ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Οπτικοποίηση και συγκριτική ανάλυση κοινοτήτων σε βιοϊατρικά δίκτυα»

Γκόντα Μαρία

Πτυχιούχος Πληροφορικής με Εφαρμογές στη Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ (2021)

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

SCHOOL OF SCIENCE
DEPARTMENT OF BIOLOGY

MASTER IN «BIOINFORMATICS»

Master Diploma Thesis

«Visualization and comparison of communities in biomedical networks»

Maria Gkonta

BSc Computer Science and Biomedical Informatics, University of Thessaly

ATHENS (2021)

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Οπτικοποίηση και συγκριτική ανάλυση κοινοτήτων σε βιοϊατρικά δίκτυα»

Τριμελής εξεταστική επιτροπή

Ερευνητής Β' Γεώργιος Παυλόπουλος (Κύριος επιβλέπων)
ΕΚΕΒΕ 'Αλέξανδρος Φλέμινγκ"

Καθηγητής Παντελής Μπάγκος (Επιβλέπων)
Τμήμα Πληροφορικής με Εφαρμογές στην Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας

Αναπληρώτρια Καθηγήτρια Βασιλική Οικονομίδου
Τομέας Βιολογίας Κυττάρου και Βιοφυσικής,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ» υπό την επίβλεψη του Ερευνητής Β' Δρ. Γεώργιου Παυλόπουλου. Με την ολοκλήρωση της διπλωματικής μου, θα ήθελα να εκφράσω την εκτίμηση, τον σεβασμό και τις ευχαριστίες μου στον Δρ. Γεώργιο Παυλόπουλο, ο οποίος έδειξε εμπιστοσύνη και μου προσέφερε την ευκαιρία να ασχοληθώ με το θέμα της διπλωματικής μου εργασίας. Τον ευχαριστώ θερμά, λοιπόν, για τη βοήθεια και την πολύτιμη καθοδήγηση που μου παρείχε κατά τη διάρκεια εκπόνησης της. Μέσα από τη συνεργασία μας, απέκτησα τα απαραίτητα εφόδια για τη συνέχεια και για αυτό τον ευχαριστώ ιδιαίτερα.

Θα ήθελα να ευχαριστήσω τον Καθηγητή Δρ. Παντελή Μπάγκο του Τμήματος Πληροφορικής με Εφαρμογές στην Βιοϊατρική, του Πανεπιστημίου Θεσσαλίας, ο οποίος δέχθηκε να είναι επιβλέπων της συγκεκριμένης διπλωματικής εργασίας. Επίσης, θα ήθελα να τον ευχαριστήσω για τις πολύτιμες συμβουλές του που πρόσφερε κατά την διάρκεια της συνεργασίας μας.

Επιπλέον, θα ήθελα να ευχαριστήσω την Αναπληρώτρια Καθηγήτρια Δρ. Βασιλική Οικονομίδου, του Τμήμα Βιολογίας, του Εθνικού και Καποδιστριακού Πανεπιστημίου για την παρουσία της στην τριμελή επιτροπή.

Επίσης, θα ήθελα να ευχαριστήσω θερμά τον Μεταδιδακτορικό Δρ. Ευάγγελο Καρατζά για τη συνεργασία στη δημιουργία του VICTOR, καθώς επίσης και για την αρωγή του καθ' όλη τη διάρκεια της διπλωματικής αλλά και για την υπομονή και τη συμβολή του στην επίλυση των προβλημάτων που προέκυπταν, αλλά και τον χρόνο που μου αφιέρωσε.

Ιδιαίτερες, φυσικά, ευχαριστίες, οφείλω και στην υπόλοιπη ομάδα που βοήθησαν στην ανάπτυξη του VICTOR και συγκεκριμένα τον Μεταδιδακτορικο Δρ. Φώτη Μπαλτούμα και την συμφοιτήτρια μου κ. Ιωάννα Χοτόβα για την συνεργασία μας. Στις ευχαριστίες μου δεν θα μπορούσα να παραλείψω και την Μεταδιδακτορικό κ. Παναγιώτα Κοντού για την προσφορά της (case study) στην αξιολόγηση του VICTOR.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και τα αδέρφια μου για τη βαθιά εμπιστοσύνη τους, καθώς επίσης και για την ψυχολογική και ηθική τους ενίσχυση όλα αυτά τα χρόνια, κατά τη διάρκεια των οποίων δεν έπαψαν λεπτό να πιστεύουν σε μένα αλλά και τους φίλους μου που μου στάθηκαν και με βοήθησαν σε όλες τις δυσκολίες.

Περίληψη

Ομαδοποίηση είναι η διαδικασία ομαδοποίησης διαφορετικών δεδομένων με βάση τις παρόμοιες ιδιότητες που εμφανίζουν. Η ομαδοποίηση έχει εφαρμογές σε διαφορετικές μελέτες που αφορούν διάφορους τομείς όπως στη θεωρία γραφημάτων, στην ανάλυση εικόνας, στην αναγνώριση προτύπων, στη στατιστική και άλλα. Σήμερα, υπάρχουν πολλοί αλγόριθμοι και εργαλεία ικανά να δημιουργήσουν αποτελέσματα ομαδοποίησης. Ωστόσο, διαφορετικοί αλγόριθμοι ή διαφορετική παραμετροποίηση αυτών μπορεί να οδηγήσει στον σχηματισμό πολύ διαφορετικών ομάδων. Με αυτόν τον τρόπο, ο χρήστης συχνά αναγκάζεται να φιλτράρει και να συγκρίνει χειροκίνητα αυτά τα αποτελέσματα, προκειμένου να αποφασίσει ποια από αυτές παράγει το ιδανικό σύμπλεγμα. Για την αυτοματοποίηση αυτής της διαδικασίας, σε αυτήν την εργασία, παρουσιάζουμε το VICTOR, την πρώτη πλήρως διαδραστική εφαρμογή οπτικής ανάλυσης που επιτρέπει τη σύγκριση και οπτικοποίηση διαφόρων αλγορίθμων ομαδοποίησης. Το VICTOR μπορεί να χειριστεί πολλαπλά αποτελέσματα συμπλέγματος ταυτόχρονα και να τα συγκρίνει χρησιμοποιώντας δέκα διαφορετικές μετρικές. Τα αποτελέσματα ομαδοποίησης μπορούν να φιλτραριστούν και να συγκριθούν μεταξύ τους με τη χρήση διαδραστικών heatmaps, bar plots, δικτύων συσχέτισης, sankey και circos plots. Η λειτουργικότητα του VICTOR αναδεικνύεται χρησιμοποιώντας τρία παραδείγματα. Στην πρώτη περίπτωση, συγκρίνουμε πέντε διαφορετικούς αλγόριθμους σε ένα σύνολο δεδομένων αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης, ενώ στο δεύτερο παράδειγμα, δοκιμάζουμε τέσσερις διαφορετικές παραμέτρους του ίδιου αλγορίθμου συμπλέγματος που εφαρμόζονται στο ίδιο σύνολο δεδομένων. Τέλος, ως τρίτο παράδειγμα, συγκρίνουμε τέσσερις διαφορετικές μετα-αναλύσεις με ιεραρχικά ομαδοποιημένα διαφορικά εκφρασμένα γονίδια που βρέθηκαν να εμπλέκονται στο έμφραγμα του μυοκαρδίου. Το VICTOR είναι διαθέσιμο στο http://bib.fleming.gr:3838/VICTOR.

Abstract

Clustering is the process of grouping different data based on the similar properties they display. Clustering has applications in different studies related to various fields such as graph theory, image analysis, pattern recognition, statistics and more. Nowadays, there are many algorithms and tools capable of generating clustering results. However, different algorithms or different parameterization may result in very different clusters. This way, the user is often forced to manually filter and compare these results in order to decide which of them produce the ideal clusters. To automate this process, in this study, we present VICTOR, the first fully interactive and dependency-free visual analytics application which allows the comparison and visualization of various clustering algorithms. VICTOR can handle multiple cluster results simultaneously and compare them using ten different metrics. Clustering results can be filtered and compared using interactive heatmaps, bar plots, correlation networks, sankey and circos plots. VICTOR's functionality is demonstrated using three examples. In the first case, we compare five different algorithms on a protein-protein interaction dataset whereas in the second example, we test four different parameters of the same clustering algorithm applied on the same dataset. Finally, as a third example, we compare four different meta-analyses with hierarchically clustered differentially expressed genes found to be involved in myocardial infarction. VICTOR is available at http://bib.fleming.gr:3838/VICTOR.

Πίνακας περιεχομένων

Κατάλογος εικόνων και πινάκων

Εισαγωγή

Το θέμα της παρούσας εργασίας αφορά την οπτικοποίηση και συγκριτική ανάλυση κοινοτήτων σε βιοϊατρικά δίκτυα. Αναμφίβολα, οι τεχνικές αναλύσεις δεδομένων έχουν αναπτυχθεί τα τελευταία χρόνια, η ομαδοποίηση έχει βρει εφαρμογή σε διαφόρους τομείς και δίνει λύσεις σε πολλά προβλήματα της σύγχρονης εποχής. Πιο συγκεκριμένα, η ομαδοποίηση είναι ιδιαίτερα χρήσιμη για την ανάλυση δεδομένων και την ανάδειξη των σχέσεων που αυτά έχουν, καθώς και την εξαγωγή χρήσιμων πληροφοριών που μπορεί να περιέχουν. Ουσιαστικά, είναι μια διαδικασία κατά την οποία τα δεδομένα χωρίζονται σε ομάδες σύμφωνα με τα κοινά χαρακτηριστικά τους. Τα τελευταία χρόνια, μια πληθώρα αλγορίθμων ομαδοποίησης έχουν προταθεί και είναι διαθέσιμοι στη βιβλιογραφία. Πολλοί από τους αλγορίθμους αυτούς μπορούν να εφαρμοστούν σε διαφορετικούς τομείς και σε διαφορετικά σετ δεδομένων. Παράλληλα με την ανάπτυξη των αλγορίθμων ομαδοποίησης έχουν δημιουργηθεί και πολλά εργαλεία τα οποία διαθέτουν διαφορετικούς αλγορίθμους ομαδοποίησης και αποδίδουν οπτικά τα αποτελέσματα αυτών. Πολλές φορές, μάλιστα, είναι δύσκολο να αποφασιστεί τι αποτελεί καλή ομαδοποίηση. Οι χρήστες χρειάζεται να θέτουν κριτήρια, κατά τέτοιο τρόπο, ώστε τα αποτελέσματα των ομαδοποιήσεων να ανταποκρίνονται στις ανάγκες τους.

Η ομαδοποίηση αναγνωρίζεται ευρέως ως χρήσιμο εργαλείο σε πολλές εφαρμογές και η ποικιλία των εφαρμογών της είναι μεγάλη σε πολλά επιστημονικά πεδία. Οι διαφορές στις υποθέσεις και το πλαίσιο μεταξύ των διαφορετικών ερευνητικών κοινοτήτων οδήγησαν στη δημιουργία διαφόρων μεθοδολογιών και αλγορίθμων ομαδοποίησης. Καθένας από τους αλγορίθμους αυτούς μπορεί να έχει διαφορετικό αποτέλεσμα, καθώς στηρίζονται σε διαφορετικές τεχνικές. Η επιλογή εξαρτάται από τη μορφή των δεδομένων και από τον ίδιο τον χρήστη. Γενικότερα, έχει δημιουργηθεί η ανάγκη εκτίμησης και πιστοποίησης των αποτελεσμάτων ομαδοποίησης. Στο στάδιο της αξιολόγησης των αποτελεσμάτων ελέγχεται η εγκυρότητα των ομάδων κι αν το τελικό αποτέλεσμα του αλγορίθμου είναι επιτυχές. Υπάρχουν πολλά διαδεδομένα μέτρα ομοιότητας που χρησιμοποιούνται για τη σύγκριση των τεχνικών ομαδοποίησης. Ωστόσο, η ανάγκη αυτοματοποίησης της διαδικασίας σύγκρισης των αποτελεσμάτων των αλγορίθμων ομαδοποίησης είναι μεγάλη γιατί σε πολλές περιπτώσεις παρατηρείται δυσκολία στην αξιολόγηση και κριτική αποτίμηση του πιο αποδοτικού αλγορίθμου. Στο πλαίσιο της παρούσας εργασίας, παρουσιάζεται η ανάπτυξη της εφαρμογής VICTOR, μιας οπτικό-αναλυτικής εφαρμογής σύγκρισης αποτελεσμάτων ομαδοποίησης. Μέσω της εφαρμογής αυτής, παρέχεται η δυνατότητα συγκριτικής ανάλυσης και οπτικοποίησης αποτελεσμάτων ομαδοποίησης. Ειδικότερα, δίνεται στον χρήστη η δυνατότητα σύγκρισης των αρχείων προερχόμενων από αλγορίθμους ομαδοποίησης, αξιοποιώντας διάφορες μετρικές σύγκρισης ομαδοποιήσεων. Καταληκτικά, παρέχονται διαφορετικοί διαδραστικοί τρόποι οπτικοποίησης αυτών.

Στο πρώτο κεφάλαιο, επιχειρείται η εννοιολογική προσέγγιση της έννοιας της «ομαδοποίησης». Η «ομαδοποίηση» έχει βρει εφαρμογή σε διάφορους τομείς, με αποτέλεσμα την ανάπτυξη ποικίλων αλγορίθμων ομαδοποιήσεις. Στη συνέχεια αυτού του κεφαλαίου, αναφέρονται οι κατηγορίες αλγορίθμων ομαδοποίησης και αναλύεται ο τρόπος λειτουργίας κάποιων βασικών αλγορίθμων που ανήκουν σε κάθε κατηγορία. Μέσω αυτού του κεφαλαίου, αναδεικνύεται η ποικιλομορφία κ η διαφορετικότητα των αλγορίθμων αυτών, τόσο στον τρόπο λειτουργίας τους όσο και στις παραγόμενες ομαδοποιήσεις. Επιπλέον, αναφέρονται κάποιοι

αλγόριθμοι, οι οποίοι βρίσκουν εφαρμογή στην ομαδοποίηση βιολογικών δικτύων, τα αποτελέσματα των οποίων θα αξιοποιηθούν για αξιολόγηση της αποδοτικότητας του VICTOR.

Στο δεύτερο κεφάλαιο, παρουσιάζονται τα βιολογικά δίκτυα. Ειδικότερα, αναλύονται κάποιες από τις βασικές κατηγορίες βιολογικών δικτύων που συναντάμε στον τομέα της Βιολογίας και της Βιοπληροφορικής. Επιπλέον, γίνεται αναλυτική περιγραφή της δομής, του περιεχομένου και του τρόπου κατασκευής των δικτύων αυτών.

Στο επόμενο κεφάλαιο, αναφέρονται ενδεικτικά κάποιες από τις πλατφόρμες ομαδοποίησης. Πρώτα, παρατίθενται εργαλεία και λογισμικά που διαθέτουν διάφορους αλγορίθμους ομαδοποίησης κι έχουν σαν αποτέλεσμα ομαδοποιήσεις, καθώς και οπτική απεικόνιση αυτών. Στη συνέχεια, παρουσιάζονται κάποια από διαθέσιμα εργαλεία σύγκρισης ομαδοποιήσεων που συναντήσαμε στη βιβλιογραφία.

Το τέταρτο κεφάλαιο της δοθείσας εργασίας θεωρείται ιδιαίτερα σημαντικό, αφού παρουσιάζεται η εφαρμογή VICTOR. Αρχικά, αναφέρονται τα διάφορα προβλήματα που συναντάμε κατά την επιλογή αλγορίθμων ομαδοποίησης και αξιολόγησης αυτών τα οποία αποτέλεσαν και κίνητρο για τον σχεδιασμό του VICTOR. Εν συνέχεια, γίνεται μια σύντομη περιγραφή της εφαρμογής.

Στο επόμενο κεφάλαιο, παρουσιάζεται η βιβλιοθήκη mclustcomp η οποία χρησιμοποιήθηκε από την εφαρμογή VICTOR. Συγκεκριμένα, η βιβλιοθήκη αυτή παρέχει διαφορετικές μετρικές σύγκρισης ομαδοποιήσεων, κάποιες από τις οποίες επιλέχθηκαν να χρησιμοποιηθούν και στο VICTOR. Μες στην ενότητα αυτή, αναφέρονται κάποιοι βασικοί υπολογισμοί που αξιοποιούνται από τις μετρικές. Επιπροσθέτως, αναλύονται όλες οι διαθέσιμες μετρικές της βιβλιοθήκης χωρισμένες σε τρεις κατηγορίες (καταμέτρηση ζευγών, αλληλοεπικάλυψη των σετ\αντιστοίχιση και θεωρία πληροφοριών). Γίνεται αναλυτική περιγραφή του τρόπου λειτουργίας τους και δίνεται ο τύπος υπολογισμού τους.

Επιπρόσθετα, στο παρακάτω κεφάλαιο περιγράφεται η μορφή των αρχείων εισόδου της εφαρμογής VICTOR, καθώς επίσης κι οι δυνατότητες φιλτραρίσματος που παρέχει για τα αρχεία αυτά ώστε να συμβαδίζουν με τις απαιτήσεις της βιβλιοθήκης. Τέλος, περιγράφονται οι διαθέσιμες οπτικοποιήσεις για το μέγεθος και το περιεχόμενο των αρχείων.

Εν συνεχεία της εργασίας, παρουσιάζονται οι πέντε διαφορετικές οπτικοποιήσεις των συγκρίσεων των αποτελεσμάτων των ομαδοποιήσεων που διατίθενται στο VICTOR. Γίνεται αναλυτική περιγραφή κάθε οπτικοποίησης ξεχωριστά και πώς αναπαρίστανται οι συγκρίσεις των ομαδοποιήσεων σε αυτές.

Στο όγδοο κεφάλαιο, αναφέρεται μια επιπλέον δυνατότητα του VICTOR. Αυτή είναι η ικανότητα υπολογισμού της μετρικής conductance των δικτύων. Ουσιαστικά, υπολογίζεται το conductance ενός αρχείου δικτύου, το οποίο σχετίζεται με ένα αποτέλεσμα ομαδοποίησης της αρεσκείας των χρηστών. Τα αποτελέσματα εμφανίζονται σε ιστόγραμμα.

Στο τελευταίο μέρος της εργασίας, παρουσιάζεται η λειτουργικότητα του VICTOR εφαρμόζοντας σε αυτό τρία διαφορετικά παραδείγματα. Τα παραδείγματα αυτά είναι: i) η σύγκριση των αποτελεσμάτων ομαδοποίησης πέντε διαφορετικών αλγορίθμων (MCL, SPICi, Louvain, Walktrap και Label Propagation), ii) η σύγκριση των αποτελεσμάτων ομαδοποίησης διαφορετικών παραμέτρων του ίδιου αλγορίθμου πάνω στο ίδιο σετ δεδομένων και τέλος iii) η σύγκριση και οπτικοποίηση τεσσάρων διαφορετικών μετα-αναλύσεων με ιεραρχικά ομαδοποιημένα διαφορικά εκφρασμένα γονίδια που βρέθηκαν να εμπλέκονται στο έμφραγμα του μυοκαρδίου. Συγκεκριμένα, σε κάθε παράδειγμα περιγράφεται η διαδικασία σύγκρισης

των αποτελεσμάτων με τη βοήθεια του VICTOR κι αναλύονται τα αποτελέσματα που προκύπτουν.

1. Ομαδοποίηση

1.1 Η έννοια της ομαδοποίησης

Ως «ομαδοποίηση» χαρακτηρίζεται η διαδικασία εκείνη κατά την οποία τα δεδομένα χωρίζονται σε ομάδες σύμφωνα με τα κοινά χαρακτηριστικά τους ή τις όμοιες ιδιότητες τους, με τέτοιο τρόπο ώστε μια ομάδα να είναι μια συλλογή από αντικείμενα τα οποία είναι όμοια μεταξύ τους και ανόμοια με αντικείμενα που ανήκουν σε άλλες ομάδες. Η ομαδοποίηση είναι μια τεχνική που εφαρμόζεται σε πολλά ερευνητικά πεδία όπως: Στη Βιολογία [1], στην Ιατρική [2], στην Πληροφορική [3], στη Φυσική, στη Χημεία, στο Marketing [4] και στη Βιοπληροφορική. Συγκεκριμένα, στο πεδίο της Βιοϊατρικής για παράδειγμα η ιεραρχική ομαδοποίηση μπορεί να χρησιμοποιηθεί για να ομαδοποιήσει δεδομένα σε ένα Heatmap, να συνάγει φυλογενέσεις ή να ομαδοποιήσει δεδομένα με βάση τις ομοιότητες τους (π.χ. ομοιότητες ακολουθίας ή μοτίβα έκφρασης γονιδίου). Μια άλλη εφαρμογή της ομαδοποίησης είναι ο εντοπισμός κοινοτήτων ή πυκνά συνδεδεμένων περιοχών σε δίκτυα [5]. Παραδείγματος χάριν, η ανίχνευση ισχυρά συνδεδεμένων στοιχείων (π.χ. πρωτεΐνες ή γονίδια σε δίκτυα αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης (PPI)). Σε πειράματα προσδιορισμού αλληλουχίας RNA μονού κυττάρου (scRNAseq) , η ομαδοποίηση με βάσει μεθόδων μείωσης διαστάσεων μπορεί να χρησιμοποιηθεί για να ομαδοποιήσει κύτταρα με παρόμοια προφίλ γονιδιακής έκφρασης. Στην Ιατρική, οι αλγόριθμοι επεξεργασίας σήματος μπορούν να χρησιμοποιηθούν για την αναγνώριση αντικειμένων σε μια εικόνα και να λειτουργούν ως αυτοματοποιημένα διαγνωστικά εργαλεία. Στο πεδίο της μεταγονιδιωματικής, η μέθοδος binning μπορεί να χρησιμοποιηθεί για την ομαδοποίηση συνεχόμενων αλληλουχιών (contings) και την ανάθεσή τους σε μεμονωμένα γονιδιώματα [6].

1.2 Είδη ομαδοποιήσεων

Οι αλγόριθμοι ομαδοποιήσεων μπορούν να χωριστούν σε τέσσερις βασικές κατηγορίες με βάση τη μέθοδο που χρησιμοποιούν για να δημιουργήσουν ομάδες. Οι κατηγορίες αυτές είναι: Ιεραρχικοί αλγόριθμοι ομαδοποίησης (hierarchical), διαμεριστικοί αλγόριθμοι ομαδοποίησης (partitional), αλγόριθμοι ομαδοποίησης βασισμένοι στην πυκνότητα (density-based) και αλγόριθμοι ομαδοποίησης βασισμένοι σε πλέγμα (grid-based). Οι ιεραρχικοί αλγόριθμοι ομαδοποιήσεων βρίσκουν ομάδες χρησιμοποιώντας προηγούμενες καθιερωμένες ομάδες, ενώ οι διαμεριστικοί αλγόριθμοι ομαδοποιήσεων δημιουργούν τις ομάδες εκείνη τη στιγμή. Οι ιεραρχικοί αλγόριθμοι μπορεί να είναι συγκεντρωτικοί ή διαχωριστικοί. Οι συγκεντρωτικοί (Agglomerative) αλγόριθμοι ξεκινούν με το κάθε στοιχείο ως ξεχωριστή ομάδα και τα συγχωνεύουν διαδοχικά σε μεγαλύτερες ομάδες [7]. Οι διαχωριστικοί (Divisive) αλγόριθμοι ξεκινούν με ολόκληρο το σετ των στοιχείων και το χωρίζουν διαδοχικά σε μικρότερες ομάδες (Εικόνα 1A).

1.2.1 Ιεραρχική ομαδοποίηση (Hierarchical clustering)

Στην ιεραρχική ομαδοποίηση, η ιεραρχία χτίζεται σταδιακά από το κάθε στοιχείο συγχωνεύοντας αυτά σε ομάδες. Ουσιαστικά, τα αντικείμενα κατηγοριοποιούνται σε μια ιεραρχία παρόμοια με ένα δέντρο σαν διάγραμμα που ονομάζεται «δενδρόγραμμα» (Εικόνα

1B), το οποίο απεικονίζει τις διαιρέσεις που γίνονται σε κάθε διαδοχικό στάδιο ανάλυσης. Αρχικά, πρέπει να υπολογιστεί ποιο στοιχείο θα συγχωνευτεί σε ποια ομάδα. Συνήθως, ομαδοποιούνται τα στοιχεία που είναι πιο κοντά μεταξύ τους. Επομένως, πρέπει να οριστεί μια απόσταση μεταξύ των στοιχείων αυτών. Για αυτό τον λόγο, βασικό βήμα στην ιεραρχική ομαδοποίηση είναι η επιλογή της μετρικής απόστασης. Μια από τις πιο συχνά χρησιμοποιούμενες μετρικές είναι η απόσταση Manhattan, η οποία ισούται με το άθροισμα της συνολικής απόστασης κάθε μεταβλητής. Μια άλλη κοινή μετρική είναι η Ευκλείδεια απόσταση, η οποία υπολογίζεται ως η ρίζα του αθροίσματος του τετραγώνου της απόστασης δύο μεταβλητών.

Η απόσταση ***Manhattan*** υπολογίζει την απόσταση που θα διανυθεί από ένα σημείο σε ένα άλλο αν ακολουθείται μια διαδρομή τύπου πλέγματος. Η απόσταση Manhattan μεταξύ δύο αντικειμένων είναι το άθροισμα των διαφορών των αντίστοιχων στοιχείων τους. Ο τύπος ανάμεσα σε ένα σημείο $\mathrm{X} = \mathrm{X}_1, \mathrm{X}_2,\ldots$ και $Y = \Upsilon_1, \Upsilon_2,\ldots$ είναι: $d = \sum_{i=1}^{n} |X_i - Y_i|$ [7].

Η **Ευκλείδεια** απόσταση μετράει το μήκος του ευθύγραμμου τμήματος των σημείων που ενώνει. Ο τύπος για αυτήν την απόσταση μεταξύ ενός σημείου $\mathrm{X} = \mathrm{X}_1, \mathrm{X}_2,\ldots$ και ενός σημείου $Y = \Upsilon_1, \Upsilon_2,\ldots$ είναι: $d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ [7]

Όσον αφορά τη συγκεντρωτική ιεραρχική ομαδοποίηση, κάθε συσσωμάτωση συμβαίνει σε μια μεγαλύτερη απόσταση ανάμεσα στις ομάδες από ό,τι στην προηγούμενη περίπτωση συσσωμάτωσης. Η διαδικασία αυτή μπορεί να σταματήσει, όταν οι ομάδες είναι πολύ μακριά για να συγχωνευτούν ή όταν υπάρχει μικρός αριθμός ομάδων. Ανάλογα με το πώς ορίζεται η απόσταση, υπάρχουν και διαφορετικές μέθοδοι όπως: Ομαδοποίηση μονής σύνδεσης, ομαδοποίηση πλήρους σύνδεσης, κεντροειδής σύνδεσης και μέσης σύνδεσης.

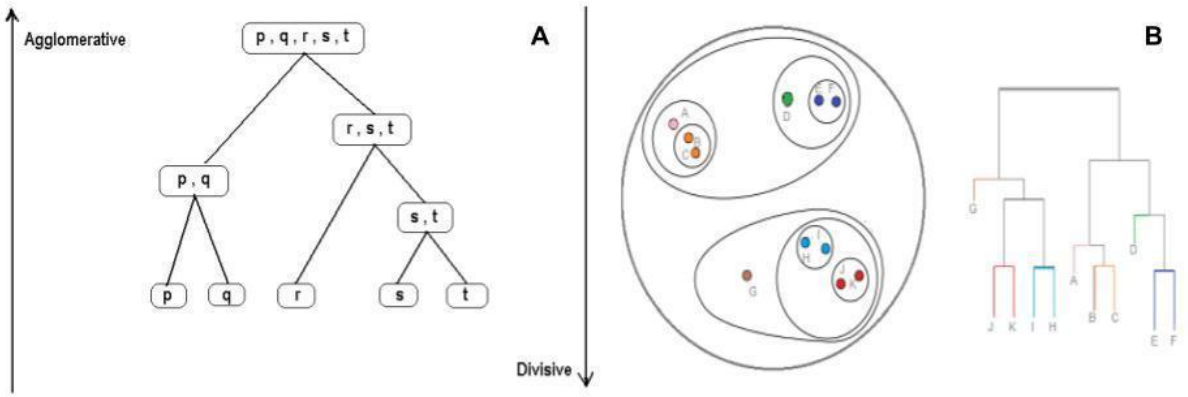
Η ομαδοποίηση **μονής σύνδεσης (single linkage)** βασίζεται στην τεχνική του κοντινότερου γείτονα. Το χαρακτηριστικό της μεθόδου είναι ότι η απόσταση μεταξύ των ομάδων ορίζεται ως η ελάχιστη απόσταση μεταξύ των στοιχείων κάθε ομάδας. Στη μέθοδο μονής σύνδεσης, η απόσταση υπολογίζεται ως: $min\{d(i,j)$, όπου τα i και j ανήκουν σε διαφορετικές ομάδες. Σε κάθε στάδιο της ιεραρχικής ομαδοποίησης, οι ομάδες που έχουν τη μικρότερη απόσταση συγχωνεύονται. Αυτό γίνεται, ούτως ώστε η νεοσυσταθείσα ομάδα, να έχει κατά μέσο όρο ελάχιστες αποστάσεις μεταξύ των σημείων.

Η ομαδοποίηση **πλήρης σύνδεσης (complete linkage)** βασίζεται στο μακρινότερο γείτονα. Αυτή η μέθοδος ομαδοποίησης είναι το αντίθετο της απλής σύνδεσης. Η απόσταση μεταξύ ομάδων ορίζεται τώρα ως η μεγαλύτερη απόσταση μεταξύ του πιο απομακρυσμένου ζεύγους αντικειμένων, τα οποία ανήκουν σε διαφορετικές ομάδες. Στην πλήρη σύνδεση, η απόσταση υπολογίζεται ως: $max\{d(i,j)$, όπου τα i και j ανήκουν σε διαφορετικές ομάδες. Η απόσταση μεταξύ δύο ομάδων δίνεται από την τιμή του μακρύτερου συνδέσμου μεταξύ των ομάδων. Σε κάθε στάδιο της ιεραρχικής ομαδοποίησης, οι ομάδες, για τις οποίες η απόσταση είναι μικρότερη, συγχωνεύονται.

Η ομαδοποίηση **κεντροειδούς σύνδεσης (centroid linkage)** βασίζεται στην απόσταση μεταξύ των κεντροειδών δύο ομάδων**.** Είναι πιθανό οι μικρότερες ομάδες να μοιάζουν περισσότερο με μια νέα μεγαλύτερη ομάδα από ό,τι με τις μεμονωμένες ομάδες τους, προκαλώντας αντιστροφή στο δενδρόγραμμα. Στην κεντροειδή σύνδεση, η απόσταση μεταξύ δύο ομάδων είναι η απόσταση μεταξύ των δύο μέσων διανυσμάτων των ομάδων. Σε κάθε

στάδιο της διαδικασίας συνδυάζουμε τις δύο ομάδες που έχουν τη μικρότερη απόσταση κεντροειδούς.

Στην ομαδοποίηση **μέσης σύνδεσης (average linkage)** η απόσταση μεταξύ δύο συστάδων ορίζεται ως ο μέσος όρος των αποστάσεων μεταξύ όλων των ζευγών αντικειμένων, όπου κάθε ζεύγος αποτελείται από ένα αντικείμενο από κάθε ομάδα. Σε κάθε στάδιο της ιεραρχικής ομαδοποίησης, οι ομάδες, για τις οποίες η απόσταση είναι η ελάχιστη, συγχωνεύονται.



Εικόνα 1: Α) τρόπος λειτουργίας συγκεντρωτικής και διαχωριστικής ιεραρχικής ομαδοποίησης [8]. Β) Ένα δενδρόγραμμα (δεξιά) που αντιπροσωπεύει εμφωλευμένες ιεραρχικές ομάδες [9].

Η ιεραρχική ομαδοποίηση είναι ιδιαίτερα χρήσιμη και στον τομέα της Βιολογίας. Κάποια παραδείγματα εφαρμογής της ιεραρχικής ομαδοποίησης στον τομέα αυτό είναι πάνω σε γονίδια και τα επίπεδα έκφρασης τους σε διάφορους τομείς για την παραγωγή δενδογραμμάτων και heatmap για την ανάλυση και οπτικοποίηση δεδομένων μικροσυστοιχιών [10]. Η ιεραρχική ομαδοποίηση μπορεί επίσης να χρησιμοποιηθεί για την ανάλυση δεδομένων γενετικής αλληλεπίδρασης που βασίζονται σε μεταλλάγματα διπλής διαγραφής [11]. Όσον αφορά τον τομέα της Βιοπληροφορικής υπάρχουν δύο μέθοδοι που χρησιμοποιούνται ευρέως για τη δημιουργία φυλογενετικών δέντρων. Αυτές είναι: Η UPGMA και η ένωση γειτόνων (neighbor joining).

Η **UPGMA** [12] είναι μια συγκεντρωτική μέθοδος ιεραρχικής ομαδοποίησης κι έχει σαν αποτέλεσμα ένα φυλογενετικό δέντρο. Το προκύπτον φυλογενετικό δέντρο είναι ένα ριζωμένο φυλογενετικό δέντρο με κοινό πρόγονο. Όταν σχεδιάζετε ένα φυλογενετικό δέντρο χρησιμοποιώντας τη μέθοδο UPGMA, οι εξελικτικοί ρυθμοί θεωρούνται ότι είναι οι ίδιοι για όλες τις γενεαλογίες και λαμβάνονται υπόψη οι αποστάσεις ανά ζεύγος για την παραγωγή ενός φυλογενετικού δέντρου. Αρχικά, κάθε είδος είναι μια ομάδα και δύο τέτοιες ομάδες με τη μικρότερη εξελικτική απόσταση σχηματίζουν ένα ζεύγος. Στη συνέχεια, υπολογίζεται η απόσταση του κοινού ζεύγους λαμβάνοντας τον μέσο όρο. Ο αλγόριθμος επαναλαμβάνει τη διαδικασία έως ότου όλα τα είδη συνδέονται σε ένα μόνο σύμπλεγμα.

