ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΙΕΡΑΡΧΗΣΗ ΓΟΝΙΔΙΑΚΩΝ ΛΙΣΤΩΝ ΔΙΑΦΟΡΙΚΗΣ ΕΚΦΡΑΣΗΣ ΜΕ ΑΝΑΛΥΣΗ ΡΥΘΜΙΣΤΙΚΩΝ ΚΑΙ ΛΕΙΤΟΥΡΓΙΚΩΝ ΔΙΚΤΥΩΝ

Αντώνιος Παπαδάκης

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΪΑΤΡΙΚΗΣ

Επιβλέπων καθηγητής: Χριστόφορος Νικολάου
Ηράκλειο, Οκτώβριος 2016

Περίληψη

Οι τεχνολογίες υψηλής απόδοσης έχουν δημιουργήσει μεγάλο όγκο δεδομένων για τις μοριακές αλληλεπιδράσεις στο εσωτερικό των κυττάρων. Η ιεράρχηση των υποψήφιων γονιδίων για εμφάνιση ασθένειας που έχουν προκύψει από τα δεδομένα αυτά αποτελεί θεμελιώδη πρόκληση για τις βιοϊατρικές επιστήμες, διευκολύνοντας την ανάπτυξη νέων διαγνωστικών μεθόδων και θεραπειών. Ένα σημαντικό υποσύνολο των υπαρχουσών τεχνικών χρησιμοποιεί τα βιολογικά δίκτυα, είτε ρυθμιστικά είτε πρωτεϊνικών αλληλεπιδράσεων, για τη γονιδιακή ιεράρχηση, βασισμένο στην παρατήρηση ότι τα προϊόντα γονιδίων που αλληλεπιδρούν το ένα με το άλλο σε υψηλό βαθμό σε ένα δίκτυο είναι πιθανό να σχετίζονται με παρόμοιες ασθένειες. Η θορυβώδης και μη ολοκληρωμένη φύση των δεδομένων πρωτεϊνικών αλληλεπιδράσεων(Protein-Protein Interactions ή PPI) όμως αποτελεί σημαντική πρόκληση για τις εφαρμογές αυτές.

Στόχος της παρούσας διατριβής είναι η ιεράρχηση υποψήφιων γονιδίων σε κλίμακα ολόκληρου γονιδιώματος χρησιμοποιώντας κατάλληλες υπολογιστικές μεθόδους. Συγκεκριμένα, γίνεται χρήση μεθόδων ιδιοδιανύσματος με σκοπό την αξιοποίηση των υπαρχόντων δεδομένων ρυθμιστικών και λειτουργικών πρωτεϊνικών αλληλεπιδράσεων για την εξαγωγή των σημαντικότερων γονιδίων από λίστες διαφορικής έκφρασης που έχουν προκύψει από πειράματα υψηλής απόδοσης έτσι ώστε να χρησιμοποιηθούν σε μεταγενέστερες πειραματικές εφαρμογές.

Abstract

High-throughput technologies have generated a huge amount of data concerning the molecular interactions that transpire inside the cells. The prioritization of the candidate genes connected to diseases which have been derived from this data constitutes a fundamental challenge for the biomedical sciences, facilitating the development of new diagnostic methods and treatments. A considerable subset of the existing techniques utilizes biological networks, regulatory or protein-protein-protein interaction ones, for gene prioritization, based on the notion that the products of genes that interact heavily in a network are more likely to associate with similar diseases. However, the noisy and incomplete nature of PPI(Protein-Protein Interactions) data remains an important challenge for these applications.

The aim of this thesis is a genome-wide candidate gene prioritization using suitable computational methods. More specifically, eigenvalue algorithms are applied in order to effectively utilize the existing regulatory and functional PPI data for the extraction of the most important genes from differential expression lists derived from high-throughput experiments so that they can be applied in future experimental applications.

Ευχαριστίες

Τα χρόνια του μεταπτυχιακού αποτέλεσαν για εμένα ένα περιπετειώδες διάστημα. Μολαταύτα, μου δόθηκε η ευκαιρία να δουλέψω σε μια μεγάλη ποικιλία εργαστηρίων στο Πανεπιστήμιο και στο Ίδρυμα Τεχνολογίας και Έρευνας και να εξοικειωθώ με τον τομέα της Βιοπληροφορικής, διευρύνοντας τους ορίζοντές των γνώσεών μου.

Υπάρχουν πολλά άτομα που με στήριξαν κατά τη διάρκεια της πορείας μου, από την εκκίνηση μέχρι τη λήξη της παρούσας εργασίας. Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα Ερευνητή της μεταπτυχιακής μου διατριβής Χριστόφορο Νικολάου για την εμπιστοσύνη του σε εμένα κατά τη διάρκειά της, την καθοδήγηση του όποτε αυτή ήταν απαραίτητη αλλά και την ανθρώπινη παρουσία του σε στιγμές ανάγκης.

Ευχαριστώ και τα υπόλοιπα μέλη της τριμελής μου επιτροπής Ιωάννη Ηλιόπουλο και Δημήτρη Τζαμαρία οι οποίοι δέχθηκαν να διαβάσουν και να αξιολογήσουν όχι μόνο την παρούσα διατριβή αλλά και τις γενικές εξετάσεις του μεταπτυχιακού όπως και την Νίκη Κρετσόβαλη και τον Ιωσηφ Παπαματθαιάκη για την αμέριστη στήριξή τους.

Ιδιαίτερα ευχαριστώ οφείλω στο Διονύση Παπαζωγονόπουλο, συνεργάτη και πολύ καλό φίλο, ο οποίος πρόσφερε τεχνική βοήθεια και συμπαράσταση κατά τη διάρκεια της διατριβής. Και άλλα πολλά ευχαριστώ σε όλα τα μέλη του εργαστηρίου για τις κοινές ευχάριστες στιγμές. Τον Αντώνη Κλωνιζάκη, τη Μαρία Μαλλιαρού, τον Άκη Λινάρδο, το Θεόφιλο Χαλκιαδάκη, την Άννα Καραβαγγέλη, το Γιώργο Κολιοπάνο και τον Ανδρέα Αγγελόπουλο.

Πολλά ευχαριστώ στους φίλους μου, χωρίς την παρουσία των οποίων οι εμπειρίες μου εδώ στην Κρήτη θα ήταν φτωχότερες. Γιώργο, Ιάσονα, Ηλία, Δήμητρα, Κατερίνα, Μπάμπη, Δημήτρη, Μαριέτα. Ευχαριστώ πάνω από όλα στη Βίρνα, η στήριξη της οποίας ήταν ανεκτίμητη και χωρίς την οποία μπορεί να μην είχα τελειώσει σήμερα.

Ιδιαίτερα, θερμά και άπειρα ευχαριστώ από τα βάθη της καρδιάς μου στους γονείς μου, Λεωνίδα και Αργυρώ, και στα αδέλφια μου, Αγγελική, Νίκο και Μαρία για την απεριόριστη αγάπη που μου έχουν προσφέρει μέχρι σήμερα..

Πίνακας Περιεχομένων

Κεφάλαιο 1

Εισαγωγή

Στόχος αυτού του κεφαλαίου είναι μια σύντομη εισαγωγή στις βασικές έννοιες της υπολογιστικής βιολογίας που αποτελούν τον πυρήνα της συγκεκριμένης εργασίας. Το μέρος 1 αποτελεί μια περιληπτική ανάλυση των τεχνολογιών μικροσυστοιχιών και γονιδιακής αλληλούχισης νέας γενιάς(next-generation sequencing). Το μέρος 2 ρίχνει μια σύντομη ματιά στη χρήση των βιολογικών δικτύων στη βιολογία συστημάτων. Το μέρος 3 ασχολείται με το πρόβλημα της ιεράρχησης γονιδίων(gene prioritization).

1.1)Τεχνολογίες ανάλυσης μεταγραφώματος

Η κατανόηση του ρόλου και της λειτουργίας των γονιδίων και των πρωτεϊνών τους είναι ένας επιστημονικός τομέας μεγάλου ενδιαφέροντος και παραμένει σημαντική πρόκληση για τις βιολογικές επιστήμες. Παράγοντας αύξησης της δυσκολίας του εγχειρήματος αυτού αποτελεί το γεγονός ότι η γενετική πληροφορία οργανώνεται σε πολλά επίπεδα υψηλής πολυπλοκότητας. Για την αποσαφήνιση της λειτουργικής σημασίας των επιπέδων αυτών, οι τεχνολογίες υψηλής απόδοσης(high-throughput technologies) έχουν προσφέρει καινοτόμα εργαλεία, με παραγωγή καινούριων δεδομένων[1]. Το πεδίο μελέτης αυτών των δεδομένων, τα οποία περιγράφουν ουσιαστικά το σύνολο των βιολογικών διεργασιών, χαρακτηρίζεται με το γενικό όρο *omics*. Το πεδίο αυτό μπορεί να διακριθεί σε περαιτέρω κατηγορίες ανάλογα με τα συγκεκριμένα στοιχεία που βρίσκονται υπό μελέτη.

Μια από τις κατηγορίες είναι τα *transcriptomics* και αναφέρεται στη μελέτη του μεταγραφώματος, δηλαδή του συνόλου των μεταγράφων σε ένα κύτταρο, και την ποσότητά τους σε ένα συγκεκριμένο αναπτυξιακό στάδιο ή κατάσταση φυσιολογίας. Στόχοι αυτής της μελέτης είναι η καταγραφή όλων των ειδών μεταγράφων, ο καθορισμός της μεταγραφικής δομής των γονιδίων και η ποσοτικοποίηση της διαφοράς στα επίπεδα έκφρασης κάθε μετάγραφου μεταξύ διακριτών καταστάσεων.

Καθώς το mRNA(messenger RNA) αποτελεί πρόδρομο μόριο στη διαδικασία της πρωτεϊνικής σύνθεσης, η συγκέντρωση μιας συγκεκριμένης αλληλουχίας mRNA σε ένα κύτταρο ή σε ένα ιστό παρέχει σημαντικές πληροφορίες για το ρυθμό με τον οποίο μεταγράφεται το γονίδιο από το οποίο αυτό προκύπτει. Για την ποσοτικοποίηση του επιπέδου έκφρασης του mRNA, οι κύριες διεργασίες που χρησιμοποιούνται βασίζονται σε προσεγγίσεις υβριδισμού(για παράδειγμα, μικροσυστοιχίες) ή προσεγγίσεις αλληλούχισης(για παράδειγμα, RNA-seq).

1.1.1)Μικροσυστοιχίες DNA

Τις τελευταίες δεκαετίες έχει παρατηρηθεί εκθετική αύξηση στην ποσότητα των βιολογικών δεδομένων, γεγονός που οφείλεται στη ραγδαία ανάπτυξη στην επιστήμη τόσο της πληροφορικής όσο και της βιολογίας[1]. Μέχρι τις αρχές της δεκαετίας του 1990, υπήρχαν αυστηροί περιορισμοί στον αριθμό των γονιδίων που μπορούσαν να παρατηρηθούν σε ένα πείραμα, με χρήση τεχνικών όπως Real Time PCR, Northern Blot ή RNase protection assays[2]. Η ανάπτυξη των τεχνολογιών

ευρείας κλίμακας που ακολούθησε από τα μέσα της δεκαετίας επέτρεψε τη μελέτη της έκφρασης χιλιάδων γονιδίων σε ένα μόνο πείραμα. Τέτοιες τεχνολογίες είναι η MPSS(*massive parallel signature sequencing)*[3], η SAGE(*serial analysis of gene expression)*[4] και η Differential Display[5]. Η κορωνίδα όμως των τεχνικών που αναπτύχθηκαν εκείνη την περίοδο είναι οι γονιδιακές μικροσυστοιχίες(DNA microarrays)[6].

Οι μικροσυστοιχίες DNA περιέχουν πολλά δείγματα DNA για την ανίχνευση και ποσοτικοποίηση των μεταγράφων χιλιάδων γονιδίων ταυτοχρόνως, ανακαλύπτοντας έτσι γονίδια που έχουν σημαντικά διαφορετικό μοτίβο έκφρασης μεταξύ δύο ιστών ή δύο χρονικών σημείων. Η ανίχνευση στηρίζεται στην ιδιότητα του DNA να προσδένεται ειδικά μέσω δημιουργίας ζευγών βάσεων με συμπληρωματικά μόρια DNA ή RNA.

Τα βασικά συστατικά μιας μικροσυστοιχίας είναι

- Ένα επίπεδο, στερεό υπόστρωμα. Αρχικά, τα υποστρώματα ήταν μεμβράνες νάιλον. Πλέον, συχνότερα χρησιμοποιείται ως υλικό γυαλί ή πυρίτιο[7].
- Χιλιάδες μονόκλωνα μόρια DNA με αλληλουχία συμπληρωματική με τα μόρια στόχος. Τα μόρια αυτά είναι συνδεδεμένα μέσω ομοιοπολικού δεσμού με το υπόστρωμα και τοποθετημένα σειριακά σε συγκεκριμένες θέσεις.

Υπάρχουν πολλά διαφορετικά είδη μικροσυστοιχιών τα οποία διαφέρουν σε πειραματικά πρωτόκολλα, μήκος ανιχνευτών και μοτίβο εναπόθεσης των ανιχνευτών ανάλογα με την εφαρμογή για την οποία είναι αυτά απαραίτητα. Έχουν αναπτυχθεί πλατφόρμες όπως:

1) Μικροσυστοιχίες cDNA για κατάρτιση προφίλ για την γονιδιακή έκφραση(gene-expression profiling)[8]
2) Ανοσοκατακρύμνηση χρωματίνης πάνω σε chip(ChIP-on-chip) για την ανακάλυψη in vivo αλληλεπιδράσεων μεταξύ DNA και πρωτεϊνών.[9]
3) Short-oligonucleotide arrays, με σύντομους ανιχνευτές μήκους περίπου 25 bp που συντίθενται in situ απευθείας πάνω στη συστοιχία.[10] Μπορούν να χρησιμοποιηθούν για genotyping(την ταυτοποίηση της γενετικής ποικιλομορφίας μεταξύ ατόμων και πληθυσμών) με την ανίχνευση SNPs(Single Nucleotide Polymorphisms).
4) Array-CGH(Comparative Genomic Hybridization) για ανίχνευση αριθμού αντιγράφων DNA. Οι αλλαγές στον αριθμό αντιγράφων DNA είναι γεγονότα κομβικής σημασίας για την ανάπτυξη καρκίνου.[11]
5) DNA barcoding arrays, όπου χρησιμοποιείται μια συγκεκριμένη περιοχή του γονιδιώματος για την ταυτοποίηση ενός είδους.[12]
6) Tiling arrays, για τον καθορισμό των τομέων του γονιδιώματος που μεταγράφονται υπό συγκεκριμένες συνθήκες[13].

Οι δύο πιο συχνά χρησιμοποιούμενες πλατφόρμες είναι οι μικροσυστοιχίες cDNA και οι μικροσυστοιχίες ολιγονουκλεοτιδίων. Για την κατασκευή συστοιχιών cDNA για έλεγχο γονιδιακής έκφρασης, πραγματοποιείται εναπόθεση κλώνων cDNA ενισχυμένων με PCR. Από την άλλη, οι μικροσυστοιχίες ολιγονουκλεοτιδίων βασίζονται στην in situ δημιουργία των ολιγονουκλεοτιδίων στη επιφάνεια με μεθόδους φωτολιθογραφίας, ηλεκτροχημικής σύνθεσης ή και εκτύπωσης ψεκασμού-μελάνης. Ανεξάρτητα από τη συγκεκριμένη μέθοδο, τα πειράματα με μικροσυστοιχίες μπορούν να διακριθούν σε πέντε στάδια[14].

1. *Επιλογή σωστών μορίων-ανιχνευτών.* Οι ανιχνευτές πρέπει να είναι ευαίσθητοι, να ανταποκρίνονται επαναλήψιμα με τα μόρια-στόχους και να είναι σωστά υπομνηματισμένοι.
2. *Κατασκευή των συστοιχιών.*

3. *Δημιουργία των ανιχνευτών*. Παρασκευάζεται mRNA και σημαίνονται τα νουκλεοτίδια με φθορίζουσες χρωστικές . Εφόσον αναγκαίο, πραγματοποιείται ένα επιπλέον βήμα για την ενίσχυση του δείγματος.
4. *Υβριδοποίηση*. Προσδένονται οι ανιχνευτές με τα γονίδια-στόχους. Ακολουθεί μία αλληλουχία πλύσεων.
5. *Ανάλυση*. Πραγματοποιείται οπτική μέτρηση της έντασης του σήματος φθορισμού με ανιχνευτή laser. Έτσι, με την τεχνική αυτή, επιτυγχάνεται μέτρηση της σχετικής γονιδιακής έκφρασης(με άλλα λόγια, το πηλίκο του σήματος έκφρασης ενός δείγματος Ά με το σήμα έκφρασης ενός δείγματος Β').[14]

Οι εφαρμογές που έχουν οι cDNA συστοιχίες δεν περιορίζονται στη απλή εύρεση διαφορικής γονιδιακής έκφρασης μεταξύ δυο ιστών, δυο ασθενών ή και δύο χρονικών καταστάσεων(πριν και μετά τη λήψη ενός φάρμακου λόγω χάρη). Έχουν δημιουργηθεί ειδικές διαγνωστικές συστοιχίες για την ταυτοποίηση ασθενειών, όπως διαφορετικά είδη καρκίνου[15][16]. Σε άλλες κλινικές εφαρμογές, υπάρχει η δυνατότητα για τη χρήση μικροσυστοιχιών για το σχεδιασμό φαρμάκων, με τη σύνδεση της διαφορικής έκφρασης ενός γονιδίου με την εμφάνιση ή τα συμπτώματα μίας ασθένειας.[17]

Η τεχνολογία των μικροσυστοιχιών βέβαια χαρακτηρίζεται από διάφορους σημαντικούς περιορισμούς. Αρχικά, η ανάγκη για χιλιάδες έως και εκατομμύρια κύτταρα για μία και μόνο μέτρηση σημαίνει ότι το αποτέλεσμα που εξάγεται στην πραγματικότητα αφορά το μέσο όρο ενός πληθυσμού κυττάρων, ο οποίος πιθανώς χαρακτηρίζεται από μεγάλη ετερογένεια. [18] Μικρότερος αριθμός κυττάρων οδηγεί σε μεγαλύτερη ομοιογένεια στο υπό εξέταση υλικό. Επειδή όμως η πειραματική διαδικασία απαιτεί μια συγκεκριμένη ποσότητα πειραματικού υλικού για να μας δώσει αξιόλογα αποτελέσματα, δημιουργείται επίσης η ανάγκη για αύξηση του διαθέσιμου RNA με PCR. Τα βήματα αυτά αύξησης του RNA μπορεί να οδηγήσουν σε απώλεια υλικού ή σε εσφαλμένα υψηλές τιμές συγκέντρωσης[19].

