
Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων

Τομεας Μηχανικης


# Αξιολόγηση Μέτρων Σημασιολογικής Ομοιότητας σε Βιοϊατρικές Οντολογίες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ


**Αλέξανδρος Ξένος**

**Επιβλέπων**: Κωνσταντίνος Σιέττος
Αναπληρωτής Καθηγητής ΕΜΠ

**Συνεπιβλέπων**: Αριστοτέλης Χατζηιωάννου
Ερευνητής Β΄ ΕΙΕ



Αθήνα, Σεπτέμβριος 2018


ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΗΧΑΝΙΚΗΣ


# Αξιολόγηση Μέτρων Σημασιολογικής Ομοιότητας σε Βιοϊατρικές Οντολογίες

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ


**Αλέξανδρος Ξένος**


Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 12 Οκτωβρίου 2018.

| ..... | ..... | ..... |
|---|---|---|
| Κωνσταντίνος Σιέττος | Αριστοτέλης Χατζηιώαννου | Φραγκίσκος Κολίσης |
| Αναπληρωτής Καθηγητής ΕΜΠ | Ερευνητής Β΄ ΕΙΕ | Ομότιμος Καθηγητής ΕΜΠ |

# Abstract

The aim of this thesis is to evaluate the semantic similarities in Biomedical Ontologies. First of all, the most important characteristics that define the terms of an ontology are described as well as the the techniques for comparing them. Furthermore several scenarios for the evaluation of metrics were adopted using the Gene Ontology. On a first level, the metrics were evaluated against a snapshot of the Gene Ontology whose pairs of terms were ranked semantically based on rules derived from its topology. On the second level the ability of metrics to locate functional modulues in complex biological networks was evaluated using Clustering Algorithms. Taking into consideration the different approaches, Resnik and Aggregatc IC measures overcome all the other metrics. Finally, measures that do not take into account the Information Content of the terms to calculate semantic similarity, such as Dice and Jaccard Coefficient should not be used in Biomedical Ontologies.

# Περίληψη

Στην παρούσα διπλωματική εργασία αξιολογήθηκαν οι μετρικές σημασιολογικής ομοίοτητας που χρησιμοποιούνται στις Βιοϊατρικές Οντολογίες για την λειτουργική σύγκριση γονιδίων και πρωτεϊνών. Αρχικά, περιγράφηκαν τα σημαντικότερα χαρακτηριστικά που προσδιορίζουν τους όρους μιας οντολογίας καθώς και οι διαφορετικές τεχνικές σύγκρισής τους, που βασίζονται σ΄ αυτά τα χατρακτηριστικά. Στην συνέχεια, κατασκευάστηκαν διάφορα σενάρια για την αξιολόγηση των μετρικών, με την χρήση της Γονιδιακής Οντολογίας (Gene Ontology). Σε πρώτο επιπέδο αξιολογήθηκαν οι μετρικές ως προς ένα στιγμιότυπο της Γονιδιακής Οντολογίας του οποίου ζευγάρια όρων ιεραρχήθηκαν σημασιολογικά με βάση κάνονες που πηγάζουν από την τοπολογία της. Σε δεύτερο επίπεδο με χρήση αλγορίθμων Ομαδοποίσης αξιολογήθηκε η ικανότητα των μετρικών να εντοπίζουν λειτουργικές ομάδες γονιδίων σε σύνθετα βιολογικά δίκτυα. Το συμπέρασμα που προκύπτει είναι ότι την καλύτερη συμπεριφορά την είχαν οι μετρικές του Resnik και ο Aggregate IC. Τέλος μέτρα που δεν χρησιμοποιούν το Information Content των όρων στον υπολογισμό της σημασιολογικής ομοιότητας όπως ο Dice Coefficient και ο Jaccard Coefficient δεν πρέπει να χρησιμοποιούνται στις Βιοϊατρικές Οντολογίες.

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ.Φραγκίσκο Κολίση που με παρότρυνε να ασχοληθώ με την Βιοπληροφορική, τον κ.Αριστοτέλη Χατζηιώαννου που μου έδωσε την δυνατότητα να ασχοληθώ με το εν λόγω αντικείμενο και τον κ.Κωνσταντίνο Σιέττο που ανέλαβε να εκπροσωπήσει την ΣΕΜΦΕ. Ιδιαίτερη μνεία αξίζει στον διδακτορικό φοιτητή του ΕΜΠ Θοδωρή Κουτσανδρέα που χωρίς την βοήθεια και την καθογήση του δεν θα μπορούσε να ολοκληρωθεί η διπλωματική εργασία. Τέλος οφείλω ένα μεγάλο ευχαριστώ στο οικογενειάκο και φιλικό μου περιβάλλον που μου παρείχαν αμέριστη συμπαράσταση και στηρίξη.

# Περιεχόμενα

# Εισαγωγή

Σκοπός αυτής της εργασίας είναι η αξιολόγηση των μετρικών σημασιολογικής ομοιότητας στις Βιοϊατρικές Οντολογίες και κυρίως στην Γονιδιακή Οντολογία. Η αξιολόγιση των μέτρικών έγινε με βάση τα δομικά χαρακτηριστικά της Γονιδιακής Οντολογίας αλλά και με βάση την συμπεριφορά τους στην σύγκριση γονιδίων όταν αυτά σχηματίζουν πολύπλοκα Βιολογικά Δίκτυα. Η εργασία χωρίστηκε στο θεωρητικό (Κεφ. 1, 2 και 3) και στο τεχνικό μέρος (Κεφ. 4, 5 και 6).

Στο Κεφάλαιο 1 μελετήθηκε η δομή των οντολογιών και τα υπάρχοντα μέτρα σημασιολογικής ομοιότητας, δηλαδή οι τρόποι που υπάρχουν για την λειτουργική σύγκριση δύο γονιδιών ή δύο γονιδιακών προϊόντων. Στο Κεφάλαιο 2 περιγράφηκαν βασικοί ορισμοί από την Θεωρία Γραφημάτων για να ορισθούν στην συνέχεια οι βασικές τοπολογικές και στατιστικές ιδιότητες των σύνθετων Βιολογικών Δικτυών. Επιπλέον περιγράφηκαν τόσο λειτουργικά όσο και τοπολογικά τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων. Στο Κεφάλαιο 3 παρουσιάστηκαν οι αλγόριθμοι ομαδοποίησης, δηλαδή οι τρόποι που υπάρχουν για την χωρισμό των δεδομένων σε διακριτές ομάδες καθώς και οι τρόποι σύγκρισης δύο διαφορετικών ομαδοποιήσεων για τα ίδια δεδομένα.

Στο Κεφάλαιο 4 ορίσθηκαν κανόνες σύμφωνοι με την ανθρώπινη αντίληψη για την ιεράρχιση της σημασιολογικής ομοιότητας των όρων. Με βάση αυτούς του κανόνες ταξινομήθηκαν σε φθίνουσα σειρά οι σημασιολογικές ομοιότητες σε ένα στιγμιότυπο της Γονιδιακής Οντολογίας και εξετάστηκε η συμπεριφορά των μετρικών ως προς αυτή την κατάταξη. Τέλος στα Κεφάλαια 5 και 6 αξιολογήθηκε η συμπεριφορά των μετρικών στα δίκτυα πρωτεϊνικών αλληλεπιδράσεων και με την χρήση αλγορίθμων ομαδοποίησης εξετάστηκε η δυνατότητα των μετρικών να εντοπίζουν λειτουργικά όμοιες ομάδες γονιδίων στα Βιολογικά Δίκτυα.

# Κεφάλαιο 1

# Βιοϊατρικές Οντολογιές και Μέτρα Σημασιολογικής Ομοιότητας

Στο κεφάλαιο αυτό θα παρουσιαστούν οι οντολογίες και κυρίως η πίο γνωστή Βιοϊατρική Οντολογία η Γονιδιακή Οντολογία (Gene Ontology). Στην συνέχεια θα περιγραφούν εκτενώς τα μέτρα σημασιολογικής ομοιότητας (semantic similarities measures) που έχουν αναπτυχθεί γύρω από τις οντολογίες.

## 1.1 Βιοϊατρικές Οντολογίες

Η Οντολογία είναι ένας τομέας της Φιλοσοφίας η πρακτική του οποίου χρονολογείται από τον Αριστοτέλη. Συνήθως αναφέρεται ως τμήμα του υπερκλάδου της φιλοσοφίας που είναι γνωστός ως μεταφυσική. Πρόκειται για την φιλοσοφική μελέτη του όντος δηλαδή την μελέτη εννοιών που σχετίζονται άμεσα με την ύπαρξη, την πραγματικότητα και τη φύση των πραγμάτων καθώς και τις σχέσεις μεταξύ τους [1]. Συχνά ασχολείται με ερωτήματα σχετικά με τις οντότητες που υπάρχουν ή που είναι πιθανόν να υπάρχουν και τον τρόπο με τον οποίο οι οντότητες αυτές μπορούν να ομαδοποιηθούν, να συνδεθούν ιεραρχικά και να υποδιαιρεθούν σύμφωνα με ομοιότητες και διαφορές.

Στις αρχές τις δεκαετίας του ΄80 οι Οντολογίες υιοθετήθηκαν από ερευνητές στον τομέα της Τεχνητής Νοημοσύνης που ήθελαν να βρουν μία παραλληλία μεταξύ της "μελέτης του τί υπάρχει" (υπάρχουσα γνώση), με αυτό που "υποθέτουμε ότι υπάρχει" (εξαγωγή συμπερασμάτων/νέας γνώσης). Δηλαδή τη μετατροπή της φυσικής γλώσσας που χρησιμοποείται για την περιγραφή ενός τομέα, σε κωδικοποιημένη γλώσσα (γλώσσα μηχανής) η οποία "υποθέτουμε ότι υπάρχει" [2] προκειμένου να επιτευχθεί μια συνεκτική περιγραφή της πραγματικότητας (υπάρχουσας γνώσης) στα έξυπνα συστήματα (intelligent systems). Στην επιστήμη των υπολογιστών, η οντολογία είναι μια μορφή αναπαράστασης γνώσης (knowledge represantation) μέσω μιας κοινής και συμφωνημένης εννοιολογικής μορφοποίησης ενός γνωστικού πεδίου, συνήθως ως ένα σύνολο εννοιών, σχέσεων και ιδιοτήτων. Με άλλα λόγια οι οντολογίες είναι ένα σχήμα που αναπαριστά κάποιο τομέα ή γνωστικό αντικείμενο χρησιμοποιώντας ένα λεξιλόγιο όρων (repre-

santation vocabulary) που δεν καθορίζεται από την γλώσσα που είναι γραμμένο αλλά από τις έννοιες που εκφράζουν/μοντελοποιούν οι όροι [3]. Οι οντολογίες αναπαριστόνται μέσω ιεραρχικών δομών (γράφων), όπου οι κόμβοι είναι οι όροι και οι ακμές οι σημασιολογικές σχέσεις μεταξύ των όρων. Οι όροι στα ψηλά επίπεδα της οντολογίας είναι γενικοί και μερικές φορές ανεξάρτητοι από το πεδίο που περιγράφουν και όσο κατεβαίνουν στην ιεραρχία τόσο πιο πολύ εξειδικεύονται. [3].

Τα τελευταία χρόνια η τεράστια πρόοδος στον τομέα των βιολογικών επιστημών και πιο συγκεκριμένα στις τεχνολογίες αλληλούχησης νέας γενιάς (next generation sequencing) οδήγησε στην αλληλούχιση των πρώτων γονιδιωμάτων με αποκορύφωμα την αλληλούχιση του ανθρώπινου γονιδιώματος. Η γονιδιωματική, η μελέτη δηλαδή των ιδιοτήτων του γονιδιώματος, αποτελεί βασική πηγή παραγωγής τεράστιου όγκου δεδομένων (big data) στη σύγχρονη βιολογία [4]. Ενώ η επεξεργασία και η ανάλυση τέτοιου όγκου δεδομένων δεν μπορούσε να γίνει χειροκίνητα, η ετερογένεια των μορφών και των δομών των δεδομένων καθιστούσε πολύ δύσκολη την ενσωμάτωση τους σε μεγάλα υπολογιστικά συστήματα [5]. Τα παραγώμενα δεδομένα για είναι χρήσιμα έπρεπε να οριστούν και να περιγραφούν μέσα από ένα κοινό σχήμα. Για το σκοπό αυτό, κρίθηκε αναγκαία η ανάπτυξη διαφόρων βιολογικών οντολογιών, οι οποίες παρέχουν σχήματα πάνω στα οποία οι οντότητες των βιολογικών συστημάτων (όπως γονίδια, πρωτεΐνες, μεταβολίτες κ.α) μπορούν να περιγραφούν (gene mapping).

Οι Βιοϊατρικές Οντολογίες είναι διαθέσιμες και ελεύθερα προσβάσιμες στο διαδύκτιο στο ηλεκτρονικό αποθετηρίο Open Biological and Biomedical Ontology(OBO) Foundry [6]. Η πρώτη χρονολογικά και πιο γνωστή Βιοϊατρική οντολογία είναι η Γονιδιακή Οντολογία (Gene Ontology) [7] η οποία περιγράφει το σύνολο της βιολογικής γνώσης σχετικά με τις γονιδιακές λειτουργίες και τον τρόπο με τον οποίο αυτές συνδέονται μεταξύ τους. Άλλες γνωστές οντολογίες είναι:
**i)** η Οντολογία Φαινοτυπικών Χαρακτηριστικών του Ανθρώπου (Human Phenotype Ontology) [8] που περιγράφει τις φαινοτυπικές ανωμαλίες που σχετίζονται με τις ανθρώπινες ασθένειες,
**ii)** η Οντολογία Ανθρώπινων Ασθενειών (Human Disease Ontology)[9] η οποία ταξινομεί τις καταγεγραμμένες ανθρώπινες ασθένειες με βάση την αίτια τους και **iii)** η Οντολογία των Φυτών (Plant Ontology) [10] η οποία περιγράφει τα ανατομικά, μορφολογικά και αναπτυξιακά χαρακτηριστικά των φυτών.

## 1.2 Γονιδιακή Οντολογία

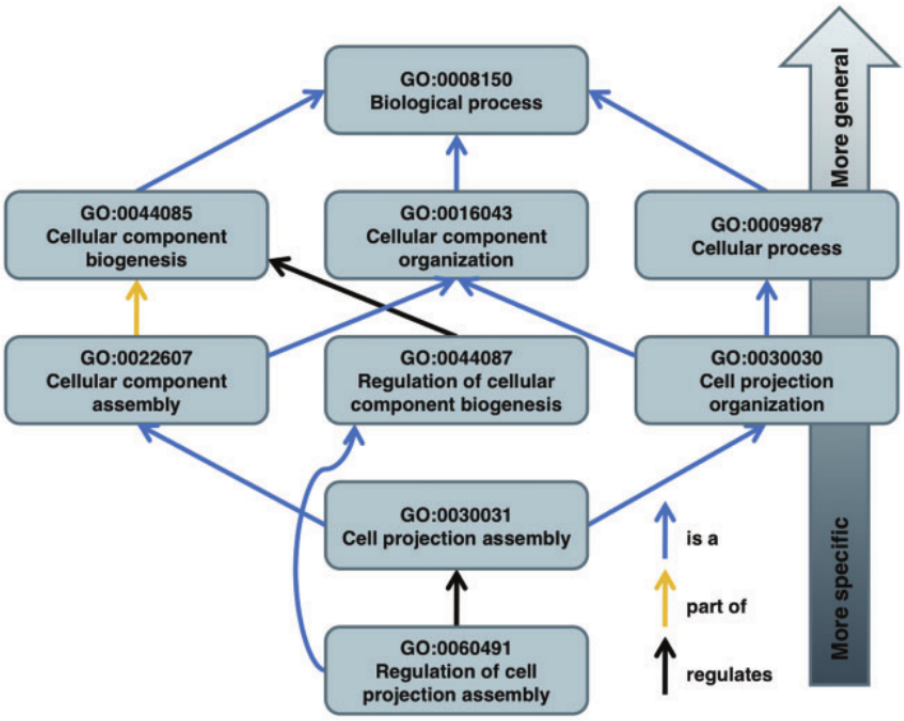
Η Γονιδιακή Οντολογία (GO) [7] αποτελεί το αποτέλεσμα της δουλειάς του αντίστοιχου Consortium, το οποίο συστήθηκε το 1998 με σκοπό τη δημιουργία μιας συνεκτικής, παγκόσμιας ονοματολογίας γονιδίων για όλους τους οργανισμούς είτε ευκαριωτικούς είτε προκαρυωτικούς[11]. Η Gene Ontology σύμφωνα με τον Ashburner [7] αποτέλεσε το πρώτο βήμα προς την επίλυση ενός σημαντικού βιολογικού προβλήματος: της ενωποίησης του συνόλου της βιολογικής γνώσης, σχετικά με τα γονίδια και τα γονιδιάκα προϊόντα για διαφορετικά είδη οργανισμών, σε ένα ενιαίο σχήμα. Η προσπάθεια αυτή δημιούργησε το υπόβαθρο για τη βαθύτερη κατανόηση των μηχανισμών και των λειτουργιών διαφόρων οργανισμών, καθώς η γνώση του βιολογικού ρόλου ενός

γονιδίου ή μιας πρωτεΐνης σε ένα οργανισμό, μπορεί να δώσει ισχυρές ενδείξεις για το ρόλο της σε άλλους. Το σχήμα της Γονιδιακής Οντολογίας αποτελείται από ένα σύνολο όρων ('GO terms') που περιγράφουν τις λειτουργίες των γονιδίων και από τις (σημασιολογικές) σχέσεις με τις οποίες συνδέονται οι λειτουργίες. Η GO αποτελείται από 3 κατηγορίες (οντολογίες) που η καθεμία περιγράφει τις γονιδιακές λειτουργίες από διαφορετική πλευρά. Οι τρεις οντολογίες είναι [11]:

1. **Η Μοριακή Λειτουργία** (Molecular Function): η οποία περιγράφει τις μοριακές λειτουργίες που εκτελούνται από γονιδιακά προϊόντα.

2. **Η Βιολογική Διαδικασία** (Biological Process): η οποία περιγράφει μεγαλύτερες διεργασίες που γίνονται χάρη σε πολλαπλές μοριακές λειτουργίες δηλαδή τις λειτουργίες των γονιδίων.

3. **Το Κυταρικό Στοιχείο** (Cellular Component): το οποίο περιγράφει σε ποιο τμήμα/ οργανίδιο του κυττάρου επιτελείται μια λειτουργία από ένα γονιδιακό προϊόν.




Σχήμα 1.1: Ένα στιγμιότυπο της Βιολογικής Διαδικασίας [12]
για τον υπογράφο του όρου 'GO:0060491'.

Η κάθε υποκατηγορία της GO (όπως και κάθε βιοϊατρική οντολογία γενικότερα) αναπαριστάται από έναν κατευθυνόμενο κυκλικό γράφο (direct acyclic graph (DAG)), όπου οι κόμβοι αντιστοιχούν στους περιγραφικούς όρους (terms) και οι ακμές στις σημασιολογικές σχέσεις μεταξύ

των όρων. Η ρίζα της οντολογίας, (root node) όπως φαίνεται και στο σχήμα 1.1, ορίζει την βασική κατηγορία. Οι οντολογίες υιοθετούν μια ψεύδο-ιεραρχική (loosely hierarchical) δομή: όσο πιο ¨ψηλά' βρισκόμαστε στην οντολογία τόσο πιο γενικοί είναι οι όροι, ενώ όσο χαμηλώνουμε στην ιεραρχία εξιδεικεύονται. Με τον τρόπο αυτό οι απόγονοι είναι πιο ειδικοί από τους προγόνους τους [13]. Ένας όρος μπορεί να έχει περισσότερους από ένα γονέα (parent term), σε αντίθεση με τις αυστηρές ιεραρχικές δομές [13], δημιουργώντας πολύπλοκα μονοπάτια στον γράφο, συνδέοντας φαινομενικά ασυσχέτιστους όρους. Οι ακμές του γράφου, δηλαδή οι πιθανές σημασιολογικές σχέσεις μεταξύ των όρων, είναι οι ακόλουθες [14]:

1. **"Είναι"** (is-a) που δηλώνει ότι ο απόγονος είναι εξειδίκευση(subtype) του προγόνου του. Η σχέση is-a είναι μεταβατική δηλαδή αν *ο A είναι B* και *ο B είναι C* τότε και *ο A είναι C*. Αν υπήρχαν μόνο is-a σχέσεις η οντολογία θα είχε δενδρική(tree) δομή συνεπώς είναι αυτή που συνδέεται με την ψέυδο-ιεραρχική δομή της.

2. **"Αποτελεί μέρος του"** (is-part) που υποδηλώνει ότι ο όρος συμμετέχει σε μια διαδικασία που έχει οριστεί προηγουμένως μέσω του όρου που συνδέεται.

3. **"Ρυθμίζει"** (regulates) που σχετίζεται με τη λειτουργία της ρύθμισης της έντασης μια διαδικασίας από μία αλλή. Η ρύθμιση μπορεί να είναι είτε θετική (positive regulates) είτε αρνητική (negative regulates).

Πέρα από τις σημασιολογικές σχέσεις που σχηματίζονται πάνω στο γράφημα της οντολογίας, ο συσχετισμός των όρων των τριών κατηγοριών της ΓΟ με γονίδια και γονιδιακά προϊόντα, δημιουργεί ένα δεύτερο επίπεδο περιγραφής τους. Η σύνδεση ενός γονιδίου με όρους της GO ορίζει τη λειτουργία του ή το σύνολο των κυτταρικών τμημάτων στα οποία λειτουργεί. Με τον τρόπο όμως αυτόν καθορίζονται και τα σύνολα των γονιδίων, με τα οποία συσχετίζεται ο κάθε οντολογικός όρος [15]. Κάθε χαρακτηρισμός γονιδίου στην GO περιέχει το όνομα του γονιδίου, τη βάση δεδομένων στην οποία βρίσκεται καταχωρυμένο (π.χ. UniProt), τον όρο ή τους όρους στους οποίους αντιστοιχίζεται, τις αντίστοιχες επιστημονικές δημοσιεύσεις και έναν κωδικό (evidence code) που συνδέεται με το είδος και την ποιότητα της αντιστοίχισης.

Οι κατηγορίες κωδικων είναι [12]: **i)**'EXP' για τις αντιστοιχίσεις που επιβεβαιώνονται πειραματικά, **ii)**"IEA' για αυτές που προκύπτουν χώρις καμία ανθρώπινη εποπτεία δηλαδή είτε αυτόματα (automatically-assigned) είτε από υπολογιστικές αναλύσεις (computational analysis), **iii)**"TAS' για αυτές που προκύπτουν από δημοσιεύσεις ή reviews, **iv)**"IC' για αυτές που προκύπτουν από την ανάλυση των δεδομένων από ειδικούς, **v)**"ND' για γονίδια που δεν είναι ακόμα πλήρως χαρακτηρισμένα και αντιστοιχίζονται σε πολύ γενικούς όρους και **vi)**"NR' για αντιστοιχίσεις που είχαν γίνει πριν την υποχρεωτική χρήση των evidence codes. Οι πιο αξιόπιστες αντιστοιχίσεις είναι αυτές που επιβεβαιώνονται πειραματικά. Ωστόσο λόγω του όγκου των διαθέσιμων δεδομένων αυτές που είναι οι πιο συχνές είναι αυτές που γίνονται ηλεκτρονικά ή υπολογιστικά οι οποίες όμως είναι πιθανόν να έχουν λάθη.

Κάθε γονίδιο αντιστοιχίζεται με όρους και από τις τρεις κατηγορίες της GO, χωρίς να λαμβάνονται υπόψιν οι σημασιολογικές σχέσεις του κάθε όρου. Αυτό έχει σαν αποτέλεσμα ένα

γονίδιο να συσχετίζεται με μια συγκεκριμένη κυτταρική λειτουργία ή κυτταρικό τμήμα, αλλά όχι με τον αντίστοιχο πατρικό όρο. Το φαινόμενο αυτό είναι αντίθετο στη δομή της οντολογίας, καθώς οτιδήποτε χαρακτηρίζει ένα συγκεκριμένο υποσύνολο, θα πρέπει να είναι μέρος του χαρακτηρισμού και του πατρικού συνόλου. Για τον λόγο αυτό, κατά τον χαρακτηρισμό της ΓΟ με γονίδια, είναι αναγκαίο να λαμβάνεται υπόψιν ο κανόνας true path rule. Ο κανόνας αυτός ορίζει πως κάθε γονίδιο που συνδέεται με έναν οντολογικό όρο, χαρακτηρίζει αυτόματα και όλους τους πρόγονους τους (ancestors) [16] . Ο true path rule πηγάζει από την μεταβατικότητα των σχέσεων is-a και part-of και την ψευδό-ιεραρχική δομή της οντολογίας. Συνεπώς κάθε γονίδιο μπορούμε να το δούμε ως ένα υπογράφημα (subgraph) πάνω στην οντολογία. Αξίζει να σημειωθεί πως η διατήρηση και η ενημέρωση της GO, η προσθήκη νέων όρων και η σύνδεσή ή η αποσύνδεσή τους με γονίδια, με βάση τις νέες επιστημονικές έρευνες, είναι μια διαρκής διαδικασία που στόχο έχει την καλύτερη περιγραφή της βιολογίκής πληροφορίας.

## 1.3 Μέτρα Σημασιολογικής Ομοιότητας

Σημαντικό ρόλο στην πρόοδο της Βιολογίας και στην κατανόηση όλο και περισσότερων αχαρτογράφητων οντοτήτων όπως γονίδια, κύταρα ή ακόμα και ολόκληρων πληθυσμών έπαιξε η συγκρίση τους με άλλες ήδη καλά χαρακτηρισμένες οντότητες. Ακόμα και σήμερα, όταν ανακαλύπτονται καινούργιες οντότητες, οι βιολόγοι τις συγκρίνουν με τις υπάρχουσες και προσπαθούν να βρουν ομοιότητες και διαφορές. Για παράδειγμα, η σύγκριση της αλληλουχίας μιας άγνωστης πρωτεΐνης σε μια βάση δεδομένων γνωστών αλληλουχιών, με τη χρήση αλγορίθμων στοίχισης (alignment algorithms), μπορεί να δώσει σημαντικές πληροφορίες για την οικογένεια που ανήκει, καθώς και ενδείξεις για τη λειτουργία της. Παρόλ΄ αυτά, οι αλγόριθμοι συνολικής λειτουργικής σύγκρισης γονιδίων ή γονιδιακών προϊόντων δεν έχουν αναπτυχθεί αρκετά, ενώ ο όποιος συσχετισμός γίνεται, βασίζεται στο κοινό προφίλ ρύθμισης και έκφρασης τους μέσα στο κύτταρο σε συγκεκριμένες συνθήκες (functional aspects).[17]

Για να συγκριθούν λειτουργικά οι βιολογικές οντότητες, δεν αρκεί να καταγραφούν και να περιγραφούν οι λειτουργίες τους σε φυσική γλώσσα. Χρειάζεται το σύνολο της γνώσης για τις γονιδιακές λειτουργίες να έχει ορισθεί αυστηρά σε ένα κοινό σχήμα το οποίο θα επιτρέπει την ποσοτικοποίηση των ομοιοτήτων και των διαφορών τους και συνεπώς την σύγκριση τους σε λειτουργικό επίπεδο. Εδω έγκειται η χρησιμότητα των βιοϊατρικών οντολογιών. Η οργάνωση της βιολογικής γνώσης σε ιεραρχικά σχήματα (όροι και σημασιολογικές σχέσεις) και ο συσχετισμός των όρων με τις βιολογικές οντότητες, αποτελεί ένα μαθηματικό εργαλείο με στοιχεία από τη Θεωρεία Συνόλων και τη Θεωρεία Γραφημάτων, το οποίο μπορεί να χρησιμοποιηθεί για τη σημασιολογική σύγκριση αυτών των οντοτήτων. Συνεπώς, αν θελούμε να συγκρίνουμε δυο γονίδια, αρκεί να συγκρίνουμε τους οντολογικούς όρους που τα χαρακτηρίζουν και στην σύνεχεια με κατάλληλες στρατηγικές να υπολογιστεί η σημασιολογική τους ομοιότητά.

Στη συνέχεια θα παρουσιαστούν οι τρόποι σημασιολογικής σύγκρισης (στρατηγικές και μετρικές) οντολογικών όρων. Η μέθοδος αυτή αποτελεί τη βάση για τη σύγκριση βιολογικών οντοτήτων, όπως είναι τα γονίδια. Συνήθως, οι μετρικές αυτές παίρνουν τιμές στο διάστημα

$[0, 1]$ όπου το 0 αντιστοιχεί σε μηδενική ομοιότητα ενώ το 1 καταδεικνύει πλήρη ομοιότητα βιολογικού περιεχομένου. Τα μέτρα σημασιολογικής ομοιότητας για τη σύγκριση δύο όρων (terms) της οντολογίας χωρίζονται σε 4 βασικές κατηγορίες [17] :

1. Τα **μέτρα ακμών** (edge-based) που υπολογίζουν την ομοιότητα με βάση την απόσταση των όρων στην οντολογία.

2. Τα **μέτρα κόμβων** (node-based) που υπολογίζουν την ομοιότητα με βάση το περιεχόμενο (information content) των όρων δηλαδή την λειτουργία που περιγράφουν.

3. Τα **υβριδικά** (hybrid) μέτρα που είναι μια μίξη των δύο προηγούμενων και λαμβάνουν υπόψιν τους τόσο την τοπολογία της οντολογίας όσο και το περιεχόμενο των όρων.

4. Τα **μέτρα συνόλων** (set-based) που υπολογίζουν την ομοιότητα με βάση ιδιότητες από την θεωρία συνόλων όπως η τομή και η ένωση του συνόλου των απογόνων/προγόνων των όρων.

### 1.3.1 Μέτρα Ακμών

Τα μέτρα ακμών δεν υπολογίζουν απευθείας την ομοιότητα δυο όρων, αλλά μετράνε την απόσταση (distance) τους πάνω στην οντολογία. Συνήθως η απόσταση υπολογίζεται ως το μήκος του μικρότερου μονοπατιού (shortest path) που συνδέει τους δύο όρους [18] ή σπανιότερα υπολογίζεται ως το μέσο μήκος όλων των μονοπατιών που τους συνδέουν. Εναλλακτικά μπορεί να υπολογισθεί ως η απόσταση του πιο κοντινού κοινού τους προγόνου (lowest common ancestor) από την ρίζα(root) της οντολογίας [19]. Η απόσταση μπορεί εύκολα να μετασχηματιστεί σε μέτρο ομοιότητας με τη χρήση ενός γραμμικού μετασχηματισμού της μορφής: $1 - distance$ [17].

Ωστόσο τα μέτρα ακμών βασίζονται σε δύο υποθέσεις[17] : α) ότι οι ακμές και οι κόμβοι είναι τυχαία κατανεμημένοι στην οντολογία και β) ότι οι ακμές στο ίδιο επίπεδο (βάθος) της οντολογίας αντιστοιχούν στον ίδιο βαθμό εξειδίκευσης και συνεπώς στην ίδια σημασιολογική απόσταση. Οι υποθέσεις αυτές μπορεί να είναι σύμφωνες με τη Θεωρία Γραφημάτων όμως δεν επιβεβαιώνονται στις βιολογικές οντολογίες. Αυτό συμβαίνει γιατί ο βαθμός εξειδίκευσης ενός όρου, δηλαδή το "σπάσιμο" των γενικότερων όρων σε πιο ειδικούς συμβαδίζει με τις τάσεις της επιστημονικής έρευνας. Συνεπώς ένας όρος μπόρει να είναι αναλύμενος σε μεγάλο βαθμό, ενώ ένας άλλος στο ίδιο επίπεδο να είναι πιο γενικός.

