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

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

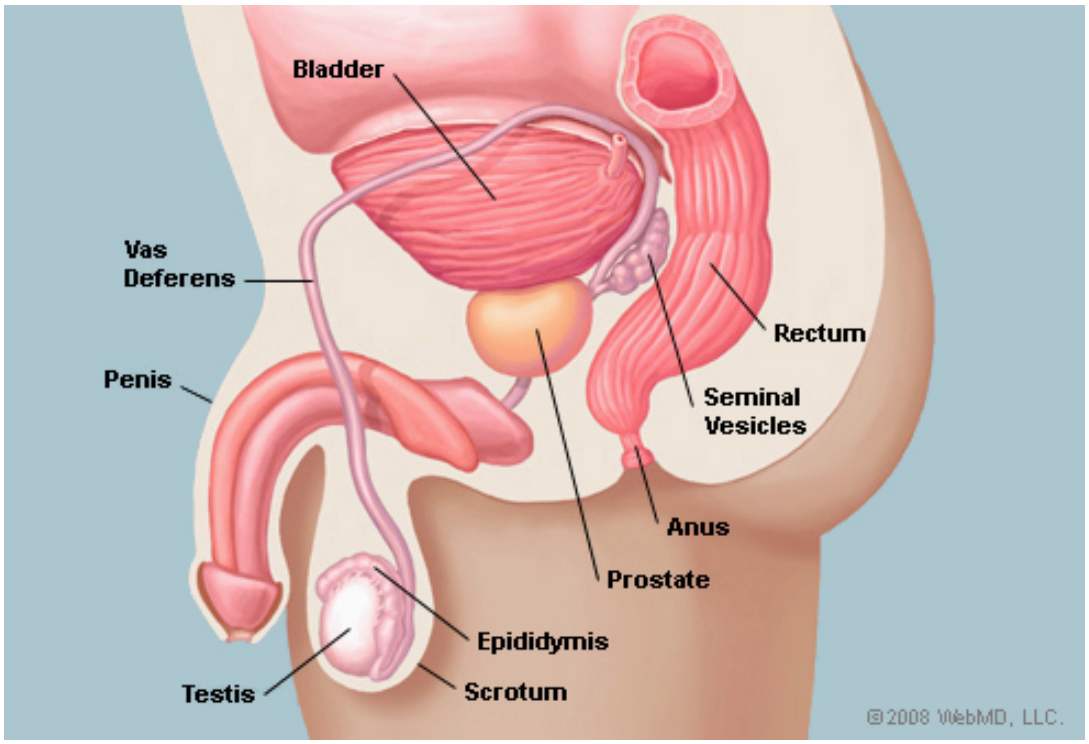
ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο Εργαστήριο επεξεργασίας ιατρικού σήματος και εικόνας του Τμήματος Τεχνολογίας Ιατρικών Οργάνων του ΤΕΙ Αθήνας υπό την επίβλεψη του Καθηγητή κ. Διονύσιου Κάβουρα. Μέρος της εργασίας κοινοποιήθηκε στο διεθνές συνέδριο, 4th International Conference on Experiments / Process / System Modeling / Simulation / Optimization (4th IC-EpsMso), Athens, Greece, 6-9 July 2011. Η Υλοποίηση του πηγαίου κώδικα που απαιτήθηκε για τις ανάγκες της εργασίας πραγματοποιήθηκε στη γλώσσα προγραμματισμού MATLAB (έκδοση 7.7.0 (R2008b) -Mathworks). Θα ήθελα στο σημείο αυτό να ευχαριστήσω τον κ.Κάβουρα για την εμπιστοσύνη της ανάθεσης του θέματος και την πολύπλευρη υποστήριξη του. Σημαντικός αρωγός τόσο στην σχεδίαση και υλοποίηση της συγκεκριμένης εργασίας όσο και στην εύρεση σετ δεδομένων και στην αξιολόγηση των αποτελεσμάτων ήταν ο Δρ. Σπύρος Κωστόπουλος.

1. ΕΙΣΑΓΩΓΗ

1.1 Προστάτης

Ο προστάτης είναι ένας μικρός εσωτερικός αδένας, στο μέγεθος ενός καρυδιού, ο οποίος βρίσκεται ακριβώς κάτω από την ουροδόχο κύστη του άντρα και περιβάλει την ουρήθρα. Ο αδένας αυτός είναι μέρος του ανδρικού αναπαραγωγικού συστήματος και λόγο της ανατομικής του θέσης (Εικόνα 1.1) βοηθάει τόσο στον έλεγχο της ούρησης όσο και στον έλεγχο της εκσπερμάτισης. Από την μια δηλαδή παίζει ένα ρόλο σφικτήρα κατά την ούρηση και από την άλλη εμπλουτίζει το σπέρμα με χρήσιμα και απαραίτητα συστατικά για την ενδυνάμωση και την προστασία του. Ο προστάτης μεγαλώνει κατά την διάρκεια της ζωής του άνδρα και κυρίως μετά την ηλικία των 40 ετών κάτω από την επίδραση της τεστοστερόνης.

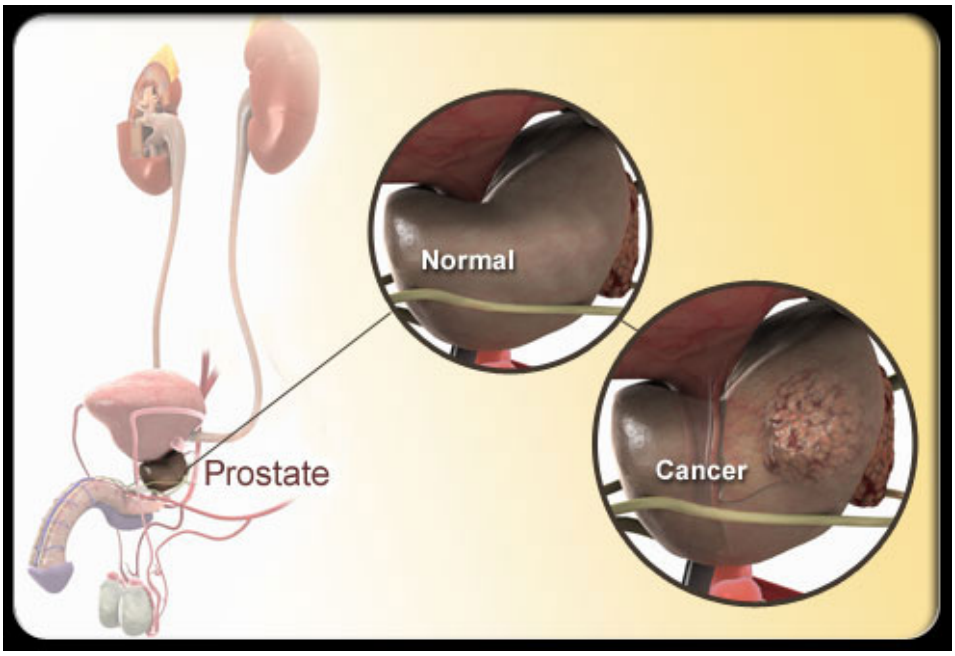


Εικόνα 1. 1: Ανατομική θέση προστάτη [1].

1.2 Καρκίνος του Προστάτη

Σε κάθε άνθρωπο υπάρχει μια φυσιολογική ανακύκλωση των κυττάρων κατά την οποία νέα κύτταρα αναπληρώνουν τα γερασμένα. Στην περίπτωση του καρκίνου, η ισορροπία κατά την διάρκεια αυτής της ανακύκλωσης χαλάει με αποτέλεσμα να παράγονται νέα κύτταρα σε πολύ μεγαλύτερο βαθμό από αυτόν του θανάτου των παλαιών. Επομένως, καρκίνο του προστάτη ονομάζουμε την δημιουργία τέτοιων κακοηθών κυττάρων (και κατά συνέπεια όγκων) τα οποία πολλαπλασιάζονται πολύ γρήγορα και επιπλέον έχουν την δυνατότητα μετάστασης δηλαδή μεταφοράς τους σε άλλο σημείο του σώματος. Ο καρκίνος του προστάτη μπορεί να αναπτυχθεί γρήγορα αλλά τις περισσότερες φορές αναπτύσσεται πολύ αργά. Σύμφωνα με μερικά επίσημα στατιστικά δεδομένα σχετικά με τον καρκίνο του προστάτη, το 2010 είχαμε 217.730 νέες περιπτώσεις και 32.045 θανάτους στην Ηνωμένες Πολιτείες Αμερικής [2], το 2008 είχαμε 37.051 νέες περιπτώσεις και 10.168 θανάτους στην Ηνωμένο Βασίλειο [3] ενώ το 2006 είχαμε 345.900 νέες περιπτώσεις και 87.400 θανάτους στην Ευρώπη [4]. Η πρόγνωση για τις

Ηνωμένες Πολιτείες Αμερικής μιλάει για 240.890 νέες περιπτώσεις καρκίνου του προστάτη και 33.720 θανάτους από αυτόν [2]. Η μορφή των όγκων που δημιουργούνται κατά των καρκίνο του προστάτη φαίνεται στην εικόνα 1.2.



Εικόνα 1. 2 : Καρκίνος του Προστάτη [5].

1.2.1 Αιτίες

Οι αιτίες που προκαλούν καρκίνο του προστάτη, δεν έχουν αναγνωριστεί με απόλυτη βεβαιότητα. Σημαντικό ρόλο στην γέννηση του καρκίνου παίζουν η ηλικία, η εθνική προέλευση αλλά και το οικογενειακό ιστορικό. Η αύξηση της ηλικίας αποτελεί βασικό παράγοντα εμφάνισης καρκίνου του προστάτη καθώς είναι εξαιρετικά σπάνια η εμφάνιση του κάτω από την ηλικία των 40 χρόνων αλλά η πιθανότητες εμφάνισης αυξάνουν ραγδαία πάνω από τα 50 χρόνια και μάλιστα παρατηρείται ότι οι 2 στους 3 καρκίνους του προστάτη που θα διαγνωστούν είναι σε ηλικίες πάνω των 65 χρόνων. Πρώτου βαθμού συγγενείς ασθενών με καρκίνο του προστάτη έχουν αυξημένο κίνδυνο να παρουσιάσουν την ασθένεια και οι ίδιοι ενώ η ασθένεια αυτή είναι πιο συχνή στους άνδρες του Δυτικού κόσμου παρά σε αυτούς στην Ασία όπου τα ποσοστά εμφάνισης παραμένουν χαμηλά. Συσχέτιση της ασθένειας υπάρχει ακόμα με ανθρώπους που επιβαρύνονται περισσότερο από την μόλυνση της ατμόσφαιρας και με αυτούς που κάνουν κακή διατροφή με μεγάλες ποσότητες λιπαρών ουσιών. Επίσης σημαντικό ρόλο στην ανάπτυξη του καρκίνου του προστάτη φέρεται να παίζουν οι ανδρικές σεξουαλικές ορμόνες και τα ανδρογόνα.

1.2.2 Συμπτώματα

Στα πρώτα στάδια του καρκίνου του προστάτη συχνά τα συμπτώματα δεν είναι εμφανή. Όσο ο όγκος μεγαλώνει, πιέζει την ουρήθρα με αποτέλεσμα να έχουμε διαταραχές στην ούρηση αλλά και στην εκσπερμάτιση ή την σεξουαλική δραστηριότητα.

Όσον αφορά την ούρηση,

- επιτακτική και συχνή ανάγκη ούρησης, κυρίως την νύχτα,
- δυσκολία και πόνος στην έναρξη και τον τερματισμό της ούρησης,
- επώδυνη με αίσθηση καψίματος ούρηση,
- ασθενή ή διακεκομμένη ροή ούρων,
- αίμα στα ούρα

όσον αφορά την εκσπερμάτιση ή την σεξουαλική δραστηριότητα,

- αίμα στο σπέρμα
- πόνος κατά την εκσπερμάτιση
- ελάττωση της ποιότητας του σπέρματος
- ελάττωση ή απώλεια της σεξουαλικής επιθυμίας αλλά και
- ελάττωση ή απώλεια της στυτικής ικανότητας.

Επίσης συχνά εμφανίζεται και πόνος στην μέση χαμηλά ή και στην σπονδυλική στήλη, πράγμα που δείχνει προχωρημένη μορφή καρκίνου με πιθανές μεταστάσεις. Κάποια από τα παραπάνω είναι συμπτώματα και της φυσιολογικής μεγέθυνσης του αδένα και όχι μόνο του καρκίνου.

1.2.3 Τεχνικές διάγνωσης

Δυστυχώς, ίσως και λόγο της άγνοιας του τι προκαλεί το καρκίνο του προστάτη, δεν υπάρχουν καθορισμένες προ-συμπτωματικές εξετάσεις που να μπορούν να διαγνώσουν με ακρίβεια το καρκίνου του προστάτη και τον εντοπισμό του κυρίως όταν αυτός βρίσκεται σε πολύ πρώιμα στάδια. Παρόλα αυτά υπάρχουν ορισμένες επιβοηθητικές εξετάσεις για την πρώιμη ανεύρεση του καρκίνου που πιθανόν να βοηθούν στο ποσοστό θνησιμότητας από την ασθένεια καθώς, όπως και σε κάθε άλλη μορφή καρκίνου έτσι και στο καρκίνο του προστάτη, η έγκαιρη και γρήγορη εύρεση της ασθένειας στα πρώιμα στάδια την καθιστά πιο εύκολη στην αντιμετώπιση. Οι βασικές επιβοηθητικές τεχνικές διάγνωσης είναι η δαχτυλική εξέταση ορθού – Digital Rectal Examination (DRE) και η μέτρηση του ειδικού προστατικού αντιγόνου – Prostate Specific Antigen (PSA) στο αίμα ενώ άλλες τεχνικές είναι η εξέταση των ούρων, η βιοψία (συνήθως ακολουθεί την διάγνωση προβλήματος), το διορθικό υπερηχογράφημα – TransRectal Ultrasound (TRUS), η εξέταση ισοτόπων στα κόκαλα – isotope bone scan, οι ακτίνες X και οι εικόνες MRI – magnetic reasonance imaging και CT-computerized tomography.

<u>DRE</u>

Η δαχτυλική εξέταση ορθού γίνεται από τον Ουρολόγο, ο οποίος προσπαθεί μέσω του ορθού να ψηλαφήσει τον προστάτη έτσι ώστε να εντοπίσει ανωμαλίες ή παθολογικά σκληρές περιοχές οι οποίες υποδεικνύουν την ύπαρξη προβλήματος και πιθανόν καρκίνο. Ο προστάτης βρίσκεται ακριβώς μπροστά από το τελικό τμήμα του ορθού και έτσι μπορεί με ευκολία να ψηλαφηθεί η οπίσθια επιφάνεια του, επιφάνεια στην οποία ξεκινούν οι περισσότεροι καρκίνοι του προστάτη. Εάν όμως το καρκίνωμα βρίσκεται στην μπροστά επιφάνεια του προστάτη η εξέταση αυτή δεν θα δείξει τίποτα και έτσι είναι αναγκαίες άλλες εξετάσεις.

PSA test

Η ποσότητα του ειδικού προστατικού αντιγόνου – PSA στο αίμα είναι από τις πιο χρήσιμες εξετάσεις για την έγκαιρη ανίχνευση του καρκίνου του προστάτη. Το PSA είναι μια γλυκοπρωτείνη που παράγεται από τα επιθηλιακά κύτταρα του προστατικού αδένα. Εντοπίζεται στα κύτταρα του προστάτη και γενικά βρίσκεται σε χαμηλές συγκεντρώσεις. Αλλοιώσεις όμως της φυσικής δομής του προστάτη, όπως μόλυνση, τραυματισμός, προστατίτιδα και καρκίνος δημιουργούν μεγαλύτερη συγκέντρωση PSA με αποτέλεσμα αυτό να μπαίνει στην κυκλοφορία του αίματος και να εντοπίζεται από το ειδικές εξετάσεις αίματος όπως το PSA test. Το μεγαλύτερο μέρος του PSA εγχέεται στο σπερματικό υγρό και βοηθάει στην ρευστοποίηση στου σπέρματος και άρα στην κινητικότητα των σπερματοζωαρίων και στην ικανότητά τους για γονιμοποίηση του ωαρίου. Σε κανονικές συνθήκες το PSA εντοπίζεται σε συγκεντρώσεις μικρότερες του 4 ng/ml πράγμα όμως που επηρεάζεται από διάφορους παράγοντες όπως η ηλικία και καταγωγή. Στο παρακάτω πίνακα βλέπουμε την διακύμανση της συγκέντρωσης του PSA ανάλογα με ηλικία και καταγωγή.

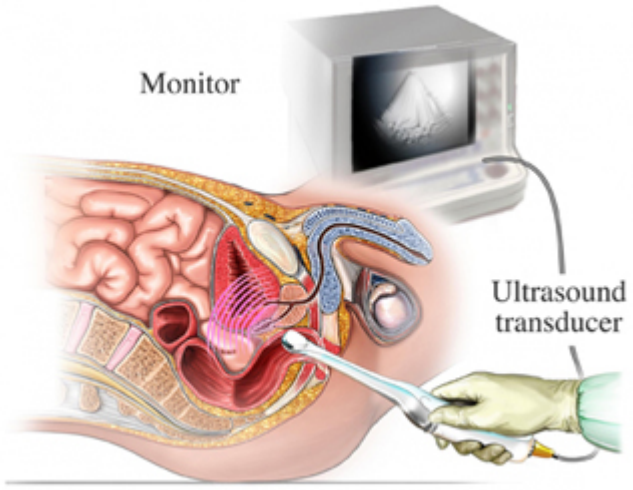
Πίνακας 1.1 : Συγκεντρώσεις PSA ανάλογα με ηλικία και εθνικότητα [4].

Ηλικία	**Asian -Americans**	**African - Americans**	**Whites**
40-49 yr	0-2.0 ng/mL	0-2.0 ng/mL	0-2.5 ng/mL
50-59 yr	0-3.0 ng/mL	0-4.0 ng/mL	0-3.5 ng/mL
60-69 yr	0-4.0 ng/mL	0-4.5 ng/mL	0-4.5 ng/mL
70-79 yr	0-5.0 ng/mL	0-5.5 ng/mL	0-6.5 ng/mL

Παρόλο την ευρεία χρήση, το PSA test δεν είναι τελείως αξιόπιστο [6-7]. Για κάθε 100 άντρες με υψηλό βαθμό συγκέντρωσης PSA μόνο γύρω στους 30 θα βρεθούν με καρκίνο μετά την βιοψία. Επίσης περίπου 15% των ανδρών με φυσιολογικά επίπεδα PSA στην πραγματικότητα έχουν την ασθένεια. Άρα ούτε το PSA test είναι αρκετό για να μάθουμε για την παρουσία ή όχι της ασθένειας.