Επιπλέον, επειδή οι αλληλουχίες των ανιχνευτών που χρησιμοποιούνται για να γεμίσουν τα πηγάδια των μικροσυστοιχιών πρέπει να είναι προκαθορισμένες, σημαίνει ότι το μεταγράφωμα του οργανισμού που εξετάζουμε πρέπει να είναι καλά μελετημένο και ότι τα επίπεδα έκφρασης γονιδίων χωρίς αντίστοιχο ανιχνευτή στο υπόστρωμα δεν είναι δυνατό να μετρηθούν. Για αυτό το λόγο, οι μικροσυστοιχίες δεν είναι κατάλληλες για την εύρεση νέων γονιδίων ή ισομορφών. Άλλα προβλήματα περιλαμβάνουν τη δύσκολη σύγκριση επιπέδων έκφρασης μεταξύ διαφορετικών πειραμάτων[2] και τα υψηλά επίπεδα “θορύβου” λόγω cross-hybridization(πρόσδεση των ανιχνευτών σε τμήματα DNA διάφορα του επιθυμητού)[20].

1.1.2)RNA-seq

Όπως είδαμε λοιπόν, οι μελέτες του μεταγραφώματος με μικροσυστοιχίες, αν και αναμφίβολης σημασίας, προσφέρουν περιορισμένη δυνατότητα για την πλήρη ταξινόμηση και ποσοτικοποίηση των μορίων RNA που μεταγράφονται από το γονιδίωμα. Αυτά περιλαμβάνουν, πέρα από τα messenger-RNA, μεγάλη ποικιλία κωδικών και μη-κωδικών μορίων(ncRNAs), όπως τα tRNAs, rRNAs, snRNAs ή snoRNAs αλλά και προϊόντα της pervasive transcription[21]. Μέχρι σχετικά πρόσφατα, η εναλλακτική προσέγγιση των τεχνολογιών αλληλούχισης χαρακτηριζόταν από σχετικά χαμηλή απόδοση και υψηλό κόστος. Όμως, η επινόηση τεχνολογιών αλληλούχισης DNA επόμενης γενιάς(Next Generation DNA Sequencing, NGS) επέτρεψε την ανάλυση RNA μέσω της μαζικής αλληλούχισης συμπληρωματικών cDNA(RNA-seq), φέρνοντας έτσι επανάσταση στη μελέτη της μεταγραφικής περιπλοκότητας.

Σαν πρώτο βήμα για την ανάλυση με RNA-seq, δημιουργείται μια βιβλιοθήκη cDNA από τον πληθυσμό των μορίων RNA που θέλουμε να εξετάσουμε. Κάθε μόριο cDNA έχει αλληλουχίες adaptors στο ένα ή και στα δύο άκρα. Στη συνέχεια, όλα τα μόρια αλληλουχούνται με μεθόδους

υψηλής απόδοσης. Με αυτό τον τρόπο προκύπτουν σύντομες αλληλουχίες(30-400 bp), από το ένα ή και από τα δύο άκρα[2]. Υπάρχει μεγάλος αριθμός τεχνολογιών που χρησιμοποιούνται για την νουκλεοτιδική αλληλούχιση κατά την πραγματοποίηση του RNA-seq, οι οποίες έχουν αναπτυχθεί από εταιρίες όπως η Illumina (GenomeAnalyzer I/II and Hiseq)[27], η Roche(454 Life Science)[2] και η Applied Biosystems(ABI SOLiD).[28] Κάθε τεχνική έχει και διαφορετικό πειραματικό πρωτόκολλο. Τελικά, οι αλληλουχίες είτε στοιχίζονται με ήδη γνωστά γονιδιώματα ή μετάγραφα είτε συναρμολογούνται de novo χωρίς χρήση έτοιμης γονιδιακής αλληλουχίας.

Τα πλεονεκτήματα του RNA-seq σε σχέση με τις μικροσυστοιχίες είναι πολλαπλά.

1) Μπορεί να ανιχνεύει μετάγραφα χωρίς την ανάγκη ύπαρξης κάποιας αντίστοιχης αλληλουχίας. Το γεγονός αυτό καθιστά την τεχνική αυτή ιδιαίτερα χρήσιμη για τη μελέτη καινούριων οργανισμών[22].
2) Έχει ανάγκη σημαντικά λιγότερης ποσότητας RNA, καθώς δε μεσολαβούν χρονοβόρα βήματα κλωνοποίησης[2].
3) Καθώς είναι ποσοτική μέθοδος, επιτρέπει τον καθορισμό της απόλυτης ποσότητας του κάθε μορίου σε ένα πληθυσμό κυττάρων.
4) Έχει, σχετικά με τις μικροσυστοιχίες, υψηλά επίπεδα επαναληψιμότητας και αναπαραγωγιμότητας[23].
5) Έχει υψηλότερη ανάλυση(μέχρι και ενός νουκλεοτιδίου) και χαμηλότερο “θόρυβο” από τις μικροσυστοιχίες.[2]
6) Έχει σημαντικό δυναμικό εύρος για την ποσοτικοποίηση της γονιδιακής έκφρασης. Σε μία μικροσυστοιχία, κάθε πηγάδι περιέχει περιορισμένο αριθμό ανιχνευτών. Αυτό μπορεί να δημιουργήσει φαινόμενα κορεσμού, δημιουργώντας προβλήματα και ανακρίβειες στη μέτρηση επιπέδων έκφρασης γονιδίων με υψηλό ρυθμό μεταγραφής. Τέτοια θέματα δεν υπάρχουν με το RNA-seq.
7) Το κόστος του μειώνεται κάθε χρόνο και σε περιπτώσεις χαρτογράφησης μεταγραφώματος μεγάλων γονιδιωμάτων είναι η φτηνότερη επιλογή.
8) Πέρα από την έκφραση γονιδίων, επιτρέπει τη διερεύνηση εναλλακτικού ματίσματος[21], την ανίχνευση έκφρασης από συγκεκριμένο αλληλόμορφο[25] και την ταυτοποίηση γεγονότων γονιδιακής σύντηξης[26].

Όλα αυτά τα πλεονεκτήματα που παρέχει η τεχνολογία του RNA-Seq έχουν επιτύχει τη δημιουργία μιας άνευ προηγούμενου ολικής ανάλυσης της δομής και της οργάνωσης του μεταγραφώματος σε πληθώρα κυτταρικών τύπων και οργανισμών. Η ευαισθησία και η υψηλή ανάλυση του RNA-Seq έχει αποκαλύψει καινούριες μεταγραφόμενες περιοχές και καινούρια προϊόντα εναλλακτικού ματίσματος. Επιπροσθέτως, είναι υπεύθυνη για τη χαρτογράφηση με ακρίβεια των 5' και 3' ορίων για μεγάλο αριθμό γονιδίων, όπως και την ανακάλυψη 5' και 3' UTR οι οποίες δεν είχαν έως τότε αναλυθεί. Η χαρτογράφηση των μεταγραφικών ορίων αποκάλυψε με τη σειρά της νέα χαρακτηριστικά της γονιδιακής οργάνωσης των ευκαρυωτικών οργανισμών[2]. Μελλοντικοί στόχοι της RNA-Seq συμπεριλαμβάνουν την ανίχνευση των αλλαγών στην έκφραση σπάνιων ισομορφών RNA από όλα τα γονίδια ενός οργανισμού όπως και η στόχευση πιο περίπλοκων μεταγραφωμάτων για τον καθορισμό της δομής τους και της δυναμικής που τα διέπει.

1.1.3)Ανάλυση διαφορικής γονιδιακής έκφρασης

Η αλλαγή στο μέσο επίπεδο έκφρασης ενός γονιδίου μεταξύ δύο διαφορετικών καταστάσεων ονομάζεται διαφορική έκφραση(differential expression)Όπως έχει ήδη εξηγηθεί, και οι δύο κύριες στρατηγικές μεταγραφικής ανάλυσης(RNA-seq, microarrays) έχουν τη δυνατότητα να μας παρέχουν μια λίστα ή ομάδες λιστών διαφορικά εκφρασμένων γονιδίων(Differentially

Expressed Genes ή DEGs). Κατά την εκπόνηση της πτυχιακής αυτής, για ανάλυση και δοκιμή των αλγορίθμων χρησιμοποιήθηκαν λίστες που προέκυψαν και από cDNA microarrays και από αλληλούχιση.

Υπάρχουν διάφοροι, διακριτοί τρόποι για τον καθορισμό της διαφορικής έκφρασης(έστω DE) ενός γονιδίου(έστω g) χρησιμοποιώντας τα επίπεδα έκφρασής του e_g. Ένας από αυτούς είναι ο λόγος των συγκεντρώσεων του γονιδίου μεταξύ δύο καταστάσεων B και A.

$$DE(g)=\frac{B(e_g)}{A(e_g)}$$

Το πηλίκο log2(log2 ratio ή fold change) θεωρεί ότι τα γονίδια που διαφέρουν περισσότερο από ένα αυθαίρετο κατώφλι(πχ όταν η έκφραση υπό μία συνθήκη είναι δύο φορές μεγαλύτερη ή δυο φορές μικρότερη από μία άλλη συνθήκη)υπόκεινται σε διαφορική έκφραση[29].

$$DE_L(g)=\log_2\left(\frac{B(e_g)}{A(e_g)}\right)$$

Στατιστικά, η απλή διαφορά στις τιμές έκφρασης δεν μπορεί να θεωρηθεί στατιστικά σημαντική από μόνη της, καθώς υφίστανται διάφοροι βιολογικοί αλλά και πειραματικοί παράγοντες κατά τη διάρκεια ενός πειράματος οι οποίοι μπορούν να οδηγήσουν σε τέτοια ποικιλομορφία. Αν και η σωστή κανονικοποίηση των δειγμάτων αφαιρεί αρκετούς από τους παράγοντες αυτούς, υπάρχουν βιολογικοί παράγοντες που είναι σχεδόν αδύνατο να αφαιρεθούν[30][31]. Επομένως, αν ληφθεί υπόψη μόνο το fold change για την κατασκευή των λιστών διαφορικής έκφρασης, οι λίστες αυτές θα περιλαμβάνουν μεγάλο αριθμό ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων[29].

Για να γίνει, λοιπόν, χρήση μικροσυστοιχιών ή RNA-seq για στατιστική ανάλυση της διαφορικής γονιδιακής έκφρασης είναι απαραίτητο να πραγματοποιηθούν πολλαπλές επαναλήψεις των πειραμάτων αυτών. Ιδανικά, τα πειράματα με μικροσυστοιχίες θα πρέπει να επαναλαμβάνονται τουλάχιστον τρεις φορές[32] και αυτά των RNA-seq τουλάχιστον έξι[33], παρά το υψηλό κόστος κάθε επανάληψης. Μετέπειτα, υπάρχει μια πληθώρα από στατιστικές μέθοδοι που μπορούν να χρησιμοποιηθούν για την κατασκευή DEG, οι οποίες χωρίζονται σε δύο κύριες κατηγορίες.

Η πρώτη κατηγορία είναι οι παραμετρικές δοκιμασίες[34], όπως το t-test[35] και το Anova[36][37] στην περίπτωση των μικροσυστοιχιών και οι αλγόριθμοι βασισμένοι σε αντίστροφη διωνυμική κατανομή[38][39][40], όπως η edgeR, η DESeq και η BaySeq στην περίπτωση του RNA-seq. Η δεύτερη κατηγορία είναι η μη παραμετρικές δοκιμασίες[41]. Όσον αφορά τις μικροσυστοιχίες, τα μη παραμετρικά τεστ περιλαμβάνουν την Kruskal-Wallis , η Wilcoxon sign rank[42] και το τεστ Mann-Whitney[43], δοκιμασίες που μπορούν να χρησιμοποιηθούν σε συστοιχίες cDNA και ολιγονουκλεοτιδίων. Παράδειγμα μη παραμετρικού τεστ για RNA-seq είναι το SAM-seq[44].

1.2)Βιολογικά δίκτυα στη βιολογία συστημάτων

Τον περασμένο αιώνα, η εστίαση της βιολογικής επιστήμης ήταν η απομόνωση και ανάλυση κάθε ενός μορίου, από τα χιλιάδες που βρίσκονται σε ένα κύτταρο, ξεχωριστά. Σήμερα, οι εξελίξεις στη βιοτεχνολογία έχουν επιτρέψει την επέκταση της επιστημονικής έρευνας σε ένα πιο ολιστικό πλαίσιο. Στη *βιολογία συστημάτων(systems biology)* στοχεύεται η κατανόηση των βιολογικών διαδικασιών σε επίπεδο συστήματος. Μία μέθοδος για την αναπαράσταση των διαφορετικών στοιχείων που απαρτίζουν ένα βιολογικό σύστημα(πρωτεΐνες, γονίδια, βιοχημικές αντιδράσεις) και των αλληλεπιδράσεων τους είναι τα βιολογικά δίκτυα. Χρήσιμες πληροφορίες για τους μηχανισμούς των βιολογικών συστημάτων προκύπτει από την υπολογιστική ανάλυση της

δομής αυτών των δικτύων.

1.2.1)Βιολογικά δίκτυα

Τα βιολογικά δίκτυα μπορούν να διακριθούν σε τρεις κύριες κατηγορίες.

1. Τα *μεταβολικά δίκτυα*[45] περιέχουν τις μεταβολικές και φυσικές διεργασίες που ελέγχουν τα βιοχημικά χαρακτηριστικά ενός κυττάρου. Βασικά δομικά στοιχεία τους είναι οι μεταβολικές χημικές αντιδράσεις και οι αλληλεπιδράσεις των ρυθμιστικών στοιχείων που τις προκαλούν.
2. Τα *δίκτυα αλληλεπίδρασης πρωτεϊνών*[46] τα οποία αποτελούνται από τις κυτταρικές πρωτεΐνες οι οποίες συνδέονται ανάλογα με τη ρυθμιστική ή λειτουργική αλληλεπίδρασή τους.
3. Τα *δίκτυα μεταγραφικής ρύθμισης[47]*. Σε αυτά, χαρακτηρίζονται οι πρωτεΐνες και τα γονίδια που αυτές ρυθμίζουν, αναλύοντας τον τρόπο με τον οποίο τα γονίδια του γονιδιώματος. Τα δίκτυα αυτά είναι κατευθυνόμενα: οι κόμβοι αντιπροσωπεύουν τα γονίδια και οι ακμές κατευθύνονται από το γονίδιο-ρυθμιστή στο γονίδιο που ρυθμίζεται.

Ανεξαρτήτως του είδους, στα βιολογικά δίκτυα φαίνεται να επικρατούν συγκεκριμένα μοτίβα δικτύων. Με άλλα λόγια, κάποια αρχιτεκτονικά μονοπάτια φαίνεται να επαναλαμβάνονται στα βιολογικά δίκτυα με μεγαλύτερη συχνότητα από την αναμενόμενη[48]. Η μελέτη των Uri Alon έδειξε ότι στα βιολογικά δίκτυα κυριαρχούν δύο κύρια είδη μοτίβων. Ένα bi-fan μοτίβο που αποτελείται από τέσσερις κόμβους και ένα feed-forward μοτίβο που αποτελείται από τρεις. Μεταγενέστερη μελέτη που πραγματοποιήθηκε στο ζυμομύκητα[49] προσδιόρισε έξι διακριτά θεμελιώδη μοτίβα των βιολογικών δικτύων: το feed-forward μοτίβο, το μοτίβο ρυθμιστικής αλυσίδας, το μοτίβο μιας εισόδου, το μοτίβο πολλαπλών εισόδων, το αυτορυθμιστικό μοτίβο και το κυκλικό μοτίβο πολλαπλών στοιχείων.