Παρότι έχουν γίνει προτάσεις για να ξεπεράστουν τα προβλήματα που αναφέρθηκαν προηγούμενως, όπως το να υπάρχουν βάρη στις ακμές ανάλογα με την εξειδίκευση τους, τα μέτρα ακμών θεωρούνται αναποτελεσματικά και δεν χρησιμοποιούνται [17]. Ενδεικτικά αναφέρονται τα παρακάτω μέτρα για δύο οντολογικούς όρους $C_1$ και $C_2$:

Ο Rada [18] όρισε την απόσταση μέταξυ δύο όρων της οντολογίας ως το μήκος του μικρότερου μονοπατιού (shortest path) που τους συνδέει:

$$dist_{Rada} = sp(C_1, C_2) \tag{1.1}$$

Ωστόσο η απόσταση του Rada δεν είναι κανονικοποιημένη, δηλαδή δεν παίρνει τιμές στο $[0, 1]$. Για να κανονικοποιηθεί αρκεί να διαιρεθεί με τη μέγιστη απόσταση που υπάρχει στην οντολογία η οποία θα είναι η απόσταση της ρίζας(root node) από τα φύλλα της οντολογίας.

Οι Pekar and Stab [19] όρισαν ως μέτρο σημασιολογικής ομοιότητας την απόσταση του πιο κοντινού κοινού πρόγονου (έστω C) από τη ρίζα(root) της οντολογίας:

$$sim_{PS} = \frac{sp(C, root)}{sp(C, root) + sp(C_1, root) + sp(C_2, root)} \tag{1.2}$$

### 1.3.2 Μέτρα Κόμβων

Πρόκειται για την πιο συνιθισμένη κατηγορία μέτρων, τα οποία βασίζονται στο σημασιολογικό περιεχόμενο των όρων. Τα πρώτα μέτρα που χρησιμοποιήθηκαν έχουν τις καταβολές στους στην επεξεργασία φυσικής γλώσσας (natural language processing) και βασίζονται στην πληροφορία που περιέχει (information content) ο κάθε όρος. Το **Information Content (IC)** [20] είναι ένα μέτρο από την Θεωρία Πληροφοριών που καταδεικνύει πόσο ειδικός ή γενικός είναι ένας όρος και ορίζεται ως ο αρνητικός λογάριθμος της πιθανότητας εμφάνισης του όρου:

$$IC(c) = -\log p(c) \tag{1.3}$$

Η πιθανότητα εμφάνισης ενός όρου μπορεί να οριστεί με δύο τρόπους. Ο πρώτος τρόπος βασίζεται στη δομή της οντολογίας (**graph corpus**). Η πιθανότητα εμφάνισης ενός όρου ισούται με το λόγο του αριθμού των απογόνων του προς το συνολικό αριθμό των όρων της οντολογίας: [21] :

$$p(c) = \frac{|descendants(c)|}{|descandants(root)|} \tag{1.4}$$

Ο δεύτερος τρόπος βασίζεται στον χαρακτηρισμό των οντολογικών όρων με ένα εξωτερικό συνολο στοιχείων, όπως είναι τα γονίδια mapping. Με βάση αυτη την προσέγγιση, η πιθανότητα εμφάνισης ενός όρου υπολογίζεται ως ο λόγος του αριθμού των γονιδίων που έχουν αντιστοιχηθεί σ΄ αυτό τον όρο, προς το μήκος του συνόλου των γονιδίων (έστω N) [21]. Να σημειωθεί ότι λόγω του κανόνα true path rule οι όροι αντιστοιχίζονται και με τα γονίδια που έχουν χαρακτηριστεί οι απογονοί τους [21].

$$p(c) = \frac{|annotation(c)| + \sum_{j \in descendants(c)} |annotation(j)|}{N} \tag{1.5}$$

Από τον τρόπο που ορίστηκε η πιθανότητα εμφάνισης ενός όρου είναι φανερό ότι όσο πιο γενικός είναι ένας όρος, δηλαδή όσο ψηλότερα βρίσκεται στην οντολογία, τοσο μεγαλύτερη είναι η

πιθανότητα εμφάνισής του. Εξ΄ ορισμού το IC που έχει ένας όρος είναι αντιστρόφος ανάλογο της πιθανότητας εμφάνισής του. Επομένως, η ρίζα της οντολογίας έχει πιθανότητα εμφάνισης ίση με τη μοναδα και μηδενικό IC. Αντιθέτως, όροι χαμηλότερα στην ιεραρχία, έχουν μεγαλύτερο IC.

Το IC ενός όρου όπως έχει ορισθεί δεν είναι κανονικοποιημένο, δηλαδή δεν παίρνει τιμές στο διάστημα [0,1]. Για να γίνει αυτό, αρκεί να διαιρεθεί το IC κάθε όρου με το μέγιστο IC που υπάρχει στην οντολογία [21] . Με βάση το σχήμα της οντολογίας, το μέγιστο IC θα το έχει ένα φύλλο που κατά σύμβαση έχει έναν απόγονο, τον εαυτό του. Συνεπώς:

$$max_{IC} = -\log\left(\frac{1}{|descandants(root)|}\right) \tag{1.6}$$

Αντίστοιχα, όταν χρησιμοποιείται η αντιστοίχιση των οντολογικών όρων με γονίδια πλήθους N, το μέγιστο IC θα το έχει ένας όρος που έχει αντιστοιχηθεί με τα λιγότερα γονιδία. Κατά σύμβαση θα έχει αντιστοιχηθεί με ένα μόνο γονίδιο. Επομένως:

$$max_{IC} = -\log\left(\frac{1}{N}\right) \tag{1.7}$$

### Μετρικές

Το πρώτο μέτρο πάνω στο οποίο βασίστηκαν και τα υπόλοιπα είναι αυτό του Resnik [20], που ορίζει ότι η ομοιότητα δύο όρων $C_1$ και $C_2$ είναι το IC που έχει ο κοινός τους πρόγονος με το μεγαλύτερο IC (Most Informative Common Ancestor - MICA):

$$sim_{Res}(C_1, C_2) = IC(C_{MICA}) \tag{1.8}$$

Ωστόσο η μετρική του Resnik δεν λαμβάνει καθόλου υπόψιν την απόσταση που έχουν οι δύο όροι από τον κοινό αυτόν πρόγονο. Για παράδειγμα, δύο ζευγάρια όρων που έχουν τον ίδιο MICA ανεξάρτητα από την θέση τους στην οντολογία έχουν την ίδια σημασιολογική ομοιότητα. Για να διορθώσουν το πρόβλημα του Resnik, ο Lin και στην συνέχεια οι Jiang and Conrath πρότειναν μετρικές - παραλλαγές του Resnik.

Ο Lin [22] από την πλευρά του πρότεινε:

$$sim_{Lin}(C_1, C_2) = \frac{2 \times IC(C_{MICA})}{IC(C_1) + IC(C_2)} \tag{1.9}$$

Οι Jiang and Conrath [23] πρότειναν μία μετρική που υπολογίζει την απόσταση των δύο όρων ως το άθροισμα των διαφορων που έχει το IC των δύο όρων από το IC του MICA:

$$dist_{JC}(C_1, C_2) = IC(C_1) + IC(C_2) - 2 \times IC(C_{MICA}) \tag{1.10}$$

Το παραπάνω μετρο, με την χρήση ενός γραμμικού μετασχηματισμού της μορφής $1 - distance$ μετατρέπεται σε μέτρο σημασιολογικής ομοιότητας.

Οι Pirro and Euzenat [24] πρότειναν μια παραλλαγή του Lin:

$$sim_{Pirro}(C_1, C_2) = \frac{IC(C_{MICA})}{IC(C_1) + IC(C_2) - IC(C_{MICA})} \tag{1.11}$$

Ωστόσο η ομοιότητα που υπολογίζουν και τα τρία αυτά μέτρα είναι ανάλογη της διαφόρας του IC των όρων με το IC του MICA και συνεπώς δεν λαμβάνουν υπόψιν την τοπολογική θέση του MICA στον γράφο της οντολογίας [17]. Για να αντιμετωπιστεί το εν λόγω πρόβλημα ο Schlicker [25] πρότεινε να σταθμιστεί το μέτρο του Lin με ένα βάρος (weight factor) της μορφής $1 - p(C_{MICA})$. Η τιμή που παίρνει το βάρος για την ρίζα της οντολογίας είναι 0 και όσο εξειδικεύονται σημασιολογικά οι όροι μεγαλώνει.

$$sim_{Rel}(C_1, C_2) = sim_{Lin}(C_1, C_2) \times (1 - p(C_{MICA})) \tag{1.12}$$

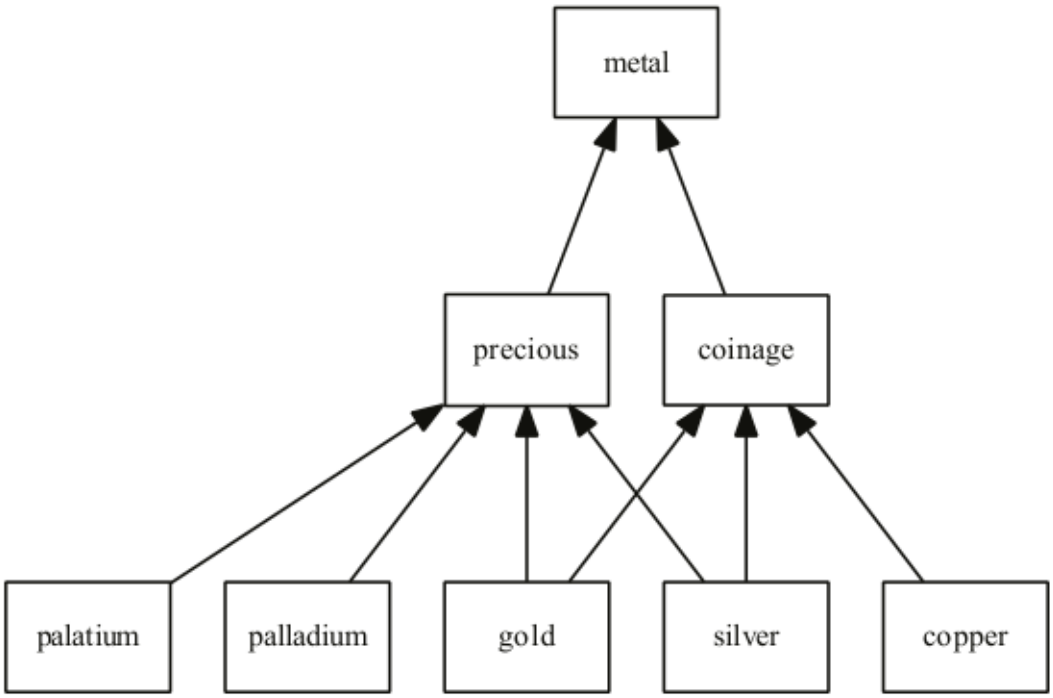
Όλες οι μετρικές που έχουν αναφερθεί λαμβάνουν υπόψιν τους μόνο το MICA. Ένα κεντρικό ερώτημα στις οντολογίες είναι το κατά πόσο η ύπαρξη πολλάπλων κοινών προγόνων (multiple parent inheritance) ή η ύπαρξη τελείως διαφορετικών προγόνων επηρεάζει την σημασιολογική ομοιότητα δυο όρων. Στην συνέχεια θα παρουσιαστούν εναλλακτικές στρατηγικές επιλογής προγόνων, οι οποίες μπορούν να χρησιμοποιηθούν στα υπάρχοντα μέτρα αντί του MICA.

### Στρατηγικές Επιλογής Προγόνων

Η πιο απλή στρατηγική είναι να λαμβάνονται υπόψιν όλοι οι διακριτοί κοινοί προγόνοι (Disjoint Common Ancestors - DCA) και στην θέση του MICA να υπολογίζεται ο μέσος όρος του IC τους. Διακριτοί κοινοί πρόγονοι ενός όρου, είναι δύο πρόγονοί του, οι οποίοι δεν έχουν μεταξύ τους σχέση προγόνου-απογόνου.

Μία άλλη προσέγγιση η οποία λαμβάνει υπόψιν της και τους μη διακριτούς προγόνους που έχουν δυο όροι είναι η Disjunctive Shared Information (DiShIn) [26] η οποία βασίζεται στο πλήθος των διακριτών μονοπατιών που έχει ένας όρος προς τους προγόνους του. Ο DiShIn όπως και ο DCA υπολογίζει αντί για το IC του MICA, το μέσο όρο του IC των διακριτών κοινών προγόνων των δύο όρων. Ωστόσο ως διακριτούς κοινούς προγόνους δεν θεώρει μόνο αυτούς που ορίζει ο DCA, αλλά επιπλέον και αυτούς που ο ένας όρος συνδέεται με περισσότερα διακριτά μονοπάτια από ότι ο άλλος.

Στο πιλοτικό παράδειγμα 1.2 και στις δύο στρατηγικές ο χρυσός (gold) και το ασήμι (silver) έχουν ως κοινούς προγόνους τους όρους πολύτιμα (precious) και κέρμα (coinage). Όμως το παλλάδιο (palladium) και ο χρυσός (gold) με την DCA στρατηγική έχουν ως κοινό πρόγονο μόνο τον όρο πολύτιμα (precious) ενώ με την DiShIn στρατηγική έχουν ως κοινούς προγόνους τους όρους πολύτιμα (precious) και μέταλλο (metal). Σε αυτό το απλό παράδειγμα η DiShIn στρατηγική χαμηλώνει την ομοιότητα όρων όπως το παλλάδιο και ο χρυσός που είναι παράλληλοι (parallel interpretations) το οποίο είναι και ο στόχος της στρατηγικής αυτής σύμφωνα με τους εμπνευστές της. [26]




Σχήμα 1.2: Παράδειγμα επιλογής προγόνων με βάση τη στρατηγική DiShIn [26]

.

Τέλος υπάρχει και μία τρίτη εναλλακτική στρατηγική η xGraSM [27] που λαμβάνει υπόψιν όλους τους κοινούς προγόνους για να βγάλει ένα παράγοντα βάρους/διόρθωσης e, τον όποιο πολλαπλασιάζει στην συνέχεια με το τελικό αποτέλεσμα των μέτρων σημασιολογικής ομοιότητας που βασίζονται στο IC:

$$e = \frac{1}{n}\left(1 + \sum_{j=1}^{n-1} \frac{IC(t_j)}{IC(C_{MICA})}\right) \tag{1.13}$$

Η ίδεα πίσω από την xGraSM στρατηγική είναι ότι ο υπολογισμός των διακριτών κοινών προγόνων είναι υπολογιστικά δαπανήρος οπότε τους κρατάει όλους και σταθμίζει την συνεισφορά τους με το μέγιστο IC, που είναι αυτό του MICA.

### 1.3.3 Μέτρα Συνόλων

Σε αυτή την κατηγορία ανήκουν τα μέτρα που υπολογίζουν την ομοιότητα δύο όρων είτε με βάση το λόγο του αριθμού των κοινών προγόνων (graph corpus) ή γονιδίων (mapping) προς τον αριθμό των συνολικών προγόνων/γονιδίων (Jaccard Coefficient και Dice Coefficient) είτε με βάση το λόγο του IC των κοινών προγόνων προς το IC των συνολικών προγόνων (Mazandu, Graph Information Content).

Παρακάτω παρουσιάζονται οι μετρικές με κριτήριο το σχήμα της οντολογίας, αλλά αντίστοιχα

μπορούν να υπολογιστούν και με βάση τον χαρακτηρισμό σε γονίδια. Έστω ότι οι πρόγονοι του όρου $C_1$ είναι το σύνολο A και του όρου $C_2$ είναι το σύνολο B.

Η μετρική του Jaccard είναι ένα στατιστικό που χρησιμοποιείται για τη σύγκριση των ομοιοτήτων και των διαφορών δύο πεπαρασμένων συνόλων:

$$sim_{Jaccard}(C_1, C_2) = \frac{|A \cap B|}{|A \cup B|} \tag{1.14}$$

Ο Pesquita et al. [28] πρότεινε μια παραλλαγή του Jaccard Coefficient που αντί να υπολογίζει τον αριθμό των κοινών και των διαφορετικών προγόνων χρησιμοποιεί το IC τους:

$$sim_{GIC}(C_1, C_2) = \frac{\sum_{j \in A \cap B} IC(C_j)}{\sum_{i \in A \cup B} IC(C_i)} \tag{1.15}$$

Ο Dice τροποποίησε τον Jaccard ώστε να υπολογίζει τον ομοιότητα ως τον λόγο του αριθμού των κοινών προγόνων με το άθροισμα όλων των προγόνων:

$$sim_{Dice}(C_1, C_2) = \frac{2|A \cap B|}{|A| + |B|} \tag{1.16}$$

Ο Mazandu [29] πρότεινε μια παραλλαγή του Dice που βάζει βάρος στον αριθμό των κοινών και των συνολικών προγόνων το IC του κάθε προγόνου.

$$sim_{Mazandu}(C_1, C_2) = \frac{2 \times \sum_{j \in A \cap B} IC(C_j)}{\sum\limits_{i \in A} IC(C_i) + \sum\limits_{k \in B} IC(C_k)} \tag{1.17}$$

### 1.3.4 Υβριδικά Μέτρα

Πρόκειται για μέτρα που λαμβάνουν υπόψιν τους τόσο το περιεχόμενο των όρων (IC) όσο και την τοπολογία της οντολογίας, δηλαδή τη θέση των όρων στον γράφο.

Η πιο απλή μετρική είναι των Mazandu et al. [30], η οποία διαιρεί το IC του MICA με το IC του όρου που βρίσκεται πιο χαμηλά στην οντολογία, δηλαδή του όρου που είναι πιο εξειδικευμένος από τους δύο:

$$sim_{Nunivers}(C_1, C_2) = \frac{2 \times IC(C_{MICA})}{max(IC(C_1), IC(C_2))} \tag{1.18}$$

Μία αρκέτα πιο σύνθετη προσέγγιση είναι το Αθροιστικό Μέτρο Πληροφορίας (Aggregate Information Content) [31]. Για να υπολογιστεί το IC ενός όρου λαμβάνει υπόψιν όλο τον υπογράφο του, δηλαδή τη σημασιολογικη συνεισφορά (semantic contribution) όλων των προγόνων του.

Με βάση τις παρατηρήσεις ότι **i)** οι όροι στα υψηλότερα επίπεδα της οντολογίας είναι πιο γενικοί και **ii)** οι όροι στα χαμηλότερα επίπεδα είναι περισσότερο εξειδικευμένοι και συνεπώς

περισσότερο μελετημένοι ορίστηκε ως γνώση (knowledge) ενός όρου t:

$$K(t) = \frac{1}{IC(t)} \tag{1.19}$$

Ομώς το μέτρο knowledge δεν παίρνει τιμές στο διάστημα $[0, 1]$. Η κανονικοποίησή του γίνεται με χρήση της λογαριθμικής συνάρτησης (logistic function) και ονομάζεται σημασιολογικό βάρος (semantic weight):

$$SW(t) = \frac{1}{1 + e^{-K(t)}} \tag{1.20}$$

Ως σημασιολογική αξία (semantic value) ενός όρου C ορίζεται το άθροισμα των semantic weights των προγόνων του.

$$SV(C) = \sum_{t \in Ancestors(C)} SW(t) \tag{1.21}$$

Τέλος, η σημασιολογική ομοιότητα δυο όρων είναι ίση με [31] :

$$sim_{AIC}(C_1, C_2) = \frac{2 \times \sum_{t \in Anc(C_1) \cap Anc(C_2)} SW(t)}{SV(C_1) + SV(C_2)} \tag{1.22}$$

### 1.3.5 Στρατηγικές Σύγκρισης Γονιδιακών Προϊόντων

Προηγουμένως περιγράφηκαν οι μετρικές που υπάρχουν για τη σύγκριση δύο όρων της οντολογίας. Οι μετρικές αυτές αποτελούν τη βάση για τη σημασιολογική σύγκριση των γονίδιων και των πρωτεϊνών που χαρακτηρίζονται από πολλούς όρους.

Έστω ότι το γονίδιο $g_1$ έχει χαρακτηριστεί από τους όρους $A =< C_1, C_2, .., C_n >$ και το γονιίδιο $g_2$ από τους όρους $B =< C_1, C_3, .., C_k >$. Η πιο απλή στρατηγική είναι να συγκρίνονται ανά δύο οι όροι που έχουν αντιστοιχηθεί τα δυο γονίδια και η μέγιστη τιμή τους [32] να ορίζεται ως η σημασιολογική τους ομοιότητα:

$$sim_{MAX}(g_1, g_2) = \max_{i \in A, j \in B} sim(i, j) \tag{1.23}$$

Ωστόσο αυτή η στρατηγική δεν λαμβάνει υπόψιν τις διαφορές στον χαρακτηρισμό των γονιδίων και είναι ευαισθήτη σε πιθανά λάθη που συμβαίνουν κατά την αντιστοίχιση όρων με γονίδια [17] . Επιπλέον αν δύο γονίδια χαρακτηρίζονται από τον ίδιο όρο, θα έχουν σημασιολογική ομοιότητα 1.

Μία άλλη στρατηγική που χρησιμοποιεί όλο τον χαρακτηρισμό των γονιδίων, ορίζει ως σημασιολογική ομοιότητα το μέσο όρο (average) της ομοιότητας που έχουν ανά δύο οι όροι που χαρακτηρίζουν αυτά τα γονίδια [33]:

$$sim_{AVG}(g_1, g_2) = \frac{1}{n \times k} \sum_{i \in A} \sum_{j \in B} sim(i, j) \tag{1.24}$$

Ένα πρόβλημα με αυτή τη στρατηγική (all vs all comparison) είναι ότι χαμηλώνει την ομοιότητα δύο γονιδίων που έχουν χαρακτηριστεί με κοινούς όρους [17] .

Δύο στρατηγικές που προσπάθησαν να απαντήσουν στα προβλήματα που έχουν οι max και average στρατηγικές είναι οι Best Match Average(BMA) [34] και Average Best Match(ABM) [35] . Οι στρατηγικές αυτές συγκρίνουν ανά δύο τους όρους που χαρακτηρίζουν τα γονίδια και για κάθε όρο του $g_1$ κρατάνε την μέγιστη ομοιότητα (best match) του με τους όρους του $g_2$ και αντιστρόφως.
Ορίζεται ως best match ενός όρου $C$ με ένα σύνολο όρων $go =< C_1, C_3, .., C_k >$:

$$sim_{BM}(C_1, go) = \underset{i \in go}{max}(sim(i, C_1)) \tag{1.25}$$

Η Best Match Average(BMA) [34] στρατηγική για την σύγκριση δύο γονιδίων $g_1 =< C_1, C_3, .., C_m >$ και $g_2 =< C_1, C_2, .., C_n >$ είναι:

$$sim_{BMA}(g_1, g_2) = \frac{\sum_{i \in g_1} sim_{BM}(i, g_2) + \sum_{j \in g_2} sim_{BM}(j, g_1)}{m + n} \tag{1.26}$$

Τέλος η Average Best Match(ABM) [35] στρατηγική αποτελεί μια παραλλαγή της BMA:

$$sim_{ABM}(g_1, g_2) = 0.5 \left( \frac{\sum_{i \in g_1} sim_{BM}(i, g_2)}{m} + \frac{\sum_{j \in g_2} sim_{BM}(j, g_1)}{n} \right) \tag{1.27}$$

Σε μελέτη [35] που έχει γίνει για την αποτίμηση των max, avg και bma στρατηγικών για την σύγκριση των πρωτεϊνων προτείνεται η bma προσέγγιση.

# Κεφάλαιο 2

# Ανάλυση Σύνθετων Δικτύων και Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων

Σε αυτό το κεφάλαιο παρουσιάζονται συνοπτικά τα σύνθετα δίκτυα: οι τοπολογικές και στατιστικές τους ιδιότητες καθώς και οι τρόποι ανάλυσης τους με χρήση στοιχείων από την Θεωρία Γραφημάτων. Στην συνέχεια περιγράφονται συγκεκριμένες κατηγορίες δικτύων, τα βασικότερα βιολογικά δίκτυα, καθώς και τα λειτουργικά και τοπολογικά χαρακτηριστικά των Δικτύων Πρωτεϊνικών Αλληλεπιδράσεων (Protein Protein Interaction (PPI) Networks).

## 2.1 Στοίχεια από την Θεωρία Γραφημάτων

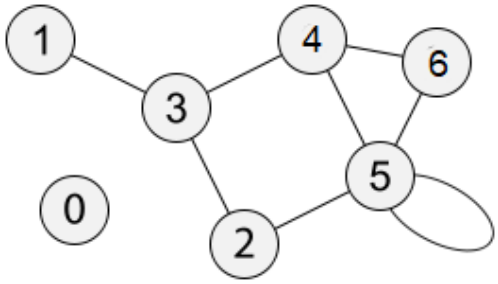
### Βασικοί Ορισμοί

Ένα δίκτυο (network) είναι ένας γράφος μεγάλης κλίμακας ο οποίος χρησιμοποιείται για να μοντελοποιήσει τις αλληλεπιδράσεις μεταβλητών σε πληθώρα τομέων όπως η βιολογία, οι μεταφορές, τα μέσα κοινωνικής δικτύωσης και οι τηλεπικοινωνίες. Διασθητικά, ένας γράφος αναπαριστά τις σχέσεις μεταξύ των στοιχείων ενός συνόλου. Τα στοιχεία του συνόλου είναι οι κόμβοι (nodes) του γράφου και οι συνδέσεις μεταξύ τους ονομάζονται ακμές (edges). Στην συνέχεια παρουσιάζονται φορμαλιστικά οι γράφοι και οι ιδιότητες τους.

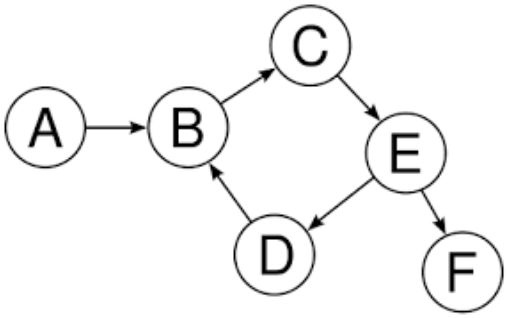
**Ορισμός 1.** *Γράφος είναι ένα διατεταγμένο ζεύγος $G = (V, E)$, όπου $V$ είναι ένα πεπερασμένο σύνολο το οποίο αντιστοιχεί στους κόμβους και $E$ το σύνολο των ακμών το οποίο είναι ένα σύνολο υποσυνόλων του $V$ το καθένα εκ των οποίων έχει δύο στοιχεία του $V$.*

Ο αριθμός των κόμβων $|V(G)|$ ονομάζεται τάξη του γραφήματος G και ο αριθμός των ακμών $|E(G)|$ σχετίζεται με την πυκνότητα του. Ένας γράφος είναι **αραιός** (sparse) όταν $|E| \sim |V|$ και **πυκνός** (dense) όταν $|E| \sim |V|^2$.

**Ορισμός 2.** *Δύο κορυφές $u$ και $v \in V$ που ενώνονται με μία ακμή $e = (u, v) \in E$ λέγονται συνδεδεμένες ή γειτονικές. Συνεπώς οι γείτονες ενός κόμβου $v$ είναι όλοι οι κόμβοι οι οποίοι*




Σχήμα 2.1: Μη κατευθηνόμενος γράφος




Σχήμα 2.2: Κατευθηνόμενος γράφος

*συνδέονται μαζί του μέσω μιας ακμής. Αν ένας κόμβος δεν συνδέεται με κανένα κόμβο λέγεται απομονωμένος (isolated).*

Οι ακμές ενός γράφου μπορεί να είναι κατευθηνόμενες (directed), δηλαδή να υπάρχουν βέλη που δείχνουν προς συγκεκριμένες κατευθύνσεις, όπως φαίνεται στο σχήμα 2.2 και μη-κατευθηνόμενες (undirected) δηλαδή απλές συμμετρικές γραμμές χωρίς προσανατολισμό όπως φαίνεται στο σχήμα 2.1. Συνέπως ένας γράφος μπορεί να είναι είτε κατευθηνόμενος είτε μη-κατευθηνόμενος.

**Ορισμός 3.** *Ως βαθμός ενός κόμβου ορίζεται το πλήθος των ακμών που "συνδέονται" σε αυτόν. Σε ένα κατευθηνόμενο γράφο ο κάθε κόμβος έχει εισερχόμενες (incoming) και εξερχόμενες (outgoing) ακμές, οι οποίες δεν είναι απαραίτητο να είναι ίσες σε αριθμό. Συνεπώς σε ένα μη-κατευθηνόμενο γράφο ο κάθε κόμβος έχει* ***έσω-βαθμό*** *και* ***έξω-βαθμό****. Αντίθετα σε ένα μη-κατευθηνόμενο γράφο οι εισερχόμενες ακμές είναι ίσες με τις εξερχομένες και ο βαθμός είναι ένας.*

Επιπλέον υπάρχουν γράφοι που έχουν self-loops δηλαδή ακμές από έναν κόμβο στον εαυτό του όπως ο κόμβος νούμερο 5 στο σχήμα 2.2. Οι γράφοι που δεν έχουν self-loops ονομάζονται **απλοί** γράφοι.

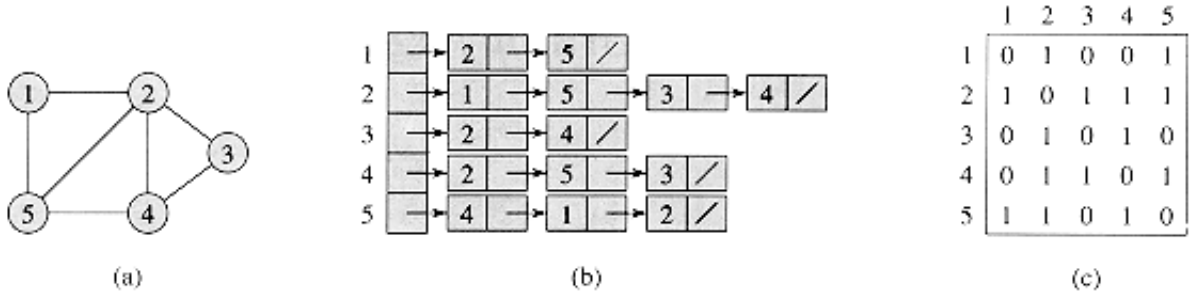
**Ορισμός 4.** ***Συνεκτικός*** *ονομάζεται ένας μη-κατευθηνόμενος γράφος αν για κάθε ζευγάρι κόμβων $u$ και $v$ υπάρχει μονοπάτι, δηλαδή μια ακολουθία ακμών που να ενώνει τους κόμβους $u$ και $v$. Ένας κατευθηνόμενος γράφος που ικανοποιεί αυτή την ίδια ιδιότητα ονομάζεται* ***ισχυρά συνεκτικός****, ενω αν την ικανοποιεί ο μη κατευθηνόμενος γράφος που του αντιστοιχεί ονομάζεται* ***ελαφρά συνεκτικός****.*

Στο σχήμα 2.1 βλέπουμε ότι υπάρχει ένας κόμβος που είναι απομονωμένος (isolated) δηλαδή δεν συνδέεται με κανένα άλλο κόμβο. Συνεπώς αυτός ο γράφος δεν είναι συνεκτικός (connected). Αντίθετα στο σχήμα 2.2 ο γράφος είναι συνεκτικός καθώς δεν υπάρχει isolated κόμβος.

