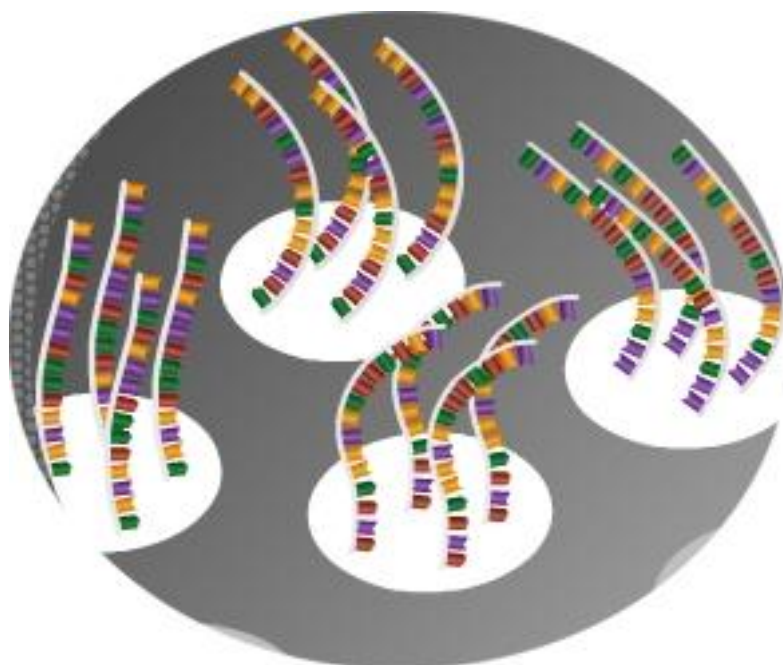




ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ: ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΘΟΔΟΙ ΑΝΑΓΝΩΡΙΣΗΣ ΚΥΨΕΛΙΔΩΝ ΣΕ ΕΙΚΟΝΕΣ ΜΙΚΡΟ-ΣΥΣΤΟΙΧΙΩΝ



ΟΡΦΑΝΟΥΔΑΚΗ ΓΕΩΡΓΙΑ
Α.Μ. 2002030018
ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ : ΖΕΡΒΑΚΗΣ ΜΙΧΑΗΛ



ΠΕΡΙΛΗΨΗ.....	7
1 ΕΙΣΑΓΩΓΗ.....	9
1.1 ΕΙΣΑΓΩΓΗ ΣΤΟ ΑΝΤΙΚΕΙΜΕΝΟ.....	9
1.1.1 ΒΙΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ.	9
1.1.2 ΔΗΜΙΟΥΡΓΙΑ ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ.....	10
1.1.3 ΔΗΜΙΟΥΡΓΙΑ ΕΙΚΟΝΑΣ.....	12
1.1.4 ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΑΣ.	13
1.2 ΠΡΟΒΛΗΜΑΤΑ ΠΟΥ ΑΝΑΚΥΠΤΟΥΝ ΚΑΤΑ ΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΤΗΣ ΕΙΚΟΝΑΣ	17
1.3 ΠΡΟΗΓΟΥΜΕΝΕΣ ΜΕΛΕΤΕΣ	19
1.4 ΣΤΟΧΟΙ.....	20
1.5 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ.....	21
1.5.1 ΤΟΠΟΘΕΤΗΣΗ ΠΛΕΓΜΑΤΟΣ	21
1.5.2 ΑΠΟ ΤΟ RGB ΣΤΟ HSI.....	23
1.6 ΔΟΜΗ ΕΡΓΑΣΙΑΣ.....	27
2 ΚΑΤΑΤΜΗΣΗ ΠΕΡΙΟΧΗΣ.....	29
2.1 ΚΑΤΑΤΜΗΣΗ ΜΕ ΒΑΣΗ ΠΕΡΙΓΡΑΜΜΑ.....	29
2.1.1 ΚΑΤΑΤΜΗΣΗ ΣΤΑΘΕΡΟΥ ΚΥΚΛΟΥ (FIXED CIRCLE SEGMENTATION).....	29
2.1.2 ΚΑΤΑΤΜΗΣΗ ΜΕΤΑΒΑΛΛΟΜΕΝΟΥ ΚΥΚΛΟΥ (ADAPTIVE CIRCLE SEGMENTATION)	29
2.2 ΚΑΤΑΤΜΗΣΗ ΒΑΣΗ ΕΝΤΑΣΗΣ.....	30
2.2.1 WATERSHED	30
2.2.2 SEEDED REGION GROWING (SRG)	32
2.2.3 MANN WHITNEY TEST	33
2.2.4 STATISTICAL REGION MERGING.....	34
2.2.5 ΑΛΓΟΡΙΘΜΟΣ ΤΥΧΑΙΟΥ ΠΕΔΙΟΥ MARKOV-(THE MARKOV RANDOM FIELD)	36
3 ΚΑΤΑΤΜΗΣΗ ΙΣΤΟΓΡΑΜΜΑΤΟΣ.....	37
3.1 ΚΑΤΑΤΜΗΣΗ ΒΑΣΗ ΕΝΤΑΣΗΣ.....	37
3.1.1 ΚΑΤΑΤΜΗΣΗ ΜΕ ΜΟΝΤΕΛΟΠΟΙΗΣΗ	37
3.1.1.1 Αλγόριθμος Otsu	38
3.1.1.2 Επαναληπτικός Otsu.....	39
3.1.1.3 Αλγόριθμος Expectation Maximization (EM).....	40
3.1.1.4 Επαναληπτική κατάτμηση με βάση τον αλγόριθμο Expectation Maximization	44
3.1.2 ΚΑΤΑΤΜΗΣΗ ΜΕ ΟΜΑΔΟΠΟΙΗΣΗ.....	44
3.1.2.1 K-means clustering	44
3.1.2.2 Υβριδικός k-means (Hybrid k-mean)	45
3.1.2.3 Παραλλαγή υβριδικού k-means.....	45
3.2 ΚΑΤΑΤΜΗΣΗ ΕΝΤΑΣΗΣ ΚΑΙ ΑΠΟΧΡΩΣΗΣ	46
3.2.1 ΔΥΣΔΙΑΣΤΑΤΗ ΚΑΤΑΤΜΗΣΗ	48
3.2.1.1 K-means δύο διαστάσεων.....	48
3.2.1.2 Επαναληπτικός k-means δύο διαστάσεων.....	48
3.2.2 ΔΙΑΔΟΧΙΚΗ ΚΑΤΑΤΜΗΣΗ.....	48

3.2.2.1	Ασυσχετίστα χαρακτηριστικά	48
3.2.2.2	Δέντρο απόφασης	49

4 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΦΑΡΜΟΓΗΣ..... 51

4.1	ΑΠΛΟΙ OTSU, K- MEANS – ΠΡΟΒΛΗΜΑΤΑ	51
4.2	ΕΜ – ΠΡΟΒΛΗΜΑΤΙΚΕΣ ΠΕΡΙΠΤΩΣΕΙΣ	54
4.3	ΕΠΑΝΑΛΗΠΤΙΚΟΣ K-MEANS – ΕΦΑΡΜΟΓΗ.....	56
4.4	ΕΠΑΝΑΛΗΠΤΙΚΟΣ OTSU – ΕΦΑΡΜΟΓΗ	59
4.5	ΕΠΑΝΑΛΗΠΤΙΚΟΣ ΔΥΣΔΙΑΣΤΑΤΟΣ K-MEANS.....	60
4.6	ΣΧΟΛΙΑ-ΣΥΜΠΕΡΑΣΜΑΤΑ	62
4.6.1	ΣΥΓΚΡΙΣΗ ΑΠΛΟΥ - ΕΠΑΝΑΛΗΠΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ	62
4.6.2	ΣΥΓΚΡΙΣΗ ΜΟΝΟΔΙΑΣΤΑΤΗΣ ΚΑΙ ΔΥΣΔΙΑΣΤΑΤΗΣ ΠΡΟΣΕΓΓΙΣΗΣ	62
4.7	ΔΙΑΔΙΚΑΣΙΑ ΣΥΜΠΛΗΡΩΣΗΣ SPOT (CENTER FILLING).....	63
4.7.1	ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΘΕΣΗΣ ΕΝΟΣ SPOT	64
4.7.2	ΑΛΓΟΡΙΘΜΟΣ ΣΥΜΠΛΗΡΩΣΗΣ ΤΟΥ SPOT	66
4.7.3	ΑΦΑΙΡΕΣΗ ΠΕΡΙΦΕΡΕΙΑΚΩΝ PIXEL.....	67
4.7.4	ΤΕΛΙΚΟΣ ΑΛΓΟΡΙΘΜΟΣ	68
4.7.5	ΕΦΑΡΜΟΓΗ ΑΛΓΟΡΙΘΜΟΥ	69

5 ΠΕΙΡΑΜΑ -ΑΠΟΤΕΛΕΣΜΑΤΑ..... 71

5.1	DATASET	71
5.2	GOLDEN TRUTH	71
5.3	ΥΠΟΛΟΓΙΣΜΟΣ ΣΦΑΛΜΑΤΟΣ – MEAN SQUARE ERROR.....	74
5.4	ΠΕΡΙΓΡΑΦΗ ΠΕΙΡΑΜΑΤΩΝ	74
5.5	ΕΙΚΟΝΕΣ ΜΕ SPOT ΕΛΕΓΧΟΥ	75
5.5.1	ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ OTSU	76
5.5.2	ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ K-MEANS.	78
5.5.3	ΑΠΟΔΟΣΗ ΕΠΑΝΑΛΗΠΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ ΕΜ.....	80
5.5.4	ΑΠΟΔΟΣΗ ΔΥΣΔΙΑΣΤΑΤΟΥ 2-MEANS ΑΛΓΟΡΙΘΜΟΥ	82
5.6	ΕΙΚΟΝΕΣ ΟΜΟΤΥΠΙΚΩΝ ΥΒΡΙΔΙΣΜΩΝ	82
5.6.1	ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ OTSU	83
5.6.2	ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ K-MEANS.	85
5.6.3	ΑΠΟΔΟΣΗ ΕΠΑΝΑΛΗΠΤΙΚΟΥ ΑΛΓΟΡΙΘΜΟΥ ΕΜ.....	87
5.6.4	ΑΠΟΔΟΣΗ ΔΥΣΔΙΑΣΤΑΤΟΥ 2-MEANS ΑΛΓΟΡΙΘΜΟΥ	89
5.7	ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ.....	90
5.8	ΑΛΓΟΡΙΘΜΟΣ ΣΥΜΠΛΗΡΩΣΗΣ SPOT	92

6 ΣΥΜΠΕΡΑΣΜΑΤΑ..... 95

6.1	ΜΟΝΟΔΙΑΣΤΑΤΗ ΚΑΙ ΔΥΣΔΙΑΣΤΑΤΗ ΚΑΤΑΤΜΗΣΗ.....	95
6.2	ΔΥΣΔΙΑΣΤΑΤΗ ΚΑΙ ΔΙΑΔΟΧΙΚΗ ΚΑΤΑΤΜΗΣΗ – ΑΣΥΣΧΕΤΙΣΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	95
6.3	ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ.....	95
6.4	ΑΠΟ RGB ΣΤΟ HSI ΣΥΣΤΗΜΑ ΑΝΑΠΑΡΑΣΤΑΣΗΣ	97
6.5	ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ	100



ΠΕΡΙΛΗΨΗ

Η μελέτη των γονιδίων, η κατανόηση της αλληλεπίδρασης μεταξύ τους, των μηχανισμών μετάφρασης και ελέγχου παραγωγής πρωτεϊνών, του τρόπου διαφοροποίησης της λειτουργίας των κυττάρων αποτελούν μια συνεχή προσπάθεια η οποία έχει πολλούς στόχους όπως τον έγκαιρο εντοπισμό ασθενειών (καρκινικά κύτταρα, διαβήτης, γενετικές ανωμαλίες) και την ανακάλυψη κατάλληλων γονιδιακών θεραπειών.

Οι Μικροσυστοιχίες αποτελούν μια από τις πιο καινούργιες βιοτεχνολογικές μεθόδους για την ανάλυση της έκφρασης γονιδίων μέσω της διαδικασίας του υβριδισμού. Εφαρμόζονται ευρέως σε βιομηχανικές, κλινικές, τοξικολογικές και φαρμακευτικές έρευνες καθώς το βασικό πλεονέκτημα τους είναι ότι μετράνε την έκφραση χιλιάδων γονιδίων ταυτόχρονα. Οι cDNA συστοιχίες αποτελούμενες από χιλιάδες ατομικές γονιδιακές ακολουθίες τυπωμένες σε ένα συμπυκνωμένο πίνακα παρέχουν ένα πρακτικό και οικονομικό εργαλείο για την μελέτη γονιδιακής έκφρασης σε πολύ μεγάλη κλίμακα.

Η επεξεργασία εικόνων μικροσυστοιχιών είναι η βασική διαδικασία με την οποία ασχολείται η παρούσα εργασία. Στόχος της επεξεργασίας αυτής είναι η εξαγωγή ενός ποσοστού έκφρασης για κάθε ένα από τα γονίδια που μελετάται, δηλαδή η αντιστοίχιση κάθε spot της εικόνας σε μια τιμή έκφρασης, σε ένα σχετικό ποσοστό υβριδισμού. Υλοποιούμε και τα τρία βήματα επεξεργασίας εικόνων μικροσυστοιχιών (τοποθέτηση πλέγματος, κατάτμηση, εξαγωγή ποσοστού έκφρασης) με μεγαλύτερη έμφαση στο στάδιο της κατάτμησης.

Στόχος μας είναι εκτός από την πληροφορία της τιμής έντασης των pixel να χρησιμοποιήσουμε επιπρόσθετα την πληροφορία της απόχρωσης κάθε pixel. Για το λόγο αυτό εργαστήκαμε στο HSI σύστημα αναπαράστασης εικόνων αντί για το RGB το οποίο χρησιμοποιούν οι περισσότερες μελέτες. Επιλέξαμε τέσσερις απλούς ήδη εφαρμοσμένους αλγόριθμους ταξινόμησης και υλοποιήσαμε μικρές παραλλαγές αυτών. Οι αλγόριθμοι που επιλέξαμε είναι: Otsu, k-means, Expectation Maximazation. Οι δύο πρώτοι αλγόριθμοι, Otsu και k-means επιλεχθήκαν για λόγους που έχουν να κάνουν με την απλότητα και την εφαρμοσιμότητα τους στην ειδική περίπτωση των εικόνων μικροσυστοιχιών μας ενώ ο αλγόριθμος EM επιλέχθηκε ως πιο πολύπλοκος και προσαρμόσιμος στις κατανομές(επιλογή αριθμού παραμέτρων και εξίσωσης κατανομής). Μελετήσαμε την αξία του χαρακτηριστικού της απόχρωσης, μετρήσαμε την απόδοση των αλγορίθμων κατάτμησης, εντοπίσαμε βελτιώσεις, υπολογίσαμε το ποσοστό έκφρασης και να το συγκρίναμε με ήδη υπάρχοντα αποτελέσματα.

Ως επέκταση της εφαρμογής των παραπάνω αλγορίθμων υλοποιήσαμε ένα αλγόριθμο συμπλήρωσης του κέντρου κάθε spot ο οποίος κάνει χρήση χωρικής πληροφορίας. Παρόμοια προσέγγιση συναντάμε και στην εργασία των **Th.Margaritis, K. Marias, Kafetzopoulos**^[34] όπου ορίζεται μια περιοχή του spot TSA η οποία χρησιμοποιείται συμπληρωματικά για την επιλογή της κλάσης του spot. Μελετήσαμε και συγκρίναμε την απόδοση των πρώτων αλγορίθμων που ασχολούνται μόνο με πληροφορίες που προκύπτουν από τις κατανομές των χαρακτηριστικών με την επιπρόσθετη εφαρμογή του αλγορίθμου συμπλήρωσης του spot.

Εφαρμόσαμε τους παραπάνω αλγορίθμους στην βάση εικόνων μικροσυστοιχιών και δεδομένων του Εθνικού Ινστιτούτου Καρκίνου (National Cancer Institute ^[6]) και συγκρίναμε τα αποτελέσματα κατάτμησης που προκύπτουν από τον αλγόριθμο επεξεργασίας εικόνας του λογισμικού GEMTools ^[24]. Οι συγκεκριμένες εικόνες δημιουργήθηκαν από την **Incyte Genomics** για λόγους σύγκρισης των παραμέτρων σάρωσης ανάμεσα σε διάφορα πειράματα.



1 ΕΙΣΑΓΩΓΗ

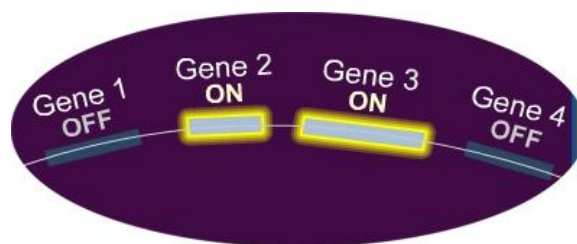
1.1 Εισαγωγή στο αντικείμενο

Η μελέτη των γονιδίων και η προσπάθεια για πλήρη αποκωδικοποίηση του ανθρώπινου γονιδιώματος αλλά και του γονιδιώματος άλλων οργανισμών είναι μια διαδικασία στην οποία έχουν επικεντρωθεί πολλοί επιστήμονες τα τελευταία χρόνια. Η κατανόηση της λειτουργίας των γονιδίων, της αλληλεπίδρασης μεταξύ τους, της κατανόησης των μηχανισμών μετάφρασης και ελέγχου παραγωγής πρωτεϊνών, του τρόπου διαφοροποίησης της λειτουργίας των κυττάρων αποτελούν μια συνεχή προσπάθεια η οποία έχει πολλούς στόχους όπως τον έγκαιρο εντοπισμό ασθενειών (καρκινικά κύτταρα, διαβήτης, γενετικές ανωμαλίες), την κατανόηση της λειτουργίας της κληρονομικότητας και την ανακάλυψη κατάλληλων γονιδιακών θεραπειών.

Οι Μικροσυστοιχίες αποτελούν μια από τις πιο καινούργιες βιοτεχνολογικές μεθόδους για την ανάλυση της έκφρασης γονιδίων μέσω της διαδικασίας του υβριδισμού. Εφαρμόζονται ευρέως σε βιομηχανικές, κλινικές, τοξικολογικές και φαρμακευτικές έρευνες καθώς το βασικό πλεονέκτημα τους είναι ότι μετράνε την έκφραση χιλιάδων γονιδίων ταυτόχρονα. Οι DNA Μικροσυστοιχίες αποτελούμενες από χιλιάδες ατομικές γονιδιακές ακολουθίες τυπωμένες σε ένα συμπυκνωμένο πινάκα πάνω σε μια γυάλινη επιφάνεια παρέχουν ένα πρακτικό και οικονομικό εργαλείο για την μελέτη γονιδιακής έκφρασης σε πολύ μεγάλη κλίμακα.

Εκτός από την από την τιμή έκφρασης διάφορων γονιδίων οι μικροσυστοιχίες έχουν διάφορες εφαρμογές όπως η μελέτη της γονιδιακής έκφρασης βακτηρίων κάτω από διαφορετικές περιβαλλοντολογικές συνθήκες^{[1],[2]}, σύγκριση της έκφρασης διάφορων γονιδίων σε καρκινικά και υγιή κύτταρα^[3]. Είναι πάρα πολύ σημαντική η συμβολή τους στην κατανόηση της λειτουργίας και της αλληλεπίδρασης των γονιδίων. Με την σύγκριση μέσω μικροσυστοιχιών των γονιδίων που εκφράζονται σε υγιή και σε καρκινικά κύτταρα μπορούν να εντοπιστούν προβληματικά γονίδια καθώς επίσης και στόχοι θεραπείας. Βέβαια με την μέθοδο των Μικροσυστοιχιών δεν εντοπίζονται όλες οι προβληματικές περιπτώσεις επειδή η παραγωγή πρωτεϊνών δεν εξαρτάται εξ ολοκλήρου από την ύπαρξη των αντίστοιχων mRNA αλλά και από τα αντιδραστήρια που ελέγχουν την μετάφραση του mRNA.

Σημαντικό ρόλο στα πειράματα μικροσυστοιχιών παίζει η **επεξεργασία εικόνας** καθώς σε κάθε πείραμα υβριδισμού οι μικροσυστοιχίες απεικονίζονται ώστε να υπολογιστούν οι τιμές εντάσεων για το κόκκινο και το πράσινο φθορίζον χρώμα και στην συνέχεια να εξαχθεί μια τιμή έκφρασης για κάθε spot, όπου κάθε spot του πίνακα μικροσυστοιχιών είναι αποτέλεσμα υβριδισμού γνωστών γονιδίων. Έτσι η ανάπτυξη της τεχνολογίας των μικροσυστοιχιών έδωσε την δυνατότητα για δημιουργία τεραστίων βάσεων δεδομένων γονιδιακής έκφρασης με δυνατότητα μελλοντικής τους επεξεργασίας μέσω διαδικασιών μοντελοποίησης και ταξινόμησης.



Εικόνα 1.1.1 Σε κάθε κύτταρο του ανθρώπινου οργανισμού κάποια γονίδια είναι ενεργοποιημένα και κάποια όχι. Η λειτουργία των κυττάρων διαφοροποιείται ανάλογα με τον σύνολο των γονιδίων που εκφράζονται.^[3]

1.1.1 Βιολογικό Υπόβαθρο.

Το ανθρώπινο γονιδίωμα αποτελείται από πάνω από 20.000 γονίδια. Κάθε κύτταρο του ανθρώπινου οργανισμού διαφοροποιείται από τα υπόλοιπα στην μορφή και στην λειτουργία παρόλο που όλα περιέχουν το ίδιο γεννητικό υλικό. Αυτό γίνεται επειδή σε κάθε κύτταρο κάποια γονίδια είναι ενεργοποιημένα και κάποια άλλα όχι. Κάθε γονίδιο περιέχει τις «οδηγίες» για την παραγωγή κάποιας πρωτεΐνης η οποία με την σειρά της εξυπηρετεί κάποιά λειτουργία. Ένα γονίδιο είναι ενεργό όταν

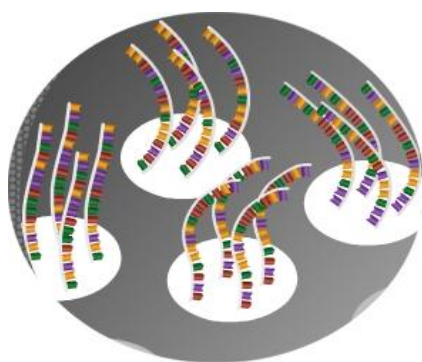
επιτρέπεται η μετάφραση του στην αντίστοιχη πρωτεΐνη, αυτό που βιολόγοι ονομάζουν έκφραση του γονιδίου. Ένας τρόπος για να διαπιστώσουμε την έκφραση ενός γονιδίου είναι η ύπαρξη του αντίστοιχου messenger RNA ή mRNA πάνω στο οποίο γίνεται η διαδικασία της μετάφρασης. Για παράδειγμα σε ένα επιδερμικό κύτταρο αναμένουμε να συναντήσουμε mRNA του γονιδίου από το οποίο παράγεται μελανίνη ενώ δεν περιμένουμε να βρούμε mRNA για την παραγωγή ινσουλίνης.

Έχοντας στην διάθεση μας ένα «κατάλογο» με τα mRNA που συναντάμε σε κάθε κύτταρο μπορούμε να συμπεράνουμε τι είδους βιολογικές διαδικασίες λαμβάνουν χώρα σε αυτό και συγκρίνοντας αυτές τις πληροφορίες με τις αντίστοιχες για άλλα είδη κυττάρων μπορούμε να κατανοήσουμε τις διαφοροποιήσεις τους.

Μια τέτοιου είδους όμως διαδικασία συνεπάγεται έλεγχο έκφρασης πάνω σε όλο το ανθρώπινο γονιδίωμα, δηλαδή σε περίπου 20.000 γονίδια. Σε αυτό το σημείο υπεισέρχονται οι μικροσυστοιχίες με την βοήθεια των οποίων μπορούμε να κάνουμε έλεγχο έκφρασης σε πάνω από 10.000 γονίδια ταυτόχρονα.

1.1.2 Δημιουργία Μικροσυστοιχιών.

Οι cDNA μικροσυστοιχίες είναι ένα από τα πιο δυνατά εργαλεία για την εξαγωγή γενετικής πληροφορίας. Για τα cDNA πειράματα απαιτείται η απομόνωση DNA από τα πειραματικά δείγματα και από τα δείγματα ελέγχου, δηλαδή εξαγωγή των άγνωστων γονιδίων και των γνωστών γονιδίων που θα μας βοηθήσουν στην σύγκριση.



Εικόνα 1.1.2 Μια μικροσυστοιχία αποτελείται από χιλιάδες spot. Κάθε ένα από τα spot περιέχει πολλά αντίτυπα από το ίδιο τμήμα της αλυσίδας του DNA που αντιστοιχεί σε κάποιο γνωστό γονίδιο. ^[3]

Οι γυάλινες πλάκες πάνω στις οποίες εναποθέτουμε τα δείγματα μας έχουν διάφορα ονόματα όπως GeneChip, GenomeChip, expression chip, gene array και παράγονται βιομηχανικά από διάφορες εταιρείες. Ήδη κυκλοφορούν στην αγορά έτοιμες μικροσυστοιχίες με τα γονίδια αναφοράς (ή γονίδια βάσης) τοποθετημένα πάνω στην γυάλινη επιφάνεια σε δεδομένη διάταξη.

Ας υποθέσουμε για παράδειγμα ότι θέλουμε να συγκρίνουμε τα γονίδια που εκφράζονται σε κάποια καρκινικά κύτταρα με αυτά των αντίστοιχων υγιών και με αυτόν τον τρόπο να κατανοήσουμε το λόγο που διαφοροποιεί την λειτουργία τους. Συγκεκριμένα στην περίπτωση του καρκίνου έχει διαπιστωθεί ότι οφείλεται σε μετάλλαξη γονιδίων τα οποία ελέγχουν την διαδικασία αύξησης διαίρεσης και θανάτωσης των κυττάρων. Τα μεταλλαγμένα γονίδια είναι γνωστά ως ογκογονίδια. Οι

διαφορετικοί τύποι καρκίνου οφείλονται σε μετάλλαξη άλλων γονιδίων. Συνεπώς μέσω της σύγκρισης των γονιδίων που εκφράζονται στα υγιή κύτταρα και δεν εκφράζονται στα καρκινικά κύτταρα μπορούμε να εντοπίσουμε τα προβληματικά γονίδια.

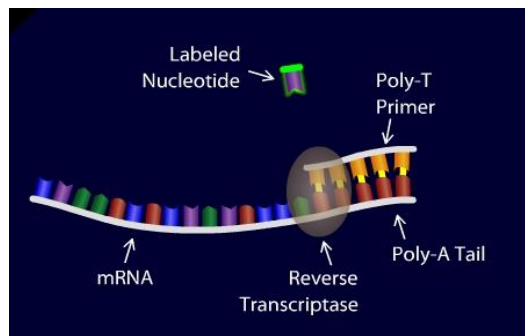
Η διαδικασία παραγωγής των μικροσυστοιχιών και απεικόνισης τους αποτελείται από τα παρακάτω βήματα:

1. Συλλογή ιστού.
2. Απομόνωση RNA.
3. Απομόνωση mRNA.
4. Δημιουργία χρωματισμένου αντίγραφου DNA.
5. Εναπόθεση του δείγματος DNA πάνω στην μικροσυστοιχία.
6. Σάρωση της μικροσυστοιχίας.

Ας πάρουμε τα βήματα ένα ένα αναλυτικά. Αρχικά συλλέγουμε κύτταρα από φυσιολογικό ιστό και κύτταρα από το καρκίνωμα. Έπειτα με στόχο να απομονώσουμε τα τμήματα του mRNA από τα δύο δείγματα μας τα διαλύουμε σε ένα μίγμα από διάφορους οργανικούς διαλύτες. Με την ανάδευση το RNA

διαλύεται και με την χρήση φυγοκεντρωτή διαχωρίζεται από τα υπόλοιπα συστατικά του κυττάρου. Λόγω μεγάλου μεγέθους τα νήματα του RNA μένουν στην επιφάνεια του διαλύματος και έτσι το συλλέγουμε. Το RNA που συλλέγουμε αποτελείται από τα τρία γνωστά RNA, mRNA (messenger RNA), tRNA (transfer RNA), rRNA (ribosomal RNA). Μόνο το mRNA μεταφέρει την πληροφορία για την παραγωγή πρωτεΐνης και ξεχωρίζει μέσω μια σειράς από μερικές εκατοντάδες αδενίνες (Poly-A tale) οι οποίες σημαίνουν το τέλος του γονιδίου και της αντιγραφής και προωθούν το mRNA έξω από τον πυρήνα. Ξεπλένοντας το RNA με ένα διάλυμα που περιέχει σταγόνες από Poly-T τμήματα, οι Poly-A ουρές του mRNA προσδένονται σε αυτά και διαχωρίζονται από τα υπόλοιπα RNA τμήματα. Με την χρήση ενός ακόμα διαλύματος το οποίο ευνοεί την αντίθετη διαδικασία του υβριδισμού παίρνουμε το καθαρό m-RNA.

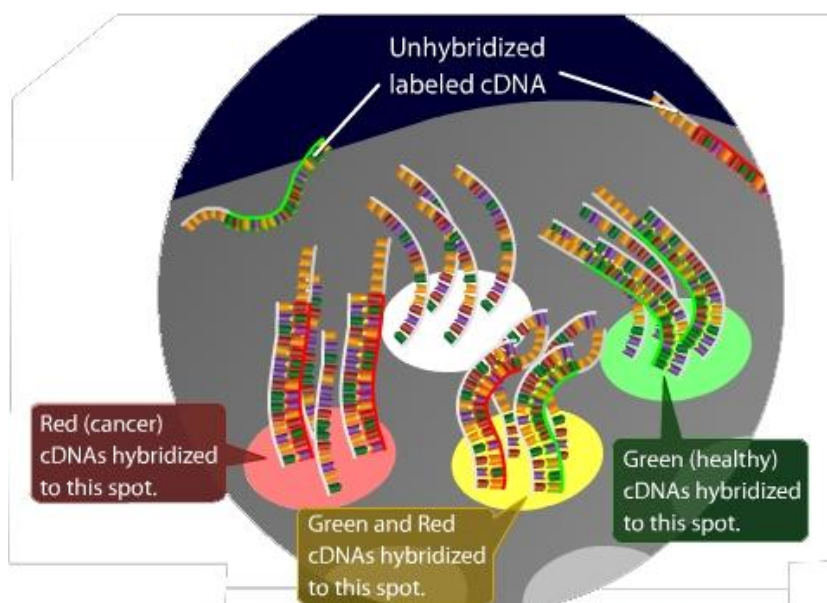
Στην συνέχεια χρωματίζουμε τα δείγματα μας χρησιμοποιώντας δύο χρωματισμένα μίγματα, συνήθως κόκκινο για τα καρκινικά και πράσινο για τα υγιή mRNA νήματα. Συγκεκριμένα τα μίγματα μας περιέχουν Poly-T primers τα οποία προσδένονται στο Poly-A άκρο του mRNA νήματος και την αντίστροφη μεταγραφάση η οποία θα συνθέσει την συμπληρωματική αλυσίδα DNA και τέλος νουκλεοτίδια που έχουν προσδεμένο πάνω τους ένα μόριο φθορίζουσας ουσίας (Εικ. 1.1.3).



Εικόνα 1.1.3 Διαδικασία παραγωγής χρωματισμένου συμπληρωματικού DNA από τα mRNA δείγματα που έχουμε συλλέξει. ^[3]

Έπειτα όλη η υπόλοιπη διαδικασία βασίζεται στην ιδιότητα του υβριδισμού που έχουν οι συμπληρωματικές αλυσίδες DNA. Εναποθέτουμε τα χρωματισμένα δείγματα DNA μας πάνω στην μικροσυστοιχεία μας. Σε κάθε spot τοποθετούμε υλικό και από τα δύο χρώματα. Εάν έχουμε υβριδισμό των γονιδίων ελέγχου με τα καρκινικά γονίδια τότε το spot παίρνει κόκκινο χρώμα. Αντίστοιχα το spot παίρνει πράσινο χρώμα εάν τα έχουμε υβριδισμό με τα υγιή γονίδια. Εάν το γονίδιο εκφράζεται και στα καρκινικά αλλά και στα υγιή κύτταρα τότε το spot παίρνει κίτρινο χρώμα (Εικ. 1.1.4).

Σαρώνουμε την εικόνα μας στα 540 nm (green) και στα 630 nm (red) και το αποτέλεσμα της σάρωσης δίνει δύο 16-bit μονοχρωματικές εικόνες. Η κάθε εικόνα περιέχει εντάσεις ενός μόνο χρώματος

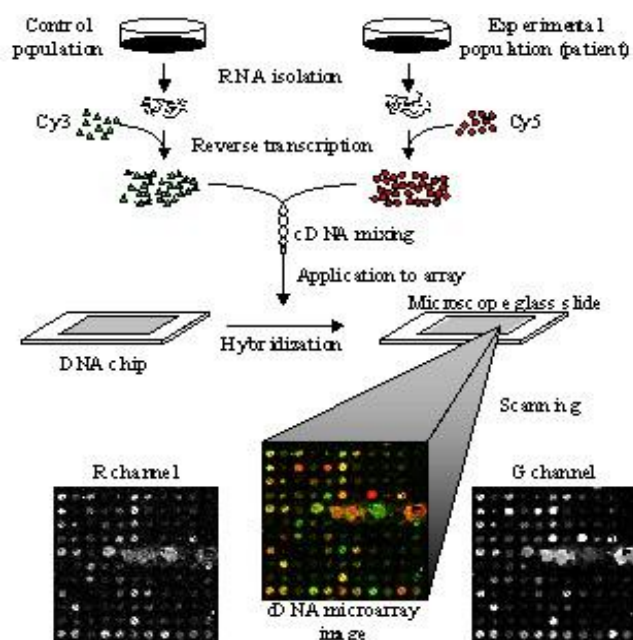


Εικόνα 1.1.4 Αποτέλεσμα υβριδισμού. Στα spot που παίρνουν κόκκινο χρώμα έχει πραγματοποιηθεί υβριδισμός ανάμεσα στα καρκινικά δείγματα και στα δείγματα ελέγχου. Με πράσινο έχουμε υβριδισμό των υγιών δειγμάτων ενώ με κίτρινο συμβαίνουν και τα δύο ταυτόχρονα ^[3]

τις οποίες στην συνέχεια συνδυάζουμε σε μια δικάναλη εικόνα. Ο λόγος της έντασης των δύο χρωμάτων για κάθε spot υποδεικνύει την σχετική αφθονία των δύο δειγμάτων, δηλαδή το σχετικό ποσοστό υβριδισμού.

Πιο συγκεκριμένα ανάλογα με το ποσοστό υβριδισμού των δειγμάτων μας με τα γονίδια βάσης (ή γονίδια αναφοράς) έχουμε και μια αντίστοιχη τιμή έντασης. Για παράδειγμα στην περίπτωση των γονιδίων που έχουν προκύψει από καρκινικά κύτταρα ανάλογα με την ποσότητα των τμημάτων του DNA που «ταίριαξαν» με γονίδια αναφοράς και ανάλογα με το ποσοστό ομοιότητας των αλυσίδων DNA δηλαδή σε ποια σημεία ενώθηκαν και σε ποια όχι έχουμε και μια ανάλογη τιμή έντασης του κόκκινου. Ομοίως με τα δείγματα που παίρνουμε από υγιή κύτταρα προκύπτει μια αντίστοιχη τιμή έντασης του πράσινου.

Κατ' αυτόν τον τρόπο κάθε μια από τις δύο εικόνες δίνει μια ποσοτικοποίηση του ποσοστού ομοιότητας των γονιδίων αναφοράς με τα δύο δείγματα μας. Η εικόνα στο κόκκινο δίνει μια ποσοτικοποίηση του υβριδισμού των καρκινικών γονιδίων με τα γονίδια αναφοράς. Αντίστοιχα η εικόνα στο πράσινο δίνει τον βαθμό υβριδισμού των υγιών γονιδίων με τα γονίδια αναφοράς. Καθώς όμως ο στόχος μας είναι να συγκρίνουμε την έκφραση των καρκινικών γονιδίων σε σχέση με την έκφραση των υγιών για κάθε spot υπολογίζουμε τον λόγο της μέσης τιμής έντασης του κόκκινου προς την μέση τιμή έντασης του πράσινου.

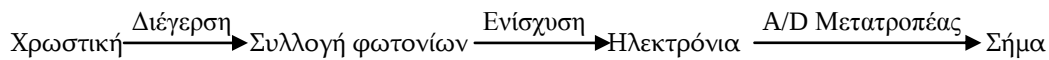


Εικόνα 1.1.5 Διάγραμμα διαδικασίας παραγωγής και απεικόνισης μικροσυστοιχίας.^[4]

1.1.3 Δημιουργία Εικόνας.

Η σάρωση της μικροσυστοιχίας μας μπορεί να γίνει με διάφορες συσκευές όπως ένα scanner, μια κάμερα CCD (charge coupled device), ένα μικροσκοπικό σαρωτή. Υπάρχουν βέβαια ειδικευμένες συσκευές γι' αυτήν την διαδικασία οι οποίες σαρώνουν την επιφάνεια της μικροσυστοιχίας και παράγουν για κάθε χρώμα μια εικόνα.

Τα βήματα που ακολουθεί ένας τυπικός σαρωτής μικροσυστοιχιών είναι Διέγερση των φθορίζοντων μορίων με εκπομπή φωτονίων, Συλλογή εκπεμπόμενου φωτός, Μετατροπή Φωτονίων σε Ηλεκτρόνια, Ενίσχυση, Μετατροπή αναλογικού σήματος σε ψηφιακό (A/D Converter).^[5]



Αρχικά ο σαρωτής χωρίζει σε ισομεγέθη περιοχές (pixels) την επιφάνεια της μικροσυστοιχίας και εκπέμπει σε πρώτη φάση φως περίπου στα 510 nm με 550nm όπου είναι το κόκκινο χρώμα εστιασμένο σε μια από τις υπο-περιοχές (pixels). Με την εκπομπή φωτός έχουμε διέγερση των φθορίζοντων μορίων και εκπομπή φωτονίων πίσω στον δέκτη του σαρωτή. Μέρος των συνολικών φωτονίων που συλλέγεται από τον φακό του σαρωτή προέρχεται από φωτόνια διέγερσης και άλλο μέρος από φωτόνια εκπομπής. Τα φωτόνια διέγερσης έχουν μικρότερο μήκος κύματος έτσι με ένα διαχωριστή πλάτους και ένα band-pass φίλτρο κρατάμε μονό τα φωτόνια εκπομπής. Ο δέκτης του σαρωτή μετατρέπει στην συνέχεια το σήμα των φωτονίων σε ηλεκτρικό σήμα και το ενισχύει. Τέλος καταλήγει σε ένα ψηφιακό σήμα μέσω ενός A/D μετατροπέα. Ο σαρωτής παράγει ένα σήμα για κάθε pixel το οποίο έχει άμεση συσχέτιση με την ένταση φθορισμού της αντίστοιχης υπο-περιοχής.

Η αντίστοιχη διαδικασία γίνεται και για το πράσινο χρώμα (630nm με 660nm). Σε ένα ακολουθιακό σαρωτή έχουμε ξεχωριστή σάρωση για κάθε χρώμα ενώ υπάρχουν σαρωτές με διπλό δέκτη που κάνουν τις δύο αυτές διαδικασίες ταυτόχρονα.

Στην διαδικασία απεικόνισης της μικροσυστοιχίας υπεισέρχονται δύο είδη θορύβου ο **θόρυβος πηγής** και ο **θόρυβος δέκτη**. Ο πηγαίος θόρυβος έχει να κάνει με το αν η γυάλινη επιφάνεια της μικροσυστοιχίας μας έχει σκόνη και από την κατεργασία του γυαλιού. Ο θόρυβος δέκτη έχει προκύπτει από την μετατροπή του σήματος από φωτεινό σε ηλεκτρικό, την ενίσχυση του και την ψηφιακή μετατροπή (κβάντιση). Ανάλογα με τις ρυθμίσεις του σαρωτή όπως την ένταση του φωτός διέγερσης την τάση ενίσχυσης έχουμε διαφορετικό τελικό αποτέλεσμα. Για παράδειγμα μεγαλύτερη ένταση στο φως διέγερσης παράγει μεγαλύτερο θόρυβο πηγής και αντίστοιχα μεγαλύτερη ενίσχυση του ηλεκτρικού σήματος παράγει επιπλέον θόρυβο στον δέκτη. Γενικά προτιμάμε να ενισχύσουμε το σήμα της πηγής με την αύξηση της έντασης του φωτός διέγερσης γιατί απλά διεγείρουμε περισσότερο τα φθορίζοντα μόρια ώστε να μας δώσουν μεγαλύτερο σήμα γιατί με την μεγαλύτερη ενίσχυση του ηλεκτρικού σήματος «παράγουμε» σήμα. Επίσης με υπερβολική διέγερση των δειγμάτων μας καταστρέφουμε το υλικό γεγονός που εξαρτάται και από το αριθμό των σαρώσεων.

Το αποτέλεσμα της σάρωσης είναι δύο tiff μονοχρωματικές εικόνες 16bit, μια με τιμές έντασης στο κόκκινο (Cy5) και μια στο πράσινο (Cy3).

1.1.4 Επεξεργασία Εικόνας.

Η επεξεργασία της εικόνας είναι η βασική διαδικασία με την οποία ασχολείται η παρούσα εργασία. Στόχος της επεξεργασίας αυτής είναι η εξαγωγή ενός ποσοστού έκφρασης για κάθε ένα από τα γονίδια που μελετάται. Κατ' επέκταση αυτό σημαίνει ότι θέλουμε να αντιστοιχίσουμε κάθε spot της εικόνας σε μια τιμή έκφρασης. Αυτό γίνεται αρχικά με υπολογισμό μιας μέσης έντασης του spot για τα δύο κανάλια, κόκκινο και πράσινο και έπειτα υπολογισμό της αναλογίας τους όπως αναλύθηκε και σε παραπάνω ενότητα.

Από την διαδικασία σάρωσης έχουμε στην διάθεση μας δύο εικόνες για κάθε μικροσυστοιχία, μια στο κανάλι Cy3 (πράσινο) και μια στο κανάλι Cy5 (κόκκινο). Συνδυάζουμε αυτές τις δύο εικόνες σε μία δυκάναλη η οποία στο σύστημα RGB έχει τιμές έντασης μόνο για τους άξονες R και G ενώ έχουμε μηδενικές τιμές έντασης για το μπλε χρώμα. Ακολουθεί μια διαδικασία επεξεργασίας της δυκάναλης εικόνας με στόχο την εξαγωγής του ποσοστού έκφρασης για κάθε spot.

Η επεξεργασία της εικόνας μπορεί να χωριστεί σε τρία βήματα:

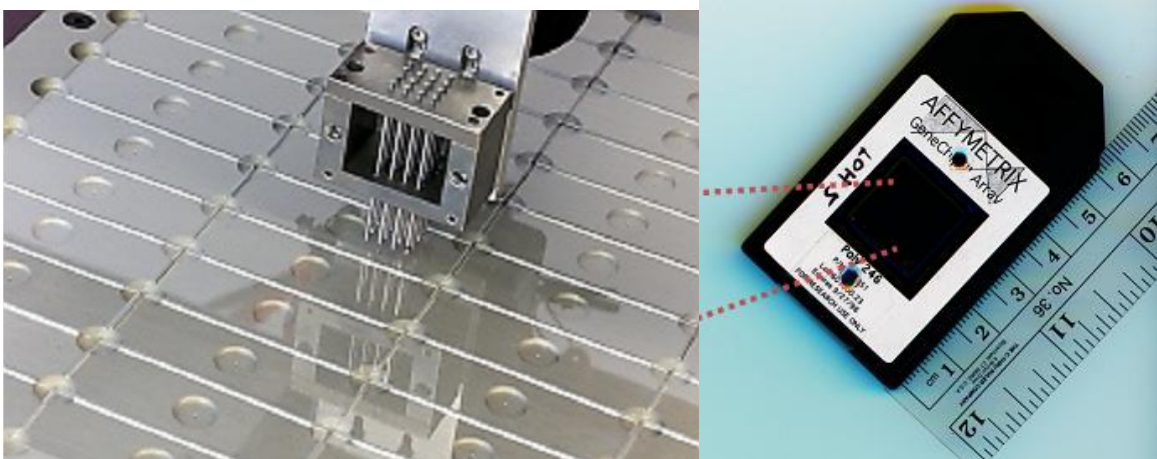
1. **Τοποθέτηση Πλέγματος.** Σχεδίαση του πίνακα της μικροσυστοιχίας μας, δηλαδή εύρεση της θέσης κάθε spot πάνω στην εικόνα.
2. **Κατάτμηση.** Διαχωρισμός των spot από το background.
3. **Υπολογισμός ποσοστού έκφρασης.** Εξαγωγή της χρήσιμης πληροφορίας η οποία είναι η

αναλογία της έντασης του κόκκινου προς το πράσινο για κάθε spot.

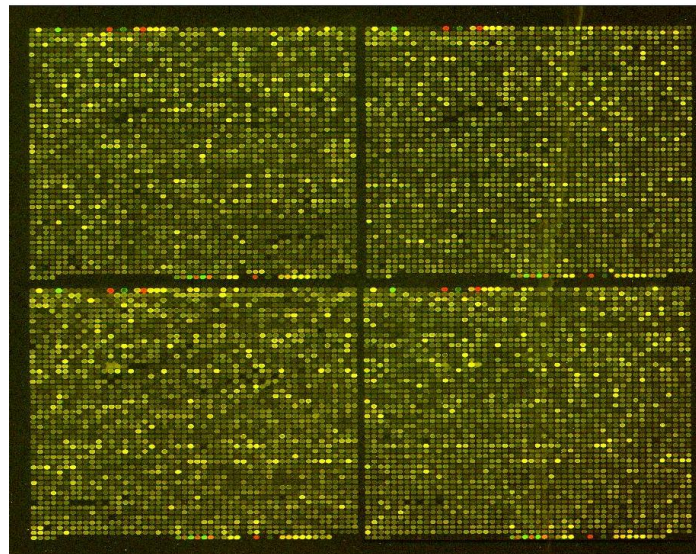
Η διαδικασία **τοποθέτησης πλέγματος** είναι πάρα πολύ σημαντική καθώς σε αυτήν βασίζονται τα δύο επόμενα βήματα. Κάθε μικροσυστοιχία έχει την δικιά της δομή ανάλογα με τον κατασκευαστή. Κάποιες μικροσυστοιχίες αποτελούνται από ένα πίνακα από spot συγκεκριμένων διαστάσεων και κάποιες αποτελούνται από ένα πίνακα μικρότερων πινάκων από spot. Έτσι η **τοποθέτηση πλέγματος** (gridding) σε μια εικόνα που αποτελείται από έναν αριθμό από μικρότερους πίνακες συνεπάγεται στην τοποθέτηση του αντίστοιχου αριθμού από υπο-πλέγματα (sub-grid).

Υπάρχουν διάφοροι αλγόριθμοι τοποθέτησης πλέγματος κάποιοι προχωράνε σε διαχωρισμό της εικόνας σε μικρότερες ώστε η κάθε μια να περιέχει μόνο ένα υπο-πίνακα ενώ άλλοι αντιμετωπίζουν την εικόνα ως μια οντότητα. Στην ουσία η διαδικασία τοποθέτησης πλέγματος μεταφράζεται στην εύρεση των διαχωριστικών γραμμών ανάμεσα στις σειρές και στις στήλες των υπο-πινάκων και κατ' επέκταση ανάμεσα στις σειρές και στις στήλες των spot. Σε τελική φάση έχουμε ένα σύνολο από συντεταγμένες οι οποίες ορίζουν τον χώρο που βρίσκεται κάθε ένα από τα spot, ο χώρος αυτός ονομάζεται *κελί ή κυψελίδα (cell)*.

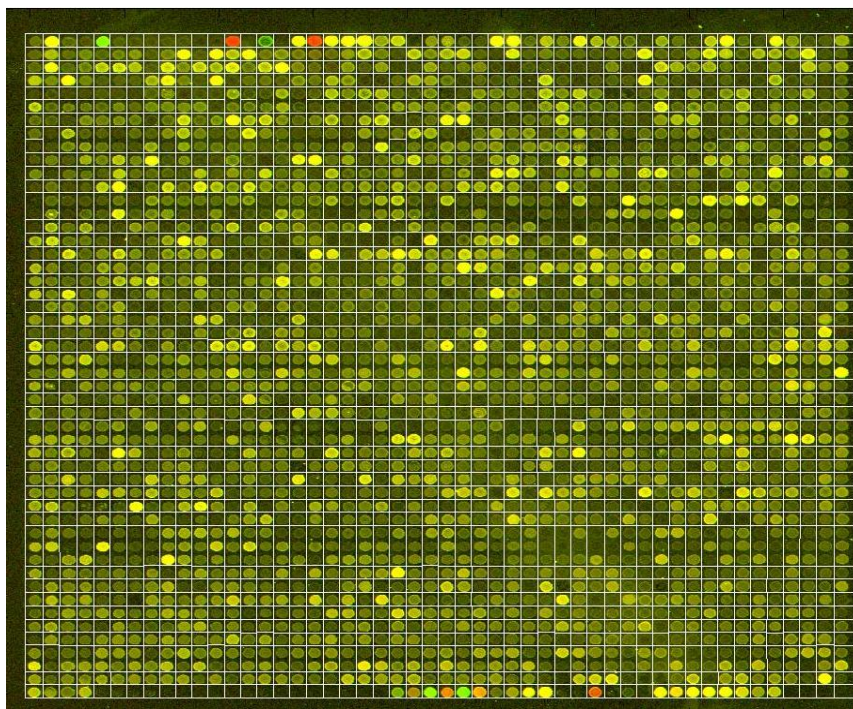
Σε έναν αλγόριθμο που αντιμετωπίζει την εικόνα ως μια οντότητα χρειάζονται διάφοροι παράμετροι πριν προχωρήσει στην διαδικασία διευθυνσιοδότησης κάθε υπο-πίνακα και κάθε spot. Οι παράμετροι θα μπορούσαν να είναι για παράδειγμα συνδυασμός των ακόλουθων μεταβλητών: οριζόντια και κάθετη απόσταση ανάμεσα στους υπο-πίνακες, οριζόντια και κατακόρυφη απόσταση ανάμεσα στα spot κάθε υπο-πίνακα, η συνολική θέση της μικροσυστοιχίας πάνω στην εικόνα. Σε εικόνες όπου ο μηχανισμός εκτύπωσης των spot δίνει για παράδειγμα μια μικρή κλίση σε κάθε σειρά ή στήλη του πίνακα τότε ίσως χρειάζονται περισσότεροι παράμετροι.



Εικόνα. 1.1.6 Δύο βασικές τεχνολογίες παραγωγής μικροσυστοιχιών. **Αριστερά: Spotted Microarrays** Κατασκευάζονται με απ' ευθείας σύνθεση φωτολιθογραφικά μέσω ακίδων οι οποίες εναποθέτουν υλικό σε υποδοχές. **Δεξιά: Affymetrix Gene Chips** Κάθε Chip μπορεί να χειριστεί μόνο ένα φλωροχρώμιο οπότε για την σύγκριση δύο δειγμάτων DNA απαιτούνται δύο chips.



Εικόνα. 1.1.8 Μια μικροσυστοιχία η οποία αποτελείται από ένα πίνακα από 2x2 υπο-πίνακες. Ο κάθε υπο-πίνακας αποτελείται από 50x50 spot. Η εικόνα αυτή προέρχεται από την βάση δεδομένων του **National Cancer Institute** <http://dc.nci.nih.gov/dataSets/geawQCandIA>



Εικόνα 1.1.9 Αποτέλεσμα τοποθέτησης πλέγματος για τον πρώτο υπο-πίνακα της μικροσυστοιχίας μας. Εάν η τοποθέτηση πλέγματος έχει γίνει σωστά μέσα σε κάθε κελί περιέχεται μόνο ένα spot.