1.2.2)PPI δίκτυα

Οι βάσεις δεδομένων με χαρακτηρισμένες πρωτεϊνικές αλληλεπιδράσεις μπορούν να χρησιμοποιηθούν για την κατασκευή ενός δικτύου πρωτεϊνικών αλληλεπιδράσεων(PPI network). Το δίκτυο πρωτεϊνικών αλληλεπιδράσεων αναπαρίσταται σαν ένας μη κατευθυνόμενος, αζύγιστος γράφος G=(V,E), όπου V ένα σύνολο κόμβων(πρωτεΐνες) και E ένα σύνολο ακμών(αλληλεπιδράσεις). Τα δίκτυα PPI μπορούν να αποτελέσουν χρήσιμες πηγές για την ανακάλυψη νέων, αξιόπιστων αλληλεπιδράσεων. Έχει προταθεί η χρήση τοπολογικών υπολογισμών που βασίζεται στη συνεκτικότητα των γειτόνων, με την υπόθεση ότι δύο πρωτεΐνες είναι πιθανότερο να αλληλεπιδρούν αν έχουν μεγάλο αριθμό κοινών γειτόνων[50]. Οι Saito et al. [51] πρότειναν υπολογισμό γενικότητας για μία αλληλεπίδραση, βασισμένο στην ιδέα ότι αλληλεπιδράσεις μεταξύ πρωτεϊνών με πολλούς κοινούς γείτονες είναι πιθανόν ψευδώς θετικές, εκτός από όσες από αυτές δημιουργούν κλειστό βρόχο ή είναι διασυνδεδεμένες σε μεγάλο βαθμό, οι οποίες έχουν μεγάλη πιθανότητα να είναι αληθώς θετικές. Αυτός είναι τοπικός υπολογισμός, που λαμβάνει υπόψη μόνο τους άμεσους γείτονες μιας πρωτεΐνης.

Την τελευταία δεκαετία, έχουν δημιουργηθεί πολλές υπολογιστικές μέθοδοι, με καταβολή από το χώρο της στατιστικής και της μηχανικής μάθησης για την αύξηση της ακρίβειας των προβλέψεων. Συνοπτικά, αυτές περιλαμβάνουν Μπεϋζιανούς ταξινομητές[52][53], τεχνητά νευρωνικά δίκτυα[54], ταξινομητές Μηχανών Διανυσματικής Υποστήριξης[55][56] και τυχαία δάση[57].

1.3)Ιεράρχηση γονιδίων

Πολλές ασθένειες παρουσιάζουν μεγάλο αριθμό συμπτωμάτων επειδή προκαλούνται από πολλά γονίδια και περιβαλλοντικούς παράγοντες που διαφέρουν από άτομο σε άτομο και από το ένα στάδιο της ασθένειας στο άλλο. Ο υψηλός αυτός βαθμός πολυπλοκότητας αντανακλά φαινόμενα επίστασης, όπου ορισμένα γονίδια επιδρούν στην έκφραση πολλών άλλων. Η αναγνώριση καινούριων γονιδίων που έχουν αιτιακό ρόλο για μία ασθένεια ή μια γενετική διαταραχής παραμένει σημαντική πρόκληση για τις βιολογικές επιστήμες. Έχουν γίνει βέβαια στο παρελθόν μεγάλες ανακαλύψεις στον τομέα αυτό, με μελέτες ανάλυσης σύνδεσης για την ταυτοποίηση περιοχών του χρωμοσώματος που σχετίζονται με τον υπό μελέτη φαινότυπο[58]. Μολαταύτα, οι τεχνολογίες αυτές επιστρέφουν μια μεγάλη λίστα γονιδίων, ελάχιστα από τα οποία έχουν πραγματική σχέση με το φαινότυπο που μας ενδιαφέρει[59].

Η πειραματική επιβεβαίωση όλων των υποψήφιων γονιδίων θα ήταν όχι μόνο πολύ δαπανηρή, αλλά και εξαιρετικά χρονοβόρα. Η πειραματική επιβεβαίωση ενός και μόνο γονιδίου υπεύθυνου για ασθένεια μπορεί να πάρει ένα χρόνο ή και παραπάνω[60]. Για αυτό το λόγο, έχει πραγματοποιηθεί σημαντική δουλειά από τη βιοπληροφορική κοινότητα πάνω στην ιεράρχηση γονιδίων(gene prioritization). Αυτή είναι μια μέθοδος για την ταυτοποίηση των πιο σημαντικών γονιδίων που σχετίζονται με ένα φαινότυπο κάνοντας χρήση τεχνικών υπολογιστικής βιολογίας. Αναφέρεται ως όρος στη βιβλιογραφία για πρώτη φορά στην έρευνα των Perez-Iratxeta *et al.*[61]

Πολλές διακριτές υπολογιστικές μέθοδοι ιεράρχησης γονιδίων έχουν ήδη επινοηθεί[61][62][63][64]. Τα κριτήρια που χρησιμοποιεί η κάθε μέθοδος για την ιεράρχηση είναι διαφορετικά. Οι περισσότερες όμως μέθοδοι βασίζονται στην αρχή guilt-by-association, σύμφωνα με την οποία δίνεται προτεραιότητα στην ταξινόμηση σε γονίδια που παρουσιάζουν ομοιότητες με άλλα γονίδια που είναι ήδη γνωστό ότι εμπλέκονται με το φαινότυπο που εξετάζεται[65]. Οι ομοιότητες αυτές δεν περιορίζονται μόνο σε δεδομένα αλληλεπίδρασης. Μπορούν να επεκταθούν σε οποιοδήποτε γενετικό δεδομένο. Η κύρια διαφορά λοιπόν των τεχνικών ιεράρχησης εμπίπτει στις πηγές δεδομένων που αυτές χρησιμοποιούν, οι οποίες κυμαίνονται από βιβλιογραφικές αναφορές και δεδομένα έκφρασης, μέχρι πρωτεϊνικές αλληλεπιδράσεις και γονιδιακή οντολογία[66][67].

1.3.1)Ροή εργασίας στην ιεράρχηση γονιδίων

Πρώτο βήμα για την ιεράρχηση είναι η κατασκευή της λίστας των υποψήφιων γονιδίων(candidate genes) που επιθυμούμε να ιεραρχηθούν. Αυτή η λίστα μπορεί να προκύπτει από πρωτογενή, πειραματικά δεδομένα(π.χ λίστες DEGs), αλλά αυτό δεν είναι απαραίτητο. Με την ευρεία διαθεσιμότητα δευτερογενών δεδομένων, υπάρχει πλέον η δυνατότητα να γίνει πρώτα ιεράρχηση και ανάλυση των δεδομένων αυτών. Έτσι, ο μελλοντικός πειραματικός σχεδιασμός γίνεται πιο στοχευμένος. Πιο συγκεκριμένα, λίστες μπορούν να προκύψουν από χρωμοσωμικές ανωμαλίες, γενετικούς τόπους υπό μελέτες GWA ή sequencing variants. Είναι δυνατό να πραγματοποιηθεί ιεράρχηση ακόμη και σε ολόκληρο το γονιδίωμα, με το μειονέκτημα ότι έτσι προκύπτουν πολύ μεγάλες λίστες γονιδίων, με μεγάλο ποσοστό ψευδά θετικών αποτελεσμάτων[67].

Δεύτερο βήμα είναι η επιλογή των κριτηρίων με τα οποία θα γίνει η ιεράρχηση. Αυτή εξαρτάται από το είδος των υποψήφιων γονιδίων και του βιολογικού ερωτήματος για το οποίο αναζητείται απάντηση. Επιπλέον, ανάλογα με το βιολογικό ερώτημα και το είδος των πειραμάτων που θα πραγματοποιηθούν μετά την ιεράρχηση, θα είναι διαφορετικός ο αριθμός τόσο των υποψήφιων γονιδίων όσο και των γονιδίων που τελικά θα χρησιμοποιηθούν σε downstream διαδικασίες. Διαφορετικά κριτήρια θα επιλεχθούν αν ερευνάται μονογονιδιακό χαρακτηριστικό σε σχέση με μια ασθένεια, για την οποία μπορεί να είναι υπεύθυνα δεκάδες γονίδια. Ένας τελευταίος

παράγοντας που πρέπει να ληφθεί υπόψη για την επιλογή σωστών κριτηρίων είναι η ήδη υπάρχουσα γνώση για το υπό μελέτη βιολογικό φαινόμενο. Πλατφόρμες ιεράρχησης που είναι κατάλληλες για την εύρεση καινούριων γονιδίων σχετικών με μονοπάτια τα οποία έχουν ήδη χαρτογραφηθεί εκτενώς μπορεί να μην προσφέρουν τη δυνατότητα χαρακτηρισμού γονιδίων όταν οι γνώσεις για τη μοριακή βάση της ασθένειας είναι περιορισμένες.

Συγκεκριμένα, τα ιεραρχικά κριτήρια έχουν κατά κύριο λόγο τη μορφή είτε λέξεων-κλειδιών(keywords) είτε γονιδίων για τα οποία είναι ήδη γνωστό ότι εμπλέκονται με ένα συγκεκριμένο φαινότυπο(seed genes). Η συλλογή των seed genes, αν και σαφώς πιο χρονοβόρα από την αντίστοιχη των keywords, επιτρέπει τη διατύπωση περίπλοκων ερωτημάτων με τρόπο άμεσο και σχετικά ευέλικτο, επιτρέποντας επίσης την εύρεση σχέσεων που μπορεί να διέφευγαν εναλλακτικά από την ανακάλυψη. Πάντως, ανεξάρτητα από τη μορφή των κριτηρίων που θα χρησιμοποιηθούν, η επιλογή τους πρέπει να γίνει προσεχτικά ώστε να έχουν τη μέγιστη δυνατή πληροφορία, με όσο το δυνατό πιο συγκεκριμένο τρόπο. Με άλλα λόγια, πρέπει να δοθεί προτεραιότητα σε seed genes που έχουν άμεση σύνδεση με τον υπό έρευνα φαινότυπο ή keywords με ισχυρή συσχέτιση με αυτόν. Για την εύρεση keywords αλλά και seed genes, χρήσιμες είναι οι βάσεις δεδομένων που περιέχουν φαινοτυπικές πληροφορίες. Μερικά δωρεάν παραδείγματα τέτοιων βάσεων είναι:

- Η OMIM(Online Mendelian Inheritance in Man), η οποία συλλέγει πληροφορίες για γενετικές διαταραχές με Μεντελική κληρονομικότητα[68]
- Η GoPubMed, η οποία χρησιμοποιεί γνώσεις γενετικής οντολογίας με σκοπό τη συσχέτιση γονιδίων και οντολογικών όρων με φαινοτύπους και βιολογικές διεργασίες, μέσω των πληροφοριών που περιέχονται στη Medline. Η αλληλεπίδραση με τη Medline πραγματοποιείται μέσω της PubMed, μηχανή έρευνας που παρέχει πρόσβαση σε μία συλλογή από παραπάνω από δεκατέσσερα εκατομμύρια περιλήψεις βιοϊατρικής επιστημονικής βιβλιογραφίας[69][70].
- Η Genetic Association Database, η οποία εστιάζει σε μελέτες συσχέτισης περίπλοκων ασθενειών[71].
- Η Phenopedia, που παρέχει πληροφορίες για τα γονίδια που σχετίζονται με μια ασθένεια ή ένα φαινότυπο με μορφή πίνακα[72].
- Η KEGG Disease, που αντλεί δεδομένα από την Kyoto Encyclopedia of Genes and Genome[73].

Τρίτο βήμα είναι η επιλογή κατάλληλης υπολογιστικής στρατηγικής. Γενικά, τα βιοπληροφορικά εργαλεία ιεράρχησης παράγουν τα αποτελέσματά τους είτε φιλτράροντας τα δεδομένα με σκοπό την απόκτηση ενός μικρότερου υποσυνόλου γονιδίων από το αρχικό, είτε προσδίδοντάς τους βαθμολογία και κατατάσσοντάς τα ανάλογα με αυτήν[74]. Τα εργαλεία που εκτελούν φιλτράρισμα(επιλέγουν μόνο τα γονίδια που πληρούν προϋποθέσεις οι οποίες έχουν καθοριστεί από το χρήστη, απορρίπτοντας όλα τα άλλα υποψήφια γονίδια. Παραδείγματα τέτοιων εργαλείων είναι το Biofilter[75] και το TEAM[76]. Ο κύριος περιορισμός της στρατηγικής αυτής, γνωστής και ως στρατηγική ab initio, είναι ο υψηλός αριθμός ψευδώς αρνητικών αποτελεσμάτων.

Οι στρατηγικές κατάταξης αποφεύγουν τον περιορισμό αυτό, κατατάσσοντας όλα τα υποψήφια γονίδια από περισσότερο σε λιγότερο υποσχόμενα, σύμφωνα με την ομοιότητά τους με τα seed genes ή τις λέξεις-κλειδιά που έχουν επιλεγεί. Οι στρατηγικές αυτές μπορούν να διακριθούν περαιτέρω σε τρεις κύριες υποκατηγορίες:

1. *Κατάρτιση προφίλ ομοιότητας(Similarity Profiling)[77]*

Η κατάταξη των υποψήφιων γονιδίων γίνεται σύμφωνα με την ομοιότητά τους με άλλα γονίδια τα οποία σχετίζονται με βεβαιότητα με το φαινότυπο. Συνήθως, γίνεται χρήση ταυτόχρονα και στρατηγικών σύντηξης δεδομένων(data fusion) από διάφορες πηγές, οι οποίες

συμπεριλαμβάνουν GO annotations[78] και λεπτομέρειες για φαινοτύπους ασθενειών[79]. Αφού συγκεντρωθούν πληροφορίες από διάφορες πηγές, γίνεται ιεράρχηση για κάθε μία από αυτές και η τελική κατάταξη βγαίνει από την ολοκλήρωση των διαφορετικών ιεραρχήσεων. Επιγραμματικά, παραδείγματα διαδικτυακών πλατφορμών που εκτελούν ιεράρχηση βάση ομοιοτήτων είναι τα Endeavour[80], ToppGene[81], SUSPECTS[82] και PROSPECTR[62].

2. *Εξόρυξη Κειμένου(Text Mining)[79]*

Με τη χρήση λέξεων-κλειδιών που συνδέονται με ένα φαινότυπο, γίνεται ανάκτηση εγγράφων που σχετίζονται με το φαινότυπο αυτό. Μετέπειτα, από το σύνολο των εγγράφων αυτών, γίνεται εξόρυξη σχετικών γονιδίων, τα οποία και κατατάσσονται μέσω στατιστικών μεθόδων βάση της σχετικότητας τους με τη ανακτημένη πληροφορία. Κύριο μειονέκτημα των στρατηγικών αυτών είναι ότι από τη φύση τους είναι περιορισμένες από την ήδη υπάρχουσα βιοϊατρική βιβλιογραφία και επομένως έχουν περιορισμένη χρησιμότητα για τη διατύπωση ριζικά νέων προβλέψεων. Μερικά παραδείγματα αποτελούν οι πλατφόρμες GLAD4U[83], GeneProspector[84] και Genie[85].

3. *Ανάλυση Δικτύων(Network Analysis)[86]*

Δημιουργείται δίκτυο των seed genes και πραγματοποιείται κατάταξη των υποψήφιων γονιδίων ανάλογα με την απόστασή τους από τα γνωστά γονίδια. Η απόσταση υπολογίζεται με χρήση πληροφοριών είτε τοπικού[87] είτε ολικού[88] δικτύου. Οι πηγές των δεδομένων σε αυτή την περίπτωση είναι μόνο δίκτυα, είτε πρωτεϊνικών αλληλεπιδράσεων(PPIN) είτε λειτουργικών συνδέσεων(για παράδειγμα η βάση STRING) . Παραδείγματα ιεράρχησης μέσω ανάλυσης δικτύου είναι τα PINTA[89] και GeneWanderer[90].

Έχοντας λοιπόν επιλέξει κατάλληλα υποψήφια γονίδια, πηγές δεδομένων και στρατηγικές προτεραιοποίησης, απαραίτητο βήμα πριν την εκκίνηση της ιεράρχησης, όπως και για κάθε πειραματική διαδικασία. είναι ο καθορισμός control. Αρχικά, πρέπει να κατασκευαστεί ένα αρνητικό training set(από keywords ή seed genes) το οποίο σχετίζεται επιβεβαιωμένα με ένα άλλο φαινότυπο. Αν η ιεράρχηση των υποψηφίων γονιδίων δίνει παρόμοια αποτελέσματα ανεξαρτήτως του training set, τότε υποδεικνύεται ότι υπάρχει κάποια συστημική προκατάληψη προς συγκεκριμένα γονίδια και ως αποτέλεσμα η ιεράρχηση δεν είναι αξιόπιστη[62]. Άλλος τρόπος ελέγχου της ποιότητας των αποτελεσμάτων είναι η σύγκριση της λίστας που προκύπτει όταν εφαρμόζεται ιεράρχηση στα υποψήφια γονίδια με αυτή που προκύπτει από ολόκληρο το γονιδίωμα. Αν τα κορυφαία γονίδια στην πρώτη λίστα δε βρίσκονται σε υψηλές θέσεις στη δεύτερη λίστα, τότε υπάρχει μεγάλη πιθανότητα η ιεράρχηση να μην ανταποκρίνεται με την πραγματικότητα. Τέλος, άλλη μια επιλογή για επιβεβαίωση της διαδικασίας είναι η διασταύρωση των κορυφαίων γονιδίων που προκύπτουν της ιεράρχησης με βάσεις δεδομένων λειτουργικού εμπλουτισμού, όπως οι κατηγορίες Gene Ontology. Οι εμπλουτισμένοι όροι θα πρέπει να σχετίζονται με το φαινότυπο ή τη βιολογική διαδικασία που βρίσκεται υπό έρευνα.

1.3.2)Διαθέσιμες πλατφόρμες ιεράρχησης

Υπάρχει πλέον πληθώρα υπολογιστικών εργαλείων ιεράρχησης. Θα ακολουθήσει μια συνοπτική περιγραφή ενός ενδεικτικού δείγματος από τα διαθέσιμα εργαλεία που περιγράφονται στη βιβλιογραφία.

