ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Σχολή Ηλεκτρονικών Μηχανικών και Μηχανικών

Υπολογιστών

Διπλωματική Εργασία

ΜΕΘΟΔΟΙ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ *"biclustering"* ΓΙΑ ΕΠΙΛΟΓΗ ΓΟΝΙΔΙΑΚΩΝ ΔΕΙΚΤΩΝ ΑΠΟ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ

ΑΛΕΒΥΖΑΚΗ ΑΝΔΡΟΝΙΚΗ

Επιβλέπων Καθηγητής :Καθηγητής Ζερβάκης Μιχάλης

Εξεταστική Επιτροπή: Καθηγητής Ζερβάκης Μιχάλης

Καθηγητής Λιάβας Αθανάσιος

Καθηγητής Πετράκης Ευριπίδης

Χανιά, Σεπτέμβριος 2014

Στη Μαίρη που έφυγε νωρίς…

Ευχαριστίες

Θα ήθελα να ευχαριστήσω

Τον Καθηγητή κύριο Ζερβάκη Μιχάλη, για την πολύτιμή βοήθεια και καθοδήγησή του σε όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας.

Τους Καθηγητές κύριο Λιάβα Αθανάσιο και κύριο Πετράκη Ευριπίδη για τη συνεισφορά τους ως μέλη της εξεταστικής επιτροπής.

Την Δρ. Obermayr για την παροχή του dataset.

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

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

Τον φίλο μου Λιοντήρη Ανδρέα για όλη τη δύναμη και το κουράγιο που μου προσέφερε μέχρι το τέλος της εργασίας.

Τον φίλο μου Παλόγο Γιάννη για τη βοήθειά του και το έναυσμα που μου έδωσε στην αρχή της εργασίας 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.

Η εργασία είναι αφιερωμένη στη Μαίρη που αν και έφυγε είναι πάντα κοντά μας!

Περίληψη

Η ανάλυση μικροσυστοιχιών DNA παρέχει τη δυνατότητα μελέτης της γονιδιακής έκφρασης. Το προφίλ της γονιδιακής έκφρασης που παράγεται, απεικονίζει το υποσύνολο των γονιδιακών μεταγραφών που εκφράζονται σε ένα κύτταρο ή έναν ιστό. Με την εξέλιξη της βιοπληροφορικής τα προφίλ της γονιδιακής έκφρασης αναλύονται με στόχο την επίλυση ζητημάτων που αφορούν γονίδια που εκφράζονται σε παθολογικές καταστάσεις. Τα δεδομένα γονιδιακής έκφρασης αναπαρίστανται με τη μορφή πινάκων όπου γραμμές αποτελούν τα γονίδια και στήλες τις διάφορες πειραματικές συνθήκες. Στόχο των διάφορων τεχνικών ομαδοποίησης αποτελεί η εξαγωγή σημαντικών βιολογικών πληροφοριών που αφορούν ομάδες γονίδιων κάτω από συγκεκριμένες συνθήκες ή αντίστροφα. Στην παρούσα εργασία πραγματοποιείται η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης δεδομένων γονιδιακής έκφρασης που προέρχονται από δείγματα παθολογίας του καρκίνου του μαστού, των ωοθηκών, του ενδομητρίου και του τραχήλου της μήτρας, συγκεντρωμένα σε ένα πίνακα δεδομένων με στόχο την τελική επιλογή γονιδιακών δεικτών από κυτταρικές σειρές. Η μεθοδολογία που προτείνεται περιλαμβάνει την εφαρμογή του αλγορίθμου διπλής κατηγοριοποίησης Cheng and Church εφαρμόζοντας σε αυτόν μια σειρά βελτιώσεων ούτως ώστε να εξάγουμε ταυτόχρονα ομάδες γονιδίων με ομοιόμορφη συμπεριφορά αλλά και συγκεκριμένο εύρος τιμών. Τα αποτελέσματα που προκύπτουν πέραν της στατιστικής τους σημαντικότητας αξιολογούνται βιολογικά εξετάζοντας τις διεργασίες και τα μονοπάτια που ενεργοποιούνται και υποδηλώνονται από τα γονίδια κάθε εξαγόμενης ομάδας και συσχετίζονται με την πραγματική παθολογία. Η βιολογική ανάλυση της μεθόδου διπλής κατηγοριοποίησης ανέδειξε: α) *ομάδες-γονιδιακών* παθολογικών δεικτών για κάθε καρκινικό τύπο και β) κύριες *ομάδες-γονιδιακών* δεικτών καρκίνου γ) 23 *πολυγονιδιακούς* καρκινικούς δείκτες κοινούς σε όλους τους τύπους δ) έναν *ειδικό* καρκινικό δείκτη για τους τέσσερις καρκινικούς τύπους και το σύνολο των τεσσάρων καρκινικών τύπων.

Abstract

DNA microarray analysis allows the study of gene expression. Gene expression generated profile illustrates the subset of gene transcripts that are expressed in a cell or a tissue. With the development of bioinformatics profiles of gene expression analyzed in order to solve issues related with genes expressed in pathological conditions. Gene expression data is represented as tables where rows refer to genes and columns to experimental conditions. Object of various clustering techniques is to extract relevant biological information on groups of genes under specific conditions or versa. Our present work refers to the method of biclustering of gene expression data classification from pathology of breast, ovarian, endometrial and cervical cancer, with the aim of selecting gene markers of cell lines. Proposed methodology involves the application of biclustering algorithm Cheng and Church with an number of improvements in order to extract simultaneously sets of co -regulated genes with small range. Results beyond their statistical significance, are biologically evaluated by examining processes and pathways of genes of each group and correlated it with actual pathology. Biological analysis has shown: a) gene-groups of cancer type specific pathologic markers, b) gene-groups of cancer markers, c) 23 multigene tumor markers and d) one specific tumor marker of each group and all cancer types.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

1.1 Ανάλυση του προβλήματος

Οι DNA μικροσυστοιχίες και η ανάλυση τους μετρούν τα επίπεδα έκφρασης μεγάλων αριθμών γονιδίων εντός μίας ομάδας διαφορετικών πειραματικών καταστάσεων. Η οπτικοποίηση αυτής της πληροφορίας αποτελεί μια από τις πιο σημαντικές και ελκυστικές για τους βιολόγους πτυχές έρευνας που στόχο έχουν τον προσδιορισμό της παθολογίας διαφόρων ασθενειών και την ανακάλυψη νέων γενετικών μονοπατιών. Συνήθως το σύνολο της γονιδιακής πληροφορίας συγκεντρώνεται σε πίνακες όπου κάθε γονίδιο αντιπροσωπεύει μία γραμμή και κάθε κατάσταση μια στήλη. Η ανάλυση αυτών των πινάκων γονιδιακής έκφρασης τοποθετείται σε δύο βασικές κατευθύνσεις: πρώτον την ομαδοποίηση γονιδίων ή καταστάσεων με βάση την πλειοψηφία των περιπτώσεων της παθολογίας και δεύτερον την κατηγοριοποίηση και πρόβλεψη νέων γονιδίων ή δειγμάτων στηριζόμενα σε ήδη γνωστή βιολογική πληροφορία. Οι τεχνικές ομαδοποίησης στοχεύουν σε αυτές τις κατευθύνσεις και έχουν αναπτυχθεί πολλαπλές μέθοδοι για τέτοιας μορφής ανάλυση με έντονη προσπάθεια στην αντιμετώπιση των διαφόρων προβλημάτων. Από τις πιο σημαντικές δυσκολίες που συναντώνται σε τέτοιες προσπάθειες είναι: η τυχαιότητα και ο θόρυβος των μετρήσεων καθώς και η έλλειψη ακρίβειας που συναντάται σε πολλές μεθόδους. Ιδιαίτερα οι διάφορες τεχνικές απλής ομαδοποίησης τείνουν να ομαδοποιούν την γενική εικόνα με αποτέλεσμα επιμέρους μικρές υποομάδες με χρήσιμη πληροφορία να μην θεωρούνται σημαντικές.

Στόχο αυτής της εργασίας αποτέλεσε η μελέτη αλγορίθμων ευαίσθητων στην πληροφορία και όχι στο θόρυβο και με ιδιαίτερη λεπτομέρεια και ακρίβεια στην ομαδοποίηση. Γι' αυτό το λόγο επιλέχθηκε η μεθοδολογία της διπλής κατηγοριοποίησης. Αποτελεί μια πιο σύγχρονη μέθοδο ταυτόχρονης ομαδοποίησης γραμμών και στηλών με βάση τις τελευταίες εξελίξεις της τεχνολογίας (state-of-the-art), που αφορά στην ομαδοποίηση δεδομένων με βάση την ένταση των δεδομένων ή το εύρος αυτών. Καινοτομία της συγκεκριμένης διπλωματικής αποτέλεσε η στόχευση και στις δύο αυτές παραμέτρους που ο συνδυασμός τους παρέχει αποτελέσματα με μεγάλη στατιστική σημαντικότητα. Πέραν όμως αυτής επιβεβαιώθηκε και η βιολογική σημαντικότητα εξετάζοντας τις διεργασίες που ενεργοποιούνται και υποδηλώνονται από κάθε εξαγόμενη ομάδα και τη συσχέτιση αυτών με την πραγματική παθολογία.

1.2 Σχετική βιβλιογραφία

Οι περισσότερες βιολογικές εφαρμογές που χρησιμοποιούν την τεχνική διπλής κατηγοριοποίησης, πραγματοποιούνται χρησιμοποιώντας την τεχνολογία των μικροσυστοιχιών που επιτρέπει την μέτρηση της γονιδιακής έκφρασης χιλιάδων γονιδίων κάτω από συγκεκριμένες πειραματικές συνθήκες. Οι περισσότεροι μελετητές που έχουν χρησιμοποιήσει τεχνικές διπλής κατηγοριοποίησης αξιοποιούν πίνακες προερχόμενους από καρκινικά κύτταρα από διάφορα στάδια της ασθένειας, δείγματα από διαφορετικούς ασθενείς με συγκεκριμένο καρκινικό τύπο, από υγιείς ανθρώπους ή ακόμα και από πρότυπους οργανισμούς ζύμης (ζυμομύκητες-yeast).

Αυτά τα σύνολα δεδομένων έχουν χρησιμοποιηθεί για τη μελέτη της λειτουργίας των διαφόρων μεθόδων διπλής κατηγοριοποίησης στοχεύοντας σε τρείς βασικές εφαρμογές:

1. Αναγνώριση γονιδίων με ανάλογη συμπεριφορά (συνρυθμιζόμενα γονίδια).

2. Λειτουργικό σχολιασμό γονιδίων.

3. Κατηγοριοποίηση δειγμάτων και γονιδίων.

Ο πίνακας που ακολουθεί παραθέτει κάποιες ενδεικτικές εφαρμογές διπλής κατηγοριοποίησης σε δεδομένα μικροσυστοιχιών τα τελευταία 14 χρόνια:

Πίνακας 1. Εφαρμογές διπλής κατηγοριοποίησης τα τελευταία 14 χρόνια.

Σύνολο δεδομένων	Εφαρμογές	Αναφορές
Πολλαπλή σκλήρυνση	3	C.Tang et al 2001 [1]
Καρκίνος του μαστού	1	A.Ben-Dor et al 2002 [1]
Yeast-stress	1/2	E.Segal et al 2003 [1]
Καρκίνος του παχέως εντέρου	1/2	T.M.Murali et al 2003[1]
Λευχαιμία	3	Q.Sheng et al 2003[1]
Yeast/Λέμφωμα	3	C.Cano et al 2007[2]
Οξεία λεμφοβλαστική λευχαιμία/λέμφωμα/yeast	1	U. Maulik & S. Bandyopadhyay 2009[3]
5 τύποι καρκίνου του πνεύμονα	1/3	C.-P. Chen et al 2014[4]

1.3 Αλγοριθμικό υπόβαθρο

Οι πίνακες γονιδιακής έκφρασης, όπως αναφέρθηκε, αναλύονται σε δύο διαστάσεις: τη διάσταση των γονιδίων και τη διάσταση των καταστάσεων. Η πρώτη διάσταση αναλύεται συγκρίνοντας τις γραμμές και η δεύτερη τις στήλες του πίνακα. Τα πιο σημαντικά σημεία μελέτης αναλύοντας τέτοιους πίνακες περιλαμβάνουν [1] :

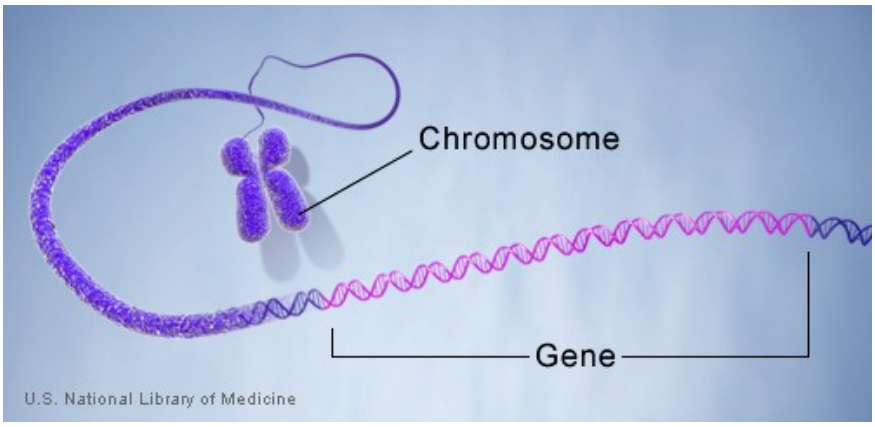
1. Ομαδοποίηση γονιδίων με βάση την έκφραση τους κάτω από διαφορετικές καταστάσεις
2. Πρόβλεψη νέων γονιδίων, με βάση την έκφραση άλλων γονιδίων με ήδη γνωστή πρόβλεψη.
3. Ομαδοποίηση καταστάσεων με βάση την έκφραση ενός αριθμού γονιδίων.
4. Πρόβλεψη ενός νέου συνολικού δείγματος, έχοντας ως γνωστή πληροφορία την έκφραση γονιδίων κάτω από συγκεκριμένες πειραματικές καταστάσεις.

Οι τεχνικές απλής ομαδοποίησης χρησιμοποιούνται με στόχο είτε να ομαδοποιήσουν γονίδια ή καταστάσεις δηλαδή είναι πιο άμεσες για τα σημεία 1 και 3. Σε αυτές όμως τις τεχνικές συναντάται η εξής δυσκολία. Ομαδοποιούν με ένα τρόπο καθολικό με αποτέλεσμα κάποιες ομάδες που παρουσιάζουν βιολογικό ενδιαφέρον σε μία μικρή ομάδα καταστάσεων να μην μπορούν να αξιοποιηθούν. Τέτοια τοπικά μοντέλα ομαδοποίησης μπορεί να αποτελούν κλειδιά για γενετικά μονοπάτια και αυτά στοχεύουν να ανακαλύψουν οι διάφορες τεχνικές διπλής κατηγοριοποίησης που πραγματοποιούν ταυτόχρονη ομαδοποίηση γραμμών και στηλών σε ένα πίνακα γονιδιακής έκφρασης.

1.4 Βιολογικό υπόβαθρο

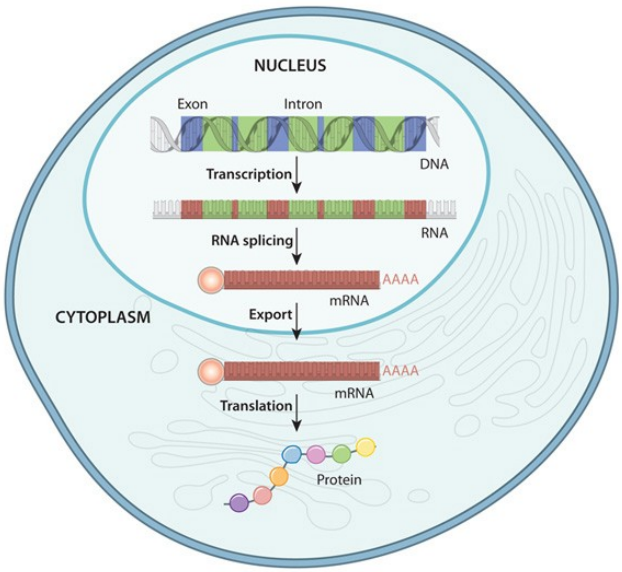
1.4.1 Γονίδιο και γονιδιακή έκφραση

Το γονίδιο είναι η βασική φυσική και λειτουργική μονάδα της κληρονομικότητας. Τα γονίδια αποτελούνται από DNA και λειτουργούν σαν οδηγοί για την παραγωγή των πρωτεϊνών. Στους ανθρώπους τα γονίδια ποικίλουν σε μέγεθος ξεκινώντας από μερικές εκατοντάδες βάσεις DNA και φτάνοντας περισσότερες από δύο εκατομμύρια. Έχει υπολογιστεί ότι οι άνθρωποι έχουν 20.000-25.000 γονίδια συνολικά. Κάθε άτομο έχει δύο αντίγραφα από κάθε γονίδιο, ένα από κάθε γονέα. Τα περισσότερα γονίδια είναι ίδια σε όλους τους ανθρώπους με εξαίρεση ένα μικρό αριθμό (λιγότερο από 1% του συνόλου) που είναι ελαφρώς διαφορετικά [1] με μικρές αλλαγές στην αλληλουχία των βάσεων του DNA. Αυτές οι διαφοροποιήσεις συμβάλλουν στα μοναδικά φυσικά χαρακτηριστικά κάθε ατόμου [5].



Εικόνα 1. Τα γονίδια φτιάχνονται από DNA. Κάθε χρωμόσωμα περιέχει πολλά γονίδια [5] .

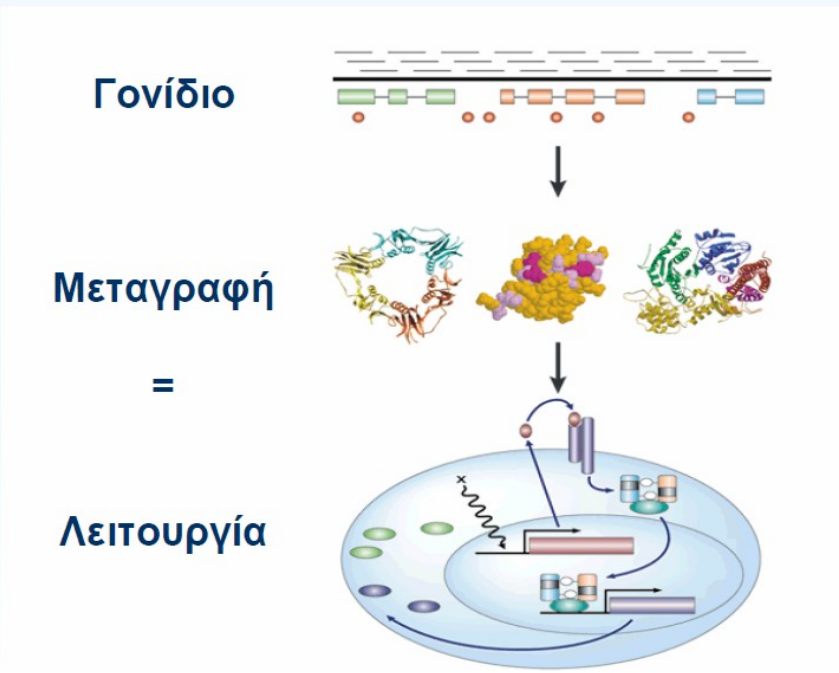
Όπως αναφέραμε τα γονίδια κωδικοποιούν τις πρωτεΐνες και οι πρωτεΐνες υπαγορεύουν την λειτουργία των κυττάρων. Ως εκ τούτου τα γονίδια που εκφράζονται σε ένα συγκεκριμένο κύτταρο, προσδιορίζουν τι είναι αυτό το κύτταρο και τι μπορεί να κάνει. Επιπλέον κάθε βήμα στη ροή πληροφοριών από το DNA στο RNA και στην πρωτεΐνη παρέχει στο κύτταρο ένα σημείο ελέγχου για την ρύθμιση της ποσότητας και του τύπου των πρωτεϊνών που κατασκευάζει. Σε οποιαδήποτε δεδομένη στιγμή, η ποσότητα μιας πρωτεΐνης σε ένα κύτταρο αντικατοπτρίζει την ισορροπία μεταξύ συνθετικών και βιοχημικών οδών της πρωτεΐνης. Από την πλευρά αυτής της ισορροπίας, υπενθυμίζεται ότι η πρωτεΐνη αρχίζει από την μεταγραφή (DNA σε RNA) και συνεχίζει με τη μετάφραση (RNA σε πρωτεΐνη) [6].



Εικόνα 2. Η διαδικασία μεταγραφής από mRNA σε πρωτεΐνη [6].

1.4.2 Προσδιορισμός 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 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) και η ανταπόκριση των κυττάρων ή ιστών σε φυσιολογικά (όπως φάρμακα και αναπτυξιακοί παράγοντες) ή παθολογικά «ερεθίσματα» (π.χ. μεταλλαγή) εκδηλώνεται με την αλλαγή αυτής [7].

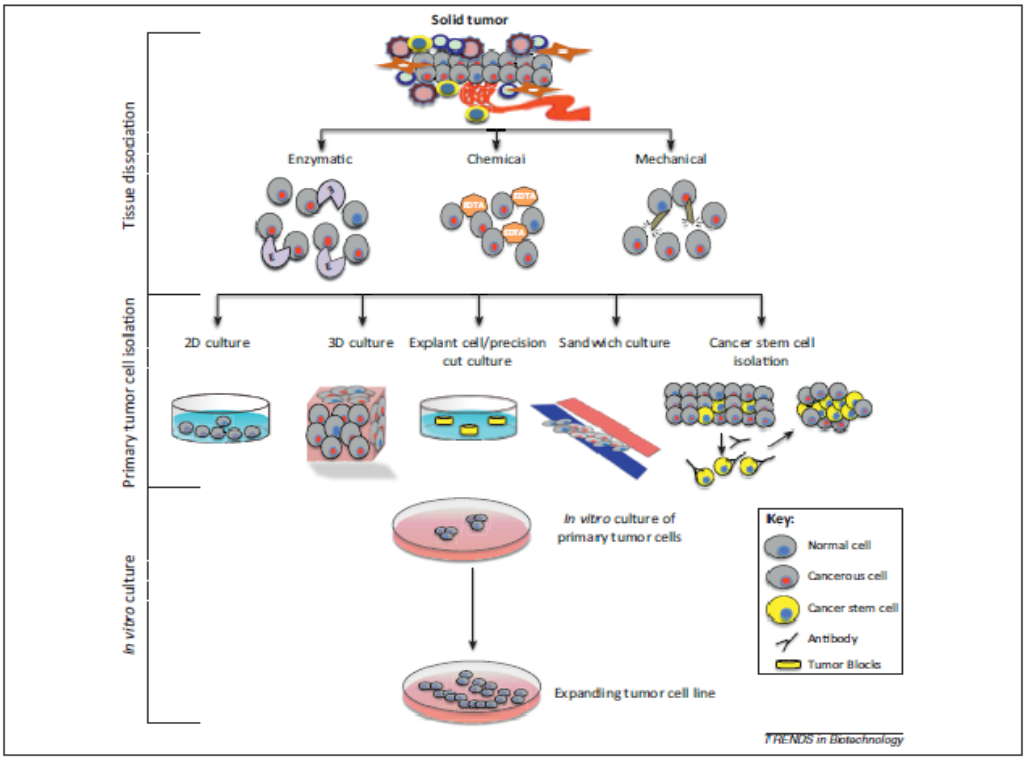


Εικόνα 3. Προσδιορισμός μεταγραφικών προτύπων [7].

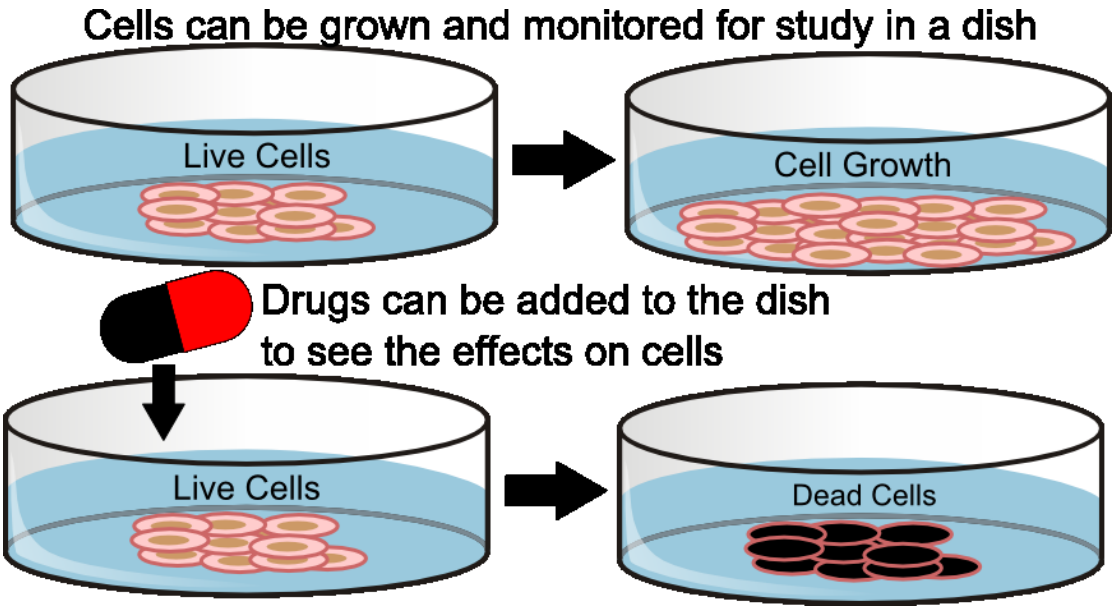
1.4.3 Κυτταρικές σειρές

Παρόλη την πρόοδο της γονιδιωματικής τα τελευταία χρόνια, εξακολουθούσε να υπάρχει η ανάγκη για αξιόπιστα προκλινικά μοντέλα για τη δοκιμή θεραπευτικών στρατηγικών στους διάφορους τύπους καρκίνου. Σταθμό σε αυτή την έρευνα αποτέλεσε η δημιουργία των κυτταρικών σειρών. Οι κυτταρικές σειρές αποτελούν τα σημερινά θεμελιώδη μοντέλα που χρησιμοποιούνται για τη μελέτη της βιολογίας του καρκίνου και της θεραπευτικής αποτελεσματικότητας των αντικαρκινικών μεθόδων [8].

Οι κυτταρικές σειρές αποτελούν ζωντανά καρκινικά κύτταρα στα λεγόμενα πιάτα καλλιέργειας (Τρυβλίο-Πέτρι). Γι' αυτό το λόγο τέτοιες μελέτες ονομάζονται *in vitro* (η λατινική έκφραση για το "σε δοκιμαστικό σωλήνα"). Οι σειρές συνήθως προέρχονται από κάποιον ασθενή με καρκίνο και κάθε σειρά που χρησιμοποιείται διαφέρει από τις άλλες είτε γιατί προέρχεται από άλλο μέρος του σώματος είτε γιατί έχει διαφορετική αντίδραση στην εκάστοτε θεραπεία.



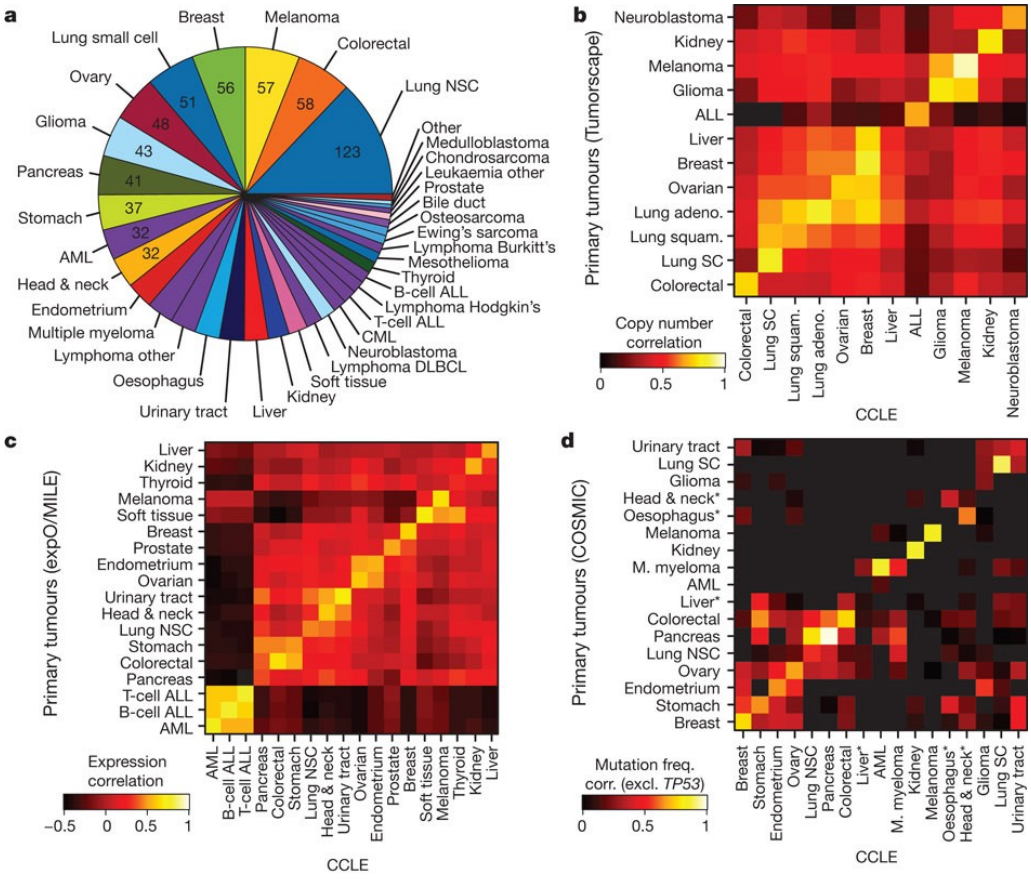
Εικόνα 4. Σχηματική αναπαράσταση της διαδικασίας παραγωγής μιας κυτταρικής σειράς [54].



Εικόνα 5. Κυτταρικές σειρές – ***in vitro*** μελέτη. Οι κυτταρικές σειρές μπορούν να μεγαλώσουν και να βρίσκονται υπό παρακολούθηση σε πιάτα καλλιέργειας (Τρυβλίο-Πέτρι) . Φάρμακα προστίθενται και μελετούνται οι αντιδράσεις των σειρών [8].

Η HeLa ήταν η πρώτη κυτταρική σειρά που δημιουργήθηκε από καρκινικά κύτταρα του τραχήλου της μήτρας προερχόμενα από την αφρικανοαμερικανή Henrietta Lacks το 1951 [7].

Εικόνα 6. Η ασθενής **Henrietta Lacks** από την οποία δημιουργήθηκε η πρώτη καρκινική σειρά [55] .



Εικόνα 7. Η Cancer Cell Line Εγκυκλοπαίδεια **(CCLE)**. **a.** Κατανομή των καρκινικών τύπων στην CCLE. **b.** Σύγκριση των DNA προφίλ (GISTIC G-scores) μεταξύ κυτταρικών σειρών και αρχικών όγκων. **c.** Σύγκριση των προφίλ έκφρασης mRNA μεταξύ κυτταρικών σειρών και αρχικών όγκων**. d.** Σύγκριση της συχνότητας μεταλλαγής μεταξύ κυτταρικών σειρών και αρχικών όγκων στην COSMIC [11].

Από τότε εκατοντάδες καρκινικές σειρές έχουν καθιερωθεί και πλέον υπάρχει συγκεντρωμένη διαθέσιμη πληροφορία για τις κυτταρικές σειρές και για δείγματα καρκινικών όγκων μαζί [8]. Αναφέρεται ότι, σταδιακά σύμφωνα με τους ειδικούς, αντινεοπλασματικά φάρμακα θα προσαρμόζονται στο γενετικό προφίλ του όγκου του ασθενούς ξεχωριστά [9]. Στον Cancer Genome Atlas (TCGA) τα γονιδιώματα και οι γονιδιακές εκφράσεις για τουλάχιστον 500 δείγματα ιστών για κάθε τύπο όγκου έχουν χαρακτηριστεί. Επιπλέον η Broad – Novartis Cancer cell Line Εγκυκλοπαίδεια (CCLE) περιέχει γενετικά προφίλ για έναν μεγάλο αριθμό κυτταρικών σειρών που χρησιμοποιούνται ως μοντέλα για διάφορους καρκινικούς τύπους [10]. Συγκεκριμένα το σύνολο των δεδομένων της εγκυκλοπαίδειας αυτής αποτελείται από 947 ανθρώπινες κυτταρικές σειρές, μαζί με τα χαρακτηριστικά των 500 από αυτών έπειτα από επίδραση 24 φαρμακολογικών ενώσεων, καθώς και 36 τύπους καρκινικών όγκων [11].

Κυτταρικές σειρές και ανθρώπινος καρκίνος

Υπάρχουν πολλοί λόγοι για τους οποίους οι σειρές αποτελούν τα βασικά μοντέλα έρευνας για τους διάφορους τύπους καρκίνου. Υπάρχουν σειρές που μοιράζονται πολλά γονιδιακά χαρακτηριστικά με τον ανθρώπινο καρκίνο. Αν και μεμονωμένες κυτταρικές σειρές δεν αποτελούν προϊόν μελέτης, ομάδες σειρών μπορούν να αποτελέσουν πολύ σημαντικά εργαλεία για την έρευνα κατά του καρκίνου. Οι σειρές έχουν την ικανότητα του εύκολου πολλαπλασιασμού, είναι σχετικά επιρρεπείς στον γενετικό χειρισμό και υπό καθορισμένες πειραματικές συνθήκες αποδίδουν σημαντικά συγκρίσιμα αποτελέσματα [12].

➢ Κυτταρικές σειρές και καρκίνος του μαστού

Ο καρκίνος του μαστού δεν είναι μια απλή ασθένεια. Αποτελεί μια συλλογή ασθενειών του στήθους με διαφορετικές ιστοπαθολογίες και γονιδιακές παραλλαγές. Μια από τις σημαντικότερες προκλήσεις για τον καρκίνο του μαστού είναι η κατανόηση και η μελέτη των μεταστατικών μηχανισμών της νόσου οι οποίοι αποτελούν και την κύρια αιτία θνησιμότητας [12]. Αναφορικά, παγκοσμίως μεταξύ 1980-2010 ο αριθμός κρουσμάτων της ασθένειας αυξήθηκε από 641000 σε 1643000 περιστατικά με 3,1% ετήσια άνοδο και ο αριθμός θνησιμότητας το 2010 έφτασε τις 425000 γυναίκες με 68000 μεταξύ 15-49 ετών στις αναπτυσσόμενες χώρες [13].

Οι κυτταρικές σειρές του καρκίνου του μαστού αποτελούν τα πιο ευρέως διαδεδομένα μοντέλα για τη μελέτη του πολλαπλασιασμού, της απόπτωσης και της μετάστασης του συγκεκριμένου καρκινικού τύπου. Η πιο αντιπροσωπευτική έρευνα παρουσιάστηκε από τον Gran και τους συνεργάτες του η οποία αναφέρει ότι ένα πάνελ 51 κυτταρικών σειρών μαστού εμφανίζουν πολλές από τις επαναλαμβανόμενες γονιδιακές ανωμαλίες που εντοπίζονται σε όγκους μαστού. Έτσι γίνεται κατανοητό ότι η αναπαράσταση των όγκων αυτών ως κυτταρικές σειρές καλλιεργημένες στο εργαστήριο δεν μεταβάλλουν σημαντικά τα κοινά γονιδιακά χαρακτηριστικά [12].

➢ Κυτταρικές σειρές και καρκίνος του τραχήλου της μήτρας

Ο καρκίνος του τραχήλου της μήτρας αποτελεί μια από τις κύριες αιτίες θανάτου των γυναικών και ένα από τα σημαντικότερα αναπαραγωγικά προβλήματα υγείας της σύγχρονης γυναίκας [14]. Βασική αιτία εμφάνισης του καρκίνου του τραχήλου αποτελεί ένα σεξουαλικώς μεταδιδόμενο νόσημα που οφείλεται στον ιό των ανθρώπινων θηλωμάτων ή παπιλοϊό (HPV) [14]. Στατιστικά, η εμφάνιση του καρκίνου του τραχήλου της μήτρας αυξήθηκε παγκοσμίως από 378.000 περιστατικά ετησίως το 1980, σε 454.000 το 2010 με 0,6% αύξηση ανά χρονιά.

Παρόλο που ο αριθμός των θανάτων εξαιτίας του συγκεκριμένου καρκινικού τύπου έχει μειωθεί, η ασθένεια σκότωσε 200.000 γυναίκες το 2010, εκ των οποίων 46.000 σε ηλικία 15-49 ετών στις αναπτυσσόμενες χώρες [13].

Η πρόσφατη *in vitro* έρευνα για τον καρκίνο του τραχήλου περιλαμβάνει την καλλιέργεια αθανατοποιημένων κυτταρικών σειρών σε κατάλληλο υπόστρωμα. Οι σειρές που χρησιμοποιούνται συνήθως περιλαμβάνουν μεταξύ άλλων τις HeLa, SiHa και Caski [15].

➢ Κυτταρικές σειρές και καρκίνος του ενδομητρίου

Ο καρκίνος του ενδομητρίου είναι η πιο συχνά διαγνωσμένη κακοήθεια του γυναικείου γεννητικού συστήματος. Δέκα έως 25 γυναίκες στις 100.000 εμφανίζουν το συγκεκριμένο τύπο καρκίνου κυρίως στις Ευρωπαϊκές χώρες (Ισπανία, Ηνωμένο Βασίλειο, Γαλλία) και στις χώρες της Βόρειας Αμερικής (χώρες των ΗΠΑ και του Καναδά), με μία μεγαλύτερη συχνότητα σ' αυτές της Αμερικής. Παρά τη συχνή εμφάνισή του, το ποσοστό θνησιμότητας από τον καρκίνο του ενδομητρίου είναι σχετικά χαμηλό και η πρόγνωση είναι ιδιαίτερα σημαντική για πλήρη ίαση, εφόσον ανιχνευθεί στα πρώτα στάδια. Ένας από τους πιο χρησιμοποιημένους ανθρώπινους όγκους στην έρευνα *in vivo* του ενδομητρίου είναι ο EnCA101/ECC-1 όγκος. Όσον αφορά στις σειρές, από την πρώτη τους περιγραφή (Nishida 1985), τα κύτταρα Ishikawa - μια ανθρώπινη κυτταρική σειρά του αδενοκαρκινώματος του ενδομητρίου που εκφράζει το οιστρογόνο και τους υποδοχείς της προγεστερόνης - αποτέλεσαν το πιο εκτεταμένο ανθρώπινο μοντέλο καλλιέργειας κυττάρων του ενδομητρίου. Τα κύτταρα Ishikawa αντιπροσωπεύουν ένα συνδυασμό *in vitro* και *in vivo* μοντέλου του ανθρώπινου όγκου του ενδομητρίου που είναι κατάλληλος για την μελέτη της ανάπτυξης του ορμονικού ελέγχου [16].

➢ Κυτταρικές σειρές και καρκίνος των ωοθηκών

Ο καρκίνος των ωοθηκών είναι η πέμπτη πιο συχνή αιτία θανάτου στις γυναίκες. Κάθε χρόνο πεθαίνουν περισσότερες από 100.000 γυναίκες στον κόσμο από αυτό τον τύπο καρκίνου. Από το 75% των γυναικών που θα διαγνωστούν με τοπικά προχωρημένη ή γενικευμένη νόσο μόνο το 30% θα επιβιώσουν 5 χρόνια μετά τη θεραπεία. Τα ποσοστά επιβίωσης έχουν αλλάξει ελάχιστα από τις αρχές του 1980 παρά τη χρήση νέων χημικοθεραπευτικών φαρμάκων. Αυτή η συνολικά κακή πρόγνωση είναι το αποτέλεσμα ενός συνδυασμού παραγόντων συμπεριλαμβανομένων της έλλειψης συμπτωμάτων σε πρώιμο στάδιο και της αντοχής σε φάρμακα σε προχωρημένο στάδιο της νόσου [17]. Ο καρκίνος των ωοθηκών χωρίζεται σε τέσσερις βασικούς υπότυπους: ορώδης, βλεννώδης, ενδομητριώδης, και διαυγοκυτταρικός. Το ορώδες καρκίνωμα είναι υπεύθυνο για το 70% των περιπτώσεων [10].

Η τεχνολογία των μικροσυστοιχιών ήδη παρέχει πολύτιμα δεδομένα έκφρασης για την ταξινόμηση του καρκίνου των ωοθηκών και τις πρώτες ενδείξεις για το ποιές μοριακές αλλαγές θα μπορούσαν να αξιοποιηθούν σε νέες θεραπευτικές στρατηγικές [18].

Εικόνα 8. Τα μοτίβα έκφρασης γονιδίων μπορούν να χρησιμοποιηθούν για την ταξινόμηση των καρκινωμάτων των ωοθηκών σε κλινικά σχετικούς υπότυπους [18].

Οι περισσότερες σειρές του καρκίνου των ωοθηκών χρησιμοποιούνται για να ταυτοποιήσουν γονίδια που παρουσιάζουν ανοσία σε φάρμακα. Οι πιο γνωστές σειρές που λειτουργούν ως μοντέλα για αυτόν τον καρκίνο και συγκεκριμένα για τον υπότυπο HGSOC είναι οι SK-OV-3, A2780, OVCAR-3, CAOV3, και IGROV1 [10].

1.4.4 Μικροσυστοιχίες DNA

Μία από τις πιο διαδεδομένες μεθόδους για την ανάλυση της γονιδιακής έκφρασης αποτελούν οι μικροσυστοιχίες γονιδίων (DNA microarray) [19]. Η τεχνολογία των μικροσυστοιχιών DNA παρέχει τη δυνατότητα συγκριτικής μελέτης μεγάλου αριθμού δειγμάτων παρέχοντας μια πιο σφαιρική θεώρηση των βιολογικών συστημάτων. Τα προς ανάλυση δείγματα μπορούν να ομαδοποιηθούν με βάση τις ομοιότητες που παρουσιάζουν στο πρότυπο της γονιδιακής έκφρασης. Ως αποτέλεσμα της ομαδοποίησης, δημιουργείται μια βάση δεδομένων με αναλυτικές πληροφορίες για τη γονιδιακή έκφραση και λεπτομερείς χάρτες γενετικών ρυθμιστικών δικτύων για επιπλέον γνώση και πληροφορία.

Στην ουσία, το μέγεθος των δυνατοτήτων των μικροσυστοιχιών στην ανάλυση της κυτταρικής λειτουργίας γίνεται πιο σαφές όταν επικεντρωνόμαστε στο πρότυπο της γονιδιακής έκφρασης και όχι σε ένα μόνο γονίδιο. Αυτό πραγματοποιείται μόνο με τη χρήση εξειδικευμένου λογισμικού, που δίνει τη δυνατότητα ομαδοποίησης γονιδίων με παρόμοιο πρότυπο έκφρασης και σχεδιασμού φυλογενετικών δέντρων που θα περιλαμβάνουν τις συγγενείς γονιδιακές ομάδες. Οι μελέτες αυτές βασίζονται στο σενάριο ότι γονίδια με παρόμοιο πρότυπο έκφρασης είναι δυνατόν να ελέγχονται από τους ίδιους ρυθμιστικούς μηχανισμούς.

Εικόνα 9. Μικροσυστοιχίες DNA [56].

Η μέθοδος των μικροσυστοιχιών DNA καλύπτει ευρύ φάσμα εφαρμογών στην ιατρική έρευνα, και κυρίως σε μελέτες που αφορούν ασθένειες όπως ο καρκίνος, και τους αντίστοιχους διαγνωστικούς κα θεραπευτικούς στόχους. Συγκεκριμένα, όσον αφορά την ασθένεια του καρκίνου, οι μικροσυστοιχίες που χρησιμοποιούνται περιλαμβάνουν γονίδια που σχετίζονται με συγκεκριμένο ιστό ή με κάποια φυσιολογική λειτουργία (π.χ. απόπτωση, αγγειογένεση).

Επιπλέον, επιτρέπουν την ταυτοποίηση μοριακών καρκινικών δεικτών μέσω της σύγκρισης πληροφοριών από την ανάλυση της γονιδιακής έκφρασης σε μεγάλο αριθμό δειγμάτων κακοήθων όγκων ή σε καρκινικές κυτταρικές σειρές. Μελλοντικά, θα μπορούσαν να προσφέρουν σημαντικές γνώσεις στη διάκριση της πρωτογενούς εστίας από τις μεταστάσεις. Οι ανθρώπινοι ιστοί αποτελούνται από τρισδιάστατες και πολύπλοκες δομές και από πολλούς κυτταρικούς πληθυσμούς που αλληλεπιδρούν μεταξύ τους. Παρόλ' αυτά, για τη μελέτη των πολύπλοκων αυτών δομών, ο ερευνητής χρειάζεται ένα ελάχιστο κλάσμα του συνολικού όγκου του ιστού. Για αυτό το σκοπό καλλιεργούνται οι προαναφερόμενες κυτταρικές σειρές με μοναδική επιφύλαξη ότι οι συνθήκες *in vitro* δεν αντιστοιχούν πάντα ακριβώς στις συνθήκες *in vivo*.

Μια άλλη εξίσου σημαντική χρήση των μικροσυστοιχιών αποτελεί η διαλεύκανση του μηχανισμού δράσης των αντικαρκινικών φαρμάκων. Η μελέτη και ανάλυση της γονιδιακής έκφρασης αποκαλύπτει τις μεταβολές που μπορεί να επιφέρει κάποιο χρησιμοποιούμενο φάρμακο και δίνει πληροφορίες για τους μοριακούς στόχους της δράσης του. Αυτή η ανάλυση θα αποβεί εξίσου σημαντική και χρήσιμη σε μοριακούς βιολόγους και σε φαρμακευτικές εταιρείες για ανάπτυξη καλύτερων φαρμάκων.

Είναι γνωστό από την κλινική πρακτική, ότι ασθενείς με μορφές καρκίνου με παρόμοια μορφολογικά χαρακτηριστικά συχνά νοσούν με διαφορετική έκβαση. Επομένως αποτελεί άμεση ανάγκη η διερεύνηση προγνωστικών παραγόντων για την σωστή κατηγοριοποίηση των ασθενών ώστε να είναι πιο εύκολη η λήψη αποφάσεων σχετικά με τη θεραπεία που θα πρέπει να χορηγηθεί.

Στόχος τα επόμενα χρόνια θα αποτελέσει η χρήση της πληροφορίας της γονιδιακής έκφρασης για τη δημιουργία του "μοριακού αποτυπώματος" κάθε όγκου, γεγονός που θα οδηγήσει σε καλύτερο καθορισμό της πρόγνωσης και διάγνωσης της νόσου και σε εξατομίκευση της θεραπείας για κάθε ασθενή [20].

1.5 Δομή της εργασίας

Η οργάνωση των κεφαλαίων που ακολουθούν, βασισμένη στο τρόπο ανάπτυξης της παρούσας εργασίας, έχει ως εξής:

Το πρώτο κεφάλαιο περιλαμβάνει το σύνολο το βιολογικού υπόβαθρου που μελετήθηκε. Σ' αυτό αναλύονται βασικές έννοιες όπως το γονίδιο και η γονιδιακή έκφραση, τα μεταγραφικά πρότυπα και οι κυτταρικές σειρές. Ακόμη γίνεται μια μεγάλη αναφορά στις μικροσυστοιχίες DNA και στους τέσσερις τύπους καρκίνου (μαστού, τράχηλου της μήτρας, ωοθηκών και ενδομητρίου).

Στο δεύτερο κεφάλαιο περιγράφεται η τεχνική διπλής κατηγοριοποίησης, η δομή και τα είδη των αλγορίθμων της καθώς και μια θεωρητική σύγκριση με την τεχνική απλής ομαδοποίησης. Επιπλέον, περιγράφεται ο αλγόριθμος Cheng and Church που εφαρμόστηκε αναπτύσσοντας τα επιμέρους βήματά του και το θεωρητικό υπόβαθρο στο οποίο στηρίζεται.

Ακολουθεί το τρίτο κεφάλαιο το οποίο περιλαμβάνει την προτεινόμενη μεθοδολογία μας με την εφαρμογή του αλγορίθμου Cheng and Church στον οποίο έχουμε επέμβει προτείνοντας βελτιώσεις που στοχεύουν σε πιο λεπτομερή αποτελέσματα.

Τα αποτελέσματα παρατίθενται στο τέταρτο κεφάλαιο με συγκεντρωτικούς πίνακες που αποδεικνύουν την στατιστική σημαντικότητά τους, καθώς και μία πρακτική σύγκρισή τους με τα αποτελέσματα μιας εφαρμογής clustering.

Το πέμπτο κεφάλαιο αποτελείται από την βιολογική αξιολόγηση των αποτελεσμάτων και την επιλογή των τελικών γονιδιακών δεικτών για τους τέσσερις τύπους καρκίνου που επεξεργαστήκαμε.

Στο έκτο κεφάλαιο περιγράφονται κάποια τελικά συμπεράσματα και μελλοντικές επεκτάσεις της δουλειάς μας. Η εργασία ολοκληρώνεται με την παράθεση δύο παραρτημάτων που περιλαμβάνουν: α. τις γραφικές με την αναλυτική συμπεριφορά των γονιδίων για κάθε ομάδα (bicluster), και β. τα αναλυτικά βιολογικά αποτελέσματα για κάθε ομάδα (κωδικοί γονιδίων, διαγράμματα με βιολογικές διεργασίες και μονοπάτια, πίνακες με 20 ομάδες διπλής κατηγοριοποίησης για κάθε κυτταρικό τύπο).

ΚΕΦΑΛΑΙΟ 2 ΤΕΧΝΙΚΗ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (Biclustering)

**Για συντομία χρησιμοποιούμε τον ευρέως διαδεδομένο όρο biclustering για την τεχνική της διπλής κατηγοριοποίησης και τον όρο biclusters για τις ομάδες διπλής κατηγοριοποίησης.*

Τα προφίλ γονιδιακής έκφρασης αποτελούν τα τελευταία χρόνια την βασική τεχνική για την λήψη του μοριακού αποτυπώματος των ιστών σε διάφορες βιολογικές συνθήκες. Με τη διαθεσιμότητα ολόκληρων γονιδιακών αλληλουχιών, οι μικροσυστοιχίες DNA επιτρέπουν την μέτρηση των επιπέδων mRNA σε χιλιάδες γονίδια ταυτόχρονα. Η μέτρηση των επιπέδων γονιδιακής έκφρασης κάτω από μία συγκεκριμένη κατάσταση αποτελεί το προφίλ αυτού του γονιδίου στην κατάσταση αυτή. Δεδομένου ενός συνόλου προφίλ γονιδιακής έκφρασης που οργανώνεται σε έναν πίνακα με γραμμές που αντιστοιχούν στα γονίδια και στήλες στις καταστάσεις (κυτταρικές σειρές, χρονικές στιγμές κ.ά.), στόχος αποτελεί η ομαδοποίηση γονιδίων και καταστάσεων σε υποσύνολα που μεταφέρουν σημαντική βιολογική πληροφορία. Τέτοιες τεχνικές ομαδοποίησης αποτελούν το διαδεδομένο clustering και μια πιο αποτελεσματική και σύγχρονη μέθοδος το biclustering [21].

2.1 Η τεχνική Ομαδοποίησης (Clustering)

**Για συντομία χρησιμοποιούμε τον ευρέως διαδεδομένο όρο clustering για την τεχνική της ομαδοποίησης .*

Σε έναν πίνακα γονιδιακής έκφρασης, αν δύο γονίδια σχετίζονται (έχουν παρόμοιες λειτουργίες), τα γονιδιακά τους προφίλ πρέπει να μοιάζουν (π.χ. μικρή ευκλείδεια απόσταση ή υψηλή συσχέτιση). Η κλασσική ομαδοποίηση ή αλλιώς clustering ομαδοποιεί τα στοιχεία ενός πίνακα έτσι ώστε μέλη του ίδιου γκρουπ να είναι παρόμοια και τα γκρουπ μεταξύ τους να διαφοροποιούνται με σαφήνεια. Δύο είναι οι βασικοί τρόποι ομαδοποίησης στη μέθοδο clustering:

- Με βάση την κατάσταση: Πολλαπλά πειράματα μικροσυστοιχιών μπορούν να ομαδοποιηθούν μαζί, με κριτήριο την ομοιότητα της γονιδιακής έκφρασης μεταξύ των πειραμάτων.
- Με βάση το γονίδιο: Τα γονίδια ομαδοποιούνται με κριτήριο την γονιδιακή τους έκφραση σε μία σειρά καταστάσεων-πειραμάτων.

Οι πιο γνωστοί και συχνοί τύποι clustering αλγορίθμων είναι οι:

- K-means/K-median
  Αποτελεί μια σχετικά απλή προσέγγιση και δίνει αρκετά σαφή αποτελέσματα αλλά κάθε αντικείμενο (γονίδιο ή κατάσταση) μπορεί να ανήκει σε ένα μόνο γκρουπ.

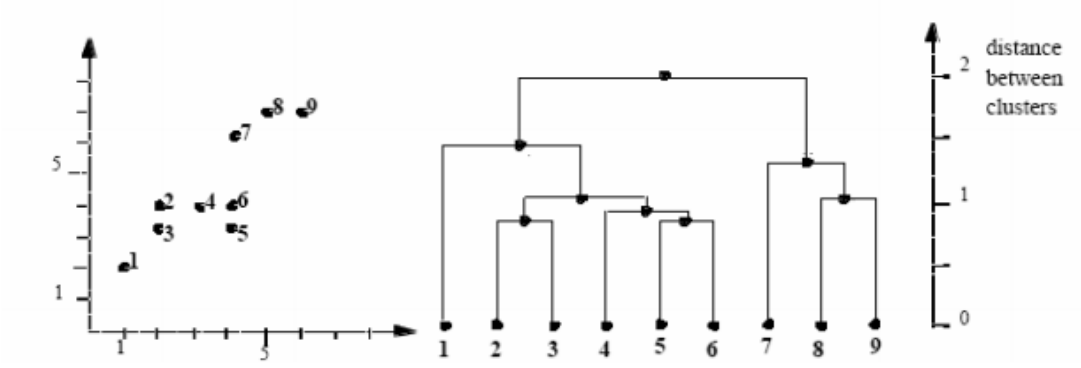
- Hierarchical clustering
  Αποτελεί επίσης μια εύκολη διαδικασία, περιλαμβάνει δύο τύπους, το agglomerative και το divisive clustering, και δεν υπάρχει πρότερη γνώση του αριθμού των ομάδων. Κάθε αντικείμενο μπορεί να ομαδοποιηθεί μόνο μια φορά και τα γονίδια ομαδοποιούνται όλα, παρόλο που κάποιες ομάδες μπορεί να είναι πιο αδύναμες και λιγότερο ακριβείς.

  Τα κύρια μειονεκτήματα αυτής της τεχνικής ομαδοποίησης είναι δύο. Αρχικά τα γονίδια ομαδοποιούνται με βάση την έκφρασή τους σε όλες τις καταστάσεις αν και μια χρήσιμη πληροφορία μπορεί να κρύβεται μόνο σε ένα επιμέρους κομμάτι των καταστάσεων. Δεύτερον, κάθε γονίδιο ομαδοποιείται μόνο σ' ένα cluster, παρόλ' αυτά ένα γονίδιο μπορεί να συμμετέχει σε πάνω από μία κυτταρικές διεργασίες [22].

2.1.2 Ιεραρχική ομαδοποίηση

(Hierarchical clustering)

Η ιδέα της μεθόδου Hierarchical Clustering, όπως αναφέρει και το όνομά της, είναι να χτίσει μία ιεραρχία από clusters δείχνοντας τις σχέσεις μεταξύ ανεξάρτητων μελών και συγχωνευμένων clusters των δεδομένων, βασιζόμενη στην ομοιότητα [23].



