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 αξιολόγηση αλγορίθμων ομαδοποίησης εφαρμοσμένοι σε βιολογικά δίκτυα»

Ανδρομάχη Παμπάλου

Πτυχιούχος Πληροφορικής με Εφαρμογές στην Βιοϊατρική,
Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ 2020

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS»

Master Diploma Thesis

«Benchmarking of clustering algorithms in biological networks»

Andromachi Pampalou

Graduate of Department of Computer Science and Biomedical Informatics, University of Thessaly

A T H E N S 2 0 2 0

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 αξιολόγηση αλγορίθμων ομαδοποίησης εφαρμοσμένοι σε βιολογικά δίκτυα»

Τριμελής εξεταστική επιτροπή

Καθηγητής Παντελής Μπάγκος (Επιβλέπων)
Εργαστήριο Μοριακής & Υπολογιστικής Βιολογίας και Γενετικής,
Τμήμα Πληροφορικής με Εφαρμογές στην Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας

Καθηγητής Ιωάννης Τρουγκάκος
Τομέας Βιολογίας Κυττάρου και Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Αναπληρωτής Καθηγητής Αλέξανδρος Γεωργακίλας
Τομέας Φυσικής,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο εργαστήριο του Δρ. Γεώργιου Παυλόπουλου, Ερευνητή Β' του Ερευνητικού Κέντρου Αλέξανδρος Φλέμινγκ, με την συνεπίβλεψη του Καθηγητή Παντελή Μπάγκου του Τμήματος Πληροφορικής με Εφαρμογές στην Βιοϊατρική του Πανεπιστημίου Θεσσαλίας , τους οποίους ευχαριστώ θερμά για την ευκαιρία που μου έδωσαν να ασχοληθώ με ένα θέμα ιδιαίτερου ενδιαφέροντος, καθώς και για όλη την βοήθεια και καθοδήγηση τους κατά την διάρκεια πραγματοποίησης της εργασίας. Ιδιαίτερες ευχαριστίες εκφράζω και προς τα υπόλοιπα μέλη της Τριμελούς Εξεταστικής Επιτροπής, τον Καθηγητή Ιωάννη Τρουγκάκο και τον Αναπληρωτή Καθηγητή Αλέξανδρο Γεωργακίλα, για τον χρόνο και την γνώση που προσέφεραν στην ολοκλήρωση της μελέτης. Τέλος είμαι ευγνώμων στην οικογένεια μου, τους φίλους μου και σε όλους όσους με στηρίζουν καθημερινά και με προτρέπουν να κυνηγάω τους στόχους μου.

Περιεχόμενα

Περίληψη

Η λειτουργία ενός οργανισμού σε μοριακό επίπεδο εξαρτάται σε μεγάλο βαθμό από τις αλληλεπιδράσεις των μακρομορίων αυτού. Η δημιουργία συμπλόκων πρωτεϊνών αποτελούν βασική μονάδα, υπεύθυνη για ποικίλους βιολογικούς μηχανισμούς μέσα στο κύτταρο. Ο προσδιορισμός, επομένως, των συγκεκριμένων αλληλεπιδράσεων καθίσταται απαραίτητος για την αποσαφήνιση τόσο της δομικής όσο και της λειτουργικής οργάνωσης των έμβιων όντων. Τα προηγούμενα χρόνια η προσπάθεια εύρεσης αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών (PPIs) περιοριζόταν σε πειραματικές μεθόδους, οι οποίες όμως, ερευνούσαν μικρά σύνολα πρωτεϊνών. Τις τελευταίες δεκαετίες, χάρη στην ανάπτυξη μεθόδων υψηλής απόδοσης, οι επιστήμονες έχουν στην διάθεσή τους σημαντικό όγκο δεδομένων, τον οποίο μπορούν με την χρήση υπολογιστικών μεθόδων, όπως η κατασκευή δικτύων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών (PPINs), να επεξεργαστούν, ώστε να εξάγουν σύμπλοκα πρωτεϊνών με πιθανή βιολογική σημασία.

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανασκόπηση και αξιολόγηση διαδεδομένων στην βιβλιογραφία αλγορίθμων ομαδοποίησης, οι οποίοι εφαρμόζονται σε βιολογικά δίκτυα και πιο συγκεκριμένα σε δίκτυα αλληλεπίδρασης πρωτεΐνης – πρωτεΐνης, για την εύρεση συμπλόκων πρωτεϊνών. Πρώτο στάδιο για την επίτευξη αυτού αποτελεί η εκτενής μελέτη βάσεων δεδομένων σχετικές με αλληλεπιδράσεις τέτοιου τύπου και η επιλογή εξ αυτών αντιπροσωπευτικών συνόλων δεδομένων συσχετίσεων. Στην συνέχεια, με την χρήση των συνόλων αυτών και των αντίστοιχων δικτύων που παράγουν, θα εκτελεστούν οι υπό μελέτη αλγόριθμοι ομαδοποίησης: Affinity Propagation, ClusterONE, MCL, MCODE, NCMine, SPICi. Ακολούθως τα αποτελέσματα που θα παραχθούν θα υποστούν στατιστικό έλεγχο, ώστε να βρεθεί η αξιοπιστία των δεδομένων που προέκυψαν, μέσω της σύγκρισης τόσο μεταξύ των αποτελεσμάτων των διαφορετικών αλγορίθμων, όσο και με ήδη γνωστά δεδομένα σχετικά με πρωτεϊνικά σύμπλοκα. Διαμέσου της μελέτης αυτής θα είναι εφικτός ο προσδιορισμός των αλγορίθμων αυτών που αποδίδουν τα βέλτιστα αποτελέσματα, και επομένως θα υπάρξει η δυνατότητα για πιο αποτελεσματική, μελλοντική, ανάπτυξη ερευνών ανάλυσης συσχετίσεων πρωτεϊνών με άλλες πρωτεΐνες.

Abstract

The function of an organism on the molecular level depends to a large extent on the interactions of its macromolecules. The formation of protein complexes is a basic unit, responsible for a variety of biological mechanisms within the cell. The identification, therefore, of the specific interactions becomes necessary in order to clarify both the structural and the functional organization of living beings. In previous years, the attempt to find protein - protein interactions (PPIs) was limited to experimental methods which, however, mainly investigated small sets of proteins. In recent decades, thanks to the development of high - throughput methods, scientists have at their disposal a significant amount of data. By using computational methods, such as protein - protein interaction networks (PPINs) this information can be processed in order to predict protein complexes with potential biological significance.

The main purpose of this thesis is to provide an overview and evaluation in some of the main clustering algorithms that are applied in biological networks and specifically in protein - protein interaction networks, in order to identify protein complexes. The first step to achieve this is the extensive study of databases containing such interactions and the selection of a representative dataset of PPIs. Thereafter in total 6 different clustering algorithms (Affinity Propagation, ClusterONE, MCL, MCODE, NCMine, and SPICi) are executed with inputs the interaction networks of those datasets. The results that are produced undergo statistical evaluation in order to verify the credibility of the methods, by comparing the results of the different algorithms both with each other and with protein complexes of a reference dataset. Through this study it is possible to determine the algorithm that returns the best results as well as the algorithms with the most related results, consequently leading to more effective future studies of protein - protein interactions.

1. Εισαγωγή

1.1 Γράφοι και Δίκτυα

1.1.1. Βασικοί ορισμοί και κατηγορίες γράφων

Η θεωρία των γράφων αρχίζει να αναπτύσσεται από την επιστήμη των διακριτών μαθηματικών το 1736 με την δημοσίευση από τον γνωστό μαθηματικό Euler του προβλήματος των Επτά Γεφυρών του Königsberg. Ωστόσο, ο όρος "γράφος" εμφανίζεται για πρώτη φορά στα πλαίσια των φυσικών επιστημών έναν αιώνα περίπου αργότερα, το 1878, από τον μαθηματικό James J. Sylvester. Έκτοτε η χρήση των γράφων για την απεικόνιση και μελέτη συστημάτων των οποίων τα μέρη συνδέονται με σχέσεις αλληλεπίδρασης εξαπλώθηκε σε πλήθος επιστημονικών κλάδων.

Ο ορισμός του γράφου περιλαμβάνει τον προσδιορισμό ενός μη μηδενικού και πεπερασμένου συνόλου κορυφών V (Vertices) και ενός πεπερασμένου συνόλου ακμών E (Edges) οι οποίες ενώνουν τα σημεία. Ως γράφος ορίζεται η σχέση *G=(V,E)*. Στις περισσότερες περιπτώσεις μία ακμή συνδέει ένα συγκεκριμένο ζευγάρι κορυφών, παρόλα αυτά υπάρχουν γράφοι που επιτρέπουν την ύπαρξη ακμών οι οποίες ενώνουν την κορυφή με τον εαυτό της δημιουργώντας έναν βρόγχο (Loop ή self-loop). Επιπλέον σε σύνθετες μορφές γράφων συχνά δύο κορυφές ενώνονται με περισσότερες της μίας ακμής (Multi- edge). Οι γράφοι που δεν περιέχουν βρόγχους και παράλληλες ακμές καλούνται απλοί (Simple graph). Η ύπαρξη παράλληλων συνδέσεων αποκτά κυρίως νόημα όταν κάθε ακμή συνοδεύεται από έναν επιπλέον παράγοντα, το βάρος W (Weight). Το βάρος δίνει μία παραπάνω πληροφορία για την σχέση ανάμεσα σε δύο κορυφές. Ένα ακόμα χαρακτηριστικό των ακμών είναι η κατεύθυνση τους και συγκεκριμένα η ύπαρξη αυτής, όπου και ο γράφος καλείται κατευθυνόμενος (Directed) ή η έλλειψη της, οπότε και ο γράφος είναι μη κατευθυνόμενος (Undirected). Στην περίπτωση των κατευθυνόμενων γράφων το σύνολο των ακμών τους είναι διατεταγμένο, ενώ στους μη κατευθυνόμενους είναι συμμετρικό. Ένας γράφος μπορεί ταυτόχρονα να έχει τόσο ακμές με κατεύθυνση, όσο και μη κατευθυνόμενες (Mixed).

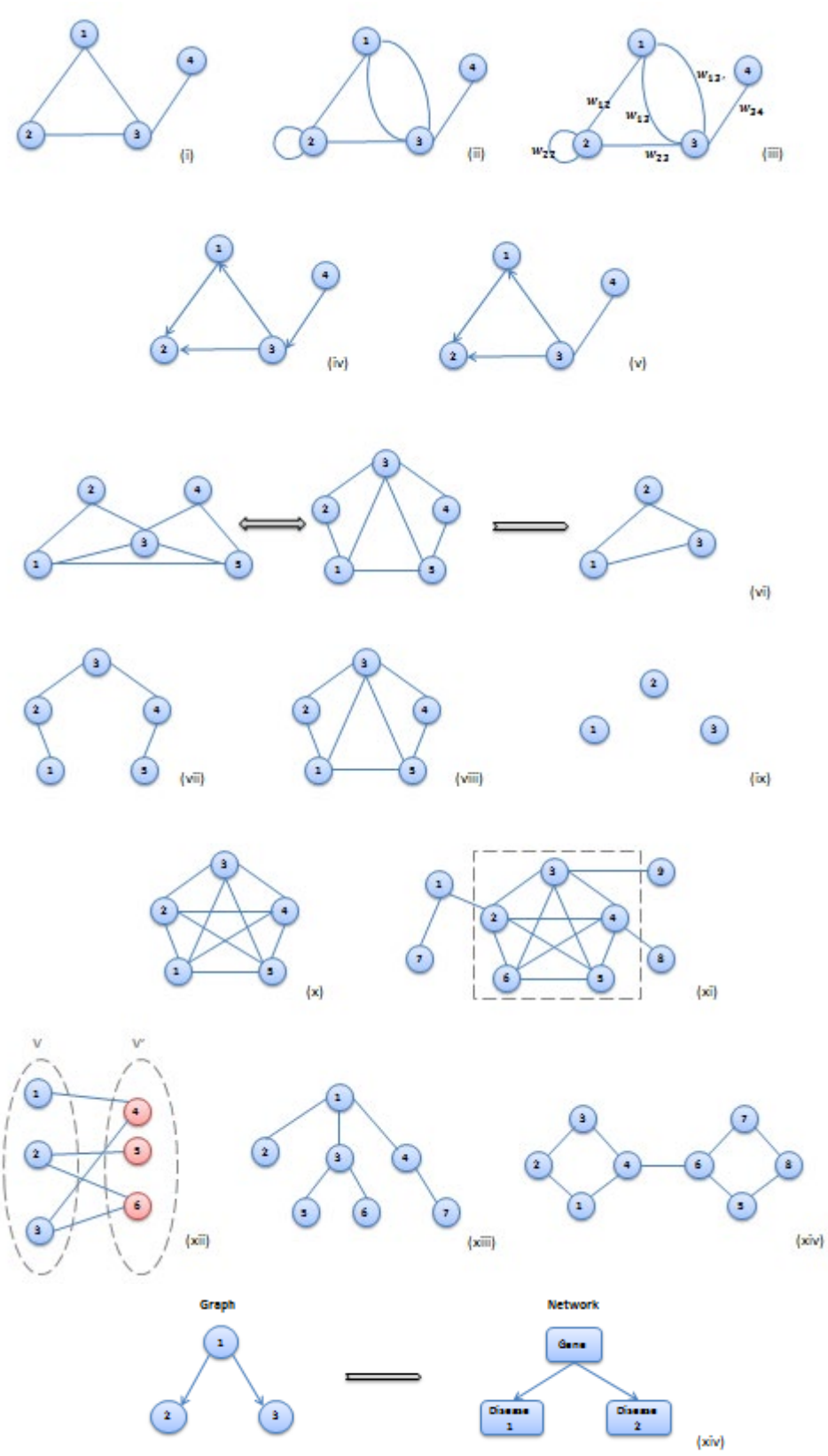
Στην πραγματικότητα ένας γράφος G αποτελεί μία διάταξη των σημείων *V(* v_1*,* v_2*,…,* v_n*)* με τρόπο που να τηρούνται οι σχέσεις των ακμών E. Επομένως είναι λογικό το ίδιο γράφημα να μπορεί να αναπαρασταθεί με διαφορετικό τρόπο, δημιουργώντας ισόμορφα (Isomorphic) αυτού. Επιπλέον σε πολλά προβλήματα μελετάται τμήμα μόνο του γράφου, λεγόμενος υπογράφος (Subgraph). Η πυκνότητα των ακμών σε κάθε γράφο διαφέρει με αποτέλεσμα να υπάρχουν γράφοι με ελάχιστο αριθμό

ακμών (Sparse), πυκνοί (Dense) γράφοι, των οποίων ο αριθμός των ακμών πλησιάζει το μέγιστο δυνατό, ακόμα και κενοί (Null), όπου οι κορυφές δεν συνδέονται μεταξύ τους. Ειδική κατηγορία αποτελούν οι πλήρης (Completed) γράφοι, των οποίων κάθε ζεύγος κορυφών συνδέεται μέσω μίας ακμής. Συχνά τα πλήρη γραφήματα περιγράφονται και ως κλίκες (Clique). Όταν σε ένα πλήρες γράφημα μπορεί να ομαδοποιηθεί μέρος τον κορυφών του βάση συγκεκριμένων μεθόδων τότε αυτές αποτελούν μία συστάδα (Cluster).

Δύο κατηγορίες γράφων με συχνή εφαρμογή στην Βιολογία και στην Βιοϊατρική είναι οι διμερής (Bipartite) γράφοι και τα δέντρα (Tree). Ένας μη κατευθυνόμενος γράφος είναι διμερής όταν οι κορυφές του μπορούν να χωριστούν σε δύο διακριτά σύνολα, έτσι ώστε κάθε ακμή να συνδέει μια κορυφή του ενός σύνολο με μια κορυφή του άλλου, ενώ ένα δέντρο σχηματίζεται από έναν μη κυκλικό συνεκτικό γράφο, όπου ως συνεκτικός (Connected) ορίζεται ο γράφος στον οποίο για κάθε ζεύγος κορυφών υπάρχει ένα μονοπάτι που τις συνδέει.

Η παρούσα εργασία επικεντρώνεται σε μία κατηγορία γράφων, τα δίκτυα. Η επιστήμη των δικτύων, παρόλο που συχνά συγχέεται, αποτελεί μέρος της θεωρίας των γράφων. Ένα δίκτυο δημιουργείται όταν στην αόριστη μαθηματική έννοια του γράφου δοθούν στις κορυφές και στις ακμές ιδιότητες, με αποτέλεσμα οι κορυφές πλέον να αναπαριστούν αντικείμενα ή οντότητες του πραγματικού κόσμου και οι ακμές τις σχέσεις μεταξύ τους. Επιπλέον στην θεωρία των δικτύων οι κορυφές συχνά αναφέρονται ως κόμβοι και οι ακμές ως σύνδεσμοι. Οι όροι αυτοί τείνουν να ταυτίζονται στην σχετική βιβλιογραφία, καθώς υποδηλώνουν τα ίδια σύνολα.

Στην Εικόνα 1 συνοψίζονται όλες οι περιπτώσεις γράφων που περιγράφθηκαν στο υποκεφάλαιο αυτό.



Εικόνα 1: Χαρακτηριστικά παραδείγματα των περιπτώσεων γράφων: i) απλός μη κατευθυνόμενος γράφος, ii) γράφος με βρόγχους και παράλληλες ακμές, iii)γράφος με βάρη, iv)κατευθυνόμενος γράφος, v) μεικτός γράφος, vi)ισόμορφοι γράφοι και υπογράφος του αρχικού, vii)αραιός γράφος, viii)πυκνός γράφος, ix)κενός γράφος, x)πλήρης γράφος, xi)κλίκα σε γράφο, xii)διμερής γράφος, xiii)δέντρο, xiv)συνεκτικός γράφος, xv) δίκτυο

1.1.2. Δομές Αναπαράστασης Γράφων

Οι επικρατέστερες δομές αναπαράστασης γράφων είναι δύο, ο πίνακας γειτνίασης (Adjacency matrix) και η λίστα γειτνίασης (Adjacency list). Θεωρώντας ως γειτονικές δύο κορυφές οι οποίες συνδέονται μέσω μίας μόνο ακμής, ο πίνακας γειτνίασης αποτελεί μία δομή με NxN γραμμές και στήλες, όπου N ο αριθμός των κορυφών του γράφου. Για το γέμισμα του πίνακα γειτνίασης *A=(a_{ij})* ακολουθείται η απλή συνθήκη:

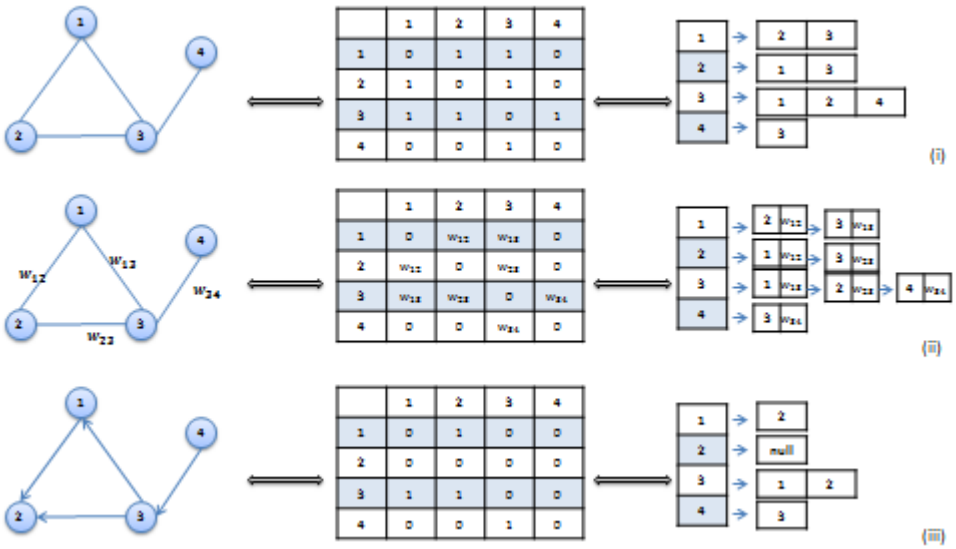
$$A_{ij} = \begin{cases} 1, \varepsilon\acute{\alpha}\nu\ a_{ij} \in E, \\ 0, \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Ενώ σε περίπτωση γράφου με βάρη η παραπάνω συνθήκη τροποποιείται, ώστε κάθε κελί του πίνακα να αποδίδει, εκτός της ύπαρξη ή μη ακμής, και το βάρος της, έχοντας:

$$A_{ij} = \begin{cases} w_{ij}, \varepsilon\acute{\alpha}\nu\ w_{ij} \in W, \\ 0, \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

Στις περιπτώσεις που ο πίνακας γειτνίασης περιγράφει ένα μη κατευθυνόμενο γράφο, τότε αυτός είναι συμμετρικός και η διαγώνιος του αποτελείται από μηδέν. Αντιθέτως, όταν ο γράφος είναι κατευθυνόμενος, τότε το άνω τριγωνικό μέρος διαφέρει του κάτω.

Παρόλο που η κατασκευή ενός πίνακα γειτνίασης είναι εύκολη και λόγω αυτού επί χρόνια αποτελεί δημοφιλή επιλογή αναπαράστασης γραφημάτων, τόσο από άποψη απαιτούμενης μνήμης, όσο και προσπέλασης και επεξεργασίας των στοιχείων του, αποτελεί μη αποδοτική λύση, ιδιαίτερα για αραιούς γράφους. Λόγω των παραπάνω η δομή που προτιμάται τα τελευταία χρόνια είναι η λίστα γειτνίασης. Μια λίστα γειτνίασης αποτελείται από ένα διάνυσμα μεγέθους Nx1, όπου κάθε κελί του αντιστοιχεί σε μία κορυφή και περιέχει σε μία λίστα όλες τις κορυφές που συνδέονται με αυτήν. Βάση της παραπάνω περιγραφής η μνήμη που χρειάζεται για την αποθήκευση των στοιχείων είναι *O(V+E)* έναντι *O(V^2)* του πίνακα γειτνίασης. Για την καλύτερη κατανόηση των μεθόδων, αποδίδονται στην Εικόνα 2 οι αναπαραστάσεις τριών χαρακτηριστικών γράφων τόσο με πίνακα γειτνίασης όσο και με λίστα.



Εικόνα 2: Παραδείγματα δημιουργίας δομών πίνακα και λίστας για την αναπαράσταση: i)μη κατευθυνόμενου γράφου, ii)γράφου με βάρη, iii)κατευθυνόμενου γράφου

1.1.3. Ιδιότητες Γράφων και Δικτύων

Για την μελέτη των χαρακτηριστικών των γράφων έχουν οριστεί κάποια μέτρα, τα οποία περιγράφουν τις ιδιότητες του. Στο υποκεφάλαιο αυτό θα περιγραφούν τα βασικότερα εξ αυτών και εκείνα τα οποία θα συναντηθούν σε μετέπειτα σημεία της εργασίας.

Από τις πιο εύκολα κατανοητές έννοιες είναι ο βαθμός (Degree). Ο βαθμός αποτελεί μέτρο κεντρικότητας (Centrality) μίας κορυφής και εννοείται ως ο αριθμός των ακμών αυτής. Στην περίπτωση των μη κατευθυνόμενων γραφημάτων ο βαθμός αυτός είναι ένας και αποτελεί το άθροισμα όλων των ακμών που ενώνονται άμεσα με την κορυφή. Στην περίπτωση των κατευθυνόμενων γραφημάτων ο συνολικός βαθμός της κορυφής προκύπτει από το άθροισμα των ακμών που δείχνουν προς αυτή συν το άθροισμα των ακμών που εκπέμπονται από εκείνη. Συχνά νόημα έχει και ο υπολογισμός του μέσου βαθμού (Average Degree) όλου του γράφου, αφού συγκρίνει των αριθμό των κορυφών με τον αριθμό των ακμών. Ο βαθμός αυτός υπολογίζεται μέσω των τύπων:

$deg_{avg} = \frac{2E}{N}$ για μη κατευθυνόμενους γράφους και $deg_{avg} = \frac{E}{N}$ για κατευθυνόμενους.

Μία από τις πιο βασικές ιδιότητες ενός γράφου είναι η κατανομή των βαθμών του (Degree Distribution) η οποία υποδηλώνει την δομή του γράφου υπολογίζοντας τον αριθμό των κορυφών με συγκεκριμένο βαθμό. Συχνά απεικονίζεται ως ένα ιστόγραμμα με τις συχνότητες των βαθμών, στο οποίο σε απλά γράφημα οι ψηλότερες στήλες αντιστοιχούν σε μικρό βαθμό. Και σε γράφους που

αναπαριστούν δίκτυα του πραγματικού κόσμου οι περισσότερες κορυφές έχουν σχετικά μικρό βαθμό, ενώ υπάρχουν και ελάχιστες κορυφές με πολύ υψηλό βαθμό, καθώς συνδέονται με πολλές άλλες. Αυτές οι κορυφές υψηλού βαθμού αναφέρονται συχνά ως κομβικά σημεία (Hub) του δικτύου. Μαθηματικά το μέτρο εκφράζεται:

$$P_{deg}(k) = \frac{\sum V_{deegre=k}}{V}$$

Σχετικό μέτρο με την ύπαρξη κομβικών σημείων σε γράφους είναι η συνεκτικότητα (Connectivity) των στοιχείων του. Η συνεκτικότητα των κορυφών ή ακμών ενός γράφου καταγράφει τον ελάχιστο αριθμό κορυφών ή ακμών αντίστοιχα που χρειάζεται να αφαιρεθούν, ώστε ο γράφος να διαχωριστεί σε μεμονωμένους υπογράφους. Συνήθως σε πολύπλοκα δίκτυα οι κόμβοι που αφαιρούνται είναι τα κομβικά σημεία του δικτύου.

Ιδιότητα των γράφων, η οποία, όπως και τα προηγούμενα μέτρα, καθορίζεται από τις ακμές του είναι η πυκνότητα (Density). Η πυκνότητα ενός γράφου βρίσκεται από το κλάσμα των ακμών που υπάρχουν στο γράφημα ως προς τον αριθμό τον ακμών που είναι εφικτό να υπάρξουν σε αυτό. Σε ένα απλό γράφο, χωρίς βρόχους και παράλληλες ακμές, ο μέγιστος αριθμός ακμών του είναι $E_{max} = \frac{n\,(n-1)}{2}$, οπότε η συνάρτηση πυκνότητας είναι:

$$den = \frac{E}{E_{max}}$$

Από την εύρεση της πυκνότητας των γράφων μπορεί να βρεθεί αν αυτοί είναι πυκνοί ή αραιοί, αφού ένα πυκνό γράφημα θα έχει den≅1. Τα βιολογικά δίκτυα έχει αποδειχθεί ότι είναι γενικά αραιά, καθώς αυτό παρέχει ένα εξελικτικό πλεονέκτημα για την επιβίωση των συστημάτων [1–4].

Ένα μέτρο που συνδέει την πληροφορία της συνεκτικότητας και της πυκνότητας ενός γράφου είναι η συγκεντρωτικότητα (Centralization) του. Η συγκεντρωτικότητα δίνει πληροφορίες για την κατανομή των ακμών στο γράφο και περιγράφεται ως:

$$centralization = \frac{n}{n-2}\left(\frac{max(deg)}{n-1} - den\right)$$

Επιπλέον πληροφορίες πάνω στην κατανομή αυτή λαμβάνονται από την ετερογένεια (Heterogenity) του γράφου, η οποία αντικατοπτρίζει και την τάση ενός δικτύου να περιέχει κομβικά σημεία. Στην βιβλιογραφία υπάρχει πλήθος προτάσεων μεθόδων υπολογισμού της ετερογένειας, ενώ μία από τις πρώτες και πιο απλές είναι βάση του τύπου:

$$heterogenity = \frac{var(deg)}{mean(deg)}$$

Ένα από τα συνήθη ερωτήματα που αναζητούνται στην μελέτη γράφων και δικτύων είναι η τάση ή πιθανότητα ύπαρξης μέσα σε αυτούς ομάδων κορυφών, οι οποίες μπορούν να διαφοροποιηθούν βάση κάποιων χαρακτηριστικών τους. Για την απάντηση αυτών των ερωτημάτων έχουν οριστεί αρκετά μέτρα, τα οποία εκφράζουν με ποικίλους τρόπους την τάση αυτή των γραφημάτων.

Ένα από τα μέτρα αυτά είναι η αρθρωτότητα (Modularity), η οποία περιγράφει την πυκνότητα των ακμών ανάμεσα στις κορυφές μίας ομάδας σε σύγκριση με τις ακμές που φεύγουν προς κόμβους εκτός της ομάδας αυτής. Πιο συγκεκριμένα, έστω η ύπαρξη μία ομάδας στοιχείων i και j το σύνολο με τα στοιχεία εκτός της ομάδας αυτής. Η αρθρωτότητα του γραφήματος δίνεται από τον τύπο:

$Q = \sum_i [\, e_{ii} - (\sum_j e_{ij})^2]$, όπου e_{ii} οι ακμές εντός του συνόλου και e_{ij} οι ακμές εκτός αυτού.

Το πρώτο σκέλος της εξίσωσης 'επιβραβεύει' την ύπαρξη πολλών ακμών ανάμεσα στις κορυφές της ομάδας, ενώ η τιμή της αρθρωτότητας μειώνεται με την ύπαρξη πολλών ακμών που φεύγουν προς κορυφές εκτός αυτής.

Παρόμοιο μέτρο με την αρθρωτότητα ενός γράφου είναι και ο συντελεστής ομαδοποίησης (Clustering Coefficient) και εκφράζει το ποσοστό των γειτονικών κορυφών που συνδέονται μεταξύ τους. Ο συντελεστής ομαδοποίησης μίας κορυφής v ορίζεται ως ο αριθμός των συνδεδεμένων τριάδων κορυφών στις οποίες περιλαμβάνεται η v, προς τον αριθμό των τριάδων που γειτονεύουν με την v. Στην περίπτωση μη κατευθυνόμενου γράφου ο συντελεστής υπολογίζεται από την εξίσωση:

$C_v = \frac{2e_v}{deg_v\,(deg_v - 1)}$, ενώ σε κατευθυνόμενους γράφους η σχέση παίρνει την μορφή:

$$C_v = \frac{e_v}{deg_v\,(deg_v - 1)}$$

Ο ολικός συντελεστής ομαδοποίησης ενός γράφου ή δικτύου βρίσκεται από τον υπολογισμό του μέσου συντελεστή ομαδοποίησης των κορυφών αυτού, δηλαδή ισούται με:

$$C_{avg} = \frac{1}{V} \sum_{i=1}^{v} C_i$$

Ένας όρος που αναφέρθηκε και σε προηγούμενο υποκεφάλαιο είναι το μονοπάτι (Path). Ως μονοπάτι δύο σημείων εννοείται η διαδοχική ακολουθία διακριτών κορυφών και ακμών που μεσολαβούν για την ένωση των δύο σημείων. Είναι δηλαδή η διαδρομή μέσα στον γράφο από μία κορυφή προς μία άλλη. Ένα από τα συνήθη προβλήματα στην θεωρεία τον γράφων είναι η εύρεση των συντομότερων

μονοπατιών των γράφων, πεδίο που έχει οδηγήσει στην ανάπτυξη ποικίλων αλγορίθμων αναζήτησης [5–9]. Το συντομότερο μονοπάτι (Shortest path) ανάμεσα σε δύο κορυφές είναι αυτό που περιλαμβάνει το μικρότερο αριθμό κορυφών και ακμών ανάμεσα τους, ενώ ο αριθμός των ακμών αυτών αποτελεί το μήκος ή απόσταση (Distance) του μονοπατιού και ισούται με *dist=N-1*, όπου N οι κορυφές της ακολουθίας. Στις περισσότερες περιπτώσεις ανάλυσης δικτύων καταγράφονται όλα τα συντομότερα μονοπάτια που σχηματίζονται σε αυτό σε πίνακες, μέσω των οποίων μπορούν να βρεθούν εύκολα τρία χαρακτηριστικά μέτρα του δικτύου: η διάμετρος (Diameter), η ακτίνα (Radius) και το μέσο μήκος του συντομότερου μονοπατιού ή αλλιώς χαρακτηριστικό μήκος μονοπατιού (Average Shortest Path Length ή Characteristic Path Length). Η διάμετρος και η ακτίνα του δικτύου απευθύνονται, αντίστοιχα, στο μεγαλύτερου μήκους συντομότερο μονοπάτι και στο μικρότερου μήκους συντομότερο μονοπάτι από όσα έχουν βρεθεί στο γράφημα. Το χαρακτηριστικό μονοπάτι αποδίδει τον μέσο όρο των συντομότερων μονοπατιών του δικτύου και μπορεί να προσφέρει μία εικόνα για την ροή μετάδοσης της πληροφορίας σε αυτό, επομένως την αποδοτικότητα του. Ο υπολογισμός της τιμής του γίνεται μέσω της σχέσης:

$$l_G = \frac{1}{V\,(V-1)} \sum_{i \neq j} dist(v_i, v_j)$$

1.1.4. Μέτρα κεντρικότητας Γράφων και Δικτύων

Έχοντας ορίσει την έννοια του μονοπατιού και της απόστασης αυτού μπορούν να περιγραφούν οι μετρικές που αποδίδουν την κεντρικότητα (Centrality) ενός γράφου. Τα μέτρα κεντρικότητας προσπαθούν να καθορίσουν την επιρροή της κάθε κορυφής στον γράφο, ποσοτικοποιώντας την ικανότητα αυτής να επικοινωνεί άμεσα ή έμμεσα με τις υπόλοιπες κορυφές. Από τις πρώτες μελέτες, οι οποίες συγκέντρωσαν και υπέδειξαν τα βασικά μέτρα κεντρικότητας που είναι απαραίτητα κατά την ανάλυση ενός δικτύου είναι των Freeman et al. [10,11], ενώ έκτοτε έχουν δημοσιευθεί δεκάδες εργασίες οι οποίες είτε προτείνουν νέες μετρικές είτε συγκρίνουν τις υπάρχοντες[3,12–17]. Στο υποκεφάλαιο αυτό θα παρουσιαστούν ορισμένα από τα βασικότερα μέτρα κεντρικότητας και ιδιαίτερα εκείνα που χρησιμοποιούνται σε μελέτες δικτύων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών, όπως η παρούσα.

Το πρώτο μέτρο ταυτίζεται με ένα χαρακτηριστικό που περιγράφθηκε ήδη σε προηγούμενο υποκεφάλαιο, τον βαθμό μίας κορυφής. Καλείται κεντρικότητα βαθμού (Degree Centrality) και η εύρεση της αποσκοπεί στην επισήμανση των κορυφών με τις πιο πολλές συνδέσεις. Η μαθηματική σχέση που την περιγράφει είναι $C_{deg\,(v)} = deg_v$, ενώ ισχύουν όλες οι ιδιότητες του βαθμού, όπως αναλύθηκαν παραπάνω.

Η εκκεντρικότητα ενός γράφου, αποτελεί ένα ακόμα μέτρο κεντρικότητας, το οποίο υπολογίζεται από την εύρεση του μεγαλύτερου μήκους συντομότερο μονοπάτι ανάμεσα σε δύο κορυφές του γράφου. Δηλαδή η εκκεντρικότητα δύο σημείων αντιστοιχεί στο $C_{ecc\,(v_i,v_j)}$ = *max dist(* v_i*,* v_j*)*. Για να έχει σημαντική επιρροή ένας κόμβος σε ένα δίκτυο θα πρέπει η τιμή εκκεντρικότητας να είναι χαμηλή, καθώς όσο μεγαλύτερη η τιμή τόσο περισσότερο θα καθυστερεί να μεταφερθεί η πληροφορία από τον δεδομένο κόμβο.

Μέτρο κεντρικότητας το οποίο χρησιμοποιείται, όμοια με το προηγούμενο, για την εύρεση των κόμβων σε ένα δίκτυο που μεταφέρουν ταχύτερα την πληροφορία είναι η κεντρικότητα εγγύτητας (Closeness Centrality). Για τον υπολογισμό της απαιτείται η εύρεση του αθροίσματος των αποστάσεων των συντομότερων μονοπατιών και συγκεκριμένα:

$C_{cl(v_i,v_j} = \frac{1}{\sum_{i\neq j} dist(v_i,v_j)}$, ενώ η κανονικοποιημένη μορφή αυτού είναι:

$C_{cl(v_i,v_j} = \frac{N-1}{\sum_{i\neq j} dist(v_i,v_j)}$

Το τελευταίο μέτρο που θα περιγραφθεί στην ενότητα αυτή είναι η ενδιάμεση κεντρικότητα (Betweenness Centrality), η οποία δείχνει την σημασία μίας κορυφής ως γέφυρα επικοινωνίας ανάμεσα σε δύο άλλες κορυφές. Η σημασία αυτή ποσοτικοποιείται μέσω του αριθμού των φορών εμφάνισης της συγκεκριμένης κορυφής στα συντομότερα μονοπάτια των άλλων δύο και βρίσκεται μέσω της συνάρτησης:

$C_{bet(v)} = \frac{\sigma_{xy}(v)}{\sigma_{xy}}$, όπου $\sigma_{xy}(v)$ είναι ο αριθμός των συντομότερων μονοπατιών ανάμεσα στις κορυφές x και y, τα οποία διέρχονται από την κορυφή v, ενώ σ_{xy} είναι το άθροισμα όλων των συντομότερων μονοπατιών των x και y.

Η εύρεση των κορυφών με υψηλή ενδιάμεση κεντρικότητα είναι καθοριστική για την ευστάθεια ενός γράφου, καθώς η αφαίρεση αυτών οδηγεί στην μη μετάδοση πληροφορίας μέσα σε ένα δίκτυο και σταδιακά στην κατάρρευση αυτού.

Κατά την εξέλιξη του κλάδου της Βιοπληροφορικής τα μέτρα κεντρικότητας έχουν χρησιμοποιηθεί, εκτός από την άμεση αξιοποίηση της πληροφορίας που δίνουν για την οργάνωση και ιεράρχηση ενός δικτύου, και ως μέρη αλγορίθμων κατασκευής, επεξεργασίας και ομαδοποίησης βιολογικών δικτύων [18–27].

1.2 Αλληλεπιδράσεις Πρωτεϊνών - Πρωτεϊνών

Η ομαλή λειτουργία και επιβίωση ενός οργανισμού στηρίζεται στην παράλληλη δράση εκατομμυρίων μορίων μέσα σε αυτόν. Από τα σημαντικότερα μακρομόρια, με βασικό ρόλο στην πλειοψηφία των διεργασιών των έμβιων όντων, είναι οι πρωτεΐνες. Για την εκτέλεση των λειτουργιών αυτών οι πρωτεΐνες δημιουργούν σχέσεις τόσο μεταξύ τους όσο και με άλλα μόρια του οργανισμού, με την πρώτη κατηγορία να αποτελεί αντικείμενο ιδιαίτερου ενδιαφέροντος πολλών μελετών[28–33]. Οι αλληλεπιδράσεις πρωτεϊνών – πρωτεϊνών (Protein – Protein Interactions ή PPIs) αποτελούν την φυσική επαφή ανάμεσα σε δύο ή περισσότερες πρωτεΐνες. Η μοριακή πρόσδεση αυτών οφείλει να προκύπτει ως αποτέλεσμα εσκεμμένων βιομοριακών δυνάμεων και να αποσκοπεί στην εκτέλεση συγκεκριμένων βιολογικών λειτουργιών. Παρόλο που πολλές αλληλεπιδράσεις απεικονίζουν σταθερά σύμπλοκα μακρομορίων, το κύτταρο δεν αποτελεί μία στατική δομή, με αποτέλεσμα πλήθος πρωτεϊνικών αλληλεπιδράσεων να σχηματίζεται μόνο για την εκτέλεση παροδικών δράσεων ή κατά την ύπαρξη συγκεκριμένου πλαισίου όπως τη φάση και την κατάσταση του κυτταρικού κύκλου, το στάδιο ανάπτυξης του οργανισμού, τις περιβαλλοντικές συνθήκες που επικρατούν, τις τροποποιήσεις που επιδέχεται μία πρωτεΐνη, καθώς και την παρουσία συμπαραγόντων ή την πρόσδεση μη πρωτεϊνικών μορίων[34].

Η ανίχνευση των αλληλεπιδράσεων αυτών πραγματοποιείται με ποικίλες μεθόδους οι οποίες μπορούν να διαχωριστούν σε εκείνες που πραγματοποιούνται in vivo, in vitro ή in silico, καθώς και βάση υψηλής ή χαμηλής απόδοσης τους. Παρακάτω θα περιγραφούν ορισμένες εκ των βασικών μεθόδων για κάθε κατηγορία.

1.2.1. Πειραματικές Τεχνικές

Διϋβριδισμός στο ζυμομύκητα (yeast -two- hybrid)

Μία από της πιο δημοφιλής μεθόδους ανίχνευσης αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών σε μεγάλη κλίμακα είναι ο διϋβριδισμός στο ζυμομύκητα, η οποία προτάθηκε αρχικά το 1989 από τους Fields Stanley και Ok-kyu Song[28]. Εκ τότε έχουν δημοσιευθεί αρκετές παραλλαγές αυτής, με βασική αρχή όλων πως όταν δύο πρωτεΐνες ενδιαφέροντος αλληλεπιδρούν ανασύσταται ένας λειτουργικός παράγοντας μεταγραφής. Βάση αυτού, δύο υβρίδια του μεταγραφικού παράγοντα κατασκευάζονται, ανάμεσα στις πρωτεΐνες προς μελέτη, τα οποία περιέχουν την περιοχή ενεργοποίησης και την περιοχή πρόσδεσης του DNA. Στη συνέχεια, εάν αυτές οι πρωτεΐνες ενδιαφέροντος βρεθούν πολύ κοντά η μία στην άλλη, θα φέρουν τους επισημασμένους τους τομείς ενεργοποίησης και σύνδεσης DNA αρκετά κοντά, ώστε να σχηματίσουν μια λειτουργική μονάδα μεταγραφής. Αυτή η λειτουργική μονάδα μπορεί ακολούθως να συνδέσει ένα στοιχείο προαγωγού και

να ενεργοποιήσει τη μεταγραφή τους. Η επιλογή της μεθόδου διϋβριδισμού συνοδεύεται από πολλά πλεονεκτήματα, όπως η απλή και οικονομική πειραματική διαδικασία, το γεγονός πως αυτή πραγματοποιείται in vivo, επομένως αντικατοπτρίζονται πραγματικές καταστάσεις των πρωτεϊνών, καθώς και ότι είναι εφικτή η ανίχνευση ακόμα και ασθενών ή παροδικών αλληλεπιδράσεων[35]. Παράλληλα όμως περιλαμβάνει μειονεκτήματα, για παράδειγμα η ικανότητα εύρεσης μόνο αλληλεπιδράσεων δύο πρωτεϊνών και ο αριθμό ψευδώς θετικών αποτελεσμάτων, ο οποίος μπορεί να ξεπεράσει το 50%[36].

Διαδοχικός Συγγενικός Καθαρισμός (Tandem Affinity Purification - TAP)

Η τεχνική του διαδοχικού συγγενικού καθαρισμού αναπτύχθηκε στα τέλη της δεκαετία του 1990[37,38] με στόχο την απομόνωση των πρωτεϊνών του ζυμομύκητα και την ταυτοποίηση των αλληλεπιδράσεων τους. Πλέον χρησιμοποιείται για την μαζική εύρεση πρωτεϊνικών σχέσεων σε ποικίλους οργανισμούς, ανάμεσα τους και ο άνθρωπος[39–42]. Η μέθοδος στηρίζεται στην χρήση ενός Tap - tag, το οποίο, στην γενική μορφή του, αποτελείται από μονάδες πρόσδεσης ανοσοσφαιρινών και ενζύμων, με ένα ειδικό σημείο αποκοπής συγκεκριμένης πρωτεάσης ανάμεσα τους. Στο άκρο του Tap - tag τοποθετείται κάθε φορά η πρωτεΐνη στόχος. Η χρήση συγκεκριμένων πρωτεασών εξασφαλίζει την αναγνώριση μόνο των πρωτεϊνών με τις αντίστοιχες ακολουθίες, μειώνοντας σημαντικά την πρόβλεψη τυχαίων αλληλεπιδράσεων. Επιπλέον δεν απαιτεί την πρότερη γνώση ούτε της δομής ούτε της λειτουργίας των πρωτεϊνών, ενώ ο προσδιορισμός αυτών γίνεται σε πραγματικές καταστάσεις, με καταγεγραμμένο πρωτόκολλο, το οποίο καθιστά τα πειράματα τόσο αντιπροσωπευτικά, όσο και αναπαράξιμα και συγκρίσιμα. Η αδυναμία της τεχνική βρίσκεται στη μη ανίχνευση αλληλεπιδράσεων οι οποίες δεν εκφράστηκαν υπό τις συγκεκριμένες συνθήκες ή πραγματοποιήθηκαν για ελάχιστο χρονικό διάστημα και δεν καταγράφηκαν από το σύστημα[43].

Φασματομετρία Μάζας (Mass Spectrometry)

Μία τεχνική, που συχνά συνδυάζεται και ακολουθεί τον διαδοχικό συγγενικό καθαρισμό, είναι η φασματομετρία μάζας[44,45], καθώς ανιχνεύει ακόμα και πολύ ασθενείς αλληλεπιδράσεις[46]. Όπως και με τις προηγούμενες μεθόδους, υπάρχουν πολλές εκδοχές της τεχνικής[47], με τη βάση όλων να αποτελεί ο ιονισμός των χημικών ενώσεων και η φόρτιση των ατόμων των μορίων. Για κάθε μόριο υπολογίζεται στον φασματογράφο το πηλίκο της φόρτισής του δια την μάζα του, με την μέτρηση να στηρίζεται στην κίνηση των ιόντων κατά την μετάδοση τους στο ηλεκτρομαγνητικό πεδίο. Η εύρεση των αλληλεπιδράσεων γίνεται μέσω της ποσότητας των ιόντων που βρίσκεται, για αυτό και η φασματομετρία μάζας κατατάσσεται στις ποσοτικές τεχνικές. Το μειονέκτημα της είναι στην χαμηλή

ακρίβεια της, αφού θεωρεί ως σχέσεις πρωτεϊνών ακόμα και τις τυχαίες αλληλεπιδράσεις.

Πρωτεϊνικές Μικροσυστοιχίες (Protein Microarrays)

Η ανίχνευση πρωτεϊνικών αλληλεπιδράσεων με την χρήση μικροσυστοιχιών πρωτεϊνών έχει γνωρίσει ιδιαίτερη άνθηση τα τελευταία είκοσι χρόνια[48–50]. Στηρίζεται στην τεχνική που πρότεινε το 1989 ο Roger Ekins[51] και αποτελεί μέθοδο υψηλής απόδοσης, η οποία πραγματοποιείται in vitro. Σε μία επίπεδη, στερεή, συνήθως γυάλινη, επιφάνεια επισυνάπτονται, σε χαρτογραφημένες περιοχές, μόρια επιλεγμένων πρωτεϊνών. Το σύνολο αυτό των επικολλημένων στην μικροσυστοιχία πρωτεϊνών έρχεται ακολούθως σε επαφή με ένα δεύτερο σύνολο πρωτεϊνών σημασμένο με φθορίζουσες ουσίες. Κατά την ύπαρξη αλληλεπίδρασης μεταξύ πρωτεϊνών των δύο συνόλων εκπέμπεται φθορίζουσα ακτινοβολία, η οποία ανιχνεύεται από έναν σαρωτή laser και επεξεργάζεται μέσω ειδικού λογισμικού. Η μέθοδος χαρακτηρίζεται από υψηλή ευαισθησία και αναλογία σήματος προς θόρυβο, ενώ απαιτεί και χαμηλή ποσότητα δείγματος για την εκτέλεση του πειράματος, υπάρχει όμως παράλληλα και μεγάλο ποσοστό πιθανότητας ύπαρξης τυχαίων συνδέσεων[52].

Οι προηγούμενες τεχνικές αποτελούν μεθόδους υψηλής απόδοσης οι οποίες, παρόλο που προσφέρουν μεγάλο όγκο πληροφορίας, έχει αποδειχθεί ότι παρουσιάζουν αδυναμίες ως προς τα ψευδώς θετικά και ψευδώς αρνητικά αποτελέσματα που προβλέπουν[36,53–55]. Για την πιο λεπτομερή και σε βάθος μελέτη των αλληλεπιδράσεων των πρωτεϊνών χρησιμοποιούνται μέθοδοι χαμηλής απόδοσης, οι οποίες μελετούν μικρά σύνολα πρωτεϊνών, συνήθως του ίδιου συμπλόκου. Τεχνικές τέτοιου τύπου είναι η κρυσταλλογραφία ακτίνων Χ, ο πυρηνικός μαγνητικός συντονισμός και η ηλεκτρονική μικροσκοπία.

Κρυσταλλογραφία ακτίνων Χ (X-ray Crystallography)

Η κρυσταλλογραφία ακτίνων Χ έχει υπάρξει στην πάροδο των χρόνων η πιο παραγωγική τεχνική για τη δομική ανάλυση πρωτεϊνών και πρωτεϊνικών συμπλοκών, ενώ εξακολουθεί να αποδίδει τα βέλτιστα αποτελέσματα όσον αφορά την ακρίβεια και την ποιότητα ανάλυσης[46]. Για την μελέτη του πρωτεϊνικού συμπλόκου απαιτείται πρώτα η κρυστάλλωση αυτού, με την διάταξη των μορίων των πρωτεϊνών στην βάση ενός πλέγματος. Η διαδικασία της κρυστάλλωσης είναι ιδιαίτερα ευαίσθητη και χρονοβόρα, αφού απαιτεί προσεκτική διατήρηση των συνθηκών στις οποίες αναπτύσσεται ο κρύσταλλος. Η μελέτη του κρυσταλλικού πλέγματος, μέσω της χρήσης ακτίνων Χ, αποκαλύπτει τη διάταξη των δομικών μερών τα οποία συγκροτούν τον κρύσταλλο. Η πληροφορία αυτή προκύπτει από τις γωνίες διάθλασης των προσκείμενων ακτινών στον κρύσταλλο, όπως αυτές

καταγράφονται σε ένα ειδικό φιλμ, σε σχήμα κουκίδων. Τα σημεία αυτά σχηματίζουν συγκεκριμένα μοτίβα, τα οποία ακολούθως δημιουργούν έναν χάρτη της κατανομής των ηλεκτρονίων των μορίων στον χώρο, ο οποίος αποδίδει με μεγάλη ακρίβεια και την δομή ολόκληρου του μορίου ή συμπλόκου. Τα μειονεκτήματα της μεθόδου προκύπτουν από την αδυναμία κρυστάλλωσης όλων των μορίων και συμπλόκων, ιδιαίτερα εκείνων με μεγάλο μοριακό βάρος και κακή διαλυτότητα, όπως οι διαμεμβρανικές πρωτεΐνες. Επιπλέον, καθώς ο κρύσταλλος 'αιχμαλωτίζει' το σύμπλοκο σε μία δεδομένη στιγμή, οι πληροφορίες που θα εξαχθούν θα απεικονίζουν τις αλληλεπιδράσεις μόνο εκείνης της κατάστασης.

Πυρηνικός Μαγνητικός Συντονισμός (Nuclear Magnetic Resonance - NMR)

Η τεχνική, η οποία μπορεί να συναγωνιστεί, σε συγκεκριμένες περιπτώσεις, την κρυσταλλογραφία ακτίνων Χ, είναι ο πυρηνικός μαγνητικός συντονισμός. Κατά τα πειράματα μαγνητικού συντονισμού οι πυρήνες των πρωτεϊνών ή των συμπλόκων πρωτεϊνών φορτίζονται με αποτέλεσμα να περιστρέφονται με μία συγκεκριμένη συχνότητα συντονισμού, η οποία διαφέρει ανάλογα με το ίδιο το άτομο αλλά με τα άτομα τα οποία περιβάλλεται. Τα διαφορετικά σήματα που αποδίδουν τα μόρια ανάλογα με την δευτεροταγή δομή τους, τις αλληλεπιδράσεις των ατόμων στον πυρήνα και τα δυναμικά χαρακτηριστικά των πολυπεπτιδικών τμημάτων αναλύονται μέσω υπολογιστικών μεθόδων, οδηγώντας στην αποκρυπτογράφηση της τριτοταγούς δομής πρωτεϊνών, αλλά και των αλληλεπιδράσεων μεταξύ πρωτεϊνών. Μέσω του μαγνητικού συντονισμού επιλύεται ένα από τα βασικά μειονεκτήματα της κρυσταλλογραφίας ακτίνων Χ, η απεικόνιση δυναμικών καταστάσεων. Επιπλέον, καθώς δεν απαιτείται η πολύπλοκη διαδικασία κατασκευής κρυστάλλων, ο χρόνος προετοιμασίας μειώνεται και τα σύμπλοκα προς μελέτη επιτρέπεται να είναι μεγαλύτερου μεγέθους. Και σε αυτή την τεχνική όμως υπάρχουν περιορισμοί ως προς το μοριακό βάρος, αφού το φάσμα μεγάλων μακρομοριακών δομών είναι ιδιαίτερα πολύπλοκο και ερμηνεύεται δύσκολα. Επιπλέον για να υπάρξει αξιόλογο ποσοστό σήματος ως προς τον θόρυβο απαιτούνται μεγάλες ποσότητες καθαρού δείγματος[56].

Ηλεκτρονική Μικροσκοπία (Electron Microscopy- EM)

Τέλος, μία ακόμα μέθοδος χαμηλής κλίμακας είναι η ηλεκτρονική μικροσκοπία, η οποία στηρίζεται στην σκέδαση ηλεκτρονίων. Η ηλεκτρονική μικροσκοπία περιλαμβάνει ως τεχνική πλήθος παραλλαγών, κατά την μελέτη όμως δομών μεγαλύτερων των 200 kDa, χρησιμοποιείται η κρύο – ηλεκτρονική μικροσκοπία. Το πρώτο συνθετικό προκύπτει από την ψύξη του δείγματος πρωτεϊνών πριν την μικροσκόπηση. Εν συνεχεία μία δέσμη ηλεκτρονίων διαπερνά το δείγμα και το στρώμα πάγου που το περιβάλει και τα ηλεκτρόνια που σκεδάζονται

καταγράφονται από έναν ανιχνευτή. Η πλήρης τρισδιάστατη δομή του μορίου ανακατασκευάζεται από πολλές δισδιάστατες προβολές του δείγματος, καθεμία από τις οποίες δείχνει το αντικείμενο από διαφορετική γωνία. Όπως φανερώνει και η περιγραφή της διαδικασίας, η μέθοδος υπερτερεί περιορισμών των προηγούμενων τεχνικών, όπως το μέγεθος της υπό μελέτης πρωτεΐνης, η ποσότητα δείγματος που χρειάζεται και η προετοιμασία αυτού. Παρόλα αυτά η ποιότητα ανάλυσης που αποδίδει είναι σημαντικά χαμηλότερη, με αποτέλεσμα να χρησιμοποιείται κυρίως σε συνδυασμό με άλλες μεθόδους[46].

1.2.2. Υπολογιστικές Τεχνικές

Οι μέθοδοι πειραματικού προσδιορισμού των αλληλεπιδράσεων των πρωτεϊνών συχνά συνοδεύεται από αυξημένες απαιτήσεις σε χρόνο, ειδικευμένο προσωπικό, αλλά και οικονομικό κόστος. Λόγω αυτών όλο και περισσότερες μέθοδοι που συνδυάζουν αρχές και τεχνικές από την επιστήμη των Υπολογιστών και της Βιοφυσικής προτείνονται για των προσδιορισμών σχέσεων πρωτεϊνών - πρωτεϊνών. Ένας βασικός διαχωρισμός μπορεί να γίνει σε μεθόδους που χρησιμοποιούν γονιδιακή ή ακολουθιακή και δομική πληροφορία πρωτεϊνών.

Μέθοδοι βασισμένες στην διατήρηση γειτονιών γονιδίων (Gene Conserved Neighborhood)

Οι τεχνικές οι οποίες αξιοποιούν πληροφορία από τα γονίδια τα οποία εκφράζουν την πρωτεΐνη, καθώς και τα μονοπάτια τους, στηρίζονται κυρίως στους στενά συνδεδεμένους εξελικτικούς μηχανισμούς αυτών. Χαρακτηριστικό παράδειγμα αποτελούν οι μελέτες διατηρημένων γονιδιακών γειτονιών (Gene Conserved Neighborhood)[57–62], οι οποίες έχουν προβλέψει, με αρκετά υψηλή ακρίβεια[63], αλληλεπιδράσεις πρωτεϊνών θεωρώντας πως αν τα γονίδια που κωδικοποιούν δύο πρωτεΐνες είναι γειτονικά στο χρωμόσωμα σε πολλά γονιδιώματα, οι αντίστοιχες πρωτεΐνες είναι πιθανό να συνδέονται λειτουργικά. Υψηλά ποσοστά απόδοσης έχει επιφέρει ιδιαίτερα η χρήση της μεθόδου σε προκαρυωτικούς οργανισμούς, των οποίων τα γονίδια συχνά οργανώνονται σε οπερόνια[64]. Από υπολογιστικής πλευράς, η υλοποίηση της διαδικασίας συχνά γίνεται με χρήση τεχνητών νευρωνικών δικτύων και την εκπαίδευση αυτών πάνω σε πληροφορίες από μεγάλες βάσεις δεδομένων[65,66].Η δυσκολία της εφαρμογής της συγκεκριμένης μεθόδου είναι πως απαιτεί τον προσδιορισμό σε πρώτο στάδιο των ορθολόγων σε άλλα γονιδιώματα και κατόπιν την πραγματοποίηση του ελέγχου για την εύρεση εκείνων που ανήκουν στην ίδια γειτονιά.

Μέθοδοι βασισμένες στην Συγχώνευση γονιδίων (Gene Fusion)

Η πιο άμεση εφαρμογή γονιδιακής πληροφορίας για τον προσδιορισμό σχέσεων πρωτεϊνών είναι μέσω μεθόδων συγχώνευσης γονιδίων (Gene Fusion). Η τεχνική αυτή συχνά καλείται και πέτρα της Rosetta ή συγχώνευση αυτοτελών δομικών περιοχών (Rosetta Stone ή Domain Fusion) και στηρίζεται στην υπόθεση πως πρωτεΐνες που κωδικοποιούνται από γονίδια των οποίων τα ομόλογα συγχωνεύονται τείνουν να έχουν σχετική λειτουργία, ειδικά εάν είναι ορθόλογα των γονιδίων αυτών[67,68]. Βασικό στάδιο της μεθόδου είναι η αλληλούχιση με αλγορίθμους των ακολουθιών των πρωτεϊνών προς μελέτη με το γονιδίωμα διαφόρων οργανισμών. Καθώς δεν είναι απαραίτητο πως οι πρωτεΐνες που μελετώνται θα έχουν βρεθεί συγχωνευμένες σε κάποια από τα γνωστά γονιδιώματα, η μέθοδος είναι πιθανό να δώσει ψευδώς αρνητικά αποτελέσματα. Γενικότερα, σε σύγκριση με όλες της υπολογιστικές μεθόδους αυτής της κατηγορίας, η συγκεκριμένη παρουσιάζει την χαμηλότερη κάλυψη[59].

Μέθοδοι βασισμένες στα Φυλογενετικά Προφίλ (Phylogenetic Profiles)

Μια ακόμη υπόθεση που έχει οδηγήσει στην δημιουργία πολλών αλγορίθμων ανίχνευσης αλληλεπιδράσεων πρωτεϊνών είναι πως πρωτεΐνες που έχουν παρόμοια ή ταυτόσημα φυλογενετικά προφίλ τείνουν να συνδέονται και λειτουργικά. Η ιδέα χρήσης των φυλογενετικών προφίλ (Phylogenetic Profiles) προτάθηκε αρχικά το 1999 στην δημοσίευση των Pellegrini et al.[67] στην οποία η ύπαρξη ή η απουσία μιας πρωτεΐνης σε κάθε γνωστό γονιδίωμα αναπαριστάται από μία συμβολοσειρά από δυαδικά ψηφία. Μέσω της μεθόδου αυτής μπορούν να βρεθούν σχέσεις ακόμα και μη χαρακτηρισμένων πρωτεϊνών.

Μέθοδοι βασισμένες στην συνεξέλιξη (Co – evolution)

Ιδιαίτερη κατηγορία μεθόδων είναι εκείνες που στηρίζονται στην συνεξέλιξη των οργανισμών (Co – evolution), εννοώντας την παράλληλη εξέλιξη οργανισμών στο ίδιο οικοσύστημα και με την απαίτηση της ύπαρξης αμοιβαίας επιλεκτικής πίεσης σε δύο ή περισσότερα είδη[69,70]. Η ταυτόχρονη προσαρμογή των ειδών στην πίεση αυτή που ασκείται από το περιβάλλον υποστηρίζεται πως μπορεί να ανιχνευθεί και στο γονιδίωμα τους με την εμφάνιση συγκεκριμένων μεταλλάξεων στις πρωτεΐνες τους. Οι συσχετισμένες αυτές μεταλλάξεις μπορούν επίσης να χρησιμοποιηθούν για τον εντοπισμό των καταλοίπων που εμπλέκονται στις θέσεις αλληλεπίδρασης των πρωτεϊνών[71].

Αγκυροβόληση (Docking)

Παρόλο που οι τεχνικές που στηρίζονται στην πληροφορία του γονιδιώματος είναι ιδιαίτερα δημοφιλής, οι τεχνικές που ανιχνεύουν αλληλεπιδράσεις πρωτεϊνών μέσω της δομής αυτών προσφέρουν περαιτέρω πληροφορίες σχετικά με τα φυσικά χαρακτηριστικά της σχέσης, αλλά και για τα κατάλοιπα στη διεπιφάνεια των πρωτεϊνών τα οποία διαμεσολαβούν στην αλληλεπίδραση. Η πιο αντιπροσωπευτική μέθοδος για την ανίχνευση πρωτεϊνικών αλληλεπιδράσεων μέσω πληροφοριών για την δομή των προσδεμένων πρωτεϊνών είναι η αγκυροβόληση (Docking)[72–75]. Γνωρίζοντας τις δομές δύο μορίων, οι αλγόριθμοι αγκυροβόλησης μπορούν να επιβεβαιώσουν αν αυτές είναι δυνατόν να αλληλεπιδράσουν μεταξύ τους, καθώς και τον προσανατολισμός με τον οποίο βελτιστοποιείται η αλληλεπίδραση, ενώ παράλληλα η ενέργεια του συμπλόκου είναι η ελάχιστη. Ως δεδομένο απαιτούνται μόνο οι συντεταγμένες των δομών των βιομορίων, ενώ στην αγκυροβόληση πρωτεΐνης - πρωτεΐνης και τα δύο μόρια θεωρούνται άκαμπτα. Ένα πρόγραμμα αγκυροβόλησης αποτελείται από έναν αλγόριθμο αναζήτησης (Search algorithm), ο οποίος παράγει ένα αριθμό από πιθανές στερεοδιατάξεις στη θέση πρόσδεσης και μία συνάρτηση βαθμονόμησης (Scoring function), η οποία βαθμολογεί την συγγένεια πρόσδεσης, βάση των στερεοχημικών περιορισμών και των ενεργειακών υπολογισμών. Το μειονέκτημα της μεθόδου είναι το υψηλό υπολογιστικό κόστος, αφού τόσο η αναζήτηση στο σύνολο των βάσεων δεδομένων δομών, όσο και η πραγματοποίηση όλων των απαραίτητων υπολογισμών έχουν αυξημένες απαιτήσεις σε χρόνο και μνήμη.

Μέθοδοι βασισμένες στην αναζήτηση στην βιβλιογραφία

Τα τελευταία χρόνια, με την αύξηση των διαδικτυακών Βάσεων Δεδομένων πρωτεϊνικών πληροφοριών, έχουν υπάρξει αρκετές προσπάθειες ανάπτυξης μεθόδων οι οποίες προβλέπουν συσχετίσεις πρωτεϊνών μέσω αναζήτησης της δημοσιευμένης βιβλιογραφίας, είτε χειροκίνητα είτε με μεθόδους εξόρυξης κειμένου[76–83]. Μεγάλο ποσοστό των βάσεων δεδομένων πρωτεϊνικών αλληλεπιδράσεων που θα παρουσιαστούν στο ακόλουθο κεφάλαιο περιέχει συσχετίσεις ανιχνευμένες μέσω αυτής της μεθόδου.

Δίκτυα αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών

Μία μέθοδος η οποία κάνει παράλληλα χρήση των πληροφοριών για αλληλεπιδράσεις πρωτεϊνών, όπως αυτές προέκυψαν από τις προηγούμενες τεχνικές και ταυτόχρονα υποδεικνύει νέα πιθανά λειτουργικά σύμπλοκα είναι τα δίκτυα αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών. Αποτελούν μία μαθηματική απεικόνιση των δυαδικών σχέσεων ανάμεσα σε ένα σύνολο πρωτεϊνών, πάνω στην οποία μπορούν να εφαρμοστούν όλες οι ιδιότητες και πρακτικές από την Θεωρία

των Γράφων, όπως αυτές περιγράφθηκαν στο προηγούμενο υποκεφάλαιο. Στην παρούσα εργασία θα δοθεί πλήθος παραδειγμάτων κατασκευής δικτύων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών, εύρεσης των βασικών χαρακτηριστικών αυτών, ομαδοποίησης τους και αξιολόγησης των προτεινόμενων συμπλόκων με ποικίλες μετρικές.

