Επεξεργασία εικόνων μικροσυστοιχιών *DNA* με χρήση σύγχρονων μεθόδων ταξινόμησης προτύπων

Αθανασιάδης Ε.[2], Γεωργιάδης Π.[2], Καλατζής Ι.[1], Κωστόπουλος Σπ.[2], Γκλώτσος Δ.[1], Σιδηρόπουλος Κ[1], Νίνος Κ.[1], και Κάβουρας Δ.[1]

[1]*Εργαστήριο Επεξεργασίας και Ανάλυσης Ιατρικής Εικόνας, Εργαστήριο Ιατρικής Φυσικής, Τμήμα Ιατρικής, Πανεπιστήμιο Πατρών.*
[2]*Εργαστήριο Επεξεργασίας Ιατρικού Σήματος και Εικόνας (Ε.Ι.Σ.Ε.), Τμήμα Τεχνολογίας Ιατρικών Οργάνων, Τεχνολογικό Εκπαιδευτικό Ίδρυμα (Τ.Ε.Ι.) Αθήνας.*

Λέξεις κλειδιά: *μικροσυστοιχίες γονιδίων (cDNA microarrays),τμηματοποίηση (segmentation)*

***Περίληψη*: Σκοπός της παρούσας εργασίας είναι η ανάπτυξη νέων μεθόδων αυτόματης διευθυνσιοδότησης και κατάτμησης εικόνων μικροσυστοιχιών συμπληρωματικού DNA (cDNA). Πιο συγκεκριμένα προτάθηκε μια νέα μέθοδο αυτόματης διευθυνσιοδότησης βασισμένη σε μετασχηματισμό κυματιδίου. Επίσης, προτάθηκαν και δύο νέες μέθοδοι τμηματοποίησης οι οποίες βασίζονται στις τεχνικές Fuzzy Gaussian Mixture Models και Wavelet Markov Random Fields. Οι αλγόριθμοι αυτοί εφαρμόστηκαν σε απομιμούμενες εικόνες *cDNA*, καθώς και σε πραγματικές. Χρησιμοποιώντας κατάλληλα μετρικά μπορέσαμε να ποσοτικοποιήσουμε τον βαθμό επιτυχίας των προτεινόμενων αλγορίθμων, σε επίπεδο τμηματοποίησης, καθώς και σε επίπεδο εξαγωγής έντασης κηλίδων. Σύμφωνα με τα αποτελέσματα μας οι προτεινόμενοι αλγόριθμοι πέτυχαν καλύτερα αποτελέσματα, συγκρινόμενη με ήδη υπάρχουσες τεχνικές καθιστώντας τες ως ένα νέο εργαλείο για ακριβέστερες μετρήσεις στο χώρο των μικροσυστοιχιών.**

I. Εισαγωγή

Τα τελευταία χρόνια η τεχνολογία των μικροσυστοιχιών (microarrays) εισήγαγε μια νέα τεχνική για την ποιοτική και ποσοτική μέτρηση της έκφρασης χιλιάδων γονιδίων ταυτοχρόνως σε ένα και μόνο πείραμα [1][2]. Το γεγονός αυτό είναι ιδιαίτερα σημαντικό διότι γνωρίζοντας ποια γονίδια εκφράζονται σε μια δεδομένη χρονική στιγμή στο κύτταρο, μπορούμε να εντοπίσουμε τις αντίστοιχες πρωτεΐνες που δημιουργούνται. Οι πρωτεΐνες είναι σημαντικές για τη σωστή λειτουργία των κυττάρων και παράγονται ως σε δύο κύρια στάδια ως εξής: στο πρώτο στάδιο της *μεταγραφής* (*transcription*) πραγματοποιείται η μεταγραφή του αντίστοιχου γονιδίου από το *DNA* σε αγγελιοφόρο *RNA* (*mRNA*). Στο δεύτερο στάδιο της *μετάφρασης* (*translation*) πραγματοποιείται στα ριβοσωμάτια η μετάφραση των μορίων του *mRNA* στα αντίστοιχα αμινοξέα, τα οποία είναι ο δομικός λίθος των πρωτεϊνών[3].