Οι Perez-Iratxeta et al.[61] ολοκλήρωσαν μία από τις πρώτες απόπειρες ιεράρχησης με έναν αλγόριθμο ο οποίος συνέδεε φαινότυπο με γονότυπο και φίλτραρε τα υποψήφια γονίδια μέσω

αναζήτησης στη Medline. Έχει χρησιμοποιηθεί για την ανακάλυψη σχέσεων που είχαν γονίδια με 455 κληρονομήσιμες ασθένειες.

Οι Rossi et al.[91] δημιούργησαν το **TOM**(Transcriptomics of OMIM). Το TOM χρησιμοποιεί πληροφορίες αλληλουχίας και χαρτογράφησης για την ταυτοποίηση υποψήφιων γονιδίων σε μία ορισμένη χρωμοσωμική περιοχή. Στη συνέχεια, τα υποψήφια γονίδια φιλτράρονται βάση της ομοιότητας λειτουργικών και οντολογικών δεδομένων με τα γονίδια που είναι ήδη γνωστό ότι σχετίζονται με την ασθένεια. Η πλατφόρμα είναι διαθέσιμη ελεύθερη στο διαδίκτυο.

Οι Adi et al.[62] κατασκεύασαν το **PROSPECTR**, ένα εργαλείο που χρησιμοποιεί μια μεγάλη ποικιλία χαρακτηριστικών αλληλουχίας για την πρόβλεψη της συσχέτισης ενός γονιδίου με μία ασθένεια. Αυτά τα χαρακτηριστικά συμπεριλαμβάνουν τη δομή των γονιδίων και των προϊόντων τους και εξελικτικά δεδομένα, όπως το ρυθμό μεταλλαξιγένεσης. Το PROSPECTR χρησιμοποιεί ταξινομητή δέντρου αποφάσεων. Η ίδια ομάδα προγραμμάτισε το SUSPECTS[82], το οποίο είναι δωρεάν διαθέσιμο στο διαδίκτυο και χρησιμοποιεί επιπλέον και δεδομένα υπομνηματισμού. Θεωρείται από τους ίδιους τους κατασκευαστές[82] βελτίωση ως προς το PROSPECTR.

Από τους Chen et al.[62] δημιουργήθηκε το **ToppGene**, το οποίο συνδυάζει δεδομένα από φαινότυπο ποντικού με υπομνηματισμό ανθρώπινων γονιδίων και βιβλιογραφικά στοιχεία. Πιο συγκεκριμένα, τα είδη των υπομνηματισμών από τα οποία αντλεί στοιχεία είναι οι GO, MEDLINE, Mammalian Phenotype, Protein Domain, Protein Interactions και Pathway.

Το **Endeavour**, το οποίο πρωταγωνιστεί γενικότερα στο πεδίο της γονιδιακής ιεράρχησης, κατασκευάστηκε από τους Aerts et al[80]. Εμπεριέχει πολλαπλές πηγές δεδομένων για γονίδια και πρωτεΐνες, οι οποίες συμπεριλαμβάνουν λειτουργική ανάλυση, πληροφορίες βιοχημικών μονοπατιών και πειράματα μικροσυστοιχιών. Τα υποψήφια γονίδια κατατάσσονται ανάλογα με την ομοιότητά τους με το training set βάση των δεδομένων αυτών.

Οι Hutz et al.[63] δημιούργησαν το **CANDID**, ένα εργαλείο ιεράρχησης για την κατάταξη γονιδίων σε σχέση με μία συγκεκριμένη ασθένεια. Χρησιμοποιεί και αυτό πολλαπλές πηγές δεδομένων, συμπεριλαμβανομένων πρωτεϊνικών αλληλεπιδράσεων, βιβλιογραφία, προφίλ έκφρασης γονιδίων, και περιγραφές πρωτεϊνικών τομέων. Το σκορ που λαμβάνει κάθε υποψήφιο γονίδιο στηρίζεται στην ομοιότητα του γονιδίου με τα χαρακτηριστικά που σχετίζονται με την υπό μελέτη ασθένεια για κάθε πηγή δεδομένων. Η τελική βαθμολογία προκύπτει ανάλογα με το “βάρος” που δίνεται από το χρήστη για κάθε πηγή δεδομένων.

Οι Radivojac et al[92] σχεδίασαν το **Phenopred**, έναν αλγόριθμο που συνδυάζει το δίκτυο των ανθρώπινων πρωτεϊνικών αλληλεπιδράσεων, τις πρωτεϊνικές αλληλουχίες, λειτουργικά και φυσικοχημικά μοριακά χαρακτηριστικά και οντολογίες που σχετίζονται με ασθένειες για τον εντοπισμό συσχετισμών μεταξύ γονιδίων και ασθενειών. Έχει χρησιμοποιηθεί ως προς την ανίχνευση τέτοιων συσχετισμών για εκατοντάδες ασθένειες.

Για την ιεράρχηση γονιδίων σε ένα συγκεκριμένο γενετικό τόπο, δημιουργήθηκε από τους George et al[93] μια μεθοδολογία γνωστή ως **CMP**(Common Module Profiling), η οποία εφαρμόζει τον αλγόριθμο SSEARCH(βασισμένο στον αλγόριθμο στοίχισης των Smith και Waterman) για να υπολογίσει την ομοιότητα μεταξύ τομέων της υποψήφιας πρωτεΐνης με τομείς πρωτεϊνών που σχετίζονται με ασθένειες.

Οι Oti et al.[94] χρησιμοποίησαν ένα PPIN για την αναζήτηση γονιδίων συσχετισμένων με μία δεδομένη ασθένεια. Αρχικά, οι ερευνητές ταυτοποίησαν τους συντρόφους αλληλεπίδρασης ενός δεδομένου γονιδίου ασθένειας σε ένα PPIN. Αν ένας από αυτούς βρίσκεται μέσα σε ένα ή παραπάνω χρωμοσωμικούς τόπους ενός γονιδίου ασθένειας, τότε θεωρείται υποψήφιο για την ασθένεια. Συνολικά, έγιναν περίπου 300 προβλέψεις υποψηφίων γονιδίων και η ακρίβεια των προβλέψεων ελέγχθηκε χρησιμοποιώντας ήδη γνωστά γονίδια ασθένειας.

Το **Genie**[95] αποτελεί μια μέθοδο ιεράρχησης βασισμένη στη βιβλιογραφία, η οποία εξορύσσει περιλήψεις της Medline και εξερευνά ορθόλογα για την ταυτοποίηση περισσότερων περιλήψεων που περιέχουν σχετικά ορθόλογα γονίδια για τη συμπλήρωση των γονιδίων

υπομελετημένων οργανισμών. Για την ανεύρεση ορθόλογων γονιδίων χρησιμοποιήθηκε Μπεϋζιανός ταξινομητής.

1.4)Στόχος

Στόχος της διατριβής είναι η κατασκευή ενός αλγορίθμου γονιδιακής ιεράρχησης λίστας διαφορικής έκφρασης υποψήφιων γονιδίων που να βασίζεται στην κεντρικότητά τους σε ένα δίκτυο πρωτεϊνικών αλληλεπιδράσεων. Το δίκτυο αυτό προκύπτει από την εξαγωγή PPIN με πειραματικά και βιβλιογραφικά επιβεβαιωμένες αλληλεπιδράσεις *Mus musculus* και *Homo sapiens* από τη βάση δεδομένων STRING. Τα γονιδιακά δεδομένα που περιέχονται στη STRING είναι αποθηκευμένα σε μορφή identifiers. Δευτερεύων, αλλά απαραίτητος, στόχος είναι η μετάφρασή τους σε μορφή συμβατή με τα εισαγόμενα δεδομένα υποψήφιων γονιδίων.

Ο αλγόριθμος ιεράρχησης είναι εμπνευσμένος από τον αλγόριθμο PageRank, ο οποίος έχει πλούσιες βιβλιογραφικές αναφορές για χρήση σε δίκτυα από ένα ευρύ φάσμα επιστημονικών τομέων. Επιπροσθέτως, αναζητήθηκε η επίτευξη οπτικοποίησης των αποτελεσμάτων τις ιεράρχησης και του δικτύου πρωτεϊνικών αλληλεπιδράσεων σε μορφή αναγνώσιμη από μη έμπειρους χρήστες αλλά και η συμβατότητα με ανεπτυγμένα προγράμματα απεικόνισης δικτύων. Ο αλγόριθμος είναι προέκταση του προγράμματος λειτουργικού εμπλουτισμού RNEA. Εφαρμόζεται και σε ρυθμιστικά δίκτυα που προκύπτουν από τη χρησιμοποίηση του RNEA στις λίστες διαφορικής έκφρασης.

Κεφάλαιο 2

Υλικά και μέθοδοι

2.1) Λογισμικό

Στην ενότητα αυτή θα αναλυθεί το λογισμικό που χρησιμοποιήθηκε κατά την εκπόνηση της διατριβής. Το κύριο εργαλείο, το οποίο χρησιμοποιήθηκε και για την εργασία του προγραμματισμού, είναι η γλώσσα R. Πέρα από την R, θα παρουσιαστεί συνοπτικά και το Cytoscape, το πρόγραμμα το οποίο ήταν υπεύθυνο για τη γραφική κατασκευή των δικτύων που προέκυψαν από τον αλγόριθμο, και η γλώσσα SQL, η οποία ήταν χρήσιμη για την εφαρμογή του αλγορίθμου ιεράρχησης PageRank.

2.1.1) R

Η R είναι γλώσσα προγραμματισμού και ταυτοχρόνως περιβάλλον για τη διαχείριση και γραφική αναπαράσταση δεδομένων. Μπορεί να θεωρηθεί μία “διάλεκτος” της γλώσσας S. Η S αναπτύχθηκε από τον John Chambers και άλλους στα Bell Telephone Laboratories ως περιβάλλον στατιστικής ανάλυσης, γραμμένο σε Fortran[96]. Το 1988 ξαναγράφτηκε στη C και συνέχισε να μετεξελίσσεται μέχρι το 1998, όποτε βγήκε η 4η έκδοση η οποία χρησιμοποιείται μέχρι και σήμερα[97].

Σε αντίθεση με την S, η R είναι δωρεάν λογισμικό, διαθέσιμο στο κοινό μέσω της GNU (General Public License). Το γεγονός ότι το συντακτικό της R ήταν παρόμοιο με την S (και άρα προσβάσιμο στους χρήστες της S) σε συνδυασμό με τη λειτουργικότητα της R σε όλα τα περιβάλλοντα χρήσης και με τη διαθεσιμότητά της ως open source λογισμικό, οδήγησε στη γρήγορη υιοθέτηση της R από την επιστημονική κοινότητα.[98]

Η R λοιπόν έχει διάφορα χαρακτηριστικά που την καθιστούν κατάλληλη για την ανάπτυξη αλγορίθμων ανάλυσης βιολογικών δεδομένων. Είναι αρκετά ευέλικτη και εύχρηστη για να αντιμετωπίσει τα πολλά και ποικιλόμορφα βιοπληροφορικά και υπολογιστικά προβλήματα, παρέχοντας ανώδυνη πρόσβαση σε μεγάλο αριθμό βάσεων δεδομένων με βιολογικό περιεχόμενο. Χωρίς άλλα υποστηρικτικά προγράμματα, το περιβάλλον λειτουργίας της επιτρέπει την εφαρμογή στατιστικής ανάλυσης σε μεγάλη ποσότητα δεδομένων και την κατασκευή με ακρίβεια μοντέλων και γραφημάτων. Τέλος, επειδή υπάρχει ήδη μεγάλος όγκος βιβλιογραφίας με χρήση της R σε έρευνες στην υπολογιστική βιολογία όπως και μία δραστήρια και αναπτυσσόμενη κοινότητα ερευνητών, είναι πιο εύκολο να βρεθούν απαντήσεις και διέξοδοι σε προβλήματα προγραμματιστικής φύσεως. Παρέχεται λοιπόν μια ανεπτυγμένη συλλογή πακέτων τα οποία απλοποιούν σημαντικά διάφορες επιμέρους διαδικασίες.

2.1.2) Cytoscape

Το Cytoscape είναι μια εφαρμογή η οποία επιτρέπει στο χρήστη να δημιουργήσει δίκτυα ή να φορτώσει δίκτυα από πίνακες ή βάσεις δεδομένων[99]. Είναι αυτόνομη εφαρμογή, προγραμματισμένη σε γλώσσα Java, της οποίας κύριος στόχος είναι η οπτικοποίηση και

παρουσίαση δικτύων μοριακών αλληλεπιδράσεων. Έχει τη δυνατότητα να δημιουργεί δίκτυα με πολλούς κόμβους, κατευθυνόμενα και μη. Επιτρέπει την επιλογή και περαιτέρω ανάλυση υποσυνόλου του δικτύου, βάσει ορισμένων από το χρήστη χαρακτηριστικών και κριτηρίων. Επιπλέον, υποστηρίζει διάφορα εύχρηστα εργαλεία γραφικής τροποποίησης των δικτύων, όπως μεγάλη ποικιλία χρωμάτων και ευελιξία στη χωροταξική τοποθέτηση των κόμβων. Τέλος, το μεγαλύτερο πλεονέκτημα του σε σχέση με αντίστοιχες εφαρμογές είναι η μεγάλη επεκτασιμότητά του, η οποία σε συνδυασμό με την πληθώρα διαθέσιμων έτοιμων αλγορίθμων από την επιστημονική κοινότητα, καθιστά δυνατή την ολοκλήρωση αναλύσεων υψηλής δυσκολίας[100].

2.1.3)SQL

Η SQL(Structured Query Language ή Γλώσσα Δομικής Αναζήτησης) είναι μια γλώσσα υπολογιστών η οποία χρησιμοποιείται για τη διαχείριση σχεσιακών βάσεων δεδομένων και την εκτέλεση πράξεων και μετασχηματισμών στις πληροφορίες που αυτή περιέχει[101]. Σχεδιάστηκε το 1974, όταν η μηχανική λογισμικού βρισκόταν ακόμη σε πρώιμα στάδια και η αυξανόμενη πολυπλοκότητα των δεδομένων είχε δημιουργήσει σημαντικά προβλήματα στον τομέα της υπολογιστικής. Κατασκευάστηκε ως μία δηλωτική γλώσσα[101], προσβάσιμη σε μη εξειδικευμένο προσωπικό. Βάση της SQL είναι η σχεσιακή άλγεβρα

Εάν ληφθούν υπόψη και οι διάφορες εναλλακτικές εκδόσεις της, η SQL είναι η πιο ευρέως χρησιμοποιούμενη γλώσσα βάσης δεδομένων. Υποστηρίζεται από τις περισσότερες σχεσιακές DBMSs, όπως η Oracle[102], η Microsoft SQL Server[103], η MySQL[104], η Sybase[105] και η Informix[106]. Ένα από τα πιο θεμελιώδη στοιχεία της γλώσσας είναι η έννοια της query(επερώτησης) η οποία επιτρέπει την επικοινωνία και το χειρισμό μιας βάσης, με τον έλεγχο των πληροφοριών που αυτή περιέχει.

2.2) Βάσεις δεδομένων βιολογικού περιεχομένου

Για τη διατύπωση μιας υπόθεσης περί των σχέσεων που υφίστανται μεταξύ των γονιδίων σε μια λίστα DEGs, είναι χρήσιμη η ανάκτηση και ο συνδυασμός πληροφοριών που είναι αποθηκευμένες σε βιολογικές βάσεις δεδομένων. Χρησιμοποιήθηκαν διάφορα είδη βάσεων για το RNEA αλλά και για τον αλγόριθμο που κατασκευάσαμε, οι οποίες μπορούν να διακριθούν σε δύο κατηγορίες ανάλογα με τον τρόπο που αντιπροσωπεύουν την πληροφορία που περιέχουν: οι σχεσιακές βάσεις και οι οντολογικές-σημασιολογικές βάσεις[107].

2.2.1)TRRUST

Η TRRUST(transcriptional regulatory relationships unravelled by sentence-based text-mining) περιέχει ρυθμιστικές αλληλεπιδράσεις μεταξύ μεταγραφικών παραγόντων και των στόχων τους[108]. Αυτές έχουν ταυτοποιηθεί από την επιμέλεια περιλήψεων της βάσης δεδομένων Medline με την εφαρμογή προσέγγισης text-mining βασισμένη σε προτάσεις. Ειδικότερα, πραγματοποιήθηκε εξαγωγή και μετά επεξεργασία προτάσεων που σχετίζονταν δυνητικά με μεταγραφική ρύθμιση. Μάλιστα, η TRRUST είναι παγκοσμίως η μεγαλύτερη βάση δεδομένων ανθρώπινων ρυθμιστικών αλληλεπιδράσεων που προέρχονται από βιβλιογραφία.

2.2.2)TRED

Η TRED(Transcriptional Regulatory Element Database) δημιουργήθηκε με σκοπό την ακριβή και πλήρη καταγραφή των cis- και trans- ρυθμιστικών στοιχείων σε υπό μελέτη θηλαστικά, συνεισφέροντας στην κατανόηση της γονιδιακής ρύθμισης[109]. Για τον υπομνηματισμό υποκινητών από το ολικό γονιδίωμα τριών διαφορετικών ειδών(ανθρώπου, ποντικού και αρουραίων)πραγματοποιήθηκε υπολογιστικά εξαγωγή γνωστών υποκινητών από διάφορες βάσεις δεδομένων, όπως η EPD και η Genbank. Ακολούθησε χειρωνακτικός έλεγχος για την αξιολόγηση

της ακρίβειας των δεδομένων και της αλγοριθμικής πρόβλεψης. Στην TRED καταγράφονται και επιπρόσθετες πληροφορίες σχετικά με τις πειραματικές επιβεβαιώσεις που έχουν γίνει για τα ρυθμιστικά στοιχεία και την πρόσδεση μεταγραφικών παραγόντων σε αυτά. Τα δεδομένα είναι οργανωμένα σε μορφή γονιδιακών ρυθμιστικών δικτύων(GRNs).

