

Πολυτεχνείο Κρήτης

Τμήμα Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών



Δίκτυα Έκφρασης για Γονιδιακούς Δείκτες Καρκίνου

Συγγραφέας:
Δημήτριος-Απόστολος
Χαλεπάκης-Ντελλής

Επιβλέπων Καθηγητής:
Καθ. Μιχάλης Ζερβάκης

Εξεταστική Επιτροπή:
Καθ. Μιχάλης Ζερβάκης
Αν.Καθ. Κατερίνα Μανιά
Κύρ.Ερευν. Δημήτρης Καφετζόπουλος

Ιούλιος, 2015

Ευχαριστίες

Θα ήθελα να ευχαριστήσω

Τον Καθηγητή κύριο Μιχάλη Ζερβάκη, για την πολύτιμη βοήθεια και καθοδήγησή του κατά τη διάρκεια όλης της εκπόνησης της διπλωματικής μου εργασίας.

Την Καθηγήτρια Κυρία Αικατερίνη Μανιά και τον Ερευνητή κύριο Δημήτριο Καφετζόπουλο για τη συνεισφορά τους ως μέλη της εξεταστικής επιτροπής.

Την Δρ. Αικατερίνη Μπέη για την τεράστια συνεισφορά της στην εκπόνηση της εργασίας αλλά και για την αμέριστη στήριξη και συμπαράσταση της όλο αυτό το χρονικό διάστημα.

Την υποψήφιο διδάκτορα Καλλιόπη Καλαντζάκη για την παροχή των πολύτιμων γνώσεων της κατά τη διάρκεια της εργασίας.

Τον υποψήφιο διδάκτορα Στέλιο Σφακιανάκη για την παροχή και βοήθεια του με το dataset και τις χρήσιμες γνώσεις του.

Την φίλη μου Βέρα Κλοκοτίνη για την καθημερινή και αμέριστη στήριξη και υπομονή της καθ' όλη τη διάρκεια της εκπόνησης της εργασίας αλλά και των σπουδών μου.

Την οικογένεια μου και όλους μου τους φίλους για την αγάπη, τη στήριξη και την εμπιστοσύνη που μου έδειξαν όλα τα χρόνια της φοιτητικής μου ζωής.

Περίληψη

Τα κύτταρα περιέχουν χιλιάδες βιολογικά μόρια, όπως γονίδια, RNA, πρωτεΐνες και μεταβολίτες, που αλληλεπιδρούν με πολύπλοκους τρόπους. Τα δίκτυα παρουσιάζουν ένα ισχυρό πλαίσιο για να αντιπροσωπεύσουν αυτές τις περίπλοκες σχέσεις και αλληλεπιδράσεις, οι οποίες είναι υπεύθυνες για διάφορες κυτταρικές λειτουργίες με τις επιδράσεις των μεμονωμένων κόμβων των μορίων. Στην παρούσα διπλωματική εργασία πραγματοποιείται η εκμάθηση δομής Bayesian δικτύων από δεδομένα γονιδιακής έκφρασης που προέρχονται από δείγματα παθολογίας του καρκίνου του μαστού. Τα Bayesian δίκτυα παρέχουν μια καλή και εύχρηστη αναπαράσταση για την έκφραση της από κοινού κατανομής πιθανότητας και για τον συμπερασμό. Η αναπαράσταση και η χρήση της θεωρίας πιθανοτήτων κάνουν τα Bayesian δίκτυα κατάλληλα για το συνδυασμό του πεδίου γνώσης και δεδομένων, την έκφραση αιτιωδών σχέσεων, την αποφυγή υπερκάλυψης ενός μοντέλου σε δεδομένα εκπαίδευσης, και την εκμάθηση από μη ολοκληρωμένα σύνολα δεδομένων. Συγκεκριμένα, για να γίνει η εκμάθηση μιας τέτοιας δομής, χρησιμοποιήθηκαν οι αλληλεπιδράσεις των 77 γονιδίων που είναι μια γονιδιακή υπογραφή που σχετίζεται με την παθολογία του καρκίνου του μαστού. Κατασκευάστηκαν δομές ξεχωριστά για τα καρκινικά και τα υγιή δείγματα ενώ η εκμάθηση των δομών έγινε σύμφωνα με τον αλγόριθμο εκμάθησης δομής K2, θεωρώντας τις μεταβλητές διακριτές αλλά και συνεχείς. Οι δομές που προέκυψαν μελετήθηκαν ως προς τις ιδιότητες Μικρού Κόσμου και Ελεύθερης Κλίμακας, που εμφανίζονται στην πλειονότητα των δικτύων του πραγματικού κόσμου. Επίσης αναζητήθηκαν σημαντικοί κόμβοι, συμπλέγματα (σύμφωνα με τον αλγόριθμο MCODE) και δομικές ενότητες (σύμφωνα με τον αλγόριθμο jActiveModules) στις δομές, τα οποία αξιολογήθηκαν στατιστικά και βιολογικά. Η στατιστική ανάλυση των δικτύων έδειξε ότι εμφανίζουν την ιδιότητα Ελεύθερης Κλίμακας που συνάδει με τη βιολογική τους διάσταση όπως επίσης και ότι υπάρχουν σημαντικοί κόμβοι, συμπλέγματα και δομικές ενότητες στα δίκτυα. Η ανάλυση των Bayesian δικτύων ανέδειξε δυναμικά υποδίκτυα με κεντρικούς κόμβους τα οποία προσφέρουν νέα γνώση σχετικά με τα βιολογικά μονοπάτια που εμπλέκονται στον καρκίνο του μαστού.

Abstract

Cells contain thousands of organic molecules, such as genes, RNA, proteins and metabolites which interact in complex ways. The networks provide a powerful framework to represent these complex relationships and interactions, which are responsible for various cellular functions with the effects of individual nodes of the molecules. In this thesis is carried out the learning of Bayesian network structure from gene expression data from breast cancer pathology samples. The Bayesian networks provide a neat and compact representation for expressing joint probability distribution and for inference. The representation and the use of probability theory makes Bayesian networks suitable for combining domain knowledge and data, expressing causal relationships, avoiding overfitting a model to training data, and learning from incomplete datasets. Specifically, to learn such a structure, have been used interactions of 77 genes, which is a gene signature associated with the pathology of breast cancer. Structures were constructed separately for both cancer and control samples, while learning structures were made according to the structure learning algorithm K2, considering the discrete and continuous variables. The resulting structures studied in properties of Small-World and Scale-Free, shown in most real world networks. Furthermore, important nodes as well as complexes (according to the algorithm MCODE) and modules (according to the algorithm jActive Modules) were searched out in the structures which were statistically and biologically evaluated. Statistical analysis showed that the networks have the property of Scale-Free, which is consistent with their biological dimension and that there are important hubs, clusters and modules in the networks. The Bayesian networks analysis dynamically highlighted subnetworks with central hubs which offer new knowledge of the biological pathways involved in cancer breast.

Περιεχόμενα

1 Εισαγωγή	4
1.1 Ανάλυση του Προβλήματος	4
1.2 Υφιστάμενη Γνώση	6
1.3 Δομή της Εργασίας	8
2 Υπόβαθρο Εργασίας	10
2.1 Βιολογικό Υπόβαθρο	10
2.1.1 Δεσοξυριβονουκλεϊκό οξύ	10
2.1.2 Χρωμοσώματα	11
2.1.3 DNA	11
2.1.4 Γονίδια	11
2.1.5 Γονιδίωμα	12
2.1.6 Χαρακτηρισμός μεταγραφικών προτύπων	13
2.1.7 Μικροσυστοιχίες DNA	14
2.1.8 Μοριακή Υπογραφή	16
2.1.9 Πρωτεϊνικές αλληλεπιδράσεις	17
2.1.10 Μοριακές αλληλεπιδράσεις	18
2.2 Μαθηματικό Υπόβαθρο	20
2.2.1 Στοιχεία θεωρίας πιθανοτήτων	20
2.2.2 Θεωρία Γράφων	21
2.3 Δίκτυα	22
2.3.1 Βιολογικά Δίκτυα	22
2.4 Ιδιότητες Δικτύων	24
2.4.1 Ορισμοί βασικών ιδιοτήτων των δικτύων	24
2.4.2 Το φαινόμενο του Μικρού Κόσμου	25
2.4.3 Ομαδοποίηση	26
2.4.4 Κατανομή Βαθμού	28
2.4.5 Δίκτυα Ελεύθερης Κλίμακας	29
2.4.6 Ενδιάμεση Κεντρικότητα	30
2.4.7 Κεντρικότητα Εγγύτητας	31
2.5 Bayesian Δίκτυα	32
2.5.1 Αναπαράσταση Bayesian Δικτύων	33
2.5.2 Ορισμός Bayesian δικτύου	34

2.5.3	Κανόνας αλυσίδας	34
2.5.4	Υπό Συνθήκη ανεξαρτησία σε Bayesian δίκτυα	36
2.5.5	d-διαχωρισμός	36
3	Ανάπτυξη Μεθοδολογίας	41
3.1	Μέθοδος	41
3.2	Δεδομένα προς επεξεργασία	41
3.3	Δημιουργία Bayesian Δικτύων	43
3.3.1	Δεσμευμένη Κατανομή Πιθανότητας	43
3.3.2	Διακριτές Μεταβλητές	45
3.3.3	Συνεχείς Μεταβλητές	51
3.4	Εκμάθηση Δομής Δικτύου	51
3.5	Οπτικοποίηση Bayesian Δικτύων	54
3.6	Ανάλυση Δικτύων	54
3.6.1	Φαινόμενο Μικρού Κόσμου	55
3.6.2	Δίκτυα Ελεύθερης Κλίμακας	55
3.6.3	Άλλες Ιδιότητες	55
3.7	Σύγκριση Δικτύων	55
3.8	Ομαδοποίηση και Ανάλυση Ομάδων	56
3.8.1	Ομαδοποίηση με τον αλγόριθμο MCODE	57
3.8.2	Ομαδοποίηση με τον αλγόριθμο jActiveModules	59
3.9	Σύγκριση Υποδικτύων	61
3.10	Αναζήτηση Κεντρικών Κόμβων	61
3.11	Προτεινόμενη Μεθοδολογία	62
4	Αποτελέσματα - Συζήτηση	65
4.1	Δημιουργία Bayesian Δικτύων	65
4.2	Ανάλυση Δικτύων	66
4.2.1	Ανάλυση Δικτύων - Φαινόμενο Μικρού Κόσμου	66
4.2.2	Ανάλυση Δικτύων - Δίκτυα Ελεύθερης Κλίμακας	67
4.2.3	Ανάλυση Δικτύων - Άλλες Ιδιότητες	71
4.2.4	Ανάλυση Δικτύων - Σύνοψη	71
4.3	Σύγκριση Δικτύων	71
4.4	Ομαδοποίηση με τον αλγόριθμο MCODE	75
4.5	Ομαδοποίηση με τον αλγόριθμο jActiveModules	79
4.6	Σύγκριση Υποδικτύων	83
4.7	Αναζήτηση Κεντρικών Κόμβων	83
4.7.1	Κεντρικοί Κόμβοι και Σύγκριση Δικτύων	85
4.7.2	Κεντρικοί Κόμβοι και Ομαδοποίηση με το εργαλείο MCODE	86
4.7.3	Κεντρικοί Κόμβοι και Ομαδοποίηση με το εργαλείο jActiveModules	86
4.7.4	Σύνοψη	87
4.8	Βιολογική Ερμηνεία των Αποτελεσμάτων	87

5 Συμπεράσματα	94
Παράρτημα Α	105
Παράρτημα Β	106
Παράρτημα C	115
Παράρτημα D	116
Παράρτημα E	119
Παράρτημα F	124

Κεφάλαιο 1

Εισαγωγή

1.1 Ανάλυση του Προβλήματος

Βιολογικά συστήματα βασισμένα σε δίκτυα ή δικτυακή βιολογία, μια ραγδαία αναπτυσσόμενη περιοχή που μελετά διάφορα βιομοριακά δίκτυα, είναι μια διεπιστημονική τομή μαθηματικών, επιστήμης υπολογιστών και βιολογίας. Τα τελευταία χρόνια, συσσωρευμένα δεδομένα υψηλής απόδοσης έχουν οδηγήσει την ολοκληρωμένη μελέτη από την περιγραφή πολύπλοκων φαινομένων στην κατανόηση απαραίτητων αρχών και από την μελέτη μεμονωμένων στοιχείων στην κατανόηση λειτουργικών δικτύων για βιομοριακά συστήματα, όπως κύτταρα, ιστοί, όργανα αλλά και ολόκληρους οργανισμούς. Το δίκτυο παρέχει ένα ποσοτικό εναλλακτικό τρόπο για την μοντελοποίηση της πολύπλοκης σχέσης ανάμεσα σε αυτούς τους στόχους, οι οποίοι έχουν εμπνεύσει ουσιαστικά μια νέα άποψη για την κυτταρική βιολογία. Η δικτυακή βιολογία στοχεύει να διερευνήσει πως τα κυτταρικά συστήματα διευκολύνουν στο να προκύψουν βιολογικοί φαινότυποι με τη μορφή δικτύων μεταξύ γονιδίων, πρωτεϊνών και μεταβολικών. Οι ραγδαίες εξελίξεις στην δικτυακή βιολογία, παράγουν αποτελεσματικά πολύτιμες μεθόδους για την ανάλυση δεδομένων υψηλής απόδοσης, επηρεάζουν σημαντικά την κατανόηση βιολογικών συστημάτων και ενσωματώνονται σε σημαντικές εφαρμογές της πρακτικής ιατρικής. Ειδικότερα, η εμφάνιση και η ανάπτυξη μιας σύνθετης νόσου, ένας από τους πιο σημαντικούς βιολογικούς φαινοτύπους, μπορεί να θεωρηθεί ως η χωροχρονική μεταβολή των αντίστοιχων βιομοριακών δικτύων. Η μελέτη που βασίζεται σε βιομοριακά δίκτυα προσφέρει ένα νέο εννοιολογικό πλαίσιο που θα μπορούσε δυνητικά να φέρει επανάσταση στις απόψεις μας σχετικά με την βιολογία και τις παθολογικές καταστάσεις [1].

Οι δυσκολίες σε αυτήν την περιοχή προκύπτουν από τα ακόλουθα τρία σημεία:

- Η εμφάνιση και η ανάπτυξη των πολύπλοκων ασθενειών περιλαμβάνουν πολλούς γενετικούς παράγοντες, επιγενετικούς παράγοντες (όπως περιβαλλοντικοί παράγοντες) και τις αλληλεπιδράσεις τους.
- Διαφορετικά είδη κυττάρων και ιστών περιπλέκονται στην ανάπτυξη και εξέλιξη των ασθενειών.

- Η μοριακή βλάβη μίας πολύπλοκης νόσου διαταράσσει την φυσιολογική συμπεριφορά των πολύπλοκων μοριακών δικτύων των γονιδίων και των πρωτεϊνών.

Πολύπλοκες ασθένειες, όπως ο καρκίνος, έχουν μελετηθεί εκτενώς για μεγάλο χρονικό διάστημα. Η κλασική βιοϊατρική που βασίζεται στην μοριακή βιολογία, κυτταρική βιολογία, γενετική και άλλη πειραματική βιολογία, έχει κάνει σημαντική πρόοδο στην αποσαφήνιση του μηχανισμού των πολύπλοκων ασθενειών. Όμως, οι ερευνητές της βιοϊατρικής αντιμετωπίζουν μεγάλη δυσκολία απέναντι σε πολύπλοκες ασθένειες καθώς η μεθοδολογία της κλασικής πειραματικής βιολογίας βασίζεται κυρίως στην μελέτη συγκεκριμένων γονιδίων και πρωτεϊνών. Επιπλέον, αντιμετωπίζει έναν οργανισμό ως γραμμικό σύστημα, το οποίο είναι απλοποιημένο και ανεπαρκές για την κατανόηση της εμφάνισης και ανάπτυξης πολύπλοκων ασθενειών. Αντίθετα, η συστημική βιολογία, μας παρέχει ένα νέο τρόπο να κατανοήσουμε τις μη-γραμμικές αλληλεπιδράσεις πολλαπλών βιομοριακών στοιχείων και να χαρακτηρίσουμε την αλλαγή του φαινοτύπου αναλύοντας την δυναμική συμπεριφορά των βιολογικών δικτύων. Αυτό ίσως ρίξει φως σε πολυπαραγοντικές ασθένειες με σύνθετη αιτιολογία ώστε να εντοπιστούν βιολογικά μονοπάτια ή δίκτυα που συνδέουν τα διαφορετικά στοιχεία ενός συστήματος και εξετάζουν πως εξελίσσονται κάτω από χρονικές και περιβαλλοντικές μεταβολές [1].

Γονίδια, RNA, πρωτεΐνες και μεταβολίτες είναι γνωστά ως βιολογικά μόρια. Τα κύτταρα περιέχουν χιλιάδες τέτοιες μοριακές οντότητες που αλληλεπιδρούν με πολύπλοκους τρόπους. Τα δίκτυα παρουσιάζουν ένα ισχυρό πλαίσιο για να αντιπροσωπεύσουν αυτές τις περίπλοκες σχέσεις και αλληλεπιδράσεις, οι οποίες είναι υπεύθυνες για διάφορες κυτταρικές λειτουργίες με τις επιδράσεις των μεμονωμένων κόμβων των μορίων.

Σε αυτή τη διπλωματική εργασία έγινε γονιδιακή ανάλυση μέσω Bayesian δικτύων. Τα Bayesian δίκτυα χρησιμοποιούν και αναπαριστούν την θεωρία πιθανοτήτων κάτι το οποίο τα κάνουν κατάλληλα για το συνδυασμό του πεδίου γνώσης και δεδομένων στις βιολογικές επιστήμες. Τα δεδομένα προήλθαν από την ανάλυση μικροσυστοιχιών με δείγματα ασθενών του καρκίνου του μαστού αλλά και υγιών ανθρώπων. Η κατασκευή της δομής των δικτύων έγινε με τη χρήση ενός αλγορίθμου που βασίζεται στην λειτουργία βαθμονόμησης. Δημιουργήθηκαν διάφορα δίκτυα από διακριτές αλλά και συνεχείς (Γκαουσιανές) μεταβλητές. Εκμεταλλευόμενοι την δομή των δικτύων, αναζητήσαμε κάποιες ιδιότητες που εμφανίζουν δίκτυα του πραγματικού κόσμου όπως Small-World και Scale-Free. Επίσης ομαδοποιήσαμε τα δίκτυα σε συμπλέγματα με τη βοήθεια του αλγορίθμου MCODE και δομικές ενότητες με τον αλγόριθμο jActiveModules έτσι ώστε να βρούμε σημαντικές μονάδες μέσα σε αυτά, συγκρίναμε τα δίκτυα αλλά και τις ενότητες με διάφορες μεθόδους και τέλος αναζητήσαμε σημαντικούς κόμβους μέσα σε αυτά.

1.2 Υφιστάμενη Γνώση

Παραδοσιακά, η μελέτη ασθενειών επικεντρωνόταν σε μεμονωμένα γονίδια ή πρωτεΐνες συσχετισμένα με ορισμένους φαινοτύπους. Προσεγγίσεις που βασίζονται σε δίκτυα, παρέχουν νέες εναλλακτικές λύσεις για την μελέτη πολύπλοκων ασθενειών συγχρόνως με τη διερεύνηση ομάδων σχετικών μορίων. Σε πρωτογενικό επίπεδο, οι πρωτεΐνες αλληλεπιδρούν με άλλες πρωτεΐνες για να σχηματίσουν σύμπλοκα πρωτεΐνης και μονοπάτια για να συμμετάσχουν σε διάφορες βιολογικές διεργασίες. Οι εν λόγω ενώσεις μεταξύ των πρωτεϊνικών μορίων ονομάζονται αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης (PPIs), οι οποίες οδηγούν σε ένα PPI δίκτυο. Επιπλέον, συχνά παρουσιάζονται οι ενδιαφέρουσες βιολογικές οντότητες και οι σχέσεις τους με λειτουργικά δίκτυα σύνδεσης. Σε ένα πάσχων βιολογικό σύστημα, οι δυσλειτουργικές διεργασίες είναι υποκείμενες σε αυτά τα διάφορα δίκτυα.

Ένας αριθμός ορισμών έχουν εισαχθεί για να περιγράψουν τις πολύπλοκες σχέσεις αυτών των κόμβων στο δίκτυο. Η τοπολογική ανάλυση είναι το πρώτο βήμα για την κατανόηση των βιολογικών λειτουργιών και των αρχών σχεδιασμού του δικτύου. Τα τοπολογικά χαρακτηριστικά των κόμβων (δηλαδή μεμονωμένων μορίων) και των ακμών (π.χ. αλληλεπιδρώντα ζεύγη) σε ένα δίκτυο, μοτίβα δικτύου και μονάδων (δηλαδή των τοπικών δομών), καθώς και οι καθολικές δομές δικτύου παρουσιάζουν πολύτιμες πληροφορίες για τον εντοπισμό γονιδίων που έχουν σχέση με κάποια νόσο, δυσλειτουργικά μονοπάτια, disease-responsive υποδίκτυα, καθώς και για τις σχέσεις της νόσου. Το δίκτυο ωφελεί επίσης στον σχεδιασμό των φαρμακευτικών στόχων και στον συνδυασμό φαρμάκων για τη θεραπεία των πολύπλοκων ασθενειών [1].

Έχουν αναπτυχθεί αρκετές μέθοδοι βασισμένες σε δίκτυα για την πρόβλεψη παθολογικών γονιδίων. Μερικές από αυτές φαίνονται στον παρακάτω Πίνακα 1.1.

Μέθοδος	Νόσος	Δεδομένα δικτύου	Λογισμικό	Αναφορά
Krauthammer κ.α.	Αλτσχάιμερ	PPI	molecular triangulation	[2]
Morrison κ.α.	γενικά	λειτουργικό δίκτυο σύνδεσης	GeneRank	[3]
Franke κ.α.	96 νόσοι	λειτουργικό δίκτυο γονιδίων	prioritiser	[4]
George κ.α.	γενικά	PPI	CPS και CMP	[5]
Aerts κ.α.	γενικά	PPI	ENDEAVOUR	[6]
Ma κ.α.	Αλτσχάιμερ	PPI	CGI	[7]
Lage κ.α.	γενικά	PPI	DGF	[8]
Kohler	γενικά	PPI	GeneWanderer	[9]
Wu κ.α.	γενικά	PPI	CIPHER	[10]

Πίνακας 1.1 – Συνοπτικός πίνακας για τις μεθόδους βασισμένες σε δίκτυα για την πρόβλεψη παθολογικών γονιδίων.

Οι τοπολογικές ιδιότητες ενός δικτύου διερευνήθηκαν αρκετά στο σύστημα της πρόβλεψης. Για παράδειγμα, το μήκος του κοντύτερου μονοπατιού ανάμεσα σε ένα υποψήφιο γονίδιο και ένα γνωστό για μια ασθένεια γονίδιο είναι μια ευρέως χρησιμοποιούμενη μέτρηση για την συσχέτιση τους με την ασθένεια. Η μοριακή τριγωνοποίηση [2] είναι μια μέθοδος που βασίζεται σε αυτό το σύστημα βαθμονόμησης.

Αρχικά κατασκευάζει ένα δίκτυο ανθρώπινων πρωτεϊνικών αλληλεπιδράσεων συσχετισμένες με την ασθένεια Αλτσχάιμερ σύμφωνα με τεχνικές από την βιβλιογραφία. Αφού αναθέσει μια τιμή αρχικής-απόδειξης σε ένα υποσύνολο παθολογικών γονιδίων και τους θέσει αρχικούς κόμβους στο δίκτυο, υπολογίζει μια τιμή δεύτερης-απόδειξης για να τα συνδυάσει τις πολλαπλές τιμές αρχικής-απόδειξης. Στη συνέχεια υπολογίζονται τα μήκη των κοντύτερων μονοπατιών από τους αρχικούς κόμβους. Το σκορ μετατρέπεται σε ένα p-value για να μετρήσει το δυσλειτουργικό ενδεχόμενο συσχετισμένο με το Αλτσχάιμερ. Η μέθοδος αυτή λειτουργεί καλά όταν υποθέσουμε ότι κοινοί φαινότυποι σχετίζονται με πρωτεϊνικές δυσλειτουργίες στο ίδιο δίκτυο. Στο [5] περιγράφεται μια άλλη μέθοδος πρόβλεψης υποψήφιων γονιδίων υπεύθυνα για κάποια ασθένεια από κοινά μονοπάτια. Χρησιμοποιεί δεδομένα του δικτύου που προέρχονται από βάσεις δεδομένων πρωτεϊνικών αλληλεπιδράσεων και μονοπατιών για να εντοπίσει σχέσεις ανάμεσα σε γονίδια. Επίσης χρησιμοποιεί το σύστημα βαθμονόμησης βάσει του μήκος των κοντύτερων μονοπατιών. Από την άλλη, αναπτύχθηκε ένα σύστημα [7] ιεράρχησης γονιδίων συσχετισμένα με την νόσο του Αλτσχάιμερ, συνδυάζοντας δίκτυα γονιδιακής έκφρασης και PPI χρησιμοποιώντας την θεωρία του τυχαίου πεδίου του Μαρκόφ. Επίσης συγκρίθηκαν συστήματα βαθμονόμησης από των κοντύτερο γείτονα και τα μήκος των κοντύτερων μονοπατιών. Για να διερευνηθεί η μέτρηση της ολικής απόστασης ανάμεσα σε υποψήφια γονίδια και τις πληροφορίες του δικτύου, προτάθηκε ένα σύστημα [9] απόκλισης του πυρήνα για να κατανοήσουν την ολοκληρωμένη συμβολή άλλων γονιδίων. Στο [4] γίνεται ένας συνδυασμός γονιδιακής οντολογίας (gene ontology-GO), γονιδιακής συν-έκφρασης και πρωτεϊνικών αλληλεπιδράσεων σύμφωνα με την Bayesian μέθοδο για να κατασκευαστεί ένα λειτουργικό δίκτυο και να ελεγχθεί η απόδοση του στην πρόβλεψη με διαφορετικούς συνδυασμούς από μια Γκαουσιανή συνάρτηση βαθμονόμησης σε 96 κληρονομικές ασθένειες. Στο [3] σχεδιάστηκε μια μέθοδος γονιδιακής ιεράρχησης για την ανάλυση δεδομένα μικροσυστοιχιών βασισμένη στον PageRank αλγόριθμο. Χρησιμοποιήθηκαν GO δίκτυα, δίκτυα γονιδιακής συν-έκφρασης και συνθετικά δίκτυα για αυτό το σκοπό. Μια συστηματική ανάλυση των συμπλοκών των ανθρώπινων πρωτεϊνών που περιλαμβάνουν γονιδιακά προϊόντα τα οποία εμπλέκονται σε διαφορετικές κατηγορίες ανθρώπινων ασθενειών δημιουργώντας ένα phenom-interactome δίκτυο παρουσιάστηκε στο [8]. Συλλέχθηκαν τα παθολογικά πρωτεϊνικά σύμπλοκα και καθορίστηκαν οι βαθμοί ομοιότητας των φαινοτύπων. Τα παθολογικά γονίδια ταξινομήθηκαν και εκλέχθηκαν σύμφωνα με την μέθοδο Bayesian πρόβλεψης. Ανάλογη δουλειά παρουσιάστηκε και στο [10]. Το ENDEAVOUR [6] προτάθηκε για να ενσωματώσει την πληροφορία από διάφορες πηγές, όπως η πληροφορία από την βιβλιογραφία, λειτουργικό σχολιασμό, μικροσυστοιχίες, πρωτεϊνικές αλληλεπιδράσεις και πολλά άλλα, ώστε να ταυτοποιηθούν τα γονίδια που είναι υπεύθυνα για μια νόσο. Δημιουργεί διαφορετική προτεραιότητα για αυτές τις πολλαπλές ετερογενείς πηγές δεδομένων, τα οποία στη συνέχεια συγχωνεύονται μαζί σε μια καθολική κατάταξη, χρησιμοποιώντας στατιστικές ταξινομήσεις.

Αρκετές υπολογιστικές μέθοδοι έχουν προταθεί για την ανίχνευση ενεργών μονοπατιών από ενσωματωμένη γονιδιακή έκφραση, πρωτεϊνικές αλληλεπιδράσεις και άλλες

διαθέσιμες πληροφορίες, μερικές από τις οποίες φαίνονται στον Πίνακα 1.2 [1].

Μέθοδος	Ενότητα	Πηγή Δεδομένων	Λογισμικό	Αναφορά
Ideker κ.α.	γενική	PPI/πρωτεϊνη-DNA/ μικροσυστοιχία	jActiveModules	[11]
Steffen κ.α.	σηματοδοτικό μονοπάτι	PPI/μικροσυστοιχία	Netsearch	[12]
Segal κ.α.	λειτουργική	PPI/μικροσυστοιχία	GeneXpress	[13]
Sohler κ.α.	γενική	λειτουργικό δίκτυο/ μικροσυστοιχία	ToPNet	[14]
Breitling κ.α.	γενική	PPI/μικροσυστοιχία/ GO	GiGA	[15]
Scott κ.α.	γενική	PPI/πρωτεϊνη-DNA/ ενδιαφέροντα γονίδια	ColorCoding	[16]
Liu κ.α.	διαβήτη	PPI/μικροσυστοιχία	GNEA	[17]
Nacu κ.α.	γενική	PPI/μικροσυστοιχία	GXNA	[18]
Bebek and Yang	σηματοδοτικό μονοπάτι	PPI/μικροσυστοιχία/ GO	Pathfinder	[19]
Dittrich κ.α.	γενική	PPI/μικροσυστοιχία	Heinz	[20]
Zhao κ.α.	σηματοδοτικό μονοπάτι	PPI/μικροσυστοιχία	ILP	[21]
Ulitsky and Shamir	γενική	PPI/μικροσυστοιχία	Matisse	[22]
Qiu κ.α.	μονοπάτι καρκίνου μαστού	PPI/μικροσυστοιχία	MILPs	[23]
Liu κ.α.	Αλτσχάιμερ	PPI/μικροσυστοιχία	Epathway	[24]

Πίνακας 1.2 – Συνοπτικός πίνακας των μεθόδων για τον εντοπισμό λειτουργικών ενοτήτων σε βιομοριακά δίκτυα.

Η συστημική βιολογία βασισμένη σε δίκτυα προσελκύει όλο και περισσότερο τους ερευνητές από τους τομείς της πειραματικής αλλά και της υπολογιστικής βιολογίας και αναμένεται να φέρει επανάσταση στην αντίληψη μας για τα κυτταρικά συστήματα [1].

1.3 Δομή της Εργασίας

Η οργάνωση των κεφαλαίων που ακολουθούν έχει ως εξής: Στο πρώτο κεφάλαιο περιγράφεται η ανάλυση του προβλήματος και το κίνητρο που μας ώθησε ώστε να χρησιμοποιήσουμε τα δίκτυα για την έρευνα μας. Επίσης σε αυτό το κεφάλαιο φαίνεται η υφιστάμενη γνώση για τη συστημική βιολογία βασισμένη σε δίκτυα και για τους διάφορους αλγόριθμους και τις μεθόδους που χρησιμοποιήσαμε. Στο δεύτερο κεφάλαιο περιγράφεται το υπόβαθρο αυτής της εργασίας ως προς την βιολογία αλλά και τα μαθηματικά-στατιστική και η θεωρία των δικτύων γενικά αλλά και των Bayesian. Στο τρίτο κεφάλαιο αναλύεται η μεθοδολογία που ακολουθήσαμε. Σε αυτό το κεφάλαιο, αρχικά περιγράφεται ο τρόπος κατασκευής μια δομής ενός Bayesian δικτύου από διακριτές και συνεχείς μεταβλητές. Επίσης περιγράφεται ο τρόπος με τον οποίο αναλύσαμε τις δομές για να εντοπίσουμε κάποιες ιδιότητες σε αυτά (Small-World, Scale-Free κ.α.) ενώ στη συνέχεια περιγράφονται μέθοδοι σύγκρισης και ομαδοποίησης των δικτύων με αλγόριθμους όπως ο MCODE και ο jActiveModules. Στο τέλος

αυτού του κεφαλαίου φαίνεται συγκεντρωτικά η προτεινόμενη μεθοδολογία. Στο τέταρτο κεφάλαιο παραθέτονται τα αποτελέσματα που προέκυψαν από την προτεινόμενη μεθοδολογία ενώ στο πέμπτο φαίνονται τα στατιστικά αλλά και βιολογικά συμπεράσματα στα οποία καταλήξαμε.

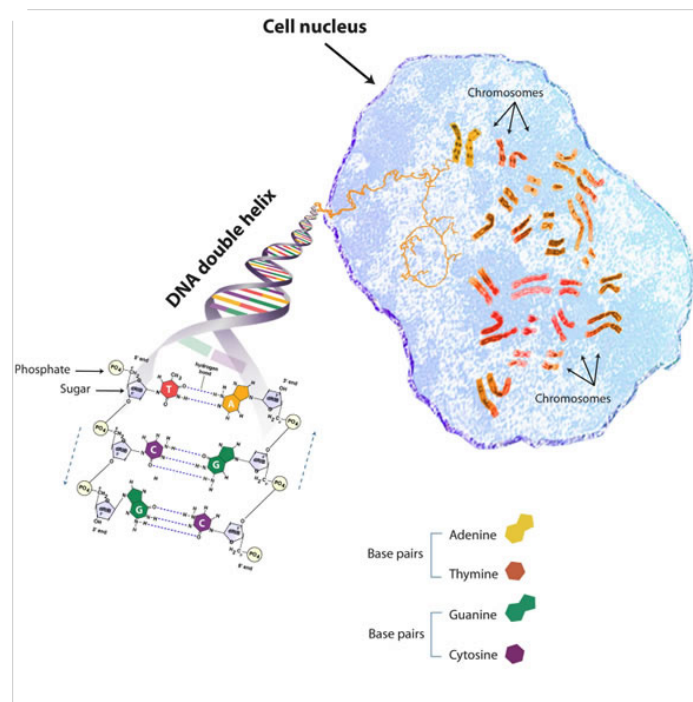
Κεφάλαιο 2

Υπόβαθρο Εργασίας

2.1 Βιολογικό Υπόβαθρο

2.1.1 Δεσοξυριβονουκλεϊκό οξύ

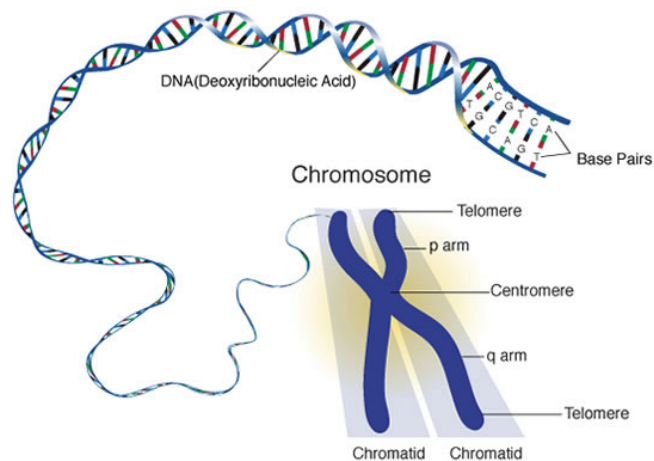
Το Δεσοξυριβονουκλεϊκό οξύ (DNA) είναι η βάση δεδομένων των χημικών πληροφοριών που φέρει το σύνολο των εντολών για το κύτταρο ως προς την φύση των πρωτεϊνών που παράγονται από αυτό, τη διάρκεια ζωής, την ωριμότητα, τη λειτουργία και τον θάνατο του. Τα γονίδια είναι οι 'υπομονάδες εργασίας' του DNA. Κάθε γονίδιο περιέχει ένα συγκεκριμένο σύνολο εντολών, όπως είναι για παράδειγμα η κωδίκευση μιας συγκεκριμένης πρωτεΐνης. [25].



Εικόνα 2.1 – Διπλή έλικά του DNA.

2.1.2 Χρωμοσώματα

Κάθε ανθρώπινο κύτταρο περιέχει 46 μόρια δίκλωνου DNA. Το DNA συσπειρώνεται και υπερ-συσπειρώνεται σε χρωμοσώματα. Κάθε χρωμόσωμα περιέχει περίπου 50 με 250 εκατομμύρια βάσεις.



Εικόνα 2.2 – Δομή χρωμοσώματος.

Τα ανθρώπινα κύτταρα περιέχουν δύο σαι χρωμοσωμάτων, το ένα σαι κληρονομείται από την μητέρα και το άλλο από τον πατέρα. Το ωάριο της μητέρας περιέχει τα μισά απο τα 46 (23) και το σπερματοζωάριο του πατέρα φέρει τα άλλα μισά 23 από τα 46 χρωμοσώματα. Με αυτό τον τρόπο το παιδί κληρονομεί όλα τα 46 χρωμοσώματα. Υπάρχουν 22 ζεύγη αυτοσωμικών χρωμοσωμάτων και ένα ζεύγος φυλετικών χρωμοσωμάτων. Οι γυναίκες έχουν δύο χρωμοσώματα X ενώ οι άνδρες έχουν ένα X και ένα Y [25].

2.1.3 DNA

Το DNA βρίσκεται στον πυρήνα κάθε κυττάρου ενός ευκαρυωτικού οργανισμού. Κάθε κύτταρο του ανθρώπινου οργανισμού (με εξαίρεση των ώριμων ερυθρών αιμοσφαιρίων, τα οποία δεν έχουν πυρήνα) περιέχει το ίδιο DNA.

Το DNA είναι δύο κλώνοι που σχηματίζουν μια διπλή έλικα. Κάθε κλώνος αποτελείται από εκατομμύρια χημικές δομικές μονάδες που καλούνται βάσεις. Υπάρχουν μόνο τέσσερις τύποι βάσεων που αποτελούν το DNA - η αδερίνη, η θυμίνη, η κυτοσίνη και η γουανίνη. Η διάταξη αυτών των βάσεων μέσα σε μια ακολουθία αλλάζει μέσω της μετάθεσης και του συνδυασμού, οδηγώντας σε μοναδικές ακολουθίες που κωδικοεύουν πρωτεΐνες [25].

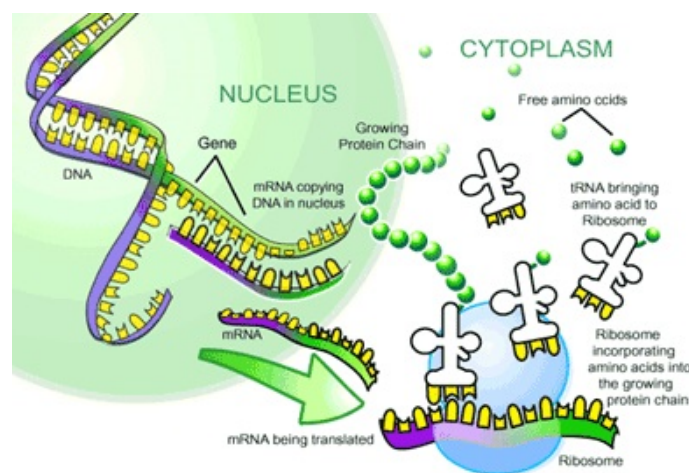
2.1.4 Γονίδια

Τα γονίδια όπως προαναφέρθηκε είναι συγκεκριμένες αλληλουχίες βάσεων του DNA, τα οποία περιέχουν αποθηκευμένη μία συγκεκριμένη γενετική πληροφορία. Υπάρχ-

ουν γονίδια τα οποία μεταγράφονται σε mRNA και μεταφράζονται σε πρωτεΐνες, καθορίζοντας την αλληλουχία των αμινοξέων σε μια πρωτεΐνη, ενώ υπάρχουν γονίδια που μεταγράφονται σε tRNA και rRNA.

Το μόριο που προκύπτει από την γονιδιακή έκφραση, RNA ή πρωτεΐνη, καλείται γονιδιακό προϊόν. Ένας μεγάλος αριθμός γονιδίων κωδικεύουν πρωτεΐνες, οι οποίες επιτελούν ποικίλους λειτουργικούς αλλά και δομικούς ρόλους. Μερικά γονίδια κωδικεύουν μόρια των μεταγραφόμενων από αυτά RNA, τα οποία διαδραματίζουν σημαντικούς ρόλους στον έλεγχο της γονιδιακής έκφρασης και στη βιοσύνθεση των πρωτεϊνών. Τα γονίδια μέσω των γονιδιακών προϊόντων τους ελέγχουν κάθε κυτταρική δραστηριότητα και κατευθύνουν τη φυσική ανάπτυξη και συμπεριφορά του οργανισμού.

Το σύνολο των γονιδίων αποτελεί τμήμα μόνο του γονιδιώματος ενός οργανισμού, που απαρτίζεται από το σύνολο του κυτταρικού DNA [26].



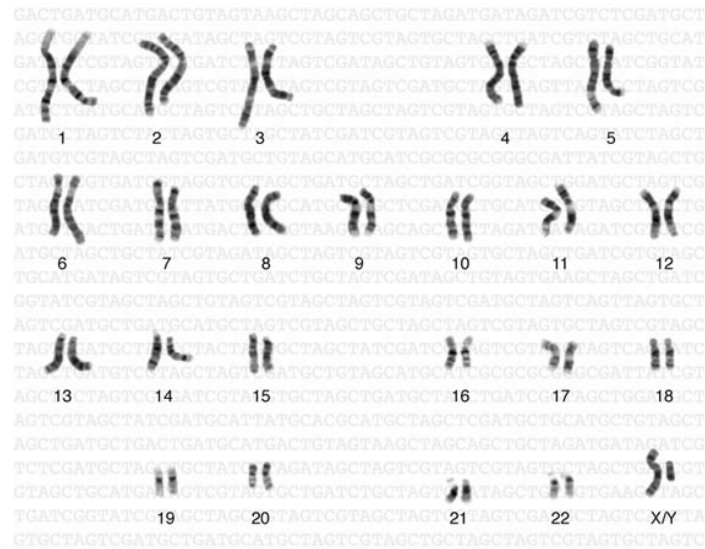
Εικόνα 2.3 – Διαδικασία δημιουργίας πρωτεΐνης.

2.1.5 Γονιδίωμα

Όλο το DNA σε ένα κύτταρο αποτελεί το ανθρώπινο *γονιδίωμα*. Υπάρχουν περίπου 20,000 σημαντικά γονίδια τα οποία βρίσκονται σε ένα από τα 23 ζεύγη χρωμοσωμάτων μέσα στον πυρήνα ή σε μακριές αλυσίδες του DNA που βρίσκονται στα μιτοχόνδρια.

Το DNA στα γονίδια αποτελεί μόνο περίπου το 2% του γονιδιώματος. Εδώ και μερικά χρόνια καθεμία από τις ακολουθίες και τα γονίδια που ανακαλύπτονται καταγράφονται προσεκτικά ως προς την ακριβή τους θέση, ακολουθίες τους και κάθε άλλη πληροφορία που υπάρχει για αυτά. Ολόκληρη η πληροφορία είναι αποθηκευμένη σε μία βάση δεδομένων που είναι δημόσια προσβάσιμη

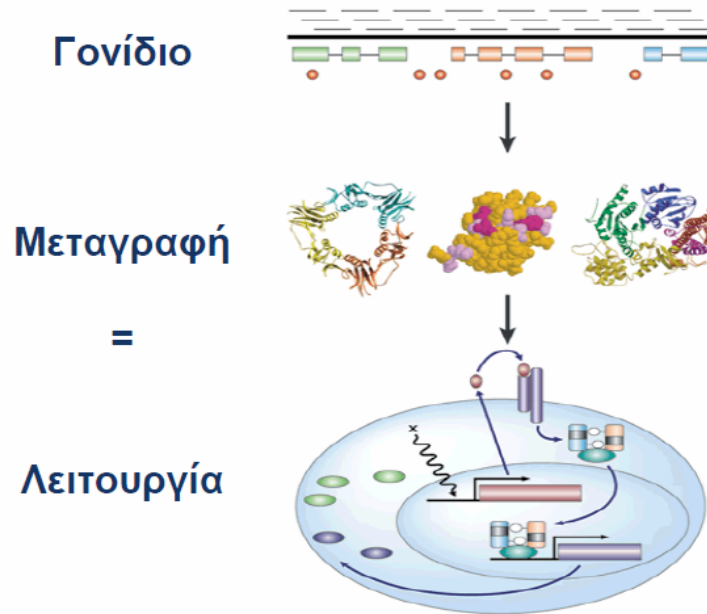
Περίπου 13.000 γονίδια έχουν χαρτογραφηθεί σε συγκεκριμένες θέσεις (loci) σε κάθε ένα από τα χρωμοσώματα. Αυτή η πληροφορία ξεκίνησε από την δουλειά που έγινε ως μέρος του Προγράμματος Αποκωδικοποίησης του Ανθρώπινου Γονιδιώματος (Human Genome Project). Η ολοκλήρωση αυτού του προγράμματος επιτεύχθηκε τον Απρίλη του 2003 αλλά ο ακριβής αριθμός των γονιδίων και πολλά άλλα γονίδια του ανθρώπινου γονιδιώματος είναι ακόμα άγνωστα.



Εικόνα 2.4 – Χρωμοσώματα του ανθρώπινου γονιδιώματος.

2.1.6 Χαρακτηρισμός μεταγραφικών προτύπων

Ο προσδιορισμός των μεταγραφικών προτύπων, δηλαδή των επιπέδων μεταγραφικής έκφρασης χιλιάδων γονιδίων σε ολόκληρο το γονιδίωμα, είναι ένα πολύ χρήσιμο εργαλείο για τους βιοεπιστήμονες σε μια προσπάθειά τους να αποκρυπτογραφήσουν τη σύνθετη οργάνωση των βιολογικών φαινομένων. Το πρότυπο γονιδιακής έκφρασης κάθε κυτταρικής σειράς, κάθε αναπτυξιακού σταδίου ή κάθε απόκρισης σε κάποια ουσία, αποτελεί ένα μοναδικό *μοριακό αποτύπωμα* το οποίο επιτρέπει την διάκριση μεταξύ διαφόρων τύπων ιστών, αναπτυξιακών σταδίων και ασθένειας (π.χ. καρκινικών ή μολυσμένων και μη κυττάρων, εμβρυικών και ενηλίκων ιστών, κυττάρων). Με πιο απλά λόγια, οι δομικές και λειτουργικές διαφορές μεταξύ διαφόρων τύπων κυττάρων ή ιστών καθορίζονται από την γονιδιακή έκφραση (ποια γονίδια και πόσο εκφράζονται). Η ανταπόκριση των κυττάρων ή ιστών σε φυσιολογικά (όπως φάρμακα και αναπτυξιακοί παράγοντες) ή παθολογικά *ερεθίσματα* (π.χ. μεταλλαγή) εκδηλώνεται με την αλλαγή αυτής [27].



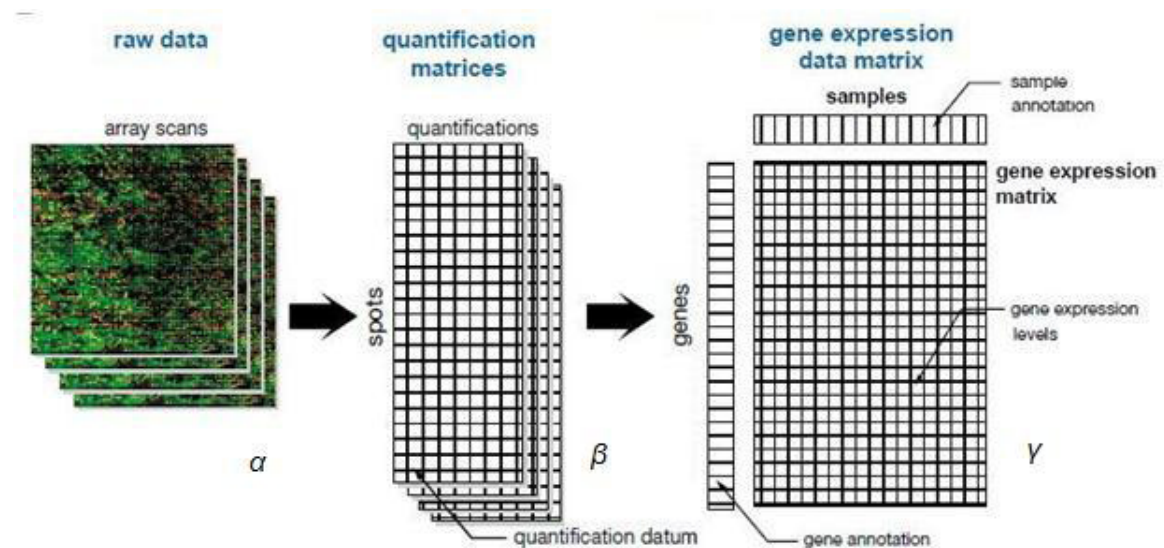
Εικόνα 2.5 – Προσδιορισμός μεταγραφικών παραγόντων.

2.1.7 Μικροσυστοιχίες DNA

Τα δεδομένα της γονιδιακής έκφρασης είναι αρκετά πολύπλοκα με την έννοια ότι έχουν νόημα μόνο στο πλαίσιο μιας αναλυτικής περιγραφής των συνθηκών κάτω από τις οποίες παράγονται, συμπεριλαμβανομένης και της ιδιαίτερης κατάστασης του συστήματος διαβίωσης που είναι υπό μελέτη και των διαταραχών στις οποίες έχει υποβληθεί. Σε αντίθεση με το γονιδίωμα ενός οργανισμού, υπάρχουν τόσα πολλά μεταγγραφα, όσα και οι κυτταρικοί τύποι που υπάρχουν αλλά και οι εκάστοτε περιβαλλοντικές συνθήκες στις οποίες αναπτύσσονται. Επιπλέον, η σύγκριση των δεδομένων γονιδιακής έκφρασης είναι σημαντικά δυσκολότερη, διότι επί του παρόντος, οι μικροσυστοιχίες δεν μετρούν επίπεδα γονιδιακής έκφρασης σε οποιεσδήποτε αντικειμενικές μονάδες. Στην πραγματικότητα, οι περισσότερες μετρήσεις αναφέρουν μόνο σχετικές αλλαγές στη γονιδιακή έκφραση, χρησιμοποιώντας μία (τιμή) αναφορά(ς) που είναι σπανίως τυποποιημένη. Τέλος, οι διαφορετικές πλατφόρμες μικροσυστοιχιών καθώς και τα πειραματικά σχέδια, παράγουν δεδομένα σε διαφορετικές μορφές και μονάδες και κανονικοποιούνται με διαφορετικούς τρόπους.

Τα ανεπεξέργαστα δεδομένα, από τα πειράματα μικροσυστοιχιών, είναι εικόνες από σαρώσεις μικροσυστοιχιών υβριδοποίησης, που πρέπει να αναλυθούν για να προσδιοριστεί και να ποσοτικοποιηθεί κάθε χαρακτηριστικό (spot) στην εικόνα. Μία ακολουθία DNA μπορεί να εντοπιστεί σε μία μικροσυστοιχία αρκετές φορές. Επιπλέον, μπορεί να εντοπιστούν αρκετές διακριτές ακολουθίες DNA που απεικονίζουν το ίδιο γονίδιο. Για να αποδοθεί μία μοναδική τιμή γι' αυτά, οι αντίστοιχες μετρήσεις πρέπει να συνδυαστούν. Επιπροσθέτως, η ίδια βιολογική συνθήκη μπορεί να μετρηθεί σε πολλαπλά αντίγραφα υβριδοποίησης, και η πληροφορία από όλα τα αντίγραφα πρέπει να συνοψιστεί για να εξαχθεί ένας πίνακας δεδομένων μοναδικών γονιδιακών εκφρά-

σεων. Τελικά, για να συγκριθούν τα δεδομένα μικροσυστοιχιών από αρκετά δείγματα, τα δεδομένα πρέπει να κανονικοποιηθούν κατάλληλα.



Εικόνα 2.6 – Εννοιολογική άποψη των τριών επιπέδων της επεξεργασίας των δεδομένων της μικροσυστοιχίας γονιδιακής έκφρασης. α. Ανεπεξέργαστα δεδομένα (row data), β. Ποσοτικοποίηση πινάκων μικροσυστοιχιών (quantification matrices), γ. Πίνακες δεδομένων γονιδιακής έκφρασης (gene expression data matrix)

Όπως φαίνεται στην Εικόνα 2.6, υπάρχουν, τουλάχιστον, τρία επίπεδα δεδομένων σχετικών με ένα πείραμα μικροσυστοιχιών:

1. οι σαρωμένες εικόνες (ανεπεξέργαστα δεδομένα, Εικόνα 2.6 α
2. οι ποσοτικές έξοδοι από τη διαδικασία ανάλυσης της εικόνας (ποσοτικοποίηση πινάκων μικροσυστοιχιών, Εικόνα 2.6 β
3. οι εξαγόμενες μετρήσεις (πίνακες δεδομένων γονιδιακής έκφρασης, Εικόνα 2.6 γ.

Το μοντέλο αυτού του τρίτου επιπέδου διαθέτει τρία τμήματα:

1. περιγραφή των γονιδίων, που μπορεί να διατίθεται σαν σύνδεσμος στις βάσεις δεδομένων της γονιδιακής ακολουθίας,
2. περιγραφή των δειγμάτων, για τα οποία δεν υπάρχουν τρέχουσες δημόσιες εξωτερικές βάσεις δεδομένων (εκτός από την ταξινόμηση των ειδών) και
3. τον πίνακα της γονιδιακής έκφρασης, στον οποίο κάθε θέση περιλαμβάνει πληροφορία που χαρακτηρίζει την έκφραση ενός συγκεκριμένου γονιδίου σε ένα συγκεκριμένο δείγμα (Εικόνα 2.6 γ) [28].

Τεχνολογίες γονιδιωματικής υψηλής απόδοσης (όπως μικροσυστοιχίες γονιδιακής έκφρασης) έχουν αλλάξει τρομερά την έρευνα στον τομέα της βιοϊατρικής στις μέρες

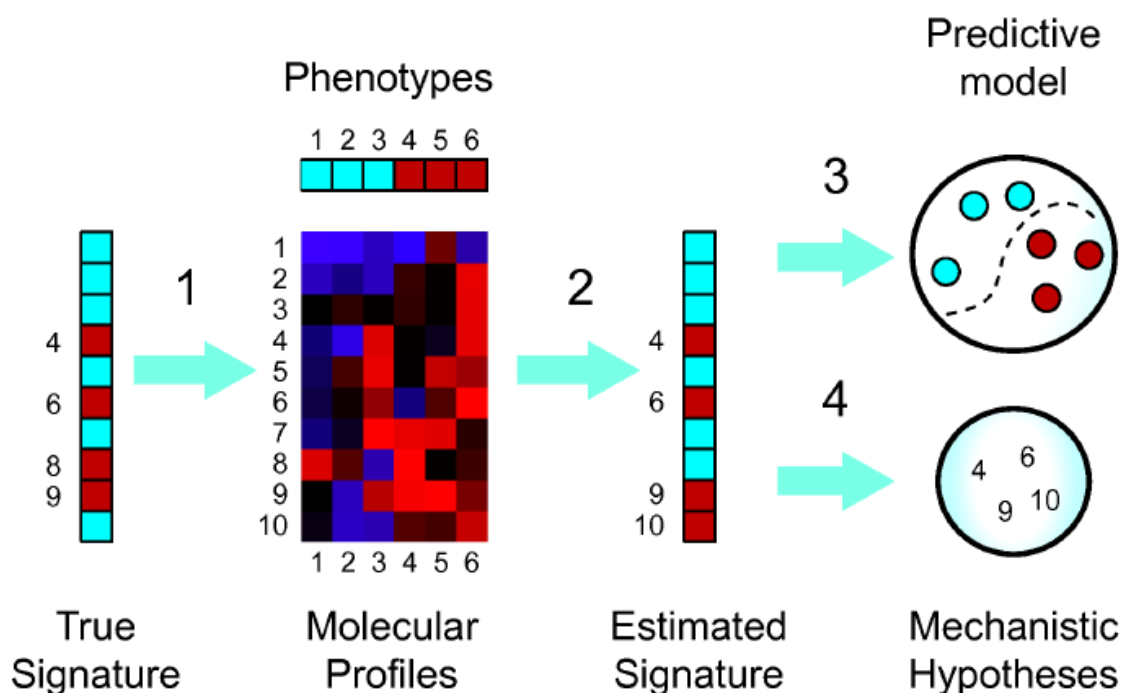
μας, κάτι το οποίο επιτρέπει στους ερευνητές να παρακολουθούν την έκφραση δεκάδων χιλιάδων γονιδίων ταυτόχρονα. Η ανάλυση δεδομένων μικροσυστοιχιών έχει επίσης γίνει μια συνήθης πρακτική σε πολλά πειραματικά εργαστήρια. Ένας μεγάλος αριθμός βιβλιογραφίας περιγράφει τις καινοτόμες ιδέες στο πλαίσιο της ανάλυσης δεδομένων από μικροσυστοιχίες. Έχει εφαρμοστεί ευρέως σε πολλούς τομείς της ιατρικής, συμπεριλαμβανομένου την διάκριση υποτύπων μιας νόσου, τον εντοπισμό υποψήφιων βιοδεικτών και την αποκάλυψη του βασικού βιομοριακού μηχανισμού μιας νόσου ή ανταπόκρισης σε φάρμακο.

Οι μικροσυστοιχίες μπορούν να παρέχουν ένα στιγμιότυπο όλης της μεταγραφικής δραστηριότητας σε ένα βιολογικό δείγμα, ενώ παράγουν επίσης μια τεράστια ποσότητα δεδομένων με εγγενή θόρυβο (θόρυβος δειγμάτων ή οργάνων), κάτι το οποίο εξακολουθεί να είναι ένα δύσκολο έργο για ερμηνεία ακόμα και με την αξιοποίηση μοντέρνων υπολογιστικών και στατιστικών εργαλείων. Σε πολλές περιπτώσεις, σημαντικά γονίδια παρουσιάζουν μικρές αλλαγές και πολλά γονίδια που επιλέγονται δεν είναι καλά σημειωμένα. Για να μεταφραστεί μια τέτοια λίστα διαφορεικά εκφρασμένων γονιδίων σε ένα λειτουργικό προφίλ, ερευνητές παρουσίασαν πολλές προσεγγίσεις για καλύτερη κατανόηση των βασικών βιολογικών φαινομένων. Ένας τρόπος για να βοηθήσει μια τέτοια ερμηνεία είναι η αναζήτηση αλλαγών σε μια ομάδα γονιδίων, όσον αφορά στις μεταξύ τους αλληλεπιδράσεις, ανάμεσα σε διαφορετικούς πληθυσμούς (π.χ. σε ασθενείς με καρκίνο του μαστού και υγιή άτομα) μέσα από τη δημιουργία ενός δικτύου γονιδιακών/πρωτεϊνικών αλληλεπιδράσεων [29].

