το όνομα του microRNA ή της λίστας του χρήστη διαφορετικά. Η δομή αυτή (που φαίνεται και στο διπλανό γράφημα) αποθηκεύεται στη μεταβλητή *"%lists"*. Σε αυτή τη δομή έχουν αφαιρεθεί microRNAs τα οποία εμφανίζονται παραπάνω από μία φορά με το ίδιο prediction software που τυχόν έχουν εισαχθεί από τον χρήστη εκ παραδρομής.



Γράφημα 4.1

Στην συνέχεια γίνεται η εξαγωγή των γονιδίων-στόχων των εκάστοτε microRNAs σε συνάρτηση με το prediction software που έχει δηλώσει ο χρήστης, μέσω των συναρτήσεων *"retrieve_target_genes"* και *"get_genes_from_db"*. Υπάρχει η δυνατότητα επιλογής από τον χρήστη ανάμεσα σε πέντε διαφορετικά prediction softwares, τα MicroT4, MicroT3_strict, MicroT3_loose, PicTar και TS5. Για κάθε μία από της τρεις κατηγορίες (MicroT, PicTar και TS5) παρέχονται από το ινστιτούτο Φλέμινγκ τα δεδομένα όπου για κάθε microRNA αναφέρονται τα γονίδια-στόχοι του καθενός με την μορφή του ensembl_id. Ειδικότερα, για το MicroT κάθε γονίδιο στόχος έχει και μια τιμή η οποία πρέπει να ξεπερνά ένα κατώφλι (threshold) ώστε να συμπεριληφθεί στη λίστα. Πλέον, η δομή *"%lists"* που είχε δημιουργηθεί νωρίτερα, εμπλουτίζεται ώστε κάθε microRNA να περιέχει τα γονίδια-στόχους.

Παρακάτω, γίνεται έλεγχος στις λίστες που έχει συντάξει ο ίδιος ο χρήστης. Αυτός, έχει τη δυνατότητα να δηλώσει ένα γονίδιο είτε με το ensembl_id ή με το όνομα του. Εάν δεν έχει δηλωθεί με μία από τις δύο μορφές που θεωρούνται έγκυρες, τότε αυτή η εγγραφή πρέπει να παραληφθεί. Η διαδικασία αυτή επιτυγχάνεται μέσω της συνάρτησης *"sanitize_user_lists"* και πλέον η δομή *"%lists"* περιέχει μόνο έγκυρα δεδομένα και έτσι μπορεί να προχωρήσει η περαιτέρω ανάλυση με ασφάλεια.

Τέλος, εάν ο χρήστης έχει εισάγει παραπάνω από μία λίστα, τότε παράγεται και η ένωση και η τομή όλων των λιστών με βάση τα γονίδια-στόχους της κάθε λίστας. Μέσω της συνάρτησης *"create_join_and_intersection"* η δομή *"%lists"* εμπλουτίζεται με τις λίστες "union" για την ένωση και "intersection" για την τομή με τα αντίστοιχα γονίδια.

$ 4.6 Στην ονοματολογία του KEGG

Από το σημείο αυτό και έπειτα, η ανάλυση έχει να κάνει με τα βιολογικά μονοπάτια και με τι βαρύτητα τα micro-RNAs συμμετέχουν σ' αυτά. Επειδή όμως τα βιολογικά μονοπάτια παρέχονται από το KEGG, ότι έχει σχέσει με αυτά είναι στην ονοματολογία του KEGG και ειδικότερα τα γονίδια που εμφανίζονται στα μονοπάτια αναφέρονται με το kegg_gene_id. Γι' αυτό θεωρείται σκόπιμο να εργαζόμαστε πλέον με βάση την ονοματολογία του KEGG. Για να γίνει όμως αυτό, θα πρέπει τα γονίδια στόχοι που εμπεριέχονται στις εκάστοτε λίστες -και εμφανίζονται με το ensembl_id- να "μεταφραστούν" σε kegg_genes_id. Μέσω της συνάρτησης "translate_lists_to_kegg_annotation" και με την χρήση της δομής *"%genes"* όπου έχει κρατηθεί η παραπάνω αντιστοιχία παράγεται μια νέα δομή όπου για κάθε λίστα υπάρχει μια εγγραφή σε ένα hash με key: το όνομα της λίστας και value: ένας πίνακας με τα γονίδια της λίστας (σε αντιστοιχία με τη δομή *"%lists"* που είχε δημιουργηθεί προηγουμένως). Πλέον, τα γονίδια κάθε λίστας εμφανίζονται με την ονοματολογία του KEGG. Επιπρόσθετα, εάν δεν υπάρχει αντιστοιχία μεταξύ ensembl_id και kegg_gene_id το γονίδιο-στόχος αφαιρείται από τη λίστα ενώ αν υπάρχει αντιστοιχία με περισσότερα από ένα kegg_genes_id τότε αυτό το γονίδιο υπολογίζεται αντίστοιχες φορές.

Επίσης, δημιουργείται και μια δομή όπου για κάθε αναγνωριστικό γονιδίου σε ονοματολογία KEGG υπάρχει η αντιστοίχιση σε ensembl_id. Η δομή αυτή αποθηκεύεται στη μεταβλητή *"$kegg_to_ensemble"* που είναι μια αναφορά σε hash.