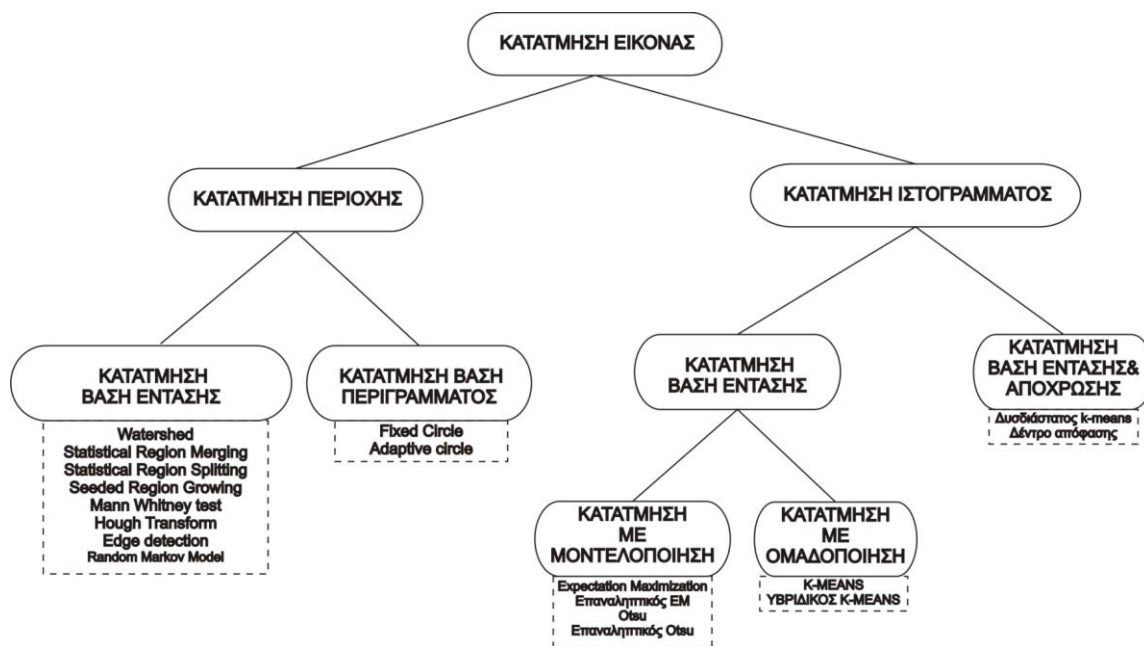
Λόγω της μεγάλης ποικιλίας στην δομή των μικροσυστοιχιών η τοποθέτηση πλέγματος απαιτεί την παρέμβαση του χρήστη. Υπάρχουν ήδη έτοιμα εμπορικά αλλά και ελεύθερα προγράμματα τα οποία πραγματοποιούν όλα τα επίπεδα της επεξεργασίας της εικόνας και στην περίπτωση της τοποθέτησης πλέγματος υποστηρίζουν και χειροκίνητη λειτουργία.

Το δεύτερο βήμα στην επεξεργασία της εικόνας είναι η διαδικασία **κατάτμησης**, δηλαδή ο διαχωρισμός των pixel της εικόνας σε δύο σύνολα: το background και το foreground. Έχει υλοποιηθεί ένας μεγάλος αριθμός αλγορίθμων κατάτμησης. Κάποιοι αλγόριθμοι βασίζονται μόνο στην τιμή των

εντάσεων R και G και στα ιστογράμματα τους και κάποιοι άλλοι συνδυάζουν τα χαρακτηριστικά αυτά με χωρική πληροφορία. Μπορούμε να χωρίσουμε τους αλγόριθμους ταξινόμησης στις τέσσερις παρακάτω κατηγορίες:

1. Κατάτμηση περιοχής
 - A. Κατάτμηση βάση περιγράμματος.
 - B. Κατάτμηση βάση έντασης
2. Κατάτμηση βάση ιστογράμματος
 - A. Κατάτμηση βάση έντασης
 - i. Μοντελοποίηση
 - ii. Κατηγοριοποίηση - Ομαδοποίηση
 - B. Κατάτμηση βάση συνδυασμού έντασης και χρώματος.
 - i. Δυσδιάστατη κατάτμηση
 - ii. Ακολουθιακή Κατάτμηση.

Στην **κατάτμηση βάση περιγράμματος** ανήκουν οι αλγόριθμοι κατάτμησης σταθερού και μεταβαλλόμενου κύκλου ^[27] τέτοιοι αλγόριθμοι υποστηρίζονται από όλα τα ήδη υπάρχοντα λογισμικά (Imagene ^[31], ScanAlyze ^[7], Genepix ^[24]). Στην **χωρική κατάτμηση βάση έντασης** ανήκουν αλγόριθμοι οι οποίοι χρησιμοποιούν την πληροφορία της έντασης σε συνδυασμό με χωρική πληροφορία για παράδειγμα τιμές έντασης των γειτονικών pixels όπως Watershed ^[8], Region Growing ^{[9],[10]}, Region Splitting, Mann Whitney test ^[11], Hough Transform ^[12], Edge detection ^[26], Seeded region Growing (SRG) ^[27]. Στην κατάτμηση με **Μοντελοποίηση** ανήκουν οι αλγόριθμοι Expectation Maximization ^{[13-17],[23]} και Otsu ^[18]. Στην κατάτμηση με ομαδοποίηση έχουμε τον αλγόριθμο k-means και τον υβριδικό k-means ^[23]. Τέλος στην κατάτμηση με **συνδυασμό έντασης και χρώματος** έχουμε τον δυσδιάστατο k-means και στην



Εικόνα 1.1.10 Κατηγοριοποίηση Αλγορίθμων Κατάτμησης. Κάτω από κάθε κατηγορία υπάρχουν παραδείγματα αλγορίθμων.

ακολουθιακή ταξινόμηση σε πρώτη φάση έχουμε κατάτμηση με βάση έντασης με έναν από τους προηγούμενους αλγορίθμους και στην συνέχεια εφαρμογή band-pass φίλτρου στο ιστογράμμα της απόχρωσης.

Το τελευταίο βήμα της επεξεργασίας εικόνων μικροσυστοιχιών είναι το στάδιο της εξαγωγής

ποσοστού έκφρασης από το οποίο προκύπτει η πληροφορία η οποία μετέπειτα μελετάται για την εξαγωγή συμπερασμάτων όσον αφορά την ομοιότητα των γονιδίων. Υπολογίζουμε για κάθε spot την τιμή έκφρασης, δηλαδή την αναλογία έκφρασης των καρκινικών προς τα υγιή δείγματα. Ο υπολογισμός γίνεται με βάση την ένταση του κόκκινου προς την ένταση του πράσινου. Για λόγους που έχουν να κάνουν με την διαφορά στην κλίμακα των τιμών για το κόκκινο και το πράσινο πραγματοποιούμε κανονικοποίηση των τιμών δηλαδή έχουμε

$$\text{Τιμή έκφρασης}_{\text{spot},i} = \frac{\text{Normalized}(R_i)}{\text{Normalized}(G_i)} \quad (1.1.2.1)$$

Όπου R_i και G_i οι μέσες τιμές των εντάσεων του κόκκινου και του πράσινου για το αντίστοιχο spot. Η κανονικοποίηση γίνεται σε κάθε κανάλι χρώματος ξεχωριστά, συγκεκριμένα γίνεται με βάση την μέγιστη μέση τιμή spot (για το αντίστοιχο χρώμα) που συναντάμε στην εικόνα. Δηλαδή έχουμε:

$$\text{Τιμή έκφρασης}_{\text{spot},i} = \frac{\frac{(R_i)}{\max_{i=1..k} R_i}}{\frac{(G_i)}{\max_{i=1..k} G_i}} \quad (1.1.2.2)$$

όπου k ο αριθμός των spot της μικροσυστοιχίας και R_i , G_i οι μέσες τιμές του κόκκινου και του πράσινου για το spot.

Εναλλακτικοί τύποι οι οποίοι χρησιμοποιούνται είναι:

$$\text{Τιμή έκφρασης}_{\text{spot},i} = \frac{\text{Normalized}(R_i)}{\text{Normalized}(G_i)} \quad \text{όταν } R_i > G_i \quad (1.1.2.3)$$

$$\text{Τιμή έκφρασης}_{\text{spot},i} = -\frac{\text{Normalized}(G_i)}{\text{Normalized}(R_i)} \quad \text{όταν } G_i > R_i \quad (1.1.2.4)$$

Επίσης σε πολλές εργασίες βρίσκει εφαρμογή ο εξής τύπος:

$$\text{Ποσοστό έκφρασης}_{\text{spot},i} = \frac{\log_2(R_i)}{\log_2(G_i)} \quad (1.1.2.5)$$

Τέλος για τον υπολογισμό του σφάλματος υπολογισμού των ποσοστών έκφρασης συμπεριλαμβάνονται στις μικροσυστοιχίες κάποια **spot ελέγχου** (yeast controls). Τα spot αυτά προέρχονται από γνωστά γονίδια συνήθως κάποιων μικροοργανισμών και τυπώνονται στις μικροσυστοιχίες σε προκαθορισμένες θέσεις και σε γνωστές αναλογίες μαζών (γνωστές αναλογίες κόκκινου προς πράσινο).

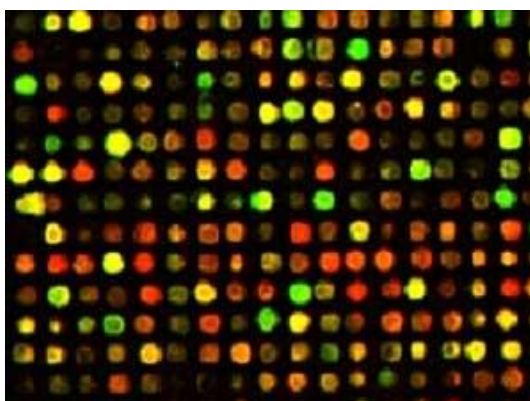
1.2 Προβλήματα που ανακύπτουν κατά την επεξεργασία της εικόνας

Η επεξεργασία εικόνων μικροσυστοιχιών είναι εμφανές ότι ασχολείται με συγκεκριμένο τύπο εικόνων με δεδομένη δομή και λογική και συγκεκριμένα προβλήματα. Όλες οι προηγούμενες μελέτες υλοποίησης αλγορίθμων κατάτμησης αντιμετώπισαν παρόμοιες δυσκολίες. Τα προβλήματα αυτά μπορούν να συνοψιστούν στα εξής:

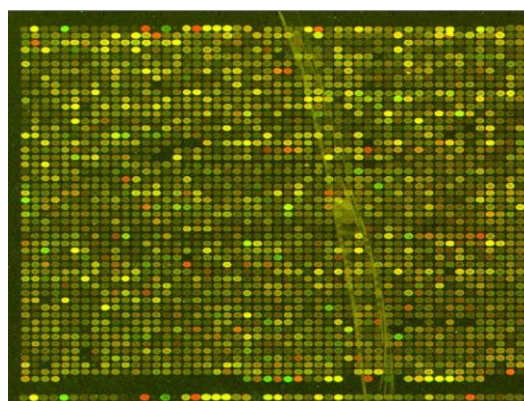
1. **Χαμηλή αντίθεση (Contrast):** Κάποιες εικόνες μικροσυστοιχιών έχουν χαμηλή αντίθεση για διάφορους λόγους που έχουν να κάνουν με την δημιουργία τους, για παράδειγμα λόγω ρυθμίσεων του σαρωτή. Σε αυτήν την περίπτωση τα spot είναι δυσδιάκριτα δηλαδή έχουν

τιμές εντάσεων κοντινές στις τιμές εντάσεων του background. Η χαμηλή αντίθεση δημιουργεί σε πρώτη φάση προβλήματα στην διαδικασία τοποθέτησης πλέγματος αλλά και μετέπειτα στον αλγόριθμο κατάτμησης καθώς είναι δύσκολο να διαχωριστούν τα pixel σε δύο κατηγορίες: spot και background. Το πρόβλημα αυτό βελτιώνεται συνήθως με την εφαρμογή φίλτρων την ενίσχυση της έντασης του foreground και υποβιβασμό της έντασης του background.

2. **Θόρυβος στο background:** Ο θόρυβος μπορεί να προέρχεται από διάφορες πηγές. Μια πηγή είναι κόκκοι σκόνης που έχει η επιφάνεια της μικροσυστοιχίας οι οποίοι διεγείρονται μαζί με τα μόρια φθορισμού και παράγουν σήματα και σχήματα που επηρεάζουν την τελική εικόνα της μικροσυστοιχίας. Μια άλλη αιτία είναι *σταγόνες* (blobs) και *μουτζούρες* (artifacts) τα οποία έχουν προκύψει από σφάλματα στον μηχανισμό εκτύπωσης των spots. Στις περιπτώσεις αυτές αυξάνεται η μέση ένταση του background με αποτέλεσμα να μειώνεται η απόσταση μεταξύ των δύο κατηγοριών (foreground και background) και η ικανότητα διάκρισης τους. Κάποιοι αλγόριθμοι ταξινόμησης και κυρίως η κατηγορία που εφαρμόζει thresholds πάνω στο ιστόγραμμα της έντασης και όσοι μοντελοποιούν τις καμπύλες των δύο κατηγοριών, ταξινομούν εσφαλμένα στο foreground κάποια pixel που ανήκουν στο background και το αντίστροφο. Αλγόριθμοι όπως edge detection ^[26] και Hough transform^[12] επιλύουν αυτό το πρόβλημα.
3. **Ανομοιογένεια μεταξύ των spot:** Τα spot που ανήκουν στην ίδια μικροσυστοιχία παρόλο που έχουν δημιουργηθεί από τον ίδιο μηχανισμό διαφέρουν σε σχήμα σε μέγεθος. Αυτό συμβαίνει λόγω ακανόνιστου υβριδισμού αλλά και μικρής ακρίβειας του μηχανισμού



Εικόνα 1.2.1 Ανομοιογένεια spot - Αλληλοεπικάλυψη spot. Τα μεγέθη και τα σχήματα των spot ποικίλουν.



Εικόνα 1.2.2 Background θόρυβος. Μια «μουτζούρα» πάνω στην μικροσυστοιχία που προκλήθηκε από ίσως από τον μηχανισμό εκτύπωσης ή λόγω σκόνης πάνω στην γυάλινη επιφάνεια.

εναπόθεσης του υλικού. Λόγω αυτής της ποικιλίας σχημάτων και μεγέθους οι αλγόριθμοι σταθερού κύκλου και μεταβαλλόμενου κύκλου έχουν κακή απόδοση ^[27].

4. **Αλληλοεπικάλυψη spot:** Περίεργα σχήματα δημιουργούνται στην περίπτωση όπου δύο spot ενώνονται λόγω μικρής απόστασης μεταξύ τους.
5. **Μεταβλητότητα στον προσανατολισμό των συστοιχιών:** Λόγω κάποιων ιδιομορφιών στους μηχανισμούς εκτύπωσης ορισμένοι υπο-πίνακες των μικροσυστοιχιών παρουσιάζουν μια κλίση, δηλαδή μια μικρή περιστροφή ή κάποια απόκλιση από την θέση τους. Παρόμοιο πρόβλημα είναι η απόκλιση που παρουσιάζουν κάποιοι πίνακες στις σειρές ή στις στήλες, δηλαδή δεν είναι τα spot στη ίδια νοητή ευθεία αλλά αποκλίνουν από αυτή. Τα προβλήματα αυτά επηρεάζουν κυρίως τους αλγόριθμους τοποθέτησης πλέγματος και επιλύονται με κάποια διόρθωση των εικόνων που προηγείται. Επίσης αλγόριθμοι τοποθέτησης πλέγματος που συνδυάζουν ανίχνευση ακμών σε συνδυασμό με τις τιμές έντασης έχουν πολύ καλύτερα αποτελέσματα.
6. **Μεταβλητότητα έντασης ανάμεσα στα pixel που ανήκουν στο background :** Το background

παίρνει τιμές από μια μεγάλη γκάμα εντάσεων. Αυτό οφείλεται σε λόγους θορύβου (μουντζούρες, σκόνη στη γυάλινη επιφάνεια κλπ.). Το πρόβλημα επιλύεται με μεταβλητές τιμές ορίων (**adaptive thresholding**). Σε πολλές μελέτες εφαρμόζεται **εξισορρόπηση background** χρησιμοποιώντας την μέση τιμή έντασης των pixel που βρίσκονται στις γωνίες της εικόνας. Σε άλλες γίνεται διόρθωση του background με σταθερή (constant) διόρθωση ή με τοπική (local) διόρθωση ^[28]. Μια πιο πολύπλοκη λύση είναι η **μοντελοποίηση** της κατανομής της έντασης και βελτιστοποίηση των σημείων απόφασης. Μια άλλη ενδιαφέρουσα προσέγγιση είναι το **TV+L μοντέλο** ^[28] το οποίο δίνει την διορθωμένη ένταση που είναι πιο κοντά στα πραγματικά δεδομένα.

7. **Μεταβλητότητα έντασης ανάμεσα στα pixel που ανήκουν στο foreground:** Αυτό οφείλεται στο γεγονός ότι τα εσωτερικά pixel του spot είναι πιο σκοτεινά από τα εξωτερικά. Το πρόβλημα αυτό επιλύεται και πάλι μεταβλητές τιμές ορίων (**adaptive thresholding**) και με μοντελοποίηση των κατανομών.
8. **Εσφαλμένη διαγραφή spot - άδεια κελιά.** Τα άδεια κελιά, δηλαδή κελιά που δεν περιέχουν γενετικό υλικό πρέπει να εντοπίζονται και να μην συνυπολογίζονται στα τελικά αποτελέσματα. Άδεια κελιά προκύπτουν από σφάλματα εναπόθεσης υλικού και εντοπίζονται μέσω της μέσης τιμής της έντασης του κελιού η οποία είναι μικρή στην περίπτωση αυτή. Συνήθως ορίζεται κάποιο όριο κάτω από το οποίο θεωρούμε ότι δεν υπάρχει spot. Πρόβλημα αντιμετωπίζουμε με τα spot που έχουν πολύ μικρές τιμές έντασης και θεωρούνται ανύπαρκτα και διαγράφονται με αποτέλεσμα να έχουμε απώλεια πληροφορίας. Για να έχουμε πιο αξιόπιστες διαγραφές εφαρμόζουμε **λογάριθμο** ώστε να απλώσουμε τις τιμές έντασης και να αναδείξουμε τις μικρές διαφοροποιήσεις. Ένας άλλος τρόπος είναι ο υπολογισμός της **συσχέτισης** (correlation) τις μεσαίες με την μέση τιμή κάθε spot ^[29]. Στην περίπτωση του **Hough Transform** ^[12] απορρίπτονται τα spot που έχουν περιφέρεια κύκλου μικρότερη από το 50%.

1.3 Προηγούμενες Μελέτες

Πολλές προσπάθειες έχουν γίνει για την ανάπτυξη αλγορίθμων κατάτμησης με στόχο τον βέλτιστο διαχωρισμό του spot από το background ώστε να γίνει όσον τον δυνατόν πιο ακριβέστερος υπολογισμός των μέσων τιμών των εντάσεων για κάθε κανάλι και της τιμής του ποσοστού έκφρασης.

Μια πρόσφατη υλοποίηση που προτείνει μια μέθοδο κατάτμησης εικόνων μικροσυστοιχιών είναι των **Th.Margaritis, K. Marias, Kafetzopoulos, "Improved Microarray Spot Segmentation by Combining two Information Channels"** ^[34] το 2006. Η εργασία αυτή ορίζει μια περιοχή στην οποία θεωρεί ότι ανήκει το spot TSA και εφαρμόζει Bayesian Information Criterion (BIC) στον λογάριθμο του γινομένου της έντασης του κόκκινου και του πράσινου $M = \log_2(R \times G)$. Έπειτα εφαρμόζει fuzzy αλγόριθμο ταξινόμησης με αριθμό κλάσεων όπως προκύπτει από το κριτήριο BIC. Από όλες της κλάσεις επιλέγεται αυτή με το μεγαλύτερο αριθμό Pixel με την προϋπόθεση ότι τουλάχιστον το 50% των pixel βρίσκεται στον χώρο του spot TSA.

Οι μικροσυστοιχίες έχουν κατασκευαστεί έτσι ώστε να περιέχουν spot ελέγχου με γνωστές αναλογίες, συγκεκριμένα κάποια από αυτά δημιουργήθηκαν ώστε να έχουν πολύ καλή ποιότητα εκτύπωσης και κάποια άλλα θορυβώδη. Ο αλγόριθμος βασίζεται στην παραδοχή ότι τα pixel που ανήκουν σε όμοια spot (spot με το ίδιο υλικό και αναλογία πράσινου προς κόκκινου, τυπωμένα σε αντίγραφα σε διάφορες θέσεις πάνω στην μικροσυστοιχία) έχουν την ίδια τιμή $\log(R/G)$ και ότι σε ένα «τέλειο» spot χωρίς θόρυβο το εύρος τιμών του λογαρίθμου αυτού πρέπει να είναι θεωρητικά μηδέν. Συγκρίνονται παραλλαγές του αλγορίθμου με έναν σύνολο από γνωστούς αλγορίθμους κατάτμησης (πχ SRG, GOGAC, ImageJ). Ως μέτρο σύγκρισης των αλγορίθμων χρησιμοποιείται το εύρος τιμών του παραπάνω λογάριθμου για τα όμοια spot.

Μια εργασία η οποία ασχολήθηκε με την σύγκριση των πιο γνωστών αλγορίθμων κατάτμησης είναι των *Antti Lehmussola, Pekka Ruusuvuori and Olli Yli-Harja, "Evaluating the performance of microarray segmentation algorithms"*^[23]. Μελετάει την απόδοση εννέα πολύ γνωστών αλγορίθμων Fixed Circle algorithm, Adaptive Circle, Seeded Region Growing, Mann Whitney test, k-means, hybrid k-means, Random Markov Model, Model based algorithm, και ένα toolbox της Matlab. Η σύγκριση γίνεται πάνω σε πραγματικά δεδομένα αλλά και με προσομοίωση.

Ένα άλλο paper το οποίο πραγματοποιεί κατάτμηση μέσω της μοντελοποίησης της συνολικής κατανομής της έντασης είναι των *Qunhua Li, Chris Fraley, Roger E. Bumgarner, Ka Yee Yeung and Adrian E. Raftery, "Donuts, scratches and blanks: robust model – based segmentation of microarray images" to 2005*^[14]. Αρχικά η εργασία ασχολείται με το πρόβλημα της τοποθέτησης πλέγματος και παρουσιάζει μια μέθοδο η οποία εντοπίζει τις θέσεις των spot μέσω του αθροίσματος των εντάσεων ανά σειρά και ανά στήλη. Έπειτα μοντελοποιεί την κατανομή της έντασης ως άθροισμα Gaussian κατανομών. Έπειτα χρησιμοποιεί χωρική πληροφορία για να αφαιρέσει τμήματα που δεν ανήκουν στο spot, αραβουργήματα (artifacts) με την ιδιότητα ότι τα τμήματα αυτά αποτελούν μικρές συνδεδεμένες ομάδες από pixel. Εφαρμόζεται ένα threshold στον αριθμό συνδεσιμότητας των pixel και με βάση το όριο αυτό αφαιρούνται ομάδες από pixel.

1.4 Στόχοι

Η παρούσα εργασία υλοποιεί και τα τρία βήματα επεξεργασίας εικόνων μικροσυστοιχιών (τοποθέτηση πλέγματος, κατάτμηση, εξαγωγή ποσοστού έκφρασης) με μεγαλύτερη έμφαση στο στάδιο της κατάτμησης.

Το στάδιο τοποθέτησης πλέγματος είναι μια απαραίτητη προ-επεξεργασία της εικόνας και καθορίζει και την απόδοση των αλγορίθμων κατάτμησης αλλά και την εξαγωγή ποσοστού έκφρασης. Η διευθυνσιοδότηση των spot γίνεται γιατί οι αλγόριθμοι κατάτμησης αντιμετωπίζουν κάθε spot σαν μια ξεχωριστή οντότητα. Θεωρούμε ότι έχει γίνει σωστή τοποθέτηση πλέγματος όταν κάθε κελί περιέχει ολόκληρο και μόνο ένα spot.

Με μεγαλύτερη έμφαση στην διαδικασία της κατάτμησης είχαμε ως στόχο εκτός από την πληροφορία της τιμής έντασης των pixel να χρησιμοποιήσουμε επιπρόσθετα την πληροφορία της απόχρωσης κάθε pixel. Για το λόγο αυτό εργαστήκαμε στο HSI σύστημα αναπαράστασης εικόνων αντί για το RGB το οποίο χρησιμοποιούν οι περισσότερες μελέτες. Επιλέξαμε τέσσερις απλούς ήδη υλοποιημένους αλγόριθμους ταξινόμησης και για λόγους που είχαν να κάνουν με τις παραδοχές μας, πραγματοποιήσαμε μικρές παραλλαγές αυτών. Οι αλγόριθμοι που επιλέξαμε είναι: Otsu, k-means, Expectation Maximization. Οι αλγόριθμοι Otsu και k-means επιλεχθήκαν για λόγους που έχουν να κάνουν με την απλότητα και την εφαρμοσιμότητα τους στην ειδική περίπτωση των εικόνων μικροσυστοιχειών μας. Ο αλγόριθμος EM επιλέχθηκε ως πιο πολύπλοκος και πιο προσαρμόσιμος (επιλογή αριθμού παραμέτρων και εξίσωσης κατανομής).

Στόχοι μας ήταν να μελετήσουμε την αξία του χαρακτηριστικού της απόχρωσης, να μετρήσουμε την απόδοση των αλγορίθμων κατάτμησης, να εντοπίσουμε τυχόν βελτιώσεις, να υπολογίσουμε το ποσοστό έκφρασης και να το συγκρίνουμε με ήδη υπάρχοντα αποτελέσματα.

Ως επέκταση της εφαρμογής των παραπάνω αλγορίθμων υλοποιήσαμε ένα αλγόριθμο συμπλήρωσης του κέντρου κάθε spot ο οποίος κάνει χρήση χωρικής πληροφορίας. Παρόμοια προσέγγιση συναντάτε και στην εργασία των *Th. Margaritis, K. Marias, Kafetzopoulos*^[34] όπου ορίζεται μια περιοχή του spot TSA η οποία χρησιμοποιείται συμπληρωματικά για την επιλογή της κλάσης του spot. Μελετήσαμε και συγκρίναμε την απόδοση των πρώτων αλγορίθμων που ασχολούνται μόνο με πληροφορίες που προκύπτουν από τις κατανομές των χαρακτηριστικών με την επιπρόσθετη εφαρμογή του αλγορίθμου συμπλήρωσης του spot.

Χρησιμοποιήσαμε την βάση εικόνων μικροσυστοιχιών και δεδομένων του Εθνικού Ινστιτούτου Καρκίνου (National Cancer Institute^[25]) καθώς επίσης και τα αποτελέσματα κατάτμησης που προκύπτουν από τον αλγόριθμο επεξεργασίας εικόνας του λογισμικού GEMTools^[24]. Οι συγκεκριμένες

εικόνες δημιουργήθηκαν από την **Incyte Genomics** για λόγους σύγκρισης των παραμέτρων σάρωσης ανάμεσα σε διάφορα πειράματα.

1.5 Προ-επεξεργασία

Παρακάτω ακολουθεί σε βήματα η προ-επεξεργασία που προηγήθηκε της διαδικασίας της κατάτμησης.

1. Ενίσχυση της αντίθεσης των εικόνων
2. Τοποθέτηση πλέγματος.
3. Μετατροπή από το RGB στο HSI σύστημα αναπαράστασης.

Σε πρώτη φάση και μόνο για την καλύτερη τοποθέτηση πλέγματος πραγματοποιούμε μια ενίσχυση στην αντίθεση (contrast) των εικόνων μας καθώς είναι γενικά πολύ σκοτεινές και αυτό επηρέαζε των αλγόριθμο τοποθέτησης πλέγματος που επιλέξαμε να υλοποιήσουμε. Η ενίσχυση αυτή έγινε με πρόγραμμα επεξεργασίας εικόνων (Corel Photo Paint).

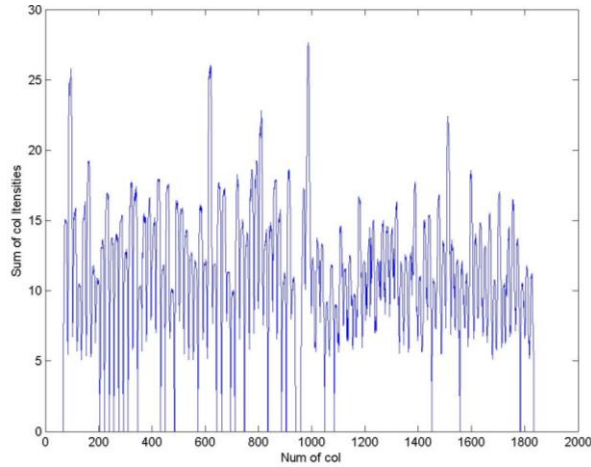
Στην συνέχεια ακολουθεί μια ενότητα όπου αναλύεται η μέθοδος τοποθέτησης πλέγματος την οποία υλοποιήσαμε σε πρώτη φάση για την διευθυνσιοδότηση των spot και μια δεύτερη ενότητα η οποία αναφέρεται στην μετάβαση από το χρωματικό μοντέλο RGB στο χρωματικό μοντέλο HSI. Η μετατροπή αυτή όπως έχουμε ήδη αναφέρει γίνεται για καλύτερη διάκριση της έντασης από το χρώμα με στόχο να χρησιμοποιήσουμε τα δύο αυτά χαρακτηριστικά σε συνδυασμό.

1.5.1 Τοποθέτηση Πλέγματος

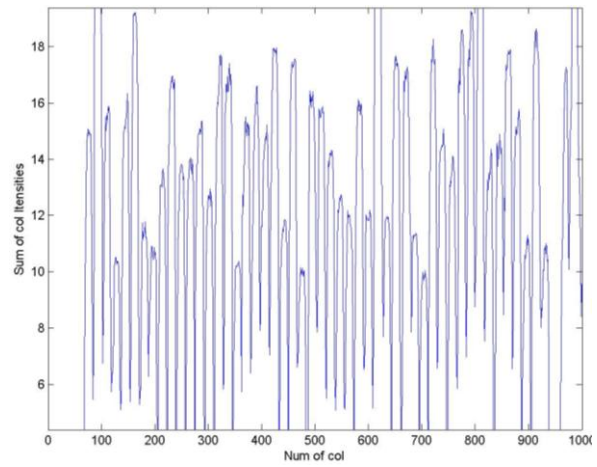
Η τοποθέτηση πλέγματος βασίζεται στο άθροισμα των εντάσεων ανά γραμμή και ανά στήλη πάνω σε ολόκληρη την εικόνα ^[14] (Donuts scratches and blanks: robust model based segmentation of Microarrays Quanhua Li, Chris Fraley, Roger E. Bumgarner ka Yee Yeung and Andrian E. Raftery).

Με απλά λόγια τα βήματα που ακολουθεί ο βασικός αλγόριθμος είναι:

1. Βρίσκουμε το **άθροισμα των εντάσεων** ανά σειρά και ανά στήλη στην εικόνα.
2. Προκύπτουν δύο ανύσματα: ένα με το άθροισμα των εντάσεων ανά σειρά και ένα με το άθροισμα των εντάσεων ανά στήλη τα οποία έχουν την μορφή που φαίνεται στην γραφική 1.3.1.
3. Ορίζουμε δύο παράθυρα ένα για την οριζόντια απόσταση, (πλάτος των spots) και ένα παράθυρο για την κατακόρυφη απόσταση (ύψος των spots). Ορίζουμε δηλαδή τις τυπικές διαστάσεις ενός spot όπου στην περίπτωση της μικροσυστοιχίας μας είναι κατά μέσο όρο 15 pixel το ύψος και 18 το πλάτος.
4. Κυλώντας τα παράθυρα βρίσκουμε τα τοπικά ελάχιστα κάθε τμήματος της γραφικής του αθροίσματος εκτιμώντας ότι αυτά τοπικά ελάχιστα αποτελούν ένδειξη κενού χώρου ανάμεσα στις σειρές και στις στήλες των spot.
5. Τοποθετούμε κάθε τοπικό ελάχιστο ως την θέση που θα «ζωγραφίσουμε» μια γραμμή ή αντίστοιχα μια στήλη στον πίνακα της μικροσυστοιχίας.



Γρ 1.5.1.1 Η γραφική του αθροίσματος των εντάσεων των στηλών της εικόνας.



Γρ 1.5.1.2 Τμήμα της γραφικής του αθροίσματος των εντάσεων των στηλών της εικόνας. Παρατηρούμε ότι υπάρχουν κατά περιοδικό τρόπο, τοπικά ελάχιστα (κενό ανάμεσα στις στήλες του microarray) και μέγιστα (τα κέντρα των spots).

Ο αλγόριθμος δέχεται ως παραμέτρους το ύψος και το πλάτος ενός τυπικού spot, την οριζόντια και κατακόρυφη απόσταση των υπο-πινάκων της μικροσυστοιχίας καθώς και παραμέτρους για την δομή της για παράδειγμα τον αριθμό των υπο-πινάκων και τις διαστάσεις των υπο-πινάκων. Οι μικροσυστοιχίες μας αποτελούνται 4 υπο-πίνακες (2x2) διαστάσεων 50x50 (2500 spots).

Για να βρούμε την θέση της μικροσυστοιχίας πάνω στην εικόνα (Σειρά, Στήλη) -δηλαδή τις συντεταγμένες του πρώτου spot πάνω αριστερά- κυλάμε και πάλι τα δύο παράθυρα ώστε:

$$\text{Στήλη}=w1 \text{ όταν } \frac{\max(\text{ΑθροισμαΣτηλών}(w1:w2)) - \text{MIN στηλών}}{\text{MAX στηλών} - \text{MIN στηλών}} > \text{Threshold} \quad (1.3.1)$$

$$\text{Σειρά}=w3 \text{ όταν } \frac{\max(\text{ΑθροισμαΣειρών}(w3:w4)) - \text{MIN σειρών}}{\text{MAX σειρών} - \text{MIN σειρών}} > \text{Threshold} \quad (1.3.2)$$

$$\text{MAX στηλών} = \max(\text{ΑθροισμαΣτηλών}) \quad (1.3.3)$$

$$\text{MAX σειρών} = \max(\text{ΑθροισμαΣειρών}) \quad (1.3.4)$$

$$\text{MIN στηλών} = \min(\text{ΑθροισμαΣτηλών}) \quad (1.3.5)$$

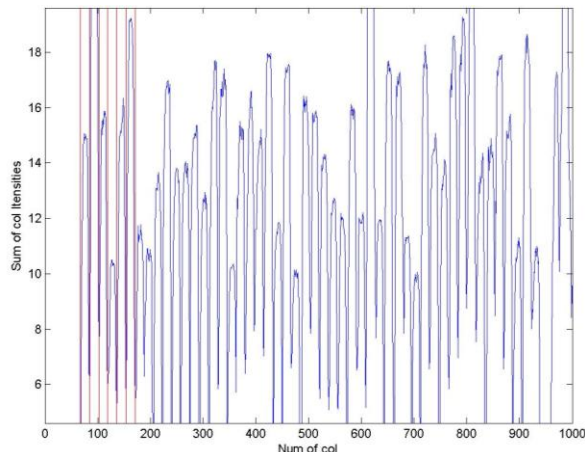
$$\text{MIN σειρών} = \min(\text{ΑθροισμαΣειρών}) \quad (1.3.6)$$

Όπου με w1:w2 ορίζουμε την θέση του παράθυρου των στηλών κατά την διάρκεια κύλισης και w3:w4 την θέση του παράθυρου των σειρών. Οι μεταβλητές MAXσειρών, MAXστηλών ορίζουν τα μέγιστα της γραφικής του αθροίσματος των σειρών και το μέγιστο της γραφικής του αθροίσματος των στηλών αντίστοιχα. Ομοίως με τις μεταβλητές MINσειρών, MINστηλών ορίζουμε τα ελάχιστα.

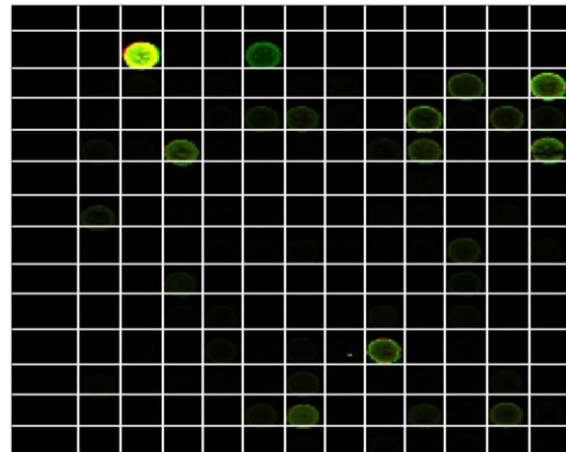
Στην απλή του εκδοχή ο αλγόριθμος αντιμετωπίζει την εικόνα ως ένα ενιαίο πλέγμα και με την βοήθεια των παραμέτρων μεταβαίνει από υπο-πίνακα σε υπο-πίνακα. Ως έξοδο δίνει τις θέσεις των καθολικών σειρών και στηλών στην εικόνα, δηλαδή ορίζει 102 στήλες και 102 σειρές.

Το μειονέκτημα του αλγορίθμου είναι ότι επειδή αθροίζει την ένταση σε όλη την έκταση της εικόνας επηρεάζεται από τυχόν αποκλίσεις των σειρών ή των στηλών που προκύπτουν από τον μηχανισμό εκτύπωσης των spot.

Σε μεγάλο ποσοστό οι εικόνες που έχουμε στην διάθεση μας είναι καλής ποιότητας, δηλαδή με ελάχιστες περιπτώσεις όπου έχουμε θόρυβο λόγο κακής εκτύπωσης των spot, όπως «μουτζούρες» ή απλωμένα spot. Όλες όμως παρουσιάζουν μια μικρή απόκλιση στην εκτύπωση των στηλών των spot σε κάποιους από τους υπο-πίνακες. Ως αποτέλεσμα ο αλγόριθμος δίνει λανθασμένα τοπικά ελάχιστα και πολλά κελιά περιέχουν τμήματα δύο ή και παραπάνω spot.



Γρ 1.5.1.3. Οι κόκκινες κατακόρυφες είναι τα πρώτα επτά τοπικά ελάχιστα στην γραφική αθροίσματος των εντάσεων των στηλών. Στα πρώτα αυτά τοπικά ελάχιστα τοποθετούμε τις πρώτες επτά στήλες του πλέγματος.



Γρ 1.5.1.4. Το αποτέλεσμα της τοποθέτησης πλέγματος σε κομμάτι του αριστερού υπο-πίνακα.

Για τους παραπάνω λόγους υλοποιήσαμε μια μικρή παραλλαγή του βασικού αλγορίθμου.

1. Ορίζουμε ένα δυσδιάστατο παράθυρο στις τυπικές διαστάσεις ενός spot (15x18).
2. «Κυλάμε» το παράθυρο σε όλη την έκταση της εικόνας και εντοπίζουμε δύο τοπικά ελάχιστα κάθε φορά ένα κατά στήλη και ένα κατά σειρά. Τα ελάχιστα αυτά προκύπτουν από την γραφική του αθροίσματος των σειρών και των στηλών αλλά μόνο σε ένα μικρό τμήμα της εικόνας, για παράδειγμα στις επόμενες τρεις σειρές και στήλες (από spot) αντίστοιχα της μικροσυστοιχίας μας.
3. Βρίσκουμε τέσσερις τιμές που ορίζουν την θέση κάθε κελιού, την θέση της πρώτης πάνω σειράς και της κάτω σειράς του spot, της δεξιάς στήλης και της αριστερής. Έτσι ορίζουμε την θέση του κάθε κελιού.

Ο τροποποιημένος αλγόριθμος δεν επηρεάζεται από τις αποκλίσεις στις θέσεις των σειρών και των στηλών.

1.5.2 Από το RGB στο HSI

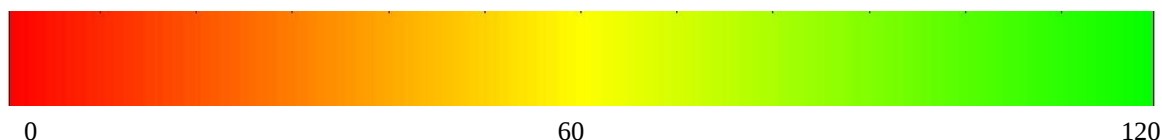
Έχουμε ήδη αναφέρει ότι ο βασικός μας στόχος είναι η βελτίωση του αποτελέσματος της διαδικασίας της κατάτμησης. Όπως είχαμε διαπιστώσει από προηγούμενες μελέτες η πληροφορία της έντασης ήταν αυτή που κυριαρχούσε στους αλγόριθμους κατάτμησης άλλωστε το τελικό ποσοστό της έκφρασης προκύπτει ως αναλογία εντάσεων, δηλαδή η τελική μας πληροφορία αφορά την ένταση. Με μια όμως απλή ματιά στις εικόνες μικροσυστοιχιών και κατανοώντας την διαδικασία δημιουργίας τους καθώς και όλη την λογική του υβριδισμού, μπορεί εύκολα κανείς να παρατηρήσει ότι και το χρώμα δίνει πολύ σημαντική πληροφορία για τα γονίδια που συμμετέχουν στον υβριδισμό.

Συγκεκριμένα ένας βιολόγος χωρίς να έχει στην διάθεση του ποσοστά έκφρασης με μια απλή ματιά στα χρώματα των spot μπορεί να εξάγει κάποια πρώτα συμπεράσματα για το ποια γονίδια

εκφράζονται και ποια όχι. Έχουμε ήδη εξηγήσει ότι στα πράσινα spot, δηλαδή τα spot όπου έχουμε μεγαλύτερη τιμή έντασης στο πράσινο απ' ότι στο κόκκινο, το ποσοστό υβριδισμού των «υγιών» γονιδίων είναι μεγαλύτερο από αυτό των καρκινικών. Πιο απλά τα τμήματα DNA που έχουμε απομονώσει από υγιή κύτταρα έχουν «ταιριάζει» σε περισσότερα σημεία τους ή σε μεγαλύτερη ποσότητα με τα τμήματα DNA βάσης (γονίδια βάσης) απ' ότι τα τμήματα DNA που απομονώθηκαν από καρκινικά κύτταρα. Το αντίστροφο έχουμε όταν κυριαρχεί η ένταση στο κόκκινο ενώ με κίτρινο έχουμε ταυτόχρονη έκφραση των καρκινικών και των υγιών γονιδίων

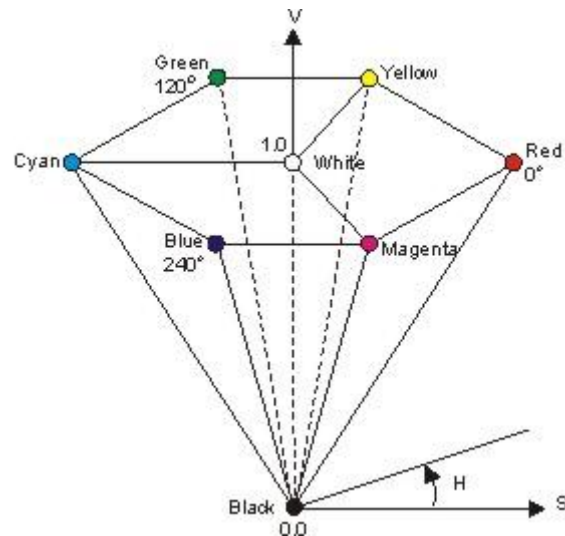
Συμπεραίνουμε λοιπόν ότι η πληροφορία του χρώματος είναι πολύ σημαντική για τον διαχωρισμό του spot από το background αλλά και για την περαιτέρω μελέτη του ποσοστού υβριδισμού των δειγμάτων μας. Για τον λόγο αυτό στην διαδικασία της κατάτμησης χρησιμοποιούμε δύο πληροφορίες την ένταση και την απόχρωση των pixel. Η πληροφορία της έντασης όπως είναι ήδη φανερό είναι η βασική πληροφορία για να ένα αξιόπιστο αποτέλεσμα και η απόχρωση έχει συμπληρωματικό ρόλο.

Τα χρώματα που συναντάμε στην εικόνα μας αφορούν μόνο ένα κομμάτι του συνολικού φάσματος ακριβώς επειδή η συνδυασμένη εικόνα προκύπτει από δύο μόνο κανάλια χρωμάτων κόκκινο και πράσινο. Ανεξάρτητα από την τιμή της έντασης τα pixel που ανήκουν στα spot παίρνουν τιμές χρώματος από το φάσμα που φαίνεται στην εικόνα 1.5.5



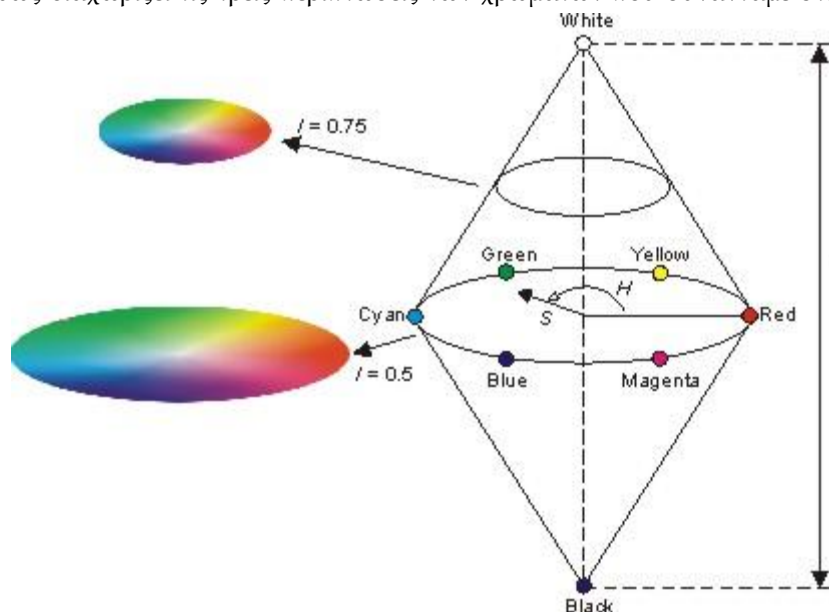
Εικόνα 1.5.15 Γκάμα χρωμάτων που συναντάμε σε εικόνες μικροσυστοιχιών. Η συνδυασμένη εικόνα μια μικροσυστοιχίας προέρχεται από δύο μόνο χρωματικά κανάλια κόκκινο και πράσινο. Για τον λόγο αυτό τα spot στην τελική μας εικόνα έχουν κόκκινο πράσινο και κίτρινο χρώμα και κατ' επέκταση τα Pixel της εικόνας παίρνουν μόνο συγκεκριμένες τιμές χρώματος. Στο σύστημα HSI έχουμε $H=0$ για κόκκινο $H=60$ για κίτρινο και $H=120$ για πράσινο.

Για να μπορέσουμε να εργαστούμε με τις δύο αυτές πληροφορίες την ένταση και την απόχρωση έπρεπε να μεταβούμε σε κάποιο σύστημα αναπαράστασης το οποίο να διαχωρίζει αυτά τα χαρακτηριστικά. Το σύστημα RGB δεν μας εξυπηρετεί καθώς συνδυάζει αυτά τα δύο χαρακτηριστικά δηλαδή έχουμε μόνο τρία χρώματα και ο διαχωρισμός αποχρώσεων γίνεται με βάση την ένταση στο κόκκινο στο πράσινο και στο μπλε.



Εικόνα 1.5.1.6 Το HSV χρωματικό μοντέλο. Τα χρώματα βρίσκονται πάνω σε μια εξαπλευρη πυραμίδα. ^[33]

Υπάρχουν διάφορα συστήματα τα οποία εξυπηρετούν στην περίπτωση των εικόνων μικροσυστοιχιών όπως HSV, L^*a^*b και HSI. Το σύστημα L^*a^*b αναπαριστά όλα τα χρώματα με τρεις τιμές: το L που είναι η τιμή της φωτεινότητας και παίρνει τιμές από 0 έως 100 και τα a , b τα οποία αντιστοιχούν σε $+a/-a$ για κόκκινο/πράσινο και $+b/-b$ για μπλέ/κίτρινο. Το γεγονός ότι οι τιμές του a μας δίνουν την πληροφορία για τα δύο κανάλια χρώματος κόκκινο και πράσινο και η τιμή $-b$ για το κίτρινο εξυπηρετεί καθώς διαχωρίζει τις τρεις περιπτώσεις των χρωμάτων που συναντάμε στις μικροσυστοιχίες



Εικόνα 1.5.1.7 Το HSI χρωματικό μοντέλο. Το μοντέλο το οποίο χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία. Αναπαριστά τα χρώματα με παρόμοιο τρόπο με αυτόν που τα αντιλαμβάνεται το ανθρώπινο μάτι διαχωρίζει τις τιμές έντασης από τις τιμές απόχρωσης.^[33]

μας. Για τον λόγο αυτό το σύστημα αυτό έχει ήδη εφαρμοστεί σε προηγούμενες μελέτες.

Το σύστημα HSV είναι παρόμοιο με το HSI και οι τιμές των χρωμάτων αναπαρίστανται με τρεις τιμές Απόχρωση (**H**ue) Πρόσμιξη (**S**aturation) και Τιμή (**V**alue). Τα χρώματα αναπαρίστανται σε εξάπλευρη πυραμιδική μορφή όπως φαίνεται στην εικόνα 1.5.6 η απόχρωση παίρνει τιμές από 0° έως

360° όπου ανά 60° έχουμε και διαφορετικό βασικό χρώμα (0° κόκκινο, 60° κίτρινο 120° πράσινο, 240° μπλε και 300° μωβ). Η Πρόσμιξη παίρνει τιμές από μηδέν έως ένα [0,1].

Στο σημείο όπου Τιμή=Πρόσμιξη=0 έχουμε το μαύρο ενώ για Τιμή=1 και Πρόσμιξη=0 έχουμε το άσπρο χρώμα. Στο σημείο όπου η Τιμή είναι ίση με μηδέν το χαρακτηριστικό της Πρόσμιξης δεν παίζει κανένα ρόλο.