Εικόνα 10. Hierarchical Clustering [24].

- Η ρίζα αναπαριστά το συνολικό σύνολο δεδομένων (dataset).
- Ενα φύλλο αναπαριστά ένα συγκεκριμένο στοιχείο του dataset.
- Ένας ολόκληρος κόμβος αναπαριστά την ένωση των στοιχείων στο υποδέντρο.
- Το ύψος ενός κόμβου αναπαριστά την απόσταση δύο παιδιών κόμβων [24].

Agglomerative Clustering

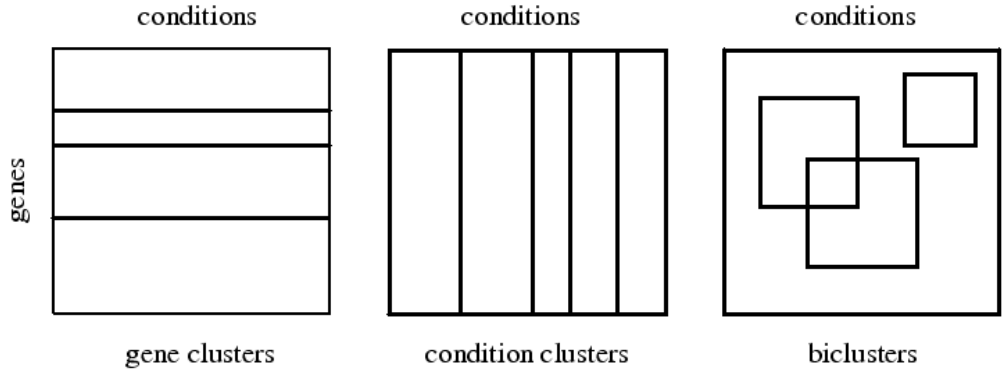
- Ξεκινά τοποθετώντας κάθε στοιχείο του πίνακα σε ένα ξεχωριστό cluster.
- Σε κάθε βήμα συγχωνεύει τα δύο πιο όμοια clusters.
- Σταματά όταν όλα τα στοιχεία ανήκουν σε ένα μόνο cluster ή όταν τερματιστεί κάποιο συγκεκριμένο κριτήριο.

Divisive Clustering

- Ξεκινά τοποθετώντας όλα τα στοιχεία του πίνακα σε ένα cluster.
- Σε κάθε βήμα χωρίζει ένα cluster σε δύο καινούργια.
- Σταματά όταν όλα τα στοιχεία ανήκουν στο δικό τους cluster ή όταν τερματιστεί κάποιο συγκεκριμένο κριτήριο [25].

2.2 Η τεχνική διπλής κατηγοριοποίησης (Biclustering)

Η έννοια του biclustering αποτελεί μια τεχνική πιο πολύπλοκη και ευέλικτη σε σχέση με την clustering τεχνική. Η τεχνική biclustering αποτελεί ένα NP-hard πρόβλημα που πρωτοσυζητήθηκε από τους Hartigan, Morgan και Peeters [26]. Από τη στιγμή που εμφανίστηκαν οι πρώτοι αλγόριθμοι biclustering το 2000 από τους Cheng και Church, το biclustering προκάλεσε έντονο ενδιαφέρον και έγινε το κέντρο της προσοχής για τους ερευνητές [27]. Η τεχνική biclustering παίρνει σαν είσοδο τα ίδια δεδομένα (πίνακας γονιδιακής έκφρασης) με την clustering μέθοδο και εξάγει ομάδες-υποπίνακες που καλούνται biclusters. **Ενώ οι clustering αλγόριθμοι ομαδοποιούν κάθε φορά με βάση τις γραμμές ή τις στήλες του πίνακα χωριστά, οι biclustering αλγόριθμοι πραγματοποιούν clustering σε δύο διαστάσεις ταυτόχρονα** [28]**.** Αυτό σημαίνει ότι σε αντίθεση με τους clustering αλγορίθμους μια bicustering μέθοδος αναγνωρίζει ομάδες γονιδίων που εμφανίζουν ανάλογη συμπεριφορά κάτω από μία συγκεκριμένη ομάδα καταστάσεων. Γι' αυτό άλλωστε, οι biclustering αλγόριθμοι είναι γνωστοί και ως co-clustering ή two way clustering [27].



Εικόνα 11. Οι δύο αριστερές εικόνες απεικονίζουν clustering ομαδοποίηση με βάση τις γραμμές και τις στήλες αντίστοιχα, ενώ η δεξιά αφορά την τεχνική biclustering (ταυτόχρονη ομαδοποίηση γραμμών και στηλών) [28].

2.3 Τύποι διπλής κατηγοριοποίησης

Ένα από τα κριτήρια επιλογής του κατάλληλου αλγορίθμου biclustering είναι ο τύπος των biclusters που δίνει ως αποτέλεσμα. Επικρατούν τέσσερις βασικοί τύποι:

- Biclusters με σταθερές τιμές
- Biclusters με σταθερές τιμές σε γραμμές ή στήλες
- Biclusters με συνεκτικές τιμές
- Biclusters με συνεκτικές εξελίξεις

Οι τρείς πρώτοι τύποι ασχολούνται με τις αριθμητικές τιμές ενός πίνακα και έχουν ως στόχο να εντοπίζουν υποομάδες γραμμών ή στηλών με παρόμοια συμπεριφορά όπως φαίνεται στις εικόνες 12, 13 και 14. Ο τέταρτος τύπος στοχεύει στην ομαδοποίηση ανεξάρτητα από τις ακριβείς αριθμητικές τιμές των στοιχείων του πίνακα αλλά τις περισσότερες φορές τις αντιμετωπίζει ως σύμβολα. Τα σύμβολα αυτά μπορεί να είναι κάποιοι συγκεκριμένοι χαρακτήρες, να ανταποκρίνονται σε μια δεδομένη σειρά ή να αναπαριστούν θετικές και αρνητικές αλλαγές που σχετίζονται με μία τιμή [1].

Ακολουθεί μια σύντομη σημειογραφία χρήσιμη για την αναλυτικότερη περιγραφή των 4 αυτών τύπων: Δεδομένου πίνακα $A=(X,Y)$ με X το σύνολο των γραμμών και Y το σύνολο των στηλών του, ένα bicluster αποτελεί έναν υποπίνακα (I,J) όπου I ένα υποσύνολο του X και J ένα υποσύνολο του Y αντίστοιχα και a_{ij} η τιμή κάθε στοιχείου του πίνακα. Ακόμη ορίζουμε τις μέσες τιμές των γραμμών και των στηλών και τη μέση τιμή όλου του πίνακα (I,J):

$$a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{ij} \quad , a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij}$$

και

$$a_{IJ} = \frac{1}{|I||J|}\sum_{i\in I, j\in J} a_{ij} = \frac{1}{|I|}\sum_{i\in I} a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{Ij}$$

2.3.1 Biclusters με σταθερές τιμές

Οι αλγόριθμοι που έχουν ως αποτέλεσμα biclusters με σταθερές τιμές τείνουν να αναδιατάσσουν τις γραμμές και τις στήλες ενός πίνακα ούτως ώστε να ομαδοποιήσουν γραμμές και στήλες με παρόμοιες τιμές. Ένα τέλειο σταθερό bicluster είναι ένας υποπίνακας *(I,J)* όπου όλες οι τιμές του είναι ίσες και για κάθε $i \in I$ και $j \in J$: $\alpha_{ij}=\mu$. Τέτοια παραδείγματα συναντάμε σε κάποιους πίνακες, τα περισσότερα δεδομένα προς επεξεργασία όμως περιέχουν θόρυβο, γεγονός που σημαίνει ότι ένα σταθερό bicluster παρουσιάζεται με τιμές $n_{ij}+\mu$, όπου το n_{ij} είναι ο σχετικός θόρυβος της τιμής μ του α_{ij} [1].

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

Εικόνα 12. Bicluster με σταθερές τιμές [25].

2.3.2 Biclusters με σταθερές τιμές σε γραμμές ή στήλες

Οι τιμές ενός βέλτιστου bicluster με σταθερές γραμμές σε έναν υποπίνακα *(I,J)* μπορούν να περιγραφούν ως εξής: $\alpha_{ij}=\mu+\alpha_i$ ή $\alpha_{ij}=\mu \times \alpha_i$ όπου μ είναι μια σταθερή τιμή εντός του bicluster και α_i αποτελεί μία προσαρμοσμένη τιμή για κάθε γραμμή $i \in I$ η οποία προστίθεται ή πολλαπλασιάζεται. Παρομοίως οι τιμές ενός βέλτιστου bicluster με σταθερές στήλες σε έναν υποπίνακα *(I,J)* μπορούν να περιγραφούν ως εξής: $\alpha_{ij}=\mu+\beta_j$ ή $\alpha_{ij}=\mu \times \beta_j$ όπου μ είναι μια σταθερή τιμή εντός του bicluster και β_j αποτελεί μία προσαρμοσμένη τιμή για κάθε στήλη $j \in J$ η οποία προστίθεται ή πολλαπλασιάζεται [1].

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

Εικόνα 13. Biclusters με σταθερές τιμές σε γραμμές και στήλες [25].

Στην παραπάνω εικόνα το αριστερό bicluster εμφανίζει σταθερές τιμές στις γραμμές του ακολουθώντας το μοντέλο $\alpha_{ij}=\mu+ \alpha_i$, όπου $\mu=0.0$ και κάθε φορά προστίθεται σε κάθε νέα γραμμή η τιμή $\alpha_i=1.0$. Αντίστοιχα το δεξί bicluster εμφανίζει σταθερές τιμές στις στήλες ακολουθώντας το μοντέλο $\alpha_{ij}=\mu+ \beta_j$, όπου $\mu=0.0$ και κάθε φορά προστίθεται σε κάθε νέα γραμμή η τιμή $\beta_i=1.0$.

2.3.3 Biclusters με συνεκτικές (coherent) τιμές

Η κατηγορία των biclusters με συνεκτικές τιμές δεν μπορεί να προσδιοριστεί εύκολα με τα απλά μοντέλα πρόσθεσης και πολλαπλασιασμού που αναφέρθηκαν στα σταθερά biclusters. Υπάρχουν πιο πολύπλοκες προσεγγίσεις που αναφέρονται στην ανάλυση της διακύμανσης μεταξύ υποομάδων του πίνακα. Αυτές οι μέθοδοι χρησιμοποιούν μια συγκεκριμένη μορφή συνδιακύμανσης μεταξύ γραμμών και στηλών σε ένα bicluster για να αξιολογήσουν την ποιότητα του αποτελέσματος. Χρησιμοποιώντας λοιπόν ένα πιο σύνθετο μοντέλο πρόσθεσης, ένα τέλειο συνεκτικό bicluster χαρακτηρίζεται από μία υποομάδα γραμμών και στηλών που οι τιμές της α_{ij} επιλέγονται με βάση την ακόλουθη έκφραση: $\alpha_{ij}=\mu+ \alpha_i + \beta_j$, όπου μ είναι μια σταθερή τιμή εντός του bicluster και α_i μία προσαρμοσμένη τιμή για κάθε γραμμή $i \epsilon I$, και β_j μία προσαρμοσμένη τιμή για κάθε στήλη $j \epsilon J$ οι οποίες προστίθενται στην μ. Μια άλλη προσέγγιση αναφέρει ότι τα biclusters με συνεκτικές τιμές μπορούν να περιγραφούν από ένα μοντέλο πολλαπλασιασμού, όπου $\alpha_{ij}= \mu' \times \alpha_i ' \times \beta_j '$ [1].

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

Εικόνα 14. Biclusters με συνεκτικές τιμές (αριστερά μοντέλο πρόσθεσης, δεξιά μοντέλο πολλαπλασιασμού) [25].

Για να γίνει πιο κατανοητή η μορφή των biclusters με συνεκτικές τιμές αναλύουμε το παράδειγμα της εικόνας 14 για το μοντέλο της πρόσθεσης:
Γνωρίζουμε ότι $\alpha_{ij}=\mu+ \alpha_i + \beta_j$
άρα για $\mu=0$

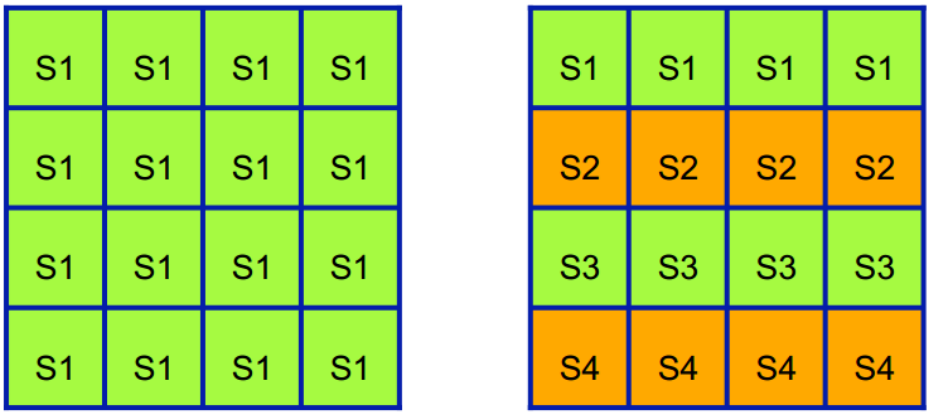
	$\beta_1=0$	$\beta_2=1$	$\beta_3=4$	$\beta_4=-1$
$\alpha_1=1$	1	2	5	0
$\alpha_2=2$	2	3	6	1
$\alpha_3=4$	4	5	8	3
$\alpha_4=5$	5	6	9	4

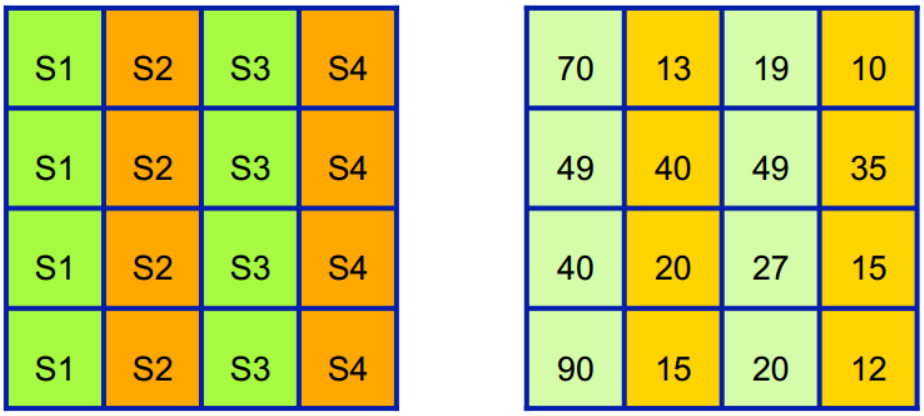
Και για το μοντέλο του πολλαπλασιασμού:
Γνωρίζουμε ότι $\alpha_{ij}= \mu' \times \alpha_i ' \times \beta_j '$
άρα για $\mu=1$

	$\beta_1=1$	$\beta_2=2$	$\beta_3=0.5$	$\beta_4=3/2$
$\alpha_1=1$	1	2	0.5	1.5
$\alpha_2=2$	2	4	3	3
$\alpha_3=4$	4	8	2	6
$\alpha_4=3$	3	6	1.5	4.5

2.3.4 Biclusters με συνεκτικές εξελίξεις (coherent evolutions)

Αυτός ο τύπος bicluster στοχεύει στην ομαδοποίηση γραμμών ή στηλών ανεξάρτητα από τις ακριβείς τιμές των στοιχείων του πίνακα [1]. Σε αντίθεση με τα biclusters με συνεκτικές τιμές τα biclusters με συνεκτικές εξελίξεις προσδιορίζουν υποσύνολα των γραμμών (γονιδίων) που οι τιμές τους αυξάνονται ή μειώνονται σταδιακά κατά μήκος μιας ομάδας στηλών ή γραμμών. Σε αντίθεση με τους άλλους τύπους, ο συγκεκριμένος τύπος είναι δύσκολο να μοντελοποιηθεί χρησιμοποιώντας κάποια μαθηματική εξίσωση.





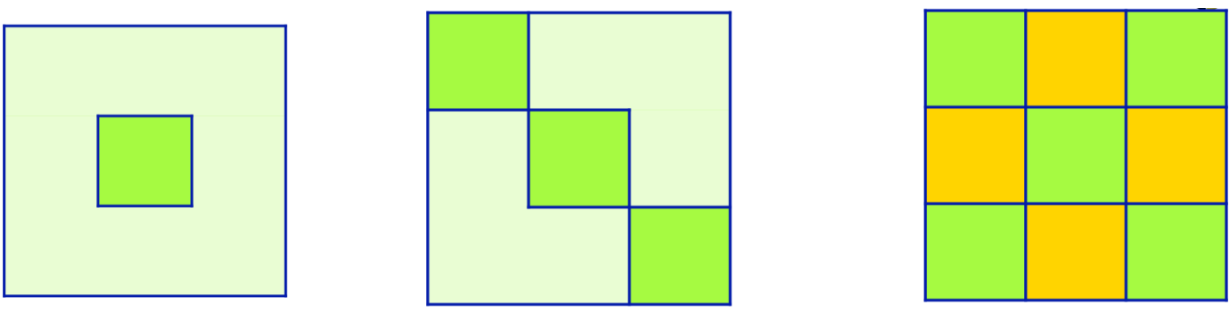
Εικόνα 15. Biclusters με συνεκτικές εξελίξεις [25].

Παρατηρώντας την εικόνα 15 με τις αριθμητικές τιμές είναι εμφανές ότι από τη μία στήλη στην επόμενη οι τιμές είτε όλες μειώνονται είτε όλες αυξάνονται όπως αναφέρθηκε παραπάνω.

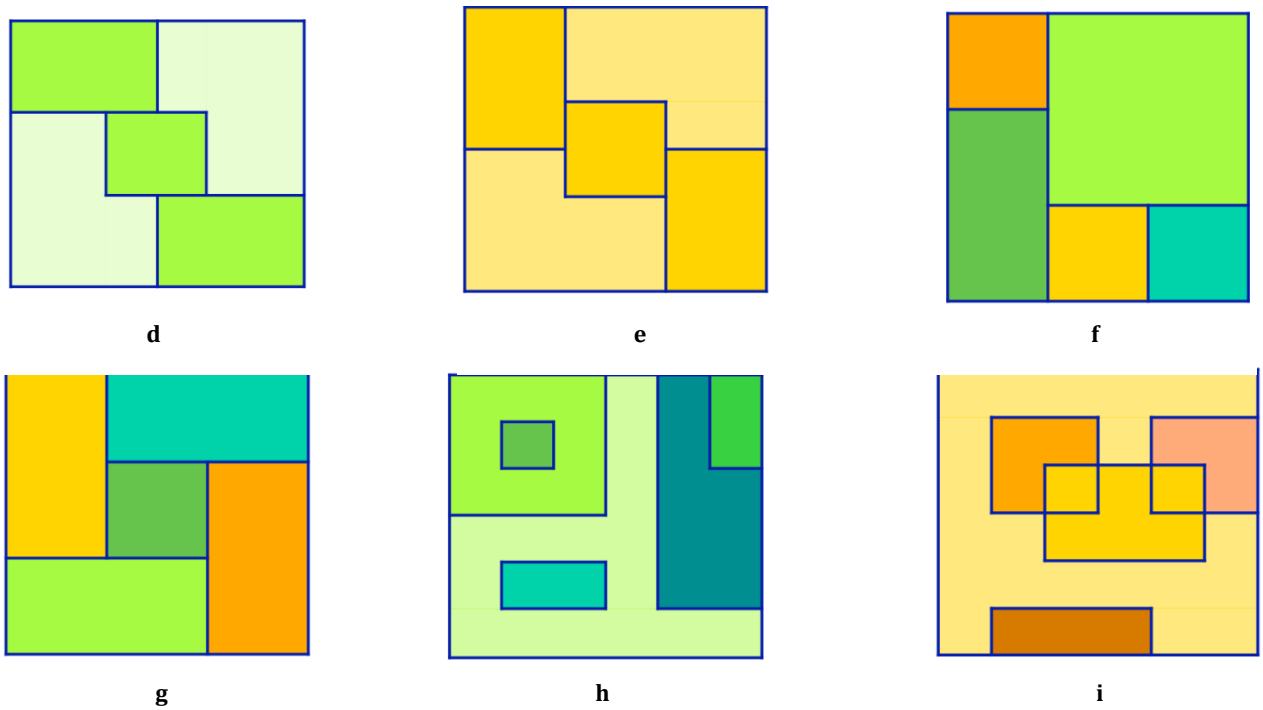
2.4 Δομή διπλής κατηγοριοποίησης

Οι δομές που μπορούν να προκύψουν από την εφαρμογή των bicluster αλγορίθμων χωρίζονται σε 2 κατηγορίες: ένα μοναδικό bicluster ή ένας μεγαλύτερος αριθμός από biclusters με διάφορες μορφές οι οποίες είναι οι ακόλουθες:

- Αποκλειστικής γραμμής και στήλης
- Μη επικαλυπτόμενο με τη μορφή σκακιέρας
- Αποκλειστικά με σειρές
- Αποκλειστικά με στήλες
- Μη-επικαλυπτόμενα με δενδροειδή μορφή
- Μη-επικαλυπτόμενα ,μη αποκλειστικά
- Επικαλυπτόμενα με ιεραρχική δομή
- Αυθαίρετα τοποθετημένα-επικαλυπτόμενα



a **b** **c**



Εικόνα 16. Δομή Bicluster: a. Μονό, b. Αποκλειστικής γραμμής και στήλης, c. Μη επικαλυπτόμενο με τη μορφή σκακιέρας, d. Αποκλειστικά με σειρές, e. Αποκλειστικά με στήλες, f. Μη-επικαλυπτόμενα με δενδροειδή μορφή, g. Μη-επικαλυπτόμενα, μη αποκλειστικά, h. Επικαλυπτόμενα με ιεραρχική δομή, i. Αυθαίρετα τοποθετημένα-επικαλυπτόμενα [25].

Ένας τρόπος για την εύρεση πολλών biclusters σε έναν πίνακα είναι η χρήση χρώματος όπου κάθε στοιχείο του πίνακα θα χρωματίζεται ανάλογα με την τιμή a_{ij}. Έτσι γίνεται πιο εύκολη η εναλλαγή γραμμών και στηλών ούτως ώστε να ομαδοποιηθούν γραμμές και στήλες που σχηματίζουν μπλοκς με παρόμοια χρώματα. Αυτά τα μπλοκς είναι υποπίνακες του αρχικού πίνακα με παρόμοιες τιμές έκφρασης και αποτελούν τα biclusters. Μια ιδανική αναδιάταξη δημιουργεί μια εικόνα με K τετράγωνα μπλοκς στη διαγώνιο του πίνακα (16.b). Αυτή η ιδέα στηρίζεται στο ότι κάθε γραμμή και στήλη του πίνακα μπορεί να ανήκει αποκλειστικά σε ένα από τα K biclusters. Στην πραγματικότητα όμως πολλές γραμμές και στήλες του πίνακα μπορεί να ανήκουν σε περισσότερα από ένα biclusters και είτε υπάρχει επικάλυψη μεταξύ των στοιχείων (16.h,i), είτε όχι (16.c,d,e,f,g) [1].

2.5 Αλγόριθμοι διπλής κατηγοριοποίησης

Δεδομένης της πολυπλοκότητας των αλγορίθμων επικρατούν 5 μέθοδοι για την εύρεση των biclusters:

- Επαναληπτικός συνδυασμός γραμμών και στηλών
- Διαίρει και βασίλευε
- Άπληστη επαναληπτική αναζήτηση
- Εξαντλητική απαρίθμηση
- Διανομή παραμέτρων αναγνώρισης

Υπάρχει ένας μεγάλος αριθμός αλγορίθμων που ακολουθούν τις πέντε αυτές μεθόδους. Παρακάτω παραθέτουμε τους πιο διαδεδομένους από αυτούς [1].

2.5.1 Επαναληπτικός συνδυασμός γραμμών και στηλών

Ο εννοιολογικά πιο απλός τρόπος εκτέλεσης bicluster χρησιμοποιώντας ήδη υπάρχουσες τεχνικές είναι η χρήση κλασσικών clustering μεθόδων στις γραμμές και τις στήλες του πίνακα, και ο συνδυασμός των αποτελεσμάτων για την εξαγωγή των biclusters. Ένας μεγάλος αριθμός συγγραφέων προτείνουν αλγορίθμους βασισμένους σε αυτή την ιδέα:

<u>Αλγόριθμος Coupled Two-Way Clustering (CTWC)</u>

Ο αλγόριθμος CTWC εντοπίζει ζευγάρια από μικρά υποσύνολα χαρακτηριστικών (F_i) και αντικειμένων (O_j) όπου μπορούν να είναι οι γραμμές ή οι στήλες του πίνακα, έτσι ώστε μόνο όταν τα χαρακτηριστικά (F_i) χρησιμοποιούνται για να ομαδοποιήσουν τα αντίστοιχα αντικείμενα (O_j), πραγματοποιούνται σταθερές και σημαντικές υποομάδες. Επιπλέον, χρησιμοποιεί μια διαδικασία για να αποφευχθεί η απαρίθμηση όλων των πιθανών συνδυασμών: μόνο υποσύνολα γραμμών ή στηλών που χαρακτηρίζονται από σταθερά συμπλέγματα σε προηγούμενες επαναλήψεις είναι υποψήφια για την επόμενη επανάληψη [1].

<u>Αλγόριθμος Interrelated Two-Way Clustering (ITWC)</u>

Ο συγκεκριμένος αλγόριθμος βασίζεται στο συνδυασμό των αποτελεσμάτων clustering που γίνεται σε μία από τις δύο διαστάσεις του πίνακα (γραμμές ή στήλες) ξεχωριστά. Σε κάθε επανάληψη πραγματοποιούνται πέντε βήματα. Στα δύο πρώτα γίνεται clustering πρώτα στις γραμμές (βήμα 1) και μετά στις στήλες (βήμα 2) και στα επόμενα 3 γίνεται ο συνδυασμός αυτών για την τελική εξαγωγή των biclusters [1].

<u>ΑλγόριθμοςDouble Conjugated Clustering (DCC)</u>

Ο αλγόριθμος αυτός πραγματοποιεί clustering σε γραμμές και στήλες του πίνακα χρησιμοποιώντας αυτοοργανωμένους χάρτες και μετρήσεις γωνιών σαν μέτρο ομοιότητας. Ο αλγόριθμος λειτουργεί επαναληπτικά κάνοντας έναν clustering κύκλο και μετατρέποντας κάθε στοιχείο στο συζυγές του, όπου ο νέος κύκλος λαμβάνει χώρα. Η διαδικασία επαναλαμβάνεται μέχρι ο αριθμός των μετατροπών των δειγμάτων πέσει κάτω από ένα συγκεκριμένο όριο και στις δύο διαστάσεις [1].

2.5.2 Διαίρει και βασίλευε

Οι αλγόριθμοι που ακολουθούν την συγκεκριμένη μέθοδο έχουν το πλεονέκτημα να είναι πολύ γρήγοροι. Όμως έχουν την τάση να χάνουν σημαντικά biclusters που αφαιρούνται πριν αναγνωριστούν [1].

<u>Αλγόριθμος Block Clustering</u>

Το block-clustering ήταν η πρώτη προσπάθεια αλγορίθμου που ακολουθεί αυτή τη μέθοδο από τον Hartigan. Το Block clustering αποτελεί ένα clustering γραμμών και στηλών από πάνω προς τα κάτω. Ο αλγόριθμος ξεκινά με όλο τον πίνακα σε ένα μπλοκ (bicluster). Σε κάθε επανάληψη βρίσκει τη γραμμή ή στήλη που παράγει τη μεγαλύτερη μείωση στην ολική - εντός του μπλοκ - διασπορά, χωρίζοντας το μπλοκ σε δύο κομμάτια. Για την εύρεση της καλύτερης διαίρεσης του μπλοκ στα δύο μέρη, οι γραμμές και οι στήλες είναι ταξινομημένες με βάση τις μέσες τιμές των γραμμών και των στηλών αντίστοιχα. Η διαίρεση συνεχίζεται μέχρι να παραχθεί ένας δοθέν αριθμός Κ από μπλοκς και η ολική διασπορά εντός των μπλοκς να φτάσει ένα συγκεκριμένο όριο [1].

2.5.3 Άπληστη επαναληπτική αναζήτηση

Η μέθοδος αυτή στηρίζεται στην ιδέα της δημιουργίας biclusters προσθέτοντας και αφαιρώντας γραμμές και στήλες με βάση κάποιο κριτήριο. Ορισμένες φορές μπορεί το αποτέλεσμα να μην περιλαμβάνει κάποια καλά biclusters, αλλά η συγκεκριμένη μέθοδος έχει το πλεονέκτημα της γρήγορης ταχύτητας [1].

Cheng and Church ή δ-biclustering

Ο αλγόριθμος Cheng and Church δοθέντος ενός πίνακα A και ενός μέγιστου επιτρεπτού mean residue score, δ>0 προσπαθεί να ανακαλύψει δ-biclusters (υποπίνακες του αρχικού πίνακα) με σκορ όχι μεγαλύτερο από του δ προσθαφαιρώντας γραμμές και στήλες από τον αρχικό πίνακα. Σε κάθε επανάληψη εξάγει από ένα bicluster του οποίου τα στοιχεία φιλτράρονται για την επόμενη επανάληψη ώστε να μην υπάρχει επικάλυψη [1].

Αλγόριθμος Floc

Ο αλγόριθμος Floc στηρίζεται στην ιδέα του Cheng and Church, αλλά εφαρμόζει στιγμιαία αναγνώριση biclusters. Ανακαλύπτει δηλαδή *K* πιθανώς επικαλυπτόμενα biclusters στιγμιαία. Αποτελείται από 2 φάσεις. Στην πρώτη εξάγονται K αρχικά biclusters προσθέτοντας κάθε στήλη/γραμμή στο καθένα με ανεξάρτητη πιθανότητα p. Η δεύτερη φάση είναι μια επαναληπτική διαδικασία που βελτιώνει την ποιότητα των προηγούμενων biclusters [1].

Αλγόριθμος Xmotifs

Ο αλγόριθμος Xmotifs βρίσκει biclusters με coherent evolutions. Η μέθοδος αυτή αναζητά γραμμές με σταθερές τιμές σε ένα σετ στηλών. Για πίνακες γονιδιακής έκφρασης, τα εξαγόμενα biclusters ονομάζονται “conserved genes expressions motifs” (συντηρημένα μοτίβα γονιδιακής έκφρασης), εξού και το όνομα του αλγορίθμου [29].

Αλγόριθμος OPSM

Ο αλγόριθμος OPSM στηρίζεται στην λειτουργία των μερικών μοντέλων. Στις περιπτώσεις που δεν είναι δυνατόν να δοκιμαστούν όλα τα πιθανά ολοκληρωμένα μοντέλα (biclusters), η ιδέα είναι η ανάπτυξη μερικών μοντέλων αρχικά, μέχρι να καταλήξουμε στα τελικά. Ένα ολοκληρωμένο μοντέλο είναι ένα ζεύγος *(J,π),* όπου *J* είναι ένα σετ από *s* στήλες και $\pi=(j_1, j_2,\ldots,j_s)$ μια γραμμική διάταξη των στηλών του *J*. Ο αλγόριθμος ξεκινά αξιολογώντας όλα τα (1,1) μερικά μοντέλα και κρατάει τα *l* καλύτερα από αυτά. Στη συνέχεια περνά στα (2,1) μοντέλα και κρατά τα *l* καλύτερα. Μετά στα (2,2), στα (3,2) κ.ο.κ. μέχρι να φτάσει στην επιλογή των *l ([s/2,s/2])* μοντέλων και να διαλέξει το καλύτερο από αυτά [1].

Αλγόριθμος Spectral

Αυτός ο αλγόριθμος που περιγράφηκε από τον Kluger κ.α. (2003) χρησιμοποιεί μια συγκεκριμένη τιμή και μία σειρά ιδιοτιμών και ιδιοδιανυσμάτων για να ανακτήσει τα biclusters από τα δεδομένα. Αυτό οδηγεί σε μία μορφή bicluster-σκακιέρας. Ο αλγόριθμος είναι πολύ ευαίσθητος σε μεταβολές των δεδομένων και γι αυτό χρειάζεται πολύ προσεκτική προ-επεξεργασία των δεδομένων πριν εφαρμοστεί. Ο αριθμός των biclusters καθορίζεται από ένα επιλεγμένο άνω όριο διακύμανσης [1].

2.5.4 Εξαντλητική απαρίθμηση

Η βασική ιδέα της μεθόδου αυτής είναι ότι τα καλύτερα biclusters μπορούν μόνο να αναγνωρισθούν χρησιμοποιώντας μια εξαντλητική απαρίθμηση όλων των πιθανών biclusters που υπάρχουν σε έναν πίνακα. Οι αλγόριθμοι που ακολουθούν τη συγκεκριμένη μέθοδο βρίσκουν τα βέλτιστα biclusters αν υπάρχουν, έχουν όμως ένα σοβαρό μειονέκτημα. Εξαιτίας της μεγάλης τους πολυπλοκότητας, μπορούν να εκτελεστούν μόνο υπό περιορισμούς στα μεγέθη των εξαγόμενων biclusters [1].

Αλγόριθμος SAMBA

Tanay κ.α. παρουσίασαν τον SAMBA, έναν αλγόριθμο που εξάγει στιγμιαία biclusters χρησιμοποιώντας εξαντλητική απαρίθμηση. Ο SAMBA αποφεύγει μια εκθετική χρονική πολυπλοκότητα, περιορίζοντας τον αριθμό των γραμμών που μπορεί να εμφανιστούν σε ένα bicluster [29].

Αλγόριθμος MDS

Ο αλγόριθμος αυτός εφαρμόζει εξαντλητική απαρίθμηση με βασικό κριτήριο τον περιορισμό ενός ελάχιστου αριθμού γραμμών και στηλών. Για να γίνει η διαδικασία πιο γρήγορη και για να αποφευχθούν επαναλήψεις κατά τους υπολογισμούς, χρησιμοποιείται ένα suffix δέντρο που απαριθμεί τους πιθανούς συνδυασμούς των σετ γραμμών και στηλών που αντιπροσωπεύουν έγκυρα biclusters [1].

2.5.5 Διανομή παραμέτρων αναγνώρισης

Αυτή η μέθοδος προσεγγίζει ένα δοθέν στατιστικό μοντέλο και προσπαθεί να αναγνωρίσει τις παραμέτρους εκείνες που χρησιμοποιούνται για την παραγωγή δεδομένων ελαχιστοποιώντας ένα συγκεκριμένο κριτήριο μέσω μιας επαναληπτικής διαδικασίας [1].

Αλγόριθμος Plaid Models

Ο συγκεκριμένος αλγόριθμος ανήκει στους Lazzeroni και Owen (2002). Τοποθετεί k επίπεδα με βάση το μαθηματικό μοντέλο :

$$\alpha_{ij} = (\mu_0 + \alpha_{i0} + \beta_{j0}) + \sum_{k=1}^{K} (\mu_k + a_{ik} + \beta_{jk}) p_{ik} \kappa_{jk} + \varepsilon_{ij}$$

Όπου μ, α, β συμβολίζουν τη μέση τιμή, τις γραμμές και τις στήλες αντίστοιχα, και ρ, κ δείχνουν αν μία γραμμή ή μία στήλη είναι μέλη του επιπέδου [29].

2.6 Συνολική σύγκριση αλγορίθμων

Ο πίνακας που ακολουθεί συνοψίζει τους περισσότερους αλγορίθμους biclustering της βιβλιογραφίας, τους τύπους, τον αριθμό και τις δομές των biclusters που εξάγουν, καθώς και τη μέθοδο που χρησιμοποιούν [1][26].

Πίνακας 2. Συνολική σύγκριση biclustering αλγορίθμων.

Αλγόριθμος	Τύπος	Δομή	Αριθμός	Μεθοδολογία
Block clustering	Constant	f	Ένα σετ τη φορά	Διαιρ-Βασιλ
δ-biclusters	Coherent Τιμές	f/g/i	Ένα τη φορά	Άπληστη
FLOC	Coherent Τιμές	i	Όλα ταυτόχρονα	Άπληστη
pClusters	Coherent Τιμές	g	Όλα ταυτόχρονα	Εξαντλ. Απαρ
Plaid Models	Coherent Τιμές	i	Ένα τη φορά	Διαν.Παρ.Αναγν
PRMs	Coherent Τιμές	i	Όλα ταυτόχρονα	Διαν.Παρ.Αναγν
CTWC	Constant στήλες	i	Ένα σετ τη φορά	Επαναλ.Συνδ
ITWC	Coherent Τιμές	d/e	Ένα σετ τη φορά	Επαναλ.Συνδ
DCC	Constant	b/c	Όλα ταυτόχρονα	Επαναλ.Συνδ
δ-patterns	Constant γραμμές	i	Όλα ταυτόχρονα	Άπληστη
Spectral	Coherent Τιμές	c	Όλα ταυτόχρονα	Άπληστη
Gibbs	Constant στήλες	d/e	Ένα τη φορά	Διαν.Παρ.Αναγν
OPSM	Coherent Evolution	a/i	Ένα τη φορά	Άπληστη
SAMBA	Coherent Evolution	i	Όλα ταυτόχρονα	Εξαντλ. Απαρ
xMOTIFS	Coherent Evolution	a/i	Όλα ταυτόχρονα	Άπληστη
OP-CLusters	Coherent Evolution	i	Όλα ταυτόχρονα	Εξαντλ. Απαρ

2.7 Εισαγωγή στον αλγόριθμο Cheng and Church

Στην ανάλυση των δεδομένων γονιδιακής έκφρασης, πέραν της ομαδοποίησης των γονιδίων που βασίζονται στην συνολική ομοιότητα, κρίνεται κάποιες φορές απαραίτητο να διασωθούν πληροφορίες που μπορεί να χάνονταν κατά τη διάρκεια του υπολογισμού των ομάδων. Στόχος ενός τέτοιου εγχειρήματος αποτελεί η αποκάλυψη της συμμετοχής ενός γονιδίου ή μίας κατάστασης σε περισσότερες από μία οδούς ή υποομάδες. Το biclustering επιτυγχάνει αυτή την προσπάθεια ομαδοποιώντας υποσύνολα γονιδίων και καταστάσεων με σκορ υψηλής ομοιότητας, η όποια αποτελεί μέτρο συνοχής για αυτά.

Ένα συγκεκριμένο σκορ που αντιπροσωπεύει τα λογαριθμισμένα δεδομένα έκφρασης αποτελεί το *mean squared residue score*. Το υπόλειμμα (residue) του στοιχείου a_{ij} σε ένα bicluster που αποτελείται από τα υποσύνολα I και J είναι το $a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$ όπου a_{iJ} είναι η μέση τιμή της *i*-οστης σειράς του bicluster, a_{Ij} η μέση τιμή της *j*-οστής στήλης, και a_{IJ} η μέση τιμή όλων των στοιχείων του bicluster. Η τιμή *mean squared residue score* αποτελεί την διακύμανση της ομάδας όλων των στοιχείων του bicluster και την μέση διακύμανση γραμμής και στήλης αντίστοιχα. Ο βασικός στόχος είναι η εύρεση μεγάλων biclusters με χαμηλό *mean squared residue score* και συγκεκριμένα μικρότερο από ένα συγκεκριμένο threshold. Μια ειδική περίπτωση για ένα βέλτιστο σκορ (μηδενικό *mean squared residue score)* είναι ένα constant bicluster με στοιχεία με ίδια, συγκεκριμένη τιμή. Όταν ένα bicluster έχει μη-μηδενικό σκορ είναι πάντα πιθανό αφαιρώντας κάποια γραμμή ή στήλη, το σκορ να μειωθεί, μέχρι το bicluster που θα απομείνει να είναι constant.

Στην περίπτωση της γονιδιακής έκφρασης, περισσότερο ενδιαφέρον δεν παρουσιάζει η εύρεση ενός μέγιστου bicluster, αλλά η ανακάλυψη ενός μεγαλύτερου αριθμού biclusters αποτελούμενα από ομάδες γονιδίων που εμφανίζουν παρόμοια ανοδική ή καθοδική συμπεριφορά κατά μήκος μιας σειράς καταστάσεων. Ένα χαμηλό *mean squared residue score* και μια μεγάλη παραλλαγή από την σταθερή μορφή αποτελούν καλά κριτήρια για την σωστή επιλογή αυτών των γονιδίων και καταστάσεων [30].

Στην τελευταία αυτή λογική βασίζεται ο αλγόριθμος Cheng and Church με συγγραφείς τους Yizong Cheng και George M. Church. Η έρευνά τους δημοσιεύτηκε το 2000 και διεξήχθη στο Lipper Center for Computational Genetics at the Harvard Medical School [31].

2.8 Βήματα αλγορίθμου

Ένας πίνακας γονιδίων-καταστάσεων αποτελείται από πραγματικούς αριθμούς, με πιθανές μηδενικές τιμές σε κάποια από τα στοιχεία του. Κάθε στοιχείο-κελί του πίνακα αντιπροσωπεύει το επίπεδο έκφρασης ενός γονιδίου κάτω από την αντίστοιχη συγκεκριμένη κατάσταση και αντιπροσωπεύεται από έναν πραγματικό αριθμό που αποτελεί τον λογάριθμο της σχετικής αφθονίας του mRNA του γονιδίου υπό αυτή την κατάσταση. Ο λογάριθμος χρησιμοποιείται για να μετατρέψει το doubling ή άλλες αλλαγές της αφθονίας των γονιδίων [30].

Έστω X ο αριθμός των γονιδίων και Y ο αριθμός των καταστάσεων, a_{ij} ο αριθμός έκφρασης του εκάστοτε γονιδίου, και I,J τα υποσύνολα των X και Y αντίστοιχα. Το σετ (I,J) ορίζει έναν υποπίνακα A_{IJ} με το ακόλουθο mean squared residue score:

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

Όπου:

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij} \quad , a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$$

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij} = \frac{1}{|I|} \sum_{i \in I} a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{Ij}$$

είναι οι μέσες τιμές των γραμμών και των στηλών και η μέση τιμή όλου του πίνακα (I,J). Ο υποπίνακας A_{ij} ονομάζεται δ-bicluster, αν $H(I,J) \leq \delta$ για κάποιο $\delta \geq 0$.

2.8.1 Αλγόριθμος 0

Διαγραφή Κόμβου

Κάθε πίνακας γονιδιακής έκφρασης περιέχει έναν υποπίνακα με βέλτιστο σκορ *(H(I,J)=0)* και κάθε ξεχωριστό στοιχείο είναι ένας τέτοιος υποπίνακας. Συγκεκριμένα όμως το είδος των biclusters που αναζητούνται πρέπει να έχουν ένα μέγιστο μέγεθος όσον αφορά και τα γονίδια και τις καταστάσεις.

Ξεκινώντας από το συνολικό πίνακα το ερώτημα είναι πως θα γίνει η σωστή επιλογή ενός υποπίνακα με χαμηλό *H score.* Μια άπληστη μέθοδος είναι να αφαιρεθούν η γραμμή ή η στήλη που θα πετύχει την μεγαλύτερη μείωση του σκορ. Αυτό απαιτεί τον υπολογισμό των σκορ όλων των υποπινάκων που θα αποτελούν αποτέλεσμα κάποιας αφαίρεσης γραμμής ή στήλης. Αυτή η μέθοδος (Αλγόριθμος 0) απαιτεί *O((n+m)nm)* χρόνο για την εύρεση ενός bicluster, όπου n και m είναι τα μεγέθη των γραμμών και στων στηλών του πίνακα αντίστοιχα [30].

Αλγόριθμος 0:Single Node Deletion

Είσοδος: Ένας πίνακας A, μία τιμή για το δ>=0 που αποτελεί το μέγιστα αποδεκτό mean squared residue score.

Έξοδος: AIJ, ένα δ-bicluster υποσύνολο του αρχικού πίνακα με score όχι μεγαλύτερο του δ.

Μέθοδος: Υπολογισμός score H για κάθε πιθανή πρόσθεση/αφαίρεση στήλης/γραμμής και επιλογή αυτών που μειώνουν όσο το δυνατόν περισσότερο το H. Ο αλγόριθμός σταματάει όταν καμία άλλη κίνηση δεν επηρεάζει το H ή αν H<=δ.

Ο αλγόριθμος 0 αποτέλεσε μια πρώτη προσέγγιση του τελικού αλγορίθμου, όχι αρκετά αποτελεσματική για μια γρήγορη ανάλυση στους περισσότερους γονιδιακούς πίνακες [30].

2.8.2 Βήματα 1 & 2

Μετέπειτα προτάθηκε το βήμα 1 με χρονική πολυπλοκότητα *O(nm)* και το βήμα 2 με χρονική πολυπλοκότητα *O(m* log*n)* ο συνδυασμός των οποίων (βήμα 3) αποτελεί μια αρκετά αποτελεσματική μέθοδο εύρεσης bicluster με χαμηλό σκορ. Η ακρίβεια και η ορθότητα αυτών των βημάτων στηρίζεται σε έναν αριθμό λημμάτων όπου οι γραμμές και οι στήλες αντιμετωπίζονται σαν σημεία σε ένα διάστημα όπου η απόσταση είναι ορισμένη [30].

Λήμμα 1. Έστω S ένα πεπερασμένο σετ από σημεία σε ένα διάστημα όπου έχει οριστεί μια μη αρνητική πραγματική συνάρτηση d, 2 στοιχείων. Έστω m(S) ένα σημείο που ελαχιστοποιεί την συνάρτηση:

$$f(s) = \sum_{x \in S} d(x, s) \quad (1).$$

Στη συνέχεια ορίζεται το μέτρο:

$$E(S) = \frac{1}{S} \sum_{x \in S} d\big(x, m(S)\big) \quad (2).$$

και η αφαίρεση οποιουδήποτε μη-άδειου υποσυνόλου

$$R \subset \{x \in S : d(x, m(S) > E(S)\} \quad (3)$$

θα συντελέσει μόνο στο $E(S - R) < E(S)$ (4).

<u>Απόδειξη</u>

Η σχέση (4) μπορεί να γραφτεί ως:

$$\frac{A'}{|S - R|} < \frac{A + B}{|S|} \qquad (5)$$

Όπου

$$A = \sum_{x \in S - R} d(x, m(S)), \quad A' = \sum_{x \in S - R} d(x, m(S - R)), \qquad B = \sum_{x \in R} d\big(x, m(S)\big) \ (6).$$

Ο ορισμός της συνάρτησης m προϋποθέτει ότι $A' \leq A$. Έτσι μια επαρκής κατάσταση για την ανισότητα (5) είναι η:

$$\frac{A}{|S-R|} < \frac{A+B}{|S|} \quad (7)$$

που είναι ισοδύναμη με:

$$E(S) = \frac{A+B}{|S|} < \frac{B}{|R|} = \frac{1}{|R|}\sum_{x\in R} d\big(x, m(S)\big) \quad (8).$$

Είναι φανερό ότι η σχέση (3) αποτελεί επαρκή όρο για αυτή την ανισότητα και επιπλέον για τη σχέση (4).

Λήμμα 2. Υποθέτουμε ότι το σετ που αφαιρέθηκε από το S είναι το

$$R \subset \{x \in S: d(x, m(S) > \alpha E(S)\} \quad (9)$$

με α≥1. Τότε ο ρυθμός μείωσης του σκορ *E(S)* μπορεί να χαρακτηριστεί ως:

$$\frac{E(S) - E(S-R)}{E(S)} > \frac{a-1}{\frac{|S|}{|R|} - 1} \quad (10).$$

Όταν ένα μοναδικό σημείο x αφαιρείται , ο ρυθμός μείωσης του σκορ έχει το όριο

$$E(S) - E(S-R) > \frac{d(x, m(S) - E(S)}{|S| - 1} \quad (11).$$

<u>Απόδειξη</u>

Χρησιμοποιώντας τη σημειογραφία του λήμματος 1 έχουμε ότι

$$\alpha E(S) = \alpha\frac{A+B}{|S|} < \frac{B}{|R|} = \frac{1}{|R|}\sum_{x\in R} d\big(x, m(S)\big) \quad (12).$$

Αυτό οδηγεί στη σχέση:

$\alpha|R|A < (|S| - a|R|)B \quad (13)$ ή $\textit{εναλλακτικά} \ |S|A < (|S| - a|R|)(A+B) \quad (14).$

Αυτές οι σχέσεις είναι ίδιες με την ακόλουθη:

$$\frac{A}{|S-R|} < \frac{|S| - a|R|}{|S-R|}\ \frac{A+B}{|S|} \quad (15).$$

Χρησιμοποιώντας την ανισότητα $A' \leq A$ και το ότι $E(S-R)=A'/|S-R|$ και $E(S)=(A+B)/|S|$ καταλήγουμε στην ανισότητα:

$$E(S-R) < \frac{|S| - a|R|}{|S-R|} E(S), \quad (16)$$

ουσιαστικά δηλαδή η σχέση (10).

Θεώρημα 1. Το σετ των γραμμών που μπορούν ολικώς ή μερικώς να αφαιρεθούν με επίδραση στη μείωση του σκορ ενός bicluster A_{IJ}, είναι:

$$R = \{i \in I, \frac{1}{|J|}\sum_{j \in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 > H(I,J)\} \quad (17)\,.$$

Απόδειξη. Έστω ότι τα σημεία του λήμματος 1 είναι $|J|$-διαστάσεων πραγματικά διανύσματα και S είναι το σετ των διανυσμάτων b_i με συνιστώσες $b_{ij} = a_{ij} - a_{iJ}$ για $i \epsilon I$ και $j \epsilon J$. Η συνάρτηση d ορίζεται ως εξής:

$$d(b_i, b_k) = \sum_{j \epsilon J} (b_{ij} - b_{kj})^2 \quad (18).$$

Σε αυτή την περίπτωση: $m(S) = \frac{1}{|I|}\sum_{i \in I} b_i$ (19)

και έχει τις συνιστώσες: $a_{Ij} - a_{IJ}$.

Παρόμοιο θεώρημα αντιστοιχεί και στις στήλες.

Το λήμμα 2 λειτουργεί σαν οδηγός στην εναλλαγή 2 τύπων διαγραφής κόμβων, αυτόν με τη διαγραφή ενός κόμβου τη φορά, και αυτόν με τη διαγραφή ενός σετ κόμβων τη φορά.

Βήμα 1:Single Node Deletion

Είσοδος: Ένας πίνακας A, μία τιμή για το δ>=0 που αποτελεί το μέγιστα αποδεκτό mean squared residue score.

Έξοδος: AIJ, ένα δ-bicluster υποσύνολο του αρχικού πίνακα με score όχι μεγαλύτερο του δ.

Μέθοδος:

1. Υπολογίζει aiJ για όλα τα i που ανήκουν στο I, aIj για όλα τα j που ανήκουν στο J, aIJ και H. Αν H<=δ επιστρέφει τον υποπίνακα AIJ αλλιώς:

2. Βρίσκει την γραμμή i που ανήκει στο I με το μεγαλύτερο

$$d(i) = \frac{1}{|J|}\sum_{j \in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

και τη στήλη j που ανήκει στο J με το μεγαλύτερο

$$d(i) = \frac{1}{|I|}\sum_{i \in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2$$

και αφαιρεί αυτήν (στήλη ή γραμμή) με τη μεγαλύτερη ποσότητα d ανανεώνοντας τα I, J.

Η ορθότητα του βήματος 1 φαίνεται από το θεώρημα 1, με την έννοια ότι κάθε αφαίρεση μειώνει το τελικό σκορ. Επειδή υπάρχει πεπερασμένος αριθμός γραμμών και στηλών προς αφαίρεση, ο αλγόριθμος τερματίζει σε όχι περισσότερες από $n+m$ επαναλήψεις. Υπάρχει όμως η περίπτωση όπου όλα τα $d(i)$ και $d(j)$ είναι όμοια με το $H(I,J)$ $\gamma\iota\alpha$ $i \in I$ και $j \in J$ και εδώ το θεώρημα 2 δεν ανταποκρίνεται. Σε αυτή την περίπτωση, η αφαίρεση μιας γραμμής ή μίας στήλης μπορεί ακόμα να μειώσει το σκορ εκτός αν είναι ήδη μηδενικό [30].

Το πρώτο στάδιο του βήματος σε κάθε επανάληψή του απαιτεί $O(nm)$ χρόνο και ο συνολικός υπολογισμός όλων των d τιμών στο στάδιο 2 χρειάζεται επίσης $O(nm)$ χρόνο. Η επιλογή της καλύτερης γραμμής και στήλης προς αφαίρεση παίρνει $O(\log n + \log m)$ χρόνο [30].

Βήμα 2: Multiple Node Deletion

Είσοδος: Ένας πίνακας A, μία τιμή για το δ>=0 που αποτελεί το μέγιστα αποδεκτό mean squared residue score και μια τιμή για το α>1, ένα threshold για τη μέθοδο Multiple node deletion [30].

Έξοδος: A_{IJ}, ένα δ-bicluster υποσύνολο του αρχικού πίνακα με score όχι μεγαλύτερο του δ.

Μέθοδος:

1. Υπολογίζει a_{iJ} για όλα τα i που ανήκουν στο I, a_{Ij} για όλα τα j που ανήκουν στο J, a_{IJ} και H. Αν H<=δ επιστρέφει τον υποπίνακα A_{IJ} αλλιώς:

2. Αφαιρεί όλες τις γραμμές για τις οποίες ισχύει ότι:

$$\frac{1}{|J|}\sum_{j\in J}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2 > aH(I,J).$$

3. Υπολογίζει ξανά τα μεγέθη aIj , aIJ και H.

4. Αφαιρεί όλες τις στήλες για τις οποίες ισχύει ότι

$$\frac{1}{|I|}\sum_{i\in I}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2 > aH(I,J).$$

5. Αν δεν έχει αφαιρεθεί τίποτα εκτελείται ο αλγόριθμος 1.

Η ορθότητα του δεύτερου αλγόριθμου στηρίζεται στο λήμμα 2. Όταν το α επιλέγεται σωστά, ο αλγόριθμος 2 (πριν την κλήση του αλγόριθμου 1) πραγματοποιεί μια σειρά επαναλήψεων σε $O(\log n + \log m)$ χρόνο που είναι συνήθως πολύ σύντομος.

2.8.3 Βήμα 3

Μετά τη διαγραφή κόμβων, το δ-bicluster μπορεί να μην αποτελεί το μέγιστο δυνατό, με την έννοια ότι θα μπορούσαν να προστεθούν σε αυτό κάποιες γραμμές ή στήλες που δεν θα αυξήσουν το σκορ. Ακολουθούν το θεώρημα 2 και λήμμα 3 ως οδηγοί για την προσθήκη κόμβων [30].

Λήμμα 3. Έστω *S*, *d*, *m(S)* και *E(S)* ορισμένα όπως και στο λήμμα 1. Τότε η πρόσθεση στο σετ S οποιασδήποτε μη-άδειας υποομάδας

$$R \subset \{x \notin S: d(x, m(S) \leq E(S)\} \quad (20)$$

δεν θα αυξήσει το σκορ E:

$$E(S+R) \leq E(S) \quad (21).$$

Απόδειξη. Η σχέση (21) μπορεί να ξαναγραφτεί ως:

$$\frac{A'}{|S+R|} \leq \frac{A-B}{|S|} \quad (22),$$

όπου

$$A = \sum_{x \in S+R} d(x, m(S)), \quad A' = \sum_{x \in S+R} d(x, m(S+R)), \qquad B = \sum_{x \in R} d\big(x, m(S)\big) \quad (23).$$

Ο ορισμός της συνάρτησης m προυποθέτει ότι $A' \leq A$. Έτσι μια επαρκής κατάσταση για την ανισότητα (22) είναι η:

$$\frac{A}{|S+R|} \leq \frac{A-B}{|S|} \quad (24)$$

που είναι ισοδύναμη με :

$$E(S) = \frac{A-B}{|S|} \geq \frac{B}{|R|} = \frac{1}{|R|} \sum_{x \in R} d\big(x, m(S)\big) \quad (25).$$

Είναι φανερό ότι η σχέση (20) αποτελεί επαρκή όρο για αυτή την ανισότητα.