2.1.8 Μοριακή Υπογραφή

Μοριακές υπογραφές είναι σύνολα από γονίδια, πρωτεΐνες, γενετικές παραλλαγές ή άλλες μεταβλητές οι οποίες μπορούν να χρησιμοποιηθούν ως δείκτες για ένα φαινότυπο ενός συγκεκριμένου γονιδίου ή ενός ιστού, όπως είναι μία καρκινική ή διαβητική κατάσταση. Αξιόπιστες μέθοδοι για την ανακάλυψη υπογραφών θα μπορούσαν να αποφέρουν πολύτιμες πληροφορίες για την κυτταρική βιολογία και τους μηχανισμούς των ασθενειών του ανθρώπου. Ο στόχος των υπογραφών είναι διττός: μπορούν να είναι χρήσιμες για τη διάγνωση ασθενειών ή για την αξιολόγηση του κινδύνου (πρόγνωση), αλλά επίσης μπορούν να προσδιορίσουν την ύπαρξη μορίων, που δεν είναι γνωστά, σε μία υποκείμενη μοριακή παθολογία (ανακάλυψη μοριακών μηχανισμών), όπως φαίνεται στην Εικόνα 2.7. Η ανακάλυψη μιας υπογραφής είναι σημαντική δεδομένου ότι μπορεί να ευθύνεται για πολλαπλές επιπτώσεις και αποτελείται μόνο από μεταβλητές που είναι πιο άμεσα συσχετισμένες με τον δοθέν φαινότυπο. Η χρήση των υπογραφών είναι ιδιαίτερα δημοφιλής για την διάγνωση του καρκίνου που βασίζεται στο προφίλ της γονιδιακής έκφρασης, και μερικές μελέτες έχουν προτείνει υπογραφές για συγκεκριμένους τύπους καρκίνου [30, 31, 32, 33, 34]. Ένα εξαιρετικό παράδειγμα είναι η υπογραφή για τον καρκίνο του μαστού που ανακαλύφθηκε από την van't Veer, το οποίο έχει επικυρωθεί και κλινικά [35].

Δεδομένου ότι δεν υπάρχει κάποιος "χρυσός" κανόνας για την ανακάλυψη υπογρα-



Εικόνα 2.7 – Ανακάλυψη υπογραφής. Μοριακές υπογραφές (1) είναι δείκτες για ένα συγκεκριμένο φαινότυπο κυττάρου ή ιστού. Οι υπογραφές ανακαλύπτονται από ένα δοθέν σύνολο μοριακών προφίλ (προφίλ γονιδιακών εκφράσεων) μαζί με ετικέτες φαινοτύπων (2). Οι υπογραφές έχουν διπλή χρήση, για πρόγνωση (3) και για ανακάλυψη μοριακών μηχανισμών (4).

φών, η επικύρωση των ανακαλυφθέντων υπογραφών συχνά ισοδυναμεί με την αναζήτηση στην επιστημονική βιβλιογραφία για τεκμηριωμένες συνδέσεις ανάμεσα στον φαινότυπο που έχει μελετηθεί και στα στοιχεία (γονίδια) της πιθανολογούμενης υπογραφής. Δεν είναι καθόλου σαφές αν ένα γονίδιο θα έπρεπε να συμπεριλαμβάνεται σε μια προβλεπόμενη υπογραφή απλά επειδή είναι κάπως "συσχετισμένη" με τον φαινότυπο [36].

2.1.9 Πρωτεϊνικές αλληλεπιδράσεις

Ο ολοκληρωμένος χάρτης των πρωτεϊνικών αλληλεπιδράσεων που μπορεί να λαμβάνουν χώρα σε έναν ζωντανό οργανισμό ονομάζεται *interactome* [37].

Ορισμός: [Πρωτεϊνικές αλληλεπιδράσεις] Πρωτεϊνικές αλληλεπιδράσεις είναι οι φυσικές επαφές μοριακής σύνδεσης ανάμεσα σε πρωτεΐνες που λαμβάνουν χώρα σε ένα κύτταρο ή σε ένα ζωντανό οργανισμό (*in vivo*). Γνωστές ως PPI (Protein-Protein Interaction).

Το ότι μία PPI συνεπάγεται φυσική επαφή ανάμεσα σε πρωτεΐνες δεν σημαίνει ότι τέτοιες επαφές είναι στατικές ή μόνιμες. Ένα άλλο απαραίτητο στοιχείο για τον προσδιορισμό της PPI είναι το βιολογικό πλαίσιο. Δεν συμβαίνουν όλες οι πιθανές αλληλεπιδράσεις σε κάθε κύτταρο κάθε στιγμή. Αντιθέτως, οι αλληλεπιδράσεις εξαρτώνται από το είδος του κυττάρου, την φάση και την κατάσταση του κυτταρικού κύκλου, το στάδιο ανάπτυξης, τις περιβαλλοντικές συνθήκες, την τροποποίηση των

πρωτεϊνών (φωσφορυλίωση), την παρουσία συμπαραγόντων και την παρουσία άλλων προσδεσμένων μορίων (binding partners) [38].

2.1.10 Μοριακές αλληλεπιδράσεις

Με την ολοκλήρωση του προγράμματος αποκωδικοποίησης του ανθρώπινου γονιδιώματος, τα βλέμματα στράφηκαν από την cataloging the "parts list" των γονιδίων και πρωτεϊνών στην χαρτογράφηση των δικτύων των αλληλεπιδράσεων που συμβαίνουν σε αυτά. Η κατανόηση αυτού του δικτύου είναι σημαντική επειδή οι πρωτεΐνες δεν λειτουργούν μεμονωμένα, αλλά αλληλεπιδρούν μεταξύ τους, με DNA, RNA και μικρά μόρια για να σχηματίσουν μοριακές μηχανές. Αυτές οι μηχανές απαρτίζονται από πολλά σύνολα, περιλαμβάνουν τόσο στατικά όσο και δυναμικά σύνολα μακρομορίων, και μεταδίδουν και λαμβάνουν το ίδιο ενδοκυτταρικά και εξωκυτταρικά σήματα.

Με την αύξηση της διαθεσιμότητας των δεδομένων για τις πρωτεϊνικές αλληλεπιδράσεις του ανθρώπου, το επίκεντρο της βιοπληροφορικής στράφηκε από την κατανόηση δικτύων κωδικοποιημένα από διάφορα είδη μοντέλων στην κατανόηση δικτύων που διέπουν ανθρώπινες ασθένειες. Το επίκεντρο της μελέτης σε τέτοια δίκτυα μπορεί να χωριστεί σε τέσσερις κατηγορίες:

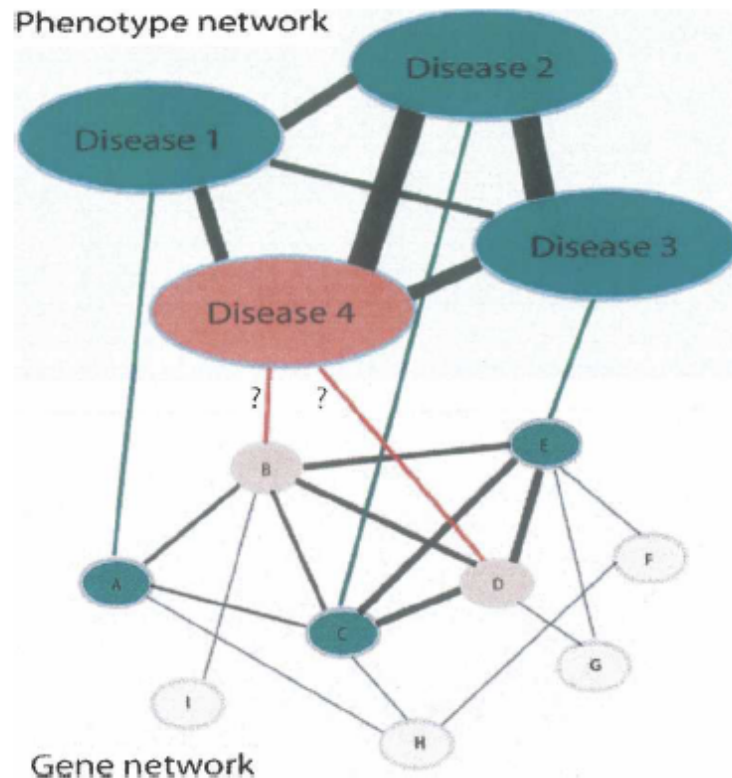
- Μελέτη ιδιοτήτων των δικτύων από γονίδια ανθρώπινων ασθενειών.
- Χρήση πρωτεϊνικών δικτύων για να εμπλακούν επιπλέον γονίδια σε μία ασθένεια.
- Προσδιορισμός των υποδικτύων που σχετίζονται με μία ασθένεια.
- Ταξινόμηση με βάση τα δίκτυα των μελετών ασθενών-μαρτύρων (case-control).

Ανάλυση δικτύων στον άνθρωπο:Ιδιότητες των παθολογικών γονιδίων

Τα παθολογικά γονίδια που συσχετίζονται με ένα μοναδικό φαινότυπο ή λειτουργία, συμπεριλαμβανομένου την εξέλιξη της νόσου, δεν είναι τυχαία τοποθετημένα στο δίκτυο. Τείνουν να έχουν υψηλή συνδεσιμότητα, να δημιουργούν συμπλέγματα και να εμφανίζονται σε κεντρικά σημεία του δικτύου. Στους μύκητες[40, 41], οι τιμές των ιδιοτήτων των δικτύων τους, όπως ο συνολικός βαθμός η μέση απόσταση τους, τείνουν να βρίσκονται ανάμεσα σε αυτές των σημαντικών και μη γονιδίων. Στους ανθρώπους, μερικές ομάδες έχουν αυξημένο βαθμό για παθολογικά γονίδια [42, 43], ενώ άλλα όχι [44].

Πρόβλεψη γονιδίων που προκαλούν ασθένειες

Ένας ακόμα λόγος για τον οποίο τα δίκτυα των βιομοριακών αλληλεπιδράσεων έχουν γίνει αιτία πολλών μελετών είναι η πρόβλεψη νέων γονιδίων που προκαλούν κάποια ασθένεια. Η υπόθεση-κλειδί σε αυτές τις μελέτες είναι ότι ένα γονίδιο που γειτνιάζει στο δίκτυο με ένα γονίδιο που προκαλεί κάποια νόσο είναι πιθανό να προκαλεί και αυτό την ίδια ή μία παρόμοια ασθένεια [44, 45]. Αυτό φαίνεται στην Εικόνα 2.8



Εικόνα 2.8 – Ένα δίκτυο γονιδίων-φαινοτύπου

Η ιδέα ότι οι πρωτεΐνες που είναι σε κοντινή απόσταση σε ένα δίκτυο προκαλούν παρόμοιες ασθένειες γίνεται όλο και πιο σημαντικός παράγοντας στο κυνήγι των παθογόνων γονιδίων. Διαφορετικές προσεγγίσεις για την αντιμετώπιση του προβλήματος της πρόβλεψης χρησιμοποιούν διαφορετικά είδη ολοκληρωμένων δεδομένων. Όλα περιλαμβάνουν ένα σύνολο από υποψήφια γονίδια που υπερθέτονται-παραθέτονται παραπλεύρως ενός συνόλου γνωστών παθολογικών γονιδίων σε ένα φυσικό ή λειτουργικό δίκτυο. "De-novo" προσεγγίσεις οι οποίες δεν εξαρτώνται από προηγούμενη γνώση παθολογικών γονιδίων δεν έχουν ακόμη αναπτυχθεί.

Προσδιορισμός των υποδικτύων που σχετίζονται με μία ασθένεια

Εκτός από την πρόβλεψη συγκεκριμένων παθολογικών πρωτεϊνών, ένα βιομοριακό δίκτυο μπορεί επίσης να χρησιμοποιηθεί για να προβλέψει υποδίκτυα που σχετίζονται με κάποια νόσο. Τα υποδίκτυα είναι σημαντικά επειδή, σε αντίθεση με τις επιμέρους πρωτεΐνες, παρέχουν συγκεκριμένες υποθέσεις όπως με τα μοριακά σύμπλοκα, τα μονοπάτια σηματοδότησης και άλλους μηχανισμούς που επηρεάζουν το αποτέλεσμα της νόσου.

Η συσχέτιση των παθολογικών γονιδίων με φυσικά ή λειτουργικά δίκτυα μπορεί να οδηγήσει στον προσδιορισμό και στην ταυτοποίηση επιπλέον γονιδίων που σχετίζονται με κάποια νόσο και να δημιουργήσει υποδίκτυα που προσφέρουν μηχανιστικές υποθέσεις για τα αίτια της νόσου. Οι αλληλεπιδράσεις σε τέτοια υποδίκτυα συχνά υποδηλώνουν λειτουργικούς καταρράκτες σηματοδότησης, μεταβολικά μονοπάτια, ή μορια-

κά σύμπλοκα τα οποία είναι είτε η αιτία είτε το αποτέλεσμα του παθολογικού φαινοτύπου. Επίσης, βοηθούν στην εξήγηση της επίδρασης που έχουν πολλοί γενετικοί και περιβαλλοντικοί παράγοντες σε μία νόσο.

Ταξινόμηση με βάση τα δίκτυα των case-control μελετών

Μια τελευταία εφαρμογή της ανάλυσης των μοριακών δικτύων είναι η χρήση τους για την βελτίωση της ταξινόμησης των νοσημάτων. Η ταξινόμηση είναι μια σημαντική τεχνική για τον προσδιορισμό βιοδεικτών που μπορεί να ξεχωρίσει ασθενείς "cases" από μάρτυρες "controls". Ιδιαίτερα στην έρευνα ασθενειών, ασθενείς εναντίων μαρτύρων χρησιμοποιούνται για να διαχωρίσουν άτομα που έχουν μια νόσο με άτομα που δεν την έχουν, για να προβλέψουν άτομα που ενδέχεται να έχουν σοβαρές συνέπειες της νόσου σε σχέση με εκείνα που οι συνέπειες είναι πιο ομαλές, ή για να γίνει διάκριση μεταξύ των διαφόρων ασθενειών που θα μπορούσαν αλλιώς να φαίνονται παρόμοιες [46].

Τα πρωτεϊνικά δίκτυα είναι μια ισχυρή πηγή πληροφοριών για την ταξινόμηση των ασθενειών. Χαρακτηριστικά, ένας υπερθέτει δεδομένα γονιδιακής έκφρασης στο δίκτυο για να προσδιοριστούν συνδέσεις, ή πιο σύνθετες δομές υποδικτύων, των οποίων η συνολική έκφραση διαφέρει μεταξύ των καταστάσεων της νόσου. Αυτά τα επιλεκτικά υποδίκτυα είναι πιο αναπαραγώγιμα από ότι τα μεμονωμένα γονίδια και μπορούν να αυξήσουν την ακρίβεια της πρόβλεψης.

2.2 Μαθηματικό Υπόβαθρο

2.2.1 Στοιχεία θεωρίας πιθανοτήτων

Παρακάτω φαίνονται μερικοί βασικοί ορισμοί από την θεωρία πιθανοτήτων που θα χρησιμοποιηθούν ευρέως στην εκπόνηση της εργασίας.

Οι δυο πιο θεμελιώδεις κανόνες στην θεωρία πιθανοτήτων είναι οι ακόλουθοι:

Ορισμός: Κανόνας αθροίσματος

$$p(X) = \sum_Y p(X, Y) \quad (2.1)$$

Ορισμός: Κανόνας γινομένου

$$p(X, Y) = p(Y|X)p(X) \quad (2.2)$$

όπου $p(X, Y)$ είναι η από κοινού πιθανότητα των X και Y . Η ποσότητα $p(Y|X)$ είναι η υπό συνθήκη πιθανότητα, την οποία θα διαβάζουμε ως την πιθανότητα του Y δοθέντος X , ενώ η ποσότητα $p(X)$ είναι η περιθώρια πιθανότητα ή απλά η πιθανότητα του X .

Το θεώρημα του Bayes (Bayes'theorem) είναι συνέπεια των προαναφερθέντων

κανόνων και είναι πολύ σημαντικός στην κατασκευή *Bayesian δικτύων* (Bayesian Networks):

Θεώρημα: [Το θεώρημα του Bayes] Το θεώρημα του Bayes σχετίζει την τρέχουσα πιθανότητα με την αρχική πιθανότητα. Είναι σημαντικό στο μαθηματικό χειρισμό της υπό συνθήκη πιθανότητας.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad (2.3)$$

Με την εφαρμογή του κανόνα του αθροίσματος, ο παρανομαστής στο θεώρημα του Bayes μπορεί να εκφραστεί με τους όρους των ποσοτήτων που εμφανίζονται στον αριθμητή:

$$p(X) = \sum_Y p(X|Y)p(Y) \quad (2.4)$$

οπότε ο παρανομαστής λειτουργεί ως η σταθερά κανονικοποίησης που εξασφαλίζει ότι το άθροισμα της υπό συνθήκη πιθανότητας του αριστερού μέλους $p(Y|X)$ πάνω σε όλες τις τιμές του Y ισούται με τη μονάδα.



Εικόνα 2.9 – Thomas Bayes

2.2.2 Θεωρία Γράφων

Παρακάτω παραθέτουμε μερικές βασικές έννοιες για την θεωρία των γράφων:

Ορισμός: [Γράφος] Ένας Γράφος G αποτελείται από κόμβους (ή κορυφές) και ακμές (ή σύνδεσμοι) μεταξύ των κόμβων. Οι ακμές μπορεί να είναι κατευθυνόμενες (έχουν ένα τόξο σε μία μόνο κατεύθυνση) ή μη κατευθυνόμενες. Ένας γράφος με όλες τις ακμές του κατευθυνόμενες ονομάζεται κατευθυνόμενος γράφος, ενώ ένας με όλες τις ακμές μη κατευθυνόμενες ονομάζεται μη κατευθυνόμενος γράφος.

Ορισμός: [Κατευθυνόμενος άκυκλος γράφος] Ένας Κατευθυνόμενος Άκυκλος Γράφος ή ΚΑΓ είναι ένας γράφος G με κατευθυνόμενες ακμές (με τόξα σε κάθε σύνδεσμο) μεταξύ των κόμβων, ο οποίος δεν παρουσιάζει κατευθυνόμενους κύκλους, δηλαδή κλειστά μονοπάτια μέσα στο γράφο, ώστε όταν ξεκινάμε από ένα κόμβο και ακολουθώντας την κατεύθυνση των συνδέσμων, να μην καταλήγουμε στον αρχικό κόμβο. Ισοδύναμα, θα λέμε ότι στους ΚΑΓ υπάρχει μία αρίθμηση κόμβων τέτοια ώστε

δεν υπάρχουν σύνδεσμοι που πηγαίνουν από οποιονδήποτε κόμβο σε οποιονδήποτε μικρότερα αριθμημένο κόμβο.

Ορισμός: [Απόγονος κόμβος] Ένας κόμβος y θα λέμε ότι είναι απόγονος ενός κόμβου αν υπάρχει μονοπάτι από τον κόμβο στον κόμβο y ακολουθώντας τις κατευθύνσεις όλων των τόξων.

Ορισμός: [Κλίκα] Πρόκειται για μία έννοια πάνω σε γραφικό μοντέλο, που ορίζεται ως το υποσύνολο των κόμβων του γράφου τέτοιο ώστε να υπάρχει σύνδεσμος μεταξύ όλων των ζευγαριών κόμβων που ανήκουν σε αυτό το υποσύνολο. Με άλλα λόγια, το σύνολο των κόμβων σε μία κλίκα είναι πλήρως συνδεδεμένο. Τα μέλη μιας κλίκας είναι γείτονες.

2.3 Δίκτυα

Η δικτυακή δομή μπορεί να παρατηρηθεί σε όλες τις εκφάνσεις της ζωής. Για παράδειγμα, ο εγκέφαλος είναι ένα δίκτυο από νευρικά κύτταρα συνδεδεμένα με άξονες, και τα κύτταρα από μόνα τους είναι ένα δίκτυο μορίων συνδεδεμένα με βιοχημικές αντιδράσεις. Οι κοινωνίες, επίσης, είναι δίκτυα ανθρώπων συνδεδεμένοι με φιλίες, οικογενειακές σχέσεις και επαγγελματικούς δεσμούς. Σε μεγαλύτερη κλίμακα, διατροφικές αλυσίδες και οικοσυστήματα μπορούν να παρουσιαστούν σαν δίκτυα ειδών. Τα δίκτυα εισχωρούν ακόμα και στην τεχνολογία: το Διαδίκτυο, τα δυναμικά δίκτυα και τα συστήματα μεταφοράς δεν είναι παρά μόνο μερικά παραδείγματα. Ακόμα και η γλώσσα που χρησιμοποιούμε για να μεταφέρουμε τις σκέψεις μας στους άλλους, είναι ένα δίκτυο, φτιαγμένο από λέξεις συνδεδεμένες με συντακτικές σχέσεις. Παρά τη σημαντικότητα και τη διεξοδικότητα των δικτύων, οι επιστήμονες γνωρίζουν λίγα πράγματα για τη δομή και τα χαρακτηριστικά τους.



Εικόνα 2.10 – Παράδειγμα Δικτύου.

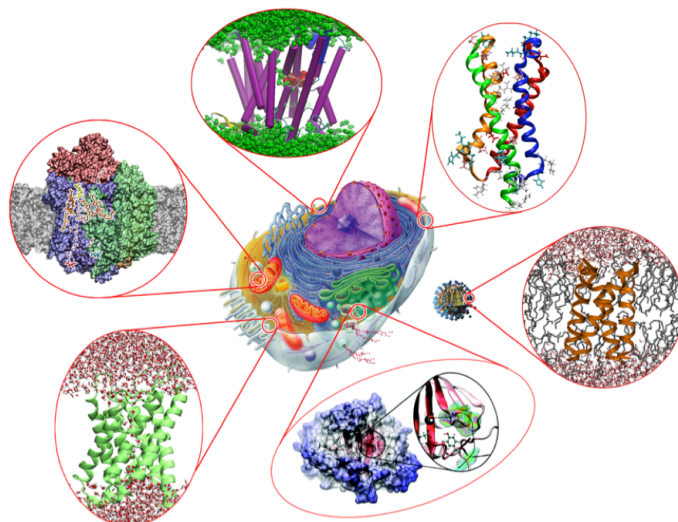
2.3.1 Βιολογικά Δίκτυα

Για την αναπαράσταση των ενδοκυτταρικών βιολογικών δικτύων τυπικά θεωρούνται κόμβοι τα μοριακά στοιχεία μέσα σε ένα κύτταρο και ακμές οι άμεσες ή έμμεσες

αλληλεπιδράσεις τους. Διαφορετικά είδη ενδοκυτταρικών μοριακών βιολογικών δικτύων μπορούν να αναπαρασταθούν από διαφορετικά είδη μαθηματικών δομών (*γράφοι*) [47].

Τα δίκτυα βιολογικών συστημάτων εμφανίζονται σε πολλές μορφές, από τροφικά πλέγματα στην οικολογία μέχρι διάφορα βιοχημικά δίκτυα στην μοριακή βιολογία. Ειδικότερα, το ευρύ φάσμα των αλληλεπιδράσεων μεταξύ γονιδίων, πρωτεϊνών και μεταβολιτών σε ένα κύτταρο αναπαρίστανται καλύτερα με πολύπλοκα δίκτυα. Κατά τη διάρκεια των τελευταίων δεκαετιών, η γονιδιωματική παρήγαγε μια τεράστια ποσότητα δεδομένων μοριακών αλληλεπιδράσεων, συμβάλλοντας έτσι σε χάρτες ειδικών κυτταρικών δικτύων. Η μελέτη του μεταγραφικού προτύπου (transcriptomics) και της πρωτεομικής (proteomics) έχουν την δυνατότητα να χρησιμοποιούν τις ήδη εκτεταμένες πηγές δεδομένων που παρέχονται από την ευρεία ανάλυση την γονιδιακής έκφρασης σε επίπεδο mRNA και πρωτεΐνης.

Τα δίκτυα μας προσφέρουν ένα νέο τρόπο να κατηγοριοποιούμε συστήματα διαφορετικής προέλευσης ενός ενιαίου πλαισίου. Αυτή η προσέγγιση έχει αποκαλύψει απρόσμενες ομοιότητες ανάμεσα στη οργάνωση διάφορων σύνθετων συστημάτων, κάτι το οποίο δείχνει ότι τα δίκτυα τα οποία τα περιγράφουν, διέπονται από αρχές και μηχανισμούς γενικής οργάνωσης. Η κατανόηση των κινητήριων δυνάμεων που περιβάλλουν διαφορετικά δίκτυα με παρόμοια τοπολογικά χαρακτηριστικά, επιτρέπει στη συστημική βιολογία να συνδυάζει τις πολυάριθμες λεπτομέρειες για τις μοριακές αλληλεπιδράσεις σε ένα ενιαίο πλαίσιο, προσφέροντας τα μέσα για την αντιμετώπιση της δομής του κυττάρου ως σύνολο [49].



Εικόνα 2.11 – Biomolecular Systems

2.4 Ιδιότητες Δικτύων

Πιθανόν το πιο απλό χρήσιμο μοντέλου δικτύου είναι ο τυχαίος γράφος. Σε αυτό το μοντέλο, μη κατευθυνόμενες ακμές τοποθετούνται τυχαία ανάμεσα σε έναν σταθερό αριθμό n κόμβων για να δημιουργηθεί ένα δίκτυο στο οποίο κάθε μια από τις $\frac{1}{2}n(n-1)$ πιθανές ακμές είναι ανεξάρτητα υπαρκτή με πιθανότητα p , και ο αριθμός των ακμών που είναι συνδεδεμένες σε κάθε κορυφή (ο βαθμός μίας κορυφής) είναι κατανομημένος σύμφωνα με μια διωνυμική κατανομή, ή μια Poisson κατανομή στο όριο μεγάλου n . Οι τυχαίοι γράφοι έχουν μελετηθεί από μαθηματικούς, και αρκετά αποτελέσματα, και κατά προσέγγιση και ακριβώς, έχουν αποδειχθεί αυστηρά. Πολλά από τα ενδιαφέροντα χαρακτηριστικά των πραγματικών δικτύων που έχουν τραβήξει το ενδιαφέρον των ερευνητών τα τελευταία χρόνια, όμως, αφορούν τις μεθόδους κατά τις οποίες δίκτυα δεν είναι τυχαίοι γράφοι. Τα πραγματικά δίκτυα είναι μη-τυχαία σε μερικές μεθόδους που προτείνουν και πιθανό μηχανισμό που θα μπορούσε να οδηγήσει σε σχηματισμό δικτύου, και πιθανές μεθόδους στις οποίες θα μπορούσαμε να αξιοποιήσουμε τη δομή του δικτύου για να επιτευχθούν συγκεκριμένοι στόχοι. Παρακάτω θα περιγράψουμε μερικά χαρακτηριστικά που φαίνεται να είναι κοινά σε διαφορετικού τύπου δίκτυα [52].

2.4.1 Ορισμοί βασικών ιδιοτήτων των δικτύων

Παρακάτω παραθέτουμε μερικούς βασικούς ορισμούς των δικτύων που χρησιμοποιήθηκαν στην ανάλυση των δικτύων μας. Θεωρούμε n τον αριθμό των κόμβων και e τον αριθμό των ακμών του δικτύου.

Ορισμός: [Πυκνότητα] Η πυκνότητα D ενός δικτύου ορίζεται ως ο λόγος του αριθμού των ακμών e προς τον αριθμό των πιθανών ακμών. Ο αριθμός των πιθανών ακμών δίνεται από τον διωνυμικό συντελεστή $\binom{n}{2}$, οπότε

$$D = \frac{2e}{n(n-1)}$$

Ορισμός: [Μέγεθος] Το μέγεθος ενός δικτύου ορίζεται ως ο αριθμός των κόμβων n ή διαφορετικά ως ο αριθμός των ακμών e ο οποίος μπορεί να κυμαίνεται από $n-1$ (σε ένα δέντρο) μέχρι e_{max} (σε ένα πλήρη γράφο).

Ορισμός: [Μέσος Βαθμός] Ο Βαθμός k ενός κόμβου είναι ο αριθμός των ακμών που συνδέονται σε αυτόν. Πολύ κοντά στην έννοια της πυκνότητας είναι και αυτή του μέσου βαθμού, $\langle k \rangle = \frac{2e}{n}$.

Ορισμός: [Μέσο μήκος μονοπατιού] Το μέσο μήκος μονοπατιού υπολογίζεται βρίσκοντας το κοντύτερο μονοπάτι ανάμεσα σε όλα τα ζευγάρια κόμβων, αθροίζοντας τα, και μετά διαιρώντας με τον συνολικό αριθμό των ζευγαριών. Αυτό μας δείχνει κατά μέσο όρο τον αριθμό των βημάτων που χρειάζεται για να πας από τον ένα κόμβο στον άλλο.

Ορισμός: [Διάμετρος] Η διάμετρος ενός δικτύου ορίζεται ως το μακρύτερο από τα υπολογισμένα κοντύτερα μονοπάτια του. Η διάμετρος είναι αντιπροσωπευτική του γραμμικού μεγέθους ενός δικτύου.

Ορισμός: [Συνδετικότητα] Η συνδετικότητα ενός γράφου αναφέρεται στον ελάχιστο αριθμό στοιχείων (κόμβοι ή ακμές) που απαιτείται να αφαιρεθούν για να αποσυνδεθούν οι υπολειπόμενοι κόμβοι ο ένας από τον άλλο. Η συνδετικότητα είναι μια σημαντική μετρική των γράφων γιατί μετράει την δύναμη του ως δίκτυο.

Ένας γράφος είναι *συνδεδεμένος* εάν υπάρχει ένα μονοπάτι ανάμεσα σε κάθε ζεύγος κόμβων. Ένας γράφος που δεν είναι συνδεδεμένος είναι *αποσυνδεδεμένος*. Ένας γράφος με ένα ή κανένα κόμβο είναι συνδεδεμένος ενώ ένας γράφος χωρίς ακμές με δύο ή περισσότερους κόμβους είναι αποσυνδεδεμένος.

2.4.2 Το φαινόμενο του Μικρού Κόσμου

Τη δεκαετία του 1960 ο Stanley Milgram πραγματοποίησε ένα πείραμα, κατά το οποίο γράμματα που περνούσαν από άτομο σε άτομο μπορούσαν να φτάσουν σε ένα συγκεκριμένο στόχο-άτομο με ένα μικρό αριθμό βημάτων. Αυτή είναι μία από τις πρώτες επιδείξεις του φαινομένου του Μικρού Κόσμου (Small World), το γεγονός δηλαδή ότι τα περισσότερα ζεύγη κορυφών στα περισσότερα δίκτυα φαίνεται να είναι συνδεδεμένα με ένα μικρό μονοπάτι δια μέσου του δικτύου. Στις μέρες μας, το φαινόμενο Μικρού Κόσμου έχει μελετηθεί και επιβεβαιωθεί σε μεγάλο αριθμό διαφορετικών δικτύων.

Θεωρούμε ένα μη-κατευθυνόμενο δίκτυο, ορίζουμε l τη μέση γεωδαισιακή (δηλαδή μικρότερη) απόσταση μεταξύ ζεύγη κορυφών σε ένα δίκτυο:

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \quad (2.5)$$

όπου d_{ij} είναι η γεωδαισιακή απόσταση από την κορυφή i στην κορυφή j .

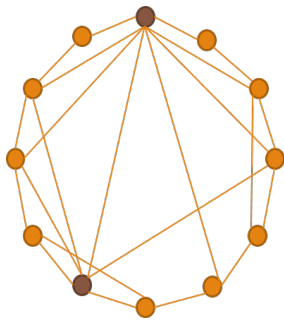
Η ποσότητα l μπορεί να μετρηθεί για ένα δίκτυο με n κορυφές και e ακμές σε χρόνο $O(en)$ με τη χρήση μίας απλής breadth-first αναζήτησης (burning algorithm) [53].

Ο ορισμός 2.5 του l είναι προβληματικός για δίκτυα που έχουν περισσότερα από ένα στοιχεία. Σε τέτοιες περιπτώσεις, υπάρχουν ζεύγη κορυφών που δεν έχουν σύνδεση μεταξύ τους. Συνήθως ορίζεται άπειρη γεωδαισιακή απόσταση σε τέτοια ζευγάρια, αλλά έτσι η τιμή του l επίσης γίνεται άπειρη. Για να αποφευχθεί αυτό το πρόβλημα συνήθως ορίζεται το l σε τέτοια δίκτυα να είναι η μέση γεωδαισιακή απόσταση ανάμεσα σε όλα τα ζεύγη που έχουν σύνδεση μεταξύ τους. Ζεύγη που υπάγονται σε δυο διαφορετικά στοιχεία εξαιρούνται από το μέσο όρο.

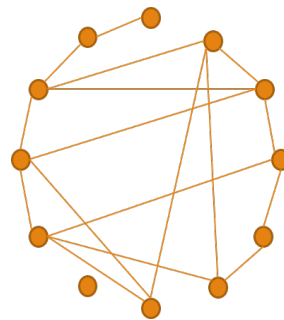
Το φαινόμενο Small-World έχει προφανείς επιπτώσεις για τη δυναμική των διαδικασιών που παίρνουν μέρος στα δίκτυα. Για παράδειγμα, αν θεωρήσουμε την διάδοση των πληροφοριών, ή οτιδήποτε άλλο, τότε το φαινόμενο Small-World υποδηλώνει ότι αυτή η διάδοση θα είναι γρήγορη στα περισσότερα δίκτυα του πραγματικού κόσμου. Αν χρειάζεται μόνο έξι βήματα για μια φήμη να διαδοθεί από ένα άτομο σε ένα άλλο, για

παράδειγμα, τότε η φήμη θα διαδοθεί πολύ πιο γρήγορα από το να έκανε εκατοντάδες βήματα.

Από την άλλη μεριά, το φαινόμενο Small-World έχει και μαθηματική βάση. Αν ο αριθμός των κορυφών μέσα σε ένα διάστημα r μιας κεντρικής κορυφής μεγαλώνει εκθετικά με το r (αυτό ισχύει για πολλά δίκτυα συμπεριλαμβανομένου και του τυχαίου γράφου) τότε η τιμή του l θα αυξάνεται όπως η $\log n$. Τα τελευταία χρόνια ο όρος "Small-World" έχει λάβει μια πιο συγκεκριμένη έννοια: δίκτυα λέγεται ότι εμφανίζουν το φαινόμενο Small-World αν η τιμή του l κλιμακώνεται λογαριθμικά ή πιο αργά με το μέγεθος του δικτύου για σταθερό μέσο βαθμό. Η λογαριθμική κλιμάκωση μπορεί να αποδειχθεί για διάφορα μοντέλα δικτύων [54, 55, 56, 57, 58] και έχει επίσης παρατηρηθεί σε διάφορα δίκτυα πραγματικού κόσμου [59, 60, 61] [52].



(2.12.1) Παράδειγμα Small-world δικτύου. Τα hubs είναι πιο σκούρα από τους άλλους κόμβους.
Μέσος βαθμός κόμβου = 1.917
Μέσο μήκος μονοπατιού = 1.803
Συντελεστής ομαδοποίησης = 0.522



(2.12.2) Τυχαίος γράφος.
Μέσος βαθμός κόμβου = 1.417
Μέσο μήκος μονοπατιού = 2.109
Συντελεστής ομαδοποίησης = 0.167

Εικόνα 2.12

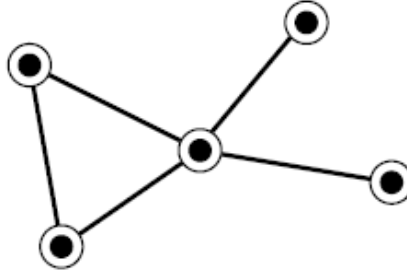
2.4.3 Ομαδοποίηση

Μια ξεκάθαρη απόκλιση από την συμπεριφορά των τυχαίων γράφων φαίνεται στην ιδιότητα των δικτύων *ομαδοποίησης*. Σε μερικά δίκτυα έχει βρεθεί ότι αν η κορυφή A είναι συνδεδεμένη στην κορυφή B και η κορυφή B στην κορυφή C , τότε υπάρχει μεγάλη πιθανότητα η κορυφή A να είναι επίσης συνδεδεμένη με την κορυφή C . Με όρους της τοπολογίας των δικτύων, ομαδοποίηση σημαίνει η παρουσία ενός αυξημένου αριθμού τριγώνων στο δίκτυο-τριάδες κορυφών όπου η κάθε μια είναι συνδεδεμένη με κάθε άλλη. Αυτό μπορεί να ποσοτικοποιηθεί ορίζοντας έναν συντελεστή ομαδοποίησης ως εξής:

$$C = \frac{3 * \text{number of triagles in the network}}{\text{number of connected triples of vertices}} \quad (2.6)$$

όπου μια "συνδεδεμένη τριάδα" σημαίνει μία μεμονωμένη κορυφή με ακμές σε μη διατεταγμένα ζευγάρια.

Ο C μετράει τον λόγο των τριάδων που έχουν την τρίτη τους ακμή συμπληρωμένη



Εικόνα 2.13 – Εικονογράφηση του ορισμού του συντελεστή ομαδοποίησης C (2.6). Αυτό το δίκτυο έχει ένα τρίγωνο και οκτώ συνδεδεμένες τριάδες, και συνεπώς έχει συντελεστή ομαδοποίησης $3 * \frac{1}{8} = \frac{3}{8}$. Η κάθε κορυφή έχει τοπικό συντελεστή ομαδοποίησης (2.8), 1, $\frac{1}{6}$, 0 και 0, και μέση τιμή (2.9) $C = \frac{13}{30}$

για να ολοκληρωθεί το τρίγωνο. Ο παράγοντας 3 στον αριθμητή αντιπροσωπεύει το γεγονός ότι κάθε τρίγωνο συνεισφέρει σε τρεις τριάδες και διασφαλίζει ότι ο C κυμαίνεται μεταξύ $0 \leq C \leq 1$. Πιο απλά, ο C είναι η μέση πιθανότητα δυο κορυφών, που η κάθε μία είναι γειτονική με μια κοινή κορυφή, να είναι και αυτές οι δυο γειτονικές. Αυτό μπορεί να γραφτεί διαφορετικά στη μορφή:

$$C = \frac{6 * \text{number of triangles in the network}}{\text{number of paths of length two}} \quad (2.7)$$

όπου ένα μονοπάτι μήκους δύο αναφέρεται σε ένα άμεσο μονοπάτι που ξεκινάει από μια καθορισμένη κορυφή.

Ένας διαφορετικός ορισμός του συντελεστή ομαδοποίησης δόθηκε από τους Watts και Strogatz [62], οι οποίοι πρότειναν τον ορισμό μιας τοπικής τιμής

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} \quad (2.8)$$

Για κορυφές με βαθμό 0 ή 1, για τις οποίες και ο αριθμητής και ο παρονομαστής είναι μηδέν, θέτουμε $C_i = 0$. Έτσι ο συντελεστής ομαδοποίησης για ολόκληρο το δίκτυο είναι ο μέσος όρος

$$C = \frac{1}{4} \sum_i C_i \quad (2.9)$$

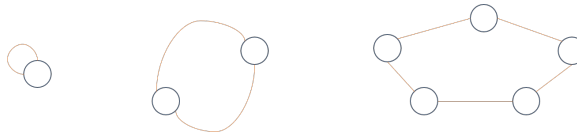
Ο τοπικός συντελεστής C_i παραπάνω έχει χρησιμοποιηθεί ευρέως στην λογοτεχνία, όπου αναφέρεται ως *πυκνότητα δικτύου* [63]. Η εξάρτηση του από τον βαθμό k_i από την κεντρική κορυφή i έχει μελετηθεί από τους Dorogontsev, Goltsev, Mendes [64] και Szabo, Alava, Kertesz [65]. Και οι δύο ομάδες βρήκαν ότι ο C_i ελαττώνεται κατά k_i περίπου όπως η k_i^{-1} για τα μοντέλα των scale-free δικτύων. Παρόμοια συμπεριφορά έχει παρατηρηθεί εμπειρικά σε δίκτυα του πραγματικού κόσμου.

Γενικά οι τιμές των συντελεστών ομαδοποίησης τείνουν να είναι σημαντικά υψηλότερες από αυτές ενός τυχαίου γράφου με ίδιο αριθμό κορυφών και ακμών [52].

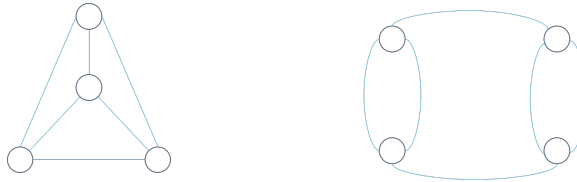
2.4.4 Κατανομή Βαθμού

Ο βαθμός μιας κορυφής σε ένα δίκτυο είναι ο αριθμός των ακμών (incident) στην εν λόγω κορυφή. Ορίζεται p_k το κλάσμα των κορυφών του δικτύου που έχουν βαθμό k . Ισοδύναμα, p_k είναι η πιθανότητα μια κορυφή που επιλέγεται τυχαία να έχει βαθμό k . Το γράφημα του p_k για οποιοδήποτε δίκτυο σχηματίζεται δημιουργώντας ένα ιστόγραμμα των βαθμών των κορυφών. Αυτό το ιστόγραμμα είναι η *Κατανομή Βαθμού* (Degree Distribution) για κάθε δίκτυο.

Σε ένα τυχαίο γράφο, ο οποίος μελετήθηκε από τον Erdos και τον Renyi [66, 67, 68], η κάθε κορυφή είναι παρούσα ή απύουσα με την ίδια πιθανότητα και ως εκ τούτου η κατανομή βαθμού είναι, όπως προαναφέρθηκε, διωνυμική ή Poisson στο όριο ενός μεγάλου μεγέθους γράφου. Τα πραγματικά δίκτυα είναι συνήθως διαφορετικά από τους τυχαίους γράφους ως προς τις κατανομές βαθμού. Εκτός της κατανομής Poisson, οι βαθμοί των κορυφών στα περισσότερα δίκτυα είναι ιδιαίτερα ασύμμετρα στην δεξιά μεριά, δηλαδή η κατανομή τους έχει μια μακριά δεξιά ουρά τιμών η οποία είναι πολύ πιο πάνω από τον μέσο όρο.



Εικόνα 2.14 – Γράφοι με βαθμό 2



Εικόνα 2.15 – Γράφοι με βαθμό 3

Ένας εναλλακτικός τρόπος παρουσίασης των δεδομένων των βαθμών είναι το γράφημα της αθροιστικής συνάρτησης κατανομής

$$P_k = \sum_{k'=k}^{\infty} p_{k'}, \quad (2.10)$$

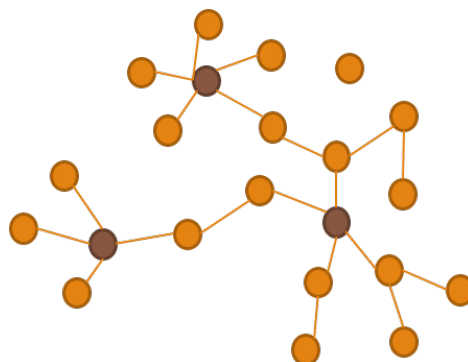
η οποία είναι η πιθανότητα ο βαθμός να είναι μεγαλύτερος ή ίσος του k . Το πλεονέκτημα αυτού του γραφήματος είναι ότι παρουσιάζονται όλα τα αρχικά δεδομένα. Σε ένα συμβατικό ιστόγραμμα με bins, οι όποιες διαφοροποιήσεις μεταξύ των τιμών των δεδομένων που εμπίπτουν στο ίδιο bin χάνονται. Στην αθροιστική συνάρτηση κατανομής δεν εμφανίζεται αυτό το πρόβλημα. Η αθροιστική συνάρτηση κατανομής επιπλέον μειώνει το θόρυβο στην ουρά. Το μειονέκτημα του γραφήματος είναι ότι δεν δίνει μια άμεση απεικόνιση της ίδιας της κατανομής βαθμού και τα γειτονικά σημεία του γραφήματος δεν είναι στατιστικά ανεξάρτητα, δυσκολεύοντας την σωστή προσαρμογή των δεδομένων.

Για κατευθυνόμενους γράφους κάθε κορυφή έχει και μέσα-βαθμό και έξω-βαθμό, και η κατανομή βαθμού επομένως γίνεται μια συνάρτηση p_{jk} δυο μεταβλητών, που αντιπροσωπεύουν το λόγο των κορυφών που έχουν ταυτόχρονα μέσα-βαθμό j και έξω βαθμό k . Σε εμπειρικές μελέτες για τους κατευθυνόμενους γράφους όπως το διαδίκτυο, ερευνητές συχνά μελετούν μεμονωμένη την κατανομή του μέσα και έξω βαθμού, δηλαδή την κατανομή που προέρχεται αθροίζοντας τα p_{jk} πάνω σε έναν από τους δείκτες της [69, 70, 71]. Αυτό, όμως, απορρίπτει πολλές από τις πληροφορίες που υπάρχουν στην από κοινού κατανομή. Έχει βρεθεί ότι οι μέσα και έξω βαθμοί είναι ισχυρά συσχετισμένοι σε μερικά δίκτυα, γεγονός που υποδηλώνει ότι υπάρχει κάτι περισσότερο που μπορεί να αντληθεί από την από κοινού κατανομή [52].

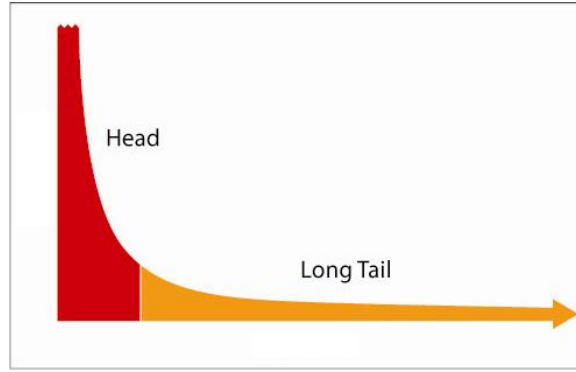
2.4.5 Δίκτυα Ελεύθερης Κλίμακας

Ένα δίκτυο ελεύθερης κλίμακας (scale-free network) είναι ένα είδος πολύπλοκου δικτύου που έχει προκαλέσει το ενδιαφέρον. Ο λόγος είναι ότι από τότε που ανακαλύφθηκε θεωρείται ότι πολλά δίκτυα πραγματικού κόσμου (real-world networks) ανήκουν σε αυτή την κατηγορία. Το δίκτυο αυτό ακολουθεί μια κατανομή με νόμο δύναμης (power law distribution) όσον αφορά τον αριθμό των συνδέσεων ανάμεσα στους κόμβους του. Λίγοι κόμβοι, οι οποίοι ονομάζονται κύριοι κόμβοι (hubs), παρουσιάζουν υψηλή συνδεσιμότητα ενώ η πλειοψηφία τους, που ονομάζονται δευτερεύοντες κόμβοι ή απλά κόμβοι (nodes), παρουσιάζουν χαμηλή συνδεσιμότητα.

Ένα δίκτυο με τοπολογία ελεύθερης κλίμακας, αποτελείται από κόμβους συνδεδεμένους μεταξύ τους (οι λεγόμενοι σύνδεσμοι ή links), σχηματίζοντας κύριους κόμβους, οι οποίοι βρίσκονται σε συγκεκριμένες θέσεις στο δίκτυο. Ένα δίκτυο ελεύθερης κλίμακας θα μπορούσε να χαρακτηριστεί με μία φράση ως "ιστός χωρίς αράχνη". Αυτό σημαίνει ότι κανένας από τους κόμβους που αποτελούν το δίκτυο δεν ευθύνεται από μόνος του για τη δομή, τον έλεγχο και την παρακολούθηση όλων των υπόλοιπων κόμβων και συνδέσμων. Με άλλα λόγια, ένας κόμβος έχει μικρή επίδραση στην ακεραιότητα του συστήματος [72].



Εικόνα 2.16 – Παράδειγμα δικτύου ελεύθερης κλίμακας. Τα hubs είναι πιο σκούρα χρωματισμένα.



Εικόνα 2.17 – Μία Power law κατανομή

2.4.6 Ενδιάμεση Κεντρικότητα

Η *κεντρικότητα* (centrality) αναφέρεται σε δείκτες οι οποίοι αναγνωρίζουν τους σημαντικότερους κόμβους σε ένα γράφο. Η *ενδιάμεση κεντρικότητα* (Betweenness centrality) ή αλλιώς κατανομή φόρτωσης (Load Distribution) είναι ένας από αυτούς τους δείκτες. Ισούται με τον αριθμό των κοντύτερων μονοπατιών από όλες τις κορυφές σε όλες τις άλλες που περνούν από αυτόν τον κόμβο. Ένας κόμβος με υψηλή ενδιάμεση κεντρικότητα έχει μεγάλη επιρροή στην "μεταφορά" αντικειμένων στο δίκτυο, υπό την προϋπόθεση ότι η μεταφορά του αντικειμένου ακολουθεί το κοντύτερο μονοπάτι. Πιο συγκεκριμένα η ενδιάμεση κεντρικότητα ενός κόμβου n δίνεται από την παρακάτω έκφραση:

$$g(n) = \sum_{s \neq n \neq t} \frac{\sigma_{st}(n)}{\sigma_{st}} \quad (2.11)$$

όπου σ_{st} είναι ο συνολικός αριθμός των κοντύτερων μονοπατιών από τον κόμβο s στον κόμβο t και $\sigma_{st}(n)$ είναι ο αριθμός των μονοπατιών που περνούν από τον n . Η ενδιάμεση κεντρικότητα ενός κόμβου μετρείται με τον αριθμό των ζευγών των κόμβων όπως συνεπάγεται από τους δείκτες άθροισης. Επομένως ο υπολογισμός προσαρμόζεται διαιρώντας με τον αριθμό των ζευγών του κόμβου μη συμπεριλαμβανομένου του n , έτσι ώστε $g \in [0, 1]$. Η διαίρεση γίνεται με $(N - 1)(N - 2)$ για κατευθυνόμενους γράφους και $(N - 1)(N - 2)/2$ για μη κατευθυνόμενους γράφους.

Η Ενδιάμεση Κεντρικότητα σε δίκτυα πρότυπα

Έχει δειχθεί ότι η ενδιάμεση κεντρικότητα ενός scale-free δικτύου ακολουθεί μια κατανομή με νόμο δύναμης που δίνεται από έναν εκθέτη φόρτωσης δ ,

$$P(g) \approx g^{-\delta} \quad (2.12)$$

από αυτό συνεπάγεται η σχέση κλιμάκωσης με τον βαθμό του κόμβου,

$$g(k) \approx k^{-r} \quad (2.13)$$

Όπου $g(k)$ είναι το μέσο "φόρτωμα" κορυφών με βαθμό k . Τα εκθετικά δ και r δεν είναι ανεξάρτητα καθώς από την 2.12 συνεπάγεται

$$P(g) = \int dk P(k) \delta(g - k^r) \quad (2.14)$$

Για μεγάλο g και επομένως μεγάλο k η έκφραση γίνεται

$$P(g \gg 1) = \int dk k^{-\gamma} \delta(g - k^r) \quad (2.15)$$

$$\sim g^{-1 - \frac{\gamma-1}{r}}$$

το οποίο αποδεικνύει την παρακάτω ισότητα:

$$r = \frac{\gamma - 1}{\delta - 1}$$

Το r φαίνεται να είναι το σημαντικό εκθετικό το οποίο δείχνει πως η ενδιάμεση κεντρικότητα εξαρτάται από την συνδετικότητα. Η κατάσταση που μεγιστοποιεί την ενδιάμεση κεντρικότητα για ένα κόμβο είναι όταν όλα τα κοντύτερα μονοπάτια περνάνε από αυτόν, το οποίο αντιστοιχεί στην δομή του δέντρου (ένα δίκτυο χωρίς ομαδοποίηση (clustering)). Σε αυτήν την περίπτωση το r φτάνει στην μέγιστη τιμή του η οποία είναι $r = 2$.

$$r = 2 \rightarrow \delta = \frac{\gamma + 1}{2}$$

Αυτή η μέγιστη τιμή του r (και επομένως ελάχιστη τιμή του δ) βάζει όρια στους εκθέτες φόρτωσης για δίκτυα με μη-μηδενική ομαδοποίηση.

$$r \leq 2 \rightarrow \delta \geq \frac{\gamma + 1}{2}$$

Σε αυτήν την περίπτωση, οι εκθέτες δ , r δεν είναι γενικοί και εξαρτώνται από τις διάφορες λεπτομέρειες όπως η μέση συνδετικότητα, η ομαδοποίηση κ.α. [73, 74, 75].

Η Ενδιάμεση Κεντρικότητα σε Πραγματικά Δίκτυα

Η κατανομή φόρτωσης των πραγματικών δικτύων με scale-free χαρακτηριστικά επίσης ακολουθεί μια κατανομή με νόμο δύναμης. Αυτό είναι ένα διαισθητικό αποτέλεσμα. Τα scale-free δίκτυα από μόνα τους δημιουργούν μικρού μήκους μονοπάτια κατά μήκος του δικτύου δημιουργώντας λίγους *hub κόμβους* με πολύ υψηλότερη συνδετικότητα από την πλειοψηφία των υπόλοιπων κόμβων του δικτύου. Αυτοί οι hub κόμβοι έχουν πολύ υψηλότερες φορτώσεις εξαιτίας της υψηλής συνδετικότητας [73, 74, 75].

2.4.7 Κεντρικότητα Εγγύτητας

Η Κεντρικότητα Εγγύτητας είναι ένα μέτρο που αποδίδει υψηλές τιμές σε κόμβους που βρίσκονται σε κοντινότερες αποστάσεις με τους υπόλοιπους του δικτύου και ορίζεται ως εξής:

$$Cl(i) = \frac{1}{\sum_j d(i, j)} \quad (2.16)$$

όπου $d(i, j)$ η απόσταση των κόμβων i και j .

Η κεντρικότητα εγγύτητας μπορεί επίσης να θεωρηθεί ως η ικανότητα κάθε κόμβου να μεταδίδει πληροφορίες σε όλους τους υπόλοιπους κόμβους του δικτύου. Η ύπαρξη μικρής μέσης απόστασης ενός κόμβου από τους υπόλοιπους, συνεπάγεται αυτόματα και καλύτερη ικανότητα στη διάδοση πληροφορίας. Είναι επομένως φυσικό επακόλουθο να αποδίδουμε υψηλή κεντρικότητα σε έναν τέτοιο κόμβο [76].

2.5 Bayesian Δίκτυα

Τα Bayesian δίκτυα παρέχουν μια καλή και εύχρηστη αναπαράσταση για την έκφραση της από κοινού κατανομής πιθανότητας και για τον συμπερασμό. Τα δίκτυα αυτά γίνονται όλο και περισσότερο σημαντικά στις βιολογικές επιστήμες για τον συμπερασμό κυτταρικών δικτύων, την μοντελοποίηση πρωτεϊνικών σηματοδοτικών μονοπατιών, συστημική βιολογία, την ολοκλήρωση δεδομένων, την ταξινόμηση και ανάλυση γενετικών δεδομένων. Η αναπαράσταση και η χρήση της θεωρίας πιθανοτήτων κάνουν τα Bayesian δίκτυα κατάλληλα για το συνδυασμό του πεδίου γνώσης και δεδομένων, την έκφραση αιτιωδών σχέσεων, την αποφυγή υπερκάλυψης ενός μοντέλου σε δεδομένα εκπαίδευσης, και για την εκμάθηση από μη ολοκληρωμένα σύνολα δεδομένων. Ο πιθανολογικός φορμαλισμός παρέχει μια φυσική απειλή για τη στοχαστική φύση των βιολογικών συστημάτων και μετρήσεων. Αυτό στοχεύει στην εισαγωγή των Bayesian δικτύων στην υπολογιστική βιολογία, εστιάζοντας στις έννοιες πίσω από τις μεθόδους για την εκμάθηση των παραμέτρων και την δομή των μοντέλων. Υπάρχουν πολλές εφαρμογές στην βιολογία όπου επιθυμούμε να γίνει κατηγοριοποίηση των δεδομένων. Για να επιλυθούν τέτοια προβλήματα, είναι απαραίτητο ένα σύνολο κανόνων το οποίο μπορεί να χρησιμοποιηθεί για πρόβλεψη. Τέτοια γνώση δεν είναι συχνά διαθέσιμη [77].

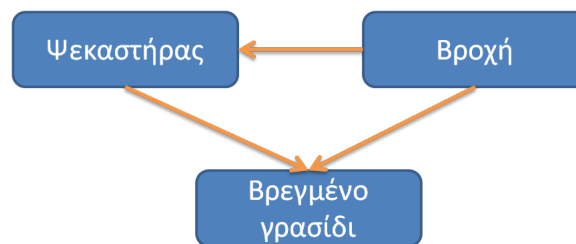
Οι προσεγγίσεις για την εκμάθηση μηχανών παράγουν συχνά καλύτερα αποτελέσματα. Ένας μεγάλος αριθμός παραδειγμάτων χρησιμοποιείται για να προσαρμόσει τις παραμέτρους ενός μοντέλου οι οποίες μετά μπορούν να χρησιμοποιηθούν για να γίνει πρόβλεψη ή ταξινόμηση στα δεδομένων. Υπάρχουν πολλοί διαφορετικοί τύποι μοντέλων και πολλές διαφορετικές προσεγγίσεις που μπορεί να απαιτηθούν για την εκπαίδευση των μοντέλων, η κάθε μια με τα θετικά και τα αρνητικά της [77].