TRUS

Σε αυτήν την εξέταση μια λεπτή κεφαλή υπερήχων εισέρχεται στον ορθό, τα ηχητικά κύματα που εκπέμπονται προσκρούουν στους διάφορους ιστούς του προστάτη και αντανακλούνται ενώ τις αντανακλάσεις αυτές τις μετατρέπουμε σε εικόνα μέσω ενός ηλεκτρονικού υπολογιστή. Η μέθοδος αυτή χρησιμοποιείται κυρίως ως οδηγός για μια επόμενη βιοψία. Στις κάτωθεν εικόνες βλέπουμε αρχική την διαδικασία του διορθικού υπερηχογραφήματος και στην συνέχεια την εικόνα-αποτέλεσμα της διαδικασίας.

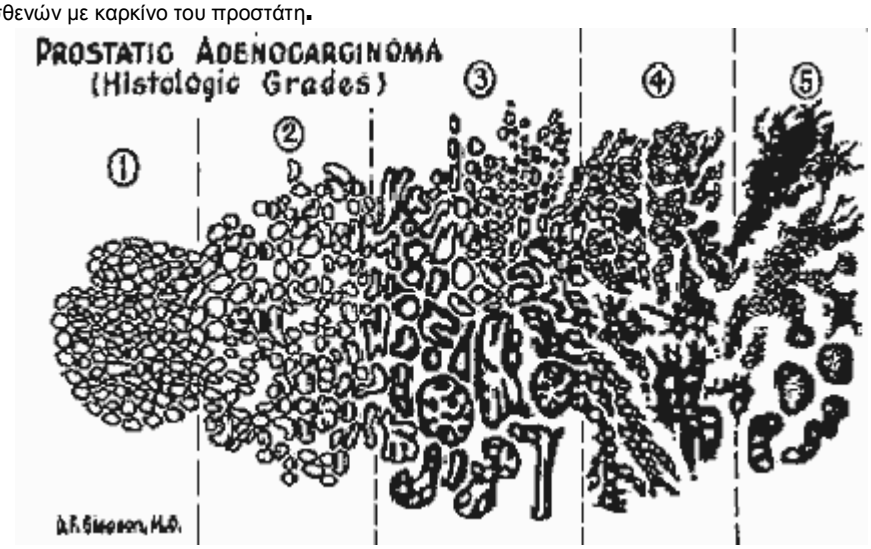


Εικόνα 1. 3 : Διορθικό υπερηχογράφημα [8].

Εικόνα 1. 4 : Εικόνα-Αποτέλεσμα διορθικού υπερηχογραφήματος [9].

Βιοψία και Gleason Grading

Βιοψία ονομάζουμε την αφαίρεση ενός δείγματος από τον ιστό, συγκεκριμένα τον προστάτη, το οποίο στην οποία εξετάζεται με το ηλεκτρονικό μικροσκόπιο για την παρουσία καρκίνου. Η βιοψία, η οποία συνήθως γίνεται υπό τοπική αναισθησία, ακολουθεί την μέθοδο TRUS και βασίζεται σε αυτήν για την καθοδήγηση της βελόνας που θα αντλήσει το δείγμα. Έχοντας τώρα τον ιστό, χρειάστηκε ένα σύστημα βαθμονόμησης αυτό σχετικά με το πόσο απέχει από έναν υγιή ιστό. Αυτό ακριβώς προσφέρει το Gleason grading system που πήρε τα αρχικά του από τον Donald F. Gleason ο οποίος το περιέγραψε αρχικά. Υπάρχουν 5 βαθμοί στο σύστημα Gleason 1 έως 5, ανάλογα με το αν είναι υγιή ή μακριά από το υγιές δείγμα αντίστοιχα. Στην κάτωθεν εικόνα φαίνεται μια απλοϊκή μορφή της βαθμονόμησης του Gleason, όπως αυτός της σχεδίασε.



Εικόνα 1. 5 : Gleason grading system [10].

Σε έναν ιστό υπολογίζονται οι δύο μεγαλύτερες περιοχές με διαφορετικό βαθμό Gleason και οι 2 αυτοί βαθμοί αθροίζονται για να κάνουν το άθροισμα Gleason. Κοινό άθροισμα σε 2 διαφορετικούς ιστούς δεν σημαίνει απαραίτητα κοινή θεραπεία και μέθοδος αντιμετώπισης καθώς το από ποιους βαθμούς προήλθε αυτό το άθροισμα παίζει κομβική σημασία. Το άθροισμα Gleason 3+4 με το 4+3 είναι διαφορετικά και χρήζουν διαφορετικής αντιμετώπισης καθώς το μεν πρώτο αποτελείται κυρίως από έναν μέτρια επιθετικά καρκίνο και σε μικρότερο βαθμό από έναν επιθετικής μορφής ενώ το δεύτερο άθροισμα αποτελείται ως επί το πλείστον από καρκίνο πιο επιθετικής μορφής.

1.2.4 Θεραπεία

Ο καρκίνος του προστάτη διακρίνεται σε 4 στάδια, όσο πιο μεγάλο το στάδιο τόσο πιο προχωρημένος είναι ο καρκίνος. Στο $1^{ο}$ στάδιο ο καρκίνος δεν εμφανίζει συμπτώματα και επιπλέον είναι τόσο μικρός που δεν μπορεί να εντοπιστεί μέσω του DRE. Συνήθως εντοπίζεται κατά λάθος, λόγο χειρουργικής επέμβασης για άλλες αιτίες. Στο $2^{ο}$ στάδιο ο καρκίνος ακόμα βρίσκεται μόνο στο προστάτη. Μπορεί να εντοπιστεί με DRE αλλά και πάλι μπορεί να μην εμφανίζει συμπτώματα. Στο $3^{ο}$ στάδιο τα καρκινικά κύτταρα έχουν διαδοθεί σε περιφερειακούς του προστάτη ιστούς. Σύνηθες φαινόμενο είναι η εμφάνιση δυσκολίας κατά την ούρηση. Στο $4^{ο}$ στάδιο ο καρκίνος έχει εμφανίσει αρκετές μεταστάσεις στα κόκαλα, στους πνεύμονες, στο συκώτι και σε άλλους ιστούς μακριά από τον προστάτη. Ο ασθενής παρουσιάζει πόνους στην μέση και στα κόκαλα, απώλεια βάρους και έντονη κούραση.

Η θεραπεία του καρκίνου του προστάτη εξαρτάται αρκετά από το στάδιο στο οποίο βρίσκεται αλλά και από την ηλικία και την γενικότερη υγεία του ασθενή. Ο καρκίνος του προστάτη πολύ συχνά αναπτύσσεται πάρα πολύ αργά και παίρνει ίσως και χρόνια για να αναπτυχθεί, συνεπώς όταν αναφερόμαστε σε ηλικιωμένους άντρες πιθανόν και να μην χρειάζεται η επιθετική και άμεση θεραπεία. Σε τέτοιες περιπτώσεις πιθανόν ο γιατρός να μείνει σε τακτική παρακολούθηση του ασθενούς ώστε να μπορεί να ελέγχει την εξέλιξη του όγκου. Στην περίπτωση νεαρότερων ανδρών εάν ο καρκίνος εντοπιστεί ενώ βρίσκεται ακόμα μόνο στον προστατικό αδένα, μια ολική αφαίρεση του προστάτη (ριζική προστατεκτομή) είναι πιθανόν να λύσει το πρόβλημα ενώ εάν ο καρκίνος εξαπλωθεί σε σημεία εκτός του προστάτη είναι αρκετά δύσκολο έως και ακατόρθωτο να αντιμετωπιστεί και να υπάρξει θεραπεία. Εκτός από την προστατεκτομή, άλλες μορφές θεραπείας μπορεί να είναι η θεραπεία μέσω ακτίνων Χ με σκοπό να σκοτώσει τα καρκινικά κύτταρα, η ορμονοθεραπεία με σκοπό να σταματήσει την αύξηση των καρκινικών κυττάρων, η χημειοθεραπεία που είναι συνδυασμός για προχωρημένα

στάδια του καρκίνου (με ήδη υπάρχουσες μεταστάσεις) αλλά και συνδυασμός των παραπάνω.

1.3 Δομή εργασίας

Η παρούσα διπλωματική εργασία αποτελείται από έξι κεφάλαια. Το κεφάλαιο ένα αποτελείται από μία εισαγωγή στις έννοιες του προστάτη καθώς και του καρκίνου του προστάτη. Το κεφάλαιο δύο περιέχει μια θεωρητική προσέγγιση περιοχών της πρωτεωμικής. Μας εισάγει στους τρόπους διαχωρισμού των πρωτεϊνών αλλά και στο πως μπορούμε να βρούμε την ταυτότητά τους. Το κεφάλαιο τρία μας περιγράφει θεωρητικά πως γίνεται ανάλυση και ταξινόμηση ενός MS φάσματος ενώ στο κεφάλαιο τέσσερα η θεωρία συναντά για πρώτη φορά την πράξη καθώς εκτός της επισκόπησης των ερευνών διάφορων ομάδων αναφέρονται τα βήματα και η στρατηγική που ακολουθήθηκε στην παρούσα εργασία. Στην συνέχεια, στο κεφάλαιο 5 υπάρχει το πειραματικό μέρος στο οποίο το σύστημα που υλοποιήθηκε δοκιμάζεται υπό διάφορες προϋποθέσεις. Κλείνοντας, το κεφάλαιο 6 αναφέρει συμπεράσματα επί των αποτελεσμάτων καθώς και πιθανές βελτιώσεις του συστήματος στο προσεχές διάστημα.

2. ΠΡΩΤΕΩΜΙΚΗ

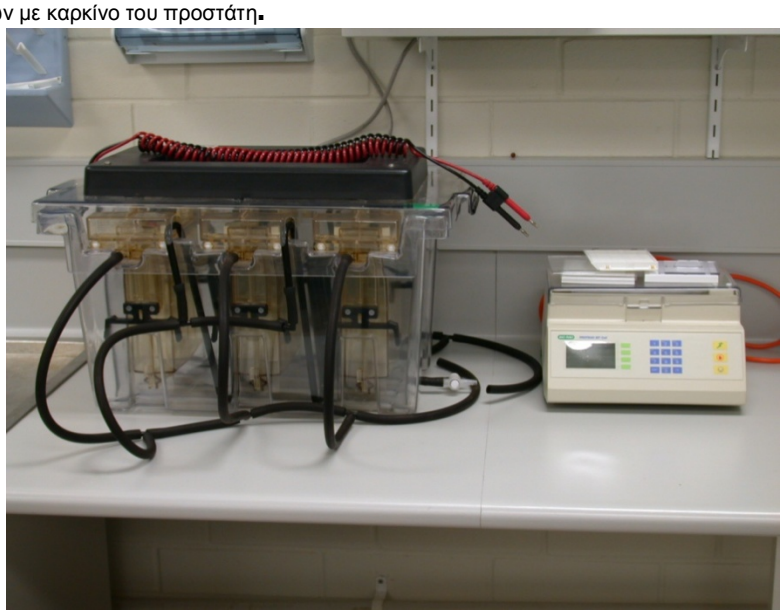
Πρωτέωμα ενός οργανισμού ορίζεται το πλήρες σύνολο των εκφρασμένων πρωτεϊνών και των αντίστοιχων μετά-μεταφραστικών καταστάσεων τους, σε μια δεδομένη χρονική στιγμή [11]. Η μελέτη του πρωτεώματος και η σύγκρισή του μεταξύ οργανισμών διαφορετικών παθολογικών καταστάσεων είναι η τρόπος που η διαφορική πρωτεωμική (differential proteomics) προσπαθεί να βρει διαγνωστικούς βιοδείκτες για διάφορες ασθένειες, όπως για παράδειγμα ο καρκίνος του προστάτη. Συνεπώς η εφαρμογή της πρωτεωμικής προσφέρει τεράστιες δυνατότητες για την κατανόηση των ασθενειών και την εύρεση θεραπευτικών στόχων για αυτές. Η πρωτεϊνική ανάλυση διακρίνεται σε 2 στάδια: τον διαχωρισμό των πρωτεϊνών και την αναγνώριση-ταυτοποίηση των πρωτεϊνών μέσω φασματομετρίας μάζας.

2.1 Τεχνικές διαχωρισμού πρωτεϊνών

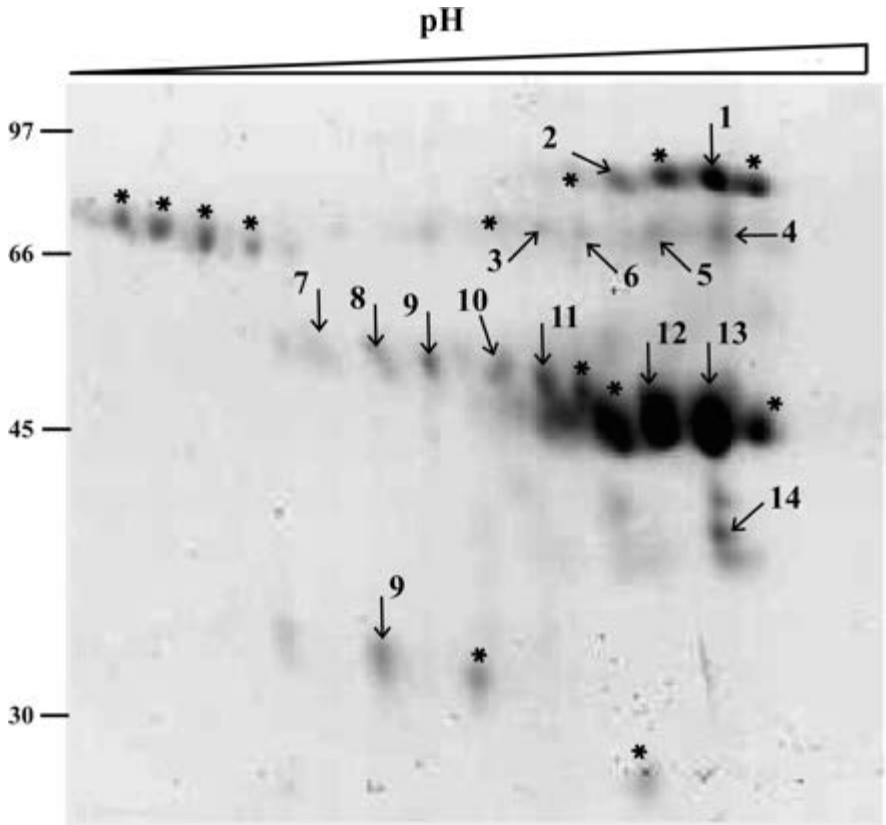
Κύριας σημασίας για την αποτελεσματική επιτέλεση της πρωτεϊνικής ανάλυσης είναι βεβαίως η σωστή συλλογή του δείγματος από ιστούς, κύτταρα, σωματικά υγρά και η φύλαξη του με τέτοιο τρόπο που να μην αλλοιώνει το δείγμα. Στην συνέχεια βρίσκεται το στάδιο του διαχωρισμού των πρωτεϊνών με βασικές τεχνικές διαχωρισμού την δισδιάστατη ηλεκτροφόρηση πηκτωμάτων – 2D gel electrophoresis (2DGE) και την υγρή χρωματογραφία – Liquid Chromatography LC.

2.1.1 Δισδιάστατη ηλεκτροφόρηση

Μέσω της 2DGE, σύνθετα μείγματα πρωτεϊνών διαχωρίζονται με βάση 2 ιδιότητες στις 2 διαστάσεις του πηκτώματος. Αρχικά γίνεται ένας πρώτος διαχωρισμός σε μία διάσταση βάση του ισοηλεκτρικού τους σημείου έτσι ώστε οι πρωτεΐνες να διαταχθούν κατά μήκος μίας γραμμής. Σε αυτό το στάδιο οι πρωτεΐνες κατανέμονται πάνω σε ένα μέσο που έχει διαβάθμιση του pH και ηλεκτρικό ρεύμα διαπερνά αυτό το μέσο δημιουργώντας δύο πόλους, έναν θετικό και έναν αρνητικό. Οι αρνητικά φορτισμένες πρωτεΐνες κινούνται, μέσω του διαλύματος pH, προς τον θετικό πόλο, οι θετικά φορτισμένες πρωτεΐνες κινούνται προς τον αρνητικό πόλο έως ότου φτάσουν στο ισοηλεκτρικό τους σημείο στο οποίο το συνολικό φορτίο τους είναι μηδέν. Στο δεύτερο στάδιο, ξαναδιαχωρίζονται οι πρωτεΐνες με βάση το μοριακό τους βάρος αυτήν την φορά. Τώρα πραγματοποιείται ηλεκτροφόρηση σε πηκτώματα πολυακρυλαμίδης, μέσω στο οποίο οι πρωτεΐνες κινούνται με διαφορετικό ρυθμό ανάλογα με το μοριακό τους βάρος. Αυτά συμβαίνει καθώς το ηλεκτρικό πεδίο εφαρμόζει ίδια δύναμη στις πρωτεΐνες και έτσι αυτές που θα κινηθούν γρηγορότερα είναι αυτές που είναι μικρότερες σε μέγεθος. Στις εικόνες που ακολουθούν, βλέπουμε έναν μηχανισμό εφαρμογής ηλεκτρικού πεδίου σε ένα πήκτωμα στην εικόνα 2.1 και την τελική μορφή ενός 2DGE στην εικόνα 2.2. Στην εικόνα 2.2 φαίνεται επίσης ο άξονας της $1^{ης}$ διάστασης (pH) και ο άξονας της $2^{ης}$ διάστασης (μοριακό βάρος).



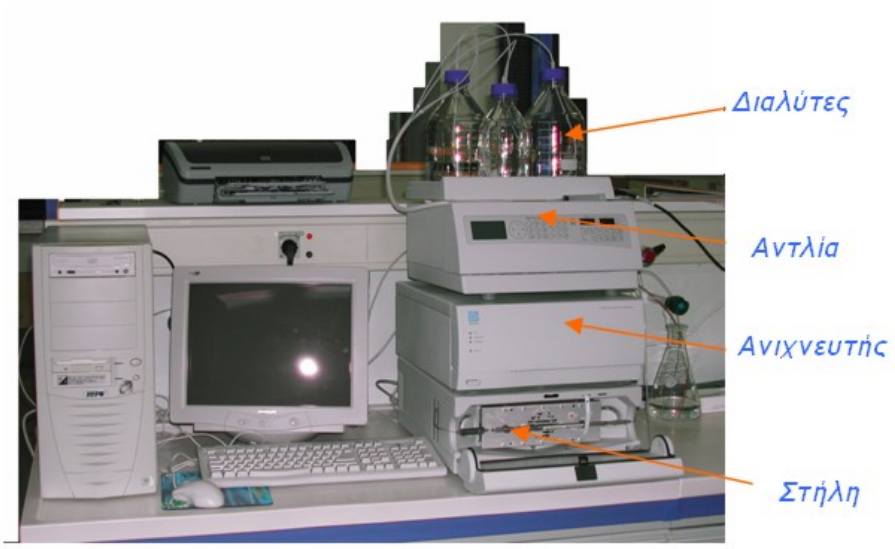
Εικόνα 2. 1 : Μηχανισμός εφαρμογής ηλεκτρικού πεδίου σε πρωτεΐνες [12].