Θεώρημα 3. Το σετ των γραμμών που μπορούν ολικώς ή μερικώς να προστεθούν με επίδραση στη μείωση του σκορ ενός bicluster A_{IJ} είναι:

$$R = \{i \notin I, \frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J)\} \quad (26)\,.$$

Απόδειξη. Παρόμοια με του θεωρήματος 2 .

Παρόμοιο θεώρημα αντιστοιχεί και στις στήλες.

Βήμα 3:Node Addition

Είσοδος: Ένας πίνακας A, και I, J που αποτελούν ένα δ-bicluster.

Έξοδος: I' και J' όπου I' ανήκει στο I και J' ανήκει στο J με την ιδιότητα ότι *H(I',J')<= H(I,J).*

Μέθοδος:

1.Υπολογίζει a_{iJ} για όλα τα i που ανήκουν στο I, a_{Ij} για όλα τα j που ανήκουν στο J, a_{IJ} και H.

2. Προσθέτει όλες τις στήλες που δεν ανήκουν στο J για τις οποίες:

$$\frac{1}{|I|}\sum_{i\in I}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J).$$

3.Υπολογίζει ξανά τα μεγέθη a_{iJ} , a_{Ij} και H.

4. Προσθέτει όλες τις γραμμές που δεν ανήκουν στο I για τις οποίες:

$$\frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \leq H(I,J).$$

5.Αν σταματάει να προστίθεται κάτι, δίνονται τα I' και J' στην έξοδο.

Το λήμμα 3 και το θεώρημα 2 εγγυώνται ότι η πρόσθεση γραμμών και στηλών στο βήμα 3 δεν θα αυξήσουν το σκορ. Παρόλ' αυτά το δ-bicluster που δίνεται στην έξοδο μπορεί να μην είναι μέγιστο για 2 λόγους. Ο πρώτος είναι ότι το λήμμα 3 δίνει μια πιθανή περίπτωση πρόσθεσης γραμμής και στήλης που δεν είναι απαραίτητα υποχρεωτική. Ο δεύτερος λόγος είναι ότι προσθέτοντας γραμμές και στήλες το σκορ μπορεί να μειωθεί κατά πολύ σε σχέση με το δ, και αυτό συμβαίνει γιατί σε κάθε επανάληψη του βήματος η πρόσθεση γίνεται με βάση το σκορ εκείνη τη δεδομένη στιγμή και όχι με βάση το ορισμένο από την αρχή δ.

Το βήμα 3 είναι πολύ αποτελεσματικό. Η χρονική του πολυπλοκότητα είναι συγκρίσιμη με αυτήν του βήματος 2 και είναι περίπου της τάξης του *O(nm)* [30].

2.8.4 Βήμα 4

Τα βήματα που περιγράφηκαν παραπάνω, περικλείονται στο βήμα 4. Οι τιμές των παραμέτρων δ, α και n καθορίζονται πριν την έναρξη του αλγορίθμου αυθαίρετα.

Σε περίπτωση μηδενικών τιμών του πίνακα γίνεται αντικατάστασή τους με τυχαίους αριθμούς που θα αποτελούν τους πρώτους υποψηφίους προς αφαίρεση στη διαδικασία αφαίρεσης κόμβων.

Επισημαίνεται ότι επειδή ο αλγόριθμος CC είναι ντετερμινιστικός, επαναληπτικές εφαρμογές του δεν θα δώσουν διαφορετικά αποτελέσματα, εκτός αν χρησιμοποιηθεί κάποιου είδους μάσκα στα αποτελέσματα. Κάθε φορά επομένως που εξάγεται ένα bicluster, τα στοιχεία του υποπίνακα αντικαθίστανται από τυχαίους αριθμούς [30].

Βήμα 4: Βρίσκοντας έναν συγκεκριμένο αριθμό biclusters

Είσοδος: Ένας πίνακας A, μία τιμή για το δ>=0 που αποτελεί το μέγιστα αποδεκτό mean squared residue score, μια τιμή για το α>1 (threshold για τη μέθοδο Multiple node deletion) και τον αριθμό των biclusters που θέλουμε να βρεθούν.

Έξοδος: n δ-biclusters στον A πίνακα. Επανάληψη n φορές:

Μέθοδος:

1. Εφαρμογή βήματος 2. Αν το μέγεθος των γραμμών ή στηλών είναι μικρό (<100), δεν πραγματοποιείται multiple node deletion. Ο πίνακας μετά από αυτό το βήμα είναι ο B.

2. Εφαρμογή βήματος 1 στον B και ο πίνακας μετά από αυτό το βήμα αποτελεί τον C.

3. Εφαρμογή βήματος 3 στον A και τον C και το αποτέλεσμα είναι το bicluster D.

4. Εξαγωγή bicluster D και αντικατάσταση των στοιχείων του στον αρχικό πίνακα με τυχαίους αριθμούς.

ΚΕΦΑΛΑΙΟ 3
ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Δεδομένα προς επεξεργασία

Η ποσοτική μέτρηση της γονιδιακής έκφρασης χρησιμοποιώντας μικροσυστοιχίες πραγματοποιήθηκε πρώτη φορά από τον Schena κ.α. το 1995 σε 45 γονίδια *Arabidopsis thaliana* και αργότερα σε χιλιάδες γονίδια ή σε ένα ολόκληρο γονιδίωμα (DeRisi κ.α. 1996, 1997) [27].

Οι μικροσυστοιχίες είναι στερεά υποστρώματα που φιλοξενούν εκατοντάδες μονοκλωνικά DNA με συγκεκριμένη αλληλουχία, τα οποία βρίσκονται σε εντοπισμένα σημεία οργανωμένα σε δίκτυα. Αυτά τα μόρια, που ονομάζονται *probes,* υβριδοποιούνται με μονοκλωνικά μόρια cDNA (στόχοι), τα οποία έχουν επισημανθεί κατά τη διάρκεια της διαδικασίας αντίστροφης μεταγραφής. Οι στόχοι αντικατοπτρίζουν την ποσότητα mRNA που απομονώνεται από ένα δείγμα που ελήφθη κάτω από μια συγκεκριμένη κατάσταση. Έτσι η ποσότητα φθορισμού που εκπέμπεται από κάθε κηλίδα είναι ανάλογη με την ποσότητα του mRNA που μεταγράφεται από την αντίστοιχη αλληλουχία DNA. Η μικροσυστοιχία σαρώνεται και η εικόνα που προκύπτει αναλύεται χρησιμοποιώντας τεχνικές επεξεργασίας σήματος και εικόνας, έτσι ώστε το σήμα από κάθε μόριο (probe) να μπορεί να ποσοτικοποιηθεί σε αριθμητικές τιμές. Αυτές οι τιμές αντιπροσωπεύουν το επίπεδο έκφρασης του γονιδίου σε μια δεδομένη κατάσταση [27].

Ένας πίνακας γονιδιακής έκφρασης επομένως προσδιορίζεται από έναν *NxM* πίνακα :

$$A = \begin{bmatrix} g(1) \\ g(2) \\ \vdots \\ g(n) \\ \vdots \\ g(N) \end{bmatrix} = \begin{bmatrix} a(1,1) & a(1,2) & \cdots & a(1,m) & \cdots & a(1,M) \\ a(2,1) & a(2,2) & \cdots & a(2,m) & \cdots & a(2,M) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a(n,1) & a(n,2) & \cdots & a(n,m) & \cdots & a(n,M) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a(N,1) & a(N,2) & \cdots & a(N,m) & \cdots & a(N,M) \end{bmatrix}$$

όπου *G={g(1), g(2), …, g(n), …g(N)}* απεικονίζει το σύνολο των γονιδίων. Κάθε στοιχείο του G αντιστοιχεί σε μία γραμμή του πίνακα *C={c(1), c(2), …, c(m), …c(M)}* που αναπαριστά το σετ των πειραματικών καταστάσεων, των χρονικών στιγμών ή των κυτταρικών σειρών [27].

Το αρχείο προς επεξεργασία που μας δόθηκε αποτελείται από έναν τέτοιο πίνακα 33096 γραμμών και 38 στηλών. Οι γραμμές αντιπροσωπεύουν τα γονίδια και οι στήλες τις καρκινικές κυτταρικές σειρές . Οι τύποι των καρκίνων από τους οποίους προέρχονται οι σειρές είναι:

- καρκίνος του μαστού
- καρκίνος του τραχήλου της μήτρας
- καρκίνος του ενδομητρίου
- καρκίνος των ωοθηκών

και στον καθένα από αυτούς ανήκουν 10, 9, 9 και 10 σειρές αντίστοιχα.

Η μέθοδος διπλής κατηγοριοποίησης εφαρμόστηκε σε ένα σύνολο δεδομένων, το οποίο χορηγήθηκε ευγενικά στο Εργαστήριο Ψηφιακής Επεξεργασίας Σήματος και Εικόνας από την Δρ. Obermayr (Ιατρικό Πανεπιστήμιο Βιέννης, Τμήμα Γενικής Γυναικολογίας και Γυναικολογικής Ογκολογίας), που αφορούσε σε τέσσερις διαφορετικούς τύπους καρκινικών κυτταρικών σειρών: α) κυτταρικές σειρές καρκίνου του μαστού, β) κυτταρικές σειρές καρκίνου του τραχήλου της μήτρας, γ) κυτταρικές σειρές καρκίνου του ενδομητρίου, και δ) κυτταρικές σειρές καρκίνου των ωοθηκών.

Στην πρωτότυπη εργασία της Δρ. Obermayr, χρησιμοποιήθηκε τεχνολογία μικροσυστοιχιών (Applied Biosystems) για την μέτρηση της γονιδιακής έκφρασης των 38 καρκινικών κυτταρικών σειρών (του μαστού, των ωοθηκών, του τραχήλου της μήτρας και του ενδομητρίου) και των 10 δειγμάτων μονοπύρηνων κυττάρων περιφερικού αίματος (PBMCs) από υγιείς γυναίκες δότριες. Ο στόχος της εργασίας των Obermayr και συνεργατών (2010) ήταν ο προσδιορισμός νέων γονιδιακών δεικτών για την ανίχνευση κυκλοφορούντων καρκινικών κυττάρων (CTCs) στο περιφερικό αίμα των θηλέων ασθενών με καρκίνο [32].

Η δική μας μελέτη, αν και καλείται επίσης να ανιχνεύσει και να προσδιορίσει καινούργιους γονιδιακούς δείκτες, διαφοροποιείται από την αρχική ως προς τα εξής σημεία: α) την ταυτότητα των γονιδιακών δεικτών, που δεν σχετίζεται με την ανίχνευση των CTCs αλλά με την ανίχνευση *ειδικών*-καρκινικών δεικτών για κάθε τύπο καρκίνου και *πολυγονιδιακών* καρκινικών δεικτών για το σύνολο των τεσσάρων καρκινικών τύπων, β) τον αριθμό των δειγμάτων που αποτελούν το σύνολο δεδομένων, όπου χρησιμοποιούνται αποκλειστικά τα δείγματα των 38 καρκινικών κυτταρικών σειρών, και δεν συμπεριλαμβάνονται τα 10 δείγματα των PBMCs από υγιείς γυναίκες δότριες και γ) τη μεθοδολογία εξόρυξης των αποτελεσμάτων, όπου χρησιμοποιείται μέθοδος διπλής κατηγοριοποίησης και δεν ακολουθείται στρατηγική «κατάρρους μείωσης» (step down strategie).

Η προεπεξεργασία των δεδομένων μας αποτελεί ένα σημαντικό βήμα πριν την εφαρμογή οποιουδήποτε αλγορίθμου. Σκοπός της κανονικοποίησης των στοιχείων του πίνακα είναι ο εντοπισμός συστηματικών διαφορών μεταξύ των συνόλων δεδομένων και η εξάλειψη των τεχνικών λαθών. Η πρόκληση της

κανονικοποίησης είναι να αφαιρεθούν όσο το δυνατόν περισσότερες τεχνικές διακυμάνσεις αφήνοντας τις αντίστοιχες βιολογικές ανέγγιχτες. Μία από τις πιο διαδεδομένες τεχνικές κανονικοποίησης είναι η λογαρίθμιση [27].

Οι τιμές του πίνακα μας ήταν ήδη λογαριθμισμένες και δεν περιείχαν μηδενικές τιμές. Το εύρος των τιμών του πινάκα κυμαίνονταν από 3,6764 (ελάχιστη τιμή) έως 19,7826 (μέγιστη τιμή).

Βιολογική ερμηνεία σύμφωνα με το πρόσθετο εργαλείο ClueGO

Για την ερμηνεία των αποτελεσμάτων έγινε αρχικά ταυτοποίηση των γονιδίων με τη χρήση κώδικα που είχε αναπτυχθεί στο εργαστήριο για την μετατροπή των κωδικών αριθμών (ABI Probe Ids) της πλατφόρμας GPL9851 σε κωδικούς γονιδίων (Entrez Gene IDs). Σ' αυτό το σημείο, θα πρέπει να αναφερθεί ότι η πλατφόρμα GPL9851 περιέχει την πληροφορία για τις αλληλουχίες του γονιδιώματος της μικροσυστοιχίας έκφρασης Human Genome Survey Microarrays Hs.v1 (Applied Biosystems) που χρησιμοποιήθηκε στην πρωτότυπη εργασία των Obermayr και συνεργατών (2010) [32].

Στη συνέχεια ακολούθησε ο σχολιασμός των "ταυτοποιημένων" γονιδίων των ομάδων διπλής κατηγοριοποίησης με τη βοήθεια του εργαλείου WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) [33] .

Τέλος, εξετάστηκαν όλες οι ομάδες διπλής κατηγοριοποίησης ως προς το βιολογικό τους περιεχόμενο με τη βοήθεια του πρόσθετου εργαλείου ClueGO [34] της πλατφόρμας Cytoscape [35], διευκολύνοντας τη βιολογική ερμηνεία των αποτελεσμάτων.

Το πρόσθετο εργαλείο ClueGO της πλατφόρμας Cytoscape, μιας πλατφόρμας ανοικτού λογισμικού για την οπτικοποίηση πολύπλοκων δικτύων και την ενσωμάτωση με οποιοδήποτε χαρακτηριστικό τύπο δεδομένων, οπτικοποιεί τους μη-περιττούς, σημαντικούς βιολογικούς όρους μεγάλων ομάδων γονιδίων σε ένα λειτουργικά ομαδοποιημένο δίκτυο. Οι λίστες των γονιδίων μπορούν να φορτωθούν από ένα αρχείο κειμένου. Οι όροι επιλέγονται από διάφορες πηγές, όπως τη γονιδιακή οντολογία (Gene Ontology, GO) [36] και τα μοριακά μονοπάτια KEGG [37], Wikipathways [38] και Reactome [39].

Το δίκτυο ClueGO δημιουργείται με στατιστική κάπα (kappa statistics) και ανακλά τις σχέσεις μεταξύ των βιολογικών όρων με βάση την ομοιότητα των συνδεδεμένων γονιδίων τους. Κατά την ανάλυση των γονιδιακών ομάδων χρησιμοποιούνται διαφορετικά κριτήρια φίλτρου και οι βιολογικοί όροι που ανάγονται στα γονίδια που σχετίζονται μεταξύ τους μπορούν να συγχωνευθούν, έτσι ώστε να μειωθεί ο πλεονασμός των βιολογικών όρων.

Η ιδιαιτερότητα και οι κοινές πτυχές των γονιδίων που συνιστούν τις ομάδες αποτυπώνονται γραφικά στα δίκτυα και διαγράμματα του ClueGO επισημαίνοντας τη βιολογική σημασία της κάθε ομάδας.

Η κατηγορία των βιολογικών διεργασιών που περιλαμβάνεται στην γονιδιακή οντολογία (GO), αναφέρεται σε διεργασίες ή ένα σύνολο από μοριακά γεγονότα με μια καθορισμένη αρχή και τέλος, τα οποία σχετίζονται με τη λειτουργία ολοκληρωμένων έμβιων μονάδων όπως κύτταρα, ιστοί, όργανα και οργανισμοί. Τα μοριακά μονοπάτια KEGG, Wikipathways και Reactome αποτελούν συλλογές από χαρτογραφημένα μονοπάτια τα οποία απεικονίζουν τις γνώσεις μας σχετικά με τις βιοχημικές αντιδράσεις που λαμβάνουν χώρα μεταξύ των μορίων σε ένα κύτταρο, οδηγώντας σε ένα συγκεκριμένο προϊόν ή μια αλλαγή σε ένα κύτταρο. Για παράδειγμα, αυτά τα μονοπάτια αφορούν στην επεξεργασία της γενετικής πληροφορίας, την μεταγωγή σήματος, το μεταβολισμό, την ανταπόκριση σε στρεσογόνες καταστάσεις, και τις ανθρώπινες ασθένειες.

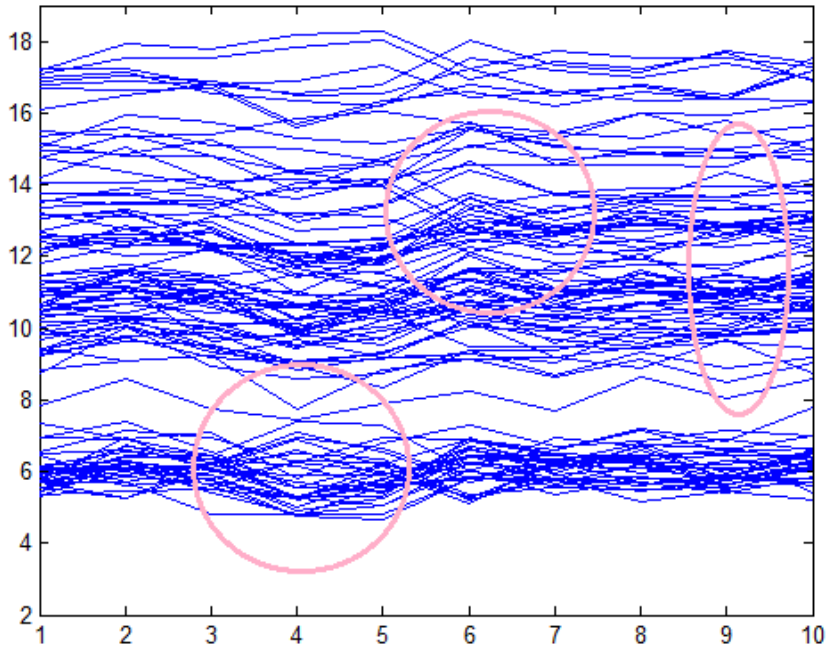
3.2 Επιλογή αλγορίθμου

Διαφορετικές μέθοδοι biclustering χρησιμοποιούν διαφορετικές έννοιες και αντιμετωπίζουν το πρόβλημα bicluster από άλλη οπτική ο καθένας. Γι' αυτό το λόγο καθίσταται αρκετά δύσκολη η επιλογή του πιο κατάλληλου αλγορίθμου για εφαρμογή [22]. Ο αλγόριθμος Cheng and Church είναι ένας από τους πιο παλιούς και αξιόπιστους αλγορίθμους. Τον επιλέξαμε γιατί αποτελεί μία μέθοδο ικανή να αντιμετωπίσει το μεγάλο dataset μας, παρέχει την δυνατότητα μη-επικάλυψης και τα biclusters που εξάγει έχουν coherent τιμές, γεγονός που μας επιτρέπει να μελετήσουμε τη συμπεριφορά των γονιδίων κατά μήκος των σειρών χωρίς οι τιμές απαραίτητα να είναι σταθερές.

Εφαρμογή Αλγορίθμου Cheng and Church

Ο αλγόριθμος Cheng and Church, όπως αναλύθηκε στο προηγούμενο κεφάλαιο, εφαρμόστηκε 5 φορές, μία για κάθε τύπο καρκίνου ξεχωριστά και μία φορά και για τους 4 τύπους μαζί για διάφορες τιμές των παραμέτρων *δ* και *α* ούτως ώστε να μελετηθούν τα αποτελέσματα και να καταλήξουμε στις πιο ενδεικτικές. Για τη μάσκα των στοιχείων των εξαγόμενων biclusters λόγω ντετερμινιστικότητας του αλγορίθμου χρησιμοποιήθηκε μια γεννήτρια τυχαίων αρνητικών αριθμών από -1 έως -200 για την αποφυγή επικάλυψης των τελικών μας αποτελεσμάτων.

Εφαρμόζοντας τον αλγόριθμο στο dataset, για τις περισσότερες τιμές των *δ* και *α*, το αποτέλεσμα ήταν αρκετά ικανοποιητικό με ορισμένες όμως παραμέτρους που θα δημιουργούσαν δυσκολία στην περαιτέρω μελέτη και επεξεργασία των αποτελεσμάτων.



Εικόνα 17. Ένα από τα πρώτα αποτελέσματα του αλγορίθμου Cheng and Church.

Η παραπάνω εικόνα αποτελεί ένα δείγμα από τα πρώτα αποτελέσματα του αλγορίθμου για δ=0.08 και α=1.2. Η γραφική απεικονίζει για την ομάδα - bicluster - την συμπεριφορά (γονιδιακή έκφραση) των γονιδίων κατά μήκος των σειρών που ανήκουν σε αυτή. Ο άξονας x απεικονίζει τον αριθμό των σειρών του bicluster, ο άξονας y τις τιμές έκφρασης των γονιδίων του, και οι μπλε γραμμές τα γονίδια της ομάδας. Προέρχεται από την πρώτη εφαρμογή του αλγορίθμου στο dataset του καρκίνου του μαστού και εμφανίζει την συμπεριφορά των γονιδίων του 12ου bicluster. Τα βασικά σημεία που αποτέλεσαν πηγή προβληματίσμού ήταν τα εξής:

Πρόβλημα πρώτο

Ενώ τα περισσότερα από τα γονίδια του bicluster εμφανίζουν κατά μήκος των σειρών ομοιόμορφη συμπεριφορά, υπάρχουν σημεία (ροζ κύκλοι) που παρουσιάζονται εμφανείς αποκλίσεις των τιμών σε συγκεκριμένες κυτταρικές σειρές.

Πρόβλημα δεύτερο

Οι τιμές των γονιδιακών εκφράσεων (άξονας y) καλύπτουν μεγάλο εύρος (18-4). Στην επεξεργασία των αποτελεσμάτων θα ήταν προτιμότερη η μελετή γονίδιων που όχι μόνο εμφανίζουν τις ίδιες διακυμάνσεις κατά μήκος των σειρών, αλλά και οι τιμές της γονιδιακής τους έκφρασης είναι σχετικά κοντινές μεταξύ τους.

Τέθηκε επομένως ως στόχος η υλοποίηση μίας καινοτομίας με σκοπό την επίλυση των ζητημάτων που τέθηκαν και την βελτίωση των αποτελεσμάτων διευκολύνοντας την περαιτέρω συζήτηση και επεξεργασία τους.

3.3 Χρησιμότητα και υλοποίηση προτεινόμενων βελτιώσεων

Στόχος αυτής της καινοτομίας είναι για κάθε bicluster να παραμείνουν τα γονίδια εκείνα που θα έχουν την ίδια συμπεριφορά μεταξύ τους σε κάθε κυτταρική σειρά της ομάδας. Συγκεκριμένα σε κάθε μία από τις σειρές να έχουν όλα τα γονίδια υψηλή ή χαμηλή έκφραση με μικρές αποκλίσεις μεταξύ των τιμών.

Παρατηρώντας τις ήδη υπάρχουσες γραφικές αναζητήθηκε σε ποια διαστήματα συσσωρεύονται τα γονίδια και παρατηρήθηκε, ότι ανά 2-3 μονάδες δημιουργούνται ομάδες, οι οποίες θα μπορούσαν η κάθε μία να αποτελεί ένα ξεχωριστό bicluster. Θεωρήθηκαν λοιπόν σαν επιλογές για κάθε κυτταρική σειρά η τιμή έκφρασης κάθε γονιδίου να απέχει το πολύ από τη μέση τιμή των τιμών της σειράς 2 ή 3 μονάδες. Έτσι μειώνονται και οι αποκλίσεις και το αντίστοιχο εύρος (στον άξονα y).

Εφαρμόστηκαν λοιπόν οι ακόλουθες ενέργειες:

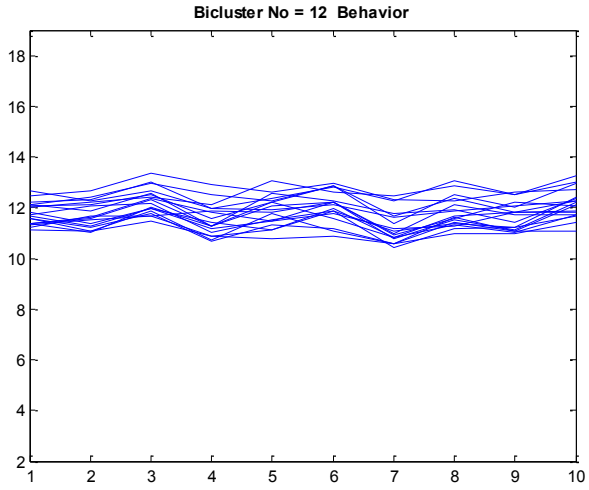
Για κάθε κυτταρική σειρά-στήλη *j* του bicluster υπολογίστηκε η μέση τιμή της:

$$\alpha_{Ij} = \frac{1}{|I|}\sum_{i \in I} a_{ij}.$$

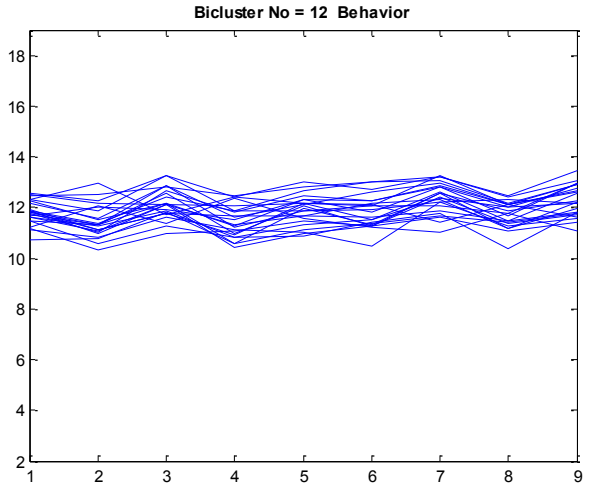
Για κάθε γονίδιο αν η τιμή του σε κάποια κυτταρική σειρά-στήλη απέκλινε περισσότερες από k μονάδες (όπου κ=2 ή 3) από τη μέση τιμή αφαιρούνταν από το συγκεκριμένο bicluster.

Για κάθε $i \in I$ *πρέπει:*

$$a_{iJ} - k \leq a_{ij} \leq a_{iJ} + k.$$



Εικόνα 18. Bicluster για k=2.



Εικόνα 19. Bicluster για k=3.

Στις νέες εικόνες έχουν φιλτραριστεί οι αποκλίσεις που συναντήθηκαν προηγουμένως (ροζ κύκλοι), και το εύρος τιμών έχει περιοριστεί στις μονάδες που είχαν οριστεί. Είναι σαφές ότι τα γονίδια που αποκλείστηκαν, αν πληρούν τις προϋποθέσεις μπορούν να συμπεριληφθούν σε επόμενο bicluster. Όσον αφορά την παράμετρο k, στη γραφική της τιμής 3 υπάρχει περισσότερη χρήσιμη πληροφορία χωρίς να επηρεάζεται το αποτέλεσμα, επομένως επιλέχθηκε αυτή για την διεξαγωγή των γραφικών.

Τελευταίο ζήτημα για την τελική επιλογή των αποτελεσμάτων αποτέλεσε η επιλογή των παραμέτρων δ και α, καθώς και ο αριθμός των biclusters (n) που θα είχε σαν έξοδο ο αλγόριθμος Cheng and Church, και θα έδιναν τα βέλτιστα αποτελέσματα.

- ***Για την παράμετρο α***

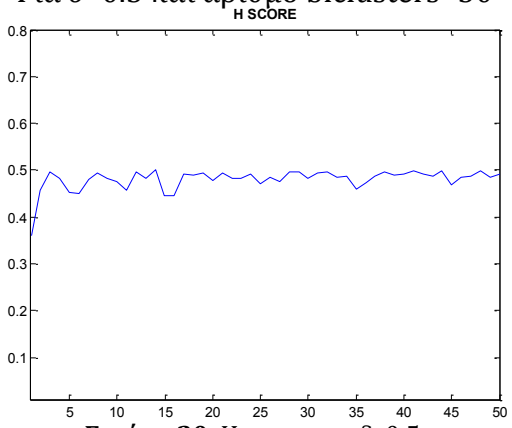
Μελετώντας τη βιβλιογραφία η τιμή ***α=1.2*** είναι ενδεικτική. Εφαρμόστηκε ο αλγόριθμος με μικρές αλλαγές του *α* της τάξεως $1.1 \leq \alpha \leq 1.9$ και το αποτέλεσμα δεν επηρεάστηκε άμεσα. Τιμές όμως πολύ μεγαλύτερες του 1.2 κόστισαν αρκετά στο χρόνο απόδοσης των αποτελεσμάτων. Έτσι αποφασίστηκε η αρχική επιλογή όπου ***α=1.2.***

- ***Για την παράμετρο δ***

Όπως αναφέρθηκε, στόχος του αλγορίθμου CC είναι η εύρεση μεγάλων biclusters με χαμηλό *mean squared residue score (H)* και συγκεκριμένα μικρότερο από ένα συγκεκριμένο threshold $\delta \geq 0$. Όσο περισσότερο προσεγγίζει το *δ* το μηδέν τόσο πιο ακριβή και βέλτιστα είναι τα εξαγόμενα biclusters.

Για το λόγο αυτό, οι τιμές του *δ* περιορίστηκαν σε μικρότερες της μονάδας που θα παρείχαν έναν αριθμό αποτελεσμάτων ικανοποιητικό για μελέτη και επεξεργασία. Ορίστηκε επομένως *δ*=0.5, 0.3, 0.1, 0.08, 0.07 και εφαρμόστηκε ο βελτιωμένος αλγόριθμος στο dataset του καρκίνου του μαστού για τα 50 πρώτα biclusters. Κύριος παράγοντας για την επιλογή του *δ* θα ήταν όλα τα biclusters που θα προέκυπταν να είχαν εξίσου ένα καλό H score:

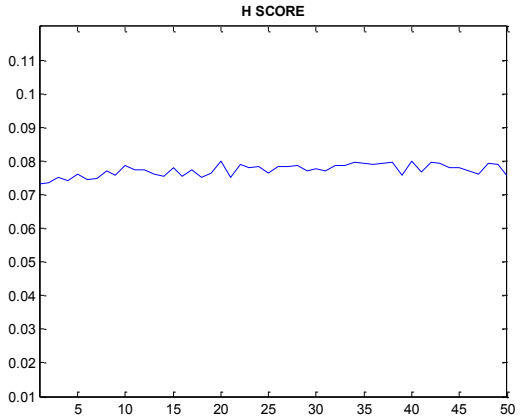
Για *δ*=0.5 και αριθμό biclusters=50



Εικόνα 20. H-score για *δ*=0.5.

Παρατηρείται ότι τα πρώτα 4 biclusters εμφανίζουν πολύ χαμηλότερο σκορ, άρα πολύ πιο καλό σε σχέση με τα υπόλοιπα 46, και ο αριθμός των συνολικά ομαδοποιημένων γονιδίων αγγίζει τα 7810, γεγονός που αποτελεί εμπόδιο στην επεξεργασία των αποτελεσμάτων και στην επιλογή τελικά συγκεκριμένων γονιδιακών δεικτών. Ομοίως απορρίφθηκαν και οι τιμές δ=0,3 και 0,1.

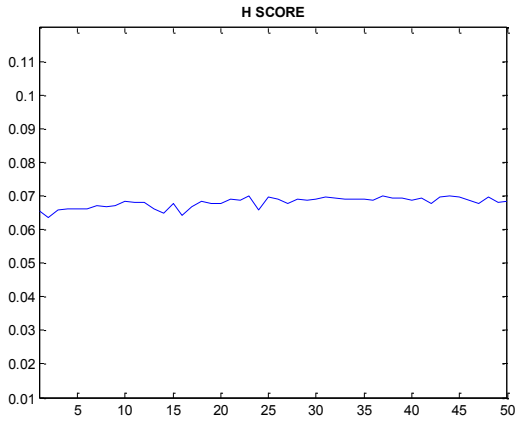
Για την τιμή δ=0.08 και αριθμό biclusters=50



Εικόνα 21. H-score για δ=0.08.

Όλα τα biclusters εμφανίζουν εξίσου καλό H-score και ο αριθμός των συνολικά ομαδοποιημένων γονιδίων είναι: 1085 με μεγαλύτερο το πρώτο bicluster με συνολικά γονίδια γύρω στα 100.

Για τιμές μικρότερες του 0.08 ενώ υπάρχει επίσης αρμονία στο H-score

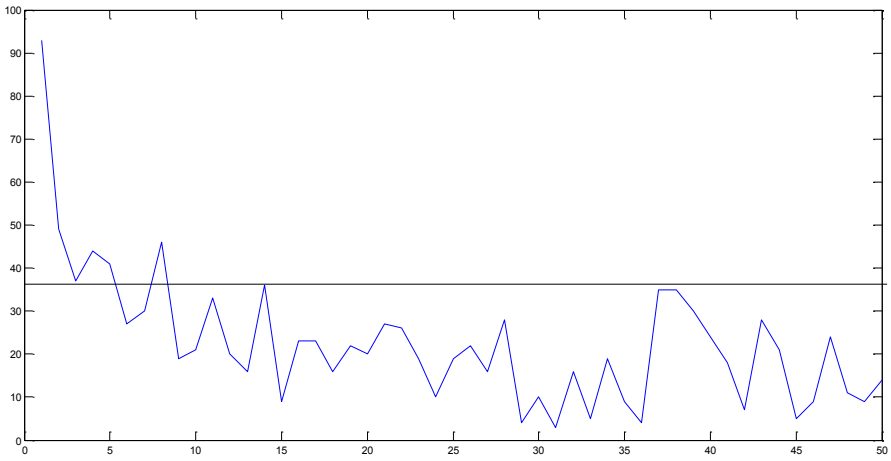


Εικόνα 22. H-score για δ=0.07

μεταξύ των biclusters,ο αριθμός των ομαδοποιημένων γονιδίων είναι αρκετά μικρός με πρώτο bicluster με αριθμό γονιδίων μικρότερο του 100.

Επομένως, είναι φανερό ότι η καλύτερη τιμή είναι η δ=0.08 που παρέχει σαν αποτέλεσμα καλό σκορ και παράλληλα ικανό αριθμό δειγμάτων προς επεξεργασία.

- **Για τον αριθμό n των biclusters**

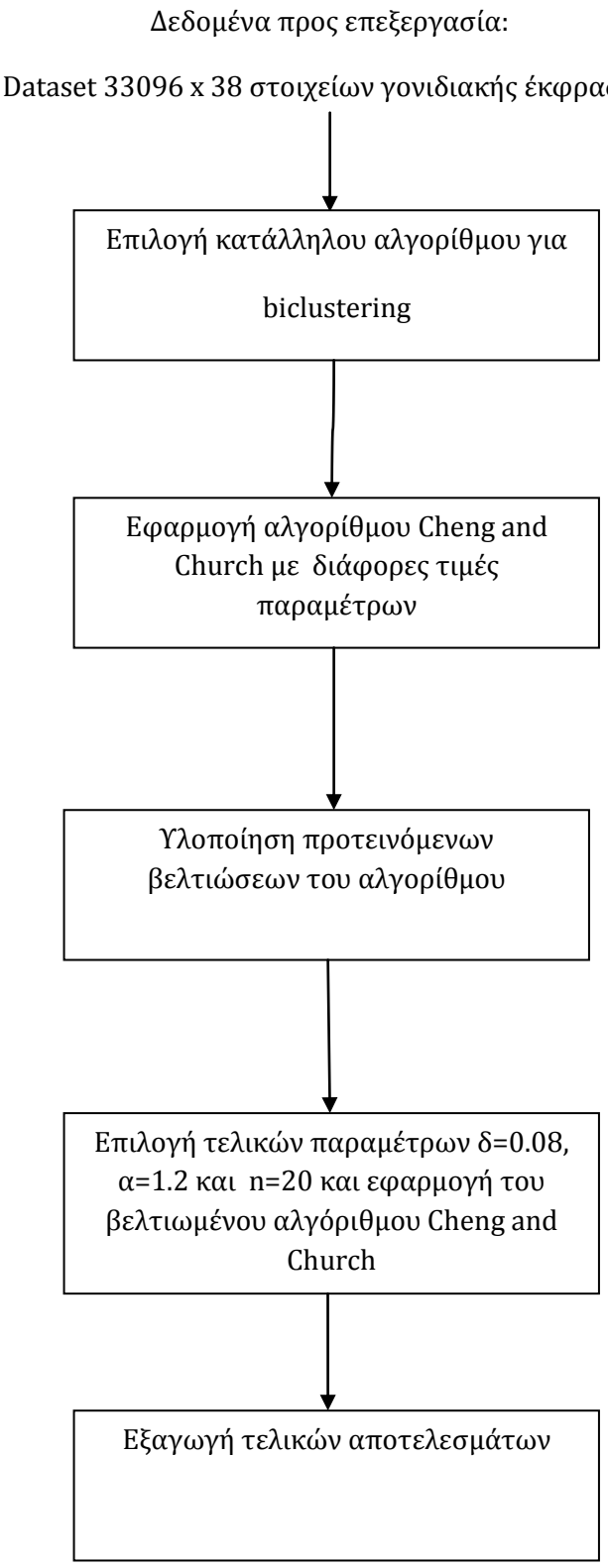


Εικόνα 23. Γραφική που απεικονίζει το σύνολο των ομαδοποιημένων γονιδίων για τα 50 πρώτα biclusters.

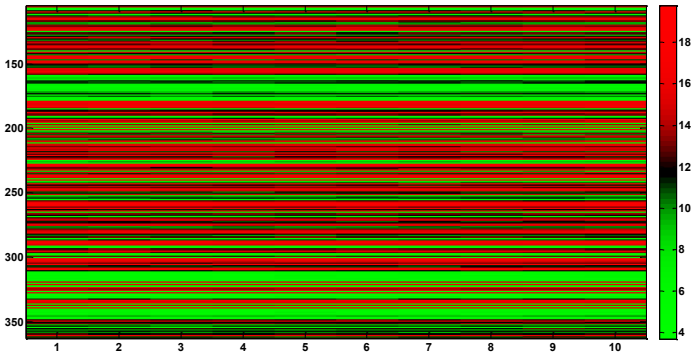
Το παραπάνω διάγραμμα απεικονίζει το σύνολο των ομαδοποιημένων γονιδίων για τα 50 πρώτα biclusters. Παρατηρείται ότι από το 15ο bicluster και μετά, ο αριθμός των γονιδίων είναι σχετικά μικρός.

Επομένως, επιλέχθηκαν για μελέτη για κάθε μία εφαρμογή του αλγορίθμου τα πρώτα 20 biclusters και η συμπεριφορά τους.

Σύνοψη προτεινόμενης μεθοδολογίας

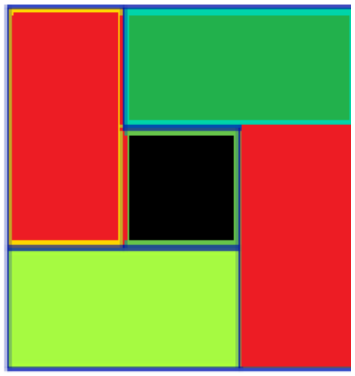


Αρχικό dataset

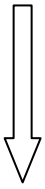


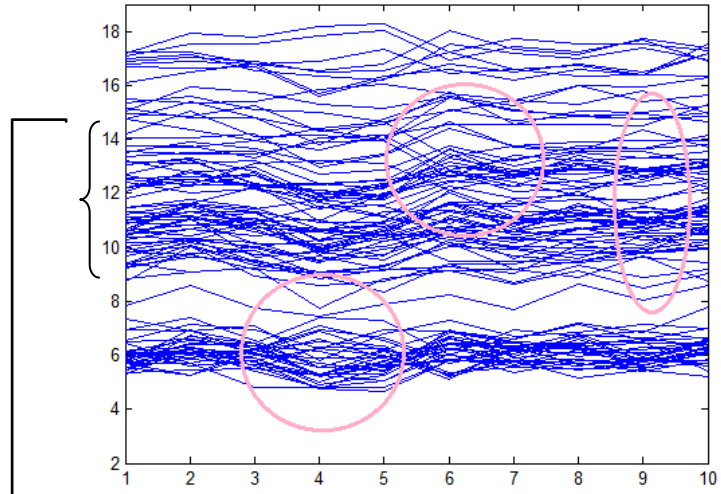
Παρατηρείται ότι τα δεδομένα εμφανίζονται αρκετά συμπαγή σε επίπεδο γονιδίων και καταστάσεων και παρατηρούνται σημεία ομοιομορφίας που χρήζουν ομαδοποίησης για τη σωστή ερμηνεία τους.

Εφαρμογή αλγορίθμου Cheng and Church

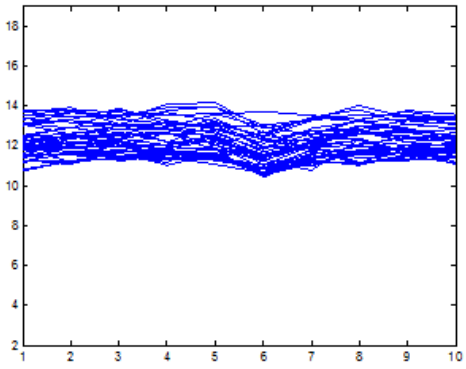


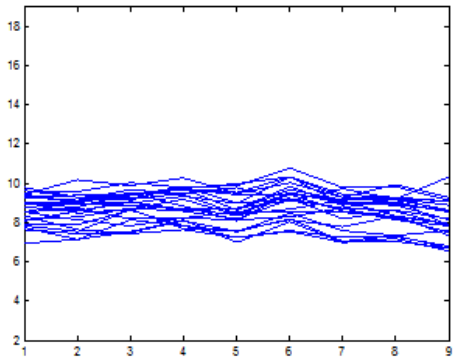
Τα εξαγόμενα biclusters δεν εμφανίζουν επικάλυψη και δεν έχουν συγκεκριμένη μορφή. Ανήκουν λοιπόν στην κατηγορία μη επικαλυπτόμενα-μη αποκλειστικά. Κάθε bicluster πρίν τις βελτιώσεις έχει την ακόλουθη μορφή συμπεριφοράς:





Εφαρμογή βελτιωμένου αλγορίθμου Chung and Church.





Στις παραπάνω εικόνες φαίνεται το αποτέλεσμα της καινοτομίας που χρησιμοποιήθηκε μειώνοντας τις αποκλίσεις και το εύρος και δημιουργώντας πιο μικρές και ομοιόμορφες ομάδες. Οι συνολικές ομάδες κάθε εφαρμογής του βελτιωμένου αλγορίθμου συγκεντρώνονται σε γραφήματα της παρακάτω μορφής.

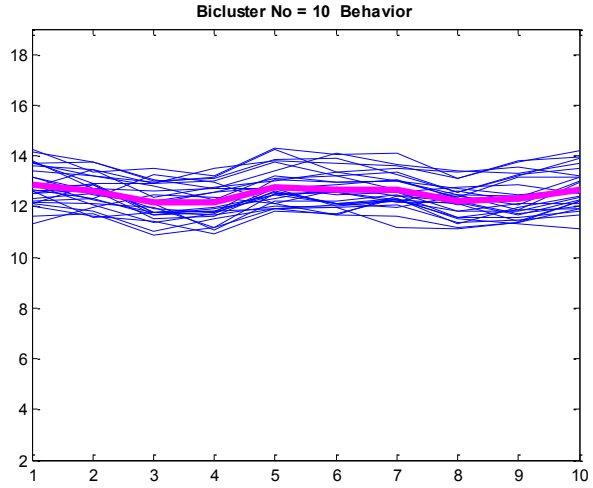
ΚΕΦΑΛΑΙΟ 4
ΑΠΟΤΕΛΕΣΜΑΤΑ

Στις παρακάτω γραφικές παρουσιάζονται συνοπτικά τα τελικά αποτελέσματα, εφαρμόζοντας τον βελτιωμένο αλγόριθμο Cheng and Church με παραμέτρους δ=0.08, α=1.2 και αριθμό biclusters=20. Η μεθοδολογία που επιλέχθηκε, εφαρμόστηκε σε 5 αρχεία, ένα για κάθε τύπο καρκίνου και ένα συγκεντρωτικό αρχείο με όλα τα γονίδια και τις σειρές μαζί, εξάγοντας τα παρακάτω αποτελέσματα:

- 5 πίνακες με τις μέσες τιμές των συμπεριφορών των 20 biclusters για κάθε μία από τις 5 εφαρμογές.
- 3 ενδεικτικές γραφικές από την εφαρμογή του αλγορίθμου στο μεγαλύτερο bicluster του καρκίνου των ωοθηκών.
- 5 συγκεντρωτικοί πίνακες με τα γονίδια και τις σειρές για τις 5 εφαρμογές του αλγορίθμου.
- Ένας τελικός πίνακας που παρουσιάζει σε ποιες από τις 5 εφαρμογές εμφανίζεται κάθε γονίδιο (που ανήκει σε τουλάχιστον ένα bicluster).

4.1 Μέσες τιμές των συμπεριφορών των 20 biclusters κατά μήκος των σειρών

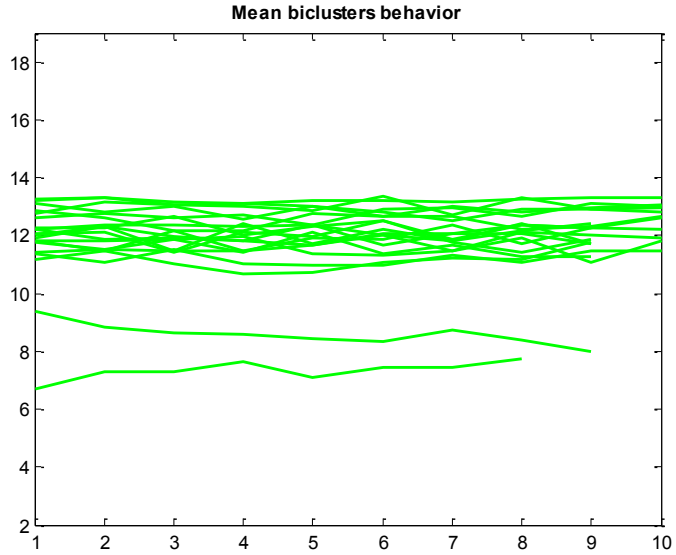
Η παρακάτω εικόνα είναι ένα δείγμα από τα αποτελέσματα για κάθε ένα από τα biclusters. Απεικονίζει τη συμπεριφορά (γονιδιακή έκφραση) των γονιδίων κατά μήκος των σειρών που ανήκουν σε αυτή την ομάδα.



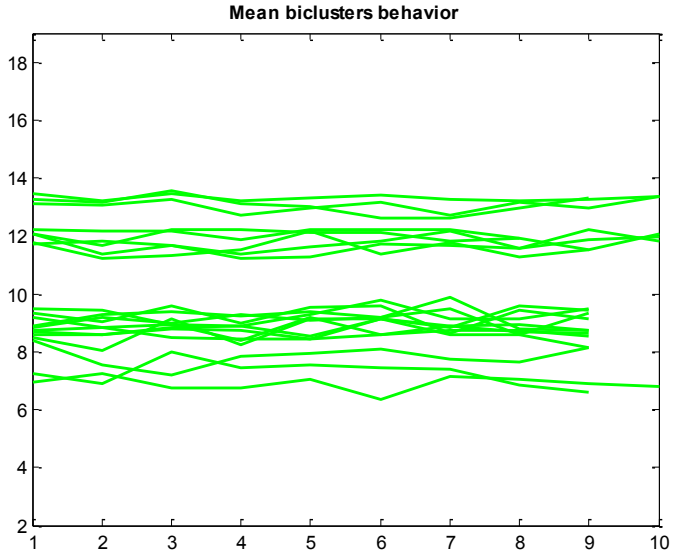
Εικόνα 24. Συμπεριφορά γονιδίων κατά μήκος των σειρών.
Με ροζ απεικονίζεται η μέση συμπεριφορά.

Αναλυτικά, ο άξονας x απεικονίζει τον αριθμό των σειρών του bicluster, ο άξονας y τις τιμές έκφρασης των γονιδίων του, και οι μπλε γραμμές τα γονίδια της ομάδας. Λόγω του μεγάλου αριθμού των καμπυλών, για καλύτερη κωδικοποίηση επιλέχθηκε η μέση τιμή των καμπυλών (η καμπύλη με τη ροζ γραμμή) για κάθε bicluster. Αναλυτικά, παρατίθενται τα γραφήματα συμπεριφοράς όλων των ομάδων στο παράρτημα Α.

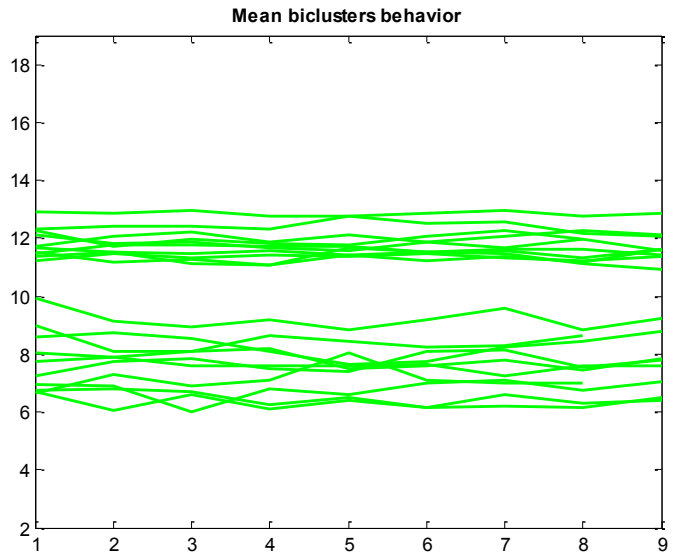
Στις γραφικές που ακολουθούν εμφανίζονται με πράσινο χρώμα οι 20 καμπύλες μέσων τιμών που συγκεντρώθηκαν για κάθε μία από τις 5 εφαρμογές του αλγορίθμου.



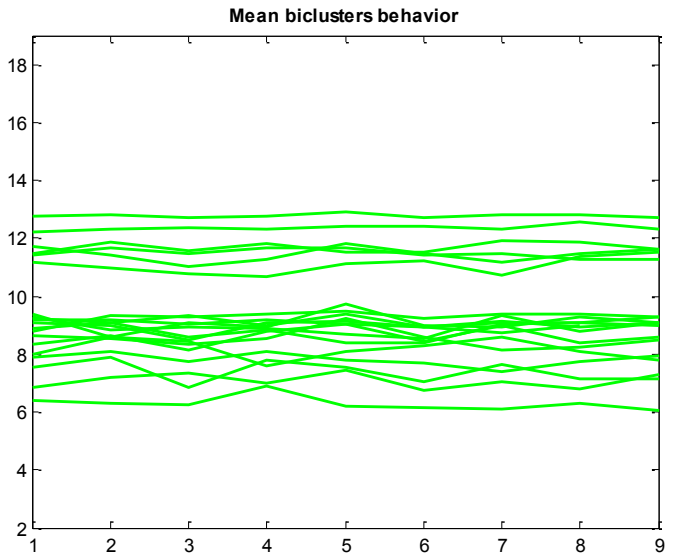
Εικόνα 25. Μέσες τιμές για τον καρκίνο του μαστού.



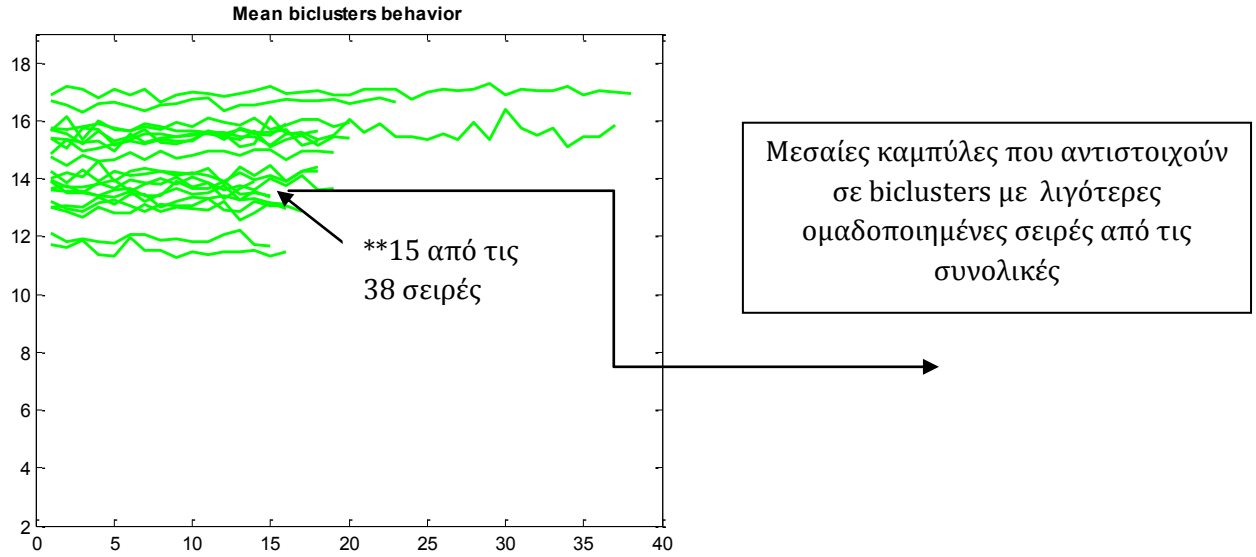
Εικόνα 26. Μέσες τιμές για τον καρκίνο των ωοθηκών.



Εικόνα 27. Μέσες τιμές για τον καρκίνο του ενδομητρίου.



Εικόνα 28. Μέσες τιμές για τον καρκίνο του τραχήλου.



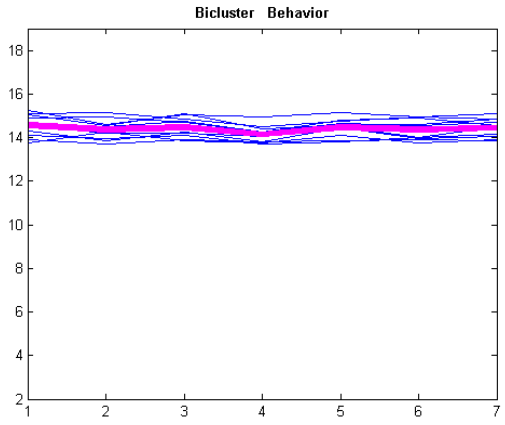
Εικόνα 29. Μέσες τιμές για ομαδοποίηση των 4 τύπων μαζί.

Όπως αναφέρθηκε, σε μια μέθοδο biclustering γίνεται ταυτόχρονη ομαδοποίηση γραμμών και στηλών. Γι΄ αυτό το λόγο, σε κάποια biclusters, τα γονίδια ομαδοποιούνται κάτω από συγκεκριμένες καταστάσεις. Σε αυτό οφείλεται το ότι στις παραπάνω εικόνες κάποιες από τις καμπύλες σταματούν νωρίτερα από τις άλλες. Τα biclusters δηλαδή που αντιστοιχούν σ'αυτές περιλαμβάνουν μικρότερο αριθμό κυτταρικών σειρών από το συνολικό.

