ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΕΘΟΔΟΙ ΓΙΑ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΠΕΙΡΑΜΑΤΟΣ ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ - ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

METHODS FOR ANALYSING MICRO-ARRAY DATA - NORMALIZATION

Επιβλέπουσα καθηγήτρια: Άρτεμις Χατζηγεωργίου

Εργασία μεταπτυχιακής φοιτήτριας: Μολόχα Ναυσικά-Μαρία

Φεβρουάριος 2017

ΠΕΡΙΛΗΨΗ

Στην παρούσα μεταπτυχιακή εργασία μας ζητήθηκε να συγκρίνουμε τις σημαντικότερες μεθόδους κανονικοποίησης στα δεδομένα από πειράματα μικροσυστοιχιών. Οι μικροσυστοιχίες είναι μια πειραματική τεχνική βασιζόμενη στον υβριδισμό των νουκλεικών οξέων που έχουν σημανθεί με συγκεκριμένη χρωστική. Η αλματώδης ανάπτυξη της συγκεκριμένης τεχνικής τα τελευταία χρόνια έχει δημιουργήσει την ανάγκη για ευέλικτους αλγορίθμους ώστε να μπορούν να διαχειρισθούν τη μεγάλη πληθώρα πειραματικών δεδομένων απατώντας ταυτόχρονα σε βιολογικά ερωτήματα.

Οι τρεις σημαντικότερο αλγόριθμοι που αναφέρονται στη βιβλιογραφία είναι οι RMA, GCRMA και MAS5.0. Μετά τη μαθηματική ανάλυση του κάθε αλγορίθμου μελετούμε ένα πείραμα της Affymetrix που επιχειρεί με συγκεκριμένα δεδομένα τη σύγκριση των τριών αλγορίθμων. Συμπερασματικά, μπορούμε να πούμε ότι από τα αποτελέσματα του πειράματος που προαναφέρθηκε η επιλογή της κατάλληλης μεθόδου κανονικοποίησης εξαρτάται από τις τοπικές συνθήκες που επικρατούν σε κάθε εργαστήριο.

Περιεχόμενα

1. Εισαγωγή στις μικροσυστοιχίες

1.1. Ιστορική αναδρομή

Οι μικροσυστοιχίες (microarrays) είναι μια τεχνολογία, η οποία εξελίσσεται με γοργούς ρυθμούς τα τελευταία δέκα με δεκαπέντε χρόνια και παρέχει το μέσο ώστε να επιτύχουμε με ταχύτητα παράλληλα πειράματα υβριδοποίησης. Με τη υπάρχουσα τεχνολογία, ένα πείραμα υβριδοποίησης μπορεί να παράγει τα πρότυπα έκφρασης εκατοντάδων χιλιάδων γονιδίων ταυτόχρονα. Ένα πείραμα μικροσυστοιχιών, όπως και τα περισσότερα πειράματα γενετικής, περιλαμβάνει συνήθως την απόκτηση και την επικύρωση μεγάλου συνόλου δεδομένων. Τα σύνολα δεδομένων αυτά περιέχουν μία ποικιλία από διαφορετικές

πληροφορίες, από τις αλληλουχίες των γονίδιων ή των κλώνων που τοποθετούνται σε μια μικροσυστοιχία, έως τις ποσοτικοποιημένες τιμές έκφρασης για κάθε γονίδιο κάτω από διαφορετικές πειραματικές συνθήκες. Η διαδικασία ενός πειράματος μικροσυστοιχιών αρχίζει από το βιολογικό του σχεδιασμό και την κατασκευή της συστοιχίας, συνεχίζεται με την λήψη της εικόνας και την επεξεργασία της, και τελειώνει με την συγκέντρωση των δεδομένων και την ανάλυση τους. Η ανάλυση των δεδομένων θα οδηγήσει πιθανώς σε μια νέα βιολογική υπόθεση, η οποία θα απαιτήσει στη συνέχεια επαλήθευση με συμπληρωματικές τεχνικές αλλά και ταυτόχρονη επανάληψη του πειράματος. Ο παραπάνω κύκλος διεργασιών έχει ονομαστεί κύκλος ζωής των μικροσυστοιχιών, και φαίνεται στο Σχήμα 1.

Σχήμα 1.: Ο κύκλος ζωής των μικροσυστοιχιών

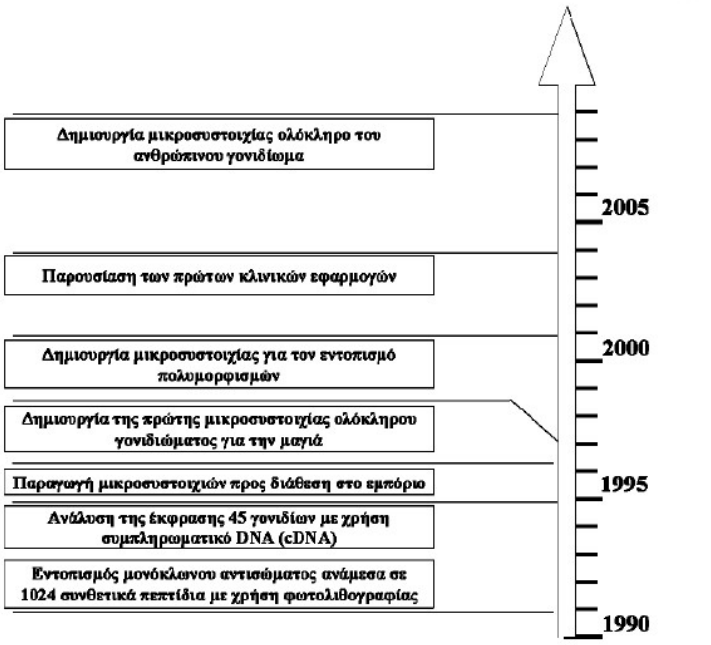
Κατά τη διάρκεια κάθε επανάληψης του κύκλου ζωής των μικροσυστοιχιών, παράγονται νέα στοιχεία, τα οποία μπορούν στη συνέχεια να χρησιμοποιηθούν είτε σε επόμενες επαναλήψεις, είτε στα επόμενα βήματα της ίδιας επανάληψης. Παραδείγματος χάριν, τα δεδομένα που συνδέονται με τη διαδικασία κατασκευής της μικροσυστοιχίας, τα οποία περιέχουν πληροφορίες για τις θέσεις των αλληλουχιών πάνω στην μικροσυστοιχία, μπορούν να χρησιμοποιηθούν στην επεξεργασία εικόνας για να αυτοματοποιήσουν πολλές πτυχές της λειτουργίας. Ωστόσο, οι μέθοδοι οι οποίες προτείνονται για την επεξεργασία εικόνας στην παρούσα εργασία είναι αυτόματες και δεν απαιτούν καμιά εκ των προτέρων γνώση που να πηγάζει από προηγούμενα βήματα. Λόγω επίσης του μεγάλου όγκου των δεδομένων που παράγονται στα πειράματα μικροσυστοιχιών, η αρωγή ευφυών πληροφοριακών συστημάτων καθίσταται απαραίτητη.

Δυο είναι οι βασικές απαιτήσεις που μπορούν να συμβάλουν στην περαιτέρω εξέλιξη των μικροσυστοιχιών. Αφενός πρέπει να υπάρχει συνεχής εξέλιξη του υλικού (Hardware) για την κατασκευή της μικροσυστοιχίας, αφετέρου να υλοποιηθούν τα υπολογιστικά εργαλεία για την ανάλυση του πλήθους των δεδομένων που θα παραχθούν, ώστε αυτά να συμβάλουν στην κατανόηση των λειτουργιών των βιολογικών συστημάτων. Με τη σταθερή πρόοδο στην ανάπτυξη του υλικού των μικροσυστοιχιών, ο διαθέσιμος σήμερα εξοπλισμός μπορεί να παράγει μικροσυστοιχίες με εκατοντάδες χιλιάδες βιολογικές αλληλουχίες. Όσον αφορά στην ανάπτυξη των κατάλληλων εργαλείων για την αξιοποίηση του μεγάλου όγκου δεδομένων, υπάρχουν δύο σημαντικά ζητήματα, με τα οποία η επιστήμη της πληροφορικής συνεισφέρει. Το πρώτο έχει να κάνει με την εξαγωγή των ποσοτικοποιημένων τιμών από την εικόνα. Η διαδικασία αυτή όπως θα δούμε εκτενώς στην παρούσα διατριβή λαμβάνει χώρα με την χρήση τεχνικών επεξεργασίας εικόνας. Το δεύτερο και εξίσου σημαντικό ζήτημα είναι η επεξεργασία των εξαγόμενων από την εικόνα δεδομένων, με σκοπό την εξόρυξη νέας γνώσης. Η παρούσα εργασία ασχολείται με το σχεδιασμό και την υλοποίηση ευφυών πληροφοριακών συστημάτων για την επεξεργασία των εικόνων των μικροσυστοιχιών.

Ως προπομπός της τεχνολογίας αυτής μπορεί να χαρακτηριστεί η εργασία των Fodor και των συνεργατών του Fodor et al το 1991, οι οποίοι χρησιμοποίησαν την φωτολιθογραφία με σκοπό να εντοπίσουν ένα μονόκλωνικο αντίσωμα ανάμεσα σε 1024 συνθετικά πεπτίδια. Η πρώτη ωστόσο εργασία μικροσυστοιχιών παρουσιάστηκε το 1995 από τους Schena *et al.* οι οποίοι ανέλυσαν την έκφραση 45 γονιδίων κάνοντας χρήση συμπληρωματικού DNA (cDNA). Η ανάλυση της έκφρασης των γονιδίων ήταν και η πρώτη εφαρμογή της τεχνολογίας αυτής. Ένα καθοριστικό βήμα για την διάδοση της τεχνολογίας έκανε η εταιρία Affymetrix, η οποία το 1996 παρήγαγε τις πρώτες μικροσυστοιχίες προς διάθεση στο εμπόριο (Lockhart etal 1996). Όσο το υλικό των μικροσυστοιχιών εξελίσσεται, δίνεται η δυνατότητα στις ερευνητικές ομάδες να αναπτύσσουν μικροσυστοιχίες με όλο και μεγαλύτερο αριθμό ανιχνευτών. Αυτό είχε ως αποτέλεσμα οι ερευνητές του Stanford να παρουσιάσουν την πρώτη μικροσυστοιχία ολόκληρου γονιδιώματος για τον ζυμομύκητα (ζύμη).