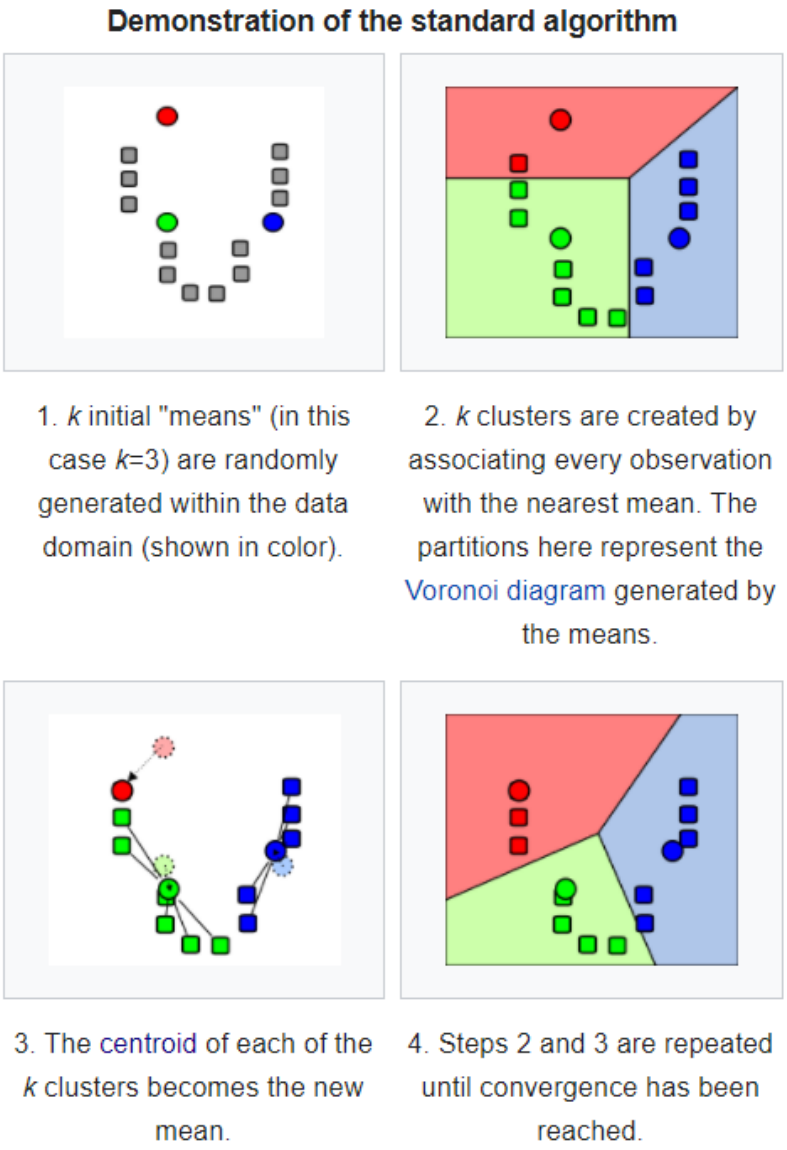
Η **ένωση γειτόνων** (neighbor joining) [13] είναι επίσης μια συγκεντρωτική ιεραρχική μέθοδος για τη δημιουργία φυλογενετικών δέντρων. Συνήθως, χρησιμοποιείται για δέντρα που βασίζονται σε δεδομένα αλληλουχίας DNA ή πρωτεϊνών. Ο αλγόριθμος απαιτεί γνώση της απόστασης μεταξύ κάθε ζεύγους τάξης, για να σχηματιστεί το δέντρο. Αφού υπολογιστεί η απόσταση, επιλέγετε το ζεύγος των γενεαλογιών με τη χαμηλότερη απόσταση, για να ενταχθεί σε έναν νέο κόμβο. Ωστόσο, αυτός ο κόμβος συνδέεται με τον κεντρικό κόμβο. Μετά από αυτό, ο αλγόριθμος υπολογίζει την απόσταση από κάθε γενεαλογία στον νέο κόμβο. Στη συνέχεια,

υπολογίζει την απόσταση κάθε γενεολογίας έξω από αυτό το ζεύγος έως τον νέο κόμβο. Τέλος, αντικαθιστά τους ενωμένους γείτονες με τον νέο κόμβο βάσει των υπολογισμένων αποστάσεων.

1.2.2 Διαμεριστική ομαδοποίηση (Partitional Clustering)

Οι αλγόριθμοι διαμέρισης βασίζονται στον καθορισμό ενός αρχικού αριθμού ομάδων και επαναπροσδιορισμό των στοιχείων μεταξύ των ομάδων. Ουσιαστικά, ο αλγόριθμος προσδιορίζει όλες τις ομάδες ταυτόχρονα, κατασκευάζοντας πολλές διαμερίσεις των δεδομένων αξιολογώντας αυτές σύμφωνα με κάποιο κριτήριο απόστασης. Οι περισσότερες εφαρμογές υιοθετούν μία από τις δύο δημοφιλείς ευρετικές μεθόδους: Τον αλγόριθμο k-means και τον αλγόριθμο k-medoids.

Ο αλγόριθμος ομαδοποίησης ***k-means*** [14] είναι ιδιαίτερα δημοφιλής λόγω της απλότητάς του και της ελαχιστοποίησης των σφαλμάτων ομαδοποίησης. Κάθε κόμβος ανήκει στην ομάδα με το κοντινότερο μέσο (mean), κέντρο ή αλλιώς κεντροειδές της. Ο αλγόριθμος αναθέτει κάθε σημείο (κόμβο) σε μια ομάδα, της οποίας το κέντρο είναι πιο κοντά σε αυτό, όποτε αρχικά θα πρέπει να οριστούν προσεκτικά τα κεντροειδή για την κάθε ομάδα. Το κέντρο είναι ο μέσος όρος όλων των σημείων της ομάδα και οι συντεταγμένες τους είναι ο αριθμητικός μέσος της κάθε διάστασης ξεχωριστά έναντι όλων των σημείων της ομάδας. Διαφορετικές αρχικές θέσεις κεντροειδών επιφέρουν διαφορετικά αποτελέσματα. Ιδανικά, τα κεντροειδή θα πρέπει να έχουν τη μεγαλύτερη δυνατή απόσταση μεταξύ τους. Αρχικά, ο αλγόριθμός k-means επιλέγει έναν αριθμό k όσες και οι ομάδες από τους n κόμβους για τον ρόλο των κέντρων αυτών. Στη συνέχεια, υπολογίζεται η απόσταση για τον κάθε κόμβο από το κάθε κέντρο και «προσκολλάται» στο πιο κοντινό του κέντρο, με αποτέλεσμα να δημιουργούνται k ομάδες. Για την κάθε ομάδα που προκύπτει, χρειάζεται να υπολογιστεί ο μέσος (mean) της κι η διαδικασία επαναλαμβάνεται έχοντας ως κέντρο, το νέο μέσο της κάθε ομάδας έως ότου δεν υπάρχουν άλλες αλλαγές. Ολόκληρη η διαδικασία χρειάζεται να επαναληφθεί από την αρχή έχοντας αυτή τη φορά διαφορετικά τυχαία k κέντρα. Κάθε φορά που μία επανάληψη ολοκληρώνεται, υπολογίζεται η ομοιομορφία (variation) των ομάδων που έχουν προκύψει. Μετά από μερικές επαναλήψεις (μέχρι να παρατηρηθεί σταθερότητα), πλέον ο αλγόριθμος μπορεί να συγκρίνει τις ομάδες της κάθε επανάληψης και να επιλέξει τη βέλτιστη. Συνοπτικά, σε κάθε επανάληψη, τα κέντρα αλλάζουν κι οι κόμβοι ανατίθενται στην ομάδα με το κοντινότερο σε αυτά κέντρο, μέχρις ότου δεν παρατηρηθούν άλλες μεταβολές (Εικόνα 2). Δυστυχώς, ο αλγόριθμος, αν και φέρνει το βέλτιστο τοπικό αποτέλεσμα, δεν εγγυάται ότι είναι και το ολικά βέλτιστο. Το αποτέλεσμα εξαρτάται από την αρχικοποίηση των κέντρων και για αυτό τον λόγο χρειάζεται να «τρέξει» πολλές φορές με διαφορετική αρχική κατάσταση. Ένα ακόμη μειονέκτημα του αλγορίθμου είναι πως υπάρχει πιθανότητα μία συστάδα να είναι κενή, με αποτέλεσμα να μην ανανεωθεί κάποιο κέντρο. Το συγκεκριμένο πρόβλημα είναι γνωστό ως «Το πρόβλημα των απόμακρων στοιχείων» [15].



Εικόνα 2: Η διαδικασία του αλγορίθμου k-means για k=3. Αρχικά έχουμε k = 3 μέσους (οι κόμβοι κόκκινος, πράσινος και μπλε) και οι πιο κοντινοί κόμβοι προσελκύονται, δημιουργώντας τις 3 ομάδες. Τα κεντροειδή της κάθε ομάδας στην τρίτη εικόνα γίνονται οι νέοι μέσοι. Επαναλαμβάνεται η διαδικασία μέχρι να μην παρατηρηθούν αλλαγές.

Στον αλγόριθμο ***k-medoids*** [7] κάθε ομάδα αναπαρίσταται από ένα από τα αντικείμενα που είναι κοντά στο κέντρο της ομάδας. Συγκεκριμένα, επιλέγονται αυθαίρετα αντικείμενα k ως τα αρχικά μεσοειδή (medoids) και επαναλαμβανόμενα αντιστοιχούνται τα αντικείμενα που παραμένουν στις ομάδες με τα κοντινότερα μεσοειδή. Επιλέγεται τυχαία ένα μη μεσοειδές αντικείμενο και υπολογίζεται το συνολικό κόστος S της ανταλλαγής αντικειμένων με το μη μεσοειδές αντικείμενο αυτό. Αν το $S < 0$, τότε αντικαθιστάται το αντικείμενο αυτό με το μη μεσοειδές και δημιουργείται ένα νέο σετ k-medoids. Η διαδικασία αυτή γίνεται μέχρι να μην αλλάξει αυτό το σετ. Η μέθοδος K-medoids είναι πιο δυνατή από τη μέθοδο k-mean παρουσία θορύβου και ακραίων τιμών, επειδή ένα μεσοειδές επηρεάζεται λιγότερο από ακραίες τιμές από ό,τι ο μέσος.

Ο αλγόριθμος ***PAM*** (Partitioning around Medoids) [16] είναι ένας από τους πρώτους αλγορίθμους k-medoids. Ο αλγόριθμος βρίσκει μια ακολουθία από αντικείμενα τα μεσοειδή

(medoids), τα οποία είναι κεντρικά τοποθετημένα σε ομάδες. Αντικείμενα που ορίζονται προσωρινά ως μεσοειδή τοποθετούνται σε ένα σύνολο S επιλεγμένων αντικειμένων. Ο στόχος του αλγορίθμου είναι να ελαχιστοποιήσει τη μέση ανισότητα των αντικειμένων στο κοντινότερο επιλεγμένο αντικείμενο. Ομοίως, μπορεί να ελαχιστοποιηθεί το άθροισμα των διαφορών μεταξύ του αντικειμένου και του πλησιέστερου επιλεγμένου αντικειμένου τους. Αρχικά, μια συλλογή από k επιλεγμένα αντικείμενα σχηματίζει το σύνολο S και στη συνέχεια γίνονται προσπάθειες για τη βελτίωση της ποιότητας της ομαδοποίησης, ανταλλάσσοντας επιλεγμένα αντικείμενα με μη επιλεγμένα.

Ο αλγόριθμος ***CLARA*** [17] αποτελεί μια βελτίωση του αλγορίθμου PAM και αναπτύχθηκε για την αντιμετώπιση των υψηλών απαιτήσεων που επιβάλλονται από τα μεγάλα σύνολα δεδομένων. Αναλυτικότερα, ο αλγόριθμος επιλέγει πολλαπλά τυχαία δείγματα από το σύνολο δεδομένων και εφαρμόζει σε αυτά τον αλγόριθμο PAM επιστρέφοντας την καλύτερη ομαδοποίηση. Ο αλγόριθμος CLARA έχει αποδειχθεί ότι παράγει λύσεις καλής ποιότητας σε λογικό υπολογιστικό χρόνο για μεγάλα σετ δεδομένων. Ωστόσο, είναι λιγότερο αποτελεσματικός, καθώς η αποτελεσματικότητά του εξαρτάται από το μέγεθος του δείγματος και τη μέθοδο δειγματοληψίας και όχι από ολόκληρο το σετ δεδομένων [8].

Για να βελτιωθεί η αποτελεσματικότητα και η επεκτασιμότητα του αλγορίθμου CLARA προτάθηκε ο αλγόριθμος ***CLARANS*** [18]. Σύμφωνα με τον CLARANS, ο αλγόριθμος PAM εφαρμόζεται πάνω σε ολόκληρο το σύνολο δεδομένων, αλλά με μια μικρή τροποποίηση. Σε κάθε επανάληψη, δεν εξετάζονται όλοι οι γείτονες του συνόλου των εσωτερικών αντιπροσώπων (medoids). Αντιθέτως, χρησιμοποιείται μόνο ένα τυχαία επιλεγμένο κλάσμα αυτών. Οι επιλεγμένοι γείτονες εξετάζονται ακολουθιακά κι αν ο τρέχων γείτονας είναι καλύτερος από το σύνολο των εσωτερικών αντιπροσώπων, τότε αντικαθίσταται από αυτόν κι η διαδικασία επαναλαμβάνεται. Εάν με αυτή τη διαδικασία δε βρεθεί μια καλύτερη λύση μετά από έναν αριθμό προσπαθειών, τότε θεωρείται ότι έχει επιτευχθεί το τοπικό βέλτιστο. Έχει αποδειχθεί πειραματικά ότι ο αλγόριθμος CLARANS είναι πιο αποτελεσματικός συγκριτικά με τους αλγορίθμους PAM και CLARA. Επιπλέον, η ποιότητα της ομαδοποίησης εξαρτάται από τις δύο προκαθορισμένες παραμέτρους: Τον μέγιστο αριθμό γειτόνων που εξετάστηκαν και τον αριθμό των τοπικών ελαχίστων που λήφθηκαν.

1.2.3 Ομαδοποίηση βασισμένη στη πυκνότητα (Density-based clustering)

Οι αλγόριθμοι ομαδοποίησης με βάση την πυκνότητα δημιουργήθηκαν για να ανακαλύπτουν αυθαίρετα σχήματα ομάδων. Έτσι, οι περισσότεροι αλγόριθμοι που βασίζονται στην έννοια της πυκνότητας δε θέτουν κανέναν περιορισμό στο σχήμα των ομάδων που θα προκύψουν αναδεικνύοντας ομάδες οποιουδήποτε σχήματος. Σε αυτή την κατηγορία, μια ομάδα θεωρείται ως μια περιοχή στην οποία η πυκνότητα των αντικειμένων υπερβαίνει ένα όριο. Ο αλγόριθμος DBSCAN, DBCLASD και ο αλγόριθμος SSN είναι τρεις τυπικοί αλγόριθμοι αυτής της κατηγορίας. Παρόλο που οι αλγόριθμοι αυτοί μοιράζονται την ίδια φιλοσοφία, διαφέρουν μεταξύ τους ως προς τον τρόπο με τον οποίο ποσοτικοποιούν την έννοια της πυκνότητας.

Ο αλγόριθμος ***DBSCAN*** [7] βασίζεται στην πυκνότητα των ομάδων. Οι ομάδες δημιουργούνται από την πυκνότητα των σημείων. Περιοχές με υψηλή πυκνότητα σημείων απεικονίζουν την ύπαρξη ομάδων, ενώ οι περιοχές με χαμηλή πυκνότητα υποδηλώνουν θόρυβο ή ομάδες ακραίων τιμών. Ο αλγόριθμος DBSCAN είναι κατάλληλος για την αντιμετώπιση μεγάλων συνόλων δεδομένων με θόρυβο και είναι σε θέση να αναγνωρίζει συστάδες με διαφορετικά μεγέθη και σχήματα. Για κάθε σημείο της ομάδας, η γειτονιά μιας

συγκεκριμένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν συγκεκριμένο αριθμό από σημεία, δηλαδή η πυκνότητα στη γειτονιά πρέπει να υπερβαίνει κάποιο προκαθορισμένο όριο. Ο αλγόριθμος χρειάζεται τρεις παραμέτρους τη k που είναι το μέγεθος της γειτονιάς, την Eps που είναι η ακτίνα που οριοθετεί την περιοχή της γειτονιάς ενός σημείου και τη $MinPts$ που είναι ο ελάχιστος αριθμός των σημείων που πρέπει να υπάρχει στη γειτονιά Eps. Για την ομαδοποίηση ενός συνόλου, ο αλγόριθμος DBSCAN αρχικά προσδιορίζει τους πλησιέστερους k γείτονες κάθε σημείου, καθώς επίσης και τον πιο μακρινό k πλησιέστερο γείτονα. Στη συνέχεια, υπολογίζεται ο μέσος όρος αυτής της απόστασης. Επιπλέον, για κάθε σημείο του σετ δεδομένων ο αλγόριθμος προσδιορίζει τα άμεσα προσβάσιμα σημεία πυκνότητας χρησιμοποιώντας το όριο Eps που παρέχεται από τον χρήστη και ομαδοποιεί τα σημεία σε σημεία του πυρήνα και σε σημεία των ορίων. Το κατώφλι Eps είναι η μέση απόσταση που υπολογίστηκε προηγουμένως. Στο τέλος, η ομαδοποίηση επαληθεύεται προκειμένου να ελεγχθεί, εάν υπάρχουν ομάδες που μπορούν να συγχωνευτούν, δηλαδή εάν δύο σημεία διαφορετικών ομάδων βρίσκονται σε απόσταση μικρότερη από το κατώφλι Eps.

Ο αλγόριθμος ***SSN*** [7] είναι και αυτός ένας αλγόριθμος ομαδοποίησης που βασίζεται στην πυκνότητα. Μπορούμε να πούμε ότι διαφέρει από τον αλγόριθμο DBSCAN στο γεγονός ότι προσδιορίζει την ομοιότητα των σημείων, υπολογίζοντας τον αριθμό των κοντινών γειτόνων που τα δύο αυτά σημεία μοιράζονται. Χρησιμοποιώντας αυτή την ομοιότητα, η πυκνότητα μεταξύ των σημείων υπολογίζεται ως το άθροισμα των ομοιοτήτων των κοντινών γειτόνων ενός σημείου. Τα σημεία που έχουν μεγάλη πυκνότητα αποτελούν τα σημεία του πυρήνα, ενώ σημεία με μικρή πυκνότητα αποτελούν σημεία θορύβου. Ο αλγόριθμός SSN χρησιμοποιεί επίσης τις παραμέτρους k, Eps και $MinPts$ και ακολουθεί μια παρόμοια διαδικασία με την παραπάνω για τον σχηματισμό των ομάδων. Πρώτα προσδιορίζονται οι k κοντινότεροι γείτονες όλων των σημείων. Στη συνέχεια, υπολογίζεται η ομοιότητα μεταξύ των ζευγών των σημείων με βάση πόσους κοντινούς γείτονες μοιράζονται τα σημεία αυτά. Χρησιμοποιώντας αυτή την ομοιότητα, η πυκνότητα κάθε σημείου μπορεί να υπολογιστεί ως ο αριθμός των γειτόνων με τους οποίους ο αριθμός των κοινών γειτόνων είναι ίσος ή μεγαλύτερος από το Eps. Τα σημεία ορίζονται ως βασικά σημεία, εάν η πυκνότητα του σημείου είναι ίση ή μεγαλύτερη από το $MinPts$. Επομένως. οι ομάδες αρχίζουν να σχηματίζονται γύρω από αυτά τα βασικά σημεία, αλλά οι ομάδες δεν περιέχουν πάντα όλα τα σημεία, καθώς υπάρχουν σημεία που θεωρούνται θόρυβος.

Ο αλγόριθμος ***DBCLASD*** [19] αποτελεί επέκταση του αλγορίθμου DBSCAN. Ουσιαστικά, είναι ένας σταδιακός αλγόριθμος που υποθέτει ότι τα στοιχεία μιας ομάδας είναι ομοιόμορφα κατανεμημένα και τα σημεία εκτός της ομάδας πιθανόν δεν ικανοποιούν αυτόν τον περιορισμό. Βάσει αυτής της υπόθεσης ο αλγόριθμός προσπαθεί να προσδιορίσει την κατανομή που ικανοποιείται από τις αποστάσεις μεταξύ πλησιέστερων γειτόνων. Αρχικά, δημιουργείται μια ομάδα από ένα στοιχείο στόχο και στη συνέχεια η ανάθεση ενός σημείου σε μια ομάδα γίνεται σταδιακά από το σύνολο των πλησιέστερων γειτόνων που ικανοποιούν την υπόθεση της ομοιόμορφης κατανομής. Ανάμεσα στα πλεονεκτήματα του αλγορίθμου είναι η δυνατότητα του να αναδεικνύει ομάδες αυθαιρέτου σχήματος και ποικίλων πυκνοτήτων.

1.2.4 Ομαδοποίηση βασισμένη στο πλέγμα (Grid-based clustering)

Η ομαδοποίηση βάσει πλέγματος χρησιμοποιείται σε πολυδιάστατα δεδομένα χωρίζοντας τον χώρο σε έναν πεπερασμένο αριθμό κελιών, σχηματίζοντας ένα πλέγμα στο οποίο εκτελούνται όλες οι λειτουργίες της ομαδοποίησης. Η κύρια στρατηγική που υιοθετείται από αυτούς τους

αλγορίθμους είναι: i) ο προσδιορισμός των υποχώρων του χώρου χαρακτηριστικών που ενδεχομένως περιέχουν τις ομάδες, ii) ο προσδιορισμός των ομάδων που βρίσκονται σε καθέναν από τους υποχώρους αυτούς και iii) η περιγραφή των ομάδων που προκύπτουν. Με τον προσδιορισμό των υποχώρων με έναν επαναληπτικό τρόπο από κάτω προς τα πάνω (bottom-up), από τους χώρους χαμηλότερης προς τους χώρους υψηλότερης διάστασης, οι αλγόριθμοι αναζητούν ομάδες σε καθέναν από αυτούς. Παρακάτω περιγράφονται λεπτομερώς οι αλγόριθμοι STING και CLIQUE.

Ο αλγόριθμος ***STING*** [20] χρησιμοποιείται για την εκτέλεση ομαδοποίησης σε χωρικά δεδομένα. Χρησιμοποιώντας μια ιεραρχική τεχνική, γίνεται διαίρεση των χωρικών περιοχών σε ορθογώνια κελιά. Πιο συγκεκριμένα, κάθε κελί σε ένα υψηλό επίπεδο διαιρείται σε μικρότερα, δηλαδή στο επόμενο κατώτερο επίπεδο. Ο αλγόριθμός STING ξεκινάει από ένα προεπιλεγμένο επίπεδο με μικρό αριθμό κελιών και για κάθε κελί στο τρέχον επίπεδο υπολογίζει το διάστημα εμπιστοσύνης, το οποίο αντικατοπτρίζει τη σχέση του κελιού με το δεδομένο. Στη συνέχεια, ο αλγόριθμος μετακινεί τα μη σχετιζόμενα κελιά κι όταν η εξέταση του τρέχοντος επιπέδου ολοκληρωθεί, προχωράει στο επόμενο κατώτερο επίπεδο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι το κατώτατο επίπεδο.

Ο αλγόριθμος ***CLIQUE*** [21] βρίσκει αυτόματα υποδιαστήματα με ομάδες υψηλής πυκνότητας. Παράγει πανομοιότυπα αποτελέσματα ανεξάρτητα από τη σειρά με την οποία παρουσιάζονται οι εγγραφές εισόδου και δεν κανονικοποιεί τα δεδομένα εισόδου. Ο αλγόριθμος CLIQUE κλιμακώνεται γραμμικά με τον αριθμό των εγγραφών εισόδου και έχει καλή επεκτασιμότητα όσο ο αριθμός των διαστάσεων των δεδομένων ή η υψηλότερη διάσταση στην οποία ενσωματώνονται οι ομάδες αυξάνεται.

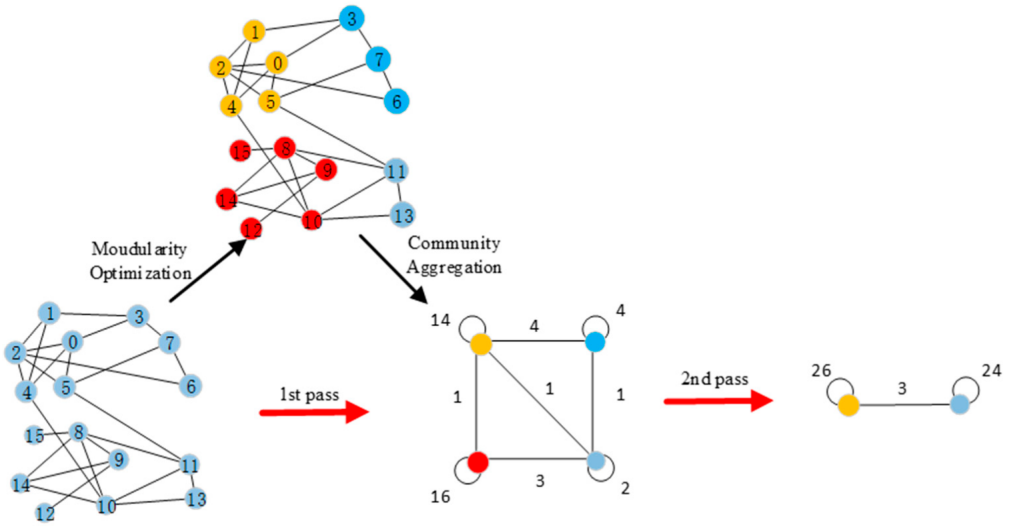
1.2.5 Ομαδοποίηση δικτύων

Για την καλύτερη κατανόηση της ομαδοποίησης με βάσει γραφημάτων και των αλγορίθμων που περιγράφονται στη συνέχεια, αξίζει να αποσαφηνιστεί η έννοια του γράφου και του τυχαίου περίπατου. Οι γράφοι όπως ορίζονται στη γλώσσα των Μαθηματικών ή αλλιώς τα δίκτυα, είναι ένας τρόπος απεικόνισης των σχέσεων μεταξύ διαφόρων οντοτήτων. Κατά κύριο λόγο, ένας γράφος αποτελείται από δύο βασικά συστατικά: Τους κόμβους και τις ακμές οι οποίες συνδέουν ένα πλήθος από κόμβους μεταξύ τους. Οι γράφοι συμβολίζονται ως εξής: G=(V,E) όπου το G προέρχεται από τη λέξη graph (γράφος), V (vertices) που αντιπροσωπεύει τους κόμβους και E (edges) που αντιπροσωπεύει τις ακμές. Ένας τυχαίος περίπατος είναι μια ακολουθία κόμβων, οπού ξεκινώντας από έναν κόμβο, οι επόμενοι επιλέγονται τυχαία από τους γειτονικούς κόμβους του κάθε κόμβου.

Όσον αφορά την ομαδοποίηση δικτύων, η ομαδοποίηση με βάση γραφήματα μπορεί να εκμεταλλευτεί τη τοπολογία ενός δικτύου για τον εντοπισμό ομάδων. Η ομαδοποίηση γράφων βασίζεται στην ομαδοποίηση των κορυφών του γραφήματος σε ομάδες λαμβάνοντας υπόψη την δομή των ακμών του με τέτοιο τρόπο ώστε να υπάρχουν πολλές ακμές μέσα στις ομάδες και λίγες μεταξύ αυτών. Αξίζει να αναφερθεί ότι οι αλγόριθμοι ομαδοποίησης δικτύων συμβάλουν στην ανάλυση των βιολογικών δικτύων βοηθώντας στην ανάδειξη των λειτουργικών ενοτήτων και πληροφοριών σχετικά με τη κυτταρική οργάνωση. Καθώς, γίνονται όλο και περισσότερες προσπάθειες για την ανάλυση των βιολογικών δικτύων ώστε να αναδειχθούν οι πληροφορίες που αναφέραμε, οι αλγόριθμοι αποτελούν σημαντικό εργαλείο για την ανάλυση αυτή [22]. Οι πιο γνωστοί αλγόριθμοι ομαδοποίησης γραφημάτων είναι οι

αλγόριθμοι: Louvain, SPICi, MCL/HipMCL και Walktrap, των οποίων τα αποτελέσματα των ομαδοποιήσεων θα συγκριθούν με τη βοήθεια του VICTOR στη συνέχεια.

Ο αλγόριθμος ***Louvain*** [23] χρησιμοποιείται για τον εντοπισμό κοινοτήτων σε μεγάλα δίκτυα και προϋποθέτει το δίκτυο να χωριστεί σε ομάδες πυκνά συνδεδεμένων κόμβων, όπου οι κόμβοι να ανήκουν σε διαφορετικές ομάδες και να είναι αραιά συνδεδεμένοι. Ο αλγόριθμος σε γρήγορο χρονικό διάστημα σχηματίζει ομάδες επιτυγχάνοντας υψηλή τιμή αρθρωτότητας (modularity), ενώ είναι αποδοτικός για δίκτυα με εκατομμύρια κόμβους. Η αρθρωτότητα μιας ομάδας είναι η κλιμακούμενη τιμή που κυμαίνεται από -1 έως 1 και μετράει την πυκνότητα των συνδέσμων εντός της ομάδας σε σύγκριση με τους συνδέσμους μεταξύ των ομάδων. Η διαδικασία ομαδοποίησης χωρίζεται σε δύο φάσεις οι οποίες επαναλαμβάνονται (Εικόνα 3). Αρχικά, για ένα δίκτυο n κόμβων κάθε κόμβος αποτελεί μία διαφορετική κοινότητα. Στη συνέχεια, κάθε κόμβος μετακινείται ιδεατά σε κάθε κοινότητα των γειτόνων του. Για κάθε μετακίνηση υπολογίζεται η νέα τιμή modularity και ο κόμβος μετακινείται στην κοινότητα που αποδίδει την μεγαλύτερη τιμή. Η διαδικασία αυτή επαναλαμβάνεται για κάθε κόμβο και μέχρι να μην παρατηρηθεί κάποια επιπλέον βελτίωση. Ένας κόμβος μπορεί να εξεταστεί πάνω από μία φορά και η σειρά εξέτασης των κόμβων επηρεάζει το αποτέλεσμα του αλγορίθμου. Στη δεύτερη φάση δημιουργείται ένα νέο δίκτυο, οι κόμβοι του οποίου ορίζονται από τις κοινότητες της πρώτης φάσης. Στο νέο δίκτυο, εφόσον είναι σταθμισμένο, οι νέες ακμές δίνονται από το άθροισμα των βαρών των συνδέσμων μεταξύ των κόμβων των κοινοτήτων. Οι ακμές μεταξύ των κόμβων της ίδιας ομάδας μπορούν να οδηγήσουν σε αυτο-επαναλαμβανόμενους βρόγχους για τη συγκεκριμένη ομάδα. Οι δύο φάσεις του αλγόριθμου μπορούν να επαναληφθούν κάθε φορά για το νέο δίκτυο που προκύπτει μέχρι να μην εμφανίζεται αλλαγή στο αποτέλεσμα του αλγορίθμου.

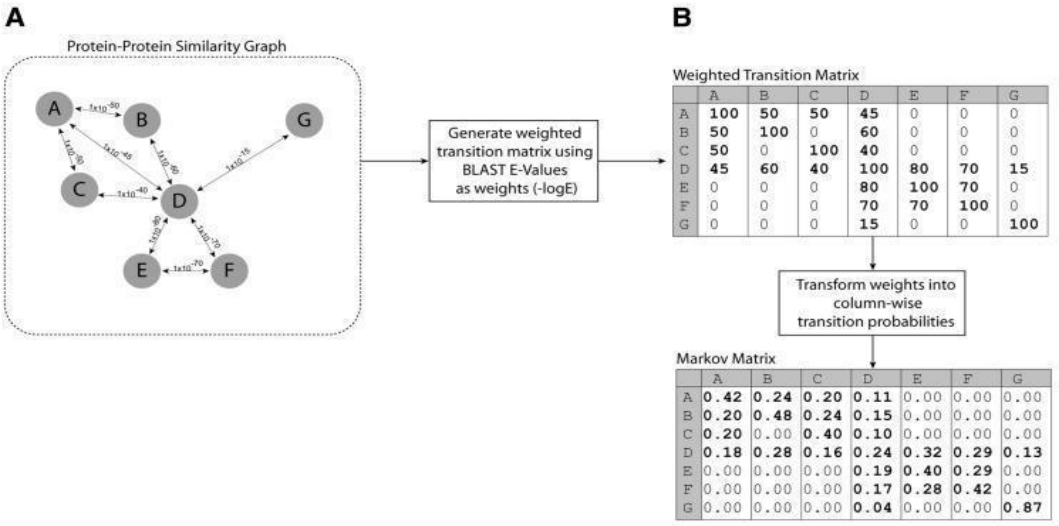


Εικόνα 3: Διαδικασία αλγορίθμου Louvain [24].

Ο αλγόριθμος ***SPICi*** [22] εφαρμόζεται σε βιολογικά δίκτυα μεγάλου μεγέθους όπως λειτουργικά δίκτυα ευκαρυωτικών οργανισμών υψηλότερου επιπέδου. Ο αλγόριθμός εντοπίζει σημεία του γράφου με υψηλή συνδεσιμότητα με βάση την τοπική τους πυκνότητα, χρησιμοποιώντας μια απλή προσέγγιση επέκτασης ομάδων ξεκινώντας από τοπικούς κόμβους που έχουν υψηλό σταθμισμένο βαθμό και προσθέτει κόμβους που διατηρούν τη

πυκνότητα της ομάδας. Λαμβάνοντας υπόψη ένα σταθμισμένο δίκτυο, ο στόχος του αλγορίθμου είναι να παράγει ένα σύνολο πυκνών υπογραφών. Στον αλγόριθμό SPICi για κάθε κόμβο υπολογίζεται ο σταθμισμένος βαθμός ως το άθροισμα όλων των τιμών των προσπιπτόντων ακμών του. Επίσης, για κάθε σύνολο κόμβων υπολογίζεται η πυκνότητα ως το άθροισμα των βαρών των άκρων μεταξύ τους, διαιρούμενο με τον συνολικό αριθμό πιθανών ακμών. Σε κάθε βήμα του αλγορίθμου έχουμε ένα τρέχον σύνολο κόμβων για το σύμπλεγμα, το οποίο αρχικά αποτελείται από τους δύο πρώτους κόμβους. Το σύνολο αυτό επεκτείνεται με γειτονικούς κόμβους που δεν επηρεάζουν την πυκνότητα του. Ένα από τα πλεονεκτήματα του αλγόριθμου SPICi είναι ότι είναι αποτελεσματικότερος από άλλες προσεγγίσεις δικτύων (π.χ.: DPClus, CFinder κ.α.). Επιπλέον, άλλο ένα πλεονεκτήματα του αλγορίθμου είναι ότι είναι ο μόνος, ο οποίος έχει την δυνατότητα να ομαδοποιεί όλα τα δίκτυα εντός ενός λογικού χρονικού διαστήματος. Παρά το γεγονός ότι ο SPICi είναι γρηγορότερος, οι ομάδες των βιολογικών δικτύων που φέρνει ως αποτέλεσμα ανακεφαλαιώνουν τις λειτουργικές μονάδες τους. Ωστόσο, η απόδοση στην ανακεφαλαιοποίηση των πρωτεϊνικών συμπλεγμάτων είναι πολύ καλή και μειώνεται μόνο σε εξαιρετικά ελλιπή δίκτυα. Τέλος, οι ομάδες που προκύπτουν από τον αλγόριθμο SPICi αποδείχθηκε ότι είναι ίδιας ποιότητας με αυτές που βρέθηκαν από αλγόριθμους τελευταίας τεχνολογίας.

Ο αλγόριθμος ***MCL*** [25] έχει σχεδιαστεί να προσπελαύνει απλούς ή σταθμισμένους γράφους. Ο αλγόριθμος MCL μπορεί να χρησιμοποιηθεί και στον τομέα της βιολογίας και συγκεκριμένα για την ομαδοποίηση βιολογικών ακολουθιών, καθώς μπορεί να αναπαραστήσει σχέσεις ομοιότητας ανάμεσα στις βιολογικές ακολουθίες. Ο αλγόριθμός εντοπίζει στο γράφο φυσικές ομάδες, που περιλαμβάνουν πολλές ακμές ανάμεσα στους κόμβους τους, μέσω τυχαίων “περιπάτων” ή προσπελάσεων. Ο αλγόριθμός MCL χρησιμοποιεί στοχαστικούς πίνακες κατά στήλη ή αλλιώς πίνακες Markov που αναπαριστούν την έννοια των τυχαίων περιπάτων μέσα σε ένα γράφο (Εικόνα 4). Βασικό πλεονέκτημα του MCL είναι ότι αποτελεί έναν γρήγορο και εύκολα επεκτάσιμο αλγόριθμο. Αναλυτικότερα, ο αλγόριθμος MCL ακολουθεί μια bootstrapping ροή και υπολογίζει τις τυχαίες μετακινήσεις ενός γράφου εισόδου, ο οποίος αντιπροσωπεύεται από έναν Μαρκοβιανό πίνακα. Έπειτα, ο τελεστής επέκτασης που τετραγωνίζει τον πίνακα εναλλάσσεται, χρησιμοποιώντας το γινόμενο του πίνακα, με τον τελεστή πληθωρισμού. Ο πληθωρισμός πραγματοποιείται αυξάνοντας κάθε είσοδο του πίνακα σε μια δεδομένη δύναμη και αναδημιουργώντας τον πίνακα ώστε να γίνει και πάλι στοχαστικός. Όλη η διαδικασία εναλλαγής συνεχίζεται μέχρι να επιτευχθεί μια κατάσταση ισορροπίας.