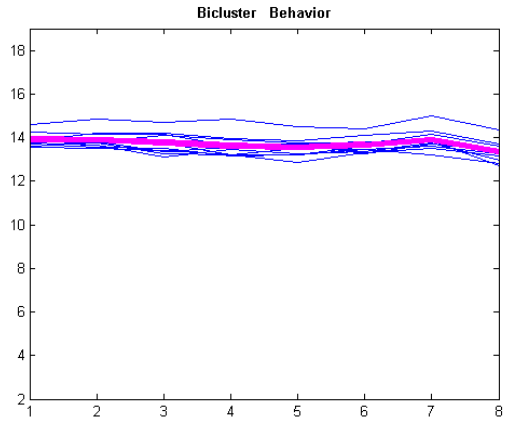
Παρατηρώντας τις 5 παραπάνω γραφικές συμπεραίνουμε ότι υπάρχουν στατιστικά συγκεκριμένες περιοχές εύρους που εμφανίζονται πιο πυκνές, συγκεντρώνοντας τις περισσότερες μεσαίες καμπύλες. Αυτές είναι οι περιοχές 11-13 και 7-9, γεγονός που εμφανίζεται σχεδόν σε όλες τις ομαδοποιήσεις και ιδιαίτερα στην γραφική που αφορά τον καρκίνο του μαστού όπου 18 από τις 20 καμπύλες ανήκουν στην πρώτη περιοχή. Επιπλέον γίνεται εμφανής η διαφοροποίηση των 4 επιμέρους ομαδοποιήσεων με την ολική ομαδοποίηση, στην οποία σε αντίθεση με τις άλλες παρατηρείται ότι υπάρχουν αρκετά biclusters στα οποία λείπει ένας μεγάλος αριθμός κυτταρικών σειρών με χαρακτηριστικό παράδειγμα ενός bicluster στο οποίο έχουν ομαδοποιηθεί οι 15 από τις 38 σειρές**.

Μελετώντας τις γραφικές που έχουν παρατεθεί στο παράρτημα Α παρατηρούμε ότι τα πρώτα biclusters από κάθε ομαδοποίηση είναι αρκετά πυκνά και περιλαμβάνουν μεγάλο αριθμό γονιδίων. Θεωρήθηκε επομένως χρήσιμο να ελεγχθεί αν εφαρμόζοντας τον βελτιωμένο αλγόριθμο μόνο στα γονίδια και στις σειρές που ανήκουν σε ένα από αυτά τα biclusters, ο αλγόριθμος θα μας έδινε εξίσου σημαντικό αποτέλεσμα τόσο στατιστικά όσο και βιολογικά.

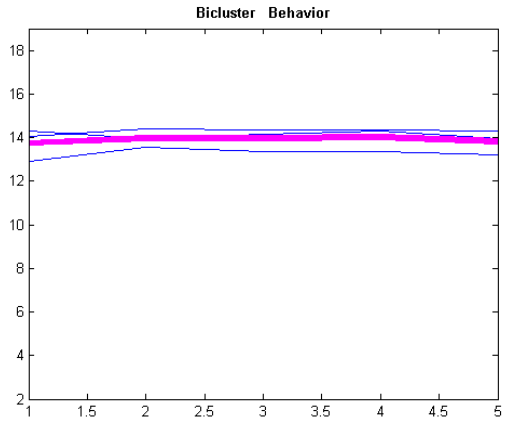
Έτσι, ο βελτιωμένος αλγόριθμος εφαρμόστηκε ενδεικτικά στο πρώτο bicluster του καρκίνου των ωοθηκών και τα αποτελέσματα ήταν τα εξής:



Εικόνα 30. Συμπεριφορά sub-bicluster 1.



Εικόνα 31. Συμπεριφορά sub-bicluster 2.



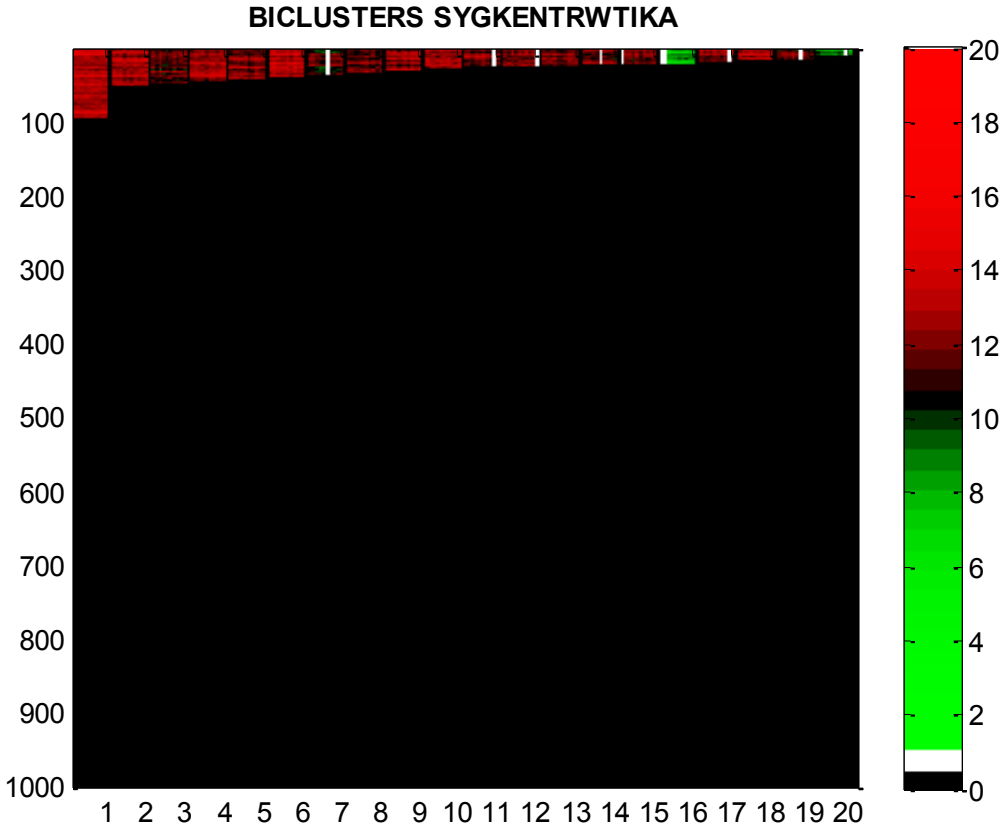
Εικόνα 32. Συμπεριφορά sub-bicluster 10.

Οι εικόνες 30, 31 και 32 αποτελούν ενδεικτικά παραδείγματα των sub-biclusters που δημιουγήθηκαν από την εφαρμογή του αλγορίθμου στο πρώτο bicluster του καρκίνου των ωοθηκών. Παρατηρούμε ότι όλα τα στοιχεία του μεγάλου bicluster, όπως η ομοιομορφία στη συμπεριφορά των γονιδίων και το μικρό εύρος εμφανίζονται και στα sub-biclusters.

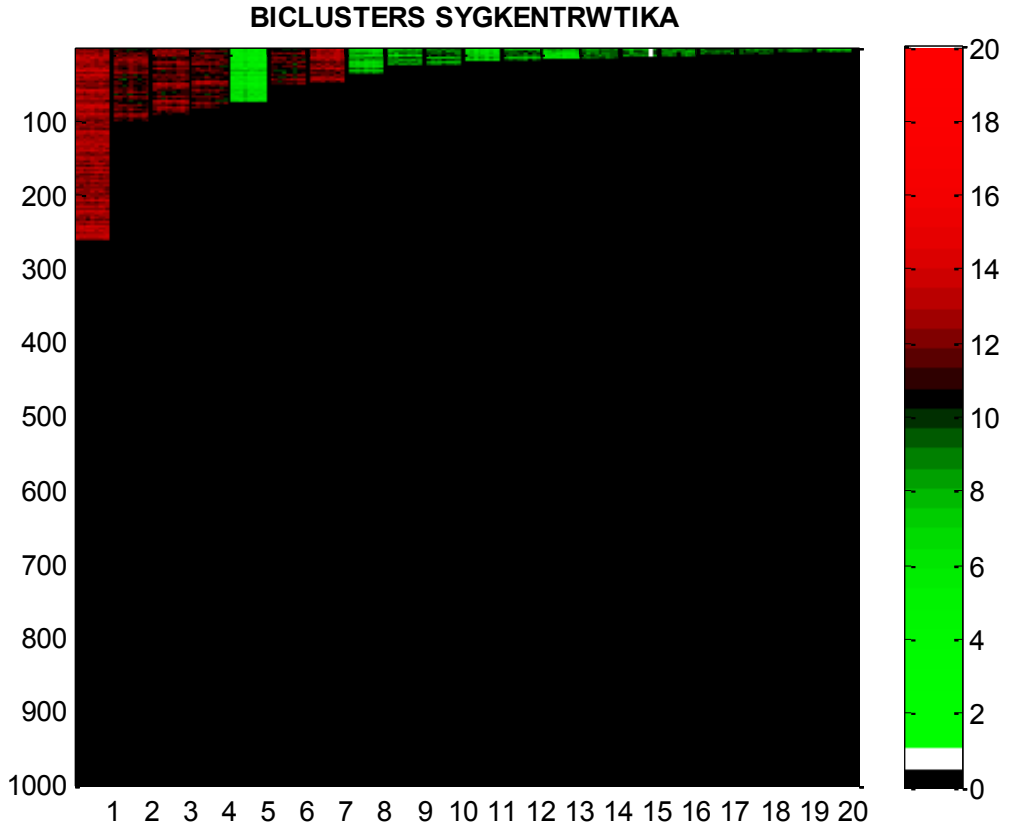
4.2 Συγκεντρωτικοί πίνακες

Στους παρακάτω πίνακες διακρίνονται συγκεντρωμένα και τα 20 biclusters για κάθε μία από τις 5 εφαρμογές του αλγορίθμου.

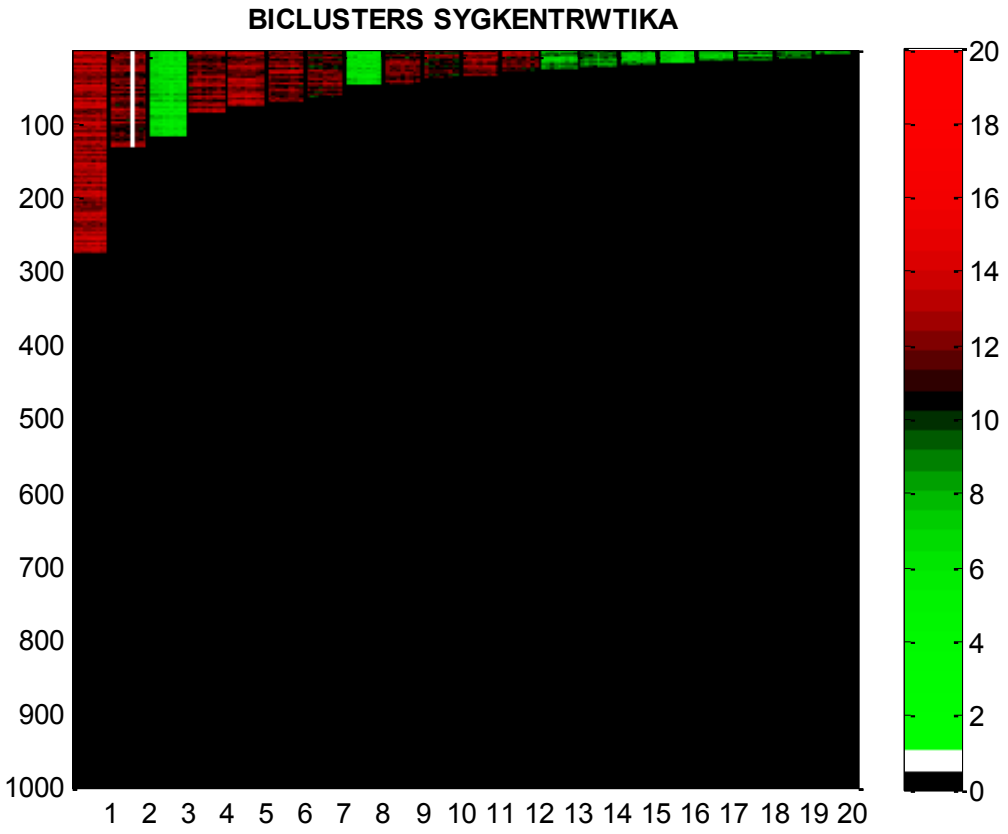
- Ο άξονας x απεικονίζει τον αριθμό των biclusters.
- Ο άξονας y απεικονίζει τον αριθμό των γονιδίων.
- Οι τιμές έκφρασης έχουν αντικατασταθεί με μικρά pixels χρώματος. Αυτές με χαμηλά επίπεδα έκφρασης καταλαμβάνουν τις πράσινες τιμές και με υψηλά επίπεδα τις κόκκινες.
- Για κάθε bicluster παρατηρείται η συμπεριφορά του σε κάθε ομαδοποιημένη κυτταρική σειρά. Με άσπρη γραμμή εμφανίζονται οι κυτταρικές σειρές που για το εκάστοτε bicluster δεν συμμετέχουν στην ομαδοποίηση.



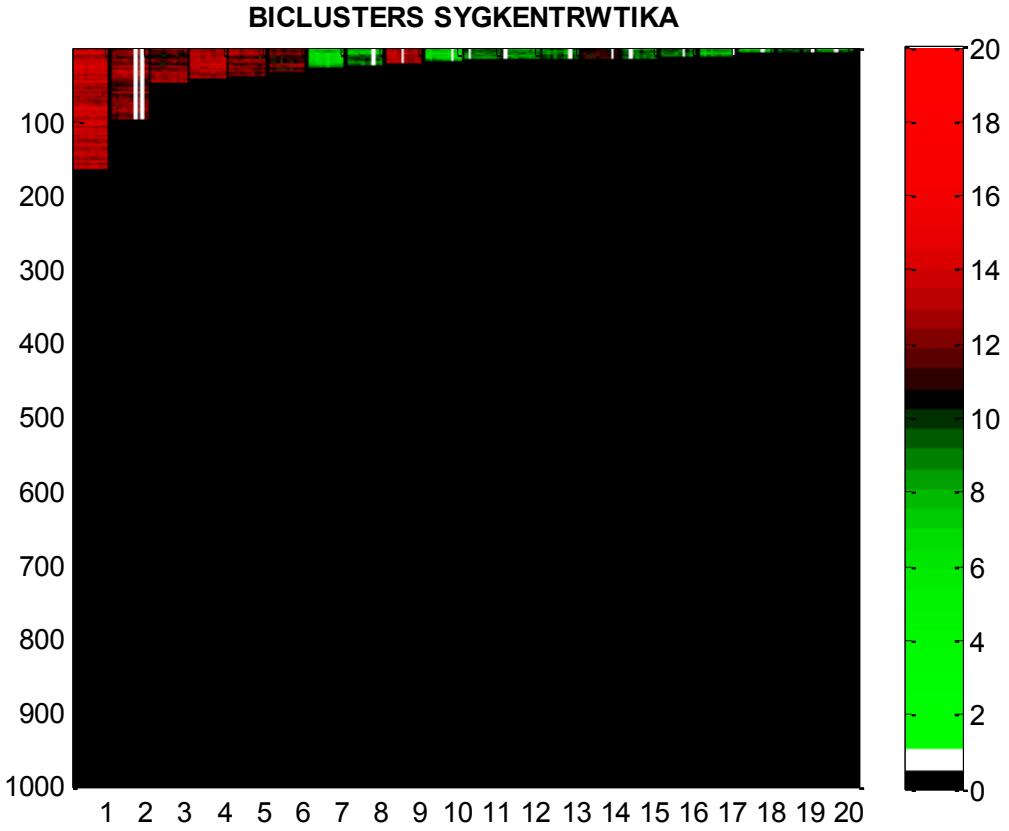
Εικόνα 33. Biclusters καρκίνου του μαστού.



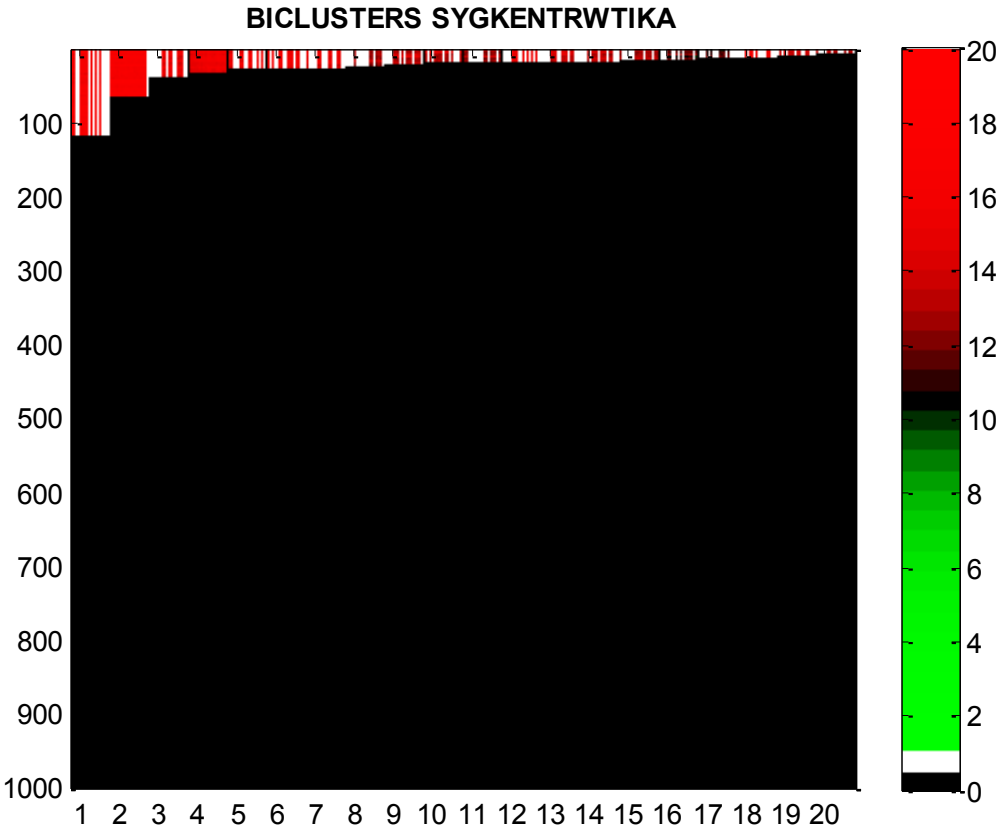
Εικόνα 34. Biclusters καρκίνου του τραχήλου της μήτρας.



Εικόνα 35. Biclusters καρκίνου του ενδομητρίου.



Εικόνα 36. Biclusters καρκίνου των ωοθηκών.



Εικόνα 37. Biclusters ολικής ομαδοποίησης.

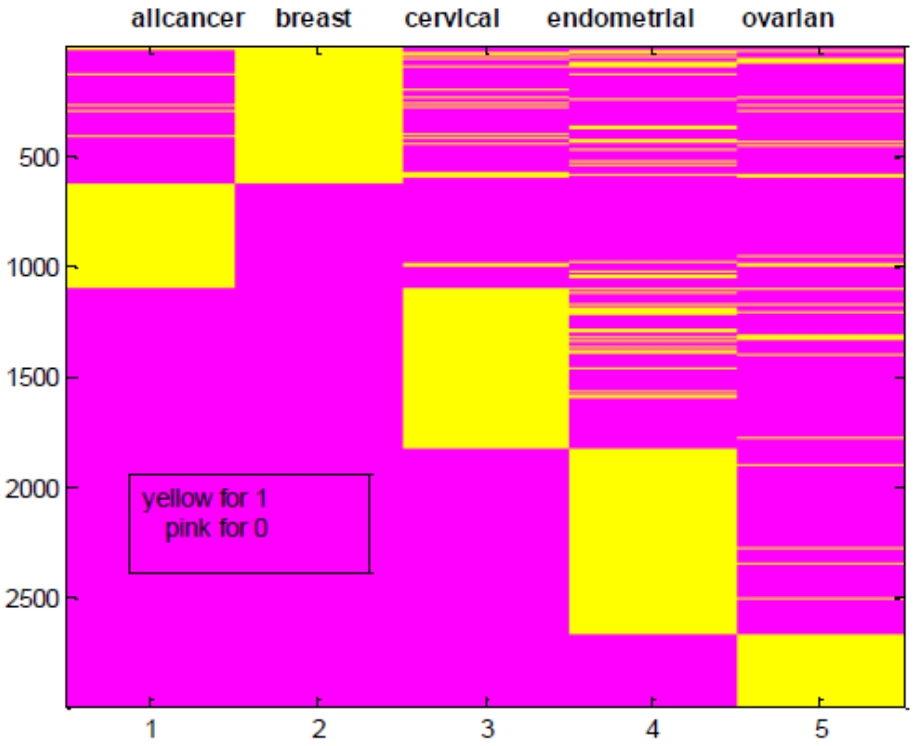
Μελετώντας τους συγκεντρωτικούς πίνακες, συμπεραίνεται ότι τα ομαδοποιημένα biclusters εμφανίζουν ομοιομορφία όσον αφορά τη χρωματική διάταξη, γεγονός που φανερώνει ότι οι ομάδες που δημιουργήθηκαν περιλαμβάνουν γονίδια με παρόμοια προφίλ έκφρασης είτε υψηλής είτε χαμηλής και επομένως κρίνεται χρήσιμη και σωστή η μείωση του εύρους των τιμών που προτάθηκε στη μεθοδολογία.

Ιδιαίτερο ενδιαφέρον παρουσιάζει ο τελευταίος συγκεντρωτικός πίνακας που αφορά την ολική ομαδοποίηση. Περιλαμβάνει τις περισσότερες άσπρες στήλες σε αντίθεση με τους υπόλοιπους που εμφανίζουν ελάχιστες (στον πίνακα του καρκίνου του τραχήλου της μήτρας λείπει μόλις μία κυτταρική σειρά από το 15ο bicluster). Παρ' όλες τις διαφορές μεταξύ των 4 καρκινικών τύπων παρατηρούμε ότι τα biclusters της ολικής ομαδοποίησης παρουσιάζουν κι αυτά εξίσου ικανοποιητική ομοιομορφία και συνεκτικότητα στη συμπεριφορά των γονιδίων κατά μήκος των σειρών.

Η παρακάτω εικόνα περιλαμβάνει στον άξονα y όλα τα γονίδια που ομαδοποιήθηκαν στις 5 εφαρμογές μας, και στον άξονα x αυτές τις 5 εφαρμογές.

Σε κάθε στήλη (allcancer, breast, cervical, endometrial, ovarian) με κίτρινο χρωματίζεται το γονίδιο εφόσον ανήκει σε κάποιο από τα 20 biclusters της ομαδοποίησης και με ροζ αν όχι.

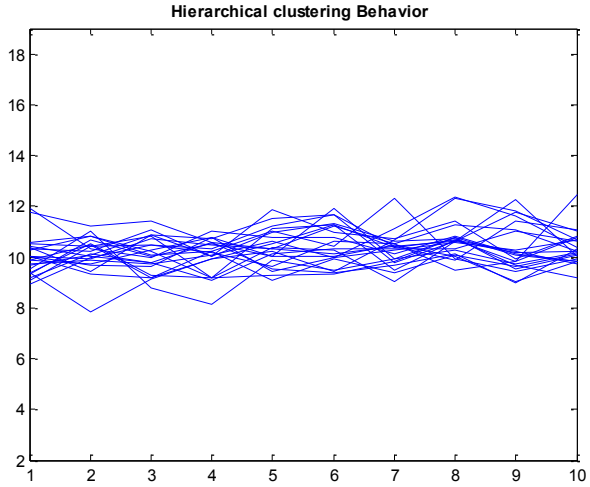
Σημειώνεται ότι ένα μόνο γονίδιο εμφανίζεται και στις 5 ομαδοποιήσεις και 27 γονίδια στις 4 από τις 5 με εξαίρεση την ολική ομαδοποίηση (allcancer).



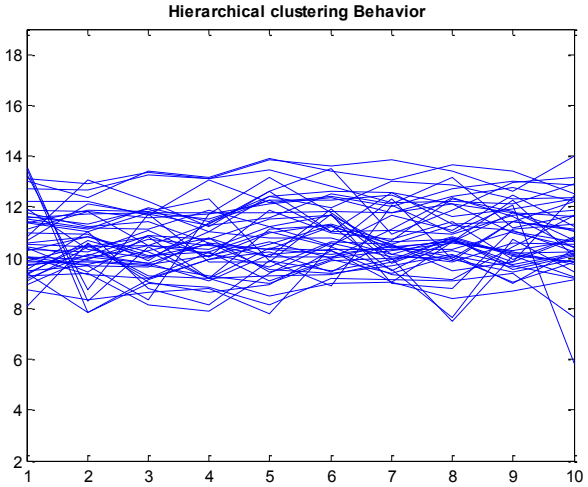
Εικόνα 38. Συγκεντρωτικός πίνακας που απεικονίζει όλα τα ομαδοποιημένα γονίδια και αναπαριστά με κίτρινο σε ποιες ομαδοποιήσεις ανήκουν και με ροζ σε ποιες όχι.

4.3 Εφαρμογή ιεραρχικής ομαδοποίησης (Hierarchical Clustering)

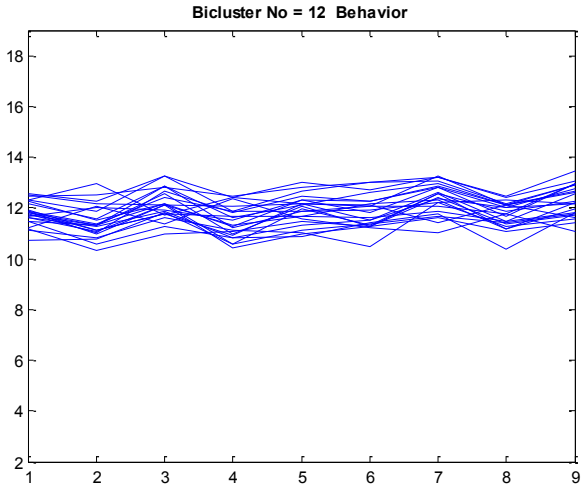
Επιθυμώντας να συγκρίνουμε τα αποτελέσματα του biclustering αλγορίθμου μας με μία τεχνική clustering για να παρατηρηθούν ομοιότητες και διαφορές μεταξύ των δύο μεθόδων εφαρμόστηκε στο dataset για τον καρκίνο του μαστού hierarchical clustering. Παρατίθενται δύο ενδεικτικά clusters από τα συνολικά αποτελέσματα του αλγορίθμου hierarchical σε σύγκριση με ένα bicluster του αλγορίθμου biclustering, και ένας συγκεντρωτικός πίνακας των πρώτων 15 clusters της ομαδοποίησης.



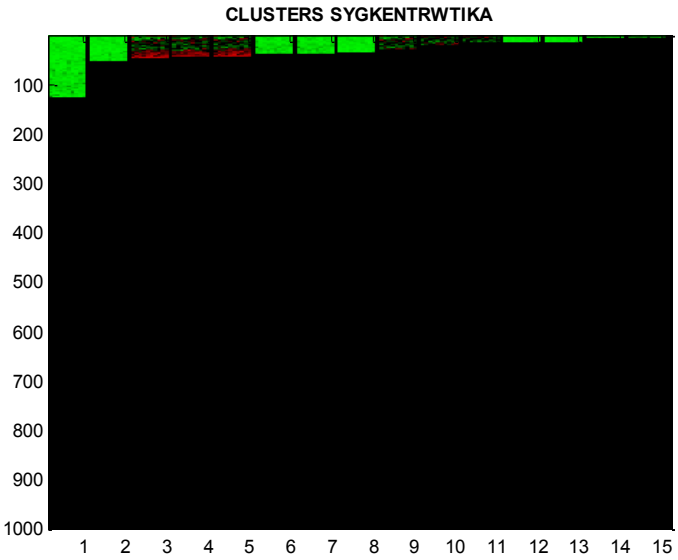
Εικόνα 39. Συμπεριφορά ενός cluster με εφαρμογή Hierarchical clustering.



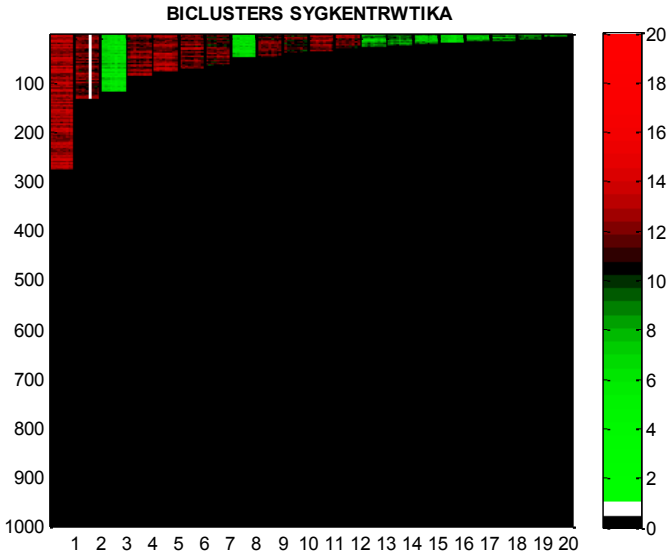
Εικόνα 40. Συμπεριφορά ενός cluster με εφαρμογή Hierarchical clustering.



Εικόνα 41. Συμπεριφορά μιας υποομάδας με εφαρμογή biclustering.



Εικόνα 42. Συγκεντρωτικός πίνακας με τα πρώτα 15 clusters έπειτα από την εφαρμογή του Hierarchical clustering.



Εικόνα 43. Συγκεντρωτικός πίνακας με τα 20 biclusters έπειτα από την εφαρμογή του αλγορίθμου Cheng and Church.

Μεταξύ των εικόνων 36, 37 και 38 φαίνεται ότι η biclustering τεχνική πετυχαίνει καλύτερη μείωση του θορύβου και η συμπεριφορά των γονιδίων παρουσιάζει μεγαλύτερη ομοιομορφία σε σχέση με τα αποτελέσματα της clustering τεχνικής. Αυτό αποδεικνύεται και από την εικόνα 39, που σε αντίθεση με την εικόνα 40 συναντώνται clusters με ανομοιόμορφη κατανομή χρωμάτων, γεγονός που φανερώνει μεγάλη απόκλιση μεταξύ των προφίλ έκφρασης των γονιδίων που ανήκουν σε αυτά.

4.3.1 Σύγκριση αποτελεσμάτων

Παρατηρώντας το αποτέλεσμα του αλγορίθμου Cheng and Church και του Hierarchical clustering συμπεραίνεται ότι επιτεύχθηκε η μειώση του θορύβου (γονίδια που δεν συνάδουν με τα υπόλοιπα) σε αντίθεση με την τεχνική hierarchical. Επιπλέον, στην biclustering τεχνική η ομαδοποίηση δεν γίνεται υποχρεωτικά για όλες τις κυτταρικές σειρές (bicluster με 9 από τις 10 σειρές του μαστού), πράγμα που συμβαίνει με την ιεραρχική ομαδοποίηση (όλα τα clusters και με τις 10 σειρές).

Έτσι ακολουθούν οι λόγοι για τους οποίους προτιμάται μια biclustering από μία clustering τεχνική:

1. Το hierarchical clustering παρείχε σημαντικές πληροφορίες όσον αφορά τα γονίδια ομαδοποιώντας τα όμως πάντα στο σύνολο των κυτταρικών σειρών. Δηλαδή, κάθε cluster όπως φαίνεται στις εικόνες συμπεριλαμβάνει και τις 10 κυτταρικές σειρές του καρκίνου του μαστού.

2. Η πληροφορία που αφορά γονίδια σε ένα υποσύνολο σειρών χάνεται και γονίδια που συμμετέχουν σε πάνω από μία ενέργειες δεν μπορούν να ομαδοποιηθούν ξανά πάνω από μία φορά.

Συμπερασματικά η biclustering μέθοδος υπερτερεί στο ότι:

- Μειώνει το θόρυβο (τιμές γονιδίων που αποκλίνουν) και παρατηρείται ομοιομορφία στις συμπεριφορές των γονιδίων εντός των biclusters.
- Μπορεί να ομαδοποιήσει ένα σετ γονιδίων που συμμετέχουν σε μία ενδιαφέρουσα κυτταρική διαδικασία σε συγκεκριμένες κυτταρικές σειρές.
- Ένα γονίδιο μπορεί να συμμετέχει σε πολλά biclusters που δεν είναι απαραίτητο να σχετίζονται μεταξύ τους.
- Παρέχει αρκετή πληροφορία ικανή να μας οδηγήσει σε χρήσιμα συμπεράσματα για κάθε τύπο καρκίνου ξεχωριστά, αλλά και για μία πιο γενική εικόνα όλων των τύπων μαζί.
- Οι τιμές έκφρασης των γονιδίων εντός των biclusters παρουσιάζουν μικρό εύρος, γεγονός που μας βοηθά στη μελέτη και την καλύτερη ανάλυση της πληροφορίας.

ΚΕΦΑΛΑΙΟ 5
ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η βιολογική αξιολόγηση των αποτελεσμάτων ακολούθησε την ανάλυση της διπλής κατηγοριοποίησης σύμφωνα με την προτεινόμενη μεθοδολογία, που όπως ήδη αναφέρθηκε, εφαρμόστηκε σ' ένα σύνολο δεδομένων που περιείχε τις τιμές έκφρασης των 33096 γονιδίων για τις 38 κυτταρικές σειρές των τεσσάρων καρκινικών τύπων (μαστού, τραχήλου της μήτρας, ενδομητρίου, και ωοθηκών) [32]. Η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης στο σύνολο δεδομένων που μελετήθηκε, είχε ως αποτέλεσμα την ανάδειξη 20 «ισχυρών» ομάδων γονιδίων για τον καθένα από τους τέσσερις καρκινικούς τύπους ξεχωριστά (του μαστού, των ωοθηκών, του τραχήλου της μήτρας και του ενδομητρίου), αλλά και 20 «ισχυρών» ομάδων γονιδίων κατά την ομαδοποίηση του συνόλου των 38 καρκινικών κυτταρικών σειρών των τεσσάρων καρκινικών τύπων.

Για την βιολογική ερμηνεία των αποτελεσμάτων διπλής κατηγοριοποίησης εξετάστηκαν ξεχωριστά όλες οι λίστες των γονιδίων που συνιστούν την κάθε ομάδα διπλής κατηγοριοποίησης για τον κάθε καρκινικό τύπο και για το σύνολο των καρκινικών τύπων με το πρόσθετο εργαλείο ClueGO του Cytoscape [34][35]. Επίσης, μελετήθηκε η λίστα που περιέχει τα κοινά γονίδια όλων των κυτταρικών τύπων.

Η βιολογική σημασία των ομάδων διπλής κατηγοριοποίησης επαληθεύτηκε με βάση τον εμπλουτισμό των βιολογικών όρων ενός εκ των τριών δομημένων ελεγχόμενων λεξιλογίων της γονιδιακής οντολογίας (τις βιολογικές διεργασίες), αλλά και των χαρτογραφημένων μονοπατιών KEGG, Wikipathways, και Reactome. Ο βαθμός του λειτουργικού εμπλουτισμού (p-values) των βιολογικών όρων υπολογίζεται χρησιμοποιώντας μια αθροιστική εκθετική (υπεργεωμετρική) κατανομή που μετρά την πιθανότητα εύρεσης του αριθμού των γονιδίων που εμπλέκονται σε ένα συγκεκριμένο βιολογικό όρο εντός της ομάδας διπλής κατηγοριοποίησης. Η σημαντικότητα μιας βιολογικής διεργασίας ή μονοπατιού καθορίστηκε από την τιμή p-value με κατώφλι το 0.05. Ως ελάχιστος αριθμός γονιδίων που συμμετέχουν σε μια βιολογική διεργασία ή μονοπάτι επιλέχθηκε ο αριθμός 2.

Στον Πίνακα 3 αποτυπώνεται: 1) ο αριθμός των γονιδίων που ταυτοποιήθηκαν σε κάθε ομάδα διπλής κατηγοριοποίησης, 2) ο συνολικός αριθμός των γονιδίων που συνιστά την κάθε ομάδα διπλής κατηγοριοποίησης, 3) το ποσοστό των γονιδίων, σε παρένθεση, που σχολιάστηκαν σύμφωνα με το WebGestalt και στο οποίο αποδόθηκαν βιολογικές διεργασίες και μονοπάτια, και 4) το άθροισμα των σημαντικότερων στατιστικά βιολογικών όρων, όπως αποδόθηκε σε κάθε ομάδα διπλής κατηγοριοποίησης σύμφωνα με το ClueGO.

ΠΙΝΑΚΑΣ 3. ΒΙΟΛΟΓΙΚΟΙ ΟΡΟΙ ΤΩΝ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ				
ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ	**κυτταρικές σειρές καρκίνου του μαστού**	**κυτταρικές σειρές καρκίνου του τραχήλου της μήτρας**	**κυτταρικές σειρές καρκίνου του ενδομητρίου**	**κυτταρικές σειρές καρκίνου των ωοθηκών**
	Γονίδια Ταυτοποιημένα/Ολικά (%)	**Γονίδια** Ταυτοποιημένα/Ολικά (%)	**Γονίδια** Ταυτοποιημένα/Ολικά (%)	**Γονίδια** Ταυτοποιημένα/ Ολικά (%)
	Βιολογικοί Όροι	**Βιολογικοί Όροι**	**Βιολογικοί Όροι**	**Βιολογικοί Όροι**
Bicluster 1	61/93 (65.6%)	208/259 (80.3%)	223/275 (81.1%)	114/163 (69.9%)
	9	36	45	13
Bicluster 2	34/49 (69.4%)	77/98 (78.6%)	118/133 (88.72%)	91/97 (**93.8%**)
	1	12	16	7
Bicluster 3	41/46 (89%)	78/90 (86.6%)	72/118 (61%)	39/46 (84.8%)
	8	6	5	1
Bicluster 4	33/44 (75%)	70/83 (84.3%)	76/84 (**90.5%**)	31/42 (73.8%)
	6	5	7	4
Bicluster 5	35/41 (85.4%)	36/73 (*49.3%*)	61/76 (80.3%)	33/39 (84.6%)
	2	1	9	5
Bicluster 6	33/37 (89.2%)	42/50 (84%)	59/70 (84.3%)	22/32 (68.75%)
	5	10	7	3
Bicluster 7	32/36 (88.8%)	40/48 (83.3%)	50/56 (89.3%)	15/27 (55.5%)
	8	6	6	NS
Bicluster 8	31/33 (**93.9%**)	29/34 (85.3%)	36/47 (76.6%)	19/24 (79.2%)
	4	1	1	1
Bicluster 9	26/30 (86.6%)	20/24 (83.3%)	36/44 (81.8%)	15/21 (71.4%)
	5	NS	5	1
Bicluster 10	24/27 (88.8%)	19/23 (82.6%)	32/37 (86.5%)	10/19 (52.6%)
	2	NS	3	1
Bicluster 11	21/23 (**91.3%**)	14/19 (73.7%)	22/32 (68.75%)	13/16 (81.25%)
	NS	NS	1	2
Bicluster 12	21/23 (**91.3%**)	14/18 (77.7%)	26/32 (81.25%)	12/15 (80%)
	NS	NS	NS	NS
Bicluster 13	19/22 (86.4%)	11/15 (73.3%)	27/29 (**93.1%**)	11/14 (78.6%)
	3	NS	NS	NS
Bicluster 14	21/21 (**100%**)	12/14 (85.7%)	17/24 (70.8%)	12/14 (85.7%)
	1	NS	1	2
Bicluster 15	16/20 (80%)	11/12 (**91.6%**)	21/22 (**95.45%**)	10/14 (71.4%)
	NS	NS	2	1
Bicluster 16	12/20 (60%)	9/11 (81.8%)	14/21 (66.6%)	13/13 (**100%**)
	NS	NS	3	NS
Bicluster 17	18/19 (**94.7%**)	7/10 (70%)	17/20 (85%)	10/12 (83.3%)
	NS	NS	NS	NS
Bicluster 18	14/16 (87.5%)	8/9 (88.8%)	17/19 (89.5%)	6/6 (**100%**)
	1	NS	1	NS
Bicluster 19	14/16 (87.5%)	6/7 (85.7%)	10/19 (52.6%)	2/6 (*33.3%*)
	NS	NS	NS	NS
Bicluster 20	4/9 (*44.4%*)	5/7 (71.4%)	12/17 (70.6%)	5/5 (**100%**)
	NS	NS	2	NS

Οι στατιστικά σημαντικοί όροι για τις βιολογικές διεργασίες GO και τα μοριακά μονοπάτια KEGG, Wikipathways και Reactome για τους τέσσερις καρκινικούς τύπους παρουσιάζονται αναλυτικά με τη μορφή πίτας (pie) στο Παράρτημα Β (Β-1.2, Β-2.2, Β-3.2, και Β-4.2). Τα στατιστικά μη-σημαντικά αποτελέσματα χαρακτηρίζονται ως NS (non-significant).

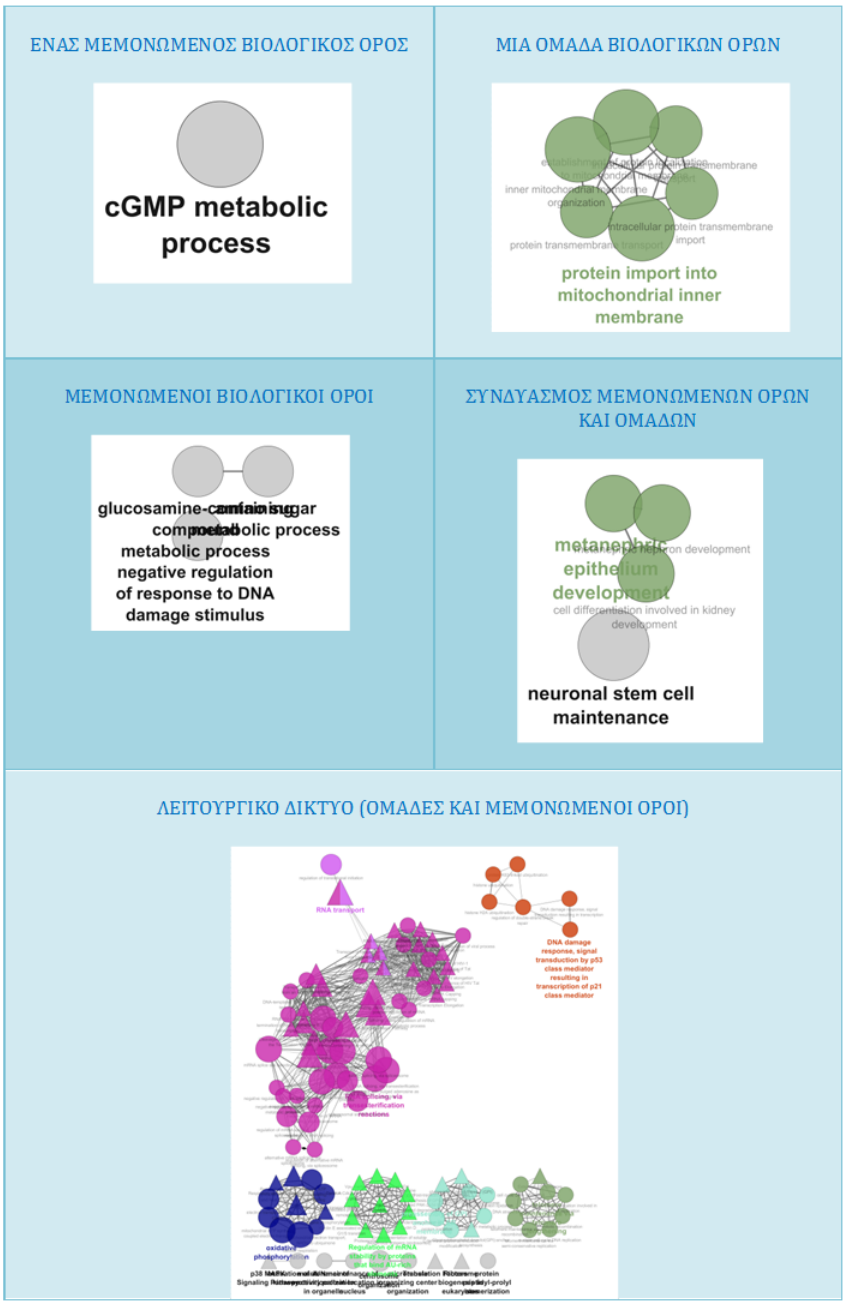
Ο Πίνακας 3 συνοψίζει τα ποσοτικά αποτελέσματα της βιολογικής ανάλυσης των 20 ομάδων διπλής κατηγοριοποίησης με το πρόσθετο εργαλείο ClueGO για τις κυτταρικές σειρές του καρκίνου του μαστού, του καρκίνου του τραχήλου της μήτρας, του καρκίνου του ενδομητρίου, και του καρκίνου των ωοθηκών, οδηγώντας στις ακόλουθες παρατηρήσεις:

- Ένας αριθμός γονιδίων διαφόρων ομάδων δεν "ταυτοποιήθηκε".
- Αρκετές ομάδες διπλής κατηγοριοποίησης δεν έδωσαν στατιστικά σημαντικό αποτέλεσμα ($p<0.05$).
- Τα πολύπλοκα λειτουργικά δίκτυα δεν συνδέονται απαραίτητα με το πλήθος των γονιδίων των ομάδων διπλής κατηγοριοποίησης.

Ο χαμηλός βαθμός λειτουργικού εμπλουτισμού ($p>0.05$) κάποιων ομάδων διπλής κατηγοριοποίησης δεν μπόρεσε να συνδεθεί άμεσα με τον μικρό αριθμό των γονιδίων που αποτελούν την εκάστοτε ομάδα ή τον αριθμό των μη "ταυτοποιημένων" ή αλλιώς των άγνωστων γονιδίων. Αντίθετα, ο εμπλουτισμός των βιολογικών όρων μιας ομάδας διπλής κατηγοριοποίησης που περικλείει και μη "ταυτοποιημένα" γονίδια επιτρέπει εν μέρει το ***λειτουργικό*** χαρακτηρισμό αυτών των γονιδίων.

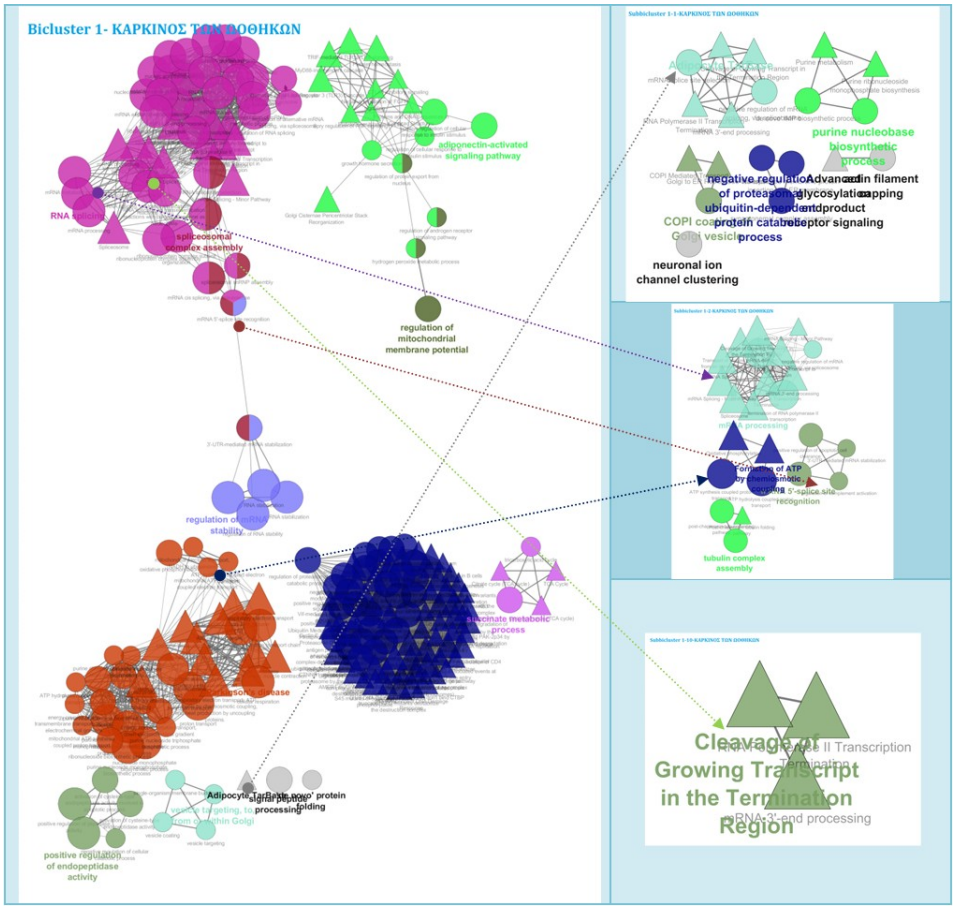
Ταυτόχρονα, ο υψηλός βαθμός λειτουργικού εμπλουτισμού ($p<0.05$) των περισσοτέρων ομάδων διπλής κατηγοριοποίησης, οδήγησε αρκετές φορές σε πολύπλοκα λειτουργικά δίκτυα, τα οποία ανακλούν όρους βιολογικών διεργασιών και μονοπατιών που σχετίζονται μεταξύ τους βάσει της ομοιότητας των συνδεδεμένων γονιδίων τους. Σ' αυτά τα δίκτυα οι βιολογικοί όροι ομαδοποιούνται σε μικρές ή μεγαλύτερες ομάδες ή συνιστούν μεμονωμένους όρους αποδίδοντας πολλές βιολογικές έννοιες σε μια ομάδα διπλής κατηγοριοποίησης ή ένα ευρύτερο πλαίσιο μιας βιολογικής διεργασίας, όπως είναι για παράδειγμα οι πολύπλοκες διαδικασίες της «γονιδιακής έκφρασης» (gene expression) που αποτελείται από πολυάριθμα εξειδικευμένα βήματα.

Ειδικότερα, οι στατιστικά σημαντικοί όροι που αποδίδονται στα γονίδια των ομάδων διπλής κατηγοριοποίησης μπορεί να έχουν τη μορφή ενός ή περισσοτέρων όρων, μιας ομάδας βιολογικών όρων, ενός συνδυασμού μεμονωμένων βιολογικών όρων και ομάδων, ή ενός μεγάλου λειτουργικού δικτύου που αποτελείται από μια ή περισσότερες λειτουργικές ομάδες και μεμονωμένους βιολογικούς όρους (Εικόνα 44).



Εικόνα 44. Παραδείγματα των ποικίλων μορφών εμφάνισης των στατιστικά σημαντικών βιολογικών όρων που αποδίδονται στα γονίδια των ομάδων διπλής κατηγοριοποίησης.

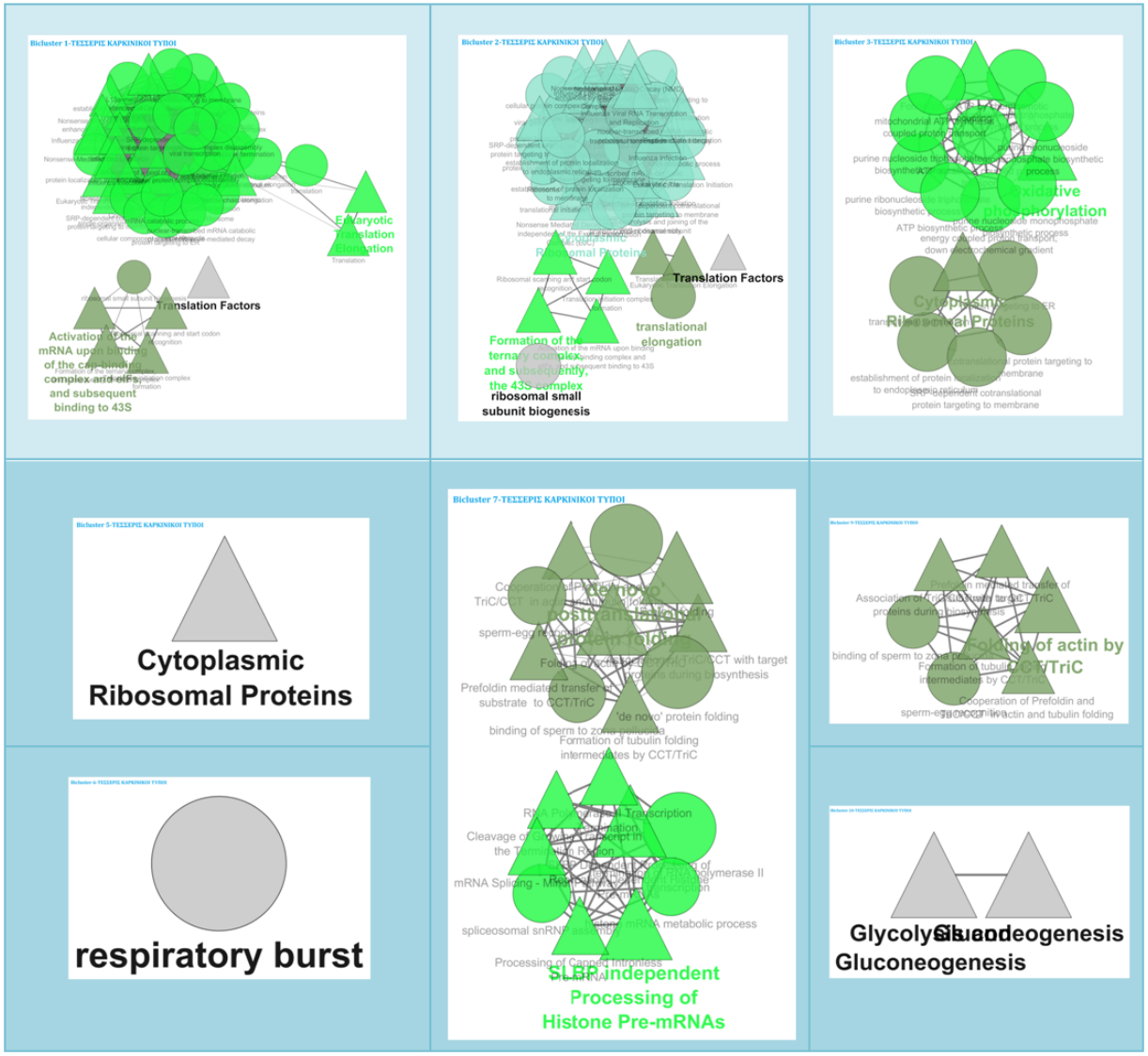
Τέλος, ένα λειτουργικό δίκτυο, όπως είναι το δίκτυο της ομάδας 1 διπλής κατηγοριοποίησης (Bicluster 1) στον καρκίνο των ωοθηκών που αποτελείται από 10 ομάδες και τρεις μεμονωμένους βιολογικούς όρους, δύναται να επιμεριστεί σε μικρότερες εξίσου «ισχυρές» υπο-ομάδες, στις οποίες αποδίδονται εμπλουτισμένοι βιολογικοί όροι της μητρικής ομάδας, όπως φαίνεται στην Εικόνα 45, ενισχύοντας την προτεινόμενη μεθοδολογία (Εικόνες 30, 31, 32).



Εικόνα 45. Λειτουργικό δίκτυο στον καρκίνο των ωοθηκών της ομάδας 1 διπλής κατηγοριοποίησης. Επιμερισμός σε υπο-ομάδες διπλής κατηγοριοποίησης. Ο κύκλος συμβολίζει τους όρους GO, ενώ το τρίγωνο συμβολίζει τα μονοπάτια KEGG, Wikipathways, και Reactome.

