Πανεπιστήμιο Θεσσαλιάς

METHODS FOR ANALYZING MICROARRAY DATA-CLUSTERING

ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΕΙΩΝ-ΣΥΣΤΑΔΟΠΟΙΗΣΗ

ΔΕΛΗΝΑΣΙΟΣ ΛΑΖΑΡΟΣ

2017

Πρόλογος

Μια μικρή εισαγωγή στην τεχνολογία μικροσυστοιχειών DNA.

Οι μικροσυστοιχείες γονιδίων (αλλιώς γνωστές ως γονιδιακό ή γενωμικό τσιπ,DNA τσιπ) είναι μια διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια και ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μια στέρεη επιφάνεια (συνήθως γυάλινη).Με την πάροδο των ετών άρχισαν να χρησιμοποιούνται κατά κόρον για την μέτρηση DNA, ή χρησιμοποιούν DNA για το σύστημα ανίχνευσής τους. Ποσοτικές ή ποιοτικές μετρήσεις με μικροσυστοιχίες γονιδίων εκμεταλευόμενες την αρχή της συμπληρωματικότητας μεταξύ νουκλεϊκών οξέων DNA-DNA ή DNA-RNA ή μεταξύ των αμινοξέων των πρωτεϊνών,υπό αυστηρά ελεγχόμενες συνθήκες θερμοκρασιας και με την χρήση φθοριζουσών ουσιών χρησιμοποιούνται για την εξέταση της γονιδιακής έκφρασης υπό ειδικές συνθήκες και για την ανίχνευση νουκλεϊκών οξέων παθογόνων οργανισμών. Στη συγκεκριμένη εργασία θα ασχοληθούμε με την ανάλυση δεδομένων μικροσυστοιχειών με τη χρήση αλγορίθμων συσταδοποίησης(clustering analysis) και κυρίως με τις δύο πιο διαδεδομένες μεθόδους, την ιεραρχική συσταδοποίηση και την συσταδοποίηση k-μέσων με τα μειονεκτήματα και τα πλεονεκτήματά τους καθώς και ανάλυση δεδομένων(προσομοίωσης) για την περαιτέρω κατανόηση τους.Επίσης θα παρετεθεί εν συντομία και δευτερογενής ανάλυση των δεδομένων για την περαιτέρω κατανόηση των μηχανισμών ενός βιολογικού συστήματος χωρίς όμως να γίνει η ανάλυσή τους.

Μικροσυστοιχίες – Εφαρμογές μικροσυστοιχειών.

Το μεγάλο πλεονέκτημα των μικροσυστοιχειών DNA είναι η ικανότητα εκτίμησης της έκφρασης δεκάδων ή εκατοντάδων χιλιάδων γονιδίων τη φορά,δηλαδή η ικανότητά τους να παρέχουν ένα προφίλ γονιδιακής έκφρασης. Η τεχνολογία των μικροσυστοιχειών DNA στηρίχθηκε στην βασική ιδιότητα των νουκλεϊκών δηλαδή στην αρχή της συμπληρωματικότητας.Οι εφαρμογές των μικροσυστοιχειών είναι πολλές, αλλά η συνηθέστερη αυτών και αυτή που εμφανίστηκε πρώτη είναι ο προσδιορισμός του προφίλ της γονιδιακής έκφρασης. Την σημερινή εποχή οι επιστήμονες διεξάγουν μελέτες μεγάλου πληθυσμού π.χ. για να καθορίσουν πόσο συχνά άτομα με μια συγκεκριμένη μετάλλαξη αναπτύσουν καρκίνο του μαστού , ή για τον εντοπισμό των αλλαγών στις αλληλουχίες γονιδίων που πιο συχνά συνδέονται με συγκεκριμένες ασθένειες.Αυτό κατέστη δυνατόν, διότι όπως ακριβώς και με τα τσιπ των υπολογιστών,στα τσιπ μικροσυστοιχειών μπορούν να τεθούν μεγάλοι αριθμοί δεδομένων (πληροφοριών) που αντιπροσωπεύουν ένα πολύ μεγάλο τμήμα του ανθρώπινου γονιδιώματος. Οι μικροσυστοιχίες μπορούν επίσης να χρησιμοποιηθούν για να μελετηθεί η έκφραση στην οποία ορισμένα γονίδια ενεργοποιούνται η απενεργοποιούνται σε κύτταρα και ιστούς. Σε τέτοιες περιπτώσεις αντί να απομονωθεί DNA από τα δείγματα, απομονώνεται RNA (το οποίο είναι αντίγραφο του DNA) για να μετρηθεί. Σήμερα οι μικροσυστοιχίες DNA χρησιμοποιούνται σε κλινικές διαγνωστικές εξετάσεις για ορισμένες ασθένειες. Η εφαρμογή τους έχει επεκταθεί και στον τομέα της φαρμακολογίας, π.χ. χρησιμοποιούνται για τον προσδιορισμό των φαρμάκων που μπορούν να συνταγογραφηθούν καλύτερα για συγκεκριμένα άτομα, γιατί τα γονίδια καθορίζουν τον τρόπο με τον οποίο το σώμα των ασθενών μπορούν να χειριστούν τα εκάστοτε φάρμακα.

Εφαρμογή μικροσυστοιχειών στην ιατρική.

Για να καθοριστεί εάν ένα άτομο έχει μια μετάλλαξη για μία συγκεκριμένη πάθηση, αρχικά ένας επιστήμονας λαμβάνει δείγμα DNA από το αίμα του ασθενούς καθώς και ένα δείγμα ελέγχου (που δεν περιέχει την μετάλλαξη στο συγκεκριμένο γονίδιο) . ΄Επειτα ο

ερευνητής με μία διαδικασία χωρίζει τους δύο συμπληρωματικούς κλώνους του DNA σε μονόκλωνα μόρια. Επόμενο βήμα είναι να κόψει τις μακριές αλυσίδες του DNA σε μικρότερα κομμάτια και στη συνέχεια ακολουθεί η επισήμανσή τους με μία φθορίζουσα ουσία. Το DNA του ατόμου είναι σημασμένο με πράσινο χρώμα ενώ το κανονικό (χωρίς μετάλλαξη) με κόκκινο χρώμα.Τα δύο σημασμένα DNA εισάγονται εντός του τσιπ και υβριδοποιούνται με συνθετικό DNA πάνω στο τσιπ. Αν το άτομο δεν έχει μετάλλαξη για το γονίδιο,τόσο τα κόκκινα όσο και τα πράσινα δείγματα θα συνδεθούν με τις αλληλουχίες του τσιπ που αντιπροσωπεύουν την αλληλουχία χωρίς την μετάλλαξη. Αν το άτομο έχει την μετάλλαξη,το DNA του ατόμου δεν θα συνδέεται σωστά με τις αλληλουχίες του DNA που αντιπροσωπεύουν την κανονική αλληλουχία (χωρίς μετάλλαξη), αλλά αντ αυτού θα προσδεθεί με την αλληλουχία του τσιπ που αντιπροσωπεύει το μεταλλαγμένο DNA.

Τα θεμελιώδη στάδια για ένα πείραμα μικροσυστοιχειών DNA έχουν ως εξής:

α)Εξασφαλίζουμε RNA από δύο ή περισσότερες ομάδες τις οποίες συγκρίνουμε.

β)Μετατροπή του RNA σε αντιπληροφορικό RNA (αRNA) ή συμπληρωματικό DNA (cDNA).

γ)Το μαρκάρισμά τους με μια φθορίζουσα ουσία.
δ)Η υβριδοποίηση των μαρκαρισμένων cDNA ή αRNA ανάμεσα σε χιλιάδες ανιχνευτές DNA και η ακινητοποίησή του σε μια στέρεη επιφάνεια στήριξης και

ε)Η καταμέτρηση της σχετικής έκφρασης του κάθε γονιδίου σε κάθε ομάδα.

Για την πραγμάτωση ενός τέτοιου πειράματος η πιο διαδεδομένη πλατφόρμα μικροσυστοιχείων σε χρήση είναι τα chip της Affymetrix.

Μέτρα ομοιότητας

Οι αλγόριθμοι συσταδοποίησης βασίζονται σε μαθηματικούς τύπους για να εξακριβωθεί κατά πόσο όμοια ή μη είναι τα στοιχεία που εξετάζουμε. Αν η απόσταση είναι μικρή ανάμεσα σε δύο φορείς έκφρασης τότε θεωρούνται όμοιοι ενώ αντίστροφα αν η απόσταση είναι μεγάλη τότε το επίπεδο ομοιότητας ανάμεσά τους είναι χαμηλό (χαμηλή μικρή συσχέτιση).
Δύο από τις πιο διαδομένες μεθόδους για τον υπολογισμό της σχέσης (χαμηλής-υψηλής)μεταξύ δύο φορέων έκφρασης είναι **η Ευκλείδια απόσταση και ο συντελεστής συσχέτισης Pearson.**

Ο τύπος της **Ευκλείδιας απόστασης** δίνεται παρακάτω

$$d_{A:B} = \sqrt{\sum_{i=1}^{n} (x_{i-} y_i)^2}$$

Για παράδειγμα σε ένα ορθοκανονικό σύστημα αξόνων δίνονται δύο σημεία Α(x_1, y_1) και Β(x_2, y_2). Ο υπολογισμός της απόστασής τους δίνεται από τον τύπο

$$d_{A:B} = \sqrt{(x_{1-} y_1)^2 + (x_{2-} y_2)^2}$$

x_i =log ratio του φορέα έκφρασης Α
y_i=log ratio του φορέα έκφρασης Β
Η μικρότερη απόσταση είναι η απόσταση ενός φορέα έκφρασης από τον ευαυτό του και ειναι 0. Αντίθετα δεν υπάρχει άνω όριο για την μεγαλύτερη τιμή της απόστασης.