Εικόνα 2. 2 : Τελική μορφή ενός 2DGE [13].

2.1.2 Υγρή χρωματογραφία

Η υγρή χρωματογραφία – Liquid Chromatography (LC) είναι μια άλλη τεχνική διαχωρισμού πρωτεϊνών με πλεονέκτημά της την ικανότητα εντοπισμού πρωτεϊνών που βρίσκονται σε χαμηλή συγκέντρωση στο διάλυμα. Διαχωρίζεται σε κινητή και σταθερή φάση. Στην κινητή φάση το δείγμα μας ανακατεύεται με έναν διαλύτη ο οποίος επιλέγεται ανάλογα με την πολικότητά του, μιας και θέλουμε διαφορετική πολικότητα σε κινητή και σταθερή φάση. Στην συνέχεια έχουμε την σταθερή φάση, όπου είναι το κύριο μέσο διαχωρισμού και βρίσκεται εντός μιας στήλης χρωματογραφίας. Οι 2 φάσεις αλληλεπιδρούν ανάλογα με τις χημικές τους ιδιότητες και έτσι κατορθώνεται ο διαχωρισμός του δείγματος στις πρωτεΐνες. Στην εικόνα 2.3 που ακολουθεί βλέπουμε ένα ολοκληρωμένο σύστημα χρωματογραφίας υψηλής απόδοσης.



Εικόνα 2. 3 : Σύστημα χρωματογραφίας υψηλής απόδοσης [14].

2.2 Ταυτοποίηση πρωτεϊνών

Από την στιγμή που θα ολοκληρωθεί ο διαχωρισμός των πρωτεϊνών έπεται η αναγνώριση-ταυτοποίηση των πρωτεϊνών με τεχνικές φασματομετρίας μάζας. Η φασματομετρία μάζας είναι μια αναλυτική τεχνική που χρησιμοποιείται για να μετρήσουμε την αναλογία μάζα προς φορτίο (mass to charge ratio- m/z) ιόντων πρωτεϊνών ή πεπτιδίων βασιζόμενη στην κίνησή τους εντός ηλεκτρικού ή μαγνητικού πεδίου. Μόρια από το δείγμα μετατρέπονται σε ιόντα σε αέρια φάση και διαχωρίζονται ανάλογα με τον λόγο m/z. Η τεχνική αυτή επιτελείται από έναν φασματογράφο μάζας ο οποίος αποτελείται από τρία βασικά μέρη : από μια πηγή ιονισμού, έναν αναλυτή μάζας και έναν ανιχνευτή – detector, πράγμα που φαίνεται και στην εικόνα 2.4.

Εικόνα 2. 4 : Βασικά μέρη ενός φασματογράφου μάζας.

Ένας φασματογράφος μάζας πρέπει να επιτελεί συνήθως τις ακόλουθες διαδικασίες :

- Παράγει ιόντα από το δείγμα μέσω της πηγής ιονισμού.
- Διαχωρίζει τα ιόντα ανάλογα με το λόγο m/z στον αναλυτή μάζας.
- Σπάει τα επιλεγμένα ιόντα και αναλύει αυτά τα θραύσματα σε έναν δεύτερο αναλυτή.
- Ανιχνεύει τα ιόντα που προέκυψαν με τον τελευταίο αναλυτή και μετράει την συγκέντρωσή τους μετατρέποντας συνήθως τα ιόντα σε ηλεκτρικό σήμα.
- Επεξεργασία αυτόν τον ηλεκτρικών σημάτων και μετάδοσή τους σε έναν υπολογιστή.

2.2.1 Πηγές ιονισμού

Στην πηγή ιονισμού, τα προς ανάλυση δείγματα ιονίζονται προτού περάσουν στον αναλυτή του φασματογράφου μάζας. Για το λόγο αυτό μια μεγάλη γκάμα μεθόδων ιονισμού έχουν προταθεί και υλοποιηθεί, άλλες πιο ενεργητικές που προκαλούν μεγάλη θραυσματοποίηση του δείγματος και άλλες πιο ήπιας μορφής. Μερικοί τρόποι-μέθοδοι ιονισμού είναι οι εξής :

- *Ιονισμός Ηλεκτρονίου – Electron Ionization EI.*
  Μέθοδος ιονισμού στην οποία ενεργητικά ηλεκτρόνια αλληλεπιδρούν με άτομα-μόρια που βρίσκονται στην αέρια φάση έτσι ώστε να σχηματιστούν ιόντα. Χρησιμοποιείται κυρίως για αέρια ή άλλα πτητικά οργανικά μόρια

- *Χημικός Ιονισμός – Chemical Ionization CI.*
  Δημιουργία ιόντων μέσω συγκρούσεων. Μέθοδος χαμηλότερης ενέργειας από το EI και άρα έχουμε μικρότερη θραυσματοποίηση και απλούστερης μορφής φάσμα.

- *Ιονισμός Πεδίου – Field ionization FI.*
  Μέθοδος κατά την οποία χρησιμοποιούνται πολύ ισχυρά ηλεκτρικά πεδία για την παραγωγή ιόντων από μόρια αέριας φάσης.

- *Εκρόφηση πεδίου - Field Desorption FD.*
  Μέθοδος που εξελίσσει το FI για μη πτητικά μόρια.

- *Εκρόφηση πλάσματος - Plasma Desorption.*
  Μέθοδος κατά την οποία ο ιονισμός του σταθερού δείγματος συμβαίνει με βομβαρδισμό ιόντων ή ουδέτερων ατόμων που έχουν προέλθει από πυρηνική σχάση.

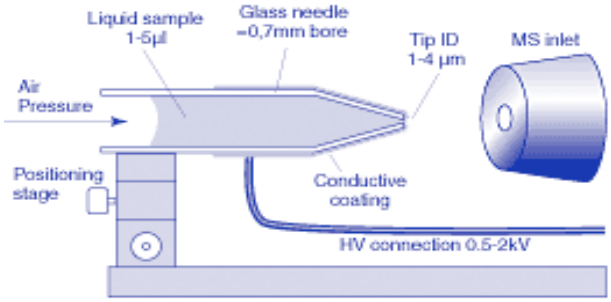
- *Εκρόφηση laser - Laser Desorption.*
  Χρήση παλμών laser για τον ιονισμό.

- *Ιονισμός Θερμικού ψεκασμού - Thermospray*
- *Ιονισμός ατμοσφαιρικής πίεσης - Atmospheric Pressure Ionization.*
- *Χημικός Ιονισμός ατμοσφαιρική πίεσης - Atmospheric Pressure Chemical Ionization.*
- *Φωτοιονισμός – Photoionization.*
- *Ιονισμός με ροή ατόμων μεγάλης ταχύτητας – Fast Atom Bombardment FAB.*
- *Ιονισμός με επιβραδυνόμενη εξαγωγή – Delayed Extraction.*
- *Ιονισμός μέσω πηγής σπινθήρα – Spark source.*
- *Θερμικός ιονισμός – Thermal Ionization.*

Οι δύο πιο συχνά χρησιμοποιούμενοι τρόποι ιονισμού είναι ο ηλεκτροψεκασμός – Electrospray Ionization ESI και ο ιονισμός εκρόφησης υποβοηθούμενος από υλικό μήτρας και laser – Matrix Assisted Laser Desorption Ionization MALDI.

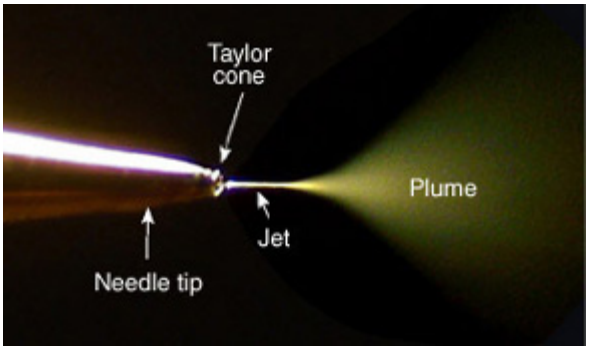
2.2.1.1 Electrospray Ionization

Η μέθοδος ιονισμού με ηλεκτροψεκασμό (Electrospray Ionization, ESI) χρησιμοποιείται συχνά σε σύγχρονα φασματόμετρα LC-MS/MS. Στην βελόνα ιονισμού εφαρμόζεται υψηλό δυναμικό, 3-4 kV, με αποτέλεσμα να σχηματίζονται σταγονίδια με ηλεκτρικά φορτία στην επιφάνειά τους. Η πυκνότητα του φορτίου αυξάνεται σε κρίσιμο σημείο και οι σταγόνες διαιρούνται σε μικρότερες μέχρι να παραχθούν μικροσκοπικά σταγονίδια. Λόγω απωθητικών δυνάμεων εισέρχονται στο φασματογράφο με τη βοήθεια φακών εστίασης και αναλύονται. Χρησιμοποιείται για την ανάλυση πολικών μορίων. Τα κύρια βήματα λοιπόν της διαδικασίας είναι η παραγωγή φορτισμένων σταγόνων στο ακροφύσιο της βελόνας ηλεκτροψεκασμού, η αποδιαλύτωση των φορτισμένων σταγόνων καθώς εξατμίζεται ο διαλύτης κατά τον ψεκασμό και ο σχηματισμός ιόντων στην αέρια φάση. Στην κάτωθεν εικόνα παρουσιάζεται μια μορφή του σχηματισμού ESI.



Εικόνα 2. 5 : Σχέδιο ηλεκτροψεκασμού [15].

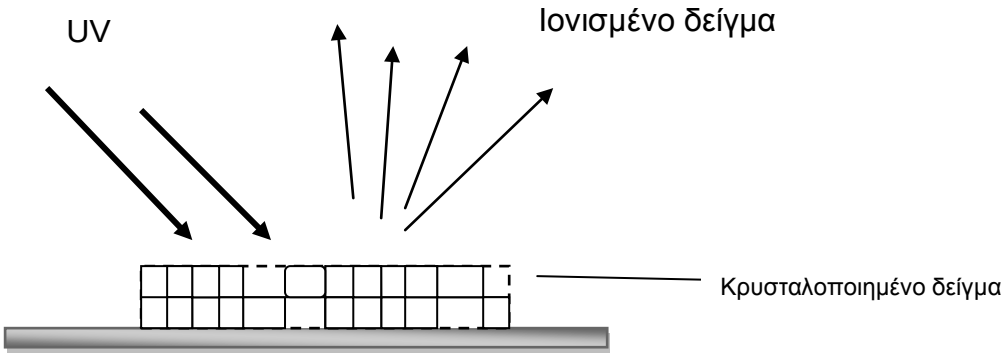
Ο ηλεκτροψεκασμός που πραγματοποιείται στην άκρη της βελόνας αν παρατηρηθεί μέσα από μικροσκόπιο φαίνεται σαν ένα κωνικό σχήμα λόγο της άπωσης των φορτισμένων με το ίδιο φορτίο σταγόνων. Αυτό το κωνικό σχήμα ονομάζεται κώνος του Taylor λόγω του ότι τα φαινόμενα αυτά περιγράφηκαν πρώτη φορά από τον G.I. Taylor το 1964.



Εικόνα 2. 6 : Κώνος του Taylor [15].

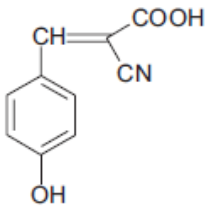
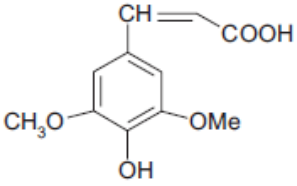
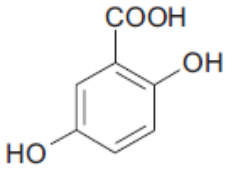
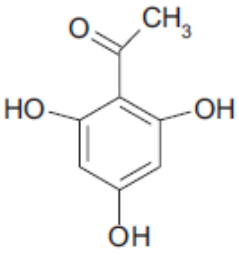
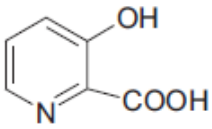
2.2.1.2 Matrix Assisted Laser Desorption Ionization

Ο ιονισμός εκρόφησης υποβοηθούμενος από υλικό μήτρας και laser – Matrix Assisted Laser Desorption Ionization MALDI παρουσιάστηκε για πρώτη φορά το 1988 από τους Karas και HillenKamp [16-18]. Η χρήση της μήτρας που βοηθάει τόσο στον ιονισμό όσο και στην εκρόφηση ήταν κύριος παράγοντας για την επιτυχία αυτής της μεθόδου ιονισμού. Το προς ανάλυση δείγμα ενώνεται με ένα οργανικό μόριο το οποίο απορροφά φως στο μήκος κύματος του laser πάνω στην μήτρα. Η κρυσταλλική αυτή δομή πάνω στην μήτρα βομβαρδίζεται με laser και η ένωση με το δείγμα μας και το οργανικό μόριο ιονίζεται και στην συνέχεια επιταχύνεται εντός του αναλυτή του φασματογράφου μάζας.



Εικόνα 2. 7 : Σχηματική μορφή της διαδικασίας MALDI.

Η ανάμιξη της μήτρας και του πολυμερούς πρέπει να γίνει σε κατάλληλο διαλύτη έτσι ώστε με την απομάκρυνση του διαλύτη να έχουμε ομοιόμορφη διασπορά των μορίων του πολυμερούς. Ο σκοπός της χρησιμοποίησης της μήτρας έχει να κάνει με την απορρόφηση της ενέργειας της ακτίνας laser και επομένως την αποφυγή αποικοδόμησης του πολυμερούς και επίσης την ελαχιστοποίηση των αλληλεπιδράσεων μεταξύ των μορίων πέρα από αυτές δείγματος-μήτρας και άρα μείωση της ενέργειας εξαέρωσης. Ουσιαστικά λοιπόν συμμετέχει στην δημιουργία των ιόντων και δίνεται η δυνατότητα για ανίχνευση συγκεντρώσεων της τάξης των picomoles [19]. Στην συνέχεια δίνονται πιθανές μήτρες, όπως αυτές παρουσιάζονται στο [19].

Matrix	Matrix Structure	Application
α-Cyano-4-hydroxy-cinnamic acid		UV laser: peptide analysis and protein digests. Analytes <10 kDa
Sinapinic acid (4-hydroxy-3,5-dimethoxycinnamic acid)		Analysis of large polypeptides and proteins >10 kDa
2,5-Dihydroxybenzoic acid (2,5 DHB)		UV laser: protein digests and proteins, oligosaccharides released from glycoproteins
2,4,6-Trihydroxyacetophenone (THAP)		UV laser: oligonucleotides <3 kDa
3-Hydroxy picolinic acid		UV laser: Oligonucleotides >3 kDa

Εικόνα 2. 8 : Πιθανά στοιχεία που χρησιμοποιούνται στην μήτρα.

2.2.2 Αναλυτές μαζών

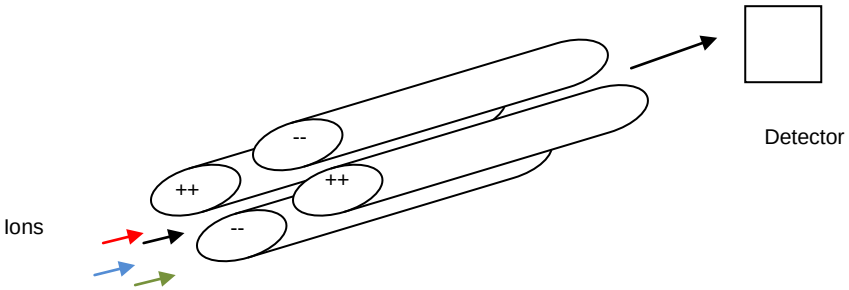
Μετά την παραγωγή των ιόντων θα πρέπει να γίνει διαχωρισμός τους βάση μάζας, η οποία με κάποιο τρόπο πρέπει να υπολογιστεί. Οι αναλυτές μάζας αυτό που ουσιαστικά υπολογίζουν είναι το λόγο m/z, μάζας προς φορτίο, καθώς τα ιόντα είναι δυνατόν να είναι πολλαπλά φορτισμένα και άρα ο λόγος αυτός είναι τμήμα της ολικής τους μάζας. Έχει σχεδιαστεί και υλοποιηθεί μεγάλη γκάμα αναλυτών μάζας, κυρίως συμβατών με τεχνικές ESI και MALDI, οι οποίες και εξελίσσονται διαρκώς παράλληλα με την εξέλιξη των μεθόδων ιονισμού. Κάθε αναλυτής μάζας έχει τα δικά του χαρακτηριστικά και εφαρμογές και παράλληλα τα δικά του πλεονεκτήματα και μειονεκτήματα. Επομένως ανάλογα με την εφαρμογή, γίνεται και η επιλογή του αναλυτή μάζας καθώς δεν υπάρχει

ο ιδανικός αναλυτής και κάθε υλοποίηση. Όλοι οι ευρέως χρησιμοποιούμενοι αναλυτές χρησιμοποιούν ηλεκτρικά και μαγνητικά πεδία για την εφαρμογή δυνάμεων πάνω σε φορτισμένα σωματίδια, σε ιόντα.

Οι τεχνικές ιονισμού ESI και MALDI έδωσαν την δυνατότητα ανάλυσης σε πολύ μεγάλα μόρια πράγμα που οδήγησε σε βελτιωμένους αναλυτές μάζας με καλύτερη διακριτική ικανότητα, ακρίβεια αλλά και μεγάλο εύρος κλίμακας m/z. Επίσης λόγω του ότι οι συγκεκριμένες μέθοδοι ιονισμού μπορεί να παραδώσουν ιόντα μετά από μικρή ή ακόμα ακόμα και καθόλου θραυσματοποίηση πολύ φασματογράφοι μάζας κάνουν χρήση δύο ή παραπάνω αναλυτών στην σειρά (MS/MS, MS/MS/MS) έτσι ώστε να επιτύχουν περαιτέρω ανάλυση και λεπτομέρειες για τα μόρια-ιόντα. Οι σειριακή αυτή χρησιμοποίηση των αναλυτών μάζας βρήκε μεγάλη εφαρμογή κυρίως με τους αναλυτές χρόνου πτήσης και της τετράπολες παγίδες ιόντων. Στην πορεία αναφέρονται βασικοί αναλυτές μάζας.