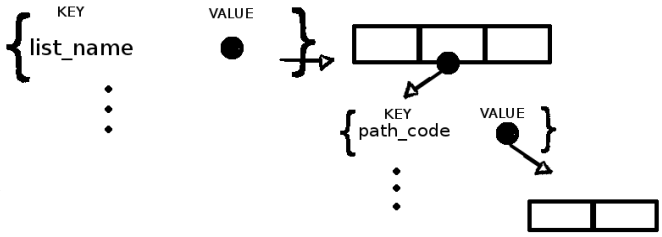
$ 4.7 Αναζήτηση στα βιολογικά μονοπάτια και υπολογισμός του Pvalue

Για τον υπολογισμό του σκορ κάθε λίστας, πρέπει να γίνει αναζήτηση σε όλα τα βιολογικά μονοπάτια και να υπολογιστεί με τι βαθμό, τα γονίδια - στόχοι των λιστών, συμμετέχουν σ' αυτά. Γι' αυτό, αρχικά, "φορτώνουμε" τα γονίδια που συμμετέχουν σε κάθε μονοπάτι σε μια δομή, που αποθηκεύουμε στη μεταβλητή *"%pathways"*, μέσω της συνάρτησης *"load_pathways"*. Η πληροφορία σχετικά με τα γονίδια που εμφανίζονται σε ένα μονοπάτι, παρέχεται από το KEGG και βρίσκεται στο αρχείο *"list"*. Η παραπάνω δομή είναι ένα hash όπου για κάθε μονοπάτι περιέχει μια εγγραφή με key: το αναγνωριστικό του μονοπατιού και value: μια αναφορά σε hash, που με τη σειρά του, για κάθε γονίδιο που εμφανίζεται στο μονοπάτι περιέχει μια εγγραφή με key: το kegg αναγνωριστικό του γονιδίου και "undefined" value.

Ακόμη, αποθηκεύουμε και την αντιστοίχιση μεταξύ αναγνωριστικού μονοπατιού και όνομα μονοπατιού σε μια δομή hash, που αποθηκεύουμε στη μεταβλητή *"%path_names"*, μέσω της συνάρτησης *"load_path_names"*. Και αυτή η πληροφορία παρέχεται από το KEGG και βρίσκεται στο αρχείο *"map_title.tab"*.

Στην συνέχεια, υπολογίζονται μέσω της συνάρτησης *"compute_parameters_for_pvalue"*,για κάθε μονοπάτι, οι παράμετροι κάθε λίστας που είναι αναγκαίοι για τον υπολογισμό του

Pvalue. Η συνάρτηση επιστρέφει μια (μάλλον πολύπλοκη) δομή που περιέχει όλη την απαιτούμενη πληροφορία για τον υπολογισμό της τιμής του p-value. Συγκεκριμένα, η δομή αποτελείται από ένα hash, που για κάθε λίστα, περιέχει μια εγγραφή με key: το όνομα της λίστας και value: μια αναφορά σε πίνακα. Ο πίνακας αυτός έχει τρεις θέσεις. Στην πρώτη, αποθηκεύεται ο συνολικός αριθμός των γονιδίων της λίστας που συμμετέχουν σε οποιοδήποτε μονοπάτι (κάθε γονίδιο - στόχος λαμβάνεται υπόψιν μόνο μία φορά). Στην τρίτη θέση, αποθηκεύεται ο συνολικός αριθμός των γονιδίων της λίστας. Στην δεύτερη θέση, αποθηκεύεται μια αναφορά σε hash που περιέχει μια εγγραφή για κάθε μονοπάτι με key: το αναγνωριστικό του μονοπατιού και value: μια αναφορά σε πίνακα, δύο θέσεων. Στην πρώτη θέση αυτού του πίνακα, αποθηκεύεται ο αριθμός των γονιδίων της λίστας που εμφανίζονται στο συγκεκριμένο μονοπάτι ενώ στη δεύτερη αποθηκεύεται ένα αλφαριθμητικό με τα kegg αναγνωριστικά αυτών των γονιδίων. Για ευκολία στην κατανόηση, η παραπάνω δομή παρουσιάζεται στο διπλανό γράφημα *(Γράφημα 4.2)*.



Γράφημα 4.2

Αφού πλέον υπάρχει όλη η απαιτούμενη πληροφορία, μπορεί να υπολογιστεί το σκορ κάθε λίστας για κάθε ξεχωριστό μονοπάτι. Για μια λίστα, μπορεί να ανατεθεί τιμή-σκορ για ένα συγκεκριμένο μονοπάτι, μόνο εάν τουλάχιστον ένα γονίδιο της λίστας συμμετέχει στο μονοπάτι αυτό. Η τιμή που πραγματικά ανατίθεται σαν σκορ σε μια λίστα είναι ο αρνητικός φυσικός λογάριθμος Pvalue, στρογγυλοποιημένος στα δύο δεκαδικά ψηφία.

Η συνάρτηση που κάνει τον παραπάνω υπολογισμό είναι η *"evaluate_pvalue"*. Η τιμή Pvalue προκύπτει από το Pearson's chi-squared-test. Το chi-squared-test παίρνει σαν είσοδο δύο διανύσματα, ας τα ονομάσουμε x_1 και x_2, δύο μεταβλητών το καθένα, ας τις ονομάσουμε x_{y1} και x_{y2} όπου y είναι ο αριθμός του διανύσματος. Η μεταβλητή x_{11} έχει σαν τιμή τον αριθμό των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι. Η μεταβλητή x_{12} έχει σαν τιμή τον συνολικό αριθμό των γονιδίων της λίστας που συμμετέχουν σε τουλάχιστον ένα μονοπάτι μείον τον αριθμό των γονιδίων της λίστας που συμμετέχουν στο συγκεκριμένο μονοπάτι. Η μεταβλητή x_{21} έχει σαν τιμή των αριθμό των γονιδίων που εμφανίζονται στο μονοπάτι (γενικά) μείον τον αριθμό των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι και τέλος, η μεταβλητή x_{22} έχει σαν τιμή τον συνολικό αριθμό όλων των γνωστών γονιδίων (υπόβαθρο/background) μείον την τιμή της x_{11} μείον την τιμή της x_{12} μείον την τιμή της x_{21}. Εάν το Pvalue που επιστρέφεται από το chi-squared-test έχει μηδενική τιμή, τότε του ανατίθεται η τιμή 10^{-280}, μιας και ο λογάριθμος δεν δέχεται αρνητικές τιμές. Όπως σημειώθηκε και προηγουμένως, το σκορ που ανατίθεται σε κάθε λίστα, είναι η αρνητική τιμή του φυσικού λογαρίθμου του Pvalue. Η τιμή που υπολογίστηκε σαν σκορ, αποθηκεύεται στην δομή που έχει περιγραφεί προηγουμένως σε ένα καινούργιο κελί του τελευταίου πίνακα.