Για να μετρήσουμε τα επίπεδα ομοιότητας της μορφής δυο φορέων έκφρασης χρησιμοποιούμε επίσης τον **συντελεστή συσχέτησης Pearson** r. Η τιμή του δίνεται από τον τύπο

$$r = \frac{\sum_{i=1}^{v}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{v}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{v}(y_i - \bar{y})^2}}$$

x_i=logratio του φορέα A
y_i=logratio του φορέα B
$\bar{x}$=μέση τιμή του logratio του φορέα A
$\bar{y}$=μέση τιμή του logratio του φορέα B

Οι τιμές του r κυμαίνονται από το -1 έως το 1. Μία τιμή κοντά στο 1 σημαίνει υψηλή συσχέτιση (επίπεδο ομοιότητας υψηλό) με ίδια μορφή. Μία τιμή κοντά στο 0 σημαίνει ότι υπάρχει χαμηλή έως καθόλου (τιμή 0) συσχέτιση ανάμεσα στους δυο φορείς. Μία τιμή κοντά στο -1 σημαίνει ισχυρή αλλά αντίθετη συσχέτιση.

Πρόβλεψη τάξεων.

Σκοπός των μελετών πρόβλεψης τάξεων είναι η ταξινόμηση των δειγμάτων σε προκαθορισμένες τάξεις βάση του γονιδιακού προφίλ(Tarca,Romero,Draghici).Έτσι χρησιμοποιούνται αλγόριθμοι ταξινόμησης , οι οποίοι βασιζόμενοι σε εξωγενή πληροφορία(πχ. Με βάση έναν πίνακα δεδομένων γονιδιακής έκφρασης) και αναζητούν στα δεδομένα ιδιότητες που να την υποστηρίζουν.

Αλγόριθμος k-εγγύτερου γείτονα (nearest neighbour analysis)

Ο αλγόριθμος k-εγγύτερου γείτονα είναι ένας από τους πιο απλούς αλγόριθμους που χρησιμοποιείται για την διάκριση ανάμεσα σε τάξεις των δεδομένων γονιδιακής έκφρασης(Cover,Hart).Τα δεδομένα (στοιχεία γονιδιακής έκφρασης) αναπαρίστανται σε ένα χώρο δεδομένων έκφρασης.Έπειτα γίνεται η αποθήκευση των στοιχείων καθώς και των τάξεων στις οποίες ανήκουν.Κατά το στάδιο της ταξινόμησης κάθε νέο στοιχείο γονιδιακής έκφρασης ταξινομείται

καθορίζοντας την Ευκλείδια απόστασή του από τους γείτονές του στο χώρο δεδομένων έκφρασης.

Ανακάλυψη τάξεων.

Συσταδοποίηση – Τύποι συσταδοποίησης.

Η μέθοδος της συσταδοποίησης στοχεύει στην ανακάλυψη τάξεων-ομάδων παρόμοιων αντικειμένων και ουσιαστικά στην εύρεση συσχετίσεων σε μεγάλου πλήθους δείγματα. Επί της ουσίας η μέθοδος της συσταδοποίησης χωρίζει τα δεδομένα που δόθηκαν σε ομάδες-συστάδες έτσι ώστε τα δεδομένα σε κάθε ομάδα να έχουν περισσότερες ομοιότητες μεταξύ τους παρά σε άλλες ομάδες. Πχ. Σε ένα εμπορικό κατά μέσο όρο ψωνίζουν 10.000 άνθρωποι την ημέρα. Για μια συγκεκριμένη μέρα θα πάρουμε τον αριθμό των πελατών και θα τους χωρίσουμε σε κατηγορίες,για παράδειγμα πελάτες που αγόρασαν τρόφιμα, ρούχα, είδη υγειηνής κ.λ.π. κ.λ.π.Μια μέθοδος συσταδοποίησης θα ήταν να χωρίσουμε τους πελάτες σε ομάδες που έχουν τις ίδιες-παρόμοιες αγορές στο εμπορικό κέντρο.

Τα βασικά βήματα για την συσταδοποίηση των δεδομένων ενός πειράματος είναι τα εξής:

α) Επιλογή των χαρακτηριστικών πάνω στις οποίες θα επιλέξουμε να εφαρμόσουμε την παραπάνω μέθοδο ομαδοποίησης και να κωδικοποιήσουμε όσο το δυνατόν περισσότερες πληροφορίες ανάλογα με την ανάλυση που μας ενδιαφέρει.

β) Η κατάλληλη επιλογή του αλγόριθμου συσταδοποίησης που θα χρησιμοποιηθεί το οποίο εξαρτάται από την εγγύτητα των μέτρων και από το κριτήριο συσταδοποίησης.

γ) Η αξιολόγηση των αποτελεσμάτων η οποία γίνεται με εφαρμογή κατάλληλων κριτηρίων και από συγκεκριμένες τεχνικές.

δ) Η σωστή ερμηνεία των αποτελεσμάτων που συνήθως γίνεται με την σύγκριση των αποτελεσμάτων με άλλα επιστημονικά στοιχεία και αναλύσεις(Halkidi,Batistaskis,Vazirgiannis).

Οι αλγόριθμοι συσταδοποίησης.

Θα αναφερθούν αρχικά ονομαστικά και έπειτα θα αναλυθούν οι δυο πιο δημοφιλείς για την ανάλυση των δεδομένων των μικροσυστοιχείων.

1) Διαιρετικοί αλγόριθμοι συσταδοποίησης.

2) Ιεραρχική συσταδοποίηση.

3) Συσταδοποίηση με βάση την πυκνότητα των δεδομένων.

4) Συσταδοποίηση πλέγματος.

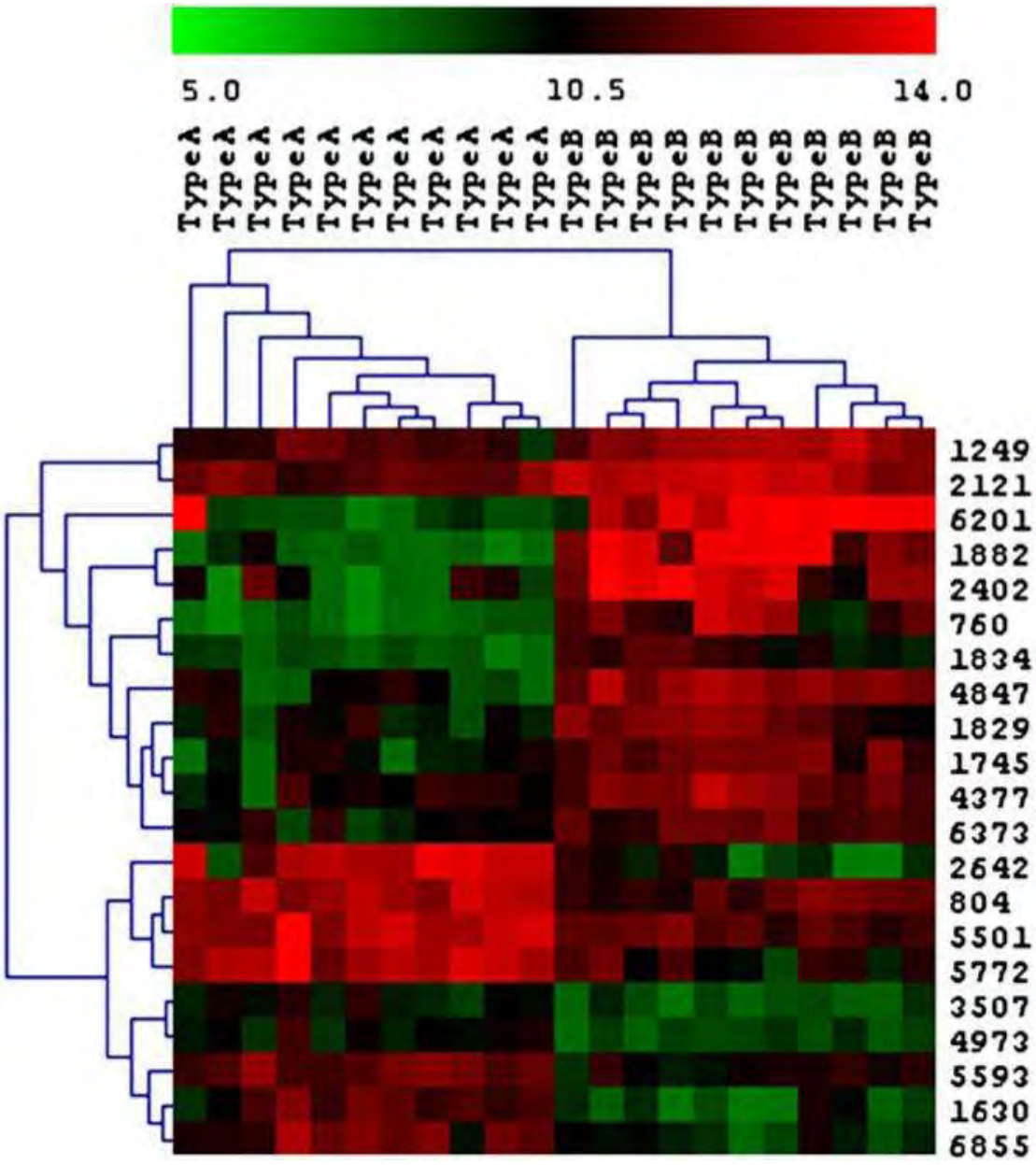
Οι δύο πιο διαδεδομένοι αλγόριθμοι που χρησιμοποιούνται στην ανάλυση των μικροσυστοιχειών είναι οι διαιρετικοί και η ιεραρχική συσταδοποίηση.

