ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΙΙ: ΑΝΑΛΥΣΗΣ, ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΑΝΑΠΤΥΞΗΣ ΔΙΕΡΓΑΣΙΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ

ΜΟΝΑΔΑ ΑΥΤΟΜΑΤΗΣ ΡΥΘΜΙΣΗΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Διερεύνηση βιολογικών πληροφοριών σε σχέση με την τοξικότητα: εφαρμογή σε πρωτεϊνικά δεδομένα νανοσωματιδίων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΛΕΞΑΝΔΡΑΣ ΡΟΥΣΣΗ

Επιβλέπων καθηγητής:

Χαράλαμπος Σαρίμβεης

Αθήνα, Σεπτέμβριος 2015

Τίτλος θέματος:

«Διερεύνηση βιολογικών πληροφοριών σε σχέση με την τοξικότητα: εφαρμογή σε πρωτεϊνικά δεδομένα νανοσωματιδίων»

Περίληψη

Η διπλωματική αυτή εργασία εστιάζει στην επιστημονική περιοχή της Βιοπληροφορικής και πιο ειδικά της Βιοστατιστικής με εφαρμογή στην νανοτεχνολογία, και συγκεκριμένα σε μια ιδιαίτερα αναπτυσσόμενη τάξη δεδομένων γνωστή ως πρωτεϊνικό στέμμα νανοϋλικών. Οι ιδιότητες αυτών των δεδομένων έχουν καταγραφεί με τεχνικές πρωτεϊνικής συσταδοποίησης με απώτερο σκοπό την πρόβλεψη της τοξικότητας των νανοϋλικών. Στην παρούσα εργασία διερευνούμε την βιολογική τους πληροφορία με χρήση όρων γονιδιακής οντολογίας (Gene Ontology, GO). Συγκεκριμένα μια σειρά μαθηματικών αλγορίθμων εφαρμόστηκαν σε τρία διαφορετικά σύνολα δεδομένων πρωτεϊνικής έκφρασης με στόχο την εύρεση των σχέσεων μεταξύ τους, την ταξινόμηση και ομαδοποίησή τους με κριτήριο τις λειτουργικές τους ιδιότητες και την ανάλυση υπερέκφρασης τους στο περιβάλλον που εξετάζονται. Για την ανάλυση που παρουσιάζεται χρησιμοποιήσαμε διαθέσιμους αλγόριθμους του ανοικτού πηγαίου κώδικα λογισμικού Bioconductor στο περιβάλλον της στατιστικής γλώσσας προγραμματισμού R.

Αρχικά μετατρέψαμε τις ταυτότητες των πρωτεϊνών/γονιδίων σε GO ταυτότητες βιολογικών διαδικασιών ή και βιολογικών μονοπατιών ώστε να δημιουργήσουμε GO γραφήματα, τα οποία απεικονίζουν τις σχέσεις των GO όρων και κατ' επέκταση των γονιδίων μεταξύ τους. Έπειτα ταξινομήσαμε τα γονίδια στις λειτουργικές τους κατηγορίες και υπολογίσαμε τις ομοιότητες τους. Τέλος, με χρήση διαφορετικών βιβλιοθηκών πραγματοποιήθηκε η ανάλυση εμπλουτισμού κατά την οποία εντοπίσαμε ποιές πρωτεΐνες, και κατά συνέπεια γονίδια, δραστηριοποιούνται περισσότερο από το σύνηθες κατά την είσοδο νανοσωματιδίων στον ανθρώπινο οργανισμό. Τα αποτελέσματα έδειξαν ότι τα γονίδια που υπερεκφράζονται σε αυτές τις συνθήκες, και δημιουργούν την βιολογική ταυτότητα του νανοσωματιδίου, αφορούν κυτταρικές λειτουργίες που σε μεγάλο βαθμό σχετίζονται με την τοξικότητα. Οι κυτταρικές αυτές λειτουργίες είναι η ενεργοποίηση του συμπληρώματος, η κυτταρική συσχέτιση, η μεταφορά λιπιδίων, η ενεργοποίηση και απόκριση του ανοσοποιητικού συστήματος καθώς και η πήξη του αίματος.

Λέξεις-Κλειδιά

Βιοστατιστική, πρωτεϊνικό στέμμα νανοϋλικών, τοξικότητα, δεδομένα πρωτεϊνικής έκφρασης, Gene Ontology, Bioconductor, R, ανάλυση εμπλουτισμού

Thesis title:

«Exploration of biological information related to toxicity: application to nanoparticle proteomics data»

Abstract

This diploma thesis focused on the scientific area of Bioinformatics and in particular Biostatistics and its application to nanotechnology, and specifically on the data class known as protein corona of nanomaterials which is a scientific area of increasing interest. The properties of these data have been recorded by protein clustering techniques in order to predict the toxicity of nanomaterials. In this work, it is their biological information that is investigated by using gene ontology terms (Gene Ontology, GO). In particular, a series of mathematical algorithms were applied to three different proteomic data sets in order to find the data relationships, their classification and grouping based on their functional properties and the over-expression analysis of proteins in the environment they are being tested. For the analysis presented, available algorithms of the open source software Bioconductor were used in the environment of statistical programming language R.

Firstly protein and gene identifiers were converted to GO biological function identifiers, in order to create GO graphs which illustrate the relationship between GO terms and hence between proteins/genes. Then, we classified genes into their functional classes and calculated their similarities. Finally, enrichment analysis was performed using different libraries to identify which proteins and therefore genes are more than usually expressed on the inlet conditions of a nanoparticle in the human organism. Our results showed that the estimated overexpressed genes which reveal the biological identity of a nanoparticle, are frequently related to cellular functions associated with toxicity. These cellular functions are: activation of complement, cell association, lipid transport, activation and response of the immune system and blood coagulation.

Keywords

Biostatistics, protein corona of nanomaterials, toxicity, protein expression data sets, Gene Ontology, Bioconductor, R, enrichment analysis

Περιεχόμενα

Λίστα Πινάκων

Λίστα Σχημάτων

Πρόλογος και ευχαριστίες

Η διπλωματική αυτή εργασία έχει ως αντικείμενο μελέτης τις βιολογικές πληροφορίες που περιέχονται σε πρωτεϊνικά δεδομένα νανοσωματιδίων χρυσού και τη συνάφεια τους με την τοξικότητα. Βασικό εργαλείο επίτευξης αυτής της διερεύνησης ήταν ο Bioconductor, ένα λογισμικό ανοικτού κώδικα που παρέχει μεθόδους για την ανάλυση και την κατανόηση δεδομένων γονιδιακής έκφρασης, βασιζόμενος στη στατιστική γλώσσα προγραμματισμού R. Τα δεδομένα διατέθηκαν από την ερευνητική δημοσίευση του Carl D. Walkey και των συνεργατών του, αλλά και μέρος τους από την ερευνητική ομάδα της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών ΕΜΠ.

Θα ήθελα να ευχαριστήσω τον επιβλέποντά μου, αναπληρωτή καθηγητή Χαράλαμπο Σαρίμβεη για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα εργασία σε ένα θέμα μεγάλου για μένα ενδιαφέροντος, την βοήθεια και καθοδήγηση του καθ’ όλη τη διάρκεια εκπόνησης αυτής της εργασίας.

Ευχαριστώ θερμά την ερευνήτρια, της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής, Γεωργία Τσιλίκη για την πολύτιμη βοήθεια της, την καθολική στήριξή της και τις συμβουλές της στις δυσκολίες που αντιμετώπισα καθ’ όλη τη διάρκεια εκπόνησης αυτής της εργασίας.

Επίσης ευχαριστώ τα μέλη της ερευνητικής ομάδας της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής για την παροχή των δεδομένων, που έκαναν πραγματοποιήσιμη αυτή την εργασία.

1 Εισαγωγή

Η νανοεπιστήμη αναγνωρίζεται ως ένα πεδίο της επιστήμης πολλά υποσχόμενο καθώς αποτελεί εργαλείο για να ξεπεραστούν αρκετές αδυναμίες σε ποικίλα επιστημονικά πεδία, όπως η φυσική, η βιολογία, η χημεία και η επιστήμη υλικών. Τα επιτεύγματα που έχουν σημειωθεί στον τομέα των νανοεπιστημών οφείλονται στις αλλαγές των ιδιοτήτων των υλικών καθώς το μέγεθος τους μειώνεται στην τάξη των νανόμετρων. Έτσι, τα υλικά αποκτούν νέες μηχανικές, χημικές, ηλεκτρικές, οπτικές, και μαγνητικές ιδιότητες.

Μια από τις πιο συναρπαστικές προοπτικές είναι η τεχνολογία των νανοσωματίδιων (nanoparticles), η οποία χρησιμοποιείται σήμερα για να λύσει πολλά περίπλοκα τεχνικά προβλήματα κυρίως στους τομείς της ιατρικής και της βιοιατρικής. Η εφαρμογή της τεχνολογίας των νανοσωματιδίων στην ιατρική προκύπτει από την ικανότητα τους να επεμβαίνουν στον κυτταρικό μηχανισμό και να έχουν πρόσβαση σε απρόσιτους στόχους όπως ο εγκέφαλος λόγω του μικρού τους μεγέθους. Κατά συνέπεια έχουν εφαρμογή σε διάφορους κλάδους της επιστήμης της βιοιατρικής όπως η χορήγηση φαρμάκων, η μεταφορά γονιδίων, η επιδιόρθωση των ιστών, η θεραπεία του καρκίνου, η διάγνωση ασθενειών και θεραπεία τους, η υπερθερμία κ.ά. [1]

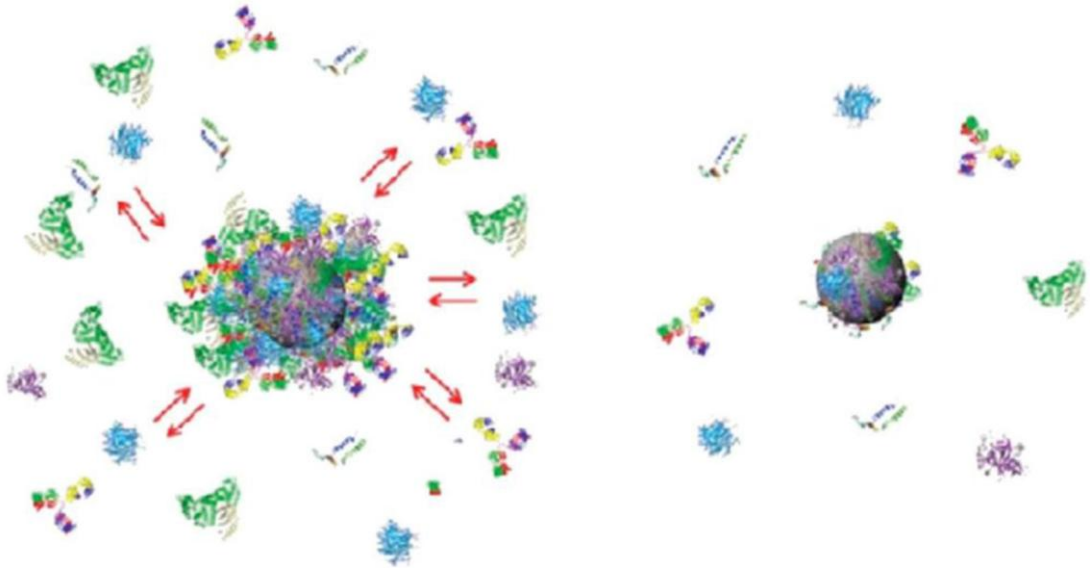
Τα νανοσωματίδια είναι σωματίδια με τουλάχιστον μια διάσταση μικρότερη από ένα εκατομμυριοστό του μέτρου (μικρό/micro) και δυνητικά όσο μικρά όσο τα άτομα ή τα μόρια. Μπορούν να έχουν άμορφη ή κρυσταλλική μορφή και οι επιφάνειες τους μπορούν να λειτουργήσουν ως μεταφορείς για σταγονίδια ή αέρια. Η κατηγοριοποίησης τους γίνεται με κριτήριο την διάμετρό τους. [2]

Τα νανοσωματίδια μπορούν να εισέλθουν στον ανθρώπινο σώμα μέσω διαφορετικών οδών, όπως με εισπνοή, με κατάποση είτε με απορρόφηση μέσω του δέρματος. Ανεξάρτητα από τον τρόπο εισαγωγής τους, το βιολογικό υγρό θα τα περικυκλώσει μόλις εισέλθουν σε ένα βιολογικό περιβάλλον. [3]

1.1 Πρωτεϊνικό στέμμα νανοσωματιδίων

Τα νανοσωματίδια όταν έρχονται σε επαφή με βιολογικό περιβάλλον, επιδιώκουν να μειώσουν την ενεργειακά υψηλή στάθμη της επιφάνειά τους προσροφώντας βιομόρια. Αυτό έχει ως αποτέλεσμα να δημιουργείται ένα σύνθετο στρώμα βιομορίων που καλύπτει την επιφάνεια τους. Πιο συγκεκριμένα, όταν ένα νανοσωματίδιο εκτίθεται σε ένα βιολογικό μέσο, λαμβάνουν χώρα φυσικές και χημικές αλληλεπιδράσεις μεταξύ της επιφάνειας του νανοσωματιδίου και διαφορετικών βιολογικών συστατικών του μέσου όπως, πρωτεΐνες, πεπτίδια και γλυκολιπίδια. Εξαιτίας των αλληλεπιδράσεων, σχηματίζεται μια διεπιφάνεια βιολογικού μέσου και νανοσωματιδίου και έτσι έρχονται σε επαφή. Η σχηματισμένη διεπιφάνεια καλύπτει την επιφάνεια του νανοσωματιδίου, τροποποιώντας έτσι την ποιότητα του και προσδίδοντάς του μια ταυτότητα μέσα στο βιολογικό πλαίσιο. Ένα υποσύνολο αυτών των πρωτεϊνών θα απορροφηθούν στην επιφάνεια του, δημιουργώντας ένα πρωτεϊνικό στέμμα (protein corona). Το πρωτεϊνικό στέμμα είναι μια «βιολογική ταυτότητα» του νανοσωματιδίου, που διαφέρει από την «συνθετική ταυτότητά» του. Η βιολογική ταυτότητα καθορίζει τις κυτταρικές αποκρίσεις όπως την κυτταρική πρόσληψη, την κινητική, την σηματοδότηση, την συσσώρευση, την μεταφορά και την τοξικότητα. Η βιολογική ταυτότητα είναι η μορφή του νανοσωματιδίου που «βλέπουν» τα συστατικά του βιολογικού συστήματος, και γι’ αυτό και καθορίζει την βιολογική του συμπεριφορά. Επίσης το πρωτεϊνικό στέμμα παρέχει πληροφορίες για την διεπιφάνεια μεταξύ του νανοσωματιδίου και του βιολογικού μέσου. [1]

Το πρωτεϊνικό στέμα μπορεί να υπάρξει σε δυο διαφορετικές μορφές στην επιφάνεια των νανοσωματιδίων, που καθορίζεται από τους τύπους των διαμορφωμένων στρωμάτων. Οι μορφές αυτές ονομάζονται «soft» και «hard» στέμματα, τα οποία αποτελούνται από χαλαρά δεσμευμένες πρωτεΐνες με μικρό χρόνο ζωής και ισχυρά δεσμευμένες πρωτεΐνες με μεγάλο χρόνο ζωής αντίστοιχα. (Σχήμα 1.1) Η σύσταση του πρωτεϊνικού στέμματος εξαρτάται από τις φυσικοχημικές ιδιότητες (δηλαδή σύσταση, μέγεθος, σχήμα, και ιδιότητες επιφάνειας) των νανοσωματιδίων και τα χαρακτηριστικά του βιολογικού περιβάλλοντος στο οποίο εκτίθεται. Η μελέτη και κατανόηση του πρωτεϊνικού στέμματος που αναθέτει την βιολογική ταυτότητα στο νανοσωματίδιο είναι σημαντική διότι έχει σημαντικές επιπτώσεις στον τομέα της νανοιατρικής, όπου εξετάζονται τοξικολογικές και φυσιολογικές αποκρίσεις των νανοσωματιδίων.



Σχήμα 1.1: Σχηματική αναπαράσταση της δομής πρωτεΐνης-νανοσωματιδίων στο πλάσμα του αίματος, η οποία επιβεβαιώνει τις διαφορετικές μορφές τους (ένα εξωτερικό ασθενώς αλληλεπιδρών στρώμα πρωτεΐνης που αναπαρίσταται με τα κόκκινα βέλη και ένα ισχυρό αργά ανταλλασσόμενο στέμμα πρωτεϊνών στα δεξιά της εικόνας) **[1]**

Η σύσταση του πρωτεϊνικού στέμματος είναι μοναδική για κάθε νανοσωματίδιο και επηρεάζεται από πολλές παραμέτρους όπως οι φυσικοχημικές ιδιότητες των νανοσωματιδίων και τα χαρακτηριστικά του περιβάλλοντος. Ο σχηματισμός του προκαλεί μια ελάττωση της ενέργειας της επιφάνειας του νανοσωματιδίου καθώς και της τοξικότητας του. Γενικότερα το πρωτεϊνικό στέμμα αλλάζει το μέγεθος, τη χημεία και τη φόρτιση της επιφάνειας του νανοσωματιδίου, επηρεάζοντας έτσι την πρόσληψη και την βιοκατανομή του κυττάρου. Το πρωτεινικό στέμμα σχηματίζεται στη διεπιφάνεια βιολογικού μέσου και νανοσωματιδίου με τη βοήθεια δυνάμεων ,όπως υδροδυναμικών, ηλεκτοδυναμικών, ηλεκτροστατικών, στερικών και δυνάμεων γεφύρωσης πολυμερών και διαλυτών. Αυτές καθορίζουν τη δομή του στέμματος που σχηματίζεται στη διεπιφάνεια, η οποία μπορεί να καθοριστεί σε φυσιολογικό περιβάλλον ή αφού απομονωθεί από αυτό. Διαφορετικές παράμετροι του πρωτεινικού στέματος όπως το πάχος του, η ταυτότητά του, η πυκνότητά του μπορούν να αναλυθούν και να ποσοτικοποιηθούν χρησιμοποιώντας διάφορα αναλυτικά εργαλεία. Για να κατανοήσουμε όμως καλύτερα τη συμπεριφορά του πρωτεϊνικού στέμματος θα πρέπει να γνωρίζουμε το ρόλο των πρωτεϊνών και τη σημασία τους στο κύκλο της κυτταρικής ζωής. [1]

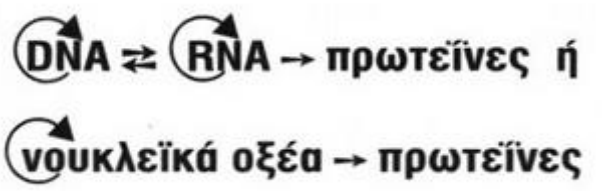
1.2 Μετάφραση: DNA→RNA→πρωτείνες

Σε κυτταρικό επίπεδο, η ζωή είναι ένα δίκτυο μοριακών αντιδράσεων, που μπορεί να οργανωθεί με ανώτερης τάξης διασυνδεδεμένα μονοπάτια. Τα μόρια συνθέτονται, διασπώνται, μεταφέρονται από την μια τοποθεσία στην άλλη, και συγκεντρώνονται σε ομάδες ή υψηλότερης τάξης δομές με άλλα μόρια. Εντατικές έρευνες της κυτταρικής σήμανσης, της κινητικότητας και άλλων πτυχών της κυτταρικής βιολογίας, σε συνδυασμό με την ανάπτυξη καταλόγων των ανθρώπινων γονιδίων και των πρωτεϊνικών προϊόντων τους, παρέχουν τη δυνατότητα να περιγράφονται πολλές κυτταρικές διεργασίες με λεπτομέρεια. Τηρώντας συνέπεια στην μορφή του κάθε βιοχημικού μονοπατιού, συγχωνεύονται όλες αυτές οι διεργασίες σε μια βάση δεδομένων, η οποία συνδέει τις ανθρώπινες πρωτεΐνες με τις μοριακές λειτουργίες τους. Έτσι δημιουργείται ένας πόρος που λειτουργεί ταυτόχρονα ως ένα αρχείο βιολογικών διεργασιών και ως έναν εργαλείο για την ανακάλυψη απροσδόκητων λειτουργικών σχέσεων σε δεδομένα γονιδιακής έκφρασης. [4]

Τα μακρομόρια (DNA,RNA,πρωτεΐνες) καθορίζουν τη δομή των κυττάρων και ελέγχουν και κυβερνούν στις περισσότερες δραστηριότητες της ζωής. Το DNA ενός οργανισμού είναι ο μοριακός «σκληρός δίσκος» που περιέχει αποθηκευμένες ακριβείς οδηγίες οι οποίες καθορίζουν τη δομή και τη λειτουργία του οργανισμού. Ταυτόχρονα περιέχει την πληροφορία για τον αυτοδιπλασιασμό του, εξασφαλίζοντας έτσι τη μεταβίβαση των γενετικών οδηγιών από ένα κύτταρο στα θυγατρικά του και από έναν οργανισμό στους απογόνους του.

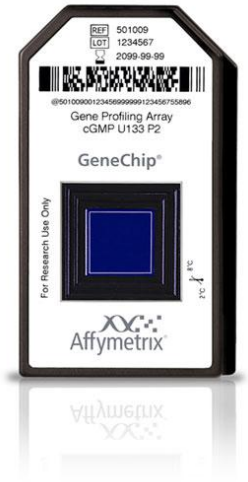
Το πρώτο βήμα για την έκφραση της πληροφορίας που υπάρχει στο DNA είναι η μεταφορά της στο RNA με τη διαδικασία της μεταγραφής. Το RNA μεταφέρει με τη σειρά του με τη διαδικασία της μετάφρασης την πληροφορία στις πρωτεΐνες που είναι υπεύθυνες για τη δομή και τη λειτουργία των κυττάρων και κατά επέκταση των οργανισμών.

Η γενετική πληροφορία είναι η καθορισμένη σειρά των αζωτούχων βάσεων των νουκλεοτιδίων, που αποτελούν τη βασική μονάδα των νουκλεϊκών οξέων DNA και RNA. Η πληροφορία υπάρχει σε τμήματα του DNA με συγκεκριμένη ακολουθία, τα γονίδια. Αυτά διαμέσου της μεταγραφής και της μετάφρασης, καθορίζουν τη σειρά των αμινοξέων στις πολυπεπτιδικές αλυσίδες. Οι πορείες της μεταγραφής και της μετάφρασης των γονιδίων αποτελούν τη γονιδιακή έκφραση. (Σχήμα 1.2) [5]



Σχήμα 1.2: Κεντρικό δόγμα της Μοριακής Βιολογίας **[5]**

Η γονιδιακή έκφραση αποκρυπτογραφείται με τη βοήθεια των DNA/RNA μικροσυστοιχιών που επιτρέπουν τη μέτρηση των επιπέδων έκφρασης χιλιάδων γονιδίων μόνο σε ένα πείραμα, δημιουργώντας έτσι μια αφθονία δεδομένων. Η τεχνολογία των DNA/RNA-μικροσυστοιχιών (microarrays) είναι μια αναπτυσσόμενη τεχνολογία, η οποία περιλαμβάνει χιλιάδες αλληλουχίες γονιδίων σε συγκεκριμένες θέσεις σε μια πλάκα (γυάλινη ή πλαστική) που ονομάζεται τσιπ γονιδίων. Ένα δείγμα που περιέχει DNA ή RNA βάφεται με φθορίζουσες χρωστικές και εκχύνεται στο τσιπ. Στη συνέχεια με τη βοήθεια λάζερ το συμπληρωματικό ζευγάρωμα των βάσεων του δείγματος και της αλληλουχίας γονιδίων του τσιπ παράγει φώς το οποίο μετράται από εικόνα του τσιπ που έχει προηγουμένως σαρωθεί. Οι περιοχές στο τσιπ που παράγουν φως αναγνωρίζουν τα γονίδια που εκφράζονται στο δείγμα. [6]



Σχήμα 1.3: Τσιπ σιλικόνης με το ανθρώπινο γονιδίωμα του κατασκευαστή Affymetrix

1.2.1 Πρωτεομική ανάλυση

Ο όρος πρωτέωμα αναφέρεται στο σύνολο των πρωτεϊνών, συμπεριλαμβανομένων των τροποποιήσεων που γίνονται σε αυτές, οι οποίες παράγονται από έναν οργανισμό ή ένα κυτταρικό σύστημα. Η πρωτεομική είναι μια μεγάλης κλίμακας ολοκληρωμένη

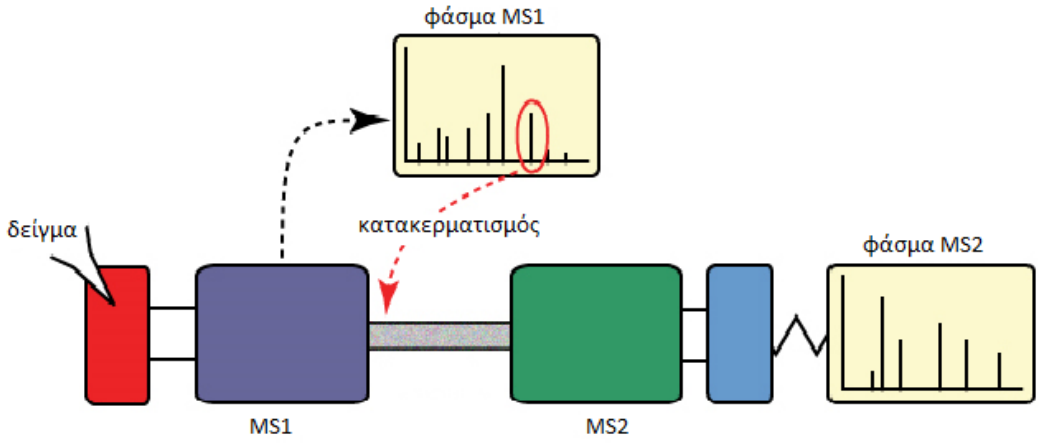
μελέτη ενός συγκεκριμένου πρωτεώματος που περιλαμβάνει πληροφορίες σχετικές με την αφθονία των πρωτεϊνών, τις τροποποιήσεις τους, τις αλληλεπιδράσεις τους με άλλα μόρια σε έναν συνολικό δίκτυο, με σκοπό την κατανόηση των κυτταρικών διεργασιών. Η πρωτεομική ανάλυση θεωρείται δυσκολότερη από την γονιδιωματική λόγω της δυναμικής φύσης των πρωτεϊνών σε αντίθεση με το γονιδίωμα το οποίο είναι στατικό. Η εφαρμογή της πρωτεομικής ανάλυσης έχει αυξηθεί τα τελευταία 15 χρόνια εξαιτίας των τεχνολογικών εξελίξεων στην φασματομετρία μάζας.

Η φασματομετρία μάζας (Mass Spectroscopy ή MS) είναι μια ευαίσθητη τεχνική που χρησιμοποιείται για την ανίχνευση, τον εντοπισμό και τον ποσοτικό προσδιορισμό μορίων με βάση τη μάζα και το φορτίο τους (m/z). Η ανάπτυξη μεθόδων, όπως ο ιονισμός μακρομορίων, ενεργοποίησε την μελέτη της δομής της πρωτεΐνης μέσω της MS. Οι επιστήμονες μέσω μαζών πρωτεϊνικών αποτυπωμάτων που ταιριάζουν σε πρωτεΐνες και πεπτίδια των βάσεων δεδομένων, μπορούν να προβλέψουν την ταυτότητα άγνωστων πρωτεϊνών. Η MS χρησιμοποιεί την αναλογία μάζας προς φορτίο των ιόντων για να αναγνωρίσει και να ποσοτικοποιήσει μόρια σε απλά και σύνθετα μίγματα. [7]

Όλα τα φασματόμετρα μάζας διαθέτουν μια πηγή ιόντων, έναν αναλυτή μάζας και έναν ανιχνευτή ιόντων. Αρχικά εισάγεται το δείγμα στο φασματόμετρο, τα μακρομόρια ιονίζονται και το φορτίο που λαμβάνουν επιτρέπει στο φασματόμετρο να επιταχύνει τα ιόντα στο υπόλοιπο σύστημα. Τα ιόντα συναντούν ηλεκτρικά ή μαγνητικά πεδία από τον αναλυτή μάζας, ο οποίος εκτρέπει τις διαδρομές των μεμονωμένων ιόντων με βάση τη μάζα και το φορτίο τους. Τα ιόντα που έχουν επιτυχώς εκτραπεί από τον αναλυτή μάζας, στη συνέχεια χτυπούν στον ανιχνευτή ιόντων. Τα φασματόμετρα μάζας συνδέονται με λογισμικό που αναλύει τα δεδομένα του ανιχνευτή και παράγει γραφικές παραστάσεις που οργανώνουν τα ανιχνευμένα ιόντα με βάση τον ξεχωριστό για το καθένα λόγο μάζας προς φορτίο (m/z) και την σχετική αφθονία (relative abundance) τους.

Μια ευρέως συνδυαστική γνωστή μέθοδος, για την ποιοτικό και ποσοτικό προσδιορισμό πρωτεϊνών, είναι η υγρή χρωματογραφία (Liquid Chromatography ή LC) με δυο διαδοχικές φασματομετρίες μάζας (tandem Mass Spectrometry/Mass Spectrometry ή tandem MS/MS). Η υγρή χρωματογραφία είναι η πιο κοινή μέθοδος διαχωρισμού για τη μελέτη βιολογικών δειγμάτων με MS/MS, διότι τα βιολογικά δείγματα είναι υγρά και μη πτητικά. Αφού το δείγμα διαχωριστεί μέσω της LC,

εισάγεται στο πρώτο φασματόμετρο. Τα διακριτά ιόντα μεγαλύτερου ενδιαφέροντος επιλέγονται με βάση την αναλογία m/z από τον πρώτο γύρο του MS και έπειτα κατακερματίζονται με χρήση μεθόδων διαχωρισμού όπως η σύγκρουση του με αδρανές αέριο. Αυτά τα κομμάτια χωρίζονται με βάση την ατομική αναλογία τους m/z στον δεύτερο γύρο του MS. Η μέθοδος MS/MS χρησιμοποιείται συνήθως για προσδιορισμό πρωτεϊνών και ολιγονουκλεοτιδίων, μιας και θραύσματά τους μπορούν να χρησιμοποιηθούν για να ταιριάξουν με αλληλουχίες νουκλεϊκών οξέων, οι οποίες εντοπίζονται σε βάσεις δεδομένων. Η πορεία της μεθόδου παρουσιάζεται στο Σχήμα 1.4. [7]



Σχήμα 1.4: Πορεία της μεθόδου MS/MS **[7]**

1.3 Ορισμός του προβλήματος και σκοπός

Το αποτέλεσμα ενός πειράματος πρωτεομικής ανάλυσης είναι μια λίστα ταυτοτήτων πρωτεϊνών, η οποία είναι το εναρκτήριο σημείο για έρευνα της βιολογικής πληροφορίας που εμπεριέχεται στα πειραματικά δεδομένα.

Στην παρούσα διπλωματική εργασία χρησιμοποιούνται μέθοδοι για να ερευνηθεί το βιολογικό προφίλ των δεδομένων πρωτεϊνών/γονιδίων μέσω της βάσης δεδομένων της γονιδιακής οντολογίας (GO). Στόχος μας είναι, στο περιβάλλον της γλώσσας προγραμματισμού R, να:

- απεικονίσουμε τις σχέσεις των γονιδίων μεταξύ τους μέσω γραφημάτων

- ταξινομήσουμε τα γονίδια στις λειτουργικές κατηγορίες που ανήκουν
- υπολογίσουμε τις λειτουργικές ομοιότητες των γονιδίων και με βάση αυτές να τα χωρίσουμε σε ομάδες
- εξετάσουμε ποια γονίδια υπερεκφράζονται στις συνθήκες που ορίζονται από το σύστημα μας με χρήση διαφορετικών βιβλιοθηκών
- εντοπίσουμε σε ποιες λειτουργίες συμμετέχουν τα υπερεκφρασμένα γονίδια και να ερευνήσουμε τη σχέση τους με την τοξικότητα

1.4 Διάρθρωση της εργασίας

Η διπλωματική εργασία συνεχίζει με το Κεφάλαιο 2 όπου παρουσιάζονται η μορφή και το περιεχόμενο των δεδομένων που μας διατίθεται. Στο κεφάλαιο 3 αναλύεται η έννοια της οντολογίας των GO όρων και το κατευθυνόμενο ακυκλικό γράφημα που απεικονίζει τις σχέσεις τους.

Στο Κεφάλαιο 4 γίνεται αναφορά στη στατιστική γλώσσα προγραμματισμού R, τα βασικά στοιχεία της, τις υπολογιστικές της δυνατότητες, καθώς και η χρήση των μεταφραστικών βιβλιοθηκών του Bioconductor, κατά την οποία οι GO ταυτότητες βιολογικών λειτουργιών μετατρέπονται σε άλλες ταυτότητες δεδομένων γονιδιακής έκφρασης.

Στο Κεφάλαιο 5 παρουσιάζονται ο στατιστικός έλεγχος υποθέσεων, αναλύεται το απλό γραμμικό μοντέλο, το γενικό γραμμικό μοντέλο και ο καθορισμός των συντελεστών τους με τις μεθόδους ελαχίστων τετραγώνων και μερικής γραμμικής παλινδρόμησης αντίστοιχα. Έτσι καταστρώνεται το μοντέλο που θα περιγράφει ικανοποιητικότερα την σχέση των περιεκτικοτήτων των πρωτεϊνών με την κυτταρική συσχέτιση των νανοσωματιδίων. Τέλος παρουσιάζεται η στατιστική τεχνική επικύρωσης του μοντέλου.

Στο κεφάλαιο 6 με χρήση των GO ταυτοτήτων που προκύπτουν από τις μεταφράσεις των δεδομένων μας καταστρώνονται τα GO γραφήματα, τα λειτουργικά προφίλ των ταυτοτήτων αυτών και η ομαδοποίηση των γονιδίων με βάση την λειτουργική ομοιότητα τους.

Στο κεφάλαιο 7 παρουσιάζεται η ανάλυση εμπλουτισμού γονιδίων είτε μέσω της προσέγγισης υπερεκπροσώπησης είτε μέσω της προσέγγισης συνολικού σκορ. Εδώ

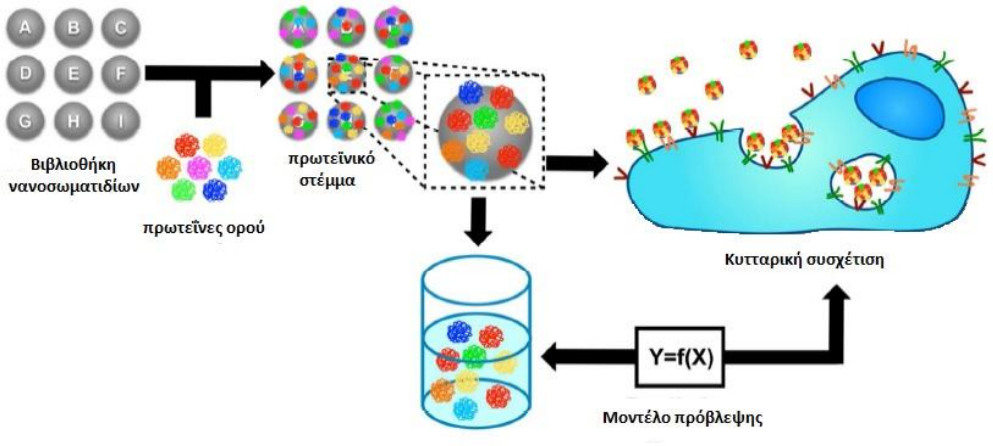
θα εντοπιστούν τα σημαντικότερα γονίδια της ανάλυσης μέσω της τιμής p value που προκύπτει από την ανάλυση αυτή και για τις δυο περιπτώσεις, τα οποία υπερεκφράζονται στις συνθήκες παρουσίας του νανοσωματιδίου στον ανθρώπινο οργανισμό.

Στο Κεφάλαιο 8 επιχειρείται η αποτίμηση των αποτελεσμάτων των αναλύσεων εμπλουτισμού. Αφού παρουσιαστούν οι εμπλουτισμένες λειτουργικές κατηγορίες στις οποίες συμμετέχουν τα γονίδια στην ερευνητική εργασία, συγκρίνονται με τα αποτελέσματα τα αναλύσεων του Κεφαλαίου 7. Στην συνέχεια εξετάζεται η σχέση των εμπλουτισμένων λειτουργικών κατηγοριών με την τοξικότητα.

2 Δεδομένα

Στην ερευνητική εργασία με τίτλο «Protein Corona Fingerprinting Predicts the Cell Association of Gold Nanoparticles» [8], χρησιμοποιήθηκε η σύσταση του αποτυπώματος του πρωτεϊνικού στέμματος για να προσδιορισθεί η κυτταρική συσχέτιση μιας βιβλιοθήκης με 105 μέλη χημικά ποικίλων νανοσωματιδίων χρυσού. Η κυτταρική συσχέτιση επιλέχθηκε ως μοντέλο βιολογικής αλληλεπίδρασης, εξαιτίας της συνάφειας της με φλεγμονώδεις αποκρίσεις, την βιοκατανομή και την τοξικότητα σε in vivo περιβάλλον. (Σχήμα 1.2)

Το συμπέρασμα της εργασίας αυτής είναι ότι το πρωτεϊνικό στέμμα ενός νανοσωματιδίου κωδικοποιεί βιολογικά πιο χρήσιμες πληροφορίες για το νανοσωματίδιο παρά οι φυσικές τους ιδιότητες. Επειδή οι αλληλεπιδράσεις νανοσωματιδίου-κυττάρου καθορίζουν τις κυτταρικές αποκρίσεις, το αποτύπωμα του πρωτεϊνικού στέμματος μπορεί να προβλέψει την ενεργοποίηση αλληλουχίας αντιδράσεων της ενδοκυτταρικής σηματοδότησης, της έκκρισης της κυτονίνης, της γονιδιακής έκφρασης, της τοξικότητας και κατά επέκταση της βιοκατανομής και της απόκρισης οργάνων. Το πρωτεϊνικό στέμμα μπορεί να εξελιχθεί σε ένα μέσο πρόβλεψης την αλληλεπίδρασης των νανοσωματιδίων με τα βιολογικά συστήματα.



Σχήμα 2.1: Από την βιβλιοθήκη νανοσωματιδίων στο μοντέλο πρόβλεψης της κυτταρικής συσχέτισης **[8]**

Κατά την πειραματική διαδικασία της παραπάνω εργασίας, δημιουργήθηκε μια βιβλιοθήκη από 105 διακριτά επιφανειακά τροποποιημένα νανοσωματίδια χρυσού διαμέτρου 15,30 και 60 nm. Οι 67 επιφανειακοί προσδέτες που χρησιμοποιήθηκαν

για την δημιουργία της βιβλιοθήκης περιλαμβάνουν μικρά μόρια, πολυμερή, πεπτίδια, επιφανειοδραστικές ουσίες και λιπίδια, που έχουν επιλεχθεί να μιμούνται την χημική σύσταση της επιφάνειας των πιο συχνά χρησιμοποιούμενων μορφών των νανοσωματιδίων.

Μετά την σύνθεση αυτή, κάθε μορφή νανοσωματιδίου επωάστηκε με μη αραιωμένο ανθρώπινο ορό για μια ώρα στους 37˚C. Έπειτα, καθαρίστηκε με φυγοκέντρηση ώστε να απομακρυνθούν οι αδέσμευτες πρωτεΐνες.

Το πρωτεϊνικό στέμμα που σχηματίστηκε γύρω από κάθε μορφή νανοσωματιδίου χαρακτηρίστηκε ποιοτικά χρησιμοποιώντας την τεχνική της ηλεκτροφόρησης σε πηκτή πολυ-ακριλαμιδίου και ποσοτικά με χρήση υψηλής ανάλυσης υγρής χρωματογραφίας και φασματομετρίας μάζας (LC-MS/MS). Κάθε μορφή νανοσωματιδίου απορρόφησε 71±22 ξεχωριστές πρωτεΐνες ορού.

Από όλη τη βιβλιοθήκη, 785 ξεχωριστές πρωτεΐνες ορού αναγνωρίστηκαν από την υγρή χρωματογραφία και τη φασματομετρία μάζας, από τις οποίες 129 ήταν κατάλληλες για σχετική ποσοτικοποίηση. Η σχετική περιεκτικότητα της κάθε μιας από τις πρωτεΐνες σε μια μορφή νανοσωματιδίου ορίζει το αποτύπωμα της πρωτεΐνης ορού για την μορφή αυτή. Η επί τοις % κατά βάρος (% w/w) σχετική περιεκτικότητα (relative abundance) της κάθε πρωτεΐνης σε μια δεδομένη μορφοποίηση του νανοσωματιδίου ορίζεται από την παρακάτω φόρμουλα :

$$RA(n)_{\%\left(\frac{w}{w}\right)} = \frac{SpC(n)}{\sum_{i=1}^{129} SpC(i)} \tag{2.1}$$

Η συνολική κυτταρική συσχέτιση (**y**), υπολογίζεται χρησιμοποιώντας τον συντελεστή ψευδο-συμμετοχής :

$$y = \frac{m_{cell}/m_{well}}{m_{cells}}, \tag{2.2}$$

όπου $\boldsymbol{m_{cell}}$ είναι η μάζα του συνολικού ατομικού χρυσού που συνδέεται με τα κύτταρα, $\boldsymbol{m_{well}}$ είναι η μάζα του συνολικού ατομικού χρυσού που συνδέεται με τα κύτταρα και αυτού που είναι ελεύθερο στο διάλυμα και $\boldsymbol{m_{cells}}$ είναι η συνολική μάζα του μαγνησίου ανά δείγμα.

Η πρόβλεψης της κυτταρικής συσχέτισης από τις πρωτεΐνες, πραγματοποιείται μέσω ενός λογαριθμικού γραμμικού μοντέλου. Ο καθορισμός του καταλληλότερου μοντέλου έγινε με τη μέθοδο μερικής παλινδρόμησης ελαχίστων τετραγώνων σε περιβάλλον διασταυρωμένης επικύρωσης (cross validation). Αυτή η μεθοδολογία όπως και όσες αναφέρονται παρακάτω παρουσιάζονται αναλυτικά στο Κεφάλαιο 5. Η ακρίβεια με τη οποία κάθε πρωτεΐνη προσαρμόζεται στο μοντέλο αυτό θα αποτελέσει το σημαντικότερο κριτήριο για την επιλογή των σημαντικότερων πρωτεϊνών/γονιδίων.

Τα δεδομένα που διατίθενται για την μελέτη της παραπάνω βιολογικής πληροφορίας είναι τρία διαφορετικά σύνολα.

Το πρώτο σύνολο δεδομένων (Πίνακας 2.1) περιέχεται σε μορφή πίνακα στο αρχείο «Scaled_RelAbund_and_LogNetCell.csv», το οποίο θα αναφέρεται στη συνέχεια ως Σύνολο Α και περιλαμβάνει:

- 129 Uniprot ταυτότητες πρωτεϊνών, πχ. P01024
- 84 ταυτότητες των μορφών των νανοσωματιδίων χρυσού, πχ. G15.AC όπου το γράμμα G φανερώνει το υλικό του νανοσωματιδίου που είναι ο χρυσός (Gold), το 15 είναι η διάμετρος του πυρήνα του νανοσωματιδίου σε νανόμετρα και το AC υποδηλώνει τον τροποποιητή της επιφάνειας του νανοσωματιδίου που είναι ο ενεργός άνθρακας(Activated Carbon).
- σχετικές περιεκτικότητες (relative abundance) $\boldsymbol{x_{ij}}$ της κάθε πρωτεΐνης σε κάθε μορφή νανοσωματιδίου
- εκτιμήσεις του γενικού γραμμικού μοντέλου της κυτταρική συσχέτισης $\mathbf{log_2}\,\widehat{\boldsymbol{y_i}}$ των πρωτεϊνών με το κύτταρο ως άθροισμα των απορροφημένων πυκνοτήτων της κάθε πρωτεΐνης

Πίνακας 2.1: Δείγμα δεδομένων Συνόλου Α

Ταυτότητα νανοσωματιδίου	**G15.AC**
$\mathbf{log_2}\,\widehat{\boldsymbol{y_i}}$	0.402466085
P01024	0.568470134
P01834	0.529881204

Το δεύτερο σύνολο δεδομένων περιέχεται στο αρχείο «OrderofProteins_PLS.csv», το οποίο θα αναφέρεται στη συνέχεια ως Σύνολο Β (Πίνακας 2.2), και περιλαμβάνει:

- Ταξινομημένη λίστα συμβολικών ονομασιών των γονιδίων ελλατούμενης σημαντικότητας με βάση τα αποτελέσματα της ανάλυσης του Carl D. Walkey και των συνεργατών του
- Τιμές του συντελεστή προσδιορισμού Q^2, οι οποίοι υπολογίζονται κατά τη διαδικασία επικύρωσης του μοντέλου.

Η ακρίβεια της προσαρμογής κάθε πρωτεΐνης στο μοντέλο μερικών ελαχίστων τετραγώνων προσδιορίζεται ποσοτικά με τον συντελεστή προσδιορισμού $Q.^2$ Τιμές του Q^2 πιο κοντά στη μονάδα, δηλώνουν ότι μια πρωτεΐνη προσδένεται πιο δυνατά στο νανοσωματίδιο, που αφορούν την συσχέτιση των πρωτεϊνών με το κύτταρο. Ο υπολογισμός των Q^2 τιμών έγινε με την μέθοδο της επαναλαμβανόμενης (iterative) μερικής παλινδρόμησης ελαχίστων τετραγώνων και με εφαρμογή της τεχνικής backwards elimination κατά την διαδικασία cross validation από τη Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής του Εθνικού Μετσοβείου Πολυτεχνείου.

Το τρίτο σύνολο δεδομένων περιέχεται στο αρχείο «ProteinList_VIPscores_fromSup.csv», το οποίο θα αναφέρεται στη συνέχεια ως Σύνολο Γ (Πίνακας 2.2) και περιλαμβάνει:

- Ταξινομημένη λίστα συμβολικών ονομασιών των γονιδίων ελλατούμενης σημαντικότητας με κριτήριο το μέγεθος της μεταβλητής σημαντικότητας για προβολή (Variable Importance in Projection ή ***VIP)***
- Λίστα των παραμέτρων b_j του γενικού γραμμικού μοντέλου για το κάθε γονίδιο
- Τιμές της μεταβλητής σημαντικότητας για προβολή ***VIP*** για κάθε γονίδιο

Αφού προσαρμοστεί το καταλληλότερο μοντέλο με τη διαδικασία της μερικής παλινδρόμησης των ελαχίστων τετραγώνων χρησιμοποιώντας όλο το σύνολο των παραμέτρων του μοντέλο, απομακρύνονται, με κριτήριο το μέγεθος ***VIP*** της κάθε πρωτεΐνης, οι παράμετροι των πρωτεϊνών με την μικρότερη σχέση με το μοντέλο. Με την απομάκρυνση των παραμέτρων αυτών δημιουργείται ένα νέο μοντέλο. Εδώ και πάλι χρησιμοποιήθηκαν τεχνικές διασταυρωμένης επικύρωσης. Η διαδικασία αυτή πραγματοποιήθηκε στα πλαίσια της ερευνητικής εργασίας του Walkey και των συνεργατών του.

Πίνακας 2.2: Δεδομένα Συνόλου Β και δεδομένα Συνόλου Γ

Σύνολο Β	
SYMBOL	Q^2
AMBP	1
HABP2	1
ITIH2	1
TTHY	1
ITIH1	1
CO3	0.6349
A1AT	0.6433
ITIH3	0.6527
CO4B	0.6465
A2AP	0.642
SPP24	0.6368
LUM	0.6362
PROC	0.654
A2MG	0.6522
ANGT	0.6578
ENPL	0.6742
PRG4	0.6684
CNDP1	0.7002
PROZ	0.7009
ANT3	0.6984
CD180	0.6904
CRP	0.6823
PLEK	0.6871
APOB	0.7158
FHR1	0.7135
KLKB1	0.7187
APOC1	0.7629
SEPP1	0.7664
HRG	0.7666
IGLL5	0.7607
HEP2	0.7583
APOE	0.7598
IGKC	0.7544
IGHG4	0.7699
CERU	0.7662
MUCB	0.7595
APOL1	0.7625
ZPI	0.7602
KNG1	0.7577
APOM	0.743

Σύνολο Γ		
SYMBOL	b_j	*VIP*
AMBP	0.0683	1.67
ITIH2	0.0586	1.66
ITIH1	0.0613	1.63
A1AT	0.1008	1.61
HABP2	0.0609	1.61
ITIH3	0.0658	1.51
CO4B	0.0549	1.51
TTHY	0.1025	1.51
A2AP	0.0412	1.5
SPP24	-0.0041	1.44
ANGT	0.0845	1.41
LUM	0.0666	1.38
CO3	-0.1102	1.37
A2MG	0.0158	1.26
ENPL	0.1244	1.21
APOB	-0.132	1.19
CNDP1	0.1161	1.19
ANT3	-0.0926	1.15
CD180	-0.0069	1.11
SEPP1	-0.0996	1.08
APOC1	0.1299	1.02
IGHG4	-0.1441	1
CERU	0.0757	1
PRG4	-0.0054	0.98
CRP	0.0052	0.98
KLKB1	-0.1077	0.98
HEP2	-0.051	0.97
FHR1	0.0644	0.97
HRG	-0.0647	0.94
CD5L	0.0826	0.93
VTNC	0.079	0.93
PROC	0.0401	0.93
ZPI	0.0625	0.92
MUCB	-0.007	0.92
APOL1	-0.0871	0.9
APOM	0.0529	0.9
KNG1	-0.049	0.87
APOE	0.0469	0.87
FA11	0.0679	0.85
PROZ	0.0119	0.82

HPTR	0.7434	KV302	-0.061	0.8
VTNC	0.7496	CO9	-0.0463	0.8
IGHM	0.7471	IGLL5	-0.0246	0.8
CD5L	0.7483	IGHM	-0.0402	0.79
FA11	0.7527	PLEK	0.0532	0.79
ITIH4	0.7612	APOA4	0.0828	0.78
PF4V	0.7575	TETN	0.0962	0.78
GELS	0.7558	GELS	0.0399	0.77
KV302	0.7544	IPSP	-0.0225	0.77
CALR	0.7635	APOC4	0.0535	0.77
FA5	0.7656	ITIH4	-0.0772	0.77
APOC4	0.7604	APOA1	0.1033	0.75
ANGI	0.7638	PF4V	0.0377	0.75
SAA4	0.7633	CBPN	0.0588	0.74
IPSP	0.7612	CALR	-0.0331	0.73
APOA1	0.7644	TSP4	-0.0391	0.73
C1QA	0.7808	FA9	0.028	0.72
TRFL	0.7792	ANGI	0.0715	0.7
CO9	0.7729	FA10	-0.004	0.7
FA10	0.7793	FA5	-0.0127	0.7
FA9	0.772	A1BG	-0.0985	0.7
C4BPA	0.772	HPTR	-0.0817	0.7
C4BPB	0.7723	COMP	-0.0381	0.69
TSP4	0.7758	AACT	0.0446	0.68
TETN	0.7976	SAA4	0.0125	0.68
CBPN	0.7946	LBP	-0.0314	0.67
AACT	0.7955	PROS	-0.0461	0.67
FA7	0.7957	TRFL	-0.0377	0.67
PROS	0.8009	IGKC	0.0226	0.67
LBP	0.7951	TSP1	0.1021	0.67
IGHG1	0.8103	C4BPB	0.037	0.66
COMP	0.8076	C4BPA	-0.0035	0.66
C1R	0.8051	C1QA	0.0791	0.66
A1BG	0.8174	FA7	0.027	0.64
TSP1	0.8236	C1R	0.0064	0.63
APOA4	0.8189	IGHG1	0.0646	0.6

Στα παραπάνω σύνολα εφαρμόζονται αλγόριθμοι για την επεξεργασία τους. Με χρήση μαθηματικών και στατιστικών μεθόδων απομονώνονται οι σημαντικές πληροφορίες που περιέχονται σε αυτά, με αποτελέσματα να εξάγεται η σημαντικότερη πληροφορία για το βιολογικό σύστημα που μελετάται.

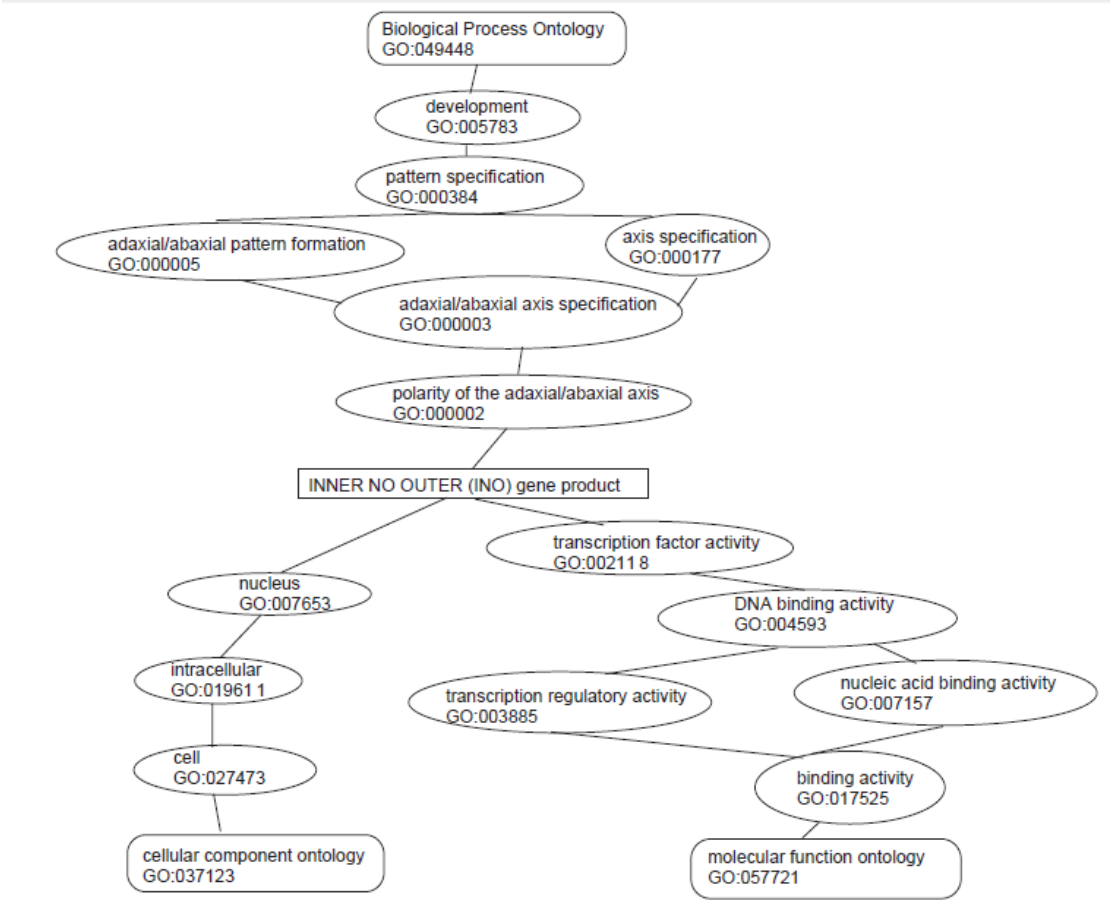
3 Οντολογία GO

Η Γονιδιακή Οντολογία (Gene Ontology ή GO) έχει στόχο την βιολογική ερμηνεία κυρίως γονιδιακών πειραμάτων (για παράδειγμα, πειραμάτων DNA και RNA μικροσυστοιχιών) αλλά και άλλων βιολογικών πειραμάτων (για παράδειγμα, πρωτεομικών). Η ερμηνεία αυτή πραγματοποιείται μέσω μιας βάσης δεδομένων μετάφρασης που δημιουργείται και συντηρείται από την κοινοπραξία επιστημόνων για την Γονιδιακή Οντολογία (Gene Ontology Consortium) [9], της οποίας ο βασικός ρόλος είναι να παράγει ένα ελεγχόμενο λεξιλόγιο (GO όρων) που μπορεί να εφαρμοστεί για όλους τους οργανισμούς, ακόμη και αν η γνώση των γονιδίων και πρωτεϊνικών ρόλων στα κύτταρα, συσσωρεύεται και αλλάζει. Η Γονιδιακή Οντολογία οργανώνεται γύρω από τρεις βασικές αρχές ή τρεις βασικές υποοντολογίες:

1. *Βιολογική Διεργασία (Biological process ή BP)*
   Περιγράφει το βιολογικό αποτέλεσμα που συνήθως περιλαμβάνει μια χημική ή φυσική μετατροπή, το οποίο επιτυγχάνεται με σύνολα μοριακών λειτουργιών. Κάποια παραδείγματα GO όρων που υπάγονται στις βιολογικές διεργασίες είναι η ανάπτυξη και συντήρηση των κυττάρων και η μετάφραση. Οι όροι που υπάγονται σε αυτή την οντολογία περιγράφουν βιολογικούς στόχους.
2. *Μοριακή λειτουργία (Molecular function ή MF)*
   Η μοριακή λειτουργία ορίζεται ως η βιοχημική δραστηριότητα (συμπεριλαμβανόμενων ειδικών συνδέσεων με συνδέτες ή δομές) ενός γονιδιακού προϊόντος. Η οντολογία αυτή περιγράφει λειτουργίες που πραγματοποιούνται από μεμονωμένα γονιδιακά προϊόντα, όπως η δραστηριότητα ενός μεταγραφικού παράγοντα σαν την ATP-άση.
3. *Κυτταρικό συστατικό (Cellular component-CC)*
   Αναφέρεται στην θέση στο κύτταρο όπου το γονιδιακό προϊόν είναι ενεργό. Περιγράφει υποκυτταρικές δομές, θέσεις και μακρομοριακά σύμπλοκα, όπως οι πυρήνες, η πυρηνική μεμβράνη, τα τελομερή κ.ά.

Οι GO όροι δεν είναι συμβατοί με όλους τους οργανισμούς. Ο όρος που περιγράφει την πυρηνική μεμβράνη ισχύει για τον ανθρώπινο οργανισμό ενώ δεν έχει υπόσταση στους μονοκύτταρους οργανισμούς που δεν διαθέτουν πυρηνική μεμβράνη. Η Γονιδιακή Οντολογία όμως δεν αποκλείει αυτούς τους όρους.[10]

Ένα δεδομένο γονιδιακό προϊόν μπορεί να αντιπροσωπεύει μια ή περισσότερες μοριακές λειτουργίες, μια ή περισσότερες βιολογικές λειτουργίες και να εμφανίζεται σε ένα ή περισσότερα κυτταρικά συστατικά. Ως συμπέρασμα από το Σχήμα 3.1, παρατηρείται ότι ένα γονίδιο δεν συνδέεται μόνο με τις μεταφράσεις του, δηλαδή τους GO όρους, αλλά συνδέεται και με τις μεταφράσεις των μεταφράσεων του, δηλαδή τους GO όρους που έχουν σχέση με αυτούς. Αυτές οι συνδέσεις δημιουργούν ένα δίκτυο όρων για κάθε γονίδιο, το οποίο ολοκληρώνεται στο μεγαλύτερο δίκτυο, το δίκτυο της οντολογίας GO.

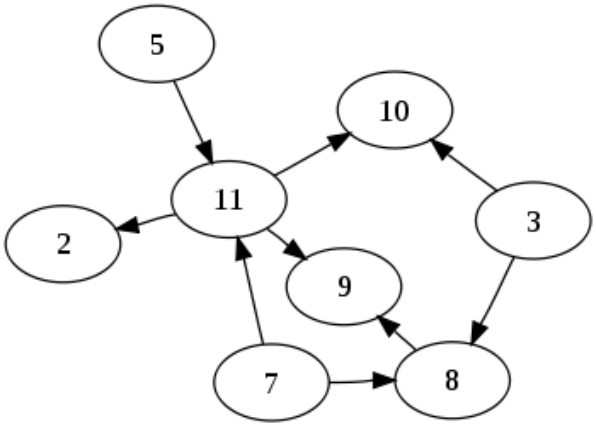


Σχήμα 3.1: Υποθετικό παράδειγμα της μετάφρασης σε GO όρους για το γονίδιο "INNER NO OUTER". Κάθε γονίδιο μεταφράζεται στις τρεις οντολογίες: MF, BP, CC **[11]**

Ένα σημαντικό σημείο είναι ότι οι πληροφορίες εδώ δεν είναι γραμμικές, παρόλο που υπάρχει ιεραρχική σχέση, δηλαδή υπάρχουν συσχετίσεις μεταξύ επιπέδων και ορών σε κάθε οντολογία. Ως αποτέλεσμα η κατάλληλη παρουσίαση για αυτές τις σχέσεις είναι ένα κατευθυνόμενο ακυκλικό γράφημα (Directed Acyclic Graph ή DAG). [12]

3.1 Κατευθυνόμενο ακυκλικό γράφημα

Στα μαθηματικά και στην πληροφορική, ένα κατευθυνόμενο ακυκλικό γράφημα είναι ένα γράφημα με μη κατευθυνόμενους κύκλους. Δηλαδή, σχηματίζεται από μια συλλογή από κόμβους (vertices/ nodes) και κατευθυνόμενες συνδέσεις (edges/ links), με κάθε σύνδεση να ενώνει έναν κόμβο με τον άλλο, έτσι ώστε να μην υπάρχει τρόπος να ξεκινήσουμε από έναν κόμβο ***v*** και ακολουθώντας μια αλληλουχία συνδέσεων να καταλήξουμε στον ίδιο κόμβο ***v***. [13]



Σχήμα 3.2: Μορφή κατευθυνόμενου ακυκλικού γραφήματος **[13]**

Το κατευθυνόμενο ακυκλικό γράφημα (Σχήμα 3.2) περιλαμβάνει τα παρακάτω στοιχεία:

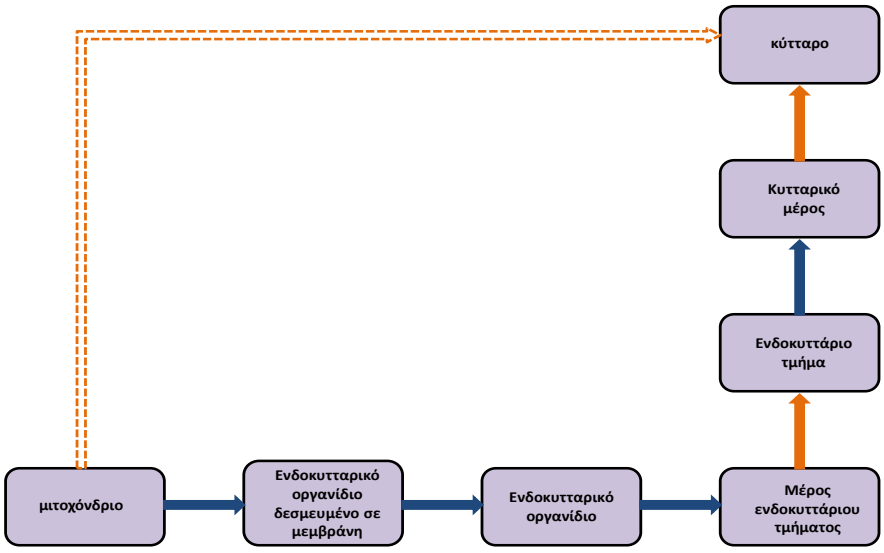
- Κόμβους: Κάθε κόμβος αντιπροσωπεύει ένα αντικείμενο ή ένα δεδομένο.
- Κατευθυνόμενες συνδέσεις: Μια κατευθυνόμενη σύνδεση (ή βέλος) από τον έναν κόμβο στον άλλο αντιπροσωπεύει κάποιου είδους σχέση μεταξύ των δυο κόμβων.
- Κόμβος ρίζα: Τουλάχιστον ένας κόμβος δεν έχει σύνδεση που να καταλήγει σε αυτόν.
- Κόμβοι φύλλα: Ένας ή περισσότεροι κόμβοι δεν θα διαθέτουν συνδέσεις που να ξεκινούν από αυτούς.

Αντίστοιχα με άλλες υπάρχουσες οντολογίες, η GO οντολογία έχει μια ιεραρχική δομή που σχηματίζει έναν κατευθυνόμενο ακυκλικό γράφημα (DAG). Για αυτό το γράφημα, θα χρησιμοποιηθούν οι έννοιες «παιδί» (child) και «γονιός» (parent). Εδώ

κάθε όρος «παιδί» μπορεί να έχει πολλούς «γονείς». Η σχέση «παιδί-γονιός» μεταξύ όρων (δηλαδή κόμβων σε ένα γράφημα) μπορεί να είναι δυο τύπων :

- «είναι ένα» («is a») : σημαίνει ότι το παιδί είναι μια περίπτωση του γονιού, π.χ. έναν μιτωτικό χρωμόσωμα είναι μια περίπτωση χρωμοσώματος.
- «μέρος του» («part of») : σημαίνει ότι το παιδί είναι ένα μέρος του γονιού, πχ. το τελομερές είναι ένα μέρος του χρωμοσώματος.

Οι όροι «παιδιά» μπορούν να έχουν διαφορετικούς τύπους σχέσεων με τους διαφορετικούς «γονείς» τους. Για δυο κόμβους GO, ***u*** και ***v***, θα πούμε ότι το ***u*** είναι «γονιός» του ***v***, αν και μόνο εάν το ***v*** έχει οποιαδήποτε τύπο σχέσης με το ***u***. Συνεπώς, ο κόμβος ***v*** είναι πιο συγκεκριμένος από τον κόμβο ***u***. Στην προκειμένη περίπτωση, το ***v*** είναι «παιδί» του ***u***. [11]



Σχήμα 3.3: Παράδειγμα σχέσεων GO όρων

Στο Σχήμα 3.3 παρουσιάζονται οι δυο τύποι σχέσεων των GO όρων. Με τις μπλε συνδέσεις εκπροσωπούνται οι σχέσεις «είναι ένα», ενώ με τις πορτοκαλί συνδέσεις οι σχέσεις «μέρος του». Το μιτοχόνδριο «είναι ένα» ενδοκυτταρικό οργανίδιο δεσμευμένο σε μεμβράνη, ενώ το κυτταρικό τμήμα είναι «μέρος του» κυττάρου.

Τέτοια γραφήματα κατασκευάζονται με τη βοήθεια των βιβλιοθηκών “GOStats” [14] και “GOSim” [15] στα υποκεφάλαια 6.2 και 6.3 αντίστοιχα.

4 *Προγραμματισμός στην R*

Η R είναι μια στατιστική γλώσσα προγραμματισμού, η οποία αναπτύσσεται ραγδαία τα τελευταία χρόνια. Η R είναι ελεύθερα διαθέσιμη στην ιστοσελίδα http://www.r-project.org/ και στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων (packages) τα οποία διατίθενται πάλι ελεύθερα από χρήστες ανά τον κόσμο. Η R εφαρμόζει μια διάλεκτο της γλώσσας S η οποία είναι διερμηνέας γλώσσα προγραμματισμού. Αυτό σημαίνει ότι οι εντολές διαβάζονται και μετά εκτελούνται αμέσως. Το μεγάλο πλεονέκτημα των διερμηνέων γλωσσών προγραμματισμού είναι ότι επιτρέπουν σταδιακή ανάπτυξη, δηλαδή μια συνάρτηση μπορεί να δημιουργηθεί, να εκτελεσθεί και μετά να δημιουργηθεί μια καινούργια συνάρτηση η οποία καλεί την προηγούμενη.

4.1 Βασικά στοιχεία της R

Ένα από τα σύμβολα που χρησιμοποιείται είναι το σύμβολο εγχώρησης "←", το οποίο καταχωρεί στις μεταβλητές συγκεκριμένες τιμές (πχ. αριθμό, διάνυσμα, πίνακα, πλαίσιο δεδομένων κ.ά.) ή αποτελέσματα πράξεων.

Κάθε έκφραση της R ερμηνεύεται από τον αξιολογητή και επιστρέφει ένα αντικείμενο δεδομένων. Τα αντικείμενα δεδομένων έχουν τις παρακάτω μορφές :

- λογική (logical)
- αριθμητική (numeric)
- μιγαδική (complex)
- κειμένου (character)

Οι μορφές είναι γραμμένες από αυτήν που παρέχει την λιγότερη πληροφορία έως εκείνη που παρέχει την περισσότερη πληροφορία. Όταν είναι ανάγκη να συνδυάσεις διαφορετικές μορφές, τότε η R χρησιμοποιεί εκείνη με την περισσότερη πληροφορία.

Τα αντικείμενα δεδομένων είναι οι διάφορες μορφές στις οποίες μπορούν να φυλαχθούν δεδομένα στην R. Οι κύριες μορφές αντικειμένων δεδομένων που υπάρχουν είναι :

- διάνυσμα (vector)
- πίνακας (matrix)
- πίνακας μεγάλης διάστασης (array)

- λίστα (list)
- παράγοντας (factor)
- χρονοσειρές (time series)
- πλαίσιο δεδομένων (data frame)

Στην παρούσα εργασία τα αντικείμενα δεδομένων θα έχουν μορφές αριθμητικές (numeric) και κειμένου (character). Αντίστοιχα οι μορφές στις οποίες θα φυλάσσονται τα δεδομένα θα είναι διανύσματα (vectors), λίστες (lists) και πλαίσια δεδομένων (data frames).

Το πιο απλό είδος αντικειμένου είναι το διάνυσμα, ένα διατεταγμένο σύνολο τιμών σε σειρά. Η εσωτερική διάταξη του διανύσματος υποδεικνύει ότι υπάρχει ένας κατάλληλος τρόπος με τον οποίο μπορούν να εξαχθούν μερικά ή όλα τα στοιχεία του. Ο πιο εύκολος τρόπος για να προσδιοριστεί ένα διάνυσμα είναι μέσω της εντολής c.

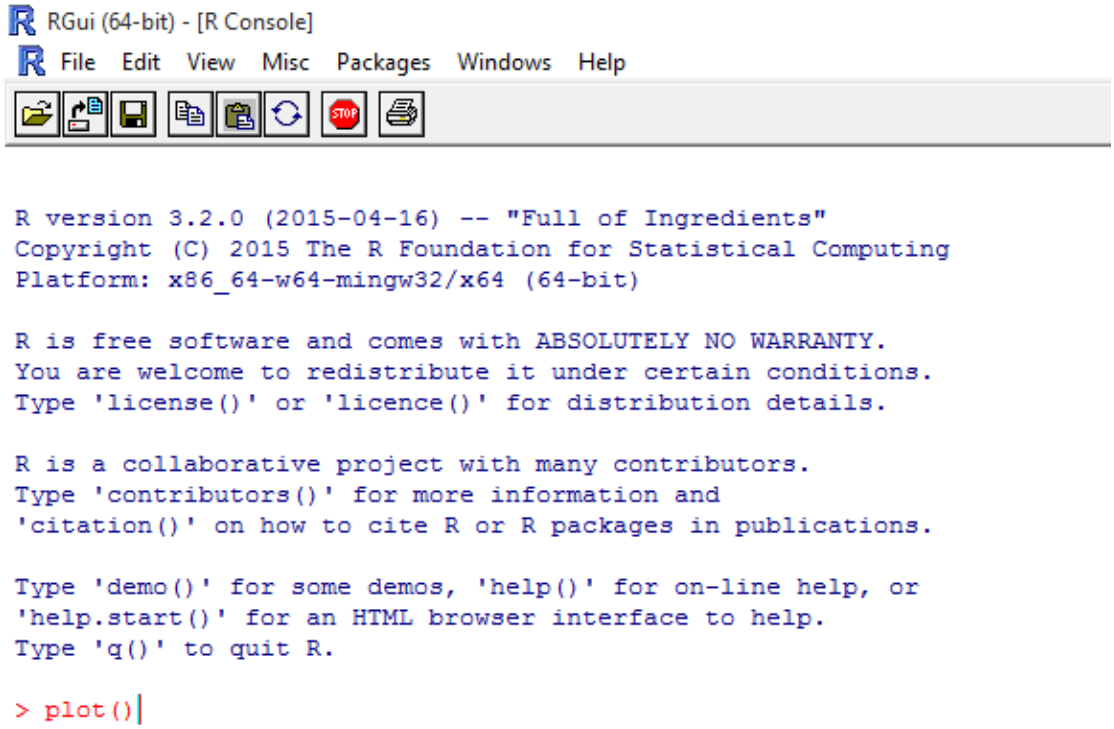
Τα αντικείμενα δεδομένων όπως το διάνυσμα είναι ατομικά, δηλαδή περιέχουν μόνο μιας μορφής δεδομένα. Όμως είναι αρκετές εκείνες οι περιπτώσεις στις οποίες υπάρχει ανάγκη να δημιουργηθούν αντικείμενα δεδομένων τα οποία περιέχουν διάφορες μορφές τιμών. Η λύση προσφέρεται μέσω των αντικειμένων λίστας (list) τα οποία αποτελούνται από διάφορες συνιστώσες, η κάθε μια από τις οποίες περιέχει διαφορετική μορφή δεδομένων.

Η μορφή των πλαισίων δεδομένων (data frames) θα χρησιμοποιηθεί εδώ για την φύλαξη των δεδομένων αποτελέσματος που προκύπτουν από τις αναλύσεις που θα πραγματοποιηθούν. Το κύριο πλεονέκτημα του πλαισίου δεδομένων είναι ότι επιτρέπει το συνδυασμό διαφορετικών μορφών σε ένα αντικείμενο. Η ιδέα του πλαισίου δεδομένων είναι η ταξινόμηση των τιμών κατά μεταβλητή (στήλη) ανεξάρτητα της μορφής τους. Έπειτα, όλες οι παρατηρήσεις ενός συγκεκριμένου συνόλου μεταβλητών ταξινομούνται σε πλαίσιο δεδομένων.

Ένα μεγάλο μέρος της ανάλυσης δεδομένων απαιτεί διάφορους μαθηματικούς υπολογισμούς. Οι υπολογιστικές δυνατότητες της R αρχίζουν από απλές πράξεις μέχρι και πολύπλοκους μαθηματικούς υπολογισμούς, όπως η μεγιστοποίηση συναρτήσεων. Για τις ανάγκες αυτής της εργασίας έχει συνταχθεί R κώδικας που εκτελεί όλους τους μαθηματικούς υπολογισμούς και τις γραφικές απεικονίσεις εκμεταλλευόμενος ρουτίνες τόσο από πακέτα της R αλλά και του Bioconductor. Ο

Bioconductor είναι λογισμικό ανοιχτού κώδικα για την βιοπληροφορική και παρέχει εργαλεία για την ανάλυση και την κατανόηση των γονιδιακών πληροφοριών.

Μετά από την εκτέλεση των στατιστικών μαθηματικών μοντέλων που πραγματοποιούνται μέσω των εντολών της εκάστοτε βιβλιοθήκης, είναι απαραίτητο να αναπαρασταθούν τα αποτελέσματα. Το οπτικό μέσο αναπαράστασης των δεδομένων αυτών είναι τα γραφήματα, που επιτρέπουν την αξιολόγηση των αναλύσεων. Η R δίνει ένα πολύ ισχυρό περιβάλλον για τη δημιουργία γραφημάτων. Η βασική εντολή για γραφική παράσταση είναι η εντολή plot (Σχήμα 4.1), η οποία έχει πολλές δυνατότητες και μπορεί να πάρει διάφορες γραφικές παραμέτρους για ορίσματα. [16]



Σχήμα 4.1: Σύνταξη της εντολής plot στο περιβάλλον της R

Η έννοια του προγραμματισμού στην R βασίζεται στη δημιουργία καινούργιων συναρτήσεων οι οποίες θα χρησιμοποιηθούν για περαιτέρω ανάπτυξη της γλώσσας. Το κύριο δομικό υλικό αυτών είναι οι υπάρχουσες συναρτήσεις (functions) της R.

Στην R έχει αναπτυχθεί η γνωστή μεθοδολογία της γραμμικής παλινδρόμησης, ένα από τα σημαντικότερα θέματα της στατιστικής θεωρίας, η οποία αναλύεται στο

Κεφάλαιο 5. Οι μέθοδοι εκτίμησης των συντελεστών του απλού λογαριθμικού γραμμικού μοντέλο και του γενικού γραμμικού μοντέλου μέσω της γραμμικής παλινδρόμησης και της μερικής παλινδρόμησης ελαχίστων τετραγώνων πραγματοποιείται στο περιβάλλον της R. Αντίστοιχα και οι έλεγχοι των αποτελεσμάτων της ανάλυσης με κριτήριο την μεταβλητή σημαντικότητας για προβολή ***VIP*** και τον συντελεστή προσδιορισμού $\boldsymbol{Q^2}$, των οποίων οι αναλυτικοί ορισμοί παρουσιάζονται αντίστοιχα στα υποκεφάλαια 5.2.5. και 5.2.7.

4.2 Μεταφραστικά πακέτα του Bioconductor

Ο Bioconductor είναι ένα λογισμικό ανοιχτού κώδικα για την βιοπληροφορική που σε συνεργασία με την R, παρέχει εργαλεία για την ανάλυση και την κατανόηση των γονιδιακών πληροφοριών. Μια από τις σημαντικές λειτουργίες του είναι ότι παρέχει πολλά μεταφραστικά πακέτα (libraries) βιολογικής πληροφορίας, ώστε μια πληροφορία να μπορεί να μεταφραστεί σε μια άλλη ή να αντιστοιχηθεί σε ισοδύναμη, λαμβάνοντας διαφορετική μορφή.

Διατίθενται πολλοί τύποι πακέτων μετάφρασης στον Bioconductor. Εδώ θα χρησιμοποιηθούν τα πακέτα που υποστηρίζονται από την βιβλιοθήκη "AnnotationDbi" [17]. Η "AnnotatioDbi" βιβλιοθήκη είναι πακέτο λογισμικού που ενεργοποιεί όλα τα μεταφραστικά πακέτα. Κάθε πακέτο περιλαμβάνει μια βάση δεδομένων.

Το πακέτο που θα χρησιμοποιηθεί στην δική μας ανάλυση είναι το "org.Hs.eg.db" [18]. Ανήκει στην κατηγορία πακέτων «organism-level» ('org') και περιλαμβάνει αντιστοιχίσεις μεταξύ αναγνωριστικών όρων του ανθρώπινου γονιδιώματος. Για παράδειγμα, η Entrez ταυτότητα [19] ενός γονιδίου αντιστοιχίζεται, με άλλους αναγνωριστικούς όρους, όπως GO ταυτότητα βιολογικής διαδικασίας [10], Uniprot ταυτότητα πρωτεΐνης [20]. Το όνομα ενός 'org' πακέτου είναι πάντα της μορφής "org.<XX>.<YY>.db". Εδώ το <XX> είναι η συντομογραφία που δηλώνει τον οργανισμό, που εδώ είναι <Hs>, άρα το ανθρώπινο είδος Homo sapiens, ενώ το <YY> είναι η συντομογραφία για τον βασικό αναγνωριστικό όρο του πακέτου, που εδώ είναι <eg>, δηλαδή η entrez gene ταυτότητα. Τα πακέτα της μορφής ".db." (database), ενημερώνονται με νέες πληροφορίες σχεδόν κάθε 6 μήνες.

Ένα άλλο πακέτο που είναι απαραίτητο για την λειτουργία πολλών αναλύσεων είναι το πακέτο "hgu95av2.db" [21]. Το πακέτο ανήκει στην κατηγορία πακέτων

«platform-level» ή «chip-level». Το όνομα του δηλώνει ότι περιέχει πληροφορίες για το ανθρώπινο γονιδίωμα “hgXXXXX” (human genome) μέσω του τσιπ σιλικόνης με κωδικό όνομα “u95av2” πάνω στο οποίο εκχύνεται το βιολογικό υλικό. Αυτό το πακέτο αποσκοπεί στο να μετατρέπει τις ετικέτες του κατασκευαστή (Affymetrix), που αντιπροσωπεύουν μια σειρά ανιχνευτές, σε ένα ευρύ φάσμα χαρακτηριστικών που αφορούν τα γονίδια. π.χ. η λειτουργία “hgu95av2ENTREZID” αντιστοιχίζει ταυτότητες ανιχνευτών του “u95av2” τσιπ του ανθρώπινου γονιδιώματος σε Entrez ταυτότητες γονιδίων.

4.2.1 Λειτουργία των μεταφραστικών πακέτων

Κάθε πακέτο αφού ενεργοποιηθεί στο περιβάλλον της R, καλεί να εμφανιστούν οι βάσεις δεδομένων που περιέχει, οι οποίες εμφανίζονται ως στήλες πχ. Entrez gene IDs, GO IDs, Uniprot IDs, Ontology κ.ά.

Στα Σχήματα 4.2 και 4.3 εμφανίζονται όλες οι πληροφορίες που περιέχουν τα πακέτα “org.Hs.eg.db” και “hgu95av2.db”, αντίστοιχα, στο περιβάλλον της R.

```
> columns(org.Hs.eg.db)
 [1] "ENTREZID"     "PFAM"         "IPI"          "PROSITE"      "ACCNUM"
 [6] "ALIAS"        "CHR"          "CHRLOC"       "CHRLOCEND"    "ENZYME"
[11] "MAP"          "PATH"         "PMID"         "REFSEQ"       "SYMBOL"
[16] "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS" "GENENAME"
[21] "UNIPROT"      "GO"           "EVIDENCE"     "ONTOLOGY"     "GOALL"
[26] "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"         "UCSCKG"
```

Σχήμα 4.2: Στήλες δεδομένων του πακέτου “org.Hs.eg.db” στο περιβάλλον της R

```
> columns(hgu95av2.db)
 [1] "PROBEID"      "ENTREZID"     "PFAM"         "IPI"          "PROSITE"
 [6] "ACCNUM"       "ALIAS"        "CHR"          "CHRLOC"       "CHRLOCEND"
[11] "ENZYME"       "MAP"          "PATH"         "PMID"         "REFSEQ"
[16] "SYMBOL"       "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
[21] "GENENAME"     "UNIPROT"      "GO"           "EVIDENCE"     "ONTOLOGY"
[26] "GOALL"        "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"         "UCSCKG"
```

Σχήμα 4.3: Στήλες δεδομένων πακέτου “hgu95av2.db” στο περιβάλλον της R

Οι στήλες αυτές αντιστοιχούν σε βάσεις δεδομένων που σκοπό έχουν να μεταφράσουν και να αντιστοιχήσουν πληροφορία μεταξύ τους με βάση το μοναδικό κωδικοποιημένο όνομα που δίνουν σε κάθε βιολογικό όρο. Οι βασικότερες κατηγορίες δεδομένων των στηλών που περιέχουν τα δύο πακέτα:

- ENTREZID: Entrez ταυτότητα γονιδίου
- CHR: χρωμόσωμα
- CHRLOC: χρωμόσωμα και θέση εκκίνησης σχετιζόμενου γονιδίου
- CHRLOCEND: χρωμόσωμα και θέση τερματισμού σχετιζόμενου γονιδίου
- ENZYME: αριθμητικός όρος κατάταξης του ενζύμου, με βάση τις χημικές αντιδράσεις που καταλύει.
- PATH: όρος που δηλώνει το μονοπάτι στην KEGG
- SYMBOL: επίσημη συμβολική ονομασία γονιδίου
- ENSEMBL: Ensembl ταυτότητα γονιδίου
- GENENAME: περιγραφική ονομασία του γονιδίου
- UNIPROT: Uniprot ταυτότητες πρωτεϊνών
- GO: GO ταυτότητες βιολογικών λειτουργιών
- EVIDENCE: κωδικές ονομασίες για την σύνδεση GO ταυτοτήτων γονιδίου με το γονίδιο
- ONTOLOGY: οντολογία, στην οποία εντάσσεται ο GO όρος

Έτσι μέσω μιας λειτουργίας της βιβλιοθήκης, όταν έχουμε ένα σύνολο δεδομένων με GO ταυτότητες γονιδίων και θέλουμε τις μεταφράσεις τους σε μία ή περισσότερες από τις παραπάνω κατηγορίες, ακολουθείται η παρακάτω διαδικασία:

1. Ορίζεται ο τύπος των δεδομένων που διαθέτονται.
2. Εισάγονται τα δεδομένα σε μορφή διανύσματος.
3. Επιλέγονται οι τύποι των δεδομένων που ζητούνται μέσω της διαδικασίας της μετάφρασης.

Εφόσον τελειώσει η διαδικασία της μετάφρασης των δεδομένων, λαμβάνεται ένας πίνακας με τις αντιστοιχίσεις των δεδομένων με τους τύπους δεδομένων που επιλέχθηκαν. Όταν δεν υπάρχει διαθέσιμη αντιστοίχιση με κάποιο δεδομένο τότε θα έχουμε την ένδειξη NA (Not Available), το οποίο δηλώνει ότι δεν υπάρχει η ζητούμενη αντιστοίχιση στις βάσεις δεδομένων.

4.2.2 Μετάφραση των γονιδίων

Για τις αναλύσεις που θα ακολουθήσουν με τη χρήση των πακέτων του Bioconductor, θα χρειαστεί να μεταφραστούν οι 129 Uniprot ταυτότητες πρωτεϊνών που περιέχονται

στο Σύνολο Α. Αρχικά θα χρησιμοποιηθεί το μεταφραστικό πακέτο “org.Hs.eg.db” για να αντιστοιχιστούν οι Uniprot ταυτότητες των πρωτεϊνών σε:

- ✓ SYMBOL
- ✓ GO
- ✓ EVIDENCE
- ✓ ONTOLOGY
- ✓ ENTREZID

Ο ακόλουθος πίνακας περιλαμβάνει τις μεταφράσεις της Uniprot πρωτεϊνικής ταυτότητας P55056, ενώ μέρος του συνολικού πίνακα των μεταφράσεων των 129 Uniprot ταυτοτήτων περιλαμβάνεται στον Πίνακα ΠΑ1.

Πίνακας 4.1: Μετάφραση της Uniprot ταυτότητας P55056

UNIPROT	SYMBOL	GO	EVIDENCE	ONTOLOGY	ENTREZID
P55056	APOC4	GO:0005319	TAS	MF	346
P55056	APOC4	GO:0006629	TAS	BP	346
P55056	APOC4	GO:0006869	IEA	BP	346
P55056	APOC4	GO:0010890	IMP	BP	346
P55056	APOC4	GO:0034361	IDA	CC	346
P55056	APOC4	GO:0034361	IMP	CC	346
P55056	APOC4	GO:0034364	IDA	CC	346
P55056	APOC4	GO:0070328	IMP	BP	346

Όπως παρατηρείται και στον παραπάνω πίνακα η Uniprot πρωτεϊνική ταυτότητα P55056 μεταφράζεται σε μια συμβολική ονομασία γονιδίου (SYMBOL) APOC4 και σε μια Entrez ταυτότητα γονιδίου 346. Αντίθετα μεταφράζεται σε πολλές GO ταυτότητες βιολογικών λειτουργιών που μάλιστα αντιστοιχούν και στις τρεις κατηγορίες οντολογιών (BP, MF, CC) και έχουν προκύψει από διαφορετικές πηγές (TAS, IEA, IMP, IDA). Κάθε GO όρος του πίνακα έχει ενεργή υπερσύνδεση στην σελίδα http://www.ebi.ac.uk στην οποία διατίθεται βάση δεδομένων που παρέχει πληροφορίες για τον κάθε όρο.

Μια GO μετάφραση περιλαμβάνει έναν GO όρο που σχετίζεται με συγκεκριμένη αναφορά που περιγράφει την ανάλυση επί της οποίας βασίζεται η σχέση μεταξύ του όρου αυτού και του γονιδιακού προϊόντος. Κάθε μετάφραση πρέπει να περιλαμβάνει αποδεικτικά στοιχεία (EVIDENCE) σε κωδική μορφή που υποδεικνύουν πως παρέχεται η μετάφραση σε έναν συγκεκριμένο GO όρο. Υπάρχουν διαφορετικά είδη

EVIDENCE, όπως το IMP (Inferred from Mutant Phenotype) και το IDA (Inferred from Direct Assay) που προκύπτουν από πειραματικά δεδομένα άρθρων, το IEA (Inferred from Electronic Annotation) που έχει εκχωρηθεί με αυτοματοποιημένες μεθόδους και το TAS (Traceable Author Statement) που προκύπτει από αναφορές συντάκτη ενός σχετικού άρθρου που αναφέρεται. [22]

Από το Σύνολο B, λαμβάνονται οι συμβολικές ονομασίες των γονιδίων (SYMBOL) που αντιστοιχούν στις πιο σημαντικές πρωτεΐνες της ανάλυσης. Για την χρήση της πληροφορίας τους κρίνεται απαραίτητο να μεταφραστούν σε άλλες μορφές για να αξιοποιηθούν στις αναλύσεις στη συνέχεια.