Τέλος οι ακμές μπορεί να έχουν βάρη (weights) w δηλαδή μια τιμή που αντιστοιχεί στο κόστος που έχει η μετακίνηση από έναν κόμβο σε έναν άλλο. Τότε ο γράφος ονομάζεται γράφος με βάρη (weighted graph).

### Αναπαράσταση Γράφων

Υπάρχουν δύο βασικοί τρόποι για την αναπαράσταση των γράφων, ο πίνακας γειτνίασης (adjacency matrix) και η λίστα γειτνίασης (adjacency list).




| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 |
| 2 | 1 | 0 | 1 | 1 | 1 |
| 3 | 0 | 1 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 | 0 | 1 |
| 5 | 1 | 1 | 0 | 1 | 0 |

Σχήμα 2.3: Οι διαφορετικές μορφές αναπαράστασης ενός γράφου. Στο (β) η λίστα γειτνίασης και στο (c) ο πίνακας γειτνίασης για το γράφο του σχήματος (α)

Ο πίνακας γειτνίασης συμβολίζεται με $A = [a_{ij}]$ και έχει διάσταση $n \times n$, όπου $a_{ij} = 1$ όταν υπάρχει η ακμή (i,j) ενώ $a_{ij} = 0$ όταν δεν συνδέονται οι δύο κόμβοι. Σε ενα μη-κατευθηνόμενο γράφο όπως ο γράφος του σχήματος 2.3 ο πίνακας γειτνίασης είναι συμμετρικός.

Η λίστα γειτνίασης όπου για κάθε κόμβο υπάρχει μία λίστα των συνδέσεων-γειτόνων του όπως φαίνεται στην εικόνα (b) του σχήματος 2.3.

Η λίστα γειτνίασης είναι περίσσοτερο space efficient από τον πίνακα ωστόσο συνήθως χρησιμοποιείται ο πίνακας καθώς διευκολύνει, με την χρήση κατάλληλων αλγορίθμων, τον υπολογισμό του πίνακα αποστάσεων(distance matrix) $D = [d_{ij}]$ δηλαδή ενός πίνακα $n \times n$ που περιέχει το μήκος των συντομότερων μονοπατιών που συνδέουν κάθε ζεύγος κόμβων του γράφου.

### Αλγόριθμος Floyd-Warshall

Ο Floyd-Warshall [36] είναι ένας αλγόριθμος δυναμικού προγραμματισμού(dynamic programming) ο οποίος υπολογίζει το distance matrix σε ένα γράφο με βάρη είτε θετικά είτε αρνητικά (χωρίς αρνητικούς κύκλους). Η ιδέα του αλγορίθμου είναι ότι σε κάθε βήμα (επανάληψη) του αλγορίθμου ελέγχεται αν η χρήση ενός κόμβου ως ενδιάμεσου μπορεί να βελτιώσει το μέχρι στιγμής βέλτιστο (μικρότερο σε μήκος) μονοπάτι μεταξύ των ζευγών των κόμβων. Στην συνέχεια παρατίθεται ο ψευδοκώδικας του αλγορίθμου.

**Algorithm 1:** Floyd-Warshall Algorithm

**input :** Adjacency Matrix A
**output:** Distance Matrix

```
1 let dist be a |V| × |V| array of minimum distances initialized to ∞.
2 dist[i][i]=0 and dist[i][j]= w[u][v] if edge (u,v) exists.
3 for k from 1 to |V|:
4   for i from 1 to |V|:
5     for j from 1 to |V|:
6       if dist[i][j] > dist[i][k] + dist[k][j]:
7         dist[i][j] = dist[i][k] + dist[k][j]
```

Με βάση των πίνακα αποστάσεων προκύπτουν οι ακόλουθες χρησίμες ιδιότητες των δικτύων:

**Ορισμός 5.** ***Εκκεντρότητα*** *e(u) μιας κορυφής u σε ένα συνεκτικό γράφο G ορίζεται η μέγιστη απόσταση $d(u, v)$ με $v \in V(G)$. Δηλαδή είναι η μέγιστη απόσταση που έχει ένας κόμβος από τους υπόλοιπους κόμβους του γράφου.*

**Ορισμός 6.** ***Διάμετρος*** *d(G) ενός γράφου G ορίζεται η μέγιστη εκκεντρότητα e(u) για $u \in V(G)$, δηλαδή το μήκος του μέγιστου shortest path και ως* ***ακτίνα*** *r(G) η μικρότερη εκκεντρότητα.*

**Ορισμός 7.** ***Μέσο μήκος*** *(average path) ενός γράφου G μεγέθους N ορίζεται ο μέσος όρος όλων των shortest paths του γράφου δηλαδή ο μέσος όρος του distance matrix:*

$$avg_{path} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} dist(i, j) \tag{2.1}$$

Τόσο η διάμετρος όσο και το μέσο μήκος ενός γράφου G αποτελούν μια πρώτη ένδειξη για την τοπολογία του. Για παράδειγμα ένας γράφος που είναι πυκνός έχει μικρότερη δίαμετρο και μέσο μήκος μονοπατιού από τον αντίστοιχο αραιό γράφο.

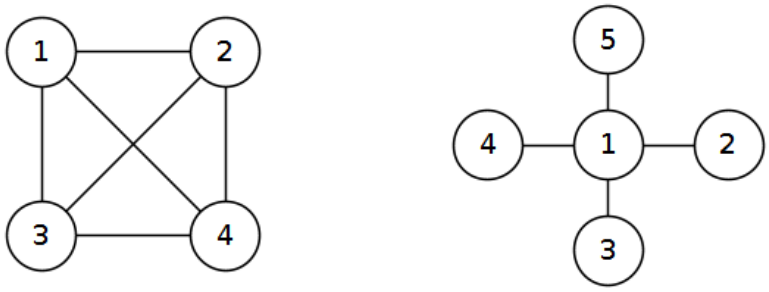
## 2.2 Χαρακτηριστικά των Κόμβων Σύνθετων Δικτύων

Σε αυτό το υποκεφάλαιο παρουσιάζονται τα βασικά χαρακτηριστικά και οι μετρικές που σχετίζονται με την ανάλυση και την αποτίμηση της σημαντικότητας των κόμβων ενός σύνθετου δικτύου.

### 2.2.1 Συντελεστής Ομαδοποίησης

Σε ένα σύνθετο δίκτυο υπάρχουν πυκνές και αραιές περιοχές. Δηλαδή, υπάρχουν κόμβοι που επικοινωνούν πιο εύκολα μεταξύ τους και δημιουργούν ομάδες, οι οποιές στην γλώσσα της Θεωρίας Γραφημάτων ονομάζονται κλίκες. Σε αυτή την υποενότητα θα ορισθεί αυστηρά η έννοια της κλίκας και θα παρουσιασθεί ο συντελεστής ομαδοποίησης, ένα μέτρο που σχετίζεται με την τάση των κόμβων να δημιουργούν κλίκες με τους γειτονές τους.

**Ορισμός 8.** *Ένας γράφος $G = (V, E)$ ονομάζεται πλήρης αν όλοι οι κόμβοι έχουν βαθμό $N-1$, δηλαδή ενώνονται ανα δύο με μια ακμή. Ένας πλήρης γράφος έχει $|E| = \frac{N(N-1)}{2}$ ακμές.*




Σχήμα 2.4: Αριστέρα μία κλίκα 4 κόμβων και δεξιά η τοπολογία αστέρα

**Ορισμός 9.** *Κλίκα ενός γράφου $G = (V, E)$ ονομάζεται ένα σύνολο κόμβων $H \subset V$ το οποίο είναι πλήρες. Τάξη της κλίκας ονομάζεται το μέγεθος του συνόλου $H$.*

**Ορισμός 10.** *Έστω ένας μη-κατευθηνόμενος γράφος $G = (V, E)$ και ένας κόμβος $v$ με βαθμό $k$. Ως συντελεστής ομαδοποίησης ορίζεται ο λόγος του αριθμού των ακμών (έστω $e$) μεταξύ των γειτόνων του $v$ με το μέγιστο αριθμό των ακμών που θα είχαν οι γειτονές του αν ήταν κλίκα:*

$$LCC(v) = \frac{2e}{k(k-1)} \tag{2.2}$$

Ο τοπικός συντελεστής ομαδοποίσης (local clustering coefficient) ενός κόμβου v είναι μια ένδειξη για το πόσο απέχουν οι κόμβοι στην γειτονιά του v από το να δημιουργήσουν κλίκα. Είναι ένα κανονικοποιημένο μέτρο καθώς παίρνει τιμές στο διάστημα $[0, 1]$, με το 1 να υποδηλώνει κλίκα και το 0 τοπολογία αστέρα (star like topology), δηλαδή ότι δεν υπάρχει καμία ακμή μεταξύ των γειτόνων του v και τυχόν αφαίρεση του κόμβου θα οδηγούσε σε πλήρη αποσύνδεση του δικτύου.

**Ορισμός 11.** *Ως μέσος συντελεστής ομαδοποίησης ο οποίος είναι μια ένδειξη για την τάση του δικτύου να οργανώνεται σε κλίκες ορίζεται ο μέσος όρος των τοπικών συντελεστών ομαδοποίησης:*

$$CC_{average} = \frac{1}{N} \sum_{i=1}^{N} \frac{E_i}{k_i(k_i-1)} \tag{2.3}$$

### 2.2.2 Κεντρικότητες

Σε αυτή την ενότητα παρουσιάζονται οι τρόποι ιεράρχησης των κόμβων σε ένα σύνθετο δίκτυο με βάση τη σημασία και την επίδραση τους. Για την ιεράρχηση των κόμβων έχει υιοθετηθεί η

έννοια της κεντρικότητας η οποία είναι μια συνάρτηση που αποδίδει σε κάθε κόμβο μια πραγματική τιμή με βάση τη σημασία του στο συνολικό δίκτυο.

Ο πιο απλός τύπος κεντρικότητας που έχει οριστεί είναι αυτός της **κεντρικότας βαθμού** (degree centrality) ο οποίος αντιστοιχεί σε κάθε κόμβο τον βαθμό που έχει στο δίκτυο. Μια απλή εξήγηση για την χρησιμότητα αυτού του μέτρου είναι ότι όσο πιο πολλές συνδέσεις με άλλους κόμβους έχει ένας κόμβος, τόσο πιο "δημοφιλής" και άρα σημαντικός είναι στη ροή της πληροφορίας μέσα στο δίκτυο. Το μέτρο αυτό μπορεί εύκολα να κανονικοποιηθεί διαιρόντας τον βαθμό του κόμβου με $|V|-1$ που είναι ο μέγιστος βαθμός που δύναται να έχει ένας κόμβος σε ένα γράφο τάξης $|V|$.

Ωστόσο οι βαθμοί των κόμβων δεν είναι ικανή και αναγκαία συνθήκη για τον προσδιορισμό των κεντρικών κόμβων ενός δικτύου. Για παράδειγμα, ένας κόμβος με μικρό βαθμό μπορεί να λειτουργεί ως γέφυρα μεταξύ δύο πυκνών περιοχών του δικτύου. Συνεπώς, προκύπτει η ανάγκη για ορισμό διαφορετικών τύπων κεντρικότητας που μπορεί να είναι υπολογιστικά πιο απαιτητικοί, αλλά παρέχουν μία καλύτερη εικόνα για την συνεισφορά των κόμβων στη ροή της πληροφορίας. Στην συνέχεια θα παρουσιασθούν οι υπόλοιπες κεντρικότητες που υπολογίζονται είτε με βάση την απόσταση που έχουν οι κόμβοι (κεντρικότητα εκκεντρότητας, κεντρικότητα εγγύτητας) είτε με βάση τον αριθμό των συντομότερων μονοπατιών που διέρχονται από αυτόν (κεντρικότητα ενδιαμεσηκότητας) [37].

**Ορισμός 12.** *Η* ***κεντρικότητα εκκεντρότητας*** *(eccentricity centrality) ενός κόμβου $v$ είναι ένα μέτρο που δείχνει πόσο εύκολα προσβάσιμος είναι ο κόμβος από τους υπόλοιπους και σχετίζεται με την μέγιστη απόσταση που έχει ο κόμβος $v$ στο δίκτυο:*

$$EC(v) = \frac{1}{max_{y \neq v} dist(v, y)} \tag{2.4}$$

Οσο μεγαλύτερη είναι η εκκεντρότητα κεντρικότητας τόσο πίο εύκολα προσβάσιμος είναι ένας κόμβος από τους άλλους κόμβους του δικτύου. Αντίθετα κόμβοι με μικρή εκκεντρότητα κεντρικότητας έχουν συνήθως ένα περιφερειακό ρόλο στην λειτουργία του δικτύου.

**Ορισμός 13.** *Η* ***κεντρικότητα εγγύτητας*** *(closenness centrality) ενός κόμβου $v$ είναι ένα μέτρο που ιεραρχεί την θέση των κόμβων σε ένα δίκτυο με βάση το κατά πόσο αυτοί είναι τοπολογικά οι πιο κεντρικού κόμβοι. Συγκεκριμένα βασίζεται στο άθροισμα των αποστάσεων του κόμβου $v$ από τους υπόλοιπους.*

$$CC(x) = \frac{1}{\sum_{y \neq v}^{N-1} dist(v, y)} \tag{2.5}$$

Με βάση τον ορισμό ως πιο κεντρικός κόμβος ενός δικτύου ορίζεται αυτός που αθροιστικά έχει τη μικρότερη απόσταση από τους υπόλοιπους κόμβους δηλαδή αυτός από τον οποιο μπορούμε

να φτάσουμε σε όλους τους υπόλοιπους κόμβους με το μικρότερο αθροιστικό κόστος.

Συνήθως χρησιμοποιείται μια τροποιημένη μορφή της κεντρικότητας εγγύτητας η οποία επιτρέπει σε ένα κόμβο που έχει βαθμό $N-1$ δηλαδή συνδέεται με όλους τους άλλους με μια ακμή να έχει κεντρικότητα εγγύτητας 1:

$$CC(x) = \frac{N-1}{\sum_{y \neq v}^{N-1} dist(v,y)} \tag{2.6}$$

Τόσο η κεντρικότητα εκκεντρότητας που αναφέρθηκε προηγουμένως, όσο και η κεντρικότητα εγγύτητας έχουν ως βασικό τους μειονέκτημα ότι ορίζονται αποκλειστικά για συνδεδεμένα δίκτυα αφού σε μη-συνδεδεμένα δίκτυα η απόσταση δύο κόμβων που δεν υπάρχει μονοπάτι μεταξύ τους ορίζεται άπειρη. Η μόνη κεντρικότητα η οποία βρίσκει εφαρμόγη και σε μη-συνδεδεμένα δίκτυα είναι η κεντρικότητα ενδιαμεσικότητας (betweenness centrality).

**Ορισμός 14.** *Η* ***κεντρικότητα ενδιαμεσικότητας*** *(betweennes centrality) ενός κόμβου $v$ είναι ένα μέτρο που υπολογίζει την ροή της πληροφορίας από τον κόμβο $v$ με βάση τον αριθμό των συντομότερων μονοπατιών (από όλους τους κόμβους του δικτύου προς όλους τους κόμβους του δικτύου) που διέρχονται από αυτόν:*

$$BC(v) = \sum_{a \neq v \neq b} \frac{|sh.paths_{ab}(x)|}{|sh.paths_{ab}|} \tag{2.7}$$

Είναι η πιο σημαντική κεντρικότητα καθώς βρίσκει τους κόμβους του δικτύου από τους οποίους διέρχονται τα περισσότερα μονοπάτια δηλαδή αυτούς που έχουν ρόλο διαμεσολαβητή. Τυχόν αφαιρεσή αυτών των κόμβων στην καλύτερη περίπτωση θα αύξανε την διάμετρο και το μέσο μήκος των μονοπατιών και στην χειρότερη περίπτωση το δίκτυο θα έπαυε να είναι συνεκτικό.

Ωστόσο το βασικό μειονέκτημα της συγκεκριμένης κεντρικότητας είναι ότι προϋποθέτει τον υπολογισμό του συνολικού αριθμού των συντομότερων μονοπατιών σε ένα δίκτυο. Να σημειωθεί πως σε ένα σύνθετο δίκτυο είναι πιθανόν να υπάρχουν περισσότερα από ένα συντομότερα μονοπάτια που να συνδέουν δύο κόμβους. Ακόμα και με τον state-of-the-art αλγόριθμο που προτάθηκε από τον Brandes [38] η πολυπλοκότητα για τον υπολογισμό της κεντρικότητας ενδιαμεσικότητας είναι $O(nm)$ για δίκτυα χωρίς βάρη και $O(nm + n^2 \log n)$ για δίκτυα με βάρη, όπου n ο αριθμός των κόμβων του δικτύου και m ο αριθμός των ακμών.

Ο Brandes υπολογίζει αναδρομικά τον συνολικό αριθμό των συντομότερων μονοπατιών του δικτύου. Ορίζει πως ο αριθμός των συντομότερων μονοπατιών που έχει ένας κόμβος v προς τον κόμβο u είναι το άθροισμα του αριθμού των μονοπατιών που έχουν οι προκατοχοί (predecessors) του προς τον κόμβο u. Το σύνολο των προκατόχων ενός κόμβου $v$ είναι οι $Pred(v) = \{t : (t,v) \in E, d(u,v) = d(u,t) + 1\}$ δηλαδή οι γείτονες του οι οποίοι απέχουν απόσταση από

τον u μειώμενη κατά 1.

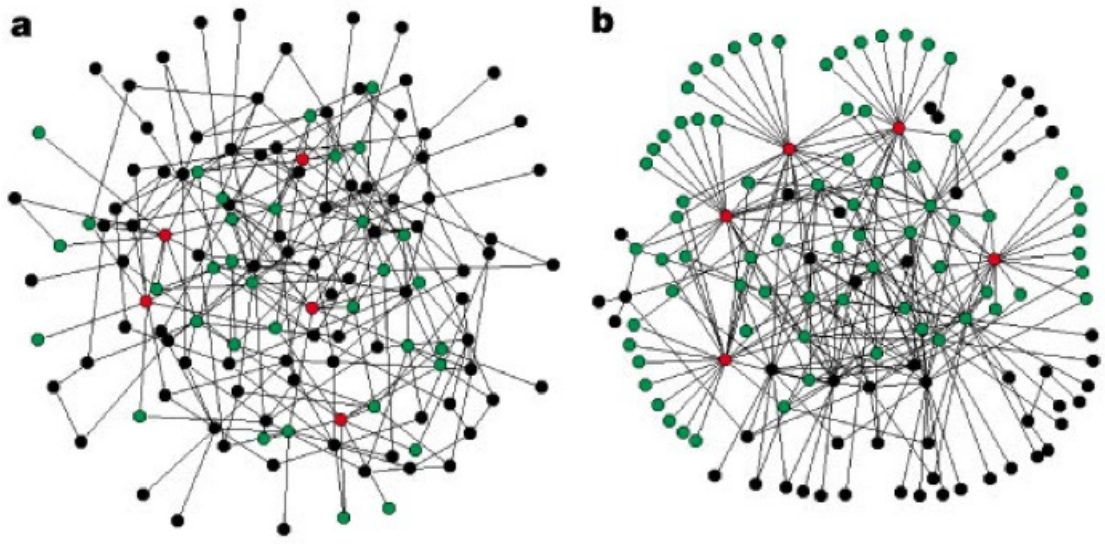
Στην συνέχεια παρατίθεται ο ψευδοκώδικας του αλγορίθμου με βάση την υλοποίηση που πρότεινε ο Brandes [38]:

**Algorithm 2:** Brandes Algorithm

**input :** Graph G=(V,E)
**output:** Betweeness Centrality $C_B[u], u \in V$

$C_B[u] \leftarrow 0, u \in V$
**for** $s \in V$ **do**
  $S \leftarrow$ Empty Stack
  Pred[w]$\leftarrow$ Empty List, $w \in V$  *//list of predecessors on shortest paths from source*
  $\sigma[t] \leftarrow 0, t \in V; \sigma[s] \leftarrow 0$  *//number of shortest paths from source to* $u \in V$
  $dist[t] \leftarrow -1, t \in V$; $dist[s] \leftarrow 0$  *//distance from source*
  $Q \leftarrow$ Empty Queue
  Q.enqueue(s)
  **while** Q *not empty* **do**
    Q.dequeue(u)
    S.push(u)
    **foreach** *neighbor w of u* **do**
      *//w found for the first time*
      **if** $dist[w] < 0$ **then**
        Q.enqueue(w)
        $dist[w] \leftarrow dist[u] + 1$
      **end**
      *//shortest path to w via u*
      **if** $dist[w] = dist[u] + 1$ **then**
        $\sigma[w] \leftarrow \sigma[w] + \sigma[u]$
        Pred[w].append(u)
      **end**
    **end**
  **end**
  $\delta[u] \leftarrow 0, u \in V$  *//dependency of source on* $u \in V$
  *//S returns vertices in order of non-increasing distance from s*
  **while** S *not empty* **do**
    S.pop(w)
    **for** $u \in Pred[w]$ **do** $\delta[u] \leftarrow \delta[u] + \frac{\sigma[u]}{\sigma[w]} \cdot (1 + \delta[w])$
    **if** $w \neq s$ **then** $C_B[w] \leftarrow C_B[w] + \delta[w]$
  **end**
**end**

## 2.3 Τοπολογικές Ιδιότητες των Δικτύων




Σχήμα 2.5: Οι διαφορετικές τοπολογίες ενός γράφου [39]. Στο (α) ένα τυχαίο δίκτυο και στο (β) ένα δίκτυο ελεύθερης κλίμακας.

**Ορισμός 15.** *Τα δίκτυα χωρίζονται σε δύο βασικές κατηγορίες τα* ***δυναμικά*** *τα οποία έχουν μεταβαλλόμενο μέγεθος και τα* ***στατικά*** *που έχουν σταθερό μεγέθος.*

Ενά κομβικό ζήτημα στην μελέτη ενός σύνθετου (complex) δίκτυου είναι ο προσδιορισμός της τοπολογίας του, η οποία συνδέεται αμέσα με την κατανομή των βαθμών του κόμβου του δικτύου. Όπως έχει ήδη αναφερθεί ο βαθμός ένος κόμβου αναφέρεται στον αριθμό των γειτόνων του. Η κατανομή των βαθμών των κόμβων $P(k)$ είναι η πιθανότητα ένας τυχαία επιλεγμένος κόμβος να έχει ακριβώς k γείτονες.

Για να διερευνηθεί η τοπολογία ενός δικτύου πρέπει αυτό να είναι συνεκτικό δηλαδή να μην υπάρχουν απομονωμένοι όροι ή ομάδες όρων. Αν το δίκτυο δεν είναι συνεκτικό πρέπει πρώτα να βρεθεί η μεγαλύτερη συνεκτική συνιστώσα (giant component) του. Για την εύρεση της αρκεί να βρούμε τους κόμβους με τους οποίους επικοινωνεί ο κόμβος με την μεγαλύτερη κεντρικότητα βαθμού δηλαδή το μεγαλύτερο βαθμό. Το οποίο γίνεται εύκολα με χρήση αλγορίθμων διάσχισης γράφων όπως ¨Η Αναζήτηση Κατά Πλάτος' (Breadth First Search) [36] η οποία ξεκινά από ένα κόμβο αφετηρίας s και εξερευνά πρώτα τους γείτονες του s και έπειτα τους γείτονες των γειτονων του. Στην συνέχεια παρατίθεται ο ψευδοκώδικας του αλγορίθμου.

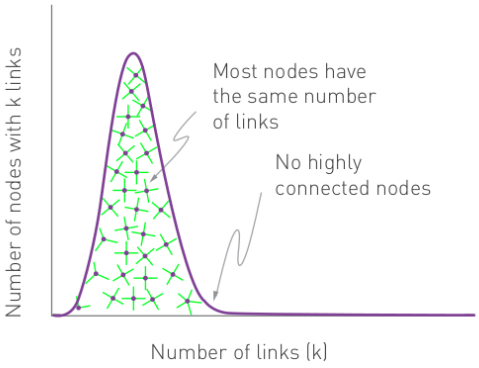
**Algorithm 3:** Breadth First Search

**input :** Graph G, Source Node S
**output:** Connected Component of S

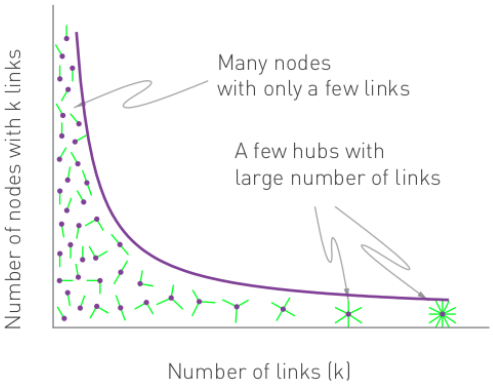
```
1 let Q be queue.
2 Q.enqueue(s)  //Inserting s in queue until all its neighbour vertices are marked.
3 mark s as visited.
4 while (Q is not empty):
5   v = Q.dequeue()  //Removing the vertex from queue,whose neighbour will be
    visited now.
6     for all neighbours w of v:
7       if w is not visited:
8         Q.enqueue(w)
9         mark w as visited.
```

### Τυχαία Δίκτυα

Η βασική τοπολογία η οποία πίστευαν οι επιστήμονες ότι διατρέχει όλα τα δίκτυα είναι το μόντελο του **Τυχαίου Δικτύου** (Random network) το οποίο προτάθηκε απο τους Paul Erdos και Alfred Renyi το 1959. Η βασική ιδέα των δύο Ούγγρων μαθηματικών ήταν ότι για να μοντελοποιήθει η συμπεριφόρα των δικτύων αρκούσε να συνδέονται οι κόμβοι με τυχαίες ακμές(random placement of links).




Σχήμα 2.6: Παράδειγμα κατανομής βαθμών σε τυχαίο δίκτυο [41]




Σχήμα 2.7: Παράδειγμα κατανομής βαθμών σε δίκτυο ελεύθερης κλίμακας [41].

Ωστόσο παρά την τυχαία τοποθέτηση των ακμών η πλειοψηφία των κόμβων έχει περίπου τον ίδιο βαθμό, κόντα στην μέση τιμή των βαθμών. Επιπλέον είναι πολύ σπάνιο ένας κόμβος να έχει σημαντικά μεγαλύτερο ή μικρότερο βαθμό από τον μέσο βαθμό. Συνεπώς όπως φαίνεται και στο σχήμα 2.6 η κατανομή των βαθμών των κόμβων ακόλουθεί κατανομή Poisson με κορυφή στην μέση τιμη των $P(k)$. Τα τυχαία δίκτυα ονομάζονται και εκθετικά καθώς η πιθανότητα

ένας κόμβος να συνδεθεί με άλλους k κόμβους μείωνεται εκθετικά για μεγάλα k.

Τα τυχαία δίκτυα χαρακτηρίζονται απο την ιδιότητα του ¨μικρού κόσμου' (small world property) $l \sim \log N$ [40] που ορίζει ότι το μέσο μήκος των μονοπατιών του δικτύου είναι ανάλογο του λογαρίθμου του μεγέθους του δικτύου.

Με γνώμονα το τυχαίο μοντέλο το 1999 ο Alberto Barabasi et al [41] προσπάθησαν να μοντελοποιήσουν το Διαδίκτυο (World Wide Web) ορίζοντας ως κόμβους τις σελίδες του διαδυκτίου και ακμές τις συνδέσεις μεταξύ τους. Περίμεναν ότι λόγω του μεγάλου όγκου και του διαφορετικού περιεχομένου των σελίδων η κατανομή των βαθμών των σελίδων θα ακολουθούσε το τυχαίο μοντέλο. Ωστόσο τα πειραματικά αποτελέσματα δεν επιβεβαίωσαν την προσδοκία των ερευνητών. Το 80% των σελίδων είχαν λιγότερες από τέσσερις ακμές και μόνο το 0.01% των σελίδων είχαν πάνω απο 1000 συνδέσεις. Συνεπώς παρατήρησαν ότι χρειάζεται ένα διαφορετικό μοντέλο για να περιγράψει σύνθετα δίκτυα σαν το World Wide Web τα οποία δεν παρουσιάζουν ομοιομορφία στην κατανομή των βαθμών των κόμβων το οποίο ονόμασαν **Ελεύθερης Κλίμακας** (Scale Free Topology)[40].

### Δίκτυα Ελεύθερης Κλίμακας

Η βασική παρατήρηση του Barabasi ήταν ότι σε αντίθεση με τα τυχαία δίκτυα που σχεδόν όλοι οι κόμβοι έχουν τον ίδιο βαθμό σε ένα δυναμικό δίκτυο μεγάλης κλίμακας σαν το Διαδίκτυο η πλειοψηφία των κόμβων έχει μικρό βαθμό και υπάρχουν ελάχιστοι κόμβοι οι οποίοι έχουν πολύ μεγάλο βαθμό [40]. Τους κόμβους αυτούς τους ονόμασε κεντρικούς κόμβους (hub nodes). Η παρουσία των οποίων εξηγείται από τον δυναμικό χαρακτήρα των δικτύων και από τον μηχανισμό της επιλεκτικής πρόσδεσης (preferential attachment) ο οποίος ορίζει ότι όταν ένας καινούργιος κόμβος εισέρχεται σε ένα δίκτυο προσδένεται σε κάποιον κόμβο με μεγάλο βαθμό δηλαδή σε ένα κεντρικό κόμβο [41]. Με αποτέλεσμα οι κεντρικοί κόμβοι με την πάροδο του χρόνου να αποκτούν όλο και περισσότερες συνδέσεις. Συνεπώς και στα δυναμικά δίκτυα ισχυέι ο κανόνας του Μαρκόβνικοφ: ¨Ο πλούσιος γίνεται πλουσιότερος¨.

Ο Barabasi απέδειξε ότι η κατανομή που μοντελοποιεί καλύτερα δυναμικά δίκτυα μεγάλης κλίμακας είναι η κατανομή νόμου δύναμης (**power-law**) η οποία φαίνεται στο σχήμα 2.7 :

$$P(k) = k^{-\gamma}, 2 < \gamma < 3 \tag{2.8}$$

**Ορισμός 16.** *Ένα δίκτυο θα λέμε ότι είναι Ελεύθερης Κλίμακας (scale-free) αν η κατανομή των βαθμών των κόμβων $P(k)$ ακολουθεί κατανομή νόμου δύναμης (power-law distribution).*

Οι κεντρικοί κόμβοι λειτουργούν ως γέφυρα (διαμεσολαβητές) στην επικοινωνία δύο κόμβων με μικρό βαθό με αποτέλεσμα τα δίκτυα ελεύθερης κλίμακας να χαρακτηρίζονται από την ιδιότητα του ¨εξαιρετικά μικρού κόσμου' (ultra small world property): $l \sim \log\log N$ [41] οπού $l$ είναι το μέσο μήκος του μονοπατιού στο δίκτυο.

