ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

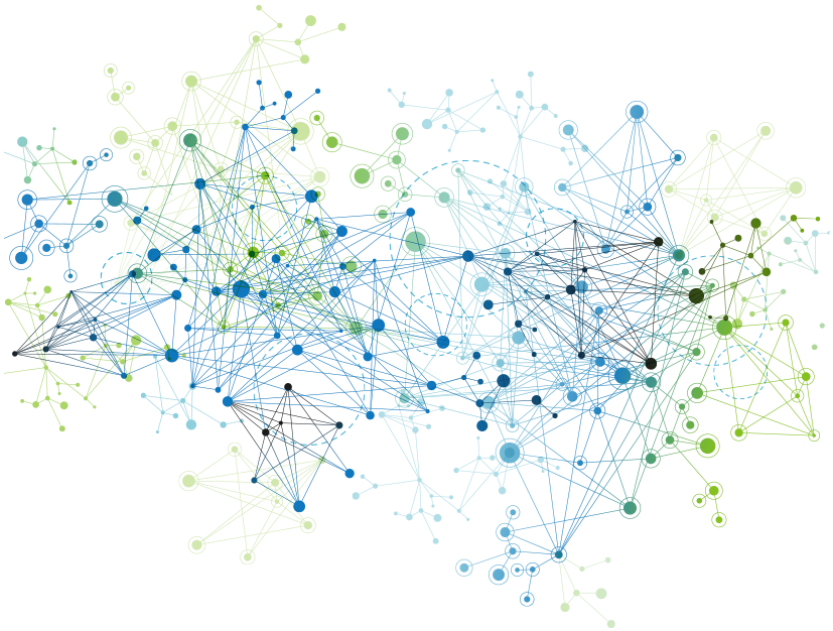
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Συγκριτική ανάλυση αλγορίθμων ομαδοποίησης σε βιοϊατρικά δίκτυα»



Ιωάννα Χοτόβα

Πτυχιούχος Τμήματος Πληροφορικής, Ο.Π.Α

ΑΘΗΝΑ 2021

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

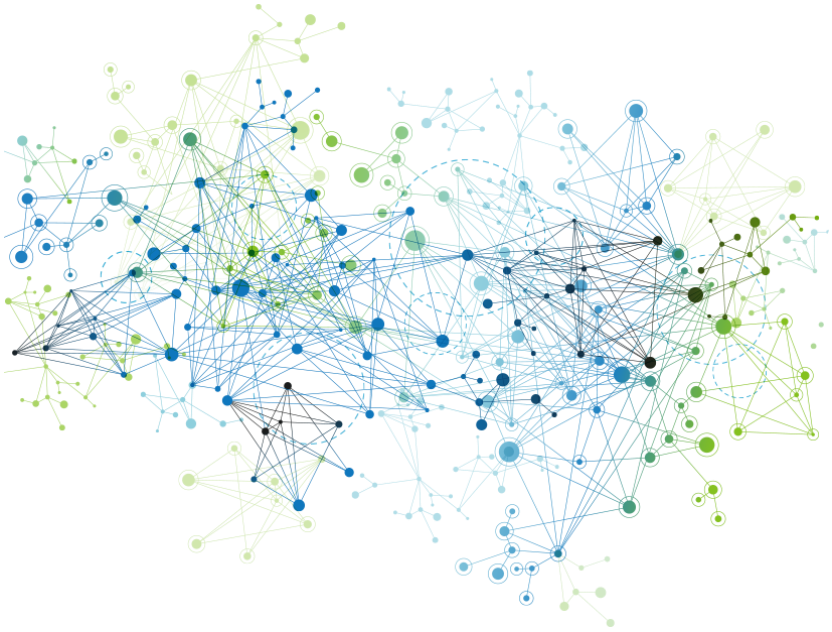
HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS»

Master Diploma Thesis

«Comparative analysis of clustering algorithms in biomedical networks»



Joana Hotova

Computer Science, A.U.E.B

A T H E N S 2 0 2 1

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

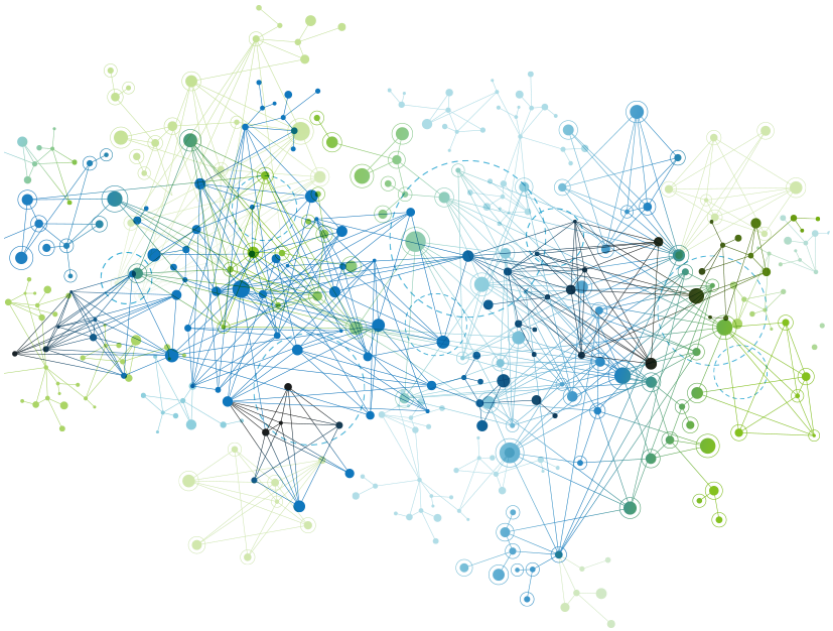
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Συγκριτική ανάλυση αλγορίθμων ομαδοποίησης σε βιοϊατρικά δίκτυα»



Τριμελής εξεταστική επιτροπή

Κύριος επιβλέπων: Δρ. Γεώργιος Παυλόπουλος
Κύριος Ερευνητής Β' – Ε.ΚΕ.Β.Ε "Αλέξανδρος Φλέμινγκ"

Επιβλέπων ΠΜΣ: Δρ. Παντελής Μπάγκος
Καθηγητής Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστημίου Θεσσαλίας

Δρ. Βασιλική Οικονομίδου
Αναπληρώτρια Καθηγήτρια Βιοφυσικής – Μοριακής Βιοφυσικής, Τμήματος Βιολογίας, Ε.Κ.Π.Α

Πίνακας περιεχομένων

Ευχαριστίες

Για την παρούσα διπλωματική εργασία θα ήθελα να εκφράσω τις εγκάρδιες ευχαριστίες μου προς τον επιβλέποντα καθηγητή, κ. Δρ. Γεώργιο Παυλόπουλο, ερευνητή στο Ε.ΚΕ.Β.Ε "Αλέξανδρος Φλέμινγκ'', για την πολύτιμη καθοδήγηση και βοήθεια του. Οι παρατηρήσεις και οι συμβουλές του, καθώς και η υπομονή που επέδειξε καθ' όλη τη περίοδο της συνεργασίας μας, ήταν τα κύρια συστατικά για την ολοκλήρωση της διπλωματικής εργασίας.

Επιπλέον, θα ήθελα να αποδώσω τις ευχαριστίες μου στα μέλη της τριμελούς επιτροπής την κα. Δρ. Βασιλική Οικονομίδου, Αναπληρώτρια Καθηγήτρια Βιοφυσικής – Μοριακής Βιοφυσικής του Τμήματος Βιολογίας Ε.Κ.Π.Α και τον κ. Δρ. Παντελή Μπάγκο, Καθηγητή στο Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, του Πανεπιστημίου Θεσσαλίας, για την συμβολή τους στην εκπαίδευση μου πάνω στον τομέα της Βιοπληροφορικής και για τα εφόδια που μου έδωσαν.

Στις ευχαριστίες μου δεν θα μπορούσα να παραλείψω τους Μεταδιδακτορικούς Ερευνητές του ΕΚΕΒΕ "Αλέξανδρος Φλέμινγκ'' Ευάγγελο Καρατζά και Φώτη Μπαλτούμα, την Μεταπτυχιακή συμφοιτήτρια μου Μαρία Γκόντα και τον Chris Bobotsis από το Πανεπιστήμιο Waterloo του Καναδά, για την σπουδαία βοήθεια τους στην παρούσα εργασία. Επίσης, η συμβολή της Μεταδιδακτορικής Ερευνήτριας του Πανεπιστημίου Θεσσαλίας στο Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Παναγιώτα Κοντού στο case study του VICTOR, ήταν ιδιαιτέρως σημαντική. Ακόμη, η βοήθεια και η υποστήριξη από την γραμματεία του Π.Μ.Σ "Βιοπληροφορική – Υπολογιστική Βιολογία", ήταν εξίσου πολύτιμη.

Τέλος, η εκπόνηση της συγκεκριμένης διπλωματικής εργασίας δεν θα μπορούσε να ήταν το ίδιο ευχάριστη και αποτελεσματική χωρίς το στήριγμα της οικογένειας μου, την βοήθεια της καλής μου φίλης Μαρίας Βατικιώτη και την αγάπη των υπόλοιπων φίλων μου, όσο κοντά ή μακριά και αν ήταν λόγω των δυσκολιών που μας επέφερε η πανδημία.

Ιωάννα Χοτόβα
Αθήνα, Ιούνιος 2021

Περίληψη

Ως ομαδοποίηση/συσταδοποίηση (clustering) βιολογικών δικτύων χαρακτηρίζουμε τη διαδικασία σύμφωνα με την οποία οι κόμβοι ενός δικτύου μπορούν να καταταχθούν σε μια κοινή ομάδα σύμφωνα με κοινά τους χαρακτηριστικά. Για το σκοπό αυτό υπάρχει μια πληθώρα αλγορίθμων ομαδοποίησης οι οποίοι μπορούν να αυτοματοποιούν αυτή τη διαδικασία ακολουθώντας διαφορετικές στρατηγικές και μεθοδολογίες. Αξίζει να σημειωθεί ότι οι περισσότεροι από τους αλγόριθμους αυτούς λαμβάνουν υπόψη την τοπολογία του δικτύου. Ως εκ τούτου, διαφορετικοί αλγόριθμοι ομαδοποίησης, μπορεί συχνά να οδηγήσουν σε διαφορετικά αποτελέσματα ακόμα και για το ίδιο σύνολο δεδομένων.

Στα πλαίσια της διπλωματικής αυτής, αρχικά συγκεντρώθηκαν πολλαπλά βιολογικά δίκτυα διαφορετικού τύπου από διάφορες βάσεις δεδομένων και μελετήθηκαν ως προς την τοπολογία τους. Οι κατηγορίες δικτύων που αναλύθηκαν είναι τα δίκτυα αλληλεπίδρασης πρωτεϊνών, τα δίκτυα συν-έκφρασης γονιδίων καθώς και τα δίκτυα ομοιότητας αλληλουχιών. Στη συνέχεια, εφαρμόστηκαν διαφορετικοί αλγόριθμοι ομαδοποίησης δεδομένων, τα αποτελέσματα των οποίων συγκρίθηκαν τόσο μεταξύ τους όσο και με ομάδες δεδομένων που προήλθαν ύστερα από εφαρμογή ροών που αφορούν το λειτουργικό εμπλουτισμό των κόμβων του εκάστοτε δικτύου. Επίσης, παρουσιάζεται το μέτρο conductance (αγωγιμότητα), με τη χρήση του οποίου φαίνεται μέσω ιστογραμμάτων, η ποιότητα της κάθε ομάδας (cluster) ενός δικτύου στο υπόλοιπο αρχικό δίκτυο. Τέλος, μέσω της δημιουργίας του εργαλείου VICTOR μπορούν να εφαρμοστούν οι διάφορες μετρικές ώστε να μελετηθούν τα αποτελέσματα των αλγορίθμων ομαδοποίησης μέσω της οπτικής ανάλυσης.

Θεματική Περιοχή: Βιολογικά Δίκτυα, Συγκριτική Ανάλυση, Βιοπληροφορική

Λέξεις Κλειδιά: Δίκτυα, Αλγόριθμοι Ομαδοποίησης Δικτύων, Λειτουργική Ανάλυση, Conductance

Abstract

As clustering of biological networks, it's called the process according to which the nodes of a network can be classified in a common group according to their common features. In order to succeed this process automatically, there are a variety of clustering algorithms based on different strategies and methodologies. Note that most of these algorithms take into account the network topology. Therefore, different clustering algorithms can often bring out different results even for the same data set.

In order to proceed with this Thesis, initially multiple biological networks of different types were collected from various databases. The collected biological networks that were analyzed are protein interaction networks, gene co-expression networks, and sequence similarity networks. Firstly, these networks were studied in terms of their topology. After that, different data clustering algorithms were applied, the results of which were compared both with each other and with data sets that came after the application of flows related to the functional enrichment of the nodes of each network. Also, as part of this Thesis, it's presented the conductance measure which shows, through histograms, the quality of each cluster in the rest of the original network. Finally, through the development of the VICTOR tool, various comparison metrics can be applied through which the results of clustering algorithms can be compared via visual analysis.

Thematic Area: Biological Networks, Comparative Analysis, Bioinformatics

Keywords: Networks, Network Clustering Algorithms, Functional Analysis, Conductance

1. Γράφοι και μαθηματική μοντελοποίηση

1.1. Η έννοια του γράφου

Τα ***δίκτυα*** ή αλλιώς ***γράφοι*** όπως ονομάζονται στη γλώσσα των μαθηματικών, είναι ένας τρόπος απεικόνισης των σχέσεων μεταξύ διάφρων οντοτήτων. Κατά κύριο λόγο, ένας γράφος αποτελείται από δύο βασικά συστατικά: τους κόμβους και τις ακμές οι οποίες συνδέουν ένα πλήθος από κόμβους μεταξύ τους.

Ο μαθηματικός συμβολισμός ενός γράφου είναι ο εξής: $G = (V, E)$ όπου το $\boldsymbol{G}$ βγαίνει από την λέξη Graph, το $\boldsymbol{V}$ (Vertices) αντιπροσωπεύει τους κόμβους και το $\boldsymbol{E}$ (Edges) αντιπροσωπεύει τις ακμές. Ένας γράφος μπορεί να περιλαμβάνει έναν αριθμό από ***υπογράφους (subgraphs)***. Ο υπογράφος απεικονίζεται ως εξής: $G' = (V', E')$ όπου G' είναι το όνομα του υπο-γράφου, V' είναι ένα υποσύνολο των κόμβων του αρχικού γράφου και E'ένα υποσύνολο των ακμών του αρχικού γράφου.

Ένας γράφος μπορεί να απεικονιστεί με πολλούς τρόπους. Αυτό δίνει την δυνατότητα σε δύο γράφους που έχουν τον ίδιο αριθμό από κόμβους και τις ίδιες συνδέσεις να λέγονται ***ισομορφικοί***. Οι ισομορφικοί γράφοι συμβολίζονται ως εξής: $G1 \simeq G2$.

Στην καθημερινή μας ζωή συναντάμε πολλούς διαφορετικούς γράφους σε μορφή δικτύου. Κάποια χαρακτηριστικά παραδείγματα είναι το δίκτυο ύδρευσης, το οδικό δίκτυο, το Διαδίκτυο, οι τηλεπικοινωνίες, τα μέσα κοινωνικής δικτύωσης κ.ο.κ. Στις βιοεπιστήμες, οι γράφοι αποτελούνται κυρίως από βιομόρια όπως οι πρωτεΐνες, το DNA ή το RNA ενώ οι ακμές δείχνουν τις σχέσεις ή τις αλληλεπιδράσεις μεταξύ των μορίων αυτών [1].

1.2. Κατηγορίες Γράφων

Στην θεωρία γράφων συναντάμε διάφορες κατηγορίες δικτύων. Οι πιο γνωστές κατηγορίες γράφων είναι: ο *κατευθυνόμενος* γράφος, ο *μη κατευθυνόμενος* γράφος, ο *γράφος με βάρη*, ο *διμερής* γράφος, τα *δέντρα*, οι *κλίκες*, το *σύμπλεγμα* (Εικόνα 1).

Κατευθυνόμενος γράφος (directed graph)*:* Σε έναν κατευθυνόμενο γράφο, οι ακμές μεταξύ των κόμβων είναι βέλη τα οποία δείχνουν την κατεύθυνση του γράφου.

Μη Κατευθυνόμενος γράφος (undirected graph)*:* Στη κατηγορία του μη κατευθυνόμενου γράφου, οι ακμές είναι απλές ευθείες χωρίς βάρη.

Γράφος με βάρη (weighted graph)*:* Στην περίπτωση του γράφου με βάρη, η κάθε ακμή έχει ένα συντελεστή βαρύτητας ο οποίος σηματοδοτεί τη σημαντικότητα της σύνδεσης.

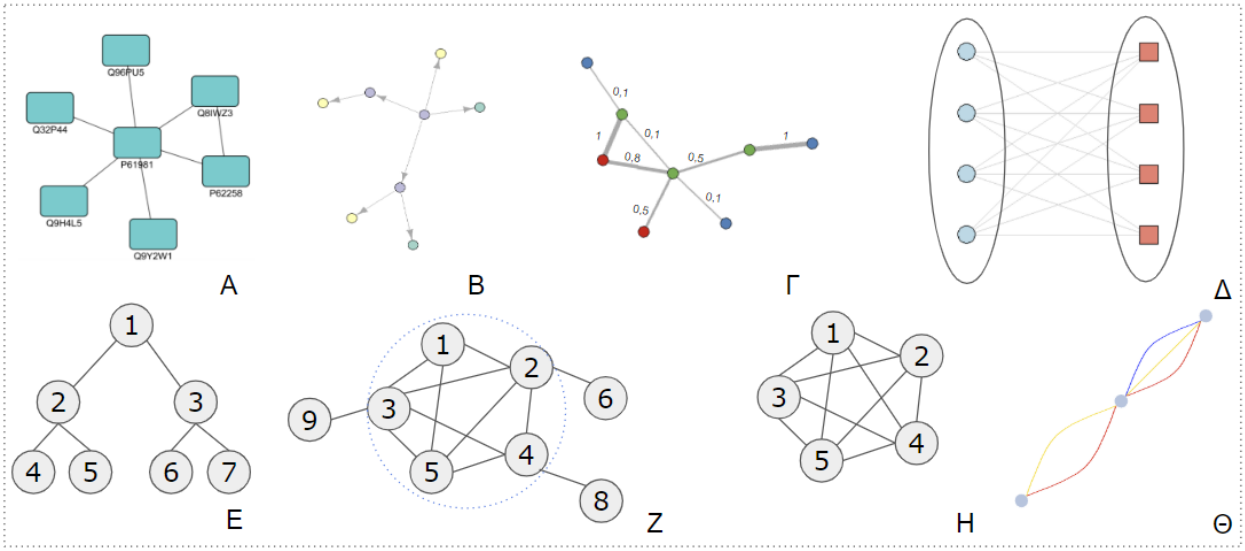
Διμερής γράφος (bipartite graph)*:* Σε έναν διμερή γράφο, το σύνολο των κορυφών μπορεί να χωριστεί σε δύο ομάδες Vκαι V'. Οι κόμβοι της μιας ομάδας μπορούν να επικοινωνούν μόνο με τους κόμβους της άλλης ομάδας ενώ δεν επιτρέπονται συνδέσεις μεταξύ κόμβων της ίδιας ομάδας.

***Δέντρα (trees)**:* Ένα δέντρο είναι ένας μη κατευθυνόμενος γράφος, στον οποίο οποιεσδήποτε δύο κορυφές συνδέονται με ένα και μόνο απλό μονοπάτι. Με άλλα λόγια κάθε συνεκτικός γράφος χωρίς κύκλους είναι ένα δέντρο.

Κλίκα (clique): Είναι ένας μη πλήρης υπο-γράφος όπου κάθε κόμβος συνδέεται με όλους τους υπόλοιπους.

Σύμπλεγμα, Συστάδα ή Ομάδα (cluster): Είναι ένας υπο-γράφος που αποτελείται από μια ομάδα κόμβων με κοινά χαρακτηριστικά μεταξύ τους.

Ως ***Συσταδοποίηση (Clustering)*** ονομάζεται η διαδικασία εκείνη κατά την οποία ένα σύνολο κόμβων χωρίζεται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση αντικειμένων σε ίδια ομάδα μεταφράζεται ως ***ομοιότητα*** των αντικειμένων αυτών και αντίστροφα. Σε περίπτωση που τα αντικείμενα του γράφου δεν ανήκουν στην ίδια ομάδα θεωρούνται ***ανόμοια*** και δεν μοιράζονται κοινά χαρακτηριστικά.



Εικόνα 1. Τύποι γράφων. Α) Ένας απλός γράφος. **Β)** Κατευθυνόμενος γράφος. **Γ)** Γράφος με βάρη. **Δ)** Διμερής γράφος. **Ε)** Γράφος σε μορφή δέντρου. **Ζ)** Σύμπλεγμα υπογράφος. **Η)** Κλίκα. **Θ)** Γράφος με πολλαπλούς τύπους ακμών.

1.3. Βασικά χαρακτηριστικά τοπολογίας δικτύων

1.3.1. Βαθμός (degree)

Ως βαθμό, ονομάζουμε τον συνολικό αριθμό των γειτονικών ακμών ενός κόμβου και συμβολίζεται deg_i. Στη περίπτωση ενός κατευθυνόμενου γράφου ο βαθμός του είναι το άθροισμα των *indegree* deg_i^{in} που είναι ο συνολικός αριθμός των ακμών που προσπίπτουν σε κάθε κόμβο, και των *outdegree* deg_i^{out}που είναι ο συνολικός αριθμός των ακμών που εκτείνεται από κάποιον κόμβο με προορισμό κάποιον άλλον $deg_i = deg_i^{in} + deg_i^{out}$. Ο μέσος βαθμός συνδεσιμότητας ορίζεται ως $deg_{avg} = \frac{\Sigma deg_i}{|V|}$.

Ο υπολογισμός του βαθμού ενός γράφου είναι από τα βασικότερα τοπολογικά χαρακτηριστικά ενός δικτύου καθώς βοηθά στον διαχωρισμό των δικτύων σε διάφορους τύπους. Τα δίκτυα που η κατανομή του βαθμού τους $p(k)$ ακολουθούν έναν νόμο ισχύος, ονομάζονται μη-κλιμακούμενα δίκτυα (scale-free networks).

1.3.2. Πυκνότητα (density)

Ως πυκνότητα ορίζουμε τον λόγο του συνολικού αριθμού των ακμών του γράφου προς τον πιθανό αριθμό των ακμών του γράφου. Σε έναν πλήρως συνδεδεμένο γράφο, ο αριθμός των πιθανών ακμών υπολογίζεται $Emax = \frac{V(V-1)}{2}$ όπου Vείναι οι κόμβοι. Κατά συνέπεια ο υπολογισμός της πυκνότητας γίνεται ως εξής: $\frac{E}{Emax} = \frac{2E}{V(V-1)}$.

Ένας γράφος χαρακτηρίζεται ως πυκνός όταν $E \simeq V^k, 2 > k > 1$, ενώ χαρακτηρίζεται ως αραιός όταν $E \simeq V$ή $E \simeq V^k, k \leq 1$.

1.3.3. Συντελεστής Ομαδοποίησης (clustering coefficient)

Ως συντελεστή ομαδοποίησης ονομάζουμε ένα μέτρο που δείχνει εάν ένας κόμβος ή ένα δίκτυο έχει την τάση να δημιουργεί υπό-ομάδες. Ο συντελεστής ομαδοποίησης ενός κόμβου υπολογίζεται από τον λόγο μεταξύ του αριθμού των ακμών μεταξύ των γειτόνων ενός κόμβου προς τον αριθμό των πιθανών ακμών μεταξύ των συγκεκριμένων γειτόνων. Ο μαθηματικός τύπος υπολογισμού του συντελεστή ενός κόμβου i είναι $Ci = \frac{2e}{k(k-1)}$, όπου k είναι το πλήθος των γειτόνων, δηλαδή ο βαθμός, ενώ e είναι οι ακμές μεταξύ των γειτόνων. Οι τιμές που μπορεί να πάρει ο συντελεστής είναι από 0, περίπτωση με χαμηλή τάση δημιουργίας ομάδων, έως και 1, περίπτωση με υψηλή τάση δημιουργίας ομάδων.

1.3.4. Απόσταση (distance)

Το μήκος του συντομότερου μονοπατιού μεταξύ δύο κόμβων ενός γράφου, ορίζεται ως απόσταση και συμβολίζεται $dist_{ij}$όπου i και j είναι οι δύο κόμβοι. Ως συντομότερο μονοπάτι, θεωρείται ο μικρότερος αριθμός ακμών που μεσολαβούν μεταξύ των δύο κόμβων. Στη περίπτωση που οι δύο κόμβοιi και j δεν συνδέονται μεταξύ τους τότε η απόσταση τους είναι ίση με το άπειρο, $dist_{ij} = \infty$.

1.3.5. Διάμετρος (diameter)

Η διάμετρος σε ένα δίκτυο είναι το μήκος του μακρύτερου μονοπατιού μεταξύ δύο κόμβων και συνδέεται με την απόσταση μέσω της σχέσης $diam_m = max(dist_{ij})$[1].

1.4. Μοντέλα δικτύων

Τα μοντέλα δικτύων είναι χρήσιμα στην διαλεύκανση της τοπολογίας των δικτύων. Υπάρχουν διάφορα μοντέλα με πιο δημοφιλή τα εξής: Erdős-Rényi [2], Watts-Strogatz [3] και Barabási – Albert [4]

1.4.1. Το μοντέλο Erdős–Rényi

Το μοντέλο ***Erdős–Rényi*** (Εικόνα 2Α) είναι από τα πιο γνωστά στην θεωρία των γράφων και έχει ως απώτερο σκοπό την περιγραφή των ιδιοτήτων ενός γράφου. Εάν υποθέσουμε πως Vείναι οι κόμβοι του

γράφου τότε η πιθανότητα σύνδεσης τους με τυχαίο τρόπο είναι: $p = \frac{2E}{V(v-1)}$ με $p \leq 1$και η κατανομή του βαθμού (degree distribution) είναι διωνυμική. Ο υπολογισμός της πιθανότητας ενός κόμβου να έχει κάποιον συγκεκριμένο βαθμό υπολογίζεται με τον εξής τύπο: $p(deg) \simeq e^{-deg_{avg}} \frac{deg_{avg}^{\ deg}}{deg!}$. Εάν η πιθανότητα p είναι μικρή τότε το δίκτυο δεν φαίνεται συνδεδεμένο, ενώ εάν $p \approx \frac{1}{V}$ το δίκτυο έχει μία μεγαλύτερη συνιστώσα η οποία περιέχει τις περισσότερες συνδέσεις του δικτύου. Στη περίπτωση που σε έναν γράφο το πλήθος των κόμβων τείνει στο άπειρο τότε η κατανομή είναι Poisson. Ο συντελεστής ομαδοποίησης (clustering coefficient) του συγκεκριμένου δικτύου δείχνει πως η πιθανότητα δύο κόμβων με κοινό γείτονα να συνδέονται μεταξύ τους είναι ίση με την πιθανότητα σύνδεσης δύο τυχαίων κόμβων. Ο υπολογισμός του συντελεστή γίνεται με τον τύπο: $C = p = \frac{deg_{avg}}{V}$. Το μοντέλο ***Erdős–Rényi*** δεν είναι ιδανικό μοντέλο όσον αφορά την κατανομή βαθμών [1].

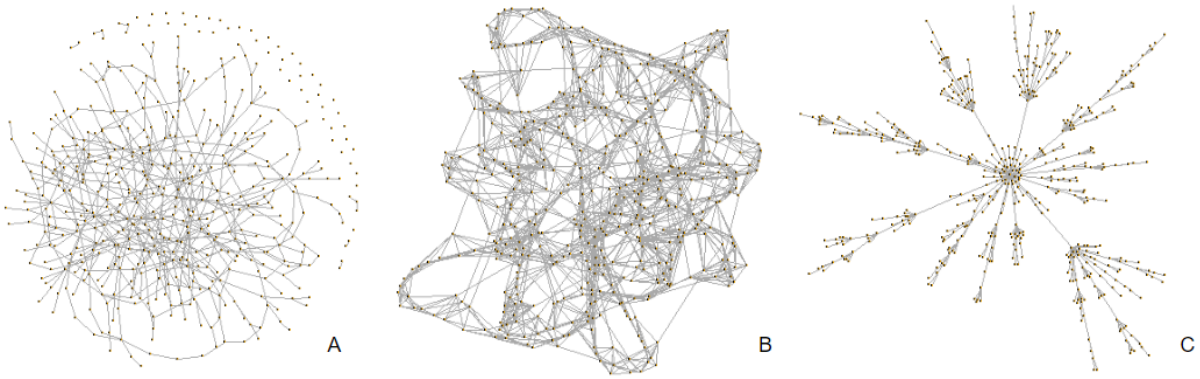
1.4.2. Το μοντέλο Watts-Strogatz

Το μοντέλο ***Watts-Strogatz*** (Εικόνα 2Β) χρησιμοποιείται για την περιγραφή τυχαίων δικτύων που η τοπολογία τους είναι παρόμοια με εκείνη ενός μικρού δικτύου (small world), δηλαδή οι περισσότεροι κόμβοι μπορούν να είναι προσβάσιμοι από οποιονδήποτε άλλο κόμβο του δικτύου μέσω ενός σύντομου μονοπατιού. Το μοντέλο των Watts και Strogatz ανέδειξε μία μοντελοποίηση δικτύων που αποτελείται από τοπικές δομές καθώς και από, κατά μέσο όρο, μικρού μήκους μονοπάτια. Χαρακτηριστικό παράδειγμα αποτελούν τα μεταβολικά δίκτυα όπου οι μεταβολίτες είναι συνδεδεμένοι μεταξύ τους μέσω μικρών μονοπατιών [5]. Σε ένα δίκτυο που βασίζεται στο μοντέλο ***Watts-Strogatz***, εάν όλοι οι κόμβοι του τοποθετηθούν σε έναν κύκλο, κάθε κόμβος θα είναι συνδεδεμένος με $\frac{V}{2}$ γείτονες. Η υψηλή τοπική ομαδοποίηση και ταυτόχρονα το μικρό μέσο μήκος διαδρομής είναι δύο κύρια χαρακτηριστικά αυτού του τύπου δικτύων [1].

1.4.3. Το μοντέλο Barabási–Albert

Το μοντέλο ***Barabási–Albert*** (Εικόνα 2Γ) χρησιμοποιείται για την περιγραφή τυχαίων μη-κλιμακούμενων (scale-free) δικτύων, των οποίων η κατανομή του βαθμού τους ακολουθεί έναν νόμο ισχύος λαμβάνοντας υπόψη την ανομοιογενή κατανομή του βαθμού ή διαφορετικά, τα δίκτυα που οι κόμβοι τους δεν έχουν συγκεκριμένο αριθμό γειτόνων. Σύμφωνα με το συγκεκριμένο μοντέλο το δίκτυο μπορεί να αναπτύσσεται συνεχώς και δεν γίνεται τυχαία η εμφάνιση νέων κόμβων. Κάθε νέος κόμβος ακολουθεί την υπάρχουσα κατανομή του δικτύου. Ένα κατανοητό παράδειγμα είναι τα κοινωνικά δίκτυα όπου, ένα άτομο-κόμβος που έχει διευρυμένο κύκλο γνωριμιών είναι πιο πιθανό να προσελκύει περισσότερα νέα άτομα-κόμβους σε σύγκριση με κάποιον που έχει στενό κύκλο γνωριμιών.

Συγκρίνοντας δίκτυα ίδιου μεγέθους και πυκνότητας, που περιγράφονται από τα μοντέλα Erdős–Rényi και Watts-Strogatz, το δίκτυο που περιγράφεται από το μοντέλο Barabási–Albert έχει μικρότερο μέσο μήκος μονοπατιού. Τα περισσότερα βιολογικά δίκτυα, όπως συμβαίνει και στην πραγματική ζωή, είναι ανθεκτικά και έναντι οποιασδήποτε αφαίρεσης κόμβων, καθώς οι βιολογικές λειτουργίες πρέπει να διατηρούνται [1].



Εικόνα 2. Μοντέλα δικτύων Α) Erdős–Rényi δίκτυο. **Β)** Watts-Strogatz δίκτυο **Γ)** Barabási–Albert scale-free δίκτυο [1]

2. Βιολογικά και βιοϊατρικά δίκτυα

Στην επιστήμη της Βιολογίας κύριο μέλημα των ερευνητών είναι η μελέτη μικροοργανισμών. Κάθε οργανισμός αποτελείται από κύτταρα - βασική δομή ζωής - και με τη χρήση των γράφων μπορεί να αναπαρασταθεί κάθε οργανισμός ως εξής: κάθε κύτταρο συμβολίζεται με έναν κόμβο και οι ακμές που συνδέουν τους κόμβους δείχνουν την μεταξύ τους σχέση. Γενικότερα, ο κόμβος ενός βιολογικού δικτύου μπορεί να αντιπροσωπεύει κύτταρα, γονίδια, πρωτεΐνες, μεταβολίτες, προσδέτες, ασθένειες και φάρμακα, ενώ οι ακμές του δικτύου εκφράζουν μία πιθανή σύνδεση μεταξύ αυτών των στοιχείων. Επιπλέον με την χρήση των γράφων μπορεί να πραγματοποιηθεί η οπτικοποίηση των σχέσεων μεταξύ των εγγραφών βιολογικών βάσεων δεδομένων [1], [6]. Τα πιο κοινά βιολογικά δίκτυα παρουσιάζονται στις παρακάτω παραγράφους.

2.1. Δίκτυα αλληλεπίδρασης πρωτεϊνών - Protein protein interactions (PPIs)

Στα δίκτυα αλληλεπίδρασης πρωτεϊνών, αναπαρίστανται οι αλληλεπιδράσεις μεταξύ των πρωτεϊνών που στόχος τους είναι η πραγματοποίηση κάποιας βιολογικής διεργασίας. Ο συγκεκριμένος τύπος δικτύου μπορεί να είναι είτε φυσικός είτε προβλεπόμενος. Η δομή των PPIs δικτύων έχει μελετηθεί σε διάφορα είδη και έχει προκύψει πως ανεξάρτητα το είδος, πολλά γνωστά πρωτεϊνικά δίκτυα είναι μη-κλιμακούμενα δίκτυα, δηλαδή μερικές κεντρικές πρωτεΐνες - κόμβοι μέσα σε ένα δίκτυο έχουν ένα μεγάλο ποσοστό των αλληλεπιδράσεων, ενώ οι περισσότερες πρωτεΐνες (μη κεντρικές) περιέχουν μόνο ένα μικρό ποσοστό αυτών [7]. Οι κεντρικοί κόμβοι συνήθως αντιπροσωπεύουν εξελικτικά συντηρημένες πρωτεΐνες, ενώ οι κλίκες έχουν υψηλή λειτουργική σημασία[8]. Οι πιο γνωστές βάσεις που φιλοξενούν PPI δίκτυα για διάφορους οργανισμούς είναι: BioGRID [9], BIND[10], DIP [11], IntAct [12].

2.2. Δίκτυα ομοιότητας αλληλουχιών - Sequence similarity networks (SSNs)

Τα δίκτυα ομοιότητας αλληλουχιών αποτελούνται από πρωτεΐνες ή γονίδια που εμφανίζονται ως κόμβοι ενώ οι ακμές παρουσιάζουν την ομοιότητα των αμινοξικών αλληλουχιών ή των νουκλεοτιδικών αλληλουχιών μεταξύ τους. Όταν δύο κόμβοι συνδέονται μεταξύ τους, σημαίνει πως οι αλληλουχίες τους - είτε πρόκειται για πρωτεΐνες είτε για γονίδια - έχουν ποσοστό ομοιότητας μεγαλύτερο ίσο από μία συγκεκριμένη τιμή που καθορίζει ο χρήστης. Με αυτόν τον τρόπο δημιουργείται ένα μικρό δίκτυο με βάρη, που αναπαριστά πιθανές λειτουργικές συσχετίσεις, μεταξύ των βιομορίων. Επιπλέον χαρακτηριστικά του συγκεκριμένου δικτύου είναι ότι πρόκειται για μη κλιμακούμενο δίκτυο (scale-free), συνήθως αραιό και πως συχνά δημιουργεί κεντρικούς κόμβους (hubs).

Τα πιο γνωστά εργαλεία [13] για τον προσδιορισμό της ομοιότητας αλληλουχιών είναι: BLAST [14], LAST [15], και FASTA3 suite [16]. Η ομαδοποίηση δικτύων ομοιότητας αλληλουχιών βοηθά στον εντοπισμό πρωτεϊνικών οικογενειών, όπου οι πρωτεΐνες έχουν παρόμοιες λειτουργίες ή συμμετέχουν σε βιολογικές διεργασίες[17].

2.3. Ρυθμιστικά δίκτυα γονιδίων - Gene regulatory networks (GRNs)

Τα ρυθμιστικά δίκτυα γονιδίων είναι μία συλλογή μοριακών ρυθμιστών οι οποίοι αλληλεπιδρούν μεταξύ τους αλλά και με τα υπόλοιπα συστατικά του κυττάρου. Με αυτόν τον τρόπο γίνεται δυνατός ο έλεγχος των επιπέδων γονιδιακής έκφρασης του mRNA και των πρωτεϊνών. Συνήθως είναι κατευθυνόμενα

δυναμικά δίκτυα όπου οι κόμβοι τους έχουν μικρό αριθμό αλληλεπιδράσεων που μπορούν να απεικονιστούν ως διμερείς γράφοι. Στα δίκτυα αυτά είναι ελάχιστοι οι κεντρικοί κόμβοι που έχουν υψηλό βαθμό συνδεσιμότητας. Ακόμη ένα χαρακτηριστικό του δικτύου είναι πως ακολουθεί έναν νόμο ισχύος για την κατανομή του βαθμού τους και είναι γνωστά και ως μη-κλιμακούμενα δίκτυα (scale-free network)) $p(k) \sim k^{-\gamma}, \gamma \approx 2$ [18]. Οι βάσεις δεδομένων KEGG [19], GTRD [20], TRANSFAC [21] είναι μερικές από τις βάσεις που περιέχουν δεδομένα για τα ρυθμιστικά γονίδια.

2.4. Δίκτυα μεταγωγής σήματος - Signal transduction networks

Τα δίκτυα μεταγωγής καταγράφουν την μοριακή σηματοδότηση, δηλαδή μία σειρά βιοχημικών διαδικασιών που λαμβάνουν χώρα είτε μέσα στο μόριο είτε από το εξωτερικό στο εσωτερικό του περιβάλλον. Όταν τα μονοπάτια σηματοδότησης που δημιουργούνται αλληλεπιδρούν μεταξύ τους σχηματίζουν κυρίως κατευθυνόμενα αραιά δίκτυα, που αποτελούνται από πολλούς κόμβους και ακμές οι οποίες αντιπροσωπεύουν τις συγκεκριμένες αλληλεπιδράσεις [22]. Οι γνωστότερες βάσεις δεδομένων που φιλοξενούν δεδομένα για τη συγκεκριμένη κατηγορία δικτύων είναι: KEGG [19] και Reactome.

2.5. Μεταβολικά δίκτυα - Metabolic networks

Τα μεταβολικά δίκτυα των οποίων οι κόμβοι είναι οι μεταβολίτες και οι ακμές οι αλληλεπιδράσεις μεταξύ των μεταβολιτών είναι συνήθως κατευθυνόμενα δίκτυα που αναπαριστούν τις χημικές αντιδράσεις του μεταβολισμού, τις μεταβολικές οδούς και τις ρυθμιστικές αλληλεπιδράσεις που καθοδηγούν αυτές τις αντιδράσεις. Πρόκειται για μικρά μη-κλιμακούμενα (scale-free) δίκτυα [5] που ακολουθούν ιεραρχίες [23]. Και στα μεταβολικά δίκτυα, οι γνωστότερες βάσεις που περιλαμβάνουν δεδομένα είναι η KEGG [19] και η Reactome.

