ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μετα - ανάλυση συντελεστών συσχέτισης και εφαρμογές στη βιολογία συστημάτων

Σασιλιόγλου Ιωάννα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Μπάγκος Παντελεήμων

Καθηγητής

Λαμία, 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Μετα - ανάλυση συντελεστών συσχέτισης και εφαρμογές στη βιολογία συστημάτων

Σασιλιόγλου Ιωάννα

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Μπάγκος Παντελεήμων

Καθηγητής

Λαμία, 2022

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 13/06/2022

Η Δηλούσα
Σασιλιόγλου Ιωάννα

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Μετα - ανάλυση συντελεστών συσχέτισης και εφαρμογές στη βιολογία συστημάτων

Σασιλιόγλου Ιωάννα

Τριμελής Επιτροπή:

Μπάγκος Παντελεήμων, Καθηγητής (επιβλέπων)

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Μπράλιου Γεωργία, Επίκουρη Καθηγήτρια

Ευχαριστίες

Η παρούσα πτυχιακή εργασία εκπονήθηκε από τον Οκτώβριο του 2021 έως τον Μάιο του 2022 στο τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου Καθηγητή κο Παντελεήμονα Μπάγκο για την εμπιστοσύνη που μου έδειξε στην εκπόνηση της παρούσας Πτυχιακής, για την συνεχή καθοδήγηση, την βοήθεια καθώς και τον χρόνο που ήταν πάντα διατεθειμένος να μου αφιερώσει, παρά το υπερβολικά φορτωμένο πρόγραμμα του.

Ακόμη, θα ήθελα να ευχαριστήσω την υποψήφια Διδάκτορα του Πληροφορικής με εφαρμογές στη Βιοϊατρική, κα Πωλίνα Γκούμπλια, για την καθοδήγηση και τον πολύτιμο χρόνο που αφιέρωσε σε κάθε δυσκολία που αντιμετώπιζα.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την αμέριστη στήριξη και αγάπη τους.

Κλείνοντας, θα ήθελα να αναφέρω ότι η πτυχιακή εργασία αυτή αφιερώνεται στην ξαδέρφη μου Πηνελόπη η οποία αποτέλεσε κινητήριο δύναμη της ενασχόλησης μου με την έρευνα.

Σασιλιόγλου Ιωάννα

Μάιος 2022

Περιεχόμενα

Λίστα Εικόνων

Λίστα Πινάκων

Λίστα Κωδίκων

Περίληψη

Η Βιοπληροφορική είναι ένα διεπιστημονικό πεδίο το οποίο αναλύει και ερμηνεύει τον μεγάλο όγκο δεδομένων που παρέχει η επιστήμη της βιολογίας συνδυάζοντας την υπολογιστική επιστήμη, τα μαθηματικά, τη στατιστική και τη μηχανική. Τις τελευταίες δεκαετίες οι αναλύσεις των γενετικών δικτύων που προκύπτουν από τις συσχετίσεις των γονιδίων και βασίζονται σε δεδομένα γονιδιακής έκφρασης (micro-array, RNA-seq) προσελκύουν όλο και περισσότερο το ενδιαφέρον της επιστημονικής κοινότητας σε διάφορους τομείς διαδραματίζοντας έναν αναντικατάστατα σημαντικό ρόλο στην κατανόηση των βιολογικών και ρυθμιστικών μηχανισμών των γονιδίων. Ωστόσο, παρατηρείται ότι οι συσχετίσεις των γονιδίων που προκύπτουν από υπολογιστικές μεθόδους επηρεάζονται από τις επιδράσεις που μπορεί να έχουν άλλα γονίδια επάνω σε αυτές. Παράλληλα, η ανάλυση δεδομένων γονιδιακής έκφρασης έθεσε όλο και μεγαλύτερες προκλήσεις στους ερευνητές για την διαχείριση και την στατιστική ερμηνεία των δεδομένων που προέκυψαν από μελέτες με αριθμό δειγμάτων *n* μικρότερο από τον αριθμό των γονιδίων *p (p>>n)*

Ο στόχος αυτής της πτυχιακής εργασίας είναι η δημιουργία ενός πλαισίου ανάλυσης δεδομένων από μελέτες με μικρό αριθμό δειγμάτων και η δημιουργία ενός δικτύου που θα βασίζεται κυρίως σε τεχνικές υπολογισμού της μερικής συσχέτισης των μεταβλητών, εφαρμόζοντας έναν συνδυασμό διαφορετικών μαθηματικών μεθόδων και υπολογιστικών πακέτων, με κυρίαρχο στόχο την κατανόηση της βιολογικής σημασίας των συστημάτων και του εντοπισμού των κρυφών συσχετίσεων μεταξύ των μεταβλητών. Παράλληλα, στοχεύει στην εύρεση των πραγματικών συσχετίσεων και των αιτιατικών σχέσεων των μεταβλητών εφαρμόζοντας μερική συσχέτιση σε δεδομένα απλής συσχέτισης και στην μετα-ανάλυση δεδομένων με τη χρήση της μερικής συσχέτισης.

Η υλοποίηση των πειραμάτων πραγματοποιείται μέσω της γλώσσας προγραμματισμού R και του R Studio (μία πλατφόρμα ανοιχτού κώδικα για τον προγραμματισμό σε γλώσσα R). Στην παρούσα εργασία χρησιμοποιήθηκε συνδυασμός ευρέως χρησιμοποιούμενων διαφορετικών υπολογιστικών πακέτων και συναρτήσεων της R για τον υπολογισμό του πίνακα συσχέτισης και μερικής συσχέτισης, καθώς και τον υπολογισμό του πίνακα συνδιακύμανσης και του αντιστρόφου του (covariance/ inverse covariance matrix). Ο υπολογισμός της μερικής συσχέτισης υλοποιήθηκε με την εφαρμογή μεθόδων υπολογισμού του Moore-Penrose ψευδο-αντίστροφου πίνακα και άλλων μεθόδων εκτίμησης οι οποίοι αποτελούν το κεντρικό βήμα και μία αποτελεσματική λύση για την διαχείριση μεγάλης κλίμακας συνόλων δεδομένων.

Η διάρθρωση της πτυχιακής ξεκινάει με τη θεωρία του συντελεστή συσχέτισης και της μερικής συσχέτισης και με την περιγραφή των δυνατοτήτων τους σε μελέτες μικρού δείγματος και στη βιολογία συστημάτων. Στην συνέχεια αναλύονται οι μέθοδοι υπολογισμού μερικής συσχέτισης σε δεδομένα με αριθμό δειγμάτων μικρότερο από τον αριθμό των μεταβλητών, καθώς και οι μέθοδοι υπολογισμού στατιστικής σημαντικότητας της μερικής συσχέτισης. Επιπλέον, γίνεται αναλυτική περιγραφή της μετα- ανάλυσης, των μοντέλων της καθώς και του υπολογισμού της με τη χρήση της μερικής συσχέτισης. Ακολουθεί η περιγραφή των δεδομένων και η ανάλυση των αποτελεσμάτων που προέκυψαν κατά την εφαρμογή του αλγορίθμου και τέλος, παρατίθενται τα συμπεράσματα και μελλοντικές επεκτάσεις.

Ο αλγόριθμος που έχει σχεδιαστεί καθώς και επιπλέον αρχεία λογισμικού που χρησιμοποιήθηκαν είναι διαθέσιμα στον παρακάτω σύνδεσμο:

https://github.com/sasioanna/DiseaseNeT

1 Εισαγωγή

1.1 Συντελεστής Συσχέτισης

Η συσχέτιση (ή αλλιώς εξάρτηση) στη στατιστική χρησιμοποιείται για την μέτρηση οποιαδήποτε σχέσης, αιτιολογικής ή όχι, μεταξύ δύο συνόλων δεδομένων ή δύο τυχαίων μεταβλητών. Ο βαθμός της γραμμικής συσχέτισης ενός ζεύγους τυχαίων μεταβλητών μετριέται με τον συντελεστή συσχέτισης. Υπάρχουν αρκετοί συντελεστές συσχέτισης, που συχνά δηλώνονται με ρ ή r, που μετρούν το βαθμό συσχέτισης. Ο πιο κοινός συντελεστής συσχέτισης είναι ο συντελεστής συσχέτισης Pearson, ο οποίος χρησιμοποιείται για τη μέτρηση της γραμμικής σχέσης μεταξύ δύο μεταβλητών. Ωστόσο, σε μια μη γραμμική σχέση, ο συντελεστής συσχέτισης Pearson μπορεί να μην αποτελεί κατάλληλο μέτρο συσχέτισης. Έτσι, έχουν αναπτυχθεί και άλλοι συντελεστές συσχέτισης όπως ο συντελεστής συσχέτισης του Spearman και ο συντελεστής συσχέτισης του Kendall, οι οποίοι στοχεύουν σε μεγαλύτερη ευαισθησία σε μη γραμμικές σχέσεις. Οι αναλύσεις συσχέτισης δεν μπορούν να ερμηνευθούν ως θεμελίωση σχέσεων αιτίου-αποτελέσματος και μπορούν να υποδείξουν μόνο πώς ή σε ποιο βαθμό οι μεταβλητές συνδέονται μεταξύ τους.

Για τον προσδιορισμό της συσχέτισης πρέπει να υπολογιστεί πρώτα η συνδιακύμανση των δύο εν λόγω μεταβλητών και η τυπική απόκλιση κάθε μεταβλητής. Ο συντελεστής συσχέτισης προσδιορίζεται διαιρώντας τη συνδιακύμανση με το γινόμενο των τυπικών αποκλίσεων των δύο μεταβλητών.

Η τυπική απόκλιση είναι ένα μέτρο της διασποράς των δεδομένων από τον μέσο όρο της. Η συνδιακύμανση είναι ένα μέτρο της κοινής μεταβλητότητας δύο τυχαίων μεταβλητών. Ωστόσο, το μέγεθός της είναι απεριόριστο, επομένως είναι δύσκολο να ερμηνευτεί. Ο συντελεστής συσχέτισης υπολογίζεται διαιρώντας τη συνδιακύμανση με το γινόμενο των δύο τυπικών αποκλίσεων όπως ακολουθεί:

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X \, \sigma_Y} \quad \textbf{(1)}$$

Το πιθανό εύρος τιμών για τον συντελεστή συσχέτισης είναι -1 έως 1. Με άλλα λόγια, οι τιμές δεν μπορούν να υπερβαίνουν το 1 ή να είναι μικρότερες από -1. Μια συσχέτιση -1 υποδηλώνει τέλεια αρνητική συσχέτιση και μια συσχέτιση 1 υποδηλώνει τέλεια θετική συσχέτιση. Εάν ο συντελεστής συσχέτισης είναι μεγαλύτερος από μηδέν, είναι θετική σχέση. Αντίθετα, εάν η τιμή είναι μικρότερη από το μηδέν, είναι αρνητική σχέση. Η τιμή μηδέν υποδηλώνει ότι δεν υπάρχει σχέση μεταξύ των δύο μεταβλητών.

Μια θετική συσχέτιση είναι μια σχέση μεταξύ δύο μεταβλητών στην οποία και οι δύο μεταβλητές κινούνται προς την ίδια κατεύθυνση. Επομένως, η αύξηση μιας μεταβλητής σχετίζεται με αύξηση της άλλης και η μείωση μιας μεταβλητής σχετίζεται με μείωση της άλλης. Αντίθετα, μια αρνητική συσχέτιση είναι μια σχέση μεταξύ δύο μεταβλητών στην οποία οι δύο μεταβλητές κινούνται προς αντίθετη κατεύθυνση. Συνεπώς, μια αύξηση σε μια μεταβλητή σχετίζεται με μείωση της άλλης.

1.2 Συντελεστής Μερικής Συσχέτισης

Ο συντελεστής μερικής συσχέτισης είναι ένα μέτρο που χρησιμοποιείται για την ποσοτικοποίηση της σχέσης μεταξύ δύο συνεχών μεταβλητών μετά τον έλεγχο των επιδράσεων άλλων μεταβλητών. Στον συντελεστή μερικής συσχέτισης αναιρείται η επίδραση που μπορεί να έχει μια τρίτη μεταβλητή επάνω σε ένα ζεύγος μεταβλητών.

Ο συντελεστής μερικής συσχέτισης παίρνει πάντα τιμές μεταξύ -1 και 1 όπου οι τιμές κοντά στο -1 ή το 1 υποδηλώνουν μια ισχυρότερη σχέση.

Ο συντελεστής συσχέτισης Pearson δεν λαμβάνει υπόψιν του τις πιθανές σχέσεις ενός ζεύγους μεταβλητών ως προς μία άλλη τρίτη μεταβλητή. Κατά συνέπεια, ο συντελεστής συσχέτισης Pearson δεν υπολογίζει την αληθινή συσχέτιση μεταξύ δύο μεταβλητών του πολλαπλού γραμμικού υποδείγματος της παλινδρόμησης, στην περίπτωση που υπάρχει κάποια άλλη μεταβλητή που πιθανόν να συσχετίζεται με τις δύο αυτές μεταβλητές. Για αυτό το λόγο καθίσταται απαραίτητη η χρήση του συντελεστή μερικής συσχέτισης.

Για τον προσδιορισμό της μερικής συσχέτισης υπολογίζονται συντελεστές μερικής συσχέτισης που περιγράφουν τη γραμμική σχέση μεταξύ δύο μεταβλητών ενώ ελέγχονται για την επίδραση μιας ή περισσότερων πρόσθετων μεταβλητών. Οι συσχετίσεις είναι μέτρα γραμμικής συσχέτισης. Δύο μεταβλητές μπορούν να συνδέονται τέλεια, αλλά εάν η σχέση δεν είναι γραμμική, ένας συντελεστής συσχέτισης δεν είναι το κατάλληλο στατιστικό στοιχείο για τη μέτρηση της συσχέτισής τους. Μερική συσχέτιση είναι η συσχέτιση μεταξύ δύο μεταβλητών μετά την αφαίρεση της επίδρασης μιας ή περισσότερων πρόσθετων μεταβλητών.

Ας υποθέσουμε ότι θέλουμε να βρούμε τη συσχέτιση μεταξύ του *X* και του *Y* ελέγχοντας το *W*. Αυτό ονομάζεται συντελεστής μερικής συσχέτισης και το σύμβολο του είναι $r_{XY.\ W}$. Σε αυτή την περίπτωση, η μερική συσχέτιση μπορεί να υπολογιστεί βάση των απλών συσχετίσεων μεταξύ των τριών μεταβλητών όπως ακολουθεί:

$$r_{p\ YX.W} = \frac{r_{XY} - r_{XW}\, r_{YW}}{\sqrt{(1-r_{XW}{}^2)(1-r_{YW}{}^2)}} \quad \textbf{(2)}$$

Όπου r_{XY} , r_{XW} , r_{YW} οι συντελεστές συσχέτισης που προέκυψαν από τον συντελεστή συσχέτισης Pearson.

Όπως και με τον τυπικό συντελεστή συσχέτισης, μια τιμή +1 υποδηλώνει μια τέλεια θετική γραμμική σχέση, μια τιμή -1 υποδηλώνει μια τέλεια αρνητική γραμμική σχέση και μια τιμή 0 καμία γραμμική σχέση.

Έχει αποδειχθεί ότι ο πίνακας των μερικών συσχετίσεων $R_p = r_{p\ ij}$ σχετίζεται με τον αντίστροφο του πίνακα συνδιακύμανσης Σ (inverse of the covariance matrix). Η παραπάνω σχέση αναπαρίσταται ως εξής:

$$r_{p\ ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\,\omega_{jj}}}\ ,\ \text{όπου}\ \Sigma^{-1} = (\omega_{ij}) \quad \textbf{(3)}$$

Η παραπάνω σχέση είναι εφαρμόσιμη μόνο στη περίπτωση που ο αριθμός των δειγμάτων *n* είναι μεγαλύτερος από τον αριθμό των μεταβλητών *p (n>p)*. Στην αντίθετη περίπτωση, όπου ο αριθμός των δειγμάτων είναι μικρότερος από τον αριθμό των μεταβλητών, ο πίνακας συνδιακύμανσης του δείγματος είναι ιδιάζον (singular) και όχι θετικά καθορισμένος. Επομένως, ο πίνακας αντίστροφης συνδιακύμανσης και κατά συνέπεια η μερική συσχέτιση δεν μπορούν πλέον να υπολογιστούν άμεσα. Επιπλέον, το μικρό μέγεθος δείγματος καθιστά επίσης άκυρες τις περισσότερες τυπικές στατιστικές δοκιμές για τη μερική συσχέτιση, καθώς αυτές συνήθως βασίζονται σε μεγάλο μέγεθος δείγματος *n* για ασύμπτωτη εγκυρότητα.

1.3 Συντελεστής Μερικής Συσχέτισης σε δεδομένα με αριθμό δειγμάτων μικρότερο από τον αριθμό των μεταβλητών

Η γενικά μικρή διαθεσιμότητα σε βιολογικά πειράματα όπως μικροσυστοιχίες και RNA-seq κατέστησε απαραίτητο τον υπολογισμό της μερικής συσχέτισης σε δεδομένα με αριθμό δειγμάτων μικρότερο από τον αριθμό των μεταβλητών (όπως γονιδίων, πρωτεϊνών, ασθενειών). Η αδυναμία υπολογισμού του πίνακα αντίστροφης συνδιακύμανσης και κατά συνέπεια της μερικής συσχέτισης για μέγεθος δείγματος που είναι μικρότερο από τον αριθμό των μεταβλητών προήγαγε την ανάπτυξη μεθόδων και εκτιμητών που οδήγησαν στην

αντιμετώπιση αυτού του προβλήματος. Οι μέθοδοι που αναπτύχθηκαν και αναλύονται παρακάτω είναι:

- Υπολογισμός μερικής συσχέτισης μέσω ενός ψευδοαντίστροφου πίνακα (Pseudoinverse)
- Υπολογισμός μερικής συσχέτισης μέσω της Ομαλοποιημένης Παλινδρόμησης (Regularized Regression)
- Υπολογισμός μερικής συσχέτισης από την εκτίμηση ενός αραιού πίνακα αντίστροφης συνδιακύμανσης χρησιμοποιώντας ποινή L1 (Estimation of sparse precision matrix using ℓ1 penalty)
- Εκτίμηση της μερικής συσχέτισης μέσω συρρίκνωσης (Shrinkage):
  1. Συρρίκωση της συνδιακύμανσης (Shrinkage covariance)
  2. Προσαρμοστική συρρίκνωση της συνδιακύμανσης (Adaptive shrinkage covariance)
- Υπολογισμός μερικής συσχέτισης χαμηλής τάξης (Low order partial correlation)

1.4 Μερική Συσχέτιση και Βιολογικά Συστήματα

Οι συσχετίσεις των βιολογικών δεδομένων είναι δυνατόν να τεκμηριωθούν και να απεικονιστούν με δίκτυα συσχέτισης. Ωστόσο, δεν είναι δυνατόν να προκύψει κάποια πληροφορία σχετικά με την αιτιότητα από τα δίκτυα συσχέτισης. Επίσης, είναι ιδιαίτερα δύσκολο να προσδιοριστούν οι πρωταρχικές εξαρτήσεις και οι αληθινές αλληλεπιδράσεις μεταξύ των δεδομένων. Ένας λόγος για τον οποίο αυτές οι αλληλεπιδράσεις είναι δύσκολο να εντοπιστούν είναι το πρόβλημα υπερπροσαρμογής που προκύπτει από τα σύνολα δεδομένων με μεγάλο όγκο μεταβλητών και μικρό αριθμό δειγμάτων. Ωστόσο, είναι δυνατό να εξεταστούν πιο στενά και οι αλληλεπιδράσεις σε σύνολα δεδομένων με μικρό αριθμό δειγμάτων. Ένας τρόπος εξαγωγής πληροφοριών σχετικά με αυτές τις αλληλεπιδράσεις είναι η χρήση της μερικής συσχέτισης. Τα δίκτυα συσχέτισης αντιπροσωπεύουν συντελεστές συσχέτισης που υπολογίζονται για ζεύγη μεταβλητών, ανεξάρτητα από όλες τις άλλες μεταβλητές. Αντίθετα, τα δίκτυα μερικής συσχέτισης αντιπροσωπεύονται από συντελεστές μερικής συσχέτισης που υπολογίζονται για ζεύγη μεταβλητών όταν λαμβάνονται υπόψη όλες οι άλλες μεταβλητές. Κατά συνέπεια, οι μερικές συσχετίσεις αντιπροσωπεύουν άμεσες συσχετίσεις, ενώ οι απλές συσχετίσεις δεν κάνουν διάκριση μεταξύ έμμεσων και άμεσων συσχετίσεων. Η μερική συσχέτιση είναι δυνατόν να εξάγει πληροφορία σχετικά με την αιτιότητα των μεταβλητών. Η ανάλυση μερικής συσχέτισης είναι μια κατάλληλη μέθοδος για την ανίχνευση της σχέσης μεταξύ μεταβλητών σε σύνολα δεδομένων όταν ο αριθμός των μεταβλητών είναι πολύ μεγαλύτερος από το μέγεθος του δείγματος. Συμπερασματικά, οι

αλληλεπιδράσεις των βιολογικών δεδομένων που προκύπτουν από τον υπολογισμό της μερικής συσχέτισης μπορούν να τεκμηριωθούν και να απεικονιστούν σε δίκτυα όπου οι κόμβοι αντικατοπτρίζουν τις μεταβλητές και οι ακμές την μεταξύ τους σχέση (συσχέτιση).

2 Θεωρητικό υπόβαθρο μεθόδων

2.1 Υπολογισμός Μερικής Συσχέτισης μέσω του Ψευδοαντίστροφου Πίνακα

2.1.1 Ψευδοαντίστοφος (Pseudoinverse)

Στην Γραμμική Άλγεβρα, η μέθοδος αποσύνθεσης ενός πίνακα A (Singular Value Decomposition, ανάλυση πίνακα σε ιδιάζουσες τιμές)**,** είναι η παραγοντοποίηση αυτού του πίνακα σε τρεις πίνακες. Είναι ουσιαστικά η αλλαγή συντεταγμένων που κάνει τον πίνακα πιο απλό, η γενίκευση της διαγωνιοποίησης.

Εάν ένας τετραγωνικός πίνακας $A \neq 0$ είναι διαγωνιοποιήσιμος, τότε υπάρχει πίνακας P, τέτοιος ώστε:

$$A = P\,D\,P^{-1} = P\,D\,P^{T},\ P^{-1} = P^{T} \quad \textbf{(4)}$$

Όπου ο πίνακας D είναι ένας διαγώνιος πίνακας με τα στοιχεία της διαγωνίου του να είναι οι ιδιοτιμές του πίνακα A και οι στήλες του P να είναι τα ιδιοδιανύσματα του πίνακα A.

Ωστόσο, δεν είναι όλοι οι πίνακες διαγωνιοποιήσιμοι. Μάλιστα, ο τυπικός ορισμός για το αντίστροφο ενός πίνακα αποτυγχάνει εάν ο πίνακας δεν είναι τετράγωνος ή ιδιάζον (singular: ένας πίνας με μηδενική ορίζουσα). Η ανάλυση ενός πίνακα σε ιδιάζουσες τιμές (SVD) είναι μία μέθοδος διαγωνιοποίησης οποιουδήποτε mxn πίνακα, ακόμη κι εάν ο πίνακας A δεν είναι τετραγωνικός όπου:

$$A = U\,\Sigma\,V^{*} \quad \textbf{(5)}$$

Ο πίνακας Σ στην μέθοδο (SVD) είναι διαγώνιος και μοναδικά ορισμένος (καθώς όλες οι τετραγωνικές ρίζες των ιδιοτιμών ονομάζονται ιδιάζουσες τιμές του πίνακα A), αλλά μπορεί να μην είναι τετραγωνικός. Τα στοιχεία της διαγωνίου του πίνακα Σ δεν είναι απαραίτητα ιδιοτιμές του πίνακα A, αλλά ιδιάζουσες τιμές που αποτελούν γενίκευση των ιδιοτιμών (δηλαδή οι τετραγωνικές ρίζες των ιδιοτιμών του). Οι πίνακες U και V^{*} στην μέθοδο (SVD) είναι τετραγωνικοί αλλά μπορεί να μην είναι της ίδιας διάστασης και όχι απαραίτητα αντίστροφοι ο ένας στον άλλο. Επίσης, οι στήλες των πινάκων U και V^{*} μπορεί να μην αποτελούν τα ιδιοδιανύσματα του πίνακα A, αλλά τα δεξιά και αριστερά ιδιάζοντα διανύσματα αντίστοιχα, με τον αστερίσκο στον πίνακα V^{*} να δηλώνει πίνακα μετάθεσης (transpose matrix), εάν ο πίνακας αποτελείται από πραγματικές τιμές. Κάθε πίνακας μπορεί να μην έχει αντίστροφο, αλλά κάθε πίνακας έχει ψευδοαντίστροφο, ακόμη κι εάν είναι μη τετραγωνικός.

