
**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ**
**"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"**


**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Σύστημα επεξεργασίας, ανάλυσης και ταξινόμησης εικόνων δισδιάστατης ηλεκτροφόρησης με τεχνικές αναγνώρισης προτύπων


**Αγγελική Δ. Θεοδόση**


**Επιβλέποντες:** **Κάβουρας Διονύσιος,** Καθηγητής

**ΑΘΗΝΑ**

**Νοέμβριος 2012**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# Σύστημα επεξεργασίας, ανάλυσης και ταξινόμησης εικόνων δισδιάστατης ηλεκτροφόρησης με τεχνικές αναγνώρισης προτύπων

**Αγγελική Δ. Θεοδόση**

**Α.Μ.:** ΠΙΒ 035

**ΕΠΙΒΛΕΠΩΝ:** **Κάβουρας Διονύσιος,** Καθηγητής

**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:** **Σαγγριώτης Εμμανουήλ** Αναπληρωτής Καθηγητής
**Βεντουράς Ερρίκος** Καθηγητής

Νοέμβριος 2012

# ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη ενός συστήματος ανάλυσης και ταξινόμησης εικόνων δισδιάστατης ηλεκτροφόρησης από ασθενείς με μυελογενή και λεμφογενή λευχαιμία, με τεχνικές ανάλυσης εικόνας και αναγνώρισης προτύπων.

Ένα σημαντικό πλεονέκτημα των εικόνων, δισδιάστατης ηλεκτροφόρησης (2-d Gel Electrophoresis), είναι η πληθώρα πρωτεϊνών που απεικονίζεται σε καθεμιά απ' αυτές, κάτι που αποτελεί ταυτόχρονα και αντικείμενο ιδιαίτερου ερευνητικού ενδιαφέροντος, καθώς θα πρέπει να εντοπιστούν από το σύνολο των κηλίδων, εκείνες οι οποίες θα μπορούσαν να αποτελέσουν πιθανούς βιοδείκτες είτε μεμονωμένα, είτε σε συνδυασμό.

Για την πρακτική εφαρμογή χρησιμοποιήθηκαν εικόνες από μια διαδικτυακή βάση δεδομένων με εικόνες από σκαναρισμένα gel δισδιάστατης ηλεκτροφόρησης, η LECB 2-D PAGE. Η συγκεκριμένη βάση είχε το πλεονέκτημα ότι οι περιοχές ενδιαφέροντος, δηλαδή, οι κηλίδες, ήταν και αυτές διαθέσιμες, καθώς δίδονταν το κέντρο βάρους κάθε μίας από αυτές.

Η εξαγωγή των περιοχών ενδιαφέροντος, έγινε με δύο τρόπους: χειροκίνητα, με επιλογή από τον χρήστη των κηλίδων που υποδεικνύονταν στη βάση δεδομένων και αυτόματα, ορίζοντας μια περιοχή ενδιαφέροντος για κάθε κηλίδα, μεγέθους 13x13 εικονοστοιχείων με κέντρο το σημείο που έδινε η βάση δεδομένων. Από τις περιοχές ενδιαφέροντος (κηλίδες/spot) υπολογίστηκαν χαρακτηριστικά υφής $1^{ης}$ και $2^{ης}$ τάξεως, δημιουργώντας για κάθε κηλίδα ένα διάνυσμα χαρακτηριστικών.

Στη συνέχεια για κάθε σύνολο ξεχωριστά, έγινε μείωση των χαρακτηριστικών πρώτα με βάση τα αποτελέσματα που προέκυψαν από την εφαρμογή ενός Man Whitney test στο οποίο εντοπίστηκαν τα τρία χαρακτηριστικά για κάθε κηλίδα (από σύνολο 22 κηλίδων) που είχαν μεγαλύτερη στατιστικά σημαντική διαφορά ($p<0.001$) μεταξύ των δυο κατηγοριών (μυελογενή και λεμφογενή λευχαιμία). Στη συνέχεια πραγματοποιήθηκε κατάταξη των χαρακτηριστικών με βάση τη συσχέτιση και το Wilcoxon τεστ. Με αυτόν τον τρόπο για κάθε εικόνα δημιουργήθηκε ένα διάνυσμα μεγέθους 16 χαρακτηριστικών. Με τα διανύσματα αυτά εκπαιδεύτηκε ένα Πιθανοκρατικό Νευρωνικό Δίκτυο (Probabilistic Neural Network PNN), και ο kNN ταξινομητής για τον οποίο δοκιμάστηκαν διάφορες τιμές του k. Οι ταξινομητές εκπαιδεύτηκαν έτσι ώστε να διαχωρίζουν με τη μεγαλύτερη δυνατή ακρίβεια τις δυο κατηγορίες, χρησιμοποιώντας το μικρότερο πλήθος χαρακτηριστικών. Η επιλογή των χαρακτηριστικών έγινε με τη μέθοδο εξαντλητικής αναζήτησης. Ο αξιολόγηση των ταξινομητών έγινε χρησιμοποιώντας τη μέθοδο αποκλεισμού ενός (Leave One Out-LOO). Επίσης, το σύστημα αξιολογήθηκε περαιτέρω, για κάθε σύνολο χωριστά, με τη μέθοδο εξωτερικής διασταυρούμενης επικύρωσης (External Cross Validation- ECV), με αποκλεισμό από το στάδιο της εκπαίδευσης του 30% του συνόλου των δειγμάτων.