Αντίστοιχα με την πάροδο των χρόνων όλο και περισσότερα πεδία βιολογικής έρευνας εύρισκαν εφαρμογή στις μικροσυστοιχίες. Έτσι η εταιρία Illumina το 2001 (Shen etal 2001) δημιούργησε την πρώτη μικροσυστοιχία για τον εντοπισμό πολυμορφισμών στο DNA

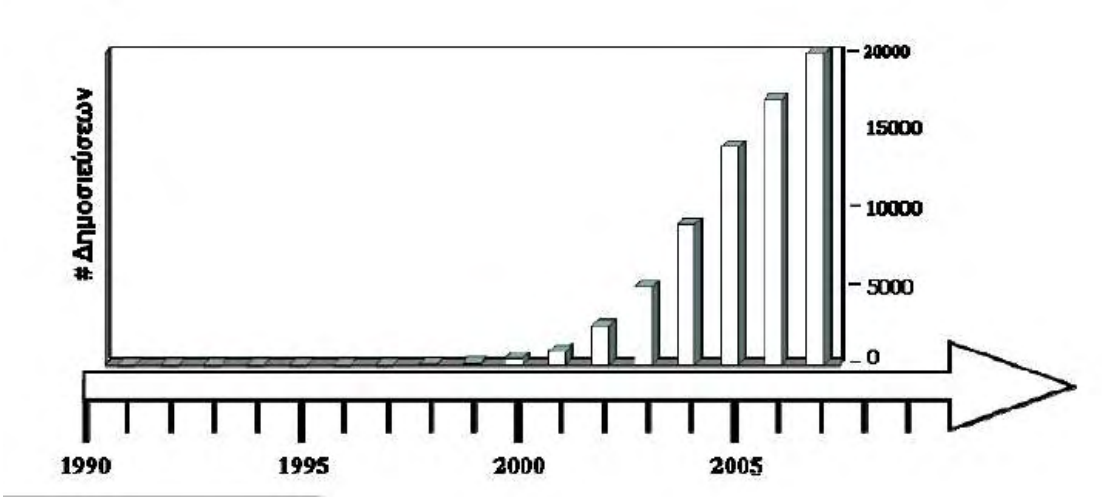
αναπτύσσοντας μία νέα τεχνολογία, τα λεγόμενα BeadChips, τα οποία είναι τρισδιάστατα πλέγματα από ανιχνευτές με σκοπό την επιτάχυνση της υβριδοποίησης, αλλά και την βελτίωση της αξιοπιστίας της. Η πρώτες κλινικές εφαρμογές των μικροσυστοιχιών εμφανίζονται το 2004, οι οποίες στόχευαν στην πρόβλεψη και την διάγνωση διαφόρων νοσημάτων. Σήμερα η εν λόγω τεχνολογία έχει καταφέρει να φτάσει σε επίπεδο διερεύνησης ολόκληρου του ανθρώπινου γονιδιώματος, όπου η επεξεργασία των παραγόμενων δεδομένων γίνεται με πολύ γοργούς ρυθμούς από εργαστήρια-επί-πλινθίου (Lab-on-Chip). Η ιστορική αναδρομή της τεχνολογίας με τα γεγονότα «ορόσημα» τα οποία αναφέρθηκαν ανωτέρω παρουσιάζονται στο Σχήμα 2.



Σχήμα 2: Η εξελικτική πορεία της τεχνολογίας των μικροσυστοιχιών

Η ραγδαία εξέλιξη της τεχνολογίας των μικροσυστοιχιών μπορεί εύκολα να φανεί και από τον αριθμό των δημοσιεύσεων σε επιστημονικά περιοδικά και σε πρακτικά διεθνών συνεδρίων. Οι δημοσιεύσεις, οι οποίες σχετίζονται με την τεχνολογία των μικροσυστοιχιών και οι οποίες αφορούν ολόκληρο τον κύκλο ζωής των μικροσυστοιχιών, αυξάνονται με γεωμετρική πρόοδο τα τελευταία δέκα περίπου χρόνια. Αυτό συνιστά μια ένδειξη ότι η βιολογική επιστήμη έχει επενδύσει πολλά στην συγκεκριμένη τεχνολογία. Ο όγκος βέβαια των δεδομένων που παράγονται απαιτεί ταυτόχρονη ενασχόληση και άλλων επιστημών με

το συγκεκριμένο αντικείμενο, όπως της πληροφορικής. Το Σχήμα 3 παρουσιάζει τον αριθμό των δημοσιεύσεων των σχετικών με την τεχνολογία μικροσυστοιχιών από την πρώτη εμφάνιση της μέχρι και σήμερα.



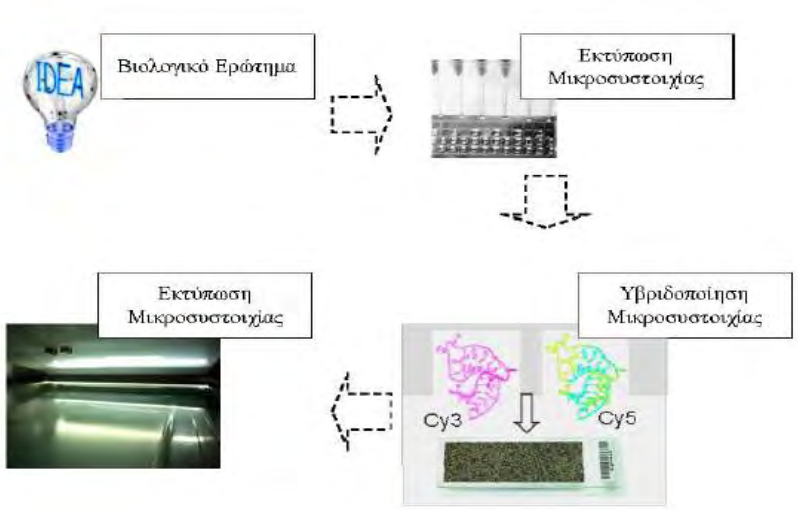
Σχήμα 3: Η ραγδαία εξέλιξη της τεχνολογίας μικροσυστοιχιών όπως αποτυπώνεται από τον αριθμό των δημοσιεύσεων τα τελευταία χρόνια.

Από την ραγδαία εξέλιξη της τεχνολογίας απορρέουν πολλές παραλλαγές και ένας μεγάλος αριθμός εφαρμογών των μικροσυστοιχιών, με σκοπό την διερεύνηση των βιολογικών λειτουργιών.

1.2. Προετοιμασία του πειράματος και βιολογικό ερώτημα

Στο κεφάλαιο αυτό περιγράφονται όλα τα βήματα, τα οποία ακολουθούνται κατά την διάρκεια του κύκλου ζωής ενός πειράματος μικροσυστοιχιών, έως ότου καταλήξουμε στο ζητούμενο της παρούσας διδακτορικής διατριβής, το οποίο είναι η επεξεργασία των εικόνων των μικροσυστοιχιών. Παρά το μεγάλο πλήθος των εφαρμογών και την ποικιλία των ειδών των μικροσυστοιχιών, ένα πείραμα ακολουθεί μια τυποποιημένη διαδικασία. Στο τέλος του κεφαλαίου, θα περιγραφούν οι βάσεις δεδομένων με εικόνες μικροσυστοιχιών, οι οποίες χρησιμοποιήθηκαν για την αξιολόγηση των μεθόδων. Όπως αναφέρθηκε στην εισαγωγή, η κατασκευή μιας μικροσυστοιχίας και τα στάδια του ίδιου του βιολογικού πειράματος είναι εκείνα τα οποία προηγούνται στον κύκλο ζωής. Το πέρας αυτών των σταδίων σηματοδοτεί την εξαγωγή της προς επεξεργασία εικόνας. Μέχρι την στιγμή της λήψης της εικόνας ακολουθούνται πέντε βασικά στάδια, όπως φαίνονται και στο Σχήμα 4.

Αρχικά διευκρινίζεται ο λόγος για τον οποίο διεξάγεται το κάθε πείραμα. Ανάλογα με το βιολογικό ερώτημα που τίθεται επιλέγεται ο τύπος των μικροσυστοιχιών που πρέπει να χρησιμοποιηθεί, το είδος των αλληλουχιών που πρέπει να απομονωθούν, ποιοι δηλαδή συγκεκριμένα θα είναι οι ανιχνευτές που θα χρησιμοποιηθούν. Στην συνέχεια λαμβάνει χώρα η κατασκευή της μικροσυστοιχίας, κατά την οποία εκτυπώνονται οι ανιχνευτές επάνω στο υπόστρωμα. Ακολουθεί η ανάμειξη και η υβριδοποίηση των φθοριζόντων δειγμάτων με τους ανιχνευτές. Τέλος η μικροσυστοιχία σαρώνεται με σκοπό την εξαγωγή των εικόνων.



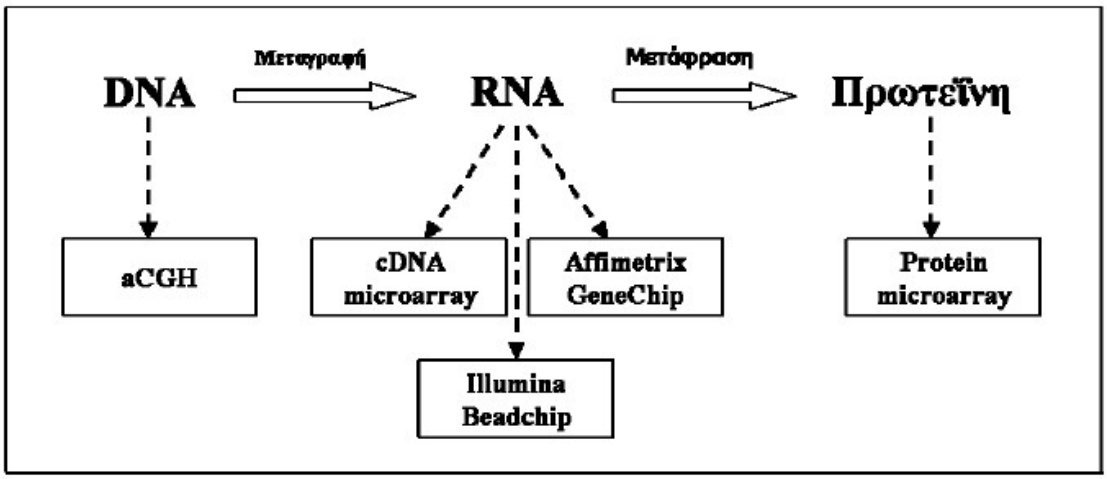
Σχήμα 4: Τα στάδια του κύκλου ζωής των μικροσυστοιχιών τα οποία προηγούνται της επεξεργασία των εξαγόμενων εικόνων.

1.2.1. Βιολογικό Ερώτημα

Από την πρώτη εμφάνιση των μικροσυστοιχιών, έχει αναπτυχθεί ένα ευρύ φάσμα εφαρμογών. Η πιο βασική από αυτές έχει να κάνει με κλινικές εφαρμογές, όπως για παράδειγμα η διάγνωση και η θεραπεία ασθενειών. Μέσω των μικροσυστοιχιών μπορεί να βρεθεί η προδιάθεση κάποιου ανθρώπου να εμφανίσει ένα νόσημα, αλλά και να προβλεφθούν επικίνδυνοι παράγοντες κατά την διάρκεια της θεραπείας του. Πέραν των κλινικών εφαρμογών οι μικροσυστοιχίες έχουν εφαρμογές στην ανακάλυψη και ανάπτυξη φαρμάκων, αλλά και στην ανάλυση των τροφών.

