**ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ**

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Διπλωματική Εργασία

# ΜΕΘΟΔΟΙ ΔΙΠΛΗΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΓΟΝΙΔΙΩΝ

## "Biclustering Methods in Gene Analysis"

ΑΝΔΡΙΤΣΟΥ ΔΑΦΝΗ

**Επιβλέπων Καθηγητής:** Καθηγητής Ζερβάκης Μιχάλης

**Εξεταστική Επιτροπή:** Καθηγητής Ζερβάκης Μιχάλης

Καθηγητής Μπάλας Κωνσταντίνος

Αναπληρωτής Καθηγητής Λαγουδάκης Μιχαήλ

Χανιά, Ιούνιος 2017

# Ευχαριστίες

*Θα ήθελα να ευχαριστήσω*

*Πρωτίστως, τον καθηγητή μου κύριο Ζερβάκη Μιχάλη, για την πολύτιμη βοήθεια και καθοδήγησή του καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Για την εμπιστοσύνη που μου έδειξε δίνοντας μου τη δυνατότητα να εκπονήσω την πτυχιακή μου εργασία στον επιστημονικό τομέα που επιθυμούσα. Επίσης, θα ήθελα να τον ευχαριστήσω για τη διάθεση του να με βοηθήσει και να μου λύσει οποιαδήποτε απορία οποιαδήποτε στιγμή το χρειαζόμουν.*

*Τους καθηγητές κύριο Μπάλα Κωνσταντίνο και κύριο Λαγουδάκη Μιχαήλ για τη συνεισφορά τους ως μέλη της εξεταστικής επιτροπής.*

*Την Δρ. Μπέη Αικατερίνη για τον πολύτιμο χρόνο που αφιέρωσε από την αρχή μέχρι το τελικό στάδιο της διπλωματικής μου εργασίας καθώς και για την απλόχερη στήριξη και ενθάρρυνσή της προς το άτομό μου.*

*Την Δρ. Obermayr για την παροχή του dataset και τον διδάκτορα Σφακιανάκη Στέλιο για την βοήθειά του στην επεξεργασία του dataset.*

*Την Αλεβυζάκη Ανδρονίκη για την βοήθειά της στο αρχικό στάδιο της εργασίας μου.*

*Θα ήθελα να ευχαριστήσω θερμά όλους τους φίλους μου, κυρίως την Έλενα, την Άννα, τη Ρένια και τη Σέβη που πίστεψαν σε εμένα και με ενθάρρυναν σε κάθε στάδιο των σπουδών μου.*

*Ιδιαίτερες ευχαριστίες θέλω να εκφράσω προς τους γονείς μου, Ευαγγελία και Κωσταντίνο για την διαχρονική συμπαράστασή τους και την υλική και ηθική στήριξη των επιλογών μου.*

# Περίληψη

Οι μικροσυστοιχίες DNA αποτελούν μία από τις πιο διαδεδομένες πειραματικές μεθόδους ταυτόχρονης ανάλυσης του τρόπου έκφρασης χιλιάδων γονιδίων σε διαφορετικά δείγματα ή σε διαφορετικές συνθήκες. Το γεγονός αυτό τις καθιστά ιδανικό εργαλείο για την ανάλυση και μελέτη καρκινικών ιστών, με στόχο την εύρεση των μοριακών μηχανισμών που διέπουν την ογκογένεση σε διάφορους τύπους καρκίνου και την περαιτέρω κατανόηση της νόσου. Τα δεδομένα γονιδιακής έκφρασης αναπαρίστανται με τη μορφή πινάκων όπου οι γραμμές αντιστοιχούν σε γονίδια και οι στήλες σε διάφορες πειραματικές συνθήκες. Στόχος των διάφορων τεχνικών ομαδοποίησης αποτελεί η εξαγωγή σημαντικών βιολογικών πληροφοριών που αφορούν ομάδες γονίδιων κάτω από συγκεκριμένες συνθήκες. Στην παρούσα εργασία επιλέγεται η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης δεδομένων γονιδιακής έκφρασης που προέρχονται από δείγματα κυτταρικών σειρών τεσσάρων τύπων καρκίνου - του μαστού, των ωοθηκών, του ενδομητρίου και του τραχήλου της μήτρας - συγκεντρωμένα σ' έναν πίνακα δεδομένων. Η μεθοδολογία που προτάθηκε, περιλάμβανε την εφαρμογή του αλγορίθμου διπλής κατηγοριοποίησης των Cheng και Church, εφαρμόζοντας σε αυτόν μια σειρά βελτιώσεων ούτως ώστε να εξάγουμε ταυτόχρονα ομάδες γονιδίων με ομοιόμορφη συμπεριφορά αλλά και συγκεκριμένο εύρος τιμών. Ο αριθμός γονιδίων (ανιχνευτών) που περιείχε ο πίνακας, ανερχόταν σε 33096, γεγονός που έκανε την μελέτη και την ανάλυση της συμπεριφοράς των γονιδίων δύσκολη. Γι' αυτό το λόγο, μειώσαμε το πλήθος των γονιδίων προς ανάλυση στα 1000 γονίδια ανά καρκινικό τύπο, καθώς επίσης, εστιάσαμε στην εξέταση των τριών πρώτων ομάδων διπλής κατηγοριοποίησης (Biclusters) που προέκυπταν κάθε φορά. Με αυτή τη διπλή κατηγοριοποίηση **υποχώρου δεδομένων**, καταφέραμε να δημιουργήσουμε ομάδες γονιδίων οι οποίες 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, όπως η συμμετοχή των γονιδίων που τις απαρτίζουν σε συγκεκριμένες βιολογικές διεργασίες (π.χ. μεταβολισμός, βιοσύνθεση μακρομορίων, αναπαραγωγή) και εξειδικευμένα μοριακά μονοπάτια (π.χ. οξειδωτική φωσφορυλίωση, μεταφορά RNA, ενδοκύττωση).

# Abstract

DNA microarrays comprise one of the most widespread experimental methods. This fact makes them a perfect tool for analysis and evaluation of cancer tissues, with the aim to find those molecular mechanisms governing the formation of tumors in different types of cancer and a further understanding of this disease. Molecular expression data are depicted in tables where the lines represent the genes and the columns represent the different experimental conditions.

The aim of the different grouping techniques is to elicit crucial biological information regarding gene groups under specific circumstances. In this essay the method selected was the double categorization of gene expression data coming from samples of cell arrays of four different types of cancer (breast, ovarian, endometrial, cervical) gathered on a data table. The methodology proposed included the application of the Cheng & Church biclustering algorithm, by applying a series of revisions to it so as to elicit gene groups with similar patterns and a specific value range simultaneously.

The number of genes (probes) included in the table rose to 33096, a fact that posed a serious burden to the study and analysis of the genes. For that reason, we brought the number of genes analyzed down to 1000 per type of cancer as well as focusing on studying the first three Biclusters each time. With this double categorization of data subspace we managed to create gene groups which showed genes with similar patterns along the cell arrays and, at the same time, their value range was small, making these groups more coherent and offering us the opportunity to elicit crucial biological information, such as the involvement of their genes in concrete biological processes (e.g., metabolism, macromolecular biosynthesis, reproduction) and specific molecular pathways (e.g., oxidative phosphorylation, RNA transfer, endocytosis).

## ΠΕΡΙΕΧΟΜΕΝΑ

## ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΕΙΚΟΝΩΝ-ΔΙΑΓΡΑΜΜΑΤΩΝ

### Κεφάλαιο 1



### Κεφάλαιο 2



### Κεφάλαιο 3



### Κεφάλαιο 4

**Κεφάλαιο 5**

**Κεφάλαιο 6**

## ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΠΙΝΑΚΩΝ

### Κεφάλαιο 1



### Κεφάλαιο 3



### Κεφάλαιο 4



### Κεφάλαιο 5



### Κεφάλαιο 6



## ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΣΧΗΜΑΤΩΝ

# 1

# Εισαγωγή

## 1.1. Ανάλυση του Προβλήματος

Η ανάπτυξη της τεχνολογίας τα τελευταία χρόνια έχει οδηγήσει σε πολύ σημαντικές εξελίξεις, οι οποίες έχουν επηρεάσει σε σημαντικό βαθμό όλες τις πτυχές της ζωής μας. Μία από αυτές τις πτυχές, είναι και αυτή της βιοϊατρικής, δηλαδή του κλάδου που προσπαθεί να αναπτύξει της δυνατότητες της ιατρικής μέσω της εξέλιξης της τεχνολογίας.

Στα πλαίσια της ανάπτυξης νέων τεχνολογιών έχουν παραχθεί εργαλεία γονιδιωματικής απαραίτητα για την κατανόηση και τη διαλεύκανση των κυτταρικών λειτουργιών, των βιοχημικών οδών και των ρυθμιστικών μηχανισμών με βάση το πρότυπο της γονιδιακής έκφρασης (gene expression profile) ενός κυττάρου. Μία από αυτές τις τεχνολογίες είναι οι μικροσυστοιχίες DNA, οι οποίες αποτελούν ένα σημαντικό εργαλείο τα τελευταία χρόνια αλλάζοντας τον τρόπο με τον οποίο η βιοϊατρική και άλλες επιστήμες αντιμετωπίζουν διάφορες βιολογικές ερωτήσεις και δίνοντας ώθηση στη μετάφραση της γονιδιακής έρευνας στην κλινική. Από την άλλη μεριά, η τεχνολογία αυτή βασίζεται σε μεγάλο βαθμό σε συγκέντρωση μεγάλου όγκου δεδομένων, τα οποία απαιτούν επεξεργασία. Τις τελευταίες δεκαετίες παρουσιάζεται ραγδαία αύξηση του όγκου των βιολογικών δεδομένων προς επεξεργασία και η ανάγκη για τον μετασχηματισμό της αδόμητης αυτής πληροφορίας σε γνώση. Αυτός ήταν ο κύριος λόγος δημιουργίας του τομέα της βιοπληροφορικής (bioinformatics). Έτσι, καθώς αναπτύσσεται αυτός ο κλάδος, κρίνεται απαραίτητο να χρησιμοποιηθούν διάφοροι σχετικοί αλγόριθμοι, ώστε να είναι εφικτά τα καλύτερα αποτελέσματα.

Όπως αναφέρθηκε, η τεχνολογία των μικροσυστοιχιών DNA επιτρέπει να παρατηρήσουμε ταυτοχρόνως χιλιάδες γονίδια και να καθορίσουμε ποια από αυτά είναι διαφορετικά εκφρασμένα σε ένα συγκεκριμένο τύπο κυττάρου ή υπό συγκεκριμένες συνθήκες. Επιπλέον, τα προφίλ γονιδιακής έκφρασης ενός νοσούντος κυττάρου/ιστού, σε σύγκριση με τους φυσιολογικούς μάρτυρες, μπορεί να προάγουν την κατανόηση της παθολογίας της νόσου και τον εντοπισμό νέων θεραπευτικών σημείων παρέμβασης, τη βελτίωση της διάγνωσης και τη διασαφήνιση της πρόγνωσης. Στην παρούσα εργασία επιλέχθηκε να εφαρμοστεί η τεχνική διπλής κατηγοριοποίησης σε δεδομένα γονιδιακής έκφρασης που προέρχονται από δείγματα καρκινικών κυτταρικών σειρών -καρκίνου του μαστού, των ωοθηκών, του ενδομητρίου και του τραχήλου της μήτρας- με στόχο την βέλτιστη ανάλυση των γονιδίων που είναι συγκεντρωμένα σε ένα πίνακα δεδομένων. Πραγματώθηκε, λοιπόν, η μελέτη αλγορίθμων ευαίσθητων στην πληροφορία και όχι στο θόρυβο και με ιδιαίτερη λεπτομέρεια και ακρίβεια στην ομαδοποίηση.

## Κίνητρο και Στόχοι

Οι πίνακες γονιδιακής έκφρασης έχουν εκτενώς αναλυθεί χρησιμοποιώντας ομαδοποίηση (clustering) σε μία από τις δύο διαστάσεις:

- Τη γονιδιακή διάσταση
- Τη διάσταση της κατάστασης

Πιο αναλυτικά :

α) ανάλυση της έκφρασης προτύπων γονιδίων, συγκρίνοντας τις γραμμές του πίνακα και,
β) ανάλυση της έκφρασης προτύπων γονιδίων από δείγματα, συγκρίνοντας τις στήλες του πίνακα.

**Οι αλγόριθμοι ομαδοποίησης**

- Μπορούν να εφαρμόσουν ομαδοποίηση είτε σε γονίδια ή σε δείγματα, έτσι η ανάλυση κατευθύνεται σε μία συγκεκριμένη πτυχή του υπό μελέτη συστήματος.
- Συνήθως, επιδιώκουν μία ασύνδετη κάλυψη του συνόλου των στοιχείων, «απαιτώντας» ότι κανένα γονίδιο ή δείγμα δεν θα ανήκει σε περισσότερες από μια ομάδες (clusters).

Πολλά ενεργά πρότυπα είναι κοινά σε μία ομάδα γονιδίων μόνο κάτω από συγκεκριμένες πειραματικές καταστάσεις. Ανακαλύπτοντας τέτοια τοπική έκφραση προτύπων μπορεί να είναι το κλειδί για την αποκάλυψη πολλών γενετικών μονοπατιών που δεν είναι εμφανή με άλλο τρόπο. Είναι ως εκ τούτου επιθυμητό να προχωρήσουμε πέρα από το παράδειγμα ομαδοποίησης και να αναπτύξουμε προσεγγίσεις που θα είναι σε θέση να ανακαλύπτουν τοπικά πρότυπα σε δεδομένα μικροσυστοιχιών.

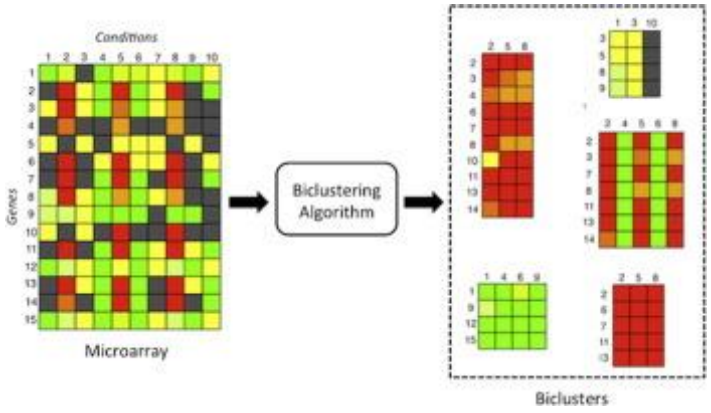
Η διπλή κατηγοριοποίηση αποτελεί μια πιο σύγχρονη μέθοδο ταυτόχρονης ομαδοποίησης γραμμών και στηλών, που αφορά στην ομαδοποίηση δεδομένων με βάση την ένταση της γονιδιακής έκφρασης (υψηλές ή χαμηλές αριθμητικές τιμές) των γονιδίων ή το εύρος της γονιδιακής έκφρασης (διακύμανση των αριθμητικών τιμών) των γονιδίων υπό συγκεκριμένες συνθήκες. Καινοτομία της συγκεκριμένης διπλωματικής αποτέλεσε η στόχευση και στις δύο αυτές παραμέτρους που ο συνδυασμός τους παρέχει αποτελέσματα με μεγάλη στατιστική σημαντικότητα.

Παρατηρούμε επίσης ότι ένας φυσικός τρόπος απεικόνισης μιας ομάδας Biclusters συνίσταται στην ανάθεση ενός διαφορετικού χρώματος για κάθε Bicluster και στην αναδιάταξη των σειρών και των στηλών του πίνακα δεδομένων, έτσι ώστε να αποκτήσουμε έναν πίνακα δεδομένων με έγχρωμα μπλοκς, όπου κάθε μπλοκ αντιπροσωπεύει ένα Bicluster (Εικόνα 1.1). Το Biclustering πρόβλημα μπορεί να διατυπωθεί ως εξής: Δοθέντος ενός πίνακα δεδομένων M, κατασκεύασε μια ομάδα Biclusters $B_{opt}$ που σχετίζονται με το M έτσι ώστε:

$$f(B_{opt}) = \max_{B \in BC(M)} f(B)$$

Όπου $f$ είναι μια αντικειμενική συνάρτηση που μετρά την ποιότητα, δηλ. το βαθμό συνοχής μιας ομάδας και τα BC (M) είναι το σύνολο όλων των δυνατών ομάδων Biclusters που σχετίζονται με το M. Αυτό το πρόβλημα είναι NP-hard.[1],[2]

Ειδικότερα, η διπλή κατηγοριοποίηση (biclustering) και οι παραλλαγές της ανήκουν στην κλάση μη-ντετερμινιστικής πολυωνυμικής (Non Deterministic Polynomial, NP) πολυπλοκότητας, δηλαδή είναι NP-πλήρης. Αυτό σημαίνει ότι τα προβλήματα που περιλαμβάνει μπορούν να επαληθευτούν εντός πολυωνυμικού χρόνου (Εικόνα 1.1). Η πολυπλοκότητα του προβλήματος διπλής κατηγοριοποίησης είναι άρρηκτα συνδεδεμένη με το εκάστοτε πρόβλημα, αλλά και την εκτιμήτρια συνάρτηση που χρησιμοποιείται για την αξιολόγηση μιας δοσμένης ομάδας διπλής κατηγοριοποίησης (Bicluster).[1],[2],[3]




**Εικόνα 1.1.** – Πρόβλημα NP-hard [3]

Στην παρούσα διπλωματική εργασία στόχος είναι η επεξεργασία δεδομένων με τον αλγόριθμο διπλής κατηγοριοποίησης ο οποίος παρουσιάστηκε το 2000 από τους **Cheng και Church.**[4]

Σκοπός είναι η μελέτη του αλγορίθμου στην αρχική του μορφή και η εκτέλεση πειραμάτων για την εύρεση Biclusters των οποίων τα γονίδια θα παρουσιάζουν λιγότερο θόρυβο (πιο μικρή απόσταση μεταξύ τους), καθώς και ομοιόμορφη δομή. Πιο συγκεκριμένα για τον υπολογισμό της διακύμανσης της ομάδας όλων των στοιχείων του Bicluster, δηλαδή της βαθμολογίας μέσου τετραγωνικού υπολείμματος (mean squared residue score) θα χρησιμοποιηθεί η απόλυτη τιμή και θα αλλάξει ο τύπος υπολογισμού ανά γονίδιο και ανά κατάσταση. Η πειραματική διαδικασία εφαρμόστηκε σε πίνακες γονιδίων-κυτταρικών σειρών προερχόμενα από μικροσυστοιχίες DNA από δείγματα καρκίνου του μαστού, του τραχήλου της μήτρας, των ωοθηκών και του ενδομητρίου.

## 1.2. Υφιστάμενη Γνώση

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. Κατά την ανάλυση τέτοιων πινάκων τα πιο σημαντικά σημεία μελέτης περιλαμβάνουν: [5]

- Ομαδοποίηση γονιδίων με βάση την έκφρασή τους κάτω από διαφορετικές καταστάσεις.
- Πρόβλεψη νέων γονιδίων, με βάση την έκφραση άλλων γονιδίων με ήδη γνωστή πρόβλεψη.

- Ομαδοποίηση καταστάσεων με βάση την έκφραση ενός αριθμού γονιδίων.
- 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 είτε 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.

- Η μέθοδος διπλής κατηγοριοποίησης παρουσιάστηκε αρχικά από τον J.A. Hartigan το 1972 [6] - αναδρομικός διαχωρισμός πίνακα μέσα σε blocks (Block clustering) - , ενώ ο όρος καθιερώθηκε το 1996 από τον Mirkin.[7]
- Όμως ο αλγόριθμος που περιγράφει τη γενική μέθοδο δόθηκε το 2000 από τους Y. Cheng και G. M. Church,[4] οι οποίοι και πρότειναν έναν αλγόριθμο διπλής κατηγοριοποίησης που βασίζεται στον υπολογισμό της διακύμανσης (variance) και τον εφάρμοσαν σε δεδομένα γονιδιακής έκφρασης και ακόμα αποτελεί τη βάση της διπλής κατηγοριοποίησης της γονιδιακής έκφρασης.

Από το 2000, ένας μεγάλος αριθμός από επιστημονικά άρθρα έχουν δημοσιευθεί σχετικά με το biclustering. Οι ενδιαφερόμενοι αναγνώστες μπορούν να ανατρέξουν στη βιβλιογραφία, σε άρθρα, όπως των [8],[9],[10],[11].

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

**Πίνακας 1.1.** – Εφαρμογές διπλής κατηγοριοποίησης τα τελευταία 14 χρόνια [12]

| Σύνολο δεδομένων | Αναφορές |
|---|---|
| Πολλαπλή σκλήρυνση | C. Tang et al 2001 [13] |
| Καρκίνος του μαστού | Ben-Dor et al 2002 [14] |
| Ζυμομύκητας-στρες | E. Segal et al 2003 [15] |
| Καρκίνος του παχέος εντέρου | T.M. Murali et al 2003 [16] |
| Λευχαιμία | Q. Sheng et al 2003 [17] |
| Ζυμομύκητας/Λέμφωμα | C. Cano et al 2007 [18] |
| Οξεία λεμφοβλαστική λευχαιμία/λέμφωμα/ ζυμομύκητας | U. Maulik & S. Bandyopadhyay 2009 [19] |
| 5 τύποι καρκίνου του πνεύμονα | C.-P. Chen et al 2014 [20] |

Η πλειοψηφία των βιολογικών εφαρμογών που εφαρμόζει την μέθοδο της διπλής κατηγοριοποίησης διεξάγεται με τη χρήση της τεχνολογίας των μικροσυστοιχιών η οποία επιτρέπει την μέτρηση της γονιδιακής έκφρασης από χιλιάδες γονίδια υπό σαφείς πειραματικές συνθήκες. Το μεγαλύτερο μέρος των μελετητών που έχουν εφαρμόσει μεθόδους διπλής κατηγοριοποίησης χρησιμοποιούν πίνακες οι οποίοι προκύπτουν από καρκινικά κύτταρα μέσω διαφόρων σταδίων της ασθενείας. Εκμεταλλεύονται επίσης δείγματα από διαφορετικούς ασθενείς και υγιή άτομα όπως και από πρότυπους οργανισμούς (π.χ. ζυμομύκητες) (Πίνακας 1.1.).

## 1.3. Δομή της Εργασίας

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

Το πρώτο κεφάλαιο περιλαμβάνει το κίνητρο και τον στόχο της παρούσας διπλωματικής εργασίας καθώς και το σύνολο του αλγοριθμικού υπόβαθρου που μελετήθηκε.

Στο δεύτερο κεφάλαιο αναλύονται βασικές έννοιες όπως το κύτταρο, το γονίδιο και η γονιδιακή έκφραση καθώς και τα διάφορα επίπεδά της, και δίνονται συνοπτικά οι έννοιες των μεταγραφικών προτύπων και των κυτταρικών σειρών. Επιπλέον, περιγράφεται το RNA, το DNA, το μοντέλο διπλής έλικας του DNA καθώς και η επιστήμη της βιοπληροφορικής.

Στο τρίτο κεφάλαιο, αρχικά διαχωρίζονται οι αλγόριθμοι κατηγοριοποίησης βάσει γονιδίων, δειγμάτων και υποδιαστημάτων. Στη συνέχεια περιγράφεται η τεχνική διπλής κατηγοριοποίησης, η δομή και τα είδη των αλγορίθμων της.

Στο τέταρτο κεφάλαιο περιγράφεται ο αλγόριθμος Cheng και Church που εφαρμόστηκε αναπτύσσοντας τα επιμέρους βήματά του και το θεωρητικό υπόβαθρο στο οποίο στηρίζεται. Περιλαμβάνεται ακόμα, η προτεινόμενη μεθοδολογία μας με την εφαρμογή του αλγορίθμου Cheng και Church στον οποίο έχουμε επέμβει προτείνοντας βελτιώσεις που στοχεύουν σε πιο λεπτομερή αποτελέσματα.

Τα αποτελέσματα παρατίθενται στο πέμπτο κεφάλαιο για τις τρεις πρώτες ομάδες γονιδίων διπλής κατηγοριοποίησης: α) για τα 1000 πρώτα από τα 33096 γονίδια στον καρκίνο του μαστού, β) για τα 1000 πρώτα από τα 33096 γονίδια στον καρκίνο του τραχήλου της μήτρας, γ) για τα 1000 πρώτα από τα 33096 γονίδια στον καρκίνο των ωοθηκών, δ) για τα 1000 πρώτα από τα 33096 γονίδια στον καρκίνο του ενδομητρίου, και ε) για το σύνολο των 33096 γονιδίων στον καρκίνο του μαστού και του τραχήλου της μήτρας.

Στο έκτο κεφάλαιο γίνεται η αξιολόγηση των αποτελεσμάτων και τέλος, στο έβδομο κεφάλαιο περιγράφονται κάποια τελικά συμπεράσματα και μελλοντικές επεκτάσεις της δουλειάς μας.

### Καινοτομία της Εργασίας

Στην παρούσα εργασία πραγματοποιείται η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης δεδομένων γονιδιακής έκφρασης που προέρχονται από δείγματα παθολογίας του καρκίνου του μαστού, των

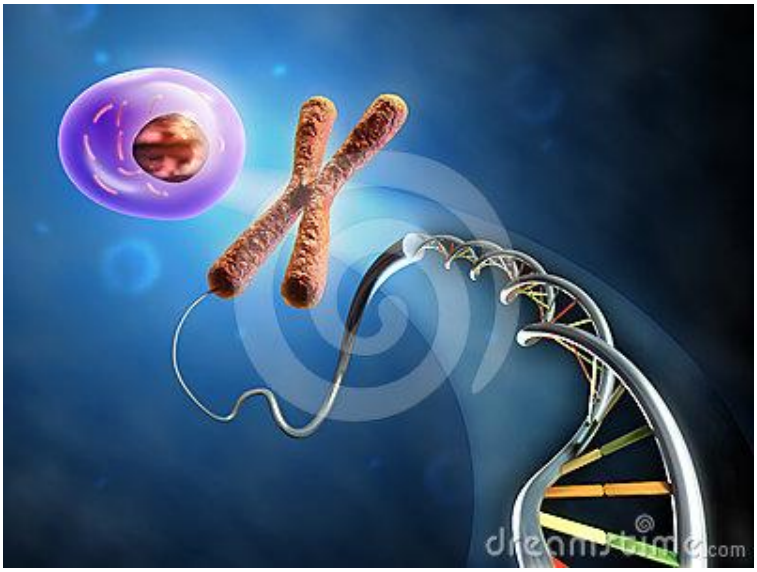
ωοθηκών, του ενδομητρίου και του τραχήλου της μήτρας. Η μεθοδολογία που προτείνεται περιλαμβάνει την εφαρμογή του αλγορίθμου διπλής κατηγοριοποίησης Cheng και Church εφαρμόζοντας σε αυτόν μια σειρά βελτιώσεων ούτως ώστε να εξάγουμε ταυτόχρονα ομάδες γονιδίων με ομοιόμορφη συμπεριφορά αλλά και συγκεκριμένο εύρος τιμών. Για να μελετήσουμε τη συμπεριφορά των γονιδίων, ώστε να διακρίνουμε τις βελτιώσεις που έπρεπε να κάνουμε στον αλγόριθμο, μειώσαμε τον αριθμό των δεδομένων από 33096 στα 1000 γονίδια και εστιάσαμε στα τρία πρώτα εξαγόμενα Biclusters. Συγκεκριμένα, οι βελτιώσεις που κάναμε στον αλγόριθμο αφορούν το d*(i)* και d*(j)*, τα οποία αντιπροσωπεύουν το «mean residue score » για κάθε γραμμή και στήλη, αντίστοιχα. Ύστερα από αρκετές δοκιμές στους τύπους, καταλήξαμε και εφαρμόσαμε τις βελτιωμένες εξισώσεις στον αλγόριθμο διπλής κατηγοριοποίησης Cheng και Church. Εκτελέσαμε τον αλγόριθμο τρεις φορές για κάθε τύπο καρκίνου. Κατά την πρώτη εκτέλεση, εξαγάγαμε τα αποτελέσματα χωρίς να έχουμε κάνει καμία αλλαγή στις εξισώσεις d*(i)* και d*(j)*. Στην δεύτερη εκτέλεση, αλλάξαμε μόνο τον τύπο της εξίσωσης του d*(i*), και τέλος, κατά την τρίτη εκτέλεση, έχοντας κάνει τις επιθυμητές αλλαγές και στις δύο εξισώσεις (d*(i)* και d*(j)*), εξαγάγαμε τα αποτελέσματα. Μελετώντας στη συνέχεια τα αποτελέσματα αυτά, παρατηρήσαμε ότι είχαμε πετύχει τον αρχικό μας στόχο σχετικά με την ομοιόμορφη ομαδοποίηση των γονιδίων και τη μείωση του εύρους τιμών σε κάθε Bicluster.

# 2

# Θεωρητικό Υπόβαθρο

## 2.1. Βιολογικό Υπόβαθρο

Το κύτταρο αποτελεί τη  δομική και λειτουργική μονάδα της ζωής. Στο κύτταρο η λειτουργία και η μορφολογία είναι αλληλένδετες έννοιες, δεδομένου ότι η λειτουργία του κυττάρου καθορίζει τη μορφολογία του κυττάρου και η μορφολογία του κυττάρου εξυπηρετεί τη λειτουργία του. Όλα τα  κύτταρα δομούνται από τις ίδιες χημικές ενώσεις και εκδηλώνουν παρόμοιες μεταβολικές διεργασίες. Τα κύτταρα περιέχουν επίσης το γενετικό υλικό του οργανισμού, το DNA (Εικόνα 2.1.).[5]



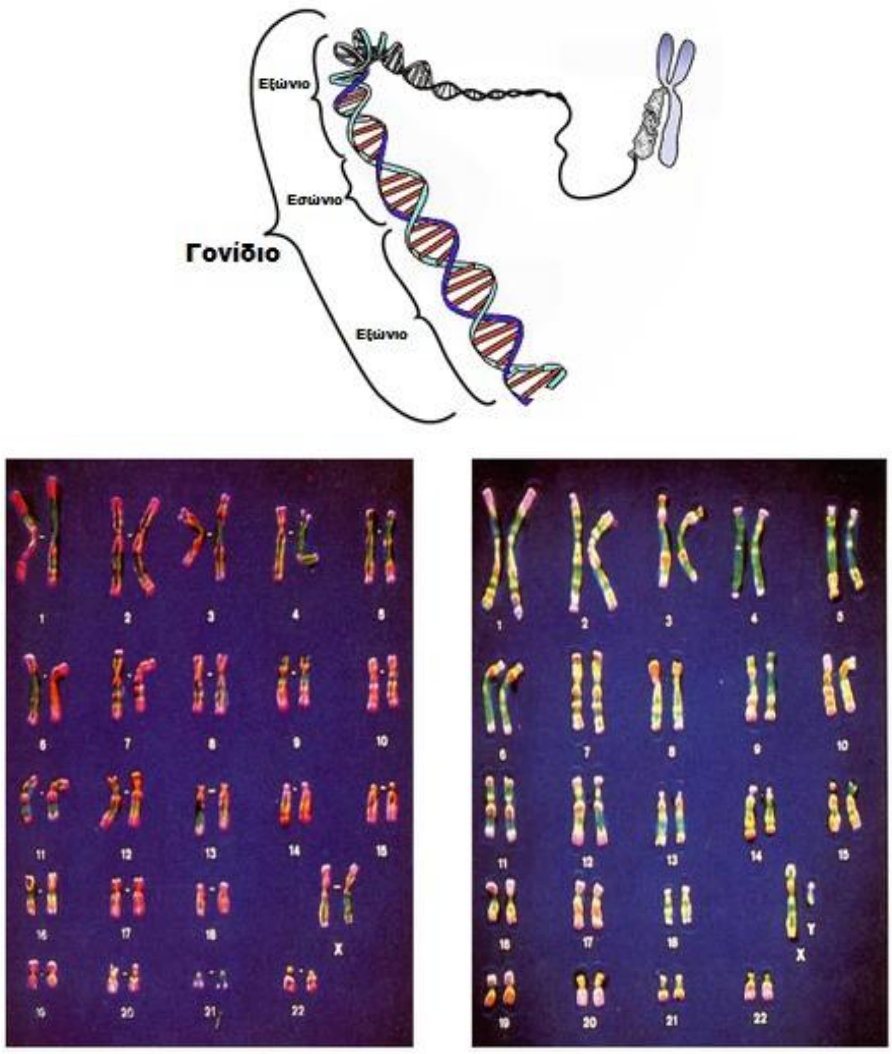
**Εικόνα 2.1.** – Από το κύτταρο στο DNA [21]

Το DNA ενός οργανισμού είναι το μοριακό πρόγραμμα που περιέχει ακριβείς οδηγίες, οι οποίες καθορίζουν τη δομή και τη λειτουργία του οργανισμού. Επιπλέον, περιέχει τις οδηγίες για τον αυτοδιπλασιασμό του. Με τον αυτοδιπλασιασμό επιτυγχάνεται η μεταβίβαση των γενετικών πληροφοριών από ένα κύτταρο (μητρικό) στα θυγατρικά του κύτταρα και από έναν οργανισμό στους απογόνους του μέσω της διαδικασίας της αντιγραφής (replication). Οι εντολές που δέχεται το κύτταρο για να επιτελέσει ορισμένες διαδικασίες μεταφέρονται μέσω των παραγομένων πρωτεϊνών.[5]

### 2.1.1. Γενετικό υλικό και Χρωμοσώματα

Η γενετική πληροφορία είναι η καθορισμένη σειρά των αζωτούχων βάσεων των νουκλεοτιδίων. Η πληροφορία υπάρχει σε τμήματα του DNA με συγκεκριμένη ακολουθία, τα γονίδια.[5] Με άλλα λόγια, γονίδια ονομάζονται οι αλληλουχίες νουκλεοτιδίων τμήματος του DNA, οι οποίες ελέγχουν τα κληρονομικά γνωρίσματα ενός οργανισμού. Οι αλληλουχίες που μεταφράζονται σε αμινοξέα αποτελούν τα εξώνια, ενώ οι ενδιάμεσες αλληλουχίες ονομάζονται εσώνια και δε μεταφράζονται σε αμινοξέα (Εικόνα 2.2 - άνω μέρος). Όλα τα κύτταρα ενός πολυκύτταρου οργανισμού έχουν το ίδιο DNA. Σε κάθε ομάδα κυττάρων όμως εκφράζονται διαφορετικά γονίδια.

Το γενετικό υλικό ενός κυττάρου αποτελεί το γονιδίωμά του. Κύτταρα στα οποία το γονιδίωμα υπάρχει σε ένα μόνο αντίγραφο, ονομάζονται απλοειδή (προκαρυωτικά κύτταρα, γαμέτες διπλοειδών οργανισμών). Τα κύτταρα στα οποία το γονιδίωμα υπάρχει σε δυο αντίγραφα, ονομάζονται διπλοειδή (σωματικά κύτταρα ανώτερων ευκαρυωτικών οργανισμών). Το γονιδίωμα καθορίζει τη γενετική κατασκευή ενός οργανισμού ή ενός κυττάρου, ή τον γονότυπό του. Ο φαινότυπος, είναι το σύνολο των χαρακτηριστικών που εμφανίζει ένας οργανισμός υπό την επιρροή ενός συνόλου περιβαλλοντικών παραγόντων.[22]




**Εικόνα 2.2.** – Γονίδιο (επάνω) και Ανθρώπινος καρυότυπος γυναίκας (κάτω-αριστερά) και άνδρα (κάτω-δεξιά) [23]

Στα ευκαρυωτικά κύτταρα το DNA κατανέμεται στον πυρήνα, στα μιτοχόνδρια και στους χλωροπλάστες (φυτικά κύτταρα). Το συνολικό μήκος του DNA στα διπλοειδή κύτταρα του ανθρώπου είναι περίπου 2 m ($6x10^9$ ζεύγη βάσεων) και συσπειρώνεται σε τέτοιο βαθμό ώστε να χωράει στον πυρήνα του κυττάρου που έχει διάμετρο 10 μm, δηλαδή το DNA είναι ισχυρά «πακεταρισμένο» στο χρωμόσωμα. Από τα παραπάνω προκύπτει ότι η σημαντικότερη λειτουργία των χρωμοσωμάτων είναι να μεταφέρουν τα γονίδια-τις λειτουργικές μονάδες της κληρονομικότητας. Το ένα χρωμόσωμα κάθε ζεύγους είναι πατρικής και το άλλο μητρικής προέλευσης και ελέγχουν τις ίδιες ιδιότητες. Από τα 23 ζεύγη τα 22 είναι μορφολογικά ίδια στα αρσενικά και στα θηλυκά άτομα και ονομάζονται αυτοσωμικά χρωμοσώματα. Το 23ο ζεύγος στα θηλυκά άτομα αποτελείται από δύο Χ χρωμοσώματα, ενώ στα αρσενικά από ένα Χ και ένα Υ χρωμόσωμα. Τα χρωμοσώματα αυτά ονομάζονται φυλετικά και σε πολλούς οργανισμούς, συμπεριλαμβανομένου και του ανθρώπου, καθορίζουν το φύλο. Στον άνθρωπο η παρουσία του Υ χρωμοσώματος καθορίζει το αρσενικό άτομο, ενώ η απουσία του το θηλυκό άτομο.[23] Η παρουσίαση του συνόλου των 46 χρωμοσωμάτων του ανθρώπου αποκαλείται καρυότυπος του ανθρώπου (Εικόνα 2.2 - κάτω μέρος).

## 2.1.2. DNA

Το DNA, όπως και το RNA, είναι ένα μακρομόριο, που αποτελείται από νουκλεοτίδια. Κάθε νουκλεοτίδιο του DNA αποτελείται από μία πεντόζη, τη δεοξυριβόζη, ενωμένη με μία φωσφορική ομάδα και μία αζωτούχο βάση. Στα νουκλεοτίδια του DNA η αζωτούχος βάση μπορεί να είναι μια από τις: αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T) (Εικόνα 2.3).

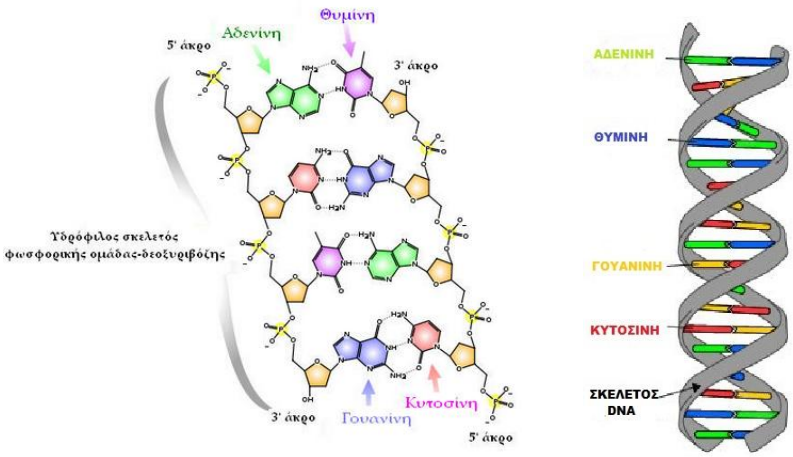

**Εικόνα 2.3.** – Τα τέσσερα διαφορετικά δεοξυριβονουκλεοτίδια του DNA [24]

Σε κάθε νουκλεοτίδιο η αζωτούχος βάση συνδέεται με τον 1' άνθρακα της δεοξυριβόζης και η φωσφορική ομάδα με τον 5' άνθρακα. Μια πολυνουκλεοτιδική αλυσίδα σχηματίζεται από την ένωση πολλών νουκλεοτιδίων με ομοιοπολικό δεσμό. Ο δεσμός αυτός δημιουργείται μεταξύ του υδροξυλίου του 3' άνθρακα της πεντόζης του πρώτου νουκλεοτιδίου και της φωσφορικής ομάδας που είναι συνδεδεμένη στον 5' άνθρακα

της πεντόζης του επόμενου νουκλεοτιδίου. Ο δεσμός αυτός ονομάζεται 3'-5' φωσφοδιεστερικός δεσμός. Με τον τρόπο αυτό η πολυνουκλεοτιδική αλυσίδα που δημιουργείται έχει ένα σκελετό, που αποτελείται από επανάληψη των μορίων φωσφορική ομάδα- πεντόζη-φωσφορική ομάδα-πεντόζη. Ανεξάρτητα από τον αριθμό των νουκλεοτιδίων από τα οποία αποτελείται η πολυνουκλεοτιδική αλυσίδα, το πρώτο της νουκλεοτίδιο έχει πάντα μία ελεύθερη φωσφορική ομάδα συνδεδεμένη στον 5' άνθρακα της πεντόζης του και το τελευταίο νουκλεοτίδιο της έχει ελεύθερο το υδροξύλιο του 3' άνθρακα της πεντόζης του. Για το λόγο αυτό αναφέρεται ότι ο προσανατολισμός της πολυνουκλεοτιδικής αλυσίδας είναι 5'→ 3' (Εικόνα 2.4).[22]

### 2.1.3. Το μοντέλο της διπλής έλικας

Το 1953 οι Watson και Crick διατύπωσαν το μοντέλο της διπλής έλικας του DNA, που αναφέρεται στη δομή του DNA στο χώρο. Σύμφωνα με το μοντέλο αυτό, το DNA αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες που σχηματίζουν στο χώρο μία δεξιόστροφη διπλή έλικα. Η διπλή έλικα έχει ένα σταθερό σκελετό, που αποτελείται από επαναλαμβανόμενα μόρια φωσφορικής ομάδας-δεοξυριβόζης ενωμένων με φωσφοδιεστερικό δεσμό. Ο σκελετός αυτός είναι υδρόφιλος και βρίσκεται στο εξωτερικό του μορίου. Προς το εσωτερικό του σταθερού αυτού σκελετού βρίσκονται οι αζωτούχες βάσεις που είναι υδρόφοβες. Οι αζωτούχες βάσεις της μιας αλυσίδας συνδέονται με δεσμούς υδρογόνου με τις αζωτούχες βάσεις της απέναντι αλυσίδας με βάση τον κανόνα της συμπληρωματικότητας. Η αδενίνη συνδέεται μόνο με θυμίνη και αντίστροφα, ενώ η κυτοσίνη μόνο με γουανίνη και αντίστροφα. Ανάμεσα στη θυμίνη και την αδενίνη αναπτύσσονται δύο δεσμοί υδρογόνου, ενώ ανάμεσα στη κυτοσίνη και τη γουανίνη αναπτύσσονται τρεις δεσμοί υδρογόνου (Εικόνα 2.4).




**Εικόνα 2.4.** – Το μοντέλο της διπλής έλικας του DNA [25]

Οι δεσμοί υδρογόνου που σχηματίζονται μεταξύ των βάσεων σταθεροποιούν τη δευτεροταγή δομή του μορίου. Οι δύο αλυσίδες ενός μορίου DNA είναι συμπληρωματικές, και αυτό υποδηλώνει ότι η αλληλουχία της μιας καθορίζει την αλληλουχία της άλλης. Η συμπληρωματικότητα έχει τεράστια σημασία

για τον αυτοδιπλασιασμό του DNA, μια ιδιότητα που το καθιστά το καταλληλότερο μόριο για τη διατήρηση και τη μεταβίβαση της γενετικής πληροφορίας.[22],[25]

### 2.1.4. RNA

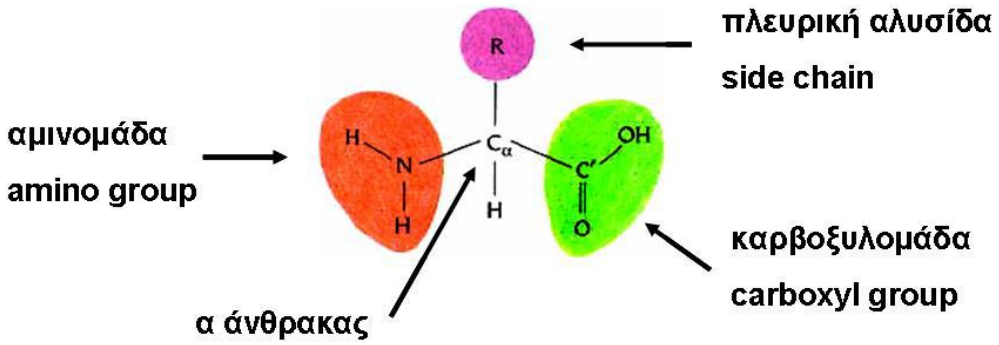
Το RNA ή διαφορετικά ριβονουκλεϊκό οξύ είναι ένα πολυμερές νουκλεϊκό οξύ το οποίο απαρτίζεται από μονομερή νουκλεοτίδια που επιτελούν βασικό ρόλο κατά τη μετάφραση του δεοξυριβονουκλεϊκού οξέος (DNA) σε πρωτεϊνικά προϊόντα. Το RNA λειτουργεί ως ο αγγελιοφόρος του DNA (αγγελιοφόρο RNA ή mRNA). Το RNA εκτός από βασικό στοιχείο των ριβοσωμάτων (rRNA) είναι μεταφορέας αμινοξέων (tRNA - διαθέτει ειδικούς υποδοχείς για την πρόσδεση αμινοξέων και μια περιοχή αντικωδικονίου για την αναγνώριση του κωδικονίου που δεσμεύεται με την ειδική ακολουθία του αγγελιοφόρου RNA με ισχυρούς δεσμούς υδρογόνου) που χρησιμοποιούνται στη διαδικασία της πρωτεϊνοσύνθεσης. Το RNA είναι σχεδόν πανομοιότυπο με το DNA. Η διαφορά τους είναι στο ότι τα μόρια RNA αποτελούνται από ριβόζη αντί για δεοξυριβόζη σαν το κύριο σάκχαρο και επίσης περιέχει τη βάση ουρακίλη αντί για τη θυμίνη που απαντάται στο DNA. Το RNA μεταγράφεται από το DNA με τη βοήθεια κυρίως ενός ενζύμου που ονομάζεται RNA πολυμεράση και στη συνέχεια επεξεργάζεται με έναν αριθμό άλλων δευτερευόντων ενζύμων. Στη συνέχεια χρησιμοποιείται σαν βάση για τη μετάφραση των γονιδίων σε πρωτεΐνες, μεταφέροντας αμινοξέα στα ριβοσώματα για να δημιουργηθούν πρωτεΐνες (πάντα βάσει της αρχής της συμπληρωματικότητας).

Στους ευκαρυωτικούς οργανισμούς το mRNA που παράγεται με τη μεταγραφή υφίσταται μια διαδικασία ωρίμανσης προτού να είναι έτοιμο να προχωρήσει στα ριβοσώματα για τη μετάφραση. Αυτό γίνεται γιατί τα περισσότερα γονίδια είναι ασυνεχή και εκτός των αλληλουχιών που μεταφράζονται υπάρχουν αλληλουχίες οι οποίες δε μεταφράζονται σε αμινοξέα. Οι αλληλουχίες που μεταφράζονται είναι τα εξώνια και εκείνες που δε μεταφράζονται είναι τα εσώνια. Έτσι, το mRNA που μόλις έχει σχηματιστεί από τη μεταγραφή ενός γονιδίου ονομάζεται πρόδρομο mRNA και περιέχει εσώνια και εξώνια. Το πρόδρομο mRNA μετατρέπεται σε mRNA με τη διαδικασία της ωρίμανσης κατά την οποία τα εσώνια κόβονται από μικρά ριβοζονουκλεοπρωτεϊνικά σωματίδια (snRNPs), που λειτουργούν ως ένζυμα. Τα εξώνια που απομένουν συρράπτονται μεταξύ τους και με αυτό τον τρόπο σχηματίζεται το ώριμο mRNA που μεταφέρεται στα ριβοσώματα για την πρωτεϊνοσύνθεση. Δύο περιοχές του ώριμου mRNA δε μεταφράζονται σε αμινοξέα, η μια βρίσκεται στο άκρο 5΄ και η άλλη στο άκρο 3΄.[22]

### 2.1.5. Πρωτεΐνες

Οι πρωτεΐνες είναι μακρομόρια, πολυμερή, με σαφώς καθορισμένη δομή και παρουσιάζουν μεγάλη ποικιλία ιδιοτήτων και σχήματος. Οι πρωτεΐνες είναι απαραίτητες σε όλες σχεδόν τις βιολογικές λειτουργίες και η σημασία τους συνοψίζεται στα εξής: ενζυμική κατάλυση, μεταφορά, αποθήκευση, κίνηση, μηχανική στήριξη, ανοσοπροστασία, δημιουργία και μετάδοση νευρικών παλμών και έλεγχος της ανάπτυξης, της διαφοροποίησης.

Οι δομικές μονάδες των πρωτεϊνών είναι 20 διαφορετικά αμινοξέα. Τα αμινοξέα αποτελούνται από μία αμινομάδα (-NH2) και μία καρβοξυλομάδα (-COOH) συνδεδεμένες σε ένα άτομο άνθρακα $C_{\alpha}$. Στο ίδιο άτομο ενώνεται και μία πλευρική αλυσίδα, η οποία διαφέρει μεταξύ των αμινοξέων και καθορίζει τις ιδιότητές τους. Τα αμινοξέα ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς για τη δημιουργία πολυπεπτιδικών αλυσίδων - πρωτεϊνών. Τα πρωτεϊνικά μόρια αποτελούνται από μια ή περισσότερες πολυπεπτιδικές αλυσίδες, που η κάθε μία τους μπορεί να αποτελείται από μερικές εκατοντάδες αμινοξέα με μια συγκεκριμένη σειρά γνωστή σαν αμινοξική ακολουθία.




**Εικόνα 2.5. –** Αμινοξύ – δομική μονάδα πρωτεϊνών

Στις πρωτεΐνες μπορούν να αναγνωριστούν 4 επίπεδα οργάνωσης:

**α. Πρωτοταγής δομή:** αντιστοιχεί στην αμινοξική ακολουθία, δηλαδή στην διάταξη των αμινοξέων σε μια πολυπεπτιδική αλυσίδα.

**β. Δευτεροταγής δομή:** αναφέρεται στην κανονική στερεοδιάταξη τμημάτων της πολυπεπτιδικής αλυσίδας.

**γ. Τριτοταγής δομή:** αναφέρεται στην τρισδιάστατη δομή.

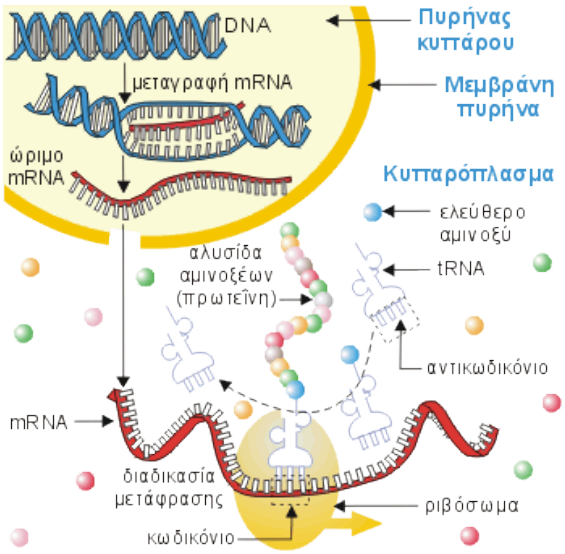
**δ. Τεταρτογενής Δομή:** η διάταξη στο χώρο των πολυπεπτιδικών αλυσίδων από τις οποίες αποτελείται.

Η ακολουθία μιας πρωτεΐνης καθορίζει την τρισδιάστατη δομή της, η οποία με τη σειρά της καθορίζει τις αλληλεπιδράσεις της με άλλες πρωτεΐνες, με νουκλεϊνικά οξέα και μικρά μόρια. Επομένως, η δομή μιας πρωτεΐνης καθορίζει τη λειτουργία της και οι ερευνητικές προσπάθειες κατευθύνονται τόσο στον καθορισμό των αμινοξικών ακολουθιών της κάθε πρωτεΐνης όσο και την εύρεση της τρισδιάστατης της δομής.[22],[26]

### 2.1.6. Γονιδιακή Έκφραση

Γονιδιακή έκφραση καλείται η διαδικασία κατά την οποία η πληροφορία από ένα γονίδιο χρησιμοποιείται για την σύνθεση ενός λειτουργικού γονιδιακού προϊόντος. Τα προϊόντα αυτά είναι συχνά πρωτεΐνες. Εκτός από τις πρωτεΐνες, ως λειτουργικό γονιδιακό προϊόν θεωρείται επίσης το RNA (π.χ. rRNA και tRNA). Ένα γονίδιο εκφράζεται μέσω των μηχανισμών της μεταγραφής και της μετάφρασης. Το πρώτο βήμα για την έκφραση της πληροφορίας που υπάρχει στο DNA είναι η μεταβίβασή της στο RNA μέσω της διαδικασίας της μεταγραφής (transcription). Στη συνέχεια, το RNA, μέσω της διαδικασίας της μετάφρασης (translation), μεταφέρει την πληροφορία στις πρωτεΐνες που είναι υπεύθυνες για τη δομή και λειτουργία των κυττάρων και κατ' επέκταση και των οργανισμών. Έτσι, οι πορείες της μεταγραφής και της μετάφρασης των γονιδίων αποτελούν τη γονιδιακή έκφραση (Εικόνα 2.6).[27] Η διαδικασία της γονιδιακής έκφρασης πραγματοποιείται σε όλους τους έμβιους οργανισμούς, ευακρυωτικούς, προκαρυωτικούς και στους ιούς, προκειμένου να παραχθεί ο μακρομοριακός μηχανισμός της ζωής. Στην γενετική η γονιδιακή έκφραση αποτελεί το θεμελιώδες επίπεδο μέσω του οποίου ο γονότυπος δίνει γέννηση στο φαινότυπο. Ο γενετικός κώδικας ερμηνεύεται μέσω της γονιδιακής έκφρασης και οι ιδιότητες των εκφραζόμενων προϊόντων μας δίνουν τον φαινότυπο του οργανισμού.

Η μέτρηση της γονιδιακής έκφρασης, δηλαδή η ποσοτικοποίηση του επιπέδου έκφρασης ενός γονιδίου μέσα σε ένα κύτταρο, αποτελεί μία τεράστια ποσότητα πληροφορίας.




**Εικόνα 2.6. -** Μηχανισμός σύνθεσης πρωτεϊνών στα ευκαρυωτικά κύτταρα [27]

Η πληροφορία της γονιδιακής έκφρασης εστιάζει είτε:

**(α)** στον καθορισμό των ιστών οι οποίοι συνδέονται με τη φυσιολογική λειτουργία ενός συγκεκριμένου γονιδίου, είτε

**(β)** στον καθορισμό των παραγόντων εκείνων που ρυθμίζουν την έκφραση ενός συγκεκριμένου γονιδίου, οι οποίοι συνήθως χωρίζονται σε διατροφικούς, ορμονικούς, ή περιβαλλοντικούς.[22]

Οπότε έχοντας την παραπάνω πληροφορία δύνανται να πραγματοποιηθεί πρόβλεψη για ένα άτομο εάν έχει προδιάθεση για κάποια μορφή καρκίνου ή να βρεθεί εάν κάποια κύτταρα παρουσιάζουν ανθεκτικότητα σε κάποιους φαρμακευτικούς παράγοντες κ.α..

Όπως αναφέρθηκε, ο όρος γονιδιακή έκφραση αναφέρεται συνήθως σε όλη τη διαδικασία με την οποία ένα γονίδιο ενεργοποιείται για να παραγάγει μια πρωτεΐνη. Όμως σε κάθε κύτταρο δεν παράγονται όλες οι πρωτεΐνες σε κάθε χρονική στιγμή. Επιπλέον, επειδή το κύτταρο χρειάζεται κάθε πρωτεΐνη σε συγκεκριμένη ποσότητα, οι πρωτεΐνες ενός κυττάρου δεν παράγονται σε ίσες ποσότητες. Έτσι, είναι απαραίτητη η ύπαρξη και η λειτουργία ενός προγράμματος ρύθμισης της γονιδιακής έκφρασης, που παρέχει τις οδηγίες για το είδος και την ποσότητα των πρωτεϊνών οι οποίες πρέπει να παραχθούν σε κάθε συγκεκριμένη χρονική στιγμή. Στα ευκαρυωτικά κύτταρα η γονιδιακή έκφραση ρυθμίζεται σε τέσσερα επίπεδα:

- **Στο επίπεδο της μεταγραφής:**

Ένας αριθμός μηχανισμών ελέγχουν ποια γονίδια θα μεταγραφούν ή και με ποια ταχύτητα θα γίνει η μεταγραφή. Το DNA των ευκαρυωτικών κυττάρων δεν οργανώνεται σε οπερόνια αλλά κάθε γονίδιο έχει τον δικό του υποκινητή και μεταγράφεται αυτόνομα. Η μεταγραφή περιλαμβάνει τη μεταφορά της γενετικής πληροφορίας από το DNA μέσω ενζυματικής σύνθεσης μιας συμπληρωματικής αλυσίδας RNA που καταλύεται από το ένζυμο της RNA πολυμεράσης. Η RNA πολυμεράση λειτουργεί (όπως και στους προκαρυωτικούς οργανισμούς) με τη βοήθεια πρωτεϊνών, που ονομάζονται μεταγραφικοί παράγοντες. Μόνο που στους ευκαρυωτικούς οργανισμούς οι μεταγραφικοί παράγοντες παρουσιάζουν τεράστια ποικιλία. Κάθε κυτταρικός τύπος περιέχει διαφορετικά είδη μεταγραφικών παραγόντων. Διαφορετικός συνδυασμός μεταγραφικών παραγόντων ρυθμίζει τη μεταγραφή κάθε γονιδίου. Μόνο όταν ο σωστός συνδυασμός των μεταγραφικών παραγόντων προσδεθεί στον υποκινητή ενός γονιδίου, αρχίζει η RNA πολυμεράση τη μεταγραφή ενός γονιδίου.