Ωστόσο η εξάρτηση των δικτύων ελεύθερης κλίμακας από τους κεντρικούς κόμβους δημιουργεί και προβλήματα. Για παράδειγμα παρουσιάζουν μεγάλη αντοχή σε τυχαίες επιθέσεις (random

attacks) δηλαδή αν αφαιρεθούν τυχαία κόμβοι από το δίκτυο δεν θα αυξηθεί σημαντικά η διάμετρος και το μέσο μήκος των μονοπατιών και συνεπώς δεν θα διαταραχθεί η ομαλή λειτουργία του. Αντίθετα αν υπάρξει στοχευμένη επίθεση στους κεντρικούς κόμβους του δικτύου είναι πιθανόν να κατακερματιστεί το δίκτυο σε πολλά επιμέρους υποδίκτυα, να δημιουργηθούν απομονωμένοι κόμβοι και συνεπώς να καταρρεύσει η επικοινωνία του[39].

Ο Barabasi υποστηρίζει ότι η τοπολογία Ελεύθερης Κλίμακας χαρακτηρίζεται απο καθολικότητα (universallity) [40]. Δηλαδή οι ιδιότητες που απορρέουν από αυτή διέπουν όλα τα σύνθετα δίκτυα Ελεύθερης Κλίμακας ανεξάρτητα από το είδος τους. Τέλος, απέδειξε ότι η πλειοψηφία των σύνθετων δικτύων είτε είναι το διαδύκτιο είτε τα βιολογικά δίκτυα που θα παρουσιασθούν στην συνέχεια μοντελοποιούνται από αυτή την τοπολογία.

## 2.4 Βιολογικά Δίκτυα

Οι ζωντανοί οργανισμοί αποτελούν εξαιρετικά πολύπλοκα συστήματα τα οποία αλληλεπιδρούν και εξελίσσονται διαρκώς σε μεταβαλλόμενα περιβάλλοντα. Η μοντελοποίηση και η μελέτη της πολυπλοκότητας των βιολογικών συστημάτων περνάει έτσι υποχρεωτικά μέσα από τα βιολογικά δίκτυα τα οποία εντάσσονται στον κλάδο της Βιολογίας Συστημάτων. Για παράδειγμα οι διάφορες αλληλεπιδράσεις μεταξύ των μορίων ενός κυττάρου δημιουργούν δίκτυα τα οποία δεν είναι ανεξάρτητα αλλά σχηματίζουν υπερδύκτια, που καθορίζουν τη συμπεριφορά του κυττάρου ή ολόκληρων οργανισμών.

Η μοντελοποίηση και η ανάλυση αυτών των δικτύων σε ό,τι αφορά τις δομικές (μέγεθος, διάφορες επιμέρους υποομάδες(functional modules)) και τις στατιστικές τους ιδιότητες μας επιτρέπει να προσεγγίσουμε σε βάθος πολύ σημαντικά ερωτήματα σχετικά με τη ρύθμιση βιολογικών διεργασιών, την οργάνωση των κυττάρων σε επίπεδο συστημάτων αλλά και την ανάδυση γενικότερων ιδιοτήτων που αντανακλούν τον τρόπο με τον οποίο τα πολύπλοκα βιολογικά συστήματα εξελίσσονται.

Τα βασικά βιολογικά δίκτυα είναι [37]:

1. **Μεταβολικά-Βιοχημικά δίκτυα** (Metabolic-Biochemical networks) που μοντελοποίουν τα ένζυμα που καταλύουν τις μεταβολικές αντιδράσεις και τους μεταβολίτες που αποτελούν τα υποστρώματα και τα προϊόντα των αντιδράσεων αυτών.

2. **Μεταγραφικά-Ρυθμιστικά δίκτυα** (Transcriptional-Regulatory networks(GRNs)) που μοντελοποιούν τον τρόπο που οι πρωτεΐνες και οι μεταγραφικοί παράγοντες (transcription factors) εμπλέκονται στην διαδικασία της έκφρασης των γονιδίων.

3. **Δίκτυα Μεταγωγής Σήματος** (Signal-Trasduction networks) που μοντελοποίουν τον τρόπο μετάδοσης του σήματος από τον έξωκυττάριο στον ενδοκυττάριο χώρο.

4. **Δίκτυα Ανθρώπινων Ασθενειών** (Human Diseases networks) που αναπαριστουν τις διάφορες ασθένειες και τις σχέσεις που υπάρχουν μεταξύ τους.

5. **Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων** (Protein Protein Interaction(PPI) Networks) που μοντελοποιούν τις διάφορες αλληλεπιδράσεις μεταξύ των πρωτεΐνων.

Με την αυστηρή έννοια του όρου όλα τα βιολογικά δίκτυα είναι Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων εφόσον περιέχουν πρωτεΐνες μεταξύ των στοιχείων τους. Για το λόγο αυτό, διακρίνουμε τα Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων από τα υπόλοιπα όταν δεν μπορούμε με σαφήνεια να αποδώσουμε τη φύση της αλληλεπίδρασης ή όταν συνηπάρχουν αλληλεπιδράσεις διαφορετικών τύπων (π.χ. ρυθμιστικές και σηματοδοτικές)[42].

Τα Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων αναπαρίστονται μέσω μη-κατευθηνόμενων γράφων όπου οι κόμβοι είναι οι πρωτεϊνες και οι ακμές οι αλληλεπιδράσεις μεταξύ των πρωτεϊνών. Είναι τα πιο μελετημένα από ολά τα βιολογικά δίκτυα που αναφέρθηκαν καθώς είναι απαραίτητα σχεδόν σε κάθε διαδικασία που συμβαίνει σε ένα κύτταρο. Το σύνολο των πρωτεϊνικών αλληλεπιδράσεων σε ένα οργανισμό ονομάζεται interactome.

Τα PPI συνηθώς δεν είναι συνεκτικά αλλά αποτελούνται από μια μεγάλη συνιστώσα (giant component) που περιέχει την πλειονότητα των συνδεδεμένων πρωτεϊνων και μερικές πρωτεϊνες που είτε είναι τελείως απομονωμένες είτε οργανώνονται σε μικρότερες συνιστώσες [43]. Η κατανομή των βαθμών των κόμβων ακολουθεί Νόμο Δύναμης και συνεπώς η τοπολογία είναι Ελεύθερης Κλίμακας. Μια εξήγηση για την τοπολογία αυτή είναι οτί κατά την διαδικασία του διπλασιασμού των γονιδίων (gene duplication) προστίθεται άλλος ένας κόμβος στο δίκτυο ο οποίος συνδέεται με τους ίδιους ή σχεδόν τους ίδιους κόμβους του γονιδίου/πρωτεΐνης που αντέγραψε [44] και συνεπώς ο βαθμός των κεντρικών κόμβων όλο και αυξάνεται.

Ωστόσο ο συντελεστής ομαδοποιήσης εξαρτάται από τον βαθμό του κόμβου το οποίο δεν αποτελεί χαρακτηριστικό ούτε των τυχαίων δικτύων ούτε των δικτύων ελεύθερης κλίμακας. Η κατανομή του συντελεστή ομαδοποίησης ακολουθεί νόμο δύναμης: $C(k) \sim k^{\beta}$ δηλαδή οι κόμβοι (πρωτεΐνες) που έχουν μικρό βαθμό τείνουν να λειτουργούν ως ομάδες (modules) [44]. Να σημειωθεί οτί έχει αποδειχθεί ότι οι πρωτεΐνες που βρίσκονται σε μια ομάδα συμμετέχουν σε κοινές βιολογικές λειτουργίες [43]. Λόγω της τάσης των πρωτεΐνων με μικρό βαθμό να δημιουργούν ομάδες οι οποίες δεν συνδέονται μεταξύ τους αλλά επικοινωνούν μέσω των κεντρικών κόμβων τα δίκτυα αυτά χαρακτηρίζονται ως Δίκτυα Ελεύθερης Κλίμακας με ιεραρχική ομαδοποίηση (hierarchical modularity) [43].

Οι ομάδες που εμφανίζονται στα Πρωτεϊνικά Δίκτυα σύμφωνα με τον Barabasi είναι ένα δομικό χαρακτηριστικό όλων των βιολογικών δικτύων [44]. Για να βρεθούν οι ομάδες των πρωτεϊνών απαιτείται χρήση αλγορίθμων ομαδοποίησης (clustering algorithms). Ο state-of-the-art[45],[46] αλγόριθμος για την εύρεση των ομάδων στα Πρωτεΐνικά Δίκτυα είναι ο Markov Clustering Algorithm [47] ο οποίος προσωμοιώνει την στοχαστική ροή πάνω στο γράφο. Στο επόμενο κεφάλαιο θα παρουσιασθούν εκτενώς οι βασικοί αλγόριθμοι ομαδοποίησης.

# Κεφάλαιο 3

# Αλγόριθμοι Ομαδοποίησης

Δεδομένου ότι στην εποχή μας ο όγκος των διαθέσιμων δεδομένων όλο και αυξάνεται η αναλύση και η ομαδοποίηση τους δεν μπορεί να γίνει χειροκίνητα. Συνεπώς οι μέθοδοι που ανιχνεύουν δομές και ομάδες στα δεδομένα γίνονται όλο και πιο σημαντικες. Υπάρχουν οι μέθοδοι επιβλεπόμενης μάθησης (Supervised Learning) οι οποίες δοθέντος ένος συνόλου δεδομένων με τις αντίστοιχες ετικέτες (labels) κάθε εγγραφής προσαρμόζουν ένα μοντέλο το οποίο κατηγοριοποιήει νέα δεδομένα σε κάποια από τις προϋπάρχουσες κλάσεις και οι μέθοδοι μη-επιβλεπόμενης μάθησης (Unsupervised Learning) για τα δεδομένα που δεν έχουν ετικέτες και η κατηγοριοποιήση τους γίνεται με βάση τα δομικά χαρακτηριστικά τους. Η πιο γνωστή μέθοδος μη-επιβλεπόμενης μάθησης είναι η **Ομαδοποίηση** (clustering), δηλαδή η διαμέριση ενός συνόλου αντικειμένων σε υποσύνολα (ομάδες) έτσι ώστε τα στοιχεία σε κάθε υποσύνολο να έχουν όμοια ή παραπλήσια χαρακτηριστικά και διαφορετικά από τα αντικείμενα των άλλων. Η ομαδοποίηση μπορεί να είναι είτε ασαφής δηλαδή ένα στοιχείο να ανήκει σε παραπάνω από μία ομάδες είτε αυστηρή στην οποία κάθε στοιχείο ανήκει σε μία και μόνο ομάδα. Στο κεφάλαιο αυτό θα παρουσιασθούν οι αλγόριθμοι αυστηρής ομαδοποίησης.

**Ορισμός 17.** *Δοθέντος ενός συνόλου διανυσμάτων* $X = X_1, X_2, X_3..., X_n$ *ζητούνται* $m$ *ομάδες* $C_1, C_2, C_3, ..., C_n$ *έτσι ώστε:* $\bigcup_{i=1}^{m} C_i = X$ *και* $C_i \neq \oslash, \forall i = 1, 2, 3...m.$

## 3.1 Κατηγορίες Αλγορίθμων Ομαδοποίησης

Οι αλγόριθμοι ομαδοποίησης χωρίζονται σε κατηγορίες με βάση την στρατηγική που χρησιμοποιούν για να ορίσουν την ομοιότητα μεταξύ των στοιχείων. Οι βασικές κατηγορίες είναι [48]:

1. **Μέθοδοι Κεντροείδων** (Centroid Method) οι οποίες αποκαλούνται και μέθοδοι διαμέρισης καθώς χωρίζουν το σύνολο των δεδομένων σε k ομάδες. Στην κατηγορία αυτή εντάσσονται επαναληπτικοί αλγόριθμοι οι οποίοι κατηγοριοποιούν τα δεδομένα με βάση την αποστασή τους από το καλύτερο αντιπρόσωπο της κάθε ομάδας το κεντροειδές. Συνεπώς το πρόβλημα της ομαδοποίσης ανάγεται στην εύρεση των κεντροειδών της κάθε

ομάδας και στην αντιστοίχηση των δεδομένων στο κοντινότερο. Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία (π.χ. k-means,k-medoids) δέχονται ως είσοδο από τον χρήστη τον αριθμό των clusters.

2. **Ιεραρχικοί Μέθοδοι** (Hierarchical Methods) οι οποίες αποκαλούνται και Μέθοδοι Συνδεσιμότητας καθώς βασίζονται στην παραδοχή ότι τα σημεία που βρίσκονται εγγύτερα στον χώρο των δεδομένων παρουσιάζουν μεγαλύτερη ομοιότητα από ότι τα σημεία που βρίσκονται πιο μακριά. Οι μέθοδοι αυτοί χωρίζονται με βάση τον τρόπο που γίνεται η ιεραρχική αποσύνθεση των δεδομένων σε δύο υποκατηγορίες τις μεθόδους συσσώρευσης (**agglomerative**) και τις μεθόδους διαίρεσης (**division**). Οι πρώτες είναι μέθοδοι από την βάση προς την κορυφή(bottom-up) δηλαδή αρχικά κάθε αντικείμενο αποτελεί μια ξεχωριστή ομάδα και στην συνέχεια όσο μειώνονται οι αποστάσεις συνενώνεται με τα αντικείμενα ή τις ομάδες που βρίσκονται πιο κοντά του, μέχρις ότου όλες οι ομάδες να συγχωνευθούν σε μια. Αντίθετα οι μέθοδοι διαίρεσης ξεκινούν θεωρόντας όλα τα σημεία μια ομάδα και σε κάθε βήμα μία ομάδα υποδιαιρείται σε δύο μέχρι να καταλήξουμε σε N ομάδες. Ο πιο γνωστός αλγόριθμος σε αυτή την κατηγοριά είναι ο Αλγόριθμος Ιεραρχικής Ομαδοποίησης.

3. **Μέθοδοι Κατανομών** (Distribution Methods) οι οποίες βασίζονται στην πιθανότητα τα στοιχεία στην ίδια ομάδα να προέρχονται από την ίδια κατανομή (π.χ Κανονική Κατανομή). Ο πίο γνωστός αλγόριθμος σε αυτή την κατηγορία είναι ο αλγόριθμος μέγιστης προσδοκίας (expectation-maximization) ο οποίος βασίζεται στην πολυπαραμετρική κανονική κατανομή (multivariate normal distribution).

4. **Μέθοδοι Πυκνότητας** (Density Methods) οι οποίες χωρίζουν τα δεδομένα με βάση την πυκνότητα των σημείων στο χώρο δεδομένων. Περιοχές με υψηλή πυκνότητα σημείων θεωρούνται ως μια ομάδα και τα σημεία στις πιο αραιές περιοχές χαρακτηρίζονται ως διαχωριστικό μεταξύ των ομάδων και θεωρούνται συνήθως θόρυβος. Ο πιο γνωστός αλγόριθμος που βασίζεται στην πυκνότητα των δεδομένων είναι ο DBSCAN.

## 3.2 Αλγόριθμος k-means

Ο αλγόριθμος k-means[49] αποτέλει ένα τυπικό παράδειγμα μεθόδου διαμέρισης. Δέχεται ως είσοδο την παράμετρο k και χωρίζει τα σημεία $X = \{x_1, x_2, ..., x_n\}$ σε k ομάδες με στόχο οι αποστάσεις εντός της ομάδας (intra-distance) να ελαχιστοποιούνται και οι αποστάσεις μεταξύ των ομάδων (inter-distance) να μεγιστοποιούνται. Η ομοιότητα των στοιχείων της κάθε ομάδας υπολογίζεται με βάση την απόσταση τους από την μέση τιμή των στοιχείων της ομάδας η οποιά στην βιβλιογραφία αναφέρεται ως κεντροείδες.

Ο αλγόριθμος ξεκινά επιλέγοντας τυχαία k σημεία από τα δεδομένα ως κεντροειδή. Κάθε ένα από τα υπόλοιπα στοιχεία ανατείθεται στην ομάδα από την οποία απέχει την μικρότερη απόσταση συνήθως χρησιμοποιείται η ευκλείδια απόσταση: $\underset{i\in[1,n]}{min} \|x_i - \mu_k\|^2$. Στην συνέχεια επαναυπολογίζει το κεντροειδές της κάθε ομάδας ως το μέσο όρο των συντεταγμένων των

στοιχείων που ανήκουν σε αυτή. Ο αλγόριθμος επαναλαμβάνει αυτά τα δύο βήματα είτε μέχρις ότου το αποτέλεσμα του clustering να μην αλλάζει δηλαδή να επιτευχθεί σύγκλιση σε μια τιμή τερματισμού είτε μετά από ένα συγκεκριμένο αριθμό επαναλήψεων. Συνήθως ως συνθήκη τερματισμού επιλέγεται η ελαχιστοποίηση του τετραγωνικού σφάλματος:

$$E = \sum_{i=1}^{k} \sum_{x_j \in X} |x_j - \mu_i|^2 \tag{3.1}$$

Στην συνέχεια παρατείθεται ο ψευδοκώδικας του αλγορίθμου:

**Algorithm 4:** K-Means Algorithm

**input :** $X = x_1, x_2, ..., x_n$: dataset containing n objects, k: the number of clusters
**output:** A set of k clusters

1 arbitrarily choose k objects from X as the initial clusters centers
2 **repeat**.
3 (re)assign each object to the cluster to which is more similar: $min(|x^i - \mu_k|^2)$
4 update the centroids, i.e. calculate the mean value $\mu_k$ of the objects for each cluster
5 **until** no change

Τα μειονεκτήματα του αλγορίθμου είναι ότι ο χρήστης πρέπει να ορίσει τον αριθμό των ομάδων και η τυχαία αρχικοποίηση των κεντροειδών η οποία μπορεί να οδηγήσει σε διαφορετικές ομαδοποιήσεις για το ίδιο σύνολο δεδομένων. Επίσης όταν υπάρχουν outliers στα δεδομένα, δηλαδή πολύ μεγάλες τιμές οι οποίες επηρεάζουν σημαντικά την μέση τιμή ο αλγόριθμος δεν δουλευέι καλά. Εντούτοις χρησιμοποιείται είτε όταν υπάρχουν ενδείξεις ότι τα clusters έχουν σφαιρικό σχήμα είτε όταν το μέγεθος του dataset είναι μικρό.

Μια παραλλαγή του k-means η οποία δεν επηρεάζεται από την ύπαρξη outliers στα δεδομένα είναι ο αλγόριθμος k-medoids ο οποίος αντί να χρησιμοποιεί ως κεντροειδές την μέση τιμή της ομάδας προτείνει να χρησιμοποιείται το σημείο το οποίο βρίσκεται πιο κοντά στο κέντρο της ομάδας το οποίο ονομάζει medoid.

## 3.3 Αλγόριθμοι Ιεραρχικής Ομαδοποίησης

Όπως έχει ήδη αναφερθεί αυτή η κατηγορία αλγορίθμων είτε ξεκινάει από n ομάδες και σε κάθε βήμα του αλγορίθμου συγχωνεύει δύο ομάδες εως ότου όλα τα στοιχεία να συνενωθούν σε μία ή αντίθετα ξεκινάει από μία ομάδα εως ότου όλα τα στοιχεία να κατακερματιστούν σε n ομάδες. Συνεπώς δημιουργείται μία ιεραρχία εμφολιασμένων ομαδοποιήσεων η οποία μπορεί να αναπαρασταθεί μέσω δενδρογραμμάτων (dendrograms) τα οποία έχουν $n - 1$ επίπεδα με το κάθε επίπεδο να αναπαριστά ένα βήμα του αλγορίθμου. Συνεπώς ένα πλεονέκτημα τους είναι ότι δεν απαιτούν ως είσοδο ένα προκαθορισμένο αριθμό ομάδων, αφού οποιοσδήποτε αριθμός μπορεί να επιτευχθεί κάνοντας τομή στο κατάλληλο επίπεδο του δενδρογράμματος.

Τόσο οι μέθοδοι συσσώρευσης όσο και οι μέθοδοι διαίρεσης βασίζονται στην απόσταση που έχουν οι δύο ομάδες. Οι μεν πρώτες σε κάθε βήμα ενώνουν τις δύο ομάδες με την μικρότερη απόσταση ενώ αντίθετα οι δεύτερες διαιρούν τις ομάδες με την μεγαλύτερη απόσταση. Συνεπώς στους ιεραρχικούς αλγορίθμους κομβικό ρόλο παίζει ο υπολογισμός της απόστασης μεταξύ δύο ομάδων η οποία βασίζεται στην απόσταση d μεταξύ δύο σημείων x και y η οποία μπορεί να υπολογισθεί με τους ακόλουθους τρόπους[48]:

- Ευκλείδια απόσταση: $d(x,y) = \|x-y\|_2 = \sqrt{\sum_i (x_i - y_i)^2}$
- Μέγιστη απόσταση: $d(x,y) = \|x-y\|_\infty = \max_i |x_i - y_i|$
- Απόσταση Μανχάταν: $d(x,y) = \|x-y\|_1 = \sum_i |x_i - y_i|$

Συνήθως επιλέγεται η ευκλείδια απόσταση. Στην συνέχεια θα παρουσιασθούν οι βασικές στρατηγικές που υπάρχουν για τον υπολογισμό της απόστασης μεταξύ δύο ομάδων [48].

Η πιο απλή στρατηγική είναι της ελάχιστης απόστασης ή απλου συνδέσμου (simple link) με βάση την οποία η απόσταση μεταξύ δύο ομάδων $C_i$ και $C_j$ υπολογίζεται με βάση την απόσταση των δύο πιο κοντινών (όμοιων) σημείων από τις διαφορετικές ομαδές:

$$d_{min}(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x,y) \tag{3.2}$$

Ωστόσο αυτή η στρατηγική είναι ευαίσθητη σε θόρυβο και σε ακραίες τιμές. Για τον λόγο αυτό προτάθηκε η στρατηγική μέγιστης απόστασης ή πλήρους συνδέσμου (complete link) η οποία υπολογίζει την απόσταση μεταξύ δύο ομάδων $C_i$ και $C_j$ με βάση την απόσταση των δύο πιο απόμακρων (ανόμοιων) σημείων από τις διαφορετικές ομαδές:

$$d_{max}(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x,y) \tag{3.3}$$

Η στρατηγική μέγιστης απόστασης τείνει να δημιουργεί μικρές και συνήθως κυκλικές ομάδες. Για τον λόγο αυτό προτάθηκε μία στρατηγική η οποία είναι η μίξη των δυο προηγούμενων: ο μέσος όρος ομάδων (group average) που υπολογίζει την μέση τιμή των αποστάσεων μεταξύ κάθε πιθανού ζεύγους σημείων από τις δύο ομάδες.

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} d(x,y) \tag{3.4}$$

Προφανώς η απόσταση που όριζει βρίσκεται ανάμεσα στην ελάχιστη και τη μέγιστη απόσταση. Τέλος έχει μικρότερη ευαισθησία σε θόρυβο και σε ακραίες τιμές (outliers) αλλά ευνοεί και αυτή τις κυκλικές ομάδες.

Μία άλλη στρατηγική είναι η απόσταση κεντρικών σημείων (distance between clusters centroid) η οποία υπολογίζει την απόσταση δύο ομάδων με βάση την απόσταση μεταξύ των κεντροειδών τους:

$$d_{mean}(C_i, C_j) = \|\mu_i - \mu_j\| \tag{3.5}$$

Τέλος υπάρχει η μέθοδος του Ward[50] η οποία ορίζει ότι η απόσταση μεταξύ δύο ομάδων $C_i$ και $C_j$, είναι ίση με την αύξηση του τετραγωνικού σφάλματος της απόστασης των στοιχείων της κάθε ομάδας από το αντίστοιχο κεντροειδές μετά τη συγχώνευση τους στην ομάδα $C_{ij}$:

$$d_{wang}(C_i, C_j) = \sum_{x \in C_i} (x - \mu_i)^2 + \sum_{y \in C_j} (y - \mu_j)^2 - \sum_{k \in C_{ij}} (k - \mu_{ij})^2 \tag{3.6}$$

Κάθε μία από τις στρατηγικές που αναφέρθηκαν έχει τα θετικά και τα αρνητικά της. Στην βιβλιοθήκη scikit-learn της Python είναι υλοποιημένες όλες οι στρατηγικές ωστόσο ως προεπιλεγμένη έχει οριστεί η μέθοδος του Ward.

Στην συνέχεια παρατίθεται ο ψευδοκώδικας του αλγορίθμου για μια στρατηγική απόστασης d:

**Algorithm 5:** Agglomerative Hierarchical Clustering

**input :** $X = \{x_1, x_2, ..., x_n\}$: dataset containing n objects, a distance function d
**output:** Hierarchical Clustering

1 Compute $N \times N$ proximity matrix $D_{ij} = d(C_i, C_j)$
2 Begin with the disjoint clustering $C = \{C_1, C_2, ..., C_n\}$ and level=0.
3 **while** $level \neq n - 1$ **:**
4  find the clusters with the smalest distance in the current clustering $(C_i, C_j) = minD_{ij}$ and merge them, level=level+1
5  Update the distance matrix D, by deleting the rows and columns corresponding to clusters $C_i$ and $C_j$ and adding a row and column corresponding to the newly formed cluster $C_{ij}$
6 **end**

## 3.4 Αλγόριθμος Affinity Propagation

Ο Affinity Propagation [51] είναι ένας σχετικά καινούργιος αλγόριθμος ο οποίος δεν απαιτεί τον προσδιορισμό του αριθμού των ομάδων και βασίζεται στην ανταλλαγή μηνυμάτων μεταξύ των δεδομένων για να εντοπίζει όμοια στοιχεία. Με βάση την λογική του k-medoids υποθέτει ότι σε κάθε ομάδα υπάρχει ένα αντιπροσωπευτικό στοιχείο (examplar) και κάθε στοιχείο στέλνει ένα μήνυμα στο examplar της κάθε ομάδας που αντιστοιχεί στην προτίμηση (affinity) που δείχνει για να ενταχθεί στην ομάδα του.

Ο αλγοριθμος δέχεται ως είσοδο ένα σύνολο στοιχείων $X = \{x_1, x_2, ..., x_n\}$ και μια συνάρτηση s που ποσοτικοποιεί την ομοιότητα που έχουν ανά δύο τα στοιχεία του X. Συνήθως επιλέγεται η αρνητική ευκλείδια απόσταση $s(i, j) = -\|x_i - x_j\|^2$ και συνεπώς δύο στοιχεία $x_i, x_j$ είναι πιο όμοια από ότι τα $x_i, x_k$ όταν $s(i, j) > s(i, k)$.

Η διαγώνιος του πίνακα ομοιότητας s δηλαδή τα στοιχεία $s(i, i)$ αντιπροσωπεύουν την πιθανότητα που έχει ένα στοιχείο να γίνει examplar. Τις τιμές των στοιχείων αυτών τις δίνει ο

χρήστης ως είσοδο στον αλγόριθμο. Οταν δεν υπάρχει κάποια πρότερη γνώση για την δομή των δεδομένων όλα τα στοιχεία της διαγώνιου παίρνουν την ίδια τιμή δηλαδή όλα τα στοιχεία θεωρούνται υποψήφια examplars. Μια τιμή κοντά στην ελάχιστη ομοιότητα $s(i,j)$ παράγει λιγότερες ομάδες, ενώ μια τιμή κοντά ή μεγαλύτερη από τη μέγιστη ομοιότητα παράγει πολλές κλάσεις. Συνήθως επιλέγεται η διάμεσος της ομοιότητας των ζευγών των στοιχείων.

Υπάρχουν δύο είδη μηνυμάτων που ανταλλάσουν τα στοιχεία ο συνδυασμός των οποίων καθορίζει ποια στοιχεία είναι examplars και σε ποια ομάδα ανήκει κάθε στοιχείο. Το πρώτο είναι ένα μήνυμα **υπευθηνότητας** (responsibility) $r(i,k)$ το οποίο στέλνει το σημείο i στον υποψηφίο examplar k και αντιστοιχεί στην προτίμηση που δείχνει το i να ενταχθεί στην ομάδα του k. Το δεύτερο είναι ένα μήνυμα **διαθεσιμότητας** (availability) $a(i,k)$ που στέλνει ο examplar k στο στοιχείο i ενημερωνοντάς το για το αν θεωρεί πως είναι κατάλληλος αντιπρόσωπος για αυτό το στοιχείο με βάση και την προτίμηση που έχουν για αυτό τα άλλα στοιχεία. Στην πρώτη επανάληψη του αλγορίθμου οι διαθεσιμότητες αρχικοποιούνται στο 0 δηλαδή $a(i,k) = 0$.

Η υπευθηνότητα $r(i,i)$ αποτελεί μια ένδειξη ότι το στοιχείο i είναι examplar. Γενικά η υπευθηνότητα υπολογίζεται με βάση την ακόλουθη σχέση:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \tag{3.7}$$

Το μήνυμα διαθεσιμότητας που στέλνει ένας examplar σε ένα στοιχείο i είναι:

$$a(i,k) \leftarrow min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} max(0, r(i',k))\right) \tag{3.8}$$

και η διαθεσιμότητα που δείχνει ένα στοιχείο προς τον εαυτό του είναι:

$$a(k,k) \leftarrow \sum_{i' \neq k} max(0, r(i',k)) \tag{3.9}$$

Ο αλγόριθμος συνεχίζει την διαδικασία ανταλλαγής μηνυμάτων είτε μέχρι ότου οι ομάδες να μην άλλαζουν είτε μέχρι ένα προκαθορισμένο αριθμό επαναλήψεων. Τέλος ως examplars ορίζει τα στοιχεία που έχουν $r(i,i) + a(i,i) > 0$ και κάθε στοιχείο k αντιστοιχίζεται στο examplar που έχει το μέγιστο άθροισμα $r(i,k) + a(i,k)$.

## 3.5 Αλγόριθμος Markov Clustering

Ο Markov Clustering [47] είναι ένας αλγόριθμος μη-επιβλεπόμενης μάθησης που χρησιμοποιείται για την ομαδοποίηση γράφων/δικτύων και βασίζεται στην προσομοίωση της στοχαστικής ροής του γράφου. Ο αλγόριθμος βασίζεται στην παραδοχή ότι σε ένα γράφο οι κόμβοι που ανήκουν σε μία ομάδα θα έχουν πολλές ακμές μεταξύ τους δηλαδή θα τείνουν να δημιουργήσουν κλίκα ενώ αντίθετα θα έχουν λίγες ακμές με κόμβους που ανήκουν σε άλλες ομάδες. Το οποίο σημαίνει ότι η διάδοση της στοχαστικής ροής είναι πιο πιθανό να γίνει από ένα κόμβο σε ένα άλλο

κόμβο της ίδιας ομάδας από ότι σε ένα μιας άλλης ομάδας.

Ο αλγόριθμος δέχεται ως είσοδο τον πίνακα γειτνίασης A ενός γράφου G και τον μετατρέπει σε στοχαστικό δηλαδή ένα πίνακα που οι στήλες του αθροίζουν στην μονάδα και τα στοιχεία του αντιστοιχούν στις πιθανότητες μετάβασης $p(i, j)$ από ένα κόμβο i σε ένα κόμβο j. Στην συνέχεια προσομοιώνει την διάδοση της ροής στο γράφο πολλαπλασιάζοντας το στοχαστικό πίνακα A με τον εαυτό του: $B = A \times A$ (expansion). Λόγω του πολλαπλασιασμού ο πίνακας B περιέχει πιθανότητες μετάβασης μεγάλες αλλά και αρκετές κόντα στο 0 τις οποίες ο αλγόριθμος μηδενίζει για να δημιουργήσει τον αραιό πίνακα $C$. Στο τελευταίο βήμα ξαναυπολογίζει τον πίνακα A ως $A = C \circ C$ (inflation) το οποίο σημαίνει ότι κάθε στοιχείο του πίνακα C πολλαπλασιάζεται με τον εαυτό του. Αυτή η διαδικασία συνεπικουρούμενη από την ακόλουθη στοχαστικοποίηση του πίνακα A ενισχυεί τις πιθανότητες μετάβασης που είναι μεγαλές και μικραίνει ή ακόμα και μηδενίζει αυτές που είναι μικρές. Στην συνέχεια επαναλαμβάνει τα προηγούμενα βήματα του αλγορίθμου μέχρι ένα προκαθορισμένο αριθμό επαναλήψεων. Στο τέλος αυτής της επαναληπτικής διαδικασίας θα προκύψουν κόμβοι που θα έχουν μεταξύ τους μεγάλες πιθανότητες μετάβασης και συνεπώς θα αποτελούν μια ομάδα.