Εικόνα 4: Α) Δίκτυο αλληλεπίδρασης πρωτεΐνης- πρωτεΐνης, οι κόμβοι αντιπροσωπεύουν τις εφτά πρωτεΐνες και οι ακμές τις BLASTp ομοιότητες. Β) Σταθμισμένος πίνακας μετάβασης και σχετική στήλη στοχαστικού μαρκοβιανού πίνακα για τις εφτά πρωτεΐνες του δικτύου [25].

Ο αλγόριθμος ***HipMCL*** [26] αποτελεί έναν παράλληλο αλγόριθμο υψηλής απόδοσης για ομαδοποίηση δικτύων μεγάλης κλίμακας. Ουσιαστικά, ο αλγόριθμος HipMCL αποτελεί μια παραλλαγή του αρχικού αλγορίθμου MCL. Παρά την αποτελεσματικότητα του MCL, η επεκτασιμότητα αποτελεί πρόβλημα, καθώς δεν μπορεί να διαχειριστεί μεγάλα δίκτυα σε προσιτό χρόνο λειτουργίας. Ο αλγόριθμος HipMCL ξεπερνά όλες αυτές τις προκλήσεις αναπτύσσοντας μαζικά παράλληλους αλγόριθμους για όλα τα κομμάτια του MCL.

Ο αλγόριθμος ***Walktrap*** [27] στηρίζεται επίσης στην ιδέα των τυχαίων περιπάτων για την ανίχνευση ομάδων σε μεγάλα δίκτυα. Αρχικά, για την κατάταξη των κόμβων σε ομάδες ορίζεται η απόσταση, η οποία παίρνει μεγάλες τιμές για δύο κόμβους που βρίσκονται σε διαφορετικές ομάδες και μικρές τιμές για κόμβους που βρίσκονται στην ίδια ομάδα. Ουσιαστικά, μέσω αυτής της απόστασης υπολογίζει και συγκρίνει την ομοιότητα των κόμβων. Έπειτα, υπολογίζει την πιθανότητα των τυχαίων περιπάτων και την απόσταση μεταξύ κόμβων και τέλος καταλήγει στην τελική απόσταση. Με βάση την απόσταση που υπολογίστηκε και τη χρήση ιεραρχικών αλγορίθμων ομαδοποίησης, μπορούν να ληφθούν διαφορετικής κλίμακας ομάδες, οι οποίες μπορούν να αναπαρασταθούν ως δέντρα, γνωστά ως δενδρογράμματα. Για την απλοποίηση του αλγόριθμου, οι γειτονικές κοινότητες χωρίζονται και υπολογίζεται συνεχώς το τετράγωνο της απόστασης ώστε αυτό να μειώνεται. Τέλος, τα αποτελέσματα αξιολογούνται με την βοήθεια της αρθρωτότητας (modularity).

Ο αλγόριθμος ***Label propagation*** [28] είναι αποδοτικός για την εύρεση σταθερών κοινοτήτων σε ένα γράφημα χρησιμοποιώντας μια επαναληπτική διαδικασία. Η μέθοδος ξεκινά δίνοντας σε κάθε κόμβο στο γράφημα μια μοναδική ετικέτα (label). Στη συνέχεια, επαναλαμβάνεται η διαδικασία κατά την οποία κάθε κόμβος υιοθετεί την πιο κοινή ετικέτα των γειτόνων του. Αυτή η διαδικασία επαναλαμβάνεται έως ότου η ετικέτα κάθε κόμβου στο γράφημα να είναι ίδια με την ετικέτα μέγιστης εμφάνισης μεταξύ των γειτόνων του. Όταν τελειώσει η διαδικασία οι κόμβοι που εμφανίζουν την ίδια ετικέτα αποτελούν και μια ομάδα. Εκτός από την εύκολη εφαρμογή και τη γρήγορη εκτέλεση του, άλλο ένα πλεονέκτημα είναι ότι ο αλγόριθμος στηρίζεται μόνο στη δομή του δικτύου και δεν χρησιμοποιεί κάποια άλλη παράμετρο ούτε βελτιστοποίηση της

αντικειμενικής λειτουργίας. Κάθε κόμβος λαμβάνει τη δική του απόφαση σχετικά με την ομάδα στην οποία ανήκει βάσει των ομάδων των άμεσων γειτόνων του, χωρίζοντας έτσι το δίκτυο σε ομάδες.

2. Βιολογικά δίκτυα

Τα **δίκτυα** ή αλλιώς **γράφοι** όπως ονομάζονται στη γλώσσα των μαθηματικών, είναι ένας τρόπος απεικόνισης των σχέσεων μεταξύ διαφόρων οντοτήτων. Κατά κύριο λόγο, ένα δίκτυο αποτελείται από δύο βασικά συστατικά: τους κόμβους και τις ακμές οι οποίες συνδέουν το πλήθος των κόμβων μεταξύ τους. Τυπικά συμβολίζουμε ένα δίκτυο ως το γράφο G που ορίζεται από το σύνολο των κόμβων V και των ακμών του E, $G\ (V, E)$. Ανάλογα με το είδος του δικτύου, η μορφή του μπορεί να ποικίλει σε ότι αφορά τα είδη των ακμών και τους περιορισμούς στη διάρθρωση των κόμβων. Πιο συγκεκριμένα, στο τομέα της Βιολογίας, οι αλληλεπιδράσεις που λαμβάνουν χώρα στο κύτταρο μπορούν να οριστούν ως δίκτυα που ονομάζονται βιολογικά δίκτυα. Βιολογικά δίκτυα συναντάμε ωστόσο σε όλα τα επίπεδα μελέτης των επιστημών της ζωής. Στη συνέχεια θα αναφέρουμε κάποιες από τις ιδιότητες των γράφων που είναι σημαντικές και για τα βιολογικά δίκτυα και που θα αναφερθούν παρακάτω.
Είδη δικτύων:

- ***Κατευθυνόμενα δίκτυα (directed)***: Οι ακμές συμβολίζονται πλέον ως βέλη των οποίων η κατεύθυνση υποδηλώνει μη-συμμετρικές σχέσεις μεταξύ κόμβων. Ουσιαστικά, σε έναν κατευθυνόμενο δίκτυο οι ακμές μεταξύ των κόμβων είναι βέλη τα οποία δείχνουν την κατεύθυνση του δικτύου.
- ***Μη Κατευθυνόμενος δίκτυα (undirected)***: Ένας δίκτυο ονομάζεται μη κατευθυνόμενο, αν τα ζεύγη των κορυφών που ορίζουν τις ακμές του, στερούνται διάταξη. Οι ακμές των δικτύων αυτών είναι συμμετρικές.
- ***Σταθμισμένο δίκτυο (weighted network):*** Στην περίπτωση του δικτύου με βάρη, η κάθε ακμή έχει ένα συντελεστή βαρύτητας ο οποίος σηματοδοτεί τη σημαντικότητα της σύνδεσης.

2.1 Δίκτυα αλληλεπίδρασης πρωτεϊνών (Protein-Protein interactions networks (PPIs))

Οι πρωτεΐνες σε ένα κύτταρο ή έναν ζωντανό οργανισμό δεν δρουν αυτόνομα, αλλά συνεργατικά, επικοινωνώντας μεταξύ τους και βοηθώντας η μία την άλλη μέσω των μεταξύ τους αλληλεπιδράσεων. Οι πρωτεϊνικές αλληλεπιδράσεις χαρακτηρίζονται από φυσικές επαφές υψηλής εξειδίκευσης που δημιουργούνται μεταξύ δύο ή περισσοτέρων μορίων πρωτεϊνών ως αποτέλεσμα βιοχημικών συμβάντων. Για αυτό τον λόγο είναι σημαντική η χαρτογράφηση των μεταξύ τους αλληλεπιδράσεων, η οποία επιτυγχάνεται με τη χρήση και την δημιουργία των δικτύων πρωτεϊνικών αλληλεπιδράσεων. Τα δίκτυα αλληλεπίδρασης πρωτεϊνών είναι πρακτικά μέσα για την απόκτηση βασικών γνώσεων και τη βελτίωση βιολογικών και βιοϊατρικών εφαρμογών [29]. Τα δίκτυα αυτά αναπαριστούν τις αλληλεπιδράσεις μεταξύ των πρωτεϊνών και αποκρυπτογραφούν τις σχέσεις μεταξύ της δομής και της λειτουργίας του δικτύου, ανακαλύπτουν νέες πρωτεϊνικές λειτουργίες, εντοπίζοντας λειτουργικά συνεκτικές ενότητες και διατηρημένα μοριακά μοτίβα αλληλεπίδρασης. Αναλυτικότερα, τα δίκτυα αλληλεπίδρασης πρωτεϊνών μπορούν να χρησιμοποιηθούν για τον εντοπισμό μοριακών και κυτταρικών μηχανισμών που ελέγχουν υγιείς και νοσούντες καταστάσεις του οργανισμού, καθώς και στο σχεδιασμό και την ανακάλυψη νέων φαρμάκων. Κάθε κόμβος του δικτύου αναπαριστά μια πρωτεΐνη και κάθε ακμή αποτελεί μια αλληλεπίδραση μεταξύ των πρωτεϊνών κόμβων. Το δίκτυο αλληλεπίδρασης

πρωτεϊνών έχει υψηλό βαθμό ετερογένειας, με ένα μικρό αριθμό πρωτεϊνών που χαρακτηρίζονται από μεγάλη συνδεσιμότητα και αναφέρονται ως κεντρικοί κόμβοι, σε αντίθεση με την πλειοψηφία των πρωτεϊνών που χαρακτηρίζονται από λίγες συνδέσεις. Οι υψηλά συνδεδεμένες πρωτεΐνες, έχει παρατηρηθεί ότι σχετίζονται με βασικά λειτουργικά γονίδια και διακρίνονται σε εκείνες που έχουν σταθερούς δεσμούς πρόσδεσης και σε εκείνες που η πρόσδεσή τους πραγματοποιείται σε διαφορετική χρονική στιγμή ή τοποθεσία.

Βάσεις δεδομένων: Yeast Proteome Database (YPD) [30], Munich Information Center for Protein Sequences (MIPS) [31], Molecular Interaction (MINT) [32], Database of Interacting Proteins (DIP) [33], BioGRID [34] και Human Protein Reference Database (HPRD) [35]

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

Τα δίκτυα ομοιότητας αλληλουχιών επιτρέπουν την οπτικοποίηση σχέσεων μεταξύ πρωτεϊνικών αλληλουχιών ή αλληλουχιών γονιδίων. Ουσιαστικά, το δίκτυο ομοιότητας αλληλουχιών είναι ένας γράφος, του οποίου οι κόμβοι αντιπροσωπεύουν αλληλουχίες νουκλεοτιδίων ή αμινοξέων ορισμένων ειδών, και οι ακμές χρησιμοποιούνται για τη σύνδεση των κόμβων που αντιστοιχούν στις αλληλουχίες με τις υψηλότερες ομοιότητες. Συνήθως, οι περισσότερο σχετικές πρωτεΐνες ή γονίδια ομαδοποιούνται σε ομάδες και υποθέτουμε ότι κάθε κόμβος του δικτύου ανήκει σε μία ομάδα, αλλά μπορεί επίσης να εξεταστεί η περίπτωση που ένας κόμβος ανήκει σε διαφορετικές ομάδες. Συγκεκριμένα, οι συνδέσεις μεταξύ δύο κόμβων σημαίνουν ότι οι αλληλεπιδράσεις τους έχουν ποσοστό ομοιότητας μεγαλύτερο ή ίσο από μία συγκεκριμένη τιμή που έχει καθοριστεί από το χρήστη. Με αυτό τον τρόπο προκύπτει ένα δίκτυο με βάρη (σταθμισμένο δίκτυο), που αναπαριστά πιθανές λειτουργικές συσχετίσεις μεταξύ των βιομορίων. Ωστόσο, μπορούμε να συναντήσουμε και μη σταθμισμένα δίκτυα ομοιότητας αλληλουχιών. Τα δίκτυα ομοιότητας αλληλουχιών μπορούμε να πούμε ότι είναι ένας αποτελεσματικός τρόπος για την οπτικοποίηση και ανάλυση σχέσεων ομοιότητας μεταξύ μελών υπεροικογένειας [36]. Επιπλέον, η ομαδοποίηση δικτύων ομοιότητας αλληλουχιών βοηθά στον εντοπισμό πρωτεϊνικών οικογενειών, όπου οι πρωτεΐνες έχουν παρόμοιες λειτουργίες ή συμμετέχουν σε βιολογικές διεργασίες.

Τα δίκτυα ομοιότητας αλληλουχιών δεν προσφέρουν μόνο έναν διαφορετικό τρόπο αναπαράστασης της ομοιότητας μεταξύ ειδών και ακολουθιών, αλλά παρέχουν επίσης έναν αριθμό εργαλείων για την ανάλυση μεταγονιδιωματικών δεδομένων [37]. Κάποια από τα χαρακτηριστικά του δικτύου ομοιότητας αλληλουχιών είναι ότι πρόκειται για μη κλιμακούμενο δίκτυο (scale-free), συνήθως αραιό και συχνά δημιουργεί κεντρικούς κόμβους (hubs).

Τα πιο γνωστά εργαλεία για τον προσδιορισμό της ομοιότητας αλληλουχιών είναι: BLAST [38] και FASTA3 suite [39].

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

Τα ρυθμιστικά δίκτυα γονιδίων είναι ο μηχανισμός που επιτρέπει στα κύτταρα να αντιδρούν σε περιβαλλοντικές αλλαγές όπως η διαθεσιμότητα ενός νέου θρεπτικού συστατικού ή η λήψη ενός (χημικό) σήματος από άλλα κύτταρα. Ουσιαστικά, τα ρυθμιστικά δίκτυα γονιδίων αναπαριστούν τις αλληλεπιδράσεις μεταξύ γονιδίων που ρυθμίζουν την ενεργοποίηση συγκεκριμένων λειτουργιών των κυττάρων. Σε αυτό το είδος δικτύου, οι κόμβοι αναπαριστούν γονίδια και οι ακμές αντιπροσωπεύουν την επιρροή που έχει κάθε γονίδιο πάνω στο άλλο [40].

Με αυτόν τον τρόπο γίνεται δυνατός ο έλεγχος των επιπέδων γονιδιακής έκφρασης του mRNA και των πρωτεϊνών. Τα ρυθμιστικά δίκτυα γονιδίων αποτελούν μια απλοποίηση μιας περίπλοκης ποσοτικής διαδικασία που περιλαμβάνει RNA, σύνθεση πρωτεϊνών και χημικές αλληλεπιδράσεις. Με τη βοήθεια των ρυθμιστικών δικτύων γονιδίων μπορούμε να εξορύξουμε αμέσως τις ρυθμίσεις μεταξύ των γονιδίων και να αποκτήσουμε βαθιές γνώσεις για διάφορες βιολογικές διεργασίες, οι οποίες θα μπορούσαν να διευκολύνουν περαιτέρω ιατρικούς τομείς όπως ο σχεδιασμός φαρμάκων ή η ανακάλυψη στόχων φαρμάκων [41]. Τα ρυθμιστικά δίκτυα γονιδίων αποτελούν κατευθυνόμενους γράφους, των οποίων οι κόμβοι δηλώνουν γονίδια ή παράγοντες μεταγραφής και οι ακμές δηλώνουν τις σχέσεις μεταξύ τους. Στα δίκτυα αυτά είναι ελάχιστοι οι κεντρικοί κόμβοι που έχουν υψηλό βαθμό συνδεσιμότητας. Για την κατασκευή τους, χρησιμοποιούνται δεδομένα από πειράματα γονιδιωματικής, μεταγραφομικής, πρωτεομικής και μεταβολομικής. Με τη χρήση κατάλληλων υπολογιστικών εργαλείων/προγραμμάτων επιτυγχάνεται η διαχείριση, η μελέτη, η ανάλυση και η οπτικοποίηση των διαφορετικών τύπων πληροφοριών που διαθέτουν. Τα υπολογιστικά μοντέλα που χρησιμοποιούνται στην ανάλυση των ρυθμιστικών δικτύων μπορούν να διακριθούν σε τρεις κατηγορίες: στα λογικά, στα συνεχή και στα στοχαστικά μοντέλα.
Βάσεις δεδομένων: KEGG [42], GTRD [43] και TRANSFAC [44]

2.4 Δίκτυα μεταγωγής σήματος (Signal transduction networks)

Τα δίκτυα μεταγωγής σήματος αναπαριστούν την μοριακή σηματοδότηση, δηλαδή μία σειρά βιοχημικών διαδικασιών που λαμβάνουν χώρα είτε μέσα στο μόριο είτε από το εξωτερικό στο εσωτερικό του περιβάλλον. Τα δίκτυα αυτά αναπαρίστανται σαν κατευθυνόμενοι γράφοι, όπου οι κόμβοι δηλώνουν στοιχεία σηματοδότησης και οι ακμές αντιπροσωπεύουν την κατεύθυνση της ροής πληροφοριών. Συνήθως, οι κόμβοι εισόδου αντιπροσωπεύουν τους συνδέτες ή τους υποδοχείς τους, οι ενδιάμεσοι κόμβοι αποτελούνται από διάφορες κινάσες και δεύτερους αγγελιοφόρους και οι κόμβοι εξόδου αντιπροσωπεύουν παράγοντες μεταγραφής, κανάλια, κυτταροσκελετό, συστατικά κινητικότητας ή κυτταρικές αποκρίσεις [45]. Οι ιδιότητες των δικτύων μεταγωγής σήματος καθορίζονται από τη τοπολογία του δικτύου. Αυτά τα δίκτυα είναι κυρίως κατευθυνόμενα και αραιά.
Βάσεις δεδομένων: MiST [46], TRANSPATH [47] και KEGG [42]

2.5 Μεταβολικά δίκτυα (Metabolic networks)

Ο μεταβολισμός είναι το άθροισμα των βιοχημικών αντιδράσεων που καταλύονται από ένζυμα σε ένα κύτταρο. Όλες αυτές οι αντιδράσεις σχηματίζουν “μεταβολικές οδούς” που συνδέονται μεταξύ τους και έτσι διαμορφώνουν ένα μεταβολικό δίκτυο. Ένα μεταβολικό δίκτυο είναι ένα πλήρες σύνολο μεταβολικών και φυσικών διεργασιών που καθορίζουν τις φυσικές και βιοχημικές λειτουργίες ενός κυττάρου. Τα μεταβολικά δίκτυα αναπαριστούν τις χημικές αντιδράσεις του μεταβολισμού, τις μεταβολικές οδούς, καθώς και τις ρυθμιστικές αλληλεπιδράσεις που καθοδηγούν αυτές τις αντιδράσεις. Τα μεταβολικά δίκτυα κατασκευάζονται από την ακολουθία του γονιδιώματος ενός οργανισμού και τα γραφήματα μπορούν να χρησιμοποιηθούν για τη μελέτη ροών μέσω των αντιδράσεων, ή για να συσχετιστεί η δομή του γραφήματος με περιβαλλοντικά χαρακτηριστικά και φαινότυπους [48]. Δεδομένου ότι τέτοια δίκτυα είναι κατευθυνόμενα γράφημα, τα στοιχεία της θεωρίας γραφημάτων μπορούν να εφαρμοστούν για τη μελέτη των ιδιοτήτων του. Όμως, παρόλο που

τα μεταβολικά δίκτυα είναι φυσικά κατευθυνόμενα γραφήματα, αναλύονται συνήθως ως μη κατευθυνόμενα γραφήματα. Τα μεταβολικά δίκτυα ακολουθούν την τοπολογία των μη κλιμακούμενων δικτύων (scale-free networks) και οι μεταβολίτες και τα ένζυμα αποτελούν τους κόμβους και οι ακμές αναπαριστούν τις διάφορες αντιδράσεις. Ένας μικρός αριθμός από μεταβολίτες λειτουργούν ως κεντρικοί κόμβο με υψηλή συνδεσιμότητα και συμμετοχή σε πολλές αντιδράσεις. Υπάρχουν πολλοί τρόποι αναπαράστασης των μεταβολικών δικτύων με τη μορφή γράφων. Οι πιο κοινοί είναι:

- Το "δίκτυο αντίδρασης", στο οποίο κάθε κόμβος αναπαριστά μια αντίδραση και οι ακμές ενώνουν τα προϊόντα μιας αντίδρασης με τα αντιδρώντα της άλλης.
- Το "δίκτυο υποστρώματος", όπου οι κόμβοι αντιπροσωπεύουν τους μεταβολίτες και οι ακμές τις μεταβολικές οδούς [49].
- Το "διμερές δίκτυο", το οποίο περιλαμβάνει δύο κατηγορίες κόμβων, μια για τους μεταβολίτες και μια για τις αντιδράσεις. Οι ακμές ενώνουν αντιδράσεις με μεταβολίτες για τα προϊόντα και μεταβολίτες με αντιδράσεις για τα υποστρώματα [50].

Βάσεις δεδομένων: KEGG [42], TRANSPATH [47], EcoCyc [51] και metaTIGER [52]

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

Τα γονίδια είναι λειτουργικές μονάδες των γενετικών υλικών. Πιστεύετε ότι είτε ένα γονίδιο εκφράζεται είτε όχι επηρεάζει τη σύνθεση των πρωτεϊνών. Τα γονίδια δεν λειτουργούν μόνα τους, αλλά αλληλεπιδρούν μεταξύ τους και επηρεάζουν από κοινού την ανθρώπινη υγεία. Μελέτες έχουν δείξει ότι κάθε γονίδιο εκτιμάται ότι κατά μέσο όρο αλληλεπιδρά με τέσσερα έως οκτώ άλλα γονίδια [53]. Τα δίκτυα γονιδίων παρέχουν τη δυνατότητα αναγνώρισης εκατοντάδων γονιδίων που σχετίζονται με σύνθετες ανθρώπινες ασθένειες και που θα μπορούσαν να χρησιμεύσουν ως σημεία για θεραπευτικές παρεμβάσεις και αυτές οι πληροφορίες είναι σημαντικές για την πρόβλεψη των λειτουργιών των νέων γονιδίων και της εύρεσης γονιδίων που παίζουν βασικούς ρόλους στις πολύπλοκες ανθρώπινες ασθένειες. Η κατασκευή ενός δικτύου συν-έκφρασης γονιδίων (GCN) είναι ένας αποτελεσματικός τρόπος για να χαρακτηριστούν τα συσχετιζόμενα πρότυπα μεταξύ των γονιδίων. Γενικότερα, τα δίκτυα γονιδιακής συν-έκφρασης είναι ένας τρόπος αναπαραστάσεις των δεδομένων από πειράματα γονιδιακής έκφρασης γονιδίων ή RNAseq. Τα δίκτυα αυτά αναπαρίστανται συνήθως σαν δίκτυα με βάρη και είναι μη κατευθυνόμενα. Οι κόμβοι αντιπροσωπεύουν τα μετάγραφα και μέσω των βαρών των ακμών βλέπουμε πόσο έντονα τα επίπεδα έκφρασης των μετάγραφων συν-εκφράζονται. Τα δίκτυα γονιδιακής συν-έκφρασης χρησιμοποιούνται στις ολοκληρωμένες προσεγγίσεις της γενετικής και της βιολογίας συστημάτων που στοχεύουν στον εντοπισμό αιτιωδών γονιδίων και των δικτύων τους [54]. Μπορούν επίσης να χρησιμοποιηθούν για τον προσδιορισμό νέων μονοπατιών μιας νόσου, πρόβλεψη νέων λειτουργιών γονιδίων και αναζήτηση πιθανών βιοδεικτών ασθένειας [55]. Τα δίκτυα συν-έκφρασης δεν είναι στατικά και μπορούν να αλλάξουν ανάλογα με το βιολογικό πλαίσιο. Η σύγκριση αυτών των δικτύων μπορεί να βοηθήσει στη βελτίωση του λειτουργικού σχεδιασμού των γονιδίων και στην ανακάλυψη αλληλεπιδράσεων γονιδίων-γονιδίων, αποκαλύπτοντας το μοριακό μηχανισμοί σύνθετων ασθενειών ή των σχέσεων μεταξύ βιολογικών διεργασιών και βοηθώντας στην επιτάχυνση της διαδικασίας επιλογής γονιδίων για στοχευμένες μεταλλακτικές μελέτες [56]. Προσεγγίσεις ικανές να συγκρίνουν πυκνά και σταθμισμένα δίκτυα, όπως τα δίκτυα γονιδιακής

συν-έκφρασης, περιλαμβάνουν τη μέτρηση της ομοιότητας μεταξύ των τοπολογικών ιδιοτήτων των δικτύων, ομαδοποίηση για τον προσδιορισμό των συντηρημένων ενοτήτων των γονιδίων και σύγκριση των βαρών των ακμών για αντιστοιχισμένους ορθολόγους.

Έχει αναπτυχθεί ένας καλός αριθμός μεθόδων για την κατασκευή δικτύων συν-έκφρασης γονιδίων, οι οποίες ακολουθούν μια προσέγγιση δύο βημάτων. Στο πρώτο βήμα, επιλέγεται ένα μέτρο συν-έκφρασης και υπολογίζεται η ομοιότητας για κάθε ζεύγος γονιδίων που χρησιμοποιούν αυτό το μέτρο. Γενικότερα, χρησιμοποιούνται διαφορετικά μέτρα συσχέτισης για την κατασκευή δικτύων, συμπεριλαμβανομένων των συσχετίσεων Pearson ή Spearman. Στη συνέχεια, προσδιορίζεται ένα κατώτατο όριο και τα ζεύγη γονιδίων που έχουν βαθμολογία ομοιότητας υψηλότερη από το επιλεγμένο κατώφλι θεωρούνται ότι έχουν σημαντική σχέση συν-έκφρασης και συνδέονται στο δίκτυο. Έτσι, δημιουργείται ένα δικτύου συν-έκφρασης γονιδίων, όπου το κάθε γονίδιο παριστάνεται με έναν κόμβο και η σχέση μεταξύ δύο γονιδίων-κόμβων παριστάνεται με μία ακμή. Τα δεδομένα εισόδου για την κατασκευή ενός δικτύου συν-έκφρασης γονιδίου συχνά αντιπροσωπεύονται ως μήτρα.

Βάσεις δεδομένων: GEO [57], ArrayExpress [58] και COXPRESdb [59]

2.7 Φυλογενετικά δίκτυα (Phylogenetic networks)

Τα φυλογενετικά δίκτυα είναι γράφοι που αναπαριστούν εξελικτικές σχέσεις μεταξύ αλληλουχιών νουκλεοτιδίων, γονιδίων, χρωμοσωμάτων, γονιδιωμάτων ή ειδών. Τα δίκτυα αυτά είναι ικανά να φιλοξενήσουν εξελικτικά γεγονότα όπως ανασυνδυασμό, υβριδοποίηση ή πλευρική μεταφορά γονιδίων [60]. Στα φυλογενετικά δίκτυα, οι κόμβοι αναπαριστούν ταξινομικές βαθμίδες και οι σχέσεις μεταξύ δύο ταξινομικών βαθμίδων απεικονίζονται με κατευθυνόμενες ή μη ακμές. Ένα φυλογενετικό δίκτυο ορίζεται ως ένα γράφημα στο οποίο τουλάχιστον μια λειτουργική ταξινομική μονάδα συνδέεται με τον κοινό πρόγονο με δύο ή περισσότερες διαδρομές. Αυτό το χαρακτηριστικό διακρίνει ένα φυλογενετικό δίκτυο από το φυλογενετικό δέντρο. Τα φυλογενετικά δέντρα είναι ένα υποσύνολο των φυλογενετικών δικτύων. Θα μπορούσαμε να πούμε ότι αλγοριθμικά τα φυλογενετικά δικτυα είναι ένα είδος ιεραρχικής ομαδοποίησης, ωστόσο κάθε αντικείμενο μπορεί να συμμετάσχει σε παραπάνω από μια ομαδα.

Υπάρχουν διαφορετικοί τύποι φυλογενετικών δικτύων όπως:

- Τα δικτυωμένα φυλογενετικά δίκτυα, τα οποία παρέχουν μια ρητή αναπαράσταση της εξελικτικής ιστορίας.
- Τα εξελικτικά φυλογενετικά δίκτυα
- Τα διαχωρισμένα φυλογενετικά δίκτυα
- Τα φυλογενετικά δέντρα

Γενικότερα τα φυλογενετικά δίκτυα μπορεί να μην περιέχουν ρίζα όπως τα διαχωρισμένα δίκτυα. Στις περισσότερες περιπτώσεις, τέτοια δίκτυα απεικονίζουν μόνο σχέσεις μεταξύ των ταξινομικών βαθμίδων χωρίς να παρέχουν πληροφορίες για την εξελικτική ιστορία. Από την άλλη πλευρά τα φυλογενετικά δίκτυα με ρίζα, όπως τα φυλογενετικά δέντρα με ρίζα, δίνουν ρητές αναπαραστάσεις της εξελικτικής ιστορίας. Αυτό σημαίνει ότι απεικονίζουν τη σειρά με την οποία τα είδη αποκλίνουν (προσδιορίζονται), συγκλίνουν (υβριδοποιούνται) και μεταφέρουν το γενετικό υλικό (οριζόντια μεταφορά γονιδίων).

Τα φυλογενετικά δίκτυα μπορούν να οπτικοποιηθούν με λογισμικό όπως το SplitsTree [61], το R-package phangorn [62] και το Dendroscope [63].

2.8 Οικολογικά δίκτυα (Ecological networks)

Ένα οικολογικό δίκτυο είναι μια αναπαράσταση των βιοτικών αλληλεπιδράσεων σε ένα οικοσύστημα. Στα οικολογικά δίκτυα, οι κόμβοι αναπαριστούν τα είδη και οι ακμές που ενώνουν δύο κόμβους αναπαριστούν τις αλληλεπιδράσεις. Οι αλληλεπιδράσεις αυτές μπορεί να είναι τροφικές, συμβιωτικές, αμοιβαίες (αμφίδρομες) και ανταγωνιστικές (παράσιτο ξενιστή) [64]. Τα οικολογικά δίκτυα μπορούν να είναι εξίσου μη σταθμισμένα αλλά και σταθμισμένα δίκτυα, όπου τα βάρη των ακμών χαρακτηρίζουν την αλληλεπίδραση. Τα οικολογικά δίκτυα χρησιμοποιούνται για την περιγραφή και τη σύγκριση των δομών των πραγματικών οικοσυστημάτων, ενώ τα μοντέλα δικτύου χρησιμοποιούνται για τη διερεύνηση των επιπτώσεων της δομής του δικτύου σε ιδιότητες όπως η σταθερότητα του οικοσυστήματος. Τα οικολογικά δίκτυα χαρακτηρίζονται από υψηλά επίπεδα πολυπλοκότητας χωρίς όμως να εμφανίζουν αστάθεια. Επίσης, τα οικολογικά δίκτυα εμφανίζουν συνδεσιμότητα με τη συμπεριφορά των μεμονωμένων οργανισμών. Η χρήση οικολογικών δικτύων καθιστά δυνατή την ανάλυση των επιπτώσεων των ιδιοτήτων του δικτύου πάνω στη σταθερότητα ενός οικοσυστήματος. Τα οικολογικά δίκτυα μπορούν επίσης να βοηθήσουν στην κατανόηση δυναμικών ιδιοτήτων των φυτικών κοινοτήτων, όπως η διαδικασία οικολογικής διαδοχής [65].

Επιπλέον, η δομή αυτών των οικολογικών δικτύων μπορεί να παρέχει πληροφορίες σχετικά με τις οικολογικές και εξελικτικές διαδικασίες που δημιουργούν και διαμορφώνουν τη βιοποικιλότητα. Μπορεί επίσης να δείξει την ευθραυστότητα των οικολογικών κοινοτήτων σε διαφορετικά είδη διαταραχών, όπως από εξαφάνιση ειδών έως εισβολή ξένων ειδών, από την κλιματική αλλαγή έως τη λαθροθηρία ειδών [66]. Αξίζει να αναφερθεί ότι η δομή των δικτύων αυτών μπορεί να αναπαρασταθεί με την μορφή πίνακα και κάθε στοιχείο του πίνακα περιγράφει τη σύνδεση μεταξύ δύο ειδών. Στη περίπτωση μη κατευθυνόμενου δικτύου χωρίς βάρη, ο πίνακας είναι συμμετρικός και αποτελείται από τις τιμές 0 και 1. Ενώ στην περίπτωση κατευθυνόμενων δικτύων με βάρη, ο πίνακας έχει τις τιμές των βαρών [64].

2.9 Επιδημιολογικά δίκτυα (Epidemiological networks)

Οι φυσικές επιδημίες περιλαμβάνουν την εξάπλωση ασθενειών μέσω κάποιου φορέα. Η κατανόηση των επιδημικών διαδικασιών οδηγεί στην κατανόηση των μηχανισμών που μπορούν να χρησιμοποιηθούν για τη διανομή περιεχομένου ή για την πρόληψη της διάδοσης του ανεπιθύμητο περιεχόμενο. Τα επιδημιολογικά δίκτυα αποτελούν χρήσιμα εργαλεία για την μελέτη τέτοιων καταστάσεων. Συγκεκριμένα, τα επιδημιολογικά δίκτυα περιγράφουν έναν πληθυσμό και τις αλληλεπιδράσεις τους. Οι κόμβοι του δικτύου αντιπροσωπεύουν άτομα και οι ακμές (συνδέσεις) απεικονίζουν αλληλεπιδράσεις μεταξύ ατόμων που θα μπορούσαν ενδεχομένως να οδηγήσουν σε μετάδοση λοίμωξης [67]. Η δομή των επιδημιολογικών δικτύων εξαρτάται από τη δομή του πληθυσμού αλλά και από το είδος της λοίμωξης, καθώς με τον ίδιο πληθυσμό, το δίκτυο θα ήταν πολύ διαφορετικό για την εξάπλωση διαφορετικών λοιμώξεων. Τα επιδημιολογικά δίκτυα μπορούν να περιγραφούν από μη κατευθυνόμενα γραφήματα. Αν και η μετάδοση της λοίμωξης είναι ένα κατευθυνόμενο συμβάν (από ένα μολυσματικό άτομο σε ευπαθή), η πιθανότητα μετάδοσης κατά μήκος ενός άκρου θα ήταν συχνά η ίδια εάν η τοποθέτηση των δύο ατόμων (ευαίσθητη και μολυσματική) αντιστραφεί. Τα επιδημιολογικά δίκτυα μέσω των δεδομένων των ασθενών μπορούν να είναι μία ωφέλιμη βάση για την ανάπτυξη υποθέσεων σχετικά με τον τρόπο που δρα κάποια ασθένεια και να αποδειχθούν χρήσιμα στην δημιουργία φαρμάκων και στην ανάπτυξη θεραπειών.

