ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΒΙΟΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ (ΠΜΣ):
«ΜΕΘΟΔΟΛΟΓΙΑ ΒΙΟΪΑΤΡΙΚΗΣ ΕΡΕΥΝΑΣ, ΒΙΟΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΔΙΑΤΡΙΒΗ

''ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ''

''STATISTICAL METHODS IN ANALYZING MICROARRAY DATA''

ΠΑΥΛΙΔΟΥ ΑΛΚΜΗΝΗ

Τριμελής Επιτροπή,
Μπατσίδης Απόστολος, Επίκουρος Καθηγητής του Πανεπιστημίου Ιωαννίνων (Επιβλέπων).
Στεφανίδης Ιωάννης, Καθηγητής του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας.
Δοξάνη Χρυσούλα, Επιστημονικός Συνεργάτης του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας

ΛΑΡΙΣΑ
ΙΑΝΟΥΑΡΙΟΣ 2019

Α.ΠΕΡΙΛΗΨΗ

Εισαγωγή

Η πρόοδος της επιστήμης έχει αναδείξει την τεχνολογία των μικροσυστοιχιών πολύ χρήσιμη για την γενετική έρευνα, καθώς μπορεί να εξετάζει ταυτόχρονα την έκφραση χιλιάδων γονιδίων και ενδείκνυται για συγκριτικές μελέτες γονιδιωμάτων.

Στόχοι

Σκοπός της παρούσας μελέτης είναι η κατανόηση της βασικής λειτουργίας και τεχνολογίας των μικροσυστοιχιών και η συνοπτική περιγραφή κάποιων εκ των στατιστικών μεθοδολογιών, οι οποίες χρησιμοποιούνται για την στατιστική ανάλυση των μικροσυστοιχιών DNA.

Μέθοδοι

Αρχικά πραγματοποιήθηκε εκτενής αναζήτηση σε βιβλιογραφικές και ηλεκτρονικές σχετικές με τις μικροσυστοιχίες, τις εφαρμογές τους, καθώς και τη στατιστική ανάλυση των αποτελεσμάτων τους. Έπειτα το ενδιαφέρον επικεντρώθηκε στην εφαρμογή του ελέγχου της ισότητας μέσων τιμών με εξαρτημένα δείγματα (t-test paired samples) για τη σύγκριση ως προς τον τρόπο έκφρασης (χρώμα) δύο γονιδίων, πιο συγκεκριμένα των BRCA1 και BRCA2 χρησιμοποιώντας δεδομένα από την έρευνα "Gene expression microarray analysis of normal breast tissue biopsies" που πραγματοποιήθηκε σε 19 δείγματα από υγιή άτομα. Τέλος, περιγράφεται συνοπτικά η μέθοδος cluster ανάλυσης (ανάλυσης κατά συστάδες).

Αποτελέσματα

Από την αρχική βιβλιογραφική αναζήτηση αποκομίσαμε γνώση σχετικά κυρίως για τις χρήσεις των μικροσυστοιχιών σήμερα, για τον τρόπο κατασκευής τους, την εξαγωγή των δεδομένων και τη στατιστική τους ανάλυση. Περιγράφτηκε ένα αποτέλεσμα σύγκρισης της έκφρασης δύο βασικών γονιδίων για την εμφάνιση του κληρονομικού καρκίνου του μαστού.

Συμπεράσματα

Οι μικροσυστοιχίες αποδεικνύονται μια τεχνολογία άριστη, πάντα σε συνδυασμό με την κατάλληλη στατιστική μέθοδο για την ανάλυσή τους, για την εξέλιξη της επιστημονικής γνώσης και θεραπείας. Έτσι, μπορούμε να βγάλουμε συμπεράσματα σχετικά με τη συσχέτιση έκφρασης των γονιδίων βάση του χρώματος με το οποίο εμφανίζονται στις μικροσυστοιχίες, υποδηλώνοντας εάν αυτά εκφράζονται ή καταστέλλονται.

ABSTRACT

Introduction

The advancement of science has highlighted microarray technology very useful for genetic research as it can simultaneously examine the expression of thousands of genes and is suitable for

comparative genomic studies.

Purpose

The aim of the present study is to understand the basic function and technology of microarrays and to describe some statistical methods by which statistical analysis of DNA microarrays is performed.

Methods

A bibliographic study was carried out and books and internet addresses were found with sources about microarrays, their applications, and statistical analysis of their results. An example of paired t-test was made to compare BRCA1 and BRCA2 genes in their mode of expression (color) in 19 samples from healthy individuals in the research "Gene expression microarray analysis of normal breast tissue biopsies". Also, the cluster analysis method was briefly described.

Results

From the research, we learned about the uses of microarrays nowadays, how they are constructed, how the data are extracted and analyzed. A comparison of the expression of two key genes for the appearance of hereditary breast cancer was described.