Εάν ισχύει η Εξίσωση 5, $A = U \Sigma V^*$, τότε ο ψευδοαντίστροφος (Moore-Penrose) του A θα είναι:

$$A^+ = V \Sigma^+ U^* \quad \textbf{(6)}$$

όπου ο Σ^+ θα σχηματίζεται από τον αντίστροφο όλων των μη μηδενικών στοιχείων. Δηλαδή, εάν ο πίνακας Σ είναι ένας mxn πίνακας, τότε ο Σ^+ θα είναι nxm πίνακας. Για τον υπολογισμό του ψευδοαντίστροφου χρησιμοποιείται η συνάρτηση pseudoinverse η οποία παρέχεται από το corpcor package της R.

2.1.2 Υπολογισμός μερικής συσχέτισης από τον ψευδοαντίστροφο πίνακα της συνδιακύμανσης.

Στην περίπτωση που η ορίζουσα του πίνακα διακύμανσης-συνδιακύμανσης είναι αριθμητικά μηδέν και συνεπώς ο πίνακας είναι ιδιάζων, ο αντίστροφος αποτυγχάνει να υπολογιστεί βάση του τυπικού ορισμού. Ωστόσο, είναι δυνατόν να υπολογιστεί ο ψευδοαντίστροφος πίνακας χρησιμοποιώντας το αντίστροφο γενικευμένου πίνακα Moore-Penrose. Ο πίνακας μερικής συσχέτισης προκύπτει από την Εξίσωση 3, όπου ω_{ij} είναι το *(i,j)* στοιχείο του πίνακα αντίστροφης διακύμανσης-συνδιακύμανσης που προέκυψε από τον ψευδοαντίστροφο πίνακα και $r_{p\ ij}$ το *(i,j)* στοιχείο του πίνακα μερικής συσχέτισης.

2.1.3 Πακέτα για υπολογισμό μερικής συσχέτισης μέσω του ψευδοαντίστροφου πίνακα της συνδιακύμανσης

2.1.3.1 Corpcor package

Το Corpcor package [1] της R παρέχει μια αποτελεσματική εκτίμηση του πίνακα συνδιακύμανσης (covariance) και της μερικής συσχέτισης (partial correlation). Οι μερικές συσχετίσεις είναι αρνητικές τυποποιημένες συγκεντρώσεις οι οποίες με τη σειρά τους είναι τα μη διαγώνια στοιχεία του πίνακα αντίστροφης συσχέτισης ή συνδιακύμανσης. Στα γραφικά μοντέλα Gauss, οι μερικές συσχετίσεις αντιπροσωπεύουν τις άμεσες αλληλεπιδράσεις μεταξύ δύο μεταβλητών, που εξαρτώνται από όλες τις υπόλοιπες μεταβλητές.

Στο παραπάνω στηρίζεται η **συνάρτηση cor2pcor(m, tol)** του πακέτου η οποία δέχεται ως είσοδο έναν πίνακα συσχέτισης Pearson ή ένα πίνακα συνδιακύμανσης (correlation ή covariance matrix) και υπολογίζει τον πίνακα μερικής συσχέτισης. Η μερική συσχέτιση υπολογίζεται μέσω του αντίστροφου πίνακα της συνδιακύμανσης στην περίπτωση που ο πίνακας της

συνδιακύμανσης είναι θετικά ορισμένος και μέσω ενός ψευδοαντίστροφου πίνακα της συνδιακύμανσης στην περίπτωση που ο πίνακας της συνδιακύμανσης είναι μη θετικά ορισμένος.

Code Corpcor – cor2pcor

```
1. library("GeneNet")
2. library("corpcor")
3.
4. data(ecoli)
5. cov_matrix<-cov(ecoli)
6. correlation<- cor(ecoli)
7.
8. partial_cor1<-cor2pcor(correlation) #input: correlation matrix
9. partial_cor2<-cor2pcor(cov_matrix) #input: covariance matrix
10.
```

Κώδικας 1: Corpcor- cor2pcor

2.1.3.2 Ppcor package

Το Ppcor package [2], [3] της R περιλαμβάνει τη συνάρτηση pcor() η οποία μπορεί να υπολογίσει τις μερικές συσχετίσεις για κάθε ζεύγος μεταβλητών. Επιπλέον, μας δίνει την τιμή p-value καθώς και στατιστικά για κάθε ζεύγος μεταβλητών εάν ο αριθμός των μεταβλητών είναι μεγαλύτερος ή ίσος με το μέγεθος του δείγματος.

Η **συνάρτηση pcor(x, method = c("pearson", "kendall", "spearman"))** δέχεται ως είσοδο έναν πίνακα δεδομένων και μια συμβολοσειρά χαρακτήρων που δείχνει ποιος συντελεστής μερικής συσχέτισης πρόκειται να υπολογιστεί. Ένα από τα "pearson" (προεπιλογή), "kendall" ή "spearman" μπορεί να επιλεχθεί. Η συνάρτηση επιστρέφει ένα πίνακα μερικής συσχέτισης.

Η διαδικασία υπολογισμού του πίνακα μερικής συσχέτισης είναι όμοια με αυτή της συνάρτησης cor2pcor() που ανήκει στο corpcor package. Συγκεκριμένα, η συνάρτηση pcor υπολογίζει τον πίνακα συνδιακύμανσης του πίνακα δεδομένων x και στην συνέχεια τον πίνακας αντίστροφης συνδιακύμανσης του. Όταν η ορίζουσα του πίνακα διακύμανσης-συνδιακύμανσης είναι αριθμητικά μηδέν, το πακέτο R ppcor υπολογίζει το ψευδοαντίστροφο χρησιμοποιώντας το αντίστροφο γενικευμένου πίνακα Moore-Penrose. Ο πίνακας μερικής συσχέτισης προκύπτει από την Εξίσωση 3.

Code ppcor - pcor

```
1. library("GeneNet")
2. library("ppcor")
```

```
3.
4. data(ecoli)
5. part_cor<-pcor(ecoli)$estimate
```

Κώδικας 2: Ppcor - pcor

2.2 Υπολογισμός Μερικής Συσχέτισης μέσω της Ομαλοποιημένης Παλινδρόμησης

Η εκτίμηση της μερικής συσχέτισης μπορεί να υπολογιστεί μέσω του ορισμού της Ομαλοποιημένης Παλινδρόμησης. Ο ορισμός αυτός [4] αναφέρει ότι για οποιαδήποτε μέθοδο παλινδρόμησης $\hat{\beta}^{(i)}{}_{reg}$ που αποδίδει (ομαλοποιημένες) εκτιμήσεις του μοντέλου γραμμικής παλινδρόμησης, ορίζεται η αντίστοιχη εκτίμηση των μερικών συσχετίσεων ως:

$$\hat{r}_{p\,ij} = sign(\hat{\beta}^{(i)}{}_{j,reg})\, min\{1, \sqrt{\hat{\beta}^{(i)}{}_{j,reg}\ \hat{\beta}^{(j)}{}_{i,reg}}\}$$

$$\text{αν } sign(\hat{\beta}^{(i)}{}_{j,reg}) = sign(\hat{\beta}^{(j)}{}_{i,reg}) \qquad \textbf{(7)}$$

και 0 διαφορετικά.

Αυτός ο ορισμός διασφαλίζει ότι οι εκτιμώμενοι συντελεστές μερικής συσχέτισης είναι πάντα καλά καθορισμένοι και ότι βρίσκονται στο διάστημα [-1, 1].

2.2.1 Μέθοδοι ομαλοποιημένης Παλινδρόμησης (Regularized Regression)

Οι μέθοδοι Παλινδρόμησης που χρησιμοποιούνται για τον υπολογισμό της Εξίσωσης 7 είναι οι μέθοδοι:

- Μερικών ελαχίστων τετραγώνων (Partial Least Squares)
- Παλινδρόμης Ridge (Ridge Regression)
- Lasso
- Προσαρμοστικού Lasso (Adaptive Lasso)

Οι μέθοδοι μη αραιής παλινδρόμησης [4], Ridge Regression και Partial Least Squares, παρουσιάζουν συντηρητική συμπεριφορά όταν συνδυάζονται με πολλαπλές δοκιμές ψευδούς ανακάλυψης (False Discovery Rate) προκειμένου να εκτιμήσουν εάν υπάρχει μια ακμή στο δίκτυο. Για δίκτυα με μεγαλύτερη πυκνότητα, η διαφορά στην απόδοση των μεθόδων μειώνεται. Για αραιά δίκτυα, το Lasso έχει τάση να επιλέγει πάρα πολλές ακμές, ενώ το προσαρμοστικό Lasso

δύο σταδίων παρέχει πιο αραιές λύσεις. Τόσο οι αραιές όσο και οι μη αραιές μέθοδοι είναι σε θέση να ανακατασκευάσουν δίκτυα με δομές συστάδων.

Τα Μερικά ελάχιστα τετράγωνα τείνουν να επιλέγουν πολλές ακμές και να κατασκευάζουν πολύ πυκνά δίκτυα. Ωστόσο, στην περίπτωση που χρησιμοποιείται η υποδειγματοληψία, η προσέγγιση συρρίκνωσης είναι πιο σταθερή από τις προσεγγίσεις που βασίζονται στην παλινδρόμηση.

2.2.1.1 Μερικά ελάχιστα τετράγωνα (Partial Least Squares)

Η παλινδρόμηση Μερικών Ελάχιστων Τετράγωνων (PLS) [4], [5] μπορεί να χρησιμοποιηθεί για τον υπολογισμό της μερικής συσχέτισης μέσω της Εξίσωσης 7. Το PLS είναι μια μέθοδος για επιβλεπόμενη μείωση διαστάσεων.

Η κύρια ιδέα του PLS είναι να δημιουργήσει μερικά ορθογώνια στοιχεία από τα αρχικά δεδομένα $X^{(\backslash i)}$ και να τα χρησιμοποιήσει ως προγνωστικούς παράγοντες σε ένα ελάχιστο τετράγωνο. Μια συνιστώσα PLS, $t = X^{(\backslash i)}w$ είναι ένας γραμμικός συνδυασμός των αρχικών προγνωστικών παραγόντων που έχουν μέγιστη συνδιακύμανση με το διάνυσμα απόκρισης $X^{(i)}$, υπό την πρόσθετη υπόθεση ότι τα στοιχεία είναι αμοιβαία ορθογώνια. Τυπικά, η k-η συνιστώσα PLS ορίζεται ως εξής:

$$\begin{aligned} \boldsymbol{w}_k &= \arg\max_{\|\boldsymbol{w}\|=1} \operatorname{cov}\left(\boldsymbol{X}^{(\backslash i)}\boldsymbol{w}, \boldsymbol{X}^{(i)}\right)^2 \\ \text{s.t.} \quad & \boldsymbol{w}^\top \boldsymbol{X}^{(\backslash i)\top}\boldsymbol{X}^{(\backslash i)}\boldsymbol{w}_l = 0 \text{ for } l < k. \end{aligned} \quad \textbf{(8)}$$

Έτσι, το PLS ρυθμίζει το πρόβλημα παλινδρόμησης συμπιέζοντας τις μεταβλητές p σε ένα μικρό αριθμό m ορθογώνιων συνιστωσών $T = (t_1, \ldots, t_m)$ και παλινδρομώντας τη μεταβλητή απόκρισης σε αυτές τις συνιστώσες. Μετά την επανακλιμάκωση (rescaling) των διανυσμάτων βάρους $w_k (k = 1, \ldots, m)$ τέτοια ώστε το t_k να έχει μήκος 1, οι συντελεστές παλινδρόμησης ορίζονται ως εξής:

$$\hat{\boldsymbol{\beta}}_{\text{pls}}^{(i)} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)\boldsymbol{T}^\top \boldsymbol{X}^{(i)}. \quad \textbf{(9)}$$

Ενώ η αρχική διατύπωση του PLS κλιμακώνεται με τον αριθμό p των μεταβλητών, είναι επίσης δυνατόν να υλοποιηθεί μόνο με κλιμάκωση του αριθμού n των παρατηρήσεων. Αυτό οδηγεί σε δραματική μείωση του χρόνου υπολογισμού για $p>>n$. Αξιοσημείωτο είναι ότι ο αριθμός των στοιχείων PLS είναι μια παράμετρος μοντέλου που πρέπει να βελτιστοποιηθεί για καθένα από τα μοντέλα παλινδρόμησης p. Οι τυπικές τεχνικές επιλογής μοντέλων είναι

κριτήρια διασταυρούμενης επικύρωσης ή πληροφόρησης που βασίζονται σε βαθμούς ελευθερίας. Στο πλαίσιο των ρυθμιστικών δικτύων γονιδίων, οι Tenenhaus et.al. προτείνουν τη χρήση του ίδιου αριθμού συνιστωσών *m* για όλα τα μοντέλα παλινδρόμησης *p*. Ακόμη, στη μελέτη χρησιμοποιείται η τεχνική διασταυρούμενης επικύρωσης με k-fold. Καθώς οι συντελεστές PLS δεν είναι αραιοί, οι λαμβανόμενες μερικές συσχετίσεις είναι γενικά μη μηδενικές. Επομένως, πρέπει να χρησιμοποιηθεί μια διαδικασία στατιστικής δοκιμής για να προσδιοριστεί ποιες ακμές είναι σημαντικές όπως αυτή των πολλαπλών δοκιμών ψευδούς ανακάλυψης (fdr).

2.2.1.1.1 Parcor package

Η **συνάρτηση pls.net(X, scale = TRUE, k = 10, ncomp = 15,verbose=FALSE)** του πακέτου parcor [6] της R υπολογίζει τον πίνακα των μερικών συσχετίσεων μέσω μιας εκτίμησης των αντίστοιχων μοντέλων παλινδρόμησης με τη χρήση των Μερικών Ελάχιστων Τετράγωνων. Η παραπάνω συνάρτηση δέχεται ως είσοδο έναν X πίνακα δεδομένων και έναν αριθμό *k* ο οποίος καθορίζει τον αριθμό των διαχωρισμών σε μια k-fold διασταυρούμενη επικύρωση (Η προεπιλεγμένη τιμή είναι k=10).

Code parcor – pls.net

```
1. library("GeneNet")
2. library("parcor")
3. data(ecoli)
4.
5. parcorr<-pls.net(ecoli, k=5)$pcor
```

Κώδικας 3: Parcor – pls.net

2.2.1.2 Παλινδρόμηση Ridge (Ridge Regression)

Η παλινδρόμηση Ridge [4] είναι μια απλή τεχνική ομαλοποιημένης παλινδρόμησης. Η ομαλοποίηση πραγματοποιείται με την προσθήκη όρου ποινής *P(β)* στο κριτήριο των ελαχίστων τετραγώνων. Η παλινδρόμηση ridge βασίζεται σε έναν όρο ποινής $l2$ της σχέσης:

$$P(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_2^2 = \lambda \sum_i \beta_i^2, \quad \textbf{(10)}$$

όπου $\lambda > 0$ υποδηλώνει την παράμετρο ποινής. Αυτό οδηγεί στη μείωση της διακύμανσης και στην αποφυγή της υπερβολικής προσαρμογής. Η λύση που

προκύπτει με τη παλινδρόμηση ridge εξαρτάται από την παράμετρο της ποινής λ. Η βέλτιστη ποσότητα ποινής της ποινής αυτής επιλέγεται με βάση την τυπική διασταυρούμενη επικύρωση k-fold. Τέλος, λόγω του ότι η παλινδρόμηση ridge δεν οδηγεί σε αραιές λύσεις, χρησιμοποιούνται μεγάλης κλίμακας πολλαπλές δοκιμές ψευδούς ανακάλυψης (fdr) για να βρεθούν οι σημαντικές ακμές.

2.2.1.2.1 Parcor package

Η συνάρτηση **ridge.net(X, lambda, plot.it = FALSE, scale = TRUE, k = 10,verbose=FALSE)** του πακέτου parcor της R [6] υπολογίζει τον πίνακα των μερικών συσχετίσεων μέσω μιας εκτίμησης των αντίστοιχων μοντέλων παλινδρόμησης με τη χρήση της παλινδρόμησης Ridge. Η παραπάνω συνάρτηση δέχεται ως είσοδο έναν X πίνακα δεδομένων και έναν αριθμό *k* ο οποίος καθορίζει τον αριθμό των διαχωρισμών σε μια k-fold διασταυρούμενη επικύρωση (Η προεπιλεγμένη τιμή είναι k=10).

Code parcor- ridge.net

```
1. library("GeneNet")
2. library("parcor")
3. data(ecoli)
4.
5. parcorr<-ridge.net(ecoli, k=5)$pcor
```

Κώδικας 4: Parcor- ridge.net

2.2.1.3 LASSO

Οι Meinshausen και Bühlmann πρότειναν την εκτίμηση των συντελεστών παλινδρόμησης της Εξίσωσης 9 με τη μέθοδο Lasso. Παρόμοια με την παλινδρόμηση Ridge, οι εκτιμώμενοι συντελεστές παλινδρόμησης επιλέγονται για να ελαχιστοποιήσουν το κριτήριο των ελαχίστων τετραγώνων με ποινές. Η παλινδρόμηση lasso βασίζεται σε μια ποινή $l1$ της μορφής:

$$P(\boldsymbol{\beta}) = \lambda \| \boldsymbol{\beta} \|_1 = \lambda \sum_i |\beta_i|, \quad \textbf{(11)}$$

όπου $\lambda > 0$ είναι η παράμετρος ομαλοποίησης. Με την ποινή $l1$ πολλοί εκτιμώμενοι συντελεστές παλινδρόμησης θα είναι ίσοι με 0. Ως αποτέλεσμα, με τη μέθοδο Lasso προκύπτει ένας αραιός εκτιμητής του πίνακα των μερικών συσχετίσεων. Κάθε τιμή $\hat{r}_{\mathrm{p\,i}j,lasso} \neq 0$ είναι μια συσχέτιση μεταξύ του *i* και *j* και μπορεί να απεικονιστεί στο δίκτυο ως ακμή. Η επιλογή της ποινής λ καθορίζεται για καθεμία από τις *p* υψηλών διαστάσεων παλινδρομήσεις διαδοχικά. Αυτό

μπορεί να πραγματοποιηθεί χρησιμοποιώντας κάποιο σχήμα διασταυρούμενης επικύρωσης (cross-validation) ή κριτήριο πληροφόρησης (information criteria). Στη μελέτη χρησιμοποιείται η ποινή oracle για τη βέλτιστη πρόβλεψη που προσδιορίζεται χρησιμοποιώντας διασταυρούμενη επικύρωση k-fold.

2.2.1.4 Two-stage adaptive LASSO

Οι Zhou και άλλοι πρότειναν την εκτίμηση των συντελεστών παλινδρόμησης της Εξίσωσης 9 με τη μέθοδο Lasso δύο σταδίων. Το Lasso είναι μόνο ασυμπτωτικά συνεπές για την επιλογή συνδιακύμανσης όταν απαιτεί ορισμένες απαραίτητες συνθήκες μεταξύ των μεταβλητών στο δίκτυο GGM. Η προσαρμοστική (adaptive) διαδικασία Lasso δύο σταδίων είναι συνεπής για την επιλογή μοντέλων υψηλών διαστάσεων σε γραφικά μοντέλα Gauss υπό μάλλον γενικές και λιγότερο περιοριστικές συνθήκες. Το προσαρμοστικό Lasso προκύπτει από το Lasso με βάρη ποινής ως:

$$P(\boldsymbol{\beta}) = \lambda \sum_{i} \widehat{\omega}_i \left| \beta_i \right|, \quad \textbf{(12)}$$

όπου τα βάρη $\widehat{\omega}_{\mathrm{i}}$ επιλέγονται με τρόπο που εξαρτάται από τα δεδομένα. Συγκεκριμένα, το προσαρμοστικό Lasso ορίζεται ως εξής:

«Έστω ότι $\hat{\beta}$ είναι ένας $\sqrt{n}$ συνεπής αρχικός εκτιμητής του β (για παράδειγμα, μπορεί να χρησιμοποιηθεί ο εκτιμητής ελαχίστων τετραγώνων $\hat{\beta}_{ols}$). Για ένα επιλεγμένο γ > 0 (η πιο συνηθισμένη επιλογή είναι γ = 1) τα βάρη ορίζονται ως $\widehat{\omega}_{\mathrm{i}} = 1/\left|\hat{\beta}_{i,ols}\right|^{\gamma}$.»

Στη συγκεκριμένη υλοποίηση χρησιμοποιείται ο εκτιμητή Lasso ως αρχικός εκτιμητής και το βάρος ορίζεται ως εξής:

$$\widehat{\omega}_{\mathrm{i}} = 1/\left|\hat{\beta}_{i,lasso}\right|. \quad \textbf{(13)}$$

Το ποσό της ποινής τόσο στο αρχικό στάδιο Lasso όσο και στο δεύτερο στάδιο Lasso με βάρη ποινής καθορίζεται μέσω k-fold cross-validation. Το προσαρμοστικό Lasso είναι τουλάχιστον τόσο αραιό όσο το Lasso. Για τον υπολογισμό της μερικής συσχέτισης χρησιμοποιείται η Εξίσωση 9 όπου μια συσχέτιση υπάρχει μόνο όταν ισχύει $\hat{r}_{p\ ij,adaptive\ lasso} \neq 0$. Ακόμη, παρατηρείται ότι για την επιλογή μοντέλου, τα βέλτιστα βάρη πρέπει να καθοριστούν σε καθεμία από τις διαιρέσεις k της διασταυρωμένης επικύρωσης. Λόγω του ότι τα βέλτιστα βάρη καθορίζονται μέσω διασταυρούμενης επικύρωσης k-fold, μια προσαρμογή lasso πρέπει να υπολογιστεί k^2 φορές, γεγονός που οδηγεί σε υψηλό υπολογιστικό κόστος.

2.2.1.4.1 Parcor package

Η συνάρτηση **adalasso.net(X, k=10, use.Gram=FALSE, both=TRUE, verbose=FALSE, intercept=TRUE)** του πακέτου parcor [6] της R υπολογίζει τον πίνακα των μερικών συσχετίσεων με βάση μια εκτίμηση των αντίστοιχων μοντέλων παλινδρόμησης μέσω lasso και προσαρμοστικού lasso αντίστοιχα. Η παραπάνω συνάρτηση δέχεται ως είσοδο έναν X πίνακα δεδομένων και έναν αριθμό k ο οποίος καθορίζει τον αριθμό των διαχωρισμών σε μια k-fold διασταυρούμενη επικύρωση. Το ίδιο k χρησιμοποιείται για την εκτίμηση των βαρών και την εκτίμηση του όρου ποινής του προσαρμοστικού lasso. (Η προεπιλεγμένη τιμή είναι k=10).

Code parcor- adalasso.net

```
1. library("GeneNet")
2. library("parcor")
3. data(ecoli)
4.
5. parcorr1<-adalasso.net(ecoli, k=5)$pcor.lasso  #lasso
6. parcorr2<-adalasso.net(ecoli,    k=5)$pcor.adalasso    #adaptive
   lasso
```

Κώδικας 5: Parcor- adalasso.net

2.3 Υπολογισμός Μερικής Συσχέτισης από την εκτίμηση ενός αραιού πίνακα αντίστροφης συνδιακύμανσης χρησιμοποιώντας ποινή L1.

2.3.1 Graphical Lasso

Το Graphical Lasso [7] είναι ένας γρήγορος και αποτελεσματικός αλγόριθμος για την εκτίμηση των πινάκων αντίστροφης συνδιακύμανσης. Το Graphical Lasso είναι ένα πλαίσιο ομαλοποίησης για την εκτίμηση του πίνακα συνδιακύμανσης Σ, υπό την υπόθεση ότι ο αντίστροφος του $\Theta = \Sigma^{-1}$ είναι αραιός. Το Θ ονομάζεται πίνακας ακριβείας ή πίνακας αντίστροφης συνδιακύμανσης. Στην περίπτωση που $\theta_{ij} = 0$, οι αντίστοιχες μεταβλητές X_i και X_j είναι υπό όρους ανεξάρτητες. Το Graphical Lasso ελαχιστοποιεί μια ℓ1-ρυθμισμένη αρνητική λογαριθμική-πιθανοφάνεια (log-likelihood):

$$\underset{\theta > 0}{minimaze}\ \mathrm{f}(\Theta) \coloneqq -\log\det(\Theta) + \mathrm{tr}(\Sigma\Theta) + \lambda\|\Theta\|_1 \quad \textbf{(14)}$$

Όπου το ο Σ είναι ο πίνακας συνδιακύμανσης του δείγματος, το $\|\Theta\|_1$ το άθροισμα των απόλυτων τιμών του Θ και το λ είναι μια παράμετρος ρύθμισης που ελέγχει το ποσοστό της συρρίκνωσης ℓ1.