2.6. Δίκτυα γονιδιακής συν-έκφρασης - Gene co-expression networks (GCN)

Τα δίκτυα γονιδιακής συν-έκφρασης είναι μη κατευθυνόμενοι γράφοι που συνδέουν γονίδια-κόμβους εφόσον υπάρχει γονιδιακή έκφραση μεταξύ τους και μπορούν να αναπαρασταθούν σε έναν πίνακα ομοιότητας γονιδίων (gene – gene similarity matrix). Τα δίκτυα αυτά έχουν την δυνατότητα να αναδείξουν ποια γονίδια είναι ενεργά ταυτόχρονα, ή στις ίδιες βιολογικές διεργασίες. Δεν μπορούν να διακρίνουν τα ρυθμιστικά από τα ρυθμιζόμενα γονίδια, αλλά μπορούν να προσδιορίσουν ποια γονίδια έχουν την τάση να δείχνουν ένα συντονισμένο μοτίβο έκφρασης σε μια ομάδα δειγμάτων.
Η δημιουργία και η ανάλυση των δικτύων γονιδιακής έκφρασης περιγράφεται ως εξής:

- Καθορισμός ενός μέτρου συν-έκφρασης και υπολογισμός ομοιότητας (similarity score) για κάθε ζεύγος. Γενικότερα, χρησιμοποιούνται διαφορετικά μέτρα συσχέτισης για την κατασκευή δικτύων, συμπεριλαμβανομένων των συσχετίσεων Pearson ή Spearman.
- Προσδιορισμός κατωφλιού (threshold) με σκοπό την σύγκριση του βαθμού ομοιότητας με την τιμή του κατωφλιού. Όσα ζευγάρια έχουν βαθμό ομοιότητας μεγαλύτερο από το κατώφλι, θεωρείται ότι έχουν σημαντική σχέση συν-έκφρασης και συνδέονται στο δίκτυο.
- Δημιουργία δικτύου συν-έκφρασης γονιδίων, όπου το κάθε γονίδιο παριστάνεται με έναν κόμβο και η σχέση μεταξύ δύο γονιδίων-κόμβων παριστάνεται με μία ακμή.
- Εντοπισμός ομάδων γονιδίων συν-έκφρασης με τη βοήθεια εργαλείων ομαδοποίησης (clustering). Ανάλογα με τον συνολικό συντελεστή ομαδοποίησης, το δίκτυο μπορεί να συγκεντρωθεί για να ανιχνεύσει λειτουργικές ενότητες.

Η μέθοδος ομαδοποίησης πρέπει να επιλεγεί με προσοχή, διότι μπορεί να επηρεάσει σημαντικά το αποτέλεσμα και το νόημα της ανάλυσης [24].

Βάσεις δεδομένων που σχετίζονται με τα δίκτυα γονιδιακής συν-έκφρασης είναι: GEO [25], ArrayExpress [26], COXPRESdb [27].

2.7. Φυλογενετικά δίκτυα - Phylogenetic networks

Ένα φυλογενετικό δίκτυο είναι ένας γράφος που χρησιμοποιείται για την αναπαράσταση των εξελικτικών σχέσεων μεταξύ αλληλουχιών νουκλεοτιδίων, γονιδίων, χρωμοσωμάτων, γονιδιωμάτων ή ειδών[28]. Η δομή αναπαράστασης τους δεν είναι ακόμα ξεκάθαρη καθώς είναι αμφισβητήσιμο εάν η παρουσίαση τους ως δέντρο είναι σωστή ή όχι. Ο ισχυρισμός που θέλει τα φυλογενετικά δίκτυα να διαφέρουν από τα φυλογενετικά δέντρα, βασίζεται στο ότι η μοντελοποίηση τους αποτελείται από πλούσια συνδεδεμένα δίκτυα, με την προσθήκη υβριδικών κόμβων (κόμβοι με δύο γονείς) σε αντίθεση με τους κόμβους δέντρων που κάθε κόμβος έχει έναν μόνο γονέα (μια ιεραρχία κόμβων) [29]. Ακόμη μία διαφοροποίηση τους είναι πως τα φυλογενετικά δέντρα είναι κατάλληλα μόνο για τη μελέτη κάθετων εξελικτικών διαδικασιών ενώ τα φυλογενετικά δίκτυα, είναι πιο γενικά και μπορούν να χρησιμοποιηθούν για τη μελέτη τόσο οριζόντιων όσο και κάθετων εξελικτικών διαδικασιών. Οι οριζόντιες διαδικασίες αντιπροσωπεύονται από δικτυώσεις στο δίκτυο, οι οποίες δεν εμφανίζονται στο δέντρο[30]. Τα φυλογενετικά δέντρα μπορούν να θεωρηθούν ένα υποσύνολο των φυλογενετικών δικτύων.

Με βάση τη θεωρία του Δαρβίνου η οποία υποστηρίζει πως όλα τα είδη που ζουν σήμερα προέρχονται από έναν κοινό πρόγονο, οι σχέσεις μεταξύ κάθε ομάδας ατόμων, ακόμη και εκείνων από διαφορετικά είδη, μπορούν να εμφανίζονται σε ένα φυλογενετικό δέντρο. Έτσι, ο στόχος της φυλογενετικής είναι η χρήση βιολογικών δεδομένων για μια συλλογή ατόμων ή ειδών, και να δημιουργηθεί ένα δέντρο που περιγράφει πώς σχετίζονται [31]. Οι πιο γνωστές μέθοδοι για την κατασκευή δέντρων είναι οι Neighbor-Joining (NJ), UPGMA και Maximum Parsimony (MP) [32].

Μερικά λογισμικά για την οπτικοποίηση των φυλογενετικών δικτύων είναι: SplitsTree [33], DendroScope[34], και το πακέτο της R, phangorn [35].

2.8. Οικολογικά δίκτυα - Ecological networks

Τα οικολογικά δίκτυα απεικονίζουν τις βιολογικές αλληλεπιδράσεις των διάφορων ειδών που ζουν μέσα σε ένα οικοσύστημα. Τα είδη του οικοσυστήματος αναπαριστώνται με κόμβους που συνδέονται σε ζεύγη. Υπάρχουν τέσσερις κατηγορίες αλληλεπιδράσεων: τροφικές, συμβιωτικές, αμοιβαίες (αμφίδρομες) και ανταγωνιστικές (παράσιτο ξενιστή)[1]. Τα δίκτυα αυτά περιγράφουν την λειτουργία του οικοσυστήματος και η μοντελοποίηση τους βοηθάει στην μελέτη πιθανών επιπτώσεων σε περίπτωση μεταβολής κάποιου στοιχείου του συστήματος.

Οι τροφικές αλυσίδες μπορούν να είναι κατευθυνόμενοι ή και μη κατευθυνόμενοι γράφοι που ακολουθούν εκθετική κατανομή και εμφανίζουν μέση χαμηλή σύνδεση, ενώ οι ποσοτικές τροφικές αλυσίδες μπορούν να αναπαρασταθούν ως γράφοι με βάρη [36].

Η παρουσία ομαδοποιημένων ειδών σε οικολογικά δίκτυα, είναι συζητήσιμη αλλά ενισχύεται από την ανάλυση μικρών, όχι καλά επιλυμένων, συγκεντρωτικών δικτύων. Οι αλυσίδες υψηλής ανάλυσης, που τα είδη δεν συγκεντρώνονται σε “τροφικά” είδη παρουσιάζουν υψηλότερο βαθμό ομαδοποίησης από τους τυχαίους ομολόγους τους [37]. Γενικότερα, ένα είδος στη μέση ενός

συμπλέγματος μπορεί να έχει τον ρόλο ενός ακρογωνιαίου λίθου και η απώλειά του θα μπορούσε να έχει μεγάλες επιπτώσεις στο δίκτυο.

2.9. Επιδημιολογικά δίκτυα - Epidemiological networks

Οι επιδημίες και η ανάλυση τους, δεν αφορούν μόνο την επιστήμη της Βιολογίας, αλλά και της Κοινωνιολογίας, καθώς οι επιδημικές ασθένειες είναι μεταδοτικές ασθένειες που επηρεάζουν άμεσα την κοινωνική εξέλιξη, με χαρακτηριστικό παράδειγμα την παρούσα έξαρση του κορονοϊού. Στα παραπάνω δύο επιστημονικά πεδία, υπάρχουν αρκετές αναλογίες στον τρόπο που διαδίδεται μια επιδημία, γεγονός που οδήγησε στην ανάπτυξη μοντέλων εξάπλωσης μόλυνσης από τη βιολογία και εφαρμογή τους στα δίκτυα υπολογιστών.

Η μελέτη ενός επιδημιολογικού δικτύου βοηθάει στον εντοπισμό των οδών μετάδοσης κάποιας ασθένειας. Επίσης μέσω ενός τέτοιου δικτύου μπορούμε να έχουμε πρόσβαση στις πληροφορίες για την επιδημιολογική δυναμική. Συνδέοντας τη δυναμική του δικτύου με δεδομένα της πραγματικής ζωής, τα δεδομένα των ασθενών μπορούν να είναι μία ωφέλιμη βάση για την ανάπτυξη υποθέσεων σχετικά με τον τρόπο που δρα κάποια ασθένεια και να αποδειχθούν χρήσιμες στην δημιουργία φαρμάκων και στην ανάπτυξη θεραπειών. Ένα δίκτυο συν-νοσηρότητας, δηλαδή η ταυτόχρονη εμφάνιση νόσων ή παθολογικών καταστάσεων στον ίδιο ασθενή, αποτελεί παράδειγμα ενός επιδημιολογικού δικτύου.

Παρόλο που πολλές φορές τα επιδημιολογικά δίκτυα “προσποιούνται” τα κοινωνικά, υπάρχουν περιπτώσεις που αναπαρίστανται ως διμερείς γράφοι [38].

Βάσεις δεδομένων οι οποίες συνδέονται είτε έμμεσα είτε άμεσα με τα επιδημιολογικά δίκτυα είναι οι εξής: KEGG (KEGG pathways, KEGG diseases) [39], HPRD [40].

2.10. Βιβλιογραφικά δίκτυα συν-αναφορών - Literature co-occurrence networks

Η συγκεκριμένη κατηγορία δικτύων παρουσιάζει την σύνδεση βιο-οντοτήτων που έχουν βρεθεί σε διάφορα κείμενα. Τα ονόματα-αναγνώρισης-οντότητας (Name Entity Recognition - NER), είναι χρήσιμα για την ταυτοποίηση βιομορίων, χημικών ενώσεων, ιστών, ασθενειών κ.α μέσα σε ένα κείμενο και να αντιστοιχηθούν με τις αντίστοιχες οντολογικές / ταξινομικές εγγραφές σε δημόσιες βάσεις δεδομένων.

Πολλά παραδείγματα κειμένων προς ανάλυση προέρχονται από τις διαδικτυακές εγκυκλοπαίδειες, Wikipedia και PubMed [1].

Το πρόγραμμα GenCLiP έχει δημιουργηθεί με σκοπό την ομαδοποίηση μιας λίστας γονιδίων στηριζόμενο στη βιβλιογραφία και δημιουργεί δίκτυα συνύπαρξης γονιδίων που σχετίζονται με συγκεκριμένες λέξεις-κλειδιά της βιβλιογραφίας [41].

2.11. Γνωσιακά δίκτυα - Knowledge networks

Τα γνωσιακά δίκτυα έχουν ως χαρακτηριστικό τα πολλαπλά άκρα, πράγμα που οφείλεται στον συνδυασμό ετερογενών πληροφοριών και μετα-δεδομένων (metadata) από διαφορετικές πηγές. Οι πηγές αυτές μπορούν να είναι δημόσιες βάσεις δεδομένων, βιολογικές βάσεις δεδομένων αλλά και η βιβλιογραφία [1]. Οι γνωστότερες βάσεις είναι οι εξής: STRING [42] (γνωστές και προβλεπόμενες πρωτεϊνικές αλληλεπιδράσεις από διάφορους οργανισμούς), STITCH [43] (γνωστές και προβλεπόμενες αλληλεπιδράσεις μεταξύ πρωτεϊνών και χημικών ενώσεων), PICKLE [44] (βάση μετα-δεδομένων για το δίκτυο με την άμεση αλληλεπίδραση των πρωτεϊνών του ανθρώπου).

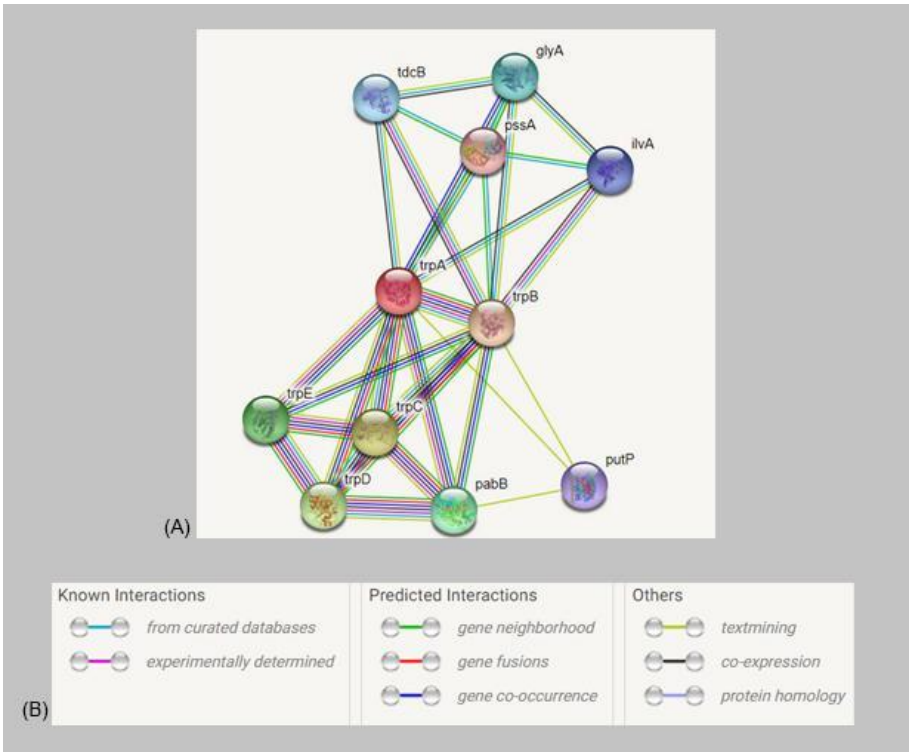
3. Βάσεις δεδομένων με βιολογικά δίκτυα

3.1. Δίκτυα πρωτεϊνικών αλληλεπιδράσεων (PPI - Protein protein Interaction Networks)

3.1.1. Βάση Δεδομένων STRING

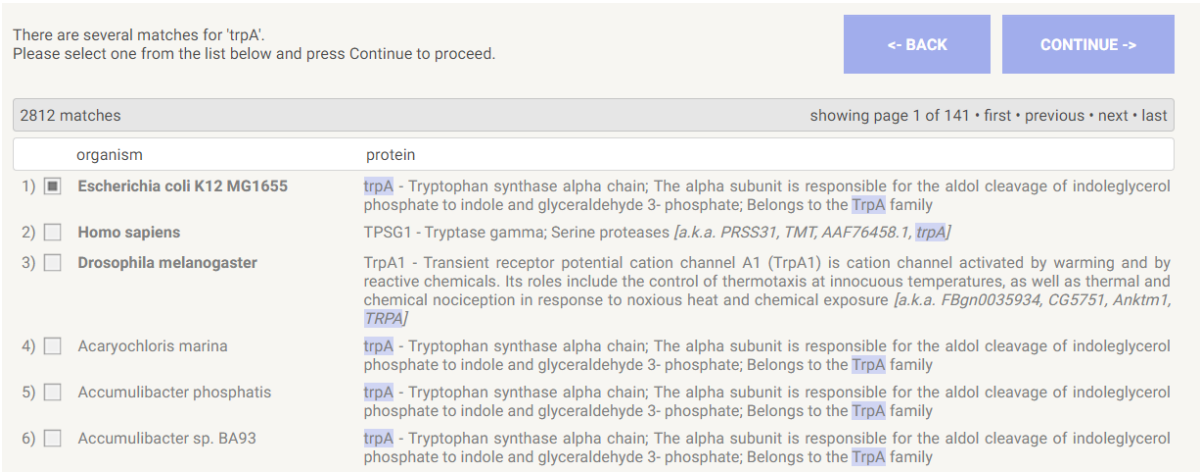
Η βάση δεδομένων ***STRING*** (The Search Tool for the Retrieval of Interacting Genes) είναι μια διαδικτυακή πλατφόρμα που δίνει πρόσβαση σε πληροφορίες σχετικά με τις άμεσες (φυσικές) ή έμμεσες (λειτουργικές) αλληλεπιδράσεις πρωτεϊνών από διάφορους οργανισμούς. Η παρούσα εκδοχή της STRING(v.11.0b) φιλοξενεί 24,584,628 πρωτεΐνες που προέρχονται από 5,090 οργανισμούς και 3,123,056,667 αλληλεπιδράσεις.

Το περιβάλλον της STRING είναι φιλικό προς τον χρήστη, ο οποίος μπορεί να ξεκινήσει μία αναζήτηση συμπληρώνοντας το όνομα μιας πρωτεΐνης, ή την αμινοξική ακολουθία της, στο αντίστοιχο πεδίο αλλά και το όνομα του οργανισμού προέλευσής της. Η πλατφόρμα έχει ήδη έτοιμα τρία παραδείγματα πρωτεϊνών τα οποία μπορεί να πατήσει ο χρήστης και να συμπληρωθούν αυτόματα τα παραπάνω δύο πεδία. Πέρα από τους δύο παραπάνω τρόπους αναζήτησης, η βάση επιτρέπει την αναζήτηση με βάση μία λίστα πρωτεϊνών ή γονιδίων ή λίστα αμινοξικών ακολουθιών ή ακόμα και ένα ολόκληρο πείραμα που δίνεται ως λίστα πρωτεϊνών [45]. Η STRING δημιουργεί ένα μη κατευθυνόμενο δίκτυο με πολλές ακμές που η κάθε μία αντιπροσωπεύει έναν διαφορετικό τύπο αλληλεπίδρασης (Εικόνα 7Β) μεταξύ των πρωτεϊνών που παρουσιάζονται ως κόμβοι (Εικόνα 3Α). Ο αριθμός των κόμβων που αλληλεπιδρούν με την ζητούμενη πρωτεΐνη είναι προεπιλεγμένος και ίσος με το δέκα. Οι πηγές προσέλευσης των σχέσεων μεταξύ των πρωτεϊνών-κόμβων μπορούν να είναι: πειραματικά δεδομένα, γνωστά μονοπάτια, πρωτεϊνικά συμπλέγματα, συστηματικές αναλύσεις συν-έκφρασης, μελέτες συσχέτισης σε ολόκληρο το γονιδίωμα, εξόρυξη δεδομένων, γονιδιακή ορθολογία.



Εικόνα 3. Παράδειγμα από τη βάση δεδομένων String. Α) Αποτέλεσμα δικτύου της πρωτεΐνης trpA (Tryptophan synthase alpha chain). Ο κόκκινος κόμβος είναι η πρωτεΐνη που αναζήτησε ο χρήστης. Οι αλληλεπιδράσεις με τις υπόλοιπες πρωτεΐνες έχουν διαφορετικό χρώμα ανάλογα με τη σημασία τους. **Β)** Επεξήγηση των αλληλεπιδράσεων με βάση το χρώμα της κάθε μιας.

Στην περίπτωση που ο χρήστης αναζητήσει μία πρωτεΐνη χωρίς να επιλέξει οργανισμό προέλευσης, τότε η STRING δημιουργεί μία σελίδα με έναν πίνακα ο οποίος περιλαμβάνει όλες τις αντιστοιχίσεις της υπό εξέταση πρωτεΐνης με διάφορους οργανισμούς. (Εικόνα 4)



Εικόνα 4. Αποτέλεσμα δικτύου της πρωτεΐνης trpA (Tryptophan synthase alpha chain) χωρίς την επιλογή οργανισμού.

Στο δίκτυο που δημιουργείται ο χρήστης έχει την δυνατότητα να αλλάξει την μορφή εμφάνισης του, επιλέγοντας και μεταφέροντας τον κάθε κόμβο στη νέα επιθυμητή θέση. Επίσης, επιλέγοντας έναν τυχαίο κόμβο του δικτύου, εμφανίζεται στην οθόνη ένα pop up παράθυρο με πληροφορίες της πρωτεΐνης-κόμβου. Μέσω του pop up παραθύρου, ο χρήστης μπορεί να οδηγηθεί στον σχολιασμό της πρωτεΐνης, στη τρισδιάστατη δομή της από τη βάση PDB [46] αλλά και σε συνδέσμους εξωτερικών πηγών όπως για παράδειγμα η UniProt [47], η Ensembl [48] και η PubMed.

Με τον ίδιο τρόπο όπως και παραπάνω ο χρήστης μπορεί να πάρει πληροφορίες για μία αλληλεπίδραση-ακμή του δικτύου. Μία αλληλεπίδραση έχει τη δυνατότητα να είναι εμφανής με πολλαπλούς τρόπους, για παράδειγμα δύο πρωτεΐνες μπορούν να συν-εκφράζονται στο ίδιο πείραμα, μπορεί να σχετίζονται εξελικτικά, μπορεί να συνυπάρχουν στη βιβλιογραφία, μπορεί να είναι προϊόντα σύντηξης, ή τα γονίδια τους μπορεί να συν-εντοπιστούν μεταξύ των γονιδιωμάτων [45].

Στη βάση δεδομένων STRING υπάρχουν οκτώ διαφορετικοί τρόποι για την απεικόνιση ενός δικτύου ανάλογα με τον κάθε τύπο σύνδεσης. Οι οχτώ αυτοί τρόποι παρουσιάζονται παρακάτω:

Δίκτυο (Network) - Σε αυτήν τη μορφή, όλες οι πρωτεΐνες και οι αλληλεπιδράσεις τους εμφανίζονται ως δίκτυο. Το δίκτυο είναι πλήρως διαδραστικό και μπορεί να εμφανιστεί είτε ως σταθμισμένο, όπου το πάχος μιας ακμής υποδεικνύει την ισχύ μιας σύνδεσης, είτε ως πολλαπλή ακμή, όπου τα χρώματα γραμμής υποδεικνύουν τους διαφορετικούς τύπους αλληλεπίδρασης.

Πειραματικά (Experiments) - Στη συγκεκριμένη περίπτωση, η βάση δείχνει τις αλληλεπιδράσεις μεταξύ πρωτεϊνών που αναφέρονται σε άλλες γνωστές πρωτογενείς βάσεις δεδομένων αλληλεπίδρασης, όπως η BIND, η DIP, η HPRD και η IntAct. Το αποτέλεσμα που εμφανίζεται είναι μια λίστα με πληροφορίες κειμένου που προέρχονται από αυτές τις βάσεις δεδομένων, μαζί με τους αντίστοιχους συνδέσμους.

Βάσεις Δεδομένων (Databases) - Στην μορφή αυτή της STRING εμφανίζονται πληροφορίες για αλληλεπιδρώντες πρωτεΐνες που βρίσκονται σε εξωτερικές βάσεις δεδομένων, όπως η Biocarta, η BioCyc, η GO και η KEGG.

Εξόρυξη δεδομένων (Text-mining) - Η STRING αναφέρεται σε πρωτεΐνες οι οποίες βρίσκονται στην βιβλιογραφία. Γίνεται χρήση λεξικών για την ακριβή Αναγνώριση Οντοτήτων Ονόματος (Name Entity Recognition - NER) και επίσης οι περιλήψεις της PubMed αναλύονται τακτικά για την εύρεση συνυπάρχοντων πρωτεϊνών με βάση κάποια πρόταση ή την ίδια την περίληψη. Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) χρησιμεύει για τη σύλληψη συνδετικών εννοιών ή συνθηκών μεταξύ των πρωτεϊνών. Ένα χαρακτηριστικό παράδειγμα είναι τα ρήματα ή φράσεις που δηλώνουν μια σημαντική σχέση μεταξύ των πρωτεϊνών. Το αποτέλεσμα της Εξόρυξης Δεδομένων είναι μία λίστα από περιλήψεις όπου οι πρωτεΐνες είναι επισημασμένες.

Συν-έκφραση (Co-expression) - Στη συγκεκριμένη επιλογή της STRING, οι πρωτεΐνες εμφανίζονται ως κόμβοι που αλληλεπιδρούν. Τα mRNA των συγκεκριμένων πρωτεϊνών έχουν παρόμοια κανονικοποιημένα πρότυπα έκφρασης. Ως έξοδος εμφανίζεται ένα heatmap το οποίο δείχνει το άνω τριγωνικό τμήμα ενός συμμετρικού πίνακα συσχέτισης και όσο πιο έντονο είναι το χρώμα μεταξύ ενός ζεύγους πρωτεϊνών, τόσο μεγαλύτερη είναι η βαθμολογία συσχέτισης.

Σύντηξη (Fusion) - Ένα γονίδιο σύντηξης είναι ένα υβριδικό γονίδιο που σχηματίζεται από τη σύντηξη δύο προηγουμένως ξεχωριστών γονιδίων που δημιουργεί ένα ενιαίο ανοιχτό πλαίσιο ανάγνωσης (Open Reading Frame - ORF). Σε αυτή τη περίπτωση δύο πρωτεΐνες συνδέονται εάν είναι προϊόν σύντηξης και έχουν βαθμό συσχέτισης με βάση τη σύντηξη των ορθολογιών τους. Εμφανίζονται μεμονωμένα γεγονότα σύντηξης γονιδίων ανά είδος, ενώ τα είδη στα οποία συμβαίνει ένα συμβάν σύντηξης συγκεντρώνονται σε ένα δενδρόγραμμα.

Γειτονικά (Neighborhood) - Παρουσιάζει γονίδια που εμφανίζονται συνεχώς σε γειτονιές ενός γονιδιώματος. Οι πρωτεΐνες που προέρχονται από γονίδια των οποίων η δια-γονιδιακή απόσταση σε ένα γονιδίωμα είναι μικρότερη από 300 ζεύγη βάσεων είναι συνδεδεμένα. Τα γονιδιώματα εμφανίζονται ως δέντρα ενώ τα γονίδια ως χρωματιστά ορθογώνια.

Συν-ύπαρξη (Co-occurrence) - Η STRING δείχνει την παρουσία ή την απουσία συνδεδεμένων πρωτεϊνών σε κάθε είδος. Οι πρωτεΐνες βρίσκονται σε ένα φυλογενετικό δέντρο ακολουθούμενο από ένα πλέγμα, που παρουσιάζει την παρουσία ή την απουσία της πρωτεΐνης σε κάποιο είδος. Η ένταση του χρώματος αντικατοπτρίζει την ποσότητα διατήρησης της ομόλογης πρωτεΐνης στο είδος.

Ο χρήστης στην βάση δεδομένων STRING έχει την δυνατότητα να επεκτείνει ή να συρρικνώσει τη περιοχή του δικτύου και να προσαρμόσει το μέγεθος του με βάση:

Βαθμό αξιοπιστίας των ακμών (Lines' Confidence Score) - Η STRING σχολιάζει τις αλληλεπιδράσεις μεταξύ των πρωτεϊνών με ένα ή περισσότερα βάρη και ο χρήστης μπορεί να οριοθετήσει από το 0 έως το 1 οποιοδήποτε βάρος των ακμών. Τα βάρη αυτά δεν εκπροσωπούν πάντα το πόσο δυνατός ή ειδικός είναι ο δεσμός, αλλά κυρίως πόσο αληθινός είναι. Η βαθμολογία του δεσμού είναι συνδυασμός των πιθανοτήτων από διαφορετικά αποδεικτικά κανάλια και την πιθανότητα της τυχαίας παρατήρησης μιας αλληλεπίδρασης.

Βαθμό Συνδεσιμότητας Κόμβων (Node's Degree of Connectivity) - Παρά το γεγονός ότι το δίκτυο που δημιουργείται περιορίζει εξ ορισμού την υπό εξέταση πρωτεΐνη να έχει γύρω της έως και δέκα πρωτεΐνες με τις οποίες αλληλεπιδρά, ο χρήστης μπορεί είτε να αυξήσει είτε να μειώσει τον αριθμό αυτό.

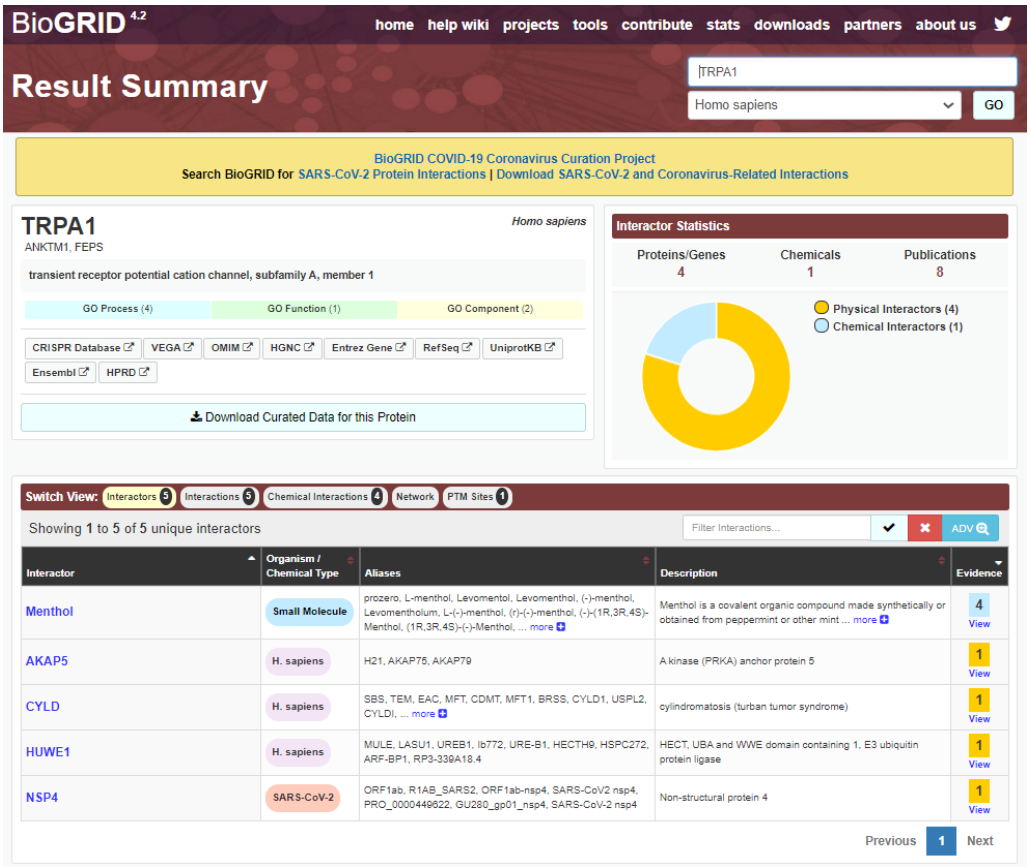
Βάθος Δικτύου (Network's Depth) - Ο χρήστης μπορεί να αυξήσει τη διάμετρο του δικτύου επιτρέποντας ένα δεύτερο επίπεδο πληροφοριών που αναφέρεται αλληλεπιδρώντα μόρια.

Μέσω της βάσης δεδομένων STRING μπορεί κανείς να εφαρμόσει κάποιον από τους διαθέσιμους αλγορίθμους ομαδοποίησης (clustering algorithms) μεταξύ των οποίων είναι ο MCL και ο k-means. Επιπλέον, μέσω της STRING μπορεί να υλοποιηθεί ο *αυτοματοποιημένος λειτουργικός εμπλουτισμός* (automated functional enrichment) και *ο λειτουργικός σχολιασμός δικτύου* (network functional annotation), καθώς ενσωματώνει μια μεγάλη ποικιλία βάσεων δεδομένων ανάμεσα στις οποίες είναι οι: Gene Ontology, KEGG, PubMed, UniProt, INTERPRO και SMART [45].

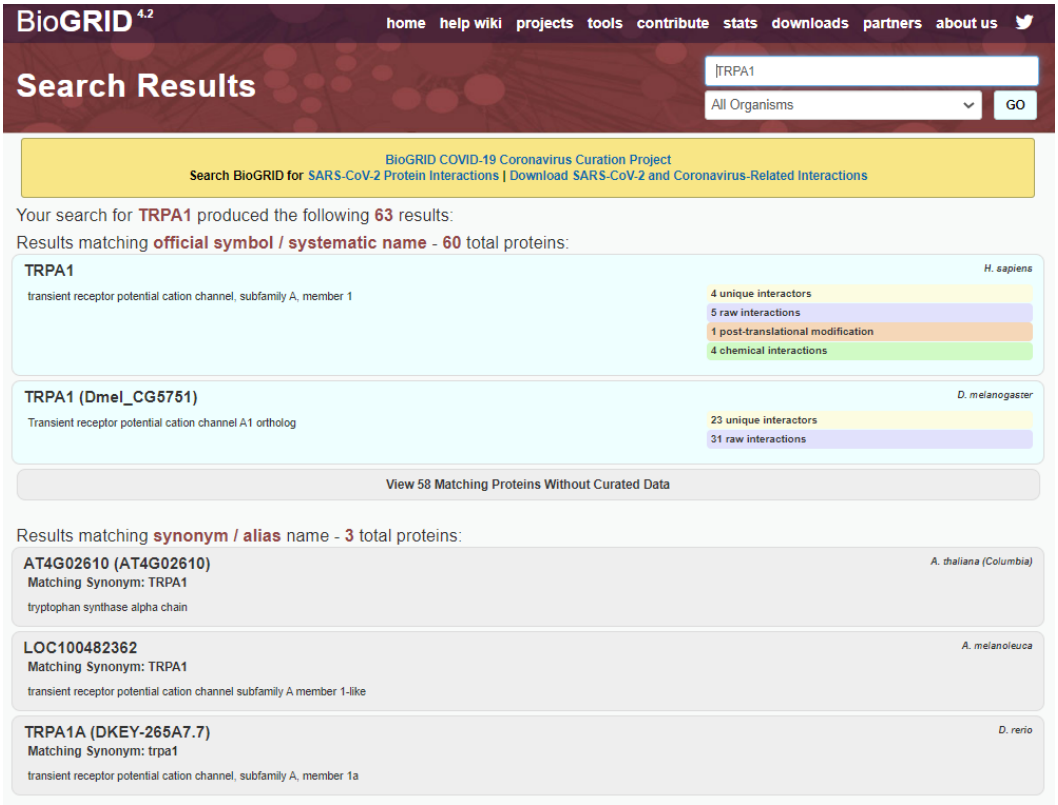
3.1.2. Βάση Δεδομένων BioGrid

Η βάση δεδομένων ***BioGRID*** (Biological General Repository for Interaction Datasets) είναι μία δημόσια βάση που αρχειοθετεί και παρέχει δεδομένα πρωτεϊνικής και γενετικής αλληλεπίδρασης που προέρχονται από οργανισμούς-μοντέλα και από τον άνθρωπο. Πλέον καλύπτει αλληλεπιδράσεις που αφορούν 71 είδη οργανισμών, ανάμεσα τους κύριοι οργανισμοί-μοντέλα αλλά και ο άνθρωπος. Η παρούσα έκδοση της BioGRID 4.2 περιέχει 75,760 δημοσιεύσεις για 1,992,321 πρωτεϊνικές και γενετικές αλληλεπιδράσεις , 29,093 χημικές αλληλεπιδράσεις και 959,750 μετα-μεταφραστικές τροποποιήσεις από βασικούς οργανισμούς-μοντέλα. Η BioGRID συνεργάζεται με άλλες βάσεις δεδομένων όπως η Entrez-Gene, SGD και FlyBase.

Ο χρήστης της πλατφόρμας μπορεί να αναζητήσει ένα σύνολο πληροφοριών για μία πρωτεΐνη που επιθυμεί, επιλέγοντας επίσης και τον οργανισμό από το αντίστοιχο πεδίο. Επίσης, ο χρήστης μπορεί να μην επιλέξει έναν συγκεκριμένο οργανισμό, αλλά να πραγματοποιήσει την αναζήτηση του σε όλους τους διαθέσιμους οργανισμούς ταυτόχρονα (Εικόνα 5, 6). Η βάση δίνει την δυνατότητα στον χρήστη να έχει πρόσβαση σε δεδομένα που προήλθαν από μελέτες υψηλής απόδοσης (High throughput - HTP) αλλά και χαμηλής απόδοσης (Low throughput - LTP). Οι δύο παραπάνω μέθοδοι, LTP και HTP βοηθούν πολλές φορές στην αύξηση των δεδομένων που φιλοξενεί η βάση [49].

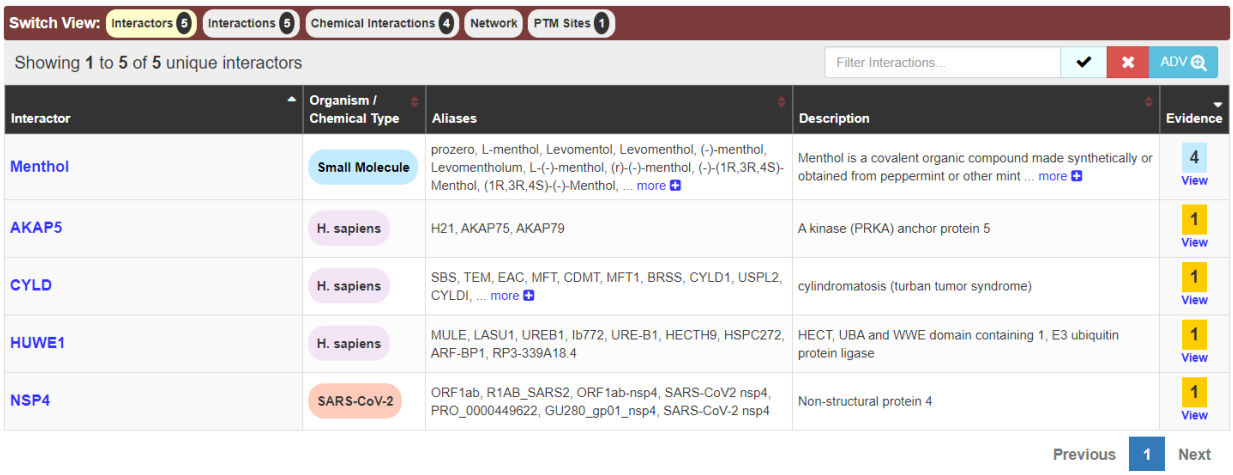


Εικόνα 5. Αποτέλεσμα αναζήτησης της πρωτεΐνης trpA1 με την επιλογή του οργανισμού Homo sapiens. Στο μπλε tab GO Process φαίνονται οι λειτουργίες στις οποίες συμμετέχει η εν λόγω πρωτείνη. Στο πράσινο tab GO Functions εμφανίζονται τα μόρια που συμμετέχουν σε μοριακές λειτουργίες και αντιστοιχούν σε δραστηριότητες που μπορούν να πραγματοποιηθούν από μεμονωμένα γονιδιακά προϊόντα. Στο κίτρινο tab GO Component εμφανίζονται οι θέσεις σε σχέση με τις κυτταρικές δομές στις οποίες ένα γονιδιακό προϊόν εκτελεί μια λειτουργία.

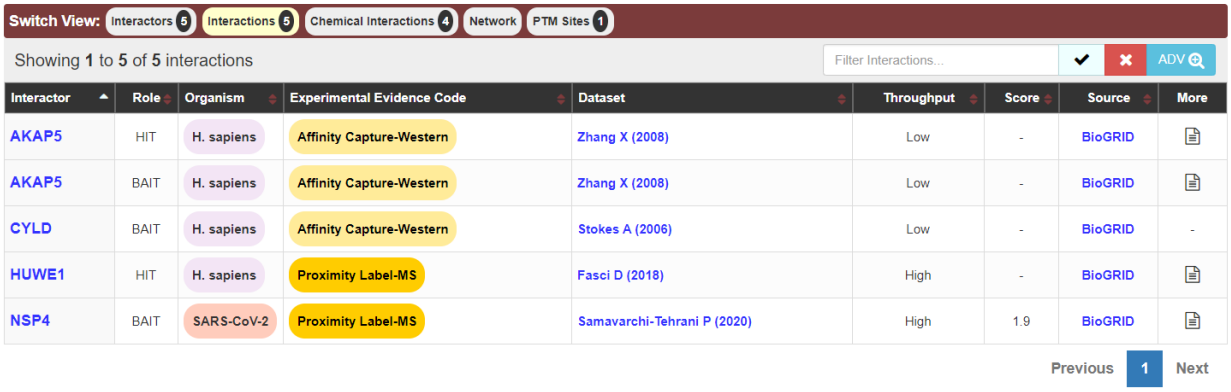


Εικόνα 6. Αποτέλεσμα αναζήτησης της πρωτεΐνης trpA1 χωρίς την επιλογή συγκεκριμένου οργανισμού.

Η διεπαφή δίνει την δυνατότητα στον χρήστη να κατεβάσει δεδομένα που σχετίζονται με την υπό εξέταση πρωτεΐνη, καθώς και συνδέσμους που τον μεταφέρουν σε άλλες βάσεις, όπως η Entrez Gene, η Ensembl ή η OMIM, οι οποίες επίσης φιλοξενούν την συγκεκριμένη πρωτεΐνη. Τέλος, ο χρήστης μπορεί να δει σε πίνακα μέσα στη πλατφόρμα τα αλληλεπιδρώντα μόρια (Interactors), τις αλληλεπιδράσεις (Interactions), τις χημικές αλληλεπιδράσεις (Chemical Interactions) στις οποίες συμμετέχει η πρωτεΐνη και το δίκτυο που δημιουργείται μεταξύ τους και με την πρωτεΐνη που έχει βάλει στην αναζήτηση ο χρήστης. (Εικόνα 7,8,9,10)



Εικόνα 7.: Τα αλληλεπιδρώντα μόρια (Interactors) της πρωτεΐνης trpA1.