Conclusion

Microarrays is proved to be an excellent technology, always in combination with the appropriate statistical method for their analysis, for the development of scientific knowledge and therapy. Thus, we can conclude the correlation of gene expression based on the color they appear in microarrays, indicating whether they are expressed or repressed.

Β.ΕΙΣΑΓΩΓΗ

Η μικροσυστοιχία DNA είναι μια διάταξη μικροσκοπικών σημείων που κατασκευάζεται με μεθόδους όπως αυτές που χρησιμοποιούνται για την κατασκευή μικροτσίπς υπολογιστών. Στην ουσία είναι αντικειμενοφόρες πλάκες μικροσκοπίου που περιέχουν μια σειριακή σειρά δειγμάτων και μπορεί να είναι DNA, RNA, πρωτεϊνική ή ιστού, με τη μικροσυστοιχία DNA να είναι η πιο συνηθισμένη. Οι μικροσυστοιχίες DNA είναι εργαλεία με τα οποία επιτυγχάνεται η ταυτοποίηση και η ποσοτικοποίηση των mRNA αντιγράφων που είναι παρόντα στα κύτταρα. Ο αριθμός των μορίων των mRNA, που προέρχεται από την αντιγραφή ενός γονιδίου, μπορεί να θεωρηθεί κατά προσέγγιση ως το επίπεδο της έκφρασης αυτού. Μια εναλλακτική τεχνολογία επιτρέπει στο DNA να συντίθεται απευθείας πάνω στην ίδια την αντικειμενοφόρο πλάκα με μια φωτολιθογραφική διαδικασία. Η μικροσυστοιχία αποτελείται από μία στερεή επιφάνεια πάνω στην οποία υπάρχουν πολυκουκλεοτίδια που ονομάζονται ανιχνευτές και τα οποία έχουν προσδεθεί ή συντεθεί σε συγκεκριμένες θέσεις. [4]

Δύο είδη της έκφρασης των μικροσυστοιχιών υπάρχουν και η κύρια διαφορά τους είναι στο τρόπο με τον οποίο τοποθετούνται οι ανιχνευτές πάνω στην ολίσθηση. Έτσι διακρίνουμε τα cDNA ή στιγματισμένες μικροσυστοιχίες τα οποία ονομάζονται έτσι γιατί οι ανιχνευτές συντίθενται κατά μέρος και τυπώνονται στην ολίσθηση. Ο όρος cDNA χρησιμοποιείται, διότι ο ανιχνευτής είναι ένα τέλειο αντίγραφο της γνήσιας αλληλουχίας και κάθε ανιχνευτής αντιπροσωπεύει ένα γονίδιο. Το άλλο είδος είναι οι ολιγονουκλεοτινικές μικροσυστοιχίες των οποίων κύριοι αντιπρόσωποι είναι οι Genechip και Affymetrix, οι εμπορικές ονομασίες των εταιρειών που τα κατασκευάζουν και οι ανιχνευτές συντίθενται απευθείας στην επιφάνεια. Ο όρος ολιγονουκλεοτιδικό απευθύνεται στο γεγονός ότι η συνθετική διαδικασία επιτρέπει να δημιουργηθούν μόνο μικρά τεμάχια έτσι ώστε το γονίδιο να μην αντιπροσωπεύεται από έναν ανιχνευτή, αλλά ως μια σειρά από αυτούς. [5,6]

Η μικροσυστοιχία DNA χρησιμοποιείται για πολλούς λόγους. Ένας λόγος είναι για να καθοριστεί εάν το DNA από ένα συγκεκριμένο άτομο περιέχει μία μετάλλαξη και για παράδειγμα πόσο συχνά άτομα με μια συγκεκριμένη μετάλλαξη αναπτύσσουν καρκίνο του μαστού. Αυτή είναι η συνηθέστερη όλων και αφορά τη γονιδιακή έκφραση. Επίσης, χρησιμοποιείται για να μελετηθεί εάν τα γονίδια ενεργοποιούνται ή απενεργοποιούνται σε κύτταρα και ιστούς. Ακόμη, τέλος, μπορεί να χρησιμοποιηθεί και για τον τομέα της φαρμακολογίας, για τον τρόπο με τον οποίο δηλαδή κάποιοι οργανισμοί μπορούν να χειριστούν κάποια φάρμακα.[9]

Ο σκοπός λοιπόν της χρήσης της τεχνολογίας των μικροσυστοιχιών είναι ο συνδυασμός δημιουργίας μιας επιστημονικής υπόθεσης και τα τεράστια ποσά δεδομένων που παράγονται με αποτέλεσμα να γίνεται αυτή η μελέτη γοητευτική. Από την άλλη μεριά απαιτούνται υψηλές τεχνικές απαιτήσεις, καθώς και απαιτήσεις οργάνων, που σε συνδυασμό με τις στατιστικές ανάγκες κάνουν τη μέθοδο αρκετά αποθαρρυντική για τους χρήστες.