Οι ταυτότητες αυτές των γονιδίων θα μεταφραστούν με τη χρήση του πακέτου “org.Hs.eg.db”, σε :

- ✓ UNIPROT
- ✓ GO
- ✓ EVIDENCE
- ✓ ONTOLOGY
- ✓ ENTREZID

Ο Πίνακας 4.2 περιλαμβάνει τις μεταφράσεις της συμβολικής ονομασίας του γονιδίου CNDP1, ενώ μέρος του συνολικού πίνακα των μεταφράσεων των 76 συμβολικών ονομασιών περιλαμβάνεται στον Πίνακα ΠΑ2.

Πίνακας 4.2: Μετάφραση της συμβολικής ονομασίας του γονιδίου CNDP1

SYMBOL	UNIPROT	GO	EVIDENCE	ONTOLOGY	ENTREZID
CNDP1	Q96KN2	GO:0004180	IEA	MF	84735
CNDP1	Q96KN2	GO:0005576	IEA	CC	84735
CNDP1	Q96KN2	GO:0006508	IEA	BP	84735
CNDP1	Q96KN2	GO:0008237	IEA	MF	84735
CNDP1	Q96KN2	GO:0016805	IEA	MF	84735
CNDP1	Q96KN2	GO:0034701	IEA	MF	84735
CNDP1	Q96KN2	GO:0046872	IEA	MF	84735

Οι Entrez ταυτότητες γονιδίων καθώς και οι GO ταυτότητες βιολογικών λειτουργιών θα χρησιμοποιηθούν για δημιουργία GO γραφημάτων καθώς και για αναλύσεις εμπλουτισμού που θα ακολουθήσουν στα Κεφάλαια 6 και 7.

5 Στατιστική ανάλυση

Η έρευνα για την βιολογική πληροφορία στην παρούσα εργασία αφορά την περιγραφική στατιστική, η οποία είναι μια συλλογή από γραφικές μεθόδους για εξερεύνηση και κατανόηση των δεδομένων.

Στις αναλύσεις που θα πραγματοποιηθούν θα χρησιμοποιούνται οι τελεστές ελέγχου για την κατασκευή των υποθετικών προτάσεων στο επίπεδο της στατιστικής. Οι κύριοι λογικοί τελεστές και τελεστές σύγκρισης που χρησιμοποιούνται είναι οι "==" (ίσο με), ">"(μεγαλύτερο από), "<" (μικρότερο από), ">=" (μεγαλύτερο ή ίσο από), "<=" (μικρότερο ή ίσο από).

Η περιγραφική στατιστική χρησιμοποιεί γραφήματα τα οποία βοηθούν στο να εξερευνηθεί αν ισχύουν οι υποθέσεις των στατιστικών μοντέλων. Μερικές ερωτήσεις που ενδιαφέρουν σε αυτές τις περιπτώσεις είναι:

- ✓ Τα δεδομένα ακολουθούν την κανονική κατανομή;
- ✓ Υπάρχουν απομακρυσμένες τιμές στα δεδομένα;
- ✓ Τα δεδομένα συγκεντρώθηκαν διαδοχικά στο χρόνο, υπάρχει ένδειξη σειριακής συσχέτισης;

Τα κύρια γραφήματα τα οποία βοηθούν στην εποπτική εξερεύνηση των δεδομένων είναι τα ακόλουθα:

- Ιστογράμματα (histograms)
- Κυριογραφήματα (boxplots)
- Γραφήματα πυκνότητας (density plots)
- QQ γραφήματα (quantile-quantile plots)

5.1 Στατιστικοί έλεγχοι υποθέσεων

Υπόθεση ονομάζεται η απόφαση που παίρνουμε για θέματα σχετικά με τους πληθυσμούς, βασιζόμενοι σε δείγματα πληθυσμών. Στατιστικός έλεγχος υποθέσεων ή έλεγχος σημαντικότητας ονομάζεται η διαδικασία που χρησιμοποιείται ώστε να αποφασίσουμε αν θα δεχτούμε ή θα απορρίψουμε τις υποθέσεις που έχουμε κάνει. Μηδενική υπόθεση $\boldsymbol{H_0}$ ορίζεται ως η υπόθεση που κάνουμε αρχικά με σκοπό να την απορρίψουμε, ενώ εναλλακτική υπόθεση $\boldsymbol{H_1}$ ορίζεται ως η ασυμβίβαστη υπόθεση σε σχέση με την μηδενική υπόθεση. Η απόφαση αν θα γίνει δεκτή ή θα απορριφθεί η

μηδενική υπόθεση $\boldsymbol{H_0}$ στηρίζεται σε στατιστικά τεστ που υπολογίζονται από τα δεδομένα του δείγματος. Απορριπτική περιοχή R της $\boldsymbol{H_0}$ ονομάζεται η περιοχή στα σημεία της οποίας η $\boldsymbol{H_0}$ απορρίπτεται.

Τα στοιχεία ενός στατιστικού ελέγχου είναι τα εξής:

- Ορίζεται η μηδενική υπόθεση $\boldsymbol{H_0}$
- Ορίζεται η εναλλακτική υπόθεση $\boldsymbol{H_1}$
- Ορίζεται το στατιστικό του ελέγχου από τα δεδομένα του δείγματος
- Ορίζεται η απορριπτική περιοχή R της $\boldsymbol{H_0}$
- Εξάγονται τα συμπεράσματα

Μονόπλευρος έλεγχος ορίζεται ως ο έλεγχος που κάνουμε όταν θέλουμε να δούμε αν ευσταθεί η υπόθεση, ότι η διαδικασία που επιλέξαμε είναι καλύτερη από μια άλλη. Δηλαδή αν $\boldsymbol{H_0}$: $\theta=\theta_o$, τότε $\boldsymbol{H_1}$: $\theta>\theta_o$ ή $\boldsymbol{H_1}$: $\theta<\theta_o$. Δίπλευρος έλεγχος ορίζεται ως ο έλεγχος που γίνεται στις τιμές των άκρων της κατανομής που μελετάμε. Δηλαδή αν $\boldsymbol{H_0}$: $\theta=\theta_o$, τότε $\boldsymbol{H_1}$: $\theta\neq\theta_o$.

Σφάλμα τύπου I ονομάζεται η απόρριψη της υπόθεσης $\boldsymbol{H_0}$ ενώ είναι σωστή. Η πιθανότητα αυτού του σφάλματος συμβολίζεται με $\boldsymbol{\alpha}$ και ονομάζεται επίπεδο σημαντικότητας (significance level) ή σημαντικότητα ενός ελέγχου. Σφάλμα τύπου II ονομάζεται η αποδοχή της $\boldsymbol{H_0}$ ενώ είναι λάθος. Η πιθανότητα αυτού του σφάλματος συμβολίζεται με $\boldsymbol{\beta}$. [23]

Οι έλεγχοι υποθέσεων χρησιμοποιούνται για να εξετάσουν την εγκυρότητα ενός ισχυρισμού που γίνεται σχετικά με τον πληθυσμό των παρατηρήσεων. Αυτός ο ισχυρισμός είναι η μηδενική υπόθεση. Ένας στατιστικός έλεγχος μπορεί να υποθέτει ότι οι παρατηρήσεις :

1. Είναι ανεξάρτητες μεταξύ τους
2. Προέρχονται από πληθυσμό από την κανονική κατανομή
3. Προέρχονται από πληθυσμούς με ίσες διασπορές

Η τιμή $\boldsymbol{p}$ (p-value) είναι μια συνάρτηση των παρατηρούμενων αποτελεσμάτων ενός δείγματος που χρησιμοποιείται για τον έλεγχο μιας στατιστικής υπόθεσης, όπως της μηδενικής. Συγκεκριμένα, πριν τον έλεγχο, επιλέγεται μια τιμή κατωφλίου, συνήθως 5% ή 1%, που συνιστά το επίπεδο σημαντικότητας του τεστ, όπως το ορίσαμε παραπάνω

Η εναλλακτική υπόθεση θα είναι αληθής, αν η μηδενική υπόθεση είναι αναληθής. Οι έλεγχοι υπόθεσης τελικά χρησιμοποιούν μια τιμή ***p***, ώστε να μετρήσουν την δύναμη των αποδεικτικών στοιχείων. Η τιμή ***p*** παίρνει τιμές μεταξύ 0 και 1 και ερμηνεύεται με τον ακόλουθο τρόπο:

- Μια μικρή τιμή ***p***, δηλαδή $p \leq \alpha$, παρέχει ισχυρές ενδείξεις εναντίον της μηδενικής υπόθεσης, άρα αυτή απορρίπτεται.
- Μια μεγάλη τιμή ***p***,δηλαδή $p > \alpha$, παρέχει ασθενείς ενδείξεις εναντίον της μηδενικής υπόθεση, άρα αυτή δεν απορρίπτεται.
- Τιμές ***p*** πολύ κοντά στο επίπεδο σημαντικότητας ***α***, θεωρούνται οριακές, και μπορούν είτε να απορρίψουν τη μηδενική υπόθεση είτε όχι.

Το αποτέλεσμα δείχνει ότι η μηδενική υπόθεση απορρίπτεται αν η τιμή ***p*** είναι πολύ κοντά στο 0, άρα μικρότερο του επιπέδου σημαντικότητας, ενώ γίνεται δεκτή αν η τιμή p είναι μεγαλύτερο του επιπέδου σημαντικότητας .

Εκτός από την τιμή ***p*** η οποία λαμβάνεται από τον έλεγχο συνήθως είναι χρήσιμο ένα διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών.[24]

5.1.1 Υπεργεωμετικό τεστ

Ένα στατιστικό τεστ που χρησιμοποιείται ιδιαίτερα στη συνέχεια είναι το υπεργεωμετρικό τεστ (hypergeometric test) που χρησιμοποιεί την υπεργεωμετρική κατανομή για να υπολογίσει τη στατιστική σημαντικότητα των επιτυχιών ***k*** που έχουν πραγματοποιηθεί (από n δειγματοληψίες) για έναν πληθυσμό. Η υπεργεωμετρική κατανομή είναι μια διακριτή κατανομή πιθανότητας που περιγράφει την πιθανότητα των ***k*** επιτυχιών σε n δειγματοληψίες, χωρίς επανάθεση, από ένα δείγμα μεγέθους N που περιέχει ακριβώς K επιτυχίες, όπου κάθε δειγματοληψία είναι επιτυχία ή αποτυχία. Αυτό το τεστ χρησιμοποιείται συνήθως για να προσδιορίσει ποιοί υποπληθυσμοί ενός δείγματος υπο-εκπροσωπούνται ή υπερ-εκπροσωπούνται.

Οι παρακάτω συνθήκες χαρακτηρίζουν την υπεργεωμετρική κατανομή:

- Το αποτέλεσμα κάθε δειγματοληψίας μπορεί να ταξινομηθεί σε μία από τις δυο αμοιβαία αποκλειόμενες κατηγορίες (πχ. επιτυχία ή αποτυχία)
- Η πιθανότητα μιας επιτυχίας αλλάζει σε κάθε δειγματοληψία, καθώς κάθε δειγματοληψία μειώνει τον πληθυσμό

Η συνάρτηση πιθανότητας της υπεργεωμετρικής κατανομής δίνεται από τη σχέση:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (5.1)$$

όπου N το μέγεθος του πληθυσμού, K ο αριθμός των επιτυχιών στον πληθυσμό, ***n*** ο αριθμός των δειγματοληψιών και ***k*** ο αριθμός των παρατηρούμενων επιτυχιών.

5.1.2 Kolmogorov-Smirnov τεστ

Ο έλεγχος καλής προσαρμογής (goodness of fit test) θεωρείται ως ακόμη ένας τρόπος για να εξακριβωθούν οι υποθέσεις της κατανομής των δεδομένων. Ένα τεστ που χρησιμοποιείται ευρέως για αυτό το σκοπό είναι το Kolmogorov-Smirnov τεστ (K-S τεστ). Ο έλεγχος Kolmogorov-Smirnov χρησιμοποιείται στην περίπτωση σύγκρισης ενός δείγματος με μια κατανομή, αλλά και στην περίπτωση σύγκρισης της κατανομής δύο δειγμάτων μεταξύ τους. Στην πρώτη περίπτωση, η μηδενική υπόθεση H_0 δηλώνει ότι τα δεδομένα ακολουθούν τη συγκεκριμένη κατανομή, ενώ στη δεύτερη δηλώνει ότι τα δύο δείγματα ακολουθούν την ίδια κατανομή. [25]

Για παράδειγμα η αξιολόγηση της υπόθεσης της κανονικότητας ενός πληθυσμού γίνεται με τον έλεγχο Kolmogorov-Smirnov.

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό X με συνάρτηση κατανομής $F(x)$. Για τον έλεγχο:

$$H_o: F(x) = F_0(x) - H_1: F(x) \geq F_O(x)$$

(η H_o ισχύει για όλα τα x ενώ η H_1 για ένα τουλάχιστον x) όπου $F_O(x)$ είναι μια πλήρως καθορισμένη συνάρτηση κατανομής συνεχούς τυχαίας μεταβλητής, χρησιμοποιείται ο έλεγχος Kolmogorov-Smirnov. Η στατιστική συνάρτηση του ελέγχου δίνεται από τη σχέση:

$$D_n = \sup_{-\infty<x<+\infty} \{|S_n(x) - F_O(x)|\} \quad (5.2)$$

όπου $S_n(x)$ είναι η εμπειρική συνάρτηση κατανομής του δείγματος. Προφανώς μεγάλες τιμές της D_n προσφέρουν ενδείξεις για την ισχύ της εναλλακτικής υπόθεσης. Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας α είναι η $\{d_n: d_n > D_{n,a}\}$, όπου d_n είναι η παρατηρούμενη τιμή της D_n και $D_{n,a}$ το άνω α ποσοστιαίο σημείο της κατανομής του D_n.

Αν και ο παραπάνω έλεγχος έχει σχεδιαστεί για συνεχείς πληθυσμούς μπορεί να χρησιμοποιηθεί και για διακριτούς πληθυσμούς στους οποίους φυσικά η $F_0(x)$ είναι συνάρτηση κατανομής διακριτής τυχαίας μεταβλητής. [26]

5.1.3 Ακριβές τεστ του Fisher

Το ακριβές τεστ του Fisher (Fisher's exact test) αρχικά βασίστηκε στον έλεγχο ανεξαρτησίας δύο μεταβλητών για πίνακες συνάφειας 2x2. Εφαρμόζεται όμως και για πίνακες συνάφειας μεγαλύτερης διάστασης.

Στην στατιστική οι πίνακες συνάφειας χρησιμοποιούνται για να περιγράψουν τη σχέση δύο μεταβλητών. Σε αυτούς καταγράφεται η συχνότητα ή η σχετική συχνότητα για την κάθε παρατήρηση των δυο μεταβλητών.

Η γενική μορφή ενός πίνακα συνάφειας 2x2 είναι η ακόλουθη:

Y \ X	1	2	Σύνολο
1	N_{11}	N_{12}	$N_{1\bullet}$
2	N_{21}	N_{22}	$N_{2\bullet}$
Σύνολο	$N_{\bullet 1}$	$N_{\bullet 2}$	$N_{\bullet\bullet}$

Μας ενδιαφέρει να ελέγξουμε κατά πόσο θεωρείται «τυχαία» η κατανομή των δεδομένων με αθροίσματα γραμμών $N_{1\bullet}$, $N_{2\bullet}$ και στηλών $N_{\bullet 1}$, $N_{\bullet 2}$. Παρατηρούμε ότι αρκεί να ελέγξουμε κατά πόσο ήταν τυχαίο το N_{11} που εμφανίστηκε διότι δεδομένων των αθροισμάτων των γραμμών και των στηλών τα N_{12}, N_{21}, N_{22} μπορούν να εξαχθούν από το N_{11}. Έστω λοιπόν ότι από το δείγμα που πήραμε βρέθηκε ότι $N_{11} = n_{11}$. Η πιθανότητα να έχει συμβεί αυτό τυχαία δεδομένου ότι $N_{1\bullet} = n_{1\bullet}$, $N_{2\bullet} = n_{2\bullet}$, $N_{\bullet 1} = n_{\bullet 1}$, $N_{\bullet 2} = n_{\bullet 2}$, δίνεται από την υπεργεωμετρική κατανομή.

$$\mathrm{p} = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{1\bullet}-n_{11}}}{\binom{n_{1\bullet}+n_{2\bullet}}{n_{1\bullet}}} \tag{5.3}$$

Το p-value του Fisher τεστ που ελέγχει την υπόθεση Ho: το αποτέλεσμα στα 4 κελιά είναι τυχαίο (δεδομένων των αθροισμάτων των γραμμών και των στηλών), είναι ίσο με την πιθανότητα να εμφανιστεί το δείγμα που εμφανίστηκε και ακόμη πιο «ακραίο» από αυτό δεδομένης της Ho, δηλαδή:

$$p - value = \sum_{i=0}^{n_{11}} \frac{\binom{n_{1\bullet}}{i}\binom{n_{2\bullet}}{n_{1\bullet}-i}}{\binom{n_{1\bullet}+n_{2\bullet}}{n_{1\bullet}}} \tag{5.4}$$

ή

$$p - value = \sum_{i=n_{11}}^{n_{1\bullet}} \frac{\binom{n_{1\bullet}}{i}\binom{n_{2\bullet}}{n_{1\bullet}-i}}{\binom{n_{1\bullet}+n_{2\bullet}}{n_{1\bullet}}} \tag{5.5}$$

Ανάλογα με το αν $n_{11} < \frac{n_{\bullet 1} n_{1\bullet}}{N}$ ή $n_{11} > \frac{n_{\bullet 1} n_{1\bullet}}{N}$ αντίστοιχα. Το παραπάνω ισχύει για μονόπλευρο έλεγχο. [27]

5.2 Ανάλυση Παλινδρόμησης: Μοντέλο κυτταρικής συσχέτισης

Η ανάλυση παλινδρόμησης είναι μια από τις σημαντικότερες στατιστικές μεθόδους. Πολλά φαινόμενα των σύγχρονων επιστημών μοντελοποιούνται μέσω αυτής. Η ανάλυση αυτή θα χρησιμοποιηθεί για να καθορίσει την σχέση μεταξύ των μεταβλητών, δηλαδή για το σχεδιασμό ενός μοντέλου, που θα ερμηνεύει ικανοποιητικά την κυτταρική συσχέτιση με την περιεκτικότητα των πρωτεϊνών στις μορφές των νανοσωματιδίων. Θεωρώντας ως εξαρτημένη μεταβλητή ***Y*** την κυτταρική συσχέτιση και ως ανεξάρτητες μεταβλητές ***X*** τις περιεκτικότητες των

πρωτεϊνών στα νανοσωματίδια, προκύπτει ένα γραμμικό λογαριθμικό μοντέλο σύνδεση τους. Σκοπός είναι ο καθορισμός του βέλτιστου μοντέλου για τα δεδομένα.

5.2.1 Απλό γραμμικό μοντέλο

Απλό γραμμικό μοντέλο ονομάζεται το μοντέλο μέσω του οποίου διερευνάται η σχέση μεταξύ δυο μεταβλητών, $\boldsymbol{y}$ εξαρτημένης και $\boldsymbol{x}$ ανεξάρτητης. Με συλλογή ενός δείγματος i από έναν πληθυσμό νανοσωματιδίων και καταγράφοντας για ένα από αυτά τις τιμές των δυο μεταβλητών, δημιουργούνται τα εξής ζεύγη τιμών:

$$(x_1, y_1), (x_2, y_2), \ldots . (x_i, y_i)$$

μέσω των οποίων μπορούμε να διερευνήσουμε τη σχέση μεταξύ των μεταβλητών $\boldsymbol{x}$ και $\boldsymbol{y}$. Για να καθοριστεί ο βαθμός στον οποίο οι μεμονωμένες πρωτεΐνες εντός του αποτυπώματος του πρωτεϊνικού πλάσματος, προβλέπουν την κυτταρική συσχέτιση, αναπτύχθηκε μια σειρά λογαριθμικών γραμμικών μοντέλων. Αυτά τα μοντέλα συσχετίζουν την σχετική περιεκτικότητα της κάθε απορροφημένης πρωτεΐνης πλάσματος με την κυτταρική συσχέτιση. Κάθε μοντέλο που ερμηνεύει μια τέτοια σχέση έχει τη μορφή:

$$\log_2 y_i = a_j \cdot x_{ij} + e_i \qquad (5.6)$$

όπου $\boldsymbol{y_i}$ είναι η παρατηρούμενη τιμή για την κυτταρική συσχέτιση των πρωτεϊνών με το κύτταρο, που χαρακτηρίζουν την μορφή του νανοσωματιδίου i και $\boldsymbol{x_{ij}}$ είναι η σχετική περιεκτικότητα της πρωτεΐνης j του πλάσματος στην μορφή του νανοσωματιδίου i, $\boldsymbol{a_j}$, η παράμετρος του μοντέλου και $\boldsymbol{e_i}$ μια ανεξάρτητη τυχαία μεταβλητή. Αυτή η τυχαία μεταβλητή ονομάζεται σφάλμα και θεωρούμε ότι ακολουθεί την κανονική κατανομή $N(0,\sigma^2)$ με $\boldsymbol{\sigma^2}$ άγνωστο. [8]

Να σημειωθεί ότι προκειμένου το μοντέλο να ερμηνευθεί στον καλύτερο δυνατό βαθμό χρησιμοποιούνται η διαδικασία του centering (κεντροποίηση) καθώς και του scaling (τυποποίηση).

Σκοπός του centering είναι στο τελικό μοντέλο να αφαιρείται ο σταθερός όρος. Αυτό συμβαίνει αφαιρώντας από κάθε μεταβλητή $\boldsymbol{x_i}$, τον αντίστοιχο μέσο $\boldsymbol{\bar{x}_i}$. Αυτός ορίζεται ως:

$$\bar{x_j} = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} x_{ij} \tag{5.7}$$

όπου n είναι ο αριθμός των πρωτεϊνών στην περίπτωσή μας. Αντίστοιχα για την μεταβλητή $\mathbf{log_2\, y}$ ισχύει:

$$log\bar{y} = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} logy_i \tag{5.8}$$

Η διαδικασία του scaling κρίνεται απαραίτητη καθώς οι μεταβλητές οφείλουν να είναι τυποποιημένες προκειμένου οι συντελεστές τους να είναι αδιάστατα μεγέθη. Έτσι θεωρούμε για τις μεταβλητές τα μεγέθη $\boldsymbol{x}^*$ και $\boldsymbol{y}^*$:

$$x_{ij}^* = \frac{x_{ij} - \bar{x_j}}{s_j} \tag{5.9}$$

όπου $\boldsymbol{s_j}$ η τυπική απόκλιση όταν,

$$s^2{}_j = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}{n-1} \tag{5.10}$$

η διασπορά των παρατηρήσεων. Ομοίως έχουμε :

$$y_i^* = \frac{y_i - \bar{y}}{s} \tag{5.11}$$

όπου $\boldsymbol{s}$ η τυπικής απόκλιση της $\boldsymbol{y}$-μεταβλητής όταν,

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{5.12}$$

[28]

Έτσι δημιουργήθηκαν μια σειρά των 129 απλών γραμμικών μοντέλων, τα οποία περιγράφουν τη συσχέτιση των A549 κυττάρων, με τη βιβλιοθήκη των νανοσωματίδιων χρυσού, ως μια λειτουργία της σχετικής περιεκτικότητας κάθε πρωτεΐνης στο αποτύπωμα του πρωτεϊνικού ορού.

5.2.2 Γενικό γραμμικό μοντέλο

Η ακρίβεια πρόβλεψης του απλού γραμμικού μοντέλου μπορεί να βελτιωθεί χρησιμοποιώντας πολλαπλές απορροφημένες πρωτεΐνες ταυτόχρονα.

Η μεταβλητή απόκρισης $\boldsymbol{log\, y_i}$, μπορεί να συνδέεται με περισσότερες από μια επεξηγηματικές μεταβλητές $\boldsymbol{x_{i1}, x_{i2}, x_{i3}\ldots x_{ij}}$. Ομοίως με το απλό γραμμικό μοντέλο μπορούμε να χρησιμοποιήσουμε ένα νέο μοντέλο που καλείται πολλαπλό γραμμικό μοντέλο, το οποίο διερευνά την εξάρτηση της $\boldsymbol{log\, y_i}$ από τις $\boldsymbol{x_{i1}, x_{i2}, x_{i3}\ldots x_{ij}}$ μεταβλητές. Αυτό θα έχει την ακόλουθη μορφή:

$$\log_2 y = b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_j \cdot x_j + \ e \qquad (5.13)$$

με $\boldsymbol{b_1, b_2,\ldots b_j}$ άγνωστες σταθερές.

Για την διερεύνηση της εξάρτησης αυτής, λαμβάνεται από τον πληθυσμό, δείγμα πρωτεϊνών μεγέθους j και για κάθε νανοσωματίδιο του δείγματος καταγράφονται οι τιμές των μεταβλητών αυτών. Δηλαδή για το i-νανοσωματίδιο καταγράφονται οι τιμές $\boldsymbol{(y_i, x_{i1}, x_{i2}, x_{i3}\ldots x_{im})}$, οδηγώντας στο μοντέλο:

$$\log_2 y_i = b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \cdots + b_j \cdot x_{im} + e_i \qquad (5.14)$$

όπου i=1,2…,84 οι μορφές των νανοσωματιδίων και m ο αριθμός των νανοσωματιδίων. Τα $\boldsymbol{e_i}$ καλούνται ομοίως σφάλματα, και είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την κανονική κατανομή $N(0,\sigma^2)$.

Οι επιφανειακά απορροφημένες πρωτεΐνες είτε ενισχύουν την κυτταρική συσχέτιση είτε την αποτρέπουν. Το πρόσημα της παραμέτρου $\boldsymbol{b_i}$ για κάθε πρωτεΐνη αντιπροσωπεύει την σχέση μεταξύ της απορροφημένης πρωτεΐνης ορού και της κυτταρικής συσχέτισης. Τιμές της παραμέτρου με θετικό πρόσημο, δηλώνουν ότι όσο

μεγαλύτερη είναι η απορρόφηση της πρωτεΐνης σε μια μορφή νανοσωματιδίου τόσο μεγαλύτερη θα είναι κυτταρική συσχέτιση. Σε αντίθεση, παράμεροι με αρνητικά πρόσημα δηλώνουν ότι όσο μεγαλύτερη η απορρόφηση της πρωτεΐνης αυτής τόσο μικρότερη θα είναι κυτταρική συσχέτιση, συμπεραίνοντας ότι η πρωτεΐνη αποτρέπει την κυτταρική συσχέτιση. Στην ερευνητική εργασία του Walkey και των συνεργατών του, από τις 64 πρωτεΐνες που χρησιμοποιήθηκαν για το μοντέλο, 39 χαρακτηρίστηκαν ως υποκινητές ενώ 25 ως αναστολείς. Οι πέντε πιο σημαντικοί υποκινητές είναι: οι ενδο-άλφα βαριές αλυσίδες αναστολείς θρυψίνης H1,H2, και H3 (ITIH1,ITIH2 και ITIH3), η α-1 μικρογλοβουλίνη (AMBP) και η πρωτεΐνη δέσμευσης της υαλουρονάνης (HABP2). Ο πιο ισχυρός αναστολέας είναι το συμπλήρωμα C3 (CO3). [8]

Το παραπάνω μοντέλο παρουσιάζεται με την μορφή πινάκων ως:

$$Y = X \cdot b + e \tag{5.15}$$

όπου $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$

όπου $\boldsymbol{Y} \in R^{nx1}$, $\boldsymbol{X} \in R^{nxm}$ και $\boldsymbol{b} \in R^{nxm}$ και $\boldsymbol{e} \in R^{nxm}$. Το $\boldsymbol{Y}$ είναι ένα διάνυσμα τιμών της κυτταρικής συσχέτισης για κάθε μορφή του νανοσωματιδίου εντός της βιβλιοθήκης των νανοσωματιδίων. Το $\boldsymbol{b}$ είναι ένα διάνυσμα των παραμέτρων του μοντέλου, $\boldsymbol{X}$ είναι ένας πίνακας των τιμών των παραμέτρων για κάθε μορφή νανοσωματιδίου, και $\boldsymbol{e}$ είναι το διάνυσμα των σφαλμάτων. Η παράμετρος $\boldsymbol{n}$ είναι ο συνολικός αριθμός των μορφών των νανοσωματιδίων στην βιβλιοθήκη. Η φόρμα του $\boldsymbol{X}$ εξαρτάται από το σετ των παραμέτρων που χρησιμοποιείται για να περιγράψει κάθε μορφή στην βιβλιοθήκη των νανοσωματιδίων. Για μοντέλα που χρησιμοποιούν το αποτύπωμα του πρωτεϊνικού ορού για να περιγράψουν την βιβλιοθήκη των νανοσωματιδίων χρυσού, $\boldsymbol{X} \in R^{84x129}$, όπου κάθε γραμμή του $\boldsymbol{X}$ είναι ένα διάνυσμα 129 στοιχείων της σχετικής περιεκτικότητας της κάθε πρωτεΐνης στο αποτύπωμα του πρωτεϊνικού ορού της κάθε μορφής. [8]

5.2.3 Μέθοδος ελαχίστων τετραγώνων

Η εκτίμηση των παραμέτρων του μοντέλου, δηλαδή των $\boldsymbol{\alpha_1, \alpha_2 \ldots \alpha_j}$ θα γίνει με τη μέθοδο των ελαχίστων τετραγώνων. Σε αυτή την μέθοδο επιλέγονται οι παράμετροι που ελαχιστοποιούν το άθροισμα των τετραγώνων των παρατηρημένων υπολοίπων $\mathbf{e_i}$. Το άθροισμα των τετραγώνων των σφαλμάτων είναι :

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (logy_i - a_j \cdot x_{ij})^2 \qquad (5.16)$$

Παραγωγίζοντας το άθροισμα αυτό ως προς $\boldsymbol{\alpha}$, βρίσκεται που ελαχιστοποιείται το άθροισμα αυτό. Από την λύση των συστημάτων που προκύπτουν υπολογίζεται η τιμή του $\boldsymbol{\hat{a}_j}$.

Η εκτίμηση $log\hat{y}_i = \hat{a}_j \cdot x_{ij}$ της ευθείας παλινδρόμησης, καλείται ευθεία των ελαχίστων τετραγώνων από τον τρόπο υπολογισμού των συντελεστών της. Εδώ θα γίνει η διάκριση μεταξύ της παρατηρούμενης τιμής του $\boldsymbol{logy_i}$ και της μέσης τιμής $\boldsymbol{log\hat{y}_i}$ που εκτιμάται. Η παρατηρούμενη τιμή είναι η πραγματική τιμή του $\boldsymbol{logy_i}$, ενώ η τιμή $\boldsymbol{log\hat{y}_i}$, είναι η τιμή που αναμένεται για το $\boldsymbol{logy_i}$, όταν δοθεί το $\boldsymbol{x_{ij}}$, με την βοήθεια της ευθείας που εκτιμήθηκε με την παραπάνω μέθοδο. Αυτές οι δυο τιμές μπορεί να μην συμπίπτουν και φυσικά όσο μικρότερη είναι η διαφορά $logy_i - log\hat{y}_i$ τόσο καλύτερο είναι το μοντέλο. Η διαφορά αυτή λέγεται κατάλοιπο ή εκτιμημένο υπόλοιπο (residual). [28]

Η στατιστική σημαντικότητα κάθε μοντέλου καθορίζεται με τη χρήση της τεχνικής CV-ANOVA. Αυτή η τεχνική χρησιμοποιεί ένα F-τεστ για να καθορίσει αν το μοντέλο έχει υπόλοιπα πρόβλεψης (prediction residuals), τα οποία είναι σημαντικά μικρότερα από την διακύμανση του κάθε $\boldsymbol{y_i}$, γύρω από το μέσο όρο του $\boldsymbol{y_i}$. Τα μοντέλα θεωρούνται στατιστικά σημαντικά εάν $p<0,05$. [8]

5.2.4 Μερική παλινδρόμηση ελαχίστων τετραγώνων

Για το γενικό γραμμικό μοντέλο εφαρμόστηκε η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης (Multiple Linear Regression) για να υπολογιστούν οι σταθερές $\boldsymbol{b_1, b_2, \ldots b_m}$, και να προταθεί το βέλτιστο μοντέλο που εκφράζει την εξάρτηση της $\boldsymbol{log\, y_i}$ από τις $\boldsymbol{x_{i1}, x_{i2}, x_{i3} \ldots x_{im}}$ μεταβλητές.

Με βάση τα δικά μας δεδομένα δηλαδή για περιεκτικότητες $\boldsymbol{x_{ij}}$ για 129 πρωτεΐνες σε 84 μορφές νανοσωματιδίων, θα απαιτούνται για την παραπάνω ανάλυση άλλες 45 μορφές νανοσωματιδίων, για να βρεθεί μια μοναδική λύση στην κατάρτιση του καταλληλότερου μοντέλου. Αυτή το φαινόμενο καλείται overfitting, κατά το οποίο ο αριθμός των ανεξάρτητων μεταβλητών $\boldsymbol{x_{ij}}$ είναι πολύ μεγαλύτερος από τον αριθμό τα παρατηρήσεων $\boldsymbol{log_2 y_i}$.

Για τους παραπάνω λόγους χρησιμοποιείται η μέθοδος μερικής παλινδρόμησης ελαχίστων τετραγώνων (Partial Least Squares Regression ή PLSR) ώστε να μειωθεί ο αριθμός των ανεξάρτητων μεταβλητών $\boldsymbol{x_{ij}}$. Όταν οι μεταβλητές $\boldsymbol{y_i}$ είναι λίγες σε αριθμό ή μη συγγραμμικές, τότε χρησιμοποιείται η Πολλαπλή Γραμμική Παλινδρόμηση, ενώ όταν μια από τις παραπάνω προϋποθέσεις δεν ισχύει, χρησιμοποιείται η PLSR. Πρόκειται για μια πρόσφατη τεχνική η οποία συνδυάζει χαρακτηριστικά τόσο της Παλινδρόμηση Κυρίων Συνιστωσών όσο και της Πολλαπλής Γραμμικής Παλινδρόμησης.

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis ή PCA) είναι μια μαθηματική μέθοδος η οποία έχει ως στόχο την «συμπίεση» διανυσμάτων σε μικρότερο αριθμό διαστάσεων. Για να το πετύχει αυτό, εκμεταλλεύεται τις συσχετίσεις ανάμεσα στις μεταβλητές των διανυσμάτων που πρόκειται να συμπιεστούν.

Έστω τα αρχικά διανύσματα είναι n διαστάσεων και έχουν μορφή: $\boldsymbol{(x_1, x_2, \dots x_n)}$. Η Ανάλυση Κυρίων Συνιστωσών μεταφέρει τα διανύσματα αυτά σε έναν άλλο χώρο, ο οποίος έχει και αυτός n διαστάσεις και είναι ο χώρος των κυρίων συνιστωσών (Principal Components ή PCs). Η Ανάλυση Κύριων Συνιστωσών μετατρέπει τα αρχικά διανύσματα στην μορφή: $\boldsymbol{(PC_1, PC_2, \dots, PC_n)}$. Έτσι τα νέα διανύσματα έχουν n κύριες συνιστώσες (PCs).

Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους, και είναι υπολογισμένες με τέτοιο τρόπο ώστε το μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος των διανυσμάτων να αντιπροσωπεύεται από όσο το δυνατό λιγότερους PCs. Πιο συγκεκριμένα οι κύριες συνιστώσες συνηθίζεται να διατάσσονται με της εξής φθίνουσα σειρά. Η πρώτη κύρια συνιστώσα $\boldsymbol{PC_1}$ είναι η κύρια συνιστώσα που εκφράζει το μεγαλύτερο ποσοστό της μεταβλητότητας του δείγματος. Και με ανάλογο τρόπο, η n-οστή κύρια συνιστώσα $\boldsymbol{PC_n}$ είναι η κύρια συνιστώσα η οποία

εκφράζει το ελάχιστο ποσοστό της μεταβλητότητας του δείγματος. Όλοι μαζί οι PC_S συνολικά εκφράζουν το 100% της μεταβλητότητας του δείγματος. [28]

Η μείωση των ανεξάρτητων μεταβλητών θα γίνει με βάση δυο σημαντικές παρατηρήσεις. Πρώτον, η απορρόφηση κάποιων ζευγαριών των πρωτεϊνών ορού έχουν υψηλή συσχέτιση με το κύτταρο και γι' αυτό μπορούν να συνδυαστούν χρησιμοποιώντας μια μοναδική παράμετρο. Δεύτερον, κάποιες πρωτεΐνες ορού στο αποτύπωμα έχουν χαμηλή συσχέτιση με το κύτταρο και μπορούν να αποκλειστούν από το μοντέλο. [8]

Στόχος της μερικής παλινδρόμησης είναι να εξάγει τις λανθάνουσες μεταβλητές (latent variables), οι οποίες ερμηνεύουν τη μέγιστη της διασπορά στην απόκριση ενώ παράλληλα οδηγεί στην καλή μοντελοποίηση των αποκρίσεων.

Τα δεδομένα περιγράφονται από i ανεξάρτητες μεταβλητές, ή αλλιώς προβλέπουσες (predictors) των j παρατηρήσεων που συλλέγονται στον πίνακα $\boldsymbol{X}$ και από μία εξαρτημένη μεταβλητών των j παρατηρήσεων σε έναν πίνακα $\boldsymbol{Y}$. Πριν την ανάλυση των δεδομένων, αυτά υπόκεινται σε διαδικασία μετασχηματισμού scaling όπως περιγράφηκε παραπάνω.

Στόχος του μοντέλου είναι η εύρεση νέων μεταβλητών που καλούνται $\boldsymbol{X-scores}$, οι οποίες είναι εκτιμήτριες των λανθάνουσων μεταβλητών. Οι νέες μεταβλητές $\boldsymbol{X-scores}$, συμβολίζονται με $\boldsymbol{t_a}$, όπου $\alpha = 1,2,\dots A$, όπου A ο αριθμός των συνιστωσών, και είναι προβλέπουσες του πίνακα $\boldsymbol{Y}$ και ταυτόχρονα μοντελοποιούν τον πίνακα $\boldsymbol{X}$.

Τα $\boldsymbol{X-scores}$, είναι A τον αριθμό και ορθογώνια. Εκτιμώνται ως γραμμικοί συνδυασμοί των αρχικών $\boldsymbol{X_i}$ μεταβλητών με τους συντελεστές ή βάρη όπως αλλιώς ονομάζονται $\boldsymbol{w_{ja}}$, $\alpha = 1,2,\dots A$.

$$t_{ia} = \sum_j x_{ij} \cdot w_{ja}^* \tag{5.17}$$

Τα $\boldsymbol{X-scores}$ έχουν τις ακόλουθες ιδιότητες:

1. Πολλαπλασιάζονται με τα φορτία (loadings) $\boldsymbol{p_{ak}}$, καλές περιλήψεις του $\boldsymbol{X}$, ώστε τα $\boldsymbol{X-υπόλοιπα}$ στην παρακάτω εξίσωση να είναι μικρά:

$$x_{ij} = \sum_{a} t_{ia} \cdot p_{aj} + e_{ij} \tag{5.18}$$

2. Τα $\boldsymbol{X - scores}$ είναι καλές προβλέπουσες του $\boldsymbol{Y}$, όπως φαίνεται στην ακόλουθη εξίσωση:

$$y_i = \sum_{a} t_{ia} \cdot c_a + f_i \tag{5.19}$$

Τα $\boldsymbol{Y - υπόλοιπα}$ $\boldsymbol{f_i}$ αποτελούν τα στοιχεία του πίνακα υπολοίπων $\boldsymbol{F}$,και ταυτόχρονα εκφράζουν τις αποκλίσεις μεταξύ των παρατηρημένων και των προβλεπόμενων τιμών. Έτσι έχουμε:

$$y_i = \sum_{j} x_{ij} \cdot b_j + f_i \tag{5.20}$$

Όπου $\boldsymbol{b_j}$, οι συντελεστές της PLS παλινδρόμησης. Μπορούν να γραφούν ως :

$$b_j = \sum_{a} w_{ja}^{*} \cdot c_a \tag{5.21}$$

Οι παραπάνω $\boldsymbol{b}$ συντελεστές δεν είναι ανεξάρτητοι εκτός και αν ο αριθμός των PLSR συνιστωσών A ισούται με τον αριθμό των $\boldsymbol{X}$ μεταβλητών j**.** Στην ειδική περίπτωση που διατίθεται μόνο μια μεταβλητή $\boldsymbol{Y}$, τότε απουσιάζει η δομή συσχέτισης στον πίνακα $\boldsymbol{X}$ και το μοντέλο εκφυλίζεται σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

Μετά από κάθε συνιστώσα α, ο πίνακας $\boldsymbol{X}$ υφίσταται ελάττωση, καθώς αφαιρείται η ποσότητα $t_{ia}^{*} \cdot p_{aj}$ από το $\boldsymbol{x_{ij}}$. Έτσι το μοντέλο παλινδρόμησης ελαχίστων τετραγώνων εκφράζεται σε βάρη $\boldsymbol{W_a}$, τα οποία αναφέρονται στα υπόλοιπα προηγούμενης διάστασης, $\boldsymbol{E_{a-1}}$, αντί να σχετίζονται με τις $\boldsymbol{X}$ μεταβλητές. Τότε αντί των παραπάνω εξισώσεων χρησιμοποιούμε τις:

$$t_{ia} = \sum_{k} w_{ja} \cdot e_{ij,a-1} \tag{5.22}$$

$$e_{ij,a-1} = e_{ij,a-2} - t_{i,a-1} \cdot p_{a-1,k} \tag{5.23}$$

$$e_{ij,0} = X_{ij} \tag{5.24}$$

Τα βάρη μπορούν να μετασχηματιστούν στα $\boldsymbol{w}^*$, τα οποία σχετίζονται άμεσα με τον $\boldsymbol{X}$. [28]

Το τελικό μοντέλο που προκύπτει θα έχει την εξής μορφή:

$$log_2(\hat{y}_i) = \sum_{j=1}^{m} b_j x_{i,j} \tag{5.25}$$

όπου m είναι ο συνολικός αριθμός των πρωτεϊνών στο αποτύπωμα του πρωτεϊνικού ορού που χρησιμοποιούνται στο μοντέλο, και $\boldsymbol{b_j}$ είναι μια παράμετρος του μοντέλου που συνδέει την σχετική περιεκτικότητα της πρωτεΐνης j με την κυτταρική συσχέτισης της μορφής του νανοσωματιδίου i.

Στην ερευνητική εργασία του Walkey και των συνεργατών του, διαπιστώνεται ότι το πολλαπλό γραμμικό μοντέλο που χρησιμοποιεί το αποτύπωμα της πρωτεΐνης του ορού προβλέπει την κυτταρική συσχέτιση με 84% μεγαλύτερη ακρίβεια σε σχέση με το απλό γραμμικό μοντέλο που χρησιμοποιεί μόνο τις πρωτεΐνες ορού. Αυτό δείχνει ότι οι ξεχωριστές πρωτεΐνες ορού στο αποτύπωμα κωδικοποιούν τις απαραίτητες πληροφορίες για το νανοσωματίδιο, που μπορούν να χρησιμοποιηθούν για να ενισχυθεί η ακρίβεια της πρόβλεψης.

5.2.5 Μεταβλητή Σημαντικότητας για προβολή: VIP

Κατά την μοντελοποίησης της PLSR διαδικασίας, απομακρύνονται οι παράμετροι, οι οποίοι έχουν μικρή σχέση με το μοντέλο κατά τη επαναλαμβανόμενη διαδικασία του jackknifing με χρήση του μέτρου ***VIP*** (Variable Importance in Projection). Μέσω

αυτού του μέτρου συσσωρεύεται η σημασία της κάθε μεταβλητής j που αντανακλάται από τα βάρη $\boldsymbol{w}$ της κάθε συνιστώσας. Το ***VIP*** μέτρο ορίζεται ως:

$$VIP_j = \sqrt{p \sum_{a=1}^{A} \left[SSE_a \left(w_{aj}/||w_a||^2\right)\right] / \sum_{a=1}^{A} (SSE_a)} \qquad (5.26)$$

Όπου $\boldsymbol{SSE_a}$ είναι το άθροισμα των τετραγώνων της α-οστής συνιστώσας, το $(w_{aj}/||w_a||^2)$ αντιπροσωπεύει την σημαντικότητας της j-οστής μεταβλητής, και $\boldsymbol{p}$ ο αριθμός των προβλέπουσων τιμών των j παρατηρήσεων.

Ο ιδανικός αριθμός των παραμέτρων επιλέγονται με χρήση της μεθόδου jackknifing. Η μέθοδος jackknifing χρησιμοποιείται στην στατιστική για την εκτίμηση της μεροληψίας και του τυπικού σφάλματος (διακύμανση) ενός στατιστικού στοιχείου, όταν ένα τυχαίο δείγμα παρατηρήσεων χρησιμοποιείται για τον υπολογισμό του.

Έστω ότι τα δεδομένα είναι οι περιεκτικότητες της πρώτης πρωτεΐνης $\boldsymbol{X_1, X_2, \ldots X_n}$. Η τεχνική αυτή αντί να δημιουργήσει ένα σύνολο τυχαίων δειγμάτων από τα δεδομένα, παράγει n δείγματα μεγέθους n-1 αποκλείοντας μια παρατήρηση κάθε φορά.

Ακολουθεί τα εξής στάδια:

- Κατασκευάζει ένα δείγμα $X_1, X_2, \ldots X_n$.
- Υπολογίζει μια συνάρτηση των δεδομένων, την $\hat{\theta}(X)$, η οποία εκτιμά μια παράμετρο θ του μοντέλου.
- Για $i = 1$ έως n

  -Κατασκευάζει ένα jackknife δείγμα $X^{-i} = \{X_1, \ldots X_{\iota-1}, X_{i+1}, .., X_n\}$, αποκλείοντας την i παρατήρηση.

  -Υπολογίζει την $\hat{\theta}_{-i}$ εφαρμόζοντας την διαδικασία εκτίμησης στο jackknife δείγμα.
- Υπολογίζει την jackknife εκτιμήτρια $\hat{\theta}_* = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{-i}$ και την jackknife εκτιμήτρια της διασποράς $\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{-i} - \hat{\theta}_*)^2$. [29]

Για κάθε σετ παραμέτρων της διαδικασίας jackknifing υπολογίζεται το ***VIP*** και επιλέγεται το σετ παραμέτρων για το οποίο το μέτρο ***VIP*** είναι το μέγιστο. Έτσι

δημιουργείται ένα νέο μοντέλο, χρησιμοποιώντας το υποβιβασμένο σε αριθμό παραμετρικό σετ. Τα δεδομένα του Συνόλου Γ προκύπτουν από την παραπάνω διαδικασία.

5.2.6 Συντελεστές Παλινδρόμησης: R^2, RMSE

Η δειγματική διασπορά των παρατηρήσεων ορίζεται ως εξής:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{5.27}$$

όπου το πρώτο άθροισμα συμβολίζεται με ***SST***, το δεύτερο με ***SSE*** και το τρίτο με ***SSR***, εκ των οποίων το ***SST*** ερμηνεύει την ολική μεταβλητότητα των $\boldsymbol{y_i}$, το ***SSR*** τη μεταβλητότητα των προβλέψεων $\boldsymbol{\hat{y}_i}$.

Τέλος το ***SSE*** ερμηνεύει τη μεταβλητότητα των $\boldsymbol{y_i}$ σε σχέση με τις αντίστοιχες τιμές που έχουμε προβλέψει μέσω του μοντέλου παλινδρόμησης. Το ποσοστό της μεταβλητότητας των $\boldsymbol{y_i}$, το οποίο ερμηνεύεται από το μοντέλο υπολογίζεται από τον συντελεστή προσαρμογής $\boldsymbol{R^2}$.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.28}$$

Ο συντελεστής προσαρμογής, παίρνει τιμές στο [0,1] και εκφράζει το ποσοστό της διασποράς της τυχαίας μεταβλητής $\boldsymbol{y}$, που εξηγείται με βάση το μοντέλο παλινδρόμησης. Όσο μεγαλύτερες τιμές παίρνει ο συντελεστής προσαρμογής τόσο καλύτερη προσαρμογή της ευθείας έχουμε, υπό την προϋπόθεση ότι το γραμμικό μοντέλο είναι το κατάλληλο. [28]

Η ρίζα μέσου τετραγωνικού σφάλματος (RMSE) είναι ένα συχνά χρησιμοποιούμενο μέτρο της διαφοράς μεταξύ των τιμών που προβλέπονται από ένα μοντέλο και των παρατηρούμενων τιμών από πειραματικά δεδομένα. Αυτές οι διαφορές ονομάζονται υπόλοιπα και η ρίζα μέσου τετραγωνικού σφάλματος τα ενσωματώνει σε ένα μέτρο.

Η RMSE ενός μοντέλου πρόβλεψης με βάση την εκτιμώμενη τιμή του μοντέλου $\hat{y}_i$ ορίζεται ως:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5.29}$$

Η RMSE ενισχύει και τιμωρεί αυστηρά τα μεγάλα σφάλματα. [30]

5.2.7 Επικύρωση του μοντέλου

Κατά τη διαδικασία της προσαρμογής του μοντέλου η PLSR χρησιμοποιείται για να υπολογιστεί η παράμετρος ***b*** κάθε μοντέλου. Στην διαδικασία αυτή το εξαρτώμενο διάνυσμα της αντίδρασης του κυττάρου ***(Y)*** εκφράζεται ως ένας γραμμικός συνδυασμός της ανεξάρτητης παραμέτρου του πίνακα ***(X)***, μέσω των κύριων συνιστωσών του χώρου ***(PCs)***. Οι προβολές της PLSR στο χώρο της PC, έχουν υπολογιστεί ώστε να μεγιστοποιούν την συνδιακύμανση μεταξύ των ***Y*** και ***X*** στο χώρο της PC. Τα χαμηλής τάξης PC περιέχουν την πλειοψηφία της πληροφορίας που είναι χρήσιμη για να εξηγηθεί η διακύμανση στο ***Y***. Ως επακόλουθο, τα υψηλής τάξης PCs μπορούν να αγνοηθούν χωρίς να χαθεί η ακρίβεια του μοντέλου. Για κάθε σετ παραμέτρων, ο ιδανικός αριθμός των PCs καθορίζονται μέσω της τεχνικής της διασταυρωμένης επικύρωσης (Cross-Validation ή CV).

Η Cross-Validation είναι μια στατιστική τεχνική επικύρωσης του μοντέλου για να αποτιμηθεί το πως τα αποτελέσματα μιας στατιστικής ανάλυσης θα γενικευθούν σε ένα νέο ανεξάρτητο σύνολο δεδομένων. Η τεχνική αυτή περιλαμβάνει επαναχρησιμοποίηση (reusing) καθώς και εκ νέου δειγματοληψία (resampling) των δεδομένων. Είναι μια μέθοδος εκτίμησης του σφάλματος πρόβλεψης όταν προσαρμόζεται ένα μοντέλο που σχετίζεται με δύο ή περισσότερες μεταβλητές.

Καθώς εφαρμόζουμε την PLSR του ***Y*** στον ***X***, ένα δείγμα παρατηρήσεων αποτελούμενο από δυο μεταβλητές, $(X_1,Y_1),\ldots,(X_n,Y_n)$, ελαχιστοποιείται το προσαρμοσμένο σφάλμα:

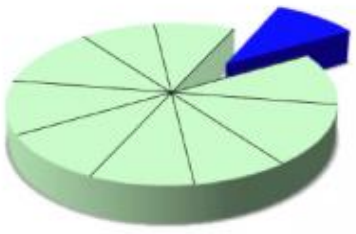
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (\hat{a} + \hat{b}X_i))^2 \quad (5.30)$$

Στην συνέχεια η πρόβλεψη της επόμενης τιμής του ***Y***, δηλαδή του $\boldsymbol{Y_{n+1}}$, καθώς δίνεται μόνο η επόμενη τιμή του ***X***, δηλαδή η $\boldsymbol{X_{n+1}}$, οδηγεί στην ελαχιστοποίηση της συνάρτησης:

$$(Y_{n+1} - (\hat{a} + \hat{b}X_{n+1}))^2 \quad (5.31)$$

όπου τα $\boldsymbol{\hat{a}}$ και $\boldsymbol{\hat{b}}$ έχουν επιλεγεί χωρίς τη χρησιμοποίηση των τιμών $\boldsymbol{X_{n+1}}$ και $\boldsymbol{Y_{n+1}}$. Μια συχνά χρησιμοποιούμενη τεχνική εκτίμησης σφάλματος πρόβλεψης είναι ο διαχωρισμός του δείγματος σε ποσοστό 75% - 25%. Το πρώτο μέρος θα χρησιμοποιηθεί για τον υπολογισμό του προσαρμοσμένου σφάλματος ενώ το δεύτερο μέρος θα χρησιμοποιηθεί για την εκτίμηση του σφάλματος πρόβλεψης. Έτσι δημιουργείται ένα δείγμα για προσαρμογή (fitting) και ένα για επικύρωση (validation), γνωστά και ως training set και testing set αντίστοιχα.

Ευρέως γνωστό παράδειγμα διασταυρωμένης επικύρωσης είναι η K-fold CV. Στην αρχή της διαδικασίας χωρίζονται τα δεδομένα σε ***K*** σύνολα με ίδιο μέγεθος. Η K-fold CV πραγματοποιείται ***K*** φορές. Σε κάθε στάδιο ένα σύνολο (fold) παίζει το ρόλο του testing set ενώ τα υπόλοιπα σύνολα (***K-1***) του training set. Συνήθως τα δεδομένα χωρίζονται σε 10 σύνολα (10-fold CV). [28]



Σχήμα 5.1: K-fold CV

Σε κάθε στάδιο περιλαμβάνεται η απομάκρυνση ενός συνόλου των δεδομένων, η προσαρμογή του μοντέλου στα υπόλοιπα σύνολα δεδομένων και μετά εφαρμόζονται τα απομακρυνόμενα σύνολα δεδομένα στο προσαρμοσμένο μοντέλο.

Τα στάδια της διαδικασίας έχουν ως εξής:

- $1^{ο}$ στάδιο: Το πρώτο σύνολο δεδομένων αποτελεί το testing set. Τα υπόλοιπα ***K-1*** αποτελούν το training set ως μια μεγάλη κατηγορία.

- 2^{o} στάδιο: Το δεύτερο σύνολο δεδομένων αποτελεί το testing set. Τα υπόλοιπα ***K-1*** αποτελούν το training set ως μια μεγάλη κατηγορία.
- …
- Κ στάδιο: Το ***K*** σύνολο δεδομένων αποτελεί το testing set. Τα υπόλοιπα ***K-1*** αποτελούν το training set ως μια μεγάλη κατηγορία.

Τέλος, συλλέγονται όλα τα σφάλματα πρόβλεψης του κάθε σταδίου, τα προσθέτουμε και αυτό δίνει το ρυθμό σφάλματος του CV. Ορίζονται τα ***K*** σύνολα $\boldsymbol{C_1, C_2, \ldots, C_k}$ όπου $\boldsymbol{C_k}$ υποδηλώνει του δείκτες των παρατηρήσεων στο μέρος ***k***. Υπάρχουν $\boldsymbol{n_k}$ παρατηρήσεις στο σύνολο ***k***. Αν ***n*** είναι ένα πολλαπλάσιο του ***K***, τότε $n_k = \frac{n}{K}$. Η εκτιμήτρια του σφάλματος για το CV είναι:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k, \tag{5.32}$$

όπου $MSE_k = \frac{1}{n_k} \sum_{i \in C_k} (y_i - \hat{y}_i)^2$,

όπου:

- $\boldsymbol{\hat{y}_i}$ είναι η εκτίμηση της παρατήρησης ***i***, που προκύπτει από τα δεδομένα με το σύνολο ***k*** απομακρυσμένο
- το μέσο τετραγωνικό σφάλμα (***MSE***) που προκύπτει από την προσαρμογή του μοντέλου στα ***K-1*** σύνολα που δεν περιλαμβάνουν τα ***K*** σύνολα. Αυτό δίνει την εκτίμηση $\boldsymbol{\hat{y}_i}$ της παρατήρησης ***i***.
- $\boldsymbol{MSE_k}$, το άθροισμα των σφαλμάτων

Τα σφάλματα του μοντέλου κατά τη διαδικασία του CV δίδονται και στη μορφή του ***RMSE*** που αναφέρθηκε στο υποκεφάλαιο 5.2.6.

Το πλεονέκτημα της μεθόδου K-fold CV είναι ότι όλα τα σύνολα δεδομένων ***K*** χρησιμοποιούνται για την προσαρμογή και την επικύρωση του μοντέλου.

Μια άλλη μέθοδος διασταυρωμένης επικύρωσης είναι η Leave-one out CV, η οποία αποτελεί ειδική περίπτωση της K-fold CV όπου ο αριθμός των συνόλων είναι ίδιος με τον αριθμό των παρατηρήσεων (K=N). Υπάρχει ένα σύνολο για κάθε παρατήρηση και έτσι κάθε παρατήρηση παίζει το ρόλο του testing set. Οι υπόλοιπες ***n-1*** παρατηρήσεις παίζουν το ρόλο του training set. [31]

Στην ερευνητική εργασία [8] χρησιμοποιήθηκε η Leave-one out CV κατά την οποία σε κάθε βήμα απομακρύνονται οι παρατηρήσεις που αφορούν μια μορφή νανοσωματιδίου από τα δεδομένα και ένα νέο μοντέλο προσαρμόζεται χρησιμοποιώντας τα υπόλοιπα δεδομένα. Το προκύπτον μοντέλο χρησιμοποιείται μετά για να προβλέψει την κυτταρική συσχέτιση της μορφής του νανοσωματιδίου που αποκλείστηκε. Ο συντελεστής προσδιορισμού της διαδικασίας ορίζεται ως:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{(LOO)i})^2}{\sum_{i=1}^{n}(y_i - \tilde{y})^2} \quad (5.33)$$

,όπου $\boldsymbol{\hat{y}_{(LOO)i}}$ είναι η τιμή της κυτταρικής συσχέτισης της μορφής του νανοσωματιδίου $\boldsymbol{i}$ που εκτιμήθηκε κατά την διαδικασία. Τα μοντέλα με $\boldsymbol{Q^2_{LOO}}$>0,7 θεωρούνται καλές προσεγγίσεις. [8]

Κατά την διαδικασία του cross-validation που πραγματοποιήθηκε από την ερευνητική ομάδα της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής, υπολογίζεται η ακρίβεια της προσαρμογής ομάδων πρωτεϊνών στο μοντέλο αυτό. Η ακρίβεια αυτή προσδιορίζεται ποσοτικά με τον συντελεστή προσδιορισμού $\boldsymbol{Q^2}$ (εξίσωση 5.33). Συγκεκριμένα υπολογίζεται με τη μέθοδο απαλοιφής μεταβλητών προς τα πίσω (backwards elimination), δηλαδή κάθε φορά απομακρύνεται μια μεταβλητή x (περιεκτικότητα πρωτεΐνης) από το μοντέλο. Το προκύπτον μοντέλο χρησιμοποιείται μετά για να προβλέψει την κυτταρική συσχέτιση της πρωτεΐνης που απομακρύνθηκε.

Το μοντέλο προσαρμόζεται κάθε φορά για κάθε σύνολο πρωτεϊνών και έτσι προκύπτουν οι συντελεστές $\boldsymbol{Q^2}$. Η διαδικασία backwards elimination ξεκινά από την έκτη σε σειρά κατάταξης πρωτεΐνη έτσι όπως προκύπτει από το απλό γραμμικό μοντέλο. Τα δεδομένα του Συνόλου Β προκύπτουν από την παραπάνω διαδικασία.

6 *Λειτουργικά προφίλ και GO γραφήματα*

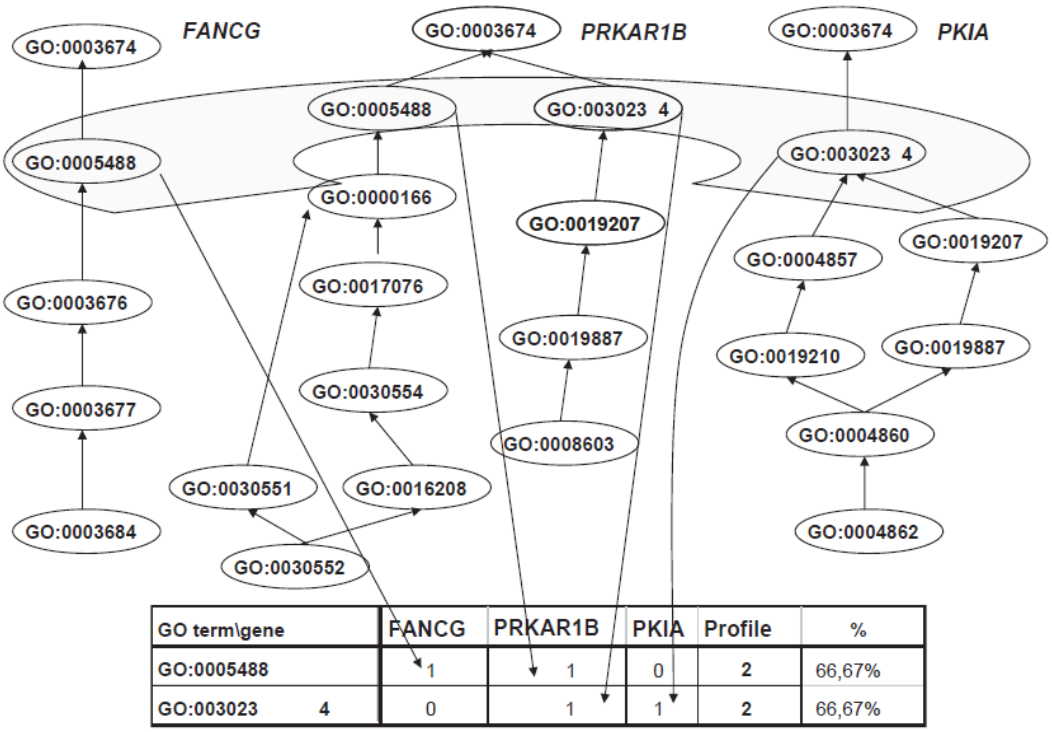
Η ανάλυση των LC-MS/MS μας παρέχει μια συλλογή από πειραματικά δεδομένα. Οι διαφορετικές τεχνικές ανάλυσης, χρησιμοποιούνται ώστε να κατανοηθούν οι βιολογικές διεργασίες πχ. στατιστικά τεστ χρησιμοποιούνται για τον εντοπισμό των σημαντικών πρωτεϊνών/γονιδίων. Άλλες μέθοδοι, ομαδοποιούν τα γονίδια ανάλογα με τα προφίλ έκφρασης τους. Η υπόθεση είναι ότι τα γονίδια με προφίλ έκφρασης παρόμοια με γνωστά γονίδια, εμπλέκονται σε παρόμοιες βιολογικές διεργασίες. Οι ερευνητές καταλήγουν σε μια μεγάλη λίστα των σημαντικών γονιδίων. Το δεύτερο βήμα είναι οι βιολόγοι να κατηγοριοποιήσουν αυτά τα γονίδια σε γνωστές βιολογικές λειτουργίες και να συνδυάσουν τα πειραματικά αποτελέσματα με βιολογική γνώση. Τέτοιες πληροφορίες παρέχονται από την Οντολογία Γονιδίων, το πιο δημοφιλές σύστημα για μετάφραση γονιδίων σε γονιδιακά προϊόντα με χρήση ελεγχόμενου και δομημένου λεξιλογίου, όπως αναλύθηκε στο Κεφάλαιο 3.

6.1 goProfiles: λειτουργικά προφίλ

Σύμφωνα με τα παραπάνω, η GO είναι μια πλούσια δομή δεδομένων, η οποία περιέχει μια μεγάλη ποσότητα πληροφοριών για την σχέση μεταξύ των GO όρων. Διαθέτοντας μια λίστα με επιλεγμένα ενδιαφέροντα γονίδια, κάποιος μπορεί να δημιουργήσει το επαγόμενο υπο-γράφημα, το οποίο δημιουργείται από τμήμα της γονιδιακής οντολογίας, οι κόμβοι του οποίου συνδέονται με τα γονίδια της λίστας, είτε άμεσα είτε μέσω άλλων κόμβων. Αυτά τα γραφήματα μπορούν να έχουν μεγάλες, σύνθετες δομές, ειδικά όταν η λίστα από την οποία προέρχονται είναι μεγάλη. Προκειμένου να απλουστευθεί αυτή η δομή, μπορεί να κοπεί σε τμήματα, ή να προβληθεί στους κόμβους, οι οποίοι είναι σε συγκεκριμένη απόσταση από τον κόμβο κορυφής. Αυτό θα δημιουργήσει έναν πίνακα συχνοτήτων, με κάθε κελί να περιέχει τον αριθμό των γονιδίων που αντιστοιχούν σε κάθε όρο στο επίπεδο, στο οποίο έχει επιλεχθεί να κοπεί. (Σχήμα 6.1) Η δομή του πλέγματος των γραφημάτων, δηλώνει ότι ένα γονίδιο μπορεί να εμφανίζεται σε πολλαπλά κελιά του πίνακα, τα οποία αποτελούν το λειτουργικό προφίλ.

Ένα λειτουργικό προφίλ μπορεί να περιγραφεί ως ένα αριθμητικό διάνυσμα με ονοματισμένα κελιά. Κάθε κελί αντιστοιχεί σε μια κατηγορία για μια δεδομένη οντολογία, συνήθως, αλλά όχι απαραίτητα, στο ίδιο επίπεδο της GO οντολογίας.

Κάθε κατηγορία μπορεί να χαρακτηριστεί από έναν μοναδικό αριθμό (π.χ. ***GO :nnnnnnn***) και ένα περιγραφικό όνομα. Όταν ένας κόμβος είναι στο επίπεδο ***k*** σημαίνει ότι το μικρότερο μονοπάτι μεταξύ αυτού και του κορυφαίου κόμβου σε κάθε GO οντολογία (MF, BP και CC) έχει ***k-1*** κόμβους. Ο αριθμός του κελιού αντιπροσωπεύει τον αριθμό των γονιδίων, των οποίων το μονοπάτι στο βασικό επίπεδο έχει έναν κόμβο σε αυτήν την κατηγορία.



GO term\gene	FANCG	PRKAR1B	PKIA	Profile	%
GO:0005488	1	1	0	2	66,67%
GO:003023 4	0	1	1	2	66,67%

Σχήμα 6.1: Ένα απλό λειτουργικό προφίλ, στο επίπεδο 2 της οντολογίας της μοριακής λειτουργίας. Για απλοποίηση αυτό βασίζεται μόνο σε τρία γονίδια, και απεικονίζει το γεγονός ότι ένα δεδομένο γονίδιο μπορεί να εμφανίζεται σε διαφορετικές κατηγορίες **[11]**

Ο Πίνακας 6.1 παρουσιάζει ένα λειτουργικό προφίλ για ένα σετ 140 γονιδίων που ταξινομούνται στο δεύτερο επίπεδο της οντολογίας μοριακής λειτουργίας. Όταν αναλύονται τα δεδομένα από τα προφίλ, υποβιβάζεται η δομή των αρχικών δεδομένων, όπως συμβαίνει σε κάθε κατηγοριοποίηση. Ένα γονίδιο ανήκει σε παραπάνω από μια λειτουργική κατηγορία όταν το άθροισμα των γονιδίων που ανήκουν σε μια κατηγορία είναι μεγαλύτερο από τον αριθμό των γονιδίων που ταξινομούνται.

Πίνακας 6.1: Λειτουργικό προφίλ για σετ 140 γονιδίων στο δεύτερο επίπεδο της οντολογίας μοριακής λειτουργίας **[11]**

Description	GOID	Sample	Theoritical
Antioxidant activity	GO:0016209	1	0.00
binding	GO:0005488	93	0.88
Catalytic activity	GO:0003824	43	0.42
Enzyme regulator activity	GO:0030234	7	0.07
Motor activity	GO:0003774	1	0.01
Signal transducer activity	GO:0004871	26	0.23
Structural molecule activity	GO:0005198	4	0.04
Transcription regulator activity	GO:0030528	12	0.14
Transporter activity	GO:0005215	16	0.09
Total		203(>140)	1.88(>1)

*Το ”Sample” αντιπροσωπεύει τον αριθμό των επιλεχθέντων γονίδιων που υπάγονται σε κάθε κατηγορία του λειτουργικού προφίλ. Το ”Theoritical ” αντιπροσωπεύει το ποσοστό των γονιδίων των επιλεχθέντων γονιδίων που υπάγονται σε κάθε κατηγορία. Παρατηρείται ότι κάθε γονίδιο ανήκει σε παραπάνω από μια κατηγορία, όπως φαίνεται και στην γραμμή ”Total”.

6.1.1 Κατανομή λειτουργικού προφίλ

Δεδομένου ενός προφίλ με $\boldsymbol{s}$ κατηγορίες, θεωρείται $\Omega = \{A_1, \dots, A_S\}$ το διάστημα των γεγονότων που αντιστοιχούν στην παρατήρηση ενός γονιδίου σε μια από τις κατηγορίες $1, \dots . s$. Δεδομένης της πιθανότητας το ίδιο γονίδιο να ανήκει σε περισσότερες από μια κατηγορίες θα πρέπει να θεωρηθεί το διάστημα των γεγονότων:

$$\Omega^* = \{A_1, A_1 \times A_2, \dots, A_1 \times A_S, A_2, A_2 \times A_3, \dots, A_{s-1} \times A_s\} \qquad (6.1)$$

όπου $\boldsymbol{A_i}$ δηλώνει ότι ένα γονίδιο έχει ταξινομηθεί στην κατηγορία i, και $\boldsymbol{A_i \times A_j}$ δηλώνει ότι ταξινομείται στις κατηγορίες i και j. Με βάση αυτή την σταυρό-δομημένη προσέγγιση, κάθε γονίδιο θα εμφανίζεται το μέγιστο μία φορά σε κάθε κατηγορία, έτσι ώστε η ταξινόμηση αυτή να μπορεί να χαρακτηρίζεται από ένα διευρυμένο προφίλ:

$$n\mathcal{P} = n(p_{11}, p_{12}, p_{1s}, \dots, p_{(s-1)s}, p_{ss})^t, \quad (6.2)$$

έτσι ώστε το διευρυμένο αυτό προφίλ να συνδέεται με την πολυωνυμική κατανομή:

$$n\widehat{P} \sim M(n; P) \quad (6.3)$$

όπου $\boldsymbol{n}$ είναι ο αριθμός των γονιδίων που διαμορφώνουν το προφίλ (για ένα επίπεδο της δεδομένης οντολογίας), $\boldsymbol{p_{ij}}$ η πιθανότητα του γονιδίου να ανήκει μόνο στην κατηγορία $\boldsymbol{A_i}$, και $p_{ij} = p_{ji}$ η πιθανότητα το γονίδιο να ανήκει ταυτόχρονα στις κατηγορίες $\boldsymbol{A_i}$ και $\boldsymbol{A_j}$.[11]

6.1.2 Από τα δεδομένα στα λειτουργικά προφίλ

Ένα λειτουργικό προφίλ «χτίζεται» από μια λίστα γονιδίων, που αντιστοιχίζεται στην οντολογία γονιδίων, πρώτα μέσω των ειδικών όρων και μετά μέσω του γραφήματος που παράγεται από αυτούς τους όρους.

Για να δημιουργηθεί ένα λειτουργικό προφίλ από ένα σετ γονιδίων απαιτούνται κάποια βήματα. Παρακάτω περιγράφονται εν συντομία :

- Για κάθε γονίδιο απαιτούνται οι αντιστοιχιζόμενοι GO όροι για μια δεδομένη οντολογία. Τα ονόματα των γονιδίων πρέπει να είναι σε κοινή μορφή όπως Entrez ταυτότητες γονιδίων.
- Εντοπισμός των σχετιζόμενων με τα γονίδια GO όρων σε ένα δεδομένο επίπεδο. Αυτό επιτυγχάνεται μέσω του επαγόμενου GO γραφήματος, από τους ειδικούς όρους μέχρι το επιθυμητό επίπεδο. Για να πραγματοποιηθεί αυτό ανακτώνται όλοι οι πρόγονοι ενός GO όρου και καθορίζονται οι GO όροι σε κάθε επίπεδο.
- Εφόσον διατίθενται οι GO όροι που συνδέονται με τα γονίδια που διαθέτουμε, και οι GO όροι που αντιστοιχούν στο επιθυμητό επίπεδο, ένας πίνακας διασταύρωσης ή αλλιώς πίνακας συνάφειας (crosstabulation) μεταξύ των δύο λιστών αποδίδει το επιθυμητό προφίλ.

Όλα αυτά τα βήματα ενσωματώνονται σε μια λειτουργία της βιβλιοθήκης του Bioconductor “goProfiles” [12], η οποία χρησιμοποιείται για να «χτίζει» τα προφίλ σε ένα επίπεδο GO των επιθυμητών οντολογιών. [11]

6.1.3 Επίπεδο οντολογίας

Ο όρος «επίπεδο» οντολογίας αντιπροσωπεύει έναν ορισμένο σταθερό αριθμό GO όρων, διαχωρίζοντας τους όρους που ανήκουν σε αυτό το επίπεδο με την κορυφή της οντολογίας. Για παράδειγμα, εάν ένας GO όρος (κόμβος) είναι στο τρίτο επίπεδο, αυτό σημαίνει ότι το συντομότερο μονοπάτι, ξεκινώντας από τον όρο αυτό μέχρι την κορυφαίο κόμβο περιλαμβάνει δύο συνδέσεις. Σε αυτή την βιβλιοθήκη χρησιμοποιείται ένας διευρυμένος ορισμός του επιπέδου.

Για μια οντολογία, ***O***, ένα επίπεδο L^O ορίζεται από ένα σύνολο κόμβων που επαληθεύουν τις παρακάτω ιδιότητες:

1. Δημιουργείται ένα διαμέρισμα του γραφήματος G^O,της οντολογίας ***O***:

$$G^O = L_{down} \cup L^O \cup L_{up} \tag{6.4}$$

   όπου L_{down}, περιλαμβάνει τους κόμβους κάτω από το L^O, το οποίο είναι οποιοδήποτε μονοπάτι που ενώνει οποιοδήποτε κόμβο στο L_{down} με τον κορυφαίο κόμβο του γραφήματος (root node) που περνά μέσα από έναν κόμβο στο L^O. Το L_{up} περιλαμβάνει τους κόμβους πάνω από το L^O, το οποίο είναι ένα οποιοδήποτε μονοπάτι που ενώνει οποιοδήποτε κόμβο στο L_{up} με τον κορυφαίο κόμβο του γραφήματος που δεν περνά μέσα από έναν κόμβο στο L^O.
2. Το «κόψιμο» του γραφήματος είναι ολοκληρωμένο, όταν όλοι οι κόμβοι στην οντολογία είναι πάνω ή κάτω του L^O.
3. Κανένας όρος στο L^O δεν περνά μέσω άλλου στοιχείου στο L^O, στο μονοπάτι του μέχρι τον κορυφαίο κόμβο. [12]

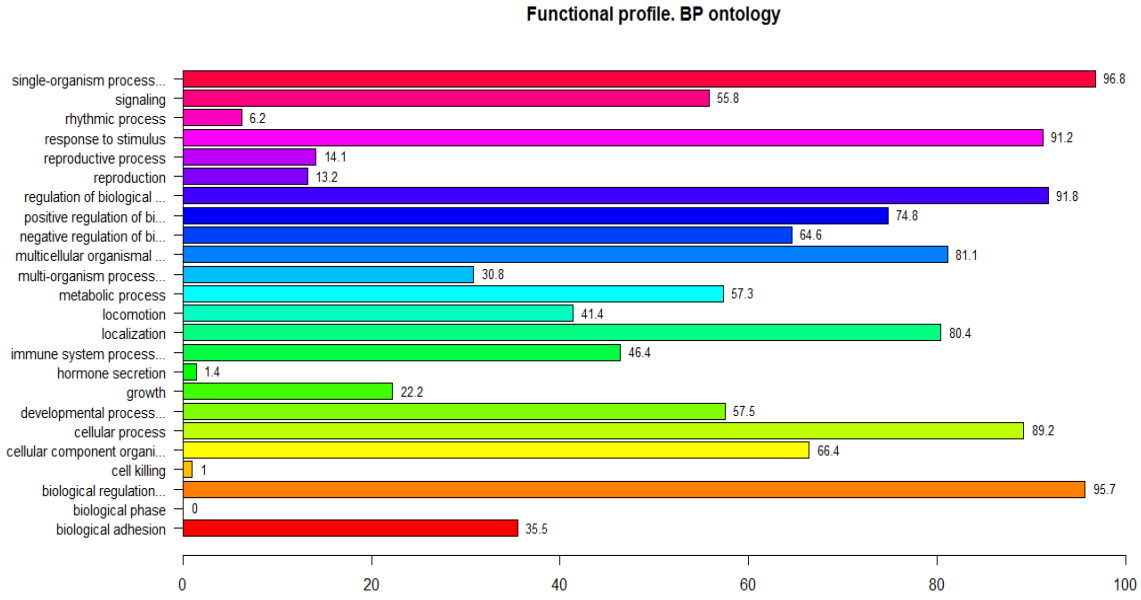
6.1.4 Σχεδιασμός λειτουργικών προφίλ

Από την βιβλιοθήκη “goProfiles” χρησιμοποιούνται δυο λειτουργίες για την δημιουργία των επιθυμητών προφίλ:

- ✓ Η πρώτη λειτουργία, πραγματοποιεί τους υπολογισμούς που απαιτούνται για την δημιουργία του λειτουργικού προφίλ για μια δεδομένη λίστα γονιδίων σε μια οντολογία και για ένα ορισμένο επίπεδο οντολογίας. Ως γονίδια εδώ εισάγονται οι Entrez ταυτότητες των γονιδίων, και γι' αυτό απαιτείται να οριστεί το μεταφραστικό πακέτο "org.Hs.eg.db", ώστε να αντιστοιχιστούν σε GO ταυτότητες βιολογικών λειτουργιών. Ως επίπεδο οντολογίας ορίζεται το δεύτερο (level=2).
- ✓ Η δεύτερη λειτουργία, σχεδιάζει σε ραβδόγραμμα το αποτέλεσμα της πρώτης λειτουργίας. Στον x άξονα εμφανίζονται οι περιγραφές των λειτουργιών των γονιδίων, ενώ στον άξονα y εμφανίζεται το ποσοστό των γονιδίων που περιγράφονται από την κατηγορία αυτή.

6.1.4.1 Λειτουργικά προφίλ Γονιδίων Συνόλου Α

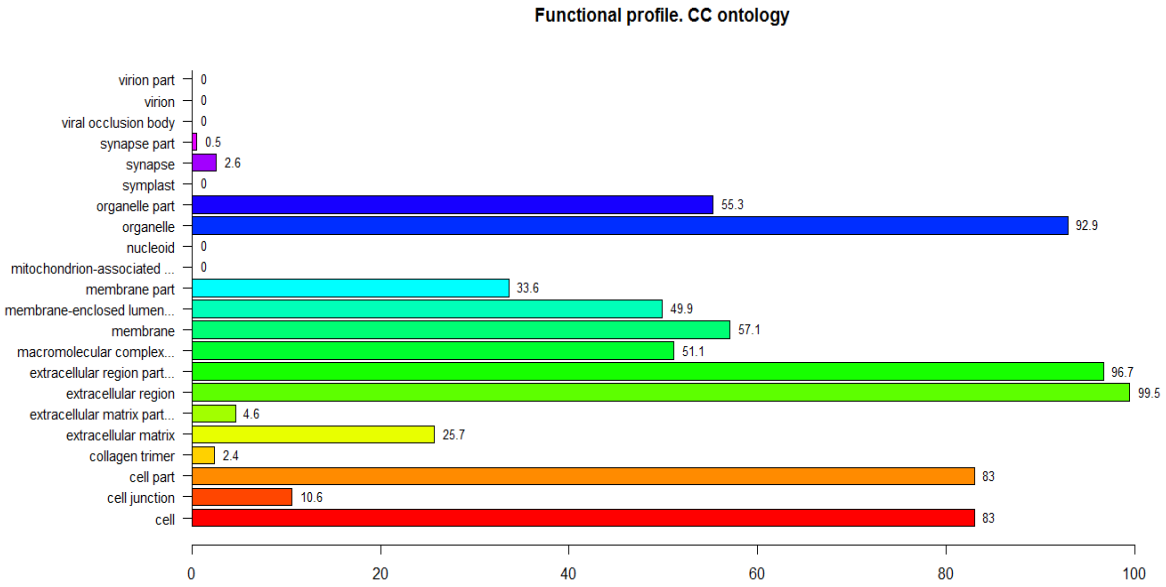
Από το Σύνολο Α, λαμβάνονται 129 Uniprot ταυτότητες πρωτεϊνών οι οποίες μεταφράζονται σε 118 Entrez ταυτότητες γονιδίων, μέσω της βιβλιοθήκης "org.Hs.eg.db" και έπειτα χρησιμοποιούνται ως είσοδος για την πρώτη λειτουργία, που αναφέρθηκε παραπάνω. Ο σχεδιασμός των προφίλ θα γίνει για τις τρεις οντολογίες.



Σχήμα 6.2: Λειτουργικό προφίλ γονιδίων Συνόλου Α σε οντολογία BP

Για το λειτουργικό προφίλ σε οντολογία BP (Σχήμα 6.2), οι κατηγορίες των βιολογικών λειτουργιών που συμμετέχουν τα γονίδια της λίστας και το αντίστοιχο ποσοστό συμμετοχή τους, είναι:

- διαδικασία ενός οργανισμού (single-organism process): 96,8%
- βιολογική διαδικασία ρύθμισης (biological regulation process): 95,7%
- ρύθμιση βιολογικής διαδικασίας (regulation of biological process): 91,8%
- απόκριση σε ερέθισμα (response to stimulus): 91,2%
- κυτταρική διαδικασία (cellular process): 89,2%
- διαδικασία πολυκύτταρου οργανισμού (multicellular organismal process): 81,1%
- εντοπισμός θέσης (localization): 80,4%
- θετική ρύθμιση βιολογικής διαδικασίας (positive regulation of biological process): 74,8%

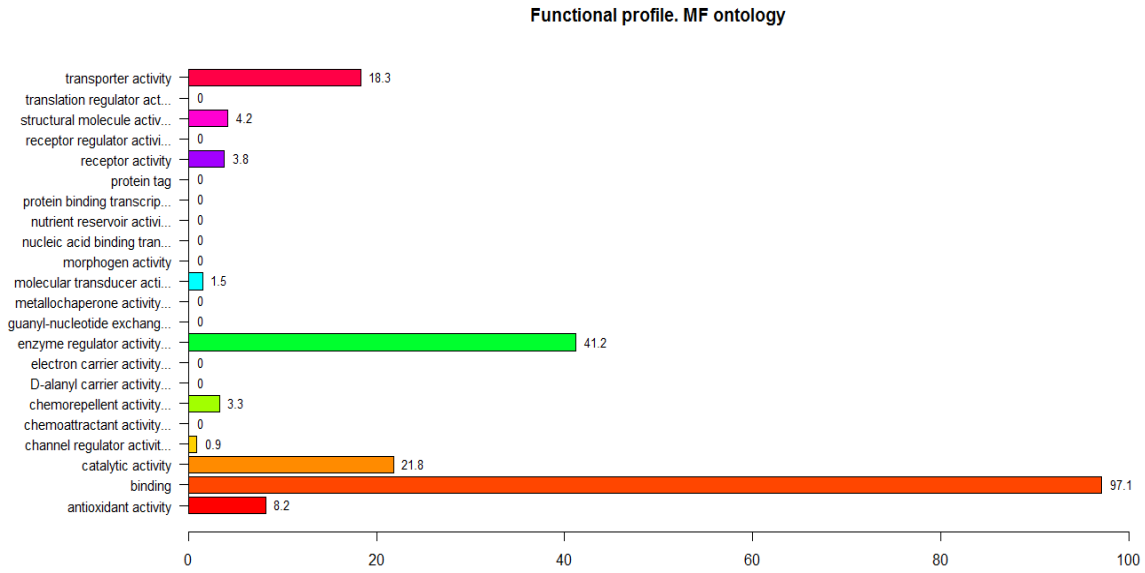


Σχήμα 6.3 : Λειτουργικό προφίλ γονιδίων Συνόλου Α σε οντολογία CC

Για το λειτουργικό προφίλ σε οντολογία CC (Σχήμα 6.3), οι κατηγορίες περιγράφουν υποκυτταρικές δομές, θέσεις και μακρομοριακά σύμπλοκα, και η παρουσία των γονιδίων σε αυτές είναι:

- εξωκυττάρια περιοχή (extracellular region): 99,5%
- τμήμα εξωκυττάριας περιοχής (extracellular region part): 96,7%

- τμήμα οργανιδίου (organelle part): 92,9%
- τμήμα κυττάρου (cell part): 83%
- κύτταρο (cell): 83%



Σχήμα 6.4: Λειτουργικό προφίλ γονιδίων Συνόλου Α σε οντολογία MF

Για το λειτουργικό προφίλ σε οντολογία MF (Σχήμα 6.4), οι κατηγορίες οι οποίες περιγράφουν τις λειτουργιές των γονιδιακών προϊόντων και η συνεισφορά των γονιδίων που εξετάζονται σε αυτές είναι:

- δέσμευση (binding): 97,1%:
- ρυθμιστής δραστηριότητας του ενζύμου (enzyme regulator activity): 41,2%

Οι παραπάνω είναι οι λειτουργικές δράσεις σε επίπεδα βιολογικής λειτουργίας, κυτταρικής σύστασης και μοριακής λειτουργίας των γονιδιακών προϊόντων της πλειοψηφίας των γονιδίων που συνδέονται με τα νανοσωματίδια και προσδιορίζουν εμμέσως τις ιδιότητες των νανοσωματιδίων.

6.1.4.2 Λειτουργικά προφίλ γονιδίων Συνόλου Β

Από το Σύνολο Β λαμβάνονται οι 76 συμβολικές ονομασίες των γονιδίων. Τα σύμβολα των γονιδίων μεταφράζονται σε Entrez ταυτότητες γονιδίων, μέσω της βιβλιοθήκης “org.Hs.eg.db”. Για αυτό το σύνολο δεδομένων θα δημιουργηθούν προφίλ για την κάθε υποομάδα των δύο ομαδοποιήσεων των δεδομένων.