Switch View: Interactors 5 | Interactions 5 | Chemical Interactions 4 | Network | PTM Sites 1

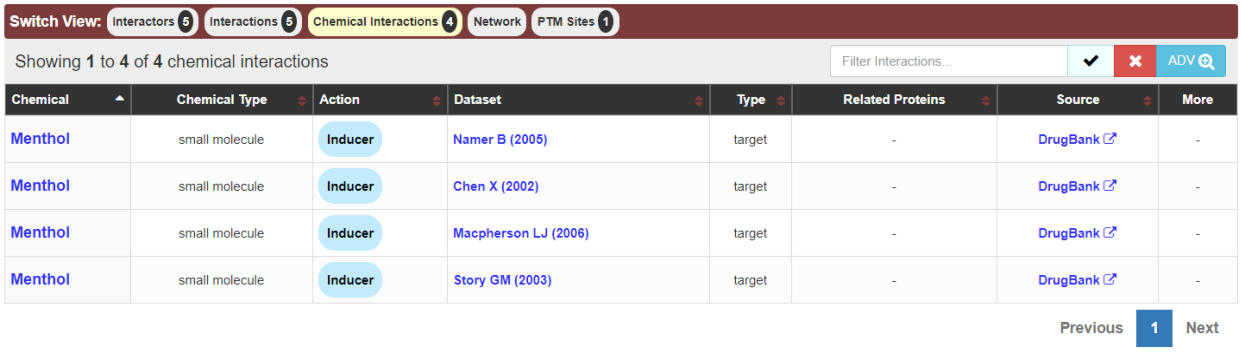
Showing 1 to 5 of 5 interactions

Filter Interactions... ADV

Interactor	Role	Organism	Experimental Evidence Code	Dataset	Throughput	Score	Source	More
AKAP5	HIT	H. sapiens	Affinity Capture-Western	Zhang X (2008)	Low	-	BioGRID	
AKAP5	BAIT	H. sapiens	Affinity Capture-Western	Zhang X (2008)	Low	-	BioGRID	
CYLD	BAIT	H. sapiens	Affinity Capture-Western	Stokes A (2006)	Low	-	BioGRID	-
HUWE1	HIT	H. sapiens	Proximity Label-MS	Fasci D (2018)	High	-	BioGRID	
NSP4	BAIT	SARS-CoV-2	Proximity Label-MS	Samavarchi-Tehrani P (2020)	High	1.9	BioGRID	

Previous 1 Next

Εικόνα 8. Οι αλληλεπιδράσεις (Interactions) των αλληλεπιδρώντων με την υπό εξέταση πρωτεΐνη trpA1.



Switch View: Interactors 5 | Interactions 5 | Chemical Interactions 4 | Network | PTM Sites 1

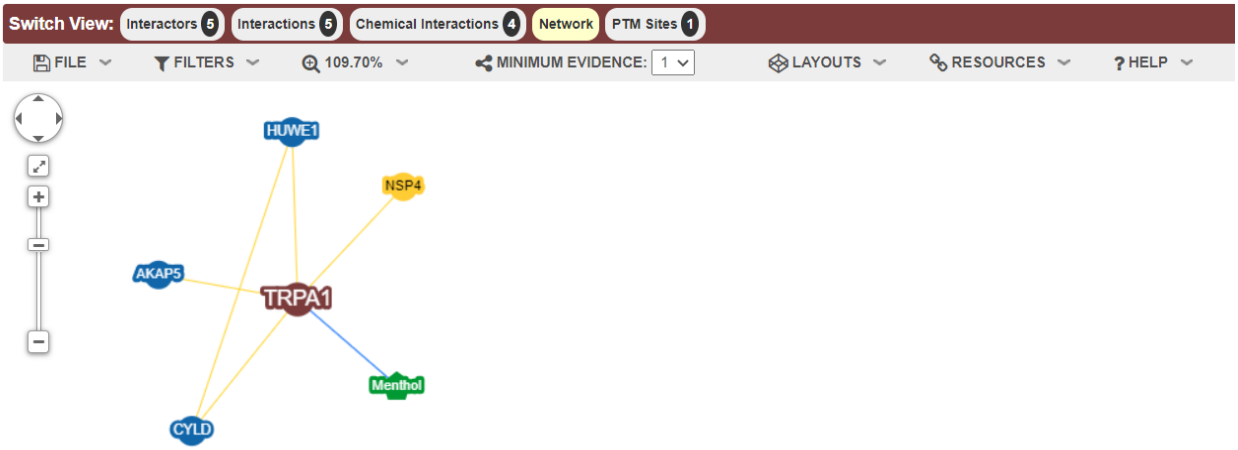
Showing 1 to 4 of 4 chemical interactions

Filter Interactions... ADV

Chemical	Chemical Type	Action	Dataset	Type	Related Proteins	Source	More
Menthol	small molecule	Inducer	Namer B (2005)	target	-	DrugBank	-
Menthol	small molecule	Inducer	Chen X (2002)	target	-	DrugBank	-
Menthol	small molecule	Inducer	Macpherson LJ (2006)	target	-	DrugBank	-
Menthol	small molecule	Inducer	Story GM (2003)	target	-	DrugBank	-

Previous 1 Next

Εικόνα 9. Οι χημικές αλληλεπιδράσεις (Chemical Interactions) στις οποίες συμμετέχει η πρωτεΐνη.



Εικόνα 10. Το δίκτυο που δημιουργείται, το οποίο είναι διαδραστικό.

Επιπλέον, ο χρήστης μπορεί να αναζητήσει με βάση:

Αναγνωριστικό γονιδίου (Gene/Identifier Search) - Πραγματοποιείται μέσω της καρτέλας "Gene" από την κύρια σελίδα αναζήτησης. Ο χρήστης μπορεί να πληκτρολογήσει τον όρο που τον ενδιαφέρει όπως για παράδειγμα STE11 και η μηχανή θα αναζητήσει αναγνωριστικά που ταιριάζουν.

Δημοσιεύσεις (Publication Search by PubMed ID / Full Text) - Ο χρήστης σε αυτή τη περίπτωση μπορεί να αναζητήσει είτε πληκτρολογώντας ένα ή περισσότερα PubMed IDs ή κείμενο, χωρίς όμως να τα συνδυάσει.

Χημική Ουσία (Chemical) - Ο χρήστης μπορεί να πληκτρολογήσει λέξεις κλειδιά, ονόματα, μοριακούς τύπους χημικών ουσιών ή και αναγνωριστικά άλλων σχετικών βάσεων δεδομένων όπως για παράδειγμα η PubChem.

Οι πρωτεϊνικές αλληλεπιδράσεις, περιγράφουν τους μοριακούς μηχανισμούς μεταξύ τους, στο εσωτερικό ενός κυττάρου, και αντικατοπτρίζουν την δυναμική όλων των κυτταρικών αποκρίσεων, ενώ οι γενετικές αλληλεπιδράσεις δείχνουν τις λειτουργικές σχέσεις που αφορούν σε ρυθμιστικά κυτταρικά μονοπάτια και μηχανισμούς. Το σύνολο των αλληλεπιδράσεων αυτών δίνει μια εικόνα σχετικά με τις λειτουργίες και τη ρύθμιση των κυττάρων, αναδεικνύοντας την χρησιμότητα της βάση δεδομένων BioGRID στο πεδίο της έρευνας.

Η BioGRID έχει αναπτυχθεί σε σημείο που να φιλοξενεί πλέον δεδομένα μετα-μεταφραστικών τροποποιήσεων (PTMs) και τον σχολιασμό των χημικών αλληλεπιδράσεων μεταξύ γονιδίων / πρωτεϊνών και βιοδραστικών μικρών μορίων για στόχους φαρμάκου πρωτεΐνης ή και αλληλεπιδράσεις γονιδίου-φαρμάκου. Ένα παράδειγμα δεδομένων PTMs είναι οι θέσεις φωσφορυλίωσης και ουβικουιτινίωσης, που προέρχονται τόσο από μεθόδους HTP όσο και από LTP μελέτες. Η συλλογή των πληροφοριών στην BioGRID διέπεται από ελεγχόμενα πειραματικά λεξιλόγια, στηρίζεται σε μεθόδους εξόρυξης δεδομένων και ελέγχεται από μια εσωτερική ειδική βάση δεδομένων που ονομάζεται Interaction Management System (IMS), που έχει σχεδιαστεί για να ενισχύσει την παραγωγικότητα της διαδικασίας συλλογής δεδομένων και να ελαχιστοποιήσει τα ανθρώπινα λάθη μέσω συνεπούς ποιοτικού ελέγχου και ελέγχου ταυτότητας των αποτελεσμάτων, ενώ ταυτόχρονα υποστηρίζει ένα περιβάλλον πολλαπλών χρηστών με πολλούς επιμελητές σε όλο τον κόσμο [49].

3.1.3. Βάση Δεδομένων DIP

Η βάση δεδομένων ***DIP*** φιλοξενεί αλληλεπιδράσεις μεταξύ πρωτεϊνών οι οποίες έχουν προκύψει πειραματικά. Η βάση συνδυάζει πληροφορίες από διάφορες πηγές έτσι ώστε να δημιουργήσει ένα σύνολο αλληλεπιδράσεων μεταξύ πρωτεϊνών.

Τα δεδομένα της βάσης προέρχονται είτε ιδιόχειρα (manually) από ειδικούς ερευνητές είτε αυτόματα μέσω υπολογιστικών μεθόδων που χρησιμοποιούν γνώση από δίκτυα αλληλεπίδρασης πρωτεϊνών προερχόμενα από το πιο αξιόπιστο υποσύνολο δεδομένων της DIP.

Η βάση δεδομένων DIP είναι μία σχεσιακή βάση δεδομένων, δηλαδή περιέχει μια συλλογή δεδομένων οργανωμένη σε σχεσιακούς πίνακες και μέσω κατάλληλων μηχανισμών ο προγραμματιστής μπορεί να υλοποιήσει διάφορες διεργασίες όπως για παράδειγμα να εισάγει ή να τροποποιήσει δεδομένα. Οι βασικοί πίνακες της DIP είναι τρεις:

- ***PROTEIN*** – φιλοξενεί πληροφορίες για κάθε πρωτεΐνη.
- ***SOURCE*** – περιέχει τις πηγές των πειραματικών πληροφοριών.
- ***EVIDENCE*** – έχει αποθηκευμένες πληροφορίες για κάθε πείραμα.

Η βάση αποτελείται από πέντε ακόμα πίνακες, δύο από αυτούς είναι ο πίνακας *INTERACTION* και ο *INT_PRT* οι οποίοι συνεργάζονται ως εξής: στον πίνακα *INTERACTION* υπάρχουν πληροφορίες σχετικά με αλληλεπιδράσεις πρωτεϊνών, ενώ ο πίνακας *INT_PRT* παίζει τον ρόλο του διαμεσολαβητή για να "επικοινωνούν" ο *INTERACTION* και ο PROTEIN. Μία εγγραφή στον πίνακα INTERACTION αντιστοιχεί σε τουλάχιστον δύο εγγραφές του πίνακα *INT_PRT*: δύο εγγραφές σε περίπτωση που εκπροσωπεί δυαδικές αλληλεπιδράσεις ενώ πάνω από δύο εγγραφές σε περίπτωση που εκπροσωπεί σύμπλοκα πολλαπλών πρωτεϊνών. Επιπρόσθετοι πίνακες στη βάση είναι οι :

- ***METHOD*** - παρέχει πληροφορίες που βοηθούν στον σχολιασμό των πειραμάτων.
- ***TOPOLOGY*** - έχει αποθηκευμένες λεπτομέρειες της τοπολογίας ενός μοριακού συμπλόκου που προέρχεται από κάθε πείραμα και καθορίζει τον τύπο αλληλεπίδρασης.
- ***LOCATION*** - περιγράφει περιοχές πρωτεϊνών που συμμετέχουν σε αλληλεπιδράσεις.

Η βάση DIP δίνει τη δυνατότητα στον χρήστη να αναζητήσει για μία συγκεκριμένη πρωτεΐνη μέσω του ονόματος της, του σχολιασμού (annotation) ή του οργανισμού προέλευσης της. Στη περίπτωση που η πρωτεΐνη δεν βρίσκεται μέσα στη βάση τότε με τη βοήθεια του BLAST και της αναζήτησης με βάση το μοτίβο της πρωτεΐνης, εμφανίζονται παρόμοιες πρωτεΐνες με εκείνη που αναζητά ο χρήστης. Το μοτίβο αλληλεπίδρασης των πρωτεϊνών μπορεί να παρέχει πληροφορίες σχετικά με τις πιθανές αλλά όχι ακόμη προσδιορισμένες αλληλεπιδράσεις της πρωτεΐνης που είναι υπό εξέταση [50]. Αναλυτικότερα στην βάση δεδομένων DIP υπάρχουν έξι τρόποι αναζήτησης για μία ή περισσότερες πρωτεΐνες και αλληλεπιδράσεις. Οι έξι διαφορετικοί τύποι αναζήτησης είναι οι εξής:

- ***Node*** - αναζήτηση μίας πρωτεΐνης με βάση τα κριτήρια που δίνει ο χρήστης στα αντίστοιχα πεδία. Το αποτέλεσμα της συγκεκριμένης αναζήτησης είναι μία λίστα πρωτεϊνών που τηρούν τα κριτήρια αυτά.
- ***BLAST*** - αναζήτηση ομοιότητας μιας πρωτεϊνικής αλληλουχίας ή ενός τμήματος της αλληλουχίας. Ως αποτέλεσμα είναι μία λίστα πρωτεϊνών ταξινομημένες με βάση το score.
- ***Motif*** - αναζήτηση στη βάση για πρωτεΐνες που περιέχουν μία αυτοτελή δομική περιοχή (domain) ή ένα μοτίβο που ορίζεται από μία από τις αντίστοιχες βάσεις δεδομένων όπως η Prosite, Pfam και SMART.
- ***Article*** - αναζήτηση για αλληλεπιδράσεις που περιγράφονται από συγκεκριμένα άρθρα.
- ***IMEx*** - αναζητούνται αλληλεπιδράσεις που προκύπτουν από πειράματα που σχολιάστηκαν σύμφωνα με τους κανόνες που υιοθετήθηκαν από το IMEx, μια κοινοπραξία που ιδρύθηκε από την DIP, με την συμμετοχή άλλων εταίρων, με σκοπό την μονιμοποίηση του σχολιασμού των πειραμάτων που αποδεικνύουν αλληλεπιδράσεις πρωτεΐνης.
- ***pathBLAST*** - αναζήτηση δικτύου πρωτεϊνικών αλληλεπιδράσεων για την εξαγωγή όλων των μονοπατιών αλληλεπιδράσεων που ταιριάζουν με το μονοπάτι που έθεσε αρχικά ο χρήστης.

3.1.4. Βάση Δεδομένων IntAct

Η ***IntAct*** είναι ένα ελεύθερα προσβάσιμο σύστημα βάσης δεδομένων που περιέχει εργαλεία ανάλυσης, αποθήκευσης και παρουσίασης δεδομένων μοριακών αλληλεπιδράσεων. Οι πληροφορίες στην IntAct αποτελούνται κυρίως από δεδομένα αλληλεπίδρασης πρωτεϊνών (PPI) που βοηθούν στη διευκρίνιση της κυτταρικής λειτουργίας, αλλά επιπλέον περιλαμβάνει και αλληλεπιδράσεις πρωτεΐνης-μικρού μορίου,

πρωτεΐνης-νουκλεϊκού οξέος και πρωτεΐνης-γονιδίου. Σε αυτές τις περιπτώσεις, οι βάσεις δεδομένων ChEBI, INSDC και Ensembl είναι πόροι αναφοράς.

Η δομή της IntAct αποτελείται από τρία κύρια συστατικά: Πείραμα, Αλληλεπίδραση και Διαδραστής.

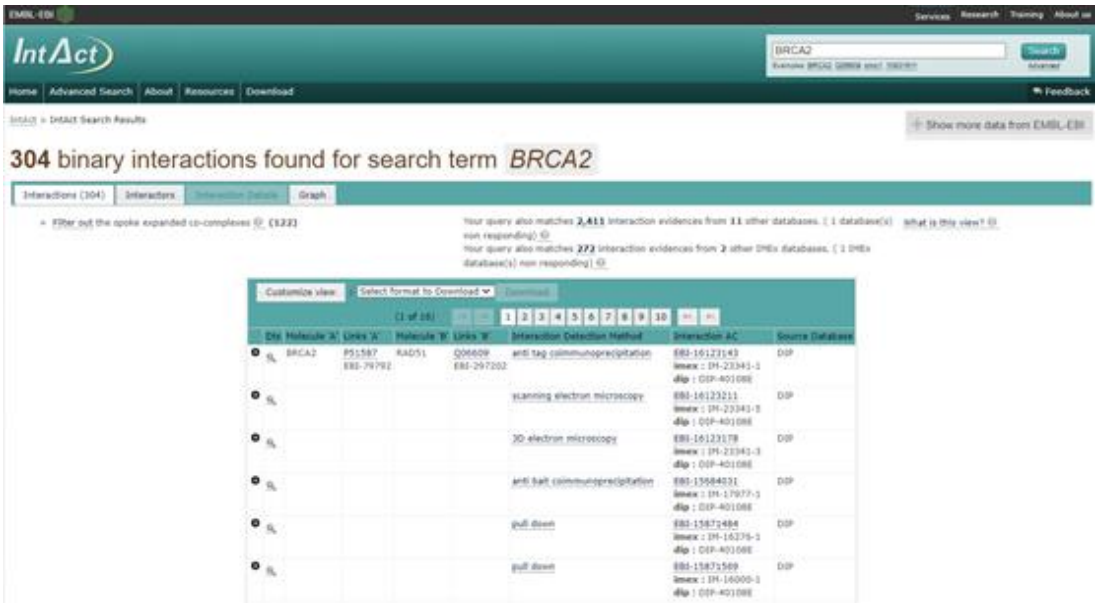
- ***Πείραμα*** - ομαδοποιεί έναν αριθμό αλληλεπιδράσεων, συνήθως από μία δημοσίευση, και ταξινομεί τις πειραματικές συνθήκες στις οποίες έχουν δημιουργηθεί αυτές οι αλληλεπιδράσεις. Ένα πείραμα μπορεί να έχει μόνο μία αλληλεπίδραση ή εκατοντάδες στην περίπτωση πειραμάτων μεγάλης κλίμακας.
- ***Διαδραστής*** - μια βιολογική οντότητα που συμμετέχει σε μια αλληλεπίδραση, συνήθως μια πρωτεΐνη, αλλά ακόμα μπορεί να είναι μια αλληλουχία DNA ή ένα μικρό μόριο.
- ***Αλληλεπίδραση*** - περιέχει έναν ή περισσότερους διαδραστές που συμμετέχουν στην αλληλεπίδραση. Η αναπαράσταση των αλληλεπιδράσεων δεν περιορίζεται μόνο σε δυαδικές αλληλεπιδράσεις.

Η IntAct είναι ενεργό μέλος του IMEx και ένα μεγάλο ποσοστό των πρωτεϊνικών δεδομένων της, είναι σχολιασμένα με βάση τις προδιαγραφές της IMEx. Οι εγγραφές της βάσης περιέχουν μια πλήρη περιγραφή των πειραματικών συνθηκών στις οποίες καταγράφηκε η αλληλεπίδραση ενώ υπάρχει και ένα υποσύνολο σχολιασμένων εγγραφών με βάση το λιγότερο ολοκληρωμένο πρότυπο MIMIx. Στην πράξη, αυτό σημαίνει ότι ενώ οι λεπτομέρειες του οργανισμού-ξενιστή, οι αλληλεπιδράσεις και οι μεθοδολογίες των συμμετεχόντων καταγράφονται, οι λεπτές λεπτομέρειες του συστήματος δεν καταγράφονται.

Η κάθε πρωτεΐνη ελέγχεται με κάθε έκδοση της UniProtKB, δηλαδή εάν μια αμινοξική ακολουθία έχει αποσυρθεί από την UniProtKB, τότε γίνεται αναζήτηση στην IntAct και η πρωτεΐνη αναδιαμορφώνεται εφόσον αυτό είναι εφικτό, διαφορετικά η ακολουθία παραμένει στην IntAct ως έχει και σε περίπτωση που η UniProtKB κυκλοφορήσει νέα έκδοση, τότε η αναζήτηση ταυτοποίησης επαναλαμβάνεται. Η αναζήτηση βασίζεται σε ένα αντίγραφο από το ίδιο γονίδιο, από τον ίδιο οργανισμό και με ομοιότητα ακολουθίας άνω του 98%.

Κάθε νέα καταχώρηση στην IntAct ελέγχεται από έναν ανώτερο επιμελητή και δεν δημοσιεύεται έως ότου εγκριθεί από αυτόν. Σε επίπεδο βάσης δεδομένων, εκτελούνται επιπλέον έλεγχοι και διορθώσεις όπου είναι απαραίτητο. Τέλος, κατά την κυκλοφορία των δεδομένων, ζητείται από τον αρχικό συγγραφέα κάθε έκδοσης να σχολιάσει την αναπαράσταση των δεδομένων τους και να προβεί σε διορθώσεις όπου χρειάζεται [12], [51].

Στην Εικόνα 11 εμφανίζεται η λίστα των αλληλεπιδράσεων με βάση τις προσαρμογές που κάνει ο χρήστης κατά την αναζήτηση του γονιδίου BRCA2. Παρά το γεγονός ότι τα δεδομένα της βάσης είναι σχολιασμένα με σκοπό να αντικατοπτρίζουν με ακρίβεια τις αλληλεπιδράσεις που αναφέρονται στην επιστημονική βιβλιογραφία, τα δεδομένα εμφανίζονται στην οθόνη ως δυαδικές αλληλεπιδράσεις. Το ίδιο συμβαίνει και στα δεδομένα που προέρχονται από ένα σύνθετο σύμπλεγμα που περιλαμβάνει περισσότερα από δύο μόρια, τα οποία ενώ αποθηκεύονται ως “σύνθετα” στη βάση δεδομένων IntAct, επεξεργάζονται και εμφανίζονται ως δυαδικά. Υπάρχουν αρκετές επιλογές λήψης που επιτρέπουν στους χρήστες να ανακτήσουν τις αλληλεπιδράσεις.



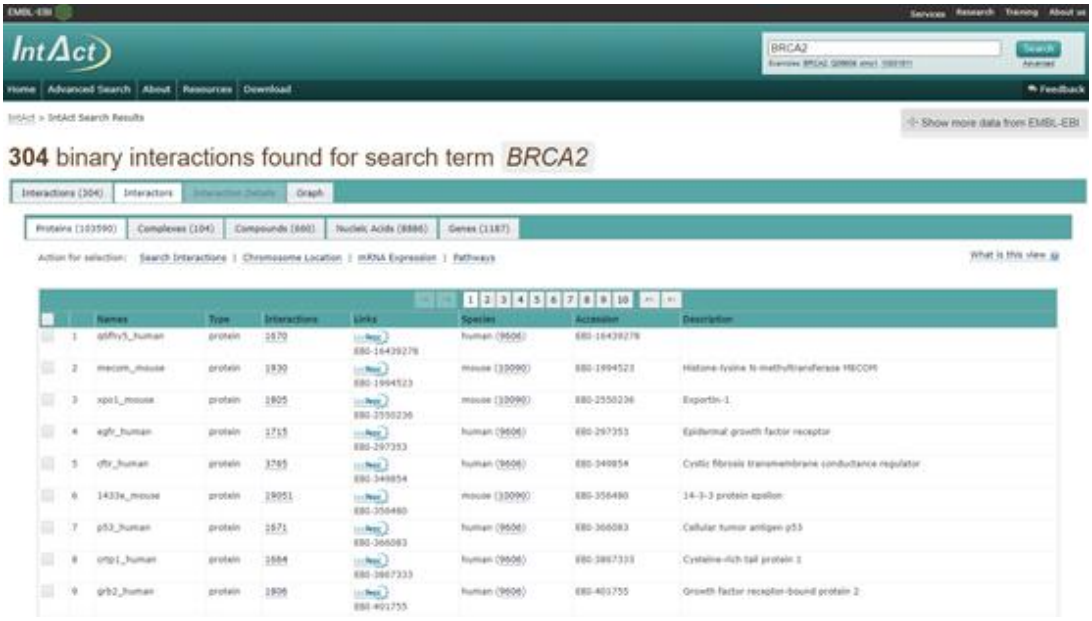
Εικόνα 11. Η λίστα των αλληλεπιδράσεων του γονιδίου BRCA2.

Ο πίνακας μπορεί να εμφανιστεί με τέσσερις διαφορετικούς τρόπους:

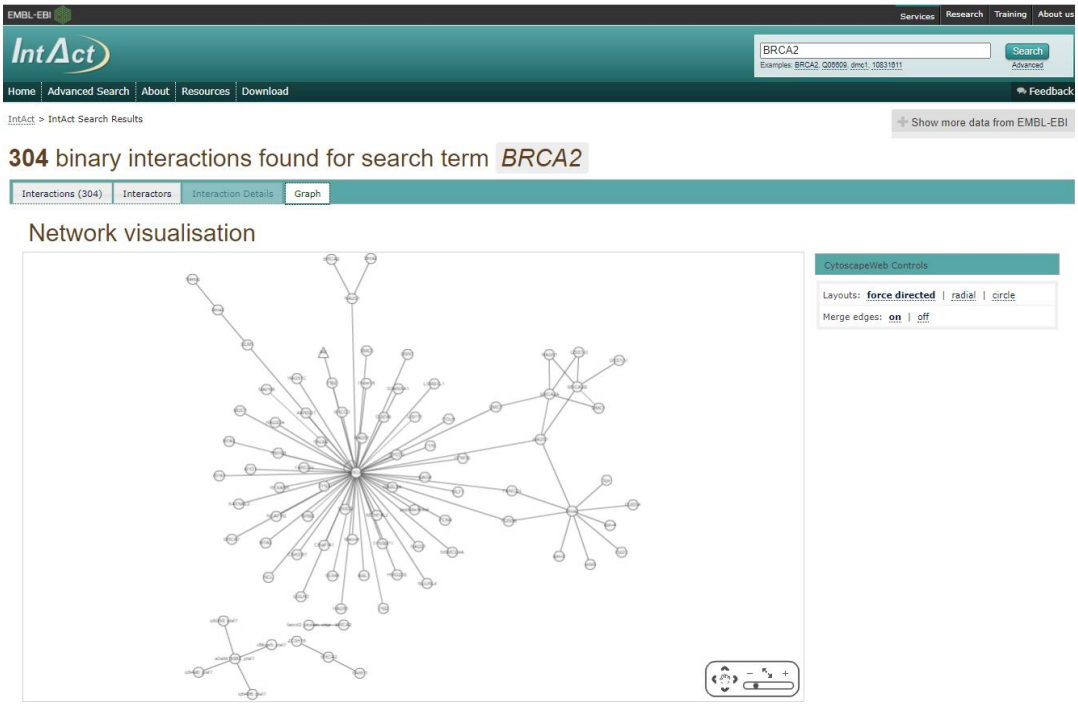
- ***Minimal*** – ονόματα μορίων και κωδικός αλληλεπίδρασης (Interaction AC).
- ***Basic*** - ονόματα μορίων, κωδικός αλληλεπίδρασης (Interaction AC), συνδέσεις μορίων, μέθοδος ανίχνευσης αλληλεπίδρασης και μέθοδος ανίχνευσης αλληλεπίδρασης.
- ***Standard*** - ονόματα μορίων, κωδικός αλληλεπίδρασης (Interaction AC), είδη μορίων, εμπιστευτικές πληροφορίες, λεπτομέρειες δημοσίευσης, λεπτομέρειες πειράματος.
- ***Expanded*** – όμοιο με το Standard προσθέτοντας επιπλέον λεπτομέρειες του πειράματος.
- ***Complete*** - περιέχει όλες τις διαθέσιμες πληροφορίες σε στήλες.

Η Εικόνα 12, εμφανίζεται εάν ο χρήστης πλοηγηθεί στην καρτέλα Interactors και δείχνει όλα τα μόρια που εμπλέκονται στο τρέχον επιλεγμένο σύνολο αλληλεπιδράσεων ανά τύπο αλληλεπιδρώντος μορίου (πρωτεΐνες, σύμπλοκα, χημικές ενώσεις, νουκλεϊκά οξέα, γονίδια).

Τέλος στην καρτέλα Graph παρουσιάζεται ο γράφος που δημιουργείται από τις αλληλεπιδράσεις του βιομορίου που αναζητά ο χρήστης (Εικόνα 13).



Εικόνα 12. Η λίστα με όσα μόρια εμπλέκονται στις αλληλεπιδράσεις με το γονίδιο BRCA2.



Εικόνα 13. Ο γράφος που σχηματίζεται από τις αλληλεπιδράσεις με το γονίδιο BRCA2.

3.2. Δίκτυα γονιδιακής συν-έκφρασης (Gene co-expression networks)

3.2.1. Η βάση δεδομένων Coexpedia

Η βάση δεδομένων ***Coexpedia***, είναι μια βάση συν-έκφρασης γονιδίων η οποία έχει τρία βασικά χαρακτηριστικά που την κάνουν να ξεχωρίζει [52]:

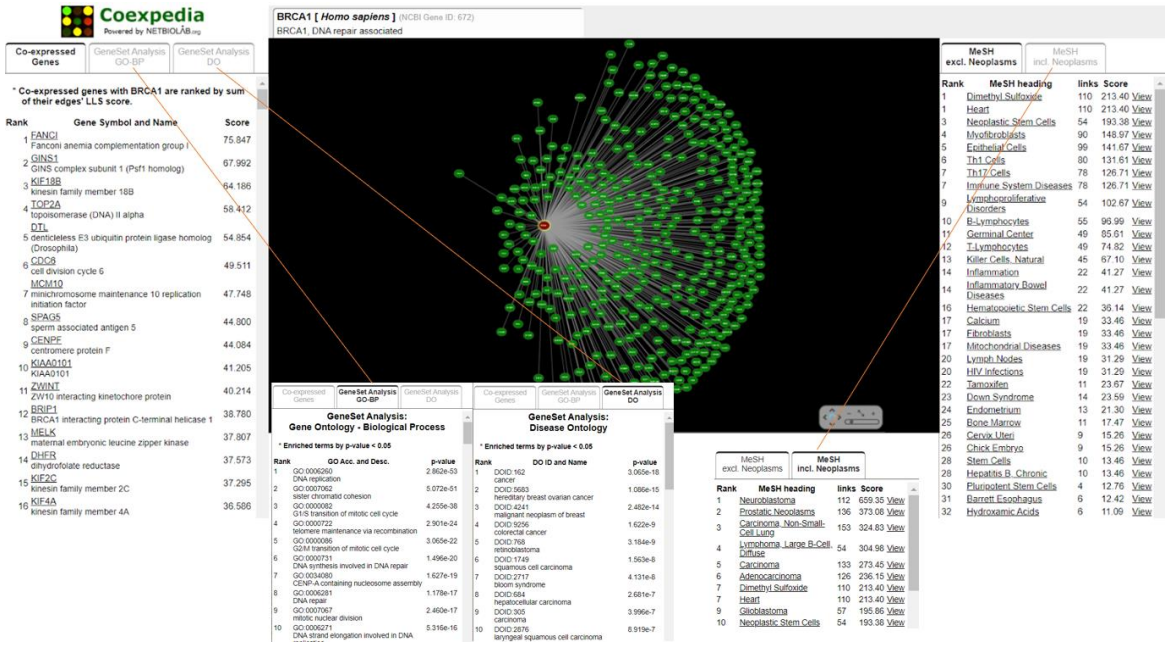
1. Οι συν-εκφράσεις που περιέχει είναι αποκλειστικά συν-λειτουργικές και αξιολογήθηκαν στατιστικά για τη λειτουργική συσχέτιση.
2. Τα δεδομένα συλλέχθηκαν από μεμονωμένες μελέτες, όπου η κάθε μελέτη στοχεύει στη διερεύνηση των γονιδιακών λειτουργιών σε σχέση με ένα συγκεκριμένο βιοϊατρικό πλαίσιο όπως για παράδειγμα μια ασθένεια.
3. Οι συν-εκφράσεις σχετίζονται με επικεφαλίδες ιατρικών θεμάτων (Medical Subject Headings - MeSH), παρέχοντας, βιοϊατρικές πληροφορίες για ασθένειες, ανατομικές και χημικές συσχετίσεις.

Οι σύνδεσμοι συν-έκφρασης της βάσης προέρχονται από δεδομένα μικροσυστοιχιών του βασικού αποθετηρίου δεδομένων Gene Expression Omnibus (GEO). Η Coexpedia φιλοξενεί οκτώ εκατομμύρια συν-εκφράσεις προερχόμενες από 384 GEO Series που αφορούσαν τον άνθρωπο και άλλες 248 που αφορούσαν τα ποντίκια. Ένα GEO Series είναι μία εγγραφή που έχει υποβληθεί στο GEO και συνοψίζει μια μελέτη.

Υπάρχουν δύο διαφορετικοί τρόποι αναζήτησης στη βάση Coexpedia:

1) Αναζήτηση εισάγοντας ένα μόνο γονίδιο: μέσω αυτής της αναζήτησης ο χρήστης μπορεί να ταυτοποιήσει ποια γονίδια της βάσης συν-εκφράζουν το γονίδιο που πληκτρολόγησε. Επιπλέον μπορεί να δει τις σχετικές λειτουργίες και τους φαινοτύπους του εισαγόμενου από τον χρήστη γονιδίου.
2) Αναζήτηση εισάγοντας πολλαπλά γονίδια: με την συγκεκριμένη αναζήτηση ο χρήστης μπορεί να μάθει ποια γονίδια της βάσης συν-εκφράζονται από κάποιο από τα εισαγόμενα γονίδια. Και σε αυτή τη περίπτωση ο χρήστης μπορεί να μάθει τις σχετικές λειτουργίες και τους φαινοτύπους των γονιδίων που αναζήτησε.

Για τις δύο διαφορετικές περιπτώσεις αναζήτησης, υπάρχουν έτοιμα παραδείγματα-γονίδια που μπορεί να δοκιμάσει ο χρήστης. Ένα από αυτά είναι το ανθρώπινο γονίδιο BRCA1 του οποίου τα αποτελέσματα αναζήτησης φαίνονται στην Εικόνα 14. Ξεκινώντας από τα αριστερά, η πρώτη στήλη που συναντά ο χρήστης εμφανίζει τα γονίδια που συν-εκφράζονται (Tab: Co-expressed Genes), τις σχετικές λειτουργίες τους (Tab: GeneSet Analysis GO-BP) και του φαινοτύπους τους (Tab: GeneSet Analysis DO). Τα αποτελέσματα είναι ταξινομημένα σύμφωνα με το αποτέλεσμα του αναγραφόμενου Score. Στο κέντρο παρουσιάζεται το δίκτυο που δημιουργείται στο οποίο διακρίνεται με κόκκινο χρώμα το γονίδιο που αναζήτησε ο χρήστης (στη περίπτωση αυτή το BRCA1). Τα γονίδια που είναι πιο κοντά στο κόκκινο γονίδιο, είναι αυτά με το υψηλότερο Score. Εάν επιλέξει κάποιον MeSH όρο από την δεξιά στήλη, τότε η αντίστοιχη σύνδεση στο δίκτυο θα εμφανιστεί πιο έντονα από τις υπόλοιπες. Τέλος, στην δεξιά στήλη εμφανίζονται MeSH όροι μεταξύ των συν-εκφρασμένων γονιδίων, επίσης ταξινομημένοι με βάση το Score [53].



Εικόνα 14. Αποτελέσματα αναζήτησης του ανθρώπινου γονιδίου BRCA1. Αριστερά στο πρώτο tab υπάρχουν τα συν-εκφρασμένα γονίδια του BRCA1, στο δεύτερο tab όπου είναι οι σχετικές λειτουργίες τους, φαίνεται πως το BRCA1 συν-εκφράζεται με τα γονίδια που είναι γνωστό ότι εμπλέκονται στην διαδικασία αντιγραφής του DNA, στην μετάβαση G1 / S του μιτωτικού κυτταρικού κύκλου και πολλές ακόμα όπου πιθανότατα εμπλέκεται και το ίδιο το BRCA1.Στο τρίτο tab που είναι για τους φαινοτύπους φαίνεται πως το BRCA1 συν-εκφράζεται με γονίδια που εμπλέκονται σε ασθένειες όπως ο καρκίνος όπου πιθανότατα και το ίδιο το BRCA1 να εμπλέκεται σε αυτές. Στην δεξιά στήλη φαίνεται πως οι MeSH όροι χωρίζονται σε δύο κατηγορίες/tabs: στο πρώτο που δεν σχετίζεται με νεοπλάσματα, όπου φαίνεται ως κορυφαίος όρος για το BRCA1, η καρδιά ("Heart") και στο δεύτερο που σχετίζεται με τα νεοπλάσματα και ως κορυφαίος όρος είναι το καρκίνωμα, μη μικροκυτταρικός πνεύμονας (Carcinoma, non-small-cell lung).

3.2.2. Η βάση δεδομένων GeneMania

Η εφαρμογή ***GeneMANIA***, είναι μία εύχρηστη, διαδικτυακή εφαρμογή με σκοπό την δημιουργία υποθέσεων σχετικά με τη γονιδιακή λειτουργία, την ανάλυση γονιδιακών λιστών και την ιεράρχηση γονιδίων για λειτουργικές δοκιμασίες. Τα δεδομένα συσχέτισης περιλαμβάνουν πρωτεΐνες και γενετικές αλληλεπιδράσεις, μονοπάτια, συν-έκφραση, συν-εντοπισμό και ομοιότητα πρωτεϊνικού τομέα. Προς το παρόν, η εφαρμογή υποστηρίζει έξι οργανισμούς. Εισάγοντας μία λίστα γονιδίων, η GeneMANIA επεκτείνει την λίστα με επιπλέον γονίδια που έχουν παρόμοια λειτουργικότητα με τα αρχικά. Η ταυτοποίηση των γονιδίων πραγματοποιείται με χρήση των διαθέσιμων δεδομένων γονιδιωματικής και πρωτεωμικής. Η εφαρμογή περιλαμβάνει επίσης βάρη που υποδεικνύουν την προγνωστική αξία κάθε επιλεγμένου συνόλου δεδομένων για το εισαγόμενο query [54].

Η εφαρμογή GeneMANIA αναζητά πολλά μεγάλα, ευρέως διαθέσιμα βιολογικά σύνολα δεδομένων για να βρει σχετικά γονίδια. Αυτά τα βιολογικά σύνολα δεδομένων, τα οποία ενημερώνονται τακτικά, περιλαμβάνουν αλληλεπιδράσεις πρωτεϊνών, αλληλεπιδράσεις πρωτεϊνών-DNA και γενετικές αλληλεπιδράσεις, μονοπάτια, αντιδράσεις, δεδομένα έκφρασης γονιδίων και πρωτεϊνών, τομείς πρωτεϊνών και φαινοτυπικά προφίλ.

Τα ονόματα των δικτύων περιγράφουν την πηγή δεδομένων και προκύπτουν είτε από την καταχώρηση PubMed που σχετίζεται με την πηγή δεδομένων, είτε από το όνομα της πηγής δεδομένων (BioGRID, PathwayCommons, Pfam) . Στα δίκτυα συν-έκφρασης, δύο γονίδια συνδέονται εάν τα επίπεδα

έκφρασής τους είναι παρόμοια μεταξύ των συνθηκών μιας μελέτης γονιδιακής έκφρασης. Τα περισσότερα από αυτά τα δεδομένα συλλέγονται από το Gene Expression Omnibus (GEO) και όλα σχετίζονται με μια δημοσίευση.

Η λειτουργικότητα της εφαρμογής αποδίδει καλύτερα, εάν η πλειοψηφία των εισαγόμενων γονιδίων σχετίζονται λειτουργικά. Η λειτουργία της λειτουργικής συσχέτισης αρκεί να καταγράφεται από κάποια λειτουργικά δίκτυα συσχέτισης της GeneMANIA. Σε περίπτωση που δεν σχετίζονται τότε δημιουργείται ένα αποσυνδεδεμένο δίκτυο και η στάθμισή του δεν θα είναι η βέλτιστη.

Εάν η λίστα αναζήτησης αποτελείται από 6 ή περισσότερα γονίδια, η εφαρμογή GeneMANIA θα υπολογίσει τα βάρη για τη συγκεκριμένη λίστα, ενώ εάν η λίστα αναζήτησης έχει λιγότερα από 6 γονίδια, η εφαρμογή θα κάνει προβλέψεις για τη λειτουργία γονιδίων με βάση μοτίβα σχολιασμών GO.

Κάθε πηγή δεδομένων δικτύου αναπαρίσταται από ένα σταθμισμένο δίκτυο αλληλεπίδρασης όπου σε κάθε ζεύγος γονιδίων αντιστοιχιστεί ένα βάρος συσχετισμού, το οποίο είτε είναι μηδέν, στη περίπτωση που δεν υπάρχει αλληλεπίδραση, είτε είναι μία τιμή μεγαλύτερη από το μηδέν που αντικατοπτρίζει τη δύναμη της αλληλεπίδρασης ή την αξιοπιστία του εντοπισμού αλληλεπίδρασης. Εάν υπάρχουν πληροφορίες δυαδικής μορφής (π.χ.: αλληλεπιδράσεις πρωτεΐνης, όπου όταν αλληλεπιδρούν δύο πρωτεΐνες, η ακμή του δικτύου τους έχει βάρος 1) τότε ως δίκτυα χρησιμοποιούνται οι άμεσες αλληλεπιδράσεις [55].

