Πληροφορικό Σύστημα Υποβοήθησης Ιατρικής Διάγνωσης Βάσει Δεδομένων Μικροσυστοιχιών DNA

Ηλίας Ν. Φλαούνας, Δημήτρης Ε. Μαρούλης,
Δημήτρης Ε. Ιακωβίδης, Σταύρος Α. Καρκάνης

Περίληψη

Σύγχρονες μεθοδολογίες αυτοματοποιημένης διάγνωσης περιλαμβάνουν την αξιοποίηση βιολογικών δεδομένων που ποσοτικοποιούν την έκφραση των γονιδίων (gene expression) των ασθενών. Μια νέα μέθοδος, που επιτρέπει την παράλληλη μέτρηση επιπέδων *έκφρασης* χιλιάδων γονιδίων είναι οι *μικροσυστοιχίες DNA* (DNA microarrays). Υλοποιήσαμε ένα σύστημα αυτόματης διάγνωσης μορφών καρκίνου που λαμβάνει ως είσοδο τις μετρήσεις γονιδιακής έκφρασης από μικροσυστοιχίες DNA ενώ είναι εκπαιδευμένο από έναν αντίστοιχο πίνακα γονιδιακής έκφρασης. Το σύστημα αυτό αποτελείται από δυο βασικά τμήματα:

Α) *Επιλογής των γονιδίων***.** Επιλέγει τα γονίδια που εκφράζονται διαφορετικά σε παθολογικές καταστάσεις. Η επιλογή γίνεται με στατιστικές μεθόδους και με κατάλληλους υπολογιστικούς αλγορίθμους.

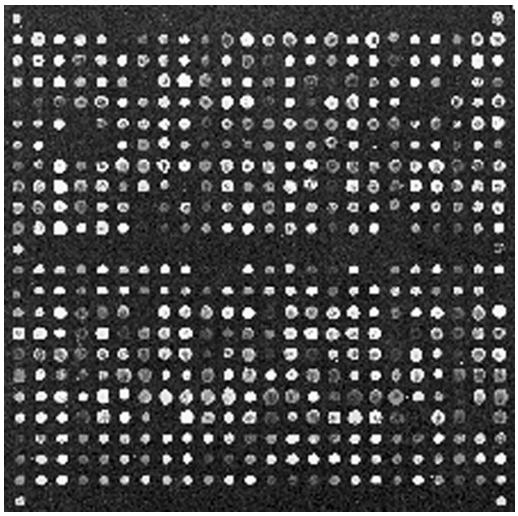
Β) *Ταξινόμησης*. Χρησιμοποιεί Μηχανές Ανυσμάτων Υποστήριξης (Support Vector Machines) για να αποφασίσει αν το δείγμα που έλαβε ως είσοδο είναι φυσιολογικό ή έχει υποστεί καρκινική μετάλλαξη.

Έγιναν δοκιμές με δυο σετ δεδομένων μικροσυστοιχιών DNA. Το πρώτο αποτελείται από 62 δείγματα (40 καρκίνου του εντέρου και 22 φυσιολογικά) όπου έχουν μετρηθεί οι εκφράσεις 2000 γονιδίων. Το σύστημα διαχωρίζει τα φυσιολογικά από τα μη φυσιολογικά δείγματα με ποσοστό επιτυχίας 91,9%. Το δεύτερο σετ δεδομένων αποτελείται από 72 δείγματα δυο διαφορετικών τύπων λευχαιμίας (47 ALL, 25 AML) και μετρήσεις της έκφρασης 7129 γονιδίων. Ο στόχος εδώ ήταν να γίνει γονιδιακός διαχωρισμός των δυο τύπων της λευχαιμίας. Το σύστημα μας εντοπίζει 4 γονίδια που είναι ικανά να διαχωρίσουν τους δύο τύπους με ποσοστό επιτυχίας 100%.