Η πρώτη ομαδοποίηση χωρίζει τα δεδομένα με βάση την ταξινομημένη με ελλατούμενη σημαντικότητα, λίστα που παρέχεται από την αρχική εργασία και ικανοποιώντας το κριτήριο ότι κάθε ομάδα θα περιέχει τουλάχιστον τα πέντε πιο σημαντικά γονίδια της ανάλυσης, δηλαδή τα γονίδια με συμβολικές ονομασίες AMBP,HABP2,ITIH2,TTHY και ITIH1. Δημιουργούνται οι εξής 5 υποομάδες :

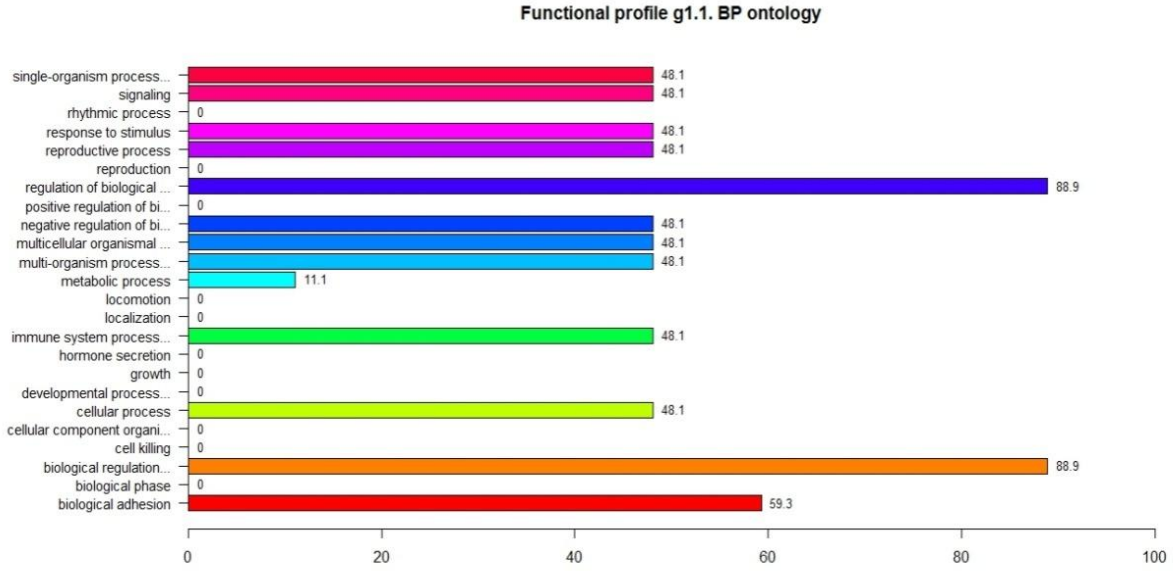
1. Τα γονίδια :
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3
2. Τα γονίδια της $1^{ης}$ ομάδας και τα γονίδια:
   A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, ENPL
3. Τα γονίδια των δύο παραπάνω ομάδων και τα γονίδια :
   PRG4, CNDP1, PROZ, ANT3, CD180, CRP, PLEK, APOB, FHR1, KLKB1, APOC1, SEPP1, HRG, IGLL5, HEP2, APOE
4. Τα γονίδια των τριών παραπάνω ομάδων και τα γονίδια :
   IGKC, IGHG4, CERU, MUCB, APOL1, ZPI, KNG1, APOM, HPTR, VTNC, IGHM, CD5L, FA11, ITH4, PF4V
5. Τα γονίδια των τεσσάρων παραπάνω ομάδων και τα γονίδια :
   GELS, KV302, CALR, FA5, APOC4, ANGI, SAA4, IPSP, APOA1, C1QA, TRFL, CO9, FA10, FA9, C4BPA, C4BPB, TSP4, TETN, CBPN, AACT, FA7, PROS, LBP, IGHG1, COMP, C1R, A1BG, TSP1, APOA4

Η δεύτερη ομαδοποίηση χωρίζει τα δεδομένα σε 6 υποομάδες με βάση τα ταξινομημένα δεδομένα ως προς το $\boldsymbol{Q^2}$ και ικανοποιώντας το κριτήριο ότι κάθε ομάδα θα περιέχει τουλάχιστον τα πέντε πιο σημαντικά γονίδια της ανάλυσης, δηλαδή αυτά που έχουν $\boldsymbol{Q^2}$ ίσο με ένα. Έτσι δημιουργούνται οι εξής 6 ομάδες :

1. Τα γονίδια με $\boldsymbol{Q^2}=1$ και με $\boldsymbol{Q^2}\leq 0{,}6349$:
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3
2. Τα γονίδια με $\boldsymbol{Q^2}=1$ και με $\boldsymbol{Q^2}<0{,}67$:
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3, A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, PRG4
3. Τα γονίδια με $\boldsymbol{Q^2}=1$ και με $\boldsymbol{Q^2}<0{,}7$:
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3, A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, PRG4, ENPL, ANT3, CD180, CRP, PLEK
4. Τα γονίδια με $\boldsymbol{Q^2}=1$ και με $\boldsymbol{Q^2}<0{,}76$:

AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3, A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, PRG4, ENPL, ANT3, CD180, CRP, PLEK, CNDP1, PROZ, APOB, FHR1, KLKB1, HEP2, APOE, IGKC, KNG1, APOM, HPTR, VTNC, IGHM, CD5L, FA11, PF4V, GELS, KV302

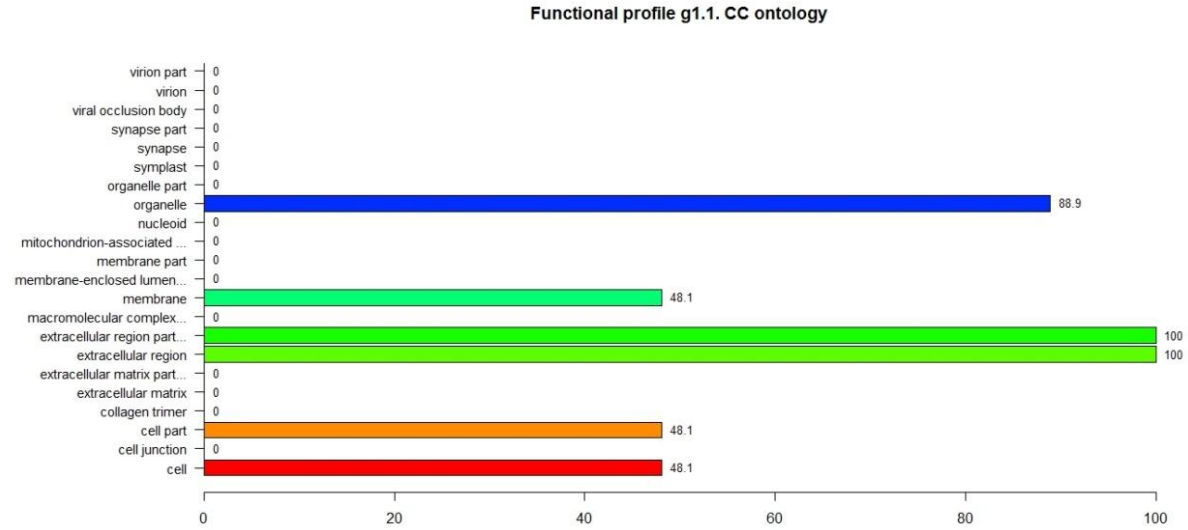
5. Τα γονίδια με Q^2=1 και με Q^2<0,78:
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3, A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, PRG4, ENPL, ANT3, CD180, CRP, PLEK, CNDP1, PROZ, APOB, FHR1, KLKB1, HEP2, APOE, IGKC, KNG1, APOM, HPTR, VTNC, IGHM, CD5L, FA11, PF4V, GELS, KV302, APOC1, SEPP1, HRG, IGLL5, IGHG4, CERU, APOL1, ZPI, ITIH4, CALR, FA5, APOC4, ANGI, SAA4, IPSP, APOA1, TRFL, CO9, FA10, FA9, C4BPA, C4BPB, TSP4
6. Τα γονίδια με Q^2=1 και με Q^2≤0,8189 :
   AMBP, HABP2, ITIH2, TTHY, ITIH1, CO3, A1AT, ITIH3, CO4B, A2AP, SPP24, LUM, PROC, A2MG, ANGT, PRG4, ENPL, ANT3, CD180, CRP, PLEK, CNDP1, PROZ, APOB, FHR1, KLKB1, HEP2, APOE, IGKC, KNG1, APOM, HPTR, VTNC, IGHM, CD5L, FA11, PF4V, GELS, KV302, APOC1, SEPP1, HRG, IGLL5, IGHG4, CERU, APOL1, ZPI, ITIH4, CALR, FA5, APOC4, ANGI, SAA4, IPSP, APOA1, TRFL, CO9, FA10, FA9, C4BPA, C4BPB, TSP4, C1QA, TETN, CBPN, FA7, PROS, LBP, IGHG1, COMP, CIR, A1BG, TSP1, APOA4



Σχήμα 6.5: Λειτουργικό προφίλ γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία BP

Σε οντολογία BP, οι βιολογικές λειτουργίες που αφορούν τα πιο σημαντικά γονίδια της ανάλυσης και το ποσοστό συμμετοχής τους σε αυτές είναι :

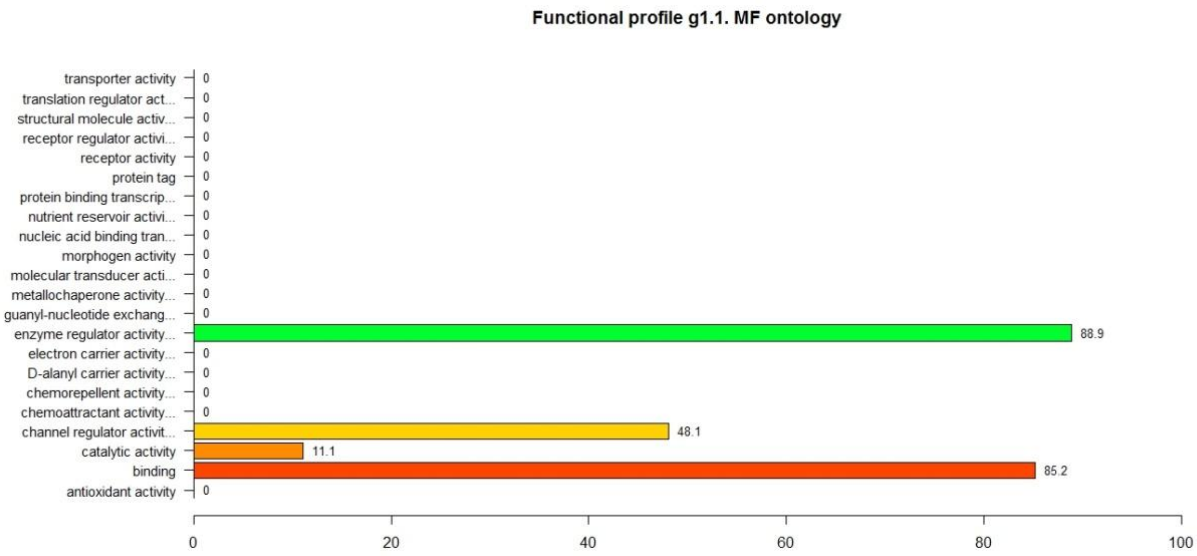
- ρύθμιση βιολογικής διαδικασίας: 88,9%
- βιολογική διαδικασία ρύθμισης: 88,9%
- βιολογική προσκόλληση **(**biological adhesion): 59,3%



Σχήμα 6.6 : Λειτουργικό προφίλ γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία CC

Ομοίως σε οντολογία CC, οι κυτταρικές δομές που σχετίζονται τα πιο σημαντικά γονίδια της ανάλυσης είναι:

- τμήμα εξωκυττάριας περιοχής: 100%
- εξωκυττάρια περιοχή: 100%
- οργανίδιο (organelle): 88,9%

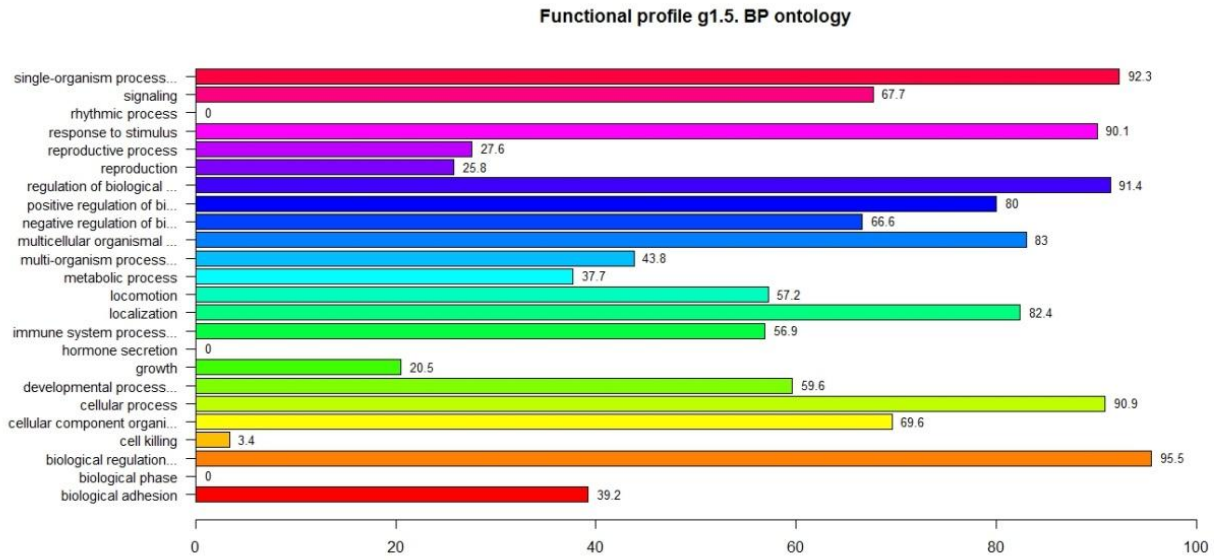


Σχήμα 6.7: Λειτουργικό προφίλ γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία MF

Σε οντολογία MF, οι λειτουργίες των γονιδιακών προϊόντων που αφορούν τα πιο σημαντικά γονίδια της ανάλυσης είναι:

- ρυθμιστής δραστηριότητας ενζύμου: 88,9%
- δέσμευση: 85,2,%
- ρυθμιστής δραστηριότητας καναλιού (channel regulator activity): 48,1%

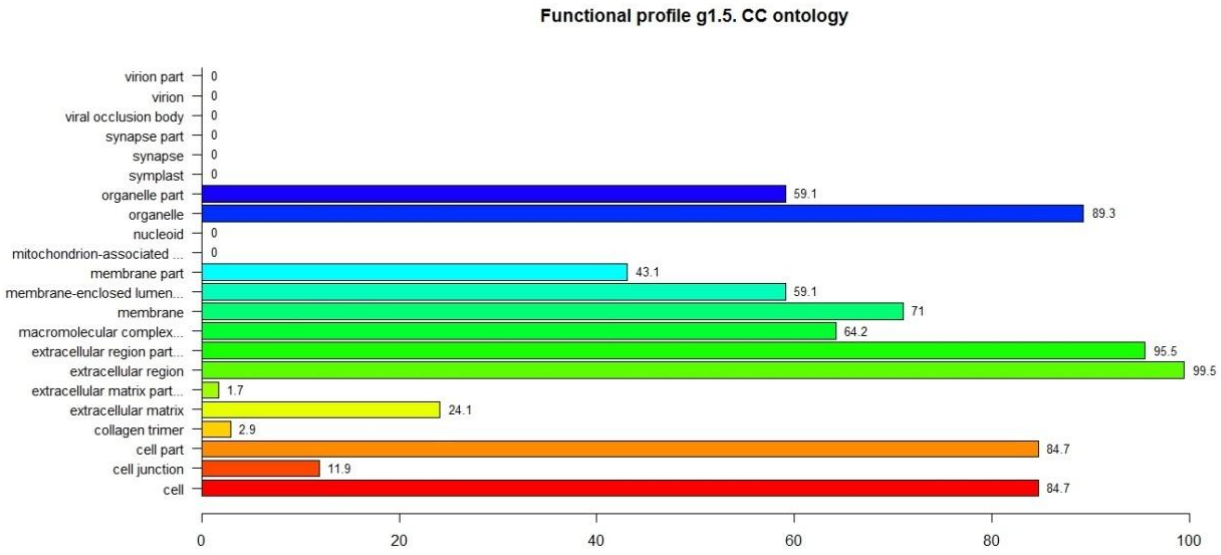
Τα προφίλ της 2ης, της 3ης και της 4ης υποομάδας περιλαμβάνονται στα Σχήματα ΠΑ1,ΠΑ2 και ΠΑ3 αντίστοιχα.



Σχήμα 6.8: Λειτουργικό προφίλ γονιδίων 5ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία BP

Σε οντολογία BP, οι βιολογικές λειτουργίες που αφορούν όλα τα γονίδια της ανάλυσης είναι:

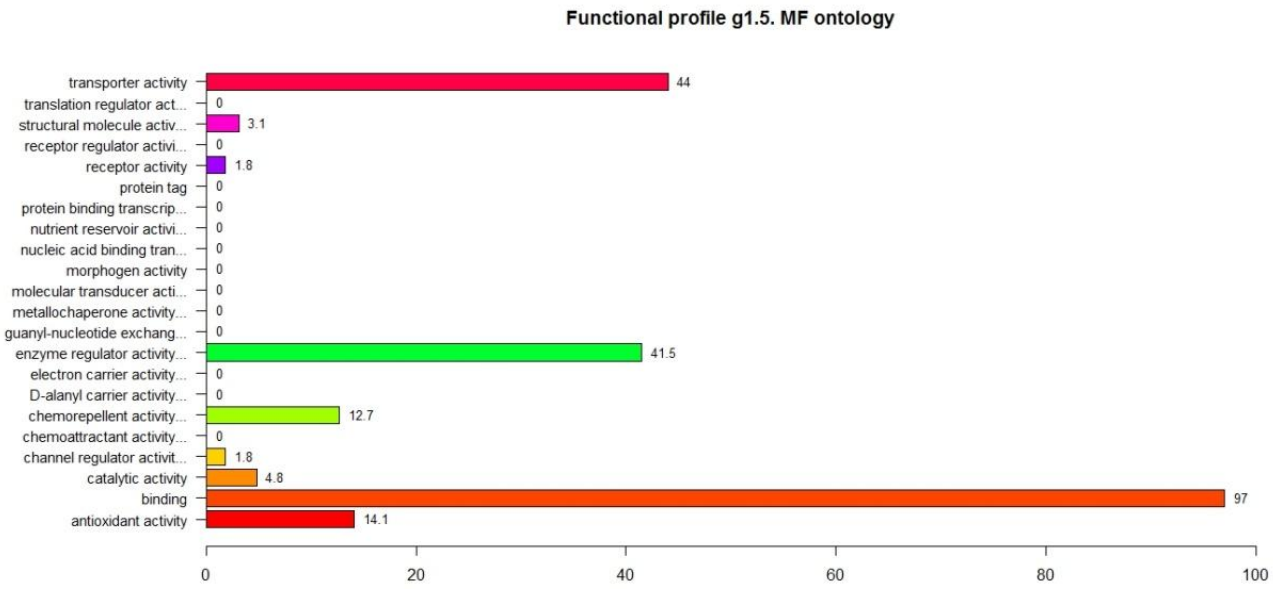
- βιολογική διαδικασία ρύθμισης: 99,5%
- διαδικασία ενός οργανισμού: 92,3%
- κυτταρική διαδικασία: 90,9%
- απόκριση σε ερέθισμα: 90,1%
- διαδικασία πολυκύτταρου οργανισμού: 83 %
- εντοπισμός: 82,4%
- θετική ρύθμιση βιολογικής διεργασίας: 80%
- οργάνωση κυτταρικού συστατικού ή βιογένεση (cellular component organization or biogenesis): 69,6%
- αρνητική ρύθμιση βιολογικής διεργασίας (negative regulation of biological process): 66,6%



Σχήμα 6.9: Λειτουργικό προφίλ γονιδίων 5ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία CC

Σε οντολογία CC, οι κυτταρικές δομές που αφορούν όλα τα γονίδια της ανάλυσης είναι:

- εξωκυττάρια περιοχή: 99,5%
- τμήμα εξωκυττάριας περιοχής: 95,5%
- οργανίδιο: 89,3%
- τμήμα του κυττάρου: 84,7%
- κύτταρο: 84,7%
- μεμβράνη (membrane): 71%
- αυλός περικλειόμενος από μεμβράνη **(**membrane-enclosed lumen): 59,1%



Σχήμα 6.10: Λειτουργικό προφίλ γονιδίων 1^{ης} υποομάδας 1^{ης} ομαδοποίησης Συνόλου Β σε οντολογία MF

Σε οντολογία MF, οι λειτουργίες των γονιδιακών προϊόντων που αφορούν όλα τα γονίδια της ανάλυσης είναι:

- δέσμευση: 97%
- ρυθμιστής δραστηριότητας ενζύμου: 41,5%
- δραστηριότητα μεταφορέα (transporter activity): 44%

Τα λειτουργικά προφίλ της 2^{ης} ομαδοποίησης για τις 6 υποομάδες περιλαμβάνονται στα Σχήματα ΠΑ4-ΠΑ9. Το Παράρτημα Γ περιέχει ευρετήριο περιγραφών των GO ταυτοτήτων των βιολογικών λειτουργιών που αφορούν τα γονίδια της ανάλυσης.

Από την δημιουργία των λειτουργικών προφίλ μπορέσαμε να δούμε ποιους βιολογικούς στόχους, ποιες βιοχημικές δραστηριότητες αλλά και κυτταρικές θέσεις αφορούν τα σύνολα των γονιδίων αλλά και ομαδοποιήσεις αυτών. Δεν μπορέσαμε να βγάλουμε συμπεράσματα για τη σχέση των λειτουργιών αυτών με την τοξικότητα καθώς αυτές ήταν πολύ γενικές.

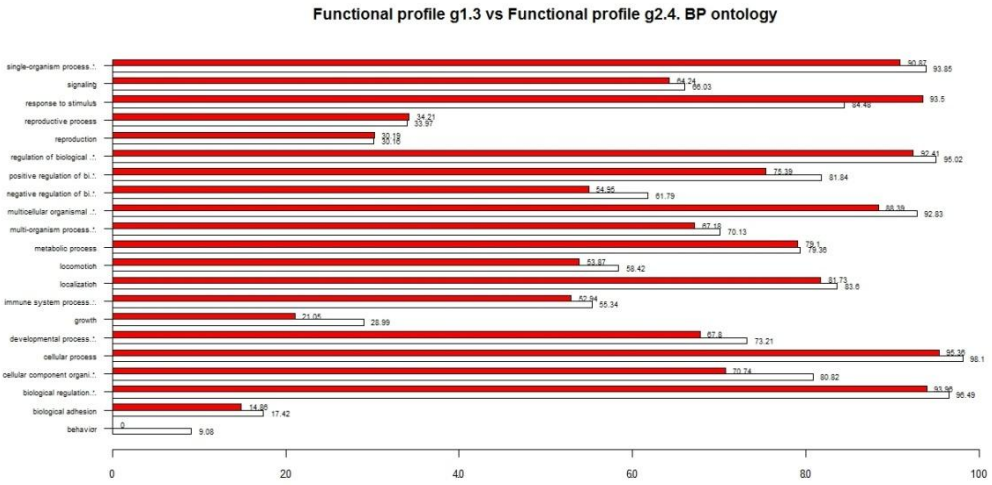
6.1.5 Σύγκριση λειτουργικών προφίλ

Για να συγκρίνουμε τις λειτουργικές διαφορές μεταξύ των δυο διαφορετικών ομαδοποιήσεων θα χρησιμοποιήσουμε μια λειτουργία που παρέχεται από την

βιβλιοθήκη “goProfiles”, με την οποία μπορούμε να δούμε τις διαφορές μεταξύ δύο λειτουργικών προφίλ στο ίδιο γράφημα. Εδώ θα επιχειρήσουμε να δούμε σε ποιο ποσοστό αλλάζουν οι κατηγορίες του λειτουργικού προφίλ ανά οντολογία μεταξύ κάποιων υποομάδων των διαφορετικών ομαδοποιήσεων.

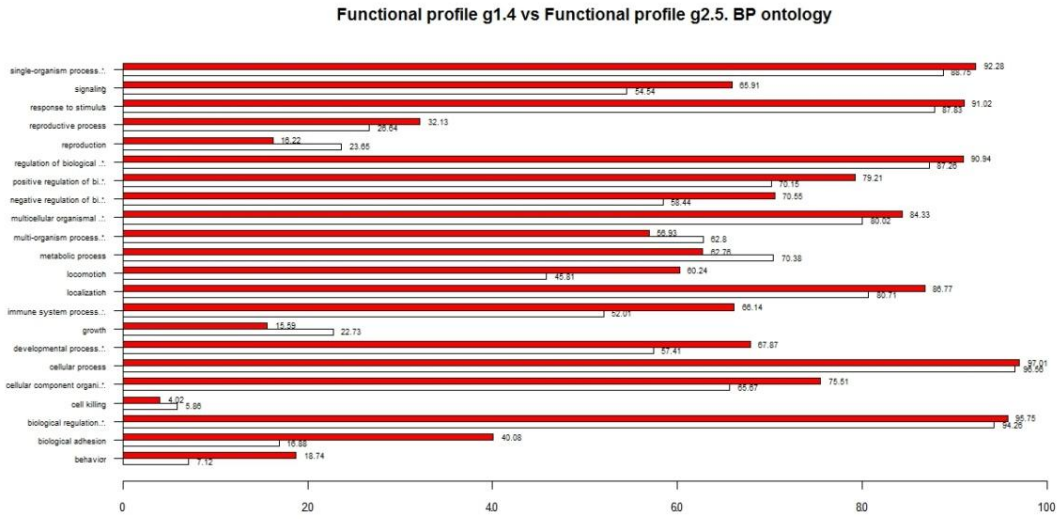
Με βάση τις ομαδοποιήσεις των γονιδίων διαπιστώνεται ότι οι υποομάδες 1, 2 κα 5 της $1^{ης}$ ομαδοποίησης περιλαμβάνουν τα ίδια γονίδια με τις υποομάδες 1, 2 και 6 της $2^{ης}$ ομαδοποίησης. Επομένως η σύγκριση των λειτουργικών προφίλ αυτών των ομάδων δεν θα έχει πρακτικό νόημα για την εξαγωγή συμπερασμάτων. Άρα θα πραγματοποιηθεί σύγκριση λειτουργικών προφίλ σε οντολογία BP μεταξύ :

- ✓ $3^{ης}$ υποομάδας της $1^{η}ς$ ομαδοποίησης με την $4^{η}$ υποομάδα της $2^{ης}$ ομαδοποίησης
- ✓ $4^{ης}$ υποομάδας της $1^{η}ς$ ομαδοποίησης με την $5^{η}$ υποομάδα της $2^{ης}$ ομαδοποίησης



Σχήμα 6.11: Συγκριτικό λειτουργικό προφίλ γονιδίων $3^{ης}$ υποομάδας $1^{ης}$ ομαδοποίησης και $4^{ης}$ υποομάδας $2^{ης}$ ομαδοποίησης Συνόλου Β σε οντολογία BP

Με βάση τα αποτελέσματα της παραπάνω σύγκρισης (Σχήμα 6.11) παρατηρούμε ότι τα γονίδια της $4^{η}$ υποομάδα της $2^{ης}$ ομαδοποίησης συμμετέχουν σε ίδιο ποσοστό στις λειτουργίες αναπαραγωγής (reproduction, reproductive process) και μεταβολικής διεργασίας (metabolic process) με τα γονίδια της 3ης υποομάδας της 1ης ομαδοποίησης. Παρατηρούμε ότι στις υπόλοιπες λειτουργικές κατηγορίες συμμετέχουν σε ποσοστά 5-10 % περισσότερο τα γονίδια της $3^{ης}$ υποομάδας, παρόλο που αυτή διαθέτει επτά λιγότερα γονίδια από την συγκρινόμενη υποομάδα.



Σχήμα 6.12: Συγκριτικό λειτουργικό προφίλ γονιδίων 4^{ης} υποομάδας 1^{ης} ομαδοποίησης και 5^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου Β σε οντολογία BP

Συμπεραίνουμε ότι τα γονίδια της 5^{ης} υποομάδα της 2^{ης} ομαδοποίησης συμμετέχουν σε αρκετά μεγαλύτερο ποσοστό στις εξής λειτουργικές κατηγορίες (Σχήμα 6.12): απόκριση (behavior), βιολογική προσκόλληση, κίνηση (locomotion). Ενώ στις παρακάτω λειτουργίες συμμετέχουν σε μικρότερο ποσοστό σε σχέση με τα γονίδια της 4^{ης} υποομάδας της 1^{η}ς ομαδοποίησης: κυτταρικός θάνατος (cell killing), ανάπτυξη (growth), μεταβολική διεργασία, αναπαραγωγή. Οι μικρές διαφορές στα ποσοστά συμμετοχής τους στις υπόλοιπες κατηγορίες οφείλονται στο ότι η 5^{η} υποομάδα διαθέτει 16 γονίδια περισσότερα από την 4^{η} .

Τα περισσότερα γονίδια της 3^{ης} υποομάδας και της 5^{ης} υποομάδας της 1^{ης} ομαδοποίησης συμμετέχουν σε μεγαλύτερο ποσοστό σε σημαντικότερες λειτουργίες όπως:

- θετική/ αρνητική βιολογική ρύθμιση διεργασίας
- οργάνωση κυτταρικού συστατικού ή βιογένεση
- κυτταρική διεργασία
- διεργασία ανοσοποιητικού συστήματος (immune system process)

Θα μπορούσαμε λοιπόν να συμπεράνουμε ότι η πρώτη ομαδοποίηση των δεδομένων εκτός από το ότι αποτελεί ταξινομημένη λίστα σημαντικότητας των γονιδίων, αποτελεί και ταξινομημένη λίστα σημαντικότητας των βιολογικών λειτουργιών στις οποίες μετέχουν αυτά.

6.2 GOStats: στατιστικά σε GO όρους

Η βιβλιοθήκη "GOStats" [14] παρέχει τη δυνατότητα να απεικονιστούν οι σχέσεις των γονιδίων με την μορφή ενός DAG γραφήματος.

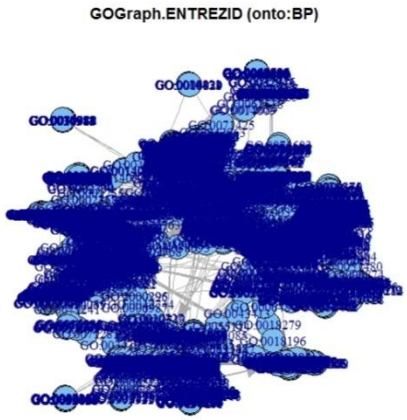
Χρησιμοποιούνται δύο λειτουργίες της βιβλιοθήκης για την παραγωγή γραφημάτων:

- ✓ Με είσοδο τις Entrez ταυτότητες των γονιδίων δημιουργείται ένα DAG γράφημα. Η κατασκευή του γίνεται ανά οντολογία. Μόλις αναγνωριστούν οι όροι που εισάγονται, τότε αναζητούνται «γονείς» των όρων, και «γονείς» των «γονέων» και ούτω καθ' εξής. Επίσης ορίζεται και το μεταφραστικό πακέτο το οποίο θα αντιστοιχίσει τις Entrez ταυτότητες των γονιδίων σε GO ταυτότητες των βιολογικών λειτουργιών, το οποίο εδώ θα είναι το "hgu95av2.db".
- ✓ Με είσοδο τους GO όρους, δημιουργείται ένα DAG γράφημα. Ακόμη ορίζεται το περιβάλλον πχ. GOMFPARENTS (δηλαδή οι «γονείς» των GO όρων σε MF οντολογία), ώστε να εντοπιστούν οι «γονείς» αυτού του όρου και να παρουσιαστεί μόνο η σχέση των GO όρων που εισάγονται με τους γονείς των GO όρων αυτών.

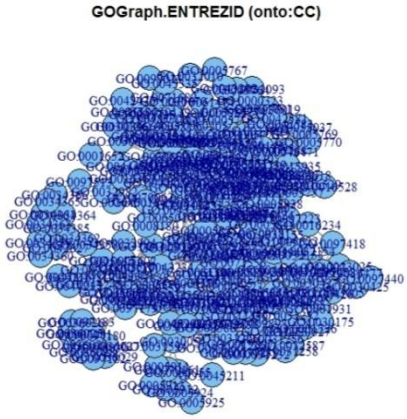
6.2.1 GO γραφήματα γονιδίων Συνόλου Α

Από το Σύνολο Α λαμβάνονται 129 Uniprot ταυτότητες πρωτεϊνών οι οποίες μεταφράζονται σε 118 Entrez ταυτότητες γονιδίων καθώς κάποιες αντιστοιχίσεις δεν διατίθενται από τη βάση δεδομένων, και σε 1089 GO ταυτότητες βιολογικών λειτουργιών, μέσω της μεταφραστικής βιβλιοθήκης "org.Hs.eg.db" και έπειτα χρησιμοποιούνται ως είσοδος για την δημιουργία των επιθυμητών γραφημάτων.

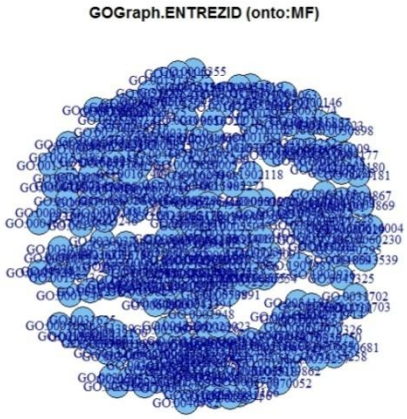
Τα παρακάτω τρία γραφήματα δημιουργούνται μέσω της 1^{ης} λειτουργίας, που απαιτεί ως είσοδο τις 118 Entrez ταυτότητες γονιδίων και παρουσιάζουν τις σχέσεις των GO όρων σε κάθε οντολογία.



Σχήμα 6.13: GO γράφημα γονιδίων Συνόλου A σε οντολογία BP



Σχήμα 6.14: GO γράφημα γονιδίων Συνόλου A σε οντολογία CC



Σχήμα 6.15: GO γράφημα γονιδίων Συνόλου A σε οντολογία MF

Τα γραφήματα (Σχήμα 6.13-6.15) παρατηρείται ότι διαθέτουν μεγάλο όγκο πληροφοριών, καθώς διατυπώνουν σχέσεις 1080 GO ταυτοτήτων για 118 γονίδια, πράγμα το οποίο καθιστά δύσκολη την εξαγωγή βιολογικών συμπερασμάτων. Με χρήση της βιβλιοθήκης "zoom" [32], είναι δυνατό να μεγεθύνουμε το γράφημα σε θέση της επιλογής μας για να το μελετήσουμε. Στο Σχήμα 6.16 παρουσιάζεται μια εικόνα της δομής GO για οντολογία BP με χρήση της λειτουργίας του zoom, όπου μπορούμε καλύτερα να παρατηρήσουμε τις σχέσεις μεταξύ των στοιχείων.

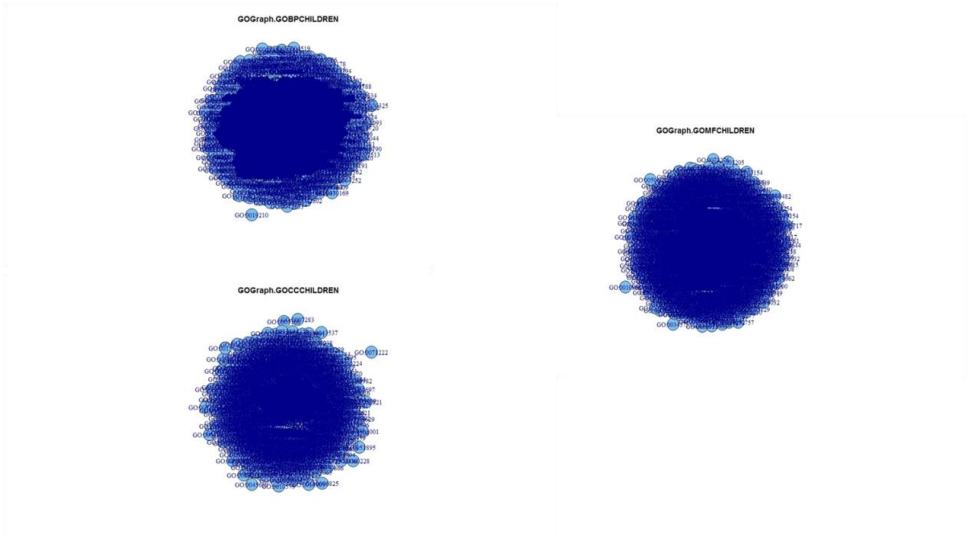
Σχήμα 6.16: GO γράφημα σε οντολογία BP με χρήση της βιβλιοθήκης zoom σε θέση της επιλογής μας

Τα γραφήματα που ακολουθούν δημιουργούνται με χρήση της 2^{ης} λειτουργίας, που απαιτεί ως είσοδο τους 1080 GO όρους και παρουσιάζουν τις σχέσεις των GO όρων με βάση το περιβάλλον που έχει οριστεί. Τα περιβάλλοντα που ορίστηκαν είναι κατά σειρά:

- ✓ Περιβάλλον GOBPANCESTOR, GOCCANCESTOR, GOMFANCESTOR: παρουσιάζει τις σχέσεις των GO όρων με τους «προγόνους» τους ανά οντολογία.
- ✓ Περιβάλλον GOBPCHILDREN, GOCCCHILDREN, GOCCCHILDREN: παρουσιάζει τις σχέσεις των GO όρων με τα «παιδιά» τους ανά οντολογία.
- ✓ Περιβάλλον GOBPPARENTS, GOCCPARENTS, GOMFPARENTS: παρουσιάζει τις σχέσεις των GO όρων με τους «γονείς» τους ανά οντολογία.

- ✓ Περιβάλλον GOBPOFFSPRING, GOCCOFFSPRING, GOMFOFFSPRING: παρουσιάζει τις σχέσεις των GO όρων με τους «απογόνους» τους ανά οντολογία.

Παρουσιάζουμε εδώ ενδεικτικά ένα γράφημα (Σχήμα 6.17) ως μια εικόνα της πολυπλοκότητας και πυκνότητας της πληροφορίας που υπάρχει στο Σύνολο Α και σαν μέτρο σύγκρισης με αυτά που ακολουθούν για το Σύνολο Β.



Σχήμα 6.17: GO γραφήματα γονιδίων Συνόλου Α σε περιβάλλοντα GOBPCHILDREN, GOCCCHILDREN, GOMFCHILDREN

6.2.2 GO γραφήματα γονιδίων Συνόλου Β

Από το Σύνολο Β λαμβάνονται οι συμβολικές ονομασίες των γονιδίων και μεταφράζονται σε Entrez ταυτότητες γονιδίων και GO ταυτότητες βιολογικών λειτουργιών, μέσω της βιβλιοθήκης "org.Hs.eg.db". Για αυτό το σύνολο δεδομένων θα δημιουργηθούν γραφήματα για κάθε υποομάδα των 2 ομαδοποιήσεων των γονιδίων που αναφέρθηκαν στο υποκεφάλαιο 6.1.4.

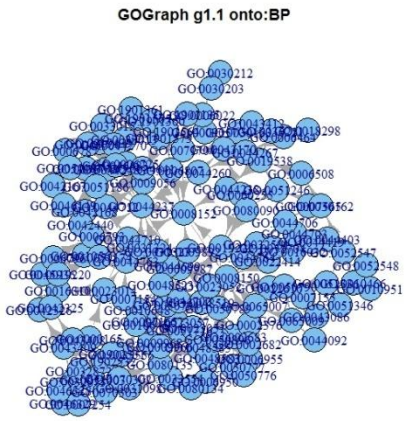
Η πρώτη λειτουργία της βιβλιοθήκης "GOstats" θα χρησιμοποιηθεί για τη δημιουργία γραφημάτων για τις δυο ομαδοποιήσεις, ενώ η δεύτερη λειτουργία μόνο για την πρώτη ομάδα με τα γονίδια με $\boldsymbol{Q^2}$=1.

Οι πληροφορίες που μπορούμε να εξάγουμε από τέτοιου είδους γραφήματα είναι:

- Τις σχέσεις που έχει ο κάθε GO όρος με τους υπόλοιπους όρους, το οποίο θα έχει πρακτικό νόημα να μελετήσουμε για μικρά σύνολα GO όρων.
- Με ποιούς GO όρους συνδέονται οι περισσότεροι κόμβοι του γραφήματος, δηλαδή ποιες βιολογικές λειτουργίες συνδέουν τους όρους μεταξύ τους. Στην παρακάτω ανάλυση θα αναφέρονται χάριν συντομίας ως GO-OUT.
- Για ποιους GO όρους δεν υπάρχει σύνδεση που να καταλήγει σε αυτούς (κόμβοι ρίζα), δηλαδή σε ποια βιολογική λειτουργία δεν καταλήγει με σύνδεση κανένας GO όρος καθώς είναι πιο ειδικοί από τις υπόλοιπες, μιας και αποτελεί μετάφραση ενός γονιδίου και όχι «γονιός» κάποιου όρου. Στην παρακάτω ανάλυση θα αναφέρονται χάριν συντομίας ως GO-IN.

6.2.2.1 GO γραφήματα της 1ης ομαδοποίησης

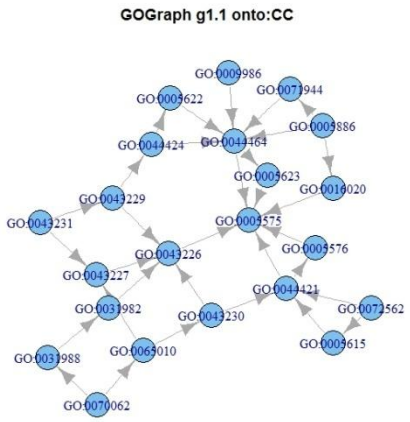
Τα γραφήματα, τα οποία δημιουργούνται με τη χρήση της 1ης λειτουργίας, για τις 5 ομάδες ανά οντολογία BP, CC, MF παρουσιάζονται παρακάτω.



Σχήμα 6.18: GO γράφημα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου B σε οντολογία BP

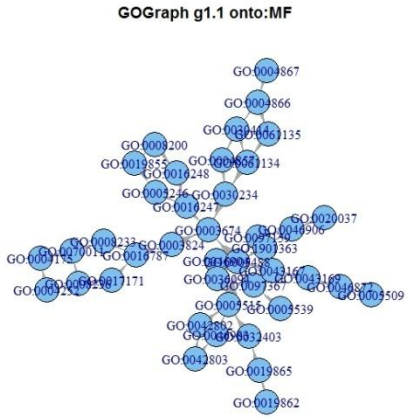
Οι GO-OUT όροι του γραφήματος (Σχήμα 6.18) σε οντολογία BP είναι: GO:0008150 (βιολογική διεργασία), GO:0008152 (μεταβολική διεργασία), GO:0019538 (μεταβολική διεργασία πρωτεϊνών), GO:0050789 (ρύθμιση βιολογικής διεργασίας), GO:0051246 (ρύθμιση πρωτεϊνικής μεταβολικής διεργασίας). Ενώ οι GO-IN όροι είναι: GO:0016032 (ιογενής διαδικασία), GO:0010951 (αρνητική ρύθμιση της

δραστηριότητας των ενδοπεπτιδάσων), GO:0018298 (σύνδεση πρωτεΐνης χρωμοφόρου), GO:0030165 (καταβολική διεργασία πρωτεϊνών), GO:0006898 (ενδοκυττάρωση με διαμεσολάβηση υποδοχέα), GO:0050777 (αρνητική ρύθμιση της απόκρισης του ανοσοποιητικού συστήματος).



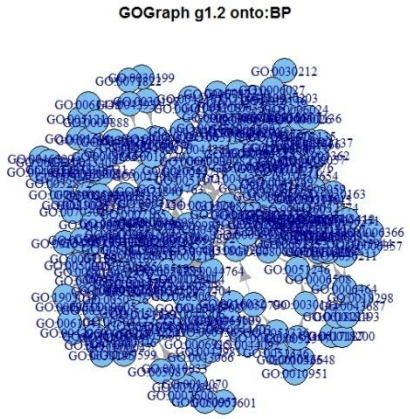
Σχήμα 6.19: GO γράφημα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία CC

Σε οντολογία CC (Σχήμα 6.19) οι GO-OUT κόμβοι είναι: GO:0044464 (τμήμα κυττάρου), GO:0005575 (κυτταρικό συστατικό), GO:0044421 (τμήμα της εξωκυττάριας περιοχής) ενώ οι GO-IN όροι είναι: GO:0070062 (εξωκυτταρικό εξώσωμα) και GO:0009956 (σχεδιασμός ακτινικής μορφής).



Σχήμα 6.20: GO γράφημα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογία MF

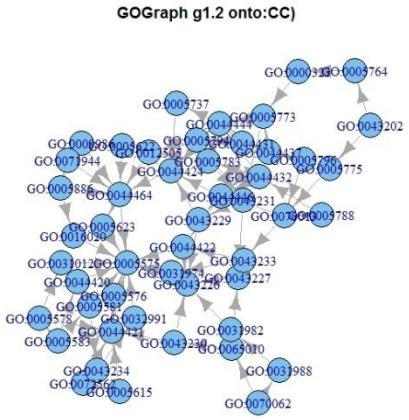
Σε οντολογία MF (Σχήμα 6.20) οι GO-OUT όροι είναι: GO:0003674 (μοριακή διεργασία), GO:0005488 (δέσμευση), ενώ οι GO-IN όροι: GO:0004867 (αναστολέας δραστηριότητας ενδοπεπτιδάσης τύπου σερίνης), GO:0004252 (δραστηριότητας ενδοπεπτιδάσης τύπου σερίνης), GO:0042803 (δραστηριότητα ομοδιμερισμού πρωτεΐνης), GO:0019862 (δέσμευση IgA).



Σχήμα 6.21: GO γράφημα γονιδίων 2^{ης} υποομάδας 1^{ης} ομαδοποίησης Συνόλου B σε οντολογία BP

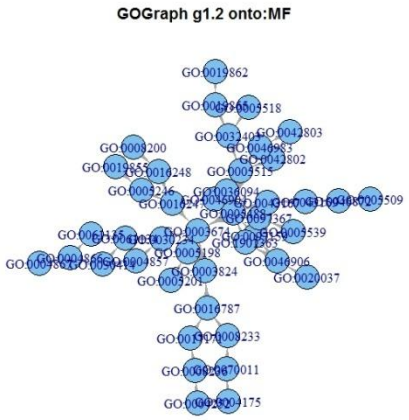
Το γράφημα της 2^{ης} υποομάδας (Σχήμα 6.21) περιγράφει σχέσεις GO όρων που προκύπτουν από 16 γονίδια και έτσι εμφανίζεται πολυπλοκότερο από της 1^{ης} υποομάδας. Εδώ οι GO-OUT όροι είναι: GO:0008150 (βιολογική διεργασία), GO:0044763 (κυτταρική διεργασία ενός οργανισμού), GO:0009987 (κυτταρική διεργασία), GO:0002376 (διεργασία του ανοσοποιητικού συστήματος), GO:0050896 (απόκριση σε ερέθισμα), GO:0048583 (ρύθμιση της απόκρισης σε ερέθισμα), GO:0050794 (ρύθμιση κυτταρικής διεργασίας), GO:0008219 (κυτταρικός θάνατος), GO:0009056 (καταβολική διεργασία), GO:0060255 (ρύθμιση μεταβολικής διεργασίας μακρομορίου), GO:0071704 (μεταβολική διεργασία οργανικού συστατικού). Οι GO-IN όροι είναι: GO:0007155 (απόπτωση), GO:0042167 (καταβολική διεργασία της αίμης), GO:0050777 (αρνητική ρύθμιση του ανοσοποιητικού συστήματος), GO:0030212 (μεταβολική διεργασία υαλουρονάνης), GO:0007601 (οπτική αντίληψη), GO:0014070 (απόκριση σε μια οργανική κυκλική ένωση), GO:0045944 (θετική ρύθμιση της μεταγραφής από την RNA πολυμεράση του υποκινητή II), GO:0051216 (ανάπτυξη χόνδρου), GO:0070848 (παράγοντας

απόκρισης στην ανάπτυξη) καθώς και οι αντίστοιχοι GO-IN όροι που αναφέρθηκαν στην $1^{η}$ υποομάδα.



Σχήμα 6.22: GO γράφημα γονιδίων $2^{ης}$ υποομάδας $1^{ης}$ ομαδοποίησης Συνόλου Β σε οντολογία CC

Σε οντολογία CC (Σχήμα 6.22) οι GO-OUT όροι είναι: GO:0005575 (κυτταρικό συστατικό), GO:0043226 (οργανίδιο), GO:0044421 (τμήμα εξωκυττάριας περιοχής), GO:0004464 (δραστηριότητα της συνθάσης λευκοκιτριένης-C4), GO:0044444 (κυτταροπλασματικό τμήμα), GO:0044446 (τμήμα ενδοκυτταρικού οργανιδίου), αντίστοιχα οι GO-IN όροι: GO:0009986 (επιφάνεια κυττάρου), GO:0005886 (μεμβράνη του πλάσματος), GO:0070062 (εξωκυτταρικό εξώσωμα), GO:0072562 (μικροσωματίδιο αίματος), GO:0005797 (συσκευή Golgi), GO:0043202 (λυσοσωμικός αυλός), GO:0005788 (ενδοπλασματικό δίκτυο).

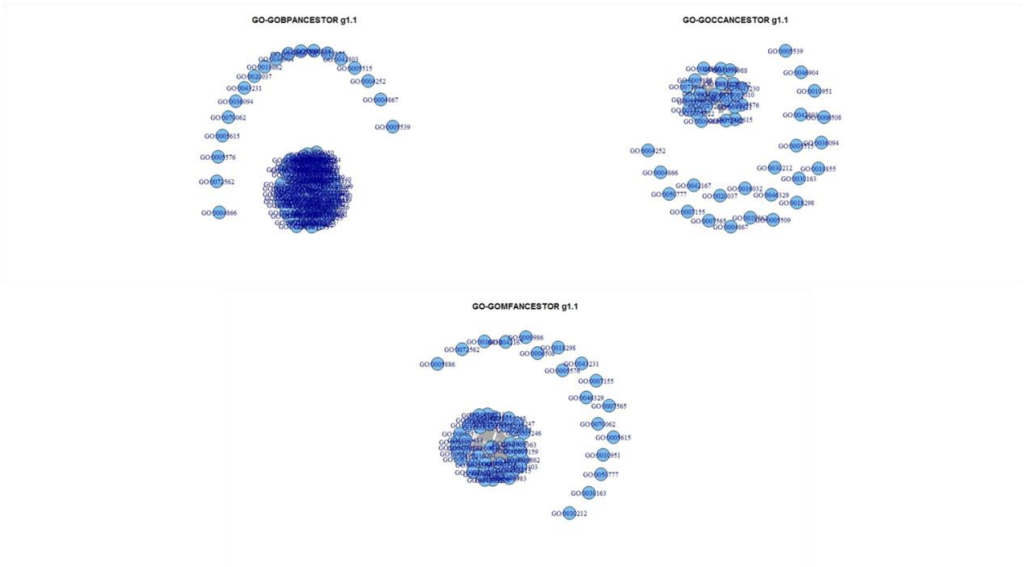


Σχήμα 6.23: GO γράφημα γονιδίων $2^{ης}$ υποομάδας $1^{ης}$ ομαδοποίησης Συνόλου B σε οντολογία MF

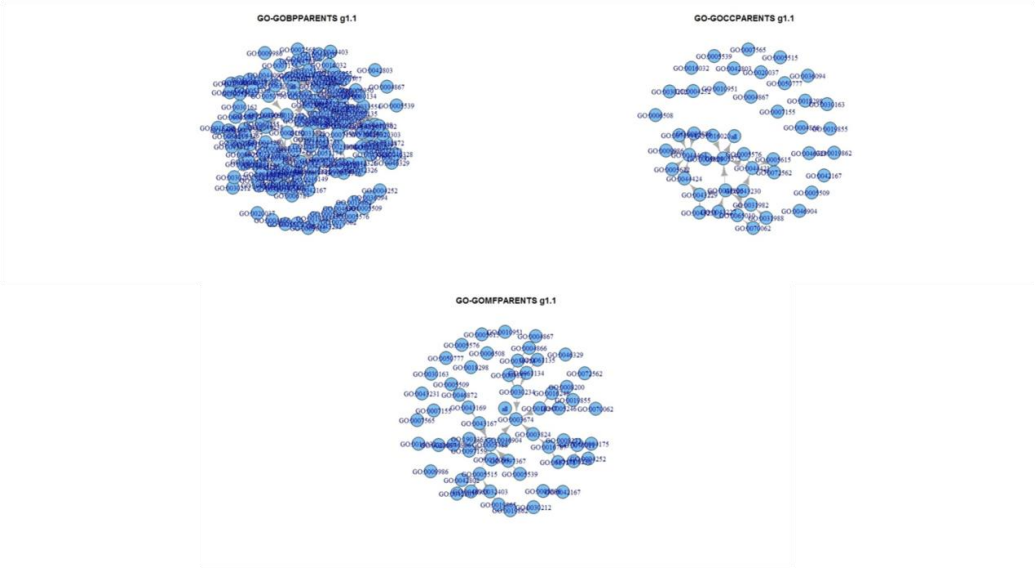
Σε οντολογία MF (Σχήμα 6.23) οι GO-OUT όροι είναι οι ίδιοι με το αντίστοιχο γράφημα της $1^{ης}$ υποομάδας που αφορούν τις λειτουργίες της δέσμευσης και της μοριακής λειτουργίας ενώ οι GO-IN όροι είναι: GO:0036094 (δέσμευση μικρού μορίου) ,GO:0004867 (αναστολέας δραστηριότητας ενδοπεπτιδάσης τύπου σερίνης), GO:0019862 (δέσμευση IgA), GO:0020037 (δέσμευση αίμης), GO:0042803 (δραστηριότητα ομοδιμερισμού πρωτεΐνης), GO:0019855 (αναστολέας δραστηριότητας καναλιού ασβεστίου).

Τα γραφήματα για την $3^{η}$, $4^{η}$ και $5^{η}$ υποομάδα διαθέτουν μεγάλο όγκο πληροφοριών καθώς, όπως έχουμε αναφέρει, περιγράφουν σχέσεις πολύ περισσότερων σε αριθμό GO όρων και δεν θα παρουσιαστούν.

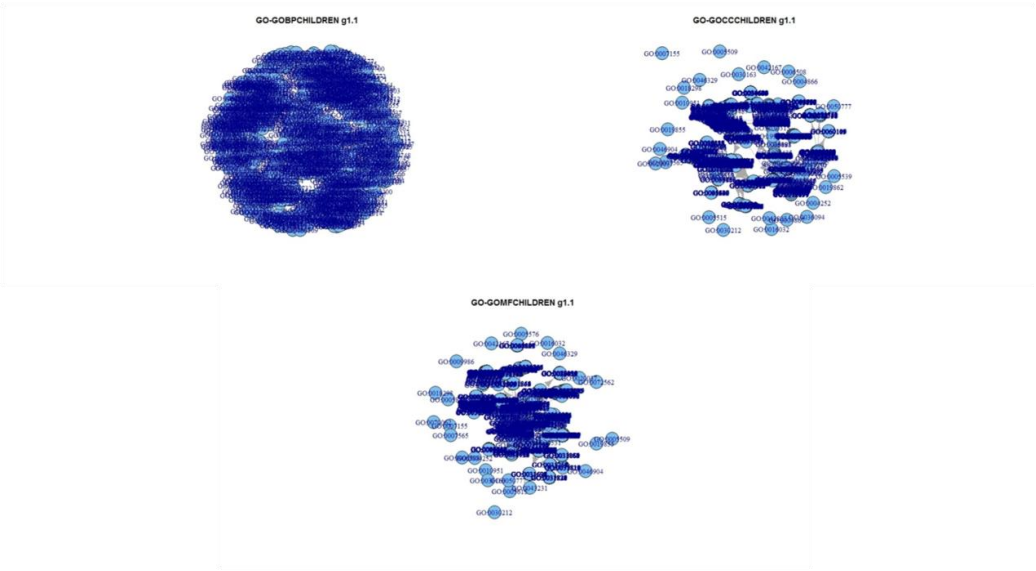
Τα γραφήματα που ακολουθούν, με τη χρήση της $2^{ης}$ λειτουργίας, γίνονται στα περιβάλλοντα που αναφέρθηκαν παραπάνω, μόνο για την πρώτη υποομάδα με $\boldsymbol{Q^2}$=1, δηλαδή εξετάζονται οι σχέσεις των πιο σημαντικών όρων με τους «προγόνους» τους, τα «παιδιά» τους, τους «γονείς» τους και τους «απογόνους» τους ανά οντολογία.



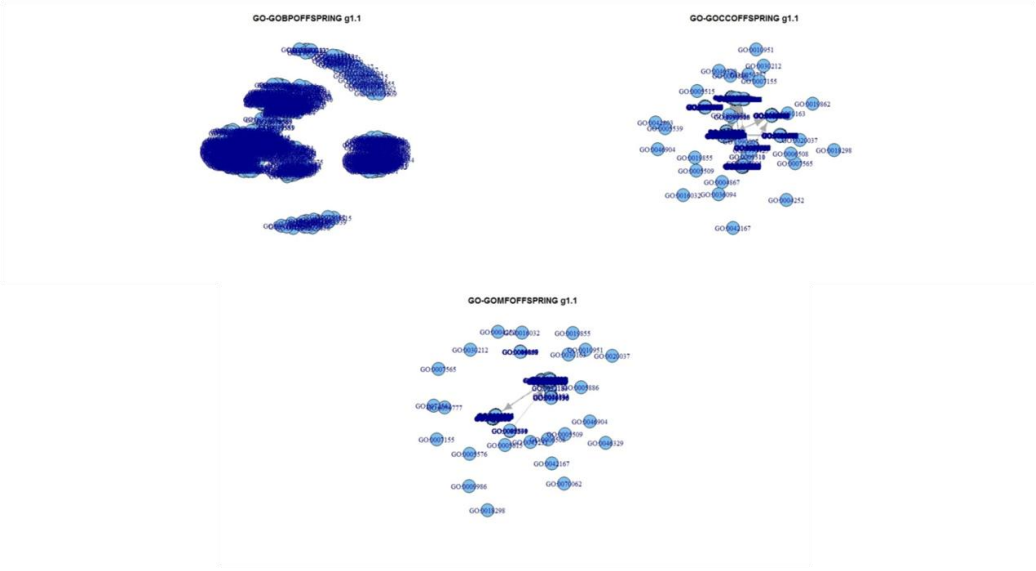
Σχήμα 6.24: GO γραφήματα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε περιβάλλοντα GOBPANCESTOR, GOCCANCESTOR, GOMFANCESTOR



Σχήμα 6.25: GO γραφήματα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε περιβάλλοντα GOBPPARENTS, GOCCPARENTS, GOMFPARENTS



Σχήμα 6.26: GO γραφήματα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε περιβάλλοντα GOBPCHILDREN, GOCCCHILDREN, GOMFCHILDREN



Σχήμα 6.27: GO γραφήματα γονιδίων 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε περιβάλλοντα GOBPOFFSPRING, GOCCOFFSPRING, GOMFOFFSPRING

Στα παραπάνω γραφήματα (Σχήμα 6.24-6.27) εντοπίζονται οι σχέσεις των GO ταυτοτήτων των σημαντικότερων γονιδίων με τους «γονείς» τους, τα «παιδιά» τους, τους «απογόνους» τους και τους «προγόνους» τους. Όπως παρατηρείται στα διαγράμματα σε περιβάλλοντα GOBPPARENTS, GOBPCHILDREN, GOCCCHILDREN, GOMFCHILDREN, οι GO ταυτότητες των γονιδίων έχουν περισσότερες σχέσεις μεταξύ τους σε σύγκριση με τα υπόλοιπα γραφήματα, όπου εκεί διακρίνονται σε όλα GO ταυτότητες, οι οποίες δεν εμφανίζουν καμία σύνδεση με τους υπόλοιπους GO όρους. Συμπεραίνεται ότι οι GO όροι εμφανίζουν άμεσες

σχέσεις μεταξύ τους μέσω των «γονιών» τους και των «παιδιών» τους και όχι τόσο μέσω μακρινότερων «συγγενών» τους.

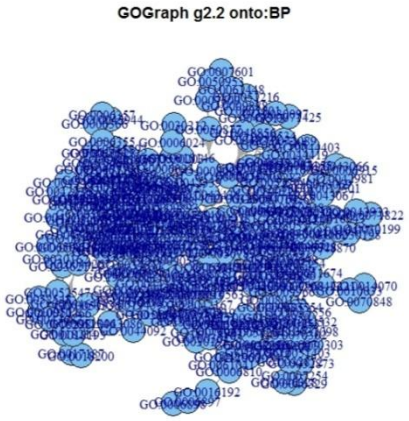
Στον παρακάτω Πίνακα 6.2 παρουσιάζονται οι GO όροι για τους οποίους οι συνδέσεις καταλήγουν σε αυτούς (GO-OUT) καθώς και οι GO όροι που δεν καταλήγει σε αυτούς καμία σύνδεση (GO-IN) στη δομή του GO γραφήματος. Στα πολύ πυκνά γραφήματα όπου συμπεριλαμβάνεται μεγάλος όγκος πληροφορίας, εξαιτίας των γενικευμένων όρων που συμπεριλαμβάνουν, εμφανίζεται η ένδειξη "-" .

Πίνακας 6.2 GO-IN και GO-OUT όροι για τα GO γραφήματα 1ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε περιβάλλοντα BP /CC /MF-ANCESTOR/ PARENTS/ CHILDREN/ OFFSPRING

Οντολογία	Περιβάλλον	GO-IN	GO-OUT
BP	ANCESTOR		
	PARENTS	GO:0050777;GO:0030212; GO:0042167;GO:0042167; GO:0042326;GO:0046329; GO:0010951	GO:0050896;GO:0050789; GO:0008150;GO:0019538; GO:0043170;GO:0044237
	CHILDREN	-	-
	OFFSPRING	-	-
CC	ANCESTOR	GO:0072562;GO:0070062; GO:0043231;GO:0005886	GO:0005576;GO:0005575; GO:0005623;GO:0044464; GO:0044421
	PARENTS	GO:0072562;GO:0070062; GO:0043231;GO:0005886	GO:0043226;GO:0005575; GO:0044464
	CHILDREN	-	-
	OFFSPRING	-	-
MF	ANCESTOR	GO:0004867;GO:0019855; GO:0020037;GO:0019862; GO:0005509;GO:0004252	GO:0005488;GO:0003674; GO:0098772;GO:0030234
	PARENTS	GO:0004252;GO:0004867; GO:0019855;GO:0005509; GO:0020037;GO:0019862	GO:0005488;GO:0003674
	CHILDREN	-	-
	OFFSPRING	-	-

6.2.2.2 GO γραφήματα της 2ης ομαδοποίησης

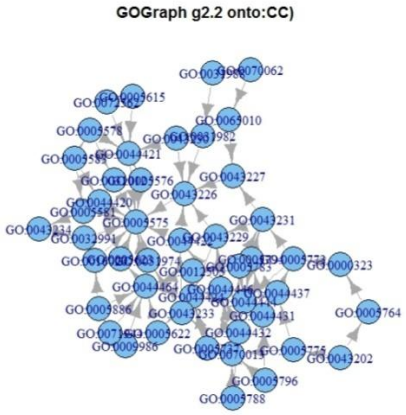
Τα γραφήματα που δημιουργούνται με τη χρήση της 1ης λειτουργίας, για τις 4 υποομάδες με βάση το Q^2, ανά οντολογία παρουσιάζονται παρακάτω. Τα GO γραφήματα για την 1η υποομάδα της 2ης ομαδοποίησης για κάθε οντολογία είναι τα ίδια με της 1ης υποομάδας της 1ης ομαδοποίησης, καθώς περιλαμβάνουν τα ίδια γονίδια, και έτσι δεν θα συμπεριληφθούν. Το ίδιο θα ισχύσει και για τα γραφήματα της 6ης υποομάδας.



Σχήμα 6.28: GO γράφημα γονιδίων 2ης υποομάδας 2ης ομαδοποίησης Συνόλου Β σε οντολογία BP

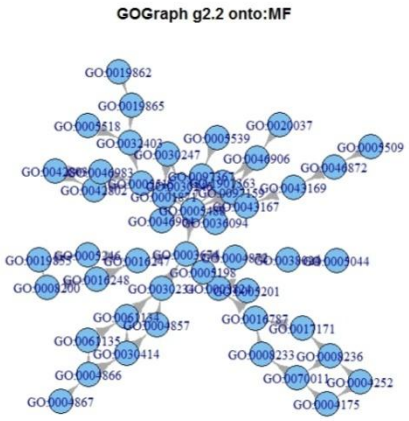
Το γράφημα του Σχήματος 6.28 παρουσιάζει σχέσεις μεταξύ περισσότερων GO όρων σε σχέση με το αντίστοιχο γράφημα του Σχήματος 6.23 της 2ης υποομάδας της 1ης ομαδοποίησης. Αυτό οφείλεται στο ότι τα γονίδια που περιλαμβάνονται σε αυτήν την ομάδα μεταφράζονται σε πιο ειδικούς GO όρους σε σύγκριση με τα γονίδια της αντίστοιχης ομάδας της 1ης ομαδοποίησης. Οι GO-IN όροι του γραφήματος είναι: GO:0016032 (ιογενής διαδικασία), GO:0010951 (αρνητική ρύθμιση της δραστηριότητας της ενδοπεπτιδάσης), GO:0018298 (σύνδεση πρωτεΐνης-χρωμοφόρου), GO:0030163 (καταβολική διεργασία πρωτεΐνης), GO:0006898 (ενδοκυττάρωση με διαμεσολάβηση υποδοχέα), GO:0007155 (προσκόλληση κυττάρου), GO:0042167 (καταβολική διεργασία της αίμης), GO:0050777 (αρνητική ρύθμιση του ανοσοποιητικού συστήματος), GO:0009405 (παθογένεση), GO:0018146 (βιοσυνθετική διεργασία θειικής κερατάνης). Ενώ οι GO-OUT όροι είναι: GO:0048583 (ρύθμιση απόκρισης σε ερέθισμα), GO:0002376 (διεργασία του

ανοσοποιητικού συστήματος), GO:0008150 (βιολογική διεργασία), GO:0009987 (κυτταρική διεργασία), GO:0031323 (ρύθμιση κυτταρικής μεταβολικής διεργασίας), GO:0019222 (ρύθμιση μεταβολικής διεργασίας), GO:0010556 (ρύθμιση βιοσυνθετικής διεργασίας μακρομορίου), GO:0044248 (κυτταρική καταβολική διεργασία), GO:0009058 (βιοσυνθετική διεργασία), GO:0044249 (κυτταρική βιοσυνθετική διεργασία), GO:0019538 (μεταβολική διεργασία πρωτεϊνών), GO:0043170 (μεταβολική διεργασία μακρομορίου), GO:0071704 (μεταβολική διεργασία οργανικού συστατικού).



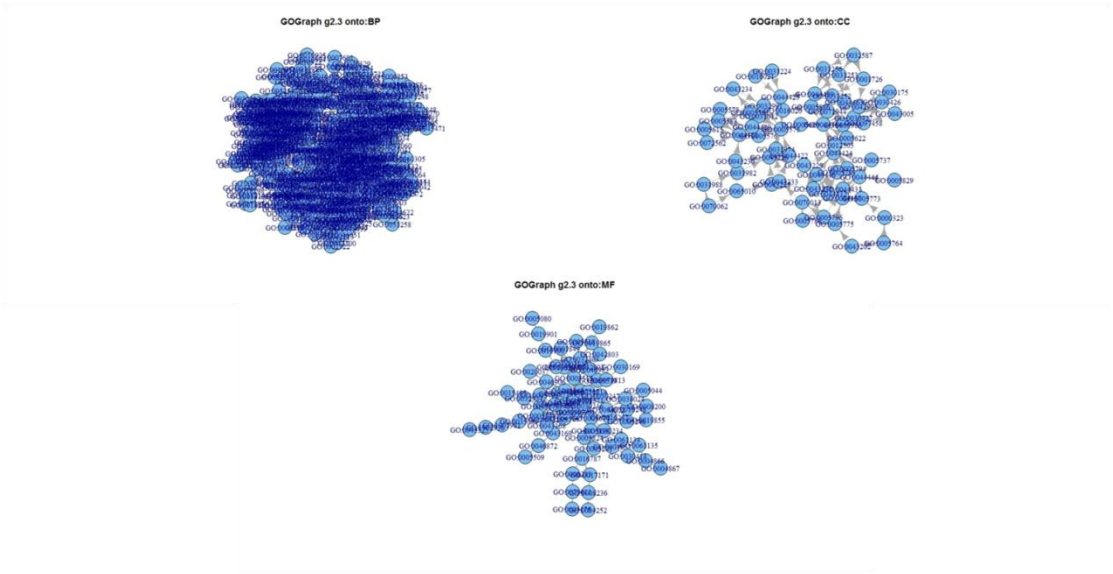
Σχήμα 6.29: GO γράφημα γονιδίων $2^{ης}$ υποομάδας $2^{ης}$ ομαδοποίησης Συνόλου Β σε οντολογία CC

Στο γράφημα του Σχήματος 6.29 οι GO-IN όροι είναι : GO:0009986 (επιφάνεια κυττάρου), GO:0005886 (πλασματική μεμβράνη), GO:0070062 (εξωκυτταρικό εξώσωμα), GO:0072562 (μικροσωματίδιο αίματος), GO:0005796 (συσκευή Golgi), GO:0043202 (αυλός ινώδους κολλαγόνου), ενώ οι GO-OUT όροι είναι: GO:0043226 (οργανίδιο), GO:0044421 (τμήμα εξωκυττάριας περιοχής), GO:0005575 (κυτταρικό συστατικό), GO:0044464 (τμήμα κυττάρου), GO:0044444 (κυτταροπλασματικό μέρος).

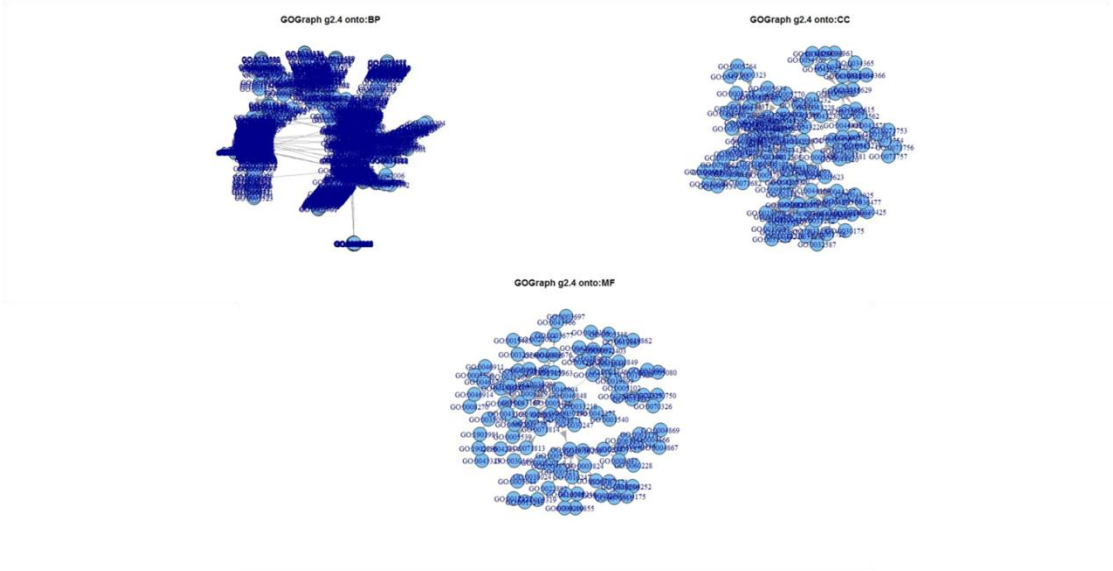


Σχήμα 6.30: GO γράφημα γονιδίων 2^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου B σε οντολογία MF

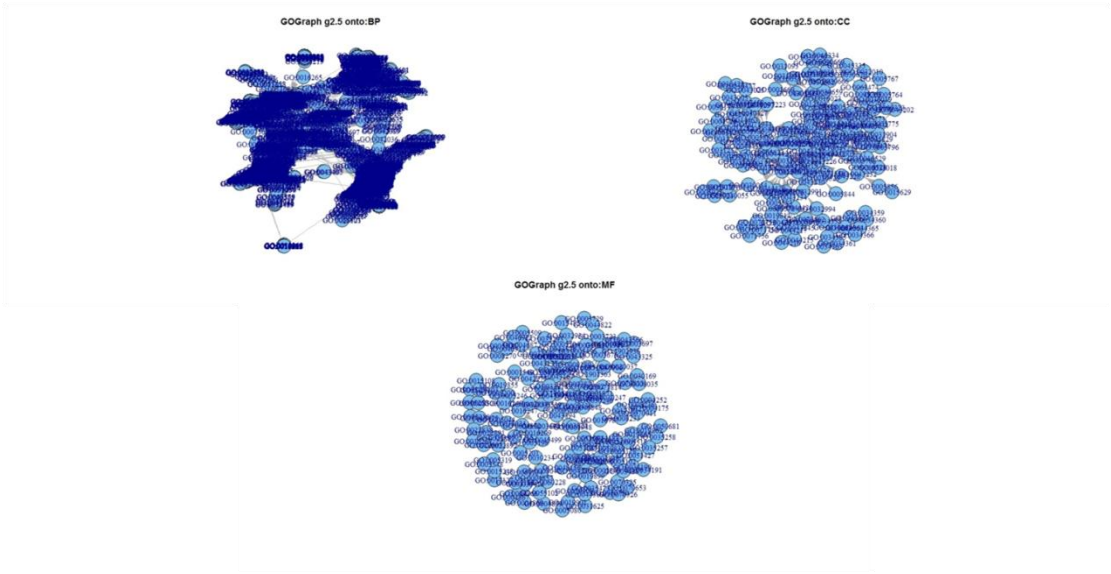
Σε οντολογία MF για το γράφημα του Σχήματος 6.30 οι GO-IN όροι είναι: GO:0036094 (σύνδεση μικρού μορίου), GO:0004867 (αναστολέας δραστηριότητας ενδοπεπτιδάσης τύπου σερίνης), GO:0019862 (δέσμευση IgA), GO:0020037 (δέσμευση αίμης), GO:0042803 (δραστηριότητα ομοδιμερισμού πρωτεΐνης), GO:0019855 (αναστολέας δραστηριότητας καναλιού ασβεστίου), GO:0046904 (δέσμευση οξαλικού ασβεστίου), GO:0004252 (δραστηριότητα ενδοπεπτιδάσης τύπου σερίνης), GO:0005539 (δέσμευση γλυκοσαμινογλυκάνης), GO:0005509 (δέσμευση ιόντος ασβεστίου), GO:0005518 (δέσμευση κολλαγόνου), GO:0005044 (δραστηριότητα scavenger υποδοχέα) ενώ οι GO-OUT όροι είναι: GO:0003674 (μοριακή λειτουργία), GO:0005488 (δέσμευση), GO:0005515 (δέσμευση πρωτεΐνης).



Σχήμα 6.31: GO γραφήματα γονιδίων 3^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου Β σε οντολογία BP, CC και MF



Σχήμα 6.32: GO γραφήματα γονιδίων 4^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου Β σε οντολογία BP, CC και MF



Σχήμα 6.33: GO γραφήματα γονιδίων 5^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου B σε οντολογία BP, CC και MF

Αντίστοιχα παρουσιάζονται στον Πίνακα 6.3 οι GO-OUT και GO-IN όροι για την 3^{η}, 4^{η} και 5^{η} υποομάδα της 2^{ης} ομαδοποίησης. Σε οντολογία BP λόγω πυκνότητας πληροφορίας δεν θα παρουσιαστούν οι αντίστοιχοι όροι.

Πίνακας 6.3: GO-IN και GO-OUT όροι για τα GO γραφήματα της 3^{ης},4^{ης} και 5^{ης} υποομάδας της 2^{ης} ομαδοποίησης του Συνόλου B

Οντολογία	Υποομάδα	GO-IN	GO-OUT
BP	3	-	-
	4	-	-
	5	-	-
CC	3	GO:0009986;GO:0070062;GO:0072562 GO:0005796;GO:0043202;GO:0005583 GO:0005788;GO:0016021;GO:0030175 GO:0030426;GO:0005829;GO:0032587	GO:0043226;GO:0005575; GO:0044421;GO:0044464; GO:0044444
	4	GO:0070062;GO:0072562;GO:0005796 GO:0043202;GO:0005583;GO:0005788 GO:0030175;GO:0030426;GO:0005829 GO:0032587;GO:0005769;GO:0005789 GO:0010008;GO:0030669;GO:0031904 GO:0034360;GO:0071682;GO:0015629 GO:0034359	GO:0044425;GO:0044459; GO:0016020;GO:0044464; GO:0005575;GO:0043226
	5	GO:0070062;GO:0072562;GO:0005796 GO:0043202;GO:0005583;GO:0030175 GO:0030426;GO:0005829;GO:0003287 GO:0005769;GO:0010008;GO:0030669 GO:0031904;GO:0034360;GO:0071682	GO:0005575;GO:0044421; GO:0044425;GO:0044464; GO:0044444;GO:0044422; GO:0044446

		GO:0015629;GO:0034359;GO:0034361 GO:0034362;GO:0043025;GO:0005635 GO:0005770;GO:0005874;GO:0031232 GO:0034365;GO:0030425;GO:0031093 GO:0005887;GO:0034366;GO:0071756 GO:0071757;GO:0036019;GO:0061474 GO:0001669;GO:0005790;GO:0033018 GO:0005844;GO:0042824;GO:0071556	
MF	3	GO:0004867;GO:0019862;GO:0020037 GO:0042803;GO:0019855;GO:0046904 GO:0004252;GO:0005539;GO:0005509 GO:0005518;GO:0005201;GO:0005044 GO:0030247;GO:0015485;GO:0033265 GO:0001849	GO:0003674;GO:0005488; GO:0005515
	4	GO:0004867;GO:0019862;GO:0020037 GO:0042803;GO:0019855;GO:0046904 GO:0004252;GO:0005509;GO:0005518 GO:0005201;GO:0005044;GO:0030247 GO:0015485;GO:0033265;GO:0001849 GO:0030169;GO:0050750;GO:0005080 GO:0043325;GO:0008201;GO:0017127 GO:0035473;GO:0046848;GO:0001540 GO:0016209;GO:0046911;GO:0070326 GO:0060228;GO:0048156;GO:0003823	GO:0003674;GO:0005488; GO:0005515
	5	GO:0004867;GO:0019862;GO:0020037 GO:0042803;GO:0019855;GO:0046904 GO:0004252;GO:0005509;GO:0005518 GO:0005201;GO:0005044;GO:0030247 GO:0015485;GO:0033265;GO:0001849 GO:0030169;GO:0050750;GO:0005080 GO:0043325;GO:0008201;GO:0017127 GO:0035473;GO:0046848;GO:0001540 GO:0016209;GO:0046911;GO:0070326 GO:0060228;GO:0048156;GO:0003823 GO:0004869;GO:0008270;GO:0003697 GO:0031210;GO:0034987;GO:0042834 GO:0004859;GO:0005504;GO:0008430 GO:0043395;GO:0005254;GO:0005506 GO:0042562;GO:0044183;GO:0051082 GO:0003729;GO:0050681;GO:0005178 GO:0031625;GO:0042056;GO:0005548	GO:0005102;GO:0005515; GO:0005488;GO:0003674; GO:0005102;GO:0043168

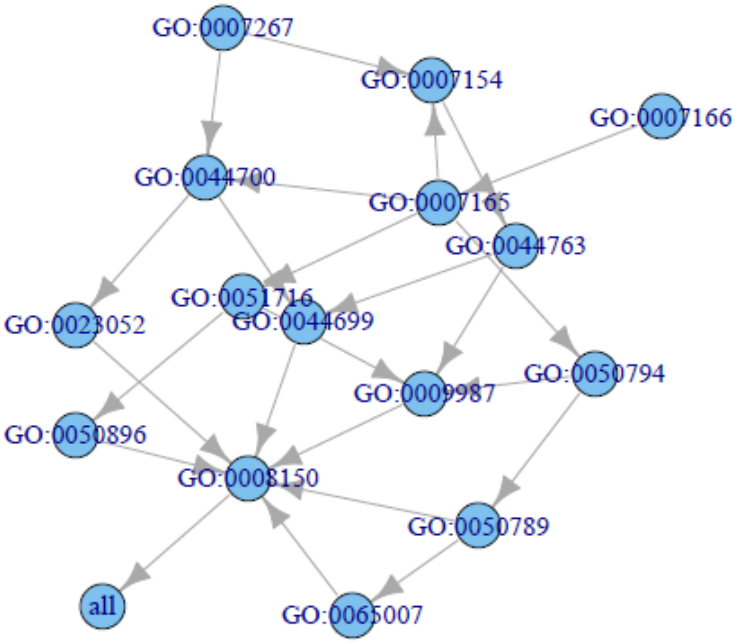
Από τα γραφήματα λάβαμε πιο ειδικούς όρους σε σύγκριση με τα λειτουργικά προφίλ. Ενώ οι GO-OUT όροι περιγράφουν λειτουργίες που περιλαμβάνονται και στα λειτουργικά προφίλ, μέσω των GO-IN όρων λάβαμε πιο ειδικούς όρους όπως η ιογενής διαδικασία, η δραστηριότητα των ενδοπεπτιδασών, η ενδοκυττάρωση με διαμεσολάβηση υποδοχέα, η ρύθμιση του ανοσοποιητικού συστήματος καθώς και οι αποπτωτικές διεργασίες που σχετίζονται με βάση τη βιβλιογραφία με την τοξικότητα.

6.3 *GOSim: ομοιότητες γονιδίων σε GO όρους*

Οι πληροφορίες που συλλέξαμε στο προηγούμενο κεφάλαιο παρότι πολύ περιεκτικές, δεν μας δίνουν κανένα μέτρο σημαντικότητας ή αξιολόγησης τους. Για το λόγο αυτό χρησιμοποιούμε τη βιβλιοθήκη "GOSim" [15] η οποία επιτρέπει τον υπολογισμό των ομοιοτήτων των γονιδίων, βασιζόμενη σε πληροφορίες θεωρητικής ομοιότητας των GO όρων. Ακόμη παρέχει λειτουργικά μέτρα ομοιότητας για τα γονίδια, τα οποία χρησιμοποιούνται για να ομαδοποιηθούν τα γονίδια με βάση την βιολογική τους λειτουργία. Αντίστοιχα, αυτά τα εργαλεία μπορούν να χρησιμοποιηθούν για να υπολογιστεί η ομοιογένεια μιας ομάδας γονιδίων αφού μεταφραστούν σε GO όρους. Τέλος παρέχει στον αναλυτή τη δυνατότητα να αναζητήσει και να βρει τους σημαντικούς υπερ-εκφρασμένους GO όρους σε ένα δείγμα γονιδίων, το οποίο θα αναλυθεί στο Κεφάλαιο 7.

6.3.1 *Ομοιότητες GO όρων*

Τα πιο γνωστά θεωρητικά μέτρα ομοιότητας διατυπώθηκαν από τον Resnik [33]. Αυτά βασίζονται στην έννοια «ελάχιστη εισφορά» (minimum subsumer), δύο GO όρων $\boldsymbol{t}$ και $\boldsymbol{t'}$, η οποία αντιπροσωπεύεται από τον χαμηλότερο κοινό πρόγονο της GO ιεραρχίας. Στο Σχήμα 6.34 ο χαμηλότερος κοινός πρόγονος των ταυτοτήτων GO:0007166 και GO:0007267 είναι ο όρος GO:0023052.



Σχήμα 6.34: GO γράφημα των ταυτοτήτων GO:0007166 και GO:0007267

Το πληροφορικό περιεχόμενο $\boldsymbol{IC_{ms}}$, το οποίο είναι μέτρο ομοιότητας μεταξύ των $\boldsymbol{t}$ και $\boldsymbol{t'}$, δίνεται από:

$$\mathrm{sim}(t, t') = \mathrm{IC_{ms}}(t, t') := \max_{t \in \mathrm{Pa}(t,t')} \mathrm{IC}(t) \tag{6.5}$$

το $\boldsymbol{Pa(t, t')}$, υποδηλώνει όλους τους κοινούς προγόνους των GO όρων $\boldsymbol{t}$ και $\boldsymbol{t'}$, ενώ το $\boldsymbol{IC(t)}$, υποδηλώνει το πληροφοριακό περιεχόμενο του όρου $\boldsymbol{t}$. Ορίζεται ως:

$$\mathrm{IC}(\hat{t}) = -\log\ \mathrm{P}(\hat{t}) \tag{6.6}$$

δηλαδή είναι ο αρνητικός λογάριθμος της πιθανότητας του παρατηρούμενου $\boldsymbol{t}$, όπως φαίνεται στην Εξίσωση 6.6. Το πληροφοριακό περιεχόμενο του κάθε GO όρου είναι ήδη προϋπολογισμένο για κάθε οντολογία. Ο υπολογισμός βασίζεται στην παρατήρηση της μέτρησης του πλήθους των φορών που εμφανίζεται ένας συγκεκριμένος GO όρος ή ένας άμεσος ή έμμεσος απόγονος του στα αντίστοιχα γονιδιακά προϊόντα. Εκτός από το μέτρο ομοιότητας του Resnik, υπάρχουν επεκτάσεις του Lin και των Jiang και Conrath, τα οποία περιλαμβάνονται στην βιβλιοθήκη αυτή. Και οι δυο παραλλαγές διαφέρουν στον τρόπο που κανονικοποιούν τα δεδομένα. Το μέτρο ομοιότητας των Jiang και Conrath ορίζεται ως:

$$\mathrm{sim}(t, t') = 1 - \min\left(1, \mathrm{IC}(T) - 2\mathrm{IC_{ms}}(t, t') + \mathrm{IC}(t')\right) \tag{6.7}$$

δηλαδή η ομοιότητα μεταξύ $\boldsymbol{t}$ και $\boldsymbol{t'}$, είναι μηδέν, αν η κανονικοποιημένη απόσταση τους είναι τουλάχιστον ένα. Η ομοιότητα του Lin μεταξύ των GO όρων ορίζεται ως:

$$\mathrm{sim}(t, t') = \frac{2\mathrm{IC_{ms}}(t, t')}{\mathrm{IC}(t) + \mathrm{IC}(t')} \tag{6.8}$$

Ένα άλλο μέτρο ομοιότητας που δημιουργήθηκε αργότερα βασίζεται στα μέτρα ομοιότητας του Resnik και του Lin. Το μέτρο αυτό ονομάζεται relevance και λαμβάνει υπόψη πόσο κοντά είναι οι GO όροι στο χαμηλότερο κοινό πρόγονό τους και πόσο ειδικός είναι ο κοινός πρόγονος τους. Το μέτρο ομοιότητας relevance ορίζεται ως:

$$\mathrm{sim}(\mathrm{t},\mathrm{t}') = \frac{2\mathrm{IC}_{\mathrm{ms}}(\mathrm{t},\mathrm{t}')}{\mathrm{IC}(\mathrm{t}) + \mathrm{IC}(\mathrm{t}')} \cdot (1 - \exp(-\mathrm{IC}_{\mathrm{ms}}(\mathrm{t},\mathrm{t}') \quad (6.9)$$

[34]

6.3.2 Ομοιότητες γονιδίων

Η βιβλιοθήκη διαθέτει εργαλεία για τον υπολογισμό της ομοιότητας των γονιδίων βασιζόμενη στην συνολική μετάφραση τους σε GO όρους.

Δεδομένων δυο γονιδίων $\boldsymbol{g}$ και $\boldsymbol{g'}$, που μεταφράζονται στους GO όρους, $\boldsymbol{t_1, t_2, \ldots, t_n}$ και $\boldsymbol{t'_1, t'_2, \ldots, t'_m}$ αντίστοιχα,ορίζεται η λειτουργική ομοιότητα μεταξύ των γονιδίων $\boldsymbol{g}$ και $\boldsymbol{g'}$ ως:

$$\mathrm{sim}_{\mathrm{gene}}(\mathrm{g},\mathrm{g}') = \max_{\substack{\mathrm{i}=1\ldots\mathrm{n} \\ \mathrm{j}=1\ldots\mathrm{m}}} \mathrm{sim}(\mathrm{t_i},\mathrm{t}'_{\mathrm{j}}) \quad (6.10)$$

όπου $\boldsymbol{sim}$ είναι το μέτρο ομοιότητας για να συγκριθούν οι GO όροι $\boldsymbol{t_i}$ και $\boldsymbol{t'_j}$.

Η προσέγγιση αυτή αναθέτει τον κάθε GO όρο $\boldsymbol{t_i}$, που αποτελεί μετάφραση του γονιδίου $\boldsymbol{g}$, στον πιο ταιριαστό GO όρο $\boldsymbol{t'_{ji}}$, που αποτελεί μετάφραση του γονιδίου $\boldsymbol{g'}$. Μπορούν πολλοί GO όροι του γονιδίου $\boldsymbol{g}$ να αναθέτονται σε ένα GO όρο του γονιδίου $\boldsymbol{g'}$. Το σκορ ομοιότητας υπολογίζεται λαμβάνοντας υπόψη τη μέση ομοιότητα των ανατεθειμένων GO όρων.

Το προκύπτον σκορ ομοιότητας, μπορεί να κανονικοποιηθεί για να αντιπροσωπεύει τον άνισο αριθμό των GO όρων και για τα δυο γονίδια. Μια από τις μεθόδους κανονικοποίησης είναι η παρακάτω και ονομάζεται sqrt (**sq**uare **r**oo**t**):

$$\mathrm{sim}_{\mathrm{gene}}(\mathrm{g},\mathrm{g}') = \frac{\mathrm{sim}_{\mathrm{gene}}(\mathrm{g},\mathrm{g}')}{\sqrt{\mathrm{sim}_{\mathrm{gene}}(\mathrm{g},\mathrm{g})\mathrm{sim}_{\mathrm{gene}}(\mathrm{g}',\mathrm{g}')}} \quad (6.11)$$

Δεδομένου ότι $sim_{gene}(g, g') \geq 0$, το αποτέλεσμα θα είναι μέτρο ομοιότητας ίσο με 1 για το γονίδιο $\boldsymbol{g}$ με τον εαυτό του και μεταξύ 0 και 1 με ένα άλλο γονίδιο.