- **Στο επίπεδο μετά τη μεταγραφή:**

Περιλαμβάνονται οι μηχανισμοί με τους οποίους γίνεται η ωρίμανση του πρόδρομου mRNA και καθορίζεται η ταχύτητα με την οποία το ώριμο mRNA αφήνει τον πυρήνα και εισέρχεται στο κυτταρόπλασμα.

- **Στο επίπεδο της μετάφρασης:**

Ο χρόνος που "ζουν" τα μόρια mRNA στο κυτταρόπλασμα δεν είναι ο ίδιος για όλα τα είδη RNA, επειδή μετά από κάποιο χρονικό διάστημα αποικοδομούνται. Επίσης, ποικίλλει και η ικανότητα πρόσδεσης του mRNA στα ριβοσώματα.

- **Στο επίπεδο μετά τη μετάφραση:**

Ακόμη και όταν πραγματοποιηθεί η πρωτεϊνοσύνθεση και παραχθεί η κατάλληλη πρωτεΐνη, μπορεί να χρειαστεί να υποστεί τροποποιήσεις, για να γίνει βιολογικά λειτουργική.

Μερικά γονίδια που κωδικοποιούν πρωτεΐνες μεταγράφονται λιγότερο ή περισσότερο συχνά, και ονομάζονται «housekeeping» γονίδια και απαιτούνται πάντα για τις βασικές αντιδράσεις. Άλλα γονίδια δεν μεταγράφονται ή, μεταγράφονται για συγκεκριμένες λειτουργίες του οργανισμού, μόνο σε ιδιαίτερες στιγμές και κάτω από ιδιαίτερες εξωτερικές συνθήκες. Το σήμα που «καλύπτει» ή «αποκαλύπτει» ένα γονίδιο μπορεί

να προέλθει από το εξωτερικού του κυττάρου, όπως μια θρεπτική ουσία ή μια ορμόνη. Πρόσθετες ρυθμιστικές ακολουθίες στο DNA υπαγορεύουν εάν ένα γονίδιο θα ανταποκριθεί στα σήματα και στη συνέχεια επηρεάζουν την μεταγραφή του γονιδίου που κωδικοποιεί την πρωτεΐνη. Η αποκωδικοποίηση του ανθρώπινου γονιδιώματος (Human Genome Project) αλλά και άλλων σημαντικών οργανισμών αποτέλεσε σταθμό στη Βιολογία, μεταβάλλοντας ριζικά τις πρακτικές της και την οπτική της έρευνας.[22]

**Προφίλ γονιδιακής έκφρασης και ανάλυση μεταγραφώματος**

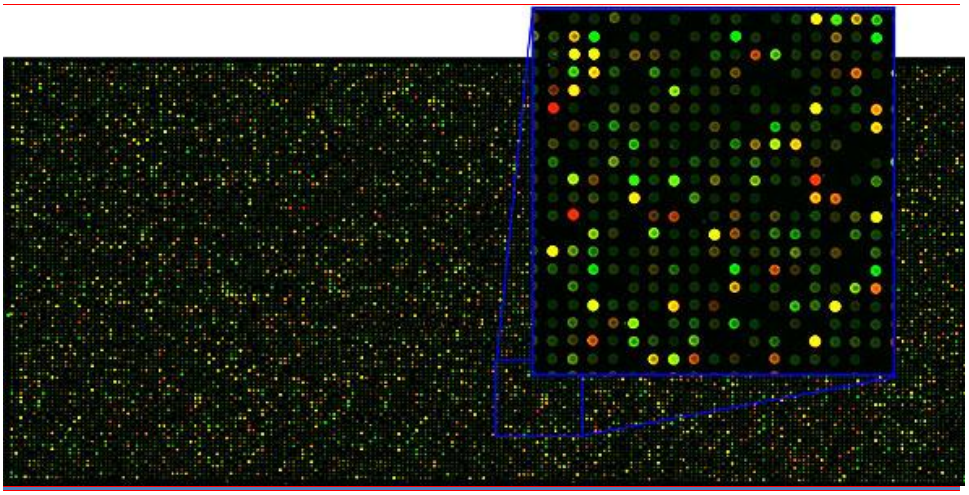
Το προφίλ γονιδιακής έκφρασης είναι ο προσδιορισμός του προτύπου γονιδίων, που εκφράζονται στο επίπεδο της μεταγραφής, κάτω από ειδικές περιστάσεις ή σε ένα ειδικό κύτταρο, το οποίο αποτυπώνει μια συνολική εικόνα της κυτταρικής λειτουργίας.

Το μεταγράφωμα μετράει το αγγελιαφόρο RNA (mRNA), δηλαδή το μήνυμα που μεταφράζεται σε πρωτεΐνη. Σε αντίθεση με το γονιδίωμα, το μεταγράφωμα δεν λέει ποιά γονίδια είναι παρόντα αλλά ποιά γονίδια μεταγράφονται και μεταφράζονται σε πρωτεΐνες. Δεν υπάρχει ακριβής συσχέτιση, επειδή υψηλά επίπεδα mRNA δεν μεταφράζονται αναγκαστικά σε αντίστοιχα υψηλά επίπεδα παραγωγής πρωτεϊνών και τα υψηλά επίπεδα παραγωγής πρωτεϊνών μπορούν να ακολουθήσουν χαμηλά επίπεδα mRNA. Επιπλέον, κάποιο mRNA δεν μεταφράζεται σε πρωτεΐνη και διαφορετικές μορφές της ίδιας πρωτεΐνης μπορεί να προκύψουν από ένα δεδομένο μετάγραφο.[28] Σήμερα, η μελέτη του μεταγραφώματος υπερτερεί της μελέτης του πρωτεώματος, και ενδείκνυται για την ποσοτικοποίηση της γονιδιακής έκφρασης σε πολλές περιπτώσεις. Η μεταγραφή των mRNA, όπως αναφέρεται παραπάνω, αποτελεί το πρώτο χρονικά βήμα, και είναι αυτό που συγκεντρώνει και τα περισσότερα πλεονεκτήματα για αξιόπιστες και αποτελεσματικές μετρήσεις γονιδιακής έκφρασης. Το mRNA απομονώνεται εύκολα και σε ποσότητες ικανές για αποτελεσματική ποσοτικοποίηση, οι μεθοδολογίες που υπάρχουν μπορούν να ενισχύσουν το δείγμα και να ταυτοποιήσουν έτσι ακόμα και μόρια mRNA που βρίσκονται σε πολύ μικρές συγκεντρώσεις, ενώ, επιπλέον, απ' όσο γνωρίζουμε η μεταγραφή είναι το στάδιο στη διαδικασία της έκφρασης γονιδίων, που υπόκειται στην πιο αυστηρή ρύθμιση. Με βάση τα παραπάνω, οι μετρήσεις σε επίπεδο mRNA, εξασφαλίζουν ομοιογένεια του δείγματος, είναι πιο εύκολες πειραματικά και μπορούν να είναι ποσοτικά ακριβείς. Επιπλέον, τα επίπεδα του mRNA αντανακλούν πιο άμεσα τις αλλαγές που επισυμβαίνουν στην έκφραση των γονιδίων, καθώς ο χρόνος που μεσολαβεί από τη στιγμή που ένα γονίδιο ενεργοποιείται ως την παραγωγή της πρωτεΐνης που κωδικοποιεί μπορεί συχνά να είναι απαγορευτικός για την μελέτη ταχείας απόκρισης σε ερεθίσματα. Τέλος, μια σειρά από μελέτες έχουν δείξει ότι σε σταθερές συνθήκες (steady state) τα επίπεδα των mRNA συσχετίζονται σε μεγάλο βαθμό με τα αντίστοιχα των πρωτεϊνών και παρά το γεγονός ότι ο βαθμός συσχέτισης είναι μειωμένος για τα mRNA που βρίσκονται σε χαμηλή συγκέντρωση, η ποσοτικοποίηση των mRNA θεωρείται ο πιο διαδεδομένος, άμεσος και συστηματικός τρόπος για τη μελέτη της γονιδιακής έκφρασης.[29] Από τις βασικές μεθοδολογίες για την πειραματική μελέτη των επιπέδων mRNA αποτελούν οι μικροσυστοιχίες DNA (DNA microarrays).

## 2.2. Εισαγωγή στην τεχνολογία των Μικροσυστοιχιών (microarrays)

Οι **μικροσυστοιχίες** (microarrays) είναι μία διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια και ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μία στερεή επιφάνεια (συνήθως

γυάλινη). Χρησιμοποιούνται για τη μέτρηση DNA ή χρησιμοποιούν DNA για το σύστημα ανίχνευσής τους. Ποσοτικές ή ποιοτικές μετρήσεις με μικροσυστοιχίες γονιδίων εκμεταλλεύονται την εκλεκτική φύση της αρχής της συμπληρωματικότητας μεταξύ νουκλεϊκών οξέων DNA-DNA ή DNA-RNA ή πρόσφατα και μεταξύ των αμινοξέων των πρωτεϊνών, υπό αυστηρά ελεγχόμενες συνθήκες θερμοκρασίας και με τη χρήση φθορίζουσων ουσιών. Οι μικροσυστοιχίες γονιδίων χρησιμοποιούνται σήμερα κατά κόρον για την εξέταση της γονιδιακής έκφρασης υπό ειδικές συνθήκες και για την ανίχνευση νουκλεϊκών οξέων παθογόνων οργανισμών π.χ. επιβλαβών ιών σε δείγματα ελέγχου. Οι μικροσυστοιχίες αποτελούν πολύτιμη μέθοδο της μοριακής βιολογίας, δεδομένου ότι δύνανται να εξετάζουν ταυτόχρονα την έκφραση χιλιάδων γονιδίων, και ενδείκνυνται για συγκριτικές μελέτες γονιδιωμάτων. Τα μειονεκτήματα της μεθόδου έγκεινται στο υψηλό κόστος και στη συχνή ανακρίβεια των αποτελεσμάτων λόγω τεχνικών προβλημάτων όπως η μη ειδική υβριδοποίηση φθορίζουσων χρωστικών σε λάθος γονίδια κτλ. Σήμερα, η χρήση μικροσυστοιχιών γονιδίων από την επιστημονική κοινότητα απαιτεί την παράλληλη χρήση και άλλων συμπληρωματικών μεθόδων για την επαλήθευση των αποτελεσμάτων του τσιπ όπως είναι το στύπωμα Northern (Northern Blot) και η ποσοτική αλυσιδωτή αντίδραση πολυμεράσης με τη χρήση αντίστροφης μεταγραφάσης σε πραγματικό χρόνο (Quantitative Real time (reverse transcriptase) PCR)).[30]



**Εικόνα 2.7.** - Μικροσυστοιχίες DNA [31]

Με τη χρήση της τεχνολογίας των μικροσυστοιχιών DNA κατέστη εφικτή η ταυτόχρονη παρακολούθηση εκατοντάδων γονιδίων κατά την διάρκεια σημαντικών βιολογικών διαδικασιών. Η καταγραφή και η εύρεση κρυμμένων προτύπων στα δεδομένα έκφρασης των γονιδίων βοήθησε πολύ στην κατανόηση των γενετικών λειτουργιών. Από την άλλη ο μεγάλος αριθμός γονιδίων και η πολυπλοκότητα των βιολογικών δικτύων αύξησε τις προκλήσεις για κατανόηση και μετάφραση των μαζικών αποτελεσμάτων, τα οποία συχνά προκύπτουν από χιλιάδες μετρήσεις. Για αυτό το λόγο με τη χρήση τεχνικών κατηγοριοποίησης (συσταδοποίησης), είναι εφικτή η εύρεση φυσικών δομών και προτύπων. Η ανάλυση συστάδων (cluster analysis) κατηγοριοποιεί τα δεδομένα σε ομάδες βάσει κοινών χαρακτηριστικών, ώστε τα μέλη κάθε ομάδας να μοιάζουν περισσότερο μεταξύ τους από ότι με μέλη άλλων ομάδων. Οι αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται ευρέως στην αναγνώριση βιολογικά όμοιων ομάδων γονιδίων και δειγμάτων.

Σε αντίθεση με τις κοινές τεχνικές στην γονιδιακή έρευνα η οποία επικεντρωνόταν στην τοπική μελέτη και συλλογή δεδομένων για μεμονωμένα γονίδια, με τη χρήση των μικροσυστοιχιών έγινε εφικτή η

ταυτόχρονη παρακολούθηση των επιπέδων έκφρασης χιλιάδων γονιδίων παράλληλα . Οι μικροσυστοιχίες διαφέρουν από την κλασσική δομή των πινάκων όπου αποτελούνται από σειρές με απλά δεδομένα, στο ότι υπολογίζουν την «ποσότητα του στόχου» έμμεσα μετρώντας μια άλλη φυσική ποσότητα – την ένταση του φθορισμού κάθε τοποθεσίας (spot) στον πίνακα αναπαράστασης φθορισμού. Έτσι στην πραγματικότητα οι σειρές δεδομένων που παράγονται από τις μικροσυστοιχίες είναι μονόχρωμες εικόνες. Μετασχηματίζοντας αυτές τις εικόνες μέσα στον «πίνακα γονιδιακής έκφρασης» (gene expression matrix), το οποίο αποτελεί μια δύσκολη διαδικασία, μπορούμε να αναλύσουμε αυτόν τον πίνακα και να προσπαθήσουμε να εξάγουμε από αυτόν κάποια συμπεράσματα και γνώσεις σχετικά με θεμελιώδης βιολογικές διαδικασίες.[32],[33]

Δύο είναι οι βασικοί τρόποι πειραματισμού:

- cDNA μικροπίνακες
- ολιγονουκλεοτιδικοί πίνακες (oligo arrays).

Τα κοινά στάδια των δύο πρωτοκόλλων είναι τα παρακάτω:

Κατασκευή chip: ένα microarray είναι ένα μικρό chip (από nylon, γυαλί ή σιλικόνη) πάνω στο οποίο προσαρμόζονται μόρια DNA σε καθορισμένο πλέγμα. Κάθε κελί του πλέγματος συνδέεται με μία ακολουθία DNA.

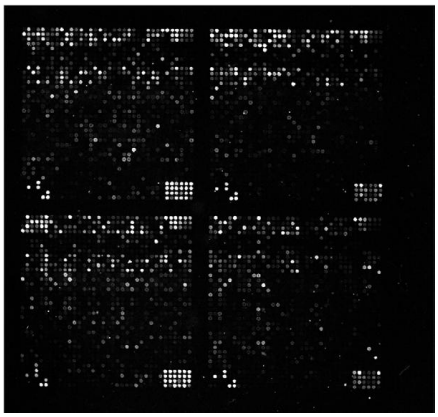
Προετοιμασία, χαρακτηρισμός και υβριδοποίηση στόχου: δύο δείγματα mRNA (ένα δείγμα ελέγχου και ένα πειραματικό δείγμα) μεταγράφονται σε cDNA (στόχοι), μαρκάρονται με ραδιοϊσότοπα ή φθοριούχες ενώσεις, και υβριδοποιούνται με τα μόρια στην επιφάνεια του πλέγματος.

Τα chips σαρώνονται για να διαβαστεί η ένταση του σήματος που εκπέμπουν από τους μαρκαρισμένους υβριδοποιημένους στόχους.

Και οι δύο μέθοδοι μετράνε την έκφραση των επιπέδων για κάθε ακολουθία DNA από τον λόγο της έντασης των σημάτων μεταξύ του δείγματος ελέγχου και του πειραματικού δείγματος.[32],[33]

Οι μικροσυστοιχίες μπορούν να χρησιμοποιηθούν για να μετρήσουν τη γονιδιακή έκφραση με πολλούς τρόπους αλλά μια από τις πιο διάσημες εφαρμογές τους είναι η σύγκριση των επιπέδων γονιδιακής έκφρασης μεταξύ δύο διαφορετικών δειγμάτων.(πχ ίδια κύτταρα κάτω από διαφορετικές συνθήκες). Οι μικροσυστοιχίες έχουν εφαρμοστεί επιτυχώς για τη λύση διαφόρων βιολογικών προβλημάτων και όσο οι συστοιχίες αυτές γίνονται περισσότερο διαθέσιμες στους ερευνητές τόσο η δημοτικότητα τους θα αυξηθεί. Για τη σωστή ανάλυση των πειραματικών δεδομένων είναι απαραίτητο να κατανοήσουμε τα πειράματα που αναπαράγουν αυτά τα δεδομένα. Τα περάματα μικροσυστοιχιών αποτελούνται από πολλά βήματα, το καθένα από τα οποία εισάγει τον δικό του θόρυβο και τη δική του διακύμανση στο σύστημα. Το τελικό αποτέλεσμα μπορεί να επηρεαστεί από το τυχαίο σφάλμα κάθε βήματος.

Για αυτόν τον λόγο ένας σωστός πειραματικός σχεδιασμός και μια προσεκτική στατιστική ανάλυση απαιτούνται για τη σωστή ερμηνεία των δεδομένων των μικροσυστοιχιών.[34]



**Εικόνα 2.8.** – Σάρωση και υβριδοποίηση μικροσυστοιχίας που περιέχει περισσότερα από 5000 γονίδια κάθε σημείο χαρακτηρίζει ένα σύνολο από πανομοιότυπα μονοκλωνικά μόρια DNA που αναπαριστούν ένα μόνο γονίδιο. Η ένταση της φωτεινότητας δείχνει την ποσότητα φθορίζουσας ουσίας που περιέχει το υβριδοποιημένο mRNA

## 2.3. Ανάλυση γονιδιακής έκφρασης από μικροσυστοιχίες

Τα πειράματα μικροσυστοιχιών καθορίζουν την έκφραση μεταγραφής των γονιδίων ενός οργανισμού υπό διαφορετικές καταστάσεις. Η ανάλυση μικροσυστοιχιών προσπαθεί να εντοπίσει ομάδες γονιδίων που παρουσιάζουν παρόμοια συμπεριφορά υπό ορισμένες συνθήκες. Τα δεδομένα που προκύπτουν από τα πειράματα με μικροσυστοιχίες έχουν συνήθως πολύ θόρυβο. Γι' αυτό τον λόγο ο αλγόριθμος κατηγοριοποίησης πρέπει να είναι ικανός να ξεχωρίζει τη χρήσιμη πληροφορία από τον θόρυβο. Τα τελευταία χρόνια, μία από τις κύριες μεθόδους για την ανάλυση των δεδομένων μικροσυστοιχιών είναι η διπλή κατηγοριοποίηση, μια πολύ ανεπτυγμένη τεχνική. Η τεχνική biclustering υπερέχει της παραδοσιακής τεχνικής ομαδοποίησης λόγω των δύο κύριων χαρακτηριστικών της: την ταυτόχρονη ομαδοποίηση γονιδίων και συνθηκών και την επικάλυψη. Ταυτόχρονη ομαδοποίηση σημαίνει ότι οι ομάδες διπλής κατηγοριοποίησης (οι ομάδες που βρέθηκαν με τους αλγορίθμους biclustering) ομαδοποιούν γονίδια με παρόμοια συμπεριφορά κάτω από έναν ορισμένο αριθμό συνθηκών (έτσι, το Bicluster θα ομαδοποιεί γονίδια και συνθήκες), ενώ οι παραδοσιακές τεχνικές ομαδοποίησης συγκεντρώνουν μόνο γονίδια με παρόμοια συμπεριφορά σε όλες τις συνθήκες(ή αντιστρόφως). Αυτό το χαρακτηριστικό καθιστά τις ομάδες διπλής κατηγοριοποίησης κατάλληλες για την αποτύπωση της βιολογικής συμπεριφοράς σε διάφορες περιστάσεις, για παράδειγμα, όταν μια ενδιαφέρουσα κυτταρική διαδικασία είναι ενεργή μόνο σε ένα υποσύνολο συνθηκών. Αν και είναι ασυνήθιστο το γεγονός ότι τα υποσύνολα των γονιδίων που ομαδοποιούνται με δύο διαφορετικές ομάδες να διατέμνονται, η αλληλοεπικάλυψη είναι ένα εγγενές χαρακτηριστικό των Biclusters. Αν δύο ομάδες διπλής κατηγοριοποίησης *B1* και *B2* που ομαδοποιούν τα γονίδια *G1* και *G2* και τις συνθήκες *C1* και *C2*, αντίστοιχα, έχουν $G1 \cap G2 \neq \emptyset$ και / ή $C1 \cap C2 \neq \emptyset$ λέγεται ότι οι *B1* και *B2* αλληλεπικαλύπτονται. Η αλληλοεπικάλυψη δίνει στις ομάδες διπλής κατηγοριοποίησης την ευελιξία να αναπαριστούν βιολογικές καταστάσεις όπως είναι τα γονίδια που συμμετέχουν σε πολλαπλά μονοπάτια που λειτουργούν κάτω από ένα υποσύνολο συνθηκών.[35] Τις τελευταίες δεκαετίες παρουσιάζεται ραγδαία αύξηση του όγκου των βιολογικών δεδομένων και η ανάγκη για τον μετασχηματισμό της αδόμητης πληροφορίας που προκύπτει από αυτά τα δεδομένα σε γνώση. Αυτός ήταν ο κύριος λόγος δημιουργίας του νέου επιστημονικού πεδίου της Βιοπληροφορικής (Bioinformatics) που θα αναλυθεί στο επόμενο υποκεφάλαιο.

## 2.4. ΒιοΠληροφορική

Η Βιοπληροφορική αποτελεί στην ουσία σύμπραξη επιστημονικών κλάδων όπως η βιολογία, η βιοχημεία, τα μαθηματικά, η επιστήμη των υπολογιστών, η τεχνολογία πληροφοριών και άλλα. Το έργο της Βιοπληροφορικής συνίσταται στην έρευνα, ανάπτυξη, ή εφαρμογή υπολογιστικών εργαλείων και προσεγγίσεων για την επέκταση της χρήσης δεδομένων βιολογίας, ιατρικής, συμπεριφοράς ή υγείας, συμπεριλαμβανομένων εκείνων για την απόκτηση, αποθήκευση, οργάνωση, αρχειοθέτηση, ανάλυση, ή οπτικοποίηση αυτών των δεδομένων.[36] Για παράδειγμα, με τη βοήθεια της Βιοπηροφορικής μέσα από τη χρήση εξελιγμένων συστημάτων βάσεων δεδομένων, αλγορίθμων ομαδοποίησης και μια σειρά από στατιστικά εργαλεία αναγνωρίζεται η ταυτότητα των βιολογικών δεδομένων, κατανοούνται βιολογικοί μηχανισμοί και ασθένειες, και σχεδιάζονται φάρμακα.[26]

Ένα από τα πιο σημαντικά ίσως ερευνητικά πεδία της Βιοπληροφορικής είναι η διαχείριση και εξόρυξη γνώσης με τις βιολογικές βάσεις δεδομένων να αποτελούν θεμελιώδες τμήμα τους.[31],[37] Οι βιολογικές βάσεις δεδομένων είναι μεγάλα, οργανωμένα συστήματα δεδομένων που συνδέονται συνήθως με κατάλληλο λογισμικό για την ενημέρωση, αναζήτηση, και ανάκτηση στοιχείων των δεδομένων που έχουν αποθηκευθεί στο σύστημα. Από αυτά τα αποθετήρια γνώσης, ένας ερευνητής μπορεί να αντλήσει τις κατάλληλες πληροφορίες (MEDLINE, GenBank)[38],[39] ή τα δεδομένα (GEO)[40] πάνω στα οποία θα εφαρμόσει την ανάλυση και τις μεθοδολογικές προσεγγίσεις για την επίλυση ενός ή πολλαπλών ερωτημάτων. Γενικά, οι βιολογικές βάσεις δεδομένων μπορούν να διακριθούν σε δυο κύριες κατηγορίες, στις πρωτογενείς και τις δευτερογενείς βάσεις δεδομένων. Η πρώτη κατηγορία απαρτίζεται από βάσεις δεδομένων με πρωτογενή στοιχεία, τα οποία απορρέουν από διάφορες πειραματικές μεθοδολογίες, ενώ η δεύτερη από βάσεις δεδομένων με στοιχεία που δημιουργούνται από ταξινομήσεις των πρωτογενών δεδομένων και είναι χρήσιμα για αναλυτικούς σκοπούς. Ενδεικτικές βιολογικές βάσεις δεδομένων της πρώτης και δεύτερης κατηγορίας είναι οι ακόλουθες:

**Πρωτογενείς βάσεις δεδομένων**

**Βάσεις δεδομένων ακολουθιών**

Βάσεις νουκλεοτιδικών ακολουθιών ελεύθερα διαθέσιμες, οι οποίες συνεργάζονται μεταξύ τους ανταλλάσσοντας εγγραφές και δημιουργώντας κοινούς κανόνες για την ταξινόμηση και το σχολιασμό των δεδομένων:

- DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/) στο Center for Information Biology (CIB).
- GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) στο National Center for Biotechnology Information (NCBI).
- EMBL_Bank (http://www.ebi.ac.uk/embl/index.html) στο European Bioinformatics Institute (EBI).

*Εξειδικευμένες βάσεις δεδομένων* που συνδυάζουν τα δεδομένα γονιδιωματικών ακολουθιών και το σχολιασμό τους με άλλα στοιχεία για τα συγκεκριμένα είδη.

- Ensembl ( http://www.ensembl.org/index.html ) αποτέλεσμα συνεργασίας του EBI και του Wellcome Trust Sanger Institute.

- Entrez Genomes (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome) στο National Center for Biotechnology Information (NCBI).

Βάσεις πρωτεϊνικών ακολουθιών που παρέχουν υψηλό επίπεδο σχολιασμού (όπως περιγραφή της λειτουργίας μιας πρωτεΐνης, μετα-μεταφραστικές τροποποιήσεις κ.λπ.) και διασυνδέσεις με άλλες βάσεις δεδομένων:

- Swiss-Prot (http://www.expasy.ch/sprot/).
- TrEMBL (http://www.ebi.ac.uk/trembl/).
- UniProt (http://www.ebi.ac.uk/uniprot/index.html/) που προέκυψε από τη συνεργασία των Swiss-Prot, TrEMBL και PIR.

**Βάσεις δεδομένων γονιδιακής έκφρασης**

Με την εξέλιξη της τεχνολογίας και τη δημιουργία νέων οικονομικότερων τσιπ μικροσυστοιχιών, αλλά και με την εμφάνιση των τεχνολογιών αλληλούχισης νέας γενιάς (next generation seguencing), τα πειράματα ανάλυσης γονιδιακής έκφρασης πραγματοποιούνται με μεγαλύτερο ρυθμό δίνοντας μεγάλο όγκο δεδομένων και δημιουργώντας την ανάγκη αποθήκευσης και ανάλυσης αυτών των δεδομένων. Οι βάσεις δεδομένων γονιδιακής έκφρασης επιτρέπουν την καταχώρηση αποτελεσμάτων από πειράματα μικροσυστοιχιών, ενώ κάποιες από αυτές προσφέρουν και επιπλέον εργαλεία ανάλυσης.[11]

- GEO (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/). Η βάση δεδομένων GEO του NCBI παρέχει δεδομένα γονιδιακής έκφρασης, τόσο από μικροσυστοιχίες, όσο και από την αλληλούχιση νέας γενιάς (next generation seguencing), ενώ στον ίδιο ιστότοπο υπάρχουν διαθέσιμα και κάποια διαδικτυακά εργαλεία που επιτρέπουν απλές αναλύσεις των δεδομένων της βάσης. Τα δεδομένα υπάρχουν τόσο σε ακατέργαστη (raw) όσο και σε επεξεργασμένη μορή (με κανονικοποιήσεις, κ.ά.).[36]

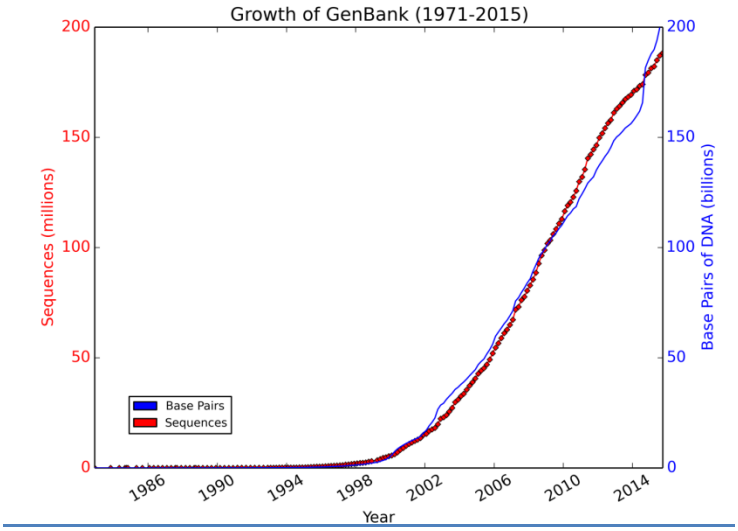
**Βιβλιογραφικές βάσεις δεδομένων**

- Η MEDLINE (US National Library of Medicine) είναι η βιβλιογραφική βάση δεδομένων της NLM (National Library of Medicine, USA) που καλύπτει τους τομείς της ιατρικής, της υγειονομικής περίθαλψης, των προκλινικών επιστημών, της βιολογίας καθώς και θεμάτων βιοϊατρικής τεχνολογίας. Περιέχει βιβλιογραφικές παραπομπές και περιλήψεις άρθρων από περισσότερα από 4800 βιοϊατρικά περιοδικά που δημοσιεύονται στις Ηνωμένες Πολιτείες και σε 70 άλλες χώρες. Η πρόσβαση στα περιεχόμενά της γίνεται από την ελεύθερη (δωρεάν) μηχανή αναζήτησης PubMed (http://www.ncbi.nlm.nih.gov/PubMed/), που είναι μέρος του συστήματος ανάκτησης πληροφοριών Entrez. Μια επιπρόσθετη βάση δεδομένων του Entrez είναι η MeSH (Medical Subject Headings) η οποία είναι ένα λεξικό όρων βιολογίας και ιατρικής. Μέσω της MeSH είναι δυνατή η άντληση ορισμών αλλά και η ανάκληση άρθρων από την PubMed.

**Δευτερογενείς βάσεις δεδομένων**

- SCOP (Structural Classification of Proteins, http://scop.mrc-lmb.cam.ac.uk/scop/). Η βάση SCOP έχει ως στόχο την ανάλυση των δομικών και εξελικτικών σχέσεων που παρατηρούνται μεταξύ όλων των πρωτεϊνών γνωστής δομής καταχωρημένων στην PDB (Protein Data Bank, http://www.rcsb.org/pdb/home/home.do). Η ταξινόμηση των πρωτεϊνών βασίζεται σ' αυτές τις δομικές και εξελικτικές σχέσεις. Τα βασικά επίπεδα ταξινόμησης είναι τέσσερα: η οικογένεια (family), η υπερ-οικογένεια (superfamily), η αναδίπλωση (fold), και η τάξη (class).

Με βάση τα παραπάνω, μπορεί να γίνει κατανοητή η παρακάτω Εικόνα, στην οποία παρουσιάζεται η αλματώδης αύξηση των βιολογικών δεδομένων από το 1971 έως το 2015.




**Εικόνα 2.9. -** Αύξηση βιολογικών δεδομένων από το 1971 έως το 2015 [41]

## 2.5. Αλγόριθμοι Κατηγοριοποίησης

Με τον όρο κατηγοριοποίηση αναφερόμαστε στη διαδικασία ομαδοποίησης αντικειμένων σε κλάσεις διαφορετικές κλάσεις, ξένες μεταξύ τους, με την προϋπόθεση ότι τα αντικείμενα τα οποία ανήκουν σε μία κλάση να παρουσιάζουν μεγαλύτερη ομοιότητα μεταξύ τους από ότι αντικείμενα άλλων κλάσεων. Σε δεδομένα γονιδιακής έκφρασης, ο διαχωρισμός των κατηγοριών βασίζεται σε μια σειρά από κριτήρια όπως είναι:

Η **κατηγοριοποίηση βάσει γονιδίων**, όπου τα γονίδια θεωρούνται ως αντικείμενα και τα δείγματα (συνθήκες) ως χαρακτηριστικά.

Η **κατηγοριοποίηση βάσει δειγμάτων**, όπου τα δείγματα θεωρούνται ως αντικείμενα , και τα χαρακτηριστικά πάνω στα οποία γίνεται η κατηγοριοποίηση είναι τα γονίδια.

Η διαφορά μεταξύ των δύο τύπων κατηγοριοποίησης είναι ότι βασίζονται σε διαφορετικά χαρακτηριστικά για να πετύχουν την κατηγοριοποίηση των δεδομένων. Παρ' όλα αυτά, αλγόριθμοι όπως ο K-means και οι ιεραρχικές προσεγγίσεις, μπορούν να εφαρμοστούν και στις δύο περιπτώσεις. Νεότερες όμως προσεγγίσεις στην μοριακή βιολογία θεωρούν ότι μόνο ένα μικρό μέρος των γονιδίων παίρνει μέρος στις κυτταρικές λειτουργίες. Έτσι έγινε απαραίτητη η εισαγωγή μίας νέας μεθόδου κατηγοριοποίησης:

Η **κατηγοριοποίηση υποδιαστημάτων (subspace clustering)**: Σ' αυτή την περίπτωση η κατηγοριοποίηση εφαρμόζεται σε ένα υποσύνολο των γονιδίων βάση ενός υποσυνόλου των δειγμάτων. Σε αυτή την προσέγγιση τα δείγματα και τα γονίδια αντιμετωπίζονται συμμετρικά, οπότε άλλοτε τα γονίδια και άλλοτε τα δείγματα θεωρούνται αντικείμενα ή χαρακτηριστικά.

Καθεμία από τις τρεις κατηγοριοποιήσεις βασίζεται σε διαφορετικό τρόπο ανάλυσης του πίνακα έκφρασης των γονιδίων και αντιμετωπίζει διαφορετικές προκλήσεις.[32]

Στην παρακάτω ενότητα θα παρουσιαστούν οι αλγόριθμοι κατηγοριοποίησης βάσει γονιδίων, δειγμάτων και υποδιαστημάτων.[42]

### Κατηγοριοποίηση Βάσει Γονιδίων

Η ιδιαιτερότητα των δεδομένων έκφρασης γονιδίων και οι ιδιόμορφες απαιτήσεις του βιολογικού τομέα, δημιουργούν στην κατηγοριοποίηση γονιδίων προκλήσεις και προβλήματα, αρκετά από τα οποία εκκρεμούν.

Η ανάλυση συστάδων είναι το πρώτο βήμα για την εξόρυξη γνώσης (data mining). Ο κύριος στόχος της κατηγοριοποίησης δεδομένων έκφρασης γονιδίων είναι η εύρεση των δομών των δεδομένων και των κατανομών των δεδομένων. Χαρακτηριστικό ενός αποτελεσματικού αλγορίθμου κατηγοριοποίησης είναι η όσο το δυνατόν μικρότερη δυνατή εξάρτηση από προηγούμενη γνώση, η οποία συνήθως δεν είναι καν διαθέσιμη πριν την ανάλυση συστάδων.

Τα δεδομένα που προκύπτουν από τα πειράματα με μικροσυστοιχίες έχουν συνήθως πολύ θόρυβο. Γι' αυτό τον λόγο ο αλγόριθμος κατηγοριοποίησης πρέπει να είναι ικανός να ξεχωρίζει τη χρήσιμη πληροφορία από τον θόρυβο.

Μεταξύ των δεδομένων γονιδιακής έκφρασης υπάρχουν ισχυρές συνδέσεις, με τις διάφορες κατηγορίες να είναι επικαλυπτόμενες ή ακόμα να εξαπλώνεται η μία πάνω στην άλλη. Οπότε οι αλγόριθμοι gene-based κατηγοριοποίησης πρέπει να μπορούν να το αντιμετωπίσουν.

Τέλος οι χρήστες μπορεί να μην ενδιαφέρονται μόνο για τις συστάδες γονιδίων, αλλά ακόμα και για τις σχέσεις μεταξύ των συστάδων, σχέσεις γονιδίων μέσα στην ίδια συστάδα κτλ. Γι' αυτό το λόγο, ένας αλγόριθμος κατηγοριοποίησης θα ήταν καλό να μην σταματά στον διαμερισμό των δεδομένων, αλλά να παρέχει και γραφική αναπαράσταση της δομής των συστάδων.[43]

Οι βασικότεροι αλγόριθμοι κατηγοριοποίησης γονιδίων είναι οι παρακάτω:[32]

- K-means.
- Αυτό-οργανωμένοι χάρτες S.O.M. (Self-organized maps).
- Ιεραρχική κατηγοριοποίηση.
- Γραφοθεωρητικές μέθοδοι κατηγοριοποίησης.
- Κατηγοριοποίηση βάσει μοντέλου.
- Κατηγοριοποίηση βάσει πυκνότητας (ιεραρχική προσέγγιση DHC).

### Κατηγοριοποίηση Βάσει Δειγμάτων

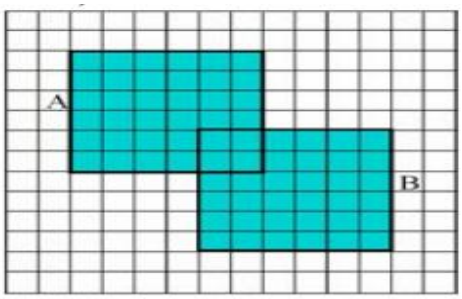
Ο πίνακας έκφρασης γονιδίων τις πιο πολλές φορές αποτελείται από μακροσκοπικούς φαινοτύπους δειγμάτων οι οποίοι συνδέονται με ασθένειες ή επίδραση φαρμάκων, όπως δείγματα ασθενειών, φυσιολογικά δείγματα και δείγματα θεραπείας. Ο κύριος στόχος της κατηγοριοποίησης βάσει δειγμάτων είναι η ανακάλυψη φαινοτυπικών δομών ή υπό-δομών των δειγμάτων. Προηγούμενες μελέτες [44] έχουν δείξει ότι ο φαινότυπος των δειγμάτων μπορεί να διακριθεί μόνο μέσα από μία μικρή υπομονάδα γονιδίων των οποίων τα

επίπεδα έκφρασης σχετίζονται στενά με την κλάση διάκρισης. Τα γονίδια αυτά αποκαλούνται και πληροφοριακά γονίδια. Τα υπόλοιπα γονίδια που απομένουν στον πίνακα έκφρασης γονιδίων θεωρούνται απλά θόρυβος. Οι κλασικές μέθοδοι κατηγοριοποίησης, όπως αυτές που αναφέρθηκαν πριν (π.χ. αυτοοργανωμένοι χάρτες, K-means κτλ.), μπορούν να εφαρμοστούν θεωρώντας όλα τα γονίδια σαν χαρακτηριστικά, όμως ο υψηλός λόγος σήματος-θορύβου (που είναι της τάξης του 1/10) μειώνει δραματικά την αξιοπιστία των αποτελεσμάτων. Για τον λόγο αυτό έχουν προταθεί καινούριοι αλγόριθμοι, οι οποίοι χωρίζονται σε δύο κύριες κατηγορίες:[44]

- Την εποπτευόμενη ανάλυση και
- Την μη-εποπτευόμενη ανάλυση.

### Κατηγοριοποίηση Υποδιαστημάτων (Subspace Clustering)

Οι παραπάνω τύποι αλγορίθμων κατηγοριοποίησης θεωρούνται όλοι δείγματα "ολικής κατηγοριοποίησης" δεδομένου ότι για ένα συγκεκριμένο σύνολο δεδομένων προς κατηγοριοποίηση, ο χώρος των χαρακτηριστικών είναι καθορισμένος και μοιράζεται από όλες τις εξαγόμενες συστάδες, με τις συστάδες να είναι αποκλειστικές και εξαντλητικές. Παρ' όλα αυτά, στην μοντέρνα μοριακή βιολογία είναι γνωστό ότι μόνο ένα μικρό μέρος των γονιδίων παίρνει μέρος στις κυτταρικές λειτουργίες. Επίσης, είναι σύνηθες ένα γονίδιο να έχει ενεργό ρόλο σε πολλαπλές λειτουργίες οι οποίες δεν είναι απαραίτητο να συνεργάζονται, οπότε ένα γονίδιο μπορεί να ανήκει σε πολλές συστάδες, ή και σε καμία. Μία σειρά μεθόδων κατηγοριοποίησης υποδιαστημάτων προτάθηκε πρόσφατα, ώστε να κάνει κατανοητή την σχέση που παρουσιάζεται στα μπλοκ -ένας υποπίνακας που ορίζεται από ένα υποσύνολο γονιδίων και ένα υποσύνολο δειγμάτων- μέσα στον πίνακα έκφρασης γονιδίων. Η κατηγοριοποίηση υποδιαστημάτων αρχικά προτάθηκε από τον Agrawal (1998) [43] για να βρει υποσύνολα αντικειμένων ώστε τα αντικείμενα να εμφανίζονται σαν μία συστάδα σε έναν υποχώρο που ορίζεται από ένα υποσύνολο των χαρακτηριστικών.




**Εικόνα 2.10. –** Υποδιάστημα συστάδων A & B [45]

Στην παραπάνω Εικόνα 2.10 παρουσιάζεται ένα παράδειγμα του υποσυνόλου των συστάδων (A, B) απλωμένου στον πίνακα έκφρασης γονιδίων. Στην κατηγοριοποίηση αυτή είναι δυνατόν το υποσύνολο των χαρακτηριστικών να διαφέρει από συστάδα σε συστάδα. Δύο υποσύνολα συστάδων μπορούν να έχουν ως γνώρισμα τον διαμοιρασμό κοινών αντικειμένων και χαρακτηριστικών, ενώ από την άλλη κάποια αντικείμενα μπορεί να μην ανήκουν σε κανένα υποσύνολο συστάδας. Για ένα πίνακα έκφρασης γονιδίων που έχει n γονίδια και m δείγματα, η πολυπλοκότητα για πλήρη συνδυασμό τους είναι $2^{n+m}$ οπότε το πρόβλημα της συνολικά βέλτιστης επιλογής μπλοκ είναι δυσκολίας NP. Οι μέθοδοι κατηγοριοποίησης υποδιαστημάτων

συνήθως ορίζουν μοντέλα για να περιγράψουν το επιθυμητό μπλοκ και στην συνέχεια υιοθετούν κάποια μέθοδο να ψάξει στον χώρο των δειγμάτων-γονιδίων.[33]

## 2.6. Τεχνικές Κατηγοριοποίησης

Σε έναν πίνακα γονιδιακής έκφρασης, αν δύο γονίδια σχετίζονται (έχουν παρόμοιες λειτουργίες), τα γονιδιακά τους προφίλ παρουσιάζουν ομοιότητα. Η ομοιότητα υπολογίζεται με τη χρήση της ευκλείδειας απόστασης ή της υψηλής συσχέτισης. Η κλασσική ομαδοποίηση ή διαφορετικά clustering ομαδοποιεί τα στοιχεία ενός πίνακα, ώστε τα στοιχεία της ίδιας ομάδας να είναι παραπλήσια ενώ από την άλλη, οι ομάδες να είναι σαφώς χωρισμένες αναμεταξύ τους. Στη μέθοδο ομαδοποίησης, οι δύο κύριες τεχνικές βασίζονται στα ακόλουθα δυο κριτήρια: **στην κατάσταση**, όπου εδώ παίζει ρόλο η ομοιότητα της γονιδιακής έκφρασης, και **στο γονίδιο, όπου** η ομαδοποίηση είναι αλληλεξάρτηση της γονιδιακής έκφρασης.

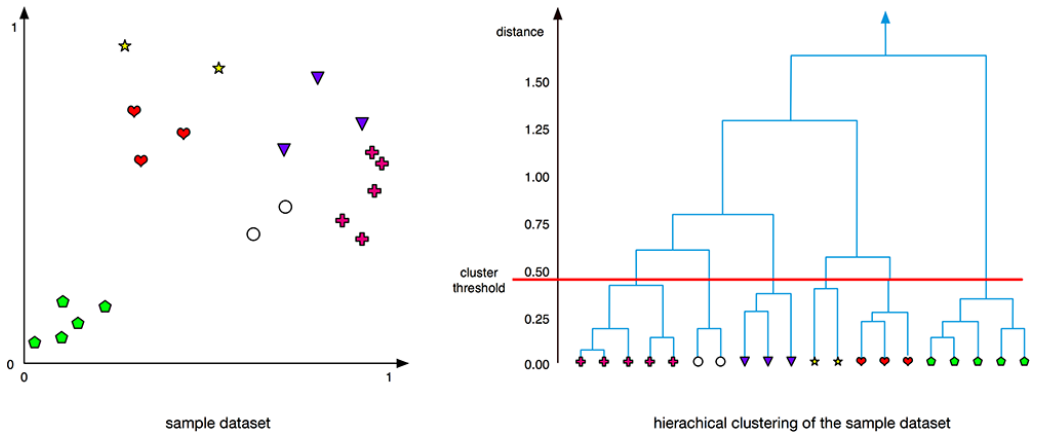
Μια από τους πιο δημοφιλείς τύπους αλγορίθμων clustering είναι:

Η **ιεραρχική ομαδοποίηση** (Hierarchical clustering) - Η ιεραρχική ομαδοποίηση εφαρμόζεται σχετικά εύκολα και περιλαμβάνει δύο είδη clustering, την ενωτική (συσσωρευτική) agglomerative και τη διαιρετική (διαχωριστική) (divisive). Στην αρχή της εφαρμογής ο αριθμός των ομάδων δεν είναι γνωστός. 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, παρ’ όλα αυτά ένα γονίδιο μπορεί να συμμετέχει σε πάνω από μία κυτταρικές διεργασίες.[46], [47]

### Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)

Αναλυτικότερα, η βασική ιδέα της τεχνικής της ιεραρχικής ομαδοποίησης (Hierarchical clustering), συνδέεται με την δημιουργία μιας ιεραρχίας από clusters η οποία παρουσιάζει τις σχέσεις (οι οποίες βασίζονται στην ομοιότητα) μεταξύ ανεξάρτητων μελών και συγχωνευμένων clusters των δεδομένων.[48]




**Εικόνα 2.11.** - Ιεραρχική ομαδοποίηση (Hierarchical clustering) [48]

Στην **ιεραρχική ομαδοποίηση** (Hierarchical clustering) πληρούνται οι ακόλουθοι κανόνες:

- Το συνολικό σύνολο δεδομένων (dataset) αντιπροσωπεύεται από μία ρίζα.
- Ένα συγκεκριμένο στοιχείο του συνόλου δεδομένων (dataset) αναπαρίσταται από ένα φύλλο.
- Ο κόμβος αντιπροσωπεύει την ένωση των στοιχείων στο υποδέντρο.
- Το ύψος ενός κόμβου αναπαριστά την απόσταση δύο παιδιών κόμβων.[49]

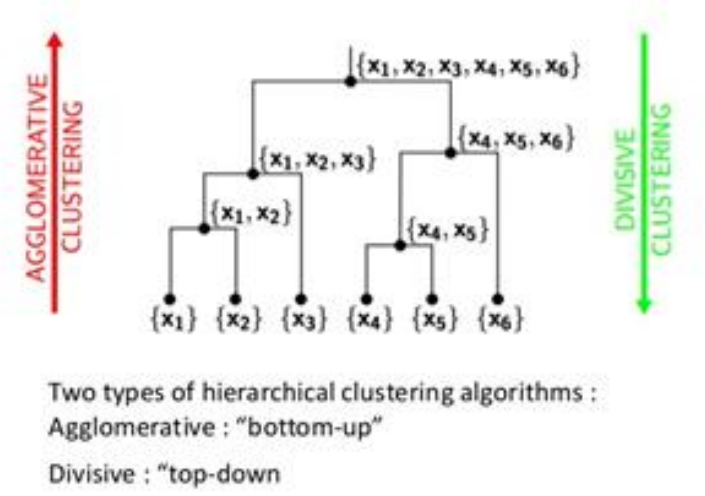
**Όπως αναφέρθηκε, οι ιεραρχικές μέθοδοι χωρίζονται σε ενωτικές και διαχωριστικές:**

**Ενωτικές ή από κάτω προς τα πάνω (Agglomerative ή bottom-up):**

- Ξεκινούν τοποθετώντας κάθε στοιχείο του πίνακα σε ένα ξεχωριστό cluster.
- Σε κάθε βήμα συγχωνεύουν τα δύο πιο όμοια clusters.
- Σταματούν όταν όλα τα στοιχεία ανήκουν σε ένα μόνο cluster ή όταν
- Τερματιστεί κάποιο συγκεκριμένο κριτήριο.

**Διαχωριστικές ή από πάνω προς τα κάτω (Divisive ή top-down):**

- Ξεκινούν τοποθετώντας όλα τα στοιχεία του πίνακα σε ένα cluster.
- Σε κάθε επανάληψη χωρίζουν ένα cluster σε δύο καινούργια.
- Ολοκληρώνονται όταν όλα τα στοιχεία ανήκουν στο δικό τους cluster ή όταν
- Τερματιστεί κάποιο συγκεκριμένο κριτήριο.[50],[51]




**Εικόνα 2.12. –** Ενωτική (Agglomerative) και Διαχωριστική (Divisive) Ομαδοποίηση [52]

### Συσταδοποίηση δυο-τρόπων (Biclustering)

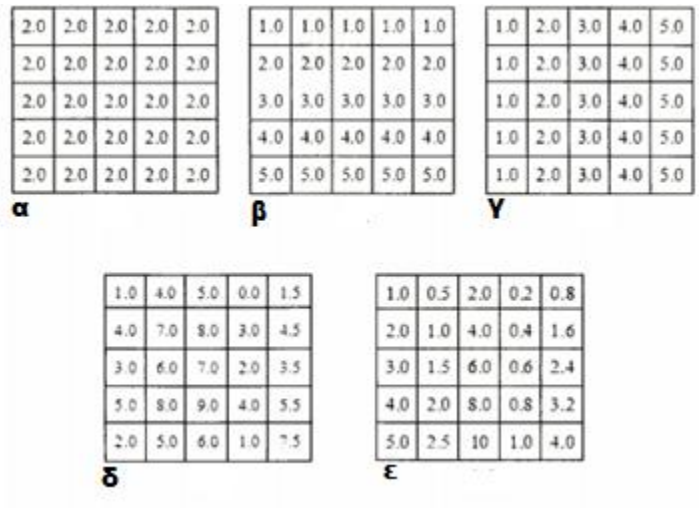
Η συσταδοποίηση δυο-τρόπων ή αλλιώς **διπλή κατηγοριοποίηση** αποτελεί μια μέθοδο ομαδοποίησης που επιτρέπει την ταυτόχρονη κατηγοριοποίηση των γραμμών και των στηλών ενός πίνακα. Ο όρος εισήχθη για πρώτη φορά από τον Mirkin (1996),[7] αν και η συγκεκριμένη τεχνική κατηγοριοποίησης παρουσιάστηκε πολύ πιο νωρίς από τον Hartigan (1975).[6] Οι περισσότεροι από αυτούς τους αλγορίθμους χρησιμοποιούνται σε εφαρμογές της βιοπληροφορικής.

#### Δημιουργία συστάδων δυο-τρόπων

Δοθέντος ενός πίνακα μεγέθους X,Y, ο αλγόριθμος διπλής κατηγοριοποίησης δημιουργεί συστάδες με τον εξής τρόπο: ένα υποσύνολο γραμμών του πίνακα που παρουσιάζει παρόμοια συμπεριφορά κατά πλάτος ενός υποσυνόλου στηλών και το αντίθετο θα τοποθετηθεί σε μια νέα συστάδα.

#### Είδη αλγορίθμων διπλής κατηγοριοποίησης

Υπάρχουν διάφορα είδη αλγορίθμων διπλής κατηγοριοποίησης τα οποία παρουσιάζονται οπτικά και μέσω των σχημάτων που ακολουθούν. Συγκεκριμένα τα είδη αυτών των αλγορίθμων είναι τα εξής: 1. διπλή κατηγοριοποίηση με σταθερές τιμές, 2. διπλή κατηγοριοποίηση με σταθερές τιμές σε γραμμές και στήλες, και 3. διπλή κατηγοριοποίηση με συναφείς τιμές.[53]




**Εικόνα 2.13.** - Είδη αλγορίθμων διπλής κατηγοριοποίησης : α) διπλή κατηγοριοποίηση σταθερής τιμής β) διπλή κατηγοριοποίηση σταθερής σειράς γ) διπλή κατηγοριοποίηση σταθερής στήλης δ) διπλή κατηγοριοποίηση συνεκτικής τιμής (προσθετική) ε) διπλή κατηγοριοποίηση συνεκτικής τιμής (μοντέλο) [54]

**Αλγόριθμοι διπλής κατηγοριοποίησης**

Το εύρος των αλγορίθμων αυτής της κατηγορίας που χρησιμοποιούνται σε εφαρμογές της βιοπληροφορικής είναι αρκετά μεγάλο. Μερικά παραδείγματα τέτοιων αλγορίθμων είναι τα ακόλουθα: ο αλγόριθμος συσταδοποίησης μπλοκ, ο CTWC, ο ITWC, ο *δ*-Bicluster, ο d-pCluster, ο d-pattem, ο FLOC, ο OPC, ο Plaid Model, ο OPSMs, ο Gibbs, ο SAMBA, ο RoBA , ο Crossing Minimization, ο cMonkey , ο PRMs και ο DCC. Ένα μεγάλο πρόβλημα που προκύπτει με όλους τους αλγορίθμους που ανήκουν σε αυτή την κατηγορία είναι ο βαθμός αξιοπιστίας των αποτελεσμάτων που δίνουν καθώς η συσταδοποίηση 2-τρόπων επιτρέπει την επικάλυψη μεταξύ διαφορετικών συστάδων.

Οι περισσότεροι από τους αναφερόμενους αλγορίθμους βασίζονται σε μια από τις πέντε (5) παρακάτω μεθόδους για την εύρεση των Biclusters:

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

Το γεγονός ότι κάποιοι από αυτούς τους αλγορίθμους δεν είναι ντετερμινιστικοί και συνάμα ότι το πρόβλημα εμπεριέχει μη επιβλεπόμενη κατηγοριοποίηση, δυσκολεύουν τον εντοπισμό σφαλμάτων στα αποτελέσματα. Μια προσέγγιση για την επίλυση του συγκεκριμένου προβλήματος είναι χρήση πολλών αλγορίθμων διπλής κατηγοριοποίησης από τους οποίους, η πλειονότητα θα ψηφίζει για να αποφασιστεί ποιο είναι το καλύτερο αποτέλεσμα κατηγοριοποίησης. Παρ' όλα αυτά οι συγκεκριμένοι αλγόριθμοι βρίσκουν αρκετά καλή εφαρμογή στο χώρο της βιολογίας και γενικότερα της βιοπληροφορικής.[3],[6],[10]

Δεδομένου ότι η παρούσα εργασία επικεντρώνεται στη μέθοδο της διπλής κατηγοριοποίησης, οι αλγόριθμοι και τα είδη της αυτά θα περιγραφούν αναλυτικότερα στο κεφάλαιο που ακολουθεί (Κεφάλαιο 3).

# 3

# Αλγόριθμοι και Τεχνικές Διπλής Κατηγοριοποίησης (Biclustering)

## 3.1. Τύποι Διπλής Κατηγοριοποίησης

Ένα από τα βασικά κριτήρια για την ορθή επιλογή του κατάλληλου αλγορίθμου διπλής κατηγοριοποίησης είναι το είδος των Biclusters που αναμένεται ως αποτέλεσμα. Έτσι, ανάλογα με τον τύπο του αλγόριθμου, οι τέσσερις κύριοι τύποι είναι:

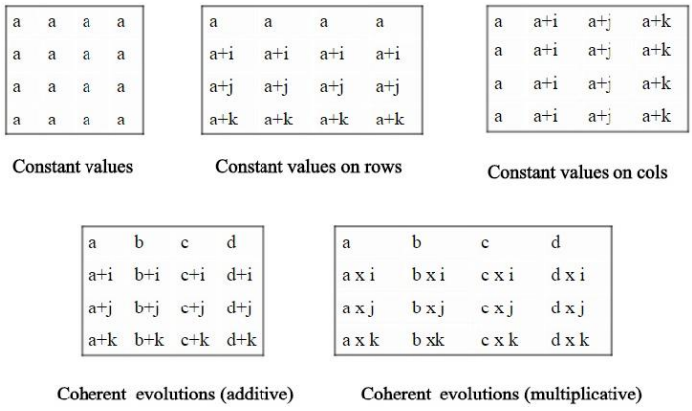
**α)** Biclusters με σταθερές τιμές,

**β)** Biclusters με σταθερές τιμές σε γραμμές ή στήλες,

**γ)** Biclusters με συναφείς τιμές, και

**δ)** Biclusters με συναφείς εξελίξεις.

Στη διπλανή εικόνα βλέπουμε την α, τη β και τη δ περίπτωση.