Τα νευρωνικά δίκτυα, για παράδειγμα, είναι ικανά να μάθουν ένα μοντέλο από τα δεδομένα εκπαίδευσης, αλλά είναι συχνά δύσκολο να εξάγουν πληροφορία για το μοντέλο, το οποίο με άλλες μεθόδους μπορεί να παρέχει πολύτιμες πληροφορίες στα δεδομένα ακόμα και να λύσει το πρόβλημα. Ένα σύνηθες πρόβλημα στην εκμάθηση μηχανών είναι η υπερκάλυψη, όπου το μοντέλο στο οποίο έχει γίνει η εκμάθηση είναι πολύπλοκο. Αυξάνοντας το μέγεθος του συνόλου δεδομένων εκπαίδευσης ίσως

μικραίνει το πρόβλημα. Όμως, αυτό υποθέτει να είναι διαθέσιμα περισσότερα δεδομένα εκπαίδευσης, κάτι το οποίο δεν είναι το ζήτημα. Επιπλέον, συχνά είναι σημαντικό να καθορίζεις την αβεβαιότητα στις παραμέτρους του μοντέλου στο οποίο έχει γίνει η εκμάθηση, ακόμα και στην επιλογή του μοντέλου. Αυτό στοχεύει στην χρήση των Bayesian δικτύων, που προσφέρει λύση σε αυτά τα θέματα. Η χρήση της Bayesian θεωρίας πιθανοτήτων παρέχει μηχανισμούς για την περιγραφή της αβεβαιότητας και για την προσαρμογή του αριθμού των παραμέτρων στο μέγεθος των δεδομένων. Η χρήση μιας γραφικής αναπαράστασης παρέχει έναν απλό τρόπο για την οπτικοποίηση της δομής του μοντέλου. Ο έλεγχος των μοντέλων μπορεί να παρέχει πολύτιμες πληροφορίες στις ιδιότητες των δεδομένων και να επιτρέπει την παραγωγή νέων μοντέλων [77].

Ίσως η πιο ενδιαφέρουσα εφαρμογή των Bayesian δικτύων στον τομέα της βιολογίας είναι η μοντελοποίηση των δικτύων και των μονοπατιών. Αυτές οι αναλύσεις συνδυάζουν όλα τα χαρακτηριστικά των Bayesian δικτύων: την ικανότητα να μαθαίνουν από μη ολοκληρωμένα δεδομένα με θόρυβο, την ικανότητα να συνδυάζουν εξειδικευμένη γνώση και δεδομένα που εξάγουν μια κατάλληλη δομή δικτύου και την ικανότητα να εκφράζουν αιτιώδεις σχέσεις [77].

Τα Bayesian δίκτυα ανήκουν στην κατηγορία των πιθανοτικών γραφικών μοντέλων. Τα πιθανοτικά γραφικά μοντέλα παρέχουν ένα πλαίσιο που υποστηρίζει και τις τρεις βασικές διαδικασίες που είναι απαραίτητες για την κατασκευή ενός έξυπνου συστήματος: αναπαράσταση, επαγωγή συμπερασμάτων και εκμάθηση. Θα ξεκινήσουμε, εισάγοντας το κομμάτι της αναπαράστασης, η οποία αποτελεί μια λογική κωδικοποίηση του πραγματικού κόσμου που μας ενδιαφέρει. Στόχος μας είναι να αναπαραστήσουμε μια από κοινού κατανομή P πάνω σε ένα σύνολο τυχαίων μεταβλητών $X = \{X_1, \dots, X_n\}$, έτσι ώστε να είμαστε σε θέση να απαντούμε σε μια ευρεία γκάμα ερωτημάτων που μας ενδιαφέρουν.



Εικόνα 2.18 – Ένα απλό Bayesian δίκτυο. Η βροχή επηρεάζει το πόσο θα ενεργοποιηθεί ψεκαστήρας. Επίσης και η βροχή αλλά και ο ψεκαστήρας επηρεάζουν την υγρασία του γρασιδιού.

2.5.1 Αναπαράσταση Bayesian Δικτύων

Με χρήση βασικών στοιχείων των πιθανοτήτων και της θεωρίας γράφων μπορούμε πλέον να κάνουμε μία πρώτη σύνδεση μεταξύ του κανόνα του Bayes και της έννοιας του άκυκλου κατευθυνόμενου γράφου, προκειμένου να περιγράψουμε τα Bayesian δίκτυα. Τα Bayesian δίκτυα αναπαριστούν τη δομή ενός ΠΓΜ (πιθανοτικού γραφικού

μοντέλου) με τη χρήση γράφων για την αποτύπωση των υποθέσεων ανεξαρτησίας μεταξύ των μεταβλητών. Τα δίκτυα αυτά είναι πολύ χρήσιμα καθώς μπορούν να διαβαστούν λεπτομέρειες του μοντέλου απευθείας από το γράφο και δίνουν το όφελος της υπολογιστικής αποδοτικότητας. Αν και τα Bayesian δίκτυα δεν μπορούν να αποτυπώσουν όλες τις δυνατές σχέσεις μεταξύ των μεταβλητών, είναι ιδανικά για την αναπαράσταση σχέσεων αιτίου-αποτελέσματος.

Στην περίπτωση που οι πιθανοτικές κατανομές που θέλουμε να αναπαραστήσουμε είναι πολυδιάστατες, χρησιμοποιούμε τα Bayesian δίκτυα ώστε να εκμεταλλευτούμε τις ιδιότητες ανεξαρτησίας μεταξύ των μεταβλητών. Ακόμα και στην απλούστερη περίπτωση που οι τυχαίες μεταβλητές $X = \{X_1, \dots, X_n\}$ είναι δυαδικές, η από κοινού κατανομή $P(X)$ θα απαιτεί τον ορισμό $2^n - 1$ αναθέσεων, το οποίο ειδικά για μεγάλα n είναι αριθμός που δύσκολα διαχειρίζεται και αποθηκεύεται στη μνήμη. Όπως και δύσκολα, μπορούμε να αποκτήσουμε τόσες πολλές τιμές από ειδικούς ή να εκτιμήσουμε τόσες πολλές παραμέτρους από διαδικασία εκμάθησης.

2.5.2 Ορισμός Bayesian δικτύου

Ορίζουμε τη σχέση μεταξύ ενός δοθέντος κατευθυνόμενου γράφου K κόμβων με την αντίστοιχη κατανομή πάνω στις μεταβλητές με τη χρήση των Bayesian δικτύων:

Ορισμός: [Bayesian Δίκτυα] Ένα Bayesian δίκτυο είναι:

- Ένας κατευθυνόμενος ακυκλικός γράφος (DAG) G του οποίου οι κόμβοι αναπαριστούν τις τυχαίες μεταβλητές x_1, \dots, x_K
- Για κάθε κόμβο x_i αντιστοιχεί μια υπό συνθήκη κατανομή πιθανότητας $P(x_i | pa_G(x_i))$
- Το Bayesian δίκτυο αναπαριστά την από κοινού κατανομή, με τη χρήση του κανόνα της αλυσίδας, ως το γινόμενο των υπό συνθήκη κατανομών των κόμβων του γράφου, με συνθήκη τους γονείς του κάθε κόμβου:

$$p(X) = P(x_1, \dots, x_K) = \prod_k^K p(x_k | pa_K) \quad (2.17)$$

όπου pa_K δηλώνει σύνολο γονιών του x και $X = \{x_1, \dots, x_K\}$.

Η εξίσωση αυτή αναπαριστά τις ιδιότητες παραγοντοποίησης μιας από κοινού κατανομής για ένα κατευθυνόμενο γραφικό μοντέλο.

2.5.3 Κανόνας αλυσίδας

Σε αυτό το σημείο, θα περιγράψουμε αναλυτικά την εφαρμογή του κανόνα της αλυσίδας σε μια τυχαία πιθανοτική κατανομή που αναπαρίσταται από ένα Bayesian δίκτυο. Έστω για παράδειγμα, μία τυχαία από κοινού κατανομή $p(a, b, c)$ πάνω σε τρεις τυχαίες μεταβλητές a, b και c . Οι μεταβλητές αυτές μπορεί να είναι είτε συνεχείς είτε διακριτές. Με την εφαρμογή του κανόνα του γινομένου από τη θεωρία πιθανοτήτων (product rule

of probability), μπορούμε να γράψουμε την από κοινού κατανομή στη μορφή:

$$p(a, b, c) = p(c|a, b)p(a, b) \quad (2.18)$$

Μια δεύτερη εφαρμογή του κανόνα του γινομένου, αυτή τη φορά στο δεύτερο όρο στο δεξιό μέλος της 2.18 μας δίνει:

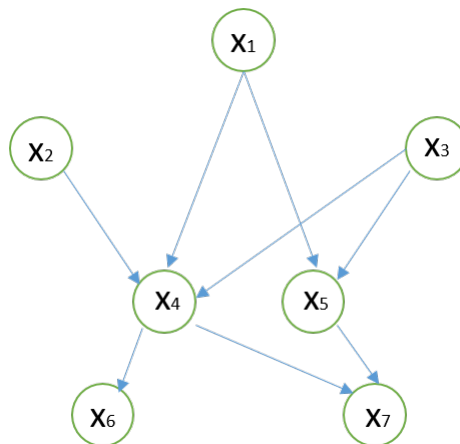
$$p(a, b, c) = p(c|a, b)p(b|a)p(a) \quad (2.19)$$

Μπορούμε τώρα να ορίσουμε το δεξιό μέλος της 2.19 ως ακολούθως: Πρώτα θεωρούμε ένα κόμβο για κάθε μία τυχαία μεταβλητή a , b και c και συσχετίζουμε κάθε κόμβο με την αντίστοιχη υπό συνθήκη κατανομή στο δεξί μέλος της 2.19. Ακολούθως, για κάθε μία υπό συνθήκη κατανομή προσθέτουμε στο γράφο κατευθυνόμενες ακμές (βέλη) που ξεκινούν από τους κόμβους που αντιστοιχούν στις μεταβλητές πάνω στις οποίες η κατανομή είναι υπό συνθήκη. Έτσι, για τον παράγοντα $p(c|a, b)$ θα υπάρχουν σύνδεσμοι από τους κόμβους a και b , ενώ για τον παράγοντα $p(a)$ δεν θα υπάρχουν εισερχόμενοι σύνδεσμοι. Εάν υπάρχει σύνδεσμος από τον κόμβο a προς τον κόμβο b θα λέμε ότι ο κόμβος a είναι γονιός του κόμβου b , και ότι ο κόμβος b είναι παιδί του κόμβου a .

Για να κάνουμε την αποσύνθεση στην 2.19 έχουμε διαλέξει μία ορισμένη σειρά στις μεταβλητές, συγκεκριμένα τις a, b, c . Αν είχαμε επιλέξει διαφορετική σειρά, θα προέκυπτε διαφορετική αποσύνθεση και επομένως διαφορετική γραφική αναπαράσταση.

Γενικεύουμε το παραπάνω παράδειγμα ως εξής: θεωρούμε την από κοινού κατανομή πάνω σε K μεταβλητές $p(x_1, \dots, x_K)$ και εφαρμόζουμε επανειλημμένα τον κανόνα γινομένου της θεωρίας πιθανοτήτων. Οπότε, η από κοινού κατανομή μπορεί να γραφεί ως γινόμενο υπό συνθήκη κατανομών, μία για κάθε μία από τις μεταβλητές

$$p(x_1, \dots, x_K) = p(x_K|x_1, \dots, x_{K-1}) \dots p(x_2|x_1)p(x_1) \quad (2.20)$$



Εικόνα 2.19 – Παράδειγμα κατευθυνόμενου άκυκλου γράφου που περιγράφει την κατανομή πάνω στις μεταβλητές x_1, \dots, x_K .

Για μια συγκεκριμένη επιλογή του , μπορούμε και πάλι να το αναπαραστήσουμε ως ένα κατευθυνόμενο γράφο που έχει κόμβους, έναν για κάθε μία υπό συνθήκη κατανομή στο δεξιό μέλος της 2.20, με κάθε ένα κόμβο να έχει εισερχόμενους συνδέσμους από όλους τους χαμηλότερα αριθμημένους κόμβους. Θα λέμε ότι αυτός ο γράφος είναι *πλήρως συνδεδεμένος* (fully connected) αφού υπάρχει σύνδεσμος μεταξύ οποιουδήποτε ζευγαριού κόμβων.

Έως τώρα ασχοληθήκαμε με τελείως γενικές από κοινού κατανομές, με αποτέλεσμα οι γραφικές αναπαραστάσεις τους να είναι πλήρως συνδεδεμένοι γράφοι. Στην επόμενη ενότητα, όμως, θα διαπιστώσουμε ότι η απουσία συνδέσμων σε ένα γράφο, η οποία υπονοεί ανεξαρτησία μεταξύ των αντίστοιχων μεταβλητών, είναι αυτή η οποία μας δίνει ενδιαφέροντα συμπεράσματα για τις ιδιότητες των κατανομών που αναπαριστούν.

Για το σκοπό αυτό, θεωρούμε ως παράδειγμα ένα μερικώς συνδεδεμένο γράφο που φαίνεται στην Εικόνα 2.19, στον οποίο δεν υπάρχει σύνδεσμος από το x_1 στο x_2 ή από το x_3 στο x_7 . Προκειμένου να πάμε από αυτό το γράφο στην αντίστοιχη από κοινού κατανομή (η οποία γράφεται ως γινόμενο ενός συνόλου υπό συνθήκη κατανομών, μίας για κάθε κόμβο στο γράφο) λαμβάνουμε υπόψη ότι κάθε μία τέτοια υπό συνθήκη κατανομή θα τίθεται υπό συνθήκη μόνο πάνω στους γονιούς του αντίστοιχου κόμβου στο γράφο. Για παράδειγμα, η μεταβλητή x_5 θα είναι υπό συνθήκη μόνο πάνω στις μεταβλητές x_1, x_3 . Οπότε η από κοινού κατανομή θα δίνεται από

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \quad (2.21)$$

2.5.4 Υπό Συνθήκη ανεξαρτησία σε Bayesian δίκτυα

Ορισμός: [Υπό Συνθήκη Ανεξαρτησία] Έστω τρεις τυχαίες μεταβλητές a, b, c και έστω ότι η από κοινού κατανομή των a, b είναι τέτοια που δεν εξαρτάται από το b , ώστε $p(a|b, c) = p(a|c)$. Θα λέμε ότι τα a, b είναι υπό συνθήκη ανεξάρτητα δοθέντος του c . Εναλλακτικά, μπορούμε να το γράψουμε ως εξής:

$$p(a, b|c) = p(a|b, c)p(b, c) = p(a|c)p(b|c). \quad (2.22)$$

Αυτό σημαίνει ότι εάν έχουμε ως συνθήκη το c , η από κοινού κατανομή παραγοντοποιείται σε γινόμενο των περιθωρίων κατανομών των a, b , όπου και αυτές είναι υπό συνθήκη του c . Άρα, δοθέντος c οι a, b είναι στατιστικά ανεξάρτητες: $a \perp b|c$

Η υπό συνθήκη ανεξαρτησία παίζει σημαντικό ρόλο στα πιθανοτικά γραφικά μοντέλα καθώς απλοποιεί τη δομή των μοντέλων αυτών και τους υπολογισμούς που χρειάζονται για να κάνουμε επαγωγή συμπερασμάτων και μάθηση παραμέτρων ή/και δομής πάνω στο μοντέλο.

2.5.5 d-διαχωρισμός

Ένα σημαντικό χαρακτηριστικό των γραφικών μοντέλων είναι ότι οι ιδιότητες της υπό συνθήκη ανεξαρτησίας μπορούν να διαβαστούν απευθείας από το γράφο, χωρίς να

χρειάζεται να κάνουμε αναλυτικούς χειρισμούς.

Το γενικό πλαίσιο μέσα στο οποίο το καταφέρνουμε να διαβάσουμε τις ιδιότητες της υπό συνθήκη ανεξαρτησίας απευθείας από το Bayesian δίκτυο ονομάζεται *d-διαχωρισμός*, όπου το *d* σημαίνει directed. Ακολούθως, θα εξετάσουμε την έννοια της *d-διαχωρισμός* και τις υποπεριπτώσεις που συναντάμε, με τη χρήση σχετικών παραδειγμάτων γράφων που τις καταδεικνύουν.



Εικόνα 2.20 – Πρώτο παράδειγμα γράφου πάνω στις a, b, c όταν δεν έχουμε και όταν έχουμε παρατηρήσει την τιμή της c

Το πρώτο από τα παραδείγματα φαίνεται στην Εικόνα 2.20 και η από κοινού κατανομή πάνω στις μεταβλητές a, b, c είναι

$$p(a, b, c) = p(a|c)p(b|c)p(c) \quad (2.23)$$

όπως προκύπτει από εφαρμογή του γενικού κανόνα 2.17. Αν καμία από τις μεταβλητές δεν παρατηρείται, μπορούμε να ελέγξουμε αν οι a, b είναι ανεξάρτητες αθροίζοντας και τα δύο μέλη της 2.23 ως προς το c . Βλέπουμε ότι δεν παραγοντοποιείται σε γινόμενο των $p(a)p(b)$ οπότε δεν είναι ανεξάρτητες. Αν θέσουμε υπό συνθήκη το c , δηλαδή το παρατηρούμε, τότε ο γράφος αναπαρίσταται όπως δείχνει η Εικόνα 2.20, με τον υπό παρατήρηση κόμβο σκιασμένο. Συνεπώς από την 2.23, μπορούμε να γράψουμε:

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = p(a|c)p(b, c) \quad (2.24)$$

οπότε έχουμε την ιδιότητα της υπό συνθήκης ανεξαρτησίας.

Αυτό το αποτέλεσμα μπορούμε να το αναπαραστήσουμε γραφικά αν θεωρήσουμε μονοπάτι από τον κόμβο a στον κόμβο b μέσω του κόμβου c . Ο κόμβος c θα λέμε ότι είναι *ουρά με ουρά* (tail-to-tail) όσον αφορά αυτό το μονοπάτι, διότι ο κόμβος αυτός είναι συνδεδεμένος με τις ουρές των δύο τόξων και η παρουσία ενός τέτοιου μονοπατιού κάνει τους δύο κόμβους να εξαρτώνται μεταξύ τους. Όμως, όταν παρατηρούμε τον κόμβο c , ο υπό συνθήκη κόμβος μπλοκάρει το μονοπάτι από το a στο b και κάνει τα a και b να είναι υπό συνθήκη ανεξάρτητα.



Εικόνα 2.21 – Το δεύτερο παράδειγμα γράφου πάνω στις a, b, c όταν δεν έχουμε και όταν έχουμε παρατηρήσει την τιμή της c

Στο δεύτερο παράδειγμα, ας θεωρήσουμε το γράφο που φαίνεται στην Εικόνα 2.21.

Η από κοινού κατανομή που αντιστοιχεί σε αυτό το γράφο μπορεί και πάλι να αποκτηθεί από την εφαρμογή του κανόνα 2.17 και είναι

$$p(a, b, c) = p(a)p(c|a)p(b|c) \quad (2.25)$$

Πρώτα ας υποθέσουμε ότι καμία μεταβλητή δεν παρατηρείται, οπότε ομοίως με πριν μπορούμε να ελέγξουμε αν οι μεταβλητές a, b είναι ανεξάρτητες, αθροίζοντας και τα δύο μέλη της 2.25 ως προς την c . Βλέπουμε ότι προκύπτει ότι $p(a, b) = p(a)p(b|a)$ το οποίο γενικά δεν παραγοντοποιείται ως γινόμενο των περιθωρίων κατανομών. Αν τώρα παρατηρήσουμε τη μεταβλητή c , έχουμε

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = p(a|c)p(b|c) \quad (2.26)$$

οπότε παρατηρούμε υπό συνθήκη ανεξαρτησία.

Ερμηνεύουμε και πάλι τα αποτελέσματα γραφικά. Ο κόμβος c θα λέμε ότι είναι *κεφάλι με ουρά* (head-to-tail) όσον αφορά αυτό το μονοπάτι, και η παρουσία ενός τέτοιου μονοπατιού κάνει τους δύο κόμβους να εξαρτώνται μεταξύ τους. Όμως, όταν παρατηρούμε τον κόμβο c , ο υπό συνθήκη κόμβος μπλοκάρει το μονοπάτι από το a στο b και κάνει τα a και b να είναι υπό συνθήκη ανεξάρτητα.



Εικόνα 2.22 – Το τρίτο παράδειγμα γράφου πάνω στις a, b, c όταν δεν έχουμε και όταν έχουμε παρατηρήσει την τιμή της c

Στο τρίτο και τελευταίο παράδειγμα, θεωρούμε το γράφο που φαίνεται στην Εικόνα 2.21. Όπως θα δούμε αυτός ο γράφος έχει περίεργη συμπεριφορά σε σχέση με τα δύο προηγούμενα παραδείγματα. Η από κοινού κατανομή που αντιστοιχεί σε αυτό το γράφο μπορεί να αποκτηθεί ως εξής από την 2.17

$$p(a, b, c) = p(a)p(b)p(c|a, b) \quad (2.27)$$

Πρώτα ας υποθέσουμε ότι καμία μεταβλητή δεν παρατηρείται. Ομοίως, ελέγχουμε αν οι μεταβλητές a, b είναι ανεξάρτητες, αθροίζοντας και τα δύο μέλη της 2.27 ως προς την c . Βλέπουμε ότι προκύπτει ότι $p(a, b) = p(a)p(b)$, οπότε χωρίς να έχουμε καμία μεταβλητή υπό συνθήκη έχουμε ανεξαρτησία. Αν τώρα παρατηρήσουμε τη μεταβλητή c , έχουμε

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)} \quad (2.28)$$

το οποίο γενικά δεν παραγοντοποιείται ως $p(a|c)p(b|c)$. Άρα, το τρίτο παράδειγμα έχει αντίθετη συμπεριφορά από τα προηγούμενα δύο.

Γραφικά, θα λέμε ότι ο κόμβος c είναι *κεφάλι-κεφάλι* (head-to-head) όσον αφορά το μονοπάτι μεταξύ των κόμβων a, b διότι συνδέει τα κεφάλια των δύο τόξων. Όταν ο κόμβος c δεν παρατηρείται, τότε μπλοκάρει το μονοπάτι και επομένως οι μεταβλητές κόμβος a και κόμβος b είναι ανεξάρτητες. Όμως, όταν παρατηρείται το c , είναι δηλαδή υπό συνθήκη, τότε ελευθερώνει το μονοπάτι και επομένως οι δύο μεταβλητές κόμβοι είναι εξαρτημένες. Επίσης, μπορεί να αποδειχθεί ότι ένα head-to-head μονοπάτι δεν θα είναι πια μπλοκαρισμένο αν είτε ο κόμβος είτε οποιοσδήποτε από τους απογόνους του, παρατηρείται.

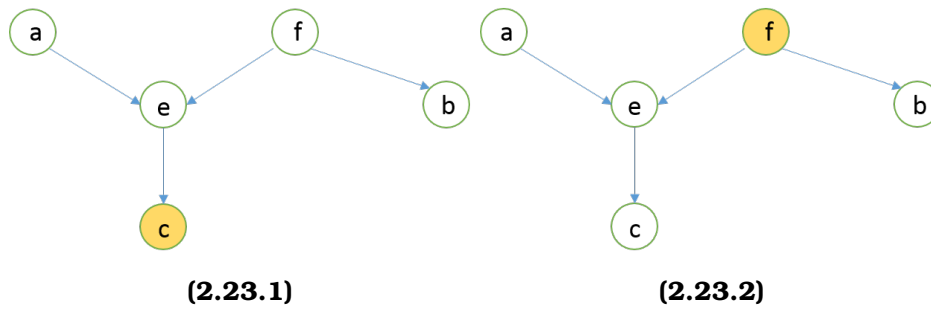
Συνολικά, συμπεραίνουμε ότι ένας tail-to-tail κόμβος ή ένας head-to-tail κόμβος αφήνει το μονοπάτι ελεύθερο, εκτός και αν παρατηρείται. Αντίθετα, ένας head-to-head κόμβος μπλοκάρει το μονοπάτι αν δεν παρατηρείται, αλλά όταν ο κόμβος και ή ένας ή περισσότεροι από τους απογόνους του παρατηρείται τότε το μονοπάτι γίνεται ελεύθερο.

Ορισμός: [d-διαχωρισμός] Έστω ένας κατευθυνόμενος γράφος στον οποίο A, B, C είναι τυχαία μη τέμνοντα σύνολα κόμβων (που η ένωσή τους είναι μικρότερη από το σύνολο των κόμβων στο γράφο). Θέλουμε να διαπιστώσουμε αν υπάρχει μία συγκεκριμένη δήλωση υπό συνθήκη ανεξαρτησίας στο δοσμένο ΚΑΓ. Για να το κάνουμε αυτό θεωρούμε όλα τα πιθανά μονοπάτια στο γράφο από οποιοδήποτε κόμβο στο A προς οποιοδήποτε κόμβο στο B . Οποιοδήποτε τέτοιο μονοπάτι λέγεται ότι είναι μπλοκαρισμένο αν συμπεριλαμβάνει έναν κόμβο τέτοιο ώστε

- τα βέλη στο μονοπάτι συναντιώνται είτε head-to-tail είτε tail-to-tail στον κόμβο και ο κόμβος είναι στο σύνολο του C
- τα βέλη συναντιώνται head-to-head στον κόμβο και ούτε ο κόμβος ούτε κάποιος από τους απογόνους του είναι στο σύνολο C

Αν όλα τα μονοπάτια είναι μπλοκαρισμένα, τότε θα λέμε ότι το A είναι d-διαχωρισμο από το B μέσω του C και η από κοινού κατανομή πάνω σε όλες τις μεταβλητές στο γράφο θα ικανοποιεί υπό συνθήκη ανεξαρτησία A, B δοθέντος του C .

Για παράδειγμα, στον γράφο που φαίνεται στην Εικόνα 2.23.1 το μονοπάτι από το a στο b δεν μπλοκάρεται από τον κόμβο f γιατί είναι tail-to-tail κόμβος και δεν παρατηρείται και ούτε μπλοκάρεται από τον e , διότι αν και ο τελευταίος είναι head-to-head κόμβος έχει έναν απόγονο που είναι στο σύνολο υπό συνθήκη. Ενώ, στο γράφο που απεικονίζει η Εικόνα 2.23.2 το μονοπάτι από το a στο b μπλοκάρεται από τον κόμβο f , καθώς είναι tail-to-tail κόμβος που παρατηρείται και η υπό συνθήκη ανεξαρτησία μεταξύ a, b δοθέντος c θα ικανοποιείται από οποιαδήποτε κατανομή που παραγοντοποιείται και αναπαρίσταται από αυτό το γράφο. Επίσης, αυτό το μονοπάτι μπλοκάρεται και από τον κόμβο e καθώς το e είναι head-to-head κόμβος και ούτε αυτός ούτε κάποιος από τους απογόνους του βρίσκεται στο σύνολο υπό συνθήκη [78, 79].



Εικόνα 2.23 – Έννοια d-διαχωρισμού

Κεφάλαιο 3

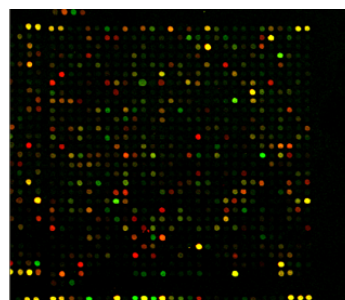
Ανάπτυξη Μεθοδολογίας

3.1 Μέθοδος

Στη μέθοδο που ακολουθήσαμε, έγινε χρήση δεδομένων γονιδιακής έκφρασης για τον εντοπισμό και την ταυτοποίηση μονοπατιών υπεύθυνα για τον καρκίνο του μαστού. Αρχικά διαλέξαμε ένα σύνολο διαφορετικά εκφραζόμενων γονιδίων (ΔΕΓ) χρησιμοποιώντας πίνακες συνόλων δεδομένων μεταξύ κανονικών (υγιών) και καρκινικών δειγμάτων. Στη συνέχεια ταυτοποιήσαμε σημαντικά ρυθμιζόμενα μονοπάτια και τελικά κατασκευάσαμε το δίκτυο των σημαντικά ρυθμιζόμενων μονοπατιών του καρκίνου του μαστού.

3.2 Δεδομένα προς επεξεργασία

Αρχικά μας δόθηκε ένα αρχείο το οποίο περιείχε 4174 γονίδια με τις τιμές έκφρασής τους για 425 καρκινικά και 104 υγιή δείγματα. Αυτά έχουν προέλθει από την προ επεξεργασία των δεδομένων μιας μικροσυστοιχίας με τη μέθοδο SAM (Significance Analysis of Microarrays). Σε ένα άλλο αρχείο που μας δόθηκε φαινόταν η κατάσταση υγείας του δείγματος ως προς τον καρκίνο. Επίσης, μας δόθηκε μια λίστα με 77 γονίδια που αποτελούν τη γονιδιακή υπογραφή, στην οποία κατέληξε ο κ. Νίκος Χλής εφαρμόζοντας ταξινόμηση (classification). Στα 77 γονίδια που αποτελούν μια γονιδιακή υπογραφή προσθέσαμε 5 γονίδια τα οποία είχαν σκοπό ελέγχου. Πιο συγκεκριμένα αυτά τα 5 γονίδια είναι σημαντικά γονίδια για τον καρκίνο του μαστού. Έχοντας καταλήξει σε 82 γονίδια προς επεξεργασία έγινε διαχωρισμός των καρκινικών και των υγιών δειγμάτων ώστε να δημιουργηθεί ένα δίκτυο για κάθε κατηγορία. Η διαδικασία προεπεξεργασίας των δεδομένων φαίνεται σχηματικά στην Εικόνα [3.1](#).



Χρήση μεθόδου SAM

4174 γονίδια
529 δείγματα



Ταξινόμηση

Γονιδιακή
υπογραφή
77 γονίδια
529 δείγματα



Προσθήκη γονιδίων ελέγχου

82 γονίδια
529 δείγματα

Εικόνα 3.1 – Διαδικασία προεπεξεργασίας δεδομένων.

3.3 Δημιουργία Bayesian Δικτύων

Τα Bayesian δίκτυα αντιπροσωπεύουν τη δομή εξάρτησης μεταξύ ποσοτήτων πολλών αλληλεπιδράσεων (π.χ., τα επίπεδα έκφρασης των διαφορετικών γονιδίων). Τα Bayesian Δίκτυα αποτελούν ένα πολλά υποσχόμενο εργαλείο για την ανάλυση προτύπων γονιδιακής έκφρασης. Πρώτον, είναι ιδιαίτερα χρήσιμα για την περιγραφή των διαδικασιών που αποτελούνται από τοπικά αλληλεπιδρούμενα συστατικά. Δηλαδή, η τιμή του κάθε συστατικού εξαρτάται άμεσα από τις τιμές ενός σχετικά μικρού αριθμού συστατικών. Δεύτερον, οι στατιστικές βάσεις για την εκμάθηση Bayesian δικτύων από τις παρατηρήσεις είναι κατανοητές και έχουν χρησιμοποιηθεί επιτυχώς σε πολλές εφαρμογές. Τέλος, τα Bayesian δίκτυα παρέχουν μοντέλα της αιτιώδους επιρροής: Αν και τα Bayesian δίκτυα είναι αυστηρά μαθηματικά ορισμένα από την άποψη των πιθανοτήτων και της υπό όρους ανεξαρτησίας, η σύνδεση μπορεί να γίνει μεταξύ αυτού του χαρακτηρισμού και της άμεσης αιτιώδους επιρροής.

Η υπόθεση Μαρκόφ Το γράφημα G αντιπροσωπεύει υποθέσεις υπό όρους ανεξαρτησίας που επιτρέπουν την από κοινού κατανομή να αναλυθεί, εξοικονομώντας παραμέτρους. Το γράφημα G κωδικοποιεί την υπόθεση Μαρκόφ:

- κάθε μεταβλητή X_i είναι ανεξάρτητη από τους μη-απογόνους της, δεδομένου των γονέων της στον γράφο G .

Με την εφαρμογή του κανόνα της αλυσίδας των πιθανοτήτων και των ιδιοτήτων της υπό όρους ανεξαρτησίας, κάθε από κοινού κατανομή που ικανοποιεί την υπόθεση Μαρκόφ μπορεί να αναλυθεί στην *product form*

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa^G(X_i)) \quad (3.1)$$

όπου $Pa^G(X_i)$ είναι το σύνολο των γονέων της X_i στο G .

3.3.1 Δεσμευμένη Κατανομή Πιθανότητας

Ένας γράφος G καθορίζει ένα *product form* σύμφωνα με την 3.1. Για να καθοριστεί πλήρως μια από κοινού κατανομή, πρέπει επίσης να οριστεί κάθε μια από τις δεσμευμένες πιθανότητες στην *product form*. Το δεύτερο μέρος των Bayesian δικτύων περιγράφει αυτές τις δεσμευμένες κατανομές $P(X_i | Pa^G(X_i))$ για κάθε μεταβλητή X_i . Ορίζουμε τις παραμέτρους που καθορίζουν αυτές τις κατανομές από το θ . Κατά τον καθορισμό των δεσμευμένων κατανομών μπορούμε να επιλέξουμε από διάφορες αναπαραστάσεις. Η επιλογή της αναπαράστασης εξαρτάται από τον τύπο των μεταβλητών:

(Έστω οι γονείς των μεταβλητών είναι $\{U_1, \dots, U_k\}$)

Διακριτές μεταβλητές Αν κάθε μια από τις X και U_1, \dots, U_k παίρνουν διακριτές τιμές από ένα πεπερασμένο σύνολο, τότε μπορούμε να αναπαραστήσουμε το $P(X|U_1, \dots, U_k)$ ως έναν πίνακα που ορίζει την πιθανότητα των τιμών των X για κάθε από κοινού εκχώρηση στο U_1, \dots, U_k . Για παράδειγμα αν όλες οι μεταβλητές είναι δυαδικές, τότε ο πίνακας ορίζει 2^k κατανομές.

Σε αυτήν την αναπαράσταση η κατανομή δεσμευμένης πιθανότητας δίνεται από:

$$P(X|U_1, \dots, U_k) = P(X_1, \dots, X_n|U_1, \dots, U_k) \quad (3.2)$$

$$= \frac{P((X_1, \dots, X_n) \cap (U_1, \dots, U_k))}{P(X_1, \dots, X_n)}. \quad (3.3)$$

Αυτό είναι μία γενική αναπαράσταση που μπορεί να περιγράψει κάθε διακριτή δεσμευμένη κατανομή. Έτσι, δεν χάνουμε την εκφραστικότητα με τη χρήση αυτής της αναπαράστασης. Η ευελιξία αυτή έχει και το μειονέκτημα της: Ο αριθμός των ελεύθερων παραμέτρων είναι εκθετικός ως προς τον αριθμό των γονέων.

Συνεχείς-Γκαουσιανές μεταβλητές Σε αντίθεση με την περίπτωση των διακριτών μεταβλητών, όταν η μεταβλητή X και οι γονείς της U_1, \dots, U_k είναι πραγματικές τιμές, δεν υπάρχει καμία αναπαράσταση που να μπορεί να αναπαραστήσει όλες τις πιθανές πυκνότητες. Μια φυσική επιλογή για τις συνεχείς κατανομές είναι η χρήση της Gaussian κατανομής. Αυτά μπορούν να αναπαρασταθούν σε ένα Bayesian δίκτυο με χρήση της γραμμικής Γκαουσιανής δεσμευμένης πυκνότητας. Σε αυτή την αναπαράσταση η δεσμευμένη πυκνότητα της X δεδομένου τους γονείς της δίνεται από:

$$P(X|u_1, \dots, u_k) \sim N(a_0 + \sum_i a_i u_i, \sigma^2). \quad (3.4)$$

Δηλαδή, η X κατανέμεται κανονικά γύρω από μια μέση τιμή που εξαρτάται γραμμικά από τις τιμές των γονέων της. Η διακύμανση αυτής της κανονικής κατανομής είναι ανεξάρτητη από τις τιμές των γονέων. Αν όλες οι μεταβλητές σε ένα δίκτυο έχουν γραμμικές Γκαουσιανές κατανομές, δεσμευμένες κατανομές, τότε η από κοινού κατανομή είναι μια πολυπαραγοντική Γκαουσιανή [80].

Υβριδικές μεταβλητές Όταν το δίκτυό μας περιλαμβάνει ένα μείγμα διακριτών και συνεχών μεταβλητών, θα πρέπει να εξετάσουμε πώς μπορούμε να αναπαραστήσουμε μια δεσμευμένη κατανομή για μια συνεχή μεταβλητή με διακριτούς γονείς, και για μια διακριτή μεταβλητή με συνεχείς γονείς. Εδώ δεν μας ενδιαφέρει η δεύτερη υπόθεση. Όταν μια συνεχής μεταβλητή έχει διακριτούς γονείς, χρησιμοποιούμε δεσμευμένη Γκαουσιανή κατανομή [80], στην οποία για κάθε από κοινού εκχώρηση στους διακριτούς γονείς της, εμείς αναπαριστούμε μια γραμμική Gaussian κατανομή της X δεδομένου των συνεχών γονέων της.

Για την δημιουργία των δυο δικτύων (ένα για τα καρκινικά και ένα για τα υγιή

δείγματα) εξετάσαμε τις περιπτώσεις των διακριτών και συνεχών-γκαουσιανών μεταβλητών.

3.3.2 Διακριτές Μεταβλητές

Σε αυτήν την περίπτωση θέλουμε να κατασκευάσουμε Bayesian δίκτυα με διακριτές μεταβλητές. Τα δεδομένα μας, δηλαδή οι τιμές γονιδιακής έκφρασης των γονιδίων είναι πραγματικές τιμές και ακολουθούν μια κανονική κατανομή (όπως θα δούμε παρακάτω). Εφόσον θέλουμε να κατασκευάσουμε δίκτυο με διακριτές μεταβλητές πρέπει τα δεδομένα μας να υποστούν διακριτοποίηση. Εμείς επιλέξαμε να γίνει η διακριτοποίηση έτσι ώστε κάθε τιμή έκφρασης να μπορεί να πάρει μία από τις τιμές ένα (1), δύο (2) ή τρία (3). Αυτό σημαίνει ότι ένα γονίδιο για να ένα συγκεκριμένο δείγμα θα χαρακτηρίζεται ως υποεκφρασμένο, κανονικό ή υπερεκφρασμένο αντίστοιχα. Για να γίνει η διακριτοποίηση των μεταβλητών εντοπίστηκαν μέσα από μια διαδικασία 2 κατώφλια A και B . Έτσι τα γονίδια για ένα συγκεκριμένο δείγμα που έχουν τιμή έκφρασης

- μικρότερη από A θεωρούνται *υποεκφρασμένα*
- μεγαλύτερη ή ίση του A και μικρότερη του B θεωρούνται *κανονικά*
- μεγαλύτερη ή ίση του B θεωρούνται *υπερεκφρασμένα*.

Διαδικασία διακριτοποίησης

Για να γίνει η διακριτοποίηση των τιμών έκφρασης έπρεπε να βρεθούν τα κατώφλια A και B σύμφωνα με τα οποία οι τιμές έκφρασης θα λάβουν μια νέα τιμή από ένα (1) έως τρία (3) όπως φαίνεται στον Πίνακα 3.1

Επίπεδο Έκφρασης	Κατώφλια για διακριτοποίηση	Τιμές έκφρασης μετά την διακριτοποίηση
Υποεκφρασμένα	$TE < A$	1
Κανονικά	$A \leq TE < B$	2
Υπερεκφρασμένα	$TE \geq B$	3

Πίνακας 3.1 – Κατώφλια διακριτοποίησης και τιμές έκφρασης μετά την διακριτοποίηση.

όπου TE οι τιμές έκφρασης των γονιδίων.

Η προσπάθεια εύρεσης των κατωφλίων αρχικά έγινε με τη βοήθεια των ιστογραμμάτων των τιμών έκφρασης όλων των δειγμάτων, αλλά και ξεχωριστά για τα καρκινικά και υγιή δείγματα. Τα ιστογράμματα των τιμών έκφρασης των γονιδίων απεικονίζουν την συχνότητα εμφάνισης περιοχών τιμών έκφρασης. Ο λόγος που χρησιμοποιούμε τα ιστογράμματα των τιμών έκφρασης για την αναζήτηση κατωφλίων είναι επειδή ψάχνουμε να βρούμε κάποιες τιμές, στις οποίες υπάρχει μεγάλη διαφοροποίηση των τιμών και τα ιστογράμματα είναι κατάλληλα για να δούμε την κατανομή των τιμών αυτών. Οι τιμές έκφρασης των δεδομένων μας είχαν αριθμητικές τιμές περίπου από -7 έως 16 . Έτσι επιλέχθηκε το μήκος των περιοχών των τιμών έκφρασης να είναι

$sp = 0.1$ ώστε να έχουμε μια αρκετά καλή ακρίβεια στο ιστόγραμμα. Κάθε περίπτωση δειγμάτων (καρκινικά και υγιή) εξετάστηκε με βάση 30 γονίδια, από τα οποία τα 15 είναι υπερεκφρασμένα (overexpressed) και τα άλλα 15 υποεκφρασμένα (underexpressed). Πιο συγκεκριμένα, αυτά τα γονίδια είναι αποδεδειγμένα βιολογικά ότι, όταν ένας ανθρώπινος οργανισμός πάσχει από τον καρκίνο του μαστού, αυτά υπό ή υπέρ εκφράζονται αντίστοιχα. Τα επιλεγμένα γονίδια φαίνονται στον Πίνακα 3.2.

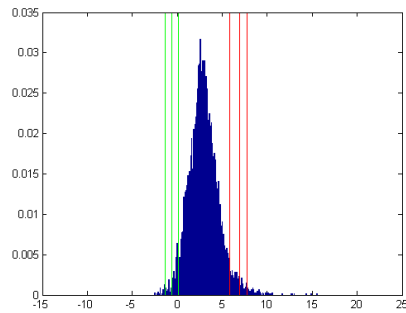
15 Επιλεγμένα Υπερεκφρασμένα Γονίδια	15 Επιλεγμένα Υποεκφρασμένα Γονίδια
GABRP	CAV1
FABP4	TIMP4
OLFM4	SLC20A1
KRT13	CD44
MMP7	HOXB6
MUC1	CTGF
GLUL	ECHDC2
S100A7	SAE1
S100A9	LCP1
SOD2	FUCA1
ING4	TNFRSF10B
CDH11	AGR2
CXCR4	ANGPT2
VEGFA	APOL1
ITM2B	EGFR

Πίνακας 3.2 – 30 επιλεγμένα γονίδια: 15 υπερεκφρασμένα και 15 υποεκφρασμένα.

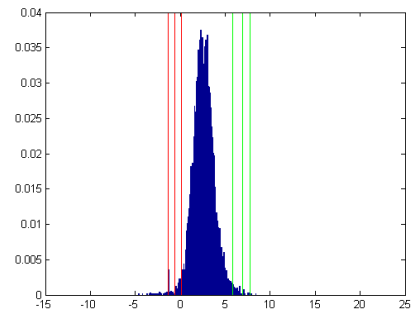
Τα γονίδια αυτά είναι ανάμεσα στα 4174 που βρίσκονται στο αρχείο των δεδομένων προς επεξεργασία.

Αυτό που προσπαθήσαμε να κάνουμε είναι να κατασκευάσουμε δυο ιστογράμματα για κάθε κατηγορία δειγμάτων (καρκινικά, υγιή και όλα), ένα για τα overexpressed και ένα για τα underexpressed γονίδια. Παρατηρώντας τα ποσοστά 95%, 98%, 99% στο ιστόγραμμα των overexpressed γονιδίων και τα ποσοστά 0, 5%, 1%, 2% στο ιστόγραμμα των underexpressed γονιδίων προσπαθήσαμε να εξάγουμε το άνω και κάτω κατώφλι αντίστοιχα. Με κόκκινες γραμμές στα ιστογράμματα της Εικόνας 3.2 φαίνονται τα παρατηρούμενα ποσοστά.

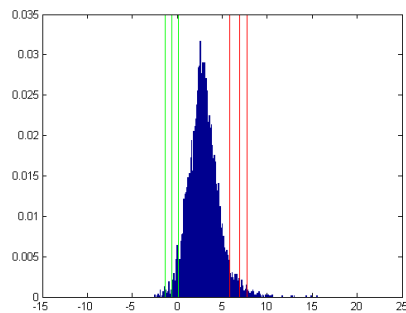
Από τα Gaussian Fit των ιστογραμμάτων που φαίνονται στην Εικόνα 3.3 παρατηρούμε ότι δεν υπάρχει μεγάλη διαφοροποίηση ανάμεσα στις δυο κατηγορίες Overexpressed και Underexpressed (κόκκινη-πράσινη καμπύλη αντίστοιχα) ώστε να συμπεράνουμε κάποια κατώφλια. Το ίδιο αρνητικά ήταν τα συμπεράσματά μας όταν προσπαθήσαμε να χρησιμοποιήσουμε τα *means* των τιμών έκφρασης κάθε γονιδίου.



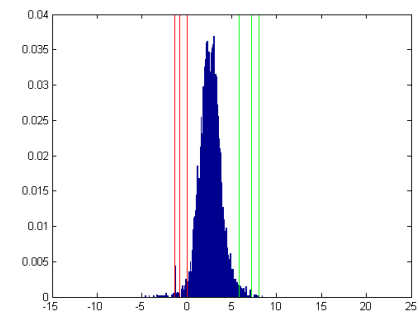
(3.2.1) Όλα τα δείγματα.
Υπερεκφρασμένα γονίδια.



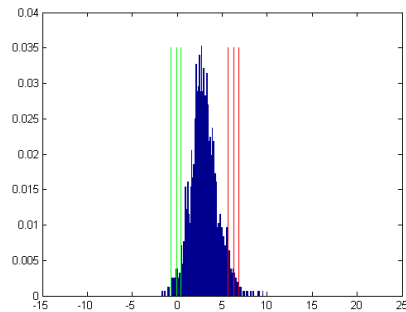
(3.2.2) Όλα τα δείγματα.
Υποεκφρασμένα γονίδια.



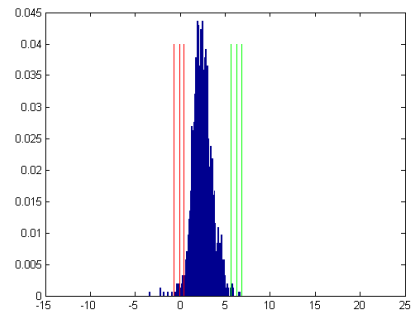
(3.2.3) Καρκινικά δείγματα.
Υπερεκφρασμένα γονίδια.



(3.2.4) Καρκινικά δείγματα.
Υποεκφρασμένα γονίδια.

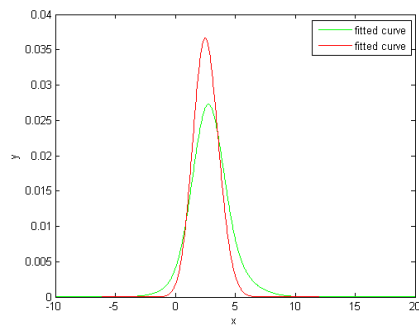


(3.2.5) Υγιή δείγματα.
Υπερεκφρασμένα γονίδια.

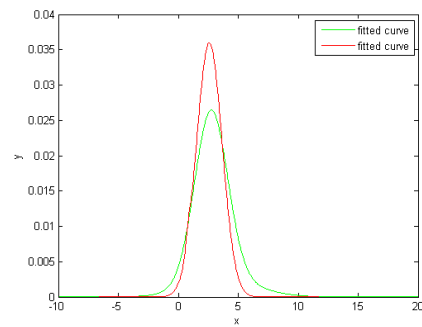


(3.2.6) Υγιή δείγματα.
Υποεκφρασμένα γονίδια.

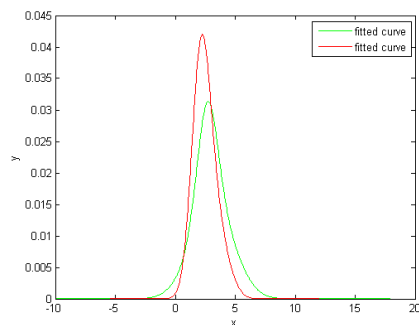
Εικόνα 3.2 – Ιστογράμματα για τη διαδικασία διακριτοποίησης με τη χρήση όλων των γονιδίων (4174)..



(3.3.1) Όλα τα δείγματα.



(3.3.2) Καρκινικά δείγματα.



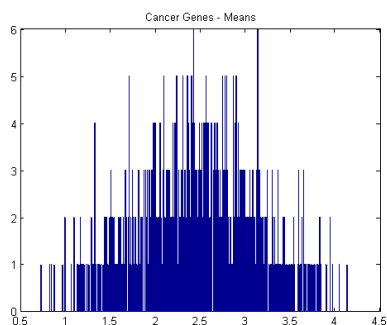
(3.3.3) Υγιή δείγματα.

Εικόνα 3.3 – Gaussian Fit για τη διαδικασία διακριτοποίησης με τη χρήση όλων των γονιδίων (4174).

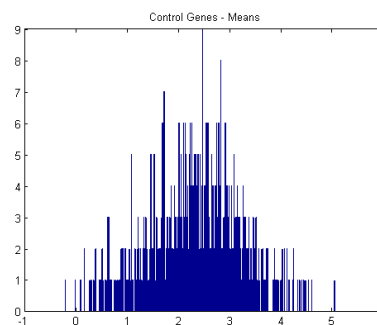
Αναζητώντας λοιπόν μια μεγαλύτερη διαφοροποίηση ανάμεσα στις δυο κατηγορίες γονιδίων, στραφήκαμε προς την επεξεργασία των γονιδίων με τα μικρότερα *q-values* που προέκυψαν από το t-test όλων των 4174 γονιδίων. Επιλέξαμε τα γονίδια με αυτές τις τιμές διότι θεωρητικά έχουν μεγαλύτερη διαφοροποίηση σε σύγκριση με γονίδια με μεγαλύτερα *q-values*. Αυτό αποδείχτηκε και πρακτικά αφού τελικά καταφέραμε να εξάγουμε κατώφλια.

Η διαδικασία ξεκινάει ξεχωρίζοντας τα 900 γονίδια με το χαμηλότερο *q-value* και υπολογίζοντας το *means* τους. Αυτό έγινε για τα καρκινικά αλλά και για τα υγιή δείγματα.

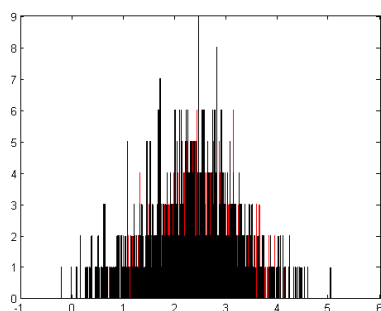
Από την Gaussian Fit (Εικόνα 3.4.4) των δυο ιστογραμμάτων των 900 γονιδίων για να καρκινικά (Εικόνα 3.4.1) και υγιή (Εικόνα 3.4.2) δείγματα, δεν μπορούμε πάλι να βγάλουμε ασφαλή συμπεράσματα για τα σημεία όπου θα ορίσουμε ως κατώφλια. Έτσι έγινε ένας διαχωρισμός στα *means*. Δημιουργήσαμε δύο ομάδες όπου στην μία ανήκουν αυτά με τις μέγιστες και στην άλλη αυτά με τις ελάχιστες τιμές. Για τα 900 γονίδια, εάν η τιμή του *mean* για τα καρκινικά δείγματα είναι μεγαλύτερη από αυτήν για τα υγιή τότε η πρώτη θα ανήκει στην ομάδα των μέγιστων και η δεύτερη στην ομάδα των ελάχιστων τιμών. Αναλόγως εάν η τιμή του *mean* για τα υγιή δείγματα είναι μεγαλύτερη από αυτήν για τα καρκινικά τότε η πρώτη θα ανήκει στην ομάδα των μέγιστων και η δεύτερη στην ομάδα των ελάχιστων τιμών.



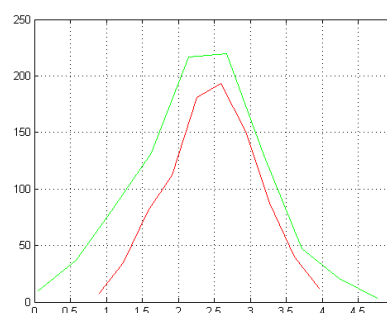
(3.4.1) Ιστόγραμμα:
Καρκινικά δείγματα.
Means 900 γονιδίων.



(3.4.2) Ιστόγραμμα:
Υγιή δείγματα.
Means 900 γονιδίων.



(3.4.3) Ιστόγραμμα:
Καρκινικά και Υγιή δείγματα.
Means 900 γονιδίων.



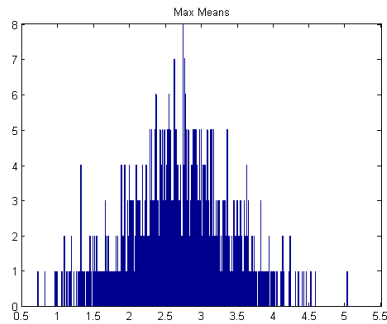
(3.4.4) Gaussian Fit:
Καρκινικά (κόκκινο) και Υγιή (πράσινο)
δείγματα.
Means 900 γονιδίων.

Εικόνα 3.4 – Ιστογράμματα και Gaussian Fit για τον εντοπισμό των κατωφλίων με τη χρήση των 900 γονιδίων με τη μεγαλύτερη διαφοροποίηση στην έκφραση.

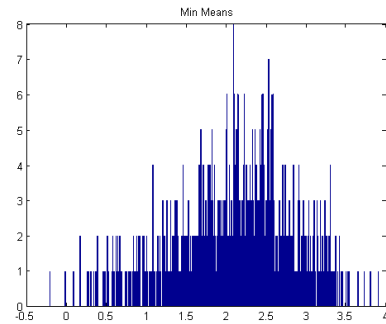
Από την Gaussian Fit που φαίνεται στην Εικόνα 3.5.4 παρατηρούμε αρκετά μεγάλη διαφοροποίηση η οποία είναι ικανή να μας δώσει δυο κατώφλια. Έτσι ορίσαμε ως το κάτω κατώφλι, το 10% των μέγιστων means και ως το πάνω κατώφλι το 90% των ελάχιστων means. Λέγοντας κάτω κατώφλι εννοούμε το κατώφλι κάτω από το οποίο αν βρίσκεται η τιμή έκφρασης ενός γονιδίου θεωρείται underexpressed, ενώ άνω κατώφλι εννοούμε το κατώφλι πάνω από το οποίο αν βρίσκεται η τιμή έκφρασης ενός γονιδίου θεωρείται overexpressed. Οι τιμές που προέκυψαν τελικά είναι 1,131 και 3,480 το κάτω και το πάνω κατώφλι αντίστοιχα. Τα κατώφλια φαίνονται στο ιστόγραμμα της Εικόνας 3.6 αλλά και στον πίνακα 3.3.

Επίπεδο Έκφρασης	Κατώφλια για διακριτοποίηση	Επίπεδο Έκφρασης μετά την διακριτοποίηση
Υποεκφρασμένα	$TE < 1,131$	1
Κανονικά	$1,131 \leq TE < 3,48$	2
Υπερεκφρασμένα	$TE \geq 3,48$	3

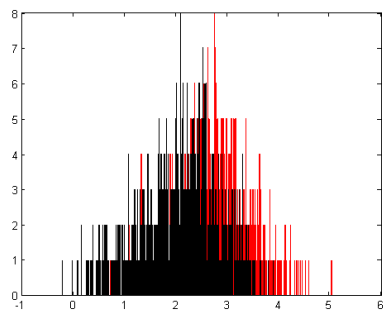
Πίνακας 3.3 – Πίνακας Κατωφλίων.



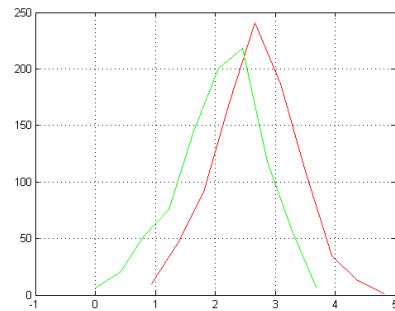
(3.5.1) Ιστόγραμμα:
Μέγιστων Means.



(3.5.2) Ιστόγραμμα:
Ελάχιστων Means.

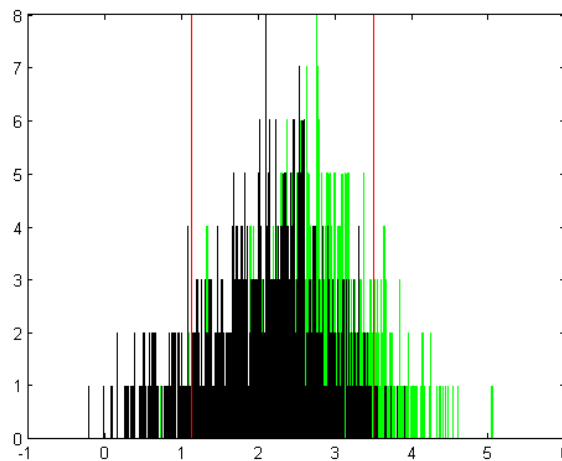


(3.5.3) Ιστόγραμμα:
Μέγιστων και Ελάχιστων Means.



(3.5.4) Gaussian Fit:
Μέγιστων και Ελάχιστων Means.

Εικόνα 3.5 – Ιστογράμματα και Gaussian Fit για τον εντοπισμό των κατώφλιών με τη χρήση των means των 900 γονιδίων με τη μεγαλύτερη διαφοροποίηση στην έκφραση.



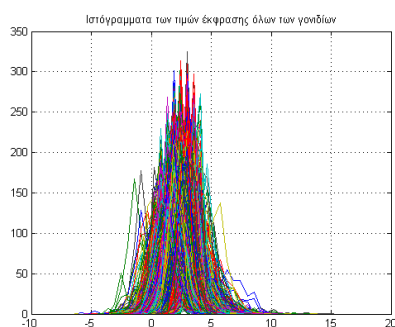
Εικόνα 3.6 – Με κόκκινες γραμμές φαίνονται τα κατώφλια που επιλέχθηκαν.

Σύμφωνα με αυτά τα κατώφλια έγινε η διακριτοποίηση των τιμών έκφρασης των δειγμάτων μας ώστε στη συνέχεια να δημιουργηθεί το Bayesian δίκτυο με διακριτές μεταβλητές.

