ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΜΟΡΙΑΚΗ ΒΙΟΛΟΓΙΑ

Χρηστίνα Κ. Δασκαλοπούλου

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική.

Πειραιάς
Οκτώβριος 2011

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΕΘΟΔΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΜΟΡΙΑΚΗ ΒΙΟΛΟΓΙΑ

Χρηστίνα Κ. Δασκαλοπούλου

Διπλωματική Εργασία
που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική.

Πειραιάς
Οκτώβριος 2011

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Επίκουρη Καθηγήτρια Βερροπούλου Γεωργία
- Λέκτορας Μπερσίμης Σωτήρης

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERISTY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

CLUSTERING TECHNIQUES IN MOLECULAR BIOLOGY

By

Christina K. Daskalopoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics.

Piraeus, Greece

October 2011

Στους Γονείς Μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά όσους συντέλεσαν στην ολοκλήρωση της παρούσας διπλωματικής εργασίας. Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα Καθηγητή κ. Κούτρα Μάρκο για τις πολύτιμες συμβουλές του, για τη γνώση και την καθοδήγηση που μου προσέφερε όχι μόνο κατά τη συγγραφή της παρούσας διπλωματικής εργασίας αλλά και καθ' όλη τη διάρκεια των μεταπτυχιακών σπουδών μου. Επίσης θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής, Επίκουρη Καθηγήτρια κ. Βερροπούλου Γεωργία και Λέκτορα κ. Μπερσίμη Σωτήρη, για τη συμμετοχή τους στην εξεταστική επιτροπή.

Τέλος θα ήθελα να ευχαριστήσω θερμά την οικογένειά μου που με στηρίζει όλα τα χρόνια της ακαδημαϊκής μου πορείας, τους φίλους μου και τους συμφοιτητές μου για την ηθική τους συμπαράσταση και υποστήριξη. Ιδιαίτερες ευχαριστίες σε Κυριακή, Γιώργο, Εύα, Πέτρο και Αγνή.

Περίληψη

Οι πρόσφατες ανακαλύψεις μικροσυστοιχιών που αποτελούνται από δεδομένα γονιδιακής έκφρασης (*gene expression*) δημιούργησαν μια νέα εποχή για τις βιολογικές επιστήμες. Ο κύριος λόγος είναι ότι πλέον δίνεται η δυνατότητα ταυτόχρονης παρακολούθησης εκατοντάδων χιλιάδων γονιδίων. Κατά συνέπεια, λόγω του τεράστιου όγκου δεδομένων που πλέον είναι διαθέσιμος, γίνεται απαραίτητη η χρήση στατιστικών τεχνικών που δίνουν τη δυνατότητα αναγνώρισης γονιδίων που παρουσιάζουν παρόμοια λειτουργική συμπεριφορά και κατάταξης αυτών σε ομογενείς ομάδες. Η πιο δημοφιλής στατιστική τεχνική που εφαρμόζεται σε τέτοιου είδους δεδομένα είναι η ανάλυση κατά συστάδες.

Στην παρούσα διπλωματική εργασία αρχικά παρουσιάζονται ορισμοί και τεχνικές της Βιολογίας ενώ στη συνέχεια επιχειρείται η παρουσίαση των δημοφιλέστερων τεχνικών ανάλυσης κατά συστάδες. Συγχρόνως παρουσιάζονται λεπτομερώς νέα μέτρα επικύρωσης που δημιουργήθηκαν με σκοπό όχι μόνο τη στατιστική αλλά και τη βιολογική ερμηνεία των συστάδων που προκύπτουν. Τέλος όλα τα θεωρητικά αποτελέσματα εφαρμόζονται και ερμηνεύονται πρακτικά στο τελευταίο κεφάλαιο όπου γίνεται μια εφαρμογή σε πραγματικά γονιδιακά δεδομένα.

Abstract

Microarrays of gene expression have provoked a tremendous advent in the Biological field because they allow the simultaneous studying of thousands of genes. However due to the enormous datasets that are available statistical techniques are necessary now more than ever. The most popular statistical technique that is being applied to such data is clustering. Cluster analysis gives the opportunity to the researcher to group data with similar functionality in homogenous clusters.

In this thesis an introduction which includes definitions and biological techniques is firstly presented and then the most popular clustering techniques are illustrated. Moreover old validation measures and new ones, which have been created in order to explain the biological meaning of clusters, are meticulously studied. Finally all the theoretical results are being applied and explained through an example to real gene expression data.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κατάλογος Συντομογραφιών

DNA	DeoxyriboNucleic Acid – Δεσοξυριβονουκλεϊκό Οξύ
RNA	RiboNucleic Acid – Ριβονουκλεϊκό Οξύ
mRNA	Messenger RNA – Αγγελιαφόρο RNA
tRNA	Transfer RNA – Μεταφορικό RNA
rRNA	Ribosomal RNA – Ριβοσωμικό RNA
snRNA	Small Nuclear RNA – Μικρό Πυρηνικό RNA
cDNA	Complementary DNA – Συμπληρωματικό DNA
CGH	Comparative Genomic Hybridization
SNP	Single Nucleotide Polymorphism
GEO	Gene Expression Omnibus
EST	Expressed Sequence Tag
UPGMA	Unweighted Pair-Group Average Method – Μέθοδος της Μη Σταθμισμένης ανά Ομάδα Απόστασης
DIANA	Divisive Analysis – Διαχωριστική Ανάλυση
DC	Divisive Coefficient – Διαχωριστικός Συντελεστής
PAM	Partition Around Medoids
CLARA	Clustering Large Applications
FANNY	Fuzzy Analysis
FOM	Figure of Merit
APN	Average Proportion of Non-Overlap – Μέτρο της Μέσης Αναλογίας μη Επικάλυψης
ADM	Average Distance between Means Measure – Μέτρο της Μέσης Απόστασης Μέσων
AD	Average Distance Measure – Μέτρο της Μέσης Απόστασης
BHI	Biological Homogeneity Index – Βιολογικός Δείκτης Ομογένειας
BSI	Biological Stability Index – Βιολογικός Δείκτης Ευστάθειας

Κεφάλαιο 1

Βιολογικές Έννοιες και Ορισμοί

1.1 Εισαγωγή στις έννοιες της Μοριακής Βιολογίας

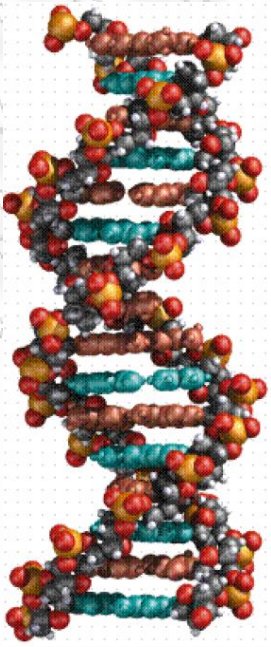
Τα τελευταία χρόνια υπάρχει μια πλειάδα νέων πληροφοριών που αφορά την επιστήμη της μοριακής βιολογίας καθώς νέες μέθοδοι και νέα δεδομένα έχουν κάνει την εμφάνισή τους. Η ελπίδα όλων είναι πως η εκμετάλλευση αυτών των νέων πληροφοριών θα οδηγήσει στη γνώση του γενετικού υλικού ιών ή στη γρήγορη διάγνωση δυσλειτουργιών, όπως ο καρκίνος, και κατά συνέπεια θα συμβάλλει στην ανακάλυψη θεραπειών σε κυτταρικό επίπεδο. Η επιστήμη η οποία βοηθά στην ανάλυση και στην εξαγωγή συμπερασμάτων από τα γιγάντια σύνολα δεδομένων που υπάρχουν είναι η στατιστική. Βασική προϋπόθεση όμως για την παρουσίαση της συγκεκριμένης διπλωματικής εργασίας αλλά και για να συνεχίσουμε την περαιτέρω ανάλυσή μας είναι η αναφορά και η επεξήγηση βασικών εννοιών όπως είναι το **DNA**, το **RNA**, το **cDNA**, το ολιγονουκλεοτίδιο (***oligonucleotide***) και οι μικροσυστοιχίες (***microarrays***).

Τα νουκλεοτίδια είναι οργανικές ενώσεις, ή αλλιώς σύνθετα οργανικά μόρια, που σχηματίζουν τη βασική μονάδα των νουκλεϊκών οξέων δηλαδή του DNA και του RNA. Αυτές οι οργανικές ενώσεις αποτελούνται από 3 διαφορετικά επιμέρους μόρια που συνδέονται μεταξύ τους με ομοιοπολικό δεσμό: μιας "πεντόζης", το οποίο είναι ένα σάκχαρο με 5 άτομα άνθρακα και καλείται είτε ριβόζη είτε δεσοξυριβόζη, ενός μορίου φωσφορικού οξέος και μιας οργανικής αζωτούχου βάσης τύπου πουρίνης ή τύπου πυριμιδίνης. Τα νουκλεοτίδια αποτελούν τις δομικές μονάδες των νουκλεϊκών οξέων όταν 3 ή και περισσότερα ενώνονται μεταξύ τους για να σχηματίσουν ένα νουκλεϊκό οξύ. Τα νουκλεϊκά οξέα που έχουν ως βάση την πεντόζη δεσοξυριβόζη ή δεοξυριβόζη (εξ ου και δε(σ)οξυριβονουκλεοτίδια), ονομάζονται DNA (*DeoxyriboNucleic Acid*) και αυτά που έχουν ως βάση την πεντόζη ριβόζη (τα ριβονουκλεοτίδια), ονομάζονται RNA (*RiboNucleic Acid*). Συνεπώς δύο είναι τα κύρια είδη νουκλεοτιδίων: τα δε(σ)οξυριβονουκλεοτίδια και τα ριβονουκλεοτίδια.

Στα δε(σ)οξυριβονουκλεοτίδια σταθερό τμήμα τους είναι η πεντόζη δε(σ)οξυριβόζη και το μόριο του φωσφορικού οξέος, ενώ το μεταβλητό μέρος είναι μια από τις παρακάτω αζωτούχες βάσεις αδενίνη, θυμίνη, κυτοσίνη, ή γουανίνη. Στα ριβονουκλεοτίδια σταθερό τμήμα τους είναι

η πεντόζη ριβόζη και το μόριο του φωσφορικού οξέος, ενώ το μεταβλητό μέρος τους είναι μία από τις 4 αζωτούχες βάσεις: αδενίνη, ουρακίλη, κυτοσίνη και γουανίνη.

Το DNA είναι ένα νουκλεϊκό οξύ που περιέχει όλες τις απαραίτητες γενετικές πληροφορίες για την ανάπτυξη και τη λειτουργία των ζωντανών οργανισμών. Το DNA αποτελείται ουσιαστικά από μια διπλή έλικα με βάσεις στο κέντρο της και σάκχαρο-φωσφορικές μονάδες κατά μήκος των πλευρών της. Τα δύο παράλληλα μέρη της «σκάλας» είναι συμπληρωματικά όπως διαπιστώθηκε από έρευνες των *Watson* και *Crick*. Πιο συγκεκριμένα η αδενίνη (*A*) ζευγαρώνει αποκλειστικά με τη θυμίνη (*T*) ενώ η κυτοσίνη (*C*) ζευγαρώνει αποκλειστικά με τη γουανίνη (*G*). Οι δεσμοί αυτοί δημιουργούνται με τη βοήθεια του υδρογόνου. Για αυτόν το λόγο, αν είναι γνωστή η ακολουθία βάσεων ενός μέρους της διπλής έλικας τότε είναι εύκολος ο υπολογισμός του συμπληρώματος της και κατά συνέπεια της ακολουθίας βάσεων του άλλου μέρους της έλικας. Ο βασικός ρόλος του DNA στο κύτταρο είναι η μακρόχρονη αποθήκευση της πληροφορίας αφού μπορεί να χαρακτηριστεί και ως *ο μοριακός σκληρός δίσκος του οργανισμού*. Παρέχει πληροφορίες οι οποίες καθορίζουν τη δομή και τη λειτουργία του οργανισμού ενώ συγχρόνως περιέχει πληροφορίες για τον αυτοδιπλασιασμό των κυττάρων εξασφαλίζοντας έτσι τη μεταβίβαση των γενετικών οδηγιών από γενιά σε γενιά. Στο παρακάτω σχήμα δίνεται η χαρακτηριστική διπλή έλικα του DNA:



Σχήμα 1.1: Η Διπλή Έλικα του DNA

Το RNA είναι ένα νουκλεϊκό οξύ πολυμερές που αποτελείται από νουκλεοτιδικά μονομερή. Τα RNA πολυνουκλεοτίδια, όπως αναφέρθηκε και προηγουμένως, έχουν για σάκχαρα ριβόζες

αντί για τις δεοξυριβόζες που έχει το DNA και αντί για τη βάση θυμίνη την ουρακίλη. Όλοι οι οργανισμοί έχουν DNA γονιδίωμα με εξαίρεση τους ιούς οι οποίοι έχουν RNA γονιδίωμα. Στο σημείο αυτό να προσθέσουμε ότι το ως γονιδίωμα (***genome***) καλούμε συχνά το αποτύπωμα ενός οργανισμού αφού η έννοια του γονιδιώματος εμπεριέχει όλες τις πληροφορίες για την κατασκευή ποικίλων δομών του κυττάρου, όπως οι πρωτεΐνες και τα μόρια RNA. Το DNA ενός οργανισμού είναι μια στατική πληροφορία και συνεπώς υπάρχουν μικρότερα κομμάτια του που κουβαλούν αυτήν την πληροφορία, τα οποία έχουν τη δυνατότητα να μεταγράφονται, να μεταφράζονται κυρίως σε πρωτεΐνες και έτσι να είναι χρήσιμα για τον οργανισμό. Αυτά τα κομμάτια DNA καλούνται γονίδια (***genes***). Πιο συγκεκριμένα τα γονίδια είναι ουσιαστικά μια λειτουργική μονάδα κληρονομικότητας που βρίσκεται στα χρωμοσώματα στον πυρήνα ενός κυττάρου και βοηθούν στο σχηματισμό ενός ενζύμου ή άλλης πρωτεΐνης. Όσον αφορά το RNA αυτό συναντάται σε τρεις διαφορετικές μορφές:

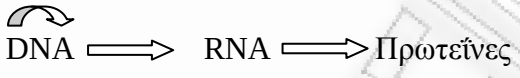
1. Το mRNA (***messenger RNA*** – αγγελιαφόρο). Αποτελεί λιγότερο από το 5% του ολικού RNA ενός κυττάρου και κωδικοποιεί την πληροφορία του DNA σε πρωτεΐνες.

2. Το tRNA (***transfer RNA*** – μεταφορικό). Αποτελεί το 15% περίπου του συνολικού RNA στο κύτταρο και μεταφέρει τα αμινοξέα στη ριβοσωμική θέση κατά την ανάπτυξη της πολυπεπτιδικής αλυσίδας στο στάδιο της μετάφρασης.

3. Το rRNA (***ribosomal RNA*** - ριβοσωμικό). Αποτελεί το 80% περίπου του συνολικού RNA του κυττάρου και μαζί με το ριβόσωμα συμβάλλει στη δημιουργία των πρωτεϊνών.

Στο σημείο αυτό αξίζει να αναφέρουμε ότι στο κυτταρόπλασμα το ριβοσωμικό RNA συνδυάζεται με ειδικές πρωτεΐνες του πλάσματος και έτσι συνθέτεται ένα νουκλεοπρωτεΐνικό σύμπλεγμα το οποίο καλείται **ριβόσωμα**. Τα ευκαρυωτικά ριβοσώματα αποτελούν το εργοστάσιο πρωτεΐνοσύνθεσης του κυττάρου και ορισμένα από αυτά μπορούν να είναι συνδεδεμένα με μία μονή αλυσίδα mRNA σε όλη τη διάρκεια της ζωής τους. Να σημειώσουμε ότι στους ευκαρυωτικούς οργανισμούς (δηλαδή στους οργανισμούς που σχηματίζουν πυρήνα) συναντάται και το snRNA (***small nuclear* RNA** - μικρό πυρηνικό) τα οποία είναι ουσιαστικά μικρά μόρια RNA που συνδέονται με τις πρωτεΐνες και σχηματίζουν μικρά ριβονουκλεοπρωτεινικά σωματίδια. Τα σωματίδια αυτά καταλύουν την ωρίμανση του mRNA.

Το πρώτο στάδιο για την έκφραση της πληροφορίας που υπάρχει στο DNA είναι η **αντιγραφή** της και έπειτα η μεταφορά της στο RNA με τη διαδικασία της **μεταγραφής**. Στη συνέχεια το *RNA* μεταφέρει με τη διαδικασία της **μετάφρασης** την πληροφορία στις πρωτεΐνες που είναι υπεύθυνες για τη δομή και τη λειτουργία των κυττάρων. Τα παραπάνω βήματα αποτελούν το κεντρικό δόγμα της μοριακής βιολογίας (*Crick,* 1958) και συνοψίζονται στο ακόλουθο σχήμα:



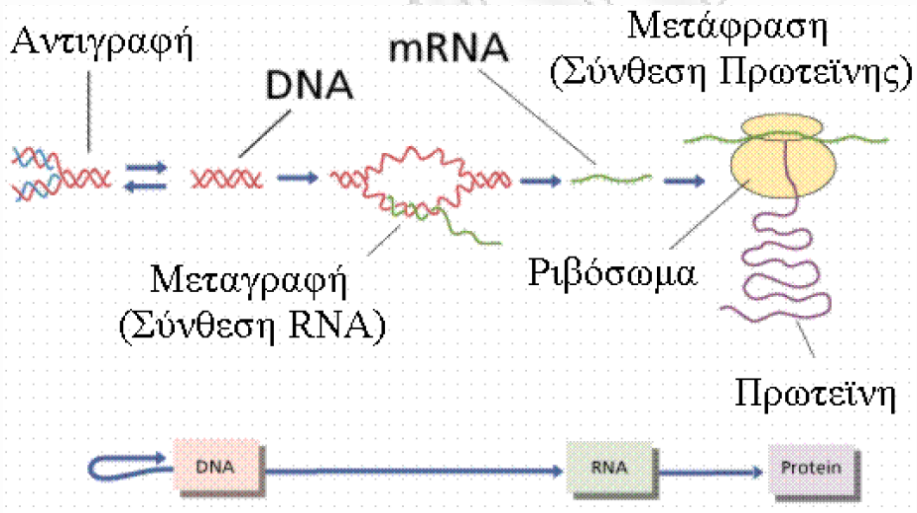
Για αρκετό καιρό οι επιστήμονες θεωρούσαν ότι όλη η ροή της γενετικής πληροφορίας γινόταν προς τη μία μόνο κατεύθυνση, δηλαδή ότι μόνο το DNA μπορούσε να μεταγραφεί σε RNA. Ωστόσο σήμερα έχει ανακαλυφθεί ότι κάποιοι ιοί έχουν RNA ως γενετικό υλικό και συνεπώς και το RNA έχει τη δυνατότητα σύνθεσης DNA. Επίσης σε ορισμένους ιούς έχει παρατηρηθεί ότι το RNA μπορεί να αυτοδιπλασιάζεται και κατά συνέπεια το παραπάνω σχήμα λαμβάνει την ακόλουθη τελική μορφή:

Αρχικά λοιπόν λαμβάνει μέρος η διαδικασία της αντιγραφής η οποία περιλαμβάνει το ξεδίπλωμα της διπλής έλικας του DNA και με τη βοήθεια ενός ενζύμου, της DNA πολυμεράσης και άλλων πρωτεϊνών δημιουργείται ένα αντίγραφο της αρχικής έλικας. Με αυτόν τον τρόπο διαιωνίζεται η γενετική πληροφορία. Κατά το στάδιο της μεταγραφής η πληροφορία του DNA μεταφέρεται στο mRNA μέσω της RNA πολυμεράσης ενώ στο στάδιο της μετάφρασης χρησιμοποιείται αυτή η πληροφορία για να κατασκευαστεί ένα πολυπεπτίδιο, δηλαδή πολλά αμινοξέα ενωμένα. Η μεταγραφή καθορίζει ποια γονίδια θα εκφραστούν, σε ποιους ιστούς και σε ποιο στάδιο.

1.2 Πρωτεΐνες

Οι **πρωτεΐνες** αποτελούν τα πιο διαδεδομένα και τα πιο πολυδιάστατα τόσο στη μορφή όσο και στη λειτουργία τους μακρομόρια. Ακόμη και σε ένα απλό κύτταρο των βακτηρίων

εντοπίζονται εκατοντάδες διαφορετικές πρωτεΐνες που κάθε μια από αυτές έχει και έναν ιδιαίτερο ρόλο. Οι πρωτεΐνες μπορεί να αποτελούν δομικό συστατικό του κυττάρου ή μπορεί να συνεργούν σε κάποια συγκεκριμένη λειτουργία του. Όσον αφορά τη δομή τους αποτελούνται από αμινοξέα τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας μια γραμμική αλυσίδα, την αλυσίδα των πολυπεπτιδίων. Η ακολουθία των αμινοξέων καθορίζεται από ένα γονίδιο και κωδικοποιείται με βάση το DNA. Ο γενετικός κώδικας κωδικοποιεί μόλις 20 αμινοξέα και για αυτό το λόγο πολλές πρωτεΐνες υφίστανται χημικές αλλαγές κατά τη **μετα-μεταγραφική** διαδικασία έτσι ώστε να μπορέσουν εν συνεχεία να λειτουργήσουν ορθά. Να συμπληρώσουμε ότι συχνά συνεργάζονται περισσότερες από μία πρωτεΐνες για να επιτύχουν μια συγκεκριμένη λειτουργία ή συσσωματώνονται έτσι ώστε να διαμορφωθούν σε σταθερά σύμπλοκα. Στο παρακάτω σχήμα δίνεται γραφικά η διαδικασία σύνθεσης πρωτεϊνών:



Σχήμα 1.2: Διαδικασία Σύνθεσης Πρωτεϊνών

1.3 Γονιδιακή έκφραση

Με τον όρο γονιδιακή έκφραση ορίζουμε τη διαδικασία μέσω της οποίας το γονίδιο μιας αλληλουχίας DNA μετατρέπεται σε λειτουργική πρωτεΐνη στο κύτταρο. Γονίδια τα οποία δεν έχουν τη δυνατότητα κωδικοποίησης πρωτεϊνών, όπως παραδείγματος χάριν τα rRNA και τα tRNA, δε μεταφράζονται και σε πρωτεΐνες. Η έναρξη της διαδικασίας ξεκινά με τη μεταγραφή και τη μετάφραση ενώ ακολουθεί το δίπλωμα της μεταφραστικής αλυσίδας, ορισμένες μετα-

μεταφραστικές τροποποιήσεις και τέλος η στόχευση. Να σημειώσουμε ότι η ποσότητα της πρωτεΐνης που τελικώς θα παραχθεί εξαρτάται από τον ιστό, το στάδιο της ανάπτυξης κατά το οποίο βρίσκεται ο οργανισμός και τη μεταβολική κατάσταση του κυττάρου. Κατά τη διαδικασία έκφρασης ενός γονιδίου παράγεται mRNA με βάση την αλληλουχία του DNA, το οποίο με τη σειρά του θα λειτουργήσει σαν οδηγός για τη σωστή τοποθέτηση των αμινοξέων στη σειρά έτσι ώστε να δημιουργηθεί μια λειτουργική πρωτεΐνη. Κατά τη διαδικασία έκφρασης λοιπόν συγκεκριμένων γονιδίων παράγεται μια πλειάδα αντιγράφων mRNA των εν λόγω γονιδίων. Συνεπώς η ποσότητα mRNA σε ένα κύτταρο ή σε έναν ιστό μπορεί να χρησιμοποιηθεί ως μέτρο της γονιδιακής έκφρασης. Ωστόσο πολλές φορές αντί για το mRNA χρησιμοποιείται το cDNA (*complementary DNA-συμπληρωματικό DNA*) το οποίο είναι ένα μόριο DNA που έχει συντεθεί από ώριμο mRNA, έτσι ώστε να είναι έτοιμο προς μετάφραση. Η σύνθεση του cDNA γίνεται από το ένζυμο αντίστροφη μεταγραφάση. Το cDNA χρησιμοποιείται ευρύτατα για την κλωνοποίηση γονιδίων ευκαρυωτικών κυττάρων σε προκαρυωτικά, ώστε να παραχθεί η πρωτεΐνη, που αυτά κωδικοποιούν, σε μεγάλες ποσότητες. Κάθε προκαρυωτικό κύτταρο που δέχεται το γονίδιο πολλαπλασιάζεται, παράγοντας έτσι ένα βακτηριακό κλώνο. Το σύνολο των βακτηριακών κλώνων που περιέχουν αντίγραφα όλων των ώριμων mRNA των γονιδίων που εκφράζονται σε έναν κυτταρικό τύπο ονομάζεται cDNA βιβλιοθήκη.

Για να καταλάβουμε όμως καλύτερα την παραπάνω διαδικασία και το λόγο ύπαρξής της θα παραθέσουμε μια ιστορία περί… ψαρέματος γονιδίων (www.scienceinschool.org). «*Μια φορά και έναν καιρό ήταν ένας ψαράς σε ένα μικρό χωριό. Κάθε μέρα ξυπνούσε και πήγαινε σε μια λίμνη για να ψαρέψει το βραδυνό του. Μια φορά έπιασε ένα λυκόψαρο, μια άλλη ένα χέλι-μερικές φορές έπιανε διαφορετικό ψάρι κάθε μέρα. Μία μέρα, αναρωτήθηκε πόσα διαφορετικά είδη ψαριών να υπήρχαν στην λίμνη και συγκεκριμένα πόσα ακριβώς ψάρια να υπήρχαν από κάθε είδος. Πώς μπορούσε να το βρει αυτό; Ήταν φανερό ότι δε θα μπορούσε να το καταφέρει με το να πιάνει μόνο ένα ψάρι κάθε φορά αφού η λίμνη μπορεί να περιείχε χιλιάδες ψάρια. Έτσι επισκέφτηκε το διαδίκτυο και βρήκε ένα βιβλίο που περιείχε 20.000 διαφορετικά είδη ψαριών γλυκού νερού, όπως και το δόλωμα που χρειάζεται κανείς για να πιάσει το καθένα από αυτά. Τελικά σκέφτηκε μια πολύπλοκη αλλά έξυπνη λύση: τοποθέτησε 20.000 διαφορετικές πετονιές, κάθε μία με πολλά αγκίστρια, μέσα στη λίμνη. Χρησιμοποιώντας την πληροφορία από το βιβλίο του τοποθέτησε ένα συγκεκριμένο είδος δολώματος στα αγκίστρια της κάθε πετονιάς έτσι ώστε να προσελκύει η κάθε μια και ένα συγκεκριμένο είδος ψαριού. Τα υπόλοιπα ήταν απλώς ένα παιχνιδάκι: περίμενε λίγο ώστε τα ψάρια να βρουν το δόλωμά τους και μάζεψε την ψαριά του*».

Φυσικά και η παραπάνω διαδικασία είναι φανταστική αλλά αξίζει να την αναφέρουμε μιας και η βασική αρχή στην οποία στηρίχθηκε ο ψαράς για να βρει πόσα ψάρια υπήρχαν μέσα στη λίμνη είναι ακριβώς η ίδια με τη βασική αρχή τεχνολογίας των μικροσυστοιχιών DNA, τεχνολογία η οποία αποτέλεσε και αποτελεί επανάσταση στον τρόπο παρακολούθησης των ζωντανών οργανισμών. Μέχρι και το 1990 οι βιολόγοι είχαν τη δυνατότητα ταυτόχρονης παρακολούθησης λίγων μόνο γονιδίων μέσα στο ίδιο κύτταρο. Κατά συνέπεια ήταν επιτακτική ανάγκη η αναζήτηση τρόπων έτσι ώστε να είναι δυνατή η ταυτόχρονη γονιδιακή έκφραση όλων των γονιδίων σε ένα κύτταρο. Για να επιτευχθεί η παραπάνω ανάγκη έπρεπε να βρεθεί το πόσα είδη mRNA υπήρχαν στο κύτταρο μια δεδομένη στιγμή και σε ποια γονίδια αντιστοιχούσαν αυτά τα μόρια mRNA. Επίσης ήταν απαραίτητο να δοθούν απαντήσεις που αφορούσαν το ποια γονίδια αλληλεπιδρούν μεταξύ τους και πόσο ενεργά (*active*) είναι αυτά κάτω από συγκεκριμένες συνθήκες. Οι ειδικοί λοιπόν δημιούργησαν κάτι παρόμοιο με τις 20.000 πετονιές του ψαρά το οποίο αποκάλεσαν μικροσυστοιχίες DNA (*DNA microarrays*).

Η κύρια υπόθεση αυτής της τεχνολογίας είναι η ακόλουθη: είναι ήδη γνωστό ότι ο φαινότυπος ενός κυττάρου σχετίζεται με την ποσότητα μετάφρασης των γονιδίων και το mRNA που υπάρχει σε κάθε κύτταρο εξαρτάται από τα ποια γονίδια μεταφράστηκαν. Συνεπώς το ποσοστό του mRNA θα μπορούσαμε να πούμε ότι είναι ο δείκτης του ***gene expression*** του κάθε γονιδίου και οι μικροσυστοιχίες βασίζονται στην ταυτόχρονη καταγραφή αυτού του *gene expression* για όλα τα γονίδια του οργανισμού. Τα *microarrays* είναι ιδιαιτέρως χρήσιμα διότι έχουν τη δυνατότητα να απαντούν στα ερωτήματα όπως τα ακόλουθα:

- Πόσο ενεργά είναι τα διαφόρου είδους γονίδια όταν βρίσκονται σε διαφορετικά κύτταρα ή διαφορετικά όργανα του σώματος.
- Πώς αλλάζει η ενεργητικότητα των γονιδίων κάτω από διάφορες συνθήκες, π.χ. στα διάφορα στάδια του κύκλου ζωής του κυττάρου, στις αλλαγές των περιβαλλοντικών συνθηκών ή στις ασθένειες.
- Ποια γονίδια φαίνεται να παρουσιάζουν παρόμοιο τρόπο έκφρασης και ποια έχουν τη δυνατότητα να συνεργάζονται μεταξύ τους.

1.4 Διαδικασία μικροσυστοιχιών

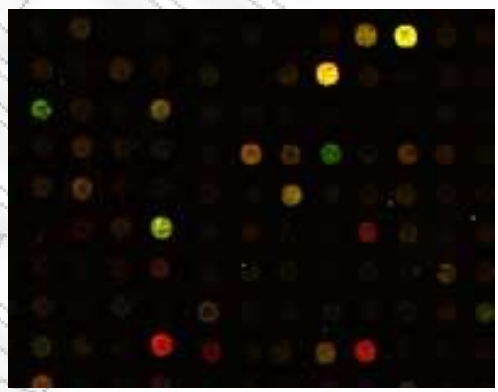
Σε ένα κομμάτι από επεξεργασμένο γυαλί ή και νάιλον ακινητοποιούνται κομμάτια DNA καθένα από τα οποία αντιστοιχεί και σε ένα γονίδιο. Αυτές οι αλληλουχίες DNA χρησιμοποιούνται ως δόλωμα για την προσέλκυση αντίστοιχων μορίων mRNA. Οι αλληλουχίες αυτές όταν εκτυπωθούν στο γυαλί φαίνονται σα μικρές κουκίδες και με αυτόν τον τρόπο δημιουργείται μια μικροσυστοιχία DNA.

Σχήμα 1.3: Εκτύπωση μιας Μικροσυστοιχίας

Η τοποθέτηση μικροσκοπικών κουκίδων στην επιφάνεια του γυαλιού γίνεται με ειδικά ρομπότ τα οποία φέρουν ειδικές κεφαλές που μεταφέρουν ένα υδατικό διάλυμα DNA και το εκτυπώνουν στο γυαλί σε μια διάταξη ειδικών μικρών κουκίδων. Ελέγχοντας κανείς τις μηχανές αυτές είναι σε θέση να ελέγξει όλα τα στοιχεία της συστοιχίας: αριθμό, μέγεθος και την απόσταση μεταξύ των κουκίδων. Σε μια πλάκα γυαλιού μπορούν να εκτυπωθούν περισσότερες από 20.000 κουκίδες κάθε μία από τις οποίες περιέχει δισεκατομμύρια αντιγράφων DNA ενός συγκεκριμένου γονιδίου.

Σχήμα 1.4: Λεπτομέρεια από μια Κεφαλή Εκτύπωσης Μικροσυστοιχιών

Σε ένα πείραμα μικροσυστοιχίας αρχικά απομονώνεται το mRNA από κύτταρα δύο καταστάσεων προς μελέτη όπως για παράδειγμα ένας κανονικός και ένας καρκινικός ιστός. Το mRNA απομονώνεται και σημαίνεται με μια φθορίζουσα χρωστική ουσία: πράσινο για τα κανονικά κύτταρα και κόκκινο για τα καρκινικά. Στη συνέχεια πραγματοποιείται η διαδικασία της υβριδοποίησης όπου τα mRNA του ελέγχου, τα πράσινα, και του τεστ, τα κόκκινα, αναμιγνύονται μαζί και τοποθετούνται στην επιφάνεια της μικροσυστοιχίας για να κολλήσουν (να γίνει η υβριδοποίηση δηλαδή) με το συμπληρωματικό ακινητοποιημένο κομμάτι DNA. Όσα μόρια mRNA δεν έχουν ακολουθία που να είναι συμπληρωματική με κανένα από τα γονίδια δε θα προσκολληθούν ενώ αυτά με τη μερικώς συμπληρωματική αλληλουχία DNA από ένα γονίδιο θα προσκολληθούν μερικώς (μη ειδική υβριδοποίηση) και στη συνέχεια θα απομακρυνθούν με τη διαδικασία του πλυσίματος. Στο τέλος της διαδικασίας πραγματοποιείται η σάρωση, όπου παρατηρείται ποια μόρια έχουν υβριδοποιηθεί και με ποια γονίδια. Επειδή τα μόρια του mRNA δεν είναι αντιληπτά με γυμνό μάτι, οι επιστήμονες έχουν τη δυνατότητα να υπολογίσουν την ποσότητα του mRNA έμμεσα υπολογίζοντας το φθορισμό της κάθε κουκίδας. Με χρήση ενός ειδικού σαρωτή με ακτίνες ***laser*** ο πράσινος και ο κόκκινος φθορισμός ανιχνεύονται ξεχωριστά και δημιουργούνται δύο εικόνες. Έπειτα γίνεται ο συνδυασμός των δύο αυτών εικόνων και το αποτέλεσμα είναι η χαρακτηριστική εικόνα με τις πολύχρωμες κουκίδες.



Σχήμα 1.5: Λεπτομέρεια από μια Σαρωμένη Μικροσυστοιχία

Μια κόκκινη κουκίδα υποδηλώνει ότι το γονίδιο είναι ενεργό στο καρκινικό κύτταρο και μια πράσινη ότι το γονίδιο είναι ενεργό στο κανονικό κύτταρο. Οι κίτρινες κουκίδες υποδηλώνουν την ύπαρξη του γονιδίου και στους δύο κυτταρικούς τύπους. Η ανάλυση των μικροσυστοιχιών ξεκινά με τη μέτρηση της έντασης φθορισμού κάθε κουκίδας. Η ένταση του σήματος που παράγεται από 1000 μόρια είναι διπλάσια από την ένταση που παράγεται από 500 μόρια. Κατά

αυτόν τον τρόπο λοιπόν οι μικροσυστοιχίες αποτελούν ένα μαζικό τρόπο για τη μελέτη της έκφρασης χιλιάδων γονιδίων από διαφορετικούς πληθυσμούς κυττάρων.

Να σημειώσουμε ότι, ξεφεύγοντας από την κλασική δομή των πινάκων πλέον αυτές οι μικροσυστοιχίες μετράνε την ποσότητα του στόχου έμμεσα αφού στην πραγματικότητα δε μπορούμε να μετρήσουμε την απόλυτη ποσότητα του mRNA που υπάρχει σε κάθε γονίδιο. Όπως αναφέρθηκε και προηγουμένως αυτό που μετράται είναι η ένταση του φθορισμού κάθε τοποθεσίας (***spot***) στον πίνακα αναπαράστασης φθορισμού σε σχέση με ένα δείγμα αναφοράς. Σε κάθε μέτρηση λοιπόν υπολογίζεται ο λόγος G_i που ορίζεται ως:

$$G_i = \log \frac{red_i}{green_i}.$$

Το *red* υποδηλώνει το επίπεδο έκφρασης του δείγματος και το *green* το επίπεδο έκφρασης του δείγματος αναφοράς για το γονίδιο *G* στο *i*-οστό πείραμα.

Οι κύριοι τρόποι πειραματισμού είναι δύο, με πιο δημοφιλή αυτόν που λειτουργεί με cDNA μικροπίνακες και περιγράφηκε προηγουμένως. Αυτός ο τρόπος δημιουργήθηκε από την ομάδα του *Pat Brown* στο *Standford*. Ο άλλος είναι με ολιγονουκλεοτιδικούς πίνακες (***oligo arrays***) και δημιουργήθηκε από το *Affymetrix Inc.* Στο δεύτερο τρόπο που περιλαμβάνει την τεχνική των *oligonucleotide arrays* αντί να τοποθετούνται ολόκληρα γονίδια στο *microarray* τοποθετούνται σύνολα από ακολουθίες DNA μήκους 25 βάσεων. Τα *oligos* αυτά συνθέτονται πάνω σε ένα ειδικό *chip* και τα δείγματα mRNA επεξεργάζονται ξεχωριστά αντί σε ζευγάρια.

Σχήμα 1.6: Affymetrix's Gene Chip

Στα βασικά βήματά τους οι δύο μέθοδοι δε διαφέρουν σημαντικά αφού και οι δύο μετρούν την έκφραση των επιπέδων για κάθε ακολουθία DNA και έτσι δημιουργούνται οι γονιδιακές

βάσεις δεδομένων. Να σημειώσουμε ότι υπάρχουν διάφοροι άλλοι τύποι *arrays*, λιγότερο δημοφιλείς όπως το ***array CGH*** (*Comparative Genomic Hybridization*) και το ***SNP array*** (*Single Nucleotide Polymorphism*) που αναζητά διαφορές στις ακολουθίες των χρωμοσωμάτων.

Μια βάση δεδομένων γονιδιακής έκφρασης αποτελείται από τον πίνακα δεδομένων γονιδιακής έκφρασης, την επισήμανση των γονιδίων και την επισήμανση των δειγμάτων. Τέτοιες μεγάλες βάσεις είναι διαθέσιμες στο κοινό μέσω του διαδικτύου ενώ άλλες μικρότερες που αναφέρονται στη βιβλιογραφία υλοποιούνται από μικρά εργαστήρια. Οι πιο δημοφιλείς εξ' αυτών είναι η ***Array Express*** από το *European Bioinformatics Insitute* και η ***NBCI's Gene Expression Omnibus* (*GEO*).** Επίσης στο κοινό είναι διαθέσιμες και γονιδιακές βάσεις δεδομένων με ***ESTs*.** Τα ESTs (***Expressed Sequence Tag***) αποτελούνται από υποακολουθίες cDNA ακολουθιών. Χρησιμοποιούνται για την ανίχνευση αντιγράφων γονιδίων, για την ανακάλυψη νέων καθώς και για τον περαιτέρω καθορισμό της γονιδιακής αλυσίδας. Ένα EST μπορεί να προκύψει από την αλληλουχία ενός κλώνου mRNA ή αλλιώς από μερικές χιλιάδες ζεύγη βάσεων που ξεκινούν από το τέλος ενός cDNA. Οι αλληλουχίες αυτές δεν είναι πολυπληθείς αφού το μέγεθός τους περιορίζεται στα 500 με 800 νουκλεοτίδια. Επειδή αυτοί οι κλώνοι που παράγονται είναι συμπληρωματικοί με το mRNA, τα ESTs ουσιαστικά αντιπροσωπεύουν τμήματα γονιδίων που εκφράζονται. Αξίζει να αναφέρουμε ότι στις βιβλιοθήκες δεδομένων μπορούν να δοθούν ως αλληλουχίες cDNA ή mRNA ενώ η κατάσταση στην οποία έχουν παραχθεί τα ESTs μπορεί να δώσει πληροφορίες για τις συνθήκες υπό τις οποίες αντιδρά το συγκεκριμένο γονίδιο (ιστός, όργανο, δυσλειτουργίες κλπ). Τέλος στην επιστημονική κοινότητα η ανίχνευση των ESTs έχει γίνει αρκετά γρήγορα με περίπου 65,9 εκατομμύρια ESTs να είναι πλέον διαθέσιμα στο ευρύ κοινό σε γονιδιακές βάσεις δεδομένων (***GenBank*** 18/6/10).