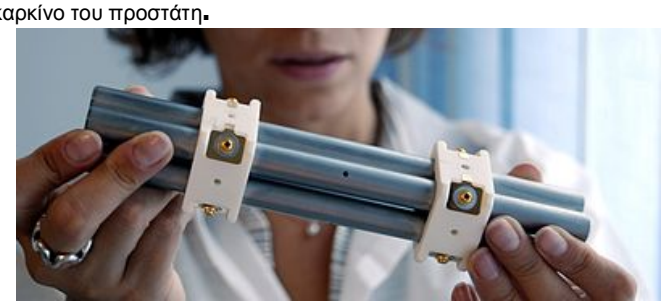
2.2.2.1 Τετράπολοι αναλυτές – Quadrupole analyzers

Η αρχή του τετραπολικού αναλυτή περιγράφηκε από την ομάδα των Paul και Steiwedel [20] το 1953 ενώ πέρασαν αρκετά χρόνια έως την εμπορευματοποίηση τους το 1994 από την ομάδα των Finnigan, Shoulders και Story [21]. Ουσιαστικά ένας τετράπολος αναλυτής είναι ένα φίλτρο μαζών καθότι επιτρέπει μόνο σε έναν λόγο m/z να περάσει μέσα από αυτόν. Αποτελείται από τέσσερα κυλινδρικά ηλεκτρόδια, παράλληλα μεταξύ τους, τα οποία σχηματίζουν ένα σταυρό. Αυτά αναλύονται σε δύο αντίθετα ζεύγη ηλεκτροδίων στα οποία έχει εφαρμοστεί ένα δυναμικό συνεχούς ρεύματος με ίδια απόλυτη τιμή αλλά αντίθετη πολικότητα, όπως φαίνεται και στην κάτωθεν εικόνα.



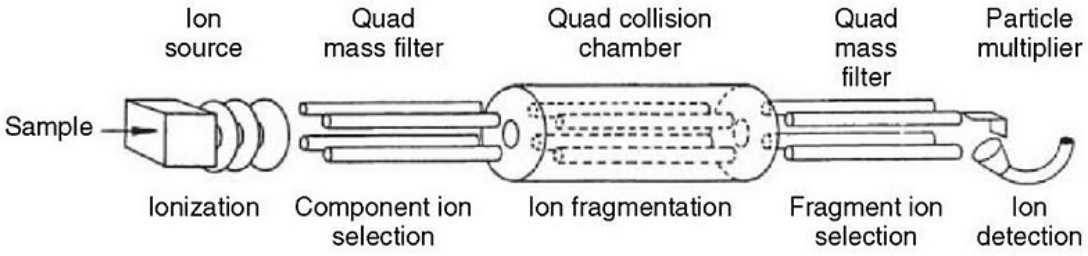
Εικόνα 2. 9 : Σχηματική αναπαράσταση τετράπολου αναλυτή.

Ένα θετικό ιόν θα κινηθεί προς το αρνητικό ηλεκτρόδιο, εάν το δυναμικό αλλάξει πριν το ιόν αποφορτιστεί πάνω στο ηλεκτρόδιο, τότε το ιόν θα αλλάξει κατεύθυνση. Αλλάζοντας λοιπόν την συχνότητα του εναλλασσόμενου δυναμικού πεδίο δίνουμε την δυνατότητα κάθε φορά να περνάει συγκεκριμένο m/z μέσα από το τετράπολο, κατευθυνόμενο προς τον αναλυτή και συνολικά μπορούμε να σαρώσουμε όλο το εύρος των μαζών γνωρίζοντας το τι κάθε φορά περνάει. Τα αποτελέσματα ενός τετράπολου αναλυτή έχουν καλή αναπαραγωγικότητα, είναι συνήθως σε συσκευές μικρές με χαμηλό κόστος αλλά έχουν περιορισμένη ανάλυση.



Εικόνα 2. 10 : Εικόνα Quadrupole [22].

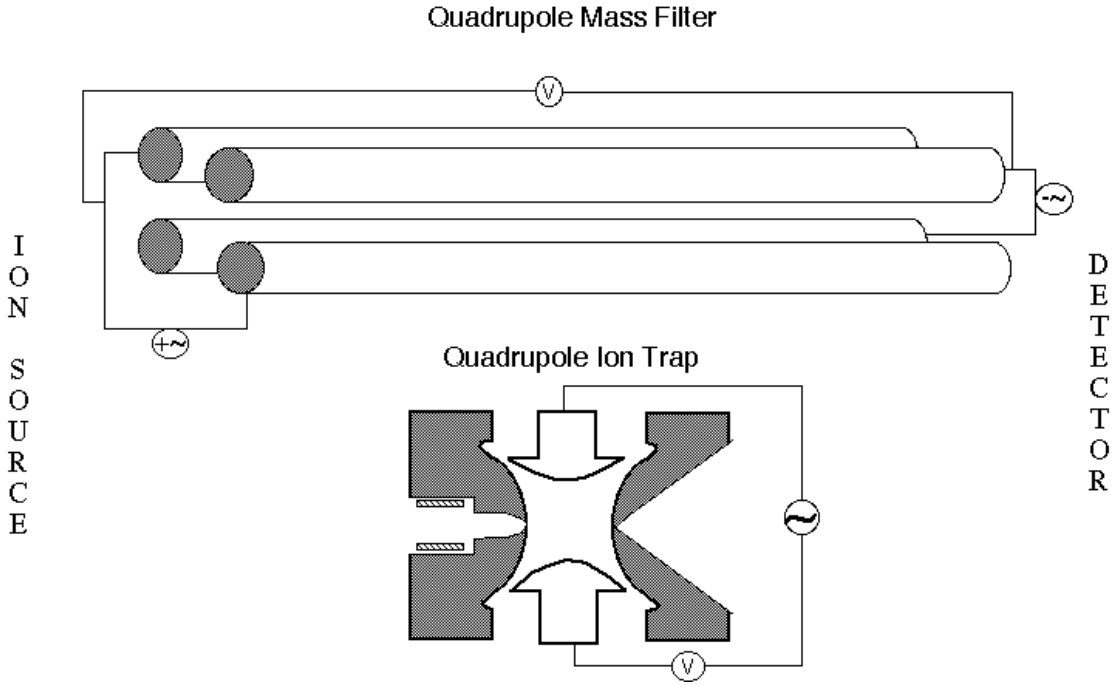
Για την βελτίωση της αναλυτικής ικανότητας μεταξύ μαζών, πολλές φορές χρησιμοποιούνται πολλαπλοί αναλυτές ή Tandem MS. Για παράδειγμα έχουμε την περίπτωση του τριπλού τετράπολου αναλυτή – Triple Quadrupole Analyzer όπου λόγω του ότι ένα μονό τετράπολο έχει αμελητέα οφέλη στην πρωτεωμική ανάλυση, βάζει τρία τετράπολα στην σειρά [23]. Σε μια τέτοια δομή, το πρώτο και το τρίτο τετράπολο συμπεριφέρονται σαν φίλτρα μαζών ενώ το δεύτερο αποτελεί ένα διαμέρισμα επιπλέον θραύσης των ιόντων [19].



Εικόνα 2. 11 : Τριπλός τετράπολος ανιχνευτής [19].

2.2.2.2 Τετράπολη παγίδα ιόντων – Quadrupole Ion Trap

Η αρχική περιγραφή της τεχνικής έγινε το 1960 [24,25] ενώ η υλοποίησή της το 1984 [26]. Αποτελείται από ένα δακτυλιοειδές ηλεκτρόδιο με σφαιρικά κάλυπτρα στην κορυφή και την βάση του. Μοιάζει με ένα συνεστραμμένο τετράπολο έτσι ώστε να είναι ένας κλειστός βρόχος ή διαφορετικά θα λέγαμε ότι πρόκειται για ένα τρισδιάστατο τετράπολο. Στην εικόνα 2.12 βλέπουμε συγκριτικό σχέδιο του τετράπολο και της τετράπολης παγίδας ιόντων.

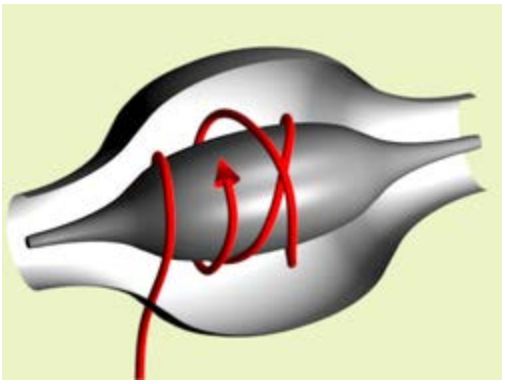


Εικόνα 2. 12 : Συγκριτικό σχέδιο τετράπολου ανιχνευτή και τετράπολης παγίδας ιόντων.

Αφού έχουν παραχθεί τα ιόντα, περιορίζεται ο αριθμός τους που εισέρχονται στην παγίδα ιόντων για να μην περιοριστεί η διακριτική ικανότητα. Στην συνέχεια εφαρμόζεται ένα δυναμικό σταθερής συχνότητας με μεταβαλλόμενο εύρος έτσι ώστε να διαχωρίσουμε τα ιόντα.

2.2.2.3 Ηλεκτροστατική παγίδα ιόντων – Orbitrap

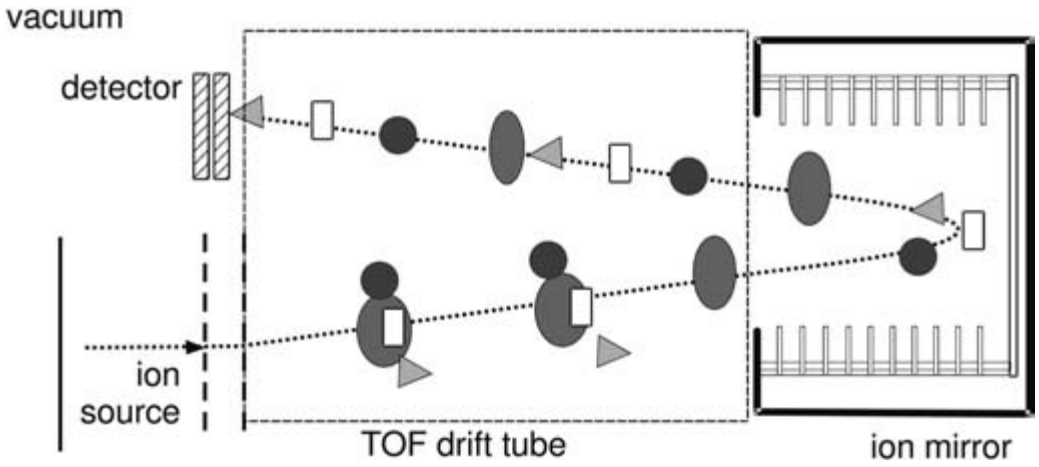
Η ηλεκτροστατική παγίδα ιόντων χρησιμοποιεί τον μετασχηματισμό Fourier για να πάρει τα φάσματα μάζας. Προτάθηκε από τον Makarov [27-28] και αποτελείται από ένα ηλεκτρόδιο με μορφή βαρελιού με ένα άλλο εσωτερικό ηλεκτρόδιο το οποίο δημιουργεί ηλεκτροστατικό πεδίο γύρω του. Τα ιόντα παγιδεύονται σε αυτό το πεδίο κινούμενα γύρω από το εσωτερικό ηλεκτρόδιο σχηματίζοντας δακτυλίους. Ανάλογα με το m/z είτε θα κινηθούν μπρος ή πίσω σε σχέση με το εσωτερικό ηλεκτρόδιο ή θα παγιδευτούν εκεί κινούμενα κυκλικά. Η ηλεκτροστατική παγίδα ιόντων έχει μεγάλη ακρίβεια και μεγάλη αναλυτική ικανότητα.



Εικόνα 2. 13 : Σχέδιο Orbitrap [28].

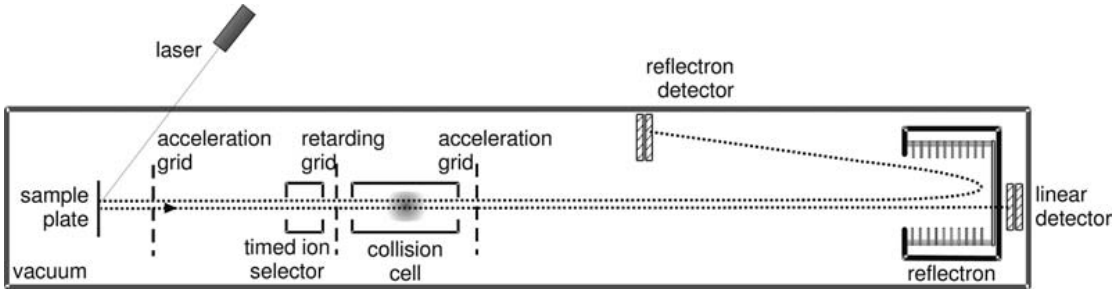
2.2.2.4 Αναλυτές χρόνου πτήσης – Time of flight analyzers TOF

Ο αναλυτής χρόνου πτήσης είναι μια μέθοδος φασματομετρίας μάζας στην οποία ο λόγος m/z του ιόντος υπολογίζεται με μια χρονική μέτρηση. Τα ιόντα επιταχύνονται από ηλεκτρικό πεδίο [29,30] με αποτέλεσμα κάθε ιόν να έχει την ίδια κινητική ενέργεια με οποιοδήποτε άλλο ιόν του ίδιου φορτίου. Ουσιαστικό λοιπόν, όσο μικρότερο το μόριο τόσο γρηγορότερα θα διανύσει την απόσταση μέσα στο σωλήνα πτήσης μέχρι τον ανιχνευτή. Ο χρόνος που το κάθε ιόν χρειάζεται για να φτάσει στον τελικό ανιχνευτή – detector, σε συγκεκριμένη απόσταση, μετριέται και στην συνέχεια από αυτόν και άλλους πειραματικούς παραμέτρους υπολογίζεται ο λόγος m/z του ιόντος. Ο αναλυτής χρόνου πτήσης είναι ένας παλμικός αναλυτής και συνήθως χρησιμοποιείται μαζί με μία πηγή ιονισμού MALDI. Στην παρακάτω εικόνα βλέπουμε το σχεδιάγραμμα ενός αναλυτή χρόνου πτήσης.



Εικόνα 2. 14 : Σχηματική μορφή TOF αναλυτή [19].

Ένας διπλός αναλυτής χρόνου πτήσης TOF/TOF συνδυασμένος με μια MALDI πηγή ιονισμού βοηθάει αρκετά στην ανάλυση της ακολουθίας των πεπτιδίων σε μία μόνο συσκευή [32]. Τυπικά, οι δύο αναλυτές χρόνου πτήσης χωρίζονται από ένα διαμέρισμα επιμέρους θραυσματοποίησης, με τον πρώτο αναλυτή να κάνει μια πρώτη διαλογή και τον δεύτερο να διαχωρίζει τα τελικά ιόντα. Ακολουθούν μια εικόνα σχεδιαγράμματος Maldi με Tandem TOF και μία με το πώς είναι ένα τέτοιο ολοκληρωμένο μηχάνημα.



Εικόνα 2. 15 : Σχεδιάγραμμα Maldi με tandem TOF [19].

Εικόνα 2. 16 : Maldi με Tandem TOF.

2.2.2.5 Ιοντικός κυκλοτρονικός συντονισμός με μετασχηματισμό Fourier – Fourier transform ion cyclotron.

Για πρώτη φορά ο ιονικός κυκλονικός συντονισμός εφαρμόστηκε στην φασματομετρία μάζας από τον Sommer [33]. Όταν ένα ιόν κινείται με μικρή ταχύτητα, η τροχιά του κάμπτεται όταν εισέλθει σε ένα μαγνητικό πεδίο. Η ταχύτητα του ιόντος και η ένταση του πεδίου μπορεί να είναι τέτοιες ώστε το ιόν να παγιδευτεί σε μια κυκλική τροχιά στο μαγνητικό πεδίο κάτι που αποτελεί την αρχή λειτουργίας του κύκλοτρου ιόντων. Πιο κάτω παρατηρούμε την μορφή ενός ιοντικού κυκλότρου από το ινστιτούτο πυρηνικής τεχνολογίας της Πορτογαλίας.

Εικόνα 2. 17: Fourier transform ion cyclotron [34].

2.2.3 Ανιχνευτές – Detectors

Η δέσμη των ιόντων αφού περάσει από τον αναλυτή ή τους αναλυτές μάζας θα πρέπει στην συνέχεια να ανιχνευτεί και να μετατραπεί σε ένα χρησιμοποιούμενο σήμα από έναν Ανιχνευτή – Detector. Υπάρχει πλειάδα διαφορετικών ανιχνευτών, και σίγουρα διαφορετικοί τρόποι κατάταξής τους σε κατηγορίες αλλά πάντα βασίζονται στην μάζα, το φορτίο ή την ταχύτητα του ιόντος. Διαφορετικοί τρόποι διαχωρισμού των ανιχνευτών που συναντά κανείς στην βιβλιογραφία είναι :

- Αν βασίζονται στην μέτρηση του φορτίου που δημιουργείται όταν το ιόν χτυπήσει μια επιφάνεια και ουδετεροποιηθεί (κύπελλο του Faraday) ή αν βασίζονται στην κινητική ενέργεια την οποία έχουν τα ηλεκτρόνια που δημιουργήθηκαν με την σύγκρουση με την επιφάνεια και τα οποία στην συνέχεια ενισχύονται για να δώσουν ένα ηλεκτρικό σήμα.
- Αν μετράνε ιόντα μιας μοναδικής μάζας κάθε φορά και άρα ανιχνεύουν την άφιξη όλων των ιόντων σειριακά σε ένα σημείο ή αν έχουν την ικανότητα να μετρήσουν πολλαπλές μάζες και άρα να ανιχνεύσουν την άφιξη πολλών ιόντων ταυτόχρονα όπως η φωτογραφική πλάκα και οι ανιχνευτές μήτρας – array detector.
- Αν επιτρέπουν την άμεση μέτρηση των φορτίων που φτάνουν στον ανιχνευτή (πχ φωτογραφική πλάκα και κύπελλο του Faraday) ή αν αυξάνουν την ένταση του σήματος που φτάνει στο ανιχνευτή (ηλεκτρονικοί πολλαπλασιαστές).

3. ΑΝΑΛΥΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ MS ΦΑΣΜΑΤΟΣ

Γενικότερα, από τις διάφορες πρωτεωμικές μελέτες, προκύπτει ότι το βασικό πλαίσιο μιας πρωτεωμικής ανάλυσης αποτελείται ως επί το πλείστον από το εξής βήματα.