3.2.3. Η βάση δεδομένων COXPRESdb

Το ***COXPRESdb*** είναι μια βάση δεδομένων που περιέχει πληροφορίες συν-έκφρασης που αν και όταν πρώτο-κυκλοφόρησε το 2007 οι πληροφορίες ήταν για ανθρώπους και ποντίκια, πλέον έχει πληροφορίες για 11 είδη ζώων. Η βάση παρέχει επιπλέον λειτουργικότητες μέσα στις οποίες είναι η αναζήτηση συν-εκφρασμένων γονιδίων χρησιμοποιώντας λειτουργικά συγγενικά πολλαπλά γονίδια, σχεδιάζοντας συν-εκφρασμένο γονιδιακό δίκτυο με πληροφορίες μονοπατιών και αλληλεπίδρασης πρωτεϊνών, και αυτόματη ανίχνευση και ανάλυση δομών υποσυστημάτων δικτύων συν-εκφρασμένων γονιδίων.

Σχετικά με τον υπολογισμό της συν-έκφρασης θα πρέπει να αναφερθεί πως η σχέση συν-έκφρασης είναι μια σύνοψη ενός συνόλου μεταγραφικών δεδομένων και επομένως η ποιότητα των δεδομένων συν-έκφρασης εξαρτάται σε μεγάλο βαθμό από αυτήν των υποκείμενων μεταγραφικών δεδομένων. Κάθε μεταγραφικό στοιχείο έχει εγγενώς κάποια τεχνική και βιολογική προδιάθεση. Για παράδειγμα, διαφορετική τεχνολογία έχει διαφορετικούς συστηματικούς θορύβους και συγκεκριμένα είδη επιλέγονται συνήθως για συγκεκριμένες έρευνες. Η σύγκριση των ανεξάρτητων δεδομένων συν-έκφρασης είναι αποτελεσματική ώστε να καταλήξουμε με λιγότερες σχέσεις συν-έκφρασης που βασίζονται στη τεχνολογική ή βιολογική προδιάθεση.

Χαρακτηριστικό του COXPRESdb είναι η ικανότητά του να συγκρίνει πολλαπλά δεδομένα συν-έκφρασης που προέρχονται από διαφορετικές τεχνολογίες μεταγραφικής και διαφορετικά είδη. Κατά συνέπεια υπάρχει σημαντική μείωση στις ψευδείς θετικές σχέσεις σε δεδομένα μεμονωμένων γονιδιακών συν-εκφράσεων [56].

Η ανάλυση μικροσυστοιχιών DNA παράγει πληροφορίες με τα σχετικά επίπεδα έκφρασης χιλιάδων γονιδίων, ταυτόχρονα. Επιπλέον, οι μεγάλες συλλογές δεδομένων μικροσυστοιχιών περιέχουν πληροφορίες σχετικά με συντονισμένες αλλαγές στα επίπεδα μεταγραφής στα συγκεκριμένα σύνολα δεδομένων ανεξάρτητα από τον αρχικό σκοπό του κάθε συνόλου δεδομένων. Η ομοιότητα της γονιδιακής συν-έκφρασης μπορεί να οριστεί χρησιμοποιώντας το μοτίβο των αλλαγών της γονιδιακής

έκφρασης μεταξύ δύο γονιδίων. Ως μέτρο συν-έκφρασης γονιδίων χρησιμοποιείται ο συντελεστής συσχέτισης του Pearson όπου η τιμή 1 υποδηλώνει ισχυρή σχέση υπό την άποψη της ρύθμισης γονιδιακής έκφρασης, ενώ η τιμή 0 δεν δείχνει καμία σχέση.

Μέσω της γονιδιακής συν-έκφρασης, παρέχονται χρήσιμες πληροφορίες για τον εντοπισμό νέου γονιδίου που σχετίζεται λειτουργικά. Παρ' όλα αυτά, αυτή η σχέση συν-έκφρασης αντικατοπτρίζει ρυθμίσεις σε επίπεδο mRNA μόνο και όχι ρυθμίσεις σε επίπεδο πρωτεΐνης. Αυτό δείχνει πως οι πληροφορίες συν-έκφρασης δεν είναι αποτελεσματικές εάν το σύστημα γονίδιο-στόχος δεν ρυθμίζεται σε επίπεδο mRNA. Γι’ αυτόν τον λόγο, στο δίκτυο συν-έκφρασης γονιδίων έχουν ενσωματωθεί ήδη γνωστές πληροφορίες αλληλεπίδρασης πρωτεϊνών. Η γονιδιακή συν-έκφραση και η αλληλεπίδραση πρωτεϊνών υποδεικνύουν άλλο επίπεδο ρύθμισης και κατά συνέπεια προσφέρουν συμπληρωματικές πληροφορίες για την κατανόηση του δικτύου γονιδιακής λειτουργίας [57].

3.2.4. Η βάση δεδομένων GeneFriends

Το ***GeneFriends*** είναι μία βάση δεδομένων για την ταυτοποίηση συν-εκφρασμένων γονιδίων με ένα ή περισσότερα γονίδια που ορίζει ο χρήστης. Είναι η πρώτη διαδικτυακή βάση δεδομένων συν-έκφρασης RNA-seq για την κοινότητα της βιοεπιστήμης. Τα αποτελέσματα που λαμβάνει ο χρήστης περιλαμβάνουν ένα δίκτυο συν-έκφρασης καθώς και μια περίληψη του λειτουργικού εμπλουτισμού (functional enrichment) μεταξύ των γονιδίων που εκφράζονται.

Το GeneFriends έχει βασιστεί σε δεδομένα μικροσυστοιχιών για τη δημιουργία δικτύων συν-έκφρασης. Οι αναλύσεις συν-έκφρασης έχουν εντοπίσει νέα γονίδια που εμπλέκονται σε ασθένειες όπως είναι ο καρκίνος και ο διαβήτης τύπου 2. Η βάση, περιέχει δεδομένα συν-έκφρασης για 44,248 ανθρώπινα γονίδια και για 114,936 μεταγραφές.

Η βάση μπορεί να χρησιμοποιηθεί για την εκχώρηση ενδεχόμενων λειτουργιών σε μη καλώς μελετημένα γονίδια χρησιμοποιώντας μια μέθοδο ονομαζόμενη: ενοχή ανά συσχέτιση (guilt-by-association) όπου διερευνάται με ποια γονίδια ένα μη καλώς μελετημένο γονίδιο, συν-εκφράζεται. Επίσης, η βάση μπορεί να εντοπίσει και να δώσει προτεραιότητα σε νέα υποψήφια γονίδια για περαιτέρω μελέτη με βάση μία λίστα αρχικών γονιδίων που σχετίζονται με μια δεδομένη ασθένεια ή βιολογική διαδικασία. Με αυτόν τον τρόπο επιτρέπεται στους ερευνητές να εντοπίσουν νέα γονίδια που σχετίζονται με τη μελέτη τους χωρίς την ανάγκη διεξαγωγής ολόκληρου πειράματος μικροσυστοιχίας ή RNA-seq. Χαρακτηριστικό παράδειγμα αποτελεί ο εντοπισμός νέων γονιδίων που σχετίζονται με τον καρκίνο που ύστερα επικυρώθηκαν πειραματικά.

Ο χρήστης μπορεί να υποβάλει στο πεδίο αναζήτησης της βάσης ένα ή περισσότερα αναγνωριστικά γονιδίων / μεταγραφής (gene/transcript IDs). Τα αποτελέσματα που εμφανίζονται περιέχουν [58]:

- Μια λίστα με τα 50 ισχυρότερα συν-εκφρασμένα γονίδια και τον αντίστοιχο σχολιασμό HGNC για κάθε γονίδιο.
- Μια λίστα με τους 25 ισχυρότερους συν-εκφρασμένους παράγοντες μεταγραφής.
- Τις 20 κορυφαίες κατηγορίες λειτουργικού εμπλουτισμού (functional enrichment) της συν-εκφρασμένης λίστας γονιδίων, ανάμεσα τους οι: GO, KEGG και OMIM.

3.3. Δίκτυα ομοιότητας αλληλουχιών (Sequence similarity networks - SSNs)

3.3.1. Εισαγωγικές έννοιες

Η στοίχιση ακολουθιών (sequence alignment) είναι μια ευρέως διαδεδομένη τεχνική σύγκρισης αλληλουχιών δύο μακρομορίων. Με την στοίχιση των ακολουθιών συγκρίνονται δύο ακολουθίες οι οποίες τοποθετούνται με τέτοιο τρόπο ώστε τα ίδια κατάλοιπα να είναι ευθυγραμμισμένα οπότε και επιτυγχάνεται μία αντιστοιχία (match), ενώ στη περίπτωση που έχουμε αναντιστοιχία (mismatch), μπορεί να ερμηνευθεί ως σημειακή μετάλλαξη. Στις περιοχές όπου τα κατάλοιπα των δύο ακολουθιών δεν σχετίζονται μεταξύ τους, τότε έχουμε κενά (gaps) τα οποία οφείλονται στις μεταλλάξεις εισαγωγής στη μία αλληλουχία ή διαγραφής στην άλλη.

Βασικός σκοπός της σύγκρισης ακολουθιών είναι ο υπολογισμός ομοιότητας των μακρομορίων από όπου προέρχονται, γεγονός που μπορεί να φανερώνει σημαντικές βιολογικές πληροφορίες.

Σε ένα δίκτυο ομοιότητας αλληλουχιών οι αλληλουχίες αναπαριστώνται ως κόμβοι και οι ομοιότητες μεταξύ τους ως ακμές οι οποίες προκύπτουν εφόσον ικανοποιούν τις συνθήκες που ορίζουμε χρησιμοποιώντας κατώφλια. Τα κατώφλια (thresholds) εξυπηρετούν στην ρύθμιση του ποσοστού ομοιότητας μεταξύ δύο ακολουθιών, της εξεταζόμενης ακολουθίας και της ακολουθίας αναφοράς (query sequence και reference sequence) [59].

Παρά το γεγονός ότι για τον εντοπισμό της λειτουργίας των γονιδίων και τη συσχέτιση νέων γονιδίων με μία λίστα αρχικών γονιδίων, υπάρχουν εργαλεία, πληροφορίες για την αλληλεπίδραση των μη κωδικοποιητικών RNA (ncRNA) είναι ελάχιστες. Γι’ αυτό τον λόγο στο GeneFriends υπάρχει ένας χάρτης συν-έκφρασης που κατασκευάστηκε από δεδομένα RNA-seq, που επιτρέπει την καλύτερη κατανόηση των ρυθμιστικών προτύπων των ncRNA σε σχέση με τα mRNAs. Με αφορμή το γεγονός ότι το RNA-seq επιτρέπει στους ειδικούς να υπολογίσουν την έκφραση διαφορετικών μεταγραφών και όχι μόνο την έκφραση σε γονιδιακό επίπεδο, έχει κατασκευαστεί ένας χάρτης συν-έκφρασης μεταγραφής. Έτσι λοιπόν, διαφορετικά αντίγραφα που προέρχονται από το ίδιο γονίδιο μπορεί να διαφέρουν στην λειτουργικότητα τους και η συν-έκφραση είναι ένας εύκολος τρόπος για την ανίχνευση διαφορετικών μοτίβων γονιδιακής έκφρασης, υποδηλώνοντας διαφορετική λειτουργικότητα [60].

3.3.2. Δημιουργία δικτύων ομοιότητας αλληλουχιών

Για να δημιουργήσουμε το δίκτυο ομοιότητας αλληλουχιών αρχικά αναζητήσαμε και αποθηκεύσαμε τα πρωτεώματα διάφορων οργανισμών όπως για παράδειγμα του ανθρώπου, του ποντικιού και της δροσόφιλα, από την βάση δεδομένων UniProt. Αφού επεξεργαστήκαμε τα αρχεία των πρωτεωμάτων δημιουργώντας για τον κάθε οργανισμό από ένα νέο αρχείο που περιείχε μέσα τους κωδικούς (accession number) και τις αντίστοιχες ακολουθίες σε fasta μορφή, τότε κάναμε χρήση του lastal. Το lastal είναι ένα πρόγραμμα που παρέχει η εφαρμογή LAST η οποία βρίσκει παρόμοιες περιοχές μεταξύ των ακολουθιών και τις στοιχίζει. Συγκεκριμένα το lastal βρίσκει τοπικές στοιχίσεις μεταξύ δύο αλληλουχιών. Στη χρήση του lastal προσδιορίσαμε τον πίνακα PAM30 (Εικόνα 15) ως τον κατάλληλο πίνακα με score αντιστοιχίας σχετικά μικρών και ισχυρών ομοιοτήτων μεταξύ ακολουθιών.

PAM30

This protein scoring scheme is good for finding strong, and short, similarities (MO Dayhoff et al. 1978). It uses this matrix:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*
A	6	-7	-4	-3	-6	-4	-2	-2	-7	-5	-6	-7	-5	-8	-2	0	-1	-13	-8	-2	-3	-6	-3	-1	-17
R	-7	8	-6	-10	-8	-2	-9	-9	-2	-5	-8	0	-4	-9	-4	-3	-6	-2	-10	-8	-7	-7	-4	-1	-17
N	-4	-6	8	2	-11	-3	-2	-3	0	-5	-7	-1	-9	-9	-6	0	-2	-8	-4	-8	6	-6	-3	-1	-17
D	-3	-10	2	8	-14	-2	2	-3	-4	-7	-12	-4	-11	-15	-8	-4	-5	-15	-11	-8	6	-10	1	-1	-17
C	-6	-8	-11	-14	10	-14	-14	-9	-7	-6	-15	-14	-13	-13	-8	-3	-8	-15	-4	-6	-12	-9	-14	-1	-17
Q	-4	-2	-3	-2	-14	8	1	-7	1	-8	-5	-3	-4	-13	-3	-5	-5	-13	-12	-7	-3	-5	6	-1	-17
E	-2	-9	-2	2	-14	1	8	-4	-5	-5	-9	-4	-7	-14	-5	-4	-6	-17	-8	-6	1	-7	6	-1	-17
G	-2	-9	-3	-3	-9	-7	-4	6	-9	-11	-10	-7	-8	-9	-6	-2	-6	-15	-14	-5	-3	-10	-5	-1	-17
H	-7	-2	0	-4	-7	1	-5	-9	9	-9	-6	-6	-10	-6	-4	-6	-7	-7	-3	-6	-1	-7	-1	-1	-17
I	-5	-5	-5	-7	-6	-8	-5	-11	-9	8	-1	-6	-1	-2	-8	-7	-2	-14	-6	2	-6	5	-6	-1	-17
L	-6	-8	-7	-12	-15	-5	-9	-10	-6	-1	7	-8	1	-3	-7	-8	-7	-6	-7	-2	-9	6	-7	-1	-17
K	-7	0	-1	-4	-14	-3	-4	-7	-6	-6	-8	7	-2	-14	-6	-4	-3	-12	-9	-9	-2	-7	-4	-1	-17
M	-5	-4	-9	-11	-13	-4	-7	-8	-10	-1	1	-2	11	-4	-8	-5	-4	-13	-11	-1	-10	0	-5	-1	-17
F	-8	-9	-9	-15	-13	-13	-14	-9	-6	-2	-3	-14	-4	9	-10	-6	-9	-4	2	-8	-10	-2	-13	-1	-17
P	-2	-4	-6	-8	-8	-3	-5	-6	-4	-8	-7	-6	-8	-10	8	-2	-4	-14	-13	-6	-7	-7	-4	-1	-17
S	0	-3	0	-4	-3	-5	-4	-2	-6	-7	-8	-4	-5	-6	-2	6	0	-5	-7	-6	-1	-8	-5	-1	-17
T	-1	-6	-2	-5	-8	-5	-6	-6	-7	-2	-7	-3	-4	-9	-4	0	7	-13	-6	-3	-3	-5	-6	-1	-17
W	-13	-2	-8	-15	-15	-13	-17	-15	-7	-14	-6	-12	-13	-4	-14	-5	-13	13	-5	-15	-10	-7	-14	-1	-17
Y	-8	-10	-4	-11	-4	-12	-8	-14	-3	-6	-7	-9	-11	2	-13	-7	-6	-5	10	-7	-6	-7	-9	-1	-17
V	-2	-8	-8	-8	-6	-7	-6	-5	-6	2	-2	-9	-1	-8	-6	-6	-3	-15	-7	7	-8	0	-6	-1	-17
B	-3	-7	6	6	-12	-3	1	-3	-1	-6	-9	-2	-10	-10	-7	-1	-3	-10	-6	-8	6	-8	0	-1	-17
J	-6	-7	-6	-10	-9	-5	-7	-10	-7	5	6	-7	0	-2	-7	-8	-5	-7	-7	0	-8	6	-6	-1	-17
Z	-3	-4	-3	1	-14	6	6	-5	-1	-6	-7	-4	-5	-13	-4	-5	-6	-14	-9	-6	0	-6	6	-1	-17
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-17
*	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	1

It sets these default lastal parameter values: **-a13 -b3**

Εικόνα 15. Ο πίνακας PAM30 με τα score στοίχισης των αμινοξέων.

Η διεργασία χωρίστηκε σε δέκα threads για να γίνονται πολλές συγκρίσεις παράλληλα. Τα αποτελέσματα της διεργασίας βασίζονται στη μορφή του BlastTab+ η οποία περιέχει: όνομα υπό εξέταση ακολουθίας, όνομα ακολουθίας αναφοράς, ποσοστό ομοιότητας, μήκος στοίχησης, αναντιστοιχίες, κενά, έναρξη εξεταζόμενης ακολουθίας, τέλος εξεταζόμενης ακολουθίας, έναρξη ακολουθίας αναφοράς, τέλος ακολουθίας αναφοράς, E-value, bit score, μήκος εξεταζόμενης ακολουθίας, μήκος ακολουθίας αναφοράς και score.

Επιθυμητές αλληλουχίες είναι όσες έχουν ποσοστό ομοιότητας (% Identity) τουλάχιστον 30% και το πηλίκο των παρακάτω σχέσεων τουλάχιστον 0.8.

$$\frac{(Align.length - Gaps)}{Length\ query\ seq}, \frac{(Align.length - Gaps)}{Length\ query\ ref}$$

4. Λειτουργική ανάλυση

4.1. Οντολογία Γονιδίων (Gene Ontology)

Η ***Οντολογία Γονιδίων (Gene Ontology / GO)*** είναι μία ευρέως γνωστή βάση δεδομένων που περιέχει πληροφορίες σχετικά με τα γονίδια, τις λειτουργίες και τις σχέσεις μεταξύ τους. Η GO αποτελείται από την οντολογία η οποία περιλαμβάνει τους όρους των γονιδίων (terms) που περιγράφουν την λειτουργία τους (gene function. Προς το παρόν, η βάση φιλοξενεί πάνω από 45,000 όρους που έχουν περίπου 134,000 σχέσεις να τους συνδέουν [61].

Βασικός στόχος της GO είναι η ύπαρξη ενός κοινού, δυναμικά δομημένου λεξιλογίου που θα ελέγχεται ώστε να μπορεί να περιγράφει τους ρόλους όχι μόνο των γονιδίων αλλά και των γονιδιακών προϊόντων όλων των οργανισμών [61], [62]. Για τον σκοπό αυτό, δημιουργήθηκαν τρεις ανεξάρτητες κατηγορίες-οντολογίες οι οποίες μπορούν να περιγράψουν όλες τις οντότητες και τις μεταξύ τους σχέσεις. Η κάθε κατηγορία έχει καλά προσδιορισμένους όρους και σχέσεις αναπαριστώντας με οργανωμένο τρόπο την βιολογική γνώση και έχοντας τον ρόλο του οδηγού για την προσθήκη νέων δεδομένων. Οι κατηγορίες αυτές είναι:

Βιολογική διαδικασία (Biological process): Η βιολογική διαδικασία αναφέρεται σε κάποιον βιολογικό στόχο όπου συνεισφέρει είτε το γονίδιο είτε το γονιδιακό προϊόν και ολοκληρώνεται με τη βοήθεια τουλάχιστον μίας ταξινομημένης ομάδας μοριακών λειτουργιών. Συνήθως στις διαδικασίες αυτές συναντάμε χημικούς ή φυσικούς μετασχηματισμούς, που σημαίνει πως κάτι εισέρχεται στην διαδικασία ενώ κάτι άλλο εξέρχεται από αυτή. Υπάρχουν όροι υψηλού και χαμηλού επιπέδου. Ένα παράδειγμα όρου υψηλού επιπέδου, βιολογικής διαδικασίας, είναι η κυτταρική ανάπτυξη και συντήρηση, ενώ ένα παράδειγμα όρου βιολογικής διαδικασίας χαμηλού επιπέδου είναι η διαδικασία της μετάφρασης.

Μοριακή λειτουργία (Molecular function): Ως μοριακή λειτουργία ορίζεται η βιοχημική δραστηριότητα ενός γονιδιακού προϊόντος ή ενός συμπλέγματος γονιδιακών προϊόντων. Στην μοριακή λειτουργία προσδιορίζεται μόνο η διαδικασία και όχι το σημείο και ο χρόνος που πραγματοποιήθηκε. Υπάρχουν ευρύτεροι και περιορισμένοι όροι μοριακής λειτουργίας. Παράδειγμα ενός ευρύ όρου είναι το ένζυμο, ενώ παράδειγμα περιορισμένου όρου είναι το πρόσδεμα του υποδοχέα Toll.

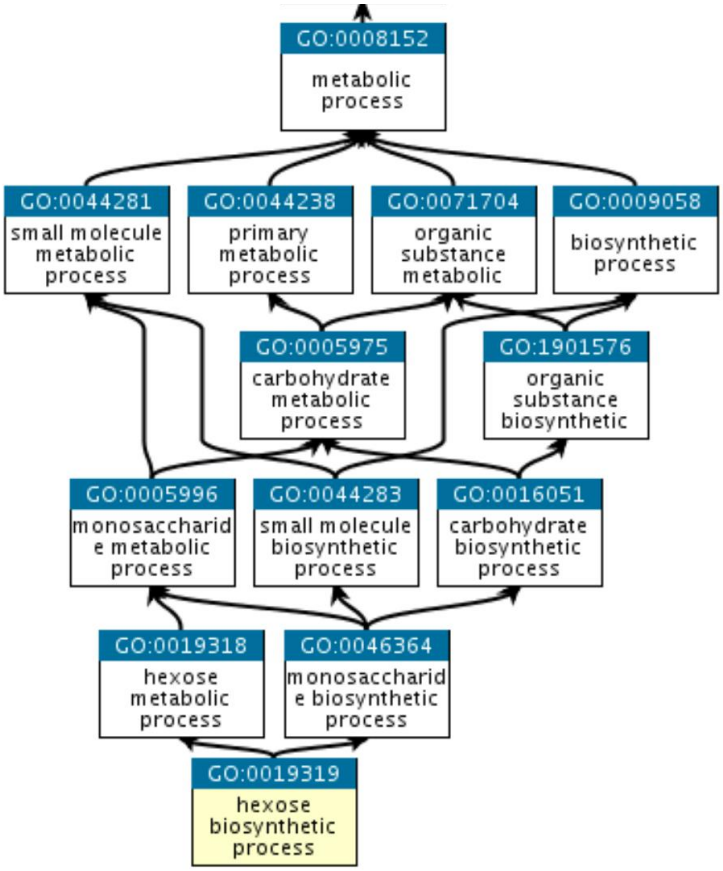
Κυτταρικό συστατικό (Cellular component): Με την έννοια αυτή, προσδιορίζεται η θέση μέσα στο κύτταρο στην οποία ένα γονιδιακό προϊόν είναι ενεργό και βοηθά στην κατανόηση της δομής των ευκαρυωτικών κυττάρων. Το κυτταρικό συστατικό περιλαμβάνει όρους όπως είναι το ριβόσωμα ή το σύμπλεγμα Golgi, διευκρινίζοντας το σημείο όπου μπορούν να βρεθούν πολλαπλά γονιδιακά προϊόντα [63].

Οι παραπάνω κατηγορίες-οντολογίες έχουν τρία χαρακτηριστικά [62]:

1. Είναι δυναμικές αφού αποτελούν ένα δίκτυο που συνεχώς αλλάζει όσο αυξάνονται οι πληροφορίες.
2. Είναι μοναδικές και ακριβείς ώστε να ενημερώνονται άμεσα όσες βάσεις είναι συνδεδεμένες μαζί τους.

3. Είναι “ελαστικές” για να μπορούν να εκπροσωπούν τις διαφορές στη βιολογία των διαφορετικών οργανισμών.

Σε έναν γράφο GO, ο κάθε όρος είναι ένας κόμβος ενώ η σχέση μεταξύ δύο όρων είναι μία ακμή που τους συνδέει. Στον γράφο επικρατεί μία ιεραρχία μεταξύ των όρων: υπάρχουν οι όροι “παιδί” και “γονέας” όπου ο πρώτος είναι πιο εξειδικευμένος από τον δεύτερο. Στην συγκεκριμένη ιεραρχία παρατηρείται μία ιδιαιτερότητα, ένας κόμβος-παιδί μπορεί να έχει πάνω από έναν κόμβο-γονέα. Ένα χαρακτηριστικό παράδειγμα είναι ο όρος της βιολογικής διεργασίας “βιοσυνθετική διαδικασία εξόζη” που έχει δύο γονείς: α) μεταβολική διαδικασία εξόζης και β) βιοσυνθετική διαδικασία μονοσακχαρίτη (Εικόνα 16). Αυτό φανερώνει πως η βιοσυνθετική διαδικασία είναι ένας υποτύπος μεταβολικής διαδικασίας και η εξόζη είναι ένας υποτύπος μονοσακχαρίτη.



Εικόνα 16. Ο όρος της βιολογική διεργασίας “βιοσυνθετική διαδικασία εξόζη” [63].

Η GO περιλαμβάνει και σχολιασμούς (GO annotations) οι οποίοι δημιουργούνται όταν ένα συγκεκριμένο γονιδιακό προϊόν συνδέεται με όρους της οντολογίας και αποδεικνύεται απο εγκεκριμένες δημοσιεύσεις. Η GO έχει πάνω από 7,000,000 σχολιασμούς σε γονίδια και γονιδιακά προϊόντα που προέρχονται από πάνω από 3,200 είδη. Το 10% των ειδών αυτών βασίζονται σε πειραματικά δεδομένα ερευνητικών έργων. Οι σχολιασμοί πραγματοποιούνται από ειδικούς βιοεπιστήμονες από όλο τον κόσμο που εξασφαλίζουν την αναγνώριση του σωστού γονιδίου και την επιλογή των κατάλληλων όρων που περιγράφουν τη βιολογία που υποστηρίζεται από τα πειραματικά ευρήματα [61].

4.2. KEGG analysis

4.2.1. Εισαγωγή στην έννοια

Η ***KEGG*** είναι μία συλλογή βάσεων δεδομένων που στοχεύει στην κατανόηση των λειτουργιών υψηλού επιπέδου και τα οφέλη του βιολογικού συστήματος, από το μοριακό επίπεδο που προκύπτει από την αλληλουχία του γονιδιώματος και από άλλες πειραματικές τεχνολογίες υψηλής απόδοσης. Η KEGG ασχολείται κυρίως με γονιδιώματα, βιολογικά μονοπάτια, ασθένειες, φάρμακα, και χημικές ουσίες. Η συλλογή αυτή, περιλαμβάνει 18 βάσεις δεδομένων χωρισμένες στις εξής κατηγορίες: Συστημικές Πληροφορίες, Γονιδιωματικές Πληροφορίες, Χημικές Πληροφορίες και Πληροφορίες Υγείας (Πίνακας 1).

ΚΑΤΗΓΟΡΙΕΣ	ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ	ΠΕΡΙΕΧΟΜΕΝΟ
Συστημικές Πληροφορίες	PATHWAY	χάρτες κυτταρικών μονοπατιών ή μονοπατιών οργανισμών
	BRITE	ιεραρχικές ταξινομήσεις λειτουργιών και οντοτήτων
	MODULE	ενότητες ή λειτουργικές μονάδες γονιδίων
Γονιδιωματικές Πληροφορίες	GENOME	πλήρη γονιδιώματα οργανισμών
	GENES	γονίδια και πρωτεΐνες
	ORTHOLOGY	μοριακές λειτουργίες που αναπαρίστανται με όρους λειτουργικών ορθολογιών
	SSDB	πληροφορίες σχετικά με τις ομοιότητες των αλληλουχιών μεταξύ όλων των γονιδίων που κωδικοποιούν πρωτεΐνες από την βάση GENES
Χημικές Πληροφορίες	COMPOUND	μεταβολίτες και άλλα μικρά μόρια
	GLYCAN	πειραματικά προσδιορισμένες δομές γλυκανών
	REACTION	βιοχημικές αντιδράσεις
	ENZYME	ένζυμα
Πληροφορίες Υγείας	NETWORK	κατηγορίες δικτύων που σχετίζονται με ασθένειες
	VARIANT	παραλλαγές ανθρώπινων γονιδίων
	DGROUPS	ομάδες φαρμάκων
	DISEASE	ανθρώπινες ασθένειες
	DRUG	φάρμακα
	MEDICUS	πηγή πληροφοριών σχετικά με την υγεία
	ENVIRON	μη επεξεργασμένα φάρμακα και ουσίες σχετικές με την υγεία

Πίνακας 1: Οι 18 βάσεις δεδομένων χωρισμένες σε κατηγορίες.

Επιπλέον η KEGG συνοδεύεται και από εργαλεία χαρτογράφησης, τα οποία επιτρέπουν την κατανόηση των λειτουργιών σε επίπεδο κυττάρου και οργανισμού από αλληλουχίες του γονιδιώματος και άλλα μοριακά σύνολα δεδομένων. Η γνώση που φιλοξενείται στην KEGG προέρχεται από πειραματικά δεδομένα που δημοσιεύονται στη βιβλιογραφία και αναπαρίστανται με όρους μοριακών δικτύων αλληλεπίδρασης και συγκεντρώθηκαν στη βάση δεδομένων KEGG PATHWAY. Σε ένα μοριακό δίκτυο, οι κόμβοι συνδέονται με λειτουργικά ορθόλογα γονίδια της βάσης δεδομένων KEGG ORTHOLOGY προκειμένου να επαναχρησιμοποιηθούν τα πειραματικά στοιχεία που παρατηρούνται σε συγκεκριμένους οργανισμούς, και στους υπόλοιπους.

Η κάθε βιολογική εγγραφή στην KEGG συνοδεύεται από ένα μοναδικό κωδικό αναγνώρισης ο οποίος ορίζεται από ένα πρόθεμα ανάλογα τη βάση και έναν πενταψήφιο αριθμό. Για παράδειγμα το πρόθεμα K αντιπροσωπεύει την βάση KEGG ORTHOLOGY (KO), οπότε το K04505 αφορά τον λειτουργικό ορθόλογο της Πρεσενιλίνης 1 στην βάση KO. Για τις βάσεις GENES, SSDB, ENZYME και VARIANT η μορφή του κωδικού αναγνώρισης είναι *βάση:καταχώριση*. Για παράδειγμα στη βάση GENES το hsa:5663 αντιστοιχεί στην ανθρώπινη Πρεσενιλίνης 1, ενώ στη βάση VARIANT το hsavar:5663v1 αντιστοιχεί σε μετάλλαξη της Πρεσενιλίνης 1. Στην KEGG θα πρέπει να τονιστεί η διαφορά μεταξύ των δεδομένων αναφοράς – κλάσεις (reference data – classes) και των δεδομένων παραλλαγής – περιπτώσεις (variation data - instances). Στο παραπάνω παράδειγμα το K04505 αποτελεί μία κλάση της Πρεσενιλίνης 1 ενώ το hsa:5663 είναι μία περίπτωση στον άνθρωπο [64].

4.2.2. Αναζήτηση στην KEGG

Η πρόσβαση στην KEGG γίνεται μέσω δύο τρόπων: μέσω flat-file format, για απλή αναζήτηση από την DBGET βάση δεδομένων ή με τη χρήση σχεσιακής βάσης δεδομένων (relational database – RDB) για πιο προηγμένα ερωτήματα σε επιλεγμένες βάσεις.

Το DBGET είναι ένα ολοκληρωμένο σύστημα ανάκτησης βάσεων δεδομένων για σημαντικές βιολογικές βάσεις δεδομένων. Αποτελεί τη βάση του συστήματος ανάκτησης για το GenomeNet και KEGG services. Βασίζεται στην flat-file μορφή των βάσεων μοριακής βιολογίας όπου η βάση αντιμετωπίζεται ως συλλογή εγγραφών. Με αυτό τον τρόπο κάθε εγγραφή/καταχώρηση, μπορεί να ανακτηθεί συνδυάζοντας το όνομα της βάσης (db) και το όνομα της εγγραφής/καταχώρησης (entry) ως εξής: *db:entry*

Η Σχεσιακή Βάση Δεδομένων (RDB) αφορά άμεσες SQL αναζητήσεις στα πεδία της KEGG τα οποία είναι προσβάσιμα μόνο για ανάγνωση [65].

4.2.3. KEGG BRITE

Το KEGG BRITE είναι μια συλλογή συστημάτων ιεραρχικής ταξινόμησης που καταγράφουν λειτουργικές ιεραρχίες διαφόρων βιολογικών αντικειμένων, ειδικά εκείνων που εκπροσωπούνται ως αντικείμενα KEGG. Τα συστήματα αυτά παρουσιάζονται ως ιεραρχικά αρχεία BRITE, γνωστά ως hierarchical text (htext) files περιλαμβάνοντας αρχεία πινάκων BRITE (BRITE table files) χρησιμοποιώντας πίνακες html. Τα αρχεία πινάκων BRITE επικεντρώνονται κυρίως σε γνωρίσματα πολλών στηλών παρά σε σχέσεις ιεραρχίας. Η συλλογή KEGG BRITE περιλαμβάνει πολλούς διαφορετικούς τύπους σχέσεων, πράγμα που το κάνει να διαφέρει από το KEGG PATHWAY το οποίο βασίζεται σε μοριακές συσχετίσεις. Οι διαφορετικοί τύποι σχέσεων που ενσωματώνει το KEGG BRITE είναι [66]:

- Γονίδια και πρωτεΐνες
- Χημικές ενώσεις και αντιδράσεις

- Φάρμακα
- Ασθένειες
- Οργανισμοί και κύτταρα

4.2.4. Χαρτογράφηση (Mapping)

Η χαρτογράφηση που γίνεται στο KEGG είναι η διαδικασία εκείνη όπου μόρια όπως είναι τα γονίδια, οι πρωτεΐνες, αλλά και μικρά μόρια χαρτογραφούνται σε δίκτυα μοριακών συσχετίσεων. Η χαρτογράφηση δεν αφορά απλά μια διαδικασία εμπλουτισμού (enrichment) των δεδομένων. Η βασική ιδέα ήταν να παράγονται αυτόματα μονοπάτια ειδικά για οργανισμούς (organism-specific pathways) από μια συλλογή λειτουργιών μεταξύ των χειροκίνητα σχολιασμένων δεδομένων γονιδιώματος και των επίσης χειροκίνητα δημιουργούμενων χαρτών μονοπατιών [67].

Υπάρχουν τρεις κύριες λειτουργίες χαρτογράφησης στην KEGG:

1. pathway mapping
2. brite mapping
3. module mapping

4.3. DAVID

Η βάση δεδομένων ***DAVID*** (Database for Annotation, Visualization and Integrated Discovery) είναι μία διαδικτυακή πηγή πληροφοριών Βιοπληροφορικής με σκοπό την διάθεση εργαλείων για την λειτουργική επεξήγηση μεγάλων γονιδιακών ή πρωτεϊνικών λιστών. Παρέχοντας ένα ολοκληρωμένο σύνολο εργαλείων λειτουργικού σχολιασμού για τη κατανόηση της βιολογικής σημασίας της εισαγόμενης λίστας από τον χρήστη, το DAVID μπορεί να εκτελέσει πολυάριθμες εργασίες, μερικές από τις οποίες είναι να:

- Ανακαλύψει εμπλουτισμένες ομάδες γονιδίων που σχετίζονται λειτουργικά.
- Ομαδοποιεί υπεράριθμους όρους σχολιασμού.
- Τονίζει αυτοτελείς δομικές περιοχές (domains) και μοτίβα μιας πρωτεΐνης.
- Καταγράφει πρωτεΐνες που αλληλεπιδρούν.
- Συνδέει τα γονίδια με τις σχετικές ασθένειες.

Ο χρήστης του DAVID συνήθως αρκεί να εισάγει μία λίστα με αναγνωριστικά γονιδίων (gene IDs), αλλά μπορεί να ανεβάσει πολλαπλές λίστες για πιο πολύπλοκες αναλύσεις. Επίσης, μπορούν να επαναφέρουν τις τιμές ορίσματος με πολλούς τρόπους, όπως για παράδειγμα με την επιλογή κατηγοριών και ειδών. Το API του DAVID που είναι βασισμένο στο URL του, δίνει την δυνατότητα στον χρήστη να έχει προγραμματιστική πρόσβαση στη βάση, αλλά λόγω των περιορισμών του URL, ο χρήστης μπορεί να εκτελεί “ελαφρού” τύπου εργασίες, οι οποίες έχουν περιορισμένο αριθμό γονιδίων στη λίστα (συνήθως περίπου στα 400 γονίδια) και δεν επιτρέπονται αλλαγές σε πολλές προεπιλεγμένες ρυθμίσεις. Σε περίπτωση που ο χρήστης θέλει να αναλύσει χιλιάδες λίστες γονιδίων, ορίζοντας ο ίδιος κάποιο όρισμα και να συγκρίνει τα αποτελέσματα του DAVID, εξαιτίας των παραπάνω περιορισμών κάτι τέτοιο δεν είναι εφικτό αλλά οι υπηρεσίες που είναι διαθέσιμες μπορούν να προγραμματιστούν για την ολοκλήρωση μίας τέτοιας διεργασίας εγκαίρως χωρίς την ανθρώπινη παρέμβαση.

Ο server πίσω από το DAVID έχει υλοποιηθεί σε Java χρησιμοποιώντας το Apache Axis2, με σκοπό να καλεί τις λειτουργικότητες του DAVID, όπως για παράδειγμα την λειτουργική ταξινόμηση γονιδίων (gene functional classification), την λειτουργική ομαδοποίηση των σχολιασμών (functional annotation

clustering), κ.α. Οι δεκάδες λειτουργίες που διαθέτει η ιστοσελίδα του DAVID χωρίζονται στις εξής κατηγορίες: Προσθήκη λίστας, Αναζήτηση, Επιλογή, Αναφορά [68].
Οι διεργασίες που εκτελούνται συχνότερα από τους χρήστες είναι:

- Αναφορά γραφήματος λειτουργικού σχολιασμού (Functional annotation chart report)
- Αναφορά πίνακα λειτουργικού σχολιασμού (Functional annotation table report)
- Αναφορά λειτουργικής ταξινόμησης γονιδίων (Gene functional classification report)
- Αναφορά ομαδοποίησης λειτουργικών σχολιασμών (Functional annotation clustering report)

4.4. g:Profiler

Το ***g:Profiler*** αποτελεί μία συλλογή εργαλείων που χρησιμοποιούνται κυρίως στις βιολογικές αναλύσεις, συγκεκριμένα για την εύρεση βιολογικών κατηγοριών εμπλουτισμένων σε λίστες γονιδίων, για τις μετατροπές μεταξύ αναγνωριστικών (IDs) γονιδίων και για αντιστοιχίσεις στα ορθόλογα τους.
Σκοπός του g:Profiler είναι η παροχή μιας αξιόπιστης υπηρεσίας, βασισμένη σε διαρκώς ενημερωμένα δεδομένα υψηλής ποιότητας, σε πολλούς τύπους στοιχείων, αναγνωριστικά και οργανισμούς. Κύρια πηγή δεδομένων του g:Profiler αποτελεί το Ensembl και υποστηρίζει 467 είδη συμπεριλαμβανομένων των σπονδυλωτών, φυτών, μυκήτων, εντόμων και των παρασίτων.