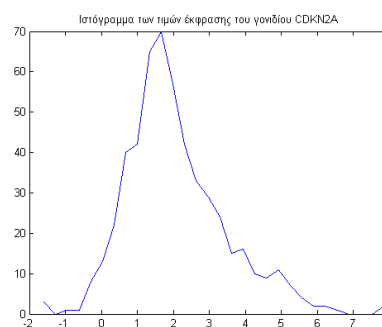
3.3.3 Συνεχείς Μεταβλητές

Οι τιμές γονιδιακής έκφρασης από καρκινικά δείγματα δεν ακολουθούν μια κανονική-Γκαουσιανή κατανομή για μεμονωμένα γονίδια, αλλά ούτε καν για ολόκληρο το πείραμα. Αντίθετα, ακολουθούν μια κατανομή με βαριά ουρά που χαρακτηρίζεται από ασυμμετρία και κύρτωση. Η μη-Γκαουσιανές κατανομές αυτών των δεδομένων, επηρεάζει την ταυτοποίηση των γονιδίων διαφορικά εκφρασμένων, τον λειτουργικό σχολιασμό και πιθανή μοριακή ταξινόμηση [81]. Για αυτό το λόγο εφαρμόζεται λογαριθμικός μετασχηματισμός με βάση το δύο (log base 2 transformation) στα δεδομένα γονιδιακής έκφρασης. Αυτό γίνεται για διάφορους λόγους: σταθεροποιεί τη διακύμανση, συμπυκνώνει το εύρος των δεδομένων και κάνει τα δεδομένα έτσι ώστε να ακολουθούν μια κανονική-Γκαουσιανή κατανομή, κάτι το οποίο επιτρέπει να εφαρμοστεί η στατιστική στα δεδομένα [82].

Τα παραπάνω επιβεβαιώνονται και με το ιστόγραμμα όλων των δεδομένων γονιδιακής έκφρασης που χρησιμοποιήσαμε εμείς (Εικόνα 3.7.1) αλλά και με το ιστόγραμμα των τιμών έκφρασης κάθε γονιδίου ξεχωριστά (Εικόνα 3.7.2 - γονίδιο CDKN2A), τα οποία φαίνεται ότι ακολουθούν μια κανονική-Γκαουσιανή κατανομή.



(3.7.1) Όλα τα γονίδια.



(3.7.2) Γονίδιο CDKN2A.

Εικόνα 3.7 – Ιστογράμματα τιμών έκφρασης.

Εφόσον γνωρίζουμε ότι τα δεδομένα μας ακολουθούν μια Γκαουσιανή κατανομή μπορούμε να τα χρησιμοποιήσουμε ως έχουν για την κατασκευή δομών Bayesian δικτύων με συνεχείς μεταβλητές.

3.4 Εκμάθηση Δομής Δικτύου

Για να γίνει η ανάλυση των δικτύων που επιθυμούμε, πρέπει πρώτα να κατασκευάσουμε τις δομές αυτών. Έχουμε επιλέξει να κατασκευάσουμε δίκτυα με διακριτές και συνεχείς μεταβλητές για κάθε τύπο δειγμάτων (καρκινικά και υγιή) για να μπορέσουμε να έχουμε μια σφαιρική άποψη και να εκμεταλλευτούμε τα διαφορετικά χαρακτηριστικά κάθε δικτύου κατασκευασμένο με διαφορετικές παραμέτρους.

Η εκμάθηση δομής των Bayesian δικτύων μπορεί να διαιρεθεί σε δυο κύριες κατηγο-

ρίες: αλγόριθμοι εκμάθησης που βασίζονται σε εξαρτημένες σχέσεις και αλγόριθμοι εκμάθησης βασισμένοι σε λειτουργίες βαθμολόγησης. Αλγόριθμοι εκμάθησης βασισμένοι στην ανάλυση εξαρτήσεων κατασκευάζουν μία δομή Bayesian δικτύου μέσω της οργάνωσης των υπό συνθήκη εξαρτωμένων σχέσεων μεταξύ κόμβων. Η διαδικασία εκμάθησης τέτοιων αλγορίθμων είναι περισσότερο διαισθητική. Η δοκιμή υπό συνθήκη ανεξαρτησιών και η αναζήτηση της δομής δικτύου μπορεί να διαχωριστεί, αλλά αυτοί οι αλγόριθμοι είναι υπερευαίσθητοι σε σφάλματα κατά τη διαδικασία της δοκιμής των υπό συνθήκη ανεξαρτησιών. Οι αλγόριθμοι εκμάθησης που βασίζονται σε λειτουργίες βαθμολόγησης βλέπουν την εκμάθηση ως την διαδικασία βελτιστοποίησης, και επιλέγουν την καλύτερη δομή δικτύου που μεγιστοποιεί την τιμή βαθμολόγησης. Γενικά δομές δικτύων η εκμάθηση των οποίων γίνεται με τέτοιους αλγόριθμους έχουν μεγάλη ακρίβεια, αλλά οι δομές έχουν την τάση να εμπίπτουν σε τοπική βελτιστοποίηση.

Το δίκτυο στο οποίο έγινε η εκμάθηση ώστε να κατασκευαστούν τελικά οι δομές, αποτελούνταν από 12 αλληλεπιδράσεις οι οποίες φαίνονται στην Εικόνα 3.8.1. Για να καταλήξουμε σε αυτές τις 12 αλληλεπιδράσεις, αναζητήσαμε στην βιβλιογραφία αλληλεπιδράσεις στις οποίες συμμετέχουν τα 82 εξεταζόμενα γονίδια και εμπλέκονται στην νόσο του καρκίνου του μαστού. Σε αυτές τις αλληλεπιδράσεις συμμετέχουν συνολικά 15 γονίδια (Εικόνα 3.8.2).

γονίδιο 1	γονίδιο 2
CDKN2A	BRCA1
CDKN2A	TAT
CDKN2A	FN1
CDKN2A	ACTA1
PPARGC1A	PARK7
EGF	ERBB2
FN1	KRT16
FN1	COMP
FN1	IGHG1
NRG1	ERBB2
FGF18	FGFR3
KRT16	IGHG1

(3.8.1) Οι 12 γνωστές αλληλεπιδράσεις των 82 γονιδίων.

γονίδιο #1	CDKN2A	γονίδιο #9	ERBB2
γονίδιο #2	BRCA1	γονίδιο #10	KRT16
γονίδιο #3	TAT	γονίδιο #11	COMP
γονίδιο #4	FN1	γονίδιο #12	IGHG1
γονίδιο #5	ACTA1	γονίδιο #13	NRG1
γονίδιο #6	PPARGC1A	γονίδιο #14	FGF18
γονίδιο #7	PARK7	γονίδιο #15	FGFR3
γονίδιο #8	EGF		

(3.8.2) Τα 15 γονίδια που συμμετέχουν στις 12 αρχικά γνωστές αλληλεπιδράσεις.

Εικόνα 3.8 – Στοιχεία αρχικού δικτύου.

Για την εκμάθηση των δομών χρησιμοποιήσαμε τον αλγόριθμο K2 συνδυαστικά με μία μέθοδο που χρησιμοποιεί έναν MWST αλγόριθμο. Ο λόγος που χρησιμοποιήθηκε αυτός ο συνδυασμός είναι επειδή έχει αποδειχθεί ότι μπορεί να είναι επωφελής στη μείωση της υπολογιστικής πολυπλοκότητας, χωρίς απώλεια πληροφοριών [83]. Ο αλγόριθμος K2 είναι ένας κλασικός αλγόριθμος εκμάθησης δομής Bayesian δικτύων

που βασίζεται σε λειτουργία βαθμολόγησης και προτάθηκε από τους Cooper και Herskovits. Ο αλγόριθμος αυτός συνδυάζει την Bayesian μέθοδο βαθμολόγησης και την στρατηγική αναζήτησης του αλγόριθμου αναζήτησης hill-climbing. Για να γίνει χρήση του αλγόριθμου K2 απαιτείται να καθοριστεί μια ταξινομημένη σειρά των κόμβων και ο μέγιστος αριθμός γονέων κάθε κόμβου [84, 85].

Ο μέγιστος αριθμός γονέων κάθε κόμβου τέθηκε ίσος με με τον συνολικό αριθμό των κόμβων-γονιδίων καθώς όταν έχουμε να κάνουμε με βιολογικά δίκτυα δεν έχει νόημα να θέσουμε κάποιο περιορισμό ως προς τους γονείς ενός γονιδίου. Ο καθορισμός μιας ταξινομημένης σειράς των κόμβων είναι ένα πιο περίπλοκο και παράλληλα καθοριστικό ζήτημα για τη δημιουργία της δομής του δικτύου. Εμείς κατασκευάσαμε τις δομές των δικτύων χρησιμοποιώντας 2 διαφορετικά ταξινομημένες σειρές των κόμβων. Η πρώτη που χρησιμοποιήθηκε ήταν αυτή που προκύπτει εφαρμόζοντας στους 82 κόμβους μια μέθοδο που χρησιμοποιεί τον αλγόριθμο MWST (Maximum Weight Spanning Tree) [86]. Αυτή η μέθοδος τοποθετεί ένα βάρος σε κάθε ακμή. Το βάρος είναι η διακύμανση του σκορ όταν ένας κόμβος γίνεται γονιός ενός άλλου. Στη συνέχεια ένας κλασικός αλγόριθμος MWST παράγει, από τον πίνακα των βαρών, ένα μη κατευθυνόμενο δέντρο το οποίο να προσανατολιστεί (να γίνει κατευθυνόμενο) δίνοντας του έναν συγκεκριμένο αρχικό κόμβο-ρίζα. Σε αυτό το δέντρο εφαρμόζεται μια τοπολογική ταξινόμηση και από αυτήν προκύπτει μια ταξινομημένη σειρά κόμβων την οποία ονομάσαμε MWST.

Ορισμός: [Δέντρο Επικάλυψης] Ένα Δέντρο Επικάλυψης (Spanning Tree) ενός γράφου n ακμών είναι ένα υποσύνολο $n - 1$ ακμών που σχηματίζουν ένα δέντρο (tree) [87, 88].

Ορισμός: [Δέντρο Επικάλυψης Μέγιστου Βάρους] Ένα Δέντρο Επικάλυψης Μέγιστου Βάρους (Maximum Weight Spanning Tree) είναι ένα spanning tree ενός γράφου με βάρη με μέγιστο βάρος. Μπορεί να υπολογιστεί χρησιμοποιώντας τον αλγόριθμο Kruskal με τις ακμές ταξινομημένες κατά φθίνουσα σειρά [89].

Ορισμός: [Αλγόριθμος Κρουσκάλ] Ο αλγόριθμος του Κρουσκάλ (Kruskal) είναι ένας αλγόριθμος που βρίσκει το spanning tree ενός γράφου με ελάχιστο μήκος. Ταξινομεί τις ακμές ενός γράφου κατά σειρά αυξανόμενου κόστους και στη συνέχεια προσθέτει επανειλημμένα ακμές οι οποίες ενώνουν χωριστά στοιχεία μέχρι ο γράφος να είναι πλήρως συνδεδεμένος [90, 91, 92, 93].

Ορισμός: [Τοπολογική Ταξινόμηση] Τοπολογική ταξινόμηση ή αλλιώς τοπολογική διάταξη ενός Κατευθυνόμενου Άκυκλου Γράφου (ΚΑΓ), ονομάζεται η γραμμική διάταξη των κόμβων, έτσι ώστε κάθε πρόγονος ενός κόμβου v προηγείται του v στη διάταξη. Κάθε ΚΑΓ μπορεί να έχει μία ή περισσότερες τοπολογικές διατάξεις. Πιο αυστηρά, μπορούμε να ορίσουμε την τοπολογική ταξινόμηση ενός ΚΑΓ $G(V, E)$, με V το σύνολο των κόμβων και E το σύνολο των ακμών ως μία γραμμική αλληλουχία των κόμβων V , έτσι ώστε αν $(u, v) \in E$, $\pi(u) < \pi(v)$, όπου $\pi(x)$ η θέση στην οποία βρίσκεται ο κόμβος x στη διάταξη.

Η δεύτερη ταξινόμηση κόμβων που χρησιμοποιήθηκε είναι ένας συνδυασμός της προηγούμενης και της τυχαίας ταξινόμησης. 15 επιλεγμένα γονίδια ταξινομήθηκαν με τη χρήση της μεθόδου MWST και τα υπόλοιπα 67 ακολούθησαν τα 15 ταξινομημένα με μια τυχαία σειρά. Η ταξινόμηση των 67 γονιδίων, παρόλο που έγινε με μία τυχειότητα, ήταν βέλτιστη για όλες τις περιπτώσεις εκμάθησης δομών που χρησιμοποιήθηκαν. Έτσι, βρέθηκε μια ακολουθία που να εξαγάγει τον μέγιστο αριθμό ακμών για όλες, η οποία και χρησιμοποιήθηκε για την εκμάθηση όλων των δομών. Αυτή η ταξινόμηση κόμβων ονομάστηκε *CUSTOM*. Πιο συγκεκριμένα για τα 15 επιλεγμένα γονίδια: αρχικά έχοντας 82 γονίδια, γνωρίζουμε από τη βιβλιογραφία ότι μεταξύ αυτών των γονιδίων υπάρχουν 12 γνωστές αλληλεπιδράσεις (Εικόνα 3.8.1). Σε αυτές τις αλληλεπιδράσεις συμμετέχουν μόνο τα 15 γονίδια από τα 82 (Εικόνα 3.8.2). Αυτά τα 15 γονίδια επιλέχθηκαν λοιπόν για να ταξινομηθούν με την μέθοδο MWST στην *CUSTOM* ταξινόμηση των κόμβων.

3.5 Οπτικοποίηση Bayesian Δικτύων

Αφού κατασκευάσαμε τις δομές των δικτύων, τις οπτικοποιήσαμε ώστε να μπορέσουμε να παρατηρήσουμε κάποιες ιδιότητες τους αλλά και να βρούμε κάποια σημαντικά στοιχεία που ίσως παρουσίαζαν όπως σημαντικοί κόμβοι και υποδίκτυα. Η οπτικοποίηση των δικτύων που εξήχθησαν από την παραπάνω διαδικασία έγινε με τη χρήση του εργαλείου *Cytoscape*.¹ Από την παραπάνω διαδικασία κατασκευής των δικτύων πήραμε την δομή των οκτώ και κατασκευάσαμε επιμέρους δίκτυα. Αρχικά κατασκευάσαμε τέσσερα δίκτυα. Έτσι, έγινε ένωση των δικτύων με διακριτές μεταβλητές και καρκινικά δείγματα (**ΚΔ**) και ένωση αυτών με συνεχείς μεταβλητές και καρκινικά δείγματα (**ΚΣ**). Αντίστοιχα δίκτυα κατασκευάστηκαν για τα υγιή δείγματα όπου έγινε ένωση των δικτύων με διακριτές μεταβλητές (**ΥΔ**) και ένωση αυτών με συνεχείς μεταβλητές (**ΥΣ**). Για την ανάλυση χρησιμοποιήθηκαν επίσης δυο πιο γενικά και μεγάλα δίκτυα, αυτά της ένωσης όλων των δικτύων για τα καρκινικά δείγματα (**ΚΕ**) και της ένωσης αυτών για τα υγιή δείγματα (**ΥΕ**). Ο λόγος που η μελέτη και η ανάλυση των δικτύων έγινε διαχωρίζοντας καρκινικά και υγιή δείγματα είναι για να εντοπιστούν οι διαφορές τους μέσω στατιστικής και ιδιοτήτων, να εξηγηθούν αυτές και να δούμε αν υπάρχει η δυνατότητα εξαγωγής συμπερασμάτων μέσω αυτών.

3.6 Ανάλυση Δικτύων

Σε αυτήν την ενότητα γίνεται ανάλυση των δικτύων που προέκυψαν, προκειμένου να διαπιστωθεί αν αυτά εμφανίζουν κάποιες ιδιότητες που θα μας βοηθήσουν στην εξαγωγή των συμπερασμάτων. Βασικές ιδιότητες που αναζητούμε είναι αυτές του φαινομένου Μικρού Κόσμου όπως και ομοιότητες με τα δίκτυα Ελεύθερης Κλίμακας.

¹<http://www.cytoscape.org>

Επίσης εξετάστηκαν βασικές ιδιότητες των δικτύων όπως ο μέσος βαθμός, ο συντελεστής ομαδοποίησης, η διάμετρος, το μέσο μήκος μονοπατιού και άλλα.

3.6.1 Φαινόμενο Μικρού Κόσμου

Για να πούμε ότι ένα δίκτυο εμφανίζει το φαινόμενο του Μικρού Κόσμου θα πρέπει το μέσο μήκος μονοπατιού των κόμβων να είναι μικρό. Πιο συγκεκριμένα η τιμή του θα πρέπει να είναι περίπου ίση με $O(\log n)$ όπου n ο αριθμός των κόμβων. Επίσης θα πρέπει να εμφανίζει υψηλό συντελεστή ομαδοποίησης συγκριτικά με ένα ισοδύναμο τυχαίο γράφο (ίδιο αριθμό κόμβων και ακμών) [62].

3.6.2 Δίκτυα Ελεύθερης Κλίμακας

Ένα δίκτυο για να θεωρείται Ελεύθερης Κλίμακας πρέπει να η κατανομή βαθμού αυτού να ακολουθεί μια κατανομή με νόμο δύναμης. Επίσης έχει παρατηρηθεί ότι και η κατανομή της ενδιάμεσης κεντρικότητας ακολουθεί μια κατανομή με νόμο δύναμης. Τα Bayesian δίκτυα είναι κατευθυνόμενα, οπότε υπάρχει μέσα και έξω βαθμός δικτύου όπου εξετάστηκαν ξεχωριστά [70].

3.6.3 Άλλες Ιδιότητες

Εκτός από αυτές τις δυο παραπάνω ιδιότητες-φαινόμενα εξετάστηκαν και άλλες ιδιότητες όπως το μέσο μήκος μονοπατιού, ο συντελεστής ομαδοποίησης, ο μέσος μέσα και έξω βαθμός και η ενδιάμεση κεντρικότητα που μας δίνουν κάποια επιπλέον στοιχεία για την δομή του δικτύου. Αυτές παραθέτονται αναλυτικά στο Κεφάλαιο 4.

3.7 Σύγκριση Δικτύων

Η αναζήτηση τυχών διαφορών ανάμεσα στις δομές δικτύων καρκινικά και υγιή δείγματα αλλά και ανάμεσα στα υποδίκτυα τους μας ώθησαν ώστε να τα συγκρίνουμε. Η σύγκριση των δικτύων έγινε μετρώντας και αξιολογώντας τις κοινές αλληλεπιδράσεις μεταξύ αυτών αλλά και σύμφωνα με μια μέθοδο σύγκρισης που προτάθηκε από τους F.Gomez και N.Diaz. Η μέθοδος αυτή παρουσιάζεται παρακάτω σε 3 βήματα [94]:

Ας υποθέσουμε ότι $G_1 = \{N_1, E_1\}$ και $G_2 = \{N_2, E_2\}$ είναι δύο γράφοι, όπου (N_1, N_2) αντιπροσωπεύουν τους κόμβους των γράφων και (E_1, E_2) τις ακμές.

Ορισμός: [Επίπεδο] Ας υποθέσουμε ότι $G = \{N, E\}$ είναι ένας γράφος και δυο κόμβοι $g_a, g_b \in N$. Το επίπεδο της σχέσης ανάμεσα στους (g_a, g_b) υπολογίζεται ως ο αριθμός των ακμών ανάμεσα στους κόμβους g_a και g_b στο G .

Ορισμός: [Χτυπήματα (Hits) στο επίπεδο l] Ο αριθμός των ακμών, όπου το επίπεδο ανάμεσα στους κόμβους που συνδεόνται κατευθείαν στο G_1 , είναι l στο G_2 .

1. Αθροιστικά χτυπήματα στο επίπεδο n , H_n , μπορεί να οριστεί ως το σταθμισμένο άθροισμα των ορθά συναγμένων ακμών στο επίπεδο n στο G_1 σύμφωνα με την πληροφορία που παρουσιάζεται στο G_2 . Θεωρείστε:

$$H_n(G_1, G_2) = \sum_{l=1}^n \frac{Hit_l}{l} \quad (3.5)$$

όπου το H_n ορίζει το άθροισμα των ακμών που ορθά συνάχθηκαν σταθμισμένα από τη σχέση τους στο δίκτυο με απόσταση (επίπεδο) n .

2. Αθροιστικές αστοχίες στο επίπεδο n , F_n , μπορεί να οριστεί ως το σταθμισμένο άθροισμα των λανθασμένα συναγόμενων ακμών στο επίπεδο n στο G_1 σύμφωνα με την πληροφορία που παρουσιάζεται στο G_2 . Θεωρείστε:

$$F_n(G_1, G_2) = \|E_l\| - H_n(G_1, G_2) \quad (3.6)$$

όπου $\|E_l\|$ είναι ο αριθμός των ακμών στο G_1 . Έτσι, το F_n ορίζει το άθροισμα των ακμών που δεν συνάχθηκαν ορθά σταθμισμένα από τη σχέση τους στο δίκτυο με απόσταση (επίπεδο) n .

3. Η εγκυρότητα του γράφου G_1 σύμφωνα με το επίπεδο του G_2 n , V_n ορίζεται ως το ποσοστό των ορθά συναγόμενων ακμών στο επίπεδο n στο G_1 .

$$V_n(G_1, G_2) = \frac{H_n(G_1, G_2)}{H_n(G_1, G_2) + F_n(G_1, G_2)} = \frac{H_n(G_1, G_2)}{\|E_l\|} \quad (3.7)$$

Αυτή η μέτρηση κυμαίνεται ανάμεσα στο 0 και στο 1. Όσο υψηλότερη είναι αυτή η τιμή τόσο πιο ποιοτικό είναι το δίκτυο σε σύγκριση με το άλλο. Αυτό μετράει τον βαθμό ορθότητας του G_1 σε σχέση με το G_2 [94].

Θέλοντας να μελετήσουμε τις διαφορές των δυο δικτύων, δημιουργήσαμε ένα δίκτυο των διαφορών των δύο δικτύων ΚΕ και ΥΕ. Στο δίκτυο που προέκυψε εφαρμόστηκε ο αλγόριθμος MCODE (Υποενότητα 3.8.1) και υπολογίστηκαν οι μέσοι όροι των βαθμών, των ενδιάμεσων κεντρικοτήτων και των κεντρικοτήτων εγγύτητας κάθε συμπλέγματος που προέκυψε. Επίσης, υπολογίστηκαν τα ίδια τοπολογικά χαρακτηριστικά για τα βιολογικά μονοπάτια που εντοπίστηκαν μέσα στα συμπλέγματα (μετά από βιολογική ανάλυση - Ενότητα 4.8) [95].

3.8 Ομαδοποίηση και Ανάλυση Ομάδων

Μετά την ανάλυση ολόκληρων των δικτύων, θεωρήσαμε απαραίτητο να μελετήσουμε και υποδίκτυα που θα προκύψουν μέσα από συγκεκριμένους αλγόριθμους. Η μελέτη τους είναι πολύ σημαντική καθώς πραγματικά δίκτυα τείνουν να δημιουργούν τέτοιες ομάδες και συνήθως ο σχηματισμός τους είναι πολύ σημαντικός για την εκάστοτε εξεταζόμενη περίπτωση(τύπος καρκίνου κτλ). Η ομαδοποίηση είναι μια διαδικασία

δημιουργίας ομάδων (*clusters*) από σύνολα αντικειμένων με τέτοιο τρόπο ώστε αντικείμενα της ίδιας ομάδας να είναι περισσότερο όμοια από αυτά μιας άλλης ομάδας. Η ανάλυση ομάδων από μόνη της δεν είναι ένας συγκεκριμένος αλγόριθμος, αλλά μια γενικότερη διαδικασία. Υπάρχουν διάφοροι αλγόριθμοι που διαφέρουν σημαντικά στην αντίληψη του τι αποτελεί μία ομάδα και μπορούν αποτελεσματικά να βρεθούν. Δημοφιλείς έννοιες για ομάδες περιλαμβάνουν ομάδες με μικρές αποστάσεις κατά μήκος των αντικειμένων τους, πυκνές περιοχές του χώρου των δεδομένων, χρονικά διαστήματα ή ιδιαίτερες στατιστικές κατανομές. Η ομαδοποίηση μπορεί ως εκ τούτου να διατυπωθεί ως ένα πολυδιάστατο πρόβλημα βελτιστοποίησης. Οι απαραίτητες ρυθμίσεις του αλγόριθμου ομαδοποίησης και των παραμέτρων (η συνάρτηση απόστασης στη χρήση ενός κατωφλίου πυκνότητας ή τον αριθμό των αναμενόμενων ομάδων) εξαρτώνται από το συγκεκριμένο σύνολο δεδομένων και τον σκοπό της χρήσης των αποτελεσμάτων [88].

3.8.1 Ομαδοποίηση με τον αλγόριθμο MCODE

Το εργαλείο MCODE χρησιμοποιεί ένα αλγόριθμο συντελεστή ομαδοποίησης για τον εντοπισμό μοριακών συμπλεγμάτων σε ένα μεγάλο δίκτυο πρωτεϊνικών αλληλεπιδράσεων που προέρχονται από ετερογενείς πειραματικές πηγές. Η ομαδοποίηση γονιδίων σε ένα βιολογικό δίκτυο έχει σκοπό την πρόβλεψη μοριακών συμπλεγμάτων από δεδομένα πρωτεϊνικών αλληλεπιδράσεων. Αυτό είναι πολύ σημαντικό επειδή παρέχει καλύτερο σχολιασμό από άλλες μεθόδους. Δεδομένου ότι οι υπομονάδες ενός μοριακού συμπλέγματος λειτουργούν ως προς τον ίδιο βιολογικό στόχο, η πρόβλεψη μιας άγνωστης πρωτεΐνης ως μέρος μιας ομάδας επιτρέπει επίσης αυξημένη εμπιστοσύνη στον σχολιασμό αυτής της πρωτεΐνης. Το MCODE επιπλέον οπτικοποιεί τις εξαγόμενες ομάδες [96].

Ο αλγόριθμος MCODE λειτουργεί σε τρία στάδια. Τοποθέτηση βαρών στις κορυφές, πρόβλεψη σύμπλοκων και προαιρετικά μετα-επεξεργασία για να φιλτράρει ή να προσθέσει πρωτεΐνες στο προκύπτοντα συμπλέγματα με ορισμένα κριτήρια συνδεσιμότητας [96].

Η χρήση αυτής της αναπαράστασης γράφων ενός βιολογικού συστήματος επιτρέπει θεωρητικές μεθόδους γράφων να εφαρμοστούν για να βοηθήσουν στην ανάλυση και επίλυση βιολογικών ζητημάτων.

Αλγόριθμοι για την εύρεση ομάδων, ή τοπικά πυκνές περιοχές, ενός γράφου είναι ένα εν εξελίξει ερευνητικό θέμα στη επιστήμη των υπολογιστών και είναι συχνά βασισμένα στην θεωρία network flow/minimum cut [97, 98] ή πιο πρόσφατα στην φασματική ομαδοποίηση [99]. Για να βρεθούν τοπικά πυκνές περιοχές σε ένα γράφο, το MCODE χρησιμοποιεί ένα σύστημα με βάρη στις κορυφές βασισμένο στον συντελεστή ομαδοποίησης, C_i , ο οποίος μετρά κατά πόσο δημιουργούν κλίκες οι γείτονες ενός κόμβου [62].

$$C_i = 2n/k_i(k_i - 1) \quad (3.8)$$

όπου k_i είναι μέγεθος του γειτονικού κόμβου του i και n είναι ο αριθμός των κόμβων στην γειτονιά.

Πρώτο στάδιο αλγόριθμου - Τοποθέτηση Βαρών στις Ακμές Το πρώτο στάδιο του αλγόριθμου MCODE τοποθετεί βάρη σε όλες τις κορυφές με βάση την τοπική πυκνότητά τους, κάνοντας χρήση του υψηλότερου k -core της γειτονιάς του κόμβου. Ένα k -core είναι ένας γράφος με ελάχιστο βαθμό k (γράφος G , για όλα τα v στο G , $\deg(v) \geq k$). Το υψηλότερο k -core ενός γράφου είναι ο κεντρικά πιο πυκνά συνδεδεμένος υπογράφος. Ορίζουμε τον όρο core-συντελεστής ομαδοποίησης ενός κόμβου v , να είναι η πυκνότητα του υψηλότερου k -core της άμεσης γειτονιάς του v (κόμβοι που συνδέονται κατευθείαν στον v) συμπεριλαμβανομένου του v (το C_i δεν συμπεριλαμβάνει το v). Ο core-συντελεστής ομαδοποίησης χρησιμοποιείται εδώ αντί του συντελεστή ομαδοποίησης, επειδή ενισχύει την τοποθέτηση βαρών των περιοχών των γράφων με πολλές διασυνδέσεις. Παράλληλα αφαιρεί τις λιγότερο συνδεδεμένες κορυφές που είναι συνήθως μέρος ενός δικτύου βιομοριακών αλληλεπιδράσεων, τα οποία είναι γνωστό ότι είναι ελεύθερης κλίμακας. Το τελικό βάρος που δίνεται σε μία κορυφή είναι το προϊόν του core-συντελεστή ομαδοποίησης κορυφής και του υψηλότερου k -core επιπέδου, k_{max} , των άμεσων γειτονιών της κορυφής.

Δεύτερο στάδιο αλγόριθμου - Πρόβλεψη Σύμπλοκων Το δεύτερο στάδιο παίρνει ως είσοδο τον γράφο με τα βάρη στις κορυφές, δημιουργεί ένα σύμπλεγμα με τις κορυφές με τα υψηλότερα βάρη και αναδρομικά κινείται προς τα έξω από τον αρχικό κόμβο, και συμπεριλαμβάνει στο σύμπλεγμα τις κορυφές των οποίων το βάρος είναι μεγαλύτερο από ένα δοθέν κατώφλι, το οποίο είναι ένα δεδομένο ποσοστό μακριά από το βάρος της αρχικής κορυφής. Αυτή είναι η παράμετρος ποσοστό βάρους κορυφής (Vertex Weight Percentage - VWP). Εάν μια κορυφή έχει συμπεριληφθεί, οι γείτονές του ελέγχονται αναδρομικά κατά τον ίδιο τρόπο για να δει αν είναι μέρος αυτού του συμπλέγματος. Μια κορυφή δεν ελέγχεται πάνω από μια φορά, δεδομένου ότι τα συμπλέγματα δεν μπορούν να επικαλύπτονται σε αυτό το στάδιο του αλγόριθμου. Αυτή η διαδικασία σταματά όταν δεν μπορούν να προστεθούν άλλες κορυφές στο σύμπλεγμα με βάση το δοθέν κατώφλι. Η διαδικασία επαναλαμβάνεται για τα επόμενα πιο υψηλά βεβαρημένα κορυφή στο δίκτυο που δεν έχει εξεταστεί. Κατά αυτό τον τρόπο εντοπίζονται οι πιο πυκνές περιοχές στο δίκτυο. Το κατώφλι του βάρους των κορυφών καθορίζει την πυκνότητα του κάθε συμπλέγματος.

Τρίτο στάδιο αλγόριθμου - Μετά Επεξεργασία Το τρίτο στάδιο είναι η μετά επεξεργασία των σύμπλοκων. Τα συμπλέγματα φιλτράρονται αν δεν περιέχουν τουλάχιστον ένα 2-core (γράφος ελάχιστου βαθμού 2). Επίσης μπορεί να μεγαλώσει το μέγεθος των συμπλεγμάτων σύμφωνα με μια παράμετρο [96].

3.8.2 Ομαδοποίηση με τον αλγόριθμο jActiveModules

Το εργαλείο jActiveModules χρησιμοποιεί μια γενική μέθοδο για την εύρεση ενεργών υποδικτύων, δηλαδή συνδεδεμένα σύνολα γονιδίων με απροσδόκητα υψηλά επίπεδα διαφορετικής έκφρασης. Όταν τα επίπεδα έκφρασης έχουν παρατηρηθεί σε πολλαπλές συνθήκες, αυτή η μέθοδος καθορίζει ποιες καταστάσεις επιδρούν σημαντικά στην γονιδιακή έκφραση κάθε ενεργού υποδικτύου. Για να επιτευχθεί αυτός ο στόχος, έχει υλοποιηθεί ένα στατιστικό σύστημα για σκορ το οποίο καταγράφει τη ποσότητα των αλλαγών της γονιδιακής έκφρασης σε ένα δοθέν υποδίκτυο. Αυτή η μέθοδος χρησιμοποιεί έναν αλγόριθμο αναζήτησης, βασισμένο σε simulated annealing, για να εντοπιστούν τα υποδίκτυα με το υψηλότερο σκορ. Ακολουθεί αναλυτικά η μέθοδος [102]:

Υπολογισμός z-σκορ Σε αυτό το στάδιο γίνεται η χρήση των p-values p_i που αντιπροσωπεύουν την σημαντικότητα της αλλαγής της γονιδιακής έκφρασης για κάθε γονίδιο i . Κάθε p_i μετατρέπεται σε ένα z-σκορ $z_i = \Phi^{-1}(1 - p_i)$, όπου Φ^{-1} είναι η αντίστροφη κανονική αθροιστική συνάρτηση κατανομής. Έτσι σε τυχαία δεδομένα, τα p-values κατανέμονται ομοιόμορφα από το 0 στο 1 και τα z-σκορ ακολουθούν μια τυπική κανονική, με τα μικρότερα p-values να αντιστοιχούν στα μεγαλύτερα z-σκορ. Για να παραχθεί ένα συνολικό z-σκορ z_A για ένα υποδίκτυο A k γονιδίων, αθροίζουμε τα z_i για όλα τα γονίδια στο υποδίκτυο:

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \quad (3.9)$$

Βαθμονόμηση z Για να καταγράψουμε σωστά την σύνδεση ανάμεσα σε έκφραση και τοπολογία δικτύου, πρέπει να ορίσουμε πότε το σκορ z_A ενός υποδικτύου είναι υψηλότερο από το αναμενόμενο σε σχέση με ένα τυχαίο σύνολο γονιδίων. Κάνοντας χρήση της Monte Carlo προσέγγισης σε ένα τυχαίο σύνολο γονιδίων μεγέθους k υπολογίζουμε τα z_A τους, τον μέσο όρο τους μ_k και την τυπική απόκλιση σ_k για κάθε k . Έτσι τα σωστά σκορ των υποδικτύων είναι:

$$s_A = \frac{z_A - \mu_k}{\sigma_k} \quad (3.10)$$

Χρησιμοποιώντας αυτήν την διόρθωση, τα σκορ s_A τυχαίων υποδικτύων θα έχουν σίγουρα $\mu = 0$ και $\sigma = 1$.

Σκορ για πολλαπλές καταστάσεις Το σύστημα για τα σκορ μπορεί να επεκταθεί για αλλαγές σε δεδομένα γονιδιακής έκφρασης μετρημένες σε πολλαπλές καταστάσεις. Σε αυτήν την περίπτωση έχουμε ένα πίνακα με p-values (γονίδια έναντι καταστάσεων) στα οποία αντιστοιχούμε z-σκορ. Κάνοντας χρήση της 3.9 παίρνουμε m διαφορετικά αθροιστικά σκορ (z_{A1}, \dots, z_{Am}) . Αυτά ταξινομούνται κατά φθίνουσα σειρά. Στη συνέχεια υπολογίζουμε την σημαντικότητα $r_{A(j)}$ του j -οστού υψηλότερου σκορ χρησιμοποιώντας μια binomial order statistic [100] ως ακολούθως: Ας υποθέσουμε ότι το $P_z = 1 -$

$\Phi(z_{A(j)})$ αντιπροσωπεύει την πιθανότητα ότι κάθε κατάσταση έχει ένα z-σκορ πάνω από $z_{A(j)}$. Τότε:

$$p_{A(j)} = \sum_{h=j}^m (P_z)^h (1 - P_z)^{m-h} \quad (3.11)$$

Κάνουμε χρήση της αντιστροφής κανονικής αθροιστικής συνάρτησης κατανομής $r_{A(j)} = \Phi^{-1}(1 - p_{A(j)})$ για να τα μετατρέψουμε πίσω σε ένα τυπικό κανονικό z-σκορ. Το μέγιστο από αυτά είναι το νέο σκορ r_A^{max} του υποδικτύου:

$$r_A^{max} = \max_j(r_{A(j)}). \quad (3.12)$$

Για την βαθμονόμηση ακολουθείται η ίδια διαδικασία όπως παρουσιάστηκε παραπάνω [102].

Αναζήτηση των υποδικτύων με το μεγαλύτερο σκορ Η παραπάνω μέθοδος μας επιτρέπει να τοποθετούμε ένα σκορ σε ένα δοθέν υποδίκτυο, αλλά πρέπει επίσης να βρούμε το υποδίκτυο ή τα υποδίκτυα με το μεγαλύτερο σκορ σε ένα ολόκληρο δίκτυο μοριακών αλληλεπιδράσεων. Επειδή το πρόβλημα της αναζήτησης του συνδεδεμένου υπογράφου με το μέγιστο σκορ είναι NP-δύσκολο, υλοποιήθηκε μια προσέγγιση βασισμένη σε Προσομειωμένη Ανόπτηση (Simulated Annealing) [101]. Στην πράξη, αυτή η προσέγγιση δεν εγγυάται ότι θα βρει το υποδίκτυο με το μέγιστο σκορ. Όμως, όλα τα υποδίκτυα με μεγάλο σκορ έχουν μεγάλο βιολογικό ενδιαφέρον ανεξάρτητα αν είναι μέγιστο το σκορ τους.

Στον αλγόριθμο που ακολουθεί, συσχετίζουμε μια ενεργό/ανενεργό κατάσταση με κάθε κόμβο. Με G_W συμβολίζουμε τον εξεταζόμενο (working) υπογράφο του γράφου G που επάγεται από τους ενεργούς κόμβους. Σε κάθε επανάληψη i , ορίζουμε s_i ως το σκορ (s_A από παραπάνω) του στοιχείου με το υψηλότερο σκορ στο G_W .

Ακολουθεί ο ψευδοκώδικας του αλγορίθμου:

Είσοδος: Ένας γράφος $G = (V, E)$ μοριακών αλληλεπιδράσεων, ένας αριθμός επαναλήψεων, και μία συνάρτηση θερμοκρασίας T_i η οποία μειώνεται γεωμετρικά από T_{start} σε T_{end} .

Έξοδος: Ένα υποδίκτυο G_W του G

1. Αρχικοποίηση G_W θέτοντας κάθε $u \in V$ ως ενεργό/ανενεργό με ίση πιθανότητα.
2. FOR $i = 1$ TO n DO
3. Διάλεξε τυχαία ένα κόμβο $u \in V$ και άλλαξε την κατάσταση του
4. Υπολόγισε το σκορ s_i για το εξεταζόμενο υποδίκτυο G_W
5. IF $(s_i > s_{i-1})$, συνέχισε να εναλλάσσεις το u
6. ELSE συνέχισε να εναλλάσσεις το u με πιθανότητα $p = e^{(s_i > s_{i-1})/T_i}$
7. Έξοδος: G_W και οι ενότητες A με το υψηλότερο σκορ

Τέλος, μειώθηκε η θερμοκρασία σε $T = 0$ μέχρι όλες να εξεταστούν όλες οι παρακείμενες πιθανότητες και το σκορ έχει γίνει τοπικά μέγιστο. Κατά τη λήξη της ανόπτησης (annealing), το υποδίκτυο A αντιπροσωπεύει ένα σηματοδοτικό ή ρυθμιστικό σύμπλεγμα με μεγάλο βιολογικό ενδιαφέρον [102].

Ευρετική μέθοδος για τη βελτίωση της ανόπτησης Η ανόπτηση επεκτάθηκε έτσι ώστε να γίνεται αναζήτηση για M υποδίκτυα ταυτόχρονα. Μια επιπλέον ευρετική μέθοδος αυξάνει την αποδοτικότητα της ανόπτησης σε ένα δίκτυο με πολλούς κεντρικούς κόμβους (Hubs). Η Προσομειωμένη Ανόπτηση δεν αποδίδει καλά σε δίκτυα με Hubs, καθώς προσθέτοντας ένα hub στο G_W αμέσως δημιουργεί μια μεγάλη ενότητα ενσωματώνοντας όλους τους κόμβους γείτονες του Hub. Εκτός εάν όλοι οι γειτονικοί κόμβοι με χαμηλό z-score είναι ανενεργοί, η ενότητα θα έχει και αυτό χαμηλό σκορ. Αυτό το πρόβλημα διορθώθηκε αλλάζοντας το βήμα 3 του αλγορίθμου: όταν προσθέτεται ένας κόμβος με βαθμό μεγαλύτερο από μια παράμετρο d_{min} καθορισμένη από το χρήστη, ταυτόχρονα αφαιρούνται όλοι οι γείτονες που δεν είναι στην ενότητα με το υψηλότερο σκορ.

3.9 Σύγκριση Υποδικτύων

Η σύγκριση των υποδικτύων έγινε ανάμεσα σε αυτά που προέκυψαν από τα διαφορετικά είδη δειγμάτων (καρκινικά και υγιή) για κάθε αλγόριθμο ομαδοποίησης ξεχωριστά. Ο λόγος που έγινε αυτό ήταν για να δούμε αν από τα δύο δίκτυα ΚΕ και ΥΕ προέκυψαν υποδίκτυα με κοινές λειτουργίες, αλληλεπιδράσεις και μονοπάτια. Η σύγκρισή τους έγινε σύμφωνα με την μέθοδο που έγινε η σύγκριση των δικτύων. Επίσης έγινε μια αναζήτηση των κοινών αλληλεπιδράσεων και γονιδίων όλων των υποδικτύων.

3.10 Αναζήτηση Κεντρικών Κόμβων

Κεντρικοί κόμβοι σε ένα γράφο ή αλλιώς *Hubs* είναι αυτοί που έχουν τον υψηλότερο βαθμό. Ο λόγος που αναζητήσαμε κεντρικούς-σημαντικούς κόμβους στα δίκτυα μας είναι επειδή τα πραγματικά δίκτυα τείνουν να έχουν τέτοιους κόμβους και συνήθως παίζουν καθοριστικό ρόλο σε συγκεκριμένες διεργασίες. Οι κόμβοι αυτοί είναι πολύ σημαντικοί για ένα γράφο ενώ φαίνεται να είναι εξίσου βιολογικά σημαντικά τα αντίστοιχα γονίδια. Προκειμένου να το διαπιστώσουμε προσπαθήσαμε να βρούμε τους hubs στα δικά μας δίκτυα. Ο βαθμός ενός κόμβου-πρωτεΐνης μετράει την τοπική συνδετικότητα του. Θέλοντας να δώσουμε στους hubs ένα πιο σημαντικό ρόλο σε όλο το δίκτυο και όχι μόνο τοπικά, χρησιμοποιήσαμε και την ενδιάμεση κεντρικότητα κόμβου στην διαδικασία αναζήτησης τους. Η ενδιάμεση κεντρικότητα κόμβου μετράει την καθολική κεντρικότητα του κόμβου και καθορίζει την κεντρικότητα μιας πρωτεΐνης σε ένα βιολογικό δίκτυο βάσει του συνολικού αριθμού των κοντύτερων μονοπατιών που διέρχονται από αυτήν την πρωτεΐνη.

Μέθοδος

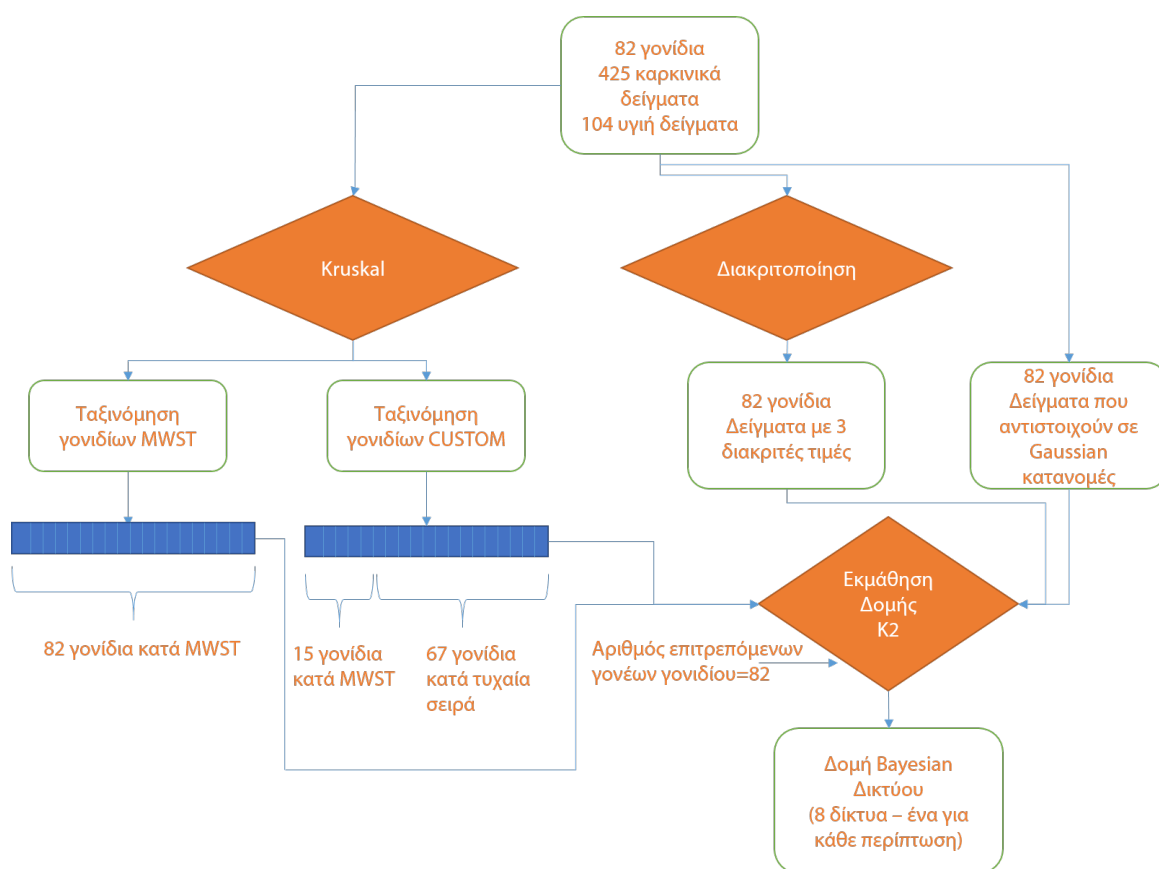
Η μέθοδος που ακολουθήσαμε φαίνεται παρακάτω:

- Κατασκευή ιστογράμματος των βαθμών των κόμβων και των ενδιάμεσων κεντρικότητων των κόμβων του δικτύου.
- Κατασκευή αθροιστικών διαγραμμάτων των βαθμών των κόμβων και των ενδιάμεσων κεντρικότητων των κόμβων του δικτύου.
- Εύρεση του σημείου όπου το αθροιστικό διάγραμμα των βαθμών αρχίζει να γίνεται ευθεία. Αυτό το σημείο καλείται ως η ελάχιστη τιμή του βαθμού όπου ένας κόμβος θεωρείται σημαντικός.
- Αντίστοιχα για την ενδιάμεση κεντρικότητα των κόμβων.
- Η τομή των κόμβων με τους υψηλότερους βαθμούς και με τις υψηλότερες ενδιάμεσες κεντρικότητες θεωρούνται τελικώς hubs.

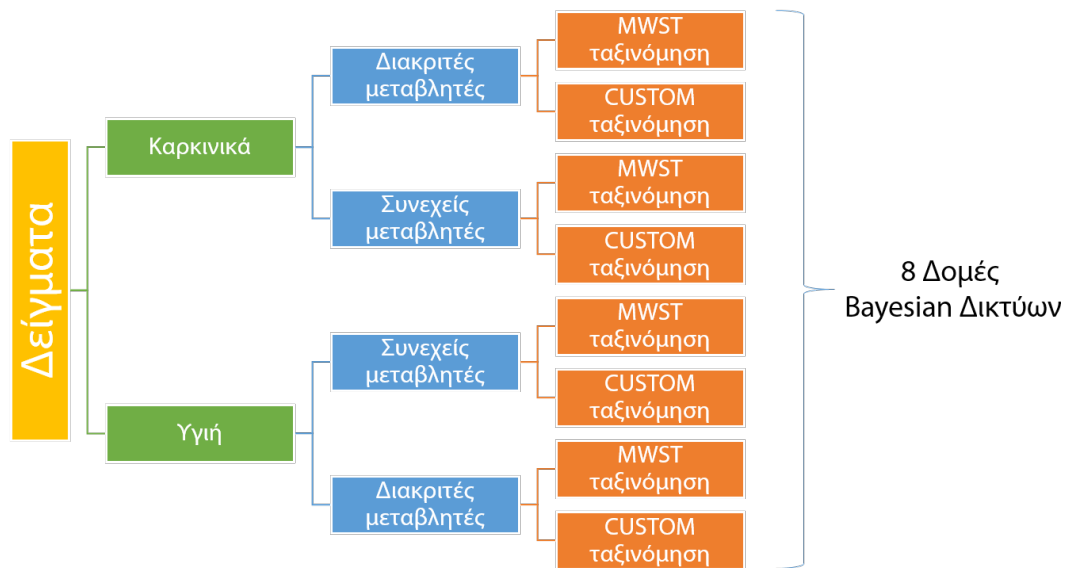
3.11 Προτεινόμενη Μεθοδολογία

Προκειμένου να φτάσουμε στους στόχους μας ακολουθήσαμε μια συγκεκριμένη διαδικασία. Αρχικά έχοντας ως δεδομένα μια γονιδιακή υπογραφή, η οποία προέκυψε μετά από μια διαδικασία επεξεργασίας των δεδομένων μιας μικροσυστοιχίας (SAM και ταξινόμηση), προσπαθήσαμε να κατασκευάσουμε την δομή Bayesian δικτύων. Σε αυτά τα δίκτυα οι σχέσεις μεταξύ των κόμβων αναπαρίστανται από δεσμευμένες κατανομές πιθανότητας οι οποίες μπορεί να είναι είτε διακριτές είτε συνεχείς (Γκαουσιανές). Για να γίνει η κατασκευή των δομών πρέπει να γίνει η επιλογή του τύπου αναπαράστασης των μεταβλητών που θα χρησιμοποιηθούν. Εμείς κατασκευάσαμε δίκτυα με δύο διαφορετικούς τύπους αναπαράστασης μεταβλητών, διακριτές και συνεχείς. Στην πρώτη περίπτωση έπρεπε να γίνει διακριτοποίηση των τιμών έκφρασης για να χαρακτηριστούν τα γονίδια υποεκφρασμένα, υπερεκφραμένα ή κανονικά. Η διακριτοποίηση έγινε με την βοήθεια των ιστογραμμάτων των means των 900 πιο διαχωρίσιμων γονιδίων. Στην δεύτερη περίπτωση οι τιμές έκφρασης χρησιμοποιήθηκαν ως έχουν καθώς τα δεδομένα μας έχουν υποστεί λογαριθμικό μετασχηματισμό με βάση το δύο, κάτι το οποίο σημαίνει ότι ακολουθούν μια κανονική-Γκαουσιανή κατανομή. Η δομή των δικτύων κατασκευάστηκε με τη χρήση του K2 αλγορίθμου σε συνδυασμό με μία μέθοδο που χρησιμοποιεί έναν MWST αλγόριθμο. Ο αλγόριθμος K2 απαιτεί τον αριθμό των γονέων που μπορεί να έχει κάθε κόμβος ο οποίος τέθηκε ίσος με τον συνολικό αριθμό των γονιδίων της υπογραφής καθώς δεν υπάρχει τέτοιος περιορισμός στη φύση. Επίσης ο αλγόριθμος αυτός απαιτεί να γνωρίζει την σειρά των κόμβων. Οι ταξινομημένες σειρές των κόμβων που χρησιμοποιήσαμε ήταν δύο. Η πρώτη ταξινομημένη σειρά κόμβων που δόθηκε ήταν αυτή που προκύπτει από την εφαρμογή ενός MWST αλγορίθμου για την παραγωγή ενός δέντρου και τη χρήση της τοπολογικής ταξινόμησης για την εξαγωγή μιας ταξινομημένης σειράς κόμβων. Η

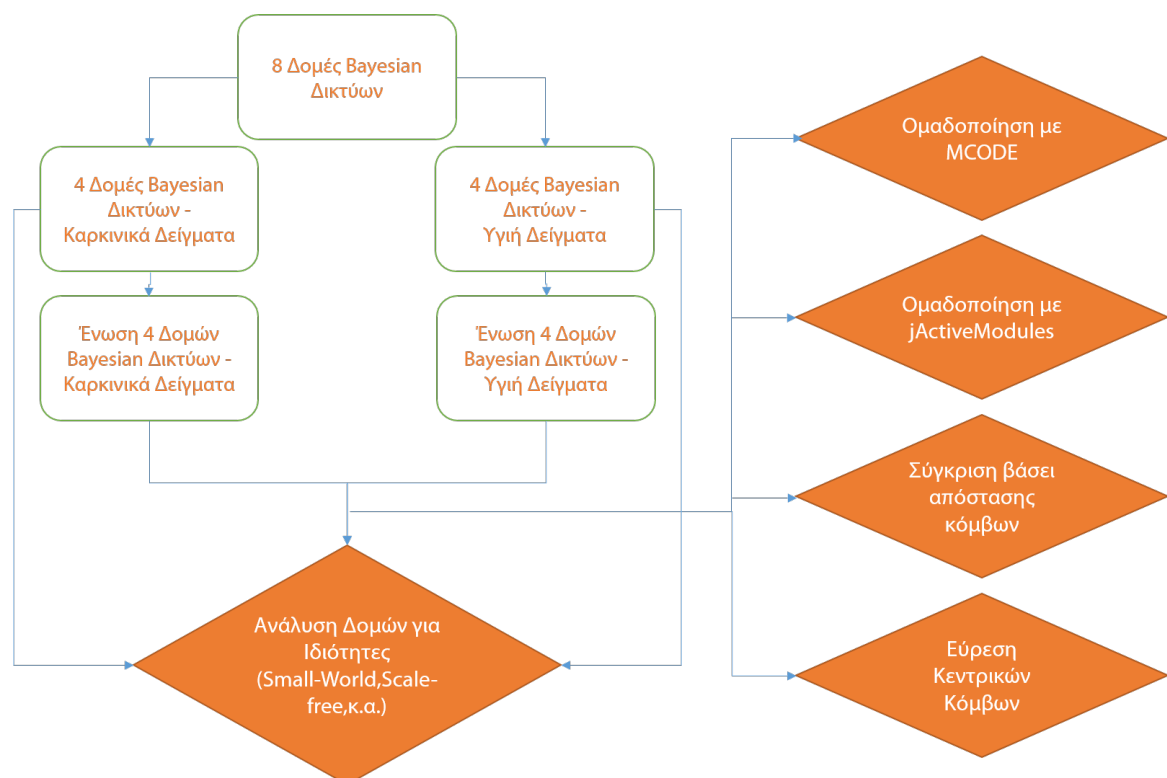
δεύτερη ταξινομημένη σειρά που χρησιμοποιήσαμε συνδυάζει την προηγούμενη και μια τυχαία. Η κατασκευή της δομής των δικτύων έγινε ξεχωριστά για τα καρκινικά και τα υγιή δείγματα. Έτσι προέκυψαν 4 δίκτυα για τα καρκινικά και 4 δίκτυα για τα υγιή δείγματα. Αυτά στην συνέχεια μελετήθηκαν ξεχωριστά αλλά και ως ενώσεις ως προς διάφορες ιδιότητες δικτύων. Επίσης έγινε οπτικοποίηση αυτών και ομαδοποίηση με διάφορους αλγορίθμους ομαδοποίησης όπως ο MCODE και ο jActiveModules. Επιπλέον αναζητήσαμε σημαντικούς κόμβους στα δίκτυα και συγκρίναμε τα δίκτυα αλλά και τα υποδίκτυα σύμφωνα με τις ακμές-αλληλεπιδράσεις τους αλλά και μια μέθοδο που μετράει τη διαφορετικότητα ανάμεσα στα δύο δίκτυα σε ένα συγκεκριμένο επίπεδο απόστασης κόμβων.



Εικόνα 3.9 – Διάγραμμα ροής διαδικασίας κατασκευής δικτύων.



Εικόνα 3.10 – Οι δομές των οκτώ Bayesian δικτύων..



Εικόνα 3.11 – Διάγραμμα ροής ανάλυσης δικτύων.

Κεφάλαιο 4

Αποτελέσματα - Συζήτηση

Σε αυτό το κεφάλαιο παραθέτονται τα αποτελέσματα που προέκυψαν ακολουθώντας την μεθοδολογία που περιγράφεται στο Κεφάλαιο 3 καθώς και σύντομος σχολιασμός αυτών.

4.1 Δημιουργία Bayesian Δικτύων

Σύμφωνα με την μεθοδολογία για την δημιουργία των δικτύων που έχει περιγραφεί στην Ενότητα 3.3 αρχικά βρέθηκαν κάποια κατώφλια για την περίπτωση των διακριτών μεταβλητών (Υποενότητα 3.3.2) και στη συνέχεια προχωρήσαμε στην εκμάθηση της δομής του δικτύου (Υποενότητα 3.4). Η εκμάθηση έγινε με τη χρήση του K2 αλγορίθμου ο οποίος απαιτεί, εκτός από τον μέγιστο αριθμό γονέων, και την αρχική ταξινόμηση των κόμβων. Για αυτό το σκοπό χρησιμοποιήθηκαν δυο ταξινομήσεις, η MWST και η Custom που είναι ένας συνδυασμός της προηγούμενης και της τυχαίας. Έτσι, έχοντας 2 διαφορετικές ταξινομήσεις κόμβων κατασκευάσαμε 8 δίκτυα. 2 (λόγω του αριθμού των διαφορετικών ταξινομήσεων) δίκτυα για την περίπτωση των διακριτών μεταβλητών-κόμβων και 2 για την περίπτωση των συνεχών (Gaussian) μεταβλητών-κόμβων. Όμως αυτή η διαδικασία έγινε ξεχωριστά για τα καρκινικά και τα υγιή δείγματα. Έτσι προκύπτουν 4 δίκτυα για την κάθε κατηγορία δειγμάτων. Στην Εικόνα 4.1 φαίνονται συνοπτικά τα στοιχεία των δικτύων που προέκυψαν για κάθε κατηγορία.