Ο τρόπος κατασκευής μικροσυστοιχιών λοιπόν είναι ο εξής: Αρχικά λαμβάνεται δείγμα DNA από το αίμα του ασθενούς, καθώς και ένα δείγμα ελέγχου (που δε περιέχει μετάλλαξη στο συγκεκριμένο γονίδιο). Για να ξεκινήσει ένα πείραμα μικροσυστοιχιών εξάγεται το RNA από τα υποκείμενα κύτταρα. Έπειτα, κάποια από αυτά τα μόρια αντικαθίστανται από άλλα τα οποία περιέχουν φθορίζουσα χρωστική ουσία. Τα αποτελέσματα των επισημασμένων μεταγραφών ονομάζονται στόχοι. Εφόσον τα δείγματα είναι έτοιμα τοποθετούνται πάνω από τη συστοιχία και αφήνονται μέσα στον θάλαμο υβριδοποίησης για μερικές ώρες. Οι επισημασμένοι στόχοι δεσμεύονται με την υβριδοποίηση απο τους ανιχνευτές πάνω στη συστοιχία με την οποία μοιράζονται επαρκή ακολουθία συμπληρωματικότητας. Μετά από αυτό το χρονικό διάστημα η συστοιχία πλένεται και με αυτόν τον τρόπο απομακρύνονται αυτοί οι στόχοι οι οποίοι δεν υβριδοποιήθηκαν. Ο τρόπος με τον οποίο το προηγούμενο βήμα πραγματοποιείται είναι η δεύτερη πιο σημαντική διαφορά ανάμεσα στα δύο είδη των μικροσυστοιχιών. Σε cDNA από 2 ιστούς μαρκαρισμένα με φθορίζουσα ουσία διαφορετικού χρώματος, συνήθως κόκκινο και πράσινο, υβριδοποιούνται σε μία μοναδική

συστοιχία. Οι δύο στόχοι ανταγωνίζονται να υβριδοποιηθούν με τους ανιχνευτές. Για προφανείς λόγους οι σημασμένες συστοιχίες ονομάζονται συστοιχίες δύο χρωμάτων. Από την άλλη μεριά, το σύστημα Affymetrix υβριδοποιεί μόνο ένα δείγμα για κάθε συστοιχία. Αυτό απαιτεί περισσότερες ολισθήσεις για κάθε πείραμα και δεν έχει το πλεονέκτημα να χρησιμοποιήσει την ανταγωνιστική υβριδοποίηση, αλλά απλοποιεί τον σχεδιασμό του πειράματος και βασίζεται σε περισσότερο ευαίσθητη τεχνολογία. Σε αυτό το σημείο, κάθε δοκιμασία της συστοιχίας μπορεί να μετρηθεί ως μαρκαρισμένος στόχος, το οποίο σύμφωνα με τις αρχικές μας υποθέσεις, πρέπει να είναι ανάλογο με το επίπεδο έκφρασης των γονιδίων που αντιπροσωπεύουν αυτόν τον ανιχνευτή. Προκειμένου να προσδιορίσουμε τη ποσότητα του δείγματος που υβριδοποιεί τη συστοιχία φωταγωγείται από ένα φώς λέίζερ το οποίο προκαλεί αναλογικά με τη ποσότητά τους την φθορίζουσα χρώση των μορίων. Έτσι, αυτός ο φθορισμός προσλαμβάνεται απο έναν σαρωτή ως μια εικόνα πλέγματος από φωτεινά σημεία, αντιστοιχώντας το καθένα απο αυτά σε έναν ανιχνευτή. Τελικά, η εικόνα θα μετασχηματιστεί σε νούμερα που θα αποτελέσει τη βάση της ανάλυσης. Το κόκκινο χρώμα πρέπει να αναφέρουμε οτι αντιστοιχεί στη διέγερση έκφρασης ενός γονιδίου, το πράσινο στη καταστολή του, ενώ η ένταση του χρώματος αντιστοιχεί στο λόγο θεραπεία/controls δηλαδή μαύρο=1 δηλαδή καμία διαφορά και κόκκινο=5 δηλαδή 5πλάσια ποσότητα mRNA στο υπό θεραπεία δείγμα σε σχέση με τα controls για το αντίστοιχο γονίδιο. [3,4,5,6,7,9]