Μια άλλη δυνατότητα είναι να χρησιμοποιηθεί η κανονικοποίηση του Lin:

$$sim(g, g') = \frac{2sim_{gene}(g, g')}{sim_{gene}(g, g) + sim_{gene}(g', g')} \tag{6.12}$$

ή η κανονικοποίηση του Tanimoto:

$$sim_{gene}(g, g') = \frac{sim_{gene}(g, g')}{sim_{gene}(g, g) + sim_{gene}(g', g') - sim_{gene}(g, g')} \tag{6.13}$$

Ένα άλλο σκορ ομοιότητας γονιδίων συγκρίνει τις μεταφράσεις των GO όρων από διαφορετικές οντολογίες. Το πρώτο βήμα της σύγκρισης αυτής των δυο γονιδίων είναι η σύγκριση των GO μεταφράσεων τους. Οι μεταφράσεις για τις διαφορετικές οντολογίες (MF και BP) εξετάζονται ξεχωριστά. Για δύο γονίδια $\boldsymbol{A}$ και $\boldsymbol{B}$ που μεταφράζονται στις ομάδες $\boldsymbol{GO^A}$ και $\boldsymbol{GO^B}$ με αντίστοιχα μεγέθη ομάδων $\boldsymbol{N}$ και $\boldsymbol{M}$, υπολογίζεται ο πίνακας ομοιότητας $\boldsymbol{S}$. Αυτός ο πίνακας περιέχει τα σκορ ομοιότητας μεταξύ των $\boldsymbol{GO^A}$ και $\boldsymbol{GO^B}$. Το σκορ αυτό υπολογίζεται από τον παρακάτω τύπο:

$$s_{ij} = sim\left(GO_i^A, GO_j^B\right), \forall i \in \{1, \dots N\}, \forall j \in \{1, \dots, M\} \tag{6.14}$$

Ο πίνακας μπορεί να υπολογιστεί με οποιοδήποτε μέτρο ομοιότητας παρουσιάστηκε παραπάνω. Ο πίνακας $\boldsymbol{S}$ δεν είναι απαραίτητα συμμετρικός μιας και το κάθε γονίδιο μεταφράζεται σε διαφορετικό πλήθος GO όρων. Οι γραμμές και οι στήλες του $\boldsymbol{S}$ αντιπροσωπεύουν δυο διαφορετικές συγκρίσεις. Τα διανύσματα γραμμής αντιπροσωπεύουν τη σύγκριση του A με το B ενώ τα διανύσματα στήλης τη σύγκριση του B με το A. Οι επιτυχίες για την σύγκριση μεταξύ των A και B ορίζονται ως οι μέγιστες τιμές των σειρών στον πίνακα ενώ οι μέγιστες τιμές στις στήλες είναι οι επιτυχίες για τη σύγκριση του B με το A.

Έχουμε αντίστοιχα :

$$rowScore = \frac{1}{N}\sum_{i=1}^{N} \max_{1 \le j \le M} s_{ij} \quad (6.15)$$

$$columnScore = \frac{1}{M}\sum_{j=1}^{M} \max_{1 \le i \le N} s_{ij}, \quad (6.16)$$

οι μεταβλητές $\boldsymbol{rowScore}$ και $\boldsymbol{columnScore}$ παίρνουν τιμές στο διάστημα [0,1].

Μια εναλλακτική είναι να υπολογιστεί το μέγιστο του $\boldsymbol{rowScore}$ και του $\boldsymbol{columnScore}$:

$$GOscore = \max\{columnScore, rowScore\}, \quad (6.17)$$

όπου $\boldsymbol{GOscore}$ είναι το γενικότερο όνομα του $\boldsymbol{MFscore}$ είτε του $\boldsymbol{BPscore}$.

Το συνολικό σκορ υπολογίζεται ορίζεται ως:

$$funSim = \frac{1}{2}\left[\left(\frac{BPscore}{\max(BPscore)}\right)^2 + \left(\frac{MFscore}{\max(MFscore)}\right)^2\right], \quad (6.18)$$

εδώ $\max(BPscore)$ και $\max(MFscore)$ υποδηλώνουν το μέγιστο πιθανό σκορ για τις αντίστοιχες οντολογίες. Αν χρησιμοποιηθεί το μέτρο ομοιότητας relevance το funSim σκορ λαμβάνει τιμές στο διάστημα [0,1]. [35]

6.3.3 Ομαδοποίηση γονιδίων

Οι υπολογιζόμενες ομοιότητες για μια δεδομένη ομάδα γονιδίων ή GO όρων μπορούν να χρησιμοποιηθούν για να ομαδοποιηθούν τα γονίδια με βάση τις λειτουργίες τους. Κατά την ομαδοποίηση των δεδομένων, εξετάζεται πόσο όμοιες είναι αυτές οι ομάδες με βάση τις GO μεταφράσεις των γονιδίων. Μια απεικόνιση της ομαδοποίησης αυτών γίνεται μέσω των ομαδοποιημένων σιλουέτων (cluster silhouettes).

Λαμβάνονται οι Entrez ταυτότητες των γονιδίων, υπολογίζονται οι γονιδιακές ομοιότητες και πραγματοποιείται μια ιεραρχική ομαδοποίηση χρησιμοποιώντας την μέθοδο του Ward.

Η ιεραρχική ανάλυση σε ομάδες (Agglomerative Hierarchical Clustering) είναι μια μέθοδος ομαδοποίησης των δεδομένων. Η μέθοδος περιλαμβάνει τα παρακάτω βήματα:

1. Αρχίζουμε με N ομάδες, με την κάθε μία να περιέχει μόνο ένα στοιχείο και έναν NxN πίνακα με αποστάσεις.
2. Βρίσκουμε στον πίνακα το ζεύγος U και V ομάδων με την μικρότερη απόσταση μεταξύ τους.
3. Ενώνουμε τις ομάδες U και V σε μια ομάδα, έστω UV. Ανανεώνουμε τον πίνακα αποστάσεων διαγράφοντας τις γραμμές και στήλες που αντιστοιχούν στις U και V και προσθέτοντας μια γραμμή και μια στήλη με τις αποστάσεις της UV από τις υπόλοιπες ομάδες.
4. Επαναλαμβάνουμε τα βήματα 2 και 3, (N-1) φόρες μέχρι να υπάρχει μόνο μια ομάδα. Καταγράφουμε τις ομάδες που δημιουργήθηκαν κατά τη διάρκεια της διαδικασίας και το επίπεδο (απόσταση) στο οποίο δημιουργήθηκε η κάθε μία.

Μια από τις επιλογές για την απόσταση μεταξύ ομάδων είναι η μέθοδος του Ward (Ward's Hierarchical Clustering). Σε κάθε ομάδα k θεωρούμε ως $\boldsymbol{ESS_k}$, το άθροισμα των τετραγώνων των αποστάσεων κάθε στοιχείου της ομάδας από τον μέσο της ομάδας και $\boldsymbol{ESS}$ το άθροισμα των $\boldsymbol{ESS_k}$. Ως απόσταση μεταξύ δύο ομάδων U και V θεωρούμε την αύξηση που θα προκύψει στο $\boldsymbol{ESS}$ από την ένωση των δύο ομάδων. [25]

Η ομαδοποίηση αυτή θα παρασταθεί με τη μορφή ενός δενδρογράμματος. Κόβεται το δένδρο της ομαδοποίησης που δημιουργείται και εξετάζονται οι ομαδοποιημένες σιλουέτες (clustering silhouettes). Οι ομαδοποιημένες σιλουέτες είναι ένας κλασσικός τρόπος να απεικονιστεί η ποιότητα μιας δεδομένης ομαδοποίησης των γονιδίων. Η τιμή της σιλουέτας για κάθε σημείο σε μια ομαδοποίηση είναι ένα μέτρο του πόσο όμοιο είναι αυτό το σημείο με τα σημεία στην δική του ομάδα εναντίον των σημείων σε άλλες ομάδες. Αυτή η τιμή κυμαίνεται μεταξύ -1 και 1, και ορίζεται ως :

$$S(i) = \frac{\min_j \left(\bar{d}_B(i,j)\right) - \bar{d}_W(i)}{\max\left(\bar{d}_W(i), \min_j(\bar{d}_B(i,j))\right)}, \quad (6.19)$$

όπου $\bar{d}_W(i)$ είναι η μέση απόσταση του i σημείου από τα άλλα σημεία στην ίδια ομαδοποίηση, $\bar{d}_B(i,j)$ είναι η μέση απόσταση του i σημείου από τα σημεία σε μια άλλη ομαδοποίηση j. Η ποιότητα μιας δεδομένης ομαδοποίησης μπορεί να εκφραστεί από την μέση τιμής της σιλουέτας των σημείων που ανήκουν στην συγκεκριμένη ομαδοποίηση. [12]

6.3.4 Λειτουργίες της βιβλιοθήκης

Για τις παραπάνω λειτουργίες της βιβλιοθήκης "GOSim", στο περιβάλλον της R, απαιτούνται να χρησιμοποιηθούν και οι παρακάτω βιβλιοθήκες:

- ✓ "GOstats" [14]
- ✓ "mclust" [36]
- ✓ "cluster" [37]

Ακόμη χρησιμοποιείται το πακέτο "Rgraphviz" [38] για να απεικονιστούν τα GO γραφήματα. Για να υπολογιστούν οι λειτουργικές ομοιότητες μεταξύ των γονιδιακών προϊόντων απαιτούνται οι Entrez ταυτότητες των γονιδίων. Η χαρτογράφηση των ταυτοτήτων αυτών στην GO οντολογία, παρέχεται από το πακέτο "GO.db" [39]. Τα γονίδια που δεν μεταφράζονται φιλτράρονται και απομακρύνονται αυτόματα, ενώ πραγματοποιούνται υπολογισμοί ομοιότητας για τα αυτά. Για κάθε υπολογισμό ορίζεται και η οντολογία για την οποία θα πραγματοποιηθεί.

Συνοψίζοντας, η βιβλιοθήκη αυτή, παρέχει μεθόδους της R για τους ακόλουθους σκοπούς :

- ✓ Δημιουργία των GO γραφημάτων, ανά οντολογία
- ✓ Υπολογισμός του πληροφοριακού περιεχομένου των GO όρων και της ομοιότητας μεταξύ τους, ανά οντολογία
- ✓ Υπολογισμός της ομοιότητας μεταξύ γονιδίων με βάση της μετάφρασης τους σε GO όρους, ανά οντολογία
- ✓ Φιλτράρισμα και εκτύπωση της μετάφρασης των GO μιας δεδομένης λίστας γονιδίων

- ✓ Αξιολόγηση μιας ομαδοποίησης γονιδίων ή GO όρων μέσω προϋπολογισμένων ομοιοτήτων

6.3.4.1 Σχεδιασμός GO γραφημάτων

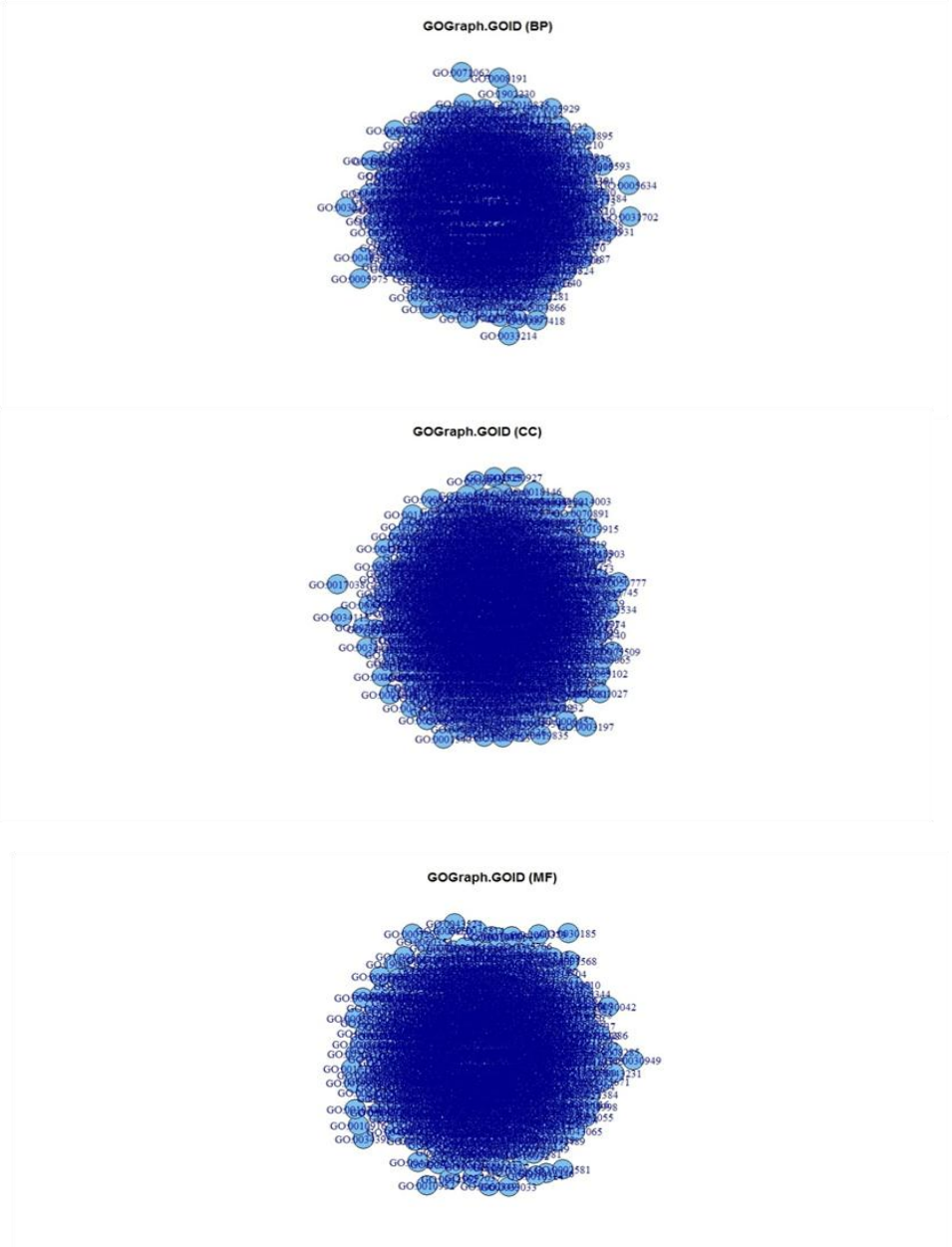
Για κάθε ανάλυση που πραγματοποιείται, αρχικά ορίζεται η οντολογία που θα βασιστούν οι επόμενοι υπολογισμοί και φορτώνεται, στο περιβάλλον της R, το περιεχόμενο των πληροφοριών όλων των GO όρων εντός αυτής της οντολογίας. Κατά το χρόνο φόρτωσης της βιβλιοθήκης η προεπιλεγμένη οντολογία είναι η BP.

Χρησιμοποιούνται δυο λειτουργίες της βιβλιοθήκης για δημιουργία GO γραφημάτων:

- ✓ Η πρώτη λειτουργία της βιβλιοθήκης απαιτεί ως είσοδο τις GO ταυτότητες, παράγοντας ένα γράφημα την σύνδεσης των GO όρων, ενώ παρέχει τη δυνατότητα να κλαδεύεται το γράφημα με βάση την τάξη των προγόνων που θα επιλεχθεί να περιλαμβάνει.
- ✓ Η δεύτερη λειτουργία αφορά την δημιουργία γραφήματος ανά γονίδιο i. Απαιτεί ως είσοδο τις Entrez ταυτότητες των γονιδίων, και σχηματίζεται ένα γράφημα που δείχνει για κάθε γονίδιο i, που εντοπίζονται οι GO όροι στην GO δομή. Αντίστοιχα και εδώ παρέχεται η δυνατότητα να κόβεται το γράφημα με βάση την τάξη των προγόνων που θα οριστεί.

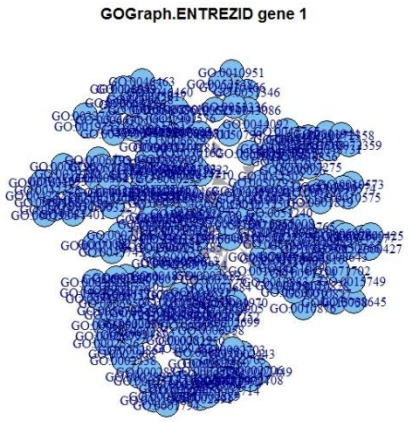
Από το Σύνολο Α λαμβάνονται Uniprot ταυτότητες πρωτεϊνών οι οποίες μεταφράζονται σε Entrez ταυτότητες γονιδίων και GO ταυτότητες, μέσω της βιβλιοθήκης “org.Hs.eg.db” και έπειτα χρησιμοποιούνται ως είσοδος στις παραπάνω δύο λειτουργίες.

Αρχικά έχουμε τα GO γραφήματα ανά οντολογία που δημιουργούνται με τις GO ταυτότητες και κλαδεύονται ώστε να περιλαμβάνουν μέχρι πρώτης τάξης προγόνους («γονείς» των όρων) σε αντίθεση με αυτά που παρουσιάστηκαν στο Κεφάλαιο 6.2 όπου αναζητούνται οι σχέσεις των όρων περιλαμβάνοντας μεγαλύτερης τάξης προγόνους («γονείς» των «γονέων» κ.ό.κ.). (Σχήμα 6.35)



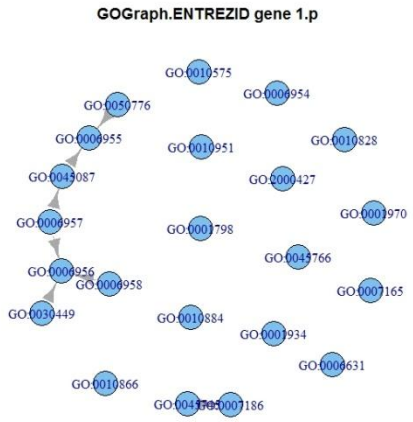
Σχήμα 6.35: GO γραφήματα γονιδίων Συνόλου A σε οντολογία BP, CC και MF

Ακολουθούν γραφήματα που παρουσιάζουν το GO γράφημα που λαμβάνεται με είσοδο τις Entrez ταυτότητες γονιδίων και δείχνουν για το πρώτο και το δεύτερο γονίδιο που εντοπίζονται οι GO όροι στην GO δομή, καθώς και τα αντίστοιχα γραφήματα για πρώτης τάξης προγόνους των γονιδίων αυτών. Η προεπιλεγμένη οντολογία είναι η BP.



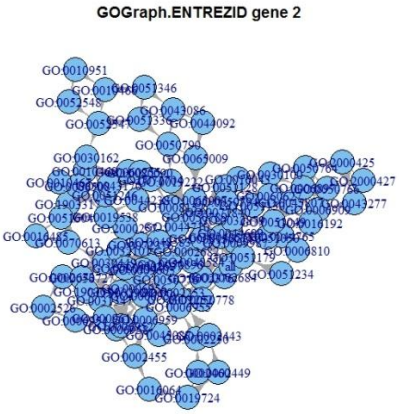
Σχήμα 6.36: GO γράφημα γονιδίων Συνόλου Α για το 1^{ο} γονίδιο σε οντολογία BP

Το γράφημα στο Σχήμα 6.36 απεικονίζει τις σχέσεις των GO όρων που αποτελούν μεταφράσεις του πρώτου γονιδίου με Entrez ταυτότητα 718. Όπως φαίνεται στο γράφημα όλοι οι GO όροι του γονιδίου αυτού έχουν σχέση μεταξύ τους. Πιο συγκεκριμένα οι περισσότεροι όροι συνδέονται με τους όρους: GO:0002367 (παραγωγή κυτοκίνης που μετέχει στην απόκριση του ανοσοποιητικού συστήματος), GO:0050896 (απόκριση σε ερέθισμα), GO:0051179 (εντοπισμός θέσης), GO:0008150 (βιολογική διεργασία), GO:0050789 (ρύθμιση βιολογικής διεργασίας), GO:0009987 (κυτταρική διεργασία), GO:0032502 (αναπτυξιακή διεργασία), GO:0031323 (ρύθμιση κυτταρικής μεταβολικής διεργασίας), GO:0019538 (μεταβολική διεργασία πρωτεΐνης), GO:0006956 (ενεργοποίηση συμπληρώματος). Ενώ οι κόμβοι φύλλα του γραφήματος είναι: GO:0001798 (θετική ρύθμιση της υπερευαισθησίας τύπου IIa), GO:0001970 (θετική ρύθμιση της ενεργοποίησης του συγκροτήματος που αφορά σε επίθεση στην μεμβράνη), GO:0006631 (μεταβολική διεργασία λιπαρού οξέος), GO:0006958 (ενεργοποίηση του συμπληρώματος, κλασσικό μονοπάτι), GO:0010951 (αρνητική ρύθμιση της δραστηριότητας των ενδοπεπτιδάσων), GO:0045766 (θετική ρύθμιση της αγγειογένεσης), GO:0006957 (ενεργοποίηση συμπληρώματος, εναλλακτικό μονοπάτι), GO:0001934 (θετική ρύθμιση της φωσφορυλίωσης πρωτεϊνών), GO:0010828 (θετική ρύθμιση μεταφοράς της γλυκόζης), GO:0010866 (ρύθμιση της βιοσυνθετικής διεργασίας των τριγλυκεριδίων), GO:0010884 (θετική ρύθμιση της αποθήκευσης των λιπιδίων).



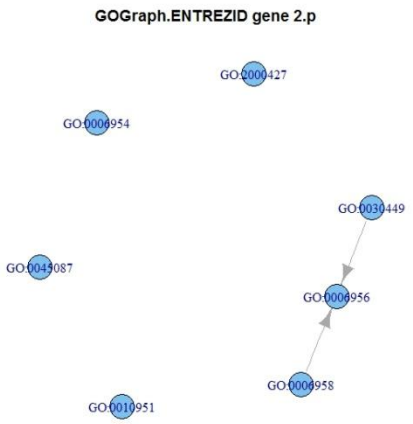
Σχήμα 6.37: GO γράφημα γονιδίων Συνόλου Α για 1ης τάξης προγόνους για το 1ο γονίδιο σε οντολογία BP

Στο Σχήμα 6.37 απεικονίζονται οι σχέσεις που έχουν μεταξύ τους οι 1ης τάξης πρόγονοι («γονείς») των GO όρων του 1ου γονιδίου. Εδώ παρατηρούμε ότι η πλειοψηφία των «γονέων» των GO όρων δεν έχουν σχέση μεταξύ τους. Με την βιολογική λειτουργική ταυτότητα της ενεργοποίησης του συμπληρώματος (GO:0006956) έχουν σχέση οι όροι GO:0006958 (ενεργοποίηση του συμπληρώματος, κλασσικό μονοπάτι), GO:0030449 (ρύθμιση της ενεργοποίησης του συμπληρώματος) και GO:0006957 (ενεργοποίηση του συμπληρώματος, εναλλακτικό μονοπάτι). Με την ταυτότητα GO:0006957 έχουν σχέση οι ταυτότητες GO:0045087 (έμφυτη ανοσοαπόκριση), GO:0006950 (απόκριση σε στρες) και GO:0050776 (ρύθμιση της απόκρισης του ανοσοποιητικού συστήματος).



Σχήμα 6.38: GO γράφημα γονιδίων Συνόλου Α για το 2ο γονίδιο σε οντολογία BP

Αντίστοιχα το γράφημα στο Σχήμα 6.38 απεικονίζει τις σχέσεις των GO όρων στις οποίες μεταφράζεται το γονίδιο με Entrez ταυτότητα 720. Οι κόμβοι φύλλα (GO-IN) του γραφήματος είναι: GO:0006958 (ενεργοποίηση του συμπληρώματος, κλασσικό μονοπάτι), GO:0010951 (αρνητική ρύθμιση της δραστηριότητας της ενδοπεπτιδάσης), GO:0030449 (ρύθμιση της ενεργοποίησης του συμπληρώματος), GO:0045087 (έμφυτη ανοσοαπόκριση), GO:2000427 (θετική ρύθμιση της αποπτωτικής κάθαρσης των κυττάρων). Οι συνδέσεις των περισσότερων GO όρων του γραφήματος καταλήγουν στου παρακάτω όρους (GO-OUT): GO:0003674 (μοριακή λειτουργία), GO:0005488 (δέσμευση), GO:0005515 (δέσμευση πρωτεΐνης), GO:0050789 (ρύθμιση βιολογικής διεργασίας), GO:0019538 (μεταβολική διεργασία πρωτεϊνών), GO:0019222 (ρύθμιση μεταβολικής διεργασίας).



Σχήμα 6.39: GO γράφημα γονιδίων Συνόλου Α για 1^{ης} τάξης προγόνους για το 2^{ο} γονίδιο σε οντολογία BP

Στο Σχήμα 6.39 απεικονίζονται οι σχέσεις των «γονέων» των GO όρων στους οποίους μεταφράζεται στο 2^{ο} γονίδιο. Μόνο τρεις όροι έχουν σχέση μεταξύ τους και αυτοί είναι οι GO:0006958 (ενεργοποίηση του συμπληρώματος, κλασσικό μονοπάτι), GO:0006956 (ενεργοποίηση του συμπληρώματος) και GO:0030449 (ρύθμιση της ενεργοποίησης του συμπληρώματος).

6.3.4.2 Υπολογισμός ομοιότητας γονιδίων και ομαδοποίησή τους

Για τον υπολογισμό των λειτουργικών ομοιοτήτων των γονιδίων απαιτείται ως είσοδος οι Entrez ταυτότητες των γονιδίων. Ακόμη θα πρέπει να επιλεχθεί η μέθοδος που θα χρησιμοποιηθεί για να υπολογιστούν οι λειτουργικές ομοιότητες των

γονιδίων, καθώς και η μέθοδος με βάση την οποία θα υπολογιστεί η ομοιότητα των GO όρων. Ορίζεται αν είναι επιθυμητή η κανονικοποίηση των δεδομένων, και αν ναι, επιλέγεται μια από τις μεθόδους που αναφέρθηκαν παραπάνω. Αφού υπολογιστεί η ομοιότητα των γονιδίων και δημιουργηθεί ένας πίνακας ομοιότητας, μπορεί να αξιολογηθεί μια ομάδα γονιδίων ή όρων όσο αφορά την ομοιότητά τους.

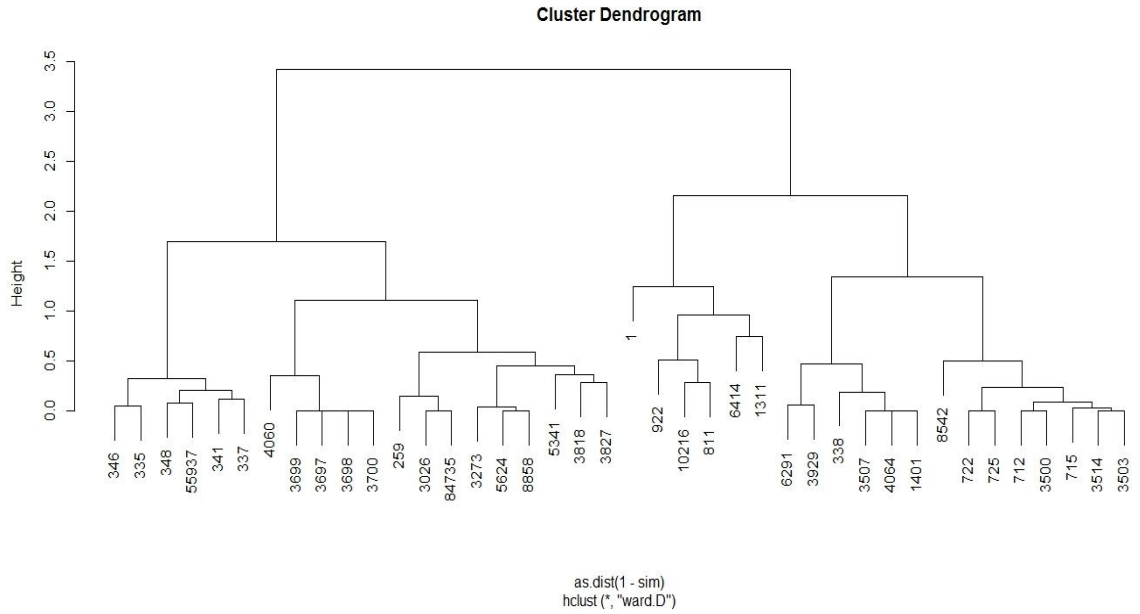
Ο υπολογισμός των λειτουργικών ομοιοτήτων των γονιδίων θα πραγματοποιηθεί για την λίστα των πιο σημαντικών γονιδίων. Έτσι λαμβάνονται οι συμβολικές ονομασίες των γονιδίων από το Σύνολο Β και μεταφράζονται σε Entrez ταυτότητες γονιδίων μέσω τις βιβλιοθήκης "org.Hs.eg.db". Έπειτα επιλέγεται η μέθοδος funSimMax για τον υπολογισμό των ομοιοτήτων των γονιδίων και η μέθοδος relevance για τον υπολογισμό των ομοιοτήτων των GO όρων . Η κανονικοποίηση των δεδομένων θα γίνει με τη μέθοδο sqrt. Μέρος του πίνακα ομοιότητας που δημιουργείται περιλαμβάνεται στον Πίνακα ΠΑ3, ενώ τμήμα του παρουσιάζεται στον Πίνακα 6.2.

Πίνακας 6.4: Σκορ ομοιότητας sim των 5 πρώτων γονιδίων

	259	3026	3698	3697	3699
259	1	0.92	0.8	0.8	0.8
3026	0.92	1	0.72	0.72	0.72
3698	0.8	0.72	1	1	1
3697	0.8	0.72	1	1	1
3699	0.8	0.72	1	1	1

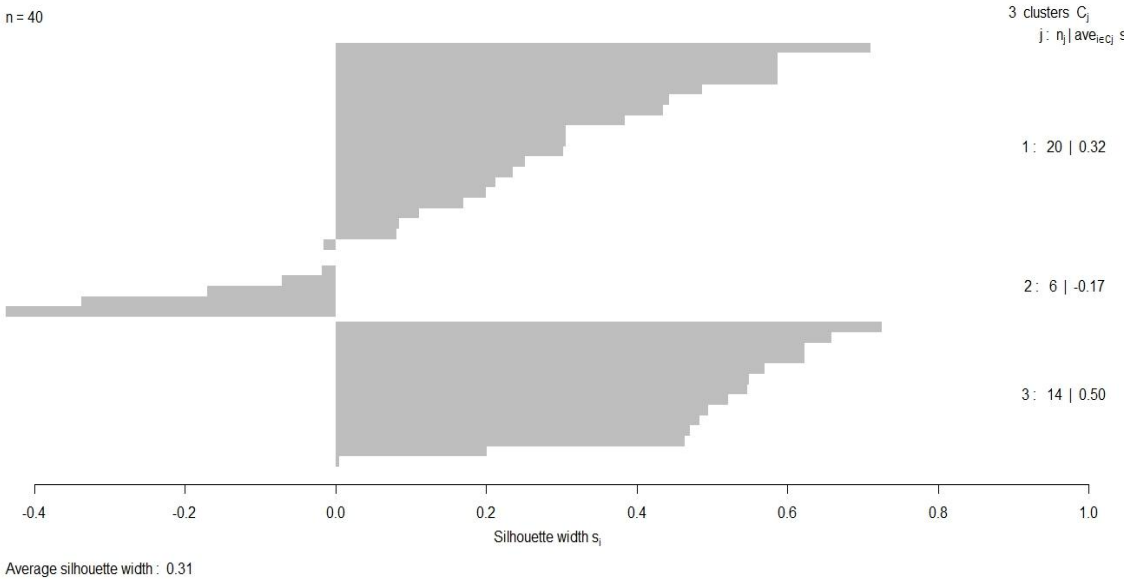
Όπως παρατηρείται στον παραπάνω πίνακα ομοιότητας το σκορ ομοιότητας sim για το γονίδιο με τον εαυτό του είναι ένα και μικρότερο ή ίσο του ένα σε σύγκριση με άλλα γονίδια.

Έχοντας τις λειτουργικές ομοιότητες των γονιδίων, υπολογίζεται μια ομαδοποίηση χρησιμοποιώντας την μέθοδο του Ward με κριτήριο την απόσταση (***1-sim***) για κάθε γονίδιο. Έτσι δημιουργείται το παρακάτω δενδρόγραμμα. (Σχήμα 6.40)



Σχήμα 6.40: Δενδρόγραμμα ομαδοποίησης γονιδίων Συνόλου Β

Το δενδρόγραμμα κόβεται επιλέγοντας να λάβουμε τρεις ομάδες των δεδομένων (3 clusters) και εξετάζεται η ποιότητα της ομαδοποίησης αυτών μέσω των μέτρων σιλουετών που δημιουργούνται.



Σχήμα 6.41: Γράφημα ποιότητας της ομαδοποίησης των γονιδίων του Συνόλου Β σε 3 ομάδες

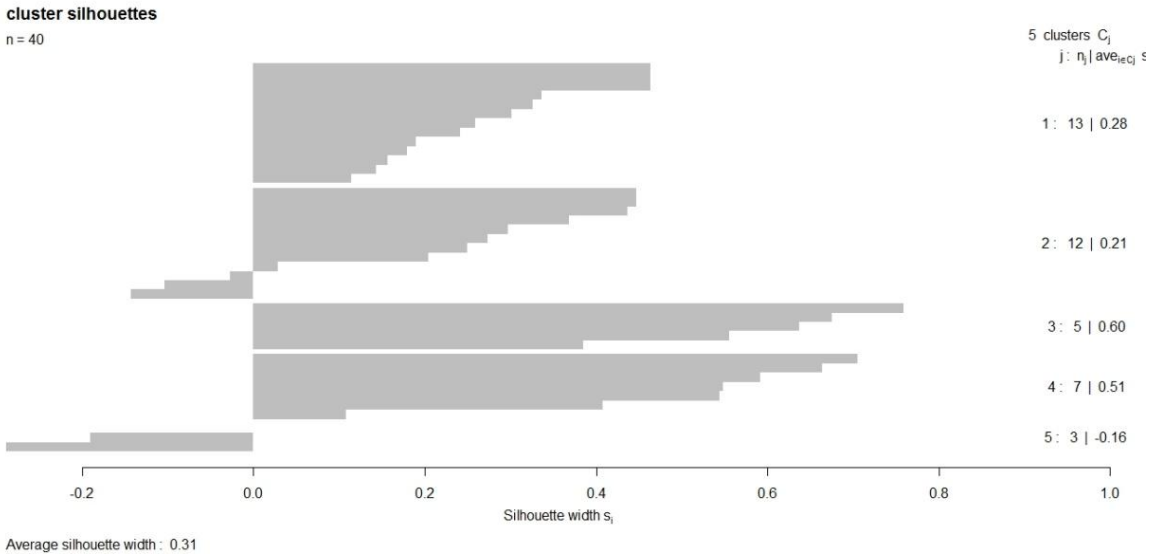
Τα γονίδια χωρίζονται σε 3 ομάδες με βάση την λειτουργικής τους ομοιότητα, όπως παρατηρείται στο γράφημα των σιλουετών (Σχήμα 6.41):

1. 20 γονίδια με S=0,32

2. 6 γονίδια με S=-0,17
3. 14 γονίδια με S=0,5

Η τιμή του ***S(i)***, όπως αναφέρθηκε και στην Εξίσωση 6.19, κυμαίνεται μεταξύ -1 και 1. Λαμβάνει την μέγιστη τιμή 1, όταν η ομαδοποίηση που πραγματοποιήθηκε με βάση την λειτουργική ομοιότητα τους είναι αρκετά ικανοποιητική, ενώ μειώνεται όταν ελαττώνεται η ποιότητα της ομαδοποίησης, δηλαδή όταν οι ομάδες δεν έχουν κοινά στις λειτουργικές τους ιδιότητες. Τιμές στο διάστημα [-1,0] φανερώνει ότι τα αντίστοιχα στοιχεία έχουν τοποθετηθεί σε λάθος ομάδα. Γι αυτό και επιλέγουμε να χωριστούν τα δεδομένα σε περισσότερες από τις 3 ομάδες, ώστε να αυξηθεί η τιμή του ***S***, και η ομαδοποίηση να είναι πιο ικανοποιητική.

Επιλέγουμε να χωριστούν τα γονίδια σε 5 ομάδες και παρατηρούμε στο παρακάτω γράφημα των σιλουέτων (Σχήμα 6.42) ότι οι τιμές των ***S*** αυξάνονται αρκετά, ενώ μία ομάδα έχει τιμή ***S*** μικρότερης του μηδενός.



Σχήμα 6.42: Γράφημα ποιότητας της ομαδοποίησης των γονιδίων του Συνόλου Β σε 5 ομάδες

Τα γονίδια χωρίζονται σε 5 ομάδες με βάση την λειτουργικής τους ομοιότητα:

1. 13 γονίδια με S=0,28
2. 12 γονίδια με S=0,21
3. 5 γονίδια με S=0,60
4. 7 γονίδια με S=0,51
5. 3 γονίδια με S=-0,16

Θεωρούμε την ομαδοποίηση αυτή πιο ικανοποιητική από την προηγούμενη αν και η $5^{η}$ ομάδα των 3 γονιδίων έχει S=-0,16. Ο χωρισμός των γονιδίων σε περισσότερες από πέντε ομάδες εξακολουθεί να μας δίνει ομάδες με τιμές ***S*** μικρότερες του μηδενός γι' αυτό και επιλέγουμε την ομαδοποίηση σε λιγότερες ομάδες ως σωστότερη.

Τα γονίδια με την μορφή των Entrez ταυτοτήτων τους χωρίζονται σε πέντε ομάδες (clusters):

C1 : 259, 3026, 3698, 3697, 3699, 4060, 5624, 84735, 8858, 5341, 3818, 3273, 3700

C2 : 10216, 3514, 3503, 8542, 3827, 922, 811, 712, 722, 725, 3500, 715

C3 : 4064, 1401, 3507, 6291, 3929

C4 : 338, 341, 348, 55937, 346, 335, 337

C5 : 6414, 1311, 1

Εφόσον C3 και C4 ομάδα λαμβάνουν ικανοποιητικές τιμές S αποτελούν ιδανικές ομάδες λειτουργικής ομοιότητας των γονιδίων που τις αποτελούν. Συνεπώς θα εξετάσουμε σε ποιες βιολογικές διεργασίες συμμετέχουν τα γονίδια αυτά και την σχέση τους με την τοξικότητα. Κρίνεται απαραίτητο λοιπόν να αντιστοιχίσουμε τις Entrez ταυτότητες σε GO ταυτότητες βιολογικών λειτουργιών και έπειτα να εντοπίσουμε τις κοινές λειτουργίες που μετέχουν τα γονίδια της κάθε ομάδας.

Η μεταφραστική βιβλιοθήκη "mygene" [40] μας δίνει τη δυνατότητα με ευκολία και αποτελεσματικότητα να μεταφράσουμε τις ταυτότητες γονιδίων (SYMBOL, Entrez, Ensembl) σε GO όρους λαμβάνοντας και την αντίστοιχη περιγραφή των βιολογικών λειτουργιών που αντιπροσωπεύουν.

Η ομάδα C3 περιέχει γονίδια που κατά κύριο λόγο συμμετέχουν στις παρακάτω βιολογικές λειτουργίες:

1. φλεγμονώδης απόκριση (inflammatory response) [41]
2. έμφυτη ανοσοαπόκριση (innate immune response) [41]
3. κυτταρική απόκριση στα λιποσακχαρίδια (cellular response to liposaccharide)
4. μονοπάτι σηματοδότησης με διαμεσολάβηση λιποσακχαριδίων (liposaccharide-mediated signaling pathway)
5. απόκριση οξείας φάσης (acute-phase response) [42]

6. ενεργοποίηση μακροφάγων ως ανοσοαπόκριση (macrophage activation, macrophage immune response) [43]
7. παραγωγή/ έκκριση ιντερλευκίνης-8 (interleukin-8 production/secretion) [44]

Οι λειτουργίες 1,2,5,6 και 7 σχετίζονται βάση της βιβλιογραφίας με την τοξικότητα.

Τα γονίδια της ομάδας C4 συμμετέχουν σε περισσότερες λειτουργίες και πιο συγκεκριμένα:

1. μεταβολικές διεργασίες ρετινοειδών (retinoid metabolic process)
2. ενδοκυττάρωση με τη διαμεσολάβηση υποδοχέα (receptor-mediated endocytosis)
3. πήξη του αίματος (blood coagulation) [45]
4. αποθήκευση/ μεταφορά/ έκκριση/ εστεροποίηση/ μεταβολική διεργασία/ καταβολική διεργασία/ βιοσυνθετική διεργασία/ ομοιόσταση της χολυστερόλης (cholesterol storage/ transport/ efflux/ esterification/ metabolic process/ catabolic process/ biosynthetic process/ homeostasis) [46]
5. αποθήκευση/ μεταφορά/ μεταβολική διεργασία/ καταβολική διεργασία/ βιοσυνθετική διεργασία/ έκκριση/ ομοιόσταση των λιπιδίων (lipid storage/ transport/ metabolic process/ catabolic process/ biosynthetic process/ efflux/ homeostasis) [8]
6. μεταφορά λιπιδίων μέσω αιματοεγκεφαλικού φραγμού (lipid transport across blood brain barrier) [47]
7. κινητοποίηση/ μεταβολική διεργασία/ καταβολική διεργασία/ ομοιόσταση τριγλυκεριδίων (triglyceride mobilization/ metabolic process/ catabolic process/ homeostasis)
8. αναδιαμόρφωση/ εκκαθάριση σωματιδίων λιποπρωτεϊνών (lipoprotein particle remodeling/ clearance) [8]
9. αποκοκκίωση/ ενεργοποίηση των αιμοπεταλίων (platelet degranulation/ activation) [48]
10. αποπτωτική διεργασία/θάνατος νευρώνα (neuron apoptic process/death) [49]
11. ρύθμιση νευρωνικής συναπτικής πλαστικότητας (regulation of neuronal synaptic plasticity)
12. μετανάστευση ενδοθυλιακών κυττάρων των αιμορφόρων αγγείων (blood vessel endothelial cell migration) [50]
13. βιοσυνθετική διεργασία λιπαρών οξέων (fatty acid biosynthetic process)

Με βάση τη βιβλιογραφία οι λειτουργίες 3, 4, 5, 6, 8, 10, 12 σχετίζονται άμεσα με την τοξικότητα. Συμπεραίνεται ότι ο υπολογισμός της λειτουργικής ομοιότητας αποτέλεσε σημαντικό εργαλείο για να εντοπίσουμε ποια γονίδια και συγκεκριμένα ποιες βιολογικές λειτουργίες των γονιδίων σχετίζονται με την τοξικότητα. Αντίθετα η δημιουργία γραφημάτων μας έδωσε πληθώρα πληροφοριών για γενικότερες λειτουργίες που συμμετέχουν αυτά χωρίς να είναι εύκολα αξιοποιήσιμες.

Οι ομαδοποιήσεις των γονιδίων θα χρησιμοποιηθούν στις αναλύσεις εμπλουτισμού που θα ακολουθήσουν στο Κεφάλαιο 7.

7 Ανάλυση εμπλουτισμού γονιδίων

Η ανάλυση εμπλουτισμού γονιδίων (Gene Enrichment Analysis), είναι ένα ισχυρό αναλυτικό εργαλείο με το οποίο ερμηνεύονται τα δεδομένα της γονιδιακής έκφρασης κυρίως αλλά και άλλων βιολογικών πειραμάτων με όμοια δομή (μετρήσεις γονίδιων σε δείγματα). Κατά την ανάλυση αυτή ο ερευνητής αναζητεί το βιολογικό ενδιαφέρον σε μια ομάδα γονιδίων.

Με την συγκέντρωση και διάδοση γνώσης για τα γνωστά γονίδια σε βάσεις δεδομένων όπως αυτή της Γονιδιακής Οντολογίας, δίνεται η δυνατότητα στους ερευνητές να αποδώσουν χαρακτηριστικά σε ομάδες γονιδίων που προκύπτουν από πειράματα ή αναλύσεις. Το αρχικό σύνολο γονιδίων μπορεί να προκύπτει από μια ανάλυση έκφρασης ή από μια πειραματική διαδικασία ή να βασίζεται σε βιολογική γνώση. Για να αναγνωριστεί η σημαντικότερη πληροφορία μέσα σε αυτό το σύνολο πρέπει να αναζητηθεί ο εμπλουτισμός του, δηλαδή να εκτιμηθεί ένα υποσύνολο των γονιδίων αυτών που παρουσιάζει σημαντική υπερ-εκπροσώπηση κάποιων βιολογικών χαρακτηριστικών.

Εφόσον προσδιοριστεί μια λίστα γονιδίων, οι ταυτότητες των οποίων αποτελούν το αντικείμενο της ανάλυσης αυτής, καθώς και οι μεταφράσεις τους σε όρους της βάσης δεδομένων, πραγματοποιείται ο εμπλουτισμός των αντιστοιχιζόμενων στα γονίδια όρων. Δεδομένου ενός συνόλου γονιδίων που αποτελούν το υπόβαθρο (όπως τα γονίδια του ανθρώπινου οργανισμού ενός τσιπ) και του συνόλου των γονιδίων της προς μελέτη λίστας, αναγνωρίζονται ποιοι όροι σχετίζονται με τα γονίδια της λίστας και εξετάζεται ο ισχυρισμός ότι αυτή η συσχέτιση είναι σημαντικά διαφορετική από ότι θα αναμενόταν κατά τύχη. Ο έλεγχος αυτός γίνεται με βάση τις αναλογίες των γονιδίων από το σύνολο που έχει ο κάθε όρος της βάσης δεομένων. Ο παρακάτω πίνακας αποτελεί παράδειγμα της ανάλυσης για ένα σύνολο 100 γονιδίων που εξετάζονται.

Πίνακας 7.1: Παράδειγμα ανάλυσης εμπλουτισμού για ένα σύνολο 100 γονιδίων

Βιολογική διεργασία	γονίδια του υποβάθρου	#γονίδια που αναμένονται τυχαία στα 100	#γονίδια που προέκυψαν στα 100
αιμόσταση	800/1000	80	80
απόκριση σε ερέθισμα	400/1000	40	30
μεταφορά γλυκόζης	50/1000	5	20

Στον Πίνακα 7.1 μπορούμε να δούμε ότι 80 από τα 100 γονίδια της λίστας σχετίζονται με την αιμόσταση, 30 από τα 100 γονίδια σχετίζονται με την απόκριση σε ερέθισμα, ενώ η πιο ενδιαφέρουσα διεργασία είναι η μεταφορά της γλυκόζης. Αυτό συμβαίνει διότι τα γονίδια που αντιστοιχίζονται σε αυτή την κατηγορία (20 γονίδια) είναι πολύ περισσότερα από το αναμενόμενο κατά τύχη να αντιστοιχίζονται (5 γονίδια), το οποίο σημαίνει ότι υπερεκπροσωπείται.

Μια κοινή στατιστική προσέγγιση είναι να χρησιμοποιηθεί ένα στατιστικό τεστ για κάθε γονίδιο και έτσι να ποσοτικοποιηθεί το ενδιαφέρον κάθε γονιδίου με μια τιμή ***p*** (p-value), να ρυθμιστούν οι τιμές ***p*** για πολλαπλές συγκρίσεις, να επιλεχθεί ένα κατάλληλο όριο αποκοπής (cut-off) ή αλλιώς επίπεδο σημαντικότητας, και να δημιουργηθεί μια λίστα των ενδιαφερόντων γονιδίων.

Πιο συγκεκριμένα η βασική διαδικασία για την ανάλυση εμπλουτισμού των γονιδίων και στις δυο διαφορετικές περιπτώσεις είναι:

- ✓ Ορίζεται μια μηδενική υπόθεση
- ✓ Εξετάζεται για κάθε γονίδιο με βάση ένα στατιστικό τεστ
- ✓ Παράγεται μια τιμή ***p*** (p-value)
- ✓ Δημιουργείται ένα επίπεδο σημαντικότητας ***α***
- ✓ Κάθε γονίδιο λαμβάνει ένα σκορ
- ✓ Με βάση το σκορ, διαμορφώνεται η λίστα των σημαντικών γονιδίων

Αυτή η «συνταγή» αναφέρεται ως οριακή προσέγγιση. Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι τα γονίδια, για τα οποία είναι γνωστό ότι είναι βιολογικά συνδεδεμένα, βαθμολογούνται ανεξάρτητα, άρα αγνοείται η βιολογική πληροφορία που αφορά την συνολική λειτουργία των γονιδίων. Παρόλα αυτά οι περισσότερες

προσεγγίσεις της ανάλυσης εμπλουτισμού των γονιδίων βασίζονται σε αποτελέσματα που παράγει αυτή η οριακή προσέγγιση. Δυο βασικές τέτοιου τύπου διαδικασίες για ανάλυση έκφρασης γονιδίων είναι, η *προσέγγιση υπερ-εκπροσώπησης* και η *προσέγγιση συνολικού σκόρ*. Και στις δυο περιπτώσεις τα σετ γονιδίων σχηματίζονται πριν τη στατιστική ανάλυση. Τα σετ σχηματίζονται από γονίδια που είναι μέρος τα ίδιων κυτταρικών συστατικών, ή είναι απαραίτητα για μια βιολογική λειτουργία ή έχουν την ίδια μοριακή λειτουργία ή αποτελούν βασικό μέρος της περίπτωσης που εξετάζεται όπως μια ασθένεια.

Πρόσφατα έχουν προταθεί πολλές μέθοδοι, που περιλαμβάνουν στην ανάλυση τους αυτή τη βιολογική πληροφορία. Μια από αυτές είναι η Ανάλυση Εμπλουτισμού Συνόλου Γονιδίων (Gene Set Enrichment Analysis ή GSEA), η οποία βασίζεται σε στατιστικά τεστ, τα οποία έχουν έλλειψη ευαισθησίας. [51]

Τα πιο συχνά χρησιμοποιούμενα στατιστικά τεστ για αυτές τις αναλύσεις είναι:

1. στατιστικό τεστ της υπεργεωμετρικής κατανομής (hypergeometric test)
2. στατιστικό τεστ της διωνυμικής κατανομής (binomial test)
3. στατιστικό τεστ Fisher (Fisher's exact test)
4. στατιστικό τεστ K-S (Kolmogorov-Smirnov test)
5. στατιστικό τεστ της κατανομής χ^2 (Chi-squared test)

7.1.1 Ανάλυση «υπερ-εκπροσώπησης»

Η διαδικασία συνοψίζεται στα παρακάτω βήματα :

1. Δημιουργία της λίστας των ενδιαφερόμενων γονιδίων
2. Για κάθε σύνολο γονιδίων, δημιουργείται ένας δυο επί δυο πίνακας, συγκρίνοντας τον αριθμό των σημαντικών γονιδίων που είναι μέλη της κατηγορίας και αυτών που δεν είναι μέλη.
3. Η σημαντικότητα της υπερεκπροσώπησης μπορεί να μετρηθεί, χρησιμοποιώντας ένα στατιστικό τεστ

Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι αγνοεί όλα τα γονίδια που δεν ανήκουν στην λίστα των σημαντικών γονιδίων. Γι' αυτό και τα αποτελέσματα θα είναι άμεσα εξαρτώμενα από το επίπεδο σημαντικότητας που θα χρησιμοποιηθεί για τη σύνταξη αυτής της λίστας.

7.1.2 Ανάλυση «αθροιστικού σκορ»

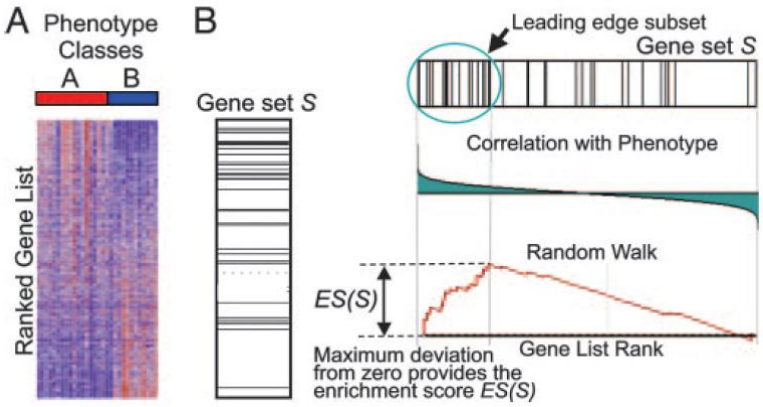
Η βασική ιδέα της ανάλυσης αυτής είναι να ορίζονται σκορ για κάθε σετ γονιδίων, που βασίζεται σε όλα τα συγκεκριμένα ανά γονίδιο σκορ για κάθε σετ.

Μία τέτοια μέθοδος είναι η GSEA που βασίζεται στο τεστ Kolmogorov-Smirnov (K-S test), το οποίο είναι γνωστό για την έλλειψη ευαισθησίας και της περιορισμένης πρακτικής χρήσης του. Γι' αυτό και ο Subramanian [52] έκανε μια κατά περίπτωση τροποποίηση του τεστ K-S, η οποία περιγράφεται παρακάτω. Το σετ των γονιδίων που θα αναλυθεί ορίζεται με βάση βιολογική γνώση, π.χ. δημοσιευμένες πληροφορίες σχετικά με βιοχημικά μονοπάτια (pathways).

Η GSEA αναλύει πειράματα με προφίλ έκφρασης από δείγματα που ανήκουν για παράδειγμα σε 2 κατηγορίες (δύο διαφορετικοί φαινότυποι που δηλώνονται με Α ή Β). Τα γονίδια μπορούν να διαταχθούν σε μια λίστα κατάταξης ***L***, με βάση την έκφρασή τους μεταξύ των δυο κατηγοριών. Ο στόχος είναι να εξαχθούν βιολογικά συμπεράσματα από αυτή τη λίστα.

Τα γονίδια κατατάσσονται βάσει του συσχετισμού τους μεταξύ της έκφρασης και της διάκρισης φαινοτυπικής κατηγορίας χρησιμοποιώντας κατάλληλο πρότυπο μέτρησης.

Ο στόχος του GSEA είναι να καθορίσει αν μέλη ενός συνόλου ***S*** γονιδίων (π.χ. γονίδια που κωδικοποιούν προϊόντα σε ένα μεταβολικό μονοπάτι, τα οποία ανήκουν στο ίδιο κυτταρογενετικό συγκρότημα ή μοιράζονται την ίδια GO κατηγορία) εμφανίζονται στην κορυφή (ή βάση) της λίστας L, περίπτωση κατά την οποία το σετ γονιδίων σχετίζεται με την διαφορά της φαινοτυπικής κατηγορίας. Η λίστα ***L*** αναφέρεται στην κατάταξη των γονιδίων όπως προκύπτει από τις τιμές του στατιστικού που έχει επιλεγχθεί (για παράδειγμα μέτρο συσχέτισης με τον φαινότυπο).



Σχήμα 7.1: Σύνοψη της μεθόδου GSEA **[52]**

Η μέθοδος (Σχήμα 7.1) απαιτεί ως είσοδο:

- ✓ Σετ δεδομένων έκφρασης ***D*** με ***N*** γονίδια και k δείγματα.
- ✓ Διαδικασία διάταξης για παραγωγή της λίστας γονιδίων ***L*** που αποτελείται από τα γονίδια ***N*** διατεταγμένα ως προς της σημαντικότητά σε σχέση με την Ho. Περιλαμβάνει μια συσχέτιση και ένα φαινότυπο ή ένα προφίλ ενδιαφέροντος C.
- ✓ Έναν εκθέτη ***p***, ώστε να καθορίζεται το βάρος του βήματος (για την GSEA p=1)
- ✓ Ανεξάρτητο σετ γονιδίων ***S*** που αποτελείται από N_H γονίδια (πχ. μεταβολικό μονοπάτι ή μια κατηγορία GO)

Τα τρία βασικά βήματα της μεθόδου GSEA είναι :

Βήμα 1^{ο}: Υπολογισμός του σκορ του εμπλουτισμού

Υπολογίζεται ένα σκορ εμπλουτισμού (enrichment score ***ES***) που αντικατοπτρίζει το βαθμό στον οποίο ένα σετ γονιδίων ***S***, υπερεκπροσωπείται στα άκρα (κορυφή ή βάση) της λίστα κατάταξης L.

Το σκορ υπολογίζεται ακολουθώντας τη λίστα από κάτω προς τα πάνω, αυξάνοντας ένα στατιστικό στοιχείο αθροίσματος όταν συναντάμε ένα γονίδιο που ανήκει στο ***S***, και μειώνοντας το όταν συναντάμε γονίδια εκτός του ***S***. Το μέγεθος της προσαύξησης εξαρτάται από την συσχέτιση του γονιδίου με το φαινότυπο. Το σκορ εμπλουτισμού είναι η μέγιστη απόκλιση από το μηδέν κατά τον τυχαίο περίπατο και αντιστοιχεί σε ένα σταθμισμένο Kolmogorov-Smirnov στατιστικό.

Η διαδικασία περιγράφεται μαθηματικά παρακάτω:

1. Διάταξη των διαθέσιμων ***N*** γονιδίων, για να μορφοποιηθεί η λίστα L={g_1,g_2…..g_N}, με βάση την συσχέτισή τους, r(g_j)=r_j, με το προφίλ ενδιαφέροντος C.
2. Υπολογισμός του μέρους των γονιδίων που ανήκουν στο ***S*** (επιτυχίες) σταθμισμένο με την συσχέτιση τους, και το κλάσμα των γονιδίων που δεν ανήκουν στο ***S*** (αποτυχίες), μέχρι μια δεδομένη θέση i, στη λίστα L.

$$P_{επιτυχία}(S,i) = \sum_{g \in S} \frac{|r_j|^p}{N_R} \tag{7.5}$$

όπου, $N_R = \sum_{g \in S} |r_j|^p$

$$P_{αποτυχία}(S,i) = \sum_{g \in S} \frac{1}{N - N_H} \tag{7.6}$$

Το σκορ εμπλουτισμού ***ES*** είναι η μέγιστη απόκλιση από το μηδέν της διαφοράς $P_{επιτυχία} - P_{αποτυχία}$. Για ένα τυχαίο S, ES(S), η διαφορά θα είναι σχετικά μικρή, αλλά αν συγκεντρώνεται στην κορυφή ή στη βάση της λίστας, το ES(S) θα είναι αντίστοιχα μεγάλο.

Βήμα 2^o :Εκτίμηση της σημαντικότητας του επιπέδου του σκορ εμπλουτισμού

Η στατιστική σημαντικότητα (ονομαστική τιμή ***p***/nominal **p *value***) των σκορ ***ES*** εκτιμάται σε σχέση με ένα ***ES*** σκορ η τιμή του οποίου προκύπτει από τυχαία κατανομή φαινοτύπων (μηδενική κατανομή). Συγκεκριμένα, αναδιατάσσονται οι φαινοτυπικές ετικέτες και επαναλαμβάνεται ο υπολογισμός του ***ES*** του σετ των γονιδίων για τα αναδιαταγμένα δεδομένα, που παράγει μια μηδενική κατανομή για το ***ES***. Η εμπειρική ονομαστική τιμή ***p***, του παρατηρούμενου ***ES***, υπολογίζεται με βάση την μηδενική κατανομή. Είναι σημαντικό, η αναδιάταξη των επισημασμένων κατηγοριών να διατηρεί μια, γονίδιο με γονίδιο, συσχέτιση και ως εκ τούτου, να παρέχει μια πιο σωστή βιολογική αξιολόγηση της σημαντικότητας.

Βήμα 3^{ο} : Ρύθμιση για πολλαπλούς ελέγχους υποθέσεων

Όταν μια ολόκληρη βάση γονιδίων αξιολογείται, προσαρμόζεται το υπολογισμένο επίπεδο σημαντικότητας, ώστε να πραγματοποιηθούν οι πολλαπλοί έλεγχοι υποθέσεων. Αρχικά, ομαλοποιείται το ***ES*** για κάθε σετ γονιδίων, ώστε να αντιπροσωπεύει το μέγεθος του συνόλου, αποδίδοντας ένα ομαλοποιημένο σκορ εμπλουτισμού (Normalized Enrichment Score ή NES). Στη συνέχεια ελέγχεται το ποσοστό των ψευδών αρνητικών περιπτώσεων υπολογίζοντας την τιμή του FDR για κάθε NES. Το FDR μας δείχνει την εκτιμώμενη πιθανότητα που ένα σύνολο με δεδομένο NES, αντιπροσωπεύει μια ψευδώς αρνητική περίπτωση.

Η διαδικασία των πολλαπλών ελέγχων υπόθεσης περιλαμβάνει :

1. Καθορισμός ***ES(S)*** για κάθε σετ γονιδίων στη βάση δεδομένων
2. Για κάθε σετ γονιδίων S και τις 1000 αναδιατάξεις π των φαινοτυπικών ετικετών, αναδιατάσσονται τα γονίδια στην ***L*** και καθορίζεται η κατανομή ***ES(S,π).***
3. Ρύθμιση της διακύμανσης για διαφορετικού μεγέθους σετ γονιδίων. Κανονικοποιείται η κατανομή ***ES(S,π)*** και η παρατηρούμενη τιμή του ***ES(S)***, επανακλιμακώνοντας χωριστά τα θετικά και αρνητικά σκορ, διαιρώντας με την μέση τιμή του ES(S,π), για να προκύψουν τα κανονικοποιημένα σκορ NES(S,π) και NES(S).
4. Υπολογισμός FDR. Ελέγχεται ο λόγος των ψευδών αρνητικών αποτελεσμάτων (σφάλμα τύπου I, όπως ορίστηκε στο Κεφάλαιο 5.1) προς το συνολικό αριθμό των σετ γονιδίων, επιτυγχάνοντας ένα σταθερό επίπεδο σημαντικότητας ξεχωριστά για θετικές/ αρνητικές τιμές των NES(S,π) και NES(S).

Δημιουργείται ένα ιστόγραμμα όλων των NES(S,π) τιμών για τα S και π. Χρησιμοποιείται αυτή η μηδενική κατανομή για να υπολογιστεί μια FDR ***q*** τιμή για ένα δεδομένο $NES(S) = NES^* \geq 0$. Το FDR είναι ο λόγος του ποσοστού όλων των (S,π) με τιμή $NES(S,\pi) \geq 0$, των οποίων $NES(S,\pi) \geq NES^*$, διαιρεμένο με το ποσοστό των παρατηρούμενων S με $NES(S) \geq 0$, των οποίων $NES(S) \geq NES^*$ και αντίστοιχα εάν $NES(S) = NES^* \leq 0$. [52]

Η τιμή q υπολογίζεται από τον παρακάτω τύπο:

$$q = \frac{|\{(S,\pi)|NES(S,\pi) \geq \alpha\}|/|\{(S,\pi)|NES(S,\pi) \geq 0\}|}{|\{S|NES(S) \geq \alpha\}|/|\{S|NES(S) \geq 0\}|} \quad (7.8)$$

7.2 Γονιδιακή Οντολογία

Η βιβλιοθήκη "clusterProfiler" [53] του Bioconductor πραγματοποιεί τη διαδικασία της ταξινόμησης των γονιδίων με βάση τις βιολογικές λειτουργίες τους και την ανάλυση εμπλουτισμού για σύνολα γονιδίων αλλά και για ομαδοποιήσεις αυτών. Για αυτές τις λειτουργίες, η βιβλιοθήκη "clusterProfiler" βασίζεται στα μεταφραστικά πακέτα του Bioconductor "GO.db" [39] και "KEGG.db" [54] ώστε να ανακτήσει δεδομένα για τους χάρτες των γονιδίων της Γονιδιακής Οντολογίας και της KYOTO εγκυκλοπαίδειας γονιδίων και γονιδιωμάτων (Kyoto Encyclopedia of genes and Genomes ή KEGG). Το μεταφραστικό πακέτο που χρησιμοποιείται για την μετάφραση των ταυτοτήτων των ανθρώπινων γονιδίων είναι το "org.Hs.eg.db" που αναλύεται στο υποκεφάλαιο 4.3.

Η βιβλιοθήκη "clusterProfiler" πραγματοποιεί την ανάλυση εμπλουτισμού για την Γονιδιακή Οντολογία και την KYOTO εγκυκλοπαίδεια των γονιδίων και των γονιδιωμάτων με υπεργεωμετρικό τεστ (hypergeometric test). Η clusterProfiler προσαρμόζει το εκτιμώμενο επίπεδο σημαντικότητας για το τεστ της πολλαπλής υπόθεσης καθώς και τις τιμές ***q*** (***q-values***) για τον έλεγχο FDR. Τα αποτελέσματα των αναλύσεων παρουσιάζονται σε διαγράμματα. Επίσης παρέχεται τη δυνατότητα της σύγκρισης των λειτουργικών προφίλ μεταξύ ομαδοποιημένων γονιδίων. Τέλος παρέχει σύγκριση βιολογικών περιεχομένων των βιοχημικών μονοπατιών της GO και της KEGG. [53]

7.2.1 Ταξινόμηση των γονιδίων στις λειτουργικές κατηγορίες τους

Επιστρέφοντας στην ανάλυση των πρωτεϊνικών δεδομένων, ταξινομούμε τα γονίδια στις βιολογικές λειτουργικές κατηγορίες ανά οντολογία, με βάση την κατανομή τους σε ένα συγκεκριμένο επίπεδο της GO κατανομής των σχέσεων μεταξύ των GO όρων. Εδώ το επίπεδο που επιλέγεται είναι το τρίτο, δηλαδή το συντομότερο μονοπάτι

μεταξύ ενός GO όρου (κόμβου) και του κορυφαίου GO όρου (κόμβου) στην κατανομή των όρων (DAG γράφημα) θα περιλαμβάνει δυο μόνο συνδέσεις μεταξύ GO όρων.

Τα γονίδια τα οποία θα χρησιμοποιηθούν για την ταξινόμηση αυτή, είναι οι 42 Entrez ταυτότητες των γονιδίων που προκύπτουν από τη μετάφραση, μέσω της βιβλιοθήκης "org.Hs.eg.db", των 76 συμβολικών ονομασιών του Συνόλου Β. Για την ταξινόμηση αυτή απαιτούνται:

- ✓ Entrez ταυτότητες γονιδίων σε μορφή διανύσματος
- ✓ Ορισμός του οργανισμού, που για την ανάλυση αυτή, είναι ο άνθρωπος
- ✓ Ορισμός της οντολογίας
- ✓ Το επίπεδο της οντολογίας, όπως περιγράφεται στο υποκεφάλαιο 6.1.3., που εδώ θα είναι το τρίτο.

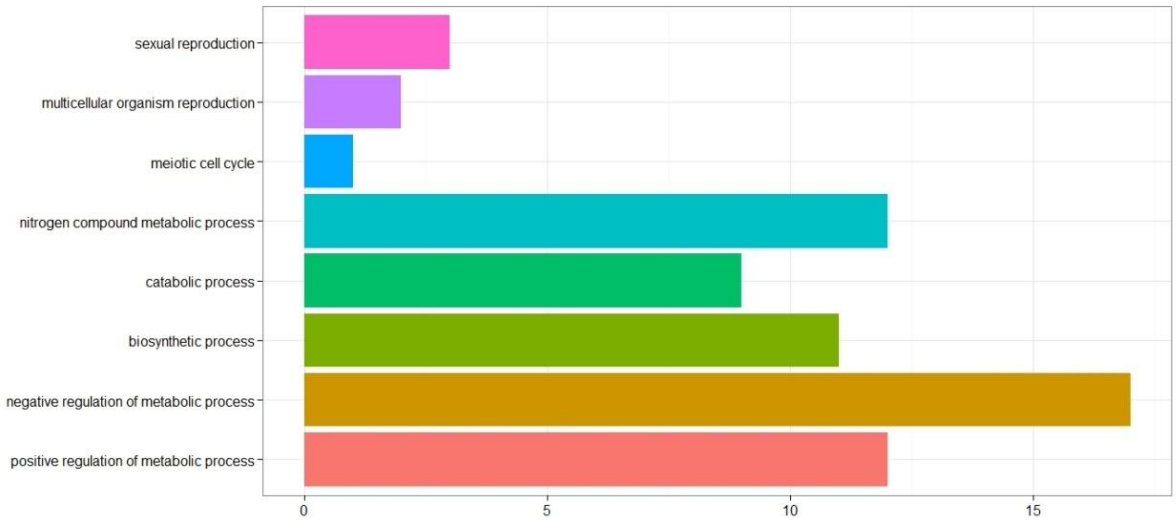
Κατά τη λειτουργία της ταξινόμησης της βιβλιοθήκης αυτής, αντιστοιχίζονται οι Entrez ταυτότητες των γονιδίων στους GO όρους, ο καθένας των οποίων διαθέτει την περιγραφή του (Description) σε κάθε οντολογία. Έπειτα μετρώνται πόσες Entrez ταυτότητες αντιστοιχίζονται στον κάθε GO όρο (***Count***) και έτσι προκύπτει η αναλογία αντιστοίχισης των γονιδίων στο σύνολο του γονιδιακού σετ της ανάλυσης (***GeneRatio***). Τέλος η λειτουργία αυτή αντιστοιχίζει τους GO όρους στις συμβολικές ονομασίες των γονιδίων (geneID). Ένα μέρος της μορφής του αποτελέσματος παρουσιάζεται στο Σχήμα 7.2, στο περιβάλλον της R. Για παράδειγμα στην πρώτη γραμμή βλέπουμε τον όρο GO:0019953 με την αντίστοιχη περιγραφή *sexual reproduction* να εμφανίζεται 3 φορές στις μεταφράσεις των ταυτοτήτων των γονιδίων με αναλογία 3/41 στα εξής γονίδια "APOB/ SEPP1/ CALR".

```
> head(summary(ggo))
                   ID                            Description Count GeneRatio                                                    geneID
GO:0019953 GO:0019953                   sexual reproduction     3      3/41                                            APOB/SEPP1/CALR
GO:0019954 GO:0019954                  asexual reproduction     0      0/41
GO:0032504 GO:0032504        multicellular organism reproduction     2      2/41                                                 APOB/CALR
GO:0032505 GO:0032505 reproduction of a single-celled organism     0      0/41
GO:0051321 GO:0051321                      meiotic cell cycle     1      1/41                                                      CALR
GO:0006807 GO:0006807        nitrogen compound metabolic process    12     12/41 AMBP/ITIH2/ITIH1/ITIH3/LUM/APOC1/HRG/APOE/ITIH4/CALR/APOA1/APOA4
```

Σχήμα 7.2: Αποτέλεσμα ταξινόμησης γονιδίων Συνόλου Β στο περιβάλλον της R

Τα αποτελέσματα ανά οντολογία παρουσιάζονται στα παρακάτω ραβδογράμματα. Σε αυτά παρουσιάζονται μόνο οι οκτώ πρώτες κατηγορίες ταξινόμησης. Στον άξονα **y**

του διαγράμματος εμφανίζονται οι βιολογικές λειτουργικές κατηγορίες ενώ στο άξονα **x** ο αριθμός των Entrez ταυτοτήτων που αντιστοιχίζονται σε αυτές.



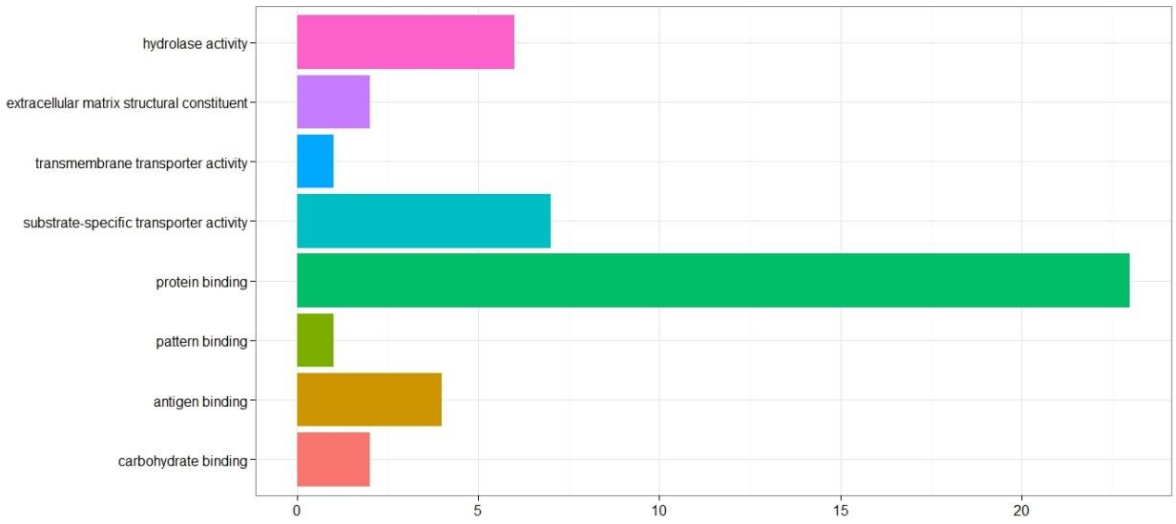
Σχήμα 7.3: Ραβδόγραμμα ταξινόμησης γονιδίων Συνόλου B σε οντολογία BP

Οι λειτουργικές κατηγορίες της οντολογίας BP, στις οποίες ταξινομούνται τα περισσότερα γονίδια της ανάλυσης, η αναλογία των γονιδίων που αντιστοιχίζονται σε κάθε κατηγορία (***GeneRatio***), καθώς και η περιγραφή των κατηγοριών παρατίθενται στον Πίνακα 7.2. Οι περιγραφές των λειτουργικών κατηγοριών διατίθενται από την διαδικτυακή εφαρμογή "AmiGO" [55], η οποία παρέχει βάσεις δεδομένων για την Γονιδιακή Οντολογία.

Πίνακας 7.2: Ταξινόμηση γονιδίων στις λειτουργικές κατηγορίες της οντολογίας BP

GOID	Λειτουργική κατηγορία	GeneRatio
GO:0044238	πρωτογενής μεταβολική διεργασία (primary metabolic process)	33/41
GO:0071704	μεταβολική διεργασία οργανικής ουσίας (organic substance metabolic process)	33/41
GO:0050789	ρύθμιση βιολογικής διεργασίας	31/41
GO:0006950	απόκριση στο στρες (response to stress)	28/41
GO:0050794	ρύθμιση κυτταρικής διεργασίας (regulation of cellular process)	27/41

Αντίστοιχα για τις λειτουργικές κατηγορίες της οντολογίας MF το αποτέλεσμα παρουσιάζεται στο Σχήμα 7.4 και στον Πίνακα 7.3.

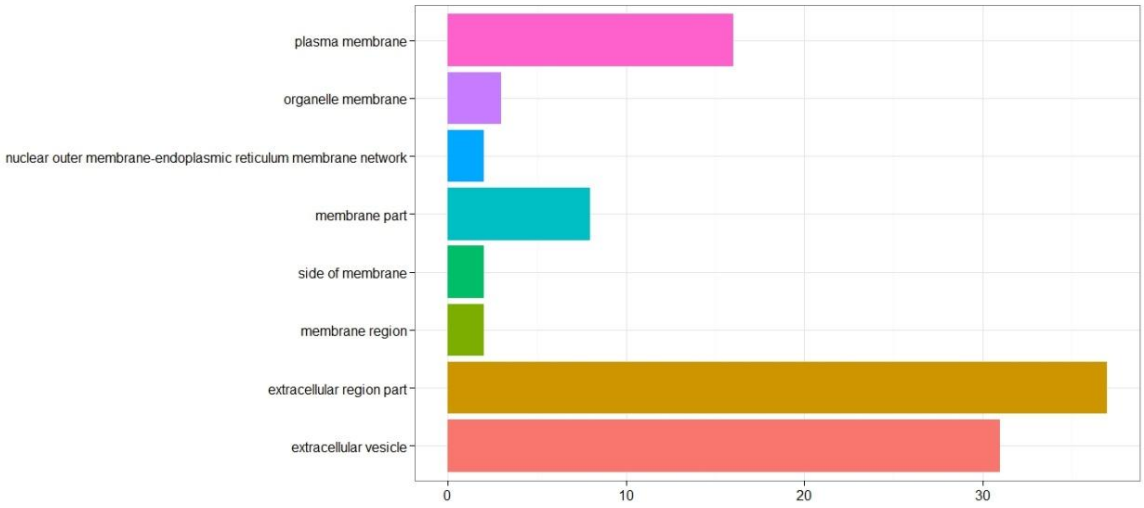


Σχήμα 7.4: Ραβδόγραμμα ταξινόμησης γονιδίων Συνόλου Β σε οντολογία MF

Πίνακας 7.3: Ταξινόμηση γονιδίων στις λειτουργικές κατηγορίες της οντολογίας MF

GOID	Λειτουργική κατηγορία	GeneRatio
GO:0005515	δέσμευση πρωτεΐνης (protein binding)	23/41
GO:0043167	δέσμευση ιόντος (ion binding)	18/41
GO:0030234	ρυθμιστής δραστηριότητας ενζύμου	11/41
GO:0008289	δέσμευση λιπιδίου (lipid binding)	11/41

Ομοίως σε οντολογία CC το αποτέλεσμα παρουσιάζεται στο Σχήμα 7.5 και στον Πίνακα 7.4.



Σχήμα 7.5: Ραβδόγραμμα ταξινόμησης γονιδίων Συνόλου B σε οντολογία CC

Πίνακας 7.4: Ταξινόμηση γονιδίων στις λειτουργικές κατηγορίες της οντολογίας CC

GOID	Λειτουργική κατηγορία	GeneRatio
GO:0044421	τμήμα εξωκυττάριας περιοχής	37/41
GO:0005615	εξωκυττάριο τμήμα	37/41
GO:0043227	οργανίδιο που οριοθετείται από μεμβράνη (membrane-bounded organelle)	32/41
GO:0043230	οργανίδιο εξωκυττάριου χώρου (extracellular organelle)	31/41
GO:1903561	κυστίδιο εξωκυττάριου χώρου (extracellular vesicle)	31/41

7.2.2 Ανάλυση εμπλουτισμού της GO

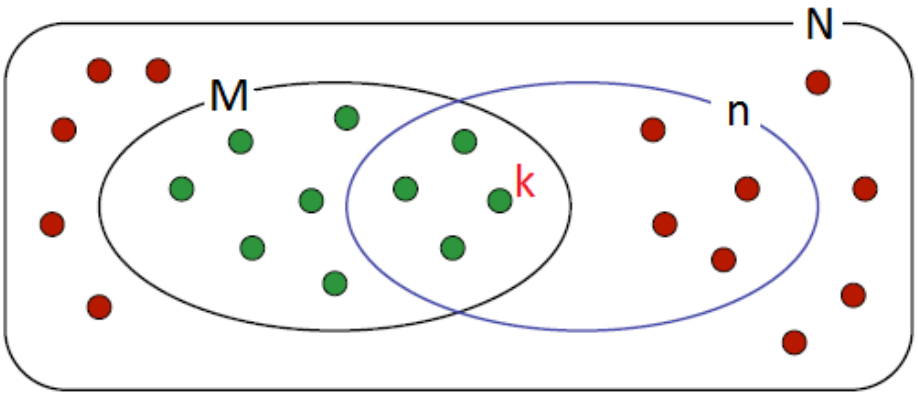
Το τεστ υπερ-εκπροσώπησης είναι μια ευρέως χρησιμοποιούμενη προσέγγιση για τον εντοπισμό των βιολογικών πληροφοριών. Εδώ χρησιμοποιείται το υπεργεωμετρικό μοντέλο για να εκτιμηθεί κατά πόσο ο αριθμός των επιλεγμένων γονιδίων που τυχόν σχετίζονται με την τοξικότητα, λόγω της εισόδου του νανοσωματιδίου στον οργανισμό, είναι μεγαλύτερος απ’ ότι αναμενόταν.

Για να καθοριστεί αν οποιοιδήποτε όροι μεταφράζουν μια συγκεκριμένη λίστα γονιδίων σε συχνότητα μεγαλύτερη από την αναμενόμενη, η βιβλιοθήκη “clusterProfiler”, υπολογίζει μια τιμή **p (p-value)** χρησιμοποιώντας την υπεργεωμετρική κατανομή:

(7.9)

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-1}}{\binom{N}{n}}$$

Σε αυτή την εξίσωση, **N** είναι ο συνολικός αριθμός γονιδίων της κατανομής του «υπόβαθρου» της βάσης δεδομένων, **M** είναι ο αριθμός των γονιδίων που μεταφράζονται στους επιθυμητούς GO όρους. **n** είναι το μέγεθος της λίστας των γονιδίων με το μεγαλύτερο ενδιαφέρον και **k** είναι ο αριθμός των γονιδίων της λίστα αυτή που αντιστοιχούν στους επιθυμητούς GO όρους. Η κατανομή του «υποβάθρου» είναι ως προεπιλογή όλα τα γονίδια, τα οποία έχουν μετάφραση, και ορίζονται μέσω μιας παραμέτρου. Οι τιμές **p** (**p-values**) προσαρμόζονται για πολλαπλή σύγκριση με τη μέθοδο BH (Benjamini & Hochberg) [56] και οι τιμές **q (q-values**) υπολογίζονται για τον FDR έλεγχο.



Σχήμα 7.6: Απεικόνιση του αποτελέσματος που δίνει το υπεργεωμετρικό μοντέλο σε μορφή συνόλων

Με βάση το Σχήμα 7.6 και την εξίσωση της τιμής p, συμπεραίνεται ότι όσο μεγαλύτερος είναι ο αριθμός των γονιδίων ***k***, δηλαδή αυτών που εκφράζονται σε βαθμό μεγαλύτερο, από ότι αναμενόταν σε ένα φυσιολογικό περιβάλλον του οργανισμού, τόσο μικρότερη θα είναι η τιμή ***p***.

Για την ανάλυση αυτή απαιτούνται:

- ✓ Entrez ταυτότητες των γονιδίων σε μορφή διανύσματος
- ✓ Τα γονίδια του «υποβάθρου», τα οποία φορτώνονται μέσω της βάσης δεδομένων της βιβλιοθήκης “org.Hs.eg.db”, είναι όλα τα γονίδια που έχουν GO μεταφράσεις.
- ✓ Ορισμός του οργανισμού, που για την ανάλυση αυτή, είναι ο άνθρωπος

- ✓ Ορισμός της οντολογίας BP ή MF ή CC
- ✓ Μέθοδος προσαρμογής των τιμών **p ,** η οποία εδώ είναι η FDR μέθοδος
- ✓ Όριο αποκοπής για τις τιμές **p** το 0.01 (**p-value**≤0.01)
- ✓ Όριο αποκοπής για τις τιμές **q** το 0.05.(**q-value**≤0.05)

Οι 42 Entrez ταυτότητες που θα χρησιμοποιηθούν θα προκύψουν από την μετάφραση των 76 συμβολικών ονομασιών των γονιδίων του Συνόλου Β.

Το όριο αποκοπής για τις τιμές **p**, χρησιμοποιείται για να ορίσει το αποτέλεσμα το οποίο βασίζεται στις τιμές **p** που υπολογίζονται, καθώς και για τις τιμές **p** οι οποίες προσαρμόζονται. Ενώ το όριο αποκοπής για τις τιμές **q** χρησιμοποιείται ώστε να ελέγξει τις τιμές **q**. [53]

Μέσω της ανάλυσης αυτής αντιστοιχίζονται οι Entrez ταυτότητες της ανάλυσης στις ταυτότητες των GO όρων (ID), οι οποίοι διαθέτουν ο καθένας μια περιγραφή (Description). Υπολογίζονται οι ταυτότητες Entrez οι οποίες αντιστοιχούν στους GO όρους (**Count**) και η αναλογία αντιστοίχισης τους στο σύνολο των γονιδίων της ανάλυσης (**GeneRatio**). Αντίστοιχα υπολογίζεται η αναλογία αντιστοίχισης των γονιδίων του υποβάθρου στους GO όρους (**BgRatio**). Η τιμή **p-value** προκύπτει με τη χρήση του υπεργεωμετρικού μοντέλου με βάση τις δυο παραπάνω αντιστοιχίσεις. Κατόπιν μέσω του ελέγχου FDR υπολογίζονται οι τιμές **p-adjust** και **q-value** για να επαληθευθεί η εγκυρότητα του αποτελέσματος. Το αποτέλεσμα περιορίζεται με βάση τις τιμές για το όρια αποκοπής που ορίστηκαν για την τιμή **p-value** και την τιμή **q-value**. Ένα μέρος της μορφής του αποτελέσματος παρουσιάζεται στο Σχήμα 7.7 στο περιβάλλον της R. Το συνολικό αποτέλεσμα για κάθε οντολογία παρατίθεται στους Πίνακες ΠΒ1, ΠΒ2 και ΠΒ3.

Με βάση το υπεργεωμετρικό μοντέλο και το όριο αποκοπής για την **p-value** που έχει τεθεί ίσο με 0,01, συμπεραίνεται ότι όσο μικρότερη υπολογιζόμενη **p-value** έχει ο κάθε GO όρος τόσο υψηλότερης τάξης εμπλουτισμού θα έχει. Άρα όσα γονίδια αντιστοιχίζονται στους GO όρους με τις χαμηλότερες τιμές **p-value**, θα υπερεκφράζονται πέρα από το σύνηθες στο περιβάλλον του οργανισμού. Εδώ οι GO όροι με τιμή p-value>0,01 θα αποκλείονται από το αποτέλεσμα.

```
> head(summary(ego))
                 ID                                          Description GeneRatio   BgRatio       pvalue     p.adjust       qvalue
GO:0072376 GO:0072376                               protein activation cascade      9/40  67/11978 7.321702e-13 1.771852e-10 7.552914e-11
GO:0006958 GO:0006958                   complement activation, classical pathway      7/40  31/11978 6.602894e-12 7.989502e-10 3.405703e-10
GO:0006897 GO:0006897                                              endocytosis     15/40 484/11978 1.599051e-11 1.021980e-09 4.356417e-10
GO:0034377 GO:0034377                         plasma lipoprotein particle assembly      6/40  18/11978 1.689223e-11 1.021980e-09 4.356417e-10
GO:0065005 GO:0065005                              protein-lipid complex assembly      6/40  19/11978 2.462853e-11 1.192021e-09 5.081254e-10
GO:0002455 GO:0002455 humoral immune response mediated by circulating immunoglobulin      7/40  42/11978 6.596520e-11 1.807872e-09 7.706456e-10
                                                                     geneID Count
GO:0072376                              CRP/IGHG1/IGKC/KLKB1/KNG1/C1QA/C1R/C4BPA/C4BPB     9
GO:0006958                                        CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB     7
GO:0006897 PRG4/CRP/AMBP/APOA1/APOB/APOC1/APOE/IGHG1/IGKC/LBP/C4BPA/C4BPB/CALR/APOL1/CD5L    15
GO:0034377                                          APOA1/APOA4/APOB/APOC1/APOE/APOM     6
GO:0065005                                          APOA1/APOA4/APOB/APOC1/APOE/APOM     6
GO:0002455                                        CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB     7
```

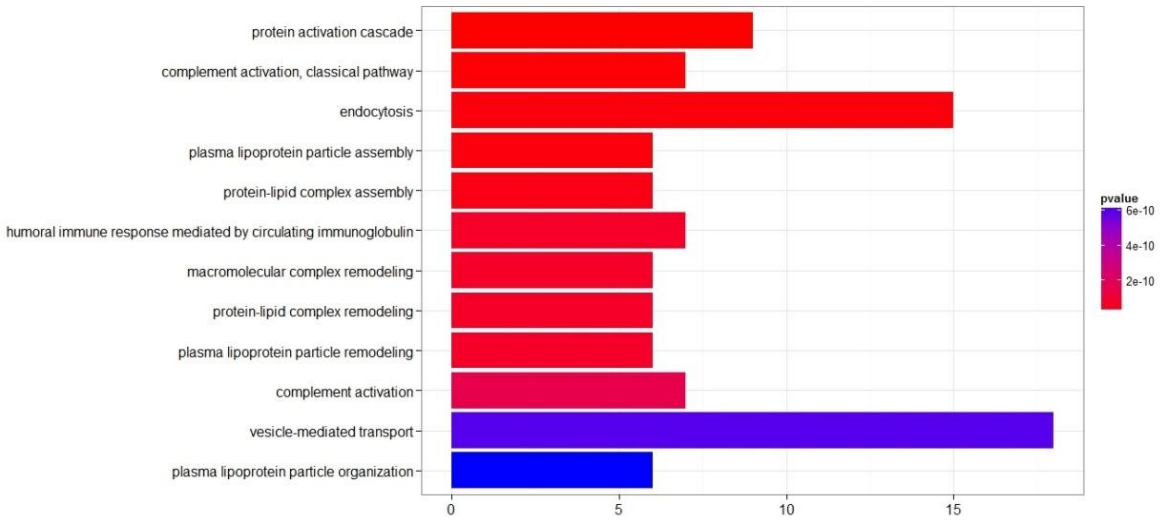
Σχήμα 7.7: Αποτέλεσμα ανάλυσης εμπλουτισμού των γονιδίων του Συνόλου Β στο περιβάλλον της R

Στο Σχήμα 7.7, παρατηρούμε στην πρώτη γραμμή ότι η ταυτότητα GO:0072376 με περιγραφή "protein activation cascade" εμφανίζει ***GeneRatio*** 9/40 και ***BgRatio*** 67/11978 με ***pvalue*** $7.32 \cdot e^{-13}$ με ***padjust*** $1.77 \cdot e^{-10}$ με ***qvalue*** $7.55 \cdot e^{-11}$. Οι αντιστοιχιζόμενες γονιδιακές ταυτότητες (SYMBOL) είναι CRP/ IGHG1/ IGKC/ KLKB1/ KNG1/ C1QA/ C1R/ C4BPA/ C4BPB με ***Count*** ίσο με 9.