		Καρκινικά Δείγματα		Υγιή Δείγματα	
Είδος Μεταβλητής	Είδος Ταξινόμησης	Κόμβοι	Ακμές	Κόμβοι	Ακμές
Διακριτές Μεταβλητές	MWST	69	69	74	71
	Custom	73	73	72	75
Συνεχείς Μεταβλητές	MWST	82	551	82	359
	Custom	82	533	82	390

Εικόνα 4.1 – Βασικά στοιχεία των δικτύων που κατασκευάστηκαν.

Παρατηρούμε από τον πίνακα 4.1 ότι οι ακμές-αλληλεπιδράσεις, που προέκυψαν από την εκμάθηση της δομής, στην περίπτωση των διακριτών μεταβλητών είναι πολύ

λιγότερες σε σύγκριση με αυτές που προέκυψαν στην περίπτωση των συνεχών μεταβλητών. Αυτό είναι απολύτως λογικό και αναμενόμενο καθώς κατά τη διαδικασία της διακριτοποίησης που πραγματοποιήθηκε στην πρώτη περίπτωση υπήρχε απώλεια πληροφορίας, κάτι το οποίο δεν ισχύει στην δεύτερη περίπτωση. Μια ακόμα σημαντική παρατήρηση που μπορούμε να κάνουμε από αυτόν τον πίνακα, είναι η μεγάλη διαφορά ακμών στην περίπτωση των συνεχών μεταβλητών ανάμεσα στα καρκινικά και τα υγιή δείγματα. Οι ακμές της δομής του δικτύου, το οποίο προέκυψε από τα καρκινικά δείγματα, είναι αρκετά περισσότερες από αυτές που προέκυψαν από τα υγιή. Δεν μπορούμε να πούμε το ίδιο στην περίπτωση των διακριτών μεταβλητών καθώς η διαφορά είναι ελάχιστη.

4.2 Ανάλυση Δικτύων

Για την ανάλυση των δικτύων δημιουργήθηκαν ενώσεις δικτύων από τα ήδη υπάρχοντα. Συγκεκριμένα δημιουργήθηκαν οι ενώσεις των δικτύων από διακριτές μεταβλητές για τα καρκινικά δείγματα (**ΚΔ**) και για τα υγιή δείγματα (**ΥΔ**). Αντίστοιχα δημιουργήθηκαν ενώσεις δικτύων από αυτά με συνεχείς μεταβλητές για τα καρκινικά (**ΚΣ**) και για τα υγιή δείγματα (**ΥΣ**). Τέλος έγιναν οι ενώσεις όλων των δικτύων από καρκινικά δείγματα (**ΚΕ**) και από υγιή δείγματα (**ΥΕ**). Στον πίνακα 4.2 φαίνονται οι αριθμοί των κόμβων και ακμών τους.

Δίκτυο	Κόμβοι	Ακμές
ΚΕ	82	843
ΥΕ	82	605
ΚΔ	74	142
ΥΔ	77	102
ΚΣ	82	765
ΥΣ	82	562

Εικόνα 4.2 – Οι ενώσεις των δικτύων που δημιουργήθηκαν και τα βασικά στοιχεία τους.

4.2.1 Ανάλυση Δικτύων - Φαινόμενο Μικρού Κόσμου

Όπως αναφέρθηκε στην Υποενότητα 3.6.1, για να πούμε ότι ένα δίκτυο εμφανίζει το φαινόμενο του Μικρού Κόσμου θα πρέπει το μέσο μήκος μονοπατιού των κόμβων να είναι περίπου ίσο με $O(\log n)$ όπου n ο αριθμός των κόμβων και επίσης θα πρέπει να εμφανίζει υψηλό συντελεστή ομαδοποίησης συγκριτικά με ένα ισοδύναμο τυχαίο γράφο (ίδιο αριθμό κόμβων και ακμών).

Παρακάτω στο Πίνακα 4.3 φαίνονται αυτές οι ιδιότητες για όλα τα δίκτυα. Το C δείχνει τον συντελεστή ομαδοποίησης του δικτύου ενώ C_{rand} ο συντελεστής ομαδοποίησης του αντίστοιχου τυχαίου γράφου. Με l φαίνεται το μέσο μήκος μονοπατιού του δικτύου.

Στον παραπάνω πίνακα φαίνονται τα αποτελέσματα για την ανάλυση των ιδιοτήτων που απαιτούνται στην εξέταση του φαινομένου Μικρού Κόσμου. Παρατηρούμε ότι

Δίκτυο	C	C _{rand}	I	log(n)	n
ΚΕ	0,152	0,126	2,37308	4,4067	82
ΥΕ	0,120	0,099	2,47681	4,4067	82
ΚΔ	0,093	0,014	1,81395	4,3041	74
ΥΔ	0,041	0,014	1,13807	4,3438	77
ΚΣ	0,147	0,117	2,44098	4,4067	82
ΥΣ	0,116	0,083	2,50858	4,4067	82

Εικόνα 4.3 – Τιμές ιδιοτήτων που σχετίζονται με το φαινόμενο του Μικρού Κόσμου για κάθε δίκτυο ξεχωριστά.

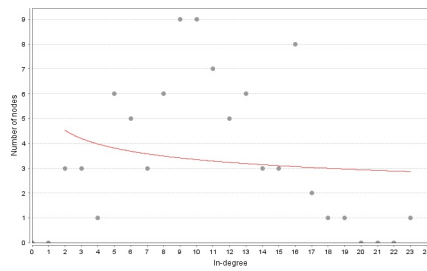
σε όλες τις περιπτώσεις ο συντελεστής ομαδοποίησης του δικτύου είναι υψηλότερος από αυτόν του αντίστοιχου τυχαίου γράφου το οποίο είναι ένα δείγμα του φαινομένου Μικρού Κόσμου. Όμως η τιμή του μέσου μήκους μονοπατιού είναι αρκετά μικρότερη από αυτήν του λογαρίθμου του αριθμού των κόμβων κάτι το οποίο μας απομακρύνει από την πιθανότητα τα δίκτυα μας να εμφανίζουν το φαινόμενο του Μικρού Κόσμου.

4.2.2 Ανάλυση Δικτύων - Δίκτυα Ελεύθερης Κλίμακας

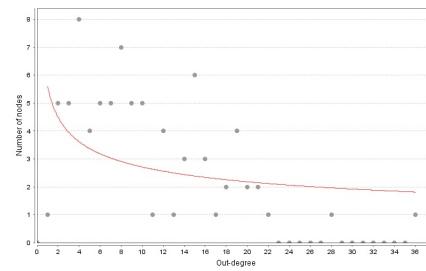
Στην Υποενότητα 3.6.2 είδαμε ότι ένα δίκτυο για να θεωρείται Ελεύθερης Κλίμακας θα πρέπει η κατανομή βαθμού αυτού να ακολουθεί μια κατανομή με νόμο δύναμης. Επίσης έχει παρατηρηθεί ότι και η κατανομή της ενδιάμεσης κεντρικότητας ακολουθεί μια κατανομή με νόμο δύναμης. Τα Bayesian δίκτυα είναι κατευθυνόμενα οπότε υπάρχει μέσα και έξω βαθμός δικτύου, όπου ο καθένας εξετάστηκε ξεχωριστά.

Ξεκινήσαμε μελετώντας τα δύο γενικά δίκτυα (ΚΕ και ΥΕ) και στη συνέχεια τα επιμέρους (ΚΔ, ΥΔ, ΚΣ, ΥΣ). Παρατηρώντας τις καμπύλες που ακολουθούν διαπιστώνουμε ότι τα δίκτυα μας μπορούν να θεωρηθούν ελεύθερης κλίμακας καθώς ιδιαίτερα οι κατανομές των βαθμών ακολουθούν μια κατανομή με νόμο δύναμης.

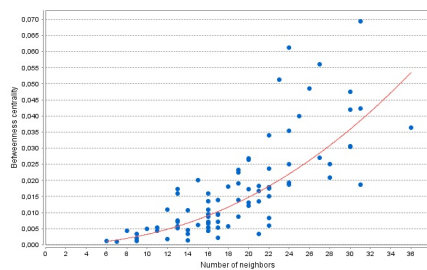
Δίκτυο Ένωσης Καρκινικών Δειγμάτων



(4.4.1) Κατανομή μέσα-βαθμού.



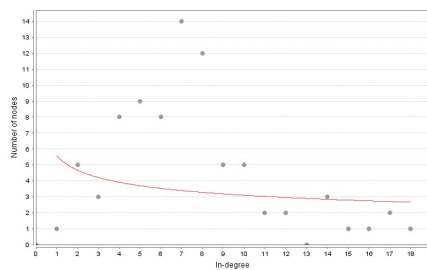
(4.4.2) Κατανομή έξω-βαθμού.



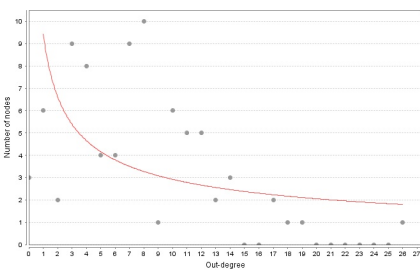
(4.4.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.4 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΚΕ δίκτυο.

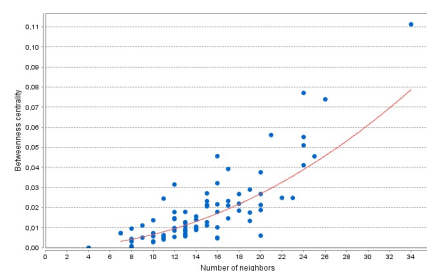
Δίκτυο Ένωσης Υγιών Δειγμάτων



(4.5.1) Κατανομή μέσα-βαθμού.



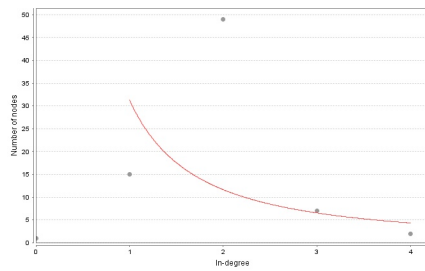
(4.5.2) Κατανομή έξω-βαθμού.



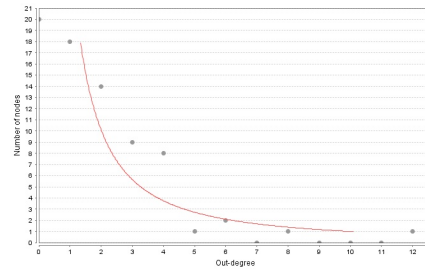
(4.5.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.5 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΥΕ δίκτυο.

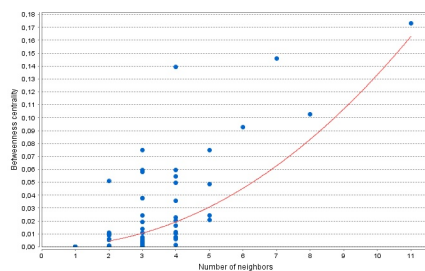
Δίκτυο Ένωσης Καρκινικών Δειγμάτων Διακριτών Μεταβλητών



(4.6.1) Κατανομή μέσα-βαθμού.



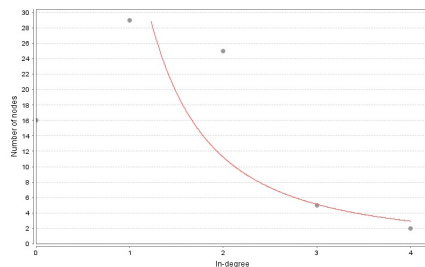
(4.6.2) Κατανομή έξω-βαθμού.



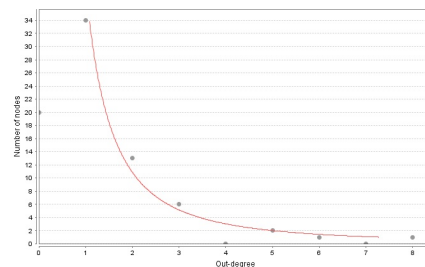
(4.6.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.6 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΚΔ δίκτυο.

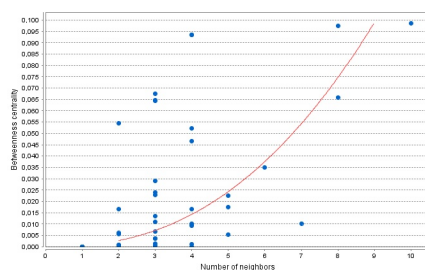
Δίκτυο Ένωσης Υγιών Δειγμάτων Διακριτών Μεταβλητών



(4.7.1) Κατανομή μέσα-βαθμού.



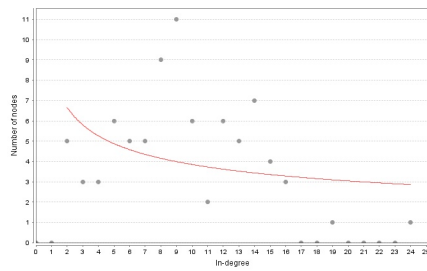
(4.7.2) Κατανομή έξω-βαθμού.



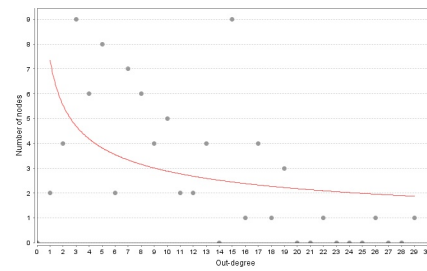
(4.7.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.7 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΥΔ δίκτυο.

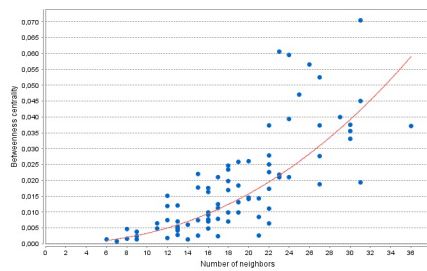
Δίκτυο Ένωσης Καρκινικών Δειγμάτων Συνεχών Μεταβλητών



(4.8.1) Κατανομή μέσα-βαθμού.



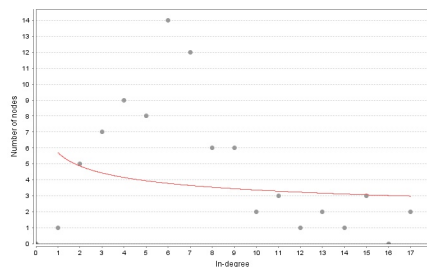
(4.8.2) Κατανομή έξω-βαθμού.



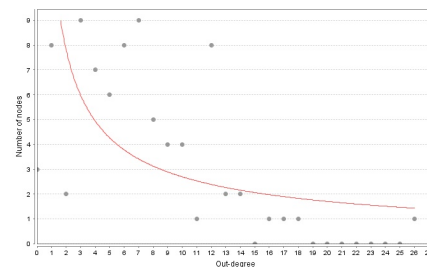
(4.8.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.8 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΚΣ δίκτυο.

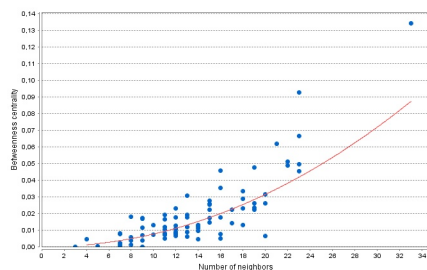
Δίκτυο Ένωσης Υγιών Δειγμάτων Συνεχών Μεταβλητών



(4.9.1) Κατανομή μέσα-βαθμού.



(4.9.2) Κατανομή έξω-βαθμού.



(4.9.3) Κατανομή ενδιάμεσης κεντρικότητας.

Εικόνα 4.9 – Διαγράμματα κατανομής βαθμού και ενδιάμεσης κεντρικότητας για το ΥΣ δίκτυο.

Συγκρίνοντας τις κατανομές βαθμού των δικτύων από διακριτές και συνεχείς μεταβλητές παρατηρούμε ότι οι πρώτες ακολουθούν μια πολύ καλύτερη κατανομή με νόμο δύναμης και ιδιαίτερα η κατανομή του έξω βαθμού. Όμως σε γενικές γραμμές, όλες οι κατανομές ακολουθούν μια κατανομής με νόμο δύναμης κάτι το οποίο μας επιτρέπει να πούμε ότι αυτά τα δίκτυα είναι ελεύθερης κλίμακας.

4.2.3 Ανάλυση Δικτύων - Άλλες Ιδιότητες

Στον παρακάτω πίνακα βλέπουμε μερικές βασικές ιδιότητες των δικτύων όπως προέκυψαν μετά από τον υπολογισμό τους.

Δίκτυο/ Ιδιότητα	Κόμβοι/ Ακμές	Μέσο Μήκος Μονοπατιού	Συντελεστής Ομαδοποίησης	Μέσος Μέσα-Βαθμός	Μέσος Έξω-Βαθμός	Μέση Ενδιάμεση Κεντρικότητα
ΚΕ	82/843	2,373	0,152	10,280	10,280	115,293
ΥΕ	82/605	2,477	0,120	7,378	7,378	123,243

Εικόνα 4.10 – Ιδιότητες Δικτύων.

Από αυτόν τον πίνακα μπορούμε να δούμε ότι το ΚΕ δίκτυο έχει μεγαλύτερο συντελεστή ομαδοποίησης παρόλο που ο αριθμός των κόμβων είναι μεγαλύτερος. Προσεγγιστικά το μέσο μήκος μονοπατιού και η μέση ενδιάμεση κεντρικότητα είναι παρόμοια ενώ οι βαθμοί του ΚΕ είναι σημαντικά μεγαλύτεροι. Σε γενικές γραμμές η τιμές του συντελεστή ομαδοποίησης είναι αρκετά χαμηλές για τα δεδομένα ενός βιολογικού δικτύου.

4.2.4 Ανάλυση Δικτύων - Σύνοψη

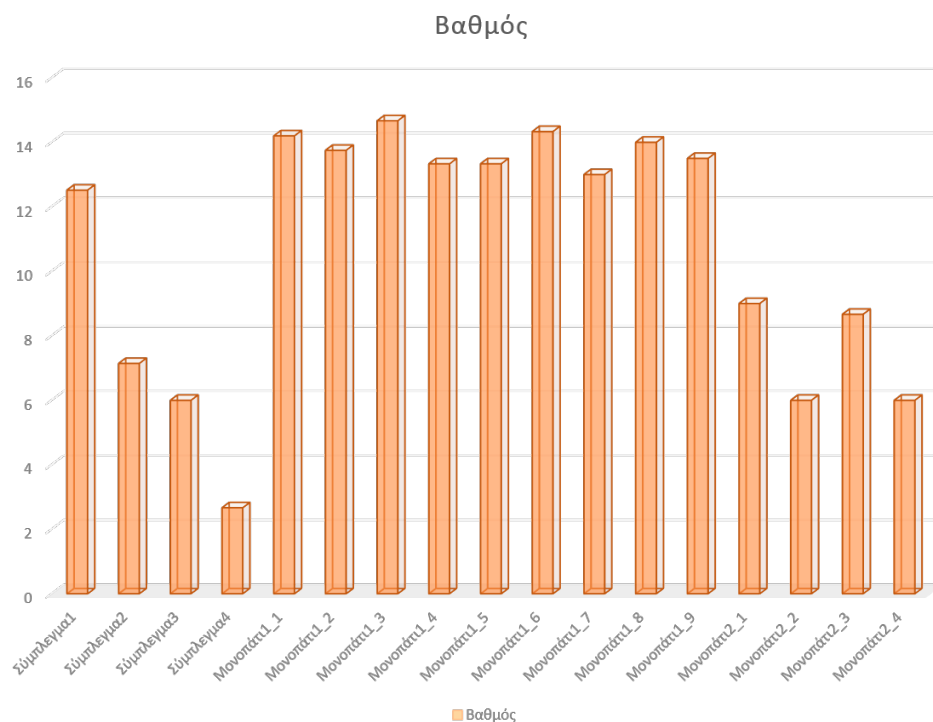
Συνοψίζοντας τις ιδιότητες που αναφέρουμε παραπάνω, μπορούμε να πούμε ότι το φαινόμενο του Μικρού Κόσμου δεν εμφανίζεται στα δίκτυα μας. Αντίθετα μπορούμε να πούμε με αρκετά μεγάλη σιγουριά ότι τα δίκτυα μας είναι Ελεύθερης Κλίμακας.

4.3 Σύγκριση Δικτύων

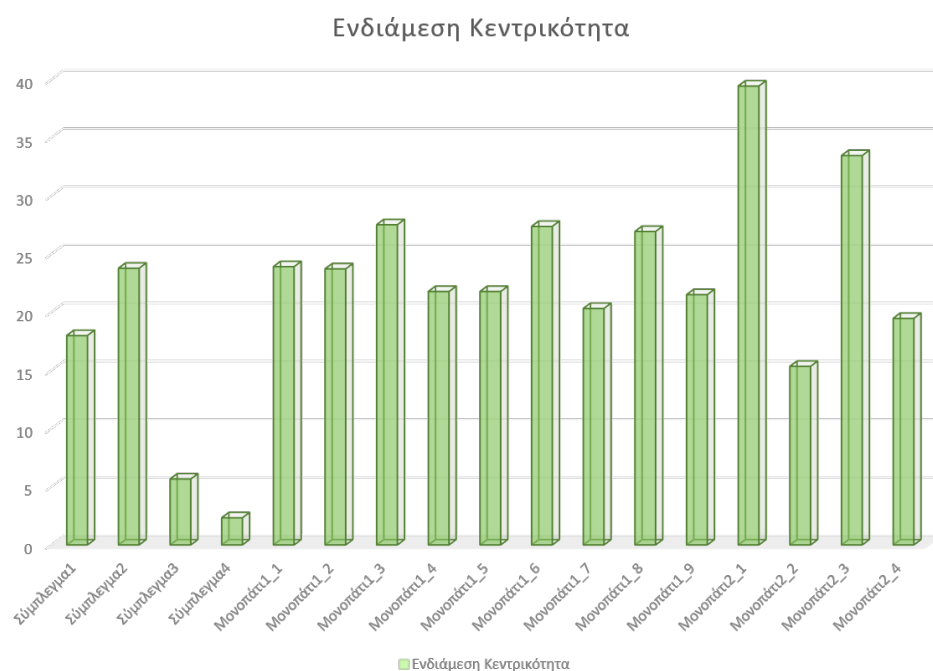
Στα πλαίσια της σύγκρισης των δύο δικτύων, μετρήθηκαν και ταυτοποιήθηκαν οι κοινές αλληλεπιδράσεις τους. Ο αριθμός των κοινών αλληλεπιδράσεων ήταν εκατόν εβδομήντα μία (171) και φαίνονται στον πίνακα της Εικόνας C.1 στο Παράρτημα C. Στη συνέχεια έγινε σύγκριση των δικτύων σύμφωνα με μια μέθοδο που μετράει τη διαφορετικότητα ανάμεσα στα δύο δίκτυα σε ένα συγκεκριμένο επίπεδο απόστασης κόμβων. Αυτή η μέτρηση κυμαίνεται ανάμεσα στο 0 και στο 1. Όσο υψηλότερη είναι αυτή η τιμή τόσο πιο ποιοτικό είναι το δίκτυο σε σύγκριση με το άλλο. Η μέθοδος περιγράφεται πλήρως στην Ενότητα 3.7.

Παρακάτω στον πίνακα φαίνονται οι τιμές εγκυρότητας που προέκυψαν με βάση την παραπάνω μέθοδο. Μπορούμε να δούμε ότι σε κάθε επίπεδο (1-4) το ΥΕ δίκτυο

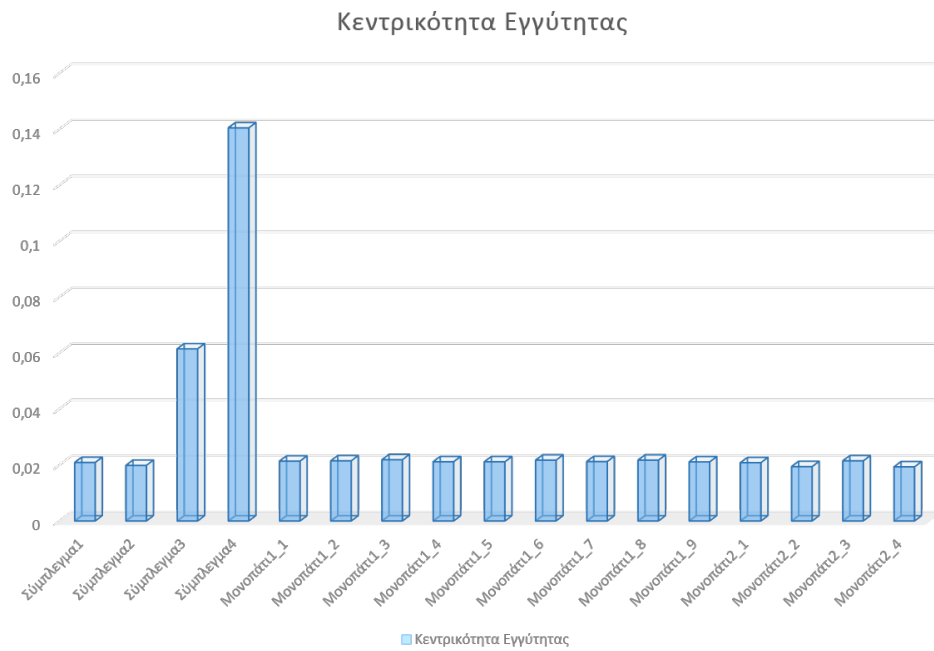
Παρακάτω φαίνονται τα διαγράμματα των μέσων όρων κάθε τοπολογικού χαρακτηριστικού που περιγράψουμε παραπάνω.



Εικόνα 4.13 – Διάγραμμα βαθμών.



Εικόνα 4.14 – Διάγραμμα Ενδιάμεσων Κεντρικοτήτων.

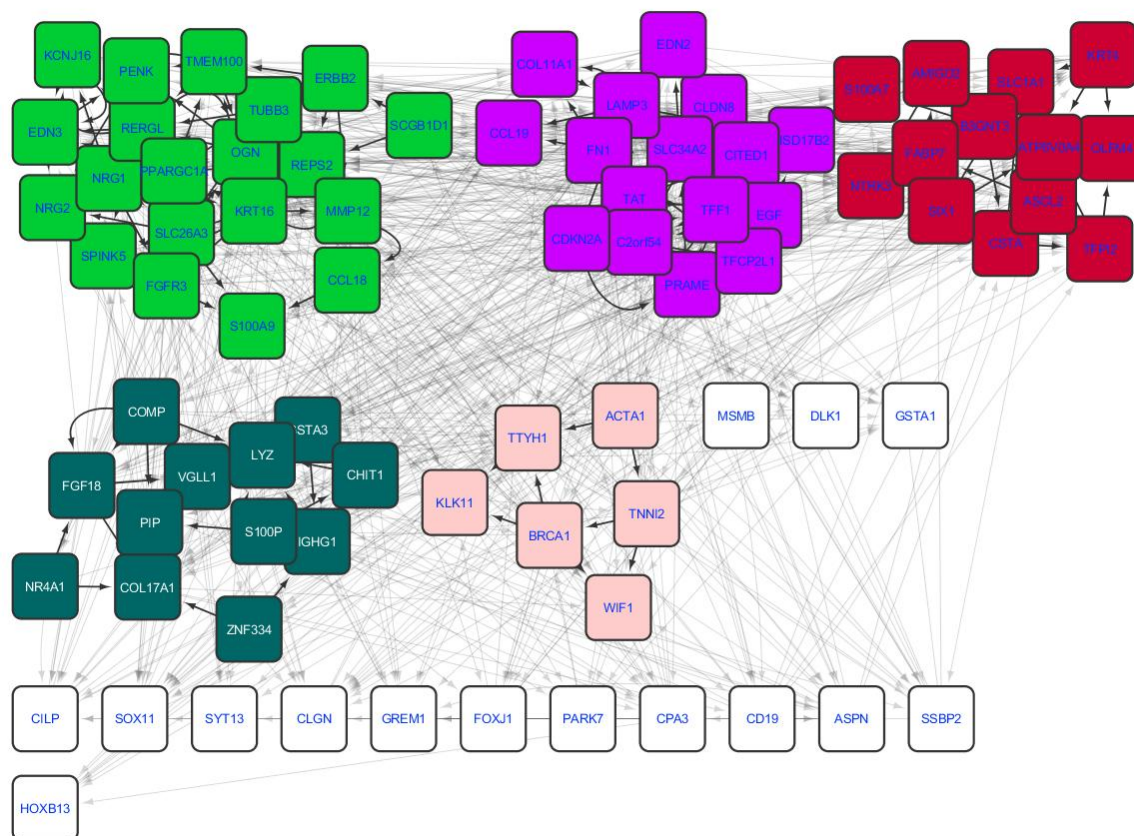


Εικόνα 4.15 – Διάγραμμα Κεντρικότητας Εγγύτητας.

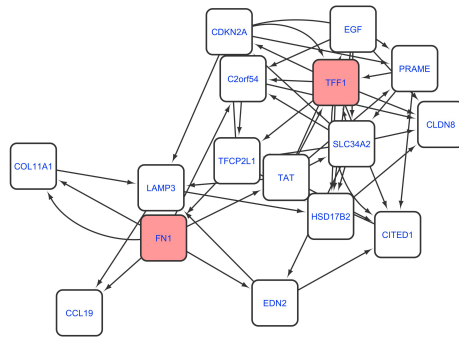
Σε αυτά τα διαγράμματα μπορούμε να δούμε ότι οι μέσοι όροι των κεντρικότητων των συμπλεγμάτων και των αντίστοιχων σημαντικών μονοπατιών κυμαίνονται στα ίδια επίπεδα. Αυτό μας δείχνει ότι όντως αυτά τα μονοπάτια είναι σημαντικά.

4.4 Ομαδοποίηση με τον αλγόριθμο MCODE

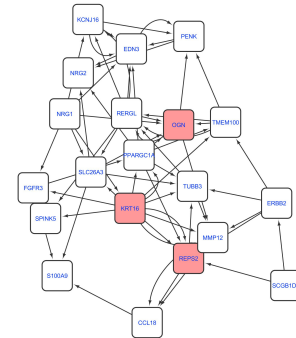
Παρακάτω παραθέτονται τα αποτελέσματα του εργαλείου MCODE, ο αλγόριθμος του οποίου αναλύθηκε στην Υποενότητα 3.8.1. Το εργαλείο MCODE εφαρμόστηκε στο ΚΕ και ΥΕ δίκτυο προκειμένου να μελετηθούν τα συμπλέγματα που θα προκύψουν.



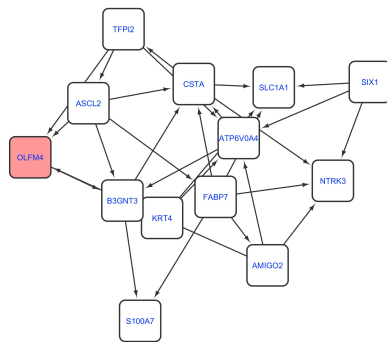
Εικόνα 4.16 – Το ΚΕ δίκτυο χωρισμένο σε 5 συμπλέγματα.



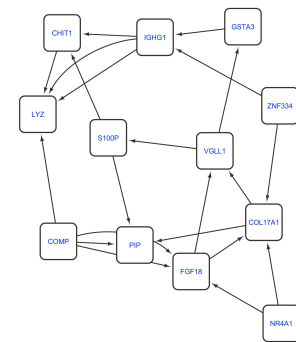
(4.17.1) Δίκτυο ΚΕ - Σύμπλεγμα 1.
Σκορ: 6.



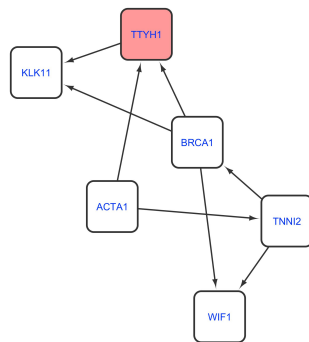
(4.17.2) Δίκτυο ΚΕ - Σύμπλεγμα 2.
Σκορ: 5.579.



(4.17.3) Δίκτυο ΚΕ - Σύμπλεγμα 3.
Σκορ: 4.833.

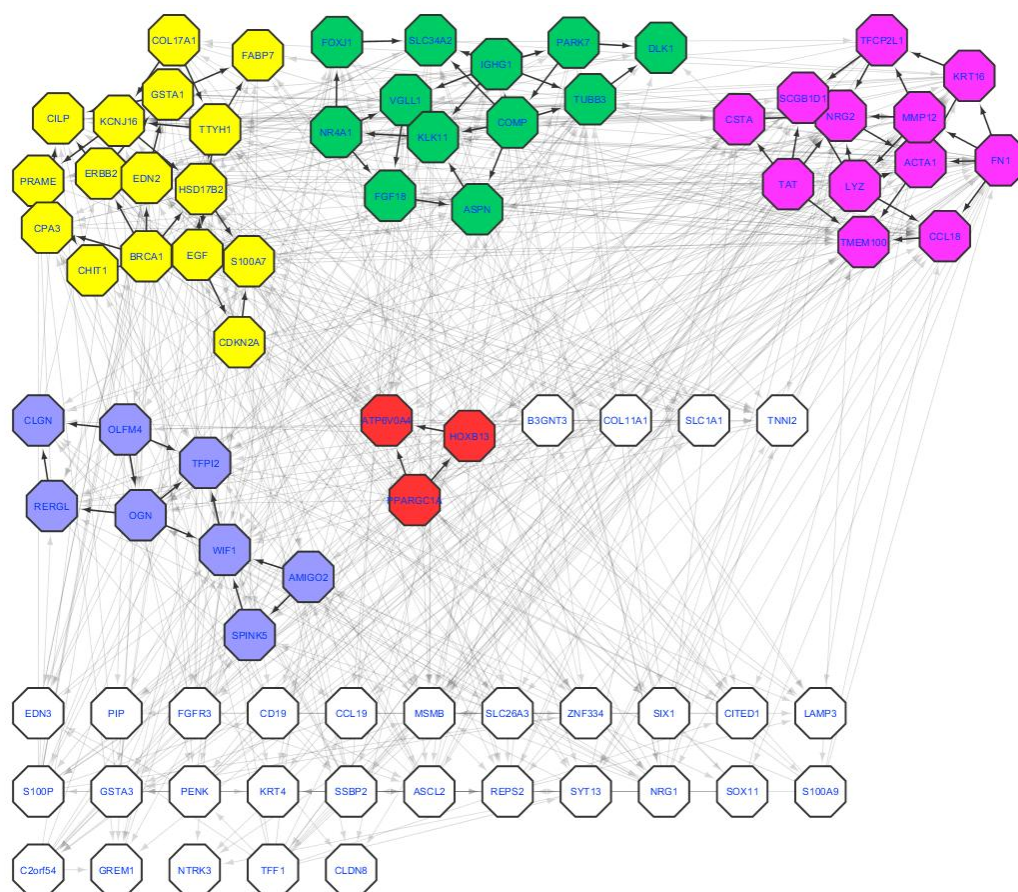


(4.17.4) Δίκτυο ΚΕ - Σύμπλεγμα 4.
Σκορ: 3.455.

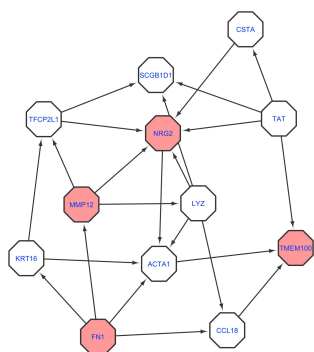


(4.17.5) Δίκτυο ΚΕ - Σύμπλεγμα 5.
Σκορ: 3.2.

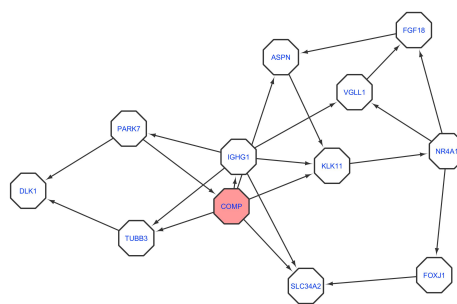
Εικόνα 4.17 – Τα 5 συμπλέγματα που προέκυψαν από την ομαδοποίηση του ΚΕ δικτύου με τον αλγόριθμο MCODE.



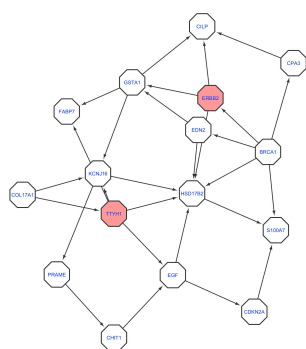
Εικόνα 4.18 – Το YE δίκτυο χωρισμένο σε 5 συμπλέγματα.



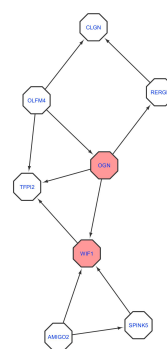
(4.19.1) Δίκτυο ΥΕ - Σύμπλεγμα 1.
Σκορ: 4.182.



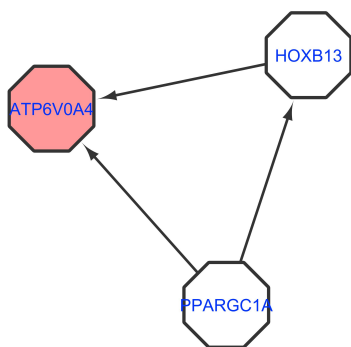
(4.19.2) Δίκτυο ΥΕ - Σύμπλεγμα 2.
Σκορ: 3.818.



(4.19.3) Δίκτυο ΥΕ - Σύμπλεγμα 3.
Σκορ: 3.733.



(4.19.4) Δίκτυο ΥΕ - Σύμπλεγμα 4.
Σκορ: 3.143.

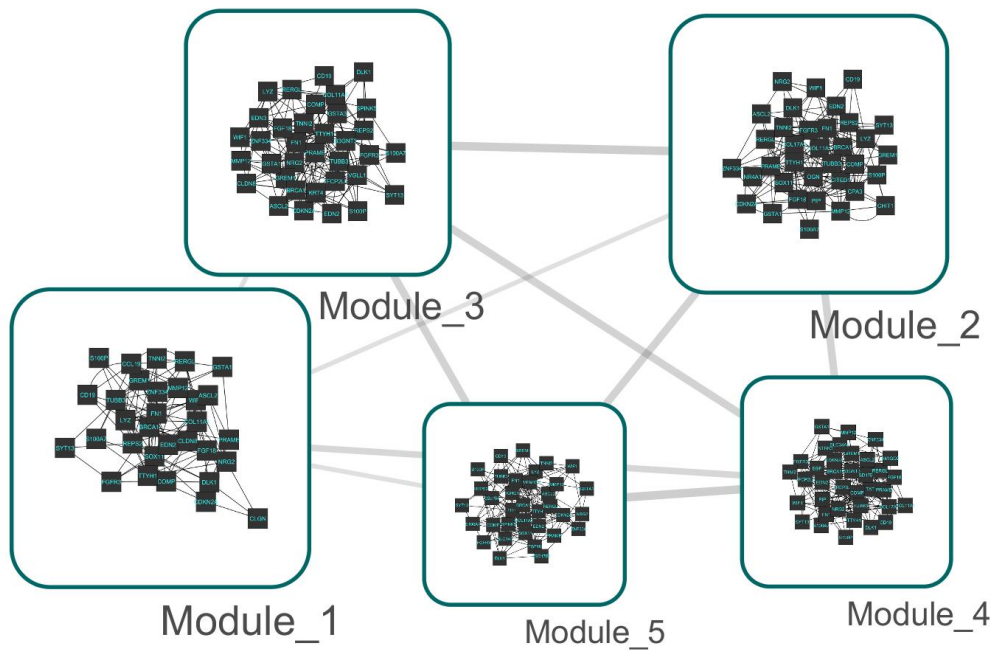


(4.19.5) Δίκτυο ΥΕ - Σύμπλεγμα 5.
Σκορ: 3.

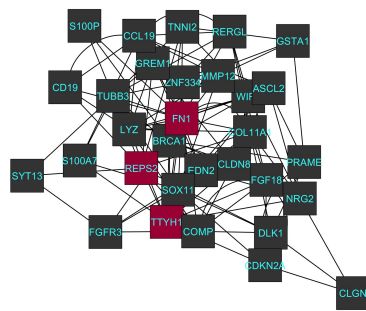
Εικόνα 4.19 – Τα 5 συμπλέγματα που προέκυψαν από την ομαδοποίηση του ΥΕ δικτύου με τον αλγόριθμο MCODE.

4.5 Ομαδοποίηση με τον αλγόριθμο jActiveModules

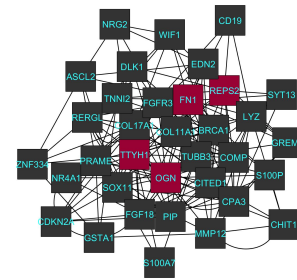
Παρακάτω παραθέτονται τα αποτελέσματα του εργαλείου jActiveModules, ο αλγόριθμος του οποίου αναλύθηκε στην Υποενότητα 3.8.2. Ως είσοδο σε αυτό το εργαλείο δώσαμε τα p-values που προέκυψαν από το t-test δύο δειγμάτων (two sample t-test). Στο t-test χρησιμοποιήθηκαν οι τιμές έκφρασης για τα καρκινικά και τα υγιή δείγματα για όλα τα 82 γονίδια. Το εργαλείο jActiveModules εφαρμόστηκε στο ΚΕ και ΥΕ δίκτυο προκειμένου να μελετηθούν τα συμπλέγματα που θα προκύψουν.



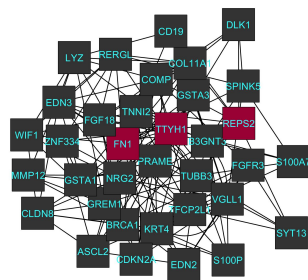
Εικόνα 4.20 – Το ΚΕ δίκτυο χωρισμένο σε 5 δομικές ενότητες.



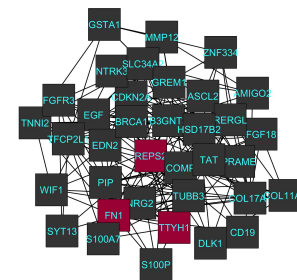
(4.21.1) Δίκτυο ΚΕ - Δομική Ενότητα 1.
Σκορ: 4.761.



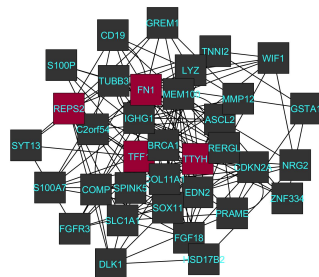
(4.21.2) Δίκτυο ΚΕ - Δομική Ενότητα 2.
Σκορ: 4.212.



(4.21.3) Δίκτυο ΚΕ - Δομική Ενότητα 3.
Σκορ: 3.873.

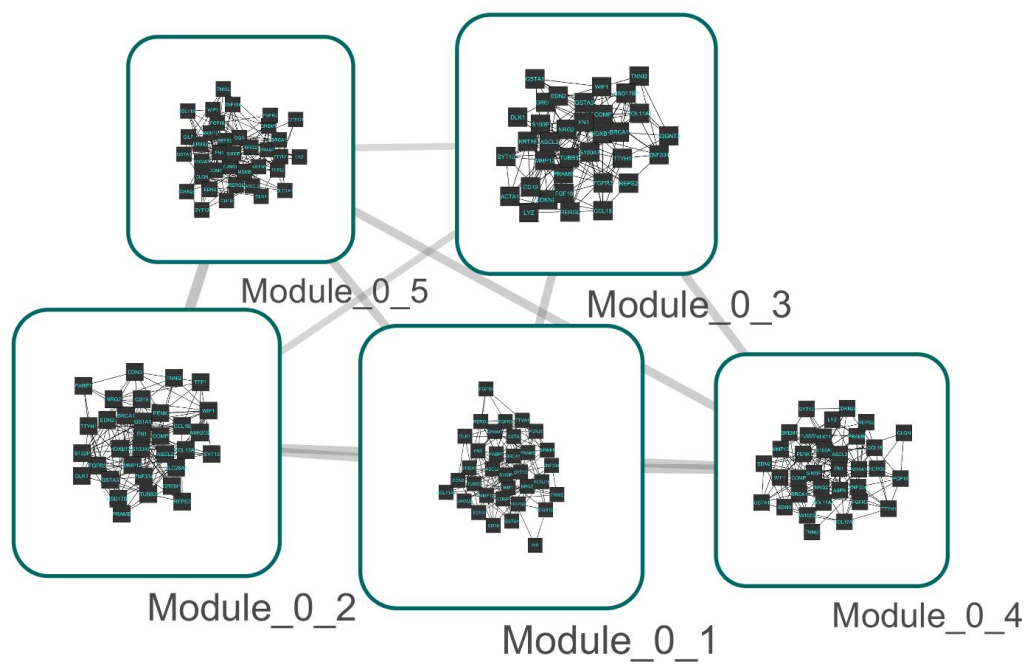


(4.21.4) Δίκτυο ΚΕ - Δομική Ενότητα 4.
Σκορ: 2.435.

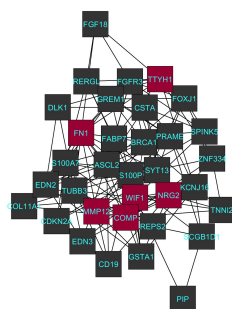


(4.21.5) Δίκτυο ΚΕ - Δομική Ενότητα 5.
Σκορ: 2.215.

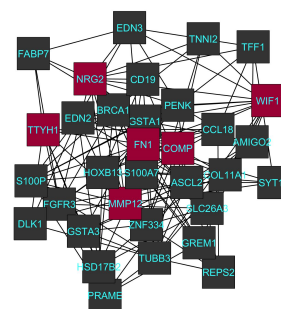
Εικόνα 4.21 – Οι 5 δομικές ενότητες που προέκυψαν από την ομαδοποίηση του ΚΕ δικτύου με τον αλγόριθμο jActiveModules.



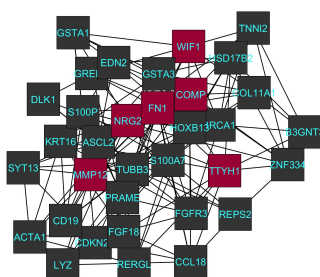
Εικόνα 4.22 – Το YE δίκτυο χωρισμένο σε 5 δομικές ενότητες.



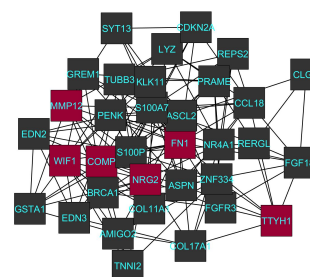
(4.23.1) Δίκτυο ΥΕ - Δομική Ενότητα 1.
Σκορ: 4.440.



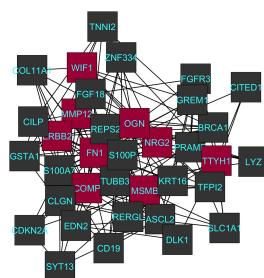
(4.23.2) Δίκτυο ΥΕ - Δομική Ενότητα 2.
Σκορ: 4.012.



(4.23.3) Δίκτυο ΥΕ - Δομική Ενότητα 3.
Σκορ: 3.819.



(4.23.4) Δίκτυο ΥΕ - Δομική Ενότητα 4.
Σκορ: 3.098.



(4.23.5) Δίκτυο ΥΕ - Δομική Ενότητα 5.
Σκορ: 2.921.

Εικόνα 4.23 – Οι 5 δομικές ενότητες που προέκυψαν από την ομαδοποίηση του ΥΕ δικτύου με τον αλγόριθμο jActiveModules.

4.6 Σύγκριση Υποδικτύων

Οι παρακάτω πίνακες μας δείχνουν τον αριθμό των κοινών αλληλεπιδράσεων που τυχόν έχουν τα υποδίκτυα που προκύπτουν εφαρμόζοντας κάποιον αλγόριθμο ομαδοποίησης στα δίκτυα ΚΕ και ΥΕ.

MCODE		Καρκινικά Δείγματα				
		Σύμπλοκο 1	Σύμπλοκο 2	Σύμπλοκο 3	Σύμπλοκο 4	Σύμπλοκο 5
Υγιή Δείγματα	Σύμπλεγμα 1	0	0	0	0	0
	Σύμπλεγμα 2	0	0	0	2	0
	Σύμπλεγμα 3	1	0	0	0	0
	Σύμπλεγμα 4	0	1	1	0	0
	Σύμπλεγμα 5	0	0	0	0	0

Εικόνα 4.24 – Αριθμός κοινών αλληλεπιδράσεων των συμπλεγμάτων που προέκυψαν από την εφαρμογή του MCODE αλγορίθμου.

Παρατηρούμε ότι, συγκρίνοντας τα συμπλέγματα που προέκυψαν από την εφαρμογή του MCODE αλγορίθμου στα δύο δίκτυα ΚΕ και ΥΕ, υπάρχουν μηδαμινές κοινές αλληλεπιδράσεις.

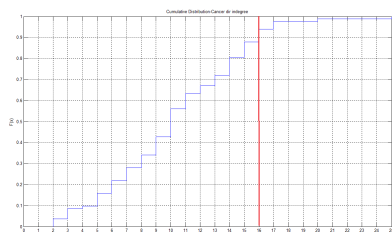
jActiveModules		Καρκινικά Δείγματα				
		Σύμπλοκο 1	Σύμπλοκο 2	Σύμπλοκο 3	Σύμπλοκο 4	Σύμπλοκο 5
Υγιή Δείγματα	Δομ.Ενότητα 1	22	22	24	22	22
	Δομ.Ενότητα 2	20	20	24	21	22
	Δομ.Ενότητα 3	23	23	26	24	24
	Δομ.Ενότητα 4	23	27	25	25	23
	Δομ.Ενότητα 5	23	29	23	22	25

Εικόνα 4.25 – Αριθμός κοινών αλληλεπιδράσεων των δομικών ενοτήτων που προέκυψαν από την εφαρμογή του jActiveModules αλγορίθμου.

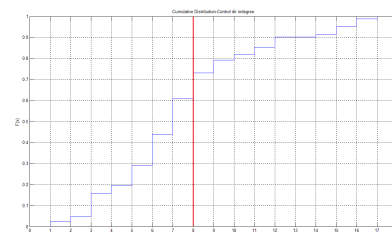
Αντίθετα, στις δομικές μονάδες που προέκυψαν εφαρμόζοντας τον jActiveModules αλγόριθμο, υπάρχουν αρκετές κοινές αλληλεπιδράσεις μεταξύ τους. Ο μεγαλύτερος αριθμός κοινών αλληλεπιδράσεων παρατηρείται μεταξύ των δομικών μονάδων 2 του ΚΕ και 5 του ΥΕ δικτύου.

4.7 Αναζήτηση Κεντρικών Κόμβων

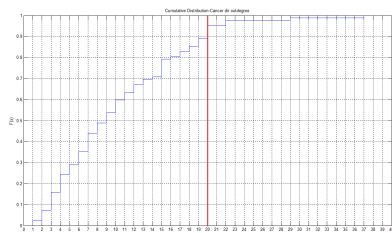
Η μέθοδος που ακολουθήσαμε για την αναζήτηση κεντρικών και σημαντικών κόμβων παρουσιάστηκε στην Υποενότητα 3.10. Εφόσον τα Bayesian δίκτυα είναι κατευθυνόμενα, οι βαθμοί κάθε κόμβου είναι δυο, ο μέσα και ο έξω. Για κάθε ένα ξεχωριστά βρέθηκαν οι πιο σημαντικοί κόμβοι και συμπεριλήφθηκαν στην τομή. Παρακάτω φαίνονται οι καμπύλες και με κόκκινες γραμμές παρουσιάζεται η ελάχιστη τιμή όπου ένας κόμβος θεωρείται σημαντικός.



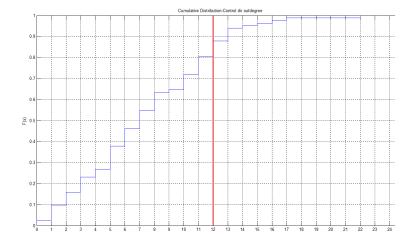
(4.26.1) Δίκτυο ΚΕ.
Μέσα-βαθμοί κόμβων.



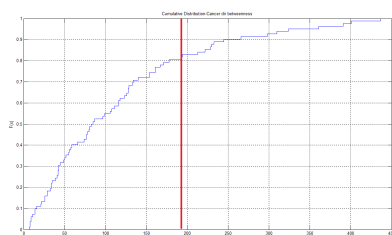
(4.26.2) Δίκτυο ΥΕ.
Μέσα-βαθμοί κόμβων.



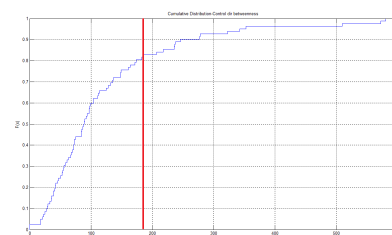
(4.26.3) Δίκτυο ΚΕ.
Έξω-βαθμοί κόμβων.



(4.26.4) Δίκτυο ΥΕ.
Έξω-βαθμοί κόμβων.



(4.26.5) Δίκτυο ΚΕ.
Ενδιάμεσες κεντρικότητες κόμβων.



(4.26.6) Δίκτυο ΥΕ.
Ενδιάμεσες κεντρικότητες κόμβων.

Εικόνα 4.26 – Αθροιστικά διαγράμματα για την εύρεση κεντρικών κόμβων.

Σύμφωνα με την παραπάνω μεθοδολογία και τα διαγράμματα προέκυψαν οι παρακάτω hubs:

Hubs KE δικτύου	Hubs YE δικτύου
KRT16	FN1
FN1	ERBB2
OGN	COMP
TFF1	ATP6V0A4
OLFM4	WIF1
TTYH1	MMP12
REPS2	OGN
	TTYH1
	NRG2
	MSMB
	TMEM100

Εικόνα 4.27 – Οι hubs που τελικά προέκυψαν.

4.7.1 Κεντρικοί Κόμβοι και Σύγκριση Δικτύων

Δίκτυο	KE		YE	
	#Hubs	Hubs	#Hubs	Hubs
Επίπεδο 1	7/7	KRT16, FN1, OGN, TFF1, OLFM4, TTYH1, REPS2	9/11	FN1, COMP, ATP6V0A4, WIF1, MMP12, OGN, TTYH1, MSMB, TMEM100
Επίπεδο 2	7/7	KRT16, FN1, OGN, TFF1, OLFM4, TTYH1, REPS2	11/11	FN1, ERBB2, COMP, ATP6V0A4, WIF1, MMP12, OGN, TTYH1, NRG2, MSMB, TMEM100
Επίπεδο 3	7/7	KRT16, FN1, OGN, TFF1, OLFM4, TTYH1, REPS2	11/11	FN1, ERBB2, COMP, ATP6V0A4, WIF1, MMP12, OGN, TTYH1, NRG2, MSMB, TMEM100
Επίπεδο 4	3/7	TFF1, OLFM4, REPS2	2/11	MMP12, NRG2

Εικόνα 4.28 – Ο αριθμός και τα ονόματα των Hubs που συμμετέχουν ανά επίπεδο σε σχέση με τον συνολικό αριθμό Hubs σε κάθε δίκτυο.

Εδώ παρατηρούμε για το KE δίκτυο ότι στα επίπεδα 1,2 και 3 συμμετέχουν όλοι οι κεντρικοί κόμβοι (hubs) στις αλληλεπιδράσεις τους, ενώ στο επίπεδο 4 συμμετέχουν μόνο οι τρεις από τις επτά. Κάτι αντίστοιχο συμβαίνει και στο YE δίκτυο αφού στο επίπεδο 1 συμμετέχουν εννιά από τους έντεκα κεντρικούς κόμβους, στο επίπεδο 2 και 3 συμμετέχουν όλοι ενώ στο επίπεδο 4 μόνο δύο από τους έντεκα.

4.7.2 Κεντρικοί Κόμβοι και Ομαδοποίηση με το εργαλείο MCODE

Δίκτυο	ΚΕ		ΥΕ	
	#Hubs	Hubs	#Hubs	Hubs
Σύμπλεγμα 1	2/7	FN1,TFF1	1/11	FN1
Σύμπλεγμα 2	3/7	KRT16,OGN,REPS2	5/11	ERBB2,MMP12,OGN, NRG2,TMEM100
Σύμπλεγμα 3	1/7	OLFM4	1/11	ATP6V0A4
Σύμπλεγμα 4	0/7		1/11	COMP
Σύμπλεγμα 5	1/7	TTYH1	2/11	WIF1,TTYH1

Εικόνα 4.29 – Ο αριθμός και τα ονόματα των Hubs που συμμετέχουν ανά σύμπλεγμα σε σχέση με τον συνολικό αριθμό Hubs σε κάθε δίκτυο.

Στα συμπλέγματα που προέκυψαν από το MCODE βρέθηκαν κάποιοι κεντρικοί κόμβοι. Τα συμπλέγματα του ΚΕ με τα υψηλότερα σκορ, τα οποία είναι τα πρώτα τρία, εμφάνισαν 2,3 και 1 από τους επτά κεντρικούς κόμβους αντίστοιχα ενώ για το ΥΕ δίκτυο οι κεντρικοί κόμβοι που εμφανίστηκαν ήταν 1,5 και 1 από τους έντεκα αντίστοιχα.

4.7.3 Κεντρικοί Κόμβοι και Ομαδοποίηση με το εργαλείο jActiveModules

Δίκτυο	ΚΕ		ΥΕ	
	#Hubs	Hubs	#Hubs	Hubs
Δομ.Ενότητα 1	3/7	FN1,TTYH1,REPS2	6/11	FN1,COMP,WIF1,MMP12, TTYH1,NRG2
Δομ.Ενότητα 2	4/7	FN1,OGN,TTYH1, REPS2	7/11	FN1,COMP,WIF1,MMP12, OGN,TTYH1,NRG2
Δομ.Ενότητα 3	3/7	FN1,TTYH1,REPS2	6/11	FN1,COMP,WIF1,MMP12, TTYH1,NRG2
Δομ.Ενότητα 4	3/7	FN1,TTYH1,REPS2	6/11	FN1,COMP,WIF1,MMP12, TTYH1,NRG2
Δομ.Ενότητα 5	4/7	FN1,TFF1,TTYH1, REPS2	7/11	FN1,COMP,WIF1,MMP12, TTYH1,NRG2,TMEM100

Εικόνα 4.30 – Ο αριθμός και τα ονόματα των Hubs που συμμετέχουν ανά σύμπλεγμα σε σχέση με τον συνολικό αριθμό Hubs σε κάθε δίκτυο.