**Εικόνα 3.1.** - Τύποι Biclusters

Στους τρεις πρώτους τύπους σημαντικό ρόλο έχουν οι αριθμητικές τιμές του πίνακα και στοχεύουν στον εντοπισμό των υποομάδων γραμμών ή στηλών με παρόμοια συμπεριφορά όπως φαίνεται στις παρακάτω εικόνες [Εικόνες 3.1, 3.2, 3.3]. Στον τέταρτο τύπο η εύρεση ομάδων επιτυγχάνεται όχι βάσει των αριθμητικών τιμών του πίνακα αλλά βάσει των συμβόλων. Τα σύμβολα αυτά μπορεί να είναι κάποιοι συγκεκριμένοι χαρακτήρες, να ανταποκρίνονται σε μια συγκεκριμένη σειρά, είτε να αναπαριστούν αλλαγές (θετικές και αρνητικές) που συνδέονται με μία τιμή.[10]

Στη συνέχεια παρουσιάζεται μια σύντομη σημειογραφία χρήσιμη για την αναλυτικότερη περιγραφή των 4 αυτών τύπων. Δεδομένου ενός πίνακα A=(*X*,*Y*) όπου *X* το σύνολο των γραμμών και *Y* το σύνολο των στηλών του, ένα Bicluster είναι ένας υποπίνακας (*I*,*J*) όπου *I* ένα υποσύνολο του *X* και *J* ένα υποσύνολο του *Y* αντίστοιχα και $\alpha_{ij}$ 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*):

$$\alpha_{iJ} = 1/|J| \sum_{j \in J} \alpha_{ij}$$

$$\alpha_{Ij} = 1/|I| \sum_{i \in I} \alpha_{ij} \text{ και}$$

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

**Biclusters με σταθερές τιμές**

Οι αλγόριθμοι από τους οποίους προκύπτουν Biclusters με σταθερές τιμές τις πιο πολλές φορές αναδιατάσσουν τις γραμμές και τις στήλες ενός πίνακα με σκοπό να ομαδοποιήσουν γραμμές και στήλες με παραπλήσιες 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, όμως τις περισσότερες φορές τα δεδομένα προς επεξεργασία εμπεριέχουν θόρυβο, πράγμα που σημαίνει ότι ένα σταθερό Bicluster παρουσιάζεται με τιμές $n_{ij}+\mu$, όπου το $n_{ij}$ είναι ο σχετικός θόρυβος της τιμής μ του $aij$.[10]

| | | | |
|---|---|---|---|
| 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 3.2. -** Bicluster με σταθερές τιμές

**Biclusters με σταθερές τιμές σε γραμμές ή στήλες**

Οι τιμές ενός βέλτιστου Bicluster με σταθερές γραμμές σε έναν υποπίνακα $(I,J)$ είναι της μορφής $\alpha ij=\mu + \alpha_i$ ή $\alpha_{ij}=\mu \times \alpha_i$ όπου μ είναι μια σταθερή τιμή εντός του Bicluster και $\alpha_i$ αποτελεί μία προσαρμοσμένη τιμή για κάθε γραμμή $i \in I$, η οποία προκύπτει είτε από πρόσθεση ή πολλαπλασιασμό. Ανάλογα, οι τιμές ενός βέλτιστου Bicluster με σταθερές στήλες σε έναν υποπίνακα (I,J) είναι της μορφής $\alpha_{ij}=\mu + \beta_j$ ή $\alpha_{ij}=\mu \times \beta_j$, όπου μ είναι μια σταθερή τιμή εντός του Bicluster και $\beta_j$ αποτελεί μία προσαρμοσμένη τιμή για κάθε στήλη $j \in J$, η οποία προκύπτει είτε από πρόσθεση ή πολλαπλασιασμό.[10]

| | | | |
|---|---|---|---|
| 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 3.3. -** Bicluster με σταθερές τιμές σε γραμμές και στήλες

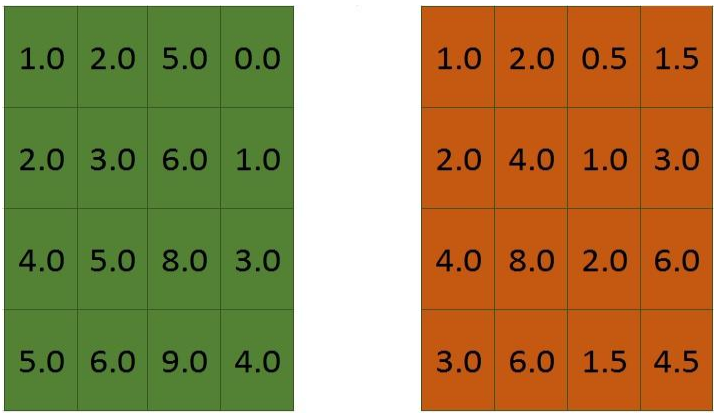
Στο αριστερό σκέλος της Εικόνας 3.3 το Bicluster εμφανίζει σταθερές τιμές στις γραμμές του ακολουθώντας το μοντέλο $\alpha_{ij}=\mu + \alpha_i$, όπου $\mu$=0.0 και κάθε φορά προστίθεται σε κάθε νέα γραμμή η τιμή

$\alpha i$=1.0. Στο δεύτερο σκέλος της Εικόνας 3.3 το 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 $\beta j$=1.0.

## Biclusters με συναφείς (coherent) τιμές

Η κατηγορία των Biclusters με συναφείς τιμές δεν ορίζεται με κάποιο απλό μοντέλο πρόσθεσης ή πολλαπλασιασμού που εφαρμόζεται στα σταθερά Biclusters. Υπάρχουν πιο σύνθετες τεχνικές που σχετίζονται με την ανάλυση της διακύμανσης μεταξύ των υποομάδων του πίνακα. Αυτές οι μέθοδοι χρησιμοποιούν μια συγκεκριμένη μορφή συνδιακύμανσης μεταξύ γραμμών και στηλών σε ένα Bicluster για την αξιολόγηση της αποτελεσματικότητας της ομαδοποίησης.

Χρησιμοποιώντας λοιπόν ένα πιο σύνθετο μοντέλο πρόσθεσης, ένα τέλειο συνεκτικό Bicluster χαρακτηρίζεται από μία υποομάδα γραμμών και στηλών που οι τιμές της $\alpha_{ij}$ επιλέγονται με βάση την ακόλουθη έκφραση: $\alpha_{ij}=\mu + \alpha_i + \beta_j$, όπου μ είναι μια σταθερή τιμή εντός του Bicluster και $\alpha_i$ μία προσαρμοσμένη τιμή για κάθε γραμμή $i \in I$, και $\beta j$ μία προσαρμοσμένη τιμή για κάθε στήλη $j \in J$,οι οποίες προστίθενται στην $\mu$. Ένα άλλο μοντέλο αναφέρει ότι τα Biclusters με συναφείς τιμές μπορούν να περιγραφούν από ένα μοντέλο πολλαπλασιασμού, όπου $\alpha ij=\mu' \times \alpha i' \times \beta j'$ .[10]




**Εικόνα 3.4.** - Biclusters με συναφείς τιμές(αριστερά μοντέλο πρόσθεσης, δεξιά μοντέλο πολλαπλασιασμού)

Για να γίνει πιο κατανοητή η μορφή των Biclusters με συναφείς τιμές παρουσιάζεται το παράδειγμα του Πίνακα 3.1 για το μοντέλο της πρόσθεσης: Γνωρίζουμε ότι $\alpha_{ij}=\mu + \alpha_i + \beta_j$, άρα για $\mu=0$

**Πίνακας 3.1.** - Μοντέλο Πρόσθεσης

| | $\beta 1=0$ | $\beta 2=1$ | $\beta 3=4$ | $\beta 4=-1$ |
|---|---|---|---|---|
| $\alpha 1=1$ | 1 | 2 | 5 | 0 |
| $\alpha 2=2$ | 2 | 3 | 6 | 1 |
| α3=4 | 4 | 5 | 8 | 3 |
| $\alpha 4=5$ | 5 | 6 | 9 | 4 |

Και για το μοντέλο του πολλαπλασιασμού βάσει ότι $aij= \mu' \times \alpha i' \times \beta j'$, άρα για $\mu=1$

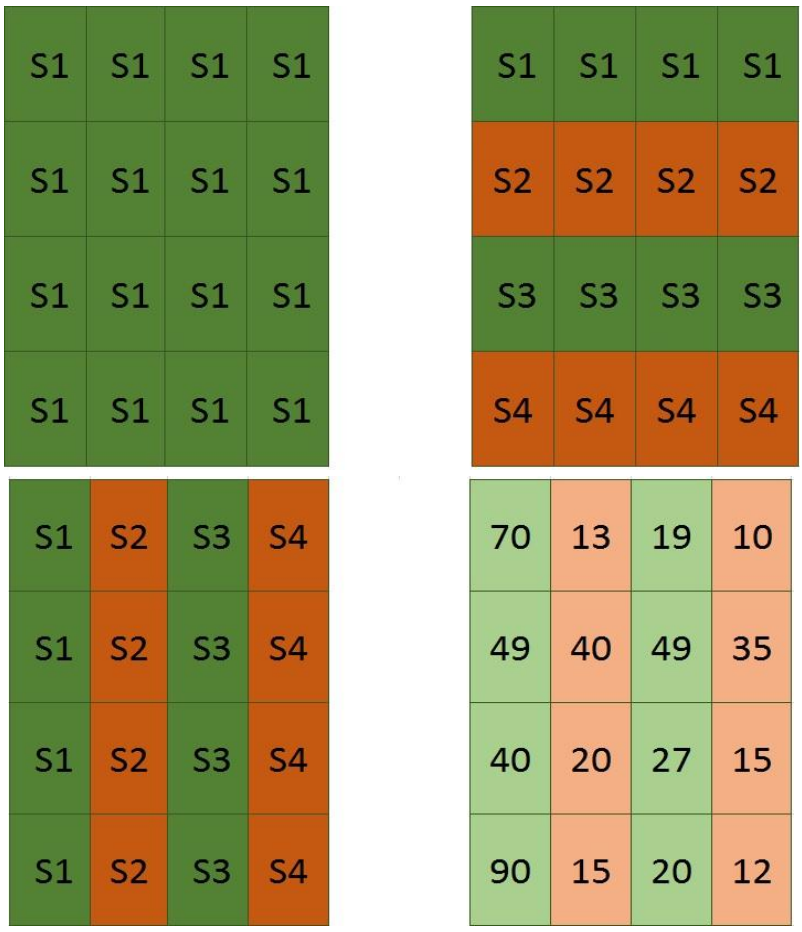
**Πίνακας 3.2.** - Μοντέλο Πολλαπλασιασμού

| | *β1=1* | *β2=2* | *β3=0,5* | *β4=3/2* |
|---|---|---|---|---|
| *α1=1* | 1 | 2 | 0,5 | 1,5 |
| *α2=2* | 2 | 4 | 3 | 3 |
| *α3=4* | 4 | 8 | 2 | 6 |
| *α4=3* | 3 | 6 | 1,5 | 4,5 |

## Biclusters με συναφείς εξελίξεις (coherent evolutions)

Αυτός ο τύπος Bicluster έχει ως στόχο την ομαδοποίηση γραμμών ή στηλών ανεξάρτητα από τις ακριβείς τιμές των στοιχείων του πίνακα.[10] Τα Biclusters με συναφείς εξελίξεις, σε αντίθεση με τα Biclusters με συναφείς τιμές, προσδιορίζουν υποσύνολα των γραμμών (γονιδίων) που οι τιμές τους είτε αυξάνονται είτε μειώνονται σταδιακά κατά μήκος μιας ομάδας στηλών ή γραμμών.

Σε αντίθεση με τους παραπάνω τύπους, αυτός ο τύπος είναι πολύπλοκο να μοντελοποιηθεί με τη χρήση μαθηματικής εξίσωσης.

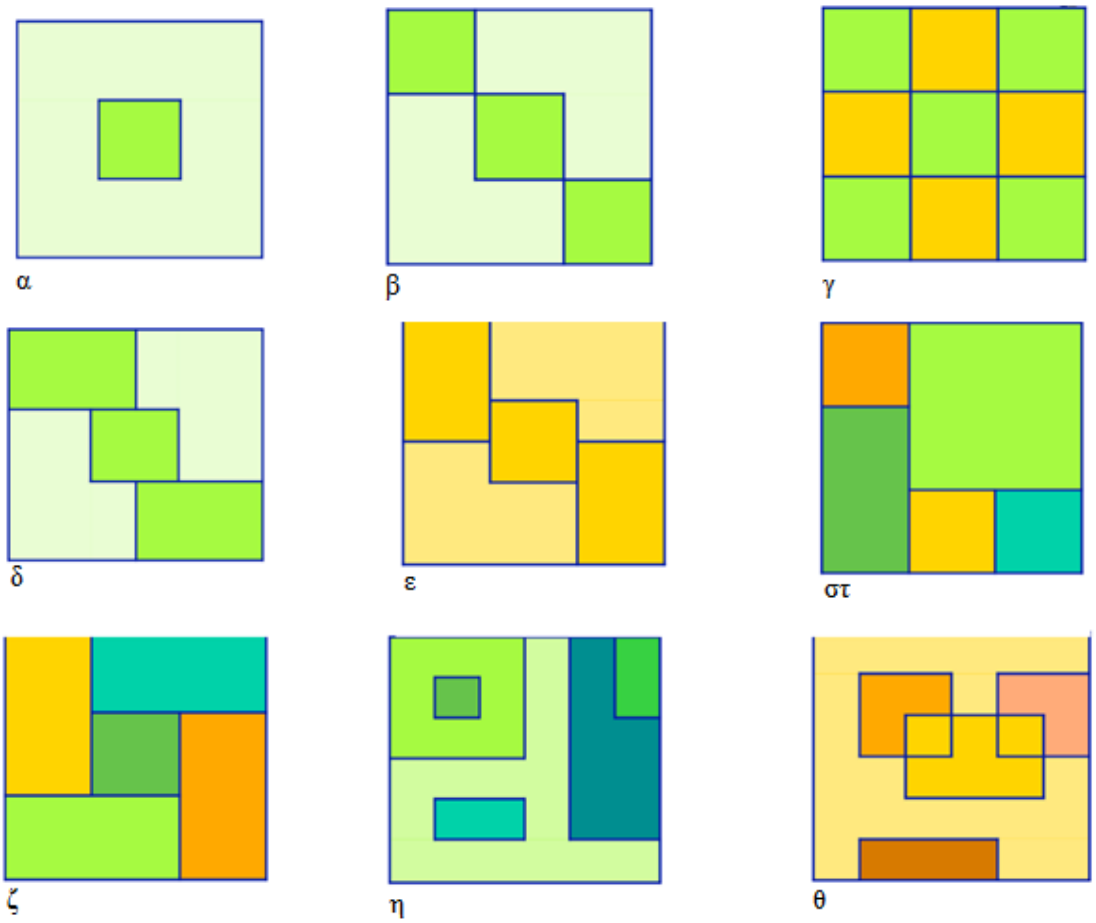



**Εικόνα 3.5.** - Biclusters με συναφείς εξελίξεις

## 3.2. Δομή Διπλής Κατηγοριοποίησης

Μετά την ολοκλήρωση εκτέλεσης των αλγορίθμων Bicluster, οι δομές οι οποίες προκύπτουν ταξινομούνται σε 2 κατηγορίες: ένα μοναδικό Bicluster ή ένας μεγαλύτερος αριθμός από Biclusters με διάφορες μορφές, οι οποίες είναι οι παρακάτω:

- Μονό.
- 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.




**Εικόνα 3.6. -** Δομές Bicluster: α. Μονό, β. Αποκλειστικής γραμμής και στήλης, γ. Μη επικαλυπτόμενο με τη μορφή σκακιέρας, δ. Αποκλειστικά με σειρές, ε. Αποκλειστικά με στήλες, στ. Μη-επικαλυπτόμενα με δενδροειδή μορφή, ζ. Μη-επικαλυπτόμενα, μη αποκλειστικά, η. Επικαλυπτόμενα με ιεραρχική δομή, θ. Αυθαίρετα τοποθετημένα-επικαλυπτόμενα [55]

Μία βοηθητική απεικόνιση για την αναζήτηση πολλών Biclusters σε έναν πίνακα είναι η χρήση του ίδιου χρώματος για κάθε στοιχείο με την ίδια τιμή $a_{ij}$. Με αυτόν τον τρόπο είναι πιο απλή η εναλλαγή γραμμών και στηλών και πιο εύκολη η ομαδοποίηση γραμμών και στηλών που σχηματίζουν «μπλοκς» με παρόμοια χρώματα. Δηλαδή τα Biclusters είναι τα «μπλοκς», δηλ. οι υποπίνακες του αρχικού πίνακα με παραπλήσιες τιμές έκφρασης. Στην ιδεατή περίπτωση η εικόνα που δημιουργείται αποτελείται από *K* τετράγωνα «μπλοκς» στη διαγώνιο του πίνακα (3.6.β). Αυτή η ιδέα στηρίζεται στο ότι κάθε γραμμή και στήλη του πίνακα μπορεί να ανήκει αποκλειστικά σε ένα από τα *K* Biclusters. Τις πιο πολλές φορές όμως πολλές γραμμές και στήλες του πίνακα μπορεί να ανήκουν σε περισσότερα από ένα Biclusters και είτε υπάρχει επικάλυψη μεταξύ των στοιχείων (3.6.η,θ), είτε όχι (3.6.γ,δ,ε,στ,ζ).[55]

## 3.3. Ποιοτικά Μέτρα Διπλής Κατηγοριοποίησης

Αρκετά ποιοτικά μέτρα για τα Biclusters έχουν προταθεί παρ 'όλα αυτά, κανένα από τα προτεινόμενα ποιοτικά μέτρα δεν είναι σε θέση να αναγνωρίσει ένα τέλειο πρότυπο μετατόπισης και κλιμάκωσης σε ένα Bicluster.

Στην ενότητα αυτή θα **παρουσιαστούν** μερικά από τα πιο γνωστά μέτρα αξιολόγησης για Biclusters, όπως περιγράφονται στη συγκριτική μελέτη των Pontes και συνεργατών (2015).[56]

**Διακύμανση (VAR)**

Ο Hartigan χρησιμοποίησε τη διακύμανση Bicluster, όπως δίνεται στην εξίσωση 3.3.1 ως συνεχές μέτρο, όπου ο στόχος του αλγόριθμού του ήταν να ελαχιστοποιήσει το άθροισμα των διακυμάνσεων του Bicluster. Ο ορισμός της Διακύμανσης είναι ο εξής:

$$\text{VAR (A)} = \frac{1}{|I|*|J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{IJ})^2 \qquad (3.3.1)$$

Μπορεί να σημειωθεί ότι η διακύμανση ανιχνεύει μόνο σταθερά Biclusters. Ως εκ τούτου, οι προσεγγίσεις biclustering με βάση την διακύμανση των μεθόδων ελαχιστοποίησης περιλαμβάνουν επίσης και άλλα κριτήρια ομοιογένειας για να ανιχνεύουν άλλους τύπους Biclusters.[56]

**Μέσο τετράγωνο υπολείμματος (Mean Square Residue, MSR)**

Ο Cheng και Church ήταν οι πρώτοι που εφάρμοσαν biclustering σε δεδομένα γονιδιακής έκφρασης. Εισήγαγαν έναν από τους πιο δημοφιλείς αλγόριθμους biclustering που συνδυάζει μια άπληστη αναζήτηση για την εύρεση Biclusters με ένα μέτρο για την αξιολόγηση της ποιότητας των εν λόγω Biclusters. Προκειμένου να αξιολογηθεί η ποιότητα των Biclusters, όπως ήδη έχει παρουσιαστεί, ο αλγόριθμος χρησιμοποιεί το μέσο τετράγωνο υπολείμματος (MSR) , το οποίο αποτελεί τη διακύμανση της ομάδας όλων των στοιχείων του εκάστοτε Bicluster. Το μέτρο αυτό αποσκοπεί στο να αξιολογήσει τη συνοχή των γονιδίων

και τις συνθήκες ενός Bicluster A που αποτελείται από σειρές I,και J στήλες. Ο ορισμός του MSR είναι ο εξής:

$$\text{MSR (A)} = \frac{1}{|I|*|J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \qquad (3.3.2)$$

Στόχος είναι η εύρεση μεγάλων Biclusters με χαμηλό αποτέλεσμα MSR και συγκεκριμένα μικρότερο από ένα συγκεκριμένο **κατώφλι** ( threshold) $\delta$ - bicluster. Εάν ένα Bicluster έχει MSR ίσο με το μηδέν, αυτό σημαίνει ότι τα γονίδια του κυμαίνονται κατά τον ίδιο ακριβώς τρόπο κάτω από το υποσύνολο των πειραματικών συνθηκών, και έτσι μπορεί να θεωρηθεί ένα τέλειο Bicluster.[56]

**Κλιμάκωση μέσου τετραγώνου υπολείμματος (Scaling Mean Squared Residue, SMSR)**

Οι Mukhopadhyay, Maulik και Bandyopadhyay ανέπτυξαν ένα μέτρο αξιολόγησης για Biclusters που είναι σε θέση να αναγνωρίζει την κλιμάκωση προτύπων. Στο έργο τους, αναλύουν τους λόγους για τους οποίους το MSR είναι σε θέση να αναγνωρίσει τη μετατόπιση προτύπων (μοτίβων) σε Biclusters αλλά **όχι** την κλιμάκωση προτύπων. Χρησιμοποιώντας τον μαθηματικό τύπο για την κλιμάκωση προτύπων, ορίζουν ένα μέτρο που στη συνέχεια αποδείχθηκε ότι μπορεί να προσδιορίσει την κλιμάκωση προτύπων. Αυτό το νέο μέτρο καλείται SMSR, από το Scalling MSR. Παρ'όλα αυτά, το SMSR **δεν** είναι ικανό να ανιχνεύσει τη μετατόπιση προτύπων.[56] **Ο ορισμός του SMSR είναι ο εξής:**

$$\text{SMSR(A)} = \frac{1}{|I|*|J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}\frac{(a_{iJ}*a_{Ij} - a_{ij}*a_{IJ})^2}{a_{iJ}{}^2 * a_{Ij}{}^2} \qquad (3.3.3)$$

**Δείκτης Σχετικότητας ή Συνάφειας (Relevance Index, RI)**

Οι Yip,Ng και Cheung πρότειναν ένα μέτρο αξιολόγησης που διαφέρει ελαφρώς από το MSR, στο οποίο η ποιότητα ενός Bicluster υπολογίζεται ως το άθροισμα των δεικτών σχετικότητας των στηλών. **Ο δείκτης σχετικότητας $R_{Ij}$ για στήλη $j \in J$ ορίζεται ως** :

$$R_{Ij} = 1 - \frac{\sigma_{Ij}{}^2}{\sigma_j{}^2} \qquad (3.3.4)$$

Όπου το $\sigma_{Ij}{}^2$ (τοπική διακύμανση) είναι η διακύμανση των τιμών στη στήλη $j$ για το Bicluster και $\sigma_{.j}{}^2$(καθολική διακύμανση) είναι το σύνολο των δεδομένων. Εφόσον ο δείκτης δίνει μια υψηλή τιμή όταν η τοπική διακύμανση είναι μικρή σε σχέση με την ολική διακύμανση, ο σχετικός δείκτης, για μια στήλη, μεγιστοποιείται αν η τοπική διακύμανση του είναι μηδέν, υπό την προϋπόθεση ότι η συνολική διακύμανση δεν είναι. Βασιζόμενοι στο δείκτη σχετικότητας, η ποιότητα ενός Bicluster μετριέται ως το άθροισμα των τιμών του δείκτη από όλες τις επιλεγμένες καταστάσεις.[56]

**Συσχέτιση βασισμένη σε μέτρα (Correlation-based Measures)**

Οι συντελεστές συσχέτισης έχουν χρησιμοποιηθεί εκτενώς σε διαφορετικά είδη αναλύσεων μικροσυστοιχιών, όπως το clustering. Αυτές οι στατιστικές μετρούν την συνολική ομοιότητα των γονιδίων

χωρίς τοποθέτηση, με οποιαδήποτε έμφαση σε συγκεκριμένα μεγέθη, λαμβάνοντας υπόψη τόσο τις αρνητικές συσχετίσεις, όσο και τις θετικές. Παρακάτω εξετάζονται οι προσεγγίσεις βάσει συσχέτισης (correlation-based) που προτείνονται για την αξιολόγηση ενός Bicluster.[56]

**Συντελεστής συσχέτισης Pearson (Pearson's Correlation Coefficient, PCC)**

Ο PCC μεταξύ δύο μεταβλητών ορίζεται ως, η συνδιακύμανση των δύο μεταβλητών διαιρούμενη από το προϊόν των τυπικών αποκλίσεων τους. Πρόκειται για ένα μέτρο της γραμμικής εξάρτησης μεταξύ δύο μεταβλητών, και δίνει μια τιμή μεταξύ +1 και -1, συμπεριλαμβανομένων και των δύο. Η τιμή 1 υποδηλώνει ότι μία γραμμική εξίσωση περιγράφει τη σχέση μεταξύ των δύο μεταβλητών τέλεια (θετική συσχέτιση), ενώ η τιμή -1 υποδηλώνει ότι όλα τα σημεία δεδομένων βρίσκονται σε μια γραμμή, για την οποία μια μεταβλητή μειώνεται καθώς η άλλη αυξάνεται (αρνητική συσχέτιση). Η τιμή 0 σημαίνει ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών.

Ο PCC είναι ένα πολύ αποτελεσματικό μέτρο, που ποσοτικοποιεί την από κοινού ρύθμιση μεταξύ των ζευγαριών των γονιδίων και επιτρέπει τη σύλληψη της **μετατόπισης** όσο και της **κλιμάκωσης** προτύπων, τα οποία θα προσδιοριστούν χωριστά από πρόσθετα και πολλαπλασιαστικά μοντέλα, αντίστοιχα. Παρ ' όλα αυτά, ο PCC **δεν είναι αποτελεσματικός** για την αναγνώριση

σταθερών Biclusters ή σταθερών μοτίβων σειράς, δεδομένου ότι αυτά τα είδη των προτύπων θα κάνουν τον παρονομαστή μηδέν.

**Ο PCC μεταξύ δύο σειρών (γονίδια) $i_1, i_2 \in I$ σε σχέση με τις στήλες (καταστάσεις) $j \in J$ ορίζεται ως:**

$$\text{PCC}(i_1,i_2) = \frac{\sum_{j=1}^{|J|}(a_{i_1 j}-a_{i_1 J})(a_{i_2 j}-a_{i_2 J})}{\sqrt{\sum_{j=1}^{|J|}(a_{i_1 j}-a_{i_1 J})^2 \sum_{j=1}^{|J|}(a_{i_2 j}-a_{i_2 J})^2}} \quad (3.3.5)$$

Όπου $a_{i_1 j}$ και $a_{i_2 j}$ συμβολίζουν τα στοιχεία στις γραμμές $i_1$ , $i_2$ και τη στήλη j και $a_{i_1 J}, a_{i_2 J}$ αντιπροσωπεύουν τα μέσα των γραμμών $i_1$και $i_2$ αντίστοιχα.

Ο PCC ποσοτικοποιεί τη συνοχή μεταξύ των ζευγαριών των γονιδίων. Ως εκ τούτου, προκειμένου να μετρηθεί η συνοχή του Bicluster, πρέπει κανείς να υπολογίσει όλα τα ζεύγη τιμών PCC μεταξύ των σειρών στο ίδιο Bicluster. Επιπροσθέτως, ο PCC έχει νόημα μόνο για τη μέτρηση της συνοχής μεταξύ των γραμμών, διότι είναι υπερβολικά περιοριστικός αν χρησιμοποιηθεί για τη μέτρηση της συνοχής μεταξύ των στηλών ταυτόχρονα. Ο PCC έχει χρησιμοποιηθεί ως έχει (μεταξύ των γραμμών) σε αρκετές έρευνες, ενώ σε άλλες μελέτε οι συγγραφείς ορίζουν ένα μέτρο που βασίζεται στον PCC (PCC-based),**το οποίο ονομάζεται Μέση Συσχέτιση (Average Corrrelation(-AC), με τη μορφή της εξίσωσης** (3.3.6) :

$$\text{AC(A)} = \frac{\sum_{i_1=1}^{|I|-1}\sum_{i_2=i_1+1}^{|I|} PCC(i_1,i_2)}{\binom{|I|}{2}} \quad (3.3.6)$$

Για να χρησιμοποιηθεί ο PCC για τη μέτρηση της συνοχής μεταξύ των στηλών, καθώς και για να ξεπεραστεί το θέμα της περιοριστικότητας, τόσο οι Yang, Dai ,και Yan στο έργο τους, όσο και οι Teng και Chan έχουν ορίσει ένα μέτρο που βασίζεται στον PCC (PCC-based), το οποίο είναι μέτρο αξιολόγησης του Bicluster όσον αφορά τη βελτίωση συσχέτισης, είτε στις γραμμές είτε στις στήλες.[56]

**Βαθμός συσχέτισης υποπίνακα (Sub-Matrix Correlation Score, SCS)**

Οι Yang, Dai και Yan χρησιμοποιούν το Pearson correlation score ως τη βάση για να καθορίσουν το μέτρο τους, υποθέτοντας ότι ένα τέλειο συσχετιζόμενο πρότυπο ικανοποιεί τέλεια γραμμική συσχέτιση σε διανύσματα γραμμής και στήλης. Σχέσεις, συσχετίσεις για γραμμές και στήλες παρουσιάζονται παρακάτω:

$$s_{row} = min_{i_1 \in \mathrm{I}} \left(S_{i_1 J}\right), S_{i_1 J} = 1 - \frac{\sum_{i_2 \neq i_1, i_2 \in \mathrm{I}} |cor(\, x_{i_1 J}, x_{i_2 J})|}{|I|-1} \qquad (3.3.7)$$

$$s_{col} = min_{j_1 \in \mathrm{J}} \left(S_{I j_1}\right), S_{I j_1} = 1 - \frac{\sum_{j_2 \neq j_1, j_2 \in \mathrm{J}} |cor(\, x_{I j_1}, x_{I j_2})|}{|J|-1} \qquad (3.3.8)$$

Όπου $cor(\, x_{i_1 J}, x_{i_2 J})$ και $cor(\, x_{I j_1}, x_{I j_2})$ αντιπροσωπεύουν τον PCC του κάθε ζεύγους γονιδίων ή καταστάσεων στο Bicluster, αντίστοιχα.

Το Srow score αντικατοπτρίζει το βαθμό συσχέτισης σχετικά με τις σειρές του Bicluster, ενώ το Scol ανακλά το βαθμό συσχέτισης σχετικά με τις στήλες, όπου και οι δύο ορισμοί είναι ασύμμετροι.[56]

**Το σκορ συσχέτισης (correlation score) του υποπίνακα ορίζεται ως η ελάχιστη τιμή των δύο αυτών scores**:

$$\mathrm{S(A)} = min(S_{row}(I,J), S_{col}(I,J)\,) \qquad (3.3.9)$$

**Μέση τιμή συσχέτισης Average Correlation Value (ACV)**

**Η ACV** προτάθηκε από τους Teng και Chan για την αξιολόγηση της ομοιογένειας ενός Bicluster ή ενός πίνακα δεδομένων με τον ακόλουθο τρόπο:

$$\mathrm{ACV} = \max \left\{ \frac{\sum_{i_1=1}^{|I|} \sum_{i_2=1}^{|I|} \left| r_{row_{i_1,i_2}} \right| - |I|}{|I|^2 - |I|}, \frac{\sum_{j_1=1}^{|J|} \sum_{j_2=1}^{|J|} \left| r_{col_{j_1,j_2}} \right| - |J|}{|J|^2 - |J|} \right\} \qquad (3.3.10)$$

Όπου $r_{row_{i_1,i_2}}, r_{col_{j_1,j_2}}$ αναφέρονται στη συσχέτιση μεταξύ κάθε ζεύγους των σειρών $i_1$,$i_2$ ή στηλών $j_1$,$j_2$ σύμφωνα με τον συντελεστή Pearson.

Η ACV (A) έχει αποδειχθεί ότι είναι στο διάστημα [0,1], όπου μία τιμή ίση με 1 σημαίνει ότι οι γραμμές ή οι στήλες του Bicluster έχουν σε μεγάλο βαθμό συνεκφραστεί, ενώ μια χαμηλή ACV σημαίνει ότι ούτε οι καταστάσεις (συνθήκες) ούτε τα γονίδια είναι παρόμοια. Επομένως, οι υψηλότερες τιμές της ACV

προτιμώνται. Οι συγγραφείς, στην εργασία τους, απέδειξαν ότι **η ACV δίνει πάντα την επιθυμητή τιμή** τόσο για το πρόσθετο όσο και για το πολλαπλασιαστικό μοντέλο, σε αντίθεση με το MSR.

Αν και η ACV παρουσιάζεται ως το κριτήριο για την αξιολόγηση Biclusters, δεν έχει χρησιμοποιηθεί για να καθοδηγήσει την αναζήτηση στον αλγόριθμό τους. Αντ' αυτού, ο αλγόριθμος βασίζεται στη χρήση ενός σταθμισμένου συντελεστή συσχέτισης.[56]

**Μέσος Όρος Spearman's Rho (Average Spearman's Rho, ASR)**

Ο ASR προτάθηκε για πρώτη φορά από τους Ayadi, Elloumi και Hao και βασίζεται στη χρήση του βαθμού συσχέτισης Spearman, η οποία μετρά την στατιστική εξάρτηση μεταξύ δύο μεταβλητών, αξιολογώντας το πόσο καλά η σχέση τους μπορεί να περιγραφεί χρησιμοποιώντας μια μονότονη συνάρτηση. Η τιμή του κυμαίνεται μεταξύ -1 και +1.

Η πιο σημαντική διαφορά μεταξύ των συσχετίσεων του Spearman και Pearson είναι ότι ο Pearson αξιολογεί γραμμικές σχέσεις, ενώ ο Spearman αξιολογεί μονοτονικές σχέσεις. Ως εκ τούτου, η συσχέτιση Spearman είναι λιγότερο ευαίσθητη σε ακραίες τιμές, που βρίσκονται στις ουρές και των δύο δειγμάτων.

Παρ' όλα αυτά, όταν τα δεδομένα διανέμονται περίπου ελλειπτικά και δεν υπάρχουν εξαιρετικά ακραίες τιμές, οι συσχετίσεις Spearman και Pearson δίνουν παρόμοιες τιμές.[56]

**Ο ASR** υπολογίζεται **όπως στην παρακάτω εξίσωση** και εξάγει μια τιμή στο διάστημα [-1,1], όπου η υψηλή /χαμηλή τιμή κοντά στο 1 / -1 δείχνει ότι τα γονίδια ή οι συνθήκες του Bicluster συσχετίζονται ισχυρώς, είτε θετικά είτε αρνητικά, αντίστοιχα.

$$\text{ASR} = 2 * \max\left\{\frac{\sum_{i\in I}\sum_{j\geq i+1, j\in I}\rho_{ij}}{|I|(|I|-1)}, \frac{\sum_{k\in J}\sum_{l\geq k+1, l\in J}\rho_{kl}}{|J|(|J|-1)}\right\} \qquad (3.3.11)$$

Όπου $\rho_{ij}$, $\rho_{kl}$ αναφέρονται στη συσχέτιση Spearman μεταξύ δύο γονιδίων ή καταστάσεων, αντίστοιχα.[56]

**Μέτρο διπλής κατηγοριοποίησης Spearman (Spearman's Biclustering Measure, SBM)**

Το SBM έχει προταθεί πρόσφατα από τους Flores, Inza, Larranaga και Calvo, τονίζοντας την ικανότητά του να αναγνωρίζει πολύπλοκη συνοχή προτύπων στα Biclusters, όπως η μετατόπιση και η κλιμάκωση, όπως και, αρνητικές συσχετίσεις.

Ο υπολογισμός του συντελεστή αυτού επιτυγχάνεται σε δύο στάδια. Πρώτον, τα δεδομένα μετατρέπονται σε τάξεις, προκειμένου να υπολογιστεί ο συντελεστής Spearman, που συμβολίζεται ως $r_{x,y}$για κάθε ζεύγος γονιδίων ή καταστάσεων. Χρησιμοποιώντας αυτούς τους συντελεστές, **το SBM ορίζεται ως εξής:**

$$\text{SBM}(A_{ij}) = \alpha(A_{ij}) * r_{A_{ij}}^{\bar{G}} * \beta(A_{ij}) * r_{A_{IJ}}^{\bar{G}} \qquad (3.3.12)$$

Όπου $r_{A_{ij}}^{\bar{G}}$ και $r_{A_{IJ}}^{\bar{G}}$ υποδηλώνουν την συνοψιζόμενη έκφραση των τάσεων που παρατηρούνται στα γονίδια και στις καταστάσεις του Bicluster, αντίστοιχα. Αυτά υπολογίζονται όπως στις εξισώσεις που ακολουθούν:

$$r_{A_{IJ}}^{\bar{G}} = \frac{2}{|I| * (|I| - 1)} * \sum_{i=1}^{|I|} \sum_{i'=i+1}^{|I|} \left| \boldsymbol{r}_{|ii|'}^{\boldsymbol{G}} \right|,$$
$$r_{A_{IJ}}^{\bar{C}} = \frac{2}{|J| * (|J| - 1)} * \sum_{j=1}^{|J|} \sum_{j'=j+1}^{|J|} \left| \boldsymbol{r}_{|jj|'}^{\boldsymbol{C}} \right| \quad (3.3.13)$$

Όπου $|\boldsymbol{r}_{|ii|'}^{\boldsymbol{G}}|$ και του $|\boldsymbol{r}_{|jj|'}^{\boldsymbol{C}}|$ αντιστοιχούν στις απόλυτες τιμές του συντελεστή συσχέτισης Spearman μεταξύ ενός ζεύγους γονιδίων $i, i'$ και ενός ζεύγους καταστάσεων $j, j'$ , αντίστοιχα.

Οι όροι $\alpha$ ($A\ ij$) και $\beta$ ($A\ ij$) στην εξίσωση (3.3.12) αντιπροσωπεύουν τους συντελεστές αξιοπιστίας, και το βάρος της επιρροής των τάσεων που παρατηρήθηκαν και στα δύο γονίδια και στις καταστάσεις, αντίστοιχα.[56]

**Μέτρα βασισμένα σε τυποποίηση (Standardisation-based Measures)**

Μία σημαντική παρατήρηση που μπορεί να εξαχθεί από την ανάλυση των Biclusters είναι ότι το εύρος των τιμών έκφρασης που λαμβάνονται από τα γονίδια μπορεί να ποικίλλει σημαντικά ανάλογα με τη μικροσυστοιχία που λαμβάνεται ως είσοδος. Συνεπώς, προκειμένου να γίνει σωστή σύγκριση μεταξύ κάθε γονιδίου και προτύπου, μια προηγούμενη διαδικασία τυποποίησης του Bicluster θα επιτρέψει στα επίπεδα έκφρασης να κλιμακωθούν σε ένα κοινό φάσμα. Ο μηχανισμός αυτός θα είναι επίσης υπεύθυνος για την ομαλοποίηση της συμπεριφοράς του γονιδίου, δεδομένου ότι η πιο σημαντική πτυχή είναι να χαρακτηρίσουμε το γονίδιο καλύτερα από την τάση του, παρά από τις αριθμητικές τιμές του.

Γονιδιακή τυποποίηση (**Gene Standardisation)** ενός A Bicluster αντιστοιχεί στο τυποποιημένο Bicluster $\hat{A}$, του οποίου το στοιχείο $\widehat{a_{ij}}$, το εξάγουμε ως εξής:

$$\widehat{a_{ij}} = \frac{\widehat{a_{ij}} - \mu_{gi}}{\sigma_{gi}}, 1 \leq i \leq |I|, 1 \leq j \leq |J| \quad (3.3.14)$$

Όπου $\sigma_{gi}$ είναι η τυπική απόκλιση όλων των τιμών έκφρασης του γονιδίου $i$ και $\mu_{gi}$ είναι ο μέσος όρος της σειράς $i$ στο A.

Μέσω της τυποποίησης, δύο ευδιάκριτα «έργα» εκτελούνται. Το πρώτο είναι να μετατοπιστεί το σύνολο των γονιδίων σε ένα παρόμοιο εύρος τιμών (κοντά στο 0 στην περίπτωση αυτή). Το δεύτερο είναι α) να ομογενοποιήσει τις τιμές έκφρασης για κάθε γονίδιο, τροποποιώντας με αυτόν τον τρόπο τις αξίες τους κάτω από όλες τις συνθήκες, και β) να εξομαλύνει τη γραφική αναπαράστασή τους, λόγω της διόρθωσης του καθολικού παράγοντα κλιμάκωσης στον παρονομαστή.[56]

Τρία διαφορετικά μέτρα έχουν οριστεί στη βιβλιογραφία, χρησιμοποιώντας τη διαδικασία της τυποποίησης για την αξιολόγηση Bicluster. Δύο από αυτά, MSA και VE, βασίζονται στη γονιδιακή τυποποίηση, ενώ το άλλο μέτρο βασίζεται σε μια κατάσταση τυποποίησης, όπως περιγράφεται παρακάτω.

**Μέγιστη τυπική περιοχή (Maximal Standard Area, MSA)**

Η ιδέα πίσω από τη MSA είναι να μετρηθεί η έκταση της περιοχής μεταξύ των μέγιστων και ελάχιστων τιμών των επιπέδων έκφρασης που λαμβάνουν τα γονίδια υπό τις καταστάσεις που περιέχονται στο Bicluster. Έτσι, αυτό που υπολογίζεται είναι η περιοχή που απεικονίζεται από τη μέγιστη διακύμανση των επιπέδων έκφρασης

για κάθε πειραματική κατάσταση. Για κάθε κατάσταση, λαμβάνονται οι ελάχιστες και οι μέγιστες τιμές του επιπέδου έκφρασης για όλα τα γονίδια που περιέχονται στο Bicluster. Αυτά τα ζεύγη τιμών ορίζουν μια ζώνη σε όλες τις καταστάσεις στο Bicluster, και η περιοχή της ζώνης αυτής είναι, ως εκ τούτου, το μέτρο MSA.

**Τα ψηλότερα και χαμηλότερα όρια ενός Bicluster για κάθε κατάσταση $j$ ορίζονται ως $M_j(A)$ και $m_j(A)$ αντίστοιχα, με τον ακόλουθο τρόπο**:

$$M_j(A) = \max_i a_{ij} \text{ ,}\forall \text{i; } m_j(A) = min_i \ a_{ij}, \forall \text{i} \qquad (3.3.15)$$

Χρησιμοποιώντας αυτά τα όρια, **η MSA ορίζεται ως η οριοθετημένη περιοχή από τα όρια της κάθε κατάστασης στο τυποποιημένο Bicluster**:

$$\text{MSA(A)} = \sum_{j=1}^{|J|-1} \left| \frac{M_j(\hat{A}) - m_j(\hat{A}) + M_{j+1}(\hat{A}) - m_{j+1}(\hat{A})}{2} \right| \qquad (3.3.16)$$

Αν τα γονίδια ενός Bicluster έχουν μια απόλυτα συνεκτική τάση, τότε η MSA (A) είναι ίση με το μηδέν. Αντίθετα, η MSA θα πάρει υψηλότερες τιμές όταν τα γονίδια είναι λιγότερο συσχετισμένα μεταξύ τους, αυτό οφείλεται στο γεγονός ότι τα $\text{M}(\hat{A})$ και $\text{m}(\hat{A})$ είναι πιο απομακρυσμένα μεταξύ τους.[56]

**<u>Εικονικό σφάλμα (Virtual Error, VE)</u>**

Το Virtual Error ακολουθεί παρόμοιες παραδοχές σχετικά με τη διαδικασία τυποποίησης στο πλαίσιο της αξιολόγησης με τη MSA. Η βασική ιδέα πίσω από το VE είναι να μετρήσει πώς τα γονίδια ακολουθούν τη γενική τάση εντός του Bicluster. Για να «πιάσει» αυτή τη γενική τάση των γονιδίων σε όλες τις καταστάσεις που εμπεριέχονται στο Bicluster, μια νέα εικονική γραμμή υπολογίζεται από τα γονίδια του Bicluster, που ονομάζεται εικονικό σχέδιο ή εικονικό γονίδιο $\rho$. Κάθε στοιχείο $\rho_j$του $\rho$ υπολογίζεται ως μέση τιμή της $j$-στήλης στην παρακάτω εξίσωση:

$$\rho_j = \frac{1}{|I|} \sum_{i=1}^{|I|} a_{ij} \qquad (3.3.17)$$

Μόλις το εικονικό γονίδιο $\rho$ έχει υπολογιστεί, και για να εκτιμηθεί πόσο καλά ένα συγκεκριμένο γονίδιο $g_i$ του Bicluster, ακολουθεί τη γενική τάση, το VE υπολογίζει τις διαφορές μεταξύ των τιμών έκφρασης του επίπεδου $g_i$ και τις τιμές του $\rho$, για κάθε πειραματική κατάσταση του Bicluster. Αυτές οι διαφορές υπολογίζονται χρησιμοποιώντας τα τυποποιημένα γονίδια στο Bicluster, καθώς και το τυποποιημένο εικονικό γονίδιο $\hat{\rho}_j$. **Ο ορισμός του VE(A) είναι ο εξής:**

$$\text{VE(A)} = \frac{1}{|I|*|J|} \sum_{i=1}^{|I|} \sum_{j=1}^{|J|} |a_{ij} - \hat{\rho}_j| \qquad (3.3.18)$$

Το VE υπολογίζει τις διαφορές μεταξύ των πραγματικών γονιδίων και των εικονικών, τη στιγμή που έχουν τυποποιηθεί. Ως εκ τούτου, όσο πιο όμοια είναι τα γονίδια, τόσο χαμηλότερη είναι η τιμή για το VE. Πράγματι, το VE είναι μηδέν για εκείνα τα Biclusters που ακολουθούν είτε μία τέλεια μετατόπιση ή μία κλιμάκωση

προτύπου. Παρ 'όλα αυτά, το VE **δεν μπορεί να αποδειχθεί ότι αναγνωρίζει και τα δύο είδη των προτύπων ταυτόχρονα**.[56]

**Μετατόπιση εικονικού σφάλματος( Transposed Virtual Error ,$VE^t$)**

Το $\mathbf{VE}^t$ υπολογίζεται παρομοίως με το VE αλλά, λαμβάνοντας υπόψη το μεταφερόμενο Bicluster. Η ιδέα εδώ είναι να δημιουργήσουμε το εικονικό μοτίβο στην "κατάσταση διάστασης" (condition dimension), η οποία έχει ονομαστεί εικονική κατάσταση. Στη συνέχεια, οι διαφορές μεταξύ των τυποποιημένων τιμών για κάθε κατάσταση και των τυποποιημένων εικονικών καταστάσεων υπολογίζονται με τον ίδιο τρόπο όπως στο VE.

Με αυτό τον τρόπο, η εικονική κατάσταση υπολογίζεται ως:

$$\rho_i = \frac{1}{|J|}\sum_{j=1}^{|J|} a_{ij} \tag{3.3.19}$$

και η τιμή Vet του bicluster A επιτυγχάνεται με τη χρήση της τυποποιημένης εικονικής κατάστασης $\hat{\rho}$ μαζί με το τυποποιημένο Bicluster όπως τα δείγματα:

$$\mathrm{VE}^t(\mathrm{A}) = \frac{1}{|I|*|J|}\sum_{i=1}^{i=|I|}\sum_{j=1}^{j=|J|}|\hat{a}_{ij} - \hat{\rho}_i| \tag{3.3.20}$$

Το $\mathbf{VE}^t$ έχει αποδειχθεί ότι είναι μηδέν για τα Biclusters με τέλεια μετατόπιση, με κλιμάκωση ή και σε συνδυασμό των δύο. Ως εκ τούτου, φαίνεται να αναγνωρίζει αποτελεσματικά τόσο τη μετατόπιση όσο και τη κλιμάκωση σε Biclusters είτε ταυτόχρονα είτε ανεξάρτητα.[56]

**Βαθμολογίες ομοιότητας (Similarity Scores)**

Ο Liu και ο Wang πρότειναν τη χρήση μιας βαθμολογίας ομοιότητας μεταξύ δύο γονιδίων και επίσης μια βαθμολογίας ομοιότητας για έναν υποπίνακα. Η πρώτη χρησιμοποιείται όταν το αναφερόμενο γονίδιο είναι γνωστό εκ των προτέρων. Όταν δεν είναι γνωστό, ένας αριθμός γονιδίων θα μπορούσε επίσης να επιλεγεί τυχαία όπως τα αναφερόμενα γονίδια.[56]

**Βαθμός ομοιότητας μεταξύ γονιδίων (Similarity Score Between Genes)**

Ο βαθμός ομοιότητας μεταξύ δύο γονιδίων (γονίδιο *i* και γονίδιο αναφοράς *i**) υπό τον όρο *j* υπολογίζεται σύμφωνα με την παρακάτω εξίσωση:

$$s_{ij} = \begin{cases} 0 & if\ d_{ij} > a * d_{avg} \\ 1 - \frac{d_{ij}}{a*d_{avg}} + \beta & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \tag{3.3.21}$$

Όπου $d_{avg}$ ορίζεται ως η μέση τιμή της απόστασης όλων των στοιχείων του πίνακα έκφρασης.

$$d_{avg} = \frac{\sum_{i\in I, j\in J} d_{ij}}{|I||J|} \tag{3.3.22}$$

και το $d_{ij}$ είναι η απόλυτη τιμή της διαφοράς έκφρασης μεταξύ του γονιδίου *I* και του γονιδίου αναφοράς *i** για την κατάσταση *j* στον πίνακα έκφρασης α:

$$d_{ij} = |a_{ij} - a_{i*j}| \quad (3.3.23)$$

Το $a * d_{avg}$ χρησιμοποιείται ως το όριο για να αγνοήσει στοιχεία με μεγάλο $d_{ij}$, ώστε να βρεθούν σταθερά Biclusters, και το $\beta$ είναι το κέρδος για τις μικρές $d_{ij}$. Με αυτό τον τρόπο, το $\beta$ διευρύνει τον βαθμό ομοιότητας για τις μικρές $d_{ij}$ και αγνοεί $d_{ij}$ που είναι μεγαλύτερα από το όριο. Σύμφωνα με την εξίσωση $s_{ij}$ s(3.3.21), η τιμή της $s_{ij}$ θα είναι πάντα μεγαλύτερη από ή ίση με μηδέν, όπου είναι η χειρότερη αξία του βαθμού ομοιότητας.[56]

**Βαθμός ομοιότητας για ένα Bicluster (Similarity Score For a Bicluster, SS)**

Για κάθε γραμμή $i \in I$, ο βαθμός ομοιότητας της σειράς $i$ στο A είναι:

$$s(i, J) = \sum_{j \in J} s_{ij} \quad (3.3.24)$$

Ομοίως για κάθε στήλη $j \in J$ ο βαθμός ομοιότητας στην A αντιστοιχεί σε :

$$s(I, j) = \sum_{i \in I} s_{ij} \quad (3.3.25)$$

Χρησιμοποιώντας αυτές τις εξισώσεις ο βαθμός ομοιότητας για το Bicluster υπολογίζεται ως η ελάχιστη τιμή του βαθμού ομοιότητας και των δύο γονιδίων και της κατάστασης των Biclusters:

$$s(A) = s(I, J) = \min \{min_{i \in I} s(i, J), min_{j \in J} s(I, j)\} \quad (3.3.26)$$

Ο στόχος όταν ψάχνουμε για Biclusters είναι να βρεθούν υπο-πίνακες με υψηλότερες τιμές για τον βαθμό ομοιότητας. Προκειμένου να βελτιωθεί η ποιότητα της παραγωγής, οι Liu και Wang εισήγαγαν επίσης ως δεύτερο κριτήριο τον μέσο όρο βαθμού ομοιότητας. Αυτός αποτελείται από τον μέσο όρο όλων των όμοιων τιμών της εξίσωσης $d_{ij}$ για όλα τα στοιχεία του Bicluster.

Παρόλο που ο τύπος του Bicluster που βρέθηκε χρησιμοποιώντας το βαθμό ομοιότητας, εξαρτάται από τις τιμές για τα διαφορετικά κατώτατα όρια ($\alpha$, $\beta$, και $\gamma$ για το μέσο όρο), αναγνωρίζονται μόνο τα σταθερά και τα πρόσθετα Biclusters.[56]

Στην εργασία αυτή βασιστήκαμε σε όλα τα παραπάνω, δηλαδή τα πιο γνωστά υφιστάμενα μέτρα αξιολόγησης για Biclusters, αλλά παράλληλα ακολουθήσαμε μια δική μας προσέγγιση, η οποία παρουσιάζεται στο επόμενο κεφάλαιο (Κεφάλαιο 4).

# 4

# Δεδομένα και Προτεινόμενη Μεθοδολογία

## 4.1. Ανάπτυξη Μεθοδολογίας

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 ενδομήτριου) 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 κυκλοφορόντων καρκινικών κυττάρων (CTCs) στο περιφερικό αίμα των θηλέων ασθενών με καρκίνο.[57],[12]

**Καρκινικές κυτταρικές σειρές** καλούνται τα καρκινικά κύτταρα που συνεχίζουν να διαιρούνται και να αυξάνονται με το χρόνο, κάτω από ορισμένες συνθήκες, σε ένα εργαστήριο. Σε σύγκριση με τα κανονικά κύτταρα, τα καρκινικά κύτταρα έχουν υποστεί κάποιες μεταλλάξεις. Αυτές είναι υπεύθυνες για την ανάπτυξη, τις μεταστάσεις και την αντίσταση που έχει στα φάρμακα ο καρκίνος. Οπότε, η χρησιμότητα των μοντέλων εξαρτάται από το κατά ποσό το μοντέλο μπορεί να περιγράψει τις αλλαγές στους πρωτεύοντες/κύριους όγκους. Γι' αυτό το λόγο στην έρευνα για τον καρκίνο χρησιμοποιούνται συλλογές από κυτταρικές σειρές που έχουν ληφθεί από όγκους επειδή αυτές οι κυτταρικές σειρές περιέχουν εκατοντάδες ή και χιλιάδες από αυτές τις μεταλλάξεις των κυττάρων που προέκυψαν στον όγκο από τον οποίο ελήφθησαν. Οι καρκινικές κυτταρικές σειρές χρησιμοποιούνται στην έρευνα για τη μελέτη της βιολογίας του καρκίνου και για τη δοκιμή των θεραπειών του καρκίνου.[58] Η *Εγκυκλοπαίδεια Καρκινικής Κυτταρικής Σειράς* (CCLE) έχει συσταθεί ως μια προσπάθεια λεπτομερούς γενετικού χαρακτηρισμού ενός μεγάλου αριθμού ανθρώπινων καρκινικών κυτταρικών σειρών (~1000), παρέχοντας δημόσια πρόσβαση ανάλυσης και απεικόνισης του αριθμού αντιγράφων του DNA, της έκφρασης του mRNA, των δεδομένων μετάλλαξης, κ.ά.[59]

Διαφορετικές μέθοδοι biclustering χρησιμοποιούν διαφορετικές έννοιες και αντιμετωπίζουν το πρόβλημα Bicluster από άλλη οπτική ο καθένας (Υποκεφάλαια 2.8, 3.1, 3.2). Γι' αυτό το λόγο καθίσταται αρκετά δύσκολη

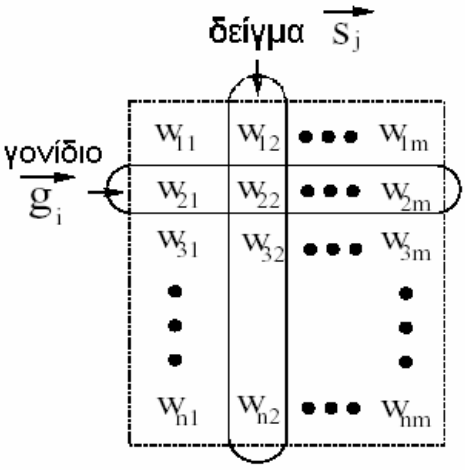
η επιλογή του πιο κατάλληλου αλγορίθμου για εφαρμογή. Ο αλγόριθμος Cheng και Church (2000) είναι ένας από τους πιο παλιούς και αξιόπιστους αλγορίθμους.[4] Τον επιλέξαμε γιατί αποτελεί μία μέθοδο ικανή να αντιμετωπίσει το μεγάλο σύνολο δεδομένων μας, παρέχει την δυνατότητα μη-επικάλυψης και τα Biclusters που εξάγει έχουν συναφείς (coherent) coherent τιμές, γεγονός που μας επιτρέπει να μελετήσουμε τη συμπεριφορά των γονιδίων κατά μήκος των σειρών χωρίς οι τιμές να είναι απαραίτητα σταθερές.

