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

Σχεδιασμός και Υλοποίηση Συστήματος αναγνώρισης προτύπων για ταξινόμηση πρωτεωμικών σημάτων φασματοσκοπίας μάζας (MS-SPECTRA) ασθενών με καρκίνο του προστάτη

Δημήτριος Κ. Σιδεράκης

Επιβλέπων: Διονύσης Κάβουρας, Καθηγητής

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σχεδιασμός και Υλοποίηση Συστήματος αναγνώρισης προτύπων για ταξινόμηση πρωτεωμικών σημάτων φασματοσκοπίας μάζας (MS-SPECTRA) ασθενών με καρκίνο του προστάτη

Δημήτριος Κ. Σιδεράκης

Α.Μ.: ΠΙΒ015

ΕΠΙΒΛΕΠΩΝ: **Διονύσιος Κάβουρας,** Καθηγητής

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Εμμανουήλ Σαγκριώτης,** Αναπληρωτής Καθηγητής
Ερρίκος Βεντούρας, Καθηγητής
Διονύσης Κάβουρας, Καθηγητής

Νοέμβριος 2011

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας διπλωματικής εργασίας ήταν να υλοποιηθεί ένα σύστημα αναγνώρισης προτύπων για το διαχωρισμό μεταξύ υγιών, καλοηθών και κακοηθών όγκων του προστάτη σε πρωτεωμικά δείγματα φασματοσκοπίας μάζας και ο εντοπισμός m/z διαστημάτων όπου πιθανόν να περιέχονται βιοδείκτες σχετιζόμενοι με τον καρκίνο του προστάτη. Για το σκοπό αυτό, χρησιμοποιήθηκαν δύο διαφορετικά σετ δεδομένων, ένα από το Εθνικό Καρκινικό Ινστιτούτο Αμερικής και ένα από το Ιατρικό κέντρο της Virginia, και τα οποία έχουν χρησιμοποιηθεί επανειλημμένα σε έρευνες σχετικά με τον καρκίνο του προστάτη. Λόγο της ιδιομορφίας των προς εξέταση φασμάτων, αρχικά απαιτήθηκε ένα στάδιο προ-επεξεργασίας τους (εξομάλυνση, εκτίμηση θορύβου, εύρεση και στοίχιση κορυφών) ώστε να καταστούν ικανά για περαιτέρω ανάλυση. Στο στάδιο αυτό πειραματιστήκαμε ενδελεχώς έτσι ώστε να καταλήξουμε στις βέλτιστες παραμέτρους για την προ-επεξεργασία των φασμάτων. Στην συνέχεια αναπτύχθηκαν πέντε διαφορετικοί ταξινομητές (MDC, KNN, Bayessian, PNN, SVM) καθώς και ένα σύστημα συνδυασμού αυτών έτσι ώστε να επιτευχθεί μέγιστη απόδοση. Για την εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών υλοποιήθηκαν οι εξαντλητική αναζήτηση, η sequential forward selection (SFS), η sequential backward selection (SBS), η sequential forward floating selection (SFFS) καθώς και η sequential backward floating selection (SBFS). Μετά από συνεχή πειραματισμό με τις παραπάνω τεχνικές και τα μοντέλα μηχανικής μάθησης, πετύχαμε υπό περιπτώσεις ακρίβεια της τάξεως του 95-98% για το πρώτο σετ δεδομένων και της τάξεως του 92-93% για το δεύτερο σετ δεδομένων. Επιπλέον, βασιζόμενοι στα χαρακτηριστικά τα οποία οι ταξινομητές χρησιμοποίησαν κατά κόρον κατά την επίτευξη της βέλτιστης απόδοσής τους, καταλήξαμε σε 6 διαστήματα m/z ως πιθανά να περιέχουν βιοδείκτες που σχετίζονται με τον καρκίνο τους προστάτη. Μετά από συσχετισμό με προηγούμενες έρευνες, εντοπίστηκαν προτεινόμενοι από άλλες ερευνητικές ομάδες βιοδείκτες εντός των προτεινόμενων από εμάς διαστημάτων m/z, κάτι που ενισχύει την θέση μας ως προς την υποψηφιότητα αυτών των διαστημάτων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Σήματος

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Πρωτεωμική, Φασματοσκοπία Μάζας, Αναγνώριση Προτύπων, βιοδείκτες, διάστημα m/z

ABSTRACT

The aim of this thesis was to implement a pattern recognition system for the discrimination amongst healthy, benign and malignant prostate tumors from proteomic mass spectroscopy samples and to identify m/z intervals of potential biomarkers associated with prostate cancer. For this reason, we used two different data sets, one from the National Cancer Institute of America and one from the East Virginia Medical School, which have been repeatedly used in researches about prostate cancer. Due to the specificity of tested spectra, initially there was a demand of pre-processing (smoothing, noise assessment, finding and peak alignment) to make them suitable for further analysis. At this stage we experimented thoroughly so as to find the optimal parameters for pre-processing of spectra. We then developed five different classifiers (MDC, KNN, Bayessian, PNN, SVM) and a system combining these so as to achieve maximum performance. For finding the optimal combination of features we implemented exhaustive search, sequential forward selection (SFS), sequential backward selection (SBS), sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS). After experimentation with these techniques and models of machine learning we achieved accuracy of 95-98% for the first set of data and of 92-93% for the second data set. Furthermore, based on the features the classifiers used when they achieved their optimal performance, we conclude at 6 different intervals of m/z as possible to contain biomarkers related to prostate cancer. After correlation with previous studies, biomarkers proposed by other research groups where found to be inside our proposed intervals of m/z, something that strengthens our position about the nomination of these intervals.

SUBJECT AREA: Signal Processing

KEYWORDS: Proteomics, Mass Spectrometry, Pattern Recognition, biomarkers, m/z interval