Βάσεις δεδομένων: KEGG [42] και HPRD [35]

2.10 Δίκτυα ασθενειών (Diseases networks)

Τα δίκτυα ασθενειών αναπαριστούν τις συνδέσεις μεταξύ ασθενειών και διαταραχών με αναφορά στη γενετική προέλευσή τους ή σε άλλα χαρακτηριστικά. Οι κόμβοι των δικτύων ασθενειών αναπαριστούν ασθένειες και οι ακμές αντιπροσωπεύουν τις σχέσεις των ασθενειών με βάση την ομοιότητά τους. Συνήθως, οι ακμές συνδέουν δύο ασθένειες που μοιράζονται τουλάχιστον ένα γονίδιο στο οποίο οι μεταλλάξεις σχετίζονται και με τις δύο ασθένειες ή πολλές φορές αν οι ασθένειες μοιράζονται κοινά συμπτώματα, καθώς και φαινοτυπικές σχέσεις [68]. Τα δίκτυα ασθενειών είναι ένας διαισθητικός και ισχυρός τρόπος για να αποκαλυφθούν κρυφές συνδέσεις μεταξύ φαινομενικά μη συνδεδεμένων βιοϊατρικών οντοτήτων όπως ασθένειες, φυσιολογικές διεργασίες, μονοπάτια σηματοδότησης και γονίδια [69]. Ένα πεδίο στο οποίο η χρήση δικτύων ασθενειών είναι ωφέλιμη είναι η εύρεση νέων ασθενειών στις οποίες μπορούν να επαναχρησιμοποιηθούν παλιά φάρμακα. Ωστόσο, η χρήση των δικτύων ασθενειών δεν περιορίζεται μόνο στις συνδέσεις ασθενειών-ασθενειών, αλλά και στις σχέσεις μεταξύ της νόσου και άλλων παραγόντων όπως τα συμπτώματά της, τα σχετικά γονίδια ή οι θεραπείες της. Τα δίκτυα αυτά ομαδοποιούνται συνήθως με βάση την κατηγορία ασθενειών. Όμως, για παράδειγμα σε ένα δίκτυο ασθενειών που βασίζεται σε συμπτώματα, οι ασθένειες ομαδοποιούνται σύμφωνα με τις κατηγορίες τους. Επιπλέον, οι ασθένειες που μοιράζονται το ίδιο σύμπτωμα είναι πιο πιθανό να μοιράζονται τα ίδια γονίδια και τις ιδίες αλληλεπιδράσεις πρωτεΐνης.
Βάσεις δεδομένων: Online Mendelian Inheritance in Man (OMIM) [70]

2.11 Νευρωνικά δίκτυα (neural networks)

Ένα νευρικό δίκτυο είναι είτε ένα βιολογικό νευρωνικό δίκτυο, που αποτελείται από πραγματικούς βιολογικούς νευρώνες, είτε ένα τεχνητό νευρωνικό δίκτυο, για την επίλυση προβλημάτων τεχνητής νοημοσύνης. Υπό αυτήν την έννοια, τα νευρικά δίκτυα αναφέρονται σε συστήματα νευρώνων, είτε οργανικής είτε τεχνητής φύσης. Τα βιολογικά νευρωνικά δίκτυα παρέχουν πληροφορίες για τον τρόπο μετάδοσης σημάτων στο νευρικό σύστημα. Ενώ, ένα τεχνητό νευρωνικό δίκτυο είναι μια σειρά αλγορίθμων που προσπαθούν να αναγνωρίσουν τις υποκείμενες σχέσεις σε ένα σύνολο δεδομένων μέσω μιας διαδικασίας που μιμείται τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Τα νευρωνικά δίκτυα βασίζονται σε μια συλλογή συνδεδεμένων μονάδων ή κόμβων που ονομάζονται τεχνητοί νευρώνες, οι οποίοι μοντελοποιούν τους νευρώνες ενός βιολογικού εγκεφάλου [71]. Κάθε σύνδεση, όπως οι συνάψεις σε έναν εγκέφαλο, μπορεί να μεταδώσει ένα σήμα σε άλλους νευρώνες. Ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα, στη συνέχεια το επεξεργάζεται και μπορεί να σηματοδοτήσει νευρώνες που συνδέονται με αυτό. Οι συνδέσεις του βιολογικού νευρώνα διαμορφώνονται ως βάρη. Ένα θετικό βάρος αντανακλά μια διεγερτική σύνδεση, ενώ οι αρνητικές τιμές σημαίνουν ανασταλτικές συνδέσεις.

3. Πλατφόρμες ομαδοποίησης

Παραπάνω αναφέρθηκαν κάποια από τους αλγορίθμους ομαδοποίησης και οι κατηγορίες που ανήκουν με βάσει τον τρόπο που ομαδοποιούν τα δεδομένα. Στην πραγματικότητα υπάρχουν πολλοί διαφορετική αλγόριθμοι ομαδοποίησης, πολλοί από τους οποίους διατίθενται σε διάφορες πλατφόρμες. Παρακάτω θα αναφέρουμε ενδεικτικά κάποιες από αυτές όπως το λογισμικό Weka, ClusterMaker και η NORMA.

Το λογισμικό εξόρυξης δεδομένων ***Weka*** [72] παρέχει μια ολοκληρωμένη συλλογή από εργαλεία οπτικοποίησης και αλγορίθμους για την ανάλυση δεδομένων και τη προγνωστική μοντελοποίηση, μαζί με γραφικές διεπαφές χρήστη για εύκολη πρόσβαση σε αυτές τις λειτουργίες. Το Weka παρέχει διάφορες βασικές διεργασίες εξόρυξης δεδομένων, όπως προεπεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση και απεικόνιση. Όσον αφορά τους αλγορίθμους ομαδοποίησης υποστηρίζει μεθόδους όπως τον αλγόριθμο προσδοκίας-μεγιστοποίησης [73], φιλτραρισμένες ομαδοποιήσεις [74], ιεραρχική ομαδοποίηση [75] και τον αλγόριθμο k-means [14] δίνοντας την δυνατότητα επιλογής ανάμεσα στην απόσταση Manhattan και την Ευκλείδεια απόσταση. Η ομαδοποίηση μέσω του λογισμικού Weka δίνει σαν αποτέλεσμα τον αριθμό των ομάδων και πόσες εμφανίσεις κάθε ομάδα περιέχει. Για ορισμένους αλγορίθμους ο αριθμός των ομάδων μπορεί να οριστεί από παραμέτρους. Όταν είναι δυνατόν από τον αλγόριθμο ομαδοποίησης που επιλέγετε, τα αποτελέσματα του αλγορίθμου αναπαρίστανται γραφικά.

Στην εικόνα (Εικόνα 5) βλέπουμε τα αποτελέσματα της εφαρμογής του αλγορίθμου k-means σε δεδομένα καιρού με την επιλογή δημιουργίας δύο ομάδων και εφαρμογή της Ευκλείδειας απόστασης. Στο λογισμικό Weka τα αποτελέσματα των ομαδοποιήσεων εμφανίζονται ως πίνακας, του οποίου οι γραμμές είναι τα ονόματα των στοιχείων και οι στήλες αντιστοιχούν στα κεντροειδή της ομάδας [76].

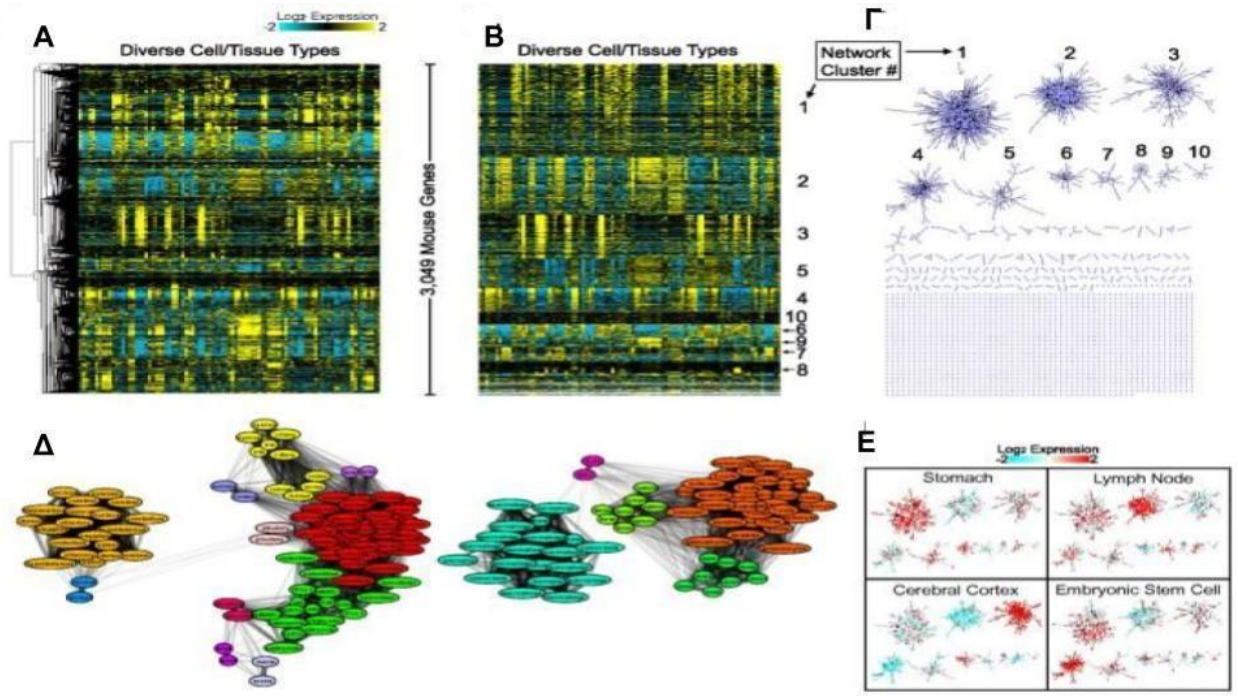
```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100
              -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2
              -A ''weka.core.EuclideanDistance -R first-last'' -I 500
              -num-slots 1 -S 10
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 16.237456311387238

Initial starting points (random):

Cluster 0: rainy,75,80,FALSE,yes
Cluster 1: overcast,64,65,TRUE,yes

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute      Full Data          0          1
                  (14.0)      (9.0)      (5.0)
=================================================
outlook            sunny      sunny   overcast
temperature      73.5714    75.8889       69.4
humidity         81.6429    84.1111       77.2
windy              FALSE      FALSE       TRUE
play                 yes        yes        yes



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 64%)
1       5 ( 36%)
```

Εικόνα 5: Αποτέλεσμα του αλγορίθμου σε δεδομένα καιρού [76] .

Σε αυτό το σημείο είναι σημαντικό να περιγραφεί και το **Cytoscape** [77], το οποίο είναι ένα πακέτο λογισμικού ανοικτού κώδικα για την οπτικοποίηση και την ανάλυση βιολογικών δικτύων. Το Cytoscape παρέχει μια εκτεταμένη διεπαφή προγραμματισμού εφαρμογών plugin (API) που επιτρέπει την επέκταση των εγγενών δυνατοτήτων του Cytoscape για να παρέχονται νέες λειτουργικότητες. Παρέχει διαφορές επεκτάσεις που υλοποιούν κάποιο είδος ομαδοποίησης. Ωστόσο, το περιβάλλον του καθένα από αυτά τα μεμονωμένα πρόσθετα είναι πολύ διαφορετικό και δεν υπάρχει αλληλεπίδραση μεταξύ τους.

Το ***ClusterMaker*** [78] το οποίο αποτελεί ένα plugin του Cytoscape παρέχει διαφόρους αλγορίθμους ομαδοποίησης και οπτικοποιήσεις που μπορούν να χρησιμοποιήσουν αυτόνομα ή να συνδυαστούν για την ανάλυση βιολογικών σετ δεδομένων. Αναλυτικότερα, διαθέτει ομαδοποίηση δικτύων και ομαδοποίηση χαρακτηριστικών. Η ομαδοποίηση δικτύων εντοπίζει πυκνά συνδεδεμένες περιοχές σε ένα δίκτυο. Υπάρχουν πολλές προσεγγίσεις ομαδοποίησης δικτύων, συμπεριλαμβανομένης της χρήσης αλγορίθμων γραφημάτων για τον εντοπισμό πυκνών περιοχών, είτε χρησιμοποιώντας μια τοπική προσέγγιση ξεκινώντας με μια γειτονιά κόμβων ή χρησιμοποιώντας μια σφαιρική προσέγγιση ξεκινώντας με ολόκληρο το γράφημα και επαναληπτικά διαμέριση του σε ομάδες χρησιμοποιώντας γραμμική άλγεβρα για να λειτουργήσει απευθείας στον πίνακα γειτνίασης. Στην ομαδοποίηση δικτύων στο ClusterMaker χαρακτηριστικοί είναι οι αλγόριθμοι MCL και Affinity propagation, MCODE [79], ομαδοποίηση κοινοτήτων (GLay) [80], φασματική ομαδοποίηση πρωτεϊνικών αλληλουχιών [81], TransClust [82] και AutoSOME [83]. Οι αλγόριθμοι αυτοί είναι χρήσιμη για τον εντοπισμό λειτουργικά συνδεδεμένων ομάδων πρωτεϊνών σε μεγάλα δίκτυα ομοιότητας πρωτεϊνών-πρωτεϊνών. Ενώ οι αλγόριθμοι συμπλέγματος χαρακτηριστικών περιλαμβάνουν ιεραρχική ομαδοποίηση, k-means, k-medoid και AutoSOME. Οι αλγόριθμοι ομαδοποίησης συμπλέγματος ομαδοποιούν τους κόμβους βάσει ομοιότητα των χαρακτηριστικών των κόμβων τους ή με βάσει ενός χαρακτηριστικού μιας ακμής.

Το ClusterMaker παρέχει επίσης 3 διαφορετικές οπτικοποιήσεις (Εικόνα 6) ανάλογα με τον αλγόριθμο που έχει επιλεχθεί. Οι οπτικοποιήσεις αυτές είναι δίκτυο που παρέχετε από το Cytoscape, δενδρόγραμμα και heatmaps. Οι δυό οπτικοποιήσεις αυτές είναι συνδεδεμένες με το δίκτυο Cytoscape, επιτρέποντας επιλογές πάνω στο δίκτυο να αντικατοπτρίζονται σε μία ή περισσότερες από τις άλλες οπτικοποιήσεις. Επιλογές πάνω στο heatmap μπορούν επίσης να αντικατοπτρίζονται στο δίκτυο και στα υπόλοιπα heatmaps. Όλοι οι αλγόριθμοι παρέχουν επίσης την επιλογή δημιουργίας ομάδας Cytoscape για κάθε σύμπλεγμα. Μια ομάδα συλλέγει ένα σύνολο κόμβων και των άκρων τους σε ένα αντικείμενο που μπορεί να εκπροσωπηθεί ως νέος κόμβος.

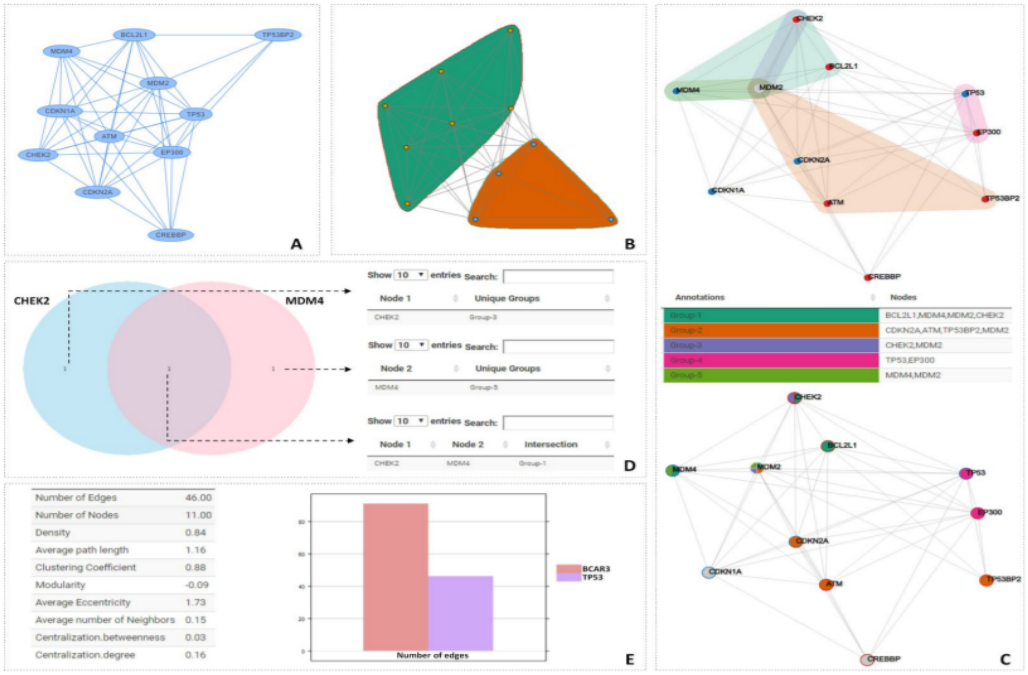


Εικόνα 6: Η ομαδοποίηση γονιδιακής έκφρασης αποκαλύπτει αλληλεπιδραστικές ενότητες πρωτεϊνών ποντικού και ασαφείς σχέσεις μεταξύ κυττάρων ποντικού και ιστών. Α) Heat map ιεραρχικής ομαδοποίησης. Β) Heat map AutoSOME ομαδοποίησης. Γ) Πρωτεϊνική αλληλεπίδραση χωρισμένη σε υποδίκτυα που αντιστοιχούν σε συμπλέγματα συν-έκφρασης που αναγνωρίζονται από το AutoSOME. Δ) Ασαφές δίκτυο ομάδων τύπων κυττάρων/ιστών στο GSE10246. Ε) Δεδομένα έκφρασης τεσσάρων τύπων κυττάρων / ιστών από το GSE10246 επικαλύπτονται στα δέκα μεγαλύτερα υποδίκτυα από το Γ [78].

Η **NORMA** [84] είναι μια εφαρμογή οπτικοποίησης που εστιάζει στην οπτικοποίηση των σχολιασμένων δικτύων και τη τοπολογική ανάλυση, ικανή να χειρίζεται πολλαπλά δίκτυα και σχολιασμούς ταυτόχρονα. Οι προυπολογισμένοι σχολιασμοί μπορούν να μεταφορτωθούν και να απεικονιστούν σε ένα δίκτυο είτε ως χρωματιστοί κόμβοι γραφημάτων πίτας είτε ως κυρτοί κύβοι (Convex Hulls) χρωματισμένοι σε στυλ διαγράμματος Venn (Venn-diagram). Σε περίπτωση που δεν υπάρχει σχολιασμός, προσφέρονται αλγόριθμοι για αυτοματοποιημένο εντοπισμό κοινοτήτων. Στη NORMA οι χρήστες έχουν τη δυνατότητα να προσαρμόσουν τις προβολές δικτύου χρησιμοποιώντας αλγόριθμους διάταξης ή να επιτρέψουν στη NORMA να τις τροποποιήσει ελαφρώς για οπτικά καλύτερο διαχωρισμό των ομάδων. Τα δίκτυα οπτικοποιούνται στην απλούστερη μορφή τους με τη χρήση της βιβλιοθήκης vizNetwork (Εικόνα 7Α) και είναι πλήρως διαδραστικά. Επιπλέον, η NORMA προσφέρει μεθόδους για αυτόματο εντοπισμό κοινοτήτων (Εικόνα 7Β). Αυτό είναι ιδιαίτερα σημαντικό σε περιπτώσεις όπου κάποιος δεν έχει εξωτερικά προκαθορισμένα αποτελέσματα. Οι μέθοδοι που προσφέρονται είναι: Fast-Greedy, Louvain, Label-Propagation, Walktrap και Betweenness. Μόλις επιλεχθεί μια μέθοδος εντοπισμού κοινοτήτων, οι χρήστες μπορούν να δουν τα αποτελέσματα διαδραστικά σε μορφή πινάκων ή ως στατικά γραφήματα (Εικόνα 7C). Μια επιπλέον δυνατότητα της NORMA είναι η απεικόνιση των σχολιασμένων δικτύων. Σχολιασμένα είναι τα δίκτυα που συνοδεύονται από προκαθορισμένες ομάδες, κοινότητες, υπογράφους, επισημασμένες περιοχές ή γειτονιές. Το επιλεγμένο δίκτυο μπορεί να οπτικοποιηθεί με:

- Convex Hulls: χρωματισμένα Convex Hulls χρησιμοποιούνται για να επισημανθούν οι κοινότητες σε ένα διάγραμμα Venn.

- Pie-chart: οι κόμβοι του δικτύου απεικονίζονται στη συνέχεια ως γραφήματα πίτας (Pie-chart), χωρίζονται σε φέτες για να απεικονιστούν οι ομάδες στις οποίες ανήκουν οι κόμβοι. Εάν ένας κόμβος για παράδειγμα ανήκει σε τέσσερις ομάδες, τότε το γράφημα πίτας θα αποτελείται από τέσσερις ίσες φέτες χρωματισμένες με διαφορετικά χρώματα.
- Διάγραμμα Venn (Εικόνα 7D)



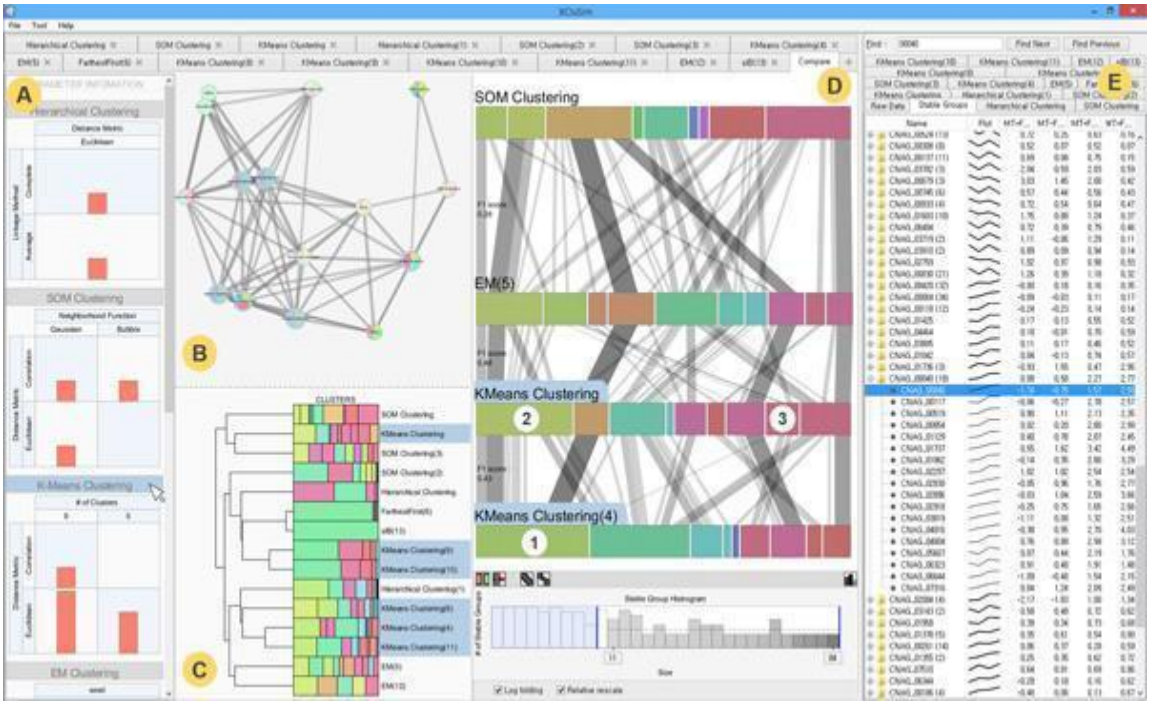
Εικόνα 7: Α) Το δίκτυο TP53. Β) Ανίχνευση κοινοτήτων χρησιμοποιώντας τον αλγόριθμο Louvain. C) Ανώτερο μέρος: Convex Hulls με χρωματιστούς κόμβους, Κάτω μέρος: Κόμβοι πίτας με χρώματα περιγράμματος. Μέσο μέρος: Οπτικοποίηση των ομάδων σε έναν διαδραστικό πίνακα. D) Διαγράμματα Venn για εμφάνιση κοινών ομάδων μεταξύ οποιουδήποτε ζεύγους επιλεγμένων κόμβων. Ε) Αριστερό μέρος: Βασική τοπολογική ανάλυση του δικτύου TP53, Δεξί μέρος: bar Chart για την άμεση σύγκριση ενός τοπολογικού χαρακτηριστικού μεταξύ δύο δικτύων. [84]

3.1 Εργαλεία σύγκρισης ομαδοποιήσεων

Παρά την πληθώρα των αλγορίθμων ομαδοποίησης, όταν αυτοί εφαρμόζονται στα δεδομένα, τα αποτελέσματα μπορεί να διαφέρουν είτε λόγω διαφορετικής επιλογής παραμέτρων είτε λόγω επιλογής διαφορετικών αλγορίθμων. Επιπλέον, στις περισσότερες περιπτώσεις προκαθορισμένες ομαδοποιήσεις, οι οποίες μπορούν να χρησιμοποιηθούν για συγκριτική αξιολόγηση σπάνια εφαρμόζονται. Επομένως, είναι δύσκολο να υποστηρίξουμε ποιος αλγόριθμος ή παραμετροποίηση λειτουργεί καλύτερα για συγκεκριμένες περιπτώσεις. Για αυτό τον λόγο, πολλές μελέτες προσπαθούν να περιγράψουν πρωτόκολλα που μπορούν να χρησιμοποιηθούν για τη σύγκριση και αξιολόγηση των αποτελεσμάτων των ομαδοποιήσεων [85] [86]. Έχουν δημιουργηθεί διάφορες υλοποιήσεις για την εξυπηρέτηση αυτού του σκοπού. Για παράδειγμα κάποιες από τις υλοποίησης αυτές είναι το XCluSim, η MatchMaker και δύο εφαρμογές των Windows σχετικές με την συγκρίση και οπτικοποίηση ομαδοποιήσεων, η HCE και το CComViz.

Το ***XCluSim*** [87] είναι μια εφαρμογή σε Java για ομαδοποίηση δεδομένων και συγκριτική ανάλυση και οπτικοποίηση αυτών, η οποία είναι μόνο διαθέσιμη κατόπιν αίτησης. Το XCluSim επιτρέπει στους χρήστες είτε να εκτελούν μέσω της εφαρμογής αλγορίθμους ομαδοποίησης

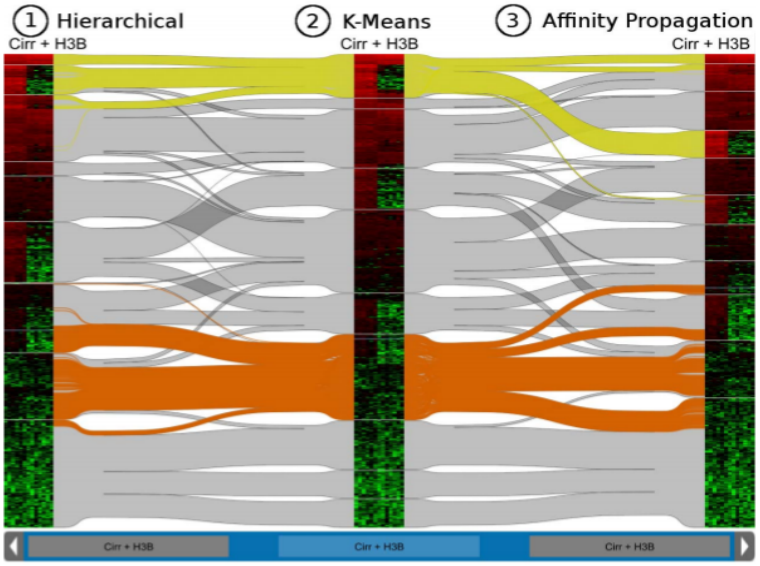
όπως ιεραρχική ομαδοποίηση, ομαδοποίηση SOM [88], ομαδοποίηση k-means και ομαδοποίηση OPTICS [89] , είτε μέσω του πακέτου Weka μέσα στην εφαρμογή, είτε να ανεβάζουν εξωτερικά αποτελέσματα ομαδοποίησης. Το XCluSim χρησιμοποιεί χρωματικά σχήματα για να αναπαραστήσει χρωματικά τις ομοιότητες των ομάδων. Για αυτο το λόγο ομαδοποιεί όλες τις ομάδες των αποτελεσμάτων των ομαδοποιήσεων χρησιμοποιώντας ομαδοποίηση HAC (hierarchical agglomerative clustering) και αναθέτει ένα χρώμα σε κάθε σύμπλεγμα με βάση το σχήμα χρωματικής κωδικοποίησης Tree Colors έτσι ώστε παρόμοια συμπλέγματα να έχουν παρόμοια χρώματα. Ο συντελεστής συσχέτισης χρησιμοποιείται ως μέτρο ομοιότητας ενός ζεύγους ομάδων, έτσι ώστε να ανατεθούν παρόμοια χρωματα σε όμοιες ομάδες. Επιπλέον, με την βοήθεια της μετρικής F- Measure (F1) [90], το XCluSim παρέχει τρεις οπτικοποιήσεις για τις συγκρίσεις των ομαδοποιήσεων. Αυτές είναι: i) ένα κατευθυνόμενο δίκτυο με γράφημα πίτας (Εικόνα 8B) ii) ένα δενδρόγραμμα (Εικόνα 8C) και iii) πολλές παράλληλες προβολές (π.χ. γραφήματα Sankey) (Εικόνα 8D). Όπως βλέπουμε εκτός από τις τρεις αυτές οπτικοποιήσεις, το XCluSim παρέχει και μια λίστα σε μορφή πίνακα (Εικόνα 8E) μέσω της οποίας οι χρήστες έχουν πρόσβαση σε πληροφορίες σχετικά με τα επιλεγμένα αποτελέσματα ομαδοποιήσεων, με κάθε αποτέλεσμα να μπορεί να εμφανιστεί σε μια ξεχωριστή καρτέλα στην προβολή λίστας πίνακα. Ωστόσο, ένα βασικό μειονέκτημα του XCluSim όπως αναφέρθηκε είναι ότι δεν είναι ελεύθερα διαθέσιμο στο κοινό αλλα μόνο κατόπιν αιτήσεις του.



Εικόνα 8: Οπτικοποιήσεις για τη σύγκριση των ομαδοποιήσεων στο XCluSim. A) Προβολή παραμέτρων πληροφοριών B) Κατευθυνόμενο δίκτυο. C) Δενδρόγραμμα. D) Παράλληλες προβολές. E) Προβολή λίστας σε πίνακα.[87]

Η ***MatchMaker*** [91] είναι μια εφαρμογή Java για οπτικοποίηση συγκρίσεων ομαδοποιήσεων και είναι διαθέσιμο ως μέρος του πλαισίου οπτικοποίησης του Caleydo [92]. Η εφαρμογή δίνει την δυνατότητα διαχωρισμού πολυδιάστατων συνόλων δεδομένων σε υποομάδες, εφαρμογή αλγορίθμων ομαδοποίησης σε αυτές τις ομάδες ξεχωριστά και στη συνέχεια σύγκριση των οπτικών αποτελεσμάτων. Αυτό επιτρέπει στους χρήστες να βρίσκουν μοτίβα στα δεδομένα

που διαφορετικά θα ήταν δύσκολο και να συγκρίνουν τα αποτελέσματα διαφορετικών αλγορίθμων ομαδοποίησης. Η MatchMaker παράγει έναν συνδυασμό από heatmaps και παράλληλες γραφικές παραστάσεις για την αναπαράσταση των ομαδοποιήσεων (Εικόνα 9). Για τον προσδιορισμό των σχέσεων σχηματίζονται ομαδοποιημένες καμπύλες και ακμές ανάμεσα σε σχετικές εγγραφές σε διαφορετικές ομάδες. Αναλυτικότερα, οι ομάδες και οι εγγραφές εντός των ομάδων ταξινομούνται σύμφωνα με τη μέση τιμή τους, με αποτέλεσμα η θέση των εγγραφών να έχει σημασία και να μην είναι τυχαία. Οι ομάδες τοποθετούνται διπλά διπλά έτσι ώστε κάθε ομάδα να είναι ισοδύναμη με έναν άξονα στη παράλληλη γραφική παράσταση συντεταγμένων. Συνδέοντας τις ίδιες εγγραφές μεταξύ ομάδων, ολοκληρώνουμε τη παράλληλη γραφική παράσταση συντεταγμένων. Ωστόσο, αντί να χρησιμοποιούνται απλές γραμμές ως άξονες, οι ομάδες αναπαρίστανται ως heatmaps. Με αυτό τον τρόπο μπορούν να συγκριθούν τιμές με τη βοήθεια χρωμάτων, να ξεχωρίσουμε το μέσο μέγεθος των ομάδων, καθώς και των εγγραφών τους και τέλος οι σχέσεις μεταξύ των ομάδων και των εγγραφών εμφανίζονται μέσω των ακμών. Τέλος, η Matchmaker παρέχει επισκόπηση και διαδραστικές ενσωματωμένες προβολές λεπτομερειών για τις μεμονωμένες ομάδες. Ώστοσο, η έκδοση Java του Caleydo έχει σταματήσει να λαμβάνει ενημερώσεις από το 2015 και η τρέχουσα διαδικτυακή έκδοση δεν υποστηρίζει το MatchMaker.

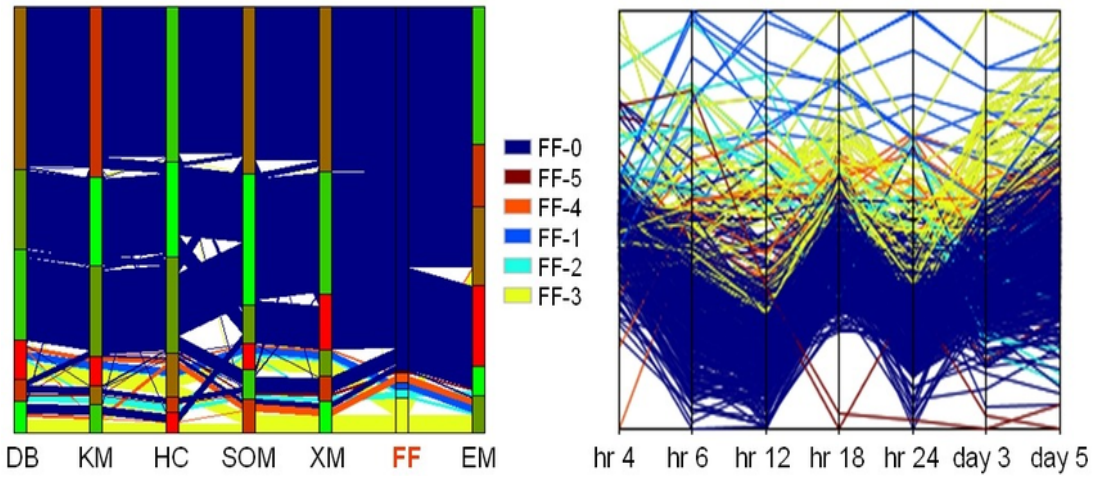


Εικόνα 9: Σύγκριση τριών αλγορίθμων ομαδοποίησης (hierarchical, k-means και Affinity Propagation) πάνω σε 1800 εγγραφές. Με τις κίτρινες και πορτοκαλί ακμές βλέπουμε ότι ο αλγόριθμος k-means ανέθεσε διαφορετικές εγγραφές σε μια ομάδα ενώ οι άλλοι δυο αλγόριθμοι λειτουργούν σύμφωνα με το επιθυμητό. [91]

Η ***HCE*** (Hierarchical Cluster Explorer) [93] είναι μια εφαρμογή που επιτρέπει την εξερεύνηση μεγάλων συνόλων δεδομένων μέσω οπτικοποιήσεων ιεραρχικής ομαδοποίησης. Η HCE δίνει τη δυνατότητα επισκόπησης όλου του συνόλου δεδομένων,, σε συνδυασμό με λεπτομερή προβολή έτσι ώστε τα μοτίβα υψηλού επιπέδου και τα σημαντικά σημεία να μπορούν να εντοπιστούν και να εξεταστούν ευκολότερα. Με τη HCE επιτρέπεται η σύγκριση ενός μόνο ζεύγους αποτελεσμάτων ομαδοποίησης και η αναπαράσταση των ομαδοποιήσεων γίνεται μέσω δενδρογραμμάτων. Ουσιαστικά αναπαριστά ένα ζευγάρι ιεραρχικών αποτελεσμάτων ομαδοποίησης παράλληλα ώστε να καταστήσει δυνατή τη σύγκριση μεταξύ των δύο αποτελεσμάτων. Οι σχέσεις μεταξύ των αντικειμένων αναπαρίστανται με συνδεομενες γραμμές. Σε ένα μεγάλο σύνολο δεδομένων αν υπάρχουν πολλά αντικείμενα που

αντιστοιχίζονται, τότε εμφανίζονται πολλές διασταυρούμενες γραμμές και αυτό αποτελεί πρόβλημα. Επίσης, το HCE εμφανίζει στα φύλλα των δενδρογραμμάτων μωσαϊκό με χρώματα ωστέ να αναπαρασταθούν τα υποκείμενα γραφικά μοτίβα.