1.3 Βάσεις Δεδομένων Αλληλεπιδράσεων Πρωτεϊνών – Πρωτεϊνών (PPIs)

Το πλήθος δεδομένων που προκύπτει από τις πειραματικές και υπολογιστικές μεθόδους, οι οποίες περιγράφθηκαν παραπάνω, αποθηκεύεται σε δημόσια αποθετήρια πληροφοριών, τις Βάσεις Δεδομένων Αλληλεπιδράσεων Πρωτεϊνών – Πρωτεϊνών (Protein – Protein Interactions Databases ή PPIs Databases). Συνήθως κάθε βάση δεδομένων περιέχει πληροφορίες για συγκεκριμένο τύπο αλληλεπιδράσεων ή για μεμονωμένους οργανισμούς, ενώ υπάρχουν και βάσεις που επικεντρώνονται στην συγκέντρωση δεδομένων σχετιζόμενα με βιολογικά μονοπάτια ή ασθένειες. Οι αλληλεπιδράσεις που εμπεριέχει κάθε σύνολο μπορεί να προκύπτουν είτε από μόνο πειραματικές μεθόδους είτε από μόνο υπολογιστικές, ενώ μπορεί ακόμα να προέρχονται και από τις δύο κατηγορίες. Επιπλέον, οι βάσεις χαρακτηρίζονται και σύμφωνα με την προέλευση των δεδομένων τους, όπου μία βάση καλείται πρωταρχική (Primary), όταν οι πληροφορίες έχουν συγκεντρωθεί και σχολιαστεί από τους ερευνητές της ίδιας, ενώ δευτερεύουσα ή βάση μεταδεδομένων (Secondary or Meta-databases) είναι εκείνη της οποίας τα δεδομένα προέρχονται προεπεξεργασμένα από άλλες βάσεις αλληλεπιδράσεων πρωτεϊνών. Δεδομένης της ποικιλίας τόσο των μεθόδων ανίχνευσης των αλληλεπιδράσεων, όσο και συγκέντρωσης αυτών, έχει καταγραφεί σημαντική διαφορά ανάμεσα στις σχέσεις που προτείνει κάθε βάση, ακόμα και για τους ίδιους οργανισμούς, μονοπάτια ή ασθένειες. Συνολικά στο Διαδίκτυο υπάρχουν διαθέσιμες πάνω από εκατό βάσεις αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών, μεγάλο μέρος των οποίων περιγράφετε συνοπτικά στον ακόλουθο συγκεντρωτικό πινάκα [Πίνακας 1], ενώ οι ηλεκτρονικές διευθύνσεις αυτών παρατίθενται στο Παράρτημα.

Databases	Description	Data source
3DID	Βάση δεδομένων με αλληλεπιδράσεις μεταξύ αυτοτελών δομικών περιοχών πρωτεϊνών, των οποίων είναι γνωστές οι υψηλής ανάλυσης τρισδιάστατες δομές. Σύμφωνα με τις πληροφορίες από την δομή, η 3DID παρέχει δεδομένα για το μοριακό υπόβαθρο, την ομοιότητα μεταξύ πρωτεϊνών μελών της ίδιας οικογένειας, καθώς και λειτουργικό χαρακτηρισμό, σύμφωνα με την βάση GO. Εκτός από δομικές αλληλεπιδράσεις, περιέχει σχέσεις πρωτεϊνών του ζυμομύκητα από πειράματα υψηλής απόδοσης.	Μεταδεδομένα
3D-Interlogs	Η 3D-Interlogs είναι βάση δεδομένων με αλληλεπιδράσεις μεταξύ διαφορετικών ειδών, των οποίων οι πληροφορίες απορρέουν από δομικά σύμπλοκα. Τα σύμπλοκα είναι επιπλέον χαρακτηρισμένα από	Μεταδεδομένα

	συνάρτηση βαθμονόμησης. σύμφωνα με τις τρισδιάστατες δομικές περιοχές αλληλεπιδρώντων ομολόγων (3D-domain Interologs). Βασιζόμενη σε αυτές τις δομές και τις συναρτήσεις score, παρέχει τη στατιστική σημαντικότητα, τα αλληλεπιδρώντα μοντέλα και τους λειτουργικούς σχολιασμούς για την εκάστοτε πρωτεΐνη προς μελέτη. Παρέχεται επιπλέον η επιλογή μέσω της βάσης να γίνει χρήση του αλγορίθμου αλληλούχισης BLAST, για τον εντοπισμό ομόλογων πρωτεϊνών από πολλαπλά είδη.	
ACSN	Η ACSN είναι βάση δεδομένων μονοπατιών η οποία εκτός από PPIs, περιέχει σηματοδοτικά μονοπάτια και διαγράμματα των μονοπατιών, κυρίως σχετιζόμενα με την νόσο του καρκίνου, αλλά και μεταβολικά. Η κυτταρική σηματοδότηση και οι μοριακοί μηχανισμοί εκφράζονται στους χάρτες σε επίπεδο βιοχημικών αλληλεπιδράσεων, σχηματίζοντας ένα μεγάλο δίκτυο αντιδράσεων, βασιζόμενο σε πάνω από 4500 επιστημονικές δημοσιεύσεις.	Πρωταρχική
AnimalTFDB	Βάση δεδομένων που συμπεριλαμβάνει ταξινομημένους και σχολιασμένους παράγοντες μεταγραφής (Transcription Factors – TFs), συμπαράγοντες μεταγραφής και παράγοντες διαμόρφωσης χρωματίνης σε 97 αλληλουχίες γονιδιωμάτων ζώων. Οι παράγοντες μεταγραφής ταξινομούνται περεταίρω σε 73 οικογένειες, σύμφωνα με την περιοχής πρόσδεσης DNA τους (DNA-binding domain - DBD) και οι συμπαράγοντες σε 83 οικογένειες και 6 κατηγορίες. Ακόμα παρέχει όλες τις βασικές πληροφορίες σχετικά με τη δομή του γονιδίου, τη λειτουργική περιοχή, το σχολιασμό από την GO, τις πρωτεϊνικές αλληλεπιδράσεις, τα ορθόλογα (Orthologs) και παράλογα (Paralogs) των μεταγραφικών παραγόντων που συμπεριλαμβάνονται στη βάση, καθώς και σχετικούς συνδέσμους σε άλλες βάσεις δεδομένων.	Μεταδεδομένα
AtPIN	Βάση δεδομένων που περιέχει πληροφορίες σχετικά με δίκτυα αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών, την αρχιτεκτονική των αυτοτελών δομικών στοιχείων, πληροφορίες για τα ορθόλογα και σχολιασμό σύμφωνα με την βάση GO, όσον αφορά το πρωτέωμα του οργανισμού Arabidopsis thaliana. Τα ζεύγη πρωτεϊνών - πρωτεϊνών προβλέπονται εφαρμόζοντας διάφορες μεθόδους ταξινόμησης με Naive Bayesian ταξινομητή. Οι υπόλοιπες πληροφορίες που εμπεριέχονται στην AtPIN αντλούνται από την σχετική δημοσιευμένη βιβλιογραφία και άλλες πηγές, από εξειδικευμένους βιολόγους.	Μεταδεδομένα
BIANA	Το BIANA είναι ένα πλαίσιο εργασίας σε γλώσσα Python, σχεδιασμένο να επιτύχει δύο βασικούς στόχους, την ενσωμάτωση πολλαπλών πηγών βιολογικών πληροφοριών, συμπεριλαμβανομένων βιολογικών οντοτήτων και των σχέσεών τους και τη διαχείριση βιολογικών πληροφοριών ως δίκτυο, στο οποίο οι οντότητες είναι κόμβοι και οι σχέσεις ακμή.	Μεταδεδομένα

BioGRID	Πρόκειται για μια από τις δημοφιλέστερες βάσεις δεδομένων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών. Παρέχει πληροφορίες τόσο για τις ίδιες τις πρωτεΐνες και όσο και για τις γενετικές αλληλεπιδράσεις αυτών, σε μια πληθώρα οργανισμών και ειδών. Οι σχέσεις που συμπεριλαμβάνει η βάση προέρχονται από μελέτες υψηλής απόδοσης, καθώς και από επικεντρωμένα πειράματα, με υψηλή αξιοπιστία. Επιπλέον η βάση θεωρείται πως περιέχει το πιο πλήρες σύνολο αλληλεπιδράσεων για τους οργανισμούς Saccharomyces cerevisiae και Schizosaccharomyces pombe, δύο ειδών ζυμομύκητα. Επί του παρόντος η BioGRID περιέχει 1.891.411 διαφορετικές αλληλεπιδράσεις πρωτεϊνών και γονιδίων από 73.056 προσεχτικά μελετημένες δημοσιεύσεις.	Πρωταρχική
CancerNet	Η CancerNet περιέχει αλληλεπιδράσεις miRNA και γονιδίων στόχων, miRNA με miRNA και πρωτεϊνών – πρωτεϊνών, όλων σχετιζόμενων με τον καρκίνο στον ανθρώπινο. Επιτρέπει, επιπλέον, την ανάκτηση μοριακών αλληλεπιδράσεων σε διάφορους τύπους καρκίνου και την ανάλυση εμπλουτισμού για την ανίχνευση σημαντικά υπερεκπροσωπούμενων κατηγοριών στην GO.	Μεταδεδομένα
CCSB Interactome DB	Βάση δεδομένων για αλληλεπίδρασης πρωτεϊνών για ένα πλήθος οργανισμών, συμπεριλαμβανομένων του ανθρώπου, ιών, Caenorhabditis elegans, saccharomyces cerevisiae και άλλων. Όλα τα σύνολα δεδομένων είναι διαθέσιμες για λήψη δωρεάν και μπορούν παράλληλα να απεικονιστούν στην διαδικτυακή διεπαφή της βάσης.	Πρωταρχική
CIDeR	Η CIDeR αποτελεί βάση δεδομένων για αλληλεπιδράσεις μορίων, συμπεριλαμβανομένων των πρωτεϊνών, τα οποία έχουν συσχετιστεί με ασθένειες. Όλες οι αλληλεπιδράσεις σχολιάζονται με μη αυτόματο τρόπο από ειδικευμένο προσωπικό και όλες οι πληροφορίες συνδέονται με τις αντίστοιχες καταχωρήσεις στην βάση PubMed.	Πρωταρχική
ComPPI	Η ComPPI παρέχει ποιοτικές πληροφορίες για τις αλληλεπιδράσεις, τις πρωτεΐνες και τους εντοπισμούς τους, από 4 διαφορετικά είδη, τα S. cerevisiae, C. elegans, D. melanogaster, H. sapiens. Οι πληροφορίες έχουν ενσωματωθεί από 7 διαφορετικές βάσεις δεδομένων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών και περιέχουν συνολικά 1.898.277 σχέσεις.	Πρωταρχική
CORNET	Είναι ένα εργαλείο που αναζητά σε όλες τις διαθέσιμες βάσεις δεδομένων αλληλεπίδρασης πρωτεϊνών τόσο για πειραματικές, όσο και υπολογιστικά, αποδεδειγμένες σχέσεις πρωτεΐνης - πρωτεΐνης. Τα αποτελέσματα μπορούν να απεικονιστούν χρησιμοποιώντας το Cytoscape Webstart, μία διαδικτυακή εκδοχή του εργαλείου Cytoscape, το οποίο θα αναλυθεί πιο λεπτομερώς παρακάτω.	Μεταδεδομένα
CORUM	Η βάση δεδομένων CORUM αποτελεί από τις πιο δημοφιλής βάσεις	Πρωταρχική

	για επιλογή συνόλων αναφοράς, σύμφωνα με τα οποία ελέγχονται αποτελέσματα αλγορίθμων πρόβλεψης πρωτεϊνικών συμπλόκων. Ο βασικός λόγος είναι πως τα σύμπλοκα πρωτεϊνών, για οργανισμούς θηλαστικών, που περιέχει, έχουν επιλεχθεί χειροκίνητα από μεμονωμένα πειράματα, που δημοσιεύονται σε επιστημονικά άρθρα, αποκλείοντας δεδομένα από πειράματα υψηλής απόδοσης.	
COXPRESdb	Βάση δεδομένων η οποία περιλαμβάνει συσχετίσεις πρωτεϊνών από δίκτυα συνεκφρασμένων γονιδίων σε διάφορους οργανισμούς.	Πρωταρχική
dbSNO 2.0	Αποτελεί μία πηγή πληροφοριών για την εξερεύνηση του δομικού περιβάλλοντος των θέσεων υποστρώματος S-νιτροζυλίωσης (SNO) και των ρυθμιστικών δικτύων των πρωτεϊνών SNO. Οι χρήστες μπορούν να αναζητήσουν μια ομάδα πρωτεϊνών ή γονιδίων και το σύστημα ανακατασκευάζει το ρυθμιστικό δίκτυο SNO.	Πρωταρχική
DEPOD	Η DEPOD είναι μια μη αυτόματα σχολιασμένη βάση δεδομένων που συλλέγει ανθρώπινες ενεργές φωσφατάσες, τα πειραματικά επαληθευμένα πρωτεϊνικά και μη-πρωτεϊνικά υποστρώματα τους και τις πληροφορίες και τις οδούς της αποφωσφορυλίωσης, στις οποίες εμπλέκονται. Παρέχει επίσης συνδέσμους προς δημοφιλείς βάσεις δεδομένων κινάσης και αλληλεπιδράσεων πρωτεϊνών για αυτές τις φωσφατάσες και τα υποστρώματα.	Πρωταρχική
DIP	Μια από τις πιο δημοφιλής βάσεις δεδομένων, η οποία καταγράφει πειραματικά υπολογισμένες αλληλεπιδράσεις μεταξύ πρωτεϊνών - πρωτεϊνών. Συνδυάζει πληροφορίες από διάφορες πηγές για να δημιουργήσει ένα ενιαίο και αξιόπιστο σύνολο σχέσεων πρωτεϊνών για πολλούς οργανισμούς. Επιπλέον παρέχει ένα σύνολο εργαλείων και λογισμικών για την ανάλυση αυτών των πληροφοριών, αλλά και την οπτικοποίηση των δικτύων. Τα δεδομένα επεξεργάζονται τόσο χειροκίνητα όσο και αυτοματοποιημένα, αξιοποιώντας γνώσεις για τα δίκτυα πρωτεϊνών – πρωτεϊνών.	Πρωταρχική
FlyBase	Μια βάση δεδομένων που παρέχει πληροφορίες σχετικά με τη Drosophila, όπως γενετικές και πρωτεϊνικές αλληλεπιδράσεις, αλλά και διαδραστικούς γονιδιωματικούς χάρτες.	Μεταδεδομένα
FlyMine	Μια ολοκληρωμένη βάση δεδομένων με γονιδιακά, πρωτεϊνικά και έκφρασης δεδομένα για τα είδη Drosophila, Anopheles και C.elegans. Περιέχει επίσης δίκτυα διαδραστικής αλληλεπίδρασης για γενετικές και φυσικές αλληλεπιδράσεις, μαζί με μεταβολικά και σηματοδοτικά δεδομένα.	Μεταδεδομένα
FunCoup	Ένα στατιστικό πλαίσιο ενσωμάτωσης δεδομένων για εύρεση λειτουργικής σύζευξης (FC) μεταξύ πρωτεϊνών. Μεταφέρει πληροφορίες από 17 μοντέλα οργανισμών, μέσω ορθολογιών που βρέθηκαν από το πρόγραμμα InParanoid. Προέρχεται από	Μεταδεδομένα

	λειτουργικούς συνδέσμους από κυρίως ακατέργαστα δεδομένα πειραμάτων υψηλής απόδοσης ή σχολιασμούς βάσεων δεδομένων και εκτιμά κάθε πληροφορία βάση συνάφειας και αξιοπιστίας.	
GenAge	Μια μη αυτόματα σχολιασμένη βάση δεδομένων γονιδίων που σχετίζονται με τη γήρανση. Τα γονίδια που περιλαμβάνονται σε αυτήν τη βάση σχετίζονται με τη μακροζωία ή τη γήρανση σε μοντέλα οργανισμών, καθώς και τη γήρανση στον άνθρωπο. Οι αλληλεπιδράσεις πρωτεϊνών – πρωτεϊνών απεικονίζονται ως δίκτυο μέσω της βάσης STRING.	Πρωταρχική
GeneMANIA	Η GeneMANIA είναι μια βάση δεδομένων η οποία, δοθέντος ενός γονιδίου προς μελέτη, βρίσκει ένα σύνολο γονιδίων, που σχετίζονται με αυτό, χρησιμοποιώντας γνώση από ένα μεγάλο σύνολο δεδομένων λειτουργικών συσχετίσεων.	Μεταδεδομένα
gpDB	Η βάση δεδομένων gpDB δημιουργήθηκε από το εργαστήριο Βιοφυσικής και Υπολογιστικής Βιολογίας, του τμήματος Βιολογίας του ΕΚΠΑ. Περιλαμβάνει πληροφορίες G-πρωτεϊνών και αλληλεπιδράσεις αυτών με GPCRs (G-protein Coupled Receptors). Υπάρχουν πλήρης πληροφορίες ακολουθίας με παραπομπές σε άλλες βάσεις δεδομένων. Επιπλέον ο ιστότοπος παρέχει εργαλεία αναζήτησης κειμένου, αλληλούχισης με τον αλγόριθμο BLAST και εύρεσης μοτίβου.	Πρωταρχική
GPS-Prot	Μια βάση δεδομένων που περιλαμβάνει πάνω από 300.000 αλληλεπιδράσεις πρωτεϊνών – πρωτεϊνών στον άνθρωπο. Οι αλληλεπιδράσεις αυτές είναι βαθμολογημένες και προέρχονται από μεγάλες βάσεις, με μη αυτοματοποιημένο σχολιασμό δεδομένων. Περιλαμβάνονται επίσης οι αλληλεπιδράσεις του ιού HIV-1.	Μεταδεδομένα
HIV-1 Interactions DB	Μια συνοπτική, αλλά λεπτομερής περίληψη όλων των γνωστών αλληλεπιδράσεων των πρωτεϊνών του HIV-1 με πρωτεΐνες κυττάρων ξενιστών, άλλων πρωτεϊνών HIV-1 ή πρωτεϊνών από οργανισμούς ασθένειας που σχετίζονται με τον HIV.	Πρωταρχική
HIVMID	Μια βάση δεδομένων που περιέχει πληροφορίες σχετικά με τους κυτταροτοξικούς, αλλά και τους βοηθητικούς αντιγονικούς καθαριστές των Τ-κυττάρων και τις θέσεις πρόσδεσης αντισωμάτων του HIV-1.	Μεταδεδομένα
HP-DPI	Η HP-DPI αποτελεί μία βάση δεδομένων από πειραματικά και υπολογιστικά προσδιορισμένες αλληλεπιδράσεις πρωτεϊνών – πρωτεϊνών στο Helicobacter pylori. Οι προβλέψεις των σχέσεων βασίζονται σε στατιστικά στοιχεία από την αλληλεπίδραση αυτοτελών δομικών στοιχείων μεταξύ τους	Μεταδεδομένα
HPID	Είναι μία συλλογή αλληλεπιδράσεων πρωτεϊνών του ανθρώπινου οργανισμού, οι οποίες είναι προ-υπολογισμένες με στατιστική μέθοδο	Μεταδεδομένα

	από υπάρχοντα δομικά και πειραματικά δεδομένα. Οι πληροφορίες αυτές προέρχονται από τις βάσεις BIND, DIP και HPRD.	
HPRD	Η πιο διαδεδομένη βάση δεδομένων με αμιγώς ανθρώπινες πρωτεϊνικές αλληλεπιδράσεις. Δίνονται για κάθε μία από τις αλληλεπιδράσεις πληροφορίες για τον τύπο πειράματος ανίχνευσης, τα επίπεδα έκφρασης, την αρχιτεκτονική της δομής της, καθώς και για πιθανές μετα-μεταφραστικές τροποποιήσεις.	Πρωταρχική
I2D	Η I2D είναι βάση δεδομένων αλληλεπιδράσεων ανθρώπινων πρωτεϊνών, όπως αυτές περιέχονται σε άλλες βάσεις πρωτεϊνικών πληροφοριών. Επιπλέον ζεύγη πρωτεϊνών προέκυψαν από χαρτογράφηση αποτελεσμάτων πειραμάτων υψηλής απόδοσης πρότυπων οργανισμών σε ανθρώπινες πρωτεΐνες.	Μεταδεδομένα
IMEx	Μια διεθνής συνεργασία μεταξύ σημαντικών δημόσιων παρόχων, όπως η DIP, IntAct, BioGRID κ.α., με δεδομένα αλληλεπίδρασης πρωτεϊνών – πρωτεϊνών, οι οποίοι διαμοιράζονται την επιμέλεια των δεδομένων. Στόχος του εγχειρήματος είναι να οριστεί ένα βασικό σύνολο κανόνων για των σχολιασμό των σχέσεων, η καταγραφή όλων των δυνατών πληροφοριών για κάθε αλληλεπίδραση, η συγκέντρωση όλων αυτών σε μία διαδικτυακή πηγή και η αποθήκευση τους σε μία μορφή αρχείου ειδική για το κατέβασμα και την επεξεργασία τους στον τοπικό υπολογιστή του κάθε χρήστη.	Μεταδεδομένα
IntAct	Από τις βάσεις δεδομένων με την μεγαλύτερη επισκεψιμότητα και η βάση με τις περισσότερες πληροφορίες για τις αλληλεπιδράσεις πρωτεϊνών – πρωτεϊνών σε πλήθος ασθενειών. Η IntAct διατηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) και επί του παρόντος περιέχει 1.118.895 δυαδικές μοριακές σχέσεις, από διάφορους οργανισμούς, προσδιορισμένες πειραματικά, μέσω βιβλιογραφία και με αναζήτηση στοχευμένων όρων.	Πρωταρχική
INTERSPIA	Είναι μια διαδικτυακή εφαρμογή για την ανάλυση και οπτικοποίηση της δυναμικής των αλληλεπιδράσεων μεταξύ πρωτεϊνών - πρωτεϊνών ανάμεσα σε διάφορα είδη. Η εφαρμογή διερευνά ένα δίκτυο αλληλεπίδρασης πρωτεΐνης - πρωτεΐνης για να βρει άμεσα ή έμμεσα σχετιζόμενες πρωτεΐνες στη βάση δεδομένων STRING, χρησιμοποιώντας τον αλγόριθμο τυχαίας διαδρομής και απεικονίζει τις ομοιότητες και τις διαφορές αλληλεπιδράσεων πρωτεΐνης - πρωτεΐνης μεταξύ πολλών ειδών χρησιμοποιώντας τις ορθολογικές πληροφορίες της πρωτεΐνης.	Μεταδεδομένα
iRefWeb	Παρέχει μια διαδικτυακή διεπαφή ερωτήματος σε μια σχεσιακή βάση δεδομένων που περιέχει την τελευταία έκδοση του δείκτη αναφοράς αλληλεπίδρασης. Ο δείκτης αναφοράς αλληλεπίδρασης ενσωματώνει δεδομένα αλληλεπίδρασης πρωτεΐνης από τις βάσεις BIND, BioGRID, IntAct, MINT, MPPI και OPHID.	Μεταδεδομένα

MatrixDB	Βάση δεδομένων που αναφέρει αλληλεπιδράσεις πρωτεΐνης - πρωτεΐνης θηλαστικών και πρωτεΐνης - υδατανθράκων, οι οποίες περιλαμβάνουν εξωκυτταρικά μόρια.	Πρωταρχική
Mentha	Η Mentha συλλέγει δεδομένα από μη αυτόματα σχολιασμένες βάσεις δεδομένων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών, οι οποίες ανήκουν στην κοινοπραξία IMEx. Τα συγκεντρωτικά δεδομένα σχηματίζουν ένα δίκτυο αλληλεπιδράσεων που περιλαμβάνει πολλούς οργανισμούς (Homo sapiens, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Escherichia coli K12, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae), καθώς και ένα επιπλέον παγκόσμιο δίκτυο, το οποίο περιλαμβάνει όλους τους οργανισμούς, συμπεριλαμβανομένων και αυτών που δεν αναφέρθηκαν πριν.	Μεταδεδομένα
MINT	Η MINT επικεντρώνεται σε πειραματικά επαληθευμένες αλληλεπιδράσεις πρωτεΐνης - πρωτεΐνης που αντλούνται από την επιστημονική βιβλιογραφία από ειδικευμένους επιμελητές. Ενσωματώνεται επίσης με την βάση HomoMINT, μη οποία περιέχει δεδομένα μοριακών αλληλεπιδράσεων, που συνάγονται από ορθολογικές πρωτεΐνες πρότυπων οργανισμών και την βάση virusMINT, με αλληλεπιδράσεις μεταξύ ανθρώπινων και ιικών πρωτεϊνών.	Πρωταρχική
Pathway Commons	Μια συλλογή από ελεύθερα διαθέσιμα μονοπάτια ποικίλων οργανισμών. Τα δεδομένα συλλέγονται από βάσεις δεδομένων συνεργατών και αντιπροσωπεύονται μέσω του προτύπου BioPAX, παρέχοντας ποικίλες πληροφορίες όπως τις βιοχημικές αντιδράσεις, τα ρυθμιστικά δίκτυα γονιδίων, γενετικές αλληλεπιδράσεις, διαδικασίες μεταφοράς και κατάλυσης, καθώς και τις φυσικές αλληλεπιδράσεις μεταξύ πρωτεϊνών, DNA, RNA και μικρών μορίων.	Μεταδεδομένα
PepBank	Μια βάση δεδομένων αλληλεπιδράσεων πρωτεΐνης – πρωτεΐνης, που βασίζεται στην εξόρυξη πληροφορίας από δημοσιεύσεις και βάσεις δεδομένων για πεπτίδια. Ποσοστό των αλληλεπιδράσεων αυτών σχολιάζεται περαιτέρω χειροκίνητα.	Πρωταρχική
PhosphoELM	Περιέχει μια συλλογή πειραματικά επαληθευμένων θέσεων σερίνης, θρεονίνης και τυροσίνης σε πρωτεΐνες ευκαρυωτικών οργανισμών. Επιπλέον παρέχονται πληροφορίες σχετικά με τη δομή, τα μοτίβα αλληλεπίδρασης και το υποκυτταρικό διαμέρισμα τους.	Πρωταρχική
PIMADb	Μια συλλογή λεπτομερειών πλευρικών αλληλεπιδράσεων όλων των συμπλόκων πρωτεΐνης που διατίθενται στην PDB. Επί του παρόντος, υπάρχουν 155.000 αλληλεπιδράσεις από 53.000 σύμπλοκα της PDB. Η βάση αναζητεί αυτοματοποιημένα την PDB για νέες εγγραφές και ενημερώνεται εβδομαδιαίως.	Μεταδεδομένα

PINA2	Είναι μια ολοκληρωμένη πλατφόρμα για την κατασκευή, το φιλτράρισμα, την ανάλυση, την οπτικοποίηση και την διαχείριση δικτύου αλληλεπίδρασης πρωτεΐνης - πρωτεΐνης. Ενσωματώνει δεδομένα αλληλεπιδράσεων από 6 βάσεις δεδομένων και δημιουργεί ένα πλήρες, μη επικαλυπτόμενο σύνολο δεδομένων για 6 μοντέλα οργανισμών.	Μεταδεδομένα
PIPs	Μια βάση δεδομένων προβλεπόμενων ανθρώπινων πρωτεϊνικών αλληλεπιδράσεων, που δημιουργήθηκε χρησιμοποιώντας έναν στοχαστικό Bayesian ταξινομητή για να υπολογίσει μια βαθμολογία αλληλεπίδρασης.	Μεταδεδομένα
PrePPI	Μια βάση δεδομένων για τις πειραματικά και υπολογιστικά προσδιορισμένες αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης (PPI) για το ανθρώπινο πρωτέωμα. Οι αλληλεπιδράσεις στη βάση δεδομένων προσδιορίζονται χρησιμοποιώντας ένα Bayesian πλαίσιο εργασίας που συνδυάζει δομικές, λειτουργικές, εξελικτικές και έκφρασης πληροφορίες. Επί του παρόντος, το PrePPI περιέχει 1.350.000 αλληλεπιδράσεις καλύπτοντας το 85% του ανθρώπινου πρωτεώματος.	Πρωταρχική
SIGNOR	Το SIGNOR είναι ένα εργαλείο που οργανώνει και αποθηκεύει σε δομημένη μορφή πληροφορίες σηματοδότησης, οι οποίες δημοσιεύονται στην επιστημονική βιβλιογραφία. Αυτές οι πληροφορίες αποθηκεύονται ως δυαδικές αιτιώδεις σχέσεις μεταξύ μορίων και μπορούν να αναπαρασταθούν γραφικά ως ροή.	Πρωταρχική
STRING	Η STRING αποτελεί μία ακόμα πολύ δημοφιλής πηγή άντλησης πληροφοριών αλληλεπίδρασης πρωτεΐνης - πρωτεΐνης. Δημιουργήθηκε ως ένα εργαλείο εύρεσης και ανάλυσης σχέσεων μεταξύ τόσο πρωτεϊνών, όσο και γονιδίων, ενώ οι σχέσεις αυτές προκύπτουν και από γνωστά σύμπλοκα και από σύμπλοκα πρωτεϊνών τα οποία έχουν προβλεφθεί υπολογιστικά. Τα ζεύγη που υποδεικνύει σχετίζονται κυρίως από προσδιορισμένες και ειδικές λειτουργικές σχέσεις ή από άμεση φυσική επαφή των πρωτεϊνών ή των γονιδίων. Οι αλληλεπιδράσεις, τέλος, απεικονίζονται μέσω διαδραστικών δικτύων, τα οποία με απλό και σαφή τρόπο δηλώνουν τον τύπο της αλληλεπίδρασης, τον τρόπο προσδιορισμού της και την στατιστική σημαντικότητα αυτής.	Μεταδεδομένα
Struct2Net	Το Struct2Net αποτελεί ένα διαδικτυακό εργαλείο, το οποίο δοσμένης από τον χρήστη μίας αμινοξικής ακολουθίας, σε FASTA μορφή, έχει την ικανότητα εύρεσης των αλληλεπιδράσεων πρωτεΐνης – πρωτεΐνης αυτής, ανάμεσα σε όλες της δομές πρωτεϊνών της βάσης PDB, χρησιμοποιώντας υπολογιστικές τεχνικές βάση πληροφορίας δομής.	Πρωταρχική
Tabloid	Πρόκειται για μια βάση δεδομένων δικτύων συσχέτισης πρωτεϊνών που δημιουργούνται χρησιμοποιώντας πληροφορίες από δημόσια διαθέσιμα πειραματικά αποτελέσματα, με βάση τη φασματομετρία	Μεταδεδομένα

	μάζας, ενώ ο βαθμός της συσχέτισης για κάθε σύνολο σχέσεων είναι βάση της Jaccard Similarity. Εκτός από απλές δυαδικές σχέσεις μεταξύ πρωτεΐνης – πρωτεΐνης, η βάση παρέχει πληροφορίες για ολόκληρα σύμπλοκα πρωτεϊνών και τον ρόλο σε μεταβολικά μονοπάτια.	
The Signaling Pathway Project	Η βάση δεδομένων Signaling Pathways Project περιλαμβάνει σχέσεις πρωτεϊνών, γονιδίων και μικρών μορίων, που εμπλέκονται σε σηματοδοτικά μονοπάτια του κυττάρου των θηλαστικών. Ο ρόλος των μορίων στα μονοπάτια χωρίζεται σε τρεις κατηγορίες: υποδοχείς, μεταγραφικοί παράγοντες ή ένζυμα. Οι πληροφορίες της βάσης προέρχονται από καλά σχολιασμένα δεδομένα άλλων βάσεων, καθώς και από σύνολα δεδομένων πειραμάτων ChIP - Seq.	Μεταδεδομένα
TRIP DB	Μία χειροκίνητα σχολιασμένη βάση δεδομένων για αλληλεπιδράσεις πρωτεΐνης - πρωτεΐνης στα κανάλια TRP των θηλαστικών. Η βάση περιέχει 706 ζεύγη πρωτεϊνών για ένα σύνολο 28 καναλιών για τα οποία δίνονται οι πληροφορίες μεθόδου εύρεσης τους, αξιολόγησης της συσχέτισης, βιοχημικών ιδιοτήτων τους, καθώς και του βιολογικού τους ρόλου.	Πρωταρχική

Πίνακας 1: Βάσεις δεδομένων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών

Όλες οι παραπάνω βάσεις που περιγράφθηκαν αποτελούν πήγες άντλησης δυαδικών σχέσεων πρωτεΐνης – πρωτεΐνης, βάση των οποίων μπορούν να κατασκευαστούν δίκτυα αλληλεπιδράσεων και εν συνεχεία αυτά να αναλυθούν, αλλά και να ομαδοποιηθούν μέσω ειδικών αλγορίθμων, για την εξαγωγή πιθανών συμπλόκων πρωτεϊνών με βιολογική σημασία. Η διαδικασία αυτή πραγματοποιήθηκε στην παρούσα διπλωματική, με στόχο την σύγκριση και αξιολόγηση μεθόδων ομαδοποίησης δικτύων αλληλεπιδράσεων πρωτεΐνης – πρωτεΐνης και παρουσιάζεται αναλυτικά στα κεφάλαια που ακολουθούν.

2. Δεδομένα και Μέθοδοι

2.1 Σύνολα Δεδομένων Αλληλεπιδράσεων Πρωτεϊνών – Πρωτεϊνών

Στην ενότητα αυτή θα γίνει μία σύντομη παρουσίαση των βασικών χαρακτηριστικών των συνόλων δεδομένων που επιλέχτηκαν προς ανάλυση στην παρούσα διπλωματική. Η περιγραφή αυτών δεν θα επεκταθεί σε τοπολογικά χαρακτηριστικά, καθώς αυτά θα περιγραφθούν αναλυτικά σε επόμενο κεφάλαιο. Επιπλέον, καθώς η πλειοψηφία των αναφερόμενων συνόλων σχηματίζουν αρκετά πολύπλοκα δίκτυα αλληλεπιδράσεων, αυτά θα απεικονιστούν, με του βοήθεια του εργαλείου Cytoscape[84], με στόχο την βέλτιστη κατανόηση των ιδιοτήτων αυτών.

Το Cytoscape είναι ένα πρόγραμμα λογισμικού ανοιχτού κώδικα για την ενσωμάτωση δικτύων βιομοριακής αλληλεπίδρασης με δεδομένα έκφρασης υψηλής απόδοσης σε ένα ενοποιημένο θεματικό πλαίσιο. Παρόλο που μπορεί να εφαρμοστεί σε οποιοδήποτε σύστημα μοριακών στοιχείων και αλληλεπιδράσεων, το Cytoscape είναι πιο ισχυρό όταν χρησιμοποιείται σε συνδυασμό με πληροφορίες από βάσεις δεδομένων με σχέσεις πρωτεϊνών-πρωτεϊνών, πρωτεϊνών-DNA ή γενετικών αλληλεπιδράσεων ανθρώπων και άλλων οργανισμών. Το λογισμικό του παρέχει βασικές λειτουργίες για να διαταχθεί το δίκτυο, να γίνουν αναζητήσεις μέσα σε αυτό, να πραγματοποιηθεί οπτική ενσωμάτωση του δικτύου με προφίλ έκφρασης, φαινοτύπους και άλλες μοριακές καταστάσεις, ενώ επιπλέον έχει την δυνατότητα εμπλουτίσει το δίκτυο με γνώσεις από βάσεις δεδομένων λειτουργικών σχολιασμών[84]. Ανάμεσα στα βασικά επιπρόσθετα εργαλεία του Cytoscape, το οποίο χρησιμοποιήθηκε και στην παρούσα εργασία για την εύρεση των βασικών τοπολογικών παραμέτρων των δικτύων αλληλεπιδράσεων πρωτεϊνών, είναι το Network Analyzer[85].

Η πρώτη ομάδα συνόλων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών αντλήθηκε από την βάση δεδομένων DIP, η οποία, όπως αναφέρθηκε και σε προηγούμενη ενότητα εκτενέστερα, αποτελεί από τις πιο διαδεδομένες πηγές πληροφοριών αλληλεπιδράσεων πρωτεϊνών. Πιο συγκεκριμένα, τα σύνολα που επιλέχθηκαν περιέχουν πληροφορίες για τις σχέσεις πρωτεϊνών από πέντε διαφορετικούς, πολύ καλά μελετημένους, πρότυπους οργανισμούς: Caenorhabditis elegans, Drosophila, Escherichia coli, Mouse, Yeast και επιπλέον PPIs για τον Homo sapiens. Όσον αφορά τους κόμβους και τις ακμές των δικτύων αυτών, το σύνολο του Caenorhabditis elegans αποτελείται από 2.423 κόμβους και 3.628 ακμές, το σύνολο της Drosophila από 972 κόμβους και 1.401 ακμές, του είδους Escherichia coli από 2.924 κόμβους

και 12.246 ακμές, ενώ τα σύνολα των οργανισμών Mouse και Yeast από 2.246 κόμβους – 2.495 ακμές και 5.124 κόμβους – 22.908 ακμές, αντίστοιχα. Τέλος, το δίκτυο που περιγράφει τις αλληλεπιδράσεις ανθρώπινων πρωτεϊνών αποτελείται από 4.615 κόμβους, οι οποίοι ενώνονται μέσω 7.417 ακμών.

Η επόμενη ομάδα δεδομένων είναι από την βάση δεδομένων IntAct και περιλαμβάνει σύνολα αλληλεπιδράσεων πρωτεϊνών σχετιζόμενα με ασθένειες και συγκεκριμένα με καρδιακά νοσήματα, διάφορους τύπους καρκίνου, τις νόσους Alzheimer και Parkinson, καθώς και για τον ιό Covid19, υπεύθυνο για την παρούσα πανδημία. Καθώς ο καρκίνος αποτελεί από τα πλέον σοβαρά νοσήματα για τον άνθρωπο και αντικείμενο ενδιαφέροντος της πλειοψηφίας των βιοϊατρικών μελετών, είναι λογικό πως σχηματίζει το μεγαλύτερο δίκτυο εκ των πέντε με 4.375 πρωτεΐνες να αλληλεπιδρούν μέσω 16.615 ακμών. Αντίστοιχα θεωρείται λογικό το σύνολο του ιού Covid19 να είναι το μικρότερο, με μόλις 646 κόμβους και 1.179 ακμές, αφού πρωτοεμφανίστηκε τους τελευταίους μήνες και οι πληροφορίες, που έχει στη διάθεση της η επιστημονική κοινότητα είναι ακόμα λίγες. Όσον αφορά τα δίκτυα των υπόλοιπων ασθενειών δεύτερο σε πλήθος κόμβων και ακμών βρίσκεται της νόσου του Parkinson με 3.222 και 6.844, ακολουθεί της νόσου του Alzheimer με 2.381 και 4.078 και τέλος το σύνολο πρωτεϊνικών αλληλεπιδράσεων σχετιζόμενο με καρδιακά νοσήματα, που αποτελείται από 1.745 κόμβους και 3.611 ακμές.

Ακολούθως επιλέχθηκαν 3 σύνολα αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών, τα οποία αποτελούν από τα πιο δημοφιλή στην σχετική βιβλιογραφία, και περιγράφουν σχέσεις πρωτεϊνών του οργανισμού ζυμομύκητα. Ο συγκεκριμένος οργανισμός αποτελεί τον συχνότερα επιλεγμένο σε σχετικές μελέτες, γεγονός που αποτέλεσε τον λόγο επιλογής του και εδώ. Συγκεκριμένα τα σύνολα που αναλύθηκαν είναι τα Gavin et al. 2002[86], δίκτυο 1.352 κόμβων και 3.210 ακμών, το δίκτυο που προτάθηκε από την επιστημονική ομάδα της ίδιας ερευνήτριας λίγα χρόνια αργότερα Gavin et al. 2006[87] και αποτελείται από 1.430 κόμβους και 6.531 ακμές και το σύνολο δεδομένων των Krogan et al. 2006[88] με 2.674 κόμβους και 7.079 ακμές. Στα τρία παραπάνω σύνολα προστέθηκαν, κατά την δεύτερη μέθοδο αξιολόγησης των αλγορίθμων, περισσότερα για την οποία θα ειπωθούν παρακάτω, δύο ακόμα σύνολα με πρωτεΐνες του οργανισμού ζυμομύκητα. Το πρώτο εκ των δύο προέρχεται από την βάση δεδομένων πρωτεϊνικών αλληλεπιδράσεων BioGRID και σχηματίζει ένα δίκτυο πρωτεϊνών με 5.640 πρωτεΐνες και 59.748 αλληλεπιδράσεις αυτών. Το τελευταίο σύνολο αυτής της ομάδας προέρχεται από ένα σύνολο αναφοράς, με πειραματικά αποδεδειγμένα δηλαδή σύμπλοκα πρωτεϊνών, από το πείραμα των Pu et al. το 2008[89] και ονομάζεται CYC2008. Το σύνολο αποτελείται από 408 σύμπλοκα ετερομερών πρωτεϊνών, ενώ ο αριθμός των πρωτεϊνών μέσα σε αυτά είναι 1.630.

Τέλος, όπως προτείνεται σε παρόμοιες μελέτες[90–93] χρησιμοποιήθηκε ένα πλήθος συνόλων, από την βάση BioGRID, από διάφορους οργανισμούς, στα οποία πριν την εκτέλεση των αλγορίθμων ομαδοποίησης σε αυτά, πραγματοποιήθηκε προ-επεξεργασία, κατά την οποία αφαιρέθηκαν οι βρόχοι (self-loops). Τα σύνολα αυτά έγινε προσπάθεια να καλύπτουν την πλειοψηφία των δικτύων, που είναι πιθανό να εκτελεστούν με τους αλγορίθμους ομαδοποίησης, τόσο σε σχέση με το μέγεθος, όσο και προς την τοπολογία αυτών, ώστε στα αποτελέσματα να υπάρξει η απαραίτητη πληρότητα και αξιοπιστία.

Ο Πίνακας 2, που ακολουθεί, συγκεντρώνει τα σύνολα δεδομένα που χρησιμοποιήθηκαν.

Όνομα	Κόμβοι	Ακμές
DIP_C.elegans	**2423**	**3628**
DIP_Drosophila	**972**	**1401**
DIP_E.coli	**2924**	**12246**
DIP_Mouse	**2246**	**2495**
DIP_Yeast	**2495**	**5124**
DIP_Homo Sapiens	**4615**	**7417**
IntAct_Cardiac	**1745**	**3611**
IntAct_Cancer	**4375**	**16615**
IntAct_Alzheimer	**2381**	**4078**
IntAct_Parkinson	**3222**	**6844**
IntAct_Covid19	**646**	**1179**
Gavin et al. 2002	**1352**	**3210**
Gavin et al. 2006	**1430**	**6531**
Krogan et al. 2006	**2674**	**7079**

BioGRID_Yeast	**5640**	**59748**
BioGRID_Arabidopsis_no_self_loops	**6580**	**32265**
BioGRID_C.elegans_no_self_loops	**1435**	**6502**
BioGRID_Chicken_Gallus_no_self_loops	**267**	**332**
BioGRID_Homo_Sapiens_no_self_loops	**23259**	**583505**
BioGRID_Mouse_no_self_loops	**14497**	**72968**
BioGRID_Rat_no_self_loops	**2956**	**5729**
BioGRID_Spompe_no_self_loops	**4500**	**77479**
BioGRID_Oryza_no_self_loops	**120**	**150**
BioGRID_Taurus_no_self_loops	**374**	**377**
Benchmark Dataset – CYC2008	**1630**	**408**

Πίνακας 2: Συγκεντρωτικός πίνακας συνόλων δεδομένων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών και συμπλόκων πρωτεϊνών που χρησιμοποιήθηκαν στην παρούσα μελέτη

2.2 Αλγόριθμοι Ομαδοποίησης (Clustering)

Τα σύνολα δεδομένων, τα οποία παρουσιάστηκαν στο προηγούμενο υποκεφάλαιο, αποτέλεσαν την είσοδο των έξι αλγορίθμων ομαδοποίησης, που συγκρίνει η παρούσα διπλωματική. Όλοι οι αλγόριθμοι που επιλέχθηκαν αποτελούν network - based προσεγγίσεις, οι οποίες χωρίς επίβλεψη (unsupervised), προσπαθούν να βρουν ομάδες (clusters) πρωτεϊνών μέσα στα δίκτυα αλληλεπιδράσεων, οι οποίες αποτελούν πιθανά σύμπλοκα (complexes) με βιολογική σημασία. Οι συγκεκριμένοι αλγόριθμοι επιλέχθηκαν με κριτήρια: α) Την διαθεσιμότητα τους, είτε online είτε του πηγαίου κώδικα τους, γεγονός που αποτέλεσε και τον βασικότερο περιορισμό· β) την δυνατότητα εφαρμογής τους σε δίκτυα στα οποία δεν διατίθεται η πληροφορία των βαρών, χωρίς αυτή η έλλειψη να επηρεάζει σημαντικά τα αποτελέσματα, αφού τα περισσότερα δεδομένα δεν περιλάμβαναν την σχετική πληροφορία· γ) να μην γίνεται χρήση σε κάποιο στάδιο της μεθόδου επιπλέον βιολογικής πληροφορίας, διότι συνήθως η λογική πίσω από τέτοιου τύπου αλγορίθμους είναι αρκετά διαφορετική από ότι στους κλασσικούς αλγορίθμους ομαδοποίησης δικτύων, οπότε και τα αποτελέσματα δεν θα ήταν συγκρίσιμα· δ) τέλος, παράλληλα με τα παραπάνω, λήφθηκαν υπόψην οι αναφορές στην βιβλιογραφία, τόσο ο αριθμός αυτών όσο και η αξιολόγηση των αποτελεσμάτων τους. Παρακάτω αναλύονται περαιτέρω οι έξι αλγόριθμοι ομαδοποίησης: Affinity Propagation, ClusterONE, MCL, MCODE, NCMine, SPICi.

2.2.1. Affinity Propagation

Ο αλγόριθμος ομαδοποίησης Affinity Propagation[94] σχεδιάστηκε το 2007 από τους Brendan Frey και Delbert Dueck και δέχεται ως είσοδο μέτρα ομοιότητας μεταξύ ζευγών κόμβων (data points) θεωρώντας κάθε ένα σημείο ως πιθανό cluster – “exemplar”. Ακολούθως η μέθοδος μεταδίδει αναδρομικά πραγματικής πληροφορίας (real-valued) μηνύματα μεταξύ των κόμβων, μέχρι να βρει ένα υπομονοπάτι, το οποίο να επιτρέπει την εύκολη μεταφορά της πληροφορίας.

Αναλυτικότερα τα βήματα του αλγορίθμου όπως περιγράφονται στην πρωτότυπη δημοσίευση είναι:

- Έστω σημεία ενδιαφέροντος x_i και x_k των οποίων υπολογίζεται η ομοιότητα μέσω της Ευκλείδειας απόστασης, δηλαδή, $s(i,k) = -||xi - xk||^2$
- Για κάθε σημείο k λαμβάνεται ως είσοδος η διαγώνιος της απόστασης, *s(k,k)*, με τέτοιον τρόπο ώστε όσο μεγαλύτερος είναι αυτός ο πραγματικός αριθμός, τόσο πιο πιθανό είναι να επιλεχθεί το συγκεκριμένο σημείο ενδιαφέροντος ως πιθανό σύμπλοκο. Η τιμή αυτή χαρακτηρίζεται ως “preference” και είναι ιδιαίτερης σημασίας καθώς επηρεάζει σημαντικά τον τελικό αριθμό των

ομάδων. Συνήθως η τιμή αρχικοποιείται στην μέση ομοιότητα όλων των ζευγών εισόδου.

- Ο αλγόριθμος συνεχίζει εναλλάσσοντας δύο βήματα μετάδοσης μηνυμάτων, τα οποία ενημερώνουν αντίστοιχα δύο πίνακες:
  - Και οι δύο πίνακες αρχικοποιούνται με μηδέν και μπορούν να θεωρηθούν ως πίνακες log-probability.
  - Ο πίνακας R (Rensposibility) έχει τιμές *r(i, k)* που ποσοτικοποιούν το πόσο καλά ταιριάζει το xk για να χρησιμεύσει ως exemplar για το xi, σε σύγκριση με άλλα σημεία, βάση της συνάρτησης: *r(I,k) ←s(i,k) - $max_{k\prime\neq k}${ a(i,k') + s(i,k') }*
  - Ο πίνακας A (Availability) περιέχει τιμές a (i, k) που αντιπροσωπεύουν πόσο "κατάλληλο" θα ήταν για το xi να επιλέξει το xk ως exemplar, λαμβάνοντας υπόψη την προτίμηση άλλων σημείων για το xk, μέσω της συνάρτησης: *a(i,k) ←min{ 0, r(k,k) +* $\sum_{i'\not\in\{i,k\}} max(0, r(i', k)\}$*}*
- Οι επαναλήψεις εκτελούνται έως ότου ή τα όρια του cluster παραμείνουν αμετάβλητα για κάποιες επαναλήψεις ή μέχρι να επιτευχθεί κάποιος προκαθορισμένος αριθμός επαναλήψεων (Number of Iterations).
- Τα πιθανά σύμπλοκα εξάγονται από τους τελικούς πίνακες, ως εκείνα που πληρούν την συνθήκη *r(i,i)+a(i,i))>0*

2.2.2. Clustering with Overlapping Neighborhood Expansion (ClusterONE)

Ο ClusterONE[95] αποτελεί μέθοδο εύρεσης επικαλυπτόμενων ομάδων πρωτεϊνών από δεδομένα αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών. Δημιουργήθηκε το 2012 από τους Tamás Nepusz, Haiyuan Yu και Alberto Paccanaro και στηρίζεται στην χρήση ενός μέτρου συνεκτικότητας (cohesiveness) για την εύρεση των ομάδων εκείνων που είναι πιο πιθανό να αποτελούν σύμπλοκα πρωτεϊνών. Συγκεκριμένα η πορεία του αλγορίθμου ακολουθεί τρία βασικά στάδια:

- Με σημείο εκκίνησης τον κόμβο-πρωτεΐνη με τον μεγαλύτερο αριθμό ακμών ξεκινάει μια “άπληστη” (greedy) διαδικασία η οποία προσθέτει ή αφαιρεί κόμβους, ώστε να βρει ομάδες με υψηλή συνεκτικότητα βάση της συνάρτησης: *f(V)* = $\frac{W^{in}(V)}{W^{in}(V)+W^{bound}(V)+p|V|}$, όπου $W^{in}(V)$ το σύνολο των βαρών των ακμών της ομάδας πρωτεϊνών V και $W^{bound}(V)$ το σύνολο των βαρών των ακμών που ενώνουν την συγκεκριμένη ομάδα με το υπόλοιπο δίκτυο. Το βήμα αυτό επαναλαμβάνεται για όλες τις πρωτεΐνες, που δεν έχουν συμπεριληφθεί σε κανένα από τα πρωτεϊνικά σύμπλοκα, που έχουν

βρεθεί μέχρι στιγμής, με προτεραιότητα εκείνες με υψηλότερο βαθμό συνεκτικότητας.

- Στο δεύτερο βήμα, ποσοτικοποιείται η έκταση της αλληλοεπικάλυψης μεταξύ κάθε ζεύγους ομάδων και συγχωνεύονται εκείνες για τις οποίες η βαθμολογία επικάλυψης είναι πάνω από ένα καθορισμένο όριο (Overlap Threshold).
- Τέλος, ο ClusterONE εξάγει τις ομάδες των πρωτεϊνών, οι οποίες περιέχουν τουλάχιστον τρεις πρωτεΐνες ή των οποίων η πυκνότητα είναι μεγαλύτερη από ένα δεδομένο όριο δ (Minimum Density). Στην εκτέλεση του αλγορίθμου μέσω της πλατφόρμας του Cytoscape είναι εφικτή η διατήρηση και ομάδων με μόλις δύο πρωτεΐνες, ώστε να υπάρχει ομοιογένεια με τα αποτελέσματα των υπολοίπων αλγορίθμων.

2.2.3. Markov Clustering Algorithm (MCL)

Ο αλγόριθμος MCL[96,97] χρησιμοποιείται γενικότερα σε προβλήματα ομαδοποίησης σε γράφους, ενώ έχει αποδειχτεί σε πλήθος εργασιών η χρησιμότητα του σε βιολογικά προβλήματα, όπως η εύρεση ομάδων πρωτεϊνών με βιολογική σημασία[95,98–101]. Παρουσιάστηκε αρχικά το 2002 ως μέρος της διπλωματικής εργασίας του Stijn van Dongen στο Πανεπιστήμιο της Ουτρέχτης, υπό την επίβλεψη των Jan van Eijck και Michiel Hazewinkel και βασίζεται σε στοχαστικά Μαρκοβιανά μαθηματικά μοντέλα. Η εύρεση των τελικών ομάδων προκύπτει με την πραγματοποίηση τυχαίων προσπελάσεων - περιπάτων μέσα στο δίκτυο και την εναλλαγή δύο μεταβλητών, επέκτασης (expansion) και εμφύησης (inflation), όπως περιγράφεται παρακάτω:

- Ως είσοδο ο αλγόριθμος δέχεται έναν μη κατευθυνόμενο γράφο, καθώς και τις παραμέτρους r και e, οι οποίες επηρεάζουν στην πορεία την εμφύσηση και την επέκταση αντίστοιχα.
- Υπολογίζεται ο Μαρκοβιανός πίνακας από τον υπολογισμό του πίνακα συσχέτισης του γράφου.
- Εκκινούνται κατόπιν οι τυχαίες προσπελάσεις, με την εναλλαγή των δυο διεργασιών:
    - Ο πίνακας συσχέτισης υψώνεται στην e-οστή δύναμη, όπως αυτή έχει οριστεί από τον χρήστη. Η πράξη αυτή αποτελεί την επέκταση (expand), η οποία επιτρέπει να έρθουν σε επαφή κόμβοι, που βρίσκονται σε μεγαλύτερη απόσταση.
    - Η διεργασία της εμφύησης (inflammation) πραγματοποιείται υψώνοντας κάθε στοιχείο του πίνακα στην μη αρνητική τιμή r και

κατόπιν κανονικοποιώντας πάλι αυτόν, ώστε το άθροισμα κάθε στήλης του να είναι 1.

- Οι παραπάνω πράξεις επαναλαμβάνονται έως ότου δεν παρατηρείται σημαντική αλλαγή των τιμών και έχει επιτευχθεί σύγκλιση του αλγορίθμου.
- Όταν ο αλγόριθμος φτάσει στο παραπάνω στάδιο, ο αρχικός γράφος έχει χωριστεί πλέον σε μικρότερους υπογράφους, οι οποίοι δεν ενώνονται μέσω μονοπατιών και αποτελούν τις τελικές ομάδες που προβλέπει ο αλγόριθμος.

2.2.4. Molecular Complex Detection (MCODE)

Η μέθοδος MCODE[102] αποτελεί μία από τις πρώτες υπολογιστικές προσεγγίσεις για την εύρεση συμπλόκων πρωτεϊνών από δίκτυα αλληλεπιδράσεων. Προτάθηκε το 2003 από τους Gary Bader και Christopher Hogue και στηρίζεται στην λογική βαθμολόγησης του κάθε κόμβου βάση των ακμών του και της πυκνότητας (density) των γειτόνων του. Αναλυτικότερα ο αλγόριθμος αποτελείται από 3 κύρια βήματα:

- Αρχικά γίνεται απόδοση βαρών στους κόμβους του δικτύου, σύμφωνα με την πυκνότητα της γειτονίας του, χρησιμοποιώντας την μετρική k-core, η οποία μετρά την πυκνότητα του κόμβου εκείνου (k) της άμεσης γειτονιάς με την υψηλότερη πυκνότητα (core- Clustering Coefficient).
- Στο δεύτερο βήμα, από το σύνολο των βαθμολογημένων κόμβων, επιλέγεται εκείνος με το υψηλότερο βάρος, ώστε να θεωρηθεί ως σημείο εκκίνησης (seed) ενός συμπλέγματος πρωτεϊνών. Ξεκινώντας από εκεί προστίθενται στο σύμπλεγμα κόμβοι, των οποίων το βάρος απέχει ένα συγκεκριμένο ποσοστό WPP από τον αρχικό. Κάθε κόμβος ελέγχεται μόνο μία φορά, καθώς δεν επιτρέπονται επικαλύψεις στο παρόν στάδιο, ενώ το βήμα αυτό ολοκληρώνεται όταν δεν υπάρχουν πλέον κόμβοι, που να μπορούν να προστεθούν στο σύμπλεγμα, πληρώντας τις προδιαγραφές. Γίνεται επανάληψη της διαδικασίας ξεκινώντας πάντα από τον κόμβο με το ψηλότερο βάρος έως ότου εξεταστούν όλοι.
- Στο τέλος του προηγούμενου σταδίου έχει δημιουργηθεί ένα πλήθος, μη επικαλυπτόμενων συμπλόκων, τα οποία μετα-επεξεργάζονται και αφαιρούνται εκείνα με λιγότερους των 2 κόμβων. Επιπλέον υπάρχει η επιλογή να επεκταθεί το δίκτυο κατά μία τιμή “fluff”, που ορίζει ο χρήστης, και επιτρέπει την δημιουργία επικαλυπτόμενων ομάδων. Οι τελικές προτεινόμενες ομάδες βαθμολογούνται βάση της πυκνότητας του υπογράφου, με το σύμπλεγμα με την υψηλότερη να ταξινομείται πρώτο.

2.2.5. NCMine

Το NCMine[25] αποτελεί παράλληλα μέθοδο ομαδοποίησης κόμβων με πιθανή λειτουργική σημασία, αλλά και εργαλείο απεικόνισης αυτών. Υλοποιήθηκε από τους Shu Tadaka και Kengo Kinoshita το 2016, ως επιπρόσθετο της πλατφόρμας Cytoscape. Η μέθοδος εξάγει υπογράφους από δίκτυα αλληλεπιδράσεων πρωτεϊνών, βαθμονομώντας τους κόμβους αυτών σύμφωνα με την κεντρικότητα (centrality) τους. Η μέθοδος, όπως αυτή προτάθηκε από τους συγγραφείς, αποτελείται από δύο κύριες φάσεις:

- Στην πρώτη φάση κάθε κόμβος λαμβάνει ένα βάρος, ανάλογα με τον βαθμό κεντρικότητας του, και ο κόμβος με το μεγαλύτερο βάρος επιλέγεται ως σημείο εκκίνησης (seed). Στον υπογράφο αυτού προστίθενται κόμβοι οι οποίοι συνδέονται άμεσα μαζί του, και παράλληλα τόσο η τιμή cliqueness του νέου γράφου που δημιουργείται, όσο και η cliqueness-change ανάμεσα στον παλιό και τον νέο γράφο, ξεπερνούν συγκεκριμένα κατώφλια. Στο τέλος αυτού του βήματος οι ομάδες που έχουν σχηματιστεί από κάθε κόμβο ταξινομούνται βάση αύξουσας σειράς των βαρών τους
- Ακολούθως γίνεται συγχώνευση των παραπάνω ομάδων σύμφωνα με την επικάλυψη τους και με τον κανόνα η Jaccard coefficient ανάμεσα σε δύο ομάδες να ξεπερνά ένα προκαθορισμένο όριο (merge threshold). Αξίζει να σημειωθεί πως η συγχώνευση των ομάδων πραγματοποιείται σύμφωνα με τη σειρά της δημιουργίας αυτών, αφού, όπως υποστηρίζουν οι συγγραφείς, αναμένεται ότι κόμβοι γειτονικών ομάδων τείνουν να έχουν παρόμοια αριθμό ακμών και κεντρικότητα (centrality).

2.2.6. Speed and Performance in Clustering (SPICi)

Ο αλγόριθμος SPICi[103] χρησιμοποιείται για την γρήγορη και με χαμηλό υπολογιστικό κόστος ομαδοποίηση βιολογικών δικτύων. Σχεδιάστηκε το 2010 από τους Peng Jiang and Mona Singh και όμοια με αλγορίθμους που περιγράφηκαν παραπάνω, δημιουργεί σταδιακά ομάδες, ξεκινώντας από κόμβους με υψηλά βάρη. Ο τρόπος που λειτουργεί το πρόγραμμα είναι:

- Δέχεται από τον χρήστη ως είσοδο δίκτυα στα οποία κάθε ακμή χαρακτηρίζεται από ένα συγκεκριμένο βάρος. Για κάθε κόμβο βρίσκει εν συνεχεία το άθροισμα των βαρών αυτού και επιλέγει να ξεκινήσει την διαδικασία από το κόμβο με το μεγαλύτερο συνολικό βάρος.
- Σε επόμενο βήμα οι γείτονες του αρχικού κόμβου κατατάσσονται σε 5 βαθμίδες (bins), ανάλογα με το άθροισμα των βαρών των ακμών τους και

επιλέγεται πάλι ο κόμβος με το μεγαλύτερο συνολικό βάρος ακμών. Το ζεύγος των δυο αυτών αρχικών κόμβων αποτελεί την αρχή μιας ομάδας.

- Για την επέκταση των ομάδων μελετάται κάθε φορά το κατά πόσο η προσθήκη ενός επόμενου γειτονικού κόμβου, που δεν ανήκει σε κάποια άλλη ομάδα, κατεβάζει την μεταβλητή υποστήριξης S (minimum support threshold) κάτω από ένα προκαθορισμένο κατώφλι. Αν η τιμή είναι κάτω του κατωφλίου ο κόμβος αυτός αφαιρείται από την ομάδα και απορρίπτεται από την λίστα εξέτασης. Σε αντίθετη περίπτωση προστίθεται και αλλάζει η τιμή της πυκνότητας της ομάδας. Αν η νέα τιμή πυκνότητας είναι μικρότερη ενός συγκεκριμένου κατωφλίου (minimum cluster density), ο κόμβος απορρίπτεται της ομάδας. Η διαδικασία αυτή επαναλαμβάνεται μέχρι όλοι οι κόμβοι να ανήκουν σε κάποια ομάδα.

2.3 Μετρικές Σύγκρισης Αποτελεσμάτων Αλγορίθμων

Στην πάροδο των χρόνων έχει προταθεί πλήθος διαφορετικών μετρικών τόσο για την σύγκριση αποτελεσμάτων διαφορετικών αλγορίθμων ομαδοποίησης δικτύων πρωτεϊνών - πρωτεϊνών, όσο και για την αξιολόγηση των αποτελεσμάτων με επικυρωμένα σύμπλοκα πρωτεϊνών[104–106]. Στην παρούσα εργασία επιλέχθηκαν να υπολογιστούν οχτώ διαφορετικά στατιστικά μέτρα, τα οποία θεωρήθηκε ότι μπορούν να αποδώσουν μία συγκεντρωτική εικόνα για την σχέση, αλλά και την ποιότητα των αποτελεσμάτων των αλγορίθμων που εκτελέστηκαν. Τα μέτρα αυτά μπορούν να διακριθούν, αρχικά σε δύο κατηγορίες[17], όπου η πρώτη περιλαμβάνει εκείνα που εξάγονται βάση της κατανομής της πληροφορίας στα σύνολα αποτελεσμάτων, ενώ η δεύτερη σχετίζεται με τις επικαλύψεις των συμπλόκων πρωτεϊνών μεταξύ των αποτελεσμάτων. Ένας επιπλέον διαχωρισμός των μετρικών μπορεί να γίνει ως προς την συμμετρικότητα αυτών. Σε αντίθεση με τα πρώτα μέτρα που θα περιγραφθούν (Normalized Mutual Information, Variation of Information, Adjusted Rand Index), τα περισσότερα είναι μη συμμετρικά, κυρίως διότι εξ ορισμού θεωρούν το ένα εκ των δύο ομαδοποιήσεων ως υπόδειγμα. Επομένως στις περιπτώσεις, όπως εδώ, που χρησιμοποιούνται για την σύγκριση διαφορετικών αλγορίθμων ομαδοποίησης είναι ορθότερο να εκτελούνται δύο φορές, με εναλλαγή κάθε φορά του πρώτου συνόλου εισόδου. Παρακάτω θα γίνει εκτενέστερη περιγραφή των μεθόδων που χρησιμοποιήθηκαν.

2.3.1. Normalized Mutual Information (NMI)

Η μετρική Mutual Information[107] περιγράφει το κατά πόσο δύο σύνολα ομάδων πρωτεϊνών είναι όμοια μεταξύ τους, και κατ' επέκταση στην περίπτωση που το ένα από αυτά αποτελείται από πειραματικά αποδεδειγμένα πρωτεϊνικά σύμπλοκα, μετράει την ορθότητα του άλλου. Το εύρος τιμών της είναι μεταξύ 0 και 1, όπου με 0 χαρακτηρίζονται τα πλήρως ανόμοια και με 1 τα ταυτόσημα σύνολα. Στην παρούσα εργασία επιλέχθηκε να χρησιμοποιηθεί η εκδοχή της μετρικής των Aaron F. McDaid, Derek Greene και Neil Hurley[108], καλούμενη Normalized Mutual Information (NMI), η οποία δίνει την δυνατότητα ελέγχου και επικαλυπτόμενων ομάδων πρωτεϊνών, ενώ, όπως απέδειξαν στην δημοσίευση τους οι συγγραφείς, η συνάρτηση κανονικοποίησης που επέλεξαν υπερτερεί προγενέστερων.

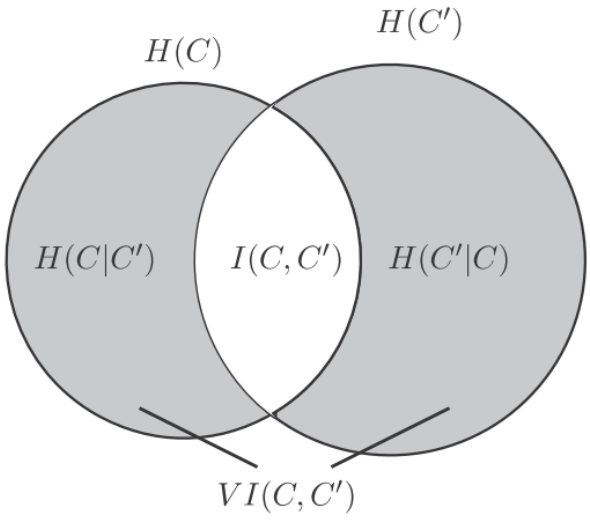
Η βασική συνάρτηση υπολογισμού της Mutual Information για δυο ομαδοποιήσεις C ΚΑΙ C' είναι: $I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P(k')}$ [1], με $P(k,k')$ την κοινή κατανομή των τυχαίων μεταβλητών (k και k') στα δύο σύνολα.

Ακολούθως η κανονικοποίηση της παραπάνω γίνεται μέσω της συνάρτησης: *NMI* = $\frac{I(C,C')}{max(H(C),H(C'))}$ [2] , όπου H(C), (H(C') η εντροπία των τυχαίων μεταβλητών στο κάθε σύνολο.

2.3.2. Variation of Information (VI)

Στην ίδια εργασία, όπου πρωτοπαρουσιάστηκε η μετρική Mutual Information, η συγγραφέας προτείνει και ένα ακόμη μέτρο σύγκρισης της ομοιότητας δύο ομαδοποιήσεων, το Variation of Information[107]. Γνωρίζοντας, λοιπόν, την εντροπία H(C) και H(C') των συνόλων C και C', ως VI ορίζεται: *VI (C, C') = H (C) + H (C') − 2I (C, C')* [3] με I(C,C') βάση της συνάρτησης [1].