Ι. Εισαγωγή

Τα τελευταία χρόνια έχουν προταθεί διάφορα πληροφορικά συστήματα υποστήριξης της ιατρικής διάγνωσης, με τη βοήθεια των οποίων είναι δυνατή η ενίσχυση των φυσικών δυνατοτήτων των ειδικών ιατρών κατά τη διαδικασία της διάγνωσης [1]. Σύγχρονες μεθοδολογίες αυτοματοποιημένης διάγνωσης περιλαμβάνουν την αξιοποίηση βιολογικών δεδομένων που ποσοτικοποιούν την *έκφραση των γονιδίων* (*gene expression*) των ασθενών. Με τον όρο γονιδιακή έκφραση εννοούνται οι δύο διαδικασίες, η *μεταγραφή* και η *μετάφραση*, με τις οποίες «αποκωδικοποιείται» η γενετική πληροφορία που είναι αποθηκευμένη στο DNA ενός οργανισμού. Η μεταγραφή καθορίζει το ποια γονίδια θα εκφραστούν, σε ποιους ιστούς και σε πιο στάδιο της ανάπτυξης ενώ η μετάφραση χρησιμοποιεί την πληροφορία που είναι αποθηκευμένη στα γονίδια για να κατασκευάσει πολυπεπτίδια με τα οποία ελέγχεται η δομή και λειτουργία των κύτταρων και κατ' επέκταση και των οργανισμών. Μεταβολές στην γονιδιακή έκφραση έχουν συσχετιστεί με διάφορες ασθένειες όπως είναι ο καρκίνος, η σκλήρυνση κατά πλάκας κ.α.

Μέχρι πρόσφατα οι βιολόγοι είχαν στη διάθεση τους τεχνικές που τους επέτρεπαν να μετρούν την έκφραση περιορισμένου αριθμού γονιδίων και για κάθε γονίδιο έπρεπε να πραγματοποιηθεί διαφορετικό πείραμα. Η τεχνολογία των *μικροσυστοιχιών DNA* (*DNA-Microarrays*) επέτρεψε για πρώτη φορά την παράλληλη μέτρηση της γονιδιακής έκφρασης εκατοντάδων έως και χιλιάδων γονιδίων με την εκτέλεση ενός και μόνο πειράματος.



Εικ. 1 Παράδειγμα εικόνας μικροσυστοιχίας DNA. Κάθε κουκίδα αντιστοιχεί σε ένα γονίδιο και η έντασή της αντιστοιχεί στο επίπεδο της έκφρασης του.

Μια μικροσυστοιχία DNA είναι ένα πλακίδιο κατασκευασμένο από ειδικό γυαλί

πάνω στο οποίο παρατάσσονται *ιχνηθέτες* σε συγκεκριμένες θέσεις, το πλήθος των οποίων μπορεί να κυμανθεί από μερικές εκατοντάδες έως πολλές χιλιάδες (~30.000). Κάθε ιχνηθέτης αποτελείται από το συμπληρωματικό DNA (cDNA) του mRNA του γονιδίου που θέλουμε να μετρηθεί. Το προς εξέταση δείγμα DNA «χρωματίζεται» με κατάλληλη χρωστική και μετά από ειδική επεξεργασία τοποθετείται πάνω στο πλακίδιο. Τα γονίδιά του αντιδρούν χημικά με τους αντίστοιχους ιχνηθέτες (*υβριδισμός*) και ελευθερώνουν στην αντίστοιχη θέση του ιχνηθέτη την χρωστική ουσία. Η ποσότητα της χρωστικής που ελευθερώνεται στην θέση ενός ιχνηθέτη είναι ανάλογη της έκφρασης του αντίστοιχου γονιδίου. Στην συνέχεια ένας οπτικός σαρωτής σαρώνει το πλακίδιο και στην έξοδό του παράγεται μια ψηφιακή εικόνα η οποία αποτελείται από ένα πλήθος κουκκίδων (Εικ. 1). Κάθε κουκκίδα αντιστοιχεί σε ένα διαφορετικό γονίδιο και η έντασή της αντιστοιχεί στο επίπεδο έκφρασης του. Η εικόνα αναλύεται με κατάλληλο λογισμικό με αποτέλεσμα την παραγωγή ενός διανύσματος, του οποίου οι μεταβλητές είναι μετρήσεις γονιδιακής έκφρασης των γονιδίων που είχαν επιλεγεί προς μέτρηση μέσω των κατάλληλων ιχνηθετών [2].