Τα διαφορετικά είδη μικροσυστοιχιών μπορούν επίσης να κατηγοριοποιηθούν ανάλογα με

της βιολογικές αλληλουχίες, οι οποίες χρησιμοποιούνται και οι οποίες υποβάλλονται σε υβριδοποίηση μαζί με τα προς εξέταση δείγματα. Στο Σχήμα 5. παρουσιάζονται τα είδη των μικροσυστοιχιών που έχουν αναπτυχθεί μέχρι σήμερα, σε αντιστοιχία με το είδος των βιολογικών αλληλουχιών.



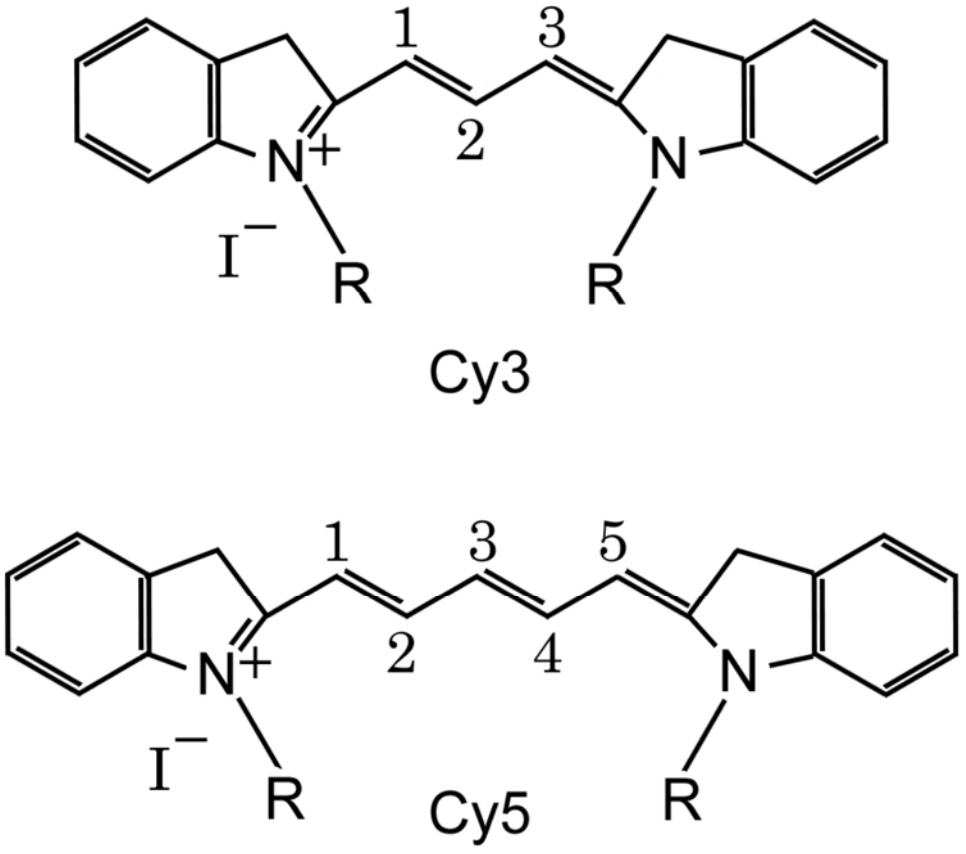
Σχήμα 5.: Κατηγοριοποίηση των ειδών των μικροσυστοιχιών ανάλογα με της βιολογικές αλληλουχίες που χρησιμοποιούνται.

Όπως αναφέρθηκε και στην εισαγωγή η πρώτη εφαρμογή των μικροσυστοιχιών αφορά στην έκφραση των γονιδίων. Οι μικροσυστοιχίες αυτές έκαναν χρήση μονόκλωνων αλληλουχιών συμπληρωματικού DNA, οι οποίες υβριδοποιούνταν με αλληλουχίες RNA. Σκοπός τους είναι η ποσοτικοποίηση των επιπέδων της έκφρασης κάθε γονιδίου μέσω του ποσοστού της υβριδοποίησης των δειγμάτων με τους ανιχνευτές. Την ίδια εφαρμογή εξυπηρετεί και το Genechip της εταιρίας Affimetrix, το οποίο διατέθηκε λίγα χρόνια αργότερα στην αγορά. Τέλος αλληλουχίες από RNA κάνει χρήση και το Beadchip τις εταιρίας Illumina. Η συγκεκριμένη όμως μικροσυστοιχία έχει δημιουργηθεί για τον εντοπισμό πολυμορφισμών στα δείγματα. Εκτός από τις μικροσυστοιχίες, η οποίες έχουν ως βάση τους το RNA, υπάρχουν και άλλες που έχουν ως βάση τους άλλες βιολογικές αλληλουχίες. Οι χρωμοσωματικές ανωμαλίες για παράδειγμα είναι ένα πεδίο που απαιτεί παραδοσιακά μεγάλη διερεύνηση, εξετάζοντας τον καρυότυπο των χρωμοσωμάτων. Τα πρώτα χρόνια αυτό συνέβαινε με την χρήση μικροσκοπίου και με την μέθοδο FISH (Fluorescence in situ Hybridization). Αργότερα και για αυτήν την εφαρμογή αναπτύχθηκαν οι μικροσυστοιχίες aCGH (array *Comparative* Genomic Hybridization), οι οποίες υβριδοποιούν αλληλουχίες DNA (Shinawi etal, 2008). Τέλος, παρά το γεγονός ότι οι DNA

και οι RNA μικροσυστοιχίες μπορούν να ποσοτικοποιήσουν με ακρίβεια τα επίπεδα έκφρασης ενός γονιδίου σε ένα κύτταρο, δεν μπορούν να μετρήσουν το ποσοστό των πρωτεϊνών, οι οποίες μεταφράζονται σε αυτό. Το γεγονός αυτό οδήγησε στην δημιουργία ενός νέου είδους μικροσυστοιχιών των πρωτεϊνικών μικροσυστοιχιών (Macbeath etal,2000).

1.2.2. Υβριδοποίηση των δειγμάτων με τους ανιχνευτές

Η βασική ιδέα σε ένα πείραμα μικροσυστοιχιών είναι ο προσδιορισμός του ποσοστού της υβριδοποίησης των ανιχνευτών με τα δείγματα. Ανάλογα με την βιολογική υπόθεση, έχουν επιλεγεί οι κατάλληλοι ανιχνευτές, οι οποίοι είναι εκτυπωμένοι επάνω στην μικροσυστοιχία από κατασκευής, όπως είδαμε στα προηγούμενα εδάφια. Στην συνέχεια λαμβάνονται και απομονώνονται τα προς εξέταση δείγματα. Ένα σημαντικό σημείο για την διαδικασία της υβριδοποίησης είναι η διάκριση των δειγμάτων. Τα δύο διαφορετικά δείγματα (π.χ. κανονικό και παθογόνο) πρέπει να είναι διακριτά μετά το πείραμα ώστε να συμπεράνουμε ποιο από τα δύο έχει υβριδοποιηθεί και ποιο όχι. Για τον λόγο αυτό, προτού λάβει χώρα η υβριδοποίηση των δειγμάτων με τους ανιχνευτές της μικροσυστοιχίας, στα προς εξέταση δείγματα προσκολλούνται φθορίζουσες ουσίες. Οι ουσίες που έχουν χρησιμοποιηθεί περισσότερο στην τεχνολογία των μικροσυστοιχιών είναι οι Cy3 και Cy5. Τις φθορίζουσες ουσίες Cy3 και Cy5 εισήγαγε πρώτος ο Ernst etal το 1989. Μια τυπική μορφή των δύο ουσιών φαίνεται στο σχήμα 6.



Σχήμα 6 : Οι δομή των φθοριζουσών ουσιών Cy3 και Cy5.

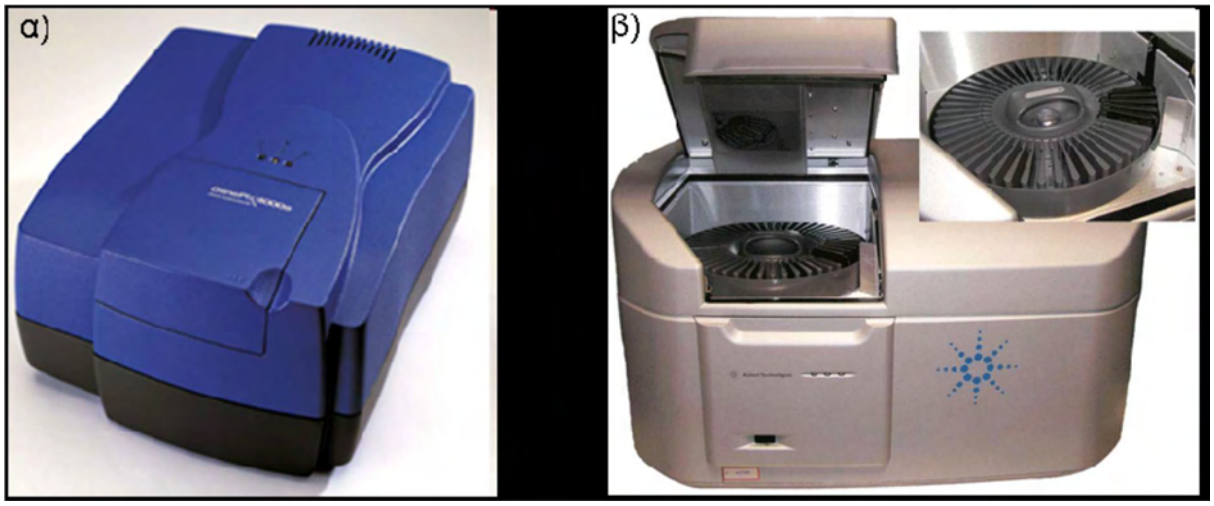
Οι ουσία Cy3 διεγείρεται κατά μέγιστο σε μήκος κύματος 550 nm ενώ εκπέμπει κατά μέγιστο σε μήκος κύματος 570 nm, δηλαδή στο παράθυρο του ορατού πράσινου του ηλεκτρομαγνητικού φάσματος. Αντίθετα η ουσία Cy5 διεγείρεται κατά μέγιστο σε μήκος κύματος 649 nm ενώ εκπέμπει κατά μέγιστο σε μήκος κύματος 670 nm, δηλαδή στο παράθυρο του ορατού κόκκινου του ηλεκτρομαγνητικού φάσματος. Πέραν των ουσιών Cy3 και Cy5 στην βιβλιογραφία έχουν χρησιμοποιηθεί και άλλες φθορίζουσες ουσίες.

Μετά την προσκόλληση των φθοριζουσών ουσιών τα δύο δείγματα αναμιγνύονται με νερό και τοποθετούνται στην μικροσυστοιχία σε όλες τις θέσεις όπου υπάρχουν ανιχνευτές.