Τέλος παρατίθεται ο ψευδοκώδικας του αλγορίθμου:

**Algorithm 6:** Markov Cluster Algorithm

**input :** Network G, Adjacency matrix A of the network
**output:** The clusters of the network

1 Calculate the column stochastic version of the adjacency matrix A.
2 **while** i $\leq$ max iterations **do**
3  $B = A \times A$, expansion by squaring the matrix
4  $C = Prune(B)$, sparsification of B by pruning low probability terms
5  $A = C \circ C$, inflation by taking power entrywise
6 **end while**

## 3.6 Σύγκριση των Ομαδοποιήσεων

Όπως αναφέρθηκε προηγουμένως υπάρχουν αρκετοί αλγόριθμοι ομαδοποίησης οι οποίοι στα ίδια δεδομένα παράγουν διαφορερικά αποτελέσματα. Συνεπώς προκύπτει το ερώτημα πόσο όμοιες είναι δύο ομαδοποιήσεις. Σε αυτό το υποκεφάλαιο θα παρουσιασθούν οι βασικοί τρόποι για την σύγκριση δύο διαφορετικών ομαδοποιήσεων. Συχνά η μία από τις δύο είναι η state-of-the-art και εξετάζεται η ομοιότητα της άλλης με αυτή.

Η πιο απλή ίδεα για την σύγκριση δύο διαφορετικών ομαδοποιήσεων βάσιζεται στον αριθμό των ζευγών των στοιχείων που ταξινομούνται με τον ίδιο τρόπο, δηλαδή που βρίσκονται είτε στην ίδια ομάδα είτε σε διαφορετική και στις δύο περιπτώσεις.

Πριν ορισθούν οι μετρικές θα περιγράφουν κάποια σύνολα ζευγών που χρησιμοποιούνται από αυτές. Έστω δύο ομαδοποιήσεις C και C' όριζονται τα ακόλουθα σύνολα ζευγών[52]:

- $n_{11} = \{$ Αριθμός των ζευγαριών που βρίσκονται στην ίδια ομάδα στο $C$ και στο $C'\}$
- $n_{00} = \{$Αριθμός των ζευγαριών που βρίσκονται σε διαφορετική ομάδα στο $C$ και στο $C'\}$
- $n_{10} = \{$Αριθμός των ζευγαριών που βρίσκονται στην ίδια ομάδα στο $C$ και σε διαφορετική στο $C'\}$
- $n_{01} = \{$Αριθμός των ζευγαριών που βρίσκονται στην ίδια ομάδα στο $C'$ και σε διαφορετική στο $C\}$

**Παρατήρηση.** *Το άθροισμα όλων των ζευγών είναι:* $n_{11} + n_{10} + n_{01} + n_{00} = \binom{n}{2} = \frac{n(n-1)}{2}$

Ο δείκτης του Rand[53] ορίζει την ομοιότητα μεταξύ δύο ομαδοποιήσεων $C$ και $C'$ ως τον λόγο των ζευγαρίων που έχουν ταξινομηθεί σωστά με τα συνολικά ζεύγη:

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)} \tag{3.10}$$

Οι Fowlkes and Mallows[54] πρότειναν έναν δείκτη για την σύγκριση ιεραρχικών ομαδοποιήσεων ο οποίος μπορεί να χρησιμοποιηθεί και στους υπόλοιπους αλγορίθμους:

$$FM(C, C') = \frac{n_{11}}{\sqrt{(n_{11} + n_{10})(n_{11} + n_{01})}} \tag{3.11}$$

Τέλος υπάρχει και ένα μέτρο που βασίζεται στην θεωρία πληροφοριών και πιο συγκεκριμένα στην κοινή πληροφορία (mutual information) $I(C, C')$ που έχουν δύο ομαδοποιήσεις:

$$I(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i, j) \log \frac{P(i, j)}{P(i)P(j)} \tag{3.12}$$

Η κοινή πληροφορία είναι ένα μέτρο που βασίζεται στην εντροπία και αποτελεί μία ένδειξη για την μείωση της αβεβαιότητας σχετικά με την ομάδα που ανήκει ένα στοιχείο όταν είναι γνωστή η ομάδα που ανήκει σε μια άλλη ομαδοποίηση του ίδιου συνόλου στοιχείων.

**Υπενθύμιση.** *Η εντροπία μίας ομαδοποίησης* $C = \{C_1, C_2, ..., C_k\}$ *ορίζεται ως:*

$$H(C) = -\sum_{k=1}^{n} P(k)\ logP(k) \tag{3.13}$$

*όπου* $P(k) = \frac{|C_i|}{n}$ *είναι η πιθανότητα ένα στοιχείο* $k$ *να ανήκει στην ομάδα* $C_i$*.*

Η Διασπορά της Πληροφορίας (Variation of Information)[55] ορίζει την ομοιότητα δύο ομαδοποιήσεων $C$ και $C'$ ως την απώλεια της πληροφορίας που υπάρχει μεταξύ τους:

$$\begin{aligned} VI(C, C') &= H(C) + H(C') - 2I(C, C') \\ &= [H(C) - I(C, C')] + [H(C') - I(C, C')] \end{aligned} \tag{3.14}$$

Δεν είναι ένα κανονικοποιημένο μέτρο ωστόσο για να γίνει αυτό αρκεί να διαιρεθεί με την μέγιστη διασπορά πληροφορίας που υπάρχει. Η οποία θα είναι $VI_{max} = \log n$ δηλαδή στην μία ομαδοποίηση το κάθε στοιχείο αποτελεί μια ομάδα και στην άλλη όλα τα στοιχεία ανήκουν σε μια ομάδα. Στην κανονικοποιημένη εκδόχη το 0 αντιστοιχεί σε δύο ίδιες ομαδοποιήσεις και το 1 στην περίπτωση που περιγράφηκε προηγουμένως για την μέγιστη διασπορά πληροφορίας.

Στην συνέχεια παρατίθεται ο ψευδοκώδικας του αλγορίθμου:

**Algorithm 7:** Variation of Information

**input :** X state of the art clustering, Y alternative clustering (List of Lists)
**output:** Similarity score of X and Y based on Variation of Information

```
1  n = total elements in cluster X
2  VI = 0
3  for cluster x in X :
4     p = len(x)/n    //probability of an element beloging to cluster x in X
5     for cluster y in Y :
6        q = len(y)/n    //probability of an element beloging to cluster y in Y
7        r = len(set(x)&set(y))/n    //probability of an element beloging to both cluster x
       and cluster y
8           if r > 0:
9              VI = −r(log(r/p) + log(r/q))
10    end
11 end
12 VI_normalized = VI/ log n
```

# Κεφάλαιο 4

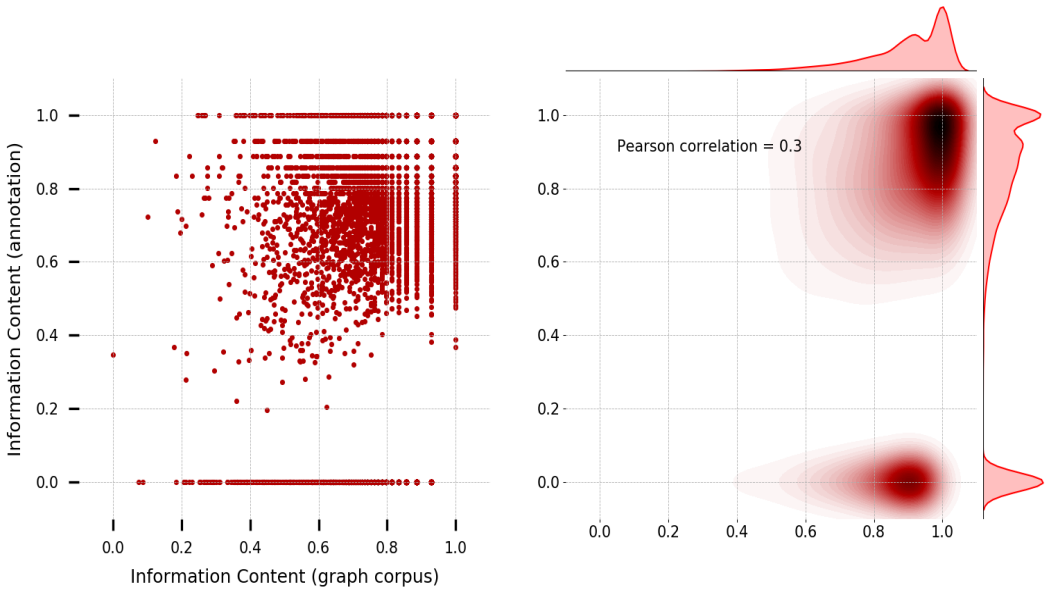
# Χαρακτηριστηκά της Γονιδιακής Οντολογίας και Τοπολογική Αποτίμηση των Μετρικών

Στο κεφάλαιο αυτό παρουσιάζεται η αναγκαίοτητα του True Path Rule στον χαρακτηρισμό των οντολογίων με εξωτερικά σύνολα, όπως είναι το σύνολο των γονιδίων. Στη συνέχεια περιγράφεται το shallow annotation problem και γίνεται μια διαισθητική αποτίμηση των μέτρων σημασιολογικής ομοιότητας σε ένα στιγμιότυπο της Γονιδιακής Οντολογίας.

## 4.1 Συσχέτιση Graph Corpus και Annotation

Στο πρώτο κεφάλαιο παρουσιάστηκε η Γονιδιακή Οντολογία GO και τα κύρια μέτρα σημασιολογικής ομοιότητας. Τα περίσσοτερα μέτρα χρησιμοποιούν την πληροφορία που φέρει ο κόμβος (Information Content - IC) η οποία μπορεί να υπολογιστεί είτε με βάση τους απογόνους που έχει ο όρος πάνω στην τοπολογία της οντολογίας (graph corpus) είτε με βάση τον αριθμό των γονιδίων που του έχουν αντιστοιχηθεί (mapping). Στην δεύτερη περίπτωση κάθε γονίδιο αντιστοιχίζεται σε συγκεκριμένους όρους (terms), συνήθως και στις 3 κατηγορίες της GO. Όπως έχει όμως αναφερθεί, θα πρέπει να υιοθετείται ο κανόνας True Path Rule στον χαρακτηρισμό των όρων, καθώς κάθε πρόγονος περιέχει την πληροφορία που φέρουν οι απογονοί του. Συνεπώς, για να υπάρχει αναλογία μεταξύ της τοπολογίας και της αντιστοίχησης με γονίδια, πρέπει ο χαρακτηρισμός των όρων να γίνεται με βάση τον True Path Rule.

Για να αποδειχθεί η αναγκαίοτητα αυτού του κανόνα θα υποτεθεί οτί δεν ισχύει και θα υπολογισθεί η συσχέτιση που υπάρχει μεταξύ των μετρήσεων του IC με βάση το graph corpus και το mapping για όλους τους όρους της GO. Αν ο κανόνας είναι αναγκαίος για την σωστή λειτουργία της οντολογίας θα πρέπει εν τη απουσία του οι δύο μετρήσεις να είναι ασυσχέτιστες. Τα αποτελέσματα της απουσίας του True Path Rule φαίνεται στην εικόνα 4.1.

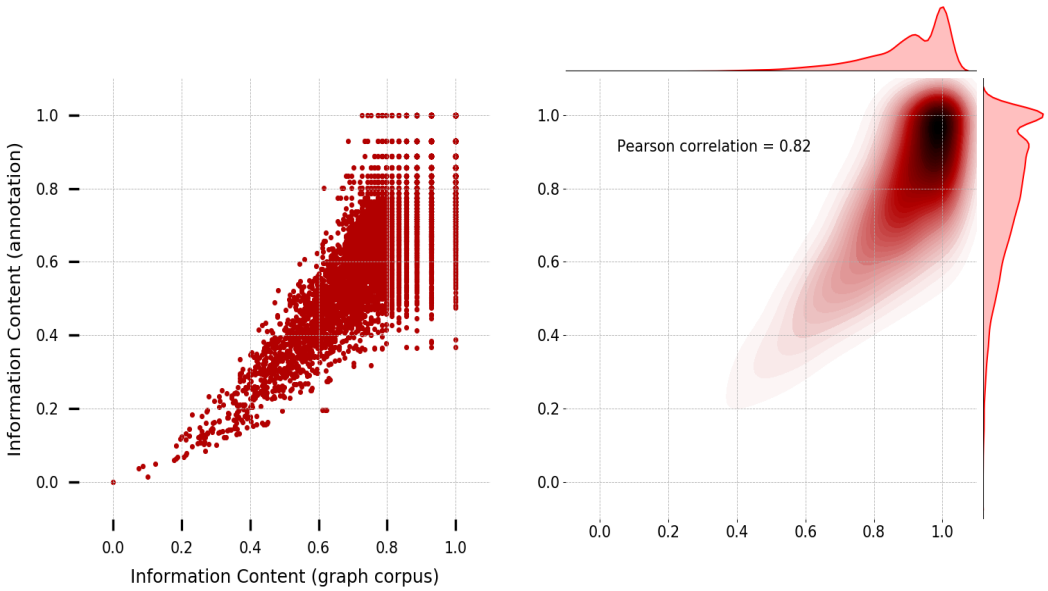



Σχήμα 4.1: Αριστερά είναι το διάγραμμα σκέδασης του IC με βάση την τοπολογία και με βάση το annotation χωρίς τον True Path Rule. Δεξιά οι κατανομές και το διάγραμμα πυκνότητας των δύο διαφορετικών μετρήσεων για το IC.

Πράγματι η συσχέτιση που υπάρχει μεταξύ των δύο διαφορετικών μετρήσεων είναι 0.32 το οποίο υποδηλώνει ότι δεν υπάρχει κάποια ομοιότητα μεταξύ τους. Επιπλεόν το διάγραμμα πυκνότητας (density plot) των δύο μετρήσεων το οποίο φαίνεται στη δεξιά εικόνα του σχήματος 4.1 αποτελείται από δύο διαφορετικές κατανομές το οποίο ενισχύει το συμπέρασμα ότι οι μεταβλητές είναι ασυσχέτιστες.

Αντίθετα όπως φαίνεται στην είκονα 4.2, όταν ισχύει ο True Path Rule ο συντελεστής συσχέτισης για τις δύο μετρήσεις είναι 0.82 δηλαδή υπάρχει έντονη γραμμική συσχέτιση μεταξύ τους. Επιπλέον το density plot τους αποτελείται από μια κατανομή το οποίο ενισχύει την υπόθεση της συσχέτισης των μετρήσεων. Συνεπώς οι δυο διαφορετικοί τρόποι για τον υπολογισμό του Information Content έχουν πολύ κοντινά αποτελέσματα.

Από τα παραπάνω συμπεραίνουμαι την αναγκαιότητα του εν λόγω κανόνα για τον σωστό υπολογισμό του IC με βάση τον χαρακτηρισμό στην GO.



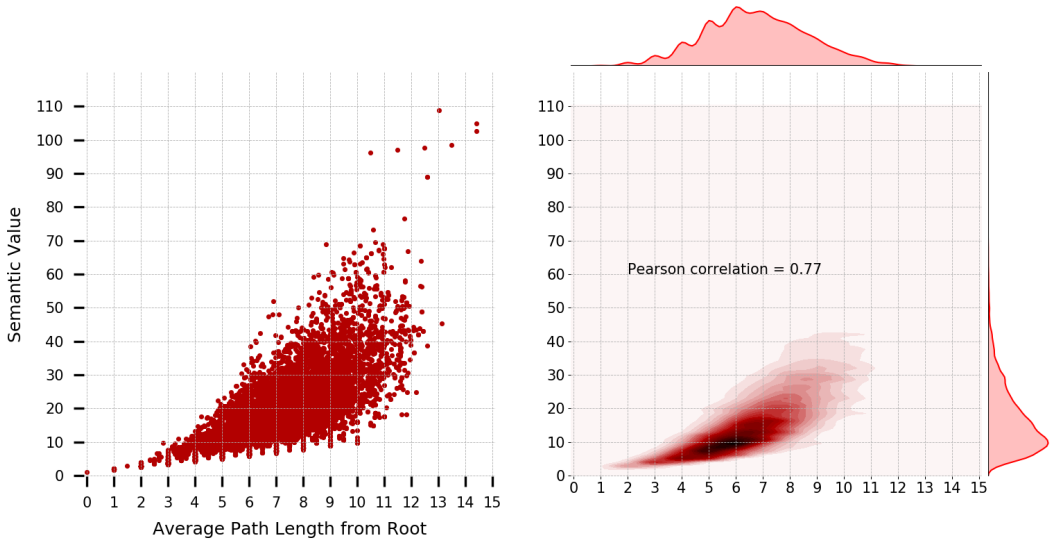

Σχήμα 4.2: Αριστερά είναι το διάγραμμα σκέδασης του IC με βάση την τοπολογία και με βάση το annotation με τον True Path Rule. Δεξιά οι κατανομές και το διάγραμμα πυκνότητας των δύο διαφορετικών μετρήσεων του IC.

## 4.2 Χαρακτηριστικά Μεγέθη των Μετρικών

Οι μετρικές χωρίζονται σε 3 βασικές κατηγορίες στα μέτρα ακμών, στα μέτρα κόμβων και στα υβριδικά μέτρα. Αύτες που χρησιμοποιούνται είναι οι δύο τελευταίες κατηγορίες δηλαδή μέτρα που είτε έμμεσα όπως το Aggregate Information Content (AIC) είτε άμεσα όπως οι υπόλοπες μετρικές πλην του Dice και του Jaccard που βασίζονται στην τομή/ένωση συνόλων χρησιμοποιούν το Information Content των όρων.

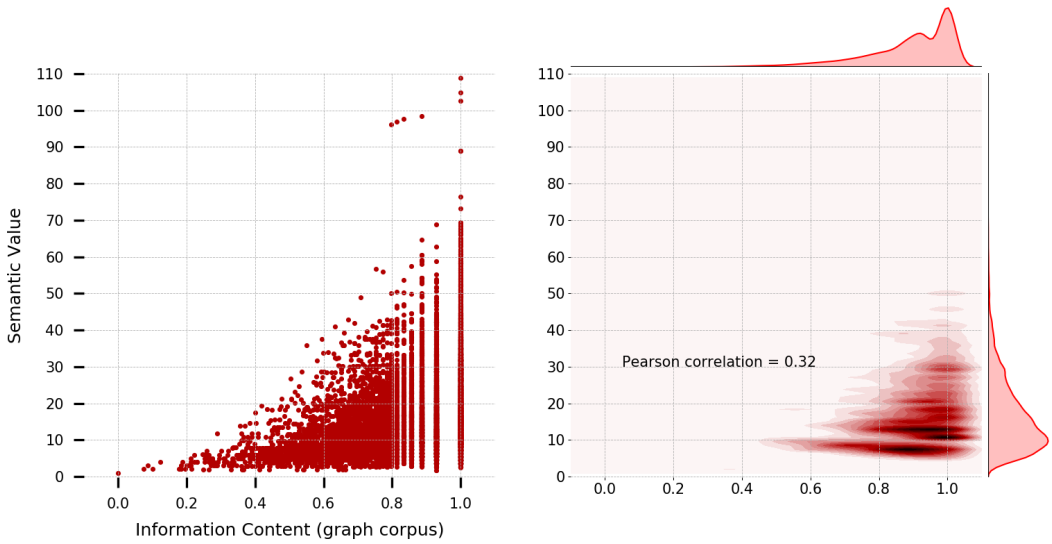
Ο AIC είναι ένα υβριδικό μέτρο καθώς λαμβάνει υπόψιν του τόσο την πληροφορία που φέρουν οι κόμβοι όσο και την θέση τους στην οντολογία. Η διαφορά του από τις υπόλοιπες μετρικές είναι ότι χρησιμοποίει το IC για να ορίσει το knowledge ενός όρου και στην συνέχεια αθροίζει το κανονικοποιήμενο knowledge (semantic weight) όλων των προγόνων του για να υπολογίσει το semantic value του. Εξ ορισμού όροι οι οποίοι έχουν πολύ μικρό IC έχουν μεγάλο semantic weight και επομένως οι όροι που βρίσκονται χαμηλότερα στην οντολογία έχουν το μεγαλύτερο semantic value.

Από τα παραπάνω προκύπτει το συμπέρασμα ότι το semantic value ενός όρου συνδέεται με την θέση του όρου στην Οντολογία. Δηλαδή περιμένουμε να σχετίζεται με την απόσταση που έχει από την ρίζα της οντολογίας και όχι με το Information Content του. Ο παραπάνω ισχυρισμός επιβαβαιώνεται στο σχήμα 4.3 όπου παρουσιάζεται το διάγραμμα σκέδασης (scatterplot) μεταξύ της απόστασης του όρου από την ρίζα της οντολογίας και του semantic value.




Σχήμα 4.3: Αριστερά είναι το διάγραμμα σκέδασης του του semantic weight με το average path από την ρίζα. Δεξιά το διάγραμμα πυκνότητας των δύο μεταβλητών.

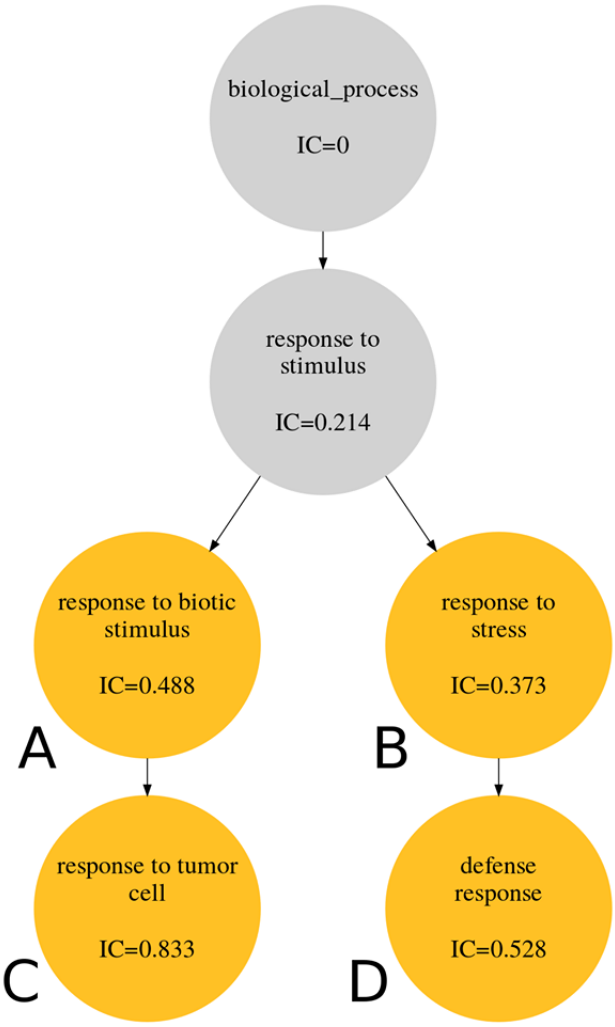
Ο συντελεστής συσχέτισης μεταξύ του semantic value και της απόστασης του όρου από την ρίζα είναι 0.77 το οποίο υποδηλώνει ότι είναι έντονα γραμμικά συσχετισμένα. Αντίθετα ο συντελέστης συσχέτισης μεταξύ του semantic value και του IC είναι 0.32 όπως φαίνεται στο σχήμα 4.4 το οποίο επιβεβαιώνει ότι τα δύο αυτά μεγέθη είναι ασυσχέτιστα.




Σχήμα 4.4: Αριστερά είναι το διάγραμμα σκέδασης του του semantic weight με το IC. Δεξιά το διάγραμμα πυκνότητας των δύο μεταβλητών.

## 4.3 Shallow Annotation Problem

Ένα πρόβλημα που έχουν όλες οι μετρικές εκτός του Resnik είναι ότι όταν δύο ζευγάρια όρων έχουν τον ίδιο κοινό πρόγονο (MICA) υπολογίζουν ως μεγαλύτερη την ομοιότητα του ζευγαριού που είναι πιο κοντά στον MICA. Το ζευγάρι των όρων που είναι ψηλότερα στην οντολογία και συνεπώς αποτελείται από πιο γενικούς όρους ορίζεται ως το πιο όμοιο. Το πρόβλημα αυτό στην βιβλιογραφία αναφέρεται ως shallow annotation problem [17].




Σχήμα 4.5: Στιγμιότυπο της Γονιδιακής Οντολογίας.

Για παράδειγμα στο σχήμα 4.5 όλες οι μετρικές όπως φαίνεται στον πίνακα 4.1 αντιστοιχούν μεγαλύτερη ομοιότητα στο ζεύγος όρων A,B από το ζεύγος C,D. Το αποτέλεσμα αυτό δεν ταιριάζει με την ανθρώπινη αντίληψη για την ομοίοτητα των εν λόγω όρων, που θεωρεί ότι οι όροι A,B είναι πολύ γενικοί καθώς βρίσκονται ψηλά στον γράφο της οντολογία και συνεπώς η ομοιότητα τους πρέπει να είναι μικρότερη ή ίση από την ομοιότητα των όρων C,D.

| **Μετρική** | C vs. D | A vs. B |
|---|---|---|
| Resnik | 0.214 | 0.214 |
| Lin | 0.314 | 0.497 |
| Jiang and Conrtath | 0.533 | 0.783 |
| Schlicker | 0.247 | 0.391 |
| Pirro and Euzenat | 0.187 | 0.331 |
| Jaccard | 0.5 | 1 |
| Dice | 0.667 | 1 |
| GIC | 0.199 | 1 |
| Mazandu | 0.332 | 1 |
| Nunivers | 0.257 | 0.439 |
| AIC | 0.535 | 0.686 |

Πίνακας 4.1: Οι σημασιολογικές ομοιότητες υπολογισμένες με όλες τις μετρικές για τα ζεύγη A,B και C,D του γράφου του σχήματος 4.5.

Από τον πίνακα 4.1 προκύπτει άλλο ένα ενδιαφέρον συμπέρασμα. Οι μετρικές που βασίζονται στον λόγο των κοινών προγόνων προς τους συνολικούς προγόνους, είτε υπολογίζοντας μόνο τον αριθμό τους (Jaccard, Dice) είτε σταθμιζοντάς τον με το IC τους (Mazandu, Graph IC) υπολογίζουν την ομοιότητα δυο όρων που έχουν ακριβώς τους ίδιους προγόνους ίση με τη μονάδα, όπως για παράδειγμα ο A με τον B. Συνεπώς δύο όροι πολύ γενικοί ή δύο όροι που είναι απόγονοι πρώτου βαθμού της ρίζας να θεωρούνται λειτουργικά ταυτόσημοι.

## 4.4 Κατάταξη των Μετρικών

Σύμφωνα με την δομή της Γονιδιακής Οντολογίας και την ανθρώπινη αντίληψη για αυτή οι παράγοντες που καθορίζουν την ομοιότητα δύο όρων κατά σειρά σημαντικότητας είναι:

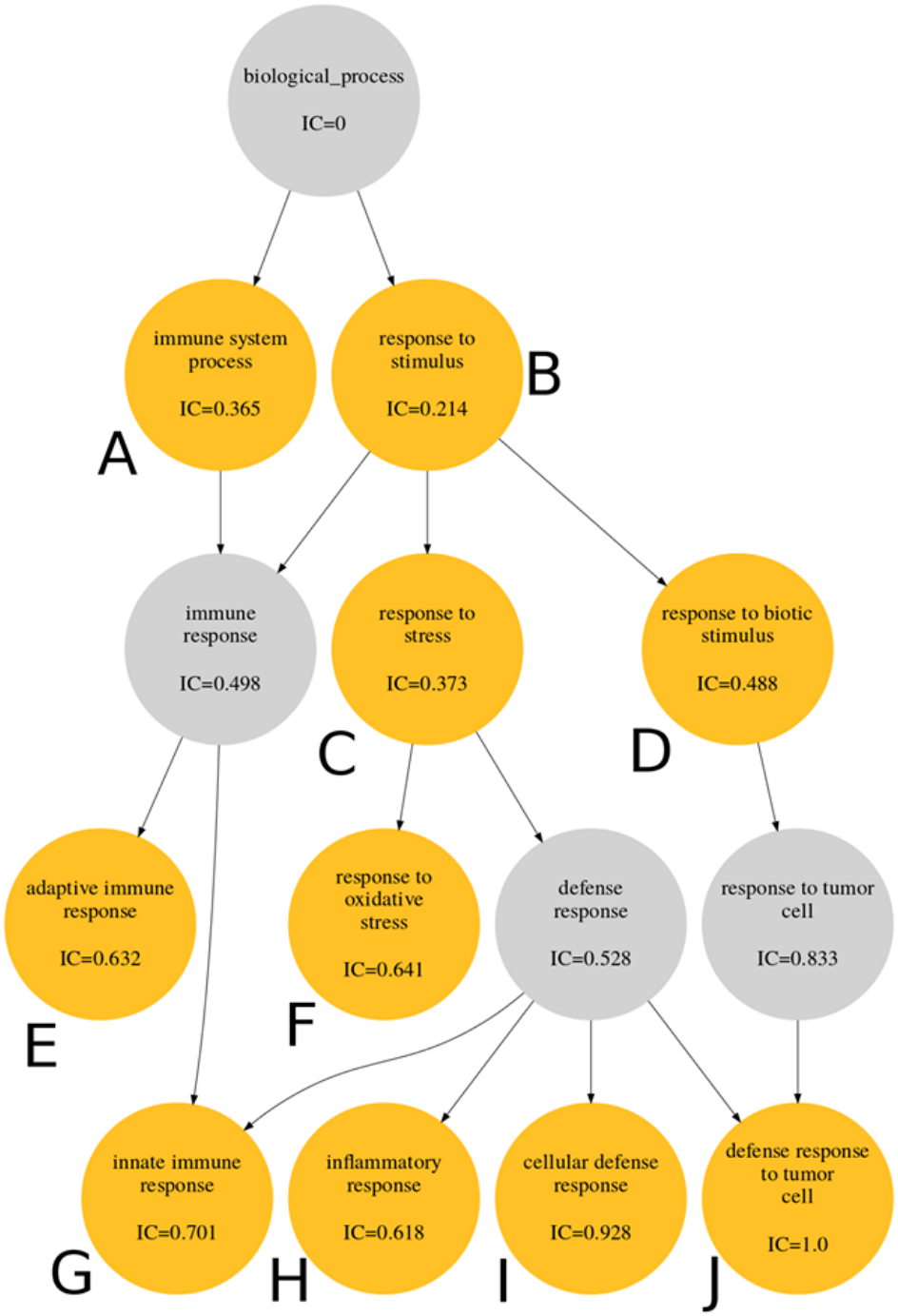
1. Το information content του πιο κοντίνου κοινού τους πρόγονου (MICA). Αν δύο όροι έχουν ως MICA έναν πολύ εξειδικευμένο όρο πρέπει να έχουν μεγάλη ομοιότητα. Αντίθετα αν έχουν ένα πολύ γενικό όρο θα πρέπει να έχουν μικρή.