Τέλος, δημιουργείται και μια βοηθητική δομή μέσω της συνάρτησης *"load_pvalues"*, που περιέχει τα σκορ κάθε λίστας, για κάθε ξεχωριστό μονοπάτι. Είναι ένα hash όπου περιέχει μια εγγραφή για κάθε μονοπάτι με key: το αναγνωριστικό του μονοπατιού και value: μια αναφορά σε hash που, με τη σειρά του, έχει μια εγγραφή για κάθε λίστα με

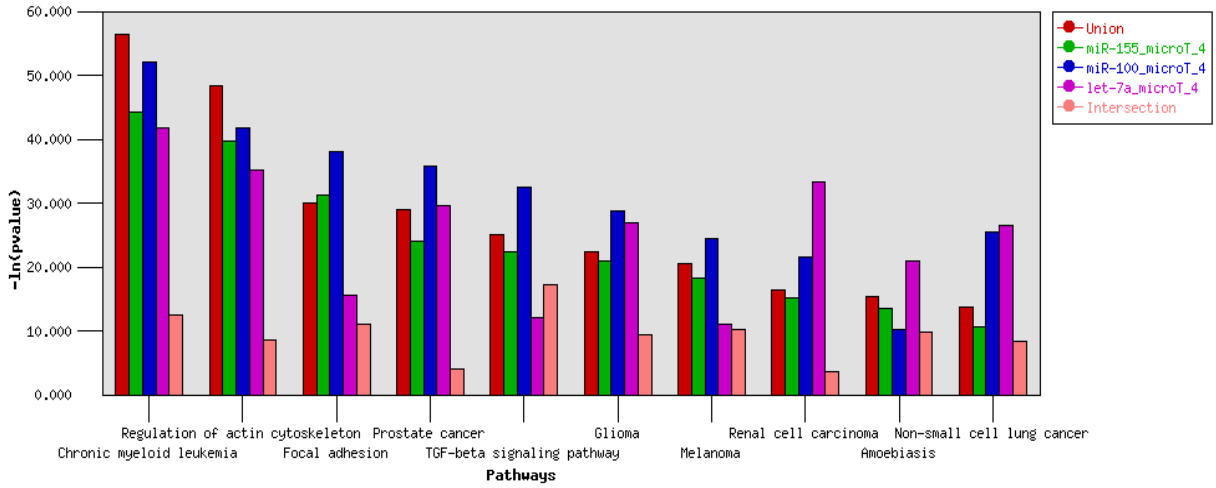
key: το όνομα της λίστας και value: το αντίστοιχο σκορ. Η παραπάνω δομή αποθηκεύεται στη μεταβλητή *“%pnp”*.

$ 4.8 Παραγωγή εξόδου και τερματισμός εφαρμογής

Στο τελευταίο, αλλά όχι λιγότερο σημαντικό κομμάτι, η εφαρμογή προσπαθεί να δημιουργήσει τις γραφικές αναπαραστάσεις των μονοπατιών με τη συμμετοχή των γονιδίων της κάθε λίστας, μια διαδικασία που περιγράφεται στο επόμενο κεφάλαιο. Επίσης, μέσω των συναρτήσεων *"print_output_to_file"* και *"create_graph"*, δημιουργούνται αντίστοιχα ένα spreadsheet αρχείο και ένα γράφημα με μπάρες που παρουσιάζουν τα αποτελέσματα με εναλλακτικό τρόπο. Στιγμιότυπα των παραπάνω φαίνονται στις εικόνες 4.1 και 4.2.

	A	B	C	D	E	
1	Path	Number of genes in >predicted_miR-155 microT_4	>predicted_miR-155 microT_4	Number of genes in >predicted_miR-100 microT_4	>predicted_miR-100 microT_4	Numbe
2	Glycolysis / Gluconeogenesis	6	5.58	0	--	
3	Citrate cycle (TCA cycle)	5	11.13	0	--	
4	Pentose phosphate pathway	1	0.16	0	--	
5	Pentose and glucuronate interconversions	3	3.12	0	--	
6	Fructose and mannose metabolism	5	8.76	1	2.44	
7	Galactose metabolism	3	4.2	0	--	
8	Ascorbate and aldarate metabolism	3	4.43	0	--	
9	Fatty acid elongation in mitochondria	1	2.11	0	--	
10	Fatty acid metabolism	10	31.97	0	--	
11	Synthesis and degradation of ketone bodies	2	7.09	0	--	
12	Primary bile acid biosynthesis	0	--	1	5.75	
13	Ubiquinone and other terpenoid-quinone biosynthesis	0	--	1	12.97	
14	Steroid hormone biosynthesis	3	1.1	0	--	
15	Oxidative phosphorylation	9	4.44	1	0.21	
16	Purine metabolism	13	8.3	3	3.3	
17	Caffeine metabolism	0	--	1	12.97	
18	Pyrimidine metabolism	9	7.53	2	2.72	
19	Alanine, aspartate and glutamate metabolism	5	10.24	0	--	
20	Glycine, serine and threonine metabolism	3	3.28	0	--	
21	Cysteine and methionine metabolism	4	5.31	0	--	
22	Valine, leucine and isoleucine degradation	11	38.23	0	--	
23	Valine, leucine and isoleucine biosynthesis	2	5.64	0	--	
24	Lysine degradation	5	6.61	1	1.92	
25	Arginine and proline metabolism	8	14.15	0	--	
26	Histidine metabolism	3	3.61	0	--	
27	Tyrosine metabolism	4	4.37	0	--	
28	Phenylalanine metabolism	1	0.69	0	--	
29	Tryptophan metabolism	3	2.09	1	2.03	
30	Phenylalanine, tyrosine and tryptophan biosynthesis	1	3.58	0	--	
31	beta-Alanine metabolism	5	12.72	0	--	
32	Taurine and hypotaurine metabolism	2	6.29	0	--	
33	Selenocompound metabolism	2	3.3	0	--	
34	D-Glutamine and D-glutamate metabolism	1	4.53	0	--	
35	D-Arginine and D-ornithine metabolism	1	17.85	0	--	
36	Glutathione metabolism	2	0.36	2	6.4	
37	Starch and sucrose metabolism	4	2.67	0	--	
38	N-Glycan biosynthesis	4	3.28	0	--	