Ιεραρχική Συσταδοποίηση.

Η ιεραρχική συσταδοποίηση αποτελεί (**Hierarchical clustering**) μία από τις πιο διαδεδομένες μεθόδους ανάλυσης μικροσυστοιχείων καθώς προσφέρει ένα διαισθητικό οπτικό αποτέλεσμα στην μορφή ενός δενδρογράμματος και διορατηκότητα του επιπέδου της σχέσης που έχουν τα στοιχεία μεταξύ τους(H.Lee,I.Saeed). Συγκεκριμένα η μέθοδος της ιεραρχικής συσταδοποίησης μπορεί και απλοποιεί μεγάλο πληθυσμό δεδομένων, ανακαλύπτει ομάδες γονιδίων με παραπλήσια έκφραση και μας δίνει μια απεικόνιση των δεδομένων σε ιεραρχική δομή.

Η ιεραρχική συσταδοποίηση μπορεί να είναι είτε συσσωρευτική είτε διαιρετική, αλλά σε ανάλυση δεδομένων μικροσυστοιχείων η συσσωρευτική μέθοδος είναι η επικρατέστερη της διαιρετικής. Συγκεκριμένα ο αλγόριθμος (συσσωρευτής συσταδοποίησης) παίρνει ένα σύνολο ανεξάρτητων στοιχείων και προοδευτικά τα συνενώνει σε μεγαλύτερης βαθμίδας συστάδες. Το πρωταρχικό βήμα για την δημιουργία ενός δενδρογράμματος είναι ο υπολογισμός της απόστασης μεταξύ των στοιχείων για να εξακριβωθεί ποια στοιχεία έχουν τον υψηλότερο βαθμό συσχέτισης μεταξύ τους. ΄Ετσι δημιουργείται ένας nxn πίνακας που δείχνει τις αποστάσεις μεταξύ των n στοιχείων που έχουν επιλεχθεί. Αφού δημιουργηθεί αυτός ο πίνακας nxn κάθε

στοιχείο θεωρείται ως μία ξεχωριστή συστάδα η οποία λαμβάνει τιμή έκφρασης την τιμή του εκάστοτε στοιχείου.



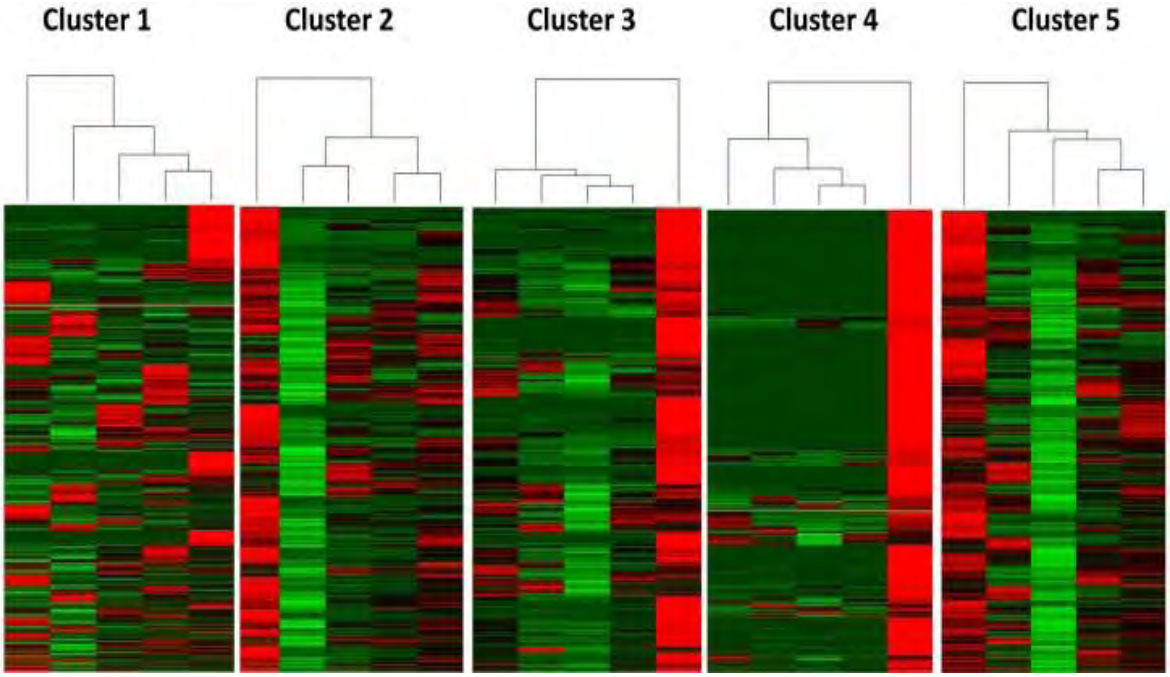
Εικόνα 1:Γραφική απεικόνηση ιεραρχικής συσταδοποίησης δεδομένων.Μικρότερα κλαδιά δείχνουν μικρότερες αποστάσεις ανάμεσα στους φορείς έκφρασης και κατά συνέπεια μεγαλύτερη συσχέτηση.

Τα βήματα έχουν ως εξής:
i) Εξακριβώνουμε ποιές (δύο) συστάδες έχουν τον υψηλότερο βαθμό ομοιότητας υπολογίζοντας την μικρότερη απόσταση που φαίνεται από τον πίνακα.
ii) Συνένωση αυτών των δύο συστάδων σε μία συστάδα υψηλότερης βαθμίδας.
iii)Επαναπροσδιορίζουμε τις αποστάσεις της νέας συστάδας με όλες τις υπόλοιπες και υλοποιείται με μια προκαθορισμένη μέθοδο γνωστή ως κανόνας συνδέσμου (linkage method).
iv)΄Επειτα αναζητούμε την αμέσως μικρότερη απόσταση και επιστρέφουμε στο βήμα i.

΄Οσον αφορά τους κανόνες συνδέσμου που θα χρησιμοποιηθούν για τον υπολογισμό της απόστασης μεταξύ των συστάδων υπάρχει **α)Ο απλός κανόνας συνδέσμου** όπου παίρνουμε την μικρότερη απόσταση ανάμεσα στα στοιχεία των συστάδων (τον μεγαλύτερο βαθμό του μέτρου ομοιότητας).**β)Κανόνας πλήρους συνδέσμου** όπου έχουμε ακριβώς την αντίθετη διαδικασία με τον κανόνα απλού συνδέσμου καθώς εδώ παίρνουμε την μεγαλύτερη απόσταση ανάμεσα στα στοιχεία των συστάδων.**γ)Κανόνας μέσου συνδέσμου** όπου παίρνουμε τον αριθμητικό μέσο των αποστάσεων των στοιχείων των δυο συστάδων. Το αποτέλεσμα αυτού του αλγόριθμου αναπαρειστάται γραφικά από ένα δενδρόγραμμα.
Συνοψίζοντας, η ιεραρχική συσταδοποίηση έχει και τα αρνητικά της όπως το ότι αδυνατεί να αναπαραστήσει τους πολλαπλούς τρόπους με τους οποίους μπορούν να μοιάζουν τα στοιχεία γονιδιακής έκφρασης μεταξύ τους (Russel Meadows, Russel 2009) όπως και η πλειονότητα των μεθόδων που εκτιμούν τα μέτρα ομοιότητας ανάμεσα στα στοιχεία γονιδιακής έκφρασης.Η ιεραρχική συσταδοποίηση θεωρείται κυρίως ως μία μέθοδος προκαταρκτικής ανάλυσης και γραφικής αναπαράστασης των δεδομένων. Παρά τις αδυναμίες της χρησιμοποιήθηκε με επιτυχία σε πολλές μελέτες.