Στο σύστημα HSI τα χρώματα είναι τοποθετημένα σε ένα δικωνικό σχήμα και αντιπροσωπεύονται από τρεις τιμές Απόχρωση (Hue) Ένταση (Intensity) και Πρόσμιξη (Saturation). Η Απόχρωση έχει την ίδια γκάμα τιμών με το προηγούμενο σύστημα [0°,360°] (0° κόκκινο, 60° κίτρινο 120° πράσινο, 240° μπλε και 300° μωβ). Η Πρόσμιξη δείχνει πόσο «γεμάτο» είναι το χρώμα και παίρνει τιμές στο διάστημα [0,1]. Οι τιμές της έντασης είναι από [0,1], το μαύρο χρώμα έχει μηδενική ένταση και το άσπρο μονάδα. Η τιμή της πρόσμιξης έχει μεγαλύτερη σημασία για τιμές εντάσεων το 0.5 απ' ότι για τιμές εντάσεων κοντά στο μηδέν ή στο ένα.

$$\begin{aligned}
 M &= \max(R, G, B) \\
 m &= \min(R, G, B) \\
 r &= \frac{(M - R)}{(M - m)} \\
 g &= \frac{(M - G)}{(M - m)} \\
 b &= \frac{(M - B)}{(M - m)} \\
 \text{Intensity Calculation} \\
 I &= \frac{(M + m)}{2} \\
 \text{Saturation Calculation} \\
 \text{If } I \leq .5 \text{ then } S &= \frac{(M - m)}{(M + m)} \\
 \text{If } I > .5 \text{ then } S &= \frac{(M - m)}{(2 - M - m)} \\
 \text{Hue Calculation} \\
 \text{If } R=M \text{ then } H &= 60 (2 + b - g) \\
 \text{If } G=M \text{ then } H &= 60 (4 + r - b) \\
 \text{If } B=M \text{ then } H &= 60 (6 + g - r) \\
 \text{If } M=m \text{ then } H &= 0
 \end{aligned}$$

Αλγ 1.5.1. Μετατροπή από το RGB στο HIS.
Τύποι υπολογισμών των τιμών (H,S,I) με βάση τις τιμές (R,G,B)

Ένας από τους λόγους για τους οποίους επιλέξαμε το χρωματικό μοντέλο HSI σε σύγκριση με το μοντέλο HSV είναι πρώτον για η μετατροπή από το RGB στο HSI είναι πιο απλή και έχουν αναπτυχθεί διάφοροι τύποι μετατροπής εκ των οποίων ένας πολύ απλός τρόπος μετατροπή τον οποίο και χρησιμοποιήσαμε. Επιπλέον το σύστημα HSI έχει ευρεία εφαρμογή σε συστήματα που περιλαμβάνουν επεξεργασία εικόνας καθώς αναπαριστά τα χρώματα με παρόμοιο τρόπο με αυτόν που τα αντιλαμβάνεται το ανθρώπινο μάτι και επειδή αναπαριστά την χρωματική πληροφορία με τρόπο ώστε να είναι δυνατή η σύγκριση χρωμάτων κάτι που στο RGB δεν γίνεται άμεσα.

Από την σάρωση κάθε μικροσυστοιχίας έχουμε δύο μονοχρωματικές εικόνες μια στο κόκκινο (Cy5) και μια στο πράσινο. (Cy3) για 70 μικροσυστοιχίες. Κατά συνέπεια για κάθε μικροσυστοιχία έχουμε στην διάθεση μας δύο πίνακες εντάσεων ένας για τον άξονα R και ένα για τον άξονα G και στο σύστημα RGB έχουμε μια εικόνα που έχει μηδενικές τιμές εντάσεων στο μπλε χρώμα.

Για κάθε pixel της εικόνας έχουμε τρεις τιμές εντάσεων (R,G,B) χρησιμοποιώντας τον Αλγ. 1.5.1 υπολογίζουμε τις τιμές της απόχρωσης της έντασης και της πρόσμιξης (H,S,I) στο σύστημα HSI. Οι αλγόριθμοι κατάτμησης εφαρμόζονται στο σύστημα HSI αλλά το ποσοστό έκφρασης υπολογίζεται με βάση τις αρχικές τιμές έντασης του κόκκινου και του πράσινου.

1.6 Δομή Εργασίας

Οι ενότητες που ακολουθούν περιγράφουν ένα σύνολο από αλγόριθμους κατάτμησης, κάποιους από τους οποίους υλοποιούμε στη συνέχεια, αναλύουμε τα αποτελέσματά τους και εξάγουμε κάποια συμπεράσματα.

Η πρώτη ενότητα (**ενότητα 2**) περιγράφει τεχνικές κατάτμησης οι οποίες χρησιμοποιούν μόνο χωρική πληροφορία ή συνδυασμό στατιστικής και χωρικής πληροφορίας (Κατάτμηση περιοχής). Στην επόμενη ενότητα (**ενότητα 3**) αναλύεται μια άλλη μεγάλη κατηγορία τεχνικών κατάτμησης όπου οι αλγόριθμοι χρησιμοποιούν ιστογράμματα και κατανομές τις έντασης για να διακρίνουν την κατηγορία του spot από αυτή του background. Στην ίδια ενότητα παραθέτουμε και κάποιες παραλλαγές γνωστών αλγορίθμων οι οποίοι έπειτα εφαρμόζονται στην βάση εικόνων μικροσυστοιχιών που έχουμε στην διάθεση μας.

Ακολουθεί μια ενότητα η οποία προχωράει σε εφαρμογή των αλγορίθμων που επιλέξαμε να τροποποιήσουμε και να εφαρμόσουμε στο σύστημα HSI. Στο σημείο αυτό γίνεται αναφορά στα προβλήματα που αντιμετωπίζει κάθε αλγόριθμος και πως επιλύονται. Στο τέλος του κεφαλαίου προστίθεται ένας αλγόριθμος ο οποίος χρησιμοποιεί χωρική πληροφορία για να ορίσει την θέση και τον χώρο στον οποίο βρίσκεται θεωρητικά το spot με στόχο να απορρίψει από το τελικό αποτέλεσμα της κατάτμησης κάποια σημεία που δεν ανήκουν σε αυτό λόγω θέσης.

Στην ενότητα 4 περιγράφονται τα πειράματα τα οποία πραγματοποιούμε με βάση τους αλγορίθμους που επιλέξαμε και υπολογίζεται το σφάλμα στην εξαγωγή ποσοστού έκφρασης. Συγκρίνουμε με το σφάλμα που προκύπτει από τα αποτελέσματα που έχουμε στην διάθεση μας από γνωστό λογισμικό.

Τέλος η ενότητα 5 συγκεντρώνει όλα τα συμπεράσματα στα οποία καταλήγουμε με βάση το σφάλμα που υπολογίσθηκε στην προηγούμενη ενότητα και πάνω στα πειράματα τα οποία πραγματοποιήσαμε.



2 ΚΑΤΑΤΜΗΣΗ ΠΕΡΙΟΧΗΣ

2.1 Κατάτμηση με βάση περίγραμμα

Οι αλγόριθμοι περιγράμματος ορίζουν μια περιοχή μέσα στην οποία θεωρούν ότι βρίσκεται το spot και αυτό γίνεται με καθορισμό του περιγράμματος του spot.

2.1.1 Κατάτμηση σταθερού κύκλου (Fixed Circle Segmentation)

Η μέθοδος αυτή υποθέτει ότι όλα τα spots έχουν το ίδιο **κυκλικό σχήμα** και τις **ίδιες διαστάσεις** και ορίζει μια **σταθερή διάμετρο** ίδια για όλα τα spot του Microarray. Εντοπίζει αρχικά το κέντρο του κάθε spot και στην συνέχεια δέχεται ως είσοδο την τιμή της διαμέτρου ή της ακτίνας των spots και ορίζει ένα κύκλο στο οποίο περιέχεται το spot.



Εικόνα 2.1.1. Τοποθέτηση μάσκας σταθερού κύκλου γύρω από το spot.

Μειονεκτήματα

Μια παράμετρος για την περιγραφή του σχήματος.

Είναι προφανές ότι πρώτον δεν μπορούμε να περιγράψουμε το σχήμα ενός spot μόνο με μια παράμετρο, την **διάμετρο** του. Τα spots ποικίλουν σε σχήματα, από κυκλικό, ελλειψοειδές μέχρι και ακανόνιστο σχήμα.

2. Ίδιες τιμές παραμέτρων για όλα τα spot.

Δεν μπορούν να περιγράψουν όλα τα spots της εικόνας με τις ίδιες τιμές παραμέτρων καθώς υπεισέρχεται όπως είπαμε το **σφάλμα του μηχανισμού εκτύπωσης** του οποίου τα ακροφύσια δεν σταθερά πάνω στην κεφαλή για να μην χαράξουν την επιφάνεια της μικροσυστοιχίας και για να μην σπασούν με αποτέλεσμα να δονούνται, δηλαδή δεν υπάρχει ακρίβεια στην θέση αλλά και στην ποσότητα του υλικού που εναποτίθεται. Υπάρχει επίσης και ο θόρυβος κατά την λήψη της φωτογραφίας αλλά και ο θόρυβος κατά την ευθυγράμμιση των δύο εικόνων (Cy3, Cy5).

Πλεονεκτήματα

Εύκολη υλοποίηση.

Η μέθοδος αυτή υλοποιείται πολύ απλά και έχει πολύ γρήγορά έναν τυπικό διαχωρισμό των spot από την εικόνα.

2.1.2 Κατάτμηση μεταβαλλόμενου κύκλου (Adaptive circle segmentation)

Στην μέθοδο αυτή έχουμε και πάλι μία παράμετρο για την περιγραφή του σχήματος του spot, την **διάμετρο**. Η διαφορά με την προηγούμενη μέθοδο είναι ότι ορίζουμε διαφορετική διάμετρο για κάθε spot. Έτσι ο κύκλος που ορίζει την περιοχή όπου θεωρούμε ότι βρίσκεται το spot εξαρτάται από το μέγεθος του και δεν είναι σταθερός.

Στα διάφορα software που χρησιμοποιούν αυτήν την μέθοδο (ImageJ^[31], ScanAlyze^[7], Genepix

[24].) ο χρήστης δίνει ως είσοδο μια μια τις διαμέτρους για κάθε spot αλλά αυτό είναι πολύ χρονοβόρο αν αναλογιστεί κανείς ότι κάθε microarray μπορεί να έχει μέχρι και 10.000 spots.

Μειονεκτήματα

Μια παράμετρος για την περιγραφή του σχήματος.

Όπως και στην προηγούμενη μέθοδο θεωρούμε ότι το spot έχει κυκλικό σχήμα κάτι το οποίο στην πραγματικότητα δεν ισχύει.

Πλεονεκτήματα: Εύκολη υλοποίηση.

2.2 Κατάτμηση βάση έντασης

Σ' αυτήν την κατηγορία αλγορίθμων κατάτμησης τίποτα δεν είναι σταθερό, ούτε το σχήμα αλλά ούτε και οι παράμετροι περιγραφής του spot είναι ίδιες σε όλη την έκταση της μικροσυστοιχίας. Χαρακτηριστικό των μεθόδων αυτών είναι ότι χρησιμοποιούν χωρική πληροφορία σε συνδυασμό με τις τιμές εντάσεων της εικόνας. Κατά κανόνα οι αλγόριθμοι που ανήκουν στην κατηγορία αυτή ορίζουν δύο σύνολα από pixel στα οποία προσθέτουν ή αφαιρούν μέλη ανάλογα με τις τιμές έντασης των γειτονικών pixel .

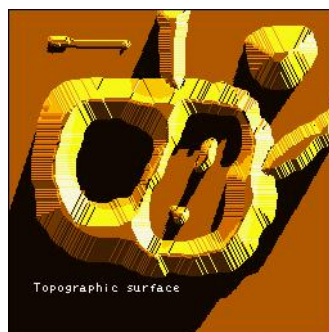
Εδώ ανήκουν δύο γνωστοί αλγόριθμοι **Watershed** (Beucher and Meyer, Vincent and Soille) και **Seeded region Growing (SRG)** (Adams and Bischof). Σε αυτούς τους αλγόριθμους επιλέγονται κάποια σημεία μέσα στην εικόνα τα οποία είναι τα **σημεία έναρξης (seeds)**. Από τα κέντρα αυτά αρχίζουν και δημιουργούνται οι διάφορες περιοχές της εικόνας.

Το **μειονέκτημα** αυτών των μεθόδων είναι με κάποιο κριτήριο πρέπει να επιλεχθούν αυτά τα κέντρα καθώς και ο αριθμός τους. Στην περίπτωση των Microarrays όμως ξέρουμε πόσα είναι αυτά τα κέντρα δηλαδή πόσες είναι οι διακριτές περιοχές της εικόνας ή αλλιώς πόσες είναι η κατηγορίες στις οποίες μπορεί να ανήκει κάθε pixel. Σε τελική φάση θέλουμε να καταλήξουμε σε δύο κατηγορίες ή περιοχές το **background** και το **spot**. Τέλος οι θέσεις των κέντρων (seeds) έχουν ήδη καθοριστεί από την διαδικασία του gridding που προηγείται και που υπολογίζει τα κέντρα των spots.

Κάποιοι άλλοι αλγόριθμοι όπως **Mann Whitney test**, **SRM (statistical region Merging)** , ξεκινάνε με σύνολα από pixel και όχι σημεία έναρξης και εφαρμόζοντας στατιστικές μεθόδους αποφασίζουν για την αναταξινόμηση κάποιων pixel καταλήγοντας σε ένα διαμερισμό της εικόνας όπου χωρίζει τα pixels σε κατηγορίες ανάλογα με τα κοινά χαρακτηριστικά τους.

2.2.1 Watershed

Ο αλγόριθμος watershed σε γενικές γραμμές αρχίζει θεωρώντας ότι τα Pixel με την μεγαλύτερη ένταση ανήκουν στο foreground και σε κάθε βήμα προσθέτει ομάδες από Pixels που έχουν την επόμενη μεγαλύτερη τιμή έντασης (flooding) μέχρις ότου να φτάσει στα σύνορα του spot σύμφωνα με κάποια



Εικ. 2.2.1. Τοπογραφική Επιφάνεια. Μια αναπαράσταση της εικόνας με βάση τις τιμές έντασης κάθε σημείου.

κριτήρια.

Η μέθοδος υδρορροής ή της πλημμύρας (flooding) λειτουργεί ως εξής. Για την κατάτμηση (segmentation) της εικόνας χρησιμοποιούμε την τοπογραφική επιφάνεια (topographic surface) της εικόνας. Η τοπογραφική επιφάνεια της εικόνας περιγράφεται από μια συνάρτηση $f(x) : Z^2 \rightarrow Z$ όπου το σημείο x είναι το σημείο της εικόνας στην θέση (i, j) και $f(x)$ η τιμή της έντασης στην κλίμακα του γκρι που έχει η εικόνα στο σημείο x . Μια τέτοια επιφάνεια φαίνεται στην παρακάτω εικόνα.

Στο επόμενο βήμα βρίσκουμε τα τοπικά ελάχιστα και μέγιστα της εικόνας. Ο αλγόριθμος αρχίζει δημιουργώντας δύο σύνολα pixels αυτά που ανήκουν σε λεκάνες ύδατος (catchment basins) και αυτά που ανήκουν σε υδροκριτική γραμμή (watershed lines).

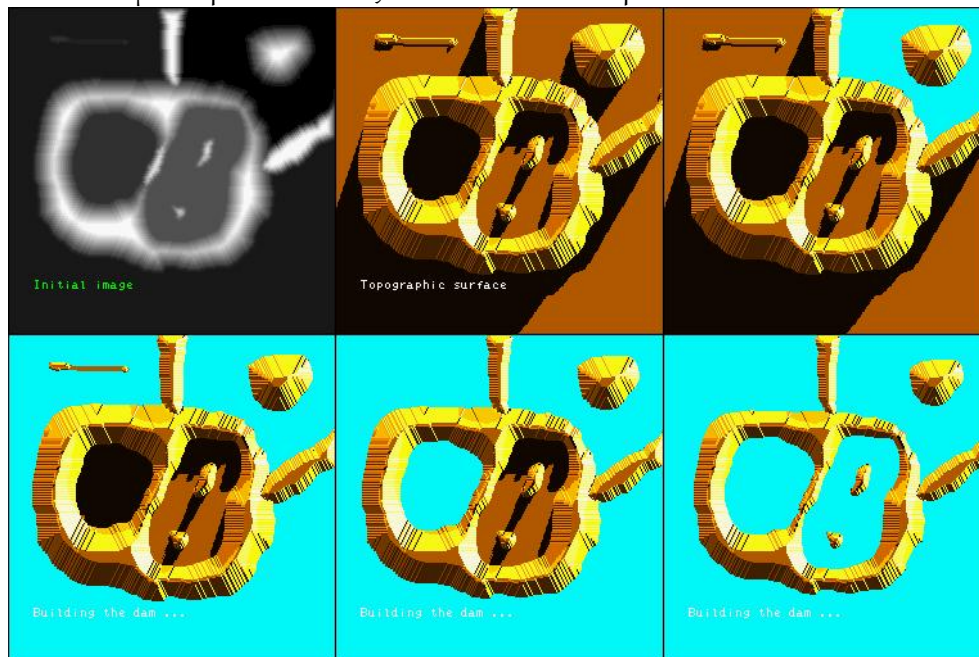
Αρχικά το σύνολο των pixels αποτελούν γραμμές υδρορροής είναι άδειο και το σύνολο των pixels που ανήκουν στην πλημμύρα είναι ίσο με όλα τα pixels που έχουν ένταση ίση με το ολικό ελάχιστο της τοπογραφικής επιφάνειας. Δηλαδή η πλημμύρα βρίσκεται σε επίπεδο $i = \text{overall min}$, $\text{flood} = i$.

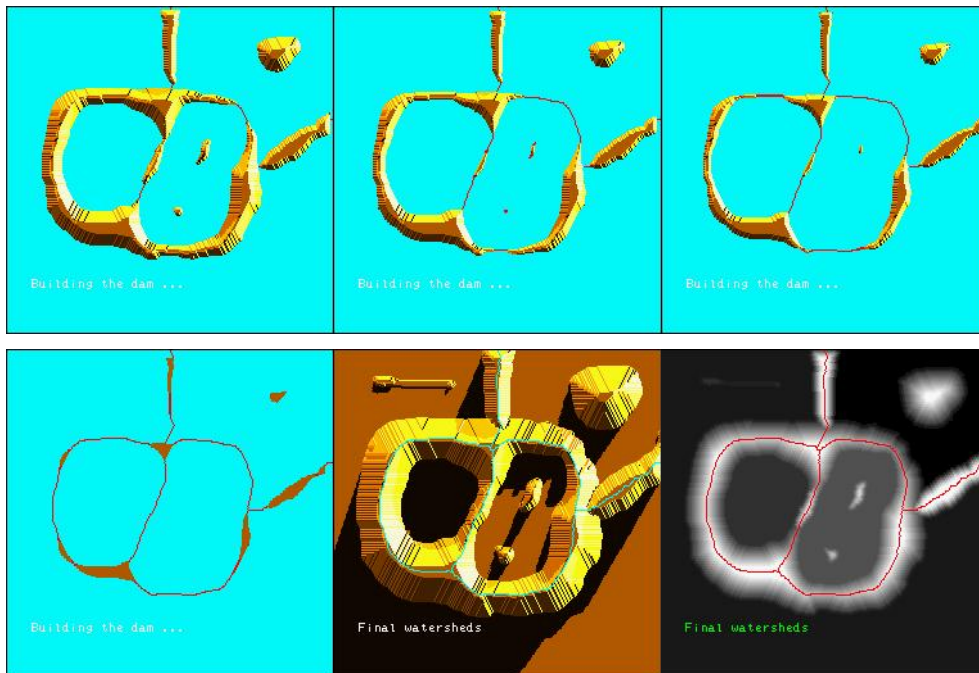
Στην συνέχεια σε κάθε επόμενο βήμα το επίπεδο της πλημμύρας αυξάνεται στην επόμενη μεγαλύτερη τιμή έντασης $\text{flood} = i+1$ και προστίθενται νέα pixels στο σύνολο της πλημμύρας, αυτά που έχουν ένταση ίση με $i+1$.

Έτσι δημιουργούνται λεκάνες «ύδατος». Εάν σε κάποια αύξηση του επιπέδου της πλημμύρας κάποιες λεκάνες ύδατος ενωθούν τότε δημιουργείται ένας υδατοφράχτης (dam) και τα pixels του φράγματος προστίθενται στο σύνολο των υδροκριτικών γραμμών. Ο αλγόριθμος τερματίζει όταν έχουν ταξινομηθεί όλα τα pixels της εικόνας.

Εάν αντί για την grey scale εικόνα χρησιμοποιηθεί η παράγωγος της τότε θεωρητικά οι λεκάνες ύδατος αντιστοιχούν στις ομογενείς grey level περιοχές της εικόνας.

Ακολουθεί μια σειρά από εικόνες που οπτικοποιούν την διαδικασία:





Μειονέκτημα

Υπερ-κατάτμηση της εικόνας (over splitting). Προκύπτουν πάνω από δύο κατηγορίες-περιοχές pixel.

2.2.2 Seeded Region Growing (SRG)

Γενικός Αλγόριθμος.

Ο αλγόριθμος SRG δέχεται αρχικά ως είσοδο έναν αριθμό από σημεία έναρξης (**seeds**) της διαδικασίας επαύξεσης περιοχής (region growing process). Στην απλή περίπτωση τα σημεία έναρξης μπορούν να είναι πράγματι απλά σημεία – pixels. Σε μια πιο σύνθετη εκδοχή τα σημεία αυτά μπορούν να αποτελούν σύνολα από Pixels τα οποία δεν είναι κατ' ανάγκη γειτονικά.

Έχοντας καθοριστεί τα σημεία έναρξης ο αλγόριθμος επαυξάνει τα δύο σύνολα του background και του foreground ταυτόχρονα μέχρι να ταξινομηθούν όλα τα Pixels.

Σε κάθε στάδιο του αλγορίθμου όλα τα Pixels τα οποία δεν έχουν ακόμα ταξινομηθεί και έχουν τουλάχιστον ένα γείτονα ο οποίος ανήκει ήδη σε κάποια κατηγορία είναι υποψήφια προς ταξινόμηση. Απ' όλα τα υποψήφια pixels δηλαδή τα γειτονικά στις περιοχές ταξινόμησης ο αλγόριθμος επιλέγει αυτό που η τιμή της έντασης (grey level) είναι πιο κοντά στην μέση τιμή έντασης της περιοχής ταξινόμησης.

Για το background για παράδειγμα έχουμε:

$$NewBackgroundPixel = \arg \min_k \{ |I_{cp,k} - I_{avg,b}| \}$$

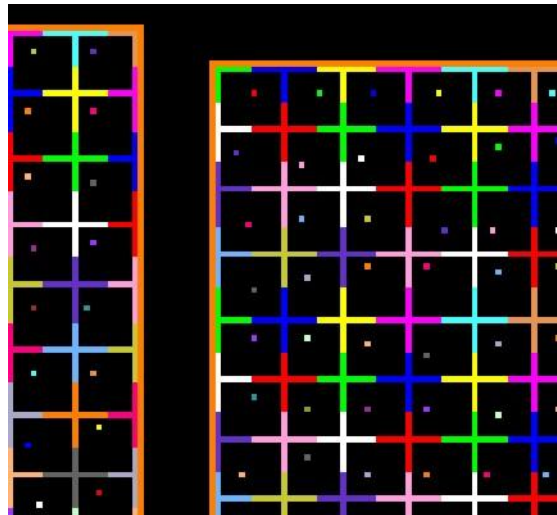
$I_{cp,k}$: η ένταση του υποψήφιου Pixel k

$I_{avg,b}$: η μέση τιμή έντασης του background

Ο αλγόριθμος επαναλαμβάνεται μέχρι να ταξινομηθούν όλα τα Pixels και για καλύτερη απόδοση μπορούν να χρησιμοποιηθούν λίστες των pixels.

Περίπτωση Μικροσυστοιχιών.

Στην περίπτωση των Microarrays έχουμε δύο περιοχές ταξινόμησης (seeded regions) το foreground και το background και ως σημεία έναρξης μπορούν να επιλεγθούν τα κέντρα των spots στην πρώτη περίπτωση και τα σημεία τομής του grid στην δεύτερη. Τα σημεία αυτά ήδη έχουν καθοριστεί από την διαδικασία του gridding η οποία προηγείται.



Εικ 2.2.2. Τα σημεία τομής του πλέγματος θεωρούνται σημεία έναρξης για το background.

Ένας άλλος τρόπος επιλογής του σημείου έναρξης της περιοχή του foreground είναι να επιλεγθεί το Pixel με την μεγαλύτερη ένταση σε ένα κομμάτι της εικόνας που βρίσκεται στο κέντρο. Αντίστοιχα για το background ως σημεία έναρξης μπορούν να είναι όλα τα σημεία τομής του πλέγματος ή αλλιώς για την εικόνα ενός κελιού του πλέγματος να επιλέξουμε ως σημεία έναρξης της περιοχής του background τα τέσσερα γωνιακά pixels. Αποφεύγεται όμως η επιλογή σημείου έναρξης που περιέχει πολλά pixels για το background πρώτον γιατί η ένταση του background έχει μεγάλη τοπική μεταβλητότητα και δεύτερον γιατί θέλουμε να υπολογίσουμε τοπικά την τιμή της έντασης προκειμένου να ταξινομήσουμε ένα νέο pixel.

2.2.3 Mann Whitney test

Συνοπτική περιγραφή και εφαρμογή του αλγορίθμου στις μικροσυστοιχίες

Ο αλγόριθμος υπολογίζει επαναληπτικά το σημείο απόφασης (threshold) ανάμεσα στο foreground και το background χρησιμοποιώντας το Mann-Whitney test. Το Mann-Whitney test είναι ένα μη παραμετρικό στατιστικό τεστ για την **εκτίμηση της στατιστικής σημασίας της διαφοράς των δύο κατανομών**.

Αρχικά ορίζεται μια κυκλική μάσκα η οποία περιέχει όλα τα pixel που πιθανά ανήκουν στο foreground διαχωρίζοντας τα έτσι από το background. Στην συνέχεια με την βοήθεια του Mann-Whitney test συγκρίνονται: ένα σύνολο από τυχαία επιλεγμένα pixel από το background και ένα σύνολο από προκαθορισμένο αριθμό pixels τα οποία βρίσκονται μέσα στην μάσκα και έχουν την χαμηλότερη ένταση. Αν η διαφορά ανάμεσα στα δύο σύνολα δεν είναι σημαντική ο αλγόριθμος αφαιρεί από το δεύτερο σύνολο ένα προκαθορισμένο αριθμό από pixel και προσθέτει νέα pixel πάλι μέσα από την μάσκα. Ο αλγόριθμος τερματίζει όταν τα δύο σύνολα διαφέρουν σημαντικά το ένα από το άλλο. Τα pixel που έχουν επιλεγθεί από την μάσκα είναι αυτά που ανήκουν στο spot-foreground και τα υπόλοιπα ανήκουν στο background.

Μαθηματική περιγραφή

Το Mann Whitney test γνωστό επίσης και ως Mann-Whitney-Wilcoxon (MWW) είναι ένα μη

παραμετρικό τεστ για να εκτιμήσει κανείς αν δύο σύνολα παρατηρήσεων (δείγματα) προέρχονται από την ίδια κατανομή. Η απαίτηση του τεστ είναι ότι πρέπει τα δύο δείγματα να είναι ανεξάρτητα και οι παρατηρήσεις να είναι συνεχείς ή τακτικές.

Το τεστ συμπεριλαμβάνει τον υπολογισμό μιας στατιστικής ποσότητας που ονομάζεται U που η κατανομή της είναι γνωστή. Για την περίπτωση που έχουμε λίγα δείγματα οι τιμές τις κατανομής μπορούν να τοποθετούν σε ένα πίνακα. Στην περίπτωση μεγάλης ποσότητας δειγμάτων γίνεται πολύ καλή προσέγγιση χρησιμοποιώντας την κανονική κατανομή.

Υπάρχουν δύο τρόποι για να γίνει το U test:

Για μικρό αριθμό δειγμάτων:

Ονόμασε το δείγμα με τις μικρότερες τιμές παρατηρήσεων ως δείγμα1 και το άλλο δείγμα ως δείγμα2.

Για κάθε παρατήρηση του δείγματος1 βρίσκουμε τον αριθμό των παρατηρήσεων από το δείγμα2 που έχουν μικρότερη τιμή απ' αυτό.

Το U είναι ο παραπάνω αριθμός.

Για μεγαλύτερο αριθμό δειγμάτων:

Ταξινόμηση της ένωσης των παρατηρήσεων των δύο δειγμάτων σε μια αύξουσα σειρά σύμφωνα με την τιμή τους.

Εύρεση του αθροίσματος των τιμών των παρατηρήσεων για το δείγμα1. Το άθροισμα των τιμών όλων των παρατηρήσεων είναι $N(N+1)/2$ όπου N ο συνολικός αριθμός των δειγμάτων.

Συνεπώς το άθροισμα των τιμών των παρατηρήσεων.

Το U υπολογίζεται ως εξής:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \text{ όπου}$$

n_1 : ο αριθμός των παρατηρήσεων του δείγματος1 και

R_1 : το άθροισμα των τιμών των παρατηρήσεων.

Ομοίως:

$$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

$$U_1 + U_2 = R_1 - \frac{n_1(n_1 + 1)}{2} + R_2 - \frac{n_2(n_2 + 1)}{2}$$

$$\text{Εφόσον } R_1 + R_2 = \frac{N(N + 1)}{2} \quad N = n_1 + n_2$$

$$U_1 + U_2 = n_1 n_2$$

Matarray (matlab toolbox)

Ένα toolbox που υπάρχει στην Matlab (Wang et al.2001). Ο αλγόριθμος συνδυάζει τοπικά δεδομένα και πληροφορίες για την ένταση. Η μέθοδος αυτή μοιάζει πολύ με τον αλγόριθμο Mann Whitney, αρχικά μια κυκλική μάσκα στόχος για το αρχικό διαχωρισμό του spot από το background. Τα pixels μέσα στην μάσκα που έχουν μεγαλύτερη ένταση από $\mu_{bkg} + 2\sigma_{bkg}$ (όπου μ_{bkg} και σ_{bkg} είναι η τοπική μέση τιμή και τυπική απόκλιση για το background) θεωρούνται ως υποψήφια Pixels του spot και χρησιμοποιούνται για τον υπολογισμό των νέων κέντρων. Επαναληπτικά υπολογίζονται τα νέα κέντρα και τα Pixels τα οποία ανήκουν στο foreground και στο background. Το τελικό αποτέλεσμα κατάτμησης προκύπτει μετά από το πολύ δύο επαναλήψεις.

2.2.4 Statistical Region Merging

Έχουμε μια εικόνα την οποία την χωρίζουμε σε στοιχειώδη τμήματα (elementary regions)

δηλαδή σε ομάδες από Pixels που μοιάζουν μεταξύ τους και αποτελούν κάποιο ξεχωριστό αντικείμενο της εικόνας. Επιλέγεται κάποιο μέτρο ομοιότητας και με βάση αυτό τα Pixels οποία μοιάζουν μεταξύ τους αποτελούν μια ξεχωριστή ομάδα.

Στην συνέχεια διαδικασία του συγκερασμού (region merging) κατά την οποία σύμφωνα με στατιστικές μεθόδους τα στοιχειώδη τμήματα της εικόνας ενώνονται σε μεγαλύτερα τμήματα.

Περιγραφή αλγορίθμου

Η κατάτμηση είναι η διαδικασία της τμηματοποίησης μιας εικόνας σε ασύνδετες και ομογενείς περιοχές. Ένας πιο τυπικός ορισμός μπορεί να δοθεί με τον ακόλουθο τρόπο: έστω I μια εικόνα και έστω H ένα συγκεκριμένο κατηγορήμα ομοιογένειας. Η τμηματοποίηση της I είναι ένας διαμερισμός P της I σε ένα σύνολο από N περιοχές $R_1, R_2, \dots, R_N$ τέτοιες ώστε:

$$\bigcup_{n=1}^N R_n = I \text{ με } R_n \cap R_m = \emptyset, \quad n \neq m$$

$$H(R_n) = \text{true} \quad \forall n$$

$$H(R_n \cup R_m) = \text{false} \quad \forall R_n \text{ and } R_m \text{ adjacent}$$

Ο αλγόριθμος Statistical Region Merging που ανήκει στην οικογένεια του region growing τεχνικών με στατιστικό έλεγχο για ένωση περιοχών. Το I είναι μια εικόνα με $|I|$ pixels που το κάθε ένα περιέχει τρεις τιμές (R,G,B) που ανήκουν στο σύνολο $\{1,2,\dots,g\}$. Το μοντέλο εξετάζει την εικόνα I σαν μια παρατήρηση ενός εντελώς άγνωστου σκηνικού Γ στο οποίο τα pixels παριστάνονται από μια οικογένεια από κατανομές από τις οποίες το επίπεδο κάθε χρώματος δειγματοληπτείται. Συγκεκριμένα η τιμή κάθε χρώματος για κάθε pixel του Γ περιγράφεται από ένα σύνολο από Q ανεξάρτητες τυχαίες μεταβλητές με τιμές στο $[0,g/Q]$. Στο Γ οι βέλτιστες περιοχές ικανοποιούν τις παρακάτω ιδιότητες ομοιογένειας.

- i. Μέσα σε οποιαδήποτε στατιστική περιοχή και για οποιοδήποτε χρωματικό κανάλι στατιστικά τα pixels έχουν την ίδια μέση τιμή για αυτό το χρωματικό κανάλι.
- ii. Η μέση τιμή των γειτονικών περιοχών είναι διαφορετική για τουλάχιστον ένα χρωματικό κανάλι.
- iii. Από αυτό το μοντέλο ο Nielsen και ο Nock πετυχαίνουν το παρακάτω κατηγορήμα συγχώνευσης.

$$P(R, R') = \begin{cases} \text{true} & \alpha \forall \alpha \in \{R, G, B\}, \left| \overline{R'_\alpha} - \overline{R_\alpha} \right| \leq b(R) + b(R') \\ \text{false} & \text{otherwise} \end{cases}$$

$$b(R) = g \sqrt{\frac{1}{2Q|R|} \left(\ln \frac{|R|}{\delta} \right)}$$

$\overline{R_\alpha}$: παρατηρούμενη μέση τιμή για το χρώμα α στην περιοχή R

$R_{|l|}$: το σύνολο των περιοχών με l pixels.

Η σειρά με την οποία γίνονται οι έλεγχοι συγχώνευσης ακολουθεί μια απλή σταθερά A .

Όταν γίνεται ένας οποιοσδήποτε έλεγχος μεταξύ δύο περιοχών, αυτό σημαίνει ότι όλοι οι έλεγχοι μέσα σε κάθε περιοχή έχουν προηγουμένως γίνει.

Στα πειράματα το A προσεγγίζεται από ένα απλό αλγόριθμο που βασίζεται στην παράγωγο των γειτονικών pixels. Συγκεκριμένα:

$$f(p, p') = \max_{a \in R, G, B} f_a(p, p')$$

Μειονεκτήματα

Υπερσυγκερασμός (Overmerging)

Κάποια από τα μεγαλύτερα σύνολα από Pixels στα οποία καταλήγει ο αλγόριθμος συγχώνευσης περιέχουν πάνω από μια πραγματικές περιοχές της εικόνας.

Πλεονεκτήματα

Γραμμική Χρονική Πολυπλοκότητα.

Εφαρμόζεται σε όλα τα συστήματα αναπαράστασης των εικόνων (RGB, HSI, κλπ).

2.2.5 Αλγόριθμος Τυχαίου πεδίου Markov-(The Markov random field)

Ο αλγόριθμος τυχαίου πεδίου Markov εφαρμόστηκε πάνω στα Microarrays για πρώτη φορά από (Demirkaya 2005) ^[32]. Η μέθοδος αυτή μοντελοποιεί τις εντάσεις του background και του foreground ως εκθετικές κατανομές. Σε συνδυασμό με την πληροφορία που προσφέρουν οι τιμές των εντάσεων των pixel ο αλγόριθμος εκμεταλλεύεται και την τοπική πληροφορία μοντελοποιώντας τα γειτονικά pixel με MRF.

Ως βάση του αλγορίθμου είναι μια αρχική ταξινόμηση των pixels σε foreground και background και η επιλογή της επηρεάζει το τελικό αποτέλεσμα του αλγορίθμου.

3 ΚΑΤΑΤΜΗΣΗ ΙΣΤΟΓΡΑΜΜΑΤΟΣ

Η τιμή έκφρασης που εξάγουμε για κάθε spot είναι ανάλογη με την τιμή της έντασης των pixel που ανήκουν στο spot. Συνεπώς τα pixel που βρίσκονται στο foreground πρέπει να έχουν παρόμοιες εντάσεις και τα pixels που ανήκουν στο background πρέπει να έχουν επίσης παρόμοιες εντάσεις. Επιπλέον τα pixels τα οποία ανήκουν σε τυχόν μουντζούρες (artifacts) τείνουν να έχουν διαφορετικές εντάσεις και από το foreground και από το background. Η βασική ιδέα των μεθόδων κατάτμησης που βασίζονται σε ιστόγραμμα είναι η μοντελοποίηση των κατανομών από τις οποίες αποτελείται το ιστόγραμμα της έντασης και η ομαδοποίηση των pixel άλογα με τις τιμές εντάσεων.

Οι μέθοδοι που χρησιμοποιούν ιστογράμματα εφαρμόζουν **φίλτρα** ή **thresholds** στην κατανομή της έντασης και της απόχρωσης. Τα όρια αυτά χωρίζουν το σύνολο των Pixel σε κατηγορίες ανάλογα με τα κοινά τους χαρακτηριστικά τους έτσι ώστε οι κατηγορίες αυτές να διαχωρίζονται όσο τον δυνατόν καλύτερα.

Σ' αυτήν την κατηγορία ανήκουν επίσης οι αλγόριθμοι που κάνουν **ταξινόμηση (unsupervised classification)** των pixel καθώς επίσης και οι μέθοδοι που κάνουν **εκτίμηση των κατανομών** και εύρεση του σημείου απόφασης.

Πιο κοινές είναι οι μέθοδοι όπου οι συνολική κατανομή προσεγγίζεται ως άθροισμα του ανάλογου αριθμού από Gaussian κατανομές (συνήθως όσος και ο αριθμός των κατηγοριών). Στόχος είναι η εύρεση των παραμέτρων κάθε κατανομής (μ , Σ). Αρχικά επιλέγεται μια συνάρτηση αξιολόγησης του σφάλματος προσέγγισης και με την ελαχιστοποίηση αυτής της συνάρτησης υπολογίζονται οι τιμές των παραμέτρων. Τέλος αυτό που μένει είναι η εύρεση του σημείου ή της ευθείας ή της επιφάνειας απόφασης. Στην περίπτωση των μικροσυστοιχιών συνήθως έχουμε ένα χαρακτηριστικό την ένταση συνεπώς έχουμε σημεία απόφασης.

3.1 Κατάτμηση βάσει έντασης

Έχοντας την εικόνα αποθηκευμένη στο σύστημα HSI προχωράμε σε κατάτμηση του κάθε κελιού του πλέγματος χρησιμοποιώντας μόνο ένα χαρακτηριστικό, την ένταση.

Θεωρούμε ότι εφόσον έχουμε δύο κατηγορίες η κατανομή της έντασης για ένα κελί του πλέγματος θα έχει δύο μέγιστα, δηλαδή θα μπορεί να προσεγγιστεί ως άθροισμα δύο Gaussian κατανομών. Δοκιμάζουμε να διαχωρίσουμε το spot από το background βρίσκοντας το σημείο το οποίο διαχωρίζει τις δύο κατηγορίες με το μικρότερο σφάλμα. Οι αλγόριθμοι που ακολουθούν βρίσκουν είτε απευθείας ένα Threshold, δηλαδή ένα σημείο απόφασης για την ταξινόμηση δειγμάτων μας (Otsu) ή τις παραμέτρους των δύο κατανομών και στην συνέχεια το σημείο απόφασης είναι το σημείο τομής των δύο κατανομών.

3.1.1 Κατάτμηση με Μοντελοποίηση

Εισάχθηκε πρώτη φορά από (Li et al. 2005) ^[14] και είναι μια μέθοδος κατάτμησης δύο βημάτων. Τα βασικά βήματα της μεθόδου είναι η ταξινόμηση των pixels βασιζόμενη σε μοντελοποίηση και τοπική εξαγωγή των συνδεδεμένων στοιχείων. Η ταξινόμηση με μοντελοποίηση υλοποιεί μια αρχική ταξινόμηση σε τρεις το πολύ κατηγορίες το **spot**, το spot που περιέχει background ή αλλιώς **αραβούργημα (artifact)** και το καθαρό **background**.

Ο αλγόριθμος αυτός βασίζεται σε μίξη Gaussian μοντέλων και ο αριθμός των κατηγοριών καθορίζεται από τα δείγματα χρησιμοποιώντας το Bayesian Information Criterion (BIC). Η αφαίρεση των τοπικά συνδεδεμένων στοιχείων χρησιμοποιείται για και να αποκλειστούν μικρές ασύνδετες κατηγορίες που υπολογίζεται ότι είναι αραβουργήματα (artifacts).

Παρόλο που υπάρχει ξεχωριστή κατηγορία για τα pixels που ανήκουν στο foreground χρησιμοποιούμε το foreground μαζί με το foreground with artifacts για να υποδηλώσουμε το foreground. Χωρίς την δεύτερη κατηγορία ο αλγόριθμος χάνει τιμές στον υπολογισμό του αποτελέσματος.

Στην μέθοδο αυτή θεωρούμε ότι τα δείγματα προέρχονται από μια μίξη κατανομών

$$f(x) = \sum_{k=1}^K p_k f_k(x)$$

Όπου p_k είναι η αναλογίες μίξης των κατανομών ($0 < p_k < 1$ για $k=1,2,\dots,K$ και $\sum_k p_k = 1$) και f_k είναι

η συνάρτηση πυκνότητας πιθανότητας των παρατηρήσεων στην κλάση k .

Στην περίπτωση που η ολική κατανομή προσεγγιστεί ως μίξη Gaussian κατανομών η κατανομή κάθε κλάσης μοντελοποιείται από μια πολυδιάστατη κανονική κατανομή με μέση τιμή μ_k και πίνακα συμεταβλητότητας και συνάρτηση πυκνότητας πιθανότητας ίση με:

$$f(x_i; \mu_k, \Sigma_k) = \frac{1}{\sqrt{|2\pi\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right)$$

Η πιθανότητα των δειγμάτων που αποτελούνται από n παρατηρήσεις υποθέτοντας ότι έχουμε μίξη k Gaussian κατανομών.

3.1.1.1 Αλγόριθμος Otsu

Για να μπορέσουμε να διαχωρίσουμε το spot από το background αρκεί να χωρίσουμε την εικόνα σε δύο ομάδες Pixels η μία θα αποτελείται από τα Pixels του background και η άλλη ομάδα από τα pixels του spot.

Για να ταξινομήσουμε τα pixels στην μια ή στην άλλη κατηγορία αρκεί να ορίσουμε ένα όριο διαχωρισμού, μια τιμή του Intensity κάτω από την οποία θα θεωρήσουμε ότι τα pixels ανήκουν στην πρώτη ομάδα (background) και πάνω από αυτήν την τιμή ότι ανήκουν στο spot.

Σύμφωνα με τον αλγόριθμο Otsu στόχος μας είναι να **ελαχιστοποιήσει** την παρακάτω διασπορά η οποία είναι αποτελεί μέτρο απόστασης των δύο κλάσεων:

$$\sigma_{Within}^2(T) = n_B(T)\sigma_B^2(T) + n_O(T)\sigma_O^2(T)$$

Όπου:

- T : το παρόν threshold
- $n_B(T) = \sum_{i=0}^{T-1} p(i)$: το άθροισμα των πιθανοτήτων των εντάσεων των pixels που ανήκουν στο background συμφωνά με το παρόν Threshold.
- $n_O(T) = \sum_{i=T}^{N-1} p(i)$: το άθροισμα των πιθανοτήτων των εντάσεων των pixels που ανήκουν στο foreground συμφωνά με το παρόν Threshold.
- $\sigma_B^2(T)$: Η διακύμανση των Pixels που ταξινομούνται στο background σύμφωνα με το παρόν threshold T
- $\sigma_O^2(T)$: Η διακύμανση των Pixels που ταξινομούνται στο foreground σύμφωνα με το παρόν threshold T .

Είναι δηλαδή το άθροισμα των διακυμάνσεων των δύο ομάδων pixels επί κάποια βάρη.

Μια απλοποίηση του παραπάνω τύπου είναι:

$$\sigma_{Between}^2(T) = n_B(T)n_O(T)(\mu_O(T) - \mu_B(T))^2 \quad (1)$$

Όπου:

- $\mu_B(T)$: Η μέση τιμή του **Intensity** των Pixels που ταξινομούνται στο background σύμφωνα με το παρόν threshold T.
- $\mu_O(T)$: Η μέση τιμή του **Intensity** που ταξινομούνται στο foreground σύμφωνα με το παρόν threshold T.

Η διακύμανση $\sigma_{Between}^2$ στην ουσία περιγράφει την απόσταση των δύο κλάσεων και στόχος μας είναι να την μεγιστοποιήσουμε, δηλαδή να βρούμε **το T το οποίο μεγιστοποιεί το τύπο (1)**. Ο αλγόριθμος αυτός παρόλο που επιστρέφει μια τιμή Threshold στην ουσία πραγματοποιεί μια ταξινόμηση των Pixels η οποία οδηγεί σε μια εκτίμηση των μέσων τιμών των δύο κατηγοριών και με την βοήθεια του παραπάνω τύπου υπολογίζει την απόσταση των δύο κατηγοριών.

Ως αρχική τιμή του Threshold θέτουμε την ελάχιστη τιμή Intensity της εικόνας του spot που έχουμε και αυξάνουμε σταδιακά την τιμή του Threshold έως ότου φτάσουμε σε ένα σημείο όπου παραπάνω παράσταση πάρει την μέγιστη της τιμή.

Για πιο γρήγορη εκτέλεση του αλγορίθμου ως επόμενη μεγαλύτερη τιμή του Threshold (βήμα αύξησης) θέτουμε την επόμενη μεγαλύτερη τιμή έντασης που συναντάμε στην εικόνα.

Η κατεύθυνση με την οποία εξετάζουμε τις τιμές της έντασης δεν έχει σημασία. Συνεπώς μπορούμε να ξεκινώντας από την μέγιστη τιμή έντασης που συναντάμε στην εικόνα και μειώνοντας την τιμή Threshold να καταλήξουμε στο ίδιο αποτέλεσμα.

3.1.1.2 Επαναληπτικός Otsu

Στον απλό αλγόριθμο Otsu κάνουμε την παραδοχή ότι έχουμε μόνο δύο κατηγορίες (background και foreground) δηλαδή ότι η συνολική κατανομή της έντασης αποτελεί γραμμικό συνδυασμό δύο μόνο κατανομών. Εισάγουμε κατ' αυτόν τον τρόπο ένα σφάλμα στην εκτίμηση των δύο κατηγοριών μας καθώς στην πραγματικότητα έχουμε περισσότερες από δύο κατηγορίες, το spot από μόνο του μπορεί να χωριστεί σε δύο κατηγορίες τα περιφερειακά και εσωτερικά pixel (τα πρώτα έχουν μεγαλύτερες τιμές εντάσεων από τα δεύτερα). Λόγω αυτής της προσέγγισης ο απλός αλγόριθμος Otsu συμπεριλαμβάνει στο foreground μονό τα περιφερειακά pixel του spot.

Ένας τρόπος όπου δεν χρειάζεται εκτιμήσουμε το πλήθος των κατηγοριών και ταυτόχρονα να επιλύσουμε το πρόβλημα αυτό είναι η **επαναληπτική εφαρμογή Otsu**. Σε κάθε επανάληψη του αλγορίθμου χωρίζουμε σε δύο ομάδες τα pixel που ανήκουν στο background όπως αυτό προέκυψε από την προηγούμενη επανάληψη. Κατ' αυτόν τον τρόπο συμπεριλαμβάνουμε σταδιακά μια νέα ομάδα εντάσεων στο spot μας μέχρι συγκεντρώσουμε μια ικανοποιητική ποσότητα από pixels (γύρω στο 50% της συνολικής εικόνας ενός κελιού).

Μία παραλλαγή της επαναληπτικής του παραπάνω αλγορίθμου είναι σε κάθε επανάληψη να πραγματοποιούμε κατάτμηση στο σύνολο των pixel της εικόνας αφού προηγουμένως θέσουμε την τιμή της έντασης των pixels που έχουν ταξινομηθεί foreground σύμφωνα με την προηγούμενη επανάληψη, ίση με την μέγιστη τιμή έντασης του background.

Σε κάθε βήμα επανάληψης προκύπτουν δύο ομάδες pixels, το background (μικρές τιμές έντασης) και το foreground (μεγάλες τιμές έντασης). Ο αλγόριθμος τερματίζει όταν το ποσοστό των Pixels που

ταξινομήθηκαν στο foreground είναι μεγαλύτερο από το 38% του μεγέθους του αντίστοιχου κελιού. Το ποσοστό αυτό προέκυψε πειραματικά.

```
[foreground, background, Threshold]=iter_Otsu(Ένταση)
[foreground, background, Threshold]=Otsu(Ένταση)
while (άθροισμα(foreground)<.38*ΜέγεθοςΕικόνας)
    Ένταση(foreground)= max(Ένταση(background))
    [foreground, background, Threshold]= Otsu (Ένταση)
end
```

Αλγ 3.1.1. Επαναληπτικός Αλγόριθμος Otsu

3.1.1.3 Αλγόριθμος Expectation Maximization (EM)

Γενικά

Ο EM είναι ένας επαναληπτικός αλγόριθμος που ξεκινάει με μια αρχική εκτίμηση των παραμέτρων θ και επαναληπτικά μεταβάλλει το θ ώστε να αυξήσει την πιθανότητα των παρατηρούμενων δεδομένων. Δουλεύει καλύτερα με περιπτώσεις όπου τα δεδομένα είναι ελλιπή ή μπορεί να θεωρηθούν ως ελλιπή. Ο EM συνήθως εφαρμόζεται σε κατανομές όπου αποτελούν μίξη άλλων κατανομών (π.χ Gaussians).

Βασική Ιδέα και Βήματα

Αν το x ήταν γνωστό τότε θα μπορούσαμε να χρησιμοποιήσουμε τον τύπο του Maximum Likelihood (ML) για να υπολογίσουμε το θ :

$$\arg \max_{\theta} (\ln p(D_x / \theta)) \quad (3.1.1)$$

Συνεπώς εάν μεγιστοποιήσουμε την πιθανότητα $p(x/\theta)$ δεδομένου ότι έχουμε τα δεδομένα x και την παρούσα εκτίμηση το θ .

- **Αρχικοποίηση:** Δίνουμε αρχική τιμή στο θ^0
- **Εκτίμηση:** Χρησιμοποιώντας τη παρούσα εκτίμηση των παραμέτρων θ και δεδομένου των παρατηρήσεων έχουμε:

$$Q(\theta; \theta') = E_{x_{\text{μην παρατηρήσιμα}}} (\ln p(D_x / \theta) / D_y, \theta')$$

Η μέση τιμή προκύπτει με βάση τις μη παρατηρήσιμες μεταβλητές.

- **Μεγιστοποίηση:** νέα εκτίμηση των παραμέτρων.

$$\theta^{i+1} = \arg \max_{\theta} Q(\theta; \theta')$$

- **Σύγκλιση:** Αν $\|\theta^{i+1} - \theta^i\| < \varepsilon$ *τερματισμός αλλιώς* πηγαίνει στο βήμα της εκτίμησης.

Μοντελοποίηση – Μίγμα κατανομών

Όπως είπαμε και σε προηγούμενες ενότητες η κατανομή των δειγμάτων μας αποτελεί άθροισμα επιμέρους κατανομών. Δηλαδή υπάρχουν υπο-μοντέλα κάθε ένα από τα οποία έχει την δική του κατανομή η οποία περιγράφει πως παράγονται τα δεδομένα όταν το μοντέλο αυτό είναι ενεργό. Κάθε υπό – μοντέλο έχει επίσης την δική του πιθανότητα ενεργοποίησης **πκ.**

Ένα τέτοιο μοντέλο που αποτελεί μίγμα κατανομών ορίζεται ως ένας γραμμικός συνδυασμός K κατανομών όπου κάθε κατανομή έχει την δική της συνάρτηση πυκνότητας πιθανότητας $p(x/\theta_k)$ δεδομένου τις δικές τις παραμέτρους $\theta_k \rightarrow (\mu_k, \Sigma_k)$.

$$p(x/\theta) = \sum_{k=1}^K p(x/\theta_k) \pi_k \quad (3.1.2)$$

Οι πιθανότητες ενεργοποίησης αθροίζουν στην μονάδα.

$$\sum_{k=1}^K \pi_k = 1 \quad (3.1.3)$$

Η διαδικασία εκτίμησης του μοντέλου σύμφωνα με τα δείγματα D_x αποτελεί διαδικασία προσέγγισης των παραμέτρων που περιγράφουν καλύτερα τα δείγματα.