Τα αποτελέσματα της παραπάνω ανάλυσης ανά οντολογία παρουσιάζονται σε:

- Ραβδογράμματα ταξινόμησης των γονιδίων στις εμπλουτισμένες λειτουργικές κατηγορίες με κλίμακα της τιμής **p**
- Χάρτης σχέσης μεταξύ των εμπλουτισμένων λειτουργικών κατηγοριών
- Χάρτης σχέσης γονιδίων με τις εμπλουτισμένες λειτουργικές κατηγορίες.

Στα ραβδογράμματα στον άξονα **x** είναι ο αριθμός των γονιδίων που αντιστοιχίζονται σε κάθε λειτουργική κατηγορία, ενώ στον άξονα **y** είναι τα ονόματα των λειτουργικών κατηγοριών. Ο χρωματισμός των ράβδων βασίζεται στις τιμές των **p** με βάση και την κλίμακα δεξιά του διαγράμματος. Η βαθμίδα των χρωμάτων είναι μεταξύ του κόκκινου και του μπλε, αυξανόμενης της τιμής του **p**. Επομένως το κόκκινο χρώμα δηλώνει χαμηλές τιμές του **p** (υψηλής τάξης εμπλουτισμό), και το μπλε δηλώνει υψηλές τιμές **p** (χαμηλής τάξης εμπλουτισμό).



Σχήμα 7.8: Ραβδόγραμμα αποτελέσματος ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία BP

Στο παραπάνω ραβδόγραμμα (Σχήμα 7.8) εμφανίζονται μόνο οι 12 πιο εμπλουτισμένες λειτουργικές κατηγορίες που περιγράφουν τα γονίδια σε οντολογία BP. Η κλίμακα της τιμής **p**, προσαρμόζεται στο διάστημα των τιμών **p** που λαμβάνουν αυτές οι δώδεκα κατηγορίες μόνο.

Στον Πίνακα 7.5 θα παρουσιαστούν οι πιο εμπλουτισμένες λειτουργικές κατηγορίες σε οντολογία BP, οι περιγραφές τους και οι υπολογιζόμενες τιμές ***p-value*** από το υπεργεωμετρικό μοντέλο.

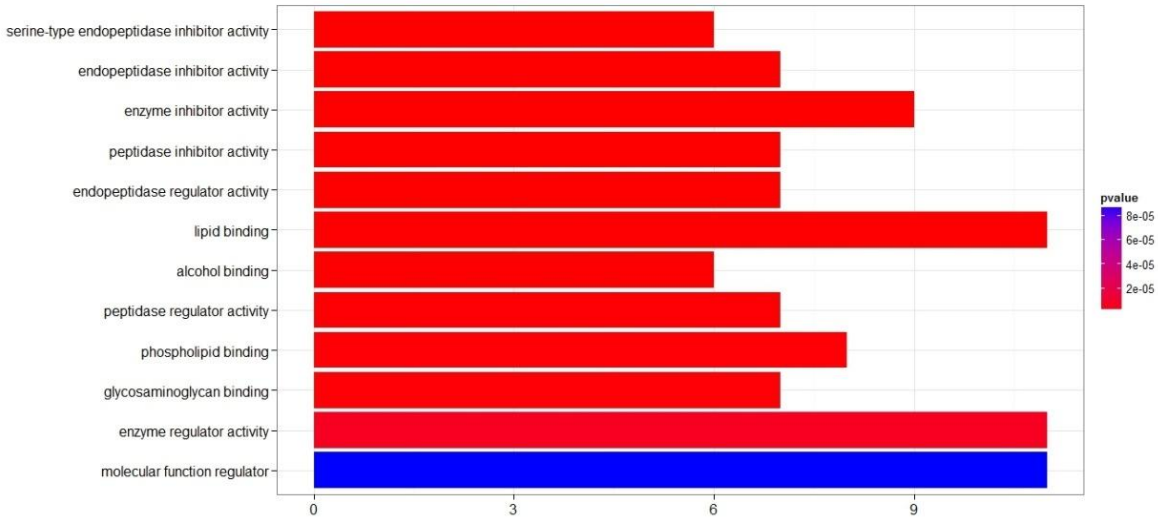
Πίνακας 7.5: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία BP

GOID	Λειτουργική κατηγορία	p-value
GO:0072376	ενεργοποίηση αλληλουχίας αντιδράσεων πρωτεΐνης (protein activation cascade)	$7{,}32 \cdot 10^{-13}$
GO:0006958	ενεργοποίηση συμπληρώματος, κλασσικό μονοπάτι (complement activation, classical pathway)	$6{,}60 \cdot 10^{-12}$
GO:0006897	ενδοκυττάρωση (endocytosis)	$1{,}6 \cdot 10^{-11}$
GO:0034377	συγκρότηση σωματιδίων λιποπρωτεΐνης πλάσματος (plasma lipoprotein particle assembly)	$1{,}7 \cdot 10^{-11}$
GO:0065005	συγκρότηση συμπλόκου πρωτεϊνών και λιπιδίων (protein-lipid complex assembly)	$2{,}46 \cdot 10^{-11}$
GO:0002455	χυμική ανοσολογική απόκριση που προκαλείται από	$6.60 \cdot 10^{-11}$

	κυκλοφορούντα μόρια ανοσοσφαιρίνης (humoral immune response mediated by circulating immunoglobulin)	
GO:0034367	αναδιαμόρφωση συμπλόκου μακρομορίων (macromolecular complex remodeling)	$6.72 \cdot 10^{-11}$
GO:0034368	αναδιαμόρφωση συμπλόκου πρωτεϊνών-λιπιδίων (protein-lipid complex remodeling)	$6.72 \cdot 10^{-11}$
GO:0034369	αναδιαμόρφωση σωματιδίου λιποπρωτεΐνης πλάσματος (plasma lipoprotein particle remodeling)	$6.72 \cdot 10^{-11}$
GO:0006956	ενεργοποίηση του συμπληρώματος (complement activation)	$1.52 \cdot 10^{-10}$

Το ***GeneRatio*** εδώ είναι αναλογία στα 40 αντί στα 42 που είναι ο αριθμός των γονιδίων, διότι δύο Entrez ταυτότητες των γονιδίων δεν αντιστοιχίζονται σε GO όρους της οντολογίας BP, επομένως αποκλείονται από την ανάλυση.

Αντίστοιχα σε οντολογία MF παρατίθεται το Σχήμα 7.9 και ο Πίνακας 7.6.



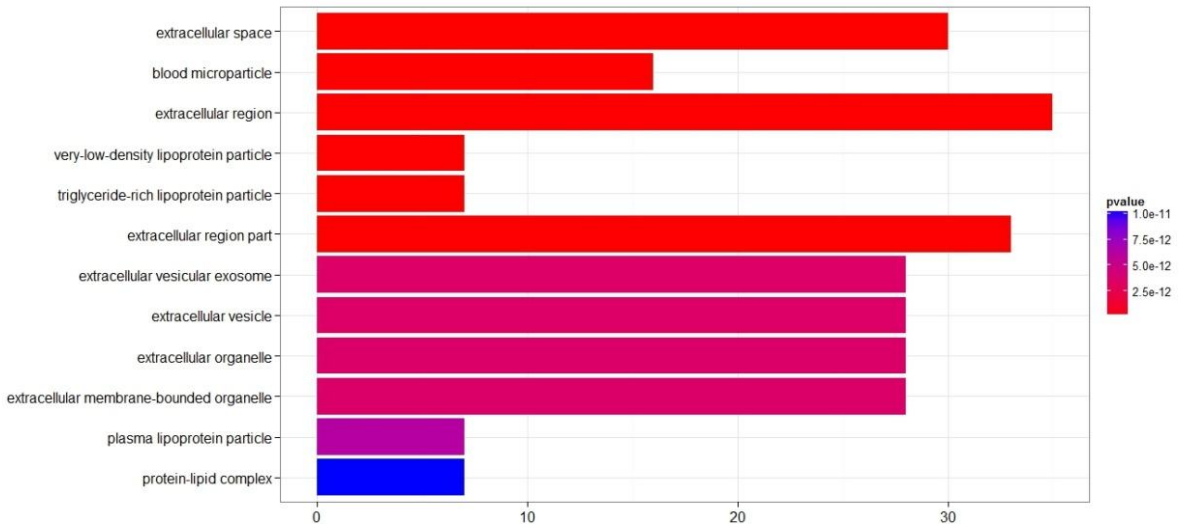
Σχήμα 7.9: Ραβδόγραμμα αποτελέσματος ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία MF

Πίνακας 7.6: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία MF

GOID	Λειτουργική κατηγορία	p-value
GO:0004867	αναστολέας δραστηριότητας ενδοπεπτιδάσης τύπου σερίνης (serine-type endopeptidase inhibitor activity)	$5{,}21 \cdot 10^{-8}$
GO:0004866	αναστολέας δραστηριότητας ενδοπεπτιδάσης (endopeptidase inhibitor activity)	$1{,}06 \cdot 10^{-7}$
GO:0004857	αναστολέας δραστηριότητας ενζύμου (enzyme inhibitor activity)	$1{,}12 \cdot 10^{-7}$
GO:0030414	αναστολέας δραστηριότητας πεπτιδάσης (peptidase inhibitor activity)	$1.19 \cdot 10^{-7}$
GO:0061135	ρυθμιστής δραστηριότητας ενδοπεπτιδάσης (endopeptidase regulator activity)	$1.33 \cdot 10^{-7}$
GO:0008289	δέσμευση λιπιδίου	$1.72 \cdot 10^{-7}$
GO:0043178	δέσμευση αλκοόλης (alcohol binding)	$2.33 \cdot 10^{-7}$
GO:0061134	ρυθμιστής δραστηριότητας πεπτιδάσης (peptidase regulator activity)	$6.02 \cdot 10^{-7}$
GO:0005543	δέσμευση φωσφολιπιδίου (phospholipid binding)	$7.81 \cdot 10^{-7}$
GO:0005539	δέσμευση γλυκοσαμινογλυκάνης (glycosaminoglycan binding)	$1.52 \cdot 10^{-7}$

Το GeneRatio εδώ είναι αναλογία στα 39 αντί στα 42 που είναι ο αριθμός των γονιδίων, διότι τρεις Entrez ταυτότητες των γονιδίων δεν αντιστοιχίζονται σε GO όρους της οντολογίας MF, επομένως αποκλείονται από την ανάλυση.

Ομοίως σε οντολογία CC το αποτέλεσμα της ανάλυσης εμπλουτισμού παρατίθεται στο Σχήμα 7.10 και στον Πίνακα 7.7.



Σχήμα 7.10: Ραβδόγραμμα αποτελέσματος ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία CC

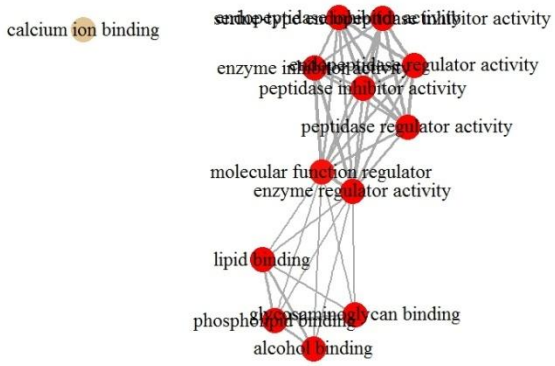
Πίνακας 7.7: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία CC

GOID	Λειτουργική κατηγορία	p-value
GO:0005615	εξωκυττάριος χώρος (extracellular space)	$1.51 \cdot 10^{-24}$
GO:0072562	μικροσωματίδιο αίματος (blood microparticle)	$9.78 \cdot 10^{-24}$
GO:0005576	εξωκυττάρια περιοχή	$1.23 \cdot 10^{-14}$
GO:0034361	σωματίδιο χαμηλής πυκνότητας λιποπρωτεΐνης (very-low-density lipoprotein particle)	$3.59 \cdot 10^{-14}$
GO:0034385	σωματίδιο λιποπρωτεΐνης πλούσιο σε τριγλυκερίδια (triglyceride-rich lipoprotein particle)	$3.59 \cdot 10^{-14}$
GO:0044421	τμήμα εξωκυττάριας περιοχής	$3{,}6 \cdot 10^{-12}$
GO:0070062	εξωκυτταρικό φυσαλιδώδες εξώσωμα (extracellular vesicular exosome)	$3.6 \cdot 10^{-12}$
GO:1903561	κυστίδιο εξωκυττάριου χώρου	$3.59 \cdot 10^{-12}$
GO:0043230	οργανίδιο εξωκυττάριου χώρου	$3.63 \cdot 10^{-12}$
GO:0065010	εξωκυττάριου χώρου οργανίδιο που οριοθετείται από μεμβράνη (extracellular membrane-bounded organelle)	$3.63 \cdot 10^{-12}$

Το ***GeneRatio*** εδώ είναι αναλογία στα 41 αντί στα 42 που είναι ο αριθμός των γονιδίων, διότι μία Entrez ταυτότητα των γονιδίων δεν αντιστοιχίζεται σε GO όρους της οντολογίας CC, επομένως αποκλείεται από την ανάλυση.

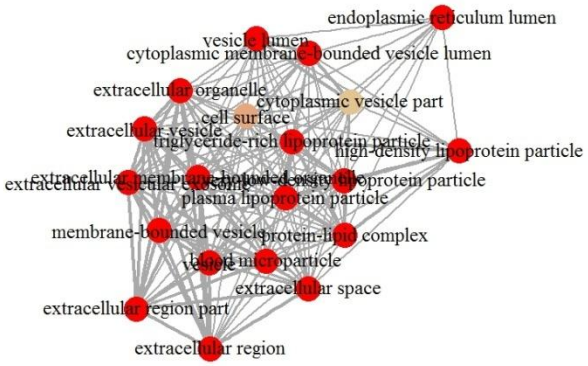
Σχήμα 7.11: Χάρτης σχέσης εμπλουτισμένων λειτουργικών κατηγοριών σε οντολογία BP

Στο χάρτη της ανάλυσης εμπλουτισμού (Σχήμα 7.11), παρατηρείται ότι όλες οι εμπλουτισμένες λειτουργικές κατηγορίες σε οντολογία BP, έχουν σχέση μεταξύ τους .



Σχήμα 7.12: Χάρτης σχέσης εμπλουτισμένων λειτουργικών κατηγοριών σε οντολογία MF

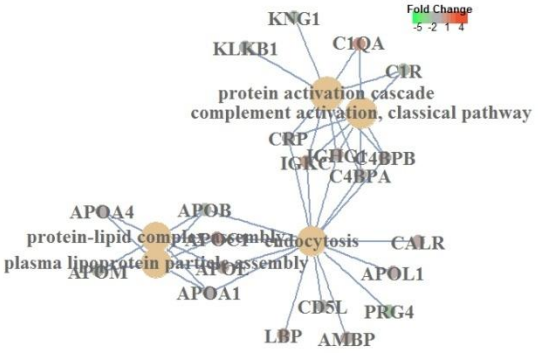
Στον χάρτη που απεικονίζεται στο Σχήμα 7.12, όλες οι εμπλουτισμένες λειτουργικές κατηγορίες σε οντολογία MF, έχουν σχέση μεταξύ τους, εκτός από την κατηγορία *calcium ion binding*, η οποία περιγράφεται ως επιλεκτική αλληλεπίδραση και μη ομοιοπολική με τα ιόντα ασβεστίου (Ca^{+2}).



Σχήμα 7.13: Χάρτης σχέσης εμπλουτισμένων λειτουργικών κατηγοριών σε οντολογία CC

Αντίστοιχα και εδώ (Σχήμα 7.13), παρατηρείται ότι όλες οι εμπλουτισμένες λειτουργικές κατηγορίες σε οντολογία CC, έχουν σχέση μεταξύ τους.

Στους χάρτες των Σχημάτων 7.14-7.16 παρουσιάζονται οι σχέσεις γονιδίων με τις πέντε λειτουργικές υψηλής τάξης εμπλουτισμού κατηγορίες ανά οντολογία. Στους χάρτες αυτούς απεικονίζεται γραφικά ότι τα ίδια γονίδια μπορούν να ανήκουν σε περισσότερες από μια λειτουργικές κατηγορίες.

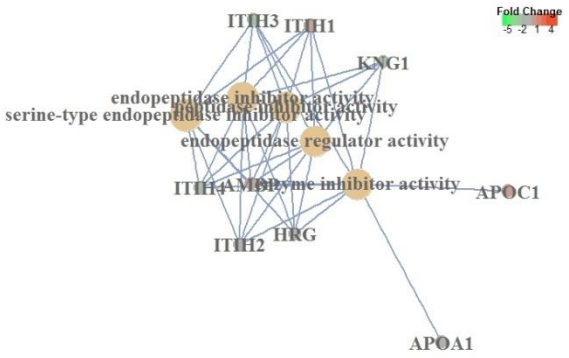


Σχήμα 7.14 Χάρτης σχέσης γονιδίων-λειτουργιών σε οντολογία BP

Πίνακας 7.8 Πίνακας αντιστοίχισης γονιδίων- βιολογικών διεργασιών

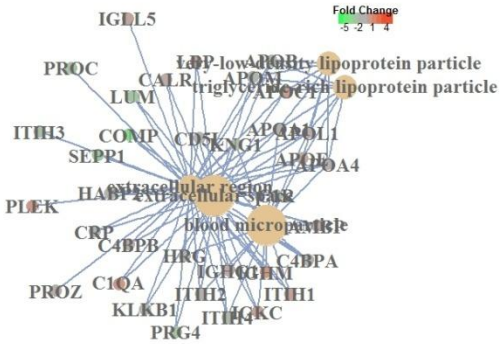
Βιολογική Λειτουργία	Γονίδια (SYMBOL)
ενεργοποίηση αλληλουχίας αντιδράσεων πρωτεΐνης	KLKB1;KNG1;C1QA;C1R;CRP;IGKC;IGHG1; C4BP4;C4BPB
ενεργοποίηση συμπληρώματος, κλασσικό μονοπάτι	C1QA;C1R;C4BPB;C4BPA;CRP;IGKC;IGHG1;KMG1
ενδοκυττάρωση	LPB;APOL1;CD5L;CALR;IGHG1;C4BPB;IGKC;C4BP4; CRP;APOB4;APOE;APOA1;APOC1;AMBP;PRG4
συγκρότηση σωματιδίων λιποπρωτεΐνης πλάσματος	APOA4;APOM;APOB;APOE;APOA1;APOC1
συγκρότηση συμπλόκου πρωτεϊνών και λιπιδίων	APOA4;APOM; APOB;APOE;APOA1;APOC1

Στον Πίνακα 7.8 παρουσιάζονται οι σχέσεις γονιδίων, με χρήση των συμβολικών τους ονομασιών, ως αποτέλεσμα μελέτης του χάρτη του Σχήματος 7.14. Λαμβάνοντας τις τιμές για το $\boldsymbol{Q^2}$ από τον Πίνακα 2.2 για κάθε γονίδιο μπορούμε να δούμε σε τι εύρος τιμών $\boldsymbol{Q^2}$ κυμαίνονται τα σύνολα γονιδίων που ανήκουν στην κάθε εμπλουτισμένη βιολογική λειτουργία. Τα γονίδια που σχετίζονται με την ενεργοποίηση αλληλουχίας αντιδράσεων πρωτεΐνης διαθέτουν τιμές στο εύρος [0.7187,0.8051], με την ενεργοποίηση του συμπληρώματος στο [0.7544,0.81], με την ενδοκυττάρωση στο [0.6684,1], και τέλος με την συγκρότηση πλάσματος από σωματίδια λιποπρωτεΐνης και την συγκρότηση συμπλόκου πρωτεϊνών και λιπιδίων, στις οποίες συμμετέχουν τα ίδια γονίδια, οι τιμές $\boldsymbol{Q^2}$ έχουν τιμές στο [0.7158,0.8189].



Σχήμα 7.15: Χάρτης σχέσης γονιδίων-λειτουργιών σε οντολογία MF

Στο χάρτη του Σχήματος 7.15 παρατηρείται ότι όλα τα γονίδια συμμετέχουν και στις πέντε εμπλουτισμένες βιοχημικές δραστηριότητες της οντολογίας MF με εξαίρεση τα γονίδια APOC1 και APOA1 που συμμετέχουν μόνο στην δραστηριότητα του αναστολέα ενζύμων. Στα γονίδια περιλαμβάνονται δύο από τα σημαντικότερα γονίδια της ανάλυσης ITH1 και AMBP (Q^2=1) ενώ οι τιμές Q^2 των υπόλοιπων γονιδίων κυμαίνονται στο διάστημα [0.6527,0.7666].



Σχήμα 7.16: Χάρτης σχέσης γονιδίων-λειτουργιών σε οντολογία CC

Στον χάρτη σε οντολογία CC (Σχήμα 7.16) παρατηρούμε ότι τα περισσότερα γονίδια εντοπίζονται στις θέσεις του εξωκυττάριου χώρου, μέρος αυτών σε μικροσωματίδια αίματος, με τα σημαντικότερα γονίδια AMBP, HABP2, ITIH1, ITH2 να περιλαμβάνονται και στις δύο θέσεις. Στα σωματίδια λιποπρωτεΐνης συμμετέχουν γονίδια που σχετίζονται και με τις θέσεις του εξωκυττάριου χώρου εκτός από τα APOM και APOL1.

7.2.3 Σύγκριση βιολογικού περιεχομένου σε ομαδοποιημένα γονίδια

Παρέχεται μέσω της βιβλιοθήκης μια λειτουργία με την οποία υπολογίζονται αυτόματα οι εμπλουτισμένες λειτουργικές κατηγορίες κάθε ομαδοποίησης γονιδίων. Έτσι μπορούμε να δούμε ποιες ομάδες γονιδίων που έχουν δημιουργηθεί με βάση τη λειτουργική τους ομοιότητα, εμφανίζονται σε ποιες εμπλουτισμένες λειτουργικές κατηγορίες.

Οι απαιτήσεις της λειτουργίας αυτής είναι:

- ✓ Ομαδοποίηση των γονιδίων σε μορφή λίστας
- ✓ Ορισμός του οργανισμού, που για την ανάλυση αυτή, είναι ο άνθρωπος

- ✓ Μέθοδος ανάλυσης εμπλουτισμού με χρήση της βιβλιοθήκης “GO.db” ή “KEGG.db”
- ✓ Ορισμός της οντολογίας

Επίσης με χρήση της μεθόδου ταξινόμησης των γονιδίων με βάση της κατανομής τους σε ένα συγκεκριμένο επίπεδο της GO δομής, μπορούν να συγκριθούν όσα αφορά την ταξινόμηση αυτή παραπάνω από μια ομαδοποιήσεις των γονιδίων.

Η λειτουργία της ταξινόμησης απαιτεί:

- ✓ Entrez ταυτότητες γονιδίων σε μορφή διανύσματος
- ✓ Ομάδες της πρώτης ομαδοποίησης σε μορφή διανύσματος
- ✓ Ομάδες της δεύτερης ομαδοποίηση σε μορφή διανύσματος
- ✓ Μέθοδος ταξινόμησης γονιδίων
- ✓ Ορισμός της οντολογίας

7.2.3.1 Ομαδοποιημένα λειτουργικά όμοια γονίδια

Έτσι μέσω της πρώτης λειτουργίας με χρήση της μεθόδου ανάλυσης εμπλουτισμού για την Γονιδιακή Οντολογία, εντοπίζονται ποιες ομάδες των γονιδίων που δημιουργήθηκαν με βάση την λειτουργικής τους ομοιότητα, εμφανίζονται σε ποιες εμπλουτισμένες λειτουργικές κατηγορίες. Η σύγκριση αυτή θα γίνει και για τις τρεις οντολογίες.

Οι Entrez ταυτότητες γονιδίων που θα χρησιμοποιηθούν θα προκύψουν από την μετάφραση των 76 συμβολικών ονομασιών τα γονιδίων του Συνόλου B με χρήση της βιβλιοθήκης “org.Hs.eg.db”. Με την μετάφραση προκύπτουν 41 Entrez ταυτότητες γονιδίων καθώς κάποιες συμβολικές ονομασίες δεν έχουν αντιστοίχιση με Entrez ταυτότητα στην βάση δεδομένων της μετάφρασης. Η ομαδοποίηση των γονιδίων που χρησιμοποιείται είναι αυτή που δημιουργήθηκε με χρήση λειτουργίας της βιβλιοθήκης “GOSim” στο υποκεφάλαιο 6.3.4.2.

Εισάγοντας την λίστα των ομαδοποιημένων γονιδίων, πραγματοποιείται η επιθυμητή σύγκριση. Η ανάλυση και εδώ θα γίνει με χρήση του υπεργεωμετρικού μοντέλου, γι’ αυτό και το αποτέλεσμα θα είναι της ίδια μορφής με την ανάλυση εμπλουτισμού που πραγματοποιήθηκε σε προηγούμενο κεφάλαιο. Η διαφορά εδώ είναι ότι τα γονίδια χωρίζονται σε ομάδες και η ανάλυση εμπλουτισμού πραγματοποιείται για κάθε ομάδα

ξεχωριστά, ώστε να μπορεί να γίνει η σύγκριση του βιολογικού περιεχομένου της κάθε ομάδας.

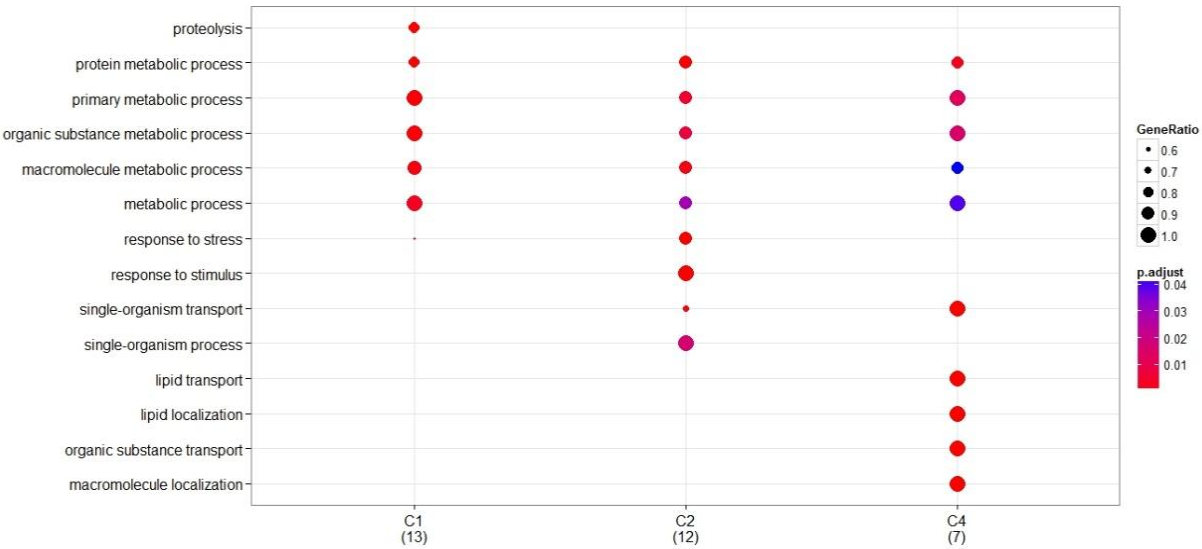
Στο Σχήμα 7.17 παρουσιάζεται η αρχή του αποτελέσματος της ανάλυσης εμπλουτισμού των γονιδίων στο περιβάλλον της R όπου στην πρώτη γραμμή για την ομάδα Cluster C1 η ταυτότητα (ID) GO:0006508 με περιγραφή “proteolysis” έχει ***GeneRatio*** 11/13 και ***BgRatio*** 1421/18585 λαμβάνει ***pvalue*** $3.4 \cdot e^{-11}$ και προσαρμοσμένη τιμή ***padjust*** $1.7 \cdot e^{-09}$ και με ***qvalue*** $3.94 \cdot e^{-10}$ με αντιστοιχιζόμενες geneID (EntrezID) 259/3026/3273/3697/3699/3700/3818/5624/84735/8858 με ***Count*** 11. Το συνολικά αποτελέσματα για κάθε οντολογία περιλαμβάνονται στους Πίνακες ΠΒ4, ΠΒ5 και ΠΒ6.

```
> head(summary(ckBP))
  Cluster         ID                                 Description GeneRatio    BgRatio       pvalue     p.adjust       qvalue
1      C1 GO:0006508                                 proteolysis     11/13 1421/18585 3.401286e-11 1.700643e-09 3.938331e-10
2      C1 GO:0070613              regulation of protein processing      7/13  370/18585 1.813752e-09 4.534379e-08 1.050067e-08
3      C1 GO:0010951 negative regulation of endopeptidase activity      6/13  225/18585 4.709118e-09 7.257181e-08 1.680610e-08
4      C1 GO:0010466     negative regulation of peptidase activity      6/13  233/18585 5.805745e-09 7.257181e-08 1.680610e-08
5      C1 GO:0016485                          protein processing      7/13  557/18585 3.069685e-08 3.069685e-07 7.108743e-08
6      C1 GO:0045861           negative regulation of proteolysis      6/13  319/18585 3.783661e-08 3.153050e-07 7.301801e-08
                                        geneID Count
1 259/3026/3273/3697/3698/3699/3700/3818/5624/84735/8858    11
2                     259/3273/3697/3698/3699/3700/3818     7
3                          259/3273/3697/3698/3699/3700     6
4                          259/3273/3697/3698/3699/3700     6
5                     259/3273/3697/3698/3699/3700/3818     7
6                          259/3273/3697/3698/3699/3700     6
```

Σχήμα 7.17 Αποτέλεσμα ανάλυσης εμπλουτισμού ανά ομάδα γονιδίων Συνόλου Β στο περιβάλλον της R

Τα αποτελέσματα της ανάλυσης παρουσιάζονται στα παρακάτω γραφήματα ανά οντολογία. Στον άξονα **y** των γραφημάτων αυτών, βρίσκονται οι εμπλουτισμένες λειτουργικές κατηγορίες. Στον άξονα **x** βρίσκονται οι ομάδες κατά σειρά C1,C2,…Cn, όπου n είναι ο αριθμός των ομάδων, για τις οποίες γίνεται η σύγκριση της συμμετοχής τους στις εμπλουτισμένες λειτουργικές κατηγορίες. Μέσα σε παρένθεση δίπλα σε κάθε ομάδα, είναι ο αριθμός των γονιδίων από τα οποία αποτελείται. Αν κάποια ομάδα δεν εμφανίζεται στον άξονα **x**, τότε συμπεραίνεται ότι τα γονίδια αυτής της ομάδας δεν συμμετέχουν στις εμπλουτισμένες λειτουργικές κατηγορίες. Δεξιά του γραφήματος παρουσιάζεται το υπόμνημα των δυο κλιμάκων του γραφήματος. Η πρώτη είναι το ποσοστό επί τις εκατό (%) της αναλογίας αντιστοίχισης των γονιδίων της κάθε ομάδας στις κατηγορίες των λειτουργιών (**GeneRatio**) . Η χρωματική κλίμακα δηλώνει το μέγεθος της προσαρμοσμένης τιμής **p**. Η βαθμίδα των χρωμάτων είναι μεταξύ του κόκκινου και του μπλε, αυξανόμενης της τιμής του **p**. Επομένως το κόκκινο χρώμα δηλώνει χαμηλές τιμές του **p** (υψηλής

τάξης εμπλουτισμού), και το μπλε δηλώνει υψηλές τιμές **p** (χαμηλής τάξης εμπλουτισμό).

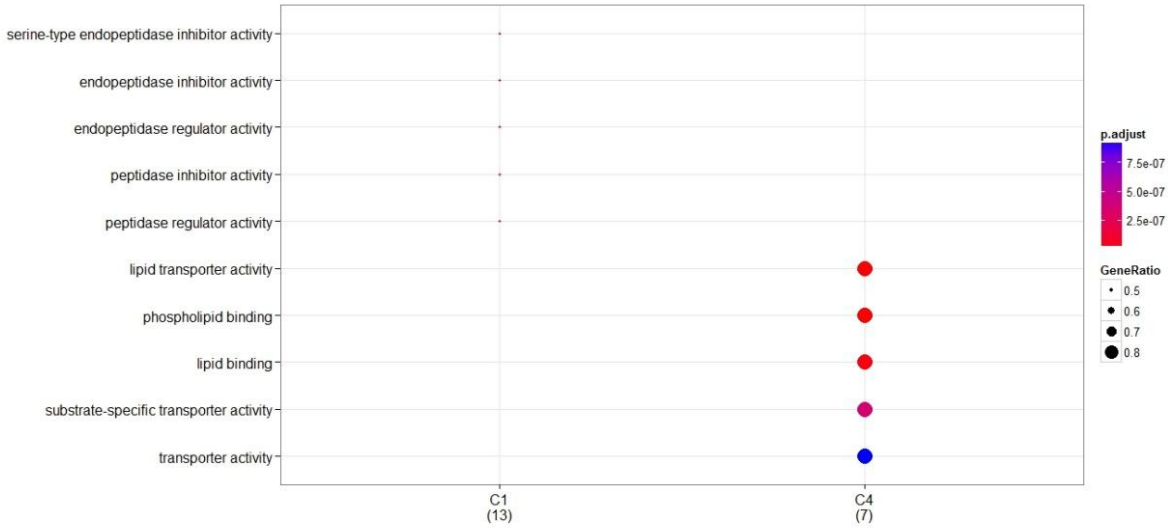


Σχήμα 7.18: Συγκριτικό γράφημα εμπλουτισμένου βιολογικού περιεχομένου μεταξύ των C1, C2 και C4 σε οντολογία BP

Σε οντολογία BP γίνεται η σύγκριση μεταξύ των C1, C2 και C4, διότι αυτές οι ομάδες μόνο περιέχουν γονίδια τα οποία αντιστοιχίζονται στις εμπλουτισμένες βιολογικές λειτουργίες.

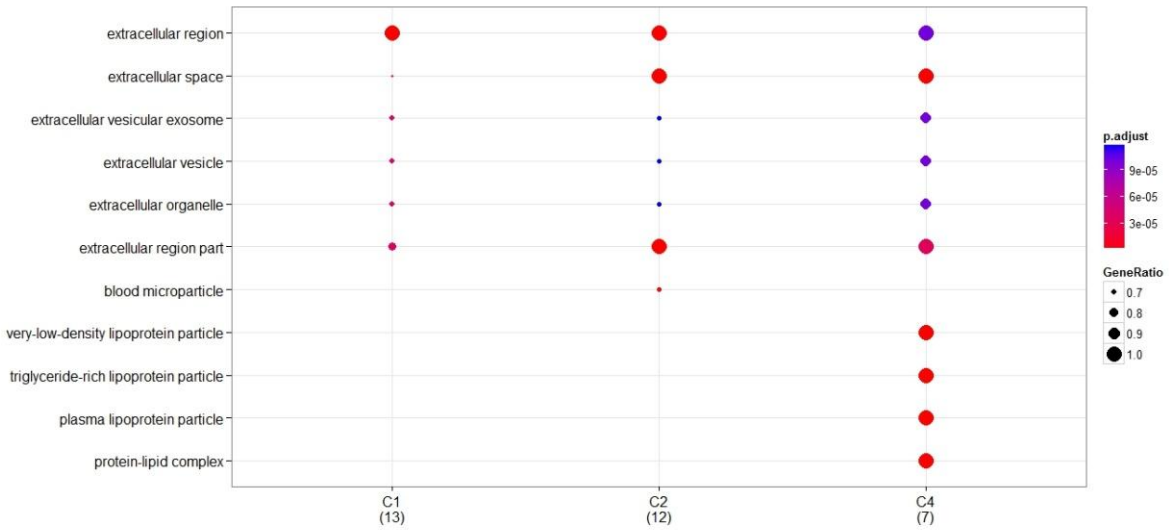
Από το γράφημα του Σχήματος 7.18, συμπεραίνεται ότι τα γονίδια όλων των ομάδων (Cluster) ανήκουν στις βιολογικές λειτουργικές κατηγορίες που αφορούν τις μεταβολικές διεργασίες, δηλαδή τις χημικές αντιδράσεις που λαμβάνουν χώρα κατά τη διάρκεια αναβολικών ή καταβολικών μονοπατιών. Συγκεκριμένα τα γονίδια της ομάδας C1 εμπλέκονται με υψηλής τάξης εμπλουτισμό στις μεταβολικές διεργασίες σε σχέση με τις άλλες ομάδες. Αντίστοιχα τα γονίδια της ομάδας C2 συμμετέχουν στις λειτουργίες, απόκριση σε ερέθισμα και απόκριση στο στρες, που αφορούν διαδικασίες που καταλήγουν σε μια αλλαγή στην κατάσταση ή στην δραστηριότητα ενός κυττάρου (σε όρους κίνησης, έκκρισης, παραγωγής ενζύμου, γονιδιακής έκφρασης) ως αποτέλεσμα ενός ερεθίσματος ή μιας διαταραχής στην κυτταρική ομοιόσταση ή εξωγενών παραγόντων όπως η θερμοκρασία ή η ακτινοβολία. Τέλος τα γονίδια της ομάδας C4 συμμετέχουν με υψηλής τάξης εμπλουτισμό σε κατηγορίες όπως η μεταφορά λιπιδίων (lipid transport) και ο εντοπισμός θέσης λιπιδίων (lipid localization), οι οποίες αφορούν την μεταφορά λιπιδίων ή τη διατήρηση τους σε μια

συγκεκριμένη θέση. Εδώ επιβεβαιώνεται ότι τα γονίδια της ομάδας C4 είναι αυτά που μετέχουν στις λειτουργίες που έχουν άμεση σχέση με την τοξικότητα. [57]



Σχήμα 7.19: Συγκριτικό γράφημα εμπλουτισμένου βιολογικού περιεχομένου μεταξύ των C1 και C4 σε οντολογία MF

Στο γράφημα του Σχήματος 7.19, παρατηρείται ότι η σύγκριση γίνεται μόνο μεταξύ των ομάδων C1 και C4, επειδή οι ομάδες C2, C3, C5 δεν περιέχουν γονίδια τα οποία να διαθέτουν εμπλουτισμένες λειτουργικές κατηγορίες στην οντολογία MF. Τα γονίδια της ομάδας C1 ανήκουν με υψηλής τάξης εμπλουτισμό σε μικρή αναλογία (**GeneRatio**) σε λειτουργίες όπως του ρυθμιστή δραστηριότητας πεπτιδάσης και του αναστολέα δραστηριότητας πεπτιδάσης, οι οποίες αντίστοιχα περιγράφουν διαδικασίες ρύθμισης της δραστηριότητας των πεπτιδάσων και αποτροπής ή μείωσης της δραστηριότητας τους. Οι πεπτιδάσες είναι ένζυμα που καταλύουν την υδρόλυση πεπτιδικών δεσμών. Αντίστοιχα τα γονίδια της ομάδας C4, συμμετέχουν σε λειτουργίες όπως η δραστηριότητα μεταφορέα λιπιδίων (lipid transporter activity) και η δέσμευση φωσφολιπιδίων οι οποίες αντίστοιχα περιγράφουν διαδικασίες ενεργοποίησης της κατευθυνόμενης κίνηση των λιπιδίων έξω ή μέσα στο κύτταρο ή μεταξύ κυττάρων και επιλεκτικής, μη ομοιοπολικής αλληλεπίδραση με φωσφολιπίδια. Και εδώ τα γονίδια της ομάδας C4 συμμετέχουν σε λειτουργίες που σχετίζονται με την τοξικότητα. [58]



Σχήμα 7.20: Συγκριτικό γράφημα εμπλουτισμένου βιολογικού περιεχομένου μεταξύ των C1 ,C2 και C4 σε οντολογία CC

Σε οντολογία CC (Σχήμα 7.20) τα γονίδια όλων των ομάδων εμφανίζονται με υψηλής τάξης εμπλουτισμό στις θέσεις εξωκυττάρια περιοχή και εξωκυττάριος χώρος, οι οποίες αναφέρονται στο χώρο έξω από την μεμβράνη του πλάσματος του κυττάρου. Ενώ παρατηρείται ότι μόνο τα γονίδια της ομάδας C4 δραστηριοποιούνται στις θέσεις σωματιδίων λιποπρωτεΐνης χαμηλής πυκνότητας, πλούσια σε τριγλυκερίδια, και του πλάσματος, τα οποία είναι τα σφαιρικά σωματίδια, που αποτελούνται από τριγλυκερίδια και μεταφέρουν λιπίδια. Ο ρόλος της μεμβράνης για το κύτταρο είναι πολύ καθοριστικός διότι αποτελεί τη δίοδο των διαφόρων συστατικών που εντοπίζονται στον εξωκυττάριο χώρο. Οι λιποπρωτεΐνης πλάσματος είναι αυτές που μπορούν να μεταφέρουν συστατικά από έξω, μέσα στο κύτταρο.

7.2.3.2 Ομαδοποιημένα γονίδια με κριτήριο το Q^2 και VIP

Μέσω της δεύτερης λειτουργίας με χρήση της μεθόδου ταξινόμησης των γονιδίων με βάση την κατανομή τους σε ένα επίπεδο της GO δομής, εμφανίζονται ποιες ομάδες γονιδίων εμφανίζονται σε κάθε κατηγορία των λειτουργιών. Εδώ θα χρησιμοποιηθούν δυο ομαδοποιήσεις γονιδίων για τις οποίες θα πραγματοποιηθεί η σύγκριση. Η λειτουργία αυτή θα πραγματοποιηθεί και για τις τρεις οντολογίες.

Η πρώτη ομαδοποίηση που θα χρησιμοποιηθεί σε όλες τις περιπτώσεις είναι η παραπάνω ομαδοποίηση των γονιδίων σε πέντε ομάδες με βάση την λειτουργικής

τους ομοιότητα. Η δεύτερη ομαδοποίηση θα αλλάζει για κάθε σύγκριση και θα είναι κατά σειρά οι εξής:

- ✓ Ομαδοποίηση των γονιδίων σε πέντε ομάδες με βάση την ταξινομημένη με ελλατούμενη σημαντικότητα, λίστα που παρέχετε από την αρχική εργασία.
- ✓ Ομαδοποίηση των γονιδίων σε πέντε ομάδες με βάση την τιμή της μεταβλητής σημαντικότητας για προβολή (Variable Importance to the Projection- **VIP**)

Η ομαδοποίηση των γονιδίων με μορφή Entrez ταυτοτήτων σε πέντε ομάδες (groups) με βάση την ταξινομημένη λίστα με ελλατούμενη σημαντικότητα θα είναι :

gA : 259, 3026, 3698, 3697

gB : 3699, 4060, 5624

gC : 10216, 84735, 8858, 4064, 1401, 5341, 338, 3818, 341, 6414, 3273, 348, 3514

gD : 3503, 8542, 3827, 55937, 3507, 922, 3700, 811

gE : 346, 6291, 335, 712, 722, 725, 3929, 3500, 1311, 715, 1, 337.

Η ομαδοποίηση των γονιδίων με μορφή Entrez ταυτοτήτων σε πέντε ομάδες με βάση την τιμή της μεταβλητής σημαντικότητας για προβολή (Variable Importance to the Projection- **VIP**) θα είναι :

gA : 259, 3026, 3698, 3697

gB : 3699, 4060

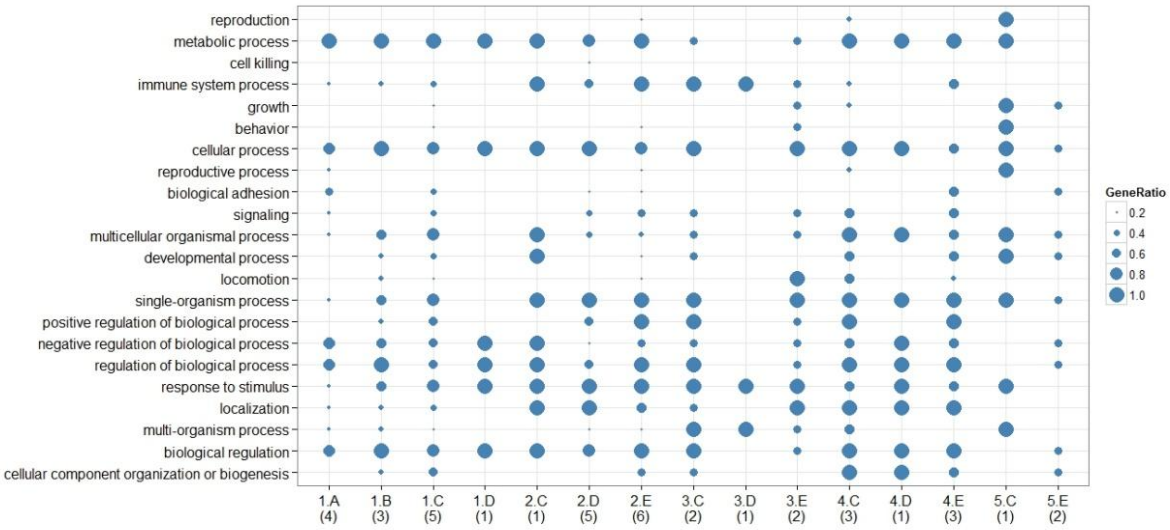
gC : 84735, 4064, 338, 341, 6414, 3503

gD : 5624, 10216, 8858, 1401, 5341, 3818, 3273, 348, 8542, 3827, 55937, 3507, 922, 3700, 811, 346, 335, 1, 337

gE : 3514, 6291, 712, 722, 725, 3929, 3500, 1311, 715

Η ταξινόμηση των ομαδοποιημένων γονιδίων και σε αυτή την ανάλυση θα γίνει με τον τρόπο που έγινε η ταξινόμηση των γονιδίων στο κεφάλαιο GOSim. Η μοναδική διαφορά εδώ είναι ότι η ταξινόμηση των γονιδίων θα γίνει για κάθε υποομάδα ξεχωριστά ώστε να πραγματοποιηθεί η σύγκριση του βιολογικού περιεχομένου της κάθε υποομάδας. Οι υποομάδες που θα δημιουργηθούν θα αποτελούνται από την τομή του συνόλου των γονιδίων της κάθε υποομάδας της δεύτερης ομαδοποίησης με το σύνολο των γονιδίων της κάθε υποομάδας της πρώτης ομαδοποίησης. Έτσι δημιουργούνται 15 νέες υποομάδες για τις οποίες πραγματοποιείται η σύγκριση.

Το παρακάτω γράφημα παρουσιάζει την σύγκριση της ομαδοποίησης με πέντε ομάδες με την ομαδοποίηση των γονιδίων σε πέντε ομάδες με βάση την ταξινομημένη λίστα με ελλατούμενη σημαντικότητα.



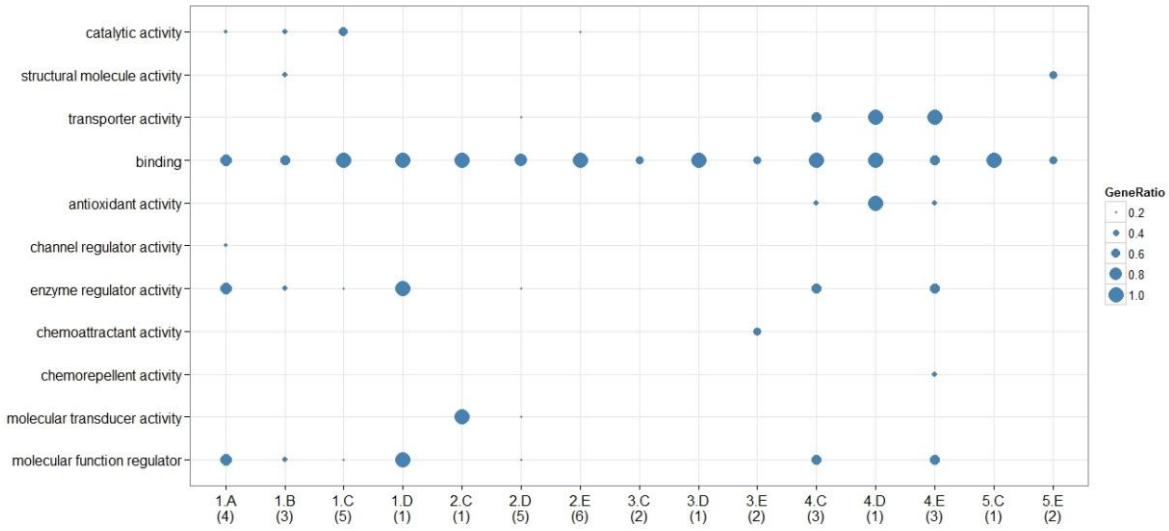
Σχήμα 7.21: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την σημαντικότητά τους σε οντολογία BP

Όπως παρατηρείται στον άξονα **x** του γραφήματος (Σχήμα 7.21) βρίσκονται οι νέες ομαδοποιήσεις των γονιδίων. Η ομαδοποίηση 1.A (4) δηλώνει ότι περιλαμβάνει τα κοινά γονίδια μεταξύ της πρώτης υποομάδας της πρώτης ομαδοποίησης και της πρώτης υποομάδας της δεύτερης ομαδοποίησης, τα οποία είναι τέσσερα. Στον άξονα **y** είναι οι κατηγορίες των βιολογικών λειτουργιών σε κάθε οντολογία. Και εδώ η αναλογία των γονιδίων που αντιστοιχίζονται στις βιολογικές λειτουργίες εκφράζεται μέσω του **GeneRatio**, η κλίμακα μεγέθους του οποίου βρίσκεται δεξιά του διαγράμματος.

Από την δημιουργία των ομαδοποιήσεων εδώ, μπορούμε να συμπεράνουμε ότι τα σημαντικότερα γονίδια της ανάλυσης, τα οποία βρίσκονται στο gA εντοπίζονται στην πρώτη υποομάδα της πρώτης ομαδοποίησης. Τα δεύτερα σε σημαντικότητα, gB, επίσης στην πρώτη υποομάδα της πρώτης ομαδοποίησης. Ως συμπέρασμα λαμβάνεται ότι τα γονίδια στις υψηλές θέσεις στην λίστα ελλατούμενης σημαντικότητας εμφανίζουν λειτουργικές ομοιότητες μεταξύ τους. Αντίθετα παρατηρούμε ότι τα γονίδια των gC, gD, gE διαμοιράζονται στις υποομάδες 2,3,4,

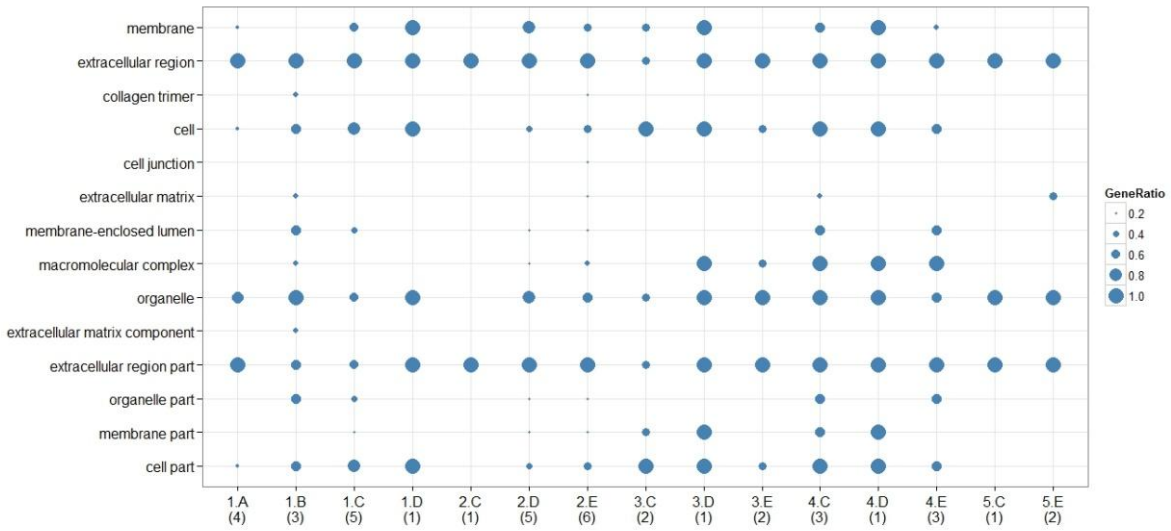
και 5. Ενώ παρατηρείται ότι τα γονίδια αυτά συμμετέχουν σε κοινές βιολογικές λειτουργίες αντίθετα διαφέρουν στην συμμετοχή τους σε άλλες.

Τα σημαντικότερα γονίδια, δηλαδή των gA και gB, συμμετέχουν σε μεγάλο βαθμό κυρίως στις παρακάτω λειτουργικές κατηγορίες: μεταβολική διεργασία, κυτταρική διεργασία, ρύθμιση βιολογικής διεργασίας. Ενώ τα γονίδια των group C, D, E συμμετέχουν σε: διεργασία ανοσοποιητικού συστήματος, ανάπτυξη, απόκριση, κίνηση, θετική ρύθμιση της βιολογικής απόκρισης (positive regulation of biological behavior), απόκριση σε ερέθισμα, εντοπισμός θέσης, οργάνωση κυτταρικού συστατικού ή βιογένεση.



Σχήμα 7.22: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την σημαντικότητά τους σε οντολογία MF

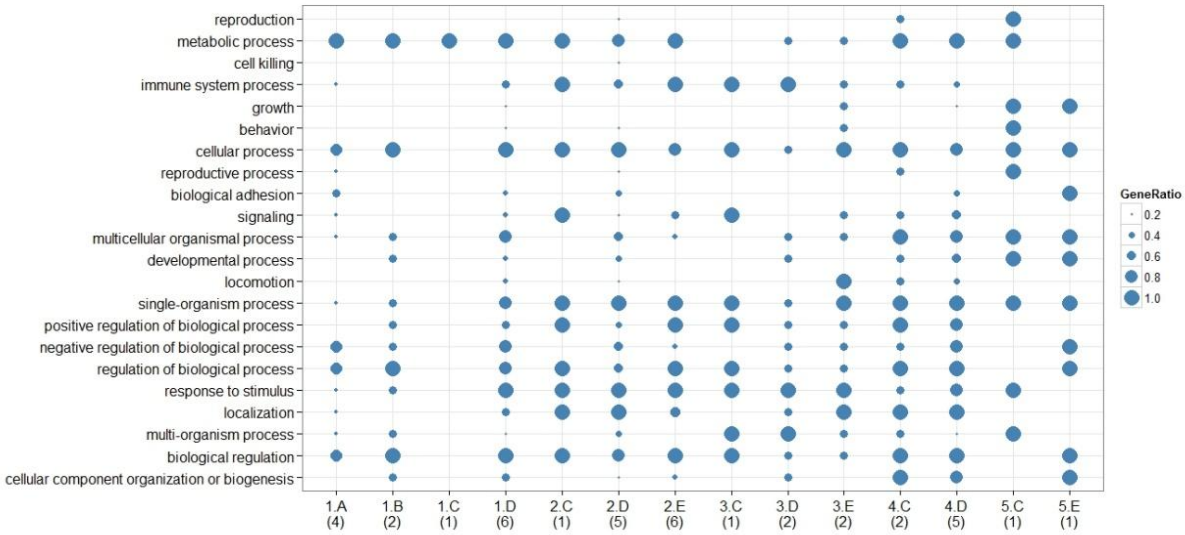
Σε αυτό το γράφημα (Σχήμα 7.22) σε οντολογία MF, δεν παρατηρείται σημαντική διαφορά μεταξύ των ομαδοποιήσεων. Οι βιοχημικές δραστηριότητες στις οποίες συμμετέχουν τα σημαντικότερα γονίδια εδώ, συμμετέχουν και τα λιγότερο σημαντικά, οι οποίες είναι: δέσμευση, ρυθμιστής δραστηριότητας ενζύμου, ρυθμιστής μοριακής λειτουργίας (molecular function regulator).



Σχήμα 7.23: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την σημαντικότητά τους σε οντολογία CC

Σύμφωνα με το γράφημα σε οντολογία CC (Σχήμα 7.23), τα γονίδια του gA λειτουργούν στις θέσεις: εξωκυττάρια περιοχή, τμήμα εξωκυττάριας περιοχής, οργανίδιο, τμήμα του κυττάρου, κύτταρο. Τα γονίδια του gB αντίστοιχα: αυλός που περιβάλλεται από μεμβράνη, τμήμα οργανιδίου. Αυτό επιβεβαιώνει ότι τα γονίδια των gA και gB εντοπίζονται στις ίδιες θέσεις. Στις θέσεις αυτές δραστηριοποιούνται και τα γονίδια των gC, gD, gE, αλλά παρουσιάζονται και στις: μεμβράνη, μέρος μεμβράνης (membrane part) και μακρομοριακό σύμπλεγμα (macromolecular complex).

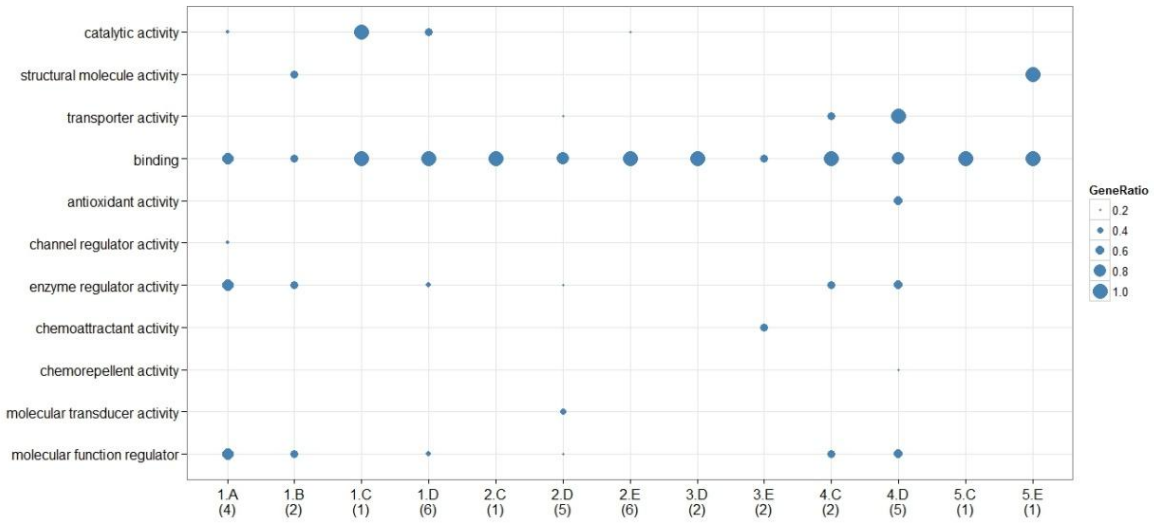
Το παρακάτω γραφήματα παρουσιάζουν την σύγκριση της ομαδοποίησης με πέντε ομάδες με την ομαδοποίηση των γονιδίων σε πέντε ομάδες με βάση την τιμή της μεταβλητής σημαντικότητας για προβολή (Variable Importance to the Projection-**VIP**).



Σχήμα 7.24: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την μεταβλητή VIP σε οντολογία BP

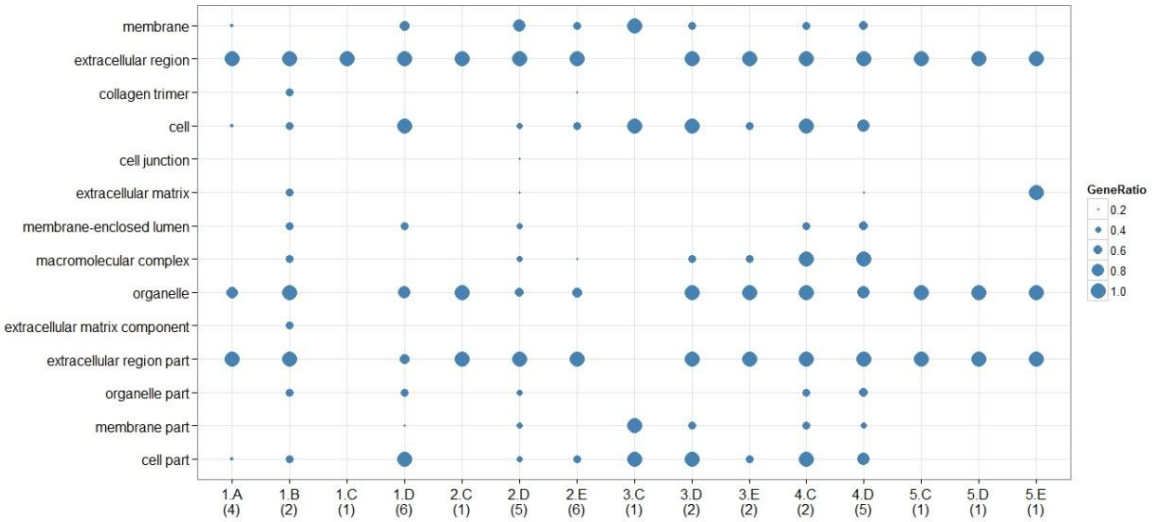
Τα σημαντικότερα γονίδια, των gA και gB, εδώ εμφανίζουν και αυτά λειτουργικές ομοιότητες μιας και ανήκουν στην πρώτη υποομάδα της πρώτης ομαδοποίησης, ενώ αντίθετα τα γονίδια gC, gD και gE διαμοιράζονται στις άλλες τέσσερις υποομάδες.

Με βάση το γράφημα σε οντολογία BP (Σχήμα 7.24), τα σημαντικότερα γονίδια συμμετέχουν σε μεγάλη αναλογία τους στις εξής λειτουργίες: μεταβολική διεργασία, κυτταρική διεργασία, ρύθμιση βιολογικής διεργασίας. Δεν συμμετέχουν σε μεγάλη αναλογία στις κατηγορίες που συμμετέχουν τα γονίδια των άλλων gC, gD, gE οι οποίες είναι οι εξής: διεργασία ανοσοποιητικού συστήματος, ανάπτυξη, αναπαραγωγική διεργασία (reproductive process), κίνηση, διεργασία ενός οργανισμού, θετική ρύθμιση βιολογικής διεργασίας, απόκριση σε ερέθισμα, εντοπισμός θέσης και οργάνωση κυτταρικού συστατικού ή βιογένεση.



Σχήμα 7.25: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την μεταβλητή VIP σε οντολογία MF

Στο γράφημα σε οντολογία MF (Σχήμα 7.25) παρατηρείται ότι όλα τα γονίδια ανήκουν στην λειτουργική κατηγορία της δέσμευσης, ενώ τα γονίδια του gA και gB ανήκουν και στις κατηγορίες ρυθμιστής δραστηριότητας ενζύμου, ρυθμιστής μοριακής δραστηριότητας (molecular regulator activity) σε σημαντική αναλογία σε σχέση με τα γονίδια των άλλων group.



Σχήμα 7.26: Συγκριτικό γράφημα βιολογικού περιεχομένου μεταξύ της ομαδοποίησης των 5 cluster και της ομαδοποίησης των γονιδίων με βάση την μεταβλητή VIP σε οντολογία CC

Με βάση το παραπάνω γράφημα (Σχήμα 7.26) σε οντολογία CC εξάγεται το συμπέρασμα ότι τα περισσότερα γονίδια δραστηριοποιούνται στις θέσεις: οργανίδιο, εξωκυττάρια περιοχή και τμήμα εξωκυττάριας περιοχής, οι οποίες είναι οι κατηγορίες που ταξινομούνται τα σημαντικότερα γονίδια σε μεγαλύτερη αναλογία. Τα λιγότερο σημαντικά γονίδια των gC, gD, gE στις θέσεις: κύτταρο, μακρομοριακό σύμπλοκο (macromolecular complex), τμήμα μεμβράνης, τμήμα κυττάρου.

Η παραπάνω σύγκριση πραγματοποιήθηκε για να μελετήσουμε τις διαφορές της ταξινόμησης των γονιδίων με βάση το $\boldsymbol{Q^2}$ και το VIP αλλά και για να συγκρίνουμε τη σημαντικότητα των γονιδίων με την συμμετοχή τους σε βιολογικές λειτουργίες χρησιμοποιώντας το αποτέλεσμα εμπλουτισμού για την ομαδοποίηση τους με βάση την λειτουργική τους ομοιότητα. Με βάση το αποτέλεσμα αυτό καθώς και το αποτέλεσμα του υποκεφαλαίου 6.3.4.2 γνωρίζουμε ότι τα γονίδια της ομάδας C3 και C4 φαίνεται να εμπλέκονται σε λειτουργίες που σχετίζονται με την τοξικότητα.

Αρχικά η σύγκριση μεταξύ των ταξινομήσεων σημαντικότητας, μας έδειξε ότι τα γονίδια που βρίσκονται σε αντίστοιχες θέσεις στις λίστες κατάταξης με βάση τα μεγέθη τιμών $\boldsymbol{Q^2}$ και VIP μετέχουν στις ίδιες λειτουργίες. Παραδείγματος χάριν τα γονίδια που βρίσκονται στη μέση της λίστας κατάταξης δηλαδή στις ομάδες gC και gD και για τα δυο μεγέθη μετέχουν στις ίδιες λειτουργίες, βιοχημικές δραστηριότητες και εντοπίζονται στις ίδιες θέσεις.

Αντίστοιχα σύγκριση με το ποιά γονίδια ανάλογα με την σημαντικότητα συμμετέχουν στις εμπλουτισμένες κατηγορίες, συμπεραίνουμε ότι τα γονίδια που βρίσκονται μεταξύ της $8^{ης}$ και της $28^{ης}$ θέσης και για τις δυο λίστες ταξινόμησης συμμετέχουν σε λειτουργίες, δραστηριότητες και θέσεις που σχετίζονται με την τοξικότητα. Αυτές περιλαμβάνουν την μεταφορά και τον εντοπισμό λιπιδίων, την δραστηριότητα μεταφοράς λιπιδίων, τη δέσμευση φωσφολιπιδίων, την λειτουργία του ανοσοποιητικού συστήματος, τις αποκρίσεις σε ερεθίσματα. Επίσης τα γονίδια αυτά δραστηριοποιούνται σε σωματίδια λιποπρωτεΐνης και στην μεμβράνη που παίζουν καθοριστικό ρόλο για την είσοδο ξένων συστατικών στο κύτταρο.

7.3 KYOTO εγκυκλοπαίδεια γονιδίων και γονιδιωμάτων

Για αυτή την ανάλυση χρησιμοποιείται επίσης το τεστ-υπερεκπροσώπησης με χρήση του υπεργεωμετρικού μοντέλου με χρήση της βιβλιοθήκης “clusterProfiler” [59], όμως η βάση δεδομένων που χρησιμοποιείται τώρα είναι η “KEGG.db”. [54]

Η βάση δεδομένων της KEGG παρέχει μια ταξινόμηση των γονιδίων με βάση το βιολογικό μονοπάτι στο οποίο ανήκουν. Κάθε όρος της KEGG αντιπροσωπεύει ένα μονοπάτι, και έχει κάποια γονίδια που αντιστοιχίζονται σε αυτό και αποτελούν τη μετάφραση του.

Εδώ θα ελεγχθεί μέσω της ανάλυσης εμπλουτισμού ποια βιοχημικά μονοπάτια (pathways) της KEGG υπερ-εκπροσωπούνται με την είσοδο του νανοσωματιδίου στον οργανισμό.

Η ανάλυση αυτή απαιτεί:

- ✓ Entez ταυτότητες γονιδίων σε μορφή διανύσματος
- ✓ Ορισμός του οργανισμού, που για την ανάλυση αυτή, είναι ο άνθρωπος
- ✓ Όριο αποκοπής για την τιμή **p (p-value)** το 0,01.(**p-value**≤0.01)

Εδώ θα χρησιμοποιηθούν οι 118 Entrez ταυτότητες των γονιδίων που προκύπτουν από την μετάφραση των 129 Uniprot ταυτοτήτων των πρωτεϊνών του Συνόλου Α, οι οποίες είναι όλες οι πρωτεΐνες που λαμβάνονται από το πρωτεϊνικό αποτύπωμα του νανοσωματιδίου σε σχετικά υψηλή περιεκτικότητα.

Μέσω της ανάλυσης αυτής αντιστοιχίζονται οι Entrez ταυτότητες της ανάλυσης στις ταυτότητες των KEGG μονοπατιών (ID), τα οποία διαθέτουν μια περιγραφή (Description). Μετρώνται οι ταυτότητες Entrez οι οποίες αντιστοιχούν στα KEGG μονοπάτια (**Count**) και υπολογίζεται η αναλογία αντιστοίχισης τους στο σύνολο των γονιδίων της ανάλυσης (**GeneRatio**). Αντίστοιχα μετράται η αναλογία αντιστοίχισης των γονιδίων του υποβάθρου στα KEGG μονοπάτια (**BgRatio**). Η τιμή **p-value** προκύπτει με τη χρήση του υπεργεωμετρικού μοντέλου με βάση τις δυο παραπάνω αντιστοιχίσεις. Κατόπιν μέσω του ελέγχου FDR υπολογίζονται οι τιμές **p-adjust** και **q-value** για να επαληθευτεί η εγκυρότητα του αποτελέσματος. Το αποτέλεσμα περιορίζεται με βάση τις τιμές για το όρια αποκοπής που ορίστηκαν για την τιμή **p-value**, και την τιμή **q-value**.

Ένα μέρος της μορφής του αποτελέσματος παρουσιάζεται στο Σχήμα 7.27, στο περιβάλλον της R. Στην πρώτη γραμμή εμφανίζεται η ταυτότητα του μονοπατιού της KEGG hsa04610 με περιγραφή “Complement and coagulation cascades” που εμφανίζεται με ***GeneRatio*** 36/70 ενώ με ***BgRatio*** 69/6899 με αντίστοιχο ***pvalue*** $1.28 \cdot e^{-57}$ και ***padjust*** $1.17 \cdot e^{-56}$ και ***qvalue*** $1.37 \cdot e^{-57}$ με αντιστοιχιζόμενες συμβολικές ταυτότητες των γονιδίων C3/ C4A/ C4B/ SERPINA1/ F2/ SERPINC1/ C4BPA/

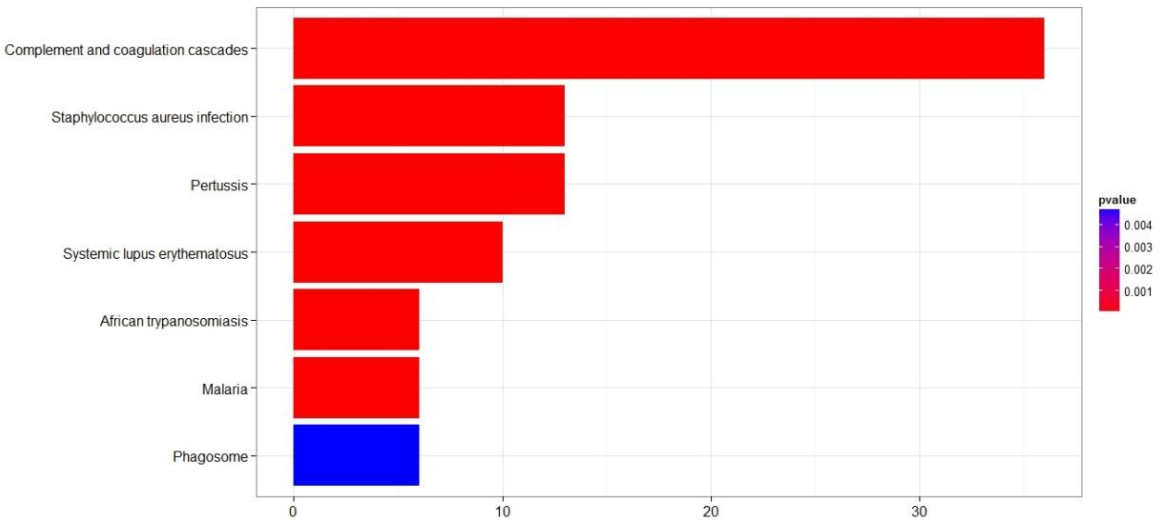
A2M/ KNG1/ SERPINA5/ SERPINF2/ CFH/ F5/ SERPIND1/ C9/ SERPING1/ F11/ KLKB1/ PROS1/ C1R/ F12/ F10/ C1QB/ PLG/ C1QC/ CFB/ F9/ C1S/ PROC/ C4BPB/ FGA/ F7/ F8/ C1QA/ C5/ MASP1 με **Count** ίσο με 36.

```
> head(summary(kk))
           ID                          Description GeneRatio  BgRatio       pvalue     p.adjust       qvalue
hsa04610 hsa04610 Complement and coagulation cascades     36/70  69/6896 1.297586e-57 1.167828e-56 1.365880e-57
hsa05150 hsa05150     Staphylococcus aureus infection     13/70  57/6896 6.529225e-15 2.938151e-14 3.436434e-15
hsa05133 hsa05133                           Pertussis     13/70  75/6896 2.936197e-13 8.808590e-13 1.030244e-13
hsa05322 hsa05322        Systemic lupus erythematosus     10/70 136/6896 9.261010e-07 1.702804e-06 1.991584e-07
hsa05143 hsa05143             African trypanosomiasis      6/70  34/6896 9.460025e-07 1.702804e-06 1.991584e-07
hsa05144 hsa05144                             Malaria      6/70  49/6896 8.726434e-06 1.308965e-05 1.530953e-06
                                                                                                                  $
hsa04610 C3/C4A/C4B/SERPINA1/F2/SERPINC1/C4BPA/A2M/KNG1/SERPINA5/SERPINF2/CFH/F5/SERPIND1/C9/SERPING1/F11/KLKB1/PROS1/C1R/F12/F10/C1QB/PLG/C1QC/CFB/F9/C1S/PROC/C4BPB/$
hsa05150                                                                                                C3/C4A/C4B/CFH/C1R/C1QB/PLG/C1Q$
hsa05133                                                                                  C3/C4A/C4B/C4BPA/SERPING1/C1R/C1QB/C1QC$
hsa05322                                                                                                C3/C4A/C4B/C9/C1$
hsa05143                                                                                                        APOA1/$
hsa05144                                                                                                      HBB/HBA$
         Count
hsa04610    36
hsa05150    13
hsa05133    13
hsa05322    10
hsa05143     6
hsa05144     6
```

Σχήμα 7.27: Αποτέλεσμα KEGG ανάλυσης εμπλουτισμού γονιδίων Συνόλου A σε περιβάλλον της R

7.3.1 Εμπλουτισμένα μονοπάτια της KEGG

Και εδώ τα αποτελέσματα της ανάλυσης εμπλουτισμού θα παρασταθούν στο παρακάτω ραβδόγραμμα και στους χάρτες που απεικονίζουν τις σχέσεις των εμπλουτισμένων μονοπατιών μεταξύ τους και τις σχέσεις μεταξύ γονιδίων και μονοπατιών.



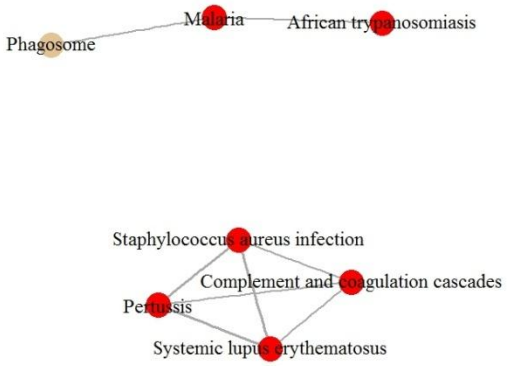
Σχήμα 7.28: Ραβδόγραμμα αποτελέσματος της KEGG ανάλυσης εμπλουτισμού γονιδίων Συνόλου A

Στον Πίνακα 7.9 παρουσιάζονται τα εμπλουτισμένα μονοπάτια της KEGG, οι ονομασίες τους, οι περιγραφές τους καθώς και η αναλογία αντιστοίχισης των γονιδίων σε αυτά και οι τιμές p-value που προκύπτουν για την κάθε μονοπάτι. Στον Πίνακα ΠΒ7 παρουσιάζεται το συνολικό αποτέλεσμα της ανάλυσης αυτής.

Πίνακας 7.9: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Α για τα εμπλουτισμένα μονοπάτια της KEGG

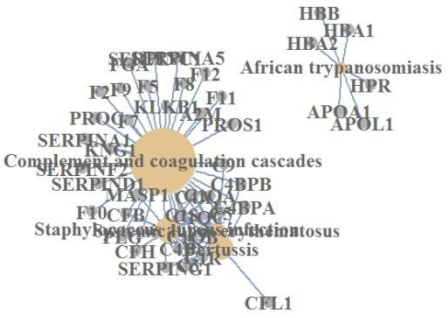
Μονοπάτι της KEGG	Όνομα μονοπατιού	GeneRatio	p-value
hsa04610	Σύστημα του συμπληρώματος και αλληλουχία αντιδράσεων πήξης (Complement and coagulation cascades)	36/70	$9.26 \cdot 10^{-58}$
hsa05150	Μόλυνση από το βακτήριο Staphylococcus aureus (Staphylococcus aureus infection)	13/70	$3.49 \cdot 10^{-15}$
hsa05133	Κοκκύτης (Pertussis)	13/70	$2.61 \cdot 10^{-13}$
hsa05322	Συστηματικός ερυθηματώδης λύκος (Systemic lupus erythematosus)	10/70	$7.42 \cdot 10^{-07}$
hsa05143	Αφρικανική τρυπανοσωμίαση (African trypanosomiasis)	6/70	$8.96 \cdot 10^{-07}$
hsa05144	Ελονοσία (Malaria)	6/70	$8.28 \cdot 10^{-06}$
hsa04145	Φαγοκυττάρωση (Phagosome)	6/70	0.004683

Το μέγεθος GeneRatio είναι αναλογία στα 70 αντί στα 118 γονίδια, διότι 48 Entrez ταυτότητες γονιδίων δεν αντιστοιχίζονται σε μονοπάτι της βάσης δεδομένων της KEGG βιβλιοθήκης.



Σχήμα 7.29: Χάρτης απεικόνισης των σχέσεων των εμπλουτισμένων βιολογικών μονοπατιών της KEGG

Στον χάρτη μπορούμε να διακρίνουμε ποια μονοπάτια συνδέονται μεταξύ τους. Παρατηρείται ότι δημιουργούνται δύο ομάδες μονοπατιών που έχουν σχέσεις μεταξύ τους. Η πρώτα ομάδα περιλαμβάνει την φαγοκυττάρωση, την ελονοσία και την αφρικανική τρυπανοσωμίαση, ενώ η δεύτερη ομάδα το σύστημα του συμπληρώματος, τη λοίμωξη του Staphylococcus aureus, την λοίμωξη του κοκκύτη, και τη νόσο του ερυθηματώδους λύκου.



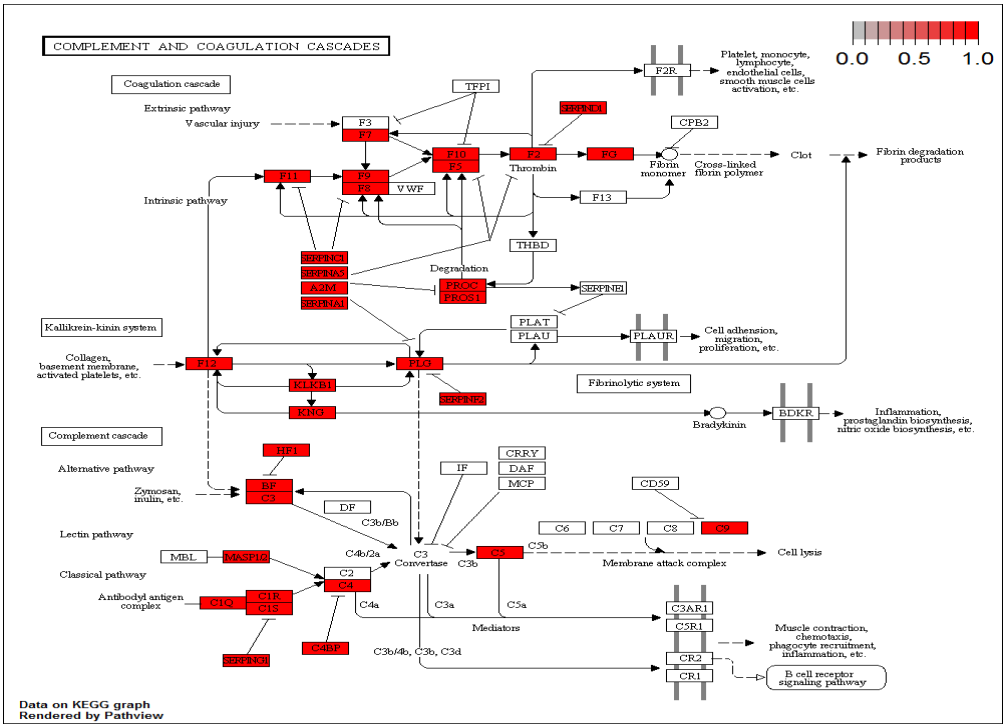
Σχήμα 7.30: Χάρτης σχέσεων γονιδίων-μονοπατιών της KEGG

Στο χάρτη (Σχήμα 7.30) παρατηρείται ότι τα γονίδια των μονοπατιών που έχουν σχέση μεταξύ τους συμμετέχουν σε παραπάνω από μια λειτουργίες, ενώ τα γονίδια που σχετίζονται με την αφρικανική τρυπανοσωμίαση δεν συμμετέχουν σε άλλες λειτουργίες. Η ανάλυση εμπλουτισμού πραγματοποιήθηκε για τις ταυτότητες

γονιδίων του Συνόλου Α, για αυτό το λόγο τα περισσότερα γονίδια που συμμετέχουν στα εμπλουτισμένα μονοπάτια της KEGG δεν ανήκουν στο σύνολο των 76 σημαντικότερων γονιδίων. Όμως στο μονοπάτι του συστήματος του συμπληρώματος και της αλληλουχίας αντιδράσεων πήξης συμμετέχουν τα γονίδια C4BPA, KNG1, KLB1, C1R, PROC, C4BPB και C1QA που μετέχουν στο Σύνολο Β και λαμβάνουν τιμές $\boldsymbol{Q^2}$ στο διάστημα [0.7577,0.8051].

Με χρήση της βιβλιοθήκης "pathview" [60] μπορούν να εικονιστούν τα μονοπάτια της KEGG. Η βιβλιοθήκη αυτή χαρτογραφεί και τοποθετεί τα δεδομένα της ανάλυσης του κάθε χρήστη στα σχετικά γραφήματα μονοπατιών που ανήκουν. Η λειτουργία αυτή απαιτεί από το χρήση να παρέχει τα γονίδια, και να ορίσει το επιθυμητό μονοπάτι. Έτσι αυτόματα γίνεται λήψη των δεδομένων του μονοπατιού του γραφήματος, αναλύεται το αρχείο των δεδομένων, χαρτογραφούνται τα δεδομένα του χρήστη στο μονοπάτι, και δημιουργείται το γράφημα του μονοπατιού με τα χαρτογραφημένα δεδομένα.

Ακολουθεί η απεικόνιση του μονοπατιού του συστήματος του συμπληρώματος και της αλληλουχία αντιδράσεων πήξης (Σχήμα 7.31) στο οποίο συμμετέχουν τα περισσότερα γονίδια της ανάλυσης. Τα γονίδια που χαρτογραφούνται στην απεικόνιση του μονοπατιού με χρώμα κόκκινο είναι αυτά που αντιστοιχίζονται με τις Entrez ταυτότητες των γονιδίων της ανάλυσης που συμμετέχουν στο μονοπάτι.



Σχήμα 7.31: Μονοπάτι hsa04610 της KEGG

7.4 Reactome

Η REACTOME είναι μια ανοιχτού κώδικα και ανοιχτής πρόσβασης βάση δεδομένων βιολογικών μονοπατιών και αντιδράσεων του ανθρώπινου οργανισμού. Η 46^{η} έκδοση της βάσης δεδομένων του REACTOME, περιγράφει 7088 ανθρώπινες πρωτεΐνες (34% του προβλεπόμενου ανθρώπινου πρωτεώματος, που είναι το σύνολο των πρωτεϊνών ενός κυττάρου που κατευθύνουν την φυσική ανάπτυξη και την συμπεριφορά του), οι οποίες συμμετέχουν σε 6744 αντιδράσεις βασιζόμενες σε δεδομένα που έχουν εξαχθεί από 15107 δημοσιευμένες έρευνες. Περιέχει βάσεις δεδομένων όπως η NCBI Gene, η Ensembl, η UNIPROT, η KEGG και η Gene Ontology. [61]

Η ανάλυση των μονοπατιών της REACTOME, πραγματοποιείται μέσω της βιβλιοθήκης “ReactomePA” [61] στο περιβάλλον της R. Τα εργαλεία της ανάλυσης μονοπατιών αναλύουν τα δεδομένα που παρέχει ο χρήστης, επιτρέποντας την χαρτογράφηση των ταυτοτήτων των δεδομένων, την απεικόνιση των μονοπατιών και την ανάλυση εμπλουτισμού. Επαναλαμβάνουμε την GEA ανάλυση διότι η κάθε βάση δεδομένων περιέχει διαφορετικές πληροφορίες και έπειτα τα αποτελέσματα εμπλουτισμού της κάθε μιας θα συγκριθούν.