2. Οι διακριτοί πρόγονοι (multiple parent inheritence) των δύο όρων. Αν δύο όροι έχουν τον ίδιο MICA και επιπλέον τους ίδιους διακριτούς κοινούς πρόγονους θα πρέπει να έχουν μεγαλύτερη ομοιότητα από ότι θα είχαν στην περίπτωση που ο μόνος κοινός τους πρόγονος είναι ο MICA. Αντίθετα δύο όροι που έχουν διαφορετικούς διακριτούς προγόνους θα πρέπει να τιμωρούνται και να έχουν μειωμένη ομοιότητα.

3. Το IC των ίδιων των όρων δηλαδή η θέση τους στην Οντολογία. Δύο όροι που βρίσκονται κοντά στη ρίζα της Οντολογίας πρέπει να έχουν μικρή ομοιότητα για να μην δημιουργείται το shallow annotation problem που περιγράφηκε προηγουμένως.

4. Οι κοινοί απόγονοι των δύο όρων. Αν δύο όροι έχουν πολλούς κοινούς απογόνους θα πρέπει να επιβραβεύονται και να έχουν λίγο μεγαλύτερη σημασιολογική ομοιότητα από ότι στην περιπτώση που δεν έχουν.

Στην συνέχεια θα εφάρμοσουμε τα εν λόγω κριτήρια για την κατάταξη σε φθίνουσα σειρά των όρων του σχήματος 4.6 με βάση την σημασιολογική τους ομοιότητα.




Σχήμα 4.6: Στιγμυότυπο της Γονιδιακής Οντολογίας για την διαισθητική κατάταξη των όρων.

Σύμφωνα με τα κριτήρια αξιολόγισης της ομοιότητας δυο οντολογικών όρων, σημαντικότερη παράμετρος είναι η εξειδίκευση του MICA καθώς όσο πιο εξειδικευμένος είναι τόσο μεγαλύτερη είναι η σημασιολογική ομοιότητα των εξεταζόμενων όρων. Ο πιο εξειδικευμένος κοινός πρόγονος του γράφου 4.6 είναι ο "defense response" τον οποίο έχουν ως MICA όλα τα ζεύγη που σχηματίζονται από τους όρους G, H, I και J. Συνεπώς, τα ζεύγη αυτά έχουν μεγαλύτερη σημασιολογική ομοιότητα από οποιοδήποτε άλλο ζεύγος όρων στο γράφο. Την μεγαλύτερο σημασιολογική ομοιότητα έχει το ζεύγος HI καθώς κάνενας από τους δύο όρους δεν έχει πολλαπλή κληρονομικότητα και όλο το σημασιολογικό τους περιεχόμενο προέρχεται αποκλειστικά από τον "defense response". Αντίθετα στα υπόλοιπα ζευγάρια υπάρχει πολλαπλή κληρονομικότητα σε τουλάχιστον έναν όρο. Συγκεκριμένα, ο όρος J έχει επιπλέον διακριτό πρόγονο τον "response to tumor cell" ο οποίος του προσδίσει ένα επιπλέον σημασιολογικό περιεχόμενο μειώνοντας έτσι την ομοιότητα των GJ, HJ, IJ σε σύγκριση με το HI. Η ομοιότητα του IJ είναι μεγαλύτερη από αυτή του HJ, γιατί αποτελείται από πιο εξειδικευμένους όρους (3ο κριτήριο αξιολόγισης). Τέλος η ομοιότητα του GJ μειώνεται κι άλλο σε σχέση με τα δυο προηγούμενα ζευγάρια, καθώς έχει ως διακρικό μη κοινό πρόγονο τον όρο ("immune response"). Τελικά τα τέσσερα παραπάνω ζεύγη ταξινομούνται ως εξής:1. HI, 2. IJ, 3. HJ, 4. GJ.

Μελετώντας τα ζεύγη GJ και GE, παρατηρείται έντονα το φαινόμενο της πολλαπλής κληρονομικότητας με παραπάνω από έναν κοινούς ή μη κοινούς διακριτούς προγόνους. Όπως προαναφέρθηκε, το GJ έχει ως κοινούς διακριτούς τους όρους "defense response" και "response to stimulus", ενώ ο G έχει επιπλέον ως διακριτό πρόγονο τον "immune response". Αντίθετα, το ζεύγος GE έχει μόνο έναν κοινό πρόγονο, τον "immune response" ενώ ο G κληρονομεί και την έννοια του "δεφενσε ρεσπονσε". Επιπλέον ο MICA του GJ ("defense response") είναι πιο εξειδικευμένος απ αυτόν του GE ("immune response"). Με γνώμονα όλα τα παραπάνω πρέπει να αποδωθει μεγαλύτερη σημασιολογική ομοιότητα στο GJ σε σχέση με το GE. Με την προσέγγιση αυτή, η ιεραρχιμένη λίστα των ζευγαριών επεκτείνεται ως εξής: 1. HI, 2. IJ, 3. HJ, 4.GJ, 5.GE.

Συνεχίζοντας στους πιο γενικούς όρους που βρίσκονται πιο ψηλά στο γράφο, συναντόνται τα ζεύγη CF, CD, DE και CE. Το ζευγάρι CF έχει την μεγαλύτερη σημασιολογική ομοιότητα καθώς ο MICA του είναι ο ίδιος ο όρος C ο οποίος έχει μεγαλύτερο information content απ τον όρο "response to stimulus", ο οποίος είναι ο MICA των υπόλοιπων τριων ζευγών. Το ζευγος CD έχει μοναδικό διακριτό πρόγονο τον "response to stimulus", ενώ η ομοιότητά του πρέπει να αυξηθεί λόγω της ύπαρξης κοινών απογόνων (4ο κριτήριο αξιολόγισης). Στα ζεύγη DE και CE παρατηρείται η πολλαπλή κληρονομικότητα του όρου E, γεγονός που μειώνει την ομοιότητά τους σε σχέση με τα δυο παραπάνω ζεύγη. Τέλος, επειδή ο όρος D είναι πιο ειδικός απ τον C συμπερένεται πως η ομοιότητα του DE θα πρέπει να είναι μεγαλυτερη απ αυτή του CE.

Φτάνοντας στην ρίζα του γράφου, το ζευγάρι AB μπορεί να θεωρηθεί απομακρυσμένο σημασιολογικά, καθώς ο μόνος κοινός πρόγονος είναι η ρίζα, η οποία έχει μηδενικό information content. Όμως η ύπαρξη κοινών απογόνων σημαίνει πως η τομή των υποκατηγοριών τους δεν είναι μηδενική. Γι αυτό τον λόγο, το AB μπορεί να θεωρηθεί ποιο όμοιο ζευγάρι απ το AD, το οποίο έχει τον ίδιο MICA, αλλά χωρίς την ύπαρξη κοινών απογόνων.

Συνοψίζοντας, η τελική λίστα κατάταξης των ζευγαριών έχει ως εξής: 1. HI, 2. IJ, 3. HJ, 4. GJ, 5. GE, 6. CF, 7. CD, 8. DE, 9. CE, 10. AB, 11. AD.

Στην συνέχεια υπολογίσθηκε για κάθε μετρική η κατάταξη των εν λόγω όρων. Να σημειωθεί ότι οι μετρικές που υποστηρίζουν διαφορετικές στρατηγικές επιλογής προγόνων (MICA, DCA, Dishin, xGrasm) υπολογίσθηκαν με όλους τους τρόπους.

Για τον έλεγχο των κατατάξεων υπάρχουν δύο τρόποι:

1. Η απόσταση Manhattan $d(x, y) = \|x - y\|_1 = \sum_i |x_i - y_i|$ η οποία αναπαριστά τα διαφορετικά rankings ως διανύσματα στο ν-διάστατο χώρο και υπολογίζει την απόσταση τους. Πρόκειται για μια αυστηρή προσέγγιση που τιμωρεί τα σημαντικά λάθη π.χ. το πρώτο στοίχειο να ιεραρχηθεί ως τελευταίο.

2. Το στατιστικό ελέγχου του Kendall Tau το οποίο ελέγχει αν ενα ζεύγος όρων x,y είναι σώστα τοποθετημένο σε σχέση με τα υπόλοιπα και αν δεν είναι υπολογίζει το μέγεθος του σφάλματος. Επιπλέον υπολογίζει τον συντελεστή συσχέτισης που έχουν τα δύο ranking και ελέγχει αν προέρχονται από την ίδια κατανομή με μηδενική υπόθεση ότι δεν προέρχονται.

Στην GO η ποιο συνηθισμένη στρατήγικη επιλογής προγόνων είναι ο MICA συνεπώς είναι πιθανό να υπάρχουν ισοπαλίες στα αποτελέσματα. Για το λόγο αυτό η παραλλαγή του δείκτη του Kendall Tau που λαμβάνει υπόψιν τις πιθανές ισοπαλίες ταιριάζει καλύτερα στην υπό μελέτη περίπτωση. Ωστόσο υπολογίσθηκε και η απόσταση Manhattan η οποία ευνοεί τις υπόλοιπες στρατηγικές επιλογής προγόνων.

Τα αποτελέσματα των δύο ελέγχων φαίνονται στον πίνακα 4.2. Με βάση το στατιστικό ελέγχου του Kendall και την απόσταση Manhattan την καλύτερη συμπεριφορά έχει ο Aggregate IC και την χειρότερη ο Dice,ο Jaccard και ο Jiang and Conrath με την Dishin στρατηγική. Επιπλέον και οι δύο μέθοδοι θεωρούν ως πολύ καλή την μετρική του Schlicker με την xGrasm στρατηγική και την μετρική του Resnik με την Dishin. Ωστόσο το στατιστικό ελέγχου θεωρεί πολύ καλό το Resnik για όλες τις παραλλαγές του κάτι το οποίο δεν συμβαίνει με την απόσταση Manhattan. Τέλος όπως ήταν αναμενόμενο ο Kendall ευνοεί την στρατηγική του MICA ενώ η απόσταση Manhattan τις υπόλοιπες στρατηγικές επιλογής προγόνων που έχουν λίγες ισοπαλίες.

Ένα γενικό συμπέρεσμα που προκύπτει και από τις δύο μεθόδους είναι ότι την καλύτερη συμπεριφορά την έχει ο Aggregate IC ο οποίος λαμβάνει υπόψιν του την τοπολογία της οντολογίας και ο Resnik ο οποίος επιστρέφει τις μεγαλύτερες τιμές σημασιολογικής ομοιότητας καθώς λαμβάνει υπόψιν μόνο το MICA των εξεταζόμενων όρων.

| Μετρικές | Manhattan Distance | Kendall Tau Correlation | Kendall Tau p-value | Ranking |
|---|---|---|---|---|
| Aggregate IC | 6 | 0.891 | 0.00014 | [1, 2, 4, 3, 6, 7, 5, 8, 9, 10, 11] |
| Sclicker xGrasm | 11 | 0.807 | 0.00059 | [3, 4, 1, 2, 5, 6, 7, 9, 8, 10, 10] |
| Resnik Dishin | 12 | 0.759 | 0.00132 | [1, 3, 6, 4, 5, 2, 7, 8, 8, 9, 9] |
| Pirro and Euzanat DCA | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Lin xGrasm | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Lin DCA | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Sclicker DCA | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Pirro and Euzanat xGrasm | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Pirro and Euzanat MICA | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Lin MICA | 17 | 0.697 | 0.00300 | [4, 5, 1, 2, 6, 3, 7, 9, 8, 10, 10] |
| Mazandu | 20 | 0.673 | 0.00468 | [1, 4, 2, 3, 6, 5, 1, 7, 7, 8, 8] |
| Nunivers xGrasm | 20 | 0.748 | 0.00167 | [3, 5, 1, 2, 5, 4, 6, 7, 7, 8, 8] |
| Graph IC | 20 | 0.673 | 0.00468 | [1, 4, 2, 3, 6, 5, 1, 7, 7, 8, 8] |
| Schlicker Dishin | 21 | 0.587 | 0.01246 | [2, 5, 6, 3, 7, 1, 4, 9, 8, 10, 10] |
| Jiang and Conrath xGrasm | 21 | 0.624 | 0.00793 | [4, 6, 1, 2, 7, 3, 5, 9, 8, 10, 10] |
| Pirro and Euzanat Dishin | 21 | 0.587 | 0.01246 | [2, 5, 6, 3, 7, 1, 4, 9, 8, 10, 10] |
| Lin Dishin | 21 | 0.587 | 0.01246 | [2, 5, 6, 3, 7, 1, 4, 9, 8, 10, 10] |
| Nunivers DCA | 22 | 0.711 | 0.00283 | [4, 5, 1, 2, 5, 3, 6, 7, 7, 8, 8] |
| Nunivers Dishin | 22 | 0.611 | 0.00974 | [2, 5, 6, 3, 7, 1, 4, 8, 8, 9, 9] |
| Nunivers MICA | 22 | 0.711 | 0.00283 | [4, 5, 1, 2, 5, 3, 6, 7, 7, 8, 8] |
| Jiang and Conrath MICA | 28 | 0.455 | 0.05163 | [5, 6, 1, 3, 9, 2, 4, 10, 7, 8, 11] |
| Jiang and Conrath DCA | 28 | 0.455 | 0.05163 | [5, 6, 1, 3, 9, 2, 4, 10, 7, 8, 11] |
| Sclicker MICA | 29 | 0.294 | 0.21152 | [6, 7, 3, 4, 9, 1, 2, 8, 5, 10, 10] |
| Resnik MICA | 35 | 0.864 | 0.00048 | [1, 1, 1, 2, 1, 3, 4, 4, 4, 5, 5] |
| Resnik xGrasm | 35 | 0.864 | 0.00048 | [1, 1, 1, 2, 1, 3, 4, 4, 4, 5, 5] |
| Resnik DCA | 35 | 0.864 | 0.00048 | [1, 1, 1, 2, 1, 3, 4, 4, 4, 5, 5] |
| Jiang and Conrath Dishin | 38 | 0.127 | 0.58579 | [3, 8, 7, 5, 10, 1, 2, 9, 6, 4, 11] |
| Jaccard Coefficient | 42 | 0.35 | 0.16371 | [1, 2, 2, 2, 3, 2, 1, 3, 3, 1, 4] |
| Dice Coefficient | 42 | 0.35 | 0.16371 | [1, 2, 2, 2, 3, 2, 1, 3, 3, 1, 4] |

Πίνακας 4.2: Η κατάταξη των μετρικών και με τις δύο στρατηγικές.

# Κεφάλαιο 5

# Ανάλυση Δικτύων από την InnateDB

Στο κεφάλαιο αυτό θα αναλυθούν πρωτεϊνικά δίκτυα από την βάση δεδομένων InnateDB. Θα εξετασθεί η τοπολογία τους και όσα χαρακτηρίζονται ως ελεύθερης κλίμακας θα χρησιμοποιηθούν για να υπολογισθεί η συσχέτιση που υπάρχει μεταξύ της σημασιολογικής ομοιότητας των όρων των δικτύων με βάση το graph corpus και το annotation.

## 5.1 InnateDB

Η InnateDB[56] είναι μία ελεύθερα προσβάσιμη βάση δεδομένων γονιδίων και πρωτεϊνών η οποία αποτελείται από πειραματικά επιβεβαιωμένα (experimentally-verified) δίκτυα πρωτεϊνικών αλληλεπιδράσεων. Πιο συγκεκριμένα περιέχει σηματοδοτικά μονοπάτια (signaling pathways) που σχετίζονται με την έμφυτη ανοσολογική απόκριση (innate immune response) των ανθρώπων, των ποντικών και των βοοειδών σε μικροβιακή λοίμωξη.

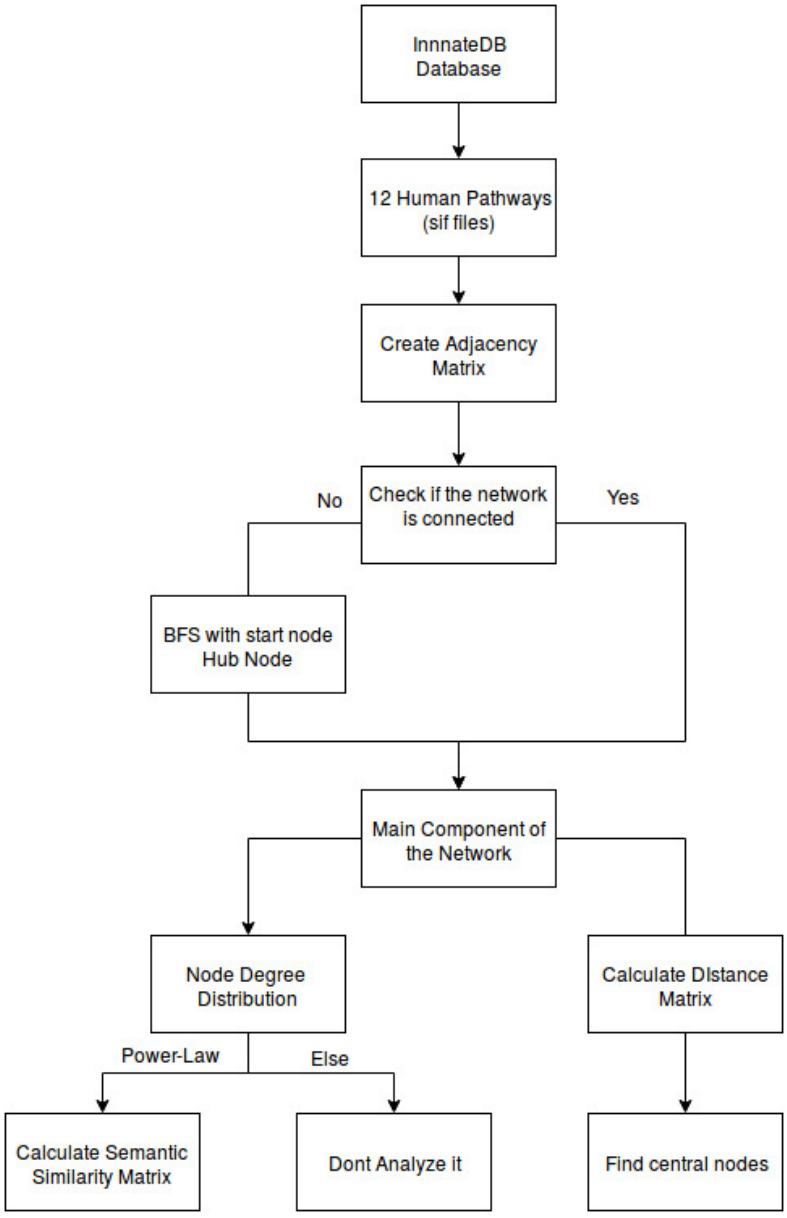
| **Όνομα Πρωτεϊνικού Δικτύου** | Αριθμός Κόμβων | Αριθμός Κόμβων Giant Component | Αριθμός Ακμών Giant Component |
|---|---|---|---|
| Chemokine Signaling Pathway | 162 | 159 | 1202 |
| Complement Cascade | 64 | 54 | 170 |
| Cytosolic DNA-Sensing Pathway | 51 | 49 | 314 |
| Jak-STAT Signaling Pathway | 130 | 128 | 888 |
| MAPK Signaling Pathway | 225 | 222 | 2106 |
| mTOR Signaling Pathway | 57 | 53 | 304 |
| Natural Killer Cell Mediated Cytotoxicity | 100 | 90 | 560 |
| NOD-like Receptor | 56 | 55 | 418 |
| Regulation of Autophagy | 24 | 21 | 106 |
| RIG-I-like receptor signaling pathway | 58 | 56 | 520 |
| Toll-like receptor signaling pathway | 94 | 88 | 798 |

Πίνακας 5.1: Αριθμός Κόμβων και Ακμών στα 11 PPI από την InnateDB.

Από την βάση δεδομένων της InnateDB επιλέχθηκαν για να αναλυθούν τα 11 διαθέσιμα σηματοδοτικά μονοπάτια για τον άνθρωπο των οποίων τα βασικά χαρακτηριστικά (αριθμός κόμβων και ακμών) φαίνονται στον πίνακα 5.1.

## 5.2 Ανάλυση Δικτύων

Τα δίκτυα είναι αποθηκευμένα στην βάση δεδομένων σε μορφή Simple Interaction Format (sif) δηλαδή ως αρχεία κειμένου που αποτελούνται από τρεις στήλες όπου η πρώτη και η τρίτη είναι οι πρωτεΐνες και η δεύτερη το είδος της αλληλεπίδρασης που έχουν. Για να μπορέσει να ελεχθεί η τοπολογία των δικτύων έπρεπε να σχεδιασθεί μια μεθολογία η οποία θα δέχεται ως είσοδο το sif αρχείο η οποία θα ελέγχει αν το δίκτυο είναι συνδεδεμένο και αν δεν είναι θα βρίσκει την μεγαλύτερη συνιστώσα του. Η μεθοδολογία αυτή υλοποιήθηκε στην γλώσσα προγραμματισμού Python και το διάγραμμη ροής της φαίνεται στο σχήμα 5.1.




Σχήμα 5.1: Workflow από το sif αρχείο στην ανάλυση του δικτύου.

Στην συνέχεια παρουσιάζονται οι κατανομές των βαθμών των 11 σηματοδοτικών μονοπατιών.


Σχήμα 5.2: Οι κατανομές των βαθμών των 11 PPI. Με κόκκινο έχουν σημανθεί τα δίκτυα που ακολουθούν τοπολογία ελεύθερης κλίμακας.

Η κατανομή των βαθμών στα μικρότερα δίκτυα δεν ακολουθεί νόμο δύναμης επομένως δεν χαρακτηρίζονται ως ελεύθερης κλίμακας. Αντίθετα τα μεγαλύτερα δίκτυα ικανοποιούν την τοπολογία της ελεύθερης κλίμακας και συνεπώς η ανάλυση θα περιορισθεί σε αυτά:

1. Το MAPK (Mitogen-activated protein kinase) signaling pathway [57] το οποίο είναι ένα σηματοδοτικό μονοπάτι με βάση την πρωτεϊνική κινάση (MAPK) και εμπλέκεται σε διάφορες κυτταρικές λειτουργίες όπως ο πολλαπλασιασμός, η διαφοροποίηση (differentiation) και η μετανάστευση (migration).

2. Το JAK-STAT (Janus kinases - Signal Transducer and Activator of Transcription proteins) signaling pathway [58] είναι ένα σηματοδοτικό μονοπάτι το οποίο μεταδιδεί πληροφορία από τον εξωκκυττάριο χώρο στον κυτταρικό πυρήνα, με αποτέλεσμα την ενεργοποίηση των γονιδίων μέσω της μεταγραφής. Η ενεργοποίηση του διεγείρει (stimulates) διάφορες κυτταρικές λειτουργίες όπως ο πολλαπλασιασμός, η διαφοροποίηση (differentiation), η μετανάστευση (migration) και η απόπτωση (apoptosis).

3. Το Chemokine signaling pathway το οποίο ενεργοποιεί το Jak-STAT signaling pathway.

4. Το Natural Killer cell mediated cytoxicity το οποίο έχει κομβικό ρόλο στην ανοσολογική απόκριση καθώς συμμετέχει στην στοχευμένη νέκρωση ενός κυτάρρου στόχου, είτε μέσω της απελευθέρωσης συσσωμάτων που περιέχουν κυτταροτοξικά μόρια είτε μεσω ενεργοποίησης υποδοχέων.

Όπως έχει ήδη αναφερθεί, η σημασιολογική ομοιότητα δύο όρων μπορεί να υπολογισθεί είτε με βάση τον αριθμό των απογόνων που έχει ένας όρος (graph corpus) είτε με βάση τον αριθμό των γονιδίων που έχει αντιστοιχηθεί σε αυτόν (annotation). Για τα 4 PPI που ακολουθούν την τοπολογία ελεύθερης κλίμακας υπολογίσθηκαν οι πίνακες σημασιολογικής ομοίοτητας, δηλαδή τετραγωνικοί πίνακες που περιέχουν τη σημασιολογική ομοιότητα ανα δύο των όρων και με τους δύο τρόπους. Για την σύγκριση των πρωτεϊνών που απαρτίζουν τα δίκτυα επιλέχθηκε η Best Match Average στρατηγική.

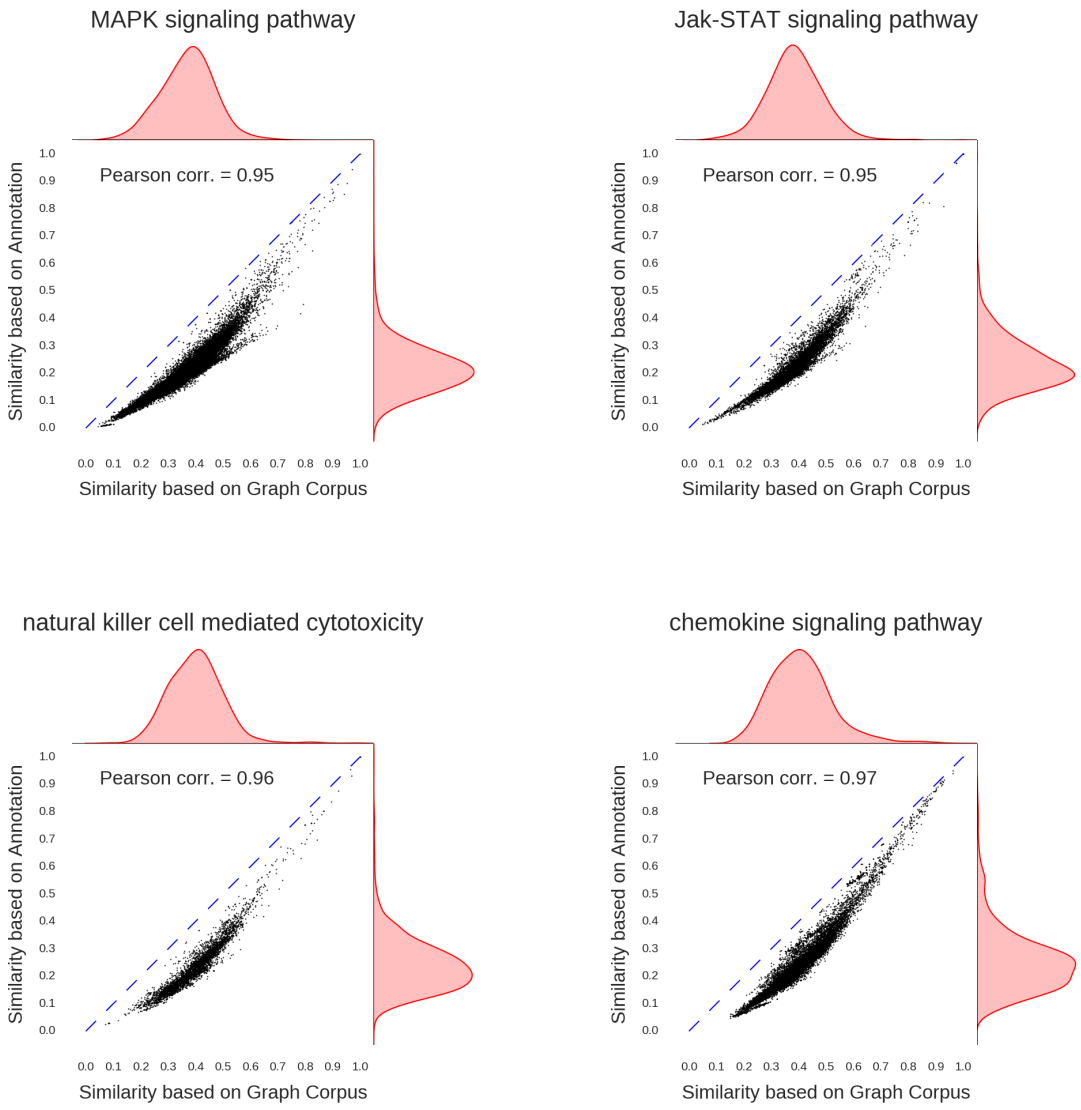
Στην συνέχεια παρουσιάζονται τα διαγράμματα σκέδασης (scatterplots) που αναπαριστούν την συσχέτιση που έχουν οι μετρικές σημασιολογικής ομοιότητας υπολογισμένες με βάση το graph corpus και το annotation στα 4 PPI. Συνεπώς επειδή και οι δύο τρόποι υπολογίζουν την ίδια ποσότητα αναμένεται έντονη γραμμική συσχέτιση μεταξύ τους.

Η συσχέτιση μεταξύ δύο μεταβλητών υπολογίζεται με βάση τον συντελεστή του Pearson:

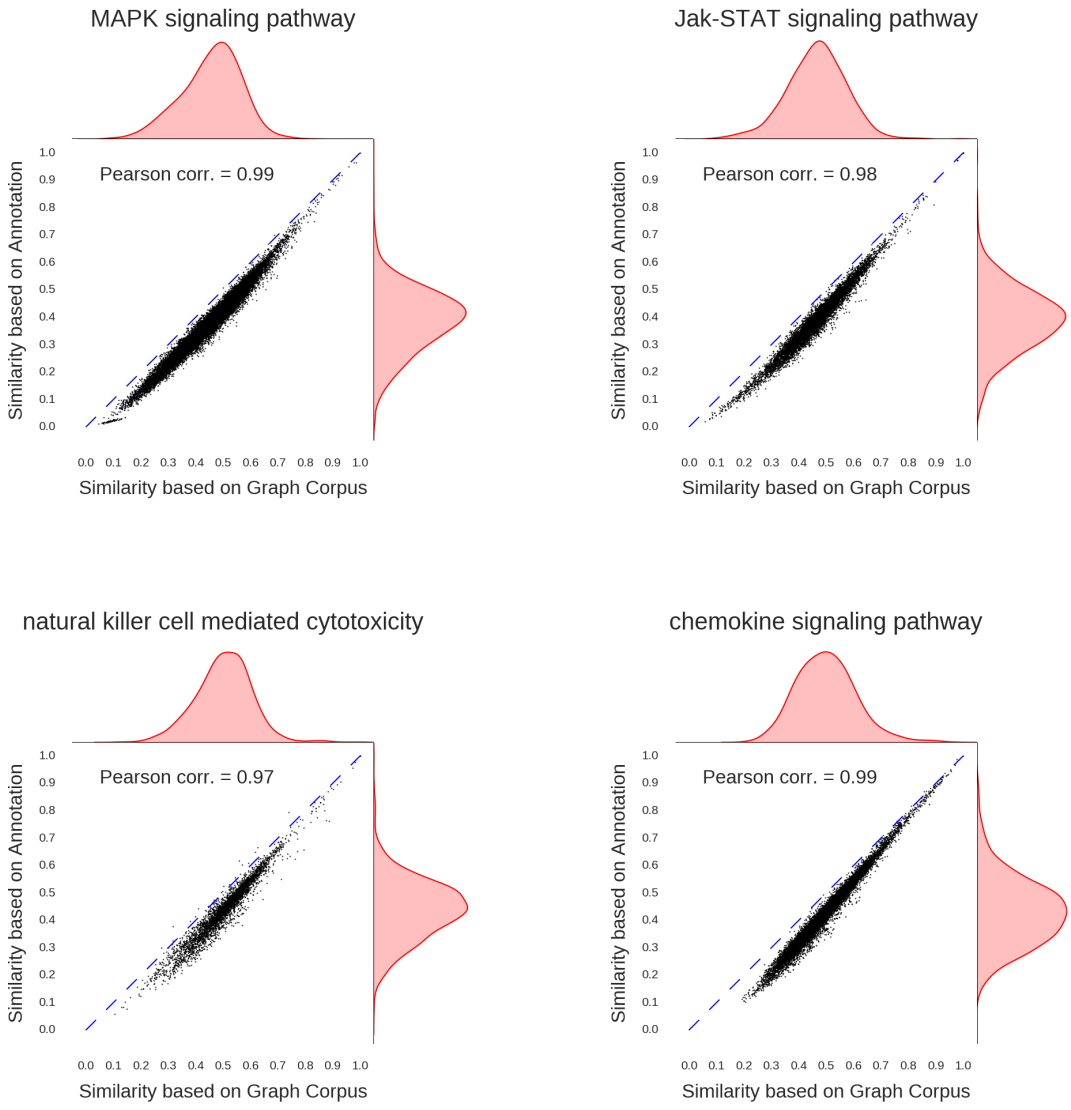
$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y} \tag{5.1}$$

Ο οποίος παίρνει τιμές στο διάστημα $[-1, 1]$ με το -1 να αντιστοιχεί σε δύο μεταβλητές αρνητικά συσχετισμένες, το 0 σε ασυσχέτιστες και το 1 σε τέλεια θετικά συσχετισμένες δηλαδή σε δύο μεταβλητές που συνδέονται με μια γραμμική σχέση της μορφής $y = x$.