- Εκτίμηση των παραμέτρων
- Εκτίμηση επιμέρους κατανομών.

Εκτίμηση Παραμέτρων χρησιμοποιώντας τον Maximum Likelihood

Όπως γνωρίζουμε δεδομένου ενός συνόλου δειγμάτων $D=(x_1, x_2, \dots, x_n)$ οι τύποι του Maximum Likelihood δίνουν την τιμή του θ η οποία μεγιστοποιεί την παρακάτω πιθανότητα:

$$p(D/\theta) = \prod_{i=1}^n p(x_i/\theta)$$

Σύμφωνα με τον τύπο 9.2 προκύπτει ο συνολικός τύπος:

$$p(D/\theta) = \prod_{i=1}^n \sum_{k=1}^K p(x_i/\theta_k) \pi_k \quad (3.1.4)$$

Δηλαδή η πιθανότητα ένα δείγμα να ανήκει στην συνολική κατανομή είναι το άθροισμα των πιθανοτήτων που δίνουν οι επιμέρους κατανομές για το συγκεκριμένο δείγμα και η πιθανότητα της συνολικής κατανομής αποτελεί το γινόμενο των πάνω σ' αυτές τις πιθανότητες όλων των δειγμάτων.

Κρυφές Μεταβλητές

Εφόσον δεν γνωρίζουμε κάθε δείγμα σε ποια επιμέρους κατανομή ανήκει έχουμε ελλιπή δεδομένα όσον αφορά την κατηγορία στην οποία ανήκει κάθε δείγμα. Συνεπώς στην περιγραφή των δειγμάτων προσθέτουμε την ελλιπή πληροφορία z_i την οποία ο EM θα προσπαθήσει να εκτιμήσει μαζί με τις παραμέτρους θ .

$$y_i = (x_i, z_i) \text{ όπου}$$

$$z_i = (z_{i1}, z_{i2}, \dots, z_{iK}) \text{ όπου } z_{ik} \text{ είναι η πιθανότητα το } x_i \text{ να ανήκει στην κατηγορία } k$$

Αν γνωρίζαμε ποια επιμέρους κατανομή είναι υπεύθυνη για την παραγωγή κάθε δείγματος τότε θα ήταν απλό να υπολογίσουμε τις βέλτιστες τιμές για το θ σύμφωνα με τον ML. Τα βήματα είναι τα εξής:

- Χρησιμοποιούμε τον EM για να εκτιμήσουμε το z_i για κάθε δείγμα i .
- Σύμφωνα με την παραπάνω εκτίμηση των ελλিপών Πληροφοριών για τις κατηγορίες των δειγμάτων βρίσκουμε τις παραμέτρους θ σύμφωνα με τον ML.
- Χρησιμοποιώντας το νέο θ υπολογίζουμε τα νέα z και επαναλαμβάνουμε.

Τελικά Βήματα Expectation Maximization

Έχοντας κάνει την παραπάνω ανάλυση και έπειτα από πράξεις προέκυψαν οι παρακάτω τύποι που μας δίνουν την εκτίμηση των ελλিপών δεδομένων όσον αφορά σε ποια κατηγορία ανήκει κάθε δείγμα και τον υπολογισμό των παραμέτρων θ που μεγιστοποιούν την πιθανότητα των δειγμάτων.

Στην γενική περίπτωση όπου έχουμε K επιμέρους κατανομές (κατηγορίες) έχουμε:

Βήμα Αρχικοποίησης:

$$\theta_k^0 = (\pi_k^0, \mu_k^0, \Sigma_k^0)$$

Βήμα Εκτίμησης:

$$E(z_{ik})^{t+1} = \frac{p(x_i / \theta_k^t) \pi_k^t}{\sum_{j=1}^K p(x_i / \theta_j^t) \pi_j^t} \quad (3.1.5)$$

Βήμα Μεγιστοποίησης:

$$\pi_k^{t+1} = \frac{1}{n} \sum_{i=1}^n E(z_{ik})^{t+1} \quad (3.1.6)$$

$$\mu_k^{t+1} = \frac{1}{n \pi_k^{t+1}} \sum_{i=1}^n E(z_{ik})^{t+1} x_i \quad (3.1.7)$$

$$\Sigma_k^{t+1} = \frac{1}{n \pi_k^{t+1}} \sum_{i=1}^n E(z_{ik})^{t+1} (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T \quad (9.8)$$

Όσον αφορά το βήμα της αρχικοποίησης χρησιμοποιούμε ως αρχικές τιμές των μέσων τιμών και των διασπορών αυτές που προκύπτουν από την ταξινόμηση του k-means.

Πιο συγκεκριμένα στην περίπτωση μας όπου έχουμε δύο κατηγορίες υπολογίζουμε την αναμενόμενη τιμή της πιθανότητας κάθε δείγματος να ανήκει στην πρώτη κατηγορία (βήμα εκτίμησης) ως εξής:

$$E(z_{i1}) = \frac{p(x_i / \theta_1^t) \pi_1^t}{\sum_{j=1}^K p(x_i / \theta_j^t) \pi_j^t} = \frac{\pi_1^t \frac{1}{\sqrt{2\pi(\sigma_1^t)^2}} e^{-\frac{1}{2(\sigma_1^t)^2}(x_i - \mu_1^t)^2}}{\sum_{j=1}^2 \pi_j^t \frac{1}{\sqrt{2\pi(\sigma_j^t)^2}} e^{-\frac{1}{2(\sigma_j^t)^2}(x_i - \mu_j^t)^2}} \quad \text{όπου } x_i \text{ είναι το } i\text{-στο δείγμα} \quad (3.1.9)$$

Αντίστοιχος είναι και ο υπολογισμός της αναμενόμενης τιμής της πιθανότητας κάθε δείγματος να ανήκει στην δεύτερη κατηγορία.

Η νέα εκτίμηση των πιθανοτήτων των δύο κατηγοριών π_1 και π_2 είναι ίση με το άθροισμα των αναμενόμενων πιθανοτήτων των δειγμάτων να ανήκουν στην πρώτη και στην δεύτερη κατηγορία αντίστοιχα.

$$\pi_1^{t+1} = \frac{1}{n} \sum_{i=1}^n E(z_{i1})^{t+1} \quad \text{και} \quad \pi_2^{t+1} = \frac{1}{n} \sum_{i=1}^n E(z_{i2})^{t+1}$$

Εύρεση Σημείου Απόφασης

Έχοντας στην διάθεση μας τις υπολογισμένες παραμέτρους κάθε επιμέρους κατανομής, μπορούμε να βρούμε τις **τα σημεία ή τις ευθείες ή τις επιφάνειες απόφασης**. Εφόσον θα εφαρμόσουμε τον EM μόνο πάνω σε μια διάσταση την ένταση και καθώς θεωρούμε ότι έχουμε δύο κατηγορίες τότε θα έχουμε απόφασης. Κάθε νέο δείγμα που μας έρχεται το ταξινομούμε σε μια από τις δύο κατηγορίες, foreground ή background, με βάση αυτό το σημείο.

Για $k=2$ λοιπόν έχουμε δύο Gaussian κατανομές με τους παρακάτω τύπους:

$$p(x/\theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2}$$

$$p(x/\theta_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2}$$

Για να βρούμε αυτές τις κατανομές αρκεί να εξισώσουμε αυτές τις δύο κατανομές πολλαπλασιάζοντας τις με την πιθανότητα ενεργοποίησης τους.

$$\pi_1 p(x/\theta_1) = \pi_2 p(x/\theta_2) \Rightarrow$$

$$\pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2} = \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2} \Rightarrow$$

$$\ln\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 = \ln\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right) - \frac{1}{2\sigma_2^2}(x-\mu_2)^2 \Rightarrow$$

$$\ln(\pi_1) - \frac{1}{2}\ln(\sigma_1^2) - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 = \ln(\pi_2) - \frac{1}{2}\ln(\sigma_2^2) - \frac{1}{2\sigma_2^2}(x-\mu_2)^2 \Rightarrow$$

$$\frac{1}{2\sigma_2^2}(x-\mu_2)^2 - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 = \ln\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \Rightarrow$$

$$(\sigma_1^2 - \sigma_2^2) \cdot x^2 + 2(\sigma_2^2\mu_1 - \sigma_1^2\mu_2) \cdot x + \sigma_1^2\mu_2 - \sigma_2^2\mu_1 - 2 \cdot \sigma_1^2\sigma_2^2 \cdot \left(\ln\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right) = 0 \Rightarrow$$

Όπότε

$$A = (\sigma_1^2 - \sigma_2^2)$$

$$B = 2(\sigma_2^2\mu_1 - \sigma_1^2\mu_2)$$

$$\Gamma = \sigma_1^2\mu_2 - \sigma_2^2\mu_1 - 2 \cdot \sigma_1^2\sigma_2^2 \cdot \left(\ln\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right)$$

και λύνοντας την εξίσωση $\mathbf{Ax}^2 + \mathbf{Bx} + \mathbf{\Gamma} = \mathbf{0}$ θα έχουμε τα σημεία απόφασης.

3.1.1.4 Επαναληπτική κατάτμηση με βάση τον αλγόριθμο Expectation Maximization

Όπως και με τον αλγόριθμο Otsu έτσι και με τον EM κάνουμε την παραδοχή ότι έχουμε δύο κατηγορίες. Μια εναλλακτική με καλύτερη προσέγγιση της συνολικής κατανομής της έντασης θα ήταν να χρησιμοποιήσουμε το Bayesian κριτήριο (BIC - Bayesian Information Criterion) για να εκτιμήσουμε τον αριθμό των κατηγοριών και έπειτα να προχωρήσουμε σε κατάτμηση της εικόνας σε k κατηγορίες. Αυτό όμως στην περίπτωση του EM σημαίνει ότι θα έπρεπε να εκτιμήσουμε τις παραμέτρους k κατανομών και να βρούμε στην καλύτερη περίπτωση $k-1$ σημεία απόφασης. Για το λόγο αυτό προτιμήσαμε αν εφαρμόσουμε και τον EM επαναληπτικά (για να εκτιμήσουμε τις παραμέτρους μόνο δύο κατηγοριών) μέχρις ότου να έχουμε ένα ικανοποιητικό αριθμό από pixels που ανήκουν στο spot.

Σε κάθε επανάληψη εάν ο αριθμός των pixel που ανήκουν στο foreground είναι μικρότερος από το 28% του μεγέθους του κελιού, ο αλγόριθμος αναταξινομεί τα Pixel του background, όπως αυτά έχουν προκύψει από την προηγούμενη ταξινόμηση, σε δύο νέες κατηγορίες. Τα Pixel με μεγαλύτερη ένταση ανήκουν στο foreground. Στην περίπτωση όπου ο αριθμός των pixel που ανήκουν στο foreground είναι μικρότερος από το 78% του μεγέθους του κελιού γίνεται αναταξινόμηση στο foreground αντί για το background.

```
[foreground,Threshold,θ]=iter_EM(Ένταση)
[foreground,Threshold,θ]= iter_EM (Ένταση)
while (άθροισμα(foreground)<.28*ΜέγεθοςΕικόνας or άθροισμα(foreground)>.78*ΜέγεθοςΕικόνας)
    if(άθροισμα(foreground)<.28*ΜέγεθοςΕικόνας)
        [foreground,Threshold,θ]= iter_EM (Ένταση(background));
    else(άθροισμα(foreground)>.78*ΜέγεθοςΕικόνας)
        [foreground,Threshold,θ]= iter_EM (Ένταση(foreground));
    end
```

Αλγ 3.1.2 Επαναληπτική κατάτμηση με βάση τον αλγόριθμο EM.

3.1.2 Κατάτμηση με Ομαδοποίηση

3.1.2.1 K-means clustering

Ο κλασικός αλγόριθμος k-means clustering και χρησιμοποιήθηκε πρώτη φορά πάνω στα Microarrays από Bozinov and Rahnenfuhrer το 2002. Η κατάτμηση γίνεται με ταυτόχρονη χρήση και των δύο καναλιών χρώματος. Έχουμε δηλαδή δυσδιάστατα δείγματα, δηλαδή δύο χαρακτηριστικά ένα η τιμή έντασης για το κόκκινο και ένα δεύτερο χαρακτηριστικό η τιμή έντασης για το πράσινο. Εφόσον θέλουμε κατάτμηση της εικόνας σε δύο τμήματα το k είναι ίσο με 2. Ως αρχικές θέσεις των κέντρων των δύο κατηγοριών θέτουμε την μέγιστη και την ελάχιστη τιμή έντασης. Ακολουθεί ταξινόμηση των δειγμάτων ανάλογα με την ευκλείδεια απόσταση των δειγμάτων από τα δύο κέντρα, Στην συνέχεια υπολογίζονται να νέα κέντρα και ο αλγόριθμος τερματίζει όταν τα δύο κέντρα πάψουν να μεταβάλλονται.

Ο αλγόριθμος k-means λειτουργεί με παρόμοιο τρόπο με αυτό που λειτουργεί ο αλγόριθμος Otsu προσπαθεί να ομαδοποιήσει τα δείγματα (pixels) σε k κατηγορίες βρίσκοντας την μέση τιμή κάθε

κατηγορίας και ελαχιστοποιώντας την συνολική απόσταση των δειγμάτων από τα κέντρα των κατηγοριών. Τα σημεία διαχωρισμού προκύπτουν από τις μεσοκάθετες των ευθειών που ενώνουν τα κέντρα των κατηγοριών.

Ο αλγόριθμος αυτός συνεπώς προχωρά σε απ' ευθείας ταξινόμηση των δειγμάτων. Χρησιμοποιεί μια απλοποίηση του προβλήματος υποθέτοντας ότι οι k κατανομές περιγράφονται μόνο από μια παράμετρο, την μέση τιμή τους. Ταξινομεί τα δείγματα με βάση την απόστασή τους από τα κέντρα των κατηγοριών και προχωρά σε νέα εκτίμηση αυτών των κέντρων.

Ορίζει k μέσες τιμές m_i και στις δίνει τυχαίες αρχικές τιμές. Στην συνέχεια κατηγοριοποιεί όλα τα δείγματα με βάση αυτές στις μέσες τιμές. Για παράδειγμα το δείγμα x ταξινομείται στην κατηγορία I εάν η ευκλείδεια απόστασή του x από το m_i είναι μικρότερη από στις ευκλείδειες του x από τα μέσα των υπόλοιπων κατηγοριών. Δηλαδή:

$$x \in c_i \quad \text{αν} \quad ||x - m_i|| < ||x - m_j|| \quad \text{για} \quad \text{κάθε} \quad j \neq i$$

```

Αρχικοποίηση των  $m_1, m_2, \dots, m_k$ 
do
    Ταξινόμηση των Pixels με βάση την
    ευκλείδεια απόσταση από τα  $m_i$ 
    Υπολογισμός των νέων μέσων  $m_1,$ 
     $m_2, \dots, m_k$ 
Until δεν υπάρχει άλλη διόρθωση των  $m_j$ 
Return  $m_1, m_2, \dots, m_k$ 

```

Αλγ 3.1.3 Αλγόριθμος k-means

3.1.2.2 Υβριδικός k-means (Hybrid k-mean)

Ο υβριδικός k-means είναι μια εκτεταμένη έκδοση του κανονικού K-means (Bozinov and Rahnenfuehrer το 2004) [30]. Ο νέος αλγόριθμος χρησιμοποιεί επαναλαμβανόμενη ταξινόμηση με στόχο να αυξηθεί ο αριθμός των pixels που ταξινομούνται στο foreground.

Όσο ο ελάχιστος εκτιμώμενος αριθμός pixels που θα πρέπει να ανήκουν στο foreground μετά το clustering δεν έχει προσεγγιστεί, τα εναπομείναντα pixels του background ταξινομούνται ξανά σε δύο κατηγορίες και η ομάδα των pixels με τις μεγαλύτερες τιμές έντασης κατηγοριοποιούνται στο foreground.

3.1.2.3 Παραλλαγή υβριδικού k-means

Οι αλγόριθμοι Otsu και k-means εφαρμοζόμενοι πάνω στο ιστόγραμμα της έντασης έχουν παρόμοια αποτελέσματα και αντιμετωπίζουν τα ίδια προβλήματα. Συγκεκριμένα έχουν λανθασμένα αποτελέσματα στις περιπτώσεις όπου έχουμε αχνά spot και όταν μέσα στο κελί λόγω κακής τοποθέτησης πλέγματος έχουμε τμήματα από διπλανά spot. Τα προβλήματα οφείλονται στην ουσία σε ένα τον λόγο ότι όταν τα pixel δεν χωρίζονται μόνο σε δύο κατηγορίες, το background και το foreground αλλά σε τρεις και ίσως και παραπάνω κατηγορίες.

Ένας τρόπος αντιμετώπισης αυτών των προβλημάτων είναι να χρησιμοποιηθεί **k-means παραπάνω τάξης** ($k=3, 4, \dots$). Σ' αυτήν την περίπτωση μπορούμε να χρησιμοποιήσουμε το Bayesian κριτήριο (BIC - Bayesian Information Criterion) ώστε να εκτιμηθεί ο αριθμός των κατηγοριών από της οποίες προέκυψε η συνολική κατανομή, δηλαδή η τιμή του k . Στην συνέχεια οι κατηγορίες με τις μεγαλύτερες τιμές έντασης συνενώνονται σε μία που αποτελεί το foreground (spot).

Ένας άλλος τρόπος όπου δεν χρειάζεται εκτιμήσουμε πόσες κατηγορίες από pixels έχουμε είναι η **επαναληπτική εφαρμογή του 2-means**. Θεωρούμε πάντα ότι έχουμε δύο κατηγορίες και σε κάθε

επανάληψη εφαρμόζουμε τον αλγόριθμο k-means πάνω στην κατανομή των pixels που έχουν ταξινομηθεί στο background σύμφωνα με την προηγούμενη επανάληψη. Κατ' αυτόν τον τρόπο συμπεριλαμβάνουμε σταδιακά μια νέα ομάδα εντάσεων στο spot μας μέχρι συγκεντρώσουμε μια ικανοποιητική ποσότητα από pixels (γύρω στο 50% της συνολικής εικόνας ενός κελιού).

Ακολουθεί μια παραλλαγή της επαναληπτικής κατάτμησης την οποία και εφαρμόζουμε. Σε κάθε επανάληψη αντί να κάνουμε κατάτμηση στα pixel του background που έχουν προκύψει από το προηγούμενο βήμα της επανάληψης εφαρμόζουμε κατάτμηση σε όλη την εικόνα αφού προηγουμένως θέσουμε την τιμή της έντασης των pixels που ανήκουν στο foreground ίση με την μέγιστη τιμή έντασης του background.

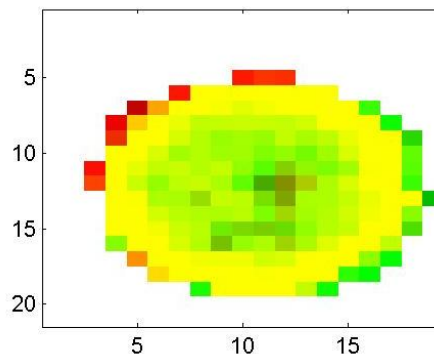
Πιο αναλυτικά, σε κάθε βήμα επανάληψης προκύπτουν δύο ομάδες pixels, το background (μικρές τιμές έντασης) και το foreground (μεγάλες τιμές έντασης). Εάν το ποσοστό των Pixels που ταξινομήθηκαν στο foreground είναι λιγότερο από το 38% του συνόλου της εικόνας τότε υποβαθμίζουμε τις τιμές έντασης τους έτσι ώστε να είναι ίσες με την μέγιστη τιμή έντασης που συναντάμε στο background (όπως έχει προκύψει από το προηγούμενο βήμα της επανάληψης) και εφαρμόζουμε νέα κατάτμηση με τις νέες τιμές εντάσεων.

```
[foreground, background, M1, M2] = iter_kmeans('Ένταση')
[foreground, background, M1, M2] = k_means('Ένταση')
while (άθροισμα(foreground) < .38 * ΜέγεθοςΕικόνας)
    Ένταση(foreground) = max(Ένταση(background))
    [foreground, background, M1, M2] = k_means('Ένταση')
end
```

Αλγ 3.1.4 Παραλλαγή επαναληπτικού αλγόριθμου k-means.

3.2 Κατάτμηση έντασης και απόχρωσης

Αφού έχουμε ένα πρώτο διαχωρισμό του spot με βάση το ιστόγραμμα της έντασης, χρησιμοποιούμε ένα ακόμα χαρακτηριστικό την τιμή της απόχρωσης (hue) για να αφαιρέσουμε κάποια pixel τα οποία ανήκουν στο spot και έχουν προκύψει από την κακή ευθυγράμμιση των δύο εικόνων (Cy3, Cy5). Δηλαδή pixel τα οποία περιέχουν μόνο τιμές του κόκκινου ή μόνο τιμές του πράσινου όπως φαίνεται και στην παρακάτω εικόνα.

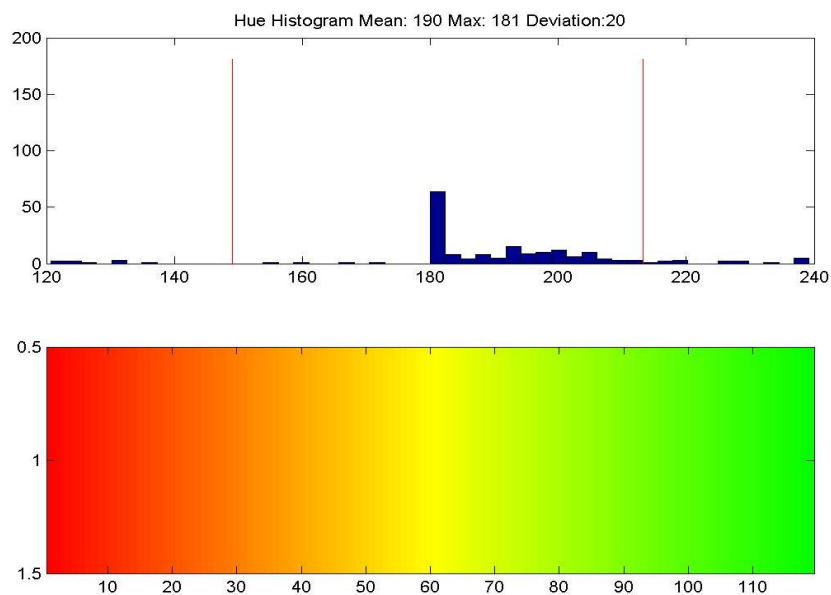


Εικόνα 3.2.1 Σφάλμα ευθυγράμμισης των δύο εικόνων. Αποτέλεσμα κατάτμησης με βάση μόνο το ιστόγραμμα της έντασης. Το spot αυτό έχει στην περιφέρεια pixel πράσινου και κόκκινου χρώματος αποτέλεσμα κακής ευθυγράμμισης των δύο εικόνων (Cy3, Cy5). Τα pixel αυτά έχουν τιμές έντασης μόνο στο κόκκινο ή μόνο στο πράσινο και εισάγουν κάποιο σφάλμα στον υπολογισμό της τιμής έκφρασης για το αντίστοιχο spot.

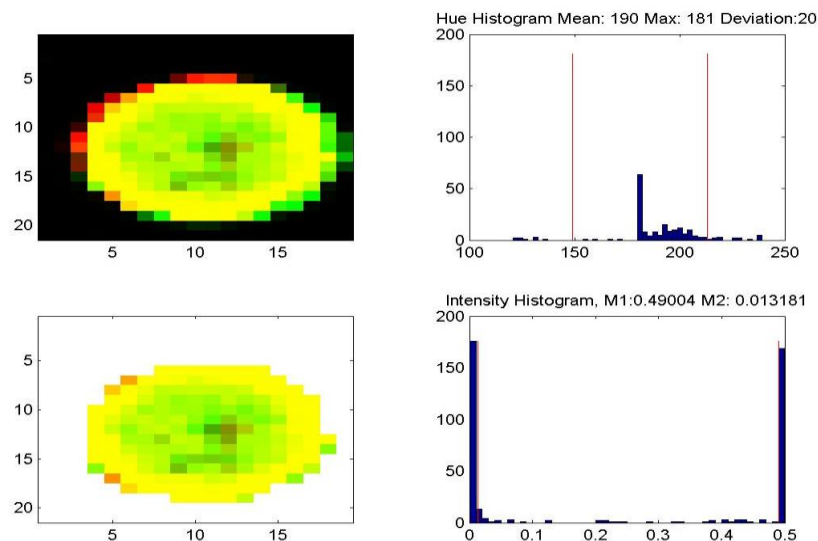
Το ιστόγραμμα της απόχρωσης (hue) έχει ένα μέγιστο γύρω από την απόχρωση του spot. Άρα μπορούμε να θεωρήσουμε ότι έχουμε μια Gaussian κατανομή με μέγιστο στην τιμή της απόχρωσης αυτής για παράδειγμα του κίτρινου. Για να διαχωρίσουμε το spot από το φόντο του αρκεί να πάρουμε όλα τα pixels που έχουν τιμή απόχρωσης κοντά στην απόχρωση του spot .

Εφόσον θεωρούμε ότι έχουμε Gaussian κατανομή για να επιλέξουμε τα pixel που μας ενδιαφέρουν αρκεί να ορίσουμε ένα band pass φίλτρο με πλάτος 2σ όπου σ η διασπορά της κατανομής της απόχρωσης (hue). Κρατάμε δηλαδή τα pixels που έχουν τιμή hue μεγαλύτερη από $\text{maxim} + \sigma$ και μεγαλύτερη από $\text{maxim} - \sigma$, όπου maxim το μέγιστο της κατανομής της απόχρωσης για ένα κελί.

$$\text{Maxim} - \sigma < \text{Hue} < \text{Maxim} + \sigma$$



Εικόνα 3.2.2 Ιστόγραμμα Απόχρωσης. Το ιστόγραμμα της απόχρωσης για το spot της εικόνας 3.2.1. Οι δύο κόκκινες ευθείες είναι τα δύο Threshold. Μέσα στα όρια των ευθειών αυτών είναι το κομμάτι του ενδιαφέροντος μας.



Εικόνα 3.2.3 Κατάτμηση βάση έντασης και απόχρωσης. Η αρχική εικόνα του κελιού μας, το αποτέλεσμα της κατάτμησης με βάση και τα δύο χαρακτηριστικά. Βλέπουμε με την εφαρμογή του φίλτρου στο ιστόγραμμα της απόχρωσης αφαιρούνται τα pixels που έχουν προκύψει από την κακή ευθυγράμμιση των δύο εικόνων.

3.2.1 Δυσδιάστατη Κατάτμηση

3.2.1.1 K-means δύο διαστάσεων

Έχουμε δύο χαρακτηριστικά τα οποία έχουν μικρό βαθμό εξάρτησης. Θα εφαρμόσουμε συνεπώς k-means στην δυσδιάστατη κατανομή των pixel λαμβάνοντας υπόψη εκτός από την ένταση την απόχρωση (hue) του κάθε pixel. Η διαφορά με τον προηγούμενο k-means είναι ότι τώρα οι μέσες τιμές των κατηγοριών είναι δυσδιάστατες. Κάνουμε και πάλι την παραδοχή ότι $k=2$.

3.2.1.2 Επαναληπτικός k-means δύο διαστάσεων

Κατ' αναλογία με τον απλό k-means έχουμε επαναληπτική εφαρμογή του δυσδιάστατου k-means για καλύτερα αποτελέσματα. Η υλοποίηση είναι ανάλογη με την παραλλαγή του υβριδικού k-means και η επανάληψη επαναλαμβάνεται όσο το σύνολο των pixel που ανήκουν στο spot είναι λιγότερο από το 30% του συνόλου των pixel. Το ποσοστό προέκυψε πειραματικά.

```
[foreground, background, M1, M2] = iter_kmeans2D(Ένταση, Απόχρωση)
[foreground, background, M1, M2] = k_means2D(Ένταση, Απόχρωση)
while (άθροισμα(foreground) < .3 * ΜέγεθοςΕικόνας)
    Ένταση(foreground) = max(Ένταση(background))
    [foreground, background, M1, M2] = k_means2D(Ένταση, Απόχρωση)
end
```

Αλγ. 3.2.1 Επαναληπτικός Δυσδιάστατος k-means

3.2.2 Διαδοχική Κατάτμηση

3.2.2.1 Ασυσχέιστα χαρακτηριστικά

Στην προηγούμενη ενότητα περιγράψαμε τεχνικές κατάτμησης με ταυτόχρονη εκμετάλλευση και των δύο χαρακτηριστικών Έντασης και η Απόχρωσης με το να θεωρήσουμε ότι έχουμε δυσδιάστατα δείγματα με τιμές (I, H) και μια δυσδιάστατη κατανομή. Στην ενότητα αυτή θα εφαρμόσουμε διαδοχική κατάτμηση των κελιών με ανεξάρτητη θεώρηση των χαρακτηριστικών δηλαδή θα εφαρμόσουμε κατάτμηση πρώτα με βάση το ένα χαρακτηριστικό και έπειτα με βάση το άλλο.

Για να μπορέσουμε να το κάνουμε αυτό και να έχουμε ισάξια αποτελέσματα με την δυσδιάστατη θεώρηση θα δείξουμε ότι τα χαρακτηριστικά έχουν μικρή συσχέτιση. Το γεγονός αυτό μας επιτρέπει να επεξεργαστούμε τα χαρακτηριστικά μας ξεχωριστά το ένα από το άλλο.

Η μικρή συσχέτιση των χαρακτηριστικών δεν σημαίνει ότι μπορούμε να θέσουμε thresholds στα δύο χαρακτηριστικά με όποια σειρά θέλουμε και ανεξάρτητα το ένα από το άλλο. Αυτό θα μπορούσαμε να το κάνουμε εάν τα χαρακτηριστικά μας ήταν ανεξάρτητα και οι κατανομές τους Gaussian κάτι το οποίο δεν μπορούμε να ισχυριστούμε.

Μπορούμε να δείξουμε όμως ότι τα δύο χαρακτηριστικά είναι πρακτικά ασυσχέιστα υπολογίζοντας των *συντελεστή συσχέτισης*.

$$\rho = \frac{C}{\sigma_x \sigma_y} \quad (3.2.2.1)$$

Όπου C είναι η συνδιακύμανση (covariance) των δύο χαρακτηριστικών και ισούται με :

$$C = E\{(X - E(X))(Y - E(Y))\} = E(XY) - E(X)E(Y) \quad (3.2.2.2)$$

Για να είναι τα δύο χαρακτηριστικά ασυσχέτιστα θα πρέπει η απόλυτη τιμή του **συντελεστή συσχέτισης** να είναι ίσο με το μηδέν, δηλαδή:

$$C = 0, \rho = 0 \text{ ή αλλιώς } E(XY) - E(X)E(Y)$$

Για να μπορέσουμε να κάνουμε την παραδοχή ότι τα δύο χαρακτηριστικά είναι ασυσχέτιστα θα πρέπει η τιμή του συντελεστή να είναι κοντά στο μηδέν

$$|\rho| \rightarrow 0$$

Χρησιμοποιήσαμε το σύνολο των εικόνων μας και υπολογίσαμε τον συντελεστή συσχέτισης:

$$|\rho| = 6.1860 \cdot 10^{-7} \cong 0$$

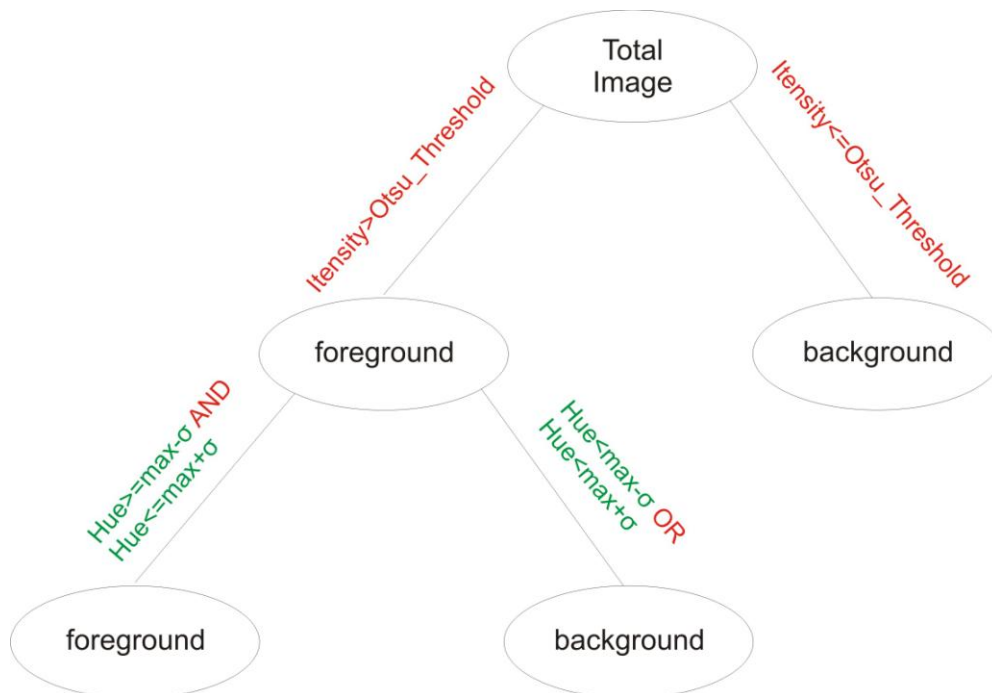
3.2.2.2 Δέντρο απόφασης

Θεωρούμε και πάλι ότι έχουμε δύο κλάσεις (κατηγορίες ταξινόμησης) το **foreground-spot** και το **background** και δύο χαρακτηριστικά την **ένταση (intensity)** και την **πρόσμιξη (hue)**.

Σε αντίθεση με την δυσδιάστατη ταξινόμηση εκμεταλλευόμαστε την μικρή συσχέτιση των δύο χαρακτηριστικών μας και θέτουμε δύο ξεχωριστά thresholds ένα για κάθε χαρακτηριστικό. Είναι σαν να θεωρούμε ότι τα δύο χαρακτηριστικά μας είναι ασυσχέτιστα κάτι το οποίο πολύ μικρό σφάλμα.

Εφαρμόζουμε κατάτμηση δύο επιπέδων όπου στο πρώτο επίπεδο διαχωρίζουμε τα pixel με βάση την τιμή της έντασης τους και στο δεύτερο επίπεδο με βάση την απόχρωση τους. Το πρώτο επίπεδο τροφοδοτεί το δεύτερο. Έχουμε δηλαδή ένα δέντρο απόφασης όπου σε πρώτη φάση θέτουμε ένα σημείο απόφασης σύμφωνα με το πρώτο χαρακτηριστικό (ένταση) και το αποτέλεσμα τροφοδοτεί το δεύτερο επίπεδο του δέντρου το οποίο θέτει ένα σημείο απόφασης συμφωνά με το δεύτερο χαρακτηριστικό (hue).

Έχουμε το παρακάτω δέντρο στο οποίο φυσικά εάν αλλάξει η σειρά εφαρμογής των thresholds θα έχουμε ένα διαφορετικό τελικό αποτέλεσμα ακριβώς λόγω της μη ανεξαρτησίας των χαρακτηριστικών. Στο πρώτο επίπεδο επιλέγουμε το χαρακτηριστικό της έντασης γιατί είναι πιο ισχυρό ενώ η απόχρωση έχει συμπληρωματικό ρόλο.



Εικ. 3.2.4 Διαδοχική Κατάτμηση. Στο παραπάνω δέντρο απόφασης στο πρώτο επίπεδο διαχωρισμού θέτουμε σημείο απόφασης στην κατανομή της έντασης σύμφωνα με τον αλγόριθμο Otsu k-means ή EM. Στο δεύτερο επίπεδο βρίσκουμε το μέγιστο της κατανομής της Απόχρωσης (Hue) και κρατάμε τα pixel που έχουν Απόχρωση σε απόσταση το πολύ σ γύρω από το μέγιστο.



4 ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΦΑΡΜΟΓΗΣ

Στην ενότητα αυτή εφαρμόζονται οι τέσσερις αλγόριθμοι που επιλέξαμε και η επαναληπτική εκδοχή τους, (απλός και επαναληπτικός k-means, απλός και επαναληπτικός Otsu, απλός και επαναληπτικός EM, δυσδιάστατος K-means) πάνω στο σύνολο των εικόνων που έχουμε στην διάθεσή μας από το Εθνικού Ινστιτούτου Καρκίνου (National Cancer Institute ^[25]).

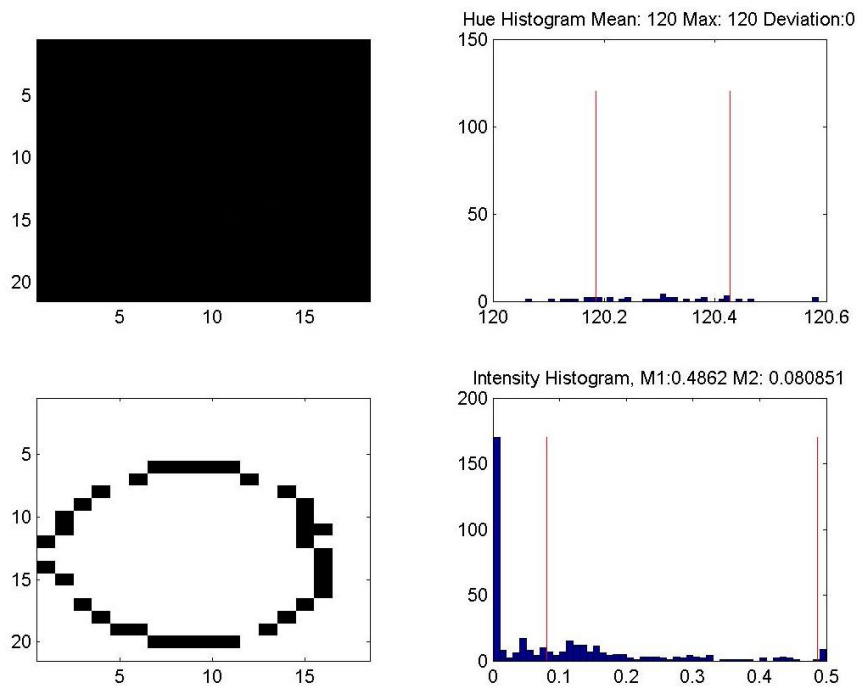
Περιγράφονται τα προβλήματα τα οποία προκύπτουν με τους απλούς αλγορίθμους και ο λόγος που αναπτύξαμε την επαναληπτική εκδοχή τους. Τέλος προστίθεται ένας αλγόριθμος ο οποίος χρησιμοποιεί χωρική πληροφορία ώστε να οριοθετήσει την περιοχή του spot. Ο αλγόριθμος αυτός εφαρμόζεται στο αποτέλεσμα της κατάτμησης με στόχο να παραλείψει ένα αριθμό από pixels στην περιφέρεια του κελιού τα οποία λόγω θορύβου έχουν υψηλές τιμές έντασης και έχουν συμπεριληφθεί λανθασμένα στο spot.

4.1 Απλοί Otsu, K- means – Προβλήματα

Ο αλγόριθμος k-means έχει παρόμοια απόδοση με τον Otsu. Δεν λειτουργεί καλά στις παρακάτω περιπτώσεις:

- Όταν το spot είναι πολύ αχνό και δύσκολα διακρίνεται από το background του.
- Όταν η τοποθέτηση πλέγματος δεν έχει γίνει σωστά και υπάρχουν κελιά που περιέχουν τμήμα κάποιου spots γειτονικού κελιού.

Στην **πρώτη περίπτωση** η κατανομή της έντασης δεν εμφανίζει δύο ευδιάκριτα μέγιστα και ο αλγόριθμος δεν προσεγγίζει καλά τις μέσες τιμές. Όταν ένα spot είναι αχνό σημαίνει ότι πολλά pixels τα οποία ανήκουν στο spot έχουν τιμές έντασης κοντά στο μηδέν, δηλαδή κοντά στις τιμές έντασης του background. Έχουμε επίσης και μια ομάδα από pixels τα οποία ανήκουν στο spot και συγκεκριμένα αυτά που βρίσκονται στην περιφέρεια του spot τα οποία έχουν μεγάλες τιμές έντασης. Στην ουσία επειδή έχουμε ένα spot με πολύ κοντινές τιμές έντασης με αυτές του background έχουμε μια συνολική κατανομή με μεγάλη συγκέντρωση γύρω από τις χαμηλές τιμές έντασης. Αυτό έχει ως αποτέλεσμα το spot να χωρίζεται σε δύο κατηγορίες: τα **περιφερειακά Pixels** και το **εσωτερικό του spot**.



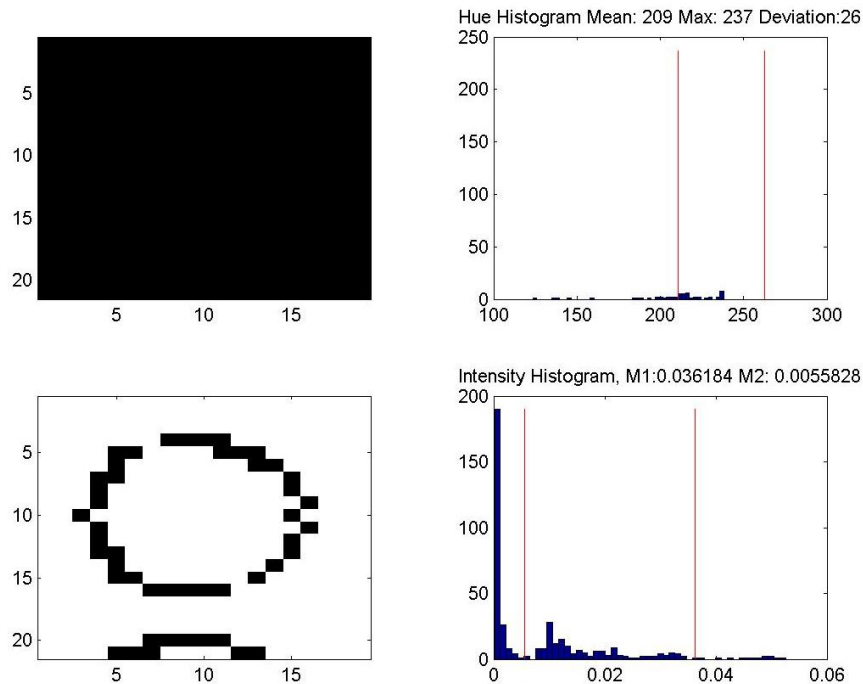
Εικόνα 4.1.1 Αχνό Spot.

Το spot είναι πολύ αχνό και η κατανομή της έντασης αποτελείται από τρεις κατανομές: το background (τιμές έντασης γύρω από το μηδέν), την περιφέρεια του spot (μεγάλες τιμές έντασης, γύρω από το 0.5) και το εσωτερικό του spot (τιμές έντασης λίγο μεγαλύτερες του μηδενός). Ο 2-means διαχωρίζει μόνο τα περιφερειακά pixel του spot.

Συνεπώς η ολική κατανομή της έντασης δεν μπορεί να προσεγγιστεί με άθροισμα μόνο δύο κατανομών. Έχουμε δηλαδή στην ουσία τρεις κατηγορίες (μπορεί και παραπάνω) και αυτός είναι ο λόγος που αποτυγχάνει ο 2-means.

Στην **δεύτερη περίπτωση** είναι πιο ξεκάθαρο γιατί η κατανομή της έντασης δεν μπορεί να προσεγγιστεί με άθροισμα δύο μόνο Gaussian. Σ' αυτήν την κατηγορία έχουμε κελιά τα οποία περιέχουν δύο spot (μπορεί και παραπάνω) τα οποία μπορεί να τύχει να έχουν παρόμοιες τιμές έντασης αλλά μπορεί να έχουν και μεγάλη διαφορά στις τιμές έντασης. Έχουμε δηλαδή δύο περιπτώσεις:

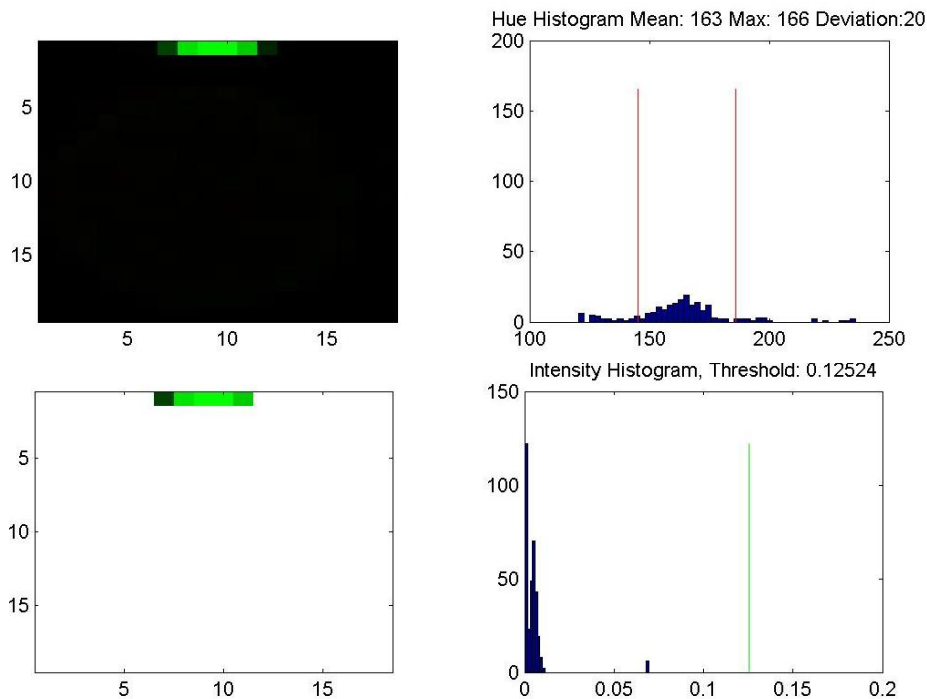
Όταν οι τιμές έντασης των γειτονικών spot είναι παρόμοιες η κατανομή μπορεί να δώσει δύο μέγιστα τα οποία να διαχωρίζουν και τα δύο spot από το background. Το μόνο σφάλμα που υπεισέρχεται σ' αυτήν την περίπτωση είναι το σφάλμα του συνυπολογισμού των γειτονικών pixels στην τιμή της έντασης του spot. Στην εικόνα 4.1.2 τα spot έχουν παρόμοιες τιμές έντασης αλλά τα pixel στο εσωτερικό τους έχουν χαμηλότερες εντάσεις με αποτέλεσμα οι αλγόριθμοι να διαχωρίζουν μόνο την περιφέρεια τους από την υπόλοιπη εικόνα.



Εικόνα 4.1.2 Pixel από γειτονικό spot με παρόμοια ένταση. Δύο spot με παρόμοιες τιμές έντασης τα οποία όμως έχουν αχνό εσωτερικό. Η κατανομή της έντασης έχει πάνω από δύο μέγιστα, δηλαδή αποτελείται από άθροισμα περισσότερων από δύο κατανομών. Ο 2-means διαχωρίζει μόνο την περιφέρεια και των δύο spot.

Όταν οι τιμές έντασης των γειτονικών spot απέχουν από αυτές του spot που εξετάζουμε τότε το ιστόγραμμα της έντασης μεταβάλλεται έτσι ώστε έχουμε πολύ μεγάλη συγκέντρωση γύρω από το μηδέν και έτσι χάνουμε την δυνατότητα διάκρισης των μικρών εντάσεων όπου και βρίσκεται το spot μας. Ο 2-means βρίσκει το ένα μέγιστο στις μεγάλες εντάσεις του γειτονικού spot ενώ τα υπόλοιπα pixels ομαδοποιούνται σε ένα μέγιστο κοντά στο μηδέν.

Για παράδειγμα στην εικόνα 4.1.3 έχουμε τρία Pixels τα οποία ανήκουν στο διπλανό spot το οποίο τυχαίνει να έχει πολύ μεγαλύτερες τιμές έντασης. Το αποτέλεσμα της συμπεριληψης αυτών των pixel στο ιστόγραμμα είναι μια κατανομή με πολύ υψηλό μέγιστο κοντά στις χαμηλές τιμές εντάσεων. Ο 2-means θεωρεί ότι μόνο τα τρία αυτά pixels ανήκουν στο foreground.



Εικόνα 4.1.3 Τμήμα από γειτονικό spot με μεγαλύτερη ένταση. Ένας αριθμός από pixel γειτονικού spot συμπεριλήφθηκαν κατά λάθος κατά την διαδικασία τοποθέτησης του πλέγματος μέσα στο κελί. Το γειτονικό spot τυχαίνει να έχει μεγαλύτερες τιμές έντασης από το spot που προσπαθούμε να διαχωρίσουμε. Λόγω της συμπερίληψης αυτών των pixel, το ιστόγραμμα της έντασης επηρεάζεται ώστε έχουμε πολύ μεγάλη συγκέντρωση γύρω από το μηδέν και έτσι χάνουμε την δυνατότητα διάκρισης των μικρών εντάσεων. Το spot που μας ενδιαφέρει έχει πολύ χαμηλότερες εντάσεις από το γειτονικό και οι αλγόριθμοι το «βλέπουν» σαν background.

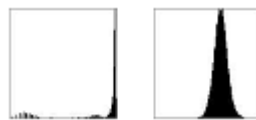
4.2 ΕΜ – Προβληματικές Περιπτώσεις

Ο ΕΜ δεν λειτουργεί καλά στις παρακάτω περιπτώσεις:

- [1] Όταν η κατανομή της έντασης είναι **μονοκόρυφος**.

Ο αλγόριθμος δεν μπορεί δώσει δύο ξεχωριστές κατανομές. Η περίπτωση αυτή πολύ σπάνια την συναντάμε, για παράδειγμα όταν δεν υπάρχει spot.

- [2] Όταν η κατανομή της έντασης έχει πάνω από μια κορυφή αλλά η μια κορυφή είναι **στα όρια του ιστογράμματος**.

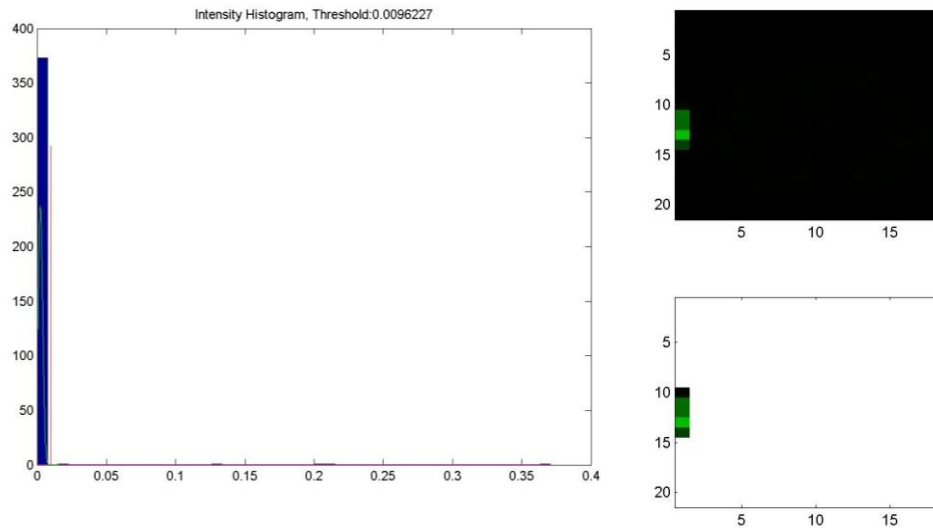


Στα όρια μονοκόρυφη

Η κατανομή η οποία βρίσκεται στα όρια προσεγγίζεται με πολύ μικρή διασπορά που τείνει στο μηδέν. Σ' αυτήν την περίπτωση δεν μπορούν να εκτιμηθούν οι παράμετροι της μιας κατανομής.