7.4.1 *Ανάλυση εμπλουτισμού*

Όπως αναφέρθηκε παραπάνω η ανάλυση εμπλουτισμού είναι μια ευρέως χρησιμοποιούμενη προσέγγιση για τα βιολογικά θέματα. Η ανάλυση που μας παρέχει η βιβλιοθήκη "ReactomePA", εφαρμόζει το υπεργεωμετρικό μοντέλο για να αξιολογήσει αν ο αριθμός των επιλεγμένων γονιδίων που συνδέονται με το «μονοπάτι αντιδράσεων» είναι μεγαλύτερος από τον αναμενόμενο. Οι τιμές **p** υπολογίζονται μέσω του υπεργεωμετρικού μοντέλου. [4]

Η ανάλυση εμπλουτισμού απαιτεί :

- ✓ Entrez ταυτότητες γονίδιων σε μορφή διανύσματος
- ✓ Υπόβαθρο γονιδίων που εδώ παρέχεται από τη βιβλιοθήκη "org.Hs.eg.db"
- ✓ Ορισμός του οργανισμού, που στην ανάλυση αυτή είναι ο άνθρωπος
- ✓ Όριο αποκοπής για την τιμή **p** το 0,01 (**p-value≤0,01**)
- ✓ Μέθοδος προσαρμογής των τιμών **p** (**p-adjust**) που εδώ είναι ο έλεγχος FDR
- ✓ Όριο αποκοπής για την τιμή **q** το 0,05, του FDR ελέγχου. (**q-value**≤0,05)

Για την ανάλυση θα χρησιμοποιηθούν οι 118 Entrez ταυτότητες που προκύπτουν από την μετάφραση των 129 Uniprot ταυτοτήτων των πρωτεϊνών του Συνόλου Α με χρήση της μεταφραστικής βιβλιοθήκης "org.Hs.eg.db".

```
> head((summary(x)))
           ID                                Description GeneRatio  BgRatio       pvalue     p.adjust       qvalue
140877 140877  Formation of Fibrin Clot (Clotting Cascade)     18/84  36/5302 4.186227e-24 1.674491e-22 3.525244e-23
166658 166658                         Complement cascade     16/84  32/5302 1.733527e-21 3.467054e-20 7.299061e-21
114608 114608                      Platelet degranulation      20/84  74/5302 2.757675e-20 3.676900e-19 7.740842e-20
76005   76005 Response to elevated platelet cytosolic Ca2+     20/84  79/5302 1.170852e-19 1.170852e-18 2.464951e-19
109582 109582                                 Hemostasis     33/84 403/5302 2.237614e-16 1.790092e-15 3.768614e-16
140837 140837   Intrinsic Pathway of Fibrin Clot Formation     10/84  16/5302 4.268800e-15 2.845867e-14 5.991298e-15
                                                                                                                  geneID
140877                                             2147/462/2/3827/5104/2153/3053/710/2160/3818/5627/2161/2159/2158/5624/2243/2155/2157
166658                                                          718/7448/722/3075/735/5627/715/713/629/716/725/2220/5199/712/727/5648
114608                                         335/1191/5265/3273/2/3827/7018/5345/2153/710/5627/5340/7094/5341/2243/7057/2157/5473/1072/2335
76005                                          335/1191/5265/3273/2/3827/7018/5345/2153/710/5627/5340/7094/5341/2243/7057/2157/5473/1072/2335
109582 335/1191/5265/338/2147/462/3273/2/3827/5104/7018/5345/2153/3053/3043/710/2160/3818/5627/2161/2159/5340/2158/7094/5624/5341/2243/7057/2155/2157/5473/1072/2335
140837                                                                                     2147/2/3827/710/2160/3818/2161/2159/2158/2157
       Count
140877    18
166658    16
114608    20
76005     20
109582    33
140837    10
```

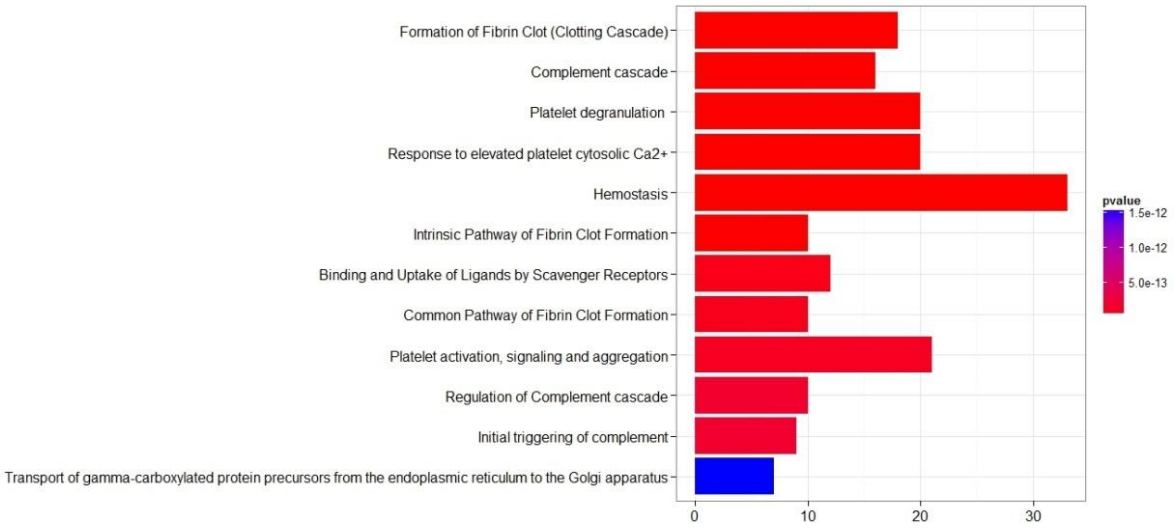
Σχήμα 7.32: Αποτέλεσμα REACTOME ανάλυσης εμπλουτισμού γονιδίων Συνόλου Α στο περιβάλλον της R

Το αποτέλεσμα της ανάλυσης εμπλουτισμού στο περιβάλλον της R απεικονίζεται στο Σχήμα 7.32. Για παράδειγμα το πρώτο μονοπάτι που εμφανίζεται στην πρώτη γραμμή του Σχήματος 7.32 με ταυτότητα 140877 με περιγραφή "Formation of Fibrin

Clot(Clotting Cascade)" έχει ***GeneRatio*** 18/84 και ***BgRatio*** 36/5302 με υπολογιζόμενο ***pvalue*** $4.19 \cdot e^{-24}$ και προσαρμοσμένη τιμή ***padjust*** $6.32 \cdot e^{-22}$ καθώς και αντίστοιχη ***qvalue*** $4.49 \cdot e^{-22}$ με αντιστοιχιζόμενες Entrez ταυτότητες (gene ID) τις 2147 /462/ 3827/ 5104/ 2153/ 3053/ 710/ 2160/ 3818/ 5627/ 2161/ 2159/ 2158/ 5624/ 2243/ 2155/ 2157 με αντίστοιχο ***Count*** 18. Το συνολικό αποτέλεσμα περιλαμβάνεται στον Πίνακα ΠΒ8.

7.4.2 Εμπλουτισμένα μονοπάτια της Reactome

Μια μορφή του αποτελέσματος λαμβάνεται στο ραβδόγραμμα του Σχήματος 7.33 στο οποίο επιλέγεται να παρουσιάζονται μόνο τα 12 μονοπάτια που αντιστοιχίζονται τα γονίδια στον άξονα **y** καθώς και η κλίμακα της τιμής **p** που αντιστοιχεί σε αυτά.



Σχήμα 7.33: Ραβδόγραμμα αποτελέσματος REACTOME ανάλυσης εμπλουτισμού γονιδίων Συνόλου Α

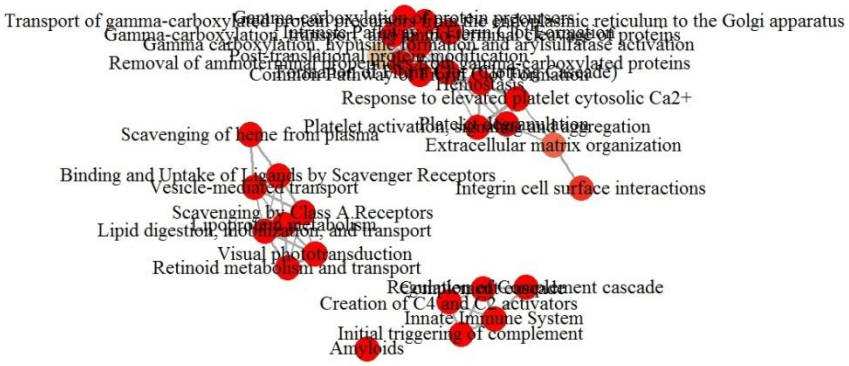
Πίνακας 7.10: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Α για τα εμπλουτισμένα μονοπάτια της REACTOME

Reactome ID	Μονοπάτι της REACTOME	GeneRatio	p-value
140877	Σχηματισμός ινώδους θρόμβου (Formation of Fibrin Clot (Clotting Cascade))	18/84	$4{,}19 \cdot 10^{-24}$
166658	Αλληλουχία διαδικασιών του συμπληρώματος (Complement cascade)	16/84	$1{,}73 \cdot 10^{-21}$

114608	Αποκοκκίωση των αιμοπεταλίων (Platelet degranulation)	20/84	$2,76 \cdot 10^{-20}$
76005	Απόκριση στο αυξημένο κυτοσυλικό Ca^{+2} των αιμοπεταλίων (Response to elevated platelet cytosolic Ca2+)	33/84	$1,17 \cdot 10^{-19}$
109582	Αιμόσταση (Hemostasis)	10/84	$2,24 \cdot 10^{-16}$
140837	Ενδογενές μονοπάτι του σχηματισμού του ινώδους θρόμβου (Intrinsic Pathway of Fibrin Clot Formation)	21/84	$4,27 \cdot 10^{-15}$
2173782	Πρόσδεση και πρόσληψη των προσδετών από Scavenger υποδοχείς (Binding and Uptake of Ligands by Scavenger Receptors)	10/84	$6,93 \cdot 10^{-14}$
140875	Κοινό μονοπάτι του σχηματισμού του ινώδους θρόμβου (Common Pathway of Fibrin Clot Formation)	9/84	$9,36 \cdot 10^{-14}$
76002	Ενεργοποίηση των αιμοπεταλίων, της σηματοδότησης και της συσσώρευση τους (Platelet activation, signaling and aggregation)	7/84	$1,27 \cdot 10^{-13}$
977606	Ρύθμιση της αλληλουχίας διαδικασιών του συμπληρώματος (Regulation of Complement Cascade)	7/84	$1,76 \cdot 10^{-13}$
166663	Αρχική ενεργοποίηση του συμπληρώματος (Initial triggering of complement)	7/84	$1,88 \cdot 10^{-13}$

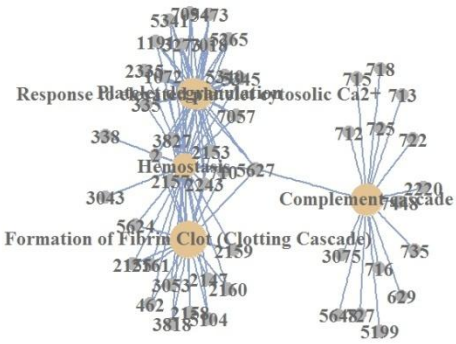
Από την περιγραφή των μονοπατιών (παρατίθεται στο Παράρτημα Γ), εξάγεται ως συμπέρασμα ότι τα μονοπάτια αυτά αφορούν την ενεργοποίηση του αμυντικού συστήματος του οργανισμού σε διαφορετικές περιπτώσεις όπως ο τραυματισμός ενός αγγείου. Γι' αυτό και είναι αναμενόμενο να είναι άμεσα συνδεδεμένες μεταξύ τους. Έτσι με τη βοήθεια της βιβλιοθήκης αυτής παρέχεται μια λειτουργία η οποία υπολογίζει τις σχέσεις των γονιδίων μεταξύ τους με χρήση του αποτελέσματος της ανάλυσης εμπλουτισμού.

Ο παρακάτω χάρτης παρουσιάζεις τις σχέσεις των εμπλουτισμένων μονοπατιών στα οποία συμμετέχουν τα γονίδια της ανάλυσης.



Σχήμα 7.34: Χάρτης απεικόνισης των σχέσεων των εμπλουτισμένων βιολογικών μονοπατιών της REACTOME

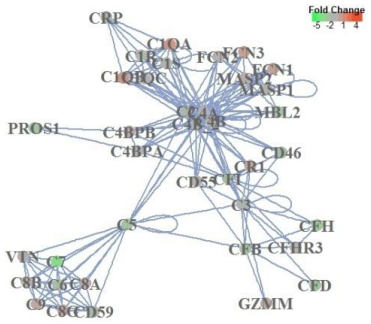
Στο χάρτη (Σχήμα 7.34) διακρίνονται τρεις ομάδες μονοπατιών που σχετίζονται μεταξύ τους, ενώ μόνο το μονοπάτι της λειτουργίας των αμυλοειδών (amyloids) δεν συνδέεται με κανένα άλλο. Η πρώτη ομάδα περιλαμβάνει την αλληλουχία διαδικασιών του συμπληρώματος, την ρύθμιση και την ενεργοποίηση τους, την έμφυτη λειτουργία του ανοσοποιητικού συστήματος καθώς και την λειτουργία των ενεργοποιητών C2 και C4. Η δεύτερη ομάδα περιλαμβάνει την πρόσδεση και πρόσληψη των προσδετών από Scavenger υποδοχείς, τον μεταβολισμό των λιποπρωτεϊνών και των ρετινοειδών, την μεταφορά λιπιδίων και την ενδοκυττάρωση από Class A Scavenger υποδοχείς. Η τρίτη ομάδα περιλαμβάνει τα υπόλοιπα μονοπάτια της ανάλυσης.



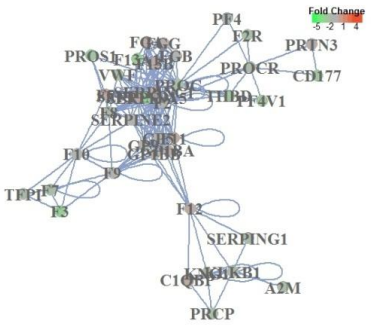
Σχήμα 7.35: Χάρτης σχέσης γονιδίων-μονοπατιών της REACTOME

Στον χάρτη του Σχήματος 7.35 απεικονίζονται οι σχέσεις γονιδίων (Entrez ταυτότητες) με τα πέντε εμπλουτισμένα λειτουργικά μονοπάτια της ανάλυση εμπλουτισμού της Reactome. Παρατηρείται ότι τα γονίδια που συμμετέχουν στον σχηματισμό του ινώδους θρόμβου, συμμετέχουν εξίσου στις λειτουργίες της αιμόστασης και αντίστοιχα τα γονίδια που συμμετέχουν στην αιμόσταση συμμετέχουν εξίσου στις λειτουργίες της αποκοκκίωση των αιμοπεταλίων και της απόκρισης στο αυξημένο κυτοσυλικό Ca^{+2} των αιμοπεταλίων. Τα γονίδια που παίρνουν μέρος στην αλληλουχία των διαδικασιών του συμπληρώματος δεν μετέχουν σε άλλα μονοπάτια εκτός από το γονίδιο 5627.

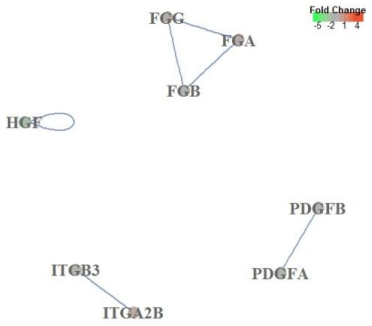
Με χρήση των λειτουργιών της βιβλιοθήκης μπορούμε να απεικονίσουμε σε χάρτη τις σχέσεις των γονιδίων που συμμετέχουν στα πέντε πιο εμπλουτισμένα μονοπάτια. Έτσι για κάθε μονοπάτι έχουμε τον αντίστοιχο χάρτη, στον οποίο εμφανίζονται οι σχέσεις των γονιδίων τα οποία συμμετέχουν σε αυτό, με χρήση των συμβολικών ονομασιών τους .



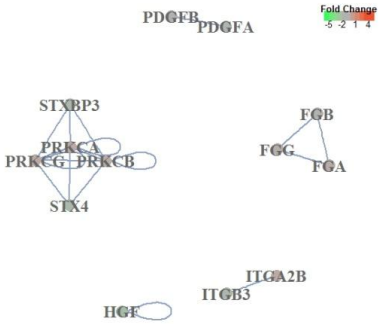
Σχήμα 7.36: Σχέσεις γονιδίων στο μονοπάτι: Αλληλουχία διαδικασιών του συμπληρώματος



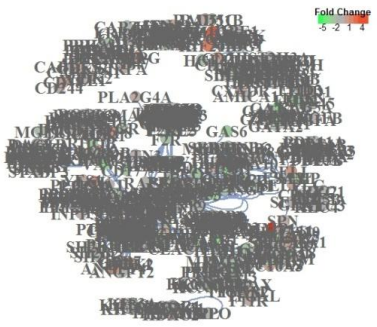
Σχήμα 7.37: Σχέσεις γονιδίων στο μονοπάτι: Σχηματισμός ινώδους θρόμβου



Σχήμα 7.38: Σχέσεις γονιδίων στο μονοπάτι: Αποκοκκίωση των αιμοπεταλίων



Σχήμα 7.39: Σχέσεις γονιδίων στο μονοπάτι: Ανταπόκριση στο αυξημένο κυτοσυλικό Ca^{+2} των αιμοπεταλίων



Σχήμα 7.40 Σχέσεις γονιδίων στο μονοπάτι : Αιμόσταση

Όπως παρατηρείται στους παραπάνω χάρτες κάποια μονοπάτια λειτουργιών περιλαμβάνουν πολλά γονίδια (Σχήμα 7.39, 7.40, 7.43) ενώ άλλα πολύ λιγότερα

(Σχήμα 7.41, 7.42). Επίσης σε κάθε χάρτη απεικονίζεται η σχέση των γονιδίων μεταξύ τους. Στα Σχήματα 7.39 και 7.40 όλα τα γονίδια συνδέονται σε ένα ευρύτερο δίκτυο, ενώ στα υπόλοιπα σχήματα εντοπίζονται ομάδες γονιδίων του ίδιου μονοπατιού να παρουσιάζουν σχέσεις ενώ άλλα όχι.

Παρατηρούμε ότι η ανάλυση της Reactome μας δίνει πιο συγκεκριμένους όρους βιολογικής πληροφορίας (Πίνακας 7.10) σε σχέση με αυτά της GO ανάλυσης (Πίνακας 7.5, 7.6). Τα αποτελέσματα εμπλουτισμού της GO, Reactome και KEGG περιλαμβάνουν την αλληλουχία διαδικασιών του συμπληρώματος και την ενεργοποίηση του, οι οποίες ενεργοποιούνται σε τοξικές συνθήκες [62]. Αντίστοιχα ενεργοποιείται και η λειτουργίας πήξης της KEGG περιλαμβάνεται και στο αποτέλεσμα της Reactome με τη μορφή του μονοπατιού του σχηματισμού του ινώδους θρόμβου και της ενεργοποίησης και συσσώρευσης των αιμοπεταλίων. [62]

7.4.3 Προφίλ των ομαδοποιημένων γονιδίων με βάση την ανάλυση εμπλουτισμού

Η βιβλιοθήκη “ReactomePa” σε συνδυασμό με την βιβλιοθήκη “clusterProfiler” μπορεί να συγκρίνει τα βιολογικά περιεχόμενα της ανάλυσης εμπλουτισμού μεταξύ των ομαδοποιημένων γονιδίων.

Η λειτουργία αυτή έχει ως απαίτηση :

- ✓ Λίστα ομαδοποίησης των γονιδίων
- ✓ Την μέθοδο με την οποία θα πραγματοποιηθεί η ανάλυση εμπλουτισμού, με χρήση του υπεργεωμετρικού μοντέλου
- ✓ Ορισμός του οργανισμού, που στην ανάλυση αυτή είναι ο άνθρωπος

Η ομαδοποίηση των γονιδίων που θα χρησιμοποιηθεί, είναι αυτή των πέντε ομάδων (cluster) των γονιδίων που έχει δημιουργηθεί με κριτήριο την ομοιότητα των λειτουργιών στις οποίες μετέχουν τα γονίδια του πρωτεϊνικού αποτυπώματος του νανοσωματιδίου, με χρήση της βιβλιοθήκης “GOSim”.

Τα γονίδια σε μορφή Entrez ταυτοτήτων χωρίζονται στις εξής πέντε ομάδες:

C1 : 718, 720, 721, 1191, 7448, 722, 325, 3075, 5197, 6291, 3078, 735, 3929, 715, 713, 714, 629, 716, 725, 81494, 5004, 5473, 2220, 1401, 5199, 1072, 712, 4064, 727, 5648

C2 : 335, 348, 336, 345, 346, 341, 344, 319, 116519, 55937, 3931

C3 : 3700, 5265, 338, 5104, 3698, 6414, 197, 8542, 3697, 259, 12, 6694, 2638, 3026, 283, 3263, 5444, 4057, 3485, 183, 7123, 1369, 1, 3699, 5315, 3483, 4060, 26998, 3959, 1311, 22883, 811, 7060, 79791

C4 : 2147, 462, 3273, 7276, 2, 3827, 7018, 5345, 2153, 3053, 350, 710, 2160, 3818, 5627, 2161, 2159, 5340, 2158, 7094, 51156, 5624, 1356, 8858, 5341, 4627, 2243, 7057, 2155, 2157, 2335, 84735

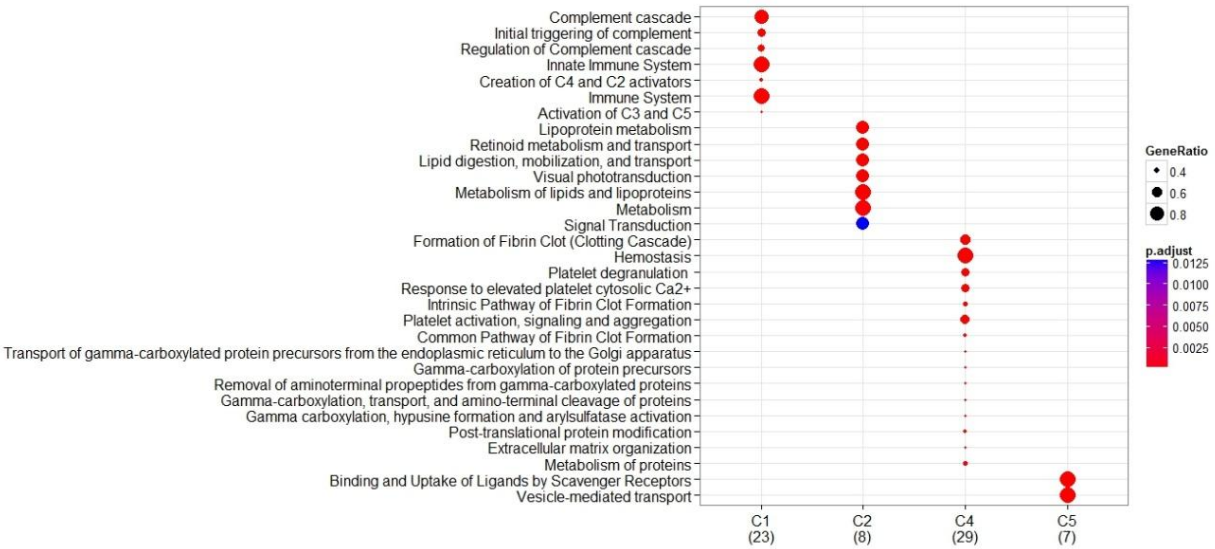
C5 : 3250, 2934, 10216, 3043, 922, 3240, 3039, 3040, 7184

```
> head((summary(res)))
  Cluster     ID                    Description GeneRatio  BgRatio       pvalue     p.adjust qvalue
1      C1 166658              Complement cascade     18/23  36/6750 2.333684e-39 1.633578e-38     NA
2      C1 166663 Initial triggering of complement     12/23  19/6750 3.644878e-27 1.275707e-26     NA
3      C1 977606 Regulation of Complement cascade     11/23  24/6750 1.003220e-22 2.340846e-22     NA
4      C1 168249            Innate Immune System     20/23 569/6750 3.346228e-19 5.855899e-19     NA
5      C1 166786 Creation of C4 and C2 activators      7/23  11/6750 6.353351e-16 8.894692e-16     NA
6      C1 168256                  Immune System     20/23 942/6750 7.674962e-15 8.954122e-15     NA
                                                                                    geneID Count
1          718/720/721/7448/722/3075/735/715/713/714/629/716/725/2220/5199/712/727/5648    18
2                                718/720/721/715/713/714/629/716/2220/5199/712/5648    12
3                                   718/720/721/7448/722/3075/735/629/725/5199/727    11
4 718/720/721/7448/722/3075/735/3929/715/713/714/629/716/725/2220/5199/1072/712/727/5648    20
5                                                 715/713/714/716/2220/712/5648     7
6 718/720/721/7448/722/3075/735/3929/715/713/714/629/716/725/2220/5199/1072/712/727/5648    20
```

Σχήμα 7.41: Αποτέλεσμα εμπλουτισμού των ομάδων των γονιδίων του Συνόλου Β στο περιβάλλον της R

Στο Σχήμα 7.41 παρουσιάζεται η αρχή του αποτελέσματος της ανάλυσης εμπλουτισμού των γονιδίων στο περιβάλλον της R όπου στην πρώτη γραμμή για την ομάδα Cluster C1 το μονοπάτι με ID 166658 με περιγραφή “Complement cascade” έχει ***GeneRatio*** 18/23 και ***BgRatio*** 36/6750 λαμβάνει ***pvalue*** $2.33 \cdot e^{-39}$ και προσαρμοσμένη τιμή ***padjust*** $1.63 \cdot e^{-38}$ με αντιστοιχιζόμενες geneID (EntrezID) 718/ 720/ 721/ 7448/ 722/ 3075/ 735/ 715/ 713/ 714/ 629/ 716/ 725/ 2220/ 5199/ 712/ 727/ 5658 με ***Count*** 18. Το συνολικό αποτέλεσμα παρουσιάζεται στον Πίνακα ΠΒ9.

Το γράφημα του αποτελέσματος της ανάλυσης αυτής απεικονίζεται στο Σχήμα 7.42..



Σχήμα 7.42: Σύγκριση των βιολογικών περιεχομένων της ανάλυσης εμπλουτισμού για τις 5 ομάδες γονιδίων του Συνόλου B

Στον άξονα **x** του διαγράμματος είναι τα μονοπάτια στα οποία συμμετέχουν τα γονίδια, ενώ στον άξονα **y** είναι οι ομάδες C1, C2, C4, C5. Η ομάδα C3 δεν εμφανίζεται στον άξονα αυτό, καθώς τα γονίδια τα οποία την αποτελούν δεν παρουσιάζουν βιολογικό ενδιαφέρον με βάση την ανάλυση εμπλουτισμού.

Η ομάδα C1 των γονιδίων στα εξής μονοπάτια με υψηλής τάξης εμπλουτισμό σε υψηλή αναλογία (GeneRatio):

- ✓ Αλληλουχία διαδικασιών του συμπληρώματος [63]
- ✓ Λειτουργία του ανοσοποιητικού συστήματος [48]

Η ομάδα C2 αντίστοιχα:

- ✓ Μεταβολισμός των λιπιδίων και των λιποπρωτεϊνών [8]
- ✓ Μεταβολισμός των ρετινοειδών και μεταφορά τους
- ✓ Πέψη των λιπιδίων, κινητοποίηση και μεταφοράς τους [8]
- ✓ Οπτικές αντιδράσεις φωτομετατροπής

Η ομάδα C4 :

- ✓ Σχηματισμός ινώδους θρόμβου [62]
- ✓ Αιμόσταση [62]
- ✓ Αποκοκκίωση των αιμοπεταλίων [62]

Τέλος, η ομάδα C5 συμμετέχει στα εξής μονοπάτια:

- ✓ Πρόσδεση και πρόσληψη των προσδετών από Scavenger υποδοχείς [64]
- ✓ Μεταφορά κυστιδίων

Αξίζει να σημειωθεί ότι στην βιβλιογραφία υπάρχουν πολλές αναφορές που σχετίζουν την τοξικότητα με τα γονίδια όλων των παραπάνω ομάδων με εξαίρεση τα γονίδια που μετέχουν στην λειτουργίας της μεταφοράς κυστιδίων, στις οπτικές αντιδράσεις φωτομετατροπής, στον μεταβολισμό των ρετινοειδών και στη μεταφορά τους.

7.5 topGO:περαιτέρω ανάλυση GO όρων

Η topGO βιβλιοθήκη έχει σχεδιαστεί για να διευκολύνει την ημι-αυτοματοποιημένη ανάλυση για τους όρους της Γονιδιακής Οντολογίας. Η διαδικασία περιλαμβάνει την είσοδο μετρήσεων της γονιδιακής έκφρασης, την διαφορική ανάλυση της έκφρασης, ανάλυση εμπλουτισμού των GO όρων, ερμηνεία και οπτικοποίηση των αποτελεσμάτων σε γραφήματα.

Ένα βασικό πλεονέκτημα της topGO είναι ότι παρέχει ένα ευρύ πλαίσιο δοκιμών για την γονιδιακή ανάλυση εμπλουτισμού. Επίσης επιτρέπεται η σύγκριση μεταξύ των διαφορετικών μεθόδων της ανάλυσης εμπλουτισμού. Για τους λόγους αυτούς παραθέτουμε εδώ αποτελέσματα συμπληρωματικά της ανάλυση που παρουσιάζουμε στο Κεφάλαιο 7.2 για την Γονιδιακή Οντολογία. Επίσης με την topGO αποσαφηνίζεται σε μεγάλο βαθμό το πυκνό τοπίο των γραφημάτων που παρουσιάσαμε στο Κεφάλαιο 6.

Υπάρχουν πολλά στατιστικά τεστ και αλγόριθμοι που διαχειρίζονται διαφορετικά τη δομή του GO γραφήματος. Ο Πίνακας 7.11 παρακάτω παρουσιάζει τη συμβατότητα μεταξύ των στατιστικών τεστ και των αλγορίθμων αυτών. [65]

Πίνακας 7.11: Αλγόριθμοι που υποστηρίζονται από την topGO και η συμβατότητα τους με τα στατιστικά τεστ **[65]**

	fisher	ks	t	globaltest	sum
classic	√	√	√	√	√
elim	√	√	√	√	√
weight	√	-	-	-	-
weight01	√	√	√	√	√
lea	√	√	√	√	√
parentchild	√	-	-	-	-

Εδώ θα αναλυθούν οι αλγόριθμοι classic, elim και weight, της GO δομής και τα στατιστικά τεστ Fisher και K-S, καθώς αυτές θα χρησιμοποιηθούν στην ανάλυση εμπλουτισμού .

Στο DAG γράφημα όπως αναφέρθηκε και σε προηγούμενα κεφάλαια, κάθε GO όρος αντιπροσωπεύεται από έναν κόμβο. Οι κόμβοι αντιστοιχούν σε ένα σετ γονιδίων. Διαφορετικά στατιστικά τεστ χρησιμοποιούνται για να βρεθεί η σημαντικότητα των GO όρων. Τα πιο συνηθισμένα τεστ υπολογίζουν τη σημαντικότητα ενός κόμβου, ανεξάρτητα από τη σημαντικότητα των γειτονικών κόμβων. Ανεξάρτητα από το στατιστικό τεστ που θα χρησιμοποιηθεί, αυτός ο τρόπος προσαρμογής του σκορ αναφέρεται ως classic μέθοδος. Στην συνέχεια παρουσιάζονται εναλλακτικές μέθοδοι.

Εάν ένας GO όρος περιλαμβάνει τα ίδια γονίδια με ένα από τα παιδιά του, η μέθοδος classic δίνει το ίδιο σκορ και στους δυο όρους. Σε αυτήν την περίπτωση το «παιδί» είναι βιολογικά πιο ενδιαφέρον μιας και η σχετιζόμενη περιγραφή του είναι πιο ειδική από την περιγραφή των προγόνων του. Για αυτό το λόγο μια πολλά υποσχόμενη ιδέα είναι να υπολογιστεί η σημαντικότητα ενός κόμβου με βάση την σημαντικότητα των παιδιών του. Η elim μέθοδος εφαρμόζει αυτή την ιδέα απομακρύνοντας όλα τα γονίδια που μεταφράζονται σε έναν σημαντικά εμπλουτισμένο κόμβο από όλους τους προγόνους τους. Η απομάκρυνση γονιδίων από έναν GO όρο μπορεί να θεωρηθεί ως ένα σύστημα στάθμισης όπου τα βάρη παίρνουν τιμές μηδέν ή ένα. Ο weight αλγόριθμος αποτελεί γενίκευση της ιδέας των βαρών στο διάστημα [0,1]. Για να αποφασιστεί αν ένας GO όρος ***u*** αντιπροσωπεύει καλύτερα τα πιο σημαντικά γονίδια, από οποιονδήποτε άλλο όρο στο γειτονικό του περιβάλλον στην GO δομή, το σκορ

εμπλουτισμού του κόμβου ***u*** συγκρίνεται με τα σκορ των παιδιών του. Τα παιδιά με ένα καλύτερο σκορ από τον ***u***, αντιπροσωπεύουν καλύτερα τα σημαντικά γονίδια. Τα γονίδια που μεταφράζονται σε αυτά τα παιδιά πρέπει να συνεισφέρουν λιγότερο στο σκορ οποιουδήποτε προγόνου του κόμβου ***u***. Έτσι σε αυτά τα γονίδια αποδίδονται μικρά βάρη σε όλους του απογόνους του κόμβου ***u***. Τα παιδιά με χαμηλότερο σκορ από το u, δεν πρέπει να λαμβάνονται ως σημαντικά. Για να επιτευχθεί αυτό, τα γονίδια τα οποία μεταφράζονται σε αυτά τα παιδιά λαμβάνουν μικρά βάρη, και το σκορ τους επαναϋπολογίζεται βασιζόμενο στα νέα βάρη τους.

Μέσω του classic αλγορίθμου υπολογίζεται μια σημαντικότητα για τον κάθε GO όρο ανεξάρτητα από τις σχέσεις τους. Έπειτα προσαρμόζονται οι τιμές ***p*** με την μέθοδο ελέγχου FDR. Μέσω των τιμών ***p*** που λαμβάνει ο κάθε όρος καθορίζεται η σημαντικότητα του κάθε γονιδίου με βάση το όριο αποκοπής της μεθόδου.

Πρώτη προσέγγιση : Αποκλεισμός γονιδίων

Η elim μέθοδος μελετά τους κόμβους στο GO γράφημα από κάτω προς τα πάνω. Ξεκινά να επεξεργάζεται τους κόμβους από το υψηλότερο επίπεδο (από κάτω) και μετά συνεχίζει με τους κόμβους στα χαμηλότερα επίπεδα. Μιας και οι κόμβοι του ίδιου επιπέδου δεν μοιράζονται καμία σύνδεση μεταξύ τους, εξετάζονται ανεξάρτητα. Η στρατηγική από κάτω προς τα πάνω θεωρεί ότι για κάθε εξεταζόμενο κόμβο, όλα τα παιδιά έχουν λάβει ένα σκορ.

Δεύτερη προσέγγιση: Στάθμιση των γονιδίων

Ο στόχος της στρατηγικής από κάτω προς τα πάνω που χρησιμοποιήθηκε στο αλγόριθμο elim είναι για να αναγνωριστούν οι κόμβοι με την πιο ειδική περιγραφή με την ελάχιστη απαιτούμενη σημαντικότητα. Ένας κόμβος θεωρείται σημαντικός αν η τιμή *p* είναι κάτω από ένα όριο . Πιο σημαντικοί κόμβοι σε υψηλότερα επίπεδα του γραφήματος μπορούν να μην ληφθούν υπόψη εξαιτίας του αποκλεισμού τους.

Μια εναλλακτική προτείνεται μέσω της weight μεθόδου. Εδώ τα σκορ σημαντικότητας των συνδεδεμένων κόμβων (ενός γονιού και ενός παιδιού) συγκρίνονται έτσι ώστε να εντοπιστεί τοπικά ο πιο σημαντικός κόμβος στο GO γράφημα. Αυτό επιτυγχάνεται με την υπο-στάθμιση των γονιδίων στους λιγότερο σημαντικούς γείτονες του. Το σκορ για μια ομάδα σταθμισμένων γονιδίων υπολογίζεται με την εφαρμογή του τεστ Fisher σε έναν σταθμισμένο πίνακα.

Στην μέθοδο του αλγορίθμου weight, οι κόμβοι επεξεργάζονται από κάτω προς τα πάνω όπως και στην μέθοδο elim. Παρόλα αυτά, για κάθε κόμβο ένα διάνυσμα των στάθμων των γονιδίων μνημονεύεται και ανανεώνεται κατά τη διάρκεια της μεθόδου. Αρχικά όλες οι στάθμες για τα γονίδια που μεταφράζονται σε έναν κόμβο θέτονται ίσες με το ένα. Ας θεωρήσουμε τον κόμβο $\boldsymbol{u}$. Η βασική αρχή της μεθόδου είναι να ενισχύσει τις διαφορές σε σημαντικότητα μεταξύ του $\boldsymbol{u}$ και των γειτόνων του. Αν ο κόμβος u είναι πιο σημαντικός, τα γονίδια που συμπεριλαμβάνονται στα παιδιά υποσταθμίζονται, αποδίδοντας μια μειωμένη σημαντικότητα στα παιδιά τους. Αν τουλάχιστον ένα παιδί είναι πιο σημαντικό από το $\boldsymbol{u}$ κόμβο, τα γονίδια που είναι κοινά για το παιδί και τον κόμβο $\boldsymbol{u}$ υποσταθμίζονται στον κόμβο $\boldsymbol{u}$ και στους απογόνους του, μειώνοντας αντίστοιχα την σημαντικότητα του κόμβου $\boldsymbol{u}$.

Η βασική λειτουργία του αλγορίθμου είναι να αποκλείει τα παιδιά του $\boldsymbol{u}$, τα οποία είναι πιο σημαντικά από το $\boldsymbol{u}$. Κάθε φορά που καλείται αυτή η λειτουργία, το σκορ του $\boldsymbol{u}$ επαναυπολογίζεται χρησιμοποιώντας τις ενημερωμένες στάθμες των γονιδίων. Με βάση το νέο σκορ, οι στάθμες για όλα τα παιδιά επαναϋπολογίζονται.

Πιο συγκεκριμένα, αν ένας κόμβος $\boldsymbol{u}$ έχει μικρότερη τιμή $\boldsymbol{p}$ από τα παιδιά του, και γι' αυτό θεωρείται σημαντικότερος, για κάθε παιδί οι παλιές στάθμες πολλαπλασιάζονται με τα βάρη που δίνει η λειτουργία $sigRatio()$. Αφού ενημερωθούν τα σκορ για όλα τα παιδιά, τελειώνει η διαδικασία για τον κόμβο $\boldsymbol{u}$.

Η εναλλακτική περίπτωση είναι τουλάχιστον ένα παιδί να έχει καλύτερο σκορ από τον κόμβο $\boldsymbol{u}$. Όλα τα γονίδια που μεταφράζονται σε κάθε ένα από τα πιο σχετικά παιδιά υποσταθμίζονται στους απογόνους του κόμβου $\boldsymbol{u}$, περιλαμβάνοντας και το ίδιο τον κόμβο. Μετά, οι στάθμες των γονιδίων πολλαπλασιάζονται με τον παράγοντα της λειτουργίας $sigRatio()$. Καλώντας πάλι την βασικότερη λειτουργία της μεθόδου για τον κόμβο $\boldsymbol{u}$ και τα λιγότερο σημαντικά παιδιά του, το σκορ για το $\boldsymbol{u}$, επαναυπολογίζεται με βάση τις νέες στάθμες. Λόγω των τροποποιημένων στάθμων για τον κόμβο $\boldsymbol{u}$, τα σκορ των παιδιών μπορεί να γίνει πιο σημαντικό, παρά το ενημερωμένο σκορ του $\boldsymbol{u}$. Έτσι η διαδικασία επαναλαμβάνεται μέχρι όλα τα απομένοντα παιδιά να έχουν χαμηλότερα σκορ από τον κόμβο $\boldsymbol{u}$. Μετά ο αλγόριθμος επεξεργάζεται τον επόμενο κόμβο. [66]

Μια ανάλυση με χρήση της βιβλιοθήκης αυτής μπορεί να χωριστεί σε τρία βήματα:

1. Προετοιμασία των δεδομένων: λίστα αναγνωριστικών όρων γονιδίων, σκορ γονιδίων, λίστα διαφορικά εκφρασμένων γονιδίων ή κριτήριο για την επιλογή γονιδίων με βάση τα σκορ τους, ακόμα και μεταφρασμένη λίστα γονιδίων και GO όρων συλλέγονται και αποθηκεύονται σε ένα αντικείμενο.
2. Εκτέλεση του τεστ εμπλουτισμού: Χρησιμοποιώντας το αντικείμενο που δημιουργήθηκε στο πρώτο βήμα, ο χρήστης μπορεί να εκτελέσει την ανάλυση εμπλουτισμού χρησιμοποιώντας οποιοδήποτε στατιστικό τεστ και αλγόριθμο.
3. Ανάλυση των αποτελεσμάτων: Τα αποτελέσματα του δεύτερου βήματος αναλύονται χρησιμοποιώντας λειτουργιές που δημιουργούν συγκεντρωτικούς πίνακες και λειτουργίες που οπτικοποιούν τα αποτελέσματα σε γραφήματα.

Η ανάλυση εμπλουτισμού της βιβλιοθήκης αυτής θα πραγματοποιηθεί για δυο σύνολα γονιδίων. Το πρώτο σύνολο γονιδίων θα είναι οι 76 συμβολικές ονομασίες των γονιδίων του Συνόλου Β καθώς και οι τιμές του $\boldsymbol{Q^2}$ του κάθε γονιδίου που εδώ θα χρησιμοποιηθεί ως σκορ. Το δεύτερο σύνολο γονιδίων θα είναι οι 76 συμβολικές ονομασίες των γονιδίων του Συνόλου Γ και οι τιμές ***VIP*** του κάθε γονιδίου θα χρησιμοποιηθεί και εδώ ως σκορ. Τα δύο Σύνολα περιλαμβάνουν τις ίδιες συμβολικές ονομασίες γονιδίων αλλά η ταξινόμηση των γονιδίων με βάση τις τιμές $\boldsymbol{Q^2}$ και ***VIP*** είναι διαφορετική. Τα γονίδια που εντοπίζονται στην κορυφή των δύο λιστών με τις μεγαλύτερες τιμές $\boldsymbol{Q^2}$ και ***VIP*** είναι τα AMBP, HABP2, ITIH2, TTHY και ITIH1.

7.5.1 Προετοιμασία δεδομένων

Το βασικότερο βήμα της ανάλυσης είναι η δημιουργία του R αντικειμένου "topGOdata". Αυτό το αντικείμενο θα μας παρέχει τις απαραίτητες πληροφορίες για την ανάλυση, όπως τα γονίδια, τα σκορ τους και το μέρος της γονιδιακής οντολογίας που θα χρησιμοποιηθεί στην ανάλυση.

Η πρώτη λειτουργία της βιβλιοθήκης που θα χρησιμοποιηθεί απαιτεί:

- ✓ Τάξη του αντικειμένου "topGOdata", που θα δημιουργηθεί
- ✓ Οντολογία της ανάλυσης
- ✓ Ονοματισμένο διάνυσμα με τις ταυτότητες των γονίδιων και τα αντίστοιχα σκορ τους
- ✓ Λειτουργία που καθορίζει ποια γονίδια είναι πιο σημαντικά με κριτήριο τα σκορ τους

- ✓ Ακέραια παράμετρος μεγαλύτερη ή ίση του 1 μέχρι το 10, που αποκλείει από την ανάλυση τους GO όρους που αντιστοιχίζονται σε λιγότερα σε αριθμό γονίδια που ορίζει αυτός ο ακέραιος.
- ✓ Λειτουργία που μεταφράζει τις ταυτότητες των γονιδίων σε GO όρους, με χρήση μεταφραστικών πακέτων της μορφής "org.XX.XX"
- ✓ Το μεταφραστικό πακέτο "org.Hs.eg.db", που θα χρησιμοποιηθεί από την παραπάνω λειτουργία
- ✓ Είδος της ταυτότητας των γονιδίων

Εφόσον η ανάλυση εμπλουτισμού θα πραγματοποιηθεί για τα παραπάνω δύο σύνολα γονιδίων, ανάλογα θα έχουμε και τη δημιουργία δύο αντικειμένων:

1. Το πρώτο αντικείμενο που δημιουργείται για οντολογία BP, θα περιέχει ονοματισμένο διάνυσμα με τις συμβολικές ονομασίες των γονιδίων με αντίστοιχο σκορ τους το $\boldsymbol{Q^2}$. Το κριτήριο επιλογής των σημαντικότερων γονιδίων είναι η τιμή $\boldsymbol{Q^2}$ να είναι μεγαλύτερη ή ίση του 0,759. Η ακέραια παράμετρος θα είναι ίση με 2, δηλαδή αποκλείονται οι GO όροι που αντιστοιχίζονται με λιγότερα από δύο γονίδια. Η λειτουργία της μετάφρασης θα γίνει μέσω της βιβλιοθήκης "org.Hs.eg.db" για να μεταφραστούν οι συμβολικές ονομασίες των γονιδίων σε GO ταυτότητες βιολογικών λειτουργιών.
2. Το δεύτερο αντικείμενο δημιουργείται και αυτό για οντολογία BP. Εδώ το ονοματισμένο διάνυσμα θα περιλαμβάνει τις συμβολικές ονομασίες των γονιδίων με αντίστοιχο σκορ τους το ***VIP***. Το κριτήριο επιλογής των σημαντικότερων γονιδίων είναι το σκορ τους ***VIP*** να είναι μεγαλύτερο ή ίσο του 0,8. Η ακέραια παράμετρος θα είναι και εδώ ίση με 2 και η μεταφραστική βιβλιοθήκη θα είναι η "org.Hs.eg.db".

Στα Σχήματα 7.43 και 7.44 παρουσιάζονται οι περιγραφές σε περιβάλλον R των παραπάνω δύο αντικειμένων αντίστοιχα.

```
> GOdata

------------------------ topGOdata object ------------------------

 Description:
   -  Test

 Ontology:
   -  BP

 76 available genes (all genes from the array):
   - symbol:  AMBP HABP2 ITIH2 TTHY ITIH1  ...
   - score :  1 1 1 1 1  ...
   - 43  significant genes.

 40 feasible genes (genes that can be used in the analysis):
   - symbol:  AMBP HABP2 ITIH2 ITIH1 APOA4  ...
   - score :  1 1 1 1 0.819  ...
   - 24  significant genes.

 GO graph (nodes with at least  2  genes):
   - a graph with directed edges
   - number of nodes = 816
   - number of edges = 1769

------------------------ topGOdata object ------------------------
```

Σχήμα 7.43: Περιγραφή αντικειμένου topGOdata, που περιλαμβάνει ως σκορ το Q^2, στο περιβάλλον της R

Στο Σχήμα 7.43, το πρώτο αντικείμενο περιγράφεται ως Test, η οντολογία στην οποία χτίζεται είναι η Biological Process (BP). Τα διαθέσιμα γονίδια (available genes) είναι 76 με συμβολικές ονομασίες τις AMBP, HABP2, ITIH2, TTHY, ITIH1… με αντίστοιχα σκορ (score) 1,1,1,1,…, εκ των οποίων με βάση το κριτήριο επιλογής των σημαντικότερων τα 43 είναι τα πιο σημαντικά (significant genes). Ενώ τα «εφικτά» γονίδια (feasible genes), τα οποία μπορούν να χρησιμοποιηθούν στην ανάλυση είναι 40, με συμβολικές ονομασίες τις AMBP, HABP2, ITIH2, ITIH1, APOA4..,με αντίστοιχα σκορ 1,1,1,1,0.819…, εκ των οποίων τα 24 λαμβάνονται ως σημαντικά. Τέλος περιγράφεται το υπογράφημα της GO δομής (GO graph) που προκύπτει από τους GO όρους που αντιστοιχίζονται σε τουλάχιστον 2 γονίδια (nodes with at least 2 genes), το οποίο είναι ένα γράφημα με κατευθυνόμενες συνδέσεις με 816 κόμβους και 1769 συνδέσεις.

```
> GOdata

------------------------ topGOdata object ------------------------

 Description:
   -  Test

 Ontology:
   -  BP

 76 available genes (all genes from the array):
   - symbol:  AMBP ITIH2 ITIH1 A1AT HABP2  ...
   - score :  1.7 1.66 1.63 1.6 1.61  ...
   - 43  significant genes.

 40 feasible genes (genes that can be used in the analysis):
   - symbol:  AMBP ITIH2 ITIH1 HABP2 ITIH3  ...
   - score :  1.7 1.66 1.63 1.61 1.51  ...
   - 23  significant genes.

 GO graph (nodes with at least  2  genes):
   - a graph with directed edges
   - number of nodes = 816
   - number of edges = 1769

------------------------ topGOdata object ------------------------
```

Σχήμα 7.44: Περιγραφή αντικειμένου topGOdata, που περιλαμβάνει ως σκορ το VIP, στο περιβάλλον της R

7.5.2 Εκτέλεση του τεστ εμπλουτισμού

Η βιβλιοθήκη topGO όπως αναφέρθηκε και παραπάνω σχεδιάστηκε για να εκτελεί διαφορετικά στατιστικά τεστ και διαφορετικούς αλγορίθμους (Πίνακα 7.11) Υπάρχουν τρία είδη στατιστικών τεστ που μπορούν να χρησιμοποιηθούν και αυτά είναι :

1. Τεστ με βάση τις μετρήσεις των γονιδίων. Αυτό το είδος τεστ είναι το πιο διαδεδομένο και απαιτεί μόνο την λίστα των γονιδίων προς ανάλυση. Τέτοια τεστ είναι το τεστ του Fisher, το υπεργεωμετρικό μοντέλο και το διωνυμικό μοντέλο.
2. Τεστ με βάση τα σκορ των γονιδίων ή την ταξινόμηση των γονιδίων. Αυτά περιλαμβάνουν το τεστ Kolmogorov-Smirnov, το t-test κ.ά.
3. Τεστ με βάση την γονιδιακή έκφραση όπως το Goerman's globaltest.

Σε συνδυασμό με τους αλγορίθμους που διαχειρίζονται τις σχέσεις των GO όρων στην δομή του GO γραφήματος, πραγματοποιείται το τεστ εμπλουτισμού.

Η λειτουργία του τεστ εμπλουτισμού απαιτεί:

- ✓ Το αντικείμενο που δημιουργήθηκε στο πρώτο βήμα που περιλαμβάνει τα δεδομένα της ανάλυσης
- ✓ Στατιστικό τεστ, υπολογισμού σημαντικότητας των γονιδίων μέσω της τιμής ***p***
- ✓ Αλγόριθμος προσαρμογής σκορ με κριτήριο την τιμή ***p*** και τις σχέσεις των GO όρων

Σε αυτή την ανάλυση εμπλουτισμού δεν χρησιμοποιείται η μέθοδος προσαρμογής των ***p*** τιμών. Σε πολλές περιπτώσεις οι τιμές ***p*** που υπολογίζονται δεν είναι τόσο μεγάλες. Η μέθοδος ελέγχου FDR μπορεί να προσαρμόσει πολύ συντηρητικές τιμές ***p*** και να μειώσει τους GO όρους που θα θεωρηθούν σημαντικοί. Υπάρχει περίπτωση σημαντικοί όροι να μην ξεπεράσουν το όριο αποκοπής της μεθόδου και να αποκλειστούν από την ανάλυση και να χαθεί με αυτόν τον τρόπο χρήσιμο μέρος της βιολογικής πληροφορίας. Επίσης μέσω των αλγορίθμων elim και weight, ορίζεται η τιμή ***p*** ενός GO όρου με κριτήριο τους γειτονικούς όρους του. Για αυτό το λόγο και τα τεστ αυτά δεν είναι ανεξάρτητα και οι τιμές ***p*** που υπολογίζονται θεωρούνται διορθωμένες και δεν απαιτείται προσαρμογής τους .

Στην δική μας ανάλυση θα πραγματοποιηθούν τεστ εμπλουτισμού για τα δύο σύνολα γονιδίων για κάθε έναν από τους εξής συνδυασμούς στατιστικών τεστ και αλγορίθμων προσαρμογής:

1. Στατιστικό τεστ Fisher με classic αλγόριθμο προσαρμογής σκορ
2. Στατιστικό τεστ Kolmogorov-Smirnov με classic αλγόριθμο προσαρμογής σκορ
3. Στατιστικό τεστ Kolmogorov-Smirnov με elim αλγόριθμο προσαρμογής σκορ
4. Στατιστικό τεστ Fisher με weight αλγόριθμο προσαρμογής σκορ

Για κάθε τεστ εμπλουτισμού ανάλογα με το στατιστικό τεστ και τον αλγόριθμο προσαρμογής, θα προκύπτουν διαφορετικοί υπερεκφρασμένοι ή αλλιώς εμπλουτισμένοι GO όροι.

```
> resultFisher

Description: Test
Ontology: BP
'classic' algorithm with the 'fisher' test
816 GO terms scored: 1 terms with p < 0.01
Annotation data:
    Annotated genes: 40
    Significant genes: 23
    Min. no. of genes annotated to a GO: 2
    Nontrivial nodes: 734
```

Σχήμα 7.45: Αποτέλεσμα στατιστικού τεστ Fisher με χρήση αλγορίθμου classic στο περιβάλλον της R

Το Σχήμα 7.45 περιγράφει το αποτέλεσμα του τεστ εμπλουτισμού με χρήση του στατιστικού τεστ Fisher με classic αλγόριθμο προσαρμογής σκορ. Όπως παρατηρούμε το αποτέλεσμα περιγράφεται ως Test σε οντολογία BP. Προσαρμόστηκε σκορ σε 816 GO όρους με έναν μόλις όρο να λαμβάνει τιμή ***p*** μικρότερη του 0.01. Τα δεδομένα μετάφρασης περιλαμβάνουν ότι μεταφράστηκαν 40 γονίδια, εκ των οποίων τα 23 είναι σημαντικά. Εδώ επισημαίνεται ότι στην ανάλυση συμπεριλαμβάνονται οι GO όροι που αντιστοιχίζονται σε τουλάχιστον 2 γονίδια και εντοπίζονται 734 κάποιας σημαντικότητας GO όροι .

7.5.2.1 Ανάλυση των αποτελεσμάτων

Αφού πραγματοποιηθεί το τεστ της ανάλυσης εμπλουτισμού των γονιδίων υπάρχουν τρεις λειτουργίες που οπτικοποιούν τα αποτελέσματά της. Αυτά είναι:

1. Γενικός πίνακας αποτελεσμάτων ανά τεστ
2. Γράφημα σύγκρισης των τιμών p για διαφορετικούς αλγορίθμους
3. Υπογράφημα της μορφής DAG, που προκύπτει από τους δέκα πιο σημαντικούς GO όρους και τις σχέσεις του με τους υπόλοιπους όρους του γραφήματος

Για την δημιουργία του γενικού πίνακα αποτελεσμάτων απαιτούνται:

- ✓ Το αντικείμενο που δημιουργήθηκε στο πρώτο βήμα που περιλαμβάνει τα δεδομένα της ανάλυσης
- ✓ Τα αποτελέσματα των τεστ που θέλουμε να παρουσιαστούν
- ✓ Η κατάταξη των GO όρων με ελλατούμενη σημαντικότητα με βάση την μέθοδο της επιλογής μας

- ✓ Η θέση κατάταξης των γονιδίων με ελλατούμενη σημαντικότητα με βάση διαφορετική μέθοδο από την παραπάνω, ώστε να είναι δυνατή η άμεση σύγκριση
- ✓ Ο αριθμός των κόμβων ξεκινώντας από την αρχή της λίστας κατάταξης για του οποίους θα παρουσιαστούν τα αποτελέσματα.

Ο γενικός πίνακας αποτελεσμάτων επιλέγουμε να έχει να αποτελέσματα των τεστ classicFisher, classicKS, elimKS. Η κατάταξη των GO όρων σε αυτόν θα γίνει με ελλατούμενη σημαντικότητα με βάση το αποτέλεσμα του τεστ elimKS και θα περιλαμβάνει την θέση κατάταξης των GO όρων με βάση το αποτέλεσμα του τεστ classicFisher. Οι παρακάτω Πίνακας 7.12 και 7.13 είναι οι γενικοί πίνακες αποτελεσμάτων για την ανάλυση εμπλουτισμού λαμβάνοντας ως σκορ το $\boldsymbol{Q^2}$ και το ***VIP*** αντίστοιχα.

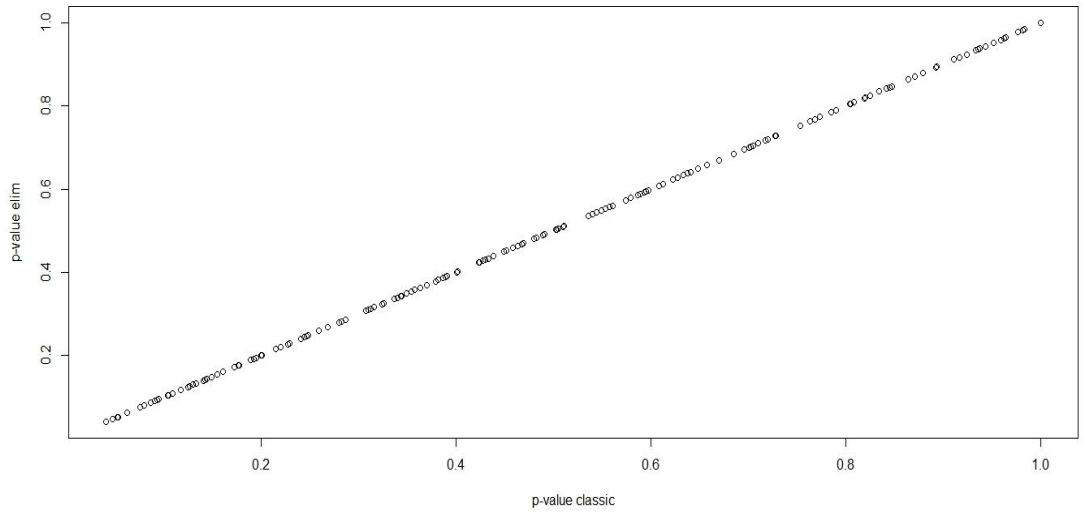
Πίνακας 7.12: Γενικός πίνακας αποτελεσμάτων για στατιστικά τεστ λαμβάνοντας ως σκορ το Q^2

GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
GO:0065008	regulation of biological quality	17	8	10.2	786	0.96	0.041	0.041
GO:0002520	immune system development	2	0	1.2	803	1	0.048	0.048
GO:0030097	hemopoiesis	2	0	1.2	804	1	0.048	0.048
GO:0048534	Hematopoietic or lymphoid organ devel..	2	0	1.2	805	1	0.048	0.048
GO:0090066	regulation of anatomical structure size	4	1	2.4	800	0.98	0.053	0.053
GO:0050896	response to stimulus	31	17	18.6	785	0.95	0.053	0.053
GO:0042060	wound healing	11	4	6.6	801	0.99	0.063	0.063
GO:0009617	response to bacterium	5	1	3	802	0.99	0.076	0.076
GO:0017187	peptidyl-glutamic acid carboxylation	2	0	1.2	806	1	0.08	0.08
GO:0018200	peptidyl-glutamic acid modification	2	0	1.2	807	1	0.08	0.08
GO:0018214	protein carboxylation	2	0	1.2	808	1	0.08	0.08
GO:0044699	single-organism process	32	19	19.2	612	0.71	0.086	0.086
GO:0006898	receptor-mediated endocytosis	10	6	6	497	0.65	0.091	0.091
GO:0010742	macrophage derived foam cell differentia...	2	0	1.2	809	1	0.094	0.094
GO:0010743	regulation of macrophage derived foam ce...	2	0	1.2	810	1	0.094	0.094
GO:0090077	foam cell differentiation	2	0	1.2	811	1	0.094	0.094
GO:0030193	regulation of blood coagulation	6	2	3.6	792	0.97	0.095	0.095
GO:0050818	regulation of coagulation	6	2	3.6	793	0.97	0.095	0.095
GO:0061041	regulation of wound healing	6	2	3.6	794	0.97	0.095	0.095
GO:1900046	regulation of hemostasis	6	2	3.6	795	0.97	0.095	0.095

Πίνακας 7.13: Γενικός πίνακας αποτελεσμάτων για στατιστικά τεστ λαμβάνοντας ως σκορ το VIP

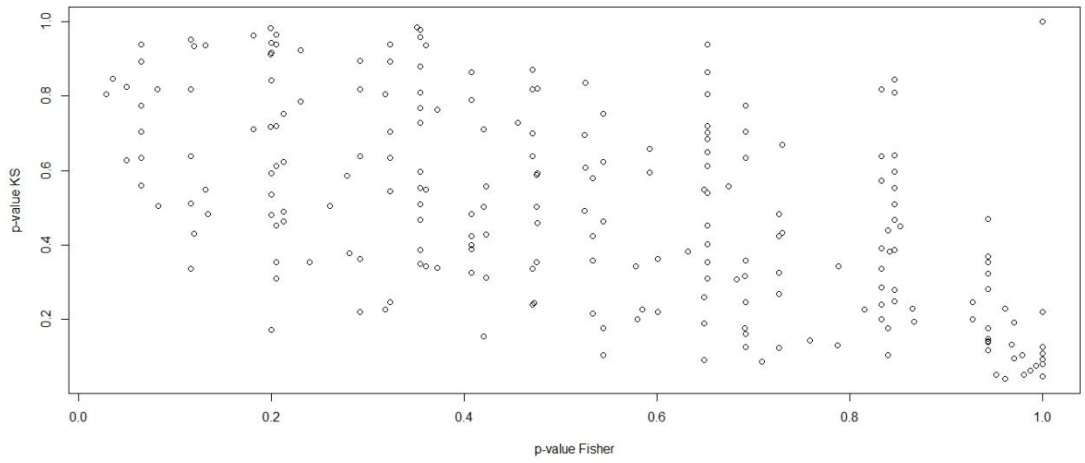
GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
GO:0006958	complement activation, classical pathway	8	2	4.6	717	0.99	0.0013	0.0013
GO:0045087	innate immune response	12	3	6.9	733	1	0.0033	0.0033
GO:0006909	phagocytosis	8	2	4.6	718	0.99	0.011	0.011
GO:0002697	regulation of immune effector process	4	0	2.3	734	1	0.0128	0.0128
GO:0044699	single-organism process	32	18	18.4	363	0.76	0.0174	0.0296
GO:0065008	regulation of biological quality	17	10	9.77	219	0.57	0.0321	0.0321
GO:0002703	regulation of leukocyte mediated immunit...	2	0	1.15	735	1	0.033	0.033
GO:0002704	negative regulation of leukocyte mediate...	2	0	1.15	736	1	0.033	0.033
GO:0002706	regulation of lymphocyte mediated immuni...	2	0	1.15	737	1	0.033	0.033
GO:0002707	negative regulation of lymphocyte mediat...	2	0	1.15	738	1	0.033	0.033
GO:0002712	regulation of B cell mediated immunity	2	0	1.15	739	1	0.033	0.033
GO:0002713	negative regulation of B cell mediated i...	2	0	1.15	740	1	0.033	0.033
GO:0002819	regulation of adaptive immune response	2	0	1.15	741	1	0.033	0.033
GO:0002820	negative regulation of adaptive immune r...	2	0	1.15	742	1	0.033	0.033
GO:0002822	regulation of adaptive immune response b...	2	0	1.15	743	1	0.033	0.033
GO:0002823	negative regulation of adaptive immune r...	2	0	1.15	744	1	0.033	0.033
GO:0002889	regulation of immunoglobulin mediated im...	2	0	1.15	745	1	0.033	0.033
GO:0002890	negative regulation of immunoglobulin me...	2	0	1.15	746	1	0.033	0.033
GO:0002920	regulation of humoral immune response	2	0	1.15	747	1	0.033	0.033
GO:0002921	negative regulation of humoral immune re...	2	0	1.15	748	1	0.033	0.033

Για την σύγκριση τιμών ***p*** απαιτούνται οι τιμές ***p*** για δυο διαφορετικά στατιστικά τεστ και παρουσιάζεται σε ένα γράφημα x,y. Μια τέτοια σύγκριση παρουσιάζεται στο παρακάτω γράφημα. (Σχήμα 7.46)



Σχήμα 7.46: Σύγκριση τιμών p μεταξύ classic και elim αλγορίθμου για το στατιστικό τεστ KS

Εδώ έχουμε επιλέξει να γίνει η σύγκριση μεταξύ των τιμών ***p*** της μεθόδου KS με classic αλγόριθμο με τις τιμές της μεθόδου KS με elim αλγόριθμο. Όπως παρατηρούμε και στον παραπάνω πίνακα, έτσι επιβεβαιώνεται και εδώ ότι οι τιμές αυτές συμπίπτουν. Αντίθετα στο παρακάτω γράφημα παρουσιάζεται η σύγκριση των τιμών p της μεθόδου Fisher με classic αλγόριθμο με τις τιμές της μεθόδου KS με classic αλγόριθμο, όπου οι τιμές ***p*** των δύο τεστ παρουσιάζουν μεγάλες διαφορές. Επομένως για τα γονίδια της ανάλυσης η χρήση διαφορετικών στατιστικών τεστ θα παίξει το σημαντικότερο ρόλο στο αποτέλεσμα του εμπλουτισμού και όχι ο αλγόριθμος προσαρμογής σκορ.

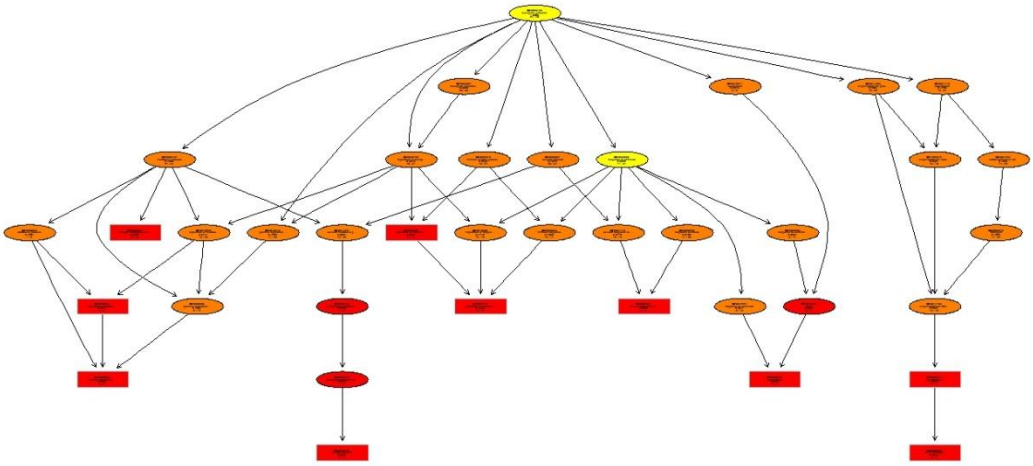


Σχήμα 7.47: Σύγκριση τιμών p μεταξύ στατιστικού τεστ KS και Fisher για classic αλγόριθμο

Για την δημιουργία των υπογραφημάτων της GO δομής απαιτούνται:

- ✓ Το αντικείμενο που δημιουργήθηκε στο πρώτο βήμα που περιλαμβάνει τα δεδομένα της ανάλυσης
- ✓ Τα σκορ της μεθόδου ανάλυσης εμπλουτισμού της επιλογής μας
- ✓ Τον αριθμό των σημαντικότερων GO όρων που θα παρασταθούν στο γράφημα.

Τα υπογραφήματα που δημιουργούνται, με βάση το αποτέλεσμα της ανάλυσης εμπλουτισμού για κάθε στατιστικό τεστ, παρουσιάζουν τα σημαντικότερους εμπλουτισμένους GO όρους και τις σχέσεις μεταξύ τους στην GO δομή. Το χρώμα του κάθε κόμβου αντιπροσωπεύει τη σημαντικότητα του. Η κλίμακα των χρωμάτων κυμαίνεται μεταξύ του κόκκινου και του κίτρινου, αυξανόμενης της τιμής p. Για την ευκολότερη επεξεργασία τους επιλέγεται οι σημαντικότεροι όροι να αντιπροσωπεύονται με τετράγωνα ενώ οι υπόλοιποι από κύκλους. Στην δική μας ανάλυση επιλέγεται να εμφανίζονται με τετράγωνα οι δέκα σημαντικότεροι GO όροι.



Σχήμα 7.48: Υπογράφημα GO δομής με χρήση κλίμακας p για το στατιστικό τεστ Fisher με classic αλγόριθμο

Το παραπάνω γράφημα (Σχήμα 7.48) προκύπτει από την ανάλυση εμπλουτισμού με χρήση της μεθόδου Fisher με classic αλγόριθμο. Τα αντίστοιχα γραφήματα για τις αναλύσεις με διαφορετικές μεθόδου και αλγορίθμους περιλαμβάνονται στα Σχήματα ΠΒ1-ΠΒ2.

Στους παρακάτω πίνακες παρουσιάζονται οι 10 σημαντικότεροι GO όροι του κάθε γραφήματος με βάση τα αποτελέσματα της ανάλυσης εμπλουτισμού για κάθε στατιστικό τεστ και για τα δυο σύνολα γονιδίων και σκορ.

Πίνακας 7.14: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο classic

GOID	Λειτουργική κατηγορία
GO:0016310	φωσφορυλίωση (phosphorylation)
GO:0006820	μεταφορά ανιόντων (anion transport)
GO:0006811	μεταφορά ιόντων (ion transport)
GO:0006935	χημειοταξία (chemotaxis)
GO:0009896	θετική ρύθμιση καταβολικής διεργασίας (positive regulation of catabolic process)
GO:0009894	ρύθμιση της καταβολικής διεργασίας (regulation of catabolic process)
GO:0006807	μεταβολική διεργασία ένωσης αζώτου (nitrogen compound metabolic process)
GO:0002682	ρύθμιση του ανοσοποιητικού συστήματος (regulation of immune system)

GO:0050776	ρύθμιση της ανοσοαπόκρισης (regulation of immune response)
GO:0033354	κύκλος της χλωροφύλλης (chlorophyll cycle)

Πίνακας 7.15: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ KS με αλγόριθμο classic με σκορ το Q^2

GOID	Λειτουργική κατηγορία
GO:0002520	ανάπτυξη του ανοσοποιητικού συστήματος (immune system development)
GO:0048534	αιματοποιητική ή λεμφοειδής ανάπτυξη οργάνου (hematopoietic or lymphoid organ development)
GO:0030097	αιμοποίηση (hemopoiesis)
GO:0018200	τροποποίηση πεπτιδυλ-γλουταμινικού οξέος (peptidyl-glutamic acid modification)
GO:0017187	καρβοξυλίωση πεπτιδυλ-γλουταμινικού οξέος (peptidyl-glutamic acid carboxylation)
GO:0009617	απόκριση στα βακτήρια (response to bacterium)
GO:0090066	ρύθμιση του μεγέθους της ανατομικής δομής (regulation of anatomical structure size)
GO:0065008	ρύθμιση της βιολογικής ποιότητας (regulation of biological quality)
GO:0042060	επούλωση πληγών (wound healing)
GO:0050896	απόκριση σε ερέθισμα

Πίνακας 7.16: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ KS με αλγόριθμο elim με σκορ το Q^2

GOID	Λειτουργική κατηγορία
GO:0002520	ανάπτυξη του ανοσοποιητικού συστήματος (immune system development)
GO:0048534	αιματοποιητική ή λεμφοειδής ανάπτυξη οργάνου (hematopoietic or lymphoid organ development)
GO:0030097	αιμοποίηση
GO:0018214	καρβοξυλίωση πρωτεΐνης (protein carboxylation)
GO:0017187	καρβοξυλίωση πεπτιδυλ-γλουταμινικού οξέος (peptidyl-glutamic acid carboxylation)
GO:0009617	απόκριση σε βακτήρια (response to bacterium)
GO:0090066	ρύθμιση του μεγέθους της ανατομικής δομής (regulation of anatomical

	structure size)
GO:0065008	ρύθμιση της βιολογικής ποιότητας
GO:0050896	απόκριση σε ερέθισμα
GO:0042060	επούλωση πληγών

Πίνακας 7.17: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο weight

GOID	Λειτουργική κατηγορία
GO:0070328	ομοιόσταση τριγλυκυρεδίων (triglyceride homeostasis)
GO:0009968	αρνητική ρύθμιση της μεταγωγής σήματος (negative regulation of signal transduction)
GO:0010873	θετική ρύθμιση της εστεροποίησης της χοληστερίνης (positive regulation of cholesterol esterification)
GO:0033700	εκροή φωσφολιπιδίων (phospholipid efflux)
GO:0034372	αναδιαμόρφωση σωματιδίων λιποπρωτεΐνης χαμηλής πυκνότητας (very-low-density lipoprotein particle remodeling)
GO:0046470	μεταβολική διεργασία φωσφατιδυλοχολίνης (phosphatidylcholine metabolic process)
GO:0006869	μεταφορά λιπιδίων
GO:0051172	αρνητική ρύθμιση της μεταβολικής διεργασίας ένωσης αζώτου (negative regulation of nitrogen compound metabolic process)
GO:0045087	έμφυτη ανοσοαπόκριση (innate immune response)
GO:0007155	προσκόλληση κυττάρων

Στους Πίνακες 7.14-7.17 παρουσιάζονται οι 10 υψηλού εμπλουτισμού GO όροι που προκύπτουν από τα τεστ Fisher-classic, KS-classic/elim, Fisher-weight, με βάση την ανάλυση εμπλουτισμού για τα 76 γονίδια λαμβάνοντας το Q^2 ως σκορ τους,. Οι εμπλουτισμένοι όροι για τα στατιστικά τεστ KS-classic και KS-elim είναι οι ίδιοι καθώς όπως παρατηρούμε ο elim αλγόριθμος δεν οδηγεί σε διαφορετικά αποτελέσματα σε σχέση με τον classic αλγόριθμο. Αντίθετα ο weight αλγόριθμος στο τεστ Fisher δίνει διαφορετικά αποτελέσματα από τον classic αλγόριθμο στο ίδιο τεστ.

Πίνακας 7.18: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο classic

GOID	Λειτουργική κατηγορία
GO:0061045	αρνητική ρύθμιση επούλωσης πληγών (negative regulation of wound healing)
GO:0030195	αρνητική ρύθμιση πήξης του αίματος (negative regulation of blood coagulation)
GO:0050819	αρνητική ρύθμιση πήξης (negative regulation of coagulation)
GO:1900047	αρνητική ρύθμιση αιμόστασης (negative regulation of hemostasis)
GO:0030162	ρύθμιση της πρωτεόλυσης (regulation of proteolysis)
GO:0043086	αρνητική ρύθμιση καταλυτικής δραστηριότητας (negative regulation of catalytic activity)
GO:0006508	πρωτεόλυση (proteolysis)
GO:0044260	μεταβολική διεργασία μακρομορίου (Cellular macromolecule metabolic process)
GO:0044092	αρνητική ρύθμιση της μοριακής λειτουργίας (negative regulation of molecular function)
GO:0044237	κυτταρική μεταβολική διεργασία

Πίνακας 7.19: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ KS με αλγόριθμο classic με σκορ το VIP

GOID	Λειτουργική κατηγορία
GO:0002443	ανοσία που προκαλείται από λευκοκύτταρα (leukocyte mediated immunity)
GO:0002449	ανοσία που προκαλείται από λεμφοκύτταρα (lymphocyte mediated immunity)
GO:0019724	ανοσία που προκαλείται από Β κύτταρα (B cell mediated immunity)
GO:0016064	ανοσοαπόκριση ανοσοσφαιρίνης (immunoglobulin mediated immune response)
GO:0002253	ενεργοποίηση της ανοσολογικής απόκρισης (activation of immune response)
GO:0006956	ενεργοποίηση συμπληρώματος
GO:0002460	προσαρμοστική ανοσολογική απάντηση που βασίζεται στον σωματικό ανασυνδυασμό των ανοσοποιητικών υποδοχέων που χτίστηκε από την υπεροικογένεια της ανοσοσφαιρίνης (adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains)
GO:0002455	χυμική ανοσολογική απόκριση που προκαλείται από κυκλοφορούντα μόρια ανοσοσφαιρίνης

GO:0006958	κλασσικό μονοπάτι ενεργοποίησης συμπληρώματος
GO:0050778	θετική ρύθμιση ανοσολογικής απόκρισης (positive regulation of immune response)

Πίνακας 7.20 Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ KS με αλγόριθμο elim με σκορ το VIP

GOID	Λειτουργική κατηγορία
GO:0045959	αρνητική ρύθμιση της ενεργοποίησης του συμπληρώματος,κλασσικό μονοπάτι (negative regulation of complement activation, classical pathway)
GO:0045916	αρνητική ρύθμιση της ενεργοποίησης του συμπληρώματος (negative regulation of complement activation)
GO:0006958	ενεργοποίησης του συμπληρώματος, κλασσικό μονοπάτι
GO:0002713	αρνητική ρύθμιση της ανοσίας που προκαλείται από Β κύτταρα (negative regulation of B cell mediated immunity)
GO:0002712	ρύθμιση της ανοσίας που προκαλείται από Β κύτταρα (regulation of B cell mediated immunity)
GO:0006909	φαγοκυττάρωση (phagocytosis)
GO:0045087	έμφυτη ανοσοαπόκριση
GO:0002697	ρύθμιση ανοσοποιητικής διεργασίας ενός τελεστή (regulation of immune effector process)
GO:0065008	ρύθμιση της βιολογικής ποιότητας
GO:0044699	διεργασία ενός οργανισμού

Πίνακας 7.21: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο weight

GOID	Λειτουργική κατηγορία
GO:0010951	αρνητική ρύθμισης της δραστηριότητας της ενδοπεπτιδάσης
GO:0030195	αρνητική ρύθμιση της πήξης του αίματος
GO:0042311	αγγειοδιαστολή (vasodilation)
GO:0051241	αρνητική ρύθμιση διεργασίας πολυκύτταρου οργανισμού (negative regulation of multicellular organismal process)
GO:0065008	ρύθμιση της βιολογικής ποιότητας
GO:0006898	ενδοκυττάρωση με μεσολάβηση υποδοχέα
GO:0042157	μεταβολική διεργασία λιποπρωτεϊνών (lipoprotein metabolic process)
GO:1901565	καταβολική διεργασία ένωσης οργανικού αζώτου (organonitrogen compound

	catabolic process)
GO:0044248	κυτταρική καταβολική διεργασία
GO:0044712	καταβολική διεργασία ενός οργανισμού

Στους Πίνακες 7.18-7.21 παρουσιάζονται οι δέκα υψηλής τάξης εμπλουτισμού GO όροι για κάθε στατιστικό τεστ της ανάλυσης εμπλουτισμού που χρησιμοποιεί ως σκορ το μέγεθος VIP. Εδώ παρατηρούμε ότι όλοι οι GO όροι του αποτελέσματος διαφέρουν για όλα τα στατιστικά τεστ.

Τα παραπάνω στατιστικά τεστ χρησιμοποιούν ως δεδομένα τις 76 συμβολικές ονομασίες των γονιδίων ενώ στην πρώτη περίπτωση λαμβάνεται ως σκορ το $\boldsymbol{Q^2}$ ενώ στην δεύτερη περίπτωση το ***VIP***. Όπως παρατηρείται παρόλο που τα πέντε σημαντικότερα γονίδια με το υψηλότερο σκορ και στις δυο περιπτώσεις είναι τα ίδια, το αποτέλεσμα της ανάλυσης εμπλουτισμού φαίνεται να διαμορφώνεται διαφορετικά. Αυτό συμβαίνει διότι η ταξινόμηση των υπόλοιπων 71 γονιδίων με βάση το σκορ τους είναι αρκετά διαφορετική. Ακόμη οι τιμές για το $\boldsymbol{Q^2}$ κυμαίνονται στο διάστημα [0.6349,1] ενώ οι τιμές για το ***VIP*** στο διάστημα [0.6,1.67]. διαπιστώνεται λοιπόν ότι και η κλίμακα των τιμών που λαμβάνονται ως σκορ διαφέρουν αρκετά μεταξύ τους προσδίδοντας στα γονίδια διαφορετικής τάξης σημαντικότητα κατά την ανάλυση εμπλουτισμού.

7.6 GOSim

Με βάση την παραπάνω ανάλυση θεωρήσαμε ότι ο αλγόριθμος προσαρμογής σκορ elim διαχειρίζεται καλύτερα τις σχέσεις των GO όρων, δίνοντας μας πιο ειδικές βιολογικές λειτουργίες των γονιδίων μας. Έτσι θα χρησιμοποιήσουμε τον αλγόριθμο αυτό στην ανάλυση εμπλουτισμού της βιβλιοθήκης “GOSim” για το Σύνολο Β των γονιδίων αλλά και για τις ομάδες λειτουργικής ομοιότητας C1 και C2 που δημιουργήθηκαν με την ίδια βιβλιοθήκη στο υποκεφάλαιο 6.3.4.2.

Η ανάλυση εμπλουτισμού δεν θα πραγματοποιηθεί για την ομάδα C5 καθώς περιλαμβάνει τρία μόνο γονίδια με καμία λειτουργική ομοιότητα μεταξύ τους καθώς και για τις ομάδες C3 και C4 μιας και οι λειτουργίες τους μελετήθηκαν στο υποκεφάλαιο 6.3.4.2. Εδώ θα χρησιμοποιηθεί το τεστ Fisher.

Η λειτουργία της βιβλιοθήκης που πραγματοποιεί την ανάλυση εμπλουτισμού απαιτεί ως είσοδο :

- ✓ λίστα Entrez ταυτοτήτων γονιδίων της κάθε ομάδας
- ✓ λίστα γονιδίων βιβλιοθήκης org.Hs.egGO
- ✓ αλγόριθμος επίλυσης elim
- ✓ όριο αποκοπής p τιμής το 0,01

Με τη βοήθεια της βιβλιοθήκης "org.Hs.eg.db" πραγματοποιείται μετάφραση των γονιδίων αυτών σε Entrez ταυτότητες γονιδίων.

Το αποτέλεσμα της ανάλυσης εμπλουτισμού περιλαμβάνει τρεις λίστες. Η πρώτη λίστα περιέχει τους GO όρους (go_id) που αποτελούν μεταφράσεις των Entrez ταυτοτήτων της εισόδου, με τις αντίστοιχες περιγραφές τους (Term). Η δεύτερη λίστα περιλαμβάνει τις εμπλουτισμένες GO ταυτότητες με αντίστοιχες τιμές ***p (p.values)*** που λαμβάνει η κάθε μια και η τρίτη περιέχει τις Entrez ταυτότητες των γονιδίων (genes) που αντιστοιχίζονται σε κάθε GO ταυτότητα.

Στα παρακάτω αποτελέσματα (Πίνακας 7.22) οι τιμές **p** είναι πολύ μικρές καταδεικνύοντας πως πρόκειται για σημαντικά ευρήματα της δεδομένης λίστας γονιδίων.

Πίνακας 7.22: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο elim για τα γονίδια του Συνόλου Β

GOID	Λειτουργική κατηγορία	p-value
GO:0034375	αναδιαμόρφωση σωματιδίου υψηλής πυκνότητας λιπο-πρωτεΐνης (high-density lipoprotein particle remodeling)	$1.99 \cdot e^{-10}$
GO:0006898	ενδοκυττάρωση με μεσολάβηση υποδοχέα (receptor-mediated endocytosis)	$5.19 \cdot e^{-10}$
GO:0033344	εκροή χοληστερόλης (cholesterol efflux)	$7.09 \cdot e^{-10}$
GO:0006958	ενεργοποίηση συμπληρώματος, κλασσικό μονοπάτι (complement activation, classical pathway)	$1.19 \cdot e^{-09}$
GO:0010873	θετική ρύθμιση εστεροποίησης της χοληστερόλης (positive regulation of cholesterol esterification)	$2.13 \cdot e^{-09}$
GO:0034380	συγκρότηση σωματιδίου υψηλής πυκνότητας λιποπρωτεΐνης (high-density lipoprotein particle assembly)	$6.36 \cdot e^{-09}$
GO:0043691	αντίστροφη μεταφορά χοληστερόλης (reverse cholesterol transport)	$7.12 \cdot e^{-08}$
GO:0042632	ομοιόσταση χοληστερόλης (cholesterol homeostasis)	$6.28 \cdot e^{-07}$
GO:0010951	αρνητική ρύθμιση της δραστηριότητας της ενδοπεπτιδάσης (negative	$1.07 \cdot e^{-06}$

	regulation of endopeptidase activity)	
GO:0034384	καθάρισμα σωματιδίου υψηλής πυκνότητας λιποπρωτεΐνης (high-density lipoprotein particle clearance)	$1.13 \cdot e^{-06}$

Στους Πίνακες 7.22 και 7.23 παρουσιάζονται οι δέκα υπερ-εκφρασμένες GO ταυτότητες των βιολογικών λειτουργιών που λαμβάνουν τα μικρότερα ***p-value*** με βάση το αποτέλεσμα του εμπλουτισμού για τον αλγόριθμο elim για τις ομάδες C1 και C2.

Πίνακας 7.23: Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο elim για τα γονίδια της ομάδας C1

GOID	Λειτουργική κατηγορία	p-value
GO:0006508	πρωτεόλυση	$1.78 \cdot 10^{-08}$
GO:0070613	ρύθμιση επεξεργασίας πρωτεΐνης (regulation of protein processing)	$3.31 \cdot 10^{-07}$
GO:0030193	ρύθμιση πήξης του αίματος	$5.11 \cdot 10^{-07}$
GO:1900046	ρύθμιση της αιμόστασης	$5.11 \cdot 10^{-07}$
GO:0050818	ρύθμιση της πήξης	$6.53 \cdot 10^{-07}$
GO:0010951	αρνητική ρύθμιση της δραστηριότητας της ενδοπεπτιδάσης	$8.51 \cdot 10^{-07}$
GO:0010466	αρνητική ρύθμιση της δραστηριότητας της πεπτιδάσης	$1.01 \cdot 10^{-06}$
GO:0061041	ρύθμιση της επούλωσης πληγών	$2.05 \cdot 10^{-06}$
GO:0030212	μεταβολική διεργασία της υαλουρονάνης (hyaluronan metabolic process)	$3.21 \cdot 10^{-06}$
GO:1903510	μεταβολική διεργασία μυκοπολυσακχαριδίων (mycopolysaccharide metabolic process)	$3.26 \cdot 10^{-06}$

Πίνακας 7.24 Οι δέκα σημαντικότεροι GO όροι της ανάλυσης εμπλουτισμού για τεστ Fisher με αλγόριθμο elim για τα γονίδια της ομάδας C2

GOID	Λειτουργική κατηγορία	p-value
GO:0006897	ενδοκυττάρωση	$1.32 \cdot 10^{-09}$
GO:0006952	απόκριση άμυνας (defense response)	$2.9 \cdot 10^{-09}$
GO:0006958	ενεργοποίηση συμπληρώματος, κλασσικό μονοπάτι	$4.41 \cdot 10^{-08}$
GO:0006955	απόκριση άμυνας	$6.19 \cdot 10^{-08}$
GO:0006959	χυμική ανοσοαπόκριση	$1.23 \cdot 10^{-07}$
GO:0002455	χυμική ανοσολογική απόκριση που προκαλείται από κυκλοφορούντα μόρια ανοσοσφαιρίνης	$1.24 \cdot 10^{-07}$

GO:0006956	ενεργοποίηση συμπληρώματος	$1.40 \cdot 10^{-07}$
GO:0072376	ενεργοποίηση αλληλουχίας αντιδράσεων πρωτεΐνης	$4.10 \cdot 10^{-07}$
GO:0016192	μεταφορά με τη διαμεσολάβηση κυστιδίου (vesicle-mediated transport)	$4.75 \cdot 10^{-07}$
GO:0002250	ανοσοαπόκριση με τη διαμεσολάβηση υποδοχέα (adaptive immune response)	$7.32 \cdot 10^{-07}$

Τα αποτελέσματα εμπλουτισμού για όλες τις ομάδες περιλαμβάνουν λειτουργίες που έχουν σχέση με την τοξικότητα. Για την ομάδα C1 οι λειτουργίες αυτές είναι η πήξη του αίματος, η αιμόσταση και η επούλωση πληγών. Για την ομάδα C2 είναι η ενδοκυττάρωση, η απόκριση άμυνας, η ενεργοποίηση του συμπληρώματος και η ανοσοαπόκριση.

8 Συμπεράσματα

Με την εξέλιξη της νανοτεχνολογίας, η χρήση των νανοσωματιδίων σε εφαρμογές της ιατρικής, της βιολογίας, της φαρμακευτικής αλλά και της γεωργίας κρίνεται όλο και περισσότερο αναγκαία. Η έκθεση του ανθρώπινου οργανισμού στα νανοσωματίδια αναμένεται να αυξηθεί στο μέλλον, για αυτό το λόγο και κρίνεται αναγκαία η πιο αποτελεσματική και πιο ασφαλή χρήση τους.

Για να βελτιωθεί η ασφάλεια των νανοσωματιδίων σε μελλοντικές βιολογικές και βιοιατρικές εφαρμογές, χρειάζεται να συλλέξουμε πληροφορίες όχι μόνο για τα νανουλικά καθαυτά αλλά και για τις απρόβλεπτες βιολογικές επιδράσεις που έχουν όταν εισέρχονται στον ανθρώπινο οργανισμό και έρχονται σε επαφή με βιολογικά υγρά και κύτταρα.

Ερευνητικές ομάδες έχουν αποδείξει ότι οι ιδιότητες των επιφανειών των νανοσωματιδίων σχετίζονται με κυτταρικές αποκρίσεις, όπως η τοξικότητα. Στην παρούσα διπλωματική εργασία μέσω της πρωτεϊνικής κορώνας που σχηματίζεται γύρω από το νανοσωματίδιο, η οποία και αποτελεί τη βιολογική του ταυτότητα, εντοπίσαμε τις βιολογικές λειτουργίες που ενεργοποιούνται και δρουν κατά την επαφή του με ορό αίματος. Αν αυτές οι λειτουργίες σχετίζονται με την απόκριση του κυττάρου σε τοξικό σώμα, τότε θα επιβεβαιωθεί η τοξικότητα των νανοσωματιδίων χρυσού που χρησιμοποιήθηκαν.

Στην ερευνητική εργασία του Walkey και των συνεργατών του [8], οι πρωτεΐνες του πρωτεϊνικού στέμματος που λαμβάνονται βρέθηκε ότι εμπλέκονται σε πέντε βασικές βιολογικές διεργασίες :

- ✓ Πήξη
- ✓ Ενεργοποίηση του συμπληρώματος
- ✓ Μεταφορά λιπιδίων
- ✓ Φλεγμονή
- ✓ Κυτταρική συσχέτιση

Στον Πίνακα 8.1 παρουσιάζουμε τους σημαντικότερους GO όρους που σχετίζονται με αυτές τις βιολογικές διεργασίες με βάση το αποτέλεσμα της ερευνητικής εργασίας του Walkey.