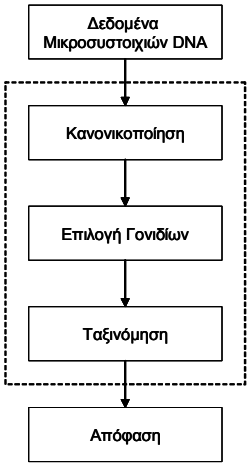
Πραγματοποιώντας μια σειρά από τέτοια πειράματα με διαφορετικά δείγματα DNA κατασκευάζεται ένας *πίνακας γονιδιακής έκφρασης*. Οι στήλες του πίνακα αντιστοιχούν σε διαφορετικά δείγματα ενώ οι γραμμές σε διαφορετικά γονίδια. Η μελέτη του πίνακα γονιδιακής έκφρασης μπορεί να αποκαλύψει πια γονίδια επηρεάζονται από μια ασθένεια ή είναι υπεύθυνα για την εκδήλωση της.

Οι πρώτες μελέτες δεδομένων μικροσυστοιχιών DNA είχαν ως στόχο την εύρεση προτύπων στην οργάνωση των γονιδίων καθώς και τον προσδιορισμό των γονιδίων των οποίων διαφοροποιείται η έκφραση σε παθολογικές ή άλλες συνθήκες. Πιο πρόσφατες μελέτες στοχεύουν στη ανάπτυξη νέων τεχνικών και συστημάτων που είναι σε θέση να ταξινομούν και να διαχωρίσουν δείγματα γονιδιακής έκφρασης με στόχο την ιατρική διάγνωση [3]. Η διάγνωση ασθενειών γίνεται εφαρμόζοντας καθοδηγούμενες (supervised) μεθόδους αναγνώρισης προτύπων [4][5]. Στη βιβλιογραφία έχουν προταθεί μεθοδολογίες αυτόματης αναγνώρισης καρκίνου του εντέρου [6], ωοθηκών [7], του μαστού [8] του πνεύμονα [9], λευχαιμίας [10] και λεμφώματος [11] οι οποίες βασίζονται σε δεδομένα γονιδιακών εκφράσεων. Ένα πρόβλημα που αντιμετωπίζεται στις μελέτες αυτές, είναι ο μικρός αριθμός δειγμάτων εκπαίδευσης σε συνδυασμό με τον υπερβολικά μεγάλο αριθμό γονιδίων (curse of dimensionality). Για την αποφυγή τέτοιων φαινομένων εφαρμόζονται επιπρόσθετα μεθοδολογίες επιλογής γονιδίων των οποίων η έκφραση διαφοροποιείται παρουσία συγκεκριμένης παθολογίας [12].

Υλοποιήσαμε ένα πληροφορικό σύστημα αυτόματης διάγνωσης μορφών καρκίνου που λαμβάνει ως είσοδο ένα διάνυσμα γονιδιακής έκφρασης και μπορεί να αποφασίσει αν το δείγμα είναι υγιές ή καρκινικό. Το σύστημα πρέπει προηγουμένως να έχει εκπαιδευτεί από έναν κατάλληλο πίνακα γονιδιακής έκφρασης. Έγιναν δοκιμές της αξιοπιστίας του συστήματος με δυο σετ δεδομένων μικροσυστοιχιών DNA. Το πρώτο αποτελείται από δείγματα, φυσιολογικά και καρκίνου του εντέρου, ενώ το δεύτερο αποτελείται από δείγματα δυο διαφορετικών τύπων λευχαιμίας Στο πρώτο σετ ο στόχος ήταν η διερεύνηση της δυνατότητας διαχωρισμού φυσιολογικών από καρκινικά δείγματα ενώ στο δεύτερο της δυνατότητας διαχωρισμού δειγμάτων μεταξύ δυο υποκατηγοριών της ίδιας ασθένειας.