Το ***CComViz*** [94] ξεπέρασε το πρόβλημα των διασταυρούμενων γραμμών επικεντρώνοντας στη σύγκριση περισσότερων από δύο ομαδοποιήσεων. Αποτελεί ένα εργαλείο οπτικοποιήσεων, το οποίο αναπαριστά τα δεδομένα σε πολυλίνες σε μια ακολουθία από άξονες που διατηρούν ανεξάρτητα σημεία δεδομένων παρόμοια με τις παράλληλες συντεταγμένες (Εικόνα 10). Το CComViz επινόησε έναν αλγόριθμο για την αναδιάταξη των ομάδων και των μελών τους για την ελαχιστοποίηση της οπτικής ακαταστασίας μεταξύ κάθε διάστασης. Επιπλέον, μέσω του "hot dimension" που διαθέτει δείχνει πώς οι εγγραφές σε μια ομάδα μιας διάστασης εμπίπτουν σε ομάδες άλλων διαστάσεων. Επίσης, μπορούμε να πούμε ότι το CComViz είναι ένα διαδραστικό οπτικό-αναλυτικό εργαλείο που εμφανίζει σταθερότητα δεδομένων, ροή δεδομένων, κατανομή πυκνότητας και ιεραρχία και συσχέτιση δεδομένων σε πολλαπλά επίπεδα μέσα σε μία γραφική οθόνη. Το CComViz εστιάζει στη σύγκρισης περισσότερων από δύο αποτελέσματα ομαδοποίησης. Τα αποτελέσματα των ομαδοποιήσεων αναπαρίστανται με παράλληλη γραφική παράσταση συντεταγμένων (Εικόνα 10), όπου οι ομαδοποιήσεις απεικονίζονται ως διαστάσεις, οι ομάδες ως κατακόρυφες θέσεις σε κάθε διάσταση και τα στοιχεία ως γραμμές. Το CComViz προς το παρόν δεν είναι δημόσια διαθέσιμο.



Εικόνα 10: CComViz και παράλληλες συντεταγμένες που δείχνουν τα χαρακτηριστικά των αποτελεσμάτων ομαδοποίησης FF [94].

4. Η εφαρμογή VICTOR

4.1 Το πρόβλημα

Η ομαδοποίηση είναι σημαντική σε πολλές επιστήμες όπως για παράδειγμα στη κοινωνιολογία, στη βιολογία και στη στατιστική, αλλά και σε πολλούς τομείς της πληροφορικής και Βιοπληροφορικής ανάλυσης. Επίσης, όσον αφορά την επεξεργασία δεδομένων η ομαδοποίηση μπορεί να συμβάλλει στη μείωση των δεδομένων και την ανάδειξη σημαντικών πληροφοριών από αυτά. Πολλές φορές χρειάζεται να αξιολογήσουμε τη μέθοδο της ομαδοποίησης. Μια ομαδοποίηση είναι καλή αν παράγει ομάδες καλής ποιότητας. Η ποιότητα εξαρτάται από τη μέτρηση της ομοιότητας και τη μέθοδο υλοποίησης της ομαδοποίησης. Η επιλογή διαφορετικών αλγορίθμων για το ίδιο σύνολο δεδομένων μπορεί να επιφέρει διαφορετικά αποτελέσματα ομαδοποιήσεων, καθιστώντας δύσκολη την απόφαση για το ποιος αλγόριθμος ταιριάζει καλύτερα στα δεδομένα. Επιπλέον, πολύ από τους αλγορίθμους ομαδοποίησης διαθέτουν διαφορές παραμέτρους, που μπορούν να επηρεάσουν τα αποτελέσματα ή να τα προσαρμόσουν στις ανάγκες των χρηστών. Επομένως, ακόμη και αν εφαρμοστεί ο ίδιος αλγόριθμος στο ίδιο σετ δεδομένων και πάλι με τη χρήση διαφορετικών παραμέτρων μπορούμε να έχουμε διαφορετικά αποτελέσματα. Τέλος, πολλοί από τους αλγορίθμους ομαδοποίησης εμφανίζουν μειονεκτήματα και δεν είναι το ίδιο αποτελεσματικοί στα διάφορα σετ δεδομένων. Για παράδειγμα, αξίζει να αναφερθεί ότι στη βιοπληροφορική ανάλυση ενός σετ δεδομένων γονιδιακής έκφρασης, αρχικά τα δεδομένα φιλτράρονται ώστε να απομακρυνθούν περιττά γονίδια. Στη συνέχεια, εφαρμόζονται είτε διαφορετικοί αλγόριθμοι ομαδοποίησης είτε ο ίδιος αλγόριθμος με διαφορετικές παραμέτρους στο σετ δεδομένων. Τέλος, οι ομάδες των ομαδοποιήσεων πιστοποιούνται για το αν τα στοιχεία έχουν ομαδοποιηθεί σωστά. Όταν η ποιότητα της ομαδοποίησης δεν είναι η επιθυμητή τότε τα παραπάνω βήματα χρειάζεται να επαναληφθούν [87]. Έτσι η διαδικασία σύγκρισης διαφορετικών αποτελεσμάτων ομαδοποίησης για την ανάδειξη του πιο αποτελεσματικού γίνεται ιδιαίτερα δύσκολη και χρονοβόρα. Για όλους αυτούς τους λόγους που αναφέρθηκαν είναι ιδιαίτερα χρήσιμη η ύπαρξη ενός εργαλείου που συγκρίνει τα αποτελέσματα των ομαδοποιήσεων και αναδεικνύει τις πιο σημαντικές ομάδες.

4.2 Εισαγωγή

Όπως αναφέραμε διαφορετικοί αλγόριθμοι ή διαφορετικές παράμετροι αυτών μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα. Πολλές φορές τα αποτελέσματα αυτά χρειάζονται να φιλτραριστούν και να συγκριθούν χειροκίνητα, έτσι ώστε να βρεθούν οι ομαδοποιήσεις με τις ιδανικές ομάδες. Στη παρούσα διπλωματική εργασία, παρουσιάζεται το VICTOR, μια οτικό-αναλυτική εφαρμογή που συγκρίνει αποτελέσματα ομαδοποιήσεων χρησιμοποιώντας διάφορες μετρικές σύγκρισης και οπτικοποίησης. Το VICTOR δημιουργήθηκε για την αυτοματοποίηση της διαδικασίας που αναφέρθηκε παραπάνω και μπορεί να χρησιμοποιηθεί για την αξιολόγηση της ποιότητας των αλγορίθμων ομαδοποίησης και των παραμέτρων τους. Η σύγκριση των αποτελεσμάτων ομαδοποίησης μπορεί να οδηγήσει στη σωστή επιλογή αλγορίθμου ομαδοποίησης ανάλογα με το σετ

δεδομένων. Καθώς, το VICTOR δεν ενσωματώνει εφαρμογές αλγορίθμων ομαδοποίησης, παραμένει ένα ευέλικτο εργαλείο γενικού σκοπού.

Το VICTOR αποτελεί ένα πλήρως διαδραστικό και ανεξάρτητο οπτικό-αναλυτικό εργαλείο που επιτρέπει τη σύγκριση και οπτικοποίηση διαφορών αλγορίθμων ομαδοποίησης. Έχει αναπτυχθεί σε R και JavaScript. Για το frontend χρησιμοποιήθηκε το πακέτο R/Shiny, HTML, CSS και JavaScript. Αναλυτικότερα, το πακέτο Shiny χρησιμοποιήθηκε ως διαμεσολαβητής για τη δημιουργία επικοινωνίας μεταξύ των μεταβλητών και των συναρτήσεων της γλώσσας R και της Javascript. Ενώ, για το backend χρησιμοποιήθηκε η βιβλιοθήκη mclustcomp. Επίσης, τα γραφήματα δημιουργήθηκαν μέσω της βιβλιοθήκης d3 JavaScript και της R.

Το VICTOR μπορεί να διαχειριστεί πολλαπλά αποτελέσματα ομαδοποιήσεων και να τα συγκρίνει ανά δύο με την βοήθεια μετρικών σύγκρισης ομαδοποιήσεων. Ένα από τα πλεονεκτήματα του VICTOR είναι ότι παρέχει δέκα διαφορετικές μετρικές σύγκρισης των ομαδοποιήσεων και ο χρήστης μπορεί να επιλέξει μια ή και όλες για την σύγκριση των ομαδοποιήσεων. Οι μετρικές σύγκρισης προέρχονται αποκλειστικά από τη βιβλιοθήκη mclustcomp, η οποία περιέχει μια συλλογή από διάφορες βιβλιογραφικά διαθέσιμες μετρικές σύγκρισης. Η βιβλιοθήκη αυτή επιστρέφει τιμές σύγκρισης των ομαδοποιήσεων χρησιμοποιώντας διαφορές σχεδιαστικές μεθόδους. Αξίζει να σημειωθεί ότι στη βιβλιοθήκη mclustcomp τα δύο διανύσματα ετικετών θα πρέπει να έχουν το ίδιο μήκος και αυτός είναι ο κύριος λόγος για τον οποίο οι χρήστες χρειάζεται να επιλέξουν ανάμεσα από την επιλογή φιλτραρίσματος, φιλτράρισμα με τομή (intersection) ή την επιλογή φιλτραρίσματος υπερσυνόλου (superset). Με τον υπολογισμό αυτών των μετρικών, οι συγκρίσεις μεταξύ των ομαδοποιήσεων μπορούν να οπτικοποιηθούν σε Bar charts, Hierarchical Heatmaps, Sankey plots, Circos plots και δίκτυα. Στη συνέχεια, γίνεται αναλυτική περιγραφή του VICTOR, των οπτικοποιήσεων του, καθώς και τα αποτελέσματα από τη χρήση του σε τρία διαφορετικά παραδείγματα σύγκρισης οπτικοποιήσεων. Επιπλέον, αξίζει να αναφερθεί ότι στην ηλεκτρονική του μορφή, το VICTOR μπορεί να χειριστεί δίκτυα έως 20.000 ακμές (για τον υπολογισμό του conductance) και σύνολα αρχείων μεγέθους έως 1MB το καθένα, κάτι που μπορεί να παρακαμφθεί κατά τη λήψη και την εκτέλεση του τοπικά.

Στον παρακάτω πίνακα (Πίνακας 1) βλέπουμε τη σύγκριση του VICTOR, με παρόμοια λογισμικά σύγκρισης ομαδοποιήσεων, τα οποία είναι το XCluSim, η MatchMaker, το CComViz και η HCE. Το VICTOR δεν αποτελεί εργαλείο ομαδοποίησης και δεν ενσωματώνει υλοποιήσεις αλγορίθμων ομαδοποίησης, ένα χαρακτηριστικό που το καθιστά ευέλικτο εργαλείο γενικού σκοπού. Ωστόσο, αυτή τη στιγμή είναι η μόνη εφαρμογή ιστού που διευκολύνει τις συγκρίσεις συνόλων συμπλέγματος, παρέχοντας στους χρήστες έως 10 μετρικές σύγκρισης ομαδοποιήσεων και 6 διαφορετικές απεικονίσεις. Αξίζει να αναφερθεί ότι το VICTOR δεν επισημαίνει μια συγκεκριμένη ομάδα ως την καλύτερη επιλογή. Είναι μάλλον ένα εργαλείο οπτικής σύγκρισης των αποτελεσμάτων ομαδοποίησης είτε από διαφορετικούς διαχωριστικούς αλγόριθμους ή από εκτελέσεις του ίδιου αλγορίθμου ομαδοποίησης αλλά με διαφορετικές παραμέτρους.

	VICTOR	XCluSim	MatchMaker	CComViz	HCE
#Μετρικές σύγκρισης	10	1	1	1	1
#Οπτικοποιήσεις σύγκρισης	6	3	2 in 1	1	1
Μετρική conductance μιας ομάδας	✔	✖	✖	✖	✖
Πολλαπλή σύγκριση	✔	✔	✔	✔	✖
Ανάλυση ομαδοποιήσεων	✖	✔	✔	✖	✔
Διαθεσιμότητα	Εφαρμογή Web, GitHub	Κατόπιν αιτήσεις	Caleydo framework (old version)	✖	Direct download
Τύπος εφαρμογής	Εφαρμογή Web, R Shiny	Standalone	Framework software	Μη καθορισμένο	Εκτελέσιμο σε Windows
Υλοποίηση	R/Javascript	Java	R/Java	Μη καθορισμένο	Εφαρμογή Windows
Τελευταία ενημέρωση	2021	2016	2010	2009	2006

Πίνακας 1: Σύγκριση εργαλείων σύγκρισης ομαδοποιήσεων.

5. Η Βιβλιοθήκη mclustcomp

Η mclustcomp είναι μια βιβλιοθήκη που παρέχει μια συλλογή από μεθόδους, οι οποίες παίζουν ρόλο παρόμοιο με αυτό των μετρικών ή των αποστάσεων στα πλαίσια του υπολογισμού της ομοιότητας δυο ομαδοποιήσεων. Συγκεκριμένα αποτελείται από 25 διαφορετικές μετρικές για την σύγκριση δύο ομαδοποιήσεων. Οι μετρικές αυτές χωρίζονται σε τρεις κατηγορίες 1) Counting Pairs, 2) Set Overlaps/Matching και 3) Mutual Information. Στις δύο πρώτες κατηγορίες γίνεται χρήση ενός πίνακα σύγχυσης (confusion matrix) (Πίνακας 2) για τον υπολογισμό των μετρικών, ενώ στην τρίτη κατηγορία αξιοποιείται η εντροπία ή μέση πληροφορία (entropy of Information). Δεδομένου δύο ομαδοποιήσεων η βιβλιοθήκη επιστρέφει τιμές σύγκρισης των ομαδοποιήσεων που αντιστοιχούν σε κάθε μετρική. Στο VICTOR χρησιμοποιούνται μόνο 10 από τις 25 μετρικές, για τους λόγους που θα αναφερθούν στην συνέχεια [95] [96].

5.1 Πίνακας σύγχυσης (Confusion Matrix)

Έστω $C = \{C_1, C_2, ..., C_\kappa\}$ μια ομαδοποίηση με k μη μηδενικές ομάδες από το ίδιο σετ δεδομένων X και $C' = \{C_1', C_2', ..., C_l'\}$ μια δεύτερη ομαδοποίηση με l μη μηδενικές ομάδες του ίδιου σετ δεδομένων. Ως Confusion Matrix $M = m_{ij}$ χαρακτηρίζεται ο $k \times l$ πίνακας μεταξύ των C και C', του οποίου η i-j είσοδος ισούται με τον αριθμό των στοιχείων της τομής των δύο ομαδοποιήσεων:

$$m_{ij} = |C \cap C'|,\ 1 \leq i \leq k \text{ και } 1 \leq j \leq l.$$

C\C'	C_1'	C_2'	. . .	C_n'	sums
C_1	m_{11}	m_{12}	. . .	m_{1n}	a_1
C_2	m_{21}	m_{22}	. . .	m_{2n}	a_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
C_n	m_{n1}	m_{n2}	. . .	m_{nn}	a_n
sums	b_1	b_2		b_n	n

Πίνακας 2: Παράδειγμα πίνακα σύγχυσης.

5.2 Μετρικές για την σύγκριση ομαδοποιήσεων

Στη συνέχεια, παρατίθενται κάποιες από τις μετρικές που χρησιμοποιούνται για τη σύγκριση των ομαδοποιήσεων. Οι μετρικές αυτές περιγράφονται αναλυτικά στις τρεις κατηγορίες που ανήκουν, όπως περιέχονται στη βιβλιοθήκη mclustcomp. Συγκεκριμένα, δίνονται πληροφορίες

για τον τρόπο λειτουργίας των μετρικών, καθώς και ο τύπος υπολογισμού αυτών. Επιπλέον, αναφέρονται τα αρνητικά που εμφανίζονται σε κάποιες από τις μετρικές, τα οποία αποτέλεσαν και λόγο για να μην χρησιμοποιηθούν στο VICTOR.

5.2.1 Κατηγορία 1: Καταμέτρηση ζευγών (Counting Pairs)

Αυτή η κατηγορία βασίζεται στην καταμέτρηση των ζευγαριών των αντικειμένων που έχουν ομαδοποιηθεί ή όχι στην ίδια ομάδα στη C και C' ομαδοποίηση. Κάθε ζεύγος από αντικείμενα ανήκει σε μια από τις παρακάτω κατηγορίες :

n_{00}: σύνολο ζευγών αντικειμένων που ανήκουν στην ίδια ομάδα και στο C και στο C'
n_{01}: σύνολο ζευγών αντικειμένων που ανήκουν στην ίδια ομάδα στο C και όχι στο C'
n_{10}: σύνολο ζευγών αντικειμένων που ανήκουν στην ίδια ομάδα στο C' και όχι στο C
n_{11}: σύνολο αντικειμένων που δεν ανήκουν στην ίδια ομάδα ούτε στο C ούτε στο C'

Οι τέσσερις κατηγορίες αυτές μπορούν να υπολογιστούν από τον πίνακα σύγχυσης και ικανοποιούν πάντα την συνθήκη: $n_{00} + n_{01} + n_{10} + n_{11} = \binom{n}{2}$, όπου n το σύνολο των στοιχείων που ανήκουν σε μία ομαδοποίηση.[95] [97]

Ο ***Rand Index*** μετράει το επίπεδο συμφωνίας μεταξύ των δύο ομαδοποιήσεων σαν ένα κλάσμα με αριθμητή το σύνολο των ζευγών στοιχείων που συμφωνούν προς όλα τα πιθανά ζευγάρια. Υπολογίζεται από τον τύπο:

$$RI(C, C') = \frac{n_{00} + n_{11}}{n_{00} + n_{11} + n_{01} + n_{10}}$$

Η αναμενόμενη τιμή της μετρικής είναι από το 0 (κανένα ζευγάρι στοιχειών δεν ανήκει στην ίδια ομάδα και στις δυο ομαδοποιήσεις) έως το 1 (οι ομαδοποιήσεις είναι ίδιες) [98] [99]. Όπως αναφέρθηκε στην μελέτη των Fowlkes and Mallows [100] , για μεγάλο αριθμό ομάδων η τιμή της μετρική πλησιάζει το 1 ακόμη και στην περίπτωση ανεξάρτητων ομαδοποιήσεων. Για αυτό τον λόγο, στο VICTOR χρησιμοποιήθηκε μόνο η μετρική Adjusted Rand Index.

Ο ***Adjusted Rand Index*** αποτελεί γειτνίαση της μετρικής Rand Index και προτάθηκε από τους Hubert and Arabie [101] ως λύση για το παραπάνω πρόβλημα και συμβολίζεται ως:

$$ARI(C, C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{mij}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$ όπου: $t_1 = \sum_{i=1}^{k}\binom{|C_i|}{2}, t_2 = \sum_{j=1}^{l}\binom{|C'_j|}{2}, t_3 = \frac{2\, t_1\, t_2}{n(n-1)}$ [97]

Αναλυτικότερα, ο Adjusted Rand Index αποτελεί κανονικοποίηση της διαφοράς της τιμής της Rand Index και της αναμενόμενης τιμής υπό την μηδενική υπόθεση. Η μηδενική υπόθεση περιγράφει ανεξαρτησία των ομαδοποιήσεων βασισμένη στην υπόθεση ότι ο Confusion Matrix κατασκευάζεται από την υπεργεωμετρική κατανομή, κατά την οποία αντικείμενα αναθέτονται τυχαία στον αρχικό αριθμό των ομάδων της κάθε ομαδοποίησης. Ο Adjusted Rand Index κυμαίνεται από $[-1, 1]$ και ισούται με -1 στην περίπτωση που η αναμενόμενη τιμή είναι

μεγαλύτερη της Rand Index. Η τιμή 0 δείχνει ανεξάρτητες μεταξύ τους ομαδοποιήσεις και η τιμή 1 όμοιες μεταξύ τους. [101]

Ο ***Jaccard Index*** ισούται με το σύνολο των στοιχείων που ανήκουν στην ίδια ομάδα και στο C και στοC' δια το σύνολο των ζευγών εκτός από αυτά που ανήκουν σε διαφορετικές ομάδες και στο Cκαι στο C'. Υπολογίζεται από τον τύπο: $J(C, C') = \frac{n_{00}}{n_{00}+n_{01}+n_{10}}$[97] και οι τιμές του κυμαίνονται από $[0, 1]$. Υψηλές τιμές δείχνουν μεγαλύτερη συμφωνία ανάμεσα στις ομαδοποιήσεις.

Η ***Overlap Coefficient*** ορίζεται ως η τομή των δύο ομαδοποιήσεων προς το μικρότερο μέγεθος αυτών. Η Overlap Coefficient κυμαίνεται από [0,1] και υπολογίζεται ως:

$$Overlap(C, C') = \frac{|C \cap C'|}{min(|C|,|C'|)} = \frac{n_{00}}{min((n_{00}+n_{01}), (n_{00}+n_{10}))}$$[96] [102]

Τα ***Wallace Criterion Type I*** και ***Wallace Criterion Type II*** είναι δύο μη συμμετρικά κριτήρια που προτάθηκαν από τον Wallace και εκφράζουν την πιθανότητα ένα ζευγάρι στοιχείων που ανήκει στην ίδια ομάδα στην ομαδοποίηση C να ανήκει στην ίδια ομάδα και στην C' για το κριτήριο τύπου I, ενω το αντίστροφο ισχύει για το κριτήριο τύπου II. Τα κριτήρια αυτά υπολογίζονται από τους τύπους:

$W_I(C, C') = \frac{n_{00}}{\sum_k n_k(n_k-1)/2}$ και $W_{II}(C, C') = \frac{n_{00}}{\sum_l n_l(n_l-1)/2}$, με n_k και n_lτα μεγέθη των ομάδων. [95]

[103]

Ο ***Fowlkes-Mallows Index*** αποτελεί συμμετρική μετρική και μπορεί να χαρακτηριστεί ως ο γεωμετρικός μέσος των δύο κριτηρίων Wallace W_I και W_{II}. Υπολογίζεται από τον τύπο:

$$F(C, C') = \sqrt{W_I(C, C')W_{II}(C, C')} = \frac{n_{00}}{\sqrt{(n_{00}+n_{01})(n_{00}+n_{10})}}$$και κυμαίνεται από $[0, 1]$.[95] [97] [100]

Η ***Chi-Squared Coefficient*** χαρακτηρίζεται από τον τύπο:

$$X(C, C') = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(m_{ij}-E_{ij})^2}{E_{ij}}$$ όπου $E_{ij} = \frac{|C_i|\,|C'_j|}{n}$ [97]

Χρησιμοποιείται για να υπολογιστεί η ανεξαρτησία μεταξύ δύο ομαδοποιήσεων. Καθώς, η Chi-Squared Coefficient δεν αποτελεί κανονικοποιημένη μετρική, κυμαίνεται σε ένα πάνω όριο που κάθε φορά εξαρτάται από το σύνολο των στοιχείων που ανήκουν σε μία ομαδοποίηση.

Η ***Mirkin Metric*** αυτή αντιστοιχεί στην απόσταση Χάμινγκ (Hamming distance) για δυαδικά διανύσματα, εάν το σύνολο όλων των ζευγαριών απαριθμείται και η ομαδοποίηση αναπαρίσταται από ένα δυαδικό διάνυσμα που ορίζεται από την απαρίθμηση αυτή.

$$M(C, C') = \sum_{i=1}^{k}|C_i|^2 + \sum_{j=1}^{l}|C'_j|^2 - 2\sum_{i=1}^{k}\sum_{j=1}^{l}m_{ij}^2$$

Η μετρική αυτή είναι ευαίσθητη στο μέγεθος των ομάδων ανάμεσα στις ομαδοποιήσεις. Συγκεκριμένα, δύο ομαδοποιήσεις που έχουν ίδιο αριθμό στοιχείων στην κάθε ομάδα μπορούν να υπολογιστούν πιο όμοιες από δύο ομαδοποιήσεις που η μια αποτελεί βελτίωση της άλλης [97]. Για αυτό τον λόγο καθώς και για την χρήση δυαδικών διανυσμάτων η μετρική αυτή δεν αξιοποιήθηκε στο VICTOR.

Η ***Partition Difference*** μετράει απλά τον αριθμό των ζευγών στοιχείων που ανήκουν σε διαφορετικές ομάδες και στις δύο ομαδοποιήσεις.

$$P(C, C') = n_{00}$$ [97]

Και αυτή η μετρική είναι ευαίσθητη στο μέγεθος των ομάδων και τον αριθμό τους. Γνωρίζοντας απλά το σύνολο των ζευγών στοιχείων που δεν ανήκουν στην ίδια ομάδα και στο C και στο C' δεν μπορούμε να κρίνουμε την ομοιότητα των δύο ομαδοποιήσεων, για αυτό τον λόγο δεν χρησιμοποιείται στο VICTOR.

Η ***Sørensen–Dice Coefficient*** εφαρμόζεται σε δυαδικά δεδομένα και ορίζεται ως:

$$DSC = \frac{2\,n_{00}}{2n_{00} + n_{01} + n_{10}}$$ [96]

Η διαφορά της Sorensen-Dice Coefficient από το Jaccard Index είναι ότι από αυτόν υπολογίζεται μόνο μια φορά τα n_{00} τόσο στον αριθμητή όσο και στον παρονομαστή. Η Sorensen-Dice Coefficient δεν ικανοποιεί την τριγωνική ανισότητα, επομένως δεν αποτελεί μετρική.

Η ***Simple Matching Coefficient*** χρησιμοποιείται για να υπολογιστεί η ομοιότητα μεταξύ δύο αντικειμένων με n δυαδικά στοιχεία και ορίζεται από τον τύπο:

$$SMC = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}}$$ [96]

Ο ***Tversky Index*** αποτελεί μια γενίκευση της Sorensen-Dice Coefficient και του Tanimoto Index

$$\mathrm{T} = \frac{n_{00}}{n_{00} + a\,n_{10} + \beta\,n_{11}}$$ [96]

Ο ***Tanimoto Index*** αποτελεί μια ειδική περίπτωση του Tversky Index με α,β=1 και ουσιαστικά ισούται με την μετρική Jaccard Index.

Οι Sorensen-Dice Coefficient, Simple Matching Coefficient, Tversky Index και Tanimoto Index δεν αξιοποιούνται από το VICTOR.

5.2.2 Κατηγορία 2: Αλληλοεπικάλυψη των σετ/Αντιστοίχιση (Set Overlaps/Matching)

Η ***F-Measure*** χρησιμοποιείται για να αξιολογηθεί η ακρίβεια μιας ομαδοποίησης με μια βέλτιστη λύση. Η F-Measure για μια ομάδα C'_j και C_i δείχνει πόσο καλά η ομάδα C'_j περιγράφει την C_i.

Ουσιαστικά, είναι ο αρμονικός μέσος όρος της ακρίβειας (precision) $p_{ij} = \frac{m_{ij}}{|C_j^{'}|}$ και της ανάκλησης (recall) $r_{ij} = \frac{m_{ij}}{|C_i|}$ και υπολογίζεται ως εξής:

$$F(C_i, C_j^{'}) = \frac{2\, p_{ij} r_{ij}}{p_{ij} + r_{ij}}$$

Η συνολική τιμή της F-Measure χαρακτηρίζεται ως το άθροισμα με βάρη της μεγαλύτερης τιμής της F-Measure για τις ομάδες της ομαδοποίησης C':

$$F(C,\ C') = F(C') = \frac{1}{n}\sum_{i=1}^{k} n_i max_{j=1}^{l} \{F(C_i, C_j^{'})\}$$[97]

Η μετρική είναι μη συμμετρική, με αποτέλεσμα να είναι δύσκολη η χρήση της στη σύγκριση δύο ομαδοποιήσεων μη γνωρίζοντας την βέλτιστη λύση.

Η ***Meila-Heckerman Measure*** είναι μια μη συμμετρική μετρική που προτάθηκε από τους Meila και Heckerman [104] για την σύγκριση αλγορίθμων ομαδοποιήσεων. Οι ομαδοποιήσεις δεν συγκρίνονται μεταξύ τους, αλλά κάθε μια συγκρίνεται με μια βέλτιστη λύση. Για τις συγκρίσεις αυτές χρησιμοποιείτε ο παρακάτω τύπος :

$$MH(C,\ C') = \frac{1}{n}\sum_{i=1}^{k} max_{c_i^{'} \epsilon C} m_{ij}$$ [97]

Το C είναι η ομαδοποίηση που προήλθε από τον αλγόριθμο και C'η βέλτιστη λύση. Η ασυμμετρία της μετρικής κάνει δύσκολη την χρήση της για σύγκριση δύο ομαδοποιήσεων. Ωστόσο, μπορεί να γενικευθεί σε μια συμμετρική μετρική τη Maximum-Match Measure που περιγράφεται παρακάτω.

Στη ***Maximum-Match Measure*** για κάθε ομάδα του C ανατίθενται μια "καλύτερη αντιστοίχιση " στοC'. Η μεγαλύτερη είσοδος του πίνακα σύγχυσης (M) θεωρείται ως η καλύτερη αντιστοίχηση των ομάδων C_a και C_b. Η a γραμμή και η b στήλη του πίνακα αντιστοιχίζονται και έπειτα διαγράφονται μέχρι το μέγεθος του πίνακα να είναι μηδέν. Η Maximum-Match Measure ισούται με το άθροισμα των αντιστοιχιών διαιρούμενο με το σύνολο των στοιχείων που ανήκουν σε μία ομαδοποίηση και ορίζεται ως εξής:

$$MM(C, C') = \frac{1}{n}\sum_{i=1}^{min\{k,l\}} m_{ij}$$[97], όπου kκαι l είναι τα μεγέθη των ομαδοποιήσεων.

Οι τιμές της κυμαίνονται από$[0, 1]$.

Η ***Van Dongen Measure*** είναι μια συμμετρική μετρική που προτάθηκε από τον Van Dongen [105] και βασίζεται στη μέγιστη τομή των δυο ομαδοποιήσεων. Ορίζεται ως εξής :

$$D(C, C') = 2n - \sum_{i=1}^{k} max_j m_{ij} - \sum_{j=1}^{l} max_i m_{ij}$$[97].

Ένα αρνητικό της μετρικής αυτής είναι ότι αγνοεί τις ομάδες των ομαδοποιήσεων εκτός της τομής. Επιπλέον, η μετρική δεν ορίζεται από σταθερά όρια και επομένως δεν χρησιμοποιείται από το VICTOR.

5.2.3 Κατηγορία 3: Θεωρία Πληροφοριών (Information theory)

Όλες οι μετρικές σε αυτή την κατηγορία βασίζονται στην εντροπία ή μέση πληροφορία και στην πιθανότητα ένα στοιχείο να ανήκει σε μια συγκεκριμένη ομάδα. Η εντροπία που αναφέρεται σε μια ομαδοποίηση C υπολογίζει την αβεβαιότητα για την ομάδα που ανήκει ένα τυχαία επιλεγμένο στοιχείο και ορίζεται ως εξής:

$$H(C) = -\sum_{i=1}^{k} P(i)\, log_2 P(i)$$ [95]

Οπού k είναι ο συνολικός αριθμός των ομάδων στην ομαδοποίηση C και $P(i) = \frac{|Ci|}{n}$ είναι η πιθανότητα ένα τυχαία επιλεγμένο στοιχείο να ανήκει στην ομάδα $C_i \in C$. Στη συνέχεια περιγράφεται η ***Αμοιβαία Πληροφορία (Mutual Information)*** μεταξύ δύο ομαδοποιήσεων, η οποία υπολογίζει πόσο μπορούμε κατά μέσο όρο να μειώσουμε την αβεβαιότητα για την ομάδα ενός τυχαία επιλεγμένου στοιχείου αν γνωρίζουμε την ομάδα που ανήκει στην άλλη ομαδοποίηση του ίδιου σετ δεδομένων (ουσιαστικά την πληροφορία που έχει μια ομαδοποίηση για την άλλη). Υπολογίζεται ως:

$$I(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i,j) log_2 \frac{P(i,j)}{P(i)P(j)}$$ [97], όπου$P(i, j) = \frac{m_{ij}}{n}$

$P(i,j)$ είναι η πιθανότητα ένα στοιχείο που έχει ομαδοποιηθεί στην ομάδα C_i στη C να έχει ομαδοποιηθεί και στην C_j στη C'. Η Αμοιβαία Πληροφορία δεν οριοθετείται από σταθερή τιμή, με αποτέλεσμα να περιορίζεται η χρήση της. Το όριο της μετρικής μεταξύ δύο ομαδοποιήσεων καθορίζεται κάθε φορά από τις εντροπίες τους $I(C, C') \leq min\{H(C), H(C')\}$. Στο VICTOR χρησιμοποιούνται οι κανονικοποιημένες εκδοχές της.

Η ***Normalized Mutual Information by Strehl and Ghosh*** [106] αποτελεί μια από τις κανονικοποιήσεις της Αμοιβαίας Πληροφορίας και ορίζεται ως κλάσμα της Αμοιβαίας Πληροφορίας $I(C, C')$ με το γεωμετρικό μέσο όρο των εντροπιών των δύο ομαδοποιήσεων. Ορίζεται ως:

$$NMI_{SG}(C, C') = \frac{I(C,C')}{\sqrt{H(C)H(C')}}$$

Κυμαίνεται από $[0, 1]$, με $NMI_{SG}(C, C') = 1$ όταν $C = C'$ και $NMI_{SG}(C, C') = 0$ αν $P(i,j) = 0$ ή $P(i,j) = P(i)\,P(j)$.

Η ***Normalized Mutual Information by Fred and Jain*** [107] είναι άλλη μια κανονικοποίηση της Αμοιβαίας Πληροφορίας που υπολογίζει την Αμοιβαία Πληροφορία των δύο ομαδοποιήσεων προς το σύνολο των εντροπιών τους. Υπολογίζεται ως :

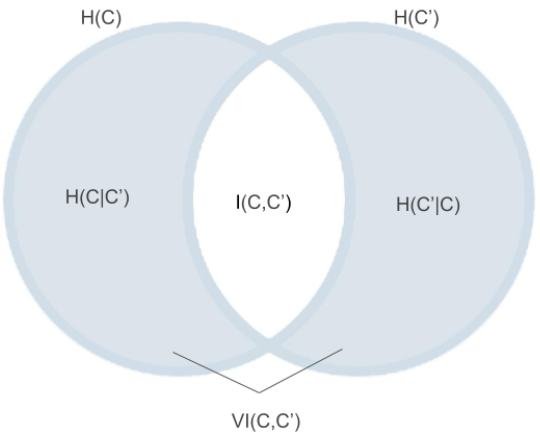
$$NMI_{FJ}(C,\ C') = \frac{2I(C,C')}{H(C) + H(C')}$$

Όπως προηγουμένως $0 \leq NMI_{FJ}(C, C') \leq 1$.[97]

Η ***Variation of Information*** είναι μια μετρική που περιγράφει το ποσοστό της πληροφορίας που χρειάζεται να προστεθεί όταν μεταβαίνουμε από την ομαδοποίηση C στην ομαδοποίηση C', καθώς επίσης και το ποσοστό της πληροφορίας που χάνουμε από την ομαδοποίηση C.[95] Υπολογίζεται ως :

$$VI(C, C') = H(C) + H(C') - 2\,I(C, C')$$

Από την εικόνα (Εικόνα 11) έχουμε $VI(C, C') = H(C|C') + H(C'|C)$. Οι $H(C|C') = H(C) - I(C, C')$ και $H(C'|C) = H(C') - I(C, C')$ αποτελούν υπο-συνθηκη εντροπίες, με την πρώτη να περιγράφει το ποσοστό της πληροφορία της ομαδοποίησης C που χάνουμε και την δεύτερη την πληροφορία που χρειαζεται να αποκτίσουμε για το C'.



Εικόνα 11: Η μετρική Variation of Information

Ομοίως και η Variation of Information δεν οριοθετείτε από σταθερή τιμή και για αυτό τον λόγο στο VICTOR χρησιμοποιείται η κανονικοποιημένη μορφή της.

Η ***Normalized Variation of Information*** υπολογίζει την μέση απουσία πληροφορίας που αναφέρεται στο C και ομοίως στο C'. Η μετρική αυτή ορίζεται από τον τύπο:

$NVI(C, C') = \frac{1}{2}\left(\frac{H(C, C') - H(C')}{H(C)} + \frac{H(C, C') - H(C)}{H(C')}\right)$, με $H(C, C')$ η από-κοινού εντροπία (joint entropy)

Η Normalized Variation of Information κυμαίνεται από $[0, 1]$ με τη τιμή 0 να περιγράφει όμοιες ομαδοποιήσεις. Στο VICTOR η μετρική αυτή αντιστράφηκε αφαιρώντας όλες τις τιμές από την τιμή 1, έτσι ώστε η χρωματικές απεικονίσεις των οπτικοποιήσεων να συμβαδίζουν με τις υπόλοιπες μετρικές.