Σε αυτό το μοντέλο, ο πίνακας ακριβείας μπορεί να εκτιμηθεί από τη συνάρτηση log-likelihood από την ακόλουθη εξίσωση [8]:

$$\Theta^* = \arg \min_{\Theta} -\log\det(\Theta) + \mathrm{tr}(\Sigma\Theta) + \lambda\|\Theta\|_1 \quad \textbf{(15)}$$

Η μερική συσχέτιση μπορεί να υπολογιστεί από την Εξίσωση 3 με την χρήση του εκτιμώμενου πίνακα αντίστροφος συνδιακύμανσης που προέκυψε από την υλοποίηση του GLASSO.

2.3.1.1 Glasso package

Ο πίνακας αντίστροφης συνδυακύμανσης μπορεί να υπολογιστεί με τη συνάρτηση glasso() του πακέτου glasso [9] της R. Ως είσοδο η **συνάρτηση glasso(s, rho)** δέχεται έναν πίνακας συνδιακύμανσης-διακύμανσης s, που μπορεί να υπολογιστεί απευθείας με τη βασική συνάρτηση "cov()" της R, και μια παράμετρο ρύθμισης λάμδα (lamda) η οποία μπορεί να εκχωρηθεί από την επιλογή "rho" στη συνάρτηση "glasso". Όταν η τιμή λάμδα αυξάνεται, ο βαθμός της αραιότητας στο δίκτυο αυξάνεται επίσης. Ο πίνακας μερικής συσχέτισης προκύπτει από την Εξίσωση 3. Η οποία υλοποιείται με τη χρήση της συνάρτηση cov2cor() του πακέτου Copcor της R.

Code glasso

```
1. library("glasso")
2. library("corpcor")
3. library("GeneNet")
4.
5. data(ecoli)
6. cov_matrix<-cov(ecoli)
7. gr_lasso<-glasso(cov_matrix, rho = .01)
8.
9. part_cor<-  -cov2cor(gr_lasso$wi)
10.  diag(part_cor)<- 1
```

Κώδικας 6: Glasso

2.3.2 SPACE (Sparse PARtial Correlation Estimation) με joint sparse regression model

Η μέθοδος αραιής μερικής συσχέτισης (SPACE) [10] προτάθηκε από τον Penget ως αποτελεσματική εναλλακτική λύση στις υπάρχουσες μεθόδους για αραιή

εκτίμηση της αντίστροφης συνδυακύμανσης. Η διαδικασία SPACE επαναλαμβάνεται μεταξύ (α) ενημέρωσης μερικών συσχετίσεων με παλινδρόμηση lasso από κοινού και (β) ξεχωριστής ενημέρωσης των μερικών διακυμάνσεων. Λαμβάνει επίσης υπόψη τη συμμετρία στην αντίστροφη συνδιακύμανση και είναι υπολογιστικά αποτελεσματική. Υπό κατάλληλες συνθήκες κανονικότητας, το SPACE αποδίδει συνεπείς εκτιμητές σε ρυθμίσεις υψηλών διαστάσεων. Όλες οι παραπάνω ιδιότητες καθιστούν το SPACE μια ελκυστική προσέγγιση βασισμένη σε παλινδρόμηση για την εκτίμηση αραιών γραφημάτων μερικής συσχέτισης.

Οι Peng, Wang και Zhu παρουσίασαν μια καινοτόμα μέθοδο για την ανίχνευση ζευγών μεταβλητών που έχουν μη μηδενικές μερικές συσχετίσεις σε δεδομένα υψηλής διάστασης-χαμηλού μεγέθους δείγματος. Το μοντέλο SPACE εκτιμά τις μερικές συσχετίσεις χρησιμοποιώντας κοινά μοντέλα αραιής παλινδρόμησης. Το SPACE υποθέτει τη συνολική αραιότητας του πίνακα μερικής συσχέτισης και χρησιμοποιεί τεχνικές αραιής παλινδρόμησης για την προσαρμογή του μοντέλου. Οι μη μηδενικές εγγραφές στον πίνακα συγκέντρωσης υποδηλώνουν εξάρτησης υπό όρους μεταξύ των δύο αντίστοιχων μεταβλητών. Επιπλέον, με την προϋπόθεση κανονικότητας μηδενικές εγγραφές στον πίνακα συσχέτισης υποδηλώνουν ανεξαρτησία των δύο μεταβλητών.

Το SPACE (Sparse PARtial Correlation Estimation) χρησιμοποιεί τεχνικές αραιής παλινδρόμησης επιβάλλοντας τη ποινή $l1$ σε μια κατάλληλη συνάρτηση απώλειας για την αντιμετώπιση του προβλήματος μικρού μεγέθους δείγματος-υψηλής διάστασης. Ακόμη, παρέχει την εκτίμηση της μερικής συσχέτισης χρησιμοποιώντας joint sparse regression models για τη δημιουργία ενός πίνακα αντίστροφης συνδιακύμανσης από τις τυχαίες παρατηρήσεις που λήφθηκαν προηγουμένως λαμβάνοντας το λάμδα ως την τιμή της ποινής. Σχετίζεται με την προσέγγιση επιλογής γειτονιάς που προτάθηκε από τους Meinshausen και Buhlmann όπου μια παλινδρόμηση lasso εκτελείται ξεχωριστά για κάθε μεταβλητή στις υπόλοιπες μεταβλητές.

Με τον έλεγχο της συνολικής αραιότητας του πίνακα μερικής συσχέτισης, ο χώρος είναι σε θέση να προσαρμόζεται αυτόματα για διαφορετικά μεγέθη γειτονιάς και έτσι να χρησιμοποιεί τα δεδομένα πιο αποτελεσματικά. Η προτεινόμενη μέθοδος χρησιμοποιεί επίσης ρητά τη συμμετρία μεταξύ των μερικών συσχετίσεων, η οποία βοηθά επίσης στη βελτίωση της αποτελεσματικότητας. Με το SPACE επιβάλλεται αραιότητα στον πίνακα μερικής συσχέτισης αντί στον πίνακα ακριβείας.

Το μοντέλο για την αραιή εκτίμηση της μερικής συσχέτισης υπολογίζεται ως εξής:

Υποθέτοντας ότι $\mathrm{Y}^{\mathrm{k}} = \left(\mathrm{y_1}^{\mathrm{k}}, \ldots, \mathrm{y_p}^{\mathrm{k}}\right)^{\mathrm{T}}$ είναι *iid* παρατηρήσεις από $(0, \Sigma)$ για $\mathrm{k} = 1, \ldots, \mathrm{n}$.

Συμβολίζουμε το δείγμα της i ης μεταβλητής ως $\mathrm{Y_i} = (\mathrm{y_i}^1, \ldots, \mathrm{y_i}^{\mathrm{n}})^{\mathrm{T}}$. Προτείνετε η ακόλουθη συνάρτηση απώλειας joint:

$$L_{\mathrm{n}}(\boldsymbol{\theta},\boldsymbol{\sigma},\mathbf{Y}) = \frac{1}{2}\left(\sum_{i=1}^{p} w_{\mathrm{i}} \left\| \boldsymbol{Y}_i - \sum_{j\neq i} \beta_{ij}\, \boldsymbol{Y}_j \right\|^2\right)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{p} w_{\mathrm{i}} \left\| \boldsymbol{Y}_i - \sum_{j\neq i} {r_p}^{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\, \boldsymbol{Y}_j \right\|^2\right) \quad \textbf{(16)}$$

Όπου ${r_p}^{ij} = sign\left(\beta_{\mathrm{ij}}\right)\sqrt{\beta_{\mathrm{ij}}\beta_{\mathrm{ji}}}$, $\theta = \left({r_p}^{12}, \ldots, {r_p}^{(p-1)p}\right)^T$, $\sigma = \left\{\sigma^{ii}\right\}^p$ όπου $\Sigma^{-1} = (\sigma_{ij})$ ο πίνακας αντίστροφης συνδιακύμανσης, $Y = \{Y^k\}^n$ $\kappa\alpha\iota$ $w = \{w_{\mathrm{i}}\}^p$ τα μη αρνητικά βάρη.

Στο space προτείνεται η εκτίμηση της μερικής συσχέτισης $\boldsymbol{\vartheta}$ ελαχιστοποιώντας μια συνάρτηση ποινής απώλειας:

$$L_{\mathrm{n}}(\boldsymbol{\theta},\boldsymbol{\sigma},\mathbf{Y}) = L_{\mathrm{n}}(\boldsymbol{\theta},\sigma,\mathbf{Y}) = J(\boldsymbol{\theta}) \quad \textbf{(17)}$$

όπου ο όρος ποινής *J(ϑ)* ελέγχει τη συνολική αραιότητα της τελικής εκτίμησης του θ και εστιάζεται στην $l1$ ποινή (Tibshirani 1996):

$$J(\boldsymbol{\theta}) = \lambda\|\theta\|_1 = \lambda \sum_{1\leq i<j\leq p} |\, r^{ij}| \quad \textbf{(18)}$$

2.3.2.1 Space package

Η **συνάρτηση space.joint(Y.m, lam1, lam2=0, sig=NULL, weight=NULL,iter=2)** του πακέτου Space [11] της R χρησιμοποιείται για την εκτίμηση της μερικής συσχέτισης με joint sparse regression models. Η συνάρτηση δέχεται ως είσοδο ένα πίνακα δεδομένων Y.m, μια αριθμητική τιμή lam1 η οποία είναι η παράμετρος ποινής ℓ1 νόρμα και μια αριθμητική τιμή lam2 η οποία στην περίπτωση που δεν καθορίζεται η παλινδρόμηση lasso χρησιμοποιείται στο μοντέλο κοινής αραιής παλινδρόμησης (Joint Sparse Regression Model). Διαφορετικά, η ελαστική καθαρή παλινδρόμηση χρησιμοποιείται στο Joint Sparse Regression Model και το lam2 χρησιμεύει ως παράμετρος ποινής ℓ2 νόρμα. Η συνάρτηση δέχεται ως είσοδο ακόμη μια αριθμητική τιμή ή διάνυσμα weight η οποία καθορίζει τα βάρη ή τον τύπο των βαρών που χρησιμοποιούνται για κάθε παλινδρόμηση στο μοντέλο, ένα αριθμητικό διάνυσμα sig το οποίο είναι το διάνυσμα του σ^{ii} (η διαγώνιος του πίνακα αντίστροφης

συνδιακύμανσης) και έναν ακέραιο αριθμό iter ο οποίος είναι ο συνολικός αριθμός αλληλεπιδράσεων του μοντέλο για την εκτίμηση του σ^{ii} και της μερικής συσχέτισης. Όταν sig= NULL και/ή weight= NULL ή 2, το iter πρέπει να είναι τουλάχιστον 2. Η συνάρτηση επιστρέφει μια λίστα με τον εκτιμώμενο πίνακα μερικής συσχέτισης και το αριθμητικό διάνυσμα της εκτιμώμενης διαγωνίου σ^{ii}.

Code space- space.joint

```
1. library("space")
2. library(Biobase)
3. library("RCy3")
4. library(dplyr)
5. library("GeneNet")
6. library(MASS)
7. library(reshape2)
8. library(reshape)
9.
10. data(ecoli)
11. n= nrow(ecoli)
12. p= ncol(ecoli)
13.
14. alpha=1
15. l1=1/sqrt(n)*qnorm(1-alpha/(2*p^2))
16. iter=3
17.
18. space<-space.joint(ecoli, lam1=l1*n*1.56, lam2=0,
    iter=3)$ParCor
19. rownames(space)<-colnames(space)<- colnames(ecoli)
20.
21. adjacenty<-melt(space)
22. adjacenty<-adjacenty[which(adjacenty$value!= 1 &
    adjacenty$value!= 0),]
```

Κώδικας 7: Space- space.joint

2.3.3 SPACE (Sparse PARtial Correlation Estimation) εκτίμηση μερικών συσχετίσεων χρησιμοποιώντας την προσέγγιση επιλογής γειτονιάς

Η μέθοδος των N. Meinshausen και P. Bühlmann [12], [13] εκμεταλλεύεται τη σύνδεση μεταξύ μερικών συσχετίσεων και συντελεστών παλινδρόμησης και εκτελεί αραιή εκτίμηση των μερικών συσχετίσεων. Η αραιή εκτίμηση υπολογίζεται με την παλινδρόμηση καθεμιάς από τις μεταβλητές p στις υπόλοιπες, βρίσκοντας έτσι τους άμεσους γείτονες κάθε κόμβου λύνοντας ένα πρόβλημα lasso (Tibshirani, 1996). Δεδομένου ενός κλιμακούμενου πίνακα

δεδομένων με κεντραρισμένες στήλες $X \in \mathbb{R}^{n \cdot p}$ με στήλες x^j , η προσέγγιση επιλογής γειτονιάς επιλύει για κάθε μεταβλητή j ορίζεται ως εξής:

$$\boldsymbol{\beta}^j = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \beta_j = 0}{\operatorname{argmin}} \left\{ n^{-1} \|\mathbf{x}^j - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \quad \textbf{(19)}$$

Όπου το $\|\beta\|_1$ υποδηλώνει το άθροισμα των απόλυτων τιμών του β και είναι η $\ell 1$ νόρμα του διανύσματος συντελεστή. Το λ είναι μια παράμετρος ρύθμισης που ελέγχει το ποσό της συρρίκνωσης $\ell 1$.

2.3.3.1 Space package

Η **συνάρτηση space.neighbor(Y.m, lam1, lam2=0)** του πακέτου Space [11] της R χρησιμοποιείται για την εκτίμηση της μερικής συσχέτισης χρησιμοποιώντας την προσέγγιση επιλογής γειτονιάς (neighborhood selection approach). Η συνάρτηση δέχεται ως είσοδο ένα πίνακα δεδομένων Y.m, μια αριθμητική τιμή lam1 η οποία είναι η παράμετρος ποινής ℓ1 νόρμα, μια αριθμητική τιμή lam2 η οποία στην περίπτωση που δεν καθορίζεται η παλινδρόμηση lasso χρησιμοποιείται στην προσέγγιση επιλογής γειτονιάς. Διαφορετικά, η ελαστική καθαρή παλινδρόμηση χρησιμοποιείται και το lam2 χρησιμεύει ως παράμετρος ποινής ℓ2 νόρμα. Η συνάρτηση επιστρέφει μια λίστα με τον εκτιμώμενο πίνακα μερικής συσχέτισης και το αριθμητικό διάνυσμα της εκτιμώμενης διαγωνίου σ^{ii}.

Code space- space.neighbor

```
1. library("space")
2. library(Biobase)
3. library("RCy3")
4. library(dplyr)
5. library("GeneNet")
6. library(MASS)
7. library(reshape2)
8. library(reshape)
9.
10.  data(ecoli)
11.  n= nrow(ecoli)
12.  p= ncol(ecoli)
13.
14.  alpha=1
15.  l1=1/sqrt(n)*qnorm(1-alpha/(2*p^2))
16.
17.  space<- space.neighbor(ecoli, lam1=l1*n*1.56, lam2=0)$ParCor
18.  rownames(space)<-colnames(space)<- colnames(ecoli)
19.
20.  adjacenty<-melt(space)
```

```
21.  adjacenty<-adjacenty[which(adjacenty$value!=      1      &
   adjacenty$value!= 0),]
```

Κώδικας 8: Space- space.neighbor

2.4 Εκτίμηση Συρρίκνωσης (Shrinkage Estimates) της μερικής συσχέτισης

Για την εκτίμηση της μερικής συσχέτισης οι Schäfer & Strimmer [1] πρότειναν έναν αναλυτικό εκτιμητή συρρίκνωσης του πίνακα συσχέτισης και του πίνακα συνδιακύμανσης για την εκτίμηση της μερικής συσχέτισης. Ο εκτιμητής συνδυάζει γραμμικά την απεριόριστη συσχέτιση δείγματος με έναν κατάλληλο στόχο συσχέτισης σε ένα σταθμισμένο μέσο όρο. Η επιλογή αυτού του στόχου απαιτεί κάποια επιμέλεια: συγκεκριμένα, επιλέγεται η συρρίκνωση των εμπειρικών συσχετίσεων $\mathrm{R} = (\mathrm{r}_{ij})$ προς τον πίνακα ταυτότητας, ενώ οι εμπειρικές αποκλίσεις παραμένουν ανέπαφες. Στην περίπτωση αυτή η αναλυτικά προσδιορισμένη ένταση συρρίκνωσης είναι:

$$\hat{\lambda}^* = \frac{\sum_{\mathrm{i}\neq\mathrm{j}} \widehat{\mathrm{var}}(\mathrm{r}_{ij})}{\sum_{\mathrm{i}\neq\mathrm{j}} \mathrm{r}_{ij}{}^2} \quad \textbf{(21)}$$

Η εκτίμηση συρρίκνωσης εμφανίζει έναν αριθμό ευνοϊκών ιδιοτήτων. Για παράδειγμα, είναι πολύ πιο αποτελεσματική, πάντα θετική, και καλά ρυθμισμένη. Ο υπολογισμός της είναι φθηνός και δεν απαιτεί παραμέτρους συντονισμού, καθώς η αναλυτικά βέλτιστη ένταση συρρίκνωσης εκτιμάται από τα δεδομένα. Επιπλέον, οι εκτιμήσεις που προκύπτουν είναι σε μια μορφή που επιτρέπει τον γρήγορο υπολογισμό του αντιστρόφου τους χρησιμοποιώντας την ταυτότητα του πίνακα Woodbury.

Οι Schäfer & Strimmer προτείνουν τις παρακάτω σχέσεις για την εκτίμηση της συνδιακύμανσης Σ^* και της συσχέτισης r^* βασιζόμενοι στην αναλυτικά προσδιορισμένη ένταση συρρίκνωσης $\hat{\lambda}^*$:

$$\Sigma^*{}_{ij} = \begin{cases} \Sigma_{ij}\,, & i = j \\ r^*{}_{ij}\sqrt{\Sigma_{ii}\,\Sigma_{jj}}\,, & i \neq j \end{cases}$$

και

$$r^*{}_{ij} = \begin{cases} 1\,, & i = j \\ r_{ij}\,min\left(1, max\left(0\,, 1-\hat{\lambda}^*\right)\right), & i \neq j \end{cases} \quad \textbf{(22)}$$

Ο εκτιμώμενος συρρικνωμένος πίνακας συνδιακύμανσης $\Sigma^*{}_{ij}$ μπορεί πλέον να αντιστραφεί ώστε να υπολογιστεί η εκτίμηση της μερικής συσχέτισης με τη χρήση του βασικού τύπου της αντίστροφης συνδιακύμανσης της Εξίσωσης 3.

2.4.1 Πακέτα για υπολογισμό μερικής συσχέτισης μέσω εκτίμησης συρρίκνωσης

2.4.1.1 Corpcor package

Η **συνάρτηση pcor.shrink(x, lambda, w, verbose=TRUE)** του πακέτου Corpcor [1] της R υπολογίζει μια συρρικνωμένη εκτίμηση της μερικής συσχέτισης. Η συνάρτηση δέχεται ως είσοδο έναν πίνακα δεδομένων X και έναν αριθμό lamda που εκφράζει την ένταση συρρίκνωσης της συσχέτισης (εύρος 0-1). Εάν δεν προσδιορίζεται το lamda (η προεπιλογή) υπολογίζεται χρησιμοποιώντας έναν αναλυτικό τύπο από τους Schafer and Strimmer (2005). Για lamda=0 οι εμπειρικές συσχετίσεις ανακτώνται. Η χρήση του pcor.shrink(x) είναι πολύ πιο γρήγορη από το cor2pcor(cor.shrink(x)).

Code corpcor- pcor.shrink

```
1. library("corpcor")
2. library("GeneNet")
3.
4. data(ecoli)
5. partial_cor<-pcor.shrink(ecoli)
```

Κώδικας 9: Corpcor- pcor.shrink

2.4.1.2 GeneNet package

Το GeneNet [14] είναι ένα πακέτο της R για την ανάλυση δεδομένων υψηλών διαστάσεων που λαμβάνονται από δοκιμές λειτουργικής γονιδιωματικής υψηλής απόδοσης, όπως μικροσυστοιχίες ή μεταβολικά προφίλ. Συγκεκριμένα, το GeneNet επιτρέπει να εξάγουμε δίκτυα συσχέτισης γονιδίων μεγάλης κλίμακας. Αυτά είναι γραφικά μοντέλα Gauss (GGMs) που αντιπροσωπεύουν πολυμεταβλητές εξαρτήσεις σε βιομοριακά δίκτυα που προκύπτουν από τον συντελεστή μερικής συσχέτισης. Επομένως, το αποτέλεσμα μιας ανάλυσης που διεξάγεται από το GeneNet είναι ένα γράφημα όπου κάθε γονίδιο αντιστοιχεί σε έναν κόμβο και οι ακμές απεικονίζουν τις άμεσες εξαρτήσεις τους.

Το GeneNet εφαρμόζει έναν συγκεκριμένο ρυθμό εκμάθησης αλγόριθμου που επιτρέπει την εκτίμηση των GGM από μικρά δείγματα δεδομένων υψηλών διαστάσεων που είναι τόσο υπολογιστικά όσο και στατιστικά αποδοτικός. Αυτή η προσέγγιση βασίζεται στην αναλυτική εκτίμηση της συρρίκνωσης των πινάκων συνδιακύμανσης και μερικής συσχέτισης και σε ένα μοντέλο επιλογής δεδομένων με τη χρήση του ψευδού ποσοστού ανακάλυψης (False Discovery Rate-FDR). Ως εκ τούτου, το GeneNet περιλαμβάνει έναν υπολογιστικό

αλγόριθμο που αποφασίζει ποιες ακμές θα συμπεριληφθούν στο τελικό δίκτυο, γεγονός που εξαρτάται από τις σχετικές τιμές των συντελεστών μερικής συσχέτισης των ζευγών.

Για το σκοπό αυτό, **η συνάρτηση ggm.estimate.pcor()** προσφέρει μια διεπαφή σε έναν εκτιμητή συρρίκνωσης μερικής συσχέτισης που εφαρμόζεται στο πακέτο corpcor, η οποία είναι στατιστικά αποτελεσματική και μπορεί να χρησιμοποιηθεί για την ανάλυση μικρών δειγμάτων δεδομένων. Από προεπιλογή, επιλέγεται η μέθοδος="static", η οποία χρησιμοποιεί τη συνάρτηση pcor.shrink() του corpcor package.

Code GeneNet- ggm.estimate.pcor

```
1. library("GeneNet")
2.
3. data(ecoli)
4. inferred.pcor <- ggm.estimate.pcor(ecoli)
```

Κώδικας 10: GeneNet- ggm.estimate.pcor

2.5 Προσαρμοστική Συρρίκνωση (Adaptive Shrinkage) της μερικής συσχέτισης

Η εμπειρική προσέγγιση συρρίκνωσης Bayes η οποία μαθαίνει προσαρμοστικά πόσο να συρρικνώνει τις συσχετίσεις συνδυάζοντας πληροφορίες σε όλα τα ζεύγη μεταβλητών χρησιμοποιείται για την επίλυση που προκαλεί ο μικρός αριθμός δειγμάτων. Ένα βασικό χαρακτηριστικό της προσαρμοστικής συρρίκνωσης, που το διακρίνει από τις περισσότερες υπάρχουσες μεθόδους, είναι η ευελιξία του στην αντιμετώπιση δεδομένων που λείπουν.

Η προσαρμοστική συρρίκνωση της συσχέτισης ορίζεται ως εξής [15]:

Έστω $(X_{np})N\times P$ υποδηλώνει έναν πίνακα δεδομένων με *N* δείγματα και *P* μεταβλητές, όπου μερικές τιμές μπορεί να λείπουν (καταγράφονται ως NA). Για κάθε ζεύγος μεταβλητών $i, j \in \{1, 2, \cdots, P\}$ έστω το R_{ij} υποδηλώνει την (άγνωστη) αληθινή τους συσχέτιση και το $\hat{R}_{ij}$ τη συσχέτιση δείγματος που υπολογίζεται χρησιμοποιώντας μόνο τα δείγματα n που έχουν τιμές δείγματος και για τις δύο μεταβλητές *i* και *j*. Επιπλέον, τα Z_{ij} και $\hat{Z}_{ij}$ υποδηλώνουν τον αντίστοιχο Fisher Z-transforms:

$$Z_{ij} = Z(R_{ij}) = \frac{1}{2}\log\left(\frac{1+R_{ij}}{1-R_{ij}}\right)$$
$$\hat{Z}_{ij} = Z(\hat{R}_{ij}).$$

(23)

Προσαρμόζουμε το παραπάνω μοντέλο για να λάβουμε τον οπίσθιο μέσο όρο των Z_{ij}, $Z_{ij}{}^{*}$ δεδομένου του $\hat{R}_{ij}$:

$$Z_{ij}^{\star} := E\left[Z_{ij}|\hat{R}_{ij}\right] \quad \textbf{(24)}$$

$Z_{ij}{}^{*}$ είναι προσαρμοστικά συρρικνωμένες εκτιμήσεις του $\hat{Z}_{ij}$ που αντιπροσωπεύουν το n_{ij}, τον αριθμό των αντιστοιχισμένων δειγμάτων μεταξύ των μεταβλητών i και j. Όσο μικρότερο είναι το n_{ij}, τόσο υψηλότερο θα ήταν το επίπεδο συρρίκνωσης στο $\hat{Z}_{ij}$.

