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 συστήματος αναγνώρισης προτύπων (PR-system) για ταξινόμηση πρωτεωμικών σημάτων φασματοσκοπίας μάζας (ms-spectra) ωοθηκών σε καλοήθη-κακοήθη

Χριστίνα Β. Κοτσιούρου

Επιβλέποντες: **Διονύσιος Κάβουρας,** Καθηγητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2010

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σχεδιασμός συστήματος αναγνώρισης προτύπων (PR-system) για ταξινόμηση πρωτεωμικών σημάτων φασματοσκοπίας μάζας (ms-spectra) ωοθηκών σε καλοήθη-κακοήθη

Χριστίνα Β. Κοτσιούρου

Α.Μ.: ΠΙΒ017

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Διονύσιος Κάβουρας,** Καθηγητής

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Εμμανουήλ Σαγκριώτης,** Καθηγητής
Ερρίκος Βεντούρας, Καθηγητής

Σεπτέμβριος 2010

ΠΕΡΙΛΗΨΗ

Στόχος της εργασίας αυτής είναι ο προσδιορισμός του υποσυνόλου των πρωτεϊνών, όπως αυτές αντιπροσωπεύονται στο σήμα φασματοσκοπίας μάζας (Mass Spectrometry – MS), που προσδίδουν στο σύστημα αναγνώρισης προτύπων (Pattern Recognition-PR-system) μέγιστη διαχωριστική ικανότητα μεταξύ καλοηθών και κακοηθών όγκων. Τελικός σκοπός της διαδικασίας αυτής είναι η ανάδειξη διαστημάτων μάζας/φορτίο τα οποία χρειάζονται περεταίρω ταυτοποίηση με τεχνικές που έχουν ως στόχο την ανάλυση επιλεγμένων μορίων με στόχο την ταυτοποίησή τους, για να διαπιστωθεί αν αποτελούν σημαντικούς βιοδείκτες για τον καρκίνο των ωοθηκών. Η συγκεκριμένη εργασία περιλαμβάνει ένα στάδιο προ-επεξεργασίας των MS-φασμάτων (εξομάλυνση, εκτίμηση θορύβου, στοίχιση κορυφών), την ανάπτυξη ενός συστήματος αναγνώρισης προτύπων, όπου θα δοκιμαστούν διάφοροι ταξινομητές αλλά και διαφορετικά σχήματα συνδυασμών ταξινομητών για την επίτευξη μέγιστης διαχωριστικής ικανότητας του PR-συστήματος και την ανάλυση και συσχετισμό των ευρημάτων με ευρήματα άλλων ερευνητών ώστε να εξαχθούν χρήσιμα συμπεράσματα που σχετίζονται με τον καρκίνο των ωοθηκών.

Η εργασία εστιάζει στον τομέα τη έγκαιρης και έγκυρης διάγνωσης του καρκίνου των ωοθηκών, επεξηγώντας αναλυτικά τόσο θεωρητικά όσο και πρακτικά τα βήματα που ακολουθούνται για την ανάπτυξη του μοντέλου ταξινόμησης. Οι αποδόσεις των διάφορων αλγορίθμων αγγίζουν πολύ ικανοποιητικά επίπεδα με ποσοστά επιτυχίας έως και 99%, ενώ ιδιαίτερη έμφαση δίνεται και στα πιο σημαντικά διαστήματα μάζας/φορτίου στα οποία καταλήγουν οι αλγόριθμο αυτοί.