Μία διαφορετική εκδοχή της ίδια μετρικής, χωρίς την άμεση συμμετοχής της I(C,C') δίνεται και από την εξίσωση *VI (C, C') = H (C|C') + H (C' |C)* [4]. Απεικονιστικά η σχέση αυτή αποδίδεται στο διάγραμμα Venn [Εικόνα 3] που ακολουθεί, όπου, όπως είναι εμφανές, η Mutual Information περιγράφει την πληροφορία που είναι κοινή και στα δύο σύνολα, ενώ η VI την πληροφορία που διαφέρει. Ο συνδυασμός, επομένως, και των δύο μετρικών στην αξιολόγηση των αποτελεσμάτων δύο αλγορίθμων ομαδοποίησης αποδίδει μία πλήρη εικόνα για την πληροφορία που εμπεριέχεται σε αυτά. Οι τιμές που αναμένονται από τις δύο μετρικές κατά την σύγκριση αλγορίθμων είναι σχετικά αντιστρόφως ανάλογες αφού ένα καλό αποτέλεσμα σύγκρισης αποδίδει υψηλό Mutual Information και χαμηλό Variation of Information.



Εικόνα 3: Σχηματική απεικόνιση της κατανομής της πληροφορίας σε δύο σύνολα C και C', όπου οι περιοχές με σκούρο χρώμα αποτελούν την Variation of Information, ενώ η λευκή περιοχή την Mutual Information (Meilă M., 2007)

2.3.3. Adjusted Rand Index (ARI)

Μία από τις πρώτες στατιστικές μεθόδους σύγκρισης και αξιολόγησης ομαδοποιήσεων δεδομένων παρουσιάστηκε το 1971 από τον William M. Rand και ονομάστηκε εξ αυτόν Rand Index[109]. Η μέθοδος υπολογίζει ανά ζεύγη το ποσοστό των συμπλόκων που έχουν ομαδοποιηθεί με τον ίδιο τρόπο και με τους δύο αλγορίθμους ομαδοποίησης ως προς το σύνολο όλων των πιθανών. Βάση ορισμού, επομένως, η τιμή αυτής κυμαίνεται στο διάστημα [0,1], με 1 να λαμβάνει όταν έχουν ομαδοποιηθεί με τον ίδιο ακριβώς τρόπο, ενώ 0 όταν δεν υπάρχει κανένα κοινό ζεύγος ομάδων.

Η αρχική συνάρτηση εύρεσης του Rand Index είναι: $R = \frac{a+b}{\binom{n}{2}}$ [5], όπου θεωρώντας S το σύνολο όλων των ζευγών στοιχείων και C,C' δύο ομαδοποιήσεις αυτού, με a συμβολίζονται οι ομάδες του S που έχουν ομαδοποιηθεί στο ίδιο υποσύνολο τόσο στο C όσο και στο C', b όσες έχουν ανατεθεί σε διαφορετικές ομάδες των δύο υποσυνόλων και $\binom{n}{2} = n*\left(\frac{n-1}{1}\right)$.

Το 1985 οι Lawrence Hubert και Phipps Arabie[110] πρότειναν μία τροποποιημένη εκδοχή της παραπάνω συνάρτησης, την Adjusted Rand Index, που επιτρέπει την μελέτη μεγαλύτερης ποικιλίας συνόλων, ενώ πλέον το αποτέλεσμα μπορεί να λάβει και αρνητική τιμή.

Ο υπολογισμός του Adjusted Rand Index γίνεται μέσω της συνάρτησης που ακολουθεί, εκφράζοντας πρώτα το σύνολο a+b ως $\sum_{ij}\binom{n_{ij}}{2}$: $ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b}{2}\right] / \binom{n}{2}}$ [6]

2.3.4. F1- score

Το F1-score[111], το οποίο συχνά συναντάται και ως F-score ή F-measure, αποτελεί από τα κλασσικά στατιστικά μέτρα σύγκρισης αποτελεσμάτων αλγορίθμων και στηρίζεται κατά μεγάλο βαθμό, όπως και τα μέτρα που ακολουθούν, στον πίνακα σύγχυσης (confusion matrix). Ο πίνακας αυτός διαχωρίζει τα αποτελέσματα σε 4 ομάδες, θεωρώντας κάθε φορά ότι ένα εκ των δυο αποτελεσμάτων είναι αναγνωρισμένο ως το σωστό. Βάση αυτού, λοιπόν, δημιουργείται ο ακόλουθος πίνακας [Πίνακας 3]:

		Predicted Complexes	
		Positive	**Negative**
Actual Complexes	**Positive**	**True Positive (TP)**	**False Negative (FN)**
	Negative	**False Positive (FP)**	**True Negative (TN)**

Πίνακας 3: Παράδειγμα πίνακα σύγχυσης

Έχοντας ως πρότυπο τον Πίνακα 3, ο τύπος του F1- score είναι: *F1* = $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$ [7]. Επιπλέον πολύ συχνά συναντάται στην βιβλιογραφία και η χρήση του τύπου *F1 = 2** $\frac{Precicion*Recall}{Precicion+Recall}$ [8], ο οποίος είναι όμοιος με τον [7], αφού ισχύει πως *Precision* = $\frac{TP}{TP+FP}$ και *Recall* = $\frac{TP}{TP+FN}$. Επομένως, βάση της εξίσωσης [8], το F1-score χαρακτηρίζεται ως ο σταθμισμένος μέσος των Precision και Recall, υπερνικώντας τις αδυναμίες αυτών.

2.3.5. Cluster-wise Sensitivity

Γενικότερα το μέτρο της Sensitivity, κατά την μελέτη της ορθότητας ενός συστήματος δείχνει το ποσοστό των αποτελεσμάτων που είναι όντως αληθή (TP), σε σύγκριση με το σύνολο των αληθών που δίνει το σύστημα ως αληθή, δηλαδή σύμφωνα με τον πίνακα σύγχυσης *SN* = $\frac{TP}{TP+FN}$, όπως αποδόθηκε παραπάνω και ο τύπος για το Recall. Κατά τον υπολογισμό της Sensitivity στην σύγκριση συμπλόκων πρωτεϊνών και κατ' επέκταση ολόκληρων συνόλων αποτελεσμάτων αλγορίθμων ομαδοποίησης, η έννοια της Sensitivity εκφράζει το ποσοστό των ομάδων του αληθές συνόλου, το οποίο έχει αντιστοιχηθεί σωστά στα αποτελέσματα του άλλου αλγορίθμου. Αριθμητικά αυτό αποδίδεται: *Sn*= $\frac{\sum_{i=1}^{n} \max_{j}\{T_{ij}\}}{\sum_{i=1}^{n} N_i}$ [9], όπου N_i οι ομάδες που ανήκουν στο θεωρούμενο αληθές σύνολο και T_{ij} οι κοινές ομάδες ανάμεσα στα δύο σύνολα αποτελεσμάτων.

2.3.6. Positive Predictive Value (PPV)

Μία μετρική παρόμοια με την προηγούμενη είναι η Positive Predictive Value, μέσω της οποίας μπορεί να βρεθεί το ποσοστό εκείνο των ομάδων που βρέθηκαν αληθής σε σχέση με τον πραγματικό αριθμό των ομάδων που υπάρχουν στο σύνολο

αναφοράς. Μέσω των τιμών του πίνακα σύγχυσης η μετρική υπολογίζεται *PPV* = $\frac{TP}{TP+FP}$, όμοια δηλαδή με τον τύπο της Precision, που αναφέρθηκε παραπάνω. Αντίστοιχα κατά την σύγκριση αποτελεσμάτων ομαδοποίησης ο τύπος που χρησιμοποιείται είναι *PPV* = $\frac{\sum_{j=1}^{m} \max_i\{T_{ij}\}}{\sum_{j=1}^{m} T_j}$ [10], με T_j να υποδηλώνει τον συνολικό αριθμό κόμβων μέσα σε κάθε ομάδα που πρόβλεψε ο αλγόριθμος πως ανήκουν στο σύνολο αναφοράς.

2.3.7. Geometric Accuracy

Η Geometric Accuracy ως μέτρο αποτελεί μία ισοστάθμιση των Sensitivity και Positive Predictive Value, αφού υπολογίζεται ως ο γεωμετρικός μέσος αυτών. Συγκεκριμένα, θεωρώντας γνωστές τις εξισώσεις [9] και [10]: *Geometric Accuracy* = $\sqrt{(Sensitivity \times PPV)}$ [11]. Η επιλογή του γεωμετρικού μέσου όρου, έναντι του αριθμητικού που χρησιμοποιείται συχνά, εξασφαλίζει πως για να ληφθεί υψηλή τιμή θα πρέπει τόσο η Sensitivity, όσο και η Positive Predictive Value να έχουν υψηλές επιδόσεις.

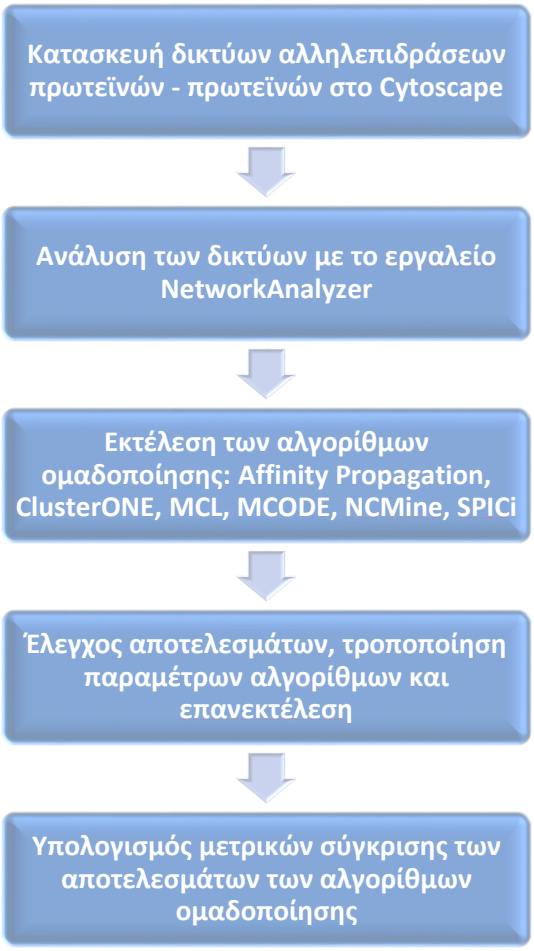
2.3.8. Maximum Matching Ratio (MMR)

Τέλος περιγράφεται η μετρική Maximum Matching Ratio[95], η οποία αντιπροσωπεύει το κλάσμα των συμπλόκων αναφοράς, τα οποία αντιστοιχίζονται από τουλάχιστον ένα προβλεπόμενο σύμπλοκο. Η τιμή αυτής δίνεται από το συνολικό βάρος της μέγιστης αντιστοίχισης, διαιρούμενο με τον αριθμό των συμπλόκων αναφοράς. Η συνάρτηση που περιγράφει την παραπάνω σχέση είναι: *MMR* = $\frac{N_r}{|R|}$ [12]. Ο αριθμητής αυτής, N_r, βρίσκεται από την σχέση $N_r = |\{c | c \in R, \exists p \in P, OS(p, r) \geq \omega\}|$, με R το σύνολο των συμπλόκων του συνόλου αναφοράς, P τα σύμπλοκα που πρόβλεψε ο δεύτερος αλγόριθμος και OS(p,r) μία βαθμολογία ταιριάσματος βάση του τύπου $OS(P,R)=|P \cap R|2|P|\times|R|$.

Για να φτάσει η μέθοδος στο στάδιο υπολογισμού του κλάσματος απαιτείται ο βαθμός ταιριάσματος να είναι μεγαλύτερος ενός προκαθορισμένου κατωφλίου ω, όπου στη παρούσα εργασία ορίστηκε ω=0.20.

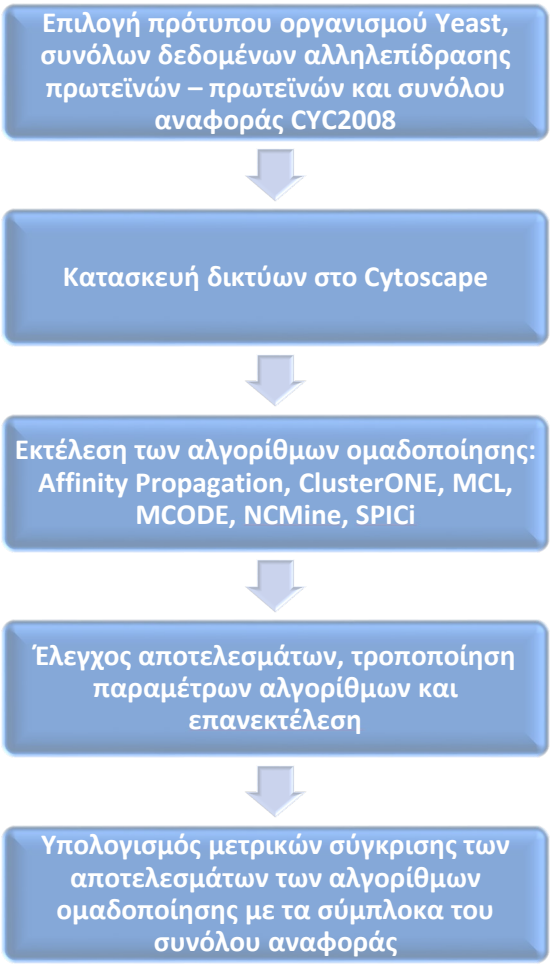
2.4 Βήματα

Έχοντας πραγματοποιηθεί η επιλογή των δεδομένων και των μεθόδων εκτέλεσης και αξιολόγησης, η σύγκριση των αλγορίθμων ομαδοποίησης πραγματοποιήθηκε με δύο τρόπους. Κατά τον πρώτο έγινε σύγκριση των αλγορίθμων μεταξύ τους, για διάφορες παραμέτρους αυτών, ώστε να βρεθεί, όπου υπήρχε, συσχέτιση αυτών. Τα βήματα που ακολουθήθηκαν κατά την διενέργεια της συγκεκριμένης μεθόδου παρουσιάζονται συνοπτικά στο παρακάτω σχήμα [Διάγραμμα 1]:



Διάγραμμα 1: Ροή διεργασιών σύγκρισης αλγορίθμων ομαδοποίησης μεταξύ τους

Ακολούθως, για την καλύτερη αξιολόγηση των αποτελεσμάτων, έγινε σύγκριση δεδομένων από πρότυπο οργανισμό με δεδομένα από σύνολο αναφοράς, ώστε να αποδειχθεί η αξιοπιστία των αλγορίθμων. Τα στάδια για την πραγματοποίηση του ελέγχου περιγράφονται στο διάγραμμα ροής [Διάγραμμα 2] που ακολουθεί:



Διάγραμμα 2: Ροή διεργασιών σύγκρισης αλγορίθμων ομαδοποίησης με δεδομένα συνόλου αναφοράς

3. Αποτελέσματα

Στην παρούσα ενότητα θα παρουσιαστούν τα αποτελέσματα κάθε σταδίου των ροών διεργασιών, που αναφέρθηκαν στο προηγούμενο κεφάλαιο. Αποτελείται από δύο κύρια υποκεφάλαια, κάθε ένα εκ των οποίων επικεντρώνεται στα βήματα μίας μεθόδου.

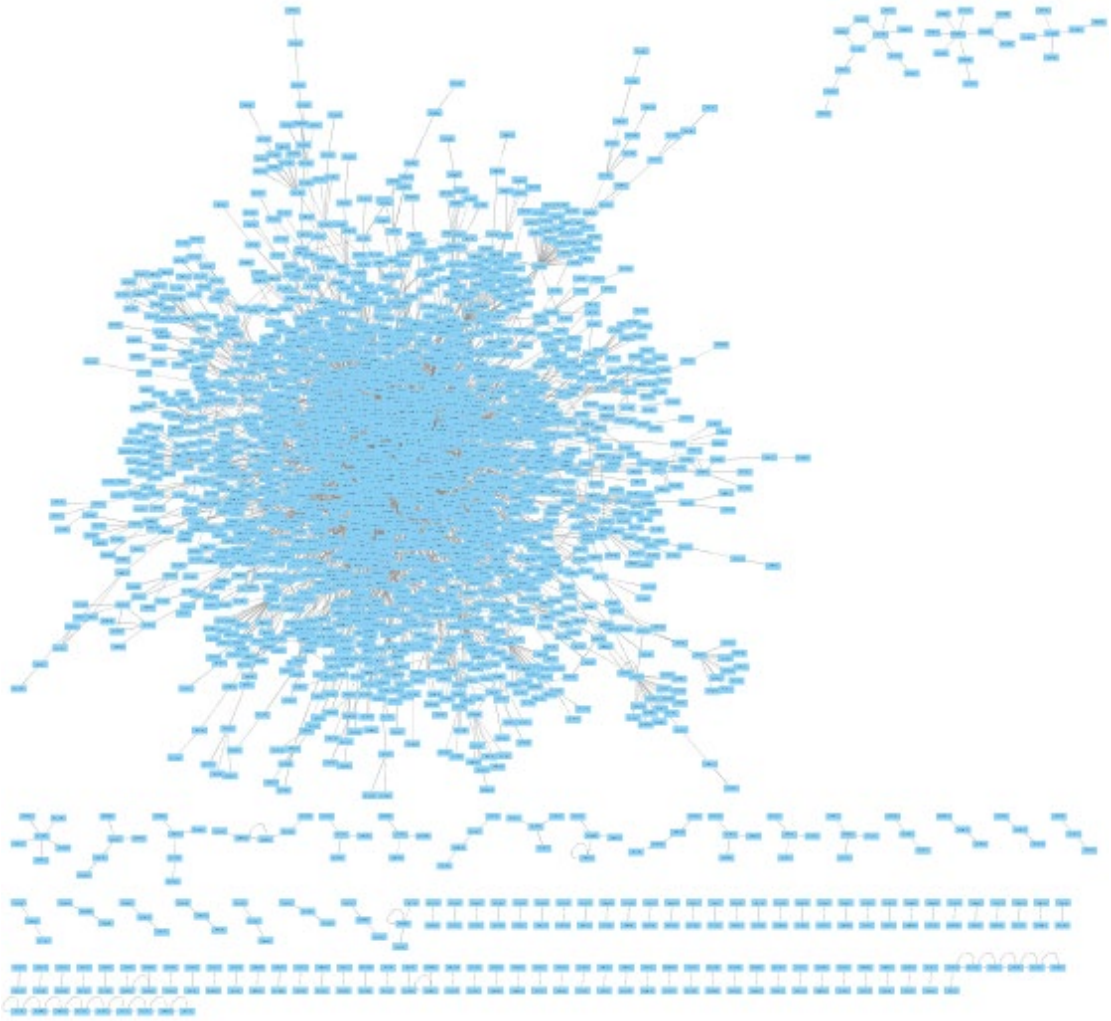
3.1 Σύγκριση αλγορίθμων ομαδοποίησης μεταξύ τους

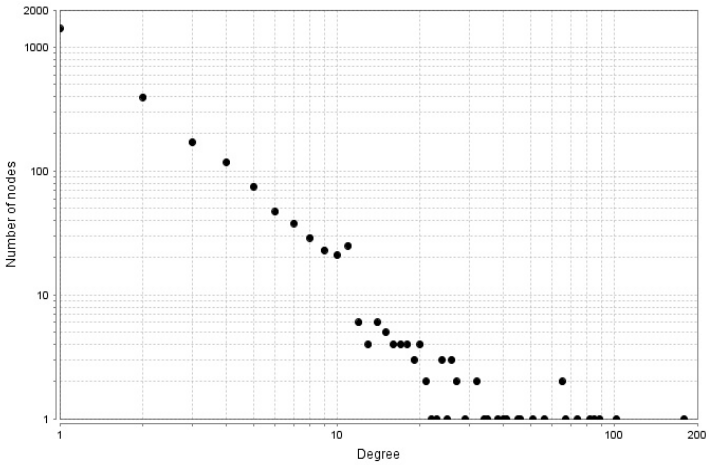
3.1.1. Κατασκευή και ανάλυση δικτύων αλληλεπιδράσεων πρωτεϊνών – πρωτεϊνών στο Cytoscape

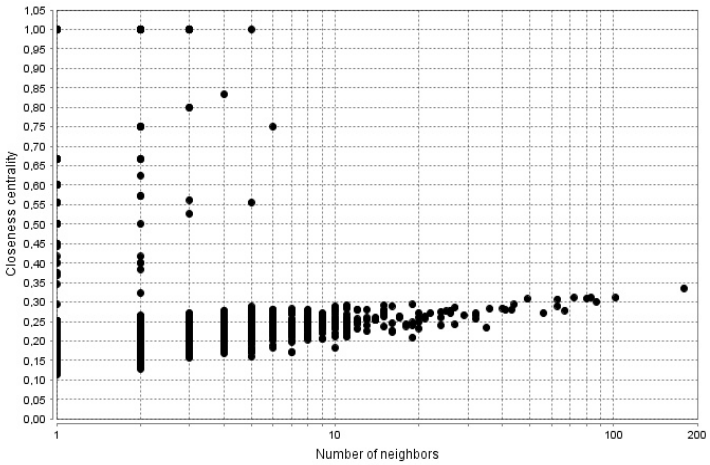
Καθώς τα σύνολα δεδομένων που επιλέχθηκαν να χρησιμοποιηθούν αποτελούνται από χιλιάδες κόμβους και ακμές, θεωρήθηκε απαραίτητη η απεικόνιση αυτών σε δίκτυα, με τη βοήθεια του προγράμματος Cytoscape. Το σχέδιο διάταξης των δικτύων, μετά από δοκιμές διαφόρων επιλογών, επιλέχθηκε το Prefuse Force-Directed Layout, το οποίο στηρίζεται στο αλγόριθμο του πακέτου εργαλείων του Jeff Heer[112]. Η συγκεκριμένη μορφοποίηση φάνηκε πως, ιδιαίτερα στα πιο μεγάλα και πολύπλοκα δίκτυα, αντικατοπτρίζει καλύτερα τη δομή τους.

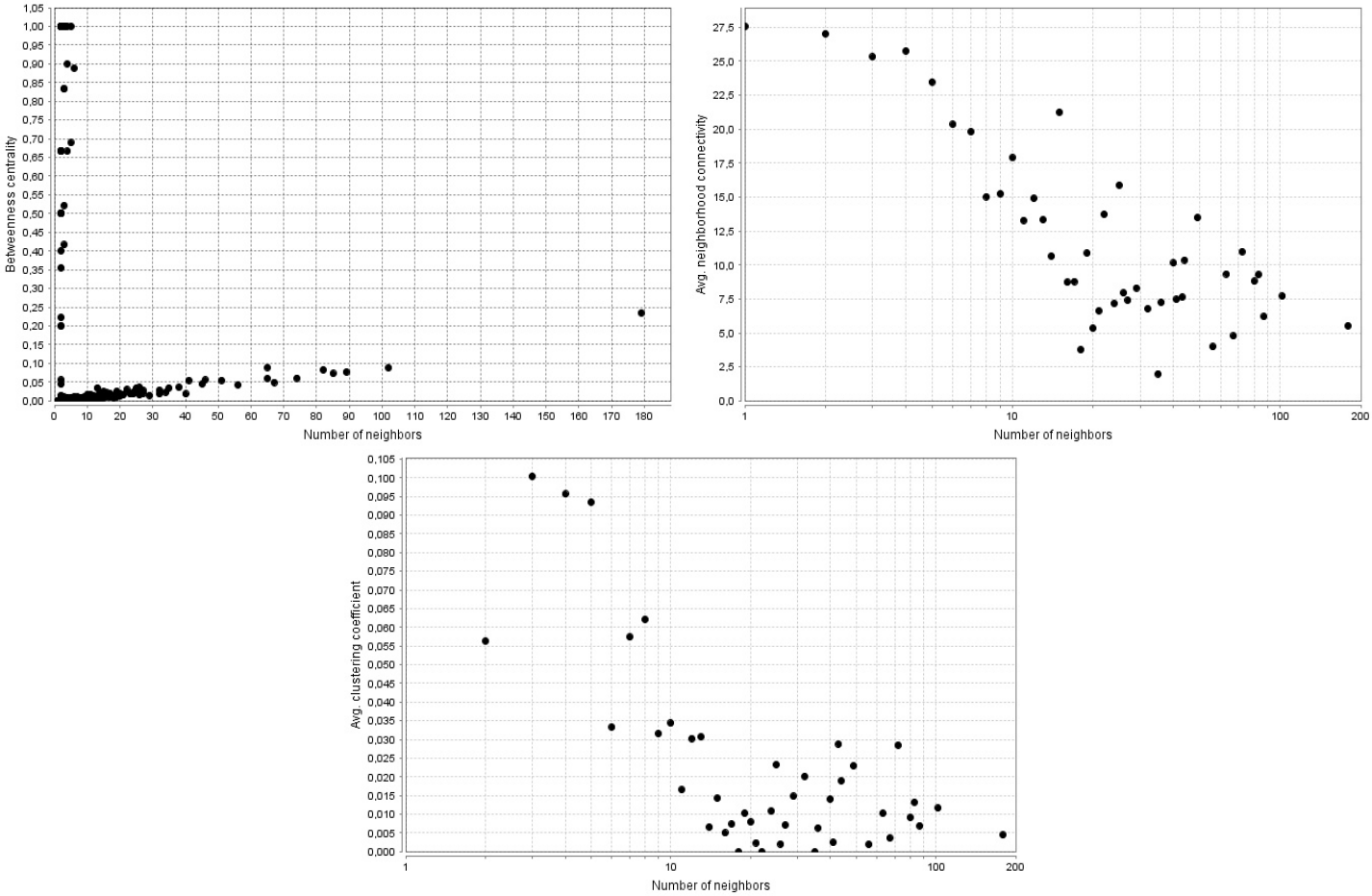
Μετά την κατασκευή των απεικονιστικών δικτύων, έγινε ανάλυση καθενός εξ αυτών με το εργαλείο NetworkAnalyzer του Cytoscape. Μέσω του NetworkAnalyzer βρέθηκαν τα βασικά περιγραφικά χαρακτηριστικά κάθε δικτύου, όπως αυτά επεξηγήθηκαν στο κεφάλαιο 1.1, ώστε να γίνουν αντιληπτά σε μεγαλύτερο βάθος οι ιδιότητες τους. Η γνώση αυτών των πληροφοριών θεωρήθηκε, επιπλέον, ότι θα οδηγήσει σε συμπεράσματα ως προς τον τρόπο που επηρεάζει η δομή του δικτύου τον κάθε αλγόριθμο ομαδοποίησης. Τα μέτρα που βρέθηκαν για κάθε δίκτυο είναι: Clustering Coefficient, Diameter, Radius, Centralization, Shortest Paths, Characteristic Path Length, Average Degree, Density, Heterogeneity, Self- Loops.

Παρακάτω παρουσιάζονται τα αποτελέσματα της απεικόνισης και ανάλυση των δικτύων αλληλεπιδράσεων, τόσο μέσω της εικόνα τους δικτύου, όσο και με γραφήματα και συγκεντρωτικούς πίνακες. Στις εικόνες των δικτύων με γαλάζια τετράγωνα συμβολίζονται οι κόμβοι - πρωτεΐνες, ενώ οι συνδέσεις του δικτύου είναι οι γκρι ευθείες που ενώνουν τους κόμβους.

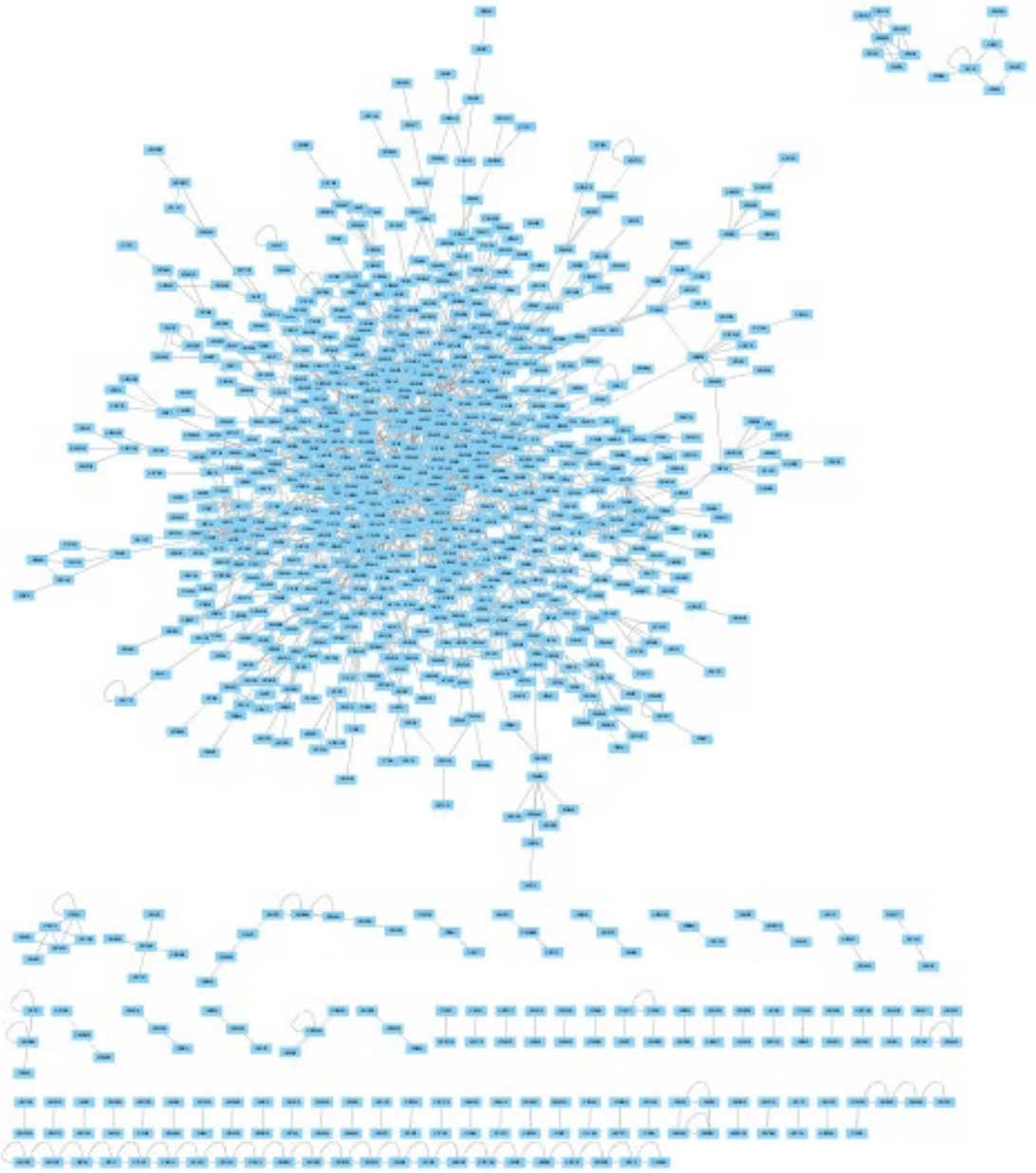


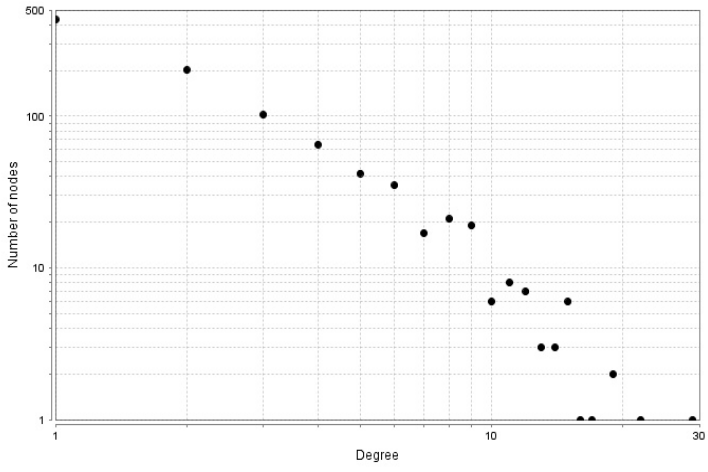

Number of nodes
2000
1000
100
10
1
1
10
100
200
Degree

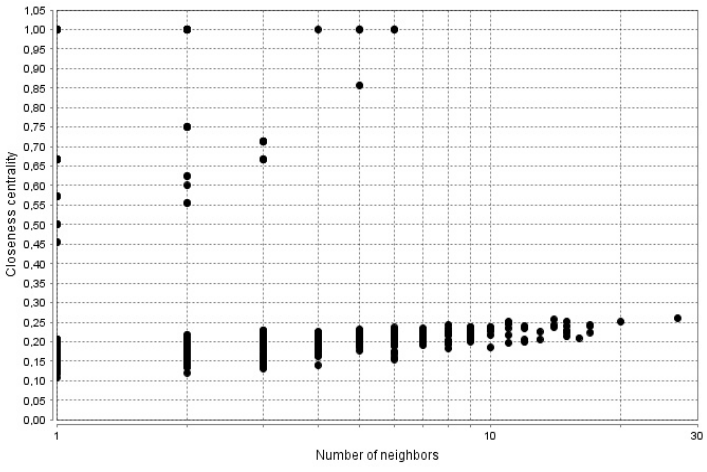

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
200
Number of neighbors

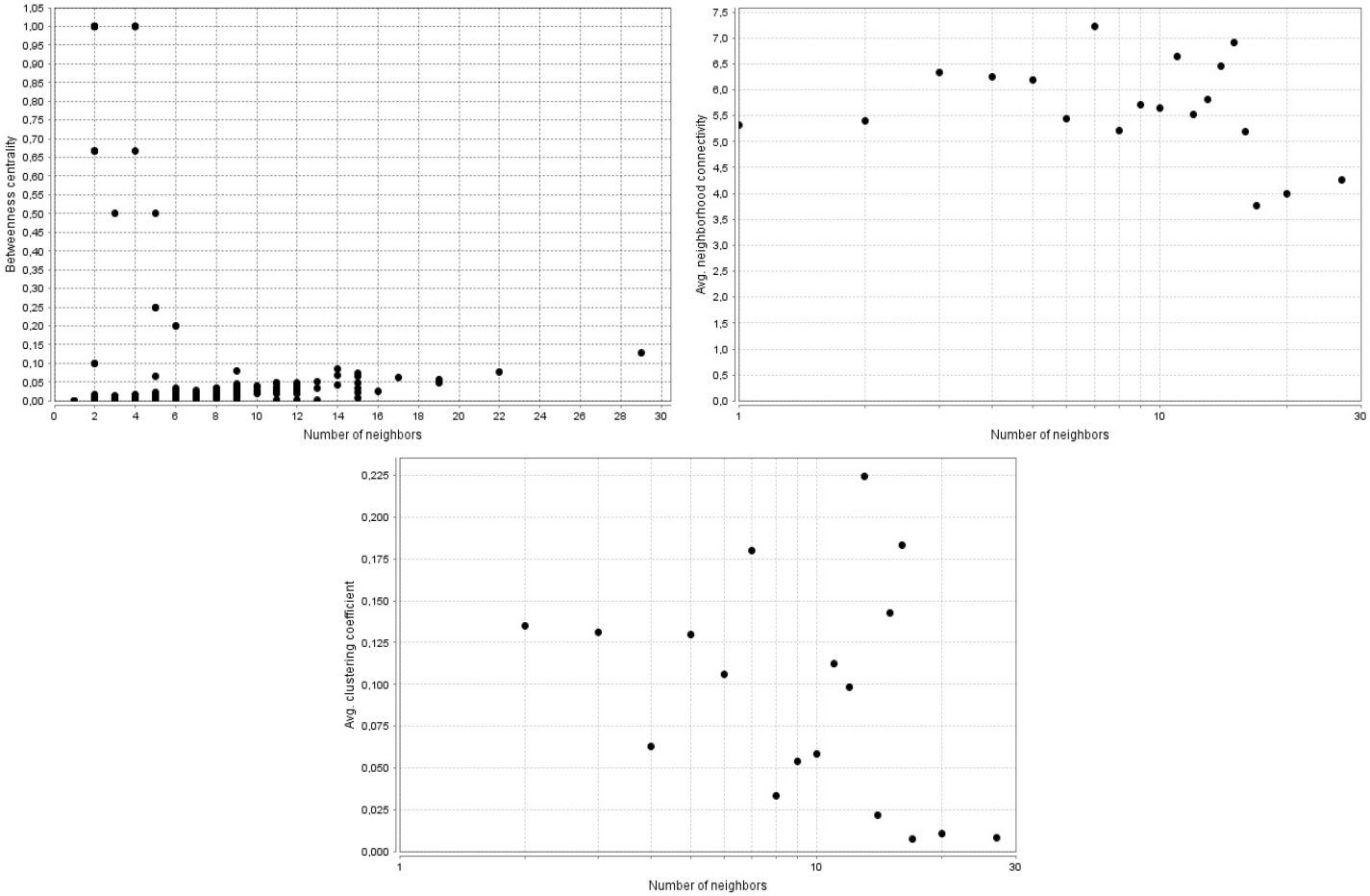


Εικόνα 4: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου C. Elegans από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

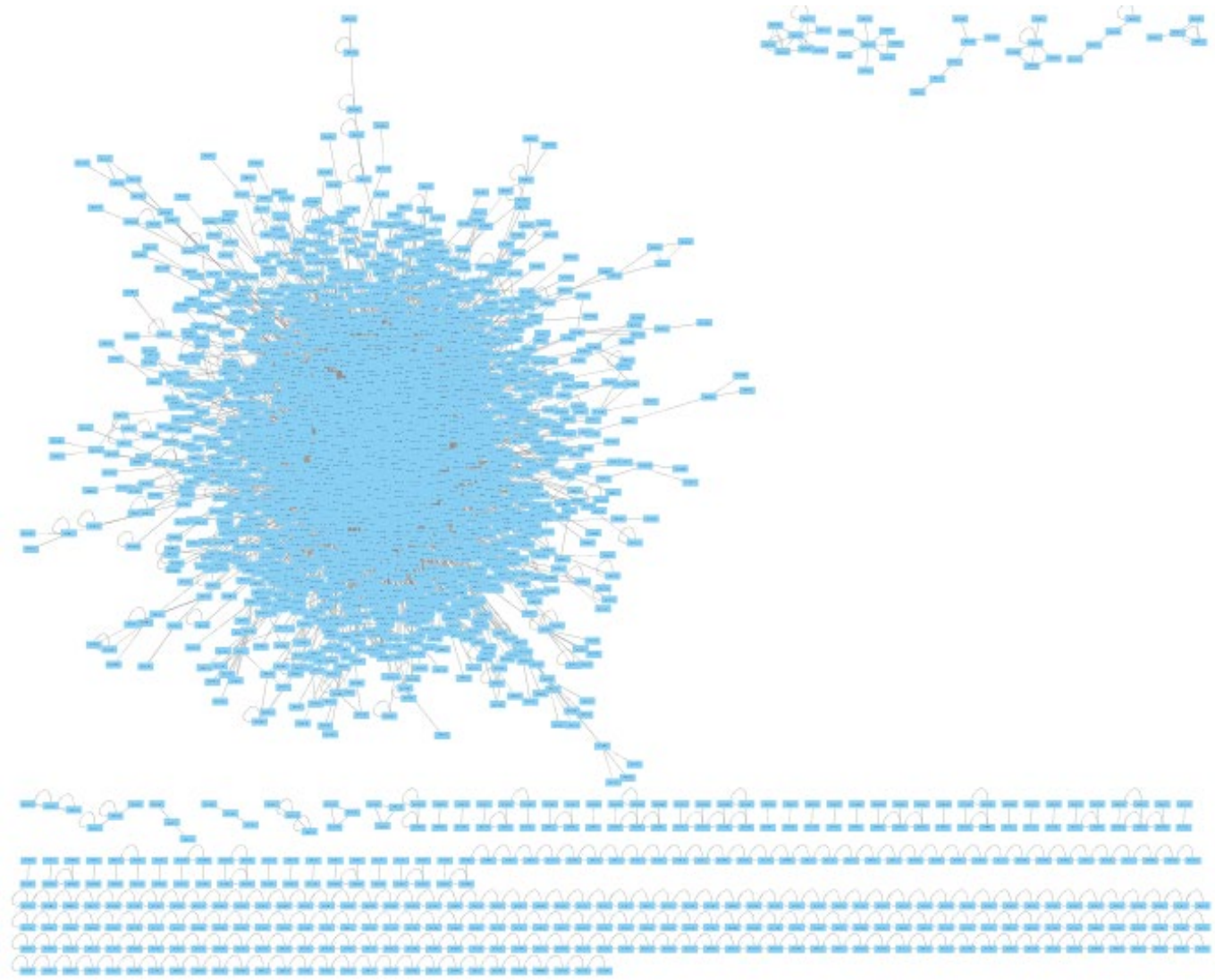


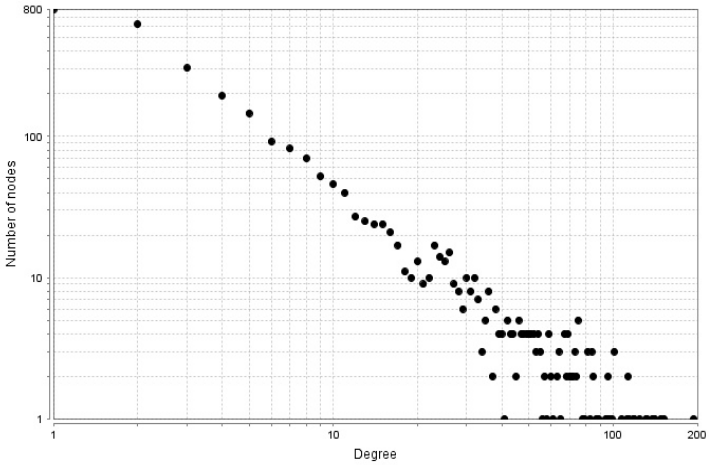

Number of nodes
500
100
10
1
1
10
30
Degree

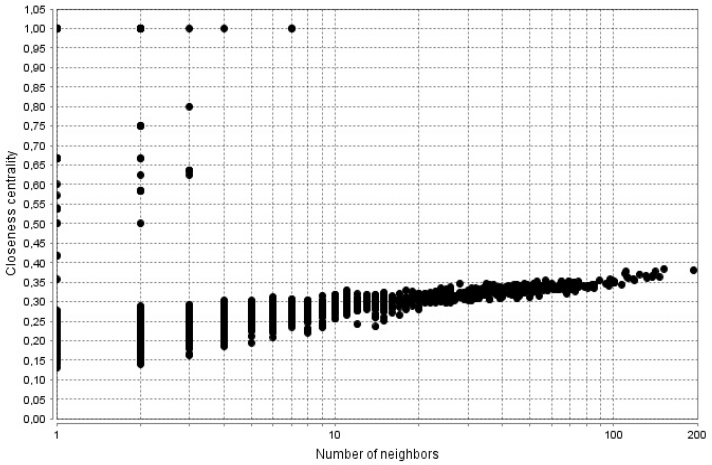

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
30
Number of neighbors

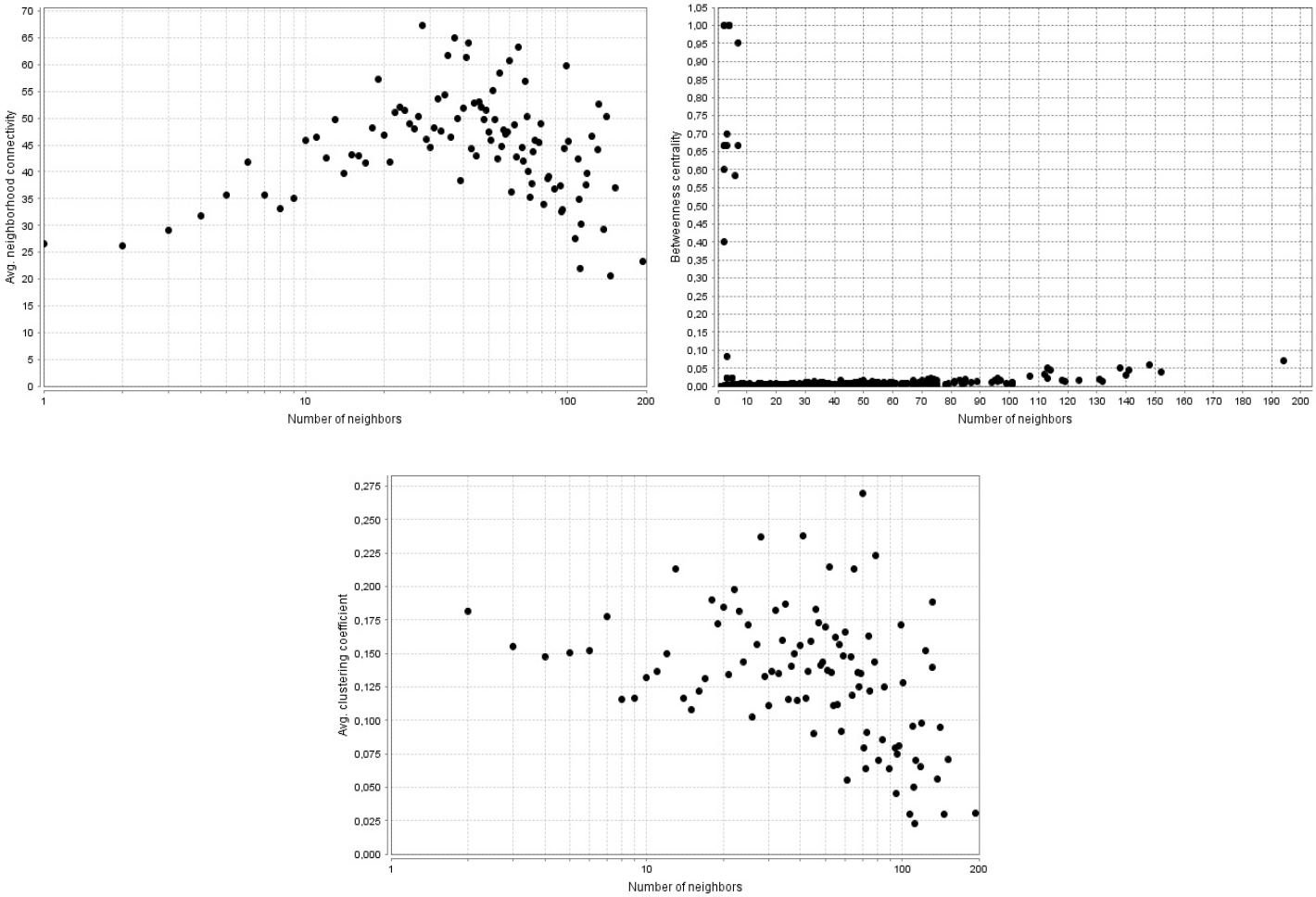


Εικόνα 5: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Drosophila από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

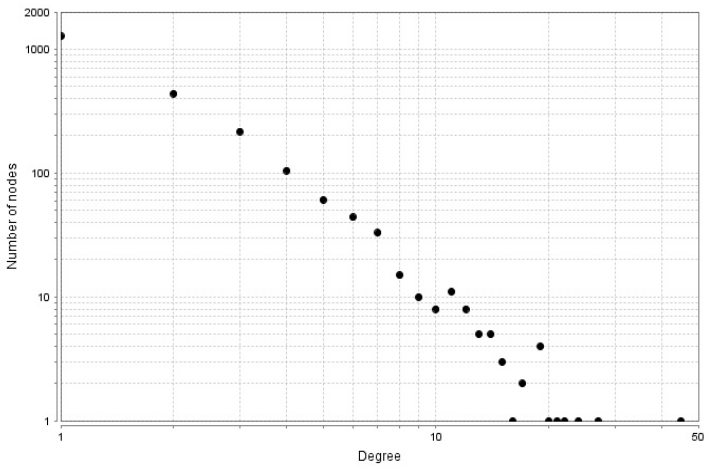


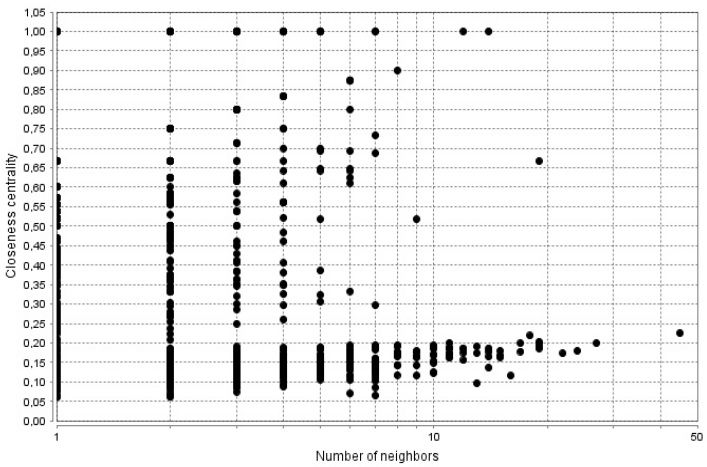

800
100
10
1
Number of nodes
1
10
100
200
Degree

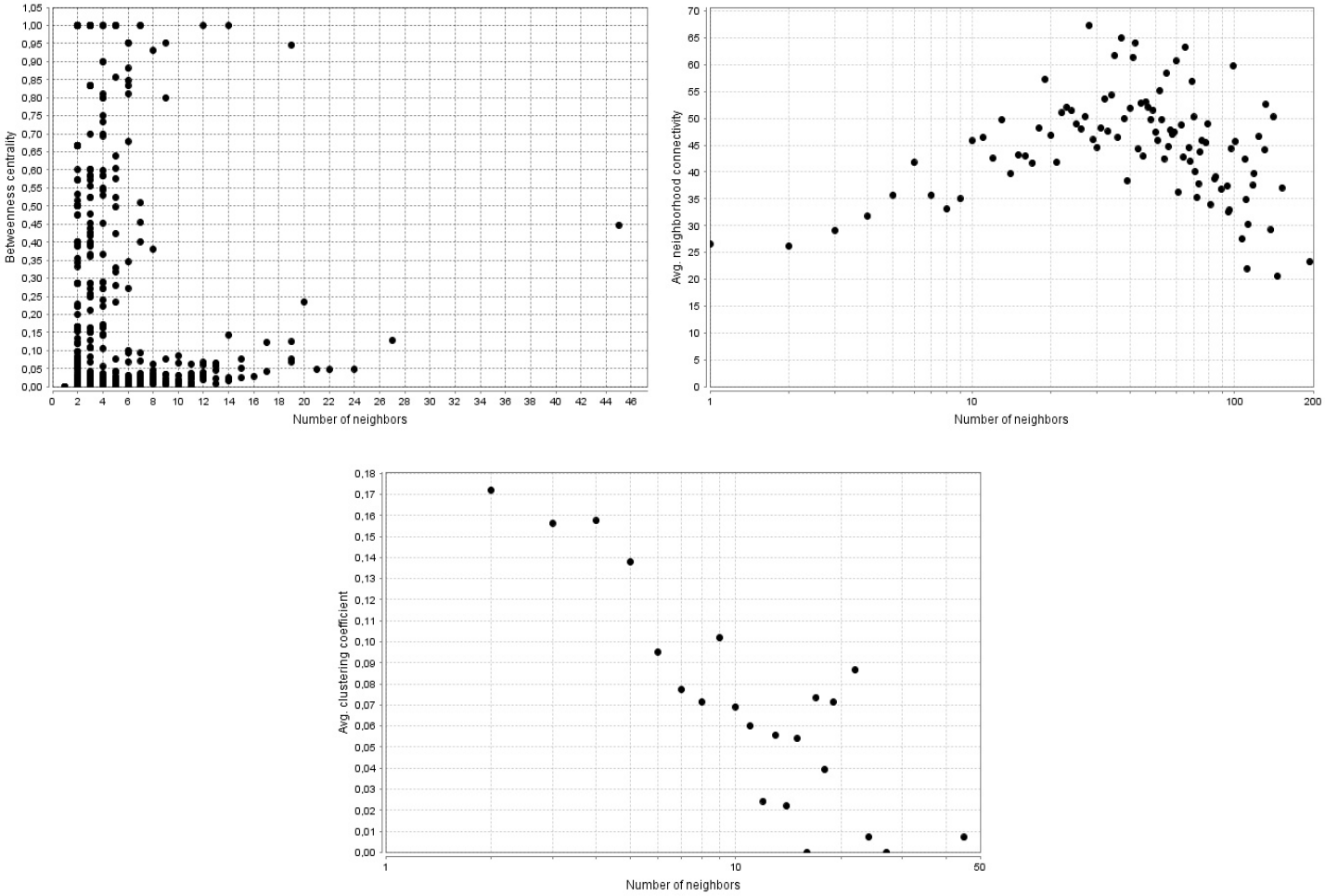

1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Closeness centrality
1
10
100
200
Number of neighbors



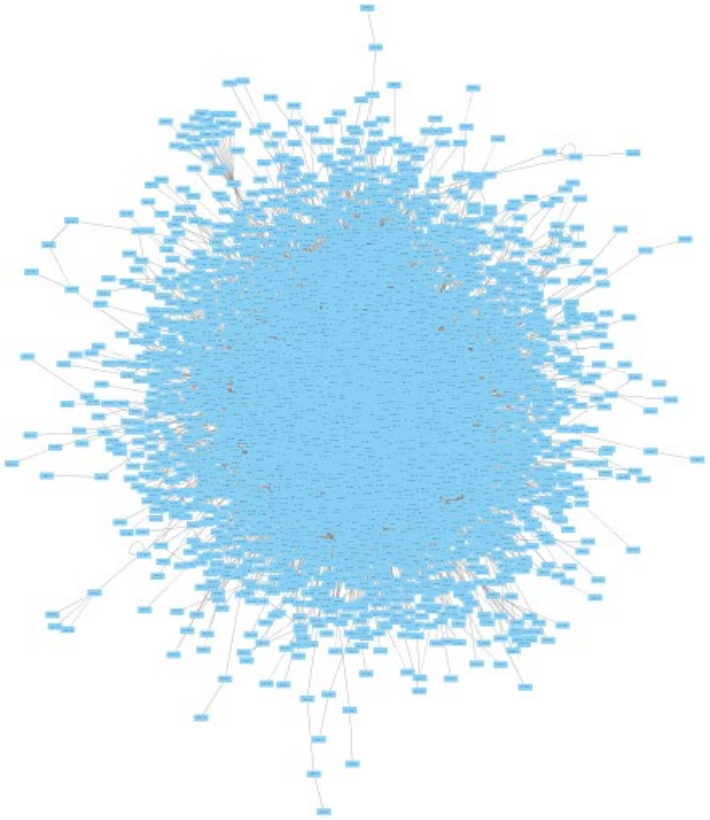
Εικόνα 6: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου E. coli από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

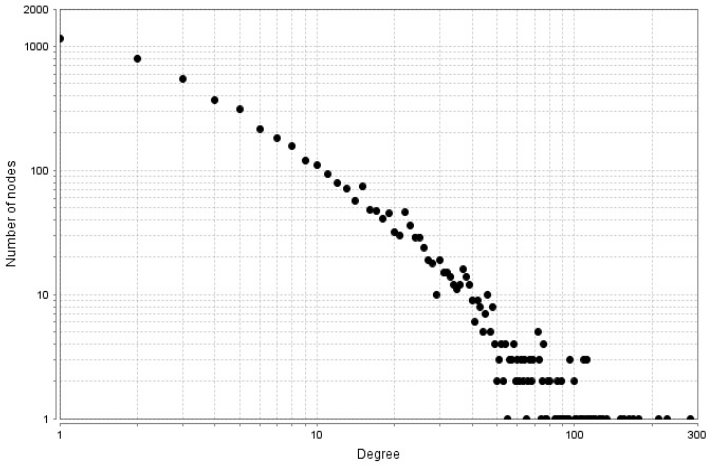

Number of nodes
2000
1000
100
10
1
1
10
50
Degree

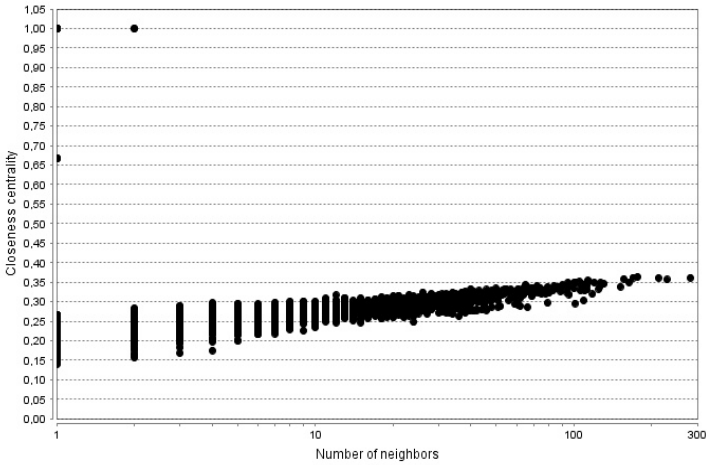

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
50
Number of neighbors

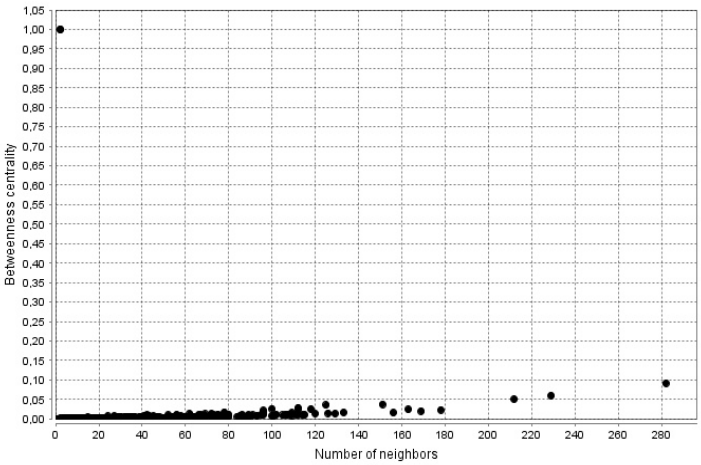


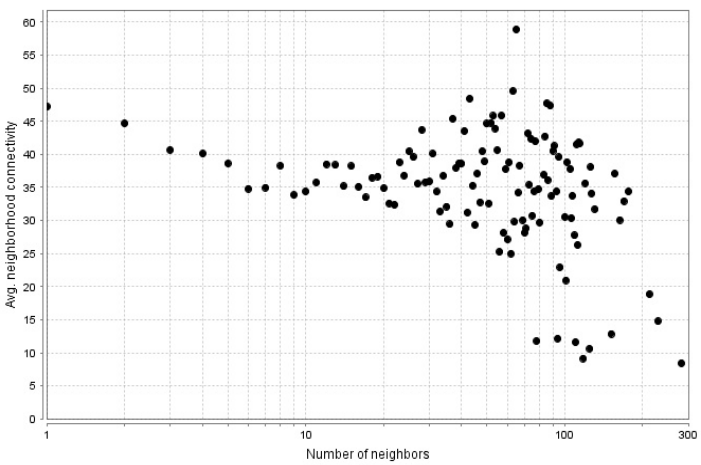
Εικόνα 7: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Mouse από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

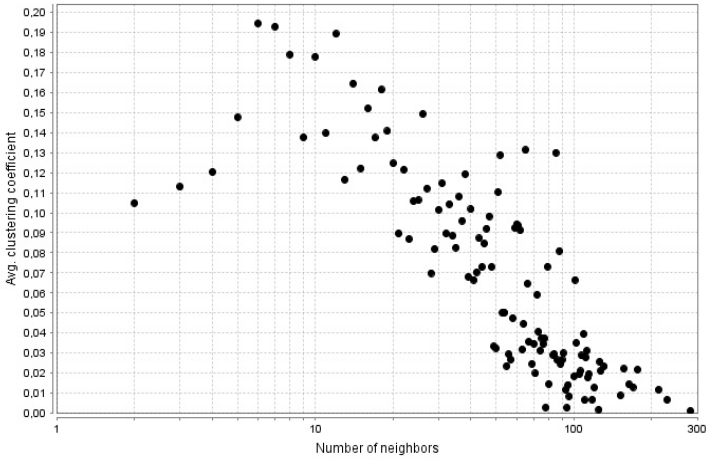



Number of nodes
Degree

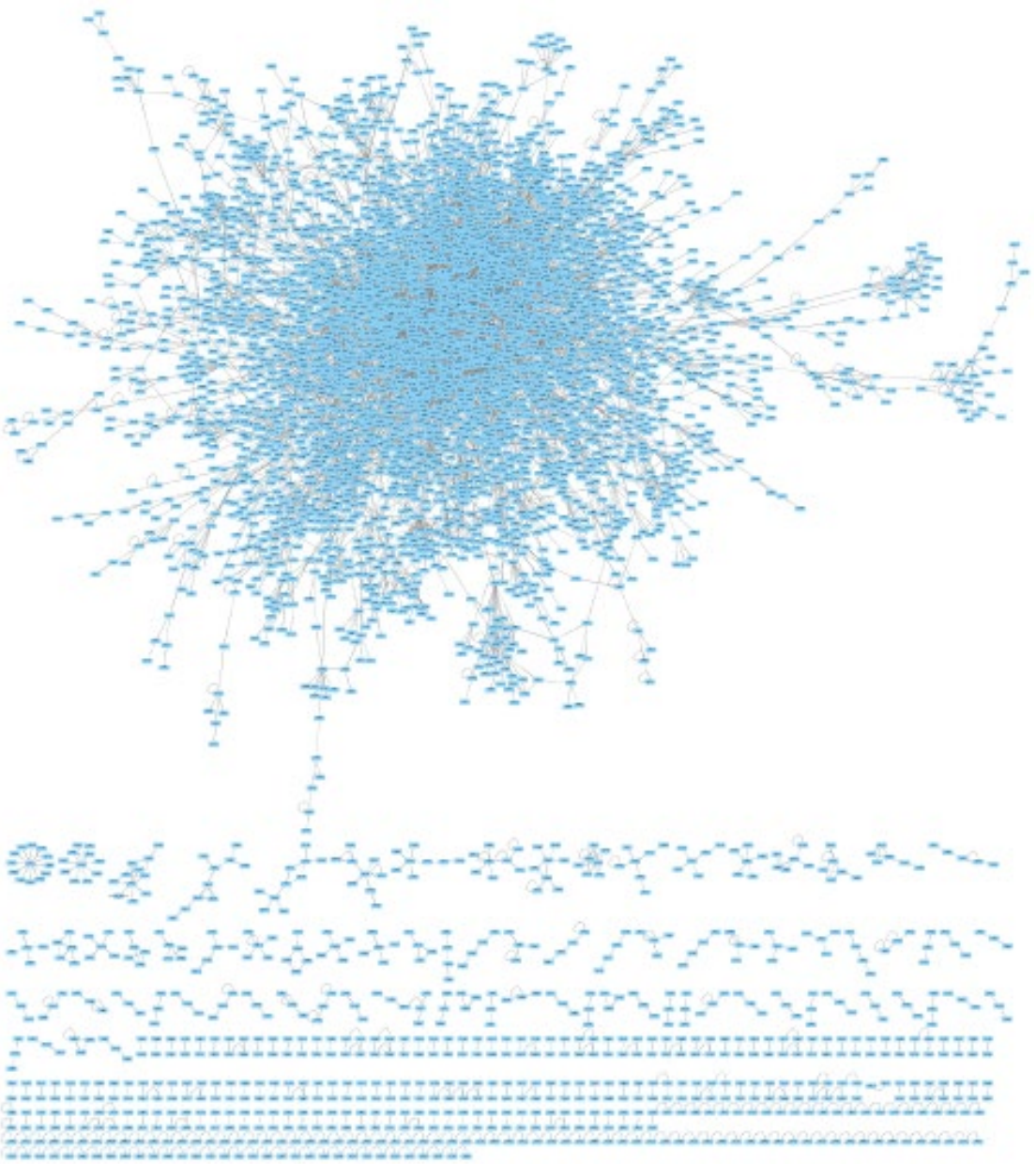

Closeness centrality
Number of neighbors

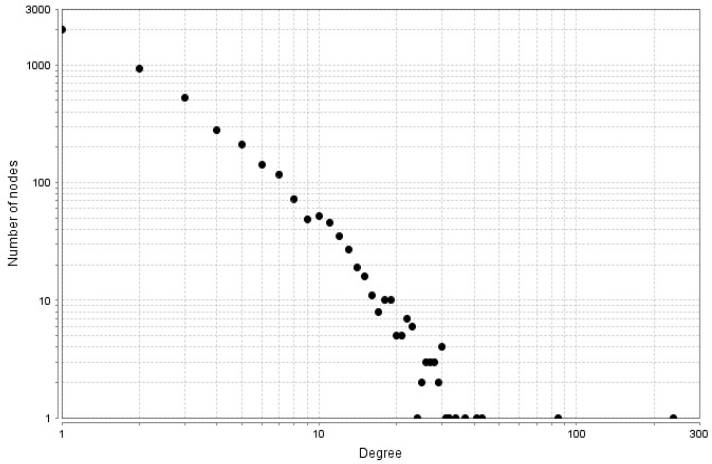

Betweenness centrality
Number of neighbors

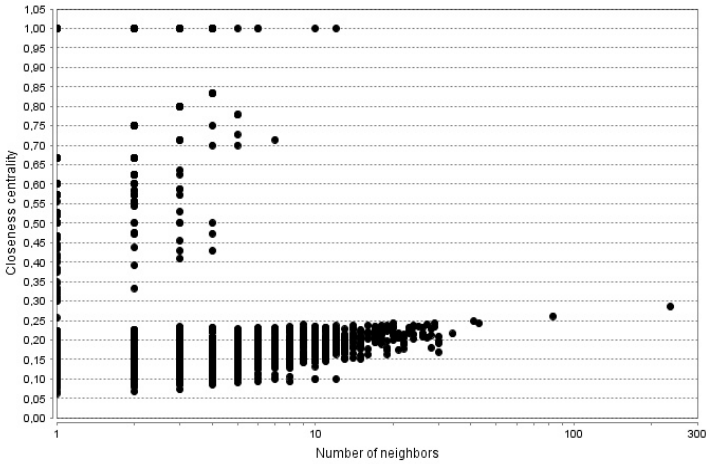

Avg. neighborhood connectivity
Number of neighbors

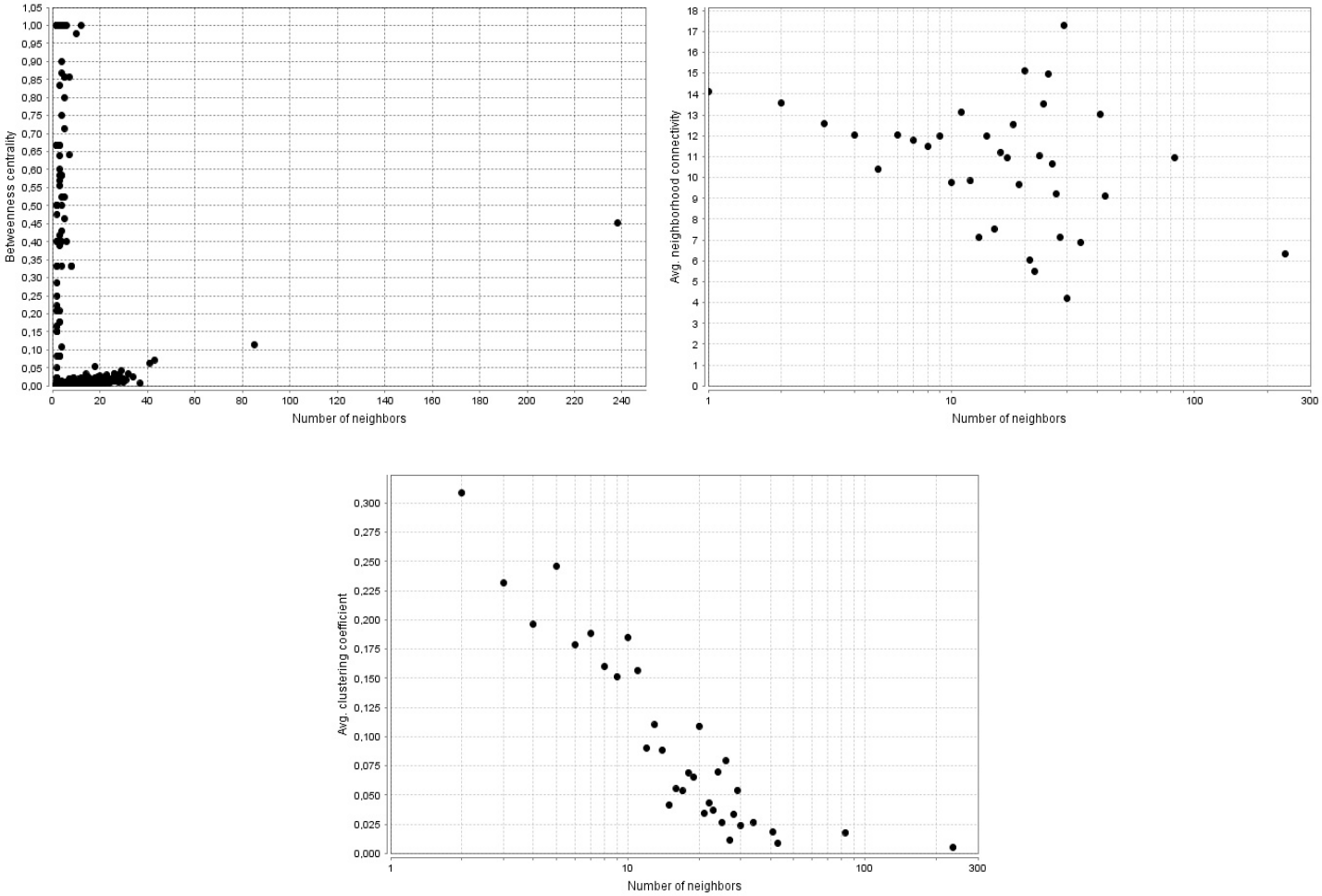


Εικόνα 8: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Yeast από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου




Number of nodes
3000
1000
100
10
1
1
10
100
300
Degree


Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
300
Number of neighbors

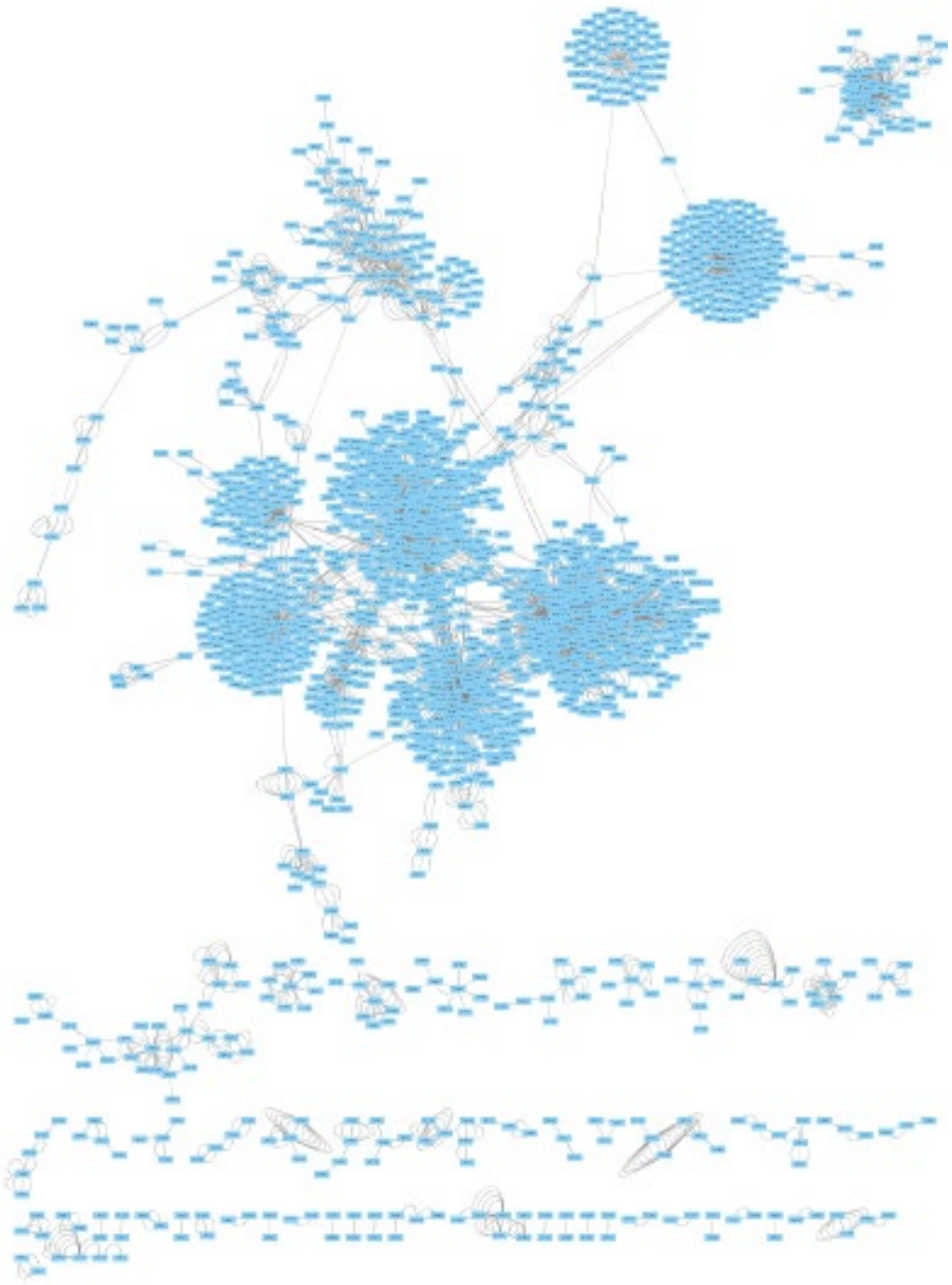


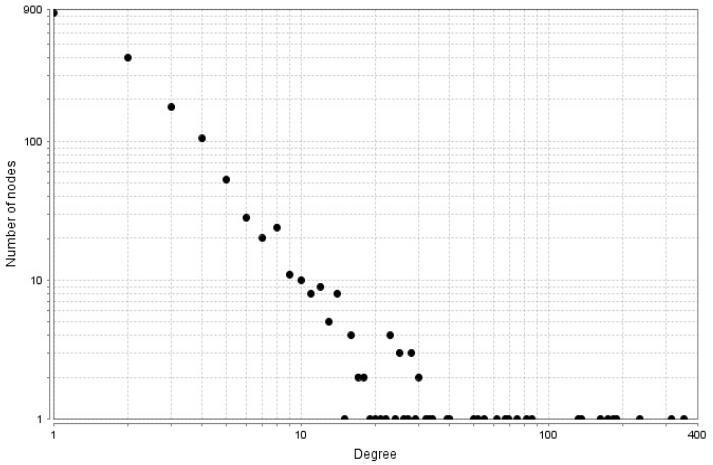
Εικόνα 9: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Homo Sapiens από την βάση DIP και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

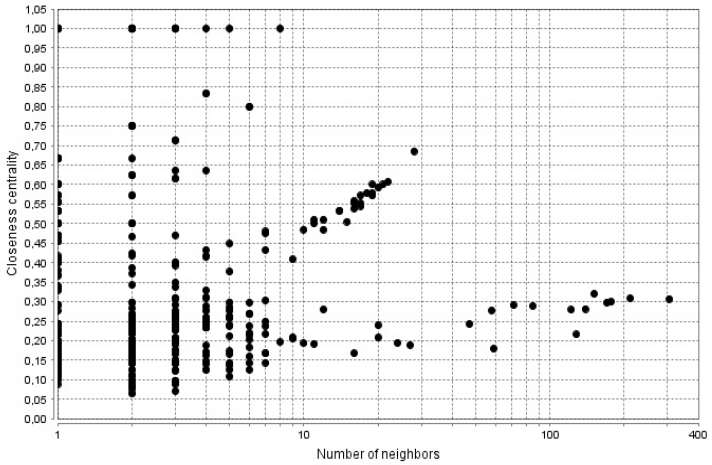
Η πρώτη ομάδα δικτύων που αναλύθηκαν προέρχονταν από την βάση DIP και περιείχαν πρωτεΐνες των οργανισμών Caenorhabditis elegans, Drosophila, Escherichia coli, Mouse, Yeast και Homo sapiens [Εικόνες 4 εώς 9]. Οι μετρικές που υπολογίστηκαν για κάθε δίκτυο συνοψίζονται στον πίνακα που ακολουθεί [Πίνακας 4]. Όπως είναι εμφανές τα σύνολα καλύπτουν περιπτώσεις με αλληλεπιδράσεις πρωτεϊνών μεσαίες κλίμακας , αφού οι κόμβοι και οι ακμές τους είναι μερικές χιλιάδες. Αρκετές ομοιότητες δείχνουν οι μετρικές Radius, Characteristic Path Length και Density. Ως προς τα υπόλοιπα μέτρα, τα σύνολα καταλαμβάνουν τιμές με μία ομοιόμορφη διακύμανση, ο οποίος ήταν και ο σκοπός της επιλογής τους, καθώς μέσω αυτού τα αποτελέσματα δεν επηρεάζονται από έλλειψη εξωτερικής ετερογένειας.

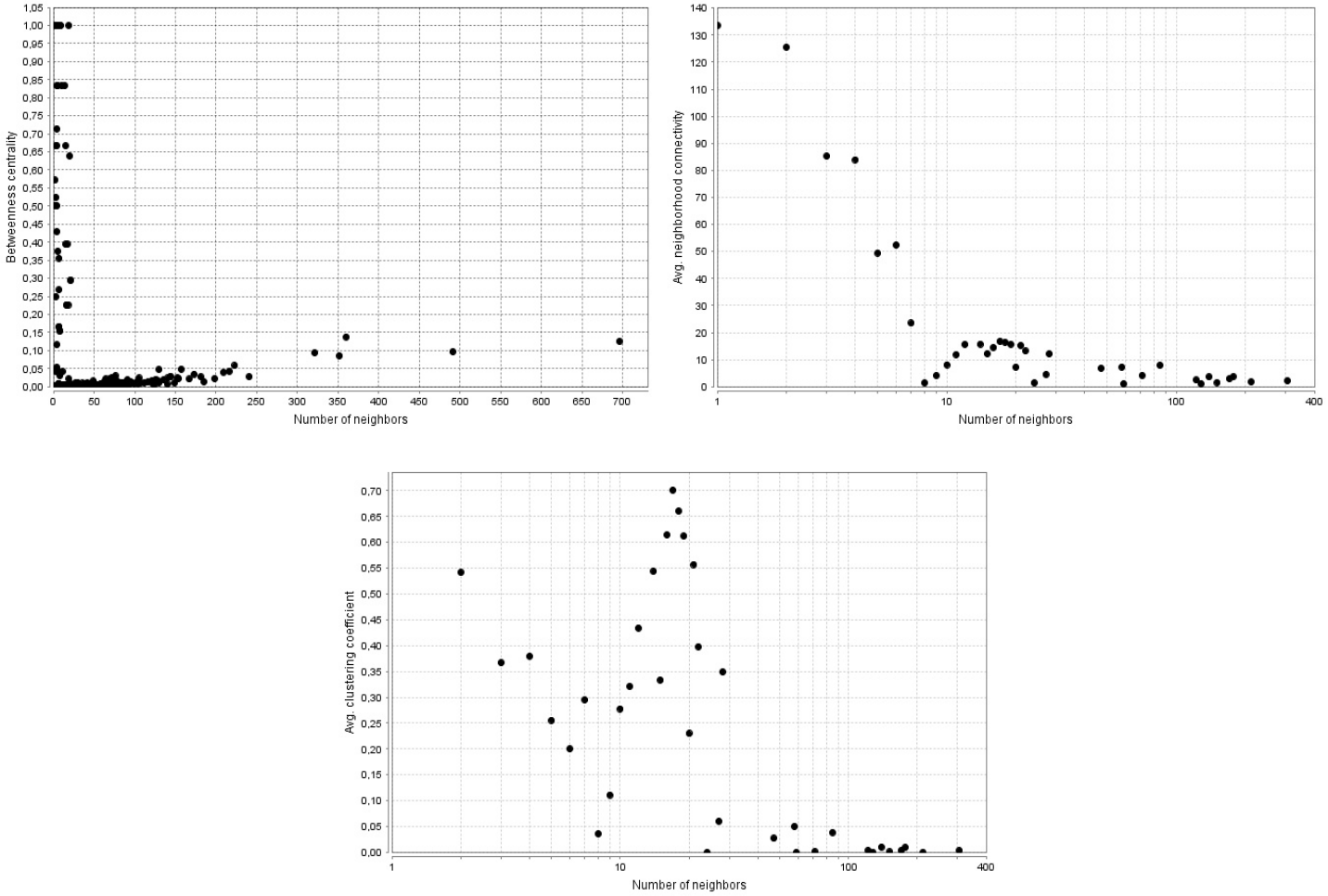
DIP	Nodes	Edges	Clustering Coef.	Diameter	Radius	Centralization	Shortest Paths	Char. Path Length	Avg. Degree	Density	Heterogeneity	Self-Loops
C. elegans	2423	3628	0.027	13	1	0.073	4582388(78%)	4.859	2.935	0.001	2.358	14
Drosophila	972	1401	0.059	13	1	0.025	609498(64%)	5.584	2.7	0.003	1.067	89
E. coli	2924	12246	0.095	11	1	0.064	6368394(74%)	3.972	7.943	0.003	2.077	634
Mouse	2246	2495	0.059	22	1	0.019	1413826(28%)	7.632	2.096	0.001	1.196	139
Yeast	5124	22908	0.099	10	1	0.053	25588484(97%)	3.98	8.802	0.002	1.768	353
Homo Sapiens	4615	7417	0.122	22	1	0.051	14362292(67%)	6.297	2.987	0.001	1.724	508

Πίνακας 4: Συγκεντρωτικός πίνακας περιγραφικών χαρακτηριστικών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών από την βάση DIP

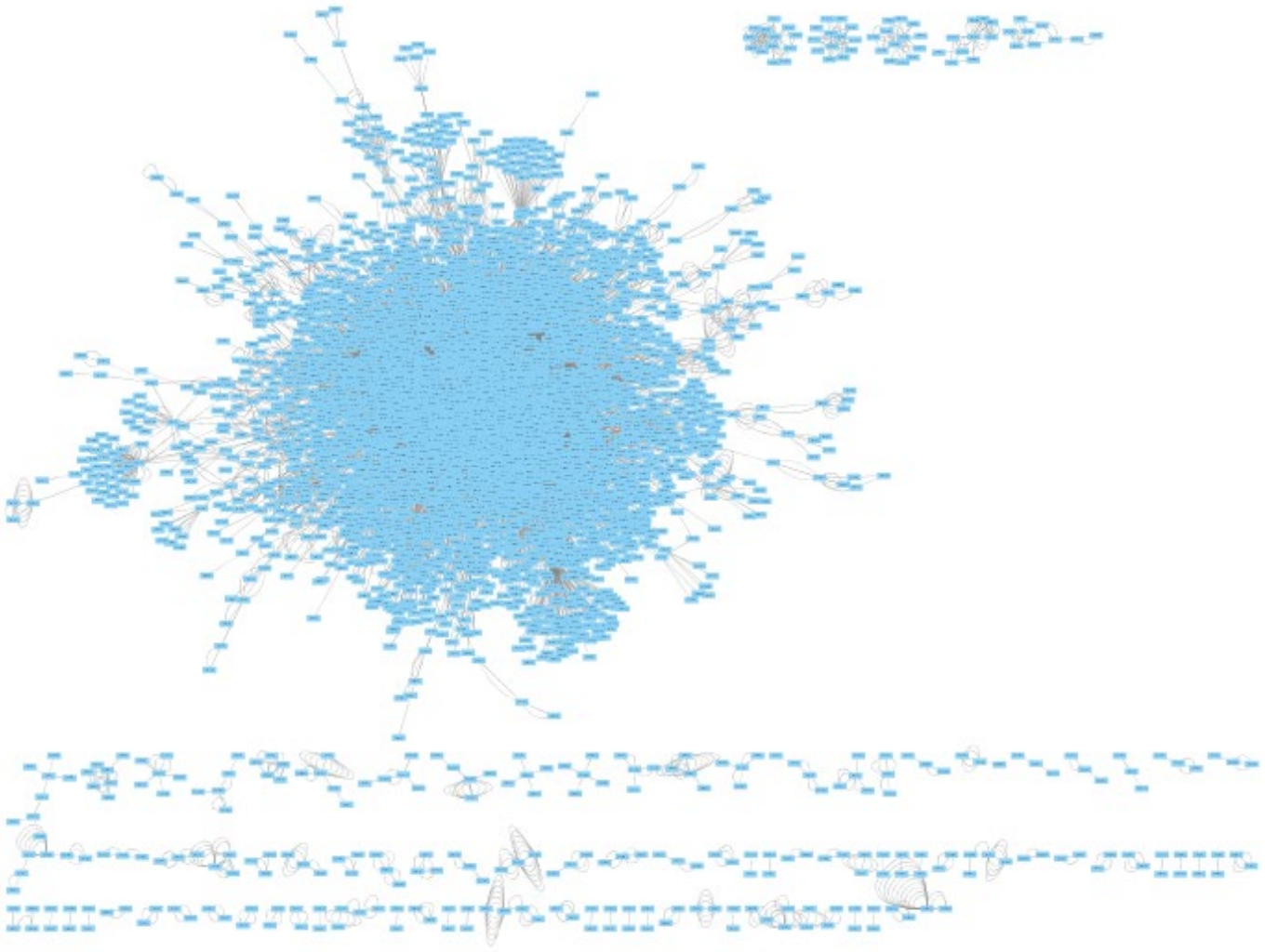


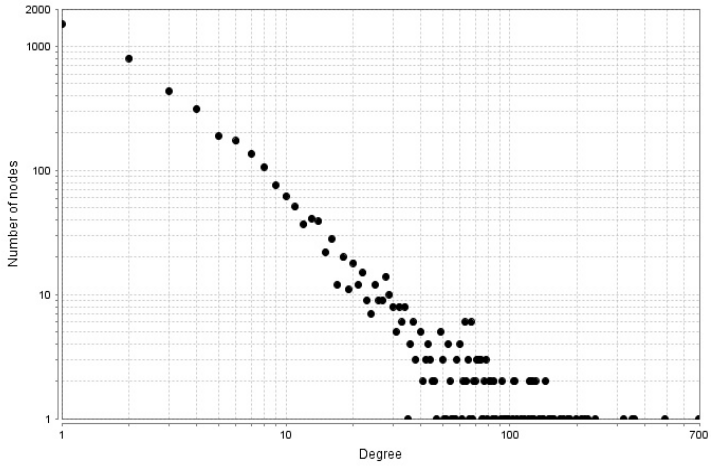

900
100
10
1
Number of nodes
1
10
100
400
Degree

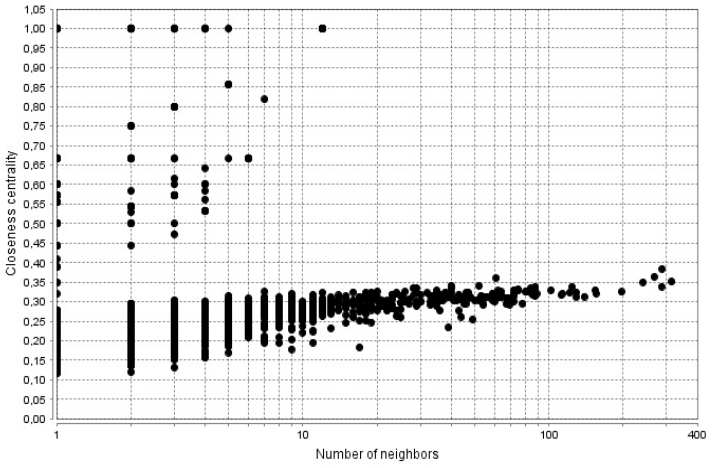

1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Closeness centrality
1
10
100
400
Number of neighbors

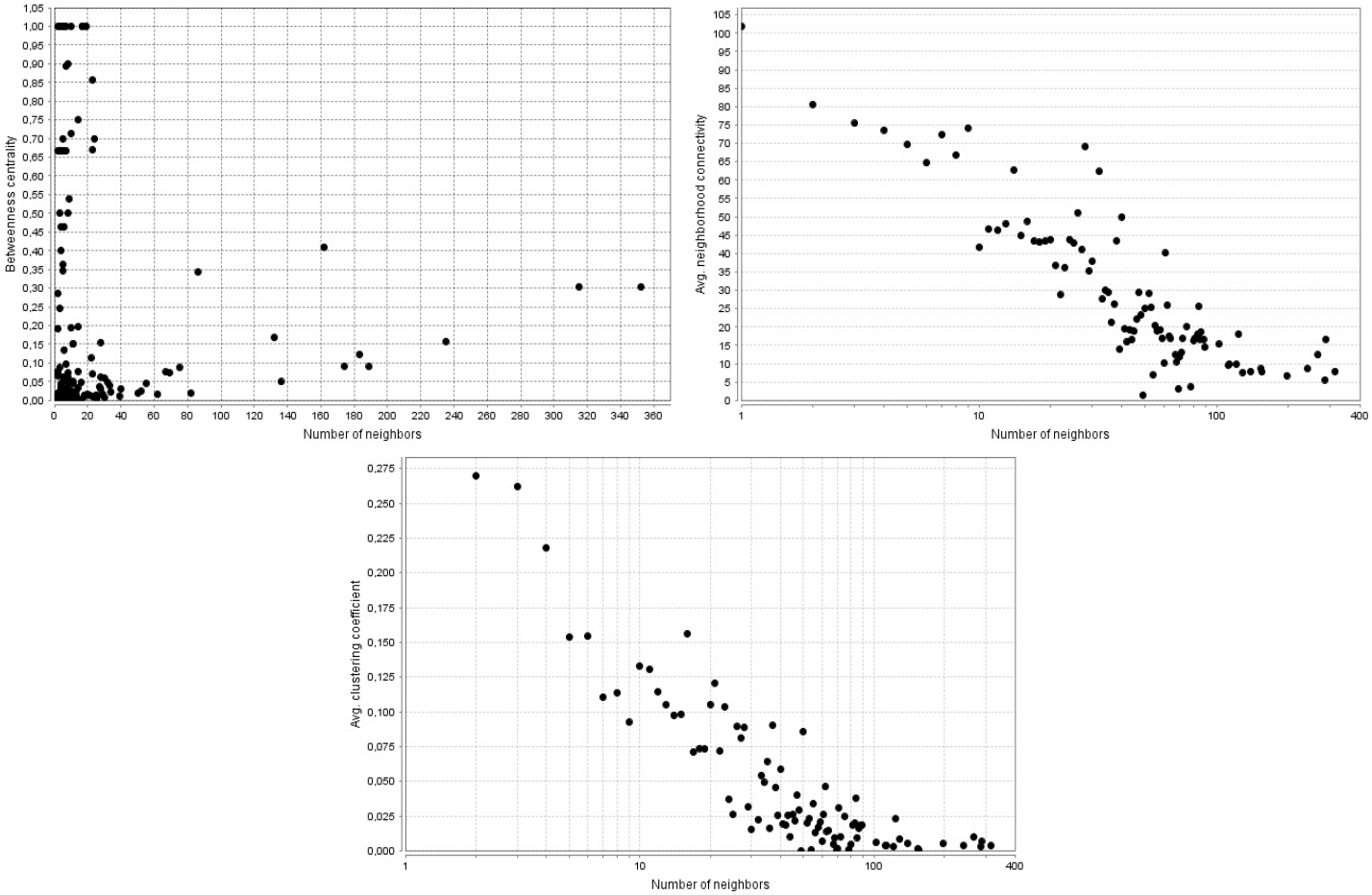


Εικόνα 10: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Cardiac από την βάση IntAct και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

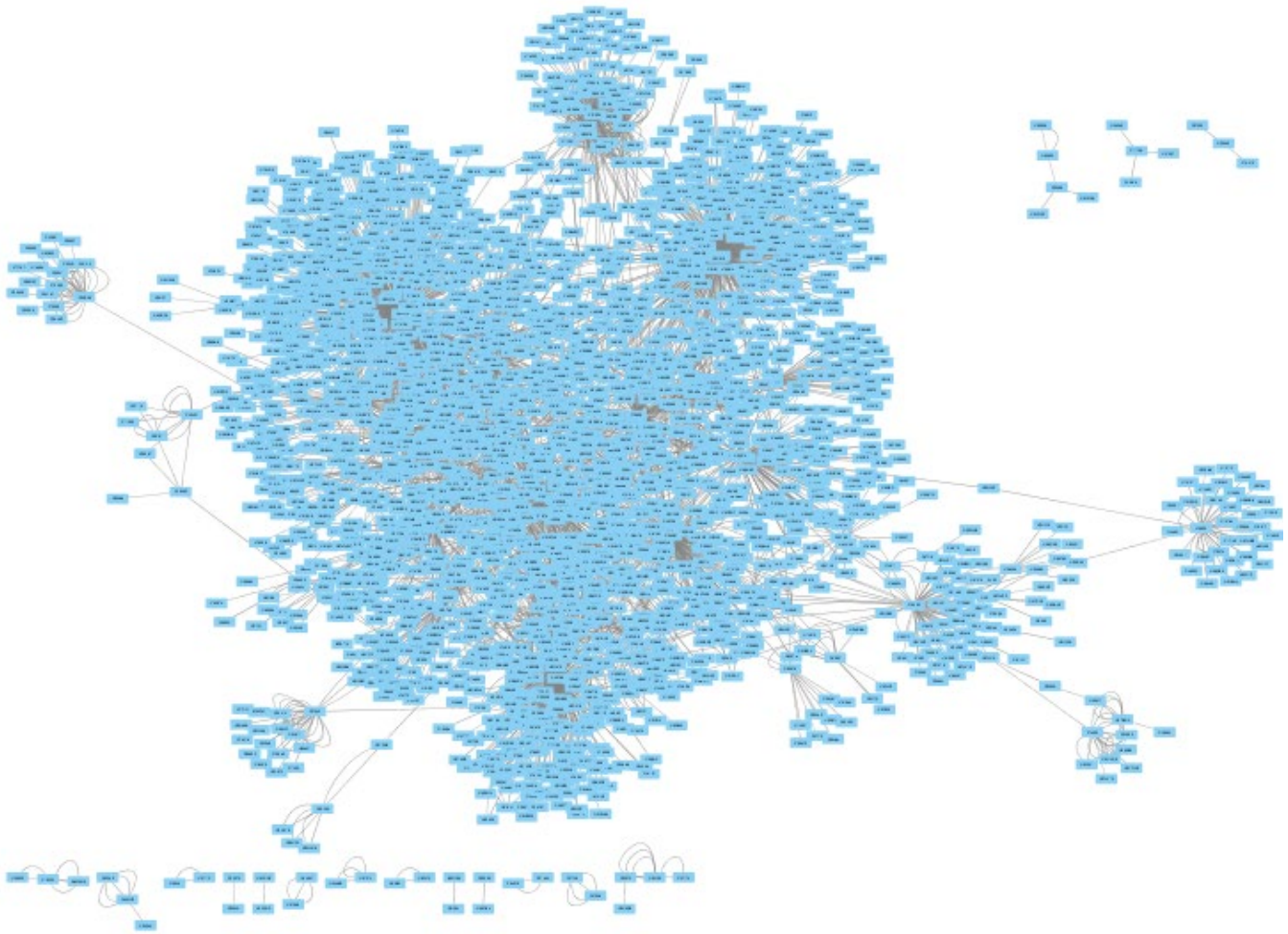


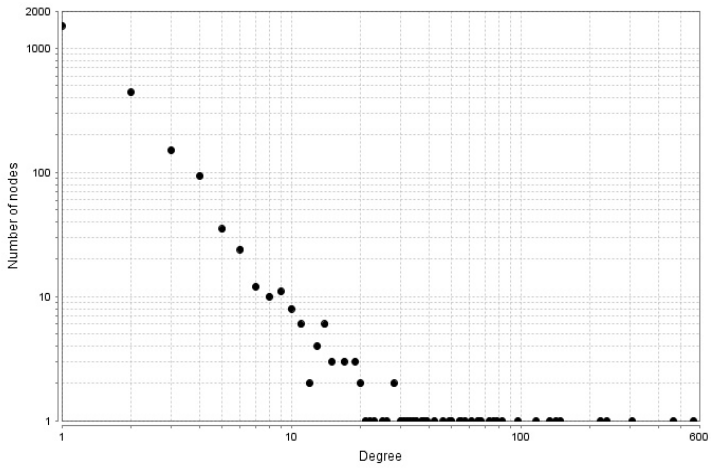

Number of nodes
2000
1000
100
10
1
1
10
100
700
Degree

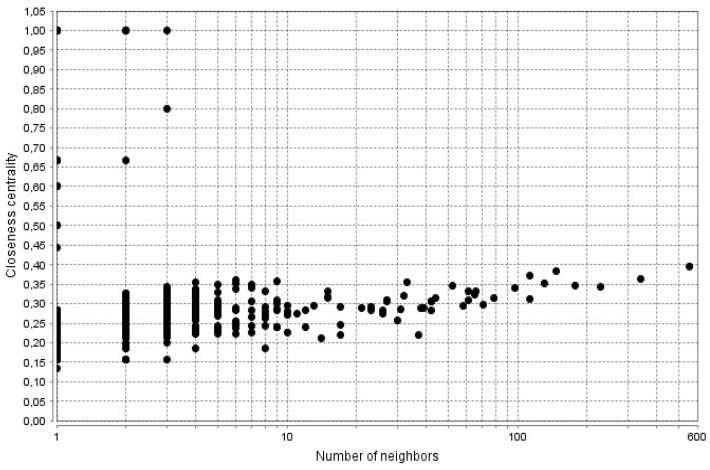

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
400
Number of neighbors

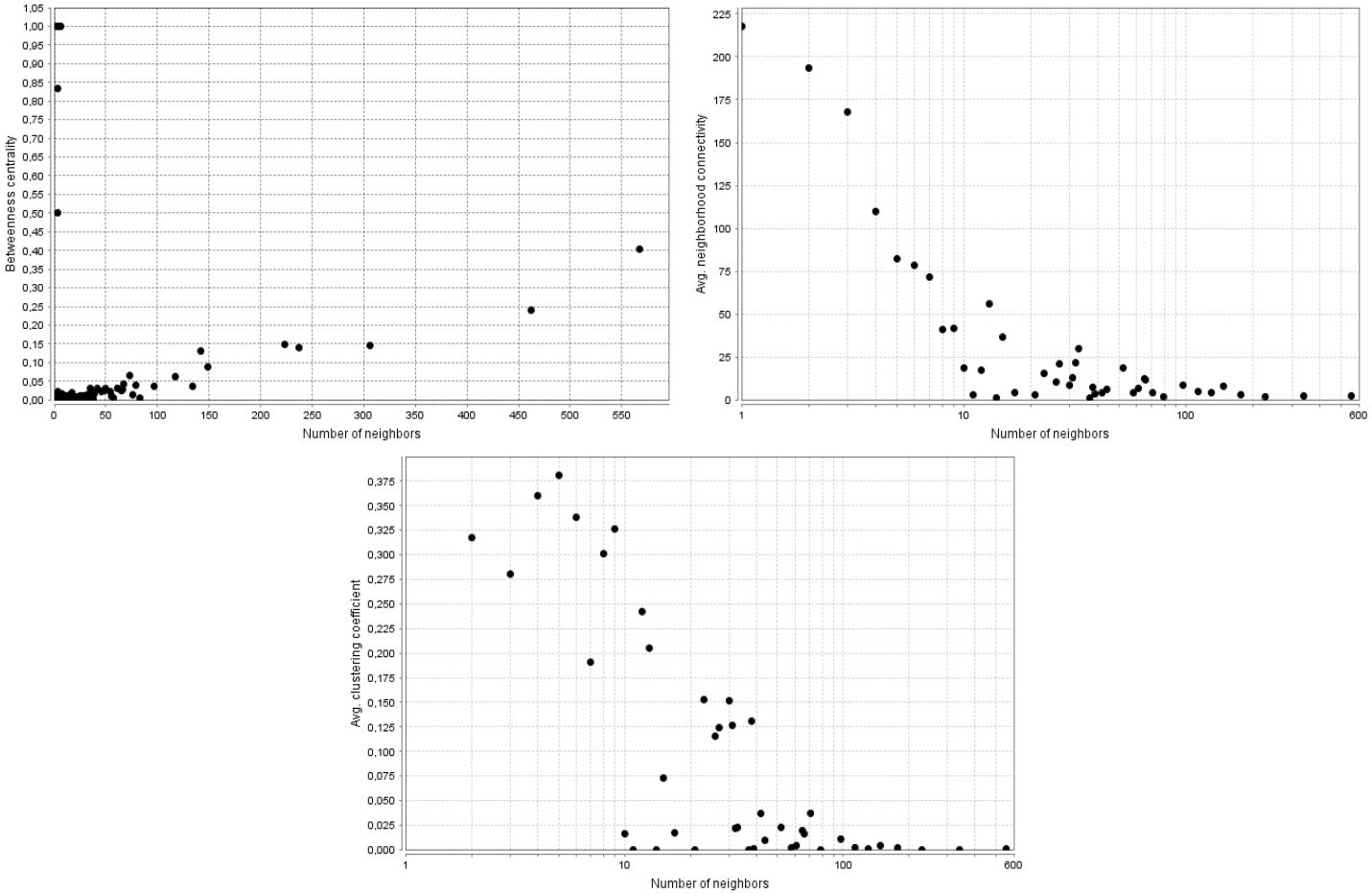


Εικόνα 11: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Cancer από την βάση IntAct και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

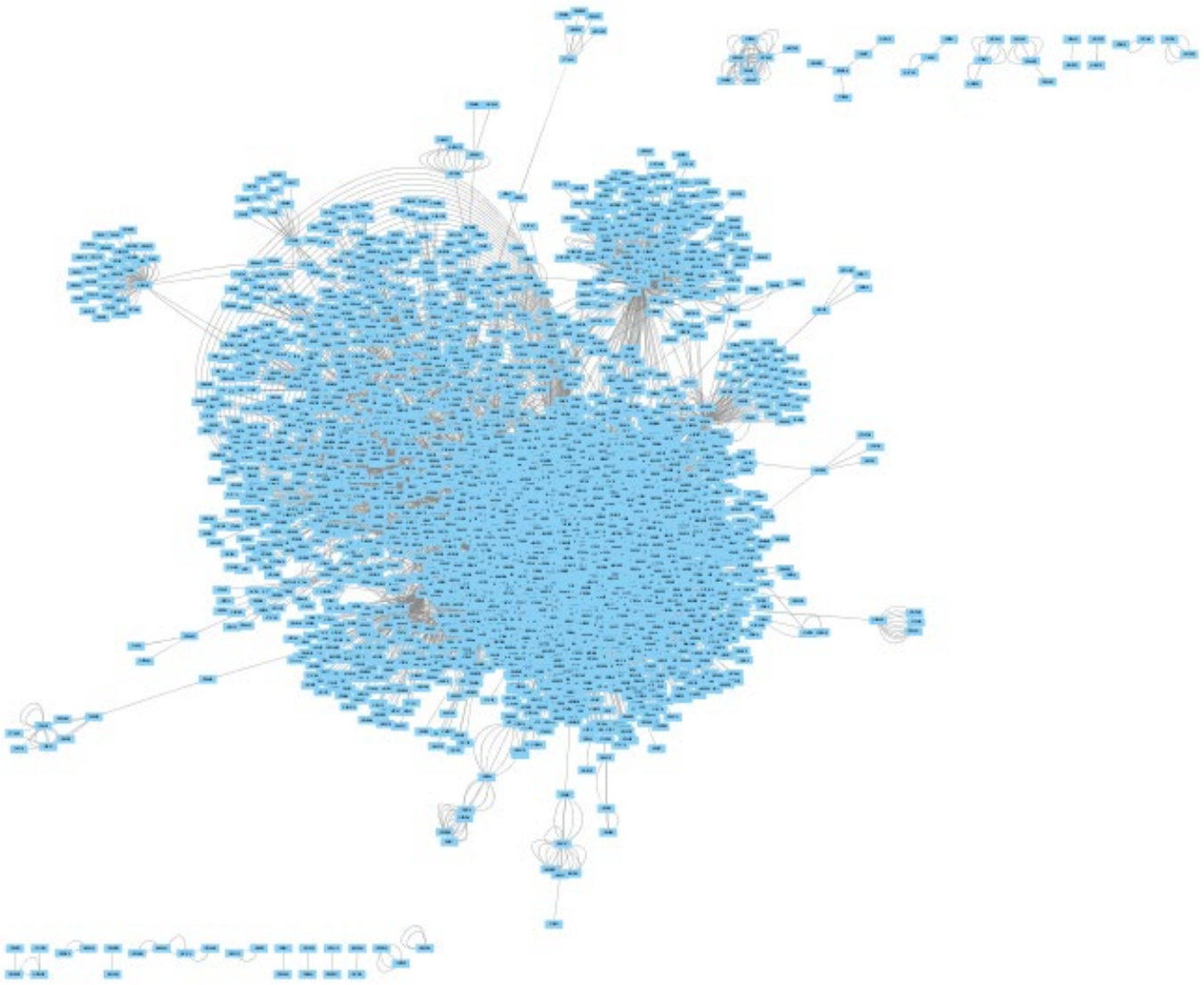


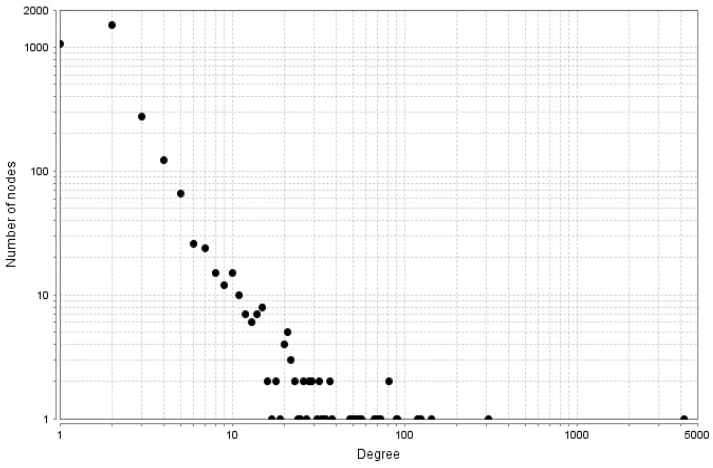

Number of nodes
2000
1000
100
10
1
1
10
100
600
Degree

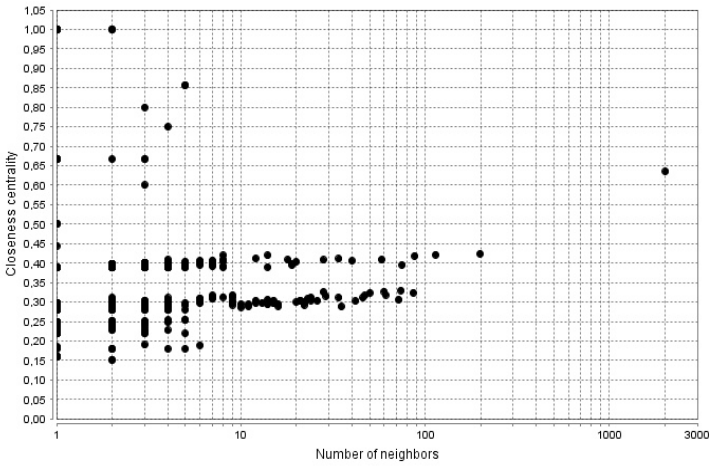

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
600
Number of neighbors

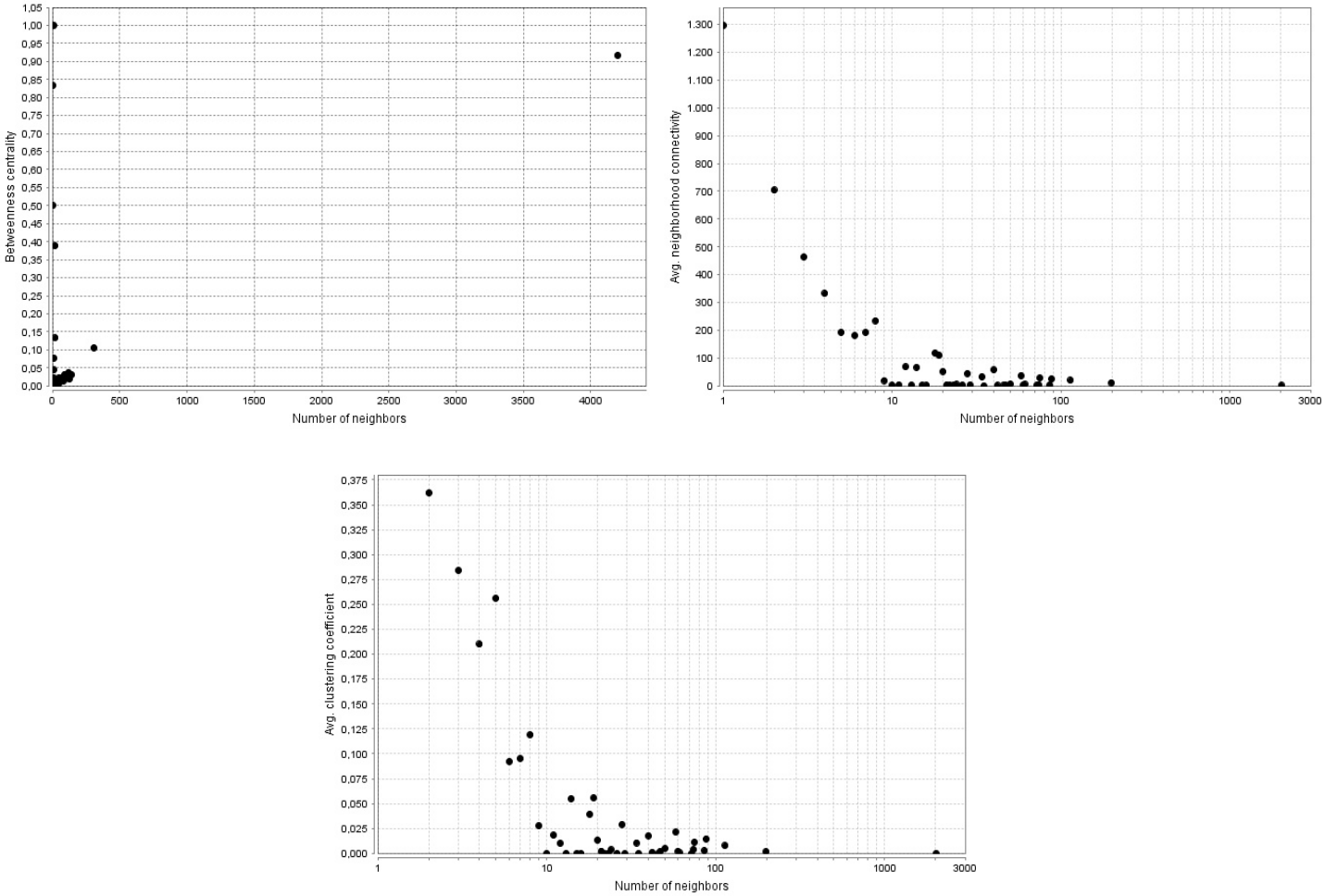


Εικόνα 12: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Alzheimer από την βάση IntAct και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

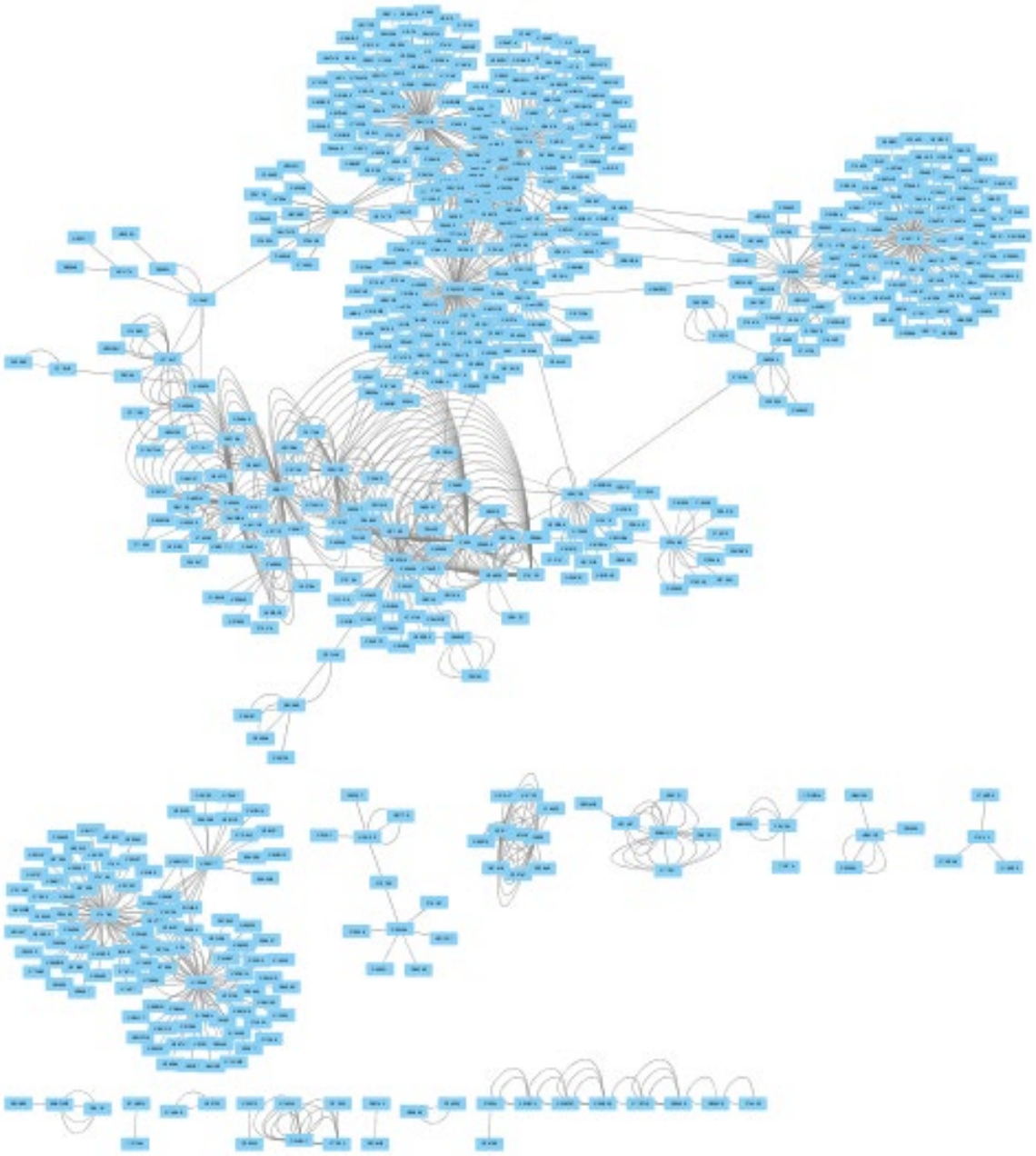


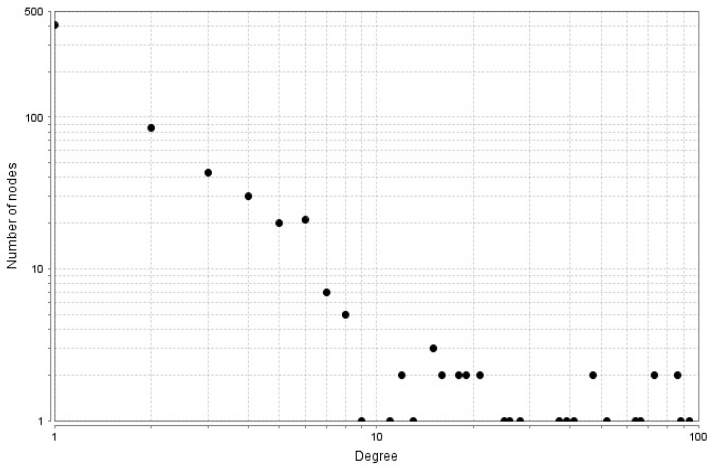

Number of nodes
2000
1000
100
10
1
1
10
100
1000
5000
Degree

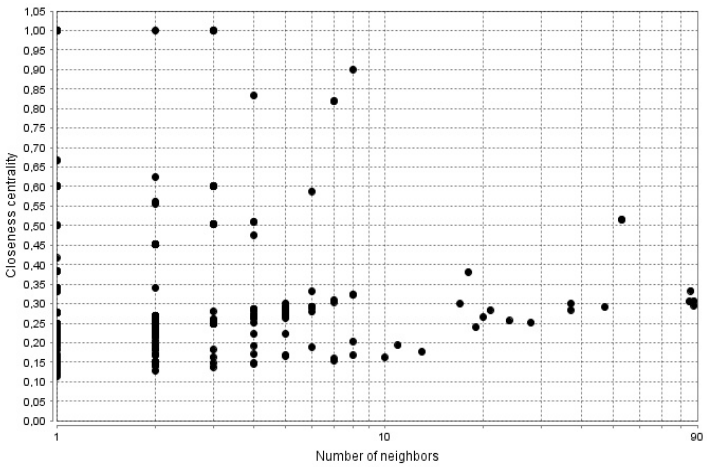

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
1000
3000
Number of neighbors

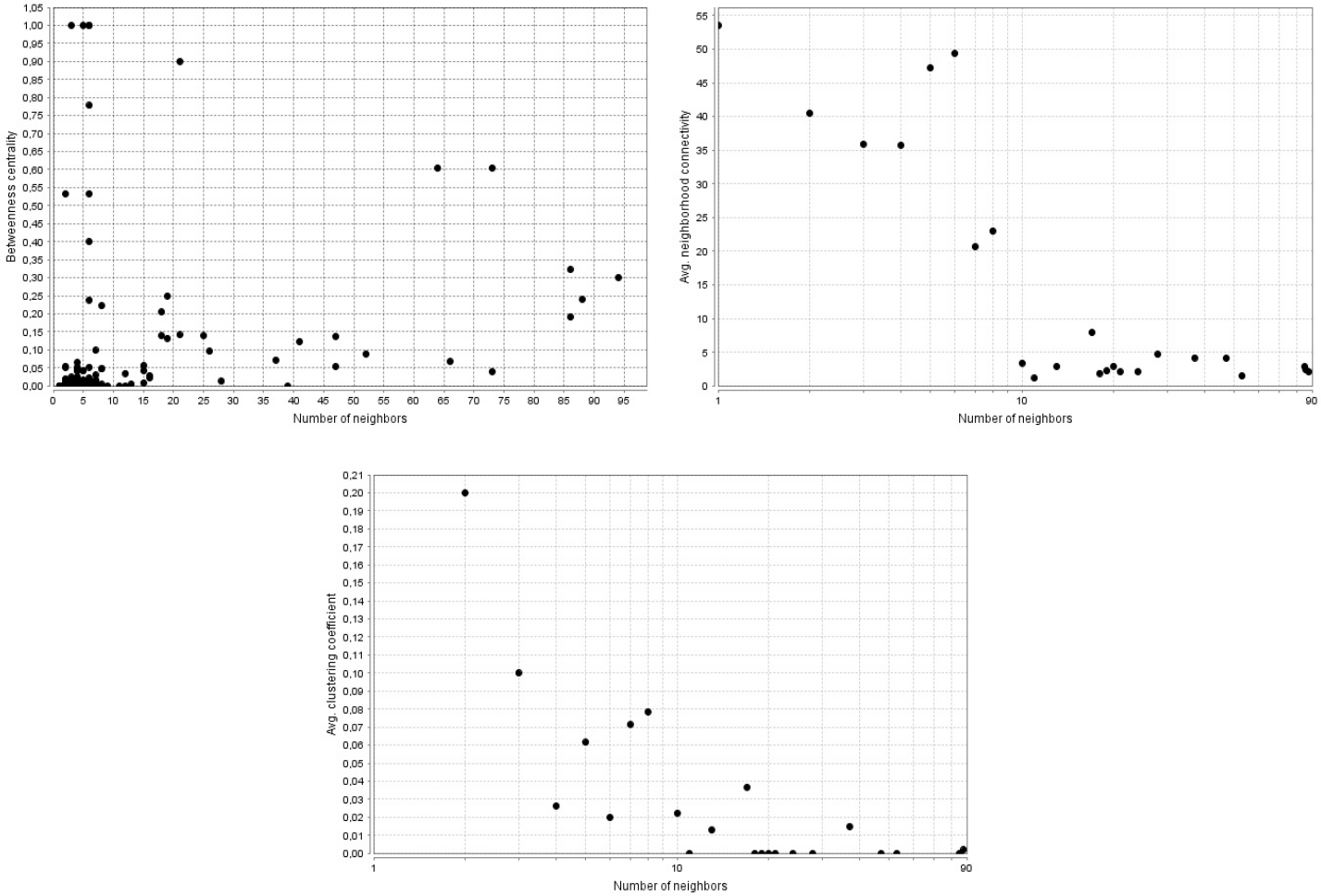


Εικόνα 13: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Parkinson από την βάση IntAct και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου




Number of nodes
500
100
10
1
1
10
100
Degree


Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
90
Number of neighbors

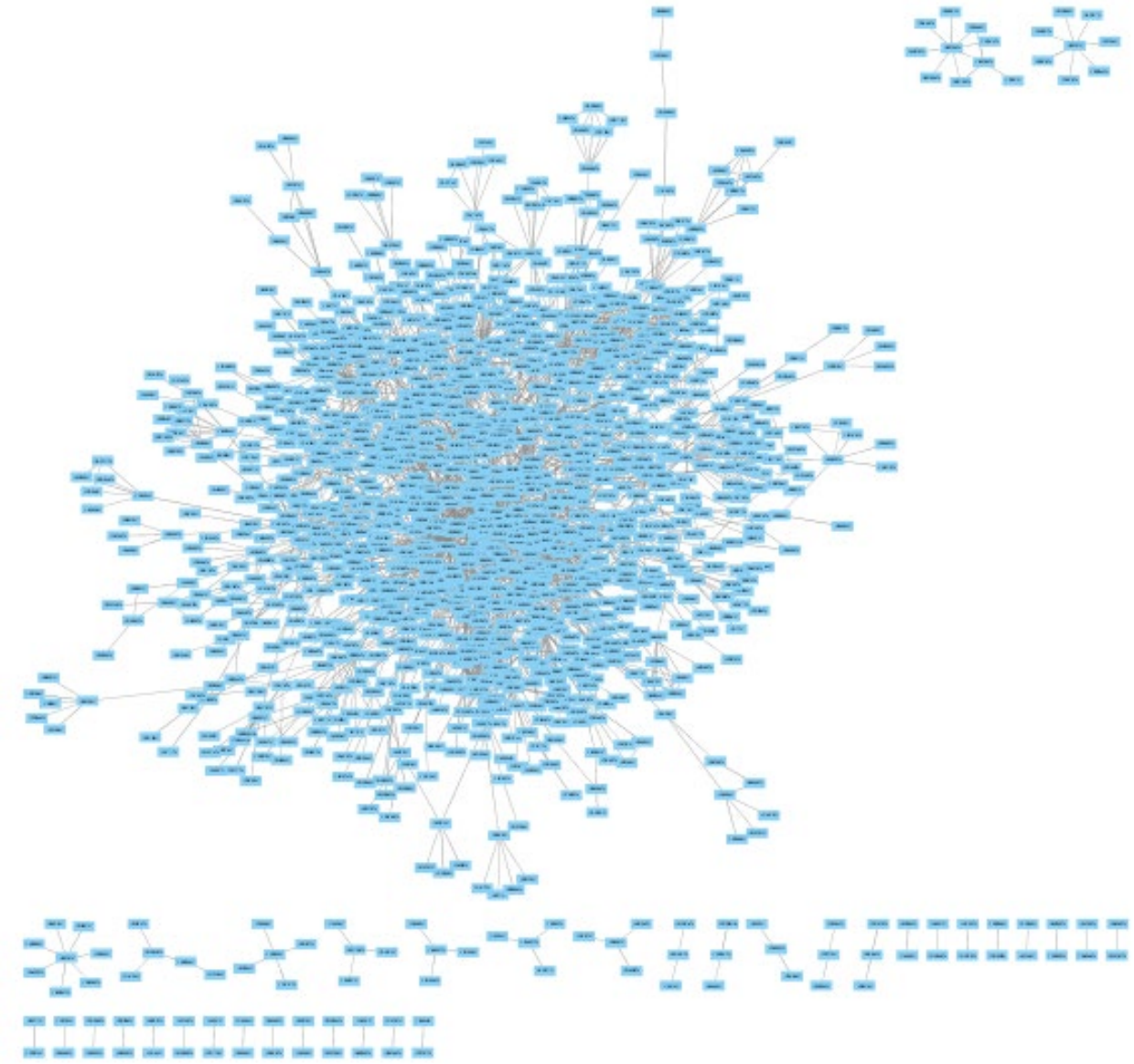


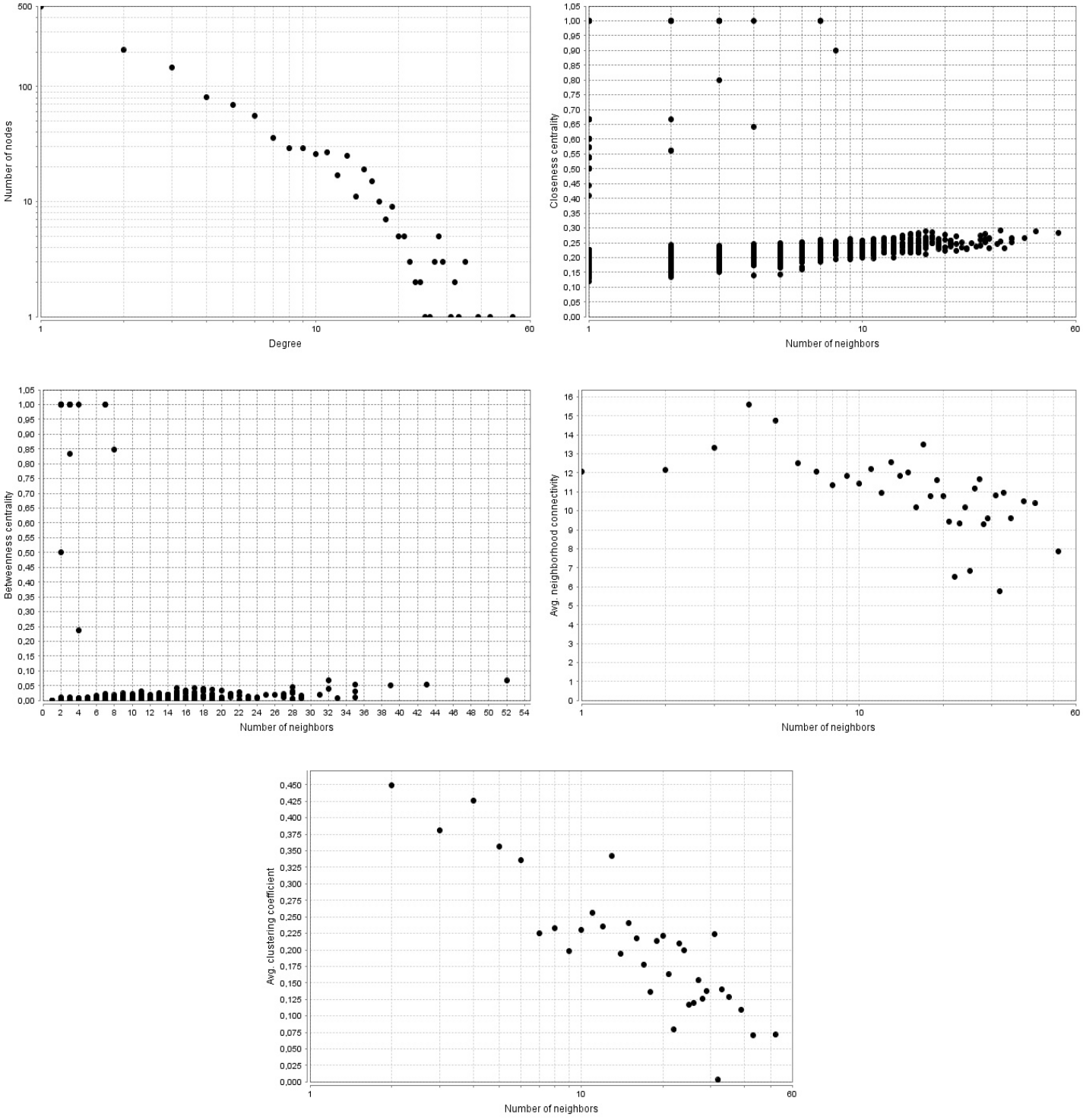
Εικόνα 14: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Covid19 από την βάση IntAct και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

Τα πέντε δίκτυα πρωτεϊνικών αλληλεπιδράσεων από την βάση IntAct αποτελούνται από λιγότερες σχέσεις πρωτεϊνών [Εικόνες 10 εώς 14], το οποίο είναι λογικό, καθώς η συσχέτιση πρωτεϊνών με μονοπάτια ασθενειών απαιτεί περαιτέρω έρευνα. Ομοιότητα, όπως και προηγουμένως, εμφανίζουν για τα πεδία της Ακτίνας, του Χαρακτηριστικού Συντομότερου Μονοπατιού και της Πυκνότητας του δικτύου, ενώ και η Κεντρικότητα Βαθμού της πλειοψηφίας τον δικτύων είναι ίδια [Πίνακας 5]. Οι μεγαλύτερες αποκλίσεις συναντώνται στο δίκτυο της νόσου Parkinson, το οποίο και οπτικά [Εικόνα 13] φανερώνει την τάση του για δημιουργία μιας πολύ μεγάλης ομάδας και μόνο ελάχιστων μικρότερων, το οποίο εκφράζεται με υψηλές τιμές Centralization και Heterogenity, χαμηλή τιμή Clustering Coefficient και πάνω από δέκα εκατομμύρια συντομότερα μονοπάτια ανάμεσα στους κόμβους του.