Συσταδοποίηση k-μέσων.

Σε περίπτωση που έχει προκαθοριστεί ένας αριθμός συστάδων στα οποία θα ομαδοποιηθούν τα στοιχεία της γονιδιακής έκφρασης τότε χρησιμποιείται ο αλγόριθμος συσταδοποίησης κ – μέσων.Στόχος του αλγόριθμου είναι η ομαδοποίηση μέσω διαμερισμού των στοιχείων γονιδιακής έκφρασης στον προκαθορισμένο αριθμό κ –συστάδων. Η τιμή του κ πρέπει να έχει ήδη καθοριστεί από τον χρήστη πριν την έναρξη της διαδικασίας.



Εικόνα 2.Πηγή:https://www.researchgate.net/:Γραφική απεικόνηση συσταδοποίησης k-μέσων(k=5).Δύο στοιχεία μέσα στην ίδια συστάδα έχουν μεγαλύτερη ομοιότητα μεταξύ τους σε σχέση με δύο στοιχεία από διαφορετικές συστάδες που έχουν μικρότερη ομοιότητα.

Ο αλγόριθμος κ – μέσων αποτελείται από τα παρακάτω βήματα.
1) Επιλέγεται ένας τυχαίος αριθμός k στοιχείων που αντιπροσωπεύει μια πρώτη προσέγγιση των κεντροειδών των συστάδων που θα σχηματιστούν.
2) ΄Ενα κεντροειδές χρησιμοποιείται για την εκπροσώπηση της κάθε συστάδας.Έπειτα υπολογίζοντας την μέση τιμή της απόστασης (μέτρα ομοιότητας) κάθε στοιχείου από τα κεντροειδή των k-συστάδων τα

αντιστοιχίζουμε στη συστάδα με της οποίας το κεντροειδές μοιάζει περισσότερο(συνήθως χρησιμοποιείται η ευκλείδια απόσταση ως μέτρο ομοιότητας).
3) Εκτελούμε τα βήματα 1, 2 για κάθε στοιχείο, με τη σειρά. Μία εννιαία επανάληψη αυτού του σταδίου περιλαμβάνει την αξιολόγηση όλων των στοιχείων.
α) Επιλογή ενός στοιχείου και εύρεση της συστάδας με την οποία υπάρχει ο υψηλότερος βαθμός ομοιότητας. Εάν τα στοιχείο δεν είναι ήδη μέλος της συστάδας μετακινείται εκεί.
β) Εάν κάποιο στοιχείο μετακινήθηκε από μία συστάδα σε μια άλλη επαναπροσδιορίζουμε τα μέτρα ομοιότητας των στοιχείων στην συστάδα από την οποία μετακινήθηκε το στοιχείο και στην συστάδα στην οποία μετακινήθηκε. ΄Επειτα συνεχίζουμε από το α.
4) Εάν δεν μετακινήθηκαν στοιχεία στην πιο πρόσφατη επανάληψη του βήματος 3 τότε όλα τα στοιχεία βρίσκονται στην ιδανικότερη συστάδα και ο αλγόριθμος έχει τελειώσει. Το αποτέλεσμα του αλγόριθμου είναι μια ομάδα απο κ – συστάδες από τις οποίες κάθε μία περιέχει στοιχεία τα οποία μοιάζουν περισσότερο με την μέση τιμή του φορέα έκφρασης που έχει χρησιμοποιηθεί στην εκάστοτε συστάδα.
Δύο από τις αδυναμίες που παρουσιάζει η συσταδοποίηση κ – μέσων είναι η επιλογή του αριθμού κ από τον χρήστη, καθώς και η τυχαία θέση των κεντροειδών στην κάθε συστάδα.Πιο συγκεκριμένα η επιλογή της αρχικής θέσης των κεντροειδών μπορεί να επηρεάζουν το τελικό αποτέλεσμα του αλγόριθμου. Αυτό έχει ως επακόλουθο την επανάληψη της όλης διαδικασίας ορισμένες φορές (κάθε φορά με χρήση των αρχικών κεντροειδών σε διαφορετικές θέσεις) και στο τέλος επιλέγεται η συσταδοποίηση για την οποία πληρείται κάποιο κριτήριο.

Αποτελέσματα.

Ακολουθεί ανάλυση δεδομένων (προσομείωσης) με Ιεραρχική και συσταδοποίηση κ-μέσων. Η ανάλυση θα γίνει μέσω του στατιστικού πακέτου SPSS.

Single linkage **Complete linkage** **Average linkage**

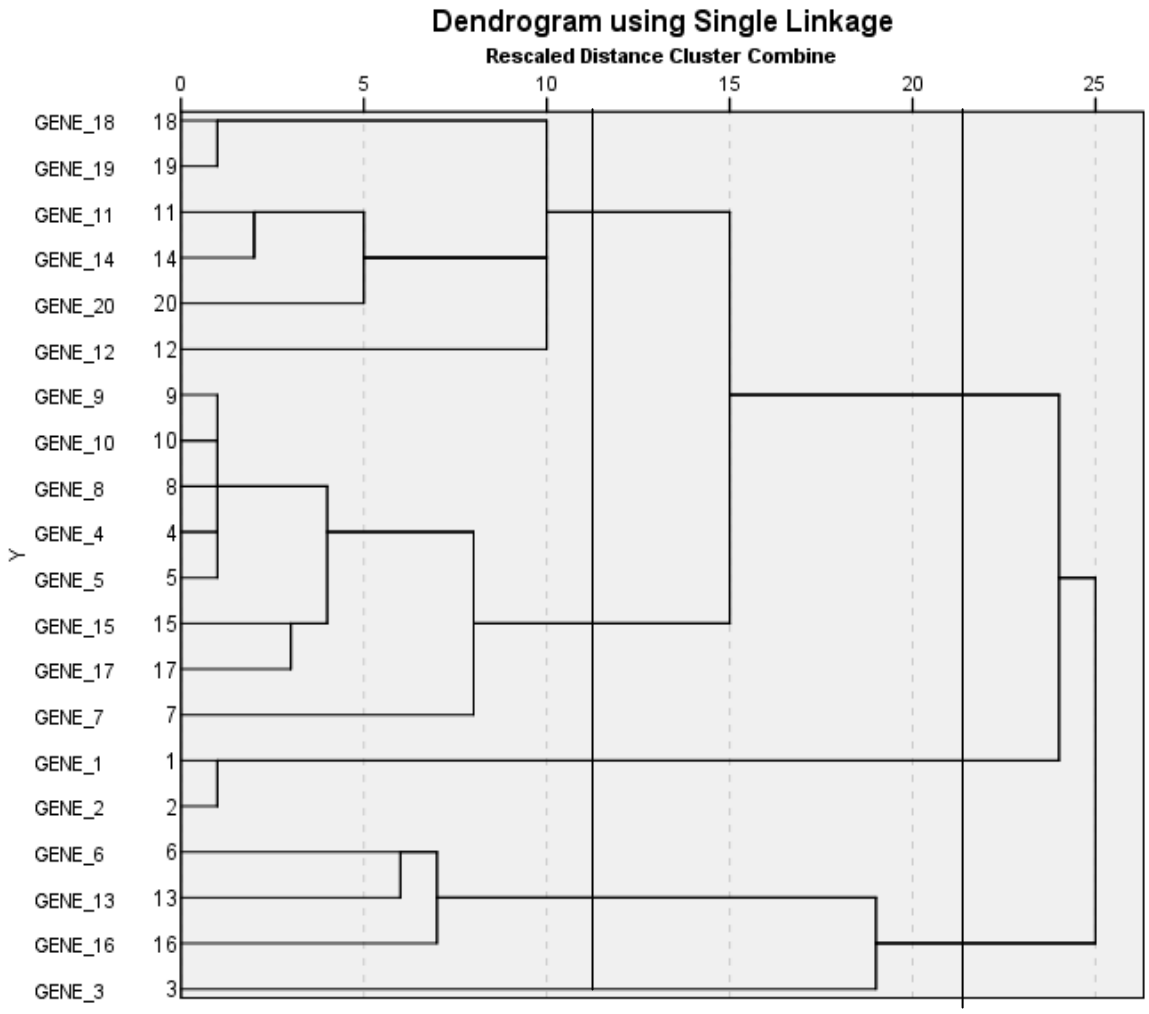
Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	18	19	,000
2	9	10	,000
3	8	9	,000
4	4	5	,000
5	1	2	,000
6	4	8	,240
7	11	14	1,302
8	15	17	1,380
9	4	15	2,610
10	11	20	3,000
11	6	13	3,620
12	6	16	4,340
13	4	7	4,650
14	11	12	6,050
15	11	18	6,060
16	4	11	9,770
17	3	6	12,000
18	1	4	15,200
19	1	3	16,420