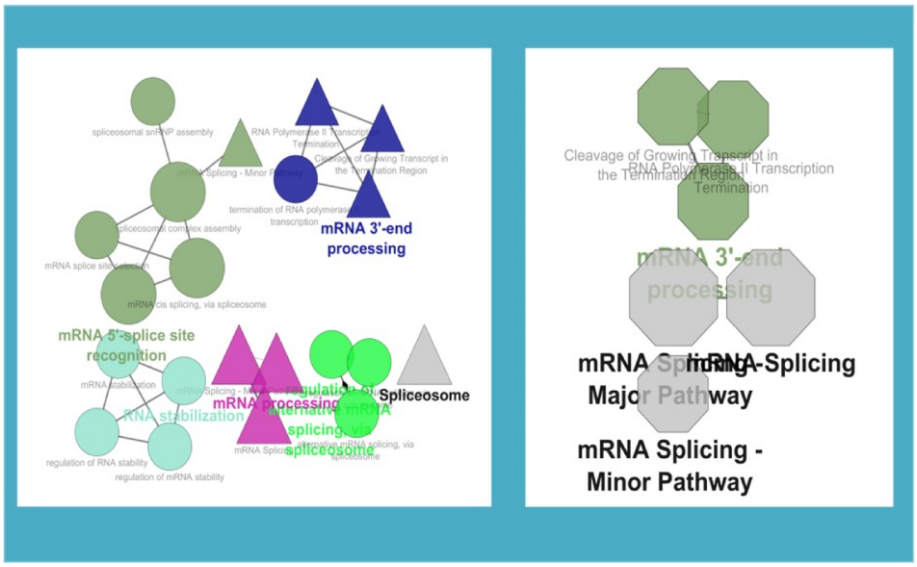
Σήμερα, γνωρίζουμε ότι οι τεχνικές διπλής κατηγοριοποίησης αποτελούν χρήσιμα εργαλεία σε μελέτες γενετικής του καρκίνου, επειδή μπορούν να εντοπίσουν υποομάδες ασθενών με καρκίνο βάσει της ομοιότητας των μοτίβων έκφρασης που μοιράζονται καθώς και υποομάδες γονιδίων που είναι ειδικές για αυτούς τους υποτύπους του καρκίνου. Με αυτό τον τρόπο δύνανται να συνδέσουν συγκεκριμένους φαινοτύπους σε γονότυπους και να χρησιμεύσουν ως βιοδείκτες [4].

Συνεπώς, οι βιολογικοί όροι που αφορούν σε κάθε ομάδα διπλής κατηγοριοποίησης για κάθε καρκινικό τύπο (Παράρτημα Β) και το σύνολο των τεσσάρων καρκινικών τύπων (Εικόνα 46) που προκύπτει από την εφαρμογή της μεθόδου διπλής κατηγοριοποίησης αποκτούν ιδιαίτερη βαρύτητα, αφού περιγράφουν συγκεκριμένες διεργασίες και μονοπάτια που χαρακτηρίζουν κάθε μεμονωμένη κυτταρική σειρά σε κάθε τύπο καρκίνου αλλά και το σύνολο των κυτταρικών σειρών σε όλους τους καρκινικούς τύπους. Η ομοιότητα του μοτίβου γονιδιακής έκφρασης και η μεγάλη εξειδίκευση των βιολογικών όρων που χαρακτηρίζουν τα γονίδια της εκάστοτε ομάδας, επιτρέπει την απόδοση του όρου «***ομάδα-γονιδιακών καρκινικών δεικτών»*** για τα γονίδια κάθε ομάδας διπλής κατηγοριοποίησης σε <u>κάθε</u> καρκινικό τύπο και το <u>σύνολο</u> των τεσσάρων καρκινικών τύπων.



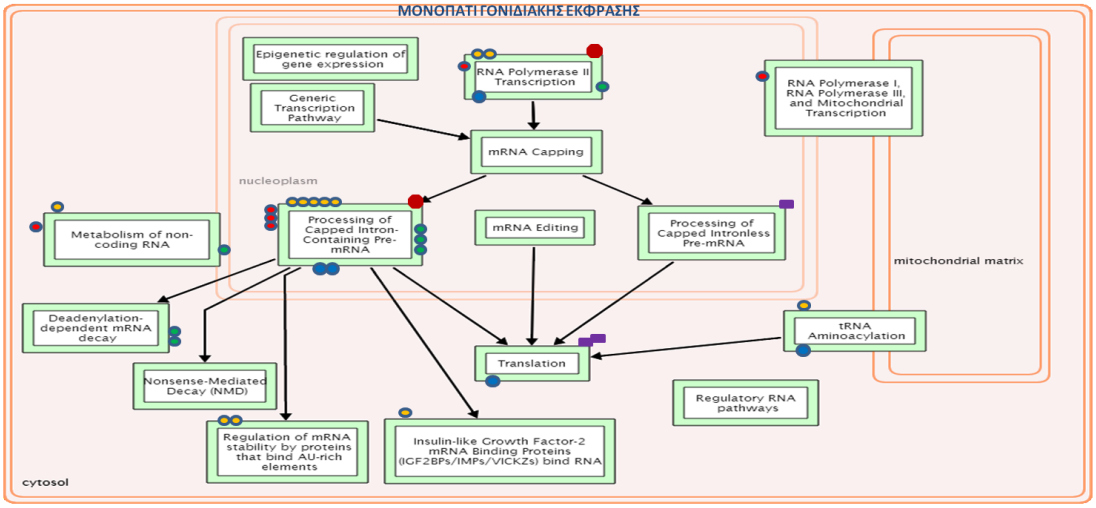
Εικόνα 46. Λειτουργικά δίκτυα από την ομαδοποίηση των τεσσάρων τύπων καρκίνου (μαστού, τραχήλου της μήτρας, ενδομητρίου, ωοθηκών).
Κύκλος: Όροι GO, Τρίγωνο: Μονοπάτια KEGG, Wikipathways, Reactome.

Εκτός από τις «***ομάδες-γονιδιακών* καρκινικών δεικτών»,** η μελέτη διπλής κατηγοριοποίησης ανέδειξε 27 κοινά γονίδια για όλους τους καρκινικούς τύπους, από τα οποία μπόρεσαν να σχολιαστούν 23 γονίδια σύμφωνα με το WebGestalt, και τα οποία συνιστούν ένα λειτουργικό δίκτυο σύμφωνα με το ClueGO (Εικόνα 47).



Εικόνα 47. Λειτουργικά δίκτυα των 23 κοινών γονιδίων των τεσσάρων τύπων καρκίνου (μαστού, τραχήλου της μήτρας, ενδομητρίου, ωοθηκών). Αριστερά: Βιολογικές διεργασίες GO και μονοπάτια KEGG, Wikipathways, και Reactome (ο κύκλος συμβολίζει τους όρους GO, ενώ το τρίγωνο συμβολίζει τα μονοπάτια). Δεξιά: Μονοπάτια Reactome.

Συνολικά, οι είκοσι «ισχυρές» ομάδες διπλής κατηγοριοποίησης που επιλέχθηκαν να μελετηθούν σε κάθε καρκινικό τύπο, στο σύνολο των τεσσάρων καρκινικών τύπων αλλά και τα 23 κοινά γονίδια του καρκίνου του μαστού, του τραχήλου της μήτρας, του ενωδομητρίου και των ωοθηκών, ανακλούν βιολογικές διεργασίες και μονοπάτια που σχετίζονται κυρίως με εξειδικευμένα στάδια της γονιδιακής έκφρασης, όπως αποτυπώνεται γραφικά στο μονοπάτι της «γονιδιακής έκφρασης» (gene expression) Reactome στην Εικόνα 48.



Εικόνα 48. Απλοποιημένη εικόνα της σύνδεσης των γεγονότων που εμπεριέχονται στην έννοια της «γονιδιακής έκφρασης» (gene expression). Μονοπάτι Reactome – Γονιδιακή Έκφραση [39]. Κόκκινος κύκλος: Καρκίνος του μαστού, Κίτρινος κύκλος: Καρκίνος του τραχήλου της μήτρας, Πράσινος κύκλος: Καρκίνος του ενδομητρίου, Μπλε κύκλος: Καρκίνος των ωοθηκών, Μωβ Παραλληλόγραμμο: Σύνολο Τεσσάρων Καρκινικών Τύπων, Κόκκινο Οκτάγωνο: 23 κοινά γονίδια.

Στην Εικόνα 49 αποτυπώνονται άλλες σημαντικές κοινές διεργασίες/μονοπάτια, όπως είναι ο μεταβολισμός και κυρίως ο μεταβολισμός των πρωτεϊνών (αναδίπλωση πρωτεϊνών και μετα-μεταφραστικές τροποποιήσεις), και ο κυτταρικός κύκλος (μίτωση). Μονοπάτια, όπως η μόλυνση από ιό και το ανοσοποιητικό σύστημα απαντώνται σε τρεις καρκινικούς τύπους, ενώ άλλα μονοπάτια όπως η απόπτωση, το σύμπλεγμα του υποδοχέα TGF-β στον καρκίνο και η επιδιόρθωση του DNA σε δύο καρκινικούς τύπους. Τέλος, συγκεκριμένα μονοπάτια χαρακτηρίζουν έναν μόνο καρκινικό τύπο, όπως είναι η αντιγραφή του DNA για τον καρκίνο του μαστού, τα σημεία ελέγχου του κυτταρικού κύκλου για τον καρκίνο του τραχήλου της μήτρας, η διατήρηση των χρωμοσωμάτων για τον καρκίνο του ενδομητρίου και η αιμόσταση για τον καρκίνο των ωοθηκών.

Εικόνα 49. Κοινά και διαφορετικά Μονοπάτια Reactome για κάθε καρκινικό τύπο, το σύνολο των τεσσάρων καρκινικών τύπων και τα 23 κοινά γονίδια.
Κόκκινος κύκλος: Καρκίνος του μαστού, Κίτρινος κύκλος: Καρκίνος του τραχήλου της μήτρας, Πράσινος κύκλος: Καρκίνος του ενδομητρίου, Μπλε κύκλος: Καρκίνος των ωοθηκών, Μωβ Παραλληλόγραμμο: Σύνολο Τεσσάρων Καρκινικών Τύπων, Κόκκινο Οκτάγωνο: 23 κοινά γονίδια.

Μελέτες γονιδιακής έκφρασης έχουν επιβεβαιώσει ότι ο καρκίνος διακρίνεται για την ύψιστη φαινοτυπική πολυπλοκότητα, και ποικίλλει σε μεγάλο βαθμό από την άποψη των υποτύπων και των εξελικτικών σταδίων. Για να χαρακτηρίσουμε με επιτυχία δείγματα καρκινικού ιστού χρειάζονται αξιόπιστα κριτήρια σε βιομοριακό επίπεδο, γεγονός που ερμηνεύεται σε *παθολογικούς δείκτες* της νόσου. Έτσι, ενώ οι *δείκτες* χρησιμεύουν κυρίως ως *δείκτες* ενός καρκινικού τύπου και/ή υποτύπων του, μπορούν επίσης να υποδείξουν τις διαταραχές σε κρίσιμες βιολογικές διεργασίες και μονοπάτια που ενέχονται στην πρόκληση του καρκίνου, αναδεικνύοντας ταυτόχρονα την αξία της μελέτης τους πέρα από το ρόλο τους ως στοιχεία ταξινόμησης [40].

Στη παρούσα μελέτη εντοπίσαμε έναν ***ειδικό* καρκινικό δείκτη**, το γονίδιο *METTL5*, που βρέθηκε να είναι το μοναδικό κοινό γονίδιο κάθε καρκινικού τύπου και του συνόλου των τεσσάρων καρκινικών τύπων, αλλά και «***ομάδων-γονιδιακών* παθολογικών δεικτών»** καθώς και **23 *πολυγονιδιακών* καρκινικών δεικτών** για όλες τις ομάδες των καρκινικών τύπων.

Στο γονίδιο *METTL5* έχει αποδοθεί η ιδιότητα της δραστικότητας μεθυλοτρανσφεράσης και ικανότητας σύνδεσης νουκλεϊκού οξέος. Αν και οι μέχρι σήμερα γνώσεις μας δεν επαρκούν για να περιγράψουμε το *METTL5*, έχει αποδειχθεί πειραματικά ότι διαφοροποιείται η έκφρασή του σε συγκεκριμένους υποτύπους καρκίνου του μαστού (ER+ vs ER-) [41], και έχει παρατηρηθεί ένα μικρό ποσοστό μετάλλαξης και παραλλαγής αριθμού αντιγράφων (CNV) του *METTL5* στον καρκίνο του μαστού, του ενδομητρίου και των ωοθηκών [42],[43],[44]. Έτσι, ο ρόλος του *METTL5* ως ***ειδικού* καρκινικού δείκτη** αξίζει να επιβεβαιωθεί πειραματικά.

Από την άλλη πλευρά, γνωρίζοντας ότι μια *ομάδα-δεικτών* είναι πολλές φορές σημαντικότερη του ενός βιοδείκτη κατά την ταξινόμηση των δειγμάτων καρκινικού ιστού [40] και αναγνωρίζοντας ότι οι βιολογικές διεργασίες και τα μονοπάτια που διέπουν τις «***ομάδες-γονιδιακών* καρκινικών δεικτών»** καθώς και τους **23 *πολυγονιδιακούς* καρκινικούς δείκτες** εμπλέκονται στην παθογένεια του καρκίνου, υποδεικνύουν την σημασία της παρούσας μελέτης.

Συμπερασματικά, τα αποτελέσματα από την εφαρμογή της μεθοδολογίας μας σε καρκινικές κυτταρικές σειρές αναδεικνύουν τον διττό ρόλο των ***δεικτών*** που ανακαλύψαμε, της ταξινόμησης και της αιτιοπαθογένειας.

Αρκετές μελέτες, έχουν εξετάσει τις ιδιότητες των καρκινικών κυτταρικών σειρών, για τους τέσσερις καρκινικούς τύπους (ωοθηκών, μαστού, τραχήλου της μήτρας και ενδομητρίου)[45][46][47][48][16], στοχεύοντας στα ποσοτικά και ποιοτικά χαρακτηριστικά των κυτταρικών σειρών προκειμένου να χρησιμοποιηθούν ως ιστο-ειδικά ανάλογα των κυττάρων προέλευσής τους για την ανάλυση της βιολογίας του όγκου αλλά και των αποτελεσμάτων της ακτινοβολίας και των χημειοθεραπευτικών φαρμάκων στα ανθρώπινα κύτταρα όγκου ωοθηκών, μαστού, τραχήλου της μήτρας και ενδομητρίου.

Οι Πίνακες 4 και 5, απεικονίζουν την παρουσία ή απουσία συγκεκριμένων ομάδων διπλής κατηγοριοποίησης για κάθε καρκινικό τύπο και για το σύνολο των τεσσάρων καρκινικών τύπων, αντίστοιχα.

Οι ανθρώπινες κυτταρικές σειρές, όπως αναφέρθηκε, χρησιμοποιούνται ευρέως στη βασική και μεταφραστική βιοϊατρική έρευνα, καθώς αποτελούν ένα απλό και αντιπροσωπευτικό μοντέλο συστήματος για διάφορες λειτουργικές μελέτες και την ταυτοποίηση διαγνωστικών εργαλείων και θεραπευτικών στόχων. Κάθε κυτταρική σειρά έχει μοναδικά χαρακτηριστικά και μπορεί να χρησιμοποιηθεί για ειδικές μελέτες [49][50].

Επομένως, είναι σημαντικό να γνωρίζουμε τα χαρακτηριστικά των κυτταρικών σειρών δίνοντας τη δυνατότητα στοχευμένων μελετών. Για παράδειγμα, η παρουσία ή η απουσία συγκεκριμένων ομάδων διπλής κατηγοριοποίησης από τις κυτταρικές σειρές A2780 και CAOV3 στον καρκίνο των ωοθηκών [Πίνακας 4], θα μπορούσε να σχετίζεται με τα ευρήματα μιας πρόσφατης μελέτης όπου χαρακτηρίζει τις συγκεκριμένες σειρές βάσει των χαρακτηριστικών τους σε όχι τόσο καλές και καλές, αντίστοιχα [Domcke]. Επίσης, η απουσία της ομάδας 4 διπλής κατηγοριοποίησης από την κυτταρική σειρά AN3CA στον καρκίνο των ωοθηκών (Πίνακας 5), θα μπορούσε να σχετίζεται με την γενετική αστάθεια που χαρακτηρίζει την συγκεκριμένη κυτταρική σειρά [51][52][53].

Συνολικά, η προτεινόμενη μεθοδολογία μας της τεχνικής διπλής κατηγοριοποίησης κατάφερε να πετύχει τους στόχους της, όπως αυτοί έχουν οριστεί από τις μέχρι σήμερα εφαρμογές της, όπως αναφέρεται στην ενότητα 1.2 (σχετική βιβλιογραφία).

Συγκεκριμένα:

- πέτυχε την ομαδοποίηση γονιδίων με παρόμοια συμπεριφορά που συμμετέχουν σε παρόμοιες λειτουργίες (συν-ρυθμιζόμενα γονίδια),
- πέτυχε τον εξειδικευμένο λειτουργικό σχολιασμό των γνωστών γονιδίων,
- επέτρεψε τον λειτουργικό σχολιασμό των άγνωστων γονιδίων ή εν μέρει γνωστών γονιδίων με άγνωστες λειτουργίες, και
- κατάφερε την κατηγοριοποίηση των κυτταρικών σειρών σύμφωνα με τα ιδιαίτερα χαρακτηριστικά των biclusters που συμμετέχουν σ' αυτές.

Πίνακας 4. ΠΑΡΟΥΣΙΑ Ή ΑΠΟΥΣΙΑ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΙΚΕΣ ΣΕΙΡΕΣ ΚΑΘΕ ΚΑΡΚΙΝΙΚΟΥ ΤΥΠΟΥ

ΤΥΠΟΣ	ΑΑ	ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ	20 ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (BICLUSTERS)

Καρκίνος του Μαστού

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ΚΑΡΚΙΝΟΣ ΤΟΥ ΜΑΣΤΟΥ	1	MCF-7																				
	2	MM231																				
	3	T47D																				
	4	Hs578T																				
	5	SKBR3																				
	6	MM435s																				
	7	ZR75-1																				
	8	BT-549																				
	9	MM453																				
	10	BT474																				

Καρκίνος του Τραχήλου της Μήτρας

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ	11	SW756																				
	12	GH354																				
	13	C-4I																				
	14	Hela																				
	15	C-33A																				
	16	CaSki																				
	17	ME180																				
	18	HT-3																				
	19	SiHa																				

Καρκίνος του Ενδομητρίου

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ΚΑΡΚΙΝΟΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ	20	Colo684																				
	21	EJ																				
	22	Ishikawa																				
	23	RL95-2																				
	24	HEC50B																				
	25	HEC1B																				
	26	EN																				
	27	KLE																				
	28	AN3CA																				

Καρκίνος των Ωοθηκών

			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ΚΑΡΚΙΝΟΣ ΤΩΝ ΩΟΘΗΚΩΝ	29	TOV-1112D																				
	30	TOV-21G																				
	31	A2780																				
	32	OVMZ-1a																				
	33	OVMZ-6																				
	34	ES2																				
	36	CaOV3																				
	37	OV-90																				
	38	NIHOVCAR3																				
	39	SKOV3																				

Πίνακας 5. ΠΑΡΟΥΣΙΑ Ή ΑΠΟΥΣΙΑ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΙΚΕΣ ΣΕΙΡΕΣ ΤΩΝ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ

ΤΥΠΟΣ	ΑΑ	ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ	20 ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (BICLUSTERS)																		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Καρκίνος του Μαστού	1	MCF-7																				
	2	MM231																				
	3	T47D																				
	4	Hs578T																				
	5	SKBR3																				
	6	MM435s																				
	7	ZR75-1																				
	8	BT-549																				
	9	MM453																				
	10	BT474																				
Καρκίνος του Τραχήλου της Μήτρας	11	SW756																				
	12	GH354																				
	13	C-4I																				
	14	Hela																				
	15	C-33A																				
	16	CaSki																				
	17	ME180																				
	18	HT-3																				
	19	SiHa																				
Καρκίνος του Ενδομητρίου	20	Colo684																				
	21	EJ																				
	22	Ishikawa																				
	23	RL95-2																				
	24	HEC50B																				
	25	HEC1B																				
	26	EN																				
	27	KLE																				
	28	AN3CA																				
Καρκίνος των Ωοθηκών	29	TOV-1112D																				
	30	TOV-21G																				
	31	A2780																				
	32	OVMZ-1a																				
	33	OVMZ-6																				
	34	ES2																				
	35	CaOV3																				
	36	OV-90																				
	37	NIHOVCAR3																				
	38	SKOV3																				

ΚΕΦΆΛΑΙΟ 6
ΣΥΜΠΕΡΆΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΈΣ ΕΠΕΚΤΆΣΕΙΣ

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήσαμε τις μεθόδους διπλής κατηγοριοποίησης σε πίνακες γονιδιακής έκφρασης για την εξαγωγή υποομάδων με στατιστική και βιολογική σημασία και την τελική επιλογή γονιδιακών καρκινικών δεικτών.

Η πειραματική διαδικασία εφαρμόστηκε σε αρχεία που περιελάμβαναν πίνακες γονιδίων-κυτταρικών σειρών προερχόμενα από μικροσυστοιχίες DNA από δείγματα καρκίνου του μαστού, του τραχήλου της μήτρας, των ωοθηκών και του ενδομητρίου. Τα αποτελέσματα που λάβαμε ήταν εξίσου σημαντικά τόσο σε στατιστικό όσο και σε βιολογικό επίπεδο. Συγκεκριμένα:

- Ο βελτιωμένος biclustering αλγόριθμος που χρησιμοποιήθηκε απέδειξε τους λόγους για τους οποίους οι τεχνικές διπλής κατηγοριοποίησης υπερτερούν σε σχέση με αυτές της απλής ομαδοποίησης, δίνοντας έμφαση στην πληροφορία, μειώνοντας το θόρυβο και παρέχοντας μας μια εικόνα για το σύνολο δεδομένων πιο συγκεκριμένη και λεπτομερής.

- Οι ομάδες που δημιουργήθηκαν, εμφάνιζαν γονίδια με όμοια συμπεριφορά κατά μήκος των κυτταρικών σειρών και παράλληλα το εύρος των τιμών τους ήταν μικρό, κάνοντας έτσι τις ομάδες πιο συνεκτικές και ικανές να εξάγουν σημαντικές βιολογικές πληροφορίες, γενικές αλλά και πιο εξειδικευμένες.

Η εξαγωγή ενός μεγάλου αριθμού ομάδων διπλής κατηγοριοποίησης (biclusters) σε πραγματικά δεδομένα μπορεί να οδηγήσει σε αποτελέσματα τα οποία είναι δύσκολο να ερμηνευθούν βιολογικά. Ως εκ τούτου, εστιάσαμε στην μελέτη των 20 πιο ≪ισχυρών≫ ομάδων διπλής κατηγοριοποίησης οδηγώντας στα ακόλουθα:

- Η μελέτη των 20 χαρακτηριστικών προτύπων συμπεριφοράς αφορούν σε μεγάλο βαθμό διαφορετικές εκφάνσεις/πτυχές της ευρύτερης διεργασίας της γονιδιακής έκφρασης.
- Η απουσία συγκεκριμένων ομάδων διπλής κατηγοριοποίησης από κάποιες καρκινικές κυτταρικές σειρές μπορεί να αποσαφηνίσει εν μέρει τις ιδιότητές τους και να αποτελέσει τον οδηγό στοχευμένης θεραπευτικής προσέγγισης.
- Η προτεινόμενη μεθοδολογική προσέγγιση παρέχει έναν *ειδικό* καρκινικό δείκτη και *ομάδες-γονιδιακών* καρκινικών δεικτών για <u>κάθε</u> καρκινικό τύπο αλλά ταυτόχρονα και ομάδα *πολυγονιδιακών* καρκινικών δεικτών που χαρακτηρίζει τους τέσσερις καρκινικούς τύπους.

6.2 Μελλοντικές επεκτάσεις

Όπως συμβαίνει με τις περισσότερες μελέτες, σε περιπτώσεις όπως η δική μας, όπου ο όγκος των δεδομένων είναι αρκετά μεγάλος, επιλέξαμε να μελετήσουμε έναν μικρό αριθμό biclusters για την καλύτερη αξιολόγηση των αποτελεσμάτων. Έτσι μελλοντικά θα μπορούσαμε να επεξεργαστούμε επιπλέον ομαδοποιήσεις που είναι πιθανό να μας παρέχουν χρήσιμη πληροφορία κι έτσι να κρίνουμε ακόμα πιο αντικειμενικά τα τωρινά μας αποτελέσματα.

Επιπλέον, ένα επόμενο βήμα θα αποτελούσε η πιο αναλυτική επεξεργασία και η ιεράρχιση των μεγάλων biclusters που μας παρείχαν αρκετή και σημαντική πληροφορία. Αυτά δηλαδή, που μελετώντας τα πιο μεθοδικά θα μπορούσαμε να καταλήξουμε σε βιολογικές διεργασίες με μεγαλύτερη λεπτομέρεια και εξειδίκευση.

Τέλος, πέραν του αλγορίθμου Cheng and Church, θα ήταν πολύ σημαντική η εφαρμογή και άλλων biclustering αλγορίθμων, για τη σύγκριση των αποτελεσμάτων με τη δική μας εφαρμογή και έτσι θα μπορούσαμε να έχουμε μία πλήρη και πιο σφαιρική εικόνα για την biclustering τεχνική.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, no. 1, pp. 24–45, 2004.

[2] C. Cano, L. Adarve, J. López, and A. Blanco, "Possibilistic approach for biclustering microarray data ," vol. 37, pp. 1426–1436, 2007.

[3] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Finding multiple coherent biclusters in microarray data using variable string length multiobjective genetic algorithm.," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 969–75, Nov. 2009.

[4] C.-P. Chen, H. Fushing, R. Atwill, and P. Koehl, "biDCG: a new method for discovering global features of DNA microarray data via an iterative re-clustering procedure.," *PLoS One*, vol. 9, no. 7, p. e102445, Jan. 2014.

[5] http://ghr.nlm.nih.gov/handbook/basics/gene.

[6] http://www.nature.com/scitable/topicpage/gene-expression-14121669.

[7] G.Patrinos, Γονιδιωματική. http://www.pharmacy.upatras.gr/index.php/el/studies/undergraduate/-mainmenu-63.

[8] http://www.unc.edu/depts/our/hhmi/hhmift_learning_modules/cancermodule/pages/modeling.html.

[9] http://ygeia.tanea.gr/default.asp?pid=8&ct=4&articleID=14523&l.

[10] S. Domcke, R. Sinha, D. a Levine, C. Sander, and N. Schultz, "Evaluating cell lines as tumour models by comparison of genomic profiles.," *Nat. Commun.*, vol. 4, p. 2126, Jan. 2013.

[11] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, V. Gregory, D. Sonkin, A. Reddy, M. Liu, L. Murray, F. Michael, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-, F. A. Mapa, J. Thibault, E. Bric-furlong, P. Raman, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. A. Jr, M. De Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, C. Robert, T. Liefeld, L. Macconaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "NIH Public Access of anticancer drug sensitivity," vol. 483, no. 7391, pp. 603–607, 2012.

[12] T. Vargo-Gogola and J. M. Rosen, "Modelling breast cancer: one size does not fit all.," *Nat. Rev. Cancer*, vol. 7, no. 9, pp. 659–72, Sep. 2007.

[13] M. H. Forouzanfar, K. J. Foreman, A. M. Delossantos, R. Lozano, A. D. Lopez, C. J. L. Murray, and M. Naghavi, "Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis.," *Lancet*, vol. 378, no. 9801, pp. 1461–84, Oct. 2011.

[14] M. Scheffner, K. Münger, J. C. Byrne, and P. M. Howley, "The state of the p53 and retinoblastoma genes in human cervical carcinoma cell lines.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 13, pp. 5523–7, Jul. 1991.

[15] M. W. Carlson, V. R. Iyer, and E. M. Marcotte, "Quantitative gene expression assessment identifies appropriate cell line models for individual cervical cancer pathways.," *BMC Genomics*, vol. 8, p. 117, Jan. 2007.

[16] G. Vollmer, "Endometrial cancer: experimental models useful for studies on molecular aspects of endometrial cancer and carcinogenesis.," *Endocr. Relat. Cancer*, vol. 10, no. 1, pp. 23–42, Mar. 2003.

[17] F. Jacob, S. Nixdorf, N. F. Hacker, and V. a Heinzelmann-Schwarz, “Reliable in vitro studies require appropriate ovarian cancer cell lines.,” *J. Ovarian Res.*, vol. 7, no. 1, p. 60, Jan. 2014.
[18] R. S. N. Fehrmann, X.-Y. Li, A. G. J. van der Zee, S. de Jong, G. J. Te Meerman, E. G. E. de Vries, and A. P. G. Crijns, “Profiling studies in ovarian cancer: a review.,” *Oncologist*, vol. 12, no. 8, pp. 960–6, Aug. 2007.
[19] Σκρέτη Γεωργία ,"Μέθοδοι οπτικοποίησης γονιδιακών δεδομένων", *Διπλωματική εργασία , Πολυτεχνείο Κρήτης*, Χανιά Ιούνιος 2012.
[20] Δ. Παπαυευαγγελίου, Σ. Σολακίδη, Β. Ζουμπουρλής ,*"Μοριακή ανάλυση των νεοπλασμάτων με τη χρήση μικροσυστοιχιών DNA" , Ιατρική επικαιρότητα ,* (2003) ,2092-2095.
[21] Amos Tanay,Roded Sharan,Ron Shamir “Biclustering Algorithms : A Survey”, *School of Computer Science, Tel-Aviv University,* pp. 1-20, May 2004 .
[22] http://www.docstoc.com/docs/84298664/Biclustering-Algorithms-for-Biological-Data-Analysis
[23] http://www.microarrays.ca/services/ ,“What Is Hierarchical Clustering ?”.
[24] http://homes.di.unimi.it/~valenti/SlideCorsi/MB0910/HierarchicalClustering.pdf , Giorgio Valentini, “Hierarchical Clustering for Gene Expression Data Analysis Clustering of Microarray Data.”
[25] http://web.ist.utl.pt/sara.madeira/ ,Sara C.Madeira *"Clustering and Biclustering Gene Expression Data".*
[26] J. L. Flores, I. Inza, P. Larrañaga, and B. Calvo, “A new measure for gene expression biclustering based on non-parametric correlation.,” *Comput. Methods Programs Biomed.*, vol. 112, no. 3, pp. 367–97, Dec. 2013.
[27] http://www.researchgate.net/publication/224806588_Biclustering_of_DNA_Microarray_Data_Theory_Evaluation_and_Applications, B. Alain, H. Ahmed, V. Panayiotis, A. B. Tchagang, Y. Pan, and F. Famili, “Biclustering of DNA Microarray Data : Data : Theory , Theory , Evaluation , and Applications,” 2010.
[28] Roded Sharan,Udi Ben Porat and Ophir Bleiberg , “Analysis of Biological Networks : Network Modules – Clustering and Biclustering ∗”, *Lecture 5,November 23* , pp. 1–21, 2006.
[29] Sebastian Kaiser, “Biclustering: Methods, Software and Application”, *Dissertation*, Munchen 7 March 2011.
[30] Yizong Cheng and George M.Church , "Biclustering of Expression Data", *Department of Genetics, Harvard Medical School, Boston ,MA 02115,Department of ECECS, University of Cincinnati ,CIncinnati, OH 45221* (2000).
[31] http://dl.acm.org/citation.cfm?id=1854814.
[32] E. Obermayr, F. Sanchez-Cabo, M.-K. M. Tea, C. F. Singer, M. Krainer, M. B. Fischer, J. Sehouli, A. Reinthaller, R. Horvat, G. Heinze, D. Tong, and R. Zeillinger, “Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients.,” *BMC Cancer*, vol. 10, no. 1, p. 666, Jan. 2010.
[33] B. Zhang, S. Kirov, and J. Snoddy, “WebGestalt: an integrated system for exploring gene sets in various biological contexts.,” *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W741–8, Jul. 2005.
[34] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, “ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.,” *Bioinformatics*, vol. 25, no. 8, pp. 1091–3, Apr. 2009.

[35] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks," no. Karp 2001, pp. 2498–2504, 2003.
[36] http://www.geneontology.org.
[37] http://www.genome.jp/kegg/ Kanehisa Laboratories.
[38] http://wikipathways.org/index.php/WikiPathways.
[39] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways.," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D428–32, Jan. 2005.
[40] P. Dao, R. Colak, R. Salari, F. Moser, E. Davicioni, A. Schönhuth, and M. Ester, "Inferring cancer subnetwork markers using density-constrained biclustering.," *Bioinformatics*, vol. 26, no. 18, pp. i625–31, Sep. 2010.
[41] Z. J. Sahab, Y.-G. Man, S. M. Semaan, R. G. Newcomer, S. W. Byers, and Q.-X. A. Sang, "Alteration in protein expression in estrogen receptor alpha-negative human breast cancer tissues indicates a malignant and metastatic phenotype.," *Clin. Exp. Metastasis*, vol. 27, no. 7, pp. 493–503, Oct. 2010.
[42] http://www.genecards.org/.
[43] http://www.ebi.ac.uk/QuickGO/GProtein?ac=Q9NRN9.
[44] http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=METTL5#dist.
[45] R. N. Buick, R. Pullano, J. M. Trent, R. N. Buick, R. Rullano, and J. M. Trent, "Comparative Properties of Five Human Ovarian Adenocarcinoma Cell Lines." ,*Cancer Res*, pp. 3668–3676, 1985.
[46] I. I. Wistuba, C. Behrens, S. Milchgrub, I. Wistuba, H. Cunningham, and D. Minna, "Comparison of features of human breast cancer cell lines and their corresponding tumors. " ,*Clin Cancer Res,* pp. 2931–2938, 1998.
[47] R. M. Neve, K. Chin, J. Fridlyand, J. Yeh, F. L. Baehner, T. Fevr, L. Clark, N. Bayani, J.-P. Coppe, F. Tong, T. Speed, P. T. Spellman, S. DeVries, A. Lapuk, N. J. Wang, W.-L. Kuo, J. L. Stilwell, D. Pinkel, D. G. Albertson, F. M. Waldman, F. McCormick, R. B. Dickson, M. D. Johnson, M. Lippman, S. Ethier, A. Gazdar, and J. W. Gray, "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.," *Cancer Cell*, vol. 10, no. 6, pp. 515–27, Dec. 2006.
[48] T. Yamori, "Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics.," *Cancer Chemother. Pharmacol.*, vol. 52 Suppl 1, pp. S74–9, Jul. 2003.
[49] P. Romano, A. Manniello, O. Aresu, M. Armento, M. Cesaro, and B. Parodi, "Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D925–32, Jan. 2009.
[50] J.-P. Gillet, S. Varma, and M. M. Gottesman, "The clinical relevance of cancer cell lines.," *J. Natl. Cancer Inst.*, vol. 105, no. 7, pp. 452–8, Apr. 2013.
[51] K. Orth, J. Hung, a Gazdar, a Bowcock, J. M. Mathis, and J. Sambrook, "Genetic instability in human ovarian cancer cell lines.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 20, pp. 9495–9, Sep. 1994.
[52] R. H. Tumor, C. Lines, M. F. Kane, M. Loda, G. M. Gaida, H. Tumor, C. Lines, M. F. Kane, M. Loda, G. M. Gaida, J. Lipman, R. Mishra, H. Goldman, and J. Milburn, "Methylation of the hMLH1 Promoter Correlates with Lack of Expression of hMLH1 in Sporadic Colon Tumors and Mismatch hMLH1 in Sporadic Colon Tumors and Mismatch Repair-defective Human Cell LInes" , *Cancer Res,* pp. 808–811, 1997.

[53] a Umar, J. C. Boyer, D. C. Thomas, D. C. Nguyen, J. I. Risinger, J. Boyd, Y. Ionov, M. Perucho, and T. a Kunkel, "Defective mismatch repair in extracts of colorectal and endometrial cancer cell lines exhibiting microsatellite instability.", *J. Biol. Chem.*, vol. 269, no. 20, pp. 14367–70, May 1994.

[54] A. Mitra, L. Mishra, and S. Li, "Technologies for deriving primary tumor cells for use in personalized cancer therapy." ,*Trends Biotechnol.*, vol. 31, no. 6, pp. 347–54, Jun. 2013.

[55] http://www.boiseweekly.com/boise/henrietta-lacks-a-genetic-cell-ebrity/Content?oid=2541495.

[56] http://www.ibbl.lu/personalised-medicine/what-is-personalised-medicine/what-is-genotyping-microarray/.

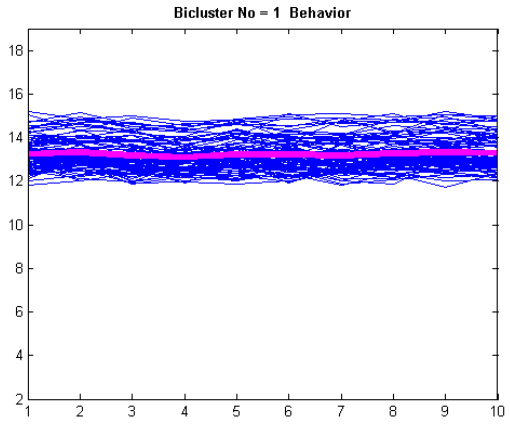
ΠΑΡΑΡΤΗΜΑ Α

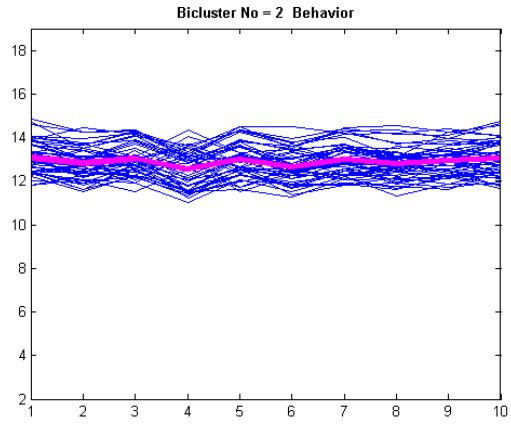
Γραφικές συμπεριφοράς γονιδίων κατά μήκος των σειρών

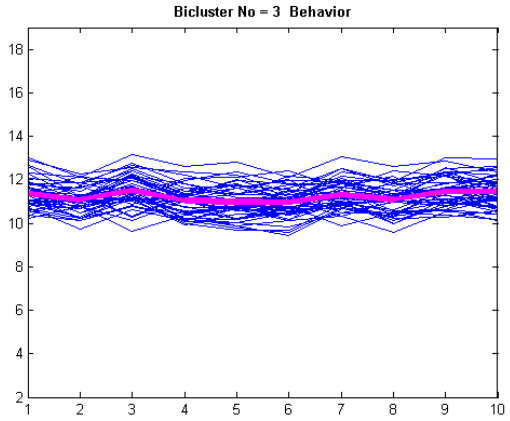
Οι γραφικές που ακολουθούν δείχνουν για κάθε ομάδα-bicluster την συμπεριφορά (γονιδιακή έκφραση) των γονιδίων κατά μήκος των σειρών που ανήκουν σε αυτή. Ο άξονας x απεικονίζει τον αριθμό των σειρών του bicluster ,ο άξονας y τις τιμές έκφρασης των γονιδίων του και οι μπλε γραμμές τα γονίδια της ομάδας.

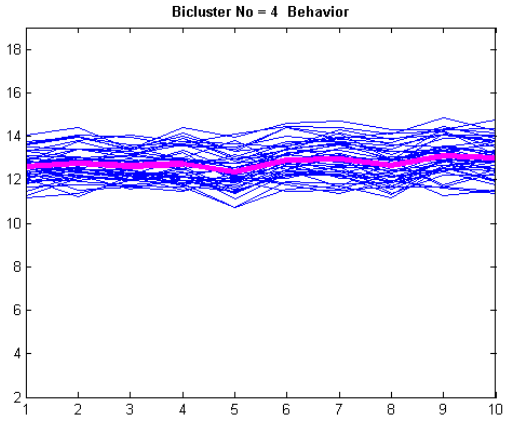
Με τη ροζ γραμμή υποδεικνύεται η μέση τιμή των συμπεριφορών των γονιδίων για κάθε bicluster.