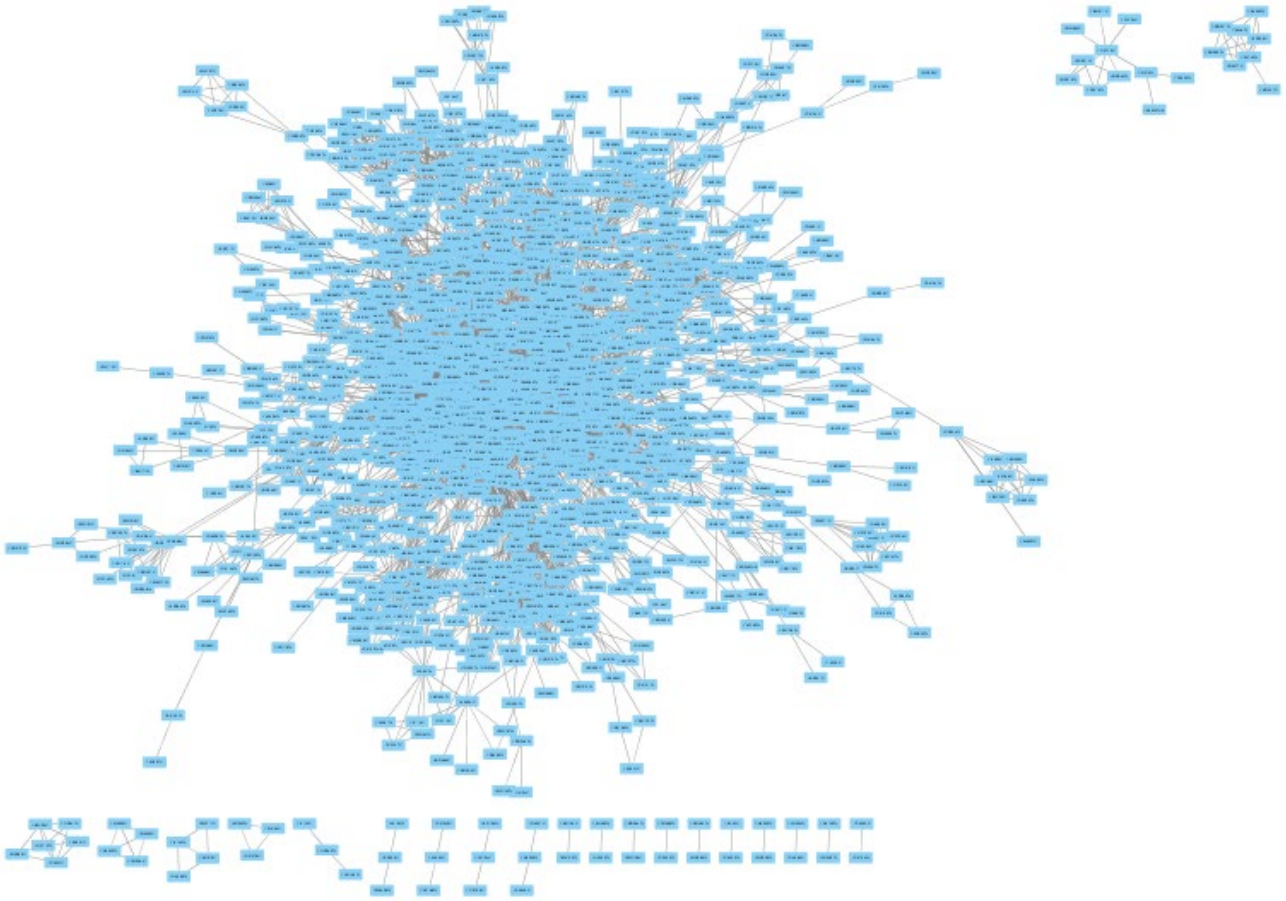
IntAct	Nodes	Edges	Clustering Coef.	Diameter	Radius	Centralization	Shortest Paths	Char. Path Length	Avg. Degree	Density	Heterogeneity	Self-Loops
Cardiac	1745	3611	0.156	19	1	0.173	2172114(71%)	4.833	2.816	0.002	4.613	70
Cancer	4375	16615	0.107	13	1	0.071	16448290(85%)	4.171	4.533	0.001	3.149	169
Alzheimer	2381	4078	0.078	10	1	0.232	5468654(96%)	3.908	2.776	0.001	5.841	54
Parkinson	3222	6844	0.062	10	1	0.626	10033168(96%)	3.047	2.502	0.001	14.389	107
Covid19	646	1179	0.035	13	1	0.133	238052(57%)	4.539	2.709	0.004	2.982	86

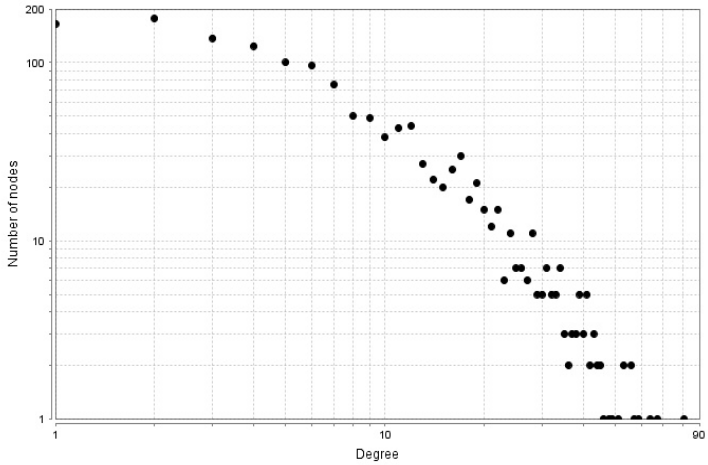
Πίνακας 5: Συγκεντρωτικός πίνακας περιγραφικών χαρακτηριστικών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών από την βάση IntAct

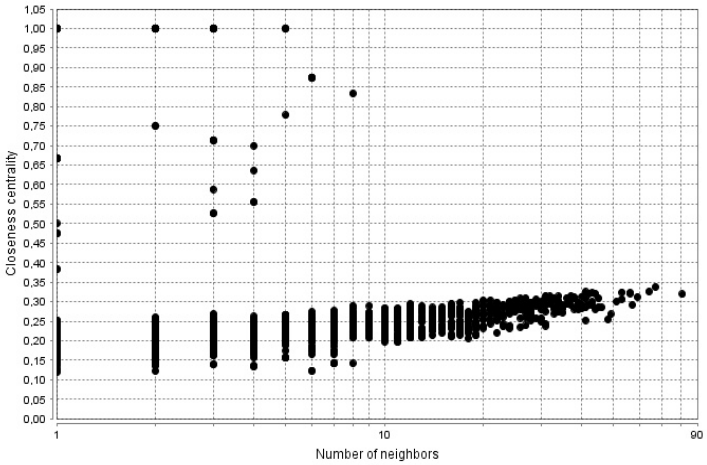


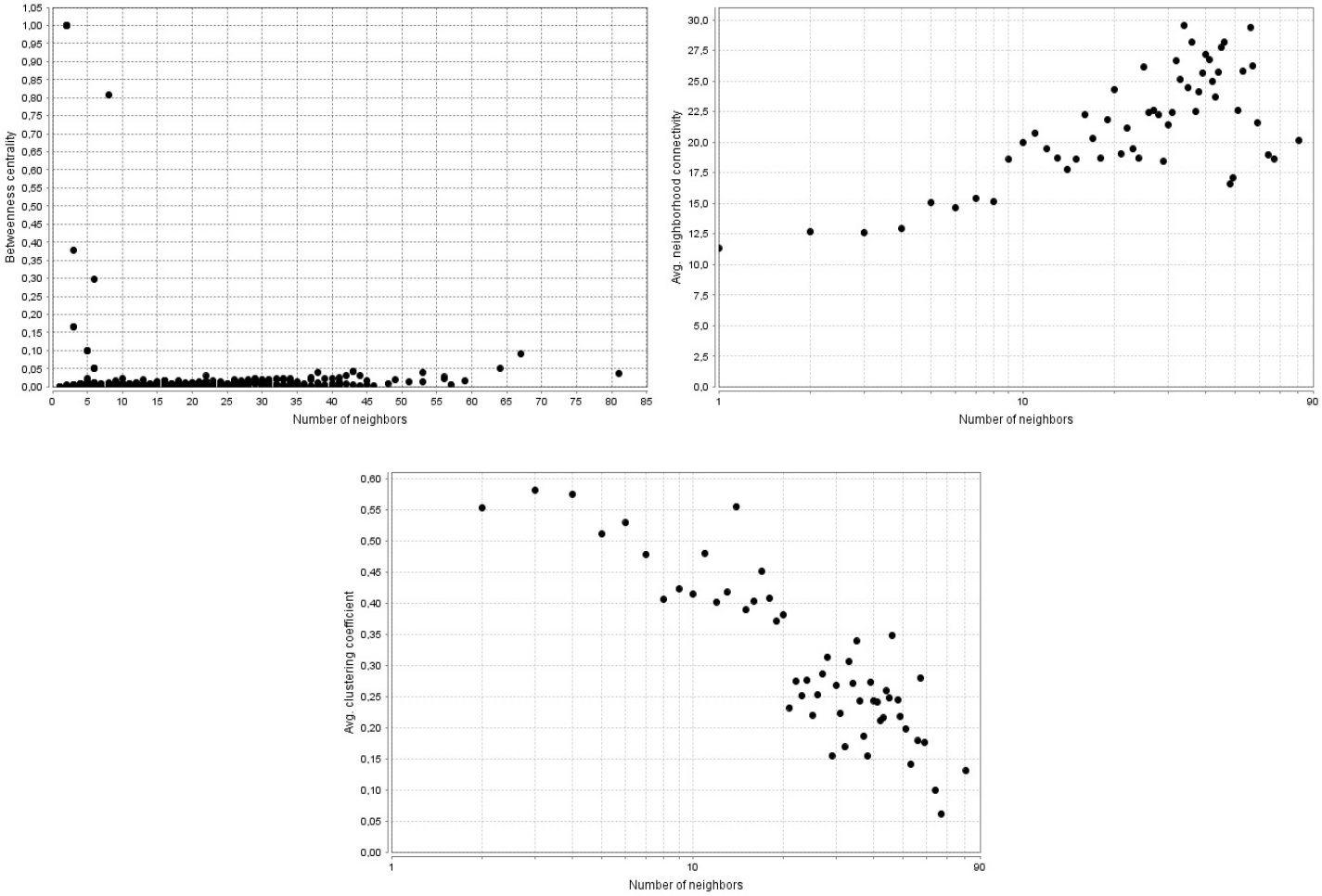


Εικόνα 15: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Gavin 2002 και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

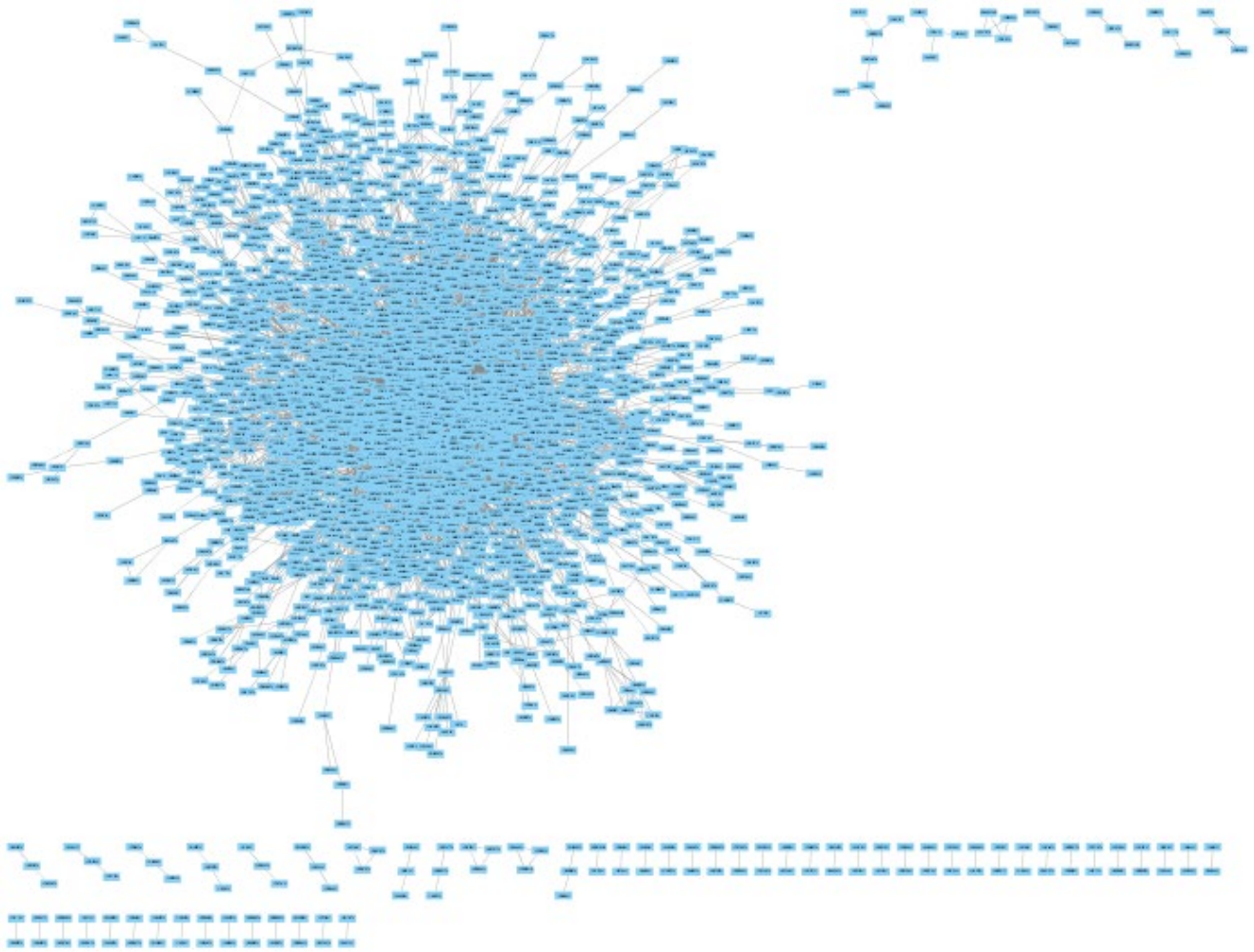


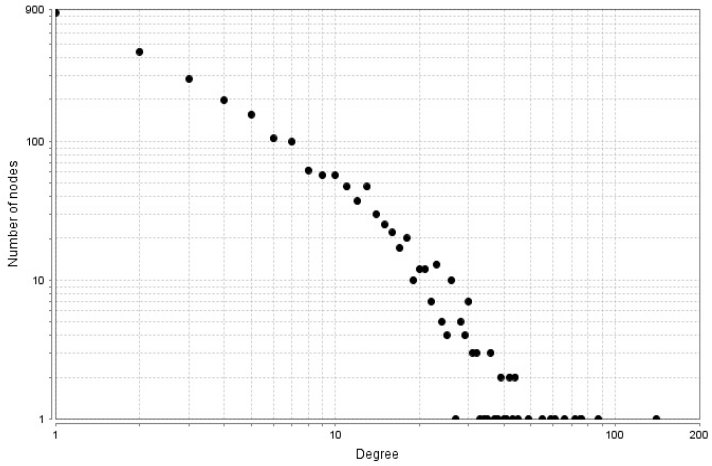

Number of nodes
200
100
10
1
1
10
90
Degree

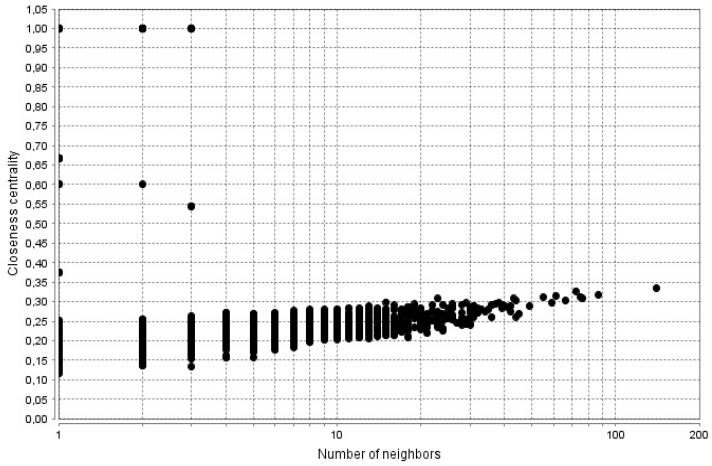

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
90
Number of neighbors

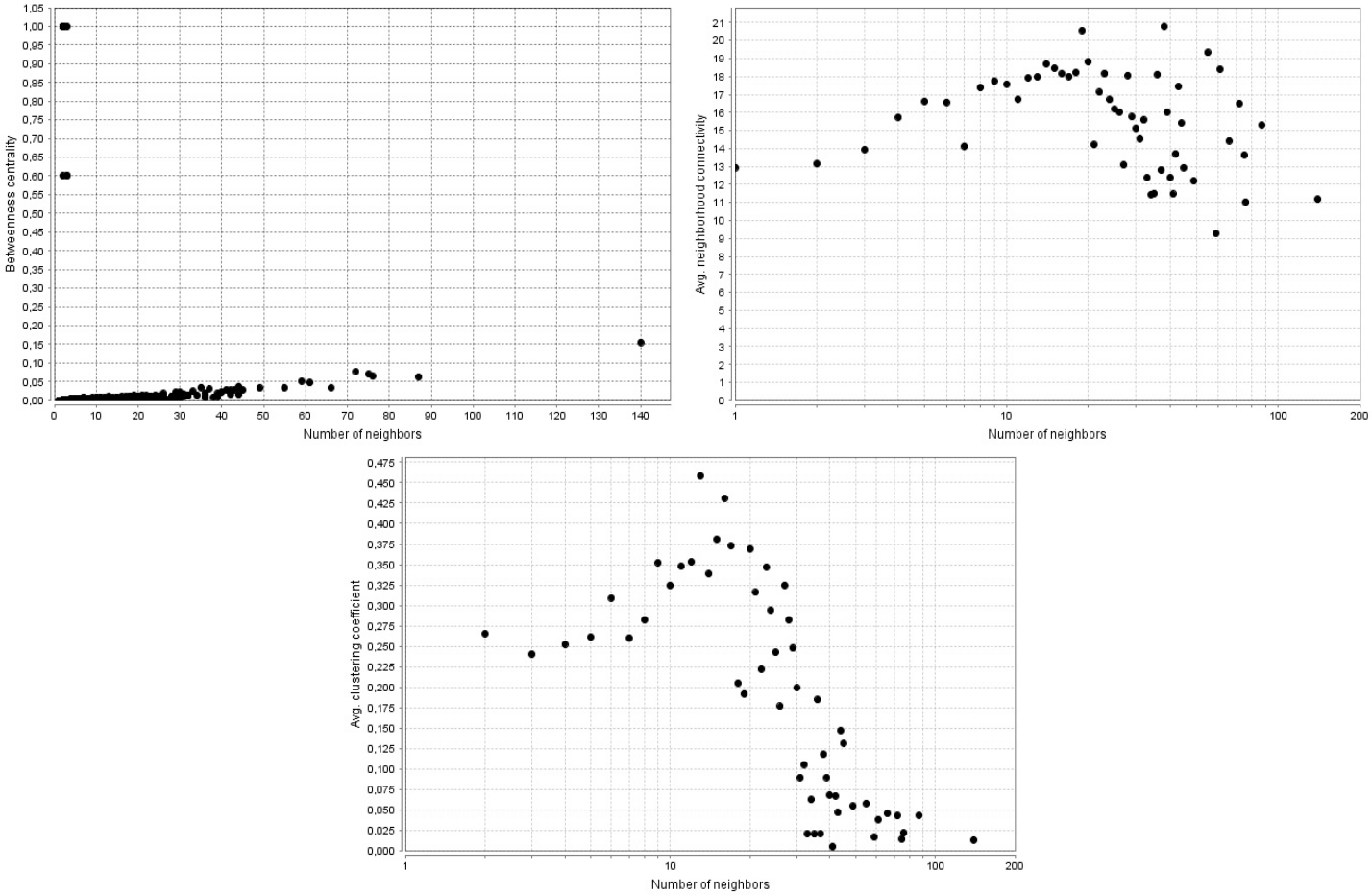


Εικόνα 16: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Gavin 2006 και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου




900
100
10
1
Number of nodes
1
10
100
200
Degree


1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Closeness centrality
1
10
100
200
Number of neighbors

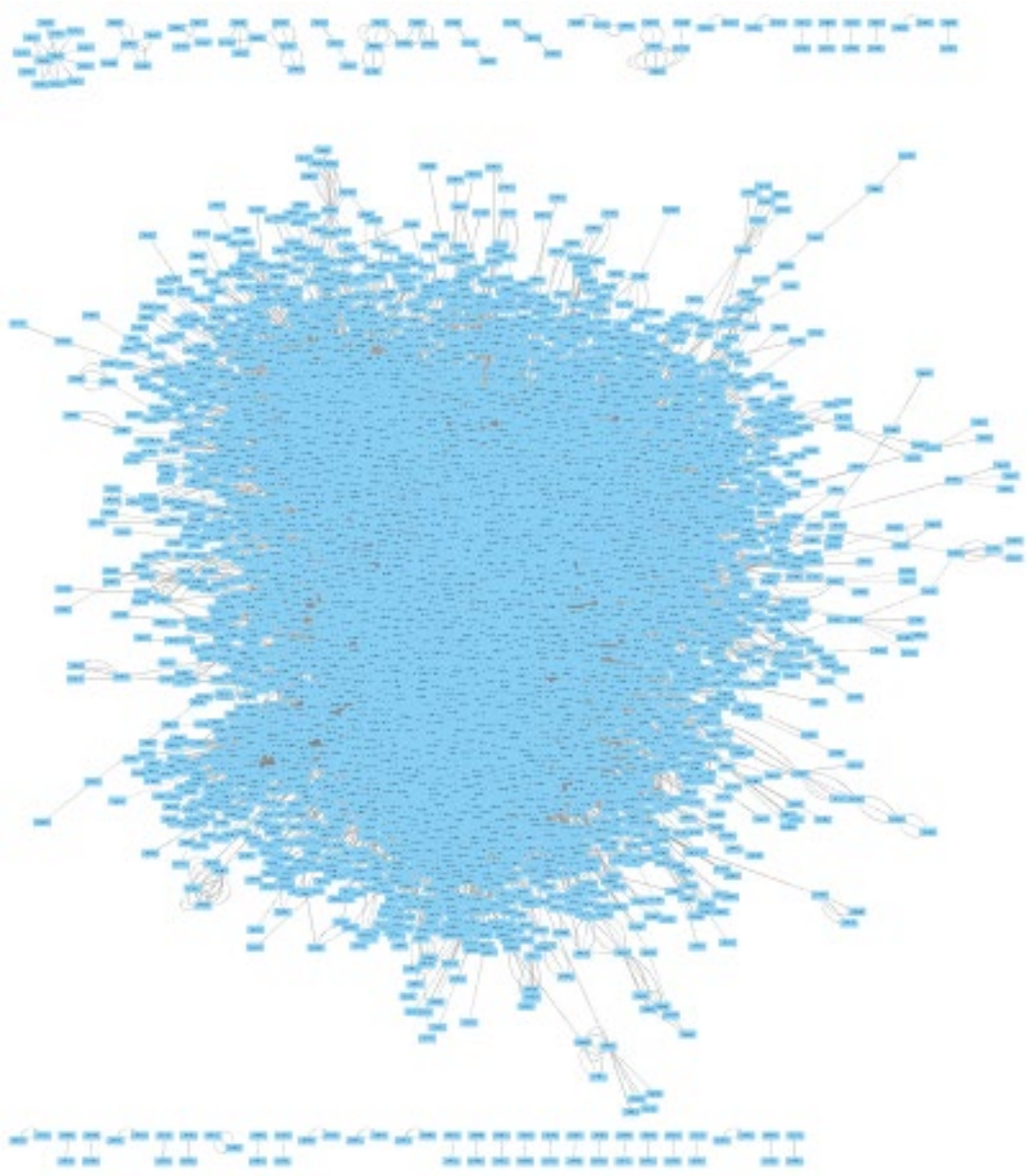


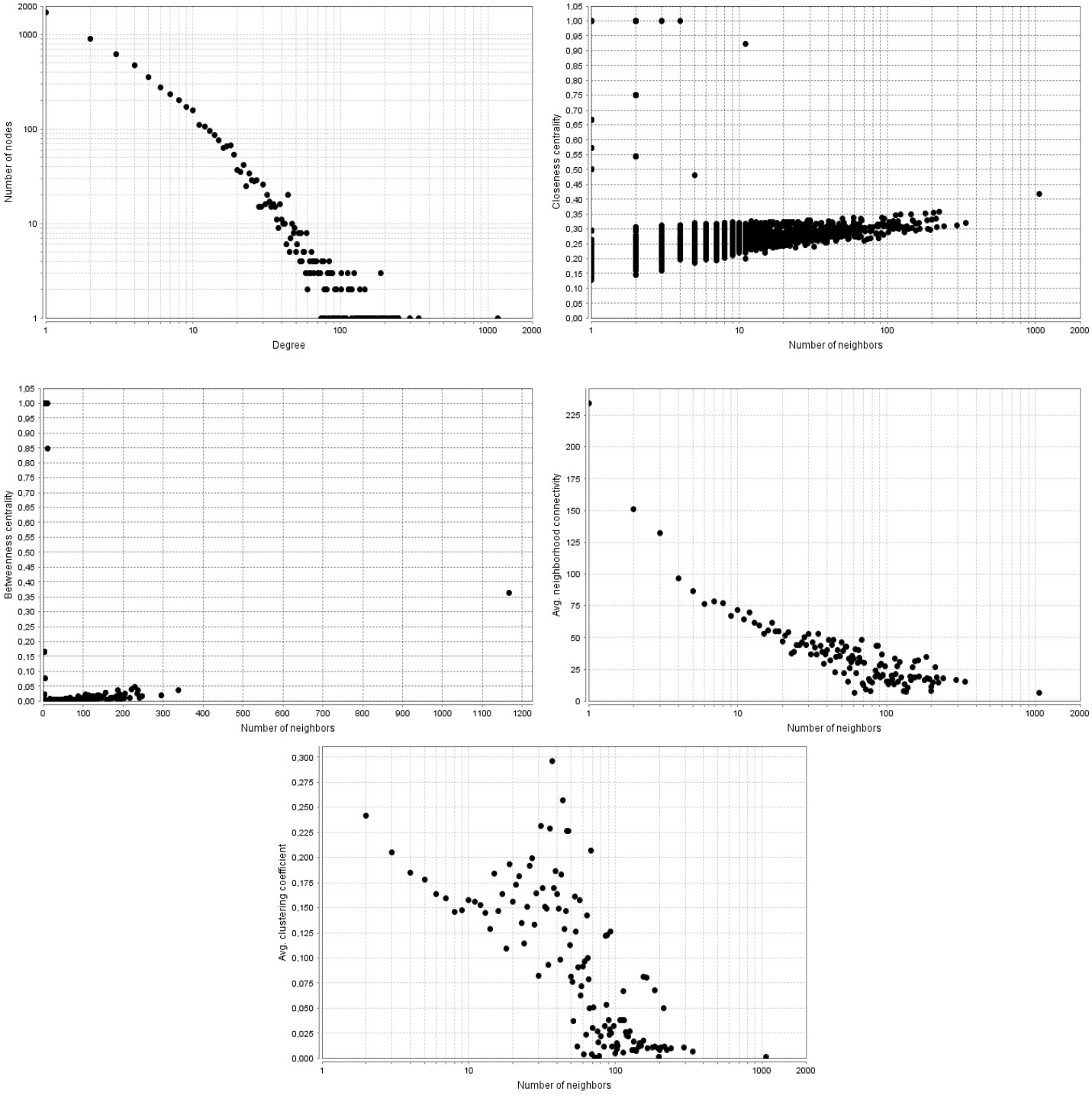
Εικόνα 17: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου Krogan 2006 και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

Τα τρία παραπάνω δίκτυα [Εικόνες 15 εώς 17] αναπαριστούν αλληλεπιδράσεις πρωτεϊνών του οργανισμού Yeast, ενώ αποτελούν και την συνήθη επιλογή συνόλων για την σύγκριση μεθόδων ομαδοποίησης πρωτεϊνικών δικτύων στην βιβλιογραφία. Παρόλο που το σύνολο των Gavin et al. του 2002 περιέχει τις μισές περίπου ακμές, σε σύγκριση με τα υπόλοιπα, και τα τρία σύνολα έχουν υψηλό βαθμό ομοιότητας σχεδόν σε όλες τις μετρικές που εξετάζονται. Η κύρια διαφορά τους συναντάται στην τάση δημιουργίας αυτοτελών ομάδων, όπου στο σύνολο Gavin et al. του 2006 είναι συγκριτικά μειωμένη, όπως δηλώνει η τιμή του Clustering Coefficient, ενώ και η κεντρικότητα βαθμού του είναι διπλάσια των υπολοίπων [Πίνακας 6].

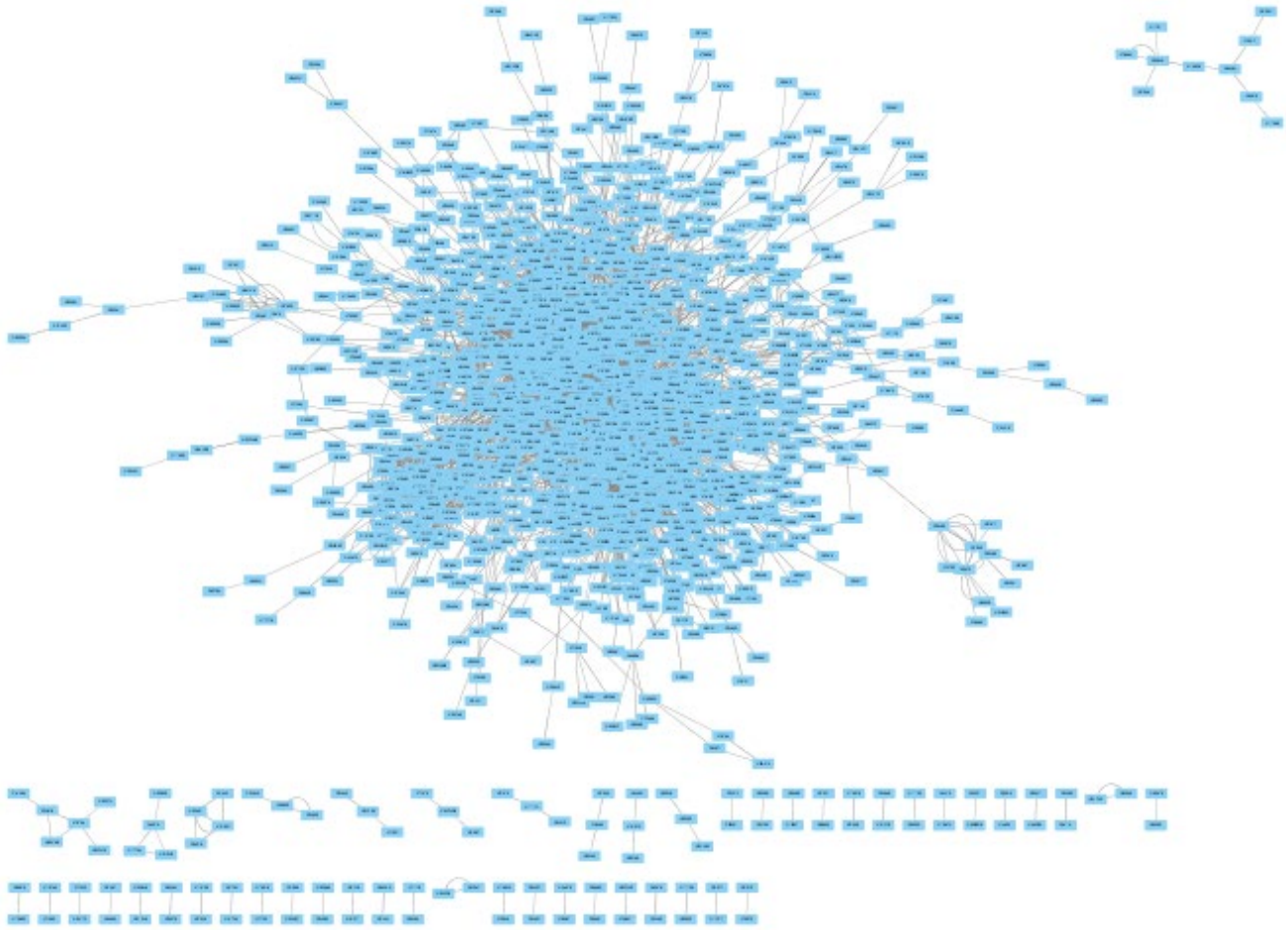
Yeast Datasets	Nodes	Edges	Clustering Coef.	Diameter	Radius	Centralization	Shortest Paths	Char. Path Length	Avg. Degree	Density	Heterogeneity	Self-Loops
Gavin 2002	1352	3210	0.217	12	1	0.035	1539204(84%)	4.928	4.749	0.004	1.229	0
Gavin 2006	1430	6531	0.416	13	1	0.05	1845798(90%)	4.484	9.134	0.006	1.077	0
Krogan 2006	2674	7079	0.19	12	1	0.05	6383448(89%)	4.736	5.292	0.002	1.427	4

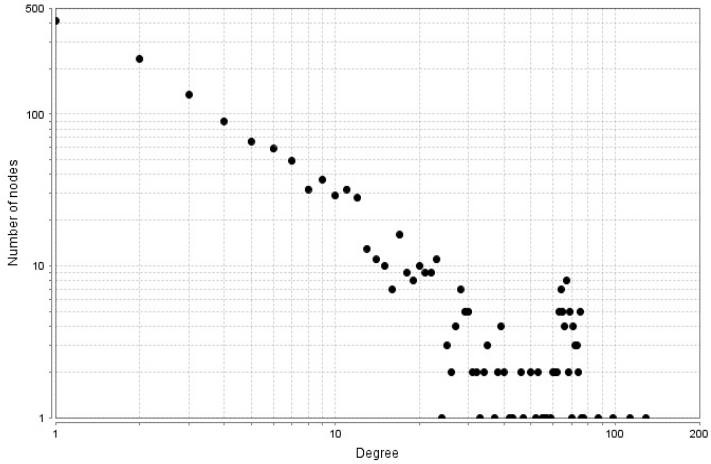
Πίνακας 6: Συγκεντρωτικός πίνακας περιγραφικών χαρακτηριστικών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών του οργανισμού Yeast

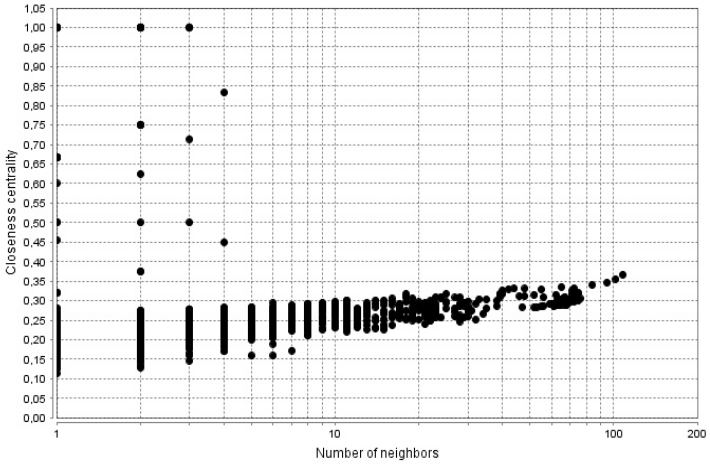


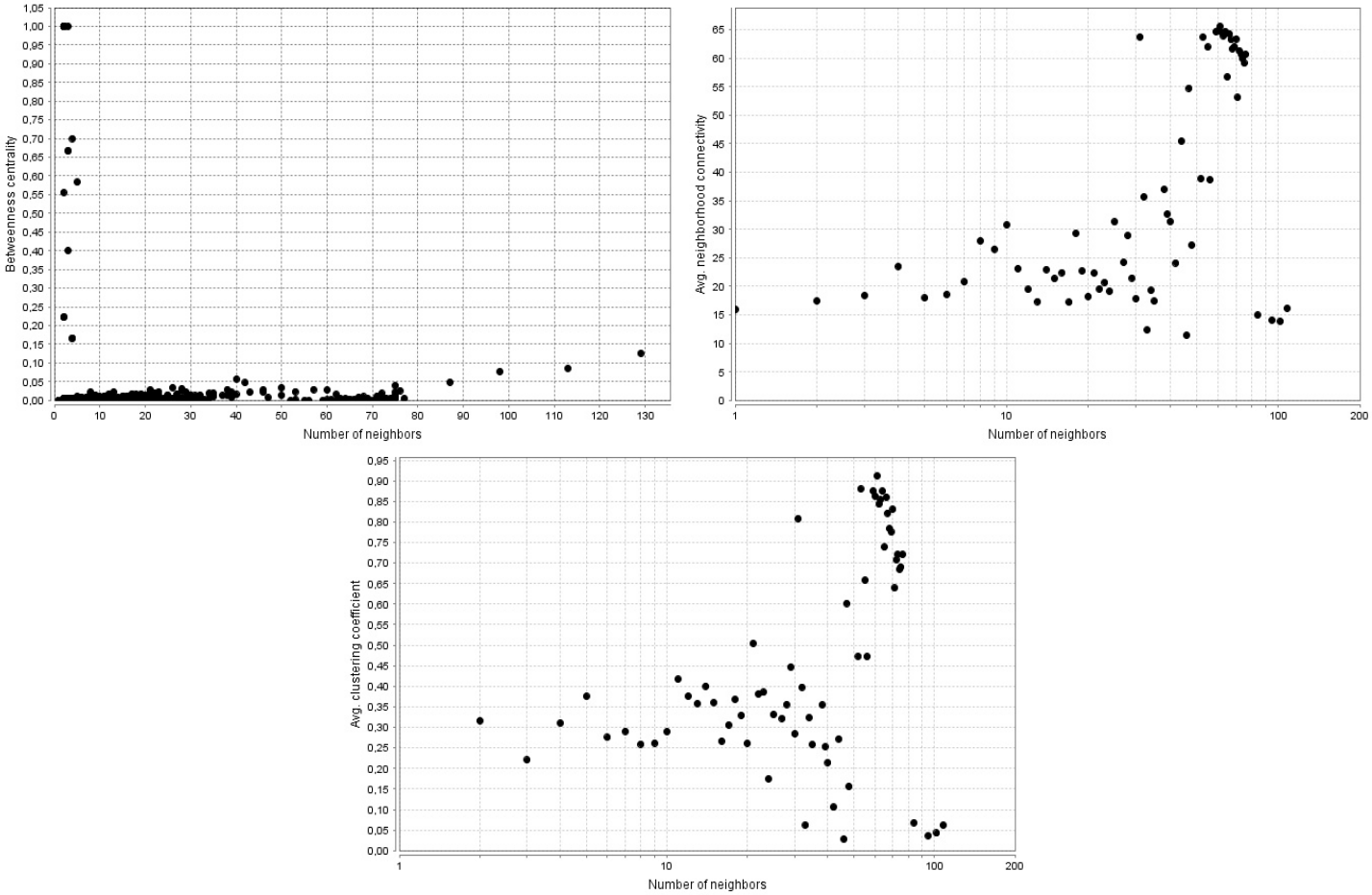


Εικόνα 18: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Arabidopsis_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

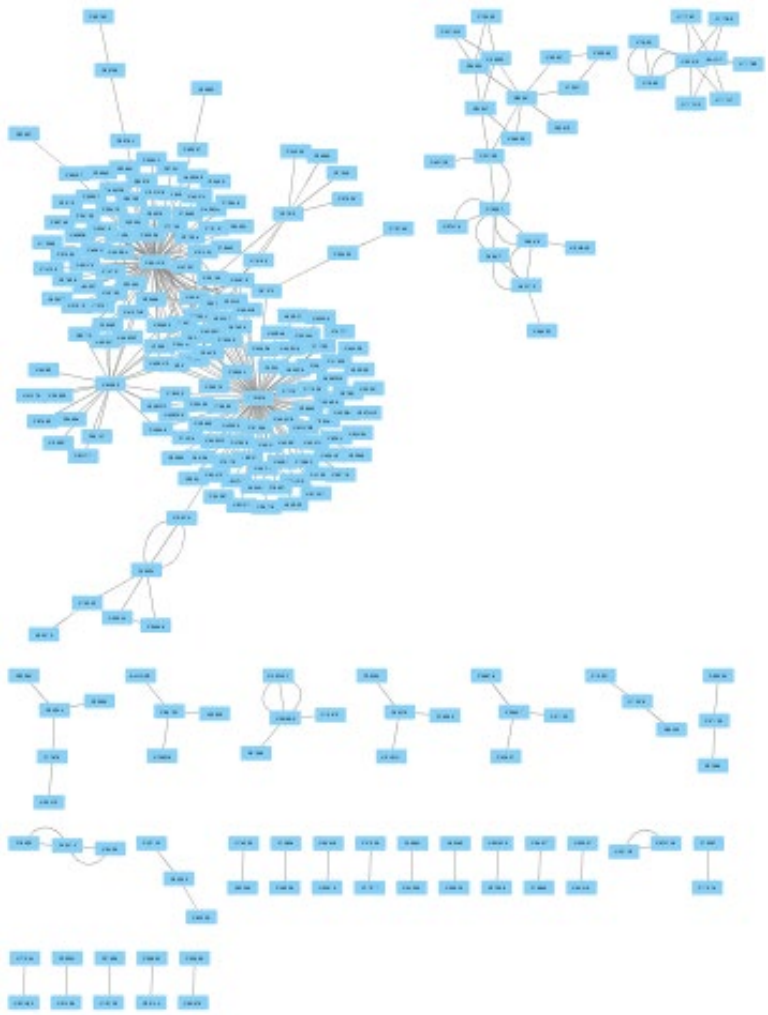


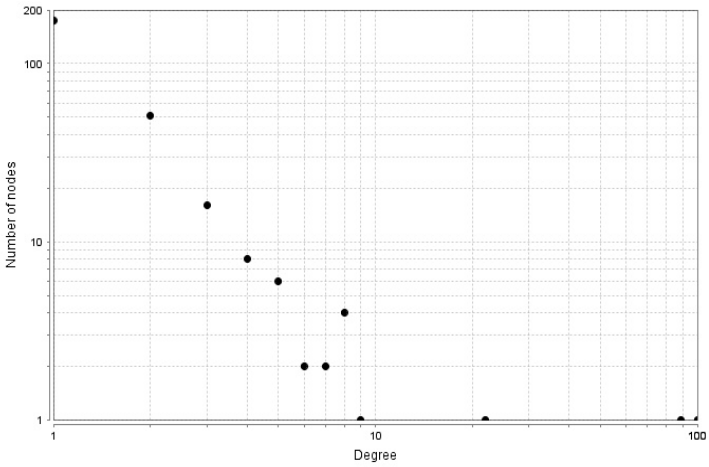

Number of nodes
500
100
10
1
1
10
100
200
Degree

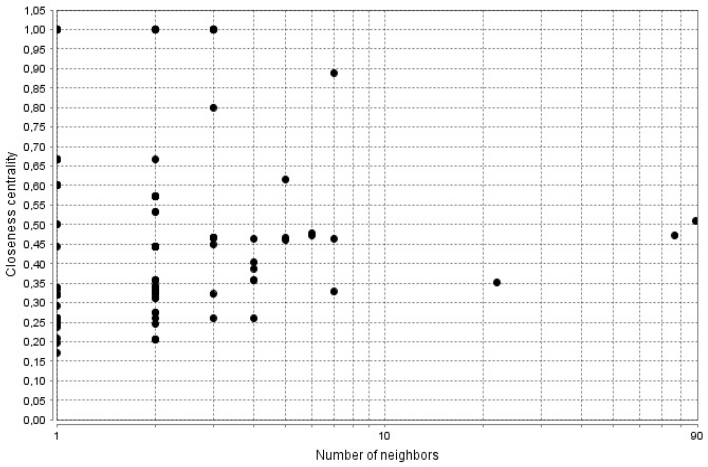

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
200
Number of neighbors

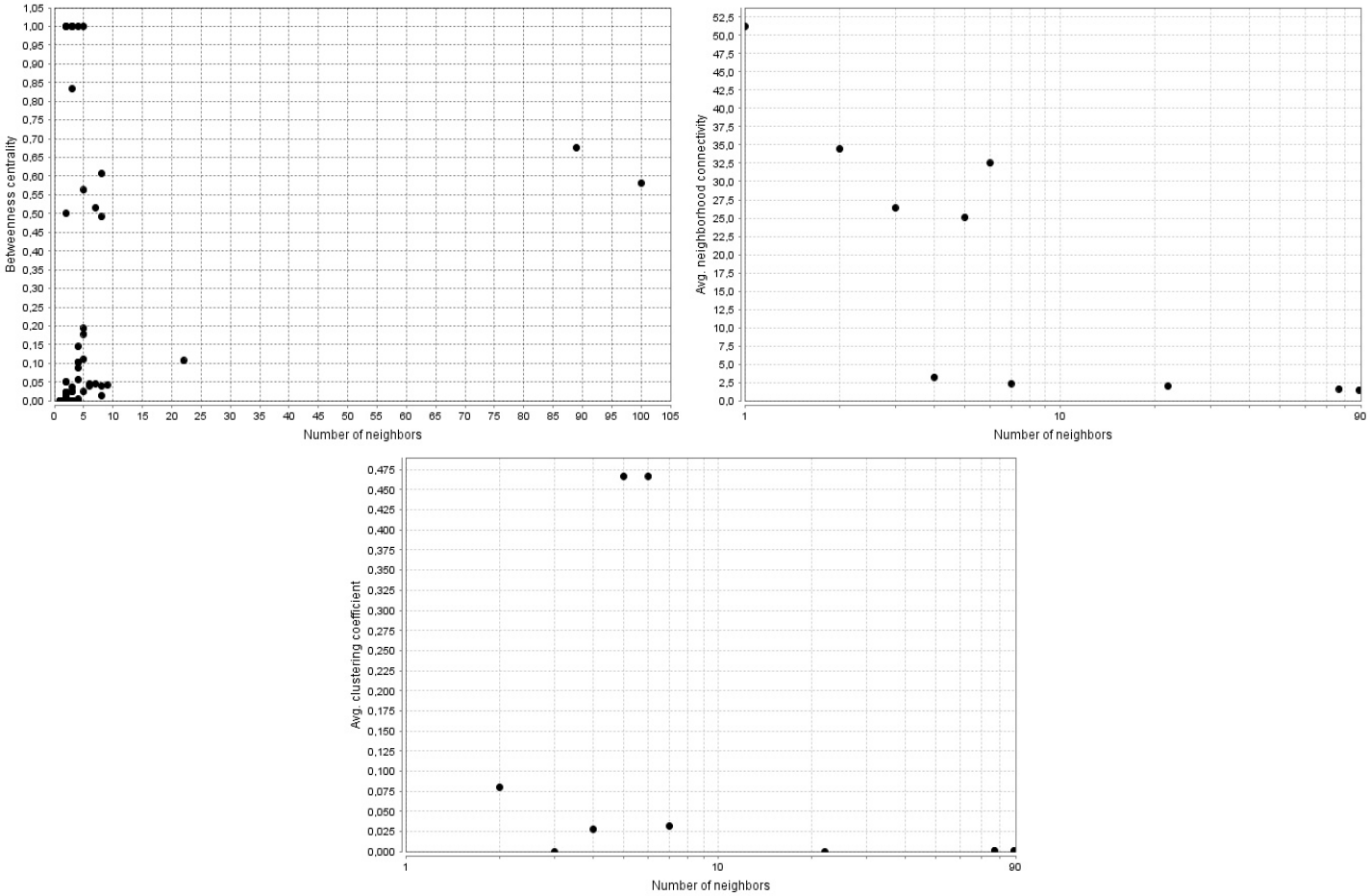


Εικόνα 19: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_C.elegans_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

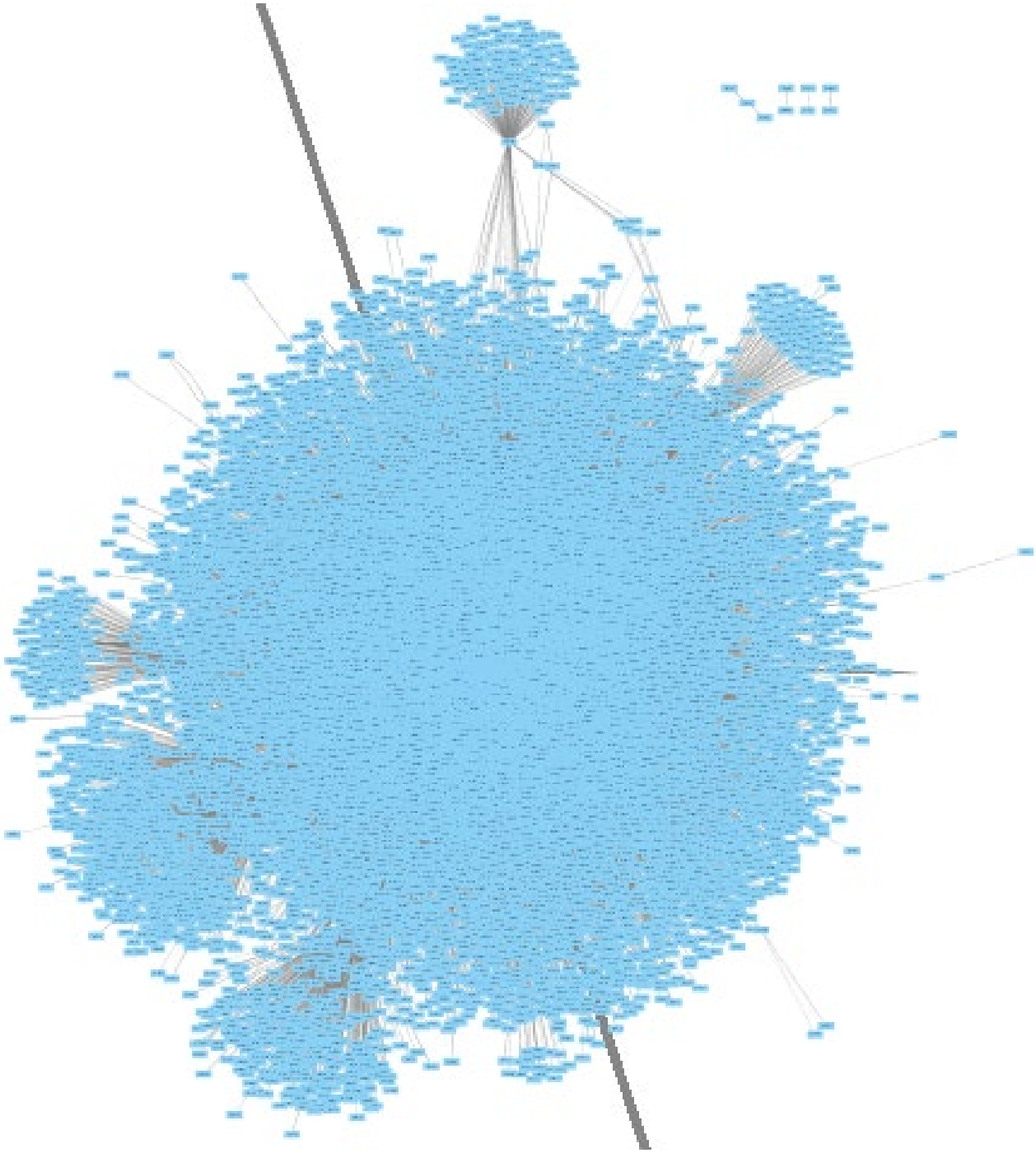


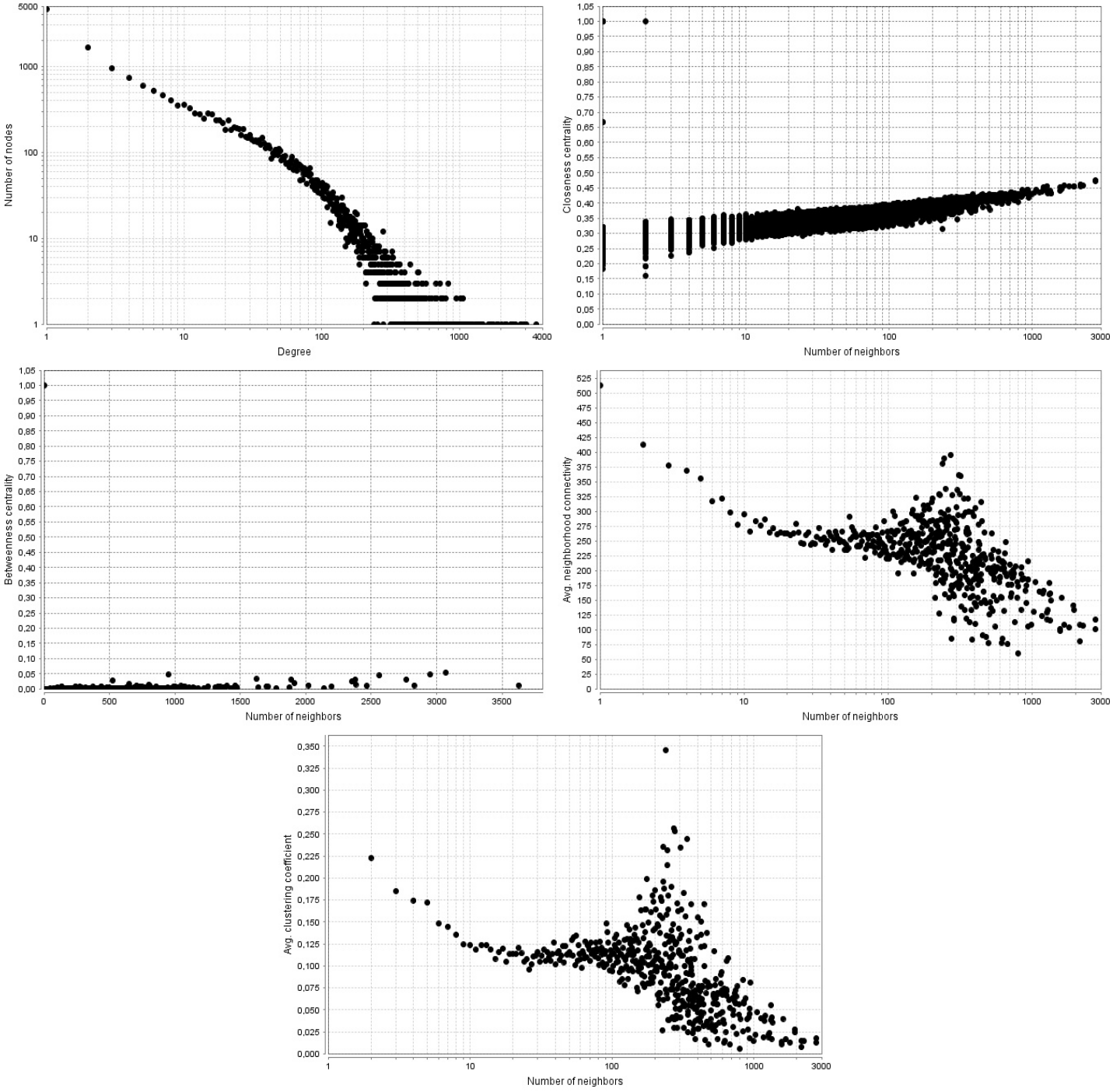

Number of nodes
200
100
10
1
1
10
100
Degree


Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
90
Number of neighbors

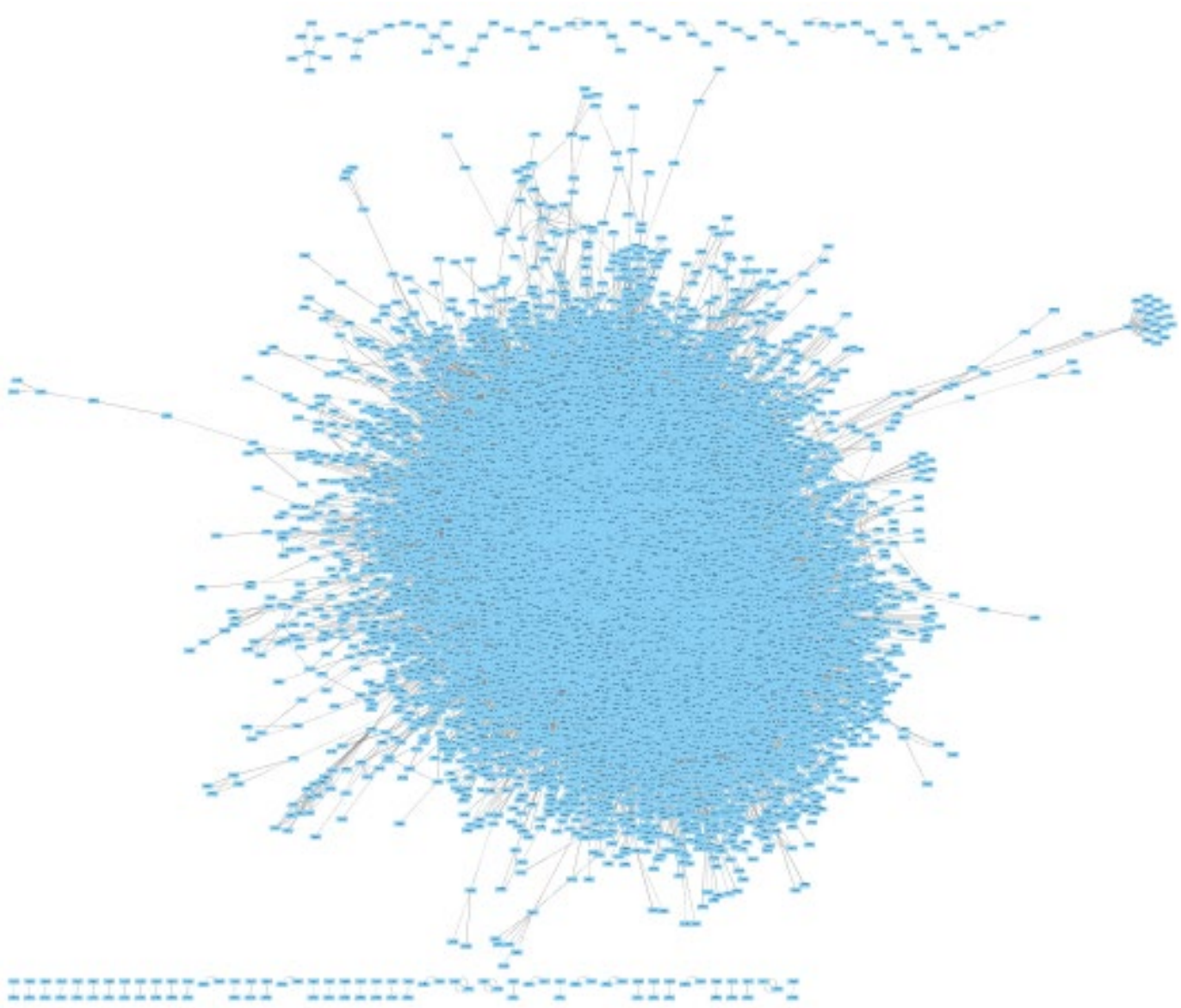


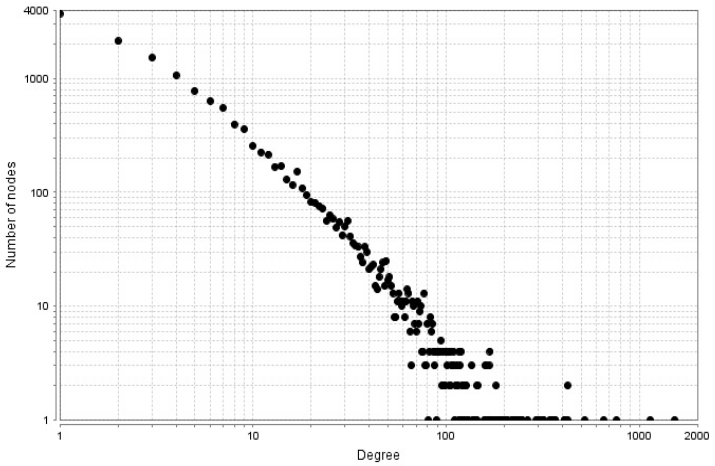
Εικόνα 20: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Chicken_Gallus_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

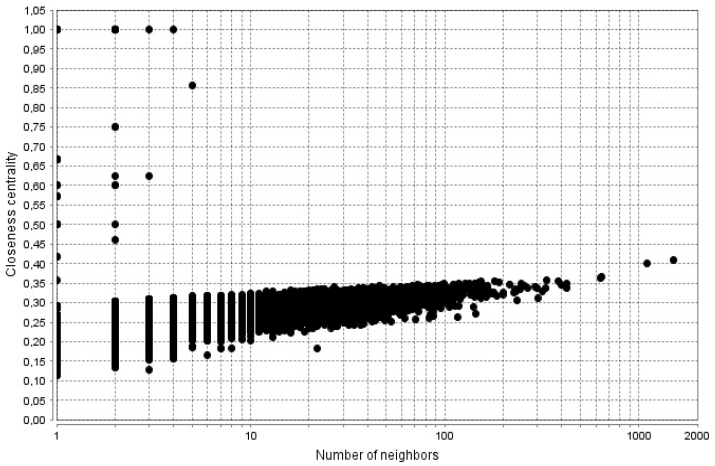


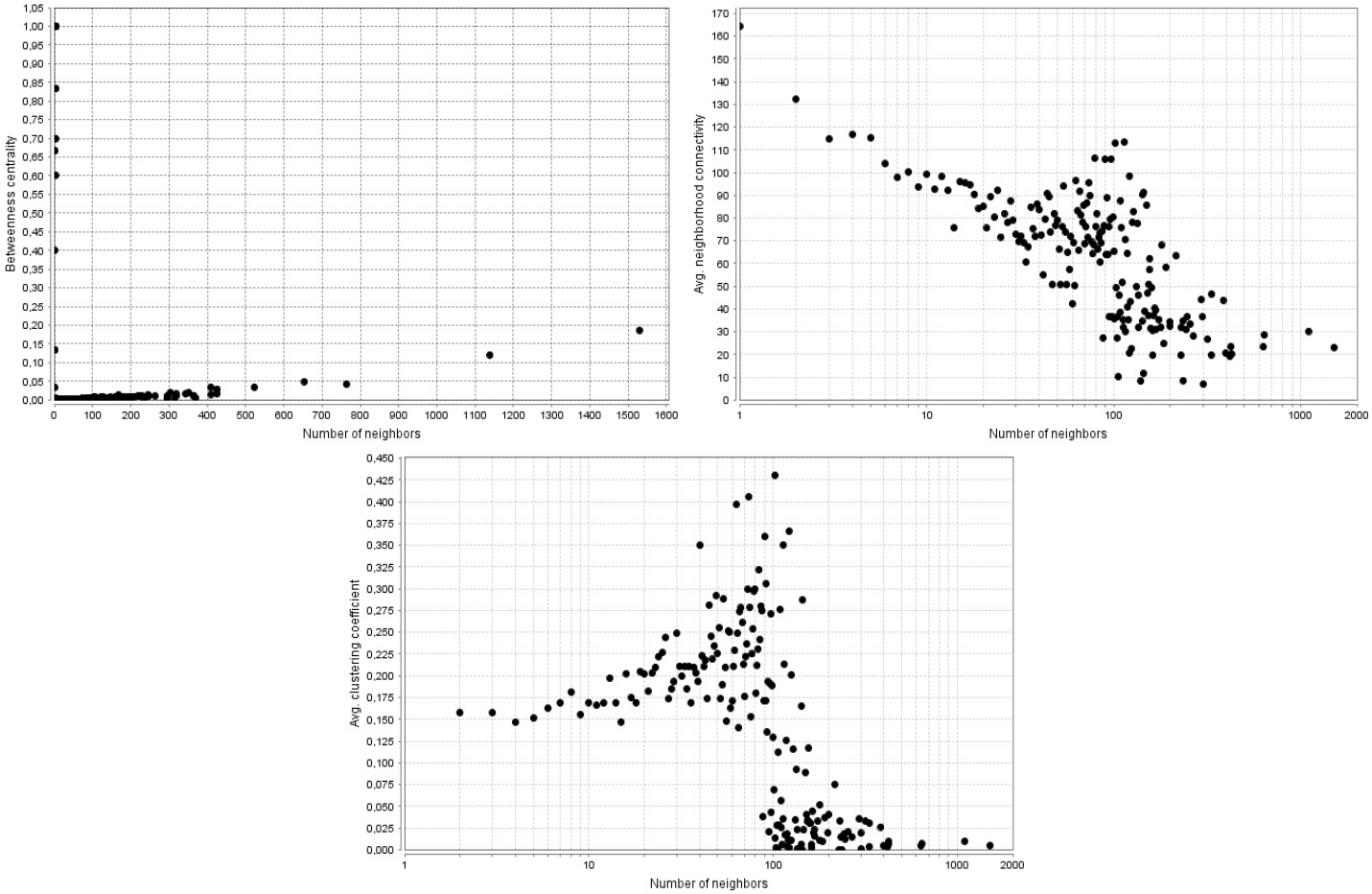


Εικόνα 21: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Homo_Sapiens_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

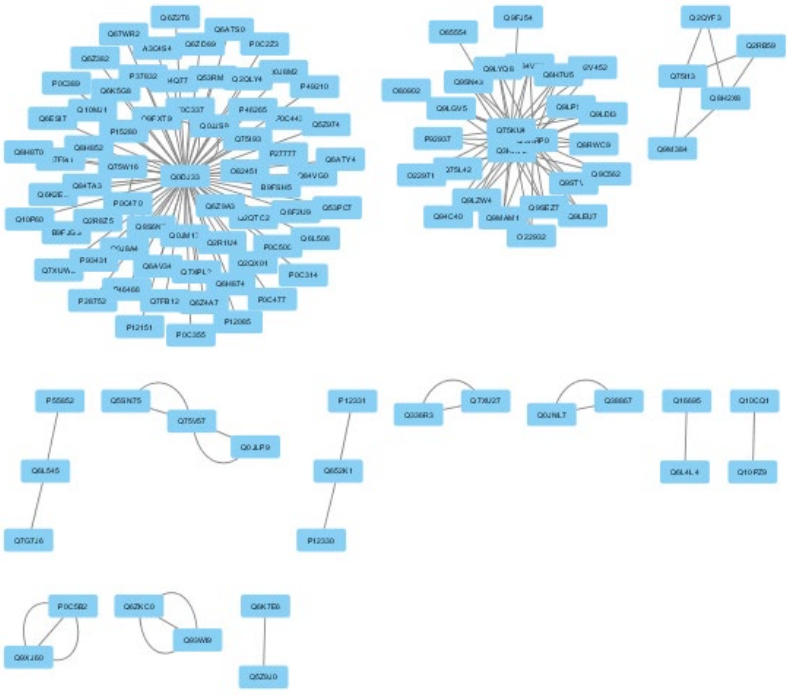


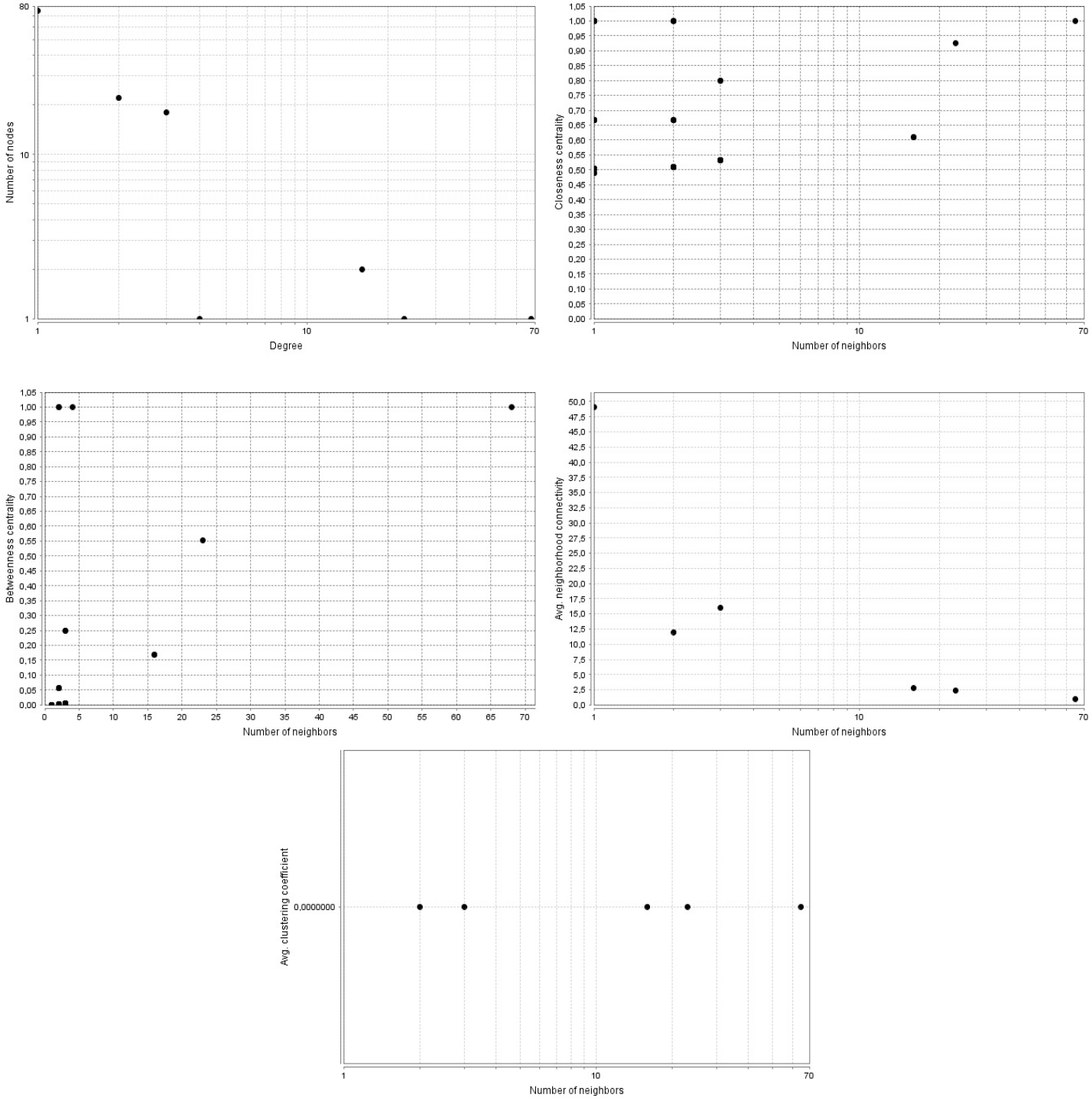

Number of nodes
4000
1000
100
10
1
1
10
100
1000
2000
Degree


Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
1000
2000
Number of neighbors