Σύμφωνα με τις αρχές της υβριδοποίησης που εξετάσαμε στο προηγούμενο κεφάλαιο, οι συμπληρωματικές αλυσίδες δειγμάτων και ανιχνευτών θα δημιουργήσουν υβρίδια. Μετά από λίγες ώρες, η διαφάνεια πλένεται ώστε να απομακρυνθούν τα δείγματα, τα οποία δεν έχουν υβριδοποιηθεί. Στο σημείο αυτό το βιολογικό πείραμα έχει ολοκληρωθεί και οι μικροσυστοιχία είναι έτοιμη για να σαρωθεί ,και να εξαχθούν οι εικόνες.

1.2.3 Σάρωση

Αφού τα δείγματα υβριδοποιηθούν η μικροσυστοιχία τοποθετείται σε σαρωτή (scanner). Η σάρωση της μικροσυστοιχίας γίνεται και στα δύο μήκη κύματος, στα οποία εκπέμπουν οι φθορίζουσες ουσίες. Για της φθορίζουσες ουσίες Cy3 και Cy5 για παράδειγμα, οι σαρωτέςσυνήθως χρησιμοποιούν μήκη κύματος εκπομπής του λέιζερ 532 nm και 635 nm ενώ ταυτόχρονα χρησιμοποιούν ζωνοπερατά φίλτρα σε μήκη κύματος 550-600 nm και 655-695 nm με σκοπό την αποφυγή του θορύβου υποβάθρου. Με τον τρόπο αυτόν καθίσταται πιο εύκολος ο διαχωρισμός των δύο δειγμάτων, τα οποία εκπέμπουν σε διαφορετικά μήκη κύματος. Σκοπός της διαδικασίας της σάρωσης είναι βέβαια η ποσοτικοποίηση της φωτεινότητας που εκπέμπουν τα δείγματα στην θέση του κάθε ανιχνευτή. Στο παρακάτω Σχήμα 7. παρουσιάζονται δύο ευρέως χρησιμοποιούμενοι σαρωτές μικροσυστοιχιών, των εταιριών Genepix και Agilent.



Σχήμα 7: Δύο πολύ γνωστοί σαρωτές μικροσυστοιχιών a) της εταιρίας Genepix και β) της εταιρίας Agilent.

1.3 Συμπεράσματα

Από τα εδάφια τα οποία προηγήθηκαν εύκολα καταλήγουμε στο συμπέρασμα ότι το πεδίο των μικροσυστοιχιών είναι πολύπλοκο. Όλα τα επιμέρους στάδια έχουν αναπτυχθεί πολύπλευρα, γεγονός το οποίο έχει άμεση επίπτωση στα επόμενα στάδια και συγκεκριμένα στα στάδια που απαιτούν ανάπτυξη λογισμικού. Έτσι και η επεξεργασία της εικόνας των μικροσυστοιχιών καλείται να καλύψει την ποικιλομορφία των μικροσυστοιχιών, που έχει προκύψει από της διαφορετικές κατασκευαστικές και

βιολογικές προσεγγίσεις. Από τα εξειδικευμένα λογισμικά πακέτα του παρελθόντος πρέπει να περάσουμε σε γενικευμένες μεθοδολογίες και σουίτες εργαλείων.

2. Κανονικοποίηση στα δεδομένα μικροσυστοιχιών

Ο σκοπός σε ένα πείραμα μικροσυστοιχιών είναι να μετρήσουμε με ακρίβεια την αλλαγή στα επίπεδα έκφρασης των γονιδίων. Όπως συμβαίνει όμως σε όλα τα βιολογικά πειράματα πρέπει να λάβουμε υπόψη όλα τα σφάλματα (τυχαία και συστηματικά). Ένα σύνηθες συστηματικό σφάλμα προκύπτει από τις διαφορετικές φθορίζουσες ουσίες που χρησιμοποιούνται για σήμανση των εξεταζόμενων αλληλουχιών στα πειράματα μικροσυστοιχιών. Σε ένα παράδειγμα αυτό-υβριδισμού τοποθετούμε στη μικροσυστοιχία το ίδιο mRNA με διαφορετική χρωστική (Cy3 και Cy5). Μετά την υβριδοποίηση παρατηρούμε ότι σπανίως οι εντάσεις είναι ίδιες και μάλιστα συνήθως παρατηρείται μεγαλύτερη ένταση στις κουκίδες χρωματισμένες με πράσινη χρωστική (Cy3) σε σχέση με την κόκκινη (Cy5). Όσο και αν οι διαφορές αυτές φαίνονται ασήμαντες σε αναζήτηση μικρών βιολογικών διαφορών σε δείγματα, μπορεί να αποτελέσουν τροχοπέδη στην αξιοπιστία των αποτελεσμάτων (Yang etal 2002).

Οι κυριότερες πηγές συστηματικών σφαλμάτων στα πειράματα μικροσυστοιχιών είναι οι εξής:

- Κατασκευή μικροσυστοιχίας – έχουν να κάνουν με την τεχνολογία που χρησιμοποιείται
- Προετοιμασία mRNA – εξαγωγή από βιολογικά δείγματα, ενσωμάτωση χρωστικών στα δείγματα η οποία δεν πετυχαίνεται πάντα.
- Υβριδισμός – προέρχεται κυρίως από συνθήκες του περιβάλλοντος όπως η θερμοκρασία και η υγρασία καθώς και μόρια σκόνης που μπορεί να προσκολλούνται στη μικροσυστοιχία

- Τα νουκλεϊκά οξέα δεν απαιτούν τέλεια συμπληρωματικότητα για να επιτευχθεί υβριδισμός. Άρα μπορεί κάποιος ανιχνευτής να έχει αυξημένο σήμα από στόχευση πολλαπλών περιοχών και να συμπεράνουμε εσφαλμένα ότι ένα γονίδιο εκφράζεται πολύ υψηλά.

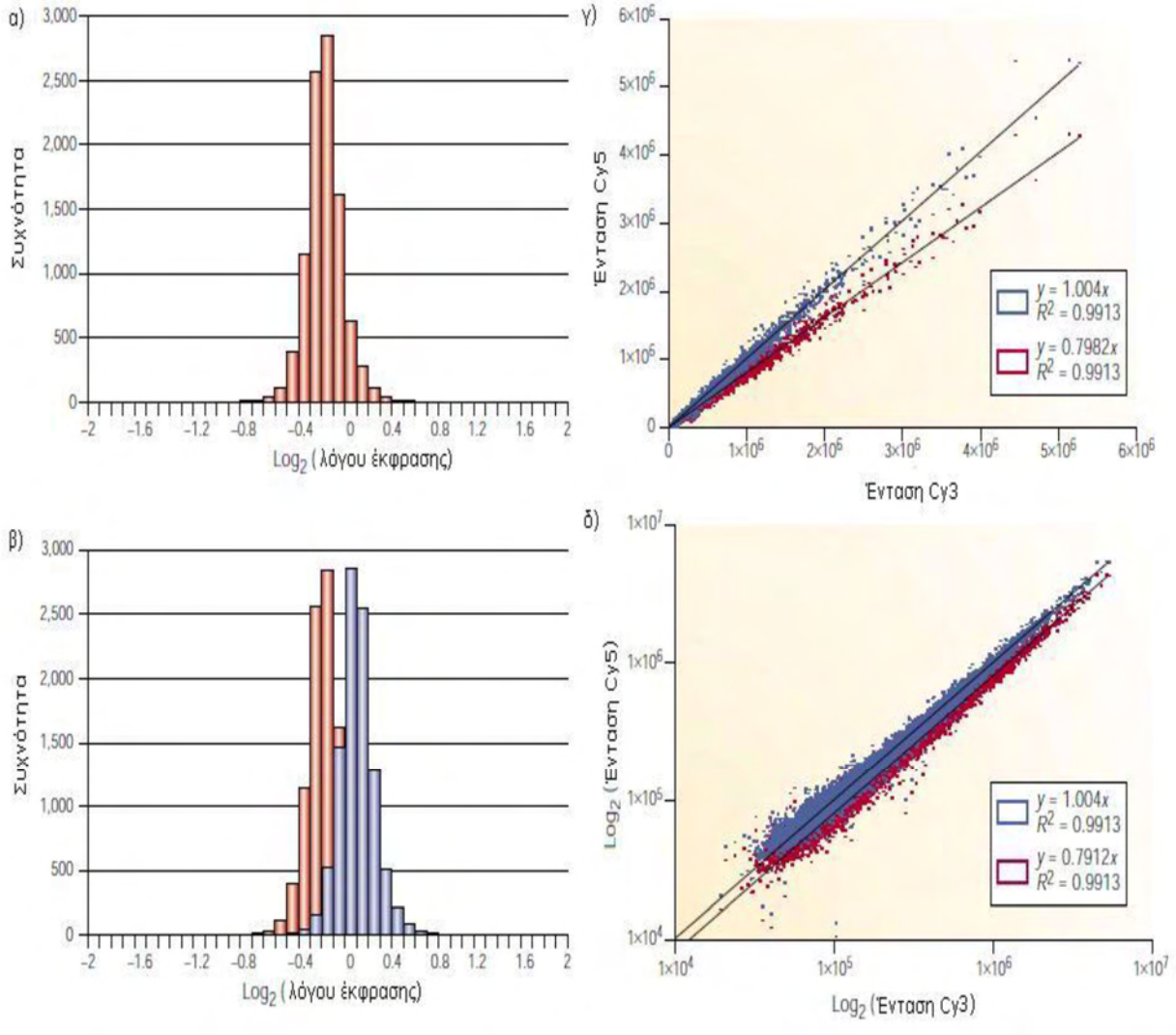
- Σάρωση – προβλήματα από το φθορισμό κατά το δέσιμο (binding) και την ένταση του σαρωτή
- Ανάλυση της εικόνας – χρησιμοποιώντας διαφορετικούς αλγορίθμους ανάλυσης εικόνας πετυχαίνουμε διαφορετικά αποτελέσματα φθορισμού (Parmigiani etal 2003).

Τα σφάλματα αυτά μεμονωμένα μπορεί να φαίνονται ασήμαντα, παρόλαυτα ο συνδυασμός τους μπορεί να αποφέρει σημαντική απόκλιση στα αποτελέσματα ενός πειράματος. Μπορούμε επομένως να περιμένουμε μεταβλητότητα στο ίδιο δείγμα mRNA όσον αφορά την έκφραση ενός γονιδίου σε διαφορετικά πειράματα. Σε μικροσυστοιχίες cDNA είναι σχετικά απλό να ποσοτικοποιήσουμε τις διαφοροποιήσεις πραγματοποιώντας ένα πείραμα αυτό-υβριδισμού στο οποίο δυο ίδια δείγματα προερχόμενα από κοινή RNA βιβλιοθήκη υβριδίζονται στην ίδια συστοιχία αλλά σημαίνονται με διαφορετικές χρωστικές.(Yang etal 2002).