Ο σκελετός της εργασίας είναι ο εξής: Στο πρώτο κεφάλαιο γίνεται μια γενική εισαγωγή στην ασθένεια του καρκίνου των ωοθηκών και στις μεθόδους πρόβλεψης της. Εισάγεται ο όρος της πρωτεωμικής και της φασματομετρίας μάζας και η σχέση της με την διάγνωση του καρκίνου των ωοθηκών. Στο δεύτερο κεφάλαιο αναλύονται οι τεχνικές προ-επεξεργασίας των φασμάτων μάζας ενώ στο τρίτο κεφάλαιο επεξηγούνται οι υπάρχουσες τεχνικές αναγνώρισης προτύπων, μοντέλων ταξινόμησης και μεθόδων επιλογής χαρακτηριστικών, εκτίμησης των μοντέλων αυτών και ερμηνεία των αποτελεσμάτων τους. Στο τέλος του θεωρητικού μέρους αναφέρονται οι υπάρχουσες έρευνες στο θέμα της ανίχνευσης καρκίνου των ωοθηκών και τα αποτελέσματα τους. Το τέταρτο κεφάλαιο αφορά το πρακτικό μέρος της εργασίας, όπου αναλύονται τα δεδομένα που χρησιμοποιήθηκαν κατά την πειραματική διαδικασία, καθώς και οι τεχνικές προ-επεξεργασίας και ταξινόμησης που εφαρμόστηκαν. Στη συνέχεια παρατίθενται με κατάλληλους πίνακες και διαγράμματα τα αποτελέσματα της εργασίας. Συγκρίνονται μεταξύ τους οι τεχνικές προ-επεξεργασίας και οι αλγόριθμοι ευθυγράμμισης κορυφών, οι διαφορετικοί τύποι δεδομένων (φάσματα υψηλής και χαμηλής ανάλυσης), οι μέθοδοι εξαγωγής χαρακτηριστικών, ταξινόμησης και εκτίμησης των ταξινομητών και οι χρόνοι επεξεργασίας. Επίσης αναδεικνύονται τα χαρακτηριστικά (διαστήματα μάζας/φορτίου m/z) που επιλέγονται από τα καλύτερα μοντέλα ταξινόμησης. Στο τελευταίο κεφάλαιο συνοψίζονται τα συμπεράσματα της έρευνας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: ταξινόμηση πρωτεωμικών σημάτων καρκίνου των ωοθηκών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: αναγνώριση προτύπων, φασματοσκοπία μάζας, καρκίνος ωοθηκών

ABSTRACT

The purpose of this project is the identification of a group of proteins that are represented in the mass spectrum, which achieves through the pattern recognition system, the maximum discriminant ability between cancer and normal samples. The final aim of this procedure is to highlight the mass/charge regions that need to be further identified with techniques which analyze the selected molecules in order to identify them and to find out if they are relevant with potential biomarkers of ovarian cancer. The current work is consisted of a stage of preprocessing of mass spectra, the development of a pattern recognition system where several classifiers and combinations of them will be tested in order find the optimum classification and finally the comparison of the results with the findings of other researchers in order to make useful conclusions relevant with ovarian cancer.

The project focuses on the area of early and accurate detection of ovarian cancer by analyzing in theory and in practice the steps that need to be done for the development of the PR system. The output of the various classifiers is really encouraging giving accuracy up to 99%, while emphasis is also given in the potential biomarkers that the algorithms conclude on.

The skeleton of this work is as followed: In the first chapter there is a general introduction about ovarian cancer and the methods of its detection. In the second chapter the methods of preprocessing of the mass spectra are analyzed, while at the third chapter the existing methodologies of pattern recognition, the classifier models, feature selection methods, the estimation alternatives and the results' interpretation are explained. At the end of the theoretical part the existing researches around this area and their corresponding results are cited. The fourth chapter refers to the practical part of this project, where the data sets, the methods of preprocessing and classification that were used are analyzed. Then, there are tables and figures with the results of this work. The techniques of preprocessing, the peak alignment algorithms, the different data sets (low and high resolution), the feature extraction, classification and evaluation methods and the computational times are compared to each other. Also, the "best" features (mass/charge regions) that are selected from the classification models are emphasized. In the last chapter there are the conclusions of the current research.

SUBJECT AREA: classification of proteomic signals of ovarian cancer

KEYWORDS: pattern recognition, mass spectrometry, ovarian cancer