Γ.ΜΕΘΟΔΟΙ

Πραγματοποιήθηκε έρευνα βιβλιογραφίας σχετικά με τις μικροσυστοιχίες DNA και τις στατιστικές μεθόδους ανάλυσής τους από τις 16/11/2018. Βρέθηκαν βάσεις δεδομένων σε ηλεκτρονικές διευθύνσεις, όπως είναι https://www.ebi.ac.uk/arrayexpress/, https://www.oncomine.org/resource/login.html και https://www.ncbi.nlm.nih.gov/gds/?term=microarrays, οι οποίες ήταν ιδιαίτερα χρήσιμες στην αναζήτηση ενημερωτικού υλικού για τις μικροσυστοιχίες, όπως και στην αναζήτηση βάσεων δεδομένων για τη περαιτέρω στατιστική ανάλυσή τους. Πραγματοποιήθηκε ταυτόχρονα έρευνα για την αναζήτηση δεδομένων. Στον τομέα των μικροσυστοιχιών η κανονικοποίηση συνηθίζεται να αποτελεί μέρος του στατιστικού σχεδιασμού. Υπόσχεται να μειώσει στο ελάχιστο τα συστηματικά λάθη των δεδομένων και να δημιουργήσει ένα καθαρότερο αρχείο δεδομένων, το οποίο στη συνέχεια μπορεί να χρησιμοποιηθεί για την ανάλυση. Πρίν τη σύγκριση ανάμεσα σε συστοιχίες δύο ανιχνευτών, η παρέκκλιση ανάμεσα στα δύο πειράματα που προκλήθηκε από τεχνικούς και βιολογικούς παράγοντες πρέπει να κανονικοποιηθεί. Κύριες πηγές τεχνικής απόκλισης σε πειράματα συστοιχιών είναι ποιοτικά και ποσοτικά στην υβριδοποίηση του μαρκαρισμένου RNA, όπως και διαφορές στα αντιδραστήρια, στο χρωματισμό, καθώς και στο χειρισμό των τσιπς. Από την άλλη μεριά, βιολογική απόκλιση

μπορεί να προκληθεί από διαφορές στο γενετικό υπόβαθρο, συνθήκες αύξησης, χρόνου, φύλου, ηλικίας κτλ.

Η κανονικοποίηση είναι μια τεχνική για τον μετασχηματισμό δύο κατανομών, ώστε να αποκτήσουν παρόμοιες στατιστικές ιδιότητες. Για την κανονικοποίηση μιας κατανομής, ορίζουμε μια κατανομή αναφοράς ίδιου μήκους και τις ταξινομούμε και τις δύο χωριστά. Τότε, η υψηλότερη τιμή της πρώτης κατανομής παίρνει την υψηλότερη τιμή της κατανομής αναφοράς, τη δεύτερη υψηλότερη τιμή της κατανομής αναφοράς κ.ο.κ., μέχρις ότου η κατανομή που μας ενδιαφέρει να αποτελεί τελικά μια αναδιάταξη των στοιχείων της κατανομής αναφοράς.

Για να κανονικοποιήσουμε δύο ή περισσότερες κατανομές μεταξύ τους, χωρίς τη χρήση μιας κατανομής αναφοράς, ταξινομούμε όπως πριν, κι έπειτα θέτουμε τον αριθμητικό μέσο όρο των κατανομών. Έτσι, η υψηλότερη τιμή σε όλες τις κατανομές γίνεται ο μέσος όρος των υψηλότερων τιμών, η δεύτερη υψηλότερη γίνεται ο μέσος όρος των δεύτερων υψηλότερων κ.ο.κ.

Η κατανομή αναφοράς θα είναι μια συνήθης στατιστική κατανομή, όπως για παράδειγμα η κατανομή Gauss ή η κατανομή Poisson.

Η κανονικοποίηση ποσοστημορίων χρησιμοποιείται συχνά στην ανάλυση δεδομένων από μικροσυστοιχίες γονιδίων.

Ακολουθεί μια σύντομη περιγραφή της μεθόδου σε ένα πολύ μικρό σύνολο δεδομένων.

Έστω 3 μικροσυστοιχίες και τα γονίδια A έως D:

A	3	2	1
B	8	6	2
C	4	2	8
D	3	4	6

Για κάθε στήλη αντιστοιχίζουμε ένα βαθμό (i-iv) από τη χαμηλότερη στην υψηλότερη τιμή:

A	i	i	i
B	iii	iii	ii
C	ii	i	iv
D	i	ii	iii

Επιστρέφουμε στο αρχικό σύνολο δεδομένων. Ταξινομούμε κάθε στήλη ξεχωριστά κατά αύξουσα σειρά. (Η πρώτη στήλη αποτελείται από τα 3,8,4,3. Μετά την ταξινόμηση θα γίνει 3,3,4,8. Παρόμοια και οι υπόλοιπες στήλες.)Το αποτέλεσμα είναι:

A	3	2	1	γίνεται A 3 2 1

B	8	6	2	γίνεται B 3 2 2
C	4	2	8	γίνεται C 4 4 6
D	3	4	6	γίνεται D 8 6 8

Τώρα υπολογίζουμε τον μέσο όρο κάθε γραμμής για να αντιστοιχίσουμε τους βαθμούς:

A (3 2 1)/3 = 2.00 = βαθμός i
B (3 2 2)/3 = 2.33 = βαθμός ii
C (4 4 6)/3 = 4.66 = βαθμός iii
D (8 6 8)/3 = 7.33 = βαθμός iv

Στη συνέχεια παίρνουμε τον πίνακα βαθμών που υπολογίσαμε νωρίτερα και αντικαθιστούμε τις τιμές τους.