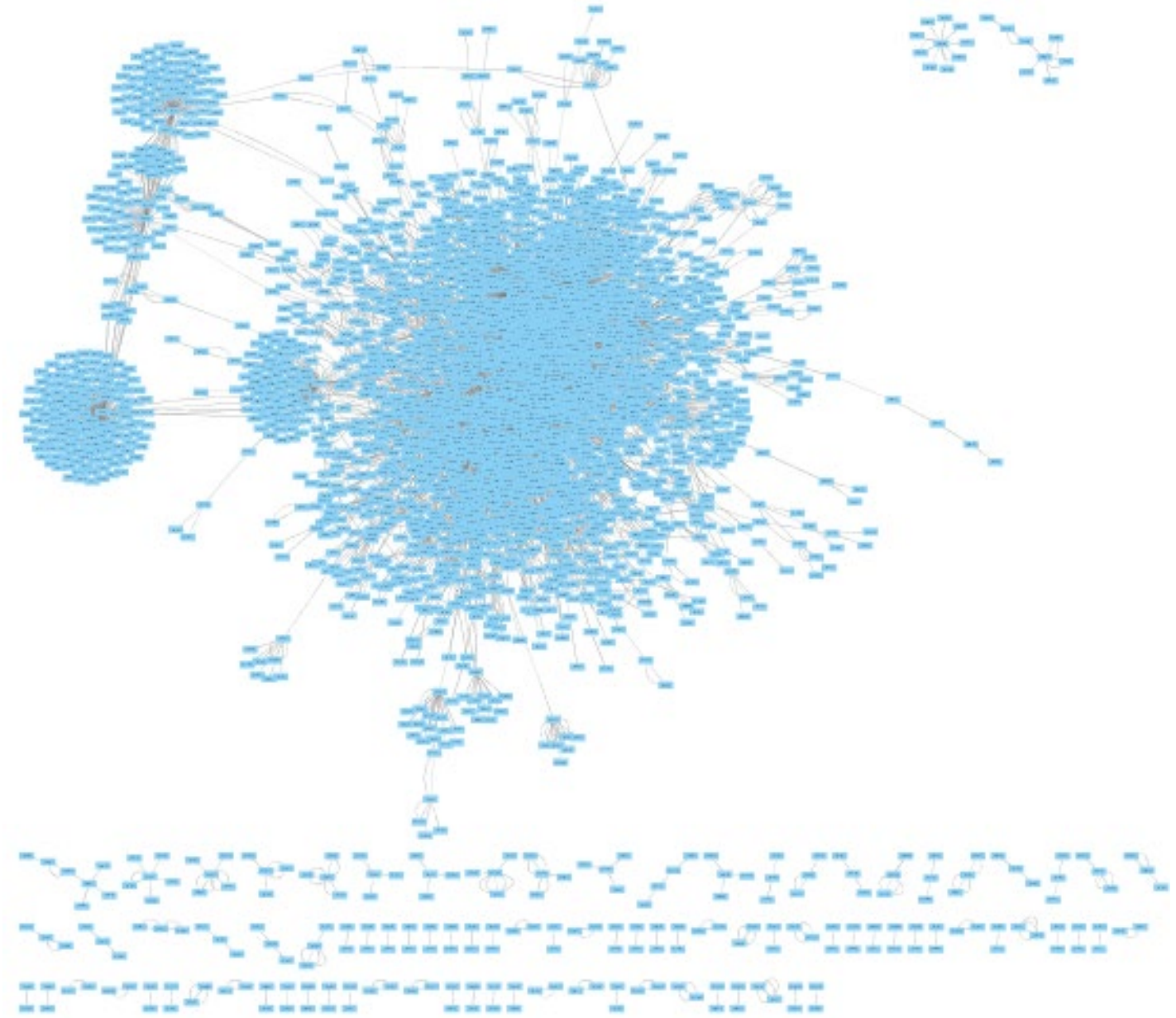


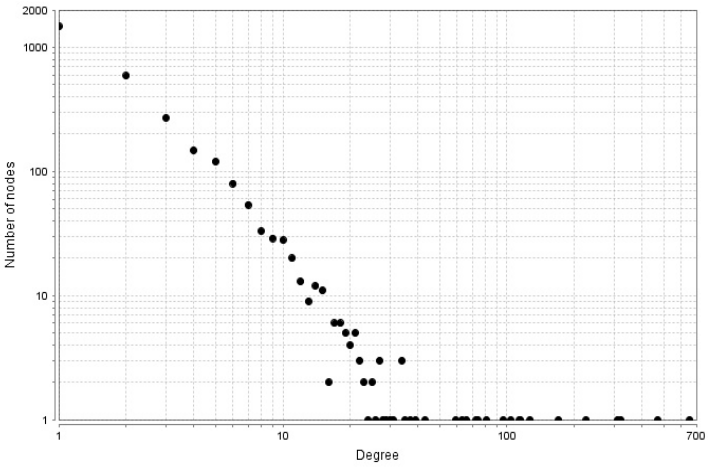
Εικόνα 22: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Mouse_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

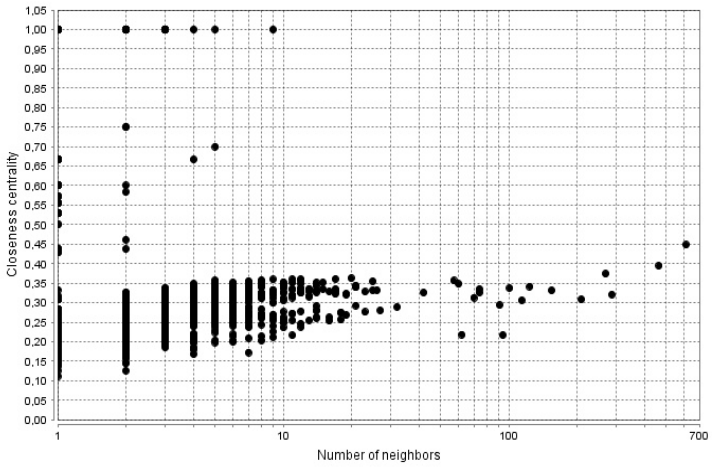


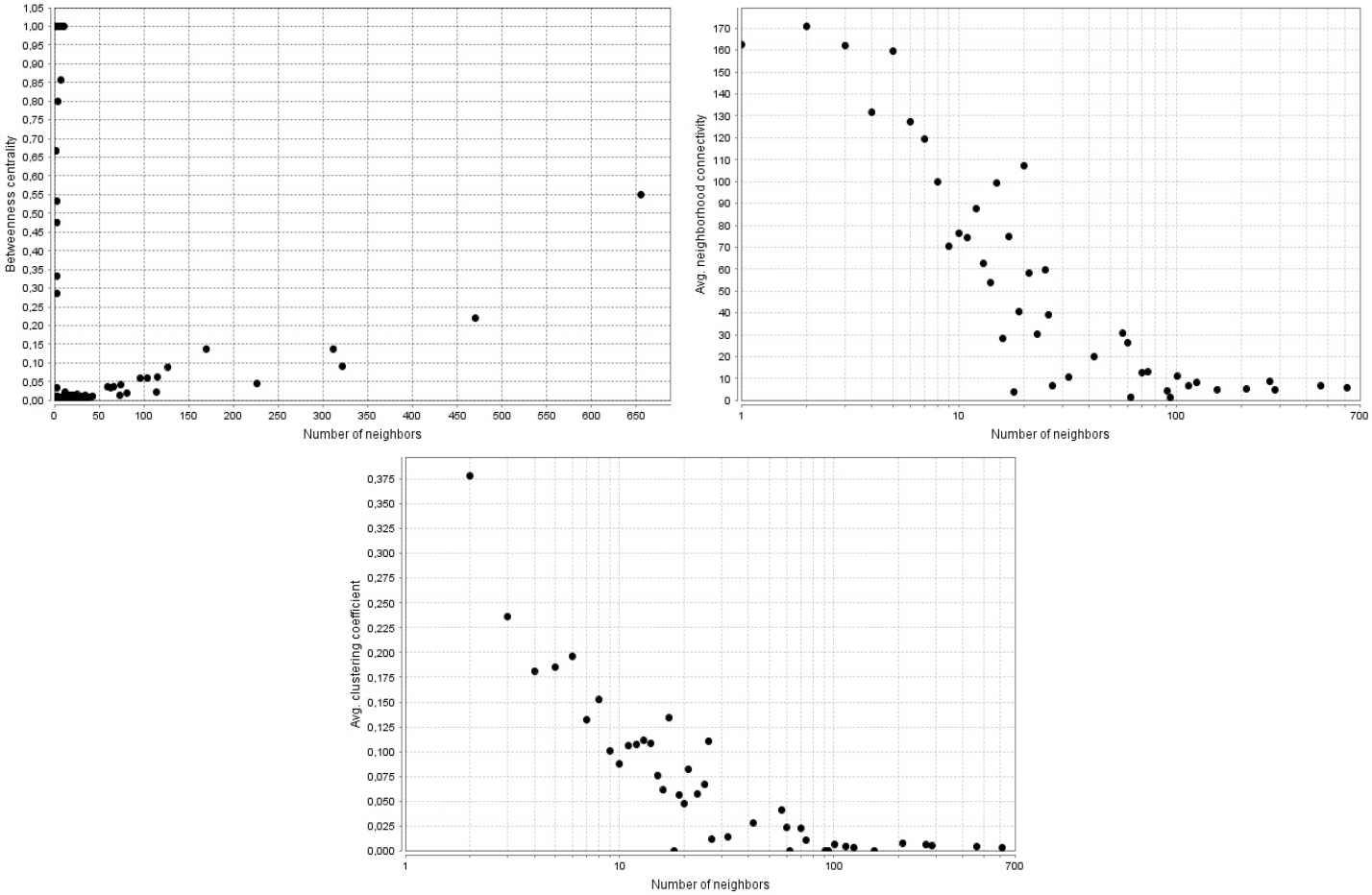


Εικόνα 23: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Oryza_no_self_loopsκαι dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

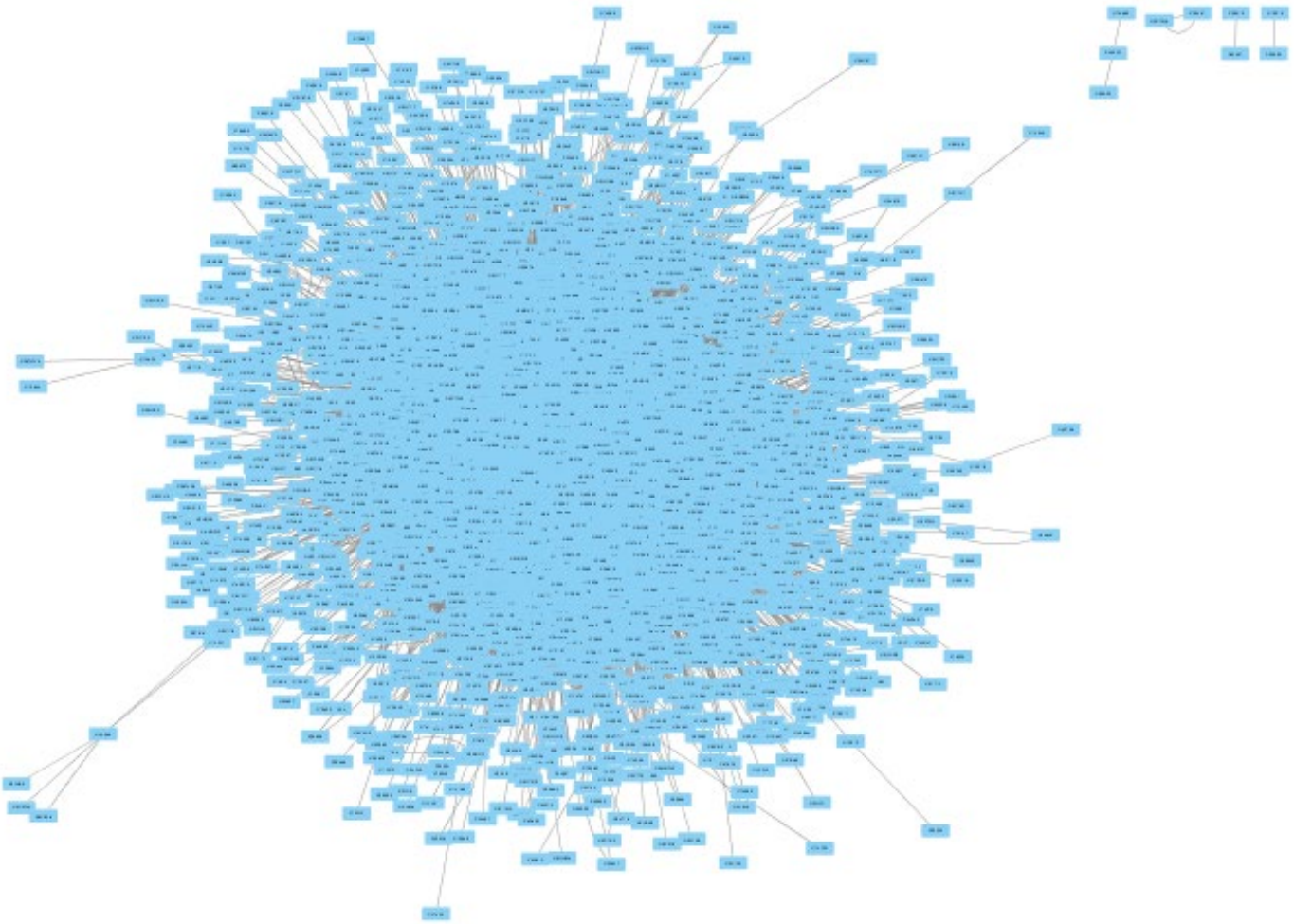


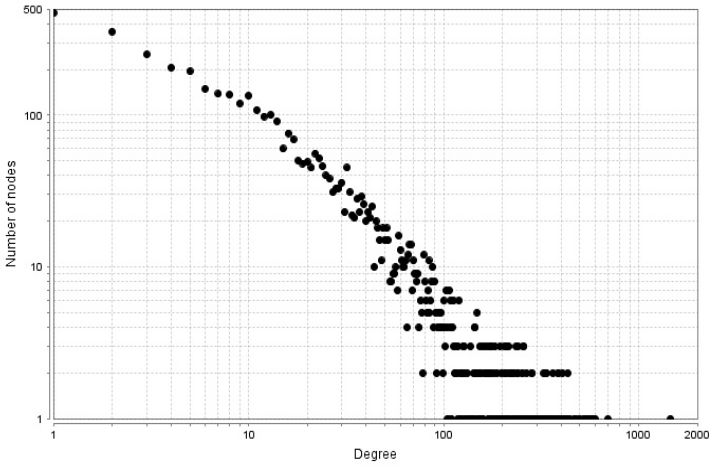

Number of nodes
2000
1000
100
10
1
1
10
100
700
Degree

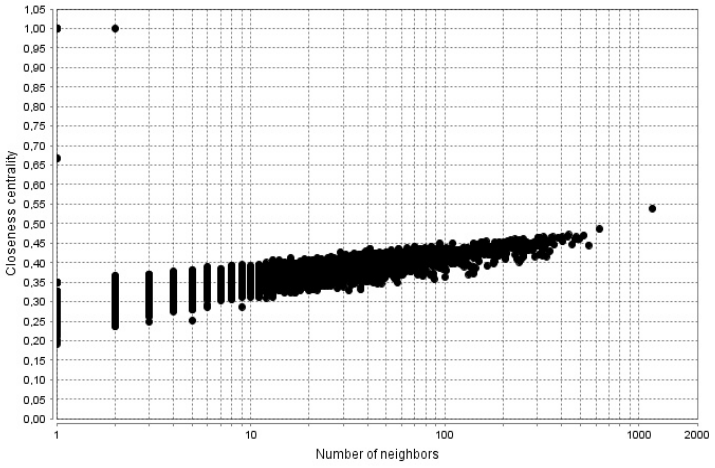

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
700
Number of neighbors

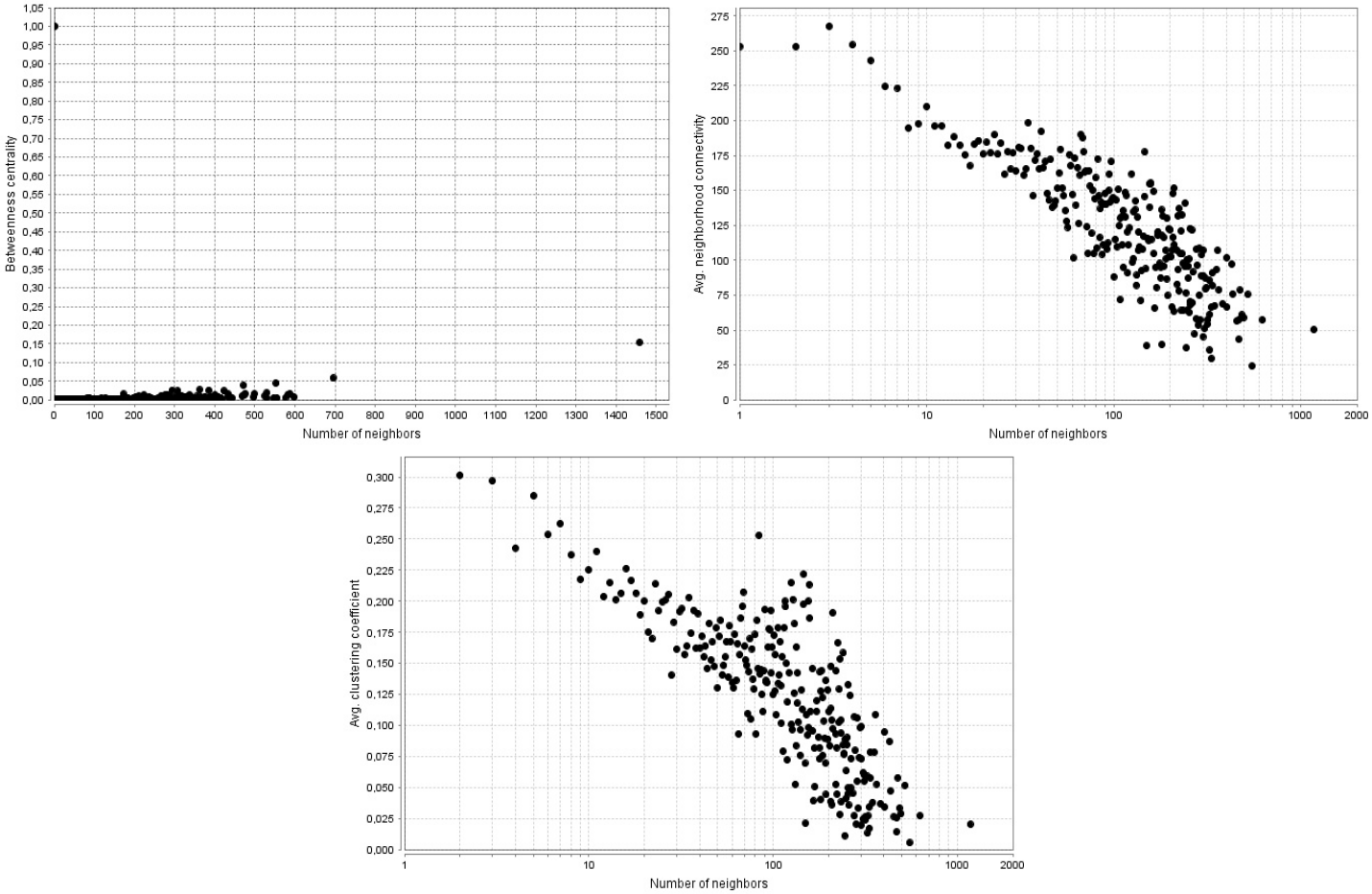


Εικόνα 24: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Rat_no_self_loopsκαι dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

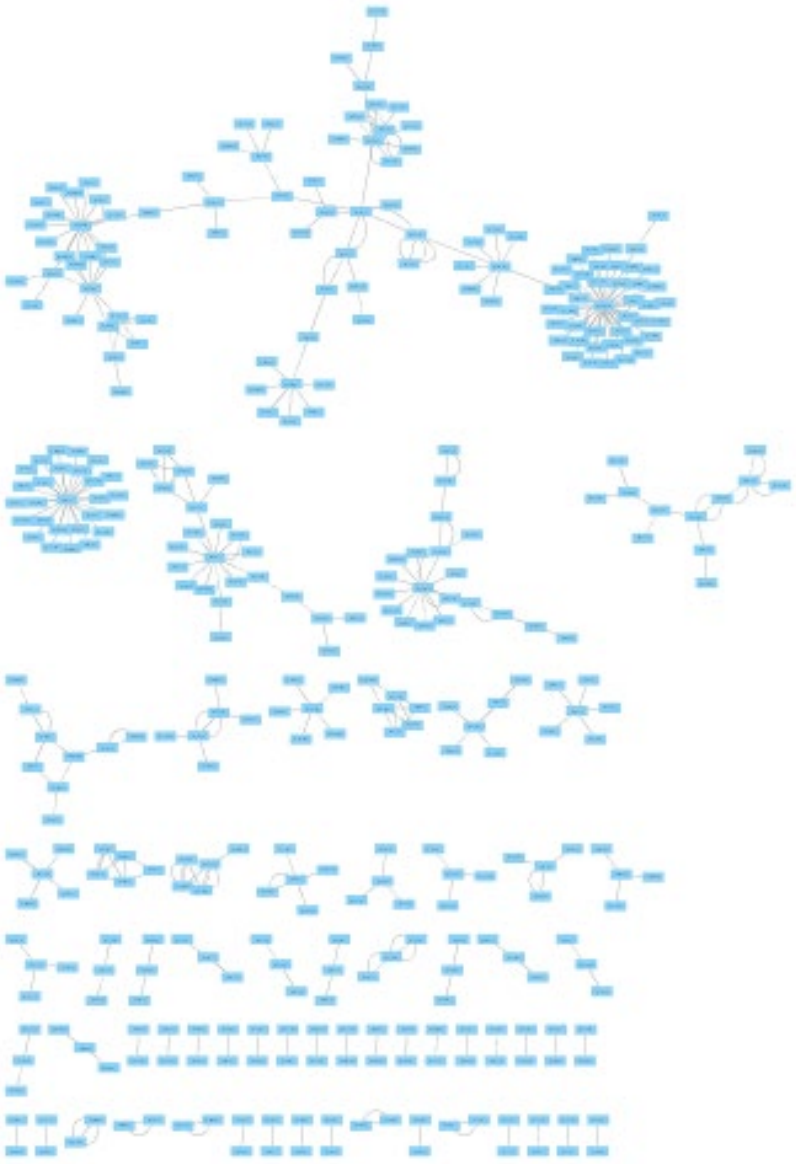


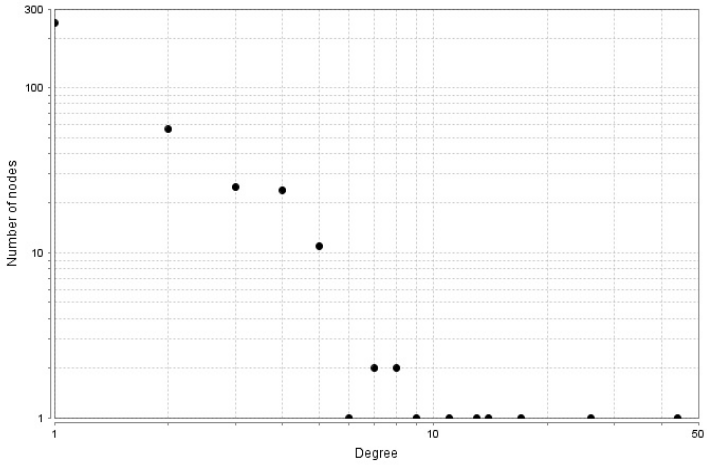

Number of nodes
500
100
10
1
1
10
100
1000
2000
Degree

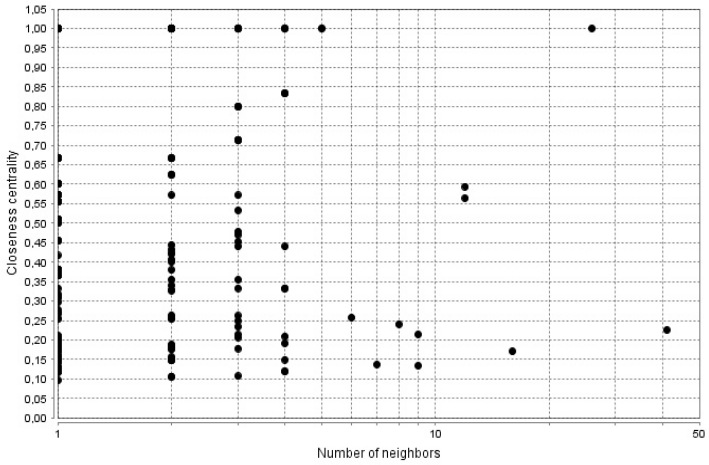

Closeness centrality
1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
1
10
100
1000
2000
Number of neighbors

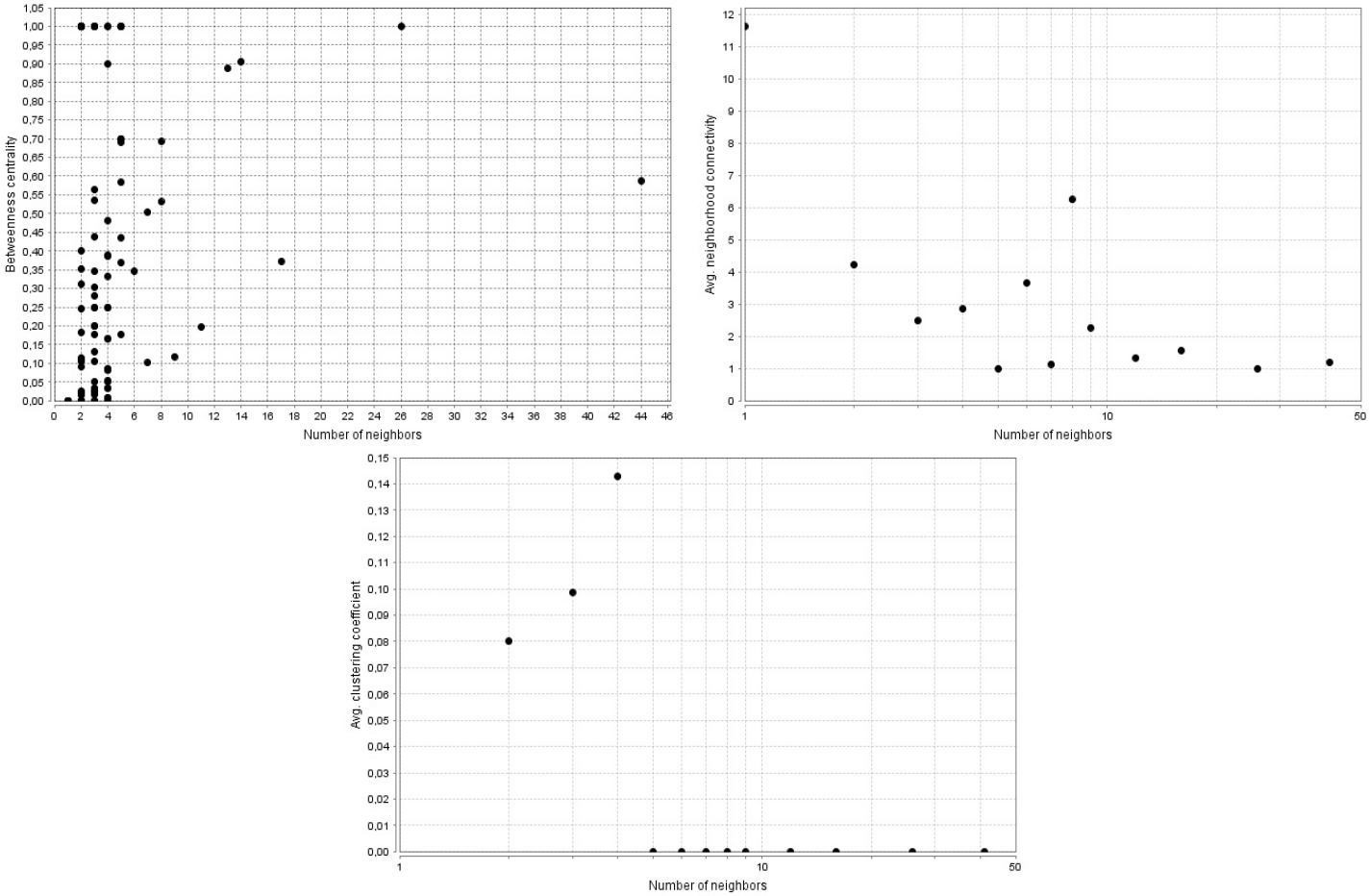


Εικόνα 25: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Spompe_no_self_loopsκαι dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου




300
100
10
1
Number of nodes
1
10
50
Degree


1,05
1,00
0,95
0,90
0,85
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
Closeness centrality
1
10
50
Number of neighbors



Εικόνα 26: Απεικόνιση Δικτύου αλληλεπιδράσεων συνόλου BioGRID_Taurus_no_self_loops και dot plots Κεντρικότητας Βαθμού, Κεντρικότητας Εγγύτητας, Ενδιάμεσης Κεντρικότητας, Συνεκτικότητας Δικτύου, Συντελεστή Ομαδοποίησης Δικτύου

Τέλος τα δίκτυα που προηγήθηκαν [Εικόνες 18 εώς 26] αποτελούν μία ομάδα προεπεξεργασμένων συνόλων της βάσης BioGRID στα οποία, όπως φαίνεται και στον συγκεντρωτικό Πίνακας 7, έχουν αφαιρεθεί σε όλα οι βρόγχοι. Τα σύνολα είναι αρκετά ανομοιογενή, ενώ καλύπτουν και ακραίες περιπτώσεις δικτύων πρωτεϊνών με πολύ μικρά σύνολα, μόλις μερικών εκατοντάδων κόμβων και ακμών, και με σύνολα με περισσότερες από εβδομήντα χιλιάδες συνδέσεις. Όπως είναι αναμενόμενο τα περισσότερα από τα μέτρα περιγραφής των δικτύων που υπολογίστηκαν διαφέρουν ανάμεσα στα σύνολα και ιδιαίτερα ο συντελεστής ομαδοποίησης, η διάμετρος, η ολική κεντρικότητα του δικτύου, ο αριθμός των συντομότερων μονοπατιών, αλλά και το χαρακτηριστικό μήκος αυτών, καθώς και η κεντρικότητα βαθμού των δικτύων. Σε σχετικά όμοια επίπεδα κυμαίνονται η πυκνότητα και η ετερογένεια των δικτύων, ενώ η μετρική της ακτίνας, όπως και σε όλα τα προηγούμενα σύνολα είναι 1. Το γεγονός πως η ακτίνα παρέμεινε ίδια σε ένα πλήθος τόσο διαφορετικών δικτύων, υποδεικνύει ότι ως μέτρο δεν προσφέρει σημαντική πληροφορία στην ανάλυση δικτύων αλληλεπιδράσεων πρωτεϊνών.

Συνολικά τα στοιχεία, που κατεγράφησαν από την επεξεργασία και ανάλυση των **23 συνόλων δεδομένων**, αποδεικνύουν πως έχει ληφθεί υπόψην ένας μεγάλος αριθμός πιθανών περιπτώσεων, ως προς **12 διαφορετικές μετρικές γράφων** και επομένως η εκτέλεση των αλγορίθμων βάση αυτών μπορεί να προσφέρει αντικειμενικά αποτελέσματα.

Pre-processed Datasets	Nodes	Edges	Clustering Coef.	Diameter	Radius	Centralization	Shortest Paths	Char. Path Length	Avg. Degree	Density	Heterogeneity	Self-Loops
BioGRID_Arabidopsis_no_self_loops	6580	32265	0.13	11	1	0.16	41725458(96%)	4.015	7.991	0.001	2.624	0
BioGRID_C.elegans_no_self_loops	1435	6502	0.236	13	1	0.069	1725544(83%)	4.354	8.662	0.006	1.783	0
BioGRID_Chicken_Gallus_no_self_loops	267	332	0.025	9	1	0.329	30332(42%)	3.049	2.187	0.008	3.304	0
BioGRID_Homo_Sapiens_no_self_loops	23259	583505	0.106	9	1	0.117	540539262(99%)	3.138	37.662	0.002	2.327	0
BioGRID_Mouse_no_self_loops	14497	72968	0.121	13	1	0.103	206080644(98%)	3.997	9.152	0.001	2.842	0
BioGRID_Oryza_no_self_loops	120	150	0	3	1	0.536	4992(34%)	1.95	2.317	0.019	2.756	0
BioGRID_Rat_no_self_loops	2956	5729	0.115	13	1	0.206	7363646(84%)	4.029	3.283	0.001	5.316	0
BioGRID_Spompe_no_self_loops	4500	77479	0.193	8	1	0.255	20164602(99%)	2.859	27.748	0.006	1.999	0
BioGRID_Taurus_no_self_loops	374	377	0.023	14	1	0.106	15552(11%)	5.543	1.775	0.005	1.601	0

Πίνακας 7: Συγκεντρωτικός πίνακας περιγραφικών χαρακτηριστικών των προεπεξεργασμένων δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών

3.1.2. Εκτέλεση αλγορίθμων και εξαγωγή μετρικών σύγκρισης των αποτελεσμάτων

Τα σύνολα δεδομένων, τα οποία περιγράφηκαν στο προηγούμενο υποκεφάλαιο, αποτέλεσαν την είσοδο στους 6 αλγορίθμους, που επιλέχθηκαν να συγκριθούν στην παρούσα διπλωματική. Μόλις ολοκληρωνόταν η εκτέλεση του κάθε αλγορίθμου για το σύνολο το δικτύων, τα αποτελέσματα ελέγχονταν και ανάλογα με τις παραμέτρους που επέτρεπε κάθε αλγόριθμος να ρυθμίσει ο χρήστης, γινόταν τροποποίηση μίας παραμέτρου κάθε φορά και επανεκτέλεση. Για κάθε αλγόριθμο ο συγκεκριμένος βρόγχος επανάληψης έγινε τέτοιο αριθμό φορών, ώστε να μπορεί να ληφθεί μια σαφή εικόνα για τα αποτελέσματα τους, αλλά και παράλληλα τα αποτελέσματα που λαμβάνονταν κάθε φορά από συγκεκριμένο αλγόριθμο και σύνολο δεδομένων να μην είναι ταυτόσημα. Ρόλο επίσης διαδραμάτισαν και οι υπολογιστικές απαιτήσεις της εκτέλεσης συγκεκριμένων συνόλων δεδομένων για κάποιες παραμέτρους αλγορίθμων. Συνολικά εξάχθηκαν **623 αποτελέσματα** από τους αλγορίθμους ομαδοποίησης.

Η σύγκριση των αποτελεσμάτων των αλγορίθμων έγινε με τις μετρικές που αναλύθηκαν στο κεφάλαιο 2.3. Για τον υπολογισμό των επτά εξ αυτών δημιουργήθηκε πρόγραμμα στην γλώσσα προγραμματισμού Python. Το πρόγραμμα δέχεται ως είσοδο αρχεία που περιέχουν σύμπλοκα πρωτεϊνών, όπως αυτά προβλέφθηκαν από τους αλγορίθμους στο προηγούμενο στάδιο και δίνει ως έξοδο στον χρήστη τις τιμές των Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value. Οι συναρτήσεις υπολογισμού αυτών στηρίζονται στις εξισώσεις που περιγράφηκαν στην παρούσα εργασία, ενώ για τις μετρικές Rand Index και Variation of Information συγκεκριμένα, έγινε χρήση της βιβλιοθήκης Scikit-learn[113]. Αξίζει να σημειωθεί το γεγονός ότι παρόλο που τα αρχεία εξόδου από τον κάθε αλγόριθμο ομαδοποίησης διαφέρουν σε σημαντικό βαθμό, μέσω του συγκεκριμένου προγράμματος σύγκρισης, η τροποποίηση των αρχείων από τον χρήστη πριν την είσοδο τους είναι ελάχιστη. Επιπλέον, ο υπολογισμό της μετρικής Normalized Mutual Information πραγματοποιήθηκε μέσω προγράμματος σε C++, το οποίο έχουν υλοποιήσει οι McDaid, Greene και Hurley. Για την εύρεση όλων των πιθανών συγκρίσεων τα προγράμματα υπολόγισαν συνολικά **127.211 στατιστικές τιμές,** καθώς το πρόγραμμα NMI χρειάστηκε να εκτελεστεί $\frac{n(n-1)}{2}$, αφού η μετρική είναι συμμετρική, ενώ καθώς η πλειοψηφία των τιμών που υπολογίζει το άλλο πρόγραμμα είναι μη συμμετρικές, χρειάστηκαν n*n εκτελέσεις, όπου n ο αριθμός των αποτελεσμάτων για κάθε αλγόριθμο.

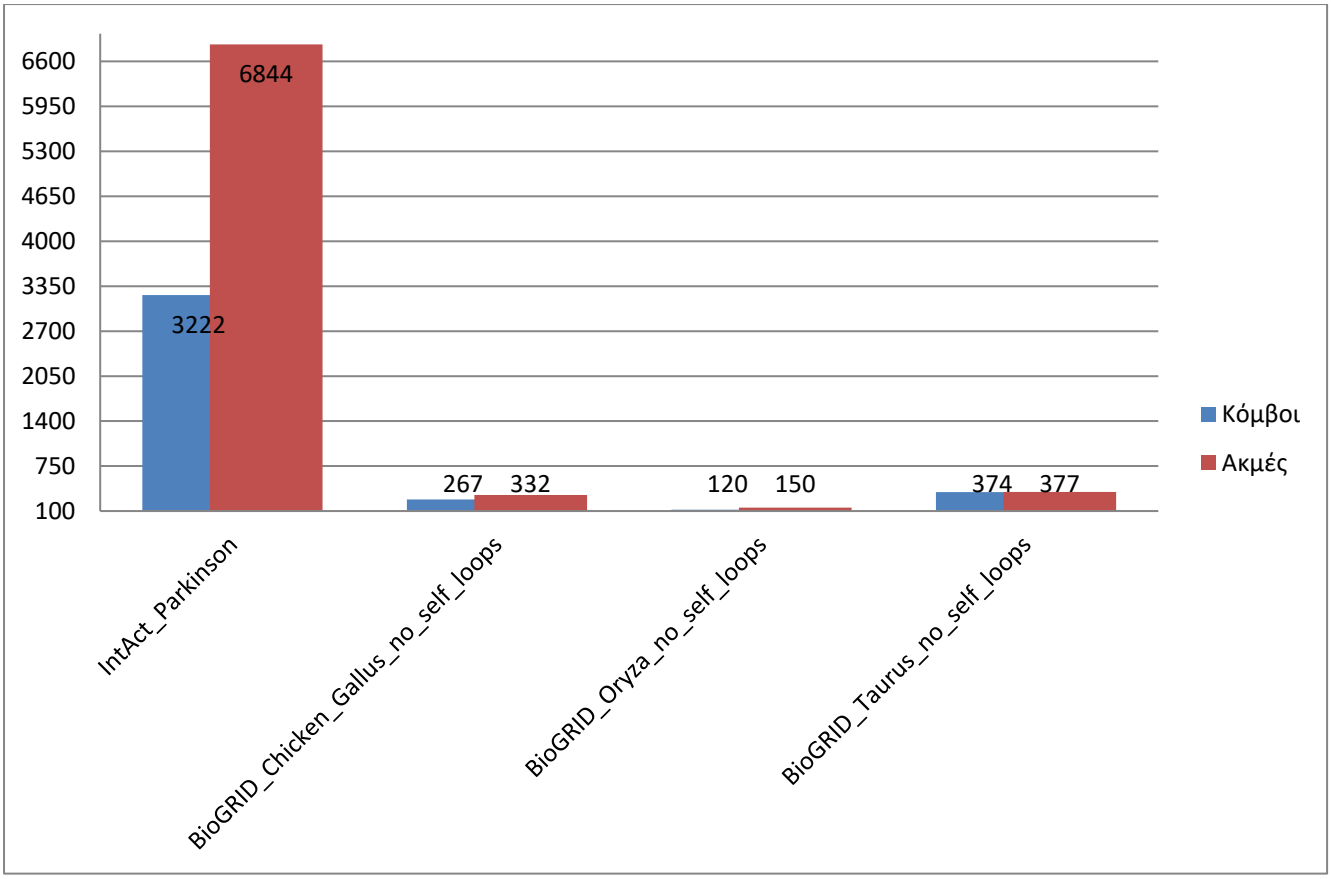
Έχοντας συγκεντρώσει το πλήθος αυτό των στατιστικών τιμών, για κάθε ζεύγος αλγορίθμων ελέχθηκε αν και με ποιες παραμέτρους, σύμφωνα με τις μετρικές, φαινόταν να υπάρχει σύγκλιση αποτελεσμάτων. Καθώς τα μέτρα που ελέγχθηκαν

ήταν αρκετά και σε πολλές συγκρίσεις δεν απέδιδε η ίδια παράμετρος αλγορίθμου την καλύτερη τιμή για όλα, ακολουθήθηκαν δύο κύρια βήματα για την επιλογή των βέλτιστων ζευγαριών. Αρχική προϋπόθεση για την εξέταση του κάθε ζεύγους ήταν οι τιμές να ξεπερνάνε ορισμένα κατώφλια, συγκεκριμένα: ARI ≥ 0.55 ή F1-score ≥ 0.75 ή VOI ≤ 0.35 ή Acc ≥ 0.75 ή Sn ≥ 0.75 ή NMI ≥ 0.55. Η επιλογή των συγκεκριμένων ορίων έγινε εμπειρικά, κατόπιν αρκετών δοκιμών, ώστε να εξασφαλιστεί πως τα ζευγάρια προς εξέταση θα παρουσίαζαν σε τουλάχιστον μία μετρική αξιόλογα αποτελέσματα. Όσα ζεύγη αλγορίθμων πληρούσαν το πρώτο κριτήριο συγκρίνονταν συνολικά, για όλες τις τιμές, ποια παράμετρος εμφάνιζε καλύτερα αποτελέσματα, δίνοντας μεγαλύτερη έμφαση στις παραμέτρους Rand Index, Variation of Information και Normalized Mutual Information, αφού θεωρούνται ως πιο αυστηρά κριτήρια[108,114–119].

Από τον παραπάνω έλεγχο προέκυψαν **126 ζεύγη** σε ένα σύνολο **14 συνδυασμών** των **6 αλγορίθμων**, καθώς οι αλγόριθμοι Affinity Propagation και NCMine δεν έδειξαν σημαντική ομοιότητα για κανένα από τα σύνολα δεδομένων και με καμία παράμετρο. Οι 14 πίνακες με τα αναφερόμενα ζεύγη παρουσιάζονται παρακάτω, σε μορφή heat map για την καλύτερη κατανόηση τους.

Affinity Propagation vs ClusterONE	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_Chicken_Gallus_no_self_loops: AFFLAMDA05-CLUSTERONEDEN02	0.559925	0.032244	4.221285	0.681497	0.709916	0.532374	0.878571	0.946666
IntAct_Parkinson: AFFLAMDA05-CLUSTERONEOVERLAP06	0.316062	**0.696675**	3.442929	0.598373	0.623265	0.467557	0.367668	0.830827
BioGRID_Oryza_no_self_loops:AFFLAMDA05-CLUSTERONEDEN02	0.615088	**1.000000**	**-0.01027**	**0.969697**	**0.970143**	**0.941176**	**0.966576**	**1.000000**
BioGRID_Taurus_no_self_loops: AFFLAMDA05-CLUSTERONEDEN02	**0.684663**	0.072219	**2.735019**	**0.863141**	**0.869747**	**0.768595**	**0.927436**	**0.984211**

Πίνακας 8 : Heatmap σύγκρισης αλγορίθμων Affinity Propagation και ClusterONE βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

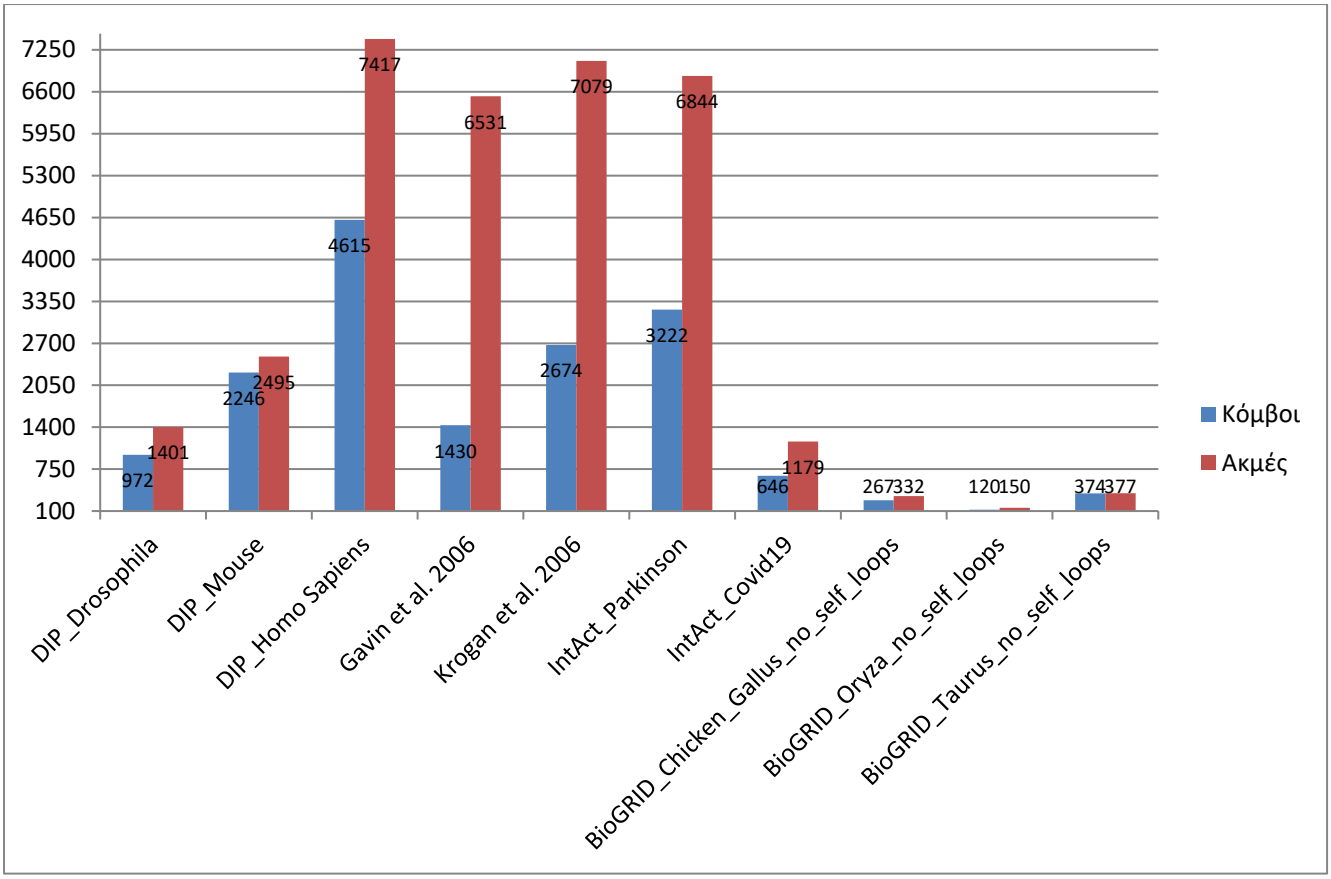


Διάγραμμα 3: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων Affinity Propagation και ClusterONE

Όπως φαίνεται και στο παραπάνω ραβδόγραμμα [Διάγραμμα 3] η σύγκριση των αλγορίθμων Affinity Propagation με ClusterONE απέδωσε ομοιότητα σε 4 σύνολα δεδομένων: BioGRID_Chicken_Gallus_no_self_loops, IntAct_Parkinson, BioGRID_Oryza_no_self_loops και BioGRID_Taurus_no_self_loops. Επιπλέον, οι παράμετροι με τις οποίες πάρθηκαν καλύτερες τιμές είναι στα τρία από αυτά ο συνδυασμός 0.5 ως Lamda parameter στον Affinity Propagation με 0.2 ως minimum density στο ClusterONE, ενώ στο σύνολο IntAct_Parkinson για τις παραμέτρους 0.5 ως Lamda parameter στον Affinity Propagation και 0.6 ως overlap threshold στον ClusterONE [Πίνακας 8]. Τα τρία πρώτα σύνολα δεδομένων αποτελούνται από παρόμοιο αριθμό κόμβων και ακμών, ενώ είναι και τα μικρότερα σύνολα από τα 23 που εκτελέστηκαν. Αντίθετα το σύνολο από την IntAct αποτελείται από δεκαπλάσιο αριθμό πρωτεϊνών και αξίζει να σημειωθεί πως μόνο η μετρική ARI δίνει αρκετά υψηλή τιμή.

Affinity Propagation vs MCL	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_Chicken_Gallus_no_self_loops: AFFLAMDA02-MCLINFL3	0.579818	0.479803	1.445826	0.783519	0.787518	0.871212	0.48058	0.711864
DIP_Drosophila:AFFLAMDA02-MCLINFL1.5	0.461928	0.486434	1.623711	0.971246	0.971648	0.944099	**0.978423**	**1.000000**
DIP_Homo_Sapiens: AFFLAMDA02-MCLINFL1.5	0.176575	0.346949	2.546474	0.935609	0.937555	0.879009	0.957109	**1.000000**
DIP_Mouse: AFFLAMDA02-MCLINFL1.5	0.56178	0.339315	2.461661	0.873034	0.880157	0.774676	0.91962	**1.000000**
GAVIN_2006:AFFLAMDA02-MCLINFL1.5	0.112892	0.17389	2.376311	0.77193	0.792825	0.628571	0.793651	**1.000000**
Intact_data_COVID19:AFFLAMDA02-MCLINFL1.5	0.19689	0.086454	2.917135	0.747253	0.772328	0.596491	0.8	**1.000000**
IntAct_Parkinson: AFFIT10-MCLINFL2	0.451899	**0.613141**	1.771048	0.914638	0.914663	0.907911	0.73023	0.921466
Krogan_BIND_2006: AFFLAMDA02-MCLINFL1.5	0.179434	0.195489	2.641261	**0.978417**	**0.978645**	**0.957746**	0.97619	**1.000000**
BioGRID_Oryza_no_self_loops: AFFLAMDA05-MCLINFL2	**0.944623**	**1.000000**	**-0.01027**	**0.999999**	**1.000000**	**1.000000**	**0.978698**	**1.000000**
BioGRID_Taurus_no_self_loops:AFFLAMDA05-MCLINFL3	**0.70984**	0.443417	**1.263135**	0.914306	0.915883	0.863636	0.868461	0.971292

Πίνακας 9: Heatmap σύγκρισης αλγορίθμων Affinity Propagation και MCL βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

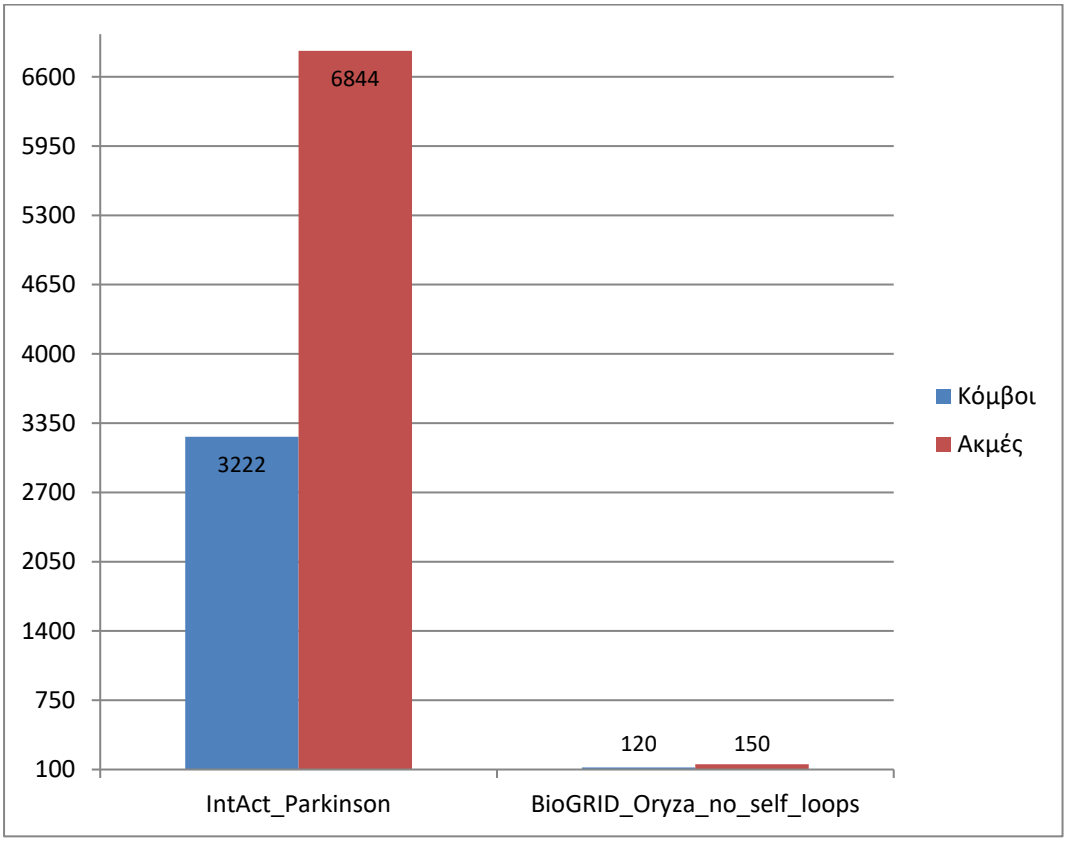


Διάγραμμα 4: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων Affinity Propagation και MCL

Ανάμεσα στους αλγορίθμους Affinity Propagation και MCL βρέθηκαν 10 συνδυασμοί, οι οποίοι δίνουν σημαντικά αποτελέσματα [Πίνακας 9]. Τα σύνολα αυτά κυμαίνονται σε ένα μεγάλο εύρος αριθμού κόμβων και ακμών, από μερικές εκατοντάδες, μέχρι και πάνω από επτά χιλιάδες ακμές [Διάγραμμα 4]. Παράλληλα, όπως φαίνεται και στον Πίνακα 9, ιδιαίτερα για τις μετρικές που υπολογίζονται μέσω των επικαλύψεων ομάδων, οι τιμές που λαμβάνονται είναι υψηλότατες. Οι παράμετροι, με τις οποίες τα αποτελέσματα των αλγορίθμων είναι πιο όμοια, είναι σε έξι σύνολα για Affinity Propagation Lamda parameter 0.2 και MCL Inflamation value 1.5 και από ένα σύνολο για τους συνδυασμούς Affinity Propagation Lamda parameter 0.2 και MCL Inflamation value 3, Affinity Propagation Lamda parameter 0.5 και MCL Inflamation value 2, Affinity Propagation Lamda parameter 0.5 και MCL Inflamation value 3 και τέλος Affinity Propagation Number of Iterations 10 και MCL Inflamation value 2.

Affinity Propagation vs MCODE	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
IntAct_Parkinson: AFFLAMDA05-MCODENODETHRES08	0.154389	0.649519	3.668021	0.662732	0.676195	0.552941	0.316326	0.826923
BioGRID_Oryza_no_self_loops: AFFLAMDA02-MCODENODETHRES02	**0.281822**	**1.000000**	**-0.01027**	**1.000000**	**1.000000**	**1.000000**	**0.944444**	**1.000000**

Πίνακας 10: Heatmap σύγκρισης αλγορίθμων Affinity Propagation και MCODE βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



Διάγραμμα 5: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων Affinity Propagation και MCODE

Κατά την σύγκριση των αλγορίθμων Affinity Propagation και MCODE μόνο δύο σύνολα εμφάνισαν κάποια ομοιότητα [Διάγραμμα 5], το IntAct_Parkinson και BioGRID_Oryza_no_self_loops, για τις παραμέτρους 0.5 Lamda parameter του Affinity Propagation με 0.8 node score threshold του MCODE και 0.2 Lamda parameter του Affinity Propagation με 0.2 node score threshold του MCODE

αντίστοιχα. Ακόμα, στο σύνολο IntAct_Parkinson λαμβάνονται υπόψην ως σημαντικά υψηλές, κατά κύριο λόγο, οι τιμές των ARI και PPV [Πίνακας 10].

Affinity Propagation vs SPICi	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_Taurus_no_self_loops:AFFLAMDA05-SPICISUP02	0.384909	0.048229	3.121123	0.788522	0.80567	0.654255	0.630434	0.992126

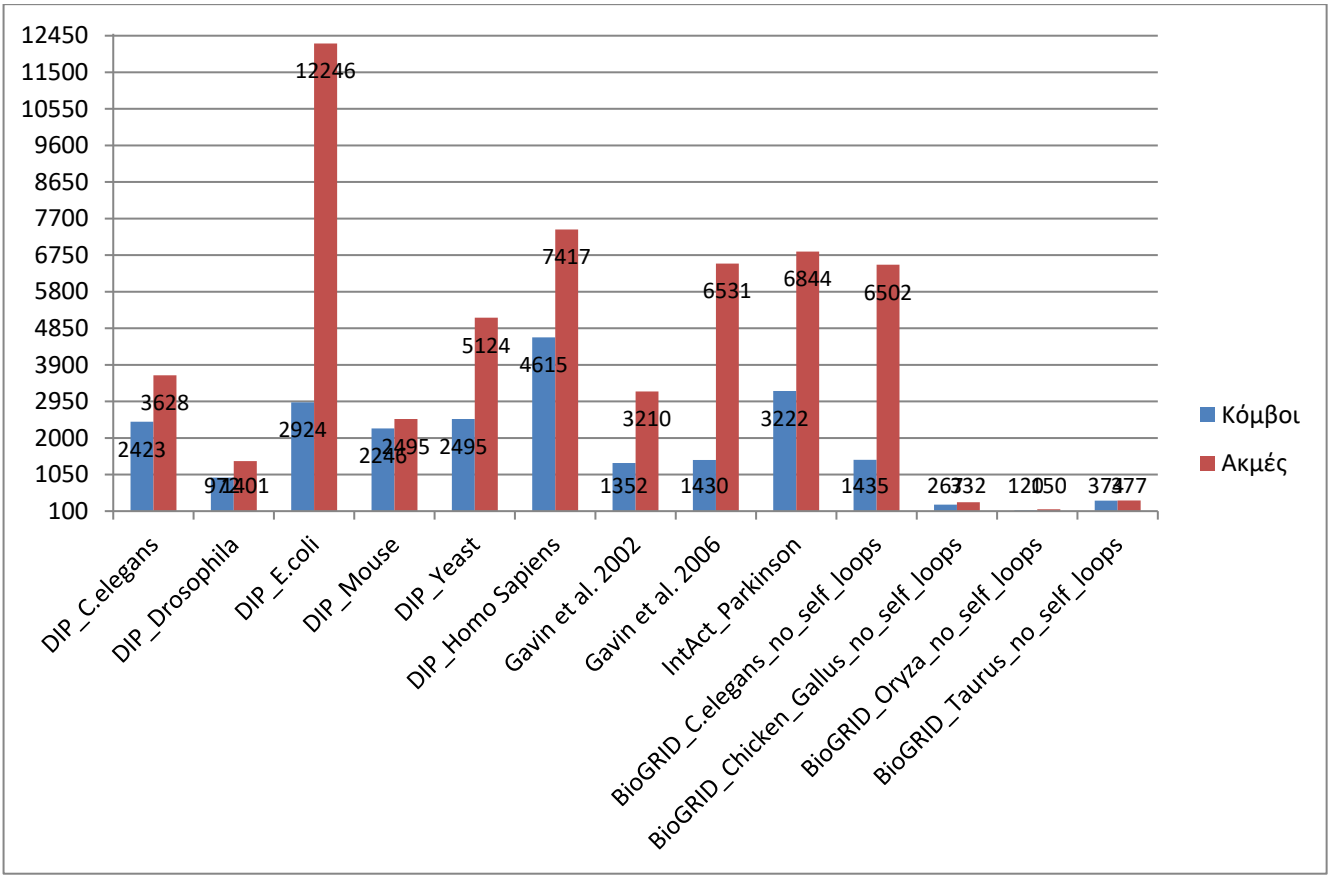
Πίνακας 11: Heatmap σύγκρισης αλγορίθμων Affinity Propagation και SPICi βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

Όμοια με το προηγούμενο συνδυασμό αλγορίθμων και το ζεύγος Affinity Propagation με SPICi παρουσίασαν ελάχιστη ομοιότητα και συγκεκριμένα μόνο στο σύνολο BioGRID_Taurus_no_self_loops, για τις παραμέτρους Lamda 0.5 του Affinity Propagation και 0.2 minimum cluster density του αλγορίθμου SPICi. Επιπλέον ικανοποιητικές τιμές μπορούν να θεωρηθούν μόνο των μετρικών F1-score, Accuracy και PPV [Πίνακας 11].