Στα συμπλέγματα που προέκυψαν από το εργαλείο jActiveModules εμφανίστηκαν

περισσότεροι κεντρικοί κόμβοι συνολικά. Με σειρά αύξοντα σκορ στο ΚΕ δίκτυο εμφανίστηκαν 3,4,3,3 και 4 από τους επτά κεντρικούς κόμβους ενώ για το ΥΕ δίκτυο εμφανίστηκαν 6,7,6,6, και 7 από τους έντεκα κεντρικούς κόμβους.

4.7.4 Σύνοψη

Συνοψίζοντας παρατηρούμε ότι υπάρχουν κάποιοι κόμβοι που εμφανίζονται στην πλειοψηφία των περιπτώσεων. Αυτοί είναι οι FN1, OGN, TTYH1, REPS2.

4.8 Βιολογική Ερμηνεία των Αποτελεσμάτων

Ο σκοπός της παρούσας εργασίας είναι να δώσει μια εικόνα λειτουργικών υποδικτύων με κεντρικούς κόμβους μέσα από τη συμπεριφορά των γονιδίων που συνδέονται σε ένα δίκτυο γονιδιακών αλληλεπιδράσεων, αναγνωρίζοντας ως "γονίδιο-γονίδιο αλληλεπίδραση" τόσο την αλληλεπίδραση γονιδίου-γονιδιακού προϊόντος όσο και την αλληλεπίδραση γονιδιακού προϊόντος-γονιδιακού προϊόντος. Στα πλαίσια αυτά εφαρμόστηκε μια Bayesian προσέγγιση για την κατασκευή γονιδιακών/πρωτεϊνικών δικτύων, ώστε να ταυτοποιηθούν τα υποδίκτυα που μπορεί να προσφέρουν νέες γνώσεις - περισσότερο από ότι τα μεμονωμένα γονίδια - σχετικά με τα λειτουργικά μοριακά μονοπάτια που εμπλέκονται στον καρκίνο του μαστού.

Αρχικά, όπως αναφέρθηκε αναλυτικά στην Ενότητα 3.2, χρησιμοποιήθηκαν δεδομένα έκφρασης 4174 γονιδίων από 954 δείγματα ιστού (529 με καρκίνο του μαστού και 425 φυσιολογικά) [103].

Για την κατασκευή των Bayesian δικτύων επιλέχθηκε μια γονιδιακή υπογραφή με δυνητική κλινική εφαρμογή που προσδιορίστηκε στο εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας [104]. Πέρα από την '77 κοινή γονιδιακή υπογραφή', επιλέχθηκαν πέντε γονίδια που αποτέλεσαν τα γονίδια αναφοράς για τον έλεγχο της ορθότητας του αλγορίθμου (NRG2) και της θετικής (BRCA1, ERBB2) ή αρνητικής (NRG2, SSBP2, PARK7) συσχέτισης με τον καρκίνο του μαστού (Παράρτημα Α, Εικόνα Α.1). Στη συνέχεια, κατά τη διαδικασία εύρεσης των κατωφλίων για τη διακριτοποίηση των δεδομένων έκφρασης, αποδόθηκαν οι όροι "υπερεκφρασμένα" ή "υποεκφρασμένα" σε 30 γονίδια, γνωστά για την αυξημένη (*GABRP*, *FABP4*, *OLFM4*, *KRT13*, *MMP7*, *MUC1*, *GLUL*, *S100A7*, *S100A9*, *SOD2*, *ING4*, *CDH11*, *CXCR4*, *VEGFA*, *ITM2B*) ή μειωμένη (*CAV1*, *TIMP4*, *SLC20A1*, *CD44*, *HOXB6*, *CTGF*, *ECHDC2*, *SAE1*, *LCP1*, *FUCA1*, *TNFRSF10B*, *AGR2*, *ANGPT2*, *APOL1*, *EGFR*) έκφρασή τους στον καρκίνο του μαστού σύμφωνα με τη βάση δεδομένων Expression Atlas [105] και πρόσφατες μελέτες [106]. Συγκεκριμένα, η βάση δεδομένων Expression Atlas παρέχει πληροφορίες σχετικά με την έκφραση των γονιδίων, των πρωτεϊνών και τις παραλλαγές ματίσματος σε διαφορετικούς τύπους κυττάρων καθώς και σε διάφορα σημεία του οργανισμού, αναπτυξιακά στάδια, ασθένειες και άλλες βιολογικές και πειραματικές συνθήκες [105].

Δεδομένης της σημασίας των αλληλεπιδράσεων στη δημιουργία δικτύων έκφρασης για γονιδιακούς δείκτες καρκίνου του μαστού, αναζητήθηκαν οι γνωστές αλληλεπιδράσεις των 82 υπό εξέταση γονιδίων στο ανοικτό αποθετήριο αλληλεπιδράσεων Bi-

oGRID (Βιολογικό Γενικό Αποθετήριο Δεδομένων Αλληλεπίδρασης, Biological General Repository for Interaction Datasets) [107], προκειμένου να χρησιμοποιηθούν ως *a priori* γνώση για την κατασκευή των Bayesian δικτύων έκφρασης. Παρόλο που η έκδοση 3.2.108 της BioGRID (προσπέλαση Ιανουάριος 2014) παρείχε συνολικά 2.197 αλληλεπιδράσεις για τα 82 επιλεγμένα γονίδια, εντοπίστηκαν μόνο 12 γνωστές αλληλεπιδράσεις ανάμεσα στα 15 από τα 82 υπό εξέταση γονίδια (Ενότητα 3.4, Εικόνα 3.8.1 και Εικόνα 3.8.2).

Για την οπτικοποίηση των δυο δικτύων αλλά και τη δημιουργία δομικών ενοτήτων (modules) και συμπλεγμάτων (clusters) από αυτά, χρησιμοποιήθηκαν κατάλληλοι αλγόριθμοι μοριακής ομαδοποίησης, jActiveModules και MCODE αντίστοιχα, μέσα από τη πλατφόρμα Cytoscape, όπως έχει περιγραφεί στην Ενότητα 3.8 (Υποενότητες 3.8.1 και 3.8.2). Στα πλαίσια της παρούσας εργασίας, η βιολογική ερμηνεία στηρίχτηκε στα αποτελέσματα που παρήχθησαν κατά τη δημιουργία των δυο Bayesian δικτύων έκφρασης για γονιδιακούς δείκτες, ενός "καρκινικού" δικτύου (με δεδομένα από ασθενείς με καρκίνο του μαστού) και ενός "φυσιολογικού" δικτύου (με δεδομένα από μη πασχόντων από καρκίνο του μαστού) καθώς και των δομικών ενοτήτων και συμπλεγμάτων που προέκυψαν από αυτά.

Το ενιαίο καρκινικό μπαγεσιανό δίκτυο που δημιουργήθηκε αποτελείται από 843 αλληλεπιδράσεις, ενώ το ενιαίο φυσιολογικό μπαγεσιανό δίκτυο αποτελείται από 605 αλληλεπιδράσεις. Σ' αυτό το σημείο θα πρέπει να τονιστεί, ότι από τις 12 γνωστές αλληλεπιδράσεις μεταξύ των 82 γονιδίων, ταυτοποιήθηκαν πέντε αλληλεπιδράσεις στο φυσιολογικό δίκτυο (FN1-CDKN2A, FN1-KRT16, FN1-COMP, ERBB2-NRG1, FGFR3-FGF18) και τρεις αλληλεπιδράσεις (ACTA1-CDKN2A, FN1-COMP, KRT16-IGHG1) στο καρκινικό δίκτυο. Για την αξιολόγηση όλων των νέων αλληλεπιδράσεων που δημιουργήθηκαν ανάμεσα στα 82 γονίδια των δυο ξεχωριστών Bayesian δικτύων, χρησιμοποιήθηκε η βάση δεδομένων HIPPIE (ανθρώπινη ολοκληρωμένη αναφορά αλληλεπίδρασης πρωτεϊνών, Human Integrated Protein-Protein Interaction reference) [108]. Μέσω της HIPPIE προσδιορίστηκαν οι γνωστές αλληλεπιδράσεις ανάμεσα στα γονίδια που συνθέτουν τις δομικές ενότητες και τα συμπλέγματα των δικτύων. Συγκεκριμένα, επιβεβαιώθηκαν τέσσερις αλληλεπιδράσεις στις δομικές ενότητες του φυσιολογικού Bayesian δικτύου, οι FN1-COMP, FN1-CDKN2A, FN1-KRT16, και FGFR3-FGF18, ενώ επιβεβαιώθηκε μόνο η αλληλεπίδραση FN1-COMP στις δομικές ενότητες του καρκινικού Bayesian δικτύου. Επιπλέον, στα συμπλέγματα του καρκινικού Bayesian δικτύου δεν εντοπίστηκε καμία γνωστή βιολογική αλληλεπίδραση, ενώ στα συμπλέγματα του φυσιολογικού Bayesian δικτύου επικυρώθηκε μόνο η αλληλεπίδραση FN1-KRT16. Η εικόνα αυτή συνάδει με την υφιστάμενη αναφορά του συνόλου των αλληλεπιδράσεων που έχουν σχολιαστεί έως σήμερα για τα 82 γονίδια, όπως προέκυψε από τις βάσεις δεδομένων BioGRID, και HIPPIE, καθώς και τη βάση δεδομένων G2SBC (Τα γονίδια στα Συστήματα του καρκίνου του μαστού, The Genes to Systems Breast Cancer) [109]. Όπως αποτυπώνεται στην Εικόνα D.1 και D.2 του Παραρτήματος D, 36 γονίδια από τα 82 υπολείπονται οποιαδήποτε αλληλεπίδραση. Συνεπώς, η πλειονότητα των αλληλεπιδράσεων που προσδιορίζονται και στα δυο δίκτυα,

μπορούν να χαρακτηριστούν ως νέες αλληλεπιδράσεις με δυνατότητα πρόβλεψης, οι οποίες ωστόσο θα πρέπει να επαληθευθούν πειραματικά.

Σημαντικό στοιχείο των δικτύων αποτελούν οι κεντρικοί κόμβοι. Ο βαθμός ως μια μετρική τοπολογία αντιπροσωπεύει τον αριθμό των συνδέσεων που έχει ένας κόμβος με έναν άλλο κόμβο στο υποδίκτυο. Ο βαθμός είναι μια σημαντική μετρική στα βιολογικά δίκτυα, η οποία αναδεικνύει γονίδια με υψηλή συνδεσιμότητα που πιθανά να διαδραματίζουν σημαντικό ρόλο στην ανάπτυξη της νόσου [110]. Επιπλέον, για τα γονίδια ασθένειας αναφέρεται από τους Cai και συν (2010) ότι βρίσκονται σε ευάλωτες θέσεις στα δίκτυα, γεγονός που αποτελεί την αιτία των αναγνωρίσιμων φαινοτύπων της νόσου που συνοδεύεται από την αποδιοργάνωσή τους [110]. Τα μόρια που εμφανίζονται σε μία κεντρική θέση ως εξαιρετικά συνδεδεμένοι καρκινικοί 'κεντρικοί κόμβοι' (hubs) έχουν συσχετιστεί στην πλειοψηφία τους (έξι από τα επτά γονίδια) με τον καρκίνο του μαστού. Πράγματι, τέτοιου είδους κεντρικές θέσεις έχουν αναφερθεί για τον καρκίνο, όπου τα γονίδια ασθένειας τείνουν να κωδικοποιούν κόμβους. Αντίστοιχα, τα γονίδια που εμφανίζονται σε μία κεντρική θέση ως εξαιρετικά συνδεδεμένοι φυσιολογικοί 'κεντρικοί κόμβοι' (hubs) δεν έχουν συσχετιστεί ακόμα, στην πλειοψηφία τους (έξι από τα 11 γονίδια) με τον καρκίνο του μαστού.

Για να αξιολογηθεί η λειτουργική συνάφεια των Bayesian δικτύων, μελετήθηκαν τα συμπλέγματα και οι δομικές ενότητες των δυο Bayesian δικτύων. Τα συμπλέγματα ή οι δομικές ενότητες που συνθέτουν το κάθε δίκτυο είναι συχνά πυκνά συνδεδεμένα/ες μεταξύ τους αντιπροσωπεύοντας λειτουργικές οντότητες στις οποίες τα εμπλεκόμενα γονίδια μπορεί να σχετίζονται μεταξύ τους όσον αφορά το λειτουργικό τους ρόλο και/ή να συνεργάζονται μεταξύ τους για την εκτέλεση ορισμένων βιολογικών διεργασιών.

Πρώτα απ' όλα διεξήχθη η ανάλυση λειτουργικού εμπλουτισμού των επιλεγμένων γονιδίων για κάθε σύμπλεγμα και μοριακή δομική ενότητα που προήλθε από την ομαδοποίηση των δυο Bayesian δικτύων, χρησιμοποιώντας το εργαλείο ανάλυσης για σύνολα γονιδιακών δεδομένων WebGestalt (Web-based Gene Set Analysis Toolkit) και επιλέγοντας την ανάλυση για την κατηγορία των μοριακών μονοπατιών με όρους της KEGG (Εγκυκλοπαίδεια Γονιδίων και Γονιδιώματος του Κιότο) και της WikiPathways (Νέο Μοντέλο Δημόσιας Βάσης Δεδομένων για τα Βιολογικά Μονοπάτια). Το εργαλείο WebGestalt εφαρμόζει το υπεργεωμετρικό τεστ στα επιλεγμένα γονίδια για τον εμπλουτισμό των όρων της KEGG και της WikiPathways, και τη μέθοδο Benjamini and Hochberg (BH) για την πολλαπλή ρύθμιση του τεστ (adjP) [111]. Τα αποτελέσματα της ανάλυσης εμπλουτισμού παρατίθενται στο Παράρτημα E (Εικόνες E.1, E.2, E.3, E.4).

Αναλυτικότερα, η εύρεση εμπλουτισμένων μονοπατιών KEGG και Wikipathways των συμπλεγμάτων και των δομικών ενοτήτων υποστηρίζουν την υφιστάμενη γνώση σχετικά με τους λειτουργικούς ρόλους των μονοπατιών που εμπλέκονται στον καρκίνο του μαστού, καθώς και στην εξέλιξη της νόσου. Για παράδειγμα, η αποκλίνουσα λειτουργία (από τη φυσιολογική λειτουργία) του σηματοδοτικού μονοπατιού EGF-EGFR έχει συσχετιστεί με ανεξέλεγκτη κυτταρική ανάπτυξη, αντίσταση στην απόπτωση

ση, αγγειογένεση, κυτταρική εισβολή και μετάσταση.

Επιπλέον, ένα σημαντικό στοιχείο των δομικών ενοτήτων κυρίως, ήταν ο εμπλουτισμός των ίδιων μονοπατιών και στα δυο δίκτυα που ήταν εν μέρει αναμενόμενος λόγω των κοινών μορίων που συμμετέχουν στα δίκτυα. Ωστόσο, παρατηρήθηκε απώλεια ή προσθήκη συγκεκριμένων γονιδίων σε κάποια από αυτά τα μονοπάτια σε κάθε δίκτυο (Παράρτημα Ε (Εικόνες Ε.1, Ε.2, Ε.3, Ε.4). Συνολικά οκτώ γονίδια (*EGF*, *ERBB2*, *FGF18*, *CDKN2A*, *NR4A1*, *CPA3*, *TUBB3*, *FGFR3*) βρέθηκε ότι υπολείπονται ή προστίθενται στα μονοπάτια των δυο δικτύων, οδηγώντας αντίστοιχα σε μεταβολές αλληλεπιδράσεων ανάμεσα στα μόρια των δικτύων. Αξίζει να τονιστεί, ότι έξι από αυτά τα γονίδια (*EGF*, *ERBB2*, *FGF18*, *CDKN2A*, *CPA3*, *FGFR3*) συμμετέχουν στα εμπλουτισμένα μονοπάτια του δικτύου διαφοροποίησης (Εικόνα 4.31).

ΜΟΝΟΠΑΤΙΑ KEGG και WikiPathways στο Υποδίκτυο Διαφοροποίησης					
MCODE - ΣΥΜΠΛΕΓΜΑΤΑ			Μέσος Όρος Ενδιάμεσης Κεντρικότητας	Μέσος Όρος Κεντροακρότητας Εγγύτητας	Μέσος Όρος Κεντροακρότητας Παθιμιάς
MCODE - ΣΥΜΠΛΕΓΜΑ 1			18	0.021	12.5
MCODE - ΣΥΜΠΛΕΓΜΑ 2			23.78	0.02	7.14
MCODE - ΣΥΜΠΛΕΓΜΑ 3			5.7	0.006	6
MCODE - ΣΥΜΠΛΕΓΜΑ 4			2.3	0.14	2.7
MCODE - ΜΟΝΟΠΑΤΙΑ	ΜΟΡΙΑ ΜΟΝΟΠΑΤΙΑ	ΣΥΝΘΕΤΟ ΓΟΝΙΔΙΟ	Μέσος Όρος Ενδιάμεσης Κεντρικότητας	Μέσος Όρος Κεντροακρότητας Εγγύτητας	Μέσος Όρος Κεντροακρότητας Παθιμιάς
Μονοπάτι 1_1	Pathways in cancer	ERBB2 FGFR3 CDKN2A FGF18 EGF	23.92	0.021	14.2
Μονοπάτι 1_2	Focal adhesion	ERBB2 COL11A1 COMP EGF	23.73	0.021	13.75
Μονοπάτι 1_3	ErbB signaling pathway	NRG2 ERBB2 EGF	27.52	0.022	14.66
Μονοπάτι 1_4	Regulation of actin cytoskeleton	FGFR3 FGF18 EGF	21.79	0.021	13.33
Μονοπάτι 1_5	MAPK signaling pathway	FGFR3 FGF18 EGF	21.79	0.021	13.33
Μονοπάτι 1_6	EGF-EGFR Signaling Pathway	ERBB2 REPS2 EGF	27.38	0.022	14.33
Μονοπάτι 1_7	ECM-receptor interaction	COL11A1 COMP	20.33	0.021	13
Μονοπάτι 1_8	Endocytosis	FGFR3 EGF	26.95	0.022	14
Μονοπάτι 1_9	Endochondral Ossification	FGFR3 FGF18	21.52	0.021	13.5
Μονοπάτι 2_1	Protein digestion and absorption	COL17A1 CPA3	39.45	0.021	9
Μονοπάτι 2_2	Chemokine signaling pathway	CCL19 CCL18	15.35	0.019	6
Μονοπάτι 2_3	Metabolic pathways	TAT ATP6V0A4 HSD17B2	33.48	0.021	8.66
Μονοπάτι 2_4	Androgen receptor signalling pathway	BRCA1 PARK7	19.47	0.019	6

Εικόνα 4.31 – Απεικόνιση των σημαντικότερων ($p \leq 0.05$) μοριακών μονοπατιών KEGG και WikiPathways που προέκυψαν από την ανάλυση εμπλουτισμού (WebGestalt) των γονιδίων που συνθέτουν τα συμπλέγματα MCODE του δικτύου διαφοροποίησης. Για κάθε μονοπάτι του συμπλέγματος αναφέρεται ο μέσος όρος της ενδιάμεσης κεντρικότητας, της κεντρικότητας εγγύτητας και της κεντρικότητας βαθμού των γονιδίων που συνθέτουν το εκάστοτε μονοπάτι.

Όπως αναφέρθηκε στην Ενότητα 4.3, κατασκευάστηκε το δίκτυο διαφοροποίησης, στο οποίο εντοπίστηκαν όλοι οι ‘κεντρικοί κόμβοι’ (hubs) των μπαγεσιανών δικτύων. Επίσης, υπολογίστηκαν οι κεντρικότητες των γονιδίων που συμμετέχουν στα συμπλέγματα του δικτύου διαφοροποίησης (Εικόνα 4.12), αλλά και οι κεντρικότητες των γονιδίων που συμμετέχουν στα 13 συνολικά μονοπάτια (Εικόνες 4.13, 4.14, 4.15 και 4.31), τα οποία βρέθηκαν σημαντικά ($p \leq 0.05$) στα συμπλέγματα 1 και 2 μέσα

από την ανάλυση εμπλουτισμού (WebGestalt) [111]. Η σημασία του υπολογισμού των κεντρικοτήτων έγκειται στην ικανότητα που παρουσιάζει ένα γονίδιο ως μια βιολογικά λειτουργική οντότητα, να διατηρεί τη συνοχή των επικοινωνούντων κόμβων στο δίκτυο. Όσο υψηλότερη είναι η τιμή, τόσο υψηλότερη είναι η συνάφεια των γονιδίων στη σύνδεση ρυθμιστικών μορίων και τη δημιουργία λειτουργικών συμπλεγμάτων [95]. Ειδικότερα, για κάθε σύμπλεγμα ή κάθε σημαντικό μονοπάτι του συμπλέγματος υπολογίστηκε ο μέσος όρος των ενδιάμεσων κεντρικοτήτων, των κεντρικοτήτων εγγύτητας και των κεντρικοτήτων βαθμού των γονιδίων που συνθέτουν το εκάστοτε μονοπάτι ή το σύμπλεγμα (Εικόνα 4.31).

Συνολικά, στην Εικόνα 4.31 φαίνεται ότι οι μεγαλύτερες τιμές παρουσιάζονται:

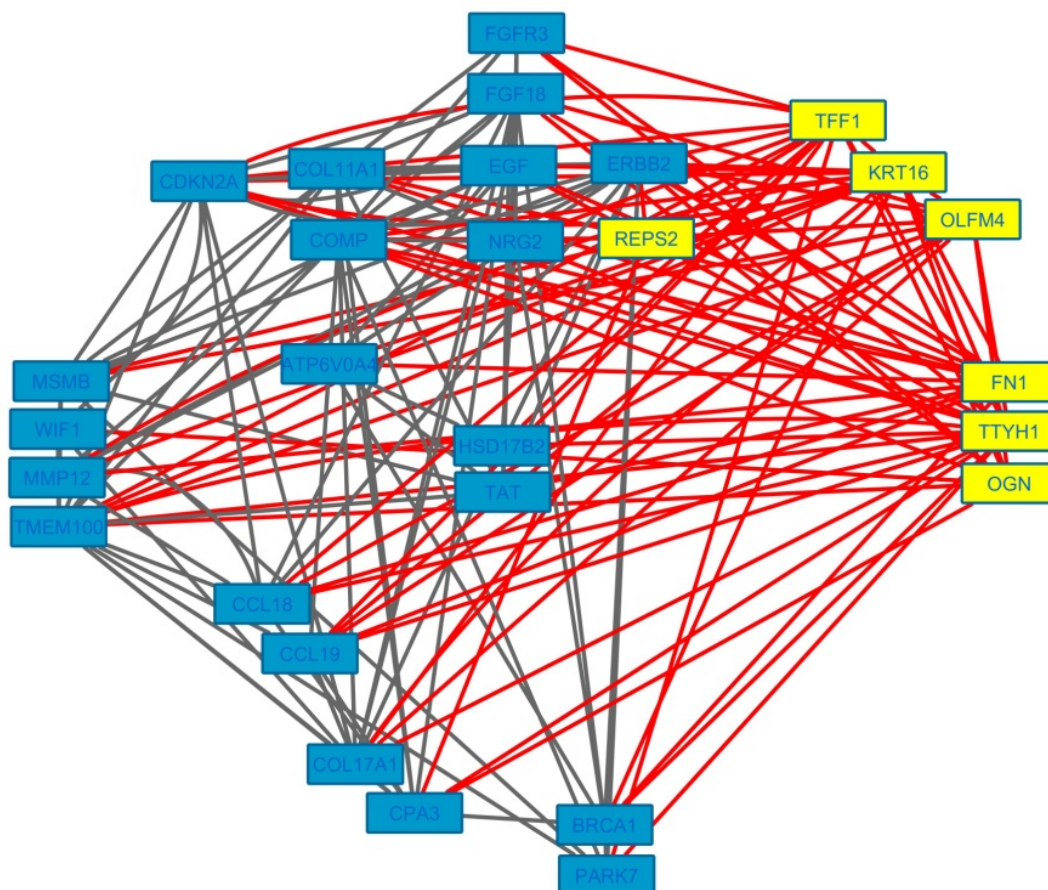
1. στο σύμπλεγμα 2, στο μονοπάτι σηματοδότησης ErbB, το μονοπάτι σηματοδότησης EGF-EGFR, τη διεργασία πέψης και απορρόφησης των πρωτεϊνών και τα μεταβολικά μονοπάτια για την ενδιάμεση κεντρικότητα,
2. στο σύμπλεγμα 4, στο μονοπάτι σηματοδότησης ErbB, το μονοπάτι σηματοδότησης EGF-EGFR και την ενδοκυττάρωση για την κεντρικότητα εγγύτητας, και
3. στο σύμπλεγμα 1, στα μονοπάτια στον καρκίνο, το μονοπάτι σηματοδότησης ErbB, το μονοπάτι σηματοδότησης EGF-EGFR και την ενδοκυττάρωση για την κεντρικότητα βαθμού.

Η τοπολογική κεντρικότητα, όπως αναφέρθηκε παραπάνω, χρησιμοποιήθηκε για να χαρακτηρίσει τη βιολογική σημασία των συμπλεγμάτων του δικτύου διαφοροποίησης καθώς και των μονοπατιών μέσα σ' αυτά. Τα αποτελέσματα έδειξαν ότι, το σύμπλεγμα 1, σύμπλεγμα 2 και σύμπλεγμα 3 ήταν σημαντικά για την κεντρικότητα βαθμού, την ενδιάμεση κεντρικότητα και την κεντρικότητα εγγύτητας, αντίστοιχα. Οι τρεις κεντρικότητες βρέθηκαν συνολικά σημαντικές για τα μονοπάτια σηματοδότησης ErbB και EGF-EGFR στο σύμπλεγμα 1, και για τη διεργασία πέψης και απορρόφησης των πρωτεϊνών και τα μεταβολικά μονοπάτια στο σύμπλεγμα 2.

Σε μια προσπάθεια αξιολόγησης της λειτουργικής συνάφειας του δικτύου διαφοροποίησης και του ρόλου του στον καρκίνο του μαστού, κατασκευάστηκαν δυο υποδίκτυα διαφοροποίησης με:

- όλους τους κεντρικούς κόμβους, 'καρκινικούς' και 'φυσιολογικούς', όπως αναδείχθηκαν από την κατασκευή των Bayesian δικτύων, και
- όλους τους κόμβους που συμμετέχουν στα εμπλουτισμένα μοριακά μονοπάτια του δικτύου διαφοροποίησης αλλά με
- διαφορετικές ακμές, αφού επιλέχθηκαν οι παρακείμενες ακμές των κεντρικών καρκινικών κόμβων για το καρκινικό υποδίκτυο (Εικόνα 4.32) και οι παρακείμενες ακμές των κεντρικών φυσιολογικών κόμβων για το φυσιολογικό υποδίκτυο (Εικόνα 4.33).

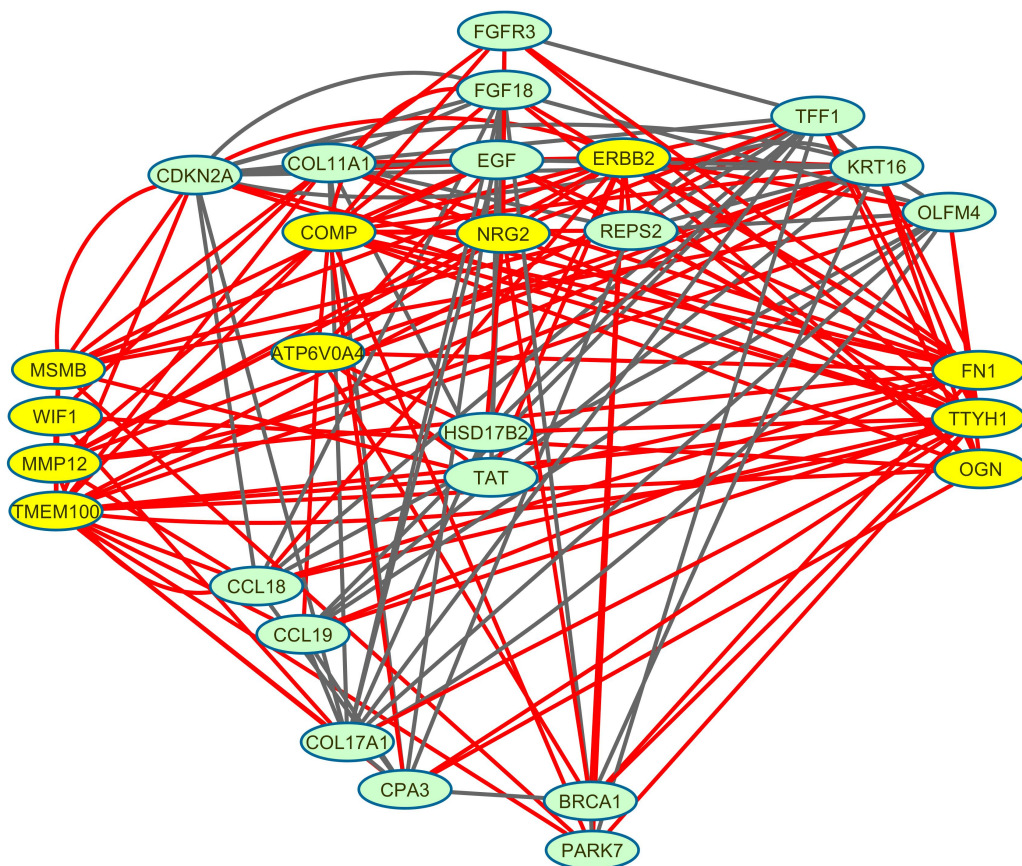
Όπως φαίνεται από τις Εικόνες 4.32 και 4.33 των δυο υποδικτύων με κεντρικούς κόμβους, αξίζει να σημειωθούν τα εξής:



Εικόνα 4.32 – Καρκινικό υποδίκτυο διαφοροποίησης με κεντρικούς καρκινικούς κόμβους. Απεικόνιση των 82 γονιδιακών/πρωτεϊνικών αλληλεπιδράσεων ανάμεσα στους κεντρικούς καρκινικούς κόμβους (7 κόμβοι σε κίτρινο χρώμα) και τα γονίδια/πρωτεΐνες που συμμετέχουν σε δεκατρία σημαντικά ($p \leq 0.05$) μοριακά μονοπάτια στο δίκτυο διαφοροποίησης. Οι κόκκινες ακμές αντιπροσωπεύουν τις παρακείμενες ακμές των κεντρικών καρκινικών κόμβων με τα γονίδια/πρωτεΐνες των μοριακών μονοπατιών. Για την οπτικοποίηση των αλληλεπιδράσεων του υποδικτύου με κεντρικούς καρκινικούς κόμβους χρησιμοποιήθηκε η πλατφόρμα Cytoscape.

1. τα γονίδια FN1, TTYH1 και OGN αποτελούν τους κοινούς κεντρικούς κόμβους των δυο υποδικτύων, και
2. ο αριθμός των αλληλεπιδράσεων μειώνεται στο καρκινικό υποδίκτυο σε σχέση με το φυσιολογικό υποδίκτυο, παρόλο που ο αριθμός των μορίων παραμένει σταθερός (Παράρτημα F, Εικόνες F.1 και F.2).

Σύμφωνα με τα παραπάνω, η παθολογική κατάσταση, ο καρκίνος του μαστού, διαφοροποιείται από τη φυσιολογική μέσα από τον αριθμό και το είδος των αλληλεπιδράσεων που λαμβάνουν χώρα ανάμεσα στα γονίδια και/ή τα γονιδιακά προϊόντα, αλλά και την τοπολογική βαρύτητα που κατέχουν τα κεντρικά μόρια σε κάθε δίκτυο, οδηγώντας σε λειτουργικές αλλοιώσεις (Παράρτημα F, Εικόνες F.1 και F.2). Η απορρύθμιση κάποιων μορίων ή/και αλληλεπιδράσεων στο φυσιολογικό υποδίκτυο



Εικόνα 4.33 – Φυσιολογικό υποδίκτυο διαφοροποίησης με κεντρικούς φυσιολογικούς κόμβους. Απεικόνιση των 135 γονιδιακών/πρωτεϊνικών αλληλεπιδράσεων ανάμεσα στους κεντρικούς φυσιολογικούς κόμβους (11 κόμβοι σε κίτρινο χρώμα) και τα γονίδια/πρωτεΐνες που συμμετέχουν σε δεκατρία σημαντικά ($p \leq 0.05$) μοριακά μονοπάτια στο δίκτυο διαφοροποίησης. Οι κόκκινες ακμές αντιπροσωπεύουν τις παρακαείμενες ακμές των κεντρικών φυσιολογικών κόμβων με τα γονίδια/πρωτεΐνες των μοριακών μονοπατιών. Για την οπτικοποίηση των αλληλεπιδράσεων του υποδικτύου με κεντρικούς φυσιολογικούς κόμβους χρησιμοποιήθηκε η πλατφόρμα Cytoscape.

και η ενδυνάμωση κάποιων άλλων μορίων και αλληλεπιδράσεων στο καρκινικό υποδίκτυο φαίνεται πως ενέχονται στους μηχανισμούς της ογκογένεσης, της εξέλιξης της νόσου και της μετάστασης.

Με αυτή την έννοια, μπορούμε να υποστηρίξουμε ότι τα υποδίκτυα διαφοροποίησης μπορούν να αποτελέσουν πρότυπα για την εξέταση των μεταβολών ανάμεσα σε δυο καταστάσεις, ενώ οι κεντρικοί καρκινικοί κόμβοι (FN1, OGN, TTYH1, KRT16, TFF1, OLFM4, REPS2) και οι κεντρικοί φυσιολογικοί κόμβοι (FN1, OGN, TTYH1, ERBB2, COMP, ATP6V0A4, WIF1, MMP12, NRG2, MSMB, TMEM100) αποτελούν πιθανούς γονιδιακούς δείκτες για τον καρκίνο του μαστού.

Κεφάλαιο 5

Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήσαμε την δομή Bayesian δικτύων που προέκυψαν από μια γονιδιακή υπογραφή συσχετισμένη με τον καρκίνο του μαστού. Τα αποτελέσματα που προέκυψαν μας δίνουν μια πειστική απάντηση για την χρήση των δικτύων, και ειδικά των Bayesian, στις βιολογικές επιστήμες.

- Η δομή των δικτύων, που προκύπτει ακολουθώντας την μέθοδο που παρουσιάζεται σε αυτήν την εργασία, φαίνεται να έχει κάποιες σημαντικές ιδιότητες που έχουν παρατηρηθεί στα βιολογικά και γενικά στα πραγματικά δίκτυα. Σε αυτήν την μελέτη επιβεβαιώνεται μια από αυτές τις ιδιότητες που είναι αυτή των δικτύων ελεύθερης κλίμακας.
- Στην δομή των δικτύων εντοπίστηκαν σημαντικοί κόμβοι (Hubs), μερικοί από τους οποίους εμφανίζονται στις περισσότερες από τις δομές δικτύων που κατασκευάσαμε.
- Στα δίκτυα και υποδίκτυα που κατασκευάσαμε υπάρχουν σημαντικά βιολογικά μονοπάτια.

Με αφετηρία την ‘77 κοινή γονιδιακή υπογραφή’ κατασκευάστηκαν δυο Bayesian δίκτυα, ένα για τα υγιή άτομα και ένα για τα νοσούντα άτομα από καρκίνο του μαστού, από όπου αναδείχθηκαν τρία κοινά κεντρικά γονίδια, *FN1*, *OGN*, και *TTYH1*, μαζί με οκτώ ‘φυσιολογικούς’ κεντρικούς κόμβους (*ERBB2*, *COMP*, *ATP6V0A4*, *WIF1*, *MMP12*, *NRG2*, *MSMB*, *TMEM100*) και τέσσερις κεντρικούς ‘καρκινικούς’ κόμβους (*KRT16*, *TFF1*, *OLFM4*, *REPS2*).

Το δίκτυο διαφοροποίησης που δημιουργήθηκε χαρακτηρίστηκε από σημαντικά τοπολογικά χαρακτηριστικά με βάση τα τρία διαφορετικά μέτρα κεντρικότητας (του βαθμού δηλ. του αριθμού συσχετίσεων με άλλους κόμβους, της εγγύτητας δηλ. του μήκους των συντομότερων μονοπατιών όλων των ζευγών των κόμβων, και του ενδιάμεσου δηλ. του αριθμού των συντομότερων διαδρομών από όλες τις κορυφές σε όλες τις άλλες που περνούν μέσα από αυτόν τον κόμβο).

Η αναζήτηση του λειτουργικού ρόλου του δικτύου διαφοροποίησης οδήγησε στην κατασκευή των υποδικτύων διαφοροποίησης, αναδεικνύοντας νέες λειτουργικές αλληλεπιδράσεις ανάμεσα στα κεντρικά μόρια και τα μόρια σημαντικών μονοπατιών, όπως

των μονοπατιών σηματοδότησης ErbB και EGF-EGFR, καθώς και των μεταβολικών μονοπατιών και της διεργασίας πέψης και απορρόφησης των πρωτεϊνών.

Αναγνωρίζοντας ότι μια παρεκκλίνουσα ενεργοποίηση των μονοπατιών, όπως αυτή του μονοπατιού EGF-EGFR, μπορεί να πυροδοτήσει διεργασίες που οδηγούν στον καρκίνο του μαστού, παρέχονται πληροφορίες σχετικά με τις μεταβολές των αλληλεπιδράσεων ανάμεσα στα δυο υποδίκτυα διαφοροποίησης.

Τα υποδίκτυα διαφοροποίησης αποτελούμενα από σημαντικούς κεντρικούς κόμβους, ‘καρκινικούς’ και ‘φυσιολογικούς’, των Bayesian δικτύων και μόρια που συμμετέχουν στα εμπλουτισμένα μοριακά μονοπάτια του δικτύου διαφοροποίησης, παρέχουν μια εικόνα δυναμικών αλλαγών ανάμεσα στα δυο δίκτυα που μπορεί να διαδραματίζουν σημαντικό ρόλο στην εξέλιξη της νόσου.

Οι κεντρικοί κόμβοι των Bayesian δικτύων προτείνονται ως πιθανοί γονιδιακοί δείκτες για τον καρκίνο του μαστού.

Βιβλιογραφία

- [1] Z.P.Liu, Y.Wang, X.S.Zhang, L.Chen, "Network-based analysis of complex diseases", *Systems Biology IET*, vol.6 , pp. 22-33, 2012.
- [2] M.Krauthammer, C.A.Kaufmann, T.C.Gilliam, A.Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease", *Proc. Natl. Acad. Sci. USA*, vol.101, pp. 15148–15153, 2004.
- [3] J.L.Morrison, R.Breitling, D.J.Higham, D.R.Gilbert, "GeneRank: using search engine technology for the analysis of microarray experiments", *BMC Bioinf.*, p. 233, vol.6, 2005.
- [4] L.Franke, H.van Bakel, L.Fokkens, E.D.de Jong, M.Egmont-Petersen, C.Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes", *Am. J. Hum. Genet.*, vol.78, pp. 1011–1025, 2006.
- [5] R.A.George, J.Y.Liu, L.L.Feng, R.J.Bryson-Richardson, D.Fatkin, M.A.Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction", *Nucleic Acids Res.*, vol.34, p. e130, 2006.
- [6] S.Aerts, D.Lambrechts, S.Maity, "Gene prioritization through genomic data fusion", *Nat. Biotechnol.*, vol.24, pp. 537–544, 2006.
- [7] X.Ma, H.Lee, L.Wang, F.Sun, "CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data", *Bioinformatics*, vol.23, pp. 215–221, 2007.
- [8] K.Lage, E.O.Karlberg, Z.M.Storling, "A human phenomeinteractome network of protein complexes implicated in genetic disorders", *Nat. Biotechnol.*, vol.25, pp. 309–316, 2007.
- [9] S.Kohler, S.Bauer, D.Horn, P.N.Robinson, "Walking the interactome for prioritization of candidate disease genes", *Am. J. Hum. Genet.*, vol.82, pp. 949–958, 2008.
- [10] X.Wu, R.Jiang, M.Q.Zhang, S.Li, "Network-based global inference of human disease genes", *Mol. Syst. Biol.*, vol.4, p. 189, 2008.

- [11] T.Ideker, O.Ozier, B.Schwikowski, A.F.Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks", *Bioinformatics*, vol.18, pp. S233-S240, 2002.
- [12] M.Steffen, A.Petti, J.Aach, P.D'haeseleer, G.Church, "Automated modelling of signal transduction networks", *BMC Bioinf.*, vol.3, p. 34, 2002.
- [13] E.Segal, H.Wang, D.Koller, "Discovering molecular pathways from protein interaction and gene expression data", *Bioinformatics*, vol.19, pp. i264-i272, 2003.
- [14] F.Sohler, D.Hanisch, R.Zimmer, "New methods for joint analysis of biological networks and expression data", *Bioinformatics*, vol.20, pp. 1517-1521, 2004.
- [15] R.Breitling, A.Amtmann, P.Herzyk, "Graph-based iterative group analysis enhances microarray interpretation", *BMC Bioinf.*, vol.5, p. 100, 2004.
- [16] J.Scott, T.Ideker, R.M.Karp, R.Sharan, "Efficient algorithms for detecting signalling pathways in protein interaction networks", *J. Comput. Biol.*, vol.13, pp. 133-144, 2006.
- [17] M.Liu, A.Liberzon, S.W.Kong, "Network-based analysis of affected biological processes in type 2 diabetes models", *PLoS Genet.*, vol.3, p. e96, 2007.
- [18] S.Nacu, R.Critchley-Thorne, P.Lee, S.Holmes, "Gene expression network analysis and applications to immunology", *Bioinformatics*, vol.23, pp. 850-858, 2007.
- [19] G.Bebek, J.Yang, "PathFinder: mining signal transduction pathway segments from protein-protein interaction networks", *BMC Bioinf.*, vol.8, p. 335, 2007.
- [20] M.T.Dittrich, G.W.Klau, A.Rosenwald, T.Dandekar, T.Muller, "Identifying functional modules in protein-protein interaction networks: an integrated exact approach", *Bioinformatics*, vol.24, pp. i223-i231, 2008.
- [21] X.M.Zhao, R.S.Wang, L.Chen, K.Aihara "Uncovering signal transduction networks from high-throughput data by integer linear programming", *Nucleic Acids Res.*, 36, p. e48, 2008.
- [22] I.Ulitsky, R.Shamir, "Identifying functional modules using expression profiles and confidence-scored protein interactions", *Bioinformatics*, vol.25, pp. 1158-1164, 2009.
- [23] Y.Q.Qiu, S.Zhang, X.S.Zhang, L.Chen, "Identifying differentially expressed pathways via a mixed integer linear programming model", *IET Syst. Biol.*, vol.3, pp. 475-486, 2009.

- [24] Z.P.Liu, Y.Wang, T.Wen, X.S.Zhang, W.Xia, L.Chen, "Dynamically dysfunctional protein interactions in the development of Alzheimer's disease.", Proc. IEEE Int. Conf. on Systems, Man and Cybernetics, San Antonio, USA, pp. 4262-4267, 2009.
- [25] DNA - Χρωμοσώματα <http://www.news-medical.net/health/What-are-Genes.aspx>
- [26] Γονίδιο <http://el.wikipedia.org/wiki>
- [27] G.Patrinou, Γονιδιωματοική <http://www.pharmacy.upatras.gr/index.php/el/studies/undergraduate/-mainmenu-63>
- [28] T.Vigdideli, "Mixture of Experts Approach in Gene Classification", Technical University of Crete, Electronics and Computer Engineering Department, 2012.
- [29] M.Sonachalam, J.Shen, H.Huang, X.Wu, "Systems biology approach to identify gene network signatures for colorectal cancer", <http://www.frontiersin.org>, vol.3, Article 80, 2012.
- [30] U.Alon, N.Barkai, DA.Notterman, K.Gish, S.Ybarra, D.Mack, AJ.Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", vol.96, pp.6745-6750, 1999.
- [31] TR.Golub, DK.Slonim, P.Tamayo, C.Huard, M.Gaasenbeek, J.Mesirov, H.Coller, M.Loh, J.Downing, M.Caligiuri, C.Bloomfield, ES.Lander "Molecular classification of cancer", class discovery and class prediction by gene expression monitoring, vol.286, pp. 531-537, Science 1999.
- [32] D.Singh, PG.Febbo, K.Ross, DG.Jackson, J.Manola, C.Ladd, P/Tamayo, AA.Renshaw, AV.D'Amico, JP.Richie, ES.Lander, M.Loda, PW.Kanto, TR.Golub, WR.Sellers, "Gene expression correlates of clinical prostate cancer behavior" vol.1, pp. 203-209, Cancer Cell, 2002.
- [33] LJ.Van't Veer, H.Dai, MJ.Vijver, De.Van, YD.He, AAM.Hart, M.Mao, HL.Peterse, K.Kooy Van Der, MJ.Marton, AT.Witteveen, GJ.Schreiber, RM.Kerkhoven, C.Roberts, PS.Linsley, R.Bernards, SH.Friend, "Gene expression profiling predicts clinical outcome of breast cancer", Nature, vol.415, pp. 530-536, 2002.
- [34] Y.Wang, JG.Klijn, Y.Zhang, AM.Sieuwerts, MP.Look, F.Yang, D.Talantov, M.Timmermans, MF.Meijer-van Gelder, J.Yu, T.Jatkoe, EM.Berns, D.Atkins, JA.Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.", Lancet, vol.365, pp. 671-679, 2005.
- [35] J.Bogaerts, F.Cardoso, M.Buyse, S.Braga, S.Loi, JA.Harrison, J.Bines, S.Mook, N.Decker, P.Ravdin, P.Therasse, E.Rutgers, LJ.Van't Veer, M.Piccart,

- "Gene signature evaluation as a prognostic tool", challenges in the design of the MINDACT trial, vol.3, pp.540-551, 2006.
- [36] R.Nilsson, J.Bjorkegren, J.Tegner, "Open Access On reliable discovery of molecular signatures", 2009.
- [37] ME.Cusick, N.Klitgord, M.Vidal, DE.Hill, Interactome: gateway into systems biology. Hum Mol Genet 14 Spec vol. 2, pp. 171-181, 2005.
- [38] J.De Las Rivas, C.Fontanillo, "Protein-Protein Interactions Essentials:Key Concepts to Building and Analyzing Interactome Networks", PLoS Comput Biol, vol.6, 2010.
- [39] J.M.Roberts, T.A.Parlikar, T.Heldt, and G.C.Verghese, "Bayesian Networks for Cardiovascular Monitoring.", In Proceedings of the 28th IEEE EMBS Annual International Conference, pp. 205-209, New York City, USA, 2006.
- [40] M.R.Said, T.J.Begley, A.V.Oppenheim, D.A.Lauffenburger, L.D.Samson, "Global network analysis of phenotypic effects: Protein networks and toxicity modulation in *Saccharomyces cerevisiae*.", Proc. Natl. Acad. Sci., vol.101, pp. 18006-18011, 2004.
- [41] R.Shachar, L.Unger, M.Kupiec, R.Ruppin, R.Sharan, "A systems-level approach to mapping the telomere-length maintenance gene circuitry.", In Proc. RECOMB satellite workshop on systems biology, Columbia University, 2007.
- [42] S.Wachi, K.Yoneda, R.Wu, "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.", Bioinformatics, vol.21, pp. 4205-4208, 2005.
- [43] P.F.Jonsson, P.A.Bates, "Global topological features of cancer proteins in the human interactome.", Bioinformatics, vol.22, pp. 2291-2297, 2006.
- [44] K.I.Goh, M.E.Cusick, D.Valle, B.Childs, M.Vidal, A.L.Barabasi, "The human disease network.", Proc. Natl. Acad. Sci., vol.104, pp. 8685-8690, 2007.
- [45] M.Oti, H.G.Brunner, "The modular nature of genetic diseases." Clin. Genet., vol.71, pp. 1-11, 2007.
- [46] J.Quackenbush, "Microarray analysis and tumor classification.", N. Engl. J. Med., vol.354, pp. 2463-2472, 2006.
- [47] A.Ma'ayan, "Introduction to Network Analysis in Systems Biology", Sci. Signal., vol. 4, p. tr5, 2011.
- [48] P.J.F.Lucas, L.C.van der Gaag, and A.Abu-Hanna, "Bayesian networks in biomedicine and health-care.", Artificial Intelligence in Medicine vol.30, 201-214, 2004.

-
- [49] S.Wuchty, E.Ravasz, A.L.Barabasi, "The Architecture of Biological Networks", Complex Systems Science in Biomedicine Topics in Biomedical Engineering International Book Series, pp. 165-181, 2006.
- [50] G.Rutledge, G.Thomsen, I.Beinlich, B.Farr, B.Sheiner, and L.Fagan, "Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.", pp. 315-319, Washington, D.C., 1989.
- [51] W.Long, "Temporal reasoning for diagnosis in a causal probabilistic knowledge base.", Artificial Intelligence in Medicine vol.8, pp. 193-215, 1996.
- [52] M.E.J.Newman, "The Structure and Function Of Complex Networks", Department of Physics, University of Michigan, Ann Arbor, MI 48109, U.S.A. and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, U.S.A., p.48, 2003.
- [53] R.K.Ahuja, T.L.Magnanti, J.B.Oracle, "Network Flows: Theory, Algorithms, and Applications", Prentice-Hall, Upper Saddle River, NJ, 1993.
- [54] B.Bollobas, "The diameter of random graphs", Trans.Amer.Math.Soc., vol.267, pp. 41-52, 1981.
- [55] B.Bollobas, "Random Graphs", Academic Press, New York, vol.2, 2001.
- [56] F.Chung, L.Lu, "The average distances in random graphs with given expected degrees", Proc.Natl.Acad.Sci., vol.99, pp. 15879-15882, 2002.
- [57] S.N.Dorogovtsev, J.F.F.Mendes, A.N.Samukhin, "Metric Structure of Random Networks", Rua do Campo Alegre, vol.687, Porto, 2003.
- [58] A.Fronczak, P.Fronczak, J.A.Holyst, "Exact Solution for Average Path Length in Random Graphs.", Faculty of Physics and Center of Excellence for Complex Systems Research, Warsaw University of Technology, 2002.
- [59] M.E.J. Newman, "Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality", Phys.Rev.E, vol.64, 2001.
- [60] M.E.J.Newman, "The structure of scientific collaboration networks.", Proc.Natl.Acad.Sci.USA, vol.98, pp. 404-409, 2001.
- [61] R.Albert and A.L.Barabasi, "Statistical mechanics of complex networks.", Rev.Modern Phys., vol.74, pp. 47-97, 2002.
- [62] D.J.Watts, S.H.Strogatz, "Collective dynamics of "small-world" networks", Nature, vol.393, pp. 440-442, 1998.
- [63] J.Scott, "Social Network Analysis: A Handbook", vol.2, 2000.
- [64] S.N.Dorogovtsev, A.V.Goltsev, J.F.F.Mendes, "Pseudofractal scale-free web", Phys.Rev.E, vol.65, 2002.

- [65] G.Szabo, M.Alava, J.Kertesz, "Structural Transitions in Scale-Free Networks", Department of Theoretical Physics, Institute of Physics, Budapest University of Technology, 2002.
- [66] P.Erdos, A.Renyi, "On random graphs", Publ.Math.Debrecen, vol.6, pp. 290-297, 1959.
- [67] P.Erdos, A.Renyi, "On the evolution of random graphs", Magyar Tud.Akad.Mat.Kutato Int.Kozl., vol.5, pp. 17-61, 1960.
- [68] P.Erdos, A. Renyi, "On the strength of connectedness of a random graph", Acta Math. Acad. Sci. Hungar., vol.12, pp. 261-267, 1961.
- [69] R.Albert, H.Jeong, A.L.Barabasi, "Diameter of the world-wide web", Nature, vol.401, pp. 130-131, 1999.
- [70] A.L.Barabasi, R.Albert, H.Jeong, "Scale-free characteristics of random networks: The topology of the World Wide Web", Phys.A, vol.281, pp. 69-77, 2000.
- [71] A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, J.Wiener, "Graph structure in the web", Computer Networks, vol.33, pp. 309-320, 2000.
- [72] A.L.Barabasi, "Linked - How Everything is Connected to Everything Else and What it Means for Business, Science and Everyday Life", Scale-Free Networks, Scientific American, pp. 52 - 56, 2003.
- [73] K.I.Goh, B.Kahng, D.Kim, "Universal Behavior of Load Distribution in Scale-Free Networks", Phys.Rev.Lett., vol.87, 2001.
- [74] S.Narayanan, "The Betweenness Centrality Of Biological Networks", Faculty of the Virginia Polytechnic Institute and State University, 2005.
- [75] Wikipedia Betweenness Centrality http://en.wikipedia.org/wiki/Betweenness_centrality
- [76] Ι.Γρηγοριάδης, "Μέτρα σε δίκτυα-ιδιότητες και εφαρμογές", Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
- [77] C.J.Needham, J.R.Bradford, A.J.Bulpitt, D.R.Westhead, "A Primer on Learning in Bayesian Networks for Computational Biology", DOI: 10.1371/journal.pcbi.0030129 Featured in PLOS Collections, 2007.
- [78] N.Friedman, M.Linial, I.Nachman, D.Pe'er, "Using Bayesian Networks to Analyze Expression Data" Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, vol.7, pp. 601-620, 2000.
- [79] D.Koller, N.Friedman, "Probabilistic Graphical Models - Principles and Techniques", Massachusetts Institute of Technology, p. 1213 2009.

- [80] S.L.Lauritzen, N.Wermuth, "Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative", Ann. Statist., vol.17, pp. 31-57, 1989.
- [81] N.F.Marko, R.J.Weil, "Non-Gaussian Distributions Affect Identification of Expression Patterns, Functional Annotation, and Prospective Classification in Human Cancer Genomes", PLOS ONE, vol.7, 2012.
- [82] M.Cowley, K.Ying, "LogTransform Documentation", Garvan Institute, 2011.
- [83] H.Bouhamed, A.Masmoudi, T.Lecroq, A.Rebai, "A new approach for Bayesian classifier learning structure via K2 Algorithm", CCIS 304, pp. 387-393, 2012.
- [84] Z.Wei, H.Xu, W.Li, X.Gui, X.Wu, "Improved Bayesian Network Structure Learning with Node Ordering via K2 Algorithm", Intelligent Computing Methodologies Lecture Notes in Computer Science, vol.8589, pp. 44-55, 2014.
- [85] F.M.Al-Akwaa, M.M.Alkhawlani, "Comparison of the Bayesian Network Structure Learning Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, vol.2, 2012.
- [86] C.Chow, C.Liu, "Approximating discrete probability distributions with dependence trees.", IEEE Transactions on Information Theory, vol.14, pp.462-467, 1968.
- [87] D.Eppstein, "Spanning Trees and Spanners." Ch. 9 in Handbook of Computational Geometry, Amsterdam, pp. 425-461, 2000.
- [88] Cluster Analysis Wikipedia http://en.wikipedia.org/wiki/Cluster_analysis
- [89] S.Pemmaraju, S.Skiena, "Computational Discrete Mathematics: Combinatorics and Graph Theory in Mathematica.", Cambridge, England: Cambridge University Press, pp. 336-337, 2003.
- [90] M.Gardner, "Mathematical Magic Show: More Puzzles, Games, Diversions, Illusions and Other Mathematical Sleight-of-Mind", Scientific American. New York: Vintage, pp. 248-249, 1978.
- [91] J.B.Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." Proc. Amer. Math. Soc., vol.7, pp. 48-50, 1956.
- [92] S.Pemmaraju, S.Skiena, "Kruskals Algorithm." in Computational Discrete Mathematics: Combinatorics and Graph Theory in Mathematica., Cambridge University Press, pp. 336-338, 2003.
- [93] Αλγόριθμος Kruskal http://en.wikipedia.org/wiki/Kruskal%27s_algorithm
- [94] F.Gomez-Vela, N.Diaz, "Gene Network Biological Validity Based on Gene-Gene Interaction Relevance", The Scientific World Journal, vol.2014, Article ID 540679, p.11, 2014.

- [95] D.Zuang, L.Jiang, Q.He, P.Zhou, T.Yue, "Identification of hub subnetwork based on topological features of genes in breast cancer.", *International Journal of Molecular Medicine*, vol.35, pp. 664-674, 2015.
- [96] D.G.Bader, C.WV.Hogue, "An automated method for finding molecular complexes in large protein interaction networks", *BMC Bioinformatics*, vol.4, 2003.
- [97] GW.Flake, S.Lawrence, CL.Giles, FM.Coetzee, "Self-Organization of the Web and Identification of Communities.", *IEEE Computer*, vol.35, pp. 66-71, 2002.
- [98] AV.Goldberg, "Finding a Maximum Density Subgraph." Technical Report UCB/CSD University of California, 1984.
- [99] A.Ng, M.Jordan, Y.Weiss, "On spectral clustering: Analysis and an algorithm.", *Advances in Neural Information Processing Systems 14: Proceedings of the 2001*, 2001.
- [100] S.M.Kendall, A.Stuart, J.K.Ord, "Kendall's Advanced Theory of Statistics", vol.5, Oxford University Press, p. 446, 1987.
- [101] S.Kirkpatrick, C.D.Gelatt, M.P.Vecchi, "Optimization by Simulated Annealing.", *Science*, vol.220, pp. 671-680, 1983.
- [102] T.Ideker, O.Ozier, B.Schwikowski, A.F.Siegel, "Discovering regulatory and signaling circuits in molecular interaction networks.", *Suppl. 1*, vol.18, pp. 233-240, 2002.
- [103] S.Sfakianakis, E.S.Bei, M.Zervakis, D.Vassou, D.Kafetzopoulos, "On the Identification of Circulating Tumor Cells in Breast Cancer.", *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 773-782, 2004.
- [104] N.K.Chlis, S.Sfakianakis, E.S.Bei, M.Zervakis, "A Generic Framework for the Elicitation of Stable and Reliable Gene Expression Signatures.", In *Proc. IEEE BIBE*, Chania Greece, 2013
- [105] Expression Atlas update - a database of gene and transcript expression from microarray and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 2014. Διαθέσιμη στον ιστότοπο: <http://www.ebi.ac.uk/gxa>
- [106] M.Schummer, A.Green, J.D.Beatty, B.Y.Karlan, S.Karlan, J.Gross, S.Thornton, M.McIntosh, N.Urban, "Comparison of breast cancer to healthy control tissue discovers novel markers with potential for prognosis and early detection.", *PLoS One*, vol. 5, no. 2, pp. e9122, 2010.
- [107] <http://thebiogrid.org/>
- [108] <http://cbdm.mdc-berlin.de/tools/hippie/network.php>
- [109] <http://www.itb.cnr.it/breastcancer/>

-
- [110] J.J.Cai, E.Borenstein, D.A.Petrov, "Broker genes in human disease.", *Genome Biol Evol*, vol. 2, pp. 815-825, 2010.
- [111] J.Wang, D.Duncan, Z.Shi, B.Zhang, "WEB-based GEne SeT AnaLysis Toolkit", *Nucleic Acids Res*, vol. 41 (Web Server issue), pp. W77-83, 2013.