A	i	i	i
B	iii	iii	ii
C	ii	i	iv
D	i	ii	iii

γίνεται:

A	2.00	2.00	2.00
B	4.66	4.66	2.33
C	2.33	2.00	7.33
D	2.00	2.33	4.66

Αυτές είναι και οι νέες κανονικοποιημένες τιμές, οι οποίες ακολουθούν την ίδια κατανομή και μπορούν πλέον να συγκριθούν με ευκολία.[32]

Παραμετρικός έλεγχος t τεστ με ανεξάρτητα δείγματα: Η στατιστική συνάρτηση που χρησιμοποιείται για τον έλεγχο της ισότητας δύο μέσων τιμών με ανεξάρτητα δείγματα καθορίζεται στη βάση της ισότητας ή μη των δύο πληθυσμιακών διακυμάνσεων.

ι) Ειδικότερα, αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων δεν απορρίπτεται (τεστ του Levene, p-τιμή>α), χρησιμοποιείται η στατιστική συνάρτηση

$$t = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \overset{H_0}{\sim} t_{n+m-2},$$

όπου $\bar{X}$ και $\bar{Y}$ οι δειγματικές μέσες τιμές και $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$, με S_1^2, S_2^2 τις δειγματικές διακυμάνσεις. Επιπλέον το 100(1-α)% Δ.Ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ είναι

$$\left(\bar{X} - \bar{Y} - t_{n+m-2,a/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{n+m-2,a/2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right).$$

ιι) Αν η υπόθεση της ισότητας των πληθυσμιακών διακυμάνσεων απορρίπτεται (τεστ του Levene, p-τιμή<α), χρησιμοποιείται η στατιστική συνάρτηση (γνωστό ως τεστ του Welch)

$$t = \frac{\bar{X} - \bar{Y}}{S} \overset{H_0}{\sim} t_v,$$

όπου $S^2 = \frac{S_1^2}{n} + \frac{S_2^2}{m}$, και $v = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$, όπου $c = \frac{S_1^2}{nS^2}$.

Επιπλέον το 100(1-α)% Δ.Ε. για τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$ είναι

$$\left(\bar{X} - \bar{Y} - t_{v,a/2} S, \bar{X} - \bar{Y} + t_{v,a/2} S\right).$$

<u>Παραμετρικός έλεγχος- Τ τεστ συγκρίσεως ζευγών</u>: Έστω ένα τυχαίο δείγμα $X_1, \ldots, X_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ_1 και διακύμανση σ_1^2. Επιπλέον έστω ένα τυχαίο δείγμα $Y_1, \ldots, Y_n$ μεγέθους n από έναν πληθυσμό με μέση τιμή μ_2 και διακύμανση σ_2^2. Επιπρόσθετα υποθέτουμε ότι τα δύο δείγματα είναι εξαρτημένα. Ενδιαφερόμαστε για τον έλεγχο, σε επίπεδο σημαντικότητας α, της ισότητας των δύο μέσων τιμών. Το πρώτο βήμα για τη μελέτη του προβλήματος είναι η δημιουργία των διαφορών $D_i = X_i - Y_i$, $i = 1, \ldots, n$.

Χρησιμοποιούμε τη στατιστική συνάρτηση $t = \frac{\bar{D}}{S_D / \sqrt{n}} \overset{H_0}{\sim} t_{n-1}$, όπου $\bar{D}$ και S_D η μέση τιμή και τυπική απόκλιση του δείγματος των διαφορών $D_i = X_i - Y_i,\ i = 1, 2, ..., n$. Επιπλέον το 100(1-α)% Δ.Ε. για την $\mu_1 - \mu_2$ είναι:

$$\left(\bar{D} - t_{n-1,a/2} \frac{S_D}{\sqrt{n}}, \bar{D} + t_{n-1,a/2} \frac{S_D}{\sqrt{n}} \right).$$

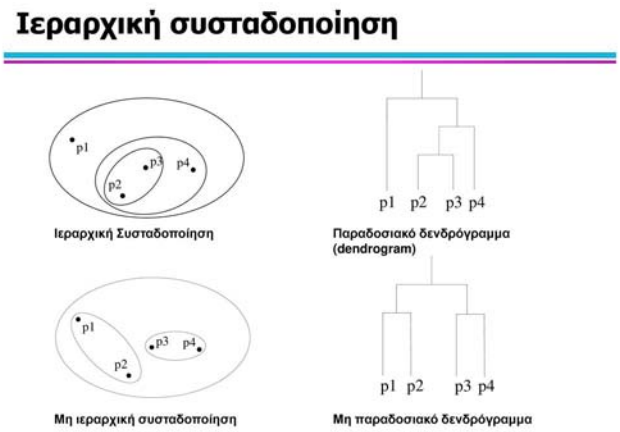
Συσταδοποίηση (ανάλυση κατά συστάδες)

Μία μέθοδος στατιστικής ανάλυσης των δεδομένων μικροσυστοιχιών είναι και η συσταδοποίηση, η οποία βοηθά στην ανακάλυψη ομάδων με παρόμοια αντικείμενα. Τα βασικά στάδια συσταδοποίησης είναι αρχικά η επιλογή των χαρακτηριστικών, η κατάλληλη επιλογή του αλγορίθμου συσταδοποίησης ανάλογα με τα κριτήρια, η αξιολόγηση των αποτελεσμάτων και τέλος η σύγκριση τους με άλλα επιστημονικά στοιχεία και αναλύσεις. Οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούνται στην ανάλυση δεδομένων μικροσυστοιχιών είναι οι διαιρετικοί αλγόριθμοι συσταδοποίησης και η ιεραρχική συσταδοποίηση.