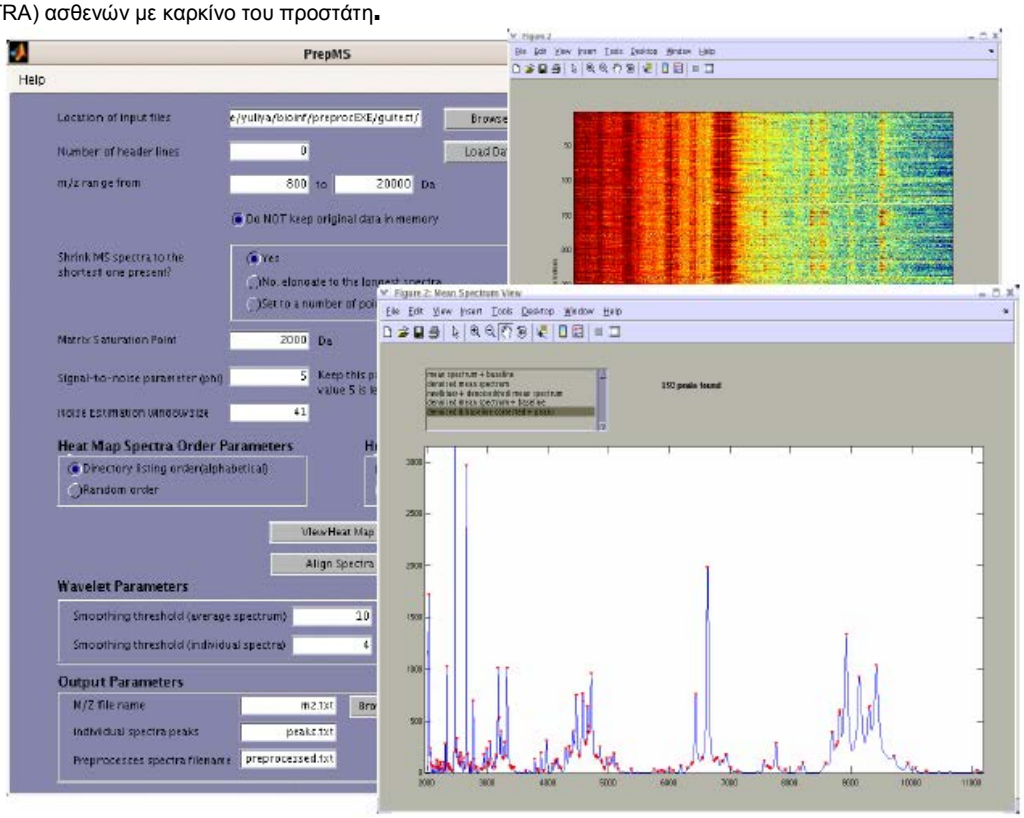
- Προεπεξεργασία φασμάτων και πιο συγκεκριμένα αφαίρεση του baseline, εξισορρόπηση του φάσματος (smoothing) και κανονικοποίηση του.
- Εξαγωγή χαρακτηριστικών (peak extraction) από το προεπεξεργασμένο φάσμα. Ανάλογα με την παθολογική κατάσταση ή το τι εξετάζεται στο συγκεκριμένο φάσμα επιλέγονται τα αντίστοιχα χαρακτηριστικά – κορυφές στο φάσμα.
- Ευθυγράμμιση των κορυφών (peak alignment) μεταξύ των διάφορων φασμάτων.
- Ελάττωση χαρακτηριστικών (feature reduction) επιλέγοντας τα χαρακτηριστικά εκείνα τα οποία μας προσφέρουν καλύτερη διαχωριστική ικανότητα μεταξύ των φασμάτων.
- Σχεδιασμός και χρήση μεθόδων μηχανικής μάθησης για την όσο το δυνατόν καλύτερη απόδοση του όλου συστήματος.

3.1 Προεπεξεργασία φάσματος

Είναι γνωστό ότι πολλές φορές πεπτίδια χαμηλής συγκέντρωσης αδυνατούν να εντοπιστούν και αυτό συμβαίνει γιατί υπερκαλύπτονται από θόρυβο και έτσι έχουμε υψηλό ποσοστό σε false positives κατά το στάδιο του peak detection. Επίσης ο χημικός, ο ιονικός και ο ηλεκτρονικός θόρυβος οδηγεί σε μια φθίνουσα καμπύλη στα MS δεδομένα που αναφέρεται ως baseline και εισάγει ένα ισχυρό σφάλμα στο peak detection. Όλοι αυτοί οι παράγοντες μας δείχνουν ότι είναι απαραίτητη μια προ-επεξεργασία των δεδομένων που προέρχονται από φασματογράφο μάζας ώστε έπειτα, «καθαρισμένα» πλέον, να χρησιμοποιηθούν για την ορθή ταύτιση των πεπτιδίων. Η ποιο αποδεκτή μορφή ενός MS σήματος δίνεται από την σχέση:

$$f(t) = B(t) + N \cdot S(t) + \varepsilon(t) \qquad (3.1)$$

Το *f(t)* είναι το εν τέλει παρατηρούμενο σήμα, το *S(t)* είναι το αναμενόμενο πραγματικό σήμα που έπρεπε να βλέπαμε ιδανικά, το *B(t)* είναι ένας προσθετικός παράγοντας βάσης που οδηγεί στο πρόβλημα του baseline, το *N* είναι ένας παράγοντας κανονικοποίησης και στο *e(t)* συναθροίζουμε όλους τους υπόλοιπους δυνατούς θορύβους που συμμετέχουν στην διαδικασία. Πλέον υπάρχουν αρκετά πακέτα τα οποία κάνουν μια ολοκληρωμένη προεπεξεργασία των φασμάτων όπως το Wave-spec [45], και το PrepMS [46], μέρος της εκτέλεσης του οποίου φαίνεται στην εικόνα 3.1



Εικόνα 3. 1: Snapshot από το εκτελέσιμο του PreMS [46].

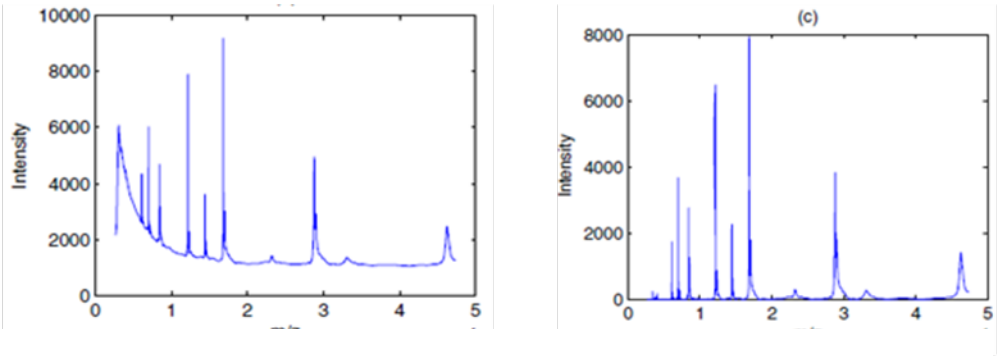
Η όλη διαδικασία της προεπεξεργασίας των δεδομένων από MS μπορεί να αναλυθεί σε κάποια συγκεκριμένα στάδια, τα οποία βεβαίως δεν είναι αναγκαία να υπάρχουν σε όλες τις περιπτώσεις και είναι αυτά που ακολουθούν :

3.1.1 Υπολογισμός και αφαίρεση βασικής γραμμής (baseline removal).

Η προεπεξεργασία των φασμάτων ξεκινά με την αφαίρεση του μονοτονικά ελαττούμενου baseline από το φάσμα έτσι ώστε να ελαττωθεί ο χημικός θόρυβος [37]. Ουσιαστικά υπολογίζεται το πραγματικό σήμα και στην συνέχεια η διαφορά παρατηρούμενου από υπολογισθέν σήμα αφαιρείται από το φάσμα Για να γίνει όλη αυτή η διεργασία:

- αρχικά χωρίζεται το φάσμα σε μικρά τμήματα
- χρησιμοποιείται το mean, minimun ή median κάθε τμήματος ως το σημείο baseline
- δημιουργείται το baseline για το φάσμα από τα επιμέρους baseline points.

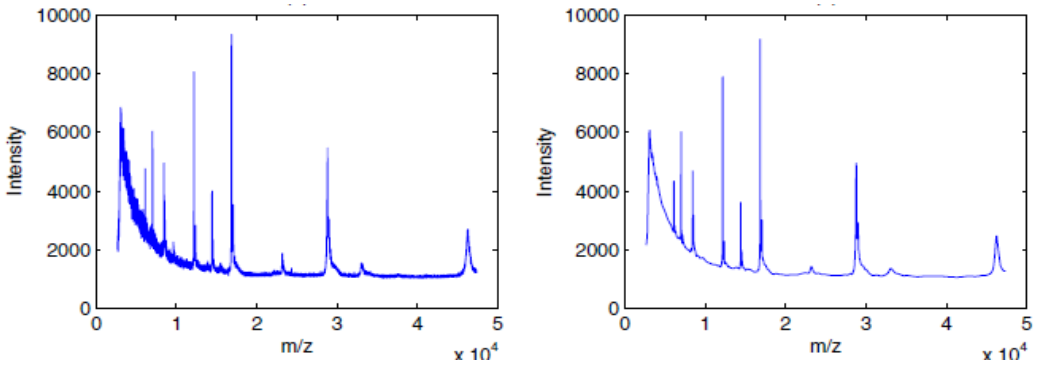
Για τον συγκεκριμένο καθαρισμό του φάσματος έχουν προταθεί και χρησιμοποιηθεί πλειάδα τεχνικών καθώς δεν υπάρχει κάποια κοινά αποδεκτή μέθοδος. Συνήθως αυτές οι μέθοδοι ομαδοποιούνται σε ευρετικές μεθόδους, που κάνουν έναν μη παραμετρικό υπολογισμό του baseline από ένα σύνολο φασμάτων και σε μεθόδους που χρησιμοποιούν ένα μαθηματικό μοντέλο βασισμένο στην ιδιότητες του φασματομέτρη μάζας και το οποίο παραμετροποιείται από ένα σύνολο φασμάτων. Γενικότερα έχουμε μεθόδους που κάνουν εξισορρόπηση σήματος μέσω τοπικής γραμμικής αναδρομής (local linear regression) [38,39], κυλιόμενα παράθυρα [40,41] αλλά και μη γραμμική προσέγγιση φίλτρων όπως το Top-Hat [42-43] ή προσεγγίσεις αυτού [44]. Όλες οι μέθοδοι πάντως βασίζονται στο μέγεθος του χρησιμοποιούμενου παραθύρου, πολύ μεγάλο παράθυρο οδηγεί σε υπεραπλούστευση του προβλήματος κάνοντας το baseline μια ευθεία γραμμή ενώ πολύ μικρό παράθυρο δημιουργεί μια αρκετά σύνθετη πρόγνωση για το baseline. Στην εικόνα που ακολουθεί ένα παράδειγμα φάσματος, πριν και μετά την αφαίρεση του baseline.



Εικόνα 3. 2 : Σχεδιάγραμμα φάσματος πριν και μετά την αφαίρεση του baseline [47].

3.1.2 Εξισορρόπηση (smoothing) φάσματος.

Στόχος σε αυτό το στάδιο είναι να μετριάσουμε την επίδραση του θορύβου e(t) και έτσι το φάσμα μας να εμφανίζεται σε μια ποιο «απαλή - smooth» μορφή. Οι τεχνικές που υπάρχουν για αυτό το στάδιο συνήθως χρησιμοποιούν παραδοσιακές τεχνικές επεξεργασίας σήματος όπως η κίνηση ενός φίλτρου μέσου όρου, ένα Savitzky – Golay φίλτρο ή ένα Gaussian φίλτρο. Συνήθως το φάσμα που παίρνουμε είναι συνεχές αλλά το δειγματολειπτούμε έτσι ώστε η επεξεργασία μας να είναι σε μια διακριτή θα λέγαμε μορφή του. Ένα φάσμα μετά από smoothing μπορεί να αποδοθεί από την σχέση y(t) = x(t) * w(t) όπου w(t) μια συνάρτηση με βάρη. Η χρήση διαφορετικών w(t) οδηγεί σε διαφορετικά φίλτρα για το πρόβλημά μας. Ποιο κάτω στην εικόνα 3.3, παρατηρούμε μια εικόνα του αρχικού φάσματος αριστερά και του ίδιου φάσματος αφού περάσει από κάποιο smoothing φίλτρο δεξιά. Βλέπουμε λοιπόν πόσο η αφαίρεση του θορύβου έχει κάνει ποιο ομαλή την μορφή του φάσματος.

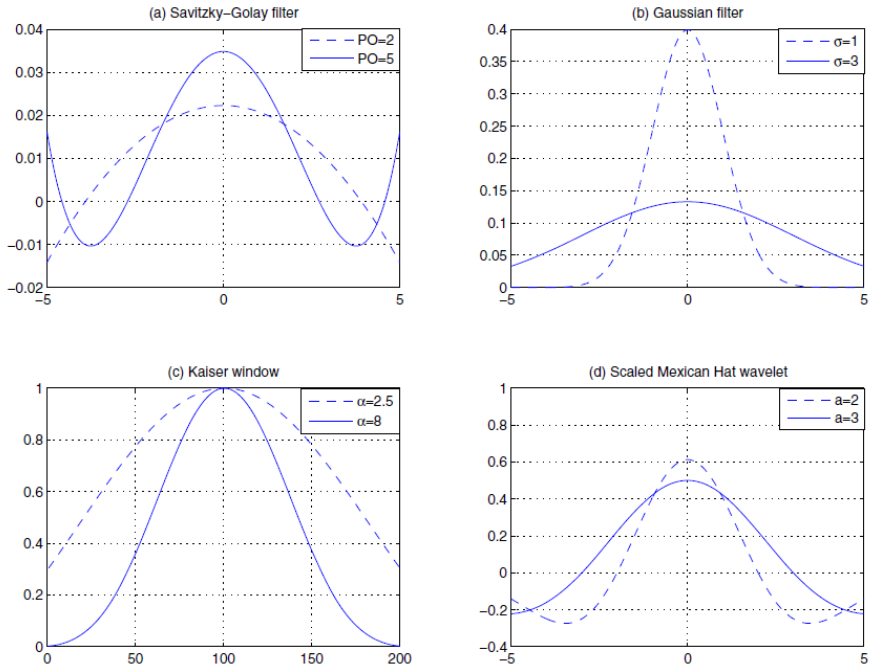


Εικόνα 3. 3 : Φάσμα πριν και μετά το smoothing [47].

Για την εξισορρόπηση του σήματος έχουν χρησιμοποιηθεί διάφορα φίλτρα [47], όπως τα :

- o Moving average filter
- o Savitzky-Golay filter
- o Gaussian filter
- o Kaiser window
- o Continuous Wavelet Transform
- o Discrete Wavelet Transform
- o Undecimated Discrete Wavelet Transform

με το τελευταίο να είναι το ποιο ευρέως χρησιμοποιούμενο. Παρακάτω βλέπουμε την μορφή κάποιων από αυτών των φίλτρων :



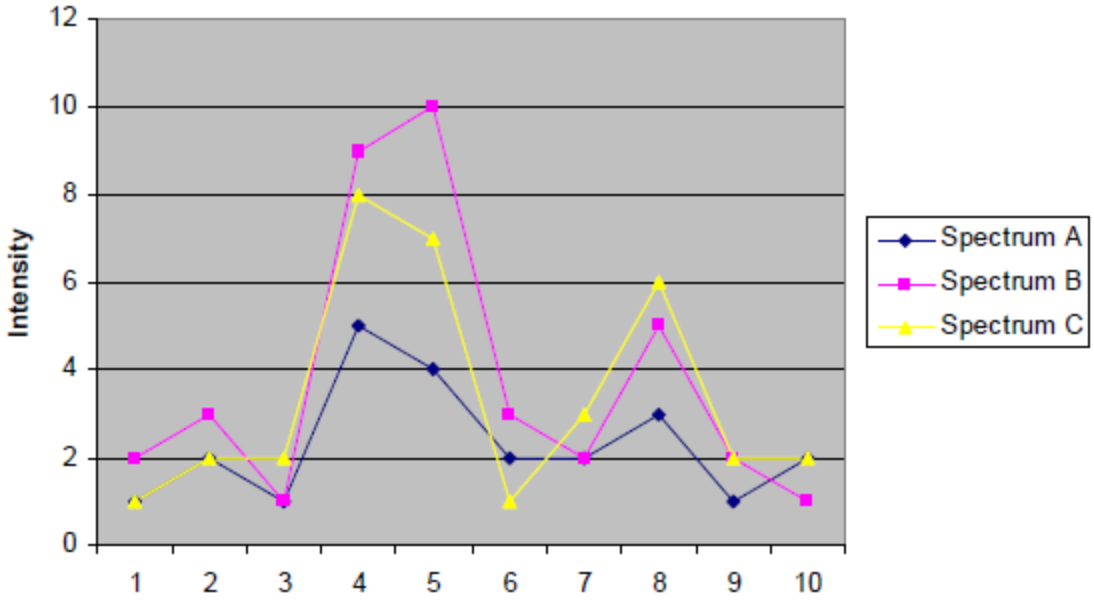
Στο (α) το PO είναι η τάξη του πολυωνύμου για το φίλτρο, στο (b) σ είναι η τυπική απόκλιση, στο (c) το α έχει να κάνει με την μορφή του Kaiser φίλτρου και στο (d) το α είναι η κλίμακα του wavelet.

Εικόνα 3. 4 : Διάφορα φίλτρα που χρησιμοποιούνται για το smoothing [47].

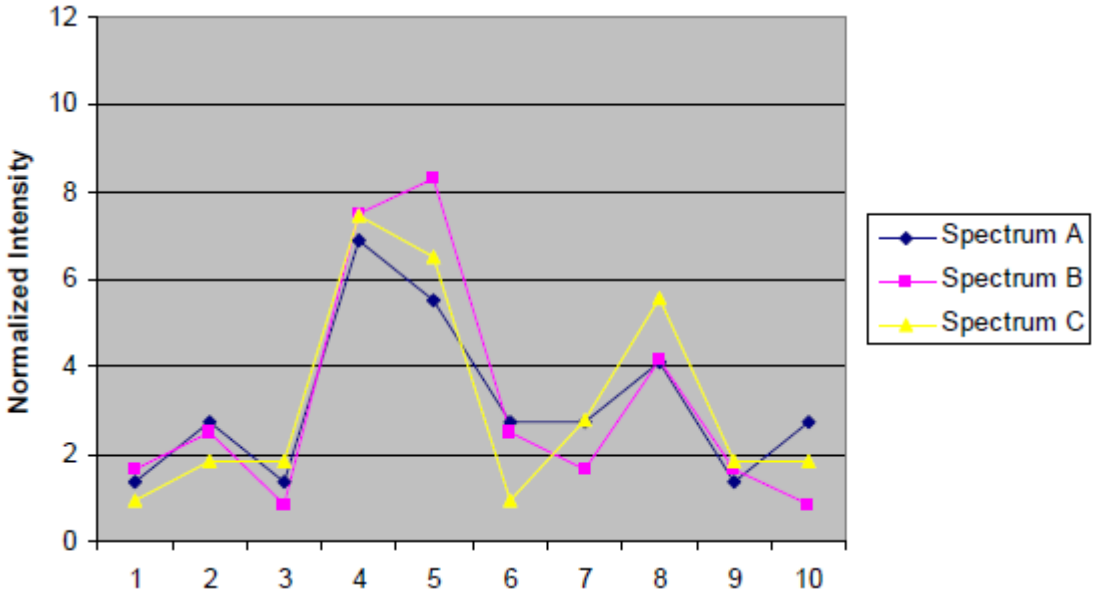
3.1.3 Κανονικοποίηση (Normalization)

Πολλές φορές σε δείγματα τα οποία είναι συναφή οι λόγοι m/z συγκεκριμένων πεπτιδίων παρουσιάζουν διαφορές που δεν θα έπρεπε να υπήρχαν. Για παράδειγμα μπορεί να μετράμε συγκεκριμένες πρωτεΐνες από ασθενείς με μια συγκεκριμένη πάθηση και τα m/z ίδιων πεπτιδίων να εμφανίζονται σε λίγο διαφορετικές θέσεις. Για να «υπερπηδήσουμε» ένα τέτοιο πρόβλημα κάνουμε την διαδικασία της κανονικοποίησης ή αλλιώς normalization. Με την κανονικοποίηση όλες οι εντάσεις των φασμάτων μεταφέρονται στο ίδιο πεδίο τιμών.