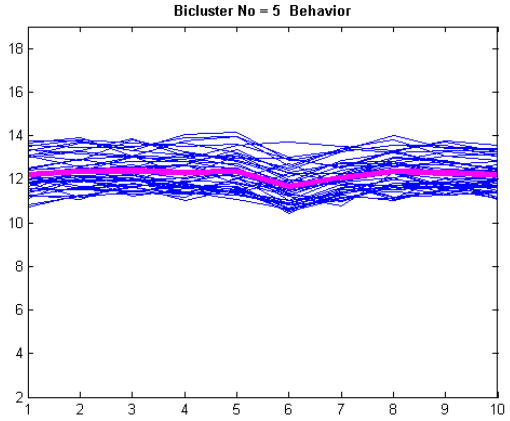
Καρκίνος του μαστού

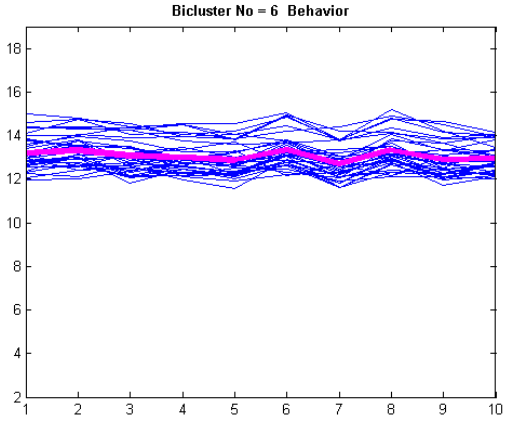


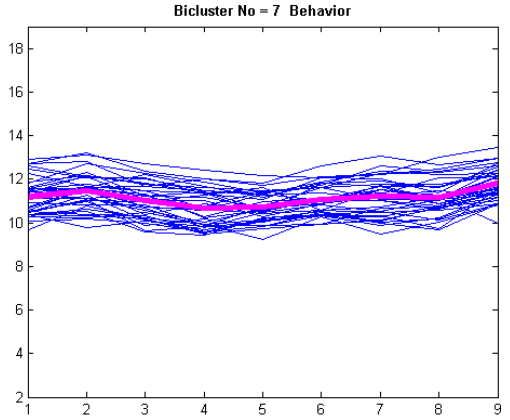


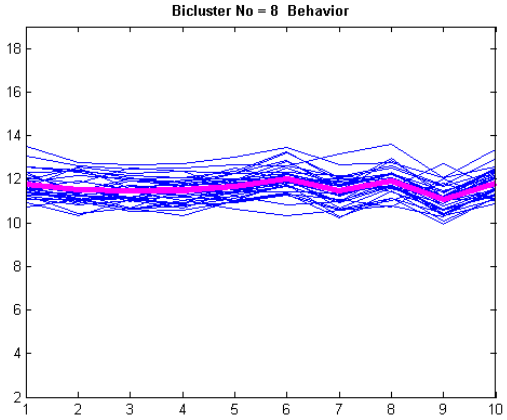

Bicluster No = 3 Behavior

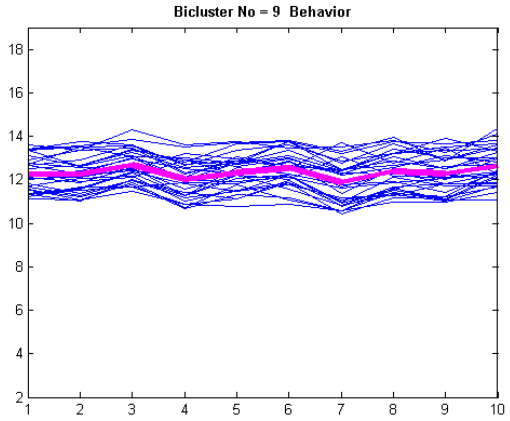

Bicluster No = 4 Behavior

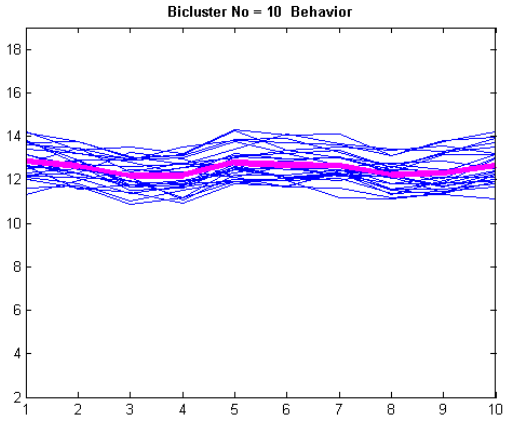

Bicluster No = 5 Behavior

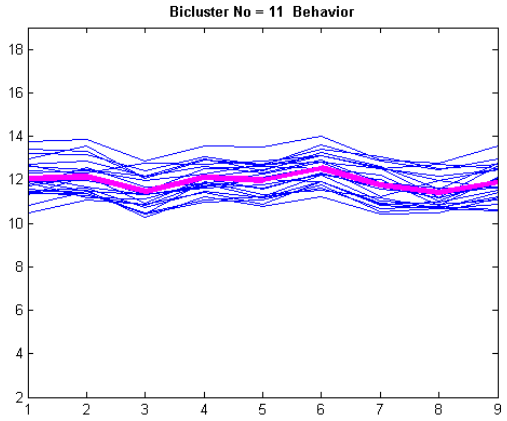

Bicluster No = 6 Behavior

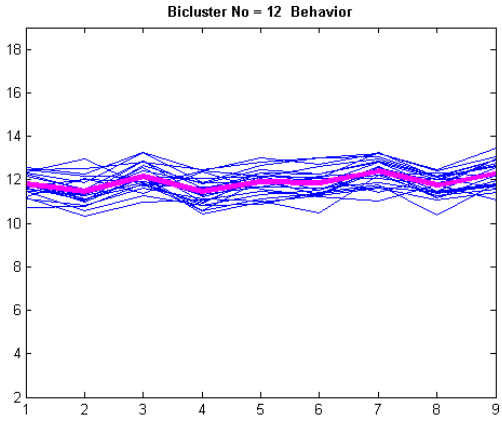

Bicluster No = 7 Behavior

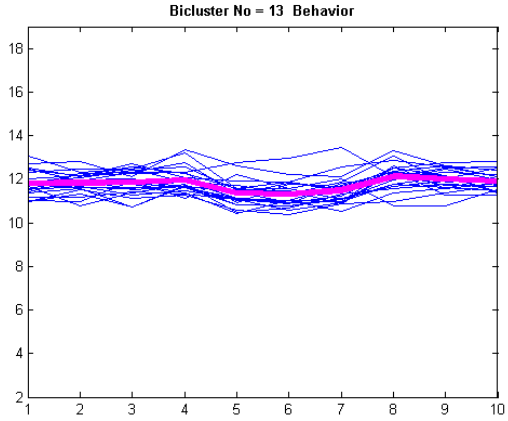

Bicluster No = 8 Behavior

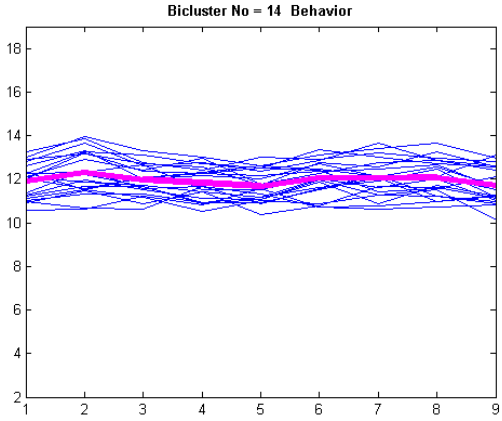

Bicluster No = 9 Behavior

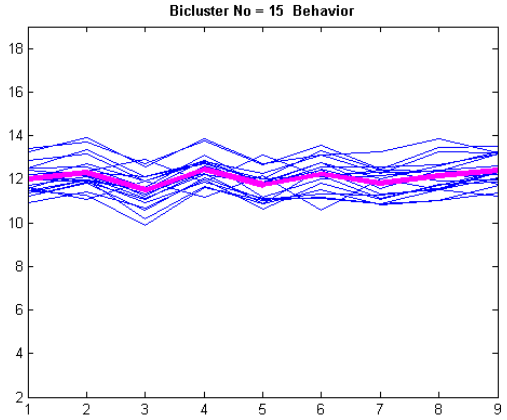

Bicluster No = 10 Behavior

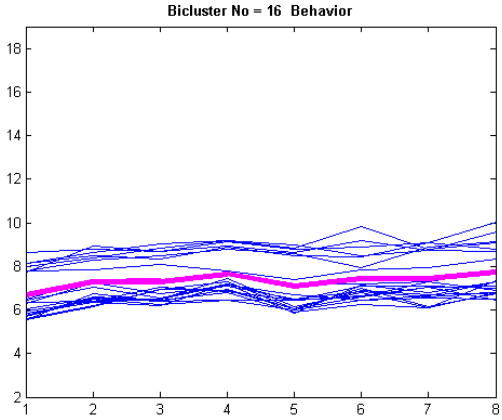

Bicluster No = 11 Behavior

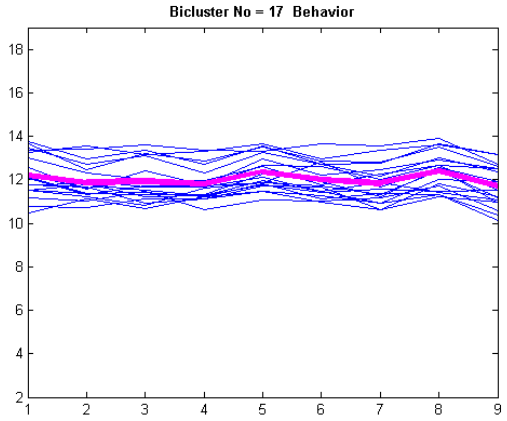

Bicluster No = 12 Behavior

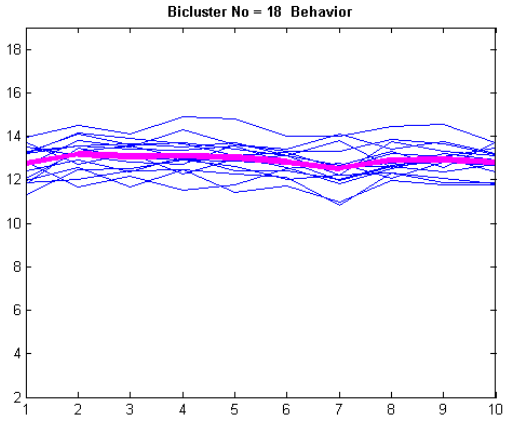

Bicluster No = 13 Behavior

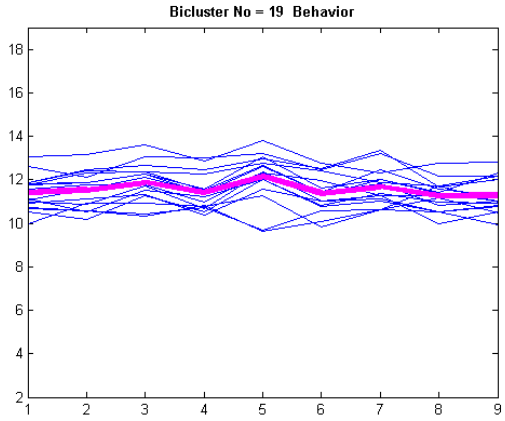

Bicluster No = 14 Behavior

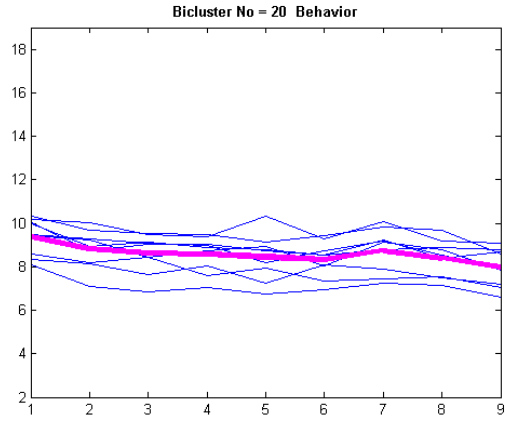

Bicluster No = 15 Behavior


Bicluster No = 16 Behavior

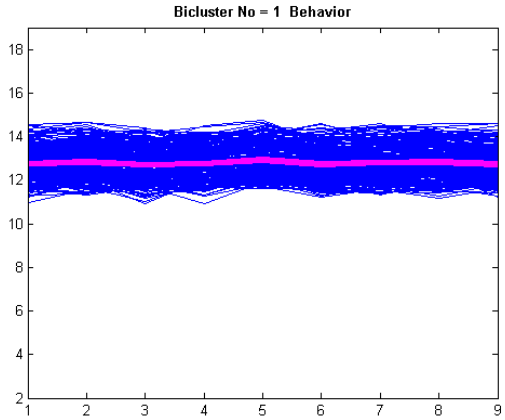

Bicluster No = 17 Behavior

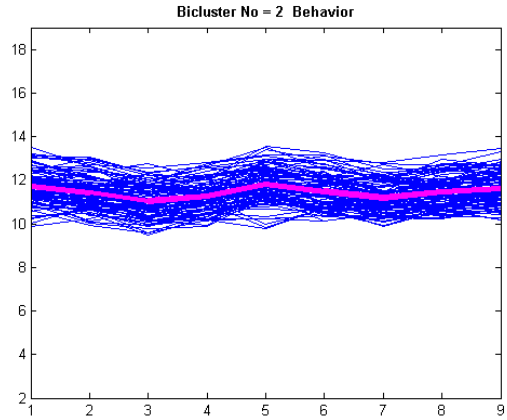

Bicluster No = 18 Behavior

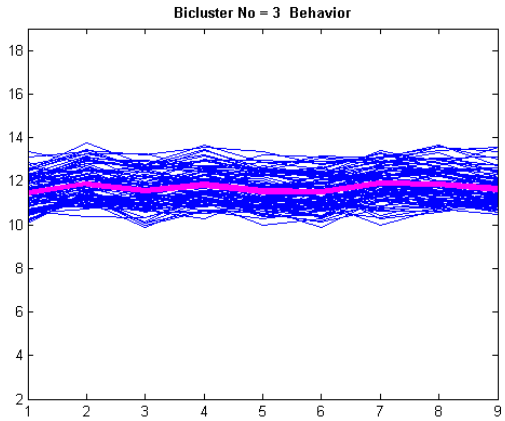


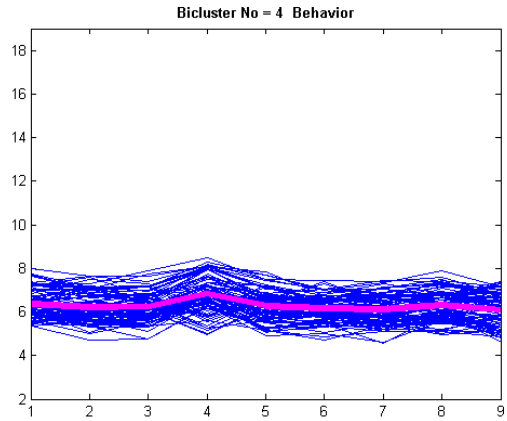


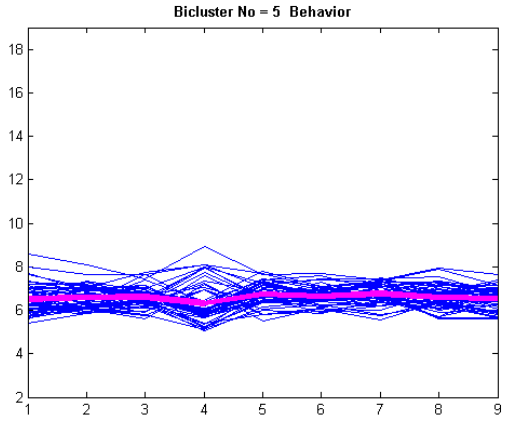
Καρκίνος του τραχήλου της μήτρας

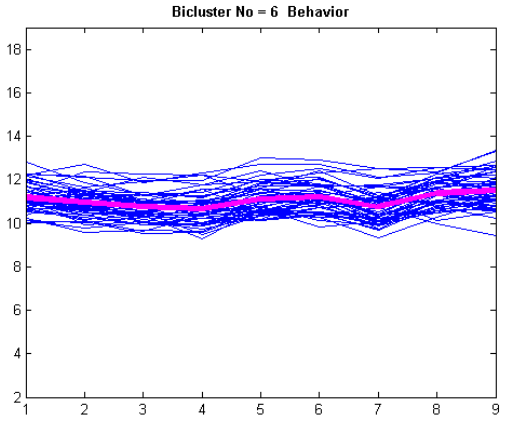


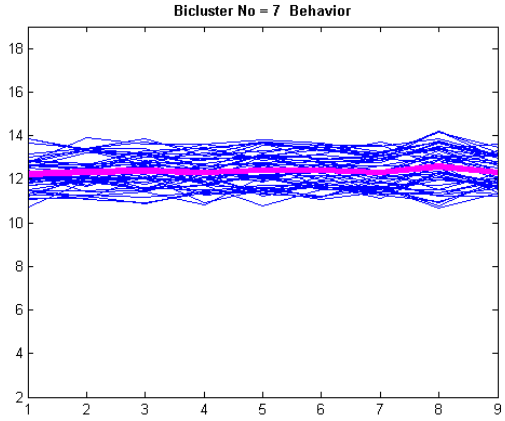


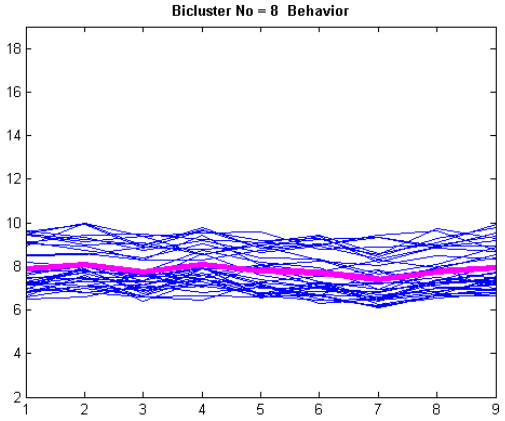


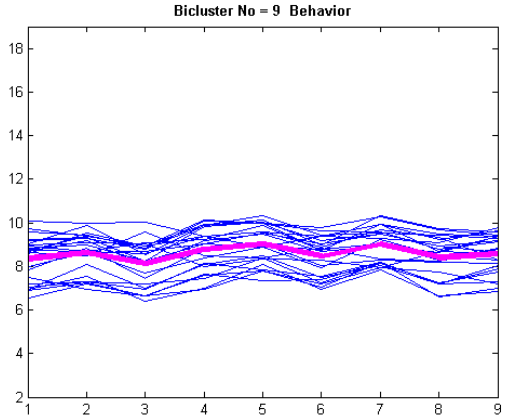


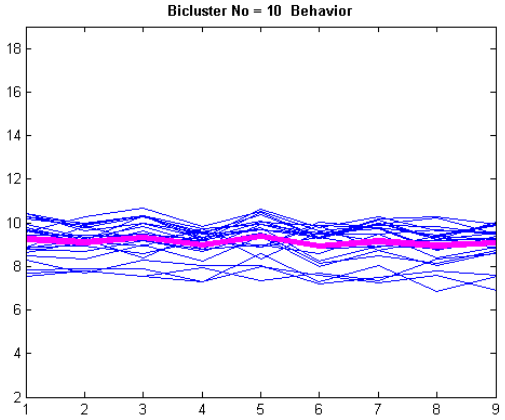

Bicluster No = 5 Behavior

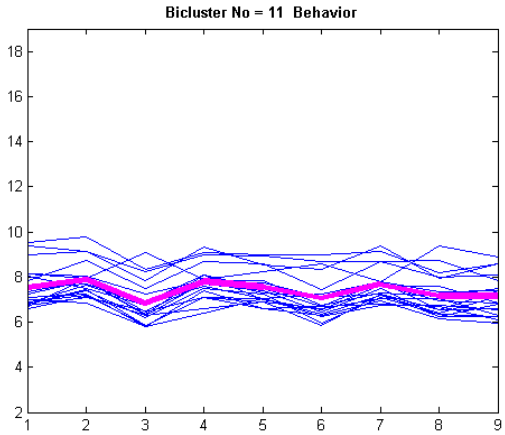

Bicluster No = 6 Behavior

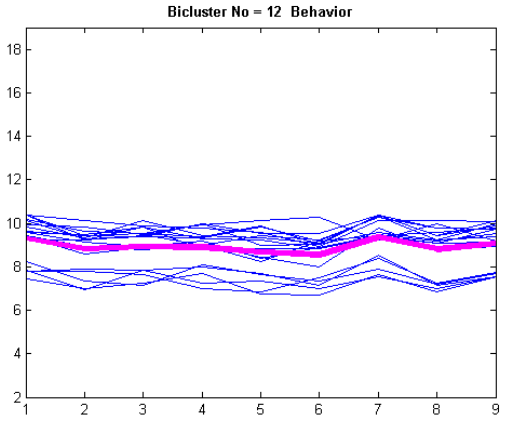

Bicluster No = 7 Behavior

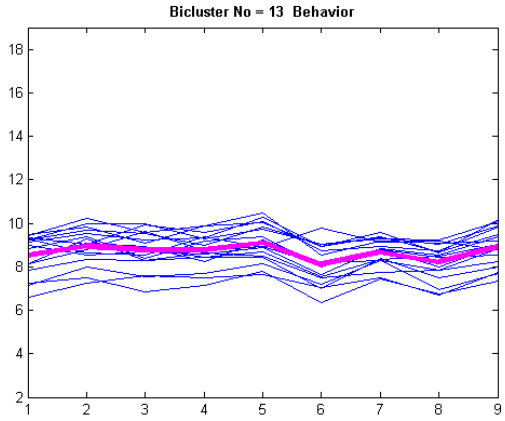

Bicluster No = 8 Behavior

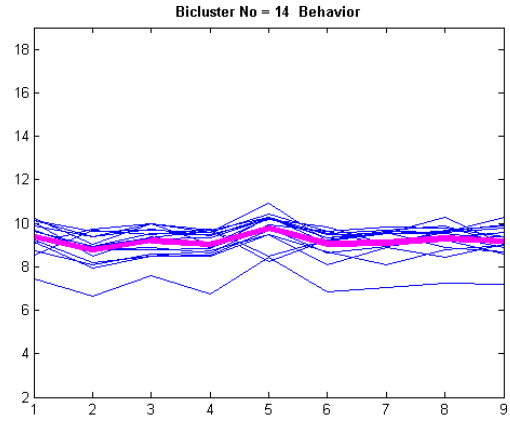

Bicluster No = 9 Behavior

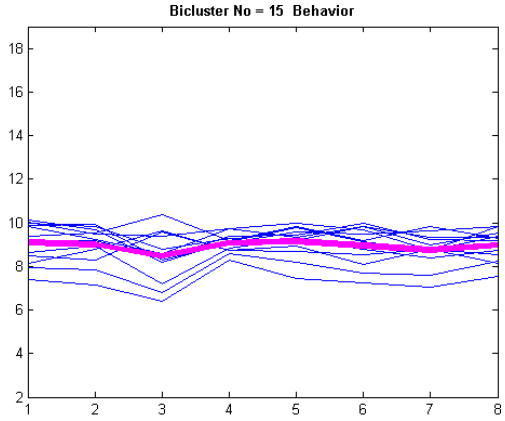

Bicluster No = 10 Behavior

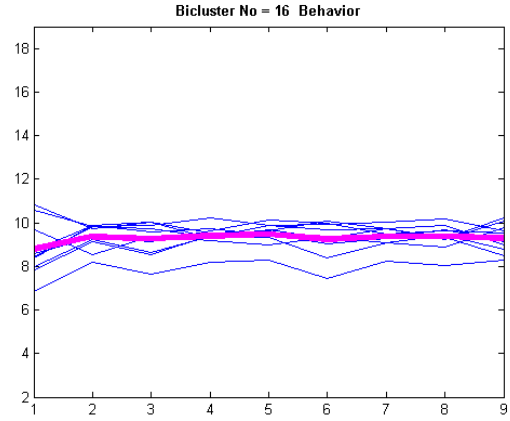

Bicluster No = 11 Behavior

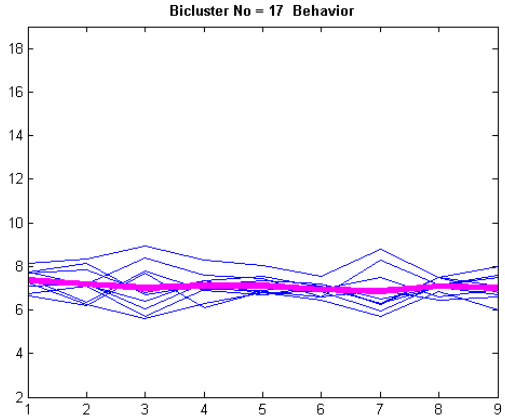

Bicluster No = 12 Behavior

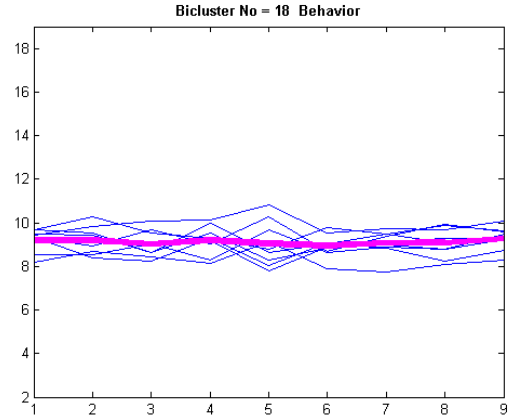

Bicluster No = 13 Behavior

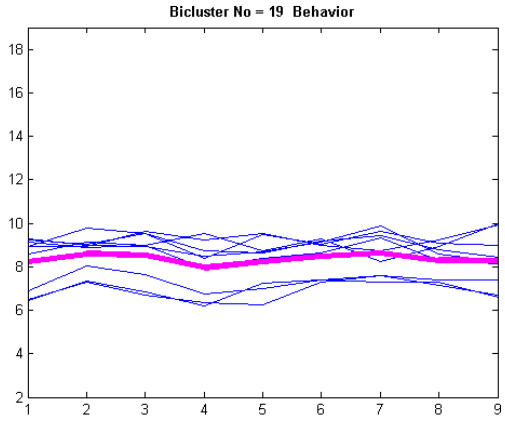

Bicluster No = 14 Behavior

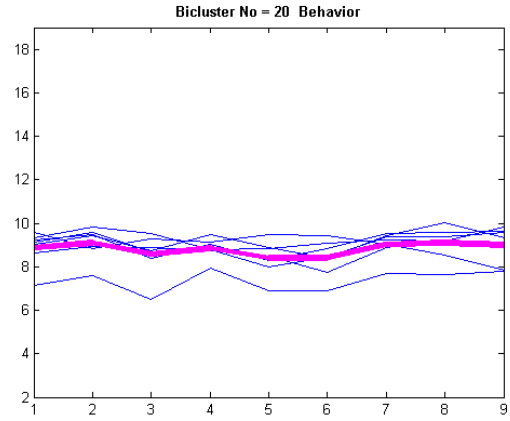

Bicluster No = 15 Behavior


Bicluster No = 16 Behavior

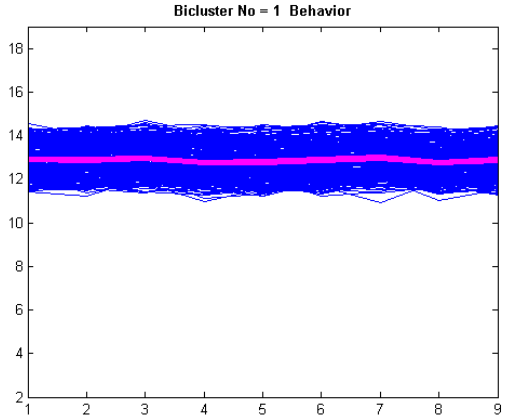

Bicluster No = 17 Behavior

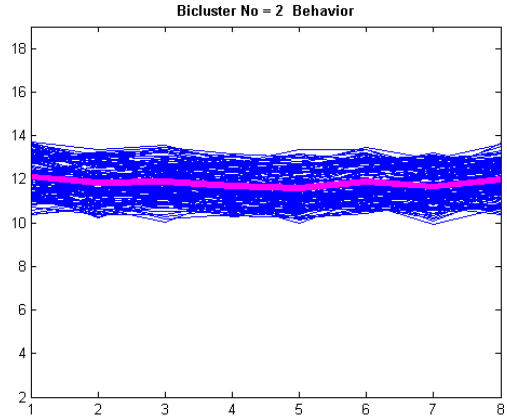

Bicluster No = 18 Behavior

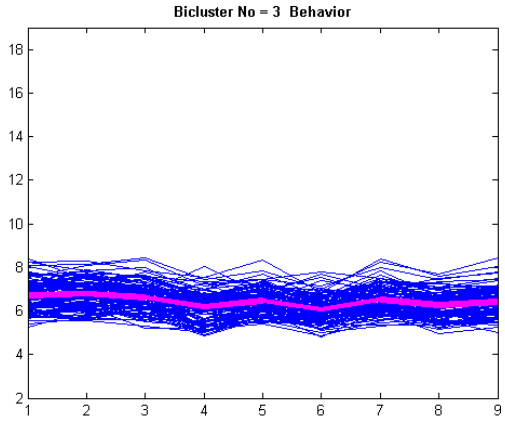

Bicluster No = 19 Behavior

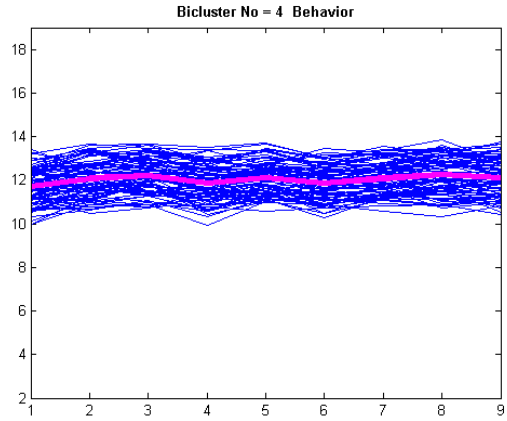

Bicluster No = 20 Behavior

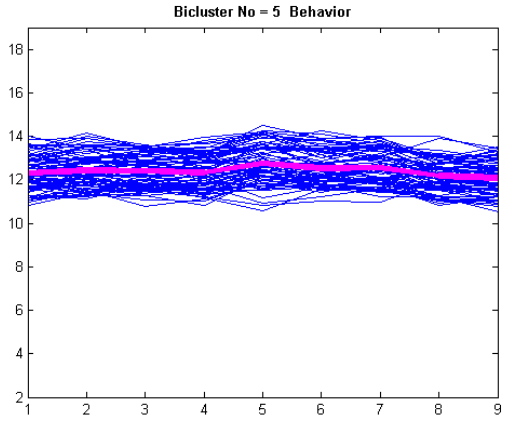
Καρκίνος του ενδομητρίου

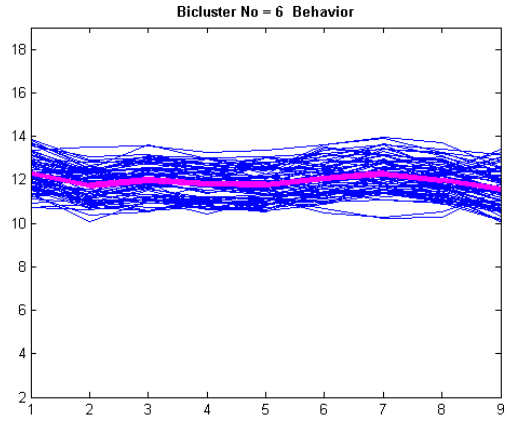


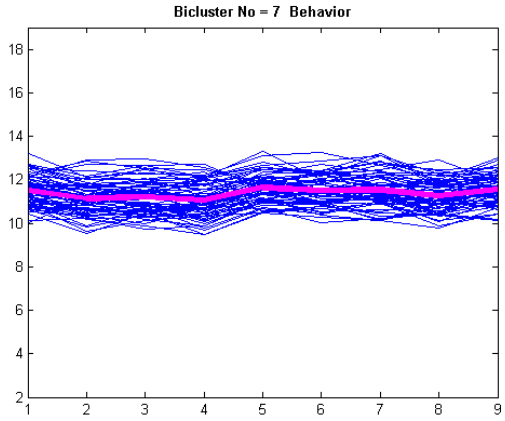


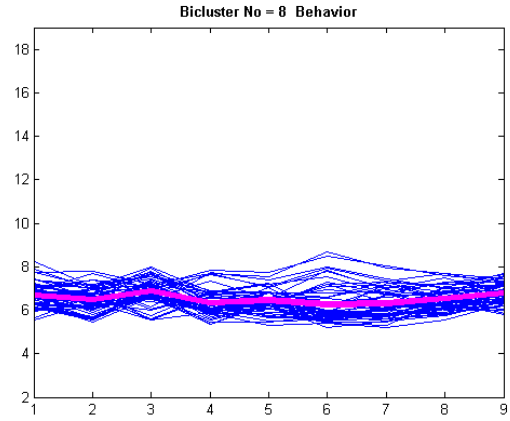


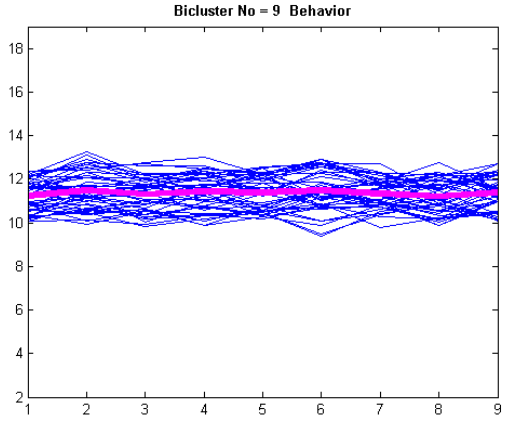


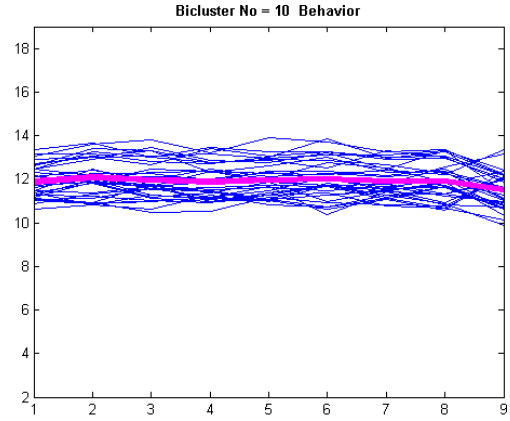


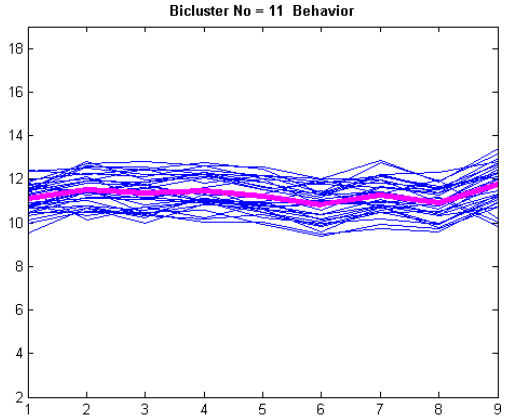


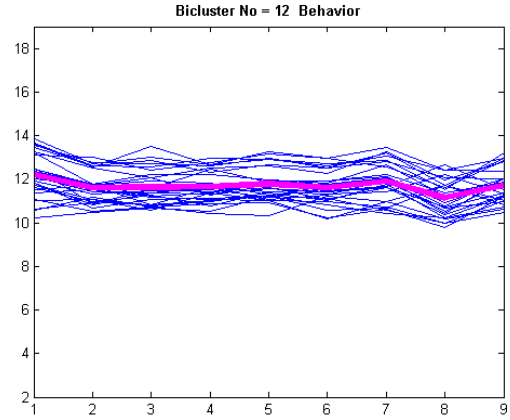

Bicluster No = 7 Behavior

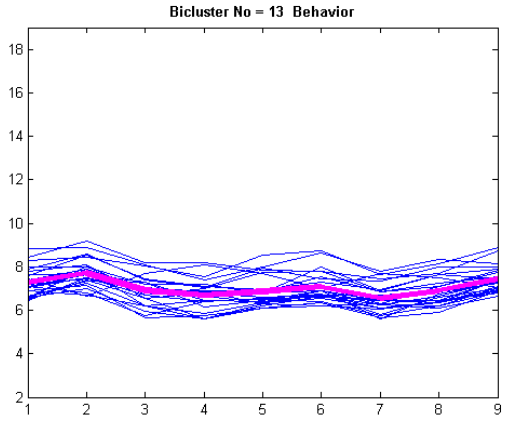

Bicluster No = 8 Behavior

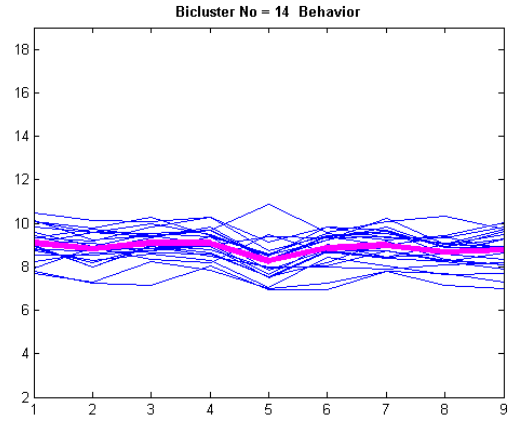

Bicluster No = 9 Behavior

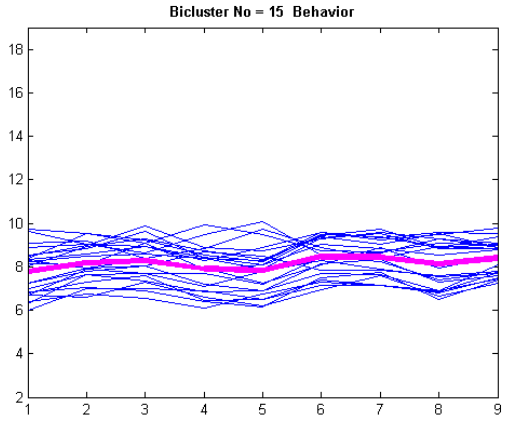

Bicluster No = 10 Behavior

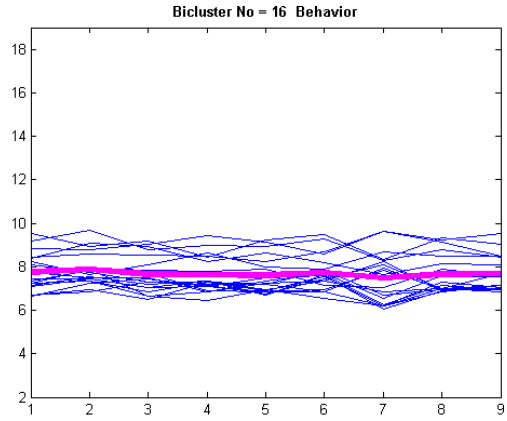

Bicluster No = 11 Behavior

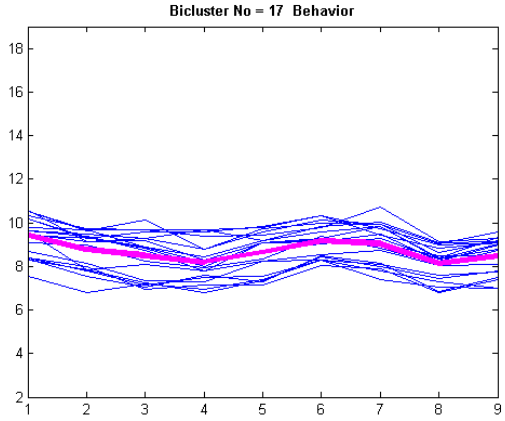

Bicluster No = 12 Behavior

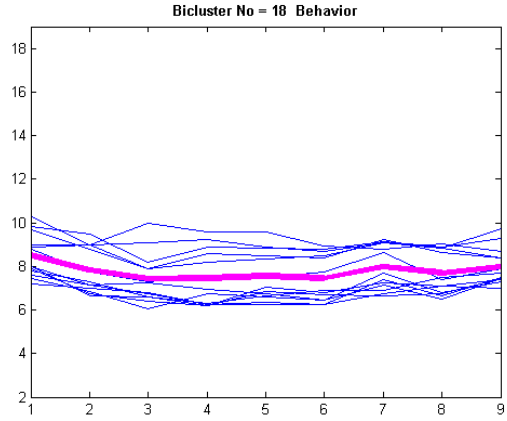

Bicluster No = 13 Behavior

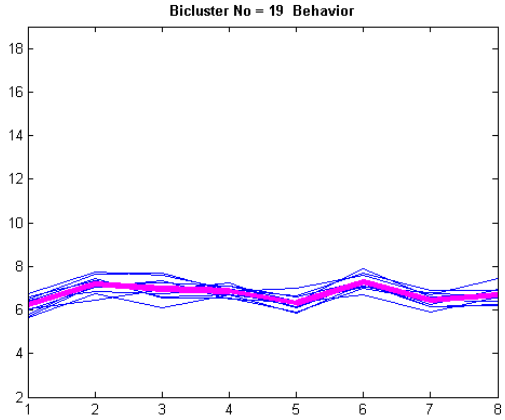

Bicluster No = 14 Behavior

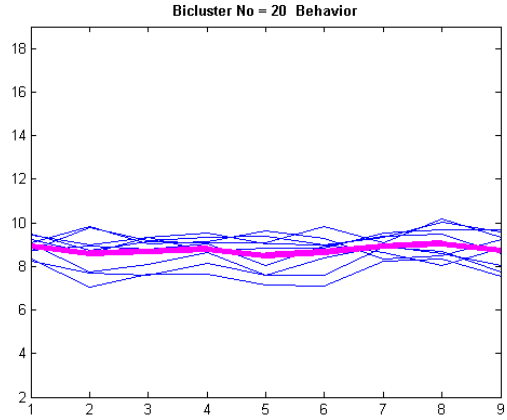

Bicluster No = 15 Behavior


Bicluster No = 16 Behavior

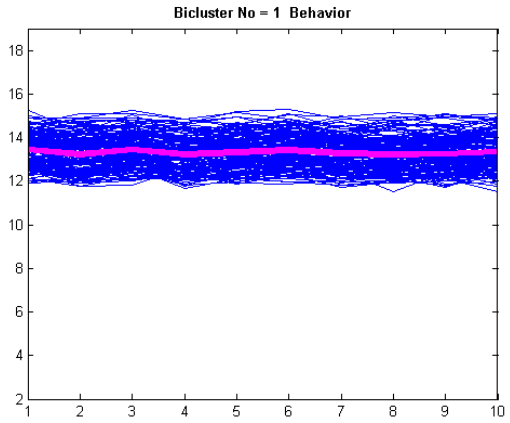

Bicluster No = 17 Behavior

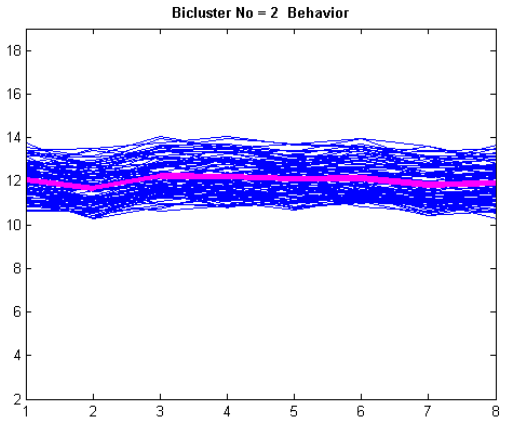

Bicluster No = 18 Behavior

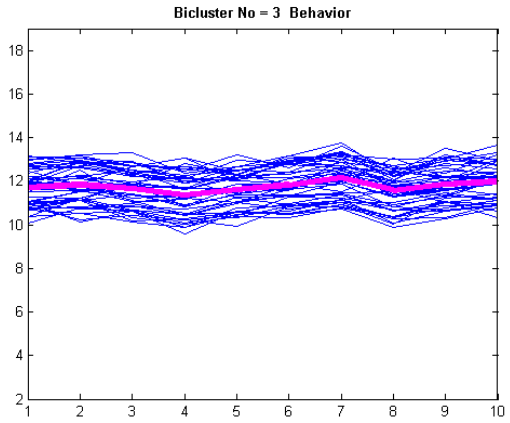

Bicluster No = 19 Behavior

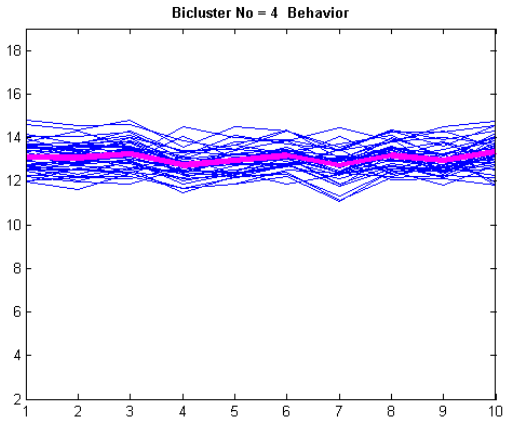

Bicluster No = 20 Behavior

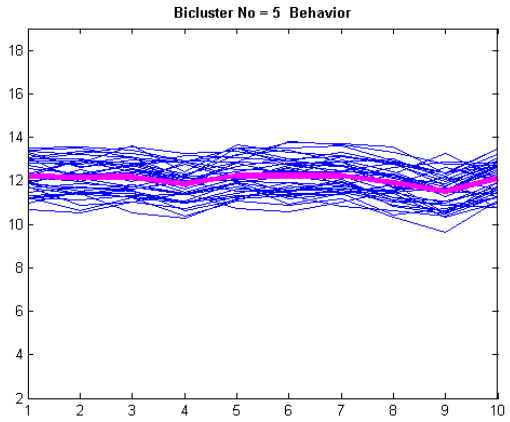
Καρκίνος των ωοθηκών

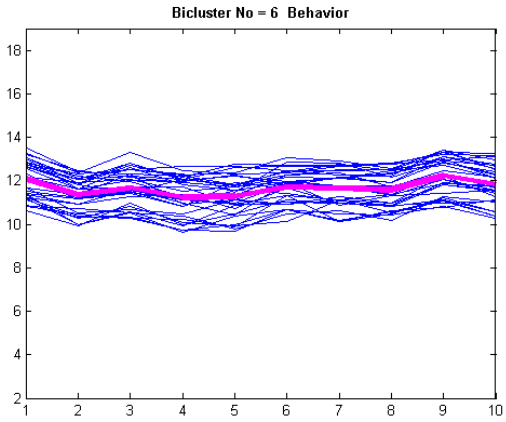


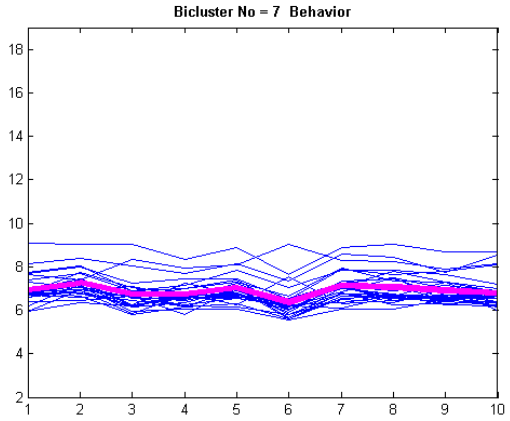


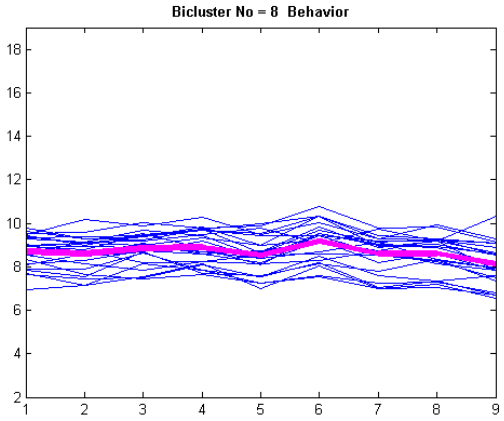


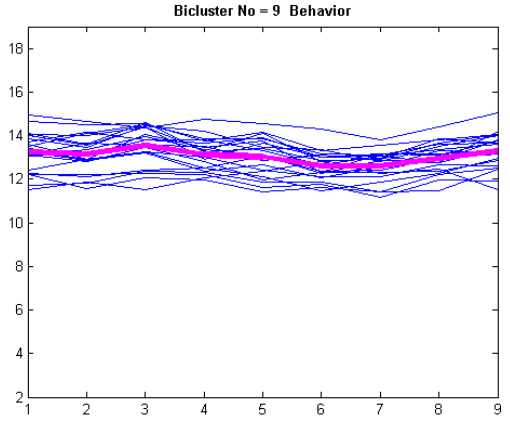


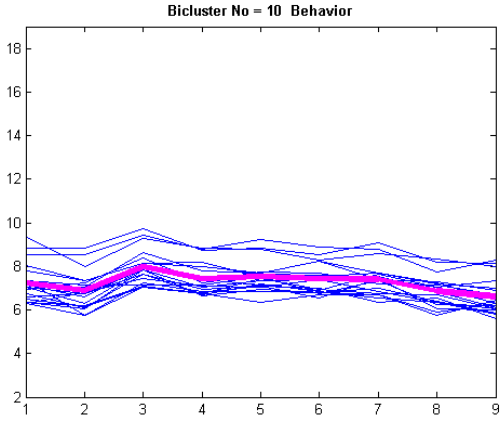


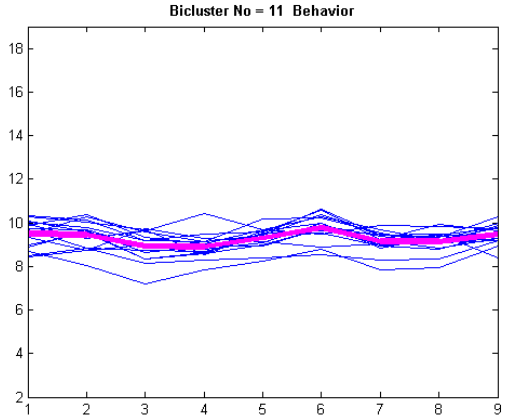


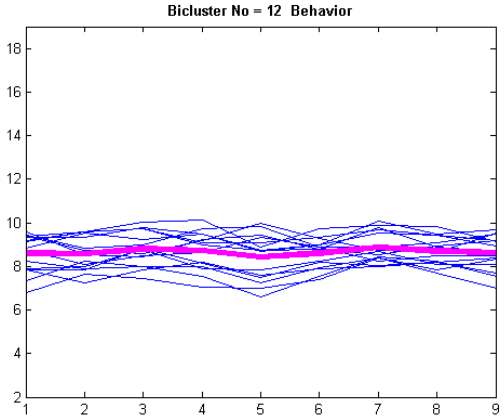

Bicluster No = 7 Behavior

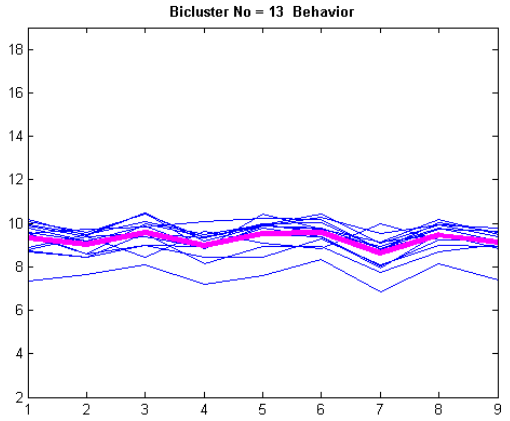

Bicluster No = 8 Behavior

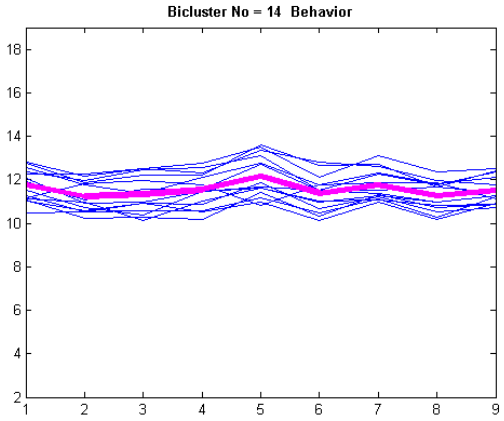

Bicluster No = 9 Behavior

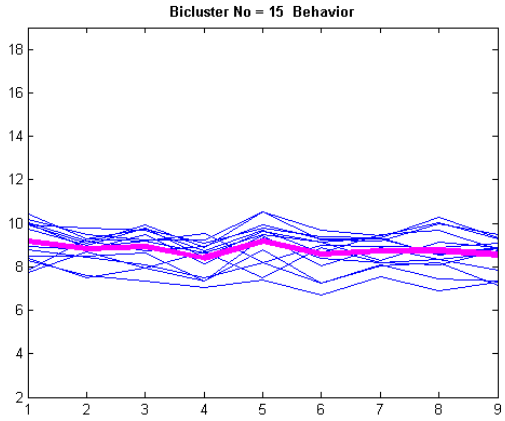

Bicluster No = 10 Behavior

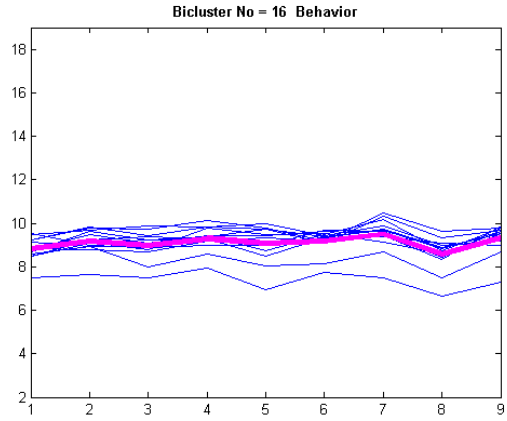

Bicluster No = 11 Behavior

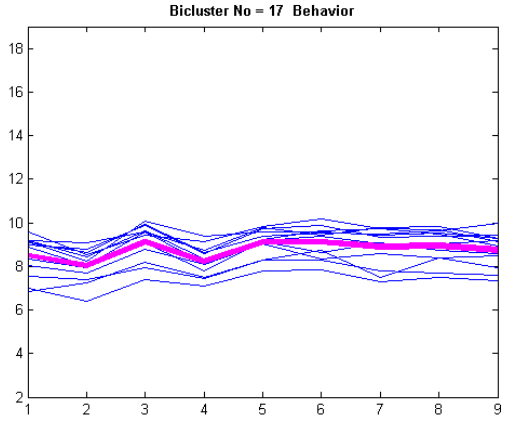

Bicluster No = 12 Behavior

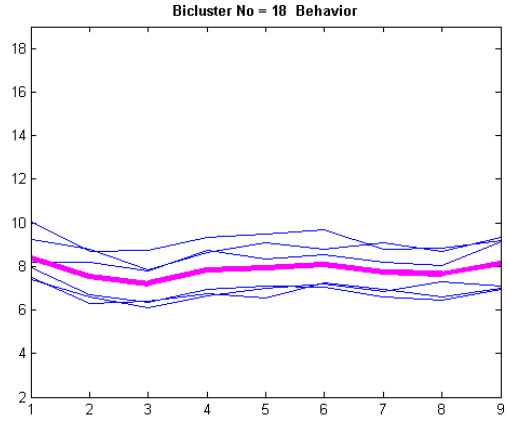

Bicluster No = 13 Behavior

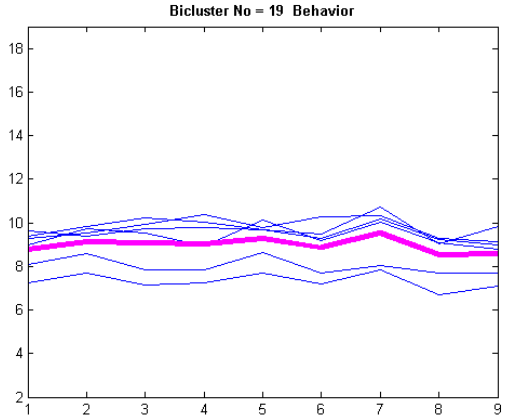

Bicluster No = 14 Behavior

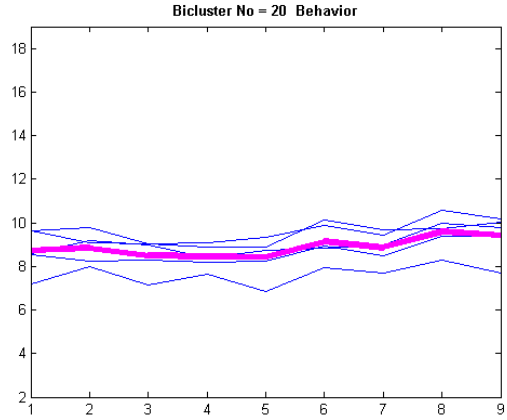

Bicluster No = 15 Behavior


Bicluster No = 16 Behavior

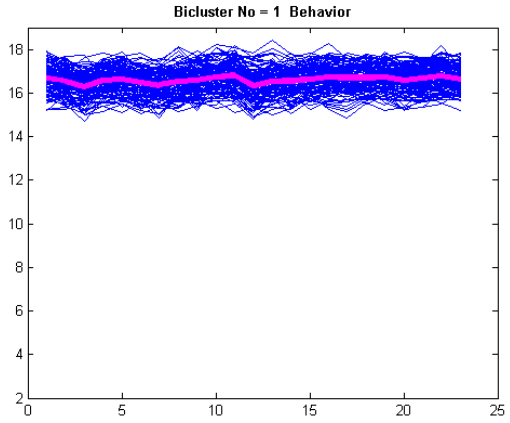

Bicluster No = 17 Behavior

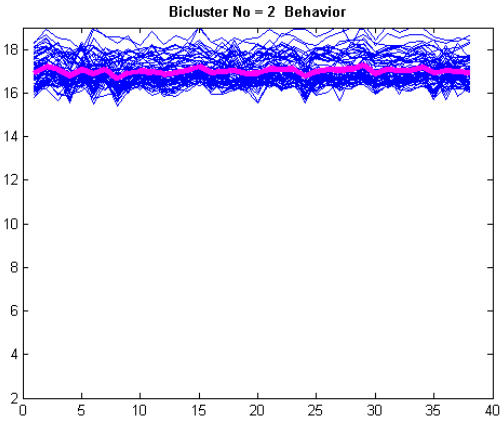

Bicluster No = 18 Behavior

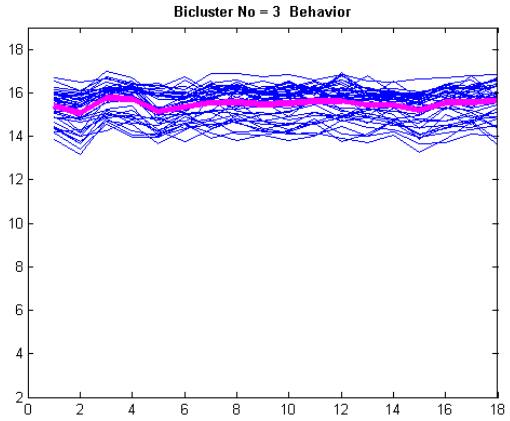

Bicluster No = 19 Behavior

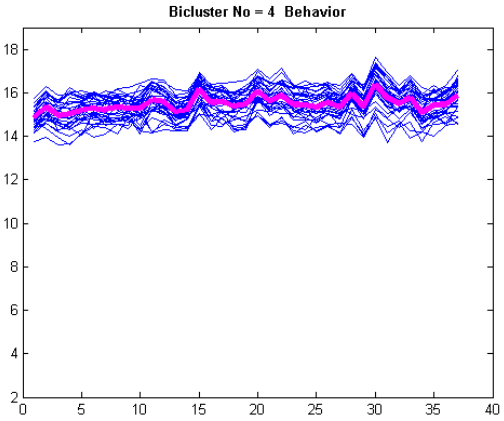

Bicluster No = 20 Behavior

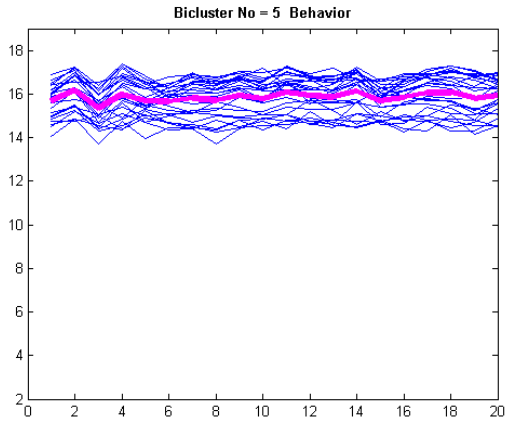
Ολική ομαδοποίηση

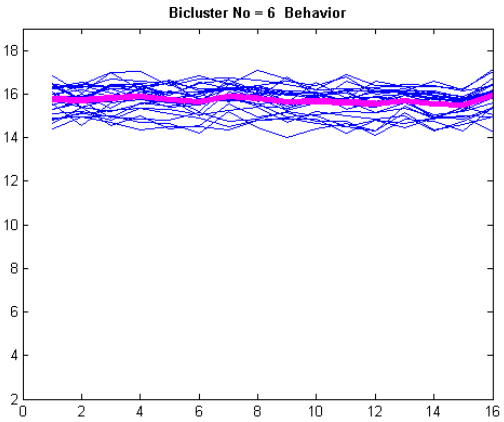


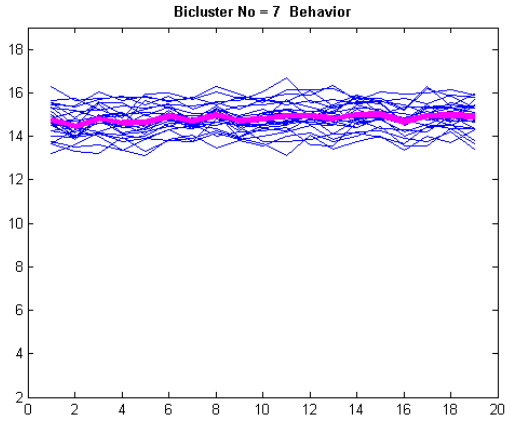


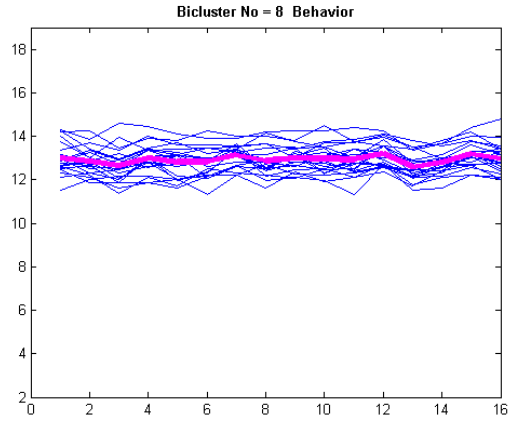


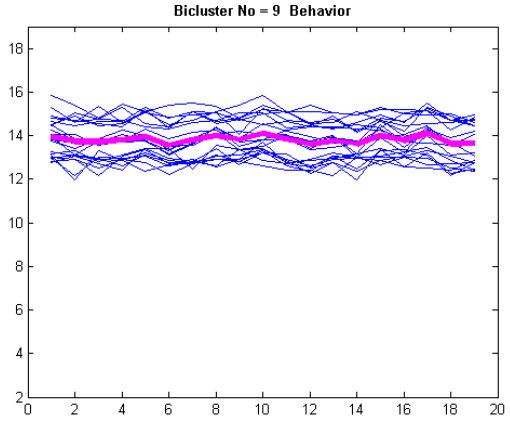


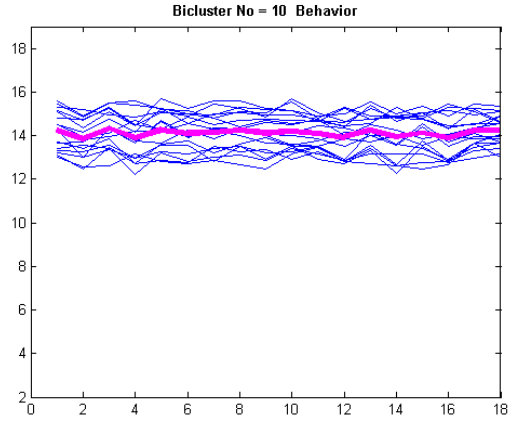


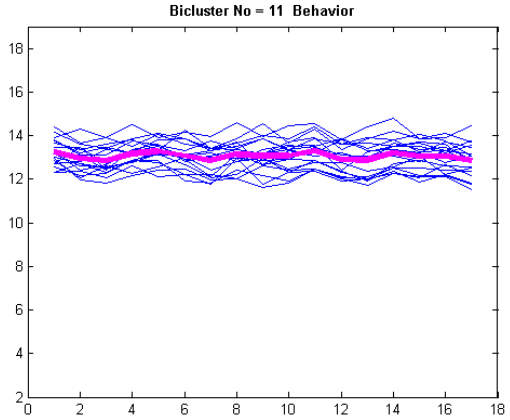


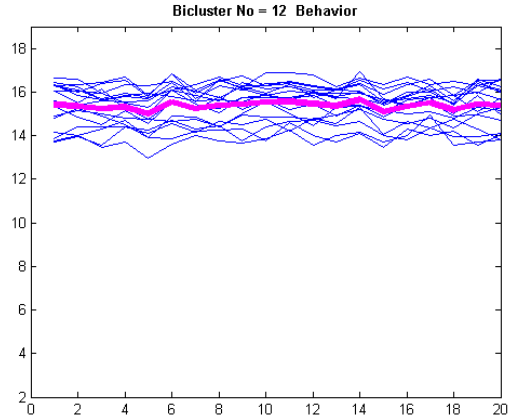

Bicluster No = 7 Behavior

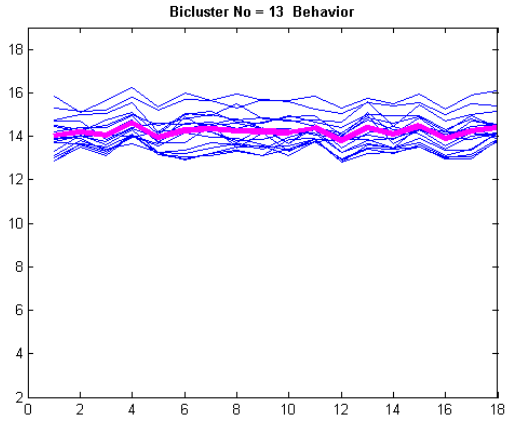

Bicluster No = 8 Behavior

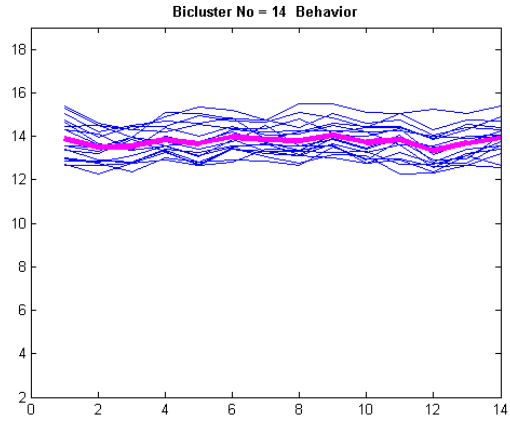

Bicluster No = 9 Behavior

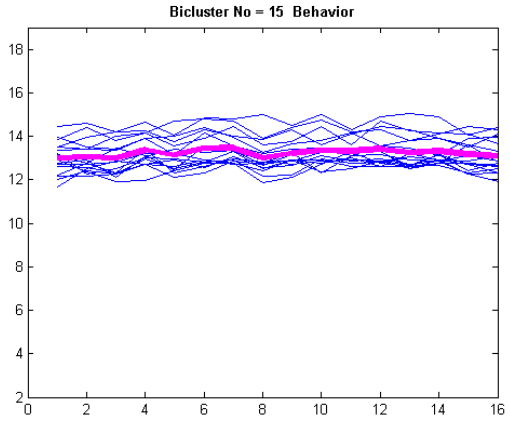

Bicluster No = 10 Behavior

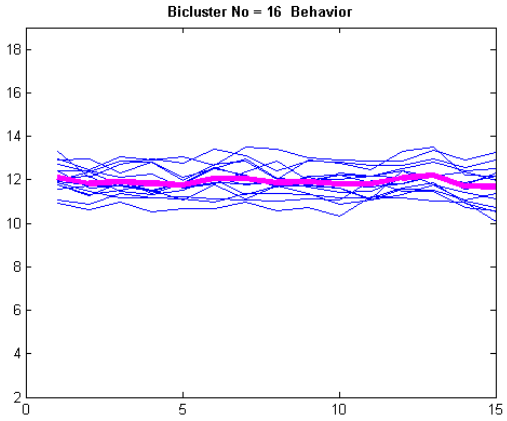

Bicluster No = 11 Behavior

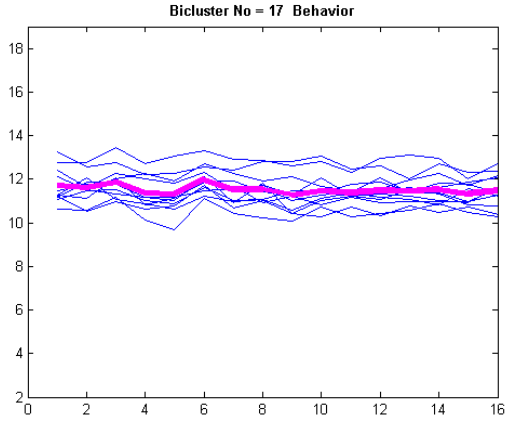

Bicluster No = 12 Behavior

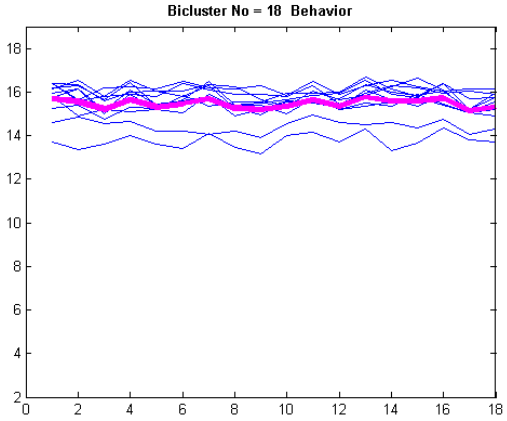

Bicluster No = 13 Behavior

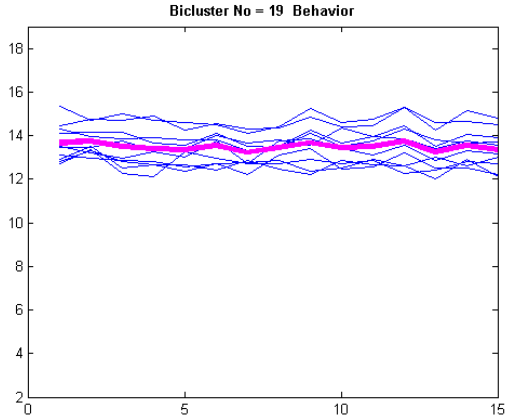

Bicluster No = 14 Behavior

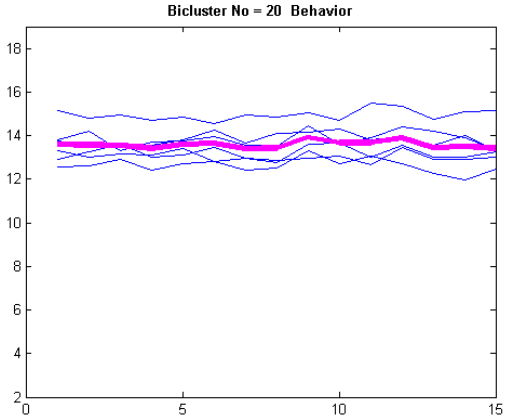

Bicluster No = 15 Behavior


Bicluster No = 16 Behavior


Bicluster No = 17 Behavior


Bicluster No = 18 Behavior


Bicluster No = 19 Behavior

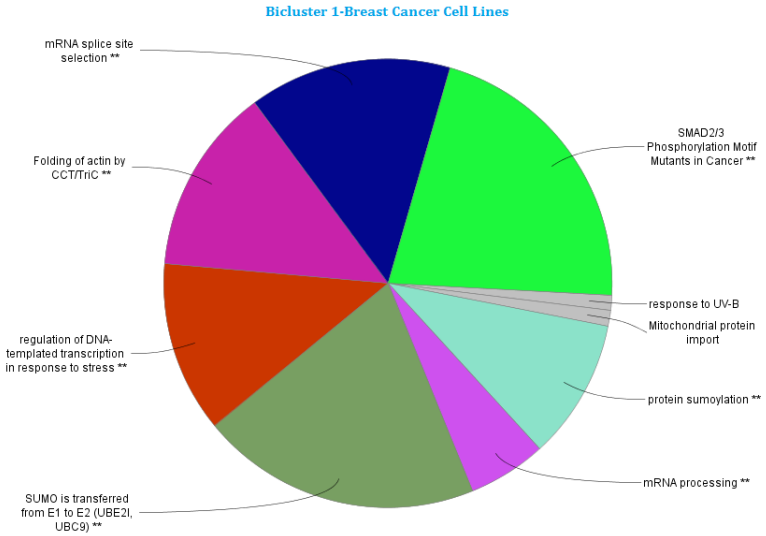

Bicluster No = 20 Behavior

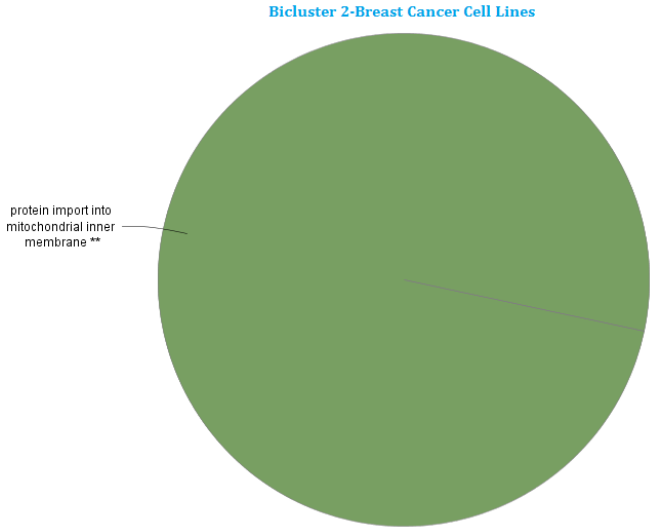
ПАРАРТНМА В

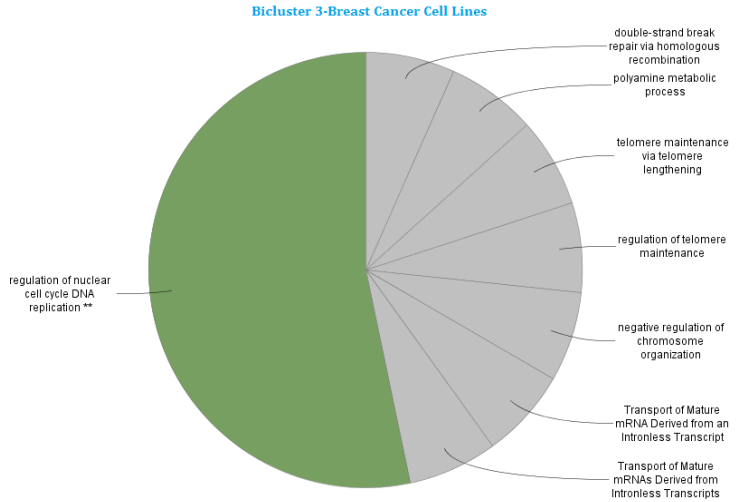
ΠΑΡΑΡΤΗΜΑ Β-1.1
ΟΜΑΔΕΣ ΓΟΝΙΔΙΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ

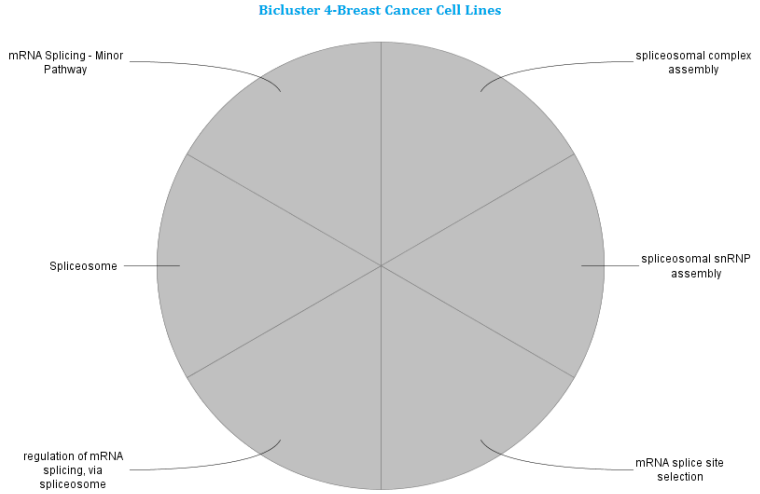
ΚΑΡΚΙΝΟΣ ΤΟΥ ΜΑΣΤΟΥ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ**
Bicluster 1	YME1L1 SNRPF UBE2I PARK7 TSEN34 CCT4 SF3B2 PA2G4 ZC3H15 RBM14 WAC EIF4E PSENEN LOC441241 CCT2 RPS15P4 DTYMK NONO HNRNPC RAB1A VDAC1 TRIM28 DHX15 NSA2 LOC390294 XPO1 BRD2 PSMD8 ARL6IP5 NDUFA11 CCT6A TSR3 UBE2D3 IK DHX9 PPP1CC SRRT SUMO1 DAD1 PAICS MSH2 CYC1 AIMP2 MTHFD1 HMGN1 RBM42 BCLAF1 TOMM40 FRG1 MATR3 PUF60 SDHB RBMX FDPS UBE2D2 HNRNPK C6orf62 SRSF3 KIAA0368 RBM39 RBMXL1
Bicluster 2	CINP IQGAP1 CDK2AP2 VTI1B SEPT2 ERP29 NDUFA13 NDUFB11 PCBP1 TIMM9 GDI2 PPP2R1A PCNP PTGES3 MRPL21 EIF5 MRPL11 ARL6IP4 TCEB2 RDBP LAMTOR1 NSA2 FAM127B ATP5G3 PSMC6 CCZ1B PSMC1P4 TRAPPC5 SSNA1 YY1 NDUFB11P1 MNF1 DDX42 SLC25A6P2
Bicluster 3	TCF25 SYF2 ATRX CDT1 TMEM97 WDR3 HNRNPA0 DBNL SLURP1 SIN3A CPSF2 POM121 TERF2IP TYW3 FUBP3 GOT1 GLE1 TRIP4 BAG5 TBC1D10B KDM3B FBXO9 USP10 MARK3 ZNF273 PRIM1 TIMM44 MNAT1 VPS28 IDE CDC16 TP53BP1 ATXN7L3 OAZ2 ZDHHC7 ZBTB11 BTBD6 ALKBH7 FAF1 THOP1 PTOV1
Bicluster 4	HDGF SF3B1 BSG AES ZNF16 POP4 PABPC4 LAMTOR5 PSMD1 YLPM1 ZNF440 C22orf28 SCRN1 TRA2B ALDH7A1 HNRNPU STRAP ZFPL1 MSH6 SF3A1 SRSF10 SNRNP200 ZNF124 METTL5 VDAC1 ISCA1 SRA1 HNRNPA3P4 HNRNPA1P30 ASNA1 ZNF224 ZNF45 HIGD2A
Bicluster 5	SBDSP1 PRPF31 DCTN2 ZNF765 RNF115 SAE1 NEMF OSTCP1 BRE EXOSC7 COPS4 HIGD1A CAPZA1 BTF3L4 MRPL54 KLHL29 EIF4E TARDBPP2 MRPL37 CLPTM1 TM9SF1 ZNF699 PUM1 MARCH5 RAB5A MRPS18A UBE2K DDX50 SERINC1 MICU1 VAPA PPP1CB OCIAD1 SREK1IP1 VPS29
Bicluster 6	ZNF706 NOSIP CSNK1A1 UBL5 TTC1 SNRPC SNRPD1 UBE2NL DNAJB6 MRPL34 TRA2B TMED9 TBCA RPL5P18 CLTB CAMTA1 MINOS1 PSMB6 LARS LRPAP1 PSMC3 UQCRC1 TUBA3C SMN2 SURF4 DDOST ARL6IP5 HDAC1 SNRPG RUVBL2 ARCN1 LARP1 CDKN3
Bicluster 7	NUP188 TM9SF4 WDR83OS ACIN1 GPSM2 LOC220906 EZH2 TJP2 PTPN2 RBM23 ZNF131 DUSP12 CDK16 TINF2 PRDM4 ORAI1 NAA35 LYPLA2 DDX24 RFXANK SRSF5 FAM96B APTX EXOSC10 SUGP1 TOM1L2 AXIN1 CDK9 SLU7 VPS4A CTBP1 PABPN1
Bicluster 8	ARHGAP11A ZNF207 MKI67IP NOB1 WDR74 PSMD10 NIT2 BRIX1 CLINT1 METTL3 NPLOC4 ATG3 RNF167 ETF1 MAP3K7 NTMT1 NSUN2 MAGOH UBAC1 SUPT5H EXOC5 YTHDF2 GLOD4 RARS TRAPPC3 MPDU1 CRK ZBTB8OS ACAD9 OAZ2 NDUFAF4
Bicluster 9	DCTN3 DRG1 RFC4 UBXN1 UBE2L3 RNPS1 PDCL MAPKAP1 MRPS18B MYL6B MTMR14 PRDX4 SNW1 PCNA UTP23 C21orf59 CDCA5 RAD23B PPP6C NABP2 MAP3K11 RBM3 CCNY HK1 ACTL6A HNRNPM
Bicluster 10	PPIF GARS IMMT OLA1 PSMA1 GADD45GIP1 DNAJB1 YIPF6 HMOX2 RPS9 SRSF1 MRPL4 UQCRFS1 BOLA3 NGDN FPGS LOC441806 FBLL1 COPS6 NDUFB7 ATRAID C12orf45 AUP1 TIMM50
Bicluster 11	NIF3L1 CENPA NDUFS1 UGP2 DDX18 UBXN6 USP39 GADD45GIP1 SNX17 GNL2 NR1H2 TPD52L2 AGPAT2 TPRKB CEBPZ PSMD14 TFPT CCT7 DGUOK MTX2 PDCL3
Bicluster 12	STAU1 TRAF7 CNOT1 TSTA3 UBE2Q1 LOC441455 DPP8 ICMT SLC25A44 RHNO1 GMNN COPZ1 ZNF24 ALYREF GDE1 RAD51 GOLGA5 RBM34 CS GSPT1 SNAPIN
Bicluster 13	BCKDK SNRNP40 ZNF587 RAF1 PNRC2 ZNF345 CYB5B WDR6 UBR5 ZNF433 RNF220 CWC27 PGM3 MGEA5 PSMB2 SRP54 ALDH7A1 TMEM9B TAF9
Bicluster 14	PGLS PTBP1 CDCA8 UBE2A GMFB SRP19 ANKIB1 AKIRIN1 NUDT1 RAB8A CCNA2 UBAP2 SF3B4 HNRNPH2 MKRN1 NUP107 DHFR HMGB2 SLC35C2 VRK1 TOP1P1
Bicluster 15	ENOPH1 HAX1 SUPT6H C1orf112 SF3A3 VTI1B MRPS34 NUCB2 RFC2 SDHA SPAG5 PTPLAD1 TFG TSPAN14 PSMC1 NUBP2
Bicluster 16	OR52K3P LEFTY2 ARFRP1 EFCAB4B C2orf18 METTL22 FAM54B LOC442075 C19orf52 C19orf12 TSTD2 TTC25
Bicluster 17	MRPS15 DYNLT1 PCMT1 UBE2J2 UFM1 PDIA4 LOC442060 EMC4 FOXJ3 RCC1 SS18L2 UCHL5 ZDHHC20 LMAN2 PDXK RBM39 MRPL18
Bicluster 18	RPF1 GLRX5 LOC441601 SAE1 ESYT1 NDUFA7 SBDS C6orf48 HNRNPR PTPLAD1 SSB POLR3B CWC15 MINOS1
Bicluster 19	VDAC2 LOC441488 ZNF833P DPPA3 ZNF540 CTNNBL1 NCOA4 PHF14 APIP RNF14 FAM120A MRPL32 EMC7 MALSU1
Bicluster 20	CLK4 SLC25A38P1 ZNF319 EIF2C3

ΠΑΡΑΡΤΗΜΑ Β-1. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ

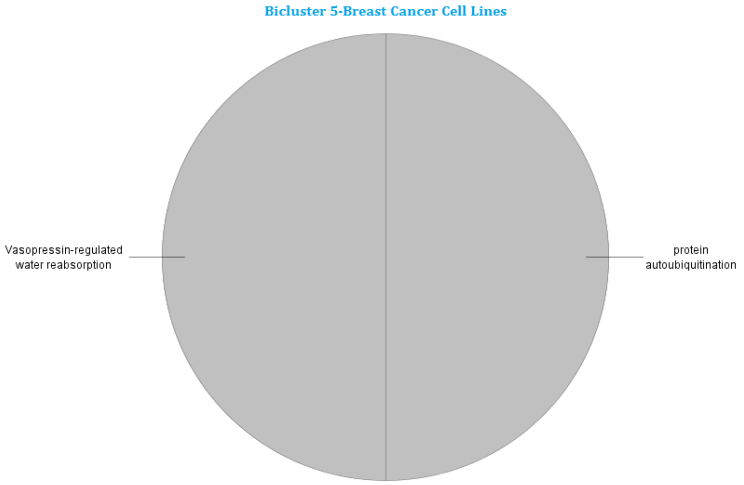


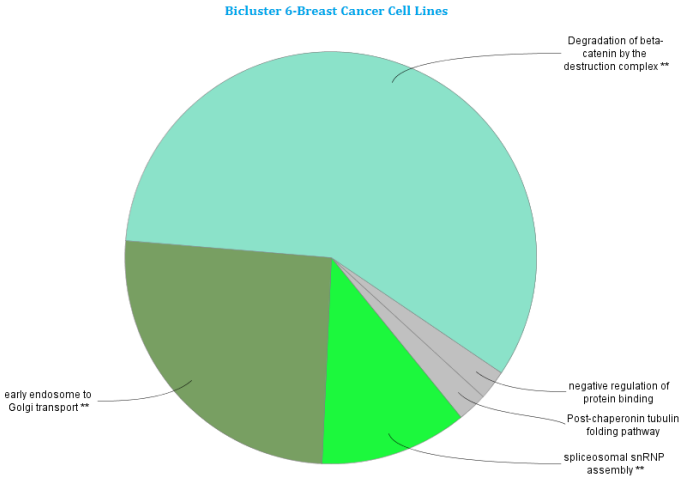


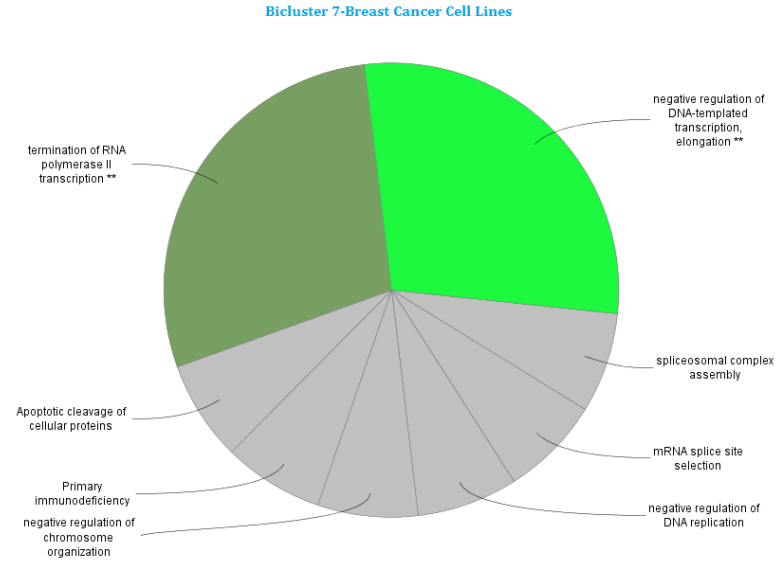




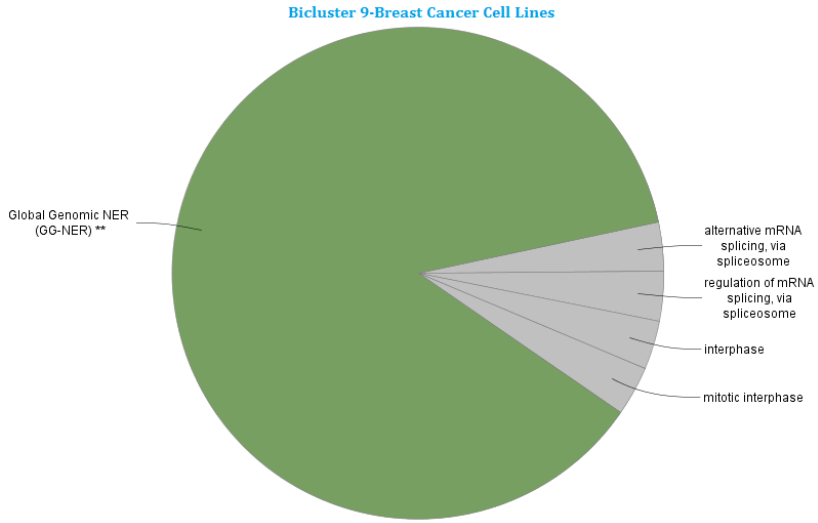
ΠΑΡΑΡΤΗΜΑ Β-1. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ
(συνέχεια)

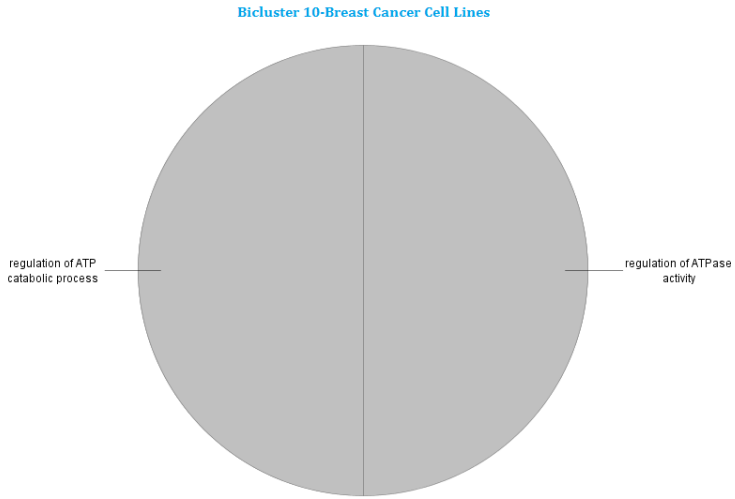


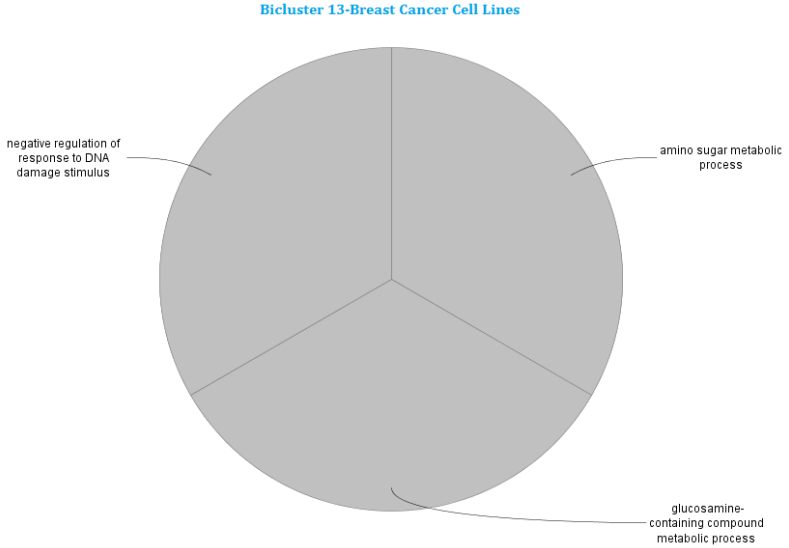


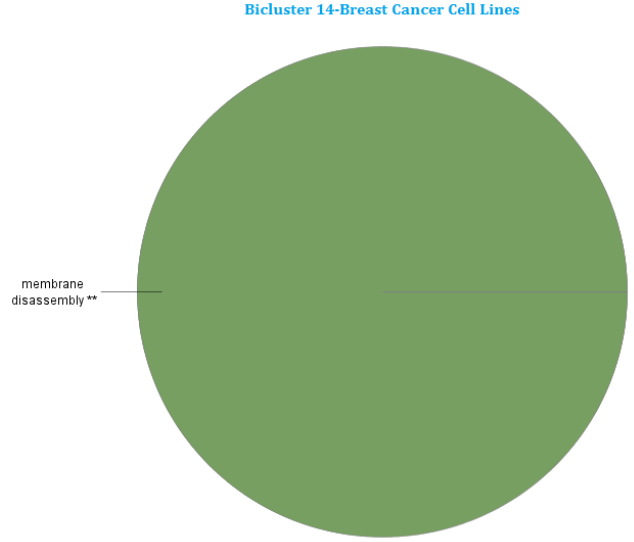


ΠΑΡΑΡΤΗΜΑ Β-1. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ
(συνέχεια)

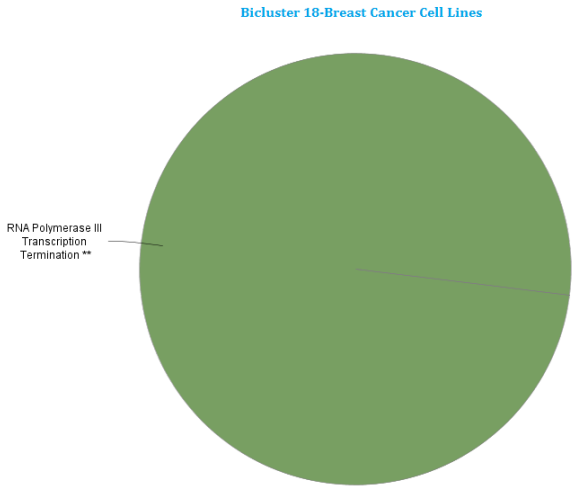








ΠΑΡΑΡΤΗΜΑ Β-1. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ (συνέχεια)



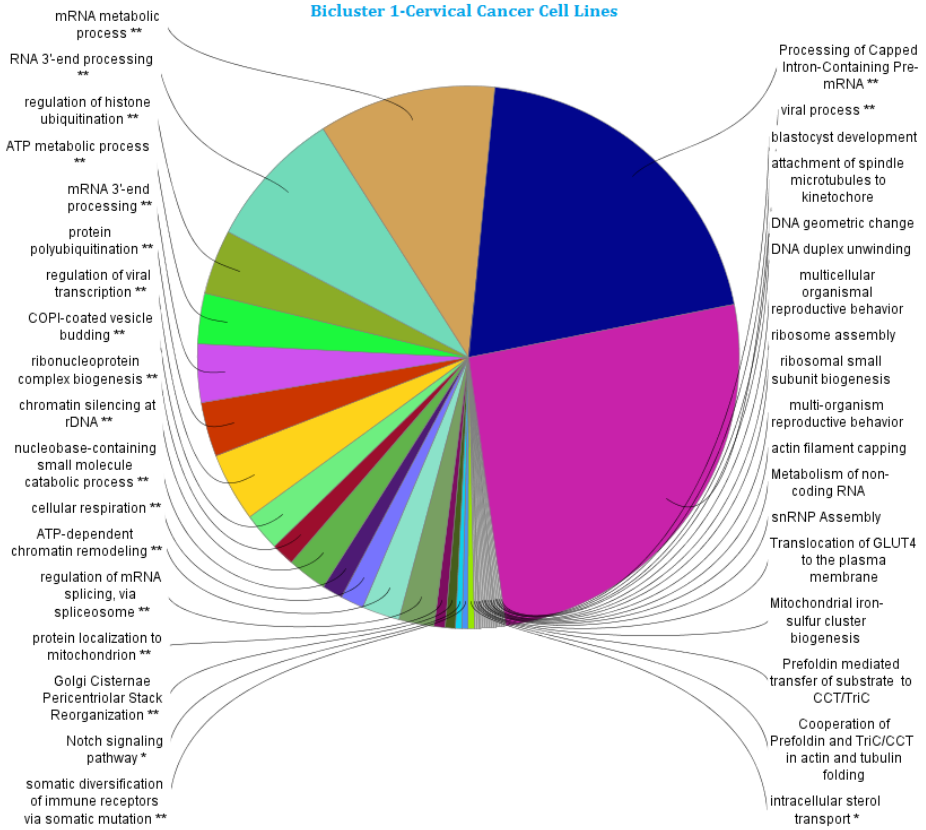
ΠΑΡΑΡΤΗΜΑ Β-1.3
ΕΝΤΟΠΙΣΜΟΣ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΤΟΥ ΜΑΣΤΟΥ

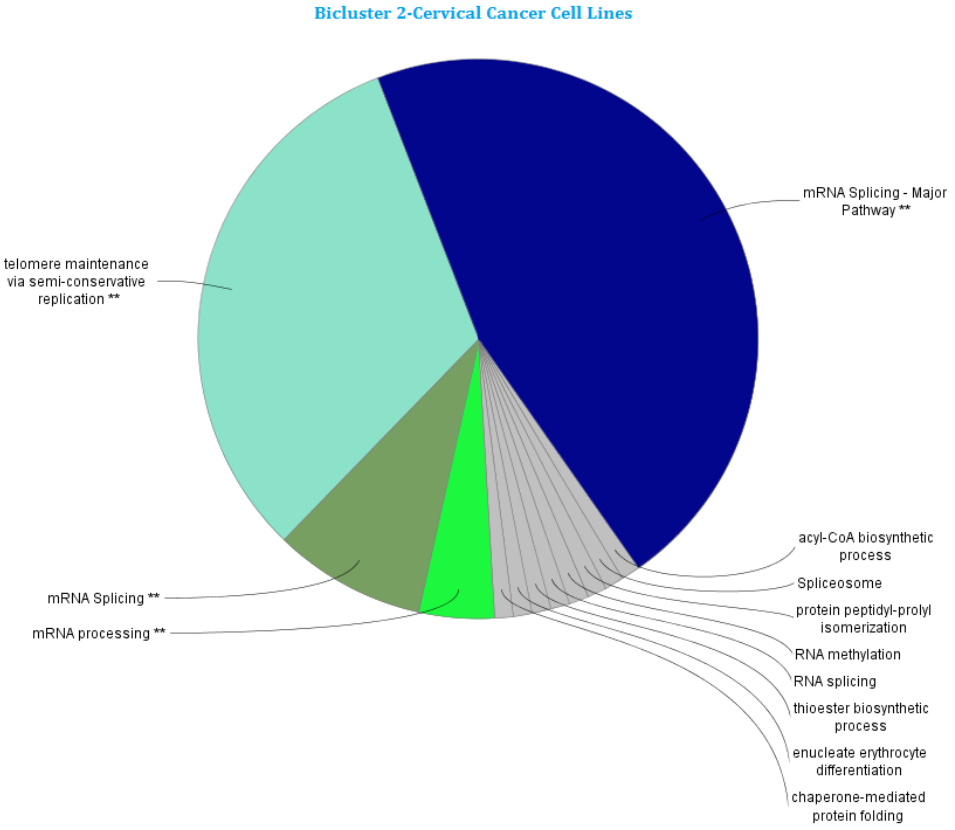
ΚΑΡΚΙΝΟΣ ΤΟΥ ΜΑΣΤΟΥ		**ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (20)**																			
Αριθμός Κυτταρικών Σειρών	Καρκινικές Κυτταρικές Σειρές	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	MCF-7																				
2	MM231																				
3	T47D																				
4	Hs578T																				
5	SKBR3																				
6	MM435s																				
7	ZR75-1																				
8	BT-549																				
9	MM453																				
10	BT474																				

ΠΑΡΑΡΤΗΜΑ Β-2.1
ΟΜΑΔΕΣ ΓΟΝΙΔΙΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ

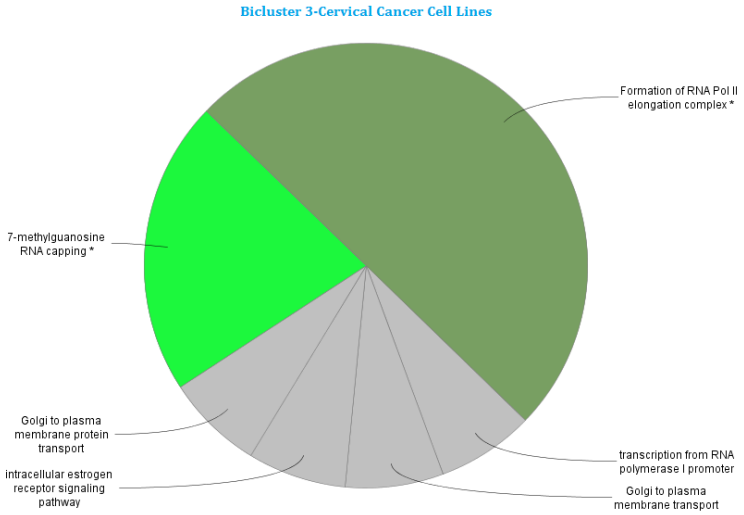
ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ**
Bicluster 1	PDCD6IP PPCS SNRNP40 HNRNPAB POLR2F MRPS15 WDR83OS UQCR10 HEXB RBM8A SNRPD2P1 RAB2A CTBP2 GPBP1 UBE2I SNRPD3 SF3B5 RAB8A OTUB1 SRSF2 SNRPD1 RBM14 UBE2NL UBE2T EIF4A3 C11orf10 DCAF12 ANAPC11 PRR13 ISCU RER1 CTTN TCEA1 NDUFA6 TFG EIF4E MINOS1 CENPA GRPEL1 PIPSL FH WDR82 PAF1 BOLA3 LOC390294 SMN2 FDX1L PSMD8 CCDC59 ZNHIT3 UBE2D3 DRG1 NDUFV2 NACAP1 MRPL9 PRPF31 DHX9 MRPL22 SEPT2 RNASEH1 RPN1 NUS1 EIF3I LSM5 C4orf3 GDI2 MBD2 EFTUD2 EIF3IP1 CAPZA1 NEK2 SRSF7 PRELID1 SRP54 TOMM22 LUC7L2 TMED2 DNAJC7 KXD1 POLR2I WDR1 DYNLT1 HDAC3 SF3B14 PPP4C SCP2 PSMA1 DIABLO PSMG2 GORASP2 PRRC1 EIF2A TRMT112 TOMM40 CCT8 MATR3 PSMC4 MNAT1 SYNCRIP RBM25 RBMX LSM3 LAMTOR4 COX7A2L SLC16A1 TMX2 C7orf50 RUVBL2 SMARCA5 PLIN3 HNRPDL TMEM14A HIST1H3I ZNF207 NDUFA5 TRAF7 YME1L1 PSMA2 DDX18 MDH2 OLA1 PMS2 TM9SF3 PTGR1 POP5 ZMAT2 SNRPC SPG21 UTP18 NUP133 PPP6C TARDBP SSBP1 TALDO1 DDX1 PSENEN COPB1 ENY2 HNRNPC MAGOH PCMT1 MARCH5 YWHAQ GLO1 GTF2A2 RAB1A VDAC1 ERAL1 RAB10 SRSF1 TIAL1 NSA2 AMZ2 NUDT1 TSR3 SREK1IP1 UBE2N COPG1 FAM120A IK JKAMP BABAM1 THOC7 SNX17 SDHAF2 MRPS34 ARF1 TBL2 RRP36 EMC4 SNW1 UTP6 ABT1 BTF3L4 H2AFY FUBP3 ANP32A PCNP NUDT5 MTX2 TMX1 PSMB6 GTF3C6 EIF4B HNRNPKP1 UBE2L3 UQCRC1 UBE2K KDELR2 FADD PFDN1 MFN2 LOC138864 TPRKB CCZ1B HNRNPH2 UNG UBE2D2 TMEM248 APOA1BP HNRNPK ASNA1 SRSF3 CAPZB PGAM1 RBMXL1 CWC15 SNAPIN
Bicluster 2	IER2 ST13 SUZ12 SLC35B2 MEAF6 RPP30 BRE MAEA RP9 TRA2B MEPCE MYCBP ZNF3 SMARCE1 DLD ZDHHC24 TACC3 SFPQ CDC5L COBRA1 FTSJ2 SP3 SMUG1 SNRNP200 HMGN4 PTBP1 C12orf65 SNRNP27 WBSCR16 BFAR RANBP2 DDX46 ELAVL1 RABL2B UBE2E3 TNIP2 PRPF4B MCM6 FAM168B TOR3A DPF2 MRPS26 WDR26 EIF2AK4 PDCD7 AKIRIN1 RBM17 RPA2 PPIG PPP1R8 FAM20B TIMELESS GTSE1 CHCHD5 LRR1 PNMA1 POLR2B PRPF40A YTHDC1 KIAA1191 ELOVL5 POLDIP2 RTF1 FAM133B SRSF5 RNPEP SLTM DHX40 PFDN4 LOC441806 OAZ2 EXOC4 ARFGAP3 KIAA0368 NAPA RBM39 TERF1
Bicluster 3	TMEM18 SLC30A5 LOC152217 ERCC3 C12orf44 ATP6V1A EMG1 MRPL30 DPP8 FBXO22 ATPAF1 MBD4 FAM32A NGDN SPIN1 FAM96AP2 SPTBN1 ETF1 RBM10 NAA35 MRPL32 KIAA1984 C1D COQ5 ISCA1P1 VBP1 FASTKD2 TOPORS DNAJA1 MRPS36 ISCA1 SERP1 ALG3 NSA2 XPO1 CEBPZ PRPF4 CYTH2 SCAND1 CRIPT FUBP1 C19orf42 CTDSPL2 CNBP MAPKAP1 SPCS2 IP6K2 CNIH IKBKAP KIAA2013 NDUFS4 TMEM179B MRPL17 UBE3C HAT1 ATRAID FAM103A1 EIF4E2 GTF3C5 CISD1 UBE2L3 RNF7 MTX1 NRD1 C11orf73 RARS LYSMD3 GFM1 GOLGA7 CDK9 CCDC43 ACAD9 G3BP2 SNRNP70 IGBP1 PTDSS1 DDX54 MRPL18
Bicluster 4	RPF1 ANP32C OSBPL9 MAP2K1 RPL7L1 TOM1L1 UBLCP1 TMEM69 RRN3 UBA2 TIMM22 LAMTOR2 ACTR2 RAN C16orf13 CPSF3 EIF4E ARL5A ORMDL1 SRSF10 COPS8 ZNF593 RPS15P4 ACBD3 METTL2A ERBB2IP RAB22A XPOT UTP11L IWS1 U2SURP ALAS1 SARS ANAPC5 AGTRAP OXLD1 PPP1R7 TAX1BP1 PPHLN1 CTCF MLH1 ARMC10 PARN RSRC1 MBNL1 LRRC47 RPN1 DDX56 MLX CCDC12 RCC2 CWC27 CAMLG COA5 TUG1 SERBP1 DDX24 MRS2 RBMX2 METTL23 ECHS1 EXOSC9 CYB5A SDHB TMEM168 ZBTB8OS MCAT DIMT1 ZC3H14 MELK
Bicluster 5	CDCP2 CLEC17A SLC9C2 FAM161B KBTBD12 SSX7 TAC1 CACNA1S CCDC164 FAM19A4 GDF9 LIMD2 CC2D2B SEZ6L NAV2 AKR1D1 PVT1 ITGAD OR3A2 AFM MYBPC2 CPXM1 OR2T8 SDR16C6P LOC387770 HAVCR1 LRIT2 CYP2D6 ZNF461 FAM57B KLK3 LOC285696 FGB IL9R CLEC12A PTGER1
Bicluster 6	TFAM DDX47 MPZL1 THOC1 BIRC6 KIF23 UBE2Q1 ASAH1 ARMC10 POGK NR1H2 AIDA UBXN4 EIF2S2 LMNB1 NUP133 PPP2R1A FBXO7 UBAP2L RNF111 ATG13 CNP RELA UBE2G2 ALDH9A1 SRPK2 CASC3 DHX29 USP8 MAPK1IP1L PIP5K1A TDP2 SON TCERG1 DHX15 SEC23IP OSBPL8 PEPD AURKA BRCC3 PITRM1 PPP4R2
Bicluster 7	CHMP2A LYRM4 POLR2K CCT4 NOP56 RNF149 ZCCHC17 COPS4 MRPL53 ANP32D POP7 C22orf28 PSMD14 MRPL40 LOC388955 ESD TIMM50 UFD1L CCT2 DTYMK ACOT9 BUB3 METTL5 ECHS1 HSD17B4 TOMM40 SEPT7 ICMT SH3GLB1 CHCHD3 BAD CNOT2 RUVBL1 SEC61A1 NSL1 PSMD3 COPS6 PPP6R3 MNF1 ACTL6A
Bicluster 8	CD93 CCDC93 ZNF304 FAM40A KAT7 PGS1 KIN TRMT44 LRRIQ4 CELA3B GSC2 SAPCD2 GLMN RPTOR GUCY1B2 SNX13 PLAC9 TBC1D22A CWF19L1 PDE10A FABP6 CRYGN OR2M3 WWC2 TPTEP1 VGLL4 SLC1A7 INSL5 SPRY1
Bicluster 9	VAMP4 MED23 TMEM115 ZNF57 UBP1 SLC35G3 GBA2 ZNF274 SH3D19 SHPRH IFT80 SP8 LYRM7 ZNF266 SUV39H2 RDH14 POLR3KP2 KLF11 AAAS MLST8
Bicluster 10	NSMCE4A CENPC1 ALG6 ARL8A FAM54A JMJD1C ZMYM5 ZNF669 NAA15 MTIF3 SREK1 FLJ23152 MANEA SGK494 PHKA1 STXBP3 TSSC4 ANKHD1 FLJ30901
Bicluster 11	TTTY9A CDRT4 CD1D ZNF691 IGSF1 STARD9 CLEC4A TRIM43 TG EME1 HOXC4 SRP14 VSX1 CXorf56
Bicluster 12	UBE2W CLP1 MFSD11 ZFYVE16 LDLRAP1 ENDOV P2RX6 ATG7 FNIP1 ASB3 YIPF1 RASL10A CRTC1 TBRG1
Bicluster 13	PCDHB16 C6orf162 GZMA FOXJ1 PCNXL2 TRBV5-4 SNN PDZD9 C3orf22 NRP2 STXBP4
Bicluster 14	SEC24B HEATR6 RNFT1 ZBTB2 CENPJ COG7 FAM149B1 MXRA8 ZNRF3 C19orf12 ZRANB1 VPRBP
Bicluster 15	NRF1 CASP6 STARD3 SYT6 VPS52 MGC2752 ZNF187 ITSN2 SMYD5 ZFAT CPSF6
Bicluster 16	PIP4K2B MTERF CDK5PS PCGF6 IFT172 CTAGE1 LOC93622 MIIP TCEB3C
Bicluster 17	PHTF2 TP53I13 INPP5A XAB2 IRAK4 NFYB DDX52
Bicluster 18	C22orf39 ALG6 TFCP2 INTS7 C1orf109 IMMP1L PIGA PHKG2
Bicluster 19	THAP1 C2orf73 TAF11 RABGAP1 TRAPPC12 NDUFB3
Bicluster 20	ANKRD18DP C10orf118 SIDT1 TYW5 CCDC94

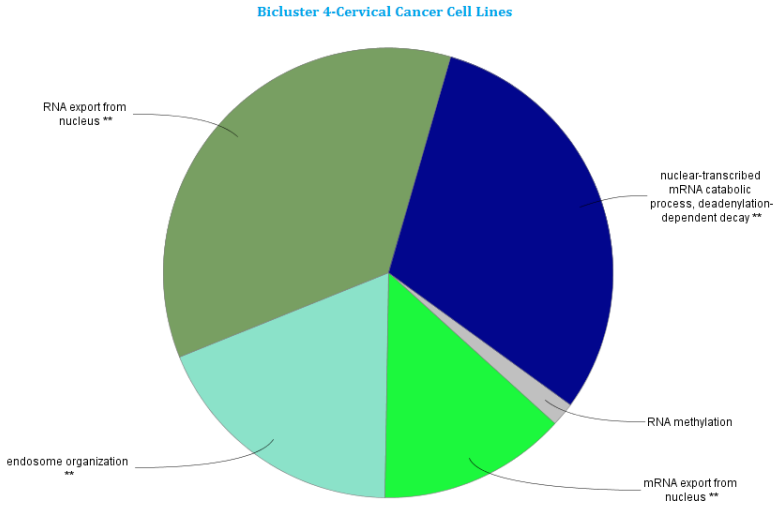
ΠΑΡΑΡΤΗΜΑ Β-2. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ

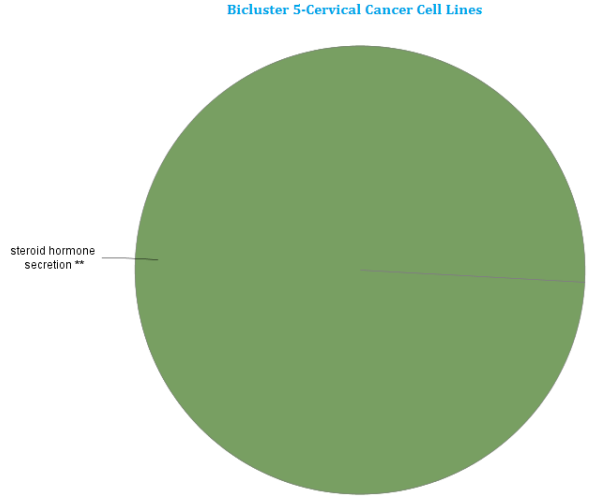


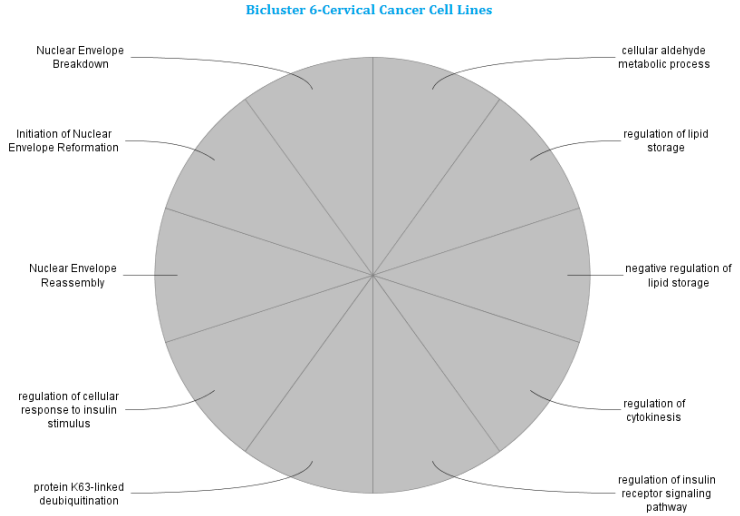


ΠΑΡΑΡΤΗΜΑ Β-2. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ (συνέχεια)

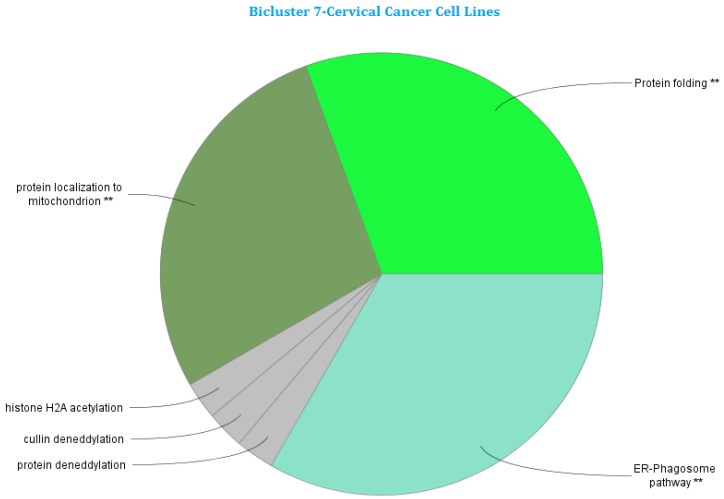


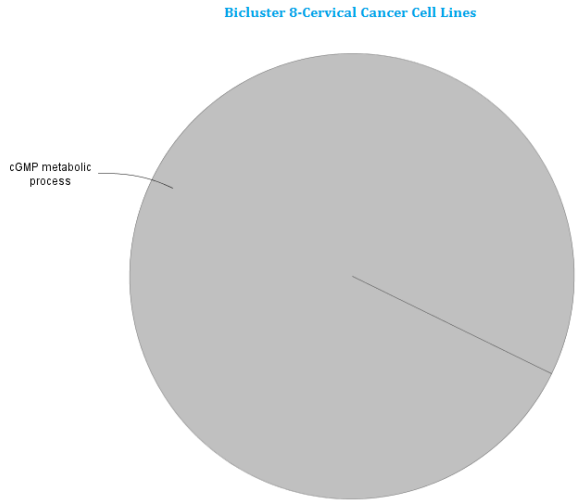






ΠΑΡΑΡΤΗΜΑ Β-2. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ (συνέχεια)





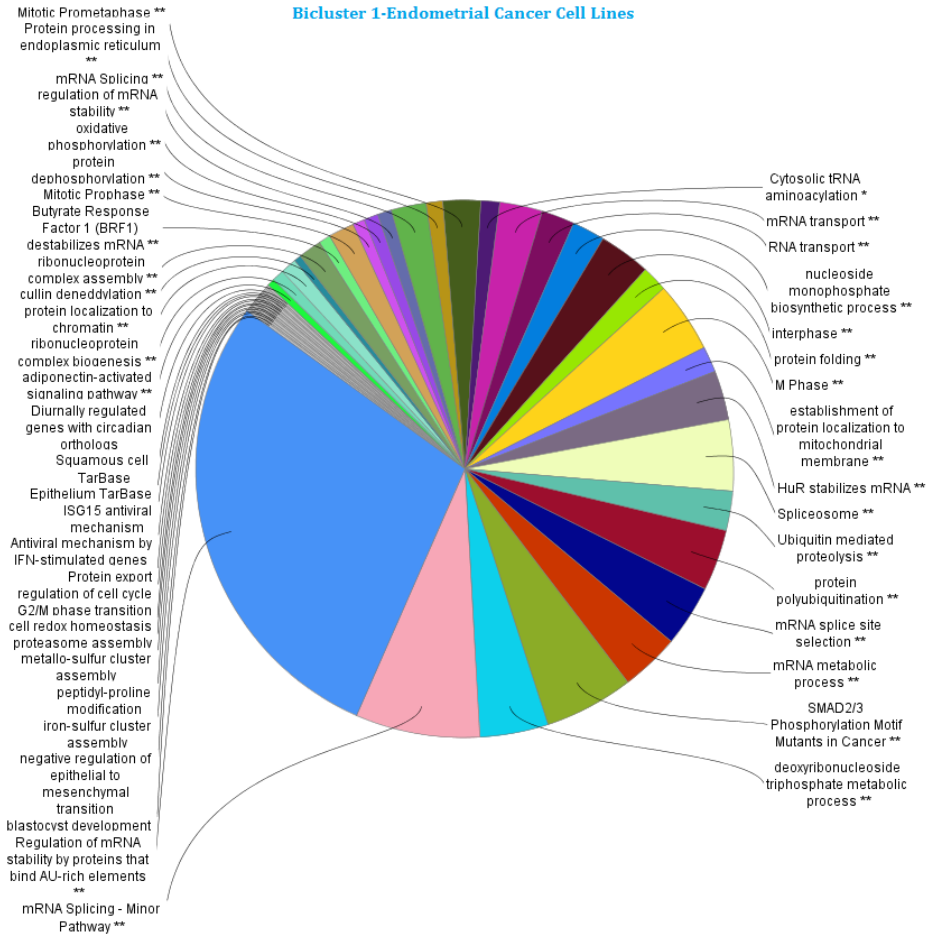
ΠΑΡΑΡΤΗΜΑ Β-2.3
ΕΝΤΟΠΙΣΜΟΣ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ

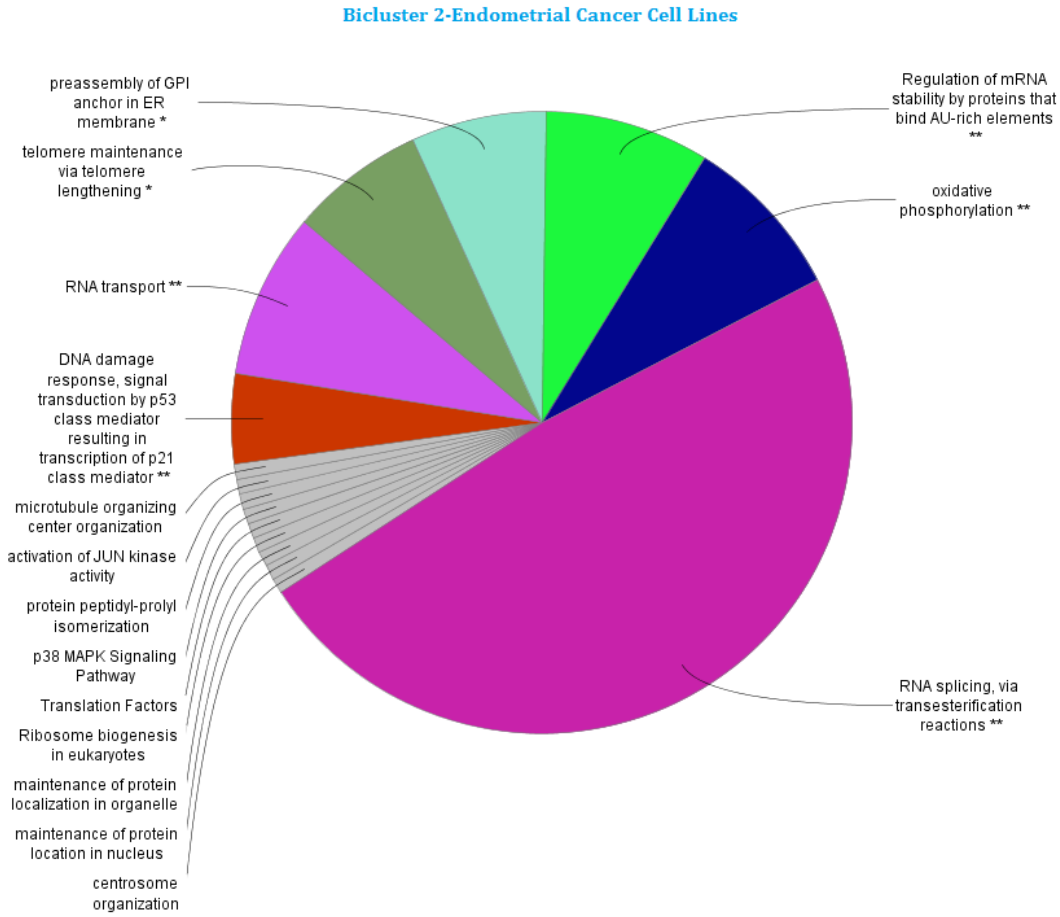
ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΡΑΧΗΛΟΥ ΤΗΣ ΜΗΤΡΑΣ		**ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (20)**																			
Αριθμός Κυτταρικών Σειρών	Καρκινικές Κυτταρικές Σειρές	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	SW756																				
2	GH354																				
3	C-4I																				
4	Hela																				
5	C-33A																				
6	CaSki																				
7	ME180																				
8	HT-3																				
9	SiHa																				

ΠΑΡΑΡΤΗΜΑ Β-3.1
ΟΜΑΔΕΣ ΓΟΝΙΔΙΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ

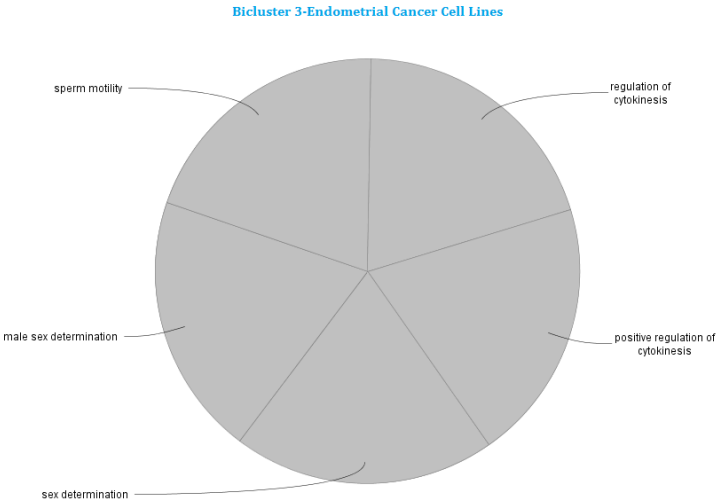
ΚΑΡΚΙΝΟΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ**
Bicluster 1	SNRNP40 DPH1 PRKCSH USP39 UBE2I NOL7 AARS C14orf119 PA2G4 PSMD1 SNRPD1 PSMA7 TIMM22 TRA2B PAK1IP1 NUP153 CCT7 SMARCE1 DLD PCBP2 DTYMK GLRX5 ELAC2 ICMT ISCA1 SON STOML2 SSU72 MRPL27 MAP4 HDAC1 DDX42 DRG1 CCZ1 PPP1R7 DHX9 CCDC47 SEC23B FKBP1A GDI2 DAD1 COPS3 MGEA5 ANP32B C19orf10 UBE2R2 NTMT1 GID8 EIF4E2 RALY PSMC3 SF3B14 BCLAF1 GORASP2 MATR3 ATP5C1 RBMX ERGIC3 LSM3 TMX2 UMPS C16orf80 ARHGAP11A RALBP1 SF3B1 ST13P5 CSNK2B RPS19BP1 UBL5 NOP56 CCNA2 TRAPPC1 DUT PSMD14 SEPHS1P4 LSM10 SFPQ MEGF6 CCT2 TAF10 EID1 PTBP1 HNRNPC DNAJA1 EDF1 GTF2A2 AMZ2 PPP2CA XPO1 UBE2N MPZL1 SF3A3 SDHAF2 RRP36 SUMO1 DNAJC8 ARF5 PAICS BTF3L4 HNRNPM PSMB6 PRPF40A UNC93B3 ENOPH1 HNRNPKP1 UBE2K NCAPG2 SRSF9 NOP10 LINC00493 CCNI ZNF12 ST13 PNRC2 ALKBH5 PPIL1 VTI1B SNRPD3 BRIX1 MAGOHB RAB8A SRSF2 PHF5A C11orf10 ANAPC11 NFU1 HNRNPU MTPN SF3A1 RBX1 INTS3 PSMG1 MRPS18A MRPS36 RHEB YTHDF2 PPM1G ADIPOR1 COPS6 NDUFA8 PSMC1 ZNHIT3 UBE2D3 CMC2 MCM6 EXOSC3 DPF2 CENPF CTDNEP1 MEA1 SRSF7 TOMM22 SRP54 UBE3C CSNK2A1P MRPS33 MRPL52 TRMT112 CBX1 ST13P2 CDCA5 SNRNP70 SLC25A5P2 KIAA0368 DYNLRB1 ANP32C CHMP5 PFDN6 TIMM9 MRPL50 NGDN IARS PPP6C TARDBP RPN2 GANAB TMEM14C ZFPL1 AUP1 CYB5R3 NARS SPCS1 MANBAL GLO1 C14orf166 RPS9 RAB10 GLOD4 ADRM1 FAM127B TIAL1 HNRNPR PSMD3 PSMD6 IK TM2D2 PDAP1 SRRM1 DCTN2 GNL2 PPP1CC EMC4 CKS1B H2AFY NDUFS8 PSMD11 UBE2G2 TOP1 NASP UBE2L3 SNRPA TARS MFN2 DEXI TRIM33 KPNA4 BIRC5 SET EIF2S1 SRSF3 LOC344382 RBMXL1 YWHAB
Bicluster 2	MAX NOP2 ARL2BP CSNK2A1 UBE2Q1 UBE2Z KAT5 TRAPPC8 EIF1B ECI1 RBM4 OTUB1 MAEA SIVA1 POP7 EIF5 EIF4E NDUFS2 SRSF10 MRPL37 LARS MRPL12 C12orf65 CLNS1A SH3GL1 PRIM1 UBTF BOLA3 SF3B4 BRD2 U2SURP NDUFS3 PRPF4 MNF1 PSMB5 SLC35A4 ADIPOR2 DPM2 DHX37 PRICKLE4 RANBP1 TIMELESS SELK MAML1 SLC25A17 LUC7L2 PREB VEGFB EXOSC2 HAUS6 POLDIP2 PSMA1 ARIH2 ODF2 MICU1 HNRPDL PHPT1 SLC39A6 MLF2 TCF12 LOC152217 UBP1 TRAF7 FOXM1 PIGH ISCA2 NDUFB11 MEAF6 FAM32A ATP6V0B ZNF24 ALKBH2 CS MRPS27 ETF1 MINPP1 PRPF19 KDM3B EBPL PPIE FASTKD2 LEPROTL1 DKC1 MRPS14 SRSF1 ATP5D DAXX SAFB PTPLAD1 SLC25A11 PRPF4B TSPAN3 PLEKHJ1 SRSF6 DCAF7 ARL2 PI4KA CKAP5 GTF2F1 CWC27 SS18L2 ANP32A CCNY METTL5 FBXO9 UQCRC1 COMMD4 NOP58 NEU1 MRPL36 ELP6 NCBP1 GOLGA7 TTC19 COQ9 RPP21 NDFIP1 DDB1
Bicluster 3	LANCL3 SLC28A3 TSPY26P TAS1R2 KLRB1 LRRC31 ANKRD26 FLJ32955 NR0B1 KCTD11 CACNG1 OR13C8 LOXHD1 ZBTB49 GRAMD2 MSLNL ADRA1A PTGFR DOHH IQUB CES4A PRDM10 PHOX2B ASB16 GPR3 SPATA9 C9orf57 TMEM150B CD69 INSRR CIB4 OR7G1 CGB5 LINC00608 CD93 LMTK3 SNTN ROBO2 POU6F2 PDZRN4 OR1A2 MST1 GAPDHS FLJ12334 RTP2 LOC283585 FBF1 LOC283854 TBX21 ZDHHC1 DCHS2 COLQ TEX33 GUCA1C SLC35G3 MLLT1 RPGRIP1 ZC3H12D APOC2 HABP2 WDR64 KRTAP9-9 C10orf112 IFNA17 KCNV2 CNGA2 SYT9 ALPK3 RBM26 HGFAC MDS2 LOC554174
Bicluster 4	VMA21 PSMA3 RWDD1 MRPS15 GPBP1L1 CHMP5 RALBP1 VTI1B MFF TSEN34 CCT4 DENR UBE2NL GOLGA5 DCAF12 PSMC2 MRPL34 LSM1 PRRC2C MND1 FAM127A SERF1A TSN RAB5A CDC23 ICMT GAR1 PNKD SNRNP27 NSA2 POLR2C PPP2CA HPRT1 NUDT21 PSMC1P4 GNG10 MAD2L1 EIF2B1 VPS4A MIS18A MED28 SNF8 TJP1 RAB6A CNIH SEPT2 HOMER1 PTPN1 ZCCHC17 KIAA0664 HIGD2B PTPN11 CKS1B GEMIN6 PCNP MRPS18C CLDND1 FMR1 SERBP1 FAM98A COPS2 BUB3 IFNGR2 RDBP DYNLT1 HSD17B4 FXR1 TARS ATP6V1D ZBTB8OS SLU7 COX7A2L FAM83D PTRH2 ARCN1 MRPL18
Bicluster 5	ZNF207 STIP1 CDV3 CNOT1 CHCHD8 NAT10 ACTR1A ZC3H15 NOL11 MRPL49 DNAJB6 C8orf59 NAE1 EMC7 ATP6V1G1 SEC61B EIF3J R3HDM1 GRWD1 KIAA0020 IWS1 LOC442060 LRRC42 C21orf59 COPS5 LUC7L3 CEBPZ MKRN1 PPA2 TSR3 PSMC5 NOLC1 TROVE2 AP2S1 PSMC1 MKI67IP WDR75 HJURP PRPF38A RRP12 DHX9 NAMPT TXNDC17 GTF2F2 MRPL17 PSMA5 UBE2L3 PPP4C DIABLO DKFZP586I1420 FRG1 ILK CDC26 CCT8 FADD PRDX4 CNOT2 SRP68 YWHAH BYSL RNF114
Bicluster 6	IMMT PTPRA FLYWCH2 TSFM ISG20L2 FBXO22 EIF4A3 UBAP2L C16orf13 PDCL3 TACC3 DCTN1 NHP2L1 EID1 CDC123 PIPSL MAGOH LSM7 SLBP KRTCAP2 TYK2 UBR4 DSN1 FASTKD5 PABPN1 MRPL9 EWSR1 ANKRD17 TM9SF4 PSMD7 CDCA8 NEMF TBL2 IKBKAP SPAG5 GTF2F1 NOL9 RNF167 AAMP BUD31 PYCR2 PEX14 HAT1 AKR7A2 CBWD3 STAU1 DDX24 HDAC3 ILF3 MYBL2 API5 C19orf70 EXOSC10 SF3B3 PWP1 GNA12 GNPAT YY1 NUBP2
Bicluster 7	TFB2M ST13 C10orf2 UBXN1 RMI2 SF3A3 RSL1D1 DDT LOC391322 SLC5A6 MIS12 SAP130 RSPRY1 DNAJB6 EFTUD2 TRIAP1 PRKRIR LOC440292 MRPL40 POLR3C PRPF6 TUG1 E2F6 IER3IP1 TRIM8 VWA5B2 NOL6 USP10 C9orf89 CNPPD1 ZC3H7A WDR12 EXO1 STUB1 TSR1 DAP3 PGP CCDC86 NSL1 ELAVL1 DNAJA3 HTATSF1 MPV17 TMX2 STK11 PPP6R3 UQCC SLC25A38 MRTO4 ALS2
Bicluster 8	DCTN3 GLT8D1 ZNHIT1 TMED4 MED31 DLGAP4 PMS2 PHF12 INO80E RPE FOXJ3 LGMN CWF19L2 SPTAN1 CASC3 VPS39 SKIV2L2 NUSAP1 PSMD4 CCNF C11orf73 WBSCR16 MRPL4 MGAT2 HAUS1 HSPB11 C1orf131 WDR46 ADNP ZRANB2 CCDC90A B3GNT1 PSME3 MAD2L1BP MBTPS1 SMAD4
Bicluster 9	RPF1 PDCD6IP DDX47 THOC1 RAE1 CDC37P1 TEX261 TMA16 ACAA1 TIMM10 NR1H2 VPS26A RBM17 SLC25A10 MAP7D3 AIMP2 EIF4E ASB6 RALY ISCA1P1 ANAPC13 FAM49B ELP3 CHCHD4 CCDC135 SHISA5 EIF2A PRKRA TRIM37 SLC50A1 RAD51 ZFAND6 TMEM248 UBE2B CSTF2 ZC3H14
Bicluster 10	ACN9 DROSHA ADO SLC35B1 ZWILCH VPS25 SEC22C SLC31A1 AP3S2 CHST12 FBXO7 EXOSC1 AP2B1 SEPHS1 ISCA1P1 POLR2G CDC123 ACBD3 ABCF2 METTL2A CDC6 ITGB1BP1 UBE2J2 MCFD2 ISCA1 ZDHHC3 SUPV3L1 GPN2 TRAPPC3 ITGB3BP FN3KRP NOLC1
Bicluster 11	PTPN18 MAP2K7 TDRD6 OR7A17 AFP HERC2P8 SPPL2B OR4N4 KRT86 PMS2P12 SIGLEC7 EPYC INTS4L1 TMEM169 GRIN2C KANSL1 TBC1D24 PARP11 B4GALNT2 CHRNA3 ASGR1 LOC440786
Bicluster 12	SEC13 KIAA1984 COX15 TOPBP1 TUFM LOC441455 DLST CRLS1 TOX4 CARHSP1 C21orf59 RBM25 PGAM4 AURKA CDC42BPB PARL TAF7 PUS1 SMPD4 CCNB2 RBM39 PNPT1 SNX1 UBE2G2 NUDT3 TMEM141
Bicluster 13	C19orf42 HIAT1 RAB14 GPKOW TBK1 GGNBP2 CKAP2 CKAP2L GLT25D1 TCERG1 ASF1B HAUS8 LOC390638 PAIP2 ADSS DPM2 PTPMT1 VPS28 PEX5 STX10 PPME1 ILVBL SBNO1 HEATR1 SERINC3 CDK2 TRIP4
Bicluster 14	NOL12 MBLAC1 ZFP64 WDR92 KCNF1 SMPD1 ZFYVE19 LINC00619 TOLLIP OSBPL11 FAM160A2 IQCE TAF4 VPREB1 BOD1L1 C10orf57 SNX14
Bicluster 15	NIF3L1 PNP SUCLG1 PAN3 CHAF1B MRPS34 HDGFRP2 MED4 SRRT RING1 CHRAC1 MPDU1 TMEM70 UBAP2 LMNB2 RBM14 AGPAT5 TCEA1 TOP1P1 C21orf33 METAP2
Bicluster 16	RTTN POLA1 TYW1 OR9G9 DZIP1L LOC148696 TET3 C20orf132 ADIG ASCC3 SNX18 PPP3CC CEBPE POLQ
Bicluster 17	CLPX B3GNT2 CDK8 THAP7 DHRS4-AS1 ERCC6L2 LOC90784 NFX1 DPP9 MAPKBP1 BRD1 CCNF AMBRA1 NSF HIRIP3 LSM14B SGSM3
Bicluster 18	KAT8 SIN3B RNMT ZMYND11 SEPT8 PRPF18 MED13 FANCL GNA12 TMEM127 TMEM38B PROSER1 GALNS SOCS6 CEP192 RBM15 C11orf57
Bicluster 19	PTPRVP SLC22A16 C2orf42 LOC100499484-C9ORF174 SPRR2C SLC52A1 NLRP9 LOC157562 GPR55 FAM75D1
Bicluster 20	STRADA PTGIR ENTPD5 PML ZXDB DDX31 NPY6R MAST2 ZNF2 NR2C1 TADA2A POC5

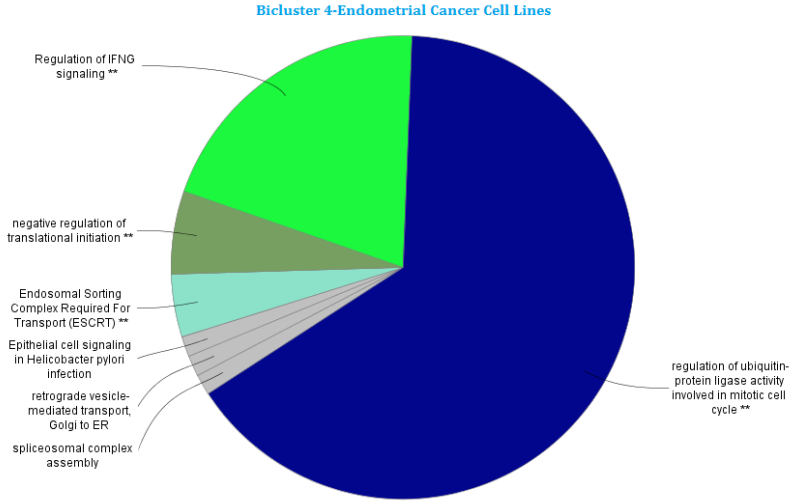
ΠΑΡΑΡΤΗΜΑ Β-3.2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ

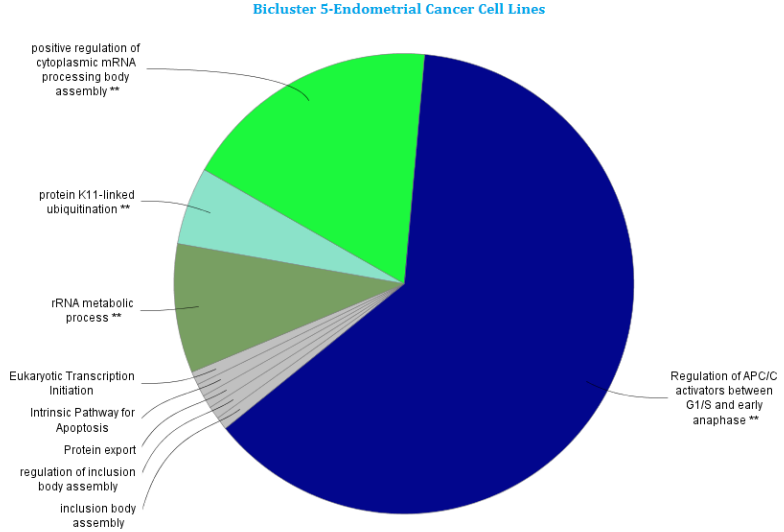


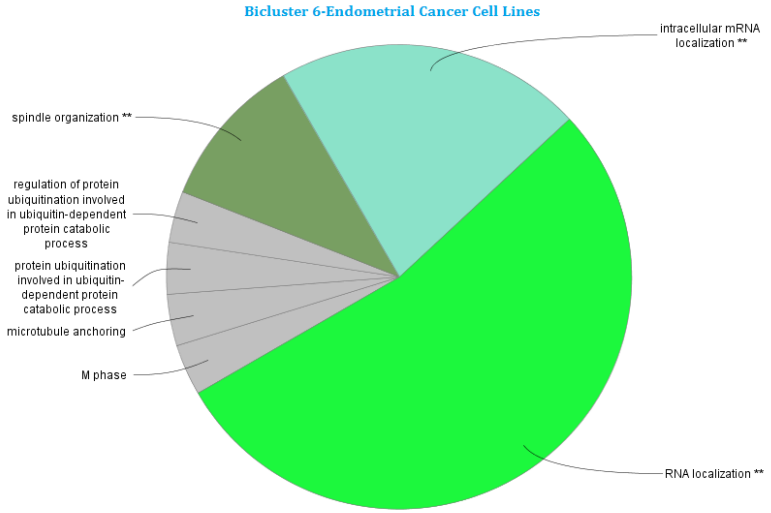


ΠΑΡΑΡΤΗΜΑ Β-3. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ (συνέχεια)

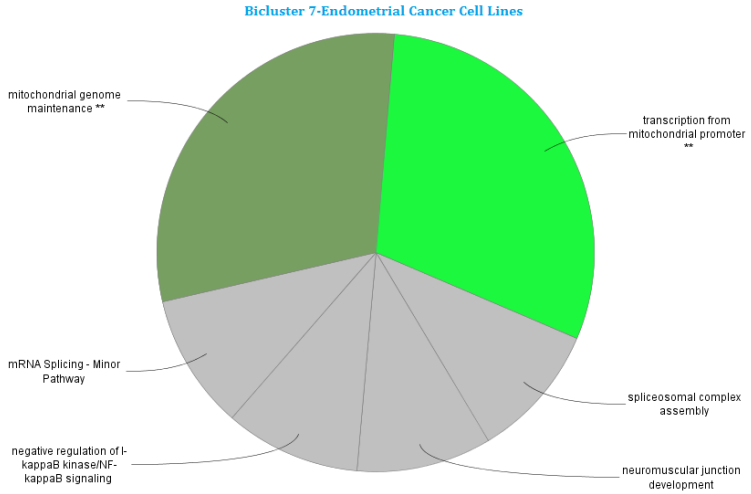


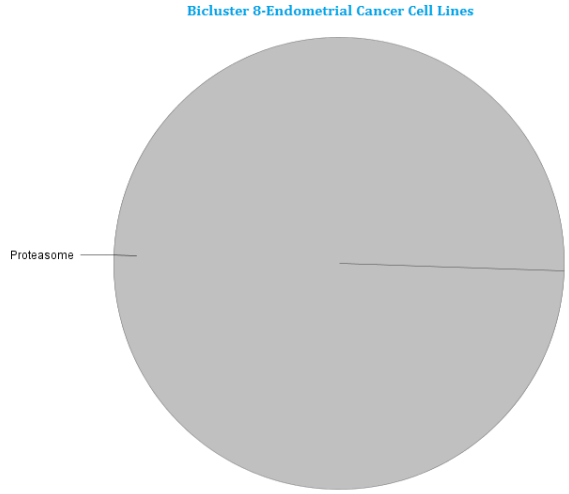


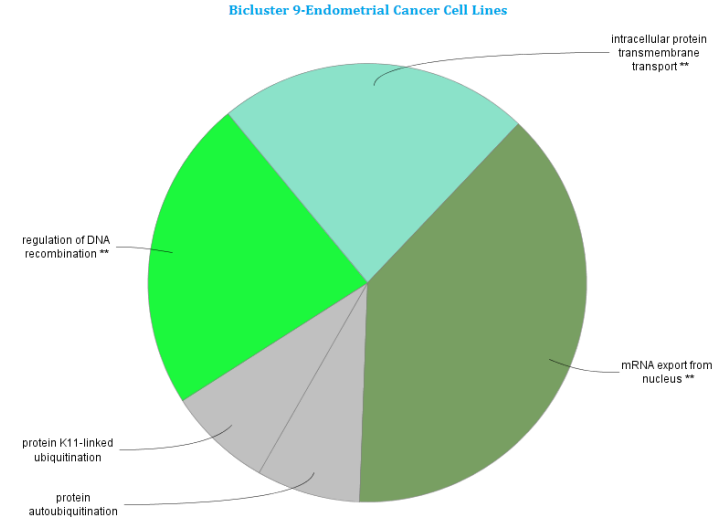


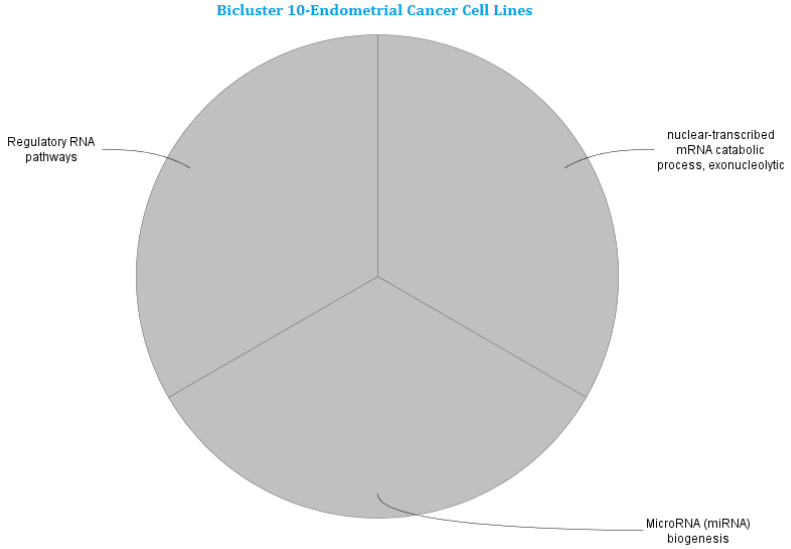


ΠΑΡΑΡΤΗΜΑ Β-3. 2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ (συνέχεια)

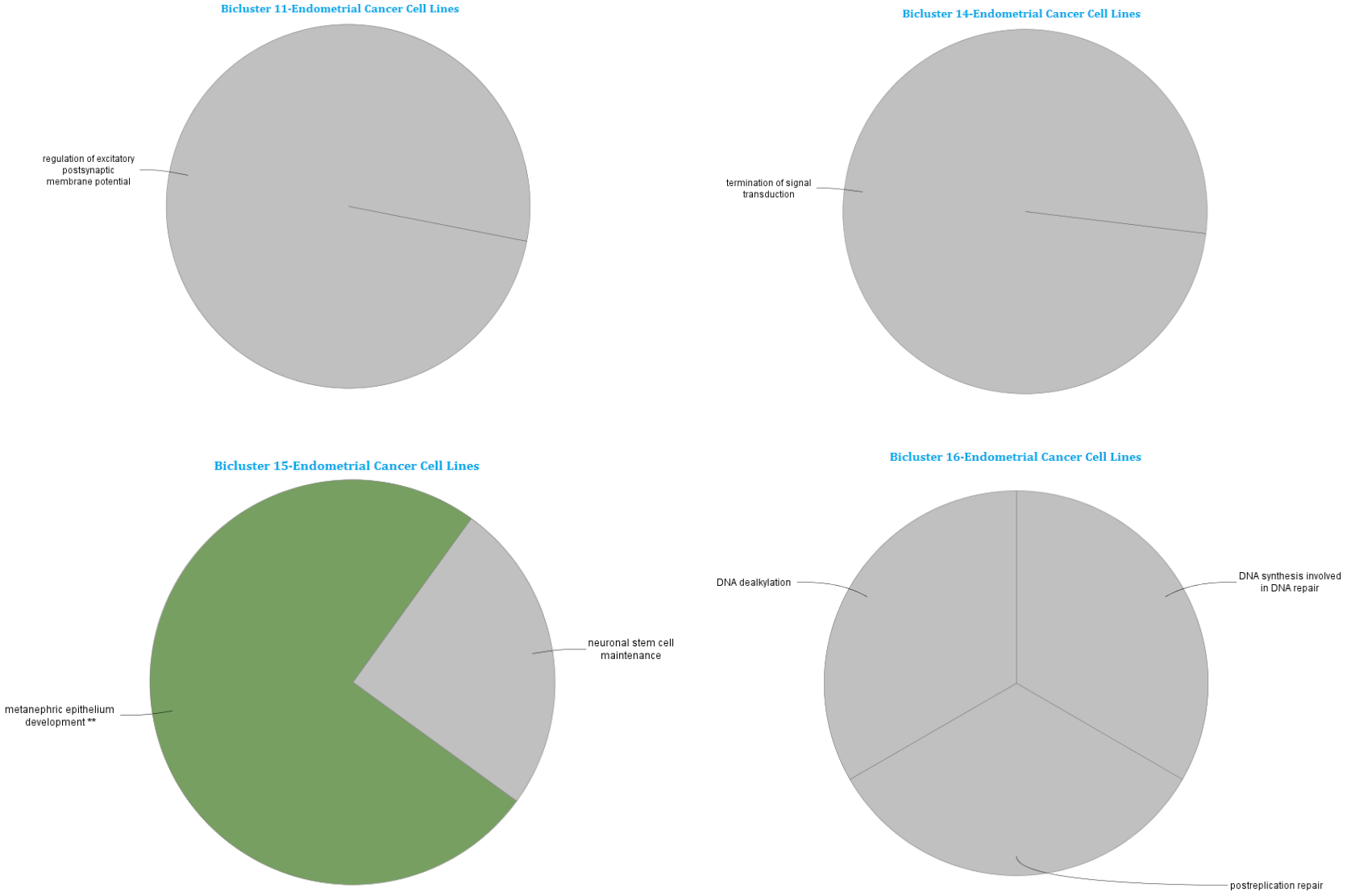




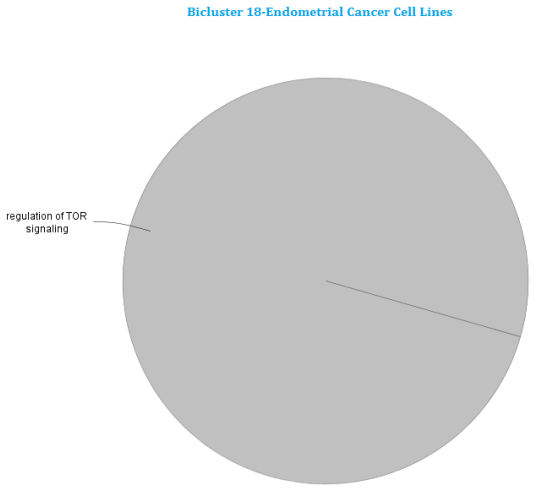


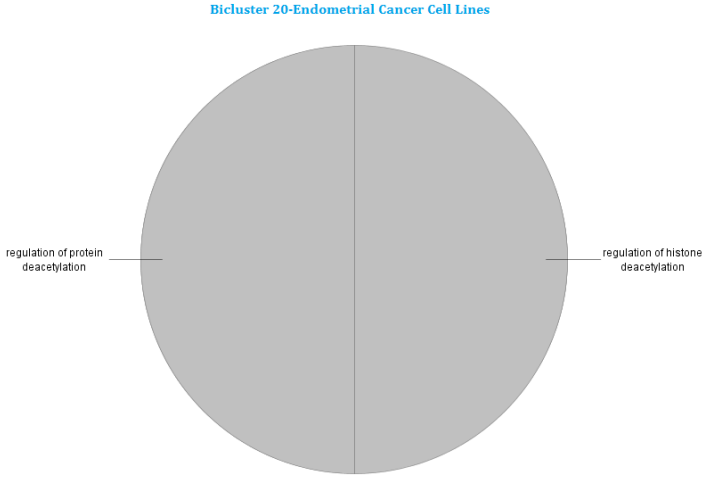


ΠΑΡΑΡΤΗΜΑ Β-3.2.
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ (συνέχεια)



ΠΑΡΑΡΤΗΜΑ Β-3.2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ (συνέχεια)





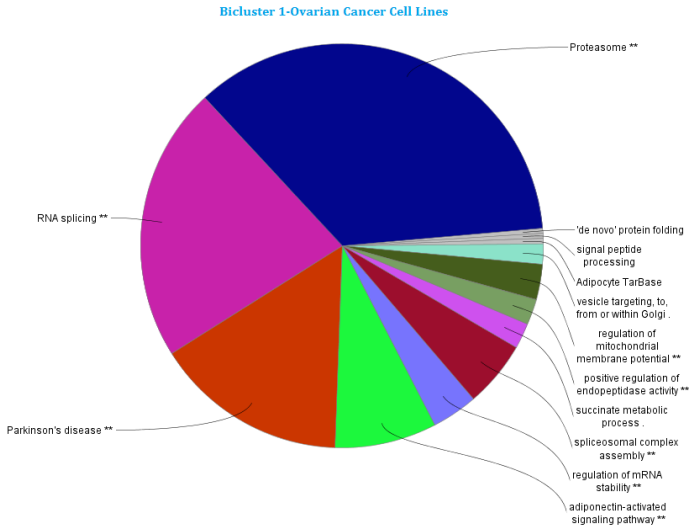
ΠΑΡΑΡΤΗΜΑ Β-3.3
ΕΝΤΟΠΙΣΜΟΣ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ

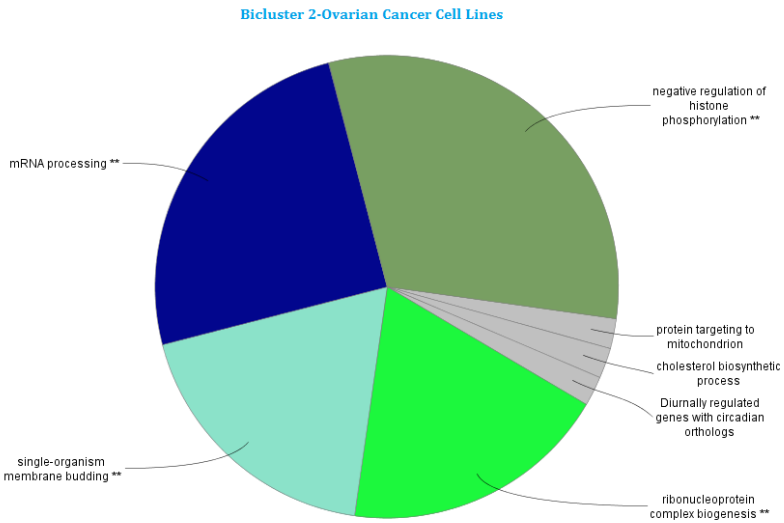
ΚΑΡΚΙΝΟΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ		ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (20)																			
Αριθμός Κυτταρικών Σειρών	Καρκινικές Κυτταρικές Σειρές	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Colo684																				
2	EJ																				
3	Ishikawa																				
4	RL95-2																				
5	HEC50B																				
6	HEC1B																				
7	EN																				
8	KLE																				
9	AN3CA																				

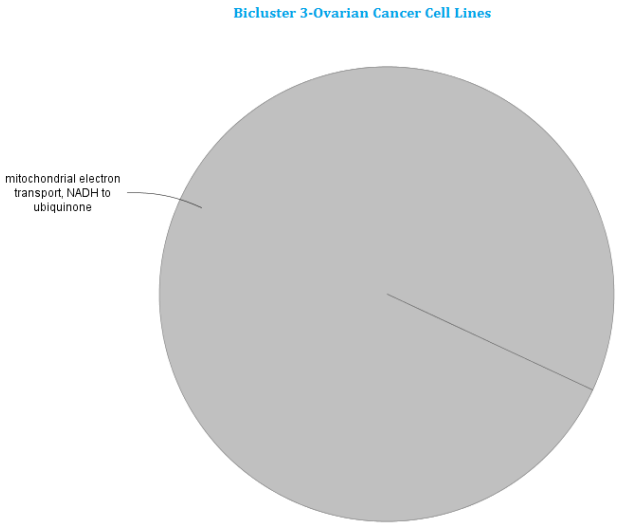
ΠΑΡΑΡΤΗΜΑ Β-4.1
ΟΜΑΔΕΣ ΓΟΝΙΔΙΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΩΝ ΩΟΘΗΚΩΝ

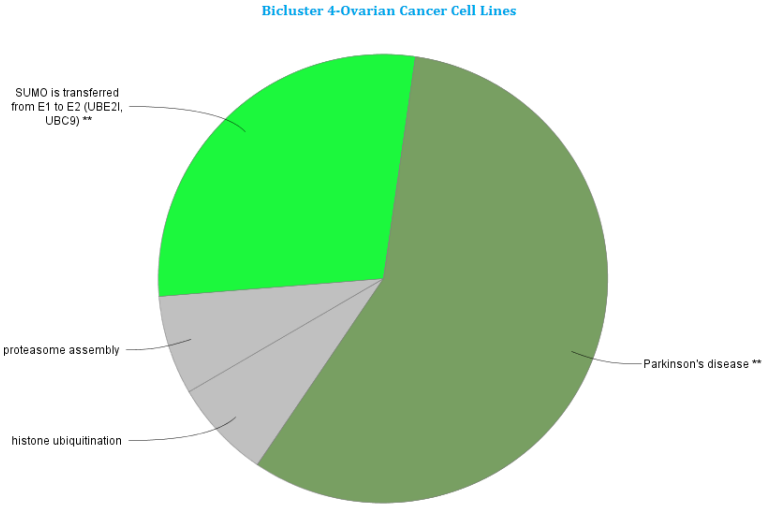
ΚΑΡΚΙΝΟΣ ΤΩΝ ΩΟΘΗΚΩΝ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΩΝ ΩΟΘΗΚΩΝ**
Bicluster 1	CCNI COPA MYEOV2 UQCR10 GNB1 PPIL1 PARK7 C14orf119 MAGOHB SNRPD1 RAD23B TRA2B CCT7 MANF MINOS1 PCBP2 SNRNP200 PIPSL UBE2V1 DHX15 YTHDF2 ATP5F1 LOC390294 DARS PPM1G PSMB2 OCIAD1 ADIPOR1 NACA2 MTCH1 SSRP1 FUBP1 COX5B NACAP1 PSMB5 AK2 PHB DHX9 MRPL22 SNF8 COX6C DNPEP EFTUD2 C22orf28 CAPZA1 SRSF7 PRELID1 TBCA STT3B SERBP1 ATP5G1 HIST1H3F BCLAF1 GORASP2 SRSF5 HDLBP SDHB SEC31A COX7A2L MZT2B BCCIP ARHGAP11A ATP5G1P5 SUCLG1 CSNK2B HNRNPA3 PAIP1 DENR CSDE1 SEC11A SNRPC ATP2A2 CAP1 MRPL49 HMGB1P1 PSMD14 VDAC3 TARDBP MRPL41 BAG6 ZFPL1 SEC11B SNRPB2 PUM1 ENY2 NMRAL1 HNRNPC RNPS1 NDUFA7 RAB1A PPP2CA NSA2 HNRNPR NAA10 ACP1 IK SRSF6 PHBP3 NDUFV1 PAICS PTPN11 PCNP CCNY HSPD1 CBWD3 PRDX3 ATP5G3 SRSF3 PSME3 RBM39 RBMXL1 ACOT7 SRSF9 NOP10
Bicluster 2	VMA21 IMMT HNRNPAB TMEM203 CHCHD8 NHP2 GGPS1 NOP56 DENR RAB8A FOXJ3 MPHOSPH10 MRPL53 PPP6C UQCRC2 RPN2 AQR TECR SFPQ MRPL37 SKIV2L2 FH VDAC2 RNPS1 METTL2A GOLPH3 ERAL1 SRSF1 NSA2 BOLA3 XPO1 DDX46 SSU72 UBE2B TOR1B C1orf43 NDUFA8 MRTO4 PACSIN2 CIAO1 KIAA0100 TM9SF4 CCZ1 UBAC2 DHX9 CTCF ARGLU1 MRPS34 ARF1 RAB6A CNIH PDCD7 LTV1 FAM192A ATIC RBM34 FBXO7 H2AFY FUBP3 GTPBP8 AZIN1 LRR1 AIMP2 MTX2 HNRNPM PREB ADAM17 RDBP PSPC1 RNF7 TUFM FBXO9 UBAC1 PUM2 MTX1 ABCF1 ARIH2 ZFR EXOSC10 PUF60 MRPL36 LCMT1 TPD52L2 GNPAT TMEM248 CSTF2 SMPD4 POLR1C PTDSS1 LINC00493 KAT6A
Bicluster 3	SEC13 ZDHHC5 SNX17 TXNDC9 NCKAP1 YIPF6 EEF1B2 AIDA ZNF358 BCAS2 KDM2A GRSF1 MRPL34 BTF3L4 CSNK1A1L NDUFS8 HARS2 ERI3 EMC3 TRIP4 PDCL3 KIAA1191 RDBP MRS2 MIER1 COQ2 UBE2K UTP11L SNRNP27 ARF4 VPS28 C21orf59 DR1 GOLGA8A TFDP1 ATP1B3 NDUFB7 DDB1 MRPL47
Bicluster 4	SDHC NDUFS6 BOLA2 SAE1 POMP UBE2I NDUFS4 RNF181 UQCRFS1 UBE2NL ARPC5 WAC LYPLA1 TFG MRP63 DNAJC7 TCEB2 RBM42 WDR1 METTL5 PSMG1 VDAC1 FRG1 CCT8 CAND1 LOC442060 GNG10 ATP5A1 CAPZB ARPP19 UBE2N
Bicluster 5	ANAPC5 PPP1R2 C12orf44 BZW1 ACTR3 CUTA NANS DUSP14 NGRN COMMD3 PA2G4 ZNF259 UTP18 PIN1 BANF1 UBL7 CNIH4 UBE2D1 SPTSSA DDX1 PDLIM2 DCTN1 SRPR IDH3G GADD4 5GIP1 NRD1 IMP3 TRAPPC3 NUP50 KIAA0368 ARCN1 CWC15 FUBP1
Bicluster 6	NIF3L1 VBP1 CHAF1B RARS2 UBXN1 PPHLN1 TRMT61B CPSF2 PAIP2 YEATS4 COPZ1 SUB1 COX14 FAM168A EMC8 ACTR2 RBBP4 LMAN1 LSM1 SERBP1 TARDBPP2KLHDC2
Bicluster 7	GPRC5D AAK1 BRAP PPEF2 KRTAP10-12 LYSMD4 TET3 RPL7A HNF1A AFF4 ADGB KNDC1 EFCAB3 CLASP1 FBXW8
Bicluster 8	MARCH9 BRAF DUSP26 SLC5A5 CCDC22 HDGFL1 VARS TIPRL CENPL CDK5RAP3 KRIT1 AGK LMBR1L OR2M3 C19orf24 TMEM209 FAM86C1 PARS2 DRG2
Bicluster 9	CYFIP1 RHOA KHDRBS1 PTBP1 GPBP1 EIF3A EMC4 HUWE1 DUT PPP2R4 LYPLA1 SSB FIP1L1 UBE2G2 SNAPIN
Bicluster 10	GLIS3 L3MBTL2 SPAG8 IGLC1 HSH2D KLK11 ATP10B KLHL32 SARM1 XKR8
Bicluster 11	EIF2C1 RSBN1 MTO1 SUV39H2 PEX16 ZNF322P1 OGT ANKRD13C MGC2752 WDR76 CDAN1 KCTD6 NFYB
Bicluster 12	NME6 PTER WBP2P1 WDR37 CTTNBP2NL FANCM SP2 NPHP3 GFRA1 NRXN2 KRT16 USP6
Bicluster 13	RFC3 LEO1 HERC1 RBAK C1orf27 SETD2 DUS2L DCP1A HEATR5A ZNF434 TMEM186
Bicluster 14	NAT14 MFSD5 NCK1 SETD3 SRP68 FAM104A PEPD MRPL12 TMED10P1 EIF2S1 EIF2B2 YY1
Bicluster 15	AMOTL2 ENTPD2 WDR92 IQCE CEP120 TUT1 PPP1R12B PCF11 SGOL1 CRLF3
Bicluster 16	MSANTD4 EML3 SPNS1 NFYA TMEM105 HENMT1 TBC1D22A TBC1D25 AGXT2L2 CTAGE1 CLN3 GABPB1 CWC25
Bicluster 17	ACSS2 UNC119B NUP50 DEM1 WDR13 EPC2 EPN1 MAP1S CLK4 SMCR7L
Bicluster 18	GRIA1 WNK4 ESCO1 LOC441167 ADA ZBTB24
Bicluster 19	CCL21 SGSM3
Bicluster 20	DNAJC14 SLAIN2 XRN1 GCFC1 SAP30L

ΠΑΡΑΡΤΗΜΑ Β-4.2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΩΝ ΩΟΘΗΚΩΝ

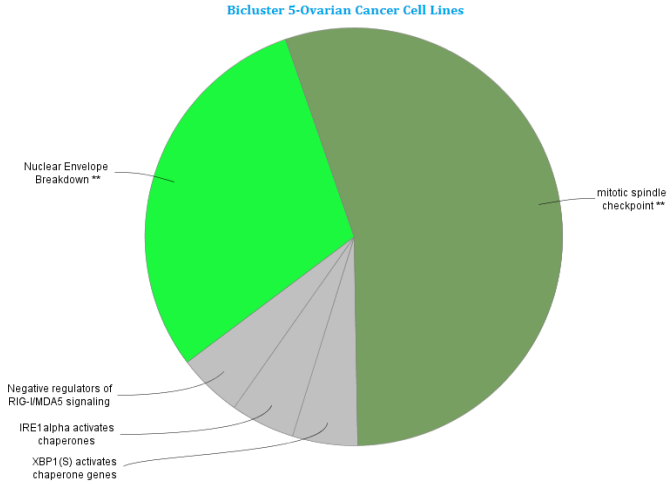


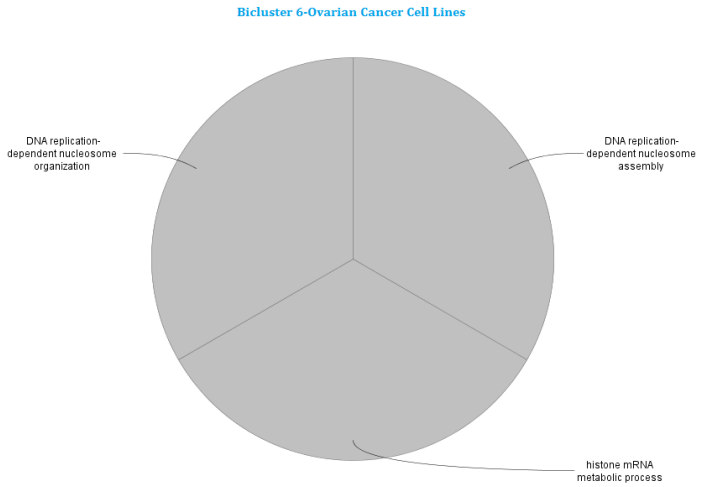


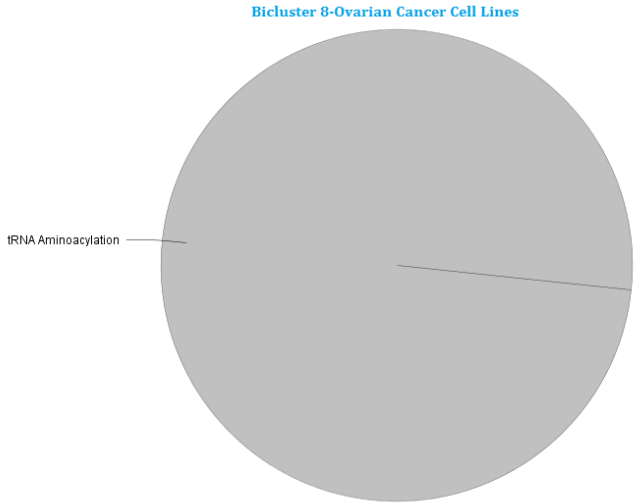


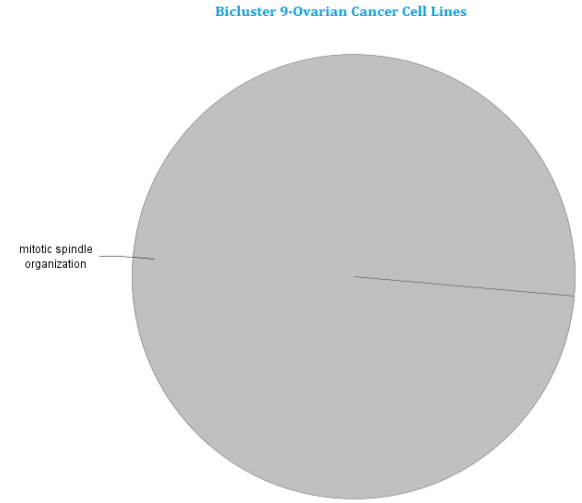


ΠΑΡΑΡΤΗΜΑ Β-4.2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΩΝ ΩΟΘΗΚΩΝ (συνέχεια)

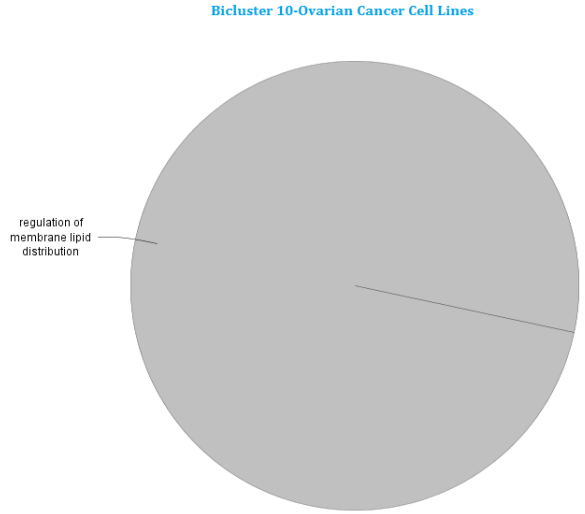


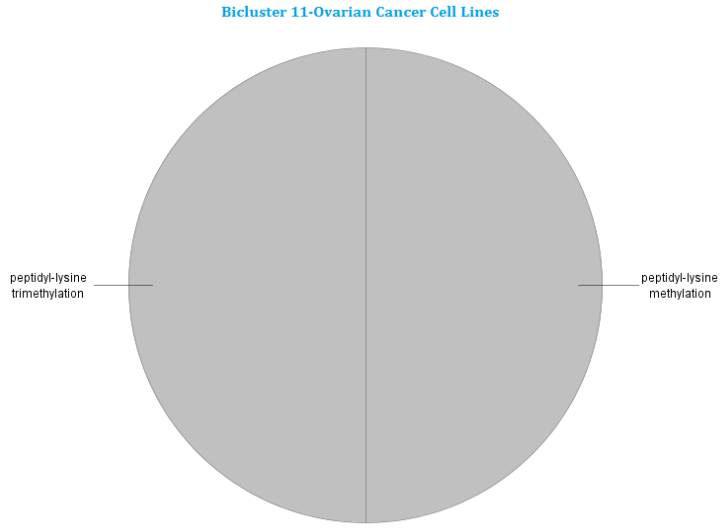


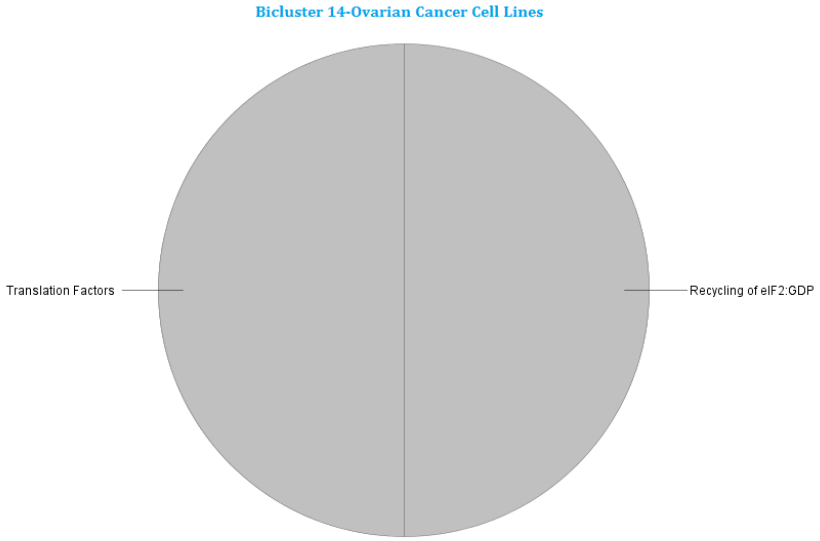


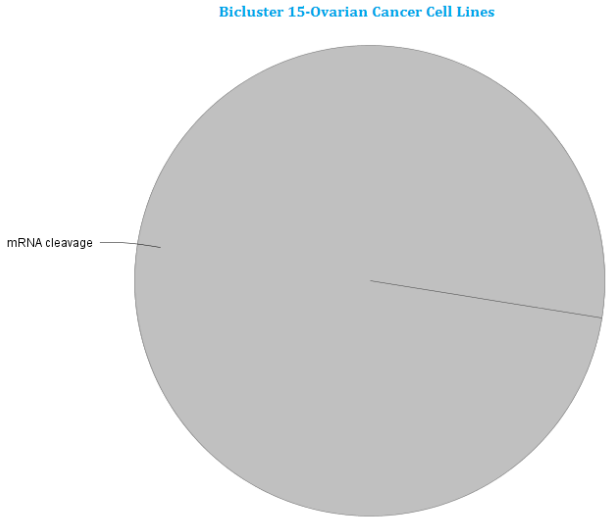


ΠΑΡΑΡΤΗΜΑ Β-4.2
ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ ΚΑΙ ΜΟΝΟΠΑΤΙΑ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΝ ΚΑΡΚΙΝΟ ΤΩΝ ΩΟΘΗΚΩΝ (συνέχεια)









ΠΑΡΑΡΤΗΜΑ Β-4.3
ΕΝΤΟΠΙΣΜΟΣ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΙΣ ΚΑΡΚΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΤΩΝ ΩΟΘΗΚΩΝ

ΚΑΡΚΙΝΟΣ ΤΩΝ ΩΟΘΗΚΩΝ		**ΟΜΑΔΕΣ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ (20)**																			
Αριθμός Κυτταρικών Σειρών	Καρκινικές Κυτταρικές Σειρές	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	TOV-1112D																				
2	TOV-21G																				
3	A2780																				
4	OVMZ-1a																				
5	OVMZ-6																				
6	ES2																				
7	CaOV3																				
8	OV-90																				
9	NIHOVCAR3																				
10	SKOV3																				

ΠΑΡΑΡΤΗΜΑ Β-5.1
ΟΜΑΔΕΣ ΓΟΝΙΔΙΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΚΑΤΑ ΤΗΝ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ

ΟΜΑΔΟΠΟΙΗΣΗ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΑΠΟ ΤΗΝ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ**
Bicluster 1	RPL21P128 EEF1G EEF1B2 RPL11 RPS10P3 RPL21P20 RPL21P44 CLIC1P1 RPL13AP23 RPL4P5 RPL37A RPL29 RPL9 RPS25 RPL31 RPL22P11 RPS9 RPL29P12 RPS27A RPL21P39 RPL21P131 RPL36P14 RPS8 RPL15 EEF1A1 RPL21P19 RPL13A RPLP1P7 RPL13AP17 RPL21 RPL6 RPS3AP5 GNB2L1 RPLP0 RPS3A RPSAP49 RPL7A RPL32 NACA3P RPL24 RPL4 SNRPD2 RPL36AP40
Bicluster 2	RPL3 RPS8 RPL23AP82 EEF1G EEF1A1 RPL5 RPL13A LOC392181 RPL23A RPS10P22 RPS18 RPS23 RPL23AP44 RPL31 RPL6 RPS10 RPL23AP65 RPL10A RPL23AP37 RPS10P29 RPS7 RPS27A RPL31P28 RPL23AP14
Bicluster 3	MTCH2 RPS5 RPL15 SNRPE RPL26P19 RPL29P33 RPSAP55 TMA7 ATP5G1 RPL21 VDAC1 RPL27A RPL29P13 RPSA ATP5F1 NDUFS5 ATP5O AURKAIP1
Bicluster 4	HNRNPA1 HNRNPC HNRNPA1P4 LOC402112 LOC120364 HNRNPA1P30 HNRNPK
Bicluster 5	RPS29 RPL38 ATP5B ERH FAU RPS26
Bicluster 6	PSMB4 SSR2 RPSA EEF1B2 RPSAP31 RPL12 RPL36AP49 PGAM1
Bicluster 7	NACAP1 UBE2L3 SLC25A3 SNRPF RPS15P9 CCT4 MAGOHB COX6A1 DDX5 TBCA BTF3 ANP32B SNRPG CCT7
Bicluster 8	COPS2 DCTN2 METTL5 VDAC2 GNB1 PPP1CC PDHB LOC442060 SUMO1 PPP2CA ZC3H15 SSU72 CHCHD1 ATP5C1 DARS CCNY ARCN1 SERBP1
Bicluster 9	C15orf23 MRPL9 KHDRBS1 CHCHD2 CCT4 TIMM9 CBX3 GDI2 COX8A H2AFY CCT6A KPNB1 LOC441241 CPSF7 SSRP1
Bicluster 10	HNRNPA1 UBE2D3 KHDRBS1 DDX39B EDF1 H2AFV EIF2S2 LOC341333 COX6A1 SET ANP32B NDUFA8
Bicluster 11	CCZ1 UQCR10 FH SNRPD2P1 PSMG1 ARF1 RARS RAD23B POP7 PPP2R4 MRPL34 PSMB2 NDUFS8 AUP1
Bicluster 12	PSMD1 RPL7 RPL23AP2 EDF1 RPL26L1 RPLP0P2 RPL7P37 HNRNPA3
Bicluster 13	ST13 EIF4B ST13P5 HNRNPD RPL7P33 NDUFA13 MZT2B SF3B2
Bicluster 14	SPCS2P TRMT112P6 C19orf43 LSM4 SPCS2 MRPL11 LOC388955 PSMC5
Bicluster 15	LARS RBX1 SRSF6 ERGIC3 UFC1 RPN2 ZNF45 EIF4E LOC375295 NOP10 YTHDF2
Bicluster 16	RNF181 SYNCRIP SCAMP3 BANF1 GMFB ACIN1 LRR1 CMC1 SLURP1 SNRNP70 JAGN1 DDX42 USP16 PREB
Bicluster 17	DRG1 MLX DDX24 RDBP CINP MRPL49 ZC3H14
Bicluster 18	RPS15AP9 LOC440577 CCT8
Bicluster 19	SDHAF2 STIP1 EIF2S2 QARS KPNB1
Bicluster 20	CAPZB SMN2 PGAM1 MDH2 COX7A2 LAMTOR5

1. Οι εικόνες με τις βιολογικές διεργασίες και τα μονοπάτια των ομάδων διπλής κατηγοριοποίησης από την ομαδοποίηση των τεσσάρων καρκινικών τύπων παρατίθενται στο κείμενο της παρούσας διπλωματικής εργασίας.

2. Ο Πίνακας που αποτυπώνει τον εντοπισμό των ομάδων διπλής κατηγοριοποίησης κατά την ομαδοποίηση των τεσσάρων καρκινικών τύπων παρατίθεται στο κείμενο της παρούσας διπλωματικής εργασίας.

ΠΑΡΑΡΤΗΜΑ Β-6.1
ΕΙΚΟΣΙ ΕΠΤΑ ΚΟΙΝΑ ΓΟΝΙΔΙΑ ΣΤΟΥΣ ΤΕΣΣΕΡΙΣ ΚΑΡΚΙΝΙΚΟΥΣ ΤΥΠΟΥΣ

ΤΕΣΣΕΡΙΣ ΚΑΡΚΙΝΙΚΟΙ ΤΥΠΟΙ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΥΣ ΤΕΣΣΕΡΙΣ ΚΑΡΚΙΝΙΚΟΥΣ ΤΥΠΟΥΣ**
One Group	IK PTBP1 METTL5 HNRNPC DHX9 MRPS34 UBE2I UBE2K SRSF1 NSA2 EMC4 RAB8A BOLA3 SNRPD1 XPO1 UBE2NL PSMD14 PPP6C BTF3L4 TRA2B SRSF3 KIAA0368 RBMXL1

1. Η εικόνα με τις βιολογικές διεργασίες και τα μονοπάτια της ομάδας διπλής κατηγοριοποίησης των τεσσάρων καρκινικών τύπων παρατίθεται στο κείμενο της παρούσας διπλωματικής εργασίας.

ΠΑΡΑΡΤΗΜΑ Β-7.1
ΕΝΑ ΚΟΙΝΟ ΓΟΝΙΔΙΟ ΣΤΟΥΣ 4 ΚΑΡΚΙΝΙΚΟΥΣ ΤΥΠΟΥΣ ΚΑΙ ΣΤΗΝ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ

ΤΕΣΣΕΡΙΣ ΚΑΡΚΙΝΙΚΟΙ ΤΥΠΟΙ ΚΑΙ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΤΕΣΣΑΡΩΝ ΚΑΡΚΙΝΙΚΩΝ ΤΥΠΩΝ	
Ομάδα Διπλής Κατηγοριοποίησης	**ΣΥΜΒΟΛΑ ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΜΑΔΩΝ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΟΥΣ ΤΕΣΣΕΡΙΣ ΚΑΡΚΙΝΙΚΟΥΣ ΤΥΠΟΥΣ**
One Gene	METTL5