Το σημαντικό είναι ότι με την τεχνολογία των microarrays λαμβάνουμε ένα στιγμιότυπο της δραστηριότητας του κυττάρου, απομονώνοντας το *mRNA* του την δεδομένη χρονική στιγμή, έτσι με αυτό τον τρόπο μπορούμε να παρατηρήσουμε όχι μόνο ποια γονίδια εκφράζονται αλλά και πόσο.

II. Υπόβαθρο

Μια από τις κυριότερες κατηγορίες μικροσυστοιχιών είναι τα *cDNA* microarrays (συμπληρωματικές μικροσυστοιχίες). Η τεχνική αυτή χρησιμοποιεί μια υάλινη βάση, η οποία αποτελείται από μικροσκοπικές οπές (spots) μέσα στις οποίες τοποθετείται συμπληρωματικό DNA από γονίδια τα οποία επιθυμούμε να ανιχνεύσουμε. Στην συνέχεια λαμβάνονται δύο δείγματα *mRNA τα οποία μετατρέπονται σε cDNA*, το ένα εκ των οποίων είναι το δείγμα αναφοράς (*control*), ενώ το δεύτερο αποτελεί το προς μελέτη δείγμα (*sample*). Στη συνέχεια τα δύο αυτά δείγματα ιχνηθετούνται με την βοήθεια δύο φθοριζουσών χρωστικών ουσιών , την Κυανίνη *Cy*3 (εκπέμπει πράσινη ακτινοβολία) και *Cy*5 (εκπέμπει κόκκινη ακτινοβολία). Στη συνέχεια, τα δύο αυτά δείγματα αναμειγνύονται και υβριδοποιούνται με την υάλινη βάση, με αποτέλεσμα όπου κάποιο γονίδιο από τα δύο δείγματα εκφράζεται, να ενώνεται με το αντίστοιχο cDNA στην υάλινη βάση. Με τη βοήθεια ενός σαρωτή δημιουργούνται δύο εικόνες 16 *bit Tiff*, μία για κάθε χρωστική (πράσινο και κόκκινο).

Για να μπορέσουμε να επεξεργαστούμε τις δύο εικόνες που δημιουργούνται ακολουθούμε

τα εξής βήματα: 1/ την διευθυνσιοδότηση ή κατασκευή πλέγματος (addressing or gridding) όπου γίνεται ο εντοπισμός κάθε κουκίδας (spot) με το αντίστοιχο φόντο (background), 2/ την κατάτμηση (segmentation), όπου γίνεται ταξινόμηση των εικονοστοιχείων εάν ανήκουν στην κουκίδα ή στο φόντο και 3/ διαδικασία εξαγωγής έντασης (intensity extraction), η οποία περιέχει υπολογισμό της έντασης για κάθε κουκίδα και κάθε κανάλι για την κουκίδα και το φόντο.

III. Μεθολογία

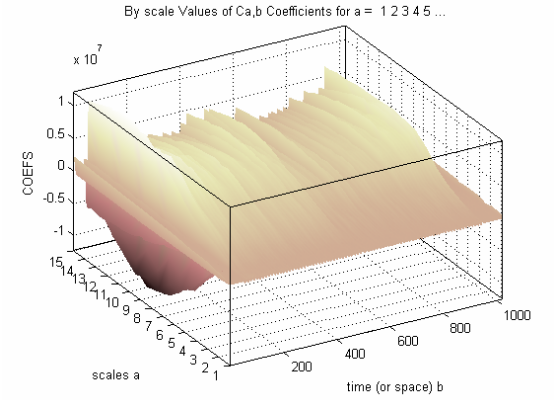
Με σκοπό την εξαγωγή συμπερασμάτων για τους αλγόριθμους τμηματοποίησης που προτείνονται, δημιουργήθηκαν απομιμούμενες εικόνες μικροσυστοιχιών συμφωνά με προτεινόμενη βιβλιογραφία[6][8]. Επίσης, οι αλγόριθμοι χρησιμοποιήθηκαν και σε πέντε πραγματικές εικόνες των δύο καναλιών η κάθε μία[15].