Η ιεραρχική συσταδοποίηση προσφέρει μια οπτική απεικόνιση σε μορφή δενδρογράμματος ομάδων γονιδίων με παρόμοια έκφραση.

Η διαδικασία της ιεραρχικής συσταδοποίησης, της γραφικής δηλαδή αναπαράστασης δεδομένων και ανάλυσης είναι η εξής: Υπολογίζουμε την απόσταση ανάμεσα στις συστάδες (μικρότερη απόσταση-μεγαλύτερη συγγένεια), ενώνουμε αυτές τις συγγενικές συστάδες με μία υψηλότερης βαθμίδας, εφαρμόζουμε τον κανόνα συνδέσμου, μια μέθοδο επαναπροσδιορισμού των αποστάσεων της νέας συστάδας και αναζητούμε την αμέσως μικρότερη απόσταση και επαναλαμβάνουμε τη διαδικασία. Ο κανόνας συνδέσμου μπορεί να είναι **απλός** και να λαμβάνει τη μικρότερη απόσταση ανάμεσα στις συστάδες, **πλήρης,** οπότε και λαμβάνει τη μεγαλύτερη απόσταση είτε **μέσου** συνδέσμου, οπότε και λαμβάνουμε το αριθμητικό μέσο των αποστάσεων δύο συστάδων και περιγράφεται ως ένα δενδρόγραμμα.

Η εικόνα απεικόνισης της ιεραρχικής συσταδοποίησης είναι απο την εργασία "Αναλυση συστάδων: βασικές έννοιες και αλγόριθμοι", κεφάλαιο 8, εξόρυξη δεδομένων απο Pang-Ning Tan, Michigan



State University, Michael Steinbach, University of Minnesota, Vipin Kumar, University of Minnesota, https://slideplayer.gr/slide/11988256/ , πρόσβαση 30/1/2019 με πηγή το βιβλίο https://www.pearson.com/us/higher-education/product/Tan-Introduction-to-Data-Mining/9780321321367.html

Χρησιμοιούμε τη συσταδοποίηση k-μέσων όταν ένας αριθμός συστάδων ομαδοποιεί τα στοιχεία της γονιδιακής έκφρασης σε προκαθορισμένο απο τον χρήστη αριθμό συστάδων (k). Πρώτα επιλέγουμε το k, υπολογίζουμε τη μέση τιμή της απόστασης κάθε στοιχείου από τα κεντροειδή των k-συστάδων και τα αντιστοιχίζουμε με τη συστάδα της οποίας το κεντρομερές μοιάζει περισσότερο. Επαναλαμβάνουμε τα βήματα αυτά εως ότου συσταδοποιηθούν όλα τα στοιχεία, επιλέγουμε μια συστάδα και βρίσκουμε αυτή με την οποία υπάρχει υψηλότερος βαθμός συγγένειας.
[4,14]

Δ.ΑΠΟΤΕΛΕΣΜΑΤΑ

Ορισμένα γονίδια, που ονομάζονται ογκοκατασταλτικά (tumor suppressors), προστατεύουν εναντίον του καρκίνου, διότι επιδιορθώνουν βλάβες που μπορεί να προκύψουν στο DNA μας τυχαία κατά τη διαδικασία συνήθων κυτταρικών διεργασιών (κυτταρική διαίρεση), ή υπό την επίδραση εξωτερικών παραγόντων (ακτινοβολίες, χημικά).

Στην κατηγορία των ογκοκατασταλτικών γονιδίων, ανήκουν τα γονίδια BRCA1 και BRCA2, τα οποία όταν φέρουν μεταλλάξεις στο περιεχόμενό τους που επηρεάζουν τη λειτουργία τους, αδυνατούν να επιδιορθώσουν αποτελεσματικά βλάβες στο DNA, με αποτέλεσμα τη συσσώρευσή τους με την πάροδο του χρόνου, που προάγει την εμφάνιση του καρκίνου.

Η παρουσία παθογόνου μετάλλαξης BRCA1/BRCA2 δεν αρκεί από μόνη της να προκαλέσει την εμφάνιση του καρκίνου. Εάν όμως ένα άτομο κληρονομήσει παθογόνο μετάλλαξη στο γονίδιο BRCA, διατρέχει αυξημένο κίνδυνο για την εμφάνιση ορισμένων μορφών καρκίνου.