Στο VICTOR επιλέχθηκαν να χρησιμοποιηθούν μόνο δέκα από τις 25 μετρικές που περιγράφηκαν παραπάνω. Όπως είδαμε κάποιες από τις μετρικές είναι ευαίσθητες σε κάποιες παραμέτρους όπως το μέγεθός της ομάδας ή τον συνολικό αριθμό των ομάδων που περιέχονται σε μια ομαδοποίηση. Πρόβλημα αποτελεί επίσης και ότι πολλές από αυτές δεν λαμβάνουν υπόψη στον υπολογισμό της τιμής το σύνολο των αντικειμένων που δεν ανήκουν στην ίδια ομάδα στις ομαδοποιήσεις. Συγκεκριμένα, επιλέχθηκαν οι μετρικές των οποίο τα όρια είναι γνωστά και κυμαίνονται από το μηδέν έως το ένα. Οι μετρικές των οποίων η τιμή για την ανάδειξη των πιο όμοιων ομαδοποιήσεων δεν είναι σταθερή απορρίφθηκαν αμέσως και διατηρήθηκαν μόνο αυτές με τις οποίες μπορούμε να ξεχωρίσουμε την καλύτερη ομαδοποίηση όταν αυτές συγκρίνονται ανά δύο. Οι επιλογές των μετρικών έγινα έτσι ώστε το VICTOR να αποτελεί μια καθολική εφαρμογή στην οποία μπορούν να συγκριθούν τόσο αποτελέσματα ομαδοποιήσεων διαφορετικών αλγορίθμων όσο και αποτελέσματα ομαδοποιήσεων που προέρχονται από διαφορετικά σύνολα δεδομένων. Τέλος, οι μετρικές που βασίζονται στην αμοιβαία πληροφορία φαίνεται να είναι ιδιαίτερα αποδοτικές, καθώς δεν εμφανίζουν προβλήματα στον υπολογισμό της τιμής τους. Και σε αυτή την περίπτωση επιλέχθηκε η κανονικοποιημένη μορφή τους, που κυμαίνεται από το μηδέν έως το ένα για την ανάδειξη της καλύτερης ομαδοποίησης.

6. Αρχεία εισόδου και δυναμικό φιλτράρισμα

6.1 Αρχεία εισόδου

Το VICTOR δέχεται σαν είσοδο πολλαπλά αρχεία ομαδοποιήσεων σε μορφή κειμένου. Τα αρχεία αυτά αποτελούνται από δύο στήλες χωρίς κεφαλίδες. Η πρώτη στήλη περιέχει τα ονόματα των ομάδων, συγκεκριμένα κάθε γραμμή αναφέρεται και σε μια ομάδα. Η δεύτερη στήλη έχει τα αντικείμενα των ομάδων χωρισμένα με κόμμα μεταξύ τους (Εικόνα 12). Ο χρήστης μπορεί οποιαδήποτε στιγμή να προσθέσει επιπλέων αρχεία ομαδοποιήσεων, καθώς και να μετονομάσει ή διαγράψει αυτά που έχει ανεβάσει ήδη. Μόλις τα αρχεία ανέβουν, εμφανίζονται βασικές πληροφορίες αυτών όπως ο αριθμός των ομάδων και των στοιχειών που ανήκουν σε κάθε ομάδα. Οι πληροφορίες αυτές οπτικοποιούνται με τη μορφή bar charts, ενώ ιστογράμματα δείχνουν την παρουσία των αντικειμένων στις διάφορες ομάδες .

Αξίζει να αναφερθεί ότι τα αρχεία των ομαδοποιήσεων πρέπει να πληρούν κάποιους περιορισμούς που προκύπτουν από τη χρήση της βιβλιοθήκης mclustcomp στο VICTOR. Αναλυτικότερα, κάθε αντικείμενο θα πρέπει να ανήκει μόνο σε μια ομάδα (δεν επιτρέπονται διπλότυπα). Επιπροσθέτως, όλες οι ομαδοποιήσεις που πρόκειται να συγκριθούν μεταξύ τους πρέπει να περιέχουν ακριβώς τον ίδιο αριθμό στοιχείων n. Οι χρήστες μέσω των διαγραμμάτων bar charts (Εικόνα 16) που διατίθενται μέσω της καρτέλας “διαχείριση αρχείων” (File Handling Tab) μπορούν να εντοπίσουν αν τα αρχεία των ομαδοποιήσεων προς σύγκριση έχουν διαφορετικό αριθμό στοιχείων μεταξύ τους. Στην περίπτωση που τα αρχεία των ομαδοποιήσεων έχουν το ίδιο μέγεθος και αποτελούνται από τον ίδιο αριθμό αντικειμένων, τότε αυτά μπορούν να συγκριθούν με οποιαδήποτε από τις διαθέσιμες μετρικές. Στην περίπτωση όμως που τα αρχεία των ομαδοποιήσεων έχουν διαφορετικό αριθμό στοιχείων μεταξύ τους, τότε μπορεί να εφαρμοστεί φιλτράρισμα σε αυτά με αποτέλεσμα ο αριθμός των στοιχείων των ομαδοποιήσεων να είναι ίσος. Στη συνέχεια, γίνεται αναλυτική περιγραφή των τριών επιλογών φιλτραρίσματος που παρέχονται από το VICTOR.

```
cluster1  F19,F21,F22,F27,F29,F33,F41,F42,F44,F47
cluster2  F02,F10,F13,F15,F18,F24,F25,F26,F46,F50
cluster3  F01,F12,F35,F40,F45
cluster4  F03,F04,F06,F07,F08,F17,F20,F36,F38,F43,F49
cluster5  F05,F09,F16,F28,F30,F32,F34,F39
cluster6  F11,F14,F23,F31,F37,F48
```

Εικόνα 12: Αρχείο ομαδοποίησης στη μορφή συμβατή με το VICTOR. Η πρώτη στήλη δείχνει τα ονόματα των ομάδων και η δεύτερη τα στοιχεία από τα οποία αποτελούνται.

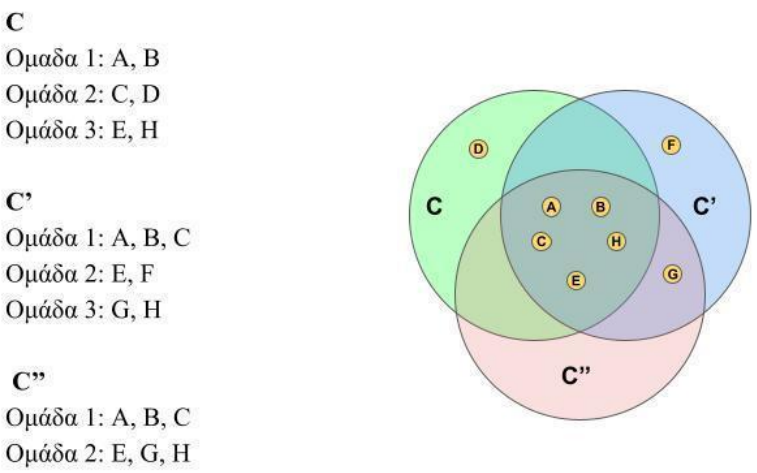
6.2 Δυναμικό φιλτράρισμα

Αρχικά, το VICTOR παρέχει τρεις διαφορετικές επιλογές φιλτραρίσματος των αρχείων των ομαδοποιήσεων έτσι ώστε αυτά να αποτελούνται από των ίδιο αριθμό αντικειμένων πριν συγκριθούν με τις διαθέσιμες μετρικές. Οι επιλογές φιλτραρίσματος είναι: i) ορισμός ορίων στον αριθμό των αντικειμένων των ομαδοποιήσεων ii) φιλτράρισμα με τομή iii) φιλτράρισμα με υπερύνολο. Πιο συγκεκριμένα ο αριθμός των στοιχείων που περιέχονται στα αρχεία μπορεί να προσαρμοστεί με την βοήθεια ορίων που ορίζονται από τον χρήστη, κρατώντας τα μεγέθη των

αρχείων εντός συγκεκριμένων τιμών. Κατά αυτών των τρόπο μονά στοιχεία μπορούν να περιοριστούν στις ομαδοποιήσεις ή να αφαιρεθούν ασαφής ομάδες με πολλά στοιχεία.

Διατίθενται και δύο επιπλέον δυνατότητες φιλτραρίσματος στο VICTOR, έτσι ώστε τα αρχεία των ομαδοποιήσεων να έχουν ίδιο μέγεθος. Αυτές είναι είτε η διατήρηση των στοιχείων της τομής των επιλεγμένων αρχείων, είτε η δημιουργία ενός σουπερ-σετ προσθέτοντας μια ομάδα που αποτελείται από ένα μονό στοιχείο σε κάθε ομαδοποίηση που δεν περιέχει αυτό το στοιχείο. Η επιλογή τομή (intersection) διατηρεί απλώς τα στοιχεία που υπάρχουν ήδη μεταξύ των ομαδοποιήσεων, ενώ η επιλογή σουπερ-σετ δημιουργεί ομάδες μονών στοιχείων για οποιοδήποτε από τα στοιχεία που δεν εντοπίζονται στις ομαδοποιήσεις.

Για να γίνει πιο κατανοητή η χρήση αυτών των δύο επιλογών, θα γίνει περιγραφή τους στην συνέχεια με την βοήθεια ενός παραδείγματος. Ας υποθέσουμε τρεις C,C' και C'' ομαδοποιήσεις με διαφορετικό αριθμό ομάδων και στοιχείων που περιέχονται στις ομάδες όπως φαίνεται στην παρακάτω εικόνα (Εικόνα 13).



Εικόνα 13: Οι τρεις ομαδοποιήσεις C,C' και C'' με τα στοιχεία που περιέχονται σε αυτές.

Αν στα αρχεία των ομαδοποιήσεων εφαρμοστεί η επιλογή τομή ώστε να έχουν το ίδιο μέγεθος, τότε αυτά θα αποτελούνται εν τέλη μόνο από τα στοιχεία που μοιράζονται. Στην παρακάτω εικόνα (Εικόνα 14) φαίνεται πως έχουν διαμορφωθεί τα αρχεία των ομαδοποιήσεων C, C' και C'' μετά το φιλτράρισμα. Ουσιαστικά και οι τρεις ομαδοποίησης αποτελούνται από τα στοιχεία A, B, C, E και H που ανήκουν στην τομή τους και τα D, F και G δεν εντοπίζονται σε καμία από τις νέες ομάδες. Ωστόσο τα στοιχεία της τομής της κάθε ομαδοποίησης ανήκουν σε διαφορετική ομάδα. Στην ομαδοποίηση C η ομάδα 1 αποτελείται από δύο στοιχεία τα A και B και η ομάδα 2 από τα C, E και H. Στην ομαδοποίηση C' η ομάδα 1 αποτελείται από τα στοιχεία A, E και H και η ομάδα 2 από τα B και C. Τέλος, στην ομαδοποίηση C'' η ομάδα 1 αποτελείται από δύο στοιχεία τα C και H και η ομάδα 2 από τα A, B και E.

C	Στοιχείο	A	B	C	E	H
	Ομάδα	1	1	2	2	2
C'	Στοιχείο	A	B	C	E	H
	Ομάδα	1	2	2	1	1
C''	Στοιχείο	A	B	C	E	H
	Ομάδα	2	2	1	2	1

Εικόνα 14: Πίνακας με τις ομαδοποιήσεις μετά την εφαρμογή της τομής ως φίλτρο.

Αν στα αρχεία των ομαδοποιήσεων εφαρμοστεί η επιλογή του σουπερ-σετ, τότε σε κάθε ομαδοποίηση θα προστεθούν επιπλέον ομάδες με ένα από τα στοιχεία που δεν περιλαμβάνονταν ήδη σε αυτά. Όπως βλέπουμε και στην παρακάτω εικόνα (Εικόνα 15) στην ομαδοποίηση C έχουν προστεθεί δύο επιπλέον ομάδες, η ομάδα 3 και η ομάδα 4 (εντόνα υπογραμμισμένο) με τα στοιχεία F και G , στην ομαδοποίηση C' προστέθηκε η ομάδα 3 με το στοιχείο D και στην ομαδοποίηση C'' δημιουργήθηκαν δύο ομάδες με τα στοιχεία D και F. Το αποτέλεσμα του φιλτραρίσματος με την επιλογή υπερ-σύνολο είναι και οι τρεις ομαδοποιήσεις πλέον αποτελούνται από το ίδιο αριθμό στοιχείων.

C	Στοιχείο	A	B	C	D	E	F	G	H
	Ομάδα	1	1	2	1	2	**3**	**4**	2
C'	Στοιχείο	A	B	C	D	E	F	G	H
	Ομάδα	1	2	2	**3**	1	1	1	1
C''	Στοιχείο	A	B	C	D	E	F	G	H
	Ομάδα	2	2	1	**3**	2	**4**	2	1

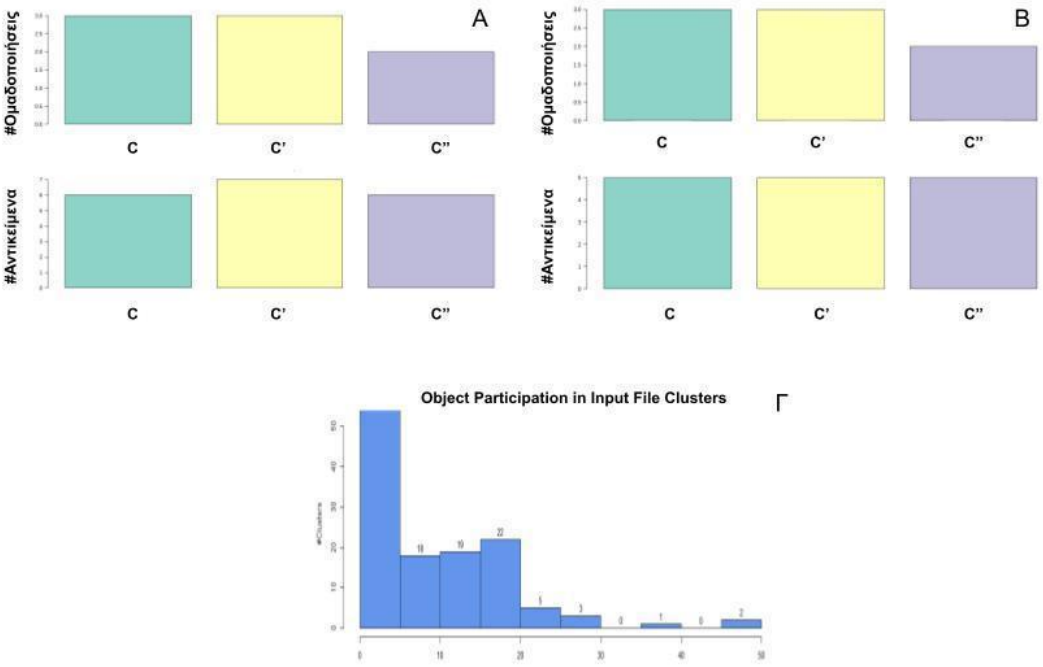
Εικόνα 15: Πίνακας με τις ομαδοποιήσεις μετά την εφαρμογή του σουπερ-σετ φίλτρου.

6.3 Οπτικοποιήσεις του περιεχομένου των αρχείων ομαδοποιήσεων

Αξίζει να σημειωθεί ότι υπάρχουν διαθέσιμες οπτικοποιήσεις στο VICTOR και για το περιεχόμενο και το μέγεθος των αρχείων τόσο στην αρχική τους μορφή όσο και μετά την επεξεργασία. Αναλυτικότερα, στατιστικά των αρχείων όπως το σύνολο των ομάδων που περιέχονται στα αρχεία των ομαδοποιήσεων, καθώς και το σύνολο των στοιχείων που περιέχονται σε κάθε ομαδοποίηση παρουσιάζονται ως bar charts (Εικόνα 16A και 16B). Στη πρώτη εικόνα βλέπουμε ότι οι ομαδοποιήσεις από το προηγούμενο παράδειγμα αρχικά αποτελούνται από διαφορετικό αριθμό αντικειμένων στις ομαδοποιήσεις (ανομοιόμορφα bar charts) και στη δεύτερη εικόνα βλέπουμε ότι ύστερα από την εφαρμογή του φιλτραρίσματος της

τομής οι ομαδοποιήσεις πλέον αποτελούνται από τον ίδιο αριθμό στοιχείων (ομοιόμορφα bar charts).

Επιπλέον, ιστογράμματα παρουσιάζουν τη παρουσία των στοιχείων ανάμεσα στις ομάδες ενός επιλεγμένου από τον χρήστη αρχείου (Εικόνα 16Γ). Στα ιστογράμματα στο κάθετο άξονα βλέπουμε τις ομάδες των ομαδοποιήσεων και στον οριζόντιο τα αντικείμενα αυτών. Τέλος, το περιεχόμενο των αρχείων των ομαδοποιήσεων εμφανίζεται ως πίνακας (Εικόνα 17) όπου ο χρήστης μπορεί να πλοηγηθεί στο περιεχόμενο του αρχείο, να αναζητήσει κάποιο συγκεκριμένο περιεχόμενο και να αποθηκεύσει το αρχείο τόσο πριν όσο και μετά το φιλτράρισμα σε διαφορές διαθέσιμες μορφές.



Εικόνα 16: Α) Διαγράμματα με το σύνολο των ομάδων και των στοιχειών στα αρχεία των ομαδοποιήσεων πριν το φιλτράρισμα. Τα αρχεία των ομαδοποιήσεων έχουν διαφορετικό μέγεθος. Β) Διαγράμματα με το σύνολο των ομάδων και των στοιχείων των ομαδοποιήσεων μετά την εφαρμογή της τομής. Τα αρχεία των ομαδοποιήσεων έχουν το ίδιο μέγεθος. Γ) Ιστόγραμμα παρουσίας των στοιχείων σε ομαδοποίηση.

Select File:

Clustering_E.txt

Show 5 entries Copy CSV Excel PDF Print Search:

Cluster	Objects	#Objects
cluster1	F19,F21,F22,F27,F29,F33,F42,F44,F47,F41	10
cluster2	F02,F10,F13,F15,F18,F24,F25,F26,F50,F46,F12,F14	12
cluster3	F01,F35,F40,F45	4
cluster4	F03,F04,F06,F07,F08,F17,F20,F38,F43,F49,F36	11
cluster5	F05,F09,F16,F28,F30,F34,F39	7

Showing 1 to 5 of 6 entries Previous 1 2 Next

Εικόνα 17: Πίνακας με το περιεχόμενο ενός αρχείου ομαδοποίησης σε μορφή συμβατή με το VICTOR.

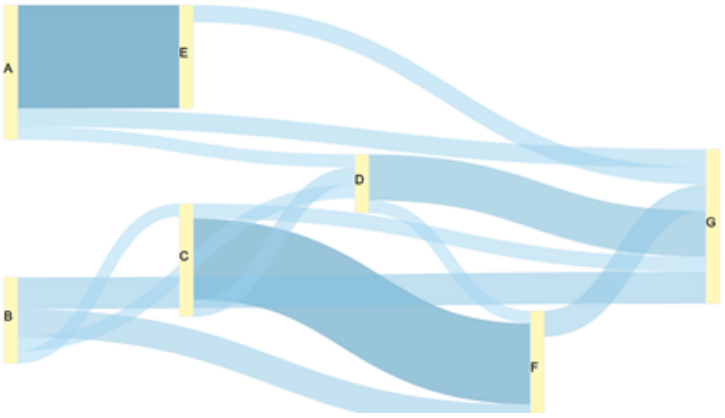
7. Οπτικοποιήσεις

Στο VICTOR διατίθενται μια πληθώρα από οπτικοποιήσεις για τη συγκριτική ανάλυση των δύο ομαδοποιήσεων. Δίνεται η δυνατότητα στο χρήστη επιλογής των αρχείων των ομαδοποιήσεων προς οπτικοποίηση καθώς και η επιλογή μιας από τις δέκα διαθέσιμες μετρικές. Με την βοήθεια της τιμής της μετρικής που επιλέχθηκε, οι χρήστες μπορούν να δουν μέσω των οπτικοποιήσεων ποια από τις ομαδοποιήσεις είναι καλύτερη σε σχέση με τις υπόλοιπες. Όλες οι διαθέσιμες οπτικοποιήσεις είναι διαδραστικές δίνοντας την δυνατότητα στον χρήστη να βλέπει ποιες οπτικοποιήσεις συγκρίνονται μεταξύ τους και τι τιμή έχει η μετρική κάθε φορά. Επιπλέον, σε κάθε οπτικοποίηση διατίθεται ολισθητής προσαρμογής της κλίμακας που κυμαίνεται η τιμή της επιλεγμένης μετρικής. Ουσιαστικά, μπορεί να οριστεί ένα πάνω και κάτω όριο, ώστε να απορρίπτονται αδύναμες συγκρίσεις των ομαδοποιήσεων και να βλέπουμε μόνο τις συγκρίσεις αυτών που είναι πιο όμοιων μεταξύ τους. Έπειτα, τα επιλεγμένα αρχεία των ομαδοποιήσεων μπορούν να συγκριθούν με την βοήθεια των οπτικοποιήσεων Sankey plots, Hierarchical Heatmap, Bar charts, networks και Circos plots.

7.1 Sankey plots

Γενικότερα, τα ***Sankey plots*** (Εικόνα 18) περιγράφουν διαγράμματα ροής όπου το πλάτος μιας γραμμής είναι ανάλογο της ροής. Κυρίως χρησιμοποιούνται για την μελέτη συστημάτων ροής ενέργειας [108] [109]. Ωστόσο, μπορούν να εφαρμοστούν και στον τομέα της Βιολογίας για την αναπαράσταση, για παράδειγμα, ομάδων κινδύνου έναντι υποτύπων κλινικού καρκίνου [110] ή ως οτικοποίηση σε άλλα λογισμικά όπως το LiveKraken το οποίο χρησιμοποιείται για ταξινομική ομαδοποίηση [111].

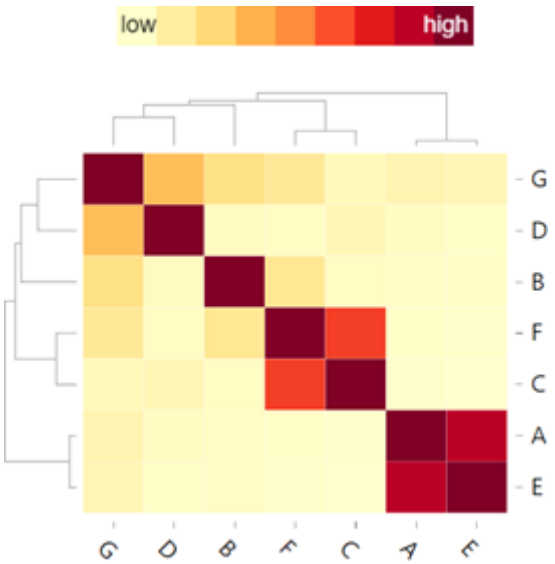
Συγκεκριμένα, στα Sankey plots στο VICTOR τα κάθετα παραλληλόγραμμα αναπαριστούν συγκρίσεις ομαδοποιήσεων, τα οποία συνδέονται με μη κατευθυνόμενες γραμμές. Το μέγεθος της γραμμής αντανακλά την ομοιότητα των δυο ομαδοποιήσεων σε σχέση με τις υπόλοιπες συγκρίσεις. Όσο μεγαλύτερο είναι το μέγεθος της γραμμής τόσο μεγαλύτερη είναι η τιμή της επιλεγμένης μετρικής για την σύγκριση των ομαδοποιήσεων. Όπως βλέπουμε και στην εικόνα (Εικόνα 18) οι ομαδοποιήσεις (A, B, C, D, E, F και G) ενώνονται όλες ανά δύο με γραμμές και το πάχος της γραμμής υποδηλώνει για παράδειγμα ότι οι ομαδοποιήσεις A και E είναι πιο όμοιες μεταξύ τους, καθώς και οι ομαδοποιήσεις C και F.



Εικόνα 18: Sankey plot. Οι ομαδοποιήσεις αναπαρίστανται από τις κίτρινες γραμμές που συνδέονται με μπλε ακμές.

7.2 Hierarchical Heatmap

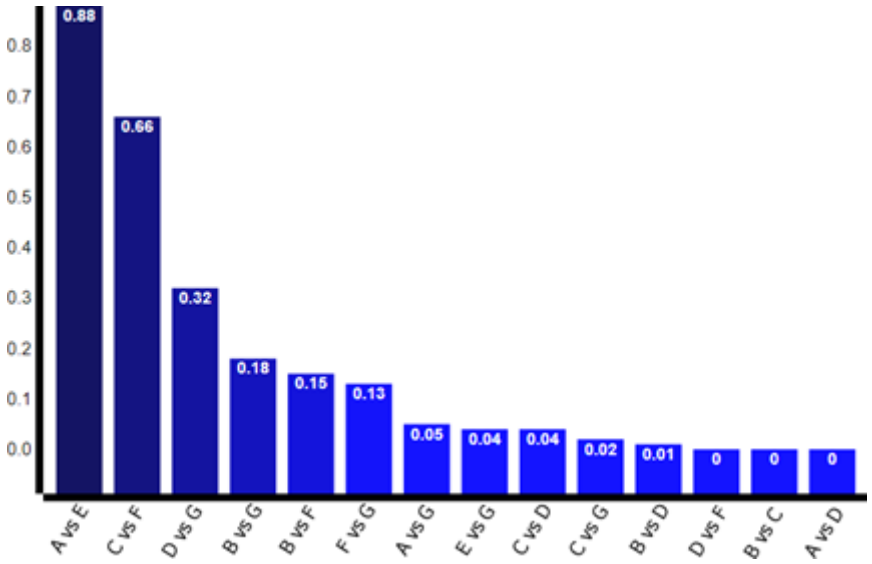
Το ***Hierarchical Heatmap*** ομαδοποιεί αντικείμενα τόσο στον x όσο και στο y άξονα με βάση την ένταση των χρωμάτων. Η ένταση του χρώματος δείχνει την ποιότητα της σύγκρισης μεταξύ των ομαδοποιήσεων. Αναλυτικότερα, το Hierarchical heatmap (Εικόνα 19) που χρησιμοποιείται στο VICTOR έχει την μορφή ενός $n \times n$ πίνακα, όπου n είναι το σύνολό των συγκρίσεων μεταξύ των ομαδοποιήσεων. Όσο πιο έντονα χρωματικά είναι ένα κελί τόσο μεγαλύτερη είναι η ομοιότητα μεταξύ των ομαδοποιήσεων. Τα χρώματα εφαρμόζονται στα κελιά με βάση την κλίμακα των τιμών της μετρικής που χρησιμοποιήθηκε για την σύγκριση των ομαδοποιήσεων. Έτσι με βαθύ κόκκινο αναδεικνύονται ομαδοποιήσεις με μεγαλύτερη ομοιότητα μεταξύ τους και με κίτρινο ομαδοποιήσεις με λιγότερη ομοιότητα. Επιπλέον, ένα δενδρόγραμμα επεκτείνεται και από τους δύο άξονες αναπαριστώντας τη διάταξη των ομάδων των ενδιάμεσων ζευγαριών σύγκρισης [112]. Το Hierarchical Heatmap είναι διαδραστικό, ο χρήστης δηλαδή μπορεί σε κάθε κελί να βλέπει την τιμή της μετρικής, καθώς και τις ομαδοποιήσεις προς σύγκριση. Μέσω της εικόνας μπορούμε να δούμε ότι με έντονο κόκκινο εντοπίζουμε τις ομαδοποιήσεις A και E, οι οποίες δείχνουν να είναι πιο όμοιες μεταξύ τους.



Εικόνα 19: Hierarchical Heatmap. Με κίτρινο βλέπουμε χαμηλές τιμές της μετρικής Adjusted Rand Index και με μπορντό υψηλές τιμές.

7.3 Bar charts

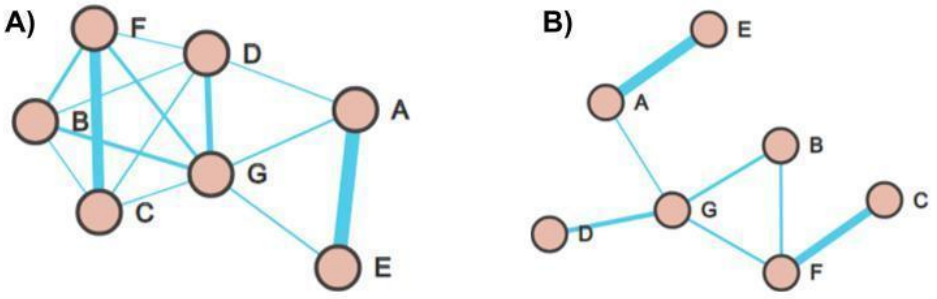
Τα ***Bar charts*** [113] αναπαριστούν δεδομένα ως ορθογώνια παραλληλόγραμμα, όπου το ύψος αυτών αντανακλά την παρατηρούμενη τιμή στον y άξονα. Τα αναπαριστομένα δεδομένα επισημαίνονται στον x άξονα. Στο VICTOR τα ορθογώνια παραλληλόγραμμα αναπαριστούν συγκρίσεις μεταξύ ομαδοποιήσεων με βάση την επιλεγμένη μετρική και αν δεν εφαρμοστεί κάποιο όριο από τον χρήστη τότε $\frac{n!}{2!(n-2)!}$ παραλληλόγραμμα εμφανίζονται, με n ο αριθμός των ομαδοποιήσεων. Σε κάθε παραλληλόγραμμο αναγράφεται η τιμή της επιλεγμένης μετρικής προς σύγκριση, στον οριζόντιο άξονα βλέπουμε σε ποιες ομαδοποιήσεις αναφέρεται κάθε παραλληλόγραμμο και στον κάθετο μια κλίμακα από την χαμηλότερη έως την υψηλότερη τιμή της μετρικής σύγκρισης. Από την εικόνα (Εικόνα 20) βλέπουμε ότι οι ομαδοποιησεις A και E και C και F έχουν τις μεγαλύτερες τιμές στη μετρική.



Εικόνα 20: Metric Bar plot. Οι κάθετοι άξονες δείχνουν την τιμή της μετρικής Adjusted Rand Index της κάθε σύγκρισης των επτά διαφορετικών ομαδοποιήσεων (A-G).

7.4 Networks

Τα ***Δίκτυα-Networks*** χρησιμοποιούνται συχνά για την αναπαράσταση δεδομένων και των σχέσεων τους σε διάφορους τομείς. Τα παρατηρούμενα δεδομένα αναπαρίστανται σαν κόμβοι και οι τιμές μεταξύ τους ως ακμές. Στο VICTOR χρησιμοποιούνται δίκτυα με βάρη (Εικόνα 21Α) για οπτικοποίηση των ομαδοποιήσεων και των μετρικών συγκρίσεων, αφού πρώτα έχει εφαρμοστεί ένας αλγόριθμος ταξινόμησης, όπως για παράδειγμα ο Fruchterman-Reingold [114], ο Reingold-Tilford [115] και ο Davidson-Harel [116]. Αν δεν έχει επιλεχθεί κάποιο όριο από τον χρήστη τότε το δίκτυο που εμφανίζεται είναι πλήρως συνδεδεμένος. Όσο πιο υψηλό το όριο που επιλέγει ο χρήστης να εφαρμόσει τόσο πιο αραιό θα είναι το γράφημα που εμφανίζεται (Εικόνα 21Β). Αναλυτικότερα, οι κόμβοι αναπαριστούν ομαδοποιήσεις και οι ακμές την τιμή της μετρικής σύγκρισης αυτών. Επομένως, το πάχος της γραμμής αναδεικνύει και την ομοιότητα μεταξύ των ομαδοποιήσεων. Και στα δύο δίκτυα βλέπουμε ότι η ακμή που ενώνει τον κόμβο A (ομαδοποίηση A) με τον κόμβο E (ομαδοποίηση E) έχει το μεγαλύτερο πάχος και υποδηλώνει όμοιες ομαδοποιήσεις.



Εικόνα 21: Α) Δίκτυο που αναπαριστά τις συγκρίσεις επτά ομαδοποιήσεις (A-G) με την μετρική Adjusted Rand Index αφού έχει εφαρμοστεί αλγόριθμος ταξινόμησης Β) Δίκτυο στο οποίο έχει εφαρμοστεί όριο στις τιμές της Adjusted Rand Index.

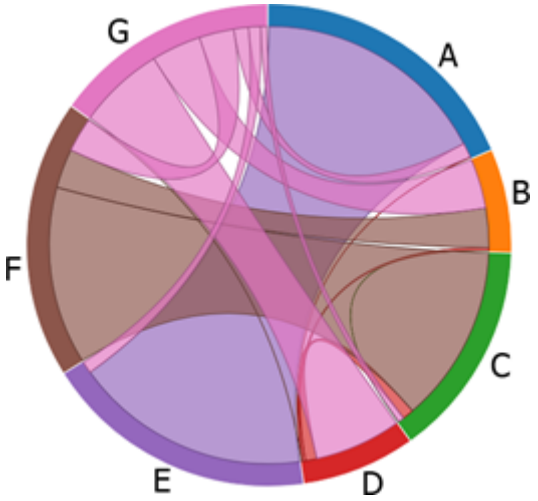
Στην οπτικοποίηση των ομαδοποιήσεων με δίκτυο, όπως αναφέρθηκε δίνεται η δυνατότητα στους χρήστες να εφαρμόσουν αλγορίθμους διάταξης για τον σχεδιασμό του δικτύου. Οι αλγόριθμοι αυτοί είναι:

- **Fruchterman-Reingold**: τοποθετεί τις κορυφές στο επίπεδο σύμφωνα με τον αλγόριθμο Fruchterman-Reingold. Ο αλγόριθμος Fruchterman-Reingold είναι ένας κατευθυνόμενος αλγόριθμος διάταξης για τον σχεδιασμό γραφημάτων. Σκοπός του είναι η τοποθέτηση των κόμβων ενός γραφήματος σε δισδιάστατο ή τρισδιάστατο χώρο έτσι ώστε όλες οι άκρες να έχουν περισσότερο ή λιγότερο ίσο μήκος και να υπάρχουν όσο το δυνατόν λιγότερες άκρες διασταύρωσης.
- **Reingold-Tilford**: εκτελεί πρώτα μια αναζήτηση πλάτους έτσι ώστε το δίκτυο να αποκτήσει ένα πιθανό δέντρο έκτασης.
- **Davidson-Harel**: σχεδιάζει το δίκτυο εφαρμόζοντας τον αλγόριθμο των Davidson και Harel. Βασίζεται στον αρχικό τρόπο λειτουργίας του αλγορίθμου διατηρώντας όμως τις συντεταγμένες εντός των ορίων του ορθογωνίου της διάταξης.
- **Circle**: είναι ένας τρόπος σχεδίασης του δικτύου τοποθετώντας τις κορυφές αυτού σε έναν κύκλο, συχνά ομοιόμορφα τοποθετημένες ώστε να σχηματίζουν κορυφές τις ενός κανονικού πολυγώνου.
- **Grid**: τοποθετεί τα στοιχεία σε ένα δισδιάστατο πλέγμα τετράγωνης μορφής.
- **Random**: για να σχεδιάσει το δίκτυο απλά τοποθετεί τυχαία τις κορυφές σε ένα τετράγωνο.
- **GEM**: τοποθετεί τις κορυφές στο επίπεδο χρησιμοποιώντας τον αλγόριθμο διάταξης κατευθυνόμενης από δύναμη GEM.
- **Graphopt**: βελτιστοποιεί τη διάταξη κορυφής μέσω του αλγορίθμου graphopt σχεδιάζοντας ετσι το δίκτυο. Το Graphopt χρησιμοποιεί φυσικές αναλογίες για τον προσδιορισμό των δυνάμεων προσέλκυσης και απώθησης μεταξύ των κορυφών και στη συνέχεια το φυσικό σύστημα προσομοιώνεται έως ότου φτάσει σε ισορροπία
- **Multidimensional Scaling**: τοποθετεί τις κορυφές σε επίπεδο χρησιμοποιώντας πολυδιάστατη κλίμακα. Αυτή η διάταξη απαιτεί έναν πίνακα απόστασης (distance matrix), όπου η τομή της γραμμής i και της στήλης j καθορίζει την επιθυμητή απόσταση μεταξύ της κορυφής i και της κορυφής j. Ο αλγόριθμος προσπαθεί να τοποθετήσει τις κορυφές σε έναν χώρο με καθορισμένη διάσταση, με τρόπο που να προσεγγίζει τις σχέσεις απόστασης που καθορίζονται στον πίνακα απόστασης.
- **Sugiyama**: προσπαθεί να εξαλείψει κύκλους και να εκχωρήσει κορυφές σε επίπεδα. Αυτός ο αλγόριθμος διάταξης έχει σχεδιαστεί για κατευθυνόμενα άκυκλα γραφήματα, όπου κάθε κορυφή αντιστοιχεί σε ένα επίπεδο. Τα επίπεδα ευρετηριάζονται από το μηδέν και οι κορυφές του ίδιου στρώματος τοποθετούνται στην ίδια οριζόντια γραμμή. Οι συντεταγμένες των κορυφών σε κάθε στρώμα αποφασίζονται από την ευρετική πρόταση των Sugiyama κ.ά. για την ελαχιστοποίηση των διασταυρώσεων των άκρων. Μπορεί επίσης να εφαρμοστεί και μη κατευθυνόμενα γραφήματα ή γραφήματα που περιέχουν κύκλους.