1.5 Στάδια μελέτης

Στη συνέχεια παραθέτουμε ένα σχεδιάγραμμα για να δείξουμε ποια είναι τα στάδια ενός πειράματος για την ανίχνευση και τη μελέτη γονιδίων όπου υπεισέρχεται και η επιστήμη της στατιστικής. Επιπλέον γίνεται φανερό που εφαρμόζεται η διαδικασία της ομαδοποίησης που αποτελεί βασικό αντικείμενο αυτής της εργασίας.

Σχήμα 1.7: Στάδια Πειραματικής Διαδικασίας

Ορισμένα από τα χαρακτηριστικά ερωτήματα που βρίσκουν απάντηση από την παραπάνω μεθοδολογία είναι τα εξής:

- Ποια γονίδια εκφράζονται περισσότερο κατά μέσο όρο; (Απαιτείται ο υπολογισμός της μέσης τιμής για κάθε γραμμή και στήλη).
- Σε ποια πειράματα παρατηρήθηκαν κατά μέσο όρο οι υψηλότερες και οι χαμηλότερες τιμές;
- Ποια γονίδια έχουν τη μεγαλύτερη ή τη μικρότερη διαφορά; (απαιτείται ο υπολογισμός της διασποράς κάθε γραμμής και στήλης).
- Ποιες από τις πειραματικές συνθήκες που εξετάστηκαν διαφέρουν περισσότερο και ποιες λιγότερο;

Επιπροσθέτως αξίζει να σημειώσουμε, όσον αφορά το κομμάτι της ομαδοποίησης, ότι στις διάφορες επιστημονικές εργασίες πραγματοποιείται είτε *clustering* γονιδίων δηλαδή αναζήτηση εκείνων των γονιδίων τα οποία μπορούν και εκφράζονται μαζί με τον ίδιο τρόπο είτε *clustering* δειγμάτων όπου στόχος είναι η εύρεση ατόμων ή θεραπειών που παρουσιάζουν παρόμοια προφίλ και συνεπώς δημιουργούν ομάδα.

Στη συνέχεια αυτής της εργασίας θα παρουσιαστούν οι κλασσικές μεθοδολογίες δημιουργίας ομάδων, τα σημαντικότερα κλασσικά μέτρα επικύρωσης αυτών αλλά και νέα μέτρα επικύρωσης τα οποία έχουν δημιουργηθεί τα τελευταία χρόνια με στόχο την εφαρμογή τους σε δεδομένα γονιδιακής έκφρασης. Στο τελευταίο κεφάλαιο θα παρουσιαστεί με αναλυτικό τρόπο η εφαρμογή των περισσοτέρων θεωρητικών αποτελεσμάτων σε δεδομένα από το χώρο της μοριακής βιολογίας με τη βοήθεια του στατιστικού πακέτου R.

Κεφάλαιο 2

Ιεραρχικές και μη Ιεραρχικές Μέθοδοι Ομαδοποίησης

2.1 Εισαγωγή

Η ανακάλυψη νέας βιολογικής γνώσης από την ανάλυση των γονιδιακών δεδομένων γίνεται με τη βοήθεια κυρίως των τεχνικών ομαδοποίησης δεδομένων σε συστάδες. Τα τελευταία χρόνια στο χώρο της βιοπληροφορικής έχει γίνει πολύ μεγάλη προσπάθεια έτσι ώστε οι κλασσικές στατιστικές μέθοδοι ομαδοποίησης δεδομένων να εφαρμοστούν στην ανάλυση των γονιδίων και να τροποποιηθούν ή να επεκταθούν έτσι ώστε να μπορούν να αντιμετωπίσουν έγκαιρα και αποτελεσματικά τυχόν ιδιαιτερότητες της συγκεκριμένης ομάδας δεδομένων.

Η εξερεύνηση πολύπλοκων συνόλων δεδομένων για τα οποία είναι διαθέσιμη λίγη ή και καθόλου πληροφορία για την κατανομή τους μπορεί να γίνει με τη χρήση μεθόδων ομαδοποίησης αυτών σε συστάδες ((*Duda*, 2001), (*Everitt*, 1993), (*Hastie*, 2001), (*Jain*,1999)). Η διαδικασία ομαδοποίησης σε συστάδες δίνεται στο επόμενο σχήμα:

Βήμα 1: Προεργασία Δεδομένων

- Επιλογή χαρακτηριστικών προς μελέτη
- Κανονικοποίηση Δεδομένων
- Επιλογή της συνάρτησης απόστασης

Βήμα 2: Ανάλυση σε Συστάδες

- Επιλογή του αλγορίθμου
- Επιλογή των παραμέτρων του αλγορίθμου
- Εφαρμογή του αλγορίθμου

Βήμα 3: Επικύρωση των συστάδων που δημιουργήθηκαν

- Επιλογή των τεχνικών επικύρωσης
- Εφαρμογή των τεχνικών επικύρωσης

Σχήμα 2.1: Διαδικασία Ομαδοποίησης

Το πρώτο βήμα περιλαμβάνει μια σειρά από τεχνικές επιλογής των δεδομένων, μετασχηματισμούς αυτών και επιλογή της κατάλληλης συνάρτησης απόστασης έτσι ώστε να εξασφαλιστεί ότι οι συστάδες που δημιουργούνται έχουν φυσική ερμηνεία. Στο δεύτερο βήμα γίνεται η εφαρμογή του κατάλληλου αλγορίθμου για την εύρεση των συστάδων ενώ στο τρίτο βήμα ελέγχονται και επικυρώνονται τα αποτελέσματα του δεύτερου βήματος. Τα αποτελέσματα της ανάλυσης σε συστάδες μπορεί να επηρεαστούν σημαντικά από τις επιλογές που γίνονται στα δύο πρώτα βήματα και συνεπώς όση διαθέσιμη πληροφορία έχουμε για τα δεδομένα πρέπει να χρησιμοποιηθεί για να ελεγχθεί αν οι επιλογές που κάναμε ήταν οι σωστές.

Να σημειώσουμε ότι παρόλο που οι τεχνικές επικύρωσης των συστάδων που προκύπτουν είναι αναπόσπαστο κομμάτι της όλης διαδικασίας, στην ανάλυση των γονιδιακών δεδομένων παραλειπόταν συχνά αυτό το βήμα με αποτέλεσμα να μην υπήρχε μέχρι προσφάτως, σημαντική πρόοδος στη βιβλιογραφία στο συγκεκριμένο κλάδο (*Dudes and Jain*, 1979). Το τελευταίο συνέβαινε επειδή ελλείψει σχετικών ερευνών οι υπάρχοντες αλγόριθμοι επικύρωσης δεν εξελίσσονταν, δεν μπορούσαν να εντοπιστούν τα δυνατά τους σημεία και οι αδυναμίες τους και κατ' επέκταση δεν μπορούσαμε να καταλάβουμε ποιος είναι ο καλύτερος αλγόριθμος ομαδοποίησης για τα συγκεκριμένης μορφής δεδομένα. Ωστόσο, τα τελευταία χρόνια με την πρόοδο της βιοπληροφορικής, νέες τεχνικές επικύρωσης έχουν δημιουργηθεί οι οποίες είναι εξειδικευμένες σε τέτοιου είδους δεδομένα (*Datta and Datta,* 2006). Οι τεχνικές αυτές θα περιγραφούν αναλυτικά στο Κεφάλαιο 3.

2.2 Μέτρα Απόστασης και Μέτρα Ομοιότητας

Οι παραδοσιακές τεχνικές ομαδοποίησης σε συστάδες χωρίζονται σε ιεραρχικές και μη ιεραρχικές ενώ υπάρχουν και οι τεχνικές που βασίζονται στην κατανομή των δεδομένων. Για τα γονιδιακά δεδομένα ο διαχωρισμός γίνεται λίγο διαφορετικά αφού στηρίζεται στην βελτιστοποίηση του κριτηρίου που χρησιμοποιείται από τον αλγόριθμο ομαδοποίησης. Στόχος της ανάλυσης κατά συστάδες είναι η δημιουργία ομάδων μέσα στις οποίες οι παρατηρήσεις διαφέρουν όσο το δυνατόν λιγότερο (ιδιότητα της συμπάγειας–*compactness*) ενώ οι παρατηρήσεις διαφορετικών ομάδων διαφέρουν όσο το δυνατόν περισσότερο.

Πριν παρουσιαστούν όμως οι διάφοροι αλγόριθμοι ομαδοποίησης θα πρέπει να εισαχθεί ένα μέτρο εγγύτητας ή ομοιότητας των δεδομένων μεταξύ τους. Συχνά δεν υπάρχει ένας σαφής διαχωρισμός για το ποιο από τα υπάρχοντα μέτρα είναι το καταλληλότερο, ωστόσο συνηθίζεται

να λαμβάνεται υπόψην το είδος της μεταβλητής (διακριτή, συνεχής, δίτιμη) και η κλίμακα μέτρησης των μεταβλητών (ονομαστική, διατακτική, διαστηματική). Η έννοια της εγγύτητας περιγράφεται με τη βοήθεια κάποιου μέτρου απόστασης.

Στο σημείο αυτό αξίζει να λάβουμε υπόψην μας ότι σε ένα κλασσικό πρόβλημα ανάλυσης δεδομένων σε συστάδες έχουμε ένα δείγμα n ατόμων ή αντικειμένων από ένα πληθυσμό, και σε κάθε άτομο παρατηρούμε p χαρακτηριστικά, δηλαδή τυχαίες μεταβλητές. Οι $n \times p$ παρατηρήσεις συγκεντρώνονται σε έναν πίνακα $X=(x_{ij})$ με n γραμμές και p στήλες. Ένας τέτοιος πίνακας καλείται πίνακας πρωτογενών δεδομένων (*raw data table*) ή πίνακας δεδομένων. Το στοιχείο (i,j) του πίνακα περιέχει την τιμή x_{ij} που παρατηρήθηκε στο i άτομο για το j χαρακτηριστικό.

$$\begin{array}{c} \\ n \text{ άτομα} \end{array} \begin{array}{c} \overset{p \text{ μεταβλητές}}{\downarrow} \\ \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array}$$

Αν συμβολίσουμε με $x_i=(x_{i1}, x_{i2}, \dots, x_{ip})$ το διάνυσμα των παρατηρήσεων για τις p μεταβλητές που αφορά το i άτομο (i=1,2,…,n), η **Ευκλείδεια απόσταση** ανάμεσα σε δύο p-διάστατες παρατηρήσεις $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ και $\mathbf{x}_j = (x_{j1}, x_{ji2}, \dots, x_{jp})'$ ορίζεται ως:

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}.$$

Η Ευκλείδεια απόσταση ικανοποιεί τις εξής ιδιότητες:

i. $d_{ij} \geq 0$ για κάθε i,j και $d_{ij}=0 \Leftrightarrow i=j$
ii. $d_{ij} \leq d_{is} + d_{sj}$ (τριγωνική ανισότητα)
iii. $d_{ij} = d_{ji}$ (συμμετρική ιδιότητα)

Θα μπορούσαμε να πούμε ότι οποιαδήποτε συνάρτηση ικανοποιεί τις παραπάνω τρεις ιδιότητες ορίζει μια απόσταση. Η Ευκλείδεια απόσταση χρησιμοποιείται συχνά για συνεχή δεδομένα και παρόλο που αποτελεί την πιο διαδεδομένη απόσταση θα πρέπει να λάβουμε

υπόψην ότι επηρεάζεται από την κλίμακα μέτρησης και από τυχόν ακραίες τιμές που υπάρχουν στα δεδομένα μας.

Ένα άλλο μέτρο απόστασης είναι η **απόσταση Minkowski** η οποία ορίζεται ως:

$$d_{ij} = \left[\sum_{r=1}^{p} \left| x_{ir} - x_{jr} \right|^{m} \right]^{1/m}.$$

Ο παραπάνω τύπος για m=1 δίνει την απόσταση **Manhattan ή City Block** ανάμεσα σε δύο σημεία στον p-διάστατο χώρο. Για m=2 έχουμε τη γνωστή μας Ευκλείδεια απόσταση. Γενικά μεταβάλλοντας τις τιμές που παίρνει το m μεταβάλλουμε την επιρροή των μεγάλων διαφορών στη διαμόρφωση της τιμής της απόστασης.

Δύο ακόμη δημοφιλή μέτρα αποστάσεων είναι η **Canberra Metric** και ο **Συντελεστής Czekanowski** τα οποία ορίζονται από τους ακόλουθους τύπους.

$$d_{ij} = \sum_{r=1}^{p} \frac{\left| x_{ir} - x_{jr} \right|}{\left| x_{ir} \right| + \left| x_{jr} \right|} \quad (\textit{Canberra metric}),$$

$$d_{ij} = 1 - \frac{2\sum_{r=1}^{p} \min(x_{ir}, x_{jr})}{\sum_{r=1}^{p} (x_{ir} + x_{jr})} \quad (\textit{Czekanowski coefficient}).$$

Τέλος υπάρχει και η **απόσταση Chebychev** η οποία σε αντίθεση με όλες τις προηγούμενες αποστάσεις δεν χρησιμοποιεί όλες τις αποκλίσεις αλλά μόνο τη μεγαλύτερη από αυτές. Η απόσταση Chebychev θεωρεί ότι δύο παρατηρήσεις είναι διαφορετικές αν παρουσιάζουν μεγάλες διαφορές σε μια τουλάχιστον μεταβλητή και δίνεται από τον τύπο:

$$d_{ij} = \max_{r=1,2,\ldots,p} \left| x_{ir} - x_{jr} \right| \quad .$$

Παράδειγμα 2.2.1

Στον Πίνακα 2.1 δίνονται οι βαθμολογίες n=4 φοιτητών σε p=5 εξαμηνιαία μαθήματα και υπολογίζεται ο πίνακας κάποιων βασικών αποστάσεων.

Φοιτητής	Βαθμολογίες στο *i* μάθημα				
	1	**2**	**3**	**4**	**5**
1	9	8	8	7	10
2	8	7	6	5	6
3	5	3	3	4	5
4	7	7	7	8	8

Πίνακας 2.1: Πίνακας Βαθμολογιών

Η Ευκλείδεια απόσταση, η απόσταση Minkowski, η απόσταση Manhattan και η απόσταση Chebychev μεταξύ των 4 φοιτητών δίνονται στους επόμενους πίνακες:

Φοιτητής	Euclidean Distance			
	1	**2**	**3**	**4**
1	.000	5.099	10.000	3.317
2	5.099	.000	6.000	3.873
3	10.000	6.000	.000	7.810
4	3.317	3.873	7.810	.000

Πίνακας 2.1.1: Ευκλείδεια Απόσταση

Φοιτητής	Minkowski (4) Distance			
	1	**2**	**3**	**4**
1	.000	4.127	6.858	2.432
2	4.127	.000	4.527	3.154
3	6.858	4.527	.000	5.423
4	2.432	3.154	5.423	.000

Πίνακας 2.1.2: Απόσταση Minkowski

Φοιτητής	Manhattan or City Block Distance			
	1	**2**	**3**	**4**
1	.000	10.000	22.000	7.000
2	10.000	.000	12.000	7.000
3	22.000	12.000	.000	17.000
4	7.000	7.000	17.000	.000

Πίνακας 2.1.3: Απόσταση Manhattan

Φοιτητής	Chebychev Distance			
	1	2	3	4
1	.000	4.000	5.000	2.000
2	4.000	.000	4.000	3.000
3	5.000	4.000	.000	4.000
4	2.000	3.000	4.000	.000

Πίνακας 2.1.4: Απόσταση Chebychev

Στη συνέχεια θα παρουσιαστούν οι μέθοδοι με τις οποίες υπολογίζονται οι αποστάσεις μεταξύ αντικειμένων ή ατόμων των οποίων οι παρατηρήσεις είναι δίτιμες μεταβλητές. Μια τέτοια περίπτωση θα μπορούσε να προκύψει όταν οι μεταβλητές καταγράφουν με 1 την παρουσία ενός συγκεκριμένου χαρακτηριστικού για ένα άτομο ενώ η τιμή 0 χρησιμοποιείται για να δηλώσει την απουσία του. Έστω ότι x_{ir} είναι το σκορ (1 ή 0) για τη *r*-οστή δίτιμη μεταβλητή στο *i* άτομο-αντικείμενο και x_{jr} είναι το σκορ (πάλι 1 ή 0) της *r*-οστής δίτιμης μεταβλητής για το *j* άτομο-αντικείμενο, $r = 1, 2, \ldots, p$. Τότε θα έχουμε ότι:

$$(x_{ir}-x_{jr})^2 = \begin{cases} 0, & x_{ir} = x_{jr} = 1 \text{ ή } x_{ir} = x_{jr} = 0 \\ 1, & x_{ir} \neq x_{jr} \end{cases} .$$

Συνεπώς η Ευκλείδεια απόσταση μας παρέχει το πλήθος των ασυμφωνιών μεταξύ δύο ατόμων όσον αφορά τα χαρακτηριστικά τα οποία μελετάμε. Αν και με αυτόν τον ορισμό θα μπορούσαμε να είχαμε μια ένδειξη για το πλήθος των ασυμφωνιών, δεν είναι κατάλληλο μέτρο διότι δίνει το ίδιο βάρος στις συμφωνίες 1-1 (παρουσία του χαρακτηριστικού και στα 2 άτομα) και στις συμφωνίες 0-0 (απουσία του χαρακτηριστικού και στα 2 άτομα). Για το λόγο αυτό έχει οριστεί μια πληθώρα μέτρων που καταφέρνουν να αντιμετωπίσουν αυτήν την ιδιαιτερότητα. Για να ορίσουμε αυτά τα μέτρα κατάλληλα θα κατασκευάσουμε έναν πίνακα συμφωνιών και ασυμφωνιών (πίνακας συνάφειας) ανάμεσα στα άτομα *i* και *j*.

		Άτομο *j*		
Άτομο *i*		**1**	**0**	**Άθροισμα**
	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	Άθροισμα	*a+c*	*b+d*	*p=a+b+c+d*

Πίνακας 2.2: Πίνακας Συνάφειας

Στον Πίνακα 2.2 *a* είναι το πλήθος του συνδυασμού (1,1), δηλαδή το πλήθος των περιπτώσεων όπου τα χαρακτηριστικά είναι παρόντα και στα 2 άτομα. Το *b* είναι το πλήθος του συνδυασμού (1,0) που σημαίνει ότι τα χαρακτηριστικά είναι παρόντα στο άτομο *i* και απόντα στο άτομο *j*. Το *c* δηλώνει το πλήθος του συνδυασμού (0,1) που σημαίνει ότι τα χαρακτηριστικά είναι απόντα στο *i* άτομο και παρόντα στο *j*. Τέλος το *d* είναι το πλήθος του συνδυασμού (0,0) που δηλώνει την απουσία του χαρακτηριστικού και από τα 2 άτομα. Με βάση τους παραπάνω συμβολισμούς τα πιο συνηθισμένα μέτρα ομοιότητας που έχουν προταθεί είναι τα εξής:

Ονομασία	**Τύπος**	**Τεκμηρίωση**
Simple matching	$\frac{a+d}{p}$	Ίσα βάρη για τους συνδυασμούς 1-1 και 0-0
Sokal and Sneath	$\frac{2(a+d)}{2(a+d)+b+c}$	Διπλάσιο βάρος για τους συνδυασμούς 1-1 κι 0-0
Rogers and Tanimoto	$\frac{a+d}{a+d+2(b+c)}$	Διπλάσιο βάρος για τις ασυμφωνίες
Russel and Rao	$\frac{a}{p}$	Δε λαμβάνει υπόψην τις συμφωνίες 0-0 στον αριθμητή
Jaccard	$\frac{a}{a+b+c}$	Δε λαμβάνονται υπόψην οι συμφωνίες 0-0
Dice and Sorensen	$\frac{2a}{2a+b+c}$	Δε λαμβάνονται υπόψην οι συμφωνίες 0-0 ενώ διπλασιάζονται τα βάρη για τις συμφωνίες 1-1
Sokal and Sneath	$\frac{a}{a+2(b+c)}$	Δε λαμβάνονται υπόψην οι συμφωνίες 0-0 ενώ διπλασιάζονται τα βάρη για τις ασυμφωνίες
Kulczynski	$\frac{a}{(b+c)}$	Λόγος των συμφωνιών προς τις ασυμφωνίες με τη συμφωνία 0-0 να μη λαμβάνεται καθόλου υπόψην

Πίνακας 2.3: Συντελεστές Ομοιότητας

Παράδειγμα 2.2.2

Έστω ότι σε δύο άτομα παρατηρούμε την ύπαρξη ή την απουσία πέντε διαφορετικών χαρακτηριστικών. Τα αποτελέσματα δίνονται στον Πίνακα 2.4.1:

Άτομα	**Χαρακτηριστικά**				
	1	**2**	**3**	**4**	**5**
1	0	1	1	1	0
2	0	0	1	1	1

Πίνακας 2.4.1: Πίνακας Χαρακτηριστικών

Ο πίνακας συνάφειας που προκύπτει είναι ο εξής:

Άτομο 2				
Άτομο 1		**1**	**0**	**Άθροισμα**
	1	2	1	3
	0	1	1	2
	Άθροισμα	3	2	5

Πίνακας 2.4.2: Πίνακας Συνάφειας

Ενδεικτικά υπολογίζουμε τους δείκτες simple matching, Russel and Rao, Jaccard , Sokal and Sneath οι οποίοι έχουν τιμή 3/5,2/5,1/2,1/3 αντίστοιχα.

Να σημειώσουμε ότι, αν έχουμε στη διάθεσή μας ένα μέτρο απόστασης, είναι δυνατόν να κατασκευάσουμε σχετικά εύκολα μέτρα ομοιότητας. Για παράδειγμα θέτουμε: $s_{ij} = \frac{1}{1+d_{ij}}$ όπου $0 < s_{ij} \leq 1$ να είναι η ομοιότητα ανάμεσα στα αντικείμενα *i* και *j* με αντίστοιχη απόσταση d_{ij}. Αντίστοιχα αν έχουμε ορίσει ένα μέτρο ομοιότητας s_{ij} τότε μπορούμε να δημιουργήσουμε την αντίστοιχη απόσταση μέσω του τύπου:

$$d_{ij} = \sqrt{2(1-s_{ij})}\,.$$

Ο παραπάνω τύπος δεν εγγυάται την ισχύ της τριγωνικής ανισότητας και κατά συνέπεια τα μέτρα απόστασης δεν μπορούν πάντα να κατασκευαστούν με την βοήθεια των μέτρων

ομοιότητας. Ωστόσο ο Gower έδειξε ότι αυτό μπορεί να γίνει αν ο πίνακας ομοιότητας $[s_{ij}]_{nxn}$ είναι μη αρνητικά ορισμένος ή αρνητικά ημιορισμένος.

Σε περίπτωση που οι μεταβλητές που μελετάμε δεν είναι όλες του ίδιου τύπου (συνεχείς, διακριτές, δίτιμες) ο Gower πρότεινε τον εξής τρόπο υπολογισμού των μέτρων ομοιότητας και απόστασης. Έστω $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jp})'$ οι παρατηρήσεις δύο ατόμων με τις μεταβλητές που καταγράφηκαν να είναι διαφορετικού τύπου. Η ομοιότητά τους μετριέται μέσω του συντελεστή:

$$s_{ij} = \frac{\sum_{r=1}^{p} w_{ij}(r) s_{ij}(r)}{\sum_{r=1}^{p} w_{ij}(r)}.$$

Όπου

- αν η μεταβλητή *r* είναι συνεχής θέτουμε:

$$s_{ij}(r) = 1 - \frac{|x_{ir} - x_{jr}|}{R_r}$$

με $R_r = \max_{i=1,\ldots,p} x_{ir} - \min_{i=1,\ldots,p} x_{ir}$ το εύρος της *r*-οστής μεταβλητής.

- αν η μεταβλητή *i* είναι διακριτή θέτουμε:

$$s_{ij}(r) = \begin{cases} 1, & \text{αν } x_{ir} = x_{jr} \\ 0, & \text{αλλιώς .} \end{cases}$$

Τα βάρη $w_{ij}(r)$ παίρνουν την τιμή 1 ή 0 ανάλογα με το αν η σύγκριση στην *r*-οστή μεταβλητή έχει νόημα ή όχι.

2.3 Ιεραρχικές Μέθοδοι Ομαδοποίησης

Οι ιεραρχικές μέθοδοι ομαδοποίησης προκύπτουν από μια σειρά διαδοχικών συνενώσεων ή από μια σειρά διαδοχικών διαιρέσεων των δεδομένων. Οι συσσωρευτικές ιεραρχικές μέθοδοι (*agglomerative hierarchical methods*) θεωρούν ότι κάθε άτομο αποτελεί και μια ομάδα οπότε

αρχικά οι συστάδες μας είναι στο πλήθος τόσες όσα είναι τα αντικείμενα που μελετούμε. Τα αντικείμενα που είναι μεταξύ τους πιο όμοια ή βρίσκονται πιο κοντά ενώνονται και σχηματίζουν ομάδες/συστάδες. Η διαδικασία αυτή ολοκληρώνεται μόλις όλα τα δεδομένα βρεθούν σε μια και μόνο συστάδα (*Gordon,* 1999).

Οι διαιρετικές μέθοδοι ομαδοποίησης (*divisive hierarchical methods*) δουλεύουν με τον ακριβώς αντίθετο τρόπο. Η αρχική μας ομάδα αποτελείται από το σύνολο των αντικειμένων και χωρίζεται σε δύο υπό-ομάδες οι οποίες διαφέρουν όσο το δυνατόν περισσότερο μεταξύ τους. Αυτές οι δύο συστάδες στη συνέχεια διαιρούνται και σε άλλες υπό-ομάδες ώσπου στο τέλος κάθε αντικείμενο να αποτελεί μια ομάδα. Τα αποτελέσματα τόσο των συσσωρευτικών όσο και των διαιρετικών μεθόδων ομαδοποίησης μπορούν να παρασταθούν γραφικά σε ένα δενδρόγραμμα. Το δενδρόγραμμα αναπαριστά τις συνενώσεις ή τις διαιρέσεις που έχουν πραγματοποιηθεί στα διαδοχικά στάδια της όλης διαδικασίας (*Kaufman, Rousseuw,* 1990).

Αρχικά θα μελετήσουμε τις ιεραρχικές μεθόδους συνένωσης οι οποίες είναι κατάλληλες για ομαδοποίηση τόσο αντικειμένων όσο και μεταβλητών και στη συνέχεια θα μελετηθούν οι διαχωριστικές μέθοδοι. Θα παρουσιαστούν:

- η μέθοδος της απλής συνένωσης ή μέθοδος του κοντινότερου γείτονα (*single linkage – nearest neighbor*)
- η μέθοδος της πλήρους συνένωσης ή μέθοδος του μακρινότερου γείτονα (*complete linkage - farthest neighbor*)
- η μέθοδος της μέσης συνένωσης (*average linkage - average distance-UPGMA*)
- η μέθοδος του *Ward*
- η μέθοδος της διαχωριστικής ανάλυσης (*Divisive Analysis - DIANA*)

Στη συνέχεια παρουσιάζονται τα βήματα που ακολουθούνται κατά τη διαδικασία συσσωρευτικής ιεραρχικής ομαδοποίησης *N* αντικειμένων.

- Αρχικά διαθέτουμε *N* συστάδες, η κάθε μια από τις οποίες περιέχει ένα μόνο αντικείμενο, και έναν *N*x*N* συμμετρικό πίνακα αποστάσεων (ή μέτρων ομοιότητας) $D=\{d_{ij}\}$.
- Στη συνέχεια αναζητούμε στον πίνακα αποστάσεων το ζεύγος με τις κοντινότερες (πιο όμοιες) συστάδες. Έστω ότι η απόσταση ανάμεσα στις πιο όμοιες συστάδες *U* και *V* είναι η d_{UV}.

- Ενώνουμε τις συστάδες *U* και *V*. Ονομάζουμε τη νεοσχηματιζόμενη συστάδα (*UV*), ενημερώνουμε τον πίνακα αποστάσεων διαγράφοντας τις γραμμές και τις στήλες που αντιστοιχούν στις κλάσεις *U* και *V* ενώ συγχρόνως προσθέτουμε μια γραμμή και μια στήλη οι οποίες δίνουν την απόσταση ανάμεσα στην κλάση (*UV*) και τις υπόλοιπες κλάσεις.
- Επαναλαμβάνουμε τα δύο προηγούμενα βήματα *N*-1 φορές συνολικά. Μόλις ο αλγόριθμος τερματίσει όλα τα *N* αντικείμενα θα βρίσκονται σε μία συστάδα.

Α) Μέθοδος της Απλής Συνένωσης (*Single Linkage*)

Οι είσοδοι στον αλγόριθμο της απλής συνένωσης είναι οι αποστάσεις ή οι ομοιότητες ανάμεσα σε ζεύγη αντικειμένων. Οι ομάδες δημιουργούνται συνενώνοντας τους κοντινότερους γείτονες, δηλαδή τα άτομα που απέχουν τη μικρότερη απόσταση ή τα άτομα που παρουσιάζουν τη μεγαλύτερη ομοιότητα. Αρχικά πρέπει να βρεθεί η μικρότερη απόσταση στον πίνακα $D=\{d_{ik}\}$ και να ενωθούν τα αντίστοιχα αντικείμενα, έστω τα *U* και *V* έτσι ώστε να προκύψει η νέα κλάση (*UV*). Οι αποστάσεις ανάμεσα στην κλάση (*UV*) και οποιαδήποτε από τις υπόλοιπες, έστω *W*, υπολογίζονται από τον τύπο:

$$d_{(UV)W}=\min\{d_{UW}, d_{VW}\}$$

όπου οι ποσότητες d_{UW} και d_{VW} είναι οι αποστάσεις ανάμεσα στους κοντινότερους γείτονες των κλάσεων *U* και *W* και των κλάσεων *V* και *W* αντίστοιχα. Τα αποτελέσματα της μεθόδου της απλής συνένωσης μπορούν να παρασταθούν γραφικά με τη βοήθεια ενός δενδρογράμματος του οποίου τα κλαδιά θα αναπαριστούν τις κλάσεις. Τα σημεία στα οποία γίνονται οι συνενώσεις των κλάσεων αναπαριστούν την απόσταση ή την ομοιότητα των δύο κλάσεων που ενώνονται. Να σημειώσουμε ότι, αν δύο διαφορετικές ομάδες έχουν κάποιο σημείο ή κάποιο σύνολο κοντινών σημείων που να τις συνδέει, τότε ο αλγόριθμος δεν έχει την δυνατότητα να τις κρατήσει ξεχωριστά και συνεπώς οι δύο αυτές ομάδες ενώνονται. Το φαινόμενο αυτό καλείται φαινόμενο της αλυσίδας (*chaining effect*) και οδηγεί σε μη συμπαγείς ομάδες. Ένα άλλο χαρακτηριστικό της μεθόδου είναι ότι συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες πολύ μικρές.

Β) **Μέθοδος της Πλήρους Συνένωσης *(Complete Linkage)***

Η μέθοδος της πλήρους συνένωσης ακολουθεί το ίδιο σκεπτικό με τη μέθοδο της απλής συνένωσης με τη μόνη διαφορά ότι ως απόσταση ανάμεσα σε δύο ομάδες/κλάσεις θεωρούμε τη μεγαλύτερη απόσταση (ή μικρότερη ομοιότητα) ανάμεσα σε δύο στοιχεία που βρίσκονται σε δύο διαφορετικές ομάδες. Ο γενικός συσσωρευτικός αλγόριθμος και πάλι ξεκινά βρίσκοντας το μικρότερο στοιχείο του πίνακα $D=\{d_{ik}\}$ και συνενώνει τα αντίστοιχα αντικείμενα, έστω τα U και V, έτσι ώστε να δημιουργηθεί η νέα κλάση (UV). Η απόσταση ανάμεσα στην κλάση (UV) και οποιαδήποτε άλλη από τις εναπομείνασες W υπολογίζεται από τον τύπο:

$$d_{(UV)W}=\max\{d_{UW}, d_{VW}\}$$

όπου d_{UW} και d_{VW} είναι οι αποστάσεις ανάμεσα στα πιο απομακρυσμένα μέλη των κλάσεων U και W και V και W αντίστοιχα. Αντίστοιχο με το φαινόμενο της αλυσίδας της απλής συνένωσης είναι το πρόβλημα που προκύπτει για την μέθοδο της πλήρους συνένωσης, στην περίπτωση ύπαρξης ενός ζεύγους στοιχείων που βρίσκονται αρκετά μακριά μεταξύ τους σε δύο διαφορετικές ομάδες. Αυτό το ζευγάρι είναι πολύ πιθανόν να αποτρέψει τον αλγόριθμο να συνενώσει αυτές τις ομάδες ενώ στην πραγματικότητα αυτές να περιέχουν όμοια στοιχεία. Οι ομάδες που δημιουργούνται με την μέθοδο της πλήρους συνένωσης είναι μεγάλες και συμπαγείς, παρατηρείται συχνά όμως το φαινόμενο μη διαχωρισμού κάποιων πολύ μικρών συμπαγών ομάδων.

Γ) **Μέθοδος της Μέσης Συνένωσης (*Average Linkage*)**

Η μέθοδος αυτή θεωρεί ως απόσταση ανάμεσα σε δύο κλάσεις το μέσο όρο των αποστάσεων ανάμεσα στα ζεύγη όλων των στοιχείων που δημιουργούνται από τις δύο ομάδες. Η μέθοδος χρησιμοποιεί είτε μέτρα απόστασης είτε μέτρα ομοιότητας ανάμεσα στα στοιχεία των δύο ομάδων, ενώ μπορεί να θεωρήσει ως στοιχεία και αυτή τα άτομα ή τις μεταβλητές. Αρχικά βρίσκουμε τη μικρότερη απόσταση στον πίνακα αποστάσεων $D=\{d_{ik}\}$, ή τη μεγαλύτερη ομοιότητα, και έστω ότι αυτή εμφανίζεται στις κλάσεις U και V. Τότε γίνεται η συνένωση των δύο αυτών κλάσεων (UV) και οι αποστάσεις ανάμεσα στην νέα κλάση και σε οποιαδήποτε από τις εναπομείνασες W υπολογίζεται από τον τύπο:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

όπου d_{ik} είναι η απόσταση ανάμεσα στο i αντικείμενο της κλάσης (UV) και το k αντικείμενο της κλάσης W, $N_{(UV)}$ και N_W είναι ο αριθμός των στοιχείων της κάθε μια από τις παραπάνω κλάσεις αντίστοιχα. Ο παραπάνω τύπος προσαρμόζεται κατάλληλα όταν εργαζόμαστε με κάποιο μέτρο ομοιότητας. Στη βιβλιογραφία η μέθοδος αυτή συναντάται και με το όνομα «*Group Average Method*» ή ως «*Unweighted Pair-Group Average Method-UPGMA*». Μια παραλλαγή της μεθόδου είναι η μέθοδος των σταθμισμένων μέσων (*Weighted Average Linkage Method*) όπου ως απόσταση ανάμεσα στις δύο ομάδες θεωρεί τον μέσο των αποστάσεων όλων των στοιχείων της μιας ομάδας με τα στοιχεία της άλλης ομάδας (*Gordon*, 1999).

Δ) Μέθοδος του Ward

Ο Ward στήριξε τη διαδικασία της ιεραρχικής ομαδοποίησης στην ελαχιστοποίηση της «έλλειψης πληροφορίας» από την συνένωση δύο ομάδων. Αυτή η μέθοδος θεωρεί ότι η προηγούμενη έλλειψη πληροφορίας δημιουργεί μια αύξηση στο άθροισμα τετραγώνων του σφάλματος, *ESS*, (*Κούτρας*, 2009). Αρχικά, για δοσμένη κλάση k, ορίζουμε ως ESS_k το άθροισμα των τετραγωνικών αποκλίσεων κάθε όρου της κλάσης από το μέσο αυτής. Αν υπάρχουν K το πλήθος κλάσεις το συνολικό άθροισμα θα ισούται με $ESS=ESS_1+ESS_2+\ldots+ESS_K$. Σε κάθε στάδιο του αλγόριθμου ελέγχεται η ένωση κάθε δυνατού ζεύγους κλάσεων και επιλέγεται εκείνος ο συνδυασμός που οδηγεί στην μικρότερη αύξηση του *ESS*. Στην αρχή μπορεί να θεωρηθεί ότι κάθε άτομο αποτελεί και μια κλάση και συνεπώς αν υπάρχουν N άτομα, θα έχουμε $ESS_k=0$, $k=1,2,\ldots,N$ και $ESS=0$. Στο τελικό βήμα όπου όλα τα άτομα έχουν ομαδοποιηθεί σε μία και μόνο κλάση, που πλέον θα περιέχει N το πλήθος αντικείμενα, η τιμή του συνολικού αθροίσματος του σφάλματος υπολογίζεται από τον τύπο:

$$ESS = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$$

όπου x_j είναι η πολυδιάστατη μέτρηση που αντιστοιχεί στο j αντικείμενο και $\bar{\mathbf{x}}$ είναι ο συνολικός μέσος. Ουσιαστικά δηλαδή η μέθοδος αυτή έχει σχεδιαστεί έτσι ώστε να ελαχιστοποιεί τη διακύμανση μέσα στις ομάδες. Να σημειώσουμε ότι η μέθοδος αυτή χρησιμοποιείται πολύ συχνά επειδή έχει μια βάσιμη στατιστική λογική ενώ συγχρόνως δημιουργεί και ομάδες με παρόμοιο αριθμό παρατηρήσεων.

Εκτός από τις παραπάνω μεθόδους υπάρχουν και άλλες συσσωρευτικές μέθοδοι που δεν παρουσιάστηκαν οι οποίες όμως ακολουθούν τον ίδιο βασικό αλγόριθμο. Ένα από τα μειονεκτήματα όλων των ιεραρχικών μεθόδων είναι ότι δεν λαμβάνουν υπόψην τους τα σφάλματα και τις διακυμάνσεις που παρουσιάζονται στα δεδομένα με αποτέλεσμα να είναι ευαίσθητες σε ακραίες παρατηρήσεις και σε «σημεία θορύβου». Επίσης οι ομάδες που δημιουργούνται σε αρχικά βήματα δεν μπορούν να χωρίσουν και συνεπώς παρατηρήσεις που ενώνονται αρχικά μένουν μαζί μέχρι το πέρας του αλγορίθμου. Να σημειώσουμε ότι από άποψη υπολογιστικού φόρτου δεν είναι ιδιαίτερα πρακτικές κυρίως όταν το πλήθος των ατόμων/αντικειμένων που μελετάμε είναι μεγάλο. Το γεγονός αυτό οφείλεται στο ότι κανείς πρέπει να αποθηκεύσει στην μνήμη του υπολογιστή ολόκληρο τον πίνακα αποστάσεων και έτσι για N το πλήθος παρατηρήσεις προκύπτουν $N(N-1)/2$ αποστάσεις.

Ε) Μέθοδος της διαχωριστικής ανάλυσης (Divisive Analysis - DIANA)

Οι διαχωριστικές μέθοδοι ανήκουν και αυτές στην κατηγορία των ιεραρχικών μεθόδων. Η κύρια διαφορά τους με τις συσσωρευτικές μεθόδους, όπως αναφέρθηκε και προηγουμένως, είναι ότι αντί σε κάθε βήμα να συνενώνουν στοιχεία κάνουν την αντίθετη ακριβώς διαδικασία δηλαδή τα διαχωρίζουν. Σε κάθε βήμα μια διαχωριστική μέθοδος διασπά μία κλάση σε άλλες δύο μικρότερες μέχρι τελικά κάθε κλάση να περιλαμβάνει μόνο ένα στοιχείο. Αυτό συνεπάγεται αν το πλήθος των αντικειμένων μας είναι n, ο αλγόριθμος που εφαρμόζεται θα περιλαμβάνει και πάλι n-1 βήματα.

Στη βιβλιογραφία οι διαχωριστικές ιεραρχικές μέθοδοι έχουν συστηματικά αγνοηθεί με αποτέλεσμα οι ιεραρχικές μέθοδοι να θεωρούνται ταυτόσημες με τις συσσωρευτικές ιεραρχικές μεθόδους. Ο κύριος λόγος για τον οποίο έχει συμβεί αυτό είναι ότι μια συσσωρευτική μέθοδος απαιτεί πολύ λιγότερο υπολογιστικό φόρτο αφού στο πρώτο βήμα, όπου εξετάζονται όλες οι πιθανές συγχωνεύσεις των n αντικειμένων, εκτελούνται $n(n-1)/2$ πράξεις αφού τόσοι είναι οι συνδυασμοί των n αντικειμένων ανά δύο. Αντιθέτως, ένας διαχωριστικός αλγόριθμός στο πρώτο

βήμα, όπου θεωρεί όλες τις πιθανές διχοτομήσεις των *n* αντικειμένων, εκτελεί 2^{n-1}-1 υπολογισμούς. Το νούμερο αυτό ακόμα και για μικρά σύνολα δεδομένων θεωρείται ιδιαίτερα μεγάλο.