Η δεύτερη περίπτωση εμφανίζεται υπό την μορφή όπου η κατανομή η οποία βρίσκεται στις χαμηλότερες εντάσεις (background) έχει μέγιστο πολύ κοντά στο μηδέν και πολύ μεγαλύτερο από το

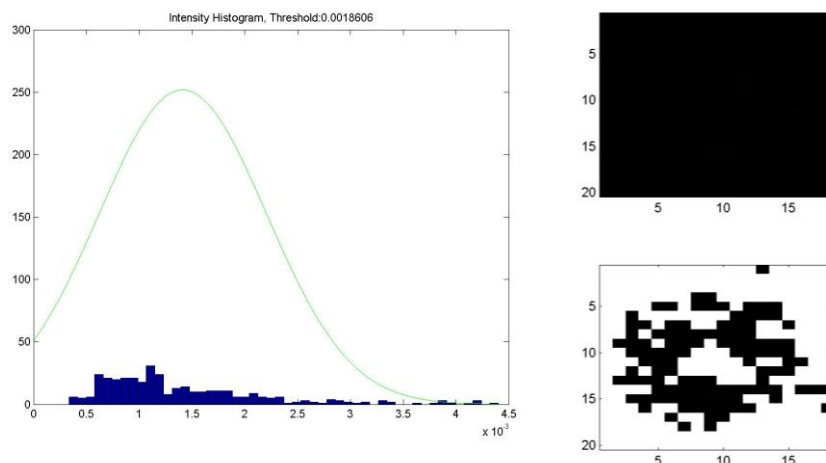
μέγιστο της δεύτερης κατανομής. Ο EM εκτιμάει ότι η διασπορά της πρώτης κατανομής (background) τείνει στο μηδέν και το μέγιστο στο άπειρο. Συνεπώς δεν έχουμε Gaussian κατανομή αλλά μια κάθετη ευθεία. Δεν μπορούμε να ορίσουμε σημείο απόφασης και έτσι χρησιμοποιούμε το αποτέλεσμα το k-means.



Εικόνα 4.2.1 Στα όρια του ιστογράμματος. Η κατανομή στις χαμηλές εντάσεις (background) έχουν πολύ μεγαλύτερο μέγιστο από την κατανομή της δεύτερης κατηγορίας. Η μια κατηγορία βρίσκεται δηλαδή στα όρια. Ο EM μας δίνει πολύ κοντινές κατανομές (magenta και green) οι οποίες δίνουν λάθος σημείο απόφασης.

Η πρώτη περίπτωση εμφανίζεται σε συνδυασμό με την δεύτερη, δηλαδή μια μονοκόρυφη κατανομή κοντά στα όρια. Ο αλγόριθμος αποτυγχάνει διότι δίνει ως αποτέλεσμα δύο κατανομές που βρίσκονται πάρα πολύ κοντά μεταξύ τους (σχεδόν ίδιες). Το πρόβλημα αυτό επιλύεται μέσω της επαναληπτικής κατάτμησης.

Επιλύουμε τα προβλήματα αυτά μέσω της επαναληπτικής κατάτμησης με EM. Στην πρώτη επανάληψη επειδή δεν έχουμε σωστό σημείο απόφασης χρησιμοποιούμε τις μέσες τιμές που δίνει ως αποτέλεσμα ομαδοποίησης ο k-means.



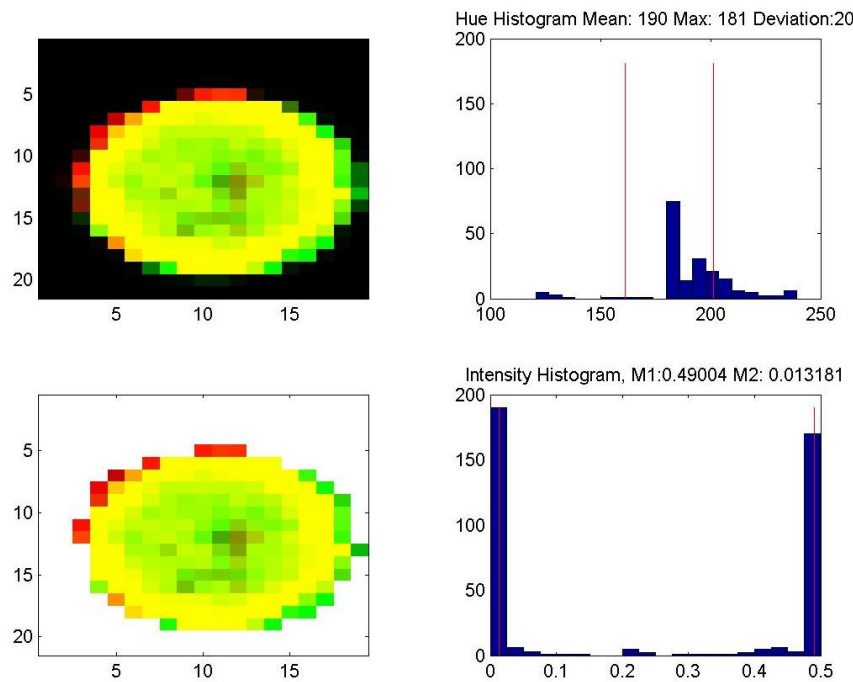
Εικόνα 4.2.2 Μονοκόρυφος κατανομή έντασης. Σ αυτήν την περίπτωση η μια από τις δύο κατανομές έχει αόριστη μέση τιμή και έτσι δεν μπορούμε να υπολογίσουμε σημείο απόφασης. Αυτό συμβαίνει διότι οι τιμές έντασης του spot είναι παρόμοιες με αυτές του background (αχνό spot).

4.3 Επαναληπτικός k-means – Εφαρμογή

Εφαρμόζουμε τον αλγόριθμο στις δύο προβληματικές περιπτώσεις που αναφέραμε στην ενότητα 4.3, σε ένα κελί όπου το spot είναι αχνό και σε ένα άλλο όπου έχουμε pixel από γειτονικό spot.

- **Απλή περίπτωση:** spot το οποίο διαχωρίζεται εύκολα από το background.

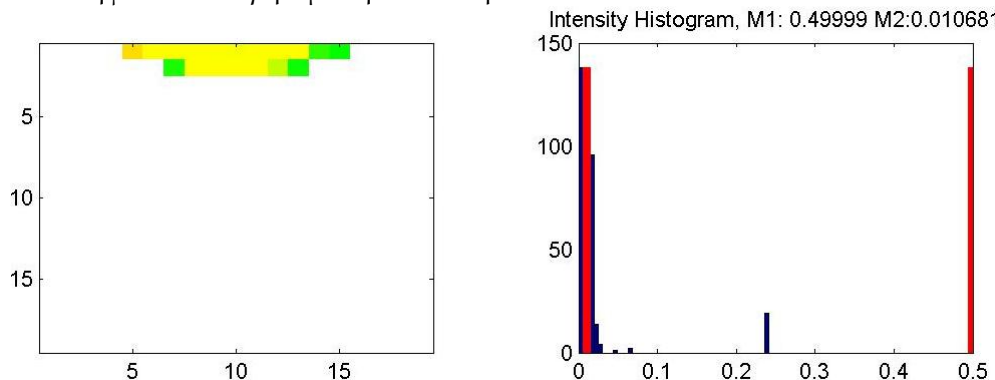
Τα αποτελέσματα είναι τα ίδια με αυτά του απλού k-means καθώς ο επαναληπτικός k-means χρειάστηκε μόνο μια επανάληψη,



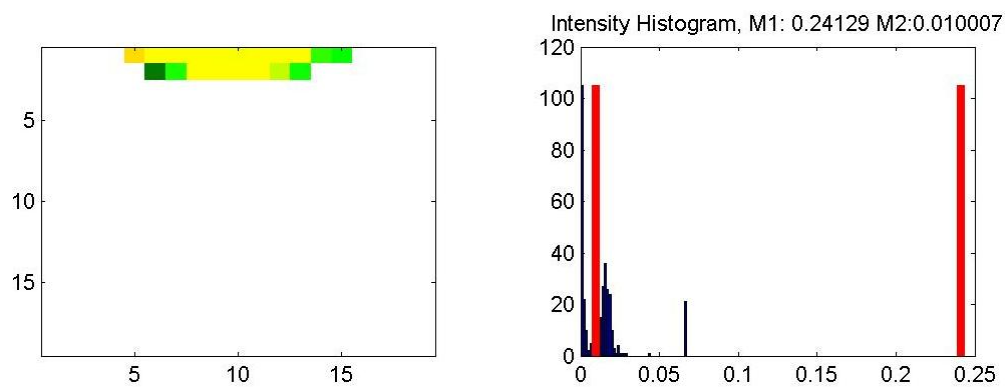
Εικόνα 4.3.1 Επαναληπτικός k-means – Εύκολα διαχωρίσιμο spot. Το αποτέλεσμα της κατάτμηση μόνο με βάση την ένταση. Τα μέγιστα που δίνει ο αλγόριθμος είναι τα ίδια με αυτά που προκύπτουν από τον απλό k-means.

- **Κελί το οποίο περιέχει pixels από γειτονικό spot.**

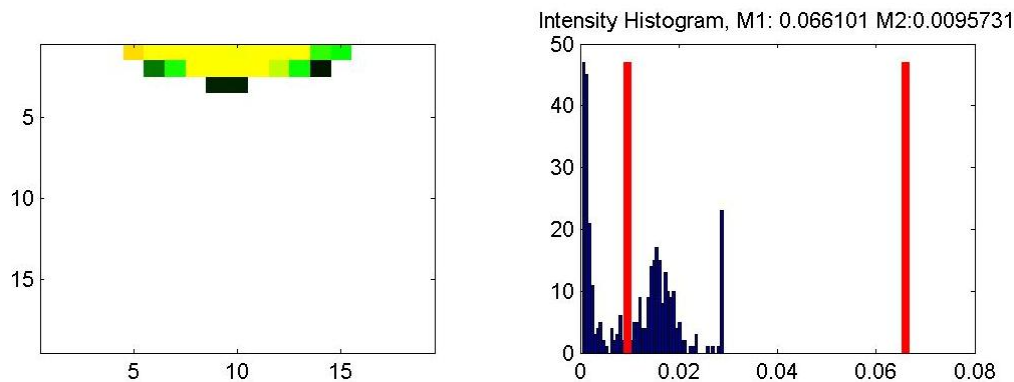
Στην περίπτωση αυτή φανταζόμαστε ότι ο αλγόριθμος θα χρειαστεί πάνω από μια επανάληψη για να «γεμίσει» το spot. Εφαρμόσαμε iterative k-means στο spot της Εικόνας 4.1.3 και τα αποτελέσματα των επαναλήψεων του αλγορίθμου ήταν τα παρακάτω:



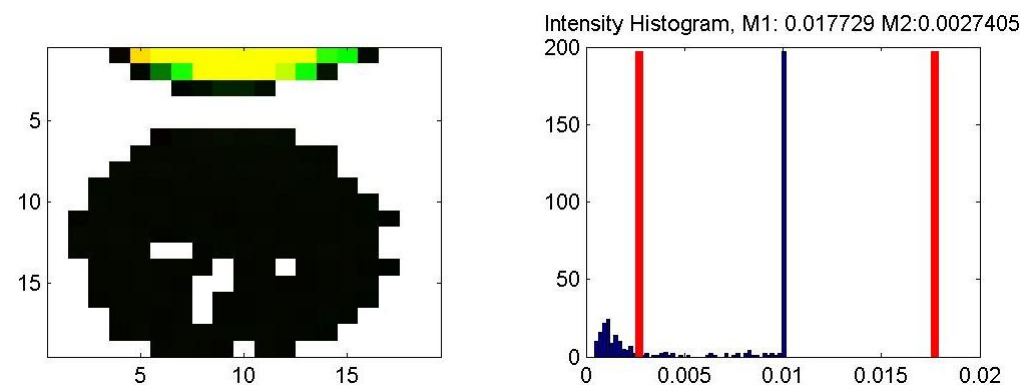
Εικόνα 4.3.2 Επαναληπτικός k-means – Pixel από γειτονικό spot - Πρώτη Επανάληψη. Το αποτέλεσμα της κατάτμηση στο πρώτο βήμα διαχωρίζει μόνο τέσσερα pixels του γειτονικού spot από το background, ότι και ο απλός k-means δηλαδή.



Εικόνα 4.3.3 Επαναληπτικός k-means – Pixel από γειτονικό spot – Δεύτερη Επανάληψη. Οι τιμές έντασης των προηγούμενων τεσσάρων Pixels τίθενται σε μια χαμηλότερη τιμή έντασης έτσι αρχίζουμε να έχουμε μεγαλύτερη ανάλυση στις χαμηλές τιμές έντασης και να διακρίνουμε τις κατανομές του ενδιαφέροντος μας.

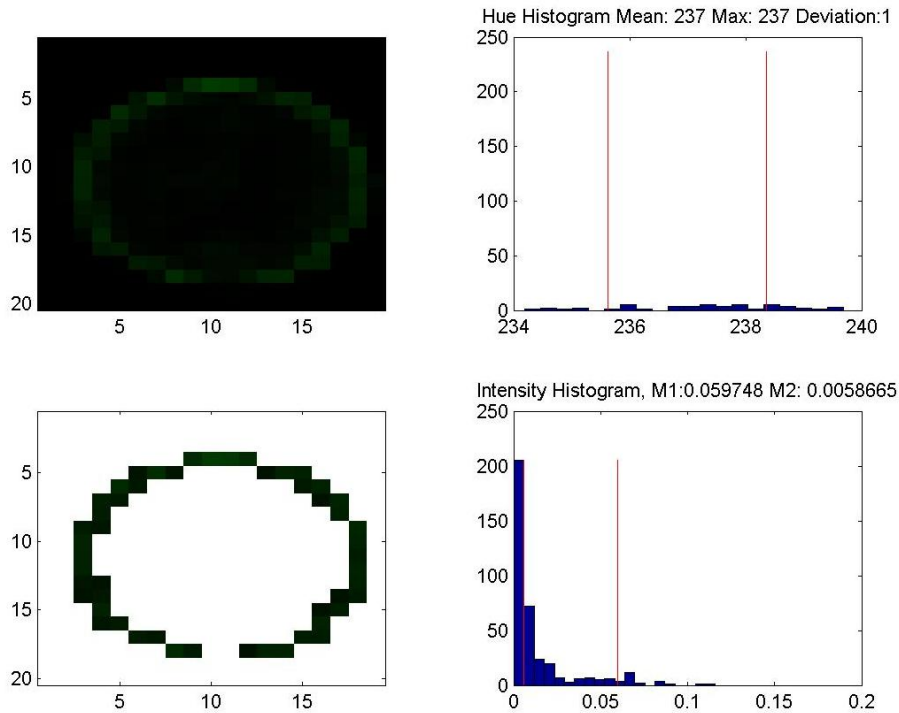


Εικόνα 4.3.4 Επαναληπτικός k-means – Pixel από γειτονικό spot – Τρίτη Επανάληψη k-means. Έχουμε πια μεγάλη ανάλυση γύρω στις τιμές έντασης που είναι πάνω από το μηδέν. Στην ουσία είναι σαν συμπύκνουμε όλα τα Pixels που ανήκουν σε spot αλλά έχουν μεγάλη διαφορά έντασης μεταξύ τους σε μια κατανομή.

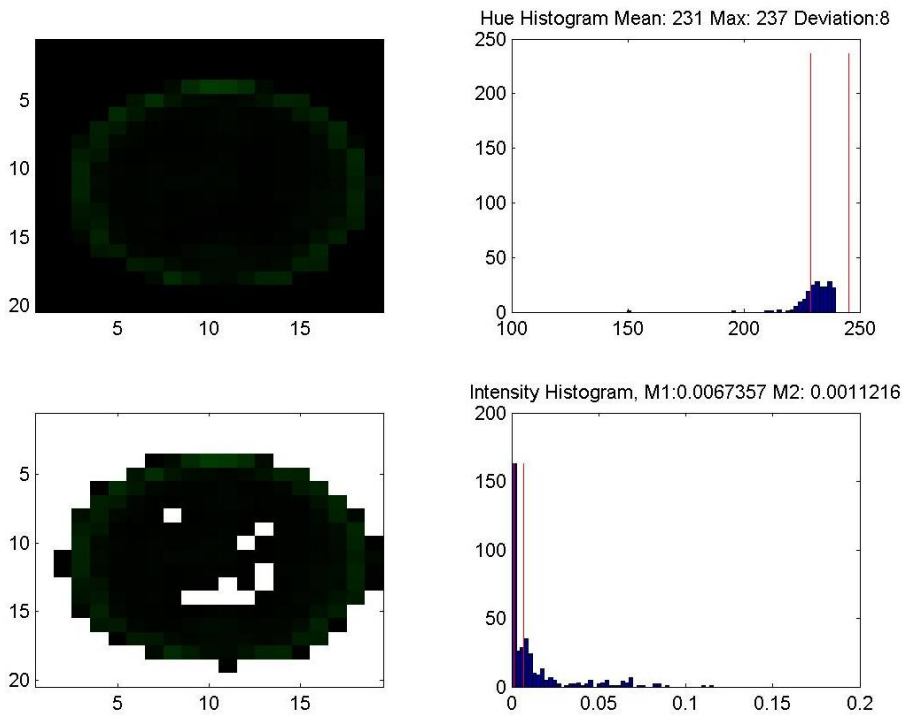


Εικόνα 4.3.5 Επαναληπτικός k-means – Pixel από γειτονικό spot – Τέταρτη Επανάληψη k-means. Ο αλγόριθμος τερματίζει και βλέπουμε ότι έχουμε ένα μέγιστο στο οποίο έχουμε συγκεντρώσει όλα τα Pixel που ανήκουν στο spot μέσω του υποβιβασμού που εξηγήσαμε στην ανάλυση του αλγορίθμου (ενότητα 3.1.2.3).

- **Αχνό Spot.** Εφαρμόζουμε επαναληπτική κατάτμηση με k-means σε ένα spot με αχνό εσωτερικό και σχετικά ευδιάκριτη περιφέρεια. Στην πρώτη εφαρμογή ο αλγόριθμος διαχωρίζει μόνο την περιφέρεια του spot από την εικόνα.



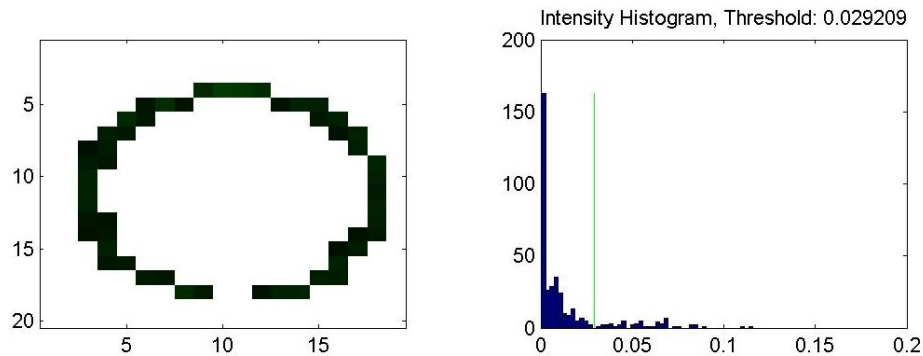
Εικόνα 4.3.6 Επαναληπτικός k-means – Spot με αχνό εσωτερικό – 1^η Επανάληψη. Τα περιφερειακά pixels και τα εσωτερικά pixels αποτελούν δύο ξεχωριστές κατανομές. Η απόσταση των αυτών των δύο κατανομών είναι μεγαλύτερη από την απόσταση της δεύτερης κατανομής από την κατανομή του background. Ο απλός k-means δίνει μόνο τα περιφερειακά Pixels ως αποτέλεσμα.



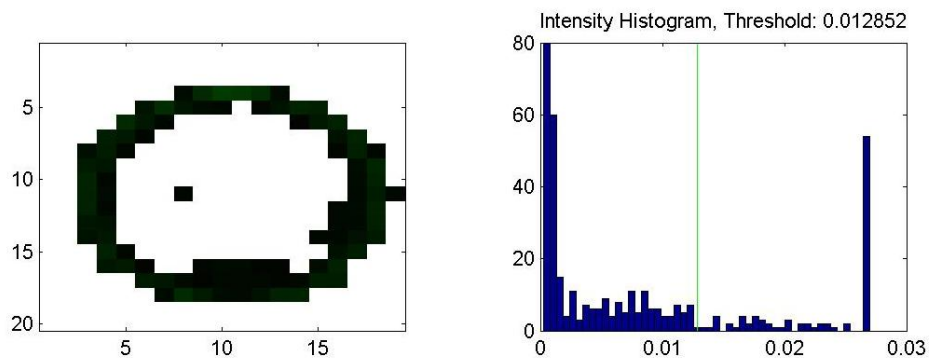
Εικόνα 4.3.7 Επαναληπτικός k-means – Spot με αχνό εσωτερικό – 3^η Επανάληψη. Το αποτέλεσμα του iterative k-means για το προηγούμενο spot. Τα περισσότερα pixels του spot συμπεριλήφθηκαν στο foreground.

4.4 Επαναληπτικός Otsu – Εφαρμογή

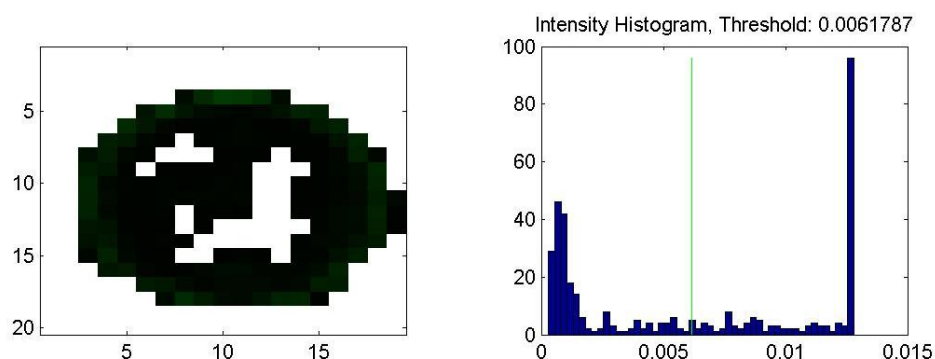
Εφαρμόζουμε τον αλγόριθμο Otsu στο παραπάνω κελί το οποίο περιέχει ένα αχνό spot. Ο απλό Otsu αλγόριθμος όπως και ο απλός k-means χωρίζει μόνο την περιφέρεια του spot από την υπόλοιπη εικόνα. Παρακάτω φαίνεται το αποτέλεσμα των τριών επαναλήψεων του επαναληπτικού Otsu.



Εικόνα 4.4.1 Επαναληπτικός Otsu – Βημα1. Το ίδιο spot το οποίο εξετάσαμε παραπάνω χρησιμοποιώντας k-means. Το αποτέλεσμα διαχωρισμού είναι ακριβώς το ίδιο με τον k-means με την διαφορά ότι αντί για μέσες τιμές έχουμε ένα threshold (πράσινη γραμμή).



Εικόνα 4.4.2 Επαναληπτικός Otsu – Βημα2. Προστίθεται άλλη μια κατηγορία από pixels στο foreground.

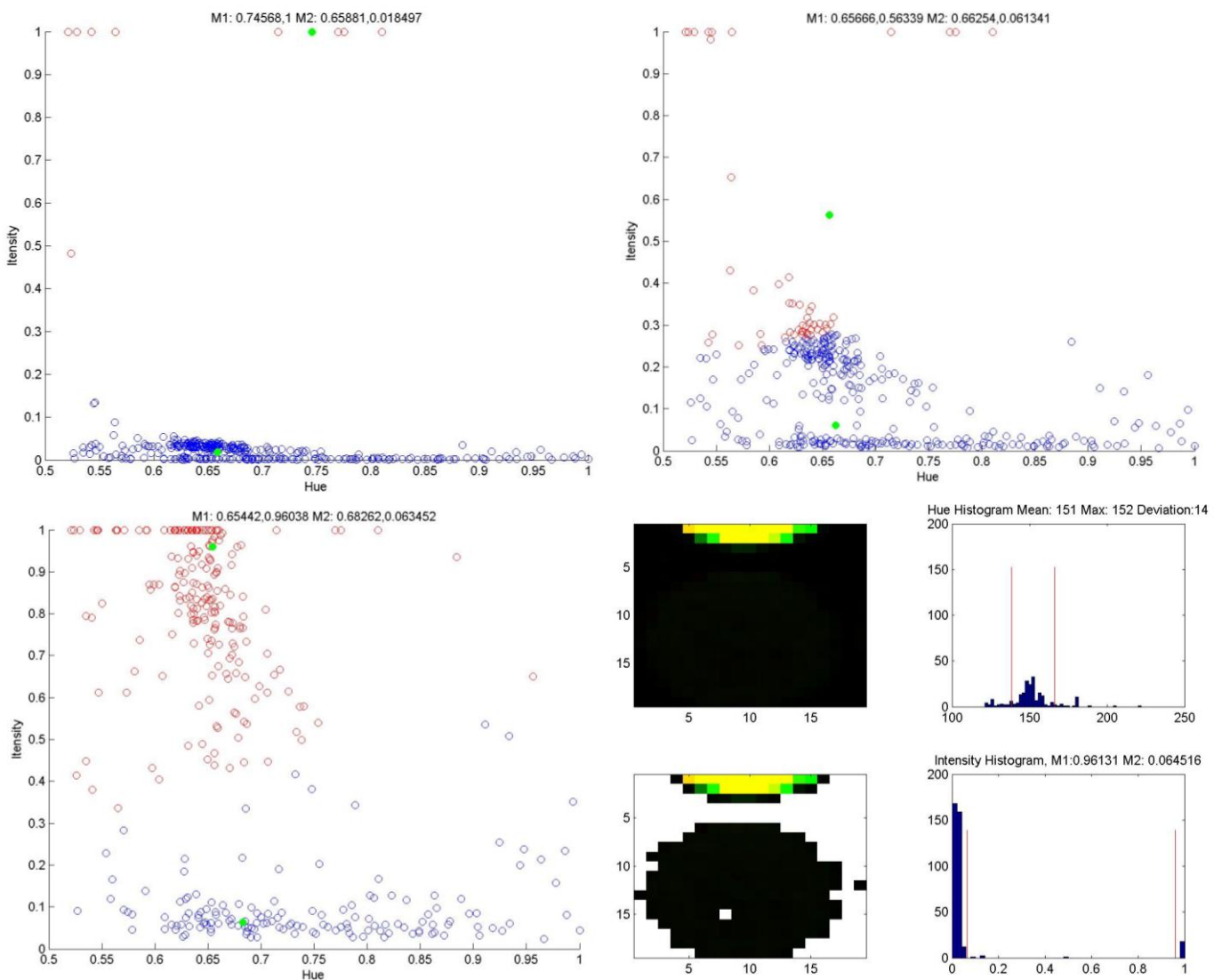


Εικόνα 4.4.3 Επαναληπτικός Otsu – Βημα3. Το τελικό αποτέλεσμα του επαναληπτικού Otsu παραλείπει μόνο μερικά pixels στο εσωτερικό του spot.

4.5 Επαναληπτικός δυσδιάστατος *k*-means

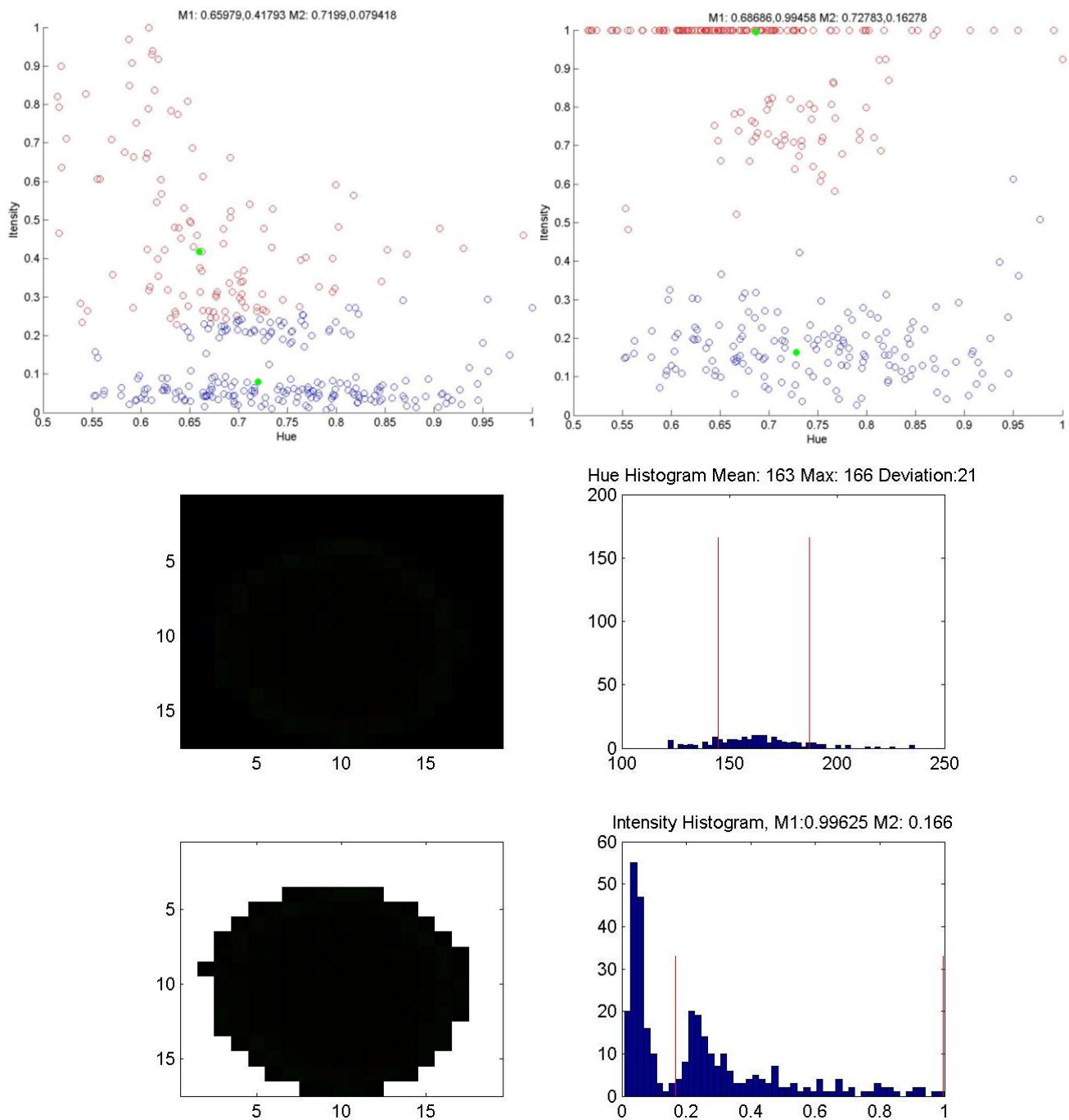
Επειδή έχουμε δυσδιάστατα δείγματα για την απεικόνιση των δύο κλάσεων χρησιμοποιούμε γραφήματα τα οποία μας δείχνουν την «κάτοψη» των κατανομών (Hue,Intensity). Με κόκκινο παριστάνεται η κλάση του foreground και με μπλε η κλάση του background, οι πράσινες κουκκίδες δείχνουν τα εκάστοτε «μέσα» των κατηγοριών. Εξετάζουμε τις ανάλογες περιπτώσεις με τους επαναληπτικούς αλγόριθμους Otsu και *k*-means.

- Κελί το οποίο περιέχει pixels από γειτονικό spot.



Εικόνα 4.5.1 Σύγκλιση Επαναληπτικού *k*-means δύο διαστάσεων– Τμήμα γειτονικού spot έχει συμπεριληφθεί μέσα στο κελί. Καθώς υποβιβάζουμε την ένταση των Pixels του foreground έχουμε μεγαλύτερη ανάλυση στις χαμηλές εντάσεις. Το αποτέλεσμα για το spot που εξετάσαμε και στους μονοδιάστατους αλγόριθμους. Συγκλίνει σε τρεις επαναλήψεις όπως και ο μονοδιάστατος *k*-means.

- **Αχνό Spot:** spot το οποίο έχει τιμές έντασης πολύ κοντινές με αυτές του background, στην ουσία η συνολική κατανομή αποτελείται από τρεις κατανομές, η μια περιγράφει τα περιφερειακά pixel και η άλλη το εσωτερικό του spot.



Εικόνα 4.5.3 Σύγκλιση Επαναληπτικού k-means δύο διαστάσεων- Αχνό spot. Ο επαναληπτικός k-means συγκλίνει σε 2 επαναλήψεις.

4.6 Σχόλια-Συμπεράσματα

- Ο k-means και γενικά οι αλγόριθμοι που χρησιμοποιούν ως μέτρο ομοιότητας των δειγμάτων την ευκλείδεια ή την mahalanobis απόσταση είναι ευμετάβλητοι στους γραμμικούς συνδυασμούς. Δηλαδή διάφορα scaling των αξόνων που προκύπτουν από γραμμικούς συνδυασμούς των αξόνων οδηγούν σε διαφορετικά αποτελέσματα ταξινόμησης.
Για να έχουμε συνεπώς τα επιθυμητά αποτελέσματα πραγματοποιήσαμε κανονικοποίηση των αξόνων πριν προχωρήσουμε στην διαδικασία ταξινόμησης.
- Όσον αφορά την προσθήκη του δεύτερου χαρακτηριστικού μπορούμε να αντιληφθούμε και από τα γραφήματα ότι **η δυσδιάστατη κατανομή αποτελείται από δύο κατανομές οι οποίες μοιάζουν να έχουν μικρή εξάρτηση ως προς την ένταση**.
- Επίσης τα δύο χαρακτηριστικά μοιάζουν να είναι ανεξάρτητα μεταξύ τους. Αυτό φαίνεται από τον πλήρη διαχωρισμό των υψηλών και χαμηλών εντάσεων ανεξάρτητα από την απόχρωση του hue. Δηλαδή παρόλο που δεν είναι ορατό με το μάτι στο background συναντάμε όλα τα χρώματα και αυτό δεν μας δίνει κάποια εξάρτηση που να λέει ότι ένα Pixel στο χρώμα κίτρινο για παράδειγμα συναντάτε μόνον στις υψηλές εντάσεις.
Συνεπώς η ακολουθιακή ταξινόμηση (ενότητα 3.2.2) αποτελεί μια απλοποίηση που δεν εισάγει μεγάλο σφάλμα στην ταξινόμηση.
- Επιπλέον το χαρακτηριστικό της απόχρωσης δεν προσφέρει σημαντική πληροφορία για τον διαχωρισμό του spot από το background. Στην ουσία μας βοηθάει μόνο στην αφαίρεση των pixel που προκύπτουν από την λάθος ευθυγράμμιση των εικόνων Cy3, Cy5.
- **Τέλος η μετατροπή από το σύστημα RGB στο HSI έχει μειώνει την τάξη του προβλήματος κατά μία διάσταση.** Ενώ στο RGB έχουμε τουλάχιστον δύο διαστάσεις μία για την ένταση του κόκκινου και μία για την ένταση του πράσινου, στο HSI έχουμε μόνο μια τιμή της έντασης. Αυτό μειώνει την πολυπλοκότητα των αλγορίθμων και μας δίνει την δυνατότητα να χρησιμοποιήσουμε την πληροφορία που μας δίνει η απόχρωση του κάθε Pixel.

4.6.1 Σύγκριση Απλού - Επαναληπτικού Αλγόριθμου

Όπως προέκυψε από την σύγκριση των επαναληπτικών και των απλών αλγορίθμων οι επαναληπτικοί αλγόριθμοι επιλύουν σε μεγάλο βαθμό τα προβλήματα τα οποία οφείλονται κυρίως στην απλοποίηση του προβλήματος και στην παραδοχή ότι έχουμε μόνο δύο κατηγορίες ($k=2$).

Παρόλο που υπάρχει μεγάλη βελτίωση στην κατάτμηση των κελιών, όπως μπορούμε να δούμε, το αποτέλεσμα της κατάτμησης παραβλέπει μια ομάδα από pixels τα οποία βρίσκονται στο εσωτερικό του spot και έχουν πολύ χαμηλή ένταση. Το πρόβλημα αυτό το επιλύουμε με ένα απλό αλγόριθμο συμπλήρωσης ο οποίος γεμίζει το εσωτερικό του spot στα σημεία που έχουν παραληφθεί pixels (ενότητα 4.7).

4.6.2 Σύγκριση Μονοδιάστατης και Δυσδιάστατης προσέγγισης

Ο δυσδιάστατος k-means δίνει αρκετά καλά αποτελέσματα αλλά παρόλο που έχουμε και την πληροφορία για την Απόχρωση ο αλγόριθμος συμπεριλαμβάνει και τα pixel από την λάθος ευθυγράμμιση των δύο εικόνων. Σε αυτήν την περίπτωση δηλαδή η απλοποίηση όπου θεωρούμε ότι έχουμε δύο κατηγορίες δεν δουλεύει καλά. Συνεπώς τα μειονεκτήματα του δυσδιάστατου k-means σε σύγκριση με τον απλό είναι:

- **Επαναλήψεις** για την σύγκλιση
- Μη εκμετάλλευση κατ' ουσία του χαρακτηριστικού της απόχρωσης.

Η αποτυχία του αλγορίθμου να αξιοποιήσει την πληροφορία που παρέχει το χαρακτηριστικό της απόχρωσης έχει να κάνει με την αξία του κάθε χαρακτηριστικού. Προφανώς η ένταση παίζει πολύ μεγαλύτερο ρόλο απ' ό τι η απόχρωση του spot. Η απόχρωση έχει συμπληρωματικό έννοια και η αξιοποίηση της πληροφορίας που μας δίνει αποκτά σημασία μόνο στην περίπτωση όπου έχουμε κακή ευθυγράμμιση των δύο εικόνων (Cy3,Cy5).

Όπως φαίνεται αλώςτε και από την κάτοψη της δυσδιάστατης κατανομής έχουμε δύο βασικές κατανομές που απλώνονται κατά μήκος όλων των χρωμάτων, η μία στις υψηλές εντάσεις (πάνω) και η άλλη στις χαμηλές (κάτω).

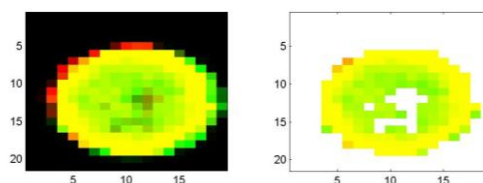
Δοκιμάσαμε ευκλείδεια mahalanobis και γωνιακή απόσταση ως μέτρο απόστασης των σημείων από τα μέσα των κατανομών. Με την ευκλείδεια απόσταση ο αλγόριθμος τείνει να συμπεριλάβει τα δείγματα μέσα σε κύκλους με αποτέλεσμα την συμπερίληψη στο foreground πολλών pixel από το background. Με την mahalanobis και την γωνιακή απόσταση είχαμε καλύτερη εκτίμηση των κατανομών αφού αυτές απλώνονται κατά μήκος του άξονα της έντασης και έχουν ελλειψοειδής μορφή. Οι mahalanobis η γωνιακή απόσταση τείνουν να συμπεριλάβουν τα δείγματα μέσα σε ελλείψεις και όχι κύκλους.

Η ακολουθιακή κατάτμηση από την άλλη έχει τα αναμενόμενα αποτελέσματα δηλαδή αφαιρεί από το spot τα μονοχρωματικά pixel.

4.7 ΔΙΑΔΙΚΑΣΙΑ ΣΥΜΠΛΗΡΩΣΗΣ SPOT (Center filling)

Στις προηγούμενες ενότητες αναλύσαμε και εφαρμόσαμε κάποιους αλγορίθμους κατάτμησης πάνω σε εικόνες μικροσυστοιχιών. Όπως είναι οπτικά εμφανές όλοι οι αλγόριθμοι ακόμα και με τις διορθώσεις και τις επαναληπτικές εφαρμογές τους στις περισσότερες περιπτώσεις παραλείπουν κάποια pixel στο εσωτερικό του spot. Αυτό γίνεται επειδή τα περιφερειακά pixel του spot έχουν μεγάλη διαφορά έντασης με το τα εσωτερικά ή κάποια διαφοροποίηση στο χρώμα.

Κατά την διαδικασία της κατάτμησης (4.1) συναντήσαμε το πρόβλημα όπου το ιστόγραμμα της έντασης δεν απαρτίζεται από δύο καμπύλες σύμφωνα με την προσέγγιση που κάναμε, μια για το background και μια για το foreground, αλλά από τουλάχιστον τρεις κατανομές. Τότε είχαμε αναφέρει ότι οι ελάχιστες κατηγορίες είναι: τα περιφερειακά pixel του spot, τα εσωτερικά pixel του spot και το background. Μια περίπτωση φαίνεται στην εικόνα 4.7.1 όπου παρόλο που το spot είναι ευδιάκριτο κάποια pixel στο εσωτερικό του δεν συμπεριλήφθηκαν λόγω χρώματος.



Εικόνα 4.7.1 Όλοι οι αλγόριθμοι κατάτμησης (στην συγκεκριμένη εικόνα το αποτέλεσμα προήλθε από την εφαρμογή του αλγορίθμου k-means) δεν συμπεριλαμβάνουν στο foreground ένα σύνολο από pixels στο εσωτερικό του spot, τα οποία έχουν χαμηλές τιμές έντασης. Ο αλγόριθμος συμπλήρωσης του κέντρου του spot προσθέτει αυτά τα pixel στο σύνολο των Pixels που ανήκουν στο spot οριοθετώντας έτσι το χώρο του spot με. Με βάση τον θεωρητικό χώρο του spot απαλείφουμε ομάδες από σημεία που δεν ανήκουν στο spot λόγω θέσης.

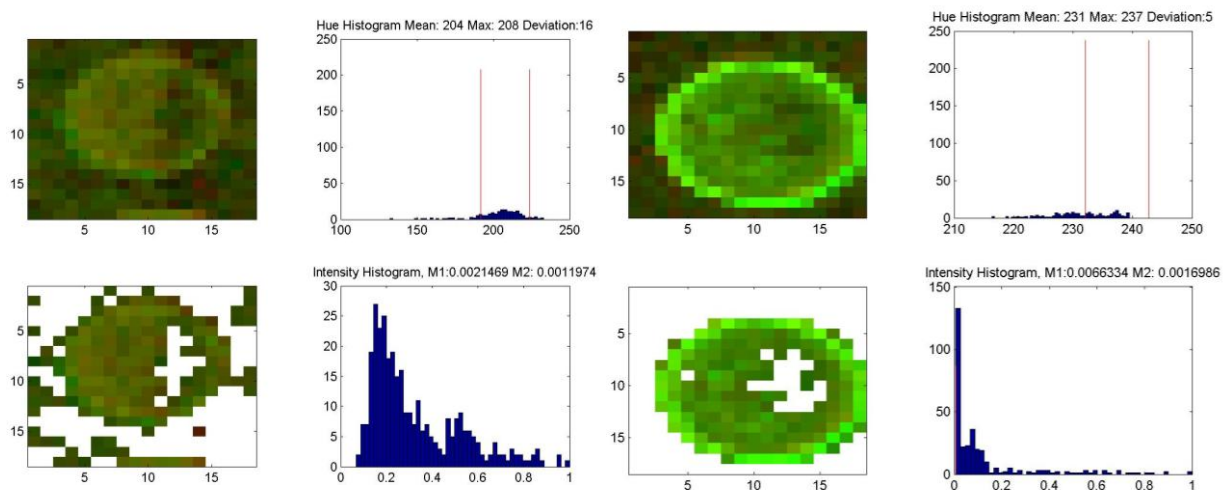
Το φαινόμενο της μεγάλης διαφοράς έντασης και χρώματος ανάμεσα στα εσωτερικά και περιφερειακά pixel του spot μπορεί να εξηγηθεί μέσω της φυσικής. Επειδή η ουσία μας είναι σε υγρή μορφή λόγω των δυνάμεων συνάφειας τείνει να πάρει σφαιρική μορφή, μορφή σταγόνας. Μια επιφάνεια που δεν είναι επίπεδη αλλά έχει καμπύλη μορφή δεν αντανακλά με τον ίδιο τρόπο σε όλα τα σημεία της τις ακτίνες φωτός. Τα περιφερειακά σημεία της σταγόνας επειδή είναι σχεδόν κάθετα με τις ακτίνες φωτός αντανακλούν με μεγαλύτερη ένταση. Αυτό σημαίνει ότι τα εσωτερικά σημεία του spot δίνουν χαμηλότερες τιμές έντασης απ' ό τι οι πραγματικές.

Θα περίμενε κανείς ότι όλα τα pixel που ανήκουν στο spot συμβάλλουν στην καλύτερη εκτίμηση του ποσοστού έκφρασης. Τα pixel όμως που βρίσκονται στο εσωτερικού του spot για τους λόγους που προαναφέραμε έχουν χαμηλή ένταση και διαφοροποιούνται στο χρώμα με αποτέλεσμα τα εισάγουν μεγαλύτερο σφάλμα από τα υπόλοιπα σημεία του spot. Συνεπώς αναμένουμε ο συνυπολογισμός αυτών των σημείων στην εξαγωγή ποσοστού έκφρασης να δώσει μεγαλύτερα σφάλματα.

Από την άλλη υπάρχει η αντίθετη περίπτωση όπου κάποια σημεία στην περιφέρεια του spot ταξινομούνται λανθασμένα στο spot για δύο λόγους: είτε λόγω χαμηλής έντασης του spot (αχνό spot) είτε επειδή ανήκουν σε γειτονικό spot αλλά βρίσκονται μέσα στο κελί λόγω κακής τοποθέτησης του πλέγματος.

Ο στόχος μας είναι να παραλείψουμε αυτά τα σημεία από το κλάση του spot κατά συνέπεια να αφαιρέσουμε από το αποτέλεσμα της κατάτμησης κάποια σύνολα από pixel. Για να το επιτύχουμε αυτό υλοποιούμε ένα αλγόριθμο ο οποίος μας δίνει μια εκτίμηση του χώρου του spot όπως θα έπρεπε να είναι και βάση αυτής της εκτίμησης κρατάμε μόνο τα σημεία που μας δίνουν πληροφορία για το spot.

Ακολουθεί η περιγραφή του αλγορίθμου εύρεσης του κέντρου του spot και στην συνέχεια οι αλγόριθμοι συμπλήρωσης του spot και αφαίρεσης των περιφερειακών pixel.



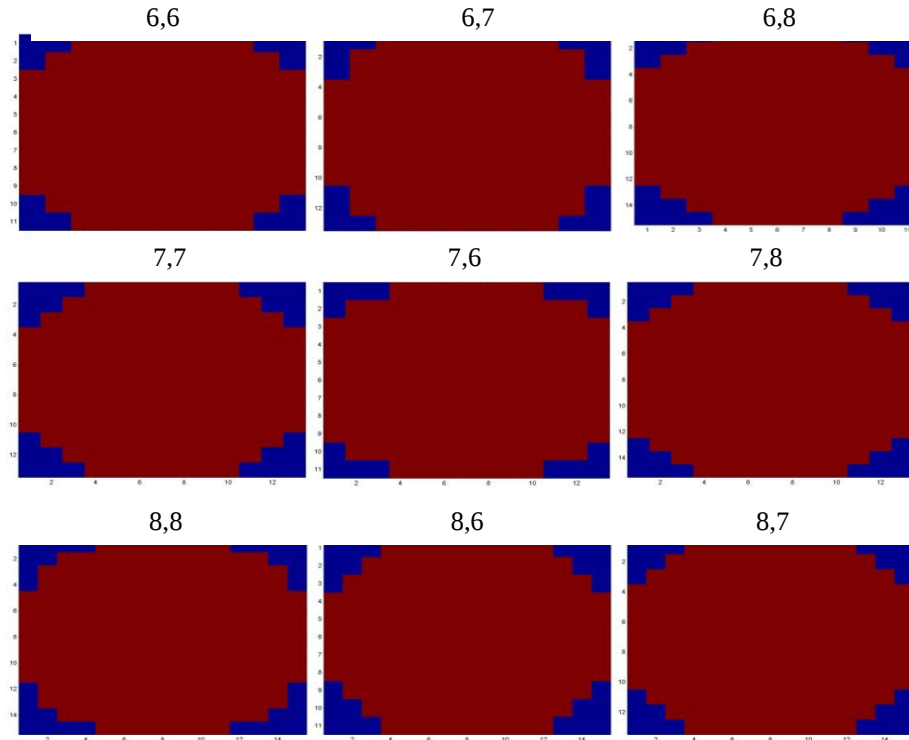
Εικόνα 4.7.2 Τα παραπάνω spot απεικονίζονται αφού έχουμε κάνει μια ενίσχυση της αντίθεσης του spot. **Αριστερά:** Το συγκεκριμένο spot έχει χαμηλές τιμές έντασης δηλαδή είναι αχνό. Στο αποτέλεσμα της κατάτμησης συμπεριλαμβάνονται pixel στην περιφέρεια από το background. **Δεξιά:** Το spot αυτό έχει ένα μικρό αριθμό από Pixel στο εσωτερικό του που έχουν χαμηλή ένταση. Τα σημεία αυτά παραλείπονται από τους αλγόριθμους κατάτμησης.

4.7.1 Υπολογισμός της θέσης ενός spot

Όλα τα pixel τα οποία ανήκουν στο background σύμφωνα με το αποτέλεσμα της κατάτμησης είναι υποψήφια για να συμπεριληφθούν στο spot. Επειδή όμως το spot αν έχει γίνει σωστά η τοποθέτηση πλέγματος (gridding) βρίσκεται στο κέντρο του κελιού στην ουσία τα υποψήφια pixel είναι αυτά που έχουν ταξινομηθεί στο background αλλά βρίσκονται σε μια κυκλική περιοχή γύρω από το κέντρο του spot.

Για να εντοπίσουμε τα υποψήφια προς «γέμισμα» pixel αρκεί να βρούμε αυτόν τον κυκλικό χώρο στον οποίο βρίσκεται το spot μας. Ο κυκλικός αυτός χώρος έχει την μορφή και τις διαστάσεις ενός τυπικού spot των εικόνων μας. Για την τοποθέτηση αυτού του χώρου πάνω στο κελί μας χρειάζεται να ξέρουμε τη θέση του κέντρου του spot μας.

Η διαδικασία εύρεσης του κέντρου του spot βασίζεται σε απλό υπολογισμό του κέντρου βάρους της εικόνας του foreground που προκύπτει από την διαδικασία της κατάτμησης. Βασίζεται στην άθροιση του αριθμού των pixels που ανήκουν στο foreground και βρίσκονται μέσα σε μια κυκλική ή ελλειψοειδής



Εικόνα 4.7.1.1 Όλες οι δυνατές μάσκες για $\alpha=6:8$ και $\beta=6:8$. Συνελίσσουμε το αποτέλεσμα της κατάτμησης με όλες τις παραπάνω μάσκες και βρίσκουμε τα αντίστοιχα κέντρα μάζας. Το τελικό κέντρο μάζας ή μέση τιμή όλων των κέντρων μάζας.

περιοχή την οποία ορίζουμε εμείς και αποτελεί την μάσκα μας. Τα μεγέθη των spot ποικίλουν από ακτίνα ίση με 6 έως 8 περίπου. Για τον λόγο αυτό παράγουμε διάφορες μάσκες κυκλικές και ελλειψοειδείς και σε διαφορά μεγέθη. Παρατηρήσαμε ότι τα spot κατά μεγάλο ποσοστό έχουν ελλειψοειδής μορφή και συγκεκριμένα με την οριζόντια διάσταση πιο μεγάλη ($\alpha > \beta$) στις περισσότερες περιπτώσεις. Για τον λόγο αυτό ετοιμάσαμε κυκλικές μάσκες με $r=6,7,8$ και ελλείψεις για όλους τους συνδυασμούς μεγάλου και μικρού άξονα $[\alpha, \beta]$ ($\alpha, \beta = 6,7,8$).