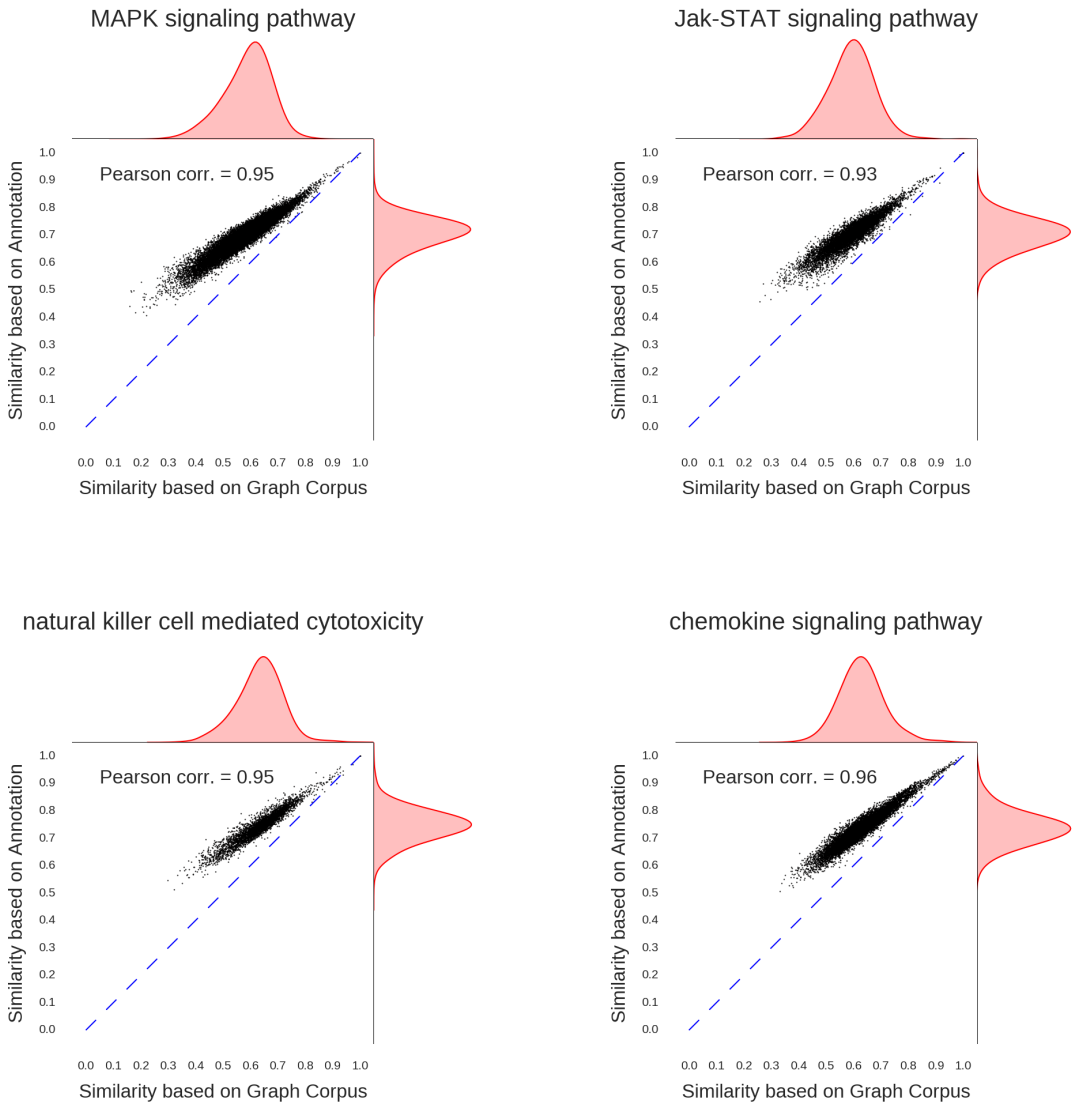
## 5.3 Διαγράμματα Σκέδασης



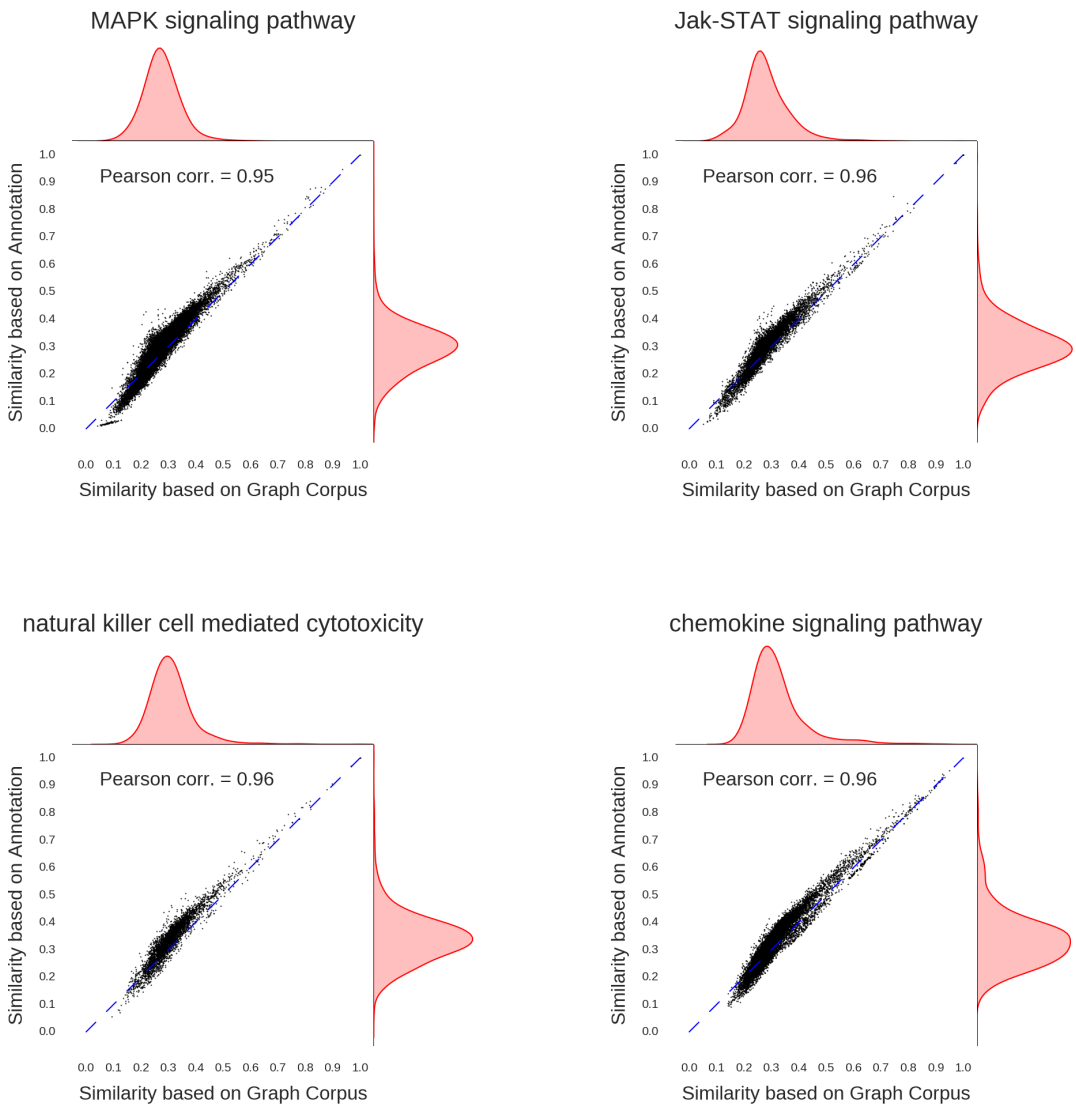

Σχήμα 5.3: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Resnik.



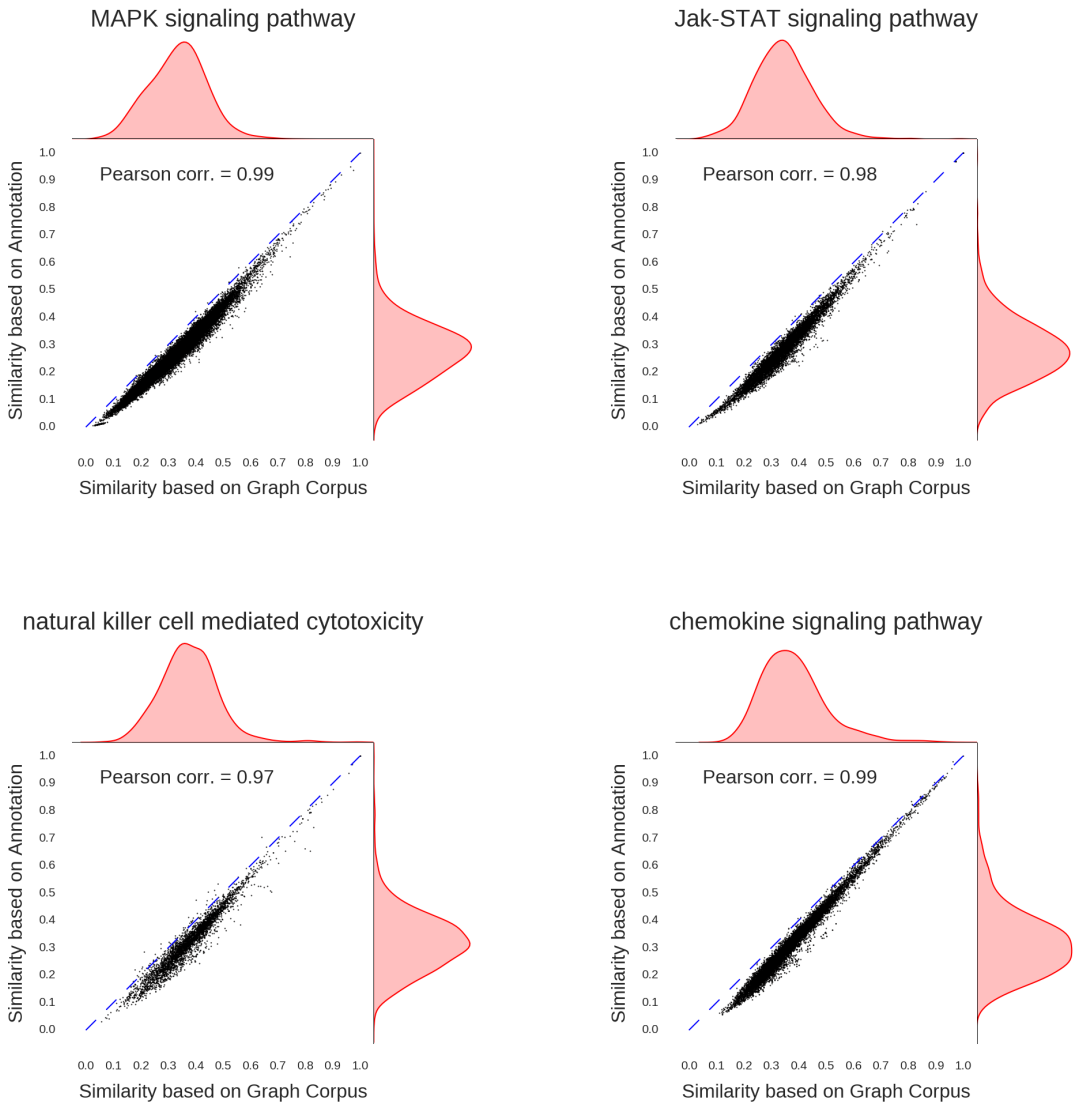

Σχήμα 5.4: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Lin.



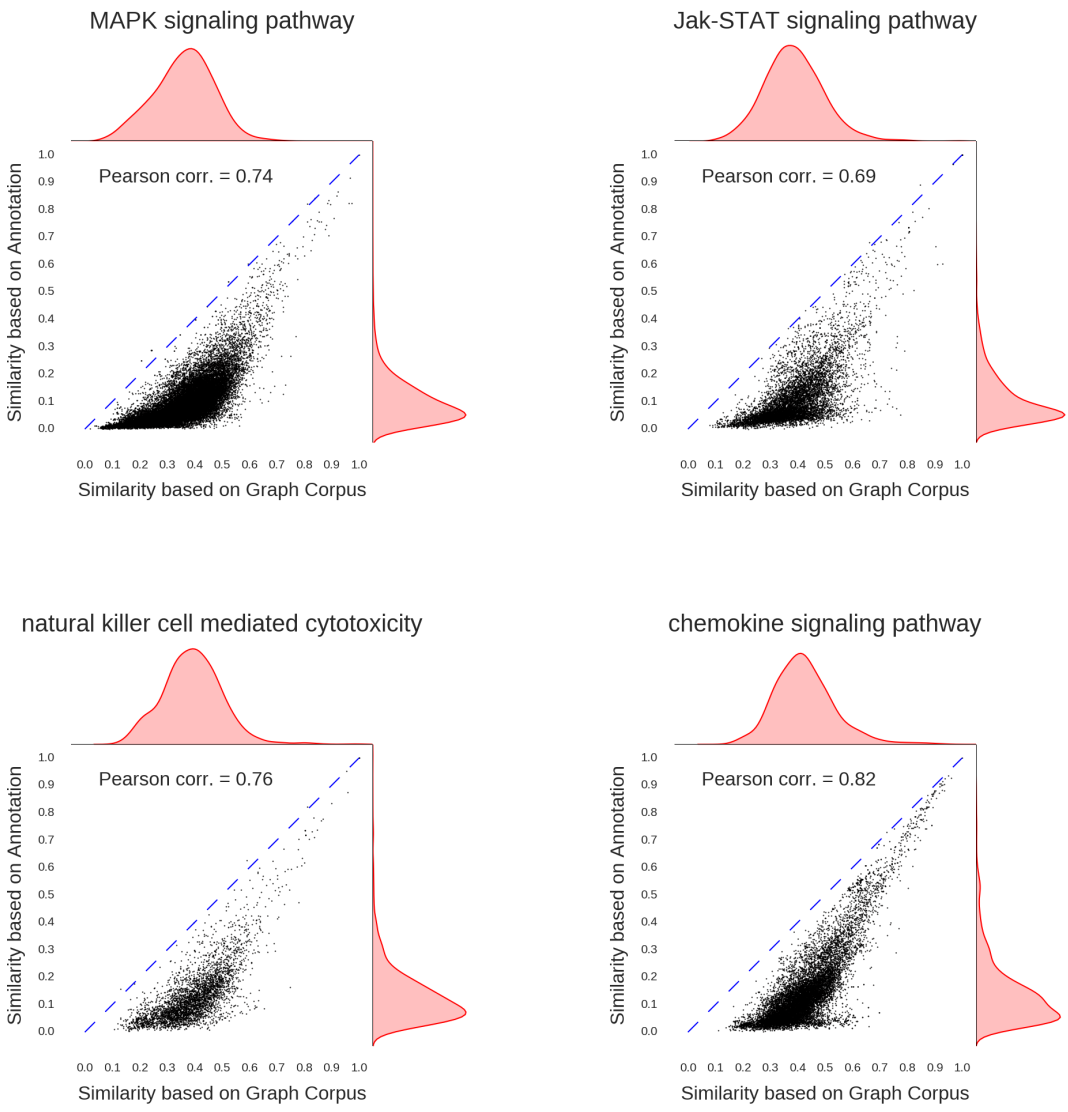

Σχήμα 5.5: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική των Jiang and Conrath.



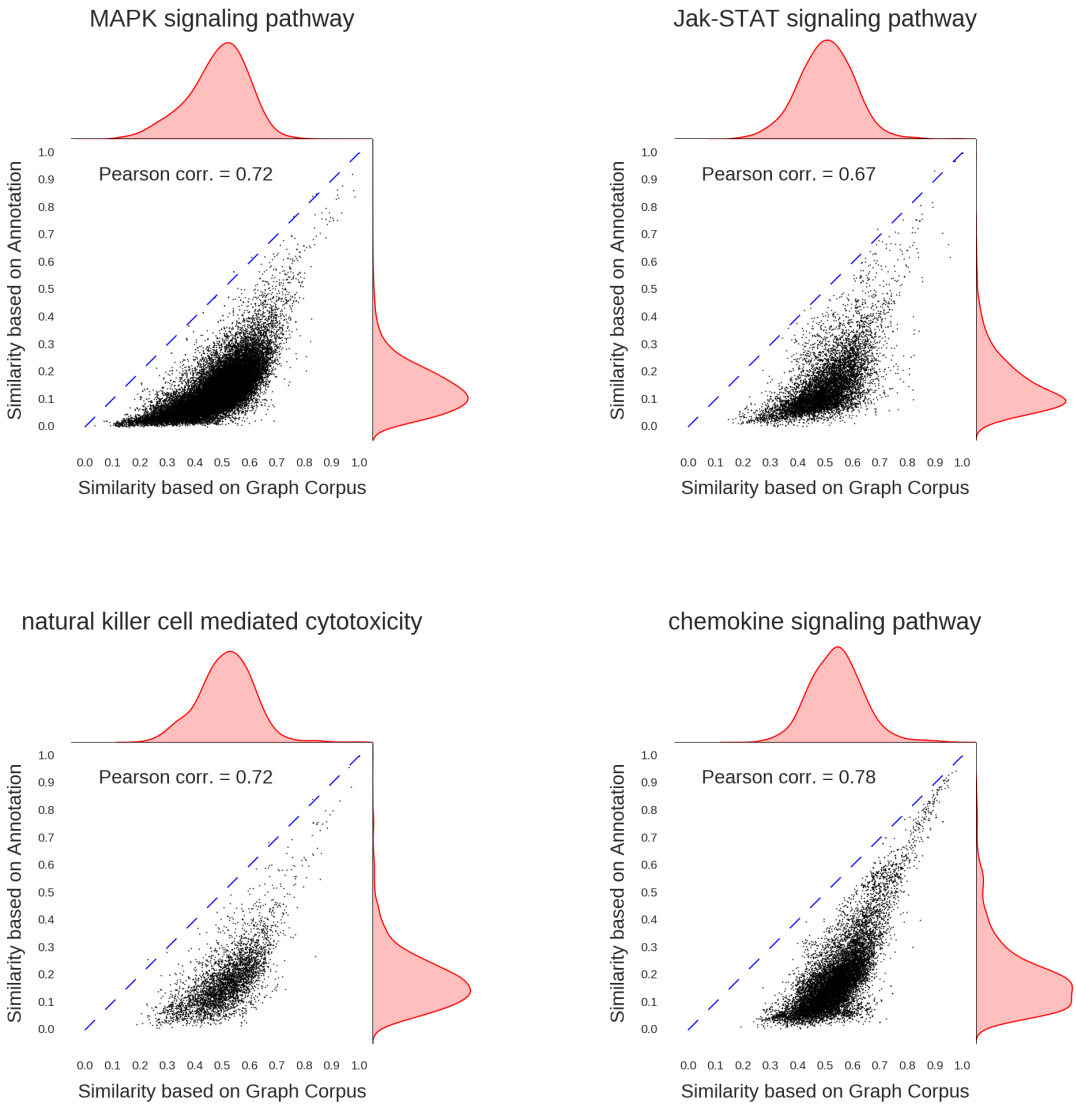

Σχήμα 5.6: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Schlicker.



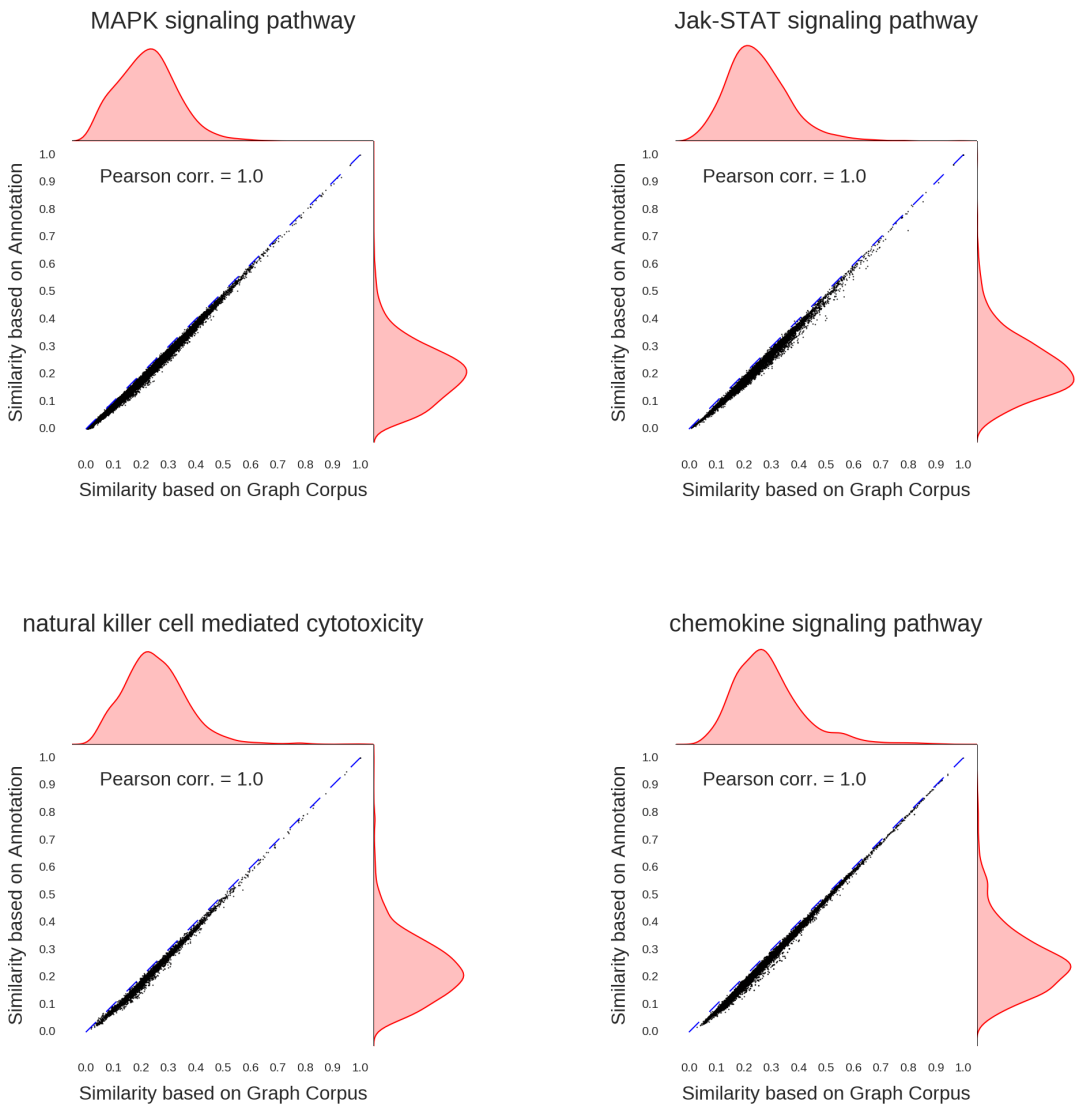

Σχήμα 5.7: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική των Pirro and Euzenat.



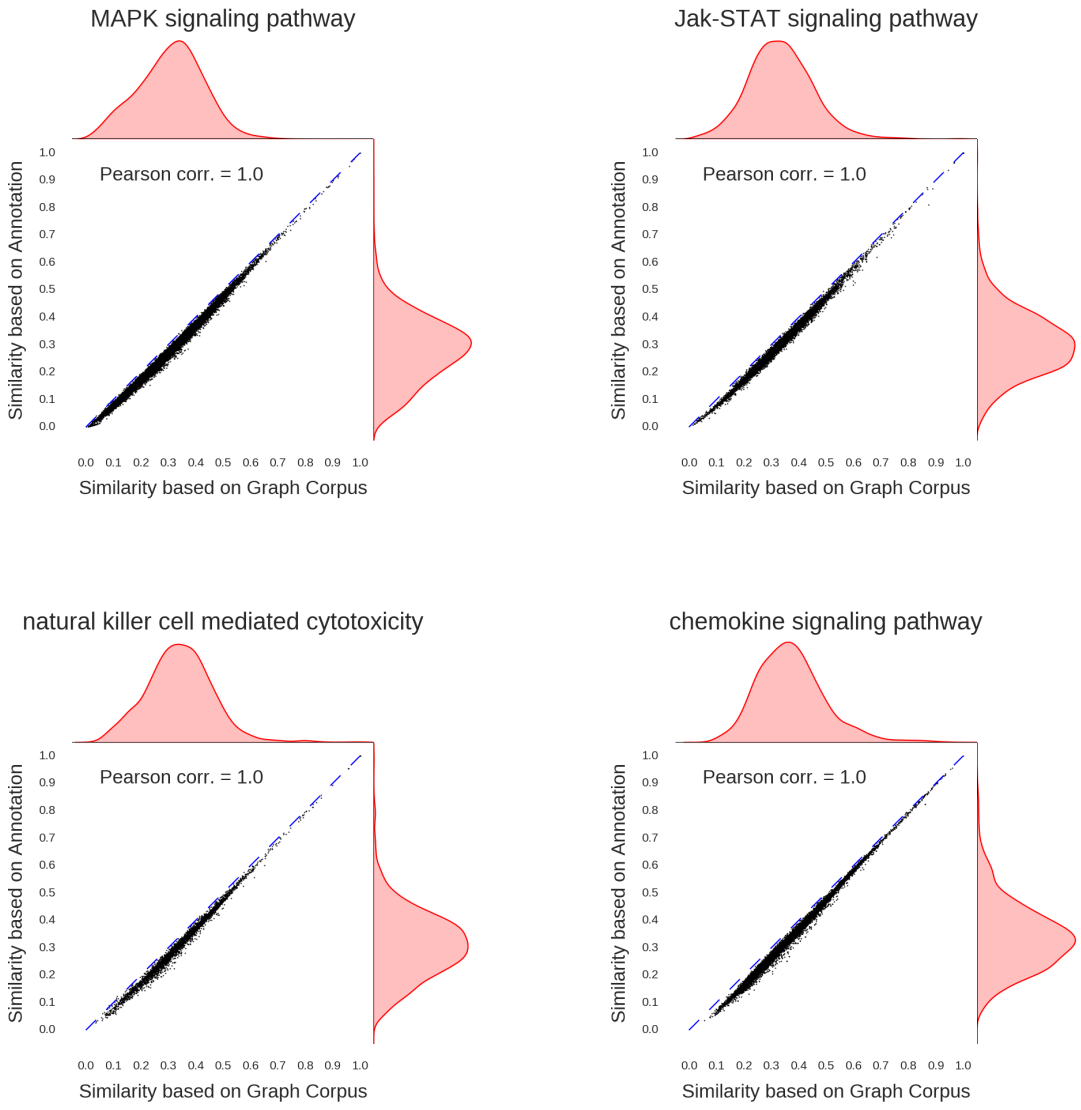

Σχήμα 5.8: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Jaccard.



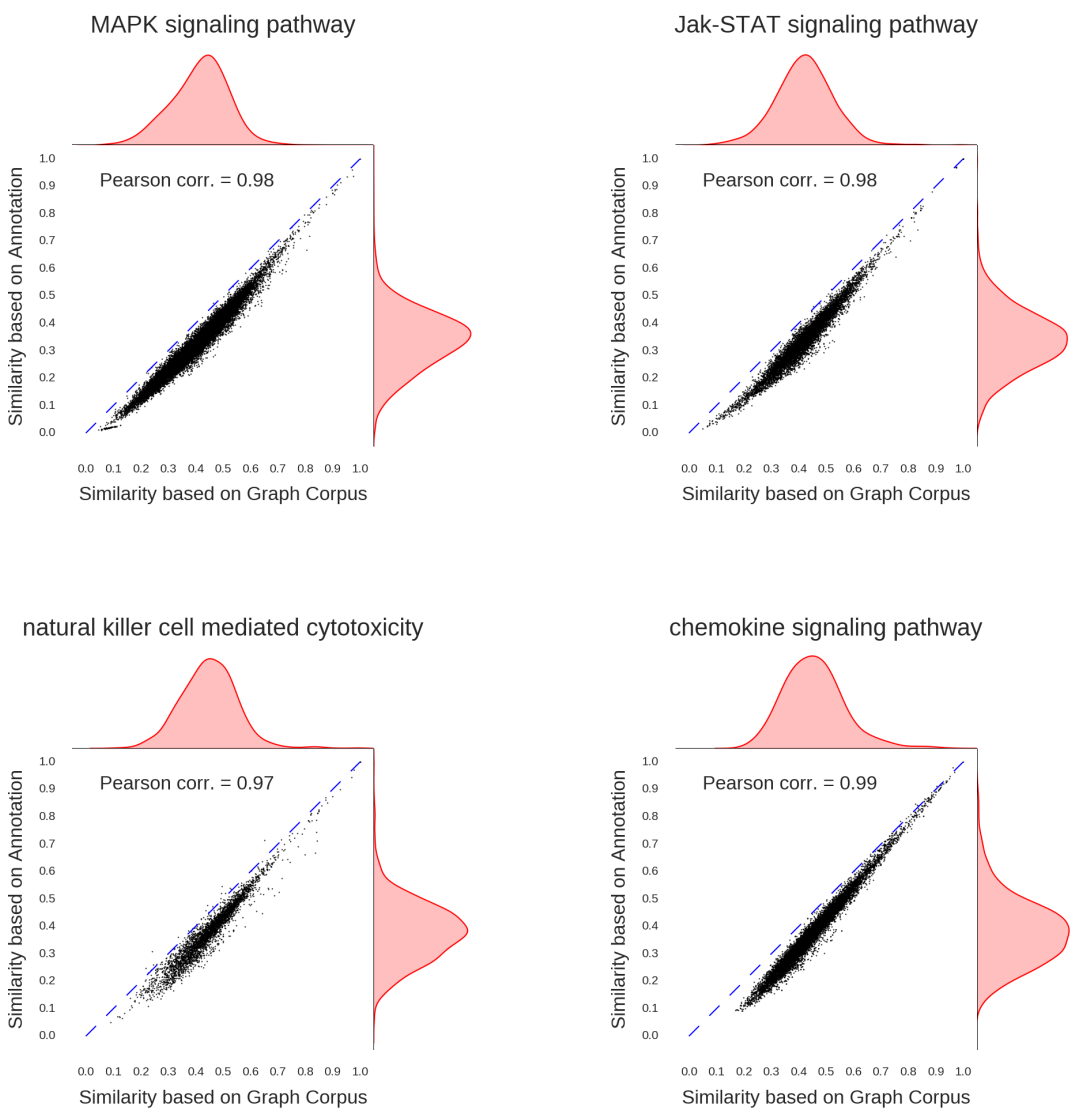

Σχήμα 5.9: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Dice.