Για δεδομένα από φασματομετρία μάζας, η συγκέντρωση μιας πρωτεΐνης μετριέται από την περιοχή κάτω από την καμπύλη της κορυφής της (Area Under Curve - AUC). Το πακέτο προγραμμάτων PROcess της BioConductor χρησιμοποιεί την μεσαία τιμή του AUC για την κανονικοποίηση ενός συνόλου φασμάτων ενώ το Ciphergen ProteinChip χρησιμοποιεί το μέσο του AUC. Στις εικόνες που ακολουθούν βλέπουμε την κανονικοποίηση 3 σημάτων από το Ciphergen ProteinChip.

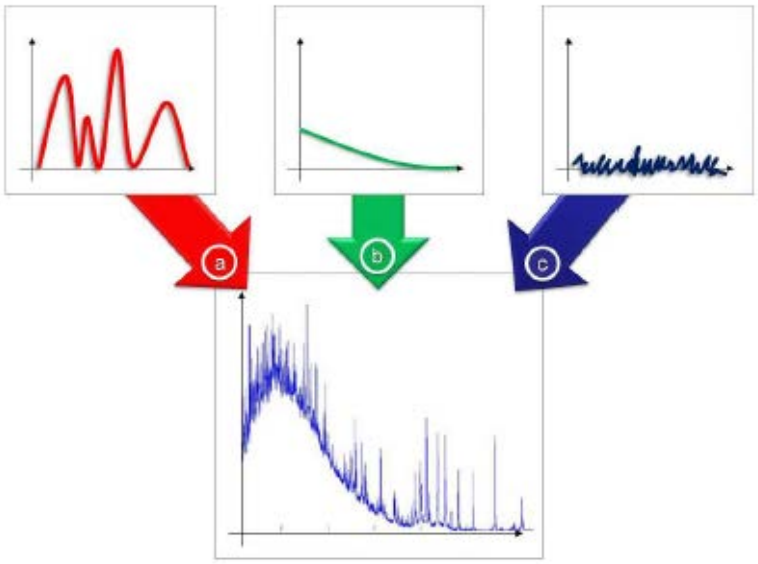


Εικόνα 3. 5 : Τρία σήματα πριν την κανονικοποίηση.



Εικόνα 3. 6 : Τα σήματα της εικόνας 3.5 μετά την κανονικοποίηση.

Γενικότερα λοιπόν, το σήμα από ένα φασματομέτρη μάζας, εκτός από το βασικό μας σήμα περιέχει και το baseline καθώς και ηλεκτρονικό ή χημικό θόρυβο κάτι που βλέπουμε στην εικόνα 3.7 που ακολουθεί.



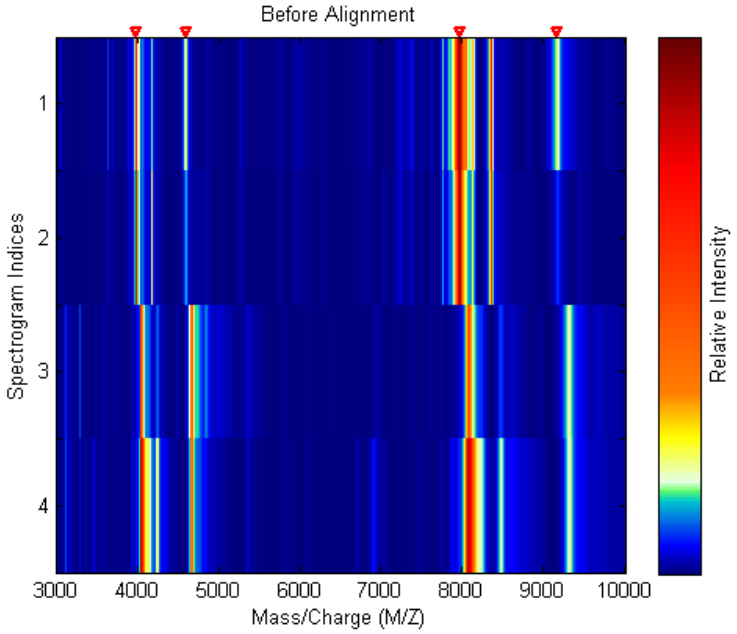
Εικόνα 3. 7 : (a)βασικό σήμα,(b) baseline,(c)θόρυβος και το αποτελεσμά τους, το τελικό φάσμα μάζας.

3.2 Εξαγωγή χαρακτηριστικών (feature extraction)

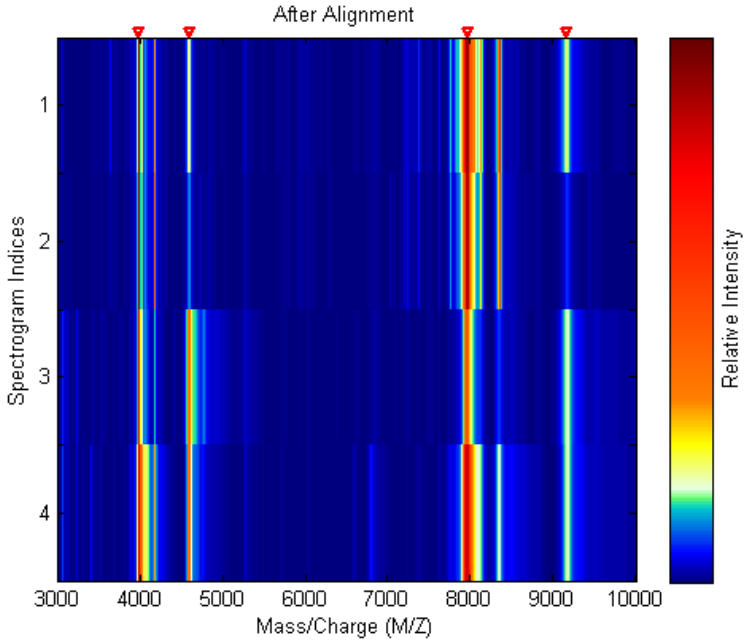
Σε αυτό το στάδιο έχουμε συνδυασμό της εύρεσης των χαρακτηριστικών που διαφοροποιούν καλύτερα της παθολογικές καταστάσεις ή γενικότερα τις ιδιότητες των προς εξέταση φασμάτων και του φιλτραρίσματος του θορύβου. Τα χαρακτηριστικά που επιλέγονται είναι αυτά που χαρακτηρίζουν όσο το δυνατόν ποιο αποδοτικά τα φάσματα. Ο απλοϊκός τρόπος επιλογής χαρακτηριστικών είναι να χρησιμοποιηθούν όλες οι κορυφές [49,50]. Οι κορυφές σε ένα φάσμα μάζας είναι αλληλοεξαρτώμενες και άρα δεν είναι απαραίτητες όλες για την ανάδειξη της πρωτεΐνης που υποδεικνύεται από αυτές. Επιπλέον η πολυπλοκότητα μιας τέτοιας αντιμετώπισης είναι πάρα πολύ υψηλή με αποτέλεσμα να μην είναι αποδοτική μια τέτοια τεχνική. Άλλη στρατηγική που χρησιμοποιήθηκε στην φασματομετρία μάζας για την εξαγωγή χαρακτηριστικών είναι το bining. Το φάσμα χωρίζεται σε τμήματα – bins ως προς την διάσταση του λόγου m/z και από κάθε τμήμα εξάγεται ένα χαρακτηριστικό. Για να εξαχθεί αυτό το χαρακτηριστικό μπορεί να χρησιμοποιηθεί το μέγιστης έντασης σημείο εντός του bin ή ακόμα και το μέσο. Γενικότερα λοιπόν, στο στάδιο αυτό, εντοπίζονται οι κορυφές ως τα τοπικά μέγιστα σε ένα συγκεκριμένο διάστημα. Στην συνέχεια υπολογίζεται ο τοπικός θόρυβος εντός ενός κινούμενου παραθύρου και κρατούνται ως χαρακτηριστικά, οι έχοντες πληροφορία κορυφές δηλαδή αυτές που βρίσκονται πάνω από τον τοπικό θόρυβο. Ο τρόπος υπολογισμού του θορύβου μπορεί να διαφέρει καθώς μπορεί να υπολογισθεί είτε ως ο μέσος όρος των εμφανιζόμενων εντάσεων εντός του παραθύρου είτε ως η ένταση η οποία εμφανίζεται συχνότερα μέσα στο παράθυρο.

3.3 Ευθυγράμμιση κορυφών (Peak Alignment)

Λόγο των εναλλαγών στις πειραματικές συνθήκες μπορεί να έχουμε διαφορετικά φάσματα ακόμα και από το ίδιο δείγμα. Οι διαφορές αυτές μπορεί βρίσκονται τόσο στην θέση εμφάνισης μιας κορυφής όσο και στο σχήμα αυτής. Επομένως πρέπει να βρεθεί ένας αποδοτικός τρόπος ευθυγράμμισης των κορυφών μεταξύ φασμάτων. Οι υπάρχουσες τεχνικές ευθυγράμμισης θα μπορούσαν να χωριστούν σε αυτές που ευθυγραμμίζουν συγκεκριμένες κορυφές (με υψηλή ένταση) και σε αυτές που ευθυγραμμίζουν τμήματα m/z.



Εικόνα 3. 8: Εικόνα φάσματος πριν από την ευθυγράμμιση, Matlab Bioinformatics Toolbox



Εικόνα 3. 9 : Εικόνα φάσματος μετά από την ευθυγράμμιση, Matlab Bioinformatics Toolbox

3.4 Μείωση χαρακτηριστικών (Peak Reduction)

Επιπλέον στάδιο στην ανάλυση ενός MS φάσματος είναι η μείωση των χαρακτηριστικών που έχουν επιλεγεί. Ένας απλοϊκός τρόπος μείωσης των χαρακτηριστικών είναι βρίσκοντας τα μέσο φάσμα κάθε κατηγορίας (πχ αν μελετείται το φάσμα παθολογικής κατάστασης και φυσιολογικής να υπολογιστεί το μέσο φάσμα από τα παθολογικά και το μέσο από τα φυσιολογικά) και στην συνέχεια να εντοπιστούν οι κορυφές εκείνες που διακρίνουν αυτά τα δύο φάσματα καλύτερα. Να βρεθούν δηλαδή οι κορυφές οι οποίες οι επί τοις εκατό διαφορά τους μεταξύ των 2 φασμάτων είναι μεγαλύτερη από ένα νούμερο. Όσο μεγαλύτερο αυτό το νούμερο, τόσο περισσότερο απέχουν τα φάσματα στο συγκεκριμένο m/z αλλά και τόσο μικρότερος ο αριθμός των

χαρακτηριστικών που μένει στο τέλος. Άλλοι τρόποι μείωσης χαρακτηριστικών βασίζονται στην χρήση κάποιου στατιστικού τεστ, όπως για παράδειγμα τα :

- o Wilcoxon rank sum test
- o T-test
- o Sign test
- o Wilcoxon signed rank test
- o Kruskal – Wallis test

Ο ουσιαστικός λόγος μείωσης των χαρακτηριστικών είναι ο εντοπισμός και η ενασχόληση μόνο με τα χαρακτηριστικά τα οποία προσφέρουν μέγιστη διαχωριστική ικανότητα μεταξύ των φασμάτων.

3.5 Σχεδίαση ταξινομητών και υπολογισμός απόδοσης συστήματος

Γενικότερα στην μηχανική μάθηση έχουν σχεδιαστεί και υλοποιηθεί πλειάδα αλγορίθμων, για διάφορους τύπους δεδομένων, οι οποίοι εκπαιδεύουν τους υπολογιστές ώστε να μπορούν να αυτοί στη συνέχεια να ταξινομήσουν τύπους – μοτίβα εντός αυτών των δεδομένων. Διακρίνονται συνήθως σε μεθόδους ταξινόμησης με επιτήρηση και μεθόδους ταξινόμησης χωρίς επιτήρηση.

Οι μέθοδοι ταξινόμησης χωρίς επιτήρηση προσπαθούν να ομαδοποιήσουν τα δεδομένα ανάλογα με το πόσο όμοια είναι αυτά και λόγο αυτού μπορούμε να τις συναντήσουμε στην βιβλιογραφία και ως μεθόδους ομαδοποίησης (clustering methods). Τέτοιες μέθοδοι ταξινόμησης δεν έχουν εφαρμοστεί ιδιαίτερα στις προσπάθειες για διάγνωση καρκίνου μέσα από ανάλυση φασμάτων MS [50,52].

Οι μέθοδοι ταξινόμησης με επιτήρηση δίνουν στον υπολογιστή συγκεκριμένες τιμές για τα χαρακτηριστικά ώστε να καταφέρουν να τα ταξινομήσουν σωστά. Βασισμένοι λοιπόν σε ένα χαρακτηριστικό των προτύπων, οι υπολογιστές προσπαθούν να ταξινομήσουν αυτό το πρότυπο σε κάποια από τις προϋπολογισθέντες κλάσεις. Τέτοιες τεχνικές χρησιμοποιούνται κατά κόρον στην προσπάθεια για διάκριση πρωτεϊνικών δειγμάτων. Στην βιβλιογραφία μπορεί κανείς να βρει προσπάθειες με χρήση δέντρων αποφάσεων [54-57], τεχνιτών νευρωνικών δικτύων [57], με χρήση του ταξινομητή k κοντινότερων γειτόνων [57,58], support vector machines SVM [59] αλλά και διάφορα άλλα σχήματα με συνδυασμούς ταξινομητών [55].

Αφού σχεδιαστούν, υλοποιηθούν και δοκιμαστούν οι ταξινομητές θα πρέπει να υπάρχουν κάποια συγκεκριμένα μέτρα για την αποτίμηση της απόδοσης. Αυτό που συνήθως γίνεται είναι η καταμέτρηση θετικών ή αρνητικών δειγμάτων που ταξινομήθηκαν σωστά και η σύγκρισή τους με τα θετικά/αρνητικά που ταξινομήθηκαν λανθασμένα. Συνεπώς ορίζονται τα :

- o True Positive – *TP* ή σωστά ταξινομημένα θετικά πρότυπα, τα δείγματα των ασθενών που ταξινομήθηκαν σωστά ως ασθενή.
- o True Negative - *TN* ή σωστά ταξινομημένα αρνητικά δείγματα, τα δείγματα των υγιών που ταξινομήθηκαν σωστά ως υγιή.
- o False Positive – *FP* ή λανθασμένα ταξινομημένα θετικά δείγματα, τα δείγματα των υγιών που ταξινομήθηκαν λανθασμένα ως ασθενή.
- o False Negative – *FN* ή λανθασμένα ταξινομημένα αρνητικά δείγματα, τα δείγματα των ασθενών που ταξινομήθηκαν λανθασμένα ως υγιή.

Παρατηρούμε ότι TP + FN = #αρνητικών δειγμάτων και TN + FP = #θετικών δειγμάτων. Με βάση τους παραπάνω όρους υπολογίζονται οι συχνότερα χρησιμοποιούμενες μετρικές της βιβλιογραφίας οι οποίες είναι οι εξής :

- **Sensitivity** ή ευαισθησία, η ικανότητα διάκρισης σωστά των περιπτώσεων που φέρουν την ασθένεια.

$$Sensitivity = \frac{TP}{TP+FN} \quad (3.2)$$

- **Specificity** ή σαφήνεια, η ικανότητα διάκρισης σωστά των περιπτώσεων που δεν φέρουν την ασθένεια.

$$Specificity = \frac{TN}{TN+FP} \quad (3.3)$$

- **Accuracy** ή ακρίβεια, η ικανότητα διάκρισης όλων συνολικά των περιπτώσεων.

$$Accuracy = \frac{TP+TN}{TN+TP+FP+FN} \quad (3.4)$$

4. ΦΑΣΜΑΤΟΜΕΤΡΙΑ ΜΑΖΑΣ ΩΣ ΔΙΑΓΝΩΣΤΙΚΟ ΤΕΣΤ

Η έγκαιρη ανίχνευση του καρκίνου του προστάτη, ενός από τους πιο συχνά εμφανιζόμενους καρκίνους που αποτελεί την δεύτερο κατά σειρά παράγοντα θανάτου στους άντρες, έχει αποδειχτεί κάτι το εξαιρετικά δύσκολο. Η εφαρμογή του PSA-test, την ανίχνευση δηλαδή της ποσότητας του προστατικού ειδικού αντιγόνου στο αίμα, βοήθησε σε αυτό τον τομέα αλλά και πάλι η αύξηση του PSA και σε περιπτώσεις που δεν υπάρχει καρκίνος του προστάτη κάνει το τεστ αυτό μη συγκεκριμένο ως προς αυτήν την ασθένεια. Επομένως η έρευνα έχει επικεντρωθεί στην εύρεση μεθόδων έγκαιρης ανίχνευσης του καρκίνου του προστάτη ώστε να συμπληρώσουν ή να αντικαταστήσουν το τεστ – PSA.