Πίνακας 8.1: Σημαντικότερες βιολογικές διεργασίες των πρωτεϊνών της εργασίας του Walkey και των συνεργατών του **[8]**

Βιολογική διεργασία (BP)	GOID	Λειτουργική κατηγορία
πήξη	GO:0050817	πήξη
	GO:0042060	επούλωση πληγών
ενεργοποίηση συμπληρώματος	GO:0030449	ρύθμιση ενεργοποίησης συμπληρώματος
	GO:0006956	ενεργοποίηση συμπληρώματος
μεταφορά λιπιδίων	GO:0006629	μεταβολική διεργασία λιπιδίων
	GO:0008610	βιοσυνθετική διεργασία λιπιδίων
	GO:0006869	μεταφορά λιπιδίων
	GO:0019915	αποθήκευση λιπιδίων
	GO:0016042	καταβολική διεργασία λιπιδίων
	GO:0097006	ρύθμιση του επιπέδου των λιποπρωτεϊνών πλάσματος (regulation of plasma lipoprotein particle levels)
	GO:0071827	οργάνωση των σωματιδίων λιποπρωτεΐνης πλάσματος
	GO:0034368	αναδιαμόρφωση συμπλόκου πρωτεΐνης λιπιδίου
	GO:0034358	σωματίδιο λιποπρωτεΐνης πλάσματος
φλεγμονή	GO:0006954	φλεγμονώδης απόκριση
	GO:0006955	ανοσοαπόκριση
	GO:0006953	απόκριση οξείας φάσης
κυτταρική συσχέτιση	GO:0006897	ενδοκυττάρωση
	GO:0007166	μονοπάτι σηματοδότησης υποδοχέα κυτταρικής επιφάνειας (cell surface receptor signaling pathway)
	GO:0005102	Δέσμευση υποδοχέα (receptor binding)

Οι παραπάνω βιολογικές διαδικασίες επιβεβαιώνεται από τη βιβλιογραφία ότι σχετίζονται με την τοξικότητα του νανοσωματιδίου. Με βάση την έρευνα των Tokuyuki Yoshida και των συνεργατών του η διαδικασία της πήξης και της φλεγμονώδους απόκρισης ενεργοποιούνται με την είσοδο νανοσωματιδίων διοξειδίου του πυριτίου στον ανθρώπινο οργανισμό ως αποτέλεσμα της τοξικότητας τους [67]. Η δραστηριοποίηση των λιπιδίων της μεμβράνης και των λιποπρωτεϊνών του πλάσματος είναι άρρηκτα συνδεδεμένες με την είσοδο τοξικών ουσιών στον ανθρώπινο οργανισμό [58]. Ο Yeon Kyung Lee και οι συνεργάτες του μελέτησαν την

επίδραση της πρωτεϊνικής κορώνας στα νανοσωματίδια για την διαμόρφωση της κυτταροτοξικότητας τους και παρατηρούν ότι η ενεργοποίηση του συμπληρώματος πραγματοποιείται κατά την είσοδο των νανοσωματιδίων στον οργανισμό μιας και θεωρείται ξένο σώμα [63]. Η κυτταρική συσχέτιση επιλέχθηκε ως μοντέλο βιολογικής αλληλεπίδρασης, από τον Walkey και τους συνεργάτες του, εξαιτίας της συνάφειας της με φλεγμονώδεις αποκρίσεις όπως η τοξικότητα. [8]

Στην παρούσα εργασία μελετήσαμε εκτεταμένα τα λειτουργικά προφίλ και τα GO γραφήματα των δεδομένων της εργασίας του Walkey [8]. Συγκεκριμένα στο Κεφάλαιο 6 μελετήσαμε ξεχωριστά τα δύο σύνολα δεδομένων πρωτεϊνών (Σύνολο Α) και γονιδίων (Σύνολο Β). Αρχικά, και για τα δυο σύνολα, ταξινομήθηκαν τα γονίδια στις κατηγορίες βιολογικής λειτουργίας, βιοχημικής δραστηριότητας και θέσης δραστηριοποίησης τους και έπειτα απεικονίστηκαν οι σχέσεις τους σε γραφήματα. Παρόλο που μελετήθηκαν και μικρές ομάδες αυτών, η πυκνότητα της βιολογικής πληροφορίας δεν μας επέτρεψε να βγάλουμε σαφή συμπεράσματα για την δραστηριοποίηση τους. Στην συνέχεια με χρήση εργαλείων μέτρησης της λειτουργικής ομοιότητας τους τα χωρίσαμε σε ομάδες και μελετήσαμε την λειτουργία των γονιδίων των ομάδων που εμφάνιζαν την μεγαλύτερη ομοιότητα. Εδώ εντοπίσαμε δυο ομάδες γονιδίων που κατά κύριο λόγο μετέχουν σε λειτουργίες που σχετίζονται με την τοξικότητα. Τέλος και προκειμένου να εμπλουτίσουμε τις μελετώμενες λίστες γονιδίων εφαρμόσαμε μεθοδολογίες ανάλυσης εμπλουτισμού. Τα αποτελέσματα αυτής της ανάλυσης παρουσιάστηκαν στο Κεφάλαιο 7 με χρήση διαφορετικών βιβλιοθηκών. Είχαν στόχο να εντοπίσουν είτε τα γονίδια είτε τα μεταβολικά μονοπάτια τα οποία υπερ-εκφράζονται στις συνθήκες εισόδου του νανοσωματιδίου στον ανθρώπινο οργανισμό, και τα οποία περιγράφουν καλύτερα την βιολογική ταυτότητα του νανοσωματιδίου. Τα αποτελέσματά μας προήλθαν είτε από ανάλυση υπερ-εκπροσώπησης γονιδίων είτε από ανάλυση αθροιστικού σκορ. Η ανάλυση υπερ-εκπροσώπησης των γονιδίων πραγματοποιήθηκε από τις βιβλιοθήκες clusterProfiler, KEGG, Reactome, topGO και GOSim, ενώ η ανάλυση αθροιστικού σκορ, λαμβάνοντας ως σκορ τις τιμές του $\boldsymbol{Q^2}$ και του ***VIP***, πραγματοποιήθηκε μόνο από την βιβλιοθήκη topGO.

Κατά την ανάλυση εμπλουτισμού της βιβλιοθήκης clusterProfiler οι εμπλουτισμένοι GO όροι που προκύπτουν για την οντολογία BP αφορούν βιολογικές διεργασίες όπως, η ενεργοποίηση του συμπληρώματος, η ενεργοποίηση αλληλουχίας

αντιδράσεων πρωτεϊνών, η ενδοκυττάρωση, η συγκρότηση πλάσματος από σωματίδια λιποπρωτεΐνης, η συγκρότηση συμπλέγματος πρωτεΐνης-λιπιδίου, καθώς και η ενεργοποίηση του ανοσοποιητικού μηχανισμού που περιλαμβάνει τη δραστηριότητα Β-λεμφοκυττάρων και της ανοσοσφαιρίνης.

Αντίστοιχα με χρήση της βιβλιοθήκης KEGG τα μονοπάτια τα οποία υπερεκφράζονται είναι το μονοπάτι του συστήματος του συμπληρώματος, το μονοπάτι αλληλουχίας αντιδράσεων πήξης και το μονοπάτι της φαγοκυττάρωσης. Τα υπόλοιπα μονοπάτια που προκύπτουν από την ανάλυση αυτή αφορούν σε ασθένειες από βακτηριακές μολύνσεις του οργανισμού όπως το Staphylococcus aureus, το Bordetella pertussis, είτε σε αυτοάνοσα νοσήματα κ.α.

Τα μονοπάτια που χαρακτηρίζονται από υψηλής τάξης εμπλουτισμό κατά την ανάλυση με χρήση της βιβλιοθήκης Reactome αφορούν διεργασίες οι οποίες ενεργοποιούνται σε περιπτώσεις αγγειακής βλάβης όπως ο σχηματισμός ινώδους θρόμβου ή η αιμόσταση, και σε διεργασίες όπως η ενεργοποίηση του συμπληρώματος, η δράση των αιμοπεταλίων και ο μεταβολισμός των πρωτεϊνών.

Κατά την ανάλυση εμπλουτισμού με χρήση της βιβλιοθήκης topGO πραγματοποιήθηκαν διαφορετικά στατιστικά τεστ και διαφορετικοί αλγόριθμοι προσαρμογής σκορ. Εδώ θεωρούμε ότι το αποτέλεσμα με χρήση του στατιστικού τεστ Kolmogorov-Smirnov με αλγόριθμο elim ως το πιο αξιόπιστο. Το K-S στατιστικό τεστ χρησιμοποιεί για την ανάλυση εμπλουτισμού τα σκορ των γονιδίων $\boldsymbol{Q^2}$ και VIP περιλαμβάνοντας έτσι την χρήσιμη πληροφορία της σημαντικότητας του κάθε γονιδίου σε σχέση με την κυτταρική συσχέτιση. Αντίστοιχα ο αλγόριθμος elim δεν εξετάζει την κάθε GO ταυτότητα χωριστά, αλλά λαμβάνοντας υπόψη την σημαντικότητα των απογόνων του, που περιέχουν σημαντικότερη πληροφορία απ' ότι οι πρόγονοι, τους οποίους αποκλείει από την ανάλυση. Τα αποτελέσματα λαμβάνοντας ως σκορ το $\boldsymbol{Q^2}$ αφορούν εμπλουτισμένες διεργασίες όπως η ρύθμιση της βιολογικής ποιότητας, η ανάπτυξη του ανοσοποιητικού συστήματος, η αιμοποίηση, η επούλωση πληγών, η πήξη ρευστών του οργανισμού και του αίματος, η καρβοξυλίωση των πρωτεϊνών και η ενδοκυττάρωση. Αντίστοιχα όταν το σκορ είναι το VIP στατιστικό οι εμπλουτισμένοι GO όροι αφορούν διεργασίες όπως η ενεργοποίηση του συμπληρώματος και η ρύθμιση αυτής, η ρύθμιση της ανοσίας του οργανισμού, η φαγοκυττάρωση καθώς και η απόκριση του ανοσοποιητικού

συστήματος μέσω της δραστηριοποίησης των Β-λεμφοκυττάρων και των λευκοκυττάρων.

Τέλος το αποτέλεσμα της ανάλυσης εμπλουτισμού της βιβλιοθήκης GOSim με χρήση των αλγορίθμων classic και elim περιλαμβάνει διεργασίες όπως η αλληλουχία αντιδράσεων ενεργοποίησης πρωτεΐνης, η ενδοκυττάρωση, η ενεργοποίηση του συμπληρώματος, η συγκρότηση πλάσματος με σωματίδια λιποπρωτεΐνης, η απόκριση της άμυνας του οργανισμού, η δραστηριοποίηση των Β-λεμφοκυττάρων, η αναδιαμόρφωση σωματιδίου λιποπρωτεΐνης και η δραστηριοποίηση των ενδοπεπτιδάσων.

Παρακάτω παρουσιάζεται ο συγκεντρωτικός πίνακας των αποτελεσμάτων της ανάλυσης εμπλουτισμού με χρήση όλων των παραπάνω βιβλιοθηκών του Bioconductor (Πίνακας 8.2). Οι πρωτεΐνες ή τα γονίδια που υπερεκφράζονται συμμετέχουν στις λειτουργίες των ευρύτερων κατηγοριών της πήξης, της ενεργοποίησης του συμπληρώματος, της μεταφοράς των λιπιδίων, της απόκρισης του ανοσοποιητικού συστήματος και της μεταβολικής δραστηριότητας των πρωτεϊνών.

Πίνακας 8.2: Σημαντικότερες βιολογικές διεργασίες των πρωτεϊνών που προκύπτουν από την ανάλυση εμπλουτισμού της κάθε βιβλιοθήκης

Βιολογική διεργασία (BP)	clusterProfiler	KeGG	Reactome	topGO	GOSim	toxicity
πήξη	GO:1903034;GO:0030193;GO:0042060;GO:0007596	hsa04610	140877;109582;140837;140875;76002	GO:0042060;GO:0030193;GO:0050817;GO:0061041;GO:1900046	GO:0030195;GO:0010543	√
ενεργοποίηση συμπληρώματος	GO:0072376;GO:0006958;GO:0006956	hsa04610	166658;977606;166663	GO:0030449;GO:0006956;GO:0006958	GO:0072376;GO:0006958;GO:0006956	√
μεταφορά λιπιδίων	GO:0034377;GO:0065005;GO:0034368;GO:0003469;GO:0071827;GO:0071825;GO:0033344;GO:0097006;GO:0042157;GO:0030301;GO:0010876	hsa04145		GO:0006629;GO:0008610;GO:0006869;GO:0019915;GO:0016042;GO:0097006	GO:0034377;GO:0065005;GO:0034368;GO:0071827;GO:0033344;GO:0010873;GO:0043691;GO:0042632;GO:0033700;GO:0001523	√
κυτταρική συσχέτιση	GO:0006897;GO:0006898		114608;76005;2173782	GO:0006898;GO:0018200;GO:0018214;GO:0006909	GO:0006897;GO:0006898	√
ανοσοποιητικό σύστημα/ανοσοαπόκριση (φλεγμονή)	GO:0002455;GO:0006959;GO:0016064;GO:0019724;GO:0006952;GO:0019724;GO:0002526	hsa05322	166786	GO:0002520;GO:0030097;GO:0048534;GO:0045087;GO:0002697;GO:0002703;GO:0002706;GO:0002712;GO:0002819;GO:0002920	GO:0002455;GO:0006952;GO:0016064;GO:0006959;GO:0019724;GO:0006953;GO:1903027;GO:0004507	√
μεταβολική διεργασία πρωτεϊνών	GO:0051248;GO:0070613;GO:0010955		159763;159740;159782;159854;163841	GO:0017187;GO:0018200;GO:0018214	GO:0010951	
Λοιμώδη νοσήματα		hsa05133;hsa05143;hsa05144				

Με βάση τον παραπάνω πίνακα οι GO όροι που εμφανίζονται με υψηλής τάξης εμπλουτισμό στην δική μας ανάλυση με χρήση των διαφορετικών βιβλιοθηκών, συμμετέχουν και στις πέντε βασικές βιολογικές διεργασίες όπως αυτές παρουσιάζονται στην ερευνητική εργασία του Walkey και συνεργατών του. [8]

Συμπεραίνεται ότι κατά κύριο λόγο στα αποτελέσματα των αναλύσεων εμπλουτισμού της κάθε βιβλιοθήκης περιλαμβάνονται οι ίδιες βιολογικές διεργασίες ή αλλιώς οι ίδιοι υπερεκφρασμένοι GO όροι. Έτσι επαληθεύεται η εγκυρότητα του αποτελέσματος, παρότι χρησιμοποιήθηκαν διαφορετικές βιβλιοθήκες και διαφορετικά στατιστικά τεστ. Επιπλέον η ανάλυση που παρουσιάζουμε εδώ κατέληξε σε πρόσθετους GO όρους που φαίνεται βιβλιογραφικά να επιβεβαιώνουν την συμμετοχή τους σε λειτουργίες τοξικότητας, για παράδειγμα η ενεργοποίηση του ανοσοποιητικού συστήματος και συγκεκριμένα των Β-λεμφοκυττάρων και των λευκοκυττάρων [62], η απόκριση της έμφυτης ανοσίας [68], η οποία περιλαμβάνει την παραγωγή αντισωμάτων, η πήξη του αίματος στις περιπτώσεις σχηματισμού ινώδους θρόμβου ή επούλωσης πληγών (αιμόσταση) [62], η κυτταρική διαφοροποίηση ως αποτέλεσμα απόκρισης σε ερεθίσματα, η φαγοκυττάρωση και η χημειοταξία [62], η αποκοκκίωση των αιμοπεταλίων, η δραστηριότητα των Scaveneger υποδοχέων [64], η αποπτωτική διεργασία και ο θάνατος των νευρώνων, η μετανάστευση των ενδοθυλιακών κυττάρων των αιμοφόρων αγγείων, η ενεργοποίηση των μακροφάγων ως απόκριση του ανοσοποιητικού συστήματος, καθώς και η μεταφορά λιπιδίων μέσω αιματοεγκεφαλικού φραγμού.

Βιβλιογραφία

[1] P. Foroozandeh and A. A. Aziz, "Merging Words of Nanomaterials and Biological Enviroment :Factors Governing Protein Corona Formation on Nanoparticles and Its Biological Consequences," *Nanoscale Research Letters,* vol. 10:221, 2015.

[2] C. Buzea, I. I. Blandino and K. Robbie, "Nanomaterials and nanoparticles: Sources and toxicity," *Biointephases,* vol. 2, no. 4, 2007.

[3] M. Lundqvist, J. Stigler, T. Cedervall και T. Berggard, «The evolution of the Protein Corona around Nanoparticles: A Test Study,» *ACSNANO,* τόμ. 5, αρ. 9, p. 7503–7509, 2011.

[4] G. Yu, «Reactome Pathway Analysis,» 2014.

[5] Β. Αλεπόρου-Μαρίνου, Α. Αργυροκαστρίτης, Α. Κομητοπούλου και Π. Παλόγλου, Βιολογία, Αθήνα: ΟΡΓΑΝΙΣΜΟΣ ΕΚΔΟΣΕΩΝ ΔΙΔΑΚΤΙΚΩΝ ΒΙΒΛΙΩΝ, 2011.

[6] V. Trevino, F. Falciani και H. A. Barrera-Saldaña, «DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research,» *Molecular Medicine,* τόμ. 13, αρ. 9-10, pp. 527-541, 2007.

[7] Thermo Fisher, «Overview of Mass Spectrometry,» Thermo Fisher SCIENTIFIC, [Ηλεκτρονικό]. Available: http://www.thermofisher.com. [Πρόσβαση 2015].

[8] C. D.Walkey, J. B.Olsen, F. Song, R. Li και H. Guo, «Protein Corona Fingerprinting Predicts the Cell Association of Gold Nanoparticles,» τόμ. 8, αρ. 3, pp. 2439-2455, 2014.

[9] The Gene Ontology Consortium, «Gene Ontology Consortium: going forward,» *Nucleic Acids Research,* τόμ. 43, pp. 1049-1056, 2014.

[10] M. Ashburner, C. Ball, J. Blake και D. Botstein, «Gene ontology: Tool for the Unification of Biology,» *Nature Genetics,* τόμ. 25, αρ. 2, pp. 25-29, 2000.

[11] A. Sánchez, M. Salicrú και J. Ocaña, «Statistical methods for the analysis of high-throughput data based on functional profiles derived from Gene Ontology,» *Elsevier,* τόμ. 137, αρ. 12, p. 3975–3989, 2007.

[12] A. Sánchez, J. Ocaña και M. Salicrú, «goProfiles: an R package for the

Statistical Analysis of Functional profiles,» 2008.

[13] Wikipedia contributors, «Directed acyclic graph,» Wikipedia, The Free Encyclopedia, 2015. [Ηλεκτρονικό]. Available: https://en.wikipedia.org/w/index.php?title=Directed_acyclic_graph&oldid=675747945. [Πρόσβαση 2015].

[14] S. Falcon και R. Gentleman, «Using GOstats to test gene lists for GO term association,» *Bioinformatics,* τόμ. 23, αρ. 2, pp. 257-258, 2007.

[15] H. Fröhlich, N. Speer, A. Poustka και T. Beißbarth, «GOSim - An R-Package for Computation of Information Theoretic GO Similarities Between Terms and Gene Products,» *BMC Bioinformatics,* τόμ. 166, αρ. 8, 2007.

[16] R Core Team, «R: A Language and Environment for Statistical Computing,» 2014. [Ηλεκτρονικό].

[17] H. Pages, M. Carlson, S. Falcon και N. Li, «AnnotationDbi: Annotation Database Interface,» 2015.

[18] M. Carlson, «org.Hs.eg.db: Genome wide annotation for Human,» 2015.

[19] L. Geer, A. Marchler-Bauer, R. Geer και L. Han, «The NCBI BioSystems database,» *Nucleic Acids Research,* αρ. 38, 2010.

[20] Magrane, Michele; UniProt Consortium , «UniProt Knowledgebase: a hub of integrated protein data,» *Database,* 2011.

[21] M. Carlson, «hgu95av2.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95av2),» 2015.

[22] The Gene Ontology Consortium, «Gene Ontology Consortium: going forward,» *Nucleic Acids Res,* τόμ. 43, αρ. D1049-D1056, 2015.

[23] Χ. Α. Χαραλαμπίδης, «Στοιχεία Πιθανοτήτων και Στατιστικής,» Αθήνα, 2004.

[24] Wikipedia contributors, «Hypergeometric distribution,» Wikipedia, The Free Encyclopedia, 2015. [Ηλεκτρονικό]. Available: https://en.wikipedia.org/w/index.php?title=Hypergeometric_distribution&oldid=678768781. [Πρόσβαση 2015].

[25] Κ. Φωκιανός και Χ. Χαραλάμπους, Εισαγωγή στην R: Πρόχειρες Σημειώσεις, 2η επιμ., Αθήνα: Πανεπιστήμιο Κύπρου, 2010.

[26] Δ. Αντζουλάκος, «Ανάλυση Δεδομένων με τη Χρήση Στατιστικών Πακέτων,»

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά, Αθήνα, 2013.

[27] Δ. Ελευθερία, «Στατιστική Ανάλυση δεδομένων ιστικών μικροσυστοιχιών,» Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών: Μαθηματικά των Υπολογιστών και των Αποφάσεων, Πανεπιστήμιο Πατρών, Πάτρα, 2014.

[28] Σ.-Κ. Σταυρινίδης, «Παλινδρόμηση Μερικών Ελαχίστων Τετραγώνων,» Αθήνα, 2011.

[29] K. H. L. S. S. Tahir Mehmood, «A review of variable selection methods in Partial Least Square Regression,» *Biostatistics,* pp. 62-69, 2012.

[30] T. Chai και R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE): Arguments against RMSE in the literature,» *Geoscientific Model Development,* pp. 1247-1250, 2014.

[31] Wikipedia contributors, «Cross-validation (statistics),» 2015. [Ηλεκτρονικό]. Available: https://en.wikipedia.org/w/index.php?title=Cross-validation_(statistics)&oldid=678327727. [Πρόσβαση 1 9 2015].

[32] C. M. Barbu, «zoom: A spatial data visualization tool,» 2013.

[33] P. Resnik, «Using Information Content to Evaluate Semantic in a Taxonomy,» *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* pp. 448-453, 1995.

[34] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu και S. Wang, «GOSemSim: an R package for measuring semantic similarity among GO terms and gene products,» *Bioinformatics,* τόμ. 26, αρ. 7, pp. 976-978, 2010.

[35] A. Shlicker, F. S. Domingues, J. Rahnenführer και T. Lengauer, «A new measure for functional similarity of gene products based on Gene Ontolgy,» *BMC Bioinformatic,* τόμ. 302, αρ. 7, 2006.

[36] C. Fraley και A. Raftery, «Model-based Clustering, Discriminant Analysis and Density,» *Journal of the American Statistical Association,* τόμ. 97, pp. 611-631, 2002.

[37] M. Maechler, P. Rousseeuw, A. Struyf και M. Hubert, «cluster: Cluster Analysis Basics and Extensions,» 2015.

[38] K. Hansen, J. Gentry, L. Long, R. Gentleman και S. Falcon, «Rgraphviz:

Provides plotting capabilities for R graph objects,» 2015.

[39] M. Carlson, «GO.db: A set of annotation maps describing the entire Gene Ontology,» 2015.

[40] A. Mark, R. Thompson και W. Chunlei, «mygene: Access MyGene.Info_ services,» 2014.

[41] P. Khanna, C. Ong, B. H. Bay και G. H. Baeg, «Nanotoxicity: An Interplay of Oxidative Stress, Inflammation and Cell Death,» *Nanomaterilas,* τόμ. 5, αρ. 3, pp. 1163-1180, 2015.

[42] K. Higashisaka, Y. Yoshioka, K. Yamashite και Y. Morishita, «Acute phase proteins as biomarkers for predicting the exposure and toxicity of nanomaterials,» *Biomaterials,* τόμ. 32, αρ. 1, 2011.

[43] D. Laskin, «Macrophages and inflamatory mediators in chemical toxicity: a battle of forces,» *Cherm Res Toxicol,* τόμ. 22, αρ. 8, pp. 1376-1385, 2009.

[44] M. Baggiolini και l. Clark-Lewis, «Interleukin-8, a chemotactic and inflammatory cytokine,» τόμ. 307, αρ. 1, pp. 97-101, 1992.

[45] A. N. Ilinskaya και M. A. Dobrovolskaia, «Nanoparticles and the blood coagulation system. Part II: safety concerns,» *Nanomedicine,* τόμ. 8, αρ. 6, pp. 969-981, 2014.

[46] K. Zhang και R. J. Kaufman, «Unfolding the toxicity of cholesterol,» *Nature Cell Biology,* τόμ. 5, pp. 769-770, 2003.

[47] W. M. Pardridge, Introduction to the Blood-Brain Barrier, Cambridge: Cambridge Univercity Press, 2006.

[48] C. Buzea, I. I. P. Blandino και K. Robbie, «Nanomaterials and nanoparticles: Sources and toxicity,» *Biointerphases,* τόμ. 2, αρ. 4, pp. MR17-MR172, 2007.

[49] Wikipedia contributors, «Neurotoxicity,» Wikipedia, The Free Encyclopedia, 2015. [Ηλεκτρονικό]. Available: https://en.wikipedia.org/w/index.php?title=Neurotoxicity&oldid=679319678. [Πρόσβαση 2015].

[50] D. Steinritz, A. Schmidt, F. Balszuweit, H. Thiermann και M. Ibrahim, «Assessment of Endothelial Cell Migration After Exposure to Toxic Chemicals,» *J Vis Exp,* τόμ. 10, αρ. 101, 2015.

[51] R. A. Irizarry, C. Wang, Y. Zhou και T. P.Speed, «Gene Set Enrichment Analysis Made Simple,» *Stat Methods Med Res,* τόμ. 18, αρ. 6, pp. 565-575, 2011.

[52] A. Subramanian, P. Tamayo, V. K.Mootha και S. Mukjerjee, «Gene set enrichment analysis:A knowledge-based approach for interpreting genome-wide expression profiles,» *PNAs,* τόμ. 102, αρ. 43, pp. 15545-15550, 2005.

[53] G. Yu, L.-G. Wang, Y. Han και Q.-Y. He, «clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters,» *OMICS,* τόμ. 16, αρ. 5, pp. 284-287, 2012.

[54] M. Carlson, «KEGG.db: A set of annotation maps for KEGG,» 2015.

[55] S. Carbon, A. Ireland, C. Mungall, S. Shu, AmiGO και P. W. W. Group, «AmiGO: online access to ontology and annotation data,» *Bioinformatics,* τόμ. 25, αρ. 2, pp. 288-289, 2009.

[56] Y. Benjamini και Y. Hochberg, «Controlling the false discovery rate: a practical and powerful approach to multiple testing,» *Journal of the Royal Statistical Society,* τόμ. 57, αρ. 1, pp. 289-300, 1995.

[57] E. Hodgson, Dictionary of Toxicology, 3rd Edition επιμ., Academic Press, 2015.

[58] E. K. Silbergeld, «Toxicology,» Encyclopedia of Occupational Health and Safety, 2015. [Ηλεκτρονικό]. Available: http://www.ilocis.org/documents/chpt33e.htm. [Πρόσβαση 2015].

[59] G. Yu, L. Wang, Y. Han και Q. He, «clusterProfiler: an R package for comparing biological themes among gene clusters.,» *OMICS: A Journal of Integrative Biology,* τόμ. 16, αρ. 5, pp. 284-287, 2012.

[60] L. Weijun, Brouwer και Cory, «Pathview:pathway based data intergration and visualization,» *Bioinformatics,* τόμ. 29, αρ. 14, pp. 1830-1831, 2013.

[61] D. Croft, A. F. Mundo, R. Haw και M. Milacic, «Reactome pathway knowledgebase,» *Nucleic Acids Res,* τόμ. 42, αρ. D, pp. 472-427, 2014.

[62] R. Gupta, Biomarkers in Toxicity, Academic Press, 2014.

[63] Y. K. Lee, E.-J. Choi και T. J. Webster, «Effect of the protein corona on nanoparticles for modulating cytotoxicity and immnotoxicity,» *International*

Journal of Nanomedicine, τόμ. 2015, αρ. 10, pp. 97-113, 2014.

[64] J. Shannahan, R. Podila, A. Aldossari και H. Emerson, «Formation of a protein corona on silver nanoparticles mediates cellular toxicity via scavenger receptors,» *Toxicological Sciences,* τόμ. 143, αρ. 1, pp. 136-146, 2015.

[65] A. Alexa και J. Rahnenführer, «Gene set enrichment analysis with topGO,» 2014.

[66] A. Alexa, J. Rahnenführer και T. Lengauer, «Improved scoring of functional groups from gene expression data by decorrelating GO graph structure,» *Bioinformatics,* τόμ. 22, αρ. 13, pp. 1600-1607, 2006.

[67] T. Yoshida, Y. Yoshioka και Y. Morishita, «Protein corona changes mediated by surface modification of amorphous silica nanoparticles suppress acute toxicity and activation of intrinsic coagulation cascade in mice,» *Nanotechnology,* τόμ. 26, αρ. 24, 2015.

[68] E. Hodgson και M. Roe, Dictionary of Toxicology, 3rd Edition επιμ., 2015.

[69] C. Kanz, P. Aldebert, N. Althorpe, W. Baker και A. Baldwin, «The EMBL Nucleotide Sequence Database,» *Nucleic Acids Res,* pp. D29-D33, 2005.

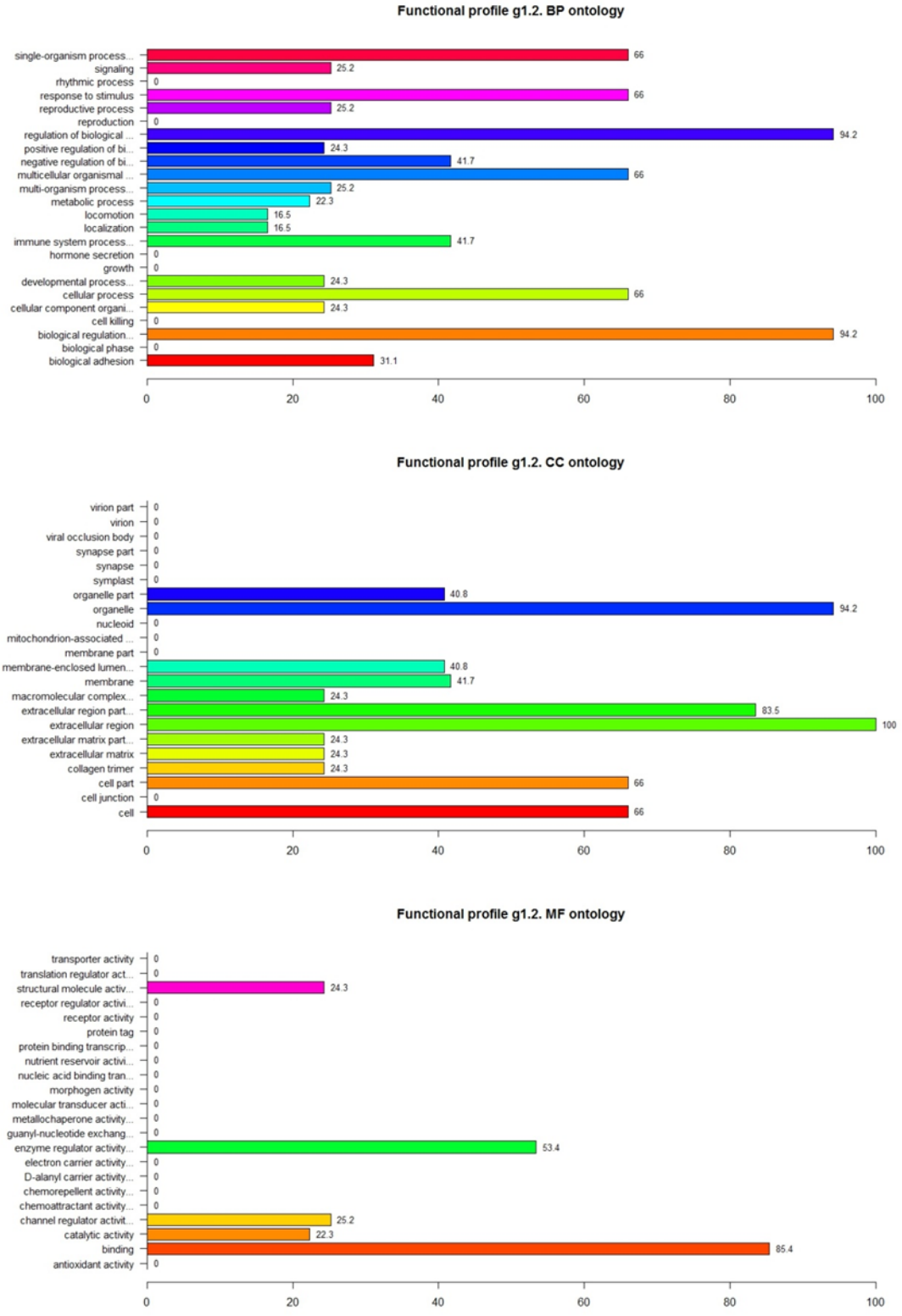
Παράρτημα Α

Πίνακας ΠΑ1: Τμήμα του πίνακα μετάφρασης που προκύπτει από τις μεταφράσεις των 129 Uniprot ταυτοτήτων των πρωτεϊνών του Συνόλου Α με χρήση της βιβλιοθήκης org.Hs.eg.db

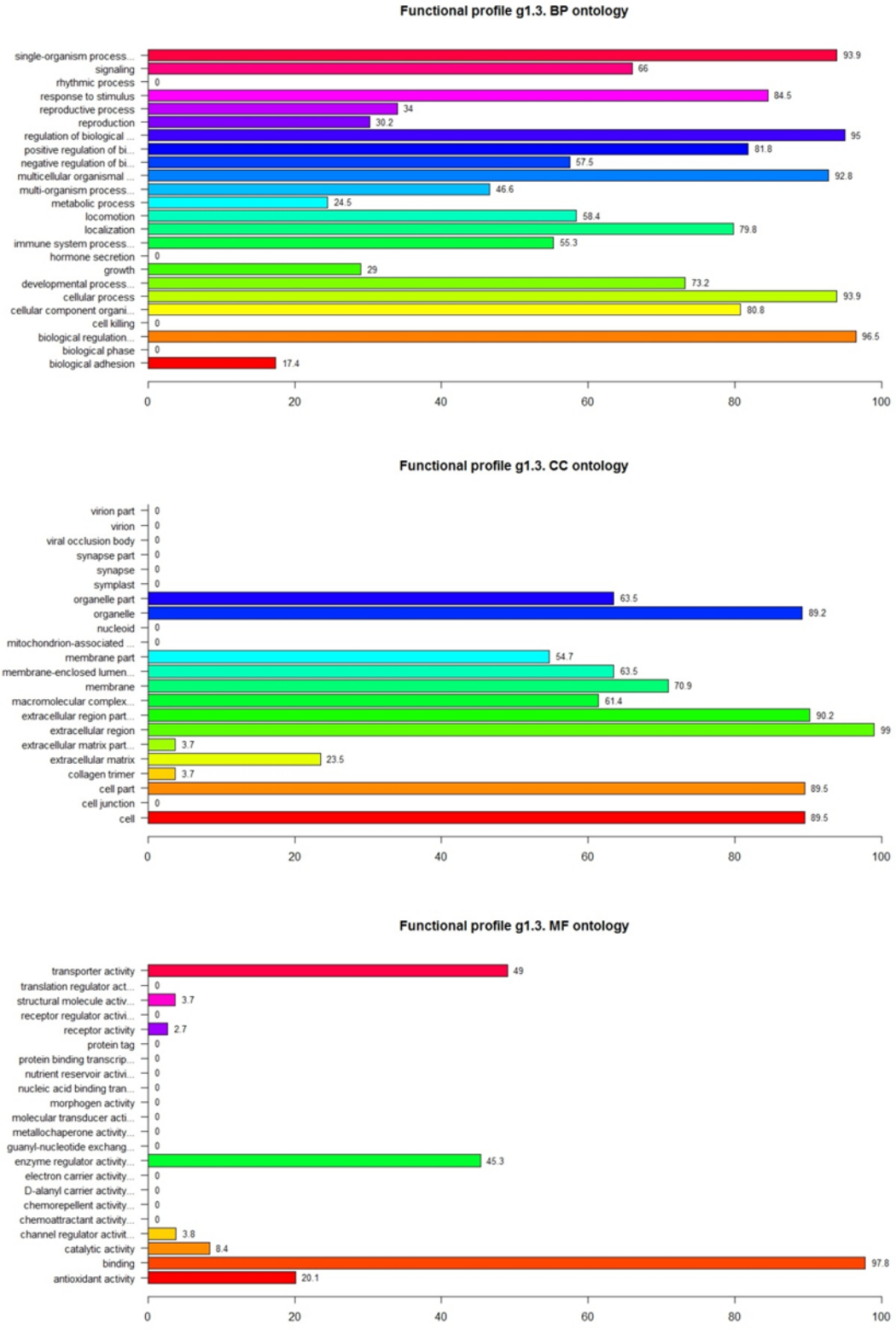
UNIPROT	GO	EVIDENCE	ONTOLOGY	ENTREZID
P01024	GO:0001798	IEA	BP	718
P01024	GO:0001934	IDA	BP	718
P01024	GO:0001970	IEA	BP	718
P01024	GO:0004866	IEA	MF	718
P01024	GO:0005102	TAS	MF	718
P01024	GO:0005515	IPI	MF	718
P01024	GO:0005576	TAS	CC	718
P01024	GO:0005615	IDA	CC	718
P01024	GO:0005886	TAS	CC	718
P01024	GO:0006631	IEA	BP	718
P01024	GO:0006954	IEA	BP	718
P01024	GO:0006955	TAS	BP	718
P01024	GO:0006956	IMP	BP	718
P01024	GO:0006956	TAS	BP	718
P01024	GO:0006957	TAS	BP	718
P01024	GO:0006958	IEA	BP	718
P01024	GO:0007165	TAS	BP	718
P01024	GO:0007186	TAS	BP	718
P01024	GO:0010575	IDA	BP	718
P01024	GO:0010828	IDA	BP	718
P01024	GO:0010866	IDA	BP	718
P01024	GO:0010884	IDA	BP	718
P01024	GO:0010951	IEA	BP	718
P01024	GO:0030449	TAS	BP	718
P01024	GO:0031715	IDA	MF	718
P01024	GO:0045087	TAS	BP	718
P01024	GO:0045745	IDA	BP	718
P01024	GO:0045766	IEA	BP	718
P01024	GO:0050776	TAS	BP	718
P01024	GO:0070062	IDA	CC	718
P01024	GO:0072562	IDA	CC	718
P01024	GO:2000427	IMP	BP	718
P01834	NA	NA	NA	NA
P0C0L4	GO:0001849	IDA	MF	720
P0C0L4	GO:0004866	IEA	MF	720
P0C0L4	GO:0005576	TAS	CC	720
P0C0L4	GO:0005886	TAS	CC	720
P0C0L4	GO:0006954	IEA	BP	720
P0C0L4	GO:0006956	IGI	BP	720
P0C0L4	GO:0006956	TAS	BP	720
P0C0L4	GO:0006958	IEA	BP	720

Πίνακας ΠΑ2: Τμήμα του πίνακα μετάφρασης που προκύπτει από τις μεταφράσεις των 76 συμβολικών ονομασιών των γονιδίων του Συνόλου Β με χρήση της βιβλιοθήκης org.Hs.eg.db

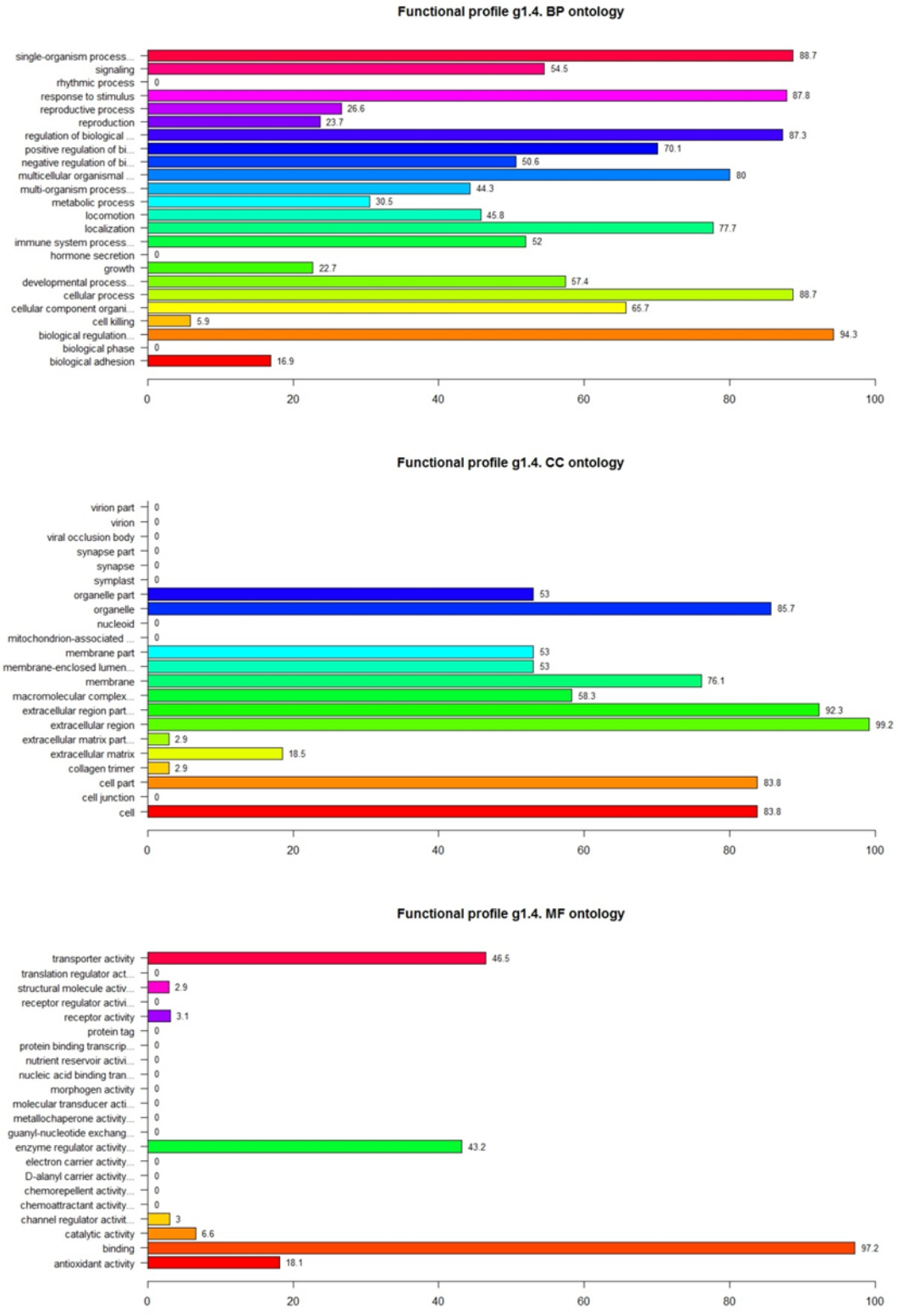
SYMBOL	UNIPROT	GO	EVIDENCE	ONTOLOGY	ENTREZID
AMBP	P02760	GO:0004867	TAS	MF	259
AMBP	P02760	GO:0005515	IPI	MF	259
AMBP	P02760	GO:0005576	NAS	CC	259
AMBP	P02760	GO:0005576	TAS	CC	259
AMBP	P02760	GO:0005615	IDA	CC	259
AMBP	P02760	GO:0005886	IDA	CC	259
AMBP	P02760	GO:0007155	NAS	BP	259
AMBP	P02760	GO:0007565	NAS	BP	259
AMBP	P02760	GO:0009986	IEA	CC	259
AMBP	P02760	GO:0010951	IEA	BP	259
AMBP	P02760	GO:0010951	TAS	BP	259
AMBP	P02760	GO:0016032	IEA	BP	259
AMBP	P02760	GO:0018298	IEA	BP	259
AMBP	P02760	GO:0019855	NAS	MF	259
AMBP	P02760	GO:0019862	IDA	MF	259
AMBP	P02760	GO:0020037	IDA	MF	259
AMBP	P02760	GO:0030163	IEA	BP	259
AMBP	P02760	GO:0036094	IEA	MF	259
AMBP	P02760	GO:0042167	NAS	BP	259
AMBP	P02760	GO:0042803	IPI	MF	259
AMBP	P02760	GO:0043231	IEA	CC	259
AMBP	P02760	GO:0046329	TAS	BP	259
AMBP	P02760	GO:0046904	NAS	MF	259
AMBP	P02760	GO:0050777	NAS	BP	259
AMBP	P02760	GO:0070062	IDA	CC	259
AMBP	P02760	GO:0072562	IDA	CC	259
HABP2	Q14520	GO:0004252	IEA	MF	3026
HABP2	Q14520	GO:0005539	TAS	MF	3026
HABP2	Q14520	GO:0005576	NAS	CC	3026
HABP2	Q14520	GO:0005615	TAS	CC	3026
HABP2	Q14520	GO:0006508	IEA	BP	3026
HABP2	Q14520	GO:0007155	TAS	BP	3026
ITIH2	P19823	GO:0004866	TAS	MF	3698
ITIH2	P19823	GO:0004867	IEA	MF	3698
ITIH2	P19823	GO:0005576	NAS	CC	3698
ITIH2	P19823	GO:0010951	IEA	BP	3698
ITIH2	P19823	GO:0010951	TAS	BP	3698
ITIH2	P19823	GO:0030212	IEA	BP	3698
ITIH2	P19823	GO:0070062	IDA	CC	3698
ITIH2	P19823	GO:0072562	IDA	CC	3698
TTHY	NA	NA	NA	NA	NA



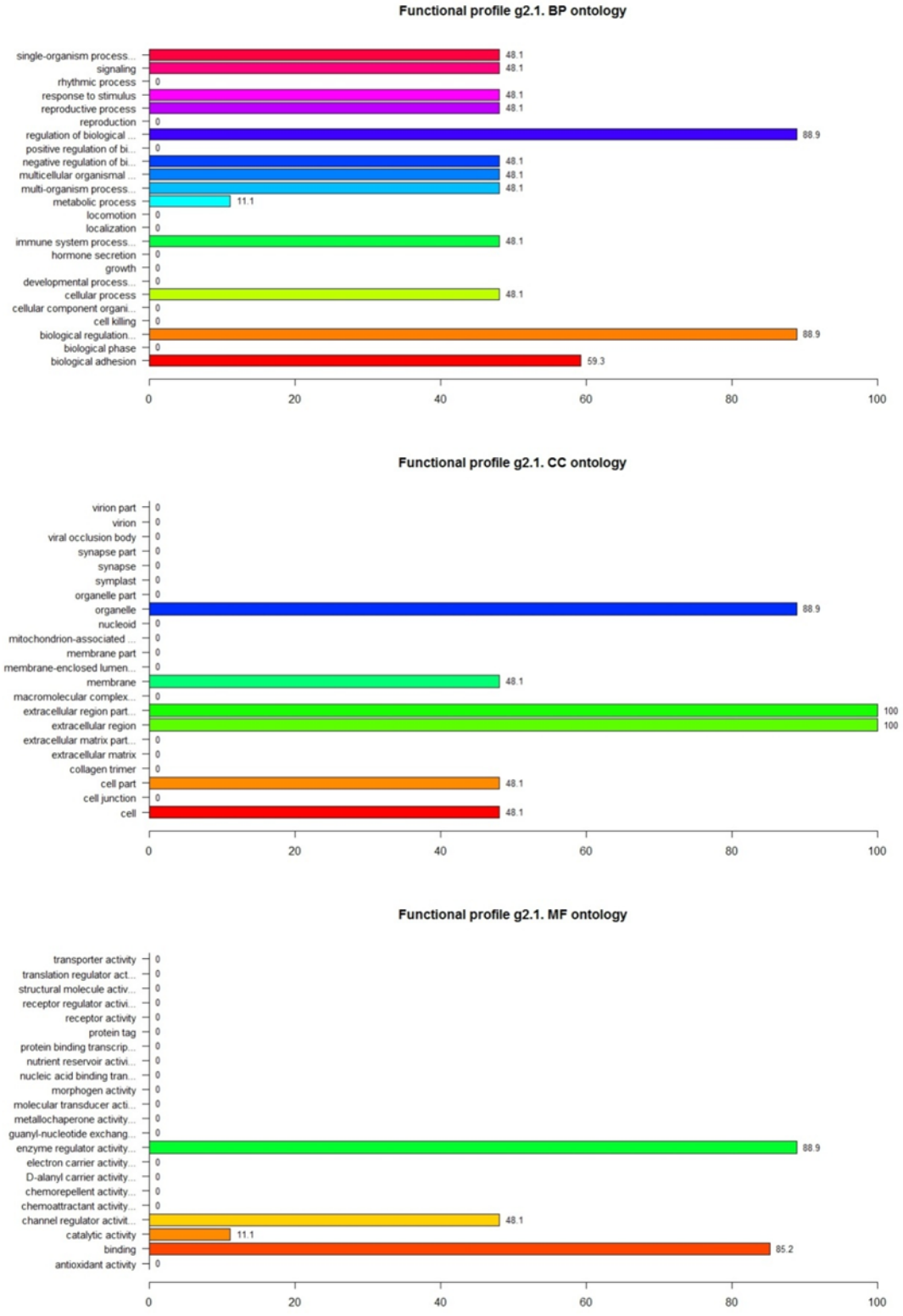
Σχήμα ΠΑ1: Λειτουργικά προφίλ γονιδίων 2^{ης} υποομάδας 1^{ης} ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (GOprofile)



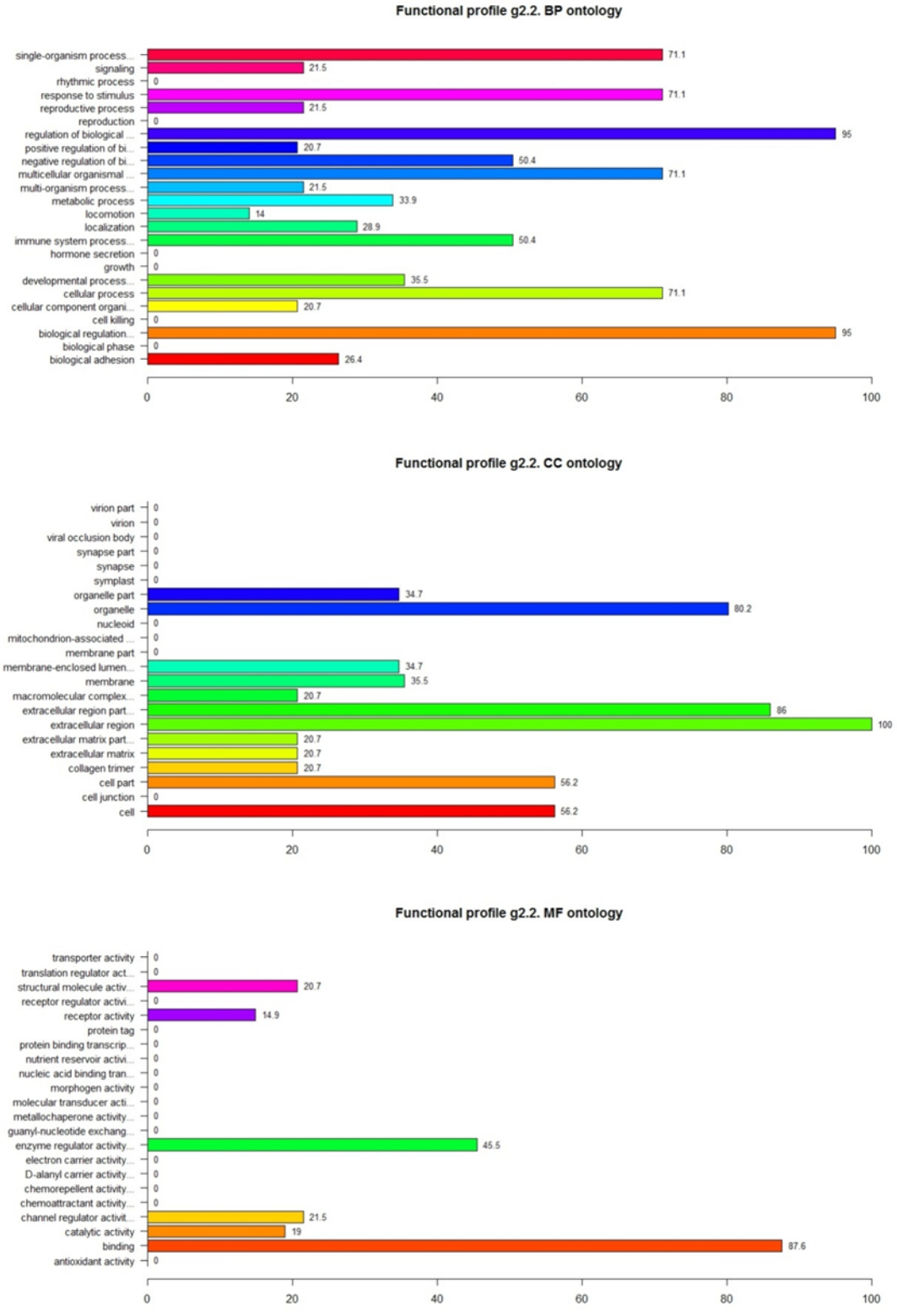
Σχήμα ΠΑ2: Λειτουργικά προφίλ γονιδίων 3ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



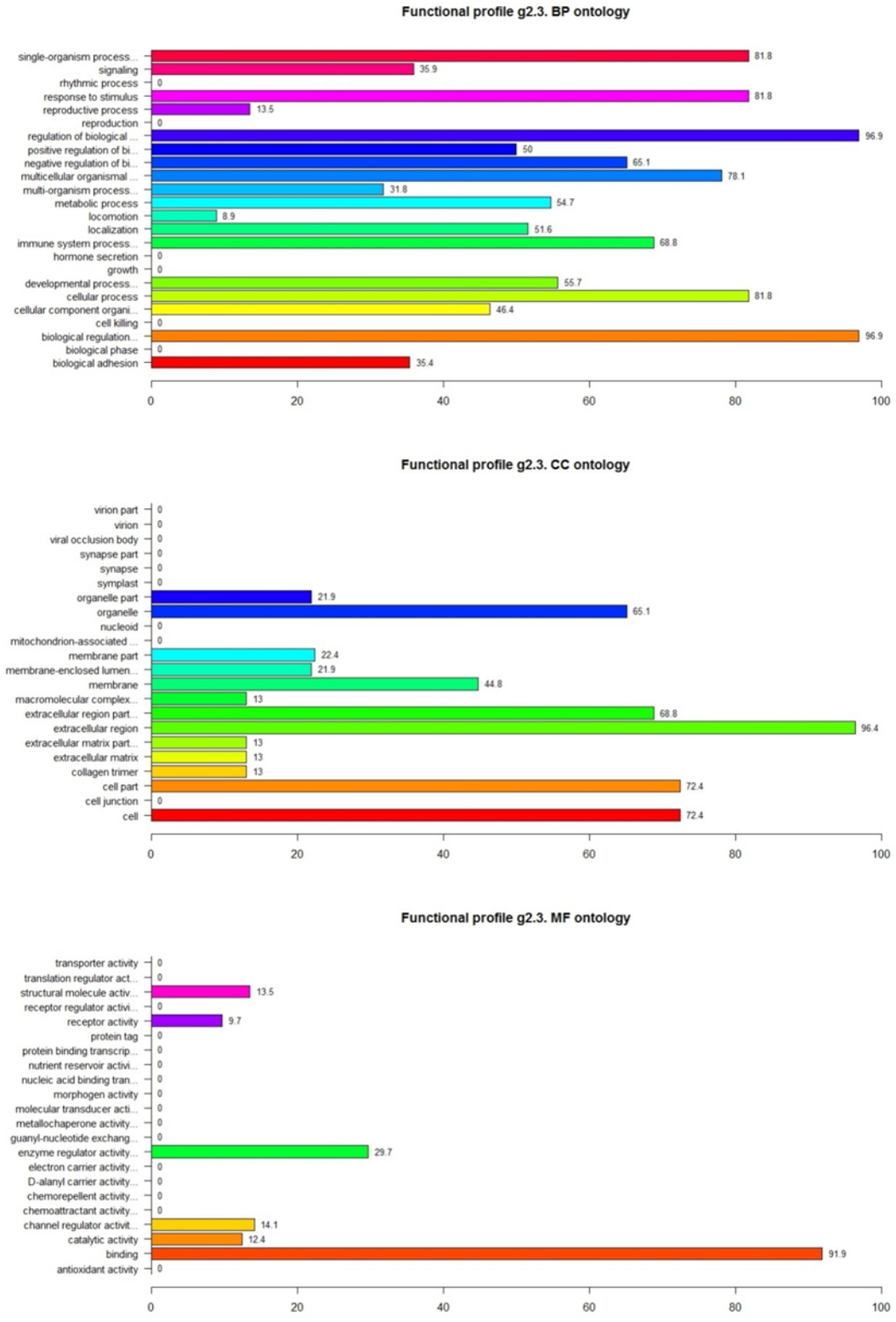
Σχήμα ΠΑ3: Λειτουργικά προφίλ γονιδίων 4ης υποομάδας 1ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



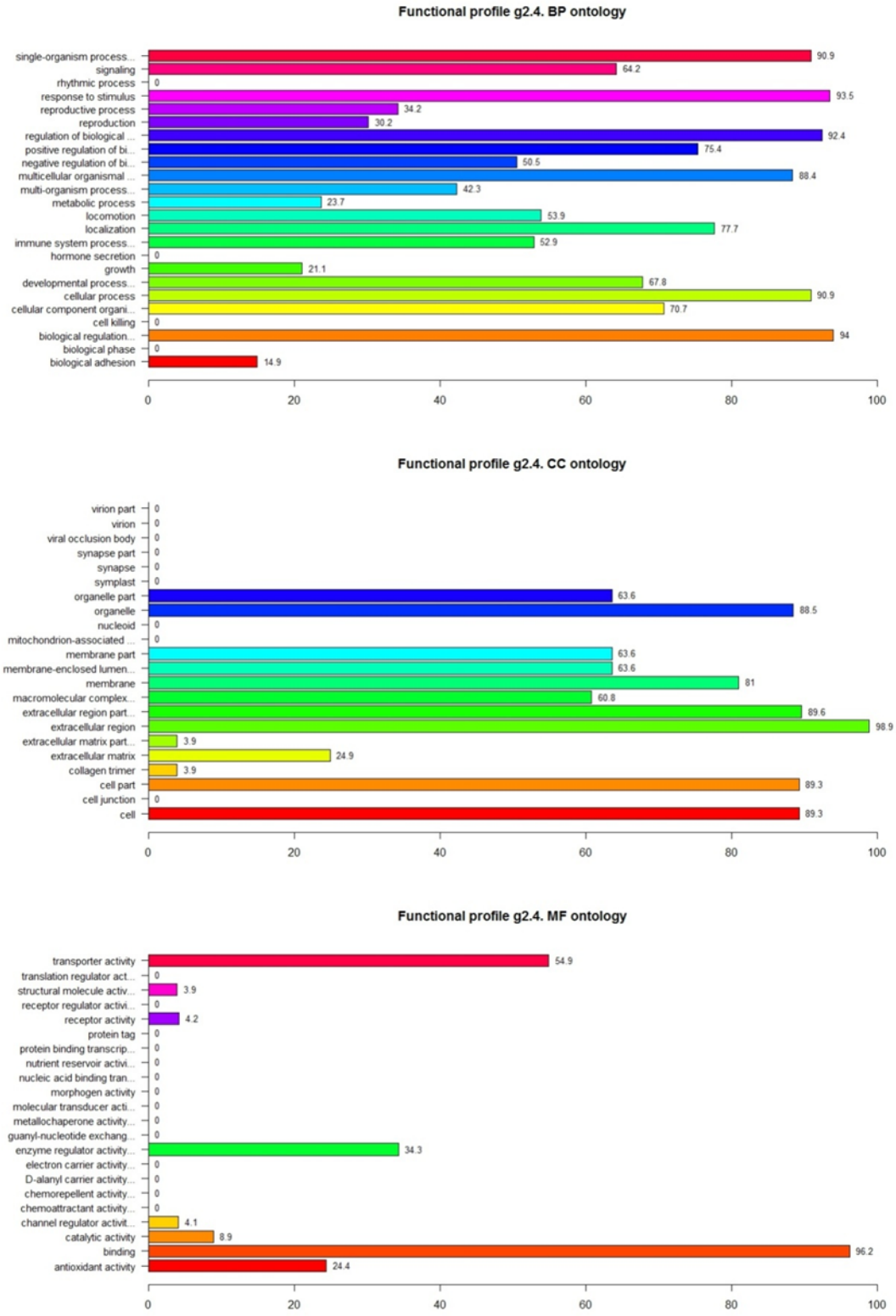
Σχήμα ΠΑ4: Λειτουργικά προφίλ γονιδίων 1ης υποομάδας 2ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



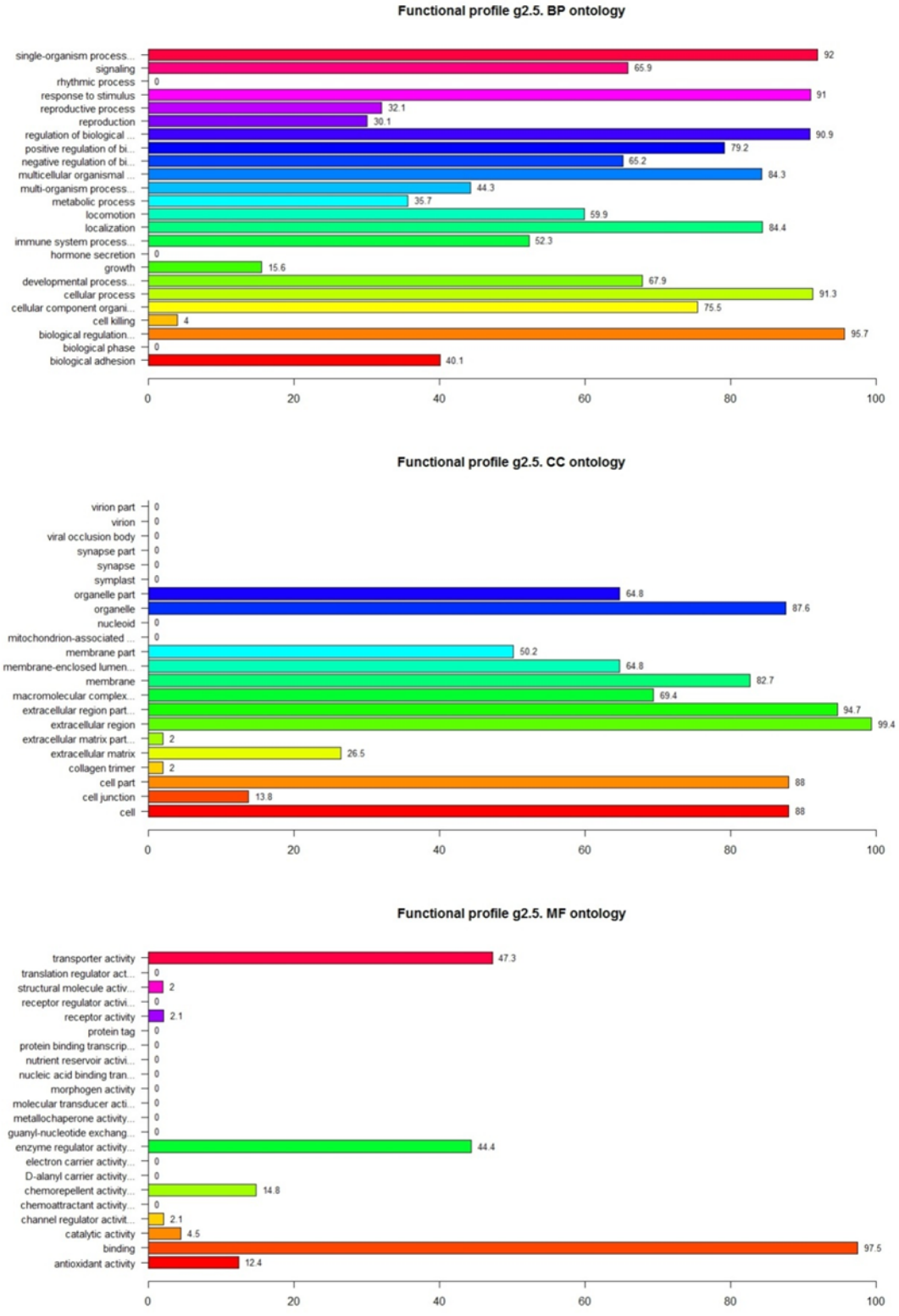
Σχήμα ΠΑ5: Λειτουργικά προφίλ γονιδίων 2^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



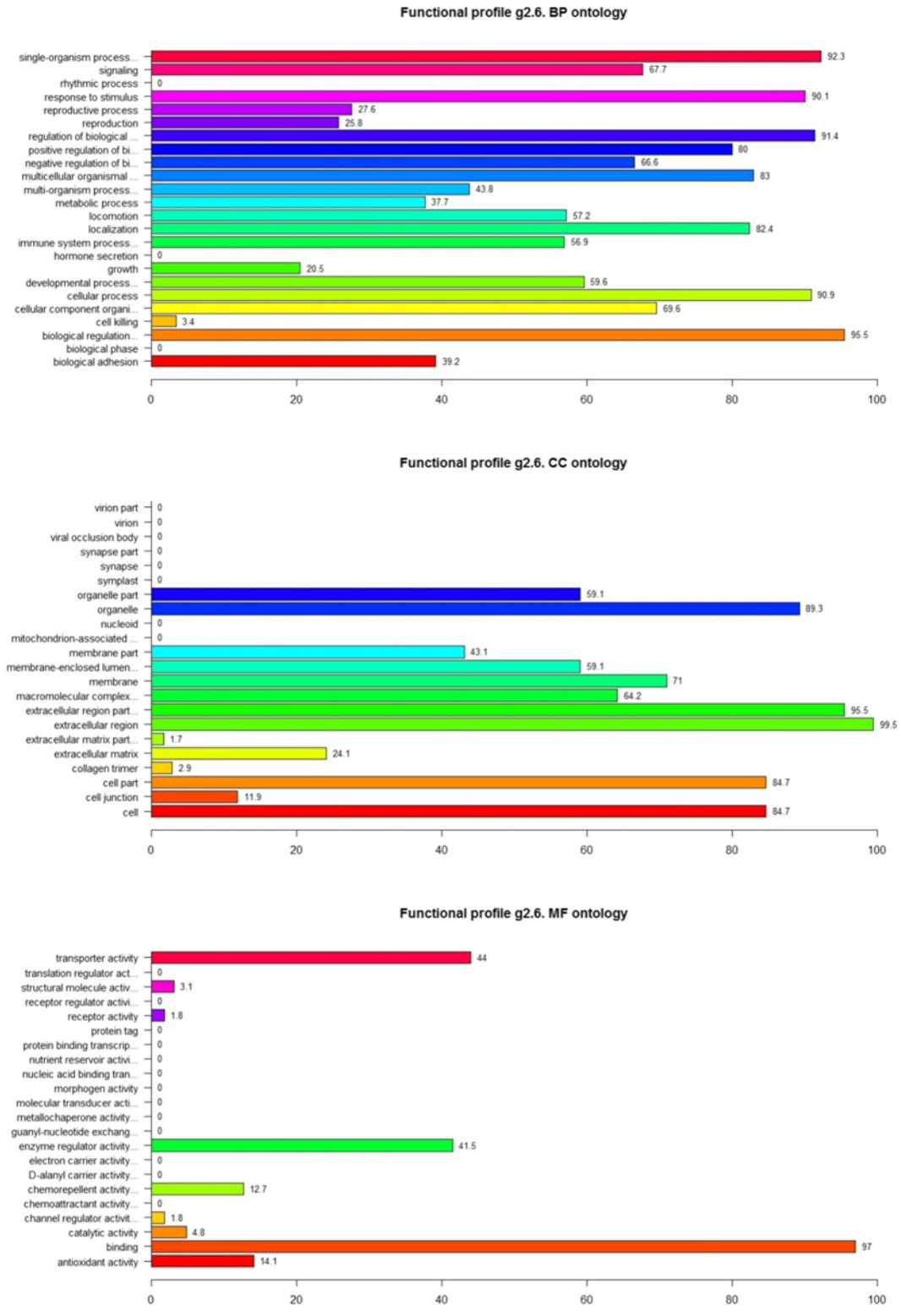
Σχήμα ΠΑ6: Λειτουργικά προφίλ γονιδίων 3^{ης} υποομάδας 2^{ης} ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



Σχήμα ΠΑ7: Λειτουργικά προφίλ γονιδίων 4ης υποομάδας 2ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



Σχήμα ΠΑ8: Λειτουργικά προφίλ γονιδίων 5ης υποομάδας 2ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)



Σχήμα ΠΑ9: Λειτουργικά προφίλ γονιδίων 6ης υποομάδας 2ης ομαδοποίησης Συνόλου Β σε οντολογίες BP, CC και MF (goProfiles)

Πίνακας ΠΑ3: Μέρος πίνακα ομοιότητας για τα γονίδια του Συνόλου Β (GOSim)

259	3026	3698	3697	3699	4060	5624	10216	84735	8858	4064	1401	5341	338	3818	341	3273	348	3514	3503	8542	3827
1.00	0.92	0.80	0.80	0.80	0.45	0.62	0.47	0.85	0.70	0.54	0.56	0.51	0.55	0.56	0.49	0.71	0.62	0.61	0.65	0.55	0.56
0.92	1.00	0.72	0.72	0.72	0.46	0.56	0.29	1.00	0.68	0.00	0.46	0.70	0.36	0.50	0.49	0.93	0.55	0.37	0.47	0.37	0.78
0.80	0.72	1.00	1.00	1.00	0.83	0.76	0.54	0.80	0.75	0.05	0.69	0.58	0.57	0.77	0.73	0.86	0.79	0.58	0.58	0.58	0.82
0.80	0.72	1.00	1.00	1.00	0.83	0.76	0.54	0.80	0.75	0.05	0.69	0.58	0.57	0.77	0.73	0.86	0.79	0.58	0.58	0.58	0.82
0.80	0.72	1.00	1.00	1.00	0.83	0.76	0.54	0.80	0.75	0.05	0.69	0.58	0.57	0.77	0.73	0.86	0.79	0.58	0.58	0.58	0.82
0.45	0.46	0.83	0.83	0.83	1.00	0.50	0.28	0.92	0.72	0.33	0.53	0.42	0.52	0.59	0.44	0.60	0.66	0.32	0.55	0.35	0.37
0.62	0.56	0.76	0.76	0.76	0.50	1.00	0.49	1.00	1.00	0.54	0.74	0.66	0.68	0.72	0.64	0.94	0.90	0.56	0.74	0.55	0.71
0.47	0.29	0.54	0.54	0.54	0.28	0.49	1.00	0.57	0.47	0.38	0.40	0.39	0.43	0.37	0.38	0.68	0.60	0.72	0.75	0.49	0.34
0.85	1.00	0.80	0.80	0.80	0.92	1.00	0.57	1.00	1.00	0.00	0.93	0.73	0.72	1.00	0.99	1.00	1.00	0.74	0.73	0.74	0.80
0.70	0.68	0.75	0.75	0.75	0.72	1.00	0.47	1.00	1.00	0.34	0.87	0.73	0.72	0.87	0.76	1.00	0.96	0.65	0.73	0.65	0.71
0.54	0.00	0.05	0.05	0.05	0.33	0.54	0.38	0.00	0.34	1.00	1.00	0.50	1.00	0.57	0.07	0.94	0.81	0.59	0.75	0.68	0.71
0.56	0.46	0.69	0.69	0.69	0.53	0.74	0.40	0.93	0.87	1.00	1.00	0.49	0.63	0.67	0.59	0.60	0.69	0.68	0.90	0.57	0.63
0.51	0.70	0.58	0.58	0.58	0.42	0.66	0.39	0.73	0.73	0.50	0.49	1.00	0.44	0.65	0.51	0.65	0.64	0.51	0.58	0.47	0.64
0.55	0.36	0.57	0.57	0.57	0.52	0.68	0.43	0.72	0.72	1.00	0.63	0.44	1.00	0.58	0.78	0.60	0.74	0.58	0.57	0.67	0.56
0.56	0.50	0.77	0.77	0.77	0.59	0.72	0.37	1.00	0.87	0.57	0.67	0.65	0.58	1.00	0.55	0.82	0.76	0.48	0.73	0.42	0.73
0.49	0.49	0.73	0.73	0.73	0.44	0.64	0.38	0.99	0.76	0.07	0.59	0.51	0.78	0.55	1.00	0.62	0.93	0.39	0.50	0.57	0.46
0.25	0.05	0.09	0.09	0.09	0.22	0.16	0.13	0.09	0.18	0.54	0.27	0.17	0.46	0.21	0.09	0.46	0.40	0.20	0.28	0.17	0.21
0.71	0.93	0.86	0.86	0.86	0.60	0.94	0.68	1.00	1.00	0.94	0.60	0.65	0.60	0.82	0.62	1.00	0.74	0.60	0.73	0.54	0.78
0.62	0.55	0.79	0.79	0.79	0.66	0.90	0.60	1.00	0.96	0.81	0.69	0.64	0.74	0.76	0.93	0.74	1.00	0.61	0.75	0.72	0.74
0.61	0.37	0.58	0.58	0.58	0.32	0.56	0.72	0.74	0.65	0.59	0.68	0.51	0.58	0.48	0.39	0.60	0.61	1.00	1.00	0.72	0.57
0.65	0.47	0.58	0.58	0.58	0.55	0.74	0.75	0.73	0.73	0.75	0.90	0.58	0.57	0.73	0.50	0.73	0.75	1.00	1.00	0.75	0.80
0.55	0.37	0.58	0.58	0.58	0.35	0.55	0.49	0.74	0.65	0.68	0.57	0.47	0.67	0.42	0.57	0.54	0.72	0.72	0.75	1.00	0.47

Παράρτημα Β

Πίνακας ΠΒ1: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία BP (clusterProfiler)

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0072376	protein activation cascade	9/40	67/11978	7.32E-13	1.77E-10	7.55E-11	CRP/IGHG1/IGKC/KLKB1/KNG1/C1QA/C1R /C4BPA/C4BPB	9
GO:0006958	complement activation, classical pathway	7/40	31/11978	6.60E-12	7.99E-10	3.41E-10	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0006897	endocytosis	15/40	484/11978	1.60E-11	1.02E-09	4.36E-10	PRG4/CRP/AMBP/APOA1/APOB/APOC1/ APOE/IGHG1/IGKC/LBP/C4BPA/C4BPB/CALR /APOL1/CD5L	15
GO:0034377	plasma lipoprotein particle assembly	6/40	18/11978	1.69E-11	1.02E-09	4.36E-10	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0065005	protein-lipid complex assembly	6/40	19/11978	2.46E-11	1.19E-09	5.08E-10	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0002455	humoral immune response mediated by circulating immunoglobulin	7/40	42/11978	6.60E-11	1.81E-09	7.71E-10	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0034367	macromolecular complex remodeling	6/40	22/11978	6.72E-11	1.81E-09	7.71E-10	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0034368	protein-lipid complex remodeling	6/40	22/11978	6.72E-11	1.81E-09	7.71E-10	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0034369	plasma lipoprotein particle remodeling	6/40	22/11978	6.72E-11	1.81E-09	7.71E-10	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0006956	complement activation	7/40	47/11978	1.52E-10	3.68E-09	1.57E-09	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0016192	vesicle-mediated transport	18/40	996/11978	6.05E-10	1.31E-08	5.58E-09	PRG4/CRP/AMBP/HRG/APOA1/APOB/APOC1 /APOE/IGHG1/IGKC/KNG1/LBP/PLEK/C4BPA /C4BPB/CALR/APOL1/CD5L	18
GO:0071827	plasma lipoprotein particle organization	6/40	31/11978	6.49E-10	1.31E-08	5.58E-09	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0071825	protein-lipid complex subunit organization	6/40	32/11978	7.97E-10	1.47E-08	6.28E-09	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0006898	receptor-mediated endocytosis	10/40	203/11978	8.53E-10	1.47E-08	6.28E-09	PRG4/AMBP/APOA1/APOB/APOC1/APOE /IGKC/CALR/APOL1/CD5L	10
GO:0033344	cholesterol efflux	6/40	36/11978	1.70E-09	2.74E-08	1.17E-08	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0006952	defense response	19/40	1253/11978	3.15E-09	4.76E-08	2.03E-08	CRP/HRG/APOA1/APOA4/APOE/IGHG1/IGHM /IGKC/ITIH4/KLKB1/KNG1/LBP/CD180/C1QA /C1R/C4BPA/C4BPB/APOL1/CD5L	19
GO:0097006	regulation of plasma lipoprotein particle levels	6/40	48/11978	1.04E-08	1.48E-07	6.30E-08	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0006959	humoral immune response	8/40	135/11978	1.20E-08	1.61E-07	6.88E-08	CRP/IGHG1/IGHM/IGKC/C1QA/C1R/C4BPA/ C4BPB	8
GO:0016064	immunoglobulin mediated immune response	7/40	89/11978	1.50E-08	1.92E-07	8.16E-08	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0019724	B cell mediated immunity	7/40	91/11978	1.76E-08	2.13E-07	9.07E-08	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0051248	negative regulation of protein metabolic process	14/40	691/11978	2.27E-08	2.55E-07	1.09E-07	PRG4/AMBP/HRG/APOA4/APOE/ITIH1/ITIH2 /ITIH3/ITIH4/KNG1/APOM/C4BPA/C4BPB/CALR	14
GO:0070613	regulation of protein processing	10/40	286/11978	2.32E-08	2.55E-07	1.09E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1 /KNG1/C4BPA/C4BPB	10
GO:0042157	lipoprotein metabolic process	7/40	97/11978	2.75E-08	2.89E-07	1.23E-07	APOA1/APOA4/APOB/APOC1/APOE/APOM	7

							/APOL1	
GO:0015918	sterol transport	6/40	60/11978	4.11E-08	3.98E-07	1.70E-07	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0030301	cholesterol transport	6/40	60/11978	4.11E-08	3.98E-07	1.70E-07	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:1903034	regulation of response to wounding	10/40	309/11978	4.82E-08	4.39E-07	1.87E-07	HRG/APOA1/APOE/KLKB1/KNG1/LBP/PLEK /PROC/C4BPA/C4BPB	10
GO:0006950	response to stress	26/40	2892/11978	4.89E-08	4.39E-07	1.87E-07	CRP/AMBP/HRG/APOA1/APOA4/APOB/APOE /IGHG1/IGHM/IGKC/ITIH4/KLKB1/KNG1/LBP /CD180/PLEK/PROC/SEPP1/C1QA/C1R/C4BPA /C4BPB/CALR/APOL1/PROZ/CD5L	26
GO:0010955	negative regulation of protein processing	9/40	256/11978	1.24E-07	9.73E-07	4.15E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1/ C4BPA/C4BPB	9
GO:0030193	regulation of blood coagulation	6/40	72/11978	1.25E-07	9.73E-07	4.15E-07	HRG/APOE/KLKB1/KNG1/PLEK/PROC	6
GO:1900046	regulation of hemostasis	6/40	72/11978	1.25E-07	9.73E-07	4.15E-07	HRG/APOE/KLKB1/KNG1/PLEK/PROC	6
GO:0010876	lipid localization	9/40	257/11978	1.28E-07	9.73E-07	4.15E-07	CRP/APOA1/APOA4/APOB/APOC1/APOE/LBP /APOM/APOL1	9
GO:0032101	regulation of response to external stimulus	12/40	548/11978	1.29E-07	9.73E-07	4.15E-07	HRG/APOA1/APOE/KLKB1/KNG1/LBP/CD180 /PLEK/PROC/C4BPA/C4BPB/CALR	12
GO:0050818	regulation of coagulation	6/40	76/11978	1.73E-07	1.27E-06	5.40E-07	HRG/APOE/KLKB1/KNG1/PLEK/PROC	6
GO:0051346	negative regulation of hydrolase activity	9/40	268/11978	1.83E-07	1.30E-06	5.56E-07	AMBP/HRG/APOA1/APOC1/ITIH1/ITIH2/ITIH3 /ITIH4/KNG1	9
GO:0002250	adaptive immune response	8/40	215/11978	4.44E-07	3.07E-06	1.31E-06	CRP/IGHG1/IGHM/IGKC/C1QA/C1R/C4BPA/C4BPB	8
GO:0019538	protein metabolic process	28/40	3726/11978	4.71E-07	3.16E-06	1.35E-06	PRG4/CRP/AMBP/HABP2/HRG/APOA1/APOA4 /APOB/APOC1/APOE/IGHG1/IGKC/ITIH1/ITIH2/ ITIH3/ITIH4/KLKB1/KNG1/LBP/APOM/PROC/C1QA /C1R/C4BPA/C4BPB/CALR/APOL1/PROZ	28
GO:0006955	immune response	16/40	1194/11978	5.33E-07	3.49E-06	1.49E-06	PRG4/CRP/AMBP/HRG/APOA1/APOA4/IGHG1/ IGHM/IGKC/LBP/CD180/C1QA/C1R/C4BPA/C4BPB /APOL1	16
GO:0002376	immune system process	20/40	1941/11978	7.24E-07	4.61E-06	1.97E-06	PRG4/CRP/AMBP/HRG/APOA1/APOA4/APOB/ IGHG1/IGHM/IGKC/LBP/CD180/PLEK/PROC/ C1QA/C1R/C4BPA/C4BPB/CALR/APOL1	20
GO:0006869	lipid transport	8/40	232/11978	7.91E-07	4.91E-06	2.09E-06	APOA1/APOA4/APOB/APOC1/APOE/LBP/APOM/ APOL1	8
GO:0061041	regulation of wound healing	6/40	102/11978	9.98E-07	6.04E-06	2.57E-06	HRG/APOE/KLKB1/KNG1/PLEK/PROC	6
GO:0006909	phagocytosis	7/40	170/11978	1.29E-06	7.42E-06	3.16E-06	CRP/IGHG1/IGKC/LBP/C4BPA/C4BPB/CALR	7
GO:0010951	negative regulation of endopeptidase activity	7/40	170/11978	1.29E-06	7.42E-06	3.16E-06	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0002449	lymphocyte mediated immunity	7/40	173/11978	1.45E-06	7.89E-06	3.36E-06	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0010466	negative regulation of peptidase activity	7/40	173/11978	1.45E-06	7.89E-06	3.36E-06	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0016485	protein processing	10/40	447/11978	1.47E-06	7.89E-06	3.36E-06	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/ KNG1/C4BPA/C4BPB	10

GO:0002526	acute inflammatory response	6/40	111/11978	1.64E-06	8.64E-06	3.68E-06	CRP/ITIH4/KLKB1/LBP/C4BPA/C4BPB	6
GO:0008203	cholesterol metabolic process	6/40	112/11978	1.73E-06	8.84E-06	3.77E-06	APOA1/APOA4/APOB/APOC1/APOE/APOL1	6
GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	7/40	178/11978	1.75E-06	8.84E-06	3.77E-06	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0009611	response to wounding	13/40	854/11978	2.21E-06	1.09E-05	4.66E-06	CRP/HRG/APOA1/APOB/APOE/KLKB1/KNG1 /LBP/PLEK/PROC/C4BPA/C4BPB/PROZ	13
GO:1903035	negative regulation of response to wounding	6/40	120/11978	2.59E-06	1.25E-05	5.35E-06	HRG/APOA1/APOE/KLKB1/KNG1/PROC	6
GO:0016125	sterol metabolic process	6/40	122/11978	2.85E-06	1.35E-05	5.77E-06	APOA1/APOA4/APOB/APOC1/APOE/APOL1	6
GO:0007596	blood coagulation	10/40	487/11978	3.18E-06	1.48E-05	6.30E-06	HRG/APOA1/APOB/APOE/KLKB1/KNG1/PLEK /PROC/C4BPB/PROZ	10
GO:0050817	coagulation	10/40	489/11978	3.29E-06	1.50E-05	6.41E-06	HRG/APOA1/APOB/APOE/KLKB1/KNG1/PLEK /PROC/C4BPB/PROZ	10
GO:0007599	hemostasis	10/40	491/11978	3.42E-06	1.51E-05	6.42E-06	HRG/APOA1/APOB/APOE/KLKB1/KNG1/PLEK /PROC/C4BPB/PROZ	10
GO:0042060	wound healing	11/40	613/11978	3.42E-06	1.51E-05	6.42E-06	CRP/HRG/APOA1/APOB/APOE/KLKB1/KNG1 /PLEK/PROC/C4BPB/PROZ	11
GO:0006954	inflammatory response	10/40	512/11978	4.96E-06	2.15E-05	9.14E-06	CRP/APOA1/APOE/ITIH4/KLKB1/KNG1/ LBP/CD180/C4BPA/C4BPB	10
GO:0002252	immune effector process	10/40	529/11978	6.63E-06	2.82E-05	1.20E-05	CRP/APOA1/IGHG1/IGKC/LBP/CD180/ C1QA/C1R/C4BPA/C4BPB	10
GO:0050776	regulation of immune response	11/40	672/11978	8.27E-06	3.45E-05	1.47E-05	CRP/AMBP/HRG/APOA1/IGHG1/IGKC/LBP /C1QA/C1R/C4BPA/C4BPB	11
GO:0002443	leukocyte mediated immunity	7/40	231/11978	9.82E-06	4.03E-05	1.72E-05	CRP/IGHG1/IGKC/C1QA/C1R/C4BPA/C4BPB	7
GO:0045861	negative regulation of proteolysis	7/40	244/11978	1.40E-05	5.66E-05	2.41E-05	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0050778	positive regulation of immune response	9/40	466/11978	1.76E-05	6.99E-05	2.98E-05	CRP/HRG/IGHG1/IGKC/LBP/C1QA/C1R/C4BPA/ C4BPB	9
GO:0002253	activation of immune response	8/40	357/11978	1.92E-05	7.50E-05	3.20E-05	CRP/IGHG1/IGKC/LBP/C1QA/C1R/C4BPA/C4BPB	8
GO:0051246	regulation of protein metabolic process	17/40	1773/11978	2.15E-05	8.26E-05	3.52E-05	PRG4/AMBP/HRG/APOA1/APOA4/APOE/ITIH1 /ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/APOM/ C4BPA/C4BPB/CALR	17
GO:0051241	negative regulation of multicellular organismal process	11/40	746/11978	2.22E-05	8.38E-05	3.57E-05	PRG4/CRP/HRG/APOA1/APOC1/APOE/KLKB1 /KNG1/LBP/PROC/CALR	11
GO:0043086	negative regulation of catalytic activity	10/40	610/11978	2.30E-05	8.58E-05	3.66E-05	AMBP/HRG/APOA1/APOC1/APOE/ITIH1/ITIH2 /ITIH3/ITIH4/KNG1	10
GO:0045087	innate immune response	11/40	752/11978	2.39E-05	8.76E-05	3.73E-05	APOA4/IGHG1/IGHM/IGKC/LBP/CD180/C1QA /C1R/C4BPA/C4BPB/APOL1	11
GO:0050878	regulation of body fluid levels	10/40	614/11978	2.44E-05	8.81E-05	3.75E-05	HRG/APOA1/APOB/APOE/KLKB1/KNG1/PLEK /PROC/C4BPB/PROZ	10