Στη συνέχεια, αντιστρέφουμε τον μετασχηματισμό $Z_{ij}{}^{*}$ σε εκτίμηση συσχέτισης $R_{ij}{}^{*}$:

$$R_{ij}^{\star} := \frac{exp(2Z_{ij}^{\star}) - 1}{exp(2Z_{ij}^{\star}) + 1}. \quad \textbf{(25)}$$

Η μερική συσχέτιση υπολογίζεται με την χρήσης της Εξίσωσης 3 με τη χρήση της προσαρμοστικής συρρίκνωσης της αντίστροφης συνδιακύμανσης που προκύπτει από την υλοποίηση του αλγόριθμο ISEE(Fan και Lv, 2016).

2.5.1 CorShrink package

Η **συνάρτηση pCorShrinkData()** του πακέτου CorShrink της R εκτελεί προσαρμοστική συρρίκνωση των αντίστροφων συνδιακυμάνσεων του δείγματος και η οποία χρησιμοποιείται για τον υπολογισμό της μερικής συσχέτισης ξεκινώντας από έναν πίνακα δεδομένων (δεν επιτρέπεται NA σε αντίθεση με το CorShrinkData). Η διαδικασία της συρρίκνωσης συνδυάζει τον υπολογισμό της αντίστροφης συνδιακύμανσης από τον αλγόριθμο ISEE του (Fan και Lv, 2016) με τη διατύπωση CorShrink. Ως είσοδο η συνάρτηση pCorShrinkData() δέχεται έναν πίνακα δεδομένων χωρίς τιμές NA.

Code CorShrink- pCorShrinkData

```
1. library("CorShrink")
2. library("GeneNet")
3.
4. data(ecoli)
5. part_Shrink<-pCorShrinkData(ecoli)
```

Κώδικας 11: CorShrink- pCorShrinkData

2.6 Μερική συσχέτιση χαμηλής τάξης (Low order partial correlation)

Η μερική συσχέτιση χαμηλής τάξης [16] χρησιμοποιείται για την επίλυση του προβλήματος που προκύπτει στην περίπτωση που το μέγεθος του δείγματος είναι μικρότερο από τον αριθμό των μεταβλητών. Σε σύγκριση με την κορεσμένη (πλήρους τάξης) μερική συσχέτιση, η μερική συσχέτιση χαμηλής τάξης μεταξύ δύο μεταβλητών λαμβάνεται μόνο υπό τον όρο ενός υποσυνόλου και όχι σε όλες τις μεταβλητές. Εάν ληφθούν υπόψη μόνο η μηδενική τάξη (συσχέτιση) και η μερική συσχέτιση πρώτης τάξης, το δίκτυο που προκύπτει ονομάζεται γράφημα «0-1». Η μερική συσχέτιση χαμηλής τάξης έχει το πλεονέκτημα ότι μπορεί να εκτιμηθεί με ακρίβεια και αποτελεσματικότητα από δεδομένα μικρού μεγέθους δείγματος και καταγράφει απλά μοτίβα ανεξαρτησίας υπό όρους. Οι de la Fuente et al. πρότειναν να υπολογιστεί η μερική συσχέτιση έως και δεύτερης τάξης για να ληφθούν υπόψη πιο περίπλοκα πρότυπα ανεξαρτησίας υπό όρους, διατηρώντας παράλληλα αποδεκτή την πολυπλοκότητα υπολογισμού. Παρά την απλότητά του, η μερική συσχέτιση χαμηλής τάξης μπορεί να χρησιμεύσει ως μια καλή προσέγγιση για να αντικατοπτρίζει τις σχέσεις ανεξαρτησίας υπό όρους μεταξύ των μεταβλητών στο δίκτυο.

Η μερική συσχέτιση χαμηλής τάξης (low order partial correlation) υπολογίζει τη μερική συσχέτιση από τη μηδενική τάξη (pearson correlation) μέχρι τη δεύτερη τάξη. Για ένα δεδομένο σύνολο δεδομένων με μεταβλητές *p*, υπολογίζει πρώτα τη μερική συσχέτιση μηδενικής τάξης (δηλαδή τη συσχέτιση pearson) και πρώτης τάξης για κάθε ζεύγος μεταβλητών. Η μερική συσχέτιση πρώτης τάξης ή συσχέτιση pearson υπολογίζεται όπως στην Εξίσωση 1.

Η μερική συσχέτιση πρώτης τάξης προκύπτει από τη συσχέτιση μηδενικής τάξης όπως στην Εξίσωση 2:

$$r_{p\,XY.\,W} = \frac{r_{\,XY} - r_{XW}\,r_{YW}}{\sqrt{(1-r_{XW}{}^2)(1-r_{YW}{}^2)}}$$

Όπου r_{XY} , r_{XW} , r_{YW} οι συντελεστές συσχέτισης που προέκυψαν από τον συντελεστή συσχέτισης Pearson.

Στη συνέχεια, η μερική συσχέτιση δεύτερης τάξης υπολογίζεται μόνο σε περιπτώσεις στις οποίες τόσο η μηδενική τάξη όσο και η μερική συσχέτιση πρώτης τάξης διαφέρουν σημαντικά από το μηδέν. Η μερική συσχέτιση δεύτερης τάξης προκύπτει από τη συσχέτιση πρώτης τάξης ως εξής:

$$r_{p\,XY.\,WZ} = \frac{r_{p\,XY.W} - r_{p\,XZ.W}\;r_{p\,YZ.W}}{\sqrt{\left(1-r_{p\,XZ.W}{}^2\right)\left(1-r_{p\,YZ.W}{}^2\right)}} \quad \textbf{(26)}$$

Όπου $r_{p\ XY.W}$, $r_{p\ XZ\ .W}$, $r_{p\ YZ\ .W}$ οι συντελεστές συσχέτισης που προέκυψαν από τη μερική συσχέτιση πρώτης τάξης.

2.6.1 RLowPC package

Η συνάρτηση LowPC() του RLowPC [17] package της R προτείνει έναν αποτελεσματικό και μαθηματικά ορθό αλγόριθμο για εξαγωγή βιολογικών δικτύων από δεδομένα υπολογίζοντας τη μερική συσχέτιση χαμηλής τάξης (Low order partial correlation). Ο αλγόριθμος είναι κατάλληλος για ένα σύνολο δεδομένων με μικρό μέγεθος δείγματος αλλά μεγάλο αριθμό μεταβλητών.

Η συνάρτηση υπολογίζει τη μερική συσχέτιση από τη μηδενική τάξη μέχρι τη δεύτερη τάξη. Υπολογίζει πρώτα τη μερική συσχέτιση μηδενικής τάξης. Οι σημαντικές ακμές της μερική συσχέτισης μηδενικής τάξης χρησιμοποιούνται για τον υπολογισμό της μερική συσχέτισης πρώτης τάξης. Ύστερα, οι σημαντικές ακμές της μερικής συσχέτισης πρώτης τάξης χρησιμοποιούνται για τον υπολογισμό της μερική συσχέτισης δεύτερης τάξης. Αυτό αυξάνει την απόδοση της συνάρτησης σε μεγάλο βαθμό, καθώς αποκλείει τα περισσότερα από τα πιθανά ζεύγη και βασίζεται μόνο στο υποσύνολο των ακμών που συνδέονται σημαντικά με τα γονίδια, πριν από τον υπολογισμό της μερικής συσχέτισης πρώτης και δεύτερης τάξης. Επιπλέον, χρησιμοποιεί τον μετασχηματισμό z του Fisher για να δημιουργήσει στατιστικά στοιχεία δοκιμής για να ορίσει ένα λογικό όριο (threshold). Για να λάβει υπόψη τις πολλαπλές δοκιμές, ελέγχει το ψευδό ποσοστό ανακάλυψης (False Discovery Rate - FDR) χρησιμοποιώντας τη διαδικασία Benjamini-Hochberg το οποίο υπολογίζεται με το fdrtool package της R. Τα αποτελέσματα της προσομοίωσης δείχνουν ότι το low order partial correlation λειτουργεί καλά κάτω από διάφορες συνθήκες που συνήθως παρατηρούνται σε πραγματικές εφαρμογές και οι ψευδείς ακμές (δηλαδή, ψευδώς θετικές) για το συμπαγόμενο δίκτυο μειώνονται σημαντικά.

Η συνάρτηση **LowPC(data.exp, cutoff = 0.05, cutat = "pval", method = "pearson", progressbar = T)** δέχεται ως είσοδο έναν πίνακα δεδομένων data.exp, έναν αριθμό αποκοπής cutoff τιμών p-value ή probability, μια συμβολοσειρά που υποδεικνύει με ποια μέθοδο πραγματοποιήθηκε η αποκοπή των αποτελεσμάτων, p-value "pval" ή probability"prob" και τέλος ένας χαρακτήρας συμβολοσειράς που ορίζει τη μέθοδο που θα χρησιμοποιηθεί για την εκτίμηση της συσχέτισης. Οι επιλογές είναι "pearson", «Spearman» και «kendall». Η συνάρτηση RLowPC() του RLowPC package της R υπολογίζει τον συντελεστή συσχέτισης αφαιρώντας την επίδραση ορισμένων περισσότερο επίφοβων γονιδίων.

Σε γενικές περιπτώσεις, η μερική συσχέτιση για ένα ζεύγος γονιδίων υπολογίζεται αφαιρώντας την επίδραση των υπόλοιπων γονιδίων που επηρεάζουν το υπό διερεύνηση ζεύγος γονιδίων. Ωστόσο, μπορεί να υπάρχει ένας αριθμός γονιδίων που δεν επηρεάζουν το ζεύγος γονιδίων. Σε αυτό στηρίχθηκε η μέθοδος RLowPC [18] η οποία χρησιμοποιείται για να βελτιώσει την δομή του δικτύου μειώνοντας τις έμμεσες ακμές που έχουν προβλεφθεί ως άμεσες. Η μέθοδος RLowPC ακολουθεί τα παρακάτω 2 βήματα:

Βήμα 1: Εισαγωγή ενός κατάλληλου σε μέγεθος δικτύου, το οποίο έχει περιθώρια βελτίωσης, για αναζήτηση έμμεσων άκρων. Για παράδειγμα, οι κορυφαίες σταθμισμένες ακμές ενός δικτύου μερικής συσχέτισης μπορεί να χρησιμοποιηθούν ως χώρος αναζήτησης. Κάθε ζεύγος γονιδίων θεωρείται ότι συνδέεται με τα πιο σχετικά γειτονικά γονίδια.

Βήμα 2: Υπολογισμός της μερικής συσχέτισης χαμηλής τάξης. Για κάθε ζεύγος γονιδίων που συνδέεται με μια ακμή στον χώρο αναζήτησης, το βάρος της ακμής επαναπροσδιορίζεται ως (α) μερική συσχέτιση μηδενικής τάξης εάν το ζεύγος γονιδίων δεν συνδέεται με το ίδιο σύνολο γειτονικών γονιδίων, (β) μερική συσχέτιση αφαιρώντας όλους τους κοινούς γείτονες ταυτόχρονα και (γ) συρρίκνωση της μερικής συσχέτισης εάν ο πίνακας συνδιακύμανσης που χρησιμοποιείται για την εκτίμηση της μερικής συσχέτισης στο (β) δεν είναι θετικός ορισμένος ή αντιστρέψιμος. Εάν ο χώρος αναζήτησης είναι πολύ μεγάλος, ενδέχεται να υπάρχει ένας αριθμός άσχετων στοιχείων ελέγχου που εμπλέκονται στη διαδικασία συρρίκνωσης του υπολογιστή στο (γ). Μια εναλλακτική είναι (δ) η διαγραφή των λιγότερο συνδεδεμένων γειτονικών γονιδίων έως ότου ο πίνακας συνδιακύμανσης στο (β) να είναι θετικός ορισμένος και αναστρέψιμος.

Η **συνάρτηση RLowPC(data.exp, edgelist, method = "pearson", pc.estimator = "shrink", progressbar = T)** δέχεται ως είσοδο έναν πίνακα δεδομένων data.exp, μια λίστα ακμών edgelist η οποία περιέχει τις πιο σημαντικές ακμές του δικτύου, έναν χαρακτήρας συμβολοσειράς method που ορίζει τη μέθοδο που θα χρησιμοποιηθεί για την εκτίμηση της συσχέτισης (οι επιλογές είναι "pearson", «Spearman» και «kendall») και μια συμβολοσειρά χαρακτήρων που υποδεικνύει τον εκτιμητή που χρησιμοποιείται για τον υπολογισμό της μερικής συσχέτισης για κάθε ζεύγος των κόμβων στη λίστα ακμών.

Code LowPC

```
1. library("RLowPC")
2. library("data.table")¨
3. library("GeneNet")
4.
5. ecoli<- data.table(data(ecoli))
6. loworder<-LowPC(ecoli, cutoff = 0.05)
```

Κώδικας 12: LowPC

Code RLowPC

```
##load library
library("RLowPC")
library("corpcor")
library("GeneNet")

##load data
data(ecoli)
data.exp<-ecoli
genes<-colnames(ecoli)

 ##filter low expressed genes
 data2anova<-
data.frame(time=factor(paste0(data.exp$experiment,'_',data.exp
$time)),data.exp[,-c(1:3)])
 data.new<-anova2de(data.exp = data2anova,ncol.idx =1,model =
'expression~time',pval.cut = 0.01)
 data.exp<-data.new$de.ts
 genes<-data.new$de.gene
 ref.adj<-ref.adj[genes,genes]


 ##infer PC network
 inf.pcor<-
abs(pcor.shrink(data.exp)[1:length(genes),1:length(genes)])
 diag(inf.pcor)<-0

 ##inf RLowPC

 reduction.sapce<-na.omit(adjmatrix2edgelist(adjmatrix =
inf.pcor,directed = F,order = T))
 inf.RLowPC.edge<-RLowPC(data.exp = data.exp,edgelist =
reduction.sapce, method = 'pearson',pc.estimator = 'shrink')
```

Κώδικας 13: RLowPC

2.7 Στατιστική Σημαντικότητα

Στην περίπτωση ενός τυχαίου δείγματος μεγέθους n ενός πολυποικιλιακού κανονικού πληθυσμού μπορούμε να εξετάσουμε τη μηδενική υπόθεση (ότι μια μερική συσχέτιση είναι ίση με μηδέν) έναντι της εναλλακτικής (ότι δεν είναι ίση με μηδέν). Αυτό εκφράζεται παρακάτω:

H0: $r_{p\,ij} = 0$ έναντι H1: $r_{p\,ij} \neq 0$

Έτσι μπορούμε να εφαρμόσουμε μια στατιστική δοκιμή για τη μερική συσχέτιση παρόμοια με αυτή που χρησιμοποιείται για μια συνηθισμένη συσχέτιση. Αυτό το στατιστικό τεστ φαίνεται παρακάτω:

$$\mathrm{t} = \mathrm{r_{p\,ij}} \sqrt{\frac{\mathrm{n-k-1}}{1-\mathrm{r_{p\,ij}}^{2}}} \quad \textbf{(27)}$$

όπου $k = n - p - 1$ οι βαθμοί ελευθερίας της μερικής συσχέτισης που προσδιορίζονται από το μέγεθος του δείγματος n και από τον αριθμό των μεταβλητών p από τις οποίες εξαρτώνται.

Για τον προσδιορισμό του p-value, είναι απαραίτητο να υπάρχει γνώση της κατανομής της στατιστικής δοκιμής με την παραδοχή ότι η μηδενική υπόθεση είναι αληθής. Το p-value μπορεί να υπολογιστεί από το t-score με τη χρήση της συνάρτησης αθροιστικής κατανομής της κατανομής t-Student με k βαθμούς ελευθερίας (cdft,k). Η συνάρτηση αυτή ορίζεται ως εξής:

$$\mathrm{cdft,k} = F(\,x \mid k\,) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\,\pi}} \frac{1}{\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}}\ dt \quad \textbf{(28)}$$

όπου k είναι οι βαθμοί ελευθερίας και Γ(·) η συνάρτηση Gamma. Το αποτέλεσμα είναι η πιθανότητα μια μεμονωμένη παρατήρηση από την κατανομή t με k βαθμούς ελευθερίας να πέσει στο διάστημα [–∞, x].

Έτσι με τη χρήση του two-tailed t-test τα p-value υπολογίζονται ως εξής:

$$p - value = 2 \cdot cdft, k(-|tscore|) \quad \textbf{(29)}$$

Ακόμη, ο υπολογισμός του FDR μπορεί να γίνει μέσω του αλγόριθμου του Benjamini Hochberg. Τα βήματα του αλγόριθμου είναι τα εξής:

1. Υπολογισμός των p-values για όλα τα tests
2. Ταξινόμηση των p-values από το μικρότερο στο μεγαλύτερο
3. Υπολογισμός των q-values ως q= συνολικός αριθμός των tests * p-value/ τάξη του p-value
4. Η τιμή του FDR για κάθε test είναι το ελάχιστο των q-values όλων των test με μεγαλύτερη τάξη από αυτό.

2.7.1 Υπολογισμός p-value και FDR της μερικής συσχέτισης μέσω των βαθμών ελευθερίας

Για τον υπολογισμός των p-value και FDR των συντελεστών μερικής συσχέτισης υπολογίζεται αρχικά το t-test σύμφωνα με την Εξίσωση 27. Τα p-value υπολογίζονται με τη μέθοδο two tailed με την Εξίσωση 29 μέσω της συνάρτησης

pt(q, df, lower.tail = FALSE) του πακέτου stats η οποία στην είσοδο της λαμβάνει το t-test στη μεταβλητή q, τους βαθμούς ελευθερίας στη μεταβλητή df και το lower.tail=FALSE για τον καθορισμό του two tail. Στη συνέχεια, το FDR υπολογίζεται μέσω του αλγορίθμου του Benjamini Hochberg με τη χρήση της συνάρτησης **p.fdr(pvalues)** του πακέτου FDRestimation.

2.7.2 Fdrtool package

Η στατιστική σημαντικότητα της μερικής συσχέτισης μπορεί να υπολογιστεί μέσω της συνάρτησης fdrtool() του πακέτου Fdrtool [19] της R. Η **συνάρτηση fdrtool(x, statistic=c("normal", "correlation", "pvalue"), plot=TRUE, color.figure=TRUE, verbose=TRUE, cutoff.method=c("fndr", "pct0", "locfdr"), pct0=0.75)** δέχεται ως είσοδο έναν πίνακα δεδομένων x (στην περίπτωση μας ένα πίνακα μερικής συσχέτισης) και ένα όρισμα statistic, που δέχεται ένα από τα "normal", "correlation", "pvalue", το οποίο προσδιορίζει τον τύπο δεδομένων x (στην περίπτωση μας "correlation"). Ακόμη η συνάρτηση δέχεται ως είσοδο ένα cutoff.method το οποίο καθορίζει ένα κατάλληλο σημείο αποκοπής. Εάν η μέθοδος cutoff είναι "fndr", τότε πρώτα τοποθετείται ένα κατά προσέγγιση μηδενικό μοντέλο και στη συνέχεια αναζητείται ένα σημείο αποκοπής με όσο το δυνατόν μικρότερο ποσοστό ψευδούς μη ανακάλυψης. Εάν το cutoff.method είναι "pct0", τότε ως σημείο αποκοπής χρησιμοποιείται ένα καθορισμένο ποσό (προεπιλεγμένη τιμή: 0,75) των δεδομένων. Εάν το cutoff.method ισούται με "locfdr", τότε χρησιμοποιείται η ευρετική του πακέτου "locfdr" για να βρεθεί η αποκοπή (μόνο για z-score και συσχετίσεις).

2.8 Συγκεντρωτικός πίνακας μεθόδων και πακέτων

Συνάρτηση	Πακέτο	Μέθοδος	Είσοδος	Μέγιστος αριθμός μεταβλητών	Υπολογιστικό Κόστος
Cor2pcor	Corpcor	Υπολογίζει τον ψευδοαντίστροφο πίνακα για να υπολογίσει τον πίνακα μερικής συσχέτισης	Πίνακας συσχέτισης ή συνδιακύμανσης	10.000	7 ώρες και 56 λεπτά
Pcor.shrink	Corpcor	Υπολογίζει μια εκτίμηση συρρίκνωσης (shrinkage estimates) της μερικής συσχέτισης	Πίνακας δεδομένων	24.000	49.14 δευτερόλεπτα
pCorShrinkData	CorShrink	Υπολογίζει την μερική συχέτιση από τη προσαρμοστική συρρίκνωση του πίνακα αντίστροφης συνδιακύμανσης	Πίνακας δεδομένων	3.500	41.48 λεπτά
ggm.estimate.pcor	GeNet	Υπολογίζει μια εκτίμηση συρρίκνωσης (shrinkage estimates) της μερικής συσχέτισης μέσω της συνάρτησης pcor.shrink	Πίνακας δεδομένων	22.000	52.09 δευτερόλεπτα
glasso	Glasso	Υπολογίζει την μερική συχέτιση από την εκτίμηση ενός αραιού πίνακα αντίστροφης συνδιακύμανσης χρησιμοποιώντας ποινή L1	Πίνακας συνδιακύμανσης	1.000	6 ώρες και 28.4 λεπτά
LowPC	RLowPC	Υπολογίζει τη μερική συσχέτιση χαμηλής τάξης	Πίνακας δεδομένων	1.000	1 ώρα και 12 λεπτά
adalasso.net	Parcor	Υπολογίζει τον πίνακα μερικής συσχέτισης βάση μιας εκτίμηση των αντίστοιχων μοντέλων παλινδρόμησης μέσω lasso και προσαρμοστικού (adaptive) lasso αντίστοιχα	Πίνακας δεδομένων	15.000	3 ώρες και 13.7 λεπτά
pls.net	Parcor	Υπολογίζει τον πίνακα μερικής συσχέτισης μέσω εκτίμησης των αντίστοιχων μοντέλων παλινδρόμησης μέσω Μερικών Ελάχιστων Τετράγωνων (Partial Least Squares)	Πίνακας δεδομένων	15.000	5 ώρες και 15 λεπτά
ridge.net	Parcor	υπολογίζει τον πίνακα μερικής συσχέτισης μέσω εκτίμησης των αντίστοιχων μοντέλων παλινδρόμησης μέσω της παλινδρόμησης Ridge (Ridge Regression)	Πίνακας δεδομένων	20.000	27 ώρες και 12 λεπτά
space.joint	SPACE	Εκτίμηση μερικής συσχέτισης με joint sparse regression model	Πίνακας δεδομένων	16.000	5 ώρες και 38 λεπτά
space.neighbor	SPACE	Εκτίμηση μερικών συσχετίσεων χρησιμοποιώντας την προσέγγιση επιλογής γειτονιάς (neighborhood selection approach)	Πίνακας δεδομένων	24.000	39.93 δευτερόλεπτα

Πίνακας 1: Συγκεντρωτικός πίνακας μεθόδων και πακέτων

3 Μετα- ανάλυση (Meta- analysis)

Η μετα- ανάλυση είναι μια στατιστική ανάλυση για την συγχώνευση των αποτελεσμάτων πολλαπλών επιστημονικών μελετών. Το πρώτο βήμα της μετα-ανάλυσης αφορά τη συλλογή των μελετών και την εξαγωγή των δεδομένων που θα χρησιμοποιηθούν στη μετα- ανάλυση. Ύστερα, ακολουθεί ο καθορισμός του μεγέθους επίδρασης το οποίο μπορεί να είναι η διαφορά μέσων τιμών, ο συντελεστής συσχέτισης ή το Odds Ratio. Τέλος, πραγματοποιείται η επιλογή ενός εκ των δύο μοντέλων που χρησιμοποιούνται για το συνδυασμό των εκτιμήσεων του μεγέθους επίδρασης των αρχικών μελετών. Τα μοντέλα αυτά είναι το μοντέλο των σταθερών επιδράσεων (fixed effect model) και το μοντέλο των τυχαίων επιδράσεων (random effect model).

Στη παρούσα πτυχιακή εργασία προτείνεται η χρήση του μερικού μεγέθους επίδρασης (partial effect sizes). Το μερικό μέγεθος επίδρασης είναι ένας δείκτης που περιγράφει το μέγεθος μιας επίδρασης μετά τον έλεγχο επίδρασης άλλων μεταβλητών στο μοντέλο και υπολογίζεται μέσω της μερικής συσχέτισης. Η χρήση του είναι ωφέλιμη στη μετα- ανάλυση πολλών μεταβλητών καθώς το μερικό μέγεθος επίδρασης μπορεί να υπολογιστεί για ένα πολύ μεγαλύτερο σύνολο εκτιμήσεων.