ClusterONE vs MCL	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:CLUSTERONEOVERLAP06-MCLINFL3	0.511498	0.486631	1.663699	0.749505	0.749949	0.776208	0.580899	0.724579
BioGRID_Chicken_Gallus_no_self_loops:CLUSTERONEOVERLAP06-MCLINFL2	0.665216	0.143062	3.388675	0.904254	0.904845	0.872727	0.832769	0.938144
DIP_Celegans:CLUSTERONEDEN02-MCLINFL3	0.553268	0.046868	2.593303	0.779022	0.779762	0.814487	0.572919	0.746518
DIP_Drosophila:CLUSTERONEOVERLAP06-MCLINFL2	0.694821	0.527469	**0.96148**	0.811874	0.812038	0.828508	0.719646	0.795895
DIP_Ecoli:CLUSTERONEOVERLAP06-MCLINFL3	0.450851	0.031722	2.371149	0.762157	0.762342	0.745746	0.635367	0.779307
DIP_Homo_Sapiens:CLUSTERONEDEN02-MCLINFL3	0.529506	0.207137	1.722815	0.784302	0.786018	0.839754	0.553004	0.735721
DIP_Mouse:CLUSTERONEDEN02-MCLINFL3	0.664976	0.592695	1.015003	0.830123	0.831347	0.877741	0.631731	0.787405

DIP_Yeast:CLUSTERONEDEN06-MCLINFL3	0.314163	0.143786	1.779703	0.763207	0.765277	0.710954	0.734855	0.823751
GAVIN_2002:CLUSTERONEDEN02-MCLINFL2	0.519844	0.507611	1.686700	0.765247	0.765254	0.762158	0.586385	0.768361
GAVIN_2006:CLUSTERONEDEN02-MCLINFL3	0.474528	0.508965	1.719191	0.793259	0.796347	0.869787	0.494294	0.729109
IntAct_Parkinson:CLUSTERONEOVERLAP06-MCLINFL3	0.366131	**0.715559**	2.558000	0.858823	0.860871	0.803435	0.70336	0.922414
BioGRID_Oryza_no_self_loops:CLUSTERONEDEN02-MCLINFL2	0.65652	**1.000000**	**-0.01027**	**1.000000**	**1.000000**	**1.000000**	**0.990385**	**1.000000**
BioGRID_Taurus_no_self_loops:CLUSTERONEDEN02-MCLINFL2	**0.712662**	0.315945	1.487874	**0.933204**	**0.933654**	**0.905109**	**0.920012**	**0.963099**

Πίνακας 12: Heatmap σύγκρισης αλγορίθμων ClusterONE και MCL βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



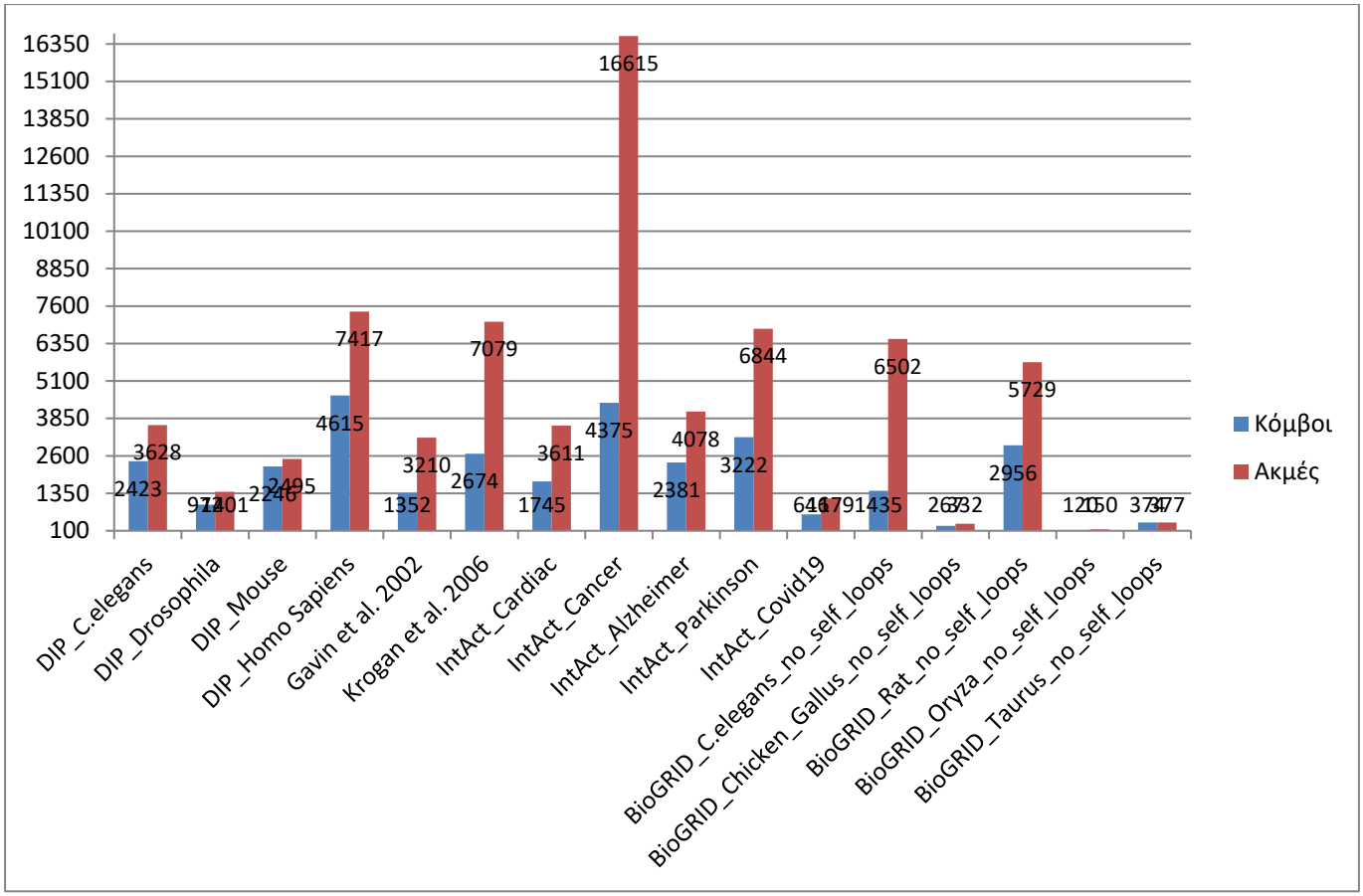
Διάγραμμα 6: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων ClusterONE και MCL

Αντιθέτως, δύο αλγόριθμοι, οι οποίοι φαίνεται να δίνουν αρκετά όμοια αποτελέσματα είναι οι ClusterONE και MCL. Οι δύο μέθοδοι βρέθηκαν σε 13 σύνολα να έχουν υψηλές τιμές στην πλειοψηφία των μετρικών που υπολογίστηκαν [Πίνακας 12]. Τα σύνολα αυτά, όπως φαίνεται και στο παραπάνω ραβδόγραμμα [Διάγραμμα 6], κυμαίνονται σε ένα αρκετά ευρύ φάσμα, ως προς τον αριθμό των κόμβων και τον ακμών τους, αντικατοπτρίζοντας ένα πλήθος διαφορετικών δικτύων αλληλεπιδράσεων, που δέχονται οι αλγόριθμοι ομαδοποίησης ως είσοδο. Όσο αφορά τις παραμέτρους με τις οποίες υπολογίστηκαν οι βέλτιστες τιμές σύγκρισης, τέσσερα σύνολα με 0.2 minimum density του ClusterONE και 3 Inflamation value του MCL, τρία σύνολα με 0.2 minimum density του ClusterONE και 2 Inflamation value του MCL, τρία σύνολα είναι με τις παραμέτρους 0.6 Overlap threshold του ClusterONE και 3 Inflamation value του MCL, δύο με 0.6 Overlap threshold του ClusterONE και 2 Inflamation value του MCL, ενώ τέλος ένα σύνολο με 0.6 minimum density του ClusterONE και 3 Inflamation value του MCL.

ClusterONE vs MCODE	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:CLUSTERONEDEN08-MCODENODETHRES02	0.16239	**0.552503**	1.015888	0.697684	0.705703	0.606516	0.52005	0.821109
BioGRID_Chicken_Gallus_no_self_loops:CLUSTERONEDEN08-MCODENODETHRES03	0.136875	0.197236	0.354465	0.857143	0.866025	0.75	0.508333	**1.000000**
DIP_Celegans:CLUSTERONEDEN08-MCODENODETHRES02	0.040995	0.177283	**0.181462**	0.567568	0.629465	0.396226	0.481027	**1.000000**
DIP_Drosophila:CLUSTERONEDEN08-MCODENODETHRES03	0.11464	0.463006	0.252929	0.682131	0.705972	0.542372	0.394531	0.918919
DIP_Homo_Sapiens:CLUSTERONEDEN05-MCODENODETHRES02	0.102137	0.215521	0.740775	0.73862	0.756223	0.608163	0.436983	0.940329
DIP_Mouse:CLUSTERONEDEN08-MCODENODETHRES02	0.113312	0.242606	0.306592	0.554874	0.617105	0.385965	0.353295	0.986667
GAVIN_2002:CLUSTERONEOVERLAP06-MCODENODETHRES02	0.048961	0.244568	1.501125	0.743169	0.750203	0.653846	0.351272	0.86076

IntAct_Alzheimers:CLUSTERONEDEN08-MCODENODETHRES03	0.019297	0.330499	0.247064	0.622977	0.623573	0.596899	0.517711	0.651438
IntAct_Cancer:CLUSTERONEOVERLAP06-MCODENODETHRES03	0.116038	0.048043	1.963245	0.745781	0.750188	0.672947	0.538744	0.836294
IntAct_Cardiac:CLUSTERONEDEN08-MCODENODETHRES03	0.156899	0.165872	0.576365	0.755187	0.768706	0.636364	**0.631963**	0.928571
Intact_data_COVID19:CLUSTERONEDEN08-MCODENODETHRES03	0.101932	0.382152	0.328564	**0.864583**	**0.870254**	**0.776119**	0.533532	0.975806
IntAct_Parkinson:CLUSTERONEDEN08-MCODENODETHRES02	0.049525	0.495737	2.673298	0.806657	0.811218	0.729411	0.472857	0.9022
Krogan_BIND_2006:CLUSTERONEDEN05-MCODENODETHRES02	0.154661	0.332425	1.006537	0.740847	0.752287	0.631193	0.547602	0.896613
BioGRID_Oryza_no_self_loops:CLUSTERONEDEN08-MCODENODETHRES02	0.511045	**1.000000**	**-0.010268**	**1.000000**	**1.000000**	**1.000000**	**0.944444**	**1.000000**
BioGRID_Rat_no_self_loops:CLUSTERONEDEN02-MCODENODETHRES03	0.06659	0.011683	1.900476	0.819627	0.823699	0.745583	0.52264	0.91
BioGRID_Taurus_no_self_loops:CLUSTERONEDEN08-MCODENODETHRES03	**0.237447**	0.412824	0.339776	0.423821	0.42465	0.398907	0.166406	0.452055

Πίνακας 13: Heatmap σύγκρισης αλγορίθμων ClusterONE και MCOD βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

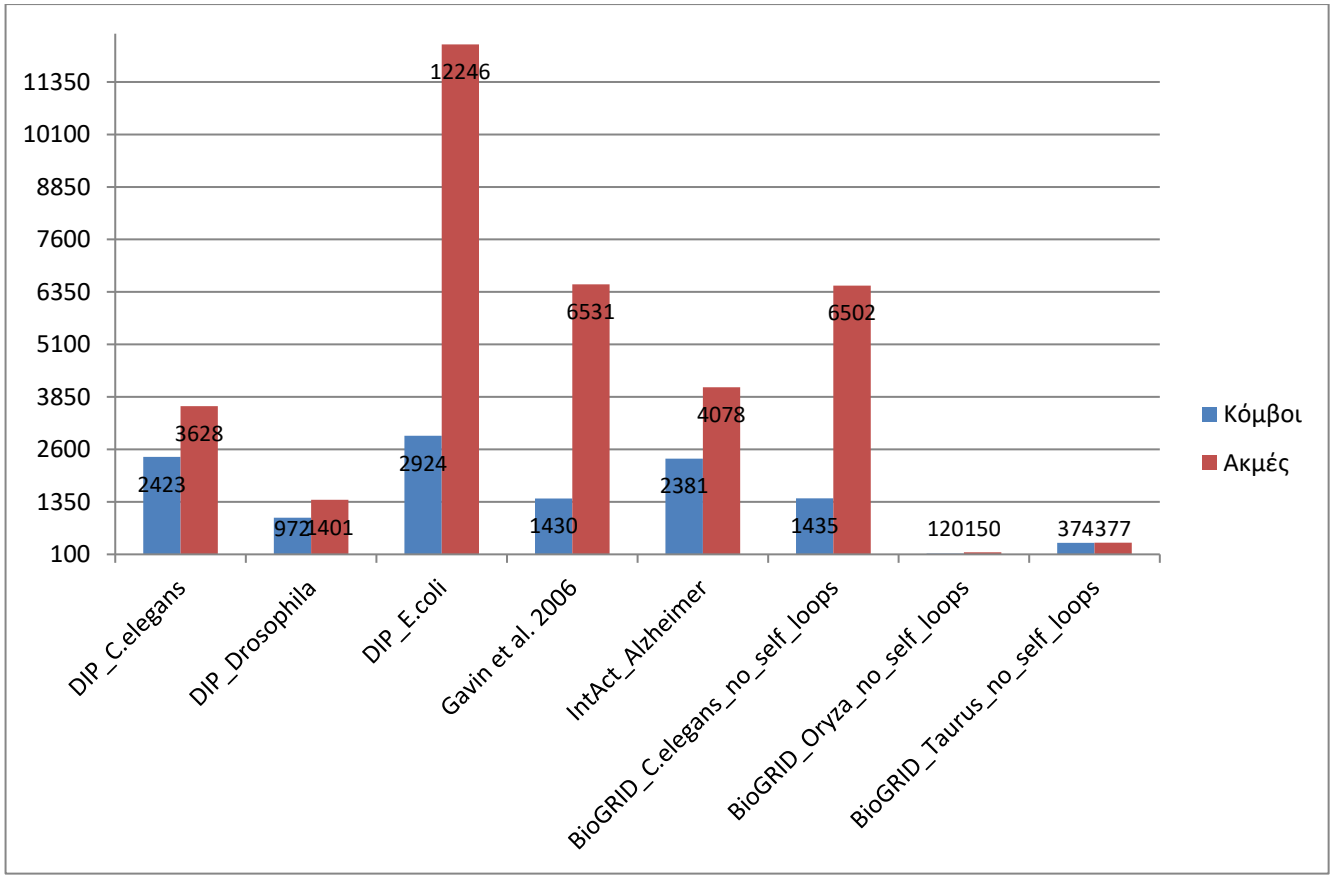


Διάγραμμα 7: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων ClusterONE και MCODE

Η σύγκριση των αλγορίθμων ClusterONE με MCODE, παρόλο που απέδωσε περισσότερα σύνολα [Διάγραμμα 7], οι τιμές των μετρικών κυμαίνονται σε αρκετά χαμηλότερα επίπεδα, σε σύγκριση με το προηγούμενο ζεύγος αλγορίθμων, όπως φαίνεται και από την ένταση των χρωμάτων στους δύο heatmaps. Στα 16 σύνολα, του συγκεκριμένου πίνακα [Πίνακας 13], οι παράμετροι των αλγορίθμων είναι 0.8 minimum density του ClusterONE με 0.3 node score threshold του MCODE σε έξι περιπτώσεις, 0.8 minimum density του ClusterONE με 0.2 node score threshold του MCODE σε πέντε, ενώ από μια περίπτωση παρατηρείται στου ακόλουθους συνδυασμούς: 0.2 minimum density του ClusterONE με 0.3 node score threshold του MCODE, 0.6 minimum density του ClusterONE με 0.2 node score threshold του MCODE, 0.5 minimum density του ClusterONE με 0.2 node score threshold του MCODE, 0.6 overlap threshold του ClusterONE με 0.3 node score threshold του MCODE και 0.6 overlap threshold του ClusterONE με 0.2 node score threshold του MCODE.

ClusterONE vs NCMine	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:CLUSTERONEDEN08-NCMINEcliq08	0.172587	**0.706561**	0.900532	0.39365	0.412848	0.563485	0.219764	0.302481
DIP_Celegans:CLUSTERONEDEN08-NCMINEcliq08	0.013187	0.166154	**0.230538**	0.295732	0.407041	0.175287	0.069461	**0.945205**
DIP_Drosophila:CLUSTERONEDEN08-NCMINEcliq08	0.042134	0.374019	0.344755	0.393748	0.467762	0.255708	0.093262	0.85567
DIP_Ecoli:CLUSTERONEDEN08-NCMINEcliq08	0.023282	0.132407	0.34761	0.547058	0.585223	0.40367	0.178344	0.848432
GAVIN_2006:CLUSTERONEDEN08-NCMINEmerge04	**0.197271**	0.523992	1.169722	0.597016	0.62435	**0.844019**	**0.494494**	0.461854
IntAct_Alzheimers:CLUSTERONEDEN08-NCMINEcliq04	0.038398	0.245643	0.29405	0.204533	0.231203	0.383212	0.042171	0.139492
BioGRID_Oryza_no_self_loops:CLUSTERONEDEN08-NCMINEcliq08	0.140336	**1.000000**	**-0.010268**	0.22069	0.342997	**0.941176**	0.023529	0.125
BioGRID_Taurus_no_self_loops:CLUSTERONEDEN02-NCMINEcliq04	0.121244	0.161142	1.251816	**0.801909**	**0.802295**	0.827586	0.270207	0.777777

Πίνακας 14: Heatmap σύγκρισης αλγορίθμων ClusterONE και NCMine βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



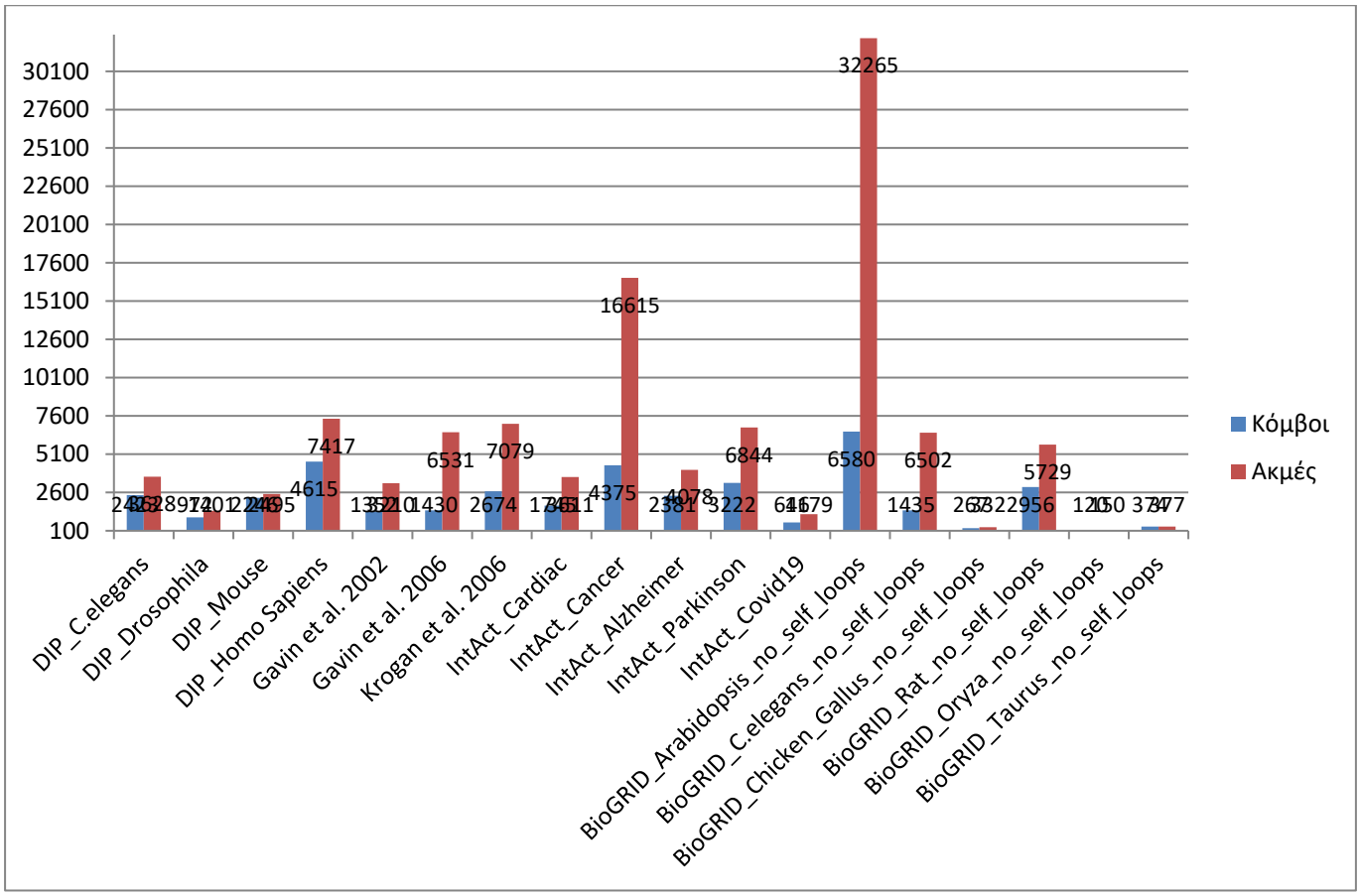
Διάγραμμα 8: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων ClusterONE και NCMine

Μέτριου επιπέδου ομοιότητα φαίνεται να υπάρχει ανάμεσα στους αλγορίθμους ClusterONE και NCMine, με την σύγκριση να υποδεικνύει 8 σύνολα δεδομένων [Διάγραμμα 8], με οριακά όμως τιμές υψηλότερες των κατωφλίων και κυρίως στα VI και PPV, με τις υπόλοιπες μετρικές να διατηρούνται σε χαμηλά επίπεδα [Πίνακας 14]. Ο συνδυασμός παραμέτρων με τα πιο πολλά σύνολα, πέντε, είναι 0.8 ClusterONE minimum density – 0.8 NCMine cliqueness threshold. Με μόλις ένα σύνολο ακλουθούν οι συνδυασμοί 0.8 ClusterONE minimum density – 0.4 NCMine cliqueness threshold, 0.8 ClusterONE minimum density – 0.4 NCMine merge threshold, 0.2 ClusterONE minimum density – 0.4 NCMine cliqueness threshold.

ClusterONE vs SPICi	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_Arabidopsis_no_self_loops:CLUSTERONEDEN02-SPICIDEN05	0.141432	0.168132	2.276403	0.702908	0.706076	0.776368	0.391936	0.642148
BioGRID_C.elegans_no_self_loops:CLUSTERONEDEN08-SPICIDEN07	0.235024	**0.748416**	0.935488	0.671218	0.68775	0.551111	0.252874	0.858268
BioGRID_Chicken_Gallus_no_self_loops:CLUSTERONEDEN02-SPICIDEN02	0.405363	0.392019	0.78513	**0.910256**	**0.910256**	**0.910256**	0.732471	0.910256
DIP_Celegans:CLUSTERONEOVERLAP06-SPICIDEN05	0.253681	0.376516	0.942116	0.753602	0.767558	0.767558	0.528595	0.930175
DIP_Drosophila:CLUSTERONEOVERLAP06-SPICISUP02	0.30761	0.488216	1.221808	0.796512	0.801528	0.716475	0.614674	0.896679
DIP_Homo_Sapiens:CLUSTERONEOVERLAP06-SPICISUP02	0.318559	0.452968	1.183870	0.770899	0.776917	0.685714	0.556499	0.88025
DIP_Mouse:CLUSTERONEOVERLAP06-SPICISUP02	0.406786	0.518582	0.975352	0.813316	0.819352	0.725424	0.64816	0.925443
GAVIN_2002:CLUSTERONEOVERLAP06-SPICIDEN07	0.207719	0.382557	1.412025	0.761034	0.761767	0.729055	0.323338	0.795948
GAVIN_2006:CLUSTERONEDEN08-SPICIDEN07	0.453475	0.594577	1.082751	0.721282	0.734353	0.607221	0.353638	0.888104
IntAct_Alzheimers:CLUSTERONEDEN08-SPICISUP02	0.137723	0.39573	**0.316838**	0.599755	0.647145	0.435986	0.404531	0.960573
IntAct_Cardiac:CLUSTERONEDEN05-SPICISUP02	0.127584	0.570436	0.398658	0.822387	0.828693	0.732227	**0.801184**	0.937867
IntAct_Cancer:CLUSTERONEDEN05-SPICIDEN07	0.123512	0.131673	1.179516	0.794972	0.794997	0.788603	0.518098	0.801444
Intact_data_COVID19:CLUSTERONEDEN05-SPICISUP02	0.265101	0.380668	0.495396	0.770037	0.790312	0.628571	0.544332	**0.993671**
IntAct_Parkinson:CLUSTERONEDEN02-SPICIDEN02	0.101223	0.214113	0.499396	0.813241	0.820107	0.720238	0.509908	0.933823
Krogan_BIND_2006:CLUSTERONEDEN08-SPICIDEN07	0.288228	0.58169	0.733273	0.651735	0.659433	0.565573	0.236573	0.768868
BioGRID_Oryza_no_self_loops:CLUSTERONEDEN02-SPICIDEN02	**0.669631**	**1.000000**	**-0.010268**	**0.936709**	**0.93859**	0.880952	**0.861141**	**1.000000**

BioGRID_Rat_no_self_loops:CLUSTERONEDEN02-SPICISUP02	0.32128	0.121448	1.413529	0.815253	0.817461	0.759468	0.693299	0.879883
BioGRID_Taurus_no_self_loops:CLUSTERONEDEN02-SPICIDEN02	0.513561	0.595989	0.797544	0.89446	0.896194	0.842105	0.758652	0.953757

Πίνακας 15: Heatmap σύγκρισης αλγορίθμων ClusterONE και SPICi βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



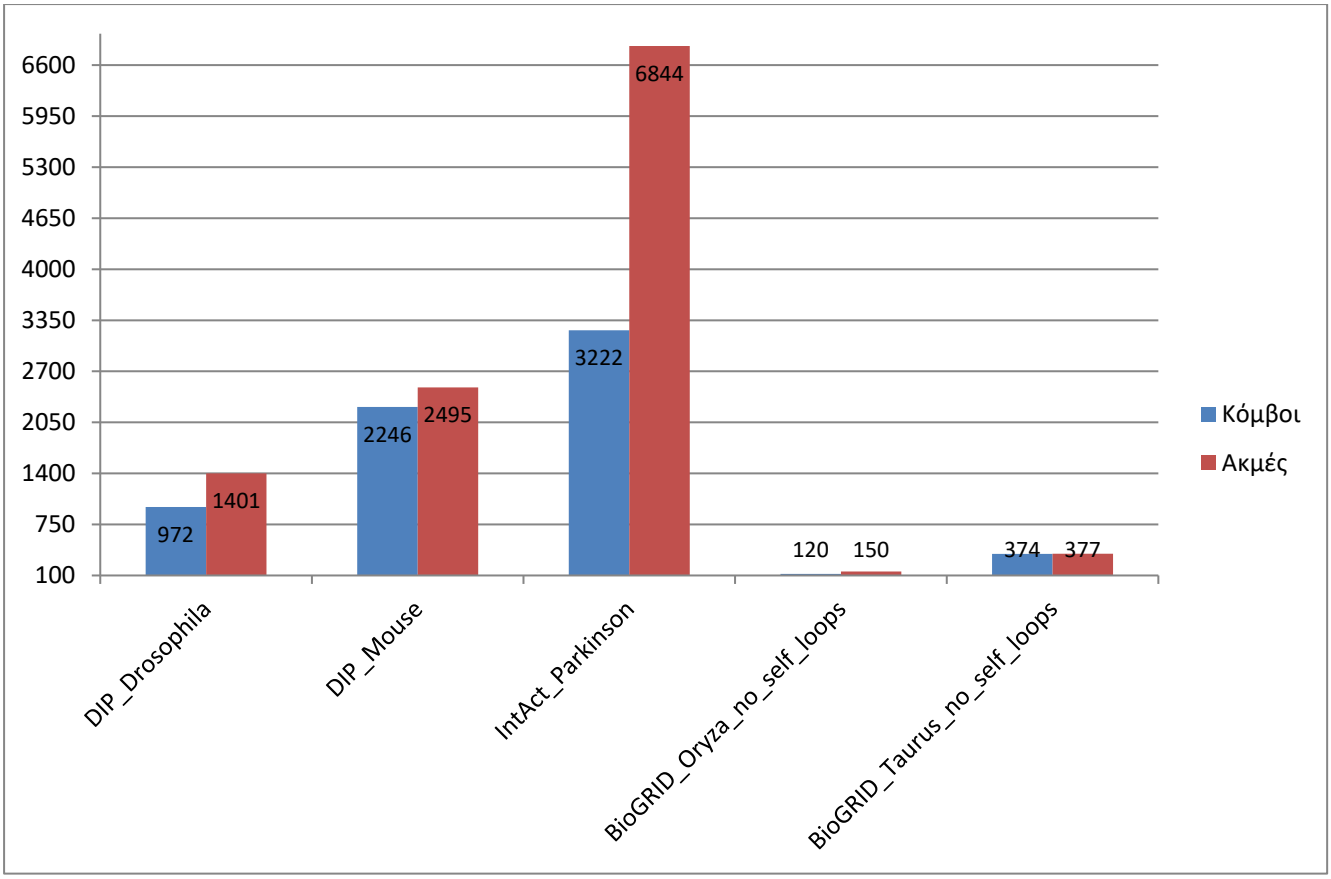
Διάγραμμα 9: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων ClusterONE και SPICi

Ο συνδυασμός αλγορίθμων ClusterONE με SPICi έδειξε την μεγαλύτερη ομοιότητα, ως προς τον αριθμό αποτελεσμάτων με υψηλές τιμές από όλες τις υπόλοιπες συγκρίσεις, αφού βρέθηκαν 18 σύνολα δεδομένων με αξιόλογες τιμές μετρικών [Διάγραμμα 9]. Ειδικά στα μέτρα F1-score, Accuracy, Cluster-wise Sensitivity, MMR και PPV τα ποσοστά είναι πολύ υψηλά, ενώ και στις υπόλοιπες 3 μετρικές είναι ψηλότερα του γενικού μέσου όρου [Πίνακας 15]. Ως προς τις παραμέτρους των αλγορίθμων ClusterONE και SPICi οι έξι διαφορετικοί συνδυασμοί παραμέτρων που εμπεριέχονται στον συγκεντρωτικό πίνακα είναι: ClusterONE minimum density 0.2

με SPICi minimum cluster density 0.2 σε τέσσερα σύνολα δεδομένων, ClusterONE overlap threshold 0.6 με SPICi minimum support threshold 0.2 σε τρία σύνολα, από δύο με τις παραμέτρους ClusterONE minimum density 0.8 με SPICi minimum cluster density 0.7 και ClusterONE minimum density 0.5 με SPICi support threshold 0.2, ενώ τέλος μία φορά συναντώνται οι συνδυασμοί: ClusterONE minimum density 0.2 με SPICi support threshold 0.2, ClusterONE minimum density 0.2 - SPICi minimum cluster density 0.5, ClusterONE minimum density 0.8 με SPICi support threshold 0.2, ClusterONE minimum density 0.8 - SPICi minimum cluster density 0.7, ClusterONE minimum density 0.5 - SPICi minimum cluster density 0.7, ClusterONE overlap threshold 0.6 - SPICi minimum cluster density 0.5 και ClusterONE overlap threshold 0.6 - SPICi minimum cluster density 0.7.

MCL vs MCODE	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
DIP_Drosophila:MCLINFL2OUTPUT-MCODENODETHRES02	0.055965	0.099484	**1.802817**	0.922297	0.923647	0.875	0.524354	0.975
DIP_MouseMCLINFL3-MCODENODETHRES02	0.029108	0.057654	2.005977	0.839795	0.848114	0.736842	0.434033	0.97619
IntAct_Parkinson:MCLINFL2-MCODENODETHRES02	**0.107797**	**0.543044**	3.646700	0.87409	0.875035	0.835294	0.408458	0.916667
Cluster File 2: BioGRID_Oryza_no_self_loopsMCL02-MCODENODETHRES02	**0.211626**	**1.000000**	**-0.01027**	**1.000000**	**1.000000**	**1.000000**	**0.944444**	**1.000000**
BioGRID_Taurus_no_self_loops:MCLINFL2-MCODENODETHRES02	0.096393	0.035694	2.427147	**0.938775**	**0.94054**	**0.884615**	**0.563468**	**1.000000**

Πίνακας 16: Heatmap σύγκρισης αλγορίθμων MCL και MCODE βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



Διάγραμμα 10: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων MCL και MCODE

Συνεχίζοντας, με την σύγκριση των αλγορίθμων MCL και MCODE, φαίνεται να υπάρχει μετρίου επιπέδου ομοιότητα με μόλις 5 σύνολα [Διάγραμμα 10]. Οι μετρικές στις οποίες τα σύνολα δίνουν καλύτερα αποτελέσματα είναι το F1-score, η Accuracy, η Cluster- wise Sensitivity και η PPV. Σχεδόν όλα τα σύνολα απέδωσαν τις καλύτερες τιμές για τις παραμέτρους Inflamation value 2 του MCL και node score threshold 0.2 του MCODE, εκτός του συνόλου DIP_Mouse, για το οποίο τα αποτελέσματα των αλγορίθμων μοιάζουν περισσότερο με τις παραμέτρους Inflamation value 3 και node score threshold 0.2 [Πίνακας 16].

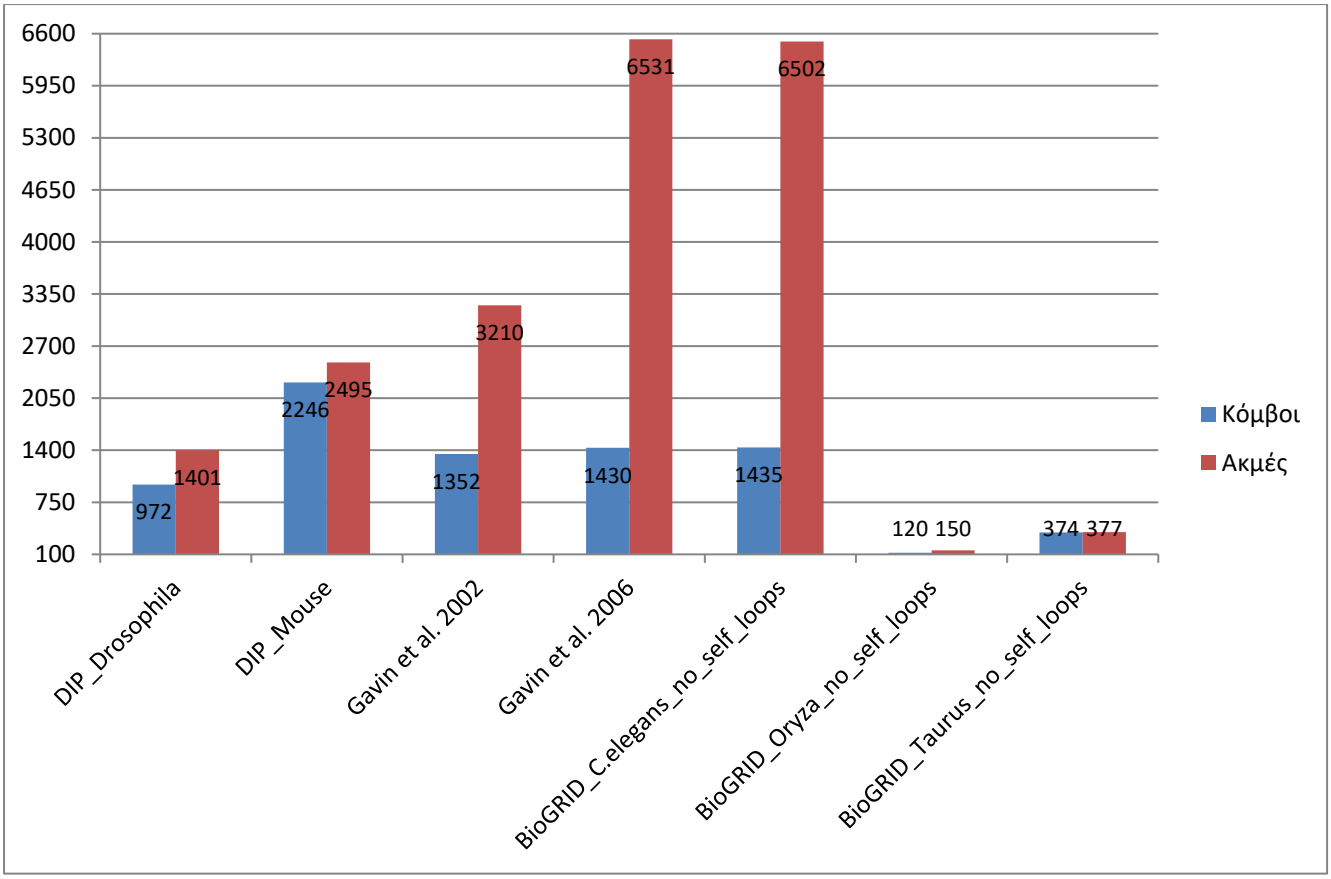
MCL vs NCMine	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_Taurus_no_self_loops:MCLINFL2-NCMINEcliq04	0.086856	0.039007	2.440232	0.770682	0.770779	0.758621	0.261681	0.783132

Πίνακας 17: Heatmap σύγκρισης αλγορίθμων MCL και NCMine βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

Ακόμη μικρότερη ομοιότητα, απ' ότι στην προηγούμενη περίπτωση που εξετάστηκε, φαίνεται να υπάρχει ανάμεσα στις μεθόδους MCL και NCMine, αφού το μόνο αποτέλεσμα με στατιστικά σημαντικές τιμές ήταν στο σύνολο BioGRID_Taurus_no_self_loops. Οι παράμετροι των αλγορίθμων αντιστοιχούν σε 2 Inflamation value για το MCL και 0.4 cliqueness threshold για το NCMine [Πίνακας 17].

MCL vs SPICi	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:MCLINFL3-SPICIDEN07	0.169157	**0.402250**	2.165372	0.762122	0.772228	0.656296	0.370019	0.908638
DIP_Drosophila:MCLINFL2-SPICISUP02	0.282570	0.338634	1.649039	0.787515	0.792687	0.706896	0.558358	0.888889
DIP_Mouse:MCLINFL3-SPICIDEN02	0.349388	0.389659	**1.596958**	0.813023	0.822781	0.704797	0.685224	0.960516
GAVIN_2002:MCLINFL2-SPICIDEN02	0.310017	0.266645	2.536442	0.803740	0.804438	0.771605	0.572167	0.838668
GAVIN_2006:MCLINFL3-SPICIDEN05	0.400369	0.331541	2.140515	0.758311	0.761253	0.697087	0.521986	0.831325
BioGRID_Oryza_no_self_loops:MCLINFL2-SPICIDEN02	**0.486109**	**1.000000**	**-0.010268**	**1.000000**	**1.000000**	**1.000000**	**0.861538**	**1.000000**
BioGRID_Taurus_no_self_loops:MCLINFL2-SPICIDEN02	**0.415627**	0.257020	1.824714	**0.923703**	**0.924903**	**0.878947**	**0.734168**	**0.973262**

Πίνακας 18: Heatmap σύγκρισης αλγορίθμων MCL και SPICi βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

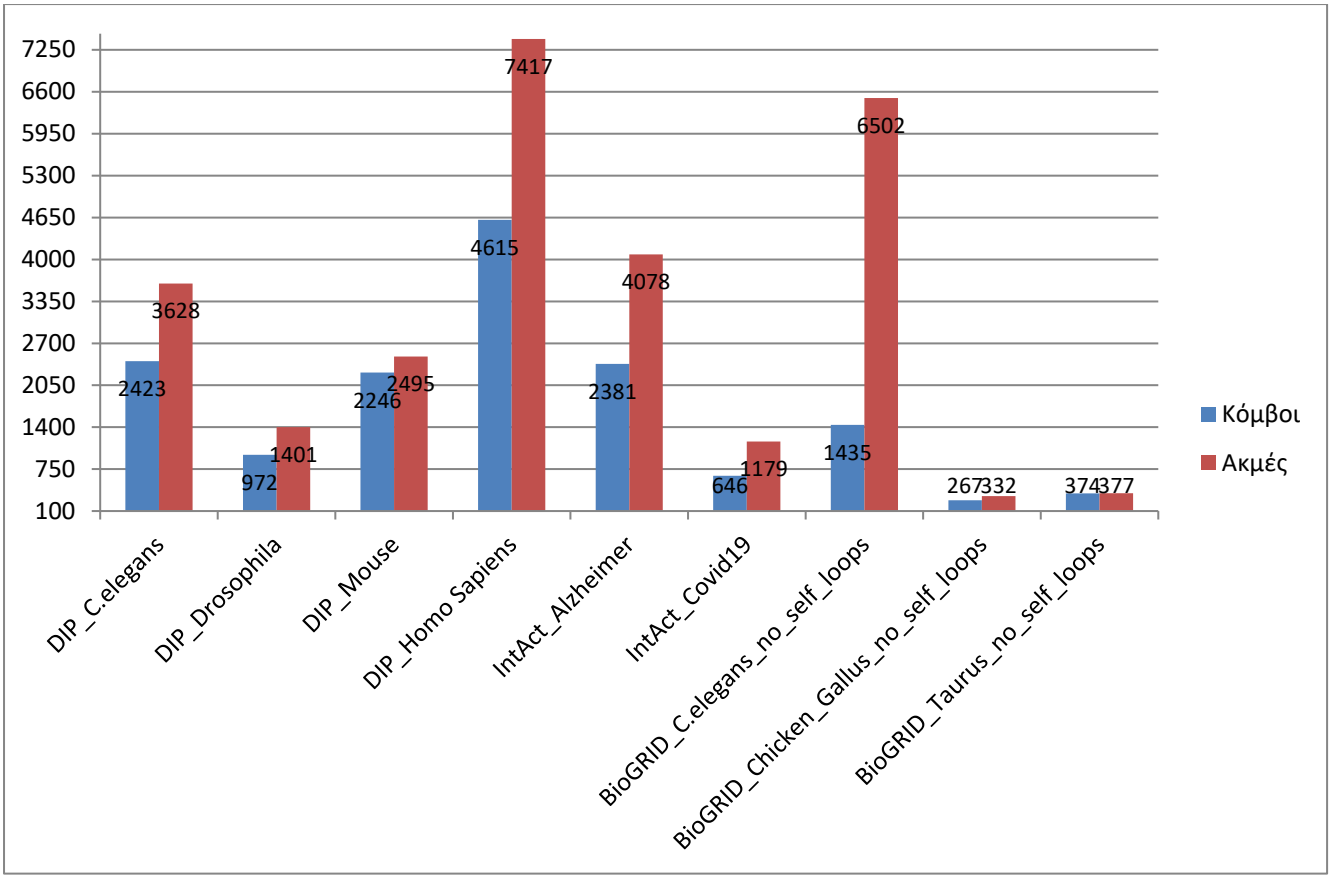


Διάγραμμα 11: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων MCL και SPICi

Οι αλγόριθμοι MCL και SPICi παρουσίασαν ομοιότητα σε 7 διαφορετικά σύνολα δεδομένων, με ιδιαίτερα υψηλές τιμές στα μικρότερα εξ αυτών, το πλήθος των κόμβων και των ακμών των οποίων φαίνεται στο παραπάνω γράφημα [Διάγραμμα 11]. Οι παράμετροι με τις οποίες τα αποτελέσματα τους συγκλίνουν περισσότερο είναι κυρίως οι MCL Inflamation value 0.2 με SPICi minimum cluster density 0.2, στα μισά σχεδόν σύνολα, ενώ στα υπόλοιπα τέσσερα οι παράμετροι ήταν 0.3 στο MCL Inflamation value και 0.2 στο SPICi minimum cluster density, 0.3 στο MCL Inflamation value και 0.5 στο SPICi minimum cluster density, 0.3 στο MCL Inflamation value και 0.7 στο SPICi minimum cluster density και τέλος 0.2 Inflamation value με την παράμετρο minimum support threshold του SPICi στο 0.2 [Πίνακας 18].

MCODE vs NCMine	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops: MCODENODETHRES05-NCMINEcliq08	0.002229	0.000721	0.700318	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**
BioGRID_Chicken_Gallus_no_self_loops: MCODENODETHRES05-NCMINEmerge08	**0.196291**	0.523551	**0.100306**	0.857143	0.866025	0.75	0.619444	**1.000000**
DIP_Celegans:MCODENODETHRES05-NCMINEcliq08	0.01436	0.2271	0.126285	0.684811	0.700404	0.566037	0.659524	0.866666
DIP_Drosophila: MCODENODETHRES03-NCMINEcliq08	0.065131	**0.595972**	0.15421	0.89001	0.891134	0.847457	0.63497	0.937063
DIP_Homo_Sapiens: MCODENODETHRES08-NCMINEcliq08	0.084398	0.292546	0.339545	0.769356	0.773676	0.695918	0.612355	0.860121
DIP_Mouse:MCODENODETHRES05-NCMINEcliq08	0.116148	0.307837	0.268174	0.799389	0.808376	0.695906	0.655675	0.939024
IntAct_Alzheimers:MCODEKCORE3-NCMINEcliq04	0.053953	0.235338	0.226963	0.59323	0.613434	0.472868	0.303652	0.795786
Intact_data_COVID19:MCODEKCORE3-NCMINEmerge08	0.087947	0.279282	0.225833	0.591255	0.635339	0.432836	0.265801	0.932584
BioGRID_Taurus_no_self_loops:MCODEKCORE3-NCMINEcliq08	0.169409	0.316671	0.235603	0.610169	0.662589	0.439024	0.333333	**1.000000**

Πίνακας 19: Heatmap σύγκρισης αλγορίθμων MCODE και NCMine βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

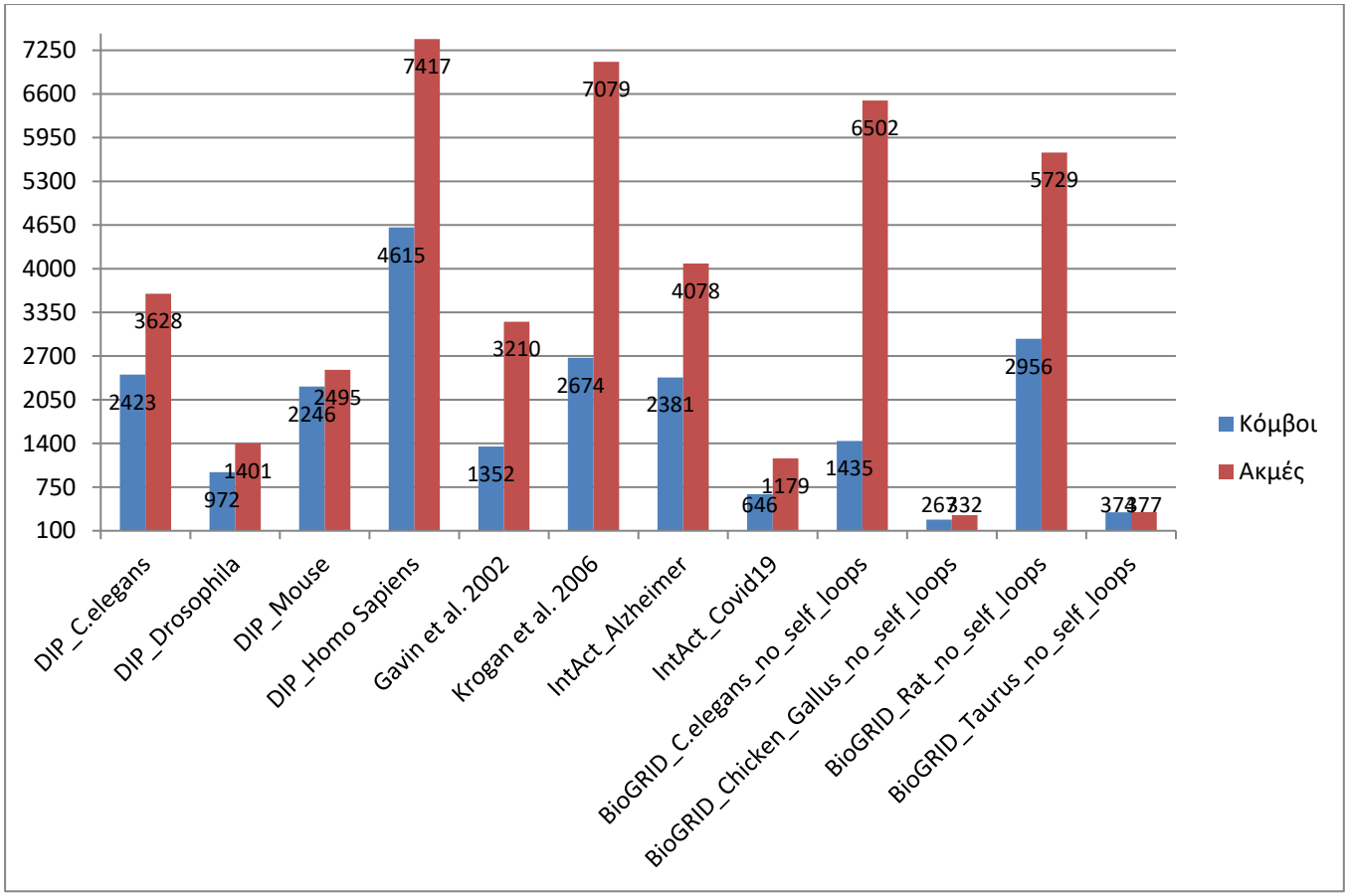


Διάγραμμα 12: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων MCL και NCMine

Παρόμοια αποτελέσματα, ως προς τον αριθμό συνόλων [Διάγραμμα 12], αλλά καλύτερα ως προς τις τιμές των μετρικών, προέκυψαν από την σύγκριση των αλγορίθμων MCODE και NCMine [Πίνακας 19]. Συνολικά εντοπίστηκαν 9 σύνολα με αξιοσημείωτες τιμές, τρία από τα οποία προέκυψαν με τις παραμέτρους node score threshold του MCODE στο 0.5 και cliqueness threshold του NCMine στο 0.8. Επιπλέον από ένα σύνολο ανά περίπτωση είχε τις παραμέτρους ρυθμισμένες στο 0.5 για το MCODE node score threshold και 0.8 για το NCMine merge threshold, 0.3 το MCODE node score threshold και στο 0.8 το NCMine cliqueness threshold , 0.8 το MCODE node score threshold και στο 0.8 το NCMine cliqueness threshold, 3 k-core threshold για το MCODE και 0.4 cliqueness threshold για το NCMine, 3 k-core threshold για το MCODE και 0.8 cliqueness threshold για το NCMine, ενώ τέλος ένα σύνολο είχε παραμέτρους MCODE k-core 3 και NCMine merge threshold 0.8.

MCODE vs SPICi	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:MCODENODETHRES02-SPICIDEN07	0.242983	**0.586925**	0.916238	0.726579	0.734435	0.634085	0.597029	0.850666
BioGRID_Chicken_Gallus_no_self_loops:MCODENODETHRES02-SPICIDEN07	0.273565	0.526021	**0.160415**	**1.000000**	**1.000000**	**1.000000**	**0.877778**	**1.000000**
DIP_Celegans::MCODENODETHRES05-SPICIDEN07	0.03091	0.144703	0.298209	0.691358	0.726844	0.528302	0.632143	**1.000000**
DIP_Drosophila:MCODENODETHRES02-SPICIDEN07	0.144387	0.470927	0.36868	0.887164	0.889693	0.825	0.662695	0.959459
DIP_Homo_Sapiens:MCODEKCORE2-SPICIDEN07	0.130374	0.302328	0.509692	0.75378	0.768583	0.630612	0.503246	0.93674
DIP_Mouse:MCODENODETHRES02-SPICIDEN07	0.0951394	0.269287	0.411666	0.763241	0.78451	0.619883	0.542303	0.992857
GAVIN_2002:MCODENODETHRES02-SPICIDEN07	0.155405	0.264763	1.329168	0.748366	0.755504	0.65812	0.373587	0.867299
IntAct_Alzheimers:MCODEKCORE3-SPICISUP02	0.162063	0.278008	0.349078	0.74749	0.750103	0.689922	0.437963	0.815533
Intact_data_COVID19:MCODENODETHRES02-SPICISUP08	**0.396975**	0.236845	0.237428	0.740242	0.763502	0.594595	0.746894	0.980392
Krogan_BIND_2006:MCODENODETHRES02-SPICIDEN07	0.174256	0.421858	0.762349	0.741906	0.754061	0.629357	0.590913	0.903475
BioGRID_Rat_no_self_loops:MCODEKCORE3-SPICISUP08	0.22101	0.023974	0.535472	0.73731	0.759901	0.59364	0.700348	0.972727
BioGRID_Taurus_no_self_loops:MCODENODETHRES02-SPICIDEN07	0.340142	0.43425	0.333232	0.951826	0.951875	0.942308	0.822727	0.961538

Πίνακας 20: Heatmap σύγκρισης αλγορίθμων MCODE και SPICi βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

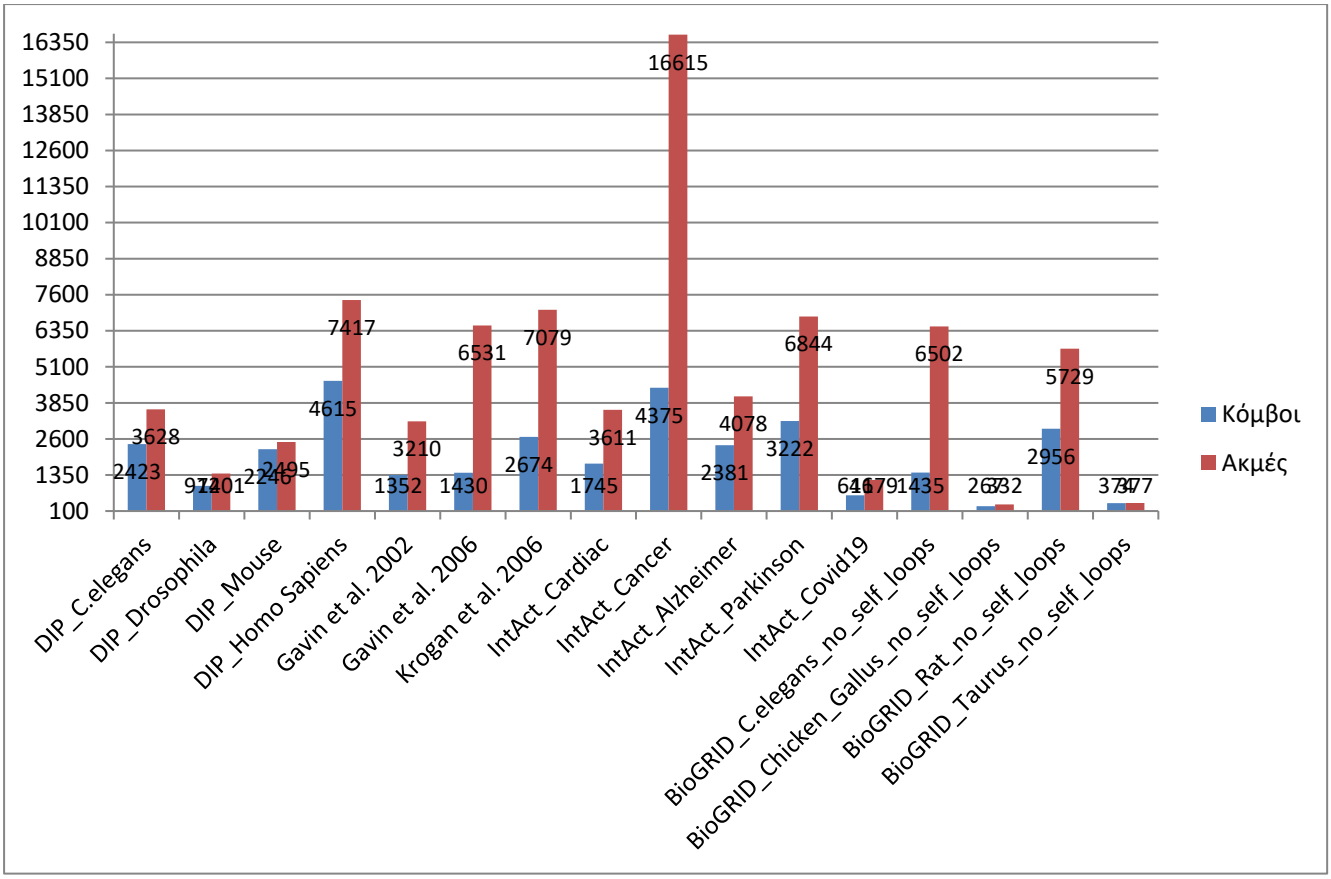


Διάγραμμα 13: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων MCODE και SPICi

Ο έλεγχος ομοιότητας των αποτελεσμάτων των αλγορίθμων MCODE και SPICi έδωσε 12 σύνολα [Διάγραμμα 13], στα οποία τα στατιστικά μέτρα, στην πλειοψηφία τους, ήταν άνω του μετρίου [Πίνακας 20]. Η μεγάλη πλειοψηφία των αποτελεσμάτων, οχτώ συνολικά, προέκυψαν με τις παραμέτρους MCODE node score threshold 0.2 και SPICi minimum cluster density 0.7. Επιπλέον από ένα αποτέλεσμα έδωσαν οι συνδυασμοί παραμέτρων node score threshold 0.5 και minimum cluster density 0.7, MCODE node score threshold 0.2 με minimum support threshold του SPICi στο 0.8, MCODE k-core threshold 3 με minimum support threshold 0.2 και k-core threshold 3, αλλά η τιμή του minimum support threshold στο 0.8.

NCMine vs SPICi	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
BioGRID_C.elegans_no_self_loops:NCMINEcliq08-SPICIDEN07	0.161147	**0.705271**	0.82195	0.698811	0.698811	0.699259	0.39938	0.698364
BioGRID_Chicken_Gallus_no_self_loops:NCMINEcliq04-SPICIDEN07	0.39562	0.367752	**0.244649**	0.647869	0.650899	0.716981	0.379192	0.590909
DIP_Celegans:NCMINEmerge04-SPICIDEN07	0.135543	0.272078	0.31608	0.613905	0.6141037	0.5986622	0.39449	0.629943
DIP_Drosophila:NCMINEcliq06-SPICIDEN07	0.257308	0.56326	0.365086	0.704012	0.704491	0.678977	0.504747	0.730965
DIP_Homo_Sapiens:NCMINEcliq04-SPICIDEN07	0.219858	0.390222	0.574586	0.700196	0.701913	0.752809	0.502713	0.654458
DIP_Mouse:NCMINEmerge04-SPICIDEN07	**0.399918**	0.517007	0.33189	0.6976	0.700659	0.638045	0.494042	0.769417
GAVIN_2002:NCMINEcliq04-SPICIDEN07	0.157952	0.394838	0.991956	0.712281	0.717284	0.807486	0.455686	0.637158
GAVIN_2006NCMINEcliq06-SPICIDEN07	0.198161	0.577252	1.118439	0.756307	0.760477	**0.844639**	0.503785	0.684703
IntAct_Alzheimers:NCMINEmerge08-SPICIDEN07	0.088124	0.291602	0.157037	0.592293	0.59387	0.552083	0.33103	0.638821
IntAct_Cancer:NCMINEcliq08-SPICISUP08	0.076614	0.169226	0.33606	0.576568	0.576598	0.582452	0.294772	0.570802
IntAct_Cardiac:NCMINEcliq08-SPICIDEN07	0.219023	0.451605	0.263112	0.614995	0.632593	0.498282	0.301664	**0.803109**
Intact_data_COVID19:NCMINEmerge08-SPICISUP08	0.154549	0.23194	0.198567	0.655632	0.664973	0.561798	0.374677	0.787097
IntAct_Parkinson:NCMINEcliq08-SPICISUP08	0.104411	0.175098	0.094028	0.575437	0.575613	0.561576	0.329496	0.59
Krogan_BIND_2006:NCMINEcliq06-SPICIDEN07	0.180406	0.560478	0.714967	0.686156	0.688527	0.74819	0.430846	0.633622
BioGRID_Rat_no_self_loops:NCMINEcliq06-SPICISUP08	0.127191	0.153311	0.305064	0.572696	0.572725	0.578498	0.333959	0.56701
BioGRID_Taurus_no_self_loops:NCMINEcliq04-SPICIDEN07	0.379625	0.468627	0.361088	**0.802898**	**0.803002**	0.816092	**0.552991**	0.790123

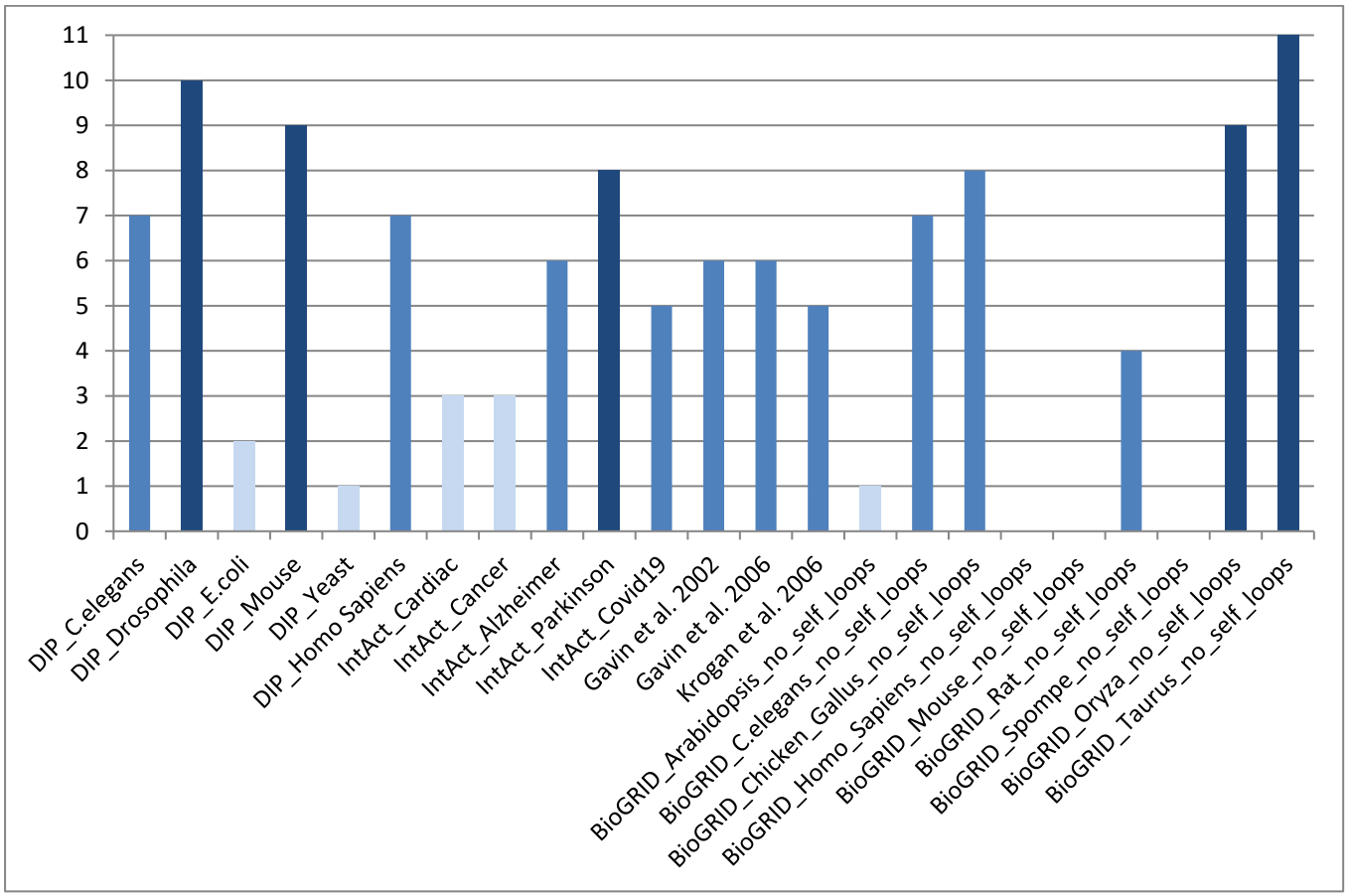
Πίνακας 21: Heatmap σύγκρισης αλγορίθμων NCMine και SPICi βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



Διάγραμμα 14: Ραβδόγραμμα αριθμού κόμβων και ακμών των δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών με σημαντική ομοιότητα στα αποτελέσματα των αλγορίθμων NCMine και SPICi

Τέλος μία σύγκριση αλγορίθμων που έδειξε από τις μεγαλύτερες ομοιότητες είναι των NCMine και SPICi. Όπως είναι εμφανές και στο παραπάνω γράφημα [Διάγραμμα 14], βρέθηκαν αξιόλογες τιμές μετρικών σε 16 από τα σύνολα, τα οποία περιλαμβάνουν ποικιλία αναλογιών κόμβων - ακμών. Ως προς τις παραμέτρους, με τις οποίες οι αλγόριθμοι παρουσιάζουν την μεγαλύτερη ομοιότητα, ξεχωρίζουν οι συνδυασμοί NCMine cliqueness threshold 0.4 με SPICi minimum cluster density 0.7 και cliqueness threshold 0.6 με minimum cluster density 0.7, με τέσσερα και τρία σύνολα αντίστοιχα. Από δύο καταγραφές στον πίνακα έχουν τα ζεύγη cliqueness threshold 0.8 στο NCMine με minimum cluster density 0.7 στο SPICi, cliqueness threshold 0.8 στο NCMine και minimum support threshold 0.8 στο SPICi, merge threshold 0.4 στο NCMine και minimum cluster density 0.7 στο SPICi. Οι υπόλοιπες καταγραφές αποτελούνται από συνδυασμούς που εμφανίζονται μόνο μια φορά στον πίνακα και συγκεκριμένα: NCMine merge threshold 0.8 – SPICi minimum cluster density 0.7, NCMine merge threshold 0.8 – SPICi minimum support threshold 0.8, NCMine cliqueness threshold 0.6 – SPICi minimum support threshold 0.8 [Πίνακας 21].

Συγκεντρωτικά τα αποτελέσματα δείχνουν μία σταθερή τάση ‘προτίμησης’ των μικρότερων συνόλων δεδομένων, αφού είναι χαρακτηριστικό πως τα μεγαλύτερα σύνολα δεν απέδωσαν καλές τιμές μετρικών σε καμία από τις συγκρίσεις των αλγορίθμων [Διάγραμμα 15]. Αντιθέτως, τα μικρότερα εμφανίστηκαν σχεδόν σε όλους τους πίνακες που αναλύθηκαν παραπάνω. Επιπλέον τρία σύνολα με αριθμό φορών εμφάνισης μεγαλύτερο του μέσου όρου είναι τα DIP_Drosophila, DIP_Mouse, IntAct_Parkinson, τα οποία, ενώ έχουν διαφορετικό αριθμό κόμβων και ακμών, έχουν παρόμοια Clustering Coefficient και Average Degree, όπως αυτά υπολογίστηκαν στο προηγούμενο στάδιο.