Οι λίστες γονιδίων που εισάγονται στα εργαλεία προκύπτουν από ένα ευρύ φάσμα πειραματικών πλατφορμών, όπου η κάθε μία πλατφόρμα έχει προεπιλεγμένους μοναδικούς τύπους αναγνωριστικών. Το g:Profiler, δεν έχει περιορισμό στο ποια αναγνωριστικά θα δεχτεί και αυτό διότι μπορεί να ανιχνεύει αυτόματα αναγνωριστικά από μία συλλογή εκατοντάδων διαφορετικών τύπων ακόμα και στη περίπτωση που είναι αναμειγμένοι μεταξύ τους. Οι μέθοδοι που χρησιμοποιούνται για την ανάλυση εμπλουτισμού ποικίλλουν μεταξύ των διαφορετικών εργαλείων. Το g:Profiler, παρέχει την πιο δημοφιλή προσέγγιση αντιπροσωπευτικής ανάλυσης, η οποία χρησιμοποιεί υπεργεωμετρικό έλεγχο για να μετρήσει τη σημασία του λειτουργικού όρου στη λίστα γονιδίων που έχει εισαχθεί. Στο g:Profiler υπάρχουν εργαλεία που παρέχουν μεθόδους που λαμβάνουν υπόψη πρόσθετες πληροφορίες κατάταξης των γονιδιακών λιστών ή χρησιμοποιούν προηγούμενες πληροφορίες από δίκτυα γονιδιακής ρύθμισης. Οι μέθοδοι αυτές συνοδεύονται από περιορισμούς και δεν υπάρχουν δεδομένα αναφοράς που να ωφελούν στην αξιολόγηση και την σύγκριση των διαφορετικών μεθόδων μεταξύ τους. Οι υπομονάδες του είναι: g:GOSt, g:Convert, g:Orth, g:SNPense.

Το g:GOSt αποτελεί πυλώνα στην εκτέλεση ανάλυσης λειτουργικού εμπλουτισμού στην εισαγόμενη λίστα γονιδίων. Χαρτογραφεί μια λίστα γονιδίων, που παρέχει ο χρήστης, σε γνωστές πηγές λειτουργικών πληροφοριών. Επιπλέον, ανιχνεύει τις στατιστικά σημαντικές εμπλουτισμένες βιολογικές διεργασίες, μονοπάτια, ρυθμιστικά μοτίβα και πρωτεϊνικά συμπλέγματα. Η βάση δεδομένων Ensembl είναι η βασική πηγή πληροφοριών σχετικά με τα γονίδια, τους τύπους των αναγνωριστικών, τους GO όρους και τις συσχετίσεις. Ο λειτουργικός εμπλουτισμός της γονιδιακής λίστας εκτιμάται χρησιμοποιώντας την καλά αποδεδειγμένη αθροιστική υπεργεωμετρική δοκιμή. Για μια λίστα γονιδίων, ελέγχονται πολλοί λειτουργικοί όροι. Για παράδειγμα, για μια λίστα ανθρώπινων γονιδίων ελέγχονται πάνω από 16.000 όροι βιολογικής διεργασίας GO. Για την ελαχιστοποίηση των ψευδώς θετικών ευρημάτων, το g:GOSt εκτελεί πολλαπλές δοκιμές διόρθωσης. Η εισαγωγή του χρήστη μπορεί να είναι είτε μία λίστα γονιδίων, είτε πολλαπλές λίστες ταυτόχρονα με τη χρήση ενός συμβόλου τύπου FASTA μπροστά από την κάθε μία λίστα.

Τα αποτελέσματα του g:GOSt, επισημαίνονται σε ένα νέο διάγραμμα Manhattan το οποίο συνοδεύεται από έναν εκτεταμένο και διαδραστικό πίνακα αποτελεσμάτων που παρέχει λεπτομέρειες για κάθε γονίδιο και όρο.

Το g:Convert εκτελεί μετατροπές μεταξύ διαφόρων γονιδίων, πρωτεϊνών, μικροσυστοιχιών και πολλών άλλων τύπων ονομάτων. Στο g:Profiler παρέχονται πάνω από 40 τύποι αναγνωριστικών, για παραπάνω από 60 είδη, τα οποία προέρχονται από το Ensembl Biomart και λαμβάνονται αντιστοιχίζοντας τα μέσω των αναγνωριστικών γονιδίων Ensembl (ENSG) ως αναφορά. Το g:Convert έχει την δυνατότητα, ως είσοδο, να δέχεται μικτούς τύπους αναγνωριστικών χωρίς να απαιτεί από τον χρήστη τον προκαθορισμό του τύπου του αναγνωριστικού, βοηθώντας έτσι την διαλειτουργικότητα και την σύνδεση εξωτερικών υπηρεσιών πέρα από το σύνολο εργαλείων του g:Profiler.

Το g:Orth χρησιμοποιώντας πληροφορίες ορθολογικής γονιδιακής χαρτογράφησης από τη βάση Ensembl, επιτρέπει στον χρήστη να ανακτά αυτόματα τα ορθολογικά γονίδια που αντιστοιχούν στη λίστα γονιδίων που εισάγει. Η χαρτογράφηση εκτελείται σε δύο στάδια: αρχικά μετατρέπει τα αναγνωριστικά των γονιδίων που εισήχθησαν σε αναγνωριστικά τύπου Ensembl ENSG και ύστερα ανακτά τις αντίστοιχες πληροφορίες ορθολογικών γονιδίων για είδη-στόχους. Η χρήση του g:Orth επικεντρώνεται στην μεταφορά γνώσης σχετικά με καλά μελετημένα μοντέλα οργανισμών σε λιγότερο μελετημένους. Η διεξαγωγή ανάλυσης εμπλουτισμού αφού έχει προηγηθεί ορθολογική χαρτογράφηση μπορεί να επιφέρει πιο κατανοητά αποτελέσματα από ό,τι όταν χρησιμοποιούνταν μόνο οι αρχικοί οργανισμοί.

Το g:SNPense δίνει την ευελιξία στον χρήστη να χαρτογραφίσει μία λίστα από SNP κωδικούς του ανθρώπου, της μορφής rs-κωδικός (π.χ.: rs7961894) σε ονόματα γονιδίων, να λάβει χρωμοσωμικές συντεταγμένες και διάφορα προβλεπόμενα αποτελέσματα τα οποία παρουσιάζονται κωδικοποιημένα χρωματικά. Η χαρτογράφηση είναι δυνατή για όσα διαφορετικά αποτελέσματα αλληλοεπικαλύπτονται με τουλάχιστον μία πρωτεΐνη που κωδικοποιεί ένα Ensembl γονίδιο. Όλα τα υπόλοιπα υποκείμενα δεδομένα λαμβάνονται από το Ensembl Variation Data.

Η πιο πρόσφατη έκδοση του g:Profiler έχει το πλεονέκτημα ότι η λίστα γονιδίων που εισάγεται μπορεί να αποτελείται από οποιαδήποτε κοινώς χρησιμοποιούμενα αναγνωριστικά γονιδίων ή πρωτεϊνών, ακόμα και αν είναι αναμειγμένα. Επιπλέον είναι δυνατή η ανάμειξη ονομάτων γονιδίων και αναγνωριστικών όπως για παράδειγμα τα αναγνωριστικά SNP ή ο αριθμός καταχώρησης μιας πρωτεΐνης, πράγμα που οφείλεται στο g:Convert. Επίσης, από τη τελευταία έκδοση του g:GOSt, το 2019, επιτρέπεται η εισαγωγή χρωμοσωμικών διαστημάτων σε μορφή αρχείου BED (Browser Extensible Data) κάνοντας δυνατή την ενσωμάτωση του σε προγράμματα που βασίζονται σε σχολιασμούς όπως το UCSC Genome Browser καθώς τέτοιου τύπου προγράμματα εξάγουν αρχεία BED.

Με βάση το πλήθος των λιστών που εισάγει ο χρήστης, αυτομάτως γίνεται αξιολόγηση του τύπου εισαγωγής και ξεκινάει η ανάλυση εμπλουτισμού για την κάθε λίστα. Τα αποτελέσματα που προκύπτουν από την ανάλυση εξαρτώνται από το πλήθος των λιστών· εάν ο χρήστης εισήγαγε μία μόνο λίστα τότε ο πίνακας αποτελεσμάτων περιλαμβάνει αποδεικτικούς κωδικούς για κάθε ζεύγος όρος-γονίδιο, ενώ σε αντίθετη περίπτωση όπου έχουν εισαχθεί πάνω από μία λίστες ο πίνακας αποτελεσμάτων εστιάζει στη σύγκριση των P-values αυτών και για την κάθε μία παράγεται από ένα διάγραμμα Manhattan [69].

4.5. Reactome

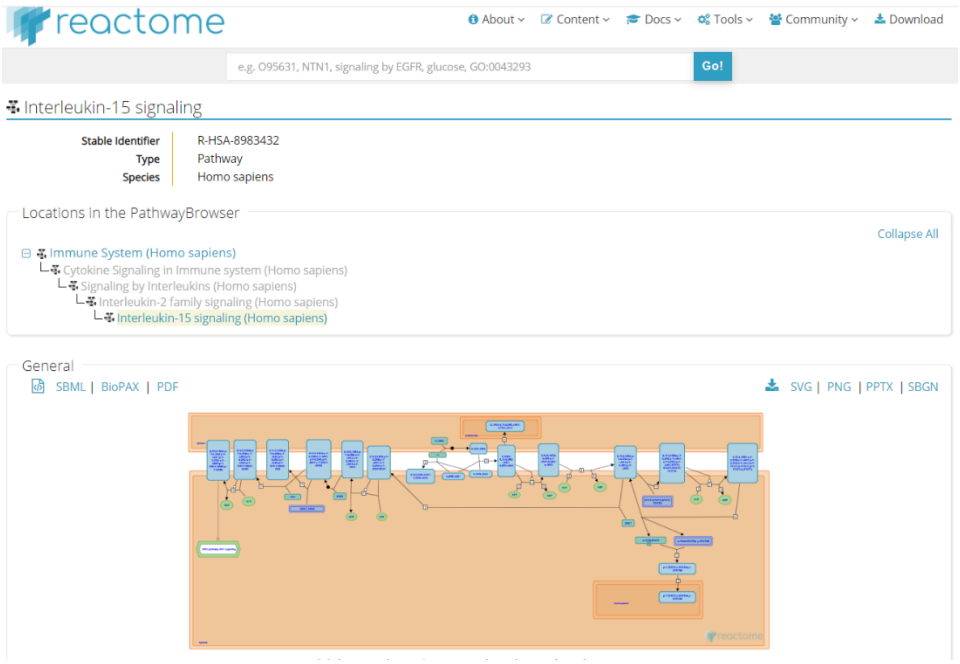
Η βάση δεδομένων ***Reactome*** περιέχει μοριακές πληροφορίες σχετικά με τη μεταγωγή σήματος, μεταφοράς, αντιγραφής DNA, μεταβολισμού και άλλων κυτταρικών διεργασιών. Η βάση είναι σε μορφή ενός ταξινομημένου δικτύου μοριακών μετασχηματισμών σε ένα ενιαίο μοντέλο δεδομένων, μια εκτεταμένη έκδοση ενός κλασικού μεταβολικού χάρτη. Η βάση συνδέει μεθοδικά, ανθρώπινες πρωτεΐνες στις μοριακές λειτουργίες τους δημιουργώντας έτσι έναν πόρο που έχει δύο τρόπους λειτουργικότητας:

1) Ως αρχείο βιολογικών διεργασιών.
2) Εργαλείο εύρεσης λειτουργικών σχέσεων δεδομένων όπως είναι τα προφίλ γονιδιακής έκφρασης.

Στην βάση υπάρχει μία επιπλέον κατηγορία φαρμάκων με σκοπό τον βέλτιστο σχολιασμό των ανθρώπινων ασθενειών. Για την καλύτερη οπτική παρουσίαση των περιεχομένων της βάσης, υπάρχει εργαλείο το οποίο καθορίζει αυτομάτως τα συστατικά των μεμονωμένων αντιδράσεων με πολλαπλές επιλογές για τη λήψη των διαγραμμάτων αντίδρασης και των σχετικών δεδομένων, αλλά και μια νέα παρουσίαση της ιεραρχίας των γεγονότων που βοηθάει οπτικά στην ερμηνεία των αποτελεσμάτων της ανάλυσης μονοπατιών.

Αναλύοντας τη ζωή σε κυτταρικό επίπεδο, παρατηρείται πως είναι ένα δίκτυο μοριακών αντιδράσεων που επιτρέπει διάφορες βιολογικές διαδικασίες. Πολλές διαδικτυακές πηγές περιέχουν πτυχές των παραπάνω πληροφοριών είτε σε επίπεδο μεμονωμένων αντιδράσεων όπως το Rhea, είτε σε επίπεδο αλληλουχιών αντίδρασης που συναντώνται σε διάφορους τομείς της βιολογίας όπως το KEGG ή το PANTHER. Η βάση Reactome ξεχωρίζει καθώς ο σχολιασμός της επικεντρώνεται μόνο στο ανθρώπινο είδος και εφαρμόζει ένα συμπαγές μοντέλο δεδομένων σε όλους τους τομείς της Βιολογίας. Οι διαδικασίες που υπάρχουν στη βάση, περιγράφονται με μοριακή λεπτομέρεια ώστε να παραχθεί ένα ταξινομημένο δίκτυο μοριακών μετασχηματισμών ως μια εκτεταμένη έκδοση ενός μεταβολικού χάρτη. Επίσης, η βάση Reactome συνδέει τις πρωτεΐνες του ανθρώπινου οργανισμού με τις μοριακές λειτουργίες τους, με αποτέλεσμα την δημιουργία ενός πόρου ο οποίος θα έχει τον ρόλο ενός αρχείου βιολογικών διεργασιών αλλά και εργαλείου εντοπισμού νέων λειτουργικών σχέσεων σε δεδομένα όπως για παράδειγμα σε μελέτες έκφρασης γονιδίων.

Η βάση δεδομένων Reactome περιέχει 10,867 ανθρώπινες πρωτεΐνες - κωδικοποιημένα γονίδια, που αντιστοιχούν περίπου στο 53% των 20,454 προβλέψιμων ανθρώπινων πρωτεϊνών - κωδικοποιημένων γονιδίων. Οι εγγραφές της βάσης υποστηρίζουν τον σχολιασμό 25,849 πρωτεϊνών συγκεκριμένων μορφών που είναι ευδιάκριτες από συν- και μετα-μεταφραστικούς μετασχηματισμούς και υποκυτταρικούς εντοπισμούς, και που λειτουργούν με πάνω από 1,800 μικρά μόρια που προκύπτουν φυσικά ως υποστρώματα, καταλύτες και ρυθμιστές σε πάνω από 11,600 αντιδράσεις, σχολιασμένες σύμφωνα με τα δεδομένα 30,398 βιβλιογραφικών αναφορών. Οι συγκεκριμένες αντιδράσεις χωρίζονται σε 1,803 μονοπάτια όπως για παράδειγμα η σηματοδότηση ιντερλευκίνης-15 (Εικόνα 17), ομαδοποιημένα σε 26 υπερ-μονοπάτια που περιγράφουν φυσιολογικές λειτουργίες των κυττάρων όπως για παράδειγμα το ανοσοποιητικό σύστημα και ο μεταβολισμός. Υπάρχει ένα ακόμη υπερ-μονοπάτι αφιερωμένο στις ασθένειες το οποίο έχει συγκεντρωμένες σε ομάδες 484 σχολιασμούς αντίστοιχων ασθενειών αυτών των φυσιολογικών κυτταρικών διεργασιών [70].



Εικόνα 17. Το μονοπάτι της σηματοδότησης της ιντερλευκίνης-15.

4.6. PANTHER

Η βάση ***PANTHER*** (Protein Analysis Through Evolutionary Relationships) αποτελεί πηγή δεδομένων σχετικά με την εξελικτική και λειτουργική ταξινόμηση των γονιδίων που κωδικοποιούν πρωτεΐνες, διάφορων οργανισμών.

Οι πρωτεΐνες ομαδοποιούνται σε δύο κύριες κατηγορίες:

1. *Εξελικτική ομαδοποίηση/ταξινόμηση* (Κατηγορίες πρωτεϊνών, Οικογένειες πρωτεϊνών, Υποοικογένεια) η οποία αντιστοιχεί στη “φυσική ταξινόμηση” των πρωτεϊνών σύμφωνα με την εξελικτική ιστορία τους. Η εξελικτική ταξινόμηση στηρίζεται σε πάνω από 15,000 φυλογενετικά δέντρα.
2. *Λειτουργική ομαδοποίηση/ταξινόμηση* (Μονοπάτια και όροι Οντολογίας Γονιδίων) όπου οι πρωτεΐνες διαχωρίζονται με βάση τις ίδιες τους τις λειτουργίες και όχι με εκείνες των οικογενειών τους. Οι πρωτεϊνικές λειτουργίες μπορούν να ταξινομηθούν είτε με βάση τη γονιδιακή οντολογία και συγκεκριμένα την μοριακή λειτουργία, τα κυτταρικά συστατικά και την βιολογική διαδικασία. Επίσης μπορούν να ταξινομηθούν και σύμφωνα με μονοπάτια, όπως είναι τα μονοπάτια σηματοδότησης και τα μεταβολικά μονοπάτια.

Οι λειτουργίες σχολιάζονται με δύο τρόπους:

1. Σχολιάζοντας ομάδες σχετικών πρωτεϊνών σε φυλογενετικό δέντρο, τα μέλη των οποίων πιστεύεται ότι προέρχονται από τον ίδιο πρόγονο, με τη βοήθεια δέντρων. Το αποτέλεσμα είναι ο σχολιασμός όλων των πρωτεϊνών που ανήκουν στην ομάδα. Οι σχολιασμοί περιέχουν όρους Οντολογίας Γονιδίων από το PANTHER GO-slim που είναι ένα υποσύνολο ολόκληρης της

Οντολογίας Γονιδίων (GO), αλλά και όρους από το PANTHER Pathway. Επίσης, στο PANTHER Pathway δημιουργείται με τη βοήθεια του CellDesigner το μοντέλο μονοπατιού.

2. Σχολιάζοντας μεμονωμένες πρωτεΐνες και χωρίζοντας ‘τες σε λειτουργικές κλάσεις είτε μέσω του GO, είτε μέσω της Reactome.

Το PANTHER παρέχει αναλυτικά δεδομένα για τις πρωτεϊνικές οικογένειες:

- Φυλογενετικό δέντρο.
- Σχολιασμοί από το GO για τους εσωτερικούς κόμβους του φυλογενετικού δέντρου.
- Hidden Markov models (HMMs) για κάθε οικογένεια ή υποοικογένεια.
- Στοίχιση πολλαπλών ακολουθιών για όλα τα μέλη της οικογένειας.
- Ορθόλογα, παράλογα και ξενόλογα (τύπος ορθόλογου όπου οι ομόλογες αλληλουχίες βρίσκονται σε διαφορετικά είδη λόγω της οριζόντιας μεταφοράς γονιδίων.)

Επιπλέον, στο PANTHER έχει προστεθεί η βάση PEREGRINE (database of gene-enhancer links), ώστε ο χρήστης να μπορεί να έχει πρόσβαση σε λίστες ενισχυτών που έχουν συσχετιστεί με την έκφραση κάθε ανθρώπινου γονιδίου που κωδικοποιεί πρωτεΐνη στο PANTHER.
Μέσω του PANTHER παρέχονται στον χρήστη τριών ειδών εργαλεία [71]:

1. Εργαλεία ταξινόμησης πρωτεϊνών: παρέχοντας δύο εργαλεία ταξινόμησης ο χρήστης του PANTHER μπορεί να τα αξιοποιήσει στις πρωτεϊνικές ακολουθίες που επιθυμεί. Τα δύο αυτά εργαλεία ταξινόμησης είναι το PANTHER HMMs και το TreeGrafter.
2. Εργαλεία ανάλυσης γονιδιακής λίστας: δίνεται η δυνατότητα στον χρήστη να ανεβάσει γονιδιακές ή πρωτεϊνικές λίστες και να υλοποιήσει στατιστικές δοκιμές ώστε να εντοπίσει εμπλουτισμένες λειτουργικές τάξεις στις λίστες αυτές.
3. Ένα εργαλείο ανάλυσης παραλλαγών κωδικοποίησης μιας πρωτεΐνης.

4.7. Webgestalt

Το ***WebGestalt*** (WEB-based GEne SeT AnaLysis Toolkit), είναι ένα εργαλείο για την ερμηνεία και ανάλυση των γονιδιακών λιστών προερχόμενες από -ωμικές (-omics) μελέτες μεγάλης κλίμακας. Το εργαλείο υποστηρίζει 342 αναγνωριστικά γονιδίων από 12 οργανισμούς και 55,175 λειτουργικές κατηγορίες. Επιπλέον, περιέχει βάσεις δεδομένων όπως ένα υποσύνολο του WikiPathways που σχετίζεται με τον καρκίνο, ενότητες δικτύου συν-έκφρασης που προκύπτουν από δεδομένα πρωτεομικής του καρκίνου (cancer proteomics) από το CPTAC (Clinical Proteomic Tumor Analysis Consortium), τη βάση δεδομένων πρωτεϊνικών συμπλεγμάτων CORUM, τη βάση δεδομένων φαινοτύπων ασθενειών OMIM και γονίδια στόχου κινάσης από το PhosphoSitePlus.
Λόγω της αυξανόμενης ανάγκης για τον προσδιορισμό σημαντικών κινασών από δεδομένα φωσφοπρωτεωμικής, δημιουργήθηκε και μία ενότητα ανάλυσης σημείων φωσφορυλίωσης (phosphorylation sites - phosphosites). Η φωσφορυλίωση είναι μία εκ των πιο μελετημένων μετα-μεταφραστικών τροποποιήσεων και είναι ιδιαίτερα σημαντική στην πλειοψηφία των κυτταρικών διεργασιών [72].

Με σκοπό την διεύρυνση του κοινού στο οποίο απευθύνεται το WebGestalt, υπάρχει ένα πακέτο στην R (R package) ονόματι WebGestaltR που και σε αυτή τη περίπτωση όπως και στη διαδικτυακή έκδοση του εργαλείου, υποστηρίζονται οι τρεις μέθοδοι ανάλυσης:

- Over Representation Analysis (ORA) - καλύπτει την ανάγκη για λειτουργική ανάλυση δεδομένων των μικροσυστοιχιών γονιδίων συν-έκφρασης. Αξιολογεί στατιστικά το κλάσμα των γονιδίων σε ένα μονοπάτι που βρίσκονται μεταξύ μίας ομάδας γονιδίων που εμφανίζουν αλλαγές στην έκφραση. Βιβλιογραφικά αναφέρεται και ως “μέθοδος πίνακα 2 × 2” (“2×2 table method”) [73].
- Gene Set Enrichment Analysis (GSEA) - χρησιμοποιείται για την ερμηνεία δεδομένων έκφρασης γονιδίων δίνοντας σημασία κυρίως σε ομάδες γονιδίων όπου τα γονίδια μοιράζονται κοινή βιολογική λειτουργία, χρωμοσωμική θέση ή ρύθμιση [74].
- Network Topology-based Analysis (NTA) [75]

Τα αποτελέσματα των μεθόδων ORA και GSEA χωρίζονται σε δύο βασικές ενότητες, στα συνοπτικά και στις εμπλουτισμένα αποτελέσματα.

Στην διεπαφή του WebGestalt ο χρήστης μπορεί να δοκιμάσει έτοιμα παραδείγματα, μπορεί να βρει πληροφορίες σχετικά με την χρήση του εργαλείου αλλά και ένα φόρουμ με απόψεις και ερωτήσεις διαφόρων χρηστών.

5. Αλγόριθμοι ομαδοποίησης δεδομένων

5.1. Εισαγωγή

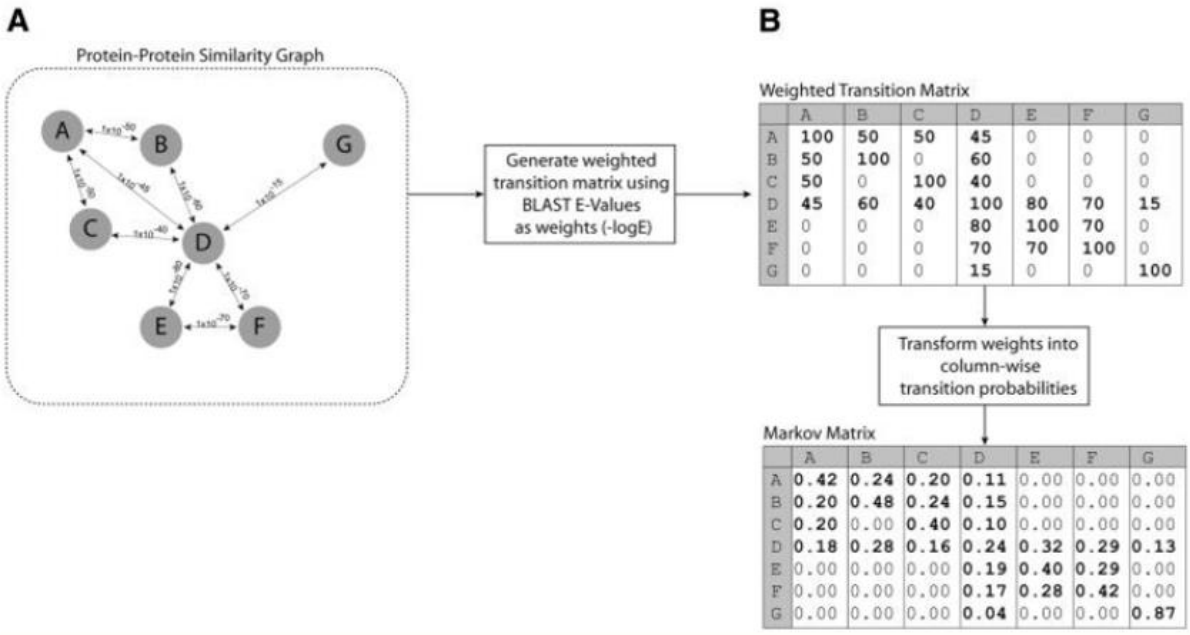
Οι αλγόριθμοι ομαδοποίησης/συσταδοποίησης συμβάλλουν αποτελεσματικά στην ανάλυση των βιολογικών δικτύων καθώς δίνουν την δυνατότητα της ανάδειξης των λειτουργικών ενοτήτων και πληροφοριών σχετικά με την κυτταρική οργάνωση. Αν και η πλειοψηφία των αλγορίθμων είναι αποδοτικοί στην ομαδοποίηση των βιολογικών δικτύων μέτριου μεγέθους, στα μεγάλα δίκτυα δυσκολεύονται αφού είτε είναι πολύ αργοί είτε αποτυγχάνουν να φέρουν αποτέλεσμα. Τα τελευταία χρόνια, γίνονται ολοένα και περισσότερες προσπάθειες ανάλυσης βιολογικών δικτύων για να απόκτηση πληροφοριών για την κυτταρική λειτουργία και οργάνωση και οι αλγόριθμοι ομαδοποίησης αποτελούν σημαντικό και το πιο δημοφιλές εργαλείο για αυτό [76].

Υπάρχουν διάφοροι τύποι αλγορίθμων, ανάλογα με τις απαιτήσεις. Οι δύο κύριες κατηγορίες αλγορίθμων ομαδοποίησης είναι οι ιεραρχικοί (hierarchical) αλγόριθμοι και οι διαιρετικοί (partitional) αλγόριθμοι ή αλλιώς αλγόριθμοι διαμέρισης. Οι αλγόριθμοι ιεραρχικής ομαδοποίησης δημιουργούν μια ιεραρχία κατατμήσεων, που αντιπροσωπεύονται από δενδρογράμματα στα οποία κάθε μέρος τους είναι εμφωλευμένο εντός του διαμερίσματος του επόμενου επιπέδου της ιεραρχίας. Οι αλγόριθμοι διαμέρισης, δημιουργούν ένα μόνο διαμέρισμα δεδομένων με ένα συγκεκριμένο ή εκτιμώμενο νούμερο μη αλληλεπικαλυπτόμενων ομάδων, με σκοπό την ανάκτηση φυσικών ομάδων που βρίσκονται στα δεδομένα [77]. Παρακάτω παρουσιάζονται μερικοί από τους πιο γνωστούς αλγορίθμους ομαδοποίησης/συσταδοποίησης, που χρησιμοποιήθηκαν στα δίκτυα της διπλωματικής.

5.2. Αλγόριθμος MCL

5.2.1. Εισαγωγή

Ο αλγόριθμος ***MCL*** αρχικά χρησιμοποιήθηκε στον τομέα της υπολογιστικής ομαδοποίησης γράφων (computational graph clustering) δίνοντας του έτσι τη δυνατότητα να χρησιμοποιηθεί και στην ομαδοποίηση βιολογικών ακολουθιών. Ο αλγόριθμος, έχει σχεδιαστεί κυρίως για απλούς και σταθμισμένους γράφους και εντοπίζει με μαθηματικούς υπολογισμούς την δομή των ομάδων εντός των γράφων. Συγκεκριμένα, υπολογίζονται ντετερμινιστικά οι πιθανότητες των τυχαίων μετακινήσεων (random walks) σε έναν γράφο ομοιότητας αλληλουχιών και χρησιμοποιεί δύο τελεστές που μετατρέπουν το ένα σετ πιθανοτήτων σε ένα άλλο. Επίσης, στηρίζεται και στον Μαρκοβιανό πίνακα, ο οποίος αντιπροσωπεύει τα μαθηματικά των τυχαίων μετακινήσεων σε έναν γράφο. Η ροή που ακολουθεί ο αλγόριθμος απαιτεί την συσχέτιση των βρόγχων με κάθε κόμβο του αρχικού γράφου. Σχετικά με τις επιλεγμένες τιμές στα βάρη των σταθμισμένων γράφων, έχει φανεί εμπειρικά πως οι “ουδέτερες” τιμές λειτουργούν καλύτερα στους υπολογισμούς του αλγορίθμου. Εφόσον το βάρος για κάθε κόμβο είναι “ουδέτερο”, συμβάλλει στο να μην αλλάξει η τιμή του σε περίπτωση που εφαρμοστεί ο τελεστής πληθωρισμού στη στήλη του Μαρκοβιανού πίνακα που σχετίζεται με τον συγκεκριμένο κόμβο. Υπάρχει η δυνατότητα επιλογής μεγαλύτερων τιμών για τα βάρη αυξάνοντας κατά συνέπεια λεπτομερώς το κοκκιώδες περιεχόμενο (cluster granularity) των ομάδων/συστάδων , αλλά γενικότερα ο αλγόριθμος δεν είναι αρκετά “ευαίσθητος” σε αλλαγές που γίνονται στα βάρη των βρόγχων (Εικόνα 18).



	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	80	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

	A	B	C	D	E	F	G
A	0.42	0.24	0.20	0.11	0.00	0.00	0.00
B	0.20	0.48	0.24	0.15	0.00	0.00	0.00
C	0.20	0.00	0.40	0.10	0.00	0.00	0.00
D	0.18	0.28	0.16	0.24	0.32	0.29	0.13
E	0.00	0.00	0.00	0.19	0.40	0.29	0.00
F	0.00	0.00	0.00	0.17	0.28	0.42	0.00
G	0.00	0.00	0.00	0.04	0.00	0.00	0.87

Εικόνα 18. Α)Γράφος ομοιότητας επτά πρωτεϊνών. Οι ομοιότητες προκύπτουν από τα E-values του BLASTp. **Β) Σταθμισμένος πίνακας μεταβάσεων και οι συσχετισμένες στοχαστικές στήλες του Μαρκοβιανού πίνακα** για τις επτά πρωτεΐνες του γράφου (Α) [78].

5.2.2. Λειτουργία αλγορίθμου MCL

Ο αλγόριθμος MCL έχει απλή διατύπωση και ακολουθεί μία bootstrapping ροή που πρόκειται για μία διαδικασία η οποία ξεκινάει από μόνη της χωρίς να χρειάζεται εξωτερική παρέμβαση για να συνεχίσει ή να αναπτυχθεί. Επιπρόσθετα πλεονεκτήματα του αλγορίθμου είναι ότι δεν “παρασύρεται” από τις ακμές που συνδέουν διαφορετικές συστάδες, είναι πολύ γρήγορος και επεκτάσιμος και τα μαθηματικά του αλγορίθμου δείχνουν πως υπάρχει μία εγγενής σχέση μεταξύ της διαδικασίας που προσομοιώνει και της δομής της συστάδας του γράφου εισόδου. Ο MCL αν και βασίζεται στις ομοιότητες μεταξύ ζευγών, μέσω επέκτασης επανασυνδυάζει τις ομοιότητες αυτές και έτσι επηρεάζεται από ομοιότητες σε επίπεδο συνόλων/σετ.

Ο αλγόριθμος καθορίζεται από τον υπολογισμό του γράφου των τυχαίων μετακινήσεων ενός γράφου εισόδου ο οποίος αντιπροσωπεύεται από έναν Μαρκοβιανό πίνακα. Έπειτα, εναλλάσσει τον τελεστή επέκτασης που τετραγωνίζει έναν πίνακα χρησιμοποιώντας το σύνηθες γινόμενο του πίνακα με τον τελεστή πληθωρισμού. Ο πληθωρισμός πραγματοποιείται αυξάνοντας κάθε είσοδο του πίνακα σε μια δεδομένη δύναμη και αναδημιουργώντας τον πίνακα ώστε να γίνει και πάλι στοχαστικός. Όλη η διαδικασία εναλλαγής συνεχίζεται μέχρις ότου επιτευχθεί μια κατάσταση ισορροπίας με τη μορφή του “διπλού αυτοδύναμου πίνακα”.

Τα βιολογικά δίκτυα όπου γίνεται εφαρμογή του MCL, παρουσιάζονται ως εξής:

- Κόμβοι - πρωτεΐνες που θα θέλαμε να αναθέσουμε σε οικογένειες
- Ακμές - ομοιότητες μεταξύ πρωτεϊνών
- Βάρη - οι τιμές τους αποδίδονται με βάση το score ομοιότητας αλληλουχίας που προκύπτει από έναν αλγόριθμο όπως είναι το BLAST.

Η διαδικασία ξεκινάει με την δημιουργία ενός Μαρκοβιανού πίνακα που θα παρουσιάζει τις πιθανότητες μετάβασης από τον έναν κόμβο-πρωτεΐνη σε άλλον όπου έχει εντοπιστεί ομοιότητα. Η κάθε στήλη του πίνακα αφορά τον κάθε κόμβο-πρωτεΐνη, κάθε είσοδος σε κάθε στήλη αντιπροσωπεύει την ομοιότητα μεταξύ πρωτεϊνών, οι τιμές στις διαγωνίους ορίζονται αυθαίρετα με μια "ουδέτερη" τιμή. Ο Μαρκοβιανός αυτός πίνακας, παρέχεται στον αλγόριθμο MCL, του οποίου η αρχική επέκταση προσομοιώνει τυχαίες μετακινήσεις που δίνουν τη δυνατότητα καταγραφής της ροής του γράφου. Σημεία με μεγάλη ροή υποδηλώνουν πως πολλές τυχαίες μετακινήσεις περνούν από αυτά. Επαναλαμβάνοντας τη διαδικασία επέκτασης και πληθωρισμού, ο αλγόριθμος προάγει τη ροή του γράφου σε σημεία όπου αυτή είναι υψηλή και αφαιρεί τη ροή του γράφου όπου αυτή είναι αδύναμη. Το πέρας της διαδικασίας φτάνει όταν έχει επιτευχθεί ισορροπία, δηλαδή περαιτέρω κύκλοι επέκτασης και πληθωρισμού δεν επηρεάζουν πια τον πίνακα.

Βιολογικά, τα μέλη μιας πρωτεϊνικής οικογένειας έχουν περισσότερες ομοιότητες μεταξύ τους από ότι με μέλη άλλων οικογενειών, γεγονός που έχει αποδειχθεί και γραφικά μέσω του αλγορίθμου οπτικοποίησης Bio-Layout. Λόγω του συγκεκριμένου χαρακτηριστικού των βιολογικών γράφων η ροή εντός των πρωτεϊνικών οικογενειών είναι ισχυρή, δηλαδή μια τυχαία μετακίνηση που ξεκινά από κάποια πρωτεΐνη σε μια οικογένεια είναι πιθανότερο να παραμείνει εντός αυτής της οικογένειας παρά να περάσει σε μια άλλη. Η ροή μεταξύ των πρωτεϊνικών οικογενειών θα είναι πιο αδύναμη από τη ροή εντός μιας οικογένειας καθώς υπάρχουν ελάχιστα μονοπάτια που να διασχίζουν δύο διαφορετικές πρωτεϊνικές οικογένειες. Τα μονοπάτια εντός μίας οικογένειας αντιπροσωπεύουν είτε σχέσεις ομοιότητας αλληλουχιών που οφείλονται σε πρωτεΐνες με πολλαπλές αυτοτελείς δομικές περιοχές (multi-domain proteins) ή σε ψευδείς θετικές ομοιότητες.

5.2.3. Απόδοση αλγορίθμου MCL

Οι ιδιότητες των γράφων βιολογικής ομοιότητας, κάνουν τους γράφους αυτούς ιδανικούς για τον αλγόριθμο MCL αφού η διαδικασία που ακολουθεί και περιγράφηκε παραπάνω, επιτρέπει στις πρωτεϊνικές οικογένειες που είναι "κρυμμένες" στο γράφημα να "εμφανιστούν" με σταδιακή "αποσυναρμολόγηση" του γράφου στα βασικά του συστατικά που εντοπίζονται από τη στοχαστική ροή.[78]

5.3. Αλγόριθμος SPICi

5.3.1. Εισαγωγή

Για τα βιολογικά δίκτυα μεσαίου μεγέθους υπάρχουν πολλοί αλγόριθμοι ομαδοποίησης οι οποίοι αποδίδουν ικανοποιητικά, αλλά στα βιολογικά δίκτυα μεγαλύτερου μεγέθους είτε είναι αργοί είτε αποτυγχάνουν. Τέτοια βιολογικά δίκτυα μεγάλου μεγέθους συνήθως δυσκολεύουν τις υπάρχουσες αλγοριθμικές προσεγγίσεις ομαδοποίησης. Ο αλγόριθμος ***SPICi*** (Speed and Performance In Clustering) είναι αποτελεσματικός στην αντιμετώπιση τέτοιου τύπου δικτύων. Ο αλγόριθμος εντοπίζει σημεία του γράφου με υψηλή συνδεσιμότητα με βάση την τοπική τους πυκνότητα. Δημιουργεί όλο και περισσότερες συστάδες ξεκινώντας από τους τοπικούς κόμβους (κόμβος-σπόρος) του γράφου με υψηλό βαθμό και ύστερα προσθέτει κατάλληλους γειτονικούς κόμβους ώστε να διατηρεί σταθερή την πυκνότητα των συστάδων.

5.3.2. Λειτουργία αλγορίθμου SPICi

Ο SPICi στηρίζεται σε ευρετική προσέγγιση για την παραγωγή των ομάδων και εγγυάται χρόνο εκτέλεσης $O(V\ logV + E)$, όπου το Vεκπροσωπεί τους κόμβους και το Eτις ακμές. Η ευρετική προσέγγιση βοηθάει στην "άπληστη" δημιουργία των ομάδων με απώτερο σκοπό την παραγωγή αποσυνδεδεμένων πυκνών υπογράφων. Σε έναν γράφο που μοντελοποιείται ως εξής: $G = (V, E)$ και με βάρη ακμών $0 < w_{u,v} \leq 1$ σε περίπτωση που δύο κόμβοι δεν έχουν ακμή μεταξύ τους τότε $w_{u,v} = 0$. Ως βαθμό σε ένα δίκτυο με βάρη, για κάθε κόμβο u ορίζεται το άθροισμα των βαρών των ακμών του: $d_w(u) = \sum_{v:(u,v)\in E} w_{u,v}$

Επίσης για ένα υποσύνολο κόμβων Sορίζεται η πυκνότητα του ως εξής: $density(S) = \frac{\sum_{u,v\in S} w_{u,v}}{|S| \cdot (|S|-1)/2}$

Ακόμη, για κάθε κόμβο u και ένα υποσύνολο κόμβων S, το support του u από το S ορίζεται όπως παρακάτω: $support\ (u, S) = \sum_{v\in S} w_{u,v}$

Ένας κόμβος ο οποίος δεν έχει συμπεριληφθεί σε μία ομάδα, εάν έχει support τόσο υψηλό ώστε η τιμή της πυκνότητας της ομάδας παραμένει υψηλότερη από το όριο που καθορίζεται από τον χρήστη, τότε συμπεριλαμβάνεται στην ομάδα. Σε διαφορετική περίπτωση, η ομάδα έχει δημιουργηθεί και οι κόμβοι της αφαιρούνται από το αρχικό δίκτυο. Με βάση τα παραπάνω, ο αλγόριθμος έχει δύο παραμέτρους, μία για το κατώφλι του support T_s και το κατώφλι της πυκνότητας T_d.