ΙΙ. Περιγραφή του Συστήματος

Η εικόνα 2 παρουσιάζει την δομή του συστήματος με το οποίο πραγματοποιείται η υποβοήθηση της ιατρικής διάγνωσης βάσει δεδομένων μικροσυστοιχιών DNA.



Εικ. 2 Δομή του προτεινόμενου συστήματος.

Το σύστημα λαμβάνει ως είσοδο ένα διάνυσμα γονιδιακής έκφρασης και στην έξοδο δίνει την κλάση στην οποία αυτό ανήκει. Τα δεδομένα αρχικά κανονικοποιούνται. Η κανονικοποίηση δίνει την δυνατότητα σύγκρισης διανυσμάτων που προέρχονται από διαφορετικά πειράματα. Συγκεκριμένα πραγματοποιούμε έναν γραμμικό μετασχηματισμό με τον οποίον οι μεταβλητές του διανύσματος αποκτούν μηδενική μέση τιμή και μοναδιαία διασπορά. Το κυρίως σύστημα αποτελείται από δυο βασικά τμήματα

α) *Επιλογής των γονιδίων* και β) *Ταξινόμησης*. Η επιλογή των κατάλληλων γονιδίων γίνεται στη φάση της εκπαίδευσης του συστήματος. Στην ίδια φάση γίνεται και η εκπαίδευση του ταξινομητή.

Α) Επιλογή Γονιδίων

Η επιλογή των κατάλληλων γονιδίων για ένα πρόβλημα ταξινόμησης δυο κλάσεων ω_α και ω_β μπορεί να γίνει με την χρήση στατιστικών κριτηρίων. Ως κατάλληλα χαρακτηρίζονται τα γονίδια που οδηγούν σε μεγάλη δια-κλασική απόσταση και μικρή ενδοκλασική διασπορά. Πολλές διαφορετικές προσεγγίσεις έχουν γίνει για το θέμα αυτό στη βιβλιογραφία. Δοκιμάσαμε τρεις διαφορετικές μεθοδολογίες προκειμένου να συγκρίνουμε την αποτελεσματικότητα και την αξιοπιστία τους.
Για κάθε γονίδιο j με έκφραση g_{ij} στο δείγμα i υπολογίζεται η απόλυτη τιμή της ποσότητας $Z(j)$:
α) «Δύναμη Πρόγνωσης» (*Prediction Strength*) του Golub [10]

$$Z(j) = \frac{\mu_j^\alpha - \mu_j^\beta}{\sigma_j^\alpha + \sigma_j^\beta} \tag{1}$$

β) *t*-test του Welch [13]

$$Z(j) = \frac{\mu_j^\alpha - \mu_j^\beta}{\sqrt{\frac{\sigma_j^{\alpha\,2}}{N_\alpha} + \frac{\sigma_j^{\beta\,2}}{N_\beta}}} \tag{2}$$

γ) Κριτήριο του Sun [14]

$$Z(j) = \frac{N_\alpha(\mu_j^\alpha - \mu_j^{a\beta})^2 + N_\beta(\mu_j^\beta - \mu_j^{a\beta})^2}{\sum_{i\in\omega_\alpha}(g_{ij} - \mu_j^\alpha) + \sum_{i\in\omega_\beta}(g_{ij} - \mu_j^\beta)} \tag{3}$$

όπου (μ^α_j, σ^α_j) and (μ^β_j, σ^β_j) αντιστοιχούν στη μέση τιμή και τυπική απόκλιση των εκφράσεων του γονιδίου j των δειγμάτων εκπαίδευσης που ανήκουν στις κλάσεις ω_α και ω_β αντίστοιχα ενώ N_α και N_β είναι το πλήθος των δειγμάτων αυτών. Τα γονίδια με το μεγαλύτερο Z είναι αυτά που επιλέγονται για την ταξινόμηση. Το πλήθος τους v προσδιορίζεται πειραματικά και εξαρτάται από το σετ δεδομένων που εξετάζεται, το κριτήριο Z και τον ταξινομητή που χρησιμοποιείται. Ερευνήσαμε την απόδοση του συστήματος για τις τιμές του v στο διάστημα 1 έως 10.

Β) Μηχανές Ανυσμάτων Υποστήριξης

Οι μηχανές ανυσμάτων υποστήριξης (Support Vector Machines – SVM) είναι δυαδικοί ταξινομητές που επιδιώκουν να βρουν το υπερεπίπεδο που διαχωρίζει τα δείγματα δυο κλάσεων ω_α και ω_β, ενώ παράλληλα μεγιστοποιεί το διαχωριστικό περιθώριο μεταξύ τους [15]. Τα διανύσματα εκπαίδευσης που ορίζουν το υπερεπίπεδο ονομάζονται *ανύσματα υποστήριξης*. Στην περίπτωση που τα δείγματα δεν διαχωρίζονται γραμμικά χρησιμοποιείται ένας *πυρήνας* που μεταφέρει τα διανύσματα από τον χώρο των χαρακτηριστικών σε έναν άλλο χώρο μεγαλύτερης διάστασης όπου εκεί είναι γραμμικά διαχωρίσιμα. Οι πιο συχνά χρησιμοποιούμενοι πυρήνες είναι ο γραμμικός, ο πολυωνυμικός, ο σιγμοειδής και ο Radial Basis (RBF). Εμείς χρησιμοποιήσαμε τον RBF διότι δίνει συνήθως καλύτερο περιθώριο για τα περισσότερα σετ δεδομένων υψηλής διάστασης τα οποία μπορούν να προσεγγιστούν με κανονικές κατανομές σαν αυτές που χρησιμοποιούνται από RBF δίκτυα [16].

III. Αποτελέσματα

Το σύστημα που αναπτύξαμε δοκιμάστηκε ως προς την αξιοπιστία του με δυο σετ δεδομένων μικροσυστοιχιών DNA. Το πρώτο αποτελείται από 62 δείγματα, 22 φυσιολογικά και 40 καρκίνου του εντέρου, όπου έχουν μετρηθεί οι εκφράσεις 2000 γονιδίων [6]. Το δεύτερο σετ δεδομένων αποτελείται από 72 δείγματα δυο διαφορετικών τύπων λευχαιμίας (47 Acute Lymphoblastic Leukemia - ALL, 25 Acute Myeloid Leukemia - AML) και μετρήσεις της έκφρασης 7129 γονιδίων [10].

Ο πίνακας Ι παρουσιάζει τα ποσοστά επιτυχίας σωστής διάγνωσης με τους τρεις τρόπους επιλογής γονιδίων για το σετ δεδομένων του καρκίνου του εντέρου και ο πίνακας ΙΙ για την λευχαιμία.

Το σύστημα διαχωρίζει τα φυσιολογικά από τα μη φυσιολογικά δείγματα στην περίπτωση του καρκίνου του εντέρου με βέλτιστο ποσοστό επιτυχίας 91,9%. Το ποσοστό αυτό επιτυγχάνεται χρησιμοποιώντας είτε την μέθοδο του Golub είτε την μέθοδο του Sun με τον ίδιο αριθμό γονιδίων δηλαδή 4. Με την μέθοδο του Welch το βέλτιστο ποσοστό που επιτυγχάνεται είναι 90,3% χρησιμοποιώντας ένα λιγότερο γονίδιο.

Οι δύο τύποι λευχαιμίας διαχωρίζονται με ποσοστό επιτυχίας 100% χρησιμοποιώντας μόλις 4 γονίδια και την μέθοδο του Welch. Οι άλλες δυο μέθοδοι παρουσιάζουν επίσης πολύ υψηλά ποσοστά επιτυχίας.

Τα πειράματα που έγιναν επιβεβαιώνουν την αντίληψη ότι δεν υπάρχει ένας βέλτιστος τρόπος επιλογής γονιδίων που να δίνει το ίδιο καλά αποτελέσματα σε κάθε σετ δεδομένων [3].