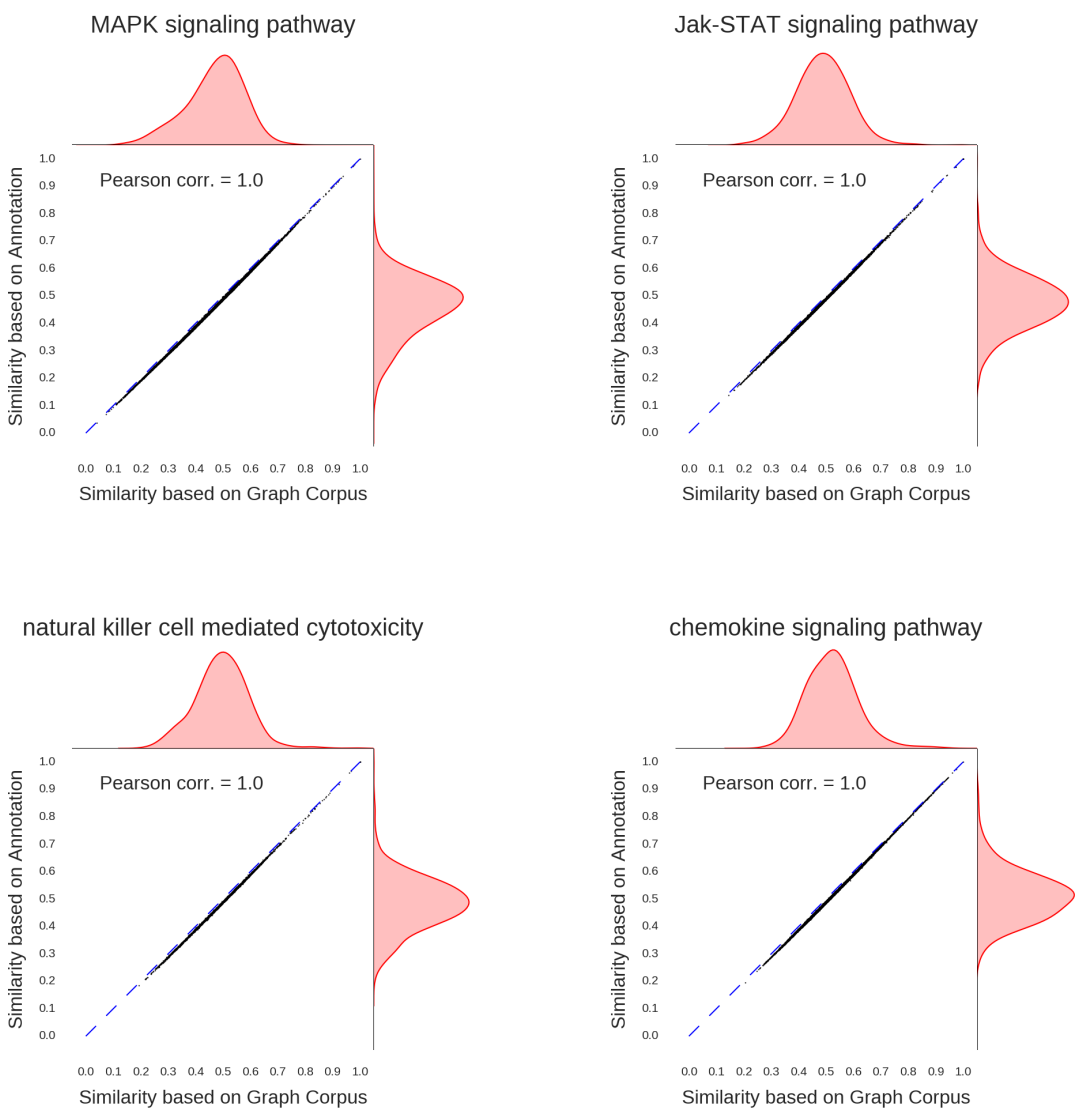

Σχήμα 5.10: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική Graph Information Content.




Σχήμα 5.11: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική του Mazandu.




Σχήμα 5.12: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική Nunivers.




Σχήμα 5.13: Scatterplots για την συσχέτιση της σημασιολογικής ομοιότητας με βάση το annotation και το graph corpus για τα 4 PPI με βάση την μετρική Aggregate IC.

## 5.4 Αποτελέσματα

Στον πίνακα 5.2 παρουσιάζονται συνοπτικά τα αποτελέσματα της συσχέτισης μεταξύ graph corpus και annotation για κάθε μία από τις μετρικές στα 4 PPI.

| **Μετρική** | MAPK Signaling Pathway | Jak-STAT Signaling Pathway | Natural Killer Cell Mediated Cytotoxicity | Chemokine Signaling Pathway | Μέσος Όρος Συσχέτισης |
|---|---|---|---|---|---|
| Resnik | 0.95 | 0.95 | 0.96 | 0.97 | 0.958 |
| Lin | 0.99 | 0.98 | 0.97 | 0.99 | 0.983 |
| Jiang and Conrtath | 0.95 | 0.93 | 0.95 | 0.96 | 0.947 |
| Schlicker | 0.95 | 0.96 | 0.96 | 0.96 | 0.958 |
| Pirro and Euzenat | 0.99 | 0.98 | 0.97 | 0.99 | 0.982 |
| Jaccard | 0.74 | 0.69 | 0.76 | 0.82 | 0.753 |
| Dice | 0.72 | 0.67 | 0.72 | 0.78 | 0.723 |
| GIC | 1 | 1 | 1 | 1 | 1 |
| Mazandu | 1 | 1 | 1 | 1 | 1 |
| Nunivers | 0.98 | 0.98 | 0.97 | 0.99 | 0.98 |
| AIC | 1 | 1 | 1 | 1 | 1 |

Πίνακας 5.2: Ο συντελεστής συσχέτισης για κάθεμια από τις μετρικές στα 4 PPI.

Οι μετρικές που χρησιμοποιούν στον υπολογισμό της ομοίοτητας δύο όρων το IC έχουν συσχέτιση μεγαλύτερη του 0.95, δηλαδή είναι σχεδόν τέλεια γραμμικά συσχετισμένες. Συνεπώς και με τους δύο τρόπους η ομοιότητα μεταξύ δύο όρων είναι σχεδόν ταυτόσημη. Αντίθετα ο Dice και ο Jaccard που δεν χρησιμοποιούν το IC των όρων αλλά μόνο τον αριθμό των προγόνων με βάση το graph corpus και τον αριθμό των γονιδίων που έχουν αντιστοιχηθεί στους όρους με βάση το annotation έχουν συντελεστή συσχέτισης 0.72 και 0.75 αντίστοιχα. Από το οποίο προκύπτει ότι οι μετρήσεις δεν έχουν έντονη γραμμική συσχέτιση και συνεπώς υπολογίζουν διαφορετικά την ομοιότητα στις δύο περιπτώσεις.

Από τα παραπάνω προκύπτει το συμπέρασμα ότι στην GO οι μετρικές είναι απαραίτητο να χρησιμοποιούν το IC για τον υπολογισμό της ομοίοτητας είτε άμεσα όπως το Aggregate IC είτε έμμεσα όπως οι υπόλοιπες. Συνεπώς οι μετρικές του Dice και του Jaccard δεν ενδείκνυται να χρησιμοποιούνται στην GO.

# Κεφάλαιο 6

# Ανάλυση Δικτύων από την **Reactome**

Στο κεφάλαιο αυτό αναλύθηκαν βιολογικά μονοπάτια από την βάση δεδομένων Reactome. Εξετάστηκε αν η κατανομή της σημασιολογικής απόστασης των κόμβων τους είναι όμοια με την κατανομή της πραγματικής απόστασης των κόμβων στο δίκτυο. Τέλος δημιουργήθηκαν τυχαία υπερδίκτυα από την συνένωση μονοπατιών φαινομεικά μηδενικής λειτουργικής συσχέτισης και εξετάστηκε η συμπεριφορά των σημασιολογικών μετρικών στην πρόβλεψη των λειτουργικών ομάδων τους.

## 6.1 **Reactome**

Η Reactome[59] είναι μια ελεύθερα προσβάσιμη βάση δεδομένων που περιέχει βιολογικά μονοπάτια που έχουν χαρακτηριστεί χειροκίνητα (manually curated). Τα βιολογικά μονοπάτια αναπαριστούν μια σειρά από αλληλεπιδράσεις μεταξύ συγκεκριμένων πρωτεϊνών και άλλων μακρομορίων που οδηγούν στην πραγματοποίηση συγκεκριμένων βιολογικών διαδικασιών. Υπάρχουν αρκετά είδη βιολογικών μονοπατιών, με τα πιο συνηθισμένα να εμπλέκονται στο μεταβολισμό, στη ρύθμιση των γονιδίων και στην εκπομπή σημάτων.

Από την βάση δεδομένων της Reactome επιλέχθηκαν μονοπάτια από διαφορετικές κατηγορίες με στόχο να έχουν όσο το δυνατόν λιγότερα κοινά γονίδια. Κάθε μονοπάτι επιτελεί μια βιολογική λειτουργία και τα γονίδια που εμπλέκονται σε αυτό συνδέονται λειτουργικά. Όπως φαίνεται στον πίνακα 6.1 επιλέχθηκαν 100 δυάδες, τρυάδες, τετράδες, πεντάδες και εξάδες βιολογικών μονοπατιών με σκοπό την συνθεσή τους και την κατασκευή βιολογικών υπερδικτύων όπου το κάθε μονοπάτι θα αντιπροσωπεύει μια ομάδα (cluster) γονιδίων.

Οι προϋποθέσεις που τέθηκαν κατα την επιλογή των δικτύων είναι:

- Κάθε δίκτυο να έχει περισσότερα από 30 γονίδια.
- Σε κάθε δίκτυο το συνολικό πλήθος γονιδίων που ανήκουν στην ίδια οικογένεια γονιδίων να μην υπερβαίνει το 30% του μεγέθους του.

- Σε κάθε n-αδα μονοπατιών πρέπει ανά δύο τα μονοπάτια να έχουν εως 3 κοινά γονίδια και η διαφορά του μεγέθους τους να μην είναι μεγαλύτερη από 30.
- Σε κάθε υποδίκτυο η μεγαλύτερη συνεκτική συνιστώσα του πρέπει να αποτελεί το 70% του μεγέθους του.

| **Αριθμός Μονοπατιών** | Εως 3 κοινά γονίδια | Διακριτά | Σύνολο |
|---|---|---|---|
| 2 | 26 | 74 | 100 |
| 3 | 56 | 44 | 100 |
| 4 | 66 | 34 | 100 |
| 5 | 77 | 23 | 100 |
| 6 | 89 | 11 | 100 |

Πίνακας 6.1: Οι ν-αδες βιολογικών μονοπατιών που επιλέχθηκαν από την Reactome.

Όπως φαίνεται από τον πίνακα 6.1 όσο αυξάνει ο αριθμός των μονοπατιών τόσο λιγότερα είναι τα διακριτά δίκτυα. Το οποίο καθιστά τα παραγώμενα υπερδίκτυα πιο ρεαλιστικά.

## 6.2 Συσχέτιση Τοπολογικής Απόστασης και Σημασιολογικής Απόστασης

Για κάθε υπερδίκτυο υπολογίσθηκε η τοπολογική απόσταση των κόμβων η οποία για να κανονικοποιηθεί αρκεί να διαιρεθεί με τη μέγιστη απόσταση που υπάρχει στο δίκτυο. Στην συνέχεια για κάθε μια μετρική υπολογίστηκε η σημασιολογική ομοιότητα των κόμβων των δίκτυών. Με χρήση του γραμμικού μετασχηματισμού $distance = 1 - similarity$ μετατράπηκε σε σημασιολογική απόσταση. Θεωρητικά, οι κεντρικοί κόμβοι με βάση την τοπολογία έχουν μικρή μέση απόσταση προς όλους τους υπόλοιπους, ενώ έχουν μεγάλη μέση σημασιολογική ομοιότητα καθώς εμπλέκονται σε πολλές λειτουργίες. Συνεπώς έχουν και μικρή σημασιολογική απόσταση. Αντίθετα τα φύλλα του δικτύου, δηλαδή οι περιφερειακοί κόμβοι έχουν μεγάλη μέση τοπολογική απόσταση και πολύ μικρή σημασιολογική ομοιότητα. Ωστόσο να σημειωθεί ότι δεν υπάρχει μία 1-1 αντιστοίχιση μεταξύ της τοπολογικής και της σημασιολογικής απόστασης καθώς η δεύτερη εξαρτάται από την ευαισθησία της κάθε μετρικής.

Από τα παραπάνω συμπεραίνεται ότι δεν μπορεί να γίνει απευθείας σύγκριση μεταξύ των κατανομών των αποστάσεων για να ελεγχθεί η συσχετισή τους. Όμως και στις δύο περιπτώσεις (τοπολογικά, σημασιολογικά) οι κόμβοι μπορούν να ταξινομηθούν σε απομακρυσμένους ή κεντρικούς, με βάση την κατάταξη της μέσης απόστασής τους, σε σχέση με την κατανομή των μέσων αποστάσεων όλων των κόμβων. Με βάση αυτή τη λογική, υπολογίστηκε η ποσοστιαία κατάταξη κάθε κόμβου, για τις δυο περιπτώσεις, και πιο συγκεκριμένα το ποσοστό των κόμβων

που έχουν μέση τιμή απόστασης μικρότερη απ αυτή του εξεταζόμενου κόμβου.

Για τη σύγκριση των δύο ιεραρχημένων κατανομών επιλέχθηκε ο μη παραμετρικός έλεγχος υποθέσεων του Wilcoxon ο οποίος ελέγχει σε επιπέδο σημαντικότητας 5% αν δύο κατανομές προέρχονται απ τον ίδιο πλυθησμό. Ως μηδενική υπόθεση ορίζει ότι οι δύο κατανομές είναι ίδιες και επιστρέφει το p-value του ελέγχου. Αν το $p - value < 0.5$ υπάρχουν ισχυρές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης. Στον ακόλουθο πίνακα φαίνεται το μέσο p-value κάθε μετρικής στις διαφορετικές ν-άδες βιολογικών μονοπατιών.

| **Average p-Value** | n=2 | n=3 | n=4 | n=5 | n=6 |
|---|---|---|---|---|---|
| Resnik | 0.858 | 0.849 | 0.869 | 0.851 | 0.810 |
| Lin | 0.858 | 0.843 | 0.836 | 0.786 | 0.742 |
| Jiang and Conrath | 0.811 | 0.816 | 0.825 | 0.737 | 0.682 |
| Schlicker | 0.825 | 0.824 | 0.862 | 0.839 | 0.832 |
| Pirro and Euzenat | 0.864 | 0.860 | 0.849 | 0.808 | 0.772 |
| Nunivers | 0.860 | 0.859 | 0.847 | 0.805 | 0.748 |
| Aggregate IC | 0.841 | 0.867 | 0.814 | 0.762 | 0.711 |
| Mazandu | 0.861 | 0.848 | 0.827 | 0.809 | 0.825 |
| Jaccard Coefficient | 0.858 | 0.868 | 0.844 | 0.792 | 0.755 |
| Dice Coefficient | 0.847 | 0.870 | 0.822 | 0.766 | 0.723 |
| Graph IC | 0.861 | 0.844 | 0.828 | 0.816 | 0.848 |
| Average | 0.849 | 0.850 | 0.838 | 0.798 | 0.768 |

Πίνακας 6.2: Το μέσο p-value που έχει κάθε μετρική στις ν-αδες βιολογικών μονοπατιών.

Από τα αποτελέσματα του στατιστικού ελέγχου προκύπτει ότι όλες οι μετρικές κατατάσσουν παρόμοια τους όρους με την τοπολογική τους κατάταξη. Στα μικρότερα υπερδίκτυα που αποτελούνται απο 2, 3 ή 4 βιολογικά μονοπάτια τα p-values είναι όλα υψηλότερα του 0.8 ενώ όσο αυξάνεται το μέγεθος των παραγόμενων δικτύων τα p-values μειώνονται. Το οποίο υποδηλώνει ότι όσο αυξάνεται το μέγεθος και η πολυπλοκότητα των δικτύων τόσο οι μετρικές δυσκολεύονται να εντοπίσουν τις σημασιολογικές ομοιότητες μεταξύ των όρων.

## 6.3 Ομαδοποίηση των Παραγόμενων Δικτύων

Κάθε υπερδίκτυο αποτελείται από την συνένωση διαφορετικών βιολογικών μονοπατιών που το καθένα περιγράφει μια συγκεκριμένη βιολογική λειτουργία. Συνεπώς κάθε δίκτυο ν-τάξης μπορεί να υποτεθεί ότι αποτελείται απο ν-ομαδες (clusters) γονιδίων οπου η κάθε μία αντιπροσωπεύει ένα βιολογικό μονοπάτι. Τις ομάδες αυτές θα τις θεωρήσουμε ως την state-of-the-art ομαδοποίηση των υπερδικτύων.

Στην συνέχεια θα ελεχθεί η ικανότητα των μετρικών να εντοπίζουν τις διαφορετικές λειτουργικές ομάδες στα υπερδίκτυα. Για να γίνει αυτό θα χρησιμοποιηθεί ο αλγόριθμος Affinity Pro-

pagation (AP) (Κεφ. 3) ο οποίος δέχεται ως είσοδο ένα πίνακα ομοιοτήτων (affinity matrix) και επιστρέφει τις διαφορετικές ομάδες για τα δεδομένα εισόδου. Ο έλεγχος της state-of-the-art ομαδοποίησης με τα αποτελέσματα του Affinity Propagation θα γίνει με την χρήση του αλγόριθμου Variation of Information (VI) (Κεφ. 3) ο οποίος ποσοτικοποιεί την διαφορά δύο ομαδοποιήσεων και επιστρεφει τιμές στο διάστημα [0,1] με το 0 να αντιστοιχεί σε δύο ίδιες.

Ο αλγόριθμος Affinity Propagation έχει δύο βασικές παραμέτρους το damping factor που αντιστοιχεί στην ταχύτητα που θα επιτευχθεί η σύγκλιση και το preference value που αντιστοιχεί στην αρχική προτίμηση που έχει ένας όρος να χαρακτηριστεί examplar της ομάδας. Εφόσον υπάρχουν οι state-of-the-art ομάδες το πρόβλημα της παρεμοτροποίησης του αλγορίθμου ανάγεται στην εύρεση της βέλτιστης δυνάτης ομαδοποίησης που υπολογίζει ο AP το οποίο θα γίνει μέσω της εξαντλητικής αναζήτησης των πιθανών ομαδοποιήσεων. Ως βέλτιστη θα χαρακτηριστεί αυτή που έχει το μικρότερο VI.

Ο υπολογισμός του VI θα γίνει για καθεμία μετρική σε κάθε κατηγορία υπερδικτύων η οποία αποτελείται απο 100 δίκτυα και συνεπώς το αποτέλεσμα θα προσεγγίζει την πραγματική μέση τιμή των μετρικών λόγω του Νόμου των Μεγάλων Αριθμών.

Στον πίνακα 6.3 φαίνεται το μέσο VI που έχει κάθε μετρική στα παραγόμενα υπερδίκτυα. Όσο αυξάνεται το μέγεθος των δικτύων τόσο αυξάνεται το VI, δηλαδή η απόσταση των μετρικών από την state-of-the-art ομαδοποίηση. Για μικρά υπερδίκτυα η συμπεριφορά των μετρικών είναι παρόμοια και δεν υπάρχουν μεγάλες διαφοροποιήσεις. Αντίθετα όσο αυξάνεται το μέγεθος των δικτύων οι μετρικές του Resnik και των Piro and Euzenat έχουν την καλύτερη συμπεριφορά. Ωστόσο επειδή τα αποτελέσματα είναι κοντινά και δεν υπάρχουν μεγάλες αποκλίσεις στην τιμή του VI θα παρουσιασθούν στον πίνακα 6.4 οι φορές που μια μετρική είχε το μικρότερο VI σε κάθε κατηγορία υπερδικτύων.

| **Average VI** | n=2 | n=3 | n=4 | n=5 |
|---|---|---|---|---|
| Graph IC | 0.153 | 0.208 | 0.234 | 0.263 |
| Schlicker | 0.168 | 0.203 | 0.239 | 0.276 |
| Resnik | 0.145 | 0.194 | 0.221 | 0.247 |
| Dice Coefficient | 0.156 | 0.201 | 0.227 | 0.259 |
| Pirro and Euzenat | 0.145 | 0.195 | 0.221 | 0.248 |
| Lin | 0.153 | 0.193 | 0.224 | 0.257 |
| Jiang and Conrath | 0.14 | 0.197 | 0.235 | 0.274 |
| Aggregate IC | 0.152 | 0.203 | 0.226 | 0.257 |
| Mazandu | 0.155 | 0.201 | 0.227 | 0.255 |
| Jaccard Coefficient | 0.152 | 0.203 | 0.23 | 0.258 |
| Nunivers | 0.15 | 0.195 | 0.225 | 0.254 |

Πίνακας 6.3: Το μέσο Variation of Information της κάθε μετρικές στα υπερδίκτυα

Με βάση τον πίνακα 6.4 προκύπτει ότι για μικρά δίκτυα υπάρχει μια ομοιομορφία στην κατανομή των ‘‘πρωτιών’’ των μετρικών. Πιο συγκεκριμένα σε υπερδίκτυα τάξης 2, 3 και 4 τις περισσότερες φορές την καλύτερη ομαδοποίηση υπολογίζει ο Jiang and Conrath και για n=5 ο Resnik.

| Count Best Clustering | n=2 | n=3 | n=4 | n=5 |
|---|---|---|---|---|
| Graph IC | 16 | 10 | 6 | 11 |
| Schlicker | 11 | 10 | 13 | 5 |
| Resnik | 22 | 14 | 11 | 18 |
| Dice Coefficient | 13 | 8 | 8 | 9 |
| Pirro and Euzenat | 18 | 11 | 9 | 11 |
| Lin | 9 | 14 | 10 | 6 |
| Jiang and Conrath | 23 | 23 | 20 | 12 |
| Aggregate IC | 15 | 7 | 5 | 9 |
| Mazandu | 11 | 8 | 4 | 7 |
| Jaccard Coefficient | 13 | 9 | 11 | 6 |
| Nunivers | 12 | 9 | 4 | 8 |

Πίνακας 6.4: Ο αριθμός των φορών που η μετρική υπολογίζει το μικρότερο VI.

Όπως φαίνεται στον πίνακα 6.5 η μετρική του Jiang and Conrath έχει την καλύτερη συμπεριφόρα όταν η τιμή του VI σε ένα υπερίδκτυο είναι μεγαλύτερη από την μέση τιμή του VI στην αντίστοιχη κατηγορία. Το οποίο οφειλείται στο ότι η εν λόγω μετρική αντιστοιχεί μεγαλύτερη ομοιότητα στους εξεταζόμενους όρους από ότι οι υπόλοιπες μετρικές με συνέπεια όταν οι υπόλοιπες δεν έχουν καλή συμπεριφορά αυτή να είναι η καλύτερη.

| Count Best Clustering Higher than Average VI | n=2 | n=3 | n=4 | n=5 |
|---|---|---|---|---|
| Graph IC | 0 | 0 | 0 | 1 |
| Schlicker | 4 | 0 | 0 | 2 |
| Resnik | 5 | 5 | 2 | 2 |
| Dice Coefficient | 3 | 3 | 2 | 1 |
| Pirro and Euzenat | 2 | 0 | 3 | 3 |
| Lin | 2 | 2 | 0 | 1 |
| Jiang and Conrath | 14 | 4 | 12 | 6 |
| Aggregate IC | 3 | 2 | 2 | 0 |
| Mazandu | 1 | 0 | 1 | 1 |
| Jaccard Coefficient | 1 | 0 | 0 | 0 |
| Nunivers | 3 | 0 | 0 | 1 |

Πίνακας 6.5: Ο αριθμός των φορών που η μετρική υπολογίζει το μικρότερο VI όταν το VI είναι μεγαλύτερο από το μέσο VI.

# Βιβλιογραφία


[1] Merriam-Webster Online Dictionary: Definiton of Ontology .

[2] Miguel-Angel Sicilia (2014). Handbook of Metadata,Semantics and Ontologies. World Scientific Publishing .

[3] B. Chandrasekaran, J. R. Josephson and V. R. Benjamins, "What are ontologies, and why do we need them?," in IEEE Intelligent Systems and their Applications, vol. 14, no. 1, pp. 20-26, Jan.-Feb. 1999. doi: 10.1109/5254.747902 .

[4] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLOS Biology 13(7) .

[5] Open Biological and Biomedical Foundry .

[6] Bogumil M. Konopka (2015), Biomedical ontologies—A review, Biocybernetics and Biomedical Engineering, Volume 35, Issue 2, Pages 75-86 .

[7] The Gene Ontology Consortium, Ashburner M, Ball CA, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000;25(1):25-29.

[8] Köhler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Research 45. Database issue(2017).

[9] Susan M. Bello, Mary Shimoyama, Elvira Mitraka et al. Disease Ontology: improving and unifying disease annotations across species. Disease Models and Mechanisms 11 (2018) .

[10] Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Preece J, Athreya B, Mungall CJ, Rensing S, Hiss M, Lang D, Reski R, Berardini TZ, Li D, Huala E, Schaeffer M, Menda N, Arnaud E, Shrestha R, Yamazaki Y, Jaiswal P. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol 54 (2013) .

[11] Gene Ontology Consortium: Ontology Documentation .

[12] Du Plessis, Louis, Nives Škunca, and Christophe Dessimoz. "The What, Where, How and Why of Gene Ontology—a Primer for Bioinformaticians." Briefings in Bioinformatics 12.6 (2011): 723–735 .

[13] Gene Ontology Consortium: Ontology Structure .

[14] Gene Ontology Consortium: Ontology Relations .

[15] Gene Ontology Consortium: GO Annotations .

[16] Valentini, G. (2011). True Path Rule Hierarchical Ensembles for Genome-Wide Gene Function Prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(3), 832–847.

[17] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology 5(7)

[18] Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. In: IEEE Transaction on Systems, Man, and Cybernetics. 19. pp 17–30

[19] Pekar, V. and Staab, S. (2002). Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In Proceedings of 19th International Conference on Computational Linguistics, COL-ING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002 .

[20] Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the 14th International Joint Conference on Artificial Intelligence. pp 448–453

[21] David Sánchez, Montserrat Batet, David Isern, Ontology-based information content computation, Knowledge-Based Systems, Volume 24, Issue 2, 2011, Pages 297-303

[22] Lin D (1998) An information-theoretic definition of similarity. In: Proc. of the 15th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann. pp 296–304.

[23] Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the 10th International Conference on Research on Computational Linguistics, Taiwan.

[24] Pirró G., Euzenat J. (2010) A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider P.F. et al. (eds) The Semantic Web – ISWC 2010. ISWC 2010. Lecture Notes in Computer Science, vol 6496. Springer, Berlin, Heidelberg

[25] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T (2006) A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics 7.

[26] Francisco M Couto,Mário J Silva (2011). Disjunctive shared information between ontology concepts: application to Gene Ontology. Journal of Biomedical Semantics.

[27] Mazandu GK, Emile RM, Mulder NJ, Nicola (2016). The 'A-DaGO-Fun' Package.

[28] Pesquita C, Faria D, Bastos H, Falcão AO, Couto FM: Evaluating GO-based Semantic Similarity Measures. In: Proceedings of the 10th Annual Bio-Ontologies Meeting (Bio-Ontologies 2007) 2007; Vienna, Austria; 2007: 37-40.

[29] Mazandu GK, Mulder NJ (2014) Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type?. PLOS ONE 9(12)

[30] Mazandu GK, Mulder NJ (2012) A topology-based metric for measuring term similarity in the Gene Ontology. Adv Bioinformatics 2012

[31] X. Song, L. Li, P. K. Srimani, P. S. Yu and J. Z. Wang, "Measure the Semantic Similarity of GO Terms Using Aggregate Information Content," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 3, pp. 468-476, May-June 2014

[32] Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, et al. (2005) Correlation between gene expression and go semantic similarity. In: IEEE/ ACM Transactions on Computational Biology and Bioinformatics vol. 2, no. 4

[33] Lord P, Stevens R, Brass A, Goble C (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283.

[34] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, and C.-F. Chen (2007), "A New Method to Measure the Semantic Similarity of GO Terms," Bioinformatics, vol. 23, pp. 1274-1281.

[35] Pesquita C, Faria D, Bastos H, Ferreira AE, Falcão AO, Couto FM (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9.Suppl 5.

[36] Cormen,Leiserson,Rivest, Stein (2009) Introduction to Algorithms, Third Edition The MIT Press

[37] Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. Using graph theory to analyze biological networks. BioData Mining. 2011;4:10.

[38] Brandes U (2001). A faster algorithm for betweenness centrality. Journal of mathematical sociology,25(2), 163-177

[39] Albert R, Jeong H, Barabási A-L (2000). Error and attack tolerance of complex networks. Nature 406, 378–382.

[40] Albert-László Barabási (2016) Network Science Chapter 4: Scale-Free Property

[41] Albert-László Barabási, Réka Albert (1999) Emergence of Scaling in Random Networks. Science Vol. 286, Issue 5439, pp. 509-512

[42] Νικολάου Χ, Χουβαρδάς Π (2015) Υπολογιστική βιολογία. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

[43] Yook, S. , Oltvai, Z. N. and Barabási, A. (2004), Functional and topological characterization of protein interaction networks. Proteomics, 4: 928-942.

[44] Barabási AL, Oltvai ZN (2004). Network biology understanding the cell's functional organization. Nature Genetics 5(2) 101-113.

[45] Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics 7

[46] James Vlasblom, Shoshana J Wodak (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC bioinformatics 10

[47] Van Dongen S (2000). Graph Clustering by Flow Simulation. In PhD Thesis University of Utrecht

[48] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc.

[49] Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C. 28 (1): 100–108

[50] Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association. 58 (301): 236–244.

[51] Brendan J. Frey; Delbert Dueck (2007). "Clustering by passing messages between data points". Science. 315 (5814): 972–976

[52] Silke Wagner, Dorothe Wagner (2007). Comparing Clusterings - An Overview. Technical Report 2006-04.

[53] Rand, William M (1972) Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association

[54] E-B Fowlkes, C-L Mallows (1983). A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association 78(383) :553–569

[55] Meila Marina (2003). Comparing Clusterings-an information based distance Journal of Multivariate Analysis 98.

[56] Breuer et al., InnateDB: systems biology of innate immunity and beyond - recent updates and continuing curation (2013). Nucl.Acids Res 41.

[57] Pearson G, Robinson F, Beers Gibson T, Xu BE, Karandikar M, Berman K, Cobb MH (2001). "Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions". Endocrine Reviews. 22 (2): 153–83.

[58] Jason S. Rawlings, Kristin M. Rosler, Douglas A. Harrison (2004). The JAK/STAT signaling pathway. Journal of Cell Science 117: 1281-1283

[59] Croft D, O'Kelly G, Wu G, et al (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research, 39