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	18	19	,000
2	9	10	,000
3	8	9	,000
4	4	5	,000
5	1	2	,000
6	4	8	,144
7	4	15	1,110
8	11	14	1,302
9	11	20	2,442
10	4	7	2,522
11	6	13	3,620
12	12	18	4,040
13	4	17	4,138
14	6	16	4,473
15	1	12	10,708
16	3	6	11,710
17	4	11	15,902
18	1	3	28,186
19	1	4	83,598

Stage	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	18	19	,000
2	9	10	,000
3	8	9	,000
4	4	5	,000
5	1	2	,000
6	4	8	,240
7	11	14	1,302
8	15	17	1,380
9	11	20	3,022
10	6	13	3,620
11	4	7	4,690
12	6	16	5,460
13	12	18	6,060
14	1	12	16,140
15	4	15	22,460
16	3	6	30,340
17	1	11	45,890
18	1	4	172,340
19	1	3	402,740

Για την ανάλυση χρησιμοποιήθηκε ο κανόνας **απλού συνδέσμου (single linnage)** καθώς και οι **κανόνες πλήρους(complete linkage)** και **μέσου συνδέσμου(average linkage)** και σαν μέτρο απόστασης η **Ευκλείδια απόσταση.** ΄Οπως φαίνεται από τους παραπάνω πίνακες τα στοιχεία με την μεγαλύτερη ομοιότητα είναι τα ακόλουθα ζευγάρια (18,19),(9,10),(8,9),(4,5),(1,2),(4,8).Οι διαφορές φαίνονται παρακάτω στους πίνακες καθώς τα ζευγάρια αλλάζουν.Για τον κανόνα απλού συνδέσμου τα ζευγάρια που ακολουθούν είναι (11,14) (15,17) (4,15) (11,20) (6,13) (6,16) (4,7) (11,12) (11,18) (4,11) (3,6) (1,4) (1,3) ,για τον κανόνα πλήρους συνδέσμου (4,15) (11,14) (11,20) (4,7) (6,13) (12,18) (4,17) (6,16) (1,12) (3,6) (4,11) (1,3) (1,4) και για τον κανόνα μέσου συνδέσμου είναι (11,14) (15,17) (11,20) (6,13) (4,7) (6,16) (12,18) (1,12) (4,15) (3,6) (1,11) (1,4) (1,3).

Το μεγάλο πλεονέκτημα της Ιεραρχικής συσταδοποίησης είναι η απεικόνιση των δεδομένων μεσω του δενδρογράμματος **(εικόνα 3-4-5).**Η χρησιμότητα του δενδρογράμματος δίνει στον ερευνητή την δυνατότητα να ανακαλύψει συστάδες ή ποιές παρατηρήσεις έχουν συσταδοποιηθεί μαζί και ομοιάζουν περισσότερο μεταξύ τους.



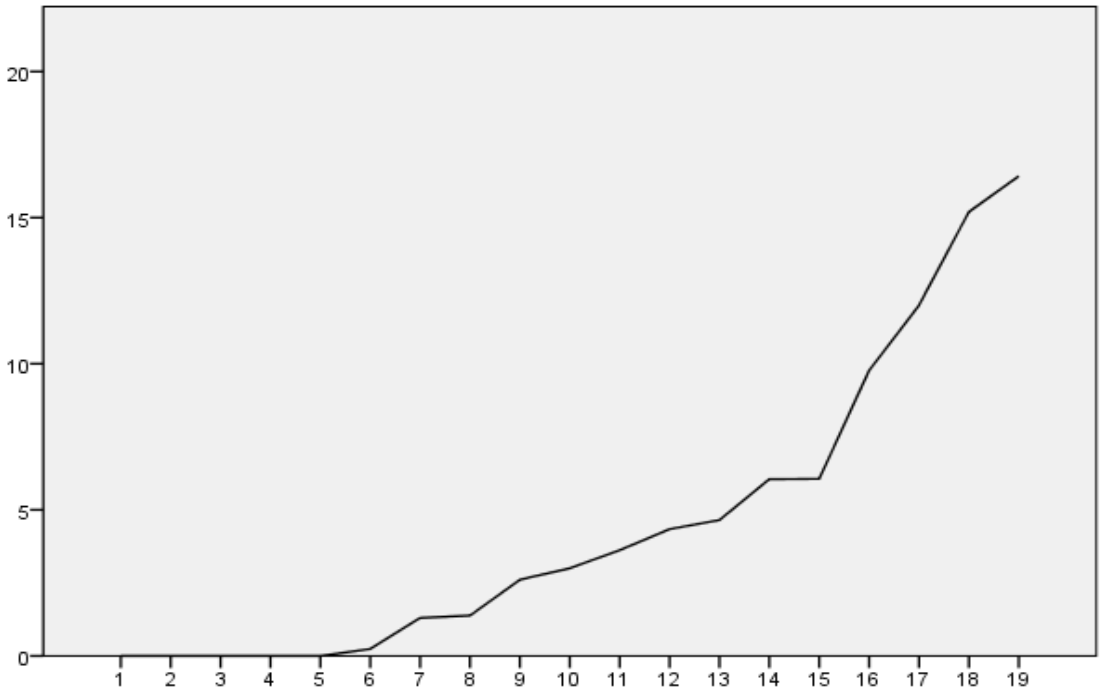
Εικόνα 3.

Το παραπάνω δενδρόγραμμα(με την μέθοδο του απλού συνδέσμου) το έχουμε χωρίσει με δύο κάθετες γραμμές. Η πρώτη γραμμή (ανάμεσα σε 20-25) χωρίζει το δενδρόγραμμα σε τρεις συστάδες $1^{η}$ (18,19,11,14,20,12,9,10,8,4,5,15,17,7), $2^{η}$ (1,2), $3^{η}$ (6,13,16,3) των οποίων τα στοιχεία ομοιάζουν μεταξύ τους. Η δεύτερη γραμμή (ανάμεσα 10-15)μας δίνει 5(που είναι και ο ιδανικός αριθμός) συστάδες $1^{η}$ (18,19,11,14,20,12), $2^{η}$ (9,10,8,4,5,15,17,7), $3^{η}$ (1,2), $4^{η}$ (6,13,16), $5^{η}$ (3) των οποίων τα στοιχεία ομοιάζουν ακόμη πιο πολύ μεταξύ τους καθώς φαίνεται και το επίπεδο συσχέτισης μεταξύ τους.

Μπορεί να είναι δύσκολο να υπολογιστούν οι διαφορές της συντελεστών αλλά αυτό γίνεται εύκολα με ένα scree plot.Στο παρακάτω γράφημα μετά την τιμή 15 φαίνεται μια μεγάλη αύξηση των

συντελεστών.Με τον ίδιο τρόπο εργαστήκαμε και στα άλλα δύο γραφήματα που ακολουθούν.