2.2.3)TFactS

Η TFactS συνδυάζει λίστες γονιδίων από διάφορες βάσεις δεδομένων, όπως τις TRED, TRRD, PAZAR και NFIregulomeDB, με συμπληρωματικές πληροφορίες από τη βιβλιογραφία[110]. Είναι διαδικτυακό εργαλείο το οποίο δέχεται υποψήφιες λίστες γονιδίων και τις συγκρίνει με καταγεγραμμένα γονίδια-στόχους για την ανίχνευση μεταγραφικών παραγόντων. Οι αλληλεπιδράσεις που περιέχει αλληλεπικαλύπτονται σε μεγάλο βαθμό από την TRED.

2.2.4)TarBase

Η TarBase είναι μια βάση δεδομένων η οποία παρέχει πληροφορίες για τις αλληλεπιδράσεις miRNA με γονίδια-στόχους[111] που έχουν προκύψει από την επιστημονική βιβλιογραφία. Οι επιβεβαιωμένες αλληλεπιδράσεις περιγράφονται όχι μόνο από το miRNA και το mRNA που αυτό στοχεύει αλλά και από το είδος του πειράματος που χρησιμοποιήθηκε για την επιβεβαίωση της αλληλεπίδρασης και το αποτέλεσμα της, δηλαδή μεταγραφική καταστολή ή καταστροφή του στόχου.

2.2.5)ORegAnno

Η OregAnno(Open Regulatory Annotation Database) σχεδιάστηκε για τη βελτίωση τη πρόσβασης σε πληροφορίες σχετικά με ρυθμιστικές περιοχές οι οποίες έχουν επιβεβαιωθεί πειραματικά[112]. Δημιουργήθηκε το 2006 από τους Montgomery et al. και επιτρέπει ανοικτό υπομνηματισμό από όλα τα μέλη της επιστημονικής κοινότητας.

Στόχος είναι η εύκολη διερεύνηση μεγάλου όγκου δεδομένων, όντας η μοναδική βάση στην οποία έχουν ταυτοχρόνως ενσωματωθεί ρυθμιστικές αλληλουχίες, σημεία πρόσδεσης μεταγραφικών παραγόντων και γονιδιακή ποικιλομορφία. Μάλιστα, αυτό επιτυγχάνεται με δομημένο τρόπο, επιτρέποντας θετικά και αρνητικά αποτελέσματα. Επιπλέον, η γρήγορη αναβάθμισή της με κάθε καινούρια ανακάλυψη που προκύπτει και η χρήση ορολογίας και γονιδιακής ταυτοποίησης ευρείας χρήσης διαβεβαιώνουν τη μέγιστη συμβατότητά της με τις παροντικές και μελλοντικές ανάγκες της κοινότητας.

2.2.6)Gene Ontology

Γενικότερα, οι οντολογίες παρέχουν στην επιστημονική κοινότητα, μέσω ενός συγκεκριμένου λεξιλογίου για την περιγραφή όρων και των μεταξύ τους σχέσεων, ένα προσβάσιμο τρόπο για τη μετάδοση της γνώσης. Η γνώση που είναι αποθηκευμένη στις σημασιολογικές-οντολογικές βάσεις δεδομένων της βιοπληροφορικής κοινότητας επιτρέπει τη διεξαγωγή μελετών ανώτερου επιπέδου στα διάφορα γονίδια που απαρτίζουν τις λίστες διαφορικής έκφρασης. Οι διάφορες σημασιολογικές πληροφορίες συνδυάζονται και αναλύονται για την ανεύρεση νέων στατιστικών αποδείξεων γύρω από τα θεμελιώδη χαρακτηριστικά των λιστών αυτών. Υπάρχουν διάφορα τέτοια εξειδικευμένα λεξιλόγια που χρησιμοποιούνται από την υπολογιστική κοινότητα για την αναπαράσταση βιολογικών οντοτήτων. Συνοπτικά, αυτά μπορούν να διακριθούν σε τρεις κατηγορίες: λειτουργική ανάλυση, ανάλυση βιοχημικής οδού και η ανάλυση βιβλιογραφίας.

Η GO βάση δεδομένων περιέχει όρους γονιδιακής οντολογίας που χρησιμοποιούνται για λειτουργική ανάλυση, ανάλυση δηλαδή η οποία σχετίζεται με τη λειτουργία των γονιδίων και των προϊόντων τους καθώς και τη μεταξύ τους αλληλεπίδραση[113]. Παρέχει μια δομή που οργανώνει τα γονίδια σε ομάδες, σύμφωνα με τρία διαφορετικά κριτήρια:

1) Τη βιολογική διαδικασία στην οποία αυτά συμμετάσχουν
2) Τη μοριακή λειτουργία την οποία αυτά πραγματοποιούν

3) Τη θέση στην οποία αυτά βρίσκονται στο κύτταρο

Η οντολογία είναι κατασκευασμένη με τη μορφή γραφήματος με κόμβους και ακμές. Οι κόμβοι αντιπροσωπεύουν λειτουργικούς όρους και οι ακμές τις ιεραρχικές σχέσεις μεταξύ των κόμβων. Οι όροι γίνονται πιο συγκεκριμένοι όσο πιο χαμηλά βρίσκεται κάποιος στο γράφημα. Κάθε πρωτεΐνη της UniProt είναι συσχετισμένη με όρους της GO για τη λειτουργία της πρωτεΐνης. Η βάση είναι διαθέσιμη από το Gene Ontology Consortium.

2.2.7)Kyoto Encyclopedia of Genes and Genome

Η KEGG είναι μια συλλογή βάσεων δεδομένων που περιέχουν πληροφορίες για γονιδιώματα, ασθένειες, φάρμακα, βιολογικές οδούς και χημικές ουσίες.[114] Πιο αναλυτικά, ανάλογα με τα δεδομένα που περιέχουν, οι βάσεις τις KEGG μπορούν να κατηγοριοποιηθούν ως εξής:

- Πληροφορίες συστήματος
    - PATHWAY — Χάρτες μονοπατιών για λειτουργίες κυττάρων και οργανισμών
    - MODULE — Λειτουργικές μονάδες γονιδίων
    - BRITE — Ιεραρχικές ταξινομήσεις βιολογικών οντοτήτων
- Πληροφορίες γονιδιωμάτων
    - GENOME — Ολόκληρα γονιδιώματα
    - GENES — Γονίδια και πρωτεΐνες
    - ORTHOLOGY — Ορθόλογες ομάδες γονιδίων
- Χημικές πληροφορίες
    - COMPOUND, GLYCAN — Χημικές ενώσεις
    - REACTION, RPAIR, RCLASS — Χημικές αντιδράσεις
    - ENZYME — Ονοματολογία ενζύμων
- Πληροφορίες υγείας
    - DISEASE — Ανθρώπινες ασθένειες
    - DRUG — Φάρμακα
    - ENVIRON — Λοιπές ουσίες που σχετίζονται με το χώρο της υγείας

2.2.8)STRING

Η STRING(*Search Tool for the Retrieval of Interacting Genes/Proteins*) είναι μία ελεύθερη βάση δεδομένων που περιέχει μια συλλογή γνωστών και προβλεπόμενων λειτουργικών συσχετισμών μεταξύ πρωτεϊνών, οι οποίες πηγάζουν από γονιδιακές πληροφορίες, πειράματα υψηλής απόδοσης, συνέκφραση και εξόρυξη κειμένου[115]. Πρόσφατα ενσωματώθηκαν στη STRING δύο νέες μέθοδοι για την πρόβλεψη λειτουργικών συσχετισμών: γεγονότα γονιδιακής σύντηξης και genetic profiling.

Η STRING καλύπτει πάνω από 1100 οργανισμούς, με εμβέλεια από προκαρυώτες μέχρι των άνθρωπο. Επιπλέον, παρέχει ένα σκορ εμπιστοσύνης, καθοδηγώντας τους χρήστες που θέλουν να βρουν μια ισορροπία ανάμεσα σε κάλυψη και ακρίβεια. Το σκορ εμπιστοσύνης κυμαίνεται από το 1 έως το 1000 και μπορεί να δημιουργηθεί για κάθε ξεχωριστή μέθοδο πρόβλεψης, κάτι που επιτρέπει εύκολη και άμεση σύγκριση των διαφορετικών μεθόδων. Επιπλέον, μπορεί επίσης να υπολογιστεί ως συνδυαστικό σκορ, ενσωματώνοντας όλες τις μεθόδους, επιτρέποντας την εξαγωγή μίας εύκολα κατανοητής βαθμολογίας για κάθε ατομικό συσχετισμό.

Ουσιαστικά, το σκορ εμπιστοσύνης της STRING δηλώνει το αναμενόμενο κλάσμα των αληθώς θετικών αλληλεπιδράσεων. Για παράδειγμα, ένα κατώφλι σκορ >=700(επιλογή όλων των συσχετισμών με σκορ μεγαλύτερο ή ίσο με το 700) σημαίνει ότι το 70% των επιλεγμένων αλληλεπιδράσεων αναμένεται να είναι αληθώς θετικές. Το σκορ δημιουργήθηκε με benchmarking κάθε ενός από τα προβλεπόμενα σκορ εμπιστοσύνης κάθε μεθόδου με την βάση δεδομένων KEGG. Η τελική έκδοση της STRING χρησιμοποιεί επτά διαφορετικά είδη αποδείξεων: γειτονικότητα, σύντηξη, συνεμφάνιση, συνέκφραση, πειράματα, βάσεις δεδομένων και εξόρυξη κειμένου.

2.3) Αλγόριθμος PageRank

Ο αλγόριθμος PageRank στη βασική του μορφή αναπτύχθηκε για τον υπολογισμό της σημασίας μιας ιστοσελίδας[116]. Στη συνέχεια, η σημασία αυτή χρησιμοποιείται για την κατάταξη των ιστοσελίδων. Πιο συγκεκριμένα, το αποτέλεσμα του αλγόριθμου είναι μία αριθμητική τιμή που αντιπροσωπεύει τη σχετικότητα της ιστοσελίδας γενικότερα στον παγκόσμιο ιστό με βάση τις άλλες σελίδες στις οποίες αυτή αναφέρεται. Ο αριθμός αυτός έχει αναλογική σχέση με τον αριθμό των συνδέσμων που υπάρχουν προς τη συγκεκριμένη ιστοσελίδα από άλλες ιστοσελίδες. Δεν αυξάνουν όμως όλοι οι σύνδεσμοι εξίσου τον αριθμό. Όσο πιο σημαντική είναι η σελίδα που έχει σύνδεσμο με την αρχική ιστοσελίδα, τόσο πιο πολύ αυξάνει τη σημασία της. Ο αλγόριθμος PageRank χρησιμοποιήθηκε από την εταιρεία Google για τον καθορισμό της σημασίας όλων των προσβάσιμων σελίδων στο παγκόσμιο δίκτυο και την κατάταξή τους στα αποτελέσματα της μηχανής αναζήτησης ανάλογα με το σκορ τους.

Αναλυτικότερα, το διαδίκτυο αντιπροσωπεύεται από ένα κατευθυνόμενο δίκτυο, στο οποίο οι κόμβοι είναι οι ιστοσελίδες. Επομένως, ένας κατευθυνόμενος σύνδεσμος μεταξύ δύο κόμβων περιγράφει το σύνδεσμο μεταξύ δύο ιστοσελίδων. Ο υπολογισμός του σκορ γίνεται με τη χρήση αλυσίδας Markov, ενός στοχαστικού μοντέλου το οποίο περιγράφει μια αλληλουχία γεγονότων. Η πιθανότητα κάθε γεγονότος εξαρτάται μόνο από την κατάσταση του προηγούμενου γεγονότος και όχι αθροιστικά από την κατάσταση όλων των προηγούμενων γεγονότων. Η αλληλουχία των γεγονότων ονομάζεται και τυχαίος βηματισμός(random walk). Στη συγκεκριμένη περίπτωση, η αλυσίδα Markov περιγράφει μια αλληλουχία επισκέψεων από σελίδα σε σελίδα κατά την οποία η πιθανότητα της επίσκεψης σε μια ιστοσελίδα εξαρτάται μόνο από την ιστοσελίδα στην οποία πραγματοποιήθηκε επίσκεψη στο προηγούμενο βήμα.

Έστω v μια ιστοσελίδα και PR(v) η τιμή PageRank της ιστοσελίδας αυτής. Έστω επίσης B_v το σύνολο που περιέχει όλες τις ιστοσελίδες με συνδέσμους προς την σελίδα v και L(v) ο αριθμός των συνδέσμων από τη σελίδα v. Ο βασικός αλγόριθμος PageRank μπορεί να εκφραστεί ως εξής

$$PR(v)=\sum_{v\in B_v}\frac{PR(v)}{L(v)}$$

Ο τελικός αλγόριθμος λαμβάνει υπόψη και ένα παράγοντα σίγησης(damping factor), ο οποίος αντιπροσωπεύει την πιθανότητα διακοπής, που αναλογεί σε οποιοδήποτε βήμα, στην πιθανότητα ένας φανταστικός διαδικτυακός χρήστης ο οποίος ακολουθεί τυχαία συνδέσμους να σταματήσει την πλοήγηση από σελίδα σε σελίδα Ο παράγοντας σίγησης d γενικά στις περισσότερες μελέτες καθορίζεται περίπου στο 0.85. Ο παράγοντας σίγησης αφαιρείται από το 1, το αποτέλεσμα διαιρείται με τον αριθμό των συνολικών σελίδων και το πηλίκο προστίθεται στο γινόμενο του παράγοντα σίγησης και του αθροίσματος των εισερχόμενων βαθμολογιών PageRank. Άρα, η τελική μορφή του PageRank έχει ως εξής:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

2.4 RNEA

Το RNEA είναι ένα πρόγραμμα που εκτελεί ανάλυση εμπλουτισμού ρυθμιστικών και λειτουργικών δικτύων. Ειδικότερα, συνδυάζει βιβλιογραφικές και πειραματικές γνώσεις για την κατασκευή ενός δικτύου αναφοράς με αλληλεπιδράσεις και εξάγει τα πιο σχετικά υποδίκτυα που σχετίζονται με μια πειραματική διαδικασία. Είναι προγραμματισμένο στην R και υποστηρίζει σαν δεδομένα εισόδου λίστες διαφορικής έκφρασης ποντικών και ανθρώπων.

Οι βάσεις δεδομένων που χρησιμοποιεί το RNEA για την κατασκευή του δικτύου αναφοράς περιλαμβάνουν τις TRED, TFactS, TRRUST, OREGANNO και TARBASE για ρυθμιστικά δεδομένα και τις GO και KEGG για λειτουργικά. Οι βάσεις αυτές έχουν όλες εξηγηθεί διεξοδικά σε προηγούμενη ενότητα. Το RNEA, λόγω της ευκολίας στη χρήση και το έτοιμο πλαίσιο ανάγνωσης λιστών διαφορικής έκφρασης, κρίθηκε κατάλληλο για την εφαρμογή του τροποποιημένου αλγορίθμου PageRank.

Κεφάλαιο 3

Ανάλυση και εφαρμογή αλγορίθμου

Σκοπός του κεφαλαίου αυτού είναι η συνοπτική παρουσίαση και περιγραφή του κώδικα που γράφτηκε ως επέκταση των δυνατοτήτων του RNEA πακέτου στα πλαίσια της εκπόνησης μεταπτυχιακής διατριβής.

3.1) Εξόρυξη δεδομένων από τη βάση STRING

Το πρώτο βήμα για την επίτευξη των στόχων που έχουν τεθεί ήταν ο εμπλουτισμός των βάσεων δεδομένων από τις οποίες αντλεί πληροφορίες το RNEA με πληροφορίες από τη STRING, την πιο ολοκληρωμένη πηγή πρωτεϊνικών αλληλεπιδράσεων. Τα δεδομένα αυτά για τις πρωτεϊνικές αλληλεπιδράσεις όλων των οργανισμών είναι διαθέσιμα ελεύθερα από την ιστοσελίδα της STRING.

Στα πλαίσια της πειραματικής διαδικασίας, έπρεπε να εξορυχθούν οι πρωτεϊνικές αλληλεπιδράσεις δύο μόνο ειδών: του Mus musculus(ποντικός) και Homo sapiens(άνθρωπος). Επιπλέον, για λόγους μείωσης του όγκου των βάσεων, αύξησης της ταχύτητας επεξεργασίας και ευκολίας στη ανάλυση, επιτεύχθηκε η διαγραφή των αριθμών μπροστά από το όνομα των πρωτεϊνών. Αυτό το πρώτο βήμα πραγματοποιήθηκε με χρήση της γλώσσας προγραμματισμού Perl. Ακολουθεί αναλυτικότερη ανάλυση του προγράμματος που εξόρυξε τα δεδομένα του Mus musculus.

```
#!/usr/bin/perl –w
use strict;
```

Κατεύθυνση του κελύφους του λειτουργικού συστήματος(Ubuntu) στο μονοπάτι όπου βρίσκεται ο μεταγλωττιστής/διερμηνέας της Perl για τη σωστή εκτέλεση του προγράμματος και εντολή για την ανεύρεση τυχόν σφαλμάτων στον κώδικα.

```
open IN, “, protein.links.detailed.v10.txt” or die “Can't
open IN file!\n”;
open OUT, “>mouse.txt” or die “Can't create OUT file\n”;
```

Άνοιγμα της βάσης δεδομένων που περιέχει τις πληροφορίες για τις σχέσεις μεταξύ των πρωτεϊνών και δημιουργία κενού αρχείου κειμένου στο οποίο θα τοποθετηθεί το υποσύνολο των δεδομένων τα οποία προορίζονται για μετέπειτα χρήση.

```
while (<IN>){
     if ($_=~ /10090.([^\s]+)\s10090.([^\s]+\s[^\s]+\s[^\s]
+\s[^\s]+\s[^\s]+\s[^\s]+\s[^\s]+\s[^\s]+\s[^\s]+)/){
     print OUT $1,” “,$2,”\n”;
}
else {print “no mouse\n”;}
}
```

Ο αριθμητικός κωδικός ταυτοποίησης των πρωτεϊνικών δεδομένων που αφορούν τον ποντικό ειναι 10090. Με τις εντολές αυτές, το πρόγραμμα διαβάζει κάθε γραμμή των δεδομένων και κρατάει μόνο αυτά που σχετίζονται με το Mus Musculus, αφαιρώντας ταυτοχρόνως τους αριθμούς που προηγούνται των ονομάτων των γονιδίων.

```
close (IN);
close (OUT);
exit;
```

Κλείσιμο των αρχείων και τερματισμός του προγράμματος.