## 4.2. Δεδομένα

Το αρχείο προς επεξεργασία που μας δόθηκε αποτελείται από έναν πίνακα 33096 γραμμών και 38 στηλών. Οι γραμμές αντιπροσωπεύουν τα γονίδια και οι στήλες τις καρκινικές κυτταρικές σειρές.[57]

Η τεχνολογία των μικροσυστοιχιών έχει πρόσβαση σε μεγάλο μέγεθος ακολουθιών DNA κάτω από πολλές συνθήκες. Ένα σύνολο δεδομένων έκφρασης γονιδίων εκφράζεται ως πίνακας έκφρασης πραγματικών τιμών,

$$M=\{wij \mid 1\leq i \leq n,\ 1 \leq j \leq m\ \}$$




| | |
|---|---|
| $n$ | αριθμός γονιδίων |
| $m$ | αριθμός δειγμάτων |
| $M$ | ένας πίνακας έκφρασης γονιδίων |
| $w_{ij}$ | κελί στον πίνακα έκφρασης γονιδίων |
| $\vec{g_i}$ | ένα γονίδιο |
| $\vec{s_j}$ | ένα δείγμα |
| $G, G', G_0, \ldots$ | ένα σετ γονιδίων |
| $S, S', S_0, \ldots$ | ένα σετ δειγμάτων |

**Εικόνα 4.1.** - Σύνολο δεδομένων εκφρασμένο ως πίνακα πραγματικών τιμών

όπου οι γραμμές είναι εκφράσεις προτύπων στα γονίδια και οι στήλες τα προφίλ των δειγμάτων. Ο αρχικός πίνακας δεδομένων περιέχει θόρυβο, χαμένες τιμές και αποκλίσεις που προκύπτουν από τις πειραματικές διαδικασίες. Άρα, απαιτείται ένα στάδιο προεπεξεργασίας, ώστε να λυθούν αυτά τα προβλήματα, πριν εφαρμόσουμε τις τεχνικές κατηγοριοποίησης. Με το στάδιο της προεπεξεργασίας δεν ασχολούμαστε σε αυτό το σημείο, δεδομένου ότι τα δεδομένα μας έχουν περάσει από αυτό το στάδιο.[32],[33]

Όπως θα αναφερθεί στη συνέχεια του κειμένου, ο αρχικός πίνακας των 33096 γραμμών και 38 στηλών διαμορφώθηκε σε 1000 γραμμές και 38 στήλες.

### 4.2.1. Επιλογή αλγορίθμου Cheng και Church

Στη δική μας εργασία, όπως αναφέρθηκε ήδη, επιλέξαμε να εφαρμόσουμε τον αλγόριθμο διπλής κατηγοριοποίησης Cheng και Church (2000).[4] Κατά την εφαρμογή του αλγορίθμου εμφανίστηκαν προβλήματα που κληθήκαμε να αντιμετωπίσουμε με μια σειρά βελτιώσεων ώστε να πετύχουμε την καλύτερη ομαδοποίηση των γονιδίων ή καταστάσεων, με βάση την πλειοψηφία των περιπτώσεων παθολογίας. Στόχοι μας ήταν: να πετύχουμε: 1) τη μείωση των αποστάσεων μεταξύ των γονιδίων, αλλά και 2) παρόμοιες διακυμάνσεις μεταξύ των γονιδίων που ομαδοποιούνται στο Bicluster.

Στην ανάλυση των δεδομένων γονιδιακής έκφρασης, πέραν της ομαδοποίησης των γονιδίων που βασίζονται στη 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 ή υποομάδες. Η διπλή κατηγοριοποίηση επιτυγχάνει αυτή την προσπάθεια ομαδοποιώντας υποσύνολα γονιδίων και καταστάσεων με σκορ υψηλής ομοιότητας, το οποίο αποτελεί μέτρο συνοχής για αυτά.

Ένα συγκεκριμένο σκορ που αντιπροσωπεύει τα λογαριθμισμένα δεδομένα έκφρασης αποτελεί τη διακύμανση της ομάδας όλων των στοιχείων του Bicluster (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 τη διακύμανση της ομάδας όλων των στοιχείων του Bicluster και τη 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 κατώφλι (threshold). Μια ειδική περίπτωση για ένα βέλτιστο σκορ (μηδενικό «mean squared residue score») είναι ένα σταθερό Bicluster (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 σταθερό (constant).

Στην περίπτωση της γονιδιακής έκφρασης, περισσότερο ενδιαφέρον δεν παρουσιάζει η εύρεση ενός μεγάλου -αναφορικά με τον αριθμό γονιδίων- 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.[4]

### 4.2.2. Εισαγωγή στη χρήση του αλγορίθμου Cheng και Church

Ένας πίνακας γονιδίων-καταστάσεων αποτελείται από πραγματικούς αριθμούς, με πιθανές μηδενικές τιμές σε κάποια από τα στοιχεία του. Κάθε στοιχείο-κελί του πίνακα ανακλά το επίπεδο έκφρασης ενός γονιδίου κάτω από μια συγκεκριμένη κατάσταση και αντιπροσωπεύεται από έναν πραγματικό αριθμό που αποτελεί τον λογάριθμο της σχετικής έκφρασης του mRNA του γονιδίου σ’ αυτή την κατάσταση. Ο λογάριθμος χρησιμοποιείται για να μειώσει το δυναμικό εύρος τιμών των γονιδίων.[4]

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 »:

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

Όπου

$$\alpha_{iJ} = \frac{1}{|J|} \sum_{j \in J} \alpha_{ij}$$

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

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

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 $\delta \geq 0$.

**<u>Αλγόριθμος 0-Διαγραφή Κόμβου</u>**

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 τον 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 απαιτεί O((n+m)nm) χρόνο για την εύρεση ενός Bicluster, όπου n και m είναι τα μεγέθη των γραμμών και των στηλών του πίνακα αντίστοιχα.[4]

Η μέθοδος που ακολουθεί ο αρχικός αυτός αλγόριθμος είναι ο 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 $H <= \delta$.

Μετέπειτα προτάθηκε ο Αλγόριθμος 1 (Μονή Διαγραφή Κόμβου (Single Node Deletion)) με χρονική πολυπλοκότητα *O(nm)* και ο Αλγόριθμος 2 (Πολλαπλή Διαγραφή Κόμβου (Multiple Node Deletion)) με χρονική πολυπλοκότητα *O(m* log*n),* ο συνδυασμός των οποίων αποτελεί μια αρκετά αποτελεσματική μέθοδο εύρεσης Bicluster με χαμηλό σκορ. Η ακρίβεια και η ορθότητα αυτών των αλγορίθμων στηρίζεται σε έναν αριθμό λημμάτων, όπου οι γραμμές και οι στήλες αντιμετωπίζονται σαν σημεία σε ένα διάστημα όπου η απόσταση είναι ορισμένη.[4]

**Αλγόριθμος 1: Μονή Διαγραφή Κόμβου (Single Node Deletion)** [4]

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

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<=$\delta$ επιστρέφει τον υποπίνακα $A_{IJ}$ αλλιώς:

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

$$d(i) = \frac{1}{|I|}\sum_{i\in I}(aij - aiJ - aIj + aIJ)^2$$

και τη στήλη j που ανήκει στο J με το μεγαλύτερο

$$d(j) = \frac{1}{|J|}\sum_{j\in J}(aij - aiJ - aIj + aIJ)^2$$

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

Η ορθότητα του βήματος 1 φαίνεται από το θεώρημα το οποίο αναφέρει ότι το σετ των γραμμών που μπορούν ολικώς ή μερικώς να αφαιρεθούν με επίδραση στη μείωση του σκορ ενός Bicluster *AIJ* , είναι:

$R = \{i \in I, \frac{1}{|J|}\sum_{j\in J}(\alpha ij - \alpha iJ - \alpha Ij + \alpha iJ)^2 > H(I,J)\}$,[60] 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.

Το πρώτο στάδιο του αλγορίθμου 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(logn + logm) χρόνο.[4]

**Αλγόριθμος 2: Πολλαπλή Διαγραφή Κόμβου (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 κατώφλι (threshold) για τη μέθοδο "Multiple node deletion".

Έξοδος: AIJ, 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, $\alpha_{Ij}$ για όλα τα j που ανήκουν στο J, $\alpha_{IJ}$ και H. Αν H<=δ επιστρέφει τον υποπίνακα $A_{IJ}$ αλλιώς:

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

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

**3.** Υπολογίζει ξανά τα μεγέθη $\alpha_{Ij}$ , $\alpha_{IJ}$ και H.

**4.** Αφαιρεί όλες τις στήλες για τις οποίες ισχύει ότι

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

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

**Αλγόριθμος 3: Προσθήκη Κόμβων (Node Addition)**

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.[4]

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 $\alpha_{iJ}$ για όλα τα i που ανήκουν στο I, $\alpha_{Ij}$ για όλα τα j που

ανήκουν στο J, $\alpha_{IJ}$ και H.

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

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

**3.**Υπολογίζει ξανά τα μεγέθη $\alpha_{iJ}$ , $\alpha_{IJ}$ και H.

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

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

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

Από τη μαθηματική ανάλυση βγαίνει το συμπέρασμα ότι η πρόσθεση γραμμών και στηλών στον αλγόριθμο 3 δεν θα αυξήσουν το σκορ. Παρόλα' αυτά το *δ*-Bicluster που δίνεται στην έξοδο μπορεί να μην είναι μέγιστο. Ο 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 τη αρχή δ.

Ο αλγόριθμος 3 είναι πολύ αποτελεσματικός. Η χρονική του πολυπλοκότητα είναι συγκρίσιμη με αυτήν του αλγορίθμου 2 και είναι περίπου της τάξης του O(nm).[4]

Οι αλγόριθμοι που περιγράφηκαν παραπάνω, περικλείονται στον αλγόριθμο 4 (Εύρεση συγκεκριμένου αριθμού biclusters). 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 Cheng και Church είναι ντετερμινιστικός, οι 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.[4],[44]

Τα προβλήματα που τέθηκαν προς αντιμετώπιση, όπως αναφέρθηκαν παραπάνω, είναι: **α)** ο εξαιρετικά μεγάλος αριθμός των ομάδων διπλής κατηγοριοποίησης που προκύπτουν κατά την αρχική εκτέλεση του αλγορίθμου για τα 33096 γονίδια για τον καθένα από τους τέσσερις καρκινικούς τύπους ξεχωριστά, παραδείγματος χάριν το πλήθος των Biclusters για τον καρκίνο του μαστού ανερχόταν σε 2411, (κεφάλαιο 5), **γι' αυτό τον λόγο χρειάστηκε να μειώσουμε το πλήθος των γονιδίων που θα μελετούσαμε σε 1000 γονίδια ανά καρκινικό τύπο, β)** το μεγάλο πλήθος των γονιδίων που ομαδοποιούνται σε καθεμιά ομάδα διπλής κατηγοριοποίησης κατά την αρχική εκτέλεση του αλγορίθμου για τα 33096 γονίδια, **γι' αυτό τον λόγο χρειάστηκε να μειώσουμε το πλήθος των εξεταζόμενων Biclusters σε τρία για κάθε μία από τις τρεις εκτελέσεις του αλγορίθμου ανά καρκινικό τύπο** και **γ)** η ανομοιομορφία μεταξύ των γονιδίων που συγκεντρώνονται σε μία ομάδα διπλής κατηγοριοποίησης, δηλαδή τα γονίδια διέφεραν στην τάση παρουσιάζοντας μεγάλες διακυμάνσεις, αλλά και στη μεταξύ τους απόσταση καλύπτοντας ένα μεγάλο εύρος τιμών, **για αυτό το λόγο πραγματοποιήσαμε κάποιες τροποποιήσεις στον υπολογισμό του d*i* (mean squared residue της γραμμής *i*) και d*j* (mean squared residue της στήλης *j*),** όπως παρουσιάζονται αναλυτικά παρακάτω.

Στην παρούσα μελέτη, ο αλγόριθμος Cheng και Church εφαρμόστηκε 18 φορές, όπως αποτυπώνεται στο σχεδιάγραμμα ροής (Σχήμα 4.1).

**Συγκεκριμένα, ο αλγόριθμος εκτελέστηκε τρεις φορές:**

- Στην αρχική του μορφή, όπως έχει προταθεί από τους Cheng & Church → 1η Εκτέλεση – *didj*.
- Με αλλαγή του *di* → 2η Εκτέλεση- *di*_new,*dj*.
- Με αλλαγή του *di* και του *dj* →3η Εκτέλεση - *di*_new,*dj*_new.
  - Για τα 1000 πρώτα γονίδια για α) τον καρκίνο του μαστού, β) τον καρκίνο του τραχήλου της μήτρας, γ) τον καρκίνο των ωοθηκών, και δ) τον καρκίνο του ενδομήτριου.
  - Για τα 33096 γονίδια για τον καρκίνο του μαστού.
  - Για τα 33096 γονίδια για τον καρκίνο του τραχήλου της μήτρας.

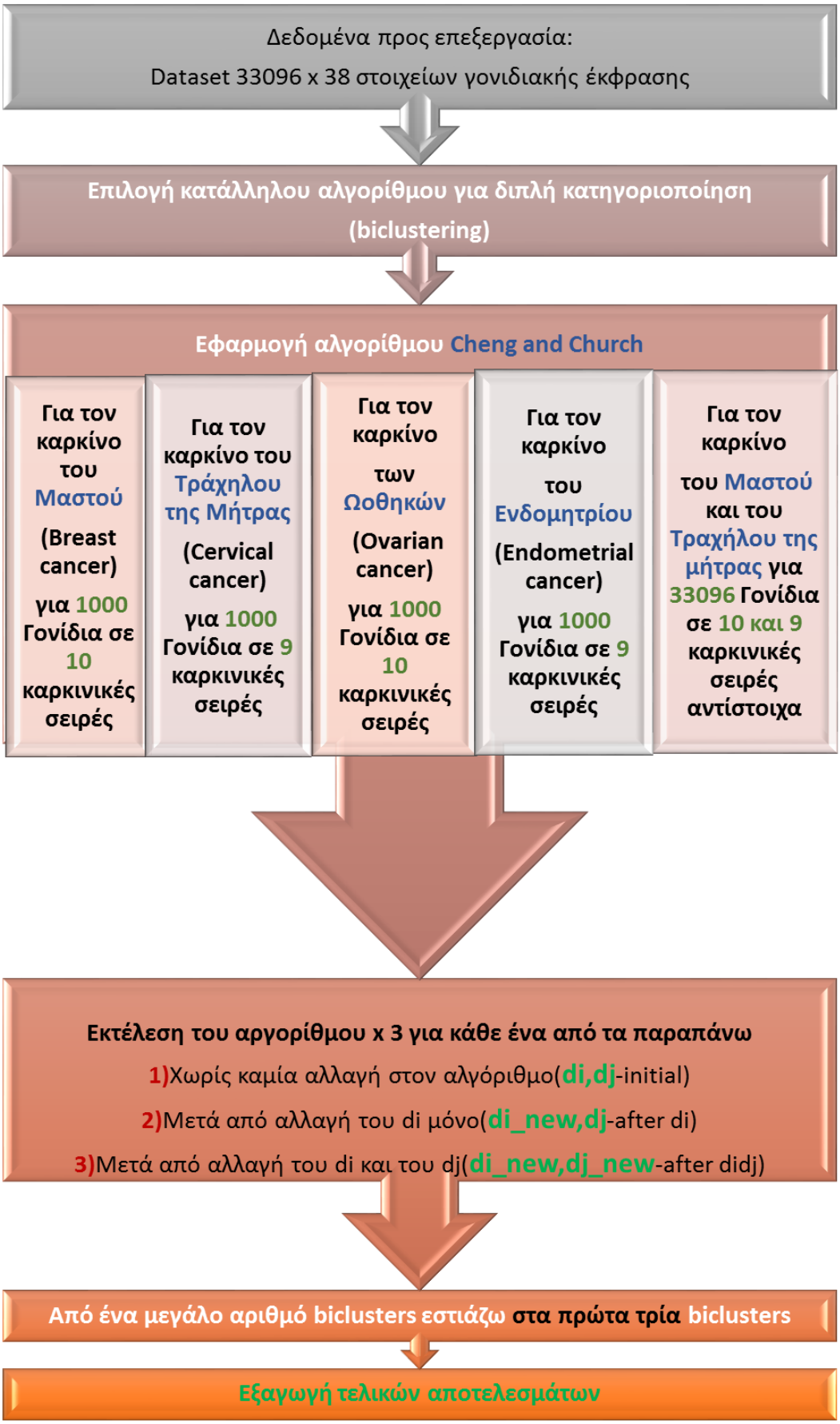
Χρειάστηκαν όλες οι παραπάνω εφαρμογές στον αλγόριθμο ώστε να επιλεγούν και να μελετηθούν τα τρία πρώτα Biclusters προκειμένου να καταλήξουμε στις πιο ενδεικτικές τιμές.

Όπως αναφέρθηκε, οι τύποι των καρκίνων από τους οποίους προέρχονται οι κυτταρικές σειρές είναι:

καρκίνος του ενδομητρίου (endometrial cancer), καρκίνος του τραχήλου της μήτρας (cervical cancer) και στον καθένα από αυτούς ανήκουν 10, 9, 9 και 10 σειρές αντίστοιχα.

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 3 πρώτων ομάδων διπλής κατηγοριοποίησης (Biclusters).

**Σχήμα 4.1.** - Σχεδιάγραμμα ροής



Σ’ αυτό το σημείο παρουσιάζεται ένας συγκεντρωτικός πίνακας με τον αριθμό των Biclusters που δημιουργούνται κατά την εκτέλεση του αλγορίθμου τόσο για τα 33096 γονίδια, όσο και για τα 1000 γονίδια για τα τρία πρώτα Biclusters κάθε φορά (Πίνακας 4.1).

**Πίνακας 4.1.** – Συγκεντρωτικά Αποτελέσματα Αριθμού Γονιδίων για Κάθε Ένα Από τα Τρία Biclusters που μελετήθηκαν

| ΚΑΡΚΙΝΟΣ ΤΟΥ ΜΑΣΤΟΥ(BREAST CANCER)-ΓΙΑ 33096 ΓΟΝΙΔΙΑ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 300 | 126 | 105 | 59 | 11 | 5 | 35 | 11 | 17 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΜΑΣΤΟΥ(BREAST CANCER)-ΓΙΑ 1000 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 30 | 18 | 27 | 4 | 6 | 2 | 9 | 7 | 5 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΩΝ ΩΟΘΗΚΩΝ(OVARIAN CANCER)-ΓΙΑ 33096 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 539 | 158 | 160 | 200 | 54 | 23 | 86 | 23 | 35 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΩΝ ΩΟΘΗΚΩΝ(OVARIAN CANCER)-ΓΙΑ 1000 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 26 | 29 | 15 | 3 | 4 | 2 | 10 | 9 | 8 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ(ENDOMETRIAL CANCER)-ΓΙΑ 33096 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 940 | 204 | 233 | 279 | 49 | 107 | 172 | 30 | 70 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΕΝΔΟΜΗΤΡΙΟΥ(ENDOMETRIAL CANCER)-ΓΙΑ 1000 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 42 | 22 | 23 | 8 | 2 | 2 | 16 | 8 | 5 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΡΑΧΗΛΟΥ(CERVICAL CANCER)-ΓΙΑ 33096 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 544 | 295 | 250 | 207 | 75 | 125 | 110 | 25 | 85 |
| **ΚΑΡΚΙΝΟΣ ΤΟΥ ΤΡΑΧΗΛΟΥ(CERVICAL CANCER)-ΓΙΑ 1000 ΓΟΝΙΔΙΑ** | | | | | | | | |
| ΑΡΧΙΚΗ ΚΑΤΑΣΤΑΣΗ(INITIAL) | | | ΑΛΛΑΓΗ DI(AFTER DI) | | | ΑΛΛΑΓΗ DI,DJ(AFTER DIDJ) | | |
| BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 | BICLUSTER1 | BICLUSTER2 | BICLUSTER3 |
| 36 | 24 | 16 | 12 | 4 | 4 | 13 | 11 | 5 |

**Παρατηρήσεις:**

Στον συγκεντρωτικό Πίνακα 4.1 παρατηρούμε ότι κατά την εκτέλεση του αλγορίθμου για το σύνολο των 33096 γονιδίων, ο όγκος των γονιδίων που ομαδοποιούνται σε ένα Bicluster είναι μεγάλος. Γι’ αυτό το λόγο, προκειμένου να διακρίνουμε και να εξάγουμε συμπεράσματα για τη συμπεριφορά των γονιδίων εκτελέσαμε τον αλγόριθμο για τα 1000 πρώτα γονίδια από τα 33096 και εστιάσαμε στη μελέτη των τριών πρώτων -στη σειρά- Biclusters. Στην ακόλουθη ενότητα αναφέρονται αναλυτικά τα προβλήματα που κληθήκαμε να αντιμετωπίσουμε καθώς και η προτεινόμενη μεθοδολογία για την επίλυσή τους.

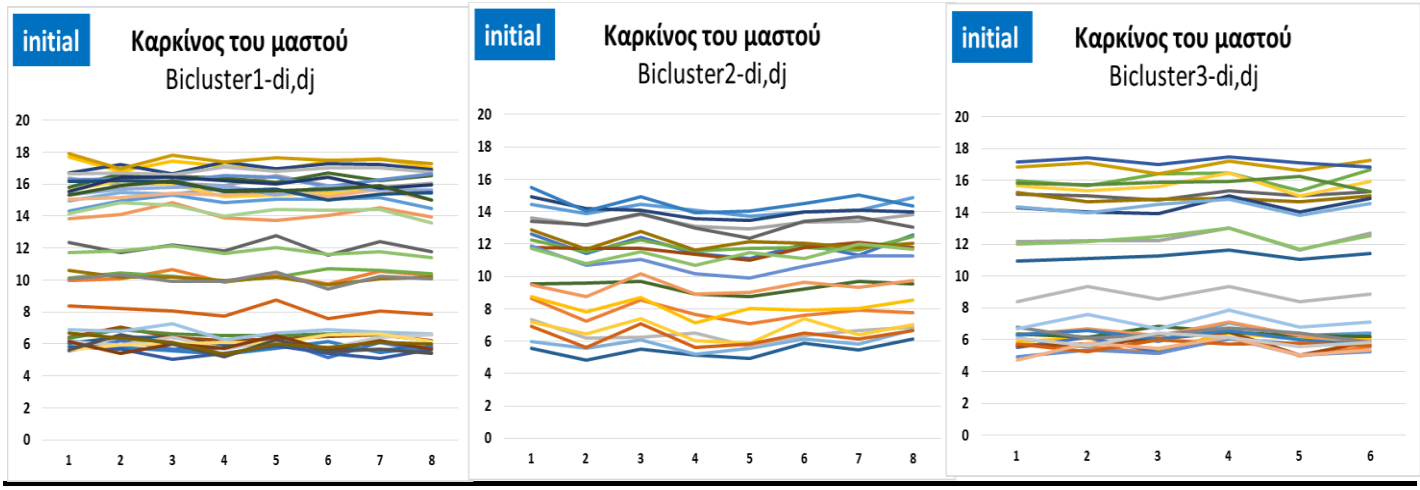
## 4.3. Προτεινόμενη Μεθοδολογία

Στην παρούσα διπλωματική εργασία, εφαρμόστηκε ο αλγόριθμος Cheng και Church, όπως αναλύθηκε στο προηγούμενο κεφάλαιο (με τις προτεινόμενες αλλαγές), τρεις φορές, για κάθε τύπο καρκίνου για συγκεκριμένες τιμές των παραμέτρων *δ* και α και με τροποποιήσεις στον υπολογισμό του *di* (mean squared residue της γραμμής *i*) και dj (mean squared residue της στήλης *j*).

Αρχικά, τρέξαμε τον αλγόριθμο Cheng και Church για τα **1000 πρώτα γονίδια** από τα 33096 **από τις κυτταρικές σειρές του** καρκίνου του μαστού (Breast cancer) **και εστιάσαμε στα τρία πρώτα Biclusters**.

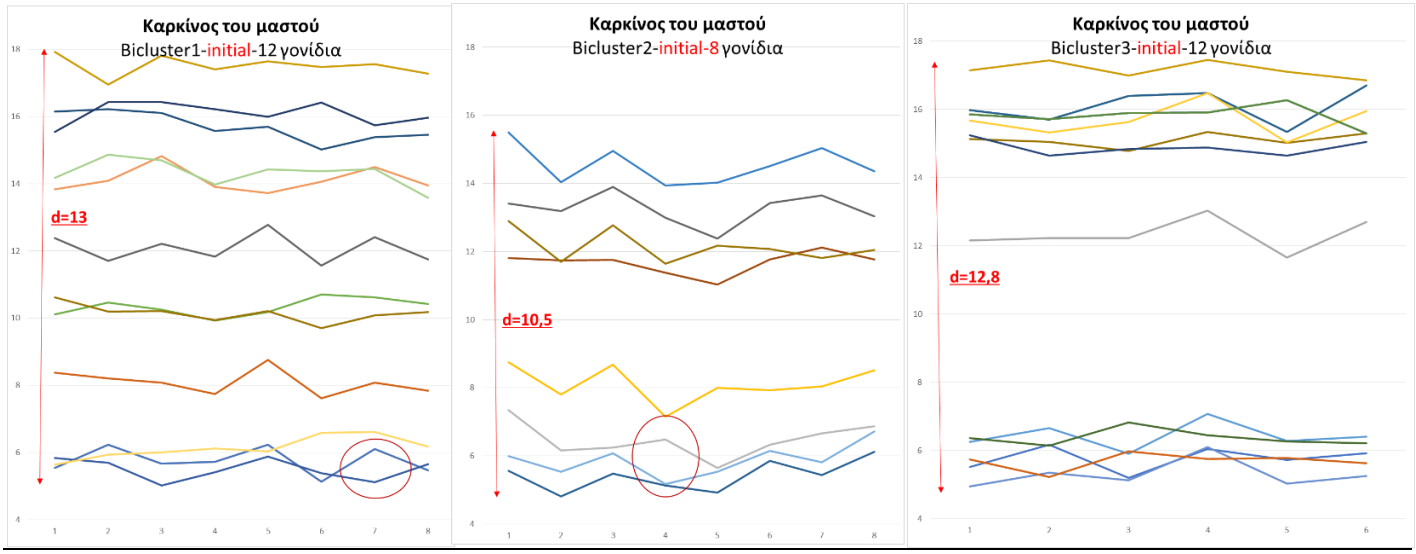
Οι παρακάτω Εικόνες (4.1.α, 4.1.β, 4.1.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 4.1.α) καθώς και δυο σημαντικά προβλήματα που προκύπτουν κατά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνες 4.1.β και 4.1.γ).

**Εικόνα 4.1.α.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di,dj*) για 1000 Γονίδια του Καρκίνου του Μαστού



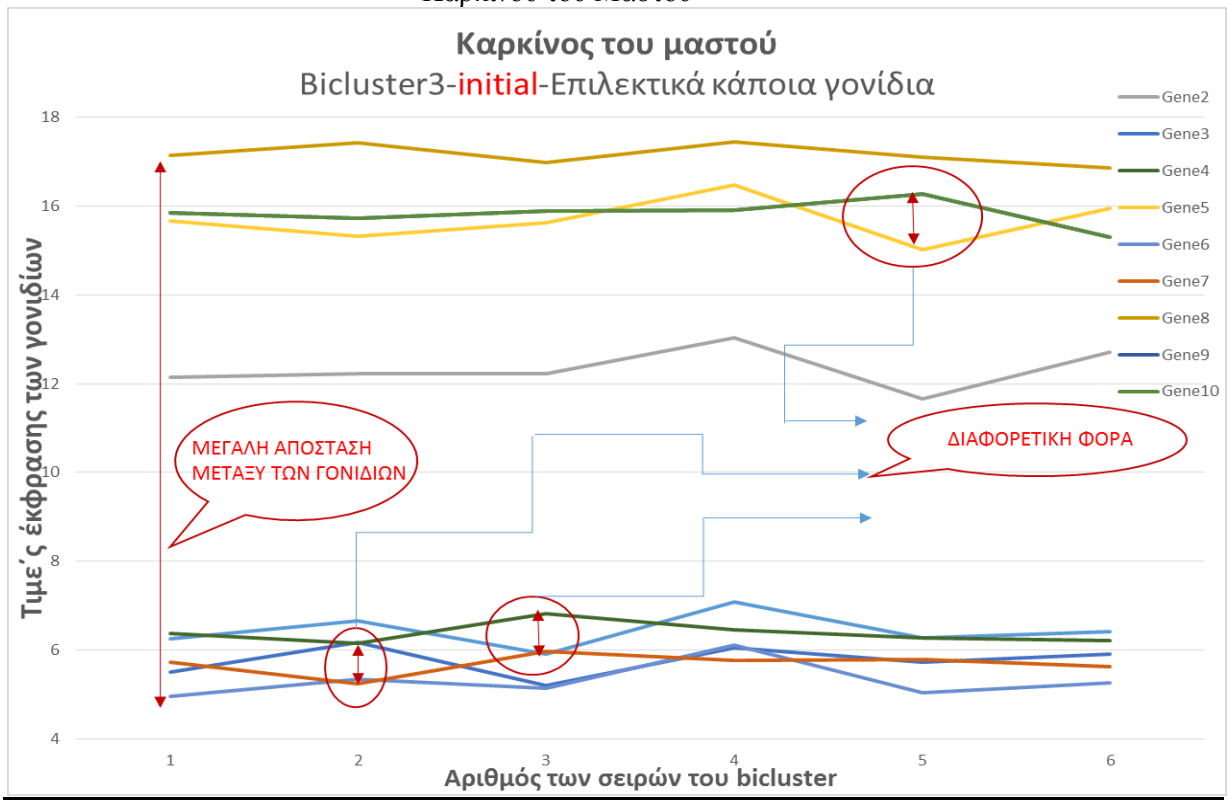

Στην Εικόνα 4.1.α παρατηρούμε το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου. Μικραίνοντας τα όρια του άξονα y (σε κλίμακα 4 έως 18 από 0 έως 20 που ήταν η αρχική) και αφαιρώντας ένα μεγάλο αριθμό γονιδίων (12 από 30 για το Bicluster 1, 8 από 18 για το Bicluster 2, και 12 από 27 για το Bicluster 3) από τα διαγράμματα, μπορούμε να διακρίνουμε καλύτερα δυο σημαντικά προβλήματα που προκύπτουν, όπως αποτυπώνονται στην Εικόνα 4.1.β.

**Εικόνα 4.1.β. -** Επιλεκτική Εμφάνιση Γονιδίων μετά την Πρώτη εκτέλεση του αλγορίθμου (*di,dj*) για 1000 Γονίδια του Καρκίνου του Μαστού




Στοχεύοντας στη μελέτη της συμπεριφοράς των γονιδίων, εστιάσαμε στην εμφάνιση 9 γονιδίων από το Bicluster 3 της Εικόνας 1.β., όπως αποτυπώνεται στην Εικόνα 4.1.γ.

**Εικόνα 4.1.γ. -** Εστιασμένη Εμφάνιση Γονιδίων μετά την Πρώτη εκτέλεση του αλγορίθμου (*di,dj*) για 1000 Γονίδια του Καρκίνου του Μαστού



Στην Εικόνα 4.1.γ διακρίνουμε τα εξής :

- Στα κυκλωμένα σημεία, τα οποία είναι ενδεικτικά**, παρατηρούμε διαφορετική τάση (different trend)** Αυτό υποδηλώνει ότι ο αλγόριθμος χρησιμοποιεί είτε απόλυτες τιμές είτε τετραγωνική ρίζα.
  - ✓ Πρέπει να αποφεύγονται κανονικοποιήσεις ή αλλαγές προσήμου.

Το δεύτερο πρόβλημα που διαφαίνεται στην Εικόνα 4.1.γ είναι **η μεγάλη απόσταση μεταξύ των καμπυλών του διαγράμματος**.

- Ο αλγόριθμος σε κάποιο σημείο χρησιμοποιεί $x_i$, $y_i$ όπου $i = 1, \dots, 10$

  Και $\text{correlation}(x, y) = \frac{<x_i, y>}{|x| * |y|}$. Ο αλγόριθμος δηλαδή αφαιρεί το μέτρο και διαιρεί με $|x| * |y|$.
  - ✓ Πρέπει λοιπόν να αποφεύγονται κανονικοποιήσεις του μέτρου.

- Αντιμετώπιση "προβλημάτων"

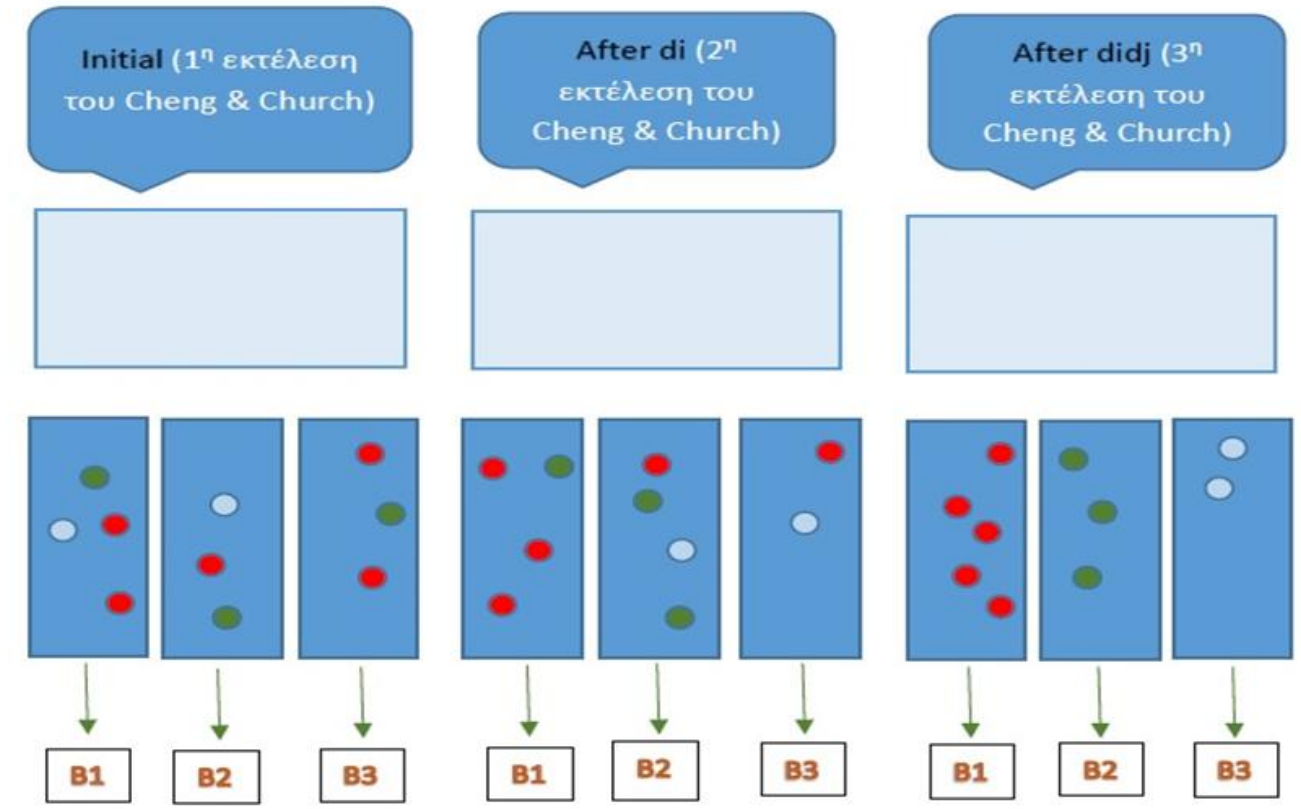
Με βάση τα παραπάνω προβλήματα, προκειμένου να εξαχθούν ταυτόχρονα ομάδες γονιδίων με ομοιόμορφη συμπεριφορά, πραγματοποιήθηκε η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης δεδομένων γονιδιακής έκφρασης πέραν της αρχικής μορφής του αλγορίθμου και με προτεινόμενες τροποποιήσεις.

Ο αλγόριθμος εφαρμόστηκε, επομένως, τρεις φορές για κάθε τύπο καρκίνου για συγκεκριμένες τιμές των δ και α και με τροποποιήσεις στον υπολογισμό του *di* (mean squared της γραμμής i) και *dj* (mean squared της στήλης *j*).

Όπως αναφέρθηκε και παραπάνω τα δεδομένα γονιδιακής έκφρασης αναπαρίστανται με τη μορφή πινάκων, όπου οι γραμμές αποτελούν τα γονίδια και οι στήλες τις διάφορες πειραματικές συνθήκες ή καταστάσεις (π.χ. διαφορετικές κυτταρικές σειρές). Η ανάλυση των πινάκων τοποθετείτε τοποθετείται σε δύο βασικές κατευθύνσεις. Στην ομαδοποίηση των γονιδίων ή καταστάσεων με βάση την πλειοψηφία των περιπτώσεων παθολογίας και στην κατηγοριοποίηση και πρόβλεψη νέων γονιδίων ή δειγμάτων στηριζόμενοι σε ήδη γνωστή βιολογική πληροφορία. Στο παρακάνω Σχήμα 4.2 φαίνεται η μετακίνηση των γονιδίων από το ένα Bicluster στο άλλο κατά την διαδικασία αλλαγής του *di* και *dj* με στόχο την καλύτερη ομαδοποίηση γονιδίων ως προς το μέγεθος έκφρασης και τάσης.

Στο Σχήμα 4.2 παρουσιάζεται με έναν απλό τρόπο τι πετυχαίνουμε με το να πραγματοποιούμε αλλαγές στο *di* και στο *dj*.

**Σχήμα 4.2.** - Σχηματική Αναπαράσταση της Συμπεριφοράς των Γονιδίων στις Τρεις Εκτελέσεις του Αλγορίθμου Cheng και Church




Με B1, B2, B3 αναφερόμαστε στο Bicluster1, Bicluster2, Bicluster3 αντίστοιχα. Ενδεικτικά, όπως αποτυπώνεται στο Σχήμα 4.2, τα γονίδια με κόκκινο χρώμα εμφανίζονται σε όλες τις ομάδες. Μετά την αλλαγή του d*i* έχει φύγει ένα «κόκκινο» γονίδιο και έχει τοποθετηθεί στο Bicluster1. Τέλος, μετά την αλλαγή και του d*j* φαίνεται πως όλα τα «κόκκινα» γονίδια έχουν ομαδοποιηθεί στο Bicluster1. Αντίστοιχη συμπεριφορά εμφανίζουν και τα «πράσινα» και τα «γαλάζια» γονίδια. Έτσι, μετά το τέλος των τριών εκτελέσεων του αλγορίθμου Cheng και Church παρατηρούμε πως έχουμε πετύχει την ομαδοποίηση γονιδίων με παρόμοια χαρακτηριστικά.

Πιο αναλυτικά, εκτελέστηκε ο αλγόριθμος Cheng και Church για κάθε καρκίνο στην κανονική του μορφή. Στη συνέχεια εκτελέστηκε ο αλγόριθμος C&C για τον κάθε τύπο καρκίνου με αλλαγή στον τρόπο υπολογισμού του d*i*.

Αντί για τον αρχικό τύπο του d*i* :

$$\mathrm{d}(i) = \frac{1}{|J|}\sum_{j\in J}\left(\boldsymbol{a}_{\mathrm{ij}} - \boldsymbol{a}_{\mathrm{iJ}} - \boldsymbol{a}_{\mathrm{Ij}} + \boldsymbol{a}_{\mathrm{IJ}}\right)^2$$

Υπολογίζει μια μέση απόσταση του γονιδίου αυτού, σε όλες τις καταστάσεις, και επιχειρεί να κρατήσει τη μέση αυτή απόσταση όσο το δυνατόν μικρότερη. Επομένως το νέο d*i* ορίζεται ως :

2η εκτέλεση
Εκτέλεση ανά γονίδιο

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

**Το d(*i*) αναφέρεται στην ΤΑΣΗ των γονιδίων. Με τον όρο $|a_{ij} - a_{iJ}|^2$ τα γονίδια, σε όλες τις χρονικές στιγμές, δεν κινούνται πολύ γρήγορα γύρω από το Μέσο Όρο του ίδιου του γονιδίου, δηλαδή δεν έχουμε σημαντικές αλλαγές. Με τον όρο$|a_{ij} - a_{Ij}|^2$ τα γονίδια μέσα στο Bicluster, σε όλες τις χρονικές στιγμές, συμπεριφέρονται σαν τον Μέσο Όρο του συνόλου. Ουσιαστικά, ο αλγόριθμος κρατάει γονίδια που έχουν ίδια ή παρόμοια τάση.**

Στη συνέχεια εκτελέστηκε ο αλγόριθμος Cheng και Church για τον κάθε τύπο καρκίνου με αλλαγή στον τρόπο υπολογισμού του d*j*.

Αντί για τον αρχικό τύπο του d*j* :

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

Για κάθε κατάσταση ή κυτταρική σειρά j, το κριτήριο d*j* επιχειρεί να κρατήσει όλα τα γονίδια του Bicluster σε μικρή απόσταση μεταξύ τους. Επομένως το νέο d*j* ορίζεται ως :

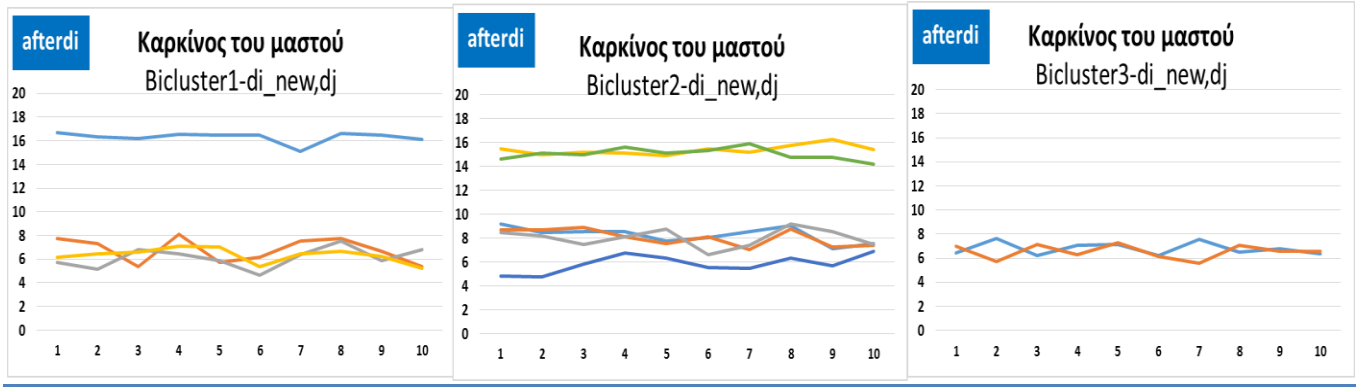
3η εκτέλεση
Εκτέλεση ανά κατάσταση

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

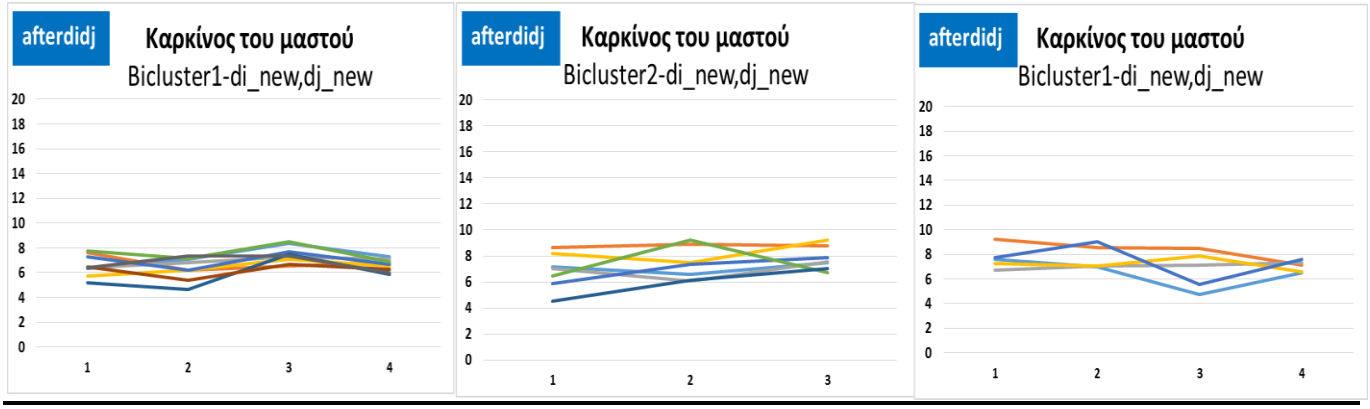
**Το d(*j*) αναφέρεται στην ΤΙΜΗ των γονιδίων. Με τον όρο $|a_{ij} - a_{iJ}|^2$ τα γονίδια, σε κάθε χρονική στιγμή, βρίσκονται κοντά στο Μέσο Όρο του ίδιου του γονιδίου. Με τον όρο$|a_{ij} - a_{Ij}|^2$ τα γονίδια μέσα στο Bicluster, σε κάθε χρονική στιγμή, είναι μαζεμένα σε ένα Μέσο Όρο. Ουσιαστικά, ο αλγόριθμος κρατάει καταστάσεις στις οποίες τα γονίδια έχουν όμοια συμπεριφορά γύρω από ένα γενικό Μέσο Όρο.**

Οι Εικόνες 4.2 και 4.3 αναδεικνύουν τις βελτιώσεις της συμπεριφοράς των γονιδίων σε κάθε Bicluster μετά την εφαρμογή των παραπάνω τροποποιήσεων παρουσιάζονται τα αποτελέσματα των γονιδίων μετά την δεύτερη και στη συνέχεια τη τρίτη εκτέλεση του αλγορίθμου.

**Εικόνα 4.2.** - Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*) για 1000 Γονίδια του Καρκίνου του Μαστού




**Εικόνα 4.3.** - Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 1000 Γονίδια του Καρκίνου του Μαστού




Στο Κεφάλαιο 5 παρουσιάζονται αναλυτικά τα αποτελέσματα της εργασίας μας για όλες τις κυτταρικές σειρές των τεσσάρων καρκινικών τύπων που μελετήσαμε.

# 5

## Αποτελέσματα

Παρακάτω παρουσιάζεται η σειρά με την οποία εξήχθησαν τα αποτελέσματα.

**1.** Αρχικά, τρέξαμε τον αλγόριθμο Cheng και Church για τα 1000 πρώτα γονίδια από τα 33096 για τους παρακάτω τύπους καρκίνου:

α. Καρκίνος του μαστού (Breast cancer)

β. Καρκίνος του τραχήλου της μήτρας (Cervical cancer)

γ. Καρκίνος των ωοθηκών (Ovarian cancer)

δ. Καρκίνος του ενδομητρίου (Endometrial cancer)

Στους τέσσερις αυτούς τύπους καρκίνου, μειώνοντας τον αριθμό των γονιδίων από 33096 σε 1000, μπορέσαμε να διακρίνουμε καλύτερα τη συμπεριφορά των γονιδίων αναφορικά με την τάση και την τιμή τους και να προβούμε σε κατάλληλες αλλαγές στον αλγόριθμο.

Τα αποτελέσματα παρουσιάζονται στη συνέχεια της εργασίας με την εξής σειρά:

- **1.α. Για τον καρκίνο του μαστού για 1000 Γονίδια, για τα γονίδια που προκύπτουν μετά την:**

  α) πρώτη εκτέλεση του αλγορίθμου (*di*,*dj* ή initial),

  β) δεύτερη εκτέλεση του αλγορίθμου (*di*_new,*dj* ή after *di*) και

  γ) τρίτη εκτέλεση του αλγορίθμου (*di*_new,*dj*_new ή after *di*_*dj*)

- **1.β. Για τον καρκίνο του τραχήλου της μήτρας για 1000 Γονίδια που προκύπτουν μετά την:**

  α) πρώτη εκτέλεση του αλγορίθμου (*di*,*dj* ή initial),

  β) δεύτερη εκτέλεση του αλγορίθμου (*di*_new,*dj* ή after *di*) και

  γ) τρίτη εκτέλεση του αλγορίθμου (*di*_new,*dj*_new ή after *di*_*dj*)

- **1.γ. Για τον καρκίνο των ωοθηκών για 1000 Γονίδια, για τα γονίδια που προκύπτουν μετά την:**

  Όλα τα γονίδια που προκύπτουν μετά την:

  α) πρώτη εκτέλεση του αλγορίθμου (*di*,*dj* ή initial),

  β) δεύτερη εκτέλεση του αλγορίθμου (*di*_new,*dj* ή after *di*) και

  γ) τρίτη εκτέλεση του αλγορίθμου (*di*_new,*dj*_new ή after *di*_*dj*)

❖ **1.δ. Για τον καρκίνο του ενδομητρίου για 1000 Γονίδια, για τα γονίδια που προκύπτουν μετά την:**

α) πρώτη εκτέλεση του αλγορίθμου (*di*,*dj* ή initial),

β) δεύτερη εκτέλεση του αλγορίθμου (*di*_new,*dj* ή after *di*) και

γ) τρίτη εκτέλεση του αλγορίθμου (*di*_new,*dj*_new ή after *di*_*dj*)

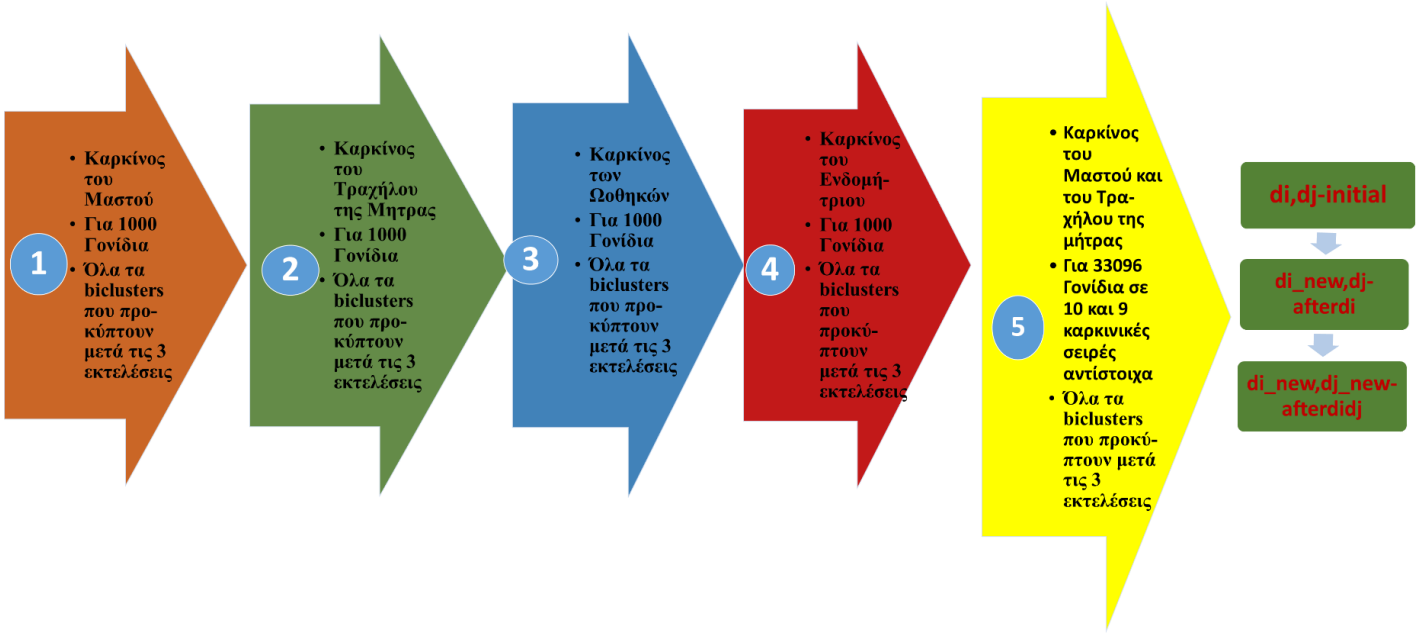
**2.** Για τον καρκίνο του μαστού και τον καρκίνο του τραχήλου της μήτρας, για 33096 Γονίδια, για τα γονίδια που προκύπτουν μετά την:

α) πρώτη εκτέλεση του αλγορίθμου (*di*,*dj* ή initial),

β) δεύτερη εκτέλεση του αλγορίθμου (*di*_new,*dj* ή after di) και

γ) τρίτη εκτέλεση του αλγορίθμου (*di*_new,*dj*_new ή after *di*_*dj*)

**Πίνακας 5.1.** – Συγκεντρωτικός πίνακας αποτύπωσης προτεινόμενης μεθοδολογίας

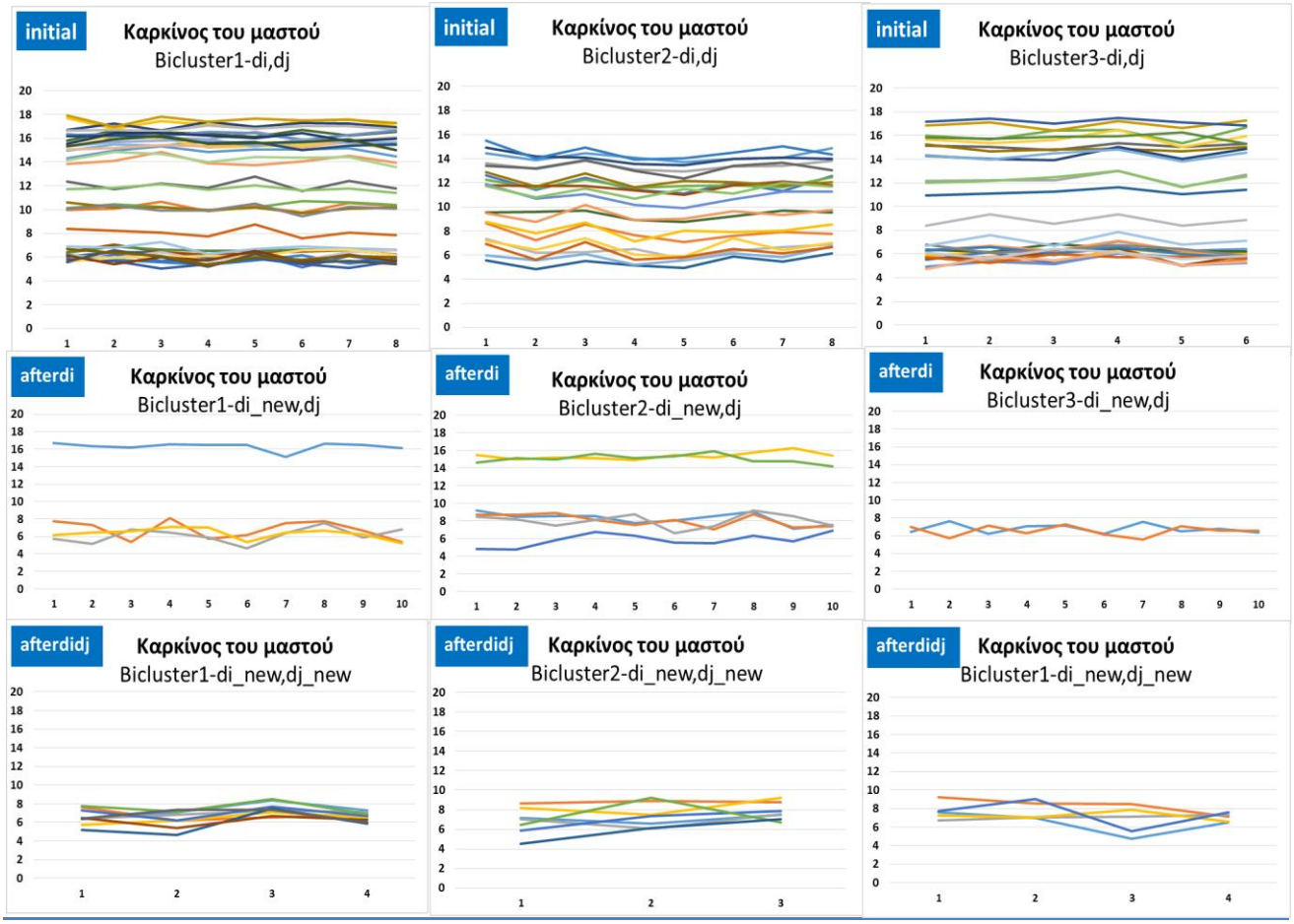



Στον Πίνακα 5.1 απεικονίζεται η σειρά με την οποία θα παρουσιαστούν τα αποτελέσματα στην ακόλουθη ενότητα.

## 5.1. Αποτελέσματα biclustering από κυτταρικές σειρές καρκίνου του μαστού

Η παρακάτω συγκεντρωτική Εικόνα 5.1 παρουσιάζει το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους για τον καρκίνο του μαστού για τα 1000 πρώτα γονίδια, σε κάθε ένα από τα τρία πρώτα Biclusters.