Scree plot for single linkage method

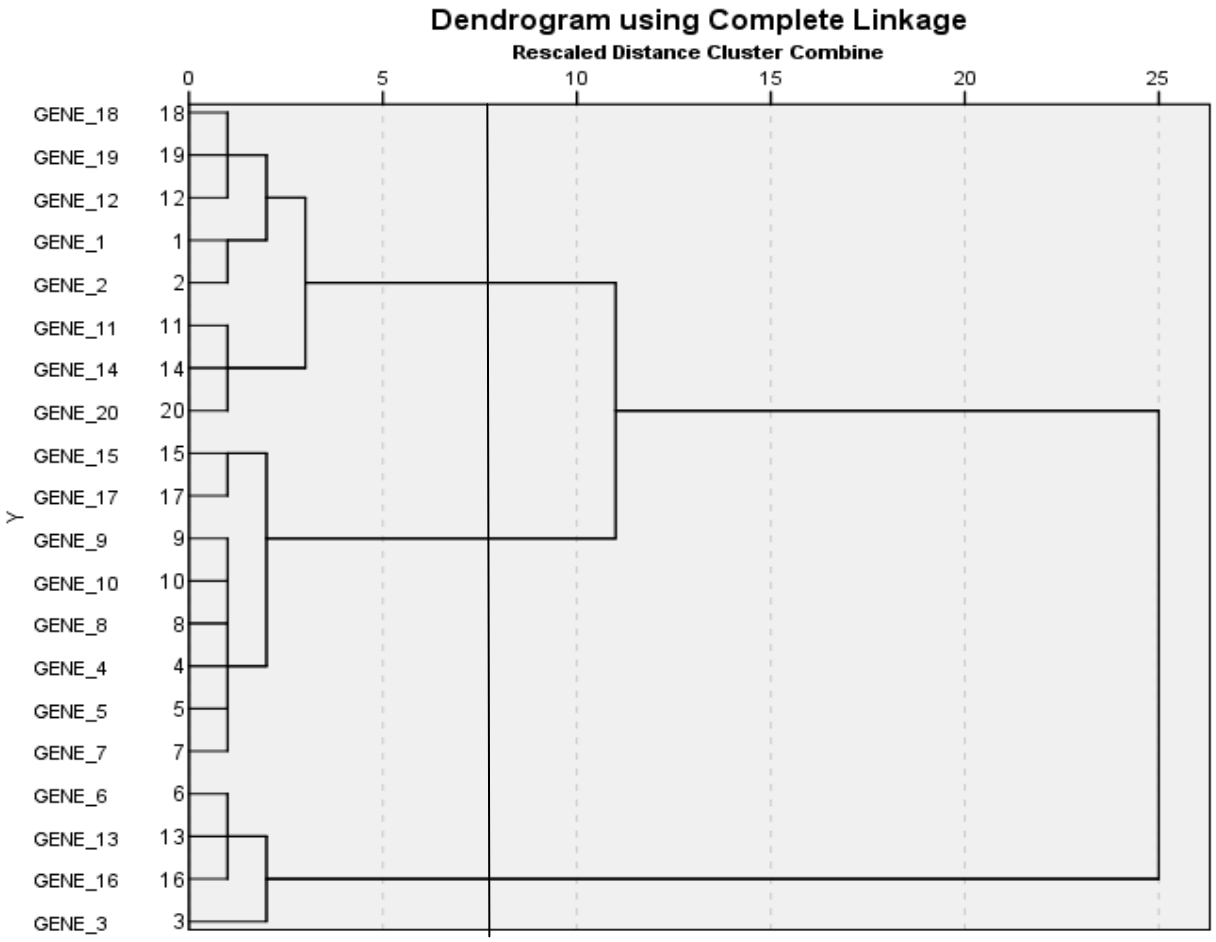


Την ίδια διαδικασία ακολουθήσαμε για το δενδρόγραμμα με την μέθοδο του πλήρους συνδέσμου και προέκυψαν οι εξής συστάδες

$1^{η}$ (18,19,12,1,2, 11,14,20)

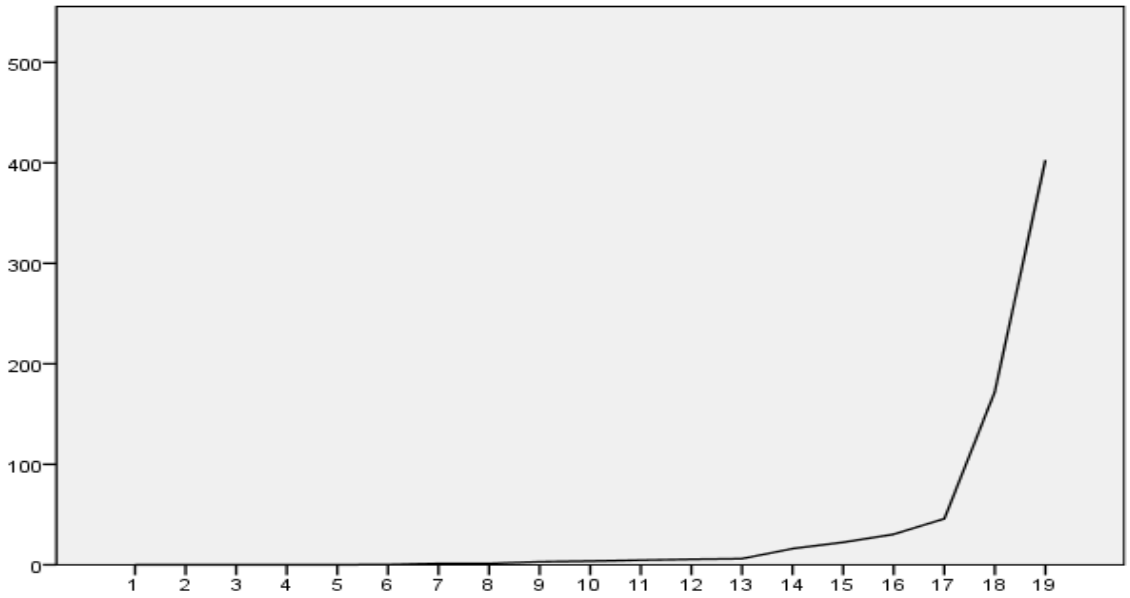
$2^{η}$ (4,5,7,8,9,10,15,17)

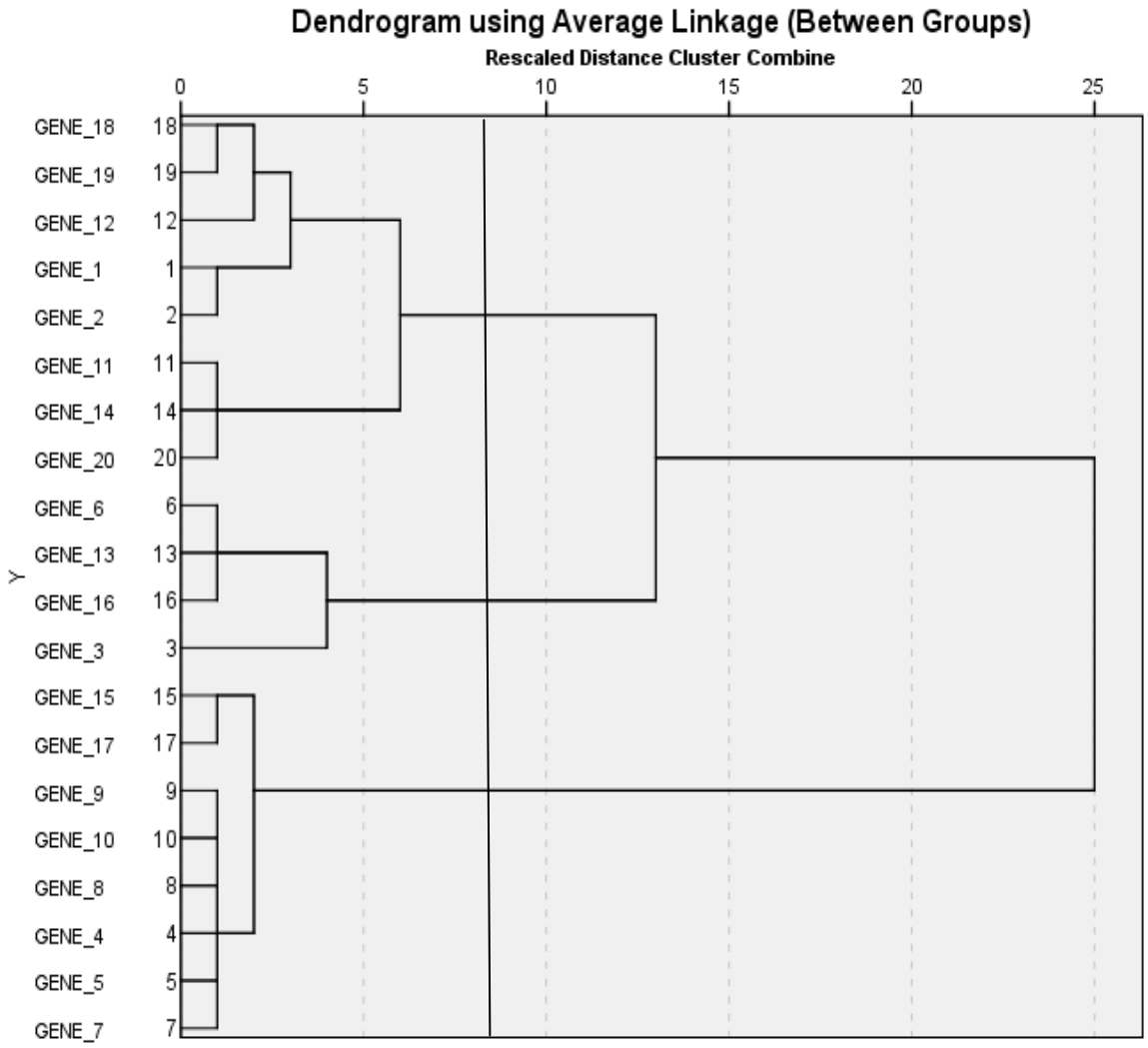
$3^{η}$ (6,13,16,3)



Εικόνα 4.

Scree plot for complete linkage method





Εικόνα 5.