Όλοι οι αλγόριθμοι διάταξης προήλθαν από τη βιβλιοθήκη igraph. Η βιβλιοθήκη igraph είναι μια συλλογή βιβλιοθηκών για τη δημιουργία και χειρισμό γραφημάτων και την ανάλυση δικτύων.

7.5 Circos plots

Τα ***Circos plots*** ή ***Chord diagrams*** αναπαριστούν αντικείμενα σε κυκλική μορφή, όπου το μέγεθος των ακμών ισούται με το σύνολο των αλληλεπιδράσεων των αντικειμένων [117]. Συγκεκριμένα, στο Circos plot που εμφανίζεται στο VICTOR (Εικόνα 22) το εξωτερικό στεφάνι αποτελείται από τις ομαδοποιήσεις, οι οποίες συνδέονται μεταξύ τους με χρωματισμένες ακμές. Κάθε ακμή αποτελεί μια σύγκριση ανάμεσα στις οπτικοποιήσεις και εμφανίζεται με διαφορετικό χρώμα. Το μέγεθος των ακμών είναι ανάλογο της τιμής της επιλεγμένης μετρικής και δείχνει την ομοιότητα ανάμεσα στις ομαδοποιήσεις. Τέλος, όσο υψηλότερο είναι το όριο που εφαρμόζεται από τον χρήστη τόσο λιγότερες ακμές θα εμφανίζονται στον κύκλο. Στην παρακάτω εικόνα βλέπουμε ότι οι ομαδοποιήσεις A και E και οι C και F συνδέονται με τις πιο χοντρές ακμές, υποδηλώνοντας μεγαλύτερη τιμή στη μετρική σύγκρισης και άρα πιο όμοιες ομαδοποιήσεις. Η επιλογή των χρωμάτων που έχει η κάθε ομαδοποίηση είναι τυχαία και κάθε ακμή παίρνει το χρώμα της μια από τις δύο ομαδοποιήσεις που συνδέει.



Εικόνα 22: Circos plot. Κάθε σύγκριση μεταξύ των ομαδοποιήσεων αναπαρίσταται με διαφορετικό χρώμα. Το μέγεθος των ακμών ορίζεται από τις τιμές της μετρικής Adjusted Rand Index.

8. Conductance

Το **conductance** είναι μια μετρική που ποσοτικοποιεί πόσο καλά συνδεδεμένο είναι ένα υπογράφημα σε σχέση με το υπόλοιπο γράφημα. Δηλαδή, υπολογίζει πόσο έντονα ένα σύνολο κόμβων συνδέεται με το υπόλοιπο γράφημα. Σύνολα κόμβων που είναι απομονωμένα από το γράφημα έχουν χαμηλό conductance και σχηματίζουν καλές ομάδες. Ο υπολογισμός της μετρικής conductance ενός γραφήματος εφαρμόζεται σε διάφορους τομείς, όπως για παράδειγμα στην αξιολόγηση της ποιότητας διαφόρων αλγορίθμων ομαδοποίησης. Συγκεκριμένα, στον τομέα των δικτύων χρησιμοποιείται ευρέως για τον εντοπισμό κοινοτήτων δικτύου και τον υπολογισμό της ποιότητας των ομάδων που εντοπίζονται από αλγορίθμους ανίχνευσης κοινοτήτων [118]. Ουσιαστικά, υπολογίζει την ποιότητα της σύνδεσης μιας ομάδας με το υπόλοιπο δίκτυο, σε σχέση με τις εσωτερικές διασυνδέσεις του δικτύου. Υψηλή τιμή conductance συνεπάγεται πολλές εξωτερικές συνδέσεις, ενώ χαμηλή τιμή conductance συνεπάγεται ότι η ομάδα αυτή των κορυφών είναι εσωτερική. Αξίζει να σημειωθεί ότι το conductance ενός υπο-γράφου είναι διαφορετικό από το conductance του γράφου, το οποίο είναι και το μικρότερο.

Όπως αναφέρθηκε το conductance μπορεί να εφαρμοστεί σε διάφορα είδη γραφημάτων και δικτύων. Η εύρεση ομάδων σε βιολογικά δίκτυα έχει αποκτήσει μεγάλο ερευνητικό ενδιαφέρον. Μια ομάδα χαρακτηρίζεται από μεγαλύτερο αριθμό συνδέσεων κόμβων εντός αυτής σε σχεση με τους κόμβους του υπόλοιπου δικτύου. Επομένως, αναμένεται χαμηλότερη τιμή conductance και έτσι ελέγχεται η ποιότητα των διάφορων αλγορίθμων ανίχνευσης κοινοτήτων (ομάδων). Το conductance αποτελεί το πιο απλό μέτρο για τον εντοπισμό μιας καλής ομάδας, καθώς στηρίζεται στη συνδεσιμότητα των κόμβων της ομάδας (κόμβοι με καλύτερη εσωτερική και όχι εξωτερική συνδεσιμότητα).

8.1 Conductance στο VICTOR

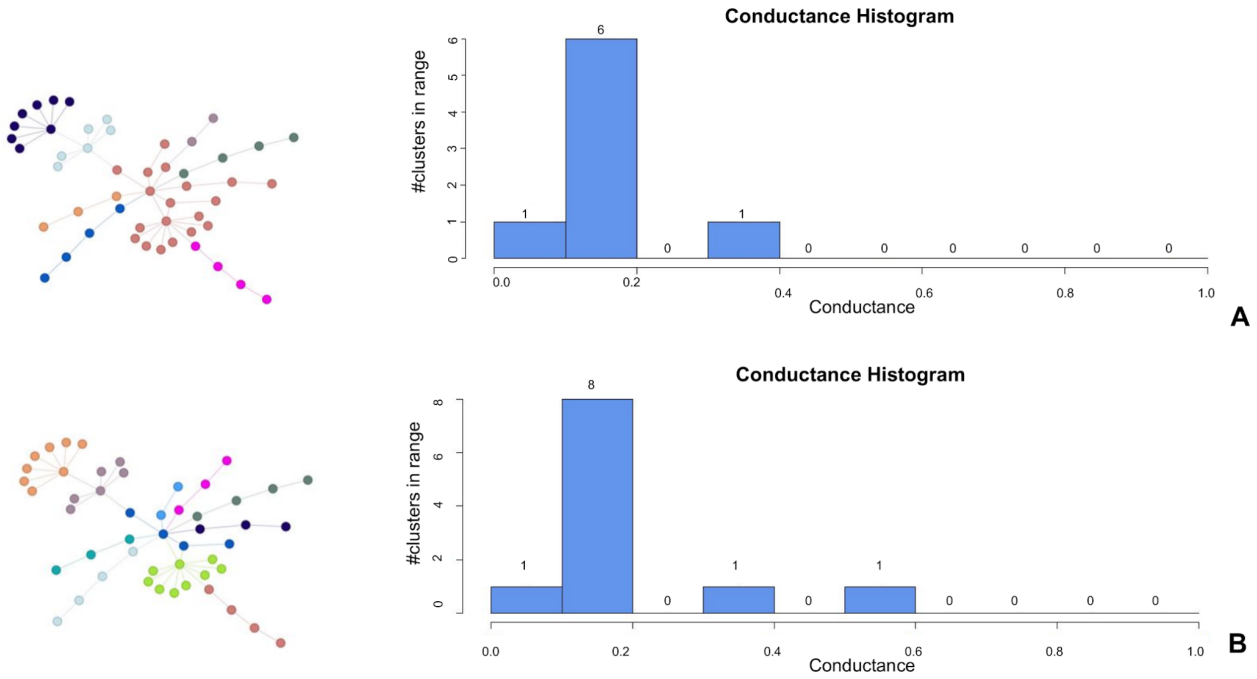
Μια επιπλέον δυνατότητα του VICTOR είναι ότι προσφέρει την δυνατότητα υπολογισμού του conductance των δικτύων. Μέσω του conductance που διατίθεται στο VICTOR, οι χρήστες μπορούν να δουν πόσο καλά συνδεδεμένη είναι μια ομάδα στο υπόλοιπο δίκτυο, σε σχέση με τις εσωτερικές συνδέσεις [119]. Για να υπολογιστεί το conductance, οι χρήστες χρειάζεται να ανεβάσουν ένα αρχείο δικτύου και να το συσχετίσουν με ένα αποτέλεσμα ομαδοποίησης της αρεσκείας τους. Το αρχείο του δικτύου πρέπει να αποτελείται από δύο ή τρεις στήλες ανάλογα αν το δίκτυο περιλαμβάνει βάρη ή όχι. Συγκεκριμένα, το αρχείο του δικτύου είναι σε μορφή text, στο οποίο οι δύο πρώτες στήλες περιγράφουν τον αρχικό κόμβο και τον κόμβο στόχο και η προαιρετική τρίτη στήλη περιγράφει τα βάρη μεταξύ των ακμών. Αξίζει να αναφερθεί ότι όλα τα αντικείμενα στο αρχείο της ομαδοποίησης πρέπει να αναπαρίστανται και στο δίκτυο.

Ας υποθέσουμε ένα δίκτυο $G(V, E)$ και αντίστοιχα μια ομαδοποίηση με n ομάδες. Προκείμενου να υπολογιστεί το conductance για κάθε ομάδα, το δίκτυο χρειάζεται να χωριστεί σε δύο υποσύνολα το S και το T για κάθε ομάδα $i \in n$. Το υποσύνολο $S_i \subseteq V(G)$ περιέχει όλες τις ακμές του δικτύου που περιλαμβάνονται και στην ομαδοποίηση. Ενώ το υποσύνολο $T_i \subseteq V(G)$ περιέχει τις ακμές του δικτύου που δεν περιλαμβάνονται στην ομαδοποίηση. Το conductance για μια ομάδα υπολογίζεται ως:

$$conductance_i = \frac{cut\ size_i}{min\ (Volume(S_i), Volume(T_i)}$$

Όπου τα $Volume(S_i)$ και $Volume(T_i)$ είναι το σύμπλεγμα των βαθμών των κόμβων για S_i και T_i αντίστοιχα και $cut\ size_i$ είναι το άθροισμα των ακραίων βαρών στη τομή των συνόλων S_i και T_i.

Το conductance του δικτύου μπορεί να περιγραφεί ως το ελάχιστο conductance των ομάδων του δικτύου. Όσο πιο κοντά η τιμή του είναι στο μηδέν, τόσο καλύτερο είναι το conductance των ομάδων. Στην εικόνα (Εικόνα 23) βλέπουμε ένα παράδειγμα από ένα δίκτυο συμπλέγματος που χρησιμοποιεί δύο διαφορετικά σύνολα συμπλέγματος. Αναλυτικότερα, σε κάθε ομάδα του δικτύου έχει δοθεί ένα χαρακτηριστικό χρώμα και ένα ιστόγραμμα απεικονίζει τη μετρική conductance. Το conductance έχει υπολογιστεί για δύο διαφορετικές ομαδοποιήσεις σε ένα τεχνητό δείγμα δικτύου. Ο οριζόντιος άξονας του κάθε ιστογράμματος δείχνει τις τιμές του conductance, ενώ ο κάθετος άξονας δείχνει τον αριθμό των ομάδων σύμφωνα με την κάθε τιμή του conductance.



Εικόνα 23: Υπολογισμός του conductance σε δύο διαφορετικές ομαδοποιήσεις από το VICTOr. Α) Οι ομάδες δημιουργήθηκαν από τον αλγόριθμο Label Propagation. Β) Οι ομάδες δημιουργήθηκαν από τον αλγόριθμο Walktrap.

9. Αποτελέσματα

Αξιοσημείωτο είναι το γεγονός ότι το VICTOR μπορεί να χρησιμοποιηθεί για την συγκριτική ανάλυση και οπτικοποίηση διαφόρων αποτελεσμάτων αλγορίθμων ομαδοποίησης. Όπως έχει προαναφερθεί το VICTOR μπορεί να εφαρμοστεί σε διαφορετικές περιπτώσεις για την σύγκριση και οπτικοποίηση των ομαδοποιήσεων. Συγκεκριμένα, αποτελεί ένα εργαλείο που μπορεί να λύσει προβλήματα όπως την ανάδειξη της καλύτερης ομαδοποίησης μεταξύ ενός σετ δεδομένων ή και περισσότερων, απλοποιώντας την διαδικασία και μειώνοντας τον χρόνο. Στο κεφάλαιο αυτό, για να αξιολογηθεί η αποτελεσματικότητα του VICTOR, καθώς και για λόγους παρουσίασης των λειτουργιών του, εφαρμόστηκαν τρεις διαφορετικές περιπτώσεις σύγκρισης ομαδοποιήσεων. Ειδικότερα, συγκρίθηκαν στο VICTOR ομαδοποιήσεις από διαφορετικά σετ δεδομένων και διαφορετικά ερευνητικά ενδιαφέροντα για την ανάδειξη της ευελιξίας της εφαρμογής. Αρχικά, συγκρίθηκαν πέντε διαφορετικοί αλγόριθμοι ομαδοποίησης δικτύων σε σετ δεδομένων αλληλεπιδράσεων πρωτεΐνης - πρωτεΐνης ζύμης. Επίσης, το ίδιο σετ δεδομένων χρησιμοποιήθηκε και για την σύγκριση των αποτελεσμάτων ομαδοποιήσεων που προήλθαν από τον ίδιο αλγόριθμο εφαρμόζοντας διαφορετικές τιμές στις παραμέτρους του. Παρόλο που οι δύο παραπάνω συγκρίσεις εφαρμόστηκαν σε σετ δεδομένων που σχετίζεται με τη Βιολογία, οι συγκρίσεις έγιναν για να αναδειχθεί η διαφορετικότητα στα αποτελέσματα μεταξύ διαφορετικών αλγορίθμων ομαδοποίησης ή διαφορετικών παραμέτρων. Ωστόσο, το VICTOR χρησιμοποιήθηκε και για την σύγκριση και οπτικοποίηση αποτελεσμάτων ομαδοποιήσεων που έχουν Βιολογική σημασία. Συγκεκριμένα, συγκρίθηκαν τέσσερις διαφορετικές μετα-αναλύσεις με ιεραρχικά ομαδοποιημένα διαφορικά εκφρασμένα γονίδια που βρέθηκαν να εμπλέκονται στο έμφραγμα του μυοκαρδίου. Στη συνέχεια, παρατίθενται περισσότερες λεπτομέρειες σχετικά με τις διαφορετικές συγκρίσεις που διεξήχθησαν από το VICTOR, καθώς και τα αποτελέσματα αυτών.

9.1 Σύγκριση διαφορετικών αλγορίθμων ομαδοποίησης

Για να εκτιμηθεί η αποτελεσματικότητα του VICTOR, αρχικά δοκιμάστηκαν σε αυτόν πέντε διαφορετικοί αλγόριθμοι ομαδοποιησης δικτύου. Οι αλγόριθμοι αυτοί εφαρμόστηκαν σε ένα δίκτυο αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης (PPI) ζύμης, το οποίο προήλθε εφαρμόζοντας χρωματογραφία συγγένειας και φασματομετρία μάζας [120]. Σύμφωνα με το NAP (Network Analysis Profiler) [121], το δίκτυο αποτελείται από 1,430 πρωτεΐνες (κόμβοι) και 6,530 αλληλεπιδράσεις (ακμές). Το NAP είναι ένα εργαλείο για την αυτοματοποίηση της δημιουργίας προφίλ δικτύου και της σύγκρισης της τοπολογίας του δικτύου. Η πυκνότητα του δικτύου (density) είναι 0,01, ο συντελεστής ομαδοποίησης (clustering coefficient) είναι 0,29 και τέλος η αρθρωτότητα (modularity) είναι 0,66. Για να παραχθούν οι ομάδες εφαρμόστηκαν πέντε διαφορετικοί αλγόριθμοι ομαδοποίησης στο δίκτυο που αναφέρθηκε παραπάνω, διατηρώντας τις προεπιλεγμένες παραμέτρους. Οι αλγόριθμοι ομαδοποίησης που εφαρμόστηκαν είναι: 1) MCL (Markov Cluster Algorithm), 2) SPICi (Speed and Performance In Clustering), 3) Louvain, 4) Walktrap και 5) Label Propagation.

Ο MCL [122] είναι ένας αλγόριθμος ομαδοποίησης γράφων ο οποίος βασίζεται στην λειτουργία της προσομοίωσης της στοχαστικής ροής (simulation of stochastic flow) σε γράφους. Ο αλγόριθμος είναι απλός και χρησιμοποιείται ήδη αρκετά στη Βιοπληροφορική λόγω της ανθεκτικότητας έναντι στο θόρυβο και της αποτελεσματικότητας του και παράγει ως

αποτέλεσμα πάρα πολλές ομάδες πρωτεϊνών. Ο αλγόριθμος SPICi αναγνωρίζει υψηλά συνδεδεμένες περιοχές με βάση την τοπική πυκνότητα τους. Στον αλγόριθμο SPICi δημιουργείται κάθε φορά μια ομάδα και η κάθε ομάδα αναπτύσσεται από το αρχικό ζεύγος πρωτεϊνών [22]. Ο αλγόριθμος Louvain [123] βρίσκει σε γρήγορο χρονικό διάστημα διαμερίσεις επιτυγχάνοντας υψηλή τιμή αρθρωτότητας (modularity) ενώ είναι αποδοτικός για δίκτυα με εκατομμύρια κόμβους. Ο αλγόριθμος Walktrap [124] εντοπίζει κοινότητες σε ένα γράφο μέσω τυχαίων περιπάτων. Ουσιαστικά, υπολογίζει την απόσταση των κόμβων βρίσκοντας τις πιο κοντινές κοινότητες μέσα από τους τυχαίους περιπάτους. Τέλος, ο αλγόριθμος Label Propagation [28] εντοπίζει δομές κοινοτήτων σε δίκτυα, αναθέτοντας μοναδικές ετικέτες στους κόμβους. Κάθε κόμβος υιοθετεί την ετικέτα σε συμφωνία με τη πλειοψηφία των γειτόνων της. Στο τέλος του αλγορίθμου, οι συνδεδεμένοι κόμβοι με την ίδια ετικέτα σχηματίζουν μια κοινότητα. Οι αλγόριθμοι αυτοί, είναι κάποιοι από τους πιο συχνούς αλγορίθμους που χρησιμοποιούνται για την ομαδοποίηση βιολογικών δικτύων. Καθώς, στο παράδειγμα μας έχουμε δίκτυο αλληλεπίδρασης πρωτεϊνών επιλέχθηκαν οι συγκεκριμένοι αλγόριθμοι για την ομαδοποίηση του δικτύου λόγω της αποτελεσματικότητας τους στην ανάδειξη ομάδων σε βιολογικά δίκτυα.

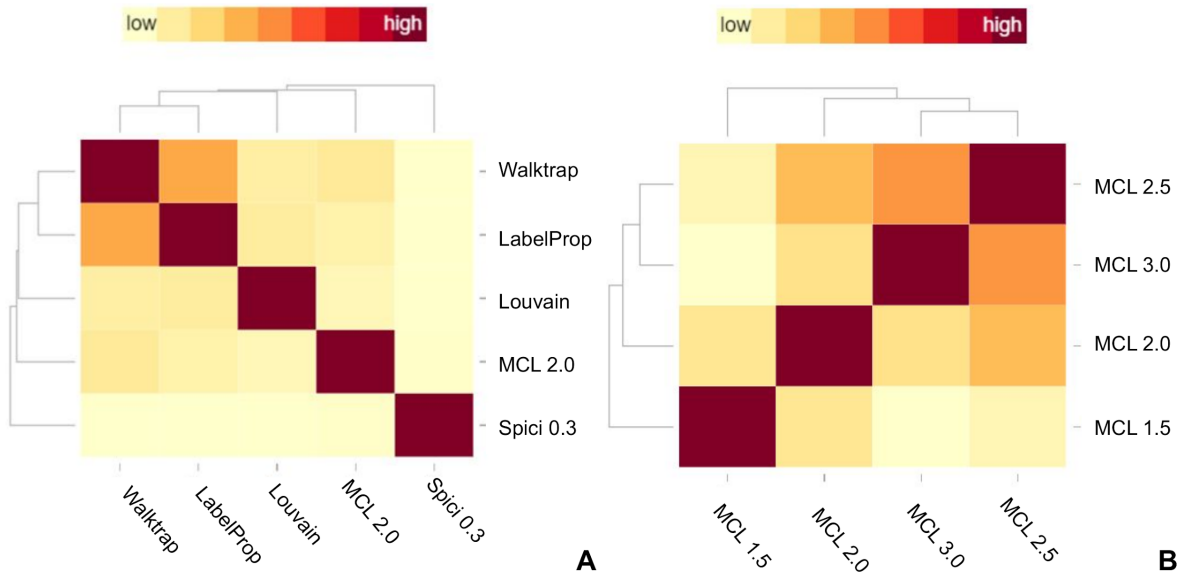
Ο αλγόριθμος MCL με τιμή πληθωρισμού 0,2 είχε ως αποτέλεσμα την δημιουργία 70 ομάδων. 66 από τις ομάδες αυτές είχαν περισσότερο από ένα μέλος αντιστοιχώντας συνολικά σε 1,426 κόμβους ενώ τέσσερις κόμβοι παρέμειναν μοναδικοί. Παρόμοια, ο SPICi αλγόριθμος, με τις παραμέτρους $\mathrm{T}_{density}$και $T_{support}$ να έχουν την τιμή $0,3$, δημιούργησε 148 ομάδες. 107 από τις ομάδες αυτές είχαν περισσότερα από τρία στοιχεία και αντιστοιχούσαν σε 928 κόμβοι και 44 από αυτές είχαν ακριβώς τρία στοιχεία (123 κόμβοι). 379 κόμβοι διανεμήθηκαν σε ομάδες που αποτελούνται λιγότερο από τρία στοιχεία. Ο αλγόριθμος Louvain δημιούργησε 193 ομάδες, όπου 146 από αυτές είχαν παραπάνω από ένα στοιχείο (1,383 κόμβοι) και 47 από αυτές αποτελούνται από μονά στοιχεία. Ο αλγόριθμος Walktrap κατέληξε σε 294 ομάδες, από τις οποίες 179 είχαν παραπάνω από ένα στοιχείο (1,315 κόμβοι) και 115 είχαν μονά στοιχεία. Τέλος, ο αλγόριθμος Label Propagation έδωσε σαν αποτέλεσμα 249 ομάδες. 1,341 κόμβοι διανεμήθηκαν σε 160 ομάδες με παραπάνω από ένα στοιχείο η κάθε μια και 89 ομάδες είχαν μονά στοιχεία.

Όπως βλέπουμε, οι ομαδοποιήσεις που προέκυψαν έχουν παράξει διαφορετικούς αριθμούς ομάδων και οι ομάδες περιέχουν διαφορετικό αριθμό στοιχείων. Καθώς, οι ομαδοποιήσεις έχουν διαφορετικά μεγέθη, για να συγκρίνουμε τα αποτελέσματα των πέντε αλγορίθμων ομαδοποίησης, αφαιρέθηκαν τα μόνα στοιχεία των ομάδων. Αυτό επιτεύχθηκε εφαρμόζοντας την επιλογή της τομής μέσω του φιλτραρίσματος που διατίθεται από το VICTOR. Έτσι διατηρήθηκαν μόνο τα στοιχεία που ανήκουν στην τομή των πέντε ομαδοποιήσεων και στη συνέχεια αυτές συγκρίθηκαν μεταξύ τους. Στην εικόνα (Εικόνα 24Α) βλέπουμε μια από τις οπτικοποιήσεις του VICTOR και συγκεκριμένα το Heatmap που προέκυψε από την σύγκριση των 5 αυτών ομαδοποιήσεων εφαρμόζοντας την μετρική Fowlkes-Mallows Index. Με την βοήθεια του Heatmap βλέπουμε ότι οι ομαδοποιήσεις που προέκυψαν από τους αλγορίθμους Walktrap και Label Propagation είναι πιο όμοιες μεταξύ τους σε σχέση με τις υπόλοιπες ομαδοποιήσεις. Αυτό μπορούμε να το καταλάβουμε γιατί το κελί που αντιστοιχεί στη σύγκριση αυτών των δύο ομαδοποιήσεων έχει πιο έντονο χρώμα (πορτοκαλί) σε σχέσει με τις υπόλοιπες συγκρίσεις (εξαιρούνται οι συγκρίσεις των αλγορίθμων με τον εαυτό τους). Ειδικότερα, η μετρική Fowlkes-Mallows Index έδωσε μια μέτρια ομοιότητα ανάμεσα στα αποτελέσματα των αλγορίθμων ομαδοποίησης Walktrap και Label Propagation με τιμή 0,408. Αξίζει να

αποσαφηνιστεί ότι τα αποτελέσματα δεν είναι καθολικά, εφαρμόζοντας τους αλγόριθμους σε ένα διαφορετικό δίκτυο PPI μπορούν να παραχθούν διαφορετικές τιμές ομοιότητας ανάμεσα στις ομαδοποιήσεις. Ωστόσο, δεν ισχύει το ίδιο για την επιλογή διαφορετικής μετρικής. Ίσως με κάποιες μικρές διαφορές στο αποτέλεσμα της μετρικής και πάλι θα αναδειχθούν οι ομαδοποιήσεις που είναι πιο όμοιες μεταξύ τους.

Ένα διαφορετικό σενάριο επιλέχθηκε για να αναπαρασταθούν οι ικανότητες του VICTOR. Αυτή τη φορά τα αποτελέσματα των ομαδοποιήσεων προήλθαν από τον ίδιο αλγόριθμο, τον αλγόριθμο MCL, με την εφαρμογή διαφορετικών παραμέτρων. Ουσιαστικά, ο αλγόριθμος MCL εφαρμόστηκε στο δίκτυο PPI της ζύμης που αναφέρθηκε παραπάνω με τέσσερις διαφορετικές παραμέτρους πληθωρισμού (τιμές 1,5, 2,0, 2,5 και 3,0). Συνήθως, στον αλγόριθμο MCL η παράμετρος αυτή προσαρμόζει την αρθρωτότητα (modularity). Η αρθρωτότητα είναι ένα μέτρο της δομής των δικτύων, που υπολογίζει τη δύναμη της διαίρεσης ενός δικτύου σε ομάδες. Τα δίκτυα με υψηλή αρθρωτότητα έχουν πυκνές συνδέσεις μεταξύ των κόμβων εντός των ομάδων, αλλά αραιές συνδέσεις μεταξύ των κόμβων σε διαφορετικές ομάδες. Επομένως, όσο μεγαλύτερη είναι η τιμή της παραμέτρου τόσο περισσότερες ομάδες παράγονται. Στην περίπτωση μας, όταν η παράμετρος είχε την τιμή 1,5, 88 ομάδες παρήχθησαν με παραπάνω από ένα στοιχείο σε αυτές και 38 ομάδες με μονά στοιχεία. Η παράμετρος με τιμή 2,0 έδωσε 66 ομάδες με παραπάνω από ένα στοιχεία και 4 ομάδες με μονά στοιχεία και η τιμή 2,5 παρήγαγε 45 ομάδες με παραπάνω από ένα στοιχείο και 13 ομάδες με μονά στοιχεία. Τέλος, η παράμετρος με τιμή 3,0 παρήγαγε 38 ομάδες, 37 από αυτές είχαν παραπάνω από 2 στοιχεία ενώ δεν υπήρχε ομάδα με μονά στοιχεία.

Τα αποτελέσματα των τεσσάρων ομαδοποιήσεων που προέκυψαν από τον αλγόριθμο MCL με την εφαρμογή διαφορετικών παραμέτρων, συγκρίθηκαν μεταξύ τους. Έτσι στην εικόνα (Εικόνα 24Β) βλέπουμε το Heatmap που προέκυψε εφαρμόζοντας την μετρική Normalized Variation of Information. Από το Heatmap μπορούμε να συμπεράνουμε ότι οι ομαδοποιήσεις που προέκυψαν από τον αλγόριθμο MCL με τιμή παραμέτρου πληθωρισμού 2,5 και 3,0 είναι πιο όμοιες μεταξύ τους σε σχέση με τις υπόλοιπες. Αναλυτικότερα, η μετρική σύγκρισης για τα αποτελέσματα των αλγορίθμων ομαδοποίησης με τιμή παραμέτρου ίση με 2,5 και 3,0 έδωσε σαν αποτέλεσμα τιμή ίση με 0,642 και βλέπουμε ότι το κελί χρωματίστηκε με έντονο χρώμα (πορτοκαλί). Για τα αποτελέσματα των αλγορίθμων ομαδοποίησης με τιμή παραμέτρου 2,0 και 2,5 η μετρική είχε την τιμή 0,555. Επομένως, βρέθηκαν δύο διαφορετικές περιπτώσεις ομαδοποιήσεων να μοιάζουν μεταξύ τους. Ωστόσο, όπως αναφέρθηκε και παραπάνω αν ο αλγόριθμος εφαρμοστεί σε διαφορετικό δίκτυο PPI τότε θα έχουμε διαφορετικά αποτελέσματα.



Εικόνα 24: Οπτικοποίησης Ιεραρχικών Heatmap. Α) Το Ιεραρχικό Heatmap που προέκυψε από πέντε διαφορετικούς αλγορίθμους εφαρμοσμένους σε ένα δίκτυο PPI ζύμης. Β) Το Ιεραρχικό Heatmap που παράχθηκε από την μετρική Normalized Variation of Information πάνω στα τέσσερα διαφορετικά αποτελέσματα ομαδοποιήσεων από τον αλγόριθμο MCL με παραμέτρους πληθωρισμού [1.5, 2.0, 2.5, 3.0].

9.2 Εφαρμογή δεδομένων γονιδιακής έκφρασης από μετα-ανάλυση του εμφράγματος του μυοκαρδίου

Το έμφραγμα του μυοκαρδίου είναι μια περίπλοκη ασθένεια με πολυπαραγοντική παθογένεση [125]. Αυτό οφείλεται στο γεγονός ότι αλληλεπιδρούν γενετικοί και περιβαλλοντικοί παράγοντες κινδύνου. Για αυτόν τον λόγο, υπάρχει ανάγκη αναγνώρισης νέων βιοχημικών και γενετικών δεικτών που σχετίζονται με το έμφραγμα του μυοκαρδίου. Έχει εντοπιστεί ότι διαφορετικά εκφρασμένα γονίδια σχετίζονται με το έμφραγμα του μυοκαρδίου (MI). Η δημιουργία μεγάλης κλίμακας προφίλ γονιδιακής έκφρασης με τεχνολογία μικροσυστοιχιών επέτρεψε την πρόβλεψη ασθενειών όπως προκαρκινικές καταστάσεις. Έτσι, έχει δημιουργηθεί ενδιαφέρον για τον προσδιορισμό των προφίλ γονιδιακής έκφρασης που βασίζονται κυρίως σε μικροσυστοιχίες (μεταγραφικά) για τη διάγνωση του εμφράγματος του μυοκαρδίου, καθώς και για τον κίνδυνο πρόβλεψη του εμφράγματος του μυοκαρδίου και του καρδιαγγειακού θανάτου. Δεδομένα έκφρασης σχετικά με διαφορετικά γονίδια που σχετίζονται με τη συχνότητα εμφάνισης του εμφράγματος του μυοκαρδίου συλλέχθηκαν από το αποθετήριο δεδομένων GEO. Το **GEO** (Gene Expression Omnibus) [126] είναι ένα δημόσιο αποθετήριο δεδομένων γονιδιακής έκφρασης που διαχειρίζεται το Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας (NCBI). Αυτά τα δεδομένα γονιδιωματικής διαλογής υψηλής απόδοσης προέρχονται από πειραματικά δεδομένα Microarray ή RNA-Seq. Ως αποτέλεσμα, προέκυψαν τέσσερα διαφορετικά σετ δεδομένων γονιδίων που σχετίζονται με το έμφραγμα του μυοκαρδίου.

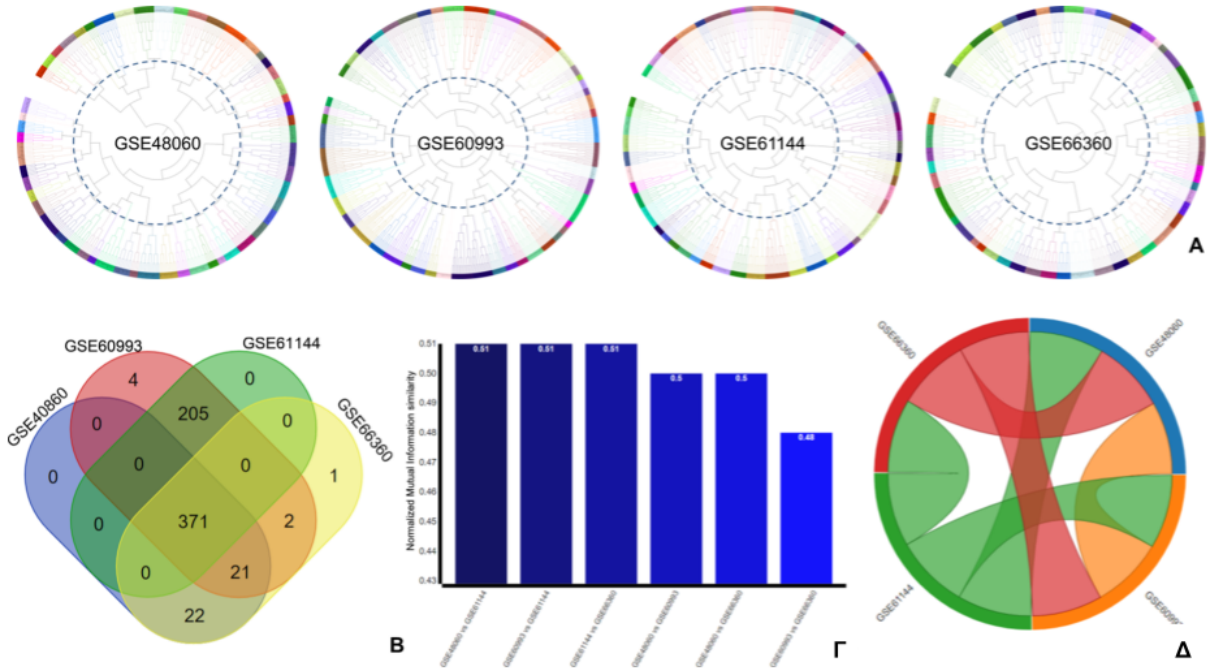
Σε αυτό το υποκεφάλαιο, παρουσιάζεται η χρήση του VICTOR για την σύγκριση και οπτικοποίηση των τεσσάρων διαφορετικών σετ δεδομένων που αφορούν γονίδια και προέρχονται από ανεξάρτητες μελέτες που έχουν χρησιμοποιηθεί σε πρόσφατα δημοσιευμένες μετα-αναλύσεις για τον εντοπισμό διαφορετικά εκφρασμένων γονιδίων στο έμφραγμα του μυοκαρδίου [125]. Σε αυτή την περίπτωση χρησιμοποιήθηκαν διαφορετικά σετ δεδομένων όπως αναφέρθηκε, τα οποία όμως είναι σχετικά μεταξύ τους (απαντούν στην ίδια

επιστημονική ερώτηση), για να συγκριθούν τα αποτελέσματα των ομαδοποιήσεων που προήλθαν εφαρμόζοντας τον ίδιο αλγόριθμο.