Ωστόσο για να μειωθεί ο υπολογιστικός φόρτος έχουν δημιουργηθεί μέθοδοι οι οποίες δεν απαιτούν την εξέταση όλων των πιθανών διαμερισμών. Μια τέτοια τεχνική περιγράφεται στον επόμενο αλγόριθμο (*Kaufman, Rousseeuw,* 1990):

1. Αρχικά βρίσκουμε το αντικείμενο που έχει τη μεγαλύτερη μέση ανομοιότητα με τα άλλα αντικείμενα. Αυτό το αντικείμενο αποτελεί από μόνο του μία κλάση, ένα είδος ομάδας αποστατών (*a splinter group).*
2. Για κάθε αντικείμενο *i* που δεν ανήκει στο *splinter group* υπολογίζεται η απόσταση D_i η οποία είναι ίση με τη διαφορά της απόσταση του αντικειμένου *i* που δεν ανήκει στο *splinter group* με κάθε άλλο αντικείμενο που επίσης δεν ανήκει στο *splinter group* και την απόσταση του αντικειμένου *i* από κάθε άλλο αντικείμενο που ανήκει στο *splinter group*. Με τύπους έχουμε την ακόλουθη σχέση:

$$D_i = [average\ d(i,j), j \notin R_{splinter\ group}] - [average\ d(i,j), j \in R_{splinter\ group}].$$

3. Στη συνέχεια βρίσκουμε το αντικείμενο *h* το οποίο έχει τη μεγαλύτερη από την παραπάνω διαφορά D_h. Αν η διαφορά D_h είναι θετική τότε το αντικείμενο *h* είναι κοντά στο *splinter group* και συνενώνεται μαζί του.
4. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρι όλες οι διαφορές D_h που προκύπτουν να είναι αρνητικές. Το σύνολο των δεδομένων τότε διαχωρίζεται σε δύο νέες κλάσεις.
5. Επιλέγουμε την κλάση με τη μεγαλύτερη διάμετρο και έπειτα η κλάση αυτή διαχωρίζεται όπως περιγράφεται στα προηγούμενα βήματα (η διάμετρος μιας κλάσης ορίζεται ως η μεγαλύτερη απόσταση ανάμεσα σε δύο αντικείμενά της).
6. Ο αλγόριθμος εκτελείται μέχρι να διαχωριστούν όλα τα αντικείμενα.

Για κάθε αντικείμενο *i* ορίζουμε ως d_i τη διάμετρο της τελευταίας κλάσης στην οποία ανήκει το αντικείμενο πριν αποσπαστεί και αποτελέσει από μόνο του μια κλάση, διαιρούμενο από τη διάμετρο της αρχικής κλάσης στην οποία ανήκουν όλα τα αντικείμενα. Με βάση αυτή τη διάμετρο ορίζεται ο διαχωριστικός συντελεστής (***Divisive Coefficient-DC***):

$$DC = \sum d_i / n$$

ο οποίος δείχνει πόσο καλός είναι ο διαχωρισμός που προέκυψε από τον αλγόριθμο. Τιμές κοντά στο 0 συνεπάγονται ότι ο διαχωρισμός δεν είναι πολύ καλός ενώ τιμές κοντά στο 1 δηλώνουν το αντίθετο.

Για τη μελέτη ενός συγκεκριμένου προβλήματος είναι καλό να χρησιμοποιούνται διάφορες μέθοδοι ομαδοποίησης και κάθε μέθοδος να εκτελείται τόσο με βάση τον πίνακα αποστάσεων όσο και με τον πίνακα ομοιοτήτων. Εάν τα αποτελέσματα διαφόρων μεθόδων είναι κατά προσέγγιση τα ίδια τότε καταλήγουμε στο συμπέρασμα ότι όντως τα δεδομένα δημιουργούν κλάσεις που είναι συνεπείς με τη φύση των δεδομένων. Η σταθερότητα του αποτελέσματος μιας μεθόδου μπορεί μερικές φορές να ελεγχθεί αν εφαρμόσουμε τον αλγόριθμο σε δεδομένα που περιέχουν μικρές αποκλίσεις από τις αρχικές μας παρατηρήσεις. Αν τα δεδομένα μας όντως δημιουργούν καλά ορισμένες ομαδοποιήσεις, τότε οι κλάσεις που δημιουργούνται πριν και μετά την αλλοίωση των δεδομένων, θα πρέπει να συμφωνούν.

2.4 Μη Ιεραρχικές Μέθοδοι Ομαδοποίησης

Οι μη ιεραρχικές μέθοδοι ομαδοποίησης έχουν σχεδιαστεί έτσι ώστε να δημιουργούν ομάδες/κλάσεις στα αντικείμενα ή στα άτομα παρά στις μεταβλητές. Ο αριθμός των κλάσεων, *k*, μπορεί να καθοριστεί από την αρχή ή να αποφασιστεί κατά τη διάρκεια διεξαγωγής του αλγορίθμου. Οι μη ιεραρχικές μέθοδοι εφαρμόζονται σε μεγαλύτερα σύνολα δεδομένων από ότι οι ιεραρχικές μέθοδοι αφού δεν απαιτούν την αποθήκευση του πίνακα αποστάσεων ή ομοιοτήτων.

Ο τρόπος λειτουργίας των περισσότερων μεθόδων είναι ένας από τους παρακάτω δύο:

i) Θεωρούμε *k* συγκεκριμένα άτομα-μητρικά σημεία (*seed points*) και γύρω από αυτά ταξινομούνται τα υπόλοιπα στοιχεία μέχρι να δημιουργηθούν οι επιθυμητές ομάδες ή

ii) Ο αλγόριθμος ξεκινά με έναν αρχικό διαχωρισμό (*initial partition*) των ατόμων σε *k* ομάδες και στη συνέχεια μετακινούνται τα άτομα μεταξύ των ομάδων μέχρι να επιτευχθεί ο καλύτερος διαμερισμός.

Α) Μητρικά Σημεία

Πριν γίνει η παρουσίαση γνωστών μη ιεραρχικών μεθόδων θα ήταν χρήσιμο να δούμε τους πιο δημοφιλείς τρόπους κατασκευής μητρικών σημείων.

- Επιλογή των *k* πρώτων στη σειρά ατόμων από τα δεδομένα.
- Αρίθμηση των ατόμων από 1 ως *n* και στη συνέχεια επιλέγονται αυτά με την αρίθμηση *n/k*, 2*n/k*, …, (*k*-1)*n/k* και *n*.
- Αρίθμηση των ατόμων από 1 ως *n*, δημιουργία *k* διαφορετικών τυχαίων αριθμών από το 1 ως το *n* και στη συνέχεια επιλογή των ατόμων που αντιστοιχούν σε αυτούς τους αριθμούς (*McRae*, 1971).
- Λογικός διαχωρισμός των ατόμων σε *k* ομάδες και στη συνέχεια επιλογή των κέντρων βάρους των ομάδων αυτών ως μητρικά σημεία (*Forgy*, 1965).
- Ο παρακάτω τρόπος προτάθηκε από τον (*Astrahan*, 1970) και προσπαθεί να επιλέξει σημεία που εκτείνονται σε όλο το μήκος των δεδομένων και συγχρόνως το κάθε μητρικό σημείο να είναι καλώς διαχωρισμένο από τα άλλα.

  - Υπολογίζουμε την πυκνότητα (*density*) για κάθε ένα αντικείμενο η οποία ορίζεται ως το πλήθος των ατόμων που βρίσκονται γύρω από αυτό σε μια καθορισμένη ακτίνα μήκους d_1.
  - Τοποθετούμε σε σειρά όλα τα σημεία με βάση την πυκνότητά τους και επιλέγουμε εκείνο με την μεγαλύτερη πυκνότητα ως το πρώτο μητρικό σημείο.
  - Στη συνέχεια επιλέγουμε διαδοχικά μητρικά σημεία με σειρά έτσι ώστε να μειώνεται η πυκνότητα και ταυτόχρονα το κάθε νέο μητρικό σημείο να απέχει τουλάχιστον μια ελάχιστη απόσταση d_2, όπου $d_2 > d_1$ από όλα τα προηγούμενα σημεία. Η διαδικασία συνεχίζεται εώς ότου όλα τα εναπομείναντα σημεία να έχουν πυκνότητα ίση με το 0 (δηλαδή απόσταση τουλάχιστον ίση με d_1 από κάθε άλλο στοιχείο).

➢ Σε περίπτωση που εφαρμόζοντας αυτήν τη διαδικασία, προκύψουν περισσότερα σημεία από όσα χρειαζόμαστε, τότε γίνεται μια ιεραρχική ομαδοποίηση των μητρικών σημείων έτσι ώστε να έχουμε ακριβώς *k*.

Να σημειώσουμε ότι, στον αλγόριθμο του *Astrahan* απαιτείται σωστή κρίση ή πολλές τυχαίες επιλογές για τον κατάλληλο προσδιορισμό των αποστάσεων d_1 και d_2. Αν το d_1 είναι πάρα πολύ μικρό τότε είναι πιθανό να υπάρχουν πολλά απομονωμένα σημεία με μηδενική πυκνότητα ενώ αν το d_1 είναι πολύ μεγάλο τότε λίγα μητρικά σημεία θα καλύψουν ολόκληρο το σύνολο των δεδομένων.

Οι Ball και Hall (1967) πρότειναν μια πιο εύκολη διαδικασία από την προηγούμενη. Αρχικά επιλέγονται τα *k* πρώτα άτομα-παρατηρήσεις ως μητρικά σημεία ενώ για κάθε παρατήρηση υπολογίζεται η απόστασή της από τα υπάρχοντα κέντρα καθώς και οι αποστάσεις όλων των κέντρων ανά δύο. Αν η μικρότερη από τις αποστάσεις της κάθε παρατήρησης από τα ήδη υπάρχοντα κέντρα είναι μεγαλύτερη από την απόσταση των δύο πιο κοντινών υπαρχόντων κέντρων τότε το κέντρο που βρίσκεται πιο κοντά στην παρατήρηση αυτή αντικαθιστάται από την παρατήρηση. Το ίδιο συμβαίνει και αν η απόσταση της παρατήρησης είναι μεγαλύτερη από τη μικρότερη απόσταση μεταξύ του συγκεκριμένου κέντρου και των υπολοίπων. Συνεπώς, όταν η παραπάνω διαδικασία πραγματοποιηθεί για όλες τις παρατηρήσεις, θα καταλήξουμε με κέντρα τα οποία θα αποτελούν τα αρχικά κέντρα. Να σημειώσουμε ότι η παραπάνω τεχνική χρησιμοποιείται από το στατιστικό πακέτο SPSS για τη δημιουργία αρχικών μητρικών σημείων στους μη ιεραρχικούς αλγόριθμους.

Β) Αρχικοί Διαμερισμοί

Σε κάποιες μη ιεραρχικές μεθόδους είναι προτιμότερος ένας αρχικός διαμερισμός των ατόμων σε *k* ομάδες από ότι ο υπολογισμός των μητρικών σημείων. Μέθοδοι που παράγουν τέτοιους αρχικούς διαμερισμούς είναι οι ακόλουθες:

- Για ένα δεδομένο σύνολο από μητρικά σημεία τοποθετούμε κάθε αντικείμενο στην ομάδα που βρίσκεται πιο κοντά σε αυτό (*Forgy*, 1965). Τα μητρικά σημεία παραμένουν στατικά μέχρι να τοποθετηθούν όλα τα άτομα σε μια ομάδα, συνεπώς το αποτέλεσμα δεν εξαρτάται από τη σειρά με την οποία επιλέγονται τα άτομα.

- Για ένα δεδομένο σύνολο μητρικών σημείων θεωρούμε ότι κάθε μητρικό σημείο αποτελεί και μια κλάση. Στη συνέχεια τοποθετούμε τα υπόλοιπα στοιχεία στην κλάση που βρίσκεται πιο κοντά και ανανεώνουμε το κέντρο βάρος της έτσι ώστε ο νέος μέσος να εκφράζει και τα νέα άτομα που περιλαμβάνονται πλέον στην κλάση (*MacQueen,* 1967).
- Στις περισσότερες περιπτώσεις η εφαρμογή ενός ιεραρχικού αλγορίθμου μπορεί να παράγει έναν ιδανικό αρχικό διαχωρισμό. Ο Wolfe (1970) χρησιμοποιεί τη μέθοδο ιεραρχικής ομαδοποίησης του Ward για την παροχή ενός αρχικού συνόλου με κλάσεις ενώ οι Lance και Williams (1976b) προτείνουν τη χρήση μιας ιεραρχικής μεθόδου σε ένα υποσύνολο των δεδομένων, με κατάλληλο μέγεθος, και στη συνέχεια χρησιμοποιούν τα αποτελέσματα σαν πυρήνα για τις υπόλοιπες κλάσεις.
- Οι ίδιοι οι αναλυτές πολλές φορές δημιουργούν εμπειρικά έναν αρχικό διαμερισμό των δεδομένων σε *k* ομάδες.

2.5 Αλγόριθμοι Υλοποίησης Μη Ιεραρχικών Μεθόδων

Οι παρακάτω μέθοδοι δουλεύουν επαναληπτικά χρησιμοποιώντας την έννοια του κέντρου βάρους μιας ομάδας (*centroid*) το οποίο ορίζεται ως η μέση τιμή κάθε μεταβλητής για όλα τα στοιχεία της ομάδας. Οι παρατηρήσεις τοποθετούνται σε ομάδες ανάλογα με την απόστασή τους από τα κέντρα των ομάδων. Αρχικά γίνεται ο υπολογισμός της απόστασης της κάθε παρατήρησης από το κέντρο της ομάδας και στη συνέχεια τοποθετείται σε εκείνη που βρίσκεται πιο κοντά. Οι διαφοροποιήσεις των μεθόδων προκύπτουν από τον τρόπο με τον οποίο γίνεται ανανέωση των κέντρων των ομάδων και η ταξινόμηση όσων παρατηρήσεων απομένουν σε αυτές.

Α) Μέθοδος του Forgy

Ο αλγόριθμος αποτελείται από τα παρακάτω βήματα:

1. Καθορίζουμε έναν αρχικό διαμερισμό των *n* ατόμων σε *k* ομάδες ή ένα σύνολο από *k* μητρικά σημεία. Στην τελευταία περίπτωση θεωρούμε ότι κάθε μητρικό σημείο αποτελεί μια ομάδα με ένα στοιχείο ενώ στην πρώτη υπολογίζουμε τα κέντρα βάρους όλων των ομάδων και θεωρούμε αυτά ως μητρικά σημεία-ομάδες.

2. Κατατάσσουμε κάθε παρατήρηση στην ομάδα εκείνη που απέχει τη μικρότερη απόσταση. Να σημειώσουμε ότι τα μητρικά σημεία παραμένουν αναλλοίωτα κατά τη διεξαγωγή ενός κύκλου του αλγορίθμου σε όλη την έκταση των δεδομένων.
3. Μετά την ολοκλήρωση ενός κύκλου σε όλα τα δεδομένα, υπολογίζουμε τα κέντρα βάρους των ομάδων που δημιουργήθηκαν και θεωρούμε αυτά ως τα νέα μητρικά σημεία.
4. Τα Βήματα 2 και 3 επαναλαμβάνονται μέχρι τη σύγκλιση του αλγορίθμου η οποία επιτυγχάνεται όταν τα νέα κέντρα δε διαφέρουν από τα παλιά.

Δεν είναι δυνατόν να γνωρίζουμε από την αρχή σε πόσες επαναλήψεις συγκλίνει ο αλγόριθμος αλλά από εμπειρικές μελέτες έχει διαπιστωθεί ότι 5 επαναλήψεις είναι συνήθως αρκετές. Σε κάθε επανάληψη του αλγορίθμου η διαδικασία ομαδοποίησης των *n* στοιχείων σε *k* ομάδες απαιτεί *nk* υπολογισμούς αποστάσεων και *n*(*k-1*) συγκρίσεις των αποστάσεων. Αφού το *k* είναι πάντα μικρότερο του *n* και ο αριθμός των επαναλήψεων είναι και αυτός μικρός κανείς μπορεί να εξετάσει πολλά σύνολα ομάδων με διαφορετικές τιμές για το *k*, για να παγιώσει τα αποτελέσματά του και συγχρόνως να κερδίσει σε χρόνο και κόστος σε σχέση με το να υλοποιούσε κάποια ιεραρχική μέθοδο.

Β) Μέθοδος MacQueen ή k-Means Method

Ο MacQueen χρησιμοποιεί τον όρο *k-means* για να περιγράψει τη διαδικασία τοποθέτησης κάθε αντικειμένου σε εκείνη την κλάση, από τις συνολικά *k*, που είναι πιο κοντά στο κέντρο βάρους της. Το κλειδί σε αυτήν τη μέθοδο είναι ότι το κέντρο βάρους των ομάδων υπολογίζεται μετά από κάθε ταξινόμηση ενός ατόμου σε μια κλάση και δεν απαιτείται η ολοκλήρωση της τοποθέτησης όλων των ατόμων σε μια κλάση όπως προϋποθέτει η μέθοδος του Forgy. Τα βήματα αυτού του αλγορίθμου για την τοποθέτηση των *n* αντικειμένων σε *k* κλάσεις είναι τα ακόλουθα:

1. Καθορίζουμε αρχικά ένα σύνολο από *k* μητρικά σημεία χρησιμοποιώντας *k* από τα *n* άτομα που είναι διαθέσιμα.
2. Τα εναπομείναντα *n-k* άτομα τα κατατάσσουμε σε εκείνη την ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από το άτομο. Έπειτα από κάθε τοποθέτηση γίνεται ο υπολογισμός του κέντρου βάρους της νέο δημιουργηθείσας ομάδας.

3. Μετά το πέρας του Βήματος 2 όλα τα άτομα θα έχουν τοποθετηθεί σε ομάδες και συνεπώς ως μητρικά σημεία θα θεωρηθούν τα δημιουργηθέντα κέντρα βάρους. Ο αλγόριθμος πραγματοποιεί μια ακόμη σάρωση κατά την οποία τοποθετούνται όλα τα άτομα στο πιο κοντινό μητρικό σημείο.

Η όλη διαδικασία απαιτεί $k(2n-k)$ υπολογισμούς αποστάσεων, $(k-1)(2n-k)$ συγκρίσεις αποστάσεων και $n-k$ ανανεώσεις κέντρων βάρους. Από υπολογιστική άποψη ο αλγόριθμος είναι πολύ πιο γρήγορος από οποιαδήποτε ιεραρχική μέθοδο αφού σε πολύ λίγες επαναλήψεις επέρχεται ο τερματισμός του. Επίσης η τελική ομαδοποίηση αποτελείται από ομάδες με περίπου ίσο αριθμό παρατηρήσεων.

Το μειονέκτημα του αλγορίθμου έγκειται στο γεγονός ότι επηρεάζεται από τα αρχικά επιλεγμένα μητρικά σημεία ή τις αρχικές διαμερίσεις τα οποία αν δεν είναι ορθά επιλεγμένα μπορεί να δημιουργήσουν μια εντελώς διαφορετική ομαδοποίηση από αυτή που υπάρχει στα δεδομένα. Για να αντιμετωπιστεί επιτυχώς αυτό το πρόβλημα θα μπορούσαμε να τρέχουμε την μέθοδο με διάφορες επιλογές έτσι ώστε να είμαστε σίγουροι πως ο αλγόριθμος δεν εγκλωβίζεται σε κάποια μη βέλτιστη λύση. Ένα ακόμη μειονέκτημα της μεθόδου είναι ότι η ύπαρξη έκτροπων παρατηρήσεων μπορεί να προκαλέσει τη δημιουργία ομάδων με πολύ απομακρυσμένα άτομα. Επίσης, αν ο πληθυσμός μας αποτελείται από k ομάδες μπορεί κάποια από αυτές να μην υπάρχει στο δείγμα μας και έτσι να οδηγηθούμε σε μια παραπλανητική ομαδοποίηση.

Να σημειώσουμε ότι υπάρχει στη βιβλιογραφία μια πληθώρα παραλλαγών του παραπάνω αλγορίθμου οι οποίες προσπαθούν να εξαλείψουν τα μειονεκτήματα του. Χαρακτηριστικά αναφέρουμε την παραλλαγή του Wishart, την παραλλαγή των σταθερών ομάδων που πρότεινε ο MacQueen καθώς και μια ακόμη παραλλαγή του ιδίου που επιτρέπει στον αριθμό των ομάδων να είναι μεταβαλλόμενος κατά την αρχική τοποθέτηση των δεδομένων στις ομάδες (*Andeberg*, 1973).

Γ) Μέθοδος Partition Around Medoids (PAM)

Μια ακόμη μέθοδος που βασίζεται στη μέθοδο *k-means* είναι η μέθοδος PAM. Η μέθοδος αυτή χρησιμοποιεί τον πίνακα αποστάσεων/ανομοιοτήτων ενός δοσμένου συνόλου δεδομένων και σε περίπτωση που αυτός δε δίνεται εξαρχής αλλά δίνεται μόνο ο nxp πίνακας με τα πλήρη δεδομένα από τις πρώτες διαδικασίες που εκτελούνται είναι ο υπολογισμός του *πίνακα*

αποστάσεων[1]. Σε σχέση με τη μέθοδο *k-means* είναι καλύτερη διότι μπορεί να ελαχιστοποιεί το άθροισμα των διαφορών ενός συνόλου και δεν αντιμετωπίζει κάποιο πρόβλημα όταν δε δίνεται ο αρχικός πίνακας δεδομένων. Ακόμη τα αντιπροσωπευτικά στοιχεία που χρησιμοποιεί ανήκουν στα δεδομένα σε αντίθεση με τα κέντρα βάρους (*centroids*) που χρησιμοποιούνται στη μέθοδο *kmeans* και δεν είναι απαραίτητο να είναι στοιχεία του αρχικού συνόλου δεδομένων (*Kaufman, Rousseeuw*, 1990). Επίσης σα μέθοδος παρέχει και ένα γράφημα, το ***silhouette plot***, το οποίο επιτρέπει στον αναλυτή να επιλέξει από μόνος του, και όχι από την αρχή, το βέλτιστο αριθμό κλάσεων.

Ο αλγόριθμος που χρησιμοποιείται χωρίζεται σε δύο βασικά στάδια. Το 1ο καλείται BUILD και περιλαμβάνει τη δημιουργία των *k* πρωταρχικών αντιπροσωπευτικών αντικειμένων τα οποία θα δώσουν και μια αρχική ομαδοποίηση των δεδομένων σε *k* συστάδες. Το 1ο αντικείμενο επιλέγεται έτσι ώστε το άθροισμα όλων των ανομοιοτήτων με τα άλλα αντικείμενα να είναι το μικρότερο δυνατό. Το αντικείμενο αυτό είναι το πιο κεντρικά τοποθετημένο στο σύνολο των δεδομένων μας και είναι αυτό που ελαχιστοποιεί περισσότερο από όλα τα άλλα την αντικειμενική συνάρτηση (βλ. ορισμό αντικειμενικής συνάρτησης παρακάτω). Στη συνέχεια επιλέγονται τα υπόλοιπα *k*-1 αντικείμενα. Τα αντικείμενα αυτά καλούνται ***medoids***. Με τον όρο αυτό καλούμε ένα αντικείμενο μιας κλάσης του οποίου η μέση ανομοιότητα/απόσταση από όλα τα άλλα αντικείμενα της ίδιας κλάσης είναι η μικρότερη. Για να βρεθεί το 1ο medoid εκτελούνται τα ακόλουθα βήματα:

1. Επιλέγουμε τυχαία ένα αντικείμενο *i*.
2. Επιλέγουμε ένα αντικείμενο *j* που δεν έχει επιλεχθεί ξανά και υπολογίζουμε τη διαφορά ανάμεσα στην ανομοιότητα D_j του αντικειμένου *j* και του πιο όμοιου αντικειμένου (με το *j*) που έχει επιλεχθεί προηγουμένως αντί του *j* και την ανομοιότητα $d(i,j)$.
3. Αν η διαφορά του βήματος 2 είναι θετική τότε το αντικείμενο *i* είναι υποψήφιο για τη θέση του 1ου Medoid. Δηλαδή $C_{ji}=\max(D_j - d(i,j),0)$.
4. Το αντικείμενο που τελικώς θα επιλεχθεί είναι αυτό το οποίο μεγιστοποιεί την ποσότητα:

$$\max_i \sum_j C_{ji} \ .$$

[1] Ο πίνακας ανομοιοτήτων/αποστάσεων ή αλλιώς *dissimilarity/distance matrix* είναι ένας τετραγωνικός συμμετρικός πίνακας όπου το στοιχείο του *ij* είναι ίσο με την τιμή του προεπιλεγμένου μέτρου ομοιότητας ή απόστασης ανάμεσα στα στοιχεία *i* και *j*.

Η παραπάνω διαδικασία εκτελείται ως ότου συγκεντρωθούν τα *k* αρχικά αντικείμενα. Στη συνέχεια κάθε αντικείμενο τοποθετείται στην κλάση που βρίσκεται το πιο κοντινό medoid. Δηλαδή το αντικείμενο *l* θα τοποθετηθεί στην κλάση v_l αν το medoid m_{v_l} είναι πιο κοντά από οποιοδήποτε άλλο medoid m_w. Η αντίστοιχη μαθηματική σχέση είναι:

$$d(l, m_{v_l}) \leq d(l, m_w) \text{ για κάθε } w=1,\ldots, k.$$

Τα *k* αντιπροσωπευτικά αντικείμενα έχουν ως στόχο την ελαχιστοποίηση της αντικειμενικής συνάρτησης. Ως αντικειμενική συνάρτηση ορίζεται το άθροισμα των ανομοιοτήτων όλων των αντικειμένων με το κοντινότερο medoid και είναι ίση με: $\sum d\left(i, m_{v_i}\right).$

Στη δεύτερη φάση του αλγορίθμου, η οποία καλείται φάση SWAP, γίνεται μια προσπάθεια βελτίωσης των ήδη επιλεγμένων σημείων και κατά συνέπεια βελτίωσης της ομαδοποίησης. Αν η αντικειμενική συνάρτηση μπορεί να ελαχιστοποιηθεί περαιτέρω με την ανταλλαγή ενός επιλεγμένου αντικειμένου με ένα μη επιλεγμένο, αυτό πραγματοποιείται.

Δ) Μέθοδος Clustering Large Applications (CLARA)

Μια παραλλαγή της μεθόδου PAM είναι η μέθοδος CLARA η οποία εφαρμόζεται σε μεγαλύτερα σύνολα δεδομένων. Η μέθοδος αυτή προσπαθεί να εντοπίσει τα *k* αντιπροσωπευτικά αντικείμενα δουλεύοντας σε υποσύνολα των δεδομένων τα οποία έχουν προκαθορισμένο μέγεθος. Συνεπώς ο συνολικός χρόνος και ο υπολογιστικός φόρτος μειώνονται κατά πολύ αφού δεν υπολογίζεται κάθε φορά ολόκληρος ο πίνακας των ανομοιοτήτων.

Και αυτή η μέθοδος εκτελείται σε 2 βήματα το BUILD και το SWAP. Αρχικά ένα δείγμα από το σύνολο των δεδομένων επιλέγεται και διαμερίζεται σε *k* κλάσεις σύμφωνα με τον αλγόριθμο που περιγράφηκε στη μέθοδο PAM. Και εδώ δίνονται τα silhouettes plots αλλά μόνο για εκείνο το σύνολο δεδομένων που δίνει την καλύτερη ομαδοποίηση.

Ε) Μέθοδος Fuzzy Analysis (FANNY)

Η μέθοδος αυτή είναι μια ακόμα παραλλαγή της PAM. Ενώ στις μεθόδους PAM και CLARA το κάθε αντικείμενο τοποθετείται αυστηρά σε μια και μόνο κλάση, στη FANNY επιτρέπεται ο

διαχωρισμός των αντικειμένων να μην είναι τόσο σαφής (κάτι που συμβαίνει ιδιαίτερα συχνά σε πραγματικά σύνολα δεδομένων).

Η μέθοδος προσπαθεί να ελαχιστοποιήσει την εξής αντικειμενική συνάρτηση:

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2}$$

όπου *d(i,j)* αποτελεί τη δοσμένη απόσταση (ή μέτρο ανομοιότητας) ανάμεσα στα αντικείμενα *i* και *j*, u_{iv} είναι η άγνωστη συμμετοχή του αντικειμένου *i* στην κλάση *v*. Οι συναρτήσεις που δείχνουν τη συμμετοχή του κάθε αντικειμένου στις κλάσεις υπόκεινται στους εξής περιορισμούς:

- $u_{iv} \geq 0$ για όλα *τα* $i=1,\ldots,n$ και όλα τα $v=1,\ldots,k$
- $\sum_{v=1}^{k} u_{iv} = 1 = 100\%$ για όλα *τα* $i=1,\ldots,n$.

Οι συναρτήσεις που δείχνουν τη συμμετοχή παίρνουν μόνο μη αρνητικές τιμές και το άθροισμα αυτών των συναρτήσεων για κάθε ένα αντικείμενο ξεχωριστά ισούται με τη μονάδα. Όταν όλα τα αντικείμενα έχουν ίση συμμετοχή σε όλες τις κλάσεις λέμε ότι η ομαδοποίηση που προέκυψε είναι πλήρως ασαφής (*fuzzy*). Αντιθέτως όταν ένα αντικείμενο έχει συμμετοχή ίση με 1 σε μία κλάση και κατά συνέπεια 0 σε όλες τις υπόλοιπες η ομαδοποίηση απολύτως σαφής.

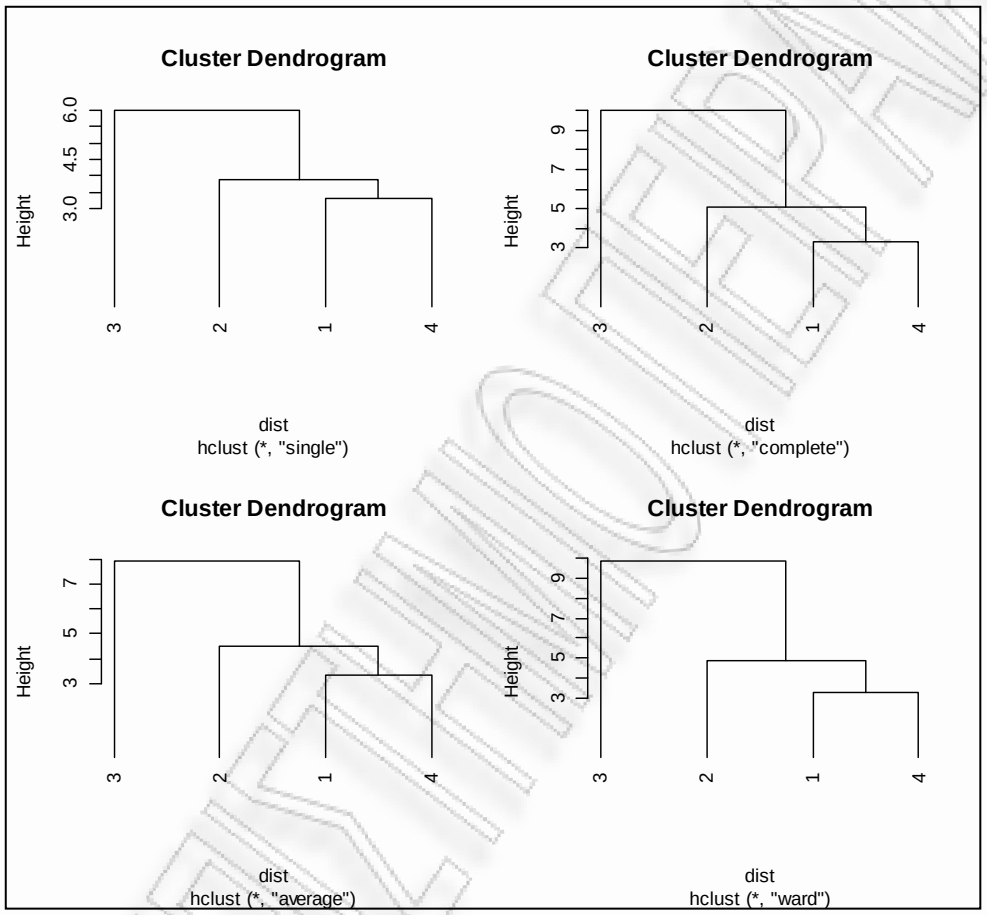
Παράδειγμα 2.5

Με βάση τα δεδομένα του Πίνακα 2.2 θα παρουσιαστεί ένα παράδειγμα των βασικών μεθόδων ομαδοποίησης με τη βοήθεια της R. Οι αποστάσεις των 4 φοιτητών με βάση την Ευκλείδεια Απόσταση είναι, όπως υπολογίστηκαν και προηγουμένως οι εξής:

Φοιτητής	**1**	**2**	**3**
2	5.099020	-	-
3	10	6	-
4	3.316625	3.872983	7.810250

Πίνακας 2.5: Πίνακας Αποστάσεων

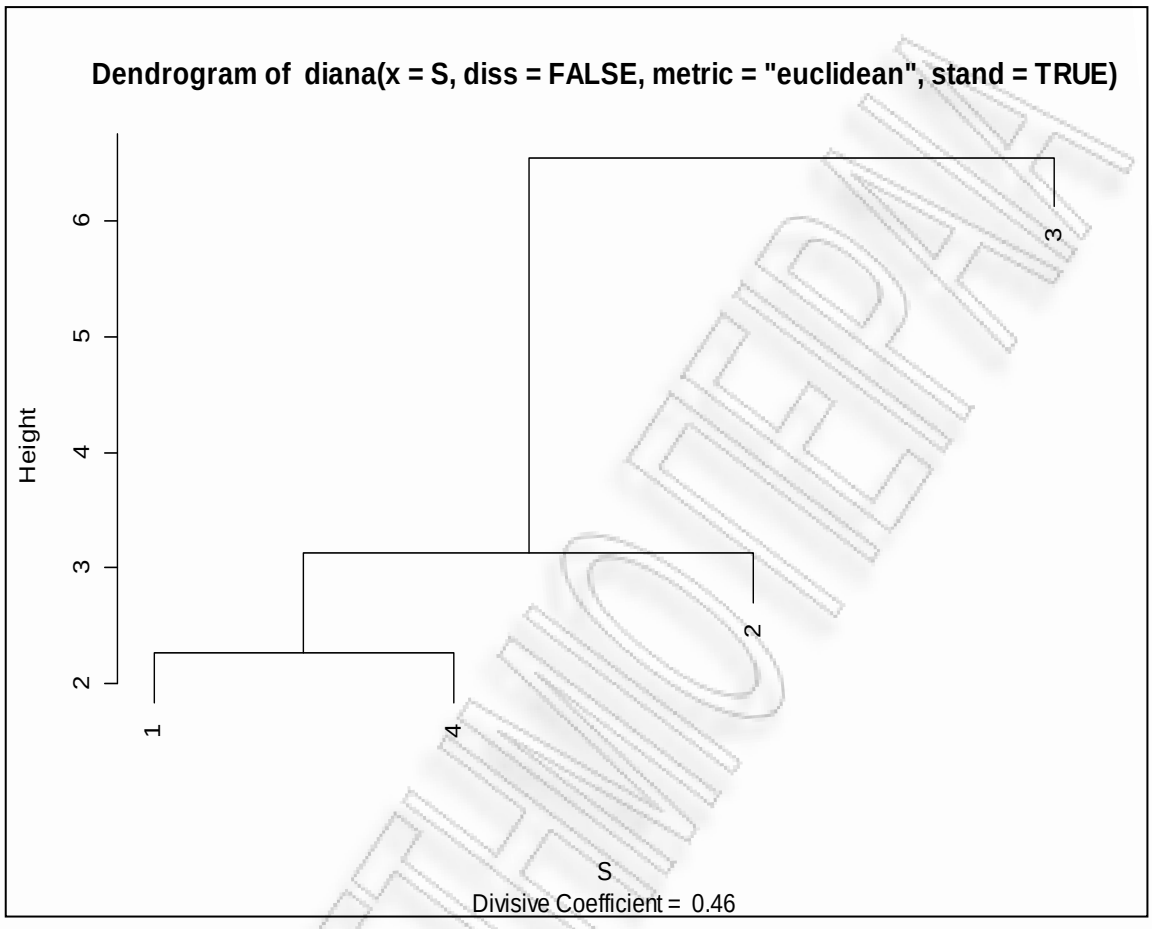
Εφαρμόζοντας τις ιεραρχικές μεθόδους της απλής συνένωσης, της πλήρους συνένωσης, της μέσης συνένωσης και της μεθόδου Ward προκύπτουν τα δενδρογράμματα που απεικονίζονται στο σχήμα 2.1:



Σχήμα 2.2: Δενδρογράμματα Ιεραρχικών Συσσωρευτικών Μεθόδων

Από τα παραπάνω δενδρογράμματα παρατηρούμε ότι στο 1ο στάδιο και σε όλες τις μεθόδους γίνεται η συνένωση του 1ου και του 4 φοιτητή. Στη συνέχεια ενώνεται ο 2ος φοιτητής και τελευταίος προσαρτάται στην ομάδα ο 3ος. Βέβαια οι συνενώσεις, ανάλογα με τη μέθοδο που χρησιμοποιείται, πραγματοποιούνται σε διαφορετικές αποστάσεις. Χαρακτηριστικά αναφέρουμε ότι με τη μέθοδο του κοντινότερου γείτονα ο φοιτητής 2 ενώνεται με τη δημιουργηθείσα κλάση των 1 και 4 σε απόσταση μικρότερη από 4 ενώ με τη μέθοδο του μακρινότερου γείτονα η συνένωση πραγματοποιείται σε απόσταση 5.

Εφαρμόζοντας τη διαχωριστική μέθοδο DIANA προκύπτει το δενδρόγραμμα του Σχήματος 2.2.



Σχήμα 2.3: Δενδρόγραμμα Ιεραρχικής Διαχωριστικής Μεθόδου

Και πάλι βλέπουμε ότι πρώτος αποσπάται ο 3ος φοιτητής, έπειτα ο 2ος ενώ τελευταίοι διαχωρίζονται οι φοιτητές 1 και 4 που είναι αυτοί οι οποίοι συνενώνονται πρώτοι στις αντίστοιχες συσωρευτικές μεθόδους. Ο διαχωριστικός συντελεστής ισούται με 0,46 που συνεπάγεται ότι ο διαχωρισμός που έχει πραγματοποιηθεί είναι αρκετά ικανοποιητικός.

Αξίζει να παρατηρήσουμε πως η συμφωνία και των 5 διαφορετικών ιεραρχικών μεθόδων είναι ένα καθαρά τυχαίο γεγονός και οφείλεται στη φύση των συγκεκριμένων δεδομένων. Οι μόνες αλλαγές που παρατηρούνται είναι στην απόσταση που γίνονται οι συγχωνεύσεις των ομάδων. Εκτελούμε την ομαδοποίηση για τα ίδια δεδομένα με βάση τη μέθοδο *k*-means και προεπιλέγουμε να δημιουργηθούν 2 κλάσεις. Τα αποτελέσματα μας οδηγούν στο εξής Clustering Vector (2 2 1 2) που συνεπάγεται ότι οι φοιτητές 1,2 και 4 αποτελούν μια ομάδα και ο φοιτητής 3 αποτελεί από μόνος του μια άλλη ομάδα.

Αν επανεκτελέσουμε την ανάλυση επιλέγοντας να δημιουργηθούν 3 αντί για 2 κλάσεις τότε θα έχουμε το Clustering Vector (3 2 1 3) που συνεπάγεται ότι τα άτομα 1 και 4 είναι αυτά που ενώνονται ενώ τα υπόλοιπα αποτελούν από μόνα τους ομάδα.

Στη συνέχεια εκτελούμε τη μέθοδο PAM επιλέγοντας και αυτή να μας δημιουργήσει 2 κλάσεις και χρησιμοποιώντας ως μέτρο απόστασης την Ευκλείδεια. Τα αποτελέσματα ταυτίζονται με τη μέθοδο *k*-means αφού ως Clustering Vector προκύπτει το (1 1 2 1). Και πάλι τα άτομα 1,2,4 αποτελούν ομάδα ενώ το 3 αποτελεί από μόνο του ομάδα. Αξίζει να παρατηρήσουμε ότι η R μας δίνει και την τιμή της αντικειμενικής συνάρτησης στα βήματα BUILD και SWAP του αλγορίθμου:

Objective function:

build swap

2.243001 1.797402

Όπως ήταν αναμενόμενο και από τη θεωρία, στο 2ο βήμα η αντικειμενική συνάρτηση έχει μικρότερη τιμή από ότι στο 1ο. Να σημειώσουμε ότι η εφαρμογή του αλγορίθμου FANNY δεν ήταν δυνατή λόγω του πολύ μικρού αριθμού δεδομένων.