Όπως αναφέρθηκε παραπάνω, ο αλγόριθμος SPICi βρίσκει τους αρχικούς κόμβους αφού πρώτα εντοπίσει τον κόμβο με το μεγαλύτερο βαθμό, έστω ο κόμβος-σπόρος u. Ύστερα, οι γειτονικοί κόμβοι του u, διαιρούνται σε 5 διαστήματα με βάση τα βαρη των κορυφών τους: (0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8] και (0.8, 1]. Η αναζήτηση γίνεται από το τελευταίο διάστημα προς το πρώτο και σε περίπτωση που κατά την αναζήτηση το εκάστοτε διάστημα δεν είναι άδειο τότε επιλέγεται ένας δεύτερος κόμβος-σπόρος v ο οποίος έχει το υψηλότερο βαθμό του συγκεκριμένου διαστήματος με αποτέλεσμα να δημιουργηθεί μία ακμή-σπόρος (u, v). Ο συγκεκριμένος ευρετικός τρόπος επιλογής των αρχικών κόμβων βασίζεται σε δύο παρατηρήσεις σχετικά με τα λειτουργικά δίκτυα:

1. Υπάρχει συσχέτιση μεταξύ του βαθμού ενός κόμβου και ενός μέτρου που αντιστοιχεί στον ολικό λειτουργικό εμπλουτισμό μεταξύ των αλληλεπιδρώντων πρωτεϊνών του δικτύου. Αυτό σημαίνει πως οι κόμβοι με υψηλό βαθμό έχουν μεγάλη σημασία για την έναρξη τοπικών αναζητήσεων σε μονάδες λειτουργικών δικτύων.
2. Η πιθανότητα δύο κόμβοι να βρίσκονται στην ίδια μονάδα είναι μεγαλύτερη στη περίπτωση που το βάρος της ακμής τους είναι υψηλό. Αυτός είναι ο λόγος που η αναζήτηση ξεκινάει από το διάστημα με τις μεγαλύτερες τιμές βαρών. Τα βάρη των ακμών των κόμβων ενός διαστήματος με τον αρχικό κόμβο-σπόρο, είναι παρόμοια. Επιλέγοντας εκείνον με το μεγαλύτερο βαθμό εξασφαλίζουμε ένα μεγαλύτερο σύνολο υποψηφίων για να συνεχίσει η διαδικασία της αναζήτησης.

Ο αλγόριθμος σε κάθε του βήμα έχει το υποσύνολο κόμβων Sγια την συστάδα, το οποίο αρχικά αποτελείται από τους δύο πρώτους κόμβους-σπόρους. Αναζητείται ο κόμβος u με την μεγαλύτερη τιμή support(u, S) μεταξύ όλων των κόμβων που δεν περιλαμβάνονται σε κάποια συστάδα αλλά είναι γείτονες με κάποιον κόμβο του S. Σε περίπτωση που το support(u, S) έχει μικρότερη τιμή από την τιμή του κατωφλιού, ο αλγόριθμος παύει να επεκτείνει την συγκεκριμένη συστάδα και την παρουσιάζει ως αποτέλεσμα. Σε αντίθετη περίπτωση, ο κόμβος u συμπεριλαμβάνεται στο υποσύνολο S και ενημερώνεται η τιμή της πυκνότητας. Εάν η πυκνότητα είναι μικρότερη από το κατώφλι της πυκνότητας

T_d, τότε ο κόμβος u περιλαμβάνεται στην συστάδα και παράγεται το υποσύνολο S. Η διαδικασία επαναλαμβάνεται μέχρι όλοι οι κόμβοι να είναι ομαδοποιημένοι.

5.3.3. Απόδοση αλγορίθμου SPICi

Ο αλγόριθμος SPICi παρουσιάζει 4 - 1,000 φορές υψηλότερη ταχύτητα από άλλες προσεγγίσεις δικτύων (π.χ.: DPClus, CFinder κ.α.) οι οποίες ολοκληρώνονταν σε διάστημα 12 ωρών σε έναν κλασικό επιτραπέζιο υπολογιστή. Επιπλέον πλεονεκτήματα του αλγορίθμου είναι ότι είναι ο μόνος από όσους έχουν εξεταστεί, ο οποίος έχει την δυνατότητα να ομαδοποιεί όλα τα δίκτυα εντός ενός λογικού χρονικού διαστήματος, έχει πολύ καλή απόδοση στην ανακεφαλαιοποίηση των πρωτεϊνικών συμπλεγμάτων, η οποία μειώνεται μόνο σε εξαιρετικά ελλιπή δίκτυα. Επίσης ο SPICi παραμένει ανεπηρέαστος στις μεταβολές πυκνών λειτουργικών δικτύων. Τέλος, παρά το γεγονός ότι ο SPICi είναι γρηγορότερος, οι συστάδες των βιολογικών δικτύων που φέρνει ως αποτέλεσμα ανακεφαλαιώνουν τις λειτουργικές μονάδες τους.

Οι ομάδες που προέκυψαν από τον αλγόριθμο SPICi έχει αποδειχθεί πως είναι ίδιας ποιότητας με αυτές που βρέθηκαν από αλγόριθμους τελευταίας τεχνολογίας. Ο SPICi είναι ιδανικός για πυκνά δίκτυα, ενώ σε αραιωμένα δίκτυα αν και εντοπίζει με ευκολία πυκνές περιοχές, για έναν σημαντικό αριθμό παραμετρικών ρυθμίσεων αρκετοί κόμβοι δεν θα συμπεριληφθούν σε κάποια ομάδα [76].

5.4. Αλγόριθμος Louvain

5.4.1. Εισαγωγή

Ο εντοπισμός συστάδων σε έναν γράφο, προϋποθέτει πως ο γράφος θα πρέπει να χωριστεί σε ομάδες πυκνά συνδεδεμένων κόμβων, όπου οι κόμβοι που ανήκουν σε διαφορετικές ομάδες να είναι αραιά συνδεδεμένοι. Η μοντελοποίηση τέτοιου τύπου προβλήματος βελτιστοποίησης έχει αρκετές υπολογιστικές δυσκολίες. Υπάρχει ποικιλία αλγορίθμων οι οποίοι ομαδοποιούν ικανοποιητικά και γρήγορα τους γράφους, λόγω του πλήθους των μεγάλων δεδομένων δικτύων και την σημαντικότητα τους τα τελευταία χρόνια. Αυτοί οι αλγόριθμοι ομαδοποίησης μπορούν να κατηγοριοποιηθούν ως εξής:

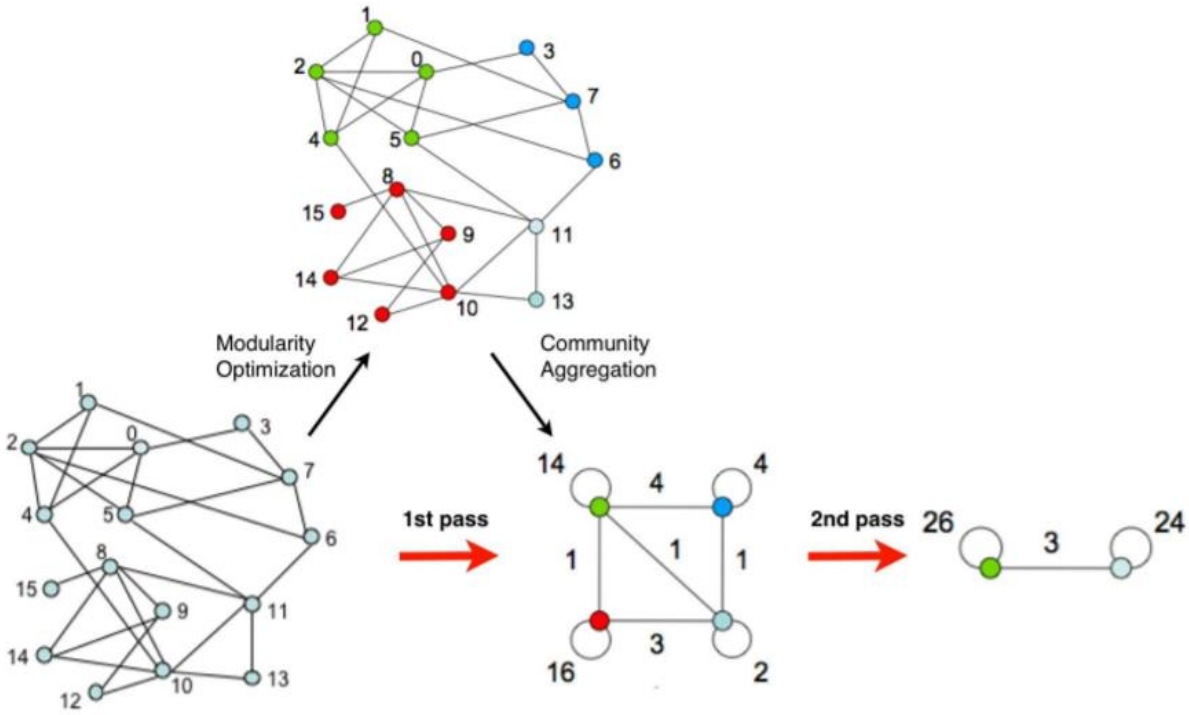
- Διαχωριστικοί αλγόριθμοι - βρίσκουν ενδο-συσταδικούς συνδέσμους και τους αφαιρούν από τον γράφο.
- Συναθροιστικοί αλγόριθμοι - συγχωνεύουν αναδρομικά παρόμοιους κόμβους / συστάδες.
- Αλγόριθμοι βελτιστοποίησης - μεγιστοποίηση αντικειμενικής συνάρτησης.

Η σύγκριση της ποιότητας των συστάδων που προκύπτουν από τους παραπάνω αλγορίθμους συνήθως μετριέται με το modularity (αρθρωτότητα). Το modularity μιας συστάδας είναι μία κλιμακούμενη τιμή μεταξύ -1 και 1 και μετράει την πυκνότητα των συνδέσμων εντός της συστάδας σε σύγκριση με τους συνδέσμους μεταξύ των συστάδων. Επίσης το modularity χρησιμεύει και στην βελτιστοποίηση, αν και αποτελεί ένα πρόβλημα δύσκολο υπολογιστικά. Για να αντιμετωπιστεί η δυσκολία αυτή, οι αλγόριθμοι προσέγγισης κρίθηκαν απαραίτητοι, ειδικά για του μεγάλους γράφους.

5.4.2. Λειτουργία αλγορίθμου Louvain

Ο αλγόριθμος ***Louvain***, εντοπίζει, μέσα σε ένα μεγάλο δίκτυο, ομάδες με υψηλό modularity, σε μικρό χρονικό διάστημα. Επίσης, παρουσιάζει μία πλήρη ιεραρχική δομή ομάδων του δικτύου, η οποία είναι σημαντική αφού δίνει πρόσβαση σε διάφορες αναλύσεις ανίχνευσης των ομάδων. Ο αλγόριθμος

αποτελείται από δύο φάσεις οι οποίες επαναλαμβάνονται. Αρχικά, στο δίκτυο υπάρχει μία ομάδα για κάθε κόμβο του και για κάθε κόμβο, έστω i, έχουμε τους γείτονες του j. Υπολογίζεται, το όφελος που θα είχε το modularity εάν αφαιρεθεί το i από την ομάδα του και μεταφερθεί στην ομάδα του jπου θα έχει το μεγαλύτερο όφελος και μεγαλύτερο από το μηδέν. Σε αντίθετη περίπτωση το i παραμένει στην αρχική του ομάδα. Η διαδικασία επαναλαμβάνεται διαδοχικά για κάθε κόμβο μέχρι να μην υπάρξει περαιτέρω πρόοδος και να τερματιστεί. Χρειάζεται να επισημανθεί πως ένας κόμβος μπορεί να εξεταστεί πάνω από μία φορές και πως η σειρά εξέτασης των κόμβων επηρεάζει το αποτέλεσμα του αλγόριθμου. Στη δεύτερη φάση του αλγορίθμου, δημιουργείται ένα νέο δίκτυο στο οποίο, τη θέση των κόμβων του έχουν οι ομάδες που δημιουργήθηκαν από την πρώτη φάση. Στο νέο δίκτυο, εφόσον είναι σταθμισμένο, τα βάρη των ακμών μεταξύ των νέων κόμβων υπολογίζονται από το άθροισμα των βαρών που έχουν οι ακμές μεταξύ των κόμβων που βρίσκονται στις αντίστοιχες ομάδες. Υπόψιν ότι οι ακμές μεταξύ των κόμβων της ίδιας ομάδας οδηγούν σε αυτο-επαναλαμβανόμενους βρόγχους για την συγκεκριμένη ομάδα στο νέο δίκτυο. Από τη στιγμή που ολοκληρωθεί η δεύτερη φάση του αλγορίθμου, μπορεί να επαναληφθεί η πρώτη στο νέο δίκτυο που προκύπτει κ.ο.κ. Οι δύο φάσεις του Louvain ονομάζονται pass. (Εικόνα 19)



Εικόνα 19. Τα βήματα που ακολουθεί ο αλγόριθμος Louvain. Το κάθε pass περιλαμβάνει τις δύο φάσεις που ακολουθεί ο αλγόριθμος. Στη πρώτη φάση το modularity βελτιστοποιείται επιτρέποντας μόνο τοπικές μεταβολές στις ομάδες. Στη δεύτερη φάση, αφότου έχουν βρεθεί οι ομάδες συγκεντρωτικά ώστε να δημιουργηθεί ένα νέο δίκτυο που θα αποτελείται από ομάδες. Η διαδικασία επαναλαμβάνεται έως ότου δεν υπάρξουν άλλες μειώσεις στο modularity [79].

5.4.3. Απόδοση αλγορίθμου Louvain

Ο αλγόριθμος Louvain σε αντίθεση με άλλους παρόμοιους αλγορίθμους χρειάζεται να αντιμετωπίσει τους περιορισμούς στο μέγεθος του δικτύου εξαιτίας της περιορισμένης χωρητικότητας και όχι του περιορισμένου χρόνου υπολογισμού. Σε ένα δίκτυο 118 εκατομμυρίων κόμβων, η εύρεση ομάδων πήρε

μόνο 152 λεπτά. Ο Louvain συνοδεύεται από μία σειρά από πλεονεκτήματα· τα βήματα του αλγορίθμου είναι διαισθητικά και εύκολα στην εφαρμογή και το αποτέλεσμα δεν χρειάζεται επίβλεψη. Επίσης, όπως έχει ήδη αναφερθεί, ο αλγόριθμος είναι πολύ γρήγορος, πράγμα που οφείλεται στο γεγονός ότι τα πιθανά οφέλη του modularity μπορούν να υπολογιστούν εύκολα με βάση τη διαδικασία που ακολουθεί ο αλγόριθμος και με το ότι το πλήθος των ομάδων μειώνεται σημαντικά ύστερα από την επανάληψη των δύο φάσεων και κατά συνέπεια ο περισσότερος χρόνος να αφιερώνεται μόνο τις πρώτες επαναλήψεις. Το πρόβλημα με τον περιορισμό της ανάλυσης του modularity, λόγω της διαισθητικότητας του αλγορίθμου, έχει αποφευχθεί. Το συγκεκριμένο πρόβλημα, προκαλείται εξαιτίας της αποτυχίας της βελτιστοποίησης του modularity να εντοπίσει ομάδες οι οποίες είναι μικρότερες από μια συγκεκριμένη κλίμακα. Η παρατήρηση αυτή, σχετίζεται μερικώς με τον αλγόριθμο Louvain, αφού στην πρώτη του φάση εκτελείται μετατόπιση μεμονωμένων κόμβων από τη μία ομάδα στην άλλη με αποτέλεσμα η πιθανότητα συγχώνευσης δύο ξεχωριστών ομάδων, μετατοπίζοντας έναν-έναν τους κόμβους, να είναι πολύ χαμηλή. Οι ομάδες μπορεί να συγχωνευθούν μετά από αρκετές επαναλήψεις των δύο φάσεων, όταν θα έχει μαζευτεί ένας αριθμός κόμβων.

Ο αλγόριθμος έχει αποδείξει την ταχύτητά του αφού δοκιμάστηκε σε μεγάλου μεγέθους δίκτυα. Συγκεκριμένα δοκιμάστηκε σε δύο δίκτυα ιστού, το ένα υπο-δίκτυο .uk, 39 εκατομμυρίων κόμβων και 783 εκατομμυρίων ακμών, ενώ το δεύτερο ήταν ένα δίκτυο 118 εκατομμυρίων κόμβων και 1 δισεκατομμυρίου ακμών που λήφθηκαν από το πρόγραμμα ανίχνευσης του Stanford WebBase. Ο χρόνος που χρειάστηκε για το πρώτο δίκτυο ήταν μόλις 12 λεπτά ενώ για το δεύτερο μόλις 152. Παρ ’όλα αυτά, ο Louvain έχει περιθώρια βελτίωσης στη ταχύτητα του χρησιμοποιώντας μερικά απλά ευρετικά στοιχεία, όπως για παράδειγμα όταν το όφελος του modularity είναι κάτω από ένα κατώφλι να διακόπτεται η πρώτη του φάση. Μία ακόμη μέθοδος για την βελτίωση της ταχύτητας είναι η αφαίρεση, από το αρχικό δίκτυο, των κόμβων με βαθμό 1 και να προστεθούν αφότου έχει ολοκληρωθεί η ομαδοποίηση [79].

5.5. Αλγόριθμος Label Propagation

5.5.1. Εισαγωγή

Ο αλγόριθμος ***Label Propagation (LPA)***, παρά το γεγονός ότι το μέγεθος των ομάδων που αναλύονται σε ένα δίκτυο όλο και μεγαλώνουν, ο σχεδόν γραμμικός χρόνος εκτέλεσης του και η εύκολη εφαρμογή του, ευνοούν τον αποτελεσματικό εντοπισμό των ομάδων. Εάν ένα δίκτυο μπορεί να χωριστεί σε ομάδες οι οποίες εντός τους έχουν πυκνές ακμές ενώ μεταξύ τους έχουν αραιές ακμές, τότε το δίκτυο έχει δομή ομάδας ή αλλιώς συσταδική δομή. Ο αλγόριθμος έχει εφαρμοστεί σε πολλά είδη δικτύων όπως ο Παγκόσμιο Ιστός, κοινωνικά δίκτυα, βιολογικά δίκτυα κ.α. Η ποιότητα των ομάδων που προκύπτουν από την εφαρμογή του αλγορίθμου μετριέται με το modularity, το οποίο εάν κυμαίνεται μεταξύ των τιμών 0.3 και 0.7 υποδεικνύει ότι το δίκτυο έχει δομή ομάδας.

5.5.2. Λειτουργία αλγορίθμου Label Propagation

Ο αλγόριθμος LPA είναι αποδοτικός ώστε να βρίσκει ομάδες σε μεγάλα δίκτυα, ενώ τρέχει γραμμικά στο πλήθος των ακμών αλλά το ίδιο γραμμικά τρέχει στο πλήθος των κόμβων στην περίπτωση των αραιών δικτύων. Ο αλγόριθμος λειτουργεί ως εξής: στον κάθε κόμβο αντιστοιχεί μία μοναδική ετικέτα (label) και κατά την διάρκεια της επαναληπτικής διαδικασίας ο καθένας “υιοθετεί” την ετικέτα σε συμφωνία με την πλειοψηφία των γειτόνων του. Ξεκινώντας από έναν κόμβο με ετικέτα, ο αλγόριθμος διαδίδει την ετικέτα

του σε κάθε βήμα και στο κάθε βήμα περιλαμβάνει όλο και παραπάνω γειτονικούς κόμβους έως ότου δημιουργηθεί μία ομάδα. Όταν τελειώσει ο αλγόριθμος, οι κόμβοι που είναι συνδεδεμένοι με την ίδια ετικέτα, αποτελούν μία ομάδα. Μία ομάδα οριοθετείται με βάση το κατώφλι που ορίζεται ως η αναλογία του αριθμού των ακμών εντός και εκτός μιας ομάδας. Ο κάθε κόμβος του δικτύου ενημερώνει την τιμή του λαμβάνοντας τον μέσο όρο της τιμής όλων των γειτόνων του. Όσο η διαδικασία προχωράει, το κενό της τιμής υποδηλώνει το όριο μεταξύ των δύο ομάδων που μόλις δημιουργήθηκαν. Η διαδικασία μπορεί να γενικευθεί και για τον εντοπισμό πολλών ομάδων αλλά χρειάζεται των αριθμό των ομάδων ως είσοδο και τείνει να βρίσκει ομάδες περίπου ίσου μεγέθους μεταξύ τους.

5.5.3. Απόδοση αλγορίθμου Label Propagation

Πέρα από την εύκολη εφαρμογή και γρήγορη εκτέλεση που ήδη έχουν αναφερθεί, στα πλεονεκτήματα του LPA είναι και το ότι χρησιμοποιεί μόνο τη δομή του δικτύου για να καθοδηγήσει τη διαδικασία του και δεν απαιτεί ούτε παραμέτρους ούτε βελτιστοποίηση της αντικειμενικής λειτουργίας. Παρά τα θετικά στοιχεία του, ο αλγόριθμος LPA έχει το μειονέκτημα ότι μπορεί να φέρει ως αποτέλεσμα διαφορετικές λύσεις (με μερικές από αυτές να είναι κακής ποιότητας) σε διαφορετικές εκτελέσεις, πράγμα που οφείλεται στην ποιότητα της λύσης του αλγορίθμου η οποία εξαρτάται από τα τοπικά ελάχιστα στα οποία φτάνει. Ο αριθμός των τοπικών ελαχίστων έχει αποδειχθεί πως είναι πολύ μεγαλύτερος από τον αριθμό των κόμβων στο υποκείμενο δίκτυο. Επίσης, ο χρόνος εκτέλεσης του LPA έχει περιθώρια βελτίωσης τα οποία είναι σημαντικά όταν πρόκειται για πολύ μεγάλα δίκτυα. Ένας τρόπος βελτίωσης είναι η αποφυγή περιττών ενημερώσεων σε κάθε επανάληψη του αλγορίθμου, διατηρώντας παράλληλα αμετάβλητη τη συνολική συμπεριφορά του αλγορίθμου. Ύστερα από πέντε επαναλήψεις, το 95% των κόμβων ομαδοποιούνται ήδη σωστά. Ο επιπλέον χρόνος που απαιτείται αφορά τις ενημερώσεις που ουσιαστικά δεν αλλάζουν τις ετικέτες, και έτσι καθυστερεί το πέρας του αλγορίθμου. Ο χρόνος μπορεί να εξοικονομηθεί με τη φύλαξη των πληροφοριών σχετικά με τα όρια των ομάδων που έχουν δημιουργηθεί. Πιο συγκεκριμένα, τα βήματα βελτιστοποίησης του αλγορίθμου LPA είναι:

1. Τη χρονική στιγμή t = 0, δημιουργείται μία λίστα ενεργών κόμβων (κόμβοι που θα άλλαζαν την ετικέτα τους σε πιθανή ενημέρωση) που περιέχει όλους τους κόμβους.
2. Γίνεται τυχαία επιλογή ενός ενεργού κόμβου από την λίστα, έστω ο κόμβος i , και γίνονται προσπάθειες να υιοθετηθεί μία νέα ετικέτα σύμφωνα με την ενημέρωση. Δεδομένου ότι μόνο οι ενεργοί κόμβοι τοποθετούνται αρχικά στη λίστα και παραμένουν στη λίστα όσο είναι ενεργοί, κάθε κόμβος που επιλέγεται για μια ενημέρωση θα αλλάξει την ετικέτα του κατά τη διάρκεια της ενημέρωσης.
3. Ελέγχεται εάν ο ενημερωμένος κόμβος μετατράπηκε σε παθητικός (κόμβοι που δεν θα άλλαζαν την ετικέτα τους σε πιθανή ενημέρωση). Εάν ναι, θα πρέπει να αφαιρεθεί από την λίστα και ύστερα να ελεγχθούν οι γείτονες του ως εξής: εάν ένας εσωτερικός γείτονας έγινε ενεργός κόμβος ορίου, τότε προστίθεται στη λίστα ενεργών κόμβων, αφαιρούνται όλοι οι προηγούμενοι ενεργοί γείτονες που μετατράπηκαν σε παθητικοί από τη λίστα ενεργών κόμβων και τέλος, προστίθεται οποιοσδήποτε προηγουμένως παθητικός γείτονας-κόμβος ορίου ο οποίος έγινε ενεργός στη λίστα ενεργών κόμβων.
4. Εάν η λίστα ενεργών κόμβων είναι άδεια τότε επέρχεται ο τερματισμός, διαφορετικά η χρονική στιγμή t αυξάνεται κατά 1 και επαναλαμβάνονται τα βήματα από το βήμα 2.

Σε αυτό το σημείο, θεωρείται βοηθητικό να διευκρινιστεί πως αναφέρεται ένας κόμβος, του οποίου όλοι οι γείτονες έχουν την ίδια ετικέτα, όπως και σε έναν εσωτερικό κόμβο. Όσοι κόμβοι δεν είναι εσωτερικοί,

ονομάζονται κόμβος ορίου. Τέλος, ένας κόμβος που δεν αλλάζει την ετικέτα του με αφορμή μια ενημέρωση αναφέρεται ως παθητικός, ενώ αντίθετα ονομάζεται ενεργός. Εξ' ορισμού λοιπόν, όλοι οι εσωτερικοί κόμβοι είναι παθητικοί ενώ ένας κόμβος ορίου μπορεί να θεωρείται ενεργός ή παθητικός ανάλογα τους γείτονες του. Έτσι, κάθε κόμβος μπορεί να είναι είτε εσωτερικός-παθητικός, είτε κόμβος ορίου-παθητικός, είτε κόμβος ορίου-ενεργητικός.

Ο τρόπος με τον οποίο λειτουργεί ο αλγόριθμος LPA, δηλαδή η μετάδοση της ετικέτας από τον έναν κόμβο στον άλλον, θυμίζει την εξάπλωση μιας επιδημίας ή μιας ιδέας. Υποθέτοντας ότι ένας κόμβος υιοθετεί πάντα την ετικέτα που έχει το μεγαλύτερο μέρος των γειτόνων του, ο LPA αγνοεί τυχόν δομές που υπάρχουν σε αυτήν τη γειτονιά κόμβων. Βέβαια, δεν αρκεί μόνο η πλειοψηφία των γειτόνων, αλλά και η ποιότητα σύνδεσης τους. Για παράδειγμα, για την εξάπλωση μιας ιδέας, εάν ένας ανεξάρτητος άνθρωπος θα την ακολουθήσει, δεν αρκεί να πειστεί μόνο από το πλήθος των ατόμων που την υιοθετούν, αλλά και από τις διασυνδέσεις μεταξύ τους. Αυτό το χαρακτηριστικό κάνει τον αλγόριθμο να θεωρείται απλός [80].

5.6. Αλγόριθμος Walktrap

5.6.1. Εισαγωγή

Τα πολύπλοκα δίκτυα όπως είναι τα βιολογικά, έχουν μεγάλη σημασία στην επιστημονική κοινότητα, και η δομή των αντίστοιχων γράφων τους δίνει αρκετές πληροφορίες για το κάθε δίκτυο. Οι γράφοι αυτοί είναι σε γενικό επίπεδο αραιοί αλλά τοπικά πυκνοί, αφού υπάρχουν ομάδες κόμβων γνωστές ως συστάδες οι οποίες συνδέονται με πολλές ακμές μεταξύ τους αλλά με λίγες προς άλλους κόμβους. Ένα βιολογικό παράδειγμα αποτελεί ένα μεταβολικό δίκτυο όπου οι συστάδες του αντιστοιχούν σε βιολογικές λειτουργίες του κυττάρου [81].

Ο αλγόριθμος ***Walktrap*** αναπτύχθηκε για την ανίχνευση συστάδων σε μεγάλα δίκτυα μέσω τυχαίων περιπάτων/μετακινήσεων, οι οποίοι χρησιμεύουν στον υπολογισμό των αποστάσεων μεταξύ κόμβων. Οι κόμβοι διαχωρίζονται σε ομάδες, με μικρές ενδο-συσταδικές και μεγάλες δια-συσταδικές αποστάσεις μεταξύ κόμβων, μέσω μίας ιεραρχικής ομαδοποίησης από κάτω προς τα πάνω. Ο αλγόριθμος αν και είναι ιδανικός για δίκτυα μεγάλης κλίμακας και έχει χαμηλή πολυπλοκότητα χρόνου, έχει χαμηλή ακρίβεια στα αποτελέσματα εντοπισμού και δεν μπορεί να ταυτοποιήσει τους επικαλυπτόμενους κόμβους. Επίσης ο αλγόριθμος έχει ανάγκη για μεγάλη μνήμη [81] [82].

5.6.2. Λειτουργία αλγορίθμου Walktrap

Οι τυχαίοι περίπατοι σε έναν γράφο έχουν την τάση να "παγιδεύονται" σε πυκνά συνδεδεμένα μέρη που αντιστοιχούν σε συστάδες. Με τη βοήθεια των τυχαίων περιπάτων μπορεί να οριστεί μία μέτρηση της δομικής ομοιότητας μεταξύ κόμβων και μεταξύ συστάδων, ορίζοντας μία απόσταση. Συγκεκριμένα, με τη χρήση της απόστασης αυτής, ο αλγόριθμος μπορεί να μετρήσει και να συγκρίνει την ομοιότητα μεταξύ των κόμβων, ύστερα βρίσκει την πιθανότητα των τυχαίων βημάτων της μετακίνησης, υπολογίζει την απόσταση μεταξύ των κόμβων και τέλος, προκύπτει η απόσταση. Με τη χρήση ιεραρχικών αλγορίθμων ομαδοποίησης και με βάση την απόσταση που προκύπτει, μπορούν να ληφθούν διαφορετικής κλίμακας συσταδικές δομές οι οποίες μπορούν να αναπαρασταθούν ως δέντρα, γνωστά ως δενδρογράμματα. Για να απλοποιηθεί η υπολογιστική πολυπλοκότητα του Walktrap, όσο οι γειτονικές συστάδες χωρίζονται, υπολογίζεται και ενημερώνεται επανειλημμένα το μέσο τετράγωνο της απόστασης και ελαχιστοποιείται,

και τέλος το modularity χρησιμοποιείται για την αξιολόγηση των αποτελεσμάτων του διαχωρισμού. Σχετικά με τον διαχωρισμό του δικτύου, ο αλγόριθμος ανήκει στην κατηγορία όσων χρησιμοποιούν ως κριτήριο την ομοιότητα των κόμβων (node similarity) [81] [82].

Για την ομαδοποίηση των κόμβων, χρειάζεται ο ορισμός της απόστασης η οποία θα έχει μεγάλες τιμές όταν δύο κόμβοι ανήκουν σε διαφορετικές συστάδες και μικρές σε αντίθετη περίπτωση. Η διαδικασία που ακολουθεί ο αλγόριθμος Walktrap ξεκινάει με έναν διαχωρισμό του γράφου, έστω $P_1 = \{\{v\}\}, v \in V\}$, σε n μέρη αποτελούμενα από έναν κόμβο και υπολογίζει τις αποστάσεις μεταξύ όλων των γειτόνων. Ύστερα, το διαμέρισμα αυτό εξελίσσεται επαναλαμβάνοντας τα εξής· σε κάθε βήμα, έστω k:

- Επιλέγονται από το P_k δύο ομάδες, C_1και C_2, σύμφωνα με την απόσταση μεταξύ των ομάδων.
- Οι δύο αυτές ομάδες συγχωνεύονται σε μία, την $C_3 = C_1 \cup C_2$ και δημιουργείται ένας νέος διαχωρισμός $P_{k+1} = (P_k \setminus \{C_1, C_2\}) \cup \{C_3\}$.
- Οι αποστάσεις μεταξύ των ομάδων ανανεώνονται.
- Ο αλγόριθμος τερματίζει μετά από $n - 1$ βήματα με $P_n = \{V\}$.

Σε κάθε βήμα ορίζεται ένας διαμερισμός P_k του γράφου, ο οποίος φέρνει ως αποτέλεσμα ένα δενδρόγραμμα. Τα φύλλα του δενδρογράμματος, αντιστοιχούν σε κόμβους του δικτύου ενώ ο κάθε εσωτερικός κόμβος του δενδρογράμματος σχετίζεται με τη συγχώνευση των ομάδων, όπως περιγράφηκε παραπάνω.

5.6.3. Απόδοση αλγορίθμου Walktrap

Ο αλγόριθμος Walktrap υπολογίζει την συσταδική δομή σε χρόνο $O(mnH)$ όπου το H αντιστοιχεί στο ύψος του δενδρογράμματος. Στο χειρότερο σενάριο η χρονική πολυπλοκότητα είναι $O(mn^2)$. Στα πραγματικά πολύπλοκα δίκτυα τα οποία είναι αραιά τότε $m = O(n)$ και το H παίρνει μικρές τιμές και τείνει στο πιο ευνοϊκό σενάριο όπου το δενδρόγραμμα είναι ισορροπημένο, $H = O(log\ n)$. Σε αυτή τη περίπτωση η πολυπλοκότητα υπολογίζεται ως εξής: $O(n^2\ log\ n)$. [82]

6. Conductance

6.1. Εισαγωγή

Το ***conductance*** (αγωγιμότητα) είναι ένα μέτρο για την ποιότητα της σύνδεσης (ή αλλιώς, το πόσο καλή είναι η σύνδεση) μίας συστάδας με το υπόλοιπο δίκτυο, σε σχέση με τις εσωτερικές διασυνδέσεις του δικτύου [83]. Συγκεκριμένα ποσοτικοποιεί την ποιότητα της σύνδεσης ενός υπο-γράφου με τον υπόλοιπο γράφο σε σχέση με τις εσωτερικές συνδέσεις του και χρησιμοποιείται για την μέτρηση της ποιότητας των αλγορίθμων ομαδοποίησης/συσταδοποίησης. Επιπλέον, άλλη μία κοινή εφαρμογή του conductance είναι για την παρακολούθηση της διάδοσης ειδήσεων ή και για την υπόδειξη ότι ένα επίκαιρο ζήτημα έχει γίνει ιδιαίτερα δημοφιλές. Χρειάζεται να διευκρινιστεί πως το conductance ενός υπο-γράφου και εκείνο ενός γράφου διαφέρουν. Το conductance ενός γράφου είναι το μικρότερο conductance μεταξύ όλων των πιθανών υπο-γράφων του αρχικού γράφου. Ακόμη και η υπολογιστική πολυπλοκότητα των δυο αυτών conductance διαφέρουν σημαντικά. Σύμφωνα με τον ορισμό του conductance δείχνει πόσο "εκτεταμένη" ή "στενά συνδεδεμένη" είναι μια συστάδα κορυφών· μεγαλύτερο conductance συνεπάγεται περισσότερες εξωτερικές συνδέσεις, ενώ όσο πιο μικρό είναι τότε η συστάδα αυτή είναι πιο "εσωστρεφής" ("inward-looking.").

Για έναν γράφο (είτε σταθμισμένο είτε όχι), έστω $G = (V, E)$, αντί για το conductance μπορούμε να αναφερθούμε σύμφωνα με την βιβλιογραφία στο sparsity (βραδύτητα) ενός cut (μέρους/μεριδίου), καθώς μία ομάδα κόμβων $S \subset V$μπορεί να θεωρηθεί ως cut που διαχωρίζει τον γράφο G σε δύο μέρη $S, V \backslash S$. Στην περίπτωση των τυχαίων περιπάτων στους γράφους, από πολλούς χρησιμοποιείται ο όρος "bottleneck ratio of a set of states" ("λόγος συμφόρησης ενός συνόλου αντικειμένων-κόμβων") για να περιγραφεί η πιθανότητα πως ένας τυχαίος περίπατος ο οποίος ορίζεται από ένα σύνολο κόμβων V, μετακινείται από ένα υποσύνολο κόμβων S προς έναν κόμβο από το $V \backslash S$. Στην περίπτωση των τυχαίων περιπάτων, το bottleneck ratio παίζει τον ίδιο ρόλο με το conductance.

Για τον υπολογισμό του conductance ενός υποσυνόλου κόμβων S, είναι απαραίτητος ο διαχωρισμός των ακμών των κόμβων του S που είναι εσωτερικές. Για παράδειγμα, για την ακμή (u, v) όπου ο κόμβος $u \in S$, χρειάζεται να εξακριβωθεί αν και το $v \in S$. Για την συγκεκριμένη διαδικασία χρειάζεται χρόνος $O(\bar{d}|S|T(|S|))$, όπου το $\bar{d}$ είναι ο μέσος βαθμός των κόμβων του συνόλου S και το $T(k)$ είναι ο χρόνος που απαιτείται να απαντηθεί ένα σύνολο ερωτημάτων από ένα σύνολο k αντικειμένων. Παρά το γεγονός πως ο υπολογισμός του conductance είναι πολυωνυμικό πρόβλημα χρόνου, ο υπολογισμός του conductance του γράφου - όπως αναφέρθηκε παραπάνω, προϋποθέτει να βρει το ελάχιστο conductance μεταξύ όλων των πιθανών υπο-γράφων του - αποτελεί ένα πρόβλημα NP [84].

6.2. Conductance και αλγόριθμοι ομαδοποίησης

Η εύρεση ομάδων/συστάδων σε ένα δίκτυο όπως είναι τα κοινωνικά δίκτυα, τα γραφήματα ιστού και τα βιολογικά δίκτυα είναι ένα πολύ σημαντικό και ενδιαφέρον ζήτημα στον επιστημονικό χώρο. Χαρακτηριστικό μίας συστάδας ενός δικτύου είναι πως οι κόμβοι της, έχουν περισσότερες ή / και ισχυρότερες αλληλεπιδράσεις μεταξύ τους παρά μεταξύ των κόμβων της συστάδας και του υπόλοιπου δικτύου. Το συγκεκριμένο χαρακτηριστικό είναι ιδιαίτερα σημαντικό αφού δίνει την αίσθηση της ύπαρξης μίας συστάδας, όπου μπορεί να βασιστεί μία αντικειμενική συνάρτηση εντοπισμού συστάδων. Για την κατανόηση της ποιότητας των διάφορων αλγορίθμων ανίχνευσης συστάδων, υπολογίζονται θεωρητικά χαμηλότερα όρια στον υπολογισμό του conductance.

Το conductance είναι η πιο απλή έννοια για την ποιότητα μιας συστάδας καθώς αντικατοπτρίζει την ιδέα “επιφάνεια-προς-όγκο” (“surface area-to-volume”) και για αυτό είναι ένα ευρέως διαδεδομένο μέτρο για τον εντοπισμό μιας καλής συστάδας ως ένα σύνολο κόμβων που έχουν καλύτερη εσωτερική - από ότι εξωτερική συνδεσιμότητα. Για την σύγκριση διαφορετικών αλγορίθμων ομαδοποίησης, η ανάλυση χωρίζεται σε δύο πτυχές. Αρχικά ιδιαίτερη σημασία έχουν οι ποιότητες των ομάδων που εντοπίζονται, δηλαδή θα πρέπει να γίνει αντιληπτό πόσο καλά αποδίδουν οι αλγόριθμοι όσον αφορά τη βελτιστοποίηση της έννοιας της ποιότητας της ομάδας. Σε δεύτερη φάση, γίνεται ποσοτικοποίηση των δομικών ιδιοτήτων των προσδιορισμένων από τους αλγορίθμους ομάδων [85].

6.3. Cheeger σταθερά και Cheeger ανισότητες

Στη Θεωρία των Γράφων, η σταθερά του Cheeger, είναι ένα αριθμητικό μέτρο που μετράει εάν ένας γράφος έχει κάποιο σημείο συμφόρησης ("bottleneck"). Ορίζεται ως εξής:

Έστω ένας μη κατευθυνόμενος γράφος $G = (V, E)$ και ένα υποσύνολο των κόμβων $S \subset V$, όπου ως ∂Sδηλώνεται η συλλογή όλων των ακμών που ξεκινάνε από έναν κόμβο του υποσυνόλου Sκαι καταλήγουν σε κόμβο εκτός του S, γνωστό ως edge boundary του S.

$$\partial S := \{\{x, y\} \in E : x \in S, y \in V \backslash S\}$$

Η σταθερά του Cheeger του γράφου G, συμβολίζεται $h(G)$και ορίζεται:

$$h(G) := min\{\frac{|\partial S|}{|S|} : S \subset V, 0 < |S| \leq \frac{1}{2}|V|\}$$

Εάν η σταθερά του Cheeger είναι μικρή και θετική τότε στον γράφο υπάρχει σημείο συμφόρησης, δηλαδή υπάρχουν δύο μεγάλα σύνολα κόμβων που μεταξύ τους υπάρχουν λίγες ακμές. Στην αντίθετη περίπτωση που η σταθερά είναι μεγάλη τότε αυτό σημαίνει πως μεταξύ των δύο συνόλων υπάρχουν πολλές ακμές [86].

Η σταθερά του Cheeger έχει ιδιαίτερη σημασία στους γράφους εκείνους που είναι αραιοί και έχουν ισχυρές ιδιότητες συνδεσιμότητας, αφού είναι μία μέθοδος μέτρησης της επέκτασης των ακμών ενός γράφου. Οι Cheeger ανισότητες σχετίζονται με το κενό των ιδιοτιμών ενός γράφου με τη σταθερά του Cheeger.

$$2h(G) \geq \lambda \frac{h^2(G)}{2\Delta(G)}$$

Όπου το $\Delta(G)$είναι ο μέγιστος βαθμός για τους κόμβους του γράφου και το λείναι το φασματικό κενό του Λαπλασιανού πίνακα του γράφου [87], [88].

6.4. Υπολογισμός conductance

Έστω $S \subset V$, το S λεγεται ***cut of a graph*** γιατί χωρίζει το V σε S & S'. Ορίζουμε πως στο υποσύνολο S είναι οι κόμβοι που ανήκουν σε κάποια συστάδα ενός γράφου, ενώ στο S' είναι όλοι οι υπόλοιποι κόμβοι του G.