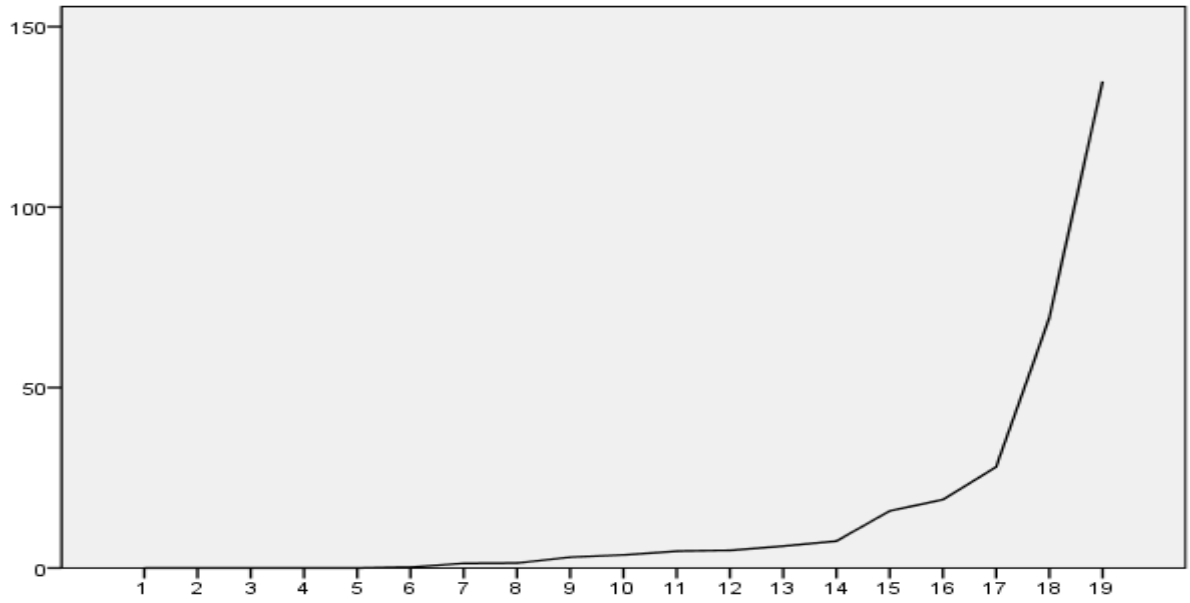
Παρόμοια από το δενδρόγραμμα με την μέθοδο του μέσου συνδέσμου προέκυψαν οι παρακάτω συστάδες

$1^{η}$ (18,19,12,1,2, 11,14,20)

$2^{η}$ (6,13,16,3)

$3^{η}$ (4,5,7,8,9,10,15,17)

Scree plot for average linkage method



K-means algorithm.

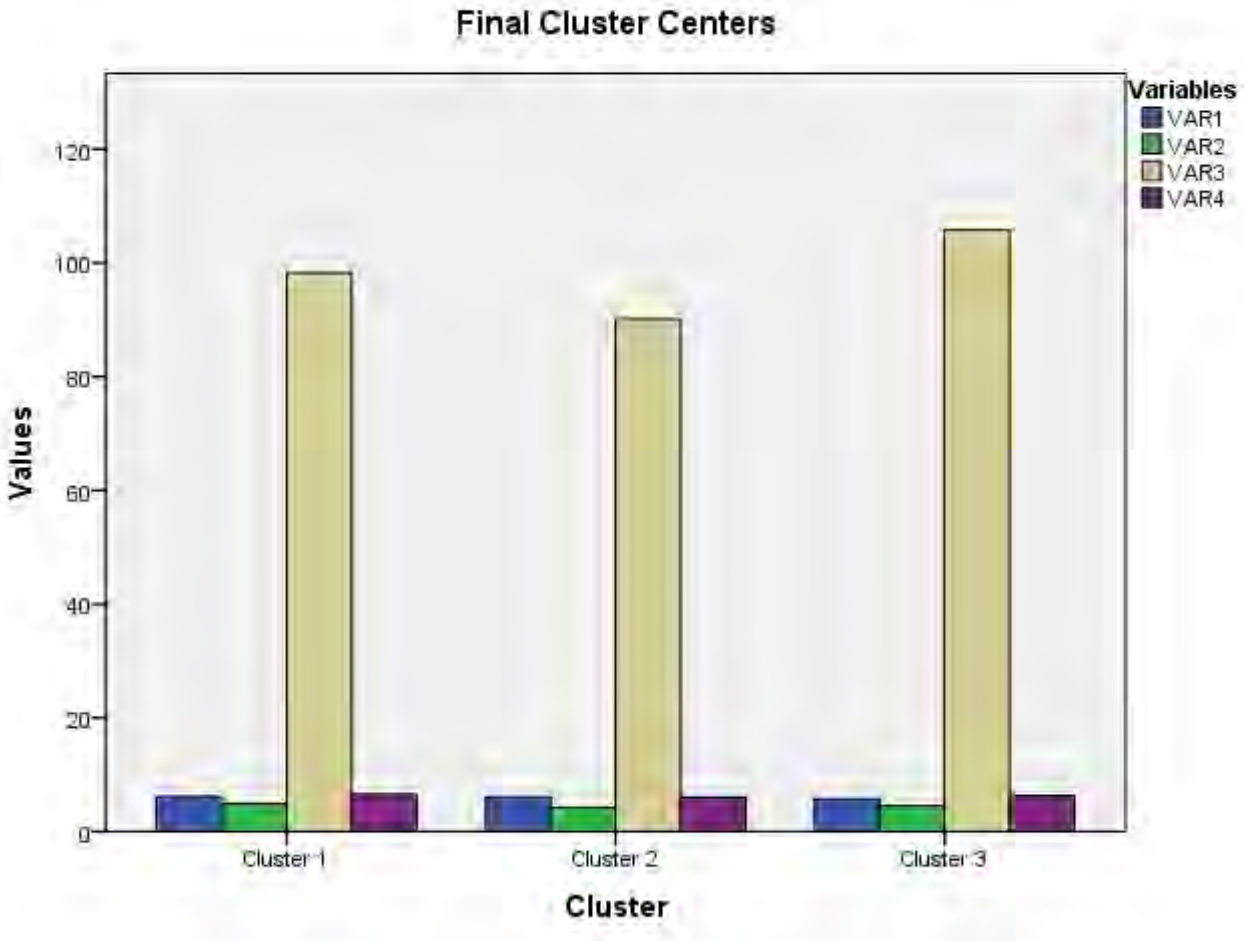
Case Number	Name	Cluster	Distance
1	GENE_1	1	3,676
2	GENE_2	1	3,676
3	GENE_3	3	3,133
4	GENE_4	2	,623
5	GENE_5	2	,623
6	GENE_6	3	,937
7	GENE_7	2	2,353
8	GENE_8	2	,280
9	GENE_9	2	,280
10	GENE_10	2	,280
11	GENE_11	1	3,256
12	GENE_12	1	,272
13	GENE_13	3	1,106
14	GENE_14	1	3,394
15	GENE_15	2	1,372
16	GENE_16	3	2,376
17	GENE_17	2	2,460
18	GENE_18	1	2,283
19	GENE_19	1	2,283

20	GENE_20	1	2,680

Με την χρήση του SPSS αναλύσαμε τα ίδια δεδομένα χρησιμοποιώντας συσταδοποίηση κ-μέσων (k = $\sqrt{N/2}$) και από τον παραπάνω πίνακα φαίνεται η ταξινόμηση του κάθε στοιχείου στην εκάστοτε συστάδα. Στην $1^{η}$ συστάδα περιέχονται τα στοιχεία [1,2,11,12,14,18,19,20] στην $2^{η}$ [4,5,7,8,9,10,15,17] και στην $3^{η}$ [3,6,13,16]. Τα στοιχεία που βρίσκονται σε κάθε συστάδα ομοιάζουν περισσότερο μεταξύ τους.

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
VAR1	,242	2	1,258	17	,192	,827
VAR2	,947	2	,261	17	3,622	,049
VAR3	345,938	2	3,831	17	90,302	,000
VAR4	,527	2	,466	17	1,130	,346

Από τον παραπάνω πίνακα της Anova φαίνεται ότι μόνο η μεταβλητή VAR3 (ρ=0,001) έχει στατιστικά σημαντικό ρόλο στον διαχωρισμό και την δημιουργία των παραπάνω συστάδων. Το ίδιο συμπέρασμα μπορεί να βγει και από το ραβδόγραμμα που ακολουθεί για την κάθε συστάδα(**εικόνα4**).



Εικόνα 6.

Σύγκριση αποτελεσμάτων.

Από τα παραπάνω αποτελέσματα των μεθόδων, η μέθοδοι του πλήρους και του μέσου συνδέσμου καθώς και οι συστάδες που προέκυψαν από την συσταδοποίηση k-μέσων είναι ακριβώς ίδιες [(1,2,11,12,14,18,19,20) , (4,5,7,8,9,10,15,17) , (3,6,13,16)] ενώ από την μέθοδο του απλού συνδέσμου προέκυψαν 5 συστάδες {$1^{η}$ (18,19,11,14,20,12), $2^{η}$ (9,10,8,4,5,15,17,7), $3^{η}$ (1,2), $4^{η}$ (6,13,16), $5^{η}$ (3) }.Η τελευταία είναι και η καταλληλότερη μέθοδος καθώς μας δίνει καλύτερη εικόνα για την συσχέτιση των δεδομένων.

Αξιολόγηση των αποτελεσμάτων.