Στο παράδειγμα που παρουσιάστηκε, όλες οι μέθοδοι έδωσαν το ίδιο αποτέλεσμα αυτό όμως δεν είναι το σύνηθες. Για παράδειγμα ας υποθέσουμε ότι οι βαθμολογίες για τους 4 φοιτητές ήταν οι εξής:

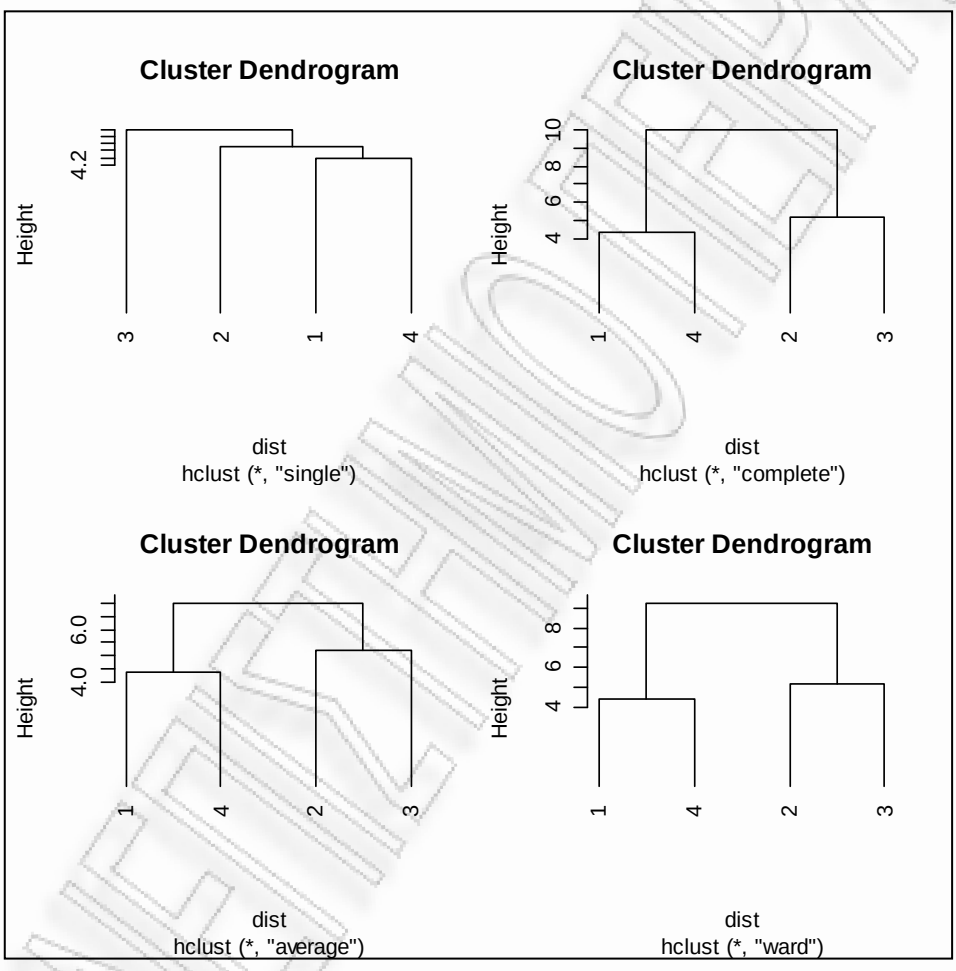
Φοιτητής	**Βαθμολογίες στο *i* μάθημα**				
	1	**2**	**3**	**4**	**5**
1	9	8	8	7	10
2	5	7	6	5	6
3	5	3	3	4	5
4	7	5	7	8	8

Πίνακας 2.6: Πίνακας Βαθμολογιών

Η Ευκλείδεια Απόσταση και τα δενδρογράμματα των ιεραρχικών συσσωρευτικών μεθόδων που προκύπτουν είναι τα εξής:

Φοιτητής	1	2	3
2	6.403124		
3	10	5.196152	
4	4.358899	4.690416	7

Πίνακας 2.6.1: Ευκλείδεια Απόσταση

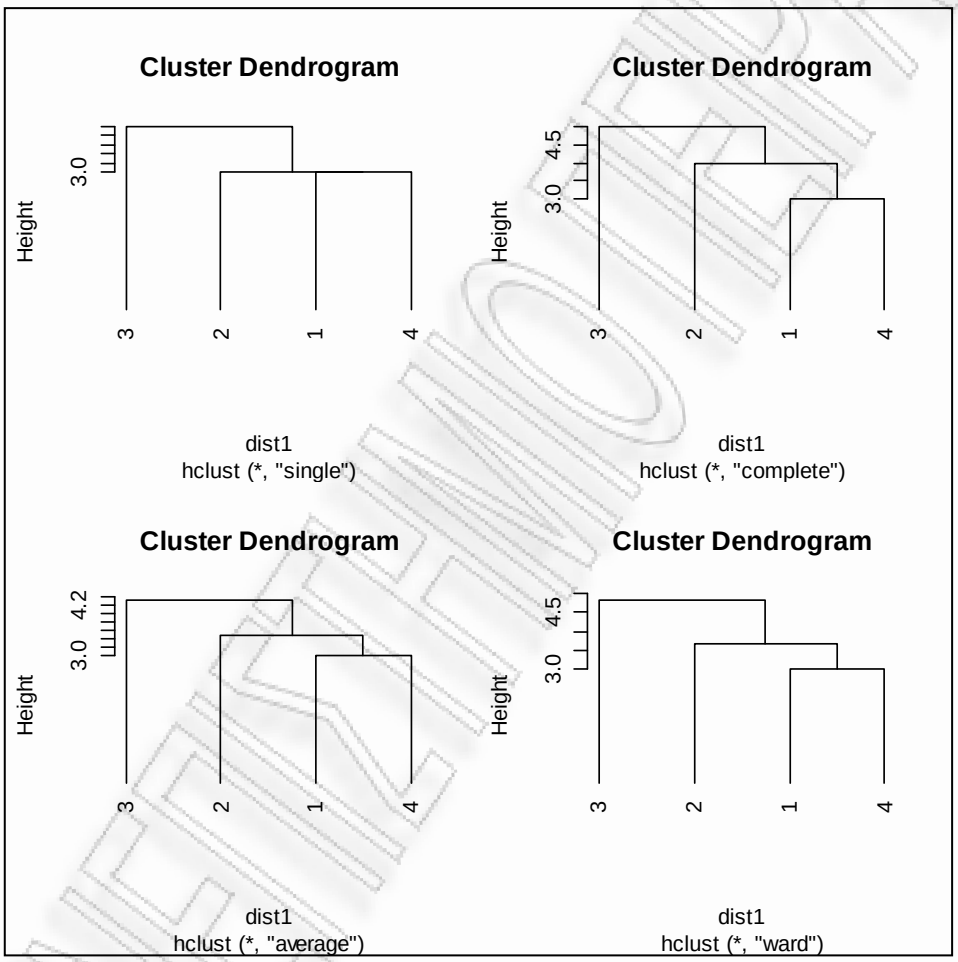


Σχήμα 2.4.1: Δενδρογράμματα Ιεραρχικών Συσσωρευτικών Μεθόδων (Ευκλείδεια Απόσταση)

Παρατηρούμε ότι με τη μέθοδο του κοντινότερου γείτονα αρχικά ενώνονται τα άτομα 1 και 4, σε μικρή απόσταση ενώνεται μαζί τους ο 2ος φοιτητής και ακολουθεί ο 3ος. Αντιθέτως με τις άλλες 3 μεθόδους είναι σαφές ότι τα δεδομένα μας χωρίζονται σε 2 κλάσεις οι οποίες αποτελούνται από τα άτομα {1,4} και {2,3}. Αξίζει να παρατηρήσουμε ότι αν χρησιμοποιήσουμε την απόσταση Chebychev τότε θα προκύψουν τα εξής αποτελέσματα:

Φοιτητής	1	2	3
2	4		
3	5	4	
4	3	3	4

Πίνακας 2.6.2: Απόσταση Chebychev

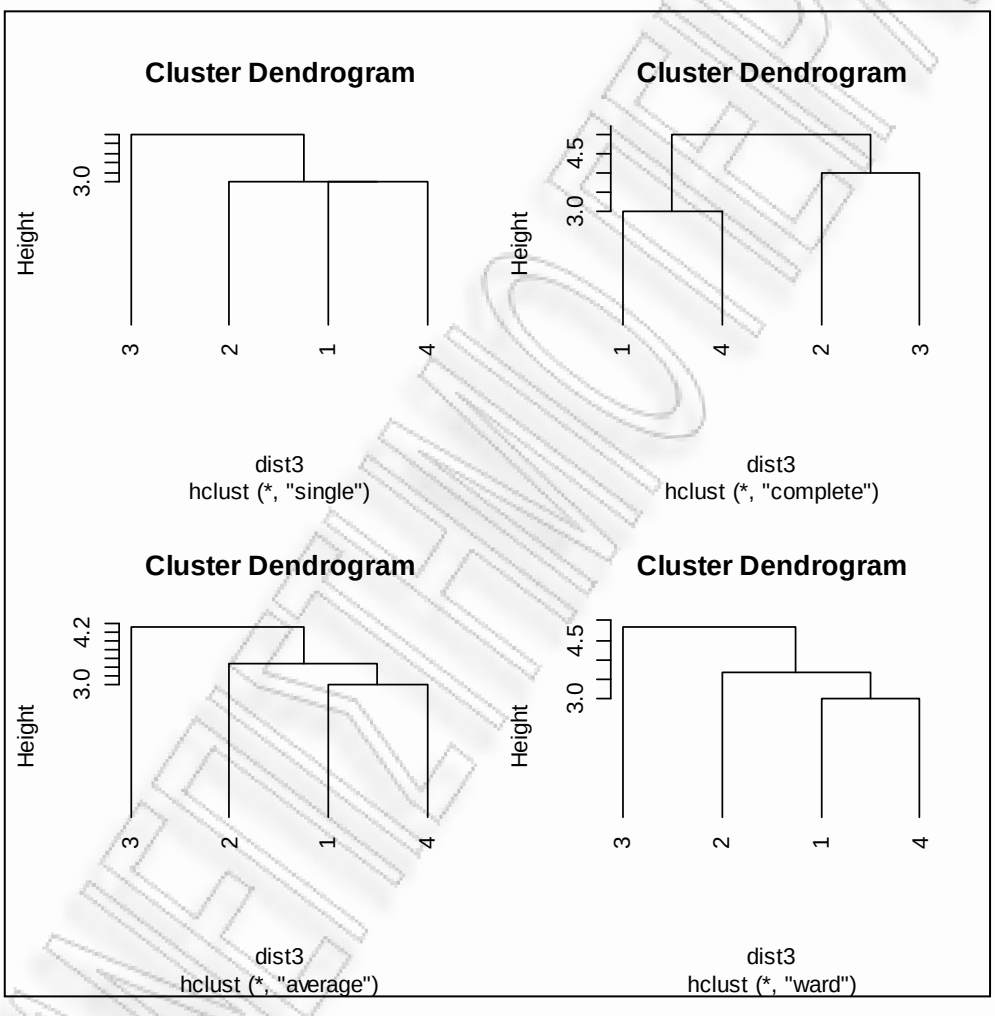


Σχήμα 2.4.2: Δενδρογράμματα Ιεραρχικών Συσσωρευτικών Μεθόδων (Απόσταση Chebychev)

Με τη μέθοδο της απλής συνένωσης δημιουργούνται ουσιαστικά 2 κλάσεις όπου η μια θα περιέχει τους φοιτητές 1, 2 και 4 , αφού συνενώνονται και οι 3 στην ίδια απόσταση, ενώ η άλλη θα περιέχει μόνο το φοιτητή 3. Αντιθέτως οι υπόλοιπες μέθοδοι συνενώνουν πρώτα τα άτομα 1 και 4 και στη συνέχεια τα άτομα 2 και 3. Με μοναδική αλλαγή λοιπόν το μέτρο της απόστασης βλέπουμε πόσο αλλάζουν τα αποτελέσματα στα οποία καταλήγουμε. Αυτό γίνεται ακόμα πιο εμφανές αν επιλέξουμε ως απόσταση την απόσταση Minkowski με m=400.

Φοιτητής	1	2	3
2	4.006937	-	-
3	5.013752	4	-
4	3	3	4.006937

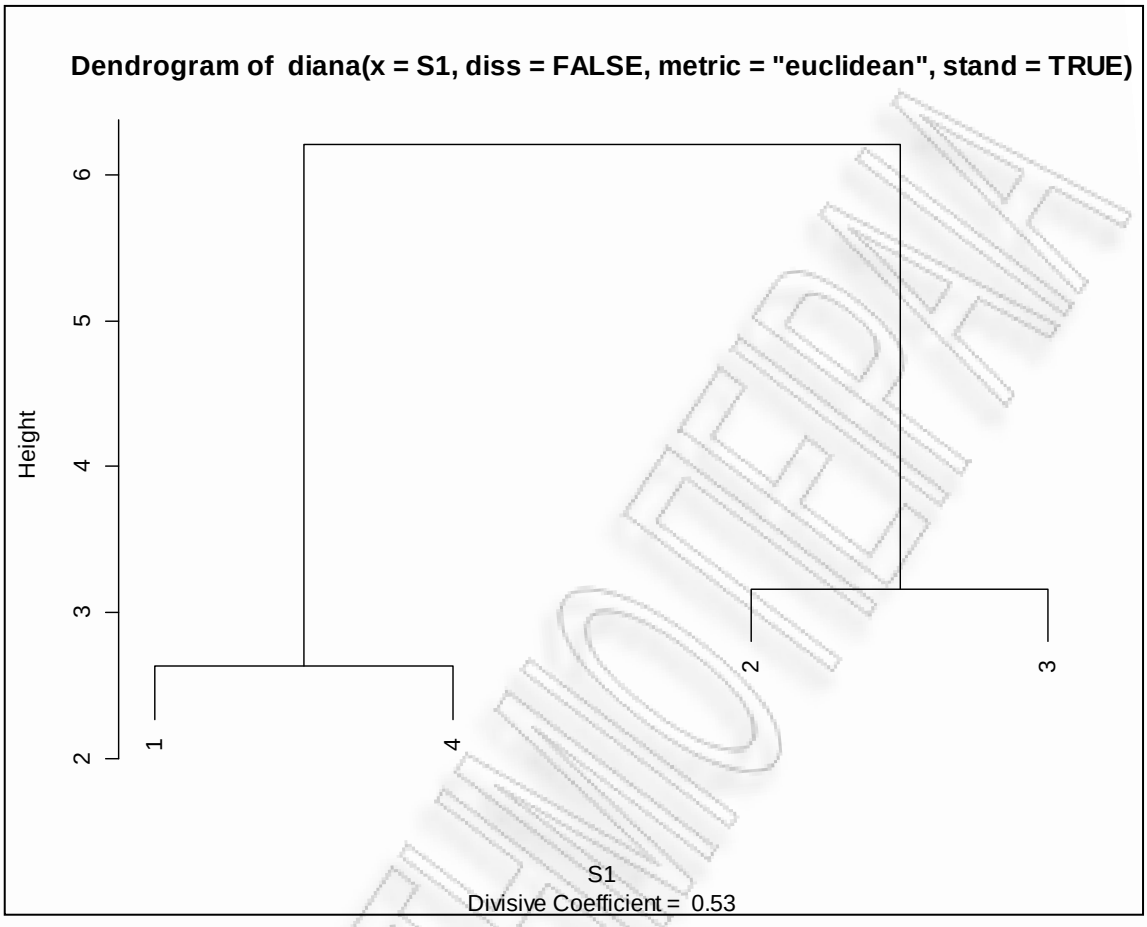
Πίνακας 2.6.3: Απόσταση Minkowski



Σχήμα 2.4.3: Δενδρογράμματα Ιεραρχικών Συσσωρευτικών Μεθόδων (Απόσταση Minkowski)

Σε αυτήν την περίπτωση η μέθοδος της απλής συνένωσης δημιουργεί δύο κλάσεις τις {1,2,4} και {3}. Η μέθοδος της πλήρους συνένωσης δημιουργεί τις κλάσεις {1,4} και {2,3} ενώ οι άλλες δύο μέθοδοι συνενώνουν πρώτα τα άτομα 1, 2, 4 και στη συνέχεια το 3.

Για τη μέθοδο DIANA σύμφωνα με το Σχήμα 2.6 προκύπτει ότι τα δεδομένα δημιουργούν τις κλάσεις {1,4} και {2,3}.



Σχήμα 2.5: Δενδρόγραμμα Ιεραρχικής Διαχωριστικής Μεθόδου

Στα ίδια αποτελέσματα με τη μέθοδο DIANA καταλήγει και η μέθοδος των *k*-means ενώ η μέθοδος PAM συμφωνεί με τα αποτελέσματα της απλής συνένωσης.

Από μελέτες προσομοίωσης έχει διαπιστωθεί ότι η καλύτερη ομαδοποίηση επιτυγχάνεται συνήθως με τη μέθοδο του Ward και τη μέθοδο της μέσης συνένωσης (οι οποίες είδαμε ότι είχαν πάντοτε σχεδόν τα ίδια αποτελέσματα). Αντιθέτως η μέθοδος με την πιο χαμηλή επίδοση είναι αυτή της απλής συνένωσης. Όπως θα διαπιστώσουμε και στη συνέχεια αυτής της εργασίας, επειδή δεν είναι σαφές ποια μέθοδος είναι η καλύτερη έχουν δημιουργηθεί διάφορα κριτήρια και δείκτες για να επικυρώνονται τα αποτελέσματα των μεθόδων. Επιπλέον δε θα πρέπει να ξεχνάμε ότι όλες οι ομαδοποιήσεις επηρεάζονται από τη φύση των δεδομένων καθώς επίσης και ότι αν τα δεδομένα μας δημιουργούν διακριτές ομάδες μεταξύ τους, τότε όλες οι μέθοδοι θα καταφέρουν να τις βρουν.

Κεφάλαιο 3

Μέτρα Επικύρωσης και Μέτρα Ευστάθειας

3.1 Εισαγωγή

Η ομαδοποίηση των δεδομένων, όπως αναλυτικά περιγράφηκε στο Κεφάλαιο 2, χρησιμοποιείται συχνά και στην ανάλυση βιολογικών δεδομένων με σκοπό τη δημιουργία ομάδων που περιέχουν γονίδια ή πρωτεΐνες που παρουσιάζουν παρόμοια συμπεριφορά και συνεπώς είναι πιθανόν να έχουν τις ίδιες βιολογικές λειτουργίες (*Eisen*, 1998), (*Bhattacherjee*, 2007). Η τελική απόφαση για το ποια μέθοδος συσταδοποίησης θα χρησιμοποιηθεί, για το αν οι κλάσεις που προκύπτουν είναι υψηλής ποιότητας ή αν υπάρχουν αντικείμενα που δεν είναι σωστά τοποθετημένα αποτελούν ερωτήματα τα οποία καλείται να απαντήσει ο ερευνητής κατά τη διαδικασία της ομαδοποίησης (*Rousseeuw*, 1987). Επιπλέον πρόβλημα είναι και η απόφαση που αφορά το τελικό πλήθος κλάσεων που «φυσικά» δημιουργούν τα δεδομένα. Ιδανικά οι κλάσεις οι οποίες έχουν προκύψει θα πρέπει να παρουσιάζουν καλές στατιστικές ιδιότητες. Αυτό συνεπάγεται ότι θα πρέπει να είναι συμπαγείς (*compact*), καλώς διαχωρισμένες (*well-separated*), συνδετικές (*connected*) και σταθερές (*stable*). Για τα δεδομένα που πραγματεύεται αυτή η εργασία όμως θα πρέπει επιπλέον οι δημιουργηθείσες κλάσεις να είναι και βιολογικώς ουσιαστικές (*biologically relevant*).

Στη διεθνή βιβλιογραφία υπάρχει μια μεγάλη ποικιλία μέτρων τα οποία χρησιμοποιούνται για την επικύρωση ή μη του αποτελέσματος που προκύπτει από την ανάλυση κατά συστάδες εξειδικευμένη στα βιολογικά σύνολα δεδομένων. Χαρακτηριστικά αναφέρουμε τις παρακάτω εργασίες οι οποίες αποτελούν οδηγό για κάποιον που μελετά το συγκεκριμένο θέμα αφού σε αυτές εισάγονται κάποια βασικά μέτρα επικύρωσης. Η πρώτη είναι η εργασία «*Validating clustering for gene expression data*» των *Yeung, Haynor* and *Ruzzo*, (2001) και η δεύτερη η «*Comparisons and validation of statistical clustering techniques fot microarray gene expression data*» των *Datta* and *Datta*, (2003). Η διαδικασία της επικύρωσης μπορεί να στηριχθεί είτε αποκλειστικά σε εσωτερικές ιδιότητες, είτε σε κάποιες εξωτερικές επιδράσεις που δέχονται τα δεδομένα τα υπό μελέτη ή ακόμα και σε κάποιες βιολογικές πληροφορίες που είναι διαθέσιμες για τα εκάστοτε σύνολα προς ανάλυση.

Οι τρεις βασικές κατηγορίες επικύρωσης είναι οι εξής:

- Εσωτερικά Μέτρα Επικύρωσης (*Internal Validation Measures*)
- Μέτρα Ευστάθειας (*Stability Measures*)
- Βιολογικά Μέτρα Επικύρωσης (*Biological Validation Measures*).

3.2 Εσωτερικά Μέτρα Επικύρωσης

Τα εσωτερικά μέτρα που θα παρουσιαστούν αντικατοπτρίζουν τη συνδεσιμότητα (*connectedness)*, την ομογένεια-συμπάγεια (*compactness*) και τη διαχωριστικότητα των κλάσεων (*separation*). Με τον όρο συνδεσιμότητα εννοούμε το βαθμό στον οποίο γειτονικές παρατηρήσεις τοποθετούνται στην ίδια κλάση και μετριέται διαμέσου της συνδετικότητας (*connectivity*). Η συμπάγεια αναφέρεται στην ομογένεια που υπάρχει στις κλάσεις μελετώντας συνήθως την εσωτερική διακύμανση των κλάσεων ενώ η διαχωριστικότητα των κλάσεων προσπαθεί να ποσοτικοποιήσει το βαθμό στον οποίο διαχωρίζονται οι κλάσεις, κυρίως με το να μετρά την απόσταση που υπάρχει ανάμεσα στα κέντρα των κλάσεων. Οι αντίστοιχοι δείκτες είναι ο δείκτης *Silhouette Width*, ο δείκτης *Dunn* και ο δείκτης *Davies-Bouldin*. Να σημειώσουμε ότι οι έννοιες της συμπάγειας και της διαχωριστικότητας είναι νοηματικά αντίθετες και η πρώτη αυξάνει με την αύξηση του αριθμού των κλάσεων ενώ η δεύτερη μειώνεται. Στη συνέχεια παρουσιάζονται αναλυτικά τα αντίστοιχα μέτρα.

Α) Συνδετικότητα

Έστω ότι N είναι ο συνολικός αριθμός παρατηρήσεων (γραμμών) σε ένα σύνολο δεδομένων και M είναι ο συνολικός αριθμός στηλών. Αν με $nn_{i(j)}$ συμβολίσουμε το j-οστό κοντινότερο γείτονα της παρατήρησης i τότε ορίζουμε:

$$x_{i,nn_i(j)} = \begin{cases} 0, \text{ όταν } i, j \text{ ανήκουν σε ίδια ομάδα,} \\ \dfrac{1}{j}, \quad \text{όταν ανήκουν σε διαφορετική.} \end{cases}$$

Η συνδετικότητα για μια συγκεκριμένη ομαδοποίηση $C=\{C_1,\ldots,C_K\}$ των N παρατηρήσεων σε K μη επικαλυπτόμενες κλάσεις δίνεται από τη σχέση:

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_i(j)}$$

όπου L είναι μια παράμετρος που συμβολίζει τον αριθμό των κοντινότερων γειτόνων που κάθε φορά θα χρησιμοποιηθούν. Το μέτρο αυτό παίρνει τιμές από το 0 ως το ∞ και όσο μικρότερη είναι η τιμή του τόσο καλύτερη είναι η ομαδοποίηση.

Β) Δείκτης Silhouette Width

Για την κατασκευή του δείκτη *silhouette* είναι απαραίτητη η ύπαρξη μιας ομαδοποίησης καθώς και ο πίνακας των αποστάσεων που έχει χρησιμοποιηθεί για την εξαγωγή της ομαδοποίησης (*Rousseeuw*, 1987). Ο ορισμός θα γίνει αρχικά για την περίπτωση όπου μας δίνονται τα μέτρα αποστάσεων και όχι τα μέτρα ομοιότητας ανάμεσα στα αντικείμενα. Έστω ένα αντικείμενο i και συμβολίζουμε με A την κλάση στην οποία ανήκει (βάσει του αλγορίθμου ομαδοποίησης που έχει χρησιμοποιηθεί προηγουμένως). Όταν η κλάση A περιλαμβάνει και άλλα αντικείμενα πέραν του i τότε ορίζουμε ως $a(i)$ τη μέση ανομοιότητα του i με τα υπόλοιπα αντικείμενα της κλάσης A. Στη συνέχεια θεωρούμε μια άλλη κλάση C, η οποία είναι διαφορετική της A, και ορίζουμε ως $d(i,C)$ τη μέση ανομοιότητα του αντικειμένου i με όλα τα αντικείμενα της κλάσης C. Αφού υπολογιστούν όλες οι αποστάσεις $d(i,C)$ για όλες τις κλάσεις πλην της A επιλέγουμε τη μικρότερη από αυτές η οποία συμβολίζεται με $b(i)$, δηλαδή

$$b(i) = \min_{C \neq A} d(i, C).$$

Η κλάση B για την οποία επιτυγχάνεται το ελάχιστο $b(i)$ καλείται γείτονας (*neighbor*) του αντικειμένου i αφού, σε περίπτωση που αυτό δεν ανήκε στην κλάση A, η αμέσως καλύτερη επιλογή του θα ήταν να ανήκει στην κλάση B. Το *silhouette* συμβολίζεται με $s(i)$ και δίνεται από τον εξής τύπο:

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & \text{αν } a(i)<b(i),\\ 0 & \text{αν } a(i)=b(i),\\ \dfrac{a(i)}{b(i)}-1, & \text{αν } a(i)>b(i).\end{cases}$$

Ισοδύναμα το *silhouette* μπορεί να εκφραστεί μέσω του τύπου:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}.$$

Όταν η κλάση *A* περιέχει ένα μόνο αντικείμενο θέτουμε *s*(*i*)=0. Από τους παραπάνω ορισμούς είναι φανερό ότι ο δείκτης *s*(*i*) παίρνει τιμές από -1 ως 1. Καλύτερη τιμή θεωρείται το 1 αφού τότε η εντός ανομοιότητα *a*(*i*) είναι πολύ μικρότερη από την μικρότερη ανομοιότητα εκτός, δηλαδή την *b*(*i*). Αντιθέτως χειρότερη τιμή είναι το -1 αφού αυτό συνεπάγεται ότι το *a*(*i*) είναι πολύ μεγαλύτερο από το *b*(*i*) και συνεπώς το αντικείμενο *i* βρίσκεται πιο κοντά στην κλάση *B* από ότι στην *A* που έχει τοποθετηθεί. Ο δείκτης *Silhouette* έχει το πλεονέκτημα ότι δεν εξαρτάται από τον αλγόριθμο από τον οποίο προέκυψε η ομαδοποίηση αλλά μόνο από την ίδια την ομαδοποίηση. Συνεπώς μπορεί να χρησιμοποιηθεί και για την επιλογή του βέλτιστου αριθμού κλάσεων αν συγκρίνουμε μεταξύ τους δείκτες που προκύπτουν από διαφορετικές ομαδοποιήσεις.

Για κάθε κλάση ορίζεται το *Μέσο Μήκος Silhouette* (*Average Silhouette Width*) ως ο μέσος όρος των *s*(*i*) για όλα τα αντικείμενα *i* που ανήκουν στην εξεταζόμενη κλάση. Επίσης ορίζεται και το *Συνολικό Μέσο Μήκος Silhouette* (*Overall Average Silhouette Width*) το οποίο είναι ο μέσος όρος *s*(*i*) όλων των αντικειμένων *i* που ανήκουν στο σύνολο των δεδομένων και το οποίο συμβολίζεται με $\bar{s}(k)$ όπου *k* είναι ο αριθμός των κλάσεων. Είναι προφανές ότι επιλέγουμε εκείνο το πλήθος κλάσεων που μεγιστοποιούν την τιμή $\bar{s}(k)$.

Αν και στα περισσότερα στατιστικά πακέτα για τον υπολογισμό του *Silhouette Width* χρησιμοποιούνται πίνακες αποστάσεων (Ευκλείδεια, *Manhattan* κλπ.) παραθέτουμε και τον ορισμό σε περίπτωση που μας δίνεται ο πίνακας ομοιοτήτων. Ως $a'(i)$ ορίζεται η μέση ομοιότητα του *i* με τα υπόλοιπα αντικείμενα της κλάσης *A*, $d'(i,C)$ η μέση ομοιότητα του αντικειμένου *i* με όλα τα αντικείμενα της κλάσης *C* και το $b'(i)$ είναι η μεγαλύτερη από τις μέσες ομοιότητες, δηλαδή:

$$b'(i) = \max_{C \neq A} d'(i, C).$$

Συνεπώς

$$s(i) = \begin{cases} 1 - \dfrac{b'(i)}{a'(i)}, \text{αν } a'(i) > b'(i), \\ 0, \text{αν } a'(i) = b'(i), \\ \dfrac{a'(i)}{b'(i)} - 1, \text{αν } a'(i) < b'(i). \end{cases}$$

Να σημειώσουμε ότι το *silhouette* ως δείκτης δεν μπορεί πάντοτε να καθορίσει το σωστό αριθμό κλάσεων κυρίως όταν ως δεδομένα χρησιμοποιούνται γονίδια. (*Famili*, 2004).

Γ) Δείκτης Dunn

Ένα ακόμη δημοφιλές μέτρο το οποίο προσπαθεί να κρατήσει υψηλά την εντός κλάσεως ομογένεια και συγχρόνως να επιλέξει την ομαδοποίηση που δημιουργεί καλώς διαχωρισμένες κλάσεις είναι ο δείκτης ***Dunn*** (*Dunn*, 1974). Το μέτρο αυτό χρησιμοποιώντας μη γραμμικές τεχνικές προσπαθεί να ενώσει τις αντιφατικές έννοιες της συμπάγειας και της διαχωριστικότητας. Ο δείκτης *Dunn* ορίζεται ως ο λόγος της μικρότερης απόστασης ανάμεσα σε παρατηρήσεις που δε βρίσκονται στην ίδια κλάση προς τη μεγαλύτερη απόσταση που παρατηρείται ανάμεσα σε παρατηρήσεις που τοποθετούνται στην ίδια κλάση. Συνεπώς αν $C=\{C_1,\ldots,C_K\}$ είναι μια ομαδοποίηση των N παρατηρήσεων σε K μη επικαλυπτόμενες κλάσεις τότε ο δείκτης δίνεται από τον τύπο:

$$D(C) = \frac{\min\limits_{C_k, C_l, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} d(i, j) \right)}{\max\limits_{C_m \in C} diam(C_m)}$$

όπου με $diam(C_m)$ συμβολίζεται η μεγαλύτερη απόσταση ανάμεσα στα αντικείμενα που έχουν ταξινομηθεί στην κλάση C_m. Ο δείκτης παίρνει τιμές από 0 ως ∞ και όσο μεγαλύτερος τόσο καλύτερη είναι η ομαδοποίηση.

Δ) Δείκτης Davies-Bouldin

Ένας ακόμη δείκτης που προσπαθεί να επιλέξει εκείνη την ομαδοποίηση που δημιουργεί συμπαγείς και καλά διαχωρισμένες κλάσεις είναι ο δείκτης Davies-Bouldin (*Davies-Bouldin*, 1974). Άν $C=\{C_1,\ldots,C_K\}$ είναι μια ομαδοποίηση των N παρατηρήσεων σε K μη επικαλυπτόμενες κλάσεις τότε ο δείκτης δίνεται από τον τύπο:

$$DB(C) = \frac{1}{k}\max_{i\neq j}\left\{\frac{diam(C_k)+diam(C_l)}{d_{i\in C_k, j\in C_l}(i,j)}\right\}$$

όπου k είναι ο αριθμός των κλάσεων που προκύπτουν από την ομαδοποίηση, $diam(C_k)$ είναι η μεγαλύτερη απόσταση ανάμεσα στα αντικείμενα της κλάσης C_k (εντός κλάσεως απόσταση-*intracluster distance*) και $d(i,j)$ είναι η απόσταση ανάμεσα στα στοιχεία i και j των κλάσεων C_k και C_l αντίστοιχα (*intercluster distance*). Μικρές τιμές του δείκτη αντιστοιχούν σε κλάσεις που είναι ομογενείς και τα κέντρα τους είναι απομακρυσμένα, δηλαδή χαρακτηρίζονται ως καλώς διαχωρισμένες. Συνεπώς η ομαδοποίηση εκείνη που δίνει τη μικρότερη τιμή στο δείκτη αυτό θεωρείται και ως η ιδανική.

3.3 Μέτρα Ευστάθειας

Η τεχνολογία των μικροσυστοιχιών, όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο, δίνει τη δυνατότητα μελέτης και επεξεργασίας τεράστιων συνόλων δεδομένων τα οποία περιέχουν καταγεγραμμένη την ταυτόχρονη έκφραση χιλιάδων γονιδίων σε διαφορετικές χρονικές στιγμές κατά τη διάρκεια μια βιολογικής διαδικασίας. Κατά συνέπεια οι βιολόγοι τα τελευταία χρόνια προσπαθούν να ομαδοποιήσουν τα γονίδια βασιζόμενοι στα επίπεδα έκφρασης τους. Νέες τεχνικές ομαδοποίησης αλλά και νέες μέθοδοι που προσπαθούν να επικυρώσουν τα αποτελέσματα της ανάλυσης κατά συστάδες έχουν δημιουργηθεί εστιάζοντας την εφαρμογή τους στα εν λόγω δεδομένα. Οι *Datta & Datta* (2003) στην εργασία τους *Comparisons and Validation of statistical clustering techniques for microarray gene expression data*, παρουσιάζουν τρία νέα διαφορετικά μέτρα ευστάθειας τα οποία ανταποκρίνονται πολύ καλά όταν τα δεδομένα που επεξεργαζόμαστε

είναι σε υψηλό βαθμό συσχετισμένα, γεγονός που συμβαίνει σε γονιδιακά δεδομένα. Οι *Yeung, Haynor* and *Ruzzo* (2001) στην εργασία τους με τίτλο: *Validating clustering for gene expression data*, εισάγουν ένα ακόμη νέο μέτρο το ***Figure Of Merit*** το οποίο μπορεί και εκτιμά τη δύναμη ενός αλγορίθμου ομαδοποίησης και συγχρόνως εκτιμά και το βέλτιστο πλήθος κλάσεων που δημιουργείται από τα δεδομένα ενώ οι *Famili, G. Liu & Z. Liu* (2004) στην εργασία τους *Evaluation and optimization of clustering in gene expression data analysis*, εισάγουν ένα ακόμη νέο μέτρο την ευστάθεια (*Stability*) η οποία εκτιμά την «ακινησία» (*Immovability*) των αντικειμένων όταν μια ήδη υπάρχουσα κλάση υποδιαιρείται σε άλλες μικρότερες.

Α) Μέτρα Ευστάθειας των Datta&Datta

Η βασική ιδέα είναι ότι, αν ένας αλγόριθμος είναι σωστός, θα πρέπει να είναι και συνεπής. Δηλαδή έστω ότι έχουμε συγκεντρώσει ένα σύνολο δεδομένων γονιδιακής έκφρασης σε διάφορες χρονικές στιγμές $T_1, T_2, ..., T_\ell$. Αυτό συνεπάγεται ότι τα δεδομένα μας είναι ℓ –διάστατα διανύσματα στον Ευκλείδειο χώρο $\Re^\ell$. Για κάθε $i = 1,2,...,\ell$ επαναλαμβάνουμε τον αλγόριθμο παραγωγής συστάδων για κάθε σύνολο δεδομένων στον $\Re^{\ell-1}$ χώρο, ο οποίος δημιουργείται αν από κάθε σύνολο αφαιρέσουμε τις παρατηρήσεις που αντιστοιχούν στον χρόνο T_i. Για κάθε γονίδιο g όπου $1 \le g \le M$ συμβολίζουμε με $C^{g,i}$ την κλάση που περιέχει το γονίδιο g και για τη δημιουργία της δεν έχει χρησιμοποιήσει τα δεδομένα που ανήκουν στον χρόνο T_i, δηλαδή από τον πίνακα δεδομένων έχει αφαιρεθεί η i στήλη . Αντιστοίχως ορίζεται ως $C^{g,0}$ η κλάση που περιέχει το γονίδιο g και για την παραγωγή της έχει χρησιμοποιηθεί ολόκληρο το σύνολο των δεδομένων. Οι Datta and Datta (2003) όρισαν τα παρακάτω μέτρα επικύρωσης:

❖ Μέτρο της Μέσης Αναλογίας μη Επικάλυψης (***Average Proportion of Non-Overlap Measure-APN***)

Το APN μετρά το μέσο ποσοστό των παρατηρήσεων που δεν τοποθετούνται στην ίδια κλάση όταν η ομαδοποίηση δημιουργείται από το πλήρες σύνολο δεδομένων και όταν δημιουργείται από το σύνολο που του έχει αφαιρεθεί μια στήλη. Αν θεωρήσουμε ότι K είναι ο αριθμός των κλάσεων που αναμένεται να δημιουργηθεί, τότε το APN δίνεται από τον εξής τύπο:

$$APN(K) = \frac{1}{M\ell}\sum_{g=1}^{M}\sum_{i=1}^{\ell}\left(1-\frac{n(C^{g,i}\cap C^{g,0})}{n(C^{g,0})}\right).$$

Όπου με $n(C)$ συμβολίζεται ο πληθάριθμος της κλάσης C, δηλαδή το σύνολο των στοιχείων που συμπεριλαμβάνονται στην κλάση C. Το μέτρο APN παίρνει τιμές στο διάστημα [0,1] και όσο μικρότερη είναι η τιμή του τόσο πιο συνεπή είναι τα αποτελέσματα της ανάλυσης κατά συστάδες.

❖ Μέτρο της Μέσης Απόστασης Μέσων (***Average Distance between Means Measure-ADM***)

Το μέτρο ADM υπολογίζει τη μέση απόσταση ανάμεσα στα κέντρα των κλάσεων για τις παρατηρήσεις που έχουν τοποθετηθεί στην ίδια κλάση από την ομαδοποίηση που προκύπτει από το σύνολο των δεδομένων και από την ομαδοποίηση που προκύπτει από τα δεδομένα που τους έχει αφαιρεθεί μία στήλη. Υπολογίζεται από τον ακόλουθο τύπο:

$$ADM(K) = \frac{1}{M\ell}\sum_{g=1}^{M}\sum_{i=1}^{\ell} d(\bar{x}_{C^{g,i}}, \bar{x}_{C^{g,0}}),$$

όπου με $\bar{x}_{C^{g,0}}$ συμβολίζεται η μέση έκφραση των γονιδίων στην κλάση $C^{g,0}$ ενώ με $\bar{x}_{C^{g,i}}$ συμβολίζεται η μέση έκφραση των γονιδίων στην κλάση $C^{g,i}$. Το μέτρο ADM συνήθως χρησιμοποιεί μόνο την Ευκλείδεια απόσταση ενώ παίρνει τιμές από 0 ως και το ∞. Όσο μικρότερη είναι η τιμή του τόσο καλύτερη είναι η ομαδοποίηση.

❖ Μέτρο της Μέσης Απόστασης ***(Average Distance Measure-AD)***

Το AD μέτρο υπολογίζει τη μέση απόσταση ανάμεσα στις παρατηρήσεις που έχουν τοποθετηθεί στην ίδια κλάση από την ομαδοποίηση που προκύπτει από ολόκληρο το σύνολο δεδομένων και από την ομαδοποίηση που στηρίζεται στα δεδομένα που τους έχει αφαιρεθεί μία στήλη. Υπολογίζεται από τον παρακάτω τύπο:

$$AD(K) = \frac{1}{M\ell}\sum_{g=1}^{M}\sum_{i=1}^{\ell}\frac{1}{n(C^{g,0})n(C^{g,i})}\times \sum_{g\in C^{g,0},g'\in C^{g,i}} d(x_g, x_{g'})$$

όπου $d(x_g, x_{g'})$ είναι η απόσταση (Ευκλείδεια, *Manhattan* κλπ) ανάμεσα στα επίπεδα έκφρασης των γονιδίων *g* και *g'*. Οι τιμές που μπορεί να πάρει το μέτρο αυτό κυμαίνονται από 0 ως το ∞ και όσο πιο μικρές είναι, τόσο πιο αξιόπιστη είναι η μέθοδος ομαδοποίησης που έχει χρησιμοποιηθεί.