GO:0015850	organic hydroxy compound transport	6/40	179/11978	2.58E-05	9.17E-05	3.91E-05	APOA1/APOA4/APOB/APOC1/APOE/APOM	6
GO:0006508	proteolysis	13/40	1074/11978	2.72E-05	9.52E-05	4.06E-05	AMBP/HABP2/HRG/APOE/ITIH1/ITIH2/ITIH3 /ITIH4/KLKB1/KNG1/PROC/C1R/PROZ	13
GO:0051336	regulation of hydrolase activity	12/40	934/11978	3.39E-05	0.000117	5.00E-05	AMBP/HRG/APOA1/APOA4/APOC1/ITIH1/ ITIH2/ITIH3/ITIH4/KNG1/PLEK/CALR	12
GO:0009892	negative regulation of metabolic process	17/40	1857/11978	3.98E-05	0.000136	5.78E-05	PRG4/AMBP/HRG/APOA1/APOA4/APOC1 /APOE/ITIH1/ITIH2/ITIH3/ITIH4/KNG1/PLEK /APOM/C4BPA/C4BPB/CALR	17
GO:0032102	negative regulation of response to external stimulus	6/40	196/11978	4.29E-05	0.000144	6.15E-05	HRG/APOA1/APOE/KLKB1/KNG1/PROC	6
GO:0052548	regulation of endopeptidase activity	7/40	291/11978	4.35E-05	0.000144	6.15E-05	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0002684	positive regulation of immune system process	10/40	674/11978	5.41E-05	0.000177	7.54E-05	CRP/HRG/IGHG1/IGKC/LBP/C1QA/C1R /C4BPA/C4BPB/CALR	10
GO:0052547	regulation of peptidase activity	7/40	304/11978	5.75E-05	0.000185	7.90E-05	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0009605	response to external stimulus	16/40	1731/11978	6.76E-05	0.000213	9.08E-05	CRP/HRG/APOA1/APOA4/APOB/APOE /IGHM/KLKB1/KNG1/LBP/CD180/PLEK /PROC/C4BPA/C4BPB/CALR	16
GO:0002682	regulation of immune system process	12/40	1002/11978	6.77E-05	0.000213	9.08E-05	CRP/AMBP/HRG/APOA1/IGHG1/IGKC /LBP/C1QA/C1R/C4BPA/C4BPB/CALR	12
GO:0006066	alcohol metabolic process	7/40	313/11978	6.91E-05	0.000214	9.14E-05	APOA1/APOA4/APOB/APOC1/APOE/ PLEK/APOL1	7
GO:0050727	regulation of inflammatory response	6/40	214/11978	7.01E-05	0.000215	9.15E-05	APOA1/APOE/KLKB1/LBP/C4BPA/C4BPB	6
GO:0080134	regulation of response to stress	11/40	908/11978	0.000134	0.000404	0.000172	AMBP/HRG/APOA1/APOE/KLKB1/KNG1 /LBP/PLEK/PROC/C4BPA/C4BPB	11
GO:0044092	negative regulation of molecular function	10/40	769/11978	0.000163	0.000486	0.000207	AMBP/HRG/APOA1/APOC1/APOE/ITIH1 /ITIH2/ITIH3/ITIH4/KNG1	10
GO:0032269	negative regulation of cellular protein metabolic process	9/40	628/11978	0.000179	0.000527	0.000225	AMBP/HRG/APOE/ITIH1/ITIH2/ITIH3/ ITIH4/KNG1/CALR	9
GO:0008202	steroid metabolic process	6/40	256/11978	0.000187	0.000547	0.000233	APOA1/APOA4/APOB/APOC1/APOE/APOL1	6
GO:0046486	glycerolipid metabolic process	6/40	258/11978	0.000196	0.000563	0.00024	APOA1/APOA4/APOB/APOC1/APOE/PLEK	6
GO:0048584	positive regulation of response to stimulus	14/40	1499/11978	0.000212	0.000604	0.000258	CRP/HRG/APOA1/IGHG1/IGKC/KLKB1/LBP /CD180/PLEK/C1QA/C1R/C4BPA/C4BPB/CALR	14
GO:0048519	negative regulation of biological process	22/40	3395/11978	0.000352	0.00099	0.000422	PRG4/COMP/CRP/AMBP/HRG/APOA1/ APOA4/APOC1/APOE/ITIH1/ITIH2/ITIH3/ ITIH4/KLKB1/KNG1/LBP/PLEK/APOM/PROC /C4BPA/C4BPB/CALR	22
GO:0030162	regulation of proteolysis	8/40	542/11978	0.000356	0.00099	0.000422	AMBP/HRG/APOE/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	8
GO:0048585	negative regulation of response to stimulus	11/40	1022/11978	0.000378	0.00104	0.000443	AMBP/HRG/APOA1/APOE/KLKB1/KNG1/PLEK /PROC/C4BPA/C4BPB/CALR	11
GO:0010605	negative regulation of macromolecule metabolic process	14/40	1593/11978	0.000404	0.001097	0.000468	PRG4/AMBP/HRG/APOA4/APOE/ITIH1/ ITIH2/ITIH3/ITIH4/KNG1/APOM/C4BPA/	14

							C4BPB/CALR	
GO:0031324	negative regulation of cellular metabolic process	14/40	1638/11978	0.000539	0.00145	0.000618	PRG4/AMBP/HRG/APOA4/APOC1/APOE /ITIH1/ITIH2/ITIH3/ITIH4/KNG1/PLEK/ APOM/CALR	14
GO:1901615	organic hydroxy compound metabolic process	7/40	440/11978	0.000561	0.001491	0.000635	APOA1/APOA4/APOB/APOC1/APOE/PLEK /APOL1	7
GO:0044283	small molecule biosynthetic process	6/40	371/11978	0.001335	0.003511	0.001497	APOA1/APOA4/APOB/APOC1/APOE/PLEK	6
GO:0048583	regulation of response to stimulus	18/40	2710/11978	0.001409	0.003666	0.001563	CRP/AMBP/HRG/APOA1/APOE/IGHG1/IGKC /KLKB1/KNG1/LBP/CD180/PLEK/PROC/C1QA /C1R/C4BPA/C4BPB/CALR	18
GO:0050896	response to stimulus	29/40	5803/11978	0.001723	0.004437	0.001891	PRG4/CRP/AMBP/HRG/APOA1/APOA4/APOB /APOE/IGHG1/IGHM/IGKC/ITIH4/KLKB1/KNG1/ LBP/LUM/CD180/PLEK/APOM/PROC/SEPP1/C1QA/ C1R/C4BPA/C4BPB/CALR/APOL1/PROZ/CD5L	29
GO:0006810	transport	20/40	3354/11978	0.002582	0.006578	0.002804	PRG4/CRP/AMBP/HRG/APOA1/APOA4/APOB/ APOC1/APOE/IGHG1/IGKC/KNG1/LBP/PLEK/ APOM/C4BPA/C4BPB/CALR/APOL1/CD5L	20
GO:0065003	macromolecular complex assembly	10/40	1101/11978	0.002705	0.006818	0.002906	CRP/HRG/APOA1/APOA4/APOB/APOC1/APOE /PLEK/APOM/CALR	10
GO:0050790	regulation of catalytic activity	13/40	1710/11978	0.002733	0.006818	0.002906	AMBP/HRG/APOA1/APOA4/APOC1/APOE/ITIH1/ ITIH2/ITIH3/ITIH4/KNG1/PLEK/CALR	13
GO:0051234	establishment of localization	20/40	3440/11978	0.003576	0.008831	0.003764	PRG4/CRP/AMBP/HRG/APOA1/APOA4/APOB/ APOC1/APOE/IGHG1/IGKC/KNG1/LBP/PLEK/ APOM/C4BPA/C4BPB/CALR/APOL1/CD5L	20

Πίνακας ΠΒ2: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία MF (clusterProfiler)

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0004867	serine-type endopeptidase inhibitor activity	6/39	64/11978	5.21E-08	7.43E-07	3.07E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4	6
GO:0004866	endopeptidase inhibitor activity	7/39	121/11978	1.06E-07	7.43E-07	3.07E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/ KNG1	7
GO:0004857	enzyme inhibitor activity	9/39	260/11978	1.12E-07	7.43E-07	3.07E-07	AMBP/HRG/APOA1/APOC1/ITIH1/ ITIH2/ITIH3/ITIH4/KNG1	9
GO:0030414	peptidase inhibitor activity	7/39	123/11978	1.19E-07	7.43E-07	3.07E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/ KNG1	7
GO:0061135	endopeptidase regulator activity	7/39	125/11978	1.33E-07	7.43E-07	3.07E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/ KNG1	7
GO:0008289	lipid binding	11/39	466/11978	1.72E-07	8.01E-07	3.31E-07	CRP/APOA1/APOA4/APOB/APOC1/APOE /IGHM/LBP/PLEK/APOM/APOL1	11
GO:0043178	alcohol binding	6/39	82/11978	2.33E-07	9.32E-07	3.85E-07	CRP/APOA1/APOA4/APOC1/APOE/IGHM	6
GO:0061134	peptidase regulator activity	7/39	156/11978	6.02E-07	2.11E-06	8.72E-07	AMBP/HRG/ITIH1/ITIH2/ITIH3/ITIH4/KNG1	7
GO:0005543	phospholipid binding	8/39	238/11978	7.81E-07	2.27E-06	9.38E-07	APOA1/APOA4/APOB/APOC1/APOE/IGHM /PLEK/APOM	8
GO:0005539	glycosaminoglycan binding	7/39	163/11978	8.10E-07	2.27E-06	9.38E-07	COMP/HABP2/HRG/APOB/APOE/IGHM/ KNG1	7
GO:0030234	enzyme regulator activity	11/39	673/11978	6.41E-06	1.63E-05	6.74E-06	AMBP/HRG/APOA1/APOA4/APOC1/APOE/ ITIH1/ITIH2/ITIH3/ITIH4/KNG1	11
GO:0098772	molecular function regulator	11/39	891/11978	8.77E-05	0.000205	8.46E-05	AMBP/HRG/APOA1/APOA4/APOC1/APOE/ ITIH1/ITIH2/ITIH3/ITIH4/KNG1	11
GO:0005509	calcium ion binding	7/39	501/11978	0.001031	0.00222	0.000918	COMP/CRP/ITIH1/PROC/C1R/CALR/PROZ	7

Πίνακας ΠΒ3: Αποτέλεσμα ανάλυσης εμπλουτισμού γονιδίων Συνόλου Β σε οντολογία CC (clusterProfiler)

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0005615	extracellular space	30/41	969/11978	1.51e-24	7.09e-23	3.97e-23	PRG4/COMP/CRP/AMBP/HABP2/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/C4BPA/C4BPB/CALR/APOL1/CD5L	30
GO:0072562	blood microparticle	16/41	98/11978	9.78e-24	2.3e-22	1.29e-22	AMBP/HRG/APOA1/APOA4/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH4/KNG1/C1R/C4BPA/APOL1/CD5L	16
GO:0005576	extracellular region	35/41	3276/11978	1.23e-14	1.93e-13	1.08e-13	IGLL5/PRG4/COMP/CRP/AMBP/HABP2/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/PLEK/APOM/PROC/SEPP1/C1QA/C1R/C4BPA/C4BPB/CALR/APOL1/PROZ/CD5L	35
GO:0034361	very-low-density lipoprotein particle	7/41	16/11978	3.59e-14	3.37e-13	1.89e-13	APOA1/APOA4/APOB/APOC1/APOE/APOM/APOL1	7
GO:0034385	triglyceride-rich lipoprotein particle	7/41	16/11978	3.6e-14	3.37e-13	1.89e-13	APOA1/APOA4/APOB/APOC1/APOE/APOM/APOL1	7
GO:0044421	extracellular region part	33/41	2903/11978	4.85e-14	3.8e-13	2.15e-13	IGLL5/PRG4/COMP/CRP/AMBP/HABP2/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/C4BPA/C4BPB/CALR/APOL1/PROZ/CD5L	33
GO:0070062	extracellular vesicular exosome	28/41	2213/11978	3.6e-12	1.70e-11	9.55e-12	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:1903561	extracellular vesicle	28/41	2213/11978	3.59e-12	1.71e-11	9.55e-12	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:0043230	extracellular organelle	28/41	2214/11978	3.63e-12	1.71e-11	9.57e-12	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:0065010	extracellular membrane-bounded organelle	28/41	2214/11978	3.63e-12	1.71e-11	9.55e-12	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:0034358	plasma lipoprotein	7/41	30/11978	6.17e-12	2.64e-11	1.48e-11	APOA1/APOA4/APOB/APOC1/APOE/APOM/APOL1	7

	particle							
GO:0032994	protein-lipid complex	7/41	32/11978	1.01e-11	3.98e-11	2.23e-11	APOA1/APOA4/APOB/APOC1/APOE/APOM/APOL1	7
GO:0034364	high-density lipoprotein particle	6/41	19/11978	2.88e-11	1.04e-10	5.83e-11	APOA1/APOA4/APOC1/APOE/APOM/APOL1	6
GO:0031988	membrane-bounded vesicle	28/41	2766/11978	9.3e-10	3.12e-09	1.75e-09	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:0031982	vesicle	28/41	2842/11978	1.8e-09	5.64e-09	3.16e-09	IGLL5/COMP/CRP/AMBP/HRG/APOA1/APOA4/APOB/APOC1/APOE/IGHG1/IGHM/IGKC/ITIH1/ITIH2/ITIH3/ITIH4/KLKB1/KNG1/LBP/LUM/APOM/SEPP1/C1QA/C1R/CALR/PROZ/CD5L	28
GO:0060205	cytoplasmic membrane-bounded vesicle lumen	6/41	71/11978	1.34e-07	3.92e-07	2.2e-07	HRG/APOA1/APOB/APOE/KNG1/CALR	6
GO:0031983	vesicle lumen	6/41	72/11978	1.45e-07	4.02e-07	2.25e-07	HRG/APOA1/APOB/APOE/KNG1/CALR	6
GO:0005788	endoplasmic reticulum lumen	6/41	156/11978	1.37e-05	3.57e-05	2e-05	APOA1/APOA4/APOB/PROC/CALR/PROZ	6
GO:0009986	cell surface	7/41	567/11978	0.002835	0.007013	0.003927	AMBP/APOA1/APOA4/APOE/IGHM/LBP/CALR	7
GO:0044433	cytoplasmic vesicle part	6/41	432/11978	0.003274	0.007695	0.004308	HRG/APOA1/APOB/APOE/KNG1/CALR	6

Πίνακας ΠΒ4: Αποτέλεσμα ανάλυσης εμπλουτισμού για τις πέντε ομαδοποιήσεις των γονιδίων του Συνόλου Β σε οντολογία BP (clusterProfiler)

Cluster	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
C1	GO:0006508	proteolysis	11/13	1421/18585	3.40E-11	1.70E-09	3.94E-10	259/3026/3273/3697/3698/3699/3700/3818/5624/84735/8858	11
C1	GO:0070613	regulation of protein processing	7/13	370/18585	1.81E-09	4.53E-08	1.05E-08	259/3273/3697/3698/3699/3700/3818	7
C1	GO:0010951	negative regulation of endopeptidase activity	6/13	225/18585	4.71E-09	7.26E-08	1.68E-08	259/3273/3697/3698/3699/3700	6
C1	GO:0010466	negative regulation of peptidase activity	6/13	233/18585	5.81E-09	7.26E-08	1.68E-08	259/3273/3697/3698/3699/3700	6
C1	GO:0016485	protein processing	7/13	557/18585	3.07E-08	3.07E-07	7.11E-08	259/3273/3697/3698/3699/3700/3818	7
C1	GO:0045861	negative regulation of proteolysis	6/13	319/18585	3.78E-08	3.15E-07	7.30E-08	259/3273/3697/3698/3699/3700	6
C1	GO:0010955	negative regulation of protein processing	6/13	334/18585	4.97E-08	3.55E-07	8.22E-08	259/3273/3697/3698/3699/3700	6
C1	GO:0051346	negative regulation of hydrolase activity	6/13	363/18585	8.14E-08	4.52E-07	1.05E-07	259/3273/3697/3698/3699/3700	6
C1	GO:0052548	regulation of endopeptidase activity	6/13	363/18585	8.14E-08	4.52E-07	1.05E-07	259/3273/3697/3698/3699/3700	6
C1	GO:0052547	regulation of peptidase activity	6/13	381/18585	1.08E-07	5.42E-07	1.26E-07	259/3273/3697/3698/3699/3700	6
C1	GO:0030162	regulation of proteolysis	6/13	681/18585	3.26E-06	1.48E-05	3.43E-06	259/3273/3697/3698/3699/3700	6
C1	GO:0051336	regulation of hydrolase activity	7/13	1188/18585	5.23E-06	2.18E-05	5.04E-06	259/3273/3697/3698/3699/3700/5341	7
C1	GO:0043086	negative regulation of catalytic activity	6/13	767/18585	6.48E-06	2.49E-05	5.77E-06	259/3273/3697/3698/3699/3700	6
C1	GO:0032269	negative regulation of cellular protein metabolic process	6/13	781/18585	7.19E-06	2.57E-05	5.95E-06	259/3273/3697/3698/3699/3700	6
C1	GO:0051248	negative regulation of protein metabolic process	6/13	852/18585	1.19E-05	3.95E-05	9.16E-06	259/3273/3697/3698/3699/3700	6
C1	GO:0019538	protein metabolic process	11/13	4880/18585	1.83E-05	5.71E-05	1.32E-05	259/3026/3273/3697/3698/3699/3700/3818/5624/84735/8858	11
C1	GO:0044092	negative regulation of molecular function	6/13	969/18585	2.47E-05	7.28E-05	1.68E-05	259/3273/3697/3698/3699/3700	6
C1	GO:0044238	primary metabolic process	13/13	9586/18585	1.82E-04	5.06E-04	1.17E-04	259/3026/3273/3697/3698/3699/3700/3818/4060/5341/5624/84735/8858	13
C1	GO:0031324	negative regulation of cellular metabolic process	7/13	2071/18585	1.97E-04	5.18E-04	1.20E-04	259/3273/3697/3698/3699/3700/5341	7

C1	GO:0050790	regulation of catalytic activity	7/13	2114/18585	2.24E-04	5.61E-04	1.30E-04	259/3273/3697/3698/3699/3700/5341	7
C1	GO:0051246	regulation of protein metabolic process	7/13	2168/18585	2.63E-04	6.27E-04	1.45E-04	259/3273/3697/3698/3699/3700/3818	7
C1	GO:1901564	organonitrogen compound metabolic process	6/13	1493/18585	2.79E-04	6.33E-04	1.47E-04	259/3697/3698/3699/3700/4060	6
C1	GO:0071704	organic substance metabolic process	13/13	9941/18585	2.92E-04	6.36E-04	1.47E-04	259/3026/3273/3697/3698/3699/3700/3818/4060/5341/5624/84735/8858	13
C1	GO:0009892	negative regulation of metabolic process	7/13	2340/18585	4.26E-04	8.88E-04	2.06E-04	259/3273/3697/3698/3699/3700/5341	7
C1	GO:0048519	negative regulation of biological process	9/13	4226/18585	4.65E-04	9.29E-04	2.15E-04	259/3273/3697/3698/3699/3700/3818/5341/5624	9
C1	GO:0043170	macromolecule metabolic process	12/13	8425/18585	5.67E-04	1.09E-03	2.53E-04	259/3026/3273/3697/3698/3699/3700/3818/4060/5624/84735/8858	12
C1	GO:0065009	regulation of molecular function	7/13	2549/18585	7.27E-04	1.35E-03	3.12E-04	259/3273/3697/3698/3699/3700/5341	7
C1	GO:0032268	regulation of cellular protein metabolic process	6/13	1970/18585	1.25E-03	2.23E-03	5.16E-04	259/3273/3697/3698/3699/3700	6
C1	GO:0010605	negative regulation of macromolecule metabolic process	6/13	2006/18585	1.38E-03	2.37E-03	5.49E-04	259/3273/3697/3698/3699/3700	6
C1	GO:0048523	negative regulation of cellular process	8/13	3881/18585	1.67E-03	2.79E-03	6.45E-04	259/3273/3697/3698/3699/3700/5341/5624	8
C1	GO:0008152	metabolic process	13/13	11474/18585	1.89E-03	3.05E-03	7.05E-04	259/3026/3273/3697/3698/3699/3700/3818/4060/5341/5624/84735/8858	13
C1	GO:0080090	regulation of primary metabolic process	9/13	5300/18585	2.73E-03	4.27E-03	9.89E-04	259/3273/3697/3698/3699/3700/3818/4060/5341	9
C1	GO:0044267	cellular protein metabolic process	8/13	4231/18585	3.01E-03	4.57E-03	1.06E-03	259/3273/3697/3698/3699/3700/5624/8858	8
C1	GO:0006950	response to stress	7/13	3522/18585	5.11E-03	7.51E-03	1.74E-03	259/3273/3700/3818/5341/5624/8858	7
C1	GO:0019222	regulation of metabolic process	9/13	6417/18585	1.14E-02	1.63E-02	3.76E-03	259/3273/3697/3698/3699/3700/3818/4060/5341	9
C1	GO:0060255	regulation of macromolecule metabolic process	8/13	5326/18585	1.36E-02	1.89E-02	4.39E-03	259/3273/3697/3698/3699/3700/3818/4060	8
C1	GO:0031323	regulation of cellular metabolic process	8/13	5581/18585	1.83E-02	2.47E-02	5.73E-03	259/3273/3697/3698/3699/3700/4060/5341	8
C2	GO:0072376	protein activation cascade	8/12	90/18585	1.07E-16	5.04E-15	4.51E-16	3500/3503/3514/3827/712/715/722/725	8
C2	GO:0006958	complement activation, classical pathway	7/12	52/18585	6.91E-16	1.62E-14	1.45E-15	3500/3503/3514/712/715/722/725	7
C2	GO:0002455	humoral immune response mediated by circulating immunoglobulin	7/12	67/18585	4.47E-15	6.52E-14	5.84E-15	3500/3503/3514/712/715/722/725	7

C2	GO:0006956	complement activation	7/12	69/18585	5.55E-15	6.52E-14	5.84E-15	3500/3503/3514/712/715/722/725	7
C2	GO:0016064	immunoglobulin mediated immune response	7/12	119/18585	2.85E-13	2.66E-12	2.39E-13	3500/3503/3514/712/715/722/725	7
C2	GO:0019724	B cell mediated immunity	7/12	122/18585	3.40E-13	2.66E-12	2.39E-13	3500/3503/3514/712/715/722/725	7
C2	GO:0006959	humoral immune response	7/12	183/18585	6.08E-12	4.08E-11	3.66E-12	3500/3503/3514/712/715/722/725	7
C2	GO:0006897	endocytosis	9/12	598/18585	7.03E-12	4.13E-11	3.70E-12	10216/3500/3503/3514/722/725/811 /8542/922	9
C2	GO:0002449	lymphocyte mediated immunity	7/12	222/18585	2.38E-11	1.12E-10	1.00E-11	3500/3503/3514/712/715/722/725	7
C2	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	7/12	222/18585	2.38E-11	1.12E-10	1.00E-11	3500/3503/3514/712/715/722/725	7
C2	GO:0002250	adaptive immune response	7/12	262/18585	7.62E-11	3.26E-10	2.92E-11	3500/3503/3514/712/715/722/725	7
C2	GO:0016192	vesicle-mediated transport	10/12	1268/18585	1.23E-10	4.81E-10	4.31E-11	10216/3500/3503/3514/3827/722/725 /811/8542/922	10
C2	GO:0002443	leukocyte mediated immunity	7/12	287/18585	1.44E-10	5.22E-10	4.68E-11	3500/3503/3514/712/715/722/725	7
C2	GO:0006952	defense response	10/12	1533/18585	8.02E-10	2.69E-09	2.41E-10	3500/3503/3514/3827/712/715/722/ 725/8542/922	10
C2	GO:0006909	phagocytosis	6/12	212/18585	1.79E-09	5.61E-09	5.03E-10	3500/3503/3514/722/725/811	6
C2	GO:0002253	activation of immune response	7/12	431/18585	2.46E-09	7.23E-09	6.48E-10	3500/3503/3514/712/715/722/725	7
C2	GO:0002684	positive regulation of immune system process	8/12	803/18585	4.98E-09	1.38E-08	1.23E-09	3500/3503/3514/712/715/722/725/811	8
C2	GO:0045087	innate immune response	8/12	926/18585	1.53E-08	3.78E-08	3.39E-09	3500/3503/3514/712/715/722/725/8542	8
C2	GO:0050778	positive regulation of immune response	7/12	561/18585	1.53E-08	3.78E-08	3.39E-09	3500/3503/3514/712/715/722/725	7
C2	GO:0006955	immune response	9/12	1443/18585	1.78E-08	4.18E-08	3.74E-09	10216/3500/3503/3514/712/715/722/725 /8542	9
C2	GO:0002252	immune effector process	7/12	632/18585	3.47E-08	7.77E-08	6.96E-09	3500/3503/3514/712/715/722/725	7
C2	GO:0002376	immune system process	10/12	2362/18585	5.59E-08	1.19E-07	1.07E-08	10216/3500/3503/3514/712/715/722/725 /811/8542	10
C2	GO:0006950	response to stress	11/12	3522/18585	1.11E-07	2.27E-07	2.03E-08	3500/3503/3514/3827/712/715/722/725 /811/8542/922	11
C2	GO:0002682	regulation of immune system process	8/12	1218/18585	1.30E-07	2.54E-07	2.28E-08	3500/3503/3514/712/715/722/725/811	8
C2	GO:0050776	regulation of immune response	7/12	807/18585	1.86E-07	3.49E-07	3.13E-08	3500/3503/3514/712/715/722/725	7
C2	GO:0048584	positive regulation of response to stimulus	8/12	1771/18585	2.33E-06	4.22E-06	3.78E-07	3500/3503/3514/712/715/722/725/811	8

C2	GO:0019538	protein metabolic process	11/12	4880/18585	3.70E-06	6.44E-06	5.77E-07	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0006810	transport	10/12	4303/18585	1.81E-05	3.04E-05	2.72E-06	10216/3500/3503/3514/3827/722/725/811/8542/922	10
C2	GO:0051234	establishment of localization	10/12	4407/18585	2.27E-05	3.68E-05	3.29E-06	10216/3500/3503/3514/3827/722/725/811/8542/922	10
C2	GO:0048583	regulation of response to stimulus	9/12	3356/18585	2.62E-05	4.11E-05	3.68E-06	3500/3503/3514/3827/712/715/722/725/811	9
C2	GO:0050896	response to stimulus	12/12	7761/18585	2.80E-05	4.24E-05	3.80E-06	10216/3500/3503/3514/3827/712/715/722/725/811/8542/922	12
C2	GO:0044710	single-organism metabolic process	10/12	5038/18585	7.99E-05	1.17E-04	1.05E-05	3500/3503/3514/3827/712/715/722/725/811/8542	10
C2	GO:0051179	localization	10/12	5335/18585	1.36E-04	1.94E-04	1.74E-05	10216/3500/3503/3514/3827/722/725/811/8542/922	10
C2	GO:0044765	single-organism transport	8/12	3640/18585	4.98E-04	6.88E-04	6.17E-05	3500/3503/3514/3827/722/725/811/8542	8
C2	GO:0048518	positive regulation of biological process	9/12	4956/18585	6.58E-04	8.83E-04	7.91E-05	3500/3503/3514/3827/712/715/722/725/811	9
C2	GO:1902578	single-organism localization	8/12	3826/18585	7.11E-04	9.28E-04	8.32E-05	3500/3503/3514/3827/722/725/811/8542	8
C2	GO:0043170	macromolecule metabolic process	11/12	8425/18585	1.16E-03	1.48E-03	1.32E-04	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0044238	primary metabolic process	11/12	9586/18585	4.34E-03	5.37E-03	4.81E-04	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0071704	organic substance metabolic process	11/12	9941/18585	6.26E-03	7.54E-03	6.76E-04	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0065007	biological regulation	11/12	10810/18585	1.44E-02	1.68E-02	1.50E-03	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0044699	single-organism process	12/12	13071/18585	1.46E-02	1.68E-02	1.50E-03	10216/3500/3503/3514/3827/712/715/722/725/811/8542/922	12
C2	GO:0008152	metabolic process	11/12	11474/18585	2.58E-02	2.89E-02	2.59E-03	10216/3500/3503/3514/3827/712/715/722/725/811/8542	11
C2	GO:0050789	regulation of biological process	10/12	10322/18585	4.55E-02	4.98E-02	4.46E-03	10216/3500/3503/3514/3827/712/715/722/725/811	10
C4	GO:0034377	plasma lipoprotein particle assembly	6/7	19/18585	3.32E-18	1.74E-16	2.08E-17	335/337/338/341/348/55937	6
C4	GO:0065005	protein-lipid complex assembly	6/7	21/18585	6.64E-18	1.74E-16	2.08E-17	335/337/338/341/348/55937	6
C4	GO:0034367	macromolecular complex remodeling	6/7	24/18585	1.65E-17	1.74E-16	2.08E-17	335/337/338/341/348/55937	6
C4	GO:0034368	protein-lipid complex remodeling	6/7	24/18585	1.65E-17	1.74E-16	2.08E-17	335/337/338/341/348/55937	6
C4	GO:0034369	plasma lipoprotein particle remodeling	6/7	24/18585	1.65E-17	1.74E-16	2.08E-17	335/337/338/341/348/55937	6
C4	GO:0071827	plasma lipoprotein particle organization	6/7	33/18585	1.35E-16	1.20E-15	1.43E-16	335/337/338/341/348/55937	6

C4	GO:0071825	protein-lipid complex subunit organization	6/7	35/18585	1.98E-16	1.50E-15	1.79E-16	335/337/338/341/348/55937	6
C4	GO:0033344	cholesterol efflux	6/7	42/18585	6.41E-16	4.25E-15	5.06E-16	335/337/338/341/348/55937	6
C4	GO:0097006	regulation of plasma lipoprotein particle levels	6/7	54/18585	3.15E-15	1.86E-14	2.21E-15	335/337/338/341/348/55937	6
C4	GO:0015918	sterol transport	6/7	69/18585	1.46E-14	7.05E-14	8.40E-15	335/337/338/341/348/55937	6
C4	GO:0030301	cholesterol transport	6/7	69/18585	1.46E-14	7.05E-14	8.40E-15	335/337/338/341/348/55937	6
C4	GO:0055088	lipid homeostasis	6/7	104/18585	1.85E-13	7.94E-13	9.46E-14	335/337/338/346/348/55937	6
C4	GO:0006869	lipid transport	7/7	287/18585	1.95E-13	7.94E-13	9.46E-14	335/337/338/341/346/348/55937	7
C4	GO:0010876	lipid localization	7/7	319/18585	4.11E-13	1.56E-12	1.86E-13	335/337/338/341/346/348/55937	7
C4	GO:0042157	lipoprotein metabolic process	6/7	127/18585	6.29E-13	2.22E-12	2.65E-13	335/337/338/341/348/55937	6
C4	GO:0015850	organic hydroxy compound transport	6/7	207/18585	1.23E-11	4.08E-11	4.86E-12	335/337/338/341/348/55937	6
C4	GO:0048878	chemical homeostasis	6/7	919/18585	9.65E-08	3.01E-07	3.59E-08	335/337/338/346/348/55937	6
C4	GO:0006629	lipid metabolic process	6/7	1237/18585	5.68E-07	1.65E-06	1.96E-07	335/337/338/341/346/348	6
C4	GO:0071702	organic substance transport	7/7	2398/18585	5.91E-07	1.65E-06	1.96E-07	335/337/338/341/346/348/55937	7
C4	GO:0033036	macromolecule localization	7/7	2526/18585	8.51E-07	2.25E-06	2.69E-07	335/337/338/341/346/348/55937	7
C4	GO:0065003	macromolecular complex assembly	6/7	1368/18585	1.03E-06	2.61E-06	3.11E-07	335/337/338/341/348/55937	6
C4	GO:0042592	homeostatic process	6/7	1412/18585	1.25E-06	3.00E-06	3.58E-07	335/337/338/346/348/55937	6
C4	GO:0044765	single-organism transport	7/7	3640/18585	1.10E-05	2.54E-05	3.02E-06	335/337/338/341/346/348/55937	7
C4	GO:0022607	cellular component assembly	6/7	2074/18585	1.22E-05	2.68E-05	3.20E-06	335/337/338/341/348/55937	6
C4	GO:0032879	regulation of localization	6/7	2102/18585	1.32E-05	2.79E-05	3.32E-06	335/337/338/341/346/348	6
C4	GO:1902578	single-organism localization	7/7	3826/18585	1.56E-05	3.18E-05	3.79E-06	335/337/338/341/346/348/55937	7
C4	GO:0044085	cellular component biogenesis	6/7	2243/18585	1.93E-05	3.79E-05	4.51E-06	335/337/338/341/348/55937	6
C4	GO:0043933	macromolecular complex subunit organization	6/7	2322/18585	2.36E-05	4.48E-05	5.33E-06	335/337/338/341/348/55937	6
C4	GO:0006810	transport	7/7	4303/18585	3.55E-05	6.49E-05	7.74E-06	335/337/338/341/346/348/55937	7
C4	GO:0051234	establishment of localization	7/7	4407/18585	4.20E-05	7.42E-05	8.84E-06	335/337/338/341/346/348/55937	7
C4	GO:0044710	single-organism metabolic process	7/7	5038/18585	1.07E-04	1.83E-04	2.18E-05	335/337/338/341/346/348/55937	7
C4	GO:0065008	regulation of biological quality	6/7	3217/18585	1.60E-04	2.57E-04	3.07E-05	335/337/338/346/348/55937	6
C4	GO:0051179	localization	7/7	5335/18585	1.60E-04	2.57E-04	3.07E-05	335/337/338/341/346/348/55937	7
C4	GO:0019538	protein metabolic process	6/7	4880/18585	1.77E-03	2.77E-03	3.30E-04	335/337/338/341/348/55937	6

C4	GO:0048518	positive regulation of biological process	6/7	4956/18585	1.94E-03	2.93E-03	3.50E-04	335/337/338/341/346/348	6
C4	GO:0080090	regulation of primary metabolic process	6/7	5300/18585	2.84E-03	4.18E-03	4.98E-04	335/337/338/341/348/55937	6
C4	GO:0016043	cellular component organization	6/7	5548/18585	3.68E-03	5.27E-03	6.28E-04	335/337/338/341/348/55937	6
C4	GO:0071840	cellular component organization or biogenesis	6/7	5663/18585	4.13E-03	5.76E-03	6.87E-04	335/337/338/341/348/55937	6
C4	GO:0044707	single-multicellular organism process	6/7	6392/18585	8.16E-03	1.11E-02	1.32E-03	335/337/338/341/348/55937	6
C4	GO:0019222	regulation of metabolic process	6/7	6417/18585	8.34E-03	1.11E-02	1.32E-03	335/337/338/341/348/55937	6
C4	GO:0044238	primary metabolic process	7/7	9586/18585	9.70E-03	1.25E-02	1.49E-03	335/337/338/341/346/348/55937	7
C4	GO:0032501	multicellular organismal process	6/7	6644/18585	1.01E-02	1.28E-02	1.52E-03	335/337/338/341/348/55937	6
C4	GO:0071704	organic substance metabolic process	7/7	9941/18585	1.25E-02	1.54E-02	1.84E-03	335/337/338/341/346/348/55937	7
C4	GO:0050789	regulation of biological process	7/7	10322/18585	1.63E-02	1.96E-02	2.34E-03	335/337/338/341/346/348/55937	7
C4	GO:0065007	biological regulation	7/7	10810/18585	2.25E-02	2.65E-02	3.16E-03	335/337/338/341/346/348/55937	7
C4	GO:0008152	metabolic process	7/7	11474/18585	3.42E-02	3.94E-02	4.69E-03	335/337/338/341/346/348/55937	7
C4	GO:0043170	macromolecule metabolic process	6/7	8425/18585	3.71E-02	4.19E-02	4.99E-03	335/337/338/341/348/55937	6

Πίνακας ΠΒ5: Αποτέλεσμα ανάλυσης εμπλουτισμού για τις πέντε ομαδοποιήσεις των γονιδίων του Συνόλου Β σε οντολογία MF (clusterProfiler)

Cluster	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
C1	GO:0004867	serine-type endopeptidase inhibitor activity	6/13	95/18585	2.53E-11	3.04E-10	1.07E-10	259/3273/3697/3698/3699/3700	6
C1	GO:0004866	endopeptidase inhibitor activity	6/13	168/18585	8.12E-10	3.12E-09	1.09E-09	259/3273/3697/3698/3699/3700	6
C1	GO:0061135	endopeptidase regulator activity	6/13	173/18585	9.70E-10	3.12E-09	1.09E-09	259/3273/3697/3698/3699/3700	6
C1	GO:0030414	peptidase inhibitor activity	6/13	175/18585	1.04E-09	3.12E-09	1.09E-09	259/3273/3697/3698/3699/3700	6
C1	GO:0061134	peptidase regulator activity	6/13	211/18585	3.20E-09	7.69E-09	2.70E-09	259/3273/3697/3698/3699/3700	6
C1	GO:0004857	enzyme inhibitor activity	6/13	346/18585	6.13E-08	1.23E-07	4.30E-08	259/3273/3697/3698/3699/3700	6
C1	GO:0030234	enzyme regulator activity	6/13	860/18585	1.25E-05	2.15E-05	7.53E-06	259/3273/3697/3698/3699/3700	6
C1	GO:0098772	molecular function regulator	6/13	1148/18585	6.45E-05	9.68E-05	3.40E-05	259/3273/3697/3698/3699/3700	6
C4	GO:0005319	lipid transporter activity	6/7	91/18585	8.13E-14	7.31E-13	1.71E-13	335/337/338/346/348/55937	6
C4	GO:0005543	phospholipid binding	6/7	319/18585	1.68E-10	7.58E-10	1.77E-10	335/337/338/341/348/55937	6
C4	GO:0008289	lipid binding	6/7	605/18585	7.91E-09	2.37E-08	5.55E-09	335/337/338/341/348/55937	6
C4	GO:0022892	substrate-specific transporter activity	6/7	1000/18585	1.60E-07	3.60E-07	8.41E-08	335/337/338/346/348/55937	6
C4	GO:0005215	transporter activity	6/7	1216/18585	5.13E-07	9.23E-07	2.16E-07	335/337/338/346/348/55937	6
C4	GO:0043168	anion binding	6/7	2614/18585	4.74E-05	7.12E-05	1.66E-05	335/337/338/341/348/55937	6
C4	GO:0043167	ion binding	6/7	5981/18585	5.62E-03	7.23E-03	1.69E-03	335/337/338/341/348/55937	6

Πίνακας ΠΒ6: Αποτέλεσμα ανάλυσης εμπλουτισμού για τις πέντε ομαδοποιήσεις των γονιδίων του Συνόλου Β σε οντολογία CC (clusterProfiler)

Cluster	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
C1	GO:0005576	extracellular region	13/13	4454/18585	8.49E-09	1.44E-07	7.15E-08	259/3026/3273/3697/3698/3699/3700/3818/4060/5341/5624/84735/8858	13
C1	GO:0005615	extracellular space	8/13	1252/18585	3.94E-07	3.35E-06	1.66E-06	259/3026/3273/3697/3698/3700/3818/4060	8
C1	GO:0070062	extracellular vesicular exosome	9/13	2759/18585	1.40E-05	3.69E-05	1.83E-05	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C1	GO:1903561	extracellular vesicle	9/13	2759/18585	1.40E-05	3.69E-05	1.83E-05	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C1	GO:0043230	extracellular organelle	9/13	2760/18585	1.40E-05	3.69E-05	1.83E-05	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C1	GO:0065010	extracellular membrane-bounded organelle	9/13	2760/18585	1.40E-05	3.69E-05	1.83E-05	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C1	GO:0044421	extracellular region part	10/13	3698/18585	1.52E-05	3.69E-05	1.83E-05	259/3026/3273/3697/3698/3699/3700/3818/4060/8858	10
C1	GO:0031988	membrane-bounded vesicle	9/13	3473/18585	9.47E-05	2.01E-04	9.97E-05	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C1	GO:0031982	vesicle	9/13	3579/18585	1.21E-04	2.29E-04	1.13E-04	259/3273/3697/3698/3699/3700/3818/4060/8858	9
C2	GO:0072562	blood microparticle	8/12	134/18585	2.86E-15	4.00E-14	1.20E-14	3500/3503/3514/3827/715/722/8542/922	8
C2	GO:0005615	extracellular space	12/12	1252/18585	8.32E-15	5.82E-14	1.75E-14	10216/3500/3503/3514/3827/712/715/722/725/811/8542/922	12
C2	GO:0044421	extracellular region part	12/12	3698/18585	3.80E-09	1.77E-08	5.33E-09	10216/3500/3503/3514/3827/712/715/722/725/811/8542/922	12
C2	GO:0005576	extracellular region	12/12	4454/18585	3.55E-08	1.24E-07	3.74E-08	10216/3500/3503/3514/3827/712/715/722/725/811/8542/922	12
C2	GO:0070062	extracellular vesicular exosome	8/12	2759/18585	6.59E-05	1.16E-04	3.48E-05	3500/3503/3514/3827/712/715/811/922	8
C2	GO:1903561	extracellular vesicle	8/12	2759/18585	6.59E-05	1.16E-04	3.48E-05	3500/3503/3514/3827/712/715/811/922	8
C2	GO:0043230	extracellular organelle	8/12	2760/18585	6.61E-05	1.16E-04	3.48E-05	3500/3503/3514/3827/712/715/811/922	8
C2	GO:0065010	extracellular membrane-bounded organelle	8/12	2760/18585	6.61E-05	1.16E-04	3.48E-05	3500/3503/3514/3827/712/715/811/922	8
C2	GO:0031988	membrane-bounded	8/12	3473/18585	3.55E-04	5.53E-04	1.66E-04	3500/3503/3514/3827/712/	8

		vesicle						715/811/922	
C2	GO:0031982	vesicle	8/12	3579/18585	4.41E-04	6.18E-04	1.86E-04	3500/3503/3514/3827/712/ 715/811/922	8
C4	GO:0034361	very-low-density lipoprotein particle	7/7	20/18585	5.11E-22	5.11E-21	1.34E-21	335/337/338/341/346/348/ 55937	7
C4	GO:0034385	triglyceride-rich lipoprotein particle	7/7	20/18585	5.11E-22	5.11E-21	1.34E-21	335/337/338/341/346/348/ 55937	7
C4	GO:0034358	plasma lipoprotein particle	7/7	38/18585	8.31E-20	5.54E-19	1.46E-19	335/337/338/341/346/348/ 55937	7
C4	GO:0032994	protein-lipid complex	7/7	40/18585	1.23E-19	6.14E-19	1.62E-19	335/337/338/341/346/348/ 55937	7
C4	GO:0034364	high-density lipoprotein particle	6/7	26/18585	2.82E-17	1.13E-16	2.96E-17	335/337/341/346/348/55937	6
C4	GO:0005615	extracellular space	7/7	1252/18585	6.20E-09	2.07E-08	5.44E-09	335/337/338/341/346/348/55937	7
C4	GO:0044421	extracellular region part	7/7	3698/18585	1.23E-05	3.51E-05	9.24E-06	335/337/338/341/346/348/ 55937	7
C4	GO:0005576	extracellular region	7/7	4454/18585	4.52E-05	1.00E-04	2.64E-05	335/337/338/341/346/348/ 55937	7
C4	GO:0032991	macromolecular complex	7/7	4651/18585	6.13E-05	1.00E-04	2.64E-05	335/337/338/341/346/348/ 55937	7
C4	GO:0070062	extracellular vesicular exosome	6/7	2759/18585	6.51E-05	1.00E-04	2.64E-05	335/337/338/341/348/55937	6
C4	GO:1903561	extracellular vesicle	6/7	2759/18585	6.51E-05	1.00E-04	2.64E-05	335/337/338/341/348/55937	6
C4	GO:0043230	extracellular organelle	6/7	2760/18585	6.52E-05	1.00E-04	2.64E-05	335/337/338/341/348/55937	6
C4	GO:0065010	extracellular membrane-bounded organelle	6/7	2760/18585	6.52E-05	1.00E-04	2.64E-05	335/337/338/341/348/55937	6
C4	GO:0031988	membrane-bounded vesicle	6/7	3473/18585	2.50E-04	3.56E-04	9.38E-05	335/337/338/341/348/55937	6
C4	GO:0031982	vesicle	6/7	3579/18585	2.97E-04	3.96E-04	1.04E-04	335/337/338/341/348/55937	6

Πίνακας ΠΒ7: Ανάλυση εμπλουτισμού γονιδίων Συνόλου Α (KEGG)

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
hsa04610	Complement and coagulation cascades	36/70	69/6899	1.28E-57	1.15E-56	1.34E-57	C3/C4A/C4B/SERPINA1/F2/SERPINC1/C4BPA /A2M/KNG1/SERPINA5/SERPINF2/CFH /F5/SERPIND1/C9/SERPING1/F11/KLKB1/PROS1 /C1R/F12/F10/C1QB/PLG/C1QC/CFB/F9/C1S/ PROC/C4BPB/FGA/F7/F8/C1QA/C5/MASP1	36
hsa05150	Staphylococcus aureus infection	13/70	57/6899	6.49E-15	2.92E-14	3.42E-15	C3/C4A/C4B/CFH/C1R/C1QB/PLG/C1QC /CFB/C1S/C1QA/C5/MASP1	13
hsa05133	Pertussis	13/70	75/6899	2.92E-13	8.76E-13	1.02E-13	C3/C4A/C4B/C4BPA/SERPING1/C1R/C1QB /C1QC/C1S/C4BPB/CFL1/C1QA/C5	13
hsa05322	Systemic lupus erythematosus	10/70	136/6899	9.22E-07	1.70E-06	1.99E-07	C3/C4A/C4B/C9/C1R/C1QB/C1QC/C1S/ C1QA/C5	10
hsa05143	African trypanosomiasis	6/70	34/6899	9.44E-07	1.70E-06	1.99E-07	APOA1/HPR/HBB/APOL1/HBA1/HBA2	6
hsa05144	Malaria	6/70	49/6899	8.70E-06	1.31E-05	1.53E-06	HBB/HBA1/HBA2/THBS1/COMP/THBS4	6
hsa04145	Phagosome	6/70	155/6899	0.004683	0.006021	0.000704	C3/C1R/THBS1/COMP/CALR/THBS4	6

Πίνακας ΠΒ8: Αποτέλεσμα ανάλυσης εμπλουτισμού για τα γονίδια του Συνόλου Α (ReactomePA)

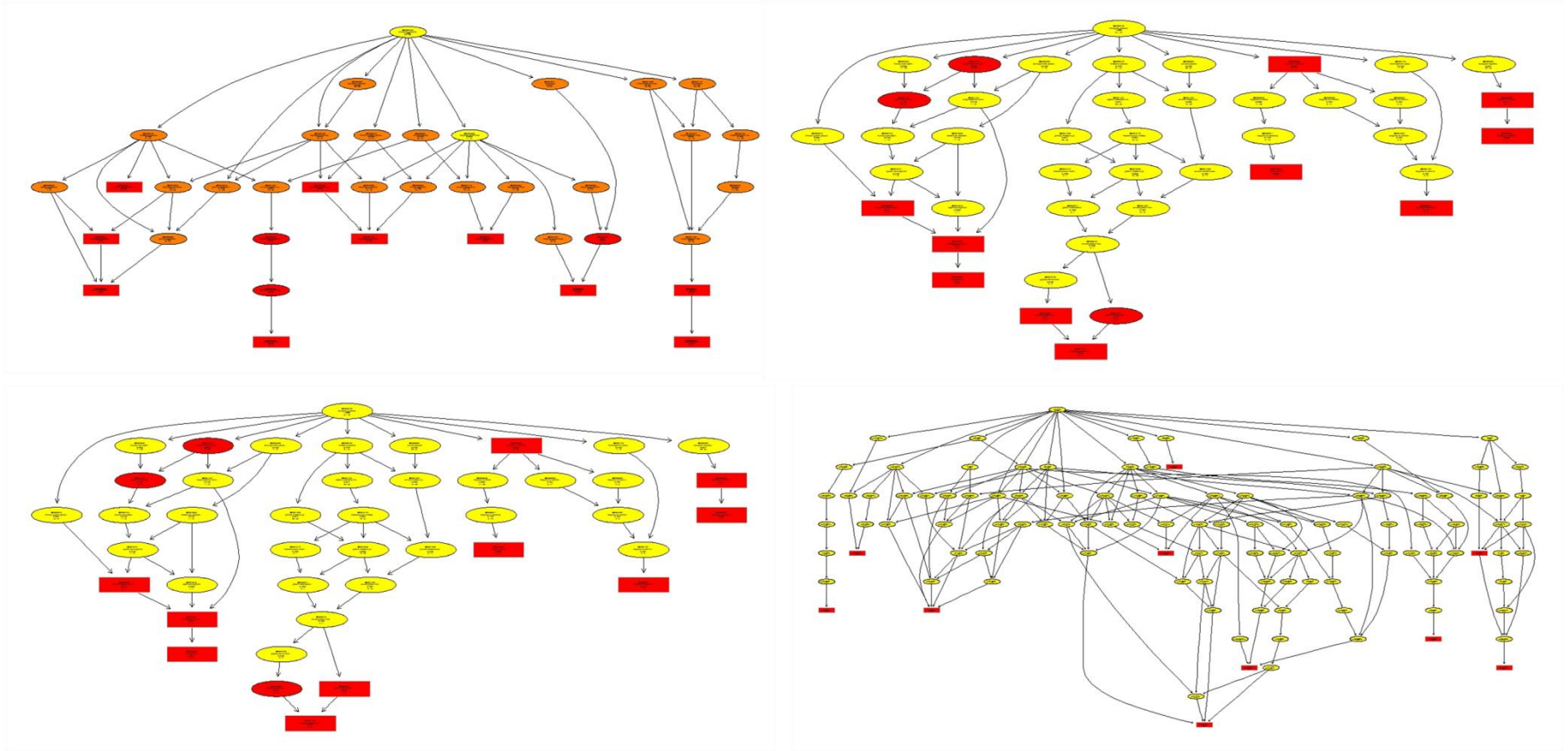
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
140877	Formation of Fibrin Clot (Clotting Cascade)	18/84	36/5302	4.19E-24	1.67E-22	3.53E-23	2147/462/2/3827/5104/2153/3053/710 2160/3818/5627/2161/2159/2158/5624 2243/2155/2157	18
166658	Complement cascade	16/84	32/5302	1.73E-21	3.47E-20	7.30E-21	718/7448/722/3075/735/5627/715/713 629/716/725/2220/5199/712/727/5648	16
114608	Platelet degranulation	20/84	74/5302	2.76E-20	3.68E-19	7.74E-20	335/1191/5265/3273/2/3827/7018/5345 2153/710/5627/5340/7094/5341/2243 7057/2157/5473/1072/2335	20
76005	Response to elevated platelet cytosolic Ca2+	20/84	79/5302	1.17E-19	1.17E-18	2.46E-19	335/1191/5265/3273/2/3827/7018/5345 2153/710/5627/5340/7094/5341/2243/ 7057/2157/5473/1072/2335	20
109582	Hemostasis	33/84	403/5302	2.24E-16	1.79E-15	3.77E-16	335/1191/5265/338/2147/462/3273/2 3827/5104/7018/5345/2153/3053/3043 710/2160/3818/5627/2161/2159/5340 2158/7094/5624/5341/2243/7057/2155 2157/5473/1072/2335	33
140837	Intrinsic Pathway of Fibrin Clot Formation	10/84	16/5302	4.27E-15	2.85E-14	5.99E-15	2147/2/3827/710/2160/3818/2161 2159/2158/2157	10
2173782	Binding and Uptake of Ligands by Scavenger Receptors	12/84	35/5302	6.93E-14	3.96E-13	8.33E-14	335/348/338/3250/3043/8542/259 3240/3263/7184/5648/811	12
140875	Common Pathway of Fibrin Clot Formation	10/84	20/5302	9.36E-14	4.68E-13	9.85E-14	2147/462/5104/2153/3053/5627 2159/5624/2243/2157	10
76002	Platelet activation, signaling and aggregation	21/84	175/5302	1.27E-13	5.65E-13	1.19E-13	335/1191/5265/2147/3273/2 3827/7018/5345/2153/710/5627 5340/7094/5341/2243/7057/2157 5473/1072/2335	21
977606	Regulation of Complement cascade	10/84	21/5302	1.76E-13	6.85E-13	1.44E-13	718/7448/722/3075/735/5627 629/725/5199/727	10

166663	Initial triggering of complement	9/84	15/5302	1.88E-13	6.85E-13	1.44E-13	718/715/713/629/716/2220/5199/712/5648	9
159763	Transport of gamma-carboxylated protein precursors from the endoplasmic reticulum to the Golgi apparatus	7/84	8/5302	1.54E-12	5.12E-12	1.08E-12	2147/5627/2159/2158/5624/8858/2155	7
159740	Gamma-carboxylation of protein precursors	7/84	9/5302	6.83E-12	1.95E-11	4.11E-12	2147/5627/2159/2158/5624/8858/2155	7
159782	Removal of aminoterminal propeptides from gamma-carboxylated proteins	7/84	9/5302	6.83E-12	1.95E-11	4.11E-12	2147/5627/2159/2158/5624/8858/2155	7
2168880	Scavenging of heme from plasma	7/84	10/5302	2.25E-11	5.99E-11	1.26E-11	335/3250/3043/8542/259/3240/3263	7
159854	Gamma-carboxylation, transport, and amino-terminal cleavage of proteins	7/84	11/5302	6.10E-11	1.53E-10	3.21E-11	2147/5627/2159/2158/5624/8858/2155	7
166786	Creation of C4 and C2 activators	6/84	10/5302	2.64E-09	6.20E-09	1.31E-09	715/713/716/2220/712/5648	6
163841	Gamma carboxylation, hypusine formation and arylsulfatase activation	7/84	25/5302	7.43E-08	1.65E-07	3.48E-08	2147/5627/2159/2158/5624/8858/2155	7
3000480	Scavenging by Class A Receptors	6/84	17/5302	1.42E-07	2.99E-07	6.30E-08	335/348/338/7184/5648/811	6
975634	Retinoid metabolism and transport	7/84	34/5302	7.41E-07	1.48E-06	3.12E-07	335/348/338/336/7276/345/55937	7
174824	Lipoprotein metabolism	6/84	22/5302	8.05E-07	1.53E-06	3.23E-07	335/348/338/336/345/3931	6
977225	Amyloids	6/84	25/5302	1.84E-06	3.34E-06	7.04E-07	335/7276/325/2934/4057/2243	6
5653656	Vesicle-mediated transport	12/84	143/5302	1.99E-06	3.47E-06	7.30E-07	335/348/338/3250/3043/8542/259/3240 /3263/7184/5648/811	12
73923	Lipid digestion, mobilization, and transport	6/84	44/5302	5.76E-05	9.60E-05	2.02E-05	335/348/338/336/345/3931	6
2187338	Visual phototransduction	7/84	69/5302	9.50E-05	0.000152	3.20E-05	335/348/338/336/7276/345/55937	7
168249	Innate Immune System	19/84	495/5302	0.000198	0.000305	6.42E-05	718/7448/722/3075/735/3929/5627/715 713/629/716/725/2220/5199/7184/1072 712/727/5648	19
216083	Integrin cell surface interactions	6/84	79/5302	0.001459	0.002161	0.000455	7448/2243/7057/4060/2335/1311	6
1474244	Extracellular matrix organization	10/84	228/5302	0.002989	0.00427	0.000899	7448/7276/2/3818/5340/2243/7057 4060/2335/1311	10
597592	Post-translational protein modification	10/84	250/5302	0.005758	0.007942	0.001672	2147/5627/2159/2158/5624/8858 7057/2155/5199/811	10

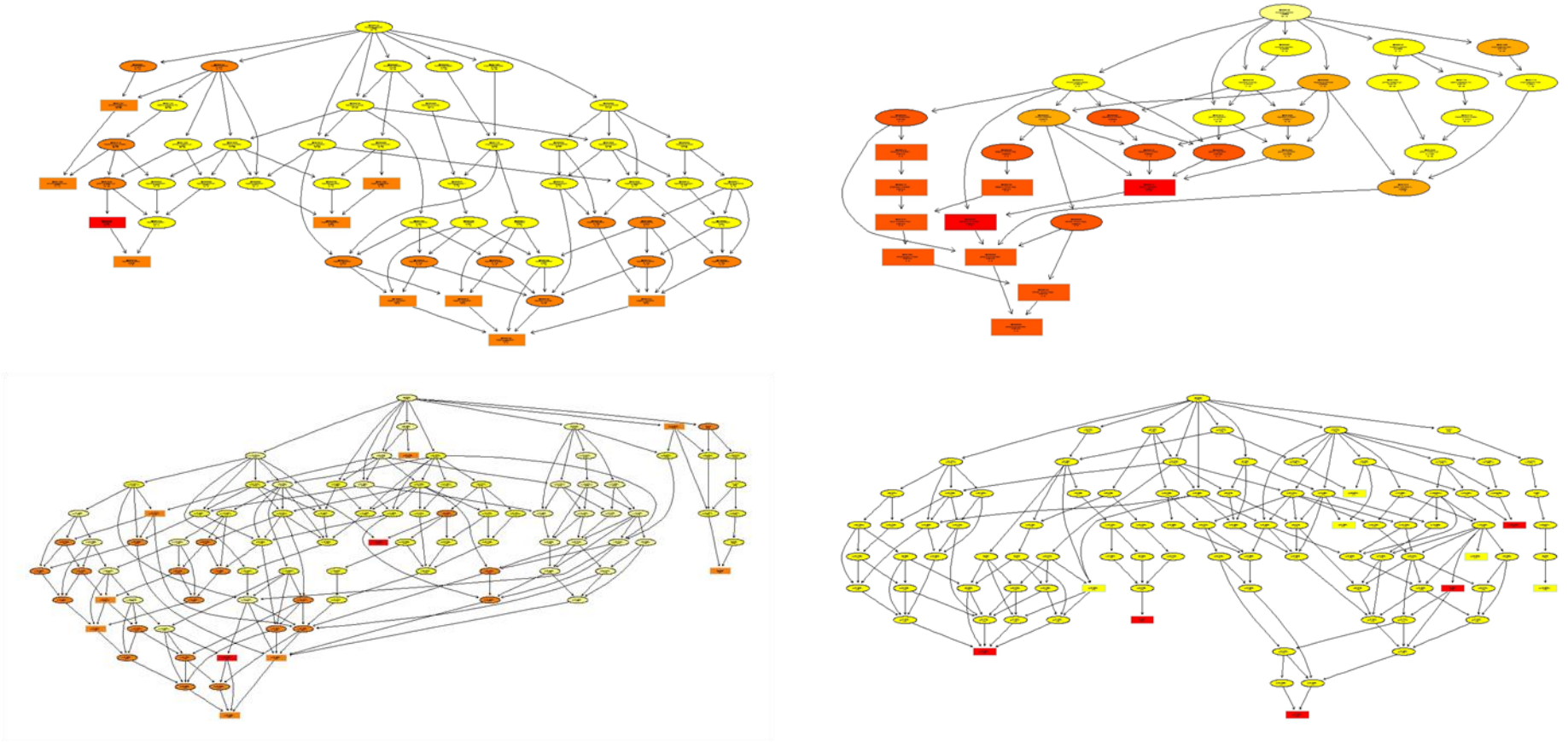
Πίνακας ΠΒ9: Αποτέλεσμα ανάλυσης εμπλουτισμού των ομαδοποιημένων γονιδίων του Συνόλου Β (ReactomePA)

Cluster	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
C1	166658	Complement cascade	18/23	36/6750	2.33E-39	1.63E-38	NA	718/720/721/7448/722/3075/735 715/713/714/629/716/725/2220/5199/71 2/727/5648	18
C1	166663	Initial triggering of complement	12/23	19/6750	3.64E-27	1.28E-26	NA	718/720/721/715/713/714/629/716 2220/5199/712/5648	12
C1	977606	Regulation of Complement cascade	11/23	24/6750	1.00E-22	2.34E-22	NA	718/720/721/7448/722/3075/735 629/725/5199/727	11
C1	168249	Innate Immune System	20/23	569/6750	3.35E-19	5.86E-19	NA	718/720/721/7448/722/3075/735 3929/715/713/714/629/716/725 2220/5199/1072/712/727/5648	20
C1	166786	Creation of C4 and C2 activators	7/23	11/6750	6.35E-16	8.89E-16	NA	715/713/714/716/2220/712/5648	7
C1	168256	Immune System	20/23	942/6750	7.67E-15	8.95E-15	NA	718/720/721/7448/722/3075/735 3929/715/713/714/629/716/725 2220/5199/1072/712/727/5648	20
C1	174577	Activation of C3 and C5	6/23	8/6750	2.15E-14	2.15E-14	NA	718/720/721/629/5199/727	6
C2	174824	Lipoprotein metabolism	6/8	23/6750	2.15E-14	1.50E-13	NA	335/348/336/345/344/3931	6
C2	975634	Retinoid metabolism and transport	6/8	40/6750	8.13E-13	2.85E-12	NA	335/348/336/345/344/55937	6
C2	73923	Lipid digestion, mobilization, and transport	6/8	48/6750	2.59E-12	6.05E-12	NA	335/348/336/345/344/3931	6
C2	2187338	Visual phototransduction	6/8	89/6750	1.22E-10	2.13E-10	NA	335/348/336/345/344/55937	6
C2	556833	Metabolism of lipids and lipoproteins	7/8	509/6750	9.98E-08	1.40E-07	NA	335/348/336/345/344/116519/3931	7
C2	1430728	Metabolism	7/8	1465/675 0	1.45E-04	1.70E-04	NA	335/348/336/345/344/116519/3931	7
C2	162582	Signal Transduction	6/8	2071/675 0	1.27E-02	1.27E-02	NA	335/348/336/345/344/55937	6
C4	140877	Formation of Fibrin Clot (Clotting Cascade)	17/29	37/6750	2.31E-33	3.70E-32	2.43E-33	2147/462/2/3827/2153/3053/710 2160/3818/5627/2161/2159/2158 5624/2243/2155/2157	17
C4	109582	Hemostasis	25/29	450/6750	3.88E-26	3.10E-25	2.04E-26	2147/462/3273/2/3827/7018/5345 /2153/3053/710/2160/3818/5627/ 2161/2159/5340/2158/7094/5624/ 5341/2243/7057/2155/2157/2335	25

C4	114608	Platelet degranulation	15/29	81/6750	2.68E-22	1.43E-21	9.40E-23	3273/2/3827/7018/5345/2153/710 5627/5340/7094/5341/2243/7057 2157/2335	15
C4	76005	Response to elevated platelet cytosolic Ca2+	15/29	86/6750	7.10E-22	2.84E-21	1.87E-22	3273/2/3827/7018/5345/2153/710 5627/5340/7094/5341/2243/7057 2157/2335	15
C4	140837	Intrinsic Pathway of Fibrin Clot Formation	10/29	17/6750	7.12E-21	2.28E-20	1.50E-21	2147/2/3827/710/2160/3818/2161 2159/2158/2157	10
C4	76002	Platelet activation, signaling and aggregation	16/29	189/6750	3.73E-18	9.95E-18	6.55E-19	2147/3273/2/3827/7018/5345/2153 /710/5627/5340/7094/5341/2243 7057/2157/2335	16
C4	140875	Common Pathway of Fibrin Clot Formation	9/29	20/6750	2.05E-17	4.68E-17	3.08E-18	2147/462/2153/3053/5627/2159 5624/2243/2157	9
C4	159763	Transport of gamma-carboxylated protein precursors from the endoplasmic reticulum to the Golgi apparatus	7/29	8/6750	9.86E-17	1.97E-16	1.30E-17	2147/5627/2159/2158/5624/8858 2155	7
C4	159740	Gamma-carboxylation of protein precursors	7/29	9/6750	4.42E-16	7.08E-16	4.66E-17	2147/5627/2159/2158/5624/8858 2155	7
C4	159782	Removal of aminoterminal propeptides from gamma-carboxylated proteins	7/29	9/6750	4.42E-16	7.08E-16	4.66E-17	2147/5627/2159/2158/5624/8858 2155	7
C4	159854	Gamma-carboxylation, transport, and amino-terminal cleavage of proteins	7/29	11/6750	4.03E-15	5.87E-15	3.86E-16	2147/5627/2159/2158/5624/8858 2155	7
C4	163841	Gamma carboxylation, hypusine formation and arylsulfatase activation	7/29	35/6750	7.67E-11	1.02E-10	6.73E-12	2147/5627/2159/2158/5624/8858 2155	7
C4	597592	Post-translational protein modification	8/29	309/6750	3.25E-05	4.00E-05	2.63E-06	2147/5627/2159/2158/5624/8858 7057/2155	8
C4	1474244	Extracellular matrix organization	7/29	249/6750	6.64E-05	7.59E-05	5.00E-06	7276/2/3818/5340/2243/7057/2335	7
C4	392499	Metabolism of proteins	10/29	645/6750	2.20E-04	0.000235	1.55E-05	2147/5627/2159/5340/2158/7094 5624/8858/7057/2155	10
C5	2173782	Binding and Uptake of Ligands by Scavenger Receptors	6/7	40/6750	2.04E-13	4.08E-13	NA	3250/3043/3240/3039/3040/7184	6
C5	5653656	Vesicle-mediated transport	6/7	184/6750	2.59E-09	2.59E-09	NA	3250/3043/3240/3039/3040/7184	6



Σχήμα ΠΒ1: GO γράφηματα ανάλυσης εμπλουτισμού λαμβάνοντας ως σκορ το Q^2 (topGO)



Σχήμα ΠΒ2: GO γραφήματα ανάλυσης εμπλουτισμού λαμβάνοντας ως σκορ το VIP (topGO)

Παράρτημα Γ

ΕΥΡΕΤΗΡΙΟ ΠΕΡΙΓΡΑΦΩΝ GO όρων

GOterm	Περιγραφή [61]
single-organism process	Μια βιολογική λειτουργία στην οποία μετέχει ένας μόνο οργανισμός
biological regulation process	Οποιαδήποτε διαδικασία διαμορφώνει ένα μετρήσιμο χαρακτηριστικό της κάθε βιολογικής διαδικασίας, της ποιότητας ή της λειτουργίας της.
regulation of biological process	Κάθε διαδικασία που διαμορφώνει τη συχνότητα, το ρυθμό ή την έκταση μιας βιολογικής διαδικασίας. Οι βιολογικές διεργασίες ρυθμίζονται από πολλά μέσα, παραδείγματα περιλαμβάνουν τον έλεγχο της γονιδιακής έκφρασης, την τροποποίηση μιας πρωτεΐνης ή την αλληλεπίδραση της με ένα άλλο μόριο πρωτεΐνης ή του υποστρώματος.
response to stimulus	Κάθε διαδικασία που καταλήγει σε μια αλλαγή στην κατάσταση ή στην δραστηριότητα ενός κυττάρου ή ενός οργανισμού (σε όρους κίνησης, έκκρισης, παραγωγής ενζύμου, γονιδιακής έκφρασης, κ.λπ.) ως αποτέλεσμα ενός ερεθίσματος. Η διαδικασία αρχίζει με την ανίχνευση του ερεθίσματος και τελειώνει με μια αλλαγή στη δραστηριότητα του κυττάρου.
cellular process	Κάθε διαδικασία που διεξάγεται στο κυτταρικό επίπεδο, αλλά όχι απαραιτήτως περιορισμένη σε ένα μόνο κύτταρο. Για παράδειγμα, η επικοινωνία των κυττάρων λαμβάνει χώρα ανάμεσα σε περισσότερα από ένα κύτταρα, αλλά διεξάγεται σε κυτταρικό επίπεδο
multicellular organismal process	Οποιαδήποτε βιολογική διεργασία, που συμβαίνει στο επίπεδο ενός πολυκύτταρου οργανισμού σχετική με τη λειτουργία του.
localization	Κάθε διαδικασία στην οποία ένα κύτταρο, μια ουσία, ή μια κυτταρική οντότητα, όπως ένα σύμπλοκο πρωτεΐνης ή οργανίδιο, μεταφέρεται ή διατηρείται σε μια συγκεκριμένη θέση.
positive regulation of biological process	Κάθε διαδικασία που ενεργοποιεί ή αυξάνει τη συχνότητα, το ρυθμό ή την έκταση μιας βιολογικής διαδικασίας.
extracellular region	Ο χώρος εξωτερικά της εξωτερικής δομής ενός κυττάρου. Για κύτταρα χωρίς εξωτερικό προστατευτικό ή εξωτερικές δομές εγκλεισμού αυτό αναφέρεται στο χώρο έξω από την μεμβράνη του πλάσματος.
extracellular region part	Κάθε συστατικό μέρος της εξωκυττάριας περιοχής, δηλαδή ο χώρος έξω από την δομή ενός κυττάρου.
organelle part	Οργανωμένη δομή, ξεχωριστής μορφολογίας και λειτουργίας. Μπορεί να εντοπίζεται στον πυρήνα, στα μιτοχόνδρια, στα πλαστίδια, στα κενοτόπια, στα κυστίδια, στα ριβοσώματα και στον κυτταρικό σκελετό, αλλά όχι στην μεμβράνη του πλάσματος.
cell part	Κάθε συστατικό μέρος του κυττάρου.
cell	Η βασική δομική και λειτουργική μονάδα όλων των οργανισμών. Περιλαμβάνει τη μεμβράνη του πλάσματος και τις εξωτερικές δομές ενθυλάκωσης όπως το κυτταρικό τοίχωμα και το περίβλημα των κυττάρων.
binding	Η εκλεκτική, μη ομοιοπολική, συχνά στοιχειομετρική, αλληλεπίδραση ενός μορίου με μία ή περισσότερες ειδικές θέσεις σε ένα άλλο μόριο.
enzyme regulator	Δέσμευση και ρύθμιση τη δράσης ενός ενζύμου

activity	
biological adhesion	Διεργασία προσκόλλησης ενός κυττάρου ή οργανισμού σε ένα υπόστρωμα ή άλλο οργανισμό.
organelle	Οργανωμένη δομή διακριτής μορφολογίας και λειτουργίας.
channel regulator activity	Ρυθμίζει την δραστικότητα ενός καναλιού. Ένα κανάλι καταλύει ενεργειακά μια ανεξάρτητη διευκολυνόμενη διάχυση, που προκαλείται από το πέρασμα μιας διαλυμένης ουσίας μέσω ενός διαμεμβρανικού υδατικού πόρου ή καναλιού.
response to stimulus	Κάθε διαδικασία που καταλήγει σε μια αλλαγή στην κατάσταση ή δραστηριότητα ενός κυττάρου ή ενός οργανισμού (σε όρους της κίνησης, έκκρισης, της παραγωγής του ενζύμου, γονιδιακής έκφρασης, κ.λπ.) ως αποτέλεσμα ενός ερεθίσματος. Η διαδικασία αρχίζει με την ανίχνευση του ερεθίσματος και τελειώνει με μια αλλαγή στην κατάσταση ή τη δραστηριότητα ή το κύτταρο ή οργανισμό.
positive regulation of biological process	Κάθε διαδικασία που ενεργοποιεί ή αυξάνει τη συχνότητα, το ρυθμό ή την έκταση μιας βιολογικής διαδικασίας.
cellular component organization or biogenesis	Μια διαδικασία που έχει σαν αποτέλεσμα την συναρμολόγηση, τη διάταξη των συστατικών μερών, ή αποσυναρμολόγηση ενός κυτταρικού συστατικού
negative regulation of biological process	Οποιαδήποτε διαδικασία που σταματά, αποτρέπει ή μειώνει τη συχνότητα, το ρυθμό ή την έκταση της βιολογικής διαδικασίας.
extracellular region part	Κάθε συστατικό μέρος της εξωκυττάριας περιοχής, ο χώρος έξω από την δομή ενός κυττάρου
membrane	Διπλό στρώμα μορίων λιπιδίων που περικλείει όλα τα κύτταρα, και σε ευκαρυωτικά κύτταρα, πολλά οργανίδια**.**
membrane-enclosed lumen	Ο όγκος που περικλείεται μέσα σε μια σφραγισμένη μεμβράνη ή μεταξύ δύο σφραγισμένων μεμβρανών.
transporter activity	Ενεργοποιεί την κατευθυνόμενη κίνηση των ουσιών (όπως μακρομόρια, μικρά μόρια, ιόντα) εντός, εκτός ενός κυττάρου, ή μεταξύ των κυττάρων.
primary metabolic process	Η χημική αντίδραση ή το μονοπάτι που περιλαμβάνει τις ενώσεις οι οποίες σχηματίζονται κατά τη διάρκεια αναβολικών ή καταβολικών μονοπατιών.
organic substance metabolic process	Η χημική αντίδραση ή το μονοπάτι που περιλαμβάνει τη συμμετοχή μιας οργανικής ουσίας, δηλαδή ενός μορίου που περιέχει άνθρακα.
response to stress	Κάθε διαδικασία που καταλήγει σε μια αλλαγή στην κατάσταση ή στην δραστηριότητα ενός κυττάρου σε όρους κίνησης, έκκρισης, παραγωγής ενζύμου, γονιδιακής έκφρασης, ως αποτέλεσμα μιας διαταραχής στην κυτταρική ομοιόσταση ή λόγω εξωγενών παραγόντων όπως η θερμοκρασία ή η ακτινοβολία.
regulation of cellular process	Κάθε διαδικασία που διαμορφώνει την συχνότητα, τον ρυθμό ή την έκταση μιας κυτταρικής διαδικασίας, που διεξάγεται στο κυτταρικό επίπεδο, αλλά δεν περιορίζεται σε ένα μόνο κύτταρο.
protein binding	Αλληλεπίδραση επιλεκτική και μη ομοιοπολική με οποιαδήποτε πρωτεΐνη ή σύμπλοκο πρωτεΐνης, που περιλαμβάνει και άλλα μη πρωτεϊνικά μόρια.
ion binding	Αλληλεπίδραση επιλεκτική και μη ομοιοπολική με ιόντα, φορτισμένα άτομα, ή και

	ομάδες αυτών.
enzyme regulator activity	Δεσμεύει και ρυθμίζει τη δράση ενός ενζύμου.
lipid binding	Αλληλεπίδραση επιλεκτική και μη ομοιοπολική με ένα λιπίδιο.
extracellular space	Το τμήμα έξω από τα κύτταρα, συνήθως έξω από την μεμβράνη πλάσματος που καταλαμβάνεται από ρευστό ή αίμα.
membrane-bounded organelle	Οργανωμένη δομή διακριτής μορφολογίας και λειτουργίας, που οριοθετείται από μια μονή η διπλή μεμβράνη διπλοστοιβάδας λιπιδίων. Μπορεί να είναι πυρήνας, μιτοχόνδριο, πλαστίδιο, κενοτόπιο ή και κυστίδιο.
extracellular organelle	Οργανωμένη δομή διακριτής μορφολογίας και λειτουργίας, που βρίσκεται εκτός του κυττάρου, όπως τα εξωκυτταρικά κυστίδια μεμβράνης.
extracellular vesicle	Οποιοδήποτε κυστίδιο είναι μέρος της εξωκυττάριας περιοχής.
protein activation cascade	Μια απόκριση σε ένα ερέθισμα που αποτελείται από μια αλληλουχία τροποποιήσεων σε μια ομάδα πρωτεϊνών. Οι τροποποιήσεις αυτές περιλαμβάνουν την πρωτεόλυση.
complement activation, classical pathway	Οποιαδήποτε διαδικασία εμπλέκεται στην ενεργοποίηση οποιουδήποτε βήματος του κλασσικού μονοπατιού της διαδικασίας των αντιδράσεων του συμπληρώματος, που επιτρέπει την άμεση θανάτωση των μικροβίων, τη διάθεση των ανοσοσυμπλεγμάτων, και τη ρύθμιση άλλων ανοσοποιητικών διαδικασιών.
endocytosis	Μια διαδικασία μεταφοράς κυστιδίων, στην οποία τα κύτταρα λαμβάνουν εξωτερικά υλικά ή συστατικά της μεμβράνης μέσω της εγκόλπωσης μιας μικρής περιοχής της μεμβράνης του πλάσματος για να σχηματιστεί μια νέα μεμβράνη με συνδεδεμένα κυστίδια.
plasma lipoprotein particle assembly	Η συσσωμάτωση και η διάταξη πρωτεϊνών και λιπιδίων για να σχηματιστεί ένα σωματίδιο πλάσματος λιποπρωτεΐνης.
protein-lipid complex assembly	Η συσσωμάτωση και η διάταξη πρωτεϊνών και λιπιδίων για να σχηματιστεί ένα σύμπλεγμα πρωτεΐνης-λιπιδίου.
serine-type endopeptidase inhibitor activity	Σταματά, αποτρέπει ή μειώνει την δραστηριότητα των ενδοπεπτιδάσων τύπου σερίνης, ένζυμα τα οποία καταλύουν την υδρόλυση των μη τερματικών πεπτιδικών δεσμών σε μια πολυπεπτιδική αλυσίδα.
endopeptidase inhibitor activity	Σταματά, αποτρέπει ή μειώνει την δραστηριότητα τα ενδοπεπτιδάσων, ένζυμα τα οποία καταλύουν την υδρόλυση των μη τερματικών πεπτιδικών δεσμών σε μια πολυπεπτιδική αλυσίδα.
enzyme inhibitor activity	Η λειτουργία που σταματάει, αποτρέπει ή μειώνει την δραστηριότητα ενός ενζύμου.
blood microparticle	Ένα κυστίδιο φωσφολιπιδίων, το οποίο προέρχεται από διαφόρους τύπους κυττάρων όπως τα αιμοπετάλια, ενδοθηλιακά κύτταρα κ.ά. Τα κυστίδια αυτά χαρακτηρίζονται ως μικροκυστίδια νουκλεϊκών οξέων.
very-low-density lipoprotein particle	Ένα σωματίδιο λιποπρωτεΐνης πλούσιο σε τριγλυκερίδια. Εντοπίζεται στο αίμα και μεταφέρει ενδογενή προϊόντα (χοληστερόλη, τριγλυκερίδια) από το ήπαρ.
triglyceride-rich lipoprotein particle	Ένα σωματίδιο λιποπρωτεϊνών του πλάσματος που έχει υδροφοβικό πυρήνα εμπλουτισμένο σε τριγλυκερίδια. Αυτά τα σωματίδια μεταφέρουν λιπίδια.

ΕΥΡΕΤΗΡΙΟ ΠΕΡΙΓΡΑΦΩΝ KEGG ΜΟΝΟΠΑΤΙΩΝ

KEGG μονοπάτι	Περιγραφή μονοπατιού[69]
Complement and coagulation cascades	Το σύστημα του συμπληρώματος είναι μια αλληλουχία αντιδράσεων στο πλάσμα του αίματος και ένας μεσολαβητής της έμφυτης ανοσίας του οργανισμού, δηλαδή ένας μη ειδικός αμυντικός μηχανισμός έναντι των παθογόνων ουσιών.
Staphylococcus aureus infection	Η μόλυνση από το “Staphylococcus aureus” βακτήριο μπορεί να προκαλέσει πολλαπλές μορφές λοιμώξεων που κυμαίνονται από επιφανειακές δερματικές λοιμώξεις σε τροφικές δηλητηριάσεις ακόμα και σε απειλητικές λοιμώξεις για τη ζωή.
Pertussis	Ο κοκκύτης είναι μια οξεία λοιμώδης νόσος του αναπνευστικού συστήματος που προκαλείται από ένα βακτήριο που ονομάζεται “Bordetella pertussis”.
Systemic lupus erythematosus	Ο συστηματικός ερυθηματώδης λύκος είναι μια αυτοάνοση νόσος που χαρακτηρίζεται από την παραγωγή αυτοαντισωμάτων IgG, έναντι αντιγόνων όπως το DNA, οι πυρηνικές πρωτεΐνες και ορισμένα κυτταροπλασματικά συστατικά.
African trypanosomiasis	Το παράσιτο “Trypanosoma brucei” είναι υπεύθυνο για την αφρικανική τρυπανοσωμίαση. Διαδίδεται από την μύγα τσε-τσε. Τα παράσιτα περνούν μέσα από το φράγμα αίματος-εγκεφάλου και μπορούν να προκαλέσουν νευρολογικές βλάβες.
Malaria	Τα πρωτόζωα “Plasmodium” είναι παράσιτα που ευθύνονται για τη μόλυνση της ελονοσίας.
Phagosome	Η φαγοκυττάρωση είναι η διαδικασία της λήψης σχετικά μεγάλων σωματιδίων από ένα κύτταρο και ένας κεντρικός μηχανισμός αναδιαμόρφωσης ιστών του κυττάρου και άμυνας εναντίον μολυσματικών παραγόντων.

ΕΥΡΕΤΗΡΙΟ ΠΕΡΙΓΡΑΦΩΝ REACTOME ΜΟΝΟΠΑΤΙΩΝ

Reactome Μονοπάτι	Περιγραφή μονοπατιού [61]
Formation of Fibrin Clot (Clotting Cascade))	Ο σχηματισμός ενός ινώδης θρόμβου στην περίπτωση ενός τραυματισμού στο τοίχωμα ενός φυσιολογικού αιμοφόρου αγγείου είναι ένα βασικό μέρος της διαδικασίας διακοπής της απώλειας αίματος μετά από αγγειακό τραυματισμό. Οι αντιδράσεις που οδηγούν στον σχηματισμό του ινώδους θρόμβου συνήθως περιγράφονται από μια αλληλουχία διαδικασιών, στις οποίες, το προϊόν κάθε βήματος είναι ένα ένζυμο ή ένας συμπαράγοντας, που απαιτείται για να προχωρήσουν αποτελεσματικά οι ακόλουθες αντιδράσεις.
Complement cascade	Το σύστημα του συμπληρώματος είναι η πρώτη γραμμή άμυνας κατά της εισβολής των μικροβίων. Αποτελείται από ένα μεγάλο αριθμό πρωτεϊνών, οι οποίες κυκλοφορούν στο αίμα σε ανενεργή μορφή. Όταν ενεργοποιηθούν, συστατικά του συμπληρώματος συγκεντρώνονται πάνω στην επιφάνεια του κυττάρου στόχου. Η ενεργοποίηση του συμπληρώματος οδηγεί σε τέσσερα κύρια αποτελέσματα: (1) την οψωνοποίηση των κυττάρων στόχων για την ενίσχυση της φαγοκυττάρωσης (2) την λύση των κυττάρων στόχων μέσω ενός συγκροτήματος του συμπλόκου της προσβολής της μεμβράνης στην επιφάνεια του κυττάρου στόχου (3) παραγωγή αναφυλατοξινών που εμπλέκονται στην φλεγμονώδη απόκριση του ξενιστή (4)στην κάθαρση των συμπλόκων αντισώματος-αντιγόνου
Platelet degranulation	Τα αιμοπετάλια λειτουργούν ως εξωκυτταρικά κύτταρα, εκκρίνοντας μια πληθώρα μορίων σε θέσεις αγγειακής βλάβης. Περιέχουν έναν αριθμό κοκκιδίων αποθήκευσης. Όταν ενεργοποιούνται τα αιμοπετάλια απελευθερώνουν μια ποικιλία πρωτεϊνών, κυρίως από τα κοκκίδια αποθήκευσης. Αυτά δρουν με τρόπο ώστε να διαμορφώνουν την κυτταρική σηματοδότηση.
Response to elevated platelet cytosolic Ca2+	Η ενεργοποίηση της φωσφολιπάσης C των κυττάρων έχει ως αποτέλεσμα την παραγωγή των δευτέρων αγγελιοφόρων του μονοπατιού της φωσφατιδυλινοσιτόλη. Αυτό οδηγεί στην αύξηση του ασβεστίου και στην ενεργοποίηση της πρωτεϊνικής κινάσης C. Μια κινάση είναι ένας τύπος ενζύμου που καταλύει τη μεταφορά φωσφορικών ομάδων από υψηλής ενέργειας μόρια φωσφορικού σε συγκεκριμένα υποστρώματα. Η πρωτεϊνική κινάση C, γνωστή και ως PKC, είναι μια οικογένεια ενζύμων πρωτεϊνικής κινάσης τα οποία εμπλέκονται στον έλεγχο της λειτουργίας άλλων πρωτεϊνών μέσω της φωσφορυλίωσης των ομάδων υδροξυλίου της σερίνης και της θρεονίνης στις πρωτεΐνες.
Hemostasis	Η αιμόσταση είναι μια φυσιολογική αντίδραση η οποία κορυφώνεται στην περίπτωση αιμορραγίας ενός τραυματισμένου αγγείου. Σε συνθήκες αγγειακού τραύματος, κυριαρχούν αγγειοσταλτικοί μηχανισμοί .Τρεις μηχανισμοί συμβάλλουν στην αναστολή της απώλειας αίματος μετά από τραυματισμό ενός αγγείου. Το αγγείο συστέλλεται, μειώνοντας την απώλεια αίματος. Τα αιμοπετάλια προσκολλώνται στην περιοχή του τραυματισμού, ενεργοποιούνται και συσσωματώνονται με ινωδογόνο σε μια δομή που περιορίζει την απώλεια αίματος, συντελώντας μια διαδικασία η οποία ονομάζεται αρχική αιμόσταση. Το ινωδογόνο παράγεται στο ήπαρ και ανήκει στις πρωτεΐνες οξείας φάσης. Πρωτεΐνες και μικρά μόρια απελευθερώνονται από τα κοκκίδια των ενεργοποιημένων αιμοπεταλλίων, διεγείροντας την διαδικασία σχηματισμού της παραπάνω δομής. Το ινωδογόνο του πλάσματος αποτελεί τη γέφυρα μεταξύ των ενεργοποιημένων αιμοπεταλίων. Έτσι κινείται η διεργασία της πήξης (δευτεροβάθμια αιμόσταση).
Intrinsic Pathway of Fibrin Clot Formation	Το ενδογενές μονοπάτι της πήξης του αίματος συνδέει αλληλεπιδράσεις μεταξύ κινινογόνου (υψηλού μοριακού βάρους κινινογόνο HK), προκαλλικρεινης (PK) και του παράγοντα XII, ώστε να ενεργοποιηθεί ο παράγοντας πήξης X από μια σειρά αντιδράσεων που είναι ανεξάρτητες από την εξωγενή οδό.
Regulation of Complement Cascade	Η ενεργοποίηση των διαδικασιών του συμπληρώματος ρυθμίζεται από μια οικογένεια συγγενών πρωτεϊνών που ονομάζονται ρυθμιστές της ενεργοποίησης του συμπληρώματος (RCA). Αυτές οι πρωτεΐνες εκφράζονται σε υγιή κύτταρα ξενιστές. Τα περισσότερα παθογόνα δεν εκφράζουν RCA πρωτεΐνες στην επιφάνειά τους, αλλά πολλά

	από αυτά έχουν βρει τρόπους να αποφεύγουν το σύστημα του συμπληρώματος απλά δεσμεύοντας τις RCA πρωτεΐνες που κυκλοφορούν στο ανθρώπινο πλάσμα. Η παγίδευση των RCA πρωτεϊνών είναι η πιο συχνά χρησιμοποιούμενη στρατηγική για την αποφυγή της απόκρισης του συμπληρώματος.
Binding and Uptake of Ligands by Scavenger Receptors	Η λειτουργία των υποδοχέων αυτών είναι να συμμετέχουν στην απομάκρυνση ξένων ουσιών και άχρηστων παραγώγων μέσω μιας ποικιλίας μορίων που λειτουργούν ως υποδοχείς. Οι υποδοχείς Scavenger δεσμεύουν τους ελεύθερους εξωκυττάριους συνδέτες. Κάποιοι είναι ειδικοί για την απομάκρυνση ενός συνδέτη ενώ άλλοι για απομάκρυνση πολλών συνδετών μαζί.
Common Pathway of Fibrin Clot Formation	Το κοινό μονοπάτι αποτελείται από μια αλληλουχία γεγονότων ενεργοποίησης που οδηγούν από την σχηματισμό του ενεργοποιημένου παράγοντα X στον σχηματισμό της ενεργούς θρομβίνης, την διάσπαση του ινωδογόνου από την θρομβίνη και το σχηματισμό του ινώδους σε ένα σταθερό, πολλαπλομερές σύμπλοκο. Η ποσοτική αλληλεπίδραση μεταξύ των ρυθμιστών των παραπάνω αντιδράσεων είναι κρίσιμη για την κανονική ρύθμιση της πήξης, διευκολύνοντας την ταχεία διαμόρφωση ενός προστατευτικού θρόμβο στη θέση της βλάβης.
Platelet activation, signaling and aggregation	Η ενεργοποίηση των αιμοπεταλίων αρχίζει την αρχική ένωση τους με μόρια συνδέτες (απελευθερώνονται ή δημιουργούνται στις θέσεις αγγειακού τραύματος) ώστε να προσελκύσουν υποδοχείς στην μεμβράνη τους. Οι αντιδράσεις της ενδοκυτταρικής σηματοδότησης ενισχύουν τους συνδέτες και τις προπηκτικές ιδιότητες των προσδεμένων αιμοπεταλίων ή των αιμοπεταλίων που κυκλοφορούν ελεύθερα. Μόλις τα αιμοπετάλια προσκολληθούν αποσυντίθενται και απελευθερώνουν παράγοντες όπως το ADP, το ATP και θρομβοξάνης A2. Έτσι ενισχύεται η απόκριση, η ενεργοποίηση και η πρόσληψη επιπλέον αιμοπεταλίων στην περιοχή και προωθείται η συσσώρευση τους.
Initial triggering of complement	Η ενεργοποίηση του συμπληρώματος πραγματοποιείται μέσω από μια αλληλουχία πρωτεολυτικών αντιδράσεων, οι οποίες εκτελούνται από τις πρωτεάσες της σερίνης.