3.1 Το μοντέλο σταθερών επιδράσεων (fixed effect model)

Το μοντέλο των σταθερών επιδράσεων μετατρέπει αρχικά τους συντελεστές μερικής συσχέτισης $r_{p\,ij}$ κάθε μεταβλητής κάθε μελέτης σε μια τυπική κανονική μέτρηση χρησιμοποιώντας Fisher's R_p *-to-Z* μετασχηματισμό πριν από τον υπολογισμό ενός σταθμισμένου μέσου όρου αυτών των μετασχηματισμένων μερικών συσχετίσεων.[20]

Ο μετασχηματισμός Fisher's R_p*-to-Z* δίνεται παρακάτω:

$$Z_{R\,p\,i} = \frac{1}{2}\ log_e \frac{(1+R_{p\,i})}{(1-R_{p\,i})} \quad \textbf{(30)}$$

Όπου $R_{p\,i}$ είναι ο συντελεστής μερικής συσχέτισης της μελέτης *i*.

Τα μετασχηματισμένα μεγέθη επιδράσεων χρησιμοποιούνται στη συνέχεια για τον υπολογισμό ενός αρχικού μέσου όρου στον οποίο κάθε συσχέτιση σταθμίζεται με το αντίστροφο της διακύμανσης εντός της μελέτης από την οποία προήλθε. Ο μέσος όρος ορίζεται ως εξής:

$$\bar{Z}_{R\,p} = \frac{\sum_{i=1}^{k} w_i\, Z_{R_p\,i}}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} (n_i - 3)\, Z_{R_p\,i}}{\sum_{i=1}^{k} (n_i - 3)}$$, **(31)**

όπου *ni* το μέγεθος του δείγματος της κάθε μελέτης και k ο αριθμός των μελετών που παίρνουν μέρος στη μετα-ανάλυση

Στη συνέχεια, ο μέσος όρος μερικής συσχέτισης που προκύπτει από τους παραπάνω υπολογισμούς μετατρέπεται και επιστρέφει πίσω στη μέτρηση των R_p μέσω του αντίστροφου του μετασχηματισμού Fisher's (μετασχηματισμός *Z* σε R_p) σύμφωνα με την παρακάτω εξίσωση:

$$\overline{R_p} = \frac{e^{\left(2\,\bar{Z}_{R_p}\right)} - 1}{e^{\left(2\,\bar{Z}_{R_p}\right)} + 1} \quad \textbf{(32)}$$

Αυτός ο μέσος όρος χρησιμοποιείται στη συνέχεια για τον υπολογισμό μιας δοκιμής της ομοιογένειας των μερικών συσχετίσεων: Χρησιμοποιείται η τετραγωνική διαφορά μεταξύ του παρατηρούμενου μετασχηματισμένου R_p της κάθε μελέτης και του μέσου μετασχηματισμένου R_p, σταθμισμένη με τη διακύμανση εντός της μελέτης. Αυτό μας δίνει τη στατιστική Q, η οποία έχει κατανομή chi-τετράγωνο με *k – 1* βαθμούς ελευθερίας κάτω από τη μηδενική υπόθεση των ομοιογενών μεγεθών επίδρασης. Η στατιστική Q υπολογίζεται από την παρακάτω εξίσωση:

$$Q = \sum_{i=1}^{k} (n_i - 3)\left(Z_{R_p\,i} - \bar{Z}_{R_p}\right)^2 \quad \textbf{(33)}$$

3.2 Το μοντέλο τυχαίων επιδράσεων (random effect model)

Το μοντέλο των τυχαίων επιδράσεων μετατρέπει αρχικά τους συντελεστές μερικής συσχέτισης $r_{p\,ij}$ κάθε μεταβλητής κάθε μελέτης σε μια τυπική κανονική μέτρηση χρησιμοποιώντας Fisher's R_p *-to-Z* μετασχηματισμό πριν από τον υπολογισμό ενός σταθμισμένου μέσου όρου αυτών των μετασχηματισμένων μερικών συσχετίσεων. Ο μετασχηματισμός Fisher's R_p*-to-Z* υπολογίζεται με την χρήση της Εξίσωσης 30.

Για τον υπολογισμό ενός σταθμισμένου μέσου των μετασχηματισμένων μερικών συσχετίσεων, τα βάρη χρησιμοποιούν μια συνιστώσα διακύμανσης που ενσωματώνει τόσο τη διακύμανση μεταξύ των μελετών όσο και τη διακύμανση εντός των μελετών. Η διακύμανση μεταξύ των μελετών συμβολίζεται με τ^2 και μια εκτίμηση της ($\hat{\tau}^2$) προστίθεται στη διακύμανση εντός της μελέτης για τον υπολογισμό του βάρους.

Η διακύμανση μεταξύ των μελετών μπορεί να εκτιμηθεί με τη χρήση της στατιστικής Q που υπολογίζεται μέσω της Εξίσωσης 33, του k (αριθμός μελετών) και της σταθερά c ως εξής:

$$\hat{\tau}^2 = \frac{Q - (k-1)}{c}, \quad \textbf{(34)}$$

όπου η σταθερά c ορίζεται ως:

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} (w_i)^2}{\sum_{i=1}^{k} w_i}, \quad \textbf{(35)}$$

με $w_i = n_i - 3$.

Στην περίπτωση που η εκτίμηση της διακύμανσης μεταξύ των μελετών έχει αρνητική τιμή παίρνει την τιμή μηδέν.

Στη συνέχεια ο σταθμισμένος μέσος όρος των μερικών συσχετίσεων όλων των μελετών υπολογίζεται μέσω της ακόλουθης Εξίσωσης:

$$\overline{Z^*}_{R_p} = \frac{\sum_{i=1}^{k} w^*_i \, Z_{R_p i}}{\sum_{i=1}^{k} w^*_i} \quad \textbf{(36)}$$

όπου το βάρος (w_i^*) ορίζεται ως:

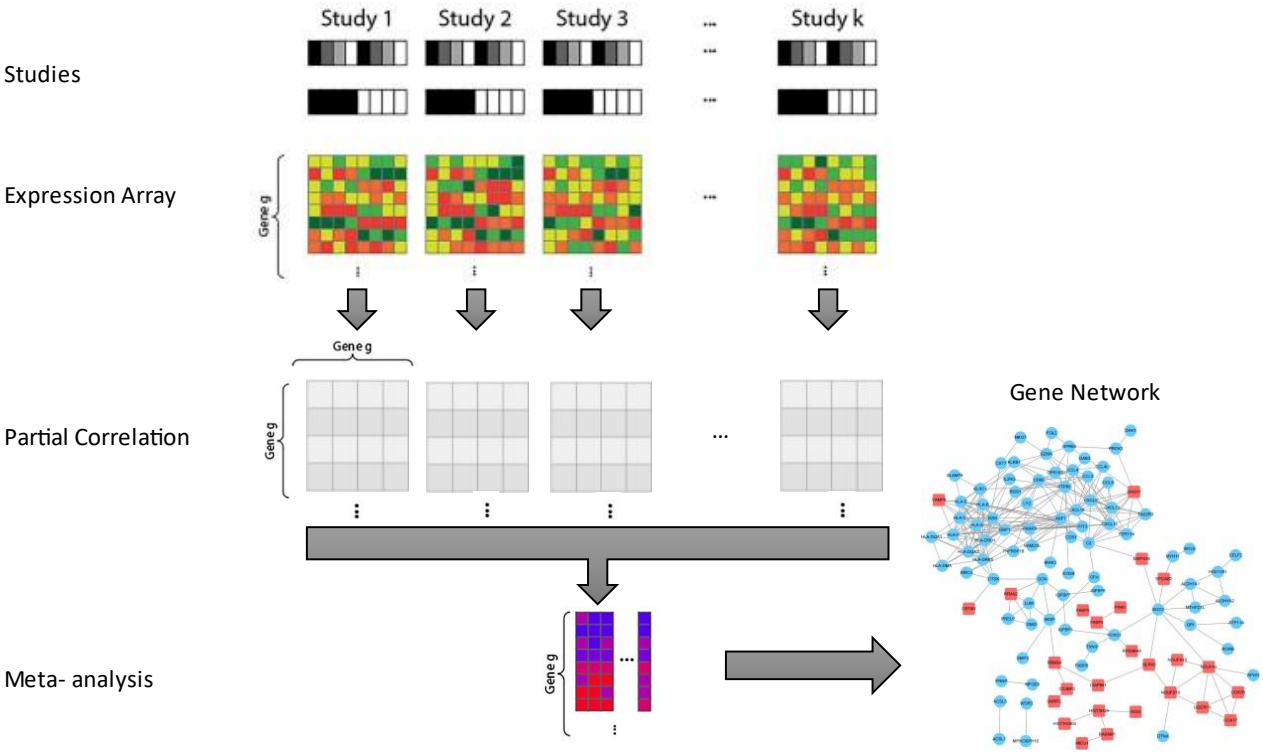
$$w_i^* = \left(\frac{1}{n_i - 3} + \hat{\tau}^2 \right)^{-1}. \quad \textbf{(37)}$$

Τέλος, ο μέσος όρος μερικής συσχέτισης που προκύπτει από τους παραπάνω υπολογισμούς μετατρέπεται και επιστρέφει πίσω στη μέτρηση των R_p μέσω του αντίστροφου του μετασχηματισμού Fisher's (μετασχηματισμός Z σε R_p) μέσω της Εξίσωσης 32.

3.3 Υλοποίηση της μετα-ανάλυσης με τη χρήση της μερικής συσχέτισης ως μέγεθος επίδρασης

Οι μέθοδοι που βρέθηκαν να είναι κατάλληλες για τον υπολογισμό του μερικού μεγέθους επίδρασης είναι η μέθοδος υπολογισμού της μερικής συσχέτισης μέσω συρρίκνωσης της συνδιακύμανσης (shrinkage estimates), η μέθοδος υπολογισμού της μερικής συσχέτισης μέσω της προσαρμοστικής συρρίκνωσης του πίνακα αντίστροφης συνδιακύμανσης και η μέθοδος αραιής εκτίμησης των μερικών συσχετίσεων χρησιμοποιώντας την προσέγγιση επιλογής γειτονιάς (neighborhood selection approach). Τα οφέλη αυτών των μεθόδων είναι το μικρό υπολογιστικό τους κόστος και το γεγονός ότι δέχονται ως είσοδο μεγάλο πλήθος παρατηρήσεων.

Το πρώτο βήμα για την υλοποίηση της μετα-ανάλυσης με τη χρήση της μερικής συσχέτισης αφορά τη συλλογή των μελετών που περιέχουν δεδομένα γονιδιακής έκφρασης (Expression Array) και προήλθαν από RNA-seq ή Μικροσυστοιχίες. Ύστερα, ακολουθεί η εφαρμογή της μερικής συσχέτισης στα δεδομένα γονιδιακής έκφρασης κάθε μελέτης και η υλοποίηση της μετα-ανάλυσης με τη χρήση της μερικής συσχέτισης ως μέγεθος επίδρασης. Τέλος, τα συγχωνευμένα αποτελέσματα των μελετών χρησιμοποιούνται για την κατασκευή δικτύων γονιδίων. Η παραπάνω μεθοδολογία απεικονίζεται παρακάτω:



Εικόνα 1: Μεθοδολογία Μετα- ανάλυσης με μερική συσχέτιση

Code Meta-analysis - corpcor

```
1. library(GEOquery)
2. library(corpcor)
3. library(space)
4. library(data.table)
5. library(GeneNet)
6. library(meta)
7. library(metafor)
8. library(dplyr)
9.
10.  #you can put any GEO GSE
11.  gse <- getGEO("GSE14764", GSEMatrix = TRUE)
12.  gse<-
   data.table(t(gse[["GSE14764_series_matrix.txt.gz"]]@assayData[
   ["exprs"]]))
13.
14.  gse2 <- getGEO("GSE26712", GSEMatrix = TRUE)
15.  gse2<-
   data.table(t(gse2[["GSE26712_series_matrix.txt.gz"]]@assayData
   [["exprs"]]))
16.
17.  #corpcor -> partial correlation
18.  pcor1<-pcor.shrink(gse)
19.  pcor1<-pcor1[1:nrow(pcor1),1:nrow(pcor1)]
20.  pcor1<-melt(pcor1)
21.  pcor1<-pcor1[which(pcor1$value!=1),]
22.  names(pcor1)<-c("from","to", "ri")
23.  pcor1$ni<-nrow(gse)
24.
25.  pcor2<-pcor.shrink(gse2)
26.  pcor2<-pcor2[1:nrow(pcor2),1:nrow(pcor2)]
27.  pcor2<-melt(pcor2)
28.  pcor2<-pcor2[which(pcor2$value!=1),]
29.  names(pcor2)<-c("from","to", "ri")
30.  pcor2$ni<-nrow(gse2)
31.
32.  #meta analysis
33.  correlations<-full_join(pcor1, pcor2, by=c("from", "to"))
34.  meta_cor<-correlations[,1:2]
35.
36.  for(i in 1:nrow(correlations)){
37.
38.    m1 <- metacor(as.numeric(correlations[ i, grepl( "ri" ,
   names( correlations) ) ]), as.numeric(correlations[ i, grepl(
   "ni" , names( correlations) ) ]),sm="ZCOR",method.tau = "HE")
39.    meta_cor$cor.fe[i]=transf.ztor(m1[["TE.fixed"]])
40.    meta_cor$cor.re[i]=transf.ztor(m1[["TE.random"]])
41.  }
42.
```

Κώδικας 14: Meta-analysis - corpcor

Code Meta-analysis - space

```
1. library(GEOquery)
2. library(corpcor)
3. library(space)
4. library(data.table)
5. library(GeneNet)
6. library(meta)
7. library(metafor)
8. library(dplyr)
9.
10.  #you can put any GEO GSE
11.  gse <- getGEO("GSE14764", GSEMatrix = TRUE)
12.  gse<-
   data.table(t(gse[["GSE14764_series_matrix.txt.gz"]]@assayData[
   ["exprs"]]))
13.  gse2 <- getGEO("GSE26712", GSEMatrix = TRUE)
14.  gse2<-
   data.table(t(gse2[["GSE26712_series_matrix.txt.gz"]]@assayData
   [["exprs"]]))
15.
16.  #space -> partial correlation
17.  n= nrow(gse)
18.  p= ncol(gse)
19.  alpha=1
20.  l1=1/sqrt(n)*qnorm(1-alpha/(2*p^2))
21.
22.  pcor1<-space.neighbor(gse, lam1=l1*n*1.56, lam2=0)$ParCor
23.  pcor1<-melt(pcor1)
24.  pcor1<-pcor1[which(pcor1$value!=1),]
25.  names(pcor1)<-c("from","to", "ri")
26.  pcor1$ni<-nrow(gse)
27.
28.  pcor2<-space.neighbor(gse2, lam1=l1*n*1.56, lam2=0)$ParCor
29.  pcor2<-melt(pcor2)
30.  pcor2<-pcor2[which(pcor2$value!=1),]
31.  names(pcor2)<-c("from","to", "ri")
32.  pcor2$ni<-nrow(gse2)
33.
34.  #meta analysis
35.  correlations<-full_join(pcor1, pcor2, by=c("from", "to"))
36.  meta_cor<-correlations[,1:2]
37.
38.  for(i in 1:nrow(correlations)){
39.    m1 <- metacor(as.numeric(correlations[ i, grepl( "ri" ,
   names( correlations) ) ]), as.numeric(correlations[ i, grepl(
   "ni" , names( correlations) ) ]),sm="ZCOR",method.tau = "HE")
40.    meta_cor$cor.fe[i]=transf.ztor(m1[["TE.fixed"]])
41.    meta_cor$cor.re[i]=transf.ztor(m1[["TE.random"]])
42.  }
```

Κώδικας 15: Meta-analysis - space

4 Δεδομένα

Η διαθεσιμότητα ηλεκτρονικών αρχείων ασθενών διευκολύνει τις μελέτες σχετικά με τη συννοσηρότητα μιας νόσου. Το παραπάνω γεγονός υποδεικνύει την πιθανότητα συνεμφάνισης δύο δεδομένων ασθενειών στο ίδιο άτομο. Η συννοσηρότητα μπορεί να θεωρηθεί ως ένας τύπος συσχέτισης ασθενειών που προκύπτει από τον ηλεκτρονικό ιατρικό φάκελο. Η συννοσηρότητα και η συσχέτιση της με άλλους τύπους συσχετίσεων ασθενειών, όπως γενετικές συσχετίσεις και εξελικτικές συσχετίσεις έχουν μελετηθεί στο παρελθόν. Στην παρούσα μελέτη χρησιμοποιήθηκαν δεδομένα συννοσηρότητας για την αξιολόγηση των συσχετίσεων ασθενειών που προήλθαν από προβλέψεις με τη χρήση μετρήσεων ομοιότητας. Οι συσχετίσεις συννοσηρότητας λήφθηκαν από το Δίκτυο Ανθρώπινων Νοσημάτων (Human Disease Network, HuDiNe)[21], οι οποίες ελήφθησαν από το ιστορικό ασθένειας 32 εκατομμυρίων Αμερικανών ασθενών. Στο HuDiNe οι ασθένειες σχολιάζονται με τη χρήση κωδικών ICD-9 (η αντιστοίχιση των κωδικών ICD-9 με τις ασθένειες παρατίθεται στο κεφάλαιο Παράρτημα της πτυχιακής). Στην παρούσα μελέτη χρησιμοποιήθηκαν τα δεδομένα συννοσηρότητας που σχολιάστηκαν με τη χρήση κωδικών ICD-9 τριψήφιου επιπέδου 13039018 ατόμων. Η ισχύς της συσχέτισης συννοσηρότητας μεταξύ ενός ζεύγους ασθενειών μπορεί να μετρηθεί με τον σχετικό κίνδυνο (Relative Risk) και τη συσχέτιση φ (phi-correlation). Η συσχέτιση φ επιλέχθηκε ως το μέτρο της συννοσηρότητας καθώς οι συσχετισμοί συννοσηρότητας που ποσοτικοποιήθηκαν με φ-συσχέτιση αναφέρθηκε ότι περιέχουν περισσότερες συνδέσεις σε διαφορετικές κατηγορίες ICD-9. Η φ-συσχέτιση μεταξύ *Di* και *Dj* ορίστηκε ως η συσχέτιση του Pearson για δυαδικές μεταβλητές μέσω της ακόλουθης εξίσως:

$$\phi_{ij} = \frac{C_{ij}N - P_iP_j}{\sqrt{P_iP_j(N-P_i)(N-P_j)}} \quad \textbf{(38)}$$

όπου *Cij* ο αριθμός των ατόμων-ασθενών που επηρεάζονται τόσο από την ασθένεια *Di* όσο και από την ασθένεια *Dj*, N ο συνολικός αριθμός ατόμων-ασθενών στον πληθυσμό και *Pi* , *Pj* οι επιπολασμοί των *Di* και *Dj* ασθενειών αντίστοιχα. Μια φ-συσχέτιση υψηλότερη από 0 υποδηλώνει ότι η συνύπαρξη των *Di* και *Dj* είναι πιο συχνά από την αναμενόμενη τυχαία. Η στατιστική σημασία της φ-συσχέτισης προσδιορίστηκε χρησιμοποιώντας ένα t-test:

$$t = \frac{\phi\sqrt{n-2}}{\sqrt{1-\phi^2}} \quad \textbf{(39)}$$

όπου *n=max(Pi,Pj)<<N* είναι ο αριθμός των παρατηρήσεων που χρησιμοποιήθηκαν για τον υπολογισμό του φ. Το *n* αντιπροσωπεύει τον πιο αυστηρό τρόπο με τον οποίο μπορεί να υπολογιστεί το t βάσει των δεδομένων μας, καθώς η χρήση *n=N* θα παράγει μεγαλύτερο αριθμό σημαντικών συσχετίσεων, οι περισσότερες από τις οποίες δεν θα είναι απαραίτητα ισχυροί προγνωστικοί παράγοντες. Χρησιμοποιήθηκαν σημαντικές συσχετίσεις σε επίπεδο 5% (t≥1,96) για τις αναλύσεις.

Τα αρχεία που παρέχονται από το HuDiNe (Data from HuDiNe (Hidalgo et al.) for the analysis of comorbidities (upf.edu)) είναι τα εξής:

Αρχείο	Ασθενείς	Αριθμός συνδέσμων (ακμών)
AllNet5.net	All patients, ICD9 5 digit level	6088553
AllNet3.net	All patients, ICD9 3 digit level	291172

Πίνακας 2: Αρχεία Hudine

Και έχουν την παρακάτω μορφή:

PDN File Structure	
Στήλη	**Περιγραφή**
1	Icd9 κωδικός ασθένεια 1
2	Icd9 κωδικός ασθένεια 2
3	Επιπολασμός ασθένειας 1
4	Επιπολασμός ασθένειας 2
5	Ταυτόχρονη εμφάνιση ασθενειών 1 και 2
6	Σχετικός κίνδυνος (Relative risk)
7	Σχετικός κίνδυνος 99% διάστημα εμπιστοσύνης (αριστερό όριο)
8	Σχετικός κίνδυνος 99% διάστημα εμπιστοσύνης (δεξιό όριο)
9	*Φ-συσχέτιση (ϕ*-correlation)
10	Τιμή t-test

Πίνακας 3: Μορφή αρχείων Hudine

Οι κωδικοί ICD9 του Hudine αντιστοιχούν στις παρακάτω κλάσεις:

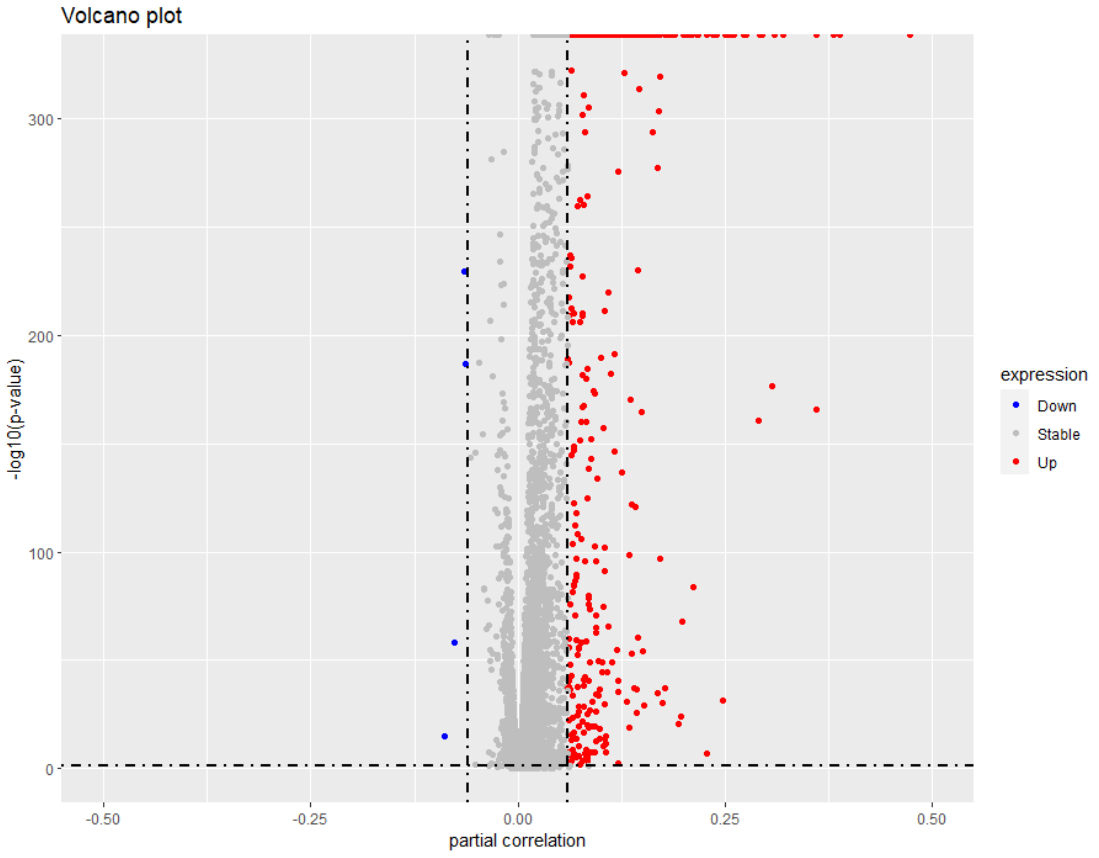
Χρώμα	Κωδικοί icd9	Κλάση ασθενειών
I	001–139	Infectious and Parasitic Diseases
II	140–239	Neoplasms
III	240–279	Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders
IV	280–289	Diseases of the Blood and Blood-forming Organs
V	290–319	Mental Disorders
VI	320–389	Diseases of the Nervous System and Sense Organs
VII	390–459	Diseases of the Circulatory System
VIII	460–519	Diseases of the Respiratory System
IX	520–579	Diseases of the Digestive System
X	580-629	Diseases of the Genitourinary System
XI	630-679	Complications of Pregnancy, Childbirth, and the Puerperium
XII	680-709	Diseases of the Skin and Subcutaneous Tissue
XIII	710–739	Diseases of the Musculoskeletal System and Connective Tissue
XIV	740–759	Congenital Anomalies
XV	760-779	Certain Conditions originating in the Perinatal Period
XVI	780–799	Symptoms, Signs and Ill-defined Conditions
XVII	800–999	Injury and Poisoning

Πίνακας 4: Κλάσεις κωδικών ICD-9

5 Αποτελέσματα και ανάλυση

Το αρχείο AllNet3.net του Hudine περιέχει 291172 συσχετίσεις μεταξύ 995 ασθενειών. Με σκοπό την εύρεση των πραγματικών συσχετίσεων και των αιτιατικών σχέσεων των ασθενειών εφαρμόστηκε μερική συσχέτιση σε δεδομένα απλής συσχέτισης. Η μερική συσχέτιση υπολογίστηκε με τη χρήση της συνάρτηση cor2pcor του πακέτου Corpcor της R, καθώς είναι η μοναδική συνάρτηση που υπολογίζει τη μερική συσχέτιση από ένα πίνακα συσχετίσεων. Από τον υπολογισμό αυτό προέκυψαν 494515 μερικές συσχετίσεις. Στην συνέχεια υπολογίστηκαν οι βαθμοί ελευθερίας, τα p-value και το FDR τους.