Συγκεκριμένα, η συσχέτιση των μεταλλάξεων BRCA με καρκίνο μαστού (συμπεριλαμβανόμενου του ανδρικού καρκίνου του μαστού) και ωοθηκών είναι γνωστή, ως Κληρονομικό Σύνδρομο Καρκίνου Μαστού. [30,31]Στη βιβλιογραφία είναι διαθέσιμη η έρευνα των Bergholtz et al. (2019) που αφορά μια μελέτη σύνδεσης μεταξύ της μαστογραφικής πυκνότητας και της έκφρασης γονιδίων σε φυσιολογικό μαστικό ιστό.

Έγινε αναζήτηση στο **https://www.ebi.ac.uk/arrayexpress/** για breast cancer+microarray data και βρέθηκε η μελέτη με επιπλέον διαθέσιμη τη βάση δεδομένων της έρευνας "Gene expression microarray analysis of normal breast tissue biopsies" στο site **https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5885/query=breast+cancer+microarray+data** η οποία και είναι κανονικοποιημένη. Τη μελέτη πραγματοποίησαν οι Bergholtz H, Lien TG, Ursin G, Holmen MM, Helland Å, Sørlie T, Haakensen VD και δημοσιεύτηκε στο *Journal of mammary gland biology and neoplasia* (2019). Περιγράφει την έκφραση 34612 γονιδίων σε 19 δείγματα υγιών ατόμων, με παράμετρο το χρώμα εμφάνισής τους στις μικροσυστοιχίες.

Χρησιμοποιώντας αυτή τη βάση δεδομένων (αφού επιτυχώς αποθηκεύτηκε ως αρχείο του SPSS) το ενδιαφέρον επικεντρώνεται στα δύο βασικά γονίδια BRCA1 και BRCA2 που ευθύνονται για τον κληρονομικό καρκίνο του μαστού. Ειδικότερα στις μεταβλητές αυτές καταγράφονται 19 δείγματα υγιών ατόμων. Στη συνέχεια θα ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά ως προς το χρώμα των δύο γονιδίων, την έκφρασή τους δηλαδή. Η διαφορά στο χρώμα σημαίνει τη διέγερση του γονιδίου (κόκκινο χρώμα), είτε την καταστολή του (πράσινο χρώμα). Καθώς οι παρατηρήσεις είναι διαθέσιμες στους ίδιους 19 υγιείς ανθρώπους, άμεσα εξάγεται το συμπέρασμα ότι πρόκειται για έλεγχο μέσων τιμών με εξαρτημένα δείγματα και επομένως η κατάλληλη στατιστική μέθοδος που πρέπει να εφαρμοσθεί είναι ο έλεγχος t-test με εξαρτημένα δείγματα (ζευγαρωτές παρατηρήσεις). Εφαρμόζοντας τον παραπάνω έλεγχο (τα αποτελέσματα παρατίθενται στη συνέχεια) προκύπτει ότι δεν υπάρχει στατιστικά σημαντική διαφορά ως προς το χρώμα των δύο γονιδίων καθώς η p-τιμή του ελέγχου είναι ίση με 0.338 (t(18)=-0.985).

Σε διαφορετική περίπτωση, εάν το αποτέλεσμα της στατιστικής μας ανάλυσης ήταν διαφορετικό, άρα και υπήρχε στατιστική διαφορά σημαντική, θα ανατρεπόταν η βασική μας πεποίθηση πως τα δύο γονίδια BRCA1 και BRCA2 σχετίζονται εξίσου με την εκδήλωση του κληρονομικού καρκίνου του μαστού.

T-Test

[DataSet30]

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	BRCA1	-,07739926	19	,419787055	,096305755
	BRCA2	,05078069	19	,750489722	,172174150

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	BRCA1 & BRCA2	19	,663	,002

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	BRCA1 - BRCA2	-,128179953	,567143570	,130111658	-,401534403	,145174498	-,985	18	,338

ΣΤ.ΑΝΑΦΟΡΕΣ

1.Αθανασιάδης, Εμμανουήλ Ιωάννης, 2010, *Επεξεργασία εικόνων μικροσυστοιχιών DNA με χρήση σύγχρονων μεθόδων ταξινόμησης προτύπων*, Πανεπιστήμιο Πατρών, Σχολή επιστημών Υγείας, Τμήμα Ιατρικής

2.Παπαρούντας, Τριαντάφυλλος Λουκάς, 2009, *Βιοπληροφορική ανάλυση δεδομένων πειραμάτων γονιδιακής έκφρασης*, Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Ιατρικής

3.Μαργαρίτης, Αθανάσιος Βύρων, 2008, *Μικροσυστοιχίες DNA για την καταγραφή και μελέτη της γονιδιακής έκφρασης: μέθοδοι ανάλυσης εικόνας και ανάλυση γονιδιακής έκφρασης για την κατανόηση του μηχανισμού δράσης μεταγραφικών παραγόντων και ιστοειδικής έκφρασης*, Πανεπιστήμιο Κρήτης. Σχολή Θετικών και Τεχνολογικών Επιστημών. Τμήμα Βιολογίας