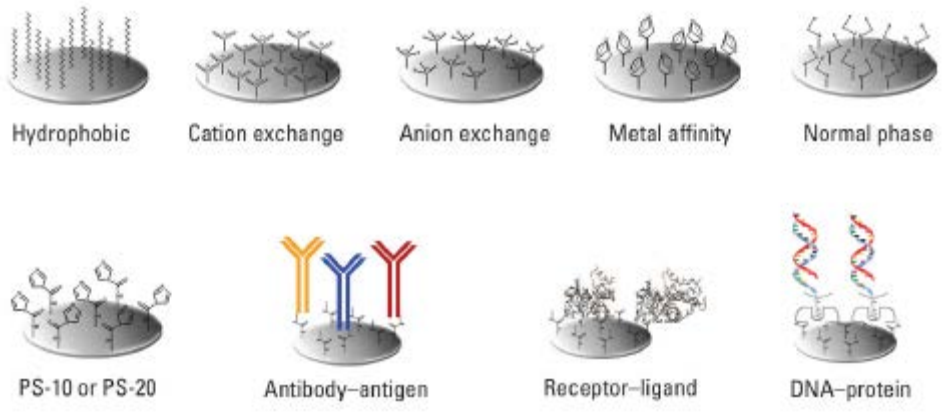
Μια από την προοπτικές που έχουν εμφανιστεί στο χώρο και μελετούνται κατά κόρον από τις διάφορες ερευνητικές ομάδες έχει να κάνει με την χρήση τεχνικών φασματομετρίας μάζας ως διαγνωστικό εργαλείο για τις διάφορες μορφές καρκίνου. Η δυνατότητα που μας προσφέρουν τέτοιες τεχνικές, να εντοπιστούν πρωτεΐνες ή πεπτίδια, και η μετέπειτα συσχέτισή τους με κάποια ασθένεια, προσφέρει κάποια προοπτική γύρω από την διάγνωση του καρκίνου. Κυριότερη τεχνική που εμφανίζεται στην βιβλιογραφία είναι η χρήση τεχνολογίας SELDI για την παραγωγή των φασμάτων τα οποία με διάφορους τρόπους στην συνέχεια (προεπεξεργασία – μηχανική μάθηση) θα οδηγήσουν στην εύρεση ύποπτων πεπτιδίων – πρωτεϊνών και στην μετέπειτα έρευνα τους ως βιοδείκτες.

4.1 Τεχνολογία SELDI

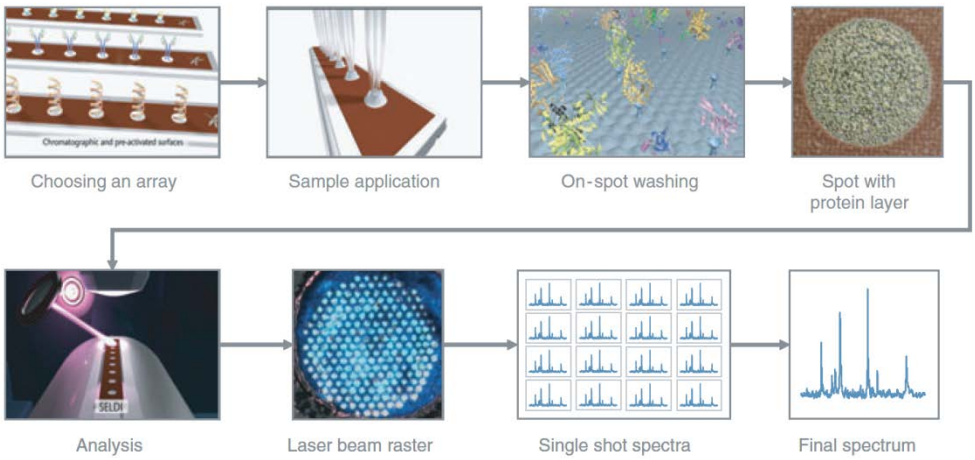
Αυτός που ουσιαστικά έκανε την αρχή στην έρευνα για την χρήση του mass spectrometry στην ανακάλυψη νέων βιοδεικτών ήταν ο Petricoin και η ομάδα του. Η ιδέα που ξεκίνησε όλη αυτήν την έρευνα ήταν ότι «πρωτεΐνες οι οποίες παράγονται από καρκινικά κύτταρα πιθανόν να εισέλθουν στην κυκλοφορία και έπειτα να καταφέρουμε να τις εντοπίσουμε, αναλύσουμε με έναν φασματογράφο μάζας και έτσι να χρησιμοποιηθούν για διαγνωστικούς σκοπούς». Το θέμα όμως είναι ότι δεν ξέρουμε αν τέτοια μόρια είναι δυνατόν να εντοπιστούν από την υπάρχοντα τεχνολογία. Επίσης, μήπως τα πεπτίδια – μόρια που ανιχνεύονται δεν είναι παράγωγα καρκινικών κυττάρων αλλά άλλων κυττάρων κάτι δηλαδή σαν επιφαινόμενο της παρουσίας του καρκίνου; Όλα φαντάζουν καινούρια σε αυτόν τον τομέα που φαίνεται να συγκεντρώνει τα βλέμματα, ίσως δικαιολογημένα ίσως όχι. Μέχρι σήμερα έχει γίνει προσπάθεια να διαγνωστούν πολλές μορφές καρκίνων μέσω mass spectra όπως οι εξής : breast, prostate, bladder, pancreatic, head, neck, lung, melanoma, liver, nasopharyngeal cancers, gliomas κ.τ.λ.

Το Seldi μπορεί να θεωρηθεί σαν εξέλιξη της τεχνικής του Maldi. Οι κύριες διαφορές τους βρίσκονται στην κατασκευή των στόχων-δειγμάτων αλλά και στο μετέπειτα αναλυτή που θα χρησιμοποιηθεί. Το βιολογικό υγρό που μας ενδιαφέρει αρχικά αλληλεπιδρά με ένα πρωτεϊνικό τσιπ έτσι ώστε να μπορέσουμε να διακρίνουμε τις πρωτεΐνες που μας ενδιαφέρουν σε πρώτη φάση. Αυτές αλληλεπιδρούν με το πρωτεϊνικό τσιπ μένοντας εκεί ενώ εμείς θα «ξεπλύνουμε» το όλο περιβάλλον για να φύγει ότι δεν χρειαζόμαστε. Στην συνέχεια θα χτυπηθεί από μια ακτίνα laser και στην συνέχεια θα αναλυθεί από ένα αναλυτή χρόνου πτήσης (συνήθως) προτού γίνει η τελική ανίχνευση των πρωτεϊνών – πεπτιδίων. Αναφερόμαστε μόνο στο Seldi (Surface enhanced laser desorption/ionization) αλλά όπως έγινε κατανοητό εννοούμε συνήθως όλο το σύστημα SELDI – TOF – MS. Τα πρωτεϊνικά τσιπ που χρησιμοποιούνται στην περίπτωση του SELDI διαφέρουν από το MALDI στο ότι προκαλούν μια ποιο συγκεκριμένη στόχευση σε πρωτεΐνες – πεπτίδια από το δείγμα. Τελικά. ο βασικός

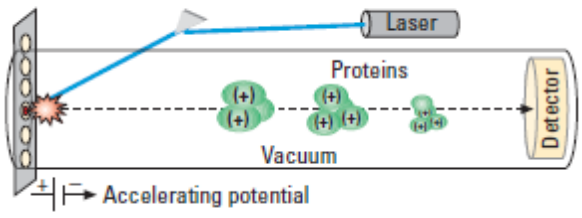
στόχος είναι να βρεθούν κάποια «διαχωριστικά» peaks και να αναγνωριστούν μέσω της τεχνολογίας του SELDI-TOF και έτσι, γνωρίζοντας τα μόρια τα οποία αναπαρίστανται από αυτά τα peaks, να γίνει χρήση φθηνότερων τεχνικών για την μετέπειτα διάγνωση σε άλλους ασθενείς. Στην εικόνα 4.1 που ακολουθεί βλέπουμε τρόπους δημιουργία των πρωτεϊνικών τσιπ, χημικούς ή βιοχημικούς ενώ στην εικόνα 4.2 έχουμε όλα τα πειραματικά στάδια σε μια ανάλυση με SELDI. Τέλος η εικόνα 4.3 δείχνει μια σχηματική αναπαράσταση της τεχνολογίας SELDI.



Εικόνα 4. 1 : Τρόποι δημιουργία πρωτεϊνικών τσιπ [60].



Εικόνα 4. 2 : Πειραματικά στάδια ανάλυσης με Seldi [61].



Εικόνα 4. 3 : Σχεδιάγραμμα τεχνολογίας SELDI [60].

Παρόλο βέβαια την προοπτική που προσφέρει η τεχνολογία SELDI στην εύρεση βιοδεικτών και στην έγκαιρη ανίχνευση του καρκίνου, έχει και τα μειονεκτήματά της [62]. Τα διακριτικά χαρακτηριστικά που εντοπίζονται εν τέλει είναι αποτέλεσμα καρκινικών καταστάσεων ή λάθη που έχουν εισαχθεί με την τεχνολογία του SELDI ; Τα αποτελέσματα τέτοιων τεχνολογιών επαληθεύονται και από διαφορετικές ομάδες ; Γιατί οι ήδη αποδεκτοί βιοδείκτες όπως το PSA δεν έχουν εντοπιστεί από τεχνολογίες SELDI; Όλα αυτά είναι μόνο κάποια από το ερωτήματα που υπάρχουν στην έρευνα περί της χρήσης τεχνικών SELDI ως διαγνωστικό εργαλείο. Επιπλέον μειονεκτήματα είναι ότι δεν βγαίνει απευθείας οι ταυτότητες της πρωτεΐνης από ένα SELDI πείραμα, το περιορισμένο δυναμικό εύρος που έχουν τέτοιες τεχνολογίες και μας αποτρέπει τουλάχιστον προς το παρόν από την ανίχνευση χαμηλής συγκέντρωσης πρωτεϊνών/πεπτιδίων αλλά και πολλά λάθη στην υπόδειξη βιοδεικτών λόγο μη βέλτιστου συστήματος.

4.2 Βιβλιογραφία

Έως σήμερα υπάρχουν άρθρα που αναφέρουν πολύ υψηλά ποσοστά σε sensitivity και specificity ειδικά αν αναλογιστούμε τα αντίστοιχα ποσοστά που βγαίνουν με χρήση των ήδη γνωστών βιοδεικτών αν μιλήσουμε συγκεκριμένα για το καρκίνο. Το μόνο σίγουρο λοιπόν είναι ότι δεν έχουμε και πολλά να περιμένουμε από τους ήδη υπάρχοντες βιοδείκτες και για αυτό ο στόχος πολλών ερευνητικών ομάδων είναι η εύρεση νέων βιοδεικτών αλλά από την άλλη όλες αυτές οι μέθοδοι που προτείνονται πρέπει να ελέγχουν προσεκτικά και να επαληθευτούν πριν περάσουν στην στάδιο της κλινικής εφαρμογής. Πολλές μελέτες έχουν πραγματοποιηθεί, που αφορούν διάφορα στάδια της πρωτεωμικής ανάλυσης όπως είναι οι προεπεξεργασία του φάσματος με αποδοτικούς τρόπους ώστε να μειωθεί στο ελάχιστο η επίδραση του χημικού και του ηλεκτρονικού θορύβου. Επίσης διάφορες τεχνικές έχουν προταθεί για την εξαγωγή χαρακτηριστικών από το φασματικό σήμα αλλά και διάφοροι τρόποι ταξινόμησης αυτού σε παθολογικό ή όχι χρησιμοποιώντας πλειάδα τεχνικών μηχανικής μάθησης.

Ποιο συγκεκριμένα τώρα, όσον αφορά την μελέτη του καρκίνου του προστάτη από πλειάδα άρθρων [66] αναφέρεται ως μια πιο δύσκολη μορφή για αξιολόγηση καθώς τα αποτελέσματα που παρουσιάζονται ευρύτερα είναι πιο πενιχρά σε σχέση με άλλους καρκίνους όπως για παράδειγμα αυτός της μήτρας. Επίσης τα δεδομένα που υπάρχουν στην βιβλιογραφία δεν είναι πλούσια και σχεδόν όλες οι ερευνητικές ομάδες περιστρέφονται γύρω από 2-3 dataset. Τα πρωτεϊνικά chip που εμφανίζονται είναι τα IMAC-Cu, WCX2 καθώς και τα υδροφοβικά H4 και C16.

Στο [54] οι ερευνητές χρησιμοποίησαν ένα δέντρο απόφασης για την διάκριση υγιών και ασθενών και στην συνέχεια χρησιμοποιήθηκε ο χώρος κάτω από την καμπύλη ROC για την αναγνώριση σημαντικών κορυφών, οι οποίες στην συνέχεια ταξινομήθηκαν μέσω ενός αλγόριθμου δέντρου απόφασης. Οι ερευνητικές αναφέρουν ποσοστά τις τάξης του 81% για sensitivity και 97% specificity. Στο [55] η ίδια ομάδα ερευνητών χρησιμοποιώντας μια διαφορετικού είδους προσέγγιση κυρίως όσον αφορά την επιλογή χαρακτηριστικών και την ταξινόμηση επιτυγχάνουν ποσοστά που κυμαίνονται μεταξύ του 97 και του 100% τόσο για sensitivity όσο και για specificity. Αυτή την φορά έκαναν χρήση των αλγορίθμων AdaBoost και Boosted Decision Stump Feature Selection (BDSFS) με την βασική διαφορά να εγγυάται στο ότι τώρα επέτρεπαν την επιλογή του ίδιου χαρακτηριστικού πολλαπλές φορές. Στο [64] η ομάδα του Wagner χρησιμοποίησε το ίδιο σετ δεδομένων με την χρήση διάφορων ταξινομητών. Ο γραμμικός SVM ήταν ο βέλτιστος φτάνοντας ακρίβεια της τάξεως του 96.4%. Στο σημείο αυτό πρέπει να αναφερθεί ότι μεταξύ των 3 προσπαθειών που ήδη αναφέρθηκαν εντοπίζουμε τις μόνες κοινές κορυφές. Οι κορυφές 7.024 και 9.656 εντοπίζονται και στις δύο προσπάθειες της

ομάδας των Qu και Adam με την κορυφή 7.820 να εντοπίζεται και αυτή αλλά για διαφορετικό λόγο(μια για διάκριση ασθενών με υγιών και μια για διάκριση υγιών από αυτούς που έχουν καλοήθη όγκο). Επιπλέον οι κορυφές 7.820 και 9.656 εμφανίζονται και από την ομάδα του Wagner. Το ίδιο πρωτεϊνικό chip χρησιμοποιεί και η ομάδα του Banez [67] αλλά αναφέρει αρκετά χαμηλά ποσοστά.

Στο [63] οι ερευνητές κάνουν χρήση γενετικών αλγορίθμων για την επιλογή χαρακτηριστικών ενώ για την ταξινόμηση χρησιμοποιούν Self Organizing Maps αναφέροντας ποσοστά της τάξεως του 95% για sensitivity και 78 έως 83% specificity. Στο [65] οι ερευνητές χρησιμοποιούν ένα μικρό δείγμα της τάξεως των 35 ατόμων, το υδροφοβικό H4 πρωτεϊνικό chip και παρουσιάζουν ποσοστά μεταξύ 64-82% για sensitivity και 67-100% για specificity μιλώντας για 3 συγκεκριμένες και μόνο κορυφές, οι οποίες δεν έχουν ξαναεμφανιστεί στην βιβλιογραφία. Στο [66] οι ερευνητές προτείνουν έναν αλγόριθμο ο οποίος παρόλο του ότι επιτυγχάνει τρομερά ποσοστά όσον αφορά τον καρκίνο της μήτρας δεν τα καταφέρνει εξίσου καλά με τα δεδομένα από καρκίνο του προστάτη. Επιπλέον το ποσοστά που επιτυγχάνει για το καρκίνο του προστάτη είναι πολύ κυμαινόμενο και κάθε φορά εξαρτάται από τον διαχωρισμό των δεδομένων που θα κάνει ο αλγόριθμος αυτός. Γενικότερα αναφέρουν ποσοστά της τάξεως του 63-78% για sensitivity και 94-98% specificity. Ο πίνακας που ακολουθεί περιέχει συνοπτικά το τι εντοπίστηκε στην βιβλιογραφία σχετικά με προσπάθειες για καρκίνο του προστάτη.

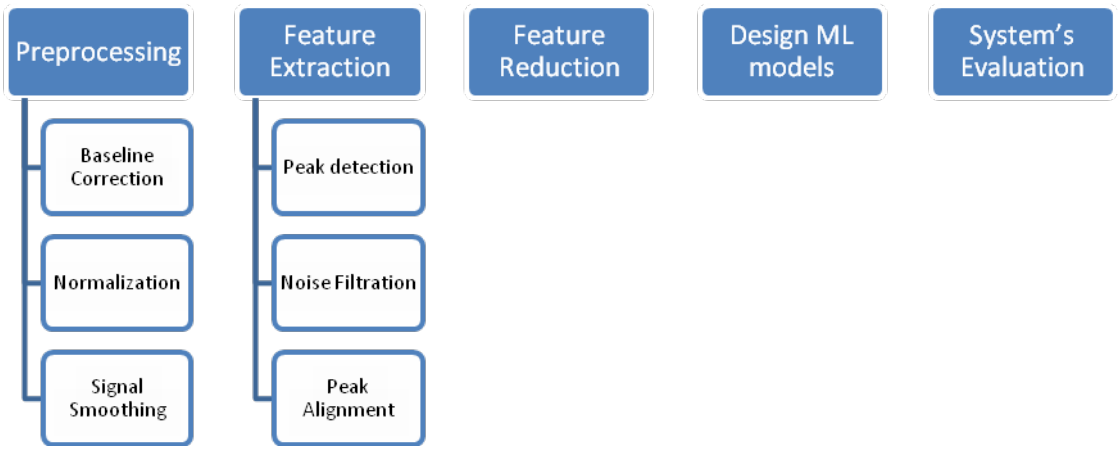
Πίνακας 4.1 : Σύγκριση αναφορών για καρκίνο του προστάτη

Ερευνητική ομάδα	Πρωτεϊνικό chip	Προτεινόμενες κορυφές m/z	Sensitivity - Specificity
Adam et al [54]	IMAC-Cu	4.475, 5.074, 5.382, **7.024**, **7.820**, 8.141, 9.149, 9.507, **9.656**	83% -97%
Qu et al [55]	IMAC-Cu	Non-cancer vs cancer : 3.963, 4.080, 6.542, 6.797, 6.949, 6.991, **7.024**, 7.885, 8.067, 8.356, **9.656**, 9.720 Healthy vs benign : 3.486, 4.071, 4.580, 5.298, 6.099, 7.054, **7.820**, 7.844, 8.943	97..100% - 97..100%
Wagner et al [64]	IMAC-Cu	3.897, 3.963, 4.080, 5.074, 6.949, **7.820**, 7.844, 8.943, **9.656**, 9.720	Acc : 70.. 96.4 %
Petricoin et al [63]	Hydrophobic C16	2.092, 2.367, 2.582, 3.080, 4.819, 5.439, 18.220	95% - 78..83%
Lehrer et al [65]	Hydrophobic H4	15.200, 15.900, 17.500	64..82% - 67..100%
Marchiori et al [66]	Hydrophobic C16		63..78% - 94..98%
Banez et al [67]	WCX2	3.972, 8.226, 13.952, 16.087, 25.167, 33.270	63%-77%
	IMAC-CU	3.960, 4.469, 9.713, 10.266, 22.832	66% - 38%

4.3 Προτεινόμενο σχήμα

Στην παρούσα εργασία πειραματιζόμαστε ευρέως με όλα τα στάδια και τους παραμέτρους της προεπεξεργασίας των δεδομένων από φασματομετρία μάζας. Στην συνέχεια αφού βρεθούν οι βέλτιστοι παράμετροι για την εξισορρόπηση του φάσματος, την αφαίρεση θορύβου αλλά και την εξαγωγή και την ευθυγράμμιση των κορυφών, χρησιμοποιούνται για να εντοπιστούν τα διαστήματα τιμών m/z τα οποία κρίνονται ως σημαντικότερα. Σε αυτά εφαρμόζεται μια μέθοδος μηχανικής μάθησης με την χρήση ταξινομητών για την εύρεση αυτών των διαστημάτων που επιτυγχάνουν καλύτερο διαχωρισμό των δειγμάτων μας σε προερχόμενα από υγιείς ή ασθενείς.