B) Figure Of Merit (FOM)

Η μέθοδος αυτή εκτιμά την ποιότητα των αποτελεσμάτων που προκύπτουν από την εκάστοτε μέθοδο ανάλυσης συστάδων χρησιμοποιώντας τη μεθοδολογία *Jackknife* η οποία προτάθηκε από τον *Efron* (1982). Κατά συνέπεια μια μέθοδος ομαδοποίησης εφαρμόζεται σε όλες τις πειραματικές συνθήκες εκτός από μία στο σύνολο των δεδομένων και η συνθήκη η οποία δε χρησιμοποιήθηκε καθορίζει την προβλεπτική ικανότητα του αλγορίθμου. Το μέτρο που χρησιμοποιείται για να εκτιμήσει ποσοτικά την προβλεπτική ικανότητα της μεθόδου καλείται *figure of merit* ενώ αξίζει να παρατηρήσουμε πως η λογική αυτή είναι παρόμοια με αυτή που χρησιμοποιήθηκε από τους *Datta* and *Datta* (2003).

Τυπικά ένα σύνολο δεδομένων βιολογικής προέλευσης αποτελείται από τις μετρήσεις των επιπέδων έκφρασης *n* γονιδίων κάτω από *m* πειραματικές συνθήκες. Έστω ότι ένας αλγόριθμος συσταδοποίησης εφαρμόζεται στα δεδομένα υπό τις πειραματικές συνθήκες 1,…, (*e*-1), (*e*+1),…, *m* και η κατάσταση *e* χρησιμοποιείται για την εκτίμηση της προβλεπτικής ικανότητας του αλγορίθμου. Υποθέτουμε ότι τα δεδομένα δημιουργούν *k* συστάδες τις C_1, C_2, …, C_k και με *R*(*g*,*e*) συμβολίζεται το επίπεδο έκφρασης του γονιδίου *g* υπό την πειραματική συνθήκη *e*. Αν με $\mu_{C_i(e)}$ συμβολίσουμε το μέσο επίπεδο έκφρασης όλων των γονιδίων της κλάσης C_i στην κατάσταση *e*, το FOM δίνεται από τον ακόλουθο τύπο:

$$FOM(e,k) = \sqrt{\frac{1}{n} * \sum_{i=1}^{k}\sum_{x\in C_i}(R(x,e) - \mu_{C_i}(e))^2}.$$

Αξίζει να παρατηρήσουμε ότι το *FOM*(*e*,*k*) παρέχει μια εκτίμηση του μέσου σφάλματος πρόβλεψης βασιζόμενο στους μέσους των κλάσεων, ενώ επιπλέον κάθε μία από τις *m* πειραματικές συνθήκες μπορεί να χρησιμοποιηθεί σα συνθήκη επικύρωσης. Οι τιμές του κυμαίνονται από 0 ως το ∞ και όσο μικρότερες είναι τόσο καλύτερος είναι ο αλγόριθμος που χρησιμοποιήθηκε. Το συνολικό *FOM*(*k*) *(Aggregated Figure Of Merit),* ισούται με:

$$FOM(k) = \sum_{e=1}^{m} FOM(e,k)$$

και αποτελεί μια εκτίμηση της συνολικής προβλεπτικής ικανότητας του αλγορίθμου κατά μήκος όλων των πειραματικών συνθηκών.

Όσο ο αριθμός των κλάσεων αυξάνεται έχει παρατηρηθεί, τόσο σε πραγματικά δεδομένα όσο και σε εργαστηριακά, ότι το FOM μειώνεται ανεξάρτητα από το ποια μέθοδος ομαδοποίησης θα χρησιμοποιηθεί. Αυτό συμβαίνει είτε επειδή οι αλγόριθμοι μπορεί να βρίσκουν κλάσεις καλύτερης ποιότητας καθώς διαχωρίζουν τις τεράστιες αρχικές κλάσεις σε μικρότερες και κατά συνέπεια πιο ομογενείς, είτε επειδή πολύ απλά το FOM τείνει να μειώνεται όσο αυξάνεται ο αριθμός των κλάσεων. Αυτό έχει ως αποτέλεσμα την αναγκαία εισαγωγή ενός προσαρμοσμένου FOM που θα ξεπερνά αυτό το πρόβλημα.

Έστω ότι ισχύει το εξής τέλειο σενάριο: τα *n* γονίδια δημιουργούν *c* κλάσεις και η *i*-οστή κλάση περιέχει $a_i n$ γονίδια όπου $0 < \alpha_i < 1$ και $\sum_{i=1}^{c} a_i = 1$. Το σενάριο θεωρείται τέλειο με την έννοια ότι σε κάθε κλάση συμπεριλαμβάνονται διαφορετικά γονίδια. Επιπλέον υποθέτουμε ότι τα επίπεδα έκφρασης των γονιδίων στην κλάση *i* υπό τη συνθήκη *e* είναι ανεξάρτητες μεταβλητές που ακολουθούν την κανονική κατανομή με μέσο $\mu_{i,e}$ και διασπορά $\sigma_{i,e}^2$. Στο επόμενο στάδιο εφαρμόζουμε έναν αλγόριθμο ομαδοποίησης στα *n* γονίδια για να αποκτήσουμε *k* κλάσεις, όπου $k \geq c$. Με βάση τις παραπάνω υποθέσεις δίνεται το ακόλουθο θεώρημα:

<u>Θεώρημα 1:</u> (*Yeung, Haynor, Ruzzo,* 2001). Το αναμενόμενο συνολικό FOM το οποίο συμβολίζεται με *FOM*(*k*) είναι ίσο με $\sqrt{\frac{n-k}{n}}\bar{\sigma}$ όπου $\bar{\sigma}$ είναι ο σταθμισμένος μέσος του $\sigma_{i,e}$ ο οποίος είναι ανεξάρτητος από το *k* και δίνεται από τον τύπο:

$$\bar{\sigma} = \sum_{e=1}^{m} \sqrt{\sum_{i=1}^{c} a_i \sigma_{i,e}^2}.$$

<u>Απόδειξη:</u> Έστω ότι οι μετρήσεις των επιπέδων έκφρασης των $a_i n$ γονιδίων στην πραγματική κλάση i και υπό τη συνθήκη e είναι οι $x_{1,e},...,x_{a_i n,e}$ και ο σταθμισμένος μέσος τους είναι ο $\bar{x}_e = \sum_{i=1}^{a_i n} \frac{x_{i,e}}{a_i n}$. Τότε η αναμενόμενη τιμή της ποσότητας $\sum_{i=1}^{a_i n} (x_{i,e} - \bar{x}_e)^2$ ισούται με $(a_i n - 1)\sigma_{i,e}^2$. Αν διαιρέσουμε την κλάση i σε $a_i k$ μικρότερες κλάσεις τότε η συνεισφορά αυτών των γονιδίων στο αναμενόμενο *FOM*(*e*,*k*) θα περιοριζόταν σε $(a_i n - a_i k)\sigma_{i,e}^2$. Κατά συνέπεια:

$$FOM(k) = \sum_{e=1}^{m} \sqrt{\frac{1}{n} \sum_{i=1}^{c} (a_i n - a_i k)\sigma_{i,e}^2} = \sqrt{\frac{n-k}{n}} \sum_{e=1}^{m} \sqrt{\sum_{i=1}^{c} a_i \sigma_{i,e}^2}. \qquad \square$$

Αν ισχύουν οι υποθέσεις του Θεωρήματος 1, ο ρυθμός με τον οποίο μειώνεται το FOM καθώς ο αριθμός των κλάσεων k αυξάνεται πρέπει να ισούται με $\sqrt{\frac{n-k}{n}}$. Έτσι καταλήγουμε ότι το προσαρμοσμένο FOM ισούται με $\frac{FOM(e,k)}{\sqrt{\frac{n-k}{n}}}$.

Οι περιορισμοί της συγκεκριμένης μεθόδου έγκεινται στο γεγονός ότι αν κάποιες πειραματικές συνθήκες περιέχουν ανεξάρτητη πληροφορία τότε το μέτρο δεν μπορεί να υπολογιστεί. Επιπλέον μη χρησιμοποιώντας διαθέσιμη πληροφορία έχει ως συνέπεια σημαντική υποβάθμιση της ποιότητας των κλάσεων που παράγονται από όλες τις μεθόδους ενώ είναι σημαντικό να τονίσουμε ότι με βάση τον ορισμό του συγκεκριμένου μέτρου δεν είναι ασφαλές να συγκρίνονται μέθοδοι ομαδοποίησης που καταλήγουν σε διαφορετικό πλήθος κλάσεων ή που χρησιμοποιούν διαφορετικές μετρικές.

Γ) Δείκτης Ευστάθειας

Με βάση τα παραπάνω μέτρα είναι εμφανές ότι είναι αναγκαία η εισαγωγή ενός επιπλέον μέτρου το οποίο θα παράγεται χρησιμοποιώντας κάθε φορά ολόκληρο το σύνολο των δεδομένων και συνεπώς θα εκμεταλλεύεται κάθε διαθέσιμη πληροφορία. Το μέτρο αυτό στηρίζεται στην ακινησία των αντικειμένων από μια κλάση όταν αυτή διαιρείται σε άλλες μικρότερες. Με τον όρο ακινησία εννοούμε το ρυθμό με τον οποίο τα αντικείμενα μιας κλάσης παραμένουν αμετάβλητα κατά τη διαδικασία μιας νέας ομαδοποίησης που παράγει $i+n$ κλάσεις ($n>0$) από τις i που υπήρχαν αρχικά.

Ας υποθέσουμε ότι σε ένα σύνολο δεδομένων εφαρμόζεται πολλές φορές ο ίδιος αλγόριθμος παραγωγής ομάδων και καταλήγουμε στην ύπαρξη ενός συνόλου ομαδοποιήσεων οι οποίες παράγουν από 2 ως n κλάσεις. Επιλέγουμε μια από αυτές τις ομαδοποιήσεις και έστω ότι $C_{c,l}$ είναι ένα σύνολο αντικειμένων στην κλάση l από τις συνολικά c κλάσεις που έχουν παραχθεί από την ομαδοποίηση που επιλέξαμε ($2 \le c \le n$). Έστω k το κατώτατο όριο για το οποίο υπολογίζεται η ευστάθεια μιας κλάσης ($0 \le k \le n\text{-}c$). Η ευστάθεια της κλάσης l δίνεται τότε από τον ακόλουθο τύπο:

$$S_{c,l} = \min_{i=c+1}^{c+k} \left\{ \max_{j=1}^{i} \left\{ \frac{|C_{c,l} \cap C_{i,j}|}{|C_{c,l}|} \right\} \right\}.$$

Ουσιαστικά υπολογίζονται οι k τιμές για το μέγιστο αριθμό αντικειμένων που επικαλύπτονται μεταξύ της κλάσης l και κάθε άλλης κλάσης που προκύπτει από την ομαδοποίηση που παράγει από $c+1$ ως $c+k$ κλάσεις. Στη συνέχεια επιλέγεται η κατώτερη από τις k τιμές και αυτή θεωρείται ως ευστάθεια της κλάσης l. Για να είναι συγκρίσιμα τα αποτελέσματα που παράγονται από διαφορετικές κλάσεις, το μέτρο αυτό κανονικοποιείται διαιρώντας με το συνολικό αριθμό αντικειμένων της κλάσης l. Κατά συνέπεια ο δείκτης αυτός μπορεί να πάρει τιμές από 0 ως 1 και όσο πιο κοντά στη μονάδα κυμαίνεται τόσο πιο ευσταθής θεωρείται η κλάση.

Στη συνέχεια δίνεται και ο αντίστοιχος ορισμός για τη *Γενική Ευστάθεια* (*General Stability*) ενός αποτελέσματος παραγωγής συστάδων που παράγει c το πλήθος συστάδες. Η γενική ευστάθεια ορίζεται ως ο μέσος όρος όλων των ευσταθειών για όλες τις κλάσεις που έχουν παραχθεί από τον αλγόριθμο, δηλαδή

$$GS_c = \frac{1}{c}\sum_{l=1}^{c} S_{c,l}.$$

Ο αριθμός *c* για τον οποίο η γενική ευστάθεια μεγιστοποιείται αποτελεί και το βέλτιστο αριθμό συστάδων που δημιουργείται από τα δεδομένα. Όσο πιο υψηλή είναι η γενική ευστάθεια τόσο καλύτερης ποιότητας είναι οι κλάσεις που παράγονται. Το μέτρο αυτό έχει τη δυνατότητα να δουλεύει για γονιδιακά δεδομένα διαφόρων ειδών και έχει αποδειχθεί πειραματικά ότι καταφέρνει να ξεχωρίζει κλάσεις που έχουν σημαντική ερμηνευτική σημασία. Επιπλέον ο υπολογισμός του γίνεται με το πέρας της ομαδοποίησης και συνεπώς στη τιμή του παίζουν ρόλο όλοι οι παράγοντες που χαρακτηρίζουν μια κλάση ως καλή (ομογένεια, διαχωριστικότητα, σχήμα) και όχι κάποιοι από αυτούς όπως για παράδειγμα συμβαίνει στον υπολογισμού του *silhouette width*.

3.4 Βιολογικά Μέτρα Επικύρωσης

Οι Datta and Datta (2006) στην εργασία τους «*Methods for evaluating clustering algorithms for gene expression data using reference set of functional classes*» παρουσιάζουν δύο νέα μέτρα τα οποία κρίνουν αν τα αποτελέσματα ενός αλγορίθμου ομαδοποίησης παράγουν βιολογικά σημαντικές κλάσεις. Το πρώτο μέτρο καλείται *Βιολογικός Δείκτης Ομογένειας* (*Biological Homogeneity Index – BHI*) και όπως προδίδει και το όνομά του χαρακτηρίζει το πόσο βιολογικά ομογενείς είναι οι δημιουργηθείσες κλάσεις. Το μέτρο αυτό χρησιμοποιείται για να κρίνει την επίδοση ενός αλγορίθμου αλλά συγχρόνως και για να συγκρίνει πολλούς αλγορίθμους μεταξύ τους που ο καθένας προσπαθεί να ανακαλύψει τις κλάσεις που δημιουργούνται από το ίδιο σύνολο δεδομένων. Το δεύτερο μέτρο καλείται *Βιολογικός Δείκτης Ευστάθειας* (*Biological Stability Index – BSI*) και μετρά πόσο συνεπής είναι ένας αλγόριθμος στο να παράγει βιολογικά ουσιώδεις κλάσεις όταν εφαρμόζεται επαναλαμβανόμενα σε παρόμοια σύνολα δεδομένων. Μια καλή ομαδοποίηση συνεπάγεται ότι και οι δύο δείκτες παρουσιάζουν σχετικά υψηλές τιμές.

Ας υποθέσουμε ότι έχουμε διαθέσιμα ένα σύνολο γονιδίων $\mathbb{G}$ και $C_1, C_2, \ldots, C_F$ είναι F λειτουργικές κλάσεις οι οποίες μπορεί και να αλληλεπικαλύπτονται. Επειδή όλα τα γονίδια δεν είναι λειτουργικώς ενεργά ορίζουμε $C := C_i \subset \mathbb{G}$ να είναι το σύνολο των ενεργά λειτουργικών γονιδίων. Στο σημείο αυτό είναι σημαντικό να παρατηρήσουμε ότι στο ευρύ κοινό είναι διαθέσιμες βιβλιοθήκες οι οποίες παρέχουν πληροφορίες για τα γονίδια και για την οργάνωση τους σε ομάδες με βάση τα βιολογικά χαρακτηριστικά τους ή τις κωδικοποιημένες τους πρωτεΐνες και κατά επέκταση βοηθούν στο να κρίνουμε ποια από αυτά είναι λειτουργικά ενεργά ή μη. Χαρακτηριστικά

αναφέρουμε την ***Gene Ontology*** (*GO*), την ***Entrez Gene*** και το ***Unigene cluster***. Επιπλέον έχουν κατασκευαστεί εξειδικευμένα λογισμικά όπως το ***AmiGO*** για την παραγωγή αυτής της πληροφορίας.

Α) Βιολογικός Δείκτης Ομογένειας – BHI

Έστω δύο ενεργά γονίδια τα *x, y* τα οποία ανήκουν στην ίδια στατιστική κλάση *D* και *C*(*x*), *C*(*y*) είναι οι λειτουργικές κλάσεις που ανήκουν τα γονίδια *x* και *y* αντίστοιχα. Ορίζουμε τη δείκτρια συνάρτηση $I(C(x)=C(y))$ να ισούται με 1 αν οι κλάσεις *C*(*x*) και *C*(*y*) συμπίπτουν. Όταν τα δύο γονίδια *x, y* ανήκουν στην ίδια στατιστική κλάση αναμένουμε να ανήκουν και στην ίδια λειτουργική κλάση. Συνεπώς η επόμενη ποσότητα μετράει το βαθμό στον οποίο οι βιολογικές κλάσεις μοιάζουν με τις στατιστικές:

$$BHI = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in D_j} I(C(x)=C(y)).$$

Στην παραπάνω έκφραση, *k* είναι ο αριθμός των στατιστικών κλάσεων και για την κλάση D_j ο αριθμός των ενεργών γονιδίων συμβολίζεται με $n_j = n(D_j \cap C)$. Αν n_j =1 ή 0 αυτό σημαίνει ότι υπάρχει ένα ή κανένα ενεργό γονίδιο στη στατιστική κλάση D_j και κατά συνέπεια στο δεύτερο άθροισμα της παραπάνω ποσότητας δεν προστίθεται καμία απολύτως τιμή.

Το μέτρο αυτό είναι ιδιαίτερα εύκολο κατά την εφαρμογή του και κατά την ερμηνεία του, αν είναι διαθέσιμες εξ' αρχής οι λειτουργικές κλάσεις. Ουσιαστικά πρόκειται για ένα μέσο όρο των ζευγαριών των γονιδίων των οποίων η βιολογική ομαδοποίηση συμπίπτει με τη στατιστική. Οι τιμές που μπορεί να πάρει κυμαίνονται από 0 ως 1 και όσο πιο κοντά βρίσκονται στο 1 τόσο πιο βιολογικά ομογενείς κλάσεις έχουν κατασκευαστεί.

Β) Βιολογικός Δείκτης Ευστάθειας – BSI

Το μέτρο αυτό προσπαθεί να εκτιμήσει την ευστάθεια ενός αλγορίθμου ομαδοποίησης επιθεωρώντας τη συνέπεια που έχουν τα παραγόμενα βιολογικά αποτελέσματα όταν τα διαθέσιμα δεδομένα μειώνονται κατά μια παρατήρηση. Να σημειώσουμε ότι το μέτρο αυτό δε σχετίζεται με τα υπόλοιπα μέτρα των Datta and Datta (2003) ή το FOM που παρουσιάστηκαν προηγουμένως καθώς εκείνα δε χρησιμοποιούν κανενός είδους βιολογική πληροφορία.

Στα πειράματα μικροσυστοιχιών κάθε γονίδιο έχει ένα επίπεδο έκφρασης το οποίο μπορεί να θεωρηθεί ως μία πολυμεταβλητή παρατήρηση στον Ευκλείδειο χώρο $\Re^p$ όπου $p>1$. Για παράδειγμα p θα μπορούσε να είναι ο αριθμός των χρονικών στιγμών κατά τα οποία έγιναν οι επαναλαμβανόμενες μετρήσεις των επιπέδων έκφρασης. Για κάθε $j=1, 2, \ldots, p$ ο αλγόριθμος ομαδοποίησης επαναλαμβάνεται για όλα τα σύνολα δεδομένων στον $\Re^{p-1}$ ο οποίος λαμβάνεται αν διαγράψουμε κάθε φορά τη μέτρηση-παρατήρηση που βρίσκεται στην i-οστή θέση του διανύσματος που εκφράζει το επίπεδο έκφρασης του κάθε γονιδίου.

Για κάθε γονίδιο g συμβολίζουμε με $D^{g,j}$ την κλάση που περιέχει το γονίδιο g και έχει δημιουργηθεί χωρίς την πληροφορία της παρατήρησης j. Αντιστοίχως με $D^{g,0}$ συμβολίζεται η κλάση που περιέχει το γονίδιο g και για την παραγωγή της έχει χρησιμοποιηθεί ολόκληρο το σύνολο δεδομένων. Για κάθε ζεύγος γονιδίων x και y σε μια βιολογική κλάση συγκρίνουμε τη στατιστική κλάση που περιέχει το γονίδιο x και έχει παραχθεί με βάση τα πλήρη δεδομένα με τη στατιστική κλάση που περιέχει το γονίδιο y και έχει παραχθεί με βάση τα μειούμενα δεδομένα κατά μία παρατήρηση. Ένας καλός αλγόριθμος οφείλει να παράγει παρόμοια αποτελέσματα με τα βιολογικά και όταν χρησιμοποιεί τα πλήρη αλλά και όταν χρησιμοποιεί τα μη πλήρη δεδομένα. Οι κλάσεις που παράγονται από τα πλήρη και τα μη πλήρη δεδομένα περιέχουν δύο λειτουργικά όμοια γονίδια και συνεπώς αναμένεται να παρουσιάζουν σημαντικές αλληλοεπικαλύψεις. Αυτό ακριβώς προσπαθεί να μετρήσει και ο παρακάτω δείκτης

$$BSI = \frac{1}{F}\sum_{i=1}^{F}\frac{1}{n(C_i)(n(C_i)-1)p}\sum_{j=1}^{p}\sum_{x\neq y\in C_i}\frac{n(D^{x,0}\cap D^{y,j})}{n(D^{x,0})}.$$

Να παρατηρήσουμε ότι αν το $n(C_i)$ ισούται με 0 ή 1 συνεπάγεται ότι υπάρχουν λιγότερο από δύο γονίδια που ανήκουν στην ίδια λειτουργική κλάση C_i και συνεπώς για την κλάση αυτή θεωρούμε ότι η συνεισφορά της στο συνολικό άθροισμα ισούται με μηδέν, δηλαδή δε λαμβάνεται καθόλου υπόψην η συγκεκριμένη κλάση στον υπολογισμό της παραπάνω ποσότητας. Το F δηλώνει το συνολικό αριθμό λειτουργικών κλάσεων. Επίσης ο δείκτης παίρνει τιμές από 0 ως 1 και όσο πιο υψηλές είναι αυτές τόσο πιο σταθερές είναι οι κλάσεις που δημιουργούν τα ενεργά γονίδια.

3.5 Ιεραρχική Συνάθροιση (*Rank Aggregation*)

Το ποια μέθοδος είναι η καλύτερη και ποιο από τα μέτρα επικύρωσης φανερώνει την πραγματική δομή των δεδομένων παραμένει ακόμη και σήμερα μια ιδιαίτερα δύσκολη απόφαση παρόλο που πολυάριθμοι αλγόριθμοι ομαδοποίησης έχουν μελετηθεί σε βάθος, έχουν εξελιχθεί ή ακόμα και έχουν οδηγήσει στη δημιουργία νέων. Συγχρόνως, ενώ υπάρχει και μια μεγάλη ποικιλία από μέτρα επικύρωσης, δεν υπάρχει κάποιο βέλτιστο που να καθορίζει με σαφήνεια ποια μέθοδος τελικά και ποια ομαδοποίηση είναι η καλύτερη δυνατή. Συνεπώς είναι απαραίτητο να οριστεί μια τεχνική η οποία να μπορεί να μπορεί να επιλέγει τον καλύτερο αλγόριθμο ανάμεσα σε πολλούς που έχουν εφαρμοστεί σε ένα συγκεκριμένο σύνολο δεδομένων συνδυάζοντας και τα αποτελέσματα που προκύπτουν από την χρήση διαφόρων μέτρων επικύρωσης.

Η ιεραρχική συνάθροιση είναι μια δημοφιλής στατιστική τεχνική αφού πολύ συχνά δημιουργούνται διατεταγμένες λίστες αντικειμένων από τις οποίες πρέπει να επιλεγεί η καλύτερη. Υπάρχουν πολλές διαφορετικές τεχνικές στη θεωρία της ιεραρχικής συνάθροισης δύο από τις πιο σημαντικές παρουσιάζονται επιγραμματικά στη συνέχεια. Η πρώτη βασίζεται στην αρχή της πλειοψηφίας και προσπαθεί να δημιουργήσει την τελική λίστα βασιζόμενη στο που ταξινομείται το κάθε αντικείμενο κατά πλειοψηφία σε σχέση με ένα άλλο. Για παράδειγμα αν το αντικείμενο «Α» ταξινομείται τις περισσότερες φορές πιο ψηλά από το αντικείμενο «Β» τότε το αντικείμενο «Α» οφείλει να ταξινομηθεί υψηλότερα από το «Β» και στη γενική λίστα. Η δεύτερη τεχνική βασίζεται στο μέσο όρο των ταξινομήσεων που υπάρχουν στις αρχικές λίστες και έτσι προσπαθεί να δημιουργήσει την τελική λίστα. Είναι πολύ πιθανόν οι δύο αυτές τεχνικές ιεραρχικής ταξινόμησης να δημιουργήσουν αντιφατικά αποτελέσματα ακόμη και όταν εφαρμόζονται στο ίδιο σύνολο δεδομένων. Εκτός όμως από αυτές τις δύο υπάρχουν και άλλες που είναι μαθηματικά πολύπλοκες και οι οποίες απαιτούν υψηλού επιπέδου υπολογιστικές μεθοδολογίες έτσι ώστε να καταλήξουν στο τελικό αποτέλεσμα. Η μέθοδος της ιεραρχικής συνάθροισης που θα παρουσιαστεί στη συνέχεια βασίζεται στον αλγόριθμο *Cross-Entropy Monte-Carlo* (*Rubinstein*, 1999).

Στόχος της μεθόδου είναι η εύρεση μιας υπέρ-λίστας που θα είναι όσο το δυνατόν πιο κοντά σε όλες τις αρχικώς διατεταγμένες λίστες ταυτοχρόνως. Κατά συνέπεια για να ορίσουμε αυτό το είδος εγγύτητας θα πρέπει να οριστεί μια αντικειμενική συνάρτηση. Έστω δ να είναι μια προτεινόμενη διατεταγμένη λίστα μήκους $k=|L_i|$, w_i το βάρος που αντιστοιχεί στην i-οστή διατεταγμένη λίστα L_i, m ο συνολικός αριθμός από λίστες και d ένα μέτρο απόστασης για το οποίο θα γίνει εκτενέστερη αναφορά παρακάτω. Η αντικειμενική συνάρτηση $\Phi(\delta)$ δίνεται από τον ακόλουθο τύπο:

$$\Phi(\delta) = \sum_{i=1}^{m} w_i d(\delta, L_i).$$

Η βασική ιδέα είναι να βρεθεί εκείνο το δ, που το συμβολίζουμε με δ^*, το οποίο θα ελαχιστοποιεί την απόσταση ανάμεσα στη λίστα δ^* και τη λίστα L_i, δηλαδή

$$\delta^* = \arg\min \sum_{i=1}^{m} w_i d(\delta, L_i).$$

Η επιλογή της απόστασης είναι πολύ κρίσιμη. Οι πλέον συνήθεις επιλογές είναι η απόσταση του *Spearman* (*Spearman Footrule Distance*) και η απόσταση Ταυ του *Kendall* (*Kendall's Tau Distance*). Οι δύο αποστάσεις δημιουργούν ελαφρώς διαφορετικές λίστες των οποίων οι διαφορές εντείνονται ανάλογα και με το πώς έχουν προκύψει οι λίστες που συγκρίνονται μεταξύ τους.

A) Απόσταση *Spearman*

Έστω $L_M = \{ A_1^M, ..., A_k^M \}$ μια διατεταγμένη λίστα που παράγεται από τους k καλύτερους αλγόριθμους και επικυρώνεται από το Μ μέτρο επικύρωσης. Έστω $M_i(1)$, …, $M_i(k)$ τα σκορ που συνδέονται με τις διατεταγμένες λίστες L_i, όπου $M_i(1)$ είναι το καλύτερο σκορ (το μεγαλύτερο ή το μικρότερο ανάλογα με το πώς ορίζεται η έννοια του καλύτερου σε κάθε αλγόριθμο), $M_i(2)$ το δεύτερο καλύτερο κ.ο.κ. Συμβολίζουμε με $r^{L_i}(A)$ το βαθμό του αντικειμένου A στη λίστα L_i. Αν το A ανήκει στις k καλύτερες διατεταγμένες λίστες, ο βαθμός του ισούται με το βαθμό διάταξης του, με το 1 να θεωρείται ως η καλύτερη διάταξη. Αν το A δεν ανήκει στην L_M τότε ο βαθμός του ισούται με $k+1$. Αντίστοιχη είναι και η ερμηνεία του $r^{\delta}(A)$. Η απόσταση *Spearman* ανάμεσα στην L_i και σε οποιαδήποτε άλλη διατεταγμένη λίστα δ υπολογίζεται από τον παρακάτω τύπο, ο οποίος είναι το άθροισμα όλων των απολύτων διαφορών ανάμεσα στους βαθμούς όλων των διατεταγμένων στοιχείων των δύο λιστών που συγκρίνονται

$$S(\delta, L_i) = \sum_{t \in L_i \cup \delta} | r^{\delta}(t) - r^{L_i}(t) |.$$

Όσο πιο μικρή είναι η τιμή της παραπάνω μετρικής τόσο πιο κοντά είναι οι δύο υπό σύγκριση λίστες. Η μέγιστη τιμή που μπορεί να πάρει η απόσταση του *Spearman* όταν συγκρίνονται δύο λίστες με *k* το πλήθος αντικείμενα είναι *k*(*k+1*) και επιτυγχάνεται όταν οι δύο λίστες δεν έχουν κανένα απολύτως κοινό στοιχείο μεταξύ τους. Η απόσταση του *Spearman* είναι δημοφιλής λόγω της απλότητάς της στη χρήση αφού πολλές φορές η μόνη διαθέσιμη πληροφορία που υπάρχει είναι ο βαθμός του κάθε αντικειμένου στις αρχικές λίστες.

Σε περίπτωση βέβαια που είναι διαθέσιμη επιπλέον πληροφορία, όσον αφορά τον τρόπο με τον οποίο κατασκευάστηκαν οι αρχικές λίστες, θα ήταν χρήσιμο να ενσωματωθεί στον τελικό αλγόριθμο της συνάθροισης (*Pihur*, 2007). Αυτό συμβαίνει επειδή πολλές φορές η ποσοτική διαφορά που εμφανίζεται στις λίστες είναι αποτέλεσμα μιας ποιοτικής διαφοράς που έχουν ενσωματωμένη τα δεδομένα. Κάτι τέτοιο γίνεται εύκολα αντιληπτό αν σκεφτούμε ένα ανάλογο παράδειγμα από το χώρο του ποδοσφαίρου όπου όταν κανείς έχει κερδίσει με 5 τέρματα διαφορά είναι πολύ πιο πειστικό ότι είναι καλύτερος από τον αντίπαλό του σε σχέση με κάποιον που έχει κερδίσει στο τελευταίο λεπτό του αγώνα από πέναλτι. Με βάση τις παρατηρήσεις αυτές είναι κατανοητό ότι έχει νόημα το να ορίσει κάποιος τη σταθμισμένη απόσταση του *Spearman (Weighted Spearman's footrule distance)* ανάμεσα στη διατεταγμένη λίστα L_i και σε οποιαδήποτε άλλη δ, μέσω του τύπου

$$WS(\delta, L_i) = \sum_{t \in L_i \cup \delta} | M(r^{\delta}(t)) - M(r^{L_i}(t) \| r^{\delta}(t) - r^{L_i}(t) |.$$

Η παραπάνω σχέση θα μπορούσε να χαρακτηριστεί ως το άθροισμα όλων των ποινών για τη μετακίνηση ενός αντικειμένου *t* από τη θέση $r^{\delta}(t)$ σε μια άλλη θέση $r^{L_i}(t)$ προσαρμοσμένο από τη διαφορά των σκορ ανάμεσα στις δύο θέσεις.

Β) Απόσταση Ταυ του Kendall

Η απόσταση του *Kendall* μετρά διαφορετικά την απόσταση ανάμεσα σε δύο διατεταγμένες λίστες αφού χρησιμοποιεί ζευγάρια αντικειμένων που προκύπτουν από την ένωση των δύο υπό μελέτη λιστών. Πιο συγκεκριμένα δίνεται από τον ακόλουθο τύπο

$$K(\delta, L_i) = \sum_{t,u \in L_i \cup \delta} K_{tu}^{p}$$

όπου:

$$K_{tu}^{p} = \begin{cases} 0, & \text{αν } r^{\delta}(t) < r^{\delta}(u), r^{L_i}(t) < r^{L_i}(u) \quad \text{ή } r^{\delta}(t) > r^{\delta}(u), r^{L_i}(t) > r^{L_i}(u), \\ 1, & \text{αν } r^{\delta}(t) > r^{\delta}(u), r^{L_i}(t) < r^{L_i}(u) \quad \text{ή } r^{\delta}(t) < r^{\delta}(u), r^{L_i}(t) > r^{L_i}(u), \\ p, & \text{αν } r^{\delta}(t) = r^{\delta}(u) = k+1, \quad \text{ή } r^{L_i}(t) = r^{L_i}(u) = k+1. \end{cases}$$

Το p είναι μια παράμετρος που παίρνει τιμές από 0 ως 1 και πρέπει να προσδιοριστεί εξ' αρχής. Αν θέσουμε το p να ισούται με 0 η μέγιστη τιμή που μπορεί να έχει η απόσταση του *Kendall* είναι k^2 και αυτό επιτυγχάνεται όταν οι δύο υπό σύγκριση λίστες δεν έχουν κανένα κοινό στοιχείο μεταξύ τους. Διαισθητικά θα μπορούσαμε να πούμε ότι αν δύο στοιχεία, το t και το u, έχουν την ίδια ακριβώς διάταξη στις δύο υπό μελέτη λίστες τότε δεν τίθεται καμία ποινή στην απόσταση (ιδανικό σενάριο). Αν το αντικείμενο t προηγείται του u στην πρώτη λίστα και έπεται του u στη δεύτερη λίστα τότε προστίθεται ως ποινή η μονάδα στην απόσταση (ανεπιθύμητο σενάριο). Όταν κανένα από τα δύο στοιχεία δεν εμφανίζονται στις λίστες τότε η διάταξη τους θεωρείται ίση με k+1 και η ποινή που τίθεται στην απόσταση ισούται με p. Εφόσον δεν υπάρχει καμία πληροφορία για την πραγματική σχέση διάταξης ανάμεσα στα δύο αντικείμενα το p μπορεί να θεωρηθεί ως αρκετά συντηρητικό παίρνοντας τιμές κοντά στο 1 ή αρκετά φιλελεύθερο παίρνοντας τιμές κοντά 0. Συνήθως επιλέγεται μια μέση λύση και το p τίθεται ίσο με 0,5.

Ανάλογα με τη σταθμισμένη απόσταση του *Spearman* ορίζεται και η *Σταθμισμένη Απόσταση του Kendall* (*Weighted Kendall's Tau*) η οποία δίνεται από τον ακόλουθο τύπο

$$WK(\delta, L_i) = \sum_{t,u \in L_i \cup \delta} |M(r^{L_i}(t)) - M(r^{L_i}(u)| \times K_{tu}^{p}.$$

Είναι φανερό ότι με αυτή την τροποποίηση η ποινή που τίθεται στην απόσταση προσαρμόζεται πλέον από την απόλυτη διαφορά των σκορ που έχουν τα στοιχεία t και u.

Είναι σημαντικό τα σκορ από κάθε λίστα L_i να κανονικοποιηθούν πριν από τον υπολογισμό των αποστάσεων έτσι ώστε τα βάρη να είναι μεταξύ τους συγκρίσιμα. Σε αντίθετη περίπτωση μπορεί να επωφελείται κάποια από τις προτεινόμενες λίστες περισσότερο αν τα βάρη που εφαρμόζονται είναι πολύ μεγάλα ή να μειώνεται η σημασία της αν τα αντίστοιχα βάρη είναι πολύ μικρά. Η

κανονικοποίηση που χρησιμοποιεί ο αλγόριθμος έτσι ώστε τα σκορ να παίρνουν τιμές στο διάστημα από 0 ως 1 είναι η ακόλουθη

$$M_i^* = \frac{M_i - \min(M_i)}{\max(M_i) - \min(M_i)},\ i=1,\ldots,n.$$

Η εισαγωγή των σταθμισμένων αποστάσεων είναι πολύ σημαντική και απαραίτητη διαδικασία. Αρχικά είναι χρήσιμες διότι μπορούν και φανερώνουν τις πραγματικές ταξινομήσεις των αντικειμένων στην κάθε λίστα. Επιπλέον σε πολλά πειράματα έχει παρατηρηθεί ότι χρησιμοποιώντας τις μη σταθμισμένες αποστάσεις των *Spearman* και *Kendall* δεν υπήρχε κάποια λίστα που να υπερείχε των άλλων. Δηλαδή δεν μπορούσε να βρεθεί καθαρός νικητής στην ταξινόμηση. Αυτό συνέβαινε επειδή πολλές από τις διατεταγμένες λίστες είχαν την ίδια αντικειμενική συνάρτηση και συνεπώς κατέληγαν στα ίδια αποτελέσματα. Έτσι ο αλγόριθμος σύγκλισης δε συνέκλινε πουθενά και ήταν αδύνατο να βρεθεί βέλτιστη λίστα. Ωστόσο με τη χρήση των σταθμισμένων αποστάσεων το πρόβλημα αυτό εξαλείφεται και στις περισσότερες περιπτώσεις προκύπτει μια βέλτιστη λύση.

Εφόσον έχει οριστεί η αντικειμενική συνάρτηση και τα μέτρα απόστασης, είναι πλέον εφικτό να βρεθεί μια λύση στο πρόβλημα βελτιστοποίησης και εύρεσης της καλύτερης λίστας δ^*. Για να λυθεί αυτό το υπολογιστικό πρόβλημα χρησιμοποιείται ο αλγόριθμος *cross entropy Monte – Carlo* που αρχικώς προτάθηκε από τον *Rubinstein* το 1997 και χρησιμοποιήθηκε για την εκτίμηση πιθανοτήτων σπάνιων γεγονότων σε σύνθετες στοχαστικές διαδικασίες. Αργότερα όμως εξελίχθηκε έτσι ώστε να είναι εφαρμόσιμος και σε προβλήματα βελτιστοποίησης (*Rubinstein,* 1999), (*Rubinstein,* 2001). Ο ενδιαφερόμενος αναγνώστης μπορεί να ανατρέξει στις αντίστοιχες εργασίες που εισάχθηκε ο αλγόριθμος αλλά και στην εργασία του *Pihur* (2007) όπου γίνεται εκτενής αναφορά στο συγκεκριμένο θέμα.

Κεφάλαιο 4

Εφαρμογή σε Γονιδιακά Δεδομένα

4.1 Εισαγωγή

Στο κεφάλαιο αυτό θα εφαρμοστούν σε πραγματικά δεδομένα οι κυριότερες μέθοδοι ομαδοποίησης και τα περισσότερα από τα μέτρα επικύρωσης που παρουσιάστηκαν θεωρητικά στα προηγούμενα κεφάλαια. Το στατιστικό πακέτο που χρησιμοποιήθηκε για την ανάλυση και την επεξεργασία των δεδομένων είναι η γλώσσα προγραμματισμού R. Τα δεδομένα προέρχονται από ένα πείραμα μικροσυστοιχιών *Affymetrix* το οποίο παρουσιάστηκε και μελετήθηκε αναλυτικά από τους *Bhattacherjee et al.* (2007). Στο πείραμα αυτό πραγματοποιήθηκε σύγκριση των επιπέδων γονιδιακής έκφρασης ανάμεσα σε κύτταρα ποντικών που προέρχονται από δύο διαφορετικούς ιστούς.

Πιο συγκεκριμένα στην παραπάνω εργασία έγινε προσπάθεια μελέτης και σύγκρισης του μοριακού αποτυπώματος που προέρχεται από δύο διαφορετικά κυτταρικά στρώματα ποντικών. Και τα δύο αυτά στρώματα ανήκουν σε εμβρυακά όντα και η συλλογή τους πραγματοποιήθηκε κατά το στάδιο ανάπτυξης του εν λόγω μέρους του σώματος. Είναι σημαντικό να αναφέρουμε ότι τα κύτταρα τα οποία μελετήθηκαν αποτελούν ιστούς που συναντώνται σε όλους τους πολυκύτταρους οργανισμούς και κατά συνέπεια τα αποτελέσματα που προέκυψαν είναι εύλογο ότι θα μπορούσαν να φανούν χρήσιμα και σε άλλους πιο σύνθετους, όπως πχ. ο άνθρωπος.