Βρίσκουμε το ποσοστό επικάλυψης κάθε μάσκας με το foreground ώστε να επιλέξουμε ως κέντρο βάρους του κελιού το σημείο που έχουμε μέγιστη επικάλυψη. Για τον υπολογισμό του ποσοστού επικάλυψης εφαρμόζουμε συνέλιξη ανάμεσα στο αποτέλεσμα της κατάτμησης (ένας πίνακας με μηδέν και ένα, ένα στις θέσεις όπου έχουμε pixel που ανήκουν στο foreground) και σε κάθε μια από τις μάσκες. Το αποτέλεσμα της συνέλιξης κανονικοποιείται από 0 έως 1 για να μπορέσει να γίνει η σύγκριση ανάμεσα στις μάσκες. Για κάθε μάσκα επιλέγουμε την θέση που μας δίνει μέγιστο ποσοστό επικάλυψης με το foreground-spot, Δηλαδή το σημείο που η συνέλιξη μας δίνει μέγιστο και αυτό είναι το κέντρο βάρους όπως το υπολογίζει η συγκεκριμένη μάσκα. Εν τέλει το κέντρο βάρους στο κελί μας είναι αυτό που προκύπτει από την μάσκα που δίνει μέγιστο ποσοστό επικάλυψης.

```

function [bestci,bestcj,besta,bestb]=FindCenter(foreground)
%Παραγωγή όλων των μάσκων
s=6;
e=8;
bestscore=-1;
bestci=-1;bestcj=-1;
for a=[s:e]
    for b=[s:e]
        mask = ellipsis(a,b);    % κατασκευή έλλειψης με μεγάλο άξονα α και μικρό β
        t=conv(foreground,mask);

        [ci,cj]=index(max(t));    % θέση της μέγιστης επικάλυψης της μάσκας με το spot
        score=max(t)/(4*a*b);    % ποσοστό επικάλυψης
        if(score>bestscore)    % επιλογή της μάσκας με την μέγιστη επικάλυψη
            besta=a;
            bestb=b;
            bestscore=score;
            bestci=ci;
            bestcj=cj;
        end
    end
end
end

```

Αλγ 4.7.1 Ο αλγόριθμος υπολογισμού του κέντρου του spot. Η συνάρτηση `ellipsis` μας δίνει ένα δυδιάστατο πίνακα με μηδενικά και άσσους σε μορφή έλλειψης (ή κύκλου αν $a=b$). Τα a, b περιγράφουν την έλλειψη μας (μεγάλος μικρός άξονας). Η Μεταβλητή `score` αποθηκεύει το ποσοστό μεγαλύτερης επικάλυψης. Επιλέγουμε την μάσκα με το μεγαλύτερο ποσοστό επικάλυψης.

4.7.2 Αλγόριθμος Συμπλήρωσης του spot

Έχοντας υπολογίσει τη θέση του spot μέσα στο κελί, δηλαδή το κέντρο του θέλουμε να υλοποιήσουμε έναν αλγόριθμο ο οποίος θα γεμίζει το εσωτερικό του spot. Δηλαδή ο αλγόριθμος αυτός εντοπίζει pixels που βρίσκονται στο κέντρο του spot και τα γεμίζει.

Όπως αναφέραμε και στην ενότητα 4.7.1 παραπάνω όλα τα pixel που ανήκουν στο background είναι αρχικά υποψήφια προς «γέμισμα». Επειδή όμως στην πραγματικότητα τα pixel που θέλουμε να γεμίσουμε βρίσκονται στο κέντρο του spot, τα κατ' ουσία υποψήφια pixel είναι αυτά που έχουν ταξινομηθεί στο background και βρίσκονται μέσα στην περιοχή του spot. Η περιοχή του spot ορίζεται με βάση το κέντρο του spot που έχουμε ήδη υπολογίσει. Είναι πολύ απλά η θέση της μάσκας πάνω στο κελί η οποία δίνει το μέγιστο άθροισμα σύμφωνα με τον αλγόριθμο υπολογισμού του κέντρου του spot.

Χρησιμοποιούμε **δύο κριτήρια** τα οποία αν ικανοποιούνται ταυτόχρονα τότε «γεμίζουμε» το υποψήφιο pixel. Το **πρώτο κριτήριο** έχει να κάνει με το αν το pixel βρίσκεται στην περιοχή του spot. Οριοθετούμε τα pixel που βρίσκονται μέσα στο spot με την βοήθεια του υπολογισμένου κέντρου βάρους του κελιού (ενότητα 4.7.1) και τις μάσκας που επιλέχθηκε ως αυτή με την μέγιστη επικάλυψη (από εικόνα 5.1.1). Το **δεύτερο κριτήριο** έχει να κάνει με την ικανοποίηση ενώ κατώτερου ορίου (`downlimit`). Το όριο αυτό αφορά τον αριθμό των γειτόνων του υποψήφιου pixel που ανήκουν στο foreground.

Ως γειτονικά pixel ορίζουμε αυτά που βρίσκονται σε μια κυκλική περιοχή γύρω από το σημείο μας. Η περιοχή αυτή ορίζεται μέσω μιας μάσκας (εικόνα 5.1.5) η οποία εφαρμόζεται στα υποψήφια pixel.



Εικόνα 4.7.2.1 Μάσκα συμπλήρωσης του spot. Η παραπάνω μάσκα ορίζει τα γειτονικά σημεία ενός pixel. Είναι περιττού βαθμού καθώς ορίζει ένα κέντρο, το υποψήφιο προς γέμισμα pixel. Δοκιμαστικά καταλήξαμε σε ένα μέγεθος 5x5 το οποίο έδωσε καλύτερο τελικό αποτέλεσμα.

Η τιμή του ορίου προέκυψε πειραματικά ίση με 11. Το μέγεθος της μάσκας είναι ίσο με $5 \times 5 = 25$ αλλά ο αριθμός των γειτονικών pixel που ορίζει είναι $25 - 4 = 21$. Άρα το ποσοστό των γειτονικών pixel που θέλουμε να είναι «γεμάτα» προκειμένου να γεμίσουμε το υποψήφιο pixel είναι τουλάχιστον $11/21 = 52\%$.

Συνεπώς ο αλγόριθμος γεμίζει τα «άδεια» pixel τα οποία βρίσκονται μέσα στο spot και ο αριθμός των γειτονικών Pixel που ανήκουν στο foreground ικανοποιούν ένα κάτω όριο.

4.7.3 Αφαίρεση περιφερειακών pixel

Προχωράμε τώρα στο σημαντικότερο κομμάτι που είναι το αντίστροφο πρόβλημα όπου κάποια Pixel τα οποία βρίσκονται στην περιφέρεια του κελιού και δεν ανήκουν στο spot τα οποία θέλουμε να αφαιρέσουμε από το σύνολο των pixels του foreground.

Τέτοιες περιπτώσεις προκύπτουν:

- Όταν το spot είναι πολύ **αχνό**, δηλαδή οι τιμές έντασης του spot είναι πολύ κοντινές με αυτές του background. Αυτό έχει ως αποτέλεσμα κατά την διάρκεια της ταξινόμησης των pixel κάποια από αυτά τα οποία βρίσκονται στη περιφέρεια του κελιού να ταξινομούνται στο spot λόγω τιμής έντασης
- Λόγω **κακής τοποθέτησης του πλέγματος** (gridding) τμήματα γειτονικών spot έχουν συμπεριληφθεί σε κάποια κελιά με αποτέλεσμα το αποτέλεσμα της κατάτμησης να περιέχει το βασικό spot αλλά και τμήματα άλλων spot συνήθως στην περιφέρεια του κελιού.
- **Κενό spot** έχουμε διάσπαρτα Pixels τα οποία δεν σχηματίζουν την εικόνα κάποιου spot.

Όσον αφορά την δεύτερη περίπτωση όπως αναφέραμε και στην ενότητα 10.2 υπάρχει μια μετατόπιση του κέντρου βάρους του spot προς το γειτονικό spot γεγονός το οποίο εισάγει μια μετατόπιση στον υπολογισμό του κέντρου του spot. Η αφαίρεση των περιφερειακών pixel του κελιού που δεν ανήκουν στο spot διορθώνει το πρόβλημα αυτό.

Ο αλγόριθμος ο οποίος αφαιρεί τα Pixel τα οποία βρίσκονται στο χώρο γύρω από το spot, δηλαδή τα σημεία που ανήκουν στο background και έχουν συμπεριληφθεί λανθασμένα στο spot είναι παρόμοιος με τον αλγόριθμο συμπλήρωσης. Χρησιμοποιεί τις ίδιες μάσκες και ένα ανώτατο όριο (UpLimit) αντί για κατώτερο.

Τα υποψήφια προς «άδειασμα» pixels είναι αυτά που ανήκουν στο foreground αλλά δεν βρίσκονται μέσα στην περιοχή του spot. Χρησιμοποιούνται και πάλι δύο κριτήρια για το «άδειασμα» ενός spot. Πρώτον εάν το pixel βρίσκεται έξω από τα όρια του spot, το οποίο ελέγχεται πάλι με την εύρεση του κέντρου του spot και δεύτερον εάν το άθροισμα των γειτονικών pixel που ανήκουν στο foreground είναι το πολύ ίσα με ένα όριο.

Το ανώτερο όριο βρέθηκε και πάλι πειραματικά ίσο με 8 δηλαδή ίσο με το $8/21=38\%$ του αριθμού

```
function [newsignal]=BgUnfillFgFill(foreground,sub_row,sub_col,uplimit,downlimit,wd,smask)
%     sub_row και sub_col δίνουν τις διαστάσεις του κελιού
%     downlimit είναι το όριο για το γέμισμα του spot ίσο με 11
%     uplimit το όριο για άδειασμα του spot ίσο με 8
%     smask η μάσκα που ορίζει τους γείτονες ενός pixel και wd το μέγεθος της

while 1
    [ci cj a b]=FindCenter(foreground);
    mask = ellipsis(a,b);    % κατασκευή έλλειψης με μεγάλο άξονα α και μικρό β (μάσκα με μέγιστη
    επικάλυψη)
    spotmasksmall = mask;
    spotmask=zeros(sub_row,sub_col);
    spotmask(ci,cj)=1;

    spotmask=conv2(spotmask,spotmasksmall); % κατασκευή ενός κελιού που περιέχει την μάσκα που ορίζει το
    χώρο του spot τοποθετημένο στο σημείο του κέντρου του
    spot
    neighbours=conv2(foreground,smask); % ο αριθμός των γειτόνων για κάθε pixel

    filled=neighbours;    %     Εντοπισμός των pixel που ικανοποιούν το κατώτερο αριθμό γειτόνων
    filled(filled<=downlimit)=0;
    filled(filled>0)=1;
    unfilled=neighbours;    %     Εντοπισμός των pixel που ικανοποιούν το ανώτερο αριθμό γειτόνων
    unfilled(unfilled<=uplimit)=0;
    unfilled(unfilled>0)=1;
    filled=filled & spotmask;    %     γέμισμα των pixel που βρίσκονται στο spot και ικανοποιούν το
    κατώτερο όριο γειτόνων
    unfilled=unfilled | spotmask;    %     γέμισμα των pixel που βρίσκονται εκτός spot και ικανοποιούν το
    ανώτερο όριο γειτόνων

    newsignal= foreground | filled;
    newsignal=newsignal & unfilled;

    if(sum(abs(newsignal-signal))<(sub_row*sub_col)*0.01)
        signal=newsignal;
        break;
    end;
    signal=newsignal;
end
```

Αλγ 4.7.3 Ο αλγόριθμος συμπλήρωσης του spot και αφαίρεσης των γειτονικών Pixel. Ο αλγόριθμος επαναλαμβάνεται μέχρι να συγκλίνει. Κριτήριο σύγκλισης είναι η μεταβολή στο σύνολο των pixel που ανήκουν στο foreground να είναι μικρότερη από 1%.

των γειτονικών Pixel.

Τα δύο όρια downlimit και uplimit αθροίζουν στο 90% αφήνοντας ένα παράθυρο 10% για την διάκριση των δύο περιπτώσεων (ανήκει στο background και δεν ανήκει).

4.7.4 Τελικός Αλγόριθμος

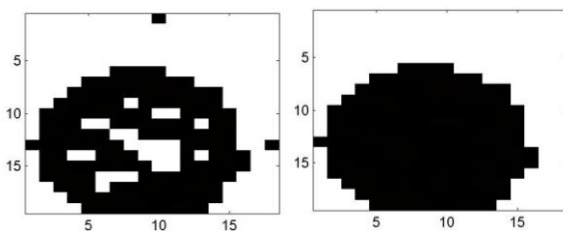
Στην τελική του μορφή η διαδικασία πλήρωσης του εσωτερικού του spot και αφαίρεσης των pixels γειτονικών spot υλοποιήθηκε ως ένας ενιαίος επαναληπτικός αλγόριθμος ο οποίος «γεμίζει» αλλά

και παράλληλα «αδειάζει» pixels με υπολογισμό της νέας θέσης του κέντρου του spot σε κάθε επανάληψη.

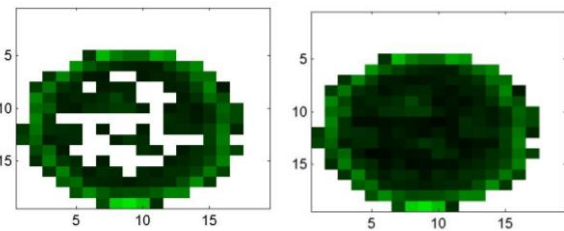
4.7.5 Εφαρμογή αλγορίθμου

Μερικές εφαρμογές του αλγορίθμου πάνω στις εξής περιπτώσεις:

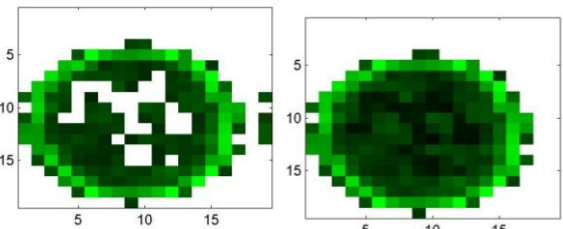
- Πολύ **αχνό spot**: Όπως αναφέραμε παραπάνω σε αυτήν την περίπτωση οι τιμές έντασης του spot έχουν μικρή διαφορά από τις τιμές έντασης του background (πολύ κοντά στο μηδέν).
- Μεγάλη διαφορά **έντασης ανάμεσα στα περιφερειακά και τα εσωτερικά pixel** του spot.
- **Κακή τοποθέτηση πλέγματος**: στο spot έχουν συμπεριληφθεί pixel από γειτονικά spot.
- **Κενό spot**: στην πραγματικότητα δεν υπάρχει spot με αποτέλεσμα η κατάτμηση τα μας δώσει διάσπαρτα pixel.



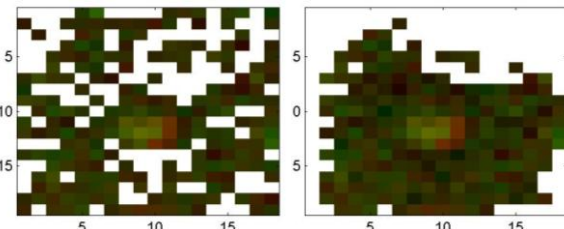
Εικόνα 4.7.5.1 Αχνό spot. Ένας αριθμός από εσωτερικά pixel του spot έχουν παραληφθεί από το spot.



Εικόνα 4.7.5.2 Μεγάλη διαφορά έντασης ανάμεσα στα εσωτερικά και στα περιφερειακά pixel του spot. Και σε αυτήν την περίπτωση έχουν παραληφθεί κάποια εσωτερικά pixel του spot αλλά εδώ οφείλεται στο γεγονός ότι το εσωτερικό του spot έχει πολύ «σκοτεινά» pixel.



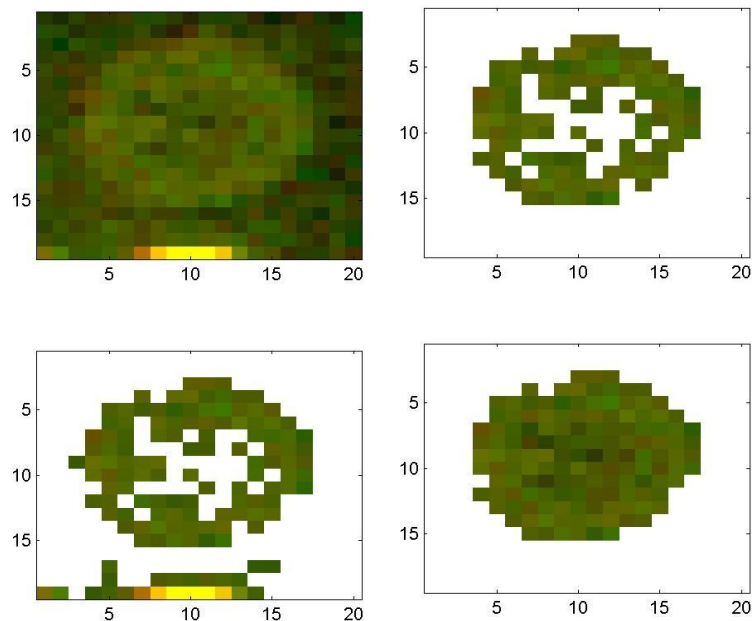
Εικόνα 4.7.5.3 Κακή τοποθέτηση πλέγματος. Τέσσερα pixel από το spot που βρίσκεται στα δεξιά του spot του ενδιαφέροντος μας έχουν συμπεριληφθεί στο foreground.



Εικόνα 4.7.5.4 Κενό spot. Κάποιο σφάλμα στην εναπόθεση του υλικού έχει ως αποτέλεσμα να μην έχουμε spot. Η κατάτμηση δίνει το παραπάνω αποτέλεσμα που περιέχει διάσπαρτα pixel και το αποτέλεσμα του center fill «σουλупώνει» κάπως την εικόνα αλλά χωρίς να έχει κάποια εφαρμογή, στο spot απορρίπτεται ως άδειο.

Στην τελευταία περίπτωση όπου έχουμε κενό spot χρησιμοποιώντας το αποτέλεσμα του ενιαίου αλγορίθμου και συγκρίνοντας την μέση τιμή έντασης του υποθετικού spot με την μέση τιμή του background μπορούμε να καταλήξουμε στο συμπέρασμα ότι δεν έχουμε spot και να θεωρήσουμε ότι έχουμε μηδενική ένταση και έκφραση.

Εν τέλει έπειτα από εφαρμογή του αλγόριθμου συμπλήρωσης του κέντρου του spot και απαλοιφής των περιφερικών σημείων που δεν ανήκουν στο spot χρησιμοποιούμε το παραπάνω αποτέλεσμα το οποίο μας δίνει την θεωρητική περιοχή του spot για να απορρίψουμε ομάδες από pixel τα οποία δεν ανήκουν στο spot. Δηλαδή στον υπολογισμό του ποσοστού έκφρασης δεν συνυπολογίζουμε τα σημεία που βρίσκονται στο εσωτερικό του spot και έχουν απορριφθεί από τους αλγόριθμους κατάτμησης αλλά ούτε και τα σημεία που βρίσκονται γύρω από το spot και λόγω θορύβου συμπεριλήφθηκαν στο foreground-spot.



Εικόνα 4.7.5.5 Οι τέσσερις περιπτώσεις διόρθωσης του αποτελέσματος κατάτμησης. **Πάνω Αριστερά:** Η συνολική εικόνα του κελιού. **Κάτω αριστερά:** Το αποτέλεσμα της κατάτμησης του αλγόριθμου επαναληπτικού kmeans (μόνο με βάση ένταση). **Πάνω δεξιά:** Το σύνολο των pixel που χρησιμοποιούμε εν τέλει στον υπολογισμό του ποσοστού έκφρασης. **Κάτω δεξιά:** Ο υπολογισμένος χώρος του spot όπως προέκυψε από την εφαρμογή των αλγορίθμων Center Fill και Unfill.

Στην περίπτωση όπου χρησιμοποιούμε και τα δύο χαρακτηριστικά (ένταση και απόχρωση) εφαρμόζουμε πρώτα τον αλγόριθμο απαλοιφής των περιφερειακών pixel και μετά το φίλτρο πάνω στην κατανομή της απόχρωσης. *Αυτό το κάνουμε για να διορθώσουμε πρώτα τον χώρο ώστε η κατανομή της απόχρωσης να μην περιέχει χρώματα άλλων spot*

5 ΠΕΙΡΑΜΑ -ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 Dataset

Χρησιμοποιήσαμε την βάση εικόνων μικροσυστοιχιών και δεδομένων του Εθνικού Ινστιτούτου Καρκίνου (National Cancer Institute ^[25]) καθώς επίσης και τα αποτελέσματα κατάτμησης που προκύπτουν από τον αλγόριθμο επεξεργασίας εικόνας του λογισμικού GEMTools ^[24]. Οι συγκεκριμένες εικόνες δημιουργήθηκαν από την **Incyte Genomics** για λόγους σύγκρισης των παραμέτρων σάρωσης ανάμεσα σε διάφορα πειράματα.

Η βάση δεδομένων αποτελείται από 70 ανεξάρτητα πειράματα στα οποία έχουμε υβριδισμούς τριών δειγμάτων RNA το πρώτο από ανθρώπινη καρδιά, το δεύτερο από πλακούντα και το τρίτο από εγκέφαλο. Στις 20 από τις 70 μικροσυστοιχίες είχαμε ομοτυπικούς υβριδισμούς: πλακούντα-πλακούντα, εγκέφαλο-εγκέφαλο, καρδιά-καρδιά. Στις υπόλοιπες 50 είχαμε spot ελέγχου, δέκα μικροσυστοιχίες για κάθε έναν από τους υβριδισμούς: εγκέφαλος-πλακούντας, πλακούντας-εγκέφαλος, καρδιά-πλακούντας, πλακούντας-καρδιά, όπου στα ζεύγη αυτά το πρώτο χρωματίστηκε με πράσινο (Cy3) και το δεύτερο με κόκκινο (Cy5).

Στην περίπτωση των ομοτυπικών υβριδισμών τα δύο δείγματα μας είναι ίδια δηλαδή το χρωματισμένο υλικό με Cy3 είναι το ίδιο με αυτό που είναι χρωματισμένο με Cy5. Κατά συνεπεία τα δύο υλικά αποτελούνται από τα ίδια τμήματα DNA άρα αναμένουμε το ίδιο ποσοστό υβριδισμού και των δύο με τα γονίδια βάσης. Αυτό σημαίνει ότι η αναλογία του κόκκινου προς πράσινο, δηλαδή το ποσοστό έκφρασης πρέπει να είναι ίσο με την μονάδα για όλα τα spot.

Η μικροσυστοιχίες αποτελούνται από περίπου 10.000 παράγωγα PCR. Σε κάθε μια τυπώθηκαν παράγωγα DNA από 10 μη κωδικές περιοχές του *Saccharomyces cerevisiae* σε τετράτυπα σε διάφορα σημεία πάνω στην μικροσυστοιχία και σε γνωστές αναλογίες υβριδισμού. Ο χρωματισμός των δειγμάτων έγινε με την παρουσία 200 ng poly-A mRNA από ανθρώπινη καρδιά ή εγκέφαλο. Κατά συνέπεια όλες οι εικόνες μας ακόμα και αυτές των ομοτυπικών υβριδισμών έχουν σε δεδομένες θέσεις spot με γνωστή αναλογία κόκκινου προς πράσινο. Συγκεκριμένα κάθε εικόνα περιέχει 10 τετράδες όμοιων spot, δηλαδή 40 spot ελέγχου.

Για την σάρωση των μικροσυστοιχιών χρησιμοποιήθηκε ο σαρωτής Axon και για την επεξεργασία της εικόνας το λογισμικό GEMTools ^[24]. Οι μικροσυστοιχίες σαρώθηκαν χρησιμοποιώντας δύο είδη ρυθμίσεων με στόχο να συγκριθούν οι παράμετροι.

5.2 Golden Truth

Για να μπορέσουμε να έχουμε μια εκτίμηση του σφάλματος για κάθε ένα από τους παραπάνω αλγορίθμους που δοκιμάσαμε χρειαζόμαστε ένα μέτρο σύγκρισης. Όπως εξηγήσαμε και παραπάνω η διαδικασία ταξινόμησης των Pixels σε background και foreground είναι unsupervised. Αυτό σημαίνει ότι δεν έχουμε την πληροφορία για την κλάση στην οποία ανήκει κάθε pixel.

Η πηγή όμως των εικόνων πάνω στις οποίες εφαρμόσαμε τους παραπάνω αλγόριθμους μας δίνει συνολικά 70 εικόνες από microarrays από τις οποίες γνωρίζουμε ότι προήλθαν από υβριδισμούς τριών πη δειγμάτων: ένα από ανθρώπινο πλακούντα, ένα από ανθρώπινο εγκέφαλο και ένα τρίτο από ανθρώπινη καρδιά. Τα είκοσι από τα εβδομήντα array έχουν προκύψει από ομοτυπικούς υβριδισμούς για παράδειγμα πλακούντας – πλακούντας, εγκέφαλος – εγκέφαλος και καρδιά – καρδιά. Τα υπόλοιπα 50 array είναι μη ομοτυπικοί υβριδισμοί με διαφορετικές αναλογίες μαζών.

Στην πρώτη περίπτωση τα δύο υλικά πη αναμείχθηκαν σε αναλογία 1:1 που σημαίνει ότι αναμένουμε για κάθε spot η αναλογία του κόκκινου προς το πράσινο να είναι κοντά στην μονάδα:

$$x_0 = \frac{Normalized(R)}{Normalized(G)} = 1 \quad (6.1)$$

Στην δεύτερη περίπτωση τα arrays περιέχουν ένα σύνολο από spots τα οποία προέκυψαν από παράγωγα PCR από δέκα μη κωδικές περιοχές του ζαχαρομύκητα *accharomyces cerevisiae*. Αυτά τα spots τυπώθηκαν σε τετραπλότυπα σε διαφορετικά σημεία πάνω στα arrays. Αντίγραφα των τμημάτων του μύκητα συμπεριλήφθηκαν σε κάθε αντίδραση κατάταξης σε προκαθορισμένα Cy3:Cy5 επίπεδα εισόδου γεγονός που σημαίνει ότι για αυτά τα spot έχουμε γνωστή την αναλογία κόκκινου προς πράσινο.

Συγκεκριμένα μας δίνεται ο παρακάτω πίνακας:

Location	GeneName	Cy3	Cy5	Cy3: Cy5
2473	Control: YCFR 43 {YC 43.2000.W}	10	30	-3
2474	Control: YCFR 22 {YC 22.2000.W}	30	10	3
2475	Control: YCFR 23 {YC 23.2000.W}	15	150	-10
2476	Control: YCFR 25 {YC 25.2000.W}	150	15	10
2477	Control: YCFR 44 {YC 44.2000.W}	10	250	-25
2478	Control: YCFR 26 {YC 26.2000.W}	250	10	25
2479	Control: YCFR 45 {YC 45.2000.W}	2	2	1
2480	Control: YCFR 6 {YC 6.2000.W}	20	20	1
2481	Control: YCFR 67 {YC 67.2000.W}	200	200	1
2482	Control: YCFR 85 {YC 85.2000.W}	2000	2000	1
4973	Control: YCFR 43 {YC 43.2000.X}	10	30	-3
4974	Control: YCFR 22 {YC 22.2000.X}	30	10	3
4975	Control: YCFR 23 {YC 23.2000.X}	15	150	-10
4976	Control: YCFR 25 {YC 25.2000.X}	150	15	10
4977	Control: YCFR 44 {YC 44.2000.X}	10	250	-25
4978	Control: YCFR 26 {YC 26.2000.X}	250	10	25
4979	Control: YCFR 45 {YC 45.2000.X}	2	2	1
4980	Control: YCFR 6 {YC 6.2000.X}	20	20	1
4981	Control: YCFR 67 {YC 67.2000.X}	200	200	1
4982	Control: YCFR 85 {YC 85.2000.X}	2000	2000	1
7473	Control: YCFR 43 {YC 43.2000.Z}	10	30	-3
7474	Control: YCFR 22 {YC 22.2000.Z}	30	10	3
7475	Control: YCFR 23 {YC 23.2000.Z}	15	150	-10
7476	Control: YCFR 25 {YC 25.2000.Z}	150	15	10
7477	Control: YCFR 44 {YC 44.2000.Z}	10	250	-25
7478	Control: YCFR 26 {YC 26.2000.Z}	250	10	25
7479	Control: YCFR 45 {YC 45.2000.Z}	2	2	1
7480	Control: YCFR 6 {YC 6.2000.Z}	20	20	1
7481	Control: YCFR 67 {YC 67.2000.Z}	200	200	1
7482	Control: YCFR 85 {YC 85.2000.Z}	2000	2000	1
9973	Control: YCFR 43 {YC 43.2000.Y}	10	30	-3
9974	Control: YCFR 22 {YC 22.2000.Y}	30	10	3
9975	Control: YCFR 23 {YC 23.2000.Y}	15	150	-10
9976	Control: YCFR 25 {YC 25.2000.Y}	150	15	10
9978	Control: YCFR 26 {YC 26.2000.Y}	250	10	-25
9977	Control: YCFR 44 {YC 44.2000.Y}	10	250	25
9979	Control: YCFR 45 {YC 45.2000.Y}	2	2	1
9980	Control: YCFR 6 {YC 6.2000.Y}	20	20	1
9981	Control: YCFR 67 {YC 67.2000.Y}	200	200	1

9982 Control: YCFR 85 {YC 85.2000.Y} 2000 2000 1

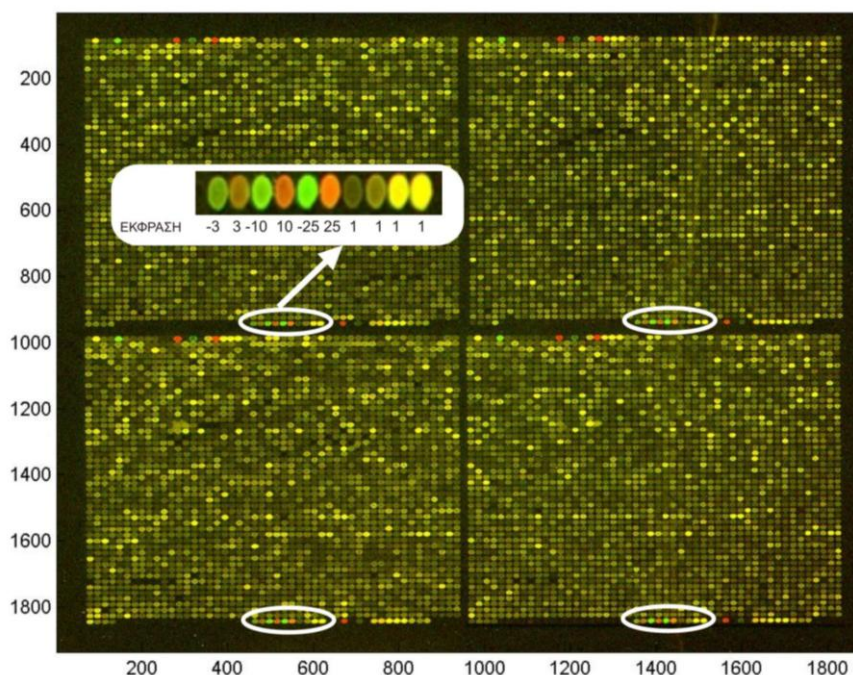
Πίνακας 5.1.1 Ο πίνακας αυτός μας δίνεται μαζί με όλα τα άλλα δεδομένα και περιέχει τις θέσεις των spot ελέγχου καθώς και τις αναλογίες πρόσμιξης, δηλαδή τις αναλογίες κόκκινου προς πράσινο. Στην περίπτωση που η τιμή έντασης του κόκκινου είναι μεγαλύτερη από την τιμή του πράσινου ($R>G$) τότε η αναλογία υπολογίζεται ως R/G . Εάν $G>R$ τότε η αναλογία είναι $-G/R$. Έτσι προκύπτουν τα αρνητικά πρόσημα στον πίνακα.

Έχουμε συνολικά 40 τέτοια spots οι θέσεις των οποίων φαίνονται στην πρώτη στήλη του παραπάνω πίνακα. Στην δεύτερη στήλη είναι το όνομα του γονιδίου και οι δύο επόμενες μας δίνουν την τιμή έντασης του κόκκινου (Cy3) και του πράσινου (Cy5). Η τελευταία στήλη είναι η τιμή έκφρασης, δηλαδή η αναλογία κόκκινου προς πράσινο με την μόνη παρατήρηση ότι έναν $R>G$ τότε η τιμή έκφρασης υπολογίζεται από τον λόγο R/G ενώ στην περίπτωση όπου έχουμε $G>R$ τότε η τιμή έκφρασης είναι ίση με $-G/R$.

Για παράδειγμα για το πρώτο spot το οποίο βρίσκεται στη θέση 2473 έχουμε ότι η τιμή έκφρασης είναι 3 άρα θα πρέπει

$$x_1 = \frac{\text{Normalized}(G)}{\text{Normalized}(R)} \cong 3 \quad (6.2)$$

Για καλύτερη κατανόηση της θέσης των spot ελέγχου παραθέτουμε την παρακάτω εικόνα.



Εικόνα 5.1.1 **Οι θέσεις των spot ελέγχου.** Μια από τις εικόνες Microarray που μελετάμε στην οποία έχουμε αυξήσει το contrast ώστε να είναι πιο ευδιάκριτα τα spot μας. Σε κάθε ένα από τα τέσσερα subarray έχουμε κυκλώσει τα δέκα spot ελέγχου. Έχουμε μεγεθύνει το πρώτο σετ (θέσεις 2473-2482) με τις αντίστοιχες τιμές έκφρασής ή αλλιώς αναλογία κόκκινου προς πράσινο (Cy3: Cy5). Όπως εξηγήσαμε οι αρνητικές τιμές αναλογίας προκύπτουν όταν έχουμε $G>R$ οπότε και η αναλογία είναι $-G/R$. Συγκεκριμένα τα spot τα οποία έχουν πράσινο χρώμα έχουν αρνητικές αναλογίες ενώ τα κόκκινα θετικές. Παρατηρούμε ότι όσο αυξάνει η κατ' απόλυτο τιμή αναλογία αυξάνει και η ένταση του χρώματος. Τα τελευταία τέσσερα spot έχουν κίτρινο χρώμα διότι η αναλογία R/G είναι ίση με την μονάδα.

5.3 Υπολογισμός Σφάλματος – Mean Square Error

Ο υπολογισμός του σφάλματος γίνεται υπολογίζοντας την αναμενόμενη τιμή της απόλυτης τιμής της διαφοράς. Συγκεκριμένα βρίσκουμε την μέση τιμή της του απόλυτου της διαφοράς των τετραγώνων των πειραματικών τιμών από τις θεωρητικές τιμές. Οι θεωρητικές τιμές όπως εξηγήσαμε στη προηγούμενη ενότητα είναι ίσες με την τελευταία στήλη του πίνακα 5.1.1 για τους μη ομοτυπικούς υβριδισμούς ενώ η θεωρητική τιμή για τους ομοτυπικούς υβριδισμούς είναι ίση με την μονάδα.

Συγκεκριμένα οι πειραματικές αναλογίες είναι υπολογισμένες ως εξής.

$$\bar{x}_{\text{πειραματικές}} = \frac{\text{Normalized}(\bar{R})}{\text{Normalized}(\bar{G})} \quad (5.2.3)$$

Οπότε το σφάλμα ή αλλιώς η απόκλιση από την πραγματική τιμή είναι:

$$\text{Error} = E\left(\left|\bar{x}_{\text{πειραματικές}} - \bar{x}_{\text{θεωρητικές}}\right|\right) \quad (5.2.4)$$

Στην περίπτωση των 20 μικροσυστοιχιών όπου έχουμε ομοτυπικούς υβριδισμούς όπως αναφέραμε και παραπάνω θεωρούμε ότι οι θεωρητικές τιμές έκφρασης των spot (κανονικοποιημένη αναλογία κόκκινου προς πράσινο) είναι όλες ίσες με την μονάδα. Πιο συγκεκριμένα επειδή στις εικόνες ομοτυπικών υβριδισμών έχουμε επιπλέον και τα spot ελέγχου που υπάρχουν και στις υπόλοιπες εικόνες και τα οποία έχουν διαφορετικές αναλογίες τα παραλείπουμε από τον υπολογισμό του σφάλματος.

Οι μικροσυστοιχίες οι οποίες μελετάμε αποτελούνται από τέσσερις υπό-πίνακες από spot όπου κάθε υπο-πίνακας διαστάσεις 50x50 δηλαδή ο αριθμός των spots είναι ίσος με $4 \times (50 \times 50) = 10.000$. Συνεπώς από κάθε υπο-πίνακα παραλείπουμε τα σαράντα spot ελέγχου (εικόνα 5.1.1)

$$\bar{x}_{\text{θεωρητικές}}^{\text{ομοτυπικά}} = \underbrace{\{1111 \dots 11\}}_{10.000 \text{ spots} - 40 \text{ spot ελέγχου}} \quad (5.2.5)$$

Στην περίπτωση όπου έχουμε μη ομοτυπικούς υβριδισμούς χρησιμοποιούμε μόνο τα 40 spots ελέγχου που έχουν τοποθετηθεί στα array μας. Οι θεωρητικές τιμές έκφρασης για αυτά τα spots είναι δοσμένες και φαίνονται στον πίνακα 5.1.1.

$$\bar{x}_{\text{θεωρητικές}}^{\text{spot ελέγχου}} = \underbrace{\left[\underbrace{-3, 3, -10, 10, -25, 25, 1, 1, 1, 1}_{10 \text{ spot}}, \dots, \underbrace{-3, 3, -10, 10, -25, 25, 1, 1, 1, 1}_{10 \text{ spot}} \right]}_{40 \text{ spots}} \quad (5.2.6)$$

5.4 Περιγραφή πειραμάτων

Πραγματοποιούμε διαφορά πειράματα με στόχο να κάνουμε κάποιες συγκρίσεις, όχι μόνο σε επίπεδο αλγορίθμων αλλά και όσον αφορά τις διάφορες προσθήκες που κάναμε και είχαν ως στόχο έναν τέλειο διαχωρισμό του spot από την υπόλοιπη εικόνα.

Συγκεκριμένα αναφερόμαστε στην προσθήκη του δεύτερου χαρακτηριστικού στην διαδικασία της ταξινόμησης, δηλαδή στον συνυπολογισμό του Hue μαζί με την ένταση και εάν αυτή προσθήκη είχε ως αποτέλεσμα μείωση του συνολικού σφάλματος.

Άλλη μια αξιολόγηση που κάνουμε αφορά τον αλγόριθμο συμπλήρωσης του spot που αναφερθήκαμε στην ενότητα 4.7. Η εφαρμογή αυτή είχε ως στόχο έναν τέλειο οπτικό διαχωρισμό του spot από το background αλλά μια τέτοια διαδικασία είναι πραγματικά αναγκαία ή περιπλέκει τα πράγματα χωρίς να επιφέρει μεγάλη διόρθωση στο σφάλμα ή αν ακόμα αυξάνει το σφάλμα;

Θέλοντας να κάνουμε τις παραπάνω συγκρίσεις εκτελέσαμε τους τρεις αλγορίθμους (Otsu, k-means, EM) σε τέσσερα πειράματα. Σε πρώτη φάση χρησιμοποιήσαμε και τα δύο χαρακτηριστικά Hue και Intensity, δηλαδή εφαρμόσαμε τους τρεις αλγορίθμους πάνω στο ιστόγραμμα της έντασης και έπειτα εφαρμόσαμε φίλτρο πάνω στο ιστόγραμμα της απόχρωσης (ενότητα 3.2). Στην δεύτερη σειρά πειραμάτων είχαμε ως μοναδικό χαρακτηριστικό την ένταση δεν εφαρμόσαμε δηλαδή το φίλτρο πάνω στην απόχρωση. Τέλος στις δύο παραπάνω περιπτώσεις αφαιρέσαμε και τον αλγόριθμο απαλοιφής των pixel που δεν ανήκουν στο spot λόγω θέσης.

Συνοψίζοντας τα πειράματά μας έχουμε:

- Μοναδικό χαρακτηριστικό η ένταση
- Συνδυασμός δύο χαρακτηριστικών έντασης και απόχρωσης (Hue - Intensity)
- Χρησιμοποιώντας μόνο το χαρακτηριστικό της έντασης και παραλείποντας τον αλγόριθμο απαλοιφής των περιφερειακών Pixel του spot.
- Με συνδυασμό των δύο χαρακτηριστικών Hue και Intensity και παραλείποντας τον αλγόριθμο απαλοιφής των περιφερειακών Pixel του spot.

Όσον αφορά τον δυσδιάστατο αλγόριθμο k-means επειδή εξ' ορισμού ο αλγόριθμος συνδυάζει τα δύο χαρακτηριστικά η μόνη σύγκριση που μπορούμε να κάνουμε είναι με την εφαρμογή του αλγορίθμου απαλοιφής των περιφερειακών και χωρίς αυτήν.

5.5 Εικόνες με Spot Ελέγχου

Όπως έχουμε ήδη αναφέρει μας δίνονται 50 εικόνες με spot ελέγχου σε γνωστές θέσεις. Εφαρμόζουμε τα πειράματα σε αυτές τις εικόνες και υπολογίζουμε το συνολικό σφάλμα (με τον τρόπο που περιγράψαμε στην ενότητα 5.2) για κάθε μια από τις εικόνες καθώς και την μέση τιμή σφάλματος πάνω στο σύνολο τους.

Ακολουθούν τα αποτελέσματα των παραπάνω πειραμάτων για τους τέσσερις αλγορίθμους. Για τους K-means, Otsu και EM συγκρίνουμε το συνολικό σφάλμα στις παρακάτω περιπτώσεις:

- όταν χρησιμοποιούμε μόνο το χαρακτηριστικό τις έντασης και όταν συνδυάζουμε τα δύο χαρακτηριστικά, ένταση και απόχρωση
- όταν χρησιμοποιούμε μόνο το χαρακτηριστικό της έντασης αλλά στην μια περίπτωση κάνουμε απαλοιφή των περιφερειακών Pixel του spot ενώ στην δεύτερη περίπτωση όχι.
- όταν συνδυάζουμε και τα δύο χαρακτηριστικά αλλά στην μια περίπτωση κάνουμε απαλοιφή των περιφερειακών Pixel του spot ενώ στην δεύτερη περίπτωση όχι.

Όσον αφορά τον δυσδιάστατο k-means κάνουμε μόνο την τελευταία σύγκριση.

Σε κάθε μια από τις συγκρίσεις παραθέτουμε και το σφάλμα (για κάθε εικόνα) το οποίο είναι υπολογισμένο με βάση τα αποτελέσματα κατάτμησης του λογισμικού GenPix τα οποία μας έχουν δοθεί μαζί με το σύνολο των εικόνων και των άλλων δεδομένων.

Στις επόμενες ενότητες τα γραφήματα προκύπτουν από τον υπολογισμό της μέσης τιμής του σφάλματος για 50 εικόνες για τις παραπάνω περιπτώσεις πειραμάτων. Στα δεξιά κάθε γραφικής παρατίθεται η μέση τιμή σφάλματος πάνω στις 50 εικόνες και πάλι για τις εκάστοτε περιπτώσεις που συγκρίνουμε.

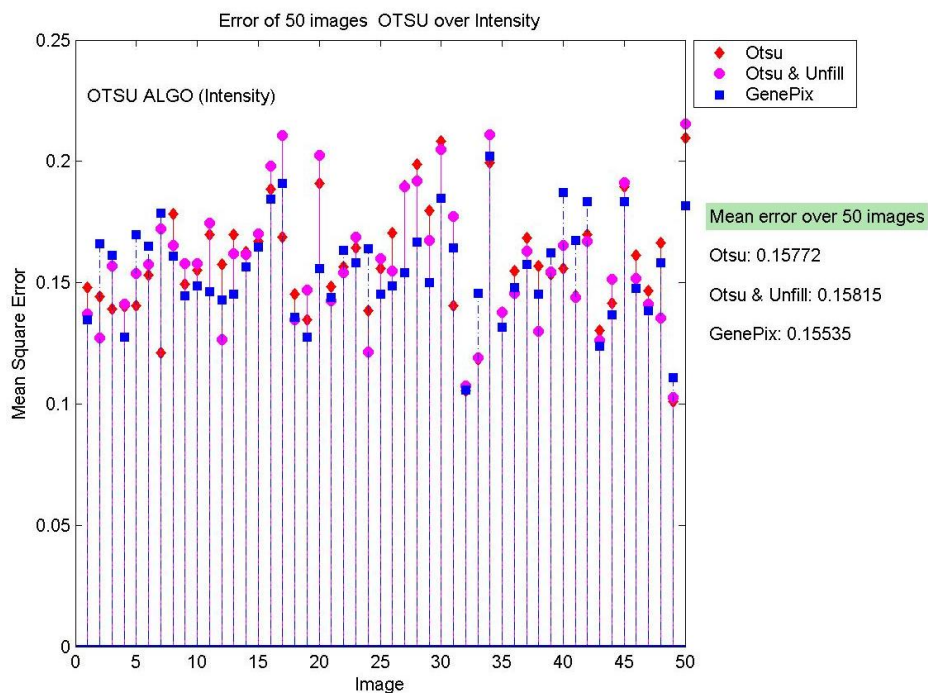
5.5.1 Απόδοση αλγορίθμου Otsu

Στην περίπτωση του αλγορίθμου Otsu (επαναληπτικού Otsu) όπως αναφέραμε και παραπάνω συγκρίνουμε την απόδοση του σε τρεις περιπτώσεις. Αρχικά παραθέτουμε την απόδοση του αλγορίθμου με την εφαρμογή ή όχι του αλγορίθμου συμπλήρωσης όταν χρησιμοποιείται μόνο το χαρακτηριστικό της έντασης και όταν συνδυάζονται τα δύο χαρακτηριστικά. Έπειτα ακολουθεί ένα γράφημα για την σύγκριση των δύο περιπτώσεων, όταν κάνουμε κατάτμηση με βάση μόνο την ένταση και όταν χρησιμοποιούμε συμπληρωματικά και την απόχρωση. *Σε αυτή την σύγκριση δεν εφαρμόστηκε ο αλγόριθμος απαλοιφής των περιφερειακών Pixel του spot (Spot Unfill).*

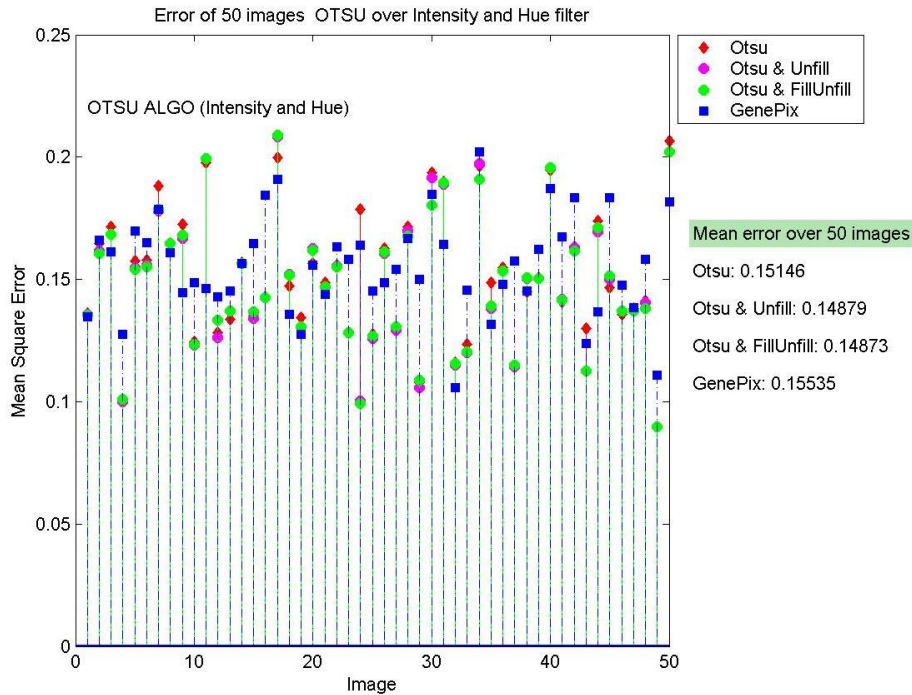
Σε όλα τα γραφήματα παραθέτουμε και το σφάλμα που προκύπτει σύμφωνα με τα δοσμένα αποτελέσματα κατάτμησης που έχουμε από την πηγή των εικόνων μας. Στα δεξιά κάθε γραφήματος παραθέτουμε την μέση τιμή σφάλματος πάνω σε όλες τις εικόνες για κάθε μια από τις περιπτώσεις που συγκρίνουμε.

Στα δύο γραφήματα που ακολουθούν Τα σημεία με κόκκινο ρόμβο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο απαλοιφής των περιφερειακών Pixel του spot. Αντίθετα στα σημεία με μωβ κύκλο δεν έχει χρησιμοποιηθεί ο αλγόριθμος. Με μπλε τετράγωνο έχουμε και πάλι το σφάλμα σύμφωνα με τα δοσμένα αποτελέσματα κατάτμησης που έχουν προκύψει από το λογισμικό GenPix.

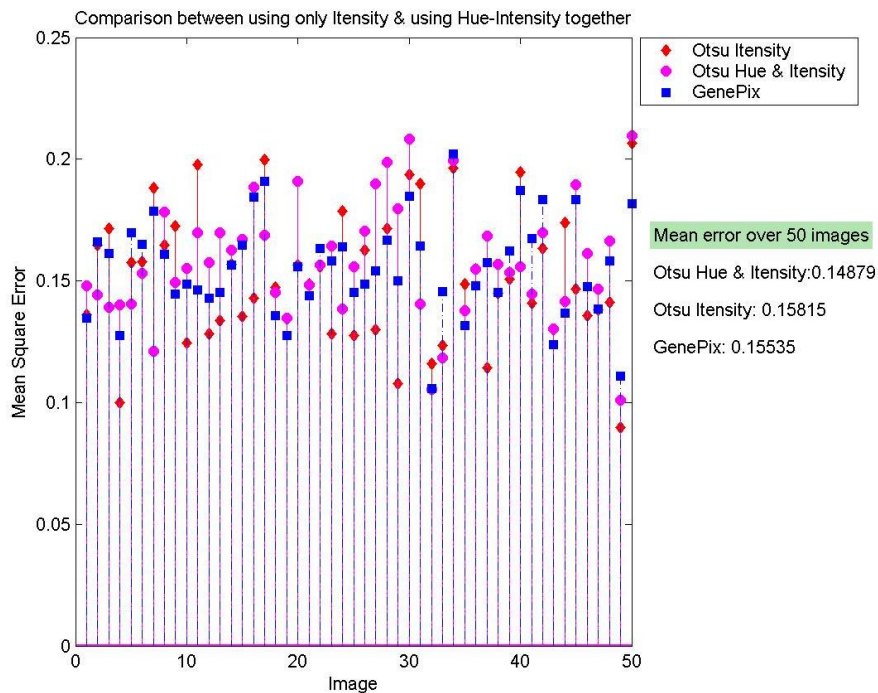
Στο τρίτο γράφημα με κόκκινο ρόμβο είναι τα μέσα σφάλματα για τις 50 εικόνες χρησιμοποιώντας και τα δύο χαρακτηριστικά και με μωβ κύκλο τα σφάλματα για τις ίδιες εικόνες όταν χρησιμοποιήσουμε μόνο την ένταση. Με μπλε τετράγωνο είναι και πάλι το σφάλμα υπολογισμένο με τα αποτελέσματα του GenPix για τις ίδιες εικόνες.



Γρ. 5.5.1 Οι γραφικές με το μωβ και κόκκινο χρώμα αντιπροσωπεύουν το σφάλμα που έχουμε για τις 50 εικόνες χρησιμοποιώντας τον αλγόριθμο Otsu. Και στις δύο περιπτώσεις εφαρμόζουμε τον **επαναληπτικό αλγόριθμο Otsu** πάνω στο ιστόγραμμα της έντασης. Δεν λαμβάνουμε υπόψη την απόχρωση. Η διαφορά στις δύο περιπτώσεις έχει να κάνει με την χρήση του αλγορίθμου απαλοιφής των περιφερειακών Pixel (Spot Unfill) του spot που περιγράψαμε στην ενότητα 4.7.3. Τα σημεία με μωβ κύκλο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill. Αντίθετα στα σημεία με κόκκινο ρόμβο δεν έχει χρησιμοποιηθεί ο αλγόριθμος συμπλήρωσης. Με μπλε τετράγωνο έχουμε το σφάλμα σύμφωνα με τα δοσμένα αποτελέσματα κατάτμησης του GenTools.



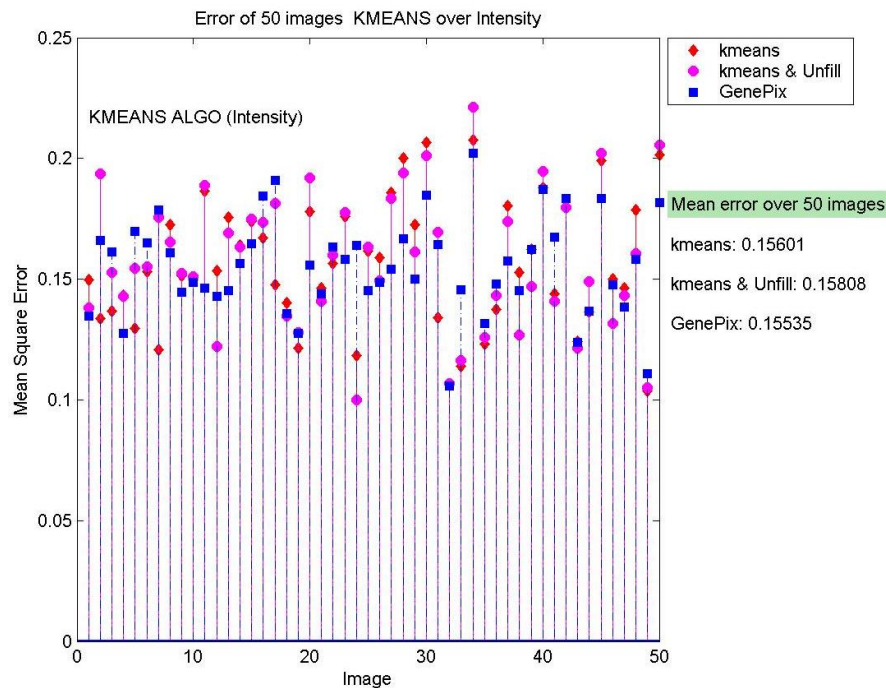
Γρ. 5.5.2 Αντίστοιχη περίπτωση με γραφική 5.5.1. Οι γραφικές με το μωβ και κόκκινο χρώμα αντιπροσωπεύουν το σφάλμα που έχουμε για τις 50 εικόνες χρησιμοποιώντας τον αλγόριθμο Otsu. Και στις δύο περιπτώσεις εφαρμόζουμε τον επαναληπτικό αλγόριθμο Otsu πάνω στο ιστόγραμμα της έντασης. Σε αυτήν την γραφική λαμβάνουμε υπόψη την απόχρωση. Εφαρμόζουμε δηλαδή ένα φίλτρο πάνω στις τιμές της απόχρωσης. Η διαφορά στις δύο περιπτώσεις έχει να κάνει και πάλι με την χρήση του αλγορίθμου Unfill που περιγράψαμε στην ενότητα 4.7.3. Τα σημεία με κόκκινο ρόμβο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill του spot. Αντίθετα στα σημεία με μωβ κύκλο δεν έχει χρησιμοποιηθεί ο αλγόριθμος Unfill. Με πράσινο κύκλο έχουμε το σφάλμα ότα χρησιμοποιούμε και Fill και Unfill. Με μπλε τετράγωνο έχουμε και πάλι το σφάλμα από GemTools.



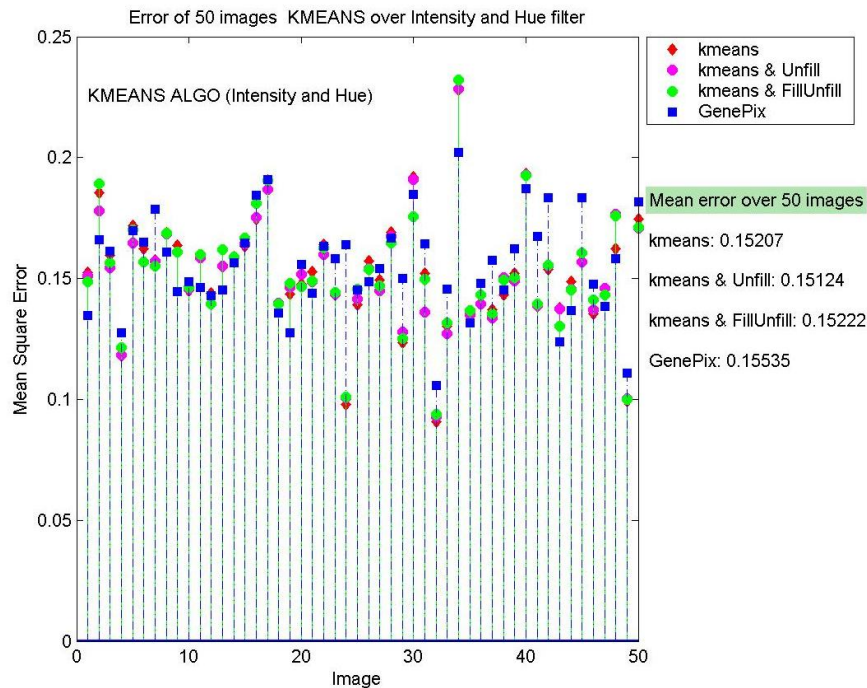
Γρ. 5.5.3 Σύγκριση μεταξύ δυοδιάστατης και μονοδιάστατης κατάτμησης. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του επαναληπτικού Otsu πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα πάνω στις 50 εικόνες για τις τρεις περιπτώσεις που συγκρίνουμε. Η σύγκριση των γίνεται για την περίπτωση όπου εφαρμόζουμε τον αλγόριθμο απαλοιφής των περιφερειακών pixel (NF) .

5.5.2 Απόδοση αλγορίθμου k-means.

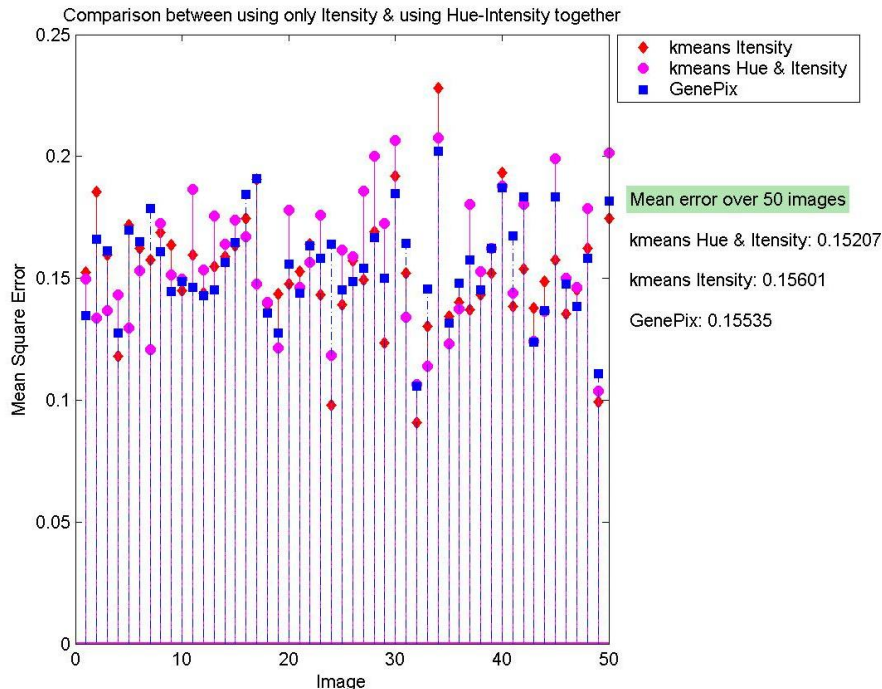
Έχουμε τα αντίστοιχα γραφήματα με παραπάνω αλλά σε αυτήν την περίπτωση πάνω στο ιστόγραμμα της έντασης εφαρμόζουμε τον επαναληπτικό αλγόριθμο k-means για $k=2$.



Γρ. 5.5.4 Οι γραφικές με το μωβ και κόκκινο χρώμα αντιπροσωπεύουν το σφάλμα που έχουμε για τις 50 εικόνες χρησιμοποιώντας τον αλγόριθμο 2-means. Και στις δύο περιπτώσεις εφαρμόζουμε τον **επαναληπτικό αλγόριθμο 2-means** πάνω στο ιστόγραμμα της έντασης. Δεν λαμβάνουμε υπόψη την απόχρωση. Η διαφορά στις δύο περιπτώσεις έχει να κάνει με την χρήση του αλγορίθμου Unfill που περιγράψαμε στην ενότητα 4.7.3. Τα σημεία με μωβ κύκλο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill του spot. Αντίθετα στα σημεία με κόκκινο ρόμβο δεν έχει χρησιμοποιηθεί ο αλγόριθμος Unfill. Με μπλε τετράγωνο έχουμε το σφάλμα σύμφωνα με τα δοσμένα από αποτελέσματα κατάταξης από GemTools.



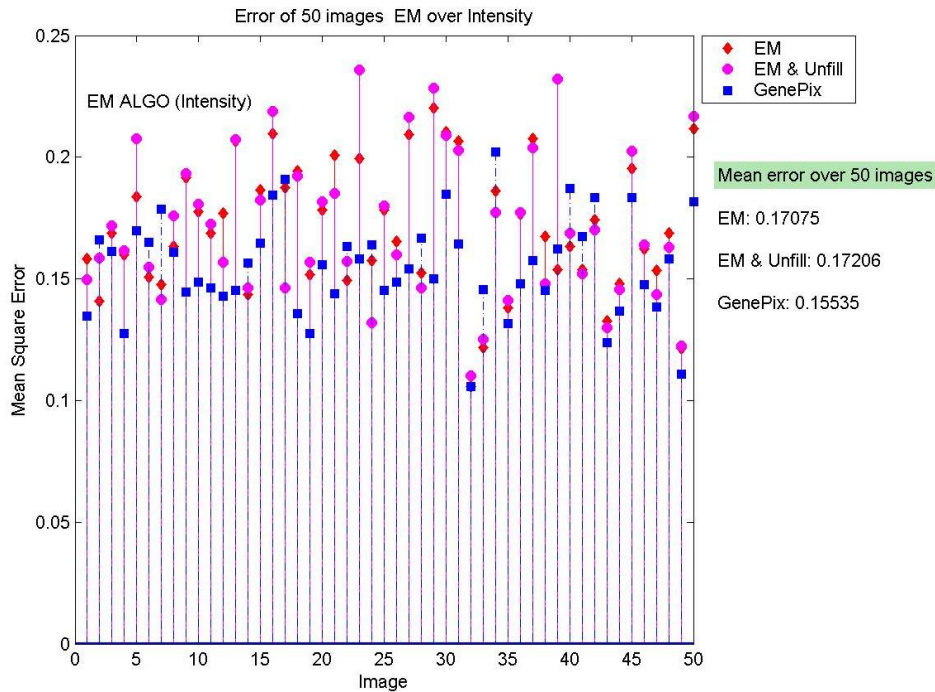
Γρ. 5.5.5 Αντίστοιχη σύγκριση με το προηγούμενο γράφημα με την διαφορά ότι εφαρμόζουμε επαναληπτικό αλγόριθμο 2-means πάνω στην ένταση σε συνδυασμό με την εφαρμογή φίλτρου πάνω στην απόχρωση. Συγκρίνουμε και πάλι την χρήση του αλγόριθμου Unfill (μωβ κύκλοι) και την μη χρήση του αλγόριθμου (κόκκινοι ρόμβοι), και με ταυτόχρονο Fill και Unfill (πράσινοι κύκλοι - θεωρητικός χώρος του spot). Στα δεξιά είναι και πάλι το συνολικό σφάλμα για τις τρεις περιπτώσεις.



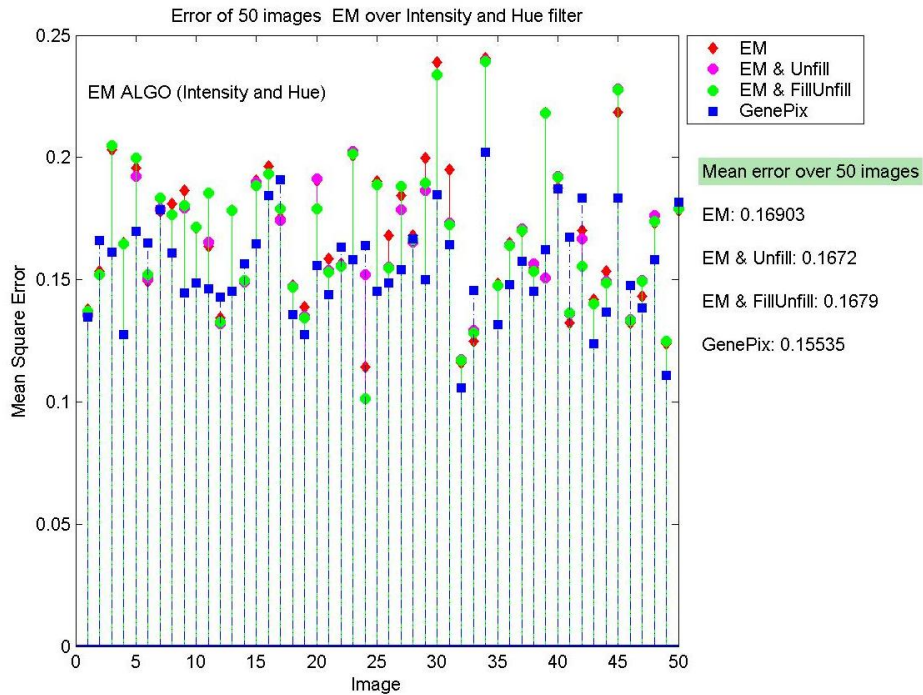
Γρ. 5.5.6 Σύγκριση μεταξύ δισδιάστατης και μονοδιάστατης κατάτμησης. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του επαναληπτικού 2-means πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα πάνω στις 50 εικόνες για τις τρεις περιπτώσεις που συγκρίνουμε. Η σύγκριση των γίνεται για την περίπτωση όπου εφαρμόζουμε τον αλγόριθμο απαλοιφής των περιφερειακών pixel (NF) .

5.5.3 Απόδοση επαναληπτικού αλγορίθμου EM

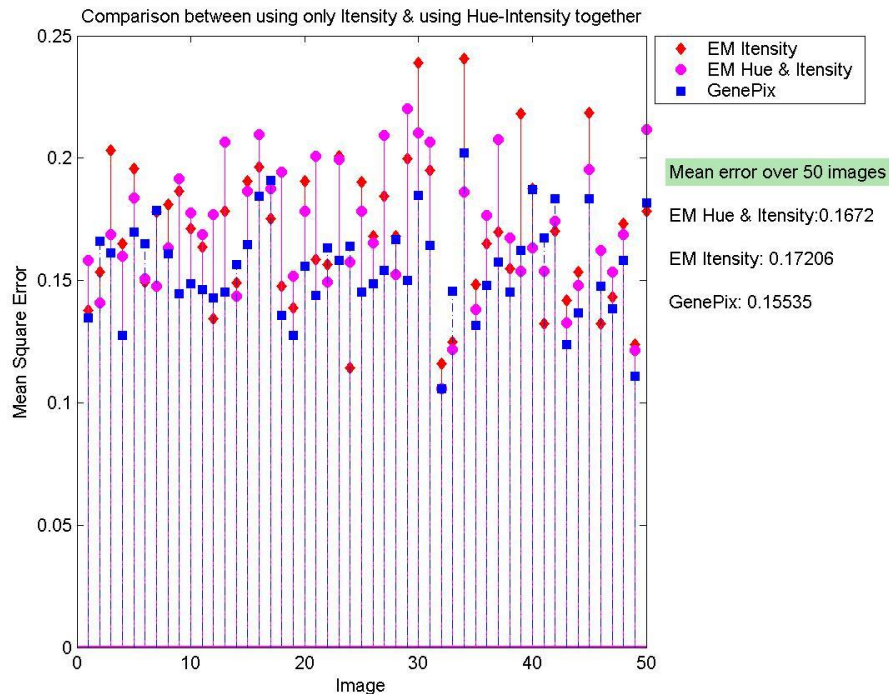
Έχουμε τα αντίστοιχα γραφήματα με δύο προηγούμενες ενότητες αλλά σε αυτήν την περίπτωση πάνω στο ιστόγραμμα της έντασης εφαρμόζουμε το επαναληπτικό αλγόριθμο Expectation Maximization.



Γρ. 5.5.7 Οι γραφικές με το μωβ και κόκκινο χρώμα αντιπροσωπεύουν το σφάλμα που έχουμε για τις 50 εικόνες χρησιμοποιώντας τον αλγόριθμο EM. Και στις δύο περιπτώσεις εφαρμόζουμε τον **επαναληπτικό αλγόριθμο EM** πάνω στο ιστόγραμμα της έντασης. Δεν λαμβάνουμε υπόψη την απόχρωση. Η διαφορά στις δύο περιπτώσεις έχει να κάνει με την χρήση του αλγορίθμου απαλοιφής των περιφερειακών pixel του spot περιγράψαμε στην ενότητα 4.7.3 (Unfill). Τα σημεία με μωβ κύκλο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill του spot. Αντίθετα στα σημεία με κόκκινο ρόμβο δεν έχει χρησιμοποιηθεί ο αλγόριθμος Unfill. Με μπλε τετράγωνο έχουμε το σφάλμα σύμφωνα με τα δοσμένα από την πηγή μας αποτελέσματα κατάτμησης.



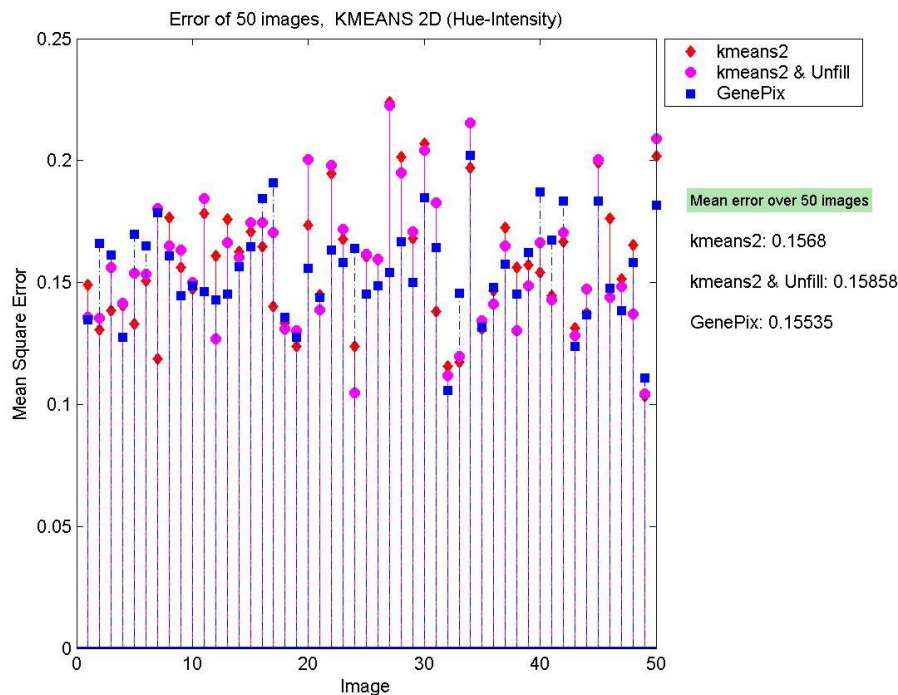
Γρ. 5.5.8 Αντίστοιχη σύγκριση με το προηγούμενο γράφημα με την διαφορά ότι εφαρμόζουμε επαναληπτικό αλγόριθμο **EM** πάνω στην ένταση σε συνδυασμό με την εφαρμογή φίλτρου πάνω στην απόχρωση. Συγκρίνουμε και πάλι την χρήση του αλγόριθμου Unfill (μωβ κύκλοι) και την μη χρήση του αλγόριθμου (κόκκινοι ρόμβοι), και με ταυτόχρονο Fill και Unfill (πράσινοι κύκλοι - θεωρητικός χώρος του spot). Στα δεξιά είναι και πάλι το συνολικό σφάλμα πάνω στις 50 εικόνες για τις τρεις περιπτώσεις.



Γρ. 5.5.9 Σύγκριση μεταξύ **δυσδιάστατης** και **μονοδιάστατης κατάτμησης**. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του **επαναληπτικού EM** πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα πάνω στις 50 εικόνες για τις τρεις περιπτώσεις που συγκρίνουμε. Η σύγκριση των γίνεται για την περίπτωση όπου εφαρμόζουμε τον αλγόριθμο απαλοιφής των περιφερειακών pixel (NF)

5.5.4 Απόδοση δυσδιάστατου 2-means αλγορίθμου

Στον αλγόριθμο αυτό έχουμε μόνο μια σύγκριση να κάνουμε καθώς εξ ορισμού ο αλγόριθμος χρησιμοποιεί και τα δύο χαρακτηριστικά, απόχρωση και ένταση. Η μόνη σύγκριση που κάνουμε αφορά την εφαρμογή του αλγόριθμου απαλοιφής των περιφερειακών pixel του spot (ενότητα 4.7.3).



Γρ. 5.5.10 Δυσδιάστατος επαναληπτικός 2-Means. Η παραπάνω γραφική προέκυψε από την εφαρμογή του δυσδιάστατου k-means πάνω στις 50 εικόνες στην μια περίπτωση χρησιμοποιώντας τον αλγόριθμο Unfill και στην άλλη όχι (κόκκινοι ρόμβοι και μωβ κύκλοι αντίστοιχα).

5.6 Εικόνες Ομοτυπικών Υβριδισμών

Όπως έχουμε ήδη αναφέρει έχουμε 20 εικόνες με ομοτυπικούς υβριδισμούς πάνω στις οποίες έχουμε τα ίδια σημεία ελέγχου που έχουμε και στους ομοτυπικούς υβριδισμούς, δηλαδή στα δεδομένα σημεία έχει τυπωθεί γεννητικό υλικό του ζαχαρομύκητα *accharomyces cerevisiae* σε δεδομένες αναλογίες. Όλα τα υπόλοιπα spot είναι σε αναλογία 1:1.

Για λόγους απλότητας δεν λαμβάνουμε υπόψη αυτά τα σημεία συγκεκριμένα την τελευταία σειρά από κάθε υπό πίνακα του microarray μας. Κάθε υπό-πίνακας έχει διαστάσεις 50x50 οπότε παραλείπουμε τα 50 τελευταία spot από τον καθένα από τους τέσσερις υπό-πίνακες. Υπολογίζουμε το σφάλμα με τον τρόπο που εξηγήσαμε στην ενότητα 5.3 λαμβάνοντας υπόψη τα υπόλοιπα spot. Βρίσκουμε το σφάλμα κάθε εικόνας και το συνολικό σφάλμα πάνω στις 20 εικόνες.

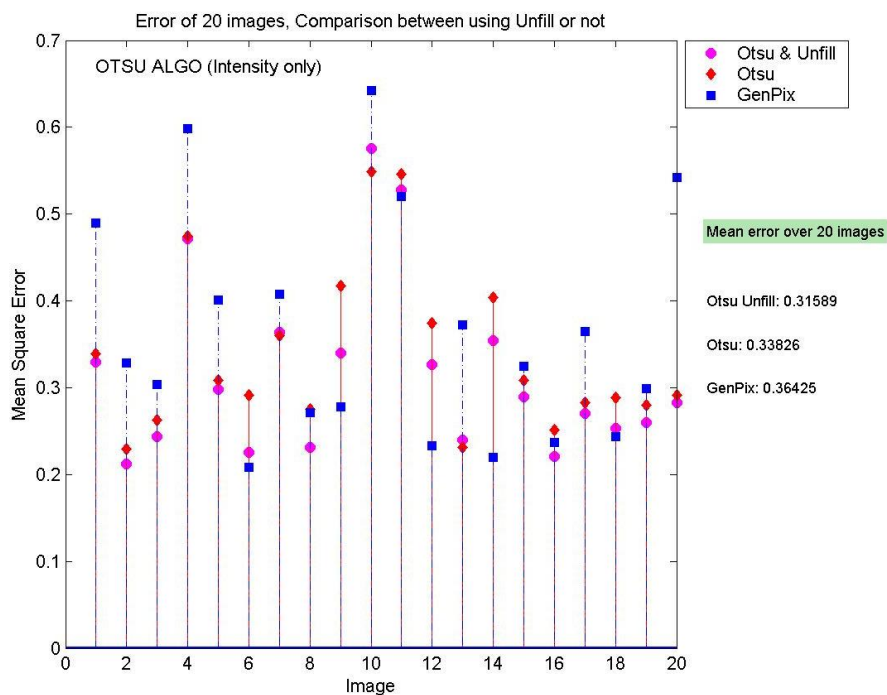
Πραγματοποιούμε τα ίδια πειράματα και τις ίδιες συγκρίσεις που περιγράψαμε στην ενότητα 5.4

5.6.1 Απόδοση αλγορίθμου Otsu

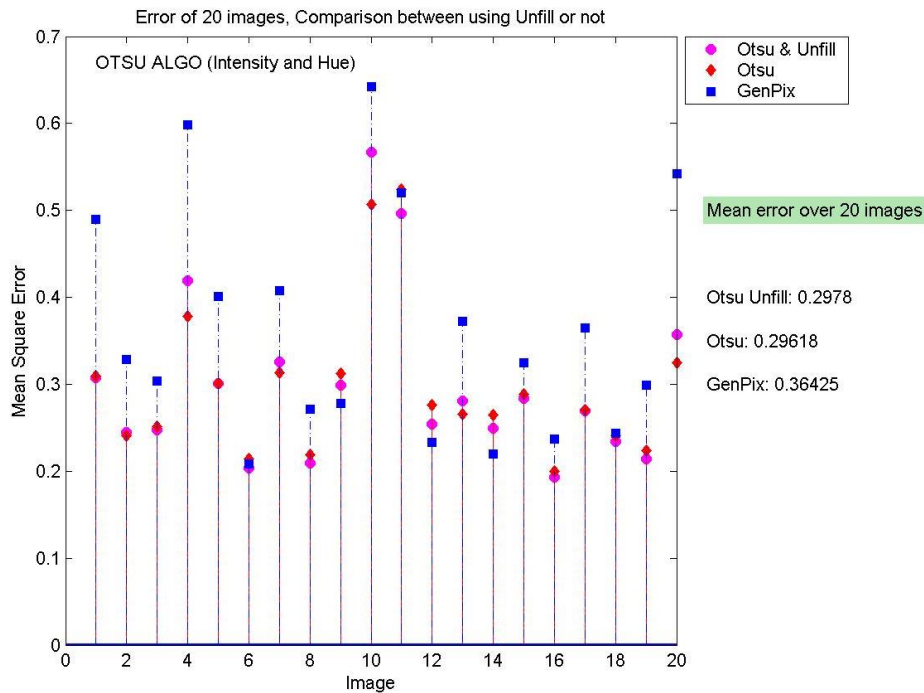
Εφαρμόζουμε τον επαναληπτικό αλγόριθμο Otsu όπως τον περιγράψαμε στην ενότητα 3.1.1.2 πάνω στο ιστόγραμμα τις έντασης και τον εφαρμόζουμε πάνω στις 20 εικόνες με ομοτυπικούς υβριδισμούς. Συγκρίνουμε το σφάλμα όταν χρησιμοποιούμε συμπληρωματικά τον αλγόριθμο συμπλήρωσης του spot και όταν δεν τον εφαρμόζουμε.

Αντίστοιχη σύγκριση κάνουμε και στην περίπτωση όπου χρησιμοποιούμε και το χαρακτηριστικό της απόχρωσης σε συνδυασμό με την ένταση (δυσδιάστατη κατάτμηση, δέντρο απόφασης ενότητα 3.2.2).

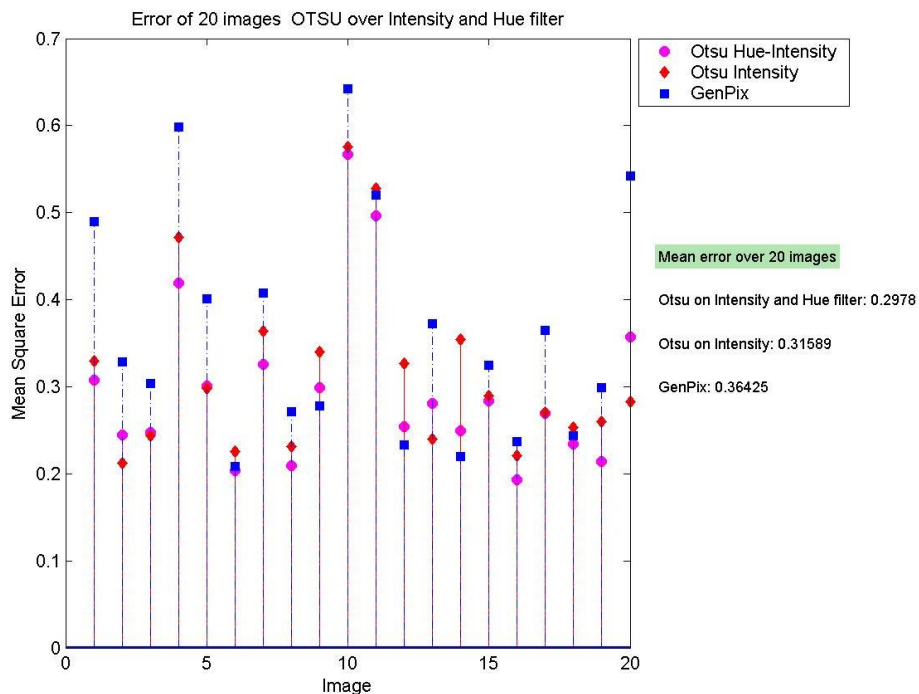
Τέλος συγκρίνουμε τα το σφάλμα όταν έχουμε μονοδιάστατη κατάτμηση με βάση την ένταση και όταν συνδυάζουμε και τα δύο χαρακτηριστικά. Όπως και στην περίπτωση των εικόνων με spot ελέγχου για την παραπάνω σύγκριση δεν χρησιμοποιούμε τον αλγόριθμο συμπλήρωσης.



Γρ. 5.6.1 Οι γραφικές με το μωβ και κόκκινο χρώμα αντιπροσωπεύουν το σφάλμα που έχουμε για τις 20 εικόνες όταν εφαρμόζουμε τον επαναληπτικό αλγόριθμο Otsu πάνω στο ιστόγραμμα της έντασης. Δεν λαμβάνουμε υπόψη την απόχρωση. Η διαφορά στις δύο περιπτώσεις έχει να κάνει με την χρήση του αλγορίθμου Unfill που περιγράψαμε στην ενότητα 4.7.3. Τα σημεία με μωβ κύκλο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill του spot. Αντίθετα στα σημεία με κόκκινο ρόμβο δεν έχει χρησιμοποιηθεί ο αλγόριθμος Unfill. Με μπλε τετράγωνο έχουμε το σφάλμα σύμφωνα με τα δοσμένα από την πηγή μας αποτελέσματα κατάτμησης.



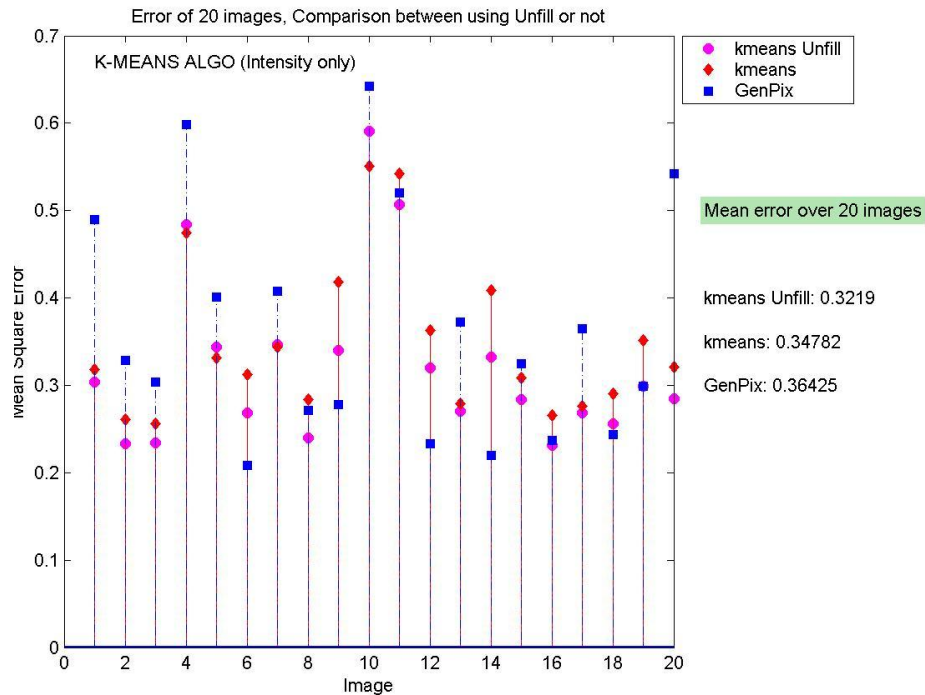
Γρ. 5.6.2 Αντίστοιχη περίπτωση με την γραφική 6.5.1. Τα σημεία με μωβ κύκλους και κόκκινους ρόμβους αντιπροσωπεύουν το σφάλμα που έχουμε για τις 20 εικόνες εάν εφαρμόσουμε επαναληπτικό αλγόριθμο Otsu πάνω στο ιστόγραμμα της έντασης. Συμπληρωματικά εφαρμόζουμε φίλτρο πάνω στο ιστόγραμμα της απόχρωσης. Η διαφορά μεταξύ των δύο έχει να κάνει και πάλι με την χρήση του αλγόριθμου Unfill που περιγράψαμε στην ενότητα 4.7. Τα σημεία με μωβ κύκλο μας δίνουν το σφάλμα για κάθε εικόνα όταν εφαρμόσουμε συμπληρωματικά τον αλγόριθμο Unfill του spot ενώ με κόκκινο ρόμβο δεν έχει χρησιμοποιηθεί ο αλγόριθμος Unfill. Με μπλε τετράγωνο έχουμε και πάλι το σφάλμα σύμφωνα με τα δοσμένα από την πηγή μας αποτελέσματα κατάτμησης.



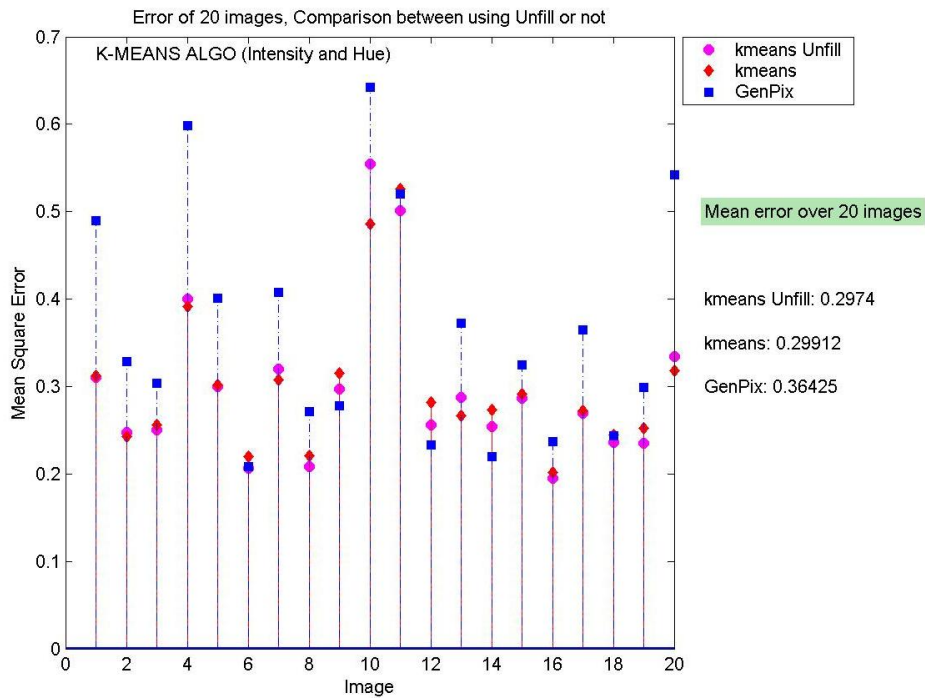
Γρ. 5.6.3 Σύγκριση μεταξύ δυσδιάστατης και μονοδιάστατης κατάτμησης. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του επαναληπτικού Otsu πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα πάνω στις 20 εικόνες για τις τρεις περιπτώσεις που συγκρίνουμε.

5.6.2 Απόδοση αλγορίθμου k-means.

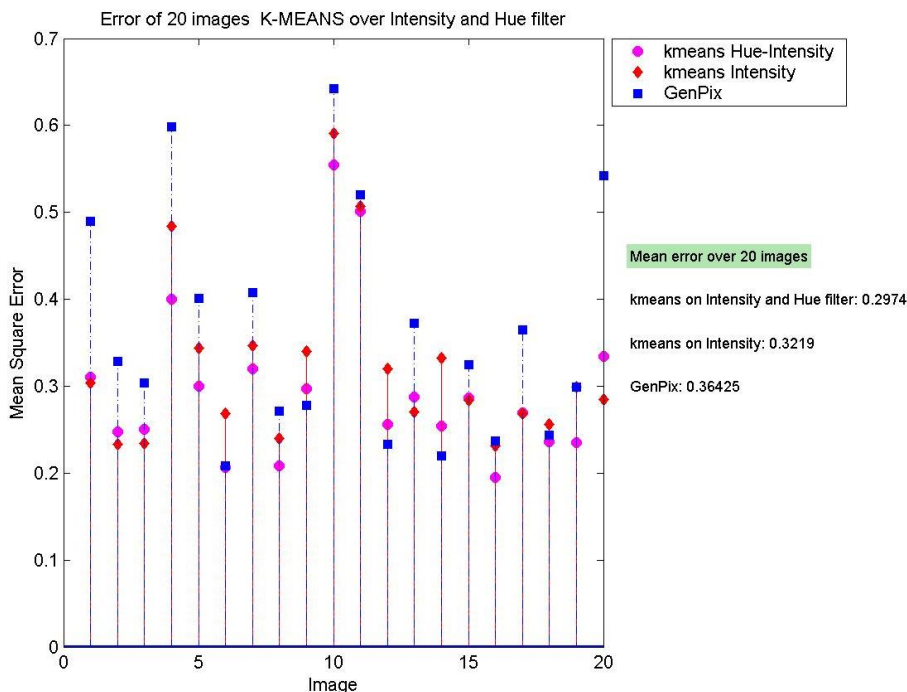
Εφαρμόζουμε επαναληπτικό αλγόριθμο k-means όπως τον περιγράψαμε στην ενότητα 3.1.2.3 πάνω στις ίδιες εικόνες. Έχουμε τις αντίστοιχες συγκρίσεις και τα ανάλογα γραφήματα με την προηγούμενη ενότητα.



Γρ. 5.6.4 Αλγόριθμος συμπλήρωσης του spot όταν έχουμε κατάτμηση μόνο βάσης έντασης. Εφαρμόζουμε επαναληπτικό αλγόριθμο k-means όπως τον περιγράψαμε στην ενότητα 5.2, κατάτμηση μόνο βάσης την έντασης. Με μωβ κύκλους έχουμε το σφάλμα για τις 20 εικόνες όταν εφαρμόζουμε συμπληρωματικά τον αλγόριθμο Unfill (ενότητα 4.7.3). Με κόκκινους ρόμβους το σφάλμα των 20 εικόνων με την διαφορά ότι δεν έχουμε Unfilling. Παραθέτουμε και το σφάλμα που υπολογίζουμε με βάση τα δοσμένα αποτελέσματα κατάτμησης (μπλε τετράγωνα). Στα δεξιά παραθέτουμε το συνολικό σφάλμα των 20 εικόνων για τις περιπτώσεις που συγκρίνουμε.



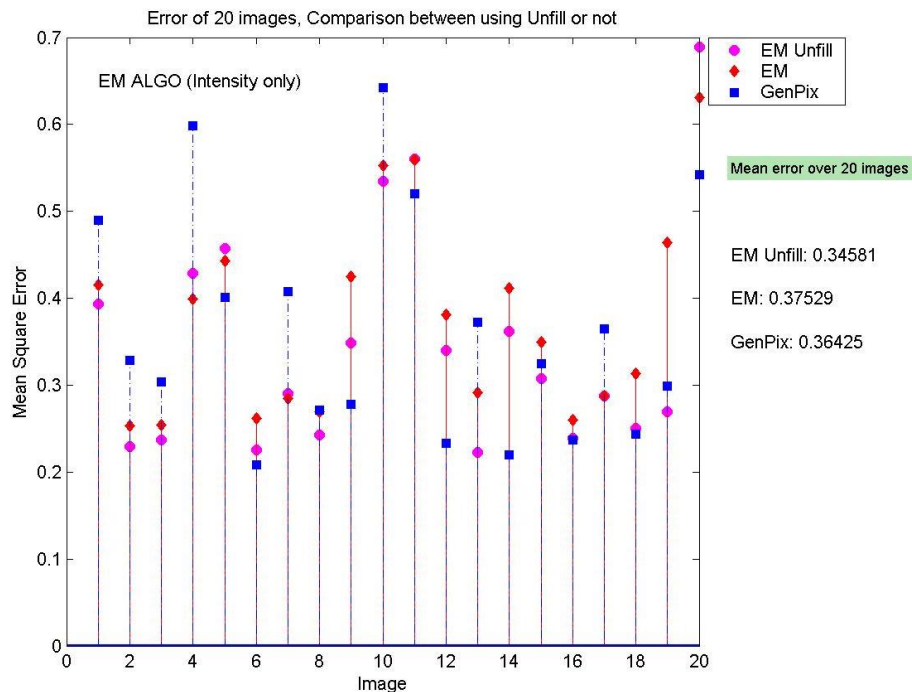
Γρ. 5.6.5 Αλγόριθμος συμπλήρωσης του spot όταν έχουμε κατάτμηση βάση έντασης και απόχρωσης μαζί. Αντίστοιχη σύγκριση με το προηγούμενο γράφημα με την διαφορά ότι εφαρμόζουμε επαναληπτικό αλγόριθμο k-means πάνω στην ένταση σε συνδυασμό με την εφαρμογή φίλτρου πάνω στην απόχρωση. Συγκρίνουμε και πάλι την εφαρμογή του αλγόριθμου Unfill (μωβ κύκλοι) και την μη εφαρμογή του (κόκκινοι ρόμβοι). Στα δεξιά είναι και πάλι το συνολικό σφάλμα για τις τρεις περιπτώσεις.



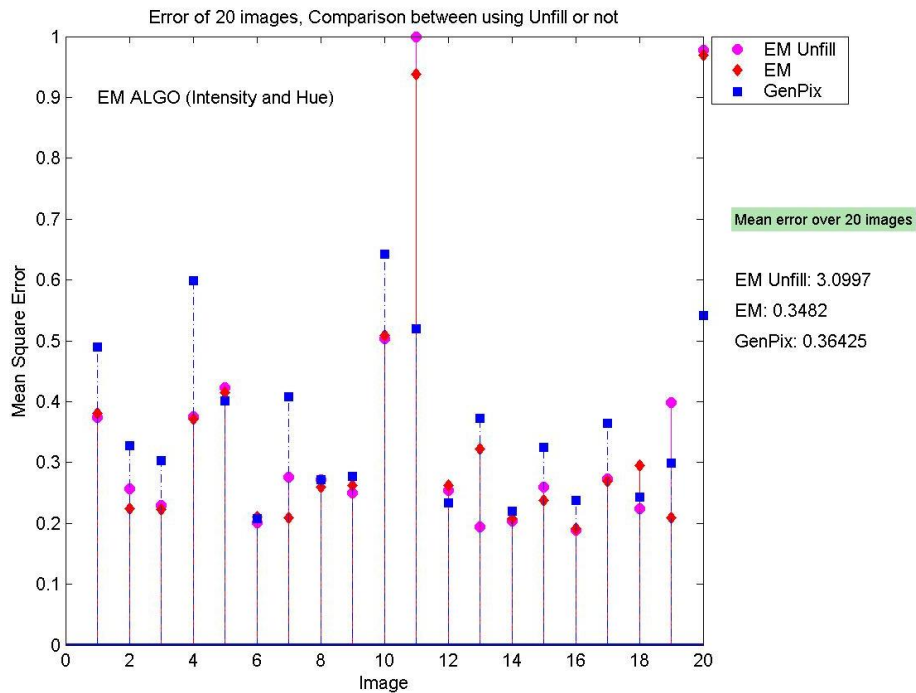
Γρ. 5.6.6 Σύγκριση μεταξύ δυοδιάστατης και μονοδιάστατης κατάτμησης. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του επαναληπτικού k-means πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα των 20 εικόνων για τις περιπτώσεις που συγκρίνουμε.

5.6.3 Απόδοση επαναληπτικού αλγορίθμου EM

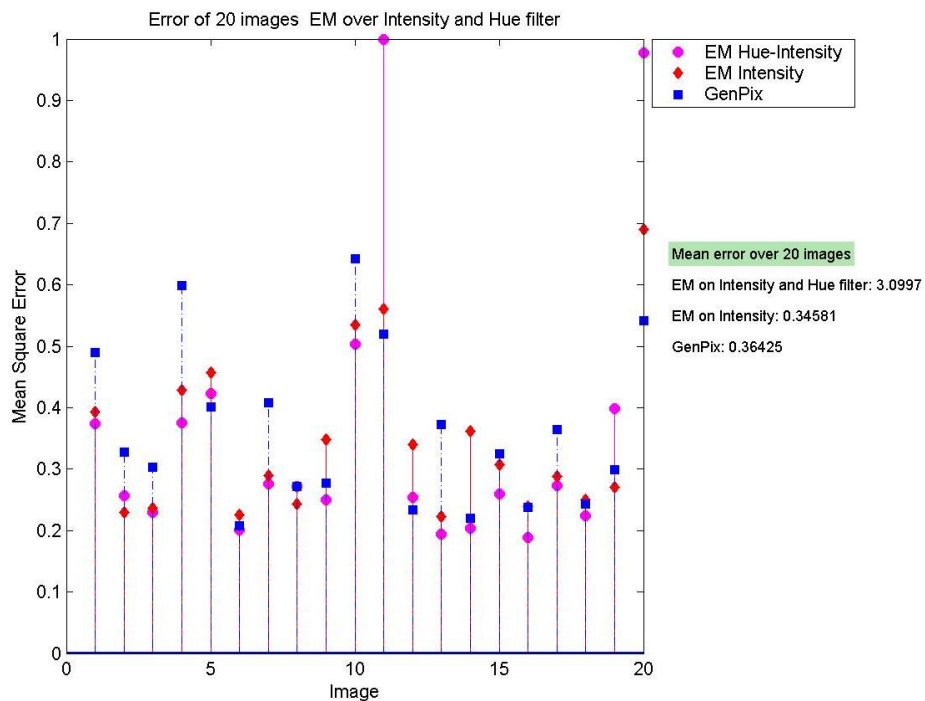
Εφαρμόζουμε το επαναληπτικό αλγόριθμο Expectation Maximization και πραγματοποιούμε τις ίδιες συγκρίσεις όπως κάναμε και στους δύο προηγούμενους αλγόριθμους.



Γρ. 5.6.7 Αλγόριθμος συμπλήρωσης των *spot* όταν έχουμε κατάτμηση μόνο βάσης έντασης. Εφαρμόζουμε επαναληπτικό αλγόριθμο EM όπως τον περιγράψαμε στην ενότητα 9.8 κατάτμηση μόνο βάσης την έντασης. Με κόκκινους ρόμβους έχουμε το σφάλμα για τις 20 εικόνες όταν εφαρμόζουμε συμπληρωματικά τον αλγόριθμο Unfill (ενότητα 4.7.3). Με μωβ κύκλους το σφάλμα των 20 εικόνων με την διαφορά ότι δεν έχουμε Unfill. Παραθέτουμε και το σφάλμα που υπολογίζουμε με βάση τα δοσμένα αποτελέσματα κατάτμησης (μπλε τετράγωνα). Στα δεξιά παραθέτουμε το συνολικό σφάλμα των 20 εικόνων για τις περιπτώσεις που συγκρίνουμε.



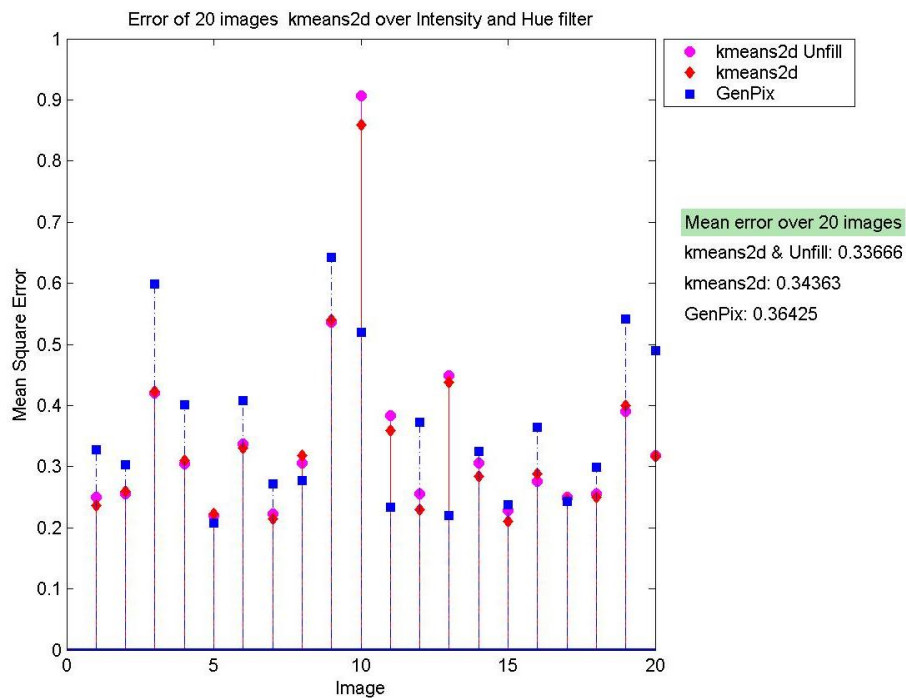
Γρ. 5.6.8 Αλγόριθμος συμπλήρωσης του spot όταν έχουμε κατάτμηση βάση έντασης και απόχρωσης μαζί. Αντίστοιχη σύγκριση με το προηγούμενο γράφημα με την διαφορά ότι εφαρμόζουμε επαναληπτικό αλγόριθμο EM πάνω στην ένταση σε συνδυασμό με την εφαρμογή φίλτρου πάνω στην απόχρωση. Συγκρίνουμε και πάλι την εφαρμογή του αλγόριθμου Unfill (Μωβ κύκλοι) και την μη εφαρμογή του (κόκκινοι ρόμβοι). Στα δεξιά είναι και πάλι το συνολικό σφάλμα για τις τρεις περιπτώσεις.