**Εικόνα 5.1. -** Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 1000 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Μαστού

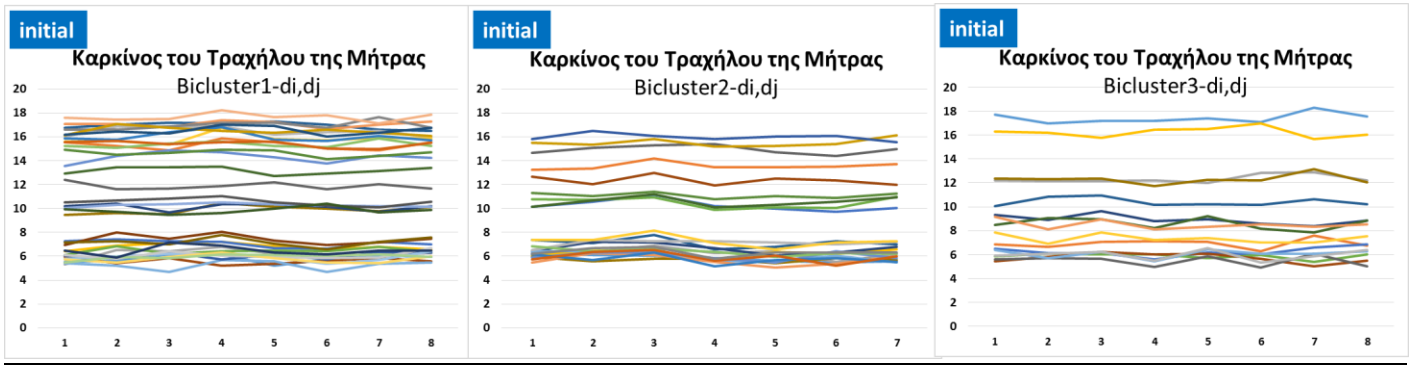



## 5.2. Αποτελέσματα biclustering από κυτταρικές σειρές του καρκίνου του τραχήλου της μήτρας

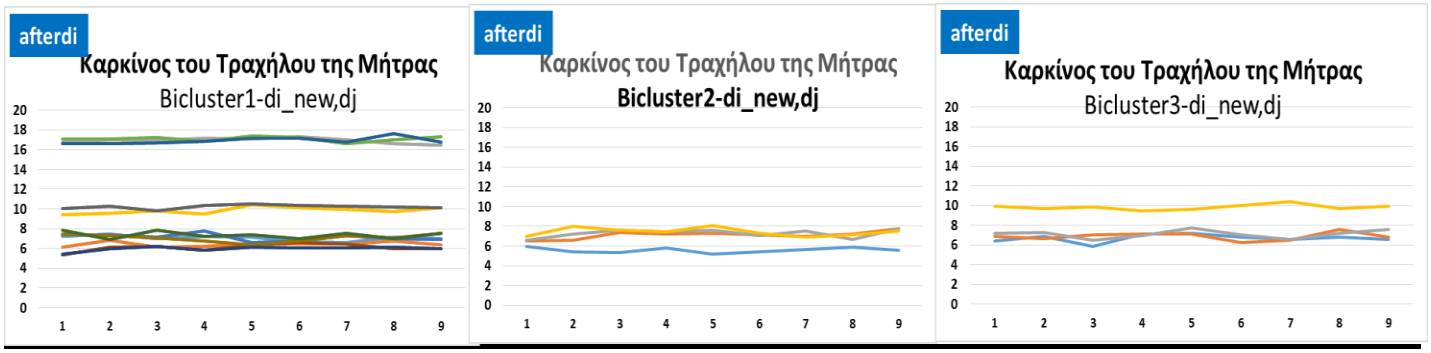
**Στη συνέχεια, τρέξαμε τον αλγόριθμο Cheng και Church** για τις κυτταρικές σειρές του **καρκίνου του τραχήλου της μήτρας, για τα** 1000 πρώτα γονίδια_**από τα 33096** και εστιάσαμε στα τρία πρώτα **B**iclusters (Cervical cancer).

Οι παρακάτω Εικόνες (5.2.α, 5.2.β, 5.2.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 5.2.α), μετά τη δεύτερη εκτέλεση του αλγορίθμου (Εικόνες 5.2.β) και τέλος, μετά την Τρίτη εκτέλεση του αλγορίθμου (Εικόνες 5.2.γ).

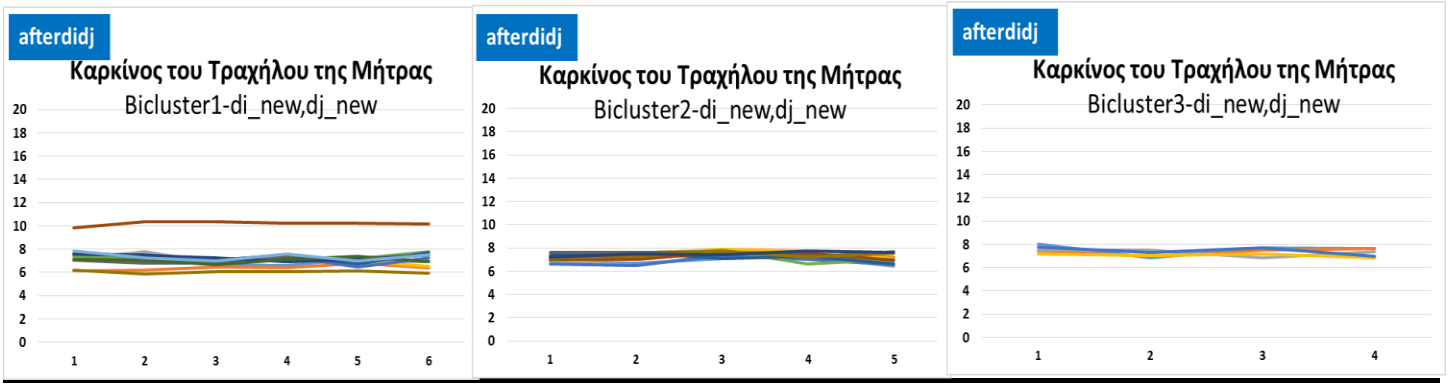
**Εικόνα 5.2.α. -** Πρώτη Εκτέλεση του Αλγορίθμου (*di,dj*) για 1000 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας



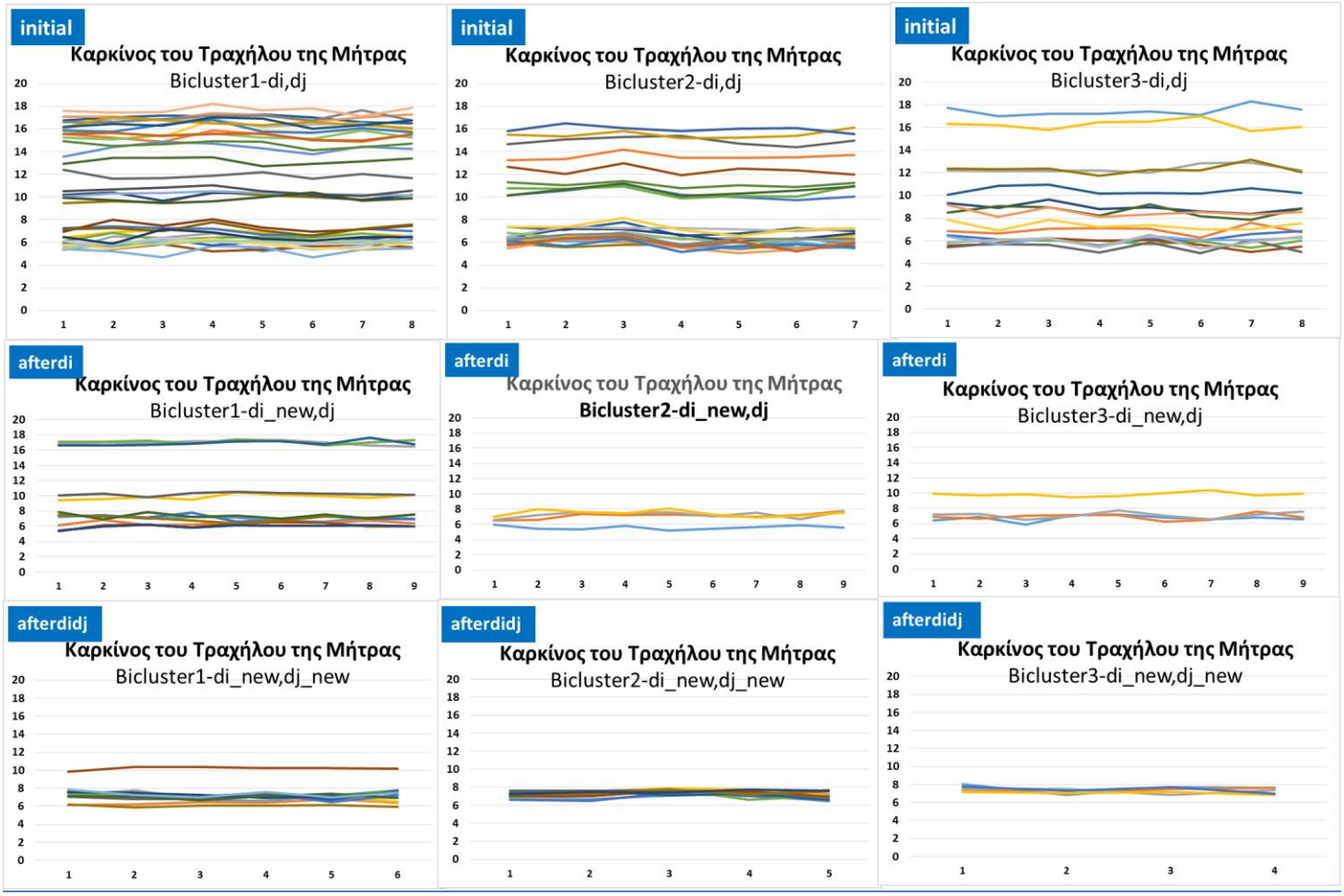

**Εικόνα 5.2.β. -** Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*) για 1000 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας




**Εικόνα 5.2.γ. -** Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 1000 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας



**Εικόνα 5.3. -** Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 1000 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Τραχήλου της Μήτρας

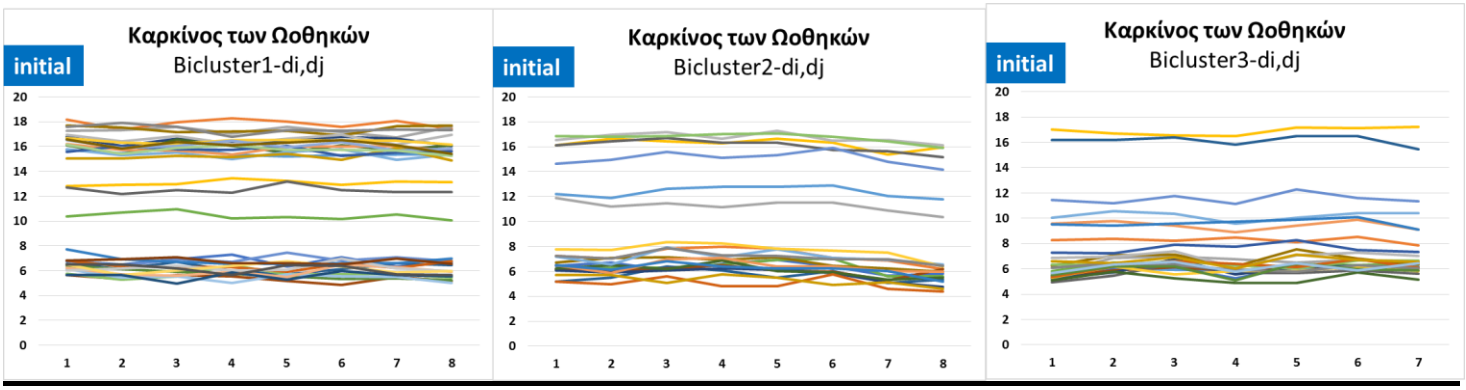



## 5.3. Αποτελέσματα biclustering από κυτταρικές σειρές καρκίνου των Ωοθηκών

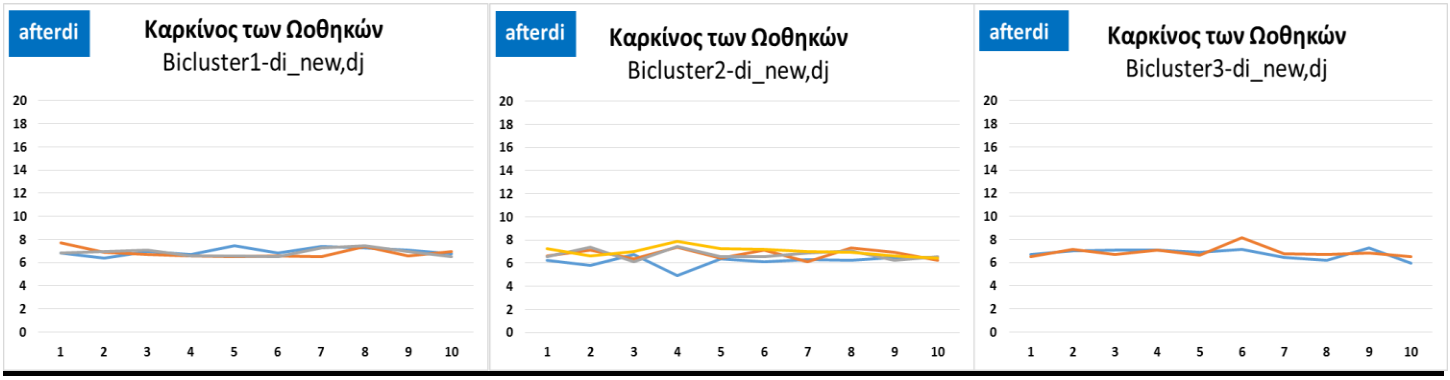
**Στη συνέχεια, τρέξαμε τον αλγόριθμο Cheng και Church** για τις κυτταρικές σειρές του **καρκίνου των ωοθηκών, για τα** 1000 πρώτα γονίδια **από τα 33096,** και εστιάσαμε στα τρία πρώτα Biclusters.

Οι παρακάτω Εικόνες (5.3.α, 5.3.β, 5.23.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 5.3.α), μετά τη δεύτερη εκτέλεση του αλγορίθμου (Εικόνες 5.3.β) και τέλος, μετά την Τρίτη εκτέλεση του αλγορίθμου (Εικόνες 5.3.γ).

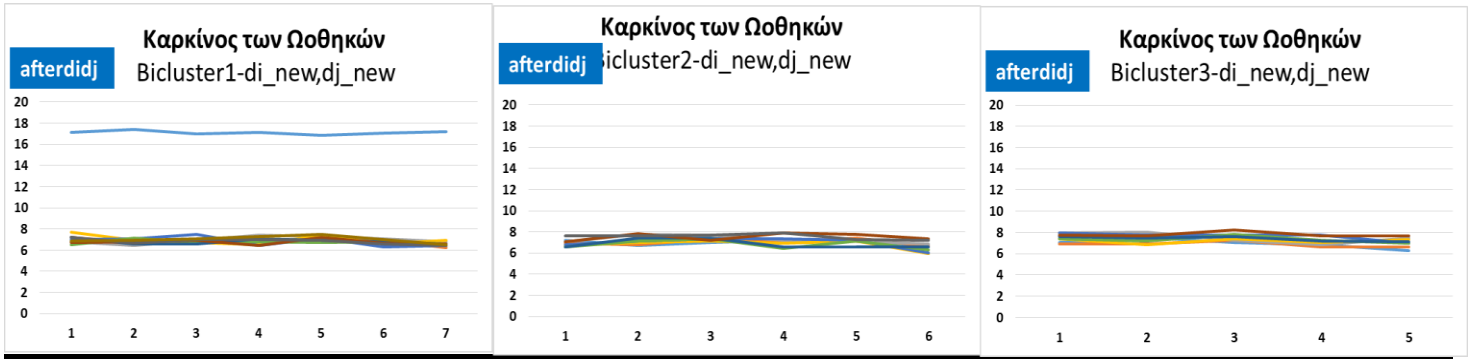
**Εικόνα 5.3.α. -** Πρώτη Εκτέλεση του Αλγορίθμου (*di*,*dj*) για 1000 Γονίδια του Καρκίνου των Ωοθηκών



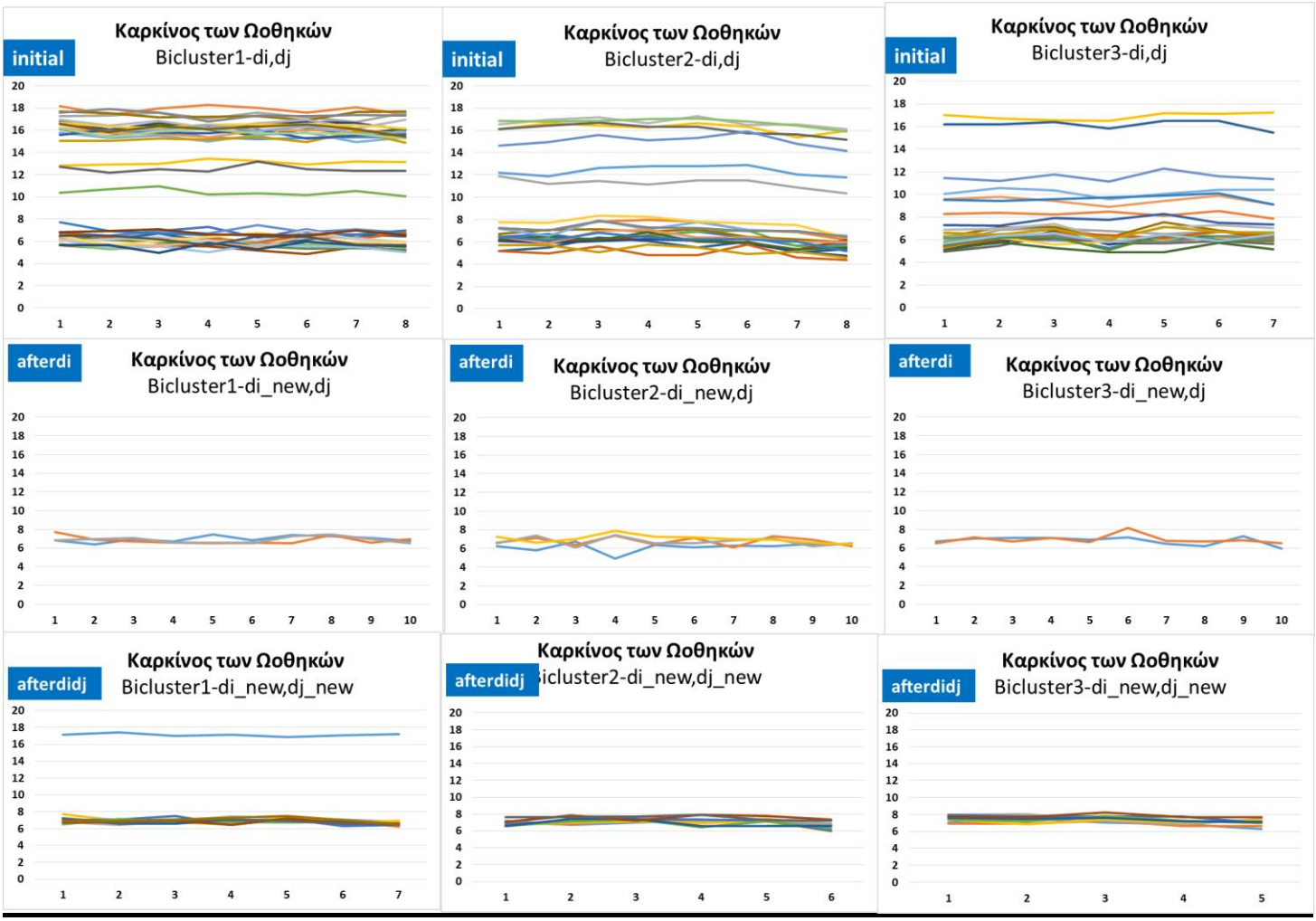

**Εικόνα 5.3.β. -** Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*) για 1000 Γονίδια του Καρκίνου των Ωοθηκών




**Εικόνα 5.3.γ. -** Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 1000 Γονίδια του Καρκίνου των Ωοθηκών



**Εικόνα 5.4.** - Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 1000 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου των Ωοθηκών

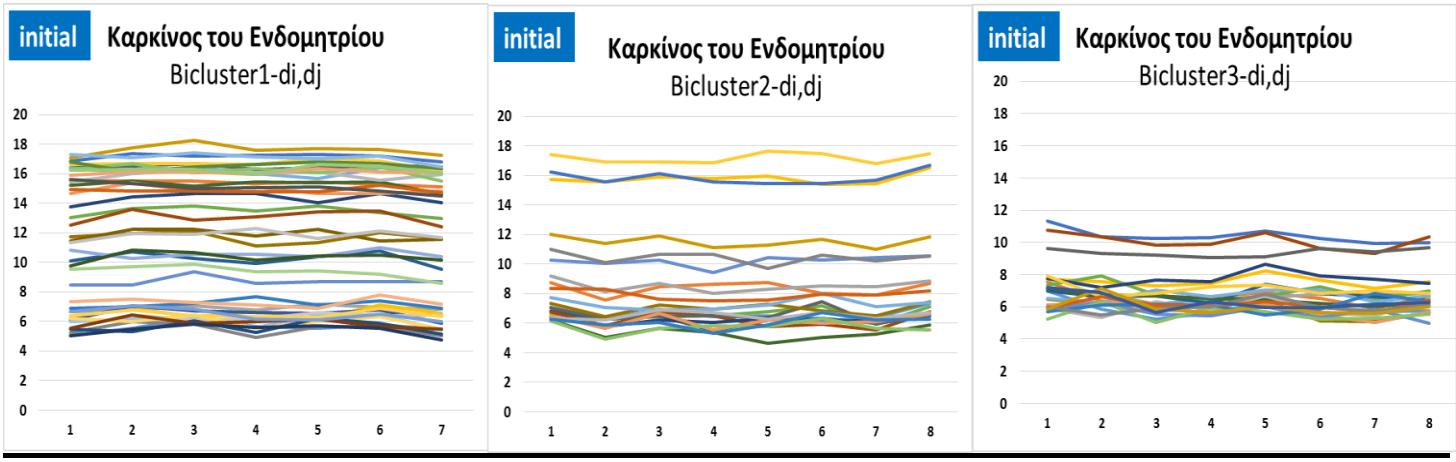



## 5.4. Αποτελέσματα biclustering από κυτταρικές σειρές καρκίνου του Ενδομητρίου

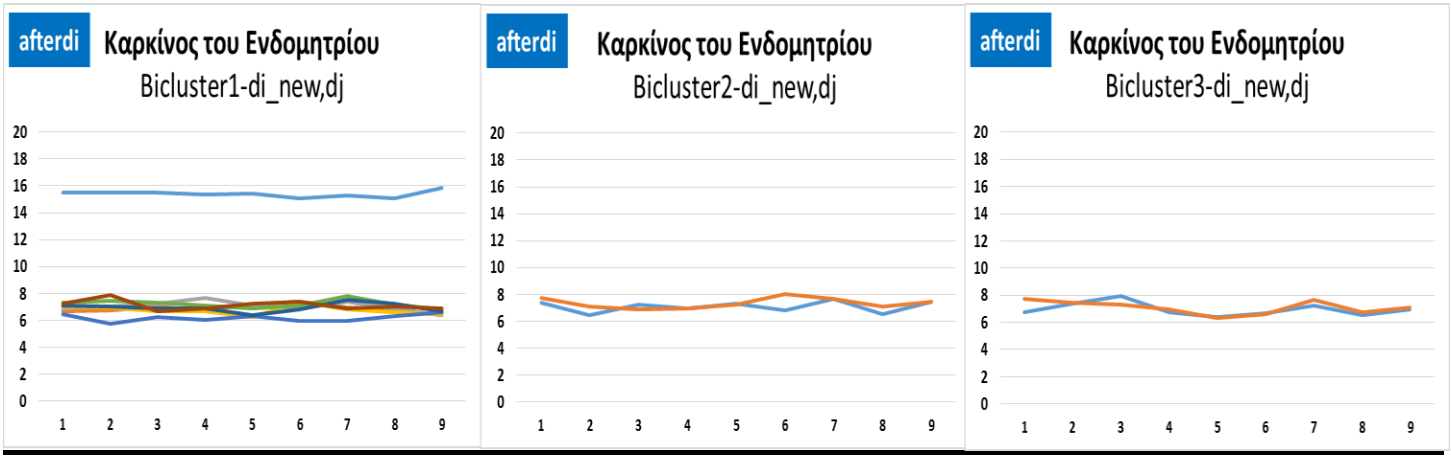
**Στη συνέχεια, τρέξαμε τον αλγόριθμο Cheng και Church** για τις κυτταρικές σειρές του **καρκίνου του ενδομητρίου, για τα** 1000 πρώτα γονίδια **από τα 33096** και εστιάσαμε στα τρία πρώτα Biclusters.

Οι παρακάτω Εικόνες (5.4.α, 5.4.β, 5.4.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 5.4.α), μετά τη δεύτερη εκτέλεση του αλγορίθμου (Εικόνες 5.4.β) και τέλος, μετά την Τρίτη εκτέλεση του αλγορίθμου (Εικόνες 5.4.γ).

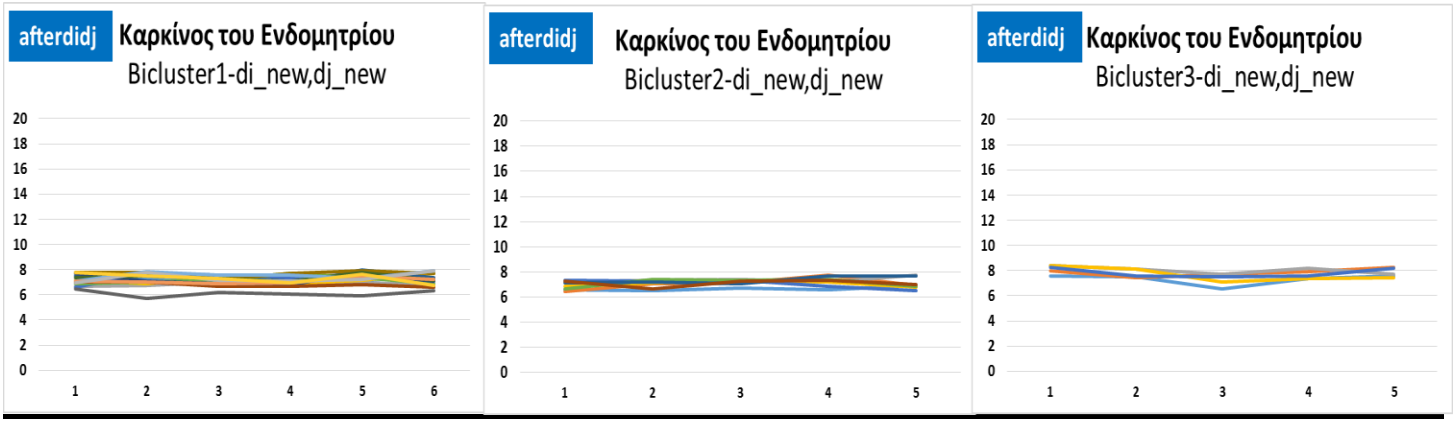
**Εικόνα 5.4.α.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di*,*dj*) για 1000 Γονίδια του Καρκίνου του Ενδομητρίου



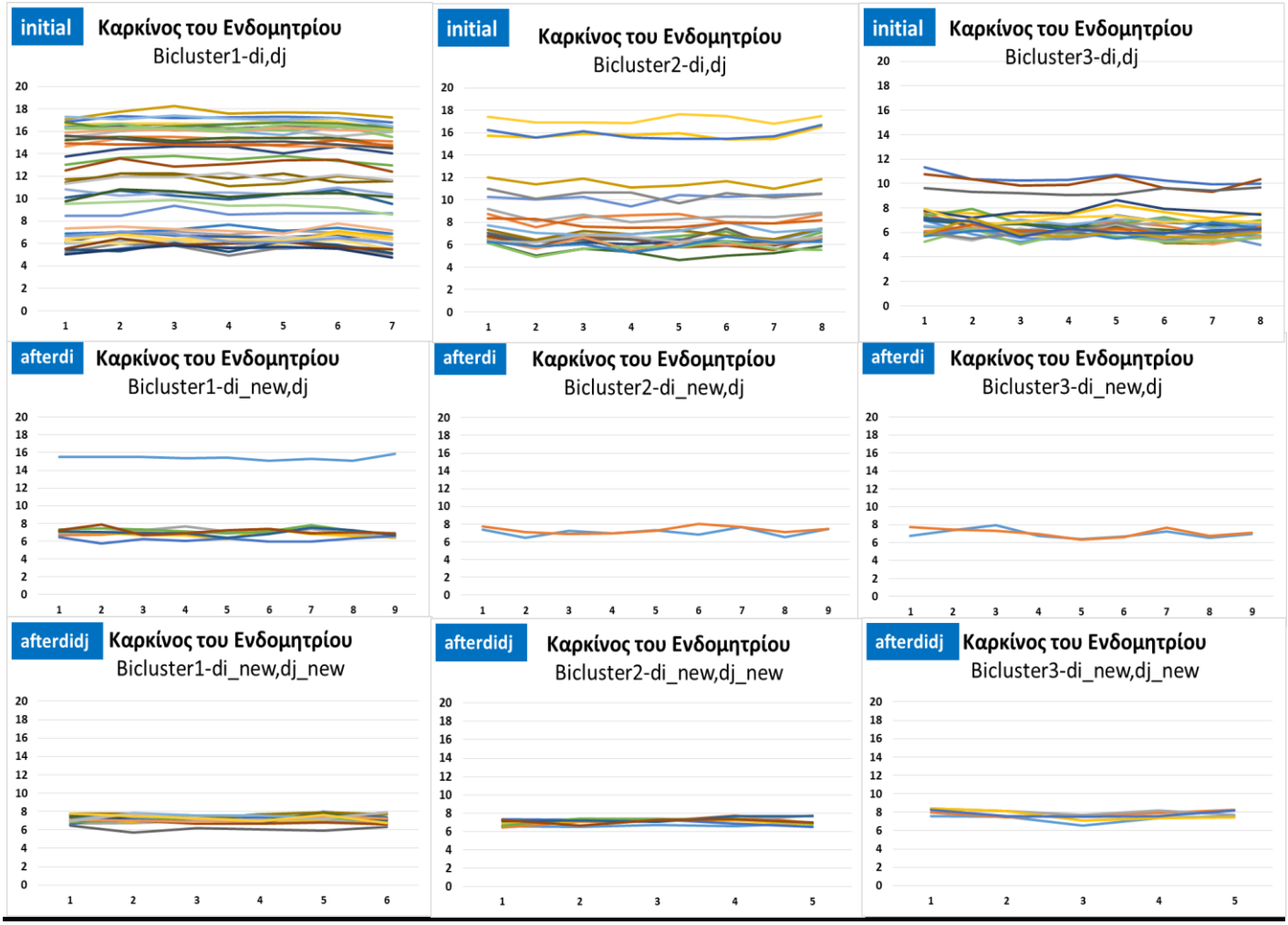

**Εικόνα 5.4.β.** - Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,d*j*) για 1000 Γονίδια του Καρκίνου του Ενδομητρίου




**Εικόνα 5.4.γ.** - Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 1000 Γονίδια του Καρκίνου του Ενδομητρίου



**Εικόνα 5.5.** - Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 1000 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Ενδομητρίου

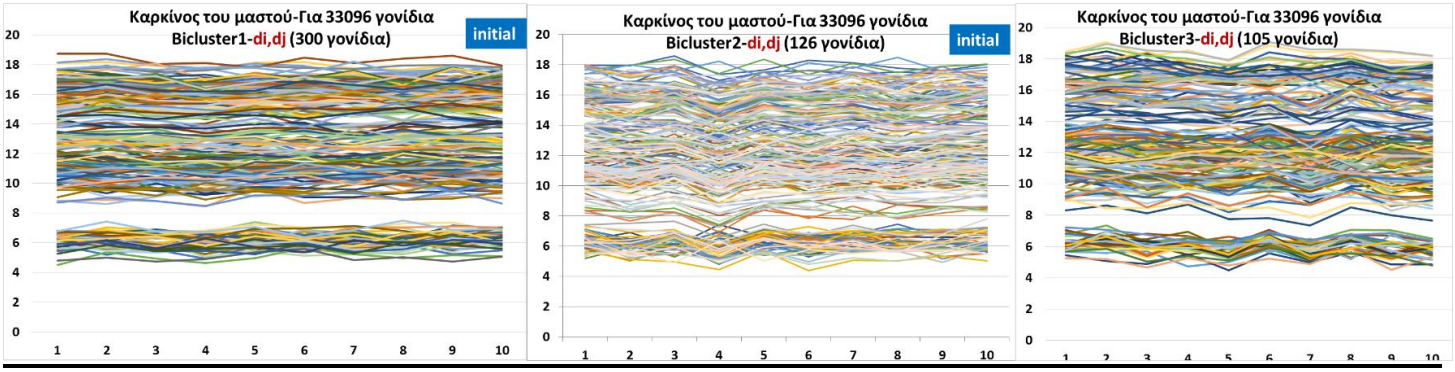



## 5.5. Αποτελέσματα Biclustering από Κυτταρικές Σειρές Καρκίνου του Μαστού για 33096 Γονίδια

Όπως έχει ειπωθεί και σε προηγούμενο κεφάλαιο, παρατηρήσαμε ότι κατά την εκτέλεση του αλγορίθμου για το σύνολο των 33096 γονιδίων, ο αριθμός τόσο των γονιδίων που ομαδοποιούνται σε ένα Bicluster, όσο και των Biclusters, είναι μεγάλος. Γι' αυτό το λόγο, προκειμένου να διακρίνουμε και να εξάγουμε συμπεράσματα για τη συμπεριφορά των γονιδίων εκτελέσαμε τον αλγόριθμο για τα 1000 πρώτα γονίδια από τα 33096 (Ενότητα 5.4.).

Σε αυτό το σημείο, τρέξαμε τον αλγόριθμο Cheng και Church για τα 33096 για τις κυτταρικές σειρές καρκίνου του μαστού, και εστιάσαμε στα τρία πρώτα Biclusters. Οι παρακάτω Εικόνες (5.5.α, 5.5.β, 5.5.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 5.8.α), μετά τη δεύτερη εκτέλεση του αλγορίθμου (Εικόνες 5.8.β) και τέλος, μετά την Τρίτη εκτέλεση του αλγορίθμου (Εικόνες 5.5.γ).

**Εικόνα 5.5.α.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di,dj*) για 33096 Γονίδια του Καρκίνου του Μαστού

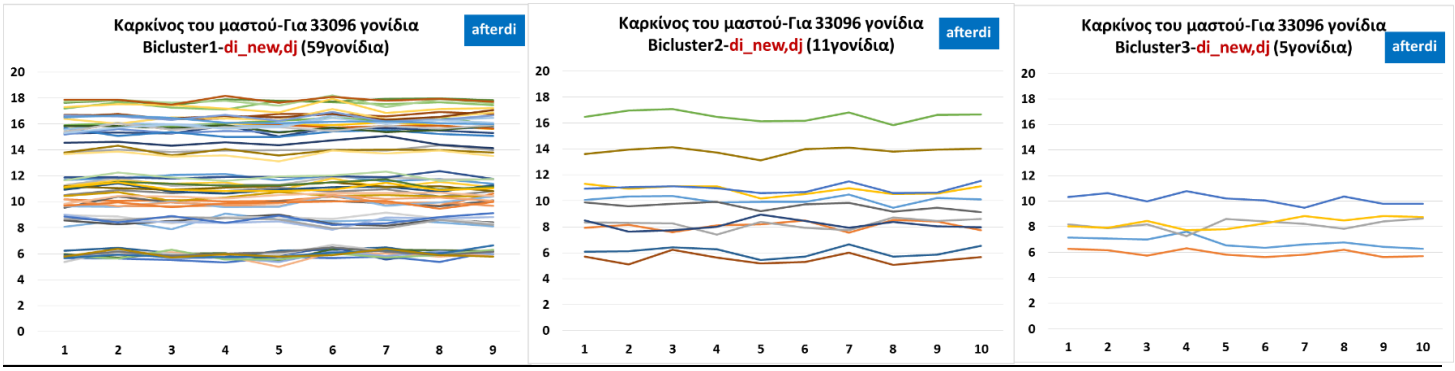



Αριστερό τμήμα: Το πρώτο διάγραμμα απεικονίζει 300 γονίδια. Λόγω του πλήθους των γονιδίων (300) που προκύπτουν στο Bicluster1 κατά την πρώτη εκτέλεση του αλγορίθμου. Στην εικόνα εμφανίζουμε MONO τα 255 πρώτα γονίδια, εφόσον το πρόγραμμα του excel επιτρέπει την εμφάνιση έως 255 στοιχείων (Εικόνα 5.5.α - αριστερό τμήμα)

Μεσαίο τμήμα: Το δεύτερο διάγραμμα απεικονίζει 126 γονίδια. Εμφάνιση των 126 γονιδίων που προκύπτουν στο δεύτερο Bicluster κατά την πρώτη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.α - μεσαίο τμήμα)

Δεξιό τμήμα: Το τρίτο διάγραμμα απεικονίζει 105 γονίδια. Εμφάνιση των 105 γονιδίων που προκύπτουν στο τρίτο Bicluster κατά την πρώτη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.α - δεξί τμήμα)

**Εικόνα 5.5.β.** - Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj)* για 33096 Γονίδια του Καρκίνου του Μαστού

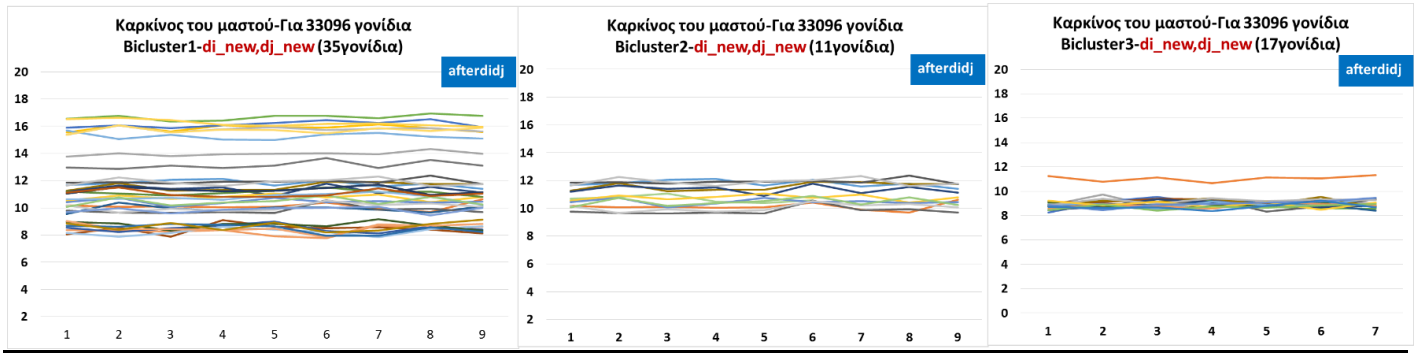



Αριστερό τμήμα: Το πρώτο διάγραμμα απεικονίζει 59 γονίδια. Εμφάνιση των 59 γονιδίων που προκύπτουν στο πρώτο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.β - αριστερό τμήμα)

Μεσαίο τμήμα: Το δεύτερο διάγραμμα απεικονίζει 11 γονίδια. Εμφάνιση των 11 γονιδίων που προκύπτουν στο δεύτερο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.β - μεσαίο τμήμα)

Δεξιό τμήμα: Το τρίτο διάγραμμα απεικονίζει 5 γονίδια. Εμφάνιση των 5 γονιδίων που προκύπτουν στο τρίτο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.β- δεξιό τμήμα)

**Εικόνα 5.5.γ.** - Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 33096 Γονίδια του Καρκίνου του Μαστού

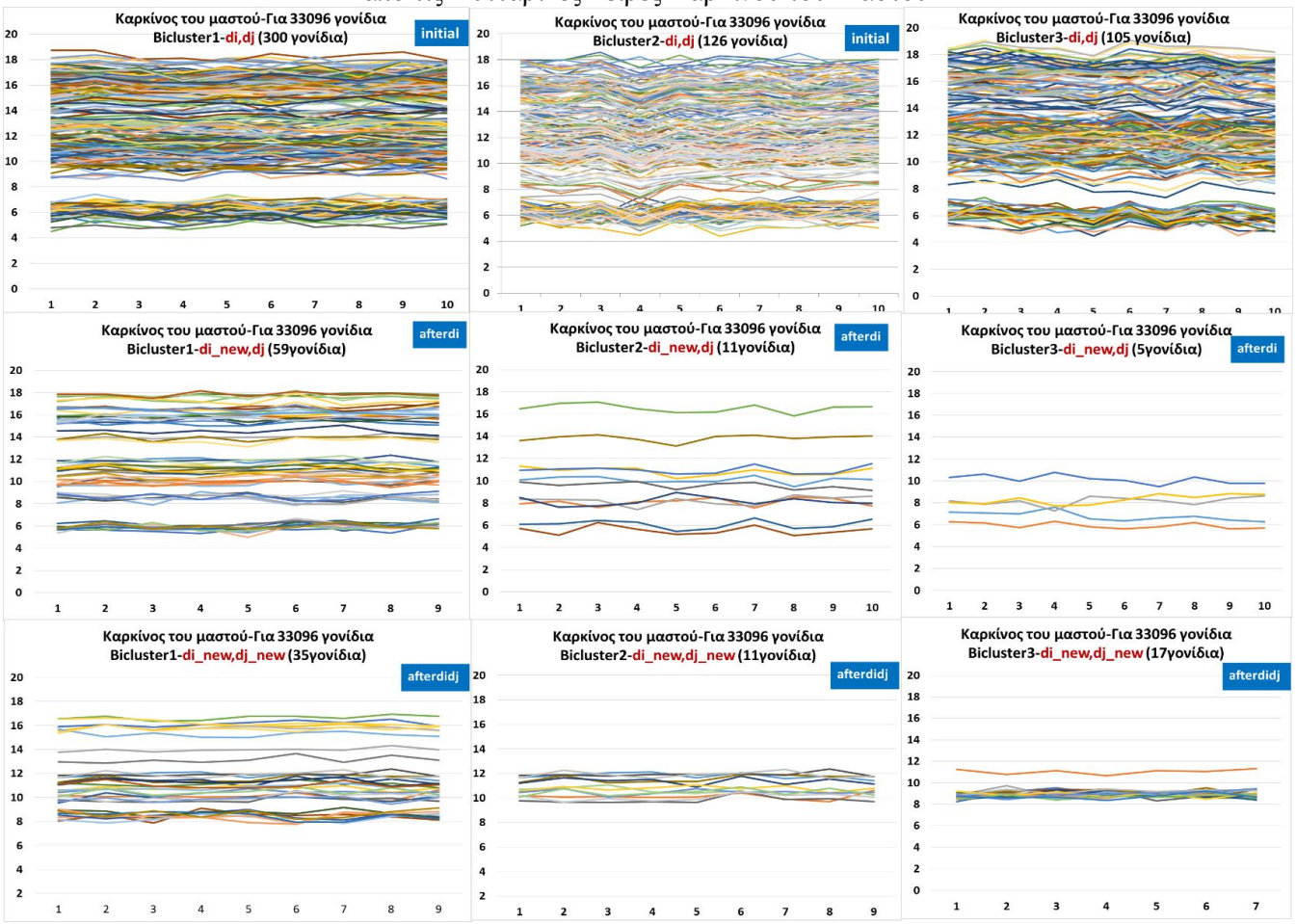



Αριστερό τμήμα: Το πρώτο διάγραμμα απεικονίζει 35 γονίδια. Εμφάνιση των 35 γονιδίων που προκύπτουν στο πρώτο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.γ - αριστερό τμήμα)

Μεσαίο τμήμα: Το δεύτερο διάγραμμα απεικονίζει 11 γονίδια. Εμφάνιση των 11 γονιδίων που προκύπτουν στο δεύτερο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.γ - μεσαίο τμήμα)

Δεξιό τμήμα: Το τρίτο διάγραμμα απεικονίζει 17 γονίδια. Εμφάνιση των 17 γονιδίων που προκύπτουν στο τρίτο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.5.γ - δεξιό τμήμα)

**Εικόνα 5.6**. - Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 33096 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Μαστού



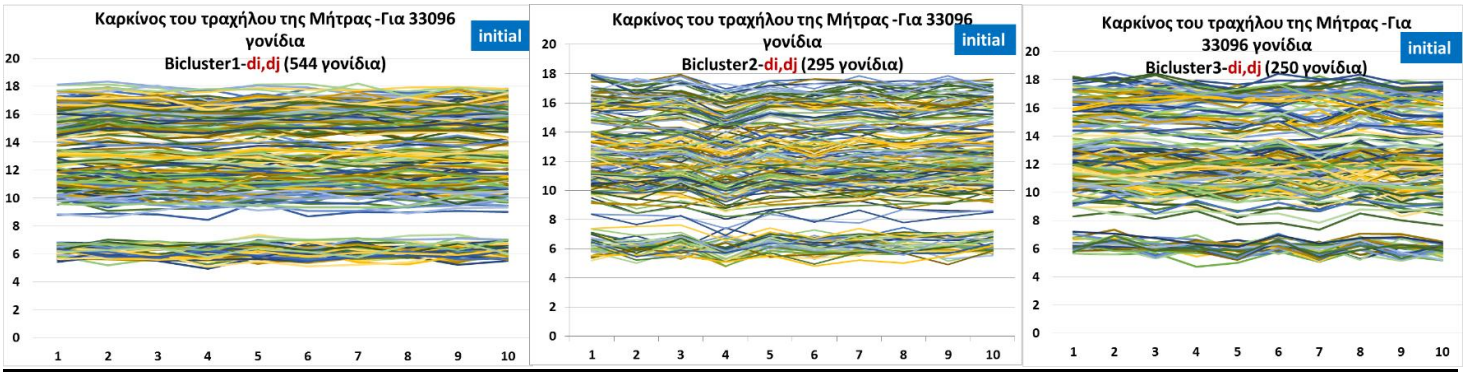
## 5.6. Αποτελέσματα Biclustering από Κυτταρικές Σειρές Καρκίνου του Τραχήλου της Μήτρας για 33096 Γονίδια

Όπως έχει ειπωθεί και σε προηγούμενο κεφάλαιο, παρατηρήσαμε ότι κατά την εκτέλεση του αλγορίθμου για το σύνολο των 33096 γονιδίων, ο όγκος των γονιδίων που ομαδοποιούνται σε ένα Bicluster και ο αριθμός των Biclusters είναι μεγάλος. Γι' αυτό το λόγο, προκειμένου να διακρίνουμε και να εξάγουμε συμπεράσματα για τη συμπεριφορά των γονιδίων εκτελέσαμε τον αλγόριθμο για τα 1000 πρώτα γονίδια από τα 33096, όπως παρουσιάστηκαν και τα αποτελέσματα αυτών των εκτελέσεων παραπάνω.

**Σε αυτό το σημείο, τρέξαμε τον αλγόριθμο Cheng and Church για τα 33096** και εστιάσαμε στα τρία πρώτα Biclusters από κυτταρικές σειρές **καρκίνου του τραχήλου της Μήτρας** .

Οι παρακάτω Εικόνες (5.6.α, 5.6.β, 5.6.γ) παρουσιάζουν το συνολικό πλήθος των γονιδίων και τη συμπεριφορά τους σε κάθε ένα από τα τρία Biclusters μετά τη πρώτη εκτέλεση του αλγορίθμου (Εικόνα 5.6.α), μετά τη δεύτερη εκτέλεση του αλγορίθμου (Εικόνες 5.6.β) και τέλος, μετά την Τρίτη εκτέλεση του αλγορίθμου (Εικόνες 5.6.γ).

**Εικόνα 5.6.α. -** Πρώτη Εκτέλεση του Αλγορίθμου (*di,dj*) για 33096 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας

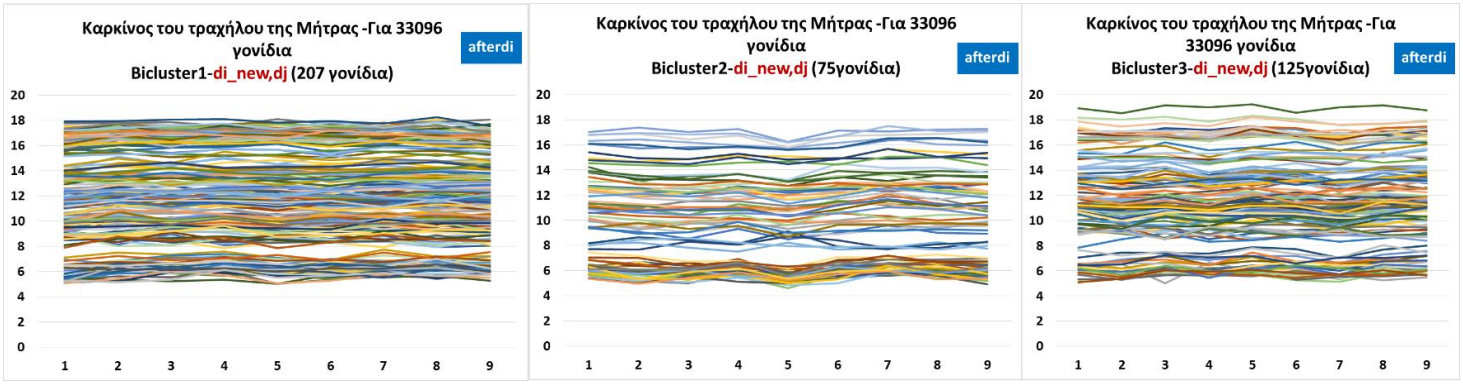



Το πρώτο διάγραμμα λόγω του πλήθους των γονιδίων (544) που προκύπτουν στο Bicluster1 κατά την πρώτη εκτέλεση του αλγορίθμου, εμφανίζει ΜΟΝΟ τα 255 πρώτα γονίδια, εφόσον το πρόγραμμα του excel επιτρέπει την εμφάνιση έως 255 στοιχείων. (Εικόνα 5.6.α - αριστερό τμήμα)

Το δεύτερο διάγραμμα λόγω του πλήθους των γονιδίων (295) που προκύπτουν στο Bicluster2 κατά την πρώτη εκτέλεση του αλγορίθμου, εμφανίζει ΜΟΝΟ τα 255 πρώτα γονίδια, εφόσον το πρόγραμμα του excel επιτρέπει την εμφάνιση έως 255 στοιχείων. (Εικόνα 5.6.α - μεσαίο τμήμα)

Το τρίτο διάγραμμα απεικονίζει 250γονίδια. Εμφάνιση των 250 γονιδίων που προκύπτουν στο τρίτο Bicluster κατά την πρώτη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.α – δεξί τμήμα)

**Εικόνα 5.6.β.** - Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj)* για 33096 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας

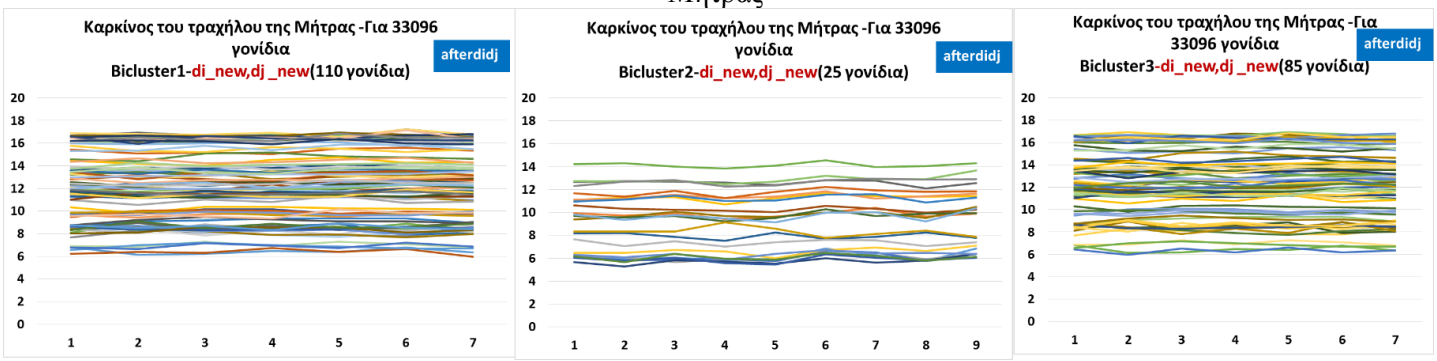



Το πρώτο διάγραμμα απεικονίζει 207 γονίδια. Εμφάνιση των 207 γονιδίων που προκύπτουν στο πρώτο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.β - αριστερό τμήμα)

Το δεύτερο διάγραμμα απεικονίζει 75 γονίδια. Εμφάνιση των 75 γονιδίων που προκύπτουν στο δεύτερο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.β – μεσαίο τμήμα)

Το τρίτο διάγραμμα απεικονίζει 125 γονίδια. Εμφάνιση των 125 γονιδίων που προκύπτουν στο τρίτο Bicluster κατά τη δεύτερη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.β - δεξί τμήμα)

**Εικόνα 5.6.γ. -** Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 33096 Γονίδια του Καρκίνου του Τραχήλου της Μήτρας

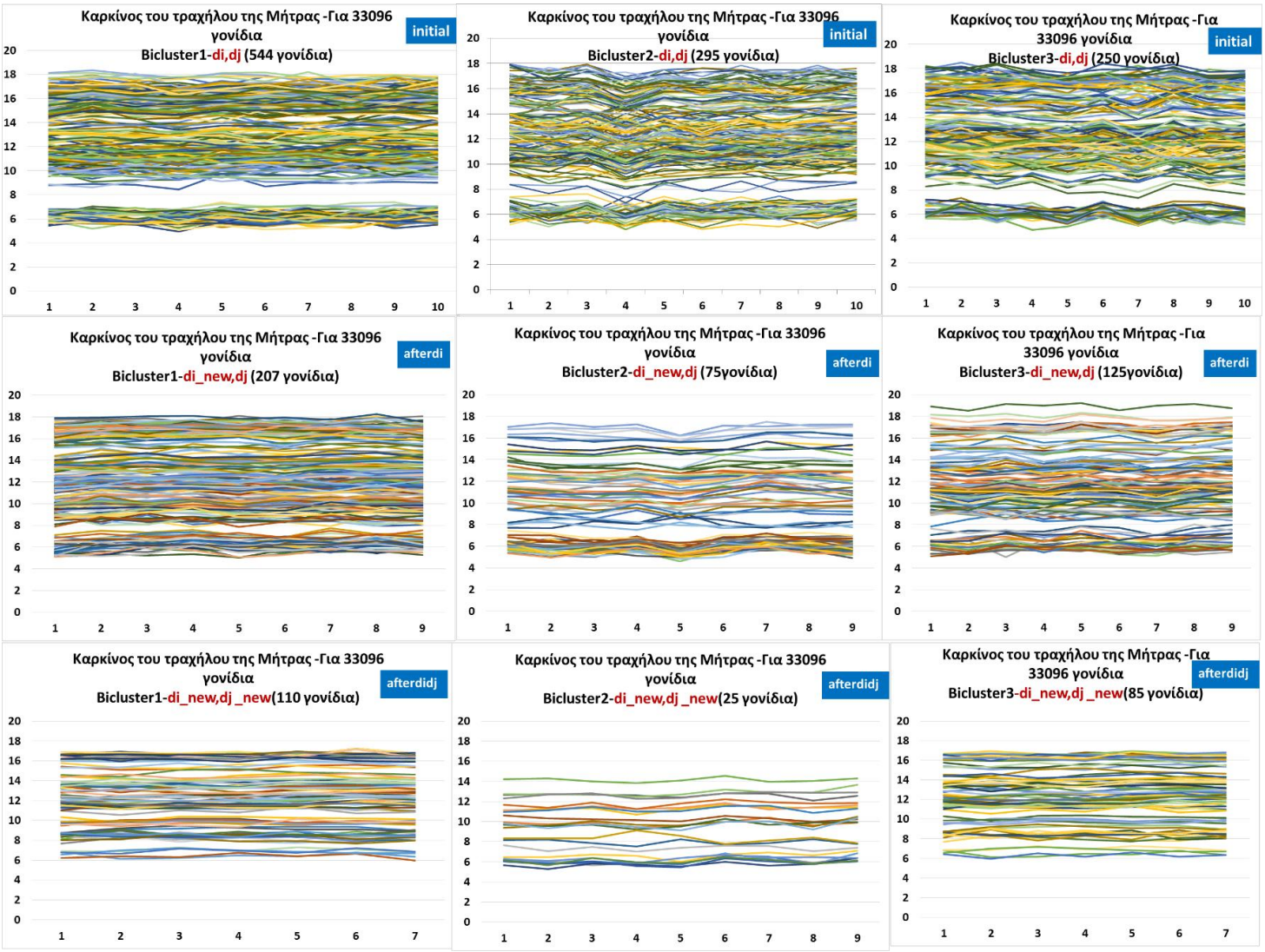



Το πρώτο διάγραμμα απεικονίζει 110γονίδια. Εμφάνιση των 110 γονιδίων που προκύπτουν στο πρώτο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.γ - αριστερό τμήμα)

Το δεύτερο διάγραμμα απεικονίζει 25 γονίδια. Εμφάνιση των 25 γονιδίων που προκύπτουν στο δεύτερο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.γ - μεσαίο τμήμα)

Το τρίτο διάγραμμα απεικονίζει 85 γονίδια. Εμφάνιση των 85γονιδίων που προκύπτουν στο τρίτο Bicluster κατά την τρίτη εκτέλεση του αλγορίθμου. (Εικόνα 5.6.γ - δεξί τμήμα)

**Εικόνα 5.7. -** Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng and Church για 33096 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Τραχήλου της Μήτρας



# 6

## Αξιολόγηση Των Αποτελεσμάτων

Η 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 33096 γονιδίων για 38 κυτταρικές σειρές τεσσάρων καρκινικών τύπων (μαστού, τραχήλου της μήτρας, ενδομητρίου, και ωοθηκών).[57]

Τα προβλήματα που τέθηκαν προς αντιμετώπιση, όπως αναφέρθηκαν, ήταν:

α) ο εξαιρετικά μεγάλος αριθμός των ομάδων διπλής κατηγοριοποίησης που προκύπτουν κατά την αρχική εκτέλεση του αλγορίθμου για τα 33096 γονίδια για τον καθένα από τους τέσσερις καρκινικούς τύπους ξεχωριστά, παραδείγματος χάριν το πλήθος των Biclusters για τον καρκίνο του μαστού ανερχόταν σε 2411, (κεφάλαιο 5),

β) το μεγάλο πλήθος των γονιδίων που ομαδοποιούνται σε καθεμιά ομάδα διπλής κατηγοριοποίησης κατά την αρχική εκτέλεση του αλγορίθμου για τα 33096 γονίδια και

γ) η ανομοιομορφία μεταξύ των γονιδίων που συγκεντρώνονται σε μία ομάδα διπλής κατηγοριοποίησης, δηλαδή τα γονίδια διέφεραν στην τάση παρουσιάζοντας μεγάλες διακυμάνσεις, αλλά και στη μεταξύ τους απόσταση καλύπτοντας ένα μεγάλο εύρος τιμών.