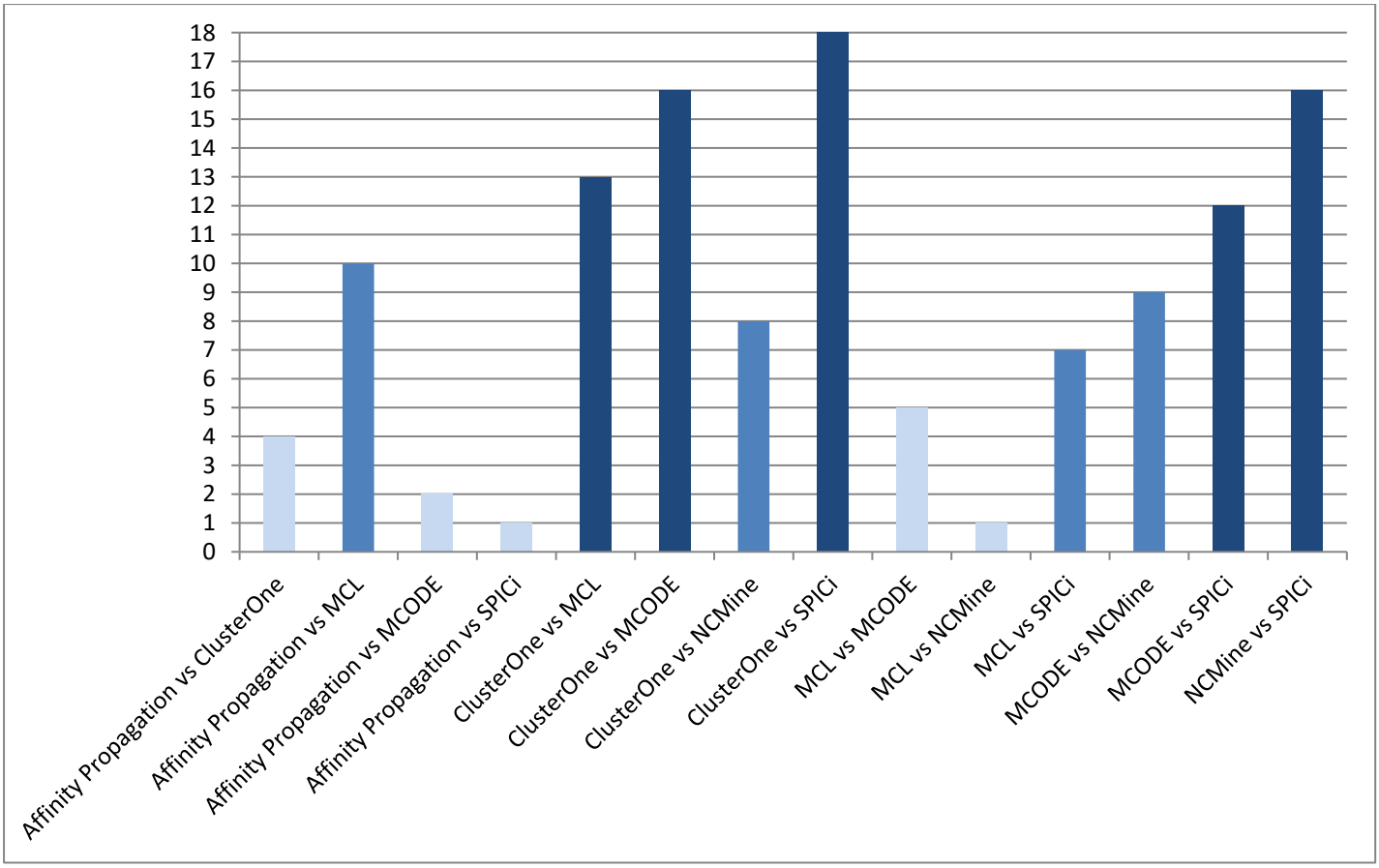
Διάγραμμα 15: Ραβδόγραμμα κατανομής συνόλων δεδομένων στα ζεύγη σύγκρισης αλγορίθμων με σημαντική ομοιότητα

Όσο αφορά τις μεθόδους ομαδοποίησης, οι οποίες κατά την σύγκριση παρουσίασαν μεγαλύτερη ομοιότητα, όπως φαίνεται και στα γραφήματα [Διάγραμμα 16 και 17] που ακολουθούν, ξεχωρίζουν τα ζευγάρια ClusterONE με SPICi, ClusterONE με MCODE, NCMine με SPICi, ClusterONE με MCL και MCODE με SPICi, ως προς τον αριθμό συνόλων στα οποία βρέθηκε ομοιότητα. Καθώς όμως μέσα στο πλήθος των συνόλων υπάρχουν και ορισμένα με μέτριες ή και κακές τιμές μετρικών, προκύπτει

ως αποτέλεσμα ο μέσος όρος των μέτρων σύγκρισης να βγαίνει πιο χαμηλός απ' ότι στις συγκρίσεις άλλων αλγορίθμων. Στον Πίνακα 22 έχουν συγκεντρωθεί για κάθε συνδυασμό αλγορίθμων ο αριθμός των συνόλων και ο μέσος όρος των μετρικών που ελέγχθηκαν. Ένα ακόμα ζεύγος αλγορίθμων, λοιπόν, που διακρίνεται μέσω αυτού, είναι οι MCL και MCODE. Οι δύο αλγόριθμοι, παρόλο που έχουν μόνο πέντε σύνολα με καλές τιμές στις μετρικές, έχουν τον ψηλότερο συνολικά μέσο όρο, αποδεικνύοντας την ύπαρξη σημαντικής ομοιότητας ανάμεσα τους.

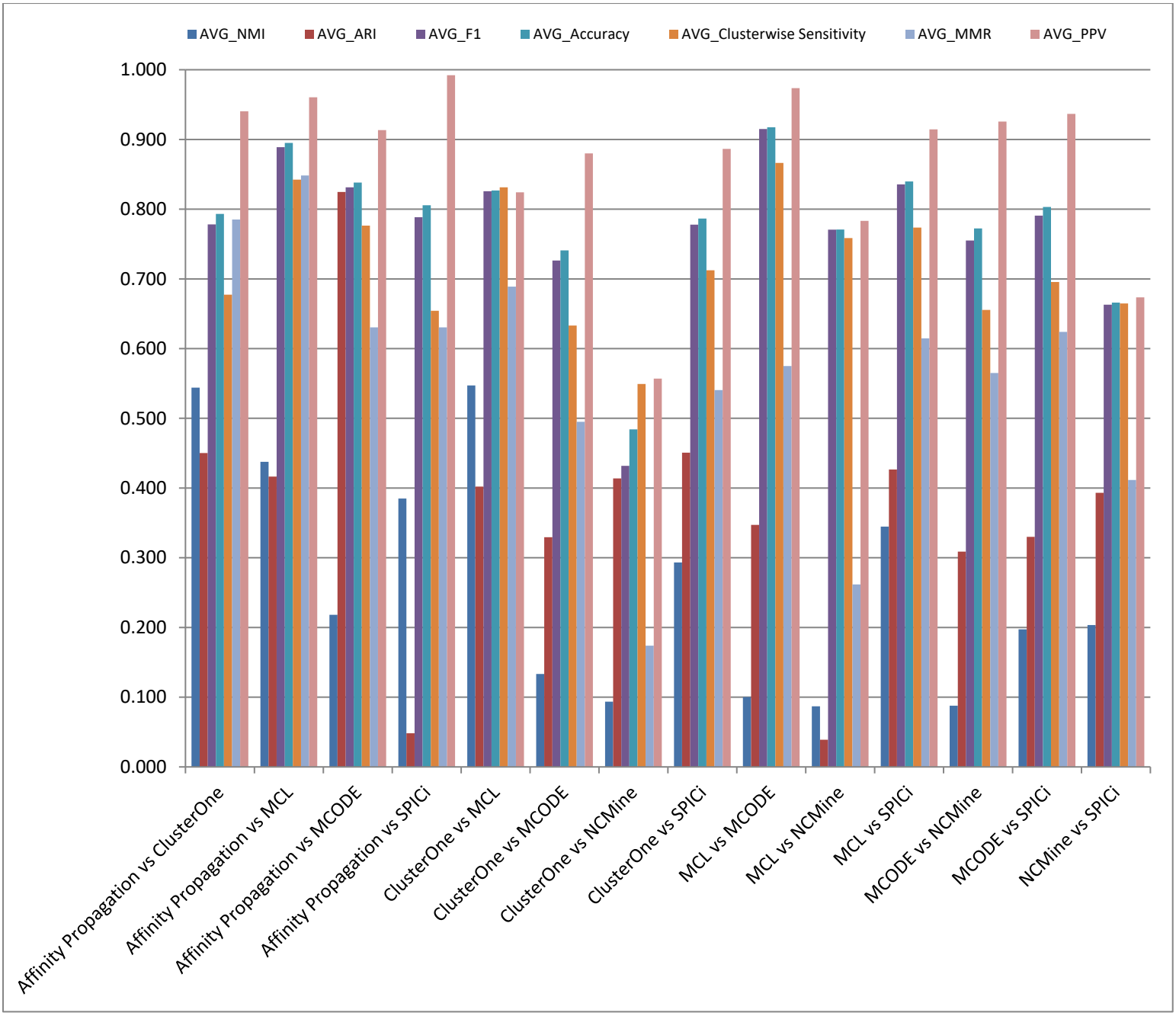
Pair of Comparing Algorithms	Number_of_Data Sets	AVG_NMI	AVG_ARI	AVG_VI	AVG_F1	AVG_Accuracy	AVG_Cluster-wise Sensitivity	AVG_MMR	AVG_PPV
Affinity Propagation vs ClusterONE	4	0.544	0.450	2.597	0.778	0.793	0.677	0.785	0.940
Affinity Propagation vs MCL	10	0.438	0.416	1.904	0.889	0.895	0.842	**0.848**	0.960
Affinity Propagation vs MCODE	2	0.218	**0.825**	1.829	0.831	0.838	0.776	0.630	0.913
Affinity Propagation vs SPICi	1	0.385	0.048	3.121	0.789	0.806	0.654	0.630	**0.992**
ClusterONE vs MCL	13	**0.547**	0.402	1.764	0.826	0.827	0.831	0.689	0.824
ClusterONE vs MCODE	16	0.133	0.329	0.836	0.726	0.741	0.633	0.495	0.880
ClusterONE vs NCMine	8	0.094	0.414	0.566	0.432	0.484	0.549	0.174	0.557
ClusterONE vs SPICi	**18**	0.293	0.451	0.924	0.778	0.787	0.712	0.541	0.887
MCL vs MCODE	5	0.100	0.347	1.974	**0.915**	**0.917**	**0.866**	0.575	**0.974**
MCL vs NCMine	1	0.087	0.039	2.440	0.771	0.771	0.759	0.262	0.783
MCL vs SPICi	7	0.345	0.427	1.700	0.835	0.840	0.774	0.615	0.914
MCODE vs NCMine	9	0.088	0.309	**0.264**	0.755	0.772	0.656	0.565	0.926
MCODE vs SPICi	12	0.197	0.330	0.518	0.791	0.803	0.695	0.624	0.937
NCMine vs SPICi	16	0.203	0.393	0.450	0.663	0.666	0.665	0.411	0.674

Πίνακας 22: Συγκεντρωτικός πίνακας αριθμού συνόλων και μέσων τιμών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



Διάγραμμα 16: Ραβδόγραμμα συνολικού αριθμού συνόλων δεδομένων για κάθε ζεύγος αλγορίθμων σύγκρισης

Μία τελευταία παρατήρηση, πάνω στην ανάλυση των αποτελεσμάτων σύγκρισης των αλγορίθμων ομαδοποίησης, αφορά τις μετρικές που χρησιμοποιήθηκαν. Όπως είχε σχολιαστεί και στο υποκεφάλαιο 3.1.2, τα μέτρα NMI, ARI, VI τείνουν να είναι πιο αυστηρά, ενώ όπως αποδείχθηκε στην παρούσα ανάλυση πιο επιεική φαίνεται να είναι τα μέτρα PPV, F1-score και Accuracy.



Διάγραμμα 17: Συγκεντρωτικό γράφημα μέσων βαθμών των μετρικών Normalized Mutual Information, Rand Index, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση όλων των αλγορίθμων μεταξύ τους

3.2 Σύγκριση αλγορίθμων ομαδοποίησης με σύμπλοκα συνόλου αναφοράς

Οι αλγόριθμοι, οι οποίοι εξετάζονται στην παρούσα εργασία, αξιολογήθηκαν, επιπλέον, και ως προς την αξιοπιστία των αποτελεσμάτων τους, μέσω της σύγκρισης αυτών με ένα σύνολο αναφοράς.

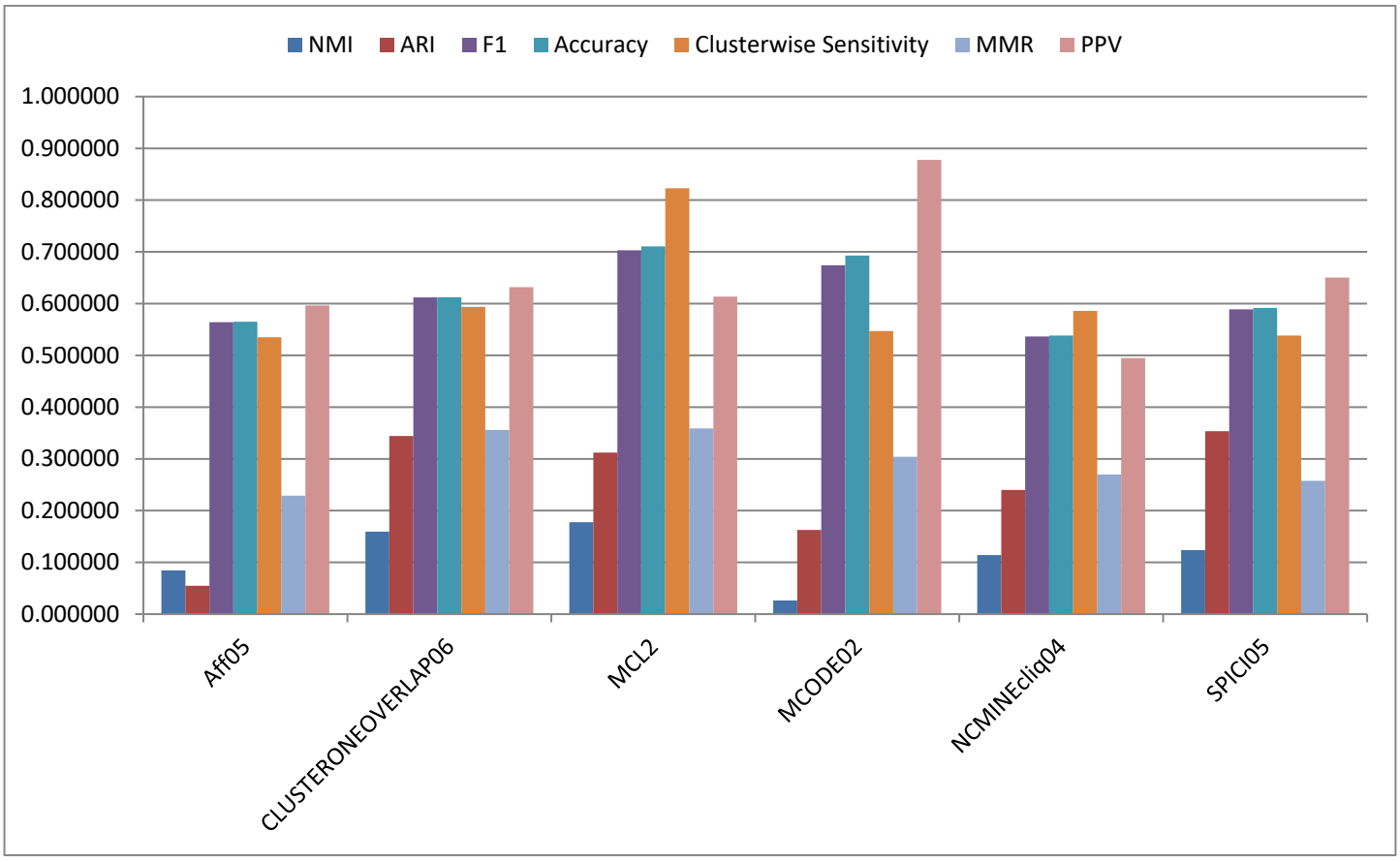
Αρχικά, για την πραγματοποίηση του ελέγχου, επιλέχθηκε ο οργανισμός προς μελέτη. Καθώς το μεγαλύτερο μέρος της σχετικής βιβλιογραφίας επικεντρώνεται στον ζυμομύκητα (Yeast), αποφασίστηκε να είναι το μοντέλο και αυτής της ανάλυσης. Τα σύνολα αλληλεπιδράσεων πρωτεϊνών συγκεκριμένα, πάνω στα οποία θα εκτελεστούν οι έξη αλγόριθμοι, είναι το Gavin et al. 2002, το Gavin et al. 2006, το Krogan et al. 2006 και ένα μεγάλο σύνολο με πρωτεϊνικές αλληλεπιδράσεις του ζυμομύκητα από την βάση BioGRID. Τα δίκτυα αυτά εισήχθηκαν στο Cytoscape και εν συνεχεία ομαδοποιήθηκαν με τους αλγορίθμους Affinity Propagation, ClusterONE, MCL, MCODE, NCMine και SPICi.

Η εκτέλεση των αλγορίθμων πραγματοποιήθηκε με την ίδια λογική, όπως στην προηγούμενη μέθοδο. Για κάθε αλγόριθμο έγινε εκτέλεση για όλα τα σύνολα δεδομένων και μέσω ελέγχου των αποτελεσμάτων πραγματοποιούταν τροποποίηση των μεταβλητών και επανεκτέλεση. Τα αποτελέσματα, με τα σύμπλοκα πρωτεϊνών που προέβλεψαν οι αλγόριθμοι, αξιολογήθηκαν βάση 408 αναγνωρισμένων συμπλόκων πρωτεϊνών που περιέχονται στον σύνολο αναφοράς CYC2008.

Για την εύρεση των μετρικών χρησιμοποιήθηκε και σε αυτήν την μέθοδο, το πρόγραμμα που κατασκευάστηκε στα πλαίσια της παρούσας εργασίας και υπολογίζει τα 7 στατιστικά μέτρα Normalized Mutual Information, Adjasted Rand Index, Variation of Information, F1-score, Geometric Accuracy, Cluster-wise Sensitivity, Maximum Matching Ratio και Positive Predictive Value, καθώς και το πρόγραμμα των McDaid, Greene και Hurley για τον υπολογισμό της Normalized Mutual Information. Συνολικά βρέθηκαν **12.096 τιμές**, από τις οποίες επιλέχθηκαν εκείνες που πληρούσαν τις προδιαγραφές που περιγράφηκαν κατά την προηγούμενη μέθοδο, δηλαδή να βρίσκεται τουλάχιστον μια από τις μετρικές άνω των καθορισμένων κατωφλίων και από τις εναπομείναντες να διατηρηθεί για κάθε αλγόριθμο και σύνολο δεδομένων η καταγραφή με τις βέλτιστες συνολικά τιμές. Το σύνολο αυτών των καταγραφών παρουσιάζονται συγκεντρωτικά στους ακόλουθους heatmaps.

Gavin 2002	NMI	ARI	VI	F1-score	Accuracy	Cluster-wise Sensitivity	MMR	PPV
AFFLamda05	0.084491	0.054920	3.647672	0.563999	0.564836	0.534884	0.228964	0.596465
CLUSTERONEOVERLAP06	0.159402	0.344076	1.763473	0.612014	0.612309	0.593577	0.355854	0.631633
MCLINFL2	**0.177690**	0.312171	2.362740	**0.702994**	**0.710569**	**0.822813**	**0.358957**	0.613636
MCODENODETHRES02	0.026422	0.162799	**1.503025**	0.674013	0.692949	0.547008	0.304181	**0.877828**
NCMINEcliq04	0.114256	0.240030	1.608072	0.536550	0.538458	0.585825	0.270026	0.494920
SPICIDEN05	0.123820	**0.353378**	1.581220	0.588982	0.591621	0.538206	0.257658	0.650338

Πίνακας 23: Heatmap σύγκρισης των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Gavin 2002 με το σύνολο αναφοράς βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



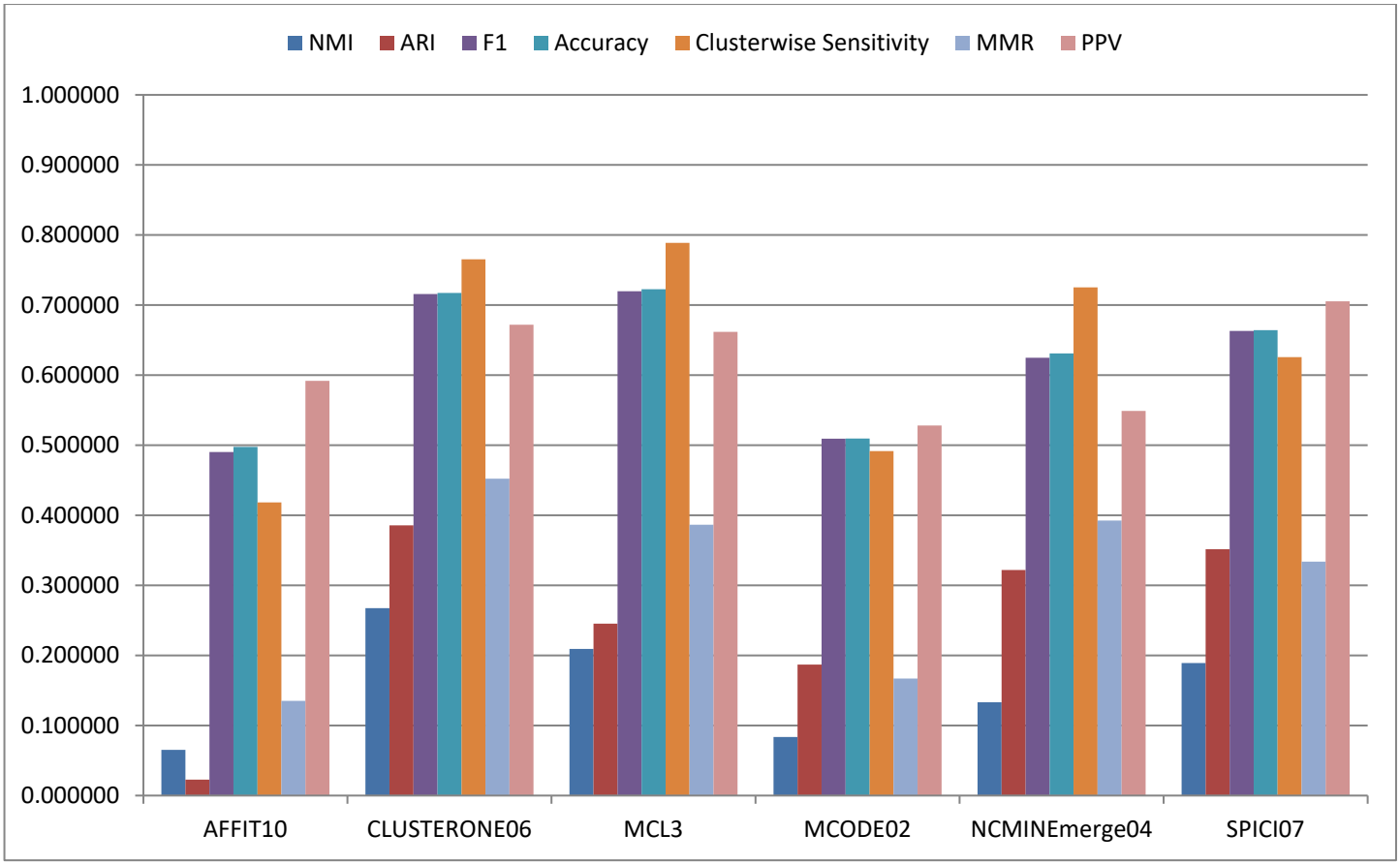
Διάγραμμα 18: Ραβδόγραμμα τιμών μετρικών Normalized Mutual Information, Rand Index, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Gavin 2002 με το σύνολο αναφοράς

Η σύγκριση των αποτελεσμάτων των αλγορίθμων για το σύνολο δεδομένων των Gavin et al. 2002 με το σύνολο αναφοράς απέδωσε καλύτερες τιμές μετρικών με τον αλγόριθμο MCL και συγκεκριμένα με την παράμετρο Inflamation value αυτού στο

0.2, ενώ με παρόμοια, λίγο χαμηλότερα, ποσά εμφανίζονται τα αποτελέσματα του MCODE με node score threshold στο 0.2. Επιπλέον οι αλγόριθμοι Affinity Propagation και NCMine παρόλο που εμφανίζεται στον πίνακα η καταγραφή με τις υψηλότερες τιμές αυτών, δεν ξεπερνάνε τα απαραίτητα κατώφλια, ώστε να θεωρηθεί ότι η ομαδοποίηση είναι αξιόπιστη [Πίνακας 23][Διάγραμμα 18].

Gavin 2006	NMI	ARI	VI	F1-score	Accuracy	Cluster-wise Sensitivity	MMR	PPV
AFFIT10	0.065247	0.022643	4.435453	0.490215	0.497605	0.418377	0.135065	0.591837
CLUSTERONEDEN06	**0.267434**	**0.385697**	1.732622	0.715691	0.717205	0.765396	**0.452179**	0.672048
MCLINFL3	0.209249	0.245401	2.291980	**0.719801**	**0.722575**	**0.788856**	0.386607	0.661863
MCODENODETHRES02	0.083515	0.186971	1.948338	0.509278	0.509604	0.491691	0.167071	0.528169
NCMINEmerge04	0.133151	0.321830	**1.581044**	0.624867	0.630947	0.725318	0.392541	0.548855
SPICIDEN07	0.189191	0.351524	1.678182	0.663060	0.664251	0.625611	0.333878	**0.705276**

Πίνακας 24: Heatmap σύγκρισης των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Gavin 2006 με το σύνολο αναφοράς βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



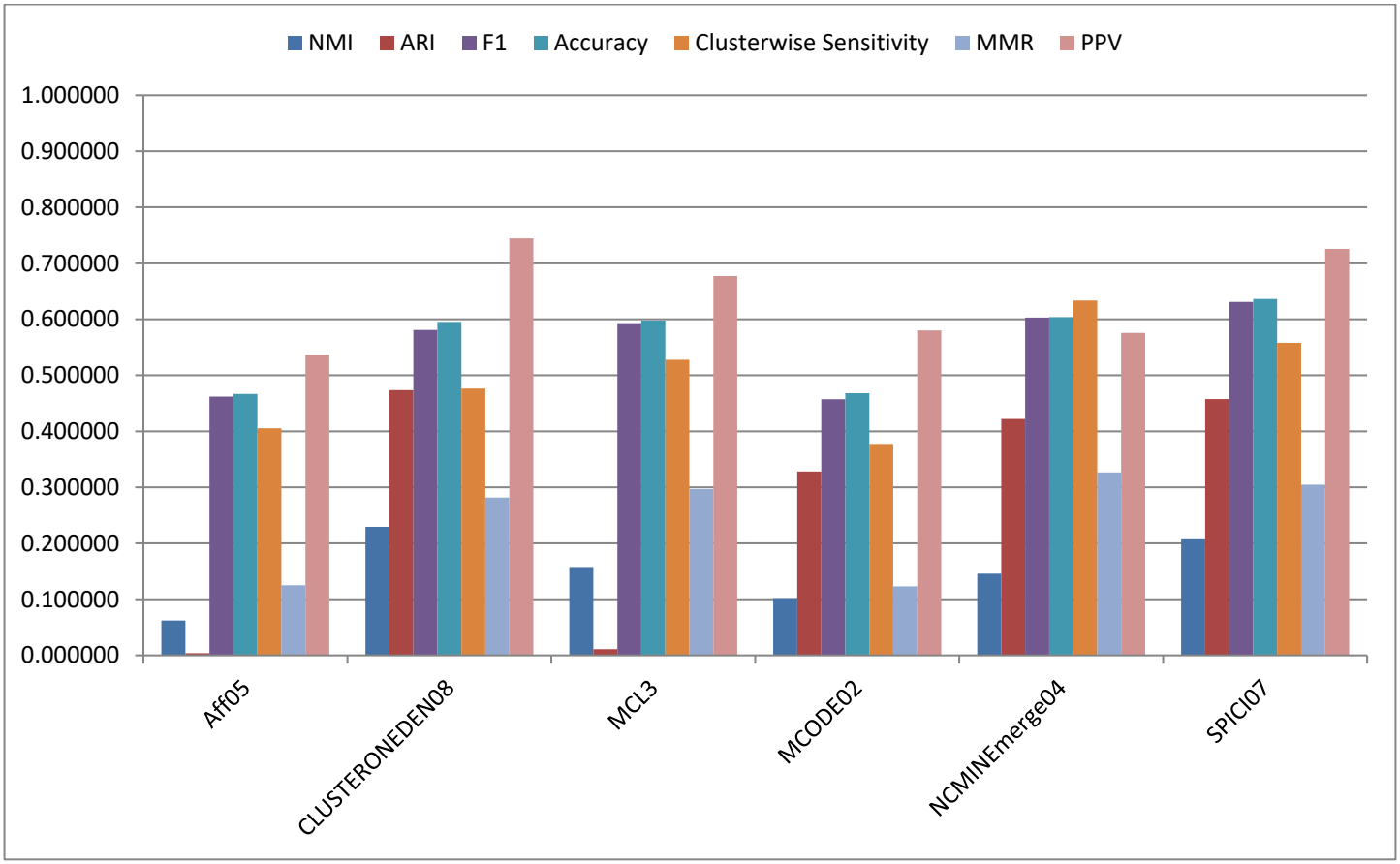
Διάγραμμα 19: Ραβδόγραμμα τιμών μετρικών Normalized Mutual Information, Rand Index, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Gavin 2006 με το σύνολο αναφοράς

Ακολούθως, το σύνολο που προτάθηκε δύο χρόνια αργότερα, το 2006, από την ερευνητική ομάδα της Gavin παρουσιάζει, όμοια με πριν, μια καλύτερη εικόνα κατά τον έλεγχο του με τον αλγόριθμο MCL, όμως με την παράμετρο Inflamation value στο 0.3 [Διάγραμμα 19]. Παράλληλα ο αλγόριθμος που ακολουθεί σε ποιότητα αποτελεσμάτων είναι ο ClusterONE με minimum density 0.6. Μία ακόμα ομοιότητα με το παλιότερο σύνολο των ερευνητών είναι πως και σε αυτή την περίπτωση ο αλγόριθμος Affinity Propagation δεν κατάφερε να αποδώσει τιμές ανώτερες των κατωφλίων, όπως και ο αλγόριθμος MCODE [Πίνακας 24].

Krogan 2006	NMI	ARI	VI	F1-score	Accuracy	Cluster-wise Sensitivity	MMR	PPV
AFFLamda05	0.062184	0.003921	5.422547	0.461900	0.466444	0.405447	0.124946	0.536616
CLUSTERONEDEN 08	**0.229403**	**0.473580**	**0.787632**	0.580867	0.595406	0.476265	0.281693	**0.744352**
MCLINFL3	0.157782	0.010873	4.153910	0.593104	0.597725	0.527626	0.296890	0.677136

MCODENODETHRES02	0.102076	0.328177	0.914802	0.457297	0.467892	0.377432	0.122983	0.580034
NCMINEmerge04	0.145716	0.421899	0.905500	0.603127	0.603820	**0.633463**	**0.326475**	0.575564
SPICIDEN07	0.208781	0.457440	0.946870	**0.630882**	**0.636337**	0.557977	0.304759	0.725702

Πίνακας 25: Heatmap σύγκρισης των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Krogan 2006 με το σύνολο αναφοράς βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value

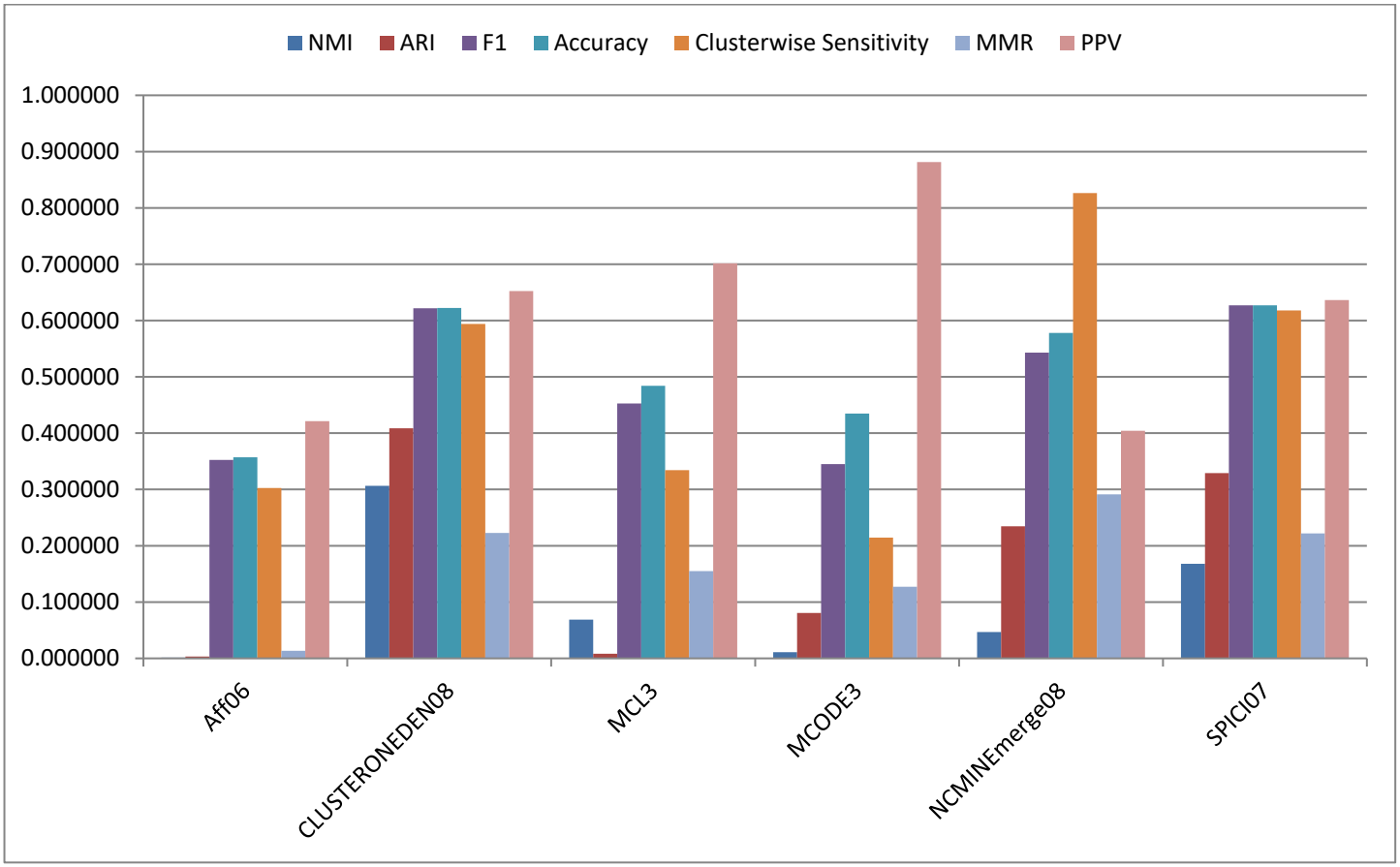


Διάγραμμα 20: Ραβδόγραμμα τιμών μετρικών Normalized Mutual Information, Rand Index, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο Krogan 2006 με το σύνολο αναφοράς

Σε αντίθεση με τον έλεγχο των συνόλων της Gavin, η αξιολόγηση των αποτελεσμάτων των αλγορίθμων για το δίκτυο πρωτεϊνών του Krogan κατάταξε τον MCL στις τελευταίες θέσεις, αφού υπερτερεί μόνο των Affinity Propagation και MCODE οι οποίοι, δεν ξεπερνάνε ούτε σε αυτή την περίπτωση τα όρια [Πίνακας 25]. Πολύ καλά αποτελέσματα φαίνεται να λαμβάνονται μέσω της μεθόδου ClusterONE για την παράμετρο minimum density 0.8, αφού στις μισές μετρικές έχει την βέλτιστη τιμή. Γενικότερα το συγκεκριμένο σύνολο δεν απέδωσε υψηλές τιμές σε κανένα μέτρο ελέγχου [Διάγραμμα 20].

BioGRID	NMI	ARI	VI	F1	Accuracy	Cluster-wise Sensitivity	MMR	PPV
AFFLamda06	0.001412	0.003218	4.297082	0.352323	0.357149	0.302728	0.013492	0.421353
CLUSTERONEDEN08	**0.306543**	**0.408678**	**0.824773**	0.621769	0.622454	0.593914	0.222660	0.652366
MCLINFL3	0.068939	0.008357	4.308880	0.452737	0.484213	0.334208	0.155064	0.701546
MCODEKCORE3	0.010898	0.080616	1.699340	0.345031	0.434832	0.214493	0.127304	**0.881517**
NCMINEmerge08	0.046875	0.234705	0.991809	0.542918	0.577977	**0.826337**	**0.291193**	0.404263
SPICIDEN07	0.167980	0.329026	1.195366	**0.627018**	**0.627084**	0.618048	0.221881	0.636251

Πίνακας 26: Heatmap σύγκρισης των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο του οργανισμού Yeast από την βάση BioGRID με το σύνολο αναφοράς βάση των μετρικών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value



Διάγραμμα 21: Ραβδόγραμμα τιμών μετρικών Normalized Mutual Information, Rand Index, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση των αποτελεσμάτων των 6 αλγορίθμων για το σύνολο του οργανισμού Yeast από την βάση BioGRID με το σύνολο αναφοράς

Τέλος έγινε έλεγχος της αξιοπιστίας των αποτελεσμάτων με είσοδο τις πρωτεϊνικές αλληλεπιδράσεις του ζυμομύκητα, όπως αυτές εμπεριέχονταν στην βάση

δεδομένων BioGRID [Διάγραμμα 21]. Ο αλγόριθμος με την συνολικά καλύτερη εικόνα είναι, όπως και σε προηγούμενες περιπτώσεις, ο ClusterONE με minimum density 0.8. Εξίσου αξιόλογα αποτελέσματα προέκυψαν και από τον έλεγχο των αλγορίθμων NCMine με merge threshold 0.8 και SPICi με minimum cluster density 0.7, ιδιαίτερα για τις μετρικές, που εξετάζουν επικαλύψεις συνόλων. Αποτελέσματα, τα οποία είναι κάτω των προκαθορισμένων ορίων βρέθηκαν με τους αλγορίθμους Affinity Propagation και MCL [Πίνακας 26].

Συγκεντρωτικά, μεγαλύτερη αξιοπιστία αποτελεσμάτων, τόσο στον έλεγχο μεμονωμένα των συνόλων, όσο και στον μέσο όρο αυτών, όπως φαίνεται και στον Πίνακα 27, εμφανίζεται στον αλγόριθμο ClusterONE. Ο ClusterONE επιτυγχάνει ποσοστά άνω του μετρίου σε όλα τα σύνολα που εξετάστηκαν, σε αντίθεση με τον Affinity Propagation, ο οποίος σε όλους τους ελέγχους βρέθηκε ανεπαρκής. Παράλληλα, όσο αφορά τους υπόλοιπους αλγορίθμους, τα αποτελέσματα του MCL μπορούν να θεωρηθούν αμφιλεγόμενα, αφού στις μισές περιπτώσεις πραγματοποιούν την βέλτιστη ομαδοποίηση και στις μισές κάκιστη. Ικανοποιητικά, και αρκετά όμοια μεταξύ τους, μπορούν να αξιολογηθούν τα αποτελέσματα των μεθόδων NCMine και SPICi, γεγονός που συνάδει με το εύρημα της προηγούμενης μεθόδου, πως οι δύο αλγόριθμοι έχουν την υψηλότερη ομοιότητα αποτελεσμάτων.

Algorithm	AVG_NMI	AVG_ARI	AVG_VI	AVG_F1-score	AVG_Accuracy	Cluster-wise Sensitivity	AVG_MMR	AVG_PPV
Affinity Propagation	0.0533	0.0212	4.4507	0.4671	0.4715	0.4154	0.1256	0.5366
ClusterONE	**0.2407**	**0.4030**	1.2771	**0.6326**	**0.6368**	0.6073	**0.3281**	0.6751
MCL	0.1534	0.1442	3.2794	0.6172	0.6288	0.6184	0.2994	0.6635
MCODE	0.0557	0.1896	1.5164	0.4964	0.5263	0.4077	0.1804	**0.7169**
NCMine	0.1100	0.3046	**1.2716**	0.5769	0.5878	**0.6927**	0.3201	0.5059
SPICi	0.1724	0.3728	1.3504	0.6275	0.6298	0.5850	0.2795	0.6794

Πίνακας 27: Συγκεντρωτικός πίνακας μέσων τιμών Normalized Mutual Information, Rand Index, Variation of Information, F1-score, Cluster-wise Sensitivity, Geometric Accuracy, Maximum Matching Ratio και Positive Predictive Value για την σύγκριση των αποτελεσμάτων όλων των αλγορίθμων με το σύνολο αναφοράς

4. Συμπεράσματα

Οι αλληλεπιδράσεις μεταξύ πρωτεϊνών είναι ουσιώδους σημασίας σχεδόν σε όλες τις βιολογικές διεργασίες ενός κυττάρου. Η κατανόηση, επομένως, των αλληλεπιδράσεων αυτών είναι απαραίτητη στην διαλεύκανση της φυσιολογίας των ζωντανών οργανισμών, στην μελέτη των ασθενειών, αλλά και στην ανάπτυξη νέων φαρμακευτικών μεθόδων.

Τα τελευταία χρόνια, χάρη στην συνεχώς αναπτυσσόμενη τεχνολογία, έχει υπάρξει εκτεταμένη αύξηση των βιολογικών δεδομένων που είναι διαθέσιμα. Το πλήθος των δεδομένων αυτών οδήγησαν στην ανάπτυξη πιο συστηματικών μεθόδων για την ανάλυση των πληροφοριών, όπως τα βιολογικά δίκτυα αλληλεπιδράσεων, τα οποία επιτρέπουν την απεικόνιση και ανάλυση των βιολογικών συστημάτων, μέσω της χρήσης τεχνικών που πηγάζουν από την θεωρία γράφων. Τα δίκτυα αλληλεπίδρασης πρωτεΐνης - πρωτεΐνης (PPINs) αποτελούν μαθηματικές απεικονίσεις των φυσικών συσχετίσεων των πρωτεϊνών στο κύτταρο. Η ανάπτυξη τέτοιων δικτύων μπορεί να χρησιμοποιηθεί στην εκχώρηση πιθανών ρόλων σε μη χαρακτηρισμένες πρωτεΐνες, στον πιο λεπτομερή προσδιορισμό των σταδίων διαφόρων βιολογικών μονοπατιών, καθώς και στον χαρακτηρισμό των σχέσεων μεταξύ πρωτεϊνών που σχηματίζουν πολύπλοκα σύμπλοκα.

Το πλήθος των πιθανοτήτων αξιοποίησης δεδομένων από τα δίκτυα αλληλεπιδράσεων πρωτεϊνών, έχει οδηγήσει στην δημοσίευση δεκάδων ερευνών, οι οποίες προτείνουν μεθόδους ομαδοποίησης των πρωτεϊνών, για την δημιουργία συμπλόκων με πιθανή βιολογική σημασία. Στην παρούσα εργασία έγινε προσπάθεια σύγκρισης και αξιολόγησης ορισμένων εξ αυτών, αποσκοπώντας στην ανάδειξη τόσο της πιο αξιόπιστης τεχνικής, όσο και πιθανών ομοιοτήτων ανάμεσα στα αποτελέσματα των διαφόρων αλγορίθμων ομαδοποίησης.

Καθοριστικό ρόλο στην αξιοπιστία των αποτελεσμάτων διαδραματίζει η επιλογή των συνόλων δεδομένων, τα οποία θα αποτελέσουν την είσοδο στις μεθόδους. Με στόχο να καλυφθεί το μεγαλύτερο δυνατό πλήθος περιπτώσεων, απομονώθηκαν 24 δίκτυα αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών, από διάφορες βάσεις δεδομένων, από ποικίλους οργανισμούς, καθώς και με ένα ευρύ φάσμα διαφορετικών τοπολογικών χαρακτηριστικών. Τα δίκτυα αυτά κατασκευάστηκαν, με την βοήθεια της πλατφόρμας του Cytoscape και αναλύθηκαν εκτενώς, αφού υπολογίστηκαν 12 περιγραφικές μετρικές για κάθε ένα από αυτά.

Το κύριο μέρος της διπλωματικής εργασίας, η εκτέλεση και η αξιολόγηση δηλαδή των αλγορίθμων, ξεκίνησε με την επιλογή αυτών, ύστερα από ανασκόπηση της

βιβλιογραφίας αλλά και βάση της διαθεσιμότητας τους. Έξι διαφορετικοί αλγόριθμοι ομαδοποίησης δικτύων επιλέχθηκαν και συγκεκριμένα οι Affinity Propagation, ClusterONE, MCL, MCODE, NCMine και SPICi.

Για την σύγκριση της ομοιότητας των αποτελεσμάτων των αλγορίθμων μεταξύ τους, εκτελέστηκαν αυτοί με είσοδο 23 σύνολα αλληλεπιδράσεων πρωτεϊνών, για διάφορες παραμέτρους του κάθε αλγορίθμου, καταλήγοντας σε ένα σύνολο 623 αποτελεσμάτων, όπου κάθε ένα περιείχε πλήθος πιθανών συμπλόκων πρωτεϊνών. Τα αποτελέσματα αυτά συγκρίθηκαν ανά ζεύγη με την εξαγωγή 8 στατιστικών μέτρων, των Normalized Mutual Information, Variation of Information, Adjusted Rand Index, F1-score, Geometric Accuracy, Cluster-wise Sensitivity, Positive Predictive Value, Maximum Matching Ratio. Οι 127.211 τιμές, που προέκυψαν, συνοψίστηκαν, ώστε για κάθε συνδυασμό αλγορίθμων σε συγκεκριμένο σύνολο πρωτεϊνών, να διατηρηθεί αυτό με τις υψηλότερες τιμές. Συνολικά σε 14 ζεύγη αλγορίθμων βρέθηκαν 126 περιπτώσεις με σημαντική ομοιότητα μεταξύ των αποτελεσμάτων. Ο μεγαλύτερος αριθμός αυτών των καταγραφών αφορούσε την σύγκριση των αλγορίθμων ClusterONE με SPICi, ClusterONE με MCODE, NCMine με SPICi, ClusterONE με MCL και MCODE με SPICi, υποδεικνύοντας πως οι συγκεκριμένοι αλγόριθμοι δίνουν παρόμοια αποτελέσματα σε ένα πλήθος διαφορετικών συνόλων δεδομένων. Απουσία ομοιότητας, για όλα τα σύνολα που εξετάστηκαν σημειώθηκε ανάμεσα στους αλγορίθμους Affinity Propagation και NCMine. Αξίζει επιπλέον να σημειωθεί η σαφή προτίμηση που υπήρχε απέναντι στα μικρότερα σύνολα δεδομένων εισόδου, έναντι των μεγαλύτερων, τα οποία δεν απέδωσαν αξιόλογες τιμές για κανένα συνδυασμό αλγορίθμων.

Επιπροσθέτως, πραγματοποιήθηκε έλεγχος της αξιοπιστίας των αποτελεσμάτων, μέσω της σύγκρισης με γνωστά σύμπλοκα από σύνολο αναφοράς. Το σύνολο αυτό επιλέχθηκε να είναι το CYC2008, το οποίο περιέχει 408 διαφορετικά σύμπλοκα πρωτεϊνών – πρωτεϊνών του πρότυπου οργανισμού Yeast. Το CYC2008 συγκρίθηκε με τις πρωτεϊνικές ομάδες που προέβλεψαν οι 6 αλγόριθμοι για 4 διαφορετικά σύνολα από τον οργανισμό Yeast, τα οποία είναι των Gavin et al. του 2002, των Gavin et al. του 2006, των Krogan et al. του 2006 και το σύνολο πρωτεϊνικών αλληλεπιδράσεων του ζυμομύκητα στην βάση δεδομένων BioGRID. Όπως και προηγουμένως, οι αλγόριθμοι εκτελέστηκαν για κάθε σύνολο με ένα πλήθος παραμέτρων και ελέγχθηκαν με τις 8 μετρικές σύγκρισης καταλήγοντας σε 12.096 στατιστικές τιμές. Τα δεδομένα αυτά ελέγχθηκαν και καταγράφηκαν σε συγκεντρωτικούς πίνακες, όπου για κάθε σύνολο δεδομένων και με κάθε αλγόριθμο σύγκρισης παρουσιάζονται τα αποτελέσματα με τις καλύτερες συνολικά τιμές. Μέσω του ελέγχου αυτού διακρίθηκε ο αλγόριθμος ClusterONE, αφού στην πλειοψηφία των δοκιμών απέδωσε καλύτερα. Αντιθέτως, αποδείχθηκε πως και στα τέσσερα σύνολα ο αλγόριθμος Affinity Propagation απέτυχε να ομαδοποιήσει ορθά τις πρωτεΐνες.

Τέλος, μέσω των 139.307 υπολογισμών στατιστικών μέτρων, που πραγματοποιήθηκαν, τόσο κατά την σύγκριση των αλγορίθμων, όσο και κατά τον έλεγχο τους, μπορούν να εξαχθούν ορισμένα, βασικά συμπεράσματα για τις μετρικές σύγκρισης. Τα μέτρα τα οποία φαίνεται να κρίνουν με μεγαλύτερη αυστηρότητα τα αποτελέσματα των αλγορίθμων είναι τα Normalized Mutual Information, Variation of Information, και Adjusted Rand Index, ενώ πιο επιεική παρουσιάζονται το F1-score, Geometric Accuracy και Positive Predictive Value. Το συγκεκριμένο συμπέρασμα συγκλίνει με παλιότερες μελέτες πάνω στην ανάλυση μέτρων σύγκρισης και αξιολόγησης αλγορίθμων ομαδοποίησης. Ο υπολογισμός τόσο των παραπάνω, όσο και των υπολοίπων μετρικών, που προαναφέρθηκαν, μπορεί να αποδώσει μια πιο πλήρη εικόνα για τους αλγορίθμους, ενώ παράλληλα γίνεται αποφυγή του αποκλεισμού αποτελεσμάτων λόγω ψευδών αρνητικών ή την ενσωμάτωση ψευδών θετικών περιπτώσεων.

Τα συμπεράσματα, τα οποία προέκυψαν από την παρούσα διπλωματική, μπορούν να αποτελέσουν χρήσιμη πληροφορία κατά την έναρξη μελλοντικών μελετών αλληλεπιδράσεων πρωτεϊνών. Επισημαίνοντας τις βασικές ιδιότητες ορισμένων από τους πιο δημοφιλής αλγορίθμους ομαδοποίησης δικτύων αλληλεπιδράσεων πρωτεϊνών - πρωτεϊνών και παράλληλα προτείνοντας βέλτιστες πρακτικές επιλογής τόσο δεδομένων εισόδου, όσο και μέτρων αξιολόγησης, δίνεται ένα υπόδειγμα εκτέλεσης αντίστοιχων πειραμάτων, τα αποτελέσματα των οποίων μπορούν να βοηθήσουν στην αποσαφήνιση βασικών βιολογικών ερωτημάτων.

5. Βιβλιογραφία

1. Demetrius, L. & Manke, T. Robustness and network evolution—an entropic principle. *Phys. Stat. Mech. Its Appl.* **346**, 682–696 (2005).
2. Leclerc, R. D. Survival of the sparsest: robust gene networks are parsimonious. *Mol. Syst. Biol.* **4**, 213 (2008).
3. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData Min.* **4**, 10 (2011).
4. Tran, T.-D. & Kwon, Y.-K. The relationship between modularity and robustness in signalling networks. *J. R. Soc. Interface* **10**, 20130771 (2013).
5. Johnson, D. B. A Note on Dijkstra's Shortest Path Algorithm. *J. ACM* **20**, 385–388 (1973).
6. Hadlock, F. O. A shortest path algorithm for grid graphs. *Networks* **7**, 323–334 (1977).
7. Jonker, R. & Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **38**, 325–340 (1987).
8. Huang, B., Wu, Q. & Zhan, F. B. A shortest path algorithm with novel heuristics for dynamic transportation networks. *Int. J. Geogr. Inf. Sci.* **21**, 625–644 (2007).
9. Chen, L. *et al.* Application of the Shortest Path Algorithm for the Discovery of Breast Cancer-Related Genes. *Curr. Bioinforma.* **11**, 51–58 (2016).
10. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1**, 215–239 (1978).
11. Freeman, L. C., Roeder, D. & Mulholland, R. R. Centrality in social networks: ii. experimental results. *Soc. Netw.* **2**, 119–141 (1979).
12. Bonacich, P. Power and Centrality: A Family of Measures. *Am. J. Sociol.* **92**, 1170–1182 (1987).
13. Friedkin, N. E. Theoretical Foundations for Centrality Measures. *Am. J. Sociol.* **96**, 1478–1504 (1991).
14. Hage, P. & Harary, F. Eccentricity and centrality in networks. *Soc. Netw.* **17**, 57–63 (1995).
15. Koschützki, D. & Schreiber, F. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regul. Syst. Biol.* **2**, 193–201 (2008).
16. Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How Correlated Are Network Centrality Measures? *Connect. Tor. Ont* **28**, 16–26 (2008).
17. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **8**, (2020).
18. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7821–7826 (2002).

19. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci.* **101**, 2658–2663 (2004).
20. Dunn, R., Dudbridge, F. & Sanderson, C. M. The Use of Edge-Betweenness Clustering to Investigate Biological Function in Protein Interaction Networks. *BMC Bioinformatics* **6**, 39 (2005).
21. Yoon, J., Blumer, A. & Lee, K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinforma. Oxf. Engl.* **22**, 3106–3108 (2006).
22. Hwang, W., Cho, Y.-R., Zhang, A. & Ramanathan, M. A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol. Biol.* **1**, 24 (2006).
23. Özgür, A., Vu, T., Erkan, G. & Radev, D. R. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* **24**, i277–i285 (2008).
24. Doncheva, N. T., Assenov, Y., Domingues, F. S. & Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **7**, 670–685 (2012).
25. Tadaka, S. & Kinoshita, K. NCMine: Core-peripheral based functional module detection using near-clique mining. *Bioinformatics* **32**, 3454–3460 (2016).
26. Liu, X. *et al.* Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinform.* **21**, 486–497 (2020).
27. Junker, B. H. & Falk, S. Analysis of Biological Networks | Wiley. *John Wiley Sons* **2**, (2011).
28. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
29. Phizicky, E. M. & Fields, S. Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94–123 (1995).
30. Jones, S. & Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13–20 (1996).
31. Marcotte, E. M. *et al.* Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science* **285**, 751–753 (1999).
32. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623–627 (2000).
33. Nooren, I. M. A. & Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492 (2003).
34. De Las Rivas, J. & Fontanillo, C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* **6**, e1000807 (2010).
35. Estojak, J., Brent, R. & Golemis, E. A. Correlation of two-hybrid affinity data with in vitro measurements. *Mol. Cell. Biol.* **15**, 5820–5829 (1995).

36. Deane, C. M., Salwiński, Ł., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics MCP* **1**, 349–356 (2002).
37. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
38. Puig, O. *et al.* The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods San Diego Calif* **24**, 218–229 (2001).
39. Knuesel, M. *et al.* Identification of novel protein-protein interactions using a versatile mammalian tandem affinity purification expression system. *Mol. Cell. Proteomics MCP* **2**, 1225–1233 (2003).
40. Guerrero, C., Tagwerker, C., Kaiser, P. & Huang, L. An integrated mass spectrometry-based proteomic approach: quantitative analysis of tandem affinity-purified in vivo cross-linked protein complexes (QTAX) to decipher the 26 S proteasome-interacting network. *Mol. Cell. Proteomics MCP* **5**, 366–378 (2006).
41. Bürckstümmer, T. *et al.* An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods* **3**, 1013–1019 (2006).
42. Fernández, E. *et al.* Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Mol. Syst. Biol.* **5**, 269 (2009).
43. Xu, X. *et al.* The tandem affinity purification method: an efficient system for protein complex purification and protein interaction identification. *Protein Expr. Purif.* **72**, 149–156 (2010).
44. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
45. Hoffmann, E. de. Mass Spectrometry. in *Kirk-Othmer Encyclopedia of Chemical Technology* (American Cancer Society, 2005). doi:10.1002/0471238961.1301191913151518.a01.pub2.
46. Russell, R. B. *et al.* A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**, 313–324 (2004).
47. Yugandhar, K., Gupta, S. & Yu, H. Inferring Protein-Protein Interaction Networks From Mass Spectrometry-Based Proteomic Approaches: A Mini-Review. *Comput. Struct. Biotechnol. J.* **17**, 805–811 (2019).
48. Lueking, A. *et al.* Protein Microarrays for Gene Expression and Antibody Screening. *Anal. Biochem.* **270**, 103–111 (1999).
49. MacBeath, G. & Schreiber, S. L. Printing Proteins as Microarrays for High-Throughput Function Determination. *Science* **289**, 1760–1763 (2000).
50. Sutandy, F. X. R., Qian, J., Chen, C.-S. & Zhu, H. Overview of protein microarrays. *Curr. Protoc. Protein Sci.* **Chapter 27**, Unit 27.1 (2013).

51. Ekins, R. P. Multi-analyte immunoassay. *J. Pharm. Biomed. Anal.* **7**, 155–168 (1989).
52. Nealon, J. O., Philomina, L. S. & McGuffin, L. J. Predictive and Experimental Approaches for Elucidating Protein-Protein Interactions and Quaternary Structures. *Int. J. Mol. Sci.* **18**, 2623 (2017).
53. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
54. Shoemaker, B. A. & Panchenko, A. R. Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Comput. Biol.* **3**, e43 (2007).
55. Lalonde, S. *et al.* Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J. Cell Mol. Biol.* **53**, 610–635 (2008).
56. Zuiderweg, E. R. P. Mapping Protein−Protein Interactions in Solution by NMR Spectroscopy. *Biochemistry* **41**, 1–7 (2002).
57. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
58. Burley, S. K. *et al.* Structural genomics: Beyond the Human Genome Project. *Nat. Genet.* **23**, 151–157 (1999).
59. Huynen, M., Snel, B., Lathe, W. & Bork, P. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res.* **10**, 1204–1210 (2000).
60. Albert, I. & Albert, R. Conserved network motifs allow protein-protein interaction prediction. *Bioinforma. Oxf. Engl.* **20**, 3346–3352 (2004).
61. Korbel, J. O., Jensen, L. J., von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* **22**, 911–917 (2004).
62. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1974–1979 (2005).
63. Ofran, Y. & Rost, B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236–239 (2003).
64. Raman, K. Construction and analysis of protein-protein interaction networks. *Autom. Exp.* **2**, 2 (2010).
65. Zhou, H. X. & Shan, Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44**, 336–343 (2001).
66. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361 (2002).

67. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96**, 4285–4288 (1999).
68. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
69. Pazos, F. & Valencia, A. Protein co-evolution, co-adaptation and interactions. *EMBO J.* **27**, 2648–2655 (2008).
70. Thompson, John N. *The Coevolutionary Process*. (1994).
71. Pazos, F. & Valencia, A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227 (2002).
72. Shoichet, B. K. & Kuntz, I. D. Protein docking and complementarity. *J. Mol. Biol.* **221**, 327–346 (1991).
73. Lawrence, M. C. & Colman, P. M. Shape Complementarity at Protein/Protein Interfaces. *J. Mol. Biol.* **234**, 946–950 (1993).
74. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information Edited by J. Thornton. *J. Mol. Biol.* **272**, 106–120 (1997).
75. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**, 989–998 (2003).
76. Marcotte, E. M., Xenarios, I. & Eisenberg, D. Mining literature for protein-protein interactions. *Bioinforma. Oxf. Engl.* **17**, 359–363 (2001).
77. Donaldson, I. *et al.* PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **4**, 1–13 (2003).
78. Huang, M. *et al.* Discovering patterns to extract protein-protein interactions from full texts. *Bioinforma. Oxf. Engl.* **20**, 3604–3612 (2004).
79. Ramani, A. K., Bunescu, R. C., Mooney, R. J. & Marcotte, E. M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
80. Jaeger, S., Gaudan, S., Leser, U. & Rebholz-Schuhmann, D. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics* **9**, S2 (2008).
81. He, M., Wang, Y. & Li, W. PPI Finder: A Mining Tool for Human Protein-Protein Interactions. *PLoS ONE* **4**, e4554 (2009).
82. Li, X., Cai, H., Xu, J., Ying, S. & Zhang, Y. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids* **38**, 1237–1252 (2010).

83. Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T. & Iliopoulos, I. Protein–protein interaction predictions using text mining methods. *Methods* **74**, 47–53 (2015).
84. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
85. Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinforma. Oxf. Engl.* **24**, 282–284 (2008).
86. Gavin, A.-C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
87. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
88. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).
89. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S. J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37**, 825–831 (2009).
90. Antti Airola *et al.* A graph kernel for protein-protein interaction extraction. *Proc. Workshop Curr. Trends Biomed. Nat. Lang. Process.*
91. Yu, L., Gao, L. & Kong, C. Identification of core-attachment complexes based on maximal frequent patterns in protein-protein interaction networks. *Proteomics* **11**, 3826–3834 (2011).
92. Wang, S. & Wu, F. Detecting overlapping protein complexes in PPI networks based on robustness. *Proteome Sci.* **11**, S18 (2013).
93. Maruyama, O. & Kuwahara, Y. RocSampler: regularizing overlapping protein complexes in protein-protein interaction networks. *BMC Bioinformatics* **18**, 491 (2017).
94. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).
95. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9**, 471–472 (2012).
96. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
97. Dongen, S.M. van & University Utrecht. Graph clustering by flow simulation.
98. Vlasblom, J. & Wodak, S. J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* **10**, 1–14 (2009).
99. Li, X., Wu, M., Kwoh, C.-K. & Ng, S.-K. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* **11 Suppl 1**, 1–19 (2010).
100. Moschopoulos, C. N. *et al.* Which clustering algorithm is better for predicting protein complexes? *BMC Res. Notes* **4**, 549 (2011).

101. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol. Clifton NJ* **804**, 281–295 (2012).
102. Bader, G. D. & Hogue, C. W. V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
103. Jiang, P. & Singh, M. SPICi: a fast clustering algorithm for large biological networks. *Bioinforma. Oxf. Engl.* **26**, 1105–1111 (2010).
104. Tripathi, S., Moutari, S., Dehmer, M. & Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinformatics* **17**, 129 (2016).
105. Ashtiani, M. *et al.* A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst. Biol.* **12**, 80 (2018).
106. Ivazeh, A., Zahiri, J., Rahgozar, M. & Srihari, S. Performance evaluation measures for protein complex prediction. *Genomics* **111**, 1483–1492 (2019).
107. Meilă, M. Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98**, 873–895 (2007).
108. McDaid, A. F., Greene, D. & Hurley, N. Normalized Mutual Information to evaluate overlapping community finding algorithms. *ArXiv11102515 Phys.* (2013).
109. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
110. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
111. Sasaki, Yutaka. The truth of the F-measure. (2007).
112. Heer, J., Card, S. K. & Landay, J. A. prefuse: a toolkit for interactive information visualization. in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05* 421 (ACM Press, 2005). doi:10.1145/1054972.1055031.
113. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
114. Milligan, G. W. & Cooper, M. C. A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivar. Behav. Res.* **21**, 441–458 (1986).
115. Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. The mutual information: detecting and evaluating dependencies between variables. *Bioinforma. Oxf. Engl.* **18 Suppl 2**, S231–S240 (2002).
116. Steinley, D. Properties of the Hubert-Arable Adjusted Rand Index. *Psychol. Methods* **9**, 386–396 (2004).
117. Amigó, E., Gonzalo, J., Artiles, J. & Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.* **12**, 461–486 (2009).

118. Emmons, S., Kobourov, S., Gallant, M. & Börner, K. Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale. *PLOS ONE* **11**, e0159161 (2016).
119. Jorge M. Santos & Mark Embrechts. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. in (2009).

6.Παράρτημα

6.1 Σύνδεσμοι Βάσεων Δεδομένων

Όνομα Βάσης Δεδομένων	Διαδικτυακός Σύνδεσμος
3DID	https://3did.irbbarcelona.org/
3D-Interlogs	http://3d-interologs.life.nctu.edu.tw/introduction.php
ACSN	https://acsn.curie.fr/ACSN2/ACSN2.html
AnimalTFDB	http://bioinfo.life.hust.edu.cn/AnimalTFDB/#!/
AtPIN	https://atpin.bioinfoguy.net/cgi-bin/atpin.pl
BIANA	http://sbi.imim.es/web/index.php/research/servers/biana?page=biana.server
BioGRID	https://thebiogrid.org/
BioPlex	https://bioplex.hms.harvard.edu/
CancerNet	http://bis.zju.edu.cn/CancerNet/
CCSB Interactome DB	http://interactome.dfci.harvard.edu/
CIDeR	http://mips.helmholtz-muenchen.de/cider
ComPPI	https://comppi.linkgroup.hu/
CORNET	https://bioinformatics.psb.ugent.be/cornet/
CORUM	http://mips.helmholtz-muenchen.de/corum/
COXPRESdb	https://coxpresdb.jp/
dbSNO 2.0	http://140.138.144.145/~dbSNO/
DEPOD	http://depod.bioss.uni-freiburg.de/
DIP	https://dip.doe-mbi.ucla.edu/dip/Main.cgi
FlyBase	https://flybase.org/
FlyMine	https://www.flymine.org/flymine/begin.do
FunCoup	http://funcoup.sbc.su.se/search/
GenAge	http://genomics.senescence.info/genes/
GeneMANIA	http://genemania.org/
gpDB	http://biophysics.biol.uoa.gr/gpDB/index.jsp
GPS-Prot	http://gpsprot.org/
HIV-1 Interactions DB	https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/
HIVMID	https://www.hiv.lanl.gov/content/immunology/
HP-DPI	http://dpi.nhri.org.tw/protein/hp/ORF/index.php
HPID	http://wilab.inha.ac.kr/hpid/
HPRD	http://www.hprd.org/
I2D	http://ophid.utoronto.ca/ophidv2.204/

IMEx	http://www.imexconsortium.org/
IntAct	https://www.ebi.ac.uk/intact/
INTERSPIA	http://bioinfo.konkuk.ac.kr/INTERSPIA/index.php
iRefWeb	http://wodaklab.org/iRefWeb/
MatrixDB	http://matrixdb.univ-lyon1.fr/
Mentha	http://mentha.uniroma2.it/index.php
MINT	https://mint.bio.uniroma2.it/
Pathway Commons	http://www.pathwaycommons.org/
PepBank	http://pepbank.mgh.harvard.edu/
PhosphoELM	http://phospho.elm.eu.org/
PIMADb	http://caps.ncbs.res.in/pimadb/
PINA2	https://omics.bjcancer.org/pina/
PIPs	http://www.compbio.dundee.ac.uk/www-pips/
PrePPI	https://bhapp.c2b2.columbia.edu/PrePPI/
SIGNOR	https://signor.uniroma2.it/
STRING	https://string-db.org/cgi/input.pl
Struct2Net	http://cb.csail.mit.edu/cb/struct2net/webserver/
Tabloid	https://iomics.ugent.be/tabloidproteome/
The Signaling Pathway Project	https://signalingpathways.org/index.jsf
TRIP DB	http://www.trpchannel.org/

Πίνακας 28: Σύνδεσμοι Βάσεων Δεδομένων Αλληλεπιδράσεων Πρωτεϊνών – Πρωτεϊνών

6.2 Ευρετήριο Εικόνων

6.3 Ευρετήριο Πινάκων

6.4 Ευρετήριο Διαγραμμάτων