Γρ. 5.6.9 Σύγκριση μεταξύ δυοδιάστατης και μονοδιάστατης κατάτμησης. Στο γράφημα αυτό συγκρίνουμε το συνολικό σφάλμα στην περίπτωση όπου έχουμε μόνο εφαρμογή του επαναληπτικού EM πάνω στην ένταση με την περίπτωση όπου εφαρμόζουμε επιπρόσθετα ένα φίλτρο πάνω στην απόχρωση (όπως το περιγράψαμε στην ενότητα 3.2). Τα σημεία με μπλε τετράγωνο αφορούν πάλι το σφάλμα από τα δοσμένα από την πηγή των εικόνων μας αποτελέσματα κατάτμησης. Στα δεξιά παραθέτουμε το συνολικό σφάλμα πάνω στις 20 εικόνες για τις τρεις περιπτώσεις που συγκρίνουμε.

5.6.4 Απόδοση δυσδιάστατου 2-means αλγορίθμου

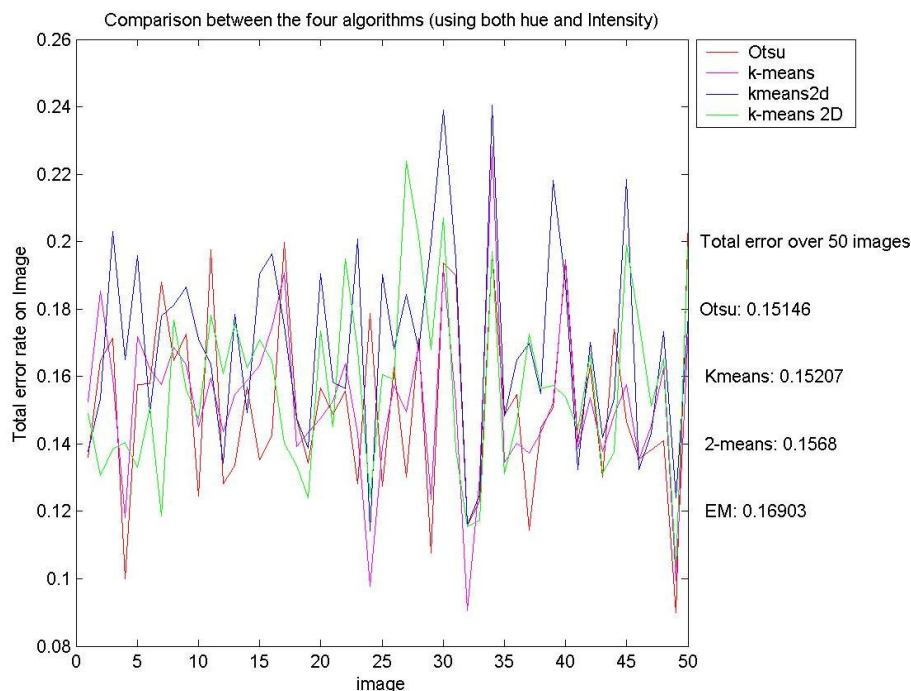
Ο δυσδιάστατος K-means όπως τον περιγράψαμε στην ενότητα 3.2.1.1 χρησιμοποιεί σε συνδυασμό τις τιμές της έντασης και της απόχρωσης σε μια δυσδιάστατη κατανομή. Επομένως η μόνη σύγκριση που μπορούμε να κάνουμε είναι σε σχέση με τον αλγόριθμο συμπλήρωσης του spot (ενότητα 4.7). Χρησιμοποιούμε δηλαδή συμπληρωματικά τον αλγόριθμο center filling και τον συγκρίνουμε με την περίπτωση όπου δεν συμπεριλαμβάνουμε στον υπολογισμό της έκφρασης τα εσωτερικά Pixel του spot.



Εικ. 5.6.10 Δυσδιάστατος επαναληπτικός 2-Means. Η παραπάνω γραφική προέκυψε από την εφαρμογή του δυσδιάστατου k-means πάνω στις 20 εικόνες στην μια περίπτωση χρησιμοποιώντας τον αλγόριθμο συμπλήρωσης και στην άλλη όχι (μωβ κύκλοι και κόκκινοι ρόμβοι αντίστοιχα). Παραθέτουμε το σφάλμα από τα δοσμένα αποτελέσματα κατάταξης και το συνολικό σφάλμα των 20 εικόνων για τις περιπτώσεις που συγκρίνουμε.

5.7 Σύγκριση αλγορίθμων

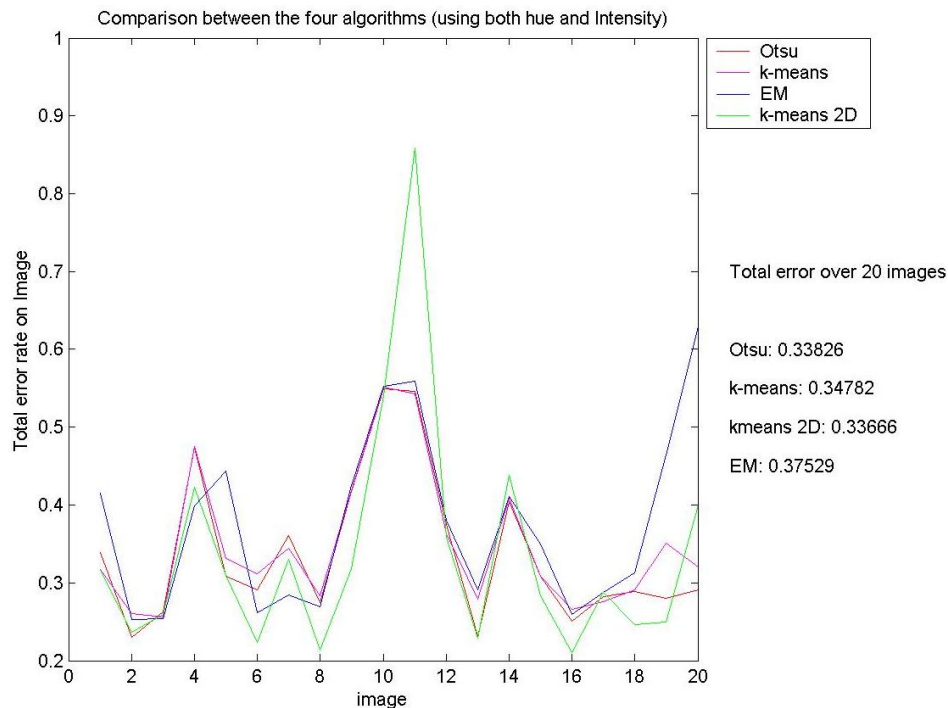
Μια άλλη βασική σύγκριση που θέλουμε να κάνουμε είναι μεταξύ των τεσσάρων αλγορίθμων. Για να έχει νόημα η σύγκριση του δυσδιάστατου K-means με τους υπόλοιπους τρεις αλγορίθμους χρησιμοποιούμε το συμπληρωματικό φίλτρο πάνω στην απόχρωση έτσι όλοι οι αλγόριθμοι να χρησιμοποιούν και τα δύο χαρακτηριστικά. Εξάλλου όπως είδαμε και στα παραπάνω αποτελέσματα η χρήση του δεύτερου χαρακτηριστικού βελτιώνει τα αποτελέσματα. Επίσης και στις τέσσερις περιπτώσεις δεν χρησιμοποιήσαμε τον αλγόριθμο του συμπλήρωσης καθώς όπως εξηγήσαμε στην ενότητα μικρότερο σφάλμα αν δε γεμίσουμε το spot. Έχουμε τα ίδιο γράφημα με τα προηγούμενα με την διαφορά ότι ενώνουμε τα σημεία για καλύτερη απεικόνιση της διαφοράς των αλγορίθμων.



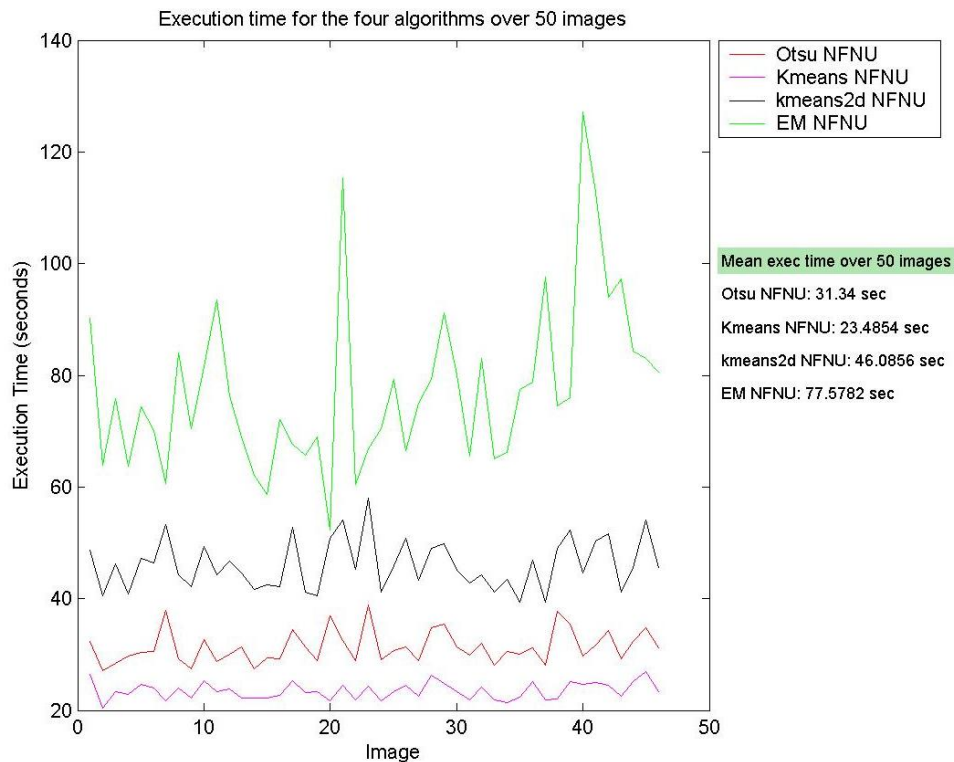
Γρ. 5.7.1 Σύγκριση των τεσσάρων αλγορίθμων μεταξύ τους. Η παραπάνω γραφικές αποτελούν στα σφάλματα πάνω στις 50 εικόνες που προκύπτουν από την εφαρμογή των αλγορίθμων Otsu, k-means, EM και διδιάστατου k-means. Για να μπορούμε να κάνουμε την σύγκριση με τον διδιάστατο k-means για τους υπόλοιπους τρεις αλγορίθμους εφαρμόζουμε επιπρόσθετα φίλτρο πάνω στο χαρακτηριστικό της απόχρωσης. Επίσης σε όλους τους αλγορίθμους παραλείψαμε την εφαρμογή του αλγόριθμου συμπλήρωσης του spot.

Σύγκριση μεταξύ των τεσσάρων αλγορίθμων σε **επίπεδο σφάλματος** και σε **επίπεδο χρόνου εκτέλεσης**. Όπως και στην περίπτωση των εικόνων με spot ελέγχου συγκρίνουμε το σφάλμα των τεσσάρων αλγορίθμων. Για να είναι ισότιμη η σύγκριση σε όλες τις περιπτώσεις αφαιρέσαμε τον αλγόριθμο center fill εφόσον δίνει και μεγαλύτερο σφάλμα όπως ήδη εξηγήσαμε. Επίσης στους αλγόριθμους EM, k-means και Otsu εφαρμόζουμε το συμπληρωματικό φίλτρο στην κατανομή της απόχρωσης.

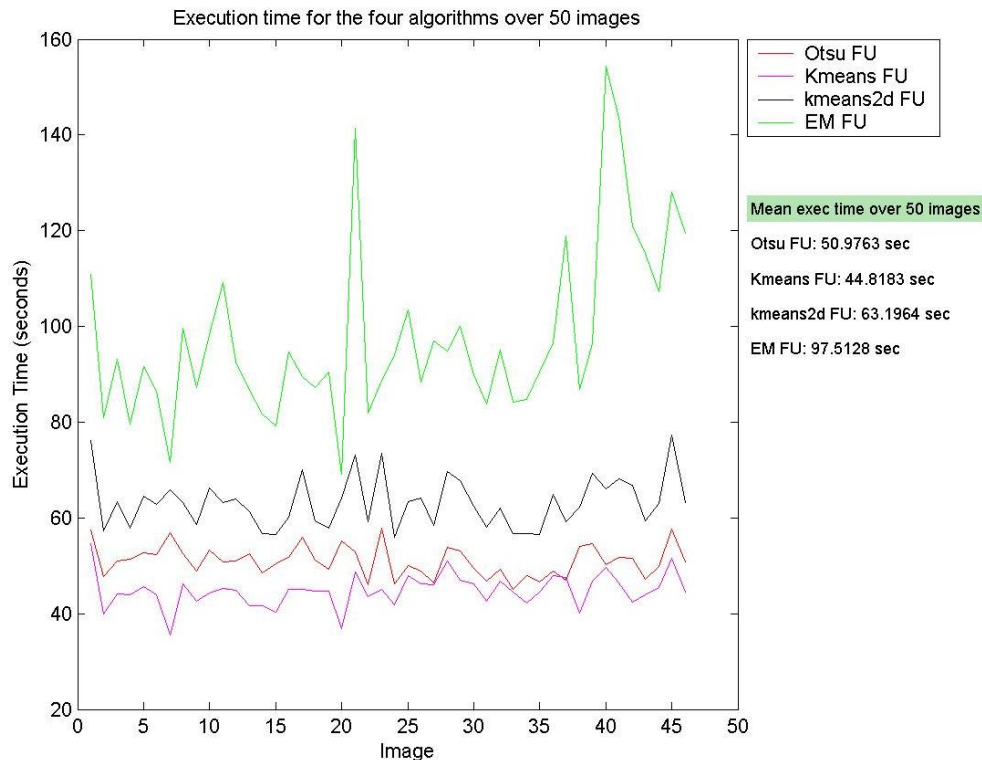
Πραγματοποιούμε μια δεύτερη σύγκριση σε **επίπεδο χρόνου εκτέλεσης** κάθε αλγορίθμου. Ο υπολογισμός των χρόνων γίνεται με τις ίδιες ρυθμίσεις όπως και στον υπολογισμό του σφάλματος: χωρίς γέμισμα του κέντρου του spot και με συνδυασμό των δύο χαρακτηριστικών.



Γρ. 5.7.2 Σύγκριση των τεσσάρων αλγορίθμων μεταξύ τους. Η παραπάνω γραφικές αποτελούν στα σφάλματα πάνω στα 20 microarrays ομοτυπικών υβριδισμών που προκύπτουν από την εφαρμογή των αλγορίθμων Otsu, k-means, EM και δυσδιάστατου k-means. Για να μπορούμε να κάνουμε την σύγκριση με τον δυσδιάστατο k-means για τους υπόλοιπους τρεις αλγορίθμους εφαρμόζουμε επιπρόσθετα φίλτρο πάνω στο χαρακτηριστικό της απόχρωσης. Επίσης σε όλους τους αλγορίθμους παραλείψαμε την εφαρμογή του αλγόριθμου συμπλήρωσης του spot. Παρατηρούμε ότι όλοι οι αλγόριθμοι έχουν παρόμοια συμπεριφορά δηλαδή δίνουν μεγαλύτερο σφάλμα στις ίδιες εικόνες.



Γρ. 5.7.3 Σύγκριση των τεσσάρων αλγορίθμων ως προς τον χρόνο εκτέλεσης τους. Εκτελούμε τους τέσσερις αλγορίθμους χωρίς τον αλγόριθμο συμπλήρωσης του spot (NFNU-No Fill No Unfill) και με συνδυασμό της έντασης και της απόχρωσης και μετράμε τον χρόνο εκτέλεσης σε δευτερόλεπτα. Οι αλγόριθμοι Otsu και K-means έχουν παρόμοιους χρόνους εκτέλεσης με τον k-means να είναι λίγο πιο γρήγορος. Ο EM είναι ο πιο αργός καθώς υπολογίζει περισσότερες παραμέτρους και ο επαναληπτικός δυσδιάστατος k-means επειδή θέλει πολλές επαναλήψεις.



Γρ. 5.7.4 Σύγκριση των τεσσάρων αλγορίθμων ως προς τον χρόνο εκτέλεσης τους. Εκτελούμε τους τέσσερις αλγόριθμους και έπειτα τον αλγόριθμο συμπλήρωσης του spot και αφαίρεσης των περιφερειακών pixel (FU-Fill Unfill) και με συνδυασμό της έντασης και της απόχρωσης και μετράμε τον χρόνο εκτέλεσης σε δευτερόλεπτα. Οι αλγόριθμοι Otsu και K-means έχουν παρόμοιους χρόνους εκτέλεσης με τον k-means να είναι λίγο πιο γρήγορος. Ο EM είναι ο πιο αργός καθώς υπολογίζει περισσότερες παραμέτρους και ο επαναληπτικός δυσδιάστατος k-means επειδή θέλει περισσότερες. Αν συγκρίνουμε με το προηγούμενο γράφημα οι χρόνοι είναι λίγο προσυζημένοι για τον λόγο ότι εφαρμόζεσαι ενδιάμεσα ο αλγόριθμος συμπλήρωσης του spot. Και σε αυτήν την περίπτωση οι αλγόριθμοι συμπεριφέρονται αναλογικά με παραπάνω.

5.8 Αλγόριθμος Συμπλήρωσης Spot

Συγκρίναμε την απόδοση των αλγορίθμων Otsu, k-means και EM στην περίπτωση όπου χρησιμοποιούμε μόνο την ένταση και όταν συνδυάζουμε την ένταση με την απόχρωση και υπολογίσαμε το σφάλμα για αυτές τις δύο περιπτώσεις όταν δεν απαλοιφή των περιφερειακών pixel του spot. Από τις γραφικές που παραθέσαμε την ενότητα 5.5 και συγκεκριμένα τις δύο πρώτες γραφικές κάθε υπο-ενότητας (5.5.1, 5.5.2, 5.5.4, 5.5.5, 5.5.7, 5.5.8) καθώς και την γραφική 5.5.10 από τον δυσδιάστατο αλγόριθμο k-means μπορούμε να διαπιστώσουμε ότι η χρήση του αλγόριθμου απαλοιφής των περιφερειακών pixel του spot αυξάνει το συνολικό σφάλμα σε όλες τις περιπτώσεις εκτός από την περίπτωση της απόχρωσης. Δηλαδή ο συνδυασμός του διορθωτικού αυτό αλγόριθμου πριν την εφαρμογή του φίλτρου στην κατανομή της απόχρωσης βελτιώνει πάρα πολύ το σφάλμα καθώς διορθώνει την κατανομή της απόχρωσης.

Στις τις εικόνες 5.5.2, 5.5.5, 5.5.8 συγκρίναμε για κάθε αλγόριθμο στην περίπτωση όπου συνδυάζουμε τα δύο χαρακτηριστικά και για τις τρεις περιπτώσεις NFNU, NF, FU, δηλαδή με καθόλου διόρθωση του χώρου του spot, με απαλοιφή των περιφερειακών και με απαλοιφή και συμπλήρωση του κέντρου αντίστοιχα. Από τα γραφήματα επιβεβαιώνεται η υπόθεση μας ότι η απαλοιφή των περιφερειακών pixel βελτιώνει την απόδοση των αλγορίθμων. Επίσης για λόγους που εξηγήσαμε στην ενότητα 4.7, η συμπερίληψη των εσωτερικών σημείων του spot αυξάνει το σφάλμα.

Στην περίπτωση των εικόνων με ομοτυπικούς υβριδισμούς και για τα αντίστοιχα γραφήματα 5.6.1, 5.6.2, 5.6.4, 5.6.5, 5.6.7, 5.6.8 παρατηρούμε και πάλι ότι η απαλοιφή των περιφερειακών pixel (NFNU) βελτιώνει γενικά το σφάλμα.



6 Συμπεράσματα

6.1 Μονοδιάστατη και δυσδιάστατη κατάτμηση

Στις γραφικές 5.5.3, 5.5.6 και 5.5.9 (αντίστοιχα 5.6.3, 5.6.6 και 5.6.9) συγκρίνουμε την απόδοση των τριών αλγορίθμων όταν έχουμε συμπληρωματική χρήση της απόχρωσης σε σχέση με την απλή περίπτωση όπου για την κατάτμηση χρησιμοποιούμε μόνο την ένταση. Σε όλες τις περιπτώσεις εφαρμόζουμε συμπληρωματικά τον αλγόριθμο απαλοιφής των περιφερειακών pixel (Unfil- NF)

Η απόχρωση όπως έχουμε αναφέρει στην ενότητα 3.2 δρα συμπληρωματικά και στην περίπτωση όπου έχουμε κακή ευθυγράμμιση των δύο εικόνων (Cy3,Cy5). Απλά θυμίζουμε ότι σε αυτήν την περίπτωση έχουμε κάποια pixel στην περιφέρεια του spot που προέρχονται μόνο από ένα κανάλι, δηλαδή είτε έχουν τιμές έντασης μόνο του κόκκινου ή μόνο του πράσινου. Αυτά τα pixel όπως είχαμε υποθέσει περιείχαν εσφαλμένες τιμές έντασης και έπρεπε να αφαιρεθούν από το spot. Η αλήθεια της παραπάνω υπόθεσης επιβεβαιώνεται από τις γραφικές αυτές καθώς και οι τρεις αλγόριθμοι δίνουν μικρότερο σφάλμα στην περίπτωση όπου για την κατάτμηση χρησιμοποιούμε και την πληροφορία που μας δίνει η απόχρωση.

6.2 Δυσδιάστατη και Διαδοχική Κατάτμηση – Ασυσχέτιστα Χαρακτηριστικά

Όπως διαπιστώθηκε στην ενότητα 3.2.2.1 τα χαρακτηριστικά της έντασης και της Απόχρωσης έχουν μικρή συσχέτιση. Το γεγονός αυτό μας επέτρεψε την διαδοχική κατάτμηση πρώτα βάση έντασης και έπειτα βάση απόχρωσης. Η γραφική 5.7.1 παρουσιάζει τα σφάλματα για την περίπτωση του δυσδιάστατου K-means και των άλλων τριών περιπτώσεων οι οποίες αποτελούν διαδοχική κατάτμηση με την μόνη αλλαγή στον αλγόριθμο που θέτει το όριο στο πρώτο επίπεδο απόφασης (με βάση την ένταση). Μέσω της γραφικής αυτής μπορεί κανείς να διαπιστώσει ότι έχουμε παρόμοια σφάλματα και στην δυσδιάστατη προσέγγιση και στην διαδοχική κατάτμηση. Το γεγονός αυτό συνηγορεί στην εκτίμηση που κάναμε ότι τα δύο χαρακτηριστικά είναι ασυσχέτιστα.

6.3 Σύγκριση αλγορίθμων

Μια γενική παρατήρηση που μπορούμε να κάνουμε είναι ότι *οι αλγόριθμοι συμπεριφέρονται με παρόμοιο τρόπο*, δηλαδή δίνουν όλοι μεγαλύτερο σφάλμα σε συγκεκριμένες εικόνες και μικρότερο σε κάποιες άλλες. Οι αλγόριθμοι Otsu k-means και δυσδιάστατος k-means δίνουν παρόμοιες τιμές σφάλματος. Ο EM δίνει λίγο μεγαλύτερο σφάλμα.

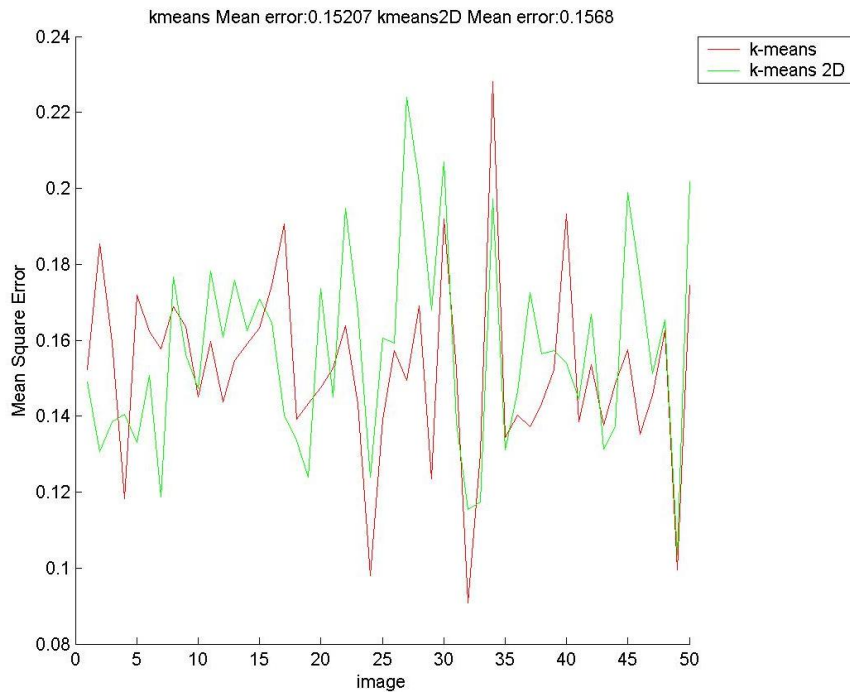
Στην περίπτωση των ομοτυπικών υβριδισμών ο k-means δύο διαστάσεων δίνει λίγο μικρότερο σφάλμα από τον K-means ενώ το αντίστροφο συμβαίνει με τις εικόνες με spot ελέγχου. Συμπερασματικά ο δυσδιάστατος και ο απλός k-means σε συνδυασμό με φίλτρο στην απόχρωση έχουν παρόμοιες επιδόσεις. *Αυτό συνηγορεί περισσότερο στην υπόθεση μας ότι τα δύο χαρακτηριστικά είναι πρακτικά ασυσχέτιστα.*

Ο αλγόριθμος Otsu αποδείχτηκε ένας πολύ λιτός και αποδοτικός αλγόριθμος. Σε σύγκριση με τον αλγόριθμο EM δίνει καλύτερα αποτελέσματα παρόλο που ο EM περιέχει περισσότερες παραμέτρους και θα έπρεπε να κάνει καλύτερη εκτίμηση των δύο καμπύλων.

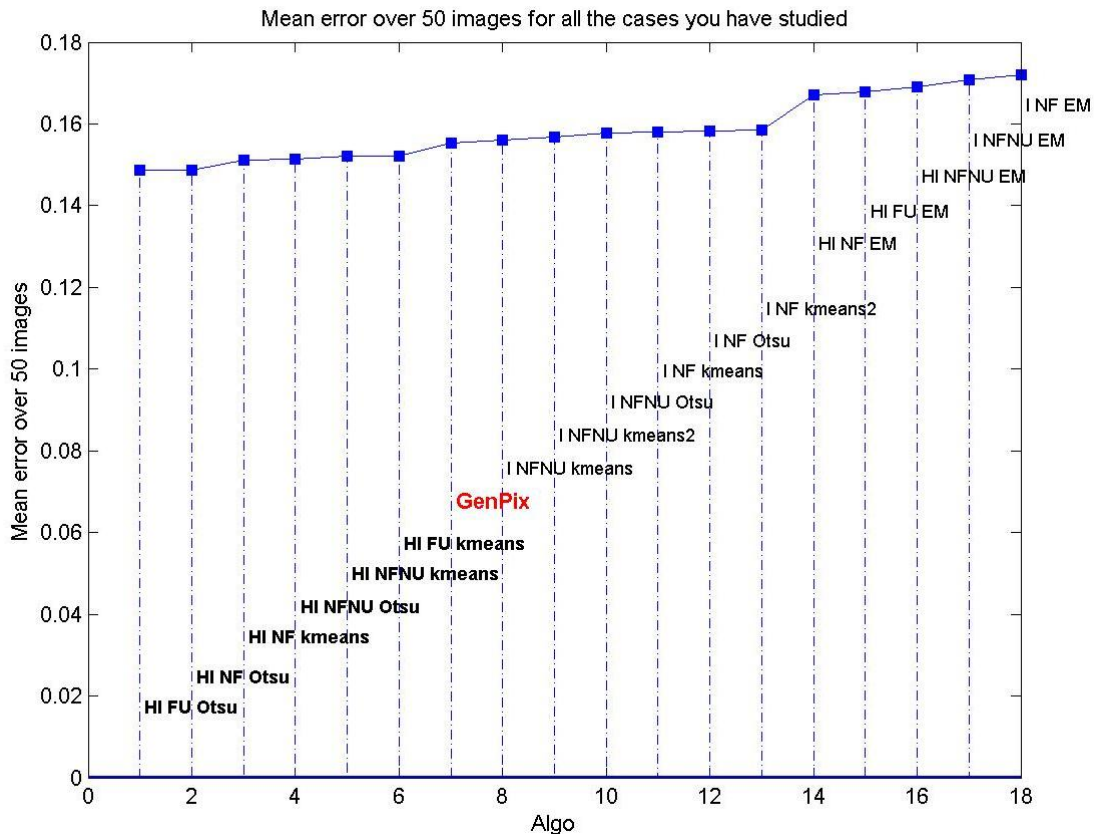
Υπάρχουν τρεις λόγοι για τους οποίους ο EM δεν είχε το μικρότερο σφάλμα όπως ίσως θα περιμέναμε. Ο πρώτος λόγος έχει να κάνει με το γεγονός ότι ο EM προσεγγίζει τις παραμέτρους δύο Gaussian καμπύλων ίσως οι καμπύλες μας να μην είναι στην πραγματικότητα Gaussian. Επίσης όπως έχουμε ήδη σημειώσει πολλές φορές το ιστόγραμμα της έντασης προσεγγίστηκε ως άθροισμα δύο Gaussian καμπύλων ενώ όπως διαπιστώσαμε υπήρχαν τουλάχιστον τρεις καμπύλες. Ο δεύτερος λόγος

έχει να κάνει με την απλότητα του προβλήματος. Δηλαδή ίσως δεν χρειαζόταν ο υπολογισμός τόσων παραμέτρων για να έχουμε ένα ικανοποιητικό αποτέλεσμα.

Επίσης ανάμεσα στους K-means και Otsu ο πρώτος δίνει μικρότερο χρόνο εκτέλεσης ενώ ο δεύτερος μικρότερο σφάλμα. Αν λοιπόν είχαμε να επιλέξουμε ανάμεσα στους τέσσερις αυτούς αλγορίθμους ο K-means είναι μια πολύ καλή επιλογή καθώς δίνει πολύ καλά αποτελέσματα και σε μικρό χρόνο εκτέλεσης.



Γρ. 6.3.1 Σύγκριση δυσδιάστατου και μονοδιάστατου k-means. Εφαρμόζουμε k-means στο ιστόγραμμα της έντασης και bandpass φίλτρο στην απόχρωση συγκρίνουμε τα αποτελέσματα με τον δυσδιάστατο k-means. Όπως φαίνεται στην παραπάνω γραφική έχουμε παρόμοια συμπεριφορά γεγονός που συνηγορεί στην υπόθεση μας ότι τα δύο χαρακτηριστικά είναι ανεξάρτητα, καθώς επίσης και στο γεγονός ότι η πληροφορία που μας δίνει η απόχρωση είναι συμπληρωματική.



Γρ. 6.7.2 Σύγκριση όλων των πειραματικών αποτελεσμάτων – Εικόνες με Spot Ελέγχου. Με την ονομασία HI σημαίνει ότι χρησιμοποιούμε και τα δύο χαρακτηριστικά, με I είναι χρησιμοποιώντας μόνο την ένταση. Το NFNU σημαίνει No Center Fill και No Unfill, δηλαδή χωρίς αλγόριθμο συμπλήρωσης και αλγόριθμο απαλοιφής, αντίστοιχα NF σημαίνει No Fill, και FU εφαρμόζουμε Fill και Unfill. Επίσης k-means είναι ο απλός k-mean και k-means 2 είναι ο δυσδιάστατος.

από άλλες μελέτες που έχουν γίνει σε μεθόδους κατάτμησης είναι ότι εμείς δουλέψαμε στο σύστημα HSI αντί του RGB. Η προσέγγιση αυτή έχει τα παρακάτω πλεονεκτήματα.

- **Μείωση της τάξης του προβλήματος κατά μία διάσταση.**

Όπως έχουμε σημειώσει πολλές φορές η *τιμή της έντασης* παίζει τον βασικό ρόλο στην διαδικασία της κατάτμησης και στην εξαγωγή του ποσοστού έκφρασης. Και στα δύο συστήματα αναπαράστασης έχουμε τρεις μεταβλητές που περιγράφουν την τιμή του χρώματος για κάθε Pixel. Στο σύστημα RGB το χρώμα περιγράφεται μόνο από τιμές εντάσεων (κόκκινου, πράσινου, μπλε). Στο HSI σύστημα εκτός από την ένταση έχουμε άλλες δύο ιδιότητες του χρώματος: απόχρωση (hue) που μας δίνει το χρωματισμό και την πρόσμιξη (Saturation) που περιγράφει το πόσο σκούρο η αραιωμένο είναι το χρώμα.

Με την μεταφορ-ά από το RGB στο HSI κατά κάποιον τρόπο συνδυάζονται όλες οι τιμές εντάσεων R,G,B σε μια I. Συγκεκριμένα στην περίπτωση μας όπου μας δίνονται δύο εικόνες μια στο κόκκινο και μια στο πράσινο έχουμε τιμές εντάσεων μόνο στους άξονες R και G και με την μετατροπή καταλήγουμε σε μια τιμή I.

Έτσι ενώ η κατάτμηση σε RGB είναι δυσδιάστατη - δύο χαρακτηριστικά ένταση κόκκινου και ένταση πράσινου - σε HSI είναι μονοδιάστατη - μόνο μια τιμή I. Κατ' αυτό τον τρόπο μπορούμε να πούμε ότι μειώνεται κατά ένα η τάξη του προβλήματος ταξινόμησης. Κατά συνέπεια έχουμε πιο απλούς αλλά και πιο γρήγορους αλγόριθμους.

- **Αξιοποίηση της πληροφορίας που μας δίνει τιμή του χρώματος (Hue).**

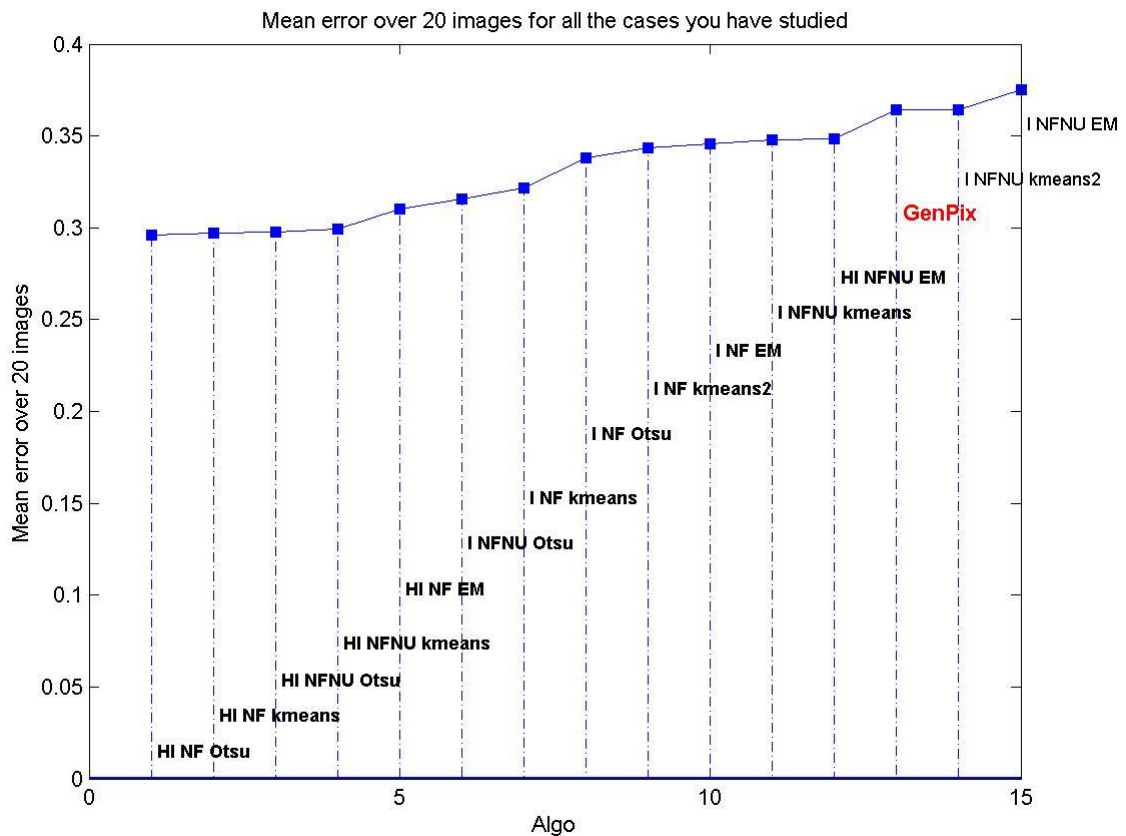
Αναφέραμε παραπάνω ότι το σύστημα HSI έχουμε την μεταβλητή Hue η οποία μας δίνει τιμή χρώματος. Έτσι στο σύστημα αυτό έχουμε την δυνατότητα στην διαδικασία της

κατάτμησης να αξιοποιήσουμε αυτή την μεταβλητή και να την συνδυάσουμε με την μεταβλητή της έντασης.

Ανεβάζουμε και πάλι την τάξη του προβλήματος κατά ένα αλλά έχουμε στην διάθεση μας περισσότερη πληροφορία. Η πληροφορία της απόχρωσης είναι συμπληρωματική και μόνο στις περιπτώσεις όπου έχουμε κακή ευθυγράμμιση των δύο εικόνων (CY3,Cy5). Με την απόχρωση αναιρούμε τα Pixel που έχουν τιμές έντασης μόνο στο πράσινο ή μονό στο κόκκινο και βελτιώνουμε με αυτό τον τρόπο το συνολικό σφάλμα.

HI_FU_Otsu	0.14873
HI_NF_Otsu	0.14879
HI_NF_kmeans	0.15124
HI_NFNU_Otsu	0.15146
HI_NFNU_kmeans	0.15207
HI_FU_kmeans	0.15222
GenPix	0.15535
I_NFNU_kmeans	0.15601
I_NFNU_Otsu	0.15772
I_NF_kmeans	0.15808
I_NF_Otsu	0.15815
HI_NF_EM	0.1672
HI_FU_EM	0.1679
HI_NFNU_EM	0.16903
I_NFNU_EM	0.17075
I_NF_EM	0.17206

Γρ. 6.7.1 Συνολικός πίνακας σφαλμάτων – Εικόνες με Spot Ελέγχου. Ισχύουν τα ίδια με την γραφική 6.7.2 Με την ονομασία HI σημαίνει ότι χρησιμοποιούμε και τα δύο χαρακτηριστικά, με I είναι χρησιμοποιώντας μόνο την ένταση. Το NFNU σημαίνει No Center Fill και No Unfill, δηλαδή χωρίς αλγόριθμο συμπλήρωσης και αλγόριθμο απαλοιφής, αντίστοιχα NF σημαίνει No Fill, και FU εφαρμόζουμε Fill και Unfill. Επίσης k-means είναι ο απλός k-mean και k-means 2 είναι ο δυδιάστατος.



Γρ. 6.7.3 Σύγκριση όλων των πειραματικών αποτελεσμάτων – Εικόνες Ομοτυπικών Υβριδισμών. Με την ονομασία HI σημαίνει ότι χρησιμοποιούμε και τα δύο χαρακτηριστικά, με I είναι χρησιμοποιώντας μόνο την ένταση. Το NFNU σημαίνει No Center Fill και No Unfill, δηλαδή χωρίς αλγόριθμο συμπλήρωσης και αλγόριθμο απαλοιφής, αντίστοιχα NF σημαίνει No Fill, και FU εφαρμόζουμε Fill και Unfill. Επίσης k-means είναι ο απλός k-mean και k-means 2d είναι ο δυοδιάστατος.

HI_NF_Otsu	0.29618
HI_NF_kmeans	0.2974
HI_NFNU_Otsu	0.2978
HI_NFNU_kmeans	0.29912
HI_NF_EM	0.30997
I_NFNU_Otsu	0.31589
I_NF_kmeans	0.3219
I_NF_Otsu	0.33826
I_NF_EM	0.34581
I_NFNU_kmeans	0.34782
HI_NFNU_EM	0.3482
GenPix	0.36425
I_NFNU_EM	0.37529

Γρ. 6.7.2 Συνολικός πίνακας σφαλμάτων – Εικόνες Ομοτυπικών Υβριδισμών. Ισχύουν τα ίδια με την γραφική 6.7.3 Με την ονομασία HI σημαίνει ότι χρησιμοποιούμε και τα δύο χαρακτηριστικά, με I είναι χρησιμοποιώντας μόνο την ένταση. Το NFNU σημαίνει No Center Fill και No Unfill, δηλαδή χωρίς αλγόριθμο συμπλήρωσης και αλγόριθμο απαλοιφής, αντίστοιχα NF σημαίνει No Fill, και FU εφαρμόζουμε Fill και Unfill. Επίσης k-means είναι ο απλός k-mean και k-means 2 είναι ο δυοδιάστατος.

6.5 Μελλοντική Εργασία

Η παρούσα διπλωματική εργασία ασχολήθηκε με κατάτμηση εικόνων μικροσυστοιχιών και την εξαγωγή ποσοστού έκφρασης για κάθε spot της μικροσυστοιχίας για το αντίστοιχο δηλαδή γονίδιο προς μελέτη. Δόθηκε μεγαλύτερη βαρύτητα στην βελτίωση του αποτελέσματος στο στάδιο της κατάτμησης της εικόνας και στην μελέτη της συμπληρωματικής χρήσης του χαρακτηριστικού της απόχρωσης.

Για την κατάτμηση της εικόνας επιλέχθηκαν τρεις αλγόριθμοι από την κατηγορία κατάτμησης μέσω ιστογράμματος: kmeans, Otsu και EM. Οι αλγόριθμοι kmeans και Otsu επιλέχθηκαν για την απλότητα τους και όπως αποδείχθηκε και από τα αποτελέσματα των πειραμάτων απέδωσαν πολύ καλύτερα και σε χρόνο αλλά και σε τιμή σφάλματος απ' ότι θα περίμενε κανείς και σε σχέση με τον τρίτο αλγόριθμο EM.

Το γεγονός αυτό οφείλεται στην μορφή των εικόνων μας η οποίες περιέχουν στην ουσία δύο σήματα: background και spot. Οι αλγόριθμοι kmeans και Otsu υλοποιούνται εύκολα και δίνουν καλά αποτελέσματα για την απλή περίπτωση όπου μια κατανομή προσεγγίζεται ως γραμμικός συνδυασμός δύο διαφορετικών κατανομών. Παρόλο που στην πραγματικότητα το ιστόγραμμα της έντασης ενός κελιού (από την εικόνα μας) αποτελείται από το λιγότερο τρεις κατανομές (background, εσωτερικό spot, περιφέρεια spot), για λόγους που έχουν να κάνουν με διάφορες πηγές θορύβου, η παραδοχή ότι έχουμε δύο κατηγορίες (δύο κατανομές) απέδωσε αρκετά καλά στις περιπτώσεις όπου το spot δεν είχε μεγάλη διασπορά τιμών και ήταν ευδιάκριτο. Επίσης η επαναληπτική εφαρμογή των αλγορίθμων έδωσε ένα πολύ ικανοποιητικό αποτέλεσμα για το σύνολο των περιπτώσεων (αχνά spot, spot με αχνό εσωτερικό, κελιά με τμήματα γειτονικών spot).

Στην περίπτωση του αλγόριθμου EM παρόλο που οι κατανομές περιγράφονται με περισσότερη λεπτομέρεια και υπολογίζονται περισσότεροι παράμετροι έχουμε υψηλότερο ποσοστό σφάλματος γεγονός που εξηγήσαμε ότι έχει να κάνει με την παραδοχή μας ότι οι κατανομές είναι Gaussian αλλά και με την προσέγγιση του αριθμού των κατηγοριών σε δύο.

Μια επέκταση θα μπορούσε να γίνει στην μοντελοποίηση των κατανομών. Θα μπορούσαμε να χρησιμοποιήσουμε εκθετικές ή πολυωνυμικές εκφράσεις όπως επίσης και μίξη αυτών. Θα μπορούσε δηλαδή να εκτιμηθεί ο αριθμός των επιμέρους κατανομών (κριτήριο BIC) και έπειτα κάθε κατανομή να υπολογιστεί με βάση την δική της συνάρτηση. Για παράδειγμα η κατανομή του background επειδή τείνει να έχει μέγιστο στο μηδέν (στα όρια του ιστογράμματος) θα μπορούσε να μοντελοποιηθεί ως εκθετική κατανομή.

Όσον αφορά την χρωματικό μοντέλο HSI το οποίο χρησιμοποιήσαμε θα μπορούσε κατ' αρχήν να εκτιμηθεί η απόδοση των αλγορίθμων σε περίπτωση όπου χρησιμοποιούσαμε πιο ακριβείς συναρτήσεις μετασχηματισμού από το ένα σύστημα στο άλλο. Επίσης μια αντίστοιχη μελέτη θα μπορούσε να γίνει χρησιμοποιώντας το χρωματικό μοντέλο HSV και μια σύγκριση ανάμεσα στα δύο συστήματα και στην καλύτερη αξιοποίηση της πληροφορίας που μας δίνει το χρώμα.



-
- [1] DeRisi, J. L., Iyer, V.R., and Brown, P.O, “Exploring the Metabolic and the Genetic Control of Gene Expression on a Genomic Scale”, *Science*, 278, 680-685, 1997
 - [2] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, Comprehensive Identification of Cell Cycle regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the cell*, 9, 3273-3297, 1998
 - [3] Genetic Science Learning Center, University of Utah, <http://learn.genetics.utah.edu/units/biotech/microarray/>,
 - [4] Rastislav Lukac personal web page <http://www.colorimageprocessing.com>
 - [5] Yee Hwa Yang, Michael J. Buckley, Sandrine Dudoit, and Terence P. Speed, “Comparison of Methods for Image Analysis on cDNA Microarray Data”
 - [6] National Cancer Institute <http://dc.nci.nih.gov/dataSets/geawQCandIA>
 - [7] Eisen, M.B., ScanAnalyze <http://rana.lbl.gov/EisenSoftware.htm>
 - [8] S. Beucher Centre de Morphologie Mathematique Ecole des Mines de Paris, “The Watershed Transformation applied to image segmentation”, *Scanning Microscopy International*, 299-314, 1991
 - [9] Richard Nock and Frank Nielsen, “Statistical Region Merging”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26 No 11, 1452-1458, 2004
 - [10] S. Battiato, G. M. Farinella, G. Puglisi IPLab – Image Processing Laboratory Dipartimento di Matematica e Informatica, University of Catania Italy, “SVG Vectorization by Statistical Region Merging”, *eurographics Italian Chapter Conference*, 2006
 - [11] Isha Dewan and Prakasa Rao, “Mann-Whitney Test for associated Sequences”, *Indian Statistical Institute, Delhi Centre*, 2002/06
 - [12] William H. Horsthemke. Jacob D. Furst, “*DNA Microarray Spot Detection Using Hough Transforms*”, DePaul University. School of Computer Science. Chicago, IL
 - [13] Aly A. Farag, Ayman El Baz, Georgy Gimel'farb, “Density Estimation Using Modified Expectation – Maximization Algorithm For a Linear Combination Of Gaussians”, *International Conference on Image Processing, Singapore*, SINGAPOUR (24/10/2004)
 - [14] Qunhua Li, Chris Fraley, Roger E. Bumgarner, Ka Yee Yeung and Adrian E. Raftery, “Donuts, scratches and blanks: robust model – based segmentation of microarray images”, *Bioinformatics, Gene expression*. Vol. 21 no.12, 2875-2882, 2005
 - [15] Yiming Wu, Kap Luk Chan, Yong Huang, “Image Texture Classification Based on Finite Gaussian Mixture Models”, *School of Electrical and Electronic Engineering Nanyang Avenue, Singapore*
 - [16] Dmitrij Csetverikov Collaborators, “*Basic Algorithms for Digital Image Analysis*” *Lecture 7*, Eötvös Loránd University, Budapest, Hungary. csetverikov@sztaki.hu
 - [17] Dr George Bebis, “The Expectation - Maximization (EM) Algorithm”, *Computer Science Department - Mathematical Methods for Computer Vision*, <http://www.cse.unr.edu/~bebis/MathMethods/EM/lecture.pdf>, 2003
 - [18] Bryan S. Morse, “Lecture 4: Thresholding”, *Brigham Young University, 1998-2000*
 - [19] S. Greenblum, M Krucoff, J Furst, D Raicu “AUTOMATED IMAGE ANALYSIS OF NOISY MICROARRAYS” Sharon Greenblum, Max Krucoff, Department of Biomedical Engineering, Northwestern University, Evanston, IL, facweb.cs.depaul.edu
 - [20] TransCripTome, An Algorithm to grid Automatically DNA Microarrays
 - [21] Alan Wee – Chung Liew, Hong Yan, Mengsu Yang, “Robust adaptive spot segmentation of DNA microarray images”, *Pattern Recognition* 36, 1251 – 1254, (2003)
 - [22] S. Battiato, G. Di Blasi, G. M. Farinella, G. Gallo, G. C. Guarnera, “*Statistical Algorithms For Microarray Image Analysis*” - *Internal IPLAB Report, October, 2005*
 - [23] Antti Lehmussola, Pekka Ruusuvaori and Olli Yli-Harja, “Evaluating the performance of microarray segmentation algorithms”. *Bioinformatics, Gene expression*. Vol.22no.232006, pages 2910–2917 doi:10.1093/bioinformatics/btl502
 - [24] GEMTools Support, <http://www.baymedia.com/gemtools/>
 - [25] National Cancer Institute, GEAW Quality Control and Image Analysis Challenge Data Set, <http://dc.nci.nih.gov/dataSets/geawQCandIA>
 - [26] Jin Hyuk Kim, Hye Young Kim, Yong Sung Lee, “A novel method using edge detection for signal extraction from cDNA microarray image analysis”, *Experimental and molecular medicine*, vol 33, No 2, 83-88, June 2001
 - [27] Yee Hwa Yang, Michael J. Buckley, Sandrine Dudoit, Terence P. Speed, “Comparison of Methods for Image Analysis on cDNA Microarray Data”, *Journal of Computational and Graphical Statistics*, Vol 11, Number 1, p 1-29
-

-
- [28] Wotao Yin, Terrence Chen, Xiang Sean Zhou and Amit Chakraborty, "Background correction for cDNA microarray images using the TV+L model", *Bioinformatics - Gene Expression*, Vol. 21, No. 10, pages 2410-2416, 2005
 - [29] Peter H. Tran, Daniel A. Peiffer, Yongchol Shin, Lauren M. Meek, James P. Brody, Ken W. Y. Cho, "Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals", Oxford University Press, *Nucleic Acids Research*, Vol. 30, No. 12, 2002
 - [30] Bozinov D., Rahnenfuher J., "Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering", *Bioinformatics*, 18, 747-756
 - [31] Imogene, <http://www.biodiscovery.com/index/imagene>
 - [32] Omer Demirkaya, Musa H. Asyali and Mohamed M. Shoukri, "Segmentation of cDNA Microarray Spots Using Markov Random Field Modeling", *Bioinformatics*, Vol 21, No 13, pages 2994-3000, 2005
 - [33] BlackIce Software <http://www.blackice.com/colorspaceHSL.htm>
 - [34] P. Soille, "Morphological Image Analysis: Principles and Applications", Springer, 1999, pp. 164-165
 - [35] Th. Margaritis, K. Marias, D. Kafetzopoulos, "Improved Microarray Spot Segmentation by Combining two Information Channels", 28th IEEE EMBS Annual Conference, April 3, 2006
-