Στην παρούσα εργασία προτείνεται μια νέα μέθοδος διευθυνσιοδότησης βασισμένη σε μετασχηματισμό κυματιδίου, καθώς και δύο νέες μέθοδοι κατάτμησης των κηλίδων βασισμένη σε Fuzzy Gaussian Mixture Models [4][5][6] και Markov Random Fileds [7][8].

Η διαδικασία της διευθυνσιοδότησης πραγματοποιήθηκε ως εξής [9]:

1/ Πραγματοποιήθηκε άθροιση των εικονοστοιχείων της εικόνας κατά τον x και y άξονα και στη συνέχεια δύο νέα σήματα δημιουργήθηκαν, ένα για κάθε άξονα.

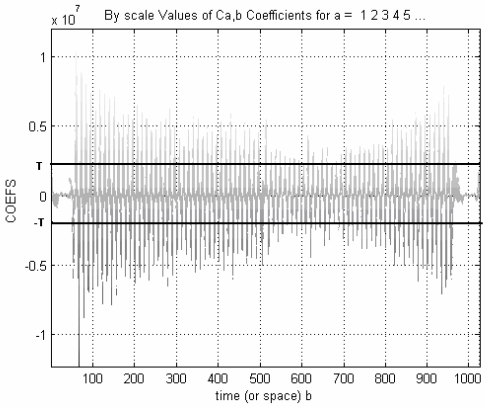
2/ Για κάθε ένα από τα δύο παραπάνω σήματα εφαρμόστηκε ο συνεχής μετασχηματισμός κυματιδίου (*Continuous Wavelet Transform*) [11][12] σε 15 επίπεδα και 15 νέα σήματα (λεπτομέρειες) δημιουργήθηκαν για κάθε ένα από τα δύο αρχικά x και y σήματα (Σχήμα 1).



Σχήμα 1: Συνεχής μετασχηματισμός κυματιδίου σε 15 επίπεδα.

3/ Στην συνέχεια πραγματοποιήθηκε μια άθροιση σημείο προς σημείο των λεπτομερειών και για κάθε x και y σήμα δημιουργώντας δύο νέα σήματα.

4/ Στα δύο αυτά σήματα εφαρμόστηκε ένα φίλτρο σκληρού κατωφλίου (*Hard Threshold*) για καταστολή του θορύβου (Σχήμα 2).



Σχήμα 2: Φιλτράρισμα σκληρού κατωφλίου για την καταστολή του θορύβου

5/ Τέλος υπολογιστήκαν τα τοπικά ελάχιστα των δύο σημάτων τα οποία αντιστοιχούν στα ενδιάμεσα σημεία μεταξύ δύο διαδοχικών κηλίδων.

Μετά τη διαδικασία της διευθυνσιοδότησης, όπου εντοπίζονται αυτόματα οι κηλίδες με τον περιβάλλοντα φόντο τους, ακολουθεί η διαδικασία της κατάτμησης σε κάθε μία κηλίδα χωριστά.

Για την πραγματοποίηση της κατάτμησης χρησιμοποιήθηκαν οι παρακάτω αλγόριθμοι: 1/ Gaussian Mixture Models (*GMM*)[4][5], 2/ Fuzzy Gaussian Mixture Models (*FGMM*)[6], /3 Markov Random Fields (*MRF*)[7], /4 Wavelet Based Markov Random Field (*WMRF*)[8].