Η εφαρμογή της μεθόδου διπλής κατηγοριοποίησης στο σύνολο δεδομένων καρκινικών κυτταρικών σειρών, είχε ως στόχο τη μελέτη ενός μικρού αριθμού ομάδων διπλής κατηγοριοποίησης και κατ' επέκταση τη μελέτη ενός μικρού αριθμού γονιδίων ώστε να είναι πιο εύκολη η παρατήρηση της συμπεριφοράς τους και μετέπειτα η εξαγωγή συμπερασμάτων με βάση αυτή. Το αποτέλεσμα αυτής της προσέγγισης ήταν η ανάδειξη των 3 πρώτων ομάδων διπλής κατηγοριοποίησης για τα 1000 πρώτα γονίδια από τα 33096 γονίδια για τις κυτταρικές σειρές των τεσσάρων τύπων καρκίνου (του μαστού, του τραχήλου της μήτρας, των ωοθηκών και του ενδομητρίου).

Αφού εκτελέστηκε ο αλγόριθμος Cheng και Church στην αρχική του μορφή (1η εκτέλεση) για κάθε καρκίνο, στη συνέχεια εκτελέστηκε ο αλγόριθμος C&C με αλλαγή στον τρόπο υπολογισμού του *di* (2η εκτέλεση). Αυτή η εκτέλεση δίνει βαρύτητα στην ομοιομορφία μεταξύ των γονιδίων με βάση την τάση τους, όπως καθορίζεται από το μέτρο που αλλάζουμε. Τέλος, αλλάζοντας και το *dj* (3η εκτέλεση) για κάθε τύπο καρκίνου επιδιώκουμε την μείωση της απόστασης μεταξύ των γονιδίων που βρίσκονται στο ίδιο Bicluster.

Στον παρακάτω συγκεντρωτικό Πίνακα 6.1, όπως έχει παρουσιαστεί και στο κεφάλαιο 5, παρουσιάζεται το συνολικό πλήθος των γονιδίων (probeid) που ομαδοποιείται σε καθένα από τα τρία Biclusters στον καρκίνο του μαστού για καθεμιά από τις τρεις εκτελέσεις του αλγορίθμου για τα 1000 πρώτα γονίδια. Με χρωματιστό

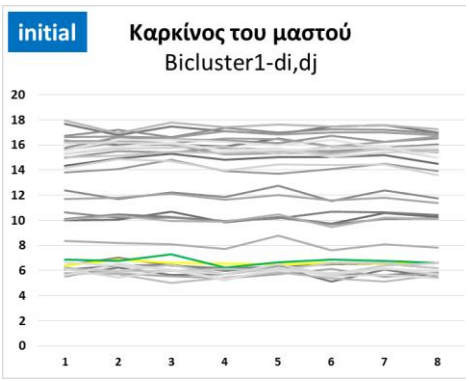
πλαίσιο παρουσιάζονται τα γονίδια (probeid) τα οποία μεταπηδούν από το ένα Bicluster σε ένα από τα άλλα δύο Biclusters κατά την εφαρμογή των τριών εκτελέσεων του αλγορίθμου. Επίσης, με έντονη πλάγια γραφή και χρωματιστό πλαίσιο παρουσιάζονται τα γονίδια τα οποία εμφανίζονται στο ίδιο Bicluster κατά τη διάρκεια δύο τουλάχιστον διαφορετικών εκτελέσεων του αλγορίθμου.

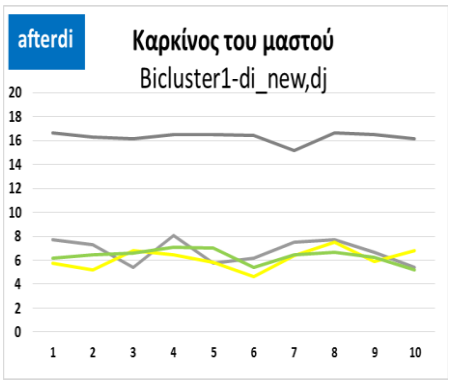
**Πίνακας 6.1**. - Παρουσίαση τριών πρώτων Biclusters του καρκίνου του μαστού - για καθεμιά από τις τρεις εκτελέσεις του αλγορίθμου στα 1000 πρώτα γονίδια. Τα γονίδια δίνονται με τη μορφή κωδικών (probeid)

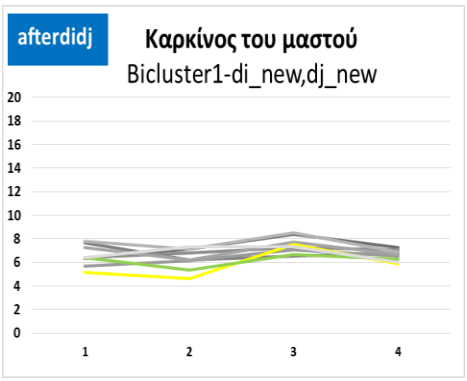
| *Πρώτη εκτέλεση του αλγορίθμου(didj)* | | |
|---|---|---|
| **Bicluster1** | **Bicluster2** | **Bicluster3** |
| 226368 | 164432 | 217298 |
| 213726 | 151787 | 165735 |
| 224284 | 223314 | 223080 |
| 224286 | 161079 | 225083 |
| 199782 | 140336 | 136093 |
| 173890 | 127051 | 156066 |
| 101695 | 137543 | 180730 |
| 147346 | 160701 | 154584 |
| 138543 | 217593 | 140229 |
| 186673 | 136075 | 173299 |
| 174944 | 230085 | 162744 |
| 155326 | 123356 | 188198 |
| 236908 | 151430 | 109012 |
| 155784 | 178114 | 200313 |
| 149734 | 182060 | 166147 |
| 228186 | 149009 | 120825 |
| 139301 | 216613 | 183840 |
| 180707 | 194406 | 174851 |
| 128561 | 121364 | 182708 |
| 181532 | 136278 | 123369 |
| 182504 | | 109258 |
| 139732 | | 140456 |
| 116808 | | 202153 |
| 180501 | | 189928 |
| 214717 | | 209910 |
| 236216 | | 163565 |
| 215964 | | 104468 |
| 128192 | | 201366 |
| 170969 | | |
| 229775 | | |
| 161284 | | |
| 139151 | | |
| 108780 | | |
| 109210 | | |
| 104969 | | |
| 118956 | | |

| *Δεύτερη εκτέλεση του αλγορίθμου(di_new,dj)* | | |
|---|---|---|
| **Bicluster1** | **Bicluster2** | **Bicluster3** |
| 139301 | 162744 | 231349 |
| 214589 | 181532 | 140490 |
| 180501 | 201489 | |
| 214717 | 236216 | |
| | 140456 | |
| | 170969 | |

| *Τρίτη εκτέλεση του αλγορίθμου(di_new,dj_new)* | | |
|---|---|---|
| **Bicluster1** | **Bicluster2** | **Bicluster3** |
| 219524 | 142719 | 151787 |
| 231349 | 181532 | 162744 |
| 111270 | 101235 | 125603 |
| 140490 | 201489 | 132754 |
| 214589 | 146732 | 115325 |
| 190994 | 109288 | |
| 180501 | 234158 | |
| 214717 | | |
| 163565 | | |



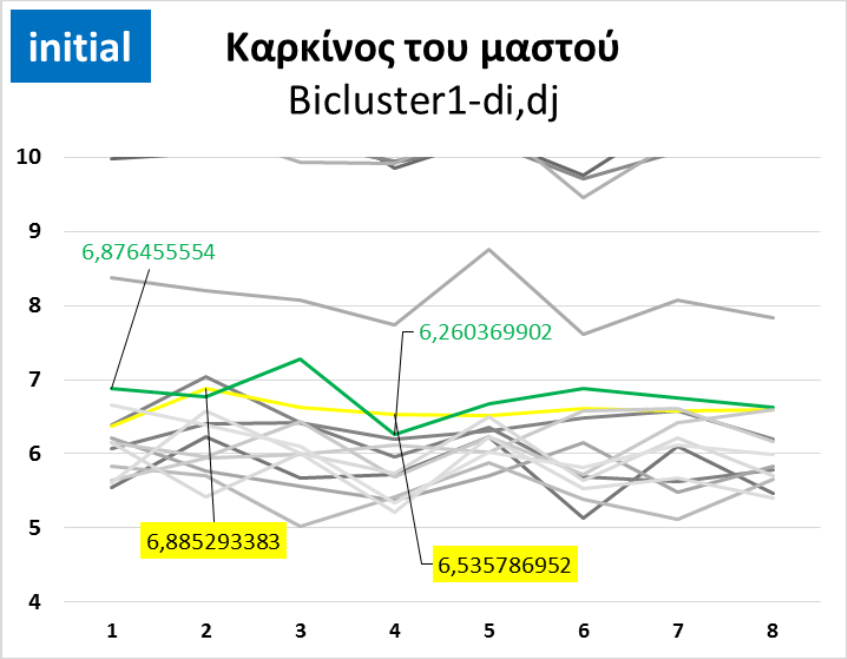

**Εικόνα 6.1**. - Συγκεντρωτική Παρουσίαση των Τριών εκτελέσεων του Αλγορίθμου Cheng και Church για 1000 Γονίδια από τις Κυτταρικές Σειρές Καρκίνου του Μαστού ( Bicluster1), εστιάζοντας στη συμπεριφορά των γονιδίων με probeid 214717 και 180501.

Για παράδειγμα, στον Πίνακα 6.1 απεικονίζεται η συμπεριφορά του γονιδίου με probeid 214717 και του γονιδίου με probeid 180501. Τα γονίδια 214717 και 180501 τυγχάνει να παραμένουν στο Bicluster1 και στις τρεις εκτελέσεις. Δηλαδή, δεν χρειάστηκε να μετακινηθούν σε άλλο Bicluster, ενώ ταυτόχρονα άλλα γονίδια που «ταίριαζαν» με τα δύο γονίδια που εξετάζουμε, στην τάση και στην τιμή, μετακινήθηκαν στο Bicluster 1.

Στα παρακάτω διαγράμματα (Εικόνες 6.1.α, 6.1.β, 6.2.α, 6.2.β, 6.3.α, 6.3.β, 6.4.α, 6.4.β, 6.5.α, 6.5.β, 6.6.α, 6.6.β), η κλίμακα του άξονα y έχει προσαρμοστεί από την τιμή 4 μέχρι την τιμή 10, ώστε να είναι πιο ευδιάκριτη η συμπεριφορά των γονιδίων όσον αφορά το εύρος της μεταξύ τους απόστασης και τη διακύμανσή τους (κορυφές). Η αλλαγή της κλίμακας έχει ως αποτέλεσμα την παρουσίαση ενός υποσυνόλου γονιδίων.

Επίσης, στις Εικόνες 6.1.α, 6.2.α, 6.3.α, 6.4.α, 6.4.β, 6.5.α, και 6.6.α απεικονίζονται οι κυτταρικές σειρές που συμμετέχουν σε κάθε Bicluster, με σκοπό την ανάδειξη του φαινομένου όπου κυτταρικές σειρές προστίθενται ή αφαιρούνται κατά την εφαρμογή των τριών διαφορετικών εκτελέσεων.

**Παρατήρηση της συμπεριφοράς των γονιδίων με κωδικό «probeid» 241717 και 180501**

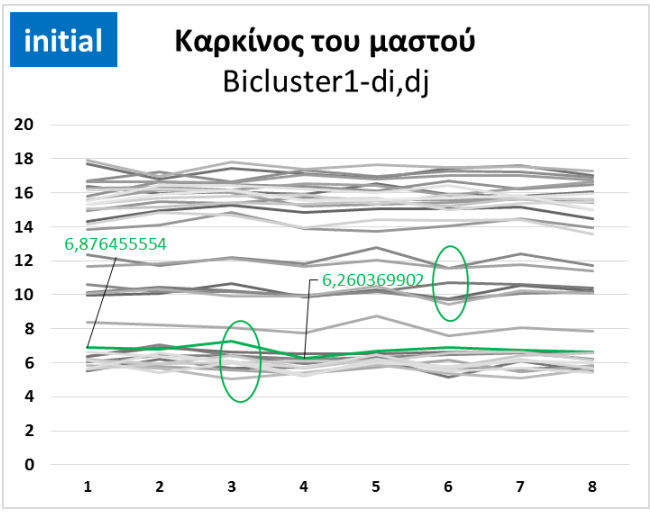



**Εικόνα 6.1.α.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di,dj*) για 1000 Γονίδια του Καρκίνου του Μαστού Ανάδειξη των μεγαλύτερων και μικρότερων τιμών των Καρκινικών Κυτταρικών Σειρών που Συμμετέχουν

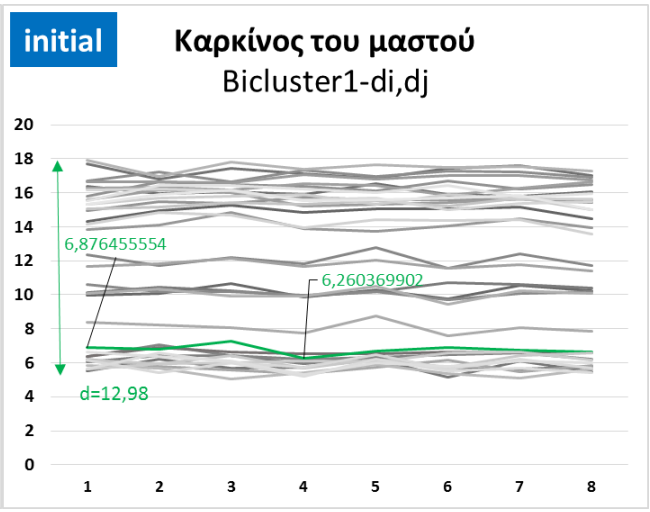
Σε πρώτη φάση, στην Εικόνα 6.1.α παρατηρείται η απόσταση των γονιδίων και η διακύμανση των γονιδίων στο Bicluster 1, κατά την πρώτη εκτέλεση του αλγορίθμου C&C για τα 1000 γονίδια του καρκίνου του μαστού.

Επίσης, στην παραπάνω Εικόνα 6.1.α απεικονίζονται οι κυτταρικές σειρές που περιλαμβάνονται στο Bicluster1 μετά την πρώτη εκτέλεση του αλγορίθμου και οι τιμές των γονιδίων (π.χ. τιμές γονιδίου με probeid 214717) σε κάθε καρκινική κυτταρική σειρά. Οι 10 κυτταρικές σειρές που απαρτίζουν τον πίνακα (των δεδομένων) στον καρκίνο του μαστού, μετά την πρώτη εκτέλεση μειώνονται στις 8.

Απομονώνοντας το γονίδιο με κωδικό 214717 παρατηρούμε: 1) την ανομοιομορφία μεταξύ των γονιδίων που συγκεντρώνονται στο Bicluster1, και 2) την μεγάλη απόσταση που παρουσιάζουν τα γονίδια μεταξύ τους. Συγκεκριμένα, τα γονίδια διαφέρουν στην τάση και παρουσιάζουν μεγάλες διακυμάνσεις (Εικόνα 6.1.α), ενώ το εύρος της απόστασης των γονιδίων στο Bicluster1 μετά την πρώτη εκτέλεση είναι d=12,98 όπως φαίνεται στην Εικόνα 6.1.γ, στην οποία παρουσιάζεται το συνολικό πλήθος των εξαγόμενων γονιδίων που ομαδοποιούνται στο Bicluster1.

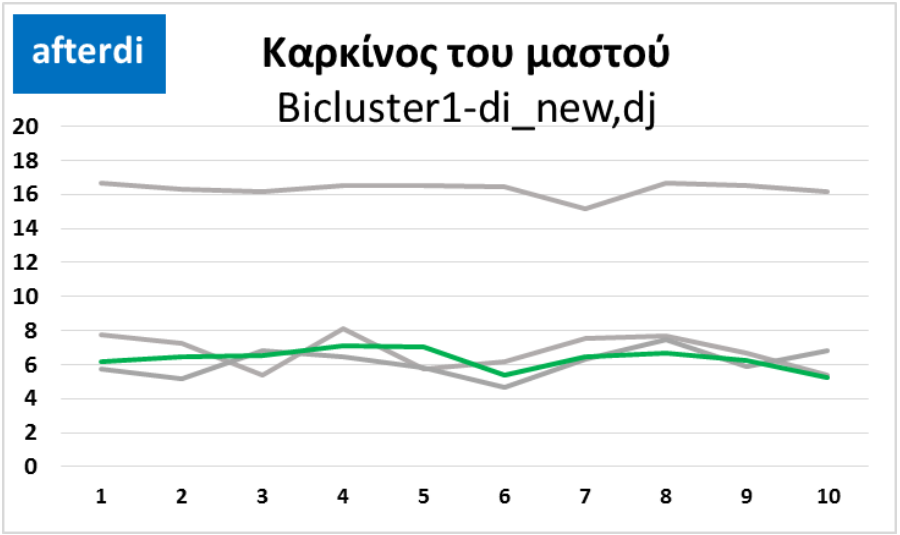



**Εικόνα 6.1.β.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di*,*dj*) για 1000 Γονίδια του Καρκίνου του Μαστού Αναδεικνύοντας την Ανομοιομορφία μεταξύ των γονιδίων




**Εικόνα 6.1.γ.** - Πρώτη Εκτέλεση του Αλγορίθμου (*di*,*dj*) για 1000 Γονίδια του Καρκίνου του Μαστού Αναδεικνύοντας το Εύρος της Απόστασης μεταξύ των γονιδίων

Στη συνέχεια, κατά τη δεύτερη εκτέλεση, δηλαδή μετά την τροποποίηση του *di* (*di*_new,*dj*) το γονίδιο το οποίο "παρακολουθείται", παραμένει στο Bicluster1, όπου συσπειρώνεται με άλλα γονίδια με παρόμοια τάση. Αυτό επιτυγχάνεται λόγω των αλλαγών που έχουμε κάνει στο μέτρο.

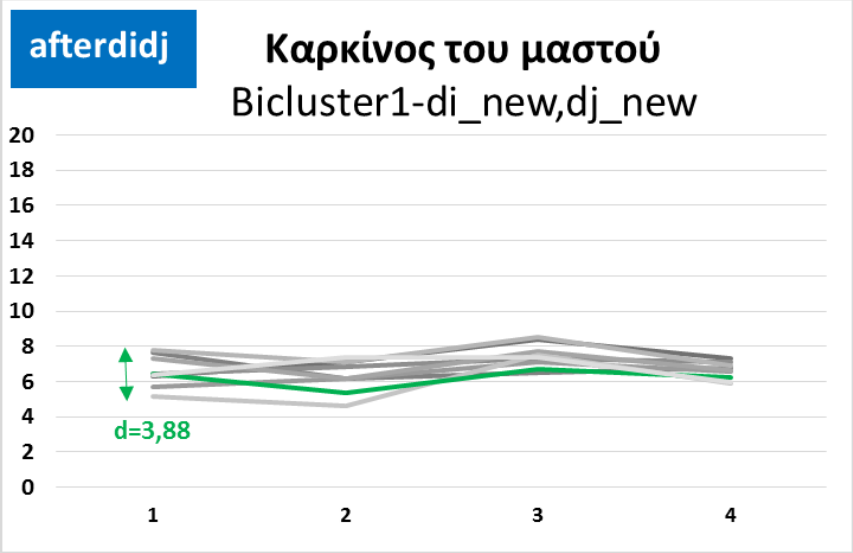



**Εικόνα 6.2.** - Δεύτερη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*) για 1000 Γονίδια του Καρκίνου του Μαστού Ανάδειξη των Καρκινικών Κυτταρικών Σειρών που Συμμετέχουν

Επίσης, στην παραπάνω Εικόνα 6.2.α απεικονίζονται οι κυτταρικές σειρές που περιλαμβάνονται στο Bicluster1 μετά τη δεύτερη εκτέλεση του αλγορίθμου και οι τιμές των γονιδίων (π.χ. τιμές γονιδίου με probeid 214717) σε κάθε καρκινική κυτταρική σειρά. Από τις 6 κυτταρικές σειρές που παρέμειναν μετά την πρώτη εκτέλεση του αλγορίθμου, μετά το πέρας της δεύτερης εκτέλεσης του αλγορίθμου παρατηρούμε ότι συμμετέχουν και οι 10 κυτταρικές σειρές .

Τέλος, μπορούμε να διακρίνουμε τη μείωση των γονιδίων από τα 30 που είχαμε στην πρώτη εκτέλεση, στα 4 γονίδια.

Ακολούθως, η παρακάτω Εικόνα 6.3.α απεικονίζει τις κυτταρικές σειρές που περιλαμβάνονται στο Bicluster1 μετά την τρίτη εκτέλεση του αλγορίθμου και οι τιμές των γονιδίων (π.χ. τιμές γονιδίου με probeid 214717) σε κάθε καρκινική κυτταρική σειρά. Ωστόσο, παρατηρείται το γεγονός ότι έχουν παραμείνει μόνο τέσσερις κυτταρικές σειρές.

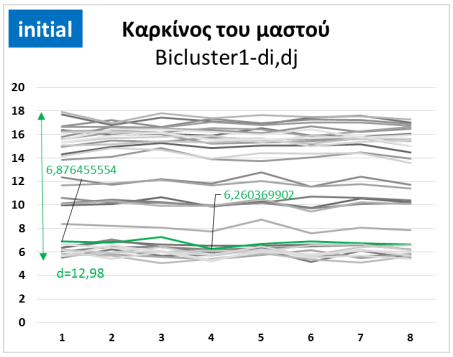



**Εικόνα 6.3.α.** - Τρίτη Εκτέλεση του Αλγορίθμου (*di*_new,*dj*_new) για 1000 Γονίδια του Καρκίνου του Μαστού Ανάδειξη των Καρκινικών Κυτταρικών Σειρών που Συμμετέχουν

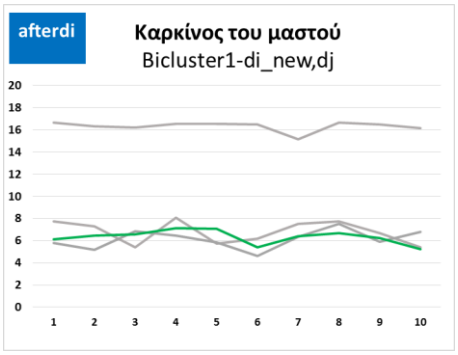
Επομένως, απομονώνοντας το γονίδιο με κωδικό 214717 παρατηρούμε την ομοιομορφία πλέον μεταξύ των γονιδίων που συγκεντρώνονται στο Bicluster1, δηλαδή τα γονίδια παρουσιάζουν παρόμοια τάση και μικρές διακυμάνσεις (Εικόνα 6.3.α), αλλά και η απόσταση που παρουσιάζουν τα γονίδια μεταξύ τους είναι πλέον μικρή. Συγκεκριμένα, το εύρος των γονιδίων στο Bicluster1 μετά την πρώτη εκτέλεση ήταν d=12,98 όπως φαίνεται στην Εικόνα 6.3.α, ενώ μετά τις τρεις εκτελέσεις, το εύρος των γονιδίων είναι d'=3,88.

Στα διαγράμματα (Εικόνες 6.3.γ, 6.3.δ, 6.3.ε), παρατηρούμε τη συμπεριφορά των γονιδίων του Bicluster1 και στις τρεις εκτελέσεις, με κλίμακα του άξονα y από την τιμή 0 μέχρι την τιμή 20, ώστε να έχουμε την πλήρη εικόνα των γονιδίων.

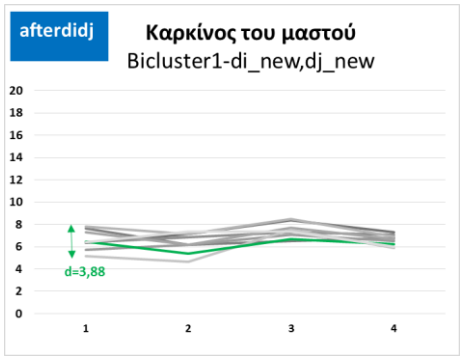
Τα διαγράμματα (Εικόνες 6.3.γ, 6.3.δ, 6.3.ε) απεικονίζουν τη συμπεριφορά των γονιδίων κατά τις τρεις εκτελέσεις του αλγορίθμου για τα 1000 πρώτα Γονίδια του Καρκίνου του Μαστού για το Bicluster1 .




**Εικόνα 6.3.β**. - Πρώτη Εκτέλεση του Αλγορίθμου (di,dj) για 1000 Γονίδια του Καρκίνου του Μαστού

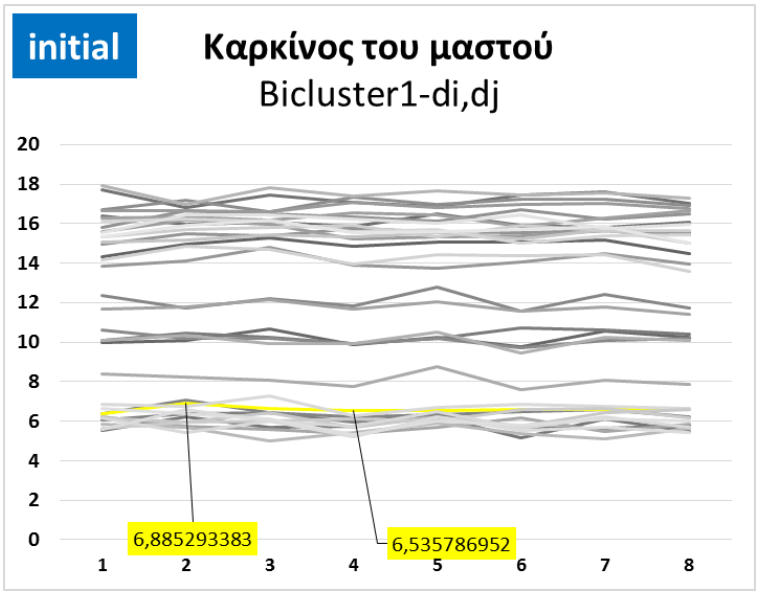



**Εικόνα 6.3.γ.** - Δεύτερη Εκτέλεση του Αλγορίθμου (di_new,dj) για 1000 Γονίδια του Καρκίνου του Μαστού

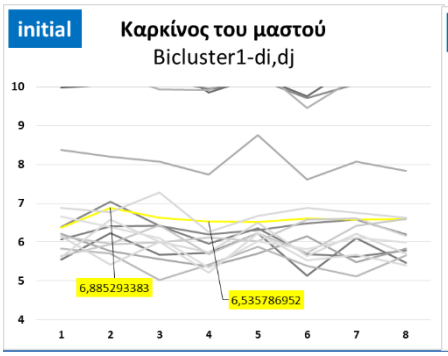



**Εικόνα 6.3.δ.** - Τρίτη Εκτέλεση του Αλγορίθμου (di_new,dj_new) για 1000 Γονίδια του Καρκίνου του Μαστού

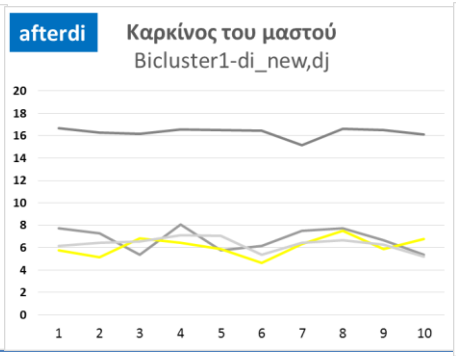
**Παρατήρηση της συμπεριφοράς του γονιδίου με αριθμό probeid 180501**



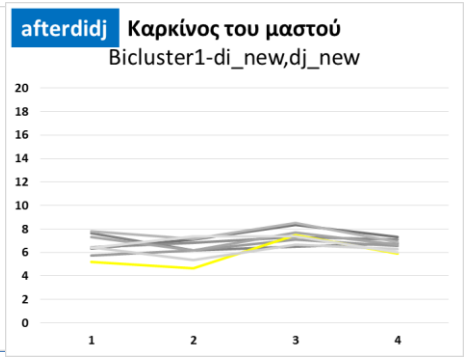

**Εικόνα 6.4.α. -** Πρώτη Εκτέλεση του Αλγορίθμου (didj) για 1000 Γονίδια του Καρκίνου του Μαστού Ανάδειξη των Καρκινικών Κυτταρικών Σειρών που Συμμετέχουν




**Εικόνα 6.4.β.** - Πρώτη Εκτέλεση του Αλγορίθμου (didj) για 1000 Γονίδια του Καρκίνου του Μαστού Αναδεικνύοντας το γονίδιο 180501




**Εικόνα 6.4.γ.** - Δεύτερη Εκτέλεση του Αλγορίθμου (di_new,dj) για 1000 Γονίδια του Καρκίνου του Μαστού Αναδεικνύοντας το γονίδιο 180501



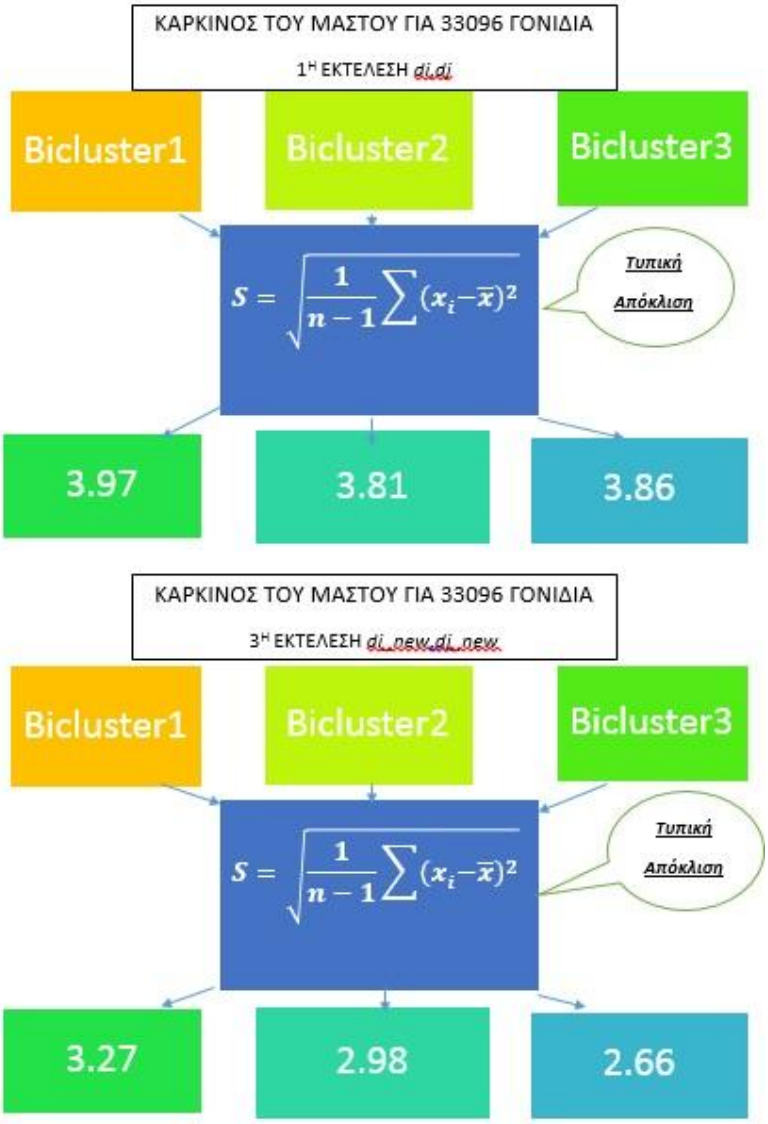

**Εικόνα 6.4.δ.** - Τρίτη Εκτέλεση του Αλγορίθμου (di_new,dj_new) για 1000 Γονίδια του Καρκίνου του Μαστού Αναδεικνύοντας το γονίδιο 18050

Σε δεύτερη φάση, στην Εικόνα 6.4.α παρατηρείται η απόσταση των γονιδίων και η διακύμανση των γονιδίων στο Bicluster1, κατά την πρώτη εκτέλεση του αλγορίθμου C&C για τα 1000 γονίδια του καρκίνου του μαστού. Επειδή το γονίδιο με probeid 180501 ακολουθεί την ίδια πορεία κατά τη διάρκεια των τριών εκτελέσεων με το γονίδιο 214717 που παρουσιάστηκε παραπάνω, δηλαδή παραμένει και στις τρεις εκτελέσεις στο Bicluster 1, παρουσιάστηκε συνοπτικά η συμπεριφορά του κατά τη διάρκεια των τριών εκτελέσεων.

## Στατιστική Επικύρωση Αποτελεσμάτων

Η επικύρωση των αποτελεσμάτων Biclusters είναι ένα σημαντικό πρόβλημα και ταυτόχρονα μια πρόκληση, λόγω των διαφορετικών κριτηρίων που χρησιμοποιούνται και των διαφορετικών στόχων που τίθενται στο πλαίσιο της διπλής κατηγοριοποίησης. Υπάρχουν αρκετές κοινές στρατηγικές επικύρωσης για τα αποτελέσματα διπλής κατηγοριοποίησης, συμπεριλαμβανομένης της επικύρωσης βάσει δεικτών, χρησιμοποιώντας επικύρωση με τη γνώση του πεδίου (domain knowledge) και στατιστικές δοκιμές.

Στην εν λόγω εργασία, για να επικυρώσουμε τα εξαγόμενα αποτελέσματα, χρησιμοποιήσαμε την Τυπική Απόκλιση. Υπολογίσαμε την Τυπική Απόκλιση στην πρώτη και στην τρίτη εκτέλεση του αλγορίθμου. Στην παρακάτω εικόνα παρουσιάζονται οι εξαγόμενες Τυπικές Αποκλίσεις. Παρατηρείται το γεγονός ότι μετά το πέρας και των τριών εκτελέσεων οι Τυπικές Αποκλίσεις έχουν μειωθεί, γεγονός που μας δείχνει ότι πετύχαμε το στόχο μας.




**Εικόνα 6.5. -** Στατιστική Επικύρωση Αποτελεσμάτων

## Οπτικοποίηση Αποτελεσμάτων Διπλής Κατηγοριοποίησης

Η οπτικοποίηση των αποτελεσμάτων διπλής κατηγοριοποίησης είναι δύσκολη λόγω των μη συνεχών και αλληλεπικαλυπτόμενων κατατάξεων τόσο στη σειρά όσο και στη στήλη. Η πιο δημοφιλής τεχνική οπτικοποίησης που αντιπροσωπεύει μια μονή ομάδα διπλής κατηγοριοποίησης είναι η τεχνική θερμικού χάρτη. Ένας θερμικός χάρτης είναι ένα ορθογώνιο πλέγμα που αποτελείται από εικονοστοιχεία, κάθε ένα από τα οποία αντιστοιχεί σε μια τιμή δεδομένων. Οι διαφορετικές κλίμακες γκρι/χρώματος αντιστοιχούν σε διαφορετικές τιμές δεδομένων. Συνήθως, όσο πιο φωτεινό είναι το χρώμα, τόσο μεγαλύτερη είναι η τιμή. Με αυτόν τον τρόπο, αν οι σειρές ή/και οι στήλες του συνόλου δεδομένων αναδιατάσσονται κατάλληλα, το μοτίβο διπλής κατηγοριοπόιησης γίνεται αντιληπτό οπτικά. Οι θερμικοί χάρτες αρκούν συνήθως για τον έλεγχο ενός μονού bicluster. Στην παρούσα εργασία, η οπτικοποίηση έγινε με τη βοήθεια της πλατφόρμας

απεικόνισης και ανάλυσης πινάκων GENE-E[61] που σχεδιάστηκε για να υποστηρίξει την εξερεύνηση οπτικών δεδομένων. Η πλατφόρμα GENE-E περιέχει επίσης εργαλεία ειδικά σχεδιασμένα για γονιδιωματικά δεδομένα. Για την παρακολούθηση των αλλαγών που επέρχονται κατά την εφαρμογή της προτεινόμενης μεθοδολογίας, έγινε προσαρμογή (adjustment) των τιμών στην κλίμακα που εμφανίζεται κατά την αρχική εκτέλεση του αλγορίθμου Cheng και Church. Στα σχήματα 6.5. – 6.9., οι εικόνες παρουσιάζονται σε σφαιρική (global) μορφή και όχι σχετική (relative)[1].[62]

Στα σχήματα 6.5.- 6.8 γίνεται οπτικά αντιληπτή η μεγάλη βελτίωση που έχει επιτευχθεί κατά την εφαρμογή της προτεινόμενης μεθοδολογίας στον **υποχώρο δεδομένων** για τον καρκίνο του μαστού, του τραχήλου της μήτρας, του ενδομητρίου, και των ωοθηκών. Αντίθετα, στο σχήμα 6.9. που αφορά στην οπτικοποίηση των αποτελεσμάτων διπλής κατηγοριοποίησης του συνόλου των δεδομένων για τον καρκίνο του μαστού δεν φαίνεται να είναι ορατή η αποτελεσματικότητα της εφαρμογής της προτεινόμενης μεθοδολογίας, γεγονός που μπορεί να οφείλεται στο μεγάλο όγκο των δεδομένων.

Πέρα από την οπτικοποίηση, η οποία επέτρεψε την επικύρωση των αποτελεσμάτων και ανέδειξε την προτεινόμενη μεθοδολογία ως κατάλληλη για εφαρμογή σ' έναν υποχώρο δεδομένων, επιχειρήθηκε παράλληλα και η βιολογική επικύρωση των αποτελεσμάτων μέσω της λειτουργικής ανάλυσης εμπλουτισμού για α) όρους γονιδιακής οντολογίας και β) βιολογικών μονοπατιών με τη βοήθεια της πλατφόρμας WebGestalt.[63] Όπως φαίνεται στους Πίνακες 6.2. και 6.3. μπόρεσαν να εξαχθούν ενδεικτικά συμπεράσματα βάσει των αποσπασματικών αποτελεσμάτων εμπλουτισμού ($p \leq 0.05$) βιολογικών διεργασιών και μοριακών μονοπατιών για τις ομάδες διπλής κατηγοριοποίησης, για τις τρεις εκτελέσεις. Αυτό οφείλεται στους παρακάτω κυρίως λόγους:

### Για τα Αποτελέσματα Υποχώρου Δεδομένων για όλους τους τύπους καρκίνου

- Στην έλλειψη σχολιασμού (annotation) των γονιδίων που συμμετέχουν στα Biclusters κατά τη δεύτερη και τρίτη εκτέλεση του αλγορίθμου.

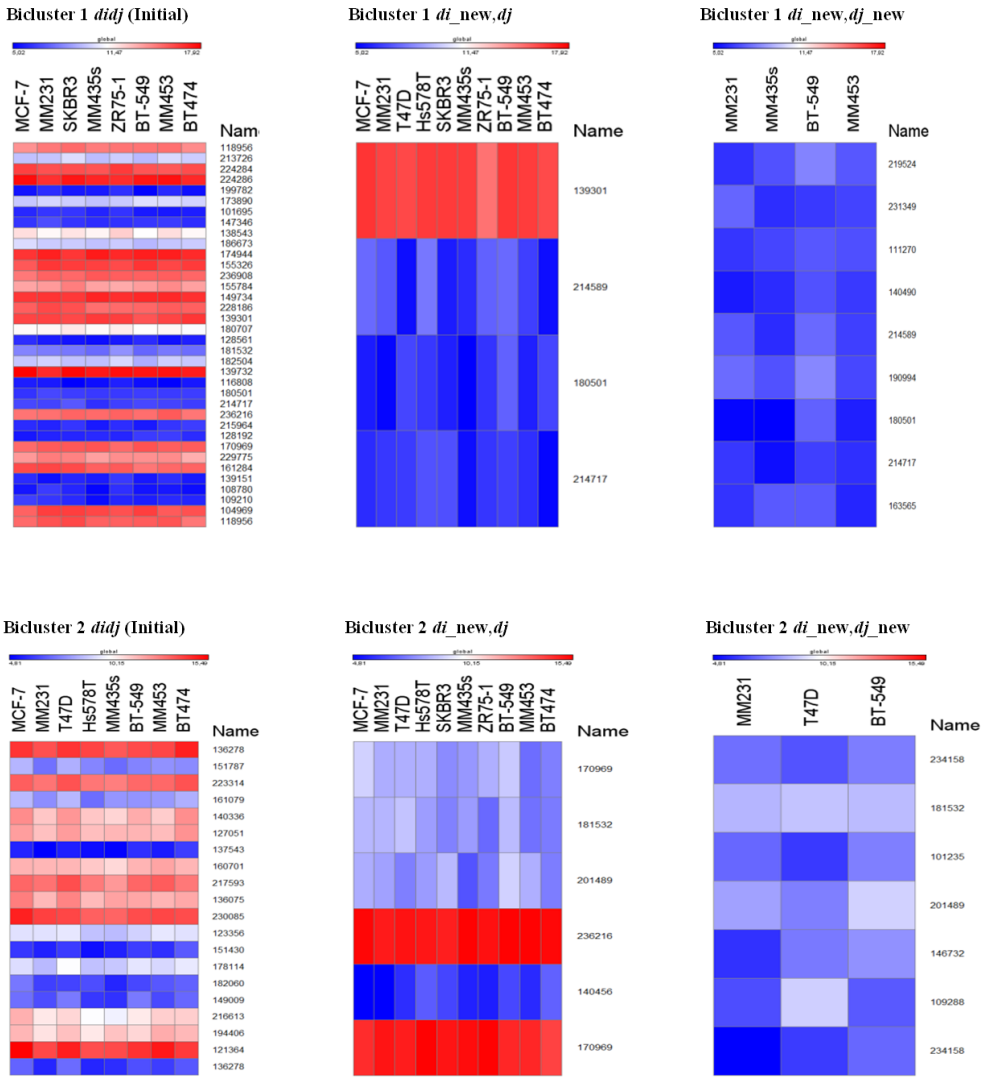
### Για τα Αποτελέσματα Συνόλου Δεδομένων για τον καρκίνο του μαστού

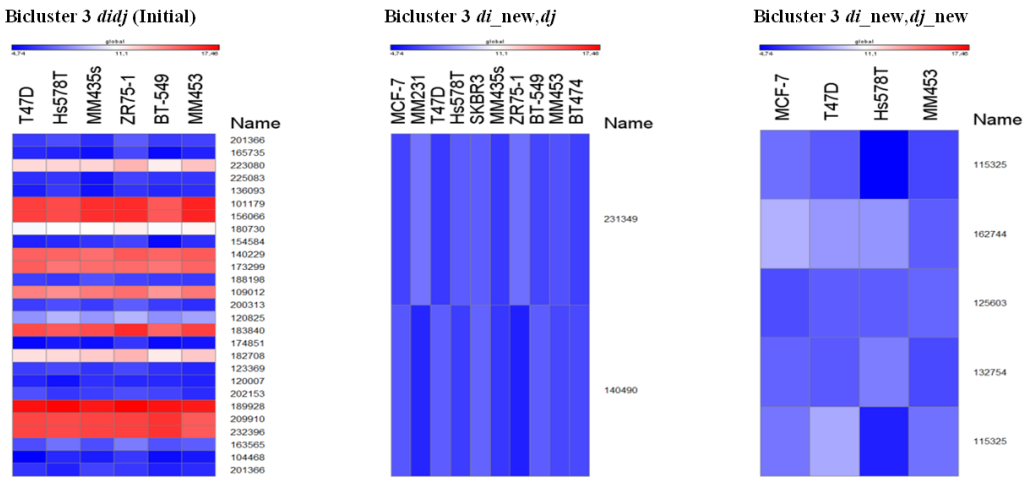
- Στην παρατήρηση ενός πολύ μικρού υποσυνόλου Biclusters (3 Biclusters) σε σχέση με τον πραγματικό αριθμό των Biclusters (π.χ. 2411 Biclusters στον καρκίνο του μαστού).

Για παράδειγμα στον πίνακα 6.2. παρουσιάζονται ικανοποιητικά αποτελέσματα για τα εμπλουτισμένα βιολογικά μονοπάτια στο Bicluster 1 και εν μέρει στο Bicluster 2, που δεν φαίνεται να συνάδουν ωστόσο με τα αποτελέσματα της οπτικοποίησης, ενώ δυστυχώς δεν μπορούμε να αξιολογήσουμε τα αποτελέσματα για το Bicluster 3 γιατί δεν βρέθηκαν εμπλουτισμένα βιολογικά μονοπάτια κατά τη $2^{η}$ και 3η εκτέλεση του αλγορίθμου. Επίσης, στον πίνακα 6.3. παρουσιάζονται ικανοποιητικά αποτελέσματα για τις εμπλουτισμένες βιολογικές διεργασίες στο Bicluster 2 που συνάδουν εν μέρει με τα αποτελέσματα της οπτικοποίησης, αλλά δεν μπορούμε να αξιολογήσουμε τα αποτελέσματα για τα Biclusters 1 και 3 γιατί δεν βρέθηκαν εμπλουτισμένες βιολογικές διεργασίες κατά τη $2^{η}$ ή/και $3^{η}$ εκτέλεση του αλγορίθμου.

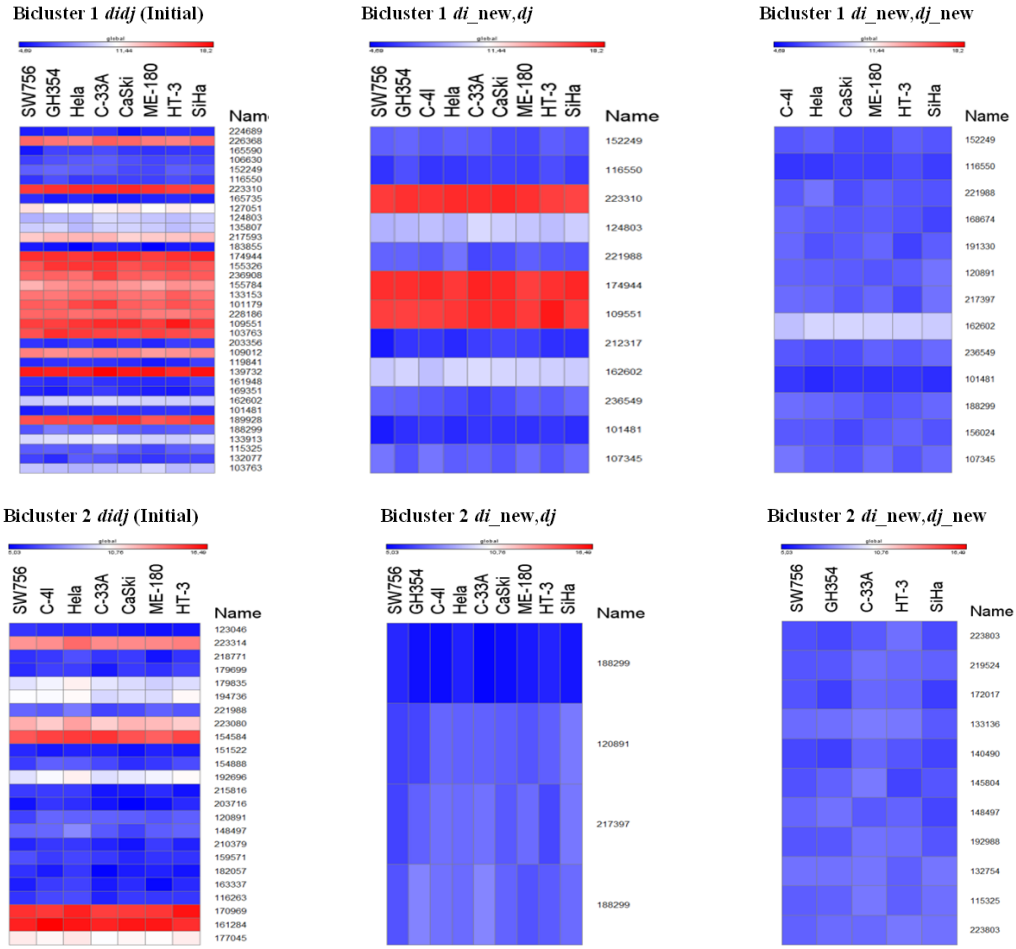
[1] **Σφαιρική (Global)**: Η πλατφόρμα GENE-E μετατρέπει τις τιμές σε χρώματα θερμικού χάρτη με τη χρήση των ελάχιστων και μέγιστων τιμών σε ολόκληρο το σύνολο δεδομένων. **Σχετική (Relative):** Η πλατφόρμα GENE-E μετατρέπει τις τιμές σε χρώματα θερμικού χάρτη χρησιμοποιώντας τις μέσες και μέγιστες τιμές για κάθε γονίδιο ή τις τυπικές αποκλίσεις από τον μέσο όρο για κάθε γονίδιο.