4.Δεληνάσιος Λάζαρος, 2017, *Μέθοδοι ανάλυσης δεδομένων μικροσυστοιχειών-συσταδοποίηση*, Πανεπιστήμιο Θεσσαλίας, Σχολή Επιστημών Υγείας, Τμήμα Ιατρικής

5.Ernst Wit and John McClure, 2004, *Statistics for Microarrays, Design, Analysis and Inference*, Department of Statistics, University of Glasgow, UK

6.Alex Sanchez and M.Carme Ruiz de Villa, 2008, *A Tutorial Review of Microarray Data Analysis*, Universitat de Barcelona, Facultat de Biologia

7.National Human Genome Research Institute, 2019, https://www.genome.gov/10000533/dna-microarray-technology/, πρόσβαση 02.01.2019

8.Μαργαρίτα-Αργεντίνα Ν. Παπακωνσταντίνου, 2007, *Έλεγχος πολλαπλών υποθέσεων σε μικροσυστοιχίες DNA*, Πανεπιστήμιο Πειραιώς, Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

9.Tomi Pasanen, Janna Saarela, Ilana Saarikko, Teemu Toivanen, Martti Tolvanen, Mauno Vihinen

and Garry Wong, 2003, *DNA Microarray Data Analysis*

10.Jennifer S.Shoemaker, Simon M.Lin, 2005, *Methods of Microarray Data Analysis IV*

11.Pierre Baldi and G.Wesley Hatfeld, 2002, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press

12.Dov Stekel, 2003, *Microarray Bioinformatics*, Cambridge University Press

13.Harvey Motulsky, 2010, *Intuitive Biostatistics*, Oxford University Press

14.Συσταδοποίηση, https://repository.kallipos.gr/bitstream/11419/2972/1/02_chapter_06.pdf, πρόσβαση 30/1/2019

15.Miami and Minseqe quidelines, https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html, πρόσβαση 30/1/2019

16.DNA microarrays: a powerful genomic tool for biomedical and clinical research, **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933257/**, πρόσβαση 30/1/2019

17.https://www.ncbi.nlm.nih.gov/gds/?term=microarrays++breast+cancer, πρόσβαση 1/2/2019

18.https://www.oncomine.org/resource/login.html, πρόσβαση 1/2/2019

19.https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html, πρόσβαση 1/2/2019

20.https://www.ebi.ac.uk/arrayexpress/search.html?query=microarrays+one+colour, πρόσβαση 1/2/2019

21.http://www.bioconductor.org/, 1/2/2019

22.Μεταλλάξεις στα γονίδια BRCA1 και BRCA2, **http://www.neaeope.gr/mutations-brca/**, πρόσβαση 2/2/2019

23.Κανονικοποίηση ποσοστημορίων, https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7_%CF%80%CE%BF%CF%83%CE%BF%CF%83%CF%84%CE%B7%CE%BC%CE%BF%CF%81%CE%AF%CF%89%CE%BD, πρόσβαση 30/1/2019

24.How to Calculate Normalized Data in SPSS, **https://www.techwalla.com/articles/how-to-calculate-normalized-data-in-spss**, πρόσβαση 30/1/2019

25.Κεφάλαιο6:Συσταδοποίηση, https://repository.kallipos.gr/bitstream/11419/2972/1/02_chapter_06.pdf, πρόσβαση 30/1/2019

26.DNA microarrays: a powerful genomic tool for biomedical and clinical research, **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933257/**, πρόσβαση 29/1/2019

27.Χρωμοσωμικές Αλλαγές, **http://www.eurogentest.org/index.php?id=521**, πρόσβαση 30/1/2019

28.https://www.heatherturner.net/turnerchapter3.pdf, πρόσβαση 1/2/2019

29.Introduction to data mining, https://www.pearson.com/us/higher-education/product/Tan-Introduction-to-Data-Mining/9780321321367.html, πρόσβαση 30/1/2019

30.Μετάλλαξη BRCA, **https://el.wikipedia.org/wiki/%CE%9C%CE%B5%CF%84%CE%AC%CE%BB%CE%BB%CE%B1%CE%BE%CE%B7_BRCA**, πρόσβαση 30/01/2019

31. Μεταλλάξεις στα γονίδια BRCA1 και BRCA2, http://www.neaeope.gr/mutations-brca/, πρόσβαση 1/2/2019

32.Κανονικοποίηση ποσοστημορίων, **https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7_%CF%80%CE%BF%CF%83%CE%BF%CF%83%CF%84%CE%B7%CE%BC%CE%BF%CF%81%CE%AF%CF%89%CE%BD**, πρόσβαση 1/2/2019

33.Chapter 3, Clustering Microarray Data, https://www.heatherturner.net/turnerchapter3.pdf, πρόσβαση 1/2/2019