Στη συνέχεια, για κάθε τμηματοποιημένη εικόνα υπολογίστηκε η μέση τιμή των εικονοστοιχείων που ανήκουν στην κηλίδα και στο φόντο αντίστοιχα. Η ένταση των πραγματικών εικόνων ομαλοποιήθηκε με την τεχνική *LOWESS* [10] ώστε τα αποτελέσματα των πέντε διαφορετικών εικόνων να μπορούν να συγκριθούν μεταξύ τους. Θα πρέπει να τονιστεί ότι στις πραγματικές εικόνες, η σύγκριση πραγματοποιήθηκε στα αποτελέματα του πράσινου καναλιού διότι αυτά αποτελούν το κανάλι αναφοράς, στο οποίο οι τιμές της έντασης δεν μεταβάλλονται στο χρόνο.

Στις απομιμούμενες εικόνες υπολογίστηκαν οι παρακάτω δείκτες τμηματοποίησης:

1/ Segmentation Matching Factor (*SMF*) [7][8][14]

$$SMF = \frac{B_{segment} \cap B_{actual}}{B_{segment} \cup B_{actual}} \qquad (1)$$

Όπου $B_{segment}$ και B_{actual} αποτελούν την τμηματοποιημένη και απομιμούμενη περιοχή της κηλίδας με το αντίστοιχο φόντο της.

2/ Coefficient of determination (r^2) [7][8][14]

$$r^2 = \frac{\sum_{i=1}^{All\ spots} (I_{segment(i)} - \bar{I}_{actual})^2}{\sum_{i=1}^{All\ spots} (I_{actual(i)} - \bar{I}_{actual})^2} \qquad (2)$$

Όπου $I_{segment}$ και I_{actual} αποτελούν την ένταση που προκύπτει από την τμηματοποίηση και την πραγματική ένταση της απομιμούμενης εικόνας αντίστοιχα.

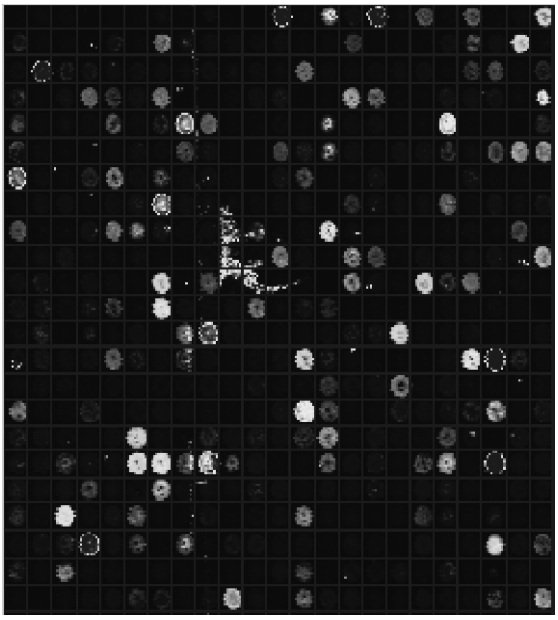
Στις πραγματικές εικόνες υπολογίστηκαν οι παρακάτω δείκτες τμηματοποίησης:

1/ Mean Absolute Error (*MAE*) [7][8][14]

$$MAE_{spot} = \frac{1}{n}\sum_{i=1}^{n}\left|I_i - \bar{I}\right| \qquad (4)$$

IV. Αποτελέσματα – Συμπεράσματα.

Όσον αφορά την διευθυνσιοδότηση, χρησιμοποιώντας οπτική ανασκόπηση για κάθε κηλίδα χωριστά σε όλες τις εικόνες δημιουργήθηκαν δύο κατηγορίες, ανάλογα με το αν οι γραμμές του πλέγματος εφάπτονταν πάνω σε κάποια κηλίδα ή όχι. Έτσι λοιπόν η προτεινόμενη μεθοδολογία ήταν ακριβής σε ποσοστό 98%. Αποτέλεσμα από την κατασκευή πλέγματος δίνεται στο Σχήμα 3.