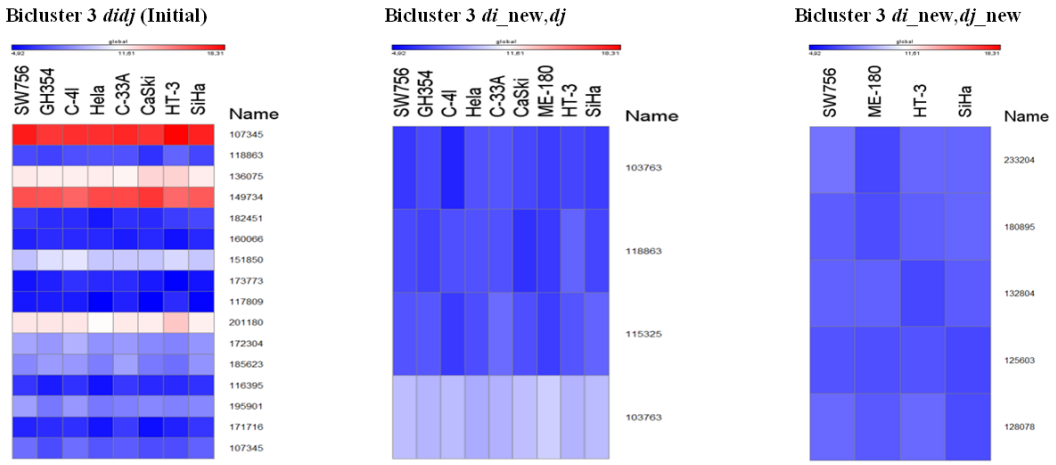
**Σχήμα 6.5.** – Οπτικοποίηση Αποτελεσμάτων Διπλής Κατηγοριοποίησης Υποχώρου Δεδομένων (1000 γονίδια - 10 κυτταρικές σειρές)
Καρκίνος του Μαστού





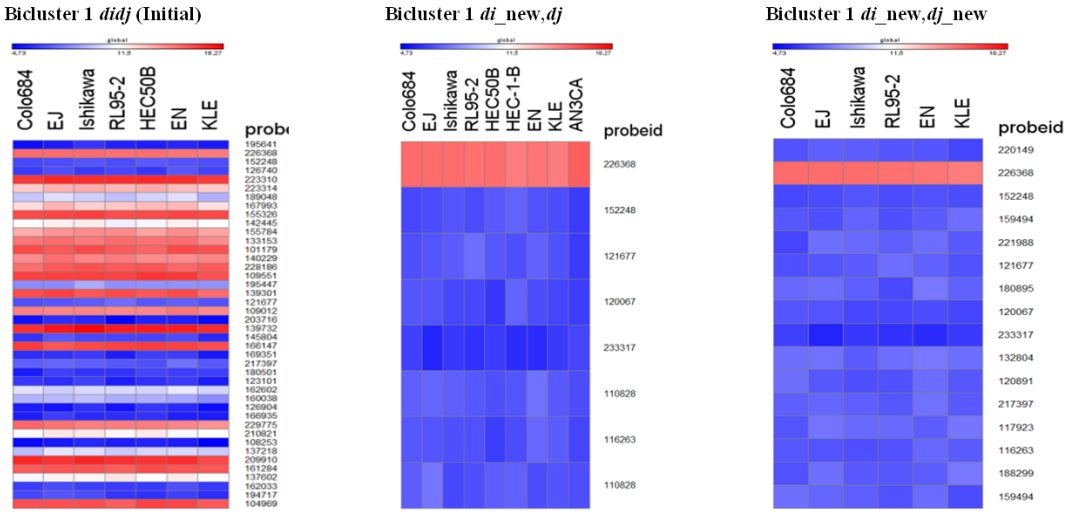

**Σχήμα 6.6.** - Οπτικοποίηση Αποτελεσμάτων Διπλής Κατηγοριοποίησης Υποχώρου Δεδομένων (1000 γονίδια - 9 κυτταρικές σειρές)
Καρκίνος του Τραχήλου της Μήτρας

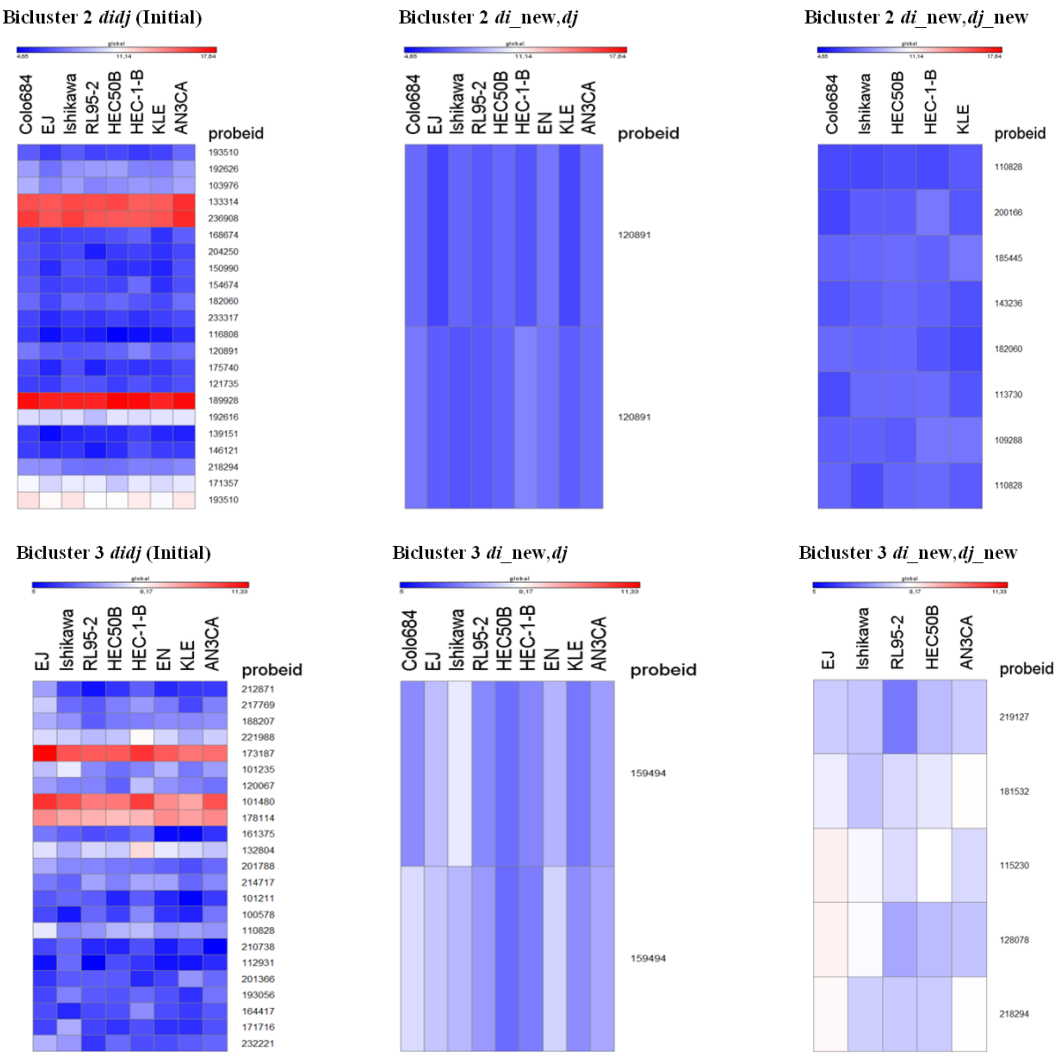





**Σχήμα 6.7.** – Οπτικοποίηση Αποτελεσμάτων Διπλής Κατηγοριοποίησης Υποχώρου Δεδομένων (1000 γονίδια - 9 κυτταρικές σειρές)

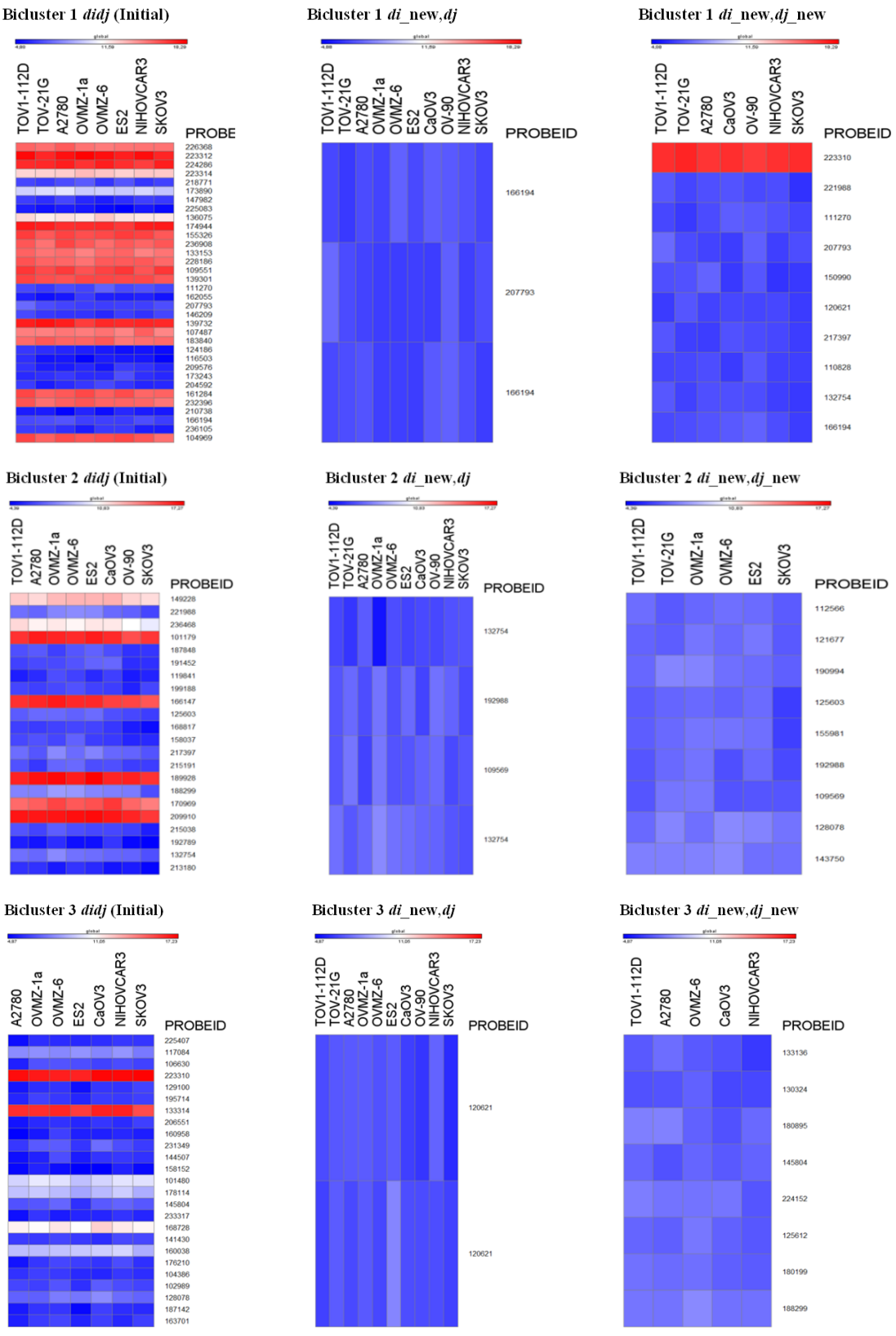
Καρκίνος του Ενδομητρίου



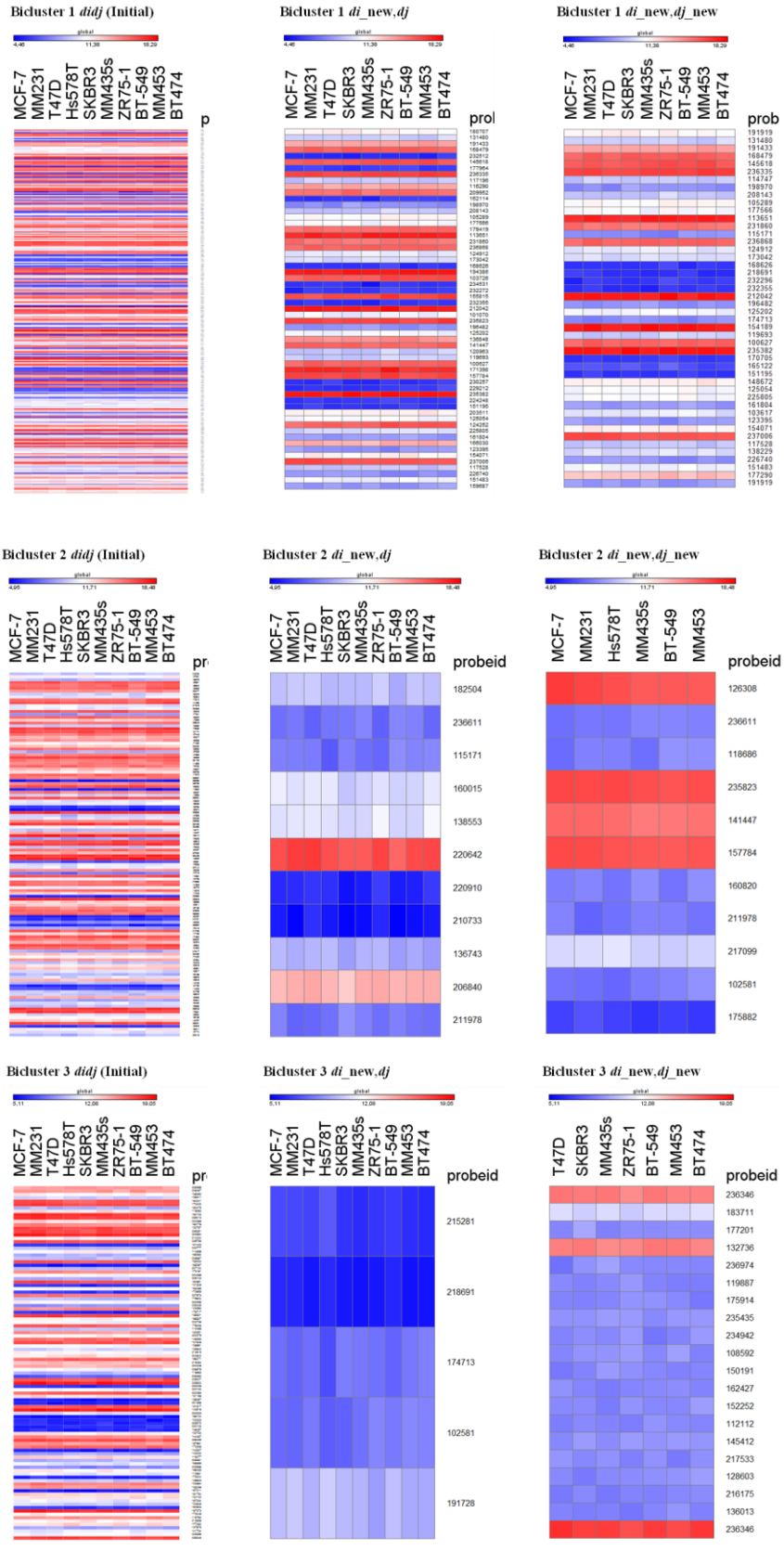



**Σχήμα 6.8.** - Οπτικοποίηση Αποτελεσμάτων Διπλής Κατηγοριοποίησης Υποχώρου Δεδομένων (1000 γονίδια - 10 κυτταρικές σειρές)

Καρκίνος των Ωοθηκών




**Σχήμα 6.9.** - Διπλή Κατηγοριοποίηση Συνόλου Δεδομένων (33096 γονίδια - 10 κυτταρικές σειρές) Καρκίνος του Μαστού


Bicluster 1 didj (Initial)
MCF-7
MM231
T47D
Hs578T
SKBR3
MM435s
ZR75-1
BT-549
MM453
BT474
Bicluster 1 di_new,dj
prol
Bicluster 1 di_new,dj_new
prob
Bicluster 2 didj (Initial)
Bicluster 2 di_new,dj
probeid
182504
236611
115171
160015
138553
220642
220910
210733
136743
206840
211978
Bicluster 2 di_new,dj_new
126308
236611
118686
235823
141447
157784
160820
211978
217099
102581
175882
Bicluster 3 didj (Initial)
Bicluster 3 di_new,dj
215281
218691
174713
102581
191728
Bicluster 3 di_new,dj_new
236346
183711
177201
132736
236974
119887
175914
235435
234942
108592
150191
162427
152252
112112
145412
217533
128603
216175
136013

**Πίνακας 6.2.** - Παρουσίαση των εμπλουτισμένων βιολογικών μονοπατιών ($p \leq 0.05$) για τα τρία Biclusters κατά την διπλή κατηγοριοποίηση του συνόλου των δεδομένων για τον καρκίνο του μαστού

**ΒΙΟΛΟΓΙΚΑ ΜΟΝΟΠΑΤΙΑ**

**Καρκίνος του Μαστού - Διπλή Κατηγοριοποίηση Συνόλου Δεδομένων**

**(33096 γονίδια – 10 κυτταρικές σειρές)**

| | |
|---|---|
| **Bicluster 1** ***didj*** | Ribosome, Spliceosome, RNA transport, Ribosome biogenesis in eukaryotes, One carbon pool by folate, Ubiquitin mediated proteolysis, Purine metabolism, Pyrimidine metabolism, ABC transporters, Oxidative phosphorylation, Metabolic pathways |
| **Bicluster 1** ***di_new,dj*** | Ribosome, **Endocytosis** |
| **Bicluster 1** ***di_new,dj_new*** | Ribosome, Spliceosome, **Endocytosis** |
| **Bicluster 2** ***didj*** | Ribosome, Spliceosome, Oxidative phosphorylation, RNA transport, Protein processing in endoplasmic reticulum, Metabolic pathways, Ubiquitin mediated proteolysis |
| **Bicluster 2** ***di_new,dj*** | - |
| **Bicluster 2** ***di_new,dj_new*** | Ribosome |
| **Bicluster 3** ***didj*** | Ribosome, Spliceosome, Proteasome, Peroxisome |
| **Bicluster 3** ***di_new,dj*** | - |
| **Bicluster 3** ***di_new,dj_new*** | - |

**Πίνακας 6.3.** - Παρουσίαση των εμπλουτισμένων βιολογικών διεργασιών ($p \leq 0.05$) για το δεύτερο Bicluster κατά την διπλή κατηγοριοποίηση του συνόλου των δεδομένων για τον καρκίνο του μαστού

## ΒΙΟΛΟΓΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ

**Καρκίνος του Μαστού - Διπλή Κατηγοριοποίηση Συνόλου Δεδομένων (33096 γονίδια - 10 κυτταρικές σειρές) - Bicluster 2**

| **Βιολογικές Διεργασίες Bcl2-BC *didj*** | |
|---|---|
| intracellular protein transport | GO:0006886 |
| mRNA metabolic process | GO:0016071 |
| viral reproduction | GO:0016032 |
| intracellular transport | GO:0046907 |
| cellular protein localization | GO:0034613 |
| cellular macromolecule localization | GO:0070727 |
| multi-organism reproductive process | GO:0044703 |
| macromolecule localization | GO:0033036 |
| **Βιολογικές Διεργασίες BC-Bcl2 *di* _new,dj** | |
| nucleobase-containing compound biosynthetic process | GO:0034654 |
| RNA biosynthetic process | GO:0032774 |
| nucleobase-containing compound metabolic process | GO:0006139 |
| RNA metabolic process | GO:0016070 |
| organic cyclic compound metabolic process | GO:1901360 |
| aromatic compound biosynthetic process | GO:0019438 |
| cellular aromatic compound metabolic process | GO:0006725 |
| cellular nitrogen compound metabolic process | GO:0034641 |
| organic cyclic compound biosynthetic process | GO:1901362 |
| heterocycle biosynthetic process | GO:0018130 |
| gene expression | GO:0010467 |
| heterocycle metabolic process | GO:0046483 |
| cellular nitrogen compound biosynthetic process | GO:0044271 |
| nitrogen compound metabolic process | GO:0006807 |
| nucleic acid metabolic process | GO:0090304 |
| virus-host interaction | GO:0019048 |
| cellular macromolecule metabolic process | GO:0044260 |
| cellular macromolecule biosynthetic process | GO:0034645 |
| macromolecule biosynthetic process | GO:0009059 |
| interaction with host | GO:0051701 |
| interspecies interaction between organisms | GO:0044419 |
| symbiosis, encompassing mutualism through parasitism | GO:0044403 |
| **Βιολογικές Διεργασίες BC-Bcl2 *di* _new,*dj* _new** | |
| cellular macromolecular complex subunit organization | GO:0034621 |
| regulation of translation | GO:0006417 |
| protein complex subunit organization | GO:0071822 |
| macromolecular complex subunit organization | GO:0043933 |
| regulation of cellular protein metabolic process | GO:0032268 |
| posttranscriptional regulation of gene expression | GO:0010608 |
| regulation of protein metabolic process | GO:0051246 |
| cellular process involved in reproduction | GO:0048610 |

| **Κοινές Βιολογικές Διεργασίες μεταξύ BC-Bcl2 *didj* και BC-Bcl2 *di* _new,*dj* _new** | |
|---|---|
| translational elongation | GO:0006414 |
| translational initiation | GO:0006413 |
| translational termination | GO:0006415 |
| cotranslational protein targeting to membrane | GO:0006613 |
| SRP-dependent cotranslational protein targeting to membrane | GO:0006614 |
| protein targeting to ER | GO:0045047 |
| establishment of protein localization to endoplasmic reticulum | GO:0072599 |
| translation | GO:0006412 |
| viral transcription | GO:0019083 |
| viral genome expression | GO:0019080 |
| nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | GO:0000184 |
| protein localization to endoplasmic reticulum | GO:0070972 |
| establishment of protein localization to organelle | GO:0072594 |
| viral infectious cycle | GO:0019058 |
| nuclear-transcribed mRNA catabolic process | GO:0000956 |
| mRNA catabolic process | GO:0006402 |
| protein targeting to membrane | GO:0006612 |
| cellular protein complex disassembly | GO:0043624 |
| viral reproductive process | GO:0022415 |
| protein targeting | GO:0006605 |
| RNA catabolic process | GO:0006401 |
| protein complex disassembly | GO:0043241 |
| protein localization to organelle | GO:0033365 |
| cellular macromolecular complex disassembly | GO:0034623 |
| macromolecular complex disassembly | GO:0032984 |
| cellular component disassembly at cellular level | GO:0071845 |
| cellular component disassembly | GO:0022411 |
| protein transport | GO:0015031 |
| establishment of protein localization | GO:0045184 |
| protein localization | GO:0008104 |
| cellular protein metabolic process | GO:0044267 |
| establishment of localization in cell | GO:0051649 |

# 7

# ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΈΣ ΕΠΕΚΤΑΣΕΙΣ

## 7.1 Συμπεράσματα

Στην παρούσα εργασία μελετήσαμε τις μεθόδους διπλής κατηγοριοποίησης σε πίνακες γονιδιακής έκφρασης για την εξαγωγή υποομάδων με στατιστική και βιολογική σημασία.

Η πειραματική διαδικασία εφαρμόστηκε σε αρχεία που περιλάμβαναν πίνακες γονιδίων-κυτταρικών σειρών προερχόμενα από μικροσυστοιχίες DNA από καρκινικές κυτταρικές σειρές του μαστού, του τραχήλου της μήτρας, των ωοθηκών και του ενδομητρίου. Τα αποτελέσματα που λάβαμε ήταν σημαντικά και επικυρώθηκαν μέσω της οπτικοποίησής τους.

Συγκεκριμένα:

Ο βελτιωμένος biclustering αλγόριθμος που χρησιμοποιήθηκε απέδειξε τους λόγους για τους οποίους, ο αρχικός αλγόριθμος έχριζε βελτίωσης και ότι ο βελτιωμένος αλγόριθμος διπλής κατηγοριοποίησης υπερτερεί, δίνοντας έμφαση στην πληροφορία, μειώνοντας το θόρυβο και παρέχοντας μας μια εικόνα για το σύνολο δεδομένων πιο συγκεκριμένη και λεπτομερής.

Σε αυτό το σημείο είναι σημαντικό να αναφερθεί το γεγονός ότι ο βελτιωμένος αλγόριθμος που προτείναμε, φαίνεται να είναι πιο αποδοτικός και να λειτουργεί καλύτερα στον υποχώρο γονιδίων. Η εξαγωγή ενός μεγάλου αριθμού γονιδίων σε μία ομάδα διπλής κατηγοριοποίησης (Biclusters) δυσκολεύει το έργο του αλγορίθμου και δεν έχουμε τόσο ικανοποιητικά αποτελέσματα.

Όσον αφορά τον υποχώρο γονιδίων, 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.

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

# Βιβλιογραφία


[1] Tanay A., Sharan R., Shamir R., "Discovering statistically significant biclusters in gene expression data", Bioinformatics, 18 Suppl 1:S136-44. 2002.
[2] Saber H.B., and Elloumi M., "A comparative study of clustering and biclustering of microarray data", International Journal of Computer Science & Information Technology (IJCSIT), vol. 6, no. 6, pp. 93-111, 2014.
[3] Graphical Abstract: http://dx.doi.org/10.1016/j.jbi.2015.06.028.
[4] Cheng Y., and Church G.M., "Biclustering of Expression Data" Proc Int Conf Intell Syst Mol Biol., vol. 8, pp. 93-103, 2000.
[5] Alberts B., "Molecular biology of the cell", 5th ed. New York: Garland Science, 2002.
[6] Hartigan J.A., "Clustering Algorithms", John Wiley & Sons, New York, 1975.
[7] Mirkin B., "Mathematical Classification and Clustering", Kluwer Academic Publishers, 1996. ISBN 0-7923-4159-7.
[8] Tanay A., Sharon R., and Shamir R., "Biclustering gene expression data", In Proceedings of ISMB 2002, pp. S136–S144, 2002.
[9] Ihmels J. Bergmann S. Barkai N. Defining transcriptional modules using large-scale gene expression data. Bioinformatics., vol. 20, no. 13, pp. 1993–2003, 2004.
[10] Madeira S.C., and Oliveira A.L., "Biclustering algorithms for biological data analysis: a survey", IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 1, no. 1, pp. 24–45, 2004.
[11] Pontes B., Giráldez R., Aguilar-Ruiz J.S., "Biclustering on expression data: A review", Journal of Biomedical Informatics, vol. 57, pp. 163-180, 2015.
[12] Αλεβυζάκη Ανδρονίκη, «Μέθοδοι διπλής κατηγοριοποίησης "biclustering" για επιλογή γονιδιακών δεικτών από κυτταρικές σειρές». Διπλωματική Εργασία, Πολυτεχνείο Κρήτης, 2010.
[13] Tang A.H., Neufeld T.P., Rubin G.M., Müller H.-A.J., "Transcriptional regulation of cytoskeletal functions and segmentation by a novel maternal pair-rule gene, lilliputian", Development vol. 128, pp. 801-813, 2001.
[14] Ben-Dor A., Chor B., Karp R., and Yakhini Z., "Discovering local structure in gene expression data: the order-preserving submatrix problem", J. Comput. Biol. vol. 10, no. 3-4, pp. 373–384, 2003.
[15] Segal E., Battle A., and Koller D., "Decomposing gene expression into cellular processes", Pac. Symp. Biocomput. vol. 8, pp. 89–100, 2003.
[16] Murali T.M., and Kasif S., "Extracting conserved gene expression motifs from gene expression data", Pac. Symp. Biocomput. vol. 8, pp. 77–88, 2003.
[17] Sheng Q., Moreau Y., and De Moor B., "Biclustering micrarray data by gibbs sampling", Bioinformatics, vol. 19, no. Suppl.2, p. ii196–ii205, 2003.
[18] Cano C., Adarve L., López J., and Blanco A., "Possibilistic approach for biclustering microarray data", Comput Biol Med., vol. 37, no. 10, pp. 1426–1436, 2007.
[19] Maulik U., Mukhopadhyay A., and Bandyopadhyay S., "Finding multiple coherent biclusters in microarray data using variable string length multiobjective genetic algorithm", IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 6, pp. 969–975, 2009.
[20] Chen C.-P., Fushing H., Atwill R., and Koehl P., "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, 2014.
[21] https://gr.dreamstime.com/ [image24298020]
[22] Lodish H., Berk A., Matsudaira P., Kaiser C.A., Krieger M., Scott M.P., Zipursky L., and Darnell J., "Molecular Cell Biology", 5th edition. W.H. Freeman & Co (Sd), 2003.
[23] http://ebooks.edu.gr/
[24] https://www.slideshare.net/annpyl/dna-rna
[25] http://mathimatabiologias.weebly.com/uploads/1/0/9/7/10979738/313406_orig.png
[26] Βιοπληροφορική-αναζήτηση πληροφορίας σε βιολογικές βάσεις δεδομένων: http://ecourses.dbnet.ntua.gr/fsr/17066/ERGASTHRIAKH%20ASKHSH%202_sumplhrwmatiko.pdf
[27] http://www.chem.uoa.gr/?cat=39
[28] http://www.sciencedirect.com/topics/page/Transcriptome
[29] https://repository.kallipos.gr/bitstream/11419/1585/1/Chapter07_geneexpression_R.pdf
[30] Μαλατράς Α., «Εφαρμογές Υπολογιστικής Νοημοσύνης σε Μικροσυστοιχίες DNA», Πτυχιακή εργασία. Πανεπιστήμιο Στερεάς Ελλάδας, 2010.
[31] microarrays dna_images

[32] Jiang D., Tang C., and Zhang A., "Cluster Analysis for Gene Expression Data: A Survey" IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370-1386, 2004.
[33] Brazma A., Parkinson H., Sarkans U., Shojatalab M., Vilo J., Abeygunawardena N., Holloway E., Kapushesky M., Kemmeren P., Lara G.G., Oezcimen A., Rocca-Serra P., Sansone S.-A., ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res., vol. 31, no. 1, pp. 68-71, 2003.
[34] Lee M.L.T., "Analysis of Microarray Gene Expression Data", Springer, 2004.
[35] Zhao H., Liew A.W.-C., Wang D.Z., and Yan, H. Biclustering analysis for pattern discovery: Current techniques, comparative studies and applications. Current Bioinformatics, vol. 7, no. 1, pp. 43-55, 2012.
[36] http://www.eng.ucy.ac.cy/cpitris/courses/ECE370/presentations/14.%20Biomaterials.pdf
[37] http://eclass.uth.gr/eclass/modules/document/file.php/DIB200/chapter2final.pdf
[38] https://www.ncbi.nlm.nih.gov/genbank/
[39] https://www.nlm.nih.gov/bsd/pmresources.html
[40] https://www.ncbi.nlm.nih.gov/geo/
[41] GenBank Statistics
[42] http://students.ceid.upatras.gr/~maurakis/Gene_Expression_Analysis.pdf
[43] Agrawal R., Gehrke J., Gunopulos D.. and Raghavan P., "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data, vol. 27, no. 2, pp. 94-105, 1998.
[44] Eisen M.B., Spellman P.T., Brown P.O., and Botstein D., "Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USA, vol. 95, no. 25, pp. 14863– 14868, 1998.
[45] https://www.slideshare.net/raedald/automatic-subspace-clustering-of-high-dimensional-data-for-data-mining-application-32900971
[46] Tchagang A.B., Pan Y., Famili F., Tewfik A.H.,and Benos P.V., "Biclustering of DNA Microarray Data: Theory, Evaluation, and Applications", Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications, IGI Global, 2011.
[47] Sharan R., Porat U.B., and Bleiberg O., "Analysis of Biological Networks: Network Modules– Clustering and Biclustering ", Lecture 5, November 23 , pp. 1–21, 2006.
[48] http://www.microarrays.ca/services/,"What Is Hierarchical Clustering ?".
[49] http://homes.di.unimi.it/~valenti/SlideCorsi/MB0910/HierarchicalClustering.pdf
[50] Maimon O., and Rokach L., "Clustering methods", in Data mining and knowledge discovery handbook. Springer US, pp. 321-352, 2005.
[51] Manning C.D., Raghavan P., and Schütze H., "Introduction to Information Retrieval", Cambridge University Press, 2008.
[52] https://www.google.gr/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwiJj9a99szUAhVFNxQKHUy4DDYQjxwIAw&url=https%3A%2F%2Fwww.slideshare.net%2Fgveress%2Fcluster-training-2013&psig=AFQjCNGxX6_dkRmkENDcCdFObjYojWyCsA&ust=1498065313474950
[53] https://www.google.gr/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0ahUKEwiJj9a99szUAhVFNxQKHUy4DDYQjxwIAw&url=https%3A%2F%2Fwww.slideshare.net%2Fgveress%2Fcluster-training-2013&psig=AFQjCNGxX6_dkRmkENDcCdFObjYojWyCsA&ust=1498065313474950
[54] https://en.wikipedia.org/wiki/Biclustering
[55] Blackshaw S., Harpavat S., Trimarchi J., Cai L., Huang H., Kuo W. P., Weber G., Brazma, and Vilo J., "Minireview: Gene Expression Data Analysis", FEBS Letters 480, pp. 17-24, 2003.
[56] Pontes B., Girldez R., and Aguilar-Ruiz J.S., "Quality Measures for Gene Expression Biclusters", Margis R, ed. PLoS ONE, vol. 10, no. 3, p. e0115497, 2015.
[57] Obermayr, Sanchez-Cabo E. F., Tea M. K. M., Singer C. F., Krainer M., Fischer M. B.,Sehouli J., Reinthaller A., Horvat R., Heinze G., Tong D., and Zeillinger R.,"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, 2010.
[58] NCI Dictionary of Cancer Terms
[59] Cancer Cell Line Encyclopedia (CCLE)
[60] Jacob F., Nixdorf S., Hacker N.F., and Heinzelmann-Schwarz V.A., "Reliable in vitro studies require appropriate ovarian cancer cell lines.," J. Ovarian Res., vol. 7, p. 60, 2014.
[61] https://software.broadinstitute.org/GENE-E/
[62] https://software.broadinstitute.org/GENE-E/doc.html
[63] http://www.webgestalt.org/webgestalt_2013/option.php

# ΠΑΡΑΡΤΗΜΑ

Στο Παράρτημα παρατίθενται τα γονίδια που ομαδοποιούνται στα τρία πρώτα Biclusters στην πρώτη, δεύτερη και τρίτη εκτέλεση του αλγορίθμου. Δίνονται οι αναλυτικοί Πίνακες για τον καρκίνο του μαστού, του τραχήλου της μήτρας, του ενδομητρίου και των ωοθηκών για τα 1000 γονίδια του "dataset".

## ***Καρκίνος του Μαστού***

### ***di,dj ->Πρώτη εκτέλεση αλγορίθμου***

| initial | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| di_dj | | | | | | | | | |
| bicluster1 | PROBEID | MCF-7 | MM231 | SKBR3 | MM435s | ZR75-1 | BT-549 | MM453 | BT474 |
| 18 | 226368 | 14,31719 | 14,95013 | 15,29338 | 14,83902 | 15,05665 | 15,05267 | 15,16847 | 14,47759 |
| 42 | 213726 | 9,975586 | 10,05667 | 10,67213 | 9,856768 | 10,26455 | 9,752279 | 10,56243 | 10,2598 |
| 105 | 224284 | 16,35968 | 15,97745 | 16,04696 | 15,87873 | 16,51371 | 15,88931 | 15,78106 | 16,05529 |
| 107 | 224286 | 17,70126 | 16,79674 | 17,45814 | 17,09883 | 16,8679 | 17,45312 | 17,58593 | 17,00214 |
| 120 | 199782 | 5,545611 | 6,226255 | 5,671115 | 5,716348 | 6,237167 | 5,131927 | 6,104767 | 5,46478 |
| 150 | 173890 | 10,10765 | 10,45769 | 10,24529 | 9,928914 | 10,17369 | 10,70491 | 10,61749 | 10,41654 |
| 153 | 101695 | 6,072557 | 6,397267 | 6,426852 | 5,952134 | 6,356477 | 5,687495 | 5,634138 | 5,790055 |
| 177 | 147346 | 6,395415 | 7,038766 | 6,417404 | 6,198541 | 6,31326 | 6,483754 | 6,572028 | 6,190864 |
| 190 | 138543 | 12,37473 | 11,7103 | 12,21802 | 11,83362 | 12,77035 | 11,55955 | 12,40636 | 11,74627 |
| 259 | 186673 | 10,62225 | 10,19384 | 10,20132 | 9,946289 | 10,20132 | 9,702818 | 10,07868 | 10,18312 |
| 273 | 174944 | 16,70533 | 17,20289 | 16,62122 | 17,35966 | 16,98538 | 17,25555 | 17,22564 | 16,93257 |
| 289 | 155326 | 15,78106 | 16,62461 | 16,47307 | 16,3555 | 16,12322 | 16,70292 | 16,22788 | 16,50926 |
| 303 | 236908 | 14,95278 | 15,50219 | 15,37988 | 15,76231 | 15,59826 | 15,56187 | 15,64221 | 15,42449 |
| 310 | 155784 | 13,82736 | 14,08613 | 14,82038 | 13,9083 | 13,71764 | 14,05261 | 14,49957 | 13,93928 |
| 320 | 149734 | 16,62461 | 16,66969 | 16,57633 | 17,08047 | 16,81259 | 16,99287 | 17,00442 | 16,77015 |
| 336 | 228186 | 15,59826 | 16,19141 | 15,98615 | 15,22927 | 15,32926 | 15,40191 | 15,65025 | 15,56187 |
| 343 | 139301 | 16,29054 | 16,31057 | 16,15504 | 16,53159 | 16,44803 | 15,86532 | 16,26508 | 16,65611 |
| 387 | 180707 | 11,68976 | 11,80061 | 12,13062 | 11,66029 | 12,03644 | 11,58405 | 11,77812 | 11,39865 |
| 390 | 128561 | 6,221719 | 5,762957 | 5,561125 | 5,375277 | 5,711893 | 6,151954 | 5,480831 | 5,835731 |
| 438 | 181532 | 8,373335 | 8,207122 | 8,078394 | 7,736715 | 8,757705 | 7,60793 | 8,076276 | 7,838595 |
| 499 | 182504 | 10,07569 | 10,3546 | 9,93524 | 9,913831 | 10,50159 | 9,463651 | 10,23471 | 10,09346 |
| 563 | 139732 | 17,923 | 16,95017 | 17,80707 | 17,40093 | 17,64809 | 17,4688 | 17,55437 | 17,27832 |
| 606 | 116808 | 5,829279 | 5,701691 | 5,020673 | 5,41716 | 5,874409 | 5,389694 | 5,113216 | 5,655472 |
| 678 | 180501 | 6,377281 | 6,885293 | 6,627092 | 6,535787 | 6,513497 | 6,610454 | 6,575171 | 6,590028 |
| 718 | 214717 | 6,876456 | 6,768432 | 7,275703 | 6,26037 | 6,673344 | 6,873969 | 6,747133 | 6,620442 |
| 730 | 236216 | 15,04358 | 15,14456 | 15,44834 | 15,33159 | 15,53159 | 15,2265 | 15,69976 | 14,99796 |
| 792 | 215964 | 6,148683 | 5,965177 | 6,433183 | 5,690075 | 6,229993 | 5,722257 | 6,421423 | 6,589592 |
| 795 | 128192 | 5,636823 | 5,939517 | 5,998356 | 6,12362 | 6,028395 | 6,584104 | 6,61524 | 6,173348 |
| 801 | 170969 | 15,31567 | 15,7042 | 15,79074 | 15,8865 | 15,31334 | 15,91738 | 15,63375 | 15,64662 |
| 803 | 229775 | 14,17414 | 14,85701 | 14,69149 | 13,97004 | 14,42462 | 14,37224 | 14,43587 | 13,5761 |
| 818 | 161284 | 16,14322 | 16,2218 | 16,10162 | 15,56673 | 15,69749 | 15,01808 | 15,39034 | 15,4559 |
| 845 | 139151 | 6,155465 | 5,426909 | 6,012651 | 5,732012 | 6,499128 | 5,648874 | 6,208971 | 5,703918 |
| 909 | 108780 | 5,605265 | 6,572742 | 6,000083 | 5,215229 | 6,237474 | 5,52839 | 5,673374 | 5,401041 |
| 915 | 109210 | 6,659545 | 6,387754 | 6,096934 | 5,33436 | 6,018268 | 5,809462 | 6,126813 | 5,99687 |
| 978 | 104969 | 15,53605 | 16,42501 | 16,4331 | 16,22309 | 15,99071 | 16,41701 | 15,73791 | 15,95591 |
| 994 | 118956 | 15,32926 | 15,90111 | 16,19141 | 15,55201 | 15,58661 | 15,69231 | 15,87653 | 14,98848 |
| [1,2,5,6,7,8,9,10] | | | | | | | | | |

| Bicluster2 | Probeid | MCF-7 | MM231 | SKBR3 | MM435s | ZR75-1 | BT-549 | MM453 | BT474 |
|---|---|---|---|---|---|---|---|---|---|
| 47 | 164432 | 14,46505 | 13,87184 | 14,46997 | 14,11255 | 13,7036 | 14,00062 | 14,07032 | 14,88423 |
| 69 | 151787 | 8,635632 | 7,239536 | 8,52605 | 7,634946 | 7,077521 | 7,59948 | 7,89729 | 7,779308 |
| 108 | 223314 | 13,5969 | 13,12667 | 13,84732 | 13,09266 | 12,91258 | 13,36127 | 13,387 | 13,83196 |
| 146 | 161079 | 8,747381 | 7,80924 | 8,677046 | 7,145436 | 7,992976 | 7,929998 | 8,043446 | 8,517164 |
| 149 | 140336 | 12,61235 | 11,4145 | 12,40398 | 11,42642 | 11,12099 | 11,96932 | 11,29658 | 12,58113 |
| 187 | 127051 | 12,24356 | 11,5182 | 12,28037 | 11,56633 | 11,72818 | 11,77746 | 11,61646 | 12,47905 |
| 193 | 137543 | 5,568 | 4,812809 | 5,485596 | 5,130313 | 4,920732 | 5,859754 | 5,432801 | 6,11472 |
| 196 | 160701 | 11,81033 | 11,7453 | 11,7571 | 11,38507 | 11,0268 | 11,76741 | 12,11732 | 11,76157 |
| 244 | 217593 | 13,41063 | 13,18044 | 13,9013 | 12,99592 | 12,38513 | 13,42249 | 13,64528 | 13,03966 |
| 256 | 136075 | 12,90104 | 11,69179 | 12,77364 | 11,63942 | 12,1696 | 12,07241 | 11,80657 | 12,04721 |
| 296 | 230085 | 14,89965 | 14,21666 | 14,07188 | 13,54719 | 13,45154 | 13,97161 | 14,07841 | 13,96269 |
| 503 | 123356 | 9,555903 | 9,573616 | 9,699554 | 8,893515 | 8,768327 | 9,23255 | 9,70536 | 9,550708 |
| 538 | 151430 | 6,002187 | 5,542083 | 6,075485 | 5,17647 | 5,541288 | 6,146669 | 5,819265 | 6,723677 |
| 540 | 178114 | 9,462366 | 8,730112 | 10,14004 | 8,890063 | 9,030732 | 9,643751 | 9,347486 | 9,746303 |
| 587 | 182060 | 7,347799 | 6,167629 | 6,248132 | 6,476599 | 5,644152 | 6,329511 | 6,667814 | 6,868262 |
| 655 | 149009 | 7,197602 | 6,459343 | 7,373792 | 6,023935 | 5,870365 | 7,385542 | 6,382085 | 7,023492 |
| 770 | 216613 | 11,88129 | 10,7117 | 11,05489 | 10,17339 | 9,921691 | 10,63467 | 11,23681 | 11,24038 |
| 814 | 194406 | 11,72418 | 10,8154 | 11,52313 | 10,70896 | 11,48223 | 11,09369 | 11,97138 | 11,69343 |
| 825 | 121364 | 15,49235 | 14,03838 | 14,94737 | 13,93928 | 14,01721 | 14,50678 | 15,0384 | 14,35543 |
| 941 | 136278 | 6,931848 | 5,631319 | 7,093789 | 5,628536 | 5,816121 | 6,51293 | 6,126555 | 6,671591 |
| [1,2,3,4,6,8,9,10] | | | | | | | | | |

| Bicluster3 | probeid | T47D | Hs578T | MM435s | ZR75-1 | BT-549 | MM453 |
|---|---|---|---|---|---|---|---|
| 6 | 217298 | 6,253944 | 6,656327 | 5,898956 | 7,070186 | 6,277799 | 6,412153 |
| 159 | 165735 | 5,735893 | 5,567824 | 5,136867 | 6,532075 | 5,005613 | 5,584856 |
| 238 | 223080 | 12,15183 | 12,22873 | 12,22808 | 13,02781 | 11,65189 | 12,70626 |
| 240 | 225083 | 5,909558 | 6,096884 | 5,248097 | 6,421529 | 6,059 | 6,090798 |
| 263 | 136093 | 5,510129 | 6,165769 | 5,198784 | 6,044713 | 5,731562 | 5,914829 |
| 313 | 101179 | 15,97745 | 15,69485 | 16,3966 | 16,47307 | 15,34586 | 16,70081 |
| 315 | 156066 | 15,97745 | 15,69485 | 16,3966 | 16,47307 | 15,34586 | 16,70081 |
| 321 | 180730 | 10,96456 | 11,09423 | 11,26086 | 11,61886 | 11,05733 | 11,40091 |
| 324 | 154584 | 5,611654 | 5,744894 | 6,053406 | 6,471021 | 5,034153 | 5,896183 |
| 326 | 140229 | 15,13648 | 15,05267 | 14,78292 | 15,33382 | 15,02597 | 15,29581 |
| 354 | 173299 | 15,23826 | 14,64175 | 14,84127 | 14,87739 | 14,64782 | 15,0492 |
| 404 | 188198 | 6,365204 | 6,147532 | 6,823609 | 6,452447 | 6,273496 | 6,216863 |
| 411 | 109012 | 14,35402 | 13,98778 | 14,50318 | 14,81061 | 13,8092 | 14,54862 |
| 550 | 200313 | 6,288208 | 6,681701 | 6,234391 | 7,102128 | 6,236535 | 5,903964 |
| 617 | 120825 | 8,398501 | 9,320165 | 8,55609 | 9,322817 | 8,375298 | 8,854036 |
| 632 | 183840 | 15,67562 | 15,32926 | 15,63375 | 16,48492 | 15,02964 | 15,95401 |
| 635 | 174851 | 4,944075 | 5,343539 | 5,132999 | 6,102083 | 5,02972 | 5,248587 |
| 665 | 182708 | 11,9843 | 12,16154 | 12,47628 | 13,03966 | 11,7103 | 12,52424 |
| 711 | 123369 | 6,325246 | 6,571876 | 6,047078 | 6,575121 | 6,012033 | 5,698496 |
| 716 | 120007 | 5,732923 | 5,229325 | 5,968539 | 5,75379 | 5,780379 | 5,631384 |
| 743 | **202153** | 6,766566 | 6,102506 | 6,314251 | 6,713627 | 6,446853 | 5,849025 |
| 774 | 189928 | 17,13934 | 17,43537 | 16,99287 | 17,45814 | 17,09883 | 16,86047 |
| 813 | 209910 | 15,85037 | 15,71945 | 15,89665 | 15,91533 | 16,26755 | 15,30499 |
| 839 | 232396 | 15,85037 | 15,71945 | 15,89665 | 15,91533 | 16,26755 | 15,30499 |
| 853 | 163565 | 6,68222 | 7,602164 | 6,686453 | 7,84799 | 6,809898 | 7,113893 |
| 863 | 104468 | 4,73585 | 5,759724 | 5,468223 | 6,154048 | 5,047554 | 5,377096 |
| 888 | 201366 | 6,079344 | 5,622237 | 6,424458 | 6,177023 | 5,565188 | 5,843548 |
| [3,4,6,7,8,9] | | | | | | | |

## *Καρκίνος του Μαστού*

### *di_new,dj ->Δεύτερη εκτέλεση αλγορίθμου*

| dinew_dj | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bicluster1 | probeid | MCF-7 | MM231 | T47D | Hs578T | SKBR3 | MM435s | ZR75-1 | BT-549 | MM453 |
| 343 | 139301 | 16,66128 | 16,30631 | 16,19141 | 16,53715 | 16,51115 | 16,45829 | 15,13869 | 16,64168 | 16,49421 |
| 462 | 214589 | 7,734273 | 7,289839 | 5,399297 | 8,090775 | 5,751375 | 6,179302 | 7,512713 | 7,718731 | 6,661609 |
| 678 | 180501 | 5,762265 | 5,173406 | 6,836401 | 6,451068 | 5,851414 | 4,637098 | 6,353068 | 7,501148 | 5,874923 |
| 718 | 214717 | 6,145841 | 6,44112 | 6,568804 | 7,100972 | 7,053609 | 5,376219 | 6,428902 | 6,670218 | 6,255565 |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | |

| Bicluster2 | probeid | MCF-7 | MM231 | T47D | Hs578T | SKBR3 | MM435s | ZR75-1 | BT-549 | MM453 |
|---|---|---|---|---|---|---|---|---|---|---|
| 395 | 162744 | 9,219379 | 8,467646 | 8,538543 | 8,513284 | 7,785826 | 8,020587 | 8,526212 | 9,064433 | 7,122914 |
| 438 | 181532 | 8,680829 | 8,653343 | 8,929641 | 8,125524 | 7,542553 | 8,101936 | 7,047804 | 8,760486 | 7,231413 |
| 448 | 201489 | 8,487513 | 8,214824 | 7,499573 | 8,090161 | 8,74074 | 6,582001 | 7,408209 | 9,212658 | 8,513943 |
| 730 | 236216 | 15,44289 | 15,00473 | 15,16472 | 15,10782 | 14,93144 | 15,50422 | 15,20032 | 15,79866 | 16,25412 |
| 740 | 140456 | 4,842105 | 4,749718 | 5,81241 | 6,768801 | 6,30261 | 5,561439 | 5,471868 | 6,341218 | 5,695855 |
| 801 | 170969 | 14,65284 | 15,13456 | 14,95476 | 15,61031 | 15,13456 | 15,30316 | 15,91908 | 14,77876 | 14,77669 |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | |

| Bicluster3 | probeid | MCF-7 | MM231 | T47D | Hs578T | SKBR3 | MM435s | ZR75-1 | BT-549 | MM453 |
|---|---|---|---|---|---|---|---|---|---|---|
| 329 | 231349 | 6,45168 | 7,646295 | 6,229041 | 7,080557 | 7,118636 | 6,186273 | 7,554411 | 6,510242 | 6,787215 |
| 371 | 140490 | 6,96837 | 5,711136 | 7,125772 | 6,299203 | 7,263096 | 6,178563 | 5,543545 | 7,096224 | 6,536871 |
| | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | |

## *Καρκίνος του Μαστού*

### *di_new,dj_new ->Τρίτη εκτέλεση αλγορίθμου*

| di_new_dj_new | | | | | |
|---|---|---|---|---|---|
| Bicluster1 | probeid | MM231 | MM435s | BT-549 | MM453 |
| 245 | 219524 | 6,313778 | 7,104665 | 8,36813 | 7,289293 |
| 329 | 231349 | 7,646295 | 6,186273 | 6,510242 | 6,787215 |
| 363 | 111270 | 6,418241 | 6,835804 | 7,2975 | 7,075434 |
| 371 | 140490 | 5,711136 | 6,178563 | 7,096224 | 6,536871 |
| 462 | 214589 | 7,289839 | 6,179302 | 7,718731 | 6,661609 |
| 553 | 190994 | 7,780951 | 7,126229 | 8,510808 | 6,882617 |
| 678 | 180501 | 5,173406 | 4,637098 | 7,501148 | 5,874923 |
| 718 | 214717 | 6,44112 | 5,376219 | 6,670218 | 6,255565 |
| 853 | 163565 | 6,378258 | 7,346787 | 7,344067 | 5,988559 |
| | | | | | |
| [2,6,8,9] | | | | | |

| Bicluster2 | probeid | MM231 | T47D | BT-549 |
|---|---|---|---|---|
| 168 | 142719 | 7,175448 | 6,612224 | 7,457644 |
| 438 | 181532 | 8,653343 | 8,929641 | 8,760486 |
| 440 | 101235 | 7,004152 | 6,055474 | 7,531619 |
| 448 | 201489 | 8,214824 | 7,499573 | 9,212658 |
| 533 | 146732 | 5,869345 | 7,373126 | 7,896661 |
| 780 | 109288 | 6,449557 | 9,199736 | 6,710841 |
| 955 | 234158 | 4,497453 | 6,105809 | 7,016225 |
| | | | | |
| [2,3,8] | | | | |

| Bicluster3 | probeid | MCF-7 | T47D | Hs578T | MM453 |
|---|---|---|---|---|---|
| 69 | 151787 | 7,571363 | 6,983927 | 4,753833 | 6,488392 |
| 395 | 162744 | 9,219379 | 8,538543 | 8,513284 | 7,122914 |
| 602 | 125603 | 6,730458 | 7,083339 | 7,112675 | 7,299382 |
| 904 | 132754 | 7,238138 | 7,022943 | 7,891657 | 6,606438 |
| 945 | 115325 | 7,712005 | 8,998439 | 5,550856 | 7,611143 |
| | | | | | |
| [1,3,4,9] | | | | | |

## *Καρκίνος του τραχήλου της μήτρας*

## *di,dj ->Πρώτη εκτέλεση αλγορίθμου*

didj

| Bicluster1 | probeid | SW756 | GH354 | Hela | C-33A | CaSki | ME-180 | HT-3 | SiHa |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 224689 | 5,5039136 | 5,6486736 | 6,1566287 | 6,1559234 | 5,2309028 | 5,7479527 | 6,0509484 | 5,9452344 |
| 18 | 226368 | 15,556685 | 15,229272 | 14,868955 | 15,867796 | 15,553903 | 15,041903 | 14,885963 | 15,561874 |
| 38 | 165590 | 5,2989428 | 6,5041562 | 6,4269136 | 6,7561068 | 6,2148508 | 6,0657973 | 6,1691286 | 6,238279 |
| 44 | 106630 | 6,4156662 | 6,8901907 | 7,1092478 | 7,1890325 | 6,821831 | 6,5629594 | 6,8059 | 6,5970332 |
| 49 | 152249 | 7,2556324 | 7,4255726 | 7,2391865 | 7,1776018 | 6,675348 | 6,605961 | 7,1831391 | 7,0102056 |
| 92 | 116550 | 6,0934023 | 6,8134234 | 6,1966267 | 6,4284483 | 6,4643673 | 6,4093507 | 6,7781048 | 6,3648074 |
| 106 | 223310 | 16,761318 | 16,992873 | 17,150353 | 17,092611 | 17,283723 | 17,004424 | 16,615897 | 16,4656 |
| 159 | 165735 | 5,9391925 | 5,3989462 | 5,8140107 | 5,1945288 | 5,3873598 | 5,6457532 | 5,8606492 | 5,5600311 |
| 187 | 127051 | 12,383706 | 11,593436 | 11,634778 | 11,887015 | 12,211357 | 11,582455 | 12,012595 | 11,668011 |
| 210 | 124803 | 9,433284 | 9,5914623 | 9,5093646 | 10,464566 | 10,147551 | 9,9777943 | 9,7560875 | 10,149089 |
| 211 | 135807 | 10,185016 | 10,386035 | 9,6589095 | 10,324026 | 10,340597 | 10,141443 | 9,7408836 | 10,169177 |
| 244 | 217593 | 12,917612 | 13,462227 | 13,450594 | 13,519456 | 12,713426 | 12,934267 | 13,113168 | 13,410631 |
| 250 | 183855 | 5,3963434 | 5,208776 | 4,7043696 | 5,6580386 | 5,5821202 | 4,6901083 | 5,3912051 | 5,4629134 |
| 273 | 174944 | 17,077839 | 17,06098 | 16,820318 | 17,364681 | 17,239012 | 16,627222 | 17,024784 | 17,294955 |
| 289 | 155326 | 16,037954 | 16,876223 | 16,876223 | 16,867904 | 16,15016 | 16,371572 | 16,491387 | 15,986147 |
| 303 | 236908 | 15,600377 | 15,669585 | 15,356016 | 16,761318 | 15,821713 | 15,692309 | 15,930585 | 15,878731 |
| 310 | 155784 | 13,539745 | 14,394271 | 14,843002 | 14,701187 | 14,307043 | 13,762306 | 14,457408 | 14,257158 |
| 312 | 133153 | 15,220577 | 15,094621 | 15,44289 | 15,57243 | 15,25053 | 15,130495 | 15,867796 | 15,216065 |
| 313 | 101179 | 15,834472 | 15,732997 | 16,369172 | 16,803813 | 15,773393 | 15,629042 | 16,080373 | 15,721235 |
| 336 | 228186 | 15,553903 | 15,661304 | 15,36272 | 15,527548 | 15,588929 | 15,025969 | 14,979705 | 15,467452 |
| 338 | 109551 | 16,615897 | 16,608639 | 16,862559 | 17,161044 | 17,153337 | 16,739178 | 17,637396 | 16,739178 |
| 343 | 139301 | 16,135774 | 17,074444 | 16,767475 | 16,50716 | 16,322663 | 16,596283 | 16,322663 | 16,06409 |
| 409 | 203356 | 6,0692505 | 5,8110248 | 6,3343603 | 5,7544304 | 6,1732642 | 5,7170775 | 6,3545069 | 6,3200238 |
| 411 | 109012 | 14,895508 | 14,500874 | 14,675607 | 14,918098 | 14,872804 | 14,125618 | 14,408398 | 14,681657 |
| 505 | 119841 | 5,7602139 | 6,0202987 | 6,2197137 | 6,0844699 | 5,8939242 | 5,3841271 | 5,7002791 | 6,4295147 |
| 563 | 139732 | 17,601485 | 17,431223 | 17,491936 | 18,19899 | 17,619068 | 17,777822 | 17,109196 | 17,835716 |
| 600 | 161948 | 6,1175966 | 5,803934 | 6,3972671 | 6,0628585 | 5,9259994 | 5,8124948 | 5,8708384 | 6,683545 |
| 658 | 169351 | 5,7231058 | 5,4487488 | 5,9394524 | 6,3057402 | 5,8341317 | 5,4881615 | 5,4302501 | 5,9826754 |
| 719 | 162602 | 10,050049 | 10,312374 | 10,361391 | 10,523458 | 10,374621 | 10,25464 | 10,202138 | 10,129886 |
| 772 | 101481 | 5,4471093 | 5,9642345 | 5,8457607 | 6,1359576 | 6,0876415 | 6,0641783 | 6,1406199 | 5,9407602 |
| 774 | 189928 | 16,165163 | 16,409861 | 16,270306 | 16,989094 | 16,89765 | 16,018858 | 16,343521 | 16,755062 |
| 789 | 188299 | 6,9468478 | 8,0099766 | 7,4654998 | 8,0572237 | 7,2810735 | 6,9269982 | 7,1301289 | 7,4889247 |
| 827 | 133913 | 10,508758 | 10,667269 | 10,807326 | 11,049491 | 10,516343 | 10,228996 | 10,093795 | 10,575521 |
| 945 | 115325 | 7,1664459 | 7,2642105 | 6,9813359 | 7,7679318 | 7,0662762 | 6,5887087 | 7,2179296 | 7,5481648 |
| 970 | 132077 | 6,4701601 | 5,8775408 | 7,1568402 | 6,8989895 | 6,3080947 | 6,1299337 | 6,4603977 | 6,4552342 |
| 988 | 103763 | 9,9362311 | 9,7216546 | 9,4493952 | 9,6243857 | 9,9805756 | 10,405236 | 9,6597138 | 9,8957957 |

[1,2,4,5,6,7,8,9]

| Bicluster2 | probeid | SW756 | C-4I | Hela | C-33A | CaSki | ME-180 | HT-3 |
|---|---|---|---|---|---|---|---|---|
| 54 | 123046 | 6,2604615 | 6,0867955 | 6,0283161 | 5,8639515 | 5,6031656 | 5,4447403 | 5,5625959 |
| 108 | 223314 | 13,250507 | 13,357549 | 14,163669 | 13,450594 | 13,435654 | 13,495538 | 13,71513 |
| 122 | 218771 | 6,1673759 | 6,3372134 | 6,8640884 | 6,2613675 | 6,4033843 | 5,515733 | 6,2307612 |
| 136 | 179699 | 6,0605367 | 6,488282 | 6,4373535 | 5,6458328 | 6,3762465 | 6,1859343 | 6,567816 |
| 188 | 179835 | 10,157576 | 10,544473 | 11,158965 | 10,212479 | 9,9924021 | 9,7347065 | 10,036018 |
| 189 | 194736 | 10,744164 | 10,712833 | 10,943116 | 9,8835131 | 10,07051 | 10,015788 | 10,983737 |
| 234 | 221988 | 7,3640563 | 7,1074731 | 7,7761992 | 6,5824233 | 6,7726323 | 7,2718647 | 6,9926969 |
| 238 | 223080 | 12,641111 | 12,038428 | 12,950012 | 11,918355 | 12,525733 | 12,355858 | 11,957329 |
| 324 | 154584 | 14,628408 | 15,054853 | 15,25637 | 15,392527 | 14,691489 | 14,394271 | 14,986364 |
| 335 | 151522 | 5,9361234 | 5,5847727 | 5,7972861 | 5,7965896 | 5,433939 | 5,8143683 | 5,6900053 |
| 340 | 154888 | 6,390746 | 7,2068961 | 7,1303375 | 6,6805298 | 6,0672843 | 6,3146393 | 6,7779422 |
| 377 | 192696 | 10,121227 | 10,659981 | 11,18284 | 10,093795 | 10,315534 | 10,55309 | 10,922014 |
| 491 | 215816 | 6,3899653 | 6,3981969 | 6,3900422 | 5,5671996 | 5,6492146 | 6,0633497 | 5,4640663 |
| 508 | 203716 | 5,4704512 | 6,2548297 | 6,0951527 | 5,4979064 | 5,0308181 | 5,3920372 | 6,0051192 |
| 623 | 120891 | 6,5295374 | 7,3545927 | 7,2493306 | 7,2700248 | 7,1550659 | 6,9694486 | 7,2185135 |
| 666 | 148497 | 7,3514019 | 7,3783108 | 8,1517736 | 7,0943583 | 6,5569517 | 7,1519138 | 7,2770573 |
| 672 | 210379 | 5,9104613 | 6,3110838 | 6,3048424 | 5,8570119 | 5,4818631 | 6,4043214 | 5,8261136 |
| 683 | 159571 | 6,844648 | 6,4133251 | 6,6130043 | 6,3629451 | 5,9622476 | 6,2332867 | 6,2076771 |
| 706 | 182057 | 6,1677562 | 5,7035211 | 6,2954765 | 5,1425854 | 5,6951074 | 5,8258961 | 5,5143382 |
| 746 | 163337 | 5,7317263 | 6,3796109 | 6,5406114 | 5,6355177 | 6,0498887 | 5,2265592 | 6,0055418 |
| 787 | 116263 | 6,2181285 | 6,67417 | 6,764491 | 5,8011927 | 6,3352916 | 6,3116288 | 6,2890676 |
| 801 | 170969 | 15,481284 | 15,313343 | 15,8155 | 15,149473 | 15,231389 | 15,359821 | 16,145245 |
| 818 | 161284 | 15,819166 | 16,486548 | 16,085887 | 15,787794 | 16,040113 | 16,049235 | 15,531591 |
| 984 | 177045 | 11,303146 | 11,030518 | 11,377406 | 10,762659 | 11,0386 | 10,898998 | 11,251174 |