Ο απότερος σκοπός ενός πειράματος μικροσυστοιχείων είναι κυρίως η κατανόηση των μηχανισμών ενός βιολογικού συστήματος.
Το αποτέλεσμα μιας μελέτης π.χ. σύγκρισης τάξεων, αν και πρόκειται για μία ορθή αρχική προσέγγιση, δεν συνεισφέρει ιδιαιτέρως προς την κατεύθυνση αυτή(Irizarry 2009). Γι αυτό, η ανάλυση δεδομένων μικροσυστοιχειών όπως οι μελέτες ανακάλυψης τάξεων πρέπει να συνοδεύονται από μία κατάλληλη δευτερογενή ανάλυση, η οποία ενσωματώνει βιολογική γνώση(Tarca,Romero,Draghisi 2006) . Για να πραγματοποιηθεί κάτι τέτοιο, δηλαδή την κατανόηση των μηχανισμών ενός βιολογικού συστήματος έχουν αναπτυχθεί οι εξής μέθοδοι, γνωστές και ως μέθοδοι ανάλυσης κατηγοριών γονιδίων. Πρόκειται για στατιστικούς ελέγχους, που αποδεικνύουν αν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ δύο δειγμάτων μιας προκαθορισμένης κατηγορίας γονιδίων. Οι μέθοδοι αυτές χωρίζονται σε δύο κύριες κατηγορίες.
α) την ανάλυση υπερεκπροσώπησης (Over-Representation Analysis ORA)
β) την ανάλυση συναθροιστικού βαθμού (Aggeregate Score Analysis – ASA).

Ανάλυση υπερ-εκπροσώπησης.

Με την συγκεκριμένη μέθοδο θεωρούμε ως είσοδο ένα κατάλογο διαφορικώς εκφραζόμενων γονιδίων και για κάθε κατηγορία γονιδίων ελέγχει εάν ο αριθμός των διαφορικώς εκφραζομένων γονιδίων της συγκεκριμένης κατηγορίας είναι σημαντικά μεγαλύτερος (ή μικρότερος) από αυτόν που θα αναμενόταν αν τα γονίδια είχαν επιλεχθεί τυχαία

(Russel Meadows, Russel 2009). Για την πραγμάτωση αυτού του ελέγχου,δηλαδή για την εξακρίβωση των στατιστικά σημαντικών κατηγοριών γονιδίων χρησιμοποιείται παραλλαγή του ελέγχου χ2 σε πίνακες συνάφειας που είναι και ο πιο αξιόπιστος για έλεγχο μεγάλου αριθμού γονιδίων (Rivals 2007).

Ανάλυση συναθροιστικού βαθμού.

Οι δύο πιο διαδεδομένες μέθοδοι είναι α)η ανάλυση εμπλουτισμού κατηγοριών γονιδίων (Gene Set Enrichment Analysis – GSEA) (Mootha 2003,Subramanian 2005), β)η ανάλυση σημαντικότητας της λειτουργίας και έκφρασης (significasnce Analysis of Function and Expression – SAFE) (Barry, Nobel and Wright 2005). Γενικότερα και οι δύο μέθοδοι αξιολογούν την κατανομή των γονιδίων μίας κατηγορίας στην κατάταξη αυτή, υπολογίζοντας στατιστικές συσσωρευτικού αθροίσματος.

Κατηγορίες γονιδίων

Επίσης πολλές μελέτες χρησιμοποιούν προκαθορισμένες κατηγορίες γονιδίων οι οποίες έχουν προκύψει από παλαιότερες μελέτες.Οι κατηγορίες αυτές είναι διαθέσιμες σε δημόσιες βάσεις βιολογικών δεδομένων.Η αντιστοίχηση γονιδίων γίνεται σε όρους γονιδιακής οντολογίας ,σε βιοχημικές οδούς...(Russel,Russel and Meadows 2009). Η γονιδιακή οντολογία(gene ontology) συμβάλει στον υπομνηματισμό των γονιδίων και των προϊόντων τους, συγκεκριμένα μας περιγράφει τις ιδιότητες των γονιδίων και των προϊόντων τους ανάμεσα στις διαφορετικές βάσεις δεδομένων μέσω μιας κοινής γλώσσας.

Σύνθεση μελετών

Οι μελέτες που αφορούν πειράματα μικροσυστοιχειών είναι γνωστό για την μειωμένη επαναληπτικότητά τους.Για αυτό και σημαντικό ρόλο στις συγκεκριμένες μελέτες αποτέλει η σύνθεση μελετών, καθώς συμβάλει στην διεξαγωγή μελετών χρησιμοποιώντας δεδομένα και στοιχεία από

παλαιότερες δημοσιευμένες μελέτες.Αυτό βοηθά στην αύξηση της επαναληπτικότητας αλλά και την ανάλυση των ίδιων δεδομένων με διαφερετικό τρόπο από τους εκάστοτε ερευνητές το οποίο με τη σειρά του συμβάλει στην γενίκευση των αποτελεσμάτων(Russel,Russel and Meadows 2009).Επίσης με την σύνθεση μελετών οι ερευνητές δεν αντιμετωπίζουν και οικονομικό πρόβλημα καθώς χρησιμοποιούν δεδομένα διαθέσιμα από παλαιότερες μελέτες.

Σύνοψη –Συμπεράσματα.

Η χρήση των αλγορίθμων συσταδοποίησης μικροσυστοιχειών έχει το πλεονέκτημα ότι μπορεί να εξετάζει ταυτόχρονα την έκφραση χιλιάδων γονιδίων και ενδείκνυται για συγκριτικές μελέτες γονιδιωμάτων ενώ σαν μειονέκτημά της είναι το υψηλό κόστος και η ανακρίβεια των αποτελεσμάτων λόγω τεχνικών προβλημάτων.Γι αυτό είναι σημαντικό να σημειωθεί ότι τα αποτελέσματα των αλγόριθμων ομαδοποίησης είναι απλές μαθηματικές ερμηνείες των δεδομένων και δεν είναι απαραίτητο να έχουν σχέση με την κατανόηση των βιολογικών μηχανισμών. Οι αλγόριθμοι που επιλέχθηκαν κατά την διάρκεια της ανάλυσης καθώς και οι παράμετροι που χρησιμοποιήθηκαν, θα έχουν σημαντική επίδραση στα συμπεράσματα που θα εξαχθούν από την ανάλυση.Η ιεραρχική συσταδοποίηση συμβάλλει κυρίως στην απεικόνιση των δεδομένων, δείνοντας στους ερευνητές αρχικά μια δυνατότητα για την περαιτέρω κατανόηση των αποτελεσμάτων σε σχέση με την συσταδοποίηση k-μέσων.Το πλεονέκτημα αυτό της ιεραρχικής συσταδοποίησης είναι και αυτό που την καθιστά καταλληλότερη για την ανάλυση δεδομένων μικροσυστοιχειών. Τα πορίσματα των παραπάνω μεθόδων δεν πρέπει να εκληφθούν ως απόλυτα γεγονότα αλλά ως υποθέσεις για περαιτέρω εξέταση και ανάλυση.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler & Leif C Groop *"PGC-1**a**-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes", Nature Genetics 34, 267 - 273 (2003)*

Rivals l. , Personnaz L. , Taing L. ,Potier MC,*"Enrichment or depletion of a GO category within a class of genes: which test?",bioinformatics 2007 ,vol23,no4 , pp: 401-7*

Aravind Subramaniana,b, Pablo Tamayoa,b, Vamsi K. Moothaa,c, Sayan Mukherjeed, Benjamin L. Eberta,e, Michael A. Gillettea,f, Amanda Paulovichg, Scott L. Pomeroyh, Todd R. Goluba,e, Eric S. Landera,c,i,j,k, and Jill P. Mesirova,k,"*Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles" vol. 102 no. 43, 15545–15550,*

William T. Barry Andrew B. Nobel Fred A. Wright ,"*Significance analysis of functional categories in gene expression studies: a structured permutation approach",Bioinformatics (2005) ,vol21 ,no9:pp: 1943-1949.*

T. Cover P. Hart,"*Nearest neighbor pattern classification"1967,IEEE Transactions on Information Theory , Vol 13, Issue: 1,pp:21-27*

Irizarry RA, Wang C, Zhou Y, Speed TP,"*Gene set enrichment analysis made simple" Stat Methods Med Res. 2009 ,vol18,no6,pp:565-75*

Adi L. Tarca, Roberto Romero, Sorin Draghici *"Analysis of microarray experiments of gene expression profiling",American Journals of Obstetrics and Gynecology, 2006 vol 195 ,no2 ,pp: 373–388.*

Maria Halkidi,Yannis Batistakis,Michalis Vazirgiannis,"*On Clustering Validation Techniques",Journal of Intelligent Information Systems December 2001, Volume 17, Issue 2, pp 107–145*

Steve Russell, Lisa Meadows, Roslin Russell, *" Microarray Technology in Practice" 1st Edition,Academic Press 2008*

Norman H. Lee, Alexander I. Saeed *,"Microarrays:An Overview" Protocols for Nucleic Acid Analysis by Nonradioactive Probes ,Volume 353 of the series Methods in Molecular Biology, pp: 265-300*

O Yim, KT Ramdeen – Quant*. " Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data"*