Η μία ομάδα κυττάρων προέρχεται από το νευρικό σύστημα του εγκεφάλου ενώ η άλλη από το μεσόδερμα[2]. Οι συγκεκριμένες κυτταρικές ομάδες συμβάλλουν στην ανάπτυξη σημαντικών ιστών όπως οι χόνδροι των αρθρώσεων, οι σκελετικοί και μη μύες, τα δόντια κλπ. Από τις δύο περιοχές συλλέχθηκαν 3 δείγματα, συνεπώς το σύνολο των δεδομένων αποτελείται από 147 γονίδια και ESTs με συνολικά 6 μετρήσεις και ο πίνακας δεδομένων έχει διάσταση 147x6. Πιο αναλυτικά η σύγκριση των επιπέδων έκφρασης έδειξε ότι υπάρχουν πάνω από 140 γονίδια που παρουσιάζουν στατιστικά διαφορετικά επίπεδα έκφρασης. Να σημειώσουμε ότι στο παράρτημα δίνεται το πλήρες σύνολο με τα δεδομένα του πειράματος. Τα δείγματα συλλέχθηκαν με τη

[2] Η μεσαία στοιβάδα μιας ομάδας κυττάρων που προέρχονται από την εσωτερική κυτταρική μάζα της βλαστοκύστης και οδηγεί σε ιστό, μύες, συνδετικούς ιστούς, νεφρούς και συναφείς δομές.

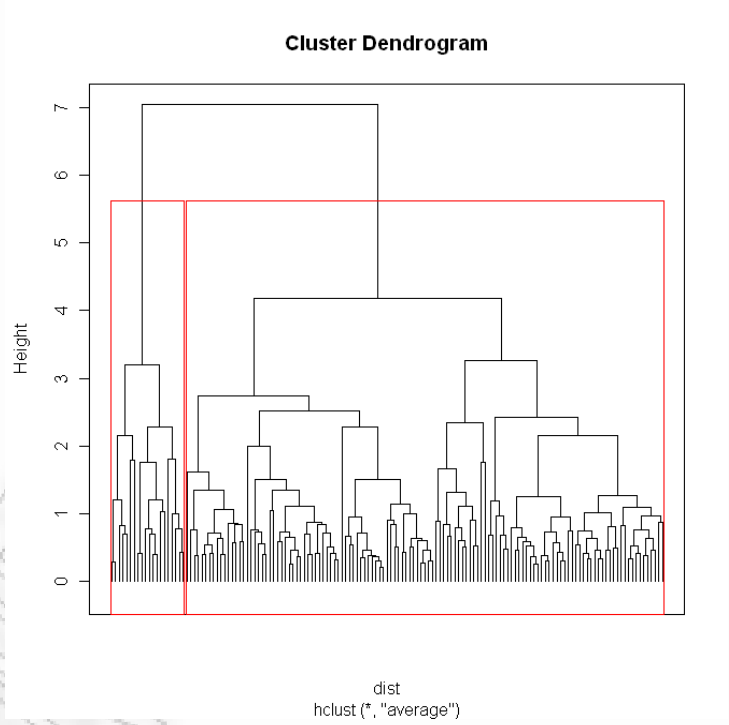
βοήθεια ειδικής τεχνολογίας λέιζερ (*Laser Capture Microdissection-LCM*) και διαχωρίστηκαν μεταξύ τους με χρήση κατάλληλου λογισμικού (*Wnt1Cre/ZEG*). Το RNA που απομονώθηκε από τους δύο πληθυσμούς των LCM κυττάρων βοήθησε στη δημιουργία της διπλής αλυσίδας του cDNA και στη μετέπειτα «*in vitro*» μετάφραση. Ο υβριδισμός των κυττάρων ώστε να παραχθούν τα *gene expression* έγινε με χρήση της τεχνολογίας *GeneChip Microarrays*. Για περισσότερες πληροφορίες καλό θα ήταν ο αναγνώστης να ανατρέξει στη συγκεκριμένη εργασία όπου περιγράφονται με πλήρη λεπτομέρεια οι διαδικασίες του πειράματος.

Τα γονίδια που τελικά ανιχνεύτηκαν διαπιστώθηκε βιολογικώς ότι ανήκουν σε συγκεκριμένες λειτουργικές κλάσεις. Πιο συγκεκριμένα με βάση την ομαδοποίηση που χρησιμοποιήθηκε στην εργασία των *Bhatacherjee et al.* (2006) 16 από αυτά ανήκουν στην κατηγορία ανάπτυξης/διαφοροποίησης (*growth/differentiation*), 7 ανήκουν στις κινάσες/φωσφατάσες (*kinases/phosphatases*), 8 στη μεταβολική ομάδα (*metabolism*), 28 στους μεταγραφικούς παράγοντες (*transcription factors*), 25 προέρχονται από διάφορες κατηγορίες (*miscellaneous*), 16 από την ομάδα που συντελεί στην κυτταρική δομή και στη συνδεσμολογία των ιστών (*ECM/receptors*) και 7 στην κατηγορία *stress-induced*. Ένα γονίδιο που ανήκει στην τελευταία κατηγορία εκφράζεται με την παρουσία μιας εξωτερικής ουσίας (ενός εισβολέα - *inducer*) η οποία μπορεί και ελέγχει το επίπεδο έκφρασής του στο μεταβολισμό αυτής της ουσίας. Για παράδειγμα η λακτόζη (*lactose*) προκαλεί την έκφραση των γονιδίων *lac* τα οποία παίρνουν μέρος στο μεταβολισμό της λακτόζης.

Το σύνολο δεδομένων έχει την τυπική μορφή που συναντάται στα σύνολα με μικροσυστοιχίες, δηλαδή οι γραμμές είναι τα γονίδια (μεταβλητές) και οι στήλες αποτελούν τα δείγματα. Στόχος είναι να δούμε ποια γονίδια εκφράζονται παρόμοια οπότε μπορούν να συνενωθούν και να δημιουργήσουν ομογενείς, συμπαγείς και βιολογικά ορθές ομάδες. Τα γονίδια που μελετήθηκαν θα παρουσιάζουν υψηλό ή χαμηλό επίπεδο έκφρασης σε έναν από τους δύο ιστούς. Συνεπώς είναι λογικό να υποθέσουμε ότι θα δημιουργούνται φυσικά από 2 ως 6 ομάδες (μία πολύ σημαντική και ισχυρή υπόθεση η οποία όμως είναι απαραίτητο να γίνει για την περαιτέρω συνέχεια αυτής της εφαρμογής). Οι μέθοδοι που θα χρησιμοποιηθούν για την ομαδοποίηση των γονιδίων είναι η ιεραρχική μέθοδος μέσης συνένωσης (UPGMA), η μέθοδος k-means, η PAM και η CLARA ενώ ως απόσταση για την παραγωγή του πίνακα αποστάσεων θα χρησιμοποιηθεί η Ευκλείδεια.

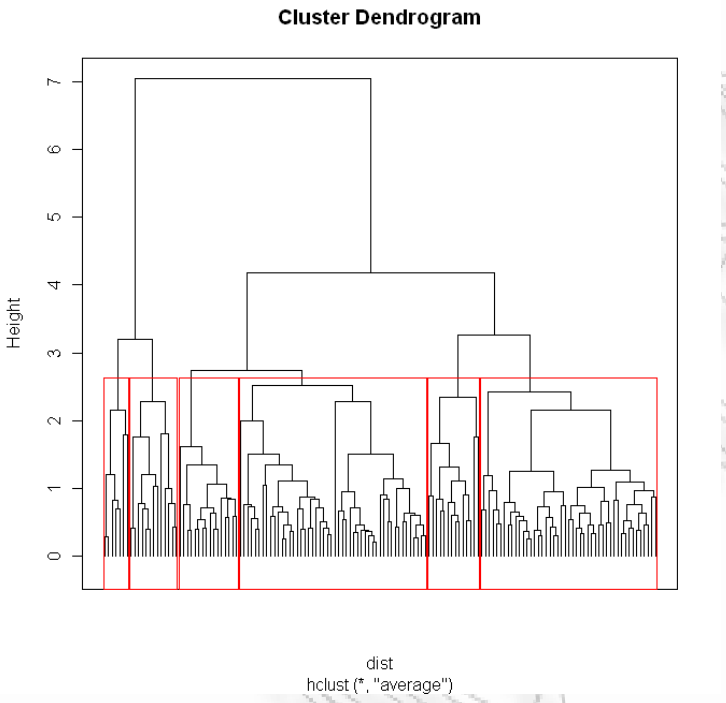
4.2 Μέθοδοι Ομαδοποίησης

Αρχικά θα εφαρμοστεί η μέθοδος της μέσης συνένωσης ή αλλιώς *Unweighted Pair Group Average Method (*UPGMA*)*. Όπως είδαμε και στο Κεφάλαιο 2 όπου παρουσιάστηκε λεπτομερώς η μέθοδος αυτή, ως απόσταση ανάμεσα στις 2 ομάδες θεωρείται ο μέσος όρος των αποστάσεων ανάμεσα σε όλα τα ζεύγη που δημιουργούνται από το σύνολο των δεδομένων. Από την εφαρμογή λοιπόν της συγκεκριμένης μεθόδου προκύπτουν τα δενδρογράμματα των σχημάτων 4.1 και 4.2. Στο σχήμα 4.1 έχουν δημιουργηθεί 2 κόκκινα πλαίσια όπου δείχνουν το διαχωρισμό των δεδομένων σε δύο ομάδες ενώ στο σχήμα 4.2 παρουσιάζεται ο διαχωρισμός που προκύπτει με 6 ομάδες. Ο διαχωρισμός της μεθόδου UPGMA με 2 κλάσεις παράγει 2 ομάδες οι οποίες αποτελούνται από 20 και από 127 γονίδια έκαστη. (Στο Παράρτημα 1 δίνονται αναλυτικά τα γονίδια που ανήκουν στην $1^{η}$ και αυτά που ανήκουν στη $2^{η}$ κλάση).



Σχήμα 4.1: Δενδρόγραμμα με Διαχωρισμό Δύο Κλάσεων

Όταν ο αριθμός των κλάσεων αρχίσει να αυξάνεται σταδιακά παρατηρούμε ότι οι κλάσεις που προκύπτουν είναι σχεδόν ισοπληθείς, γεγονός που γίνεται εμφανές από την ομαδοποίηση με 6 ή και περισσότερες κλάσεις. Πιο συγκεκριμένα για 6 ομάδες οι δημιουργηθείσες κλάσεις περιέχουν 50, 47, 13, 7, 14, 16 γονίδια.



Σχήμα 4.2: Δενδρόγραμμα με Διαχωρισμό Έξι Κλάσεων

Αξίζει να παρατηρήσουμε ότι και πάλι εύκολα διακρίνονται 2 μεγάλες υποομάδες των δεδομένων. Η μία αποτελείται από τις κλάσεις που περιέχουν μεγάλο πλήθος γονιδίων, κοντά στο 50, και η δεύτερη από 3 μικρότερες που περιέχουν περίπου 15 γονίδια η καθεμία. Συνεπώς διαφαίνεται και από αυτόν το διαχωρισμό ότι τα γονίδια τείνουν να διαμοιράζονται σε δύο ομάδες. Είναι συνεπώς πολύ πιθανό τα γονίδια να διαχωρίζονται σε αυτά που παρουσιάζουν πολύ υψηλό επίπεδο έκφρασης και στους δύο ιστούς και σε αυτά με χαμηλό επίπεδο έκφρασης. Ωστόσο με την εφαρμογή μιας και μόνο μεθόδου δε θα ήταν ορθό να επικαλεστούμε ότι προκύπτει κάποιο εμφανές και ασφαλές συμπέρασμα. Αντιθέτως οφείλουμε να επαληθεύσουμε τα αποτελέσματα χρησιμοποιώντας και άλλες μεθόδους.

Στη συνέχεια εφαρμόζεται η μέθοδος k-means για 6 κλάσεις και τα αντίστοιχα πλήθη στοιχείων στις κλάσεις που δημιουργούνται είναι τα εξής: 25, 27, 17, 31, 25, 22. Όπως είναι αναμενόμενο και από τη θεωρία αυτής της μεθόδου οι κλάσεις που προκύπτουν είναι σχεδόν ισοπληθείς. Τα κέντρα τους δίνονται στον Πίνακα 4.1 που ακολουθεί όπου M1 είναι οι μετρήσεις που προέρχονται από τα γονίδια του μεσοδέρματος κατά το 1ο δείγμα, NC1 είναι οι μετρήσεις που προέρχονται από τα γονίδια του νευρικού συστήματος του εγκεφάλου κατά το 1ο δείγμα κ.ο.κ.

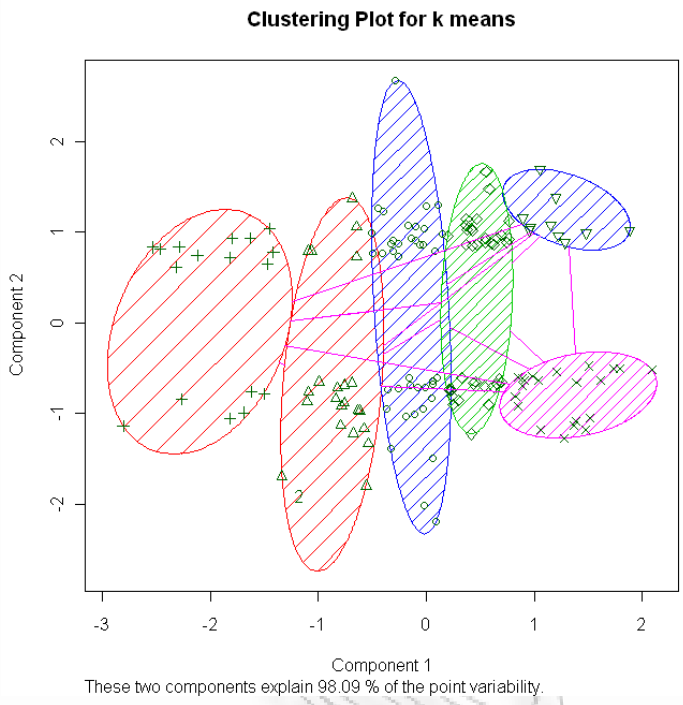
Clusters	M1	M2	M3	NC1	NC2	NC3
1	5.04847	5.04447	5.12336	5.87182	5.93097	5.93381
2	3.65139	3.58811	3.75587	3.26243	3.17971	3.21355
3	7.59714	7.64060	7.56169	7.82268	7.83692	7.83957
4	5.25348	5.22329	5.34966	4.50817	4.48843	4.42449
5	3.96747	4.02434	3.97561	4.63744	4.81829	4.84915
6	6.51655	6.37308	6.62334	5.86994	5.80353	5.70737

Πίνακας 4.1: Μέσοι των Κλάσεων (k-means Μέθοδος 6 Κλάσεων)

Ο λόγος του αθροίσματος των τετραγώνων της μεταβλητότητας ανάμεσα στις κλάσεις προς το συνολικό άθροισμα μεταβλητότητας εντός των κλάσεων είναι ο ακόλουθος και χαρακτηρίζεται ως ιδιαίτερα ικανοποιητικός

$$\frac{SSE(Between\_Clusters)}{SSE(Within\_Clusters)} = 87.7\%$$

Στο Σχήμα 4.3 παρουσιάζεται η δισδιάστατη απεικόνιση των σημείων στις 6 παραγόμενες κλάσεις του παραπάνω αλγόριθμου. Αξίζει να παρατηρήσουμε ότι η δισδιάστατη απεικόνιση γίνεται με τη βοήθεια της θεωρίας της Ανάλυσης Κυρίων Συνιστωσών (***Principal Component Analysis***). Με τη μεθοδολογία των Κυρίων Συνιστωσών από τις αρχικές μεταβλητές δημιουργείται ένας μικρός αριθμός από γραμμικούς συνδυασμούς οι οποίοι είναι ασυσχέτιστοι μεταξύ τους αλλά και ταυτόχρονα περιέχουν τη μέγιστη δυνατή πληροφορία που υπάρχει στις αρχικές μεταβλητές. Έτσι παίρνοντας τις δύο κύριες συνιστώσες μπορούμε να έχουμε ένα αξιόπιστο γράφημα στις δύο διαστάσεις για τα δεδομένα μας. Από το ακόλουθο γράφημα παρατηρούμε ότι οι κλάσεις που έχουν δημιουργηθεί δεν είναι καλώς διαχωρισμένες ούτε ιδιαίτερα συμπαγείς. Το συμπέρασμα αυτό αποτελεί ένδειξη ότι η μέθοδος αυτή ίσως και να μην είναι η πλέον κατάλληλη για την ομαδοποίηση των δεδομένων μας.



Σχήμα 4.3: Δισδιάστατη Απεικόνιση σε Έξι Κλάσεις(k-means)

Στη συνέχεια εφαρμόζεται η μέθοδος PAM για 2 κλάσεις. Τα αρχικά σημεία − γονίδια στα οποία στηρίχτηκε ο αλγόριθμος για να παραχθεί η ομαδοποίηση είναι αυτά που φαίνονται στον Πίνακα 4.2

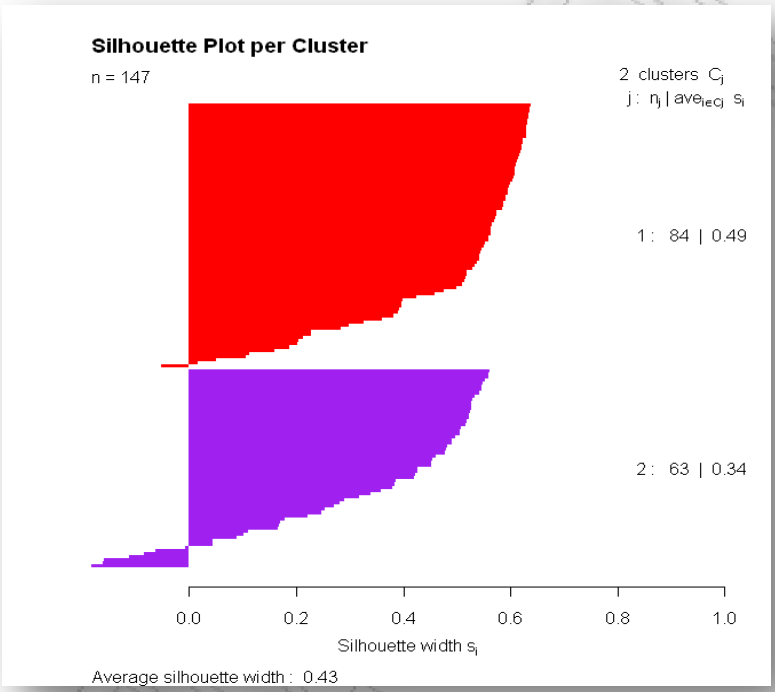
Genes	M1	M2	M3	NC1	NC2	NC3
1451266_at	3.859249	4.184658	4.026063	4.606266	4.484175	4.821662
1426083_a_at	6.358946	6.419124	6.587548	5.591812	5.905448	6.023786

Πίνακας 4.2: Medoids για τον Αλγόριθμο PAM (2 Κλάσεις)

Τα χαρακτηριστικά των κλάσεων που παράγονται δίνονται στον Πίνακα 4.3 ενώ στο Σχήμα 4.4 που ακολουθεί παρουσιάζεται το μέσο *silhouette width*, το οποίο για την πρώτη κλάση λαμβάνει ικανοποιητική τιμή (0.49) ενώ για τη δεύτερη κλάση δε θα μπορούσαμε να ισχυριστούμε κάτι αντίστοιχο αφού η τιμή (0.34) είναι αρκετά μικρή.

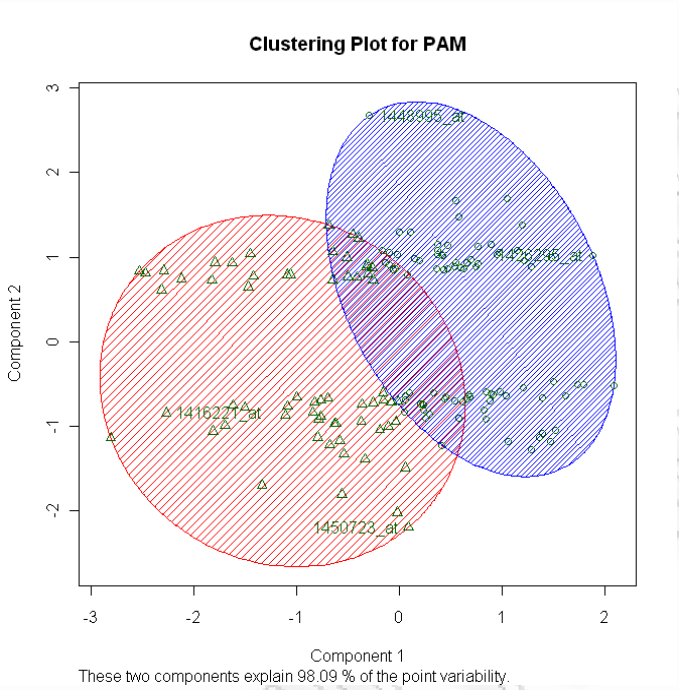
Cluster	Size	Maximum Dissimilarity	Average Dissimilarity	Diameter	Separation
1	84	5.011350	1.966399	8.586815	0.3741614
2	63	6.623064	2.366056	9.575991	0.3741614

Πίνακας 4.3: Χαρακτηριστικά Κλάσεων του Αλγορίθμου PAM (2 Κλάσεις)



Σχήμα 4.4: Μέσο Silhouette Width Παραγόμενων Κλάσεων (PAM - 2 Κλάσεις)

Από το Σχήμα 4.5 που δείχνει τη δισδιάστατη απεικόνιση της συγκεκριμένης ομαδοποίησης παρατηρούμε ότι πολλά στοιχεία ανήκουν και στις δύο κλάσεις, ένδειξη ότι ο διαχωρισμός δεν είναι απόλυτα ορθός. Επιπλέον από τον Πίνακα 4.3 παρατηρούμε ότι η μέγιστη μέση απόσταση (διάμετρος) μέσα στις κλάσεις είναι αρκετά μεγάλη ενώ συγχρόνως η τιμή της διαχωριστικότητας είναι αρκετά μικρή. Λαμβάνοντας όλα τα παραπάνω υπόψην καταλήγουμε ότι ίσως ο διαχωρισμός αυτός να μην είναι ο πλέον κατάλληλος.



Σχήμα 4.5: Δισδιάστατη Απεικόνιση σε Δύο Κλάσεις (PAM)

Αξίζει να αναφερθεί ότι ο δεύτερος καλύτερος διαχωρισμός πραγματοποιείται όταν η μέθοδος εφαρμόζεται για 6 κλάσεις με το μέσο *silhouette width* να ισούται με 0.42. Τα αρχικά σημεία στα οποία στηρίζεται η μέθοδος για να ξεκινήσει το διαχωρισμών των στοιχείων δίνονται στον Πίνακα 4.4.

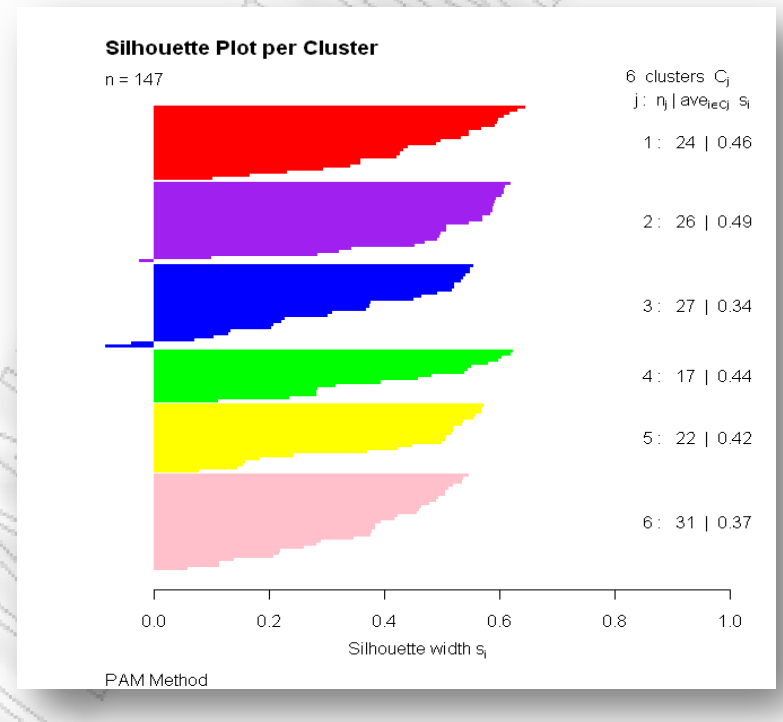
Genes	M1	M2	M3	NC1	NC2	NC3
1415904_at	5.16897	5.22609	5.0634	5.88058	5.87937	5.79193
1450846_at	4.14816	4.03093	4.13241	4.89098	4.69105	4.74963
1427314_at	3.828	3.70583	3.99351	3.27248	3.1151	3.36816
1448269_a_at	7.3231	6.97341	7.41557	7.54015	8.0456	8.0827
1427256_at	6.57374	6.49808	6.44669	6.01597	5.72017	5.60348
1424243_at	5.20181	5.21651	5.2409	4.4468	4.5014	4.59732

Πίνακας 4.4: Medoids για τον Αλγόριθμο PAM (6 Κλάσεις)

Στον Πίνακα 4.5 παρουσιάζονται τα χαρακτηριστικά των κλάσεων που παράγονται. Ενώ στο Σχήμα 4.6 που ακολουθεί παρατηρούμε ότι 4 από τις 6 δημιουργηθείσες κλάσεις παρουσιάζουν μέσο *silhouette width* άνω του 0.40, γεγονός που χαρακτηρίζεται ως ιδιαίτερα καλό.

Clusters	Size	Maximum Dissimilarity	Average Dissimilarity	Diameter	Separation	Silhouette Width
1	24	2.202733	0.8723474	2.648419	0.5421362	0.4599485
2	26	1.983297	0.8686915	3.073029	0.5421362	0.4890489
3	27	2.843144	1.3696168	4.137259	0.5539984	0.3380276
4	17	3.668133	1.7017081	4.845496	1.0661971	0.4411212
5	22	2.260796	1.0805617	3.442686	0.3952844	0.4163668
6	31	2.016246	0.9997256	3.135645	0.3952844	0.3708618

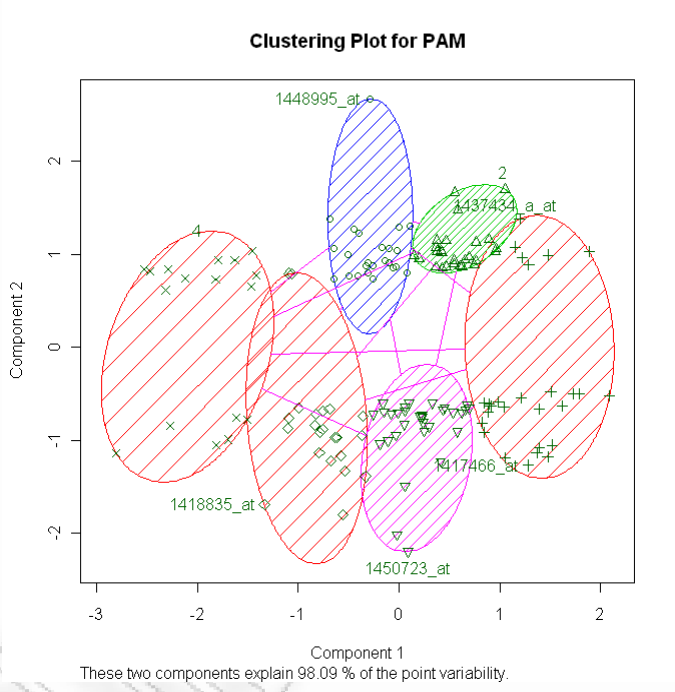
Πίνακας 4.5: Χαρακτηριστικά Κλάσεων του Αλγορίθμου PAM (6 Κλάσεις)



Σχήμα 4.6: Μέσο Silhouette Width Παραγόμενων Κλάσεων (PAM - 6 Κλάσεις)

Από τη δισδιάστατη απεικόνιση παρατηρούμε ότι οι περισσότερες κλάσεις είναι καλά διαχωρισμένες από τις γειτονικές τους, γεγονός το οποίο επαληθεύεται και από τις τιμές της

διαχωριστικότητας οι οποίες έχουν αυξηθεί σημαντικά. Επίσης υπάρχει μια σχετική ομογένεια σε κάθε κλάση αφού και η διάμετρος είναι σημαντικά μικρότερη για κάθε κλάση από ότι ήταν με το διαχωρισμό σε 2 κλάσεις (4.845496 εδώ έναντι 9.575991 που ήταν προηγουμένως οι τιμές για τις μεγαλύτερες διαμέτρους). Το γεγονός αυτό βέβαια ήταν αναμενόμενο αφού όσο μεγαλώνει ο αριθμός των κλάσεων τόσο μεγαλώνει και η ομογένειά τους. Συνεπώς το αν τελικά η μέθοδος αυτή είναι η πλέον κατάλληλη θα κριθεί και από τους δείκτες που θα σχετίζονται με τα διάφορα μέτρα επικύρωσης που παρουσιάστηκαν στο Κεφάλαιο 3. Να σημειώσουμε επίσης ότι ο διαχωρισμός των γονιδίων με βάση αυτήν την ομαδοποίηση παρέχεται στο Παράρτημα 2.



Σχήμα 4.7: Δισδιάστατη Απεικόνιση σε Έξι Κλάσεις (PAM)

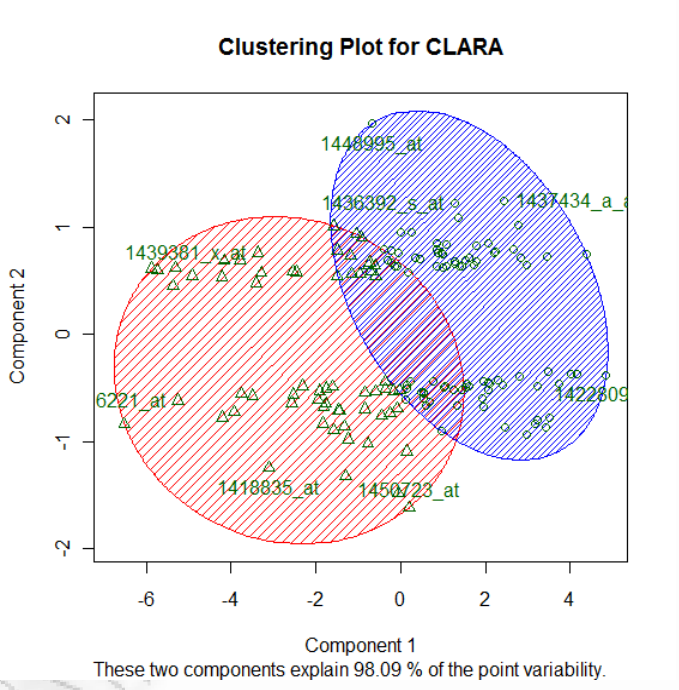
Τελευταία μέθοδος ομαδοποίησης είναι η CLARA η οποία επιτυγχάνει το καλύτερο μέσο *silhouette width* για 2 κλάσεις με αντίστοιχη τιμή 0.5. Δεύτερος καλύτερος ακολουθεί ο διαμερισμός σε 4 κλάσεις με τιμή 0.45 και τρίτος καλύτερος ο διαχωρισμός με 6 κλάσεις και μέσο *silhouette width* 0.44. Τα αρχικά σημεία–γονίδια που στηρίχτηκε η μέθοδος δίνονται στον Πίνακα 4.6 ενώ ορισμένα χαρακτηριστικά των παραγόμενων κλάσεων δίνονται στον Πίνακα 4.7 και στο δισδιάστατο γράφημα που ακολουθεί. Επίσης η τιμή της αντικειμενικής συνάρτησης είναι 2.13768.

Medoids:						
Genes	**M1**	**M2**	**M3**	**NC1**	**NC2**	**NC3**
1451266_at	3.859249	4.184658	4.026063	4.606266	4.484175	4.821662
1426083_a_at	6.358946	6.419124	6.587548	5.591812	5.905448	6.023786

Πίνακας 4.6: Medoids για τον Αλγόριθμο CLARA (2 Κλάσεις)

Clusters	Size	Maximum Dissimilarity	Average Dissimilarity	Isolation	Silhouette Width
1	84	5.01135	1.966399	1.062668	0.5379367
2	63	6.623064	2.366056	1.404436	0.4218881

Πίνακας 4.7: Χαρακτηριστικά Κλάσεων του Αλγορίθμου CLARA (2 Κλάσεις)



Σχήμα 4.8: Δισδιάστατη Απεικόνιση σε Δύο Κλάσεις (CLARA)

Αξίζει να παρατηρήσουμε ότι τα αποτελέσματα που δίνει η μέθοδος βασίζονται στο καλύτερο δείγμα γονιδίων το οποίο επιλέγεται αυτόματα μέσω του αλγορίθμου της μεθόδου. Συνεπώς οι παρακάτω δείκτες θα είναι ελαφρώς καλύτεροι από τους δείκτες που θα προκύψουν όταν χρησιμοποιηθούν όλα τα δεδομένα. Χαρακτηριστικά αναφέρουμε ότι το *Μέσο Silhouette Width* για το καλύτερο δείγμα των δύο κλάσεων ισούται με 0,4957 ενώ στην επόμενη παράγραφο θα παρατηρήσουμε ότι ο δείκτης αυτός πάλι για 2 κλάσεις θα έχει ελαφρώς χαμηλότερη τιμή. Στη

συνέχεια θα γίνει μια προσπάθεια εύρεσης του καλύτερου αλγορίθμου ομαδοποίησης με χρήση των μέτρων επικύρωσης.

4.3 Εσωτερικά Μέτρα Επικύρωσης

Όπως αναφέρθηκε και προηγουμένως στην παράγραφο αυτή θα χρησιμοποιήσουμε μερικά από τα μέτρα επικύρωσης που παρουσιάστηκαν στο προηγούμενο κεφάλαιο για να αποφασίσουμε ποια από τις μεθόδους παράγει την καλύτερη ομαδοποίηση. Αρχικά θα ασχοληθούμε με τα εσωτερικά μέτρα επικύρωσης. Πιο συγκεκριμένα για κάθε μία από τις 4 μεθόδους ομαδοποίησης θα υπολογιστούν οι τιμές της συνδετικότητας, του δείκτη *Dunn* και του *Silhouette Width*. Να υπενθυμίσουμε ότι όσο μεγαλύτερες είναι οι τιμές των δύο τελευταίων δεικτών τόσο καλύτερη είναι η ομαδοποίηση. Το αντίθετο ισχύει για τη συνδετικότητα. Επίσης να επισημάνουμε ότι οι τιμές του *Silhouette Width* για τη μέθοδο CLARA παράγονται χρησιμοποιώντας τη διαθέσιμη πληροφορία από όλα τα σύνολα των δεδομένων που έχει δημιουργήσει ο αλγόριθμος και όχι μόνο από το καλύτερο δείγμα (γεγονός που συνέβαινε στους υπολογισμούς της προηγούμενης παραγράφου) και το πλήθος των υποσυνόλων των δεδομένων στα οποία δουλεύει ισούται με 5 (k=5). Τα αποτελέσματα δίνονται στον Πίνακα 4.8:

Number of Clusters		**2**	**3**	**4**	**5**	**6**
UPGMA	**Connectivity**	5.327	14.2528	20.752	27.0726	30.6194
	Dunn	0.1291	0.0788	0.0857	0.0899	0.0899
	Silhouette	0.5133	0.4195	0.37	0.3343	0.3233
k-means	**Connectivity**	13.2548	17.6651	37.398	43.2655	50.6095
	Dunn	0.0464	0.0873	0.0777	0.0815	0.0703
	Silhouette	0.4571	0.4182	0.3615	0.3367	0.3207
PAM	**Connectivity**	18.7917	27.9651	30.9302	44.9671	32.9667
	Dunn	0.0391	0.0597	0.051	0.0761	0.0816
	Silhouette	0.4271	0.3489	0.3563	0.353	0.4152
CLARA	**Connectivity**	18.7028	27.9651	44.8234	35.5159	26.1238
	Dunn	0.0287	0.0597	0.066	0.0761	0.0857
	Silhouette	0.4257	0.3489	0.3304	0.3636	0.3836

Πίνακας 4.8: Εσωτερικά Μέτρα Επικύρωσης

Αυτό που παρατηρούμε είναι ότι οι καλύτερες τιμές για τα παραπάνω μέτρα επικύρωσης επιτυγχάνονται με την εφαρμογή της ιεραρχικής μεθόδου UPGMA με 2 κλάσεις. Πιο συγκεκριμένα οι δείκτες *Connectivity, Dunn* και *Silhouette* έχουν τις τιμές που δίνονται στον Πίνακα 4.9:

Measure	Score	Method	Clusters
Connectivity	5.327	UPGMA	2
Dunn	0.1291	UPGMA	2
Silhouette	0.5133	UPGMA	2

Πίνακας 4.9: Καλύτερες Τιμές των Εσωτερικών Μέτρων Επικύρωσης

Η τιμή του *Silhouette Width* θα μπορούσε να χαρακτηριστεί ως ικανοποιητική αφού ξεπερνά το 0.5 ωστόσο δε μπορεί να θεωρηθεί ως ιδανική αφού η τιμή δεν πλησιάζει τη μονάδα. Η τιμή του δείκτη *Dunn* και η τιμή της συνδετικότητας είναι αρκετά καλές.

4.4 Εσωτερικά μέτρα επικύρωσης των *Datta&Datta* και το μέτρο *Figure of Merit*

Στο Κεφάλαιο 3 περιγράφηκαν αναλυτικά ορισμένα μέτρα επικύρωσης των μεθόδων παραγωγής συστάδων. Στη συγκεκριμένη παράγραφο θα υπολογιστούν ορισμένα από αυτά τα μέτρα έτσι ώστε να καταλήξουμε με ασφάλεια στην εύρεση του καλύτερου αλγορίθμου. Πιο συγκεκριμένα θα υπολογιστούν για τις 4 διαφορετικές μεθόδους ομαδοποίησης οι τιμές των εσωτερικών μέτρων επικύρωσης των *Datta&Datta* και η τιμή του *Figure of Merit*. Στον Πίνακα 4.10 δίνονται οι τιμές των 4 εσωτερικών μέτρων επικύρωσης:

Number of Clusters		2	3	4	5	6
UPGMA	APN	0.0478	0.1288	0.1755	0.1689	0.1516
	AD	3.243	2.6814	2.2571	2.0642	1.8732
	ADM	0.4283	1.0953	0.807	0.6196	0.5867
	FOM	1.0658	0.8678	0.7451	0.6823	0.6371
k-means	APN	0.0603	0.0726	0.3146	0.2485	0.247
	AD	2.9001	2.2923	2.2529	1.9978	1.8389
	ADM	0.3196	0.3101	1.0621	0.7151	0.67
	FOM	0.9745	0.7548	0.7114	0.6528	0.6074

PAM	APN	0.1318	0.2376	0.3658	0.3029	0.0486
	AD	3.0382	2.5993	2.4492	2.084	1.5272
	ADM	0.6372	0.9733	1.3172	1.0164	0.1401
	FOM	1.0092	0.8391	0.7663	0.649	0.5158
CLARA	APN	0.1099	0.2199	0.2798	0.3108	0.3061
	AD	2.9902	2.5945	2.3069	2.1053	1.9024
	ADM	0.4907	0.9201	0.9264	1.0816	1.0271
	FOM	1.0103	0.8251	0.6923	0.6671	0.5239

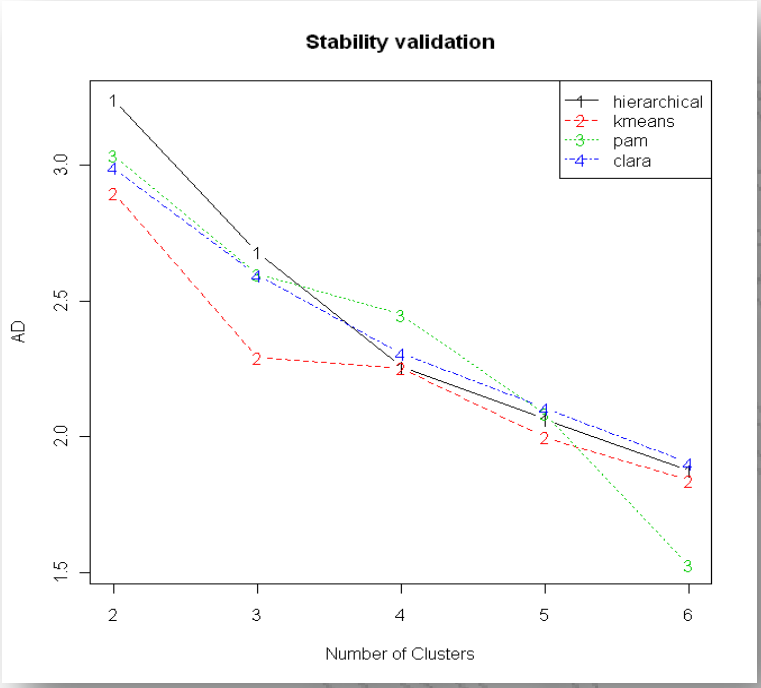
Πίνακας 4.10: Μέτρα Ευστάθειας

Να υπενθυμίσουμε ότι όσο πιο χαμηλές είναι η τιμές των παραπάνω δεικτών τόσο καλύτερη η μέθοδος ομαδοποίησης. Οι καλύτερες τιμές για 3 από τα 4 μέτρα επιτυγχάνονται με τη μέθοδο PAM με διαχωρισμό 6 κλάσεων ενώ το μέτρο APN δίνει ως καλύτερη μέθοδο τη μέθοδο UPGMA με 2 κλάσεις. Αξίζει να παρατηρήσουμε ότι το μέτρο APN δίνει πολύ χαμηλή τιμή και στη μέθοδο PAM με 6 κλάσεις. Πιο συγκεκριμένα η μέθοδος UPGMA με 2 κλάσεις και η PAM με 6 παρουσιάζουν APN με αντίστοιχες τιμές 0.0478 και 0.0486 κατά συνέπεια θα μπορούσαμε να πούμε ότι σε όλα τα μέτρα ευστάθειας ως καλύτερη μέθοδος θεωρείται η PAM με 6 κλάσεις.