2.1 Σκοπός κανονικοποίησης

Σε ένα πείραμα μικροσυστοιχιών υπάρχουν κάποια εκτιμώμενα επίπεδα στα οποία για τις τιμές στην έκφραση των γονιδίων. Η απόκλιση των τιμών αυτών οφείλεται ορισμένες φορές σε συστηματικά σφάλματα τα οποία μπορούμε να αποφύγουμε με τη μέθοδο της κανονικοποίησης. Με τον τρόπο αυτό μπορούμε να απαντήσουμε ευκολότερα στο αρχικό βιολογικό ερώτημα που τέθηκε κατά το σχεδιασμό του πειράματος αφού οι βιολογικές διαφορές γίνονται ευκολότερα αντιληπτές μεταξύ δεδομένων σε διαφορετικά πλακίδια.

Η βασική αρχή της κανονικοποίησης δεδομένων μικροσυστοιχιών που έχουν χρωματιστεί με δύο διαφορετικές χρωστικές (π.χ. Cy3 και Cy5) είναι η εξισορρόπηση των εντάσεων των χρωστικών αυτών. Μολονότι η κανονικοποίηση δεν είναι αρκετή ως μέθοδος για να εξαλείψει όλα τα συστηματικά σφάλματα που μπορεί να προκύψουν σε ένα πείραμα μικροσυστοιχιών παίζει πολύ σημαντικό ρόλο στα πρώιμα στάδια της ανάλυσης των δεδομένων σε τέτοια πειράματα. Αυτό φαίνεται από τη διαφοροποίηση αποτελεσμάτων τα οποία έχουν προκύψει με διαφορετικές μεθόδους κανονικοποίησης. Ανάλυση των πειραματικών δεδομένων με μεθόδους όπως η ομαδοποίηση (clustering) και τα νευρωνικά δίκτυα (neural networks) βασίζονται στην μέθοδο της αρχικής κανονικοποίησης (Yang etal 2001).



Σχήμα 8: Δεδομένα πριν και μετά την κανονικοποίηση. Α. Ιστόγραμμα των λογαριθμισμένων με βάση το 2 τιμών των λόγων έκφρασης που προκύπτουν από τον υβριδισμό δυο όμοιων δειγμάτων. Συνήθως παρατηρούνται υψηλότερα επίπεδα έκφρασης στο πράσινο κανάλι με αποτέλεσμα οι τι΄μές να είναι συγκεντρωμένες αριστερά του μηδενός. Β. Τα ίδια δεδομένα του προηγούμενου ιστογράμματος πριν (κόκκινο χρώμα) και μετά (μπλε χρώμα) την κανονικοποίηση. Η κανονικοποιημένη κατανομή των δεδομένων έχει μετατοπιστεί προς τα δεξιά και γύρω από το μηδέν. Γ. Γραφική παράσταση των επιπέδων έκφρασης των γονιδίων για το δύο κανάλια και για το ίδιο πείραμα πριν (κόκκινο χρώμα) και μετά (μπλέ χρώμα) την κανονικοποίση. Δ. Γραφική παράσταση των λογαριθμισμένων εντάσεων για τα δύο κανάλια και για το ίδιο πείραμα πριν (κόκκινο χρώμα) και μετά (μπλε χρώμα) την κανονικοποίηση (Quackenbush etal 2001).

2.2 Απεικόνιση δεδομένων σε μια μικροσυστοιχία

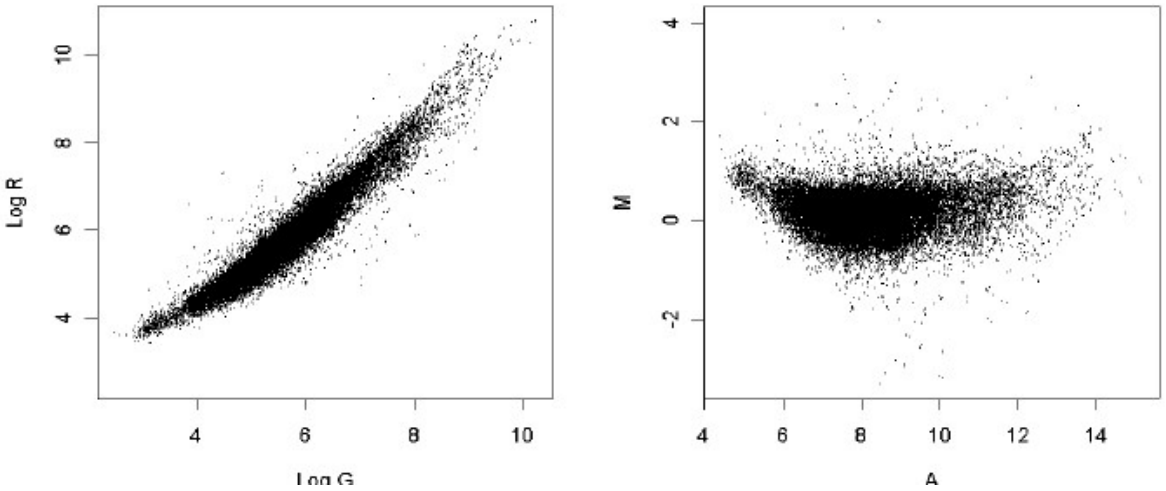
Η απεικόνιση στα δεδομένα ενός πειράματος μικροσυστοιχιών γίνεται με την απεικόνιση της γραφικής παράστασης του λογαρίθμου της έντασης της κόκκινης χρωστικής ($\log_2$R) έναντι του λογαρίθμου της έντασης της πράσινης χρωστικής ($\log_2$G). Επειδή η γραφική αυτή παράσταση δίνει μια μη ρεαλιστική συμφωνία για καλύτερη απεικόνιση χρησιμοποιείται ο λογαριθμικός λόγος των εντάσεων $M = \log_2(\frac{R}{G})$ έναντι της μέσης λογαριθμικής έντασης $A = \log_2(\sqrt{R*G})$.

Μία γραφική παράσταση M vs A ισοδυναμεί με μια περιστροφή 45^0 του συστήματος ($\log_2$R, $\log_2$G) με μια καινούρια διαβάθμιση των συντεταγμένων. Εάν θέσουμε M' και A' τις περιστρεμμένες συντεταγμένες τότε ισχύει:

$$A = A'/\sqrt{2}, M = M'/\sqrt{2}$$

Επομένως μια γραφική παράσταση M vs A είναι μια διαφορετική αναπαράσταση των (R,G) δεδομένων υπό όρους των λογαριθμικών εντάσεων M – A οι οποίοι είναι και οι τιμές που ενδιαφέρουν τα περισσότερα βιολογικά πειράματα (Εικόνα). Οι γραφικές παραστάσεις M vs A είναι περισσότερο χρήσιμες από τις ($\log_2$R,$\log_2$G) όσον αφορά την αναγνώριση ψεύτικων κουκκίδων, την ανίχνευση εξαρτώμενων προτύπων από την ένταση στους λογαριθμικούς λόγους κλπ. (Yang 2001).

Ένας ακόμη τρόπος να απεικονίσουμε μεροληψίες που οφείλονται στην ένταση σήματος είναι να δημιουργήσουμε ένα διάγραμμα με άξονες το $\log_2(R/G)$ και το $\log_{10}(R*G)$. Τα διαγράμματα αυτά είναι γνωστά και ως R – I (Ratio – Intensity) και αποκαλύπτουν σε μεγάλο βαθμό τις παρεκκλίσεις στις λογαριθμικές αναλογίες (Causton, Quackenbush & Brazma 2003, Quackenbush etal 2002).



Σχήμα 9 . α) Γραφική παράσταση (log_2R,log_2G) β) MA καμπύλη τα γονίδια που έχουν εκφραστεί φαίνονται καλύτερα

2.3 Επιλογή Γονιδίων για κανονικοποίηση

Σε κάθε βιολογικό πείραμα μικροσυστοιχιών πρέπει να διαλέξουμε το σύνολο των γονιδίων το οποίο θα υποστεί κανονικοποίηση για να προχωρήσουμε σε περαιτέρω ανάλυση. Την απόφαση αυτή την επηρεάζουν διάφοροι παράγοντες όπως η αναλογία των γονιδίων που είναι εκφρασμένη διαφορικά στα μαρκαρισμένα δείγματα με κόκκινη ή πράσινη χρωστική καθώς και η διαθεσιμότητα των DNA δειγμάτων ελέγχου. Ακολουθούν όλες οι περιπτώσεις επιλογής γονιδίων:

2.3.1 Όλα τα γονίδια της διάταξης

Τις περισσότερες φορές σε ένα πείραμα μικροσυστοιχιών το βιολογικό ερώτημα είναι συγκεκριμένο και επομένως μόνο ένα μικρό ποσοστό των γονιδίων είναι εκφρασμένο διαφορικά. Τα υπόλοιπα γονίδια έχουν σταθερές εντάσεις και μπορούν να χρησιμοποιηθούν ως δείκτες των σχετικών εντάσεων των δυο χρωστικών. Επομένως μπορούμε να κανονικοποιήσουμε όλα τα γονίδια σε μια μικροσυστοιχία εάν ισχύουν οι παρακάτω παράμετροι:

- Ένα μικρό ποσοστό γονιδίων διαφέρει σημαντικά μεταξύ των δυο δειγμάτων που υβριδοποιούνται
- Αν υπάρχει συμμετρία ανάμεσα στα υπερ- και υπο- εκφρασμένα γονίδια

2.3.2 Γονίδια ελέγχου

Στην τεχνική αυτή χρησιμοποιείται σύνολο γονιδίων. Τα γονίδια αυτά μπορεί να είναι γονίδια διαχείρισης (housekeeping genes) ή εξωγενή γονίδια ελέγχου (exogenous controls). Τα γονίδια διαχείρισης (housekeeping genes) συνήθως έχουν σταθερή έκφραση κάτω από διαφορετικές συνθήκες. Είναι γενικότερα πολύ δύσκολο να βρεθούν γονίδια τα οποία έχουν σταθερή έκφραση υπό διαφορετικές συνθήκες. Μπορούμε όμως να βρούμε σύνολα γονιδίων που προσωρινά έχουν σταθερή έκφραση κάτω από συγκεκριμένες πειραματικές συνθήκες. Αυτά τα συγκεκριμένα όμως γονίδια έχουν ιδιαίτερα υψηλό επίπεδο έκφρασης, πράγμα το οποίο τα καθιστά μη αντιπροσωπευτικά για το σύνολο των γονιδίων.