Το Volcano plot του οποίου ο x άξονας αναπαριστά τους συντελεστές μερικής συσχέτισης και ο y άξονας αναπαριστά τον αρνητικό λογάριθμο των p-value απεικονίζεται παρακάτω. Η σημαντικές θετικές συσχετίσεις (με p-value μικρότερο του 0.05 και συντελεστή μερικής συσχέτισης μεγαλύτερο του 0.06) αναπαρίστανται με κόκκινο χρώμα ενώ οι σημαντικές αρνητικές συσχετίσεις (με p-value μικρότερο του 0.05 και συντελεστή μερικής συσχέτισης μεγαλύτερο του -0.06) αναπαρίστανται με μπλε χρώμα. Η γραμμή στον x άξονα είναι στην τιμή -log(0.05) και οι γραμμές στον y είναι στις τιμές -0.06 και 0.06.



Εικόνα 2: Volcano plot μερικής συσχέτισης ασθενειών

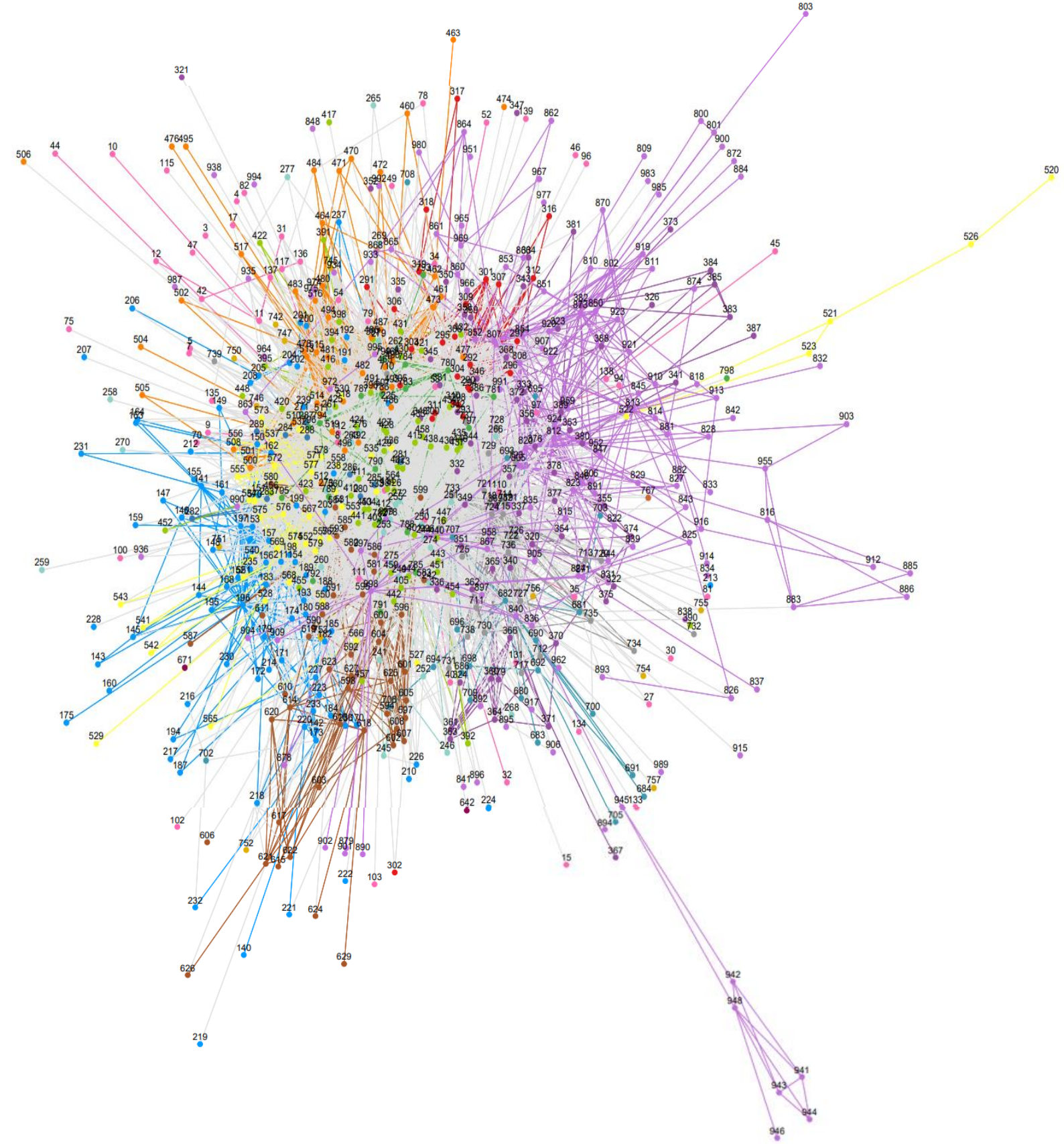
Οι σημαντικές θετικές συσχετίσεις είναι αυτές με τιμές p-value μικρότερες από 0.05 και μερική συσχέτιση μεγαλύτερη από 0.06. Οι σημαντικές αρνητικές συσχετίσεις είναι αυτές με τιμές p-value μικρότερες από 0.05 και μερική συσχέτιση μικρότερη από -0.06. Η θετική συσχέτιση υποδηλώνει την δυνατότητα της μιας ασθένειας να προκαλέσει την άλλη και η αρνητική συσχέτιση υποδηλώνει την δυνατότητα της μιας ασθένειας να αποτρέψει την εμφάνιση της άλλης.
Οι 8691 μερικές συσχετίσεις (6386 θετικές και 2305 αρνητικές) μεταξύ 714 ασθενειών είχαν adjusted FDRs μικρότερο από 0.05 και οι 7652 μερικές συσχετίσεις (5722 θετικές και 1930 αρνητικές) μεταξύ 696 ασθενειών είχαν adjusted FDRs μικρότερο από 0.01. Όλες οι μερικές συσχετίσεις που προέκυψαν από το κόψιμο με FDR είχαν p-value μικρότερο από 0.01.

Παρακάτω ακολουθεί ένας πίνακα που αναπαριστά τον αριθμό των συσχετίσεων και των ασθενειών για p-value μικρότερα από 0.05, p-value μικρότερα από 0.01, fdr μικρότερα από 0.05 και fdr μικρότερα από 0.01:

	Correlations (edges)	Diseases (nodes)	Positive correlations	Negative correlations	Partial correlations (edges)	Diseases of partial correlations (nodes)	Positive partial correlations	Negative correlations partial correlations
Without cut off	291172	995	205990	85182	494515	995	177607	316908
p-value <0.05	35093	877	29657	5436	14435	799	9826	4609
p-value <0.01	28805	835	24694	4111	11108	756	7878	3230
fdr <0.05	26728	828	23065	3663	8691	714	6386	2305
fdr <0.01	23491	798	20487	3004	7652	696	5722	1930

Πίνακας 5: Αριθμός συσχετίσεων και μερικών συσχετίσεων ασθενειών

Οι 5722 σημαντικές θετικές μερικές συσχετίσεις (ακμές) μεταξύ 695 μοναδικών ασθενειών (κόμβων) απεικονίζονται στο παρακάτω βιολογικό δίκτυο το οποίο σχηματίστηκε με τη χρήση των συναρτήσεων network() και ggnet2() των πακέτων network και ggnet της R:



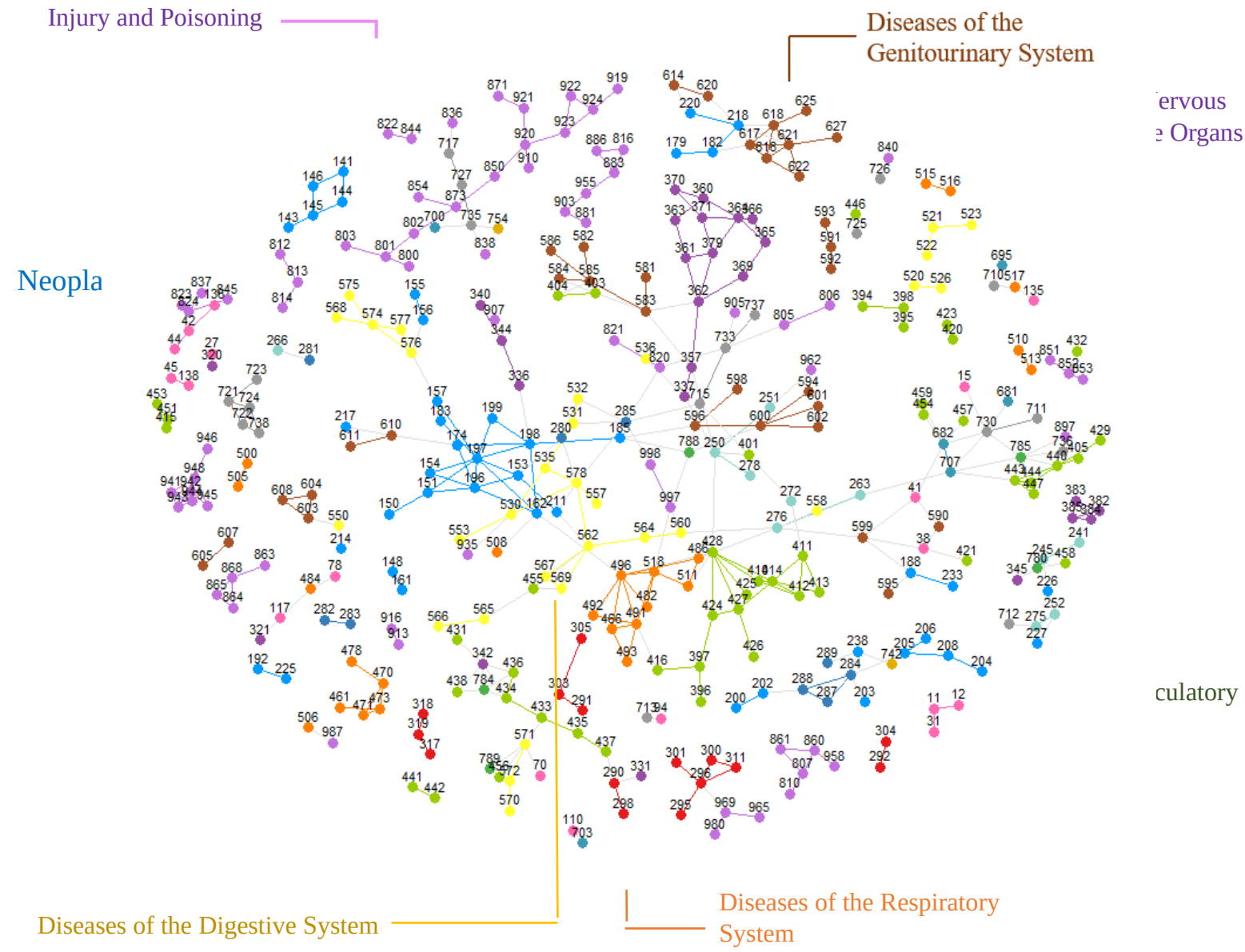
Εικόνα 3: Δίκτυο ασθενειών μερικής συσχέτισ

Οι διαφορετικοί χρωματισμοί των κόμβων συμβολίζουν την κλάση της ασθένειας στην οποία ανήκει κάθε κόμβος. Με αντίστοιχο χρώμα χρωματίζονται οι ακμές που συνδέουν κόμβους κοινής κλάσης. Συγκεκριμένα:

Χρώμα	Κωδικοί icd9	Κλάση ασθενειών
	001–139	Infectious and Parasitic Diseases
	140–239	Neoplasms
	240–279	Endocrine, Nutritional and Metabolic Diseases, and Immunity Disorders
	280–289	Diseases of the Blood and Blood-forming Organs
	290–319	Mental Disorders
	320–389	Diseases of the Nervous System and Sense Organs
	390–459	Diseases of the Circulatory System
	460–519	Diseases of the Respiratory System
	520–579	Diseases of the Digestive System
	580-629	Diseases of the Genitourinary System
	630-679	Complications of Pregnancy, Childbirth, and the Puerperium
	680-709	Diseases of the Skin and Subcutaneous Tissue
	710–739	Diseases of the Musculoskeletal System and Connective Tissue
	740–759	Congenital Anomalies
	760-779	Certain Conditions originating in the Perinatal Period
	780–799	Symptoms, Signs and Ill-defined Conditions
	800–999	Injury and Poisoning

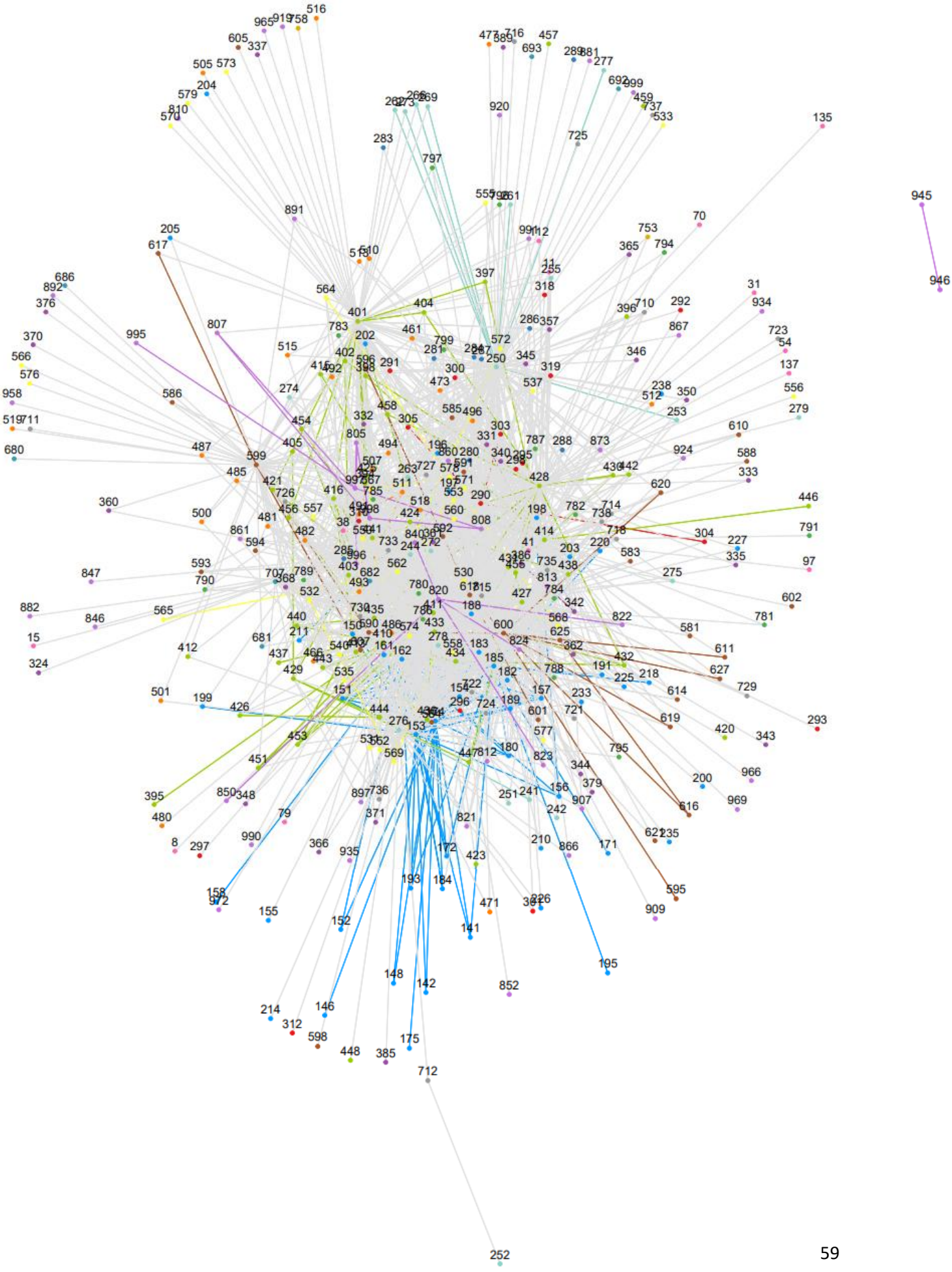
Πίνακας 6: Κλάσεις ασθενειών με το χρώμα που αναπαρίστανται στους κόμβους του δικτύου

Το παραπάνω δίκτυο των 5722 θετικών μερικών συσχετίσεων μπορεί να απεικονιστεί καλύτερα αν κόψουμε τις μερικές συσχετίσεις με τιμή μικρότερη από 0.06. Κατά αυτόν τον τρόπο, προκύπτει ένα δίκτυο με 406 ακμές (μερικές συσχετίσεις) και 358 κόμβους (ασθένειες) όπως απεικονίζεται παρακάτω:



Εικόνα 4: Δίκτυο ασθενειών μερικής συσχέτισης με τις κλάσεις των ασθενειών

Ακόμη, οι 1930 αρνητικές συσχετίσεις μεταξύ 372 ασθενειών απεικονίζονται στο παρακάτω δίκτυο:



Εικόνα 5: Δίκτυο ασθενειών αρνητικής μερικής συσχέτισης

Code file help.R

```
1. #edge list to undirect symmetric adjacency matrix
2.
3. edgelist_to_undirect_adjmatrix <- function(edgelist, genes) {
4.
5.
6.   edgelist$from<-as.character(edgelist$from)
7.   edgelist$to<-as.character(edgelist$to)
8.   genes<-genes[order(genes)]
9.
10.  mat<-matrix(0,ncol=length(genes),nrow=length(genes))
11.  rownames(mat)<-colnames(mat)<-genes
12.
13.  for(i in 1:nrow(edgelist)){
14.    mat[edgelist[i,1],edgelist[i,2]]<-edgelist[i,3]
15.  }
16.
17.  m=ifelse(!t(mat)[lower.tri(mat)], mat[lower.tri(mat)],
   t(mat)[lower.tri(mat)])
18.  mat[lower.tri(mat)]=m
19.  mat[upper.tri(mat)]=t(mat)[upper.tri(mat)]
20.
21.
22. return(mat)
23. }
24.
25. #create edge list from matrix
26. # if symmetric duplicates are removed
27. adjmatrix_to_edgelist<-
   function(mat,symmetric=TRUE,diagonal=FALSE,text=FALSE){
28.
29.  mat<-as.matrix(mat)
30.  id<-is.na(mat) # used to allow missing
31.  mat[id]<-"nna"
32.  if(symmetric){mat[lower.tri(mat)]<-"na"} # use to allow missing
   values
33.  if(!diagonal){diag(mat)<-"na"}
34.  obj<-melt(mat)
35.  obj<-obj[!obj$value=="na",]
36.  obj$value[obj$value=="nna"]<-NA
37.  if(!text){obj$value<-as.numeric(as.character(obj$value))}
38.  return(obj)
39.  # remove duplicates
40. }
```

Κώδικας 16: file help.R

Code Disease Network

```
1. library(corpcor)
2. library(reshape2)
3. library(reshape)
4. library("igraph")
5. library(dplyr)
```

```
library(FDRestimation)
library(readxl)
library(ggplot2)
library(ggnet)
library(corrplot)
library(network)

setwd("C:/Users/ioann/OneDrive/Υπολογιστής/Πτυχιακή Εργασία/codes")

source("help.r")

test1 <- read.table(file='AllNet3.tsv', sep = '\t', quote =
   "",header = TRUE)
test1<-test1[,c("icd9.1", "icd9.2", "n1", "n2", "pval", "pval.adj"
   )]
names(test1)<-c("from","to", "n1", "n2", "weight.cor", "test.cor")

test1$pvalue<-2*pt(q=abs(test1$test.cor), df=(max(test1$n1,
   test1$n2)-1), lower.tail=FALSE) #pvalue
test1$fdrs<-p.fdr(pvalues = test1$pvalue)$fdrs #adjusted FDR

#data frame of diseases code and prelevance
n1<-unique(test1[,c("from","n1")])
n2<-unique(test1[,c("to","n2")])
names(n1)<-names(n2)<-c("name", "n")
diseases<-dplyr::union(n1,n2, by="name")

rm(list="n1","n2")
test1<-test1[order(test1$from),]
diseases<-diseases[order(diseases$name),]

#edge list to undirect adjacency matrix
matrix<-edgelist_to_undirect_adjmatrix(test1[,c("from", "to",
   "weight.cor")], diseases$name)
test1$from<-as.character(test1$from)
test1$to<-as.character(test1$to)

diag(matrix)<-1

part_cor<-cor2pcor(matrix)
row.names(part_cor)<-colnames(part_cor)<-row.names(matrix)
rm(matrix)

#adjacency matix to non symmetrix edge list

part_cor<-adjmatrix_to_edgelist(part_cor)
colnames(part_cor)<-c("from","to","weight.pcor")
#add prelevance at partial correlation edge list

colnames(part_cor)[1] <- "name"
part_cor<-dplyr::inner_join(part_cor, diseases, by="name")
names(part_cor)<-c("from","name", "weight.pcor", "n1")
part_cor<-dplyr::inner_join(part_cor, diseases, by="name")
names(part_cor)<-c("from","to", "weight.pcor", "n1", "n2")
part_cor$from<-as.character(part_cor$from)
part_cor$to<-as.character(part_cor$to)

#combine of the two edge list (correlation and partial correlation)
test<-full_join(test1,part_cor, by=c("from", "to", "n1", "n2"))
```

```
61.
62. for(i in 1:nrow(test)){
63.   test$n.common[i]<-length(union(union(test1[which( test1$to %in%
    test$from[i]), "from"], test1[which( test1$from %in% test$from[i]),
    "to"]), union(test1[which( test1$to %in% test$to[i]), "from"],
    test1[which( test1$from %in% test$to[i]), "to"])))
64.
65. }
66.
67. test$degrees<-data.frame(pmax(test$n1, test$n2)-test$n.common-1)
68. test$test.pcor<-test$weight.pcor*sqrt((test$degrees)/(1-
    (test$weight.pcor^2)))
69.
70. test$pvalue<-2*pt(q=abs(test$test.pcor), df=test$degrees,
    lower.tail=FALSE)
71. test$fdrs<-p.fdr(pvalues = test$pvalue)$fdrs #adjusted p-val for FDR
72.
73. t<-test[which(test$fdrs<0.01),]
74. diseases<-data.frame(union(t$from, t$to))
75. names(diseases)<-"id"
76. t<-t[which(t$weight.pcor<0),]
77.
78.
79. #volcano plot
80. test_notna<-test
81. test_notna$Combined<-paste(test_notna$from,test_notna$to,sep="-")
82. test_notna$expression = ifelse(test_notna$pvalue < 0.05 &
    abs(test_notna$weight.pcor) >= 0.06, ifelse(test_notna$weight.pcor>
    0.06 ,'Up','Down'), 'Stable')
83.
84. test_notna$delabel <- NA
85. test_notna$delabel[test_notna$expression != "Stable"] <-
    test_notna$Combined[test_notna$expression != "Stable"]
86.
87. p<-ggplot(data = test_notna,
88.           aes(x = weight.pcor,
89.               y = -log10(pvalue),
90.               colour=expression)) +
91.   geom_point() +
92.   scale_color_manual(values=c("blue", "grey","red"))+
93.   xlim(c(-0.5, 0.5)) +
94.   geom_vline(xintercept=c(-0.06,0.06),lty=4,col="black",lwd=0.8) +
95.   geom_hline(yintercept = -log10(0.05),lty=4,col="black",lwd=0.8)+
96.   labs(x="partial correlation",
97.        y="-log10(p-value)",
98.        title="Volcano plot")
99.
100. #with labels
101. ggplot(data = test_notna,
102.         aes(x = weight.pcor,
103.             y = -log10(pvalue),
104.             colour=expression,
105.             label=delabel)) +
106.   geom_point() +
107.   scale_color_manual(values=c("blue", "grey","red"))+
108.   xlim(c(-0.5, 0.5)) +
109.   geom_vline(xintercept=c(-0.06,0.06),lty=4,col="black",lwd=0.8) +
110.   geom_hline(yintercept = -log10(0.05),lty=4,col="black",lwd=0.8)+
111.   labs(x="partial correlation",
```

```
        y="-log10(p-value)",
        title="Volcano plot")+
   theme_minimal() +
   geom_text()

t<-t[,c("from","to", "weight.pcor")]
names(t)<-c("from","to", "weight")
t<-t[order(t$from),]
mygraph <- graph.data.frame(t)

mat<-get.adjacency(mygraph, sparse = FALSE, attr='weight', names =
   diseases$id)

#to undirect
mat=as.matrix(mat)
m=ifelse(!t(mat)[lower.tri(mat)], mat[lower.tri(mat)],
   t(mat)[lower.tri(mat)])
mat[lower.tri(mat)]=m
mat[upper.tri(mat)]=t(mat)[upper.tri(mat)]

t<-melt(mat)
names(t)<-c("from","to", "weight")
t<-t[which(t$weight!=0),]
t<-t[order(t$from),]

my_data <- read_excel("classes.xlsx")

for(i in 1:nrow(my_data)){
   t[ t$from >= my_data$from[i] & t$from <= my_data$to[i],
   "value"]<-my_data$classes[i]
}


t$from<-as.character(t$from)
t$to<-as.character(t$to)

nw <- network(t[,c("from", "to")], directed = T, matrix.type =
   "edgelist")
mm.col <- c("Infectious and Parasitic Diseases" = "#ff69b4",
   "Neoplasms" = "#0099ff", "Endocrine, Nutritional and Metabolic
   Diseases, and Immunity Disorders"="#8DD3C7","Diseases of the Blood
   and Blood-forming Organs"="#377EB8", "Mental Disorders"=
   "#E41A1C","Diseases of the Nervous System and Sense
   Organs"="#984EA3", "Diseases of the Circulatory System"="#99CC00",
   "Diseases of the Respiratory System"="#FF7F00", "Diseases of the
   Digestive System"="#FFFF33",  "Diseases of the Genitourinary System"
   ="#A65628" ,"Diseases of the Skin and Subcutaneous Tissue"
   ="#3B9AB2", "Diseases of the Musculoskeletal System and Connective
   Tissue" ="#999999","Congenital Anomalies"="#E1AF00","Symptoms, Signs
   and Ill-defined Conditions"="#4DAF4A","Injury and
   Poisoning"="#C271DC")

for(i in 1:length(network.vertex.names(nw))){
   nw[["val"]][[i]]$value <- as.character(unique(t[
   which(t$from==network.vertex.names(nw)[i]),"value"]))

}

# create plot for ggnet2
```

```
154.
155. ggnet2(nw, color = mm.col[ nw %v% "value" ],
156.         label = TRUE,
157.         size = 4, vjust = -0.6, label.size = 10, edge.color =
   c("color", "gray88"))
158.
159.
160. icd9 <- read_excel(file.choose())
161. icd9$code<-as.character(icd9$code)
162.
163. diseases$id<-diseases[order(diseases$id),]
164.
165. for (i in 1:nrow(diseases)) {
166.   diseases$name[i]<-icd9[which(icd9$code %in% diseases$id[i]),
   "disease"]
167.
168. }
```