Το ***volume*** (όγκος) ενός κόμβου v ορίζεται ως εξής: $vol\ v\ =\ \sum_u W_{v,u}$.

Παρομοίως ορίζεται και το volume ενός συνόλου S: $vol\ S\ =\ \sum_{v \in S} vol\ v$.

Το volume είναι το άθροισμα των βαθμών (degrees) των κόμβων του S. Χρειάζεται ο υπολογισμός του volume και για το υποσύνολο S και για το S'. Εάν ο γράφος είναι κατευθυνόμενος τότε αντί για το άθροισμα των degrees, υπολογίζεται το άθροισμα των out-degrees.

Το volume ενός cut ορίζεται ως εξής: $vol\ \partial S\ =\ \sum_{u \in S, v \in S'} W_{u,v}$. Το cut, το συναντάμε και ως ***cut size*** και είναι το άθροισμα των βαρών των ακμών μεταξύ των δύο υποσυνόλων Sκαι S'. Εάν ο γράφος είναι χωρίς βάρη, όπως ήδη γνωρίζουμε, το βάρος κάθε ακμής είναι 1.

Ο υπολογισμός του cut size χρησιμοποιεί την έννοια του ***edge boundary*** που αναφέρθηκε στην προηγούμενη ενότητα. Τα edge boundaries είναι οι ακμές που έχουν μόνο ένα άκρο στο δεδομένο σύνολο από κόμβους.

Υπολογισμός του edge boundary:

1. Για κάθε κόμβο (έστω n_1) που ανήκει στο S

    2. Για κάθε κόμβο γενικά του γράφουG (έστω n_2) που συνδέεται με το n_1

        3. Αν το n_2υπάρχει στο S'→ Επιστρέφει το ζεύγος κόμβων (n_1,n_2) και παίρνουμε το άθροισμα των βαρών των ζευγών αυτών.

Το volume του cut είναι συμμετρικό για κάθε πλευρά του διαχωρισμού, γεγονός που οφείλεται από την συμμετρία του W.

$vol\ \partial S\ =\ \sum_{u \in S, v \in S'} W_{u,v}\ =\ \sum_{u \in S, v \in S'} W_{v,u} = vol\ \partial S'$, άρα $vol\ \partial S = vol\ \partial S'$.

Ο τύπος που ορίζει το conductance του cut S είναι: $\phi_G(S)\ =\ \frac{vol\ \partial S}{min\ (vol\ S, vol\ S')}$,

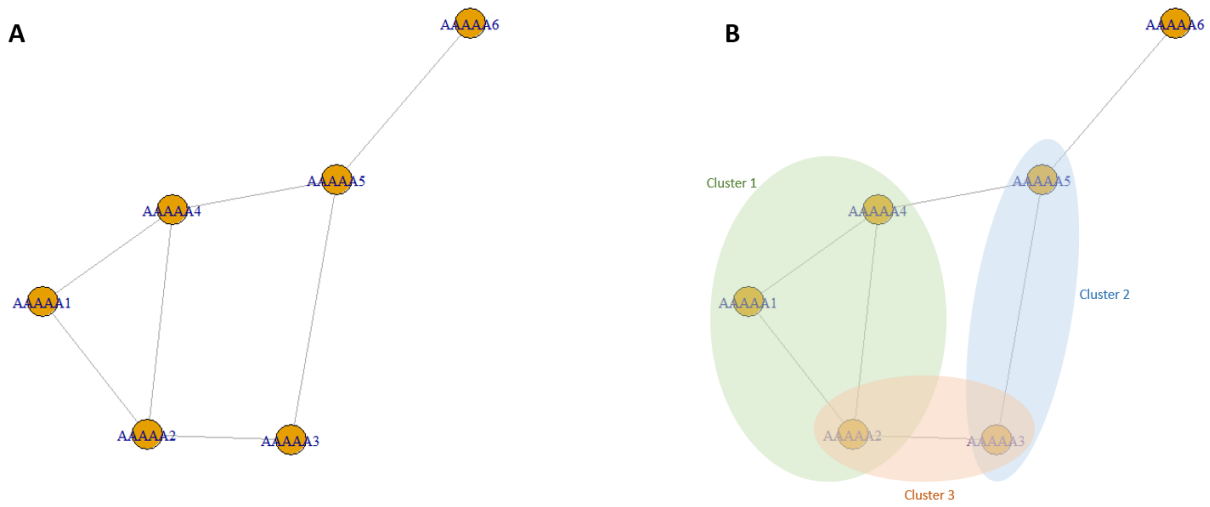
ενώ ο τύπος που ορίζει το conductance ολόκληρου του γράφου είναι: $\phi_G\ = min_{S \subset V}\ \rho_G(S)$ [83], [89].

Σε αυτό το σημείο χρειάζεται να διευκρινιστεί πως το conductance ορίζεται για ένα σύνολο (set) κόμβων και όχι για πολλαπλά σύνολα. Το conductance μετράει πόσο καλή είναι μία συστάδα και όχι πόσο καλά ένας ολόκληρος γράφος μπορεί να χωριστεί. Ιδανικά, καλό είναι να έχουμε χαμηλό conductance (κοντά στο 0) και αυτό επιτυγχάνεται εφόσον έχουμε μικρό αριθμό ακμών που συνδέουν τα Sκαι S' και ταυτοχρόνως υψηλό αριθμό στα volume των S και S'.

Παράδειγμα:

Στην Εικόνα 20Α βλέπουμε έναν δειγματικό γράφο G ενώ στην Εικόνα 20Β φαίνονται τρεις υποθετικές συστάδες (Cluster 1, Cluster 2, Cluster 3) για τον συγκεκριμένο γράφο.

Για την κάθε συστάδα, διευκρινίζονται οι ομάδες S και S', υπολογίζονται τα αντίστοιχα cut size, volume(S), volume (S') και τέλος το conductance το καθενός.



Εικόνα 20. Α) Δειγματικός μη κατευθυνόμενος και μη σταθμισμένος γράφος *G*
Β) Τρεις υποθετικές συστάδες/ομάδες του γράφου

Για το Cluster 1:

$$S = \{"AAAAA1", "AAAAA2", "AAAAA4"\}$$
$$S' = \{"AAAAA6", "AAAAA3", "AAAAA5"\}$$
$$cut\ size\ = 2$$
$$volume(S)\ = 8$$
$$volume(S')\ = 6$$
$$Conductance_{Cluster1}\ = \frac{2}{min(8,6)} = 0.333$$

Για το Cluster 2:

$$S = \{"AAAAA3", "AAAAA5"\}$$
$$S' = \{"AAAAA6", "AAAAA1", "AAAAA2", "AAAAA4"\}$$
$$cut\ size\ = 3\ (από\ "AAAAA4" \rightarrow "AAAAA5", "AAAAA6" \rightarrow "AAAAA5"\ \&\ "AAAAA3" \rightarrow "AAAAA2")$$
$$volume(S)\ = 5$$
$$volume(S')\ = 9$$
$$Conductance_{Cluster2}\ = \frac{3}{min(5,9)} = 0.6$$

Για το Cluster 3:

$S = \{"AAAAA3", "AAAAA2"\}$

$S' = \{"AAAAA6"\}$

$cut\ size\ = 3$

$volume(S)\ = 5$

$volume(S')\ = 9$

$Conductance_{Cluster3}\ = \frac{3}{min(5,9)} = 0.6$

Άρα το conductance του γράφου είναι: $Conductance_G\ = min\{0.333,0.6,0.6\}\ =\ 0.333$.

7. Συγκριτική ανάλυση δικτύων

7.1. Εισαγωγή

Τα δίκτυα που χρησιμοποιήθηκαν στην συγκεκριμένη διπλωματική εργασία, ήταν τριών διαφορετικών τύπων και προερχόντουσαν από πολλούς διαφορετικούς οργανισμούς, όπως εξηγείται παρακάτω. Τα αρχεία των δικτύων αφού πρώτα συλλέχθηκαν, μετατράπηκαν όλα σε μία συγκεκριμένη μορφή αρχείου ώστε να περάσουν όλα από την διαδικασία του λειτουργικού σχολιασμού. Τα αρχεία έπειτα από τον λειτουργικό σχολιασμό κάθε δικτύου, ουσιαστικά είναι αρχεία ομαδοποίησης των δικτύων (clustering files) τα οποία χρησιμοποιήθηκαν ώστε να υπολογιστεί το conductance της κάθε ομάδας εντός του δικτύου. Τα αποτελέσματα του conductance παρουσιάζονται για διευκόλυνση σε μορφή ιστογράμματος και φαίνεται για το κάθε clustering αρχείο, πόσες ομάδες έχουν “καλό” conductance, δηλαδή μικρότερο του 0.5, και πόσες όχι.

7.2. Συλλογή δικτύων

Για τις ανάγκες της συγκεκριμένης διπλωματικής εργασίας συλλέχθηκαν 24 πρωτεϊνικά δίκτυα (PPIs), 5 δίκτυα γονιδιακής συν-έκφρασης (GCNs) και 18 δίκτυα ομοιότητας αλληλουχιών (SSNs). Τα πρωτεϊνικά δίκτυα προήλθαν κυρίως από τις βάσεις BioGrid, DIP και IntAct, τα δίκτυα γονιδιακής συν-έκφρασης προήλθαν από τις βάσεις GeneMania, Co-Expedia και Gene Friends, και τέλος τα δίκτυα ομοιότητας αλληλουχιών προέκυψαν με βάση τη διαδικασία που περιγράφηκε στην ενότητα 3.3.2. Μερικοί από τους οργανισμούς από όπου προήλθαν τα βιομόρια των δικτύων είναι: *Arabidopsis thaliana, Homo Sapiens, Mus musculus, Drosophila melanogaster, Saccharomyces cerevisiae, Oryza Sativa, Sus scrofa domesticus* και *Gallus gallus domesticus.*

Κάθε δίκτυο δεδομένων υποβλήθηκε σε λειτουργικό εμπλουτισμό (functional enrichment) μέσω του εργαλείου DAVID. Για να ανέβει το κάθε αρχείο και να χρησιμοποιηθεί επιτυχώς το συγκεκριμένο εργαλείο, χρειάστηκαν ορισμένες τροποποιήσεις στην μορφή των αρχείων που φιλοξενούσαν τα δίκτυα. Η αρχική μορφή των δικτύων ήταν:

From	To
A	B
C	B
C	C

Όπως φαίνεται στην παραπάνω μορφή υπάρχει επικεφαλίδα που δημιουργεί δύο στήλες (From To) και η κάθε στήλη διαχωρίζεται από την άλλη με ένα tab (\t). Επίσης παρατηρούμε πως στο δίκτυο αυτό υπάρχουν και αυτο-επαναλαμβανόμενοι βρόγχοι βιομορίων-κόμβων (C C). Από το δίκτυο αρχικά αφαιρέσαμε την επικεφαλίδα, ύστερα τους αυτο-επαναλαμβανόμενους βρόγχους, και τέλος κρατήσαμε σε μία μόνο στήλη τους μοναδικούς κόμβους του (A B C).

7.3. Λειτουργικός Εμπλουτισμός (Functional Enrichment)

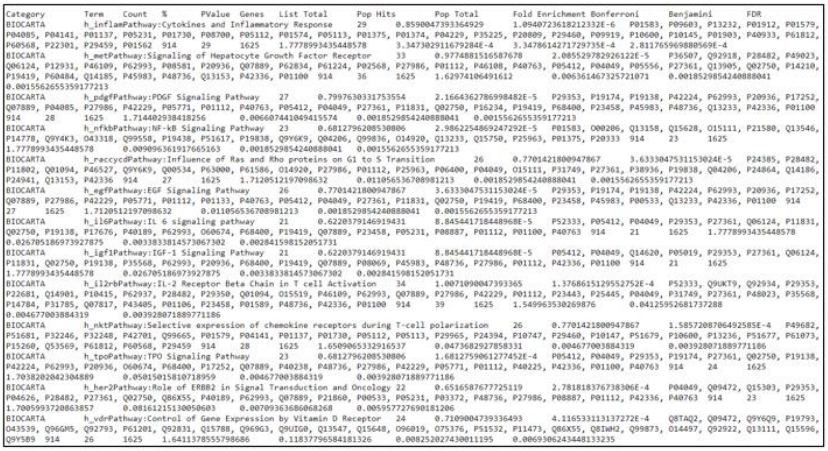
7.3.1. DAVID

Το αρχείο με του μοναδικούς κόμβους που δημιουργήθηκε ήταν στην κατάλληλη μορφή για να μπορέσει να μεταφορτωθεί στο DAVID και ύστερα να επιλεχθεί από την drop-down λίστα του εργαλείου ο τύπος του identifier που περιέχει το αρχείο. Αφού ανέβηκε το αρχείο επιτυχώς, από την κάθε κατηγορία σχολιασμού (annotation categories) του Functional Annotation Summary που εμφανιζόταν κατεβάσαμε όλες τις προεπιλεγμένες από το εργαλείο. Για παράδειγμα για το αρχείο με τους μοναδικούς κόμβους από το δίκτυο πρωτεϊνών του *Homo Sapiens* από την βάση DIP, πήραμε τα εξής:

- Disease → OMIM_DISEASE
- Functional_Categories → COG_ONTOLOGY, UP_KEYWORDS, UP_SEQ_FEATURE
- Gene_Ontology → GOTERM_BP_DIRECT, GOTERM__CC_DIRTECT, GOTERM_MF_DIRECT
- Pathways → BBID, BIOCARTA, KEGG_PATHWAY
- Protein_Domains → INTERPRO, PIR_SUPERFAMILY, SMART

Για κάθε κατηγορία από τις 13 παραπάνω, προέκυπτε ένα αρχείο της μορφής που φαίνεται στην Εικόνα 21Α και Εικόνα 21Β.

A



```
Category        Term    Count   %       PValue  Genes   List Total      Pop Hits        Pop Total       Fold Enrichment Bonferroni      Benjamini       FDR
BIOCARTA        h_inflamPathway:Cytokines and Inflammatory Response     29      0.8590047393364929      1,09E+10        P01583, P09603, P13232, P01912, P01579, P04085, P04141, P01137, P05231, P01730, P08700, P05112, P01574, P05113, P01375, P01374, P04229, P35225, P20809, P29460, P09919, P10600, P10145, P01903, P40933, P61812, P60568, P22301, P29459, P01562  914     29      1625    1.7778993435448578      3.347302911679284E-4    3.347861427172973SE-4   2.811765969880569E-4
BIOCARTA        h_metPathway:Signaling of Hepatocyte Growth Factor Receptor     33      0.9774881516587678      2,09E+11        P36507, Q92918, P28482, P49023, Q06124, P12931, P46109, P62993, P08581, P20936, Q07889, P62834, P61224, P02568, P27986, P01112, P46108, P40763, P05412, P04049, P05556, P27361, Q13905, Q02750, P14210, P19419, P60484, Q14185, P45983, P48736, Q13153, P42336, P01100  914     36      1625    1.629741064916120       0.006361467325721071    0.0018529854240888041   0.0015562655359177213
BIOCARTA        h_pdgfPathway:PDGF Signaling Pathway     27      0.7997630331753554      2,17E+11        P29353, P19174, P19138, P42224, P62993, P20936, P17252, Q07889, P04085, P27986, P42229, P05771, P01112, P40763, P05412, P04049, P27361, P11831, Q02750, P16234, P19419, P68400, P23458, P45983, P48736, Q13233, P42336, P01100  914     28      1625    1.7144029384818256      0.006607441049415574    0.0018529854240888041   0.0015562655359177213
BIOCARTA        h_nfkbPathway:NF-kB Signaling Pathway    23      0.6812796208530806      2,99E+11        P01583, O00206, Q13158, Q15628, O15111, P21580, Q13546, P14778, Q9Y4K3, O43318, Q99558, P19438, P51617, P19838, Q9Y6K9, Q04206, Q99836, O14920, Q13233, Q15750, P25963, P01375, P20333  914     23      1625    1.7778993435448578      0.009096361917665163    0.0018529854240888041   0.0015562655359177213
BIOCARTA        h_raccycdPathway:Influence of Ras and Rho proteins on G1 to S Transition     26      0.7701421800947867      3.6333047531153024E-5   P24385, P28482, P11802, Q01094, P46527, Q9Y6K9, Q00534, P63000, P61586, O14920, P27986, P01112, P25963, P06400, P04049, O15111, P31749, P27361, P38936, P19838, Q04206, P24864, Q14186, P24941, Q13153, P42336  914     27      1625    1.7120512197098632      0.011056536708981213    0.0018529854240888041   0.0015562655359177213
BIOCARTA        h_egfPathway:EGF Signaling Pathway       26      0.7701421800947867      3.6333047531153024E-5   P29353, P19174, P19138, P42224, P62993, P20936, P17252, Q07889, P27986, P42229, P05771, P01112, P01133, P40763, P05412, P04049, P27361, P11831, Q02750, P19419, P68400, P23458, P45983, P00533, Q13233, P42336, P01100  914     27      1625    1.7120512197098632      0.011056536708981213    0.0018529854240888041   0.0015562655359177213
BIOCARTA        h_il6Pathway:IL 6 signaling pathway      21      0.6220379146919431      8.845441718448968E-5    P52333, P05412, P04049, P29353, P27361, Q06124, P11831, Q02750, P19138, P17676, P40189, P62993, O60674, P68400, P19419, Q07889, P23458, P05231, P08887, P01112, P01100, P40763  914     21      1625    1.7778993435448578      0.026705186973927875    0.0033833814573067302   0.002841598152051731
BIOCARTA        h_igf1Pathway:IGF-1 Signaling Pathway    21      0.6220379146919431      8.845441718448968E-5    P05412, P04049, Q14620, P05019, P29353, P27361, Q06124, P11831, Q02750, P19138, P35568, P62993, P20936, P68400, P19419, Q07889, P00069, P45983, P48736, P27986, P01112, P42336, P01100  914     21      1625    1.7778993435448578      0.026705186973927875    0.0033833814573067302   0.002841598152051731
BIOCARTA        h_il2rbPathway:IL-2 Receptor Beta Chain in T cell Activation     34      1.0071090047393365      1.3768615129552752E-4   P52333, Q9UKT9, Q92934, P29353, P22681, Q14901, P10415, P62937, P28482, P29350, Q01094, O15519, P46109, P62993, Q07889, P27986, P42229, P01112, P23443, P25445, P04049, P31749, P27361, P48023, P35568, P14784, P31785, Q07817, P43405, P01106, P23458, P01589, P48736, P42336, P01100  914     39      1625    1.5499635302690876      0.041259526817372880.0046770038843119   0.003928071889771186
BIOCARTA        h_nktPathway:Selective expression of chemokine receptors during T-cell polarization     26      0.7701421800947867      1.5857208706492585E-4   P49682, P51681, P32246, P32248, P42701, Q99665, P01579, P04141, P01137, P01730, P05112, P05113, P29965, P24394, P10747, P29460, P10147, P51679, P10600, P13236, P51677, P61073, P15260, Q53569, P61812, P60568, P29459  914     28      1625    1.6509065332916537      0.047368292778583310.0046770038843119   0.003928071889771186
BIOCARTA        h_tpoPathway:TPO Signaling Pathway       23      0.6812796208530806      1.6812759061277452E-4   P05412, P04049, P29353, P19174, P27361, Q02750, P19138, P42224, P62993, P20936, O60674, P68400, P17252, Q07889, P40238, P48736, P27986, P42229, P05771, P01112, P40225, P42336, P01100, P40763  914     24      1625    1.7038202042304889      0.0501501581018959      0.0046770038843119      0.003928071889771186
BIOCARTA        h_her2Pathway:Role of ERBB2 in Signal Transduction and Oncology  22      0.6516587677725119      2.7818183767383O6E-4    P04049, Q09472, Q15303, P29353, P04626, P28482, P27361, Q02750, Q86X55, P40189, P62993, Q07889, P21860, P00533, P05231, P03372, P48736, P27986, P08887, P01112, P42336, P40763  914     23      1625    1.7005993720863857      0.08161215130050603     0.007093636806826800.005957727690181206
BIOCARTA        h_vdrPathway:Control of Gene Expression by Vitamin D Receptor    24      0.7109004739336493      4.1165331131372772E-4   Q8TAQ2, Q09472, Q9Y6Q9, P19793, O43539, Q96GM5, Q92793, P61201, Q92831, Q15788, Q96K63, Q9UIG0, Q13547, Q15648, O96019, O75376, P51532, P11473, Q86X55, Q8TAW2, Q99873, O14497, Q92922, Q13111, Q15596, Q9Y5B9  914     26      1625    1.6411378555798686      0.11837796584181326     0.008252027430011195    0.006930624344813323
```

B

Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrich	Bonferron	Benjamini	FDR
BIOCARTA	h_inflamP	29	0.8590047393364929	1,09E+10	P01583, P(	914	29	1625	########	3,35E+11	3,35E+12	2,81E+11	
BIOCARTA	h_metPath	33	0.9774881516587678	2,09E+10	P36507, Q(	914	36	1625	########	0.0063614	0.0018529	0.0015562655359177213	
BIOCARTA	h_pdgfPat	27	0.7997630331753554	2,17E+11	P29353, P1	914	28	1625	########	0.0066074	0.0018529	0.0015562655359177213	
BIOCARTA	h_nfkbPat	23	0.6812796208530806	2,99E+11	P01583, O(	914	23	1625	########	0.0090963	0.0018529	0.0015562655359177213	
BIOCARTA	h_raccycd	26	0.7701421800947867	3,63E+11	P24385, P2	914	27	1625	########	0.0110565	0.0018529	0.0015562655359177213	
BIOCARTA	h_egfPath	26	0.7701421800947867	3,63E+11	P29353, P1	914	27	1625	########	0.0110565	0.0018529	0.0015562655359177213	
BIOCARTA	h_il6Pathw	21	0.6220379146919431	8,85E+10	P52333, P(	914	21	1625	########	0.0267051	0.0033833	0.002841598152051731	
BIOCARTA	h_igf1Path	21	0.6220379146919431	8,85E+10	P05412, P(	914	21	1625	########	0.0267051	0.0033833	0.002841598152051731	
BIOCARTA	h_il2rbPat	34	10.071.090.047.393.300	1,38E+12	P52333, Q(	914	39	1625	########	0.0412595	0.0046770	0.003928071889771186	
BIOCARTA	h_nktPath	26	0.7701421800947867	1,59E+12	P49682, P(	914	28	1625	########	0.0473682	0.0046770	0.003928071889771186	
BIOCARTA	h_tpoPath	23	0.6812796208530806	1,68E+12	P05412, P(	914	24	1625	########	0.0501501	0.0046770	0.003928071889771186	
BIOCARTA	h_her2Pat	22	0.6516587677725119	2,78E+11	P04049, Q(	914	23	1625	########	0.0816121	0.0070936	0.005957727690181206	
BIOCARTA	h_vdrPath	24	0.7109004739336493	4,12E+11	Q8TAQ2, C	914	26	1625	########	0.1183779	0.0082520	0.006930624344813323	5
BIOCARTA	h_il1rPath	29	0.8590047393364929	4,23E+11	Q9H0E2, P	914	33	1625	########	0.1213684	0.0082520	0.006930624344813323	5
BIOCARTA	h_il2Pathw	21	0.6220379146919431	4,58E+12	P52333, P(	914	22	1625	########	0.1309169	0.0082520	0.006930624344813323	5
BIOCARTA	h_cytokine	21	0.6220379146919431	4,58E+12	P01583, P(	914	22	1625	########	0.1309169	0.0082520	0.006930624344813323	5
BIOCARTA	h_aktPath	21	0.6220379146919431	4,58E+12	P01241, P(	914	22	1625	########	0.1309169	0.0082520	0.006930624344813323	5
BIOCARTA	h_keratinc	39	115.521.327.014.218	6,49E+11	P10415, P2	914	48	1625	########	0.1800971	0.0106079	0.008909283032175884	
BIOCARTA	h_g2Pathw	23	0.6812796208530806	6,59E+11	O96017, P(	914	25	1625	########	0.1825922	0.0106079	0.008909283032175884	
BIOCARTA	h_atmPath	20	0.5924170616113744	7,52E+10	P05412, O(	914	21	1625	########	0.2056939	0.0109618	0.009206561581461778	
BIOCARTA	h_insulinP	20	0.5924170616113744	7,52E+10	P05412, P(	914	21	1625	########	0.2056939	0.0109618	0.009206561581461778	
BIOCARTA	h_tcrPathv	36	1.066.350.710.900.470	9,13E+11	P02593, P(	914	44	1625	########	0.2437453	0.0126931	0.010660597190096165	
BIOCARTA	h_hivnefP	47	13.921.800.947.867.200	0.0012326	Q13158, Q	914	61	1625	########	0.3143842	0.0153742	0.012912373561907628	

Εικόνα 21. Α) Αρχείο μετά το Functional Annotation του *H.Sapiens* από την BIOCARTA.
Β) Αρχείο μετά το Functional Annotation του *H.Sapiens* από την BIOCARTA, μέσω του Microsoft Excel.

Όπως είναι κατανοητό τα αρχεία αυτά είναι πολύ μεγάλα και χρειάστηκε να φιλτραριστούν για να μπορέσουμε να τα χρησιμοποιήσουμε. Αρχικά κρατήσαμε τις στήλες Term, PValue, Genes. Από τα δεδομένα που απομένουν κρατήσαμε μόνο όσα έχουν PValue μικρότερο ή ίσο 0.05. Τέλος αφαιρέσαμε την επικεφαλίδα του αρχείου και την στήλη PValue, οπότε έμειναν μόνο τα δεδομένα από τις στήλες Term και Genes που διαχωρίζονταν από ένα tab. Τέλος μεταξύ των βιομορίων θα πρέπει να μην υπάρχουν κενά, όπως για παράδειγμα φαίνεται παρακάτω:

First Multivalent Nuclear Factor Q15797,Q14901,Q8N726,P04637,P36897

Το DAVID χρησιμοποιεί μια στατιστική βαθμολογία Kappa για τη μέτρηση των σχέσεων μεταξύ των όρων σχολιασμού βάσει των βαθμών των αντίστοιχων γονιδίων και ενός νέου αλγορίθμου ασαφούς ομαδοποίησης για να ομαδοποιεί παρόμοια, περιττά και ετερογενή περιεχόμενα σχολιασμών, προερχόμενα από την ίδια ή διαφορετική πηγή, σε ομάδες σχολιασμών (annotation groups) [90].

7.3.2. g:Profiler

Πέρα από τον εμπλουτισμό με τη χρήση του DAVID σε όλα τα αρχεία δικτύων, επαναλάβαμε την διαδικασία και με τη χρήση του g:Profiler. Συγκεκριμένα χρησιμοποιήσαμε τον παρακάτω κώδικα στη γλώσσα R:

```
library(gprofiler2)

args <- commandArgs(trailingOnly = TRUE)
j <- as.integer(args[1]) # index of sources below
inFile <- args[2]
organism <- args[3]
outFolder <- args[4]
inData <- read.table(inFile, header=F)

sources <- c("GO:MF", "GO:CC", "GO:BP", "KEGG", "REAC", "WP", "TF", "MIRNA", "CORUM", "HPA", "HP")
annotation_names <- c("GO_MF", "GO_CC", "GO_BP", "KEGG", "REAC", "WP", "TF", "MIRNA", "CORUM", "HPA", "HP")

cat(print(sources[j]))
outFile <- file(paste(outFolder, "/annotation_", annotation_names[j], ".txt", sep=""), "w")
gostres <- gost(as.list(inData), sources = sources[j], evcodes = T, organism = organism)
for (i in 1:length(gostres$result$intersection)){
  cat(sprintf("%s\t%s\n", gostres$result$term_id[i], gostres$result$intersection[i]), file = outFile)
}
close(outFile)
```

Όπως φαίνεται από τον κώδικα, οι βάσεις δεδομένων όπου βασίζεται το enrichment (sources) είναι οι παρακάτω:

- Gene Ontology → GO molecular function (MF), GO cellular component (CC), GO biological process (BP)
- Biological Pathways → KEGG, Reactome, WikiPathways
- Regulatory Motifs in DNA → TRANSFAC, miRTarBase
- Protein Databases → Human Protein Atlas, CORUM
- Human Phenotype Ontology → HP

Για να υλοποιηθεί επιτυχώς ο κώδικας, δέχεται, όπως και το DAVID, το αρχείο με τους μοναδικούς κόμβους και την συντομογραφία του ονόματος του αντίστοιχου οργανισμού, δηλαδή για τον ανθρώπινο οργανισμό (*Homo Sapiens*) το σύντομο όνομα είναι το *hsapiens* σύμφωνα με τη λίστα οργανισμών της σελίδας του g:Profiler [91].

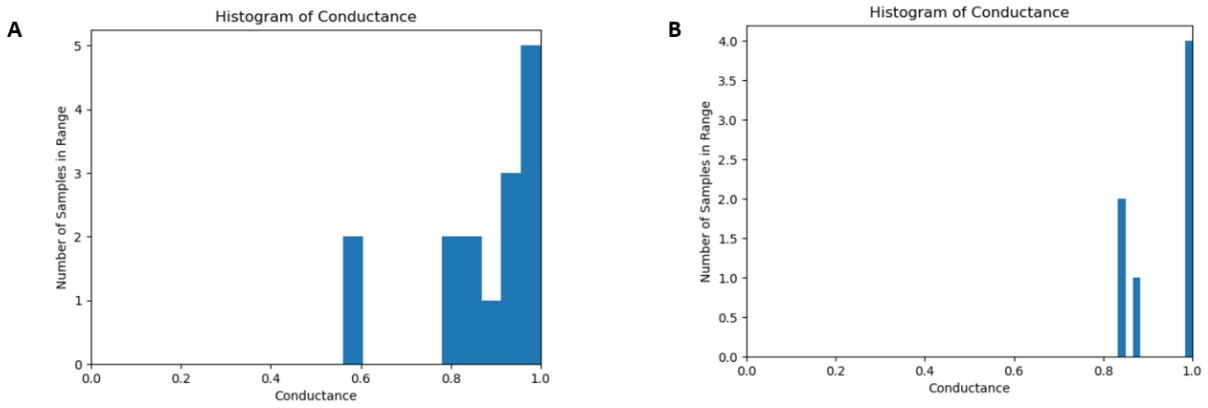
Το αρχείο του functional enrichment που προκύπτει είναι ακριβώς στη επιθυμητή μορφή που αναφέρθηκε στην ενότητα 7.2.1. Το g:Profiler αντιστοιχεί βιομόρια σε γνωστές λειτουργικές πηγές πληροφοριών και ανιχνεύει στατιστικά σημαντικά εμπλουτισμένους όρους [92].

Ουσιαστικά με την χρήση του DAVID και του g:Profiler, δημιουργήσαμε για το κάθε βιολογικό δίκτυο, αρχεία με συστάδες που αποτελούνται από τους κόμβους του κάθε δικτύου. Η συγκεκριμένη συσταδοποίηση, βασίζεται στη συσχέτιση των βιομορίων (πρωτεΐνες, γονίδια κλπ) με διάφορους βιολογικούς όρους/ρόλους.

7.4. Χρήση του Conductance

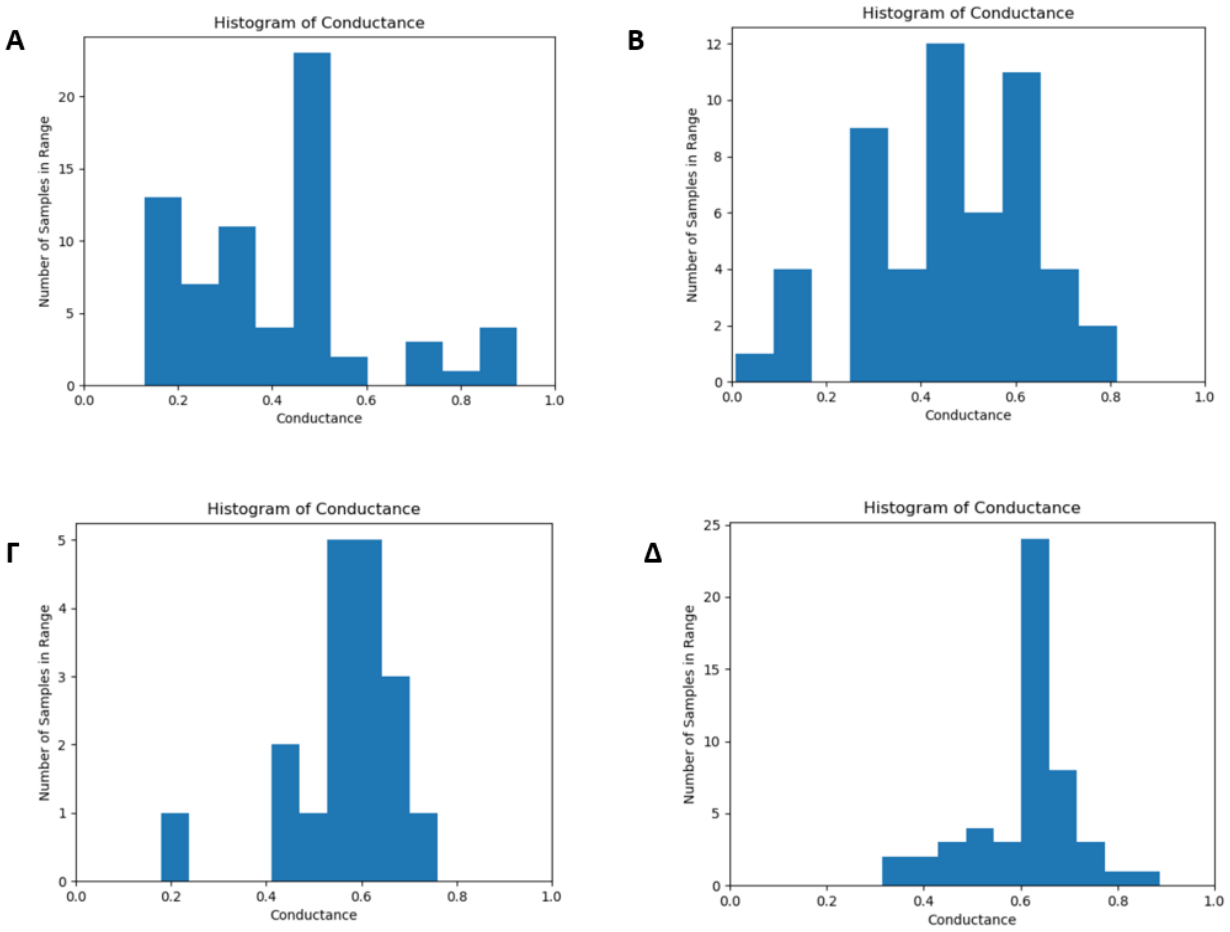
Η χρήση του conductance για το καθένα clustering αρχείο που προέκυπτε είτε από το DAVID είτε από το g:Profiler, μας βοήθησε να δούμε μέσω ιστογραμμάτων, την ποιότητα της κάθε συστάδας (cluster) στο υπόλοιπο αρχικό δίκτυο. Στον κώδικα του conductance, ο οποίος ήταν γραμμένος στην γλώσσα Python, δίναμε ως είσοδο δύο αρχεία για να μπορέσει να υπολογιστεί το conductance: το αρχικό δίκτυο (έχοντας αφαιρέσει μόνο την επικεφαλίδα “From To”) και κάποιο από τα αρχεία του functional enrichment, που έχουν τα βιομόρια ομαδοποιημένα. Τα αρχεία με τις ομάδες (ή αλλιώς τα ground truth αρχεία) που δώσαμε περισσότερη βαρύτητα, ήταν όσα βασίζονταν στις εξής βάσεις: GO-biological process (BP), GO-molecular function (MF) και KEGG Pathways.

Στην Εικόνα 22(A,B) φαίνονται τα ιστογράμματα που προέκυψαν από το πρωτεϊνικό δίκτυο του *Gallus gallus* και τα αντίστοιχα annotated αρχεία του από τις βάσεις GO-molecular function (MF) και KEGG. Στο γράφημα, οριζοντίως φαίνεται ο βαθμός του conductance για την κάθε ομάδα (cluster) ενώ καθέτως φαίνεται το πλήθος των ομάδων που έφερε το εκάστοτε conductance βαθμό. Η ιδανικότερη τιμή του conductance είναι μικρότερη ή ίση του 0.5. Στην Εικόνα 22A φαίνεται πως την μικρότερη τιμή conductance, αλλά μεγαλύτερη του 0.5, έχουν 2 μόνο ομάδες. Σε αυτή τη περίπτωση από τη συγκεκριμένη ομαδοποίηση, αφού δεν υπάρχει καμία ομάδα με την ιδανική τιμή conductance, καμία ομάδα δεν είναι “καλή” για το δίκτυο μας. Το ίδιο ισχύει και για την Εικόνα 22B.



Εικόνα 22. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα, για τον οργανισμό *Gallus gallus.*
Σύγκριση του αρχικού δικτύου PPI με το **A)** annotation αρχείο του GO-MF, **B)** annotation αρχείο του KEGG.

Στην Εικόνα 23(Α,Β,Γ,Δ) φαίνονται τα ιστογράμματα από το δίκτυο ομοιότητας αλληλουχιών του οργανισμού *Mus musculus* και τα αντίστοιχα annotated αρχεία του από τις βάσεις GO-biological process (BP), GO-molecular function (MF), KEGG και Reactome. Στην Εικόνα 23Α φαίνεται πως η πλειοψηφία των ομάδων έχει conductance μικρότερο του 0.5. Στην Εικόνα 23Β επίσης βλέπουμε πολλές συστάδες με τιμή conductance μικρότερη του 0.5, αλλά υπάρχει μία συστάδα με την τιμή του conductance να ανήκει στο εύρος [*0.0* , *0.1*]. Η συγκεκριμένη συστάδα φαίνεται να είναι η “καλύτερη” από όλες τις υπόλοιπες του συγκεκριμένου αρχείου. Στην Εικόνα 23Γ και στην Εικόνα 23Δ φαίνεται πως και εκεί υπάρχουν “καλές” συστάδες, αλλά το μεγαλύτερο μέρος των συστάδων παίρνει τιμές πάνω από 0.5.



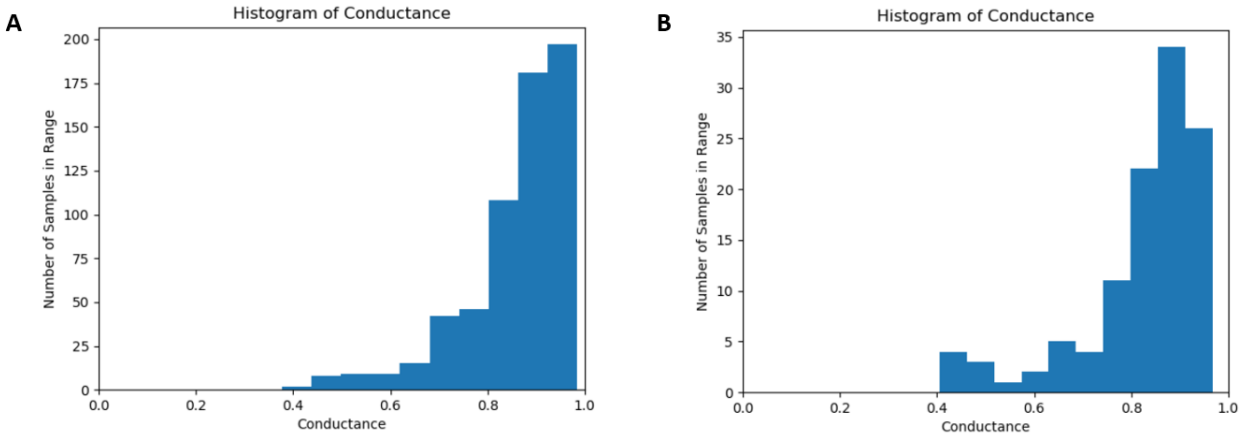
Εικόνα 23. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα, για τον οργανισμό *Mus Musculus.* Σύγκριση του αρχικού δικτύου SSN με το **Α)** annotation αρχείο του GO-BP, **Β)** annotation αρχείο του GO-MF, **Γ)** annotation αρχείο του KEGG, **Δ)** annotation αρχείο του Reactome.