Μπορούμε σε ένα πείραμα μικροσυστοιχιών να κάνουμε χρήση εξωγενών γονιδίων ελέγχου αντί για των γονιδίων διαχείριση (housekeeping genes). Αυτά μπορεί να είναι δυο ειδών γονίδια:

- Γονίδια τα οποία έχουν εμβολιαστεί στο δείγμα (spiked controls). Στη μέθοδο αυτή αλληλουχίες συνθετικού DNA ή αλληλουχίες DNA από διαφορετικό οργανισμό τοποθετούνται στη μικροσυστοιχία και συμπεριλαμβάνονται στα δύο διαφορετικά mRNA. Κατά τον τρόπο αυτό θα έχουν ίσες εντάσεις φθορισμού για την κόκκινη και την πράσινη χρωστική και μπορούν να χρησιμοποιηθούν για κανονικοποίηση.
- Τιτλοδοτημένες ακολουθίες ελέγχου (titration series). Στη μέθοδο αυτή κουκίδες που αποτελούνται από διαφορετικές συγκεντρώσεις του ίδιου γονιδίου ή αλληλουχίες έκφρασης επισήμανσης (expressed sequence tag – EST) τυπώνονται στη μικροσυστοιχία. Αυτές οι κουκίδες έχουν την ίδια ένταση φθορισμού για την κόκκινη και την πράσινη χρωστική σε όλο το εύρος των εντάσεων (Yang, Quackenbush & Brazma 2003).
- Γονίδια δεξαμενής δειγμάτων (microarray sample pool data). Συνήθως χρησιμοποιούνται για να εξαλείψουν επιδράσεις εξαρτώμενες από την ένταση. Στα γονίδια δεξαμενής διάφοροι replicates αναμιγνύονται δημιουργώντας μια κοινή δεξαμενή(pool). Τα συγκεκριμένα γονίδια μπορεί να αποτελέσουν έναν πιθανό ανιχνευτή για οποιοδήποτε σημασμένο στόχο. Οι συγκεντρώσεις των καταλαμβάνουν όλο σχεδόν το εύρος των εντάσεων.

2.3.3 Γονίδια σταθερής κατάταξης

Γονίδια τα οποία δεν έχουν εκφραστεί διαφορικά μπορούν να αναγνωριστούν με μαθηματικά κριτήρια και όχι μόνο με βιολογικά και συγκεκριμένα με κατάλληλους αλγορίθμους. Η πρώτη εφαρμογή της ιδέας αυτής έγινε από τους (Tseng etal 2001), από τους οποίους προτάθηκε μια

μέθοδος σταθερής κατάταξης (rank – invariant meth od) για την επιλογή του συνόλου των γονιδίων πάνω στο οποίο θα βασιστεί η κανονικοποίηση (σύνολο γονιδίων σταθερής κατάταξης – rank invariant gene set).

Όταν οι κατατάξεις για την πράσινη και κόκκινη χρωστική ενός γονιδίου *κ* είναι περίπου ίσες και αν η κατάταξη των δύο καναλιών (Cy3 και Cy5) δεν είναι εξαιρετικά χαμηλή ή υψηλή τότε το γονίδιο μπορεί να θεωρηθεί μη διαφορικά εκφρασμένο και επομένως μπορεί να συμπεριλήφθη στο σύνολο των γονιδίων σταθερής κατάταξης και να χρησιμοποιηθεί στην κανονικοποίηση.

Οι (Tseng etal, 2001) όρισαν δυο κατώφλια d και l τα οποία πρέπει να ικανοποιούν τα γονίδια για να συμπεριληφθούν στο σύνολο γονιδίων σταθερής κατάταξης:

$$| Rank(R_k) - Rank(G_k) | < d \text{ και } l < Rank\left\{R_k + G_k / 2\right\} < G_A - l$$

Όπου R_K και G_k οι τιμές έκφρασης του γονιδίου k στο δείγμα ελέγχου και στο δείγμα αναφοράς αντίστοιχα, G_A ο συνολικός αριθμός γονιδίων Rank(I) η κατάταξη της έντασης φθορισμού I για το σύνολο των γονιδίων G_A.

3. Ο αλγόριθμος RMA

Ο αλγόριθμος RMA της Affymetrix Gene Chips ® πραγματοποιεί διαδοχικά τις παραπάνω διαδικασίες. Η διαφορά του συγκεκριμένου αλγορίθμου από όλους τους υπολοίπους έγκειται στο γεγονός ότι είναι σχεδιασμένος για να χρησιμοποιείται με τη νέα τεχνολογία των chip όπου πραγματοποιείται μόνο Perfect Match υβριδισμός μεταξύ ολιγονουκλεοτιδίων στόχων και ανιχνευτών. Δηλαδή, ο αλγόριθμος RMA υπολογίζει τις τιμές έντασης φθορισμού από την υβριδοποίηση των ολιγονουκλεοτιδίων στόχων με τα ολιγονουκλεοτίδια ανιχνευτές που υπάρχουν στο chip. Τα ολιγονουκλεοτίδια ανιχνευτές (probes) είναι συμπληρωματικά των ολιγονουκλεοτίδια στόχων που προέρχονται από τα προς μελέτη δείγματα. Η διόρθωση υποβάθρου που χρησιμοποιείται στον RMA είναι μια μη γραμμική διόρθωση, η οποία πραγματοποιείται σε κάθε chip ξεχωριστά και βασίζεται στην κατανομή των τιμών των PM ανιχνευτών που βρίσκονται στο chip της Affymetrix. Ο θόρυβος υπολογίζεται βάση των τιμών φθορισμού 17.000 περίπου ανιχνευτών που χρησιμοποιούνται ως αναφορά. Οι τιμές έντασης φθορισμού των PM ανιχνευτών αποτελούν μίξη του θορύβου υποβάθρου του οποίου το σήμα (N) έχει κανονική κατανομή και του σήματος που προσπαθούμε να απομονώσουμε αφαιρώντας των θόρυβο. Ο θόρυβος υποβάθρου (E) στον αλγόριθμο RMA υπολογίζεται ως ο μέσος όρος του 23

επιθυμητού σήματος (S) υπό τη συνθήκη των παρατηρούμενων τιμών των PM ανιχνευτών (O) χρησιμοποιώντας μια μέθοδο εκτίμησης πιθανότητα πυρήνα:

$$E(S \mid O = o) = a + b\frac{\varphi\left(\frac{\alpha}{b}\right) - \varphi\left(\frac{o-\alpha}{b}\right)}{\Phi\left(\frac{\alpha}{b}\right) - \Phi\left(\frac{o-\alpha}{b}\right) - 1}$$

$$a = o - \mu - \sigma^2\alpha, b = \sigma$$

Όπου η αναμενόμενη τιμή (E) μιας τυχαίας μεταβλητής είναι ο σταθμισμένος μέσος όρος όλων των πιθανών τιμών που αυτή η τυχαία μεταβλητή μπορεί να πάρει. Θόρυβος υποβάθρου (E), ένταση σήματος υβριδισμού (S), ένταση σήματος υβριδισμού των 17.000 ανιχνευτών αναφοράς (O).

Για την πραγματοποίηση της διαδικασίας της κανονικοποίησης ο RMA αλγόριθμος χρησιμοποιεί τη μέθοδο της ποσοστημόριας κανονικοποίησης (quantile normalization). Στην ποσοστημόρια κανονικοποίηση σε δυο ή περισσότερες κατανομές πού δεν έχουν κατανομή αναφοράς εξάγεται ο μέσος όρος. Έτσι, η υψηλότερη τιμή σε όλες τις περιπτώσεις των κατανομών γίνεται ο μέσος όρος των υψηλότερων τιμών, η δεύτερη υψηλότερη τιμή καθίσταται ο μέσος όρος των αμέσως χαμηλότερων τιμών και ούτω καθεξής. Η συγκεκριμένη τεχνική θεωρείται πολύ αποτελεσματική κάνοντας τις κατανομές πανομοιότυπες μεταξύ τους(Bolstad etal 2003).

Η διαδικασία της συνόψισης βασίζεται στην υπόθεση ότι ο λογάριθμος των παρατηρούμενων τιμών εντάσεων των PM ανιχνευτών ακολουθούν ένα γραμμικό προσθετικό μοντέλο. Η παραγόμενη αριθμητική ακολουθία περιλαμβάνει 3 σήματα: τον όρο ομοιότητας της έντασης μεταξύ των ανιχνευτών, έναν όρο εξαρτώμενο από το γονίδιο (η τιμή έκφρασης) και έναν όρο σφάλματος. Ο όρος της συγγένειας/ομοιότητας μεταξύ των ανιχνευτών αθροίζει στο 0, ενώ οι τιμές εξαρτώμενες από το γονίδιο υπολογίζονται με τη χρήση του αλγορίθμου median polish. Με αυτό τον τρόπο ο RMA προβλέπει και προστατεύει από την πρόσμιξη ακραίων τιμών (outliers) με αυτές του μέσου όρου.

4. Ο αλγόριθμος GCRMA

Ο αλγόριθμος GCRMA χρησιμοποιεί τόσο την ποσοτική κανονικοποίηση όσο και τη μέθοδο της μέσης σύνοψης που χρησιμοποιείται στον αλγόριθμο RMA. Διαφέρει όμως στη διαδικασία για το

υπόβαθρο. Για να περιγραφούν οι παρατηρήσεις των εντάσεων του σήματος για κάθε γονίδιο στη μικροσυστοιχία χρησιμοποιείται ένα στοχαστικό μοντέλο. Πιο συγκεκριμένα το μοντέλο είναι:

$$PM_{ni} = O_{ni} + N_{1ni} + S_{ni}$$

$$MM_{ni} = O_{ni} + N_{2ni}$$

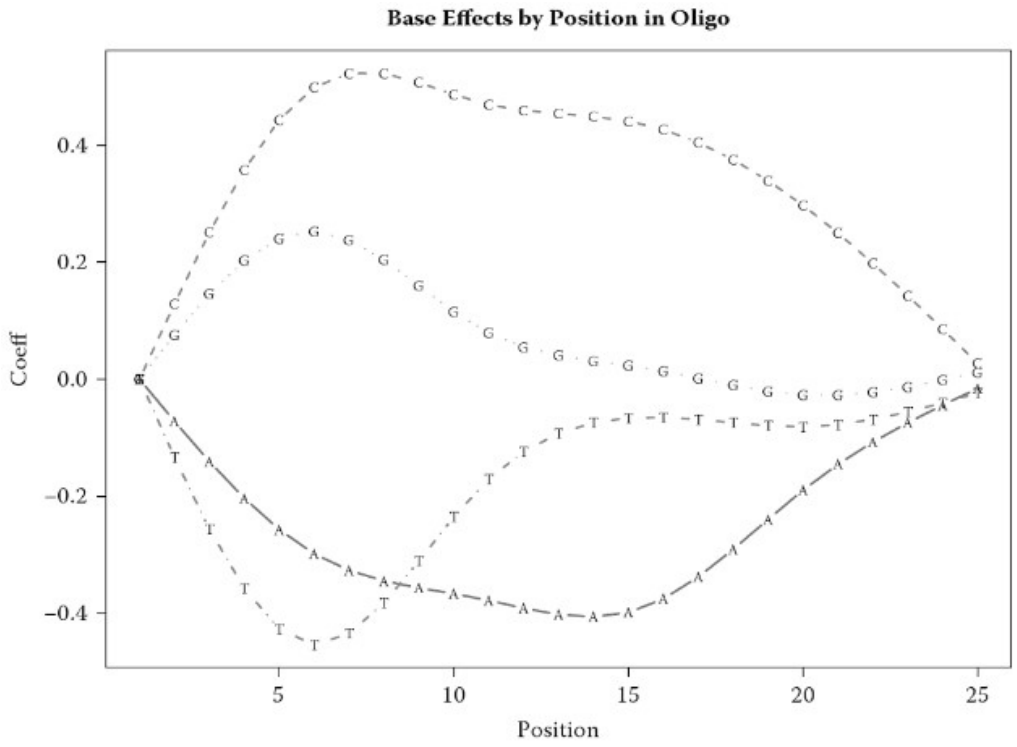
Όπου το O_{ni} είναι ο οπτικός θόρυβος, N_1 και N_2 μη συγκεκριμένος υβριδισμός και S_{ni} είναι μια ποσότητα ανάλογη της έκφρασης του RNA στο δείγμα. Επιπλέον, το μοντέλο προυποθέτει ότι το O ακολουθεί μια κανονική κατανομή $N(\mu_o,\sigma_o^2)$ και ότι οι λογάριθμοι $\log_2(N_{lni})$ και $\log_2(N_{2ni})$ ακολουθούν μια δυσδιάστατη κανονική κατανομή με ίσες μεταβλητές σ_N^2 και συσχέτιση 0,7 σταθερή σε όλα τα ζεύγη που υβριδοποιούνται. Ο μέσος της κατανομής εξαρτάται από την αλληλουχία που υβριδοποιείται. Ο οπτικός θόρυβος θεωρείται ότι είναι ανεξάρτητος.