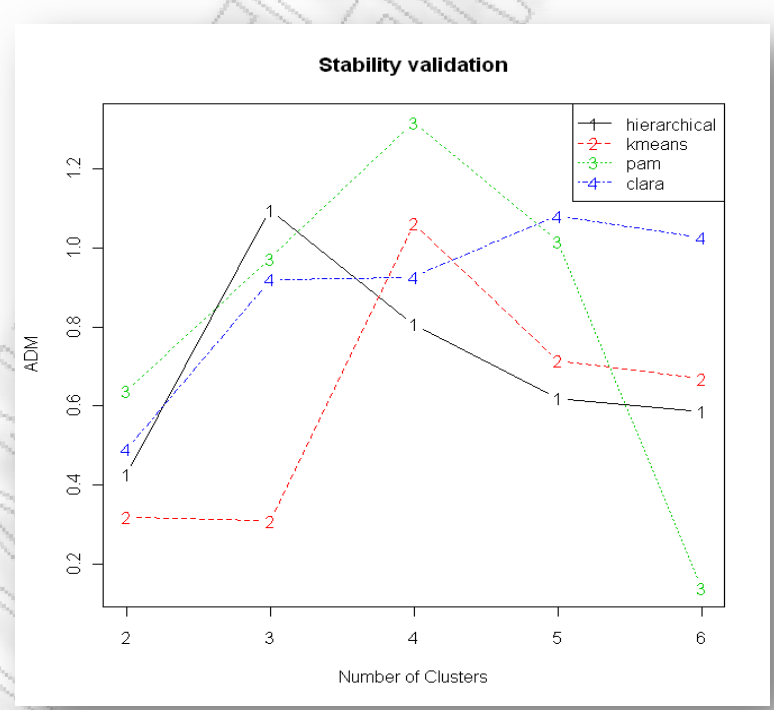
Measure	Score	Method	Clusters
APN	0.0478	UPGMA	2
AD	1.5272	PAM	6
ADM	0.1401	PAM	6
FOM	0.5158	PAM	6

Πίνακας 4.11: Καλύτερες Τιμές των Μέτρων Ευστάθειας

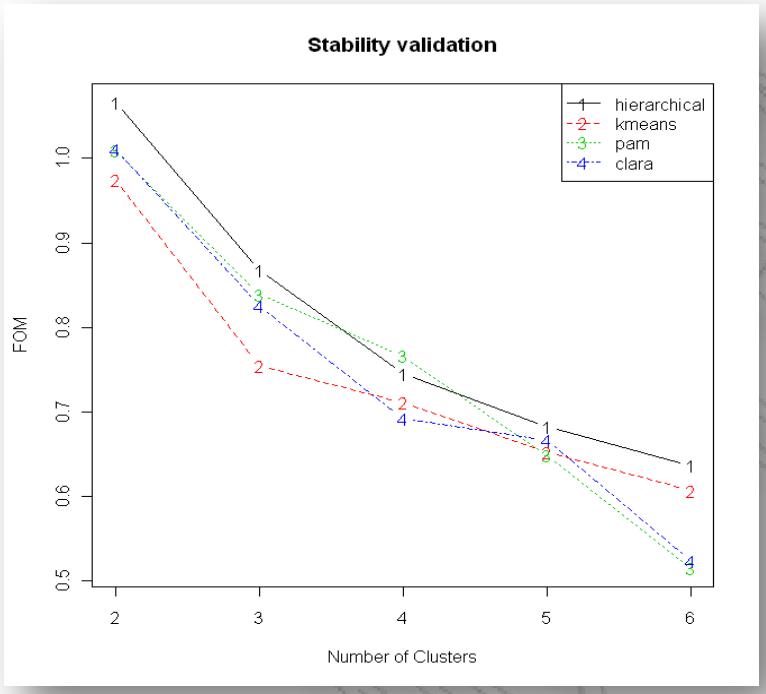
Στη συνέχεια δίνονται οι γραφικές παραστάσεις για κάθε ένα από τα μέτρα επικύρωσης. Όπως γίνεται αντιληπτό στα σχήματα 4.9 και 4.10 για τα μέτρα AD και ADM αντίστοιχα η υπεροχή του PAM είναι εμφανής. Από το σχήμα 4.11 παρατηρούμε ότι το μέτρο FOM δίνει και πολύ χαμηλή τιμή και στη μέθοδο CLARA με 6 κλάσεις ενώ το σχήμα 4.12 επικυρώνει την εκτίμηση μας ότι παρόλο που η μέθοδος UPGMA με 2 κλάσεις έχει την καλύτερη τιμή για το μέτρο APN και η μέθοδος PAM με 6 κλάσεις είναι με διαφορά από τις υπόλοιπες η επόμενη καλύτερη.



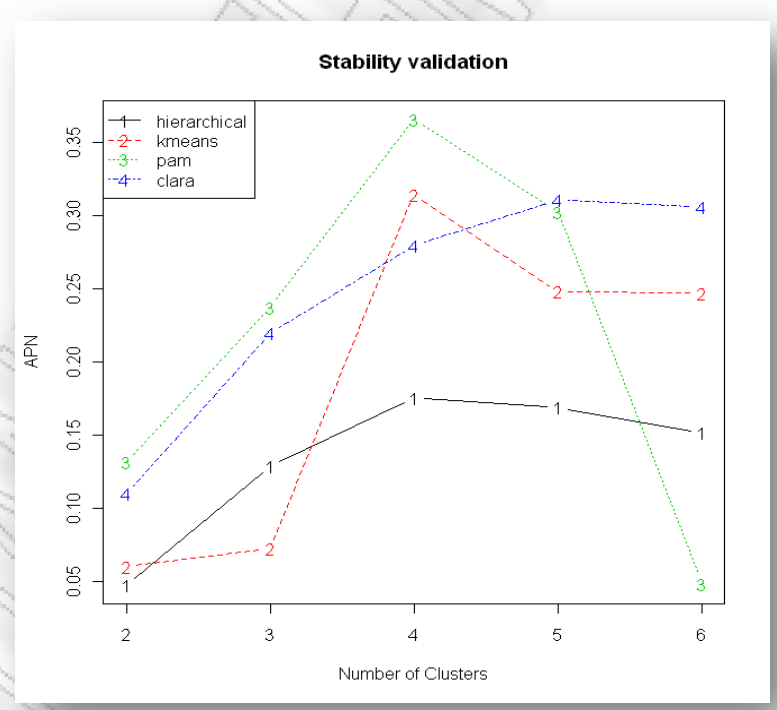
Σχήμα 4.9: Μέτρο Ευστάθειας - AD



Σχήμα 4.10: Μέτρο Ευστάθειας - ADM



Σχήμα 4.11: Μέτρο Ευστάθειας – FOM



Σχήμα 4.12: Μέτρο Ευστάθειας - APN

4.5 Βιολογικά Μέτρα Επικύρωσης

Για τη βιολογική επικύρωση θα πρέπει να βρεθούν οι τιμές των δεικτών *BHI* και *BSI*. Για τον υπολογισμό αυτών των δεικτών υπάρχουν δύο επιλογές: για την πραγματοποίηση της πρώτης απαιτείται η χρήση εξειδικευμένων στατιστικών πακέτων που περιέχουν πληροφορίες για την κατηγοριοποίηση των γονιδίων σε διάφορες λειτουργικές κλάσεις. Κάποια χαρακτηριστικά πακέτα είναι το ***FatiGO*** (*Al-Shahour et al*., 2004) και το **FunCat** (*Ruepp et al.*, 2004). Ο δεύτερος τρόπος με τον οποίο μπορεί να γίνει η παραγωγή των παραπάνω δεικτών είναι με τη χρήση του στατιστικού λογισμικού **Bioconductor** (*Gentleman et al.*, 2004). Το πακέτο που είναι διαθέσιμο πλέον μέσω αυτού του στατιστικού λογισμικού χρησιμοποιεί την κατηγοριοποίηση των γονιδίων που παρέχεται μέσω της βάσης GO (*Gene Ontology*). Το λογισμικό αυτό δίνει τη δυνατότητα χρήσης εξειδικευμένων εργαλείων για την ανάλυση και επεξεργασία γονιδιακών δεδομένων τα οποία όμως για να λειτουργήσουν απαιτούν το λειτουργικό περιβάλλον της R. Το Bioconductor είναι ελεύθερα διαθέσιμο στο κοινό, παρέχει περισσότερα από 460 πακέτα και η επίσημη ιστοσελίδα του είναι η εξής: http://www.bioconductor.org. Ωστόσο για τα δεδομένα της συγκεκριμένης εργασίας η κατηγοριοποίηση είναι άμεσα διαθέσιμη μέσω R αφού το συγκεκριμένο σύνολο δεδομένων είναι το ίδιο το οποίο μελετήθηκε βιολογικώς στο πείραμα του *Bhattacherjee* (2007) και συνεπώς δε θα χρειαστεί να ανατρέξουμε σε άλλες βάσεις. Τα αποτελέσματα παρουσιάζονται στους Πίνακες 4.12 και 4.13 και στα Σχήματα 4.13 και 4.14.

Number of Clusters		**2**	**3**	**4**	**5**	**6**
UPGMA	BHI	0.172	0.1788	0.163	0.1866	0.1778
	BSI	0.6756	0.4635	0.3053	0.2973	0.2506
k-means	BHI	0.1721	0.1826	0.1899	0.2026	0.1898
	BSI	0.5173	0.3492	0.2653	0.2267	0.1938
PAM	BHI	0.1763	0.182	0.1888	0.1877	0.1911
	BSI	0.5158	0.3496	0.2632	0.2055	0.1727
CLARA	BHI	0.1804	0.182	0.1825	0.1959	0.1751
	BSI	0.514	0.3496	0.2536	0.21	0.1792

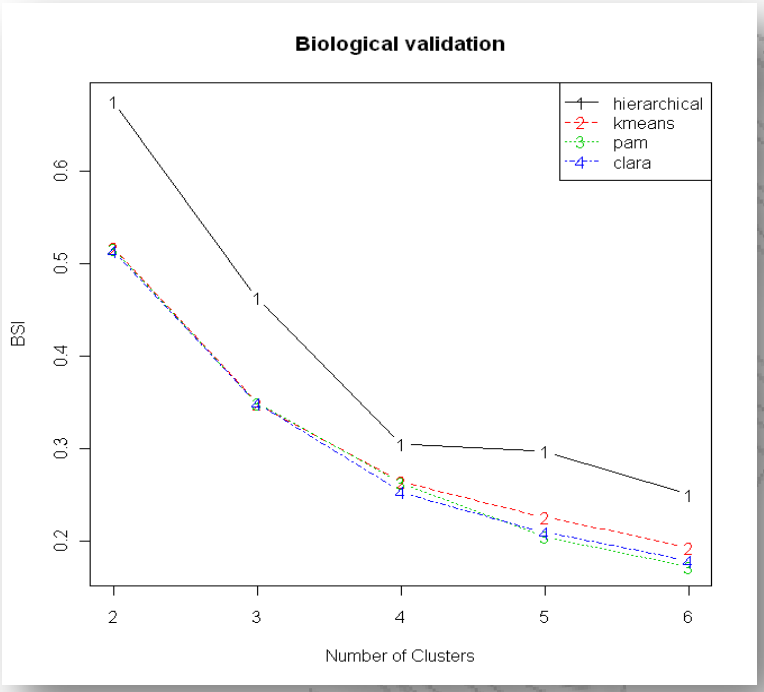
Πίνακας 4.12: Βιολογικά Μέτρα Επικύρωσης

Measure	Score	Method	Clusters
BHI	0.2026	k-means	5
BSI	0.6756	UPGMA	2

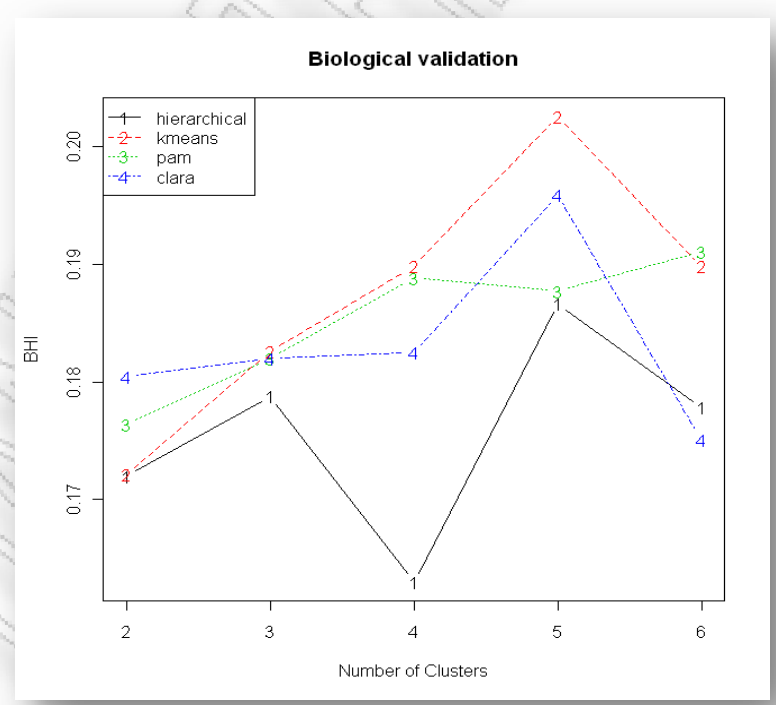
Πίνακας 4.13: Καλύτερες Τιμές των Βιολογικών Μέτρων Επικύρωσης

Ο δείκτης BHI παρουσιάζει την καλύτερη τιμή του για τη μέθοδο k-means με 5 κλάσεις. Ωστόσο οι τιμές του είναι σχετικά μεγάλες και για τις μεθόδους PAM και k-means με 6 κλάσεις κάτι που έρχεται σε συμφωνία και με τα αποτελέσματα των μέτρων ευστάθειας. Οι τιμές του δείκτη BSI κρίνονται σχεδόν για όλες τις μεθόδους ιδιαίτερα ικανοποιητικές αφού η μικρότερη τιμής τους είναι το 0.20 με τις καλύτερες να εμφανίζονται στις ομαδοποιήσεις με 2 κλάσεις. Κατά συνέπεια οι κλάσεις που έχουν δημιουργηθεί μπορούν να χαρακτηριστούν ως βιολογικά σταθερές αλλά όχι βιολογικά ομογενείς. Αυτό ήταν αναμενόμενο αφού βιολογικώς τα γονίδια γνωρίζουμε εκ των προτέρων ότι διαχωρίζονται σε 7 κλάσεις. Παρόλα αυτά στη συγκεκριμένη εφαρμογή αναζητείται κάποιος άλλος υποβόσκων διαχωρισμός των γονιδίων, που βασίζεται στη τιμή του *gene expression*, και πιθανόν να είναι χρήσιμος για περαιτέρω μελέτη και επεξεργασία.

Στα Σχήματα 4.14 και 4.15 που ακολουθούν παρατηρούμε ότι όσο αυξάνεται ο αριθμός των κλάσεων τόσο μικραίνει ο δείκτης BSI ενώ καλύτερη τιμή για διαχωρισμούς από 2 ως 6 κλάσεις παρουσιάζει η ιεραρχική μέθοδος UPGMA. Αντιθέτως όσον αφορά το δείκτη BHI, καλύτερη τιμή παρουσιάζει η μέθοδος k-means ενώ η ιεραρχική μια έχει από τις χαμηλότερες τιμές.



Σχήμα 4.13: Βιολογικός Δείκτης Ευστάθειας – BSI



Σχήμα 4.14: Βιολογικός Δείκτης Ομογένειας - BHI

4.6 Ιεραρχική Συνάθροιση

Από την παραπάνω ανάλυση γίνεται αντιληπτό ότι τα διάφορα μέτρα επικύρωσης δεν καταλήγουν σε κάποιο καθολικό συμπέρασμα. Το αντίθετο, δηλαδή η πλήρης συμφωνία, θα ήταν ίσως ανησυχητική για την εγκυρότητα των δεδομένων. Ωστόσο είναι εμφανές πως οι μέθοδοι που ξεχωρίζουν και παρουσιάζουν τα καλύτερα αποτελέσματα είναι η UPGMA με 2 κλάσεις και η PAM με 6 κλάσεις. Παρόλα αυτά για να είμαστε σε θέση να καταλήξουμε σε κάποιο συμπέρασμα, έτσι ώστε να βρεθεί η καλύτερη δυνατή μέθοδος, είναι απαραίτητη η χρήση της ιεραρχικής συνάθροισης. Η τελευταία θα μας δώσει την καλύτερη δυνατή λίστα που θα λαμβάνει υπόψη της τα αποτελέσματα όλων των μέτρων επικύρωσης που παράγονται από την εφαρμογή όλων των μεθόδων. Χρησιμοποιώντας το αντίστοιχο πακέτο στην R για την εύρεση της βέλτιστης λίστας για όλες τις μεθόδους με αριθμό κλάσεων που κυμαίνεται από 2 ως 6 έχουμε τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4.14:

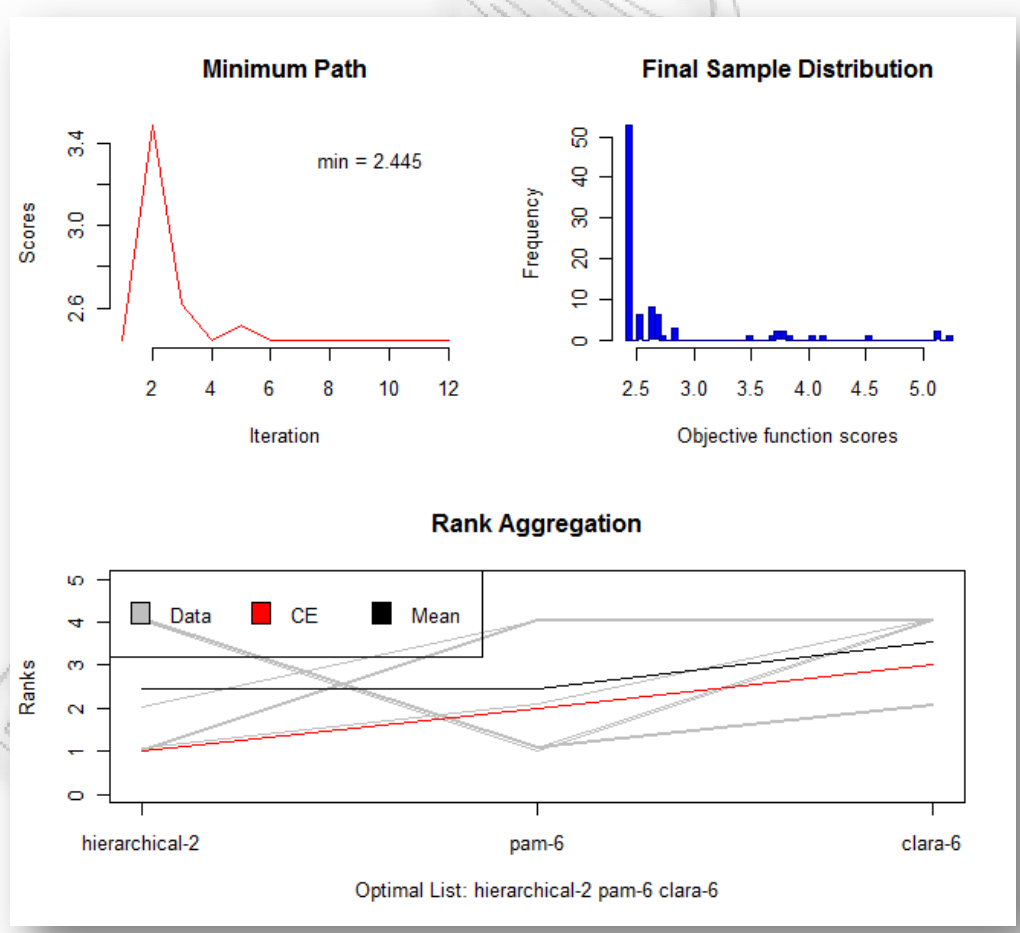
Measures/List	1	2	3
APN	hierarchical-2	pam-6	kmeans-2
AD	pam-6	kmeans-6	hierarchical-6
ADM	pam-6	kmeans-3	kmeans-2
FOM	pam-6	clara-6	kmeans-6
Connectivity	hierarchical-2	kmeans-2	hierarchical-3
Dunn	hierarchical-2	hierarchical-5	hierarchical-6
Silhouette	hierarchical-2	kmeans-2	pam-2
BHI	hierarchical-4	hierarchical-2	kmeans-2
BSI	pam-6	clara-6	kmeans-6

Πίνακας 4.14: Βέλτιστη Λίστα Ιεραρχικής Συνάθροισης (2-6 Κλάσεις)

Η καλύτερη λίστα που προκύπτει είναι η ακόλουθη: UPGMA–2 PAM–6 CLARA–6. Συνεπώς καταλήγουμε πως ως καλύτερη μέθοδος για τα περισσότερα μέτρα επικύρωσης κρίνεται η ιεραρχική μέθοδος με 2 ομάδες ενώ ακολουθεί η PAM με 6 κλάσεις. Ορισμένα χαρακτηριστικά του αλγορίθμου δίνονται στον Πίνακα 4.15 ενώ στο Σχήμα 4.16 δίνεται το πλήθος των επαναλήψεων μέχρι να επέλθει η σύγκλιση του αλγορίθμου το οποίο ισούται με 6, οι τιμές που λαμβάνει η αντικειμενική συνάρτηση και το πως ταξινομούνται οι διάφορες μέθοδοι κατά μέσο όρο.

Algorithm	**CE**
Distance	Spearman
Score	2.445375

Πίνακας 4.15: Χαρακτηριστικά Αλγορίθμου (2-6 Κλάσεις)



Σχήμα 4.15: Γραφήματα του Αλγορίθμου της Ιεραρχικής Συνάθροισης (2-6 Κλάσεις)

Στη συνέχεια ακολουθεί παρόμοια ανάλυση με την προηγούμενη με μοναδική διαφορά ότι έχουμε θέσει ως αρχική συνθήκη για τον αλγόριθμο τη δημιουργία 3 ως 6 κλάσεων και όχι 2 ως 6. Αξίζει να παρατηρήσουμε ότι πλέον ως καλύτερος διαχωρισμός επιλέγεται η μέθοδος PAM με 6 κλάσεις ενώ ακολουθεί η k-means με 3 και η UPGMA με 3.

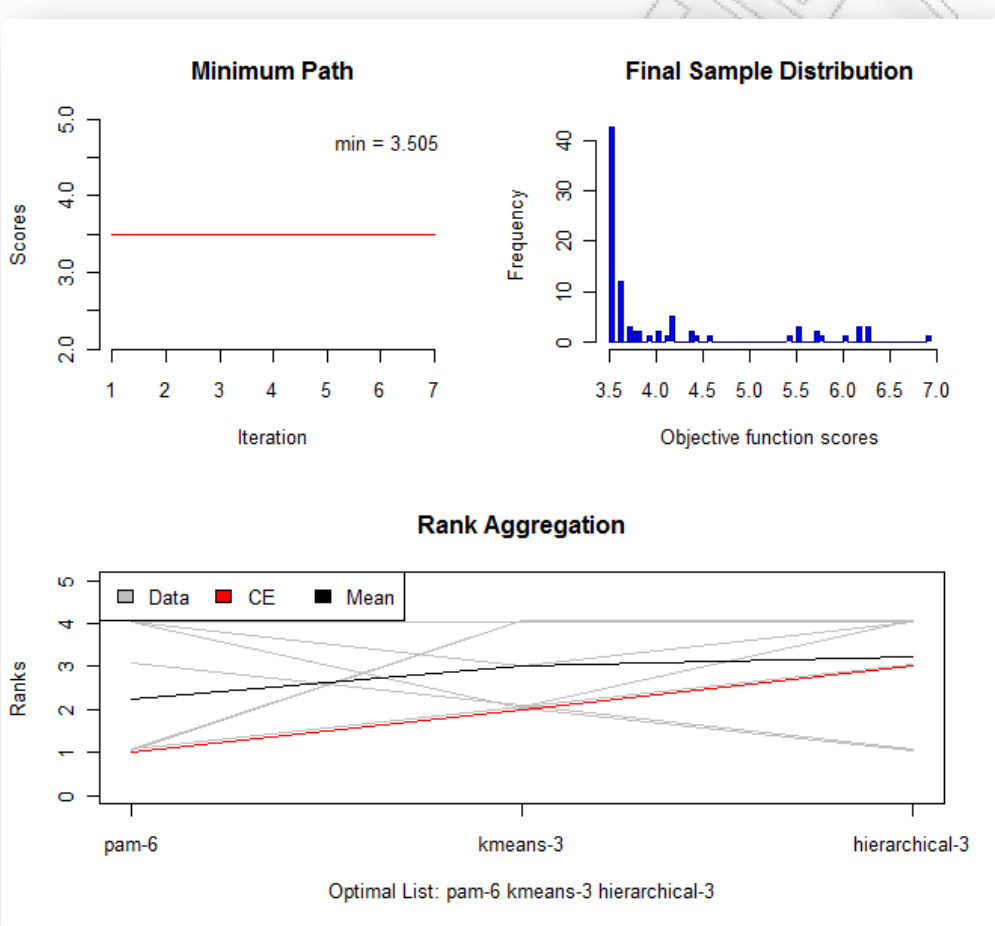
Measures/List	1	2	3
APN	pam-6	kmeans-3	hierarchical-3
AD	pam-6	kmeans-6	hierarchical-6
ADM	pam-6	kmeans-3	hierarchical-6
FOM	pam-6	clara-6	kmeans-6
Connectivity	hierarchical-3	kmeans-3	hierarchical-4
Dunn	hierarchical-5	hierarchical-6	kmeans-3
Silhouette	hierarchical-3	kmeans-3	pam-6
BHI	hierarchical-4	clara-6	hierarchical-6
BSI	pam-6	clara-6	kmeans-6

Πίνακας 4.16: Βέλτιστη Λίστα Ιεραρχικής Συνάθροισης (3-6 Κλάσεις)

Η τιμή της αντικειμενικής συνάρτησης αυξάνεται και ισούται με 3.504908 όπως παρατηρούμε από τα αποτελέσματα του Πίνακα 4.17 ενώ το Σχήμα 4.17 μας πληροφορεί πως ο αλγόριθμος συγκλίνει από την πρώτη εφαρμογή του και όχι στις 6 επαναλήψεις όπως προηγουμένως. Ωστόσο αξίζει να παρατηρήσουμε ότι από το δεύτερο γράφημα του ίδιου Σχήματος η διασπορά των τιμών της αντικειμενικής συνάρτησης είναι μεγαλύτερη από πριν με τιμές που κυμαίνονται από 3.5 ως και 7 ενώ με τη χρήση του προηγούμενου αλγορίθμου οι τιμές αυτές εκτείνονται από 2.5 ως 5.5 με τη μεγαλύτερη συγκέντρωση να παρατηρείται στις χαμηλότερες τιμές (βλ. Σχήμα 4.16).

Algorithm	CE
Distance	Spearman
Score	3.504908

Πίνακας 4.17: Χαρακτηριστικά Αλγορίθμου (3-6 Κλάσεις)



Σχήμα 4.16: Γραφήματα του Αλγορίθμου της Ιεραρχικής Συνάθροισης (3-6 Κλάσεις)

4.7 Τελικά Συμπεράσματα

Στην εφαρμογή που παρουσιάστηκε πραγματοποιήθηκε η στατιστική ανάλυση γονιδιακών δεδομένων που προέρχονται από δύο διαφορετικούς ιστούς του προσώπου ποντικών λίγων ημερών. Στόχος ήταν η συσταδοποίηση αυτών των γονιδίων σε ομάδες έτσι ώστε να βρεθεί αν

υπάρχουν γονίδια στους δύο διαφορετικούς ιστούς που παρουσιάζουν παρόμοια επίπεδα έκφρασης. Να σημειώσουμε για ακόμη μια φορά ότι τα δεδομένα τα οποία χρησιμοποιήθηκαν στα πλαίσια δημοσιευμένων μελετών έχουν παρουσιάσει στατιστικά διαφορετικά επίπεδα έκφρασης. Από τα αποτελέσματα έγινε αντιληπτό πως καλύτερη μέθοδος ομαδοποίησης είναι η Ιεραρχική Μέθοδος Ομαδοποίησης Μέσης Απόστασης με 2 κλάσεις. Από τους πίνακες του Παραρτήματος 1 παρατηρούμε ότι η μέθοδος ουσιαστικά μπορεί και διαχωρίζει τον πληθυσμό σε γονίδια με ιδιαίτερα υψηλό επίπεδο έκφρασης (6.60 ως 8.85) και σε γονίδια με χαμηλό δείκτη έκφρασης (2.38- 6.599).

Η μέθοδος PAM, η οποία είναι η αμέσως καλύτερη, έχει την ικανότητα όχι μόνο να διαχωρίζει τα γονίδια με βάση τα πολύ υψηλά ή πολύ χαμηλά επίπεδα έκφρασής τους αλλά συγχρόνως κατατάσσει τα επίπεδα έκφρασής τους με βάση τον ιστό στον οποίο βρίσκονται. Δηλαδή στο Παράρτημα 2 όσα γονίδια έχουν χρωματιστεί καφέ σημαίνει ότι παρουσιάζουν μικρό επίπεδο έκφρασης και στους δύο ιστούς ενώ όσα είναι ροζ παρουσιάζουν υψηλό επίπεδο έκφρασης επίσης και στους δύο ιστούς. Επιπλέον όσα είναι γαλάζια παρουσιάζουν γενικά χαμηλό επίπεδο έκφρασης αλλά έχουν μεγαλύτερο επίπεδο έκφρασης στα νευρικά κύτταρα από ότι στα κύτταρα του μεσοδέρματος. Το κίτρινο χρώμα χαρακτηρίζει γονίδια που έχουν γενικά χαμηλό επίπεδο έκφρασης και το επίπεδο έκφρασης τους στα κύτταρα του μεσοδέρματος είναι μεγαλύτερο από ότι το επίπεδο έκφρασης στα νευρικά κύτταρα. Τέλος υπάρχουν και τα γονίδια με γενικά μεγάλο επίπεδο έκφρασης και στους δύο ιστούς και τα οποία διαχωρίζονται στα πορτοκαλί που παρουσιάζουν χαμηλότερο επίπεδο έκφρασης στα μεσοδέρμια κύτταρα και μεγαλύτερο στα νευρικά κύτταρα και τα πράσινα που έχουν την ακριβώς αντίθετη συμπεριφορά.

Η στατιστική ανάλυση βέβαια σταματά στο σημείο αυτό διότι ας μην ξεχνάμε ότι είναι μια επιστήμη άκρως απαραίτητη και σημαντική για όλες τις επιστήμες η οποία όμως δεν μπορεί να καταλήξει σε κάποιο συμπέρασμα από μόνη της. Καταλήξαμε σε διαχωρισμούς που είναι σωστοί, ομογενείς, στατιστικά εμπεριστατωμένοι και μαθηματικά ερμηνευμένοι με βάση τις αριθμητικές τιμές του *gene expression*. Ωστόσο η πραγματική χρησιμότητα του παραπάνω διαχωρισμού μπορεί να φανεί μόνο με τη βοήθεια και άλλων ερμηνευτών όπως είναι ένας μοριακός βιολόγος ή ακόμα και ένας γενετιστής. Οι ειδικότητες αυτές θα μπορούσαν να απαντήσουν σε ερωτήματα όπως τα ακόλουθα: γιατί ορισμένα γονίδια έχουν τόσο χαμηλό επίπεδο έκφρασης και στους 2 ιστούς; Μήπως τα γονίδια αυτά δεν είναι λειτουργικά χρήσιμα και σε προβληματικούς ιστούς θα μπορούσαν να αντικατασταθούν με άλλα που θα εκτελούσαν παρόμοιες εργασίες; Μήπως όταν ανιχνευτεί ότι κάποιοι οργανισμοί πάσχουν από κάποια

σοβαρή νόσο όπως ο καρκίνος θα ήταν ωφέλιμο να ασχολούμαστε με γονίδια που παρουσιάζουν υψηλό επίπεδο έκφρασης; Τέλος μήπως γονίδια που παρουσιάζουν και στους δύο ιστούς χαμηλό *gene expression* θα μπορούσαν να αντικατασταθούν με άλλα που έχουν φανερά μεγαλύτερο *gene expression*;

4.7 Σύνοψη

Οι πρόσφατες ανακαλύψεις στο πεδίο των μικροσυστοιχιών έχουν κάνει δυνατή την ταυτόχρονη καταγραφή και παρακολούθηση του επιπέδου έκφρασης εκατοντάδων χιλιάδων γονιδίων. Τα *gene expression* που έχουν δημιουργηθεί από τα πειράματα των μικροσυστοιχιών αποτελούν το εφαλτήριο για τη δημιουργία σημαντικής προόδου στο χώρο της μοριακής βιολογίας και της γενετικής. Το γεγονός ότι τα *gene expression* μπορούν να ομαδοποιηθούν είτε με βάση τα γονίδια είτε με βάση τα δείγματα σε συνδυασμό με την πληθώρα των αλγορίθμων ομαδοποίησης καθώς και των μέτρων επικύρωσης αυτών, κάνει πολύ δύσκολη την επιλογή τελικού «νικητή». Οι ερευνητές συνήθως επιλέγουν μερικούς υποψήφιους αλγορίθμους και στη συνέχεια συγκρίνουν τα αποτελέσματα αυτών μεταξύ τους χρησιμοποιώντας ποικίλα μέτρα επικύρωσης. Μέχρι στιγμής όμως και πάλι δεν υπάρχει κάποιο μέτρο που να θεωρείται ως το βέλτιστο. Η εμπειρία του ερευνητή αλλά και η χρήση πολλών τεχνικών επικύρωσης και πολλών διαφορετικών αλγορίθμων είναι αυτά που συντελούν στην ύπαρξη ενός τελικού αποτελέσματος.

Η ομαδοποίηση σε συστάδες είναι μια ανεπιτήρητη (*unsupervised*) τεχνική διότι πριν ξεκινήσει η διαδικασία της υλοποίησής της η γενική δομή των δεδομένων είναι κάτι το άγνωστο ενώ και ο αριθμός των συστάδων που θα δημιουργηθεί «φυσικά» από τα δεδομένα παραμένει μια μη διαθέσιμη πληροφορία. Ωστόσο όταν κανείς προσπαθεί να ομαδοποιήσει *gene expression*, μερική (*partial*) πληροφορία είναι συνήθως διαθέσιμη επειδή τα περισσότερα γονίδια έχουν πλέον μελετηθεί από την επιστημονική κοινότητα. Το επόμενο στάδιο λοιπόν είναι η δημιουργία αλγορίθμων που θα συμπεριλαμβάνουν αυτήν την πληροφορία και συνεπώς τα αποτελέσματα της ομαδοποίησης θα εμπεριέχουν και βιολογική σημασία. Με αυτόν τον τρόπο η ανάλυση σε συστάδες θα μπορούσε να μη θεωρείται πλέον ως μια καθαρά ανεπιτήρητη διαδικασία αλλά ως μια διαδραστική εξερεύνηση του εκάστοτε συνόλου δεδομένων.