Κώδικας 17: Disease Network

6 Συμπεράσματα και μελλοντικές επεκτάσεις

Σε σύγκριση με την συμβατική μέθοδο της απλής συσχέτισης, η προτεινόμενη μερική συσχέτιση προσφέρει καλύτερα αποτελέσματα σε βιολογικά δίκτυα αντικατοπτρίζοντας τις αιτιατικές συσχετίσεις και τις πραγματικές αλληλεπιδράσεων μεταξύ των μεταβλητών. Αυτή η ιδιότητα καθιστά τη μέθοδο μια πολλά υποσχόμενη εναλλακτική λύση για την κατασκευή δικτύων πρωτεϊνικών αλληλεπιδράσεων, δικτύων ασθενειών και άλλων βιολογικών δικτύων που μπορεί να δώσουν μια εικόνα για τον μηχανισμό σύνθετων ασθενειών.

Η επαλήθευση των αποτελεσμάτων που προέκυψαν μπορεί να πραγματοποιηθεί μέσω της υπάρχουσας βιβλιογραφίας. Με μια πρωταρχική σύγκριση βρέθηκαν κοινές αιτιατικές συσχετίσεις ασθενειών μεταξύ των συσχετίσεων ασθενειών που υπολογίστηκαν με τη μέθοδο της μερικής συσχέτισης και των συσχετίσεων ασθενειών της ήδη υπάρχουσας βιβλιογραφίας. Ένα απόσπασμα των κοινών αιτιατικών συσχετίσεων μεταξύ των ασθενειών παρατίθεται στον ακόλουθο πίνακα: [22]–[26]

Diseases from Bibliography 1	Diseases from Bibliography 2	PMID	Diseases from Partial Correlation 1	Diseases from Partial Correlation 2	Causal Bibliography	Causal – Partial Correlation
hypothyroidism	major depressive disorder (MDD)	35507366	Acquired hypothyroidism	Affective psychoses	YES	YES
SBP	Renal cell carcinoma (RCC)	35312764	Essential hypertension	Malignant neoplasm of kidney and other and unspecified urinary organs	YES	YES
BMI	coronary heart disease	28500271	Obesity and other hyperalimentation	Other forms of chronic ischemic heart disease	YES	YES
Type 2 Diabetes	Hypertension	30646822	Diabetes mellitus	Essential hypertension	YES	YES
Endometriosis	clear cell ovarian cancer (CCOC)	35492879	Endometriosis	Benign neoplasm of ovary	YES	YES

Πίνακας 7: Επαλήθευση των αποτελεσμάτων μέσω της υπάρχουσας βιβλιογραφίας

Η μελλοντική μελέτη θα επικεντρωθεί στη σύγκριση των σημαντικών συσχετίσεων ασθενειών που προέκυψαν με τη χρήση της μερικής συσχέτισης με τις συσχετίσεις ασθενειών της υπάρχουσας βιβλιογραφίας, στην εφαρμογή του αλγορίθμου σε δίκτυα ασθενειών που προέκυψαν από τις βάσεις OMIM, GWAS Catalog and GAD, στη δημιουργία διμερών δικτύων ασθενειών-γονιδίων και στη

δημιουργία υπολογιστικής μεθόδου και server για υλοποίηση της μετα-ανάλυσης με μερική συσχέτιση.

7 Βιβλιογραφία

[1] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. 1–30, 2005, doi: 10.2202/1544-6115.1175.

[2] S. Kim, “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients,” *Communications for Statistical Applications and Methods*, vol. 22, no. 6, pp. 665–674, Nov. 2015, doi: 10.5351/csam.2015.22.6.665.

[3] K. Maintainer and S. Kim, “Package ‘ppcor’ Type Package Title Partial and Semi-Partial (Part) Correlation,” 2015.

[4] N. Krämer, J. Schäfer, and A. L. Boulesteix, “Regularized estimation of large-scale gene association networks using graphical Gaussian models,” *BMC Bioinformatics*, vol. 10, Nov. 2009, doi: 10.1186/1471-2105-10-384.

[5] A. Tenenhaus, V. Guillemot, X. Gidrol, and V. Frouin, “Gene association networks from microarray data using a regularized estimation of partial correlation based on PLS regression,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 251–262, 2010, doi: 10.1109/TCBB.2008.87.

[6] N. Kraemer and J. S. Maintainer, “Package ‘parcor’ Title Regularized estimation of partial correlation matrices,” 2015. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/384/

[7] R. Mazumder and T. Hastie, “The graphical lasso: New insights and alternatives,” *Electronic Journal of Statistics*, vol. 6, pp. 2125–2149, 2012, doi: 10.1214/12-EJS740.

[8] P. Zhang, B. R. Southey, and S. L. Rodriguez-Zas, “Co-expression networks uncover regulation of splicing and transcription markers of disease,” 2020.

[9] J. Friedman, T. Hastie, and R. Tibshirani, “Title Graphical Lasso: Estimation of Gaussian Graphical Models,” 2019.

[10] J. Peng, P. Wang, N. Zhou, and J. Zhu, “Partial correlation estimation by joint sparse regression models,” *J Am Stat Assoc*, vol. 104, no. 486, pp. 735–746, Jun. 2009, doi: 10.1198/jasa.2009.0126.

[11] P. Wang, <pwang@fhcrc Org>, J. Zhu, and M. P. Wang, “Package ‘space’ Title Sparse PArtial Correlation Estimation,” 2015.

[12] G. Yoon, I. Gaynanova, and C. L. Müller, “Microbial networks in SPRING - Semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data,” *Frontiers in Genetics*, vol. 10, no. JUN, 2019, doi: 10.3389/fgene.2019.00516.

[13] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the Lasso,” *Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006, doi: 10.1214/009053606000000281.

[14] J. Schaefer, R. Opgen-Rhein, K. S. Maintainer, and K. Strimmer, “Package ‘GeneNet’ Title Modeling and Inferring Gene Networks,” 2021. [Online]. Available: https://strimmerlab.github.io/software/genenet/

[15] K. K. Dey and M. Stephens, “CorShrink : Empirical Bayes shrinkage estimation of correlations, with applications”, doi: 10.1101/368316.

[16] Y. Zuo, G. Yu, M. G. Tadesse, and H. W. Ressom, “Biological network inference using low order partial correlation,” *Methods*, vol. 69, no. 3, pp. 266–273, Oct. 2014, doi: 10.1016/j.ymeth.2014.06.010.

[17] M. Wenbin Guo, “Package ‘RLowPC’ Title Inference of co-expression gene network using relevance low order partial correla-tion from large scale expression data,” 2016.

[18] W. Guo *et al.*, “Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size,” *BMC Systems Biology*, vol. 11, no. 1, Jun. 2017, doi: 10.1186/s12918-017-0440-2.

[19] J. Schäfer and K. Strimmer, “An empirical Bayes approach to inferring large-scale gene association networks,” *Bioinformatics*, vol. 21, no. 6, pp. 754–764, Mar. 2005, doi: 10.1093/bioinformatics/bti062.

[20] A. P. Field, “Is the meta-analysis of correlation coefficients accurate when population correlations vary?,” *Psychological Methods*, vol. 10, no. 4, pp. 444–467, Dec. 2005, doi: 10.1037/1082-989X.10.4.444.

[21] C. A. Hidalgo, N. Blumm, A. L. Barabási, and N. A. Christakis, “A Dynamic Network Approach for the Study of Human Phenotypes,” *PLoS Computational Biology*, vol. 5, no. 4, Apr. 2009, doi: 10.1371/journal.pcbi.1000353.

[22] D. S. Tylee *et al.*, “An Atlas of Genetic Correlations and Genetically Informed Associations Linking Psychiatric and Immune-Related Phenotypes.,” *JAMA Psychiatry*, May 2022, doi: 10.1001/jamapsychiatry.2022.0914.

[23] K. Alcala *et al.*, “The relationship between blood pressure and risk of renal cell carcinoma.,” *Int J Epidemiol*, Mar. 2022, doi: 10.1093/ije/dyac042.

[24] C. E. Dale *et al.*, “Causal Associations of Adiposity and Body Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis.,” *Circulation*, vol. 135, no. 24, pp. 2373–2388, Jun. 2017, doi: 10.1161/CIRCULATIONAHA.116.026560.

[25] D. Sun *et al.*, “Type 2 Diabetes and Hypertension.,” *Circ Res*, vol. 124, no. 6, pp. 930–937, 2019, doi: 10.1161/CIRCRESAHA.118.314487.

[26] S. Mortlock *et al.*, “A multi-level investigation of the genetic relationship between endometriosis and ovarian cancer histotypes.,” *Cell Rep Med*, vol. 3, no. 3, p. 100542, Mar. 2022, doi: 10.1016/j.xcrm.2022.100542.

8 Παράρτημα

Οι κωδικοί των ασθενειών αντιστοιχούν στις παρακάτω ασθένειες:

Κωδικός	Ασθένεια
1	Cholera
2	Typhoid and paratyphoid fevers
3	Other salmonella infections
4	Shigellosis
5	Other food poisoning (bacterial)
6	Amebiasis
7	Other protozoal intestinal diseases
8	Intestinal infections due to other organisms
9	Ill-defined intestinal infections
10	Primary tuberculous infection
11	Pulmonary tuberculosis
12	Other respiratory tuberculosis
13	Tuberculosis of meninges and central nervous system
14	Tuberculosis of intestines, peritoneum, and mesenteric glands
15	Tuberculosis of bones and joints
16	Tuberculosis of genitourinary system
17	Tuberculosis of other organs
18	Miliary tuberculosis
20	Plague
21	Tularemia
22	Anthrax
23	Brucellosis
24	Glanders
25	Melioidosis
26	Rat-bite fever
27	Other zoonotic bacterial diseases
30	Leprosy
31	Diseases due to other mycobacteria
32	Diphtheria
33	Whooping cough
34	Streptococcal sore throat and scarlet fever
35	Erysipelas
36	Meningococcal infection
37	Tetanus
38	Septicemia
39	Actinomycotic infections
40	Other bacterial diseases
41	Bacterial infection in conditions classified elsewhere and of unspecified site
42	Human immunodeficiency virus [HIV] disease

45	Acute poliomyelitis
46	Slow virus infection of central nervous system
47	Meningitis due to enterovirus
48	Other enterovirus diseases of central nervous system
49	Other non-arthropod-borne viral diseases of central nervous system
50	Smallpox
51	Cowpox and paravaccinia
52	Chickenpox
53	Herpes zoster
54	Herpes simplex
55	Measles
56	Rubella
57	Other viral exanthemata
58	Other human herpesvirus
59	Other poxvirus infections
60	Yellow fever
61	Dengue
62	Mosquito-borne viral encephalitis
63	Tick-borne viral encephalitis
64	Viral encephalitis transmitted by other and unspecified arthropods
65	Arthropod-borne hemorrhagic fever
66	Other arthropod-borne viral diseases
70	Viral hepatitis
71	Rabies
72	Mumps
73	Ornithosis
74	Specific diseases due to Coxsackie virus
75	Infectious mononucleosis
76	Trachoma
77	Other diseases of conjunctiva due to viruses and Chlamydiae
78	Other diseases due to viruses and Chlamydiae
79	Viral infection in conditions classified elsewhere and of unspecified site
80	Louse-borne [epidemic] typhus
81	Other typhus
82	Tick-borne rickettsioses
83	Other rickettsioses
84	Malaria
85	Leishmaniasis
86	Trypanosomiasis
87	Relapsing fever
88	Other arthropod-borne diseases
90	Congenital syphilis
91	Early syphilis, symptomatic
92	Early syphilis, latent
93	Cardiovascular syphilis
94	Neurosyphilis

95	Other forms of late syphilis, with symptoms
96	Late syphilis, latent
97	Other and unspecified syphilis
98	Gonococcal infections
99	Other venereal diseases
100	Leptospirosis
101	Vincent's angina
102	Yaws
103	Pinta
104	Other spirochetal infection
110	Dermatophytosis
111	Dermatomycosis, other and unspecified
112	Candidiasis
114	Coccidioidomycosis
115	Histoplasmosis
116	Blastomycotic infection
117	Other mycoses
118	Opportunistic mycoses
120	Schistosomiasis [bilharziasis]
121	Other trematode infections
122	Echinococcosis
123	Other cestode infection
124	Trichinosis
125	Filarial infection and dracontiasis
126	Ancylostomiasis and necatoriasis
127	Other intestinal helminthiases
128	Other and unspecified helminthiases
129	Intestinal parasitism, unspecified
130	Toxoplasmosis
131	Trichomoniasis
132	Pediculosis and phthirus infestation
133	Acariasis
134	Other infestation
135	Sarcoidosis
136	Other and unspecified infectious and parasitic diseases
137	Late effects of tuberculosis
138	Late effects of acute poliomyelitis
139	Late effects of other infectious and parasitic diseases
140	Malignant neoplasm of lip
141	Malignant neoplasm of tongue
142	Malignant neoplasm of major salivary glands
143	Malignant neoplasm of gum
144	Malignant neoplasm of floor of mouth
145	Malignant neoplasm of other and unspecified parts of mouth
146	Malignant neoplasm of oropharynx
147	Malignant neoplasm of nasopharynx

148	Malignant neoplasm of hypopharynx
149	Malignant neoplasm of other and ill-defined sites within the lip, oral cavity, and pharynx
150	Malignant neoplasm of esophagus
151	Malignant neoplasm of stomach
152	Malignant neoplasm of small intestine, including duodenum
153	Malignant neoplasm of colon
154	Malignant neoplasm of rectum, rectosigmoid junction, and anus
155	Malignant neoplasm of liver and intrahepatic bile ducts
156	Malignant neoplasm of gallbladder and extrahepatic bile ducts
157	Malignant neoplasm of pancreas
158	Malignant neoplasm of retroperitoneum and peritoneum
159	Malignant neoplasm of other and ill-defined sites within the digestive organs and peritoneum
160	Malignant neoplasm of nasal cavities, middle ear, and accessory sinuses
161	Malignant neoplasm of larynx
162	Malignant neoplasm of trachea, bronchus, and lung
163	Malignant neoplasm of pleura
164	Malignant neoplasm of thymus, heart, and mediastinum
165	Malignant neoplasm of other and ill-defined sites within the respiratory system and intrathoracic organs
170	Malignant neoplasm of bone and articular cartilage
171	Malignant neoplasm of connective and other soft tissue
172	Malignant melanoma of skin
173	Other malignant neoplasm of skin
174	Malignant neoplasm of female breast
175	Malignant neoplasm of male breast
176	Kaposi's sarcoma
179	Malignant neoplasm of uterus, part unspecified
180	Malignant neoplasm of cervix uteri
181	Malignant neoplasm of placenta
182	Malignant neoplasm of body of uterus
183	Malignant neoplasm of ovary and other uterine adnexa
184	Malignant neoplasm of other and unspecified female genital organs
185	Malignant neoplasm of prostate
186	Malignant neoplasm of testis
187	Malignant neoplasm of penis and other male genital organs
188	Malignant neoplasm of bladder
189	Malignant neoplasm of kidney and other and unspecified urinary organs
190	Malignant neoplasm of eye
191	Malignant neoplasm of brain
192	Malignant neoplasm of other and unspecified parts of nervous system
193	Malignant neoplasm of thyroid gland
194	Malignant neoplasm of other endocrine glands and related structures
195	Malignant neoplasm of other and ill-defined sites
196	Secondary and unspecified malignant neoplasm of lymph nodes
197	Secondary malignant neoplasm of respiratory and digestive systems
198	Secondary malignant neoplasm of other specified sites
199	Malignant neoplasm without specification of site

200	Lymphosarcoma and reticulosarcoma
201	Hodgkin's disease
202	Other malignant neoplasm of lymphoid and histiocytic tissue
203	Multiple myeloma and immunoproliferative neoplasms
204	Lymphoid leukemia
205	Myeloid leukemia
206	Monocytic leukemia
207	Other specified leukemia
208	Leukemia of unspecified cell type
209	Neuroendocrine tumors
210	Benign neoplasm of lip, oral cavity, and pharynx
211	Benign neoplasm of other parts of digestive system
212	Benign neoplasm of respiratory and intrathoracic organs
213	Benign neoplasm of bone and articular cartilage
214	Lipoma
215	Other benign neoplasm of connective and other soft tissue
216	Benign neoplasm of skin
217	Benign neoplasm of breast
218	Uterine leiomyoma
219	Other benign neoplasm of uterus
220	Benign neoplasm of ovary
221	Benign neoplasm of other female genital organs
222	Benign neoplasm of male genital organs
223	Benign neoplasm of kidney and other urinary organs
224	Benign neoplasm of eye
225	Benign neoplasm of brain and other parts of nervous system
226	Benign neoplasm of thyroid gland
227	Benign neoplasm of other endocrine glands and related structures
228	Hemangioma and lymphangioma, any site
229	Benign neoplasm of other and unspecified sites
230	Carcinoma in situ of digestive organs
231	Carcinoma in situ of respiratory system
232	Carcinoma in situ of skin
233	Carcinoma in situ of breast and genitourinary system
234	Carcinoma in situ of other and unspecified sites
235	Neoplasm of uncertain behavior of digestive and respiratory systems
236	Neoplasm of uncertain behavior of genitourinary organs
237	Neoplasm of uncertain behavior of endocrine glands and nervous system
238	Neoplasm of uncertain behavior of other and unspecified sites and tissues
239	Neoplasm of unspecified nature
240	Simple and unspecified goiter
241	Nontoxic nodular goiter
242	Thyrotoxicosis with or without goiter
243	Congenital hypothyroidism
244	Acquired hypothyroidism
245	Thyroiditis

246	Other disorders of thyroid
249	Secondary diabetes mellitus
250	Diabetes mellitus
251	Other disorders of pancreatic internal secretion
252	Disorders of parathyroid gland
253	Disorders of the pituitary gland and its hypothalamic control
254	Diseases of thymus gland
255	Disorders of adrenal glands
256	Ovarian dysfunction
257	Testicular dysfunction
258	Polyglandular dysfunction and related disorders
259	Other endocrine disorders
260	Kwashiorkor
261	Nutritional marasmus
262	Other severe protein-calorie malnutrition
263	Other and unspecified protein-calorie malnutrition
264	Vitamin A deficiency
265	Thiamine and niacin deficiency states
266	Deficiency of B-complex components
267	Ascorbic acid deficiency
268	Vitamin D deficiency
269	Other nutritional deficiencies
270	Disorders of amino-acid transport and metabolism
271	Disorders of carbohydrate transport and metabolism
272	Disorders of lipoid metabolism
273	Disorders of plasma protein metabolism
274	Gout
275	Disorders of mineral metabolism
276	Disorders of fluid, electrolyte, and acid-base balance
277	Other and unspecified disorders of metabolism
278	Obesity and other hyperalimentation
279	Disorders involving the immune mechanism
280	Iron deficiency anemias
281	Other deficiency anemias
282	Hereditary hemolytic anemias
283	Acquired hemolytic anemias
284	Aplastic anemia
285	Other and unspecified anemias
286	Coagulation defects
287	Purpura and other hemorrhagic conditions
288	Diseases of white blood cells
289	Other diseases of blood and blood-forming organs
290	Senile and presenile organic psychotic conditions
291	Alcoholic psychoses
292	Drug psychoses
293	Transient organic psychotic conditions

294	Other organic psychotic conditions (chronic)
295	Schizophrenic psychoses
296	Affective psychoses
297	Paranoid states
298	Other nonorganic psychoses
299	Psychoses with origin specific to childhood
300	Neurotic disorders
301	Personality disorders
302	Sexual deviations and disorders
303	Alcohol dependence syndrome
304	Drug dependence
305	Nondependent abuse of drugs
306	Physiological malfunction arising from mental factors
307	Special symptoms or syndromes, not elsewhere classified
308	Acute reaction to stress
309	Adjustment reaction
310	Specific nonpsychotic mental disorders following organic brain damage
311	Depressive disorder, not elsewhere classified
312	Disturbance of conduct, not elsewhere classified
313	Disturbance of emotions specific to childhood and adolescence
314	Hyperkinetic syndrome of childhood
315	Specific delays in development
316	Psychic factors associated with diseases classified elsewhere
317	Mild mental retardation
318	Other specified mental retardation
319	Unspecified mental retardation
320	Bacterial meningitis
321	Meningitis due to other organisms
322	Meningitis of unspecified cause
323	Encephalitis, myelitis, and encephalomyelitis
324	Intracranial and intraspinal abscess
325	Phlebitis and thrombophlebitis of intracranial venous sinuses
326	Late effects of intracranial abscess or pyogenic infection
327	Organic sleep disorders
330	Cerebral degenerations usually manifest in childhood
331	Other cerebral degenerations
332	Parkinson's disease
333	Other extrapyramidal disease and abnormal movement disorders
334	Spinocerebellar disease
335	Anterior horn cell disease
336	Other diseases of spinal cord
337	Disorders of the autonomic nervous system
338	Pain, not elsewhere classified
339	Other headache syndromes
340	Multiple sclerosis
341	Other demyelinating diseases of central nervous system