Η μέθοδος με την οποία ο GCRMA συμπεριλαμβάνει πληροφορίες σχετικά με την αλληλουχία που υβριδοποιείται υπολογίζει μια συγγένεια η οποία βασίζεται σε ένα άθροισμα από συγγένειες εξαρτώμενες από τη θέση πάνω στη μικροσυστοιχία. Πιο συγκεκριμένα η συγγένεια μιας αλληλουχίας ορίζεται ως:

$$A = \sum_{k=1}^{25} \sum_{b\in\{A,C,G,T\}} \mu_b(k) 1_{\beta_\kappa=\xi}$$

Όπου $\mu_b(\kappa)$ μοντελοποιείται σαν εξίσωση με 5 βαθμούς ελευθερίας.

Η εικόνα δείχνει μια προσέγγιση του $\mu_b(k)$ για μια μοναδική μικροσυστοιχία τύπου U133A.



Σχήμα 10 : Βάσεις νουκλεοτιδίων σε συγκεκριμένες θέσεις χρησιμοποιώντας δεδομένα από ένα U133A τσιπ μικροσυστοιχίας

Στην πραγματικότητα αυτά υπολογίζονται χρησιμοποιώντας τα δεδομένα από το σύνολο της μικροσυστοιχίας σε ένα πείραμα ή βασίζονται σε κάποιες προσεγγίσεις που δίνονται από ένα NBS πείραμα το οποίο σχεδιάστηκε και εκτελέστηκε από τους δημιουργούς του αλγορίθμου GCRMA. Οι μέσοι για τις τυχαίες μεταβλητές N_1 και N_2 στον αλγόριθμο υπολογίζονται χρησιμοποιώντας μια ομαλή εξίσωση ή τις συγγένειες υβριδοποίησης.

Οι παράμετροι του οπτικού θορύβου μ_o,σ_o^2 υπολογίζονται ακολούθως:

Η μεταβλητότητα που οφείλεται στον οπτικό θόρυβο είναι πολύ μικρότερη από τη μεταβλητότητα της υβριδοποίησης και επομένως θεωρείται ουσιαστικά σταθερή. Για περισσότερη ευκολία θεωρείται 0. Μετά την παραδοχή αυτή οι μέσες τιμές υπολογίζονται χρησιμοποιώντας την μικρότερη PM ή MM ένταση στη μικροσυστοιχία λαμβάνοντας υπόψη έναν παράγοντα συσχέτισης για να αποφύγουμε τις αρνητικές τιμές. Έτσι, όλες οι τάσεις διορθώνονται αφαιρώντας την τάση μ_o. Για να υπολογίσουμε την $h(A_{ni})$ τοποθετούμε μια καμπύλη σε μια γραφική παράσταση στην οποία σχετίζονται οι διορθωμένοι log(MM) των εντάσεων με τις MM υβριδοποιήσεις. Οι αρνητικές τιμές από αυτή τη γραφική παράσταση χρησιμοποιούνται για τον υπολογισμό του σ_N^2. Τέλος, για τη διαδικασία διόρθωσης του υποβάθρου υπολογίζουμε την τιμή S με δεδομένα τις παρατηρούμενες εντάσεις PM, MM για το συγκεκριμένο πείραμα(Irizarry etal 2003).

5 Ο αλγόριθμος MAS5.0

Ο αλγόριθμος MAS5.0 είναι σχεδιασμένος να λειτουργεί σε ένα μεμονωμένο chip μικροσυστοιχίας. Χρησιμοποιούνται τόσο οι εντάσεις PM όσο και οι MM. Αποτελείται από μια ρύθμιση υποβάθρου που αναφέρεται σε συγκεκριμένη τοποθεσία της μικροσυστοιχίας, διόρθωση των εντάσεων PM με βάση τις εντάσεις MM. Η κανονικοποίηση αποτελείται από μια γραμμική ρύθμιση η οποία διεκπεραιώνεται μετά τον υπολογισμό του συνολικού αθροίσματος. Ο σκοπός του βήματος αυτού είναι η αφαίρεση του συνολικού θορύβου υποβάθρου. Κάθε μικροσυστοιχία διαιρείται σε ένα σύνολο από πλέγματα. Εξ' ορισμού χωρίζεται σε 16 ίσες σε μέγεθος περιοχές τοποθετημένες ώστε να σχηματίζουν ένα πλέγμα 4x4. Για κάθε πλέγμα υπολογίζεται μια τιμή υποβάθρου (b_k) και μια τιμή θορύβου (n_k) χρησιμοποιώντας τον μέσο και την τυπική απόκλιση του 2% των εντάσεων στη συγκεκριμένη περιοχή πλέγματος. Μετά, κάθε ένταση ρυθμίζεται χρησιμοποιώντας το σταθμισμένο μέσο όρο για κάθε τιμή έντασης. Τα βάρη εξαρτώνται από την απόσταση από το κέντρο του πλέγματος. Ειδικότερα τα βάρη δίνονται από τους παρακάτω τύπους:

$$w_k(x,y) = \frac{1}{d_k^2(x,y) - \delta}$$

Όπου δ_k^2 (x,y) είναι η Ευκλείδεια απόσταση από μια τοποθεσία (x,y) στο κέντρο της περιοχής k. Η τιμή του δ είναι 100. Για να αποφύγουμε τις αρνητικές τιμές για τις περιοχές με χαμηλή ένταση χρησιμοποιούμε τους εκτιμητές θορύβου. B(x,y) είναι ο β σταθμισμένος όρος του b_k στην τοποθεσία (x,y), N(x,y) είναι ο σταθμισμένος μέσος όρος του n_k στην τοποθεσία (x,y). Θέτουμε $P_{x,y}$ την ένταση της τοποθεσίας (x,y) στη μικροσυστοιχία, η σωστή τιμή δίνεται από τον τύπο

$$P_{x,y} = \max(P_{x,y} - B(x,y), N_f N(x,y))$$

Όπου N_j είναι μια μεταβλητή παράμετρος που θέτουμε εξ ορισμού 0,5.

Αρχικά η χρήση των MM εντάσεων ήταν να ρυθμίζουν τις PM εντάσεις αφαιρώντας την ένταση MM από το αντίστοιχο PM. Όμως αυτό δεν είναι πάντα εφικτό σε μια τυπική συστοιχία διότι το 30% των εντάσεων MM έχουν μεγαλύτερη τιμή από την PM ένταση. Επομένως, όταν οι πραγματικές MM εντάσεις αφαιρούνται από τις PM εντάσεις είναι πιθανό να προκύψουν αρνητικές τιμές. Ένα ακόμη πλεονέκτημα τις μεθόδου είναι ότι οι αρνητικές τιμές αποκλείουν τη χρήση λογαρίθμων που έχουν αποδειχθεί χρήσιμοι σε πολλούς υπολογισμούς μικροσυστοιχιών. Για να διορθωθεί η αρνητική αυτή επίδραση των πραγματικών τιμών εντάσεων εισάγουμε τη μεταβλητή IM(Ideal Mismatch) που

έχει εξ ορισμού θετική τιμή. Η ιδέα είναι να χρησιμοποιείται η τιμή των εντάσεων MM όταν είναι μικρότερες από τις PM και στις περιπτώσεις που αυτό δεν ισχύει, μια τιμή μικρότερη της PM έντασης.

Καταρχήν υπολογίζεται μια ποσότητα που ονομάζεται Specific Background (SB) για κάθε αλληλουχία που έχει υβριδοποιηθεί. Ο υπολογισμός γίνεται παίρνοντας τον μέσο όρο του λόγου των λογαρίθμων PM και MM.

Αν i είναι η αλληλουχία και n είναι το ζεύγος αλληλουχιών μετά την υβριδοποίηση τότε για το ζεύγος με συντεταγμένες i και n το ιδανικό mismatch δίνεται από:

$$IM_{ni} = \begin{cases} MM_{ni} \\ \dfrac{PM_{ni}}{2^{SB_n}} \\ \dfrac{PM_{ni}}{2^{\tau_c/(1+(\tau_c - SB_n)/\tau_s)}} \end{cases}$$

Όταν $MM_{ni}<PM_{ni}$

Όταν $MM_{ni}>PM_{ni}$ και $SB_n>\tau_c$

Όταν $MM_{ni}>PM_{ni}$ και $SB_n<\tau_c$

Όπου τ_c και τ_s είναι σταθερές, τ_c είναι η παράμετρος αντίθεσης (contrast parameter) και η τιμή της είναι 0,03 εξ ορισμού και τ_s είναι η παράμετρος κλιμάκωσης (scaling factor) η τιμή της οποίας είναι 10. Η ρυθμιζόμενη ένταση PM δίνεται αφαιρώντας τον αντίστοιχο IM.