Σχήμα 3: Κατασκευή πλέγματος

Σύγκριση ανάμεσα στην απόδοση του GMM, FGMM, MRF και WMRF στις απομιμούμενες εικόνες σε διαφορετικά επίπεδα θορύβου πραγματοποιήθηκε και τα αποτελέσματα παρατίθενται στους παρακάτω Πίνακες .

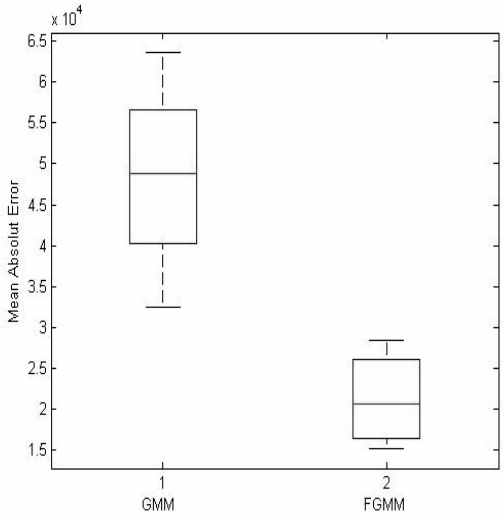
Πίνακας 1: SMF σε απομιμούμενες εικόνες

SNR	*GMM*	*FGMM*	*MRF*	*WMRF*
1db	*63.73*	*66.07*	*92.15*	*92.66*
3db	*68.59*	*73.50*	*92.96*	*93.53*
5db	*74.20*	*78.84*	*94.24*	*94.56*
7db	*79.39*	*84.48*	*96.30*	*96.69*
9db	*82.34*	*89.62*	*97.20*	*97.47*

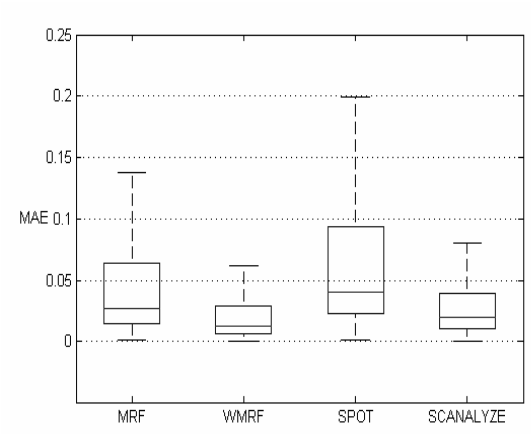
Πίνακας 2: r^2 σε απομιμούμενες εικόνες

SNR	*GMM*	*FGMM*	*MRF*	*WMRF*
1db	*0.89*	*0.95*	*0.91*	*0.96*
3db	*0.95*	*0.96*	*0.96*	*0.96*
5db	*0.91*	*0.98*	*0.97*	*0.98*
7db	*0.91*	*0.98*	*0.97*	*0.98*
9db	*0.91*	*0.99*	*0.99*	*0.99*

Στη συνέχεια ακολουθούν αποτελέσματα τα οποία προέκυψαν από τη χρήση πραγματικών εικόνων μικροσυστοιχιών. Πιο συγκεκριμένα στη περίπτωση των GMM και FGMM τα αποτελέσματα δίνονται στο Σχήμα 4, ενώ στη περίπτωση MRF και WMRF στο Σχήμα 5 αντίστοιχα. Θα πρέπει να τονιστεί ότι όσο πιο μικτή είναι η τιμή του MAE, τόσο πιο καλός είναι ο αλγόριθμος τμηματοποίησης.



Σχήμα 4. Αποτελέσματα MAE σε πραγματικές εικόνες με χρήση GMM και FGMM.



Σχήμα 5. Αποτελέσματα MAE σε πραγματικές εικόνες με χρήση MRF και WMRF καθώς και με τα ευρέως διαδεδομένα προγράμματα SPOT[16] και SCANALYZE[17].