342	Hemiplegia and hemiparesis
343	Infantile cerebral palsy
344	Other paralytic syndromes
345	Epilepsy
346	Migraine
347	Cataplexy and narcolepsy
348	Other conditions of brain
349	Other and unspecified disorders of the nervous system
350	Trigeminal nerve disorders
351	Facial nerve disorders
352	Disorders of other cranial nerves
353	Nerve root and plexus disorders
354	Mononeuritis of upper limb and mononeuritis multiplex
355	Mononeuritis of lower limb
356	Hereditary and idiopathic peripheral neuropathy
357	Inflammatory and toxic neuropathy
358	Myoneural disorders
359	Muscular dystrophies and other myopathies
360	Disorders of the globe
361	Retinal detachments and defects
362	Other retinal disorders
363	Chorioretinal inflammations and scars and other disorders of choroid
364	Disorders of iris and ciliary body
365	Glaucoma
366	Cataract
367	Disorders of refraction and accommodation
368	Visual disturbances
369	Blindness and low vision
370	Keratitis
371	Corneal opacity and other disorders of cornea
372	Disorders of conjunctiva
373	Inflammation of eyelids
374	Other disorders of eyelids
375	Disorders of lacrimal system
376	Disorders of the orbit
377	Disorders of optic nerve and visual pathways
378	Strabismus and other disorders of binocular eye movements
379	Other disorders of eye
380	Disorders of external ear
381	Nonsuppurative otitis media and Eustachian tube disorders
382	Suppurative and unspecified otitis media
383	Mastoiditis and related conditions
384	Other disorders of tympanic membrane
385	Other disorders of middle ear and mastoid
386	Vertiginous syndromes and other disorders of vestibular system
387	Otosclerosis

388	Other disorders of ear
389	Hearing loss
390	Rheumatic fever without mention of heart involvement
391	Rheumatic fever with heart involvement
392	Rheumatic chorea
393	Chronic rheumatic pericarditis
394	Diseases of mitral valve
395	Diseases of aortic valve
396	Diseases of mitral and aortic valves
397	Diseases of other endocardial structures
398	Other rheumatic heart disease
401	Essential hypertension
402	Hypertensive heart disease
403	Hypertensive renal disease
404	Hypertensive heart and renal disease
405	Secondary hypertension
410	Acute myocardial infarction
411	Other acute and subacute form of ischemic heart disease
412	Old myocardial infarction
413	Angina pectoris
414	Other forms of chronic ischemic heart disease
415	Acute pulmonary heart disease
416	Chronic pulmonary heart disease
417	Other diseases of pulmonary circulation
420	Acute pericarditis
421	Acute and subacute endocarditis
422	Acute myocarditis
423	Other diseases of pericardium
424	Other diseases of endocardium
425	Cardiomyopathy
426	Conduction disorders
427	Cardiac dysrhythmias
428	Heart failure
429	Ill-defined descriptions and complications of heart disease
430	Subarachnoid hemorrhage
431	Intracerebral hemorrhage
432	Other and unspecified intracranial hemorrhage
433	Occlusion and stenosis of precerebral arteries
434	Occlusion of cerebral arteries
435	Transcient cerebral ischemia
436	Acute but ill-defined cerebrovascular disease
437	Other and ill-defined cerebrovascular disease
438	Late effects of cerebrovascular disease
440	Atherosclerosis
441	Aortic aneurysm and dissection
442	Other aneurysm

443	Other peripheral vascular disease
444	Arterial embolism and thrombosis
445	Atheroembolism
446	Polyarteritis nodosa and allied conditions
447	Other disorders of arteries and arterioles
448	Diseases of capillaries
451	Phlebitis and thrombophlebitis
452	Portal vein thrombosis
453	Other venous embolism and thrombosis
454	Varicose veins of lower extremities
455	Hemorrhoids
456	Varicose veins of other sites
457	Noninfective disorders of lymphatic channels
458	Hypotension
459	Other disorders of circulatory system
460	Acute nasopharyngitis [common cold]
461	Acute sinusitis
462	Acute pharyngitis
463	Acute tonsillitis
464	Acute laryngitis and tracheitis
465	Acute upper respiratory infections of multiple or unspecified sites
466	Acute bronchitis and bronchiolitis
470	Deviated nasal septum
471	Nasal polyps
472	Chronic pharyngitis and nasopharyngitis
473	Chronic sinusitis
474	Chronic disease of tonsils and adenoids
475	Peritonsillar abscess
476	Chronic laryngitis and laryngotracheitis
477	Allergic rhinitis
478	Other diseases of upper respiratory tract
480	Viral pneumonia
481	Pneumococcal pneumonia [Streptococcus pneumoniae pneumonia]
482	Other bacterial pneumonia
483	Pneumonia due to other specified organism
484	Pneumonia in infectious diseases classified elsewhere
485	Bronchopneumonia, organism unspecified
486	Pneumonia, organism unspecified
487	Influenza
488	Influenza due to identified avian influenza virus
490	Bronchitis, not specified as acute or chronic
491	Chronic bronchitis
492	Emphysema
493	Asthma
494	Bronchiectasis
495	Extrinsic allergic alveolitis

496	Chronic airways obstruction, not elsewhere classified
500	Coalworkers' pneumoconiosis
501	Asbestosis
502	Pneumoconiosis due to other silica or silicates
503	Pneumoconiosis due to other inorganic dust
504	Pneumopathy due to inhalation of other dust
505	Pneumoconiosis, unspecified
506	Respiratory conditions due to chemical fumes and vapors
507	Pneumonitis due to solids and liquids
508	Respiratory conditions due to other and unspecified external agents
510	Empyema
511	Pleurisy
512	Pneumothorax
513	Abscess of lung and mediastinum
514	Pulmonary congestion and hypostasis
515	Postinflammatory pulmonary fibrosis
516	Other alveolar and parietoalveolar pneumopathy
517	Lung involvement in conditions classified elsewhere
518	Other diseases of lung
519	Other diseases of respiratory system
520	Disorders of tooth development and eruption
521	Diseases of hard tissues of teeth
522	Diseases of pulp and periapical tissues
523	Gingival and periodontal diseases
524	Dentofacial anomalies, including malocclusion
525	Other diseases and conditions of the teeth and supporting structures
526	Diseases of the jaws
527	Diseases of the salivary glands
528	Diseases of the oral soft tissues, excluding lesions specific for gingiva and tongue
529	Diseases and other conditions of the tongue
530	Diseases of esophagus
531	Gastric ulcer
532	Duodenal ulcer
533	Peptic ulcer, site unspecified
534	Gastrojejunal ulcer
535	Gastritis and duodenitis
536	Disorders of function of stomach
537	Other disorders of stomach and duodenum
538	Gastrointestinal mucositis (ulcerative)
540	Acute appendicitis
541	Appendicitis, unqualified
542	Other appendicitis
543	Other diseases of appendix
550	Inguinal hernia
551	Other hernia of abdominal cavity, with gangrene
552	Other hernia of abdominal cavity, with obstruction, but without mention of gangrene

553	Other hernia of abdominal cavity without mention of obstruction or gangrene
555	Regional enteritis
556	Ulcerative colitis
557	Vascular insufficiency of intestine
558	Other noninfective gastroenteritis and colitis
560	Intestinal obstruction without mention of hernia
562	Diverticula of intestine
564	Functional digestive disorders, not elsewhere classified
565	Anal fissure and fistula
566	Abscess of anal and rectal regions
567	Peritonitis
568	Other disorders of peritoneum
569	Other disorders of intestine
570	Acute and subacute necrosis of liver
571	Chronic liver disease and cirrhosis
572	Liver abscess and sequelae of chronic liver disease
573	Other disorders of liver
574	Cholelithiasis
575	Other disorders of gallbladder
576	Other disorders of biliary tract
577	Diseases of pancreas
578	Gastrointestinal hemorrhage
579	Intestinal malabsorption
580	Acute glomerulonephritis
581	Nephrotic syndrome
582	Chronic glomerulonephritis
583	Nephritis and nephropathy, not specified as acute or chronic
584	Acute renal failure
585	Chronic renal failure
586	Renal failure, unspecified
587	Renal sclerosis, unspecified
588	Disorders resulting from impaired renal function
589	Small kidney of unknown cause
590	Infections of kidney
591	Hydronephrosis
592	Calculus of kidney and ureter
593	Other disorders of kidney and ureter
594	Calculus of lower urinary tract
595	Cystitis
596	Other disorders of bladder
597	Urethritis, not sexually transmitted, and urethral syndrome
598	Urethral stricture
599	Other disorders of urethra and urinary tract
600	Hyperplasia of prostate
601	Inflammatory diseases of prostate
602	Other disorders of prostate

603	Hydrocele
604	Orchitis and epididymitis
605	Redundant prepuce and phimosis
606	Infertility, male
607	Disorders of penis
608	Other disorders of male genital organs
610	Benign mammary dysplasias
611	Other disorders of breast
612	Deformity and disproportion of reconstructed breast
614	Inflammatory disease of ovary, fallopian tube, pelvic cellular tissue, and peritoneum
615	Inflammatory diseases of uterus, except cervix
616	Inflammatory disease of cervix, vagina, and vulva
617	Endometriosis
618	Genital prolapse
619	Fistula involving female genital tract
620	Noninflammatory disorders of ovary, fallopian tube, and broad ligament
621	Disorders of uterus, not elsewhere classified
622	Noninflammatory disorders of cervix
623	Noninflammatory disorders of vagina
624	Noninflammatory disorders of vulva and perineum
625	Pain and other symptoms associated with female genital organs
626	Disorders of menstruation and other abnormal bleeding from female genital tract
627	Menopausal and postmenopausal disorders
628	Infertility, female
629	Other disorders of female genital organs
630	Hydatidiform mole
631	Other abnormal product of conception
632	Missed abortion
633	Ectopic pregnancy
634	Spontaneous abortion
635	Legally induced abortion
636	Illegally induced abortion
637	Unspecified abortion
638	Failed attempted abortion
639	Complications following abortion and ectopic and molar pregnancies
640	Hemorrhage in early pregnancy
641	Antepartum hemorrhage, abruptio placentae, and placenta previa
642	Hypertension complicating pregnancy, childbirth, and the puerperium
643	Excessive vomiting in pregnancy
644	Early or threatened labor
645	Prolonged pregnancy
646	Other complications of pregnancy, not elsewhere classified
647	Infective and parasitic conditions in the mother classifiable elsewhere but complicating pregnancy, childbirth, and the puerperium
648	Other current conditions in the mother classifiable elsewhere but complicating pregnancy, childbirth, and the puerperium
649	Other conditions or status of the mother complicating pregnancy, childbirth, or the puerperium

650	Normal delivery
651	Multiple gestation
652	Malposition and malpresentation of fetus
653	Disproportion
654	Abnormality of organs and soft tissues of pelvis
655	Known or suspected fetal abnormality affecting management of mother
656	Other fetal and placental problems affecting management of mother
657	Polyhydramnios
658	Other problems associated with amniotic cavity and membranes
659	Other indications for care or intervention related to labor and delivery and not elsewhere classified
660	Obstructed labor
661	Abnormality of forces of labor
662	Long labor
663	Umbilical cord complications
664	Trauma to perineum and vulva during delivery
665	Other obstetrical trauma
666	Postpartum hemorrhage
667	Retained placenta or membranes, without hemorrhage
668	Complications of the administration of anesthetic or other sedation in labor and delivery
669	Other complications of labor and delivery, not elsewhere classified
670	Major puerperal infection
671	Venous complications in pregnancy and the puerperium
672	Pyrexia of unknown origin during the puerperium
673	Obstetrical pulmonary embolism
674	Other and unspecified complications of the puerperium, not elsewhere classified
675	Infections of the breast and nipple associated with childbirth
676	Other disorders of the breast associated with childbirth, and disorders of lactation
677	Late effect of complication of pregnancy, childbirth, and the puerperium
678	Other fetal conditions
679	Complications of in utero procedures
680	Carbuncle and furuncle
681	Cellulitis and abscess of finger and toe
682	Other cellulitis and abscess
683	Acute lymphadenitis
684	Impetigo
685	Pilonidal cyst
686	Other local infections of skin and subcutaneous tissue
690	Erythematosquamous dermatosis
691	Atopic dermatitis and related conditions
692	Contact dermatitis and other eczema
693	Dermatitis due to substances taken internally
694	Bullous dermatoses
695	Erythematous conditions
696	Psoriasis and similar disorders
697	Lichen
698	Pruritus and related conditions

700	Corns and callosities
701	Other hypertrophic and atrophic conditions of skin
702	Other dermatoses
703	Diseases of nail
704	Diseases of hair and hair follicles
705	Disorders of sweat glands
706	Diseases of sebaceous glands
707	Chronic ulcer of skin
708	Urticaria
709	Other disorders of skin and subcutaneous tissue
710	Diffuse diseases of connective tissue
711	Arthropathy associated with infections
712	Crystal arthropathies
713	Arthropathy associated with other disorders classified elsewhere
714	Rheumatoid arthritis and other inflammatory polyarthropathies
715	Osteoarthrosis and allied disorders
716	Other and unspecified arthropathies
717	Internal derangement of knee
718	Other derangement of joint
719	Other and unspecified disorder of joint
720	Ankylosing spondylitis and other inflammatory spondylopathies
721	Spondylosis and allied disorders
722	Intervertebral disc disorders
723	Other disorders of cervical region
724	Other and unspecified disorders of back
725	Polymyalgia rheumatica
726	Peripheral enthesopathies and allied syndromes
727	Other disorders of synovium, tendon, and bursa
728	Disorders of muscle, ligament, and fascia
729	Other disorders of soft tissues
730	Osteomyelitis, periostitis, and other infections involving bone
731	Osteitis deformans and osteopathies associated with other disorders classified elsewhere
732	Osteochondropathies
733	Other disorders of bone and cartilage
734	Flat foot
735	Acquired deformities of toe
736	Other acquired deformities of limbs
737	Curvature of spine
738	Other acquired deformity
739	Nonallopathic lesions, not elsewhere classified
740	Anencephalus and similar anomalies
741	Spina bifida
742	Other congenital anomalies of nervous system
743	Congenital anomalies of eye
744	Congenital anomalies of ear, face, and neck
745	Bulbus cordis anomalies and anomalies of cardiac septal closure

746	Other congenital anomalies of heart
747	Other congenital anomalies of circulatory system
748	Congenital anomalies of respiratory system
749	Cleft palate and cleft lip
750	Other congenital anomalies of upper alimentary tract
751	Other congenital anomalies of digestive system
752	Congenital anomalies of genital organs
753	Congenital anomalies of urinary system
754	Certain congenital musculoskeletal deformities
755	Other congenital anomalies of limbs
756	Other congenital musculoskeletal anomalies
757	Congenital anomalies of the integument
758	Chromosomal anomalies
759	Other and unspecified congenital anomalies
760	Fetus or newborn affected by maternal conditions which may be unrelated to present pregnancy
761	Fetus or newborn affected by maternal complications of pregnancy
762	Fetus or newborn affected by complications of placenta, cord, and membranes
763	Fetus or newborn affected by other complications of labor and delivery
764	Slow fetal growth and fetal malnutrition
765	Disorders relating to short gestation and unspecified low birthweight
766	Disorders relating to long gestation and high birthweight
767	Birth trauma
768	Intrauterine hypoxia and birth asphyxia
769	Respiratory distress syndrome
770	Other respiratory conditions of fetus and newborn
771	Infections specific to the perinatal period
772	Fetal and neonatal hemorrhage
773	Hemolytic disease of fetus or newborn, due to isoimmunization
774	Other perinatal jaundice
775	Endocrine and metabolic disturbances specific to the fetus and newborn
776	Hematological disorders of fetus and newborn
777	Perinatal disorders of digestive system
778	Conditions involving the integument and temperature regulation of fetus and newborn
779	Other and ill-defined conditions originating in the perinatal period
780	General symptoms
781	Symptoms involving nervous and musculoskeletal systems
782	Symptoms involving skin and other integumentary tissue
783	Symptoms concerning nutrition, metabolism, and development
784	Symptoms involving head and neck
785	Symptoms involving cardiovascular system
786	Symptoms involving respiratory system and other chest symptoms
787	Symptoms involving digestive system
788	Symptoms involving urinary system
789	Other symptoms involving abdomen and pelvis
790	Nonspecific findings on examination of blood
791	Nonspecific findings on examination of urine

792	Nonspecific abnormal findings in other body substances
793	Nonspecific abnormal findings on radiological and other examination of body structure
794	Nonspecific abnormal results of function studies
795	Nonspecific abnormal histological and immunological findings
796	Other nonspecific abnormal findings
797	Senility without mention of psychosis
798	Sudden death, cause unknown
799	Other ill-defined and unknown causes of morbidity and mortality
800	Fracture of vault of skull
801	Fracture of base of skull
802	Fracture of face bones
803	Other and unqualified skull fractures
804	Multiple fractures involving skull or face with other bones
805	Fracture of vertebral column without mention of spinal cord lesion
806	Fracture of vertebral column with spinal cord lesion
807	Fracture of rib(s), sternum, larynx, and trachea
808	Fracture of pelvis
809	Ill-defined fractures of bones of trunk
810	Fracture of clavicle
811	Fracture of scapula
812	Fracture of humerus
813	Fracture of radius and ulna
814	Fracture of carpal bone(s)
815	Fracture of metacarpal bone(s)
816	Fracture of one or more phalanges of hand
817	Multiple fractures of hand bones
818	Ill-defined fractures of upper limb
819	Multiple fractures involving both upper limbs, and upper limb with rib(s) and sternum
820	Fracture of neck of femur
821	Fracture of other and unspecified parts of femur
822	Fracture of patella
823	Fracture of tibia and fibula
824	Fracture of ankle
825	Fracture of one or more tarsal and metatarsal bones
826	Fracture of one or more phalanges of foot
827	Other, multiple, and ill-defined fractures of lower limb
828	Multiple fractures involving both lower limbs, lower with upper limb, and lower limb(s) with rib(s) and sternum
829	Fracture of unspecified bones
830	Dislocation of jaw
831	Dislocation of shoulder
832	Dislocation of elbow
833	Dislocation of wrist
834	Dislocation of finger
835	Dislocation of hip
836	Dislocation of knee
837	Dislocation of ankle

838	Dislocation of foot
839	Other, multiple, and ill-defined dislocations
840	Sprains and strains of shoulder and upper arm
841	Sprains and strains of elbow and forearm
842	Sprains and strains of wrist and hand
843	Sprains and strains of hip and thigh
844	Sprains and strains of knee and leg
845	Sprains and strains of ankle and foot
846	Sprains and strains of sacroiliac region
847	Sprains and strains of other and unspecified parts of back
848	Other and ill-defined sprains and strains
850	Concussion
851	Cerebral laceration and contusion
852	Subarachnoid, subdural, and extradural hemorrhage, following injury
853	Other and unspecified intracranial hemorrhage following injury
854	Intracranial injury of other and unspecified nature
860	Traumatic pneumothorax and hemothorax
861	Injury to heart and lung
862	Injury to other and unspecified intrathoracic organs
863	Injury to gastrointestinal tract
864	Injury to liver
865	Injury to spleen
866	Injury to kidney
867	Injury to pelvic organs
868	Injury to other intra-abdominal organs
869	Internal injury to unspecified or ill-defined organs
870	Open wound of ocular adnexa
871	Open wound of eyeball
872	Open wound of ear
873	Other open wound of head
874	Open wound of neck
875	Open wound of chest (wall)
876	Open wound of back
877	Open wound of buttock
878	Open wound of genital organs (external), including traumatic amputation
879	Open wound of other and unspecified sites, except limbs
880	Open wound of shoulder and upper arm
881	Open wound of elbow, forearm, and wrist
882	Open wound of hand except finger(s) alone
883	Open wound of finger(s)
884	Multiple and unspecified open wound of upper limb
885	Traumatic amputation of thumb (complete) (partial)
886	Traumatic amputation of other finger(s) (complete) (partial)
887	Traumatic amputation of arm and hand (complete) (partial)
890	Open wound of hip and thigh
891	Open wound of knee, leg [except thigh], and ankle

892	Open wound of foot except toe(s) alone
893	Open wound of toe(s)
894	Multiple and unspecified open wound of lower limb
895	Traumatic amputation of toe(s) (complete) (partial)
896	Traumatic amputation of foot (complete) (partial)
897	Traumatic amputation of leg(s) (complete) (partial)
900	Injury to blood vessels of head and neck
901	Injury to blood vessels of thorax
902	Injury to blood vessels of abdomen and pelvis
903	Injury to blood vessels of upper extremity
904	Injury to blood vessels of lower extremity and unspecified sites
905	Late effects of musculoskeletal and connective tissue injuries
906	Late effects of injuries to skin and subcutaneous tissues
907	Late effects of injuries to the nervous system
908	Late effects of other and unspecified injuries
909	Late effects of other and unspecified external causes
910	Superficial injury of face, neck, and scalp except eye
911	Superficial injury of trunk
912	Superficial injury of shoulder and upper arm
913	Superficial injury of elbow, forearm, and wrist
914	Superficial injury of hand(s) except finger(s) alone
915	Superficial injury of finger(s)
916	Superficial injury of hip, thigh, leg, and ankle
917	Superficial injury of foot and toe(s)
918	Superficial injury of eye and adnexa
919	Superficial injury of other, multiple, and unspecified sites
920	Contusion of face, scalp, and neck except eye(s)
921	Contusion of eye and adnexa
922	Contusion of trunk
923	Contusion of upper limb
924	Contusion of lower limb and of other and unspecified sites
925	Crushing injury of face, scalp, and neck
926	Crushing injury of trunk
927	Crushing injury of upper limb
928	Crushing injury of lower limb
929	Crushing injury of multiple and unspecified sites
930	Foreign body on external eye
931	Foreign body in ear
932	Foreign body in nose
933	Foreign body in pharynx and larynx
934	Foreign body in trachea, bronchus, and lung
935	Foreign body in mouth, esophagus, and stomach
936	Foreign body in intestine and colon
937	Foreign body in anus and rectum
938	Foreign body in digestive system, unspecified
939	Foreign body in genitourinary tract

940	Burn confined to eye and adnexa
941	Burn of face, head, and neck
942	Burn of trunk
943	Burn of upper limb, except wrist and hand
944	Burn of wrist(s) and hand(s)
945	Burn of lower limb(s)
946	Burns of multiple specified sites
947	Burn of internal organs
948	Burns classified according to extent of body surface involved
949	Burn, unspecified
950	Injury to optic nerve and pathways
951	Injury to other cranial nerve(s)
952	Spinal cord injury without evidence of spinal bone injury
953	Injury to nerve roots and spinal plexus
954	Injury to other nerve(s) of trunk excluding shoulder and pelvic girdles
955	Injury to peripheral nerve(s) of shoulder girdle and upper limb
956	Injury to peripheral nerve(s) of pelvic girdle and lower limb
957	Injury to other and unspecified nerves
958	Certain early complications of trauma
959	Injury, other and unspecified
960	Poisoning by antibiotics
961	Poisoning by other anti-infectives
962	Poisoning by hormones and synthetic substitutes
963	Poisoning by primarily systemic agents
964	Poisoning by agents primarily affecting blood constituents
965	Poisoning by analgesics, antipyretics, and antirheumatics
966	Poisoning by anticonvulsants and anti-Parkinsonism drugs
967	Poisoning by sedatives and hypnotics
968	Poisoning by other central nervous system depressants and anesthetics
969	Poisoning by psychotropic agents
970	Poisoning by central nervous system stimulants
971	Poisoning by drugs primarily affecting the autonomic nervous system
972	Poisoning by agents primarily affecting the cardiovascular system
973	Poisoning by agents primarily affecting the gastrointestinal system
974	Poisoning by water, mineral, and uric acid metabolism drugs
975	Poisoning by agents primarily acting on the smooth and skeletal muscles and respiratory system
976	Poisoning by agents primarily affecting skin and mucous membrane, ophthalmological, otorhinolaryngological, and dental drugs
977	Poisoning by other and unspecified drugs and medicinals
978	Poisoning by bacterial vaccines
979	Poisoning by other vaccines and biological substances
980	Toxic effect of alcohol
981	Toxic effect of petroleum products
982	Toxic effect of solvents other than petroleum-based
983	Toxic effect of corrosive aromatics, acids, and caustic alkalis
984	Toxic effect of lead and its compounds (including fumes)
985	Toxic effect of other metals

986	Toxic effect of carbon monoxide
987	Toxic effect of other gases, fumes, or vapors
988	Toxic effect of noxious substances eaten as food
989	Toxic effect of other substances, chiefly nonmedicinal as to source
990	Effects of radiation, unspecified
991	Effects of reduced temperature
992	Effects of heat and light
993	Effects of air pressure
994	Effects of other external causes
995	Certain adverse effects, not elsewhere classified
996	Complications peculiar to certain specified procedures
997	Complications affecting specified body systems, not elsewhere classified
998	Other complications of procedures, not elsewhere classified
999	Complications of medical care, not elsewhere classified

Πίνακας 8: Κωδικοί ICD-9 των ασθενειών που μελετήθηκαν