Παράρτημα 1

Ομαδοποίηση Γονιδίων σε Δύο Κλάσεις

Αποτελέσματα Ιεραρχικής Μεθόδου Μέσης Συνένωσης

FC	ID	M1	M2	M3	NC1	NC2	NC3	average gene expression
Miscellaneous	1422809_at	2,618148	2,57681	2,910598	2,099962	1,996473	2,124693	2,387780667
ECM/Receptors	1426295_at	2,351928	2,139469	2,5	3,010052	2,943602	3,102831	2,674647
EST	1449755_at	3,305476	2,812376	3,092469	2,643515	2,281415	2,544622	2,779978833
Transcription	1422018_at	3,07787	3,474349	2,943245	2,574116	2,494935	2,572165	2,856113333
EST	1433512_at	3,276028	3,494691	3,250888	3,037903	2,670424	2,315488	3,007570333
EST	1419833_s_at	3,940726	3,197243	3,757793	2,983312	2,51489	2,425794	3,136626333
Metabolism	1417572_at	3,427221	3,725257	3,209826	2,755829	2,888949	2,939537	3,157769833
EST	1431830_at	3,597294	3,525438	4,102375	3,008936	2,336887	2,570999	3,1903215
EST	1451103_at	2,679357	2,631145	3,247621	3,609388	3,582333	3,480991	3,205139167
Growth/Differentiation	1422979_at	3,967541	3,284824	4,140504	3,246692	2,611315	2,570999	3,303645833
Miscellaneous	1426371_at	3,705948	3,681074	3,619078	2,894868	2,892135	3,052953	3,307676
Transcription	1455056_at	3,642517	3,452545	4,504064	2,893563	2,91006	2,619018	3,336961167
EST	1417466_at	3,867962	4,730846	3,615999	2,878687	2,709133	2,864377	3,444500667
Unknown	1419361_at	2,958426	3,214419	3,304664	3,853808	3,566219	3,8772	3,462456
EST	1427314_at	3,828001	3,705832	3,993505	3,272484	3,1151	3,36816	3,547180333
Miscellaneous	1448131_at	3,252774	3,057946	3,324683	3,810313	3,968095	3,886858	3,5501115
Miscellaneous	1418486_at	3,105656	3,108106	3,089423	3,627988	4,120963	4,468995	3,586855167
Miscellaneous	1452003_at	3,394721	2,819723	3,495647	4,14837	4,074527	3,922893	3,642646833
Growth/Differentiation	1423176_at	4,347285	4,684069	3,937137	3,070039	3,328943	3,086045	3,742253
Transcription	1417447_at	4,188767	3,870036	4,193139	3,663919	3,47529	3,224656	3,769301167
Miscellaneous	1437434_a_at	2,875112	3,379619	3,2398	3,877053	4,459629	4,850978	3,780365167
Miscellaneous	1424107_at	4,21014	3,981831	4,315948	3,296243	3,5156	3,774206	3,848994667
Unknown	1451317_at	3,545845	3,680879	3,271709	4,34998	4,289833	4,15938	3,882937667
Stress-induced	1417667_a_at	3,379465	3,879171	3,395554	4,149116	4,145721	4,487829	3,906142667
Miscellaneous	1453030_at	4,281004	4,071795	4,387243	3,94096	3,54622	3,427108	3,942388333
Growth/Differentiation	1418471_at	4,261568	4,337538	4,262944	3,650103	3,672141	3,627875	3,968694833
Unknown	1419686_at	4,263081	4,156305	4,534461	3,639833	3,734857	3,510968	3,973250833
Stress-induced	1416481_s_at	3,546616	3,565047	3,626266	3,971116	4,533655	4,669095	3,985299167
Growth/Differentiation	1449773_s_at	4,187133	4,350387	4,853885	3,581607	3,621754	3,555297	4,0250105
EST	1435357_at	4,318275	4,330118	4,380836	3,560367	3,771357	3,792306	4,025543167
EST	1423852_at	4,532577	4,46477	4,440453	3,332766	3,508629	4,058757	4,056325333
ECM/Receptors	1416164_at	3,777285	3,744346	3,884532	4,584148	4,526107	4,367286	4,147284
EST	1439373_x_at	3,88854	3,674642	3,701682	4,147006	4,68368	4,828718	4,154044667
EST	1420643_at	3,867962	3,978505	3,727608	4,487708	4,484665	4,487534	4,172330333
Miscellaneous	1425534_at	4,539276	4,685715	4,496225	3,746535	3,849266	4,070839	4,231309333
Transcription	1424749_at	3,893333	3,814262	3,950241	4,533216	4,852018	4,433374	4,246074

EST	1451536_at	4,529505	4,545451	4,693118	3,631097	3,962975	4,182425	4,2574285
ECM/Receptors	1418084_at	4,319774	4,944365	4,565816	4,032451	3,890178	3,877236	4,271636667
Stress-induced	1434920_a_at	4,014939	3,564511	4,281244	4,432186	4,724101	4,766909	4,297315
EST	1452869_at	4,732801	4,644186	4,636211	3,96301	4,008731	3,9505	4,322573167
Miscellaneous	1451266_at	3,859249	4,184658	4,026063	4,606266	4,484175	4,821662	4,3303455
EST	1424212_at	4,018167	4,276643	3,861436	4,691209	4,64464	4,691122	4,3638695
Transcription	1427208_at	4,846981	4,522761	4,964488	4,287357	3,95517	3,668834	4,374265167
Growth/Differentiation	1448147_at	3,471347	4,129992	3,964431	4,474737	5,185631	5,177967	4,400684167
Metabolism	1429859_a_at	4,747454	4,836539	4,750263	4,198176	3,882843	4,204036	4,436551833
Transcription	1436392_s_at	3,867962	4,052354	3,474651	4,995836	5,056199	5,183585	4,438431167
Unknown	1450846_at	4,148161	4,030931	4,132405	4,890975	4,691047	4,749631	4,440525
Miscellaneous	1448143_at	4,083928	3,940565	4,297338	4,531557	4,659544	5,130691	4,440603833
EST	1423924_s_at	4,323924	4,347854	4,018901	4,733006	4,824678	4,98251	4,538478833
Unknown	1449376_at	4,809409	4,992339	4,816016	4,250109	3,993233	4,415811	4,546152833
EST	1418382_at	4,132713	4,084161	4,191685	4,434847	5,217284	5,225519	4,5477015
Miscellaneous	1422671_s_at	4,83264	4,787638	5,040072	4,299753	4,126242	4,299753	4,564349667
EST	1423124_x_at	4,292644	4,338523	4,15142	5,102978	5,064133	4,584442	4,589023333
Transcription	1429177_x_at	5,264923	4,639179	5,487279	4,236399	3,972734	3,956995	4,592918167
Miscellaneous	1448630_a_at	4,147493	4,371285	4,221351	4,504064	5,085161	5,289102	4,603076
EST	1435327_at	4,09493	3,949511	4,533295	5,339269	4,937019	4,830057	4,6140135
ECM/Receptors	1451532_s_at	4,389743	4,073278	4,40278	4,775284	4,998313	5,371796	4,668532333
Kinases/Phosphatases	1439148_a_at	4,24532	4,316414	4,093031	5,258339	5,192129	4,910329	4,669260333
Unknown	1456174_x_at	4,414424	4,482568	3,950702	4,84004	5,195778	5,162156	4,674278
ECM/Receptors	1416949_s_at	4,469444	4,146272	4,489389	4,634892	5,235221	5,093204	4,678070333
Transcription	1434831_a_at	4,835092	5,225802	4,999323	4,224547	4,346019	4,618982	4,708294167
Transcription	1426510_at	5,252326	5,282724	5,000021	3,862231	4,540778	4,555123	4,748867167
Transcription	1423319_at	5,132724	5,057613	5,529812	4,051826	4,625114	4,474746	4,8119725
Miscellaneous	1436528_at	5,259188	5,246742	5,094706	4,568253	4,511278	4,28624	4,8277345
Unknown	1460359_at	5,141025	5,027044	5,384229	4,559971	4,414799	4,498077	4,837524167
Unknown	1427672_a_at	4,993094	5,043411	5,47181	4,723959	4,34267	4,466635	4,840263167
EST	1424243_at	5,201806	5,216505	5,240903	4,4468	4,501398	4,597319	4,867455167
EST	1423123_at	4,438293	4,602599	4,502685	5,591108	5,286885	4,973777	4,8992245
Miscellaneous	1449697_s_at	4,52883	4,273115	4,942308	5,238129	5,356962	5,446057	4,9642335
Transcription	1450723_at	5,947581	5,752591	6,001086	5,027349	4,187176	3,111375	5,004526333
Transcription	1418516_at	5,121712	5,57127	5,211858	4,746511	4,775323	4,640271	5,0111575
Metabolism	1449059_a_at	4,44783	4,554605	4,574175	5,201016	5,597579	5,755704	5,021818167
Metabolism	1415787_at	4,993574	4,449732	4,814784	5,219723	5,506886	5,358577	5,057212667
EST	1435129_at	5,804189	5,129053	6,187772	4,301913	4,493597	4,435727	5,0587085
Transcription	1417624_at	5,339459	5,498137	5,232613	4,92529	4,971791	4,426459	5,065624833
Miscellaneous	1416812_at	5,403627	5,306198	5,496999	4,410019	4,86	4,927776	5,0674365
EST	1422850_at	5,237206	5,225268	5,841708	4,82	4,826692	4,476587	5,0712435
Transcription	1425779_a_at	6,108686	5,960381	5,999322	4,537256	4,067935	4,251332	5,154152
ECM/Receptors	1423110_at	4,61462	4,738261	4,530825	5,99025	5,703172	5,385871	5,160499833

EST	1451415_at	5,573594	5,601971	5,36964	4,424701	4,940855	5,058487	5,161541333
Growth/Differentiation	1425494_s_at	5,476976	5,429618	5,873447	4,955877	4,760259	4,606452	5,1837715
Stress-induced	1456434_x_at	4,568506	5,053994	4,712471	5,732924	5,495584	5,560326	5,187300833
EST	1439962_at	4,760489	5,280959	4,592463	5,536098	5,50495	5,543961	5,203153333
ECM/Receptors	1421045_at	4,855392	4,995958	4,860355	5,54211	5,529276	5,619332	5,233737167
EST	1449699_s_at	5,561725	5,318318	5,789635	5,075555	4,935555	4,783231	5,244003167
Metabolism	1452110_at	5,592074	5,44547	6,133715	4,862119	4,851557	4,823724	5,2847765
Transcription	1424186_at	4,775008	5,147855	4,720375	5,632075	5,704995	5,76051	5,290136333
Transcription	1448601_s_at	4,90124	5,165775	4,745772	5,689453	5,564776	5,674928	5,290324
EST	1434326_x_at	4,772312	5,049196	5,09365	5,721894	5,639231	5,77263	5,3414855
ECM/Receptors	1449885_at	5,7989	5,512948	5,61633	5,202085	5,234982	4,690486	5,342621833
Growth/Differentiation	1438312_s_at	5,812789	5,423305	5,713232	4,716954	5,233278	5,249413	5,358161833
Growth/Differentiation	1421180_at	4,869024	5,134643	4,88132	5,690545	5,879712	5,841125	5,382728167
ECM/Receptors	1448943_at	5,783536	6,187841	5,57001	5,095353	5,04965	4,643202	5,388265333
Transcription	1426485_at	5,857678	5,861506	5,631833	5,569766	5,029083	4,910993	5,476809833
Kinases/Phosphatases	1449630_s_at	5,17633	5,016422	5,410689	5,783389	5,771345	5,810368	5,494757167
Metabolism	1415904_at	5,168966	5,226089	5,063399	5,880583	5,879372	5,791927	5,501722667
Miscellaneous	1436886_x_at	5,363896	4,981026	5,388551	5,780188	5,93506	5,931014	5,563289167
Growth/Differentiation	1448995_at	4,706812	4,528291	4,325836	5,568435	6,915079	7,353144	5,566266167
Growth/Differentiation	1418424_at	5,09343	5,228219	5,268354	6,06908	5,965593	5,801478	5,571025667
Miscellaneous	1419430_at	6,218961	6,093435	6,194866	5,177967	4,821261	4,940498	5,574498
EST	1452665_at	5,090524	4,943403	5,582417	6,252147	5,985614	5,668496	5,587100167
Transcription	1421773_at	5,958226	5,948796	5,87165	5,641757	5,294736	4,983477	5,616440333
Transcription	1451332_at	6,103973	6,011468	5,997703	5,254292	5,348126	5,038787	5,625724833
Miscellaneous	1449623_at	5,29499	5,1803	5,054241	6,297913	6,204239	6,071913	5,683932667
Miscellaneous	1460260_s_at	5,463789	5,425168	5,293497	6,043178	5,981109	6,001029	5,701295
ECM/Receptors	1452671_s_at	5,370125	4,54681	5,70481	6,407555	6,310487	6,195847	5,755939
EST	1426845_at	5,339848	5,426816	5,749674	6,057577	6,105274	6,220465	5,816609
Kinases/Phosphatases	1424474_a_at	5,402011	4,888081	5,883447	6,493211	6,276636	6,045476	5,831477
EST	1455517_at	6,435079	6,698146	6,158235	5,248448	5,223501	5,311428	5,845806167
Transcription	1417129_a_at	6,693854	6,019203	7,036972	5,770116	5,148925	4,541272	5,868390333
EST	1456393_at	6,307372	6,314294	6,514434	5,666934	5,526444	5,045101	5,895763167
Kinases/Phosphatases	1429514_at	6,439456	6,784651	5,919455	5,684244	5,374218	5,525489	5,9545855
Transcription	1418649_at	6,353283	6,077029	6,745278	5,406186	5,59572	5,683083	5,976763167
Miscellaneous	1415993_at	5,653472	5,792886	5,369757	5,956055	6,632036	6,678829	6,013839167
Miscellaneous	1448466_at	5,673765	5,58515	5,890892	6,216657	6,29384	6,433541	6,015640833
ECM/Receptors	1426628_at	6,574904	5,951698	7,049645	5,736507	5,547328	5,344793	6,034145833
Miscellaneous	1423095_s_at	6,376106	6,051813	6,619945	5,523294	5,881419	5,894885	6,057910333
Miscellaneous	1428922_at	5,326943	5,49893	5,629814	6,795194	6,535522	6,622577	6,068163333
ECM/Receptors	1427256_at	6,573736	6,498076	6,446686	6,01597	5,720167	5,60348	6,143019167
Growth/Differentiation	1426083_a_at	6,358946	6,419124	6,587548	5,591812	5,905448	6,023786	6,147777333
Unknown	1452183_a_at	6,622879	6,05783	6,966532	5,88162	5,788236	5,746224	6,177220167
Growth/Differentiation	1448823_at	6,55415	7,026365	6,423227	5,757455	5,740969	5,585081	6,181207833

EST	1450941_at	6,474793	6,251261	6,870609	5,778282	6,00763	5,994492	6,229511167
Miscellaneous	1418479_at	6,595568	6,269588	6,879275	6,078799	6,029391	5,643819	6,249406667
Metabolism	1434485_a_at	6,710259	6,688321	6,854566	5,839616	6,194749	6,475638	6,460524833
Kinases/Phosphatases	1417425_at	6,184555	6,389695	6,123599	6,723693	6,965548	6,970946	6,559672667
Transcription	1420886_a_at	6,964868	6,848927	6,85528	6,325325	6,392266	6,094601	6,580211167
ECM/Receptors	1417964_at	6,235028	6,204409	6,314183	7,008706	6,979112	6,857127	6,599760833
Transcription	1420011_s_at	6,938305	6,834526	7,063821	6,491878	6,118042	6,152163	6,599789167
Transcription	1418835_at	7,689747	6,769066	8,219896	6,53566	6,074489	6,10586	6,899119667
Kinases/Phosphatases	1450070_s_at	6,662147	6,707125	6,687644	7,340441	7,395206	7,3853	7,029643833
ECM/Receptors	1439381_x_at	6,760322	6,867398	6,244526	7,617077	7,599799	7,381005	7,0783545
Metabolism	1418560_at	6,882713	6,556255	6,96708	7,284192	7,414086	7,47502	7,096557667
Transcription	1427120_at	7,289739	7,519645	7,449258	7,11201	6,788077	6,601185	7,126652333
Growth/Differentiation	1448259_at	7,60115	7,563963	7,573535	6,852338	6,966401	7,105625	7,277168667
Growth/Differentiation	1416855_at	6,901971	6,841808	6,950886	7,651976	7,682734	7,852736	7,313685167
Growth/Differentiation	1437455_a_at	7,608199	7,819271	7,897766	6,887217	6,886286	7,158846	7,376264167
ECM/Receptors	1450857_a_at	7,002372	7,510378	6,814198	8,035897	7,953074	7,858245	7,529027333
ECM/Receptors	1438651_a_at	7,925312	8,220663	7,664084	7,47644	7,086332	6,829566	7,533732833
Kinases/Phosphatases	1448269_a_at	7,323101	6,97341	7,415568	7,540145	8,045597	8,082695	7,563419333
Stress-induced	1428942_at	7,685824	7,820177	7,330754	8,288514	8,223115	8,39759	7,957662333
Growth/Differentiation	1416221_at	8,433042	8,540299	8,342565	8,006506	7,716163	7,768855	8,134571667
Stress-induced	1422557_s_at	7,817395	7,54272	7,905092	8,905423	8,544108	8,374363	8,181516833
Transcription	1437163_x_at	7,940359	8,126173	7,643307	8,644427	8,53761	8,373675	8,210925167
Unknown	1416181_at	7,981006	7,972705	8,164087	8,471948	8,932492	8,998186	8,420070667
Transcription	1451418_a_at	8,054738	8,035023	8,270256	8,466145	8,954269	9,25102	8,505241833
Transcription	1437223_s_at	9,281908	9,273121	9,228188	8,404881	8,502308	8,378771	8,844862833

Παράρτημα 2

Ομαδοποίηση Γονιδίων σε Έξι Κλάσεις

Αποτελέσματα Μεθόδου Partition Around Medoids

FC	ID	M1	M2	M3	NC1	NC2	NC3	average gene expression
Miscellaneous	1422809_at	2,618148	2,57681	2,910598	2,099962	1,996473	2,124693	2,387780667
ECM/Receptors	1426295_at	2,351928	2,139469	2,5	3,010052	2,943602	3,102831	2,674647
EST	1449755_at	3,305476	2,812376	3,092469	2,643515	2,281415	2,544622	2,779978833
Transcription	1422018_at	3,07787	3,474349	2,943245	2,574116	2,494935	2,572165	2,856113333
EST	1433512_at	3,276028	3,494691	3,250888	3,037903	2,670424	2,315488	3,007570333
EST	1419833_s_at	3,940726	3,197243	3,757793	2,983312	2,51489	2,425794	3,136626333
Metabolism	1417572_at	3,427221	3,725257	3,209826	2,755829	2,888949	2,939537	3,157769833
EST	1431830_at	3,597294	3,525438	4,102375	3,008936	2,336887	2,570999	3,1903215
EST	1451103_at	2,679357	2,631145	3,247621	3,609388	3,582333	3,480991	3,205139167
Growth/Differentiation	1422979_at	3,967541	3,284824	4,140504	3,246692	2,611315	2,570999	3,303645833
Miscellaneous	1426371_at	3,705948	3,681074	3,619078	2,894868	2,892135	3,052953	3,307676
Transcription	1455056_at	3,642517	3,452545	4,504064	2,893563	2,91006	2,619018	3,336961167
EST	1417466_at	3,867962	4,730846	3,615999	2,878687	2,709133	2,864377	3,444500667
Unknown	1419361_at	2,958426	3,214419	3,304664	3,853808	3,566219	3,8772	3,462456
EST	1427314_at	3,828001	3,705832	3,993505	3,272484	3,1151	3,36816	3,547180333
Miscellaneous	1448131_at	3,252774	3,057946	3,324683	3,810313	3,968095	3,886858	3,5501115
Miscellaneous	1418486_at	3,105656	3,108106	3,089423	3,627988	4,120963	4,468995	3,586855167
Miscellaneous	1452003_at	3,394721	2,819723	3,495647	4,14837	4,074527	3,922893	3,642646833
Growth/Differentiation	1423176_at	4,347285	4,684069	3,937137	3,070039	3,328943	3,086045	3,742253
Transcription	1417447_at	4,188767	3,870036	4,193139	3,663919	3,47529	3,224656	3,769301167
Miscellaneous	1437434_a_at	2,875112	3,379619	3,2398	3,877053	4,459629	4,850978	3,780365167
Miscellaneous	1424107_at	4,21014	3,981831	4,315948	3,296243	3,5156	3,774206	3,848994667
Unknown	1451317_at	3,545845	3,680879	3,271709	4,34998	4,289833	4,15938	3,882937667
Stress-induced	1417667_a_at	3,379465	3,879171	3,395554	4,149116	4,145721	4,487829	3,906142667
Miscellaneous	1453030_at	4,281004	4,071795	4,387243	3,94096	3,54622	3,427108	3,942388333
Growth/Differentiation	1418471_at	4,261568	4,337538	4,262944	3,650103	3,672141	3,627875	3,968694833
Unknown	1419686_at	4,263081	4,156305	4,534461	3,639833	3,734857	3,510968	3,973250833
Stress-induced	1416481_s_at	3,546616	3,565047	3,626266	3,971116	4,533655	4,669095	3,985299167
Growth/Differentiation	1449773_s_at	4,187133	4,350387	4,853885	3,581607	3,621754	3,555297	4,0250105
EST	1435357_at	4,318275	4,330118	4,380836	3,560367	3,771357	3,792306	4,025543167
EST	1423852_at	4,532577	4,46477	4,440453	3,332766	3,508629	4,058757	4,056325333
ECM/Receptors	1416164_at	3,777285	3,744346	3,884532	4,584148	4,526107	4,367286	4,147284
EST	1439373_x_at	3,88854	3,674642	3,701682	4,147006	4,68368	4,828718	4,154044667

EST	1420643_at	3,867962	3,978505	3,727608	4,487708	4,484665	4,487534	4,172330333
Miscellaneous	1425534_at	4,539276	4,685715	4,496225	3,746535	3,849266	4,070839	4,231309333
Transcription	1424749_at	3,893333	3,814262	3,950241	4,533216	4,852018	4,433374	4,246074
EST	1451536_at	4,529505	4,545451	4,693118	3,631097	3,962975	4,182425	4,2574285
ECM/Receptors	1418084_at	4,319774	4,944365	4,565816	4,032451	3,890178	3,877236	4,271636667
Stress-induced	1434920_a_at	4,014939	3,564511	4,281244	4,432186	4,724101	4,766909	4,297315
EST	1452869_at	4,732801	4,644186	4,636211	3,96301	4,008731	3,9505	4,322573167
Miscellaneous	1451266_at	3,859249	4,184658	4,026063	4,606266	4,484175	4,821662	4,3303455
EST	1424212_at	4,018167	4,276643	3,861436	4,691209	4,64464	4,691122	4,3638695
Transcription	1427208_at	4,846981	4,522761	4,964488	4,287357	3,95517	3,668834	4,374265167
Growth/Differentiation	1448147_at	3,471347	4,129992	3,964431	4,474737	5,185631	5,177967	4,400684167
Metabolism	1429859_a_at	4,747454	4,836539	4,750263	4,198176	3,882843	4,204036	4,436551833
Transcription	1436392_s_at	3,867962	4,052354	3,474651	4,995836	5,056199	5,183585	4,438431167
Unknown	1450846_at	4,148161	4,030931	4,132405	4,890975	4,691047	4,749631	4,440525
Miscellaneous	1448143_at	4,083928	3,940565	4,297338	4,531557	4,659544	5,130691	4,440603833
EST	1423924_s_at	4,323924	4,347854	4,018901	4,733006	4,824678	4,98251	4,538478833
Unknown	1449376_at	4,809409	4,992339	4,816016	4,250109	3,993233	4,415811	4,546152833
EST	1418382_at	4,132713	4,084161	4,191685	4,434847	5,217284	5,225519	4,5477015
Miscellaneous	1422671_s_at	4,83264	4,787638	5,040072	4,299753	4,126242	4,299753	4,564349667
EST	1423124_x_at	4,292644	4,338523	4,15142	5,102978	5,064133	4,584442	4,589023333
Transcription	1429177_x_at	5,264923	4,639179	5,487279	4,236399	3,972734	3,956995	4,592918167
Miscellaneous	1448630_a_at	4,147493	4,371285	4,221351	4,504064	5,085161	5,289102	4,603076
EST	1435327_at	4,09493	3,949511	4,533295	5,339269	4,937019	4,830057	4,6140135
ECM/Receptors	1451532_s_at	4,389743	4,073278	4,40278	4,775284	4,998313	5,371796	4,668532333
Kinases/Phosphatases	1439148_a_at	4,24532	4,316414	4,093031	5,258339	5,192129	4,910329	4,669260333
Unknown	1456174_x_at	4,414424	4,482568	3,950702	4,84004	5,195778	5,162156	4,674278
ECM/Receptors	1416949_s_at	4,469444	4,146272	4,489389	4,634892	5,235221	5,093204	4,678070333
Transcription	1434831_a_at	4,835092	5,225802	4,999323	4,224547	4,346019	4,618982	4,708294167
Transcription	1426510_at	5,252326	5,282724	5,000021	3,862231	4,540778	4,555123	4,748867167
Transcription	1423319_at	5,132724	5,057613	5,529812	4,051826	4,625114	4,474746	4,8119725
Miscellaneous	1436528_at	5,259188	5,246742	5,094706	4,568253	4,511278	4,28624	4,8277345
Unknown	1460359_at	5,141025	5,027044	5,384229	4,559971	4,414799	4,498077	4,837524167
Unknown	1427672_a_at	4,993094	5,043411	5,47181	4,723959	4,34267	4,466635	4,840263167
EST	1424243_at	5,201806	5,216505	5,240903	4,4468	4,501398	4,597319	4,867455167
EST	1423123_at	4,438293	4,602599	4,502685	5,591108	5,286885	4,973777	4,8992245
Miscellaneous	1449697_s_at	4,52883	4,273115	4,942308	5,238129	5,356962	5,446057	4,9642335
Transcription	1450723_at	5,947581	5,752591	6,001086	5,027349	4,187176	3,111375	5,004526333
Transcription	1418516_at	5,121712	5,57127	5,211858	4,746511	4,775323	4,640271	5,0111575
Metabolism	1449059_a_at	4,44783	4,554605	4,574175	5,201016	5,597579	5,755704	5,021818167
Metabolism	1415787_at	4,993574	4,449732	4,814784	5,219723	5,506886	5,358577	5,057212667
EST	1435129_at	5,804189	5,129053	6,187772	4,301913	4,493597	4,435727	5,0587085

Transcription	1417624_at	5,339459	5,498137	5,232613	4,92529	4,971791	4,426459	5,065624833
Miscellaneous	1416812_at	5,403627	5,306198	5,496999	4,410019	4,86	4,927776	5,0674365
EST	1422850_at	5,237206	5,225268	5,841708	4,82	4,826692	4,476587	5,0712435
Transcription	1425779_a_at	6,108686	5,960381	5,999322	4,537256	4,067935	4,251332	5,154152
ECM/Receptors	1423110_at	4,61462	4,738261	4,530825	5,99025	5,703172	5,385871	5,160499833
EST	1451415_at	5,573594	5,601971	5,36964	4,424701	4,940855	5,058487	5,161541333
Growth/Differentiation	1425494_s_at	5,476976	5,429618	5,873447	4,955877	4,760259	4,606452	5,1837715
Stress-induced	1456434_x_at	4,568506	5,053994	4,712471	5,732924	5,495584	5,560326	5,187300833
EST	1439962_at	4,760489	5,280959	4,592463	5,536098	5,50495	5,543961	5,203153333
ECM/Receptors	1421045_at	4,855392	4,995958	4,860355	5,54211	5,529276	5,619332	5,233737167
EST	1449699_s_at	5,561725	5,318318	5,789635	5,075555	4,935555	4,783231	5,244003167
Metabolism	1452110_at	5,592074	5,44547	6,133715	4,862119	4,851557	4,823724	5,2847765
Transcription	1424186_at	4,775008	5,147855	4,720375	5,632075	5,704995	5,76051	5,290136333
Transcription	1448601_s_at	4,90124	5,165775	4,745772	5,689453	5,564776	5,674928	5,290324
EST	1434326_x_at	4,772312	5,049196	5,09365	5,721894	5,639231	5,77263	5,3414855
ECM/Receptors	1449885_at	5,7989	5,512948	5,61633	5,202085	5,234982	4,690486	5,342621833
Growth/Differentiation	1438312_s_at	5,812789	5,423305	5,713232	4,716954	5,233278	5,249413	5,358161833
Growth/Differentiation	1421180_at	4,869024	5,134643	4,88132	5,690545	5,879712	5,841125	5,382728167
ECM/Receptors	1448943_at	5,783536	6,187841	5,57001	5,095353	5,04965	4,643202	5,388265333
Transcription	1426485_at	5,857678	5,861506	5,631833	5,569766	5,029083	4,910993	5,476809833
Kinases/Phosphatases	1449630_s_at	5,17633	5,016422	5,410689	5,783389	5,771345	5,810368	5,494757167
Metabolism	1415904_at	5,168966	5,226089	5,063399	5,880583	5,879372	5,791927	5,501722667
Miscellaneous	1436886_x_at	5,363896	4,981026	5,388551	5,780188	5,93506	5,931014	5,563289167
Growth/Differentiation	1448995_at	4,706812	4,528291	4,325836	5,568435	6,915079	7,353144	5,566266167
Growth/Differentiation	1418424_at	5,09343	5,228219	5,268354	6,06908	5,965593	5,801478	5,571025667
Miscellaneous	1419430_at	6,218961	6,093435	6,194866	5,177967	4,821261	4,940498	5,574498
EST	1452665_at	5,090524	4,943403	5,582417	6,252147	5,985614	5,668496	5,587100167
Transcription	1421773_at	5,958226	5,948796	5,87165	5,641757	5,294736	4,983477	5,616440333
Transcription	1451332_at	6,103973	6,011468	5,997703	5,254292	5,348126	5,038787	5,625724833
Miscellaneous	1449623_at	5,29499	5,1803	5,054241	6,297913	6,204239	6,071913	5,683932667
Miscellaneous	1460260_s_at	5,463789	5,425168	5,293497	6,043178	5,981109	6,001029	5,701295
ECM/Receptors	1452671_s_at	5,370125	4,54681	5,70481	6,407555	6,310487	6,195847	5,755939
EST	1426845_at	5,339848	5,426816	5,749674	6,057577	6,105274	6,220465	5,816609
Kinases/Phosphatases	1424474_a_at	5,402011	4,888081	5,883447	6,493211	6,276636	6,045476	5,831477
EST	1455517_at	6,435079	6,698146	6,158235	5,248448	5,223501	5,311428	5,845806167
Transcription	1417129_a_at	6,693854	6,019203	7,036972	5,770116	5,148925	4,541272	5,868390333
EST	1456393_at	6,307372	6,314294	6,514434	5,666934	5,526444	5,045101	5,895763167
Kinases/Phosphatases	1429514_at	6,439456	6,784651	5,919455	5,684244	5,374218	5,525489	5,9545855
Transcription	1418649_at	6,353283	6,077029	6,745278	5,406186	5,59572	5,683083	5,976763167
Miscellaneous	1415993_at	5,653472	5,792886	5,369757	5,956055	6,632036	6,678829	6,013839167
Miscellaneous	1448466_at	5,673765	5,58515	5,890892	6,216657	6,29384	6,433541	6,015640833

ECM/Receptors	1426628_at	6,574904	5,951698	7,049645	5,736507	5,547328	5,344793	6,034145833
Miscellaneous	1423095_s_at	6,376106	6,051813	6,619945	5,523294	5,881419	5,894885	6,057910333
Miscellaneous	1428922_at	5,326943	5,49893	5,629814	6,795194	6,535522	6,622577	6,068163333
ECM/Receptors	1427256_at	6,573736	6,498076	6,446686	6,01597	5,720167	5,60348	6,143019167
Growth/Differentiation	1426083_a_at	6,358946	6,419124	6,587548	5,591812	5,905448	6,023786	6,147777333
Unknown	1452183_a_at	6,622879	6,05783	6,966532	5,88162	5,788236	5,746224	6,177220167
Growth/Differentiation	1448823_at	6,55415	7,026365	6,423227	5,757455	5,740969	5,585081	6,181207833
EST	1450941_at	6,474793	6,251261	6,870609	5,778282	6,00763	5,994492	6,229511167
Miscellaneous	1418479_at	6,595568	6,269588	6,879275	6,078799	6,029391	5,643819	6,249406667
Metabolism	1434485_a_at	6,710259	6,688321	6,854566	5,839616	6,194749	6,475638	6,460524833
Kinases/Phosphatases	1417425_at	6,184555	6,389695	6,123599	6,723693	6,965548	6,970946	6,559672667
Transcription	1420886_a_at	6,964868	6,848927	6,85528	6,325325	6,392266	6,094601	6,580211167
ECM/Receptors	1417964_at	6,235028	6,204409	6,314183	7,008706	6,979112	6,857127	6,599760833
Transcription	1420011_s_at	6,938305	6,834526	7,063821	6,491878	6,118042	6,152163	6,599789167
Transcription	1418835_at	7,689747	6,769066	8,219896	6,53566	6,074489	6,10586	6,899119667
Kinases/Phosphatases	1450070_s_at	6,662147	6,707125	6,687644	7,340441	7,395206	7,3853	7,029643833
ECM/Receptors	1439381_x_at	6,760322	6,867398	6,244526	7,617077	7,599799	7,381005	7,0783545
Metabolism	1418560_at	6,882713	6,556255	6,96708	7,284192	7,414086	7,47502	7,096557667
Transcription	1427120_at	7,289739	7,519645	7,449258	7,11201	6,788077	6,601185	7,126652333
Growth/Differentiation	1448259_at	7,60115	7,563963	7,573535	6,852338	6,966401	7,105625	7,277168667
Growth/Differentiation	1416855_at	6,901971	6,841808	6,950886	7,651976	7,682734	7,852736	7,313685167
Growth/Differentiation	1437455_a_at	7,608199	7,819271	7,897766	6,887217	6,886286	7,158846	7,376264167
ECM/Receptors	1450857_a_at	7,002372	7,510378	6,814198	8,035897	7,953074	7,858245	7,529027333
ECM/Receptors	1438651_a_at	7,925312	8,220663	7,664084	7,47644	7,086332	6,829566	7,533732833
Kinases/Phosphatases	1448269_a_at	7,323101	6,97341	7,415568	7,540145	8,045597	8,082695	7,563419333
Stress-induced	1428942_at	7,685824	7,820177	7,330754	8,288514	8,223115	8,39759	7,957662333
Growth/Differentiation	1416221_at	8,433042	8,540299	8,342565	8,006506	7,716163	7,768855	8,134571667
Stress-induced	1422557_s_at	7,817395	7,54272	7,905092	8,905423	8,544108	8,374363	8,181516833
Transcription	1437163_x_at	7,940359	8,126173	7,643307	8,644427	8,53761	8,373675	8,210925167
Unknown	1416181_at	7,981006	7,972705	8,164087	8,471948	8,932492	8,998186	8,420070667
Transcription	1451418_a_at	8,054738	8,035023	8,270256	8,466145	8,954269	9,25102	8,505241833
Transcription	1437223_s_at	9,281908	9,273121	9,228188	8,404881	8,502308	8,378771	8,844862833

Παράρτημα 3

Χρήσιμες Εντολές για την Παραγωγή Αποτελεσμάτων μέσω R

```
**Basic Points of R programming**

library(clValid)
data(mouse)
express <- mouse[, c("M1", "M2", "M3", "NC1", "NC2", "NC3")]
 rownames(express) <- mouse$ID
express


***hierachical average euclidean distance****

dist=dist(express, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
dist

caver <- hclust(dist, method = "aver")
plot(caver, hang = -1)
rect.hclust(caver, 2)
rect.hclust(caver, 6)

**k-means clustering**

k=kmeans(express, 6, iter.max = 1000, nstart = 10,algorithm = "MacQueen")
print(k)

library(cluster)
clusplot(express, k$cluster, main="Clustering Plot for k
means",plotchar=TRUE,color=TRUE,labels=1,shade=TRUE)

# K-Means Clustering with 6 clusters
k6means <- kmeans(express, 6)
k6means

library(cluster)
# Cluster Plot against 1st 2 principal components vary parameters for most readable graph

clusplot(express, k6means$cluster, main="Clustering Plot for k means",
plotchar=TRUE,color=TRUE,labels=1,shade=TRUE)

**PAM clustering**

PAM2=pam(express,2, diss = FALSE, metric = "euclidean",medoids = NULL, stand = FALSE,
cluster.only = FALSE,do.swap = TRUE)
```

```
PAM2
plot(PAM2)
si=silhouette(PAM2)
plot(si) # silhouette plot
plot(si, col = c("red", "purple"))
clusplot(express, PAM2$cluster, main="Clustering Plot for PAM",
plotchar=TRUE,color=TRUE,labels=1,shade=TRUE)


** Pam clustering with 6 clusters
PAM=pam(express,6, diss = FALSE, metric = "euclidean",medoids = NULL, stand = FALSE,
cluster.only = FALSE,do.swap = TRUE)
PAM
plot(PAM)
summary(PAM)
print(PAM)
si=silhouette(PAM)
plot(si) # silhouette plot
plot(si, col = c("red", "purple","blue","green","yellow","pink"))#

clusplot(express, PAM$cluster, main="Clustering Plot for PAM",
plotchar=TRUE,color=TRUE,labels=1,shade=TRUE)

**CLARA clustering**

clara=clara(express,2, metric = "euclidean", stand = FALSE,samples=5)
plot(clara)
si=silhouette(clara,full=TRUE)

plot(si, col = c("purple"))
summary(clara)
print(si)
summary(si)
print(clara)

library(cluster)
clusplot(express, clara$cluster, main="Clustering Plot for CLARA",
plotchar=TRUE,color=TRUE,labels=1,shade=TRUE)

##By default, for clara() partitions, the silhouette is just for the best random subset used.
##Use full = TRUE to compute  (and later possibly plot) the full silhouette.

**stability measures**
stab <- clValid(express, 2:6, clMethods = c("hierarchical",
"kmeans","pam","clara"),method=c("average"), metric="euclidean", validation = "stability")

summary(stab)
plot(stab,legendLoc = "topleft")
```

```
**biological measures**

fc <- tapply(rownames(express), mouse$FC, c)
fc <- fc[!names(fc) %in% c("EST", "Unknown")]
bio <- clValid(express, 2:6, clMethods = c("hierarchical", "kmeans","pam","clara"),
method=c("average"), metric="euclidean", validation = "biological", annotation = fc)
summary(bio)

optimalScores(bio)

plot(bio, measure = "BHI", legendLoc = "topleft")
plot(bio, measure = "BSI",legendLoc = "topright")

****RANK AGGREGATION 2:6******

result <- clValid(express, 2:6, clMethods = c("hierarchical","kmeans", "pam","clara"),
method=c("average"), metric="euclidean",validation = c("internal",
"stability","biological"),annotation = fc)

res <- getRanksWeights(result)
print(res$ranks[, 1:3], quote = FALSE)

if (require("RankAggreg"))
{CEWS <- RankAggreg(x = res$ranks, k = 3, weights = res$weights,distance="Spearman",
seed = 1000, verbose = FALSE)
CEWS}

plot(CEWS)

****RANK AGGREGATION 3:6******

result <- clValid(express, 3:6, clMethods = c("hierarchical","kmeans", "pam","clara"),
method=c("average"), metric="euclidean",validation = c("internal",
"stability","biological"),annotation = fc)

res <- getRanksWeights(result)
print(res$ranks[, 1:3], quote = FALSE)

if (require("RankAggreg"))
{CEWS <- RankAggreg(x = res$ranks, k = 3, weights = res$weights,distance="Spearman",
seed = 1000, verbose = FALSE)
CEWS}

plot(CEWS)
```

Ελληνική Βιβλιογραφία

1. Ανθοπούλου Ο., Μαυράκης Α., Τσιράκης Ν. Ανάλυση Γονιδιακής Έκφρασης, Εργασία για το μάθημα «Εισαγωγή στη Βιοπληροφορική».

2. Δημητρακοπούλου Κ. (2007), Διπλωματική Εργασία για το ΔΠΜΣ Βιοϊατρική Τεχνολογία: «Αναγνώριση Λειτουργικών Υποδομών στο Πρωτεϊνικό Δίκτυο του Saccharomyces Cerevisiae Συνδυάζοντας Δεδομένα Έκφρασης Γονιδίων και Αλληλεπίδρασης Πρωτεϊνών».

3. Κούτρας Μ. (2009), Εφαρμοσμένη Πολυμεταβλητή Ανάλυση: Ανάλυση κατά Συστάδες. Πανεπιστημιακές Εκδόσεις, Πανεπιστήμιο Πειραιά.

Ξένη Βιβλιογραφία

1. Anderbeg M. (1973) *Cluster Analysis for applications.* Academic Press, New York & London.

2. Barrett J.C., Kawasaki E.S. (2003) *Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression.* DDT, 8(3), 134:141.

3. Berrar D., Dubitzky W., Granzow M. (2003) *A Practical Approach to Microarray Data Analysis,* KLUWER ACADEMIC PUBLISHERS.

4. Bhattacherjee V., Mukhopadhyay P., Singh S., Johnson C., Philipose J. T., Warner C. P., Greene R. M., Pisano M. M. (2006) *Neural Crest and Mesoderm Lineage-Dependent Gene Expression in Orofacial Development.* Differantiation, 75:463-477.

5. Brazma A., Vilo J. (2000) *Minireview: Gene Expression Data Analysis.* FEBS Letters 480, 17:24.

6. Brock G., Pihur V., Datta S, Datta S. (2008) *clValid, an R package for Cluster Validation.* Journal of Statistical Software, 25(4), 1:22.

7. Cobb K. (2006) *Microarrays: the Search for Meaning in a Vast Sea of Data*. Biomedical Computational Review, www.biomedicalcomputationreview.org, 16:23.

8. Datta S. (2003) *Statistical Techniques for Microarray Data: A Partial Overview.* Communications in Statistics - Theory and Methods, 32(1), 263:280.

9. Datta S., Datta S. (2003) *Comparisons and validation of statistical clustering techniques for microarray gene expression data*. Bioinformatics, 19(4), 459-466.

10. Datta S., Datta S. (2006) *Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.* BMC Bioinformatics, 7:397.

11. Dietz K., Gail M., Krickeberg K., Samet J., Tsiatis A.(2003) *The Analysis of Gene Expression Data: Methods and Software,* Springer.

12. Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998) *Cluster Analysis and Display of Genome-Wide Expression Patterns.* Proc. Natl. Acad. Sci. USA, 95. 14863:14868.

13. European Molecular Biology Laboratory, URL: http://www.embl.it .

14. Everitt B.S., Dunn G. (1991) *Applied Multivariate Data Analysis*, Arnold, New York.

15. Famili F., Liu G., Liu Z. (2004) *Evaluation and optimization of clustering in gene expression data analysis.* Bioinformatics, 20(10), 1535-1545.

16. Gordon A.D. (1999) *Classification* (2nd Edition), Chapman & Hall/CRC.

17. Handl J., Knowles J., Kell D. (2005) *Computational cluster validation in post-genomic data analysis.* Bioinformatics, 21(15), 3201-3212.

18. Jiang D., Tang C., Zhang A. (2004) *Cluster Analysis for Gene Expression Data: A Survey.* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 16(11), 1370:1386.

19. Kaufman L., Rousseeuw P. (1990) *Finding groups in data.* Wiley Series In Probability and Mathematical Statistics.

20. Koutsos A., Manaia A., Willingale-Theune J. (2009) *Fishing for Genes: DNA microarrays in the Classroom.* Science in School, 12, 44:49.

21. Koutsos A., Manaia A., Willingale-Theune J. (2010) *Introduction to DNA microarrays.* European Learning Laboratory for Life Sciences.

22. Oksanen J., (2010) *Cluster Analysis: Tutorial with R.*

23. Pihur V., Datta S., Datta S. (2007) *Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach.* Bioinformatics, 23(13), 1607-1615.

24. Pihur V., Datta S., Datta S. (2009) *RankAggreg, an R package for weighted Rank Aggregation.* BMC Bioinformatics, 10:62.

25. Ridge B., URL: http://ridge.icu.ac.jp/biobk/BioBookPROTSYn.html#Links.

26. Rousseeuw J. (1986) *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics, 20, 53-65.

27. Yeung K.Y, Haynor D.R, Ruzzo W.L. (2001) *Validating clustering for gene expression data*. Bioinformatics, 17, 309-318.