Για τον PNN ταξινομητή και την LOO μέθοδο αξιολόγησης, η συνολική ακρίβεια (accuracy) του συστήματος ταξινόμησης, φάνηκε να φτάνει το 98.1% ενώ με τη μέθοδο ECV η απόδοση του ταξιομητή κυμαίνεται στο 89.4±4.5%. Επιπλέον, για το σύνολο αυτό έγινε και ταξινόμηση με τον kNN ταξινομητή, για τα 16 αυτά χαρακτηριστικά, και με τις δύο μεθόδους LOO, αλλά και ECV. Η ολική ακρίβεια ταξινόμησης στην περίπτωση του kNN βρέθηκε 96.3% για k=3 και k=5, με τη LOO μέθοδο, ενώ με την ECV, υπολογίστηκε 89.5±4.9% για k=3.

Το πείραμα επαναλήφθηκε και για το δεύτερο σύνολο περιοχών ενδιαφέροντος όπου ο PNN κατάφερε ακρίβεια 96.3%, με την LOO μέθοδο αξιολόγησης και 88.6±3.6% με μέθοδο αξιολόγησης την ECV. Τα αντίστοιχα ποσοστά του kNN, ταξινομητή είναι 97.2%

με την LOO μέθοδο, τόσο για k=3, k=5 και k=7. Με την μέθοδο ECV, ο kNN για k=3 είχε απόδοση 91.1±4.6%.

Ιδιαίτερα σημαντικά χαρακτηριστικά με ικανότητα διάκρισης μεταξύ των δύο κατηγοριών, φάνηκε να δίνουν οι κηλίδες {12, 22, 8} που αποτελούσαν τον κορμό σε όλα τα σύνολα που προέκυψαν για όλους τους ταξινομητές και των δύο συνόλων περιοχών ενδιαφέροντος. Επιπλέον και με σειρά συχνότητας εμφάνισης, ιδιαίτερη σημασία φάνηκε να έχουν οι κηλίδες {13, 17 και 5}, χωρίς αυτό να σημαίνει ότι στα σύνολα που προτάθηκαν από τους ταξινομητές, δεν υπήρχαν και άλλες κηλίδες, οι οποίες είχαν συχνότητα εμφάνισης από δύο και λιγότερες φορές στο πλήθος των δέκα επαναλήψεων.

Όσον αφορά τα χαρακτηριστικά υφής τα οποία ήταν πιο συχνά εμφανιζόμενα στα σύνολα, αυτά ήταν ο μέσος και το εύρος της ενέργειας (mean Energy, range Energy) και της αντίθεσης, (mean Contrast, range Contrast) καθώς επίσης και ο μέσος της ομοιογένειας (mean Homogeneity).

**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ**: Επεξεργασία Εικόνας

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ**: Δισδιάστατη ηλεκτροφόρηση Μυελογενείς Λευχαιμία, Λεμφογενής Λευχαιμία, Αναγνώριση προτύπων, Βιοδείκτες, Πιθανοκρατικό Νευρωνικό Δίκτυο, kNN.

# ABSTRACT

The aim of the present study was the design and development of a computer based system for the classification of two dimensional gel electrophoresis images, from patients with either Myeloid or Lymphoid Leukemia, using image analysis methods and pattern recognition techniques.

A significant advantage of two-dimensional polyacrylamide gel electrophoresis of proteins images (2D-gel electrophoresis), is the plethora of proteins presented on a single gel, where as at the same time it may be difficulty to detect protein-spots of high significance (possible biomarkers). This paper presents the design and development of an image processing and analysis system for the detection of 2D gels spots that can effectively distinguish between patients with Myeloid Leukemia (ML) and Lymphoid Leukemia (LL). The publicly available LECB 2-D PAGE gel images database was employed. These images were produced by scanned 2D electrophoresis gels and the sites of the particular spots of interest were known.

Regions of Interest (ROIs) were extracted both manually and automatically by defining an area of 13x13 pixels, the center of which was stored in the database as centroid of the spot. Thus, two sets of ROIs were obtained, one manually and one automatically. For each spot a number of textural features were calculated based on first and second order statistical measures.

Feature reduction took place by applying the Man Whitney test for the detection of statistical significant differences ($p<0.001$). Thus, for each one of the 22 spots, three features were retained having the higher statistical significant differences between the two categories (ML and LL). Then, the features were ranked based on the combination criterion of correlation and Wilcoxon statistical test.

Consequently, based on the retained feature vectors, a Probabilistic Neural Network (PNN) classifier and a kNN classifier were trained so as to discriminate the two classes (ML-LL). Feature selection was based on the exhaustive search method. The system classification accuracy was evaluated by the Leave One Out (LOO) method, and the External Cross Validation method (ECV).

The overall accuracy of the system, when the PNN classifier and the LOO method were employed was found to be 98.1%. Regarding the ECV method, the mean accuracy was 89.4±4.5%. Concerning the kNN classifier and the LOO method the overall accuracy was 96.3% for with k=3, while employing the ECV evaluation method an average of 89.5±4.9% was attained for k=3.

Regarding the automatically obtained ROIS, results were 96.3% for the PNN classifier and the LOO method, and when the PNN was evaluated by the ECV method the overall accuracy was reduced to 88.6±3.6%. Employing the LOO method for the evaluation of kNN classifier with k=3, the overall accuracy was 97.2%, whereas employing the ECV method the mean overall classification accuracy was 91.1±4.6%.

Features extracted from spots {8, 12, 22}, as defined in the original online database, seem to have significant differentiating capabilities between the ML and LL 2D-gel images. Those features were the average and range of the Energy, the average and range of the Contrast, and the average of Homogeneity.



**SUBJECT AREA**: Image Processing

**KEYWORDS**: 2D Gel Electrophoresis, Myeloid Leukemia, Lymphoid Leukemia, Pattern Recognition, Biomarkers, PNN, kNN