[1,3,4,5,6,7,8]

| Bicluster3 | probeid | SW756 | GH354 | C-4I | Hela | C-33A | CaSki | HT-3 | SiHa |
|---|---|---|---|---|---|---|---|---|---|
| 103 | 223312 | 17,695863 | 16,9746 | 17,167315 | 17,164593 | 17,382179 | 17,06098 | 18,307484 | 17,543301 |
| 172 | 118863 | 6,8654793 | 6,6613918 | 7,0667775 | 7,1037585 | 7,0758282 | 6,2684029 | 7,5674798 | 6,7716146 |
| 256 | 136075 | 12,229301 | 12,13014 | 12,136236 | 12,184896 | 12,006205 | 12,84008 | 12,868284 | 12,193915 |
| 320 | 149734 | 16,301211 | 16,196096 | 15,778411 | 16,470544 | 16,494209 | 16,955271 | 15,645068 | 16,021126 |
| 351 | 182451 | 6,4634986 | 6,1067434 | 6,1097399 | 5,5997289 | 6,1823656 | 6,0265526 | 6,5926419 | 6,8458697 |
| 430 | 160066 | 5,8401153 | 6,0855577 | 6,0432835 | 6,0148574 | 5,6962245 | 5,9903697 | 5,3917207 | 6,0122305 |
| 432 | 151850 | 10,032368 | 10,839931 | 10,961283 | 10,183791 | 10,228648 | 10,136258 | 10,627039 | 10,193396 |
| 614 | 173773 | 5,4640663 | 5,8088829 | 6,237605 | 6,0241773 | 6,0750597 | 5,6403743 | 5,0232437 | 5,4900634 |
| 682 | 117809 | 5,582758 | 5,691659 | 5,6703544 | 4,9430577 | 5,8586223 | 4,9187942 | 6,0878815 | 5,0303012 |
| 731 | 201180 | 12,355179 | 12,324213 | 12,346168 | 11,746266 | 12,237772 | 12,184896 | 13,171494 | 12,062419 |
| 829 | 172304 | 9,310312 | 8,9155523 | 9,6120867 | 8,8028246 | 8,9627377 | 8,572419 | 8,3956762 | 8,8686033 |
| 854 | 185623 | 8,5008586 | 9,0594775 | 8,9312181 | 8,2350808 | 9,2109289 | 8,1676981 | 7,8631886 | 8,8346931 |
| 882 | 116395 | 6,3720036 | 5,699461 | 6,2350065 | 5,4214359 | 6,432977 | 6,057131 | 6,0641783 | 6,2828348 |
| 897 | 195901 | 9,1491575 | 8,1189274 | 8,9600987 | 8,1028501 | 8,3084145 | 8,5355312 | 8,3423835 | 8,5254744 |
| 973 | 171716 | 5,9071753 | 6,0926236 | 6,2185824 | 5,4928146 | 6,5300633 | 5,3285285 | 5,8551242 | 6,3851517 |
| 1000 | 107345 | 7,844192 | 6,927098 | 7,8328134 | 7,2221772 | 7,3571121 | 6,9895693 | 6,9954964 | 7,5336043 |

[1,2,3,4,5,6,8,9]

## Καρκίνος του τραχήλου της μήτρας

### di_new,dj ->Δεύτερη εκτέλεση αλγορίθμου

| di_new,dj | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Biclusteri | probeid | SW756 | GH354 | C-4I | Hela | C-33A | CaSki | ME-180 | HT-3 | SiHa |
| 49 | 152249 | 7,2556324 | 7,4255726 | 7,0260866 | 7,2391865 | 7,1776018 | 6,675348 | 6,605961 | 7,1831391 | 7,0102056 |
| 92 | 116550 | 6,0934023 | 6,8134234 | 6,1339317 | 6,1966267 | 6,4284483 | 6,4643673 | 6,4093507 | 6,7781048 | 6,3648074 |
| 106 | 223310 | 16,761318 | 16,992873 | 16,932573 | 17,150353 | 17,092611 | 17,283723 | 17,004424 | 16,615897 | 16,4656 |
| 210 | 124803 | 9,433284 | 9,5914623 | 9,8105594 | 9,5093646 | 10,464566 | 10,147551 | 9,9777943 | 9,7560875 | 10,149089 |
| 234 | 221988 | 7,3640563 | 7,3476925 | 7,1074731 | 7,7761992 | 6,5824233 | 6,7726323 | 7,2718647 | 6,9926969 | 6,92741 |
| 273 | 174944 | 17,077839 | 17,06098 | 17,239012 | 16,820318 | 17,364681 | 17,239012 | 16,627222 | 17,024784 | 17,294955 |
| 338 | 109551 | 16,615897 | 16,608639 | 16,648786 | 16,862559 | 17,161044 | 17,153337 | 16,739178 | 17,637396 | 16,739178 |
| 551 | 212317 | 5,3460891 | 6,139169 | 6,2208492 | 5,8530042 | 6,3726268 | 6,5747708 | 6,4603028 | 5,9973073 | 5,99219 |
| 719 | 162602 | 10,050049 | 10,312374 | 9,8098455 | 10,361391 | 10,523458 | 10,374621 | 10,25464 | 10,202138 | 10,129886 |
| 720 | 236549 | 7,4425682 | 7,3287883 | 7,0360753 | 6,7838702 | 6,3687385 | 6,7847416 | 7,3366123 | 7,0931715 | 7,5441673 |
| 772 | 101481 | 5,4471093 | 5,9642345 | 6,1967716 | 5,8457607 | 6,1359576 | 6,0876415 | 6,0641783 | 6,1406199 | 5,9407602 |
| 1000 | 107345 | 7,844192 | 6,927098 | 7,8328134 | 7,2221772 | 7,3571121 | 6,9895693 | 7,5530775 | 6,9954964 | 7,5336043 |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

| Bicluster2 | probeid | SW756 | GH354 | C-4I | Hela | C-33A | CaSki | ME-180 | HT-3 | SiHa |
|---|---|---|---|---|---|---|---|---|---|---|
| 159 | 165735 | 5,9391925 | 5,3989462 | 5,34104 | 5,8140107 | 5,1945288 | 5,3873598 | 5,6457532 | 5,8606492 | 5,5600311 |
| 623 | 120891 | 6,5295374 | 6,6117393 | 7,3545927 | 7,2493306 | 7,2700248 | 7,1550659 | 6,9694486 | 7,2185135 | 7,7782864 |
| 661 | 217397 | 6,5714892 | 7,2341783 | 7,5866759 | 7,486672 | 7,6314021 | 7,0574531 | 7,4964565 | 6,692654 | 7,7362114 |
| 789 | 188299 | 6,9468478 | 8,0099766 | 7,6111433 | 7,4654998 | 8,0572237 | 7,2810735 | 6,9269982 | 7,1301289 | 7,4889247 |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

| Bicluster3 | probeid | SW756 | GH354 | C-4I | Hela | C-33A | CaSki | ME-180 | HT-3 | SiHa |
|---|---|---|---|---|---|---|---|---|---|---|
| 44 | 106630 | 6,4156662 | 6,8901907 | 5,8842316 | 7,1092478 | 7,1890325 | 6,821831 | 6,5629594 | 6,8059 | 6,5970332 |
| 172 | 118863 | 6,8654793 | 6,6613918 | 7,0667775 | 7,1037585 | 7,0758282 | 6,2684029 | 6,4555176 | 7,5674798 | 6,7716146 |
| 945 | 115325 | 7,1664459 | 7,2642105 | 6,4911273 | 6,9813359 | 7,7679318 | 7,0662762 | 6,5887087 | 7,2179296 | 7,5481648 |
| 988 | 103763 | 9,9362311 | 9,7216546 | 9,8782942 | 9,4493952 | 9,6243857 | 9,9805756 | 10,405236 | 9,6597138 | 9,8957957 |
| | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

## Καρκίνος του τραχήλου της μήτρας

### di_new,dj_new ->Τρίτη εκτέλεση αλγορίθμου

| di_new,dj_new | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bicluster1 | probeid | C-4I | Hela | CaSki | ME-180 | HT-3 | SiHa |
| 49 | 152249 | 7,0260866 | 7,2391865 | 6,675348 | 6,605961 | 7,1831391 | 7,0102056 |
| 92 | 116550 | 6,1339317 | 6,1966267 | 6,4643673 | 6,4093507 | 6,7781048 | 6,3648074 |
| 234 | 221988 | 7,1074731 | 7,7761992 | 6,7726323 | 7,2718647 | 6,9926969 | 6,92741 |
| 342 | 168674 | 7,4809475 | 7,1709002 | 7,037937 | 7,1603067 | 6,9564847 | 6,5262645 |
| 509 | 191330 | 7,4928272 | 6,9077857 | 7,1172795 | 7,3937337 | 6,4906461 | 7,2274741 |
| 623 | 120891 | 7,3545927 | 7,2493306 | 7,1550659 | 6,9694486 | 7,2185135 | 7,7782864 |
| 661 | 217397 | 7,5866759 | 7,486672 | 7,0574531 | 7,4964565 | 6,692654 | 7,7362114 |
| 719 | 162602 | 9,8098455 | 10,361391 | 10,374621 | 10,25464 | 10,202138 | 10,129886 |
| 720 | 236549 | 7,0360753 | 6,7838702 | 6,7847416 | 7,3366123 | 7,0931715 | 7,5441673 |
| 772 | 101481 | 6,1967716 | 5,8457607 | 6,0876415 | 6,0641783 | 6,1406199 | 5,9407602 |
| 789 | 188299 | 7,6111433 | 7,4654998 | 7,2810735 | 6,9269982 | 7,1301289 | 7,4889247 |
| 847 | 156024 | 7,0959633 | 7,042771 | 6,6325935 | 7,1431925 | 7,3662256 | 6,9547596 |
| 1000 | 107345 | 7,8328134 | 7,2221772 | 6,9895693 | 7,5530775 | 6,9954964 | 7,5336043 |
| [3,4,6,7,8,9] | | | | | | | |

| Bicluster2 | probeid | SW756 | GH354 | C-33A | HT-3 | SiHa |
|---|---|---|---|---|---|---|
| 172 | 118863 | 6,8654793 | 6,6613918 | 7,0758282 | 7,5674798 | 6,7716146 |
| 245 | 219524 | 6,9933466 | 7,0079331 | 7,5889966 | 7,3853697 | 7,2401637 |
| 261 | 172017 | 6,926658 | 6,486689 | 7,3822668 | 7,2808587 | 6,4238303 |
| 272 | 133136 | 7,4919227 | 7,5574718 | 7,8574376 | 7,7323969 | 7,0680994 |
| 371 | 140490 | 6,6485941 | 6,481245 | 7,3411922 | 7,0022584 | 6,5775083 |
| 569 | 145804 | 6,9062355 | 7,2519258 | 7,7897533 | 6,598699 | 6,9973845 |
| 666 | 148497 | 7,3514019 | 7,6312422 | 7,0943583 | 7,2770573 | 6,6092993 |
| 768 | 192988 | 6,9915397 | 7,0320979 | 7,6214427 | 7,5304628 | 6,9806237 |
| 904 | 132754 | 7,6118188 | 7,6046634 | 7,6840971 | 7,2112607 | 7,709287 |
| 945 | 115325 | 7,1664459 | 7,2642105 | 7,7679318 | 7,2179296 | 7,5481648 |
| 982 | 223803 | 7,2359009 | 7,4773407 | 7,4454309 | 7,7696152 | 7,6125504 |
| [1,2,5,8,9] | | | | | | |

| Bicluster3 | | | | | |
|---|---|---|---|---|---|
| | probeid | SW756 | ME-180 | HT-3 | SiHa |
| 295 | 233204 | 8,0055847 | 6,8137238 | 7,6743726 | 7,6191485 |
| 419 | 180895 | 7,4157829 | 6,9701876 | 7,4893019 | 7,6257569 |
| 596 | 132804 | 7,4968769 | 7,5022686 | 6,820095 | 7,3395559 |
| 602 | 125603 | 7,1482264 | 7,0567261 | 7,1587824 | 6,8531651 |
| 886 | 128078 | 7,7477107 | 7,3307188 | 7,7053705 | 6,9397308 |
| [1,7,8,9] | | | | | |

## *Καρκίνος των ωοθηκών*

## *di,dj ->Πρώτη εκτέλεση αλγορίθμου*

didj

| Bicluster1 | PROBEID | TOV1-112D | TOV-21G | A2780 | OVMZ-1a | OVMZ-6 | ES2 | NIHOVCAR | SKOV3 |
|---|---|---|---|---|---|---|---|---|---|
| 18 | 226368 | 15,7379129 | 15,465121 | 16,02816 | 15,356016 | 15,1792954 | 15,329264 | 15,353424 | 15,377771 |
| 103 | 223312 | 18,1794136 | 17,351643 | 17,96271 | 18,293143 | 18,02913709 | 17,590251 | 18,077758 | 17,409074 |
| 107 | 224286 | 17,2837228 | 17,32466 | 17,574964 | 17,077839 | 17,58592857 | 17,161044 | 16,741962 | 17,601485 |
| 108 | 223314 | 12,800542 | 12,925963 | 12,986568 | 13,455832 | 13,23912734 | 12,917051 | 13,16329 | 13,119356 |
| 122 | 218771 | 6,74944958 | 6,4806563 | 6,9204249 | 7,2827909 | 6,345454271 | 7,114367 | 6,344109 | 6,6661717 |
| 150 | 173890 | 10,3879201 | 10,7037 | 10,94591 | 10,22203 | 10,3110408 | 10,172476 | 10,543374 | 10,077205 |
| 229 | 147982 | 6,54364443 | 6,1168583 | 6,7338716 | 5,9788519 | 5,703983691 | 6,1352692 | 6,5926419 | 6,4525014 |
| 240 | 225083 | 5,67692581 | 5,3638811 | 5,63402 | 5,5403504 | 5,165510018 | 4,8809046 | 5,5228107 | 5,1634987 |
| 256 | 136075 | 12,7296776 | 12,15337 | 12,500029 | 12,288318 | 13,15919313 | 12,488998 | 12,311977 | 12,318726 |
| 273 | 174944 | 17,7012599 | 17,505743 | 17,146659 | 17,202891 | 17,27832253 | 16,916038 | 17,619068 | 17,70907 |
| 289 | 155326 | 16,7988501 | 16,072855 | 16,618584 | 16,310572 | 16,49738913 | 16,793959 | 16,646626 | 16,068002 |
| 303 | 236908 | 16,1255554 | 15,352013 | 16,559007 | 16,037954 | 15,52754771 | 16,074999 | 15,742935 | 16,095703 |
| 312 | 133153 | 15,8044849 | 15,324819 | 15,50583 | 14,973255 | 15,57465469 | 15,768305 | 14,935221 | 15,341162 |
| 336 | 228186 | 16,2298187 | 15,735666 | 15,699758 | 15,357498 | 16,00275343 | 16,011063 | 16,053518 | 15,620514 |
| 338 | 109551 | 16,9380756 | 16,409861 | 16,823723 | 16,203067 | 16,63016826 | 16,932573 | 16,15016 | 16,952144 |
| 343 | 139301 | 16,6612839 | 16,306307 | 16,191409 | 16,537146 | 16,51114586 | 16,458293 | 16,494209 | 16,135774 |
| 363 | 111270 | 6,8242922 | 6,418241 | 6,9320904 | 6,7052613 | 7,485121714 | 6,8358044 | 7,0754341 | 6,7667486 |
| 428 | 162055 | 5,74633398 | 5,311201 | 5,5620134 | 6,4645307 | 5,728808434 | 5,6909537 | 5,7842491 | 5,4409522 |
| 443 | 207793 | 7,71357644 | 6,9137429 | 6,692654 | 6,5501878 | 6,544998993 | 6,6085389 | 6,5899093 | 6,9672976 |
| 489 | 146209 | 6,84327644 | 6,36466 | 6,2830006 | 6,2694761 | 5,945114198 | 6,5530937 | 6,2203707 | 6,7689434 |
| 563 | 139732 | 17,5803341 | 17,879865 | 17,590251 | 16,772149 | 17,33941308 | 17,262928 | 17,359659 | 17,343609 |
| 591 | 107487 | 15,0548525 | 15,029642 | 15,272012 | 15,125877 | 15,3968738 | 14,923273 | 15,883249 | 14,865719 |
| 632 | 183840 | 15,5850621 | 16,059265 | 15,764491 | 15,747878 | 16,05926459 | 15,262877 | 15,502191 | 15,601904 |
| 638 | 124186 | 6,37254893 | 6,1123819 | 5,9202118 | 5,773219 | 5,573973994 | 5,3336591 | 5,408761 | 5,2402993 |
| 659 | 116503 | 6,15684152 | 5,4558809 | 5,5285137 | 4,994669 | 5,722318612 | 5,5746473 | 5,5235702 | 4,9947583 |
| 713 | 209576 | 6,25304379 | 5,6859401 | 5,6207292 | 5,9418353 | 5,472996456 | 6,5060455 | 6,2759358 | 5,9191895 |
| 766 | 173243 | 6,00843452 | 6,1925757 | 6,4235914 | 6,0473048 | 6,064376887 | 6,9935194 | 6,1919034 | 5,9849011 |
| 810 | 204592 | 6,48643917 | 5,669728 | 6,0518329 | 6,3899653 | 6,782556861 | 6,349585 | 5,8843376 | 5,9684849 |
| 818 | 161284 | 16,5051173 | 15,871562 | 16,070343 | 16,478134 | 15,92099323 | 16,347399 | 15,690157 | 15,819166 |
| 839 | 232396 | 16,1651629 | 15,407491 | 15,865316 | 16,090322 | 15,7457014 | 15,737913 | 15,692309 | 15,244569 |
| 866 | 210738 | 5,6510328 | 5,6654219 | 4,987273 | 5,8775408 | 5,265967828 | 5,9550258 | 5,6571059 | 5,5255084 |
| 925 | 166194 | 6,82901832 | 6,9385628 | 7,1002474 | 6,5898522 | 6,594568192 | 6,526559 | 6,974084 | 6,5459399 |
| 952 | 236105 | 6,50377135 | 6,4994536 | 6,1653725 | 5,6614856 | 6,453911737 | 6,4181078 | 5,7040492 | 5,6612455 |
| 978 | 104969 | 16,5590074 | 15,838353 | 16,353252 | 16,109584 | 16,31057239 | 16,548714 | 16,095703 | 15,416587 |

[1,2,3,4,5,6,9,10]

| Bicluster2 | PROBEID | TOV1-112D | A2780 | OVMZ-1a | OVMZ-6 | ES2 | CaOV3 | OV-90 | SKOV3 |
|---|---|---|---|---|---|---|---|---|---|
| 123 | 149228 | 12,2222882 | 11,890098 | 12,635037 | 12,776238 | 12,79095007 | 12,89723 | 12,047212 | 11,793411 |
| 234 | 221988 | 7,22217715 | 7,0130978 | 7,8113831 | 7,9480129 | 7,831337249 | 7,0974625 | 6,8644533 | 6,202928 |
| 290 | 236468 | 11,8981584 | 11,211995 | 11,463834 | 11,127235 | 11,52892636 | 11,505089 | 10,872294 | 10,365279 |
| 313 | 101179 | 16,1095845 | 16,656108 | 16,427677 | 16,296234 | 16,64662645 | 16,301211 | 15,383752 | 15,969388 |
| 347 | 187848 | 6,44728773 | 6,6978857 | 6,0488941 | 6,582136 | 6,895678912 | 6,371944 | 5,6596918 | 5,801819 |
| 476 | 191452 | 6,10305033 | 6,306172 | 6,2758745 | 6,5270428 | 6,9299607 | 7,0460671 | 5,5760726 | 5,6163786 |
| 505 | 119841 | 5,18980376 | 5,5070351 | 6,3800987 | 6,0797093 | 5,48521495 | 6,0130302 | 5,0505876 | 5,3085594 |
| 516 | 199188 | 6,15887805 | 6,0285313 | 6,2305944 | 6,4899958 | 6,091332378 | 5,9884336 | 5,2568269 | 6,1624231 |
| 589 | 166147 | 16,1213074 | 16,412396 | 16,700807 | 16,322663 | 16,33030579 | 15,723546 | 15,614817 | 15,168473 |
| 602 | **125603** | 6,73045762 | 7,0833394 | 7,1126754 | 6,9023799 | 7,135374475 | 6,4303997 | 6,2258493 | 5,9626038 |
| 604 | 168817 | 6,04815701 | 5,9130626 | 6,0657973 | 6,2034021 | 6,147852941 | 5,8481287 | 5,1059471 | 4,7327548 |
| 626 | 158037 | 6,24821223 | 6,4797547 | 6,1808723 | 6,8699071 | 5,997443953 | 5,8878844 | 5,3507019 | 5,48613 |
| 661 | 217397 | 7,18835719 | 6,5624594 | 7,9225171 | 7,138358 | 7,738561642 | 7,0507938 | 6,9243375 | 6,559212 |
| 704 | 215191 | 6,49412503 | 5,8403009 | 6,9663005 | 7,0889862 | 6,365723843 | 6,4112389 | 5,9957761 | 5,9743844 |
| 774 | 189928 | 16,5137135 | 16,94107 | 17,157058 | 16,638988 | 17,2735678 | 16,463155 | 16,521896 | 16,087828 |
| 789 | 188299 | 7,74884783 | 7,6950231 | 8,3387915 | 8,2474016 | 7,791589618 | 7,6556814 | 7,5223007 | 6,3745247 |
| 801 | 170969 | 14,6528442 | 14,954761 | 15,610309 | 15,134559 | 15,30316111 | 15,919082 | 14,778764 | 14,178687 |
| 813 | 209910 | 16,8325925 | 16,801756 | 16,829647 | 17,013202 | 17,07444358 | 16,793959 | 16,451459 | 15,922733 |
| 848 | 215038 | 6,5199507 | 6,1137038 | 6,7929495 | 6,364213 | 6,232347708 | 6,2486367 | 6,0216375 | 5,1738954 |
| 899 | 192789 | 5,18023373 | 4,9411434 | 5,5766812 | 4,7939848 | 4,80686671 | 5,7667137 | 4,6068318 | 4,3868291 |
| 904 | 132754 | 7,2381382 | 7,022943 | 7,8916574 | 7,2783088 | 7,211839636 | 6,9931589 | 6,9553409 | 6,425302 |
| 936 | 213180 | 5,69786013 | 5,6824846 | 5,0730648 | 5,7369181 | 5,492882021 | 4,9240999 | 5,1398744 | 4,5837914 |

[1,3,4,5,6,7,8,10]

| Bicluster3 | PROBEID | A2780 | OVMZ-1a | OVMZ-6 | ES2 | CaOV3 | NIHOVCAR | SKOV3 |
|---|---|---|---|---|---|---|---|---|
| 24 | 225407 | 5,23709672 | 6,0124393 | 5,9500202 | 5,8580595 | 6,008528877 | 6,2291831 | 6,2055952 |
| 40 | 117084 | 8,28385356 | 8,3987696 | 8,2030092 | 8,4775048 | 8,134211481 | 8,528927 | 7,8663875 |
| 44 | 106630 | 5,64964571 | 6,8774408 | 6,9523566 | 6,7775997 | 6,474521434 | 6,7214014 | 6,473569 |
| 106 | 223310 | 17,0247838 | 16,698362 | 16,548714 | 16,484922 | 17,15705819 | 17,095959 | 17,225637 |
| 170 | 129100 | 5,44215342 | 6,0068649 | 6,153336 | 5,2619563 | 6,188607064 | 6,1317216 | 6,5634151 |
| 264 | 195714 | 5,83503315 | 6,5082452 | 6,252999 | 5,8164987 | 6,348979684 | 6,2787551 | 6,4356725 |
| 267 | 133314 | 16,1983577 | 16,193707 | 16,384526 | 15,840708 | 16,50925634 | 16,511146 | 15,433908 |
| 275 | 206551 | 5,37222815 | 5,8995255 | 6,1266121 | 5,8287946 | 6,188034382 | 5,9754226 | 6,4668804 |
| 323 | 160958 | 4,93096138 | 5,4813576 | 6,4661248 | 5,6550354 | 5,679420835 | 5,8756055 | 5,5996623 |
| 329 | 231349 | 6,22904141 | 7,0805567 | 7,1186365 | 6,1862733 | 7,554411129 | 6,7872152 | 6,3694206 |
| 392 | 144507 | 5,27321649 | 5,7254398 | 6,7762651 | 5,5569779 | 6,136019727 | 6,2498495 | 6,1048332 |
| 471 | 158152 | 5,09554693 | 5,7802341 | 5,2542778 | 4,8822762 | 4,869196072 | 5,703022 | 5,1478208 |
| 512 | 101480 | 10,0516612 | 10,558189 | 10,376918 | 9,5686182 | 10,03200004 | 10,410123 | 10,383646 |
| 540 | 178114 | 9,57170525 | 9,7743367 | 9,4405785 | 8,8965897 | 9,419266098 | 9,8764612 | 9,1264661 |
| 569 | 145804 | 6,87806705 | 6,9794568 | 7,3857148 | 6,0649866 | 7,077212073 | 7,3067525 | 7,0181504 |
| 594 | 233317 | 5,28634618 | 6,1874563 | 5,5788593 | 5,8367945 | 5,805014068 | 6,1736692 | 5,8036519 |
| 612 | 168728 | 11,4252714 | 11,203862 | 11,745807 | 11,15573 | 12,25827925 | 11,611021 | 11,361696 |
| 688 | 141430 | 6,32383698 | 6,2992032 | 6,5919967 | 6,2977393 | 6,018812186 | 6,6628707 | 6,64085 |
| 738 | 160038 | 9,51692542 | 9,4273255 | 9,5895526 | 9,7373683 | 9,881684339 | 10,093795 | 9,1064176 |
| 833 | 176210 | 5,37194397 | 6,1844682 | 6,5860576 | 6,4042564 | 6,161553588 | 6,7394751 | 6,0737498 |
| 842 | 104386 | 6,10009516 | 6,4510679 | 6,276815 | 5,9218597 | 5,793690574 | 6,2685309 | 6,019778 |
| 857 | 102989 | 6,58575324 | 6,4664855 | 6,9082504 | 6,0447133 | 7,107032571 | 6,7193564 | 6,5610367 |
| 886 | 128078 | 7,28796263 | 7,239008 | 7,8833555 | 7,7373301 | 8,278780779 | 7,5141783 | 7,3180064 |
| 938 | 187142 | 5,53718161 | 6,1700964 | 6,1894162 | 5,0775803 | 6,452026601 | 6,0029157 | 5,8929534 |
| 959 | 163701 | 5,6691163 | 6,3350283 | 6,5723332 | 5,764264 | 6,458719587 | 5,8523537 | 6,4292283 |

[3,4,5,6,7,9,10]

## Καρκίνος των ωοθηκών
## di_new,dj ->Δεύτερη εκτέλεση αλγορίθμου

| di_new,dj | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bicluster1 | PROBEID | TOV1-112D | TOV-21G | A2780 | OVMZ-1a | OVMZ-6 | ES2 | CaOV3 | OV-90 | NIHOVCAR | SKOV3 |
| 363 | 111270 | 6,8242922 | 6,418241 | 6,9320904 | 6,7052613 | 7,485121714 | 6,8358044 | 7,3875571 | 7,2975004 | 7,0754341 | 6,7667486 |
| 443 | 207793 | 7,71357644 | 6,9137429 | 6,692654 | 6,5501878 | 6,544998993 | 6,6085389 | 6,5426531 | 7,4140354 | 6,5899093 | 6,9672976 |
| 925 | 166194 | 6,82901832 | 6,9385628 | 7,1002474 | 6,5898522 | 6,594568192 | 6,526559 | 7,2665967 | 7,4900908 | 6,974084 | 6,5459399 |
| | | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | | |

| Bicluster2 | PROBEID | TOV1-112D | TOV-21G | A2780 | OVMZ-1a | OVMZ-6 | ES2 | CaOV3 | OV-90 | NIHOVCAR | SKOV3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 621 | 165634 | 6,24406083 | 5,7958449 | 6,7238172 | 4,8975079 | 6,365203519 | 6,1545623 | 6,2973073 | 6,229092 | 6,5054146 | 6,5253312 |
| 768 | 192988 | 6,60686297 | 7,1156351 | 6,3811376 | 7,3652956 | 6,457141076 | 7,1103088 | 6,1067902 | 7,3084732 | 6,9262492 | 6,2721607 |
| 851 | 109569 | 6,57953032 | 7,3933363 | 6,1282793 | 7,422378 | 6,549703618 | 6,5671335 | 6,8731348 | 7,0839132 | 6,2752227 | 6,5750695 |
| 904 | 132754 | 7,2381382 | 6,6501032 | 7,022943 | 7,8916574 | 7,278308838 | 7,2118396 | 6,9931589 | 6,9553409 | 6,6064383 | 6,425302 |
| | | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | | |

| Bicluster3 | PROBEID | TOV1-112D | TOV-21G | A2780 | OVMZ-1a | OVMZ-6 | ES2 | CaOV3 | OV-90 | NIHOVCAR | SKOV3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 602 | 125603 | 6,73045762 | 7,0137296 | 7,0833394 | 7,1126754 | 6,902379875 | 7,1353745 | 6,4303997 | 6,2258493 | 7,2993816 | 5,9626038 |
| 629 | 120621 | 6,50568204 | 7,1436759 | 6,709385 | 7,088367 | 6,61450838 | 8,1463563 | 6,7959849 | 6,7256407 | 6,8036859 | 6,5109328 |
| [1,2,3,4,5,6,7,8,9,10] | | | | | | | | | | | |

## Καρκίνος των ωοθηκών
## di_new,dj_new ->Τρίτη εκτέλεση αλγορίθμου

| di_new,dj_new | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Bicluster1 | PROBEID | TOV1-112D | TOV-21G | A2780 | CaOV3 | OV-90 | NIHOVCAR | SKOV3 |
| 106 | 223310 | 17,1570582 | 17,405011 | 17,024784 | 17,157058 | 16,84930004 | 17,095959 | 17,225637 |
| 234 | 221988 | 7,22217715 | 6,757904 | 7,0130978 | 7,0974625 | 6,864453346 | 6,9597366 | 6,202928 |
| 363 | 111270 | 6,8242922 | 6,418241 | 6,9320904 | 7,3875571 | 7,297500361 | 7,0754341 | 6,7667486 |
| 443 | 207793 | 7,71357644 | 6,9137429 | 6,692654 | 6,5426531 | 7,414035362 | 6,5899093 | 6,9672976 |
| 475 | 150990 | 6,99351944 | 7,0657832 | 7,5185434 | 6,4555176 | 7,151673526 | 6,2844284 | 6,4270261 |
| 629 | 120621 | 6,50568204 | 7,1436759 | 6,709385 | 6,7959849 | 6,725640715 | 6,8036859 | 6,5109328 |
| 661 | 217397 | 7,18835719 | 6,5645507 | 6,5624594 | 7,0507938 | 6,924337464 | 6,7814736 | 6,559212 |
| 850 | 110828 | 6,72590257 | 6,7840097 | 6,933981 | 6,4666078 | 7,222909272 | 6,6639513 | 6,5977938 |
| 904 | 132754 | 7,2381382 | 6,6501032 | 7,022943 | 6,9931589 | 6,955340949 | 6,6064383 | 6,425302 |
| 925 | 166194 | 6,82901832 | 6,9385628 | 7,1002474 | 7,2665967 | 7,490090836 | 6,974084 | 6,5459399 |
| | | | | | | | | |
| [1,2,3,7,8,9,10] | | | | | | | | |

| Bicluster2 | PROBEID | TOV1-112D | TOV-21G | OVMZ-1a | OVMZ-6 | ES2 | SKOV3 |
|---|---|---|---|---|---|---|---|
| 37 | 112566 | 7,20101988 | 6,7020432 | 7,0299309 | 7,2306854 | 7,141359646 | 6,7395817 |
| 389 | 121677 | 6,93482685 | 6,7941012 | 7,3777676 | 7,0947042 | 7,443192215 | 6,6787354 |
| 553 | 190994 | 7,04689711 | 7,7809507 | 7,7282356 | 7,2392983 | 7,126228975 | 6,8276937 |
| 602 | 125603 | 6,73045762 | 7,0137296 | 7,1126754 | 6,9023799 | 7,135374475 | 5,9626038 |
| 741 | 155981 | 6,76334867 | 7,1038317 | 7,3703777 | 7,3575752 | 7,222283944 | 6,0529684 |
| 768 | 192988 | 6,60686297 | 7,1156351 | 7,3652956 | 6,4571411 | 7,110308839 | 6,2721607 |
| 851 | 109569 | 6,57953032 | 7,3933363 | 7,422378 | 6,5497036 | 6,567133502 | 6,5750695 |
| 886 | 128078 | 7,08596857 | 7,820268 | 7,239008 | 7,8833555 | 7,737330078 | 7,3180064 |
| 927 | 143750 | 7,65058129 | 7,6360598 | 7,6930596 | 7,9135036 | 7,263978295 | 7,223674 |
| | | | | | | | |
| [1,2,4,5,6,10] | | | | | | | |

| Bicluster3 | PROBEID | TOV1-112D | A2780 | OVMZ-6 | CaOV3 | NIHOVCAR3 |
|---|---|---|---|---|---|---|
| 272 | 133136 | 7,0674754 | 7,5522646 | 7,071667 | 6,8624547 | 6,316743597 |
| 288 | 130324 | 6,90423372 | 6,9275144 | 7,275436 | 6,6212345 | 6,614169313 |
| 419 | 180895 | 7,99619397 | 8,0429035 | 7,2522509 | 6,8778556 | 7,467860777 |
| 569 | 145804 | 7,38660904 | 6,8780671 | 7,3857148 | 7,0772121 | 7,306752463 |
| 578 | 224152 | 7,96894867 | 7,7239716 | 7,7246916 | 7,7413579 | 6,984970036 |
| 654 | 125612 | 7,38884662 | 7,2248598 | 7,7998048 | 7,2645463 | 7,019960828 |
| 691 | 180199 | 7,6332486 | 7,4964565 | 7,5844839 | 7,204759 | 7,145835637 |
| 789 | 188299 | 7,74884783 | 7,6950231 | 8,2474016 | 7,6556814 | 7,652811016 |
| | | | | | | |
| [1,3,5,7,9] | | | | | | |

## *Καρκίνος του ενδομητρίου*

## *di,dj ->Πρώτη εκτέλεση αλγορίθμου*

| didj | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Bicluster1 | probeid | Colo684 | EJ | Ishikawa | RL95-2 | HEC50B | EN | KLE |
| 17 | 195641 | 5,1111741 | 5,5256815 | 6,2073252 | 5,6442586 | 5,538481 | 5,7861768 | 5,452781 |
| 18 | 226368 | 15,529672 | 15,508389 | 15,494832 | 15,352013 | 15,44289 | 15,281114 | 15,084199 |
| 62 | 152248 | 6,6588879 | 6,7606902 | 7,0314495 | 6,7706361 | 7,1738612 | 6,9229677 | 6,8042669 |
| 76 | 126740 | 6,2783272 | 6,8723946 | 6,3174935 | 6,6619712 | 6,1738533 | 7,1022037 | 6,5448181 |
| 106 | 223310 | 16,871212 | 17,32931 | 17,206458 | 17,228569 | 17,299845 | 17,167315 | 16,786591 |
| 108 | 223314 | 13,013436 | 13,613256 | 13,809197 | 13,445186 | 13,791853 | 13,368319 | 12,977188 |
| 139 | 189048 | 10,074756 | 10,730731 | 10,255854 | 9,9010824 | 10,377425 | 10,757879 | 9,5047641 |
| 179 | 167993 | 12,508707 | 13,564157 | 12,865101 | 13,070306 | 13,390217 | 13,453833 | 12,385763 |
| 289 | 155326 | 16,40326 | 16,531586 | 16,698362 | 16,308579 | 16,371572 | 16,58984 | 16,443786 |
| 291 | 142445 | 11,729471 | 11,964825 | 12,13014 | 11,128061 | 11,323464 | 11,981681 | 11,61386 |
| 310 | 155784 | 13,754493 | 14,426497 | 14,635047 | 14,680114 | 14,059613 | 14,641752 | 14,01721 |
| 312 | 133153 | 15,192173 | 15,487503 | 15,156363 | 15,465121 | 15,359821 | 15,448341 | 14,612691 |
| 313 | 101179 | 16,398967 | 15,997415 | 16,430609 | 15,980208 | 15,687813 | 16,42183 | 16,245134 |
| 326 | 140229 | 14,637152 | 15,407491 | 14,851502 | 15,125877 | 14,628408 | 14,637152 | 14,848991 |
| 336 | 228186 | 15,429874 | 16,016084 | 16,217193 | 16,040113 | 16,123224 | 15,574655 | 15,96562 |
| 338 | 109551 | 16,648786 | 16,702922 | 16,687635 | 16,601811 | 16,992873 | 16,854327 | 16,196096 |
| 341 | 195447 | 8,4864011 | 8,4771213 | 9,3417365 | 8,5577348 | 8,6641375 | 8,6774294 | 8,6884463 |
| 343 | 139301 | 16,313029 | 16,656108 | 16,06409 | 16,361515 | 16,068002 | 16,43635 | 15,497016 |
| 389 | 121677 | 6,8971672 | 7,0236581 | 7,216451 | 7,6950231 | 7,0833394 | 7,3664419 | 6,9123757 |
| 411 | 109012 | 14,940218 | 14,837265 | 14,848991 | 14,770729 | 14,770729 | 15,19776 | 14,725384 |
| 508 | 203716 | 5,2335087 | 5,9795079 | 5,8239169 | 4,9255248 | 5,6565792 | 5,6305603 | 5,0401484 |
| 563 | 139732 | 17,080473 | 17,769008 | 18,26978 | 17,601485 | 17,678258 | 17,63115 | 17,243605 |
| 569 | 145804 | 6,1187663 | 7,0810447 | 6,6894647 | 6,594081 | 6,5586342 | 6,7865585 | 5,8728612 |
| 589 | 166147 | 16,820318 | 16,125555 | 16,451459 | 16,648786 | 16,79885 | 16,661284 | 16,349895 |
| 658 | 169351 | 6,0010496 | 6,1479011 | 6,301294 | 5,5729056 | 6,4517794 | 6,4287657 | 5,4207847 |
| 661 | 217397 | 7,3276732 | 7,4869472 | 7,2915851 | 7,1056025 | 6,8709033 | 7,7892857 | 7,1579957 |
| 678 | 180501 | 5,5227405 | 6,4467373 | 6,1163987 | 6,1318028 | 6,5977252 | 6,3799027 | 6,4561231 |
| 708 | 123101 | 6,2374339 | 5,9581039 | 6,4537342 | 5,5295333 | 6,0122746 | 6,3007162 | 5,4994347 |
| 719 | 162602 | 10,818644 | 10,277792 | 10,529423 | 10,544473 | 10,3796 | 10,988977 | 10,361109 |
| 738 | 160038 | 9,5311983 | 9,710946 | 9,8764612 | 9,3727259 | 9,4313041 | 9,1705085 | 8,5482997 |
| 747 | 126904 | 5,4988471 | 5,2913961 | 6,0177754 | 5,2725438 | 6,2023198 | 5,8492619 | 5,1511688 |
| 751 | 166935 | 5,5566714 | 6,4164375 | 5,8771759 | 5,9910683 | 6,1266121 | 5,634876 | 5,4797122 |
| 803 | 229775 | 15,6388 | 15,345858 | 15,00233 | 15,052667 | 15,082471 | 14,832302 | 14,482925 |
| 807 | 210821 | 11,463224 | 12,225078 | 12,20957 | 11,788789 | 12,217098 | 11,419131 | 11,53896 |
| 808 | 108253 | 5,0108387 | 5,4539594 | 5,7968901 | 5,5689841 | 5,7093051 | 5,5598782 | 4,7330118 |
| 812 | 137218 | 9,7702605 | 10,847045 | 10,652312 | 10,134408 | 10,421055 | 10,501591 | 10,172476 |
| 813 | 209910 | 17,32466 | 17,095959 | 17,400933 | 17,136389 | 16,995934 | 17,206458 | 16,482851 |
| 818 | 161284 | 15,908055 | 16,107068 | 16,125555 | 15,922733 | 16,361515 | 16,109584 | 16,085887 |
| 819 | 137602 | 11,330985 | 11,938639 | 11,885545 | 12,306417 | 11,628154 | 12,12709 | 11,653752 |
| 846 | 162033 | 6,3635584 | 6,7969208 | 6,391524 | 6,308339 | 6,281298 | 6,9063843 | 6,3040547 |
| 859 | 194717 | 6,6789207 | 7,0503957 | 6,830727 | 6,1677562 | 6,0959799 | 6,4642426 | 5,96013 |
| 978 | 104969 | 16,215126 | 16,273059 | 16,287786 | 15,932692 | 16,615897 | 16,505117 | 15,986147 |
| | | | | | | | | |
| [1,2,3,4,5,7,8] | | | | | | | | |

| Bicluster2 | probeid | Colo684 | EJ | Ishikawa | RL95-2 | HEC50B | HEC-1-B | KLE | AN3CA |
|---|---|---|---|---|---|---|---|---|---|
| 157 | 200991 | 7,0869994 | 6,2714974 | 6,9419974 | 6,4261323 | 6,4971081 | 6,1687011 | 6,4028236 | 7,363035 |
| 224 | 192626 | 8,7224211 | 7,5934796 | 8,440655 | 8,655606 | 8,7297236 | 7,996336 | 7,9132044 | 8,6933109 |
| 255 | 103976 | 9,1725433 | 8,1470978 | 8,7188068 | 8,0438878 | 8,2838536 | 8,5237084 | 8,4486049 | 8,8850951 |
| 267 | 133314 | 15,728419 | 15,538683 | 15,896653 | 15,760015 | 15,951415 | 15,368623 | 15,435765 | 16,524772 |
| 303 | 236908 | 16,227882 | 15,56437 | 16,114664 | 15,588929 | 15,474013 | 15,453389 | 15,650253 | 16,700807 |
| 342 | 168674 | 6,649468 | 6,2484744 | 6,4489835 | 6,5001185 | 6,7697091 | 7,1581074 | 5,9597488 | 7,1189493 |
| 460 | 204250 | 6,8473047 | 6,2779407 | 6,5172446 | 5,4595956 | 6,2232295 | 6,3396344 | 5,9702058 | 6,501812 |
| 475 | 150990 | 6,7869299 | 5,80253 | 6,760312 | 6,4953922 | 5,7753209 | 5,9243125 | 5,5671996 | 6,7696534 |
| 525 | 154674 | 7,0402551 | 6,2992032 | 6,5136902 | 6,4994536 | 6,3948869 | 7,4352667 | 5,9082504 | 6,6894647 |
| 587 | 182060 | 7,3560048 | 6,4485402 | 7,2504991 | 6,9605568 | 7,3139773 | 6,8415312 | 6,500823 | 7,4523215 |
| 594 | 233317 | 6,4631543 | 5,7393685 | 6,2220042 | 6,0649384 | 6,3434593 | 5,9534705 | 6,3135562 | 6,6205092 |
| 606 | 116808 | 6,2171797 | 5,0460441 | 5,6700971 | 5,3722282 | 4,6478598 | 5,0163802 | 5,2705486 | 5,8879292 |
| 623 | 120891 | 7,7135764 | 7,0774356 | 6,8826173 | 6,9501737 | 7,2295778 | 8,0382707 | 7,1096767 | 7,4236666 |
| 628 | 175740 | 6,5050343 | 5,6541913 | 6,6696273 | 5,5356355 | 6,2056698 | 5,9830776 | 6,0906863 | 6,5911424 |
| 714 | 121735 | 6,3268032 | 6,2120675 | 6,8884389 | 6,7437042 | 6,2585239 | 6,7410979 | 6,3422946 | 6,7155538 |
| 774 | 189928 | 17,431223 | 16,900555 | 16,910156 | 16,829647 | 17,637396 | 17,486833 | 16,801756 | 17,480562 |
| 794 | 192616 | 10,262443 | 10,027021 | 10,282683 | 9,4492 | 10,458359 | 10,269028 | 10,457694 | 10,534256 |
| 845 | 139151 | 6,1652667 | 4,9118934 | 5,6582641 | 5,8237829 | 5,7756016 | 6,3099143 | 5,6665548 | 5,5240777 |
| 879 | 146121 | 6,2609963 | 5,8633256 | 6,049218 | 5,3389123 | 5,8671812 | 6,7511494 | 6,2313374 | 6,2663329 |
| 983 | 218294 | 8,3393261 | 8,2868653 | 7,5959366 | 7,5316195 | 7,5560139 | 7,9455809 | 7,8872607 | 8,1741964 |
| 993 | 171357 | 10,988122 | 10,119898 | 10,664937 | 10,645323 | 9,7078827 | 10,586558 | 10,23447 | 10,565823 |
| 996 | 193510 | 12,032497 | 11,371232 | 11,916066 | 11,117876 | 11,301751 | 11,691295 | 11,016279 | 11,821102 |
| [1,2,3,4,5,6,8,9] | | | | | | | | | |

| Bicluster3 | probeid | EJ | Ishikawa | RL95-2 | HEC50B | HEC-1-B | EN | KLE | AN3CA |
|---|---|---|---|---|---|---|---|---|---|
| 87 | 212871 | 7,0027306 | 5,8432566 | 5,229679 | 5,6700338 | 6,2135212 | 5,5371121 | 5,7116181 | 5,7381456 |
| 110 | 217769 | 7,5413575 | 6,3502041 | 6,1350855 | 6,6074705 | 6,9140767 | 6,5436444 | 5,9006317 | 6,6245088 |
| 223 | 188207 | 7,1736272 | 6,8324637 | 6,1831469 | 6,517802 | 6,600089 | 6,763429 | 6,7252353 | 6,8211362 |
| 234 | 221988 | 7,7244349 | 7,5622545 | 7,3191789 | 7,4435761 | 8,2580316 | 7,6797671 | 7,1736272 | 7,576373 |
| 258 | 173187 | 11,330466 | 10,349652 | 10,261676 | 10,27968 | 10,730179 | 10,258141 | 9,9389199 | 9,9976542 |
| 440 | 101235 | 7,3673481 | 7,9190782 | 6,735483 | 6,3958205 | 6,6873867 | 7,2368048 | 6,5002653 | 6,9681271 |
| 468 | 120067 | 6,9855304 | 6,6461854 | 6,6608854 | 6,2227059 | 7,3991351 | 6,8230454 | 6,6202121 | 6,4308971 |
| 512 | 101480 | 10,779109 | 10,37095 | 9,8459431 | 9,8796856 | 10,615672 | 9,6349405 | 9,3243784 | 10,356617 |
| 540 | 178114 | 9,6260404 | 9,3142117 | 9,1938268 | 9,0428157 | 9,1098155 | 9,6338853 | 9,401789 | 9,6819716 |
| 580 | 161375 | 6,5022121 | 6,2079134 | 5,9845144 | 6,0583849 | 6,390617 | 5,1379942 | 5,0781986 | 5,6577152 |
| 596 | 132804 | 7,7841682 | 7,1906824 | 7,6787299 | 7,5546939 | 8,6723006 | 7,9234166 | 7,7044079 | 7,4410702 |
| 607 | 201788 | 7,1363193 | 6,7148077 | 6,6739469 | 6,514197 | 6,4269136 | 6,2227059 | 6,0120334 | 6,3264602 |
| 718 | 214717 | 6,5222119 | 6,2518387 | 7,0300078 | 6,6249589 | 7,0698539 | 7,028636 | 6,3010485 | 6,7028398 |
| 791 | 101211 | 6,1041049 | 6,2963366 | 6,1495145 | 5,474103 | 6,1540479 | 5,4938681 | 5,0385147 | 5,7258703 |
| 832 | 100578 | 5,9223731 | 5,3248769 | 6,3137089 | 5,8017509 | 6,6901487 | 5,6725221 | 5,4393782 | 6,4783779 |
| 850 | 110828 | 7,9022577 | 6,6460489 | 6,862236 | 7,2629405 | 7,351295 | 6,8498907 | 7,0041524 | 6,860414 |
| 866 | 210738 | 5,9190295 | 6,2792293 | 5,5224101 | 5,4650809 | 6,0047675 | 5,3523292 | 5,8562814 | 5,0027499 |
| 878 | 112931 | 5,2250679 | 6,3332071 | 5,0220056 | 5,9568851 | 5,6853648 | 5,2377408 | 5,2653461 | 5,5512165 |
| 888 | 201366 | 5,6974139 | 6,120568 | 6,1685428 | 6,122185 | 5,5094792 | 5,8606917 | 6,861864 | 6,3255871 |
| 898 | 193056 | 5,8883844 | 6,6240893 | 6,1214572 | 6,0652503 | 6,3227678 | 6,0241248 | 5,6378451 | 6,4450715 |
| 968 | 164417 | 5,9928459 | 5,5110119 | 5,9212678 | 6,0157877 | 6,777077 | 5,9492738 | 5,75769 | 6,1104709 |
| 973 | 171716 | 5,8561994 | 7,1410934 | 5,8560815 | 5,66837 | 6,0033537 | 5,6416677 | 5,6157709 | 6,0224963 |
| 977 | 232221 | 7,2139535 | 6,9077857 | 5,6757342 | 6,3370514 | 5,9760561 | 5,9838907 | 6,1627703 | 6,2847304 |
| | | | | | | | | | |
| [2,3,4,5,6,7,8,9] | | | | | | | | | |

## *Καρκίνος του ενδομητρίου*

## *di_new,dj ->Δεύτερη εκτέλεση αλγορίθμου*

| di_new,dj | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bicluster1 | probeid | Colo684 | EJ | Ishikawa | RL95-2 | HEC50B | HEC-1-B | EN | KLE | AN3CA |
| 18 | 226368 | 15,529672 | 15,508389 | 15,494832 | 15,352013 | 15,44289 | 15,06966 | 15,281114 | 15,084199 | 15,862873 |
| 62 | 152248 | 6,6588879 | 6,7606902 | 7,0314495 | 6,7706361 | 7,1738612 | 7,4170241 | 6,9229677 | 6,8042669 | 6,3560372 |
| 389 | 121677 | 6,8971672 | 7,0236581 | 7,216451 | 7,6950231 | 7,0833394 | 6,8761049 | 7,3664419 | 6,9123757 | 6,3542239 |
| 468 | 120067 | 7,0232991 | 6,9855304 | 6,6461854 | 6,6608854 | 6,2227059 | 7,3991351 | 6,8230454 | 6,6202121 | 6,4308971 |
| 594 | 233317 | 6,4631543 | 5,7393685 | 6,2220042 | 6,0649384 | 6,3434593 | 5,9534705 | 5,9309365 | 6,3135562 | 6,6205092 |
| 661 | 217397 | 7,3276732 | 7,4869472 | 7,2915851 | 7,1056025 | 6,8709033 | 7,0658682 | 7,7892857 | 7,1579957 | 6,8261939 |
| 787 | 116263 | 7,0970036 | 6,9996069 | 6,9001645 | 6,9141648 | 6,4049578 | 6,7954382 | 7,5178362 | 7,210215 | 6,6421391 |
| 850 | 110828 | 7,2613601 | 7,9022577 | 6,6460489 | 6,862236 | 7,2629405 | 7,351295 | 6,8498907 | 7,0041524 | 6,860414 |
| | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

| Bicluster2 | probeid | Colo684 | EJ | Ishikawa | RL95-2 | HEC50B | HEC-1-B | EN | KLE | AN3CA |
|---|---|---|---|---|---|---|---|---|---|---|
| 587 | 182060 | 7,3560048 | 6,4485402 | 7,2504991 | 6,9605568 | 7,3139773 | 6,8415312 | 7,6760452 | 6,500823 | 7,4523215 |
| 623 | 120891 | 7,7135764 | 7,0774356 | 6,8826173 | 6,9501737 | 7,2295778 | 8,0382707 | 7,6449719 | 7,1096767 | 7,4236666 |
| | | | | | | | | | | |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

| Bicluster3 | probeid | Colo684 | EJ | Ishikawa | RL95-2 | HEC50B | HEC-1-B | EN | KLE | AN3CA |
|---|---|---|---|---|---|---|---|---|---|---|
| 440 | 101235 | 6,7635828 | 7,3673481 | 7,9190782 | 6,735483 | 6,3958205 | 6,6873867 | 7,2368048 | 6,5002653 | 6,9681271 |
| 849 | 159494 | 7,7465691 | 7,4836609 | 7,3100363 | 6,93126 | 6,3518133 | 6,5724569 | 7,6436388 | 6,7709386 | 7,1090072 |
| [1,2,3,4,5,6,7,8,9] | | | | | | | | | | |

## *Καρκίνος του ενδομητρίου*

## *di_new,dj_new ->Τρίτη εκτέλεση αλγορίθμου*

| di_new,dj_new | | | | | | | |
|---|---|---|---|---|---|---|---|
| Bicluster1 | | | | | | | |
| | probeid | Colo684 | EJ | Ishikawa | RL95-2 | EN | KLE |
| 13 | 220149 | 6,9447845 | 7,3246709 | 7,2787453 | 7,0199608 | 7,0204756 | 6,6130043 |
| 18 | 226368 | 15,529672 | 15,508389 | 15,494832 | 15,352013 | 15,281114 | 15,084199 |
| 62 | 152248 | 6,6588879 | 6,7606902 | 7,0314495 | 6,7706361 | 6,9229677 | 6,8042669 |
| 217 | 214216 | 7,1207535 | 6,8887465 | 7,3970981 | 6,903977 | 7,183954 | 7,6873471 |
| 234 | 221988 | 6,5946664 | 7,7244349 | 7,5622545 | 7,3191789 | 7,6797671 | 7,1736272 |
| 389 | 121677 | 6,8971672 | 7,0236581 | 7,216451 | 7,6950231 | 7,3664419 | 6,9123757 |
| 419 | 180895 | 7,5135915 | 7,65543 | 7,3016502 | 6,7333294 | 7,9476062 | 7,3628113 |
| 468 | 120067 | 7,0232991 | 6,9855304 | 6,6461854 | 6,6608854 | 6,8230454 | 6,6202121 |
| 594 | 233317 | 6,4631543 | 5,7393685 | 6,2220042 | 6,0649384 | 5,9309365 | 6,3135562 |
| 596 | 132804 | 7,7468732 | 7,7841682 | 7,1906824 | 7,6787299 | 7,9234166 | 7,7044079 |
| 623 | 120891 | 7,7135764 | 7,0774356 | 6,8826173 | 6,9501737 | 7,6449719 | 7,1096767 |
| 661 | 217397 | 7,3276732 | 7,4869472 | 7,2915851 | 7,1056025 | 7,7892857 | 7,1579957 |
| 669 | 117923 | 7,0579746 | 7,8138656 | 7,5690601 | 7,5606345 | 7,2774489 | 7,8776361 |
| 787 | 116263 | 7,0970036 | 6,9996069 | 6,9001645 | 6,9141648 | 7,5178362 | 7,210215 |
| 789 | 188299 | 6,946637 | 7,7455558 | 7,1430367 | 7,0991817 | 7,2361376 | 7,9243734 |
| 849 | 159494 | 7,7465691 | 7,4836609 | 7,3100363 | 6,93126 | 7,6436388 | 6,7709386 |
| [1,2,3,4,7,8] | | | | | | | |

| Bicluster2 | probeid | Colo684 | Ishikawa | HEC50B | HEC-1-B | KLE |
|---|---|---|---|---|---|---|
| 172 | 118863 | 6,5593872 | 6,5203354 | 6,7121432 | 6,5491017 | 6,9444085 |
| 243 | 200166 | 6,4491008 | 7,0714341 | 7,0919503 | 7,709141 | 6,9097542 |
| 328 | 185445 | 7,2027013 | 7,3276732 | 7,4037297 | 7,1951637 | 7,7291127 |
| 368 | 143236 | 6,8458036 | 7,1207535 | 7,3217905 | 7,1574453 | 6,7624223 |
| 587 | 182060 | 7,3560048 | 7,2504991 | 7,3139773 | 6,8415312 | 6,500823 |
| 645 | 113730 | 6,6061656 | 7,3890528 | 7,3487619 | 7,426929 | 6,8743731 |
| 780 | 109288 | 7,1368994 | 7,1705418 | 7,0288774 | 7,6835506 | 7,6585742 |
| 850 | 110828 | 7,2613601 | 6,6460489 | 7,2629405 | 7,351295 | 7,0041524 |
| | | | | | | |
| [1,3,5,6,8] | | | | | | |

| Bicluster3 | probeid | EJ | Ishikawa | RL95-2 | HEC50B | AN3CA |
|---|---|---|---|---|---|---|
| 116 | 219127 | 7,5580813 | 7,4747797 | 6,5354105 | 7,352694 | 7,5579332 |
| 438 | 181532 | 7,9707598 | 7,4392251 | 7,7187311 | 7,881333 | 8,2323641 |
| 608 | 115230 | 8,3862721 | 8,0889528 | 7,7353895 | 8,1837973 | 7,7163123 |
| 886 | 128078 | 8,3740923 | 8,1413744 | 7,1051917 | 7,3554995 | 7,4534147 |
| 983 | 218294 | 8,2868653 | 7,5959366 | 7,5316195 | 7,5560139 | 8,1741964 |
| | | | | | | |
| [2,3,4,5,9] | | | | | | |