Όπως φαίνεται και αποτελέσματα, τόσο στις πραγματικές όσο και στις απομιμούμενες εικόνες, το WMRF κατάφερε να επιτύχει τα καλύτερα αποτελέσματα, συγκρινόμενος με τις υπόλοιπες τεχνικές τμηματοποίησης.

Ευχαριστίες

Στο Σημείο αυτό θα θέλαμε να ευχαριστήσουμε το Ίδρυμα Κρατικών Υποτροφιών (Ι.Κ.Υ.) για τη χρηματοδότηση της παραπάνω εργασίας.

Αναφορές

[1] Y.H. Yang, M. J. Buckley, S. Duboit and T.P.Speed , "Comparison of methods for Image Analysis on cDNA Microarray Data", *Journal of Computational and Graphical Statistics*, vol. 11, pp 108-136, 2002.

[2] M. Schena, D. Shalon, R.W. Davis and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarrray", *Science* 270, pp. 467-470, 1995.

[3] Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale", SCIENCE, Vol 278, pg 680, 1997.

[4] K. Blekas, N.P. Galatsanos and I. Georgiou, "An unsupervised Artifact Correction Approach for the Analysis of DNA Microarray Images", *Proc. IEEE International Conf. on Image Processing (ICIP)*, vol 2, pp 165-168, 2003.

[5] K. Blekas, N.Galatsanos, A. Likas, and I.E. Lagaris, "Mixture Model Analysis of DNA Microarray Images", *IEEE Transactions on Medical Imaging*, vol 24. pp. 901-907, 2005.

[6] Emmanouil I. Athanasiadis, Dionisis A. Cavouras, Panagiota P. Spyridonos, Dimitris Th. Glotsos, Ioannis K. Kalatzis, and George C. Nikiforidis, "Segmentation of complementary DNA microarray images using the Fuzzy Gaussian Mixture Model technique", *IEEE TITB*, in press

[7] S.Z. Li, *Markov Random Field Modeling in Computer vision*. New York: Springer- Verlag, 2001.

[8] Emmanouil I. Athanasiadis, Dionisis A. Cavouras, Dimitris Th. Glotsos, Pantelis V. Georgiadis, Ioannis K. Kalatzis, and George C. Nikiforidis, "Segmentation of Complementary DNA Microarray Images by Wavelet-based Markov Random Field Model", *IEEE TITB*, submitted.

[9] Emmanouil Athanasiadis, Dionisis Cavouras, Panagiota Spyridonos, Ioannis Kalatzis and George Nikiforidis, "An Automatic Microarray Image Gridding Technique Based on Continuous Wavelet Transform", *Lecture Notes in Computer Science*, vol. 4673, pp 854-870, 2007.

[10] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terence P. Speed, "Normalization for cDNA microarray data : a robust composite method addressing single and multiple slide systematic variation", *Nucleic Acid Research*, vol. 30, No. 4 e15, 2002

[11] Burrus CS, Gopinath RA, Guo H, Introduction to Wavelets and Wavelet Transfroms. Prentice Hall, Englewood Cliffs, NJ. 199

[12] James E. Fowler, "The Redundant Discrete Wavelet Transform and Additive Noise", IEEE Signal Processing Letters, Vol. 12, NO. 9, 2005.

[13] The MathWorks, Inc. Software, MATLAB®.

[14] Lehmussola,A. et al., "Evaluating the performance of microarray segmentation algorithms", *Bioinformatics*, vol. 22, 2910–2917, 2006.

[15] Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale", SCIENCE, Vol 278, pg 680, 1997.

[16] M.B. Eisen, ScanAlyze (1999). Available: http://rana.lbl.gov/EisenSoftware.htm

[17] M.J. Buckley (2000), The Spot user's guide. CSIRO Mathematical and Information Science. Available: http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm