ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ ΑΝΑΓΝΩΡΙΣΗΣ ΠΡΟΤΥΠΩΝ ΣΕ ΦΑΣΜΑΤΑ ΜΑΖΑΣ ΚΑΡΚΙΝΟΥ ΤΟΥ ΠΡΟΣΤΑΤΗ

Σ. Κωστόπουλος[1], Δ. Γκλώτσος[1], Π. Ασβεστάς[1], Γ. Σακελλαρόπουλος[2], Ι. Καλατζής[1]
[1] Τμήμα Μηχανικών Βιοϊατρικής Τεχνολογίας Τ.Ε. ΤΕΙ Αθήνας
[2] Εργαστήριο Ιατρικής Φυσικής, Τμήμα Ιατρικής, Πανεπιστήμιο Πάτρας

ΠΕΡΙΛΗΨΗ

Παρά την ευρεία κλινική χρήση του Ειδικού Προστατικού Αντιγόνου (PSA), που οφείλεται στην υψηλή του ευαισθησία, εν τούτοις η ειδικότητά του είναι χαμηλή, ελαττώνοντας τη σημαντικότητά του ως διαγνωστικού εργαλείου για πρώιμη διάγνωση του καρκίνου του προστάτη. Στην παρούσα εργασία υλοποιήθηκε σύστημα αναγνώρισης προτύπων σε ελεύθερα διατιθέμενα προστατικά δεδομένα φασματοσκοπίας μάζας, με σκοπό την προσπάθεια διάκρισης μεταξύ κακοήθων και μη περιστατικών καρκίνου του προστάτη. Τα φάσματα MS υπέστησαν την κατάλληλη επεξεργασία και ανάλυση με εξαγωγή διαστημάτων λόγων m/z, τα οποία χρησιμοποιήθηκαν ως είσοδοι στο σύστημα ταξινόμησης. Τα αποτελέσματα έδειξαν ότι η μέθοδος μπορεί να χρησιμοποιηθεί για την ταξινόμηση φασμάτων MS και κατά συνέπεια θα μπορούσε να συμβάλλει στην πιθανή ανεύρεση βιοδεικτών για τον καρκίνο του προστάτη.

Λέξεις κλειδιά: *Αναγνώριση προτύπων, Φασματοσκοπία μάζας, Καρκίνος του προστάτη*

Εισαγωγή

Η ανάλυση δεδομένων που προέρχονται από μελέτες φασματοσκοπίας μάζας (Mass Spectrometry, MS) βοηθά στην κατανόηση της συσχέτισης των πρωτεϊνών ή/και των πεπτιδίων με διάφορες ασθένειες, καθώς και στην πρώιμη διάγνωση του καρκίνου μέσω ανίχνευσης των πρωτεϊνών που εμπλέκονται στις σχετικές βιοχημικές διαδικασίες.

Ο βασικός δείκτης για την ύπαρξη καρκίνου του προστάτη είναι το Ειδικό Προστατικό Αντιγόνο (PSA). Παρά την ευρεία του κλινική χρήση, που οφείλεται στην υψηλή του ευαισθησία, εν τούτοις η ειδικότητά του είναι χαμηλή, ελαττώνοντας τη σημαντικότητά του ως διαγνωστικού εργαλείου για πρώιμη διάγνωση του καρκίνου του προστάτη.

Στις περισσότερες σχετικές ερευνητικές μελέτες με συστήματα αναγνώρισης προτύπων, το ποσοστό ορθής ταξινόμησης μεταξύ προστατικών φασμάτων MS που προέρχονται από φυσιολογικές και καρκινικές περιπτώσεις είναι πολύ υψηλό. Ωστόσο, μεγάλο ερευνητικό ενδιαφέρον παρουσιάζει η ταυτοποίηση της ομάδας πρωτεϊνών, όπως αυτές αντιπροσωπεύονται στα φάσματα MS μέσω λόγων μάζας προς φορτίο (m/z), που παρέχει τη μέγιστη διαχωριστική ικανότητα μεταξύ των δύο κατηγοριών και, ταυτόχρονα, συνδέεται ευθέως με τον καρκίνο του προστάτη, ώστε να μπορεί να χρησιμοποιηθεί ως αξιόπιστος βιοδείκτης για τη συγκεκριμένη πάθηση.

Η φασματοσκοπία μάζας (mass spectrometry, MS) είναι μια πολλά υποσχόμενη μέθοδος που μπορεί να συμβάλει στην αναγνώριση βιοδεικτών διάφορων ασθενειών σε πρώιμα στάδια [1], λόγω της ταχείας ανάπτυξής της τα τελευταία χρόνια.

Στην παρούσα εργασία αναπτύχθηκε σύστημα επεξεργασίας και ανάλυσης φασμάτων MS και υλοποιήθηκε σύστημα αναγνώρισης προτύπων με σκοπό τη διερεύνηση της δυνατότητας διάκρισης μεταξύ φασμάτων MS που προέρχονται από κακοήθη καρκίνο του προστάτη από περιπτώσεις καλοήθους ή φυσιολογικής μορφής.

Υλικό

Χρησιμοποιήθηκαν τα δεδομένα από το ερευνητικό πρόγραμμα του Εθνικού Κέντρου Έρευνας Καρκίνου των Ηνωμένων Πολιτειών Αμερικής, “Clinical Proteomics Program”, σχετικά με την δημιουργία προφίλ, εξερεύνησης και αναγνώρισης βιοδεικτών στον καρκίνο του προστάτη [2].

Οι ασθενείς από τα οποία προέρχονται τα δεδομένα είχαν όλοι ηλικία μεγαλύτερη ή ίση των 50 ετών, στην περίοδο 1996-2001. Τα διαθέσιμα δεδομένα προέρχονται επίσης από το Καθολικό Πανεπιστήμιο της Χιλής (Σαν Ντιέγκο, Χιλή), από το Εθνικό Κέντρο Καρκίνου των Ηνωμένων Πολιτειών Αμερικής, καθώς και από το Simone Protective Cancer Institute (Lawrenceville, NJ).

Τα δεδομένα αυτά αποτελούνται από 322 φάσματα μάζας υποδιαιρεμένα σε 4 κατηγορίες (63 φυσιολογικές περιπτώσεις με PSA<1, 190 καλοήθεις περιπτώσεις με PSA>4, 26 κακοήθεις περιπτώσεις με PSA από 4 έως 10 και 43 κακοήθεις περιπτώσεις με PSA>10).

Η απόκτηση των δεδομένων έγινε χρησιμοποιώντας το πρωτεϊνικό τσιπ H4 και φασματογράφο μάζας SELDI-TOF χαμηλής ανάλυσης Ciphergen PBS1. Κάθε φάσμα εξέτασης αποτελείται από ένα αρχείο χαρακτήρων ASCII με τιμές διαχωριζόμενες από κόμμα (csv). Κάθε φάσμα αποτελείται από ένα σύνολο περίπου 15200 σημείων, στα οποία αποτυπώνεται η αντίστοιχη τιμή του λόγου μάζας προς φορτίου (m/z). Από τα φάσματα έχει γίνει εξαγωγή γραμμής βάσης.

Μέθοδος

Από τα διαθέσιμα δεδομένα δημιουργήθηκαν δύο κλάσεις για είσοδο στο σύστημα αναγνώρισης προτύπων. Η 1η κλάση (**HL-BE**) δημιουργήθηκε με συνένωση των διανυσμάτων χαρακτηριστικών που προήλθαν από τις 63 φυσιολογικές περιπτώσεις με PSA<1 (HL) και από τις 190 καλοήθεις περιπτώσεις με PSA>4 (BE). Η 2η κλάση (**ML**) δημιουργήθηκε με συνένωση των διανυσμάτων χαρακτηριστικών που προήλθαν από τις 26 κακοήθεις περιπτώσεις με PSA από 4 έως 10 και 43 κακοήθεις περιπτώσεις με PSA>10.

Τα φάσματα MS υπέστησαν προεπεξεργασία που περιελάμβανε επαναδειγματοληψία [3], η αφαίρεση γραμμής βάσης [4], η κανονικοποίηση [3] και η εξομάλυνση [5-7] με αφαίρεση θορύβου υποβάθρου [8]. Τα στάδια αυτά ήταν απαραίτητα ώστε στη συνέχεια να είναι δυνατή η ανίχνευση των κορυφών των φασμάτων και η ευθυγράμμιση των κορυφών. Στη συνέχεια ακολούθησε ανίχνευση των κορυφών των φασμάτων MS [2,9,10] και ευθυγράμμισή τους [11], με δημιουργία των τελικών λόγων m/z [12], που αποτελούν τα χαρακτηριστικά του συστήματος ταξινόμησης. Για ευκολία αναφοράς, τα χαρακτηριστικά αριθμούνται διαδοχικά από το 1 έως το 170. Οι αντιστοιχίες μεταξύ των αυξόντων αριθμών των χαρακτηριστικών και των λόγων m/z βρίσκονται στο Παράρτημα Ι.

Το σύστημα ταξινόμησης που χρησιμοποιήθηκε περιλαμβάνει τον ταξινομητή Πιθανοκρατικού Νευρωνικού Δικτύου (Probabilistic Neural Network, PNN) [13] με τη γκαουσσιανή συνάρτηση ενεργοποίησης (activation function), όπως υλοποιήθηκε σε προηγούμενο παραδοτέο του Υποέργου. Η τιμή της παραμέτρου εξομάλυνσης ήταν ίση με 0.25 στην ομάδα δεδομένων εκπαίδευσης και 0.75 στην ομάδα ελέγχου.
Ο ταξινομητής Πιθανοκρατικού Νευρωνικού Δικτύου (PNN, Probabilistic Neural Network) είναι ένας μη παραμετρικός ταξινομητής, ο οποίος μπορεί να αντικαταστήσει τον Μπαεζιανό (Bayesian) ταξινομητή στις περιπτώσεις κλάσεων που τα μέλη τους δεν ακολουθούν την κανονική κατανομή ή που δεν είναι γνωστή η συνάρτηση πυκνότητας πιθανότητας.
Για σταθερή παράμετρο εξομάλυνσης σ, η συνάρτηση διάκρισης του ταξινομητή PNN με γκαουσσιανή συνάρτηση ενεργοποίησης παίρνει τη μορφή:

$$f_C(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{(\mathbf{x}-\mathbf{x}_{Ci})^T(\mathbf{x}-\mathbf{x}_{Ci})}{2\sigma^2}\right)$$

όπου n είναι το πλήθος των προτύπων εκπαίδευσης, $\mathbf{x}_{Ci}$ είναι η i-οστό πρότυπο εκπαίδευσης της κατηγορίας C, d είναι η διάσταση του χώρου των προτύπων και σ η παράμετρος εξομάλυνσης.
Ως κριτήρια επιλογής βέλτιστων συνδυασμών χαρακτηριστικών χρησιμοποιήθηκαν το ολικό ποσοστό επιτυχίας (overall accuracy, OA) και το εμβαδόν κάτωθεν της καμπύλης χαρακτηριστικού λειτουργικού δέκτη (area under the receiver operating characteristic curve, AuROC), όπως αυτά λαμβάνονται από τη συνάρτηση του χρησιμοποιούμενου ταξινομητή, καθώς και το κριτήριο διαχωρισιμότητας κλάσεων $J3$, το οποίο ορίζεται μέσω των πινάκων

διασποράς «εντός-των-κλάσεων» (within-class scatter matrix) και «μεταξύ-των-κλάσεων» (between-class scatter matrix) [14].

Για κάθε ολικό ποσοστό επιτυχούς ταξινόμησης, υπολογίστηκε και ο αντίστοιχος πίνακας αληθείας, στον οποίον οι γραμμές αντιστοιχούν στις κλάσεις όπου ανήκουν τα πρότυπα και οι στήλες στις κλάσεις στις οποίες αυτά ταξινομήθηκαν. Επίσης, υπολογίστηκαν και παρουσιάζονται οι δείκτες εγκυρότητας κριτηρίου (ευαισθησία και ειδικότητα) καθώς και οι προγνωστικές αξίες (θετική, αρνητική). Για την αξιολόγηση κάθε συνδυασμού χαρακτηριστικών, με εφαρμογή συνάρτησης ταξινομητή, χρησιμοποιήθηκε η μέθοδος της διαδοχικής αφαίρεσης ενός προτύπου (leave-one-out, LOO) [14].

Οι βέλτιστοι συνδυασμοί χαρακτηριστικών ανά μέθοδο αξιολόγησης ερευνήθηκαν με συνδυασμό των τεχνικών εξαντλητικής έρευνας (exhaustive search) και διαδοχικής εμπρόσθιας επιλογής (sequential forward selection, SFS) [14]. Συγκεκριμένα, πραγματοποιήθηκε πλήρης έρευνα με μέγιστη διάσταση διανύσματος χαρακτηριστικών = 2 και ακολούθως διαδοχική εμπρόσθια επιλογή μέχρι διάσταση διανύσματος χαρακτηριστικών = 6 (EXS-SFS). Επίσης, εφαρμόστηκε και η μέθοδος διαδοχικής εμπρόσθιας επιλογής με επίπλευση (sequential forward floating selection, SFFS), η οποία, σε σχέση με την SFS, έχει το πλεονέκτημα της μη υποχρεωτικής διατήρησης ενός χαρακτηριστικού από το ένα βήμα στο επόμενο [14].

Για την τελική αξιολόγηση του συστήματος εφαρμόστηκε η μέθοδος της Εξωτερικής Διασταυρούμενης Επικύρωσης. Σύμφωνα με τη μέθοδο αυτή, αρχικά κάθε κλάση διαιρέθηκε με τυχαίο τρόπο σε σύνολο εκπαίδευσης (2/3 των προτύπων κάθε κλάσης) και σύνολο ελέγχου (1/3 των προτύπων κάθε κλάσης). Στη συνέχεια, το σύνολο εκπαίδευσης χρησιμοποιήθηκε για την εύρεση της βέλτιστης ομάδας λόγων m/z μέσω των κριτηρίων διαχωρισμού κλάσεων για τους συνδυασμούς χαρακτηριστικών που προέκυψαν από τις μεθόδους EXS-SFS και SFFS. Η βέλτιστη ομάδα χαρακτηριστικών αξιολογήθηκε ως προς την ικανότητά της για την κατηγοριοποίηση αγνώστων φασμάτων MS μέσω της ταξινόμησης των φασμάτων MS της ομάδας ελέγχου. Η διαδικασία επαναλήφθηκε 10 φορές και ελήφθη η μέση τιμή του ολικού ποσοστού ορθής ταξινόμησης των προτύπων της ομάδας ελέγχου.

Αποτελέσματα

Οι τιμές των κριτηρίων διαχωρισμού των 2 κατηγοριών φασμάτων MS, για κάθε μια από τις μεθόδους αξιολόγησης και για κάθε μέθοδο επιλογής του βέλτιστου συνδυασμού χαρακτηριστικών, δίνονται στους παρακάτω πίνακες.

Κριτήριο διαχωρισιμότητας κλάσεων J3

Πίνακας 1. Μέγιστη τιμή του κριτηρίου *J*3 του βέλτιστου συνδυασμού για κάθε διάσταση του διανύσματος χαρακτηριστικών με τη συνδυαστική τεχνική EXS-SFS.

Διάσταση	***J*3(max)**	**Συνδυασμός**
1	1.1117	101
2	1.1530	15,28
3	1.1552	15,28,101
4	1.1648	15,28,101,33
5	1.1589	15,28,101,33,5
6	1.1475	15,28,101,33,5,22

Πίνακας 2. Μέση τιμή και τυπική απόκλιση των ποσοστών επιτυχούς ταξινόμησης της ομάδας ελέγχου με τον ταξινομητή PNN για 10 επαναλήψεις της ECV, με χρήση του βέλτιστου συνδυασμού χαρακτηριστικών όπως αυτός προέκυψε με χρήση του κριτηρίου *J*3 και την EXS-SFS.

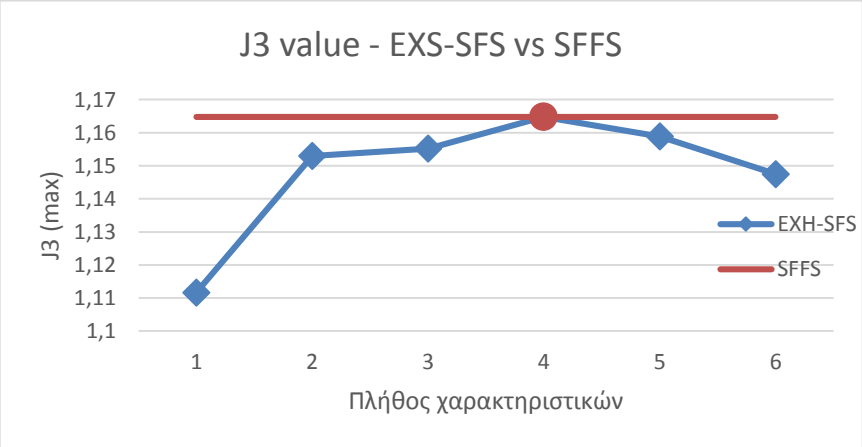
	OA (%)	**Ειδικότητα (%)**	**Ευαισθησία (%)**
Μέση τιμή	78.9	78.2	81.3
Τυπική απόκλιση	6.8	7.4	23.6

Πίνακας 3. Μέγιστη τιμή του κριτηρίου *J*3 για το βέλτιστο συνδυασμό χαρακτηριστικών της τεχνικής SFFS.

Διάσταση	***J*3(max)**	**Συνδυασμός**
4	1.1648	15,28,101,33

Πίνακας 4. Μέση τιμή και τυπική απόκλιση των ποσοστών επιτυχούς ταξινόμησης της ομάδας ελέγχου με τον ταξινομητή PNN για 10 επαναλήψεις της ECV, με χρήση του βέλτιστου συνδυασμού χαρακτηριστικών όπως αυτός προέκυψε με χρήση του κριτηρίου *J*3 και την SFFS.

	OA (%)	**Ειδικότητα (%)**	**Ευαισθησία (%)**
Μέση τιμή	78.4	76.0	87.4
Τυπική απόκλιση	9.3	11.7	9.0



Διάγραμμα 1: Σύγκριση των μέγιστων τιμών του κριτηρίου *J*3 για τους βέλτιστους συνδυασμούς χαρακτηριστικών, όπως αυτοί προκύπτουν από τις τεχνικές EXS-SFS και SFFS, όπως προκύπτει από τους Πίνακες 1 και 3.

Μέθοδος Διαδοχικής Παράλειψης Ενός Προτύπου

Πίνακας 5. Ποσοστό επιτυχίας (OA) του βέλτιστου συνδυασμού χαρακτηριστικών, όπως προέκυψε με την τεχνική EXS-SFS, για κάθε διάσταση του διανύσματος χαρακτηριστικών.

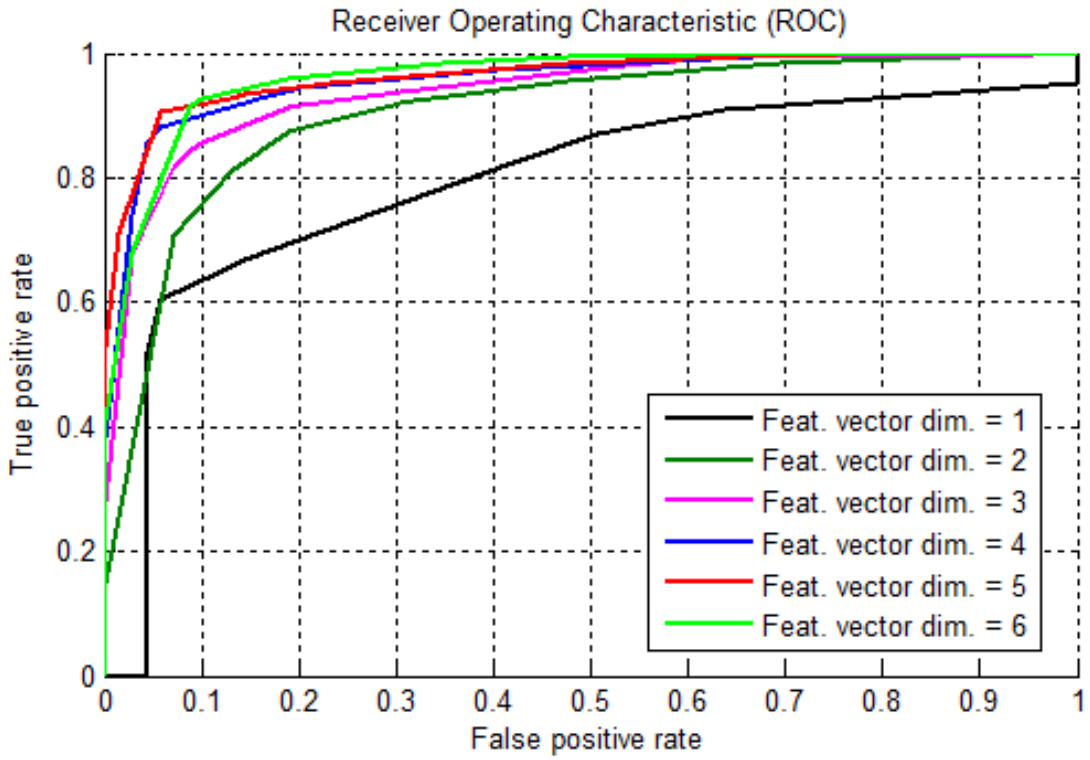
Διάσταση	OA (%)	Συνδυασμός
1	78.6	10
2	79.2	7,14
3	80.8	7,14,53
4	80.4	7,14,53,63
5	81.1	7,14,53,63,72
6	83.5	7,14,53,63,72,29

Πίνακας 6. Μέγιστη τιμή του εμβαδού κάτωθεν της καμπύλης ROC (AuROC) του βέλτιστου συνδυασμού χαρακτηριστικών, όπως προέκυψε με την τεχνική EXS-SFS, για κάθε διάσταση του διανύσματος χαρακτηριστικών.

Διάσταση	AuROC	Συνδυασμός
1	0.7853	31
2	0.8876	31,74
3	0.9265	31,74,16
4	0.9471	31,74,16,72
5	0.9573	31,74,16,72,73
6	0.9547	31,74,16,72,73,53

Πίνακας 7. Μέση τιμή και τυπική απόκλιση των ποσοστών επιτυχούς ταξινόμησης της ομάδας ελέγχου με τον ταξινομητή PNN για 10 επαναλήψεις της ECV, με χρήση του βέλτιστου συνδυασμού χαρακτηριστικών όπως αυτός προέκυψε με χρήση του OA με LOO του ταξινομητή PNN και την EXS-SFS.

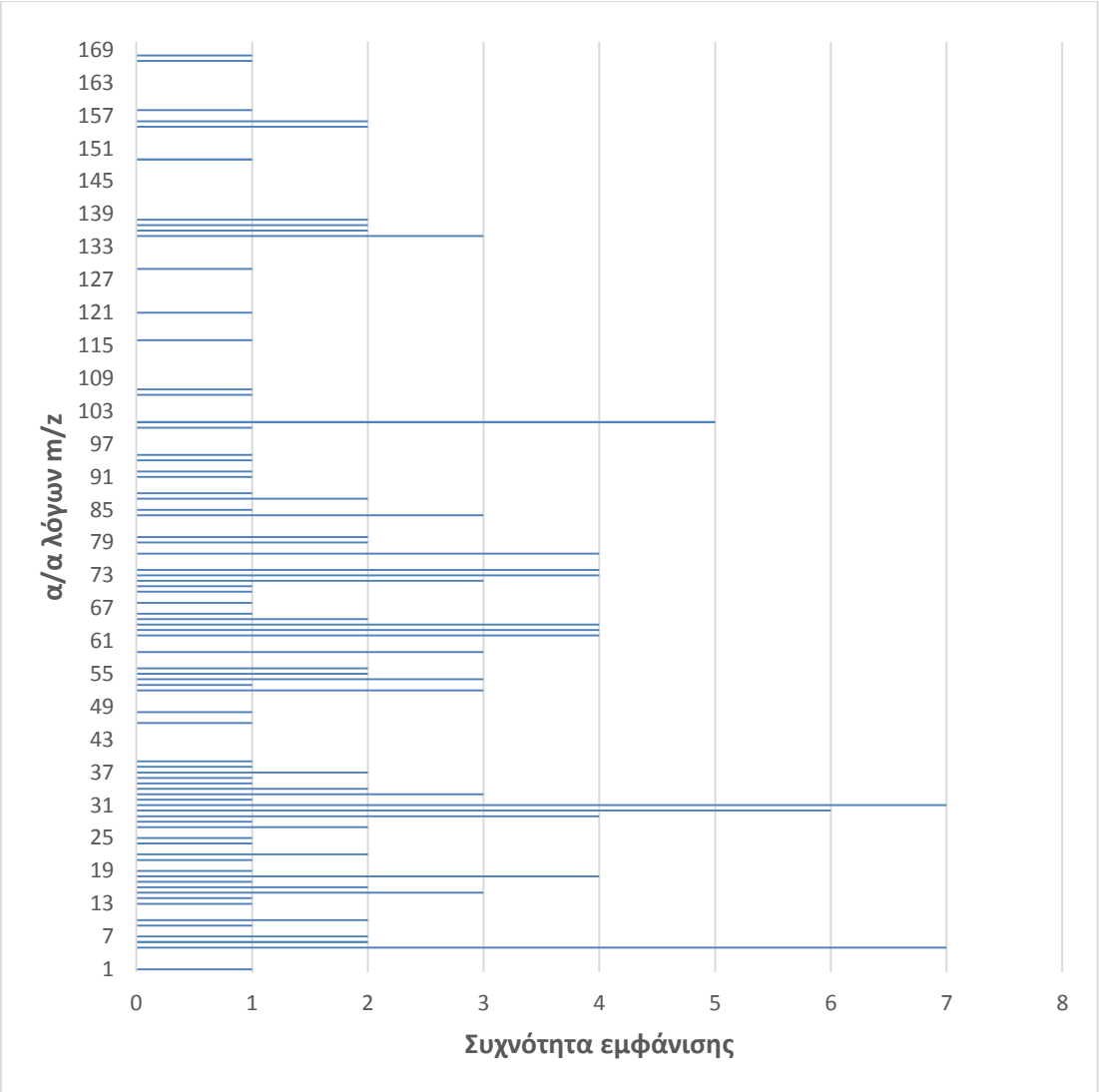
	OA (%)	Ειδικότητα (%)	Ευαισθησία (%)
Μέση τιμή	75.5	76.8	70.4
Τυπική απόκλιση	3.2	5.8	20.7



Διάγραμμα 2. Καμπύλες ROC, όπως προκύπτουν από τα στοιχεία του Πίνακα 6.

Πίνακας 8. Μέση τιμή και τυπική απόκλιση των ποσοστών επιτυχούς ταξινόμησης της ομάδας ελέγχου με τον ταξινομητή PNN για 8 επαναλήψεις της ECV, με χρήση του βέλτιστου συνδυασμού χαρακτηριστικών όπως αυτός προέκυψε με χρήση του OA με LOO του ταξινομητή PNN και την SFFS.

	OA (%)	**Specificity (%)**	**Sensitivity (%)**
Μέση τιμή	80.8	84.4	67.4
Τυπική απόκλιση	3.3	5.3	10.1



Διάγραμμα 3. Συχνότητα εμφάνισης των διαστημάτων των λόγων m/z στα συχνότερα εμφανιζόμενα βέλτιστα διανύσματα χαρακτηριστικών, που βρέθηκαν με την τεχνική SFFS δίνοντας το υψηλότερο ολικό ποσοστό επιτυχίας με τη μέθοδο LOO.

Συζήτηση

Το πολύ μεγάλο πλήθος χαρακτηριστικών, που είναι εγγενές για τα δεδομένα φασμάτων MS (λόγοι m/z), καθιστά ιδιαιτέρως δύσκολη την ερευνά για το βέλτιστο συνδυασμό τους που μπορεί να οδηγήσει σε καλύτερη ταξινόμηση και κατ' επέκταση να οδηγήσει, σε συνδυασμό με άλλες παραμέτρους, στην πιθανή ανεύρεση βιοδεικτών για τον καρκίνο του προστάτη. Η μέθοδος εξαντλητικής έρευνας είναι πρακτικώς αδύνατον να εφαρμοστεί. Για παράδειγμα, στην περίπτωση των δεδομένων που χρησιμοποιήθηκαν στην παρούσα έρευνα, το πλήθος των λόγων m/z μετά την προεπεξεργασία και ανάλυση των φασμάτων MS ήταν 170, που οδηγεί σε συνολικό πλήθος συνδυασμών προς διερεύνηση της τάξης του 10^{51}.

Για τον ανωτέρω λόγο, και ανεξαρτήτως της υπολογιστικής ισχύος, η χρήση υποβέλτιστων μεθόδων στην έρευνα για τον κατά το δυνατόν καλύτερο συνδυασμό χαρακτηριστικών είναι αναγκαία.

Χρησιμοποιήθηκαν οι τεχνικές EXS-SFS και η SFFS. Η SFFS αποδείχθηκε ιδιαίτερα ταχεία στα συγκεκριμένα δεδομένα, ενώ ταυτόχρονα ανέδειξε βέλτιστους συνδυασμούς μεγάλου πλήθους χαρακτηριστικών, κάτι πολύ δύσκολο να ανευρεθεί με την EXS-SFS, πολύ περισσότερο δε με τη μέθοδο της εξαντλητικής έρευνας. Επίσης, οι συνδυασμοί που ευρέθησαν

με την SFFS έδωσαν το υψηλότερο μέσο ποσοστό επιτυχούς ταξινόμησης, τόσο στο στάδιο εκπαίδευσης, όσο (το σημαντικότερο) και στο στάδιο ελέγχου.

Η τεχνική SFFS αποδείχθηκε ιδιαίτερα ταχεία, ενώ έδωσε τον ίδιο συνδυασμό ως βέλτιστο με αυτόν που προέκυψε από την EXS-SFS βάσει του κριτηρίου *J*3. Επίσης, με την τεχνική SFFS μπόρεσε να αναδειχθεί, βέλτιστος συνδυασμός 30 χαρακτηριστικών, κάτι πολύ δύσκολο να ανευρεθεί με την EXS-SFS, πολύ περισσότερο δε με τη μέθοδο της εξαντλητικής έρευνας.

Το υψηλότερο μέσο ποσοστό επιτυχούς ταξινόμησης στην ομάδα ελέγχου (80.8%) επιτεύχθηκε με τον ταξινομητή Πιθανοκρατικού Νευρωνικού Δικτύου, με συνδυασμό χαρακτηριστικών ο οποίος βρέθηκε με χρήση της μεθόδου παράλειψης ενός προτύπου, ερευνώντας τους συνδυασμούς χαρακτηριστικών με την τεχνική SFFS. Οι συνδυασμοί που αναδείχθηκαν σε κάθε επανάληψη με την προηγούμενη διαδικασία αποτελούντο από ιδιαιτέρως υψηλό πλήθος χαρακτηριστικών (από 11 έως 31 χαρακτηριστικά, μέση τιμή 20 χαρακτηριστικά).

Η παραπάνω διαδικασία (OA/PNN με LOO, έρευνα με SFFS) έδωσε επίσης και την υψηλότερη μέση τιμή ειδικότητας (84.4%), αποτέλεσμα που κρίνεται σημαντικό δεδομένης της χαμηλής ειδικότητας στην ανίχνευση του καρκίνου του προστάτη με τη χρήση του βασικότερου κλινικού δείκτη, του ειδικού προστατικού αντιγόνου (PSA) [1]. Παρόλ' αυτά, η υψηλότερη τιμή ευαισθησίας επετεύχθη με το κριτήριο *J*3 (87.4% με έρευνα μέσω SFFS), με πολύ μικρό συγκριτικά πλήθος χαρακτηριστικών (από 1 έως 6, μέση τιμή 4). Αυτό το αποτέλεσμα μπορεί να οδηγήσει στο πιθανό συμπέρασμα ότι η δομή των δεδομένων είναι αρκετά απλή, τουλάχιστον όσον αφορά στο διαχωρισμό τους για την ανίχνευση των κακοήθων περιστατικών. Το συμπέρασμα αυτό μπορεί να υποστηριχθεί από την απλή στατιστικά κατασκευή του κριτηρίου *J*3 (μεγαλύτερη διασπορά μεταξύ των κατηγοριών σε σχέση με τη διασπορά εντός των δεδομένων κάθε κατηγορίας).

Συμπεράσματα

Στην παρούσα εργασία παρουσιάζονται τα αποτελέσματα έρευνας για βέλτιστο συνδυασμό λόγων m/z φασμάτων MS, που μπορεί να οδηγήσουν σε διαχωρισμό μεταξύ των δύο κατηγοριών στις οποίες χωρίστηκαν τα δεδομένα που χρησιμοποιήθηκαν (υγιείς/καλοήθεις – κακοήθεις περιπτώσεις. Έγινε σύγκριση μεταξύ υποβέλτιστων μεθόδων αναζήτησης συνδυασμού χαρακτηριστικών (λόγων m/z) με χρήση κριτηρίων διαχωρισμού κλάσεων και ταξινομητή πιθανοκρατικού νευρωνικού δικτύου.
Χρησιμοποιήθηκε η μέθοδος της εξωτερικής διασταυρούμενης επικύρωσης, με χωρισμό των δεδομένων σε ομάδα εκπαίδευσης και ομάδα ελέγχου με λόγο πλήθους στοιχείων 2:1. Ο εντοπισμός του βέλτιστου συνδυασμού χαρακτηριστικών, ο οποίος ήταν μεγάλης διαστατικότητας, αναδείχθηκε με την τεχνική SFFS με τη μέθοδο παράλειψης ενός προτύπου για τον ταξινομητή που χρησιμοποιήθηκε (PNN), ο οποίος έδωσε τόσο το υψηλότερο ολικό ποσοστό επιτυχίας όσο και την υψηλότερη ειδικότητα. Η υψηλότερη ευαισθησία επετεύχθη με χρήση του κριτηρίου *J*3, με σημαντικά μικρότερη διαστατικότητα του διανύσματος χαρακτηριστικών.

Ευχαριστίες

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: **ΑΡΧΙΜΗΔΗΣ ΙΙΙ**. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου

Αναφορές

[1] McDavid K., Lee J., Fulton J. P., Tonita J. and Thompson T. D. (2004). "Prostate cancer incidence and mortality rates and trends in the United States and Canada", *Public Health Rep*, 119 : 174-86.

[2] Petricoin Iii E.F., Ornstein D.K., Paweletz C.P., Ardekani A., Hackett P.S., Hitt B.A., Velassco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC and Liotta LA. (2002). "Serum proteomic patterns for detection of prostate cancer". *Journal of the National Cancer Institute*, 94 : 1576-8.

[3] Hilario M., Kalousis A., Pellegrini C. and Muller M. (2006). "Processing and classification of protein mass spectra". *Mass Spectrom Rev*, 25 : 409-449.

[4] Andrade L.and Manolakos E. (2003). "Signal background estimation and baseline correction algorithms for accurate DNA sequencing". *Bioinformatics*, 35 : 229–243, VLSI, special issue.

[5] Cleveland W.S. (1979). "Robust locally weighted regression and smoothing scatterplots". *J. Am. Stat. Assoc.*, 74 : 829–836.

[6] Savitzky A. and Golay M. (1964). "Smoothing and differentiation of data by simplified least squares procedure". *Anal Chem*, 36 : 1627–1639.

[7] Coombes K.R., Tsavachidis S., Morris J.S., Baggerly K.A., Hung M.C. and Kuerer H.M., (2005). "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform". *Proteomics*, 5 : 4107–4117.

[8] Jeffries N., (2005). "Algorithms for alignment of mass spectrometry proteomic data". *Bioinformatics*, 21 : 3066-3073.

[9] Prados J., Kalousis A., Sanchez J.C., Allard L., Carrette O. and Hilario M. (2004). "Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents". *Proteomics*, 4 : 2320-2332.

[10] Wu B., Abbott T., Fishman D., McMurray W., Mor G., Stone K., Ward D., Williams K. and Zhao H. (2003). "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data". *Bioinformatics*, 19 : 1636-1643.

[11] Wang X. (2006). "Feature extraction in the analysis of proteomic mass spectra". *Proteomics*, 6 : 2095-2100.

[12] Wasinger V.C. and Zeng M., Yau Y. (2013). "Current status and advances in quantitative proteomic mass spectrometry". *International Journal of Proteomics*, 2013:180605.

[13] Specht, D.F., (1990). "Probabilistic Neural Networks". *Neural Networks* 3 : 109-118.

[14] Theodoridis S. and K. Koutroumbas K., (2003). "Pattern Recognition". *2nd edition Academic Pr*

Παράρτημα Ι

Αντιστοιχίες αυξόντων αριθμών αναφοράς των χαρακτηριστικών και εύρους λόγων m/z

α/α χαρακτ.	Εύρος m/z		α/α χαρακτ.	Εύρος m/z	
1	0,022436	0,025319	34	1130,6556	1135,6821
2	0,025319	0,031607	35	1209,1693	1215,0177
3	0,031607	0,063727	36	1222,8375	1228,7188
4	0,063727	0,068526	37	1304,4364	1310,5105
5	0,201084	0,415257	38	1325,4178	1331,5405
6	0,415257	0,427372	39	1396,3322	1402,6164
7	0,427372	1,035717	40	1497,8499	1505,0823
8	2,343917	2,372580	41	1517,4174	1524,6968
9	2,372580	2,401417	42	1541,5056	1548,8425
10	3,520448	3,555556	43	1590,9955	1597,7030
11	11,22793	11,29056	44	1602,9297	1610,4112
12	26,63904	26,73546	45	1635,9785	1643,5366
13	43,54016	43,66341	46	1679,2932	1686,9506
14	49,91590	50,04785	47	1696,1625	1702,3177
15	65,32444	65,62650	48	1773,1177	1781,7737
16	73,08498	73,40446	49	1813,4276	1822,1813
17	89,02972	89,38230	50	1862,2364	1871,1069
18	93,48699	93,84828	51	1911,6933	1920,6808
19	102,5390	102,9174	52	2002,5162	2011,7144
20	116,5988	117,0022	53	2034,3812	2042,8085
21	196,7599	197,5461	54	2047,8733	2058,0216
22	204,4258	205,2272	55	2069,8931	2080,0957
23	221,5840	222,4183	56	2162,6202	2173,0486
24	225,2106	226,3324	57	2230,8533	2241,4448
25	226,6133	227,7386	58	2242,3285	2252,9472
26	242,0399	243,2028	59	2313,5933	2324,3791
27	337,6784	339,0517	60	2336,9944	2347,8346
28	355,7492	357,5116	61	2436,3804	2448,3720
29	462,3529	464,3617	62	2462,2449	2474,2998
30	507,5540	510,0802	63	2514,3835	2526,5653
31	524,5134	527,0813	64	2593,6157	2605,9876
32	1030,6691	1035,4684	65	2698,2611	2710,8798
33	1046,9115	1051,1432	66	2830,7366	2844,6565
α/α χαρακτ.	**Εύρος m/z**		**α/α χαρακτ.**	**Εύρος m/z**	
67	2875,6006	2889,6302	103	7618,9701	7656,4877
68	2890,6336	2904,6998	104	8060,1740	8095,4022
69	2935,9679	2950,1438	105	8225,2294	8264,2092
70	3111,4551	3125,0046	106	8427,9028	8469,0770
71	3284,5436	3300,6087	107	9272,2475	9317,2341

72	3356,6045	3371,7608	108	10613,366	10665,348
73	3452,4397	3468,9098	109	10919,355	10972,079
74	3470,0092	3485,4190	110	10974,035	11028,851
75	3578,5954	3596,4826	111	11030,811	11085,768
76	3671,8696	3689,9882	112	11087,734	11142,833
77	3710,4248	3728,6381	113	11144,803	11200,043
78	3749,1814	3766,3439	114	11202,019	11257,400
79	3928,4172	3947,1572	115	11259,381	11314,904
80	3976,5278	3995,3821	116	11316,890	11372,554
81	4046,2741	4059,3447	117	11374,545	11430,351
82	4232,3946	4253,0625	118	11490,295	11546,384
83	4466,2515	4487,4820	119	11548,389	11604,620
84	4650,2245	4673,1632	120	11606,630	11663,002
85	4841,8060	4856,1029	121	11782,233	11841,060
86	4989,6779	5013,4380	122	11843,091	11902,070
87	5235,2199	5260,9107	123	11904,106	11963,236
88	5592,2357	5620,1861	124	11965,278	12024,560
89	5621,5855	5649,6091	125	12026,607	12086,040
90	5651,0121	5679,1089	126	12088,092	12147,677
91	5769,4869	5797,8763	127	12149,734	12209,471
92	5799,2976	5827,7602	128	12211,533	12271,421
93	5829,1852	5857,7209	129	12397,870	12458,213
94	5859,1496	5887,7585	130	12460,296	12520,790
95	6011,5707	6040,5489	131	12522,879	12583,525
96	6106,0046	6135,2093	132	12585,619	12648,515
97	6204,1144	6235,0262	133	12650,614	12713,672
98	6236,5001	6267,4924	134	12781,107	12844,490
99	6268,9701	6300,0428	135	12846,605	12910,149
100	6691,1159	6723,2166	136	12912,270	12975,976
101	6894,1674	6928,3043	137	12978,103	13041,971
102	7466,5954	7503,7364	138	13044,103	13108,133
α/α χαρακτ.	**Εύρος m/z**				
139	13243,107	13307,623			
140	13309,776	13374,454			
141	13376,613	13441,453			
142	13443,617	13510,789			
143	13792,107	13860,143			
144	13862,341	13930,549			
145	14291,947	14361,203			
146	14603,790	14676,058			
147	15150,145	15223,750			
148	15226,053	15302,152			
149	15387,699	15464,200			
150	15706,533	15783,821			
151	15786,166	15863,650			

152	16026,274	16104,344			
153	16106,712	16184,978			
154	16356,396	16437,658			
155	16440,051	16521,520			
156	16523,919	16605,596			
157	16608,001	16689,885			
158	16692,296	16774,387			
159	17214,789	17300,608			
160	17303,063	17389,102			
161	17571,715	17658,418			
162	17660,898	17747,821			
163	17932,303	18019,890			
164	18400,212	18491,471			
165	18494,009	18585,500			
166	18684,870	18776,832			
167	19043,766	19136,606			
168	19139,188	19234,848			
169	19439,893	19536,300			
170	19538,909	19635,561			