Εικόνα 4.1 : Spreadsheet αρχείο αποτελεσμάτων



Εικόνα 4.2 : Παρουσίαση αποτελεσμάτων με γράφημα

Κεφάλαιο 5
Δομή της κατασκευής και υλοποίηση της γραφική αναπαράστασης

$ 5.1 Εισαγωγή

Το interface της εφαρμογής είναι το κομμάτι όπου ο χρήστης αλληλεπιδρά με αυτή και ο σχεδιασμός του είναι μεγάλης σημασίας μιας και θα πρέπει να αφήνει μια πολύ καλή εμπειρία τόσο όσον αφορά στην κατανόηση της από το χρήστη όσο και στην ευκολία χρήσης της και την σωστή παρουσίαση των αποτελεσμάτων. Αφού, δε, είναι μια web-based εφαρμογή, αυτή θα πρέπει να είναι προσβάσιμη από το διαδίκτυο μέσω φυλλομετρητών (web-browsers). Για το λόγο αυτό, τα λογισμικά που χρησιμοποιούνται για την κατασκευή του interface είναι τα: "html", "javascript και jquery", "css" και "php", που ανήκουν στις τεχνολογίες ιστού, αλλά και η Perl καθώς η υλοποίηση των παραπάνω είναι αναπόσπαστο κομμάτι της εφαρμογής πυρήνα που εξετάστηκε στο προηγούμενο κεφάλαιο.

Το παρόν κεφάλαιο, εκτός από την παρουσίαση και την ανάλυση της δουλειάς που έχει γίνει σχετικά με αυτό το κομμάτι έχει και ως στόχο να λειτουργήσει και σαν εγχειρίδιο για κάποιον που θέλει να χειριστεί (κατανοήσει/παραμετροποιήσει) τον κώδικα της εφαρμογής.

$ 5.2 Γραφική αναπαράσταση βιολογικών μονοπατιών

Τα βιολογικά μονοπάτια παρουσιάζονται σαν συνδεδεμένοι γράφοι που περιέχουν, μεταξύ άλλων, τα γονίδια που εμφανίζονται στο μονοπάτι ως κορυφές του γράφου. Τα γονίδια, παρουσιάζονται στο γράφο ως μικρές ορθογώνιες παραλληλόγραμμες περιοχές που από το σημείο αυτό και έπειτα θα αναφέρονται ως "κουτιά". Σε κάθε κουτί/κορυφή του γράφου μπορεί να συμμετέχουν περισσότερα από ένα γονίδιο. Σε ορισμένα μονοπάτια, τα γονίδια μπορεί να εμφανίζονται ως ακμές (σύνδεση) μεταξύ δύο κορυφών. Η αναπαράσταση των μονοπατιών, δηλαδή οι γράφοι, αλλά και πληροφορίες σχετικά με τις συντεταγμένες των διάφορων στοιχείων που εμφανίζονται σ' αυτά παρέχονται από το KEGG.

$ 5.3 Για κάθε λίστα (εκτός της ένωσης)

Από τον πυρήνα της εφαρμογής, καλείται η συνάρτηση *"create_graphic_representation"* που προσπαθεί να κατασκευάσει την γραφική αναπαράσταση των μονοπατιών σε αρχεία που να είναι αναγνωρίσιμα από τους web-browsers. Εκεί, για κάθε λίστα εκτός της ένωσης (που αποτελεί μια ειδική περίπτωση που αναλύεται παρακάτω) και για κάθε μονοπάτι στο οποίο συμμετέχει τουλάχιστον ένα γονίδιο της κάθε λίστας, εξάγονται αρχικά, μέσω της συνάρτησης *"load_coordinates"* και με τη χρήση του αντίστοιχου .xml αρχείου, οι συντεταγμένες των σημείων όπου τα γονίδια εμφανίζονται πάνω στον γράφο, σε μια δομή που αποτελείται από ένα hash που έχει μια εγγραφή για κάθε γονίδιο του μονοπατιού με key: το αναγνωριστικό του γονιδίου και value: μια αναφορά σε πίνακα όπου στην πρώτη θέση αυτού κρατούνται οι συντεταγμένες του γονιδίου

πάνω στον γράφο και στη δεύτερη θέση υπάρχει ο σύνδεσμος(link) προς την ιστοσελίδα του KEGG όπου αναφέρονται περισσότερες λεπτομέρειες σχετικά με το συγκεκριμένο γονίδιο. Η δομή αυτή αποθηκεύεται στη μεταβλητή *"%coords"*. Ακόμη, κρατάμε και στη μεταβλητή *"$genes_in_path"* (που είναι μια αναφορά σε hash) τα γονίδια της λίστας που εμφανίζονται στο μονοπάτι. Έπειτα, καλείται η συνάρτηση *"create_html_represantation"* η οποία θα κατασκευάσει το αρχείο τύπου "php" του μονοπατιού.

Εκεί, δημιουργείται αρχικά μια δομή (hash), που αποθηκεύεται στη μεταβλητή *"$same_coords"*, όπου για κάθε κουτί του γράφου υπάρχει μια αντιστοίχιση με τα αναγνωριστικά των γονιδίων που συμμετέχουν σ' αυτό. Στην συνέχεια, παράγεται ο javascript και html κώδικας του αρχείου όπου, εκτός από τον υπόλοιπο σχεδιασμό της ιστοσελίδας, τοποθετείται η εικόνα που περιέχει το γράφο του μονοπατιού και ο χάρτης της εικόνας (image-map), μέσω του αρχείου ".conf", που δίνει τη δυνατότητα να δημιουργηθούν "clickable" περιοχές πάνω στην εικόνα ώστε να υπάρχει αλληλεπίδραση μεταξύ των στοιχείων του γράφου και του χρήστη. ***Τα αρχεία conf έχουν ήδη παραχθεί, από scripts που δεν αποτελούν κύριο μέρος αυτής της εργασίας, και περιέχουν το image-map κάθε μονοπατιού.*

Όσων αφορά τα γονίδια της λίστας που συμμετέχουν στο κάθε μονοπάτι, αυτά μπορούν να αναπαρασταθούν πάνω στο γράφο με δύο διαφορετικούς τρόπους. Ο πρώτος, αντικαθιστά το αρχικό κουτί του γράφου, με ένα κουτί ίδιων διαστάσεων, με κίτρινο φόντο και με το όνομα του γονιδίου της λίστας που συμμετέχει εκεί και έχει ως στόχο να δείξει ότι σε εκείνο το σημείο υπάρχει γονίδιο της λίστας. Θα αναφέρονται από εδώ και στο εξής ως "annot-κουτιά". Ο δεύτερος, δημιουργεί ένα αντίστοιχο κουτί, ελαφρώς μεγαλύτερων διαστάσεων, με κόκκινο φόντο και με το όνομα του γονιδίου της λίστας που συμμετέχει σε εκείνο το σημείο και έχει ως στόχο να μπορεί ο χρήστης να εντοπίσει ένα συγκεκριμένο γονίδιο της λίστας πάνω στο γράφο του μονοπατιού. Θα αναφέρονται από εδώ και στο εξής ως "highlights". Οποιαδήποτε στιγμή, καθένας από τους δύο τρόπους μπορεί να είναι ενεργός/ανενεργός ενώ στην περίπτωση που είναι ενεργά και τα δύο κουτιά, τότε το annot-κουτί εμφανίζεται μπροστά από το highlight δημιουργώντας ένα κόκκινο περίγραμμα γύρω από αυτό. Έτσι, στην συνέχεια παράγεται ο κώδικας για κάθε γονίδιο της λίστας που συμμετέχει στο μονοπάτι για κάθε έναν από τους δύο τρόπους. Τα κουτιά υλοποιούνται ως DIVS, με απόλυτες συντεταγμένες ως προς την εικόνα του μονοπατιού, που είναι αποθηκευμένες στη δομή *"%coords"* που περιγράψαμε νωρίτερα και είναι αρχικά αόρατα περιμένοντας από τον χρήστη να τα ενεργοποιήσει. Ειδικότερα, αν σε ένα κουτί συμμετέχει παραπάνω από ένα γονίδιο της λίστας, τότε μπροστά από το όνομα του γονιδίου μπαίνει και το σύμβολο "+" για να επιδείξει αυτή ακριβώς την περίπτωση. Τα annot-κουτιά που παράγονται ανήκουν στην κλάση "annot" ενώ τα highlights ανήκουν σε τρεις κλάσεις: την κλάση highlight, την κλάση που περιγράφεται από το όνομα του γονιδίου και την κλάση που περιγράφεται από το όνομα της λίστας. Εντάσσοντας τα κουτιά σε κλάσεις μας επιτρέπεται να τα χειριστούμε ως προς το ποια θέλουμε να είναι εμφανή και ποια όχι όπως περιγράφεται παρακάτω. Επίσης, για κάθε κουτί παράγεται και ο κώδικας του "tooltip", το οποίο είναι μια περιοχή που εμφανίζεται όταν ο χρήστης αφήσει το ποντίκι πάνω το κουτί και περιέχει τα γονίδια της λίστας που συμμετέχουν στο συγκεκριμένο κουτί όπως και το όνομα της λίστας. Ο κώδικας των highlights όλων των μονοπατιών και για κάθε λίστα, αποθηκεύεται στη μεταβλητή *"$union_highlights"*, που είναι μια αναφορά σε hash, ώστε να χρησιμοποιηθεί στην κατασκευή της γραφικής αναπαράστασης της ένωσης των λιστών αφού η ένωση πρέπει να περιέχει κάθε στοιχείο των υπόλοιπων λιστών.

Τέλος, δημιουργείται ο κώδικας για την κατασκευή του πάνελ από το οποίο ο χρήστης μπορεί να χειρίζεται την εμφάνιση των γονιδίων πάνω στο γράφο. Αρχικά, δημιουργείται το κομμάτι που εμφανίζει/εξαφανίζει τα annot-κουτιά. Δημιουργούνται δύο κουμπιά (buttons) με όνομα *SHOW/HIDE* που είναι συνδεδεμένα με δύο jquery handlers όπου εμφανίζουν/εξαφανίζουν (slide-down/slide-up) πάνω στον γράφο, τα DIVS της κλάσης annot. Αμέσως μετά, δημιουργείται το κομμάτι που χειρίζεται τα highlights. Αρχικά, δημιουργούνται δύο κουμπιά που είναι συνδεδεμένα με δύο jquery handlers και εμφανίζουν/εξαφανίζουν τα highlights που ανήκουν στην κλάση που περιγράφεται από το όνομα της λίστας. Αυτό έχει ως στόχο, την εμφάνιση των highlights συγκεκριμένης λίστας. Έπειτα, δημιουργούνται δύο κουμπιά που είναι συνδεδεμένα με δύο jquery handlers οι οποίοι εμφανίζουν/εξαφανίζουν τα DIVS που ανήκουν στην κλάση "highlights" και έχει ως στόχο την εμφάνιση όλων των highlighs της λίστας πάνω στο γράφο. Όπως γίνεται μάλλον αντιληπτό, οι δύο παραπάνω λειτουργίες δεν επιτελούν κάποια ξεχωριστή λειτουργία στην περίπτωση που εμφανίζεται μονοπάτι λίστας πέρα της ένωσης αλλά παίζουν σημαντικό ρόλο σε διαφορετική περίπτωση. Τέλος, δημιουργείται το κομμάτι από όπου μπορεί να γίνει highlight πάνω στο γράφο, κάθε ξεχωριστό γονίδιο της λίστας που ανήκει σ' αυτό. Έτσι, κατασκευάζονται για κάθε τέτοιο γονίδιο, δύο κουμπιά που είναι συνδεδεμένα με δύο jquery handlers οι οποίοι εμφανίζουν/εξαφανίζουν το higlight που περιγράφεται από το όνομα του συγκεκριμένου γονιδίου.

$ 5.4 Για τη λίστα της ένωσης

Εφόσον η ένωση των λιστών περιέχει κάθε άλλη λίστα της εισόδου του χρήστη, θα πρέπει να δίνεται η δυνατότητα να διαχωρίζονται τα γονίδια που ανήκουν στις υπόλοιπες λίστες και για αυτόν τον λόγο να μπορούν να γίνονται highlight πάνω στον γράφο τα γονίδια κάθε λίστας ξεχωριστά. Γι αυτό, στο κομμάτι του πάνελ που χειρίζεται τα highlights της κάθε λίστας, δημιουργούνται δύο κουμπιά για κάθε λίστα, που είναι συνδεδεμένα με δύο jquery handlers που εμφανίζουν/εξαφανίζουν τα DIVS που ανήκουν στην εκάστοτε κλάση που περιγράφεται από το όνομα της λίστας.
Επιπλέον, αν σε ένα annot-κουτί του γράφου συμμετέχουν πάνω από μία λίστες (πέραν της ένωσης) τότε το φόντο του κουτιού αλλάζει από κίτρινο σε πορτοκαλί για να επιδείξει αυτή την ιδιαιτερότητα. Πλέον και τα tooltips που εμφανίζονται περιέχουν και το ποιες λίστες, πέρα από την ένωση, συμμετέχουν στο συγκεκριμένο κουτί.
Κατά τα άλλα, η δημιουργία της γραφικής αναπαράστασης της ένωσης ακολουθεί τη δομή της δημιουργίας των υπολοίπων που περιγράφηκε στην προηγούμενη ενότητα.

$ 5.5 Παρουσίαση των αποτελεσμάτων

Το τελευταίο κομμάτι που πρέπει αν υλοποιηθεί είναι η παραγωγή της εξόδου που παρουσιάζει τα αποτελέσματα της ανάλυσης στο χρήστη, δηλαδή τα σκορ των μονοπατιών για την εκάστοτε λίστα και χωρίζεται σε δύο μέρη. Στο πρώτο κατασκευάζεται η αρχική σελίδα των αποτελεσμάτων που περιέχει συνοπτικά τα αποτελέσματα για όλες τις λίστες και στο δεύτερο, όπου κατασκευάζεται η παρουσίαση των αποτελεσμάτων για κάθε λίστα ξεχωριστά με αναλυτικές πληροφορίες.

Το πρώτο κομμάτι αρχίζει να υλοποιείται με την κλήση της συνάρτησης *"create_html_start_page"* από την εφαρμογή πυρήνα και κατασκευάζει το αρχείο με τον html κώδικα που περιέχει και τον πίνακα των αποτελεσμάτων. Ο πίνακας αυτός, που ένα στιγμιότυπο του φαίνεται στην επόμενη εικόνα, περιέχει μια εγγραφή για κάθε μονοπάτι που εμφανίζεται στις λίστες και για κάθε λίστα περιέχει τον αριθμό των γονιδίων της λίστας που συμμετέχουν στο μονοπάτι και το σκορ της λίστας για το μονοπάτι. Για την κατασκευή του πίνακα χρησιμοποιούνται δεδομένα που βρίσκονται στις μεταβλητές *"$parameters_of_every_list"* και *"%path_names"* που περιέχουν όλα τα σκορ και τα ονόματα των μονοπατιών αντίστοιχα (περιγράφονται αναλυτικά στο προηγούμενο κεφάλαιο). Κάθε όνομα μονοπατιού είναι σύνδεσμος προς την γραφική αναπαράσταση του μονοπατιού της ένωσης που κατασκευάστηκε νωρίτερα. Ο πίνακας συνδέεται με το jquery plugin *"table_sorter"* ώστε ο χρήστης να μπορεί να διατάξει είτε σε αύξουσα ή σε φθίνουσα σειρά οποιοδήποτε πεδίο του πίνακα. Επίσης, σ' αυτήν την σελίδα εμφανίζεται και το γράφημα με μπάρες που είχε δημιουργηθεί προηγουμένως, ένας σύνδεσμος προς το αρχείο .xls καθώς και σύνδεσμοι που οδηγούν στα αποτελέσματα κάθε λίστας ξεχωριστά.

Το δεύτερο κομμάτι, αρχίζει να υλοποιείται με την κλήση της συνάρτησης *"create_html_result_page"*, όπου κατασκευάζεται μια σελίδα για κάθε λίστα μέσω της συνάρτησης *"create_html_page"*. Εκεί, εμπεριέχεται ένας πίνακας όπου έχει μια εγγραφή για κάθε μονοπάτι που εμφανίζεται τουλάχιστον ένα γονίδιο της λίστας. Η κάθε εγγραφή έχει τα πεδία : Όνομα μονοπατιού *(KEGG PATHWAY)*, τα ονόματα των γονιδίων της λίστας, στην ονοματολογία του KEGG, που συμμετέχουν στο συγκεκριμένο μονοπάτι *(GENE NAME)*, ο αριθμός αυτών των γονιδίων *(FOUND GENES)*, το σκορ *(-ln(p-value))*, τα ονόματα των γονιδίων με τον ensembl αναγνωριστικό τους *(ENSEMBL GENE ID)* και τέλος το αναγνωριστικό του μονοπατιού *(KEGG PATHWAY ID)*. Ο πίνακας συνδέεται και αυτός με το *"table_sorter jquery plugin"* ώστε ο χρήστης να μπορεί να διατάξει τον πίνακα σε φθίνουσα ή αύξουσα σειρά. Για την κατασκευή του πίνακα χρησιμοποιούνται δεδομένα που βρίσκονται στις μεταβλητές *"%parameters_of_every_list", "%path_names", "$keggid_to_name"* και *"$kegg_to_ensemble"* που περιγράφονται στο προηγούμενο κεφάλαιο.

$ 5.6 Αποθήκευση της εξόδου

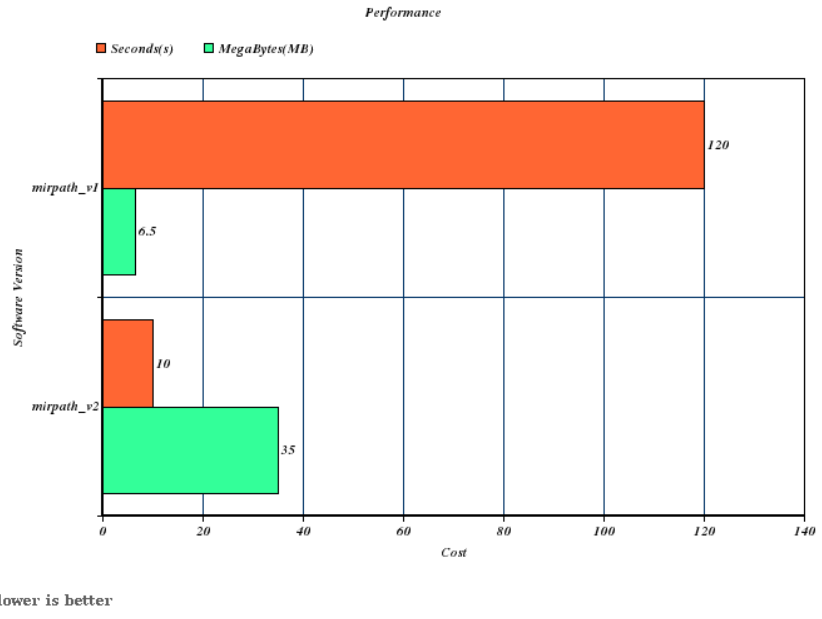
Τα αρχεία που παράγονται από την εφαρμογή, αποθηκεύονται στον web-server στον φάκελο που έχει δημιουργηθεί με όνομα το μοναδικό αναγνωριστικό που έχει δοθεί στην δουλειά του χρήστη. Ειδικότερα, εκεί δημιουργείται ένας υποφάκελος για κάθε λίστα, όπου τοποθετούνται τα αρχεία σχετικά με την εκάστοτε λίστα, τόσο τα αρχεία για κάθε μονοπάτι όσο και το αρχείο με την παρουσίαση των αποτελεσμάτων. Η αρχική σελίδα, το xls αρχείο και η εικόνα με το γράφημα, τοποθετούνται στην κορυφή του φακέλου.

Κεφάλαιο 6
Συμπεράσματα

Αρχικός στόχος της εργασίας ήταν η ανάπτυξη μιας εφαρμογής που να προσδιορίζει τον βαθμό που ένα ή περισσότερα microRNAs επηρεάζουν, ξεχωριστά ή σε συνεργασία, τα βιολογικά μονοπάτια. Η εφαρμογή που υλοποιήθηκε φαίνεται να υπηρετεί στο έπακρο τον συγκεκριμένο στόχο κάνοντας μια ταχεία ανάλυση των δεδομένων και παρουσιάζοντας τα αποτελέσματα αυτής, μέσα από ένα φιλικό προς τον χρήστη περιβάλλον, τόσο με αριθμητικό όσο και με οπτικό τρόπο κάνοντας ευκολότερη την ερμηνεία τους από τον χρήστη.

Ως προς τον τομέα της επίδοσης, η εφαρμογή επιτελεί τους υπολογισμούς σε ένα μέσο χρόνο 10 δευτερολέπτων παράγοντας μια έξοδο μέσου μεγέθους 35MB. Σε σύγκριση με παρόμοια εφαρμογή, που έχει ένα μέσο χρόνο υπολογισμού γύρω στα 2 λεπτά και μια μέση έξοδο των 6.5 MB και γνωρίζοντας ότι στο υπολογιστικό κόστος μεγαλύτερη βαρύτητα έχει σήμερα ο χρόνος παρά το μέσο αποθήκευσης* η εφαρμογή θεωρείται σαφώς βελτιωμένη σε σχέση με την προκάτοχο της.

Τέλος, ένας σημαντικός παράγοντας πιστεύω ότι είναι η συγγραφή του κώδικα. Έχει συνταχθεί ένας σωστά δομημένος (ανοιχτός) κώδικας που είναι εύκολος στην κατανόηση, πράγμα που θα βοηθήσει στην εύκολη , τυχόν, τροποποίηση της εφαρμογής ώστε να επιτελεί και άλλες λειτουργίες. Σαν ένα ενδεικτικό παράδειγμα θα μπορούσε να γίνει πολύ εύκολα το εξής: Η εφαρμογή μπορεί να κάνει ανάλυση για δύο είδη (άνθρωπος, ποντίκι). Με την απόκτηση των αντίστοιχων δεδομένων (γονίδια-στόχοι, μονοπάτια) και με μια πολύ μικρή διαφοροποίηση του κώδικα, η εφαρμογή μπορεί να είναι θέση να επιτελέσει ανάλυση για οποιοδήποτε άλλο γνωστό είδος, πχ. Drosophila melanogaster (fruit fly), Pan troglodytes (chimpanzee) κ.α.



Εικόνα Σ.1: Γράφημα απόδοσης

**Μπορούν να βρεθούν εύκολα και φθηνά, τεράστια μεγέθη αποθηκευτικών μέσων ενώ "ο χρόνος είναι χρήμα"*

Παράρτημα

Goodness-of-fit (GOF) Tests

Έστω ότι ένας πειραματιστής έχει κάποια δεδομένα και μια ιδέα της διαδικασίας από την οποία αυτά παρήχθησαν. Αυτά τα δεδομένα πρέπει να έχουν, σύμφωνα με την ιδέα και το αντίστοιχο μοντέλο, μια πιθανοτική κατανομή μιας συγκεκριμένης οικογένειας. Εάν αυτό δεν είναι αληθές τότε η αρχική εντύπωση (ιδέα) είναι λανθασμένη. Έτσι, ο πειραματιστής χρειάζεται να ελέγξει αν τα δεδομένα προέρχονται από μια συγκεκριμένη πιθανοτική κατανομή ώστε να ισχυροποιήσει την αρχική του υπόθεση. Αυτόν ακριβώς τον ρόλο επιτελούν τα Goodness-of-fit (GOF) tests.