Ο αλγόριθμος MAS5.0 χρησιμοποιεί μια μέθοδο άθροισης για κάθε μικροσυστοιχία. Η μέθοδος αυτή είναι πιο συγκεκριμένα αυτή που προτάθηκε από τους Hubbell, Lia και Mei

$$u_{nij} = \frac{\log_2(y_{nij}) - M}{cS + \varepsilon}$$

Όπου M είναι ο μέσος του $\log_2(y_{nij})$ και S ο μέσος της απόλυτης απόκλισης (MAD) από τον μέσο M. Εδώ το c=S σταθερά και το ε=0,0001 το οποίο επιλέγεται για να εξαληφθεί το πρόβλημα της διαίρεσης με το). Τα βάρη υπολογίζονται από τους παρακάτω τύπους:

$$w(u) = \begin{cases} 0 \\ (1-u^2) \end{cases}$$

Όταν |u|>1	
Όταν |u|<1	

Τέλος η έκφραση δίνεται από τον τύπο

$$\beta_{nij} = \frac{\sum_{i=1}^{l_n} w(u_{nij}) \log_2(y_{nij})}{\sum_{i=1}^{l_n} w(u_{nij})}$$

Στον αλγόριθμο MAS5.0 η κανονικοποίηση γίνεται μετά τον υπολογισμό των εντάσεων για κάθε μικροσυστοιχία. Μια κλιμακωτή κανονικοποίηση χρησιμοποιείται για να ρυθμιστεί κάθε μικροσυστοιχία. Πιο συγκεκριμένα, για κάθε μικροσυστοιχία υπολογίζεται ένας μέσος για τις εντάσεις αφαιρώντας το 2% του μεγίστου και του ελαχίστου των δεδομένων. Μετά επιλέγεται μία συγκεκριμένη ένταση είτε ο μέσος όρος της μικροσυστοιχίας είτε μια τιμή που έχει επιλεχθεί από πριν. Κάθε τιμή έκφρασης στη μικροσυστοιχία ρυθμίζεται με βάση τον παρακάτω τύπο

$$\varphi_j = \frac{mean_n(B_{nj})}{T}$$

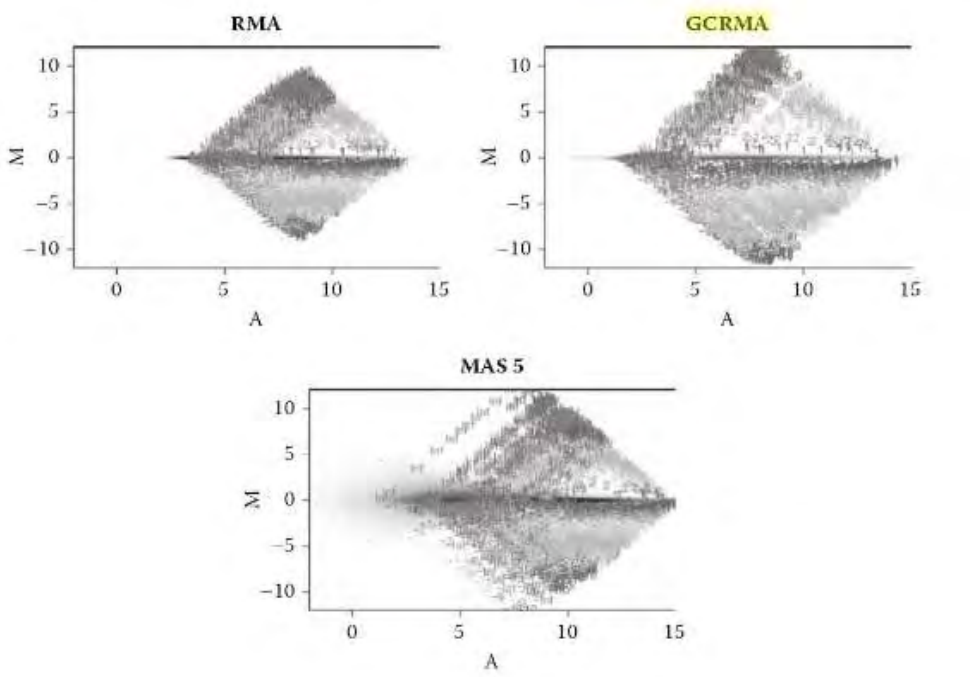
Όπου T είναι η επιλεγμένη ένταση(Hubbell etal 2002).

6. Συμπεράσματα

Από τα προαναφερθέντα συμπεραίνουμε πως διαφορετικοί αλγόριθμοι παράγουν διαφορετικά αποτελέσματα όσον αφορά τις εντάσεις σε ένα πείραμα μικροσυστοιχιών, για να αναγνωρίσουμε βιολογικά σημαντικές εντάσεις γονιδίων. Για να βρούμε πιο εύκολα τις διαφορές των αλγορίθμων πρέπει να πραγματοποιήσουμε ένα πείραμα μικροσυστοιχιών με συγκεκριμένες DNA αλληλουχίες. Ένα τυπικό τέτοιο πείραμα αποτελείται από αλληλουχίες των οποίων η συγκέντρωση πάνω στη μικροσυστοιχία έχει πειραχθεί ώστε το RNA υποβάθρου να μην είναι το ίδιο. Η Affymetrix έχει δώσει δυο διαφορετικές τέτοιες μελέτες, μία στη U95 πλατφόρμα μικροσυστοιχιών και μία στην πλατφόρμα U133A. Οι μελέτες είναι διαθέσιμες στη σελίδα της Affymetrix (www.affymetrix.com)

Αυτό που παρατηρούμε στις μελέτες είναι συνήθως η ακρίβεια, δηλαδή τα γονίδια τα οποία έχουν εκφραστεί. Στην εικόνα δείχνεται μια σύγκριση των τριών αλγορίθμων (RMA, GCRMA,

MAS5.0)χρησιμοποιώντας την πλατφόρμα μικροσυστοιχιών U133A. Στα αριστερά 3 γραφικές παραστάσεις MA δείχνουν πως οι εντάσεις διαφοροποιούνται στο πειραματικό data-set.



Σχήμα 11. Σύγκριση των αλγορίθμων RMA, GCRMA και MAS5.0 χρησιμοποιώντας dataset με τοποθετημένα γονίδια στην πλατφόρμα U133A.

Όπως φαίνεται από το διάγραμμα ο αλγόριθμος MAS5.0, είναι πολύ πιο θορυβώδης στις μικρότερες εντάσεις από τους RMA και GCRMA. Αυτό κάνει πιο δύσκολη τη διάκριση των διαφορικά εκφρασμένων γονιδίων. Στα δεξιά της Εικόνας, φαίνεται μια γραφική παράσταση με τις γενικές τάσεις που δημιουργήθηκε αναπαριστώντας τον λογάριθμο (log_2) της έκφρασης με το λογάριθμο log_2 της συγκέντρωσης. Η γραφική αυτή παράσταση μας δείχνει ότι ο MAS5.0 είναι περισσότερο γραμμικός σε όλο το εύρος των συγκεντρώσεων ενώ ο RMA, είναι ο λιγότερο γραμμικός. Η μεγαλύτερη κλίση του GCRMA δείχνει ότι οι εκφράσεις των γονιδίων που προκύπτουν από την εφαρμογή του αλγορίθμου αυτού θα είναι λιγότερο εξασθενημένες όσον αφορά τις εντάσεις(Cope etal 2004, Irizarry etal 2003).

Η κανονικοποίηση των δεδομένων σε ένα πείραμα μικροσυστοιχιών είναι απαραίτητη λόγω των μεταβλητών συνθηκών του πειράματος. Αν και υπάρχουν πολλές μέθοδοι κανονικοποίησης μας ζητήθηκε να αναλύσουμε τις 3 πιο βασικές (RMA, GCRMA και MAS5.0). Η επιλογή της κατάλληλης μεθόδου για την εξαγωγή των πιο αξιόπιστων αποτελεσμάτων εξαρτάται από τις τοπικές συνθήκες

που επικρατούν σε κάθε εργαστήριο. Πρέπει να ελεγχθούν τα δείγματα οπτικά πριν και μετά την εφαρμογή των κανονικοποιήσεων για να καταλάβουμε αν η κανονικοποίηση λειτούργησε σωστά.

Ο ορθότερος τρόπος απεικόνισης των δεδομένων είναι η κλίμακα $\log_2$. Πρέπει να χρησιμοποιηθούν διαφορετικών ειδών έλεγχοι(controls) π.χ. αρνητικός έλεγχος (negative control), τοποθετημένα δείγματα ελέγχου(spiked-in controls) και τα housekeeping γονίδια, οι οποίοι θα μας βοηθήσουν να ανιχνεύσουμε πιθανά προβλήματα με τις διαδικασίες της υβριδοποίησης και της κανονικοποίησης. Όταν το επιτρέπει ο σχεδιασμός του πειράματος, η κανονικοποίηση πρέπει να γίνεται στο σύνολο των γονιδίων του δείγματος ή πρέπει να χρησιμοποιηθεί ένας αρκετά μεγάλος αριθμός μη εκφρασμένων γονιδίων ελέγχου.

7 Βιβλιογραφικές Αναφορές

[1] Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003,19(2):185–193. 10.1093/bioinformatics/19.2.185

[2] Cope, L., Irizarry, R., Jaffee, H., Wu, Z. and Speed, T. (2004). A benchmark for Affymetrix Genechip expression measures. Bioinformatics 20 323–331.

[3] Helen Causton, John Quackenbush, Alvis Brazma - 2009 - Science. A Beginner's Guide Helen Causton, John Quackenbush, Alvis Brazma. © 2003 by Blackwell Science Ltd a Blackwell Publishing company

[4] Ernst LA, Gupta RK, Mujumdar RB, Waggoner AS. Cyanine dye labelling reagents for sulfhydryl groups. Cytometry. 1989;10(1):3-10.

[5] Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991;251(4995):767-773.

[6] Hubbell E, Liu WM, Mei R: Robust estimators for expression analysis. *Bioinformatics* 2002, 18: 1585–1592. 10.1093/bioinformatics/18.12.1585

[7] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, 4: 249–264. 10.1093/biostatistics/4.2.249

[8] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 1996;14(13):1675-80.

[9] MacBeath G, Schreiber SL. Printing Proteins as Microarrays for High-Throughput Function Determination. Science. 2000;289(5485):1760-1763.

[10] Parmigiani, G., Garrett, E. S., Irizarry, R. A., and S. L. Zeger, S. L. (eds.) (2003). *The Analysis of Gene Expression Data: Methods and Software.* Springer, New York. 3/2003

[11] Quackenbush S.L., Casey R.N., Murcek R.J., Paul T.A., Work T.M., Limpus C.J., Chaves A., Dutoit L., Perez J.V., Aguirre A.A., Spraker T.R., Horrocks J.A., Horrocks J.A., Vermeer L.A., Balazs G.H. & Casey J.W. 2001. Quantitative analysis of herpesvirus sequences from normal tissue and ibropapillomas of marine turtles with real-time PCR. Virology 287:105-111.

[12] Schena M, Shalon D, Davis RW, Brown PO. Quantitative motoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270:467-470.

[13] Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham GE, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A. High-throughput SNP genotyping on universal bead arrays. Mutation Research. 2005;573:70-82.

[14] Shinawi M, Cheung SW. The array CGH and its clinical applications. Drug Discov Today. 2008;17-18:760-70.

[15] Tseng G.C., Oh M.K., Rohlin L., Liao J.C., Wong W.H. Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations, and assessment of gene effects. Nucleic Acids Res. 2001;29:2549–2557

[16] Yang YH(1), Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.