Παράρτημα Α

Σε αυτό το παράρτημα παραθέτουμε πίνακες με τα δεδομένα που επεξεργαστήκαμε για να δημιουργήσουμε τα δίκτυα όπως και τα αποτελέσματα μας σε πίνακες.

Στον πίνακα [Α.1](#) φαίνονται τα 77 γονίδια της γονιδιακής υπογραφής που χρησιμοποιήσαμε ως δεδομένα συν τα 5 γονίδια ελέγχου που προσθέσαμε εμείς (82 γονίδια συνολικά).

#γονίδιο	ID γονιδίου	γονίδιο	#γονίδιο	gene ID	γονίδιο	#γονίδιο	gene ID	γονίδιο	#γονίδιο	gene ID	γονίδιο
#1	1029	CDKN2A	#22	1908	EDN3	#43	3868	KRT16	#64	6362	CCL18
#2	10331	B3GNT3	#23	1950	EGF	#44	4069	LYZ	#65	6363	CCL19
#3	10381	TUBB3	#24	2064	ERBB2	#45	430	ASCL2	#66	6495	SIX1
#4	1047	CLGN	#25	2173	FABP7	#46	4321	MMP12	#67	6505	SLC1A1
#5	10481	HOXB13	#26	2261	FGFR3	#47	4435	CITED1	#68	6664	SOX11
#6	10562	OLFM4	#27	2302	FOXJ1	#48	4477	MSMB	#69	672	BRCA1
#7	10568	SLC34A2	#28	2335	FN1	#49	4916	NTRK3	#70	6898	TAT
#8	10648	SCGB1D1	#29	23532	PRAME	#50	4969	OGN	#71	7031	TFF1
#9	10891	PPARGC1A	#30	23635	SSBP2	#51	50617	ATP6V0A4	#72	7136	TNNI2
#10	11005	SPINK5	#31	26585	GREM1	#52	51442	VGLL1	#73	79785	RERGL
#11	11012	KLK11	#32	27074	LAMP3	#53	5179	PENK	#74	7980	TFPI2
#12	1118	CHIT1	#33	2938	GSTA1	#54	5304	PIP	#75	79919	C2orf54
#13	11197	WIF1	#34	2940	GSTA3	#55	54829	ASPN	#76	8483	CILP
#14	11315	PARK7	#35	29842	TFCP2L1	#56	55273	TMEM100	#77	8788	DLK1
#15	1301	COL11A1	#36	3084	NRG1	#57	55713	ZNF334	#78	8817	FGF18
#16	1308	COL17A1	#37	3164	NR4A1	#58	57348	TTYH1	#79	9073	CLDN8
#17	1311	COMP	#38	3294	HSD17B2	#59	57586	SYT13	#80	9185	REPS2
#18	1359	CPA3	#39	347902	AMIGO2	#60	58	ACTA1	#81	930	CD19
#19	1475	CSTA	#40	3500	IGHG1	#61	6278	S100A7	#82	9542	NRG2
#20	1811	SLC26A3	#41	3773	KCNJ16	#62	6280	S100A9			
#21	1907	EDN2	#42	3851	KRT4	#63	6286	S100P			

Εικόνα Α.1 – Τα 77 γονίδια της γονιδιακής υπογραφής που επεξεργαστήκαμε συν τα 5 γονίδια ελέγχου που προσθέσαμε (κόκκινο χρώμα).

Παράρτημα Β

Παρακάτω παραθέτονται οι πίνακες των αλληλεπιδράσεων όλων των δικτύων που προέκυψαν από την εκμάθηση δομής.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ - ΚΑΡΚΙΝΙΚΩΝ ΔΕΙΓΜΑΤΩΝ - ΔΙΑΚΡΙΤΩΝ ΜΕΤΑΒΛΗΤΩΝ – MWST ΤΑΞΙΝΟΜΗΣΗ											
γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2
B3GNT3	pp	CSTA	FN1	pp	COL11A1	MSMB	pp	CLGN	TNNI2	pp	BRCA1
B3GNT3	pp	S100A7	FN1	pp	GREM1	OGN	pp	RERGL	TNNI2	pp	REPS2
TUBB3	pp	FN1	PRAME	pp	SLC1A1	ATP6VOA4	pp	S100P	RERGL	pp	WIF1
OLFM4	pp	COL17A1	LAMP3	pp	CCL18	VGLL1	pp	TFF1	RERGL	pp	NTRK3
SLC34A2	pp	OLFM4	LAMP3	pp	CCL19	PENK	pp	DLK1	RERGL	pp	CCL19
SLC34A2	pp	CPA3	LAMP3	pp	CD19	PIP	pp	SCGB1D1	TFPI2	pp	MSMB
SLC34A2	pp	CLDN8	GSTA1	pp	TFCP2L1	ASPN	pp	OGN	CILP	pp	OGN
SCGB1D1	pp	ERBB2	GSTA3	pp	GSTA1	ASPN	pp	CILP	REPS2	pp	SYT13
WIF1	pp	KCNJ16	TFCP2L1	pp	TNNI2	TMEM100	pp	RERGL	CD19	pp	IGHG1
WIF1	pp	NRG2	NRG1	pp	AMIGO2	TTYH1	pp	VGLL1			
COL11A1	pp	COMP	IGHG1	pp	LYZ	ACTA1	pp	TUBB3			
COMP	pp	ASPN	KCNJ16	pp	EDN3	ACTA1	pp	NRG1			
CPA3	pp	FGF18	KCNJ16	pp	PENK	S100A7	pp	S100A9			
CSTA	pp	LAMP3	KRT16	pp	B3GNT3	S100A7	pp	SOX11			
SLC26A3	pp	PRAME	KRT16	pp	PPARGC1A	S100A9	pp	HSD17B2			
SLC26A3	pp	CITED1	KRT16	pp	FABP7	CCL18	pp	S100A9			
SLC26A3	pp	PIP	KRT16	pp	TMEM100	CCL19	pp	CD19			
EDN2	pp	ASCL2	LYZ	pp	MMP12	TAT	pp	SLC26A3			
EDN2	pp	ACTA1	ASCL2	pp	C2orf54	TFF1	pp	CDKN2A			
FABP7	pp	TTYH1	MMP12	pp	CCL18	TNNI2	pp	KRT16			

Εικόνα Β.1 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των καρκινικών δειγμάτων, διακριτών μεταβλητών και MWST ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ- ΚΑΡΚΙΝΙΚΩΝ ΔΕΙΓΜΑΤΩΝ- ΔΙΑΚΡΙΤΩΝ ΜΕΤΑΒΛΗΤΩΝ– CUSTOM ΤΑΞΙΝΟΜΗΣΗ											
γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2
CDKN2A	PP	SCGB1D1	FN1	PP	COL11A1	MSMB	PP	CLGN	TAT	PP	SLC26A3
CDKN2A	PP	ERBB2	FN1	PP	COMP	MSMB	PP	TFPI2	TFF1	PP	CITED1
CDKN2A	PP	PRAME	GSTA3	PP	GSTA1	OGN	PP	CPA3	TFF1	PP	SOX11
CDKN2A	PP	KRT16	NRG1	PP	ACTA1	OGN	PP	SSBP2	RERGL	PP	SPINK5
CDKN2A	PP	MSMB	IGHG1	PP	LYZ	OGN	PP	AMIGO2	RERGL	PP	COL17A1
CDKN2A	PP	TFF1	IGHG1	PP	CD19	OGN	PP	ASPN	RERGL	PP	KCNJ16
B3GNT3	PP	GSTA3	KCNJ16	PP	WIF1	OGN	PP	CILP	RERGL	PP	OGN
B3GNT3	PP	S100A9	KCNJ16	PP	EDN3	VGLL1	PP	SLC34A2	RERGL	PP	CCL19
OLFM4	PP	CLDN8	KRT16	PP	B3GNT3	VGLL1	PP	KRT4	FGF18	PP	CDKN2A
PPARGC1A	PP	KLK11	KRT16	PP	PPARGC1A	VGLL1	PP	SYT13	FGF18	PP	NRG1
COL17A1	PP	OLFM4	KRT16	PP	GSTA1	PENK	PP	DLK1	CD19	PP	CSTA
COMP	PP	ASPN	KRT16	PP	TFCP2L1	ASPN	PP	GREM1	CD19	PP	LAMP3
COMP	PP	CILP	KRT16	PP	MMP12	ASPN	PP	SLC1A1	CD19	PP	CCL19
COMP	PP	FGF18	KRT16	PP	TTYH1	TMEM100	PP	RERGL			
SLC26A3	PP	PIP	KRT16	PP	TNNI2	TTYH1	PP	FABP7			
EDN3	PP	NTRK3	KRT16	PP	REPS2	TTYH1	PP	VGLL1			
EDN3	PP	PENK	LYZ	PP	CCL18	TTYH1	PP	TMEM100			
FABP7	PP	HSD17B2	ASCL2	PP	C2orf54	ACTA1	PP	EDN2			
FABP7	PP	NRG2	MMP12	PP	CHIT1	S100A9	PP	S100A7			
FN1	PP	TUBB3	MMP12	PP	CCL18	S100P	PP	ATP6V0A4			

Εικόνα Β.2 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των καρκινικών δειγμάτων, διακριτών μεταβλητών και custom ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ - ΚΑΡΚΙΝΙΚΩΝ ΔΕΙΓΜΑΤΩΝ – ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ – ΜWST ΤΑΞΙΝΟΜΗΣΗ																			
γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6	γονίδιο2	γονίδιο1	αυδ6
B3GNT3	pp	CLGN	SPINK5	pp	DLK1	SLC26A3	pp	OGN	PRAME	pp	OGN	KCNJ16	pp	S100P	ATP6V0A4	pp	CSTA	SIX1	pp
B3GNT3	pp	CSTA	KLK11	pp	OLFM4	SLC26A3	pp	PIP	PRAME	pp	ASPN	KRT4	pp	PPARGC1A	ATP6V0A4	pp	VGLL1	SIX1	pp
B3GNT3	pp	ASCL2	KLK11	pp	PPARGC1A	SLC26A3	pp	SYT13	PRAME	pp	SLC1A1	KRT4	pp	ERBB2	ATP6V0A4	pp	TTYH1	SLC1A1	pp
B3GNT3	pp	VGLL1	KLK11	pp	FOXJ1	SLC26A3	pp	S100A9	PRAME	pp	BRCA1	KRT4	pp	KRT16	ATP6V0A4	pp	S100P	SLC1A1	pp
B3GNT3	pp	S100A7	KLK11	pp	ASCL2	SLC26A3	pp	CLDN8	PRAME	pp	TFE1	KRT4	pp	ATP6V0A4	ATP6V0A4	pp	CCL19	SLC1A1	pp
B3GNT3	pp	S100A9	KLK11	pp	NTRK3	EDN2	pp	TUBB3	PRAME	pp	DLK1	KRT4	pp	VGLL1	VGLL1	pp	TFE1	SLC1A1	pp
B3GNT3	pp	CCL18	KLK11	pp	TTYH1	EDN2	pp	CSTA	SSBP2	pp	FN1	KRT4	pp	S100P	PENK	pp	EDN3	SLC1A1	pp
B3GNT3	pp	C2orf54	KLK11	pp	S100A9	EDN2	pp	FN1	SSBP2	pp	LAMP3	KRT4	pp	CCL19	PENK	pp	DLK1	SLC1A1	pp
TUBB3	pp	FN1	KLK11	pp	BRCA1	EDN2	pp	GREM1	SSBP2	pp	AMIGO2	KRT4	pp	SLC1A1	PENK	pp	NRG2	SLC1A1	pp
TUBB3	pp	GREM1	CHIT1	pp	CSTA	EDN2	pp	LAMP3	SSBP2	pp	IGHG1	KRT4	pp	C2orf54	PIP	pp	CDKN2A	SLC1A1	pp
TUBB3	pp	MMP12	CHIT1	pp	LAMP3	EDN2	pp	NRG1	SSBP2	pp	ASCL2	KRT4	pp	CILP	PIP	pp	TUBB3	SLC1A1	pp
TUBB3	pp	SYT13	CHIT1	pp	IGHG1	EDN2	pp	ASCL2	SSBP2	pp	OGN	KRT4	pp	CLDN8	PIP	pp	CLGN	SOX11	pp
TUBB3	pp	S100A7	CHIT1	pp	LYZ	EDN2	pp	OGN	SSBP2	pp	S100A7	KRT4	pp	REPS2	PIP	pp	SCGB1D1	SOX11	pp
TUBB3	pp	S100P	CHIT1	pp	MMP12	EDN2	pp	VGLL1	GREM1	pp	IGHG1	KRT16	pp	CDKN2A	PIP	pp	SPINK5	BRCA1	pp
TUBB3	pp	SOX11	CHIT1	pp	CCL18	EDN2	pp	ACTA1	GREM1	pp	LYZ	KRT16	pp	B3GNT3	PIP	pp	COL17A1	BRCA1	pp
TUBB3	pp	RERGL	WIF1	pp	EDN3	EDN3	pp	TFE1	GREM1	pp	MMP12	KRT16	pp	TUBB3	PIP	pp	COMP	BRCA1	pp
CLGN	pp	CDKN2A	WIF1	pp	KCNJ16	EDN3	pp	NRG2	GREM1	pp	ASPN	KRT16	pp	OLFM4	PIP	pp	ERBB2	BRCA1	pp
CLGN	pp	LAMP3	WIF1	pp	LYZ	EGF	pp	PARK7	LAMP3	pp	CDKN2A	KRT16	pp	PPARGC1A	PIP	pp	FN1	BRCA1	pp
HOBX13	pp	CPA3	WIF1	pp	NTRK3	EGF	pp	FGFR3	LAMP3	pp	HS17B2	KRT16	pp	SPINK5	PIP	pp	PRAME	TAT	pp
HOBX13	pp	SLC1A1	WIF1	pp	NRG2	EGF	pp	HS17B2	LAMP3	pp	IGHG1	KRT16	pp	KLK11	PIP	pp	TFCP2L1	TAT	pp
OLFM4	pp	B3GNT3	PARK7	pp	CLGN	EGF	pp	PIP	LAMP3	pp	LYZ	KRT16	pp	COL17A1	PIP	pp	CITED1	TAT	pp
OLFM4	pp	COL17A1	PARK7	pp	TTYH1	EGF	pp	FGF18	LAMP3	pp	CCL19	KRT16	pp	EDN2	PIP	pp	TTYH1	TAT	pp
OLFM4	pp	EDN3	COL11A1	pp	COMP	EGF	pp	CLDN8	LAMP3	pp	SOX11	KRT16	pp	FABP7	PIP	pp	S100P	TAT	pp
OLFM4	pp	FN1	COL11A1	pp	LAMP3	ERBB2	pp	TUBB3	LAMP3	pp	CD19	KRT16	pp	IGHG1	PIP	pp	CITED1	TAT	pp
OLFM4	pp	ASCL2	COL11A1	pp	OGN	ERBB2	pp	TFCP2L1	GSTA1	pp	TFCP2L1	KRT16	pp	MMP12	PIP	pp	TFE1	TAT	pp
OLFM4	pp	CILP	COL11A1	pp	ASPN	ERBB2	pp	ATP6V0A4	GSTA1	pp	MMP12	KRT16	pp	ATP6V0A4	ASPN	pp	LAMP3	TAT	pp
OLFM4	pp	DLK1	COL11A1	pp	TMEM100	ERBB2	pp	ASPN	GSTA1	pp	ZNF334	KRT16	pp	TMEM100	ASPN	pp	ASCL2	TAT	pp
OLFM4	pp	FGF18	COL11A1	pp	SOX11	ERBB2	pp	TMEM100	GSTA1	pp	TNNI2	KRT16	pp	TTYH1	ASPN	pp	OGN	TAT	pp
SLC34A2	pp	CDKN2A	COL11A1	pp	RERGL	ERBB2	pp	SIX1	GSTA3	pp	GSTA1	KRT16	pp	S100A7	ASPN	pp	S100P	TFE1	pp
SLC34A2	pp	B3GNT3	COL11A1	pp	CILP	ERBB2	pp	SOX11	GSTA3	pp	IGHG1	KRT16	pp	FGF18	ASPN	pp	CILP	TNNI2	pp
SLC34A2	pp	OLFM4	COL11A1	pp	FGF18	ERBB2	pp	C2orf54	GSTA3	pp	KRT4	LYZ	pp	MMP12	ASPN	pp	NRG2	TNNI2	pp
SLC34A2	pp	PPARGC1A	COL17A1	pp	CDKN2A	ERBB2	pp	DLK1	GSTA3	pp	ATP6V0A4	LYZ	pp	CCL18	TMEM100	pp	WIF1	TNNI2	pp
SLC34A2	pp	PARK7	COL17A1	pp	WIF1	ERBB2	pp	REPS2	GSTA3	pp	TTYH1	ASCL2	pp	CSTA	TMEM100	pp	PARK7	TNNI2	pp
SLC34A2	pp	CPA3	COL17A1	pp	COL11A1	FABP7	pp	TTYH1	GSTA3	pp	NRG2	ASCL2	pp	LAMP3	TMEM100	pp	ASCL2	TNNI2	pp
SLC34A2	pp	ERBB2	COL17A1	pp	CPA3	FGFR3	pp	SLC34A2	TFCP2L1	pp	CDKN2A	ASCL2	pp	IGHG1	TMEM100	pp	NTRK3	TNNI2	pp
SLC34A2	pp	SSBP2	COL17A1	pp	FN1	FGFR3	pp	COMP	TFCP2L1	pp	B3GNT3	ASCL2	pp	CCL19	TMEM100	pp	PENK	TNNI2	pp

SLC34A2	pp	GSTA1	COL17A1	pp	AMIGO2	FGFR3	pp	CSTA	TFCP2L1	pp	TUBB3	ASCL2	pp	C2orf54	TMEM100	pp	TTYH1	TNNI2	pp
SLC34A2	pp	GSTA3	COL17A1	pp	NTRK3	FGFR3	pp	EDN2	TFCP2L1	pp	HOBX13	MMP12	pp	CCL18	TMEM100	pp	CCL19	TNNI2	pp
SLC34A2	pp	TFCP2L1	COL17A1	pp	VGLL1	FGFR3	pp	LAMP3	TFCP2L1	pp	FOXJ1	CITED1	pp	HOBX13	TMEM100	pp	SOX11	TNNI2	pp
SLC34A2	pp	HS17B2	COL17A1	pp	S100A9	FGFR3	pp	TFCP2L1	TFCP2L1	pp	FN1	CITED1	pp	OLFM4	TMEM100	pp	RERGL	TNNI2	pp
SLC34A2	pp	AMIGO2	COL17A1	pp	SOX11	FGFR3	pp	NRG1	TFCP2L1	pp	KRT16	CITED1	pp	CHIT1	TMEM100	pp	CD19	RERGL	pp
SLC34A2	pp	KCNJ16	COL17A1	pp	RERGL	FGFR3	pp	NR4A1	TFCP2L1	pp	ASCL2	CITED1	pp	COL17A1	ZNF334	pp	COL17A1	RERGL	pp
SLC34A2	pp	KRT16	COL17A1	pp	FGF18	FGFR3	pp	KRT16	TFCP2L1	pp	VGLL1	CITED1	pp	EDN2	ZNF334	pp	LAMP3	TFPI2	pp
SLC34A2	pp	ASCL2	COL17A1	pp	CD19	FGFR3	pp	MSMB	TFCP2L1	pp	S100A7	CITED1	pp	TFCP2L1	ZNF334	pp	IGHG1	TFPI2	pp
SLC34A2	pp	VGLL1	COMP	pp	ATP6V0A4	FGFR3	pp	VGLL1	TFCP2L1	pp	SLC1A1	CITED1	pp	KCNJ16	ZNF334	pp	ASCL2	TFPI2	pp
SLC34A2	pp	SIX1	COMP	pp	ASPN	FGFR3	pp	SYT13	TFCP2L1	pp	TFE1	CITED1	pp	NTRK3	TTYH1	pp	VGLL1	TFPI2	pp
SLC34A2	pp	TFE1	COMP	pp	CILP	FGFR3	pp	S100A9	TFCP2L1	pp	TNNI2	CITED1	pp	CCL18	SYT13	pp	TFE1	TFPI2	pp
SLC34A2	pp	TNNI2	COMP	pp	DLK1	FGFR3	pp	SOX11	NRG1	pp	CDKN2A	CITED1	pp	TFE1	ACTA1	pp	TUBB3	TFPI2	pp
SLC34A2	pp	CLDN8	CPA3	pp	COMP	FGFR3	pp	BRCA1	NRG1	pp	CLGN	CITED1	pp	TFPI2	ACTA1	pp	COL11A1	TFPI2	pp
SLC34A2	pp	REPS2	CPA3	pp	FN1	FGFR3	pp	TFE1	NRG1	pp	WIF1	CITED1	pp	FGF18	ACTA1	pp	COL17A1	TFPI2	pp
SCGB1D1	pp	CDKN2A	CPA3	pp	GREM1	FOXJ1	pp	B3GNT3	NRG1	pp	PARK7	CITED1	pp	REPS2	ACTA1	pp	NRG1	TFPI2	pp
SCGB1D1	pp	OLFM4	CPA3	pp	IGHG1	FOXJ1	pp	S100P	NRG1	pp	COL11A1	MSMB	pp	CLGN	ACTA1	pp	MMP12	C2orf54	pp
SCGB1D1	pp	CPA3	CPA3	pp	LYZ	FOXJ1	pp	C2orf54	NRG1	pp	COL17A1	MSMB	pp	SLC34A2	ACTA1	pp	OGN	CILP	pp
SCGB1D1	pp	FABP7	CPA3	pp	OGN	FN1	pp	WIF1	NRG1	pp	GREM1	MSMB	pp	EDN2	ACTA1	pp	TTYH1	CILP	pp
SCGB1D1	pp	PRAME	CPA3	pp	ASPN	FN1	pp	COL11A1	NRG1	pp	LAMP3	MSMB	pp	S100P	ACTA1	pp	S100P	CILP	pp
SCGB1D1	pp	SSBP2	CPA3	pp	TMEM100	FN1	pp	EDN3	NRG1	pp	AMIGO2	MSMB	pp	SLC1A1	ACTA1	pp	TFE1	CILP	pp
SCGB1D1	pp	CITED1	CPA3	pp	CCL19	FN1	pp	FABP7	NRG1	pp	OGN	MSMB	pp	TFE1	ACTA1	pp	C2orf54	CILP	pp
SCGB1D1	pp	TTYH1	CPA3	pp	SOX11	FN1	pp	GREM1	NR4A1	pp	COL17A1	MSMB	pp	TNNI2	S100A7	pp	IGHG1	CILP	pp
SCGB1D1	pp	TFE1	CPA3	pp	CILP	FN1	pp	AMIGO2	NR4A1	pp	TTYH1	NTRK3	pp	EDN3	S100A7	pp	S100A9	CILP	pp
SCGB1D1	pp	REPS2	CPA3	pp	FGF18	FN1	pp	ASCL2	NR4A1	pp	FGF18	NTRK3	pp	FABP7	S100A7	pp	CCL19	DLK1	pp
PPARGC1A	pp	TUBB3	CSTA	pp	LAMP3	FN1	pp	CCL19	HS17B2	pp	FABP7	NTRK3	pp	TTYH1	S100A7	pp	SOX11	DLK1	pp
PPARGC1A	pp	WIF1	CSTA	pp	SYT13	FN1	pp	C2orf54	HS17B2	pp	TFE1	NTRK3	pp	C2orf54	S100A9	pp	HS17B2	DLK1	pp
PPARGC1A	pp	HS17B2	CSTA	pp	CD19	FN1	pp	CILP	HS17B2	pp	NRG2	OGN	pp	FABP7	S100P	pp	B3GNT3	FGF18	pp
PPARGC1A	pp	OGN	SLC26A3	pp	CLGN	PRAME	pp	HOBX13	AMIGO2	pp	FABP7	OGN	pp	LAMP3	S100P	pp	CHIT1	FGF18	pp
PPARGC1A	pp	TMEM100	SLC26A3	pp	KLK11	PRAME	pp	SLC34A2	AMIGO2	pp	IGHG1	OGN	pp	MMP12	S100P	pp	TFE1	FGF18	pp
PPARGC1A	pp	SOX11	SLC26A3	pp	CPA3	PRAME	pp	PPARGC1A	AMIGO2	pp	MMP12	OGN	pp	VGLL1	CCL18	pp	S100A9	FGF18	pp
PPARGC1A	pp	FGF18	SLC26A3	pp	EGF	PRAME	pp	COL17A1	AMIGO2	pp	NTRK3	OGN	pp	PENK	CCL19	pp	CCL18	FGF18	pp
PPARGC1A	pp	CLDN8	SLC26A3	pp	PRAME	PRAME	pp	EDN3	AMIGO2	pp	ATP6V0A4	OGN	pp	TMEM100	CCL19	pp	CD19	FGF18	pp
SPINK5	pp	B3GNT3	SLC26A3	pp	GREM1	PRAME	pp	SSBP2	AMIGO2	pp	VGLL1	OGN	pp	S100P	SIX1	pp	CLGN	FGF18	pp
SPINK5	pp	COL17A1	SLC26A3	pp	NRG1	PRAME	pp	GSTA1	AMIGO2	pp	C2orf54	OGN	pp	SOX11	SIX1	pp	HOBX13	FGF18	pp
SPINK5	pp	CSTA	SLC26A3	pp	NR4A1	PRAME	pp	GSTA3	IGHG1	pp	LYZ	OGN	pp	RERGL	SIX1	pp	GREM1	CLDN8	pp
SPINK5	pp	FN1	SLC26A3	pp	HS17B2	PRAME	pp	KRT4	IGHG1	pp	CCL18	OGN	pp	DLK1	SIX1	pp	NTRK3	CLDN8	pp
SPINK5	pp	NTRK3	SLC26A3	pp	KRT16	PRAME	pp	ASCL2	KCNJ16	pp	VGLL1	ATP6V0A4	pp	B3GNT3	SIX1	pp	ATP6V0A4	CLDN8	pp
SPINK5	pp	VGLL1	SLC26A3	pp	MSMB	PRAME	pp	CITED1	KCNJ16	pp	PENK	ATP6V0A4	pp	CLGN	SIX1	pp	CCL18	CLDN8	pp

Εικόνα Β.3 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των καρκινικών δειγμάτων, συνεχών μεταβλητών και MWST ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ - ΚΑΡΚΙΝΙΚΩΝ ΔΕΙΓΜΑΤΩΝ – ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ – CUSTOM ΤΑΞΙΝΟΜΗΣΗ																				
γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2	γονίδιο1	αυθ6	γονίδιο2
CDKN2A	pp	SCGB1D1	COL17A1	pp	PIP	ERBB2	pp	OLFM4	NRG1	pp	SLC26A3	ASCL2	pp	C2orf54	TTYH1	pp	FABP7	TAT	pp	SLC26A3
CDKN2A	pp	COL17A1	COL17A1	pp	CCL19	ERBB2	pp	TFPC2L1	NRG1	pp	EDN3	ASCL2	pp	NRG2	TTYH1	pp	SSBP2	TAT	pp	EGF
CDKN2A	pp	ERBB2	COMP	pp	SLC34A2	ERBB2	pp	ASCL2	NRG1	pp	FGFR3	MMP12	pp	CHIT1	TTYH1	pp	LYZ	TAT	pp	PRAME
CDKN2A	pp	PRAME	COMP	pp	PPARGC1A	ERBB2	pp	MMP12	NRG1	pp	LAMP3	MMP12	pp	GREM1	TTYH1	pp	NTRK3	TAT	pp	LAMP3
CDKN2A	pp	LAMP3	COMP	pp	COL11A1	ERBB2	pp	TMEM100	NRG1	pp	AMIGO2	MMP12	pp	LYZ	TTYH1	pp	VGLL1	TAT	pp	IGHG1
CDKN2A	pp	IGHG1	COMP	pp	CPA3	ERBB2	pp	S100P	NRG1	pp	KCNJ16	MMP12	pp	CCL18	TTYH1	pp	PIP	TAT	pp	KCNJ16
CDKN2A	pp	KRT16	COMP	pp	EDN3	ERBB2	pp	SIX1	NRG1	pp	KRT16	MMP12	pp	CCL19	TTYH1	pp	TMEM100	TAT	pp	PENK
CDKN2A	pp	VGLL1	COMP	pp	FABP7	ERBB2	pp	SOX11	NRG1	pp	OGN	CITED1	pp	TFPC2L1	TTYH1	pp	ZNF334	TAT	pp	PIP
CDKN2A	pp	PIP	COMP	pp	FGFR3	ERBB2	pp	C2orf54	NRG1	pp	ACTA1	CITED1	pp	NTRK3	TTYH1	pp	S100A7	TAT	pp	TFF1
CDKN2A	pp	TMEM100	COMP	pp	LAMP3	ERBB2	pp	DLK1	NRG1	pp	S100P	CITED1	pp	CCL18	TTYH1	pp	SLC1A1	TAT	pp	RERGL
CDKN2A	pp	TYTH1	COMP	pp	TFPC2L1	ERBB2	pp	REPS2	NRG1	pp	TFPI2	MSMB	pp	CLGN	TTYH1	pp	TFF1	TAT	pp	REPS2
CDKN2A	pp	SOX11	COMP	pp	KRT16	FABP7	pp	CSTA	NRG1	pp	CLDN8	MSMB	pp	OLFM4	TTYH1	pp	NRG2	TAT	pp	NRG2
CDKN2A	pp	TFF1	COMP	pp	LYZ	FABP7	pp	SLC26A3	HSO17B2	pp	CLDN8	MSMB	pp	SLC1A1	SYT13	pp	TFPC2L1	TFF1	pp	CLGN
B3GNT3	pp	CLGN	COMP	pp	OGN	FABP7	pp	FOXJ1	AMIGO2	pp	B3GNT3	MSMB	pp	TFPI2	ACTA1	pp	CDKN2A	TFF1	pp	COL17A1
B3GNT3	pp	TFPC2L1	COMP	pp	ATP6V0A4	FABP7	pp	HSO17B2	IGHG1	pp	B3GNT3	NTRK3	pp	HSO17B2	ACTA1	pp	TUBB3	TFF1	pp	FABP7
B3GNT3	pp	S100A7	COMP	pp	PIP	FABP7	pp	AMIGO2	IGHG1	pp	CHIT1	NTRK3	pp	CLDN8	ACTA1	pp	CLGN	TFF1	pp	FOXJ1
B3GNT3	pp	S100A9	COMP	pp	ASPN	FABP7	pp	KCNJ16	IGHG1	pp	PARK7	OGN	pp	CHIT1	ACTA1	pp	COL17A1	TFF1	pp	TFPC2L1
B3GNT3	pp	CCL18	COMP	pp	TYTH1	FABP7	pp	NTRK3	IGHG1	pp	CPA3	OGN	pp	CPA3	ACTA1	pp	EDN2	TFF1	pp	HSO17B2
TUBB3	pp	PIP	COMP	pp	CILP	FABP7	pp	OGN	IGHG1	pp	FGFR3	OGN	pp	PRAME	ACTA1	pp	KRT16	TFF1	pp	CITED1
TUBB3	pp	SYT13	COMP	pp	FGF18	FABP7	pp	S100A7	IGHG1	pp	SSBP2	OGN	pp	SSBP2	ACTA1	pp	LYZ	TFF1	pp	MSMB
TUBB3	pp	TFPI2	COMP	pp	HOXB13	FABP7	pp	SLC1A1	IGHG1	pp	LYZ	OGN	pp	LAMP3	ACTA1	pp	ASCL2	TFF1	pp	OGN
CLGN	pp	LAMP3	CPA3	pp	SCGB1D1	FABP7	pp	REPS2	IGHG1	pp	ASCL2	OGN	pp	NR4A1	ACTA1	pp	MMP12	TFF1	pp	SYT13
HOXB13	pp	PRAME	CPA3	pp	COL11A1	FGFR3	pp	PPARGC1A	IGHG1	pp	MMP12	OGN	pp	MMP12	ACTA1	pp	OGN	TFF1	pp	S100A7
HOXB13	pp	CITED1	CPA3	pp	GREM1	FGFR3	pp	LAMP3	IGHG1	pp	TYTH1	OGN	pp	PENK	ACTA1	pp	VGLL1	TFF1	pp	S100P
HOXB13	pp	SLC1A1	CPA3	pp	LAMP3	FGFR3	pp	TFPC2L1	IGHG1	pp	CCL18	OGN	pp	ASPN	ACTA1	pp	PIP	TFF1	pp	SOX11
OLFM4	pp	SCGB1D1	CPA3	pp	LYZ	FGFR3	pp	NR4A1	IGHG1	pp	RERGL	OGN	pp	SLC1A1	ACTA1	pp	TNNI2	TFF1	pp	RERGL
OLFM4	pp	CLDN8	CPA3	pp	ASPN	FGFR3	pp	PIP	IGHG1	pp	CD19	OGN	pp	SOX11	ACTA1	pp	C2orf54	TFF1	pp	TFPI2
SLC34A2	pp	B3GNT3	CPA3	pp	SYT13	FGFR3	pp	SYT13	KCNJ16	pp	OLFM4	OGN	pp	TNNI2	ACTA1	pp	CLDN8	TFF1	pp	C2orf54
SLC34A2	pp	OLFM4	CPA3	pp	REPS2	FGFR3	pp	S100A9	KCNJ16	pp	WIF1	OGN	pp	CILP	S100A9	pp	OLFM4	TFF1	pp	CLDN8
SLC34A2	pp	CPA3	CSTA	pp	HOXB13	FGFR3	pp	SOX11	KCNJ16	pp	COL17A1	OGN	pp	DLK1	S100A9	pp	KLK11	TFF1	pp	REPS2
SLC34A2	pp	EDN2	CSTA	pp	SCGB1D1	FGFR3	pp	BRCA1	KCNJ16	pp	EDN3	ATP6V0A4	pp	B3GNT3	S100A9	pp	HSO17B2	TNNI2	pp	WIF1
SLC34A2	pp	SSBP2	CSTA	pp	SPINK5	FGFR3	pp	TFF1	KCNJ16	pp	CITED1	ATP6V0A4	pp	TUBB3	S100A9	pp	S100A7	TNNI2	pp	GSTA1
SLC34A2	pp	GSTA1	CSTA	pp	NTRK3	FOXJ1	pp	KLK11	KCNJ16	pp	PENK	ATP6V0A4	pp	CSTA	S100P	pp	B3GNT3	TNNI2	pp	KRT4
SLC34A2	pp	HSO17B2	CSTA	pp	SLC1A1	FOXJ1	pp	COL11A1	KCNJ16	pp	CLDN8	ATP6V0A4	pp	GSTA3	S100P	pp	TUBB3	RERGL	pp	COL11A1
SLC34A2	pp	AMIGO2	CSTA	pp	TFPI2	FOXJ1	pp	SLC1A1	KRT4	pp	OLFM4	VGLL1	pp	B3GNT3	S100P	pp	CHIT1	RERGL	pp	COL17A1
SLC34A2	pp	MSMB	SLC26A3	pp	TUBB3	FN1	pp	TUBB3	KRT4	pp	CILP	VGLL1	pp	CLGN	S100P	pp	FOXJ1	RERGL	pp	SLC26A3

SLC34A2	pp	PIP	SLC26A3	pp	CLGN	FN1	pp	PPARGC1A	KRT4	pp	CLDN8	VGLL1	pp	SLC34A2	S100P	pp	KRT4	RERGL	pp	EDN3
SLC34A2	pp	C2orf54	SLC26A3	pp	COL17A1	FN1	pp	SPINK5	KRT16	pp	TUBB3	VGLL1	pp	SPINK5	S100P	pp	MSMB	RERGL	pp	LAMP3
PPARGC1A	pp	TUBB3	SLC26A3	pp	PRAME	FN1	pp	COL11A1	KRT16	pp	SLC34A2	VGLL1	pp	FABP7	S100P	pp	OGN	RERGL	pp	KCNJ16
PPARGC1A	pp	CLGN	SLC26A3	pp	HSO17B2	FN1	pp	COMP	KRT16	pp	PPARGC1A	VGLL1	pp	PRAME	S100P	pp	ATP6V0A4	RERGL	pp	OGN
PPARGC1A	pp	SLC34A2	SLC26A3	pp	OGN	FN1	pp	EDN2	KRT16	pp	SPINK5	VGLL1	pp	LAMP3	S100P	pp	PIP	RERGL	pp	ZNF334
PPARGC1A	pp	KLK11	SLC26A3	pp	PIP	FN1	pp	FABP7	KRT16	pp	KLK11	VGLL1	pp	GSTA3	S100P	pp	C2orf54	RERGL	pp	CCL19
PPARGC1A	pp	WIF1	SLC26A3	pp	SYT13	FN1	pp	SSBP2	KRT16	pp	CHIT1	VGLL1	pp	TFCP2L1	CCL18	pp	OLFM4	RERGL	pp	CILP
PPARGC1A	pp	CPA3	SLC26A3	pp	CLDN8	FN1	pp	GREM1	KRT16	pp	COL17A1	VGLL1	pp	HSO17B2	CCL18	pp	S100A7	C2orf54	pp	SPINK5
PPARGC1A	pp	EDN3	SLC26A3	pp	NRG2	FN1	pp	LYZ	KRT16	pp	CSTA	VGLL1	pp	KCNJ16	CCL18	pp	S100A9	C2orf54	pp	EDN3
PPARGC1A	pp	HSO17B2	EDN2	pp	CSTA	FN1	pp	ASPN	KRT16	pp	SLC26A3	VGLL1	pp	KRT4	CCL18	pp	CCL19	C2orf54	pp	TFCP2L1
PPARGC1A	pp	KRT4	EDN2	pp	FOXJ1	FN1	pp	TYTH1	KRT16	pp	EDN2	VGLL1	pp	ATP6V0A4	CCL19	pp	OLFM4	C2orf54	pp	KRT4
PPARGC1A	pp	OGN	EDN2	pp	GREM1	FN1	pp	ACTA1	KRT16	pp	FABP7	VGLL1	pp	S100P	CCL19	pp	LAMP3	C2orf54	pp	NTRK3
PPARGC1A	pp	TMEM100	EDN2	pp	CITED1	FN1	pp	S100P	KRT16	pp	FGFR3	VGLL1	pp	TFF1	CCL19	pp	CILP	C2orf54	pp	S100A9
PPARGC1A	pp	TYTH1	EDN2	pp	MSMB	FN1	pp	SOX11	KRT16	pp	GSTA1	VGLL1	pp	TNNI2	SIX1	pp	CLGN	C2orf54	pp	SOX11
PPARGC1A	pp	SOX11	EDN2	pp	SLC1A1	FN1	pp	TAT	KRT16	pp	TFCP2L1	VGLL1	pp	NRG2	SIX1	pp	HOXB13	C2orf54	pp	CLDN8
PPARGC1A	pp	REPS2	EDN2	pp	NRG2	FN1	pp	TNNI2	KRT16	pp	AMIGO2	PENK	pp	DLK1	SIX1	pp	SLC34A2	C2orf54	pp	CD19
PPARGC1A	pp	NRG2	EDN3	pp	OLFM4	FN1	pp	RERGL	KRT16	pp	VGLL1	PENK	pp	NRG2	SIX1	pp	WIF1	DLK1	pp	CLGN
SPINK5	pp	B3GNT3	EDN3	pp	WIF1	FN1	pp	C2orf54	KRT16	pp	TYTH1	PIP	pp	SCGB1D1	SIX1	pp	GREM1	DLK1	pp	OLFM4
SPINK5	pp	S100A7	EDN3	pp	GSTA1	FN1	pp	CILP	KRT16	pp	SIX1	PIP	pp	SLC1A1	SIX1	pp	LAMP3	DLK1	pp	PRAME
SPINK5	pp	DLK1	EDN3	pp	LYZ	FN1	pp	CD19	KRT16	pp	BRCA1	ASPN	pp	WIF1	SIX1	pp	MSMB	DLK1	pp	HSO17B2
KLK11	pp	S100A7	EDN3	pp	NTRK3	PRAME	pp	SLC1A1	KRT16	pp	TNNI2	ASPN	pp	COL11A1	SIX1	pp	ATP6V0A4	FGF18	pp	OLFM4
CHIT1	pp	LYZ	EDN3	pp	CLDN8	PRAME	pp	TFPI2	KRT16	pp	CILP	ASPN	pp	GREM1	SIX1	pp	CCL18	FGF18	pp	COL17A1
CHIT1	pp	CCL18	EDN3	pp	NRG2	SSBP2	pp	COL11A1	KRT16	pp	REPS2	ASPN	pp	CILP	SIX1	pp	SLC1A1	FGF18	pp	SLC26A3
PARK7	pp	CLGN	EGF	pp	TUBB3	SSBP2	pp	CPA3	LYZ	pp	COL11A1	ASPN	pp	CLDN8	SIX1	pp	C2orf54	FGF18	pp	EDN3
PARK7	pp	ASCL2	EGF	pp	SLC34A2	SSBP2	pp	EDN3	LYZ	pp	LAMP3	TMEM100	pp	OLFM4	SOX11	pp	TUBB3	FGF18	pp	NRG1
PARK7	pp	TYTH1	EGF	pp	PARK7	SSBP2	pp	TFCP2L1	LYZ	pp	CITED1	TMEM100	pp	NTRK3	SOX11	pp	HOXB13	FGF18	pp	KRT16
PARK7	pp	BRCA1	EGF	pp	COL17A1	SSBP2	pp	S100A7	LYZ	pp	CCL18	TMEM100	pp	OGN	SOX11	pp	COL11A1	FGF18	pp	OGN
PARK7	pp	TFF1	EGF	pp	CSTA	SSBP2	pp	REPS2	LYZ	pp	CCL19	TMEM100	pp	PENK	SOX11	pp	TFCP2L1	FGF18	pp	VGLL1
COL11A1	pp	DLK1	EGF	pp	FABP7	GREM1	pp	CLDN8	ASCL2	pp	B3GNT3	TMEM100	pp	CCL19	SOX11	pp	MMP12	FGF18	pp	TMEM100
COL11A1	pp	NRG2	EGF	pp	FGFR3	GSTA3	pp	GSTA1	ASCL2	pp	OLFM4	TMEM100	pp	RERGL	SOX11	pp	S100A9	FGF18	pp	BRCA1
COL17A1	pp	OLFM4	EGF	pp	HSO17B2	TFCP2L1	pp	CLGN	ASCL2	pp	KLK11	ZNF334	pp	COL17A1	SOX11	pp	DLK1	FGF18	pp	TAT
COL17A1	pp	SPINK5	EGF	pp	ASCL2	TFCP2L1	pp	GSTA1	ASCL2	pp	CSTA	ZNF334	pp	LAMP3	BRCA1	pp	KLK11	FGF18	pp	RERGL
COL17A1	pp	WIF1	EGF	pp	MSMB	TFCP2L1	pp	HSO17B2	ASCL2	pp	EDN2	ZNF334	pp	GSTA1	BRCA1	pp	WIF1	REPS2	pp	B3GNT3
COL17A1	pp	CPA3	EGF	pp	PIP	TFCP2L1	pp	PIP	ASCL2	pp	FABP7	ZNF334	pp	TFCP2L1	BRCA1	pp	FOXJ1	REPS2	pp	TUBB3
COL17A1	pp	HSO17B2	EGF	pp	RERGL	NRG1	pp	WIF1	ASCL2	pp	SSBP2	ZNF334	pp	S100A9	BRCA1	pp	OGN	REPS2	pp	OLFM4
COL17A1	pp	AMIGO2	EGF	pp	C2orf54	NRG1	pp	PARK7	ASCL2	pp	ZNF334	TYTH1	pp	KLK11	BRCA1	pp	TMEM100	REPS2	pp	KLK11
COL17A1	pp	CITED1	ERBB2	pp	B3GNT3	NRG1	pp	COL11A1	ASCL2	pp	CCL19	TYTH1	pp	CPA3	BRCA1	pp	TYTH1	REPS2	pp	COL11A1
COL17A1	pp	NTRK3		pp	TUBB3	NRG1	pp	COL17A1	ASCL2	pp	SOX11	TYTH1	pp	EDN3	BRCA1	pp	S100P	REPS2	pp	GREM1

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ – ΥΓΕΙΩΝ ΔΕΙΓΜΑΤΩΝ - ΔΙΑΚΡΙΤΩΝ ΜΕΤΑΒΛΗΤΩΝ – MWST ΤΑΞΙΝΟΜΗΣΗ											
γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2
B3GNT3	pp	SLC34A2	EDN2	pp	DLK1	KRT16	pp	TFCP2L1	TFF1	pp	MSMB
B3GNT3	pp	COL11A1	EDN3	pp	COL17A1	ASCL2	pp	CSTA	RERGL	pp	CLDN8
B3GNT3	pp	TTYH1	EDN3	pp	FABP7	ASCL2	pp	AMIGO2	TFPI2	pp	CITED1
TUBB3	pp	S100P	EGF	pp	CDKN2A	MMP12	pp	LAMP3	C2orf54	pp	COMP
CLGN	pp	ERBB2	ERBB2	pp	COMP	MMP12	pp	KCNJ16	C2orf54	pp	GREM1
HOXB13	pp	NRG2	FGFR3	pp	HOXB13	MSMB	pp	CLGN	CILP	pp	FN1
OLFM4	pp	TNNI2	FOXJ1	pp	GREM1	OGN	pp	RERGL	CILP	pp	NRG1
SLC34A2	pp	ERBB2	FOXJ1	pp	ASCL2	ATP6V0A4	pp	B3GNT3	CILP	pp	OGN
SLC34A2	pp	HSD17B2	PRAME	pp	GREM1	ATP6V0A4	pp	SIX1	FGF18	pp	ACTA1
SLC34A2	pp	KRT16	SSBP2	pp	CHIT1	VGLL1	pp	OLFM4	CD19	pp	SSBP2
SCGB1D1	pp	KCNJ16	SSBP2	pp	ASCL2	PENK	pp	TAT	NRG2	pp	EDN3
PPARGC1A	pp	TUBB3	SSBP2	pp	FGF18	PIP	pp	WIF1			
WIF1	pp	NRG2	SSBP2	pp	REPS2	TMEM100	pp	NTRK3			
COL11A1	pp	C2orf54	GSTA1	pp	GSTA3	TTYH1	pp	EGF			
COL11A1	pp	CLDN8	GSTA3	pp	VGLL1	S100A7	pp	ASPN			
COL17A1	pp	ZNF334	TFCP2L1	pp	SCGB1D1	S100A7	pp	SYT13			
COL17A1	pp	TTYH1	TFCP2L1	pp	C2orf54	S100A9	pp	CCL18			
COMP	pp	EDN2	NRG1	pp	RERGL	S100P	pp	PRAME			
CPA3	pp	CILP	IGHG1	pp	ATP6V0A4	SOX11	pp	MMP12			
CSTA	pp	KRT4	KRT16	pp	EGF	BRCA1	pp	LAMP3			

Εικόνα Β.5 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των υγιών δειγμάτων, διακριτών μεταβλητών και MWST ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ – ΥΓΕΙΩΝ ΔΕΙΓΜΑΤΩΝ - ΔΙΑΚΡΙΤΩΝ ΜΕΤΑΒΛΗΤΩΝ – CUSTOM ΤΑΞΙΝΟΜΗΣΗ											
γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2	γονίδιο1	σύνδεση	γονίδιο2
CDKN2A	pp	HOXB13	ERBB2	pp	MSMB	LYZ	pp	CCL18	TTYH1	pp	SIX1
CDKN2A	pp	NTRK3	ERBB2	pp	TNNI2	ASCL2	pp	CSTA	TTYH1	pp	RERGL
PPARGC1A	pp	TUBB3	FABP7	pp	RERGL	ASCL2	pp	FOXJ1	ACTA1	pp	SYT13
WIF1	pp	PIP	FGFR3	pp	HOXB13	ASCL2	pp	SSBP2	ACTA1	pp	FGF18
COL11A1	pp	CLDN8	FN1	pp	KRT16	ASCL2	pp	AMIGO2	S100P	pp	TUBB3
COL17A1	pp	EDN3	FN1	pp	S100P	MMP12	pp	LAMP3	S100P	pp	PRAME
COMP	pp	EDN2	FN1	pp	C2orf54	CITED1	pp	TFPI2	CCL18	pp	S100A9
COMP	pp	NRG1	FN1	pp	CILP	MSMB	pp	CLGN	BRCA1	pp	LAMP3
COMP	pp	IGHG1	SSBP2	pp	CHIT1	OGN	pp	CILP	TAT	pp	PENK
COMP	pp	KCNJ16	SSBP2	pp	WIF1	VGLL1	pp	GSTA3	RERGL	pp	OGN
COMP	pp	ASPN	SSBP2	pp	MMP12	ASPN	pp	GREM1	RERGL	pp	CLDN8
COMP	pp	C2orf54	SSBP2	pp	REPS2	ASPN	pp	CITED1	C2orf54	pp	SPINK5
CSTA	pp	KRT4	GREM1	pp	PRAME	ASPN	pp	S100A7	C2orf54	pp	COL11A1
EDN2	pp	DLK1	GSTA3	pp	GSTA1	TTYH1	pp	B3GNT3	C2orf54	pp	GREM1
EDN3	pp	NRG2	TFCP2L1	pp	SCGB1D1	TTYH1	pp	OLFM4	NRG2	pp	WIF1
EGF	pp	CDKN2A	NRG1	pp	OGN	TTYH1	pp	COL17A1			
EGF	pp	COMP	KRT16	pp	SLC34A2	TTYH1	pp	FABP7			
EGF	pp	TTYH1	KRT16	pp	EGF	TTYH1	pp	HSD17B2			
ERBB2	pp	PARK7	KRT16	pp	ERBB2	TTYH1	pp	ATP6V0A4			
ERBB2	pp	ASCL2	KRT16	pp	TFCP2L1	TTYH1	pp	ZNF334			

Εικόνα Β.6 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των υγιών δειγμάτων, διακριτών μεταβλητών και custom ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ – ΥΓΕΙΩΝ ΔΕΙΓΜΑΤΩΝ – ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ – ΜWST ΤΑΞΙΝΟΜΗΣΗ														
γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2
B3GNT3	pp	OLFM4	PARK7	pp	CITED1	GSTA1	pp	AMIGO2	OGN	pp	DLK1	SIX1	pp	MSMB
B3GNT3	pp	SCGB1D1	PARK7	pp	DLK1	GSTA1	pp	KCNJ16	ATP6V0A4	pp	B3GNT3	SIX1	pp	TMEM100
B3GNT3	pp	EGF	COL11A1	pp	TUBB3	GSTA1	pp	MMP12	ATP6V0A4	pp	SLC34A2	SLC1A1	pp	COMP
B3GNT3	pp	FABP7	COL11A1	pp	COMP	GSTA1	pp	VGLL1	ATP6V0A4	pp	PPARGC1A	SLC1A1	pp	TTYH1
B3GNT3	pp	HSD17B2	COL11A1	pp	OGN	GSTA1	pp	NRG2	ATP6V0A4	pp	CHIT1	SLC1A1	pp	RERGL
B3GNT3	pp	VGLL1	COL11A1	pp	C2orf54	GSTA3	pp	TUBB3	ATP6V0A4	pp	GSTA1	SOX11	pp	EDN2
B3GNT3	pp	TTYH1	COL11A1	pp	CLDN8	GSTA3	pp	HSD17B2	ATP6V0A4	pp	TFCP2L1	SOX11	pp	SSBP2
B3GNT3	pp	SLC1A1	COL17A1	pp	AMIGO2	GSTA3	pp	KRT16	ATP6V0A4	pp	KRT4	SOX11	pp	MMP12
B3GNT3	pp	TNNI2	COL17A1	pp	KCNJ16	GSTA3	pp	ASCL2	ATP6V0A4	pp	VGLL1	BRCA1	pp	EDN2
B3GNT3	pp	TFPI2	COL17A1	pp	TMEM100	GSTA3	pp	TTYH1	ATP6V0A4	pp	S100A7	BRCA1	pp	ERBB2
TUBB3	pp	PRAME	COL17A1	pp	TTYH1	GSTA3	pp	CCL19	ATP6V0A4	pp	SIX1	BRCA1	pp	GREM1
TUBB3	pp	OGN	COL17A1	pp	SLC1A1	GSTA3	pp	CILP	ATP6V0A4	pp	BRCA1	BRCA1	pp	HSD17B2
TUBB3	pp	S100P	COL17A1	pp	CLDN8	TFCP2L1	pp	SCGB1D1	VGLL1	pp	OLFM4	BRCA1	pp	OGN
TUBB3	pp	TFPI2	COMP	pp	EGF	TFCP2L1	pp	OGN	VGLL1	pp	CHIT1	BRCA1	pp	ASPN
TUBB3	pp	DLK1	CPA3	pp	SLC34A2	NRG1	pp	CHIT1	VGLL1	pp	CSTA	BRCA1	pp	SLC1A1
CLGN	pp	SSBP2	CPA3	pp	FN1	NRG1	pp	CSTA	VGLL1	pp	SLC26A3	TAT	pp	CLGN
CLGN	pp	KRT16	CPA3	pp	SSBP2	NRG1	pp	FN1	VGLL1	pp	ERBB2	TAT	pp	SCGB1D1
CLGN	pp	ASCL2	CPA3	pp	OGN	NRG1	pp	KRT4	VGLL1	pp	MMP12	TAT	pp	ASCL2
HOXB13	pp	CLGN	CPA3	pp	CCL18	NRG1	pp	NRG2	VGLL1	pp	MSMB	TAT	pp	MSMB
HOXB13	pp	WIF1	CPA3	pp	CILP	NR4A1	pp	COL11A1	VGLL1	pp	FGF18	TAT	pp	TFF1
HOXB13	pp	EDN2	CSTA	pp	DLK1	NR4A1	pp	FOXJ1	PENK	pp	EDN2	TAT	pp	TFPI2
HOXB13	pp	LAMP3	SLC26A3	pp	FN1	NR4A1	pp	PRAME	PENK	pp	EDN3	TAT	pp	REPS2
HOXB13	pp	HSD17B2	SLC26A3	pp	HSD17B2	NR4A1	pp	VGLL1	PENK	pp	FN1	TAT	pp	NRG2
HOXB13	pp	MMP12	SLC26A3	pp	ZNF334	NR4A1	pp	PENK	PENK	pp	LYZ	TFF1	pp	ERBB2
HOXB13	pp	ATP6V0A4	SLC26A3	pp	S100A7	NR4A1	pp	FGF18	PENK	pp	ASCL2	TFF1	pp	KRT16
HOXB13	pp	PENK	SLC26A3	pp	BRCA1	NR4A1	pp	NRG2	PENK	pp	MSMB	TFF1	pp	CITED1
HOXB13	pp	CD19	SLC26A3	pp	REPS2	AMIGO2	pp	ACTA1	PENK	pp	OGN	TFF1	pp	MSMB
OLFM4	pp	CLGN	EDN2	pp	DLK1	IGHG1	pp	B3GNT3	PENK	pp	S100P	TFF1	pp	SYT13
OLFM4	pp	SCGB1D1	EDN3	pp	COL17A1	IGHG1	pp	TUBB3	PENK	pp	CCL18	TFF1	pp	CCL19
OLFM4	pp	EDN3	EGF	pp	CDKN2A	IGHG1	pp	SLC34A2	PENK	pp	TAT	TFF1	pp	TFPI2
OLFM4	pp	FN1	ERBB2	pp	PPARGC1A	IGHG1	pp	KLK11	PENK	pp	TFF1	TNNI2	pp	ERBB2
OLFM4	pp	CITED1	ERBB2	pp	COMP	IGHG1	pp	PARK7	PIP	pp	SCGB1D1	TNNI2	pp	OGN
OLFM4	pp	OGN	ERBB2	pp	FN1	IGHG1	pp	NRG1	PIP	pp	WIF1	TNNI2	pp	TMEM100
OLFM4	pp	PIP	ERBB2	pp	HSD17B2	IGHG1	pp	KCNJ16	PIP	pp	LYZ	TNNI2	pp	ZNF334
OLFM4	pp	TFPI2	ERBB2	pp	KRT16	IGHG1	pp	ATP6V0A4	PIP	pp	TMEM100	RERGL	pp	TMEM100
SLC34A2	pp	KRT16	ERBB2	pp	ASPN	IGHG1	pp	VGLL1	PIP	pp	TAT	RERGL	pp	FGF18

SCGB1D1	pp	PRAME	ERBB2	pp	TMEM100	IGHG1	pp	CD19	ASPN	pp	ZNF334	TFPI2	pp	CITED1
PPARGC1A	pp	TUBB3	FABP7	pp	TTYH1	KCNJ16	pp	PRAME	ASPN	pp	TTYH1	CILP	pp	CSTA
PPARGC1A	pp	EDN2	FGFR3	pp	HOXB13	KCNJ16	pp	OGN	TMEM100	pp	NTRK3	CILP	pp	FN1
PPARGC1A	pp	CITED1	FGFR3	pp	FOXJ1	KRT16	pp	COL17A1	SYT13	pp	CSTA	CILP	pp	NRG1
PPARGC1A	pp	TMEM100	FGFR3	pp	PRAME	KRT16	pp	TFCP2L1	SYT13	pp	NTRK3	CILP	pp	MMP12
PPARGC1A	pp	RERGL	FGFR3	pp	KRT16	KRT16	pp	OGN	ACTA1	pp	TMEM100	CILP	pp	OGN
PPARGC1A	pp	TFPI2	FGFR3	pp	ASCL2	KRT16	pp	TTYH1	S100A7	pp	CDKN2A	CILP	pp	TNNI2
SPINK5	pp	CSTA	FGFR3	pp	BRCA1	KRT16	pp	FGF18	S100A7	pp	EDN3	CILP	pp	NRG2
SPINK5	pp	SLC26A3	FGFR3	pp	C2orf54	LYZ	pp	SCGB1D1	S100A7	pp	FN1	DLK1	pp	TTYH1
SPINK5	pp	FOXJ1	FGFR3	pp	FGF18	LYZ	pp	PRAME	S100A7	pp	GREM1	REPS2	pp	TUBB3
SPINK5	pp	AMIGO2	FOXJ1	pp	SLC34A2	LYZ	pp	ACTA1	S100A7	pp	MMP12	REPS2	pp	SCGB1D1
SPINK5	pp	IGHG1	FOXJ1	pp	GREM1	LYZ	pp	S100A9	S100A7	pp	MSMB	REPS2	pp	ZNF334
KLK11	pp	CPA3	FOXJ1	pp	LAMP3	LYZ	pp	CCL18	S100A7	pp	ASPN	CD19	pp	CLGN
KLK11	pp	LAMP3	FOXJ1	pp	ASCL2	LYZ	pp	NRG2	S100A7	pp	S100P	CD19	pp	CPA3
KLK11	pp	TFCP2L1	FOXJ1	pp	SYT13	ASCL2	pp	COMP	S100A7	pp	REPS2	CD19	pp	SSBP2
KLK11	pp	NR4A1	FOXJ1	pp	C2orf54	ASCL2	pp	ERBB2	S100A9	pp	PPARGC1A	CD19	pp	LAMP3
KLK11	pp	MSMB	FN1	pp	TUBB3	ASCL2	pp	SYT13	S100A9	pp	EDN2	CD19	pp	KRT16
KLK11	pp	ATP6V0A4	FN1	pp	COL11A1	ASCL2	pp	DLK1	S100A9	pp	FN1	CD19	pp	ASCL2
KLK11	pp	PIP	FN1	pp	COL17A1	ASCL2	pp	REPS2	S100A9	pp	OGN	CD19	pp	MMP12
KLK11	pp	TMEM100	FN1	pp	COMP	MMP12	pp	CHIT1	S100A9	pp	TTYH1	CD19	pp	S100P
KLK11	pp	TAT	FN1	pp	GREM1	MMP12	pp	FN1	S100A9	pp	CCL18	CD19	pp	NRG2
CHIT1	pp	EGF	FN1	pp	CITED1	MMP12	pp	LAMP3	S100A9	pp	REPS2	NRG2	pp	COMP
CHIT1	pp	TMEM100	FN1	pp	ASPN	MMP12	pp	TFCP2L1	S100P	pp	EDN2	NRG2	pp	CSTA
WIF1	pp	COL11A1	FN1	pp	SYT13	MMP12	pp	OGN	S100P	pp	PRAME	NRG2	pp	FABP7
WIF1	pp	SLC26A3	FN1	pp	S100P	MMP12	pp	CCL19	S100P	pp	KRT4	NRG2	pp	ACTA1
WIF1	pp	FOXJ1	FN1	pp	FGF18	MMP12	pp	BRCA1	S100P	pp	RERGL	NRG2	pp	BRCA1
WIF1	pp	FN1	PRAME	pp	CHIT1	MMP12	pp	NRG2	CCL18	pp	TMEM100	NRG2	pp	TNNI2
WIF1	pp	TFCP2L1	PRAME	pp	SLC1A1	CITED1	pp	TTYH1	CCL18	pp	CCL19			
WIF1	pp	NRG1	SSBP2	pp	EGF	MSMB	pp	CLGN	CCL18	pp	CILP			
WIF1	pp	AMIGO2	SSBP2	pp	HSD17B2	MSMB	pp	ERBB2	CCL19	pp	KRT16			
WIF1	pp	KCNJ16	SSBP2	pp	AMIGO2	MSMB	pp	TFCP2L1	CCL19	pp	TTYH1			
WIF1	pp	PENK	SSBP2	pp	ACTA1	MSMB	pp	AMIGO2	SIX1	pp	CDKN2A			
WIF1	pp	SOX11	SSBP2	pp	FGF18	MSMB	pp	NTRK3	SIX1	pp	B3GNT3			
WIF1	pp	TFPI2	GSTA1	pp	PPARGC1A	MSMB	pp	SLC1A1	SIX1	pp	SLC34A2			
WIF1	pp	NRG2	GSTA1	pp	WIF1	MSMB	pp	TFPI2	SIX1	pp	PPARGC1A			
PARK7	pp	CHIT1	GSTA1	pp	COMP	MSMB	pp	REPS2	SIX1	pp	PARK7			
PARK7	pp	COMP	GSTA1	pp	FABP7	NTRK3	pp	CDKN2A	SIX1	pp	FN1			
PARK7	pp	SSBP2	GSTA1	pp	GSTA3	OGN	pp	RERGL	SIX1	pp	KRT16			

Εικόνα Β.7 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των υγιών δειγμάτων, συνεχών μεταβλητών και MWST ταξινόμηση κόμβων.

ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΟΥ – ΥΓΕΙΩΝ ΔΕΙΓΜΑΤΩΝ – ΣΥΝΕΧΩΝ ΜΕΤΑΒΛΗΤΩΝ – CUSTOM ΤΑΞΙΝΟΜΗΣΗ																	
γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2
CDKN2A	pp	MSMB	SLC26A3	pp	ZNF334	FN1	pp	CILP	MMP12	pp	LYZ	ACTA1	pp	KLK11	C2orf54	pp	CSTA
CDKN2A	pp	SYT13	EDN2	pp	HOXB13	FN1	pp	FGF18	MMP12	pp	MSMB	ACTA1	pp	TMEM100	C2orf54	pp	FOXJ1
CDKN2A	pp	S100A7	EDN2	pp	EDN3	PRAME	pp	SCGB1D1	MMP12	pp	SYT13	ACTA1	pp	RERGL	C2orf54	pp	GSTA1
CDKN2A	pp	S100P	EDN2	pp	GSTA1	PRAME	pp	SLC1A1	MMP12	pp	S100A7	ACTA1	pp	FGF18	FGF18	pp	CLGN
CDKN2A	pp	CCL18	EDN2	pp	HSD17B2	SSBP2	pp	CPA3	MMP12	pp	CCL19	S100P	pp	TUBB3	FGF18	pp	SSBP2
CDKN2A	pp	SIX1	EDN2	pp	KRT4	SSBP2	pp	GSTA3	MMP12	pp	CILP	S100P	pp	SPINK5	FGF18	pp	ASPN
B3GNT3	pp	HSD17B2	EDN2	pp	PENK	SSBP2	pp	HSD17B2	MMP12	pp	NRG2	S100P	pp	COL17A1	FGF18	pp	TMEM100
B3GNT3	pp	CILP	EDN2	pp	S100A9	SSBP2	pp	AMIGO2	CITED1	pp	OLFM4	S100P	pp	EDN2	FGF18	pp	TTYH1
TUBB3	pp	WIF1	EDN2	pp	DLK1	GREM1	pp	WIF1	CITED1	pp	PIP	S100P	pp	PRAME	FGF18	pp	CCL18
TUBB3	pp	PRAME	EDN3	pp	WIF1	GSTA1	pp	WIF1	CITED1	pp	TFPI2	S100P	pp	GSTA1	REPS2	pp	TUBB3
TUBB3	pp	GSTA3	EGF	pp	CDKN2A	GSTA1	pp	CILP	MSMB	pp	CLGN	S100P	pp	KCNJ16	REPS2	pp	LAMP3
TUBB3	pp	LYZ	EGF	pp	B3GNT3	GSTA1	pp	NRG2	MSMB	pp	NTRK3	S100P	pp	KRT4	REPS2	pp	CCL18
TUBB3	pp	MSMB	EGF	pp	SPINK5	GSTA3	pp	GSTA1	MSMB	pp	SLC1A1	S100P	pp	MMP12	CD19	pp	TUBB3
TUBB3	pp	TFPI2	EGF	pp	PARK7	GSTA3	pp	HSD17B2	MSMB	pp	TFPI2	S100P	pp	ZNF334	CD19	pp	CLGN
TUBB3	pp	DLK1	EGF	pp	COMP	TFCP2L1	pp	SCGB1D1	OGN	pp	TUBB3	S100P	pp	S100A7	CD19	pp	LAMP3
CLGN	pp	CILP	EGF	pp	FGFR3	TFCP2L1	pp	KLK11	OGN	pp	WIF1	S100P	pp	RERGL	CD19	pp	MMP12
HOXB13	pp	LAMP3	EGF	pp	HSD17B2	TFCP2L1	pp	WIF1	OGN	pp	CPA3	S100P	pp	TFPI2	CD19	pp	CILP
HOXB13	pp	PENK	EGF	pp	IGHG1	TFCP2L1	pp	GSTA1	OGN	pp	CSTA	S100P	pp	CD19	CD19	pp	NRG2
HOXB13	pp	ASPN	EGF	pp	ATP6V0A4	TFCP2L1	pp	MSMB	OGN	pp	EDN3	CCL18	pp	S100A7	NRG2	pp	WIF1
HOXB13	pp	CCL18	EGF	pp	TTYH1	TFCP2L1	pp	NRG2	OGN	pp	LYZ	CCL18	pp	S100A9	NRG2	pp	TFPI2
OLFM4	pp	SCGB1D1	EGF	pp	TNNI2	NRG1	pp	SPINK5	OGN	pp	PENK	CCL18	pp	CCL19			
SLC34A2	pp	CPA3	EGF	pp	CD19	NRG1	pp	PARK7	OGN	pp	TFPI2	CCL18	pp	CILP			
SLC34A2	pp	EDN2	ERBB2	pp	PPARGC1A	NRG1	pp	FABP7	OGN	pp	CILP	CCL19	pp	GSTA3			
SLC34A2	pp	KRT4	ERBB2	pp	COL11A1	NRG1	pp	AMIGO2	OGN	pp	DLK1	SIX1	pp	ATP6V0A4			
SLC34A2	pp	MSMB	ERBB2	pp	COMP	NRG1	pp	KCNJ16	ATP6V0A4	pp	B3GNT3	SIX1	pp	TMEM100			
SLC34A2	pp	ATP6V0A4	ERBB2	pp	SLC26A3	NRG1	pp	KRT4	ATP6V0A4	pp	CHIT1	SIX1	pp	SOX11			
PPARGC1A	pp	TUBB3	ERBB2	pp	GSTA1	NRG1	pp	TNNI2	ATP6V0A4	pp	EDN2	SIX1	pp	NRG2			
PPARGC1A	pp	HOXB13	ERBB2	pp	NRG1	NRG1	pp	NRG2	ATP6V0A4	pp	TFCP2L1	SOX11	pp	EDN2			
PPARGC1A	pp	CSTA	ERBB2	pp	HSD17B2	NR4A1	pp	COL17A1	ATP6V0A4	pp	HSD17B2	SOX11	pp	KRT4			
PPARGC1A	pp	EDN2	ERBB2	pp	ASCL2	NR4A1	pp	KCNJ16	VGLL1	pp	B3GNT3	SOX11	pp	MMP12			
PPARGC1A	pp	FOXJ1	ERBB2	pp	OGN	NR4A1	pp	LYZ	VGLL1	pp	OLFM4	BRCA1	pp	B3GNT3			
PPARGC1A	pp	GSTA1	ERBB2	pp	TMEM100	NR4A1	pp	CCL18	VGLL1	pp	CHIT1	BRCA1	pp	WIF1			
PPARGC1A	pp	HSD17B2	ERBB2	pp	ZNF334	NR4A1	pp	REPS2	VGLL1	pp	CSTA	BRCA1	pp	CPA3			
PPARGC1A	pp	ATP6V0A4	ERBB2	pp	CCL18	NR4A1	pp	NRG2	VGLL1	pp	GSTA3	BRCA1	pp	SLC26A3			
PPARGC1A	pp	TMEM100	ERBB2	pp	TAT	HSD17B2	pp	S100A7	VGLL1	pp	MSMB	BRCA1	pp	FGFR3			
PPARGC1A	pp	SIX1	ERBB2	pp	TNNI2	AMIGO2	pp	SPINK5	VGLL1	pp	CCL18	BRCA1	pp	GREM1			

PPARGC1A	pp	TFPI2	ERBB2	pp	CILP	AMIGO2	pp	WIF1	VGLL1	pp	CCL19	BRCA1	pp	ATP6VOA4	
PPARGC1A	pp	REPS2	FABP7	pp	SPINK5	AMIGO2	pp	MSMB	PENK	pp	WIF1	BRCA1	pp	ASPN	
SPINK5	pp	WIF1	FABP7	pp	CITED1	IGHG1	pp	PARK7	PENK	pp	NRG2	BRCA1	pp	TMEM100	
KLK11	pp	WIF1	FGFR3	pp	HOXB13	IGHG1	pp	OGN	PIP	pp	SCGB1D1	BRCA1	pp	S100A7	
KLK11	pp	LAMP3	FGFR3	pp	SPINK5	IGHG1	pp	TTYH1	ASPN	pp	KLK11	BRCA1	pp	SLC1A1	
KLK11	pp	MSMB	FGFR3	pp	SLC26A3	IGHG1	pp	CCL18	ASPN	pp	S100A7	BRCA1	pp	TNNI2	
CHIT1	pp	TFPI2	FGFR3	pp	FABP7	IGHG1	pp	CD19	ASPN	pp	CCL18	TAT	pp	B3GNT3	
PARK7	pp	CHIT1	FGFR3	pp	ASCL2	KCNJ16	pp	HSD17B2	ASPN	pp	CCL19	TAT	pp	KLK11	
PARK7	pp	SSBP2	FGFR3	pp	TTYH1	KRT4	pp	NRG2	TMEM100	pp	B3GNT3	TAT	pp	COL11A1	
PARK7	pp	CITED1	FGFR3	pp	C2orf54	KRT16	pp	COL17A1	TMEM100	pp	SSBP2	TAT	pp	CSTA	
PARK7	pp	MSMB	FGFR3	pp	NRG2	KRT16	pp	EGF	TMEM100	pp	MSMB	TAT	pp	SLC26A3	
PARK7	pp	SIX1	FOXJ1	pp	SPINK5	KRT16	pp	ERBB2	TMEM100	pp	NTRK3	TAT	pp	EDN2	
PARK7	pp	DLK1	FOXJ1	pp	WIF1	KRT16	pp	FGFR3	TMEM100	pp	S100P	TAT	pp	MSMB	
COL11A1	pp	WIF1	FOXJ1	pp	LAMP3	KRT16	pp	TFCP2L1	TMEM100	pp	TFF1	TAT	pp	ATP6VOA4	
COL11A1	pp	GSTA3	FOXJ1	pp	TFCP2L1	KRT16	pp	AMIGO2	TMEM100	pp	RERGL	TAT	pp	PIP	
COL11A1	pp	S100A7	FOXJ1	pp	SYT13	KRT16	pp	KCNJ16	ZNF334	pp	FOXJ1	TAT	pp	ASPN	
COL17A1	pp	EDN3	FOXJ1	pp	SLC1A1	KRT16	pp	OGN	ZNF334	pp	GSTA3	TAT	pp	TMEM100	
COL17A1	pp	CITED1	FN1	pp	CDKN2A	KRT16	pp	VGLL1	ZNF334	pp	HSD17B2	TAT	pp	TFF1	
COMP	pp	TUBB3	FN1	pp	KLK11	KRT16	pp	TTYH1	ZNF334	pp	AMIGO2	TAT	pp	TFPI2	
COMP	pp	SLC34A2	FN1	pp	COL11A1	KRT16	pp	ACTA1	ZNF334	pp	MSMB	TAT	pp	NRG2	
COMP	pp	KLK11	FN1	pp	COMP	KRT16	pp	FGF18	ZNF334	pp	ASPN	TFF1	pp	EDN3	
COMP	pp	COL11A1	FN1	pp	CPA3	LYZ	pp	PIP	ZNF334	pp	TNNI2	TFF1	pp	AMIGO2	
COMP	pp	ASCL2	FN1	pp	SLC26A3	LYZ	pp	CCL18	ZNF334	pp	REPS2	TFF1	pp	CITED1	
COMP	pp	MMP12	FN1	pp	EGF	ASCL2	pp	TUBB3	TTYH1	pp	B3GNT3	TFF1	pp	MSMB	
COMP	pp	TMEM100	FN1	pp	FGFR3	ASCL2	pp	CLGN	TTYH1	pp	OLFM4	TFF1	pp	PENK	
COMP	pp	ZNF334	FN1	pp	GREM1	ASCL2	pp	SCGB1D1	TTYH1	pp	SLC34A2	TFF1	pp	SYT13	
COMP	pp	SLC1A1	FN1	pp	NRG1	ASCL2	pp	CSTA	TTYH1	pp	COL17A1	TFF1	pp	CCL18	
COMP	pp	TNNI2	FN1	pp	MMP12	ASCL2	pp	EDN2	TTYH1	pp	FABP7	TFF1	pp	NRG2	
COMP	pp	TFPI2	FN1	pp	CITED1	ASCL2	pp	FOXJ1	TTYH1	pp	KCNJ16	TNNI2	pp	WIF1	
COMP	pp	CLDN8	FN1	pp	ATP6VOA4	ASCL2	pp	PRAME	TTYH1	pp	VGLL1	RERGL	pp	CLGN	
CPA3	pp	TUBB3	FN1	pp	TTYH1	ASCL2	pp	SSBP2	TTYH1	pp	ASPN	RERGL	pp	HOXB13	
CPA3	pp	TFPI2	FN1	pp	SYT13	ASCL2	pp	GSTA3	TTYH1	pp	CCL18	RERGL	pp	PRAME	
CSTA	pp	SPINK5	FN1	pp	ACTA1	ASCL2	pp	SYT13	TTYH1	pp	SIX1	RERGL	pp	OGN	
CSTA	pp	DLK1	FN1	pp	S100P	ASCL2	pp	DLK1	TTYH1	pp	SLC1A1	RERGL	pp	S100A7	
CSTA	pp	NRG2	FN1	pp	CCL18	ASCL2	pp	REPS2	TTYH1	pp	RERGL	RERGL	pp	CCL19	
SLC26A3	pp	SPINK5	FN1	pp	SOX11	MMP12	pp	HOXB13	SYT13	pp	MSMB	TFPI2	pp	LAMP3	
SLC26A3	pp	HSD17B2	FN1	pp	BRCA1	MMP12	pp	CHIT1	SYT13	pp	NTRK3	C2orf54	pp	TUBB3	
SLC26A3	pp	OGN	FN1	pp	C2orf54	MMP12	pp	LAMP3	ACTA1	pp	SPINK5	C2orf54	pp	COL11A1	

Εικόνα Β.8 – Οι αλληλεπιδράσεις που προέκυψαν για την περίπτωση των υγιών δειγμάτων, συνεχών μεταβλητών και custom ταξινόμηση κόμβων.

Παράρτημα C

Σε αυτόν τον πίνακα φαίνονται οι κοινές αλληλεπιδράσεις που προέκυψαν συγκρίνοντας τα δίκτυα ΚΕ και ΥΕ. Ο συνολικός αριθμός τους είναι 171.

ΚΟΙΝΕΣ ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΔΙΚΤΥΩΝ – ΚΑΡΚΙΝΙΚΩΝ-ΥΓΕΙΩΝ ΔΕΙΓΜΑΤΩΝ														
γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2	γονίδιο1	σύνδ	γονίδιο2
CDKN2A	pp	MSMB	EDN3	pp	TFF1	NRG1	pp	AMIGO2	OGN	pp	CILP	TAT	pp	REPS2
TUBB3	pp	FN1	EDN3	pp	NRG2	NRG1	pp	KCNJ16	OGN	pp	DLK1	TAT	pp	NRG2
TUBB3	pp	TFPI2	EGF	pp	PARK7	NRG1	pp	OGN	ATP6V0A4	pp	B3GNT3	TFF1	pp	CITED1
OLFM4	pp	B3GNT3	EGF	pp	FGFR3	NR4A1	pp	COL17A1	ATP6V0A4	pp	TTYH1	TFF1	pp	MSMB
OLFM4	pp	SCGB1D1	EGF	pp	HSD17B2	NR4A1	pp	FGF18	VGLL1	pp	B3GNT3	TFF1	pp	SYT13
OLFM4	pp	FN1	ERBB2	pp	ASCL2	IGHG1	pp	B3GNT3	VGLL1	pp	GSTA3	TFF1	pp	TFPI2
SLC34A2	pp	B3GNT3	ERBB2	pp	ASPN	IGHG1	pp	PARK7	VGLL1	pp	ATP6V0A4	TNNI2	pp	B3GNT3
SLC34A2	pp	CPA3	ERBB2	pp	TMEM100	IGHG1	pp	TTYH1	PENK	pp	NRG2	TNNI2	pp	WIF1
SLC34A2	pp	EDN2	FABP7	pp	TTYH1	IGHG1	pp	CCL18	PIP	pp	SCGB1D1	TNNI2	pp	NRG1
SLC34A2	pp	ERBB2	FABP7	pp	NRG2	IGHG1	pp	CD19	PIP	pp	CITED1	TNNI2	pp	BRCA1
SLC34A2	pp	HSD17B2	FGFR3	pp	BRCA1	KCNJ16	pp	COL17A1	ASPN	pp	GREM1	RERGL	pp	CCL19
SLC34A2	pp	MSMB	FOXJ1	pp	SLC1A1	KCNJ16	pp	S100P	TMEM100	pp	NTRK3	TFPI2	pp	OLFM4
SCGB1D1	pp	PRAME	FOXJ1	pp	C2orf54	KRT4	pp	ATP6V0A4	TMEM100	pp	RERGL	C2orf54	pp	SPINK5
SCGB1D1	pp	REPS2	FN1	pp	WIF1	KRT16	pp	SLC34A2	ZNF334	pp	COL17A1	C2orf54	pp	TFCP2L1
PPARGC1A	pp	TUBB3	FN1	pp	COL11A1	KRT16	pp	COL17A1	TTYH1	pp	VGLL1	FGF18	pp	KRT16
PPARGC1A	pp	HSD17B2	FN1	pp	COMP	KRT16	pp	FGFR3	TTYH1	pp	ZNF334	FGF18	pp	VGLL1
PPARGC1A	pp	TMEM100	FN1	pp	GREM1	KRT16	pp	TFCP2L1	TTYH1	pp	SLC1A1	FGF18	pp	ASPN
PPARGC1A	pp	REPS2	FN1	pp	ASPN	KRT16	pp	AMIGO2	ACTA1	pp	KRT16	FGF18	pp	TMEM100
WIF1	pp	KCNJ16	FN1	pp	TTYH1	KRT16	pp	VGLL1	ACTA1	pp	LYZ	FGF18	pp	TTYH1
WIF1	pp	NRG2	FN1	pp	ACTA1	KRT16	pp	TTYH1	S100P	pp	TUBB3	FGF18	pp	RERGL
COL11A1	pp	COMP	FN1	pp	S100P	KRT16	pp	SIX1	S100P	pp	KRT4	CLDN8	pp	COL17A1
COL11A1	pp	OGN	FN1	pp	SOX11	LYZ	pp	MMP12	CCL18	pp	S100A7	REPS2	pp	TUBB3
COL17A1	pp	FN1	FN1	pp	C2orf54	LYZ	pp	CCL18	CCL18	pp	S100A9	REPS2	pp	CCL18
COL17A1	pp	AMIGO2	FN1	pp	CILP	ASCL2	pp	CSTA	CCL18	pp	CCL19			
COL17A1	pp	CITED1	PRAME	pp	ASCL2	ASCL2	pp	SSBP2	SIX1	pp	SLC34A2			
COMP	pp	SLC34A2	PRAME	pp	SLC1A1	MMP12	pp	CHIT1	SIX1	pp	MSMB			
COMP	pp	ASPN	SSBP2	pp	CPA3	MMP12	pp	CCL19	SIX1	pp	ATP6V0A4			
CPA3	pp	FN1	SSBP2	pp	AMIGO2	CITED1	pp	OLFM4	SOX11	pp	MMP12			
CPA3	pp	CILP	SSBP2	pp	REPS2	CITED1	pp	TFPI2	BRCA1	pp	WIF1			
CSTA	pp	SPINK5	LAMP3	pp	CD19	MSMB	pp	CLGN	BRCA1	pp	HSD17B2			
CSTA	pp	SYT13	GSTA1	pp	TFCP2L1	MSMB	pp	SLC1A1	BRCA1	pp	MMP12			
SLC26A3	pp	HSD17B2	GSTA1	pp	MMP12	MSMB	pp	TFPI2	BRCA1	pp	OGN			
SLC26A3	pp	OGN	GSTA3	pp	GSTA1	OGN	pp	CPA3	BRCA1	pp	TMEM100			
EDN2	pp	ASCL2	GSTA3	pp	TTYH1	OGN	pp	MMP12	TAT	pp	SLC26A3			
EDN3	pp	OLFM4	TFCP2L1	pp	FOXJ1	OGN	pp	PENK	TAT	pp	PENK			
EDN3	pp	WIF1	NRG1	pp	WIF1	OGN	pp	TNNI2	TAT	pp	PIP			
FNX3	pp	PENK	NRG1	pp	PARK7	OGN	pp	RFRG1	TAT	pp	TFF1			

Εικόνα C.1 – Οι κοινές αλληλεπιδράσεις που εμφάνισαν τα δίκτυα ΚΕ και ΥΕ.

Παράρτημα D

Παρακάτω παρατίθενται οι πίνακες με τα γονίδια που σχετίζονται ή δεν έχουν συσχετιστεί ακόμα με τον καρκίνο του μαστού σύμφωνα με τη βάση δεδομένων G2SBC.

ΓΟΝΙΔΙΑ - ΣΧΕΤΙΖΟΜΕΝΑ ΜΕ ΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ			
Κωδικός Γονιδίου	Συμβολικό Γονιδίο	Αριθμός Μεταλλάξεων Αλληλοπεριβάσεων	Περιγραφή
5304	PIP	1	prolactin-induced protein
57586	SYT13	2	synaptotagmin XIII
6505	SLC1A1	3	solute carrier family 1 (neuronal/epithelial high affinity glutamate transporter, system Xag), member 1
2261	FGFR3	7	fibroblast growth factor receptor 3
672	BRCA1	124	breast cancer 1, early onset
7136	TNNI2	2	troponin I type 2 (skeletal, fast)
1475	CSTA	4	cystatin A (stefin A)
29842	TFCP2L1	0	transcription factor CP2-like 1
27074	LAMP3	0	lysosomal-associated membrane protein 3
2940	GSTA3	0	glutathione S-transferase alpha 3
10568	SLC34A2	0	solute carrier family 34 (sodium phosphate), member 2
57348	TTYH1	0	twisty homolog 1 (Drosophila)
4435	CITED1	5	Chip/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1
23532	PRAME	1	preferentially expressed antigen in melanoma
1907	EDN2	0	endothelin 2
23635	SSBP2	3	single-stranded DNA binding protein 2
3868	KRT16	0	keratin 16
4321	MMP12	0	matrix metalloproteinase 12 (macrophage elastase)
1047	CLGN	1	calnequin
1029	CDKN2A	15	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
6363	CCL19	2	chemokine (C-C motif) ligand 19
2302	FOXJ1	0	forkhead box J1
1950	EGF	11	epidermal growth factor (beta-urethane)
9185	REPS2	8	RALBP1 associated Eps domain containing 2
6664	SOX11	2	SOX (sex determining region Y)-box 11
7980	TFPI2	3	tissue factor pathway inhibitor 2
2173	FABP7	0	fatty acid binding protein 7, brain
2335	FN1	52	fibronectin 1
6362	CCL18	0	chemokine (C-C motif) ligand 18 (pulmonary and activation-regulated)
6278	S100A7	5	S100 calcium binding protein A7
51442	VGLL1	0	vestigial like 1 (Drosophila)
930	CD19	14	CD19 molecule
8817	FGF18	0	fibroblast growth factor 18
11315	PARK7	0	Parkinson disease (autosomal recessive, early onset) 7
4916	NTRK3	8	neurotrophic tyrosine kinase, receptor, type 3
3084	NRG1	6	neuregulin 1
6280	S100A9	8	S100 calcium binding protein A9
4969	OGN	0	osteglycin
7031	TFF1	2	trefoil factor 1
3164	NR4A1	11	nuclear receptor subfamily 4, group A, member 1
54829	ASPN	0	asporin
4069	LYZ	6	lysozyme (renal amyloidosis)
2064	ERBB2	36	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
1308	COL17A1	0	collagen, type XVII, alpha 1

Εικόνα D.1 – Τα 44 γονίδια που σχετίζονται με τον καρκίνο του μαστού. Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που χαρακτηρίστηκαν κεντρικοί κόμβοι στο καρκινικό (κόκκινη) ή φυσιολογικό (πράσινη) δίκτυο, ή και στα δυο (μπλε) δίκτυα. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα γονίδια που εντοπίστηκαν στο καρκινικό (ανοιχτό μωβ) ή φυσιολογικό (ανοιχτό πράσινο) δίκτυο, ή και στα δυο (τυρκουάζ) δίκτυα.

ΓΟΝΙΔΙΑ - ΜΗ ΣΧΕΤΙΖΟΜΕΝΑ ΜΕ ΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ			
Κωδικός Γονιδίου	Σύμβολο Γονιδίου	Αριθμός Προτεινόμενων Αλληλεπιδράσεων	Περιγραφή
430	ASCL2	7	achaete-scute complex homolog 2 (Drosophila)
11197	WIF1	0	WNT inhibitory factor 1
10441	ROXB13	1	homeobox 13
3500	IGHG1	4	immunoglobulin heavy constant gamma 1 (430 markers)
10491	PPARGC1A	20	peroxisome proliferator-activated receptor gamma, coactivator 1 alpha
8788	DLE1	2	delta-like 1 homolog (Drosophila)
347902	AMIGO2	1	adhesion molecule with Ig-like domain 2
1301	COL11A1	1	collagen, type XI, alpha 1
1411	SLC26A3	1	solute carrier family 26, member 3
9073	CLDN8	1	claudin 8
2938	GSTA1	1	glutathione S-transferase alpha 1
3294	HSD17B2	0	hydroxysteroid (17-beta) dehydrogenase 2
10562	OLFM4	1	olfactomedin 4
6286	S100P	6	S100 calcium binding protein P
3851	KRT4	0	keratin 4
8483	CDP	0	cartilage intermediate layer protein, nucleotide pyrophosphatylhydrolase
79785	REGL	0	REGL/BAS-like
10381	TUBB3	0	tubulin, beta 3
11012	KLEC1	0	keratin-related peptidase 11
6898	TAT	0	tyrosine aminotransferase
1359	CPA3	0	carboxypeptidase A3 (mam cell)
6495	SIX1	7	SIX homeobox 1
26585	GREM1	0	gremlin 1, cysteine knot superfamily, homolog (Xenopus laevis)
1311	COMP	1	cartilage oligomeric matrix protein
1908	EDN3	2	endothelin 3
55273	TMEM100	0	transmembrane protein 100
58	ACTA1	29	actin, alpha 1, skeletal muscle
10331	KCNT3	0	GTP-binding protein beta-1,3-AT-activated guanine nucleotide exchange factor 3
55713	ZNF334	0	zinc finger protein 334
3773	KCNJ16	0	potassium inwardly-rectifying channel, subfamily J, member 16
4477	MSMB	0	microseminoprotein, beta-
5179	PENK	0	proenkephalin
9542	NRG2	2	neuregulin 2
50617	ATP6V0A4	0	ATPase, H+ transporting, lysosomal V0 subunit a4
79919	CZorf54	0	chromosome 2 open reading frame 54
11005	SPINK5	1	serine peptidase inhibitor, Kazal type 5
10648	SCGB1D1	0	secretoglobulin, family 1D, member 1
1118	CHIT1	0	chitinase 1 (chitinase)

Εικόνα D.2 – Τα 38 γονίδια που δεν έχουν συσχετιστεί ακόμα με τον καρκίνο του μαστού. Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που χαρακτηρίστηκαν κεντρικοί κόμβοι στο καρκινικό (κόκκινη) ή φυσιολογικό (πράσινη) δίκτυο. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα γονίδια που εντοπίστηκαν στο καρκινικό (ανοιχτό μωβ) ή φυσιολογικό (ανοιχτό πράσινο) δίκτυο.

Παράρτημα Ε

Παρακάτω παρατίθενται οι πίνακες με τα σημαντικότερα μοριακά μονοπάτια που προέκυψαν από την ανάλυση εμπλουτισμού (WebGestalt) των γονιδίων που συνθέτουν τα εκάστοτε συμπλέγματα MCODE ή τις δομικές ενότητες jActiveModules.

ΜΟΡΙΑΚΑ ΜΟΝΟΠΑΤΙΑ ΣΤΑ ΣΥΜΠΛΕΓΜΑΤΑ MCODE			
MONOPATHIA WikiPathways			
Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική
ΣΥΜΠΛΕΓΜΑ 1-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
Focal Adhesion	3	1301 2335 1950	adjP=0.0002
Senescence and Autophagy	2	2335 1029	adjP=0.0018
Regulation of Actin Cytoskeleton	2	2335 1950	adjP=0.0020
ΣΥΜΠΛΕΓΜΑ 2-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
ErbB signaling pathway	3	9542 2064 3084	adjP=8.38e-06
EGF-EGFR Signaling Pathway	2	2064 9185	adjP=0.0029
ΣΥΜΠΛΕΓΜΑ 3-ΦΥΣΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ			
ErbB signaling pathway	2	2064 1950	adjP=0.0006
Focal Adhesion	2	2064 1950	adjP=0.0025
EGF-EGFR Signaling Pathway	2	2064 1950	adjP=0.0025
MONOPATHIA KEGG			
Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική
ΣΥΜΠΛΕΓΜΑ 1-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
Pathways in cancer	3	2335 1029 1950	adjP=0.0006
Focal adhesion	3	1301 2335 1950	adjP=0.0006
ECM-receptor interaction	2	1301 2335	adjP=0.0008
Regulation of actin cytoskeleton	2	2335 1950	adjP=0.0033
Cytokine-cytokine receptor interaction	2	6363 1950	adjP=0.0047
ΣΥΜΠΛΕΓΜΑ 2-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
ErbB signaling pathway	3	2064 9542 3084	adjP=2.65e-05
Pathways in cancer	2	2064 2261	adjP=0.0099
ΣΥΜΠΛΕΓΜΑ 3-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
Metabolic pathways	2	50617 10331	adjP=0.0442
ΣΥΜΠΛΕΓΜΑ 4-ΚΑΡΚΙΝΙΚΟ ΔΙΚΤΥΟ			
MAPK signaling pathway	2	3164 8817	adjP=0.0024
ΣΥΜΠΛΕΓΜΑ 2-ΦΥΣΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ			
Phagosome	2	1311 10381	adjP=0.0016
MAPK signaling pathway	2	3164 8817	adjP=0.0024
ΣΥΜΠΛΕΓΜΑ 3-ΦΥΣΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ			
Pathways in cancer	3	2064 1029 1950	adjP=0.0004
Protein digestion and absorption	2	1308 1359	adjP=0.0006
ErbB signaling pathway	2	2064 1950	adjP=0.0006

Εικόνα Ε.1 – Τα μοριακά μονοπάτια WikiPathways και KEGG των συμπλεγμάτων MCODE για το καρκινικό και το φυσιολογικό δίκτυο. Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που διαφοροποιούνται στα κοινά μονοπάτια, δηλαδή συμμετέχουν αποκλειστικά στο καρκινικό δίκτυο (κόκκινη) ή το φυσιολογικό (μωβ) δίκτυο. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα διαφορετικά μονοπάτια ανάμεσα στο καρκινικό και το φυσιολογικό δίκτυο.

ΜΟΝΟΠΑΤΙΑ KEGG ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΚΑΡΚΙΝΙΚΟΥ ΔΙΚΤΥΟΥ (1/2)				ΜΟΝΟΠΑΤΙΑ KEGG ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΦΥΣΙΟΛΟΓΙΚΟΥ ΔΙΚΤΥΟΥ (1/2)			
Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική	Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική
ECM-receptor interaction				ECM-receptor interaction			
ΔΕ 1-ΚΑ	3	1301 2335 1311	adjP=0.0003	ΔΕ 1-ΦΔ	3	1301 2335 1311	adjP=0.0004
ΔΕ 2-ΚΑ	3	1301 2335 1311	adjP=0.0002	ΔΕ 2-ΦΔ	3	1301 2335 1311	adjP=0.0003
ΔΕ 3-ΚΑ	3	1301 2335 1311	adjP=0.0006	ΔΕ 3-ΦΔ	3	1301 2335 1311	adjP=0.0005
ΔΕ 4-ΚΑ	3	1301 2335 1311	adjP=0.0001	ΔΕ 4-ΦΔ	3	1301 2335 1311	adjP=0.0004
ΔΕ 5-ΚΑ	3	1301 2335 1311	adjP=0.0005	ΔΕ 5-ΦΔ	3	1301 2335 1311	adjP=0.0002
Pathways in cancer				Pathways in cancer			
ΔΕ 1-ΚΑ	4	2261 2335 1029 8817	adjP=0.0004	ΔΕ 1-ΦΔ	4	2261 2335 1029 8817	adjP=0.0004
ΔΕ 2-ΚΑ	4	2261 2335 1029 8817	adjP=0.0003	ΔΕ 2-ΦΔ	2	2261 2335 —(1029 8817)	adjP=0.0258
ΔΕ 3-ΚΑ	4	2261 2335 1029 8817	adjP=0.0006	ΔΕ 3-ΦΔ	4	2261 2335 1029 8817	adjP=0.0006
ΔΕ 4-ΚΑ	5	2261 2335 1029 8817 1950	adjP=6.32e-05	ΔΕ 4-ΦΔ	4	2261 2335 1029 8817	adjP=0.0005
ΔΕ 5-ΚΑ	4	2261 2335 1029 8817	adjP=0.0005	ΔΕ 5-ΦΔ	5	2064 2261 2335 1029 8817	adjP=4.84e-05
Focal adhesion				Focal adhesion			
ΔΕ 1-ΚΑ	3	1301 2335 1311	adjP=0.0009	ΔΕ 1-ΦΔ	3	1301 2335 1311	adjP=0.0011
ΔΕ 2-ΚΑ	3	1301 2335 1311	adjP=0.0012	ΔΕ 2-ΦΔ	3	1301 2335 1311	adjP=0.0021
ΔΕ 3-ΚΑ	3	1301 2335 1311	adjP=0.0016	ΔΕ 3-ΦΔ	3	1301 2335 1311	adjP=0.0015
ΔΕ 4-ΚΑ	4	1301 2335 1311 1950	adjP=0.0001	ΔΕ 4-ΦΔ	3	1301 2335 1311	adjP=0.0012
ΔΕ 5-ΚΑ	3	1301 2335 1311	adjP=0.0014	ΔΕ 5-ΦΔ	4	2064 1301 2335 1311	adjP=0.0001
Regulation of actin cytoskeleton				Regulation of actin cytoskeleton			
ΔΕ 1-ΚΑ	3	2261 2335 8817	adjP=0.0009	ΔΕ 1-ΦΔ	3	2261 2335 8817	adjP=0.0011
ΔΕ 2-ΚΑ	3	2261 2335 8817	adjP=0.0012	ΔΕ 2-ΦΔ	2	2261 2335 —(8817)	adjP=0.0131
ΔΕ 3-ΚΑ	3	2261 2335 8817	adjP=0.0016	ΔΕ 3-ΦΔ	3	2261 2335 8817	adjP=0.0015
ΔΕ 4-ΚΑ	4	2261 2335 8817 1950	adjP=0.0001	ΔΕ 4-ΦΔ	3	2261 2335 8817	adjP=0.0012
ΔΕ 5-ΚΑ	3	2261 2335 8817	adjP=0.0014	ΔΕ 5-ΦΔ	3	2261 2335 8817	adjP=0.0019

Εικόνα Ε.2 – Τα μοριακά μονοπάτια KEGG των δομικών ενότητων jActiveModules για το καρκινικό και το φυσιολογικό δίκτυο (1/2). Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που διαφοροποιούνται στα κοινά μονοπάτια, δηλαδή συμμετέχουν αποκλειστικά στο καρκινικό δίκτυο (κόκκινη) ή το φυσιολογικό (μωβ) δίκτυο. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα διαφορετικά μονοπάτια ανάμεσα στο καρκινικό και το φυσιολογικό δίκτυο. ΔΕ-Δομική Ενότητα, ΚΔ-Καρκινικό Δίκτυο, ΦΔ-Φυσιολογικό Δίκτυο.

ΜΟΝΟΠΑΤΙΑ KEGG ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΚΑΡΚΙΝΙΚΟΥ ΔΙΚΤΥΟΥ (2/2)				ΜΟΝΟΠΑΤΙΑ KEGG ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΦΥΣΙΟΛΟΓΙΚΟΥ ΔΙΚΤΥΟΥ (2/2)			
Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική	Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική
Phagosome				Phagosome			
ΔΕ 1-ΚΑ	2	1311 10381	adjP=0.0061	ΔΕ 1-ΦΑ	2	1311 10381	adjP=0.0073
ΔΕ 2-ΚΑ	2	1311 10381	adjP=0.0069	ΔΕ 2-ΦΑ	2	1311 10381	adjP=0.0078
ΔΕ 3-ΚΑ	2	1311 10381	adjP=0.0075	ΔΕ 3-ΦΑ	2	1311 10381	adjP=0.0077
ΔΕ 4-ΚΑ	2	1311 10381	adjP=0.0083	ΔΕ 4-ΦΑ	2	1311 10381	adjP=0.0065
ΔΕ 5-ΚΑ	2	1311 10381	adjP=0.0077	ΔΕ 5-ΦΑ	2	1311 10381	adjP=0.0079
MAPK signaling pathway				MAPK signaling pathway			
ΔΕ 1-ΚΑ	2	2261 8817	adjP=0.0159	ΔΕ 1-ΦΑ	2	2261 8817	adjP=0.0189
ΔΕ 2-ΚΑ	3	2261 3164 8817	adjP=0.0019	ΔΕ 2-ΦΑ	none	none	none
ΔΕ 3-ΚΑ	2	2261 8817	adjP=0.0200	ΔΕ 3-ΦΑ	2	2261 8817	adjP=0.0205
ΔΕ 4-ΚΑ	3	2261 8817 1950	adjP=0.0027	ΔΕ 4-ΦΑ	3	2261 3164 8817	adjP=0.0020
ΔΕ 5-ΚΑ	2	2261 8817	adjP=0.0200	ΔΕ 5-ΦΑ	2	2261 8817	adjP=0.0211
Protein digestion and absorption				Protein digestion and absorption			
ΔΕ 1-ΚΑ	none	none	none	ΔΕ 1-ΦΑ	none	none	none
ΔΕ 2-ΚΑ	3	1301 1308 1359	adjP=0.0002	ΔΕ 2-ΦΑ	none	none	none
ΔΕ 3-ΚΑ	none	none	none	ΔΕ 3-ΦΑ	none	none	none
ΔΕ 4-ΚΑ	2	1301 1308	adjP=0.0032	ΔΕ 4-ΦΑ	2	1301 1308	adjP=0.0024
ΔΕ 5-ΚΑ	2	1301 6505	adjP=0.0029	ΔΕ 5-ΦΑ	2	1301 6505	adjP=0.0030
Glutathione metabolism				Glutathione metabolism			
ΔΕ 1-ΚΑ	none	none	none	ΔΕ 1-ΦΑ	none	none	none
ΔΕ 2-ΚΑ	none	none	none	ΔΕ 2-ΦΑ	2	2940 2938	adjP=0.0021
ΔΕ 3-ΚΑ	2	2940 2938	adjP=0.0016	ΔΕ 3-ΦΑ	2	2940 2938	adjP=0.0015
ΔΕ 4-ΚΑ	none	none	none	ΔΕ 4-ΦΑ	none	none	none
ΔΕ 5-ΚΑ	none	none	none	ΔΕ 5-ΦΑ	none	none	none
Metabolism of xenobiotics by cytochrome P450				Metabolism of xenobiotics by cytochrome P450			
ΔΕ 1-ΚΑ	none	none	none	ΔΕ 1-ΦΑ	none	none	none
ΔΕ 2-ΚΑ	none	none	none	ΔΕ 2-ΦΑ	2	2940 2938	adjP=0.0025
ΔΕ 3-ΚΑ	2	2940 2938	adjP=0.0021	ΔΕ 3-ΦΑ	2	2940 2938	adjP=0.0022
ΔΕ 4-ΚΑ	none	none	none	ΔΕ 4-ΦΑ	none	none	none
ΔΕ 5-ΚΑ	none	none	none	ΔΕ 5-ΦΑ	none	none	none
Drug metabolism - cytochrome P450				Drug metabolism - cytochrome P450			
ΔΕ 1-ΚΑ	none	none	none	ΔΕ 1-ΦΑ	none	none	none
ΔΕ 2-ΚΑ	none	none	none	ΔΕ 2-ΦΑ	2	2940 2938	adjP=0.0025
ΔΕ 3-ΚΑ	2	2940 2938	adjP=0.0021	ΔΕ 3-ΦΑ	2	2940 2938	adjP=0.0022
ΔΕ 4-ΚΑ	none	none	none	ΔΕ 4-ΦΑ	none	none	none
ΔΕ 5-ΚΑ	none	none	none	ΔΕ 5-ΦΑ	none	none	none
ErbB signaling pathway				ErbB signaling pathway			
ΔΕ 4-ΚΑ	2	9542 1950	adjP=0.0034	ΔΕ 5-ΦΑ	2	2064 9542	adjP=0.0031
Gap junction				Gap junction			
ΔΕ 4-ΚΑ	2	1950 10381	adjP=0.0034	None	none	none	none
Endocytosis				Endocytosis			
ΔΕ 4-ΚΑ	2	2261 1950	adjP=0.0131	None	none	none	none

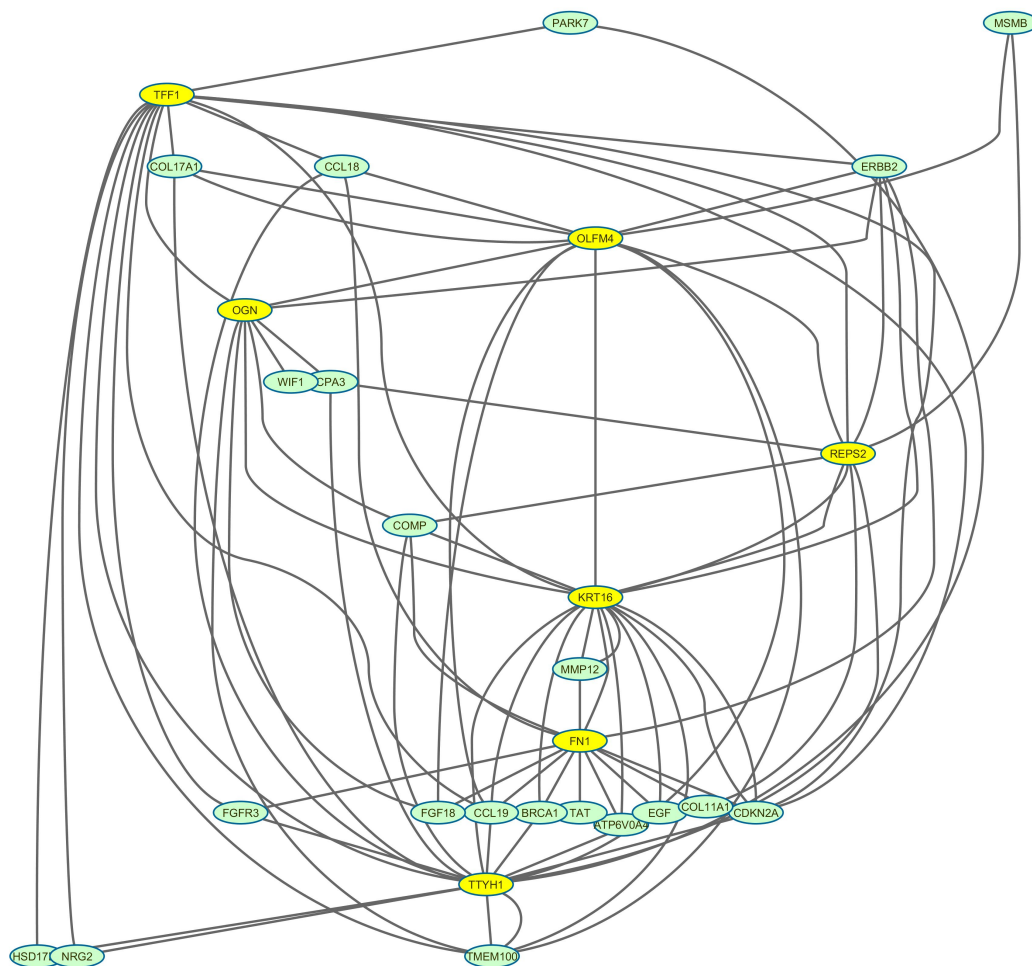
Εικόνα Ε.3 – Τα μοριακά μονοπάτια KEGG των δομικών ενοτήτων jActiveModules για το καρκινικό και το φυσιολογικό δίκτυο (2/2). Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που διαφοροποιούνται στα κοινά μονοπάτια, δηλαδή συμμετέχουν αποκλειστικά στο καρκινικό δίκτυο (κόκκινη) ή το φυσιολογικό (μωβ) δίκτυο. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα διαφορετικά μονοπάτια ανάμεσα στο καρκινικό και το φυσιολογικό δίκτυο. ΔΕ-Δομική Ενότητα, ΚΔ-Καρκινικό Δίκτυο, ΦΔ-Φυσιολογικό Δίκτυο.

ΜΟΝΟΠΑΤΙΑ WikiPathways ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΚΑΡΚΙΝΙΚΟΥ ΔΙΚΤΥΟΥ				ΜΟΝΟΠΑΤΙΑ WikiPathways ΣΤΙΣ ΔΟΜΙΚΕΣ ΕΝΟΤΗΤΕΣ jActiveModules ΤΟΥ ΦΥΣΙΟΛΟΓΙΚΟΥ ΔΙΚΤΥΟΥ			
Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική	Μοριακό Μονοπάτι	Αριθμός Γονιδίων	Κωδικός Γονιδίου	Στατιστική
Focal Adhesion				Focal Adhesion			
ΔΕ 1-ΚΑ	3	1301 2335 1311	adjP=0.0008	ΔΕ 1-ΦΔ	3	1301 2335 1311	adjP=0.0010
ΔΕ 2-ΚΑ	3	1301 2335 1311	adjP=0.0013	ΔΕ 2-ΦΔ	3	1301 2335 1311	adjP=0.0012
ΔΕ 3-ΚΑ	3	1301 2335 1311	adjP=0.0013	ΔΕ 3-ΦΔ	3	1301 2335 1311	adjP=0.0008
ΔΕ 4-ΚΑ	4	1301 2335 1311 1950	adjP=8.65e-05	ΔΕ 4-ΦΔ	3	1301 2335 1311	adjP=0.0014
ΔΕ 5-ΚΑ	3	1301 2335 1311	adjP=0.0015	ΔΕ 5-ΦΔ	4	2064 1301 2335 1311	adjP=0.0002
Regulation of Actin Cytoskeleton				Regulation of Actin Cytoskeleton			
ΔΕ 1-ΚΑ	3	2261 2335 8817	adjP=0.0008	ΔΕ 1-ΦΔ	3	2261 2335 8817	adjP=0.0010
ΔΕ 2-ΚΑ	3	2261 2335 8817	adjP=0.0013	ΔΕ 2-ΦΔ	2	2261 2335 8817	adjP=0.0085
ΔΕ 3-ΚΑ	3	2261 2335 8817	adjP=0.0013	ΔΕ 3-ΦΔ	3	2261 2335 8817	adjP=0.0008
ΔΕ 4-ΚΑ	4	2261 2335 1950 8817	adjP=8.65e-05	ΔΕ 4-ΦΔ	3	2261 2335 8817	adjP=0.0014
ΔΕ 5-ΚΑ	3	2261 2335 8817	adjP=0.0015	ΔΕ 5-ΦΔ	3	2261 2335 8817	adjP=0.0009
Endochondral Ossification				Endochondral Ossification			
ΔΕ 1-ΚΑ	2	2261 8817	adjP=0.0018	ΔΕ 1-ΦΔ	2	2261 8817	adjP=0.0023
ΔΕ 2-ΚΑ	2	2261 8817	adjP=0.0025	ΔΕ 2-ΦΔ	none	none	none
ΔΕ 3-ΚΑ	2	2261 8817	adjP=0.0025	ΔΕ 3-ΦΔ	2	2261 8817	adjP=0.0021
ΔΕ 4-ΚΑ	2	2261 8817	adjP=0.0037	ΔΕ 4-ΦΔ	2	2261 8817	adjP=0.0033
ΔΕ 5-ΚΑ	2	2261 8817	adjP=0.0030	ΔΕ 5-ΦΔ	2	2261 8817	adjP=0.0027
Senescence and Autophagy				Senescence and Autophagy			
ΔΕ 1-ΚΑ	2	2335 1029	adjP=0.0034	ΔΕ 1-ΦΔ	2	2335 1029	adjP=0.0041
ΔΕ 2-ΚΑ	2	2335 1029	adjP=0.0043	ΔΕ 2-ΦΔ	none	none	none
ΔΕ 3-ΚΑ	2	2335 1029	adjP=0.0043	ΔΕ 3-ΦΔ	2	2335 1029	adjP=0.0041
ΔΕ 4-ΚΑ	2	2335 1029	adjP=0.0075	ΔΕ 4-ΦΔ	2	2335 1029	adjP=0.0057
ΔΕ 5-ΚΑ	2	2335 1029	adjP=0.0052	ΔΕ 5-ΦΔ	2	2335 1029	adjP=0.0067
ErbB signaling pathway				ErbB signaling pathway			
ΔΕ 4-ΚΑ	2	9542 1950	adjP=0.0037	ΔΕ 5-ΦΔ	2	9542 2064	adjP=0.0027
EGF-EGFR Signaling Pathway				EGF-EGFR Signaling Pathway			
ΔΕ 4-ΚΑ	2	9185 1950	adjP=0.0125	ΔΕ 5-ΦΔ	2	2064 9185	adjP=0.0112
Striated Muscle Contraction				Striated Muscle Contraction			
None	none	none	none	ΔΕ 3-ΦΔ	2	7136 58	adjP=0.0008

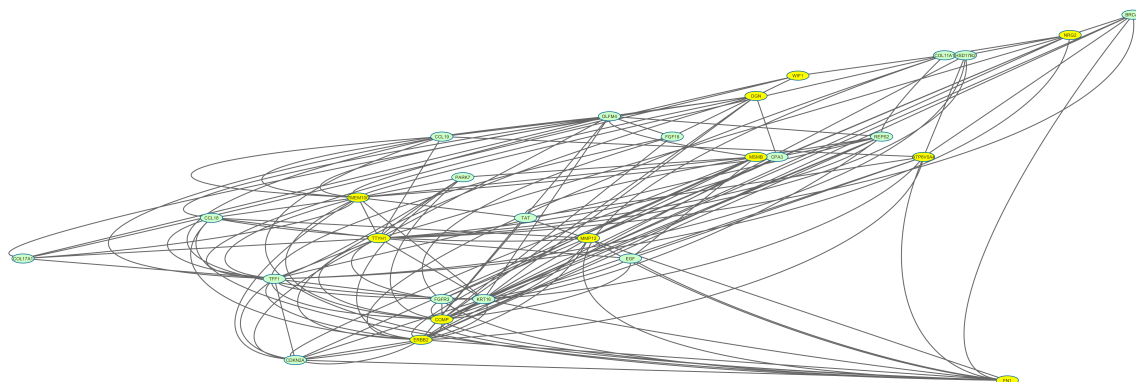
Εικόνα Ε.4 – Τα μοριακά μονοπάτια WikiPathways των δομικών ενοτήτων jActiveModules για το καρκινικό και το φυσιολογικό δίκτυο. Με έντονη γραμματοσειρά σημειώνονται τα γονίδια που διαφοροποιούνται στα κοινά μονοπάτια, δηλαδή συμμετέχουν αποκλειστικά στο καρκινικό δίκτυο (κόκκινη) ή το φυσιολογικό (μωβ) δίκτυο. Με έγχρωμο υπόβαθρο υπογραμμίζονται τα διαφορετικά μονοπάτια ανάμεσα στο καρκινικό και το φυσιολογικό δίκτυο. ΔΕ-Δομική Ενότητα, ΚΔ-Καρκινικό Δίκτυο, ΦΔ-Φυσιολογικό Δίκτυο.

Παράρτημα F

Παρακάτω παρατίθενται οι εικόνες με τις σημαντικότερες αλληλεπιδράσεις που εντοπίστηκαν στο υποδίκτυο διαφοροποίησης του καρκινικού σε σχέση με το φυσιολογικό δίκτυο με γνώμονα τους κεντρικούς κόμβους (καρκινικούς, φυσιολογικούς) και τα μοριακά μονοπάτια.



Εικόνα F.1 – Απεικόνιση των 82 γονιδιακών/πρωτεϊνικών αλληλεπιδράσεων ανάμεσα στους κεντρικούς καρκινικούς κόμβους (7 κόμβοι σε κίτρινο χρώμα) και τα γονίδια/πρωτεΐνες που συμμετέχουν σε εννέα σημαντικά ($p < 0.05$) μοριακά μονοπάτια στο υποδίκτυο διαφοροποίησης. Για την οπτικοποίηση των αλληλεπιδράσεων του υποδικτύου με κεντρικούς καρκινικούς κόμβους επιλέχθηκε η ιεραρχική διάταξη της πλατφόρμας Cytoscape.



Εικόνα F.2 – Απεικόνιση των 135 γονιδιακών/πρωτεϊνικών αλληλεπιδράσεων ανάμεσα στους κεντρικούς φυσιολογικούς κόμβους (11 κόμβοι σε κίτρινο χρώμα) και τα γονίδια/πρωτεΐνες που συμμετέχουν σε εννέα σημαντικά ($p < 0.05$) μοριακά μονοπάτια στο υποδίκτυο διαφοροποίησης. Για την οπτικοποίηση των αλληλεπιδράσεων του υποδικτύου με κεντρικούς φυσιολογικούς κόμβους επιλέχθηκε η ιεραρχική διάταξη της πλατφόρμας Cytoscape.