Πίνακας I
Αποτελέσματα ταξινόμησης (%) στο σετ δεδομένων του καρκίνου του εντέρου

Αριθμός Γονιδίων	Golub	Welch	Sun
1	82,2	82,2	82,2
2	87,1	87,1	87,1
3	87,1	**90,3**	87,1
4	**91,9**	90,3	**91,9**
5	91,9	90,3	91,9
6	91,9	90,3	90,3
7	91,9	88,7	91,9
8	90,3	88,7	91,9
9	90,3	88,7	90,3
10	90,3	88,7	88,7

Πίνακας II
Αποτελέσματα ταξινόμησης (%) στο σετ δεδομένων της Λευχαιμίας

Αριθμός Γονιδίων	Golub	Welch	Sun
1	94,4	91,6	94,4
2	95,8	95,8	**95,8**
3	95,8	94,4	95,8
4	97,2	**100**	95,8
5	97,2	100	94,4
6	97,2	97,2	94,4
7	97,2	97,2	94,4
8	97,2	98,6	93,0
9	**98,6**	98,6	94,4
10	98,6	98,6	94,4

IV. Συμπεράσματα

Σε αυτή την εργασία παρουσιάσαμε ένα καινούριο σύστημα για τον διαχωρισμό κλάσεων σε δεδομένα μικροσυστοιχιών DNA. Βασίζεται σε ένα σχήμα που αποτελείται από μια μονάδα επιλογής των κατάλληλων γονιδίων που μπορούν να οδηγήσουν στο διαχωρισμό των κλάσεων και σε έναν ταξινομητή Μηχανής Ανυσμάτων Υποστήριξης. Το σύστημα δοκιμάστηκε σε δυο σετ δεδομένων δίνοντας πολύ ικανοποιητικά αποτελέσματα τόσο στο διαχωρισμό φυσιολογικών από δείγματα καρκίνου του εντέρου όσο και στον διαχωρισμό μεταξύ δυο υποκατηγοριών της λευχαιμίας. Στην πρώτη περίπτωση το ποσοστό επιτυχίας έφτασε το 91,9% και στην δεύτερη το 100%. Τα αποτελέσματα δείχνουν ότι το προτεινόμενο σύστημα μπορεί να χρησιμοποιηθεί με επιτυχία για την υποβοήθηση της ιατρικής διάγνωσης.

Αναφορές

[1] B. Kaplan, "Evaluating informatics applications – clinical decision support systems literature review", *Int J Medical Informatics*, vol. 64, pp. 15-37, 2001.

[2] M. K. Deyholos, and D. W. Galbraith, "High-Density Microarrays for Gene Expression Analysis," *Cytometry*, vol. 43, pp. 229-238, Mar. 2001.

[3] D. Slonim, "From patterns to pathways: gene expression data analysis comes of age", *Nature Genetics*, vol. 32, pp. 502-508, 2002.

[4] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 4-37, Jan. 2000.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," Journal of the American Statistical Association, vol. 97, No. 457, pp. 77-87, Mar. 2002.

[6] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl.Acad. Sci. USA*, vol. 96, pp. 6745–6750, 1999.

[7] M. Schummer et al., "Comperative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas", Gene, Vol. 238, pp. 375-385, 1999.

[8] C.M. Perou et al., "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers", Proc. Natl.Acad. Sci. USA, Vol. 96, pp. 9212-9217, 1999.

[9] Beer D., Kardia S. et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma", Nature Medicine, Vol. 8, pp. 816-824, 2002.

[10] T.R. Golub, et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Vol. 286, pp. 531-537, 1999.

[11] Alizadeh, "Identification of Molecularly and Clinically Distinct Types of Diffuse Large B-Cell Lymphoma By Gene Expression Data", Nature, Vol. 403, pp. 503-511, 2000.

[12] S. Dudoit et al., "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", Technical report 578, Dep. of Statistics, Univ. of California at Berkley, 2000.

[13] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, pp. 546-554, Apr. 2002.

[14] M. Sun, M. Xiong, "A mathematical programming approach for gene selection and tissue classification", *Bioinformatics*, vol. 19, pp. 1243-1251, Oct. 2003.

[15] V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, 1995.

[16] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Kluwer Academic Publishers, Boston, 1998.