Με παρόμοιο τρόπο, γίνεται εξόρυξη των ανθρώπινων δεδομένων.

Τα αρχεία mouse.txt και human.txt περιέχουν όλες τις αλληλεπιδράσεις μεταξύ πρωτεϊνών του ποντικού και του ανθρώπου αντίστοιχα. Αποφασίστηκε, για λόγους όγκου δεδομένων και ακριβείας να διατηρηθούν μόνο οι αλληλεπιδράσεις με σκορ μεγαλύτερο του 900, δηλαδή 90% των οποίων υπολογίζεται ότι είναι αληθώς θετικές.

```
genes <-read.table(“mouse.txt”, header= FALSE)
tab1rows = read.table(“protein.links.detailed.v10.txt”,
header=TRUE, nrows=1)
names(genes) <-colnames(tab5rows)
overnine<-subset(genes, combined_score <=900)
write.table(overnine, “Mouse_TF.tsv” row.names=FALSE)
```

Επιλογή του υποσυνόλου των γονιδίων στα οποία το τελικό σκορ είναι πάνω από 900 και αποθήκευση του σε νέο πίνακα, με όνομα Mouse_TF.

3.2)Αναζήτηση των υποψήφιων γονιδίων στη βάση δεδομένων STRING

Με τη βάση δεδομένων που περιέχει τις πληροφορίες περί πρωτεϊνικών αλληλεπιδράσεων έτοιμη, επόμενο βήμα είναι η αναζήτηση των υποψήφιων γονιδίων για ιεράρχηση στον πίνακα που έχει κατασκευαστεί. Για να γίνει η αναζήτηση αυτή, πρέπει τα δεδομένα των υποψήφιων γονιδίων και τα δεδομένα της βάσης STRING να βρίσκονται στην ίδια μορφή. Όμως, όπως έχει ήδη αναφερθεί, οι πρωτεΐνες δεν είναι αποθηκευμένες με τα ονόματά τους, αλλά με τη χρήση STRING identifiers. Αντιθέτως, τα υποψήφια γονίδια είναι με τη μορφή των ονομάτων τους. Έτσι, υπήρχαν δύο επιλογές.

- Η μετατροπή των δεδομένων του πίνακα STRING, ο οποίος περιέχει το υποσύνολο των πρωτεϊνικών αλληλεπιδράσεων που μας ενδιαφέρουν, από identifiers σε ονόματα.
- Η μετατροπή των δεδομένων των υποψηφίων γονιδίων από ονόματα σε identifiers.

Η πρώτη επιλογή έχει το πλεονέκτημα ότι, εφόσον πραγματοποιηθεί μία φορά δεν είναι απαραίτητο να επαναληφθεί, καθώς η βάση θα είναι ήδη στη μορφή που απαιτείται για την αναζήτηση. Μολαταύτα, απορρίφθηκε για δύο λόγους. Πρώτον, απαιτούσε τη διαχείριση πολύ μεγάλου όγκου δεδομένων, γεγονός το οποίο ήταν δύσκολο και εξαιρετικά χρονοβόρο να πραγματοποιηθεί με τους περιορισμούς των διαθέσιμων συσκευών. Δεύτερον, κάθε identifier αντιστοιχεί σε πολλά ονόματα, γεγονός που σημαίνει ότι ο αριθμός των στηλών του πίνακα θα έπρεπε να αυξηθεί κατά περίπου μία τάξη μεγέθους, αυξάνοντας αναλόγων το μέγεθος του RNEA.

Για την επίτευξη λοιπόν της δεύτερης επιλογής, έγινε χρήση του αρχείου protein.aliases.txt(http://stringdb.org/newstring_download/protein.aliases.v9.1.txt.gz) που παρέχεται από την ιστοσελίδα της STRING.

Περιέχει 4 στήλες:

- species_ncbi_taxon_id: η ταξονομική ταυτότητα που έχει δοθεί από το NCBI
- protein_id: identifier της πρωτεΐνης
- alias: το όνομα της πρωτεΐνης
- source: η πηγή των δεδομένων

Με παρόμοιο τρόπο με την ενότητα 3.1, έγινε εξόρυξη μόνο των δεδομένων που αφορούσαν τον άνθρωπο και τον ποντικό, στα αρχεία humanalias.txt και mousealias.txt αντιστοίχως.

```
String_ref = paste("ReferenceFiles/",species,"_string.tsv",
sep="");
 String_alias =
paste("ReferenceFiles/",species,"_stringalias.txt", sep="");

stringdb <-read.table(String_ref, header= TRUE)
alias<-read.table(String_alias, header= FALSE, sep="\t",quote =
"")
```

Αποθήκευση του λεξιλογίου σε μεταβλητή μέσα στην R.

```
candidategenes<-paste0("^",DE_genes,"$")
```

Η μεταβλητή DE_genes περιέχει τα υποψήφια γονίδια. Η εντολή paste επιτρέπει την σύνδεση διανυσμάτων αφού πρώτα τα μετατρέπει σε χαρακτήρες. Χρησιμοποιώντας την με αυτόν τον τρόπο, προστίθενται κανονικές εκφράσεις στην αρχή και στο τέλος του ονόματος κάθε γονιδίου. Οι κανονικές εκφράσεις “^” και “$” σηματοδοτούν την αρχή και το τέλος μιας συμβολοσειράς αντίστοιχα. Έτσι, στο επόμενο βήμα, χρησιμοποιώντας τη μεταβλητή candidategenes θα αναζητηθούν αποκλειστικά τα υποψήφια γονίδια και όχι άλλα που μπορεί να τα εμπεριέχουν στην αρχή ή στο τέλος τους.

```
translde<-alias[grep(paste(candidategenes,collapse='|'),
alias[[2]], ignore.case=TRUE),]
```

Ταίριασμα κάθε υποψήφιου γονιδίου με τον Ensemble identifier του.

```
DEgenesens<-unique(as.character(translde[,1]))
translde1<-stringdb[grep(paste(DEgenesens,collapse='|'),
stringdb[[1]], ignore.case=TRUE),]
DEgenesstring<-translde1[,1:2]
```

Εύρεση των γονιδίων με τα οποία αλληλεπιδρούν τα υποψήφια γονίδια από τη βάση STRING.

```
de1<-merge(DEgenesstring, translde, by.x=c("protein1"),
by.y=c("V1"))
 de1<-merge(de1, translde, by.x=c("protein2"), by.y=c("V1"))
 de1<-de1[,3:4]
```

Μετάφραση των αλληλεπιδρώντων γονιδίων από Ensembl identifiers πίσω σε ονόματα.

```
de2<- data.frame(tolower(as.matrix(de1)))
de1<-unique(de2)
colnames(de1)<-colnames(Network)
de1$Source<-as.character(de1$Source)
de1$Target<-as.character(de1$Target)
de1[de1==""]<-NA
de1<-de1[complete.cases(de1),]
de1$Source <- gsub("^(\\w)(\\w+)", "\\U\\1\\L\\2",
                   de1$Source, perl = TRUE)
de1$Target <- gsub("^(\\w)(\\w+)", "\\U\\1\\L\\2",
                 de1$Target, perl = TRUE)
```

Απομάκρυνση των επαναλήψεων στα αποτελέσματα και μετατροπή τους σε μορφή κατάλληλη για ανάλυση δικτύου.

Το τελικό output θα έχει αυτήν τη μορφή:

Source	Target
Serpine1	Plaur
Serpinb2	Plaur
Tnfaip3	Tnf
Malt1	Tnf
Tnf	Tnf

Πίνακας 1: Δεδομένα εξόδου από την εξαγωγή υποδικτύου από τη STRING.

3.3) Ιεράρχηση γονιδίων με εφαρμογή αλγόριθμου PageRank

Από το προηγούμενο βήμα έχει εξαχθεί ένα υποδίκτυο από τη βάση δεδομένων STRING το οποίο περιέχει τα στατιστικά σημαντικά υποψήφια γονίδια και τα γονίδια με τα οποία αυτά αλληλεπιδρούν. Είναι γενικότερα ένα PPIN, το οποίο μπορεί να αντιπροσωπευθεί ως ένα μη κατευθυνόμενο γράφημα, όπου οι πρωτεΐνες είναι οι κόμβοι και οι αλληλεπιδράσεις είναι οι ακμές.

```
require(sqldf)
```

Το sqldf είναι ένα πακέτο της R το οποίο επιτρέπει την εφαρμογή δηλώσεων SELECT σε πίνακες αποθηκευμένους στην R. Οι δηλώσεις SELECT επιτρέπουν την ανάκτηση μηδέν ή παραπάνω σειρών από ένα πίνακα, ανάλογα με επιπλέον κριτήρια που έχουν τεθεί, με συγκεκριμένη ομαδοποίηση και σειρά.

```
netou=sqldf("SELECT Source, COUNT(*) outs FROM de1 GROUP BY 1")
```

Στις δηλώσεις SELECT, οι πίνακες που περιέχουν τα δεδομένα που πρέπει να κρατηθούν στο τελικό αποτέλεσμα υποδεικνύονται με τη λέξη FROM. Στη συγκεκριμένη περίπτωση επιλέγεται το υποδίκτυο των υποψηφίων γονιδίων από τη βάση δεδομένων της STRING

```
netpr=sqldf("SELECT Source vertex, 1.0 pagerank FROM de1 UNION
SELECT Target, 1.0 FROM de1")
     for (i in 1:100)
     {
       netx1=sqldf("SELECT vertex, pagerank/outs factor FROM netou
a INNER JOIN netpr b ON (a.Source=b.vertex)")
       netpr=sqldf("SELECT a.vertex,
0.15+SUM(0.85*COALESCE(factor,0)) AS pagerank
      FROM netpr a LEFT OUTER JOIN de1 b ON (a.vertex= b.Target)
LEFT OUTER JOIN netx1 c
      ON (b.Source= c.vertex) GROUP BY 1")
     }
  netprsort<-netpr[with(netpr, order(-pagerank)),]
```

Εφαρμόζεται ο αλγόριθμος PageRank στο δίκτυο και ιεραρχούνται τα γονίδια βάση της τιμής του

```
write.csv(netprsort,file=paste(output,"String sorted by
pagerank.csv"),quote=F,row.names=F);
write.csv(netprsort,file=paste(output,"String sorted by
pagerank.csv"),quote=F,row.names=F);
```

Δημιουργούνται τα αρχεία που περιέχουν το δίκτυο που έχει εξαχθεί από τη STRING και την ιεραρχημένη λίστα με τις PageRank τιμές. Με παρόμοιο τρόπο μπορεί να ιεραρχηθεί το ρυθμιστικό δίκτυο που προκύπτει από το RNEA.

3.4) Γραφική κατασκευή δικτύων

Αν και τα δίκτυα που παράγονται από την ανάλυση μπορούν να απεικονιστούν με διάφορα προγράμματα, όπως το Cytoscape, για διευκόλυνση των χρηστών παρέχεται η δημιουργία ενός μη γραφικά περίπλοκου δικτύου, επιτρέποντας το σχηματισμό μιας γενικότερης εικόνας από τα δεδομένα. Ο αλγόριθμος της γραφικής κατασκευής δικτύων εφαρμόζεται τόσο στο υποδίκτυο της STRING όσο και στο δίκτυο(ρυθμιστικό, λειτουργικό ή ολικό) που μπορεί να παράξει το RNEA. Η συγκεκριμένη εφαρμογή έχει το πλεονέκτημα της απεικόνισης των σκορ PageRank απευθείας πάνω στο γράφημα. Ενδεικτικά, παρατίθεται η γραφική κατασκευή υποδικτύου αλληλεπιδράσεων της STRING.

```
require(igraph)
```

Το πακέτο igraph περιέχει ρουτίνες για απλά γραφήματα και ανάλυση δικτύων.

```
par(family="serif", cex=1, ps=15, bg="white", col.lab="black",
col.axis="black")
```

Ορίζονται ορισμένες παράμετροι για την ευκρίνεια και ευαναγνωσιμότητα της οπτικής απεικόνισης του δικτύου.

```
g=graph.edgelist(as.matrix(net))
names=data.frame(vertex=V(g)$name)
V(g)$name=sqldf("SELECT a.vertex||' (PR='||ROUND(b.pagerank,2)||')' as name
from names a inner join netpr b ON (a.vertex=b.vertex)")$name
plot(g, edge.arrow.size=1, vertex.color="gray90", edge.color="black")
```

Κατασκευάζεται γραφικά το δίκτυο.

```
Networkname=paste(species,"_string_network",sep="")
dev.copy(png,Networkname, width=1500, height=1500, res=300)
dev.off()
```

Η οπτική απεικόνιση του δικτύου εξάγεται ως εικόνα, άμεσα προσβάσιμη από το χρήστη χωρίς ανάγκη περαιτέρω επεξεργασίας.

3.5)Εκτέλεση του RNEA στο terminal

Παρατίθενται οι απαραίτητες εντολές για την εκτέλεση του RNEA μέσω terminal

>> cd RNEA/

>> sudo R

>> source("RNEA.R")

>>RNEA()

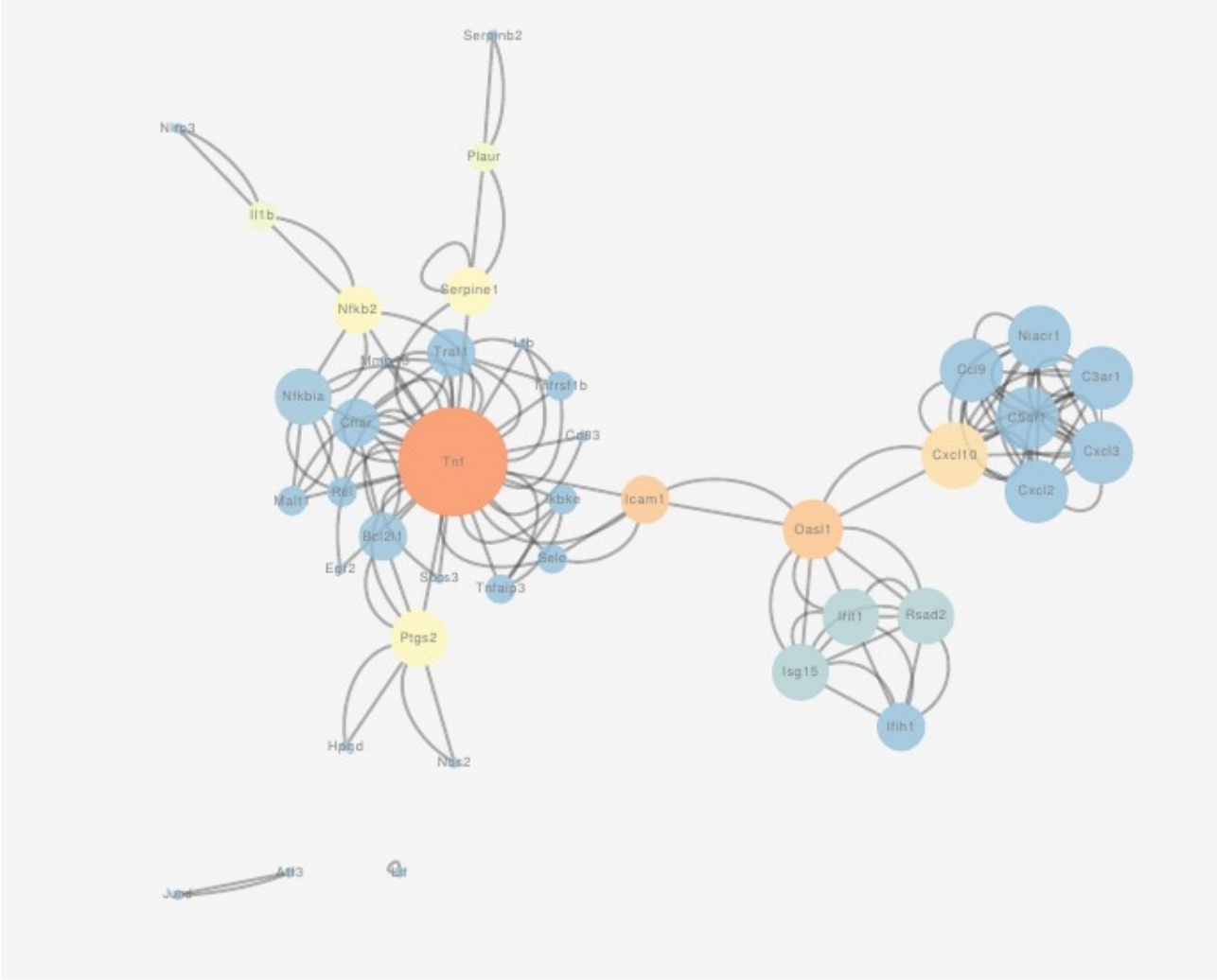
Στη συνάρτηση, απαιτείται εισαγωγή 8 arguments από το χρήστη.

1) filename
2) identifier(default="GeneName" or "RefSeq"
3) species ("Human" or "Mouse")
4) FC_threshold (default=1)
5) PV_threshold (default=0.05)
6) output (default="Output")
7) network (default="global" or "regulatory" or "functional")
8) type_of_output(default="html" or "csv").

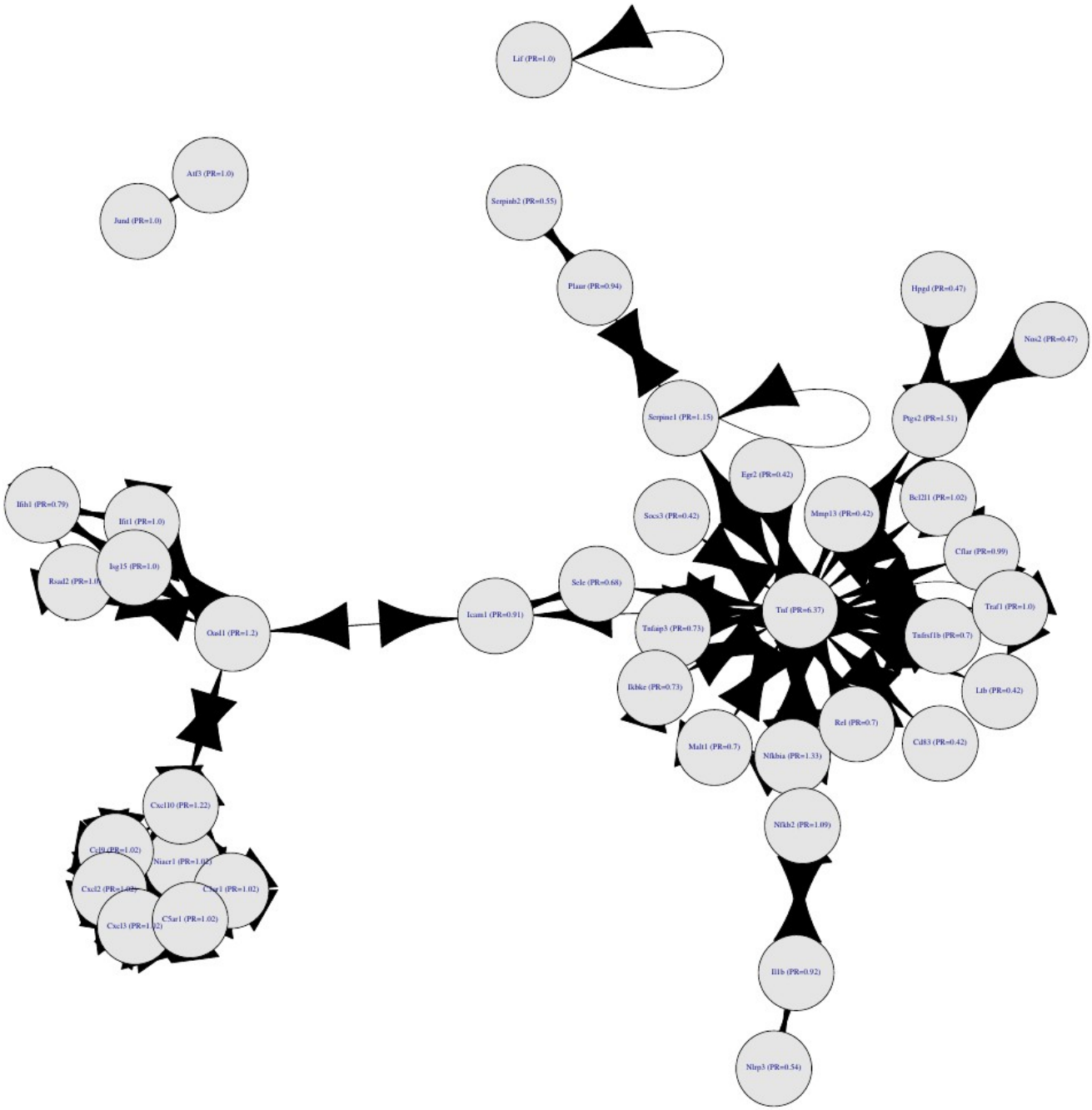
3.6) Δοκιμασία σε δεδομένα διαφορικής έκφρασης ποντικού

Για δοκιμή της μεθόδου, έγινε επιλογή δεδομένων διαφορικής έκφρασης από Mus musculus. Πιο συγκεκριμένα, έγινε επιλογή της διαφορικής απόκρισης σε φλεγμονή RAW264.7 μακροφάγων υπό LPS διέγερση[118]. Χρησιμοποιώντας τα τυπικά κριτήρια του RNEA($\log_2$FC>=1, p-value <=0,05), προέκυψαν 121 γονίδια με διαφορική έκφραση.

Το δίκτυο πρωτεϊνικών αλληλεπιδράσεων που προκύπτει από εξαγωγή από τη STRING φαίνεται στην εικόνα 1 μετά από επεξεργασία από Cytoscape. Στην εικόνα 2 φαίνεται η γραφική απεικόνιση του δικτύου που προκύπτει αυτόματα από την R. Αν και αισθητικά λιγότερο εντυπωσιακή, έχει το πλεονέκτημα της άμεσης αποτύπωσης του PageRank σκορ απευθείας πάνω στους κόμβους. Τα κορυφαία ιεραρχημένα γονίδια καταγράφονται στον πίνακα 1.



Εικόνα 1: Δίκτυο πρωτεϊνικών αλληλεπιδράσεων για ενδεικτική περίπτωση *Mus musculus* μακροφάγων διεγερμένων με LPS μέσω Cytoscape. Το μέγεθος των κόμβων είναι ανάλογο με το βαθμό σύνδεσής τους ενώ το χρώμα τους τείνει προς μπλε για χαμηλές τιμές ενδιάμεσης κεντρικότητας ακμών και κόκκινο για υψηλές τιμές ενδιάμεσης κεντρικότητας.



Εικόνα 2: Εικόνα 1: Δίκτυο πρωτεϊνικών αλληλεπιδράσεων για ενδεικτική περίπτωση *Mus musculus* μακροφάγων διεγερμένων με LPS μέσω igraph. Στους κόμβους αναγράφεται το όνομα της πρωτεΐνης και το PageRank σκορ.

Γονίδιο	PageRank
Tnf	6.36811273653727
Ptgs2	1.5118492404999
Nfkbia	1.32790975005802
Cxcl10	1.22379491171913
Oasl1	1.19752064127252
Serpine1	1.14604886710387
Nfkb2	1.09363477910022
C3ar1	1.02378400341778
C5ar1	1.02378400341778
Ccl9	1.02378400341778
Cxcl2	1.02378400341778
Cxcl3	1.02378400341778
Niacr1	1.02378400341778
Bcl2l1	1.02370203858184
Rsad2	1.00431951557814
Ifit1	1.00431951557814
Isg15	1.00431951557814
Traf1	1.00184721322364

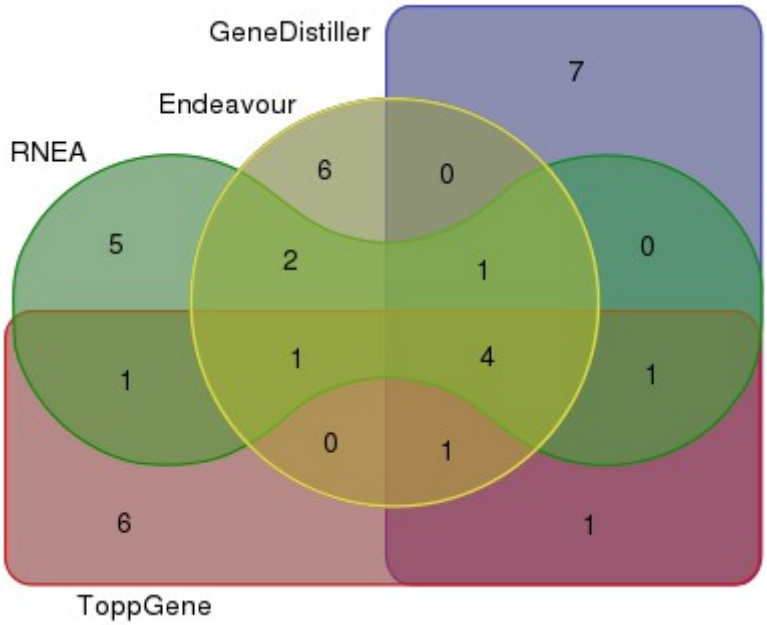
Πίνακας 2: Τα κορυφαία γονίδια μετά από την ιεράρχηση ανάλογα με το υψηλότερο PageRank σκορ.

Την υψηλότερη βαθμολογία κατέχει το Tnf το οποίο εκφράζει πρωτεΐνη(Tumor Necrosis Factor-α) με εκτενέστατη βιβλιογραφία σχετική με το ρόλο της στις φλεγμονές[119][120][121]. Από όλα τα υπόλοιπα που βρίσκονται στα δέκα κορυφαία, το Ptgs2[122][123], το Nfkbia[124], το Cxcl10[125], το Oasl1[126], το Serpine1[127], το Nfkb2[128], το C3ar1[129], το C5ar1[130], το Cccl9[131], όλα είχαν συσχέτιση με τη φλεγμονώδης απόκριση στον ποντικό είτε άμεσα με πειράματα που έχουν διεξαχθεί σε ποντίκια είτε έμμεσα με τα ορθόλογα τους στον άνθρωπο.

Για δοκιμασία της μεθόδου μας σε σχέση με παρόμοια υπολογιστική δουλειά, έγινε σύγκριση με τρεις από τις πιο δημοφιλείς πλατφόρμες ιεράρχησης

Position	RNEA	Endeavour	GeneDistiller	ToppGene
1	Tnf	Tnfrsf1b	Il1b	Pdgfb
2	Ptgs2	Tnf	Icam1	Serpine1
3	Nfkbia	Ccl9	Ptgs2	Tnf
4	Cxcl10	Cxcl2	Nos2	Bcl2l1
5	Oasl1	Cxcl10	Cxcl10	Nfkbia
6	Serpine1	Nfkbia	Nfkbia	Cxcl10
7	Nfkb2	Pik3r5	Nfkb2	Itga5
8	C3ar1	Mefv	Cxcl2	Cflar
9	C5ar1	Tnfaip3	Tnfaip3	Ptgs2
10	Ccl9	Ifih1	Cd83	Icam1
11	Cxcl2	Nfkb2	Serpine1	Cdkn1a
12	Cxcl3	Ikbke	Rel	Nfkb2
13	Niacr1	Cxcl3	Mmp13	Malt1
14	Bcl2l1	Traf1	Socs3	Sdc4
15	Rsad2	Ptgs2	Tnfaip2	Tnfaip3

Πίνακας 3: Συγκριτικός πίνακας των αποτελεσμάτων του αλγορίθμου μας με αποτελέσματα ιεράρχησης των υποψηφίων γονιδίων από Endeavour, GeneDistiller και ToppGene. Στις τελευταίες δύο πλατφόρμες, έγινε μετατροπή των γονιδίων σε ανθρώπινα μέσω ορθολόγων.



Εικόνα 3: Venn διάγραμμα των κορυφαίων αποτελεσμάτων των 4 διαφορετικών μεθόδων ιεράρχησης.

Τα αποτελέσματα που προκύπτουν από την RNEA ιεράρχηση είναι ξεκάθαρα συγκρίσιμα με αυτά των υπόλοιπων πλατφορμών, αφού τα κορυφαία γονίδια είναι σε μεγάλο βαθμό ίδια. Η ύπαρξη όμως μοναδικών γονιδίων υποδεικνύουν καινούρια ερευνητικά μονοπάτια που θα έμεναν ανεξερεύνητα με την εφαρμογή των ήδη υπαρχουσών τεχνικών.

Κεφάλαιο IV

Συμπεράσματα

Στη δουλειά αυτή, πρώτα αξιοποιήθηκε ο πλούτος πρωτεϊνικών αλληλεπιδράσεων που περιέχονται στη STRING και συνδυάστηκε με λειτουργικές, ρυθμιστικές και οντολογικές πληροφορίες από άλλες διαθέσιμες βάσεις δεδομένων. Στη συνέχεια, χρησιμοποιήθηκε μέθοδος ιεράρχησης βασισμένος στις τοπολογικές ιδιότητες των βιολογικών δικτύων. Οι περισσότερες προϋπάρχουσες μέθοδοι ιεράρχησης αφήνουν ανεκμετάλλευτο τον ισχυρό συνδυασμό ανάλυσης δικτύων με μια συνδυαστική, πολυεπίπεδη πηγή γονιδιακών δεδομένων. Αν και δεν έχει γίνει χρησιμοποίηση της μεθόδου για την ανακάλυψη νέων άγνωστων γονιδίων στα πλαίσια της διατριβής, με την εφαρμογή της μεθόδου σε πραγματικά δεδομένα διαφαίνεται η ισχύς της στον εντοπισμό σημαντικών γονιδίων.

Καλό θα ήταν να αναφερθούν ορισμένοι περιορισμοί της παρούσας τεχνικής. Αρχικά, δεν έχει γίνει πλήρης ενσωμάτωση του δικτύου που προκύπτει από τη STRING με το δίκτυο που δημιουργεί το RNEA. Με την ολοκλήρωση της ενσωμάτωσης αυτής, θα μπορέσει να εκτελεστεί μία, ολοκληρωμένη ιεράρχηση, η οποία θα είναι πιο πλούσια πληροφοριακά και με μεγαλύτερες δυνατότητες επιτυχίας. Επιπλέον, υπάρχει περιθώριο για την προσθήκη επιπλέον πληροφοριών στο βιολογικό δίκτυο. Βάσεις δεδομένων που περιέχουν άφθονες πληροφορίες για την ιστοειδική έκφραση και τη συσχέτιση ασθενειών με γονίδια, φαινότυπους αλλά και ποικίλες σημασιολογικές και οντολογικές πληροφορίες, έχουν μείνει ανεκμετάλλευτες. Επομένως, υπάρχει περιθώριο για διεύρυνση των δικτύων και σε κάλυψη γονιδίων αλλά και σε αναλυτική ισχύ. Επιπλέον, υπάρχουν περιθώρια δημιουργίας διαδικτυακής πλατφόρμας, η οποία θα είναι πιο προσβάσιμη στο μέσο χρήστη από την τωρινή μορφή. Μολαταύτα, ακόμα και αν ληφθούν υπόψη αυτοί οι περιορισμοί, η ολιστική προσέγγιση που παρέχεται από την υπάρχουσα δουλειά μπορεί να θεωρείται ένα ισχυρό εργαλείο για την υπολογιστική ανάλυση δεδομένων από πειράματα υψηλής απόδοσης.

Τα τελευταία χρόνια τα βιολογικά δίκτυα έχουν αποδειχθεί ότι είναι ιδιαίτερα χρήσιμα σε πολλαπλούς τομείς της μοριακής βιολογίας. Αποτελεί ελπίδα ότι η μελέτη αυτή θα ενθαρρύνει μελλοντικές μελέτες για την ενσωμάτωση ποικίλων βιολογικών δεδομένων σε δίκτυα, για την εξήγηση των φαινότυπων των ασθενειών και την εύρεση νέων θεραπειών.

Βιβλιογραφία

1. Howe, Doug, et al. "Big data: The future of biocuration." *Nature* 455.7209 (2008): 47-50.
2. Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature reviews genetics* 10.1 (2009): 57-63.
3. Brenner, Sydney, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." *Nature biotechnology* 18.6 (2000): 630-634.
4. Velculescu, Victor E., et al. "Serial analysis of gene expression." *Science* 270.5235 (1995): 484.
5. Liang, Peng, and Arthur B. Pardee. "Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction." *Science* 257.5072 (1992): 967-971.
6. Fodor, S. P., et al. "Light-directed, spatially addressable parallel chemical synthesis." *Science* 251.4995 (1991): 767-773.
7. Trevino, Victor, Francesco Falciani, and Hugo A. Barrera-Saldaña. "DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research." *Molecular Medicine* 13.9-10 (2007): 527.
8. Schena, Mark, et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270.5235 (1995): 467.
9. Takahashi, Kohta, Shigeaki Saitoh, and Mitsuhiro Yanagida. "Application of the chromatin immunoprecipitation method to identify in vivo protein-DNA associations in fission yeast." *Science Signaling* 2000.56 (2000): pl1-pl1.
10. Sapolsky, Ronald J., et al. "High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays." *Genetic analysis: biomolecular engineering* 14.5 (1999): 187-192.
11. Redon, Richard, et al. "Global variation in copy number in the human genome." *nature* 444.7118 (2006): 444-454.
12. Moritz, Craig, and Carla Cicero. "DNA barcoding: promise and pitfalls." *PLoS Biol* 2.10 (2004): e354.
13. Bertone, Paul, et al. "Global identification of human transcribed sequences with genome tiling arrays." *Science* 306.5705 (2004): 2242-2246.
14. Southern, Edwin M. "DNA microarrays." *DNA Arrays: Methods and Protocols* (2001): 1-15.
15. Bullinger, Lars, et al. "Gene-expression profiling identifies distinct subclasses of core binding factor acute myeloid leukemia." *Blood* 110.4 (2007): 1291-1300.
16. Tibshirani, Robert, et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proceedings of the National Academy of Sciences* 99.10 (2002): 6567-6572.
17. Marton, Matthew J., et al. "Drug target validation and identification of secondary drug target effects using DNA microarrays." *Nature medicine* 4.11 (1998): 1293-1301.
18. Forster, T., D. Roy, and P. Ghazal. "Experiments using microarray technology: limitations and standard operating procedures." *Journal of Endocrinology* 178.2 (2003): 195-204.
19. Boelens, Mirjam C., et al. "Microarray amplification bias: loss of 30% differentially expressed genes due to long probe–poly (A)-tail distances." *BMC genomics* 8.1 (2007): 1.
20. Okoniewski, Michał J., and Crispin J. Miller. "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations." *BMC bioinformatics* 7.1 (2006): 276.
21. Berretta, Julia, and Antonin Morillon. "Pervasive transcription constitutes a new level of eukaryotic genome regulation." *EMBO reports* 10.9 (2009): 973-982.
22. Vera, J. Cristobal, et al. "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing." *Molecular ecology* 17.7 (2008): 1636-1647.
23. Cloonan, Nicole, et al. "Stem cell transcriptome profiling via massive-scale mRNA sequencing." *Nature methods* 5.7 (2008): 613-619.
24. Pan, Qun, et al. "Deep surveying of alternative splicing complexity in the human

transcriptome by high-throughput sequencing." *Nature genetics* 40.12 (2008): 1413-1415.
25. Degner, Jacob F., et al. "Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data." *Bioinformatics* 25.24 (2009): 3207-3212.
26. Edgren, Henrik, et al. "Identification of fusion genes in breast cancer by paired-end RNA-sequencing." *Genome biology* 12.1 (2011): 1.
27. Marioni, John C., et al. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome research* 18.9 (2008): 1509-1517.
28. Cloonan, Nicole, et al. "Stem cell transcriptome profiling via massive-scale mRNA sequencing." *Nature methods* 5.7 (2008): 613-619.
29. Leung, Yuk Fai, and Duccio Cavalieri. "Fundamentals of cDNA microarray data analysis." *TRENDS in Genetics* 19.11 (2003): 649-659.
30. Nadon, Robert, and Jennifer Shoemaker. "Statistical issues with microarrays: processing and analysis." *TRENDS in Genetics* 18.5 (2002): 265-271.
31. Soneson, Charlotte, and Mauro Delorenzi. "A comparison of methods for differential expression analysis of RNA-seq data." *BMC bioinformatics* 14.1 (2013): 1.
32. Lee, Mei-Ling Ting, et al. "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations." *Proceedings of the National Academy of Sciences* 97.18 (2000): 9834-9839.
33. Schurch, Nicholas J., et al. "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?." *RNA* 22.6 (2016): 839-851.
34. Konishi, Tomokazu. "Parametric analysis of RNA-seq expression data." *Genes to Cells* (2016).
35. Lönnstedt, Ingrid, and Terry Speed. "Replicated microarray data." *Statistica sinica* (2002): 31-46.
36. Vijay, Nagarjun, et al. "Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments." *Molecular ecology* 22.3 (2013): 620-634.
37. Park, Taesung, et al. "Statistical tests for identifying differentially expressed genes in time-course microarray experiments." *Bioinformatics* 19.6 (2003): 694-703.
38. Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26.1 (2010): 139-140.
39. Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome biology* 11.10 (2010): 1.
40. Hardcastle, Thomas J., and Krystyna A. Kelly. "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." *BMC bioinformatics* 11.1 (2010): 422.
41. Conover, William Jay, and W. J. Conover. "Practical nonparametric statistics." (1980).
42. Tusher, Virginia Goss, Robert Tibshirani, and Gilbert Chu. "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of the National Academy of Sciences* 98.9 (2001): 5116-5121.
43. Wu, Thomas D. "Analysing gene expression data from DNA microarrays to identify candidate genes." *The Journal of pathology* 195.1 (2001): 53-65.
44. Li, Jun, and Robert Tibshirani. "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data." *Statistical methods in medical research* 22.5 (2013): 519-536.
45. Ravasz, Erzsébet, et al. "Hierarchical organization of modularity in metabolic networks." *science* 297.5586 (2002): 1551-1555.
46. Han, Jing-Dong Jackie. "Understanding biological functions through molecular networks."

Cell research 18.2 (2008): 224-237.
47. Ihmels, Jan, et al. "Revealing modular organization in the yeast transcriptional network." *Nature genetics* 31.4 (2002): 370-377.
48. Milo, Ron, et al. "Network motifs: simple building blocks of complex networks." *Science* 298.5594 (2002): 824-827.
49. Lee, Tong Ihn, et al. "Transcriptional regulatory networks in Saccharomyces cerevisiae." *science* 298.5594 (2002): 799-804.
50. Goldberg, Debra S., and Frederick P. Roth. "Assessing experimentally derived interactions in a small world." *Proceedings of the National Academy of Sciences* 100.8 (2003): 4372-4376.
51. Saito, Rintaro, Harukazu Suzuki, and Yoshihide Hayashizaki. "Interaction generality, a measurement to assess the reliability of a protein–protein interaction." *Nucleic acids research* 30.5 (2002): 1163-1168.
52. Howson, Colin, and Peter Urbach. *Scientific reasoning: the Bayesian approach*. Open Court Publishing, 2006.
53. Jansen, Ronald, et al. "A Bayesian networks approach for predicting protein-protein interactions from genomic data." *Science* 302.5644 (2003): 449-453.
54. Eom, Jae-Hong, and Byoung-Tak Zhang. "Prediction of protein interaction with neural network-based feature association rule mining." *International Conference on Neural Information Processing*. Springer Berlin Heidelberg, 2006.
55. Lewis, Darrin P., Tony Jebara, and William Stafford Noble. "Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure." *Bioinformatics* 22.22 (2006): 2753-2760.
56. Bock, Joel R., and David A. Gough. "Whole-proteome interaction mining." *Bioinformatics* 19.1 (2003): 125-134.
57. Qi, Yanjun, J*udith Klein-Seetharaman, and Ziv Bar-Joseph. "Random forest similarity for protein-protein interaction prediction from multiple sources." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. 2004.
58. Marazita, Mary L., et al. "Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35." *The American Journal of Human Genetics* 75.2 (2004): 161-173.
59. Jorde, Lynn B. "Linkage disequilibrium and the search for complex disease genes." *Genome research* 10.10 (2000): 1435-1444.
60. Page, Grier P., et al. "“Are we there yet?”: Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits." *The American Journal of Human Genetics* 73.4 (2003): 711-719.
61. Perez-Iratxeta, Carolina, Peer Bork, and Miguel A. Andrade. "Association of genes to genetically inherited diseases using data mining." *Nature genetics* 31.3 (2002): 316-319.
62. Chen, Jing, et al. "Improved human disease candidate gene prioritization using mouse phenotype." *BMC bioinformatics* 8.1 (2007): 1.
63. Seelow, Dominik, Jana Marie Schwarz, and Markus Schuelke. "GeneDistiller—distilling candidate genes from linkage intervals." *PLoS One* 3.12 (2008): e3874.
64. Hutz, Janna E., et al. "CANDID: a flexible method for prioritizing candidate genes for complex human traits." *Genetic epidemiology* 32.8 (2008): 779.
65. Wolfe, Cecily J., Isaac S. Kohane, and Atul J. Butte. "Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks." *BMC bioinformatics* 6.1 (2005): 1.
66. Zhu, Mengjin, and Shuhong Zhao. "Candidate gene identification approach: progress and challenges." *Int J Biol Sci* 3.7 (2007): 420-427.
67. Moreau, Yves, and Léon-Charles Tranchevent. "Computational tools for prioritizing

candidate genes: boosting disease gene discovery." *Nature Reviews Genetics* 13.8 (2012): 523-536.
68. Hamosh, Ada, et al. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." *Nucleic acids research* 33.suppl 1 (2005): D514-D517.
69. Dietze, Heiko, et al. "Gopubmed: Exploring pubmed with ontological background knowledge." *Bioinformatics for Systems Biology*. Humana Press, 2009. 385-399.
70. Roberts, Richard J. "PubMed Central: The GenBank of the published literature." *Proceedings of the National Academy of Sciences* 98.2 (2001): 381-382.
71. Becker, Kevin G., et al. "The genetic association database." *Nature genetics* 36.5 (2004): 431-432.
72. Yu, Wei, et al. "Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations." *Bioinformatics* 26.1 (2010): 145-146.
73. Kanehisa, Minoru, et al. "KEGG for representation and analysis of molecular networks involving diseases and drugs." *Nucleic acids research* 38.suppl 1 (2010): D355-D360.
74. Tiffin, Nicki. "Conceptual thinking for in silico prioritization of candidate disease genes." *In Silico Tools for Gene Discovery* (2011): 175-187.
75. Bush, William S., Scott M. Dudek, and Marylyn D. Ritchie. "Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, 2009.
76. Franke, Lude, et al. "TEAM: a tool for the integration of expression, and linkage and association maps." *European journal of human genetics* 12.8 (2004): 633-638.
77. Gill, Nivit, Shailendra Singh, and Trilok C. Aseri. "Computational disease gene prioritization: an appraisal." *Journal of Computational Biology* 21.6 (2014): 456-465.
78. Schlicker, Andreas, Thomas Lengauer, and Mario Albrecht. "Improving disease gene prioritization using the semantic similarity of Gene Ontology terms." *Bioinformatics* 26.18 (2010): i561-i567
79. Krallinger, Martin, Alfonso Valencia, and Lynette Hirschman. "Linking genes to literature: text mining, information extraction, and retrieval applications for biology." *Genome biology* 9.2 (2008):
80. Tranchevent, Léon-Charles, et al. "ENDEAVOUR update: a web resource for gene prioritization in multiple species." *Nucleic acids research* 36.suppl 2 (2008): W377-W384.
81. Chen, Jing, et al. "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization." *Nucleic acids research* 37.suppl 2 (2009): W305-W311.
82. Adie, Euan A., et al. "SUSPECTS: enabling fast and effective prioritization of positional candidates." *Bioinformatics* 22.6 (2006): 773-774.
83. Jourquin, Jérôme, et al. "GLAD4U: deriving and prioritizing gene lists from PubMed literature." *BMC genomics* 13.8 (2012): 1.
84. Yu, Wei, et al. "Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases." *BMC bioinformatics* 9.1 (2008): 1.
85. Fontaine, Jean-Fred, et al. "Genie: literature-based gene prioritization at multi genomic scale." *Nucleic acids research* 39.suppl 2 (2011): W455-W461.
86. Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." *Nature Reviews Genetics* 12.1 (2011): 56-68.
87. Jiang, Rui, Mingxin Gan, and Peng He. "Constructing a gene semantic similarity network for the inference of disease genes." *BMC systems biology* 5.2 (2011): 1.
88. Erten, Sinan, and Mehmet Koyutürk. "Role of centrality in network-based prioritization of disease genes." *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer Berlin Heidelberg, 2010.
89. Nitsch, Daniela, et al. "PINTA: a web server for network-based gene prioritization from

expression data." *Nucleic acids research* 39.suppl 2 (2011): W334-W338.
90. Köhler, Sebastian, et al. "Walking the interactome for prioritization of candidate disease genes." *The American Journal of Human Genetics* 82.4 (2008): 949-958.
91. Rossi, Simona, et al. "TOM: a web-based integrated approach for identification of candidate disease genes." *Nucleic acids research* 34.suppl 2 (2006): W285-W292.
92. Radivojac, Predrag, et al. "An integrated approach to inferring gene–disease associations in humans." *Proteins: Structure, Function, and Bioinformatics* 72.3 (2008): 1030-1037.
93. George, Richard A., et al. "Analysis of protein sequence and interaction data for candidate disease gene prediction." *Nucleic acids research* 34.19 (2006): e130-e130.
94. Oti, Martin, et al. "Predicting disease genes using protein–protein interactions." *Journal of medical genetics* 43.8 (2006): 691-698.
95. Fontaine, Jean-Fred, et al. "Genie: literature-based gene prioritization at multi genomic scale." *Nucleic acids research* 39.suppl 2 (2011): W455-W461.
96. Becker, Richard A., and John M. Chambers. *S: an interactive environment for data analysis and graphics*. CRC Press, 1984.
97. Chambers, John M. *Programming with data: A guide to the S language*. Springer Science & Business Media, 1998.
98. Chambers, John. *Software for data analysis: programming with R*. Springer Science & Business Media, 2008.
99. Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003): 2498-2504.
100. Smoot, Michael E., et al. "Cytoscape 2.8: new features for data integration and network visualization." *Bioinformatics* 27.3 (2011): 431-432.
101. Date, Chris J., and Hugh Darwen. *A Guide To Sql Standard*. Vol. 3. Reading: Addison-Wesley, 1997.
102. Harrison, Guy. *Oracle SQL High-Performance Tuning*. Prentice Hall Professional Technical Reference, 2000.
103. Delaney, Kalen. *Inside Microsoft SQL Server 2000*. Microsoft Press, 2000.
104. Welling, Luke, and Laura Thomson. *PHP and MySQL Web development*. Sams Publishing, 2003.
105. McGoveran, David. *Guide to SYBASE and SQL Server*. Addison-Wesley Longman Publishing Co., Inc., 1999.
106. Hellerstein, Joseph M., Ron Avnur, and Vijayshankar Raman. "Informix under control: Online query processing." *Data Mining and Knowledge Discovery* 4.4 (2000): 281-314.
107. Martinez-Cruz, Carmen, Ignacio J. Blanco, and M. Amparo Vila. "Ontologies versus relational databases: are they so different? A comparison." *Artificial Intelligence Review* 38.4 (2012): 271-290.
108. Han, Heonjong, et al. "TRRUST: a reference database of human transcriptional regulatory interactions." *Scientific reports* 5 (2015).
109. Jiang, C., et al. "TRED: a transcriptional regulatory element database, new entries and other development." *Nucleic acids research* 35.suppl 1 (2007): D137-D140.
110. Essaghir, Ahmed, et al. "Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data." *Nucleic acids research* 38.11 (2010): e120-e120.
111. Sethupathy, Praveen, Benoit Corda, and Artemis G. Hatzigeorgiou. "TarBase: A comprehensive database of experimentally supported animal microRNA targets." *Rna* 12.2 (2006): 192-197.
112. Griffith, Obi L., et al. "ORegAnno: an open-access community-driven resource for regulatory annotation." *Nucleic acids research* 36.suppl 1 (2008): D107-D113..
113. Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." *Nature*

genetics 25.1 (2000): 25-29.
114. Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* 28.1 (2000): 27-30.
115. Szklarczyk, Damian, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." *Nucleic acids research* (2014): gku1003.
116. Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).
117. Chouvardas, Panagiotis, George Kollias, and Christoforos Nikolaou. "Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis." *BMC Bioinformatics* 17.5 (2016): 319.
118. Sabari, Benjamin R., et al. "Intracellular crotonyl-CoA stimulates transcription through p300-catalyzed histone crotonylation." *Molecular cell* 58.2 (2015): 203-215.
119. Bradley, J. R. "TNF-mediated inflammatory disease." *The Journal of pathology* 214.2 (2008): 149-160.
120. Pasparakis, Manolis, et al. "TNF-mediated inflammatory skin disease in mice with epidermis-specific deletion of IKK2." *Nature* 417.6891 (2002): 861-866.
121. Popa, Calin, et al. "The role of TNF-α in chronic inflammatory conditions, intermediary metabolism, and cardiovascular risk." *Journal of lipid research* 48.4 (2007): 751-762.
122. Hla, Timothy, and Karen Neilson. "Human cyclooxygenase-2 cDNA." *Proceedings of the National Academy of Sciences* 89.16 (1992): 7384-7388.
123. Kim, Sangwon F., Daniel A. Huri, and Solomon H. Snyder. "Inducible nitric oxide synthase binds, S-nitrosylates, and activates cyclooxygenase-2." *Science* 310.5756 (2005): 1966-1970.
124. Auphan, Nathalie, et al. "Immunosuppression by glucocorticoids: inhibition of NF-kappaB activity through induction of IkappaB synthesis." *Science* 270.5234 (1995): 286.
125. Angiolillo, Anne L., et al. "Human interferon-inducible protein 10 is a potent inhibitor of angiogenesis in vivo." *The Journal of experimental medicine* 182.1 (1995): 155-162.
126. Lee, Myeong Sup, et al. "OASL1 inhibits translation of the type I interferon-regulating transcription factor IRF7." *Nature immunology* 14.4 (2013): 346-355.
127. Xu, Xia, et al. "Plasminogen activator inhibitor-1 promotes inflammatory process induced by cigarette smoke extraction or lipopolysaccharides in alveolar epithelial cells." *Experimental lung research* 35.9 (2009): 795-805.
128. Hoesel, Bastian, and Johannes A. Schmid. "The complexity of NF-κB signaling in inflammation and cancer." *Molecular cancer* 12.1 (2013):
129. Bao, Lihua, et al. "Signaling through up-regulated C3a receptor is key to the development of experimental lupus nephritis." *The Journal of Immunology* 175.3 (2005): 1947-1955.
130. Skokowa, Julia, et al. "Macrophages induce the inflammatory response in the pulmonary Arthus reaction through Gαi2 activation that controls C5aR and Fc receptor cooperation." *The Journal of Immunology* 174.5 (2005): 3041-3050.
131. Youn, Byung-S., et al. "A novel chemokine, macrophage inflammatory protein-related protein-2, inhibits colony formation of bone marrow myeloid progenitors." *The Journal of Immunology* 155.5 (1995): 2661-2667.