7.5. Συμπεράσματα

Στα δίκτυα γονιδίων, τα σχολιασμένα (annotated) αρχεία που είχαν συστάδες με ικανοποιητικό conductance ήταν από τους οργανισμούς: *Anolis carolinensis, Drosophila melanogaster, Saccharomyces cerevisiae.* Η συσταδοποίηση αυτών των αρχείων, βασίστηκε στις βάσεις GO-biological process (BP), GO-molecular function (MF) και GO-cellular component (CC). Πιο συγκεκριμένα:

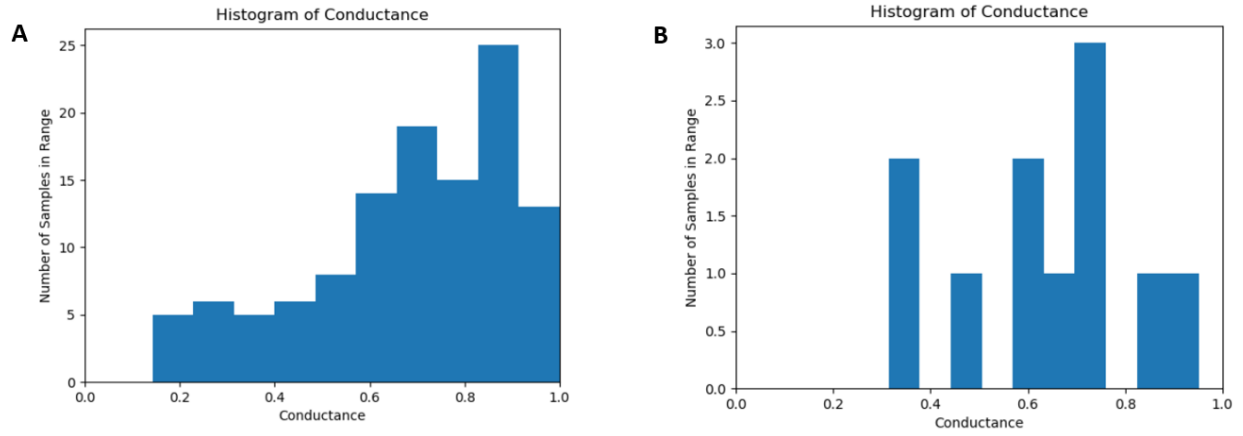
- *Anolis carolinensis* - GO-BP
- *Drosophila melanogaster* - GO-BP, GO-CC
- *Saccharomyces cerevisiae* - GO-BP, GO-CC, GO-MF

Όπως φαίνεται στην Εικόνα 24(A,B), όπου παρουσιάζονται τα ιστογράμματα του conductance για τα αρχεία GO-BP και GO-CC του *Drosophila melanogaster*, δυστυχώς είναι ελάχιστες οι συστάδες με conductance μικρότερο του 0.5. Το ίδιο ισχύει και για τα αρχεία των υπόλοιπων δύο οργανισμών.

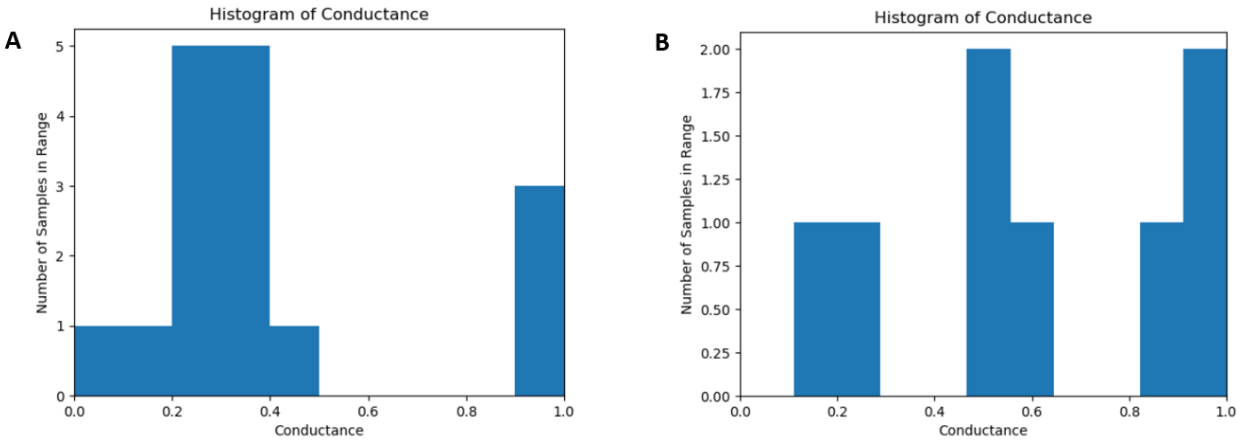


Εικόνα 24. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα, για τον οργανισμό *Drosophila melanogaster.*
Σύγκριση του αρχικού δικτύου GCN με το **A)** annotation αρχείο του GO-BP, **B)** annotation αρχείο του GO-CC.

Στα πρωτεϊνικά δίκτυα, από τα σχολιασμένα αρχεία με τις συστάδες, παρόλο που παρατηρήθηκαν συστάδες με χαμηλό conductance, όπως επιθυμούμε, λίγα ήταν τα αρχεία αυτά που έδωσαν μεγάλο πλήθος συστάδων με χαμηλό conductance. Για παράδειγμα, στον οργανισμό *Caenorhabditis elegans* (από τη βάση Biogrid), παρατηρήθηκε πως "καλό" conductance είχαν τα αρχεία που η ομαδοποίηση βασίστηκε στις βάσεις: GO-BP, GO-CC, GO-MF, KEGG, αλλά στις GO-CC και KEGG υπήρχαν αρκετές ομάδες με αρκετά καλό conductance, αφού ήταν όλο και πιο κοντά στο 0 (Εικόνα 25 A,B). Στην οργανισμό *Oryza Sativa,* επίσης σημειώθηκε αρκετά χαμηλό conductance στα: GO-BP, GO-CC, GO-MF, όπου ειδικά το GO-BP και GO-MF η πλειοψηφία των ομάδων είχαν πολύ "καλό" conductance (Εικόνα 26 A,B).

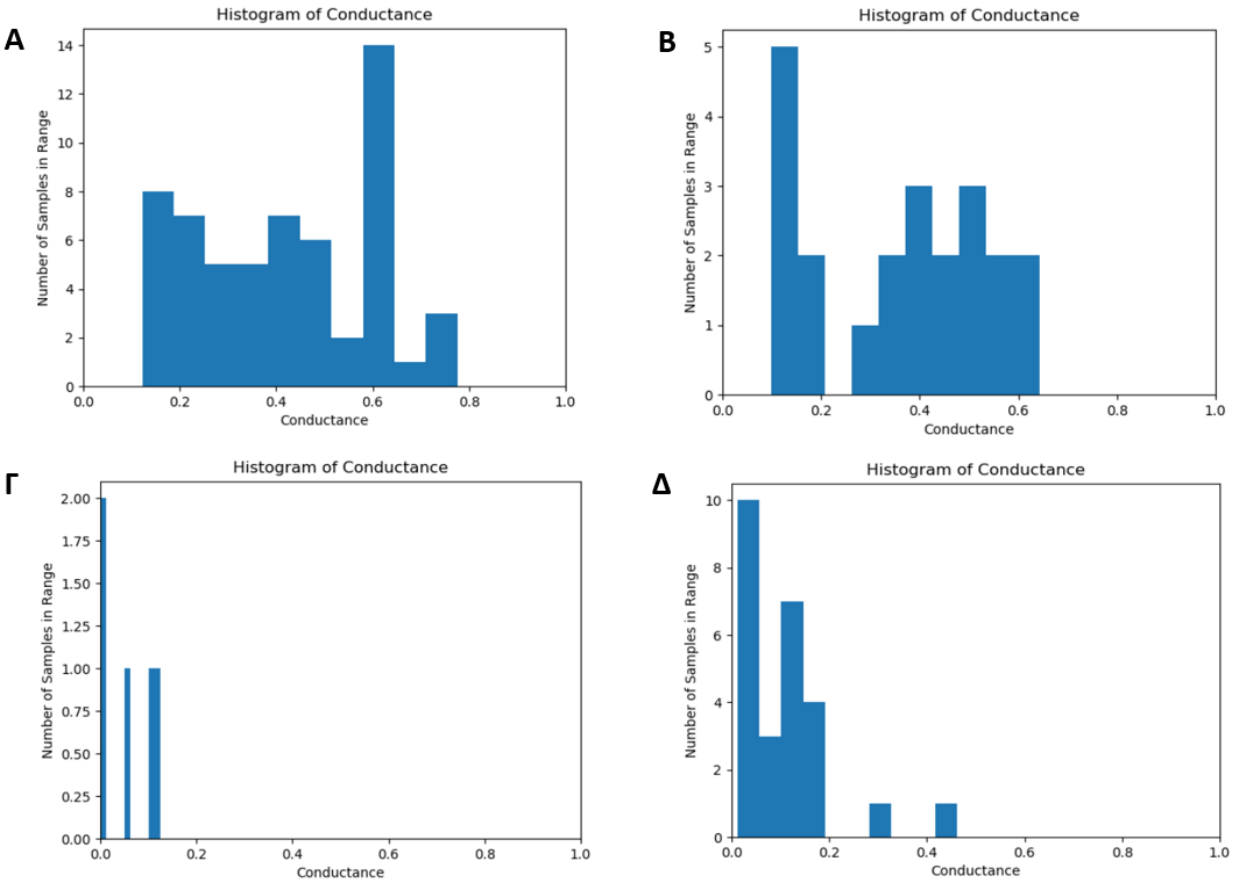


Εικόνα 25. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα, για τον οργανισμό *Caenorhabditis elegans.*
Σύγκριση του αρχικού δικτύου PPI με το **A)** annotation αρχείο του GO-CC, **B)** annotation αρχείο του KEGG.



Εικόνα 26. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα, για τον οργανισμό *Oryza Sativa*. Σύγκριση του αρχικού δικτύου PPI με το **A)** annotation αρχείο του GO-BP, **B)** annotation αρχείο του MF.

Από την άλλη, τα δίκτυα ομοιότητας αλληλουχιών φάνηκε να έχουν τις περισσότερες ομάδες των σχολιασμένων αρχείων τους με αρκετά χαμηλό conductance, πράγμα ιδιαίτερα ικανοποιητικό. Συγκεκριμένα αυτό το φαινόμενο παρατηρήθηκε στους οργανισμούς: *Arabidopsis Thaliana, Zea mays, Canis lupus familiaris, Drosophila melanogaster, Saccharomyces cerevisiae, Culicidae, Mus Musculus, Octopoda, Oryza Sativa, Ovis aries* και *Danio rerio* (Εικόνα 27 Α,Β,Γ,Δ).



Εικόνα 27. Αποτελέσματα του conductance αποτυπωμένα σε ιστογράμματα. Σύγκριση του αρχικών δικτύων SSN με το **Α)** annotation αρχείο του GO-MF για τον οργανισμό *Zea mays*, **Β)** annotation αρχείο του KEGG για τον οργανισμό *Drosophila melanogaster*, **Γ)** annotation αρχείο του GO-CC για τον οργανισμό *Culicidae*, **Δ)** annotation αρχείο του GO-BP για τον οργανισμό *Ovis Aries*.

8. Το Conductance σαν μέρος της εφαρμογής VICTOR

8.1. Εισαγωγή στο VICTOR

Όπως φάνηκε και στο Κεφάλαιο 5, υπάρχουν πολλοί αλγόριθμοι ομαδοποίησης δικτύων και ο καθένας μπορεί για το ίδιο δίκτυο να δώσει διαφορετική ομαδοποίηση, αφού ο κάθε αλγόριθμος διαφέρει από τον άλλο ή μπορεί ακόμα και ο ίδιος αλγόριθμος να φέρει διαφορετικά αποτελέσματα εφόσον δέχεται ως είσοδο παραμέτρους. Το διαδικτυακό εργαλείο ***VICTOR*** προσφέρεται για την σύγκριση των αποτελεσμάτων των διάφορων ομαδοποιήσεων ενός δικτύου, μέσω ενός διαδραστικού περιβάλλοντος οπτικής ανάλυσης των αλγορίθμων ομαδοποίησης.

8.2. Λειτουργικότητα του VICTOR

Μέσω του VICTOR, ο χρήστης έχει την ευχέρεια να επιλέξει ανάμεσα σε μια ποικιλία γραφημάτων και συγκεκριμένα: ραβδογραμματα (bar plots), ιστογράμματα (histograms), δίκτυα (networks), διαγράμματα ροής Sankey (sankey plots), διαγράμματα χορδών/διαγράμματα ακτινικού δικτύου (circos plots) και ιεραρχικοί χάρτες θερμότητας (hierarchical heatmaps). Το εργαλείο είναι γραμμένο στις γλώσσες R, Shiny και JavaScript, και οι υπολογισμοί βασίστηκαν στην βιβλιοθήκη mClustComp.

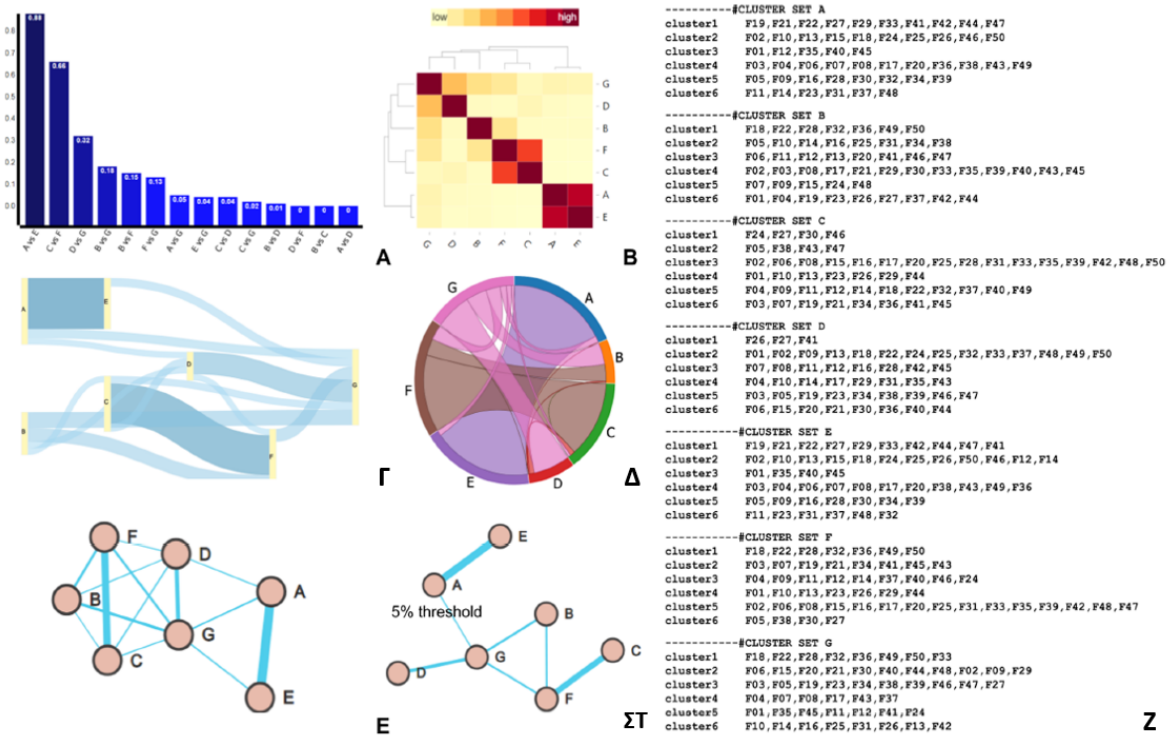
Το VICTOR, υποστηρίζει 10 τύπους μετρικών ώστε να συγκρίνει τις διάφορες ομάδες. Οι 10 τύποι μετρικών είναι οι εξής:

1. Adjusted Rand Index
2. Jaccard Index
3. Overlap coefficient
4. Wallace criteria type 1
5. Wallace criteria type 2
6. Fowlkes-Mallows Index
7. Maximum-Match Measure
8. Normalized Mutual information by Strehl & Ghosh
9. Normalized Mutual information by Fred & Jain
10. Normalized Variation of Information

Οι μετρικές αυτές χωρίζονται σε τρεις κατηγορίες: i) Counting Pairs (μετρικές 1-4), ii) Set Overlaps/Matching (μετρική 5) και iii) Mutual Information (μετρικές 6-10).

Το VICTOR μπορεί να δεχτεί ως είσοδο πολλαπλά αρχεία ομαδοποίησης (clustering files) στη μορφή που περιγράφηκε παραπάνω και μετά το upload τους, εμφανίζονται σε γραφήματα τα σχετικά στατιστικά δεδομένα όπως είναι ο αριθμός των ομάδων και ο αριθμός των μελών της κάθε ομάδας. Ο χρήστης θα πρέπει να προσέξει πως το κάθε μέλος μιας ομάδας θα πρέπει να ανήκει μόνο σε μία ομάδα, για αυτό και χρειάζονται οι διεργασίες που αναφέρθηκαν λεπτομερώς στο προηγούμενο κεφάλαιο. Επίσης, όλες οι ομάδες μεταξύ των εισαγόμενων αρχείων θα πρέπει να έχουν ακριβώς τα ίδια στοιχεία, πράγμα που μπορεί να φανεί στο tab “Compare Clusterings”. Σε αντίθετη περίπτωση, για να μπορέσει ο χρήστης να συνεχίσει την σύγκριση, μπορεί να χρησιμοποιήσει κάποια φίλτρα από το tab “File Handling”. Με αυτό τον τρόπο, ο χρήστης μπορεί να οριοθετήσει τον αριθμό των αντικειμένων της κάθε ομάδας,

διεργασία που βοηθάει πέρα από το να είναι όλες οι ομάδες ίσες, να γίνει επίσης απαλοιφή των ομάδων με μόνο ένα μέλος (Εικόνα 28 A-Z).



Εικόνα 28. Τα οπτικοποιημένα αποτελέσματα στο VICTOR. Η σύγκριση βασίζεται σε ένα δείγμα από δεδομένα και χρησιμοποιήθηκε για τις συγκρίσεις η μετρική Adjusted Rand Index **Α)** Bar plot, **Β)** Hierarchical Heatmap **Γ)** Sankey Plot **Δ)** Circos Plot, **Ε)** Οπτικοποίηση δικτύου (Network) μετά την εφαρμογή ενός force-directed layout, **ΣΤ)** Οπτικοποίηση δικτύου (Network) μετά την εφαρμογή ορίου 5% στις τιμές των ακμών του δικτύου, **Ζ)** Δείγμα των ομαδοποιημένων δεδομένων που μπορούν να εισαχθούν στο εργαλείο. [93]

Για την οπτικοποίηση της σύγκρισης μεταξύ των αρχείων χρησιμοποιούνται τα προαναφερθέντα γραφήματα που αναλύονται παρακάτω [93]:

Sankey plots - περιγράφουν διαγράμματα ροής όπου το πλάτος της ακμής/βέλους είναι ανάλογο με τον ρυθμό της ροής. Στον τομέα της Βιολογίας χρησιμοποιούνται κυρίως για να σχεδιαστούν για παράδειγμα ριψοκίνδυνες ομάδες έναντι κλινικών υπο-τύπων καρκίνου. Στο VICTOR, τα κάθετα ορθογώνια αντιστοιχούν στα αρχεία ομαδοποίησης τα οποία συνδέονται από μη κατευθυνόμενες λωρίδες, το ύψος των οποίων είναι ανάλογο με την ομοιότητα ανά ζεύγη, συγκριτικά με τις άλλες τιμές ομοιότητας.

Heatmap - πρόκειται για δισδιάστατο χρωματιστό πίνακα ($n \times n$) που κάθε απόχρωση του κόκκινου χρώματος του, αναφέρεται στην "έντασης" της τιμής ενός ζεύγους που συγκρίνεται. Για τους άξονες του πίνακα υπάρχει ένα δενδρόγραμμα παρουσιάζοντας την διάταξη των ομάδων των ενδιάμεσων όμοιων ζευγών.

Bar charts - αντιπροσωπεύουν τα δεδομένα ως ορθογώνια των οποίων το ύψος είναι ανάλογο με την αντίστοιχη τιμή του άξονα y. Ο άξονας x αφορά τα παρατηρούμενα αναγνωριστικά των δεδομένων. Κάθε

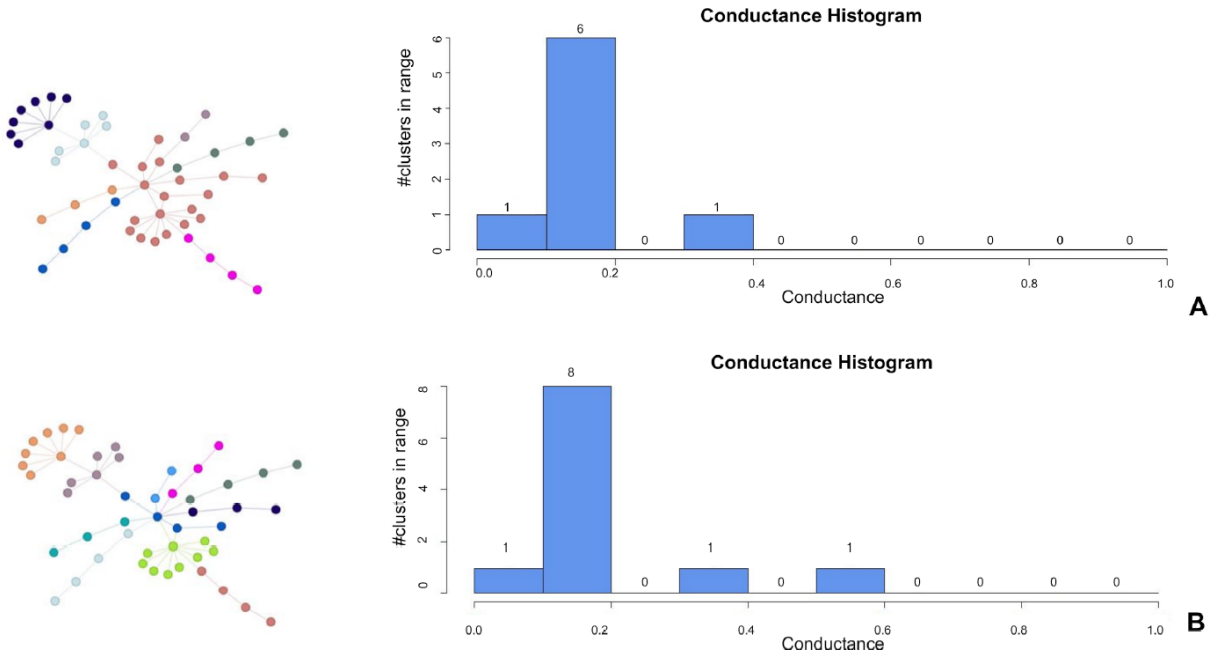
γραφική παράσταση είναι η ομοιότητα των ζευγών μεταξύ των αρχείων ομαδοποίησης σύμφωνα με την μετρική που επιλέγει ο χρήστης.

Networks - όπως έχει ήδη αναφερθεί στην συγκεκριμένη διπλωματική εργασία, η οπτικοποίηση των δεδομένων με χρήση δικτύων γίνεται με χρήση κόμβων (δεδομένα) και ακμών (τιμές ζευγών).

Circos plots - τοποθετώντας τα δεδομένα κυκλικά, κατά μήκος ενός τόξου, του οποίου το μήκος του είναι ανάλογο με το άθροισμα των συνολικών αλληλεπιδράσεων ενός αντικειμένου. Οι συνδέσεις των ζευγών αναπαρίστανται ως "κορδέλες"/χορδές, χρωματιστές σύμφωνα με ένα από τα δύο αντικείμενα που αλληλεπιδρούν. Στο VICTOR, κατά μήκος του τόξου βρίσκονται τα δεδομένα των αρχείων ομαδοποίησης, και οι ομοιότητες των ζευγών αναπαρίστανται από τις χρωματιστές "κορδέλες".

8.3. Conductance στο VICTOR

Στο εργαλείο VICTOR, υπάρχει ένα tab, όπου μπορεί να υπολογιστεί το conductance των ομάδων ενός δικτύου. Ο χρήστης, αφού μεταφορτώσει ένα αρχείο ομαδοποίησης στο tab "File Handling", μπορεί στο tab "Conductance" να ανεβάσει το αντίστοιχο μη-κατευθυνόμενο δίκτυο στη μορφή που περιεγράφηκε στο Κεφάλαιο 7. Στη περίπτωση που ο γράφος είναι σταθμισμένος, προστίθεται στο αρχείο μία επιπλέον στήλη από δεξιά που αντιστοιχεί στα βάρη. Αρχικά σε μορφή γράφου εμφανίζεται στη παράθυρο "Network" το δίκτυο έχοντας χρωματισμένους με κοινό χρώμα τους κόμβους που ανήκουν στην ίδια ομάδα, ενώ στο παράθυρο "Conductance" εμφανίζεται το αντίστοιχο ιστόγραμμα (Εικόνα 29 A, B) [93].



Εικόνα 29. Ο υπολογισμός του conductance στο VICTOR. Υπολογίστηκε το conductance για δύο διαφορετικά αρχεία ομαδοποίησης ενός δειγματικού δικτύου. Από την δεξιά πλευρά φαίνεται η οπτικοποίηση σε μορφή δικτύου με κοινό χρώμα μεταξύ των κόμβων που ανήκουν στην ίδια ομάδα, ενώ από αριστερά είναι το αντίστοιχο ιστόγραμμα conductance. **Α)** Ο αλγόριθμος που χρησιμοποιήθηκε για την ομαδοποίηση ήταν ο Label Propagation, **Β)** Ο αλγόριθμος που χρησιμοποιήθηκε για την ομαδοποίηση ήταν ο Walktrap. [93]

Βιβλιογραφία

[1] M. Koutrouli, E. Karatzas, D. Paez-Espino, and G. A. Pavlopoulos, “A Guide to Conquer the Biological Network Era Using Graph Theory,” *Front. Bioeng. Biotechnol.*, vol. 8, p. 34, Jan. 2020, doi: 10.3389/fbioe.2020.00034.

[2] B. Bollobás and B. Béla, *Random Graphs*. Cambridge University Press, 2001.

[3] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998, doi: 10.1038/30918.

[4] null Barabasi and null Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999, doi: 10.1126/science.286.5439.509.

[5] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, “The large-scale organization of metabolic networks,” *Nature*, vol. 407, no. 6804, pp. 651–654, Oct. 2000, doi: 10.1038/35036627.

[6] W. Gao, H. Wu, M. K. Siddiqui, and A. Q. Baig, “Study of biological networks using graph theory,” *Saudi J. Biol. Sci.*, vol. 25, no. 6, pp. 1212–1219, Sep. 2018, doi: 10.1016/j.sjbs.2017.11.022.

[7] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, “Protein-protein interaction networks (PPI) and complex diseases,” *Gastroenterol. Hepatol. Bed Bench*, vol. 7, no. 1, pp. 17–31, 2014.

[8] V. Spirin and L. A. Mirny, “Protein complexes and functional modules in molecular networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 21, pp. 12123–12128, Oct. 2003, doi: 10.1073/pnas.2032324100.

[9] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets,” *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D535-539, Jan. 2006, doi: 10.1093/nar/gkj109.

[10] G. D. Bader, D. Betel, and C. W. V. Hogue, “BIND: the Biomolecular Interaction Network Database,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248–250, Jan. 2003, doi: 10.1093/nar/gkg056.

[11] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, “DIP: the database of interacting proteins,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 289–291, Jan. 2000, doi: 10.1093/nar/28.1.289.

[12] H. Hermjakob *et al.*, “IntAct: an open source molecular interaction database,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D452-455, Jan. 2004, doi: 10.1093/nar/gkh052.

[13] A. R. Ekre and R. V. Mante, “Genome sequence alignment tools: A review,” in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Feb. 2016, pp. 677–681. doi: 10.1109/AEEICB.2016.7538378.

[14] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[15] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, “Adaptive seeds tame genomic sequence comparison,” *Genome Res.*, vol. 21, no. 3, pp. 487–493, Mar. 2011, doi: 10.1101/gr.113985.110.

[16] “Flexible sequence similarity searching with the FASTA3 program package - PubMed.” https://pubmed.ncbi.nlm.nih.gov/10547837/ (accessed Sep. 23, 2020).

[17] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function,” *Mol. Syst. Biol.*, vol. 3, p. 88, 2007, doi: 10.1038/msb4100129.

[18] A. Vázquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabási, “The topological relationship between the large-scale attributes and local interaction patterns of complex networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 52, pp. 17940–17945, Dec. 2004, doi: 10.1073/pnas.0406024101.

[19] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[20] “GTRD: a database on gene transcription regulation-2019 update - PubMed.”

https://pubmed.ncbi.nlm.nih.gov/30445619/ (accessed Sep. 23, 2020).
[21] V. Matys *et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 374–378, Jan. 2003, doi: 10.1093/nar/gkg108.
[22] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam, "Reconstruction of cellular signalling networks and analysis of their properties," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 2, pp. 99–111, Feb. 2005, doi: 10.1038/nrm1570.
[23] J. Gagneur, D. B. Jackson, and G. Casari, "Hierarchical analysis of dependency in metabolic networks," *Bioinforma. Oxf. Engl.*, vol. 19, no. 8, pp. 1027–1034, May 2003, doi: 10.1093/bioinformatics/btg115.
[24] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Brief. Bioinform.*, vol. 19, no. 4, pp. 575–592, Jul. 2018, doi: 10.1093/bib/bbw139.
[25] T. Barrett *et al.*, "NCBI GEO: archive for functional genomics data sets--update," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D991-995, Jan. 2013, doi: 10.1093/nar/gks1193.
[26] H. Parkinson *et al.*, "ArrayExpress--a public database of microarray experiments and gene expression profiles," *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D747-750, Jan. 2007, doi: 10.1093/nar/gkl995.
[27] T. Obayashi, Y. Okamura, S. Ito, S. Tadaka, I. N. Motoike, and K. Kinoshita, "COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D1014–D1020, Jan. 2013, doi: 10.1093/nar/gks1014.
[28] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.
[29] "Characterization of Reticulate Networks Based on the Coalescent with Recombination." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2582979/ (accessed Oct. 06, 2020).
[30] D. A. Morrison, "Phylogenetic Networks: A Review of Methods to Display Evolutionary History," *Annu. Res. Rev. Biol.*, pp. 1518–1543, Jan. 2014, doi: 10.9734/ARRB/2014/8230.
[31] E. Gross, C. Long, and J. Rusinko, "Phylogenetic Networks," *ArXiv190601586 Q-Bio*, Jun. 2019, Accessed: Oct. 06, 2020. [Online]. Available: http://arxiv.org/abs/1906.01586
[32] Z. Yang, "Phylogenetic analysis using parsimony and likelihood methods," *J. Mol. Evol.*, vol. 42, no. 2, pp. 294–307, Feb. 1996, doi: 10.1007/BF02198856.
[33] D. H. Huson and D. Bryant, "Application of phylogenetic networks in evolutionary studies," *Mol. Biol. Evol.*, vol. 23, no. 2, pp. 254–267, Feb. 2006, doi: 10.1093/molbev/msj030.
[34] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, "Dendroscope: An interactive viewer for large phylogenetic trees," *BMC Bioinformatics*, vol. 8, no. 1, p. 460, Dec. 2007, doi: 10.1186/1471-2105-8-460.
[35] K. P. Schliep, "Estimating phylogenetic trees with phangorn," p. 14.
[36] L. Danon *et al.*, "Networks and the epidemiology of infectious disease," *Interdiscip. Perspect. Infect. Dis.*, vol. 2011, p. 284909, 2011, doi: 10.1155/2011/284909.
[37] T. C. Ings *et al.*, "Review: Ecological networks – beyond food webs," *J. Anim. Ecol.*, vol. 78, no. 1, pp. 253–269, 2009, doi: 10.1111/j.1365-2656.2008.01460.x.
[38] G. A. Pavlopoulos, P. I. Kontou, A. Pavlopoulou, C. Bouyioukos, E. Markou, and P. G. Bagos, "Bipartite graphs in systems biology and medicine: a survey of methods and applications," *GigaScience*, vol. 7, no. 4, pp. 1–31, 01 2018, doi: 10.1093/gigascience/giy014.
[39] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D355-360, Jan. 2010, doi: 10.1093/nar/gkp896.
[40] G. R. Mishra *et al.*, "Human protein reference database--2006 update.," *Nucleic Acids Res.*, vol. 34, no. Database issue, Jan. 2006, Accessed: Jul. 13, 2020. [Online]. Available: https://jhu.pure.elsevier.com/en/publications/human-protein-reference-database-2006-update-6

[41] Z.-X. Huang, H.-Y. Tian, Z.-F. Hu, Y.-B. Zhou, J. Zhao, and K.-T. Yao, "GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords," *BMC Bioinformatics*, vol. 9, no. 1, p. 308, Jul. 2008, doi: 10.1186/1471-2105-9-308.
[42] A. Franceschini *et al.*, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D808-815, Jan. 2013, doi: 10.1093/nar/gks1094.
[43] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D380-384, Jan. 2016, doi: 10.1093/nar/gkv1277.
[44] A. Gioutlakis, M. I. Klapa, and N. K. Moschonas, "PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology," *PloS One*, vol. 12, no. 10, p. e0186039, 2017, doi: 10.1371/journal.pone.0186039.
[45] M. Koutrouli, P. Hatzis, and G. A. Pavlopoulos, "Exploring Networks in the STRING and Reactome Database," in *Systems Medicine*, Elsevier, 2021, pp. 507–520. doi: 10.1016/B978-0-12-801238-3.11516-8.
[46] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Mol. Biol.*, vol. 10, no. 12, Art. no. 12, Dec. 2003, doi: 10.1038/nsb1203-980.
[47] T. UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 46, no. 5, pp. 2699–2699, Mar. 2018, doi: 10.1093/nar/gky092.
[48] D. R. Zerbino *et al.*, "Ensembl 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D754–D761, 04 2018, doi: 10.1093/nar/gkx1098.
[49] "BioGRID interaction database: 2019 update | Nucleic Acids Research | Oxford Academic." https://academic.oup.com/nar/article/47/D1/D529/5204333 (accessed Oct. 28, 2020).
[50] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D449–D451, Jan. 2004, doi: 10.1093/nar/gkh086.
[51] S. Kerrien *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D841–D846, Jan. 2012, doi: 10.1093/nar/gkr1088.
[52] "COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH) - PubMed." https://pubmed.ncbi.nlm.nih.gov/27679477/ (accessed Mar. 04, 2021).
[53] "Coexpedia." https://www.coexpedia.org/tutorial.php (accessed Mar. 22, 2021).
[54] D. Warde-Farley *et al.*, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W214-220, Jul. 2010, doi: 10.1093/nar/gkq537.
[55] "Help · GeneMANIA." http://pages.genemania.org/help/ (accessed Mar. 22, 2021).
[56] T. Obayashi, Y. Kagaya, Y. Aoki, S. Tadaka, and K. Kinoshita, "COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D55–D62, Jan. 2019, doi: 10.1093/nar/gky1155.
[57] "Overview | COXPRESdb." https://coxpresdb.jp/static/overview.shtml (accessed Mar. 22, 2021).
[58] S. van Dam, T. Craig, and J. P. de Magalhães, "GeneFriends: a human RNA-seq-based gene and transcript co-expression database," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1124-1132, Jan. 2015, doi: 10.1093/nar/gku1042.
[59] H. J. Atkinson, J. H. Morris, T. E. Ferrin, and P. C. Babbitt, "Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies," *PLoS ONE*, vol. 4, no. 2, Feb. 2009, doi: 10.1371/journal.pone.0004345.
[60] P. Raina, I. Lopes, K. Chatsirisupachai, Z. Farooq, and J. P. de Magalhães, "GeneFriends 2021:

Updated co-expression databases and tools for human and mouse genes and transcripts," Genomics, preprint, Jan. 2021. doi: 10.1101/2021.01.10.426125.
[61] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.
[62] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.
[63] "Gene Ontology overview," *Gene Ontology Resource*. http://geneontology.org/docs/ontology-documentation/ (accessed Jan. 28, 2021).
[64] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, "KEGG: integrating viruses and cellular organisms," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D545–D551, Jan. 2021, doi: 10.1093/nar/gkaa970.
[65] "Searching KEGG." https://www.genome.jp/kegg/kegg1d.html (accessed Mar. 22, 2021).
[66] "BRITE Functional Hierarchies." https://www.genome.jp/kegg/kegg3b.html (accessed Mar. 22, 2021).
[67] "KEGG Mapping." https://www.genome.jp/kegg/kegg1b.html (accessed Mar. 22, 2021).
[68] X. Jiao *et al.*, "DAVID-WS: a stateful web service to facilitate gene/protein list analysis," *Bioinformatics*, vol. 28, no. 13, pp. 1805–1806, Jul. 2012, doi: 10.1093/bioinformatics/bts251.
[69] U. Raudvere *et al.*, "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.
[70] B. Jassal *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D498–D503, Jan. 2020, doi: 10.1093/nar/gkz1031.
[71] H. Mi *et al.*, "PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D394–D403, Jan. 2021, doi: 10.1093/nar/gkaa1106.
[72] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W199–W205, Jul. 2019, doi: 10.1093/nar/gkz401.
[73] P. Khatri, M. Sirota, and A. J. Butte, "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges," *PLoS Comput. Biol.*, vol. 8, no. 2, Feb. 2012, doi: 10.1371/journal.pcbi.1002375.
[74] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.
[75] J. Wang *et al.*, "Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction," *Mol. Cell. Proteomics MCP*, vol. 16, no. 1, pp. 121–134, Jan. 2017, doi: 10.1074/mcp.M116.060301.
[76] P. Jiang and M. Singh, "SPICi: a fast clustering algorithm for large biological networks," *Bioinformatics*, vol. 26, no. 8, pp. 1105–1111, Apr. 2010, doi: 10.1093/bioinformatics/btq078.
[77] K. Krishna and M. Narasimha Murty, "Genetic K-means algorithm," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 29, no. 3, pp. 433–439, Jun. 1999, doi: 10.1109/3477.764879.
[78] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.
[79] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.
[80] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven Label Propagation Algorithm," in *2011 IEEE Network Science Workshop*, Jun. 2011, pp. 188–195. doi: 10.1109/NSW.2011.6004645.

[81] P. Pons and M. Latapy, “Computing Communities in Large Networks Using Random Walks,” in *Computer and Information Sciences - ISCIS 2005*, Berlin, Heidelberg, 2005, pp. 284–293. doi: 10.1007/11569596_31.
[82] “An algorithm Walktrap-SPM for detecting overlapping community structure | International Journal of Modern Physics B.” https://www.worldscientific.com/doi/10.1142/S0217979217501211 (accessed May 13, 2021).
[83] D. Gleich, “MS&E 337 - Information Networks,” p. 24.
[84] S. Galhotra, A. Bagchi, and S. Bedathur, “Tracking the Conductance of Rapidly Evolving Topic-Subgraphs,” p. 12.
[85] J. Leskovec, K. J. Lang, and M. W. Mahoney, “Empirical Comparison of Algorithms for Network Community Detection,” *ArXiv10043539 Phys.*, Apr. 2010, Accessed: Jun. 02, 2021. [Online]. Available: http://arxiv.org/abs/1004.3539
[86] B. Mohar, “Isoperimetric numbers of graphs,” *J. Comb. Theory Ser. B*, vol. 47, no. 3, pp. 274–291, Dec. 1989, doi: 10.1016/0095-8956(89)90029-4.
[87] R. R. Montenegro and P. Tetali, *Mathematical Aspects of Mixing Times in Markov Chains*. Now Publishers Inc, 2006.
[88] S. O. Gharan, “Lecture 17: Cheeger’s Inequality and the Sparsest Cut Problem,” p. 9.
[89] “networkx.algorithms.boundary — NetworkX 1.10 documentation.” https://networkx.org/documentation/networkx-1.10/_modules/networkx/algorithms/boundary.html#edge_boundary (accessed May 19, 2021).
[90] “DAVID Bioinformatics Resources.” https://david.ncifcrf.gov/content.jsp?file=functional_annotation.html#geneenrich (accessed May 31, 2021).
[91] “g:Profiler – a web server for functional enrichment analysis and conversions of gene lists.” https://biit.cs.ut.ee/gprofiler/page/organism-list (accessed May 31, 2021).
[92] “g:Profiler – a web server for functional enrichment analysis and conversions of gene lists.” https://biit.cs.ut.ee/gprofiler/gost (accessed May 31, 2021).
[93] E. Karatzas et al., “VICTOR: A visual analytics web application for comparing cluster sets,” Comput. Biol. Med., vol. 135, p. 104557, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104557.

Παράρτημα - Δημοσιεύσεις

Δημοσίευση

VICTOR: A visual analytics web application for comparing cluster sets

Evangelos Karatzas, Maria Gkonta, Joana Hotova, Fotis A. Baltoumas, Panagiota I. Kontou, Christopher J. Bobotsis, Pantelis G. Bagos, Georgios A. Pavlopoulos

Computers in Biology and Medicine, Elsevier, 4 June 2021,
https://www.sciencedirect.com/science/article/abs/pii/S0010482521003516?via%3Dihub

Preprint

VICTOR: A visual analytics web application for comparing cluster sets

Evangelos Karatzas, Maria Gkonta, Joana Hotova, Fotis A. Baltoumas, Panagiota I. Kontou, Christopher J. Bobotsis, Pantelis G. Bagos, Georgios A. Pavlopoulos

Biorxiv, 23 March 2021, https://doi.org/10.1101/2021.03.22.436502

Εφαρμογή VICTOR
http://bib.fleming.gr:3838/VICTOR/

Πηγαίος Κώδικας VICTOR
https://github.com/PavlopoulosLab/VICTOR