Πιο συγκεκριμένα, έγινε επεξεργασία τεσσάρων σετ δεδομένων μικροσυστοιχιών (GSE48060 [127], GSE60993 [128], GSE61144 [128], GSE66360 [129]) προερχόμενα από το αποθετήριο δεδομένων GEO (Gene Expression Omnibus). Από αυτά αναγνωρίστηκαν 626 διαφορετικά εκφρασμένα γονίδια σε ασθενής με έμφραγμα του μυοκαρδίου, αφού συγκρίθηκαν με υγιή άτομα χρησιμοποιώντας ένα κατώφλι 0,01 μιας FDR προσαρμοσμένης p-value. Χρησιμοποιήθηκε ο αλγόριθμος Quantile normalization από την εφαρμογή Expander [130] για την κανονικοποίηση των τιμών έκφρασης κάθε συνόλου δεδομένων. Σε ένα επόμενο βήμα, τα γονίδια κάθε σετ δεδομένων ομαδοποιήθηκαν χρησιμοποιώντας μια ιεραρχική ομαδοποίηση μέσης σύνδεσης (average linkage hierarchical clustering). Με αποτέλεσμα, να παραχθούν τέσσερα ιεραρχικά δέντρα (ένα για κάθε σετ δεδομένων) σε μορφή Newick. Τα δέντρα αυτά οπτικοποιήθηκαν σε μορφή κυκλικού δέντρου (Εικόνα 25Α), χρησιμοποιώντας το iTOL [131] και χρησιμοποιήθηκαν για την παραγωγή ομάδων. Πιο συγκεκριμένα αυτό επιτεύχθηκε εφαρμόζοντας ένα κατώφλι στην απόσταση του κάθε ιεραρχικού δέντρου, έτσι ώστε κάθε ομάδα να αποτελείται τουλάχιστον από τέσσερα γονίδια. Έτσι, τα φύλλα των δέντρων αναπαριστούν τα γονίδια, τα οποία είναι ομαδοποιημενα σε ομάδες και είναι χρωματίζονται σύμφωνα με το κατώφλι (κυκλική διακεκομμένη γραμμή). Μέσω αυτής της διαδικασίας παράχθηκαν 54 ομάδες από το GSE48060 σετ δεδομένων, 52 ομάδες από το GSE60993, 56 ομάδες από το GSE61144 και 52 ομάδες από το GSE66360 σετ δεδομένων με 414, 603, 576 και 417 γονίδια αντίστοιχα.

Στην εικόνα (Εικόνα 25Β) βλέπουμε ένα διάγραμμα Venn με τα αλληλοεπικαλυπτόμενα γονίδια μεταξύ των τεσσάρων διαφορετικών σετ δεδομένων μαζί με τα ιεραρχικά δέντρα που παρήχθησαν με την ιεραρχική ομαδοποίηση μέσης σύνδεσης (Εικόνα 25Α). Όπως φαίνεται από το διάγραμμα Venn 371 από τα 626 εκφρασμένα γονίδια ανήκουν και στα τέσσερα διαφορετικά σετ δεδομένων ($\sim 59,26\%$), 21 από αυτά είναι κοινά σε τρία από τα σετ δεδομένων (GSE48060, GSE60993 και GSE66360) και 205 γονίδια είναι κοινά μόνο σε δύο (GSE60993, GSE61144).

Στη συνέχεια, χρησιμοποιώντας τη μετρική συγκρίσης Normalized Mutual Information by strehl and Ghosh (NMI1) οι ομοιότητες των ζευγών των στοιχείων ανάμεσα στις ομαδοποιήσεις των τεσσάρων σετ δεδομένων οτικοποιήθηκαν με την βοήθεια του VICTOR. Στην εικόνα (Εικόνα 25Γ) βλέπουμε τα bar charts που προέκυψαν από τα αρχεία των ομαδοποιήσεων των τεσσάρων διαφορετικών σετ δεδομένων εφαρμόζοντας την μετρική NMI1 και επιπλέον στην εικόνα βλέπουμε το circos plot που παράχθηκε με την χρήση της ίδιας μετρικής (Εικόνα 25Δ). Στο επίπεδο των ομαδοποιήσεων, τα σετ των δεδομένων χρησιμοποιώντας την μετρική ομοιότητας NMI1 δείχνουν ομοιότητα που κυμαίνεται $\sim 50\%$. Η ομοιότητα που προέκυψε χρησιμοποιώντας μια από τις μετρικές του VICTOR συμφωνεί με την ομοιότητα του διαγράμματος Venn (Εικόνα 25Γ, 25Δ). Τα αποτελέσματα αυτά προήλθαν χρησιμοποιώντας όλα τα μοναδικα εκφρασμένα γονίδια (ένωση) για την σύγκριση των ομαδοποιήσεων. Ωστόσο, παρόμοια αποτελέσματα παρήχθησαν χρησιμοποιώντας τα εκφραζόμενα γονίδια της τομής τους, δηλαδή ομοιότητα σε επίπεδο $\sim 48\%$.



Εικόνα 25: Σύγκριση τεσσάρων σετ δεδομένων εκφρασμένων γονιδίων που αφορούν το έμφραγμα του μυοκαρδίου. Α) Οι ιεραρχικές ομαδοποιήσεις των τεσσάρων σετ δεδομένων. Τα αποτελέσματα των ομαδοποιήσεων είναι σε μορφή κυκλικών δέντρων. Β) Διάγραμμα Venn που αναπαριστά την αλληλοεπικάλυψη των σετ δεδομένων. Γ) Bar chart από το VICTOR χρησιμοποιώντας την μετρική NMI. Δ) Circos plot που αναπαριστά τις σχέσεις μεταξύ των τεσσάρων σετ δεδομένων.

Βιβλιογραφία

[1] R. Nugent and M. Meila, 'An overview of clustering applied to molecular biology', *Methods Mol. Biol. Clifton NJ*, vol. 620, pp. 369–404, 2010, doi: 10.1007/978-1-60761-580-4_12.

[2] T. Villmann and C. Albani, 'Clustering of Categoric Data in Medicine — Application of Evolutionary Algorithms', in *Computational Intelligence. Theory and Applications*, Berlin, Heidelberg, 2001, pp. 619–627. doi: 10.1007/3-540-45493-4_62.

[3] G. B. Coleman and H. C. Andrews, 'Image segmentation by clustering', *Proc. IEEE*, vol. 67, no. 5, pp. 773–785, May 1979, doi: 10.1109/PROC.1979.11327.

[4] 'Cluster Analysis in Marketing Research: Review and Suggestions for Application - Girish Punj, David W. Stewart, 1983'. https://journals.sagepub.com/doi/abs/10.1177/002224378302000204 (accessed Apr. 12, 2021).

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

[6] R. Xu and D. C. Wunsch, 'Clustering algorithms in biomedical research: a review', *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 120–154, 2010, doi: 10.1109/RBME.2010.2083647.

[7] T. S. Madhulatha, 'AN OVERVIEW ON CLUSTERING METHODS', *IOSR J. Eng.*, vol. 02, no. 04, pp. 719–725, Apr. 2012, doi: 10.9790/3021-0204719725.

[8] 'Hierarchical Clustering | solver'. https://www.solver.com/xlminer/help/hierarchical-clustering-intro (accessed May 14, 2021).

[9] LaptrinhX, 'Hierarchical Clustering', *LaptrinhX*, Nov. 15, 2016. https://laptrinhx.com/hierarchical-clustering-3345648568/ (accessed May 14, 2021).

[10] M. Schuldiner *et al.*, 'Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile', *Cell*, vol. 123, no. 3, pp. 507–519, Nov. 2005, doi: 10.1016/j.cell.2005.08.031.

[11] M. Costanzo *et al.*, 'The genetic landscape of a cell', *Science*, vol. 327, no. 5964, pp. 425–431, Jan. 2010, doi: 10.1126/science.1180823.

[12] W. H. E. Day and H. Edelsbrunner, 'Efficient algorithms for agglomerative hierarchical clustering methods', *J. Classif.*, vol. 1, no. 1, pp. 7–24, Dec. 1984, doi: 10.1007/BF01890115.

[13] N. Saitou and M. Nei, 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Mol. Biol. Evol.*, vol. 4, no. 4, pp. 406–425, Jul. 1987, doi: 10.1093/oxfordjournals.molbev.a040454.

[14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, 'An efficient k-means clustering algorithm: analysis and implementation', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002, doi: 10.1109/TPAMI.2002.1017616.

[15] A. Likas, N. Vlassis, and J. J. Verbeek, 'The global k-means clustering algorithm', *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, Feb. 2003, doi: 10.1016/S0031-3203(02)00060-2.

[16] G. Amato, C. Gennaro, V. Oria, and M. Radovanović, Eds., *Similarity Search and Applications: 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings*, vol. 11807. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-32047-8.

[17] Y. Aboubi, H. Drias, and N. Kamel, 'BAT-CLARA: BAT-inspired algorithm for Clustering LARge Applications', *IFAC-Pap.*, vol. 49, no. 12, pp. 243–248, Jan. 2016, doi: 10.1016/j.ifacol.2016.07.607.

[18] R. T. Ng and J. Han, 'CLARANS: a method for clustering objects for spatial data mining', *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 5, pp. 1003–1016, Sep. 2002, doi: 10.1109/TKDE.2002.1033770.

[19] X. Xu, M. Ester, H.-P. Kriegel, and J. Sander, ‘A distribution-based clustering algorithm for mining in large spatial databases’, in *Proceedings 14th International Conference on Data Engineering*, Feb. 1998, pp. 324–331. doi: 10.1109/ICDE.1998.655795.
[20] N. D. Dat, V. N. Phu, V. T. N. Tran, V. T. N. Chau, and T. A. Nguyen, ‘STING Algorithm Used English Sentiment Classification in a Parallel Environment’, *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 07, p. 1750021, Jan. 2017, doi: 10.1142/S0218001417500215.
[21] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, ‘Automatic Subspace Clustering of High Dimensional Data’, *Data Min. Knowl. Discov.*, vol. 11, no. 1, pp. 5–33, Jul. 2005, doi: 10.1007/s10618-005-1396-1.
[22] P. Jiang and M. Singh, ‘SPICi: a fast clustering algorithm for large biological networks’, *Bioinformatics*, vol. 26, no. 8, pp. 1105–1111, Apr. 2010, doi: 10.1093/bioinformatics/btq078.
[23] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, ‘Fast Unfolding of Communities in Large Networks’, *J. Stat. Mech. Theory Exp.*, vol. 2008, Apr. 2008, doi: 10.1088/1742-5468/2008/10/P10008.
[24] C. Liu, W. Xiong, X. Zhang, and Z. Liu, ‘A Method of Node Layout of a Complex Network Based on Community Compression’, *Future Internet*, vol. 11, no. 12, Art. no. 12, Dec. 2019, doi: 10.3390/fi11120250.
[25] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, ‘An efficient algorithm for large-scale detection of protein families’, *Nucleic Acids Res.*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002.
[26] ‘HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks’. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5888241/ (accessed Apr. 23, 2021).
[27] ‘An algorithm Walktrap-SPM for detecting overlapping community structure | International Journal of Modern Physics B’. https://www.worldscientific.com/doi/abs/10.1142/S0217979217501211 (accessed Apr. 23, 2021).
[28] U. N. Raghavan, R. Albert, and S. Kumara, ‘Near linear time algorithm to detect community structures in large-scale networks’, *Phys. Rev. E*, vol. 76, no. 3, p. 036106, Sep. 2007, doi: 10.1103/PhysRevE.76.036106.
[29] N. Safari-Alighiarloo, M. Taghizadeh, M. Rezaei-Tavirani, B. Goliaei, and A. A. Peyvandi, ‘Protein-protein interaction networks (PPI) and complex diseases’, *Gastroenterol. Hepatol. Bed Bench*, vol. 7, no. 1, pp. 17–31, 2014.
[30] W. E. Payne and J. I. Garrels, ‘Yeast Protein Database (YPD): a database for the complete proteome of Saccharomyces cerevisiae’, *Nucleic Acids Res.*, vol. 25, no. 1, pp. 57–62, Jan. 1997, doi: 10.1093/nar/25.1.57.
[31] H. W. Mewes, J. Hani, F. Pfeiffer, and D. Frishman, ‘MIPS: A database for protein sequences and complete genomes’, *Nucleic Acids Res.*, vol. 26, no. 1, pp. 33–37, Jan. 1998, doi: 10.1093/nar/26.1.33.
[32] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, ‘MINT: a Molecular INTeraction database’, *FEBS Lett.*, vol. 513, no. 1, pp. 135–140, Feb. 2002, doi: 10.1016/s0014-5793(01)03293-8.
[33] I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, ‘DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions’, *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, Jan. 2002, doi: 10.1093/nar/30.1.303.
[34] ‘BioGRID interaction database: 2019 update | Nucleic Acids Research | Oxford Academic’. https://academic.oup.com/nar/article/47/D1/D529/5204333 (accessed May 31, 2021).
[35] R. Goel, H. C. Harsha, A. Pandey, and T. S. K. Prasad, ‘Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis’, *Mol. Biosyst.*, vol. 8, no. 2, pp. 453–463, Feb. 2012, doi: 10.1039/c1mb05340j.
[36] J. N. Copp, E. Akiva, P. C. Babbitt, and N. Tokuriki, ‘Revealing Unexplored

Sequence-Function Space Using Sequence Similarity Networks', *Biochemistry*, vol. 57, no. 31, pp. 4651–4662, Aug. 2018, doi: 10.1021/acs.biochem.8b00473.
[37] H. Xing, S. W. Kembel, and V. Makarenkov, 'Transfer index, NetUniFrac and some useful shortest path-based distances for community analysis in sequence similarity networks', *Bioinformatics*, vol. 36, no. 9, pp. 2740–2749, May 2020, doi: 10.1093/bioinformatics/btaa043.
[38] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, 'NCBI BLAST: a better web interface', *Nucleic Acids Res.*, vol. 36, no. Web Server issue, pp. W5-9, Jul. 2008, doi: 10.1093/nar/gkn201.
[39] W. R. Pearson, 'Flexible sequence similarity searching with the FASTA3 program package', *Methods Mol. Biol. Clifton NJ*, vol. 132, pp. 185–219, 2000, doi: 10.1385/1-59259-192-2:185.
[40] R. Barbuti, R. Gori, P. Milazzo, and L. Nasti, 'A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models', *J. Membr. Comput.*, vol. 2, no. 3, pp. 207–226, Oct. 2020, doi: 10.1007/s41965-020-00046-y.
[41] E. Liu, L. Li, and L. Cheng, 'Gene Regulatory Network Review', in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 155–164. doi: 10.1016/B978-0-12-809633-8.20218-5.
[42] M. Kanehisa and S. Goto, 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.
[43] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, 'GTRD: a database on gene transcription regulation—2019 update', *Nucleic Acids Res.*, vol. 47, no. D1, pp. D100–D105, Jan. 2019, doi: 10.1093/nar/gky1128.
[44] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, 'TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites', *Nucleic Acids Res.*, vol. 24, no. 1, pp. 238–241, Jan. 1996, doi: 10.1093/nar/24.1.238.
[45] I. Nassiri, A. Masoudi-Nejad, M. Jalili, and A. Moeini, 'Nonparametric Simulation of Signal Transduction Networks with Semi-Synchronized Update', *PLOS ONE*, vol. 7, no. 6, p. e39643, 2012, doi: 10.1371/journal.pone.0039643.
[46] V. M. Gumerov, D. R. Ortega, O. Adebali, L. E. Ulrich, and I. B. Zhulin, 'MiST 3.0: an updated microbial signal transduction database with an emphasis on chemosensory systems', *Nucleic Acids Res.*, vol. 48, no. D1, pp. D459–D464, Jan. 2020, doi: 10.1093/nar/gkz988.
[47] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov, and E. Wingender, 'TRANSPATH: an integrated database on signal transduction and a tool for array analysis', *Nucleic Acids Res.*, vol. 31, no. 1, pp. 97–100, Jan. 2003, doi: 10.1093/nar/gkg089.
[48] A. Weber Zendrera, N. Sokolovska, and H. A. Soula, 'Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature', *BMC Bioinformatics*, vol. 20, no. 1, p. 499, Dec. 2019, doi: 10.1186/s12859-019-3112-y.
[49] G. Chalancon, K. Kruse, and M. M. Babu, 'Metabolic Networks, Structure and Dynamics', in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer, 2013, pp. 1263–1267. doi: 10.1007/978-1-4419-9863-7_561.
[50] 'Graph-Based Analysis of Metabolic Networks | SpringerLink'. https://link.springer.com/chapter/10.1007/978-3-662-04747-7_12 (accessed May 24, 2021).
[51] P. D. Karp *et al.*, 'The EcoCyc Database', *EcoSal Plus*, vol. 6, no. 1, May 2014, doi: 10.1128/ecosalplus.ESP-0009-2013.
[52] J. W. Whitaker, I. Letunic, G. A. McConkey, and D. R. Westhead, 'metaTIGER: a metabolic evolution resource', *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D531–D538, Jan. 2009, doi: 10.1093/nar/gkn826.
[53] J. Li *et al.*, 'Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design', *Sci. Rep.*, vol. 8, no. 1, p. 622, Dec. 2018, doi: 10.1038/s41598-017-18705-z.

[54] H. N. Kadarmideen and N. S. Watson-haigh, ‘Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data’, *Bioinformation*, vol. 8, no. 18, pp. 855–861, Sep. 2012, doi: 10.6026/97320630008855.
[55] S. Xiang *et al.*, ‘Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer’s disease patients’, *BMC Med. Genomics*, vol. 11, no. S6, p. 115, Dec. 2018, doi: 10.1186/s12920-018-0431-1.
[56] K. Ovens, F. Maleki, B. F. Eames, and I. McQuillan, ‘Juxtapose: a gene-embedding approach for comparing co-expression networks’, *BMC Bioinformatics*, vol. 22, no. 1, p. 125, Dec. 2021, doi: 10.1186/s12859-021-04055-1.
[57] T. Barrett *et al.*, ‘NCBI GEO: archive for functional genomics data sets--10 years on’, *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D1005-1010, Jan. 2011, doi: 10.1093/nar/gkq1184.
[58] H. Parkinson *et al.*, ‘ArrayExpress—a public database of microarray experiments and gene expression profiles’, *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D747–D750, Jan. 2007, doi: 10.1093/nar/gkl995.
[59] T. Obayashi, Y. Okamura, S. Ito, S. Tadaka, I. N. Motoike, and K. Kinoshita, ‘COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals’, *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D1014–D1020, Jan. 2013, doi: 10.1093/nar/gks1014.
[60] M. Arenas, G. Valiente, and D. Posada, ‘Characterization of Reticulate Networks Based on the Coalescent with Recombination’, *Mol. Biol. Evol.*, vol. 25, no. 12, pp. 2517–2520, Dec. 2008, doi: 10.1093/molbev/msn219.
[61] D. H. Huson, ‘SplitsTree: analyzing and visualizing evolutionary data’, *Bioinforma. Oxf. Engl.*, vol. 14, no. 1, pp. 68–73, 1998, doi: 10.1093/bioinformatics/14.1.68.
[62] K. P. Schliep, ‘phangorn: phylogenetic analysis in R’, *Bioinformatics*, vol. 27, no. 4, pp. 592–593, Feb. 2011, doi: 10.1093/bioinformatics/btq706.
[63] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, ‘Dendroscope: An interactive viewer for large phylogenetic trees’, *BMC Bioinformatics*, vol. 8, no. 1, p. 460, Nov. 2007, doi: 10.1186/1471-2105-8-460.
[64] P. Landi, H. O. Minoarivelo, Å. Brännström, C. Hui, and U. Dieckmann, ‘Complexity and stability of ecological networks: a review of the theory’, *Popul. Ecol.*, vol. 60, no. 4, pp. 319–345, Oct. 2018, doi: 10.1007/s10144-018-0628-3.
[65] G. Losapio, A. Montesinos-Navarro, and H. Saiz, ‘Perspectives for ecological networks in plant ecology’, *Plant Ecol. Divers.*, vol. 12, no. 2, pp. 87–102, Mar. 2019, doi: 10.1080/17550874.2019.1626509.
[66] ‘Ecological Networks’, *obo*. https://www.oxfordbibliographies.com/view/document/obo-9780199830060/obo-9780199830060-0091.xml (accessed May 25, 2021).
[67] A. Lloyd and S. Valeika, ‘Network Models In Epidemiology: An Overview’, Sep. 2007, doi: 10.1142/9789812771582_0008.
[68] A. Halu, M. De Domenico, A. Arenas, and A. Sharma, ‘The multiplex network of human diseases’, *Npj Syst. Biol. Appl.*, vol. 5, no. 1, p. 15, Dec. 2019, doi: 10.1038/s41540-019-0092-5.
[69] E. P. García del Valle, G. Lagunes García, L. Prieto Santamaría, M. Zanin, E. Menasalvas Ruiz, and A. Rodríguez-González, ‘Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources’, *J. Biomed. Inform.*, vol. 94, p. 103206, Jun. 2019, doi: 10.1016/j.jbi.2019.103206.
[70] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, ‘Online Mendelian Inheritance in Man (OMIM)’, *Hum. Mutat.*, vol. 15, no. 1, pp. 57–61, 2000, doi: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G.
[71] H. M. Cartwright, ‘Artificial Neural Networks in Biology and Chemistry—The Evolution of a New Analytical Tool’, in *Artificial Neural Networks: Methods and Applications*, D. J. Livingstone, Ed. Totowa, NJ: Humana Press, 2009, pp. 1–13. doi: 10.1007/978-1-60327-101-1_1.

[72] ‘The WEKA data mining software: an update: ACM SIGKDD Explorations Newsletter: Vol 11, No 1’. https://dl.acm.org/doi/10.1145/1656274.1656278 (accessed May 03, 2021).
[73] T. K. Moon, ‘The expectation-maximization algorithm’, *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996, doi: 10.1109/79.543975.
[74] C. Tran, J.-Y. Kim, W.-Y. Shin, and S.-W. Kim, ‘Clustering-Based Collaborative Filtering Using an Incentivized/Penalized User Model’, *IEEE Access*, vol. 7, pp. 62115–62125, 2019, doi: 10.1109/ACCESS.2019.2914556.
[75] S. C. Johnson, ‘Hierarchical clustering schemes’, *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967, doi: 10.1007/BF02289588.
[76] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench*. Morgan Kaufmann, 2016.
[77] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, ‘Cytoscape 2.8: new features for data integration and network visualization’, *Bioinforma. Oxf. Engl.*, vol. 27, no. 3, pp. 431–432, Feb. 2011, doi: 10.1093/bioinformatics/btq675.
[78] J. H. Morris *et al.*, ‘clusterMaker: a multi-algorithm clustering plugin for Cytoscape’, *BMC Bioinformatics*, vol. 12, no. 1, p. 436, Nov. 2011, doi: 10.1186/1471-2105-12-436.
[79] G. D. Bader and C. W. Hogue, ‘An automated method for finding molecular complexes in large protein interaction networks’, *BMC Bioinformatics*, vol. 4, no. 1, p. 2, Jan. 2003, doi: 10.1186/1471-2105-4-2.
[80] M. E. J. Newman and M. Girvan, ‘Finding and evaluating community structure in networks’, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 69, no. 2 Pt 2, p. 026113, Feb. 2004, doi: 10.1103/PhysRevE.69.026113.
[81] T. Nepusz, R. Sasidharan, and A. Paccanaro, ‘SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale’, *BMC Bioinformatics*, vol. 11, p. 120, Mar. 2010, doi: 10.1186/1471-2105-11-120.
[82] T. Wittkop *et al.*, ‘Partitioning biological data with transitivity clustering’, *Nat. Methods*, vol. 7, no. 6, pp. 419–420, Jun. 2010, doi: 10.1038/nmeth0610-419.
[83] A. M. Newman and J. B. Cooper, ‘AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number’, *BMC Bioinformatics*, vol. 11, p. 117, Mar. 2010, doi: 10.1186/1471-2105-11-117.
[84] M. Koutrouli, E. Karatzas, K. Papanikolopoulou, and G. A. Pavlopoulos, ‘NORMA-The network makeup artist: a web tool for network annotation visualization’, Bioinformatics, preprint, Mar. 2020. doi: 10.1101/2020.03.05.978585.
[85] ‘On Clustering Validation Techniques | SpringerLink’. https://link.springer.com/article/10.1023/A:1012801612483 (accessed May 04, 2021).
[86] ‘Clustering algorithms: A comparative approach’. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0210236 (accessed May 04, 2021).
[87] S. L’Yi *et al.*, ‘XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data’, *BMC Bioinformatics*, vol. 16 Suppl 11, p. S5, 2015, doi: 10.1186/1471-2105-16-S11-S5.
[88] T. Kohonen, ‘The self-organizing map’, *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990, doi: 10.1109/5.58325.
[89] ‘OPTICS: ordering points to identify the clustering structure: ACM SIGMOD Record: Vol 28, No 2’. https://dl.acm.org/doi/abs/10.1145/304181.304187?casa_token=tKSolgQ9BrgAAAAA:aL8DiDII0XpKo6jlsOSuHxcWgCuAwR2Fg5u3j9CmHC-jmYuMQGOJ1_NxioWkj0qPD0CUz6qhpYo2 (accessed May 18, 2021).
[90] Y. Nan, K. M. Chai, W. S. Lee, and H. L. Chieu, ‘Optimizing F-measure: A Tale of Two Approaches’, *ArXiv12064625 Cs*, Jun. 2012, Accessed: May 04, 2021. [Online]. Available: http://arxiv.org/abs/1206.4625
[91] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg, ‘Comparative analysis of multidimensional, quantitative data’, *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1027–1035, Dec. 2010, doi: 10.1109/TVCG.2010.138.
[92] A. Lex, M. Streit, E. Kruijff, and D. Schmalstieg, ‘Caleydo: Design and evaluation of a

visual analysis framework for gene expression data in its biological context', in *2010 IEEE Pacific Visualization Symposium (PacificVis)*, Mar. 2010, pp. 57–64. doi: 10.1109/PACIFICVIS.2010.5429609.
[93] J. Seo and B. Shneiderman, 'Interactively exploring hierarchical clustering results [gene identification]', *Computer*, vol. 35, no. 7, pp. 80–86, Jul. 2002, doi: 10.1109/MC.2002.1016905.
[94] J. Zhou, S. Konecni, and G. Grinstein, 'Visually comparing multiple partitions of data with applications to cluste', Jan. 2009, vol. 7243, p. 72430. doi: 10.1117/12.810093.
[95] M. Meilă, 'Comparing Clusterings by the Variation of Information', in *Learning Theory and Kernel Machines*, vol. 2777, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 173–187. doi: 10.1007/978-3-540-45167-9_14.[2]\uc0\u160{} K. You and M. K. You, 'Package \uc0\u8220{}mclustcomp\uc0\u8221{}", 2018.
[97] S. Wagner and D. Wagner, 'Comparing Clusterings - An Overview', p. 19.
[98] L. I. Kuncheva and S. T. Hadjitodorov, 'Using diversity in cluster ensembles', in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, The Hague, Netherlands, 2004, vol. 2, pp. 1214–1219. doi: 10.1109/ICSMC.2004.1399790.
[99] W. M. Rand, 'Objective Criteria for the Evaluation of Clustering Methods', *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.
[100] E. B. Fowlkes and C. L. Mallows, 'A Method for Comparing Two Hierarchical Clusterings', *J. Am. Stat. Assoc.*, vol. 78, no. 383, pp. 553–569, Sep. 1983, doi: 10.1080/01621459.1983.10478008.
[101] L. Hubert and P. Arabie, 'Comparing partitions', *J. Classif.*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/BF01908075.
[102] V. M.K and K. K, 'A Survey on Similarity Measures in Text Mining', *Mach. Learn. Appl. Int. J.*, vol. 3, no. 1, pp. 19–28, Mar. 2016, doi: 10.5121/mlaij.2016.3103.
[103] D. L. Wallace, 'A Method for Comparing Two Hierarchical Clusterings: Comment', *J. Am. Stat. Assoc.*, vol. 78, no. 383, pp. 569–576, 1983, doi: 10.2307/2288118.
[104] M. Meilă and D. Heckerman, 'An Experimental Comparison of Model-Based Clustering Methods', *Mach. Learn.*, vol. 42, no. 1, pp. 9–29, Jan. 2001, doi: 10.1023/A:1007648401407.
[105] S. van Dongen, 'Performance criteria for graph clustering and Markov cluster experiments', . *Introduction*, p. 38.
[106] A. Strehl and J. Ghosh, 'Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions', p. 35.
[107] L. N. F. Ana and A. K. Jain, 'Robust data clustering', in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Jun. 2003, vol. 2, p. II–II. doi: 10.1109/CVPR.2003.1211462.
[108] 'THE THERMAL EFFICIENCY OF STEAM ENGINES. REPORT OF THE COMMITTEE APPOINTED TO THE COUNCIL UPON THE SUBJECT OF THE DEFINITION OF A STANDARD OR STANDARDS OF THERMAL EFFICIENCY FOR STEAM ENGINES: WITH AN INTRODUCTORY NOTE. (INCLUDING APPENDIXES AND PLATE AT BACK OF VOLUME). | Minutes of the Proceedings of the Institution of Civil Engineers'. https://www.icevirtuallibrary.com/doi/10.1680/imotp.1898.19100 (accessed Mar. 18, 2021).
[109] M. Schmidt, 'The Sankey Diagram in Energy and Material Flow Management', *J. Ind. Ecol.*, vol. 12, no. 1, pp. 82–94, 2008, doi: https://doi.org/10.1111/j.1530-9290.2008.00004.x.
[110] Y. Jang, J. Seo, I. Jang, B. Lee, S. Kim, and S. Lee, 'CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data', *Bioinforma. Oxf. Engl.*, vol. 35, no. 24, pp. 5341–5343, Dec. 2019, doi: 10.1093/bioinformatics/btz516.
[111] S. H. Tausch *et al.*, 'LiveKraken—real-time metagenomic classification of illumina

data', *Bioinformatics*, vol. 34, no. 21, pp. 3750–3752, Nov. 2018, doi: 10.1093/bioinformatics/bty433.
[112] A. Fernández and S. Gomez, 'Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms', *J. Classif.*, vol. 25, pp. 43–65, Feb. 2008, doi: 10.1007/s00357-008-9004-x.
[113] M. Streit and N. Gehlenborg, 'Bar charts and box plots', *Nat. Methods*, vol. 11, no. 2, p. 117, Feb. 2014, doi: 10.1038/nmeth.2807.
[114] T. M. J. Fruchterman and E. M. Reingold, 'Graph drawing by force-directed placement', *Softw. Pract. Exp.*, vol. 21, no. 11, pp. 1129–1164, 1991, doi: https://doi.org/10.1002/spe.4380211102.
[115] E. M. Reingold and J. S. Tilford, 'Tidier Drawings of Trees', *IEEE Trans. Softw. Eng.*, vol. SE-7, no. 2, pp. 223–228, Mar. 1981, doi: 10.1109/TSE.1981.234519.
[116] R. Davidson and D. Harel, 'Drawing graphs nicely using simulated annealing', *ACM Trans. Graph.*, vol. 15, no. 4, pp. 301–331, Oct. 1996, doi: 10.1145/234535.234538.
[117] P. Flajolet and M. Noy, 'Analytic Combinatorics of Chord Diagrams', in *Formal Power Series and Algebraic Combinatorics*, Berlin, Heidelberg, 2000, pp. 191–201. doi: 10.1007/978-3-662-04166-6_17.
[118] E. Marchiori, *Local Network Community Detection with Continuous Optimization of Conductance and Weighted Kernel K-Means Twan van Laarhoven*. 2016.
[119] M. Ramanath, 'Tracking the Conductance of Rapidly Evolving Topic-Subgraphs', *Maya Ramanath*, May 23, 2019. /~ramanath/publication/dblp-journalspvldb-galhotra-bbrj-15/ (accessed May 20, 2021).
[120] A.-C. Gavin *et al.*, 'Proteome survey reveals modularity of the yeast cell machinery', *Nature*, vol. 440, no. 7084, pp. 631–636, Mar. 2006, doi: 10.1038/nature04532.
[121] T. Theodosiou *et al.*, 'NAP: The Network Analysis Profiler, a web tool for easier topological analysis and comparison of medium-scale biological networks', *BMC Res. Notes*, vol. 10, no. 1, p. 278, Jul. 2017, doi: 10.1186/s13104-017-2607-8.
[122] S. van Dongen and C. Abreu-Goodger, 'Using MCL to Extract Clusters from Networks', in *Bacterial Molecular Networks: Methods and Protocols*, J. van Helden, A. Toussaint, and D. Thieffry, Eds. New York, NY: Springer, 2012, pp. 281–295. doi: 10.1007/978-1-61779-361-5_15.
[123] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, 'Fast unfolding of communities in large networks', *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.
[124] P. Pons and M. Latapy, 'Computing Communities in Large Networks Using Random Walks', in *Computer and Information Sciences - ISCIS 2005*, Berlin, Heidelberg, 2005, pp. 284–293. doi: 10.1007/11569596_31.
[125] P. Kontou *et al.*, 'Identification of gene expression profiles in myocardial infarction: a systematic review and meta-analysis', *BMC Med. Genomics*, vol. 11, no. 1, p. 109, Nov. 2018, doi: 10.1186/s12920-018-0427-x.
[126] T. Barrett and R. Edgar, 'Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*', *Methods Mol. Biol. Clifton NJ*, vol. 338, pp. 175–190, 2006, doi: 10.1385/1-59745-097-9:175.
[127] R. Suresh *et al.*, 'Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction', *J. Mol. Cell. Cardiol.*, vol. 74, pp. 13–21, Sep. 2014, doi: 10.1016/j.yjmcc.2014.04.017.
[128] H.-J. Park *et al.*, 'Assessment and diagnostic relevance of novel serum biomarkers for early decision of ST-elevation myocardial infarction', *Oncotarget*, vol. 6, no. 15, pp. 12970–12983, May 2015, doi: 10.18632/oncotarget.4001.
[129] E. D. Muse *et al.*, 'A Whole Blood Molecular Signature for Acute Myocardial Infarction', *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, Sep. 2017, doi: 10.1038/s41598-017-12166-0.
[130] T. A. Hait *et al.*, 'The EXPANDER Integrated Platform for Transcriptome Analysis', *J. Mol. Biol.*, vol. 431, no. 13, pp. 2398–2406, Jun. 2019, doi: 10.1016/j.jmb.2019.05.013.

[131] I. Letunic and P. Bork, ‘Interactive Tree Of Life (iTOL) v4: recent updates and new developments’, *Nucleic Acids Res.*, vol. 47, no. W1, pp. W256–W259, Jul. 2019, doi: 10.1093/nar/gkz239.

Παράρτημα Ι

Δημοσίευση

VICTOR: A visual analytics web application for comparing cluster sets

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

Computers in Biology and Medicine, Elsevier, 04 June 2021, https://doi.org/10.1016/j.compbiomed.2021.104557

Preprint

VICTOR: A visual analytics web application for comparing cluster sets

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

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

- **VICTOR application**
  http://bib.fleming.gr:3838/VICTOR.

- **VICTOR source code**
  https://github.com/PavlopoulosLab/VICTOR.