Pearson's chi-squared test

Το Pearson's chi-squared test είναι το δημοφιλέστερο GOF τεστ που ανήκει σε μια γενικότερη οικογένεια τέτοιων τεστ που έχουν σαν βάση τη chi-squred κατανομή και πρώτο-ερευνήθηκε από τον Karl Pearson. Το τεστ αποδεικνύει εάν μια συχνοτική κατανομή διαφέρει ή όχι από την chi-square κατανομή. Η τιμή του τεστ (test-statistic) δίνεται από τον τύπο :

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

όπου x^2 είναι η τιμή του τεστ που ασυμπτωτικά προσεγγίζει την chi-square κατανομή

O_i είναι η συχνότητα που παρατηρείται

E_i είναι η θεωρητική συχνότητα που αναμενόταν.

Στη συνέχεια μπορεί να υπολογιστεί η τιμή p-value συγκρίνοντας την τιμή του τέστ με την κατανομή chi-square.

Τι είναι η Perl

Η Perl είναι μια δημοφιλής, ανοιχτού κώδικα, γλώσσα προγραμματισμού που δημιουργήθηκε από τον Larry Wall. Η πρώτη έκδοση της εκδόθηκε το 1987 και από τότε επεκτείνεται και βελτιώνεται από μια μεγάλη κοινότητα προγραμματιστών.

Ένας πίνακας (array) στην Perl είναι μια μεταβλητή που περιέχει μια ταξινομημένη συλλογή δεδομένων.

Ένα hash στην Perl είναι μια δομή δεδομένων στην οποία μπορεί να αποθηκευτεί και εν συνεχεία να ανακτηθεί οποιοσδήποτε αριθμός τιμών/δεδομένων. Σε αντιδιαστολή με του πίνακες, όπου τα δεδομένα του δεικτοδοτούνται με αριθμούς, στα hashes η δεικτοδότηση γίνεται με βάση το όνομα. Κάθε εγγραφή/δεδομένο στο hash αποτελείται από δύο πεδία. Το πεδίο "key" που η τιμή του αποτελεί και τη δεικτοδότηση της εγγραφής και το πεδίο "value" όπου περιέχονται τα πραγματικά δεδομένα της

εγγραφής.
Μια αναφορά (reference) στην Perl είναι ένας δείκτης προς δεδομένα και όχι τα δεδομένα αυτά καθαυτά.

Html και Javascript

Η Html είναι η κυρίαρχη γλώσσα προγραμματισμού που χρησιμοποιείται για την κατασκευή ιστοσελίδων. Ένας web-browser διερμηνεύει τον κώδικα Html και παρουσιάζει τα περιεχόμενα που αυτός περιγράφει.
Η JavaScript είναι γλώσσα προγραμματισμού η οποία έχει σαν σκοπό την παραγωγή δυναμικού περιεχομένου και την εκτέλεση κώδικα ιστοσελίδων στην πλευρά του πελάτη (client-side). Αρχικά χρησιμοποιήθηκε για προγραμματισμό από την πλευρά του πελάτη (client), που ήταν ο φυλλομετρητής (browser) του χρήστη, και χαρακτηρίστηκε σαν client-side γλώσσα προγραμματισμού. Αυτό σημαίνει ότι η επεξεργασία του κώδικα Javascript και η παραγωγή του τελικού περιεχομένου HTML δεν πραγματοποιείται στο διακομιστή, αλλά στο πρόγραμμα περιήγησης των επισκεπτών (web-browser), ενώ μπορεί να ενσωματωθεί σε στατικές σελίδες HTML. Αντίθετα, άλλες γλώσσες όπως η PHP εκτελούνται στην πλευρά του διακομιστή (server-side γλώσσες προγραμματισμού).

Βιβλιογραφία

- G. L. Papadopoulos, P. Alexiou, M. Maragkakis, M. Reczko, A. G. Hatzigeorgiou . "DIANA-mirPath: Integrating human and mouse microRNAs in pathways" . *Bioinformatics 25.15 (2009), 1991 – 1993*

- Παναγιώτης Αλεξίου. *"Ο ρόλος των micro-RNAs στον καρκίνο". Διδακτορική διατριβή, τμήμα βιολογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2011.*

- Yan-Qinghang and Jagath C. Rajapakse. "Machine Learning in Bioinformatics". *Wiley, 2009*

- William C.S. Cho. "MicroRNAs in Cancer Translational Research". *Springer, 2011*

- Priscilla E. GreenWood and Mikhail S. Nikhulin. "A Guide To Chi-Squared Testing". Wiley, 1996

- Riccardo Russo. "Statistics for the behavioural sciences: an introduction". *Psychology Press, 2003*

- Randal L. Schwartz and Tom Phoenix. "Learning Perl". *O'Reilly , 2008*

- David R. Brooks. "Guide to HTML, JavaScript and PHP: For Scientists and Engineers". *Springer 2011*

- Robin Nixon. "Learning PHP, MySQL, and JavaScript". O'Reilly , 2009

- National Human Genome Research Institute. "Biological Pathways". *March 23, 2011. August 15, 2011.* <*http://www.genome.gov/27530687*>

- Wikipedia. "Pearson's chi-square test". *April 13, 2011. April 16, 2011.* <*http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test*>

- Official documentation for the Perl programming language. "Tutorials". *Perl 5 Porters. March 29, 2011* <*http://perldoc.perl.org/index-tutorials.html*>

- The Comprehensive Perl Archive Network (CPAN). <*http://www.cpan.org/index.html*>

- KEGG: Kyoto Encyclopedia of Genes and Genomes. <*http://www.kegg.jp/*>

- w3schools. "Learn to create websites". <*http://www.w3schools.com/*>