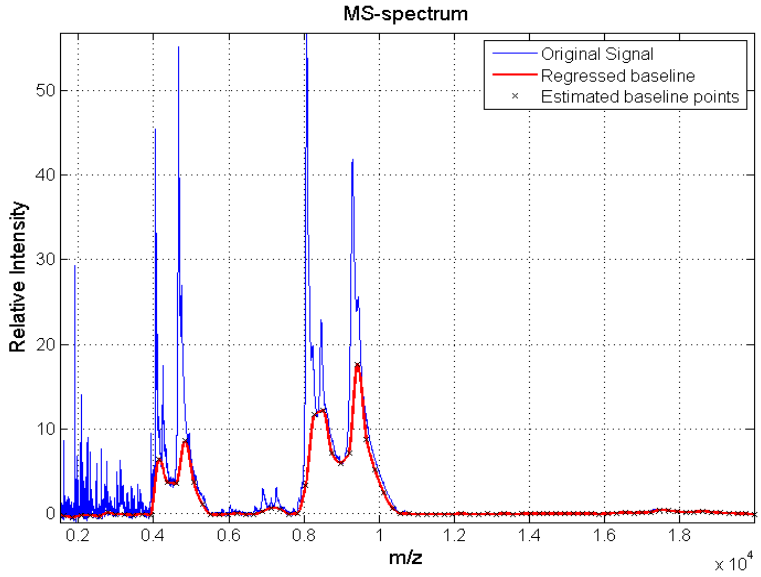
Δοκιμάζονται δύο διαφορετικά σετ δεδομένων, το (7-3-02) από τον δικτυακό χώρο του Εθνικού Καρκινικού Ινστιτούτου Αμερικής και το δεύτερο [68] να περιέχεται σε βιβλιοθήκη του R όπου έχει προστεθεί στις 08-02-2011 από τους Lixin Gong, William Constantine και Yu Alex Chen και είναι το σετ δεδομένων που χρησιμοποιήθηκε και από την ομάδα των Adam και Qu [54-55] και έχει εξαχθεί από την ομάδα του ιατρικού κέντρου στη Virginia. Όσον αφορά το πρώτο σετ δεδομένων δημιουργήθηκε χρησιμοποιώντας το πρωτεϊνικό chip H4. Έγινε χρήση του φασματογράφου μάζας τύπου SELDI και περιλαμβάνουν 63 φάσματα τα οποία αντιστοιχούν σε υγιείς με (PSA<1), 190 φάσματα καλοήθειας με (PSA>4), 26 που προήλθαν από ασθενείς που έχουν επίπεδα PSA από 4 έως 10 (PSA 4-10) και 43 από ασθενείς με επίπεδο PSA>10. Κάθε ένα πρωτεωμικό φάσμα περιέχει 15.154 τιμές μάζας/φορτίου καθώς και τις αντίστοιχες τιμές έντασης. Όσον αφορά το δεύτερο σετ δεδομένων αφού αφαιρεθούν τα διπλότυπα τα οποία περιέχει καταλήγουμε σε ένα σετ από 81 δείγματα υγιών, 78 με καλοήθη όγκο και 168 με καρκίνο του προστάτη. Στην πορεία θα αναφερθούν τα βήματα του αλγορίθμου τα οποία παρουσιάζονται και στην εικόνα 4.4.



Εικόνα 4. 4 : Βασικά βήματα αλγορίθμου

4.3.1 Υπολογισμός και αφαίρεση baseline

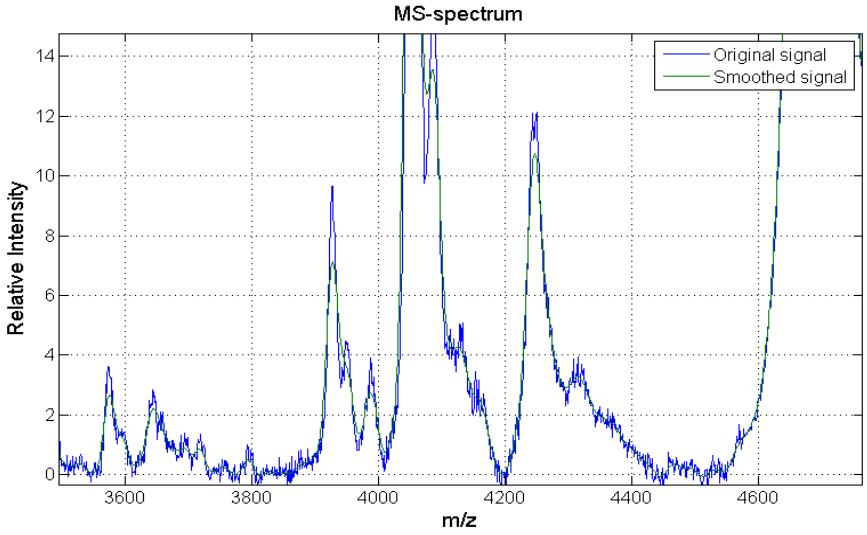
Ο τρόπος με τον οποίον πειραματιστήκαμε όσον αφορά τις παραμέτρους είναι η αλλαγή κάθε φορά μιας παραμέτρου και η καταγραφή της τελικής απόδοσης του συστήματος. Δοκιμάζοντας πλειάδα τιμών καταλήγουμε στην τιμή εκείνη η οποία μας επιτρέπει να έχουμε μέγιστη απόδοση στο όλο σύστημα. Για την αφαίρεση του baseline λοιπόν, χρησιμοποιήθηκε ένα κυλιόμενο παράθυρο μεγέθους 230 m/z με βήμα κίνησης 230 m/z ώστε να μην είναι αλληλεπικαλυπτόμενα. Επίσης μέθοδος αναδρομής ήταν η shape-preserving piecewise cubic interpolation ενώ μέθοδος εκτίμησης η quantile.



Εικόνα 4. 5 : Αφαίρεση baseline

4.3.2 Εξισορρόπηση (smoothing) φάσματος.

Για την εξισορρόπηση του φάσματος χρησιμοποιήθηκε Lowess φίλτρο εξομάλυνσης με μέγεθος παραθύρου 0.20% του αριθμού των στοιχείων m/z ενώ για την βαθμονόμηση των εντάσεων χρησιμοποιήθηκε η συνάρτηση πυρήνα tricubic.



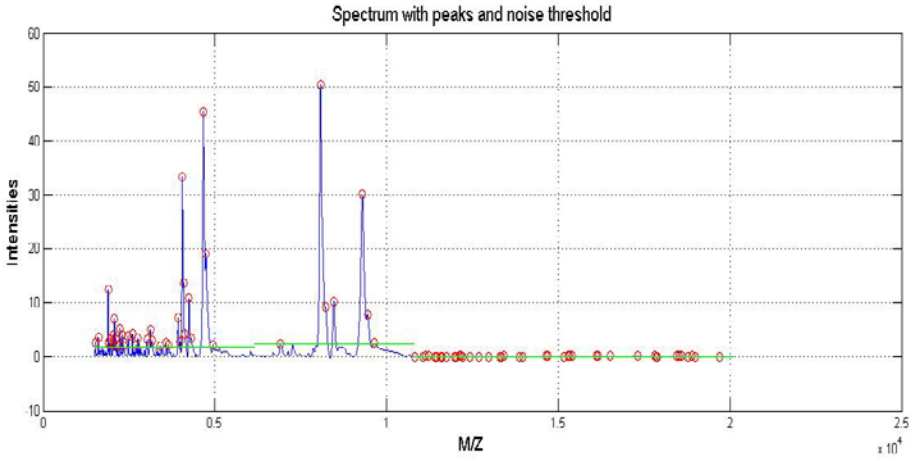
Εικόνα 4. 6: Εξισορρόπηση φάσματος

4.3.3 Αφαίρεση θορύβου και εξαγωγή χαρακτηριστικών

Σε ένα φάσμα μαζών εμφανίζονται και υψηλές και χαμηλές τιμές εντάσεων πράγμα το οποίο παραπέμπει στην προσπάθεια εύρεσης του θορύβου τοπικά και όχι στο σύνολο του φάσματος [69]. Συνεπώς ένα μη επικαλυπτόμενο κινούμενο παράθυρο χρησιμοποιήθηκε εντός του οποίου υπολογιζόταν κάθε φορά ο θόρυβος. Το μέγεθος του παραθύρου ήταν μεταβλητό αλλά μετά πλειάδα πειραματισμών καταλήξαμε σε βέλτιστο μέγεθος παραθύρου ίσο με 25% του συνολικού μεγέθους του φάσματος κάτι το οποίο επαληθεύει και προγενέστερες έρευνες επί του θέματος [69]. Όσον αφορά τον τρόπο υπολογισμού του θορύβου, αρχικά υπολογίζονται οι τιμές εντάσεων εντός του παραθύρου που είναι κάτω από το $90^{στό}$ percentile [69] και στην συνέχεια για τις τιμές αυτές υπολογίζεται η μέση τιμή και η τυπική απόκλιση οι οποίες εάν αθροιστούν μας δίνουν τον τελικό τοπικό θόρυβο. Συνεπώς ο τοπικός θόρυβος δίνεται από την σχέση:

$$Local\ Noise = \ mean(percentile(90) + std\,(percentile(90)) \qquad (4.1)$$

Τελικά τα χαρακτηριστικά τα οποία θα κρατηθούν για το συγκεκριμένο βήμα και άρα θα προωθηθούν στην περαιτέρω επεξεργασία είναι αυτά τα οποία έχουν ένταση μεγαλύτερη από την τοπική τιμή του θορύβου πράγμα το οποίο φαίνεται και στην εικόνα 4.7.



Εικόνα 4. 7 : Εξαγωγή χαρακτηριστικών και αφαίρεση θορύβου

4.3.4 Ευθυγράμμιση κορυφών (Peak Alignment)

Η αντιμετώπιση που ακολουθήθηκε στο παρόν βήμα του αλγορίθμου διαφοροποιεί την μέθοδο μας από την πλειοψηφία των άλλων μεθόδων που έχουν παρουσιαστεί από διάφορες ερευνητικές ομάδες. Στο σημείο αυτό εισάγεται η έννοια του διαστήματος m/z τιμών ως χαρακτηριστικό του προβλήματός μας και όχι μιας μεμονωμένης τιμής m/z. Ουσιαστικά λοιπόν η ευθυγράμμιση υλοποιείται καθώς μια ένταση αντιστοιχεί για εμάς σε ένα διάστημα τιμών m/z και δεν μας απασχολεί αυτήν κάθε αυτή η τιμή του m/z που παρουσιάζει την συγκεκριμένη ένταση. Για να διατηρηθεί ένα διάστημα τιμών m/z και να προωθηθεί στην περαιτέρω επεξεργασία θα πρέπει να εμφανίζεται τιμή σε αυτό το διάστημα για τουλάχιστον το 30% των φασμάτων μίας τουλάχιστον κλάσης.

Σε αυτό το σημείο θα πρέπει να τονιστούν τα διάφορα αντικείμενα πειραματισμού για το συγκεκριμένο βήμα. Αρχικά λοιπόν το μέγεθος του διαστήματος τιμών m/z ήταν μεταβλητό, ασφαλώς βέβαια εξαρτώμενο από την τιμή m/z και όχι σταθερού μεγέθους,

το οποίο στην πορεία υπολογίστηκε για βέλτιστο το is ±1/120 επί την τιμή m/z της εκάστοτε τιμής m/z. Δηλαδή το διάστημα τιμών για εμάς ήταν το:

$$Interval = [MZ - \frac{1}{120} * MZ \quad MZ + \frac{1}{120} * MZ\ ] \tag{4.2}$$

Στην πορεία πειραματιστήκαμε με το ποιος θα ήταν ο βέλτιστος αντιπρόσωπος του κάθε διαστήματος τιμών m/z και αυτά τα οποία εξετάστηκαν ήταν η μέγιστη τιμή εντάσεων που εμφανίζονται σε αυτό το διάστημα για κάποιο συγκεκριμένο φάσμα ή η μέση τιμή των εντάσεων που εμφανίζονται σε αυτό το διάστημα για συγκεκριμένο φάσμα. Μετά από πειραματισμούς βέλτιστος αντιπρόσωπος θεωρήθηκε η μέγιστη τιμή των εντάσεων που εμφανίζονται εντός του διαστήματος για ένα συγκεκριμένο φάσμα.

4.3.5 Μείωση χαρακτηριστικών

Για να εντοπιστούν και να ληφθούν υπόψη στην περαιτέρω επεξεργασία μόνο τα χαρακτηριστικά τα οποία προσφέρουν μέγιστη διαχωριστική ικανότητα εφαρμόστηκε το στατιστικό τεστ Wilcoxon. Με αυτόν τον τρόπο διαστήματα m/z με σημαντική στατιστική διαφορά ($p<0.05$) μεταξύ των εξεταζόμενων κλάσεων προωθήθηκαν σε περαιτέρω ανάλυση για τον τελικό σχεδιασμό και την υλοποίηση του συστήματος.

4.3.6 Σχεδίαση ταξινομητών και υπολογισμός απόδοσης συστήματος

Για τις ανάγκες του προβλήματος υλοποιήθηκαν 5 διαφορετικοί ταξινομητές, οι εξής :

I. Ταξινομητής ελάχιστης ευκλείδειας απόστασης

Ο βέλτιστος κατά Bayes ταξινομητής απλοποιείται αρκετά όταν ισχύουν οι κάτωθεν προϋποθέσεις :

- Οι κλάσεις είναι ισοπίθανες
- Τα δεδομένα στις κλάσεις ακολουθούν Gaussian κατανομή
- Ο πίνακας συνδιακύμανσης είναι ο ίδιος για κάθε κλάση
- Ο πίνακας συνδιακύμανσης είναι διαγώνιος, με όλα τα στοιχεία κατά μήκος της διαγωνίου είναι ίδια.

Κάτω από αυτές τις προϋποθέσεις ο βέλτιστος κατά Bayes ταξινομητής ταυτίζεται με το ταξινομητή ελάχιστης ευκλείδειας απόστασης σύμφωνα με τον οποίον ένα άγνωστο στοιχείο x ανήκει στην κλάση ω_i όταν

$$\left\|x - m_i\right\| \equiv \sqrt{(x - m_i)^T (x - m_i)} < \left\|x - m_j\right\|, \quad \forall\, i \neq j \tag{4.3}$$

Όπου m_i ,m_j το μέσο στοιχείο της κάθε κλάσης. Ουσιαστικά η παραπάνω σχέση μας λέει ότι ο ταξινομητής θα ταξινομήσει κάθε νέο στοιχείο x στην κλάση της οποίας μέσο είναι κοντινότερα στο στοιχείο x. Επομένως για την περίπτωσή μας υπολογίζεται αρχικά το μέσο στοιχείο από τα παθολογικά δείγματα, τα δείγματα των ασθενών με καρκίνο του

προστάτη, και το μέσο στοιχείο από τα φυσιολογικά δείγματα. Στην συνέχεια κάθε νέο προς ταξινόμηση στοιχείο θα εξεταστεί σχετικά με αυτά τα μέσα και ανάλογα με το σε ποιο είναι κοντίτερα εκεί θα ταξινομηθεί.

II. Ταξινομητής k κοντινότερων γειτόνων (kNN)

Στο ταξινομητή κοντινότερων γειτόνων πάλι χρησιμοποιείται η ευκλείδεια απόσταση αλλά αυτήν την φορά δεν εξετάζεται το μέσο στοιχείο της κλάσης αλλά εντοπίζουμε τους k κοντινότερους γείτονες συγκριτικά με το προς ταξινόμηση στοιχείο (τα k δηλαδή στοιχεία με την μικρότερη ευκλείδεια απόσταση από το προς ταξινόμηση στοιχείο). Αφού εντοπιστούν, εξετάζεται στο σε ποια κλάση ανήκουν τα περισσότερα από αυτά και αντίστοιχα σε αυτήν την κλάση ταξινομείται και το νέο πρότυπο.

III. Bayesian ταξινομητής.

Ο ταξινομητής αυτός βασίζεται στο θεώρημα του Bayes υποθέτοντας ότι κάθε χαρακτηριστικό από ένα πρότυπο μιας κλάσης είναι ανεξάρτητο από οποιοδήποτε άλλο χαρακτηριστικό του προτύπου. Ένας τέτοιος ταξινομητής θα οδηγούσε το άγνωστο πρότυπο $x = [x_1, x_2, x_3, \ldots\ldots, x_l]^T$ στην κλάση

$$\omega_m = \arg\, max\omega_i \quad \Pi_{j=1}^{l} p(x_j|\omega_i) \tag{4.4}$$

Στην περίπτωσή μας δοκιμάστηκε k=1,3,5 και 7.

IV. Πιθανοκρατικός νευρωνικός ταξινομητής ή Probabilistic Neural Network (PNN)

Η συνάρτηση διάκρισης για έναν PNN ταξινομητή δίνεται από τις κάτωθεν σχέσεις.

$$g_j(x) = \frac{1}{(2\pi)^{p/2}\sigma^p N_j} \sum_{i=1}^{Nj} w(y) \tag{4.5}$$

όπου $w(y_i)$ είναι συνάρτηση του $y_i = \frac{1}{\sigma}\sqrt{\|x - x_i\|^2}$

Μερικές κατάλληλες συναρτήσεις του y_i είναι οι :

(Gaussian)

$$w(y) = e^{\frac{y^2}{2}} \quad \rightarrow \quad g_j(x) = \frac{1}{(2\pi)^{p/2}\sigma^p N_j} \sum_{i=1}^{Nj} e^{\frac{\|x-x_i\|^2}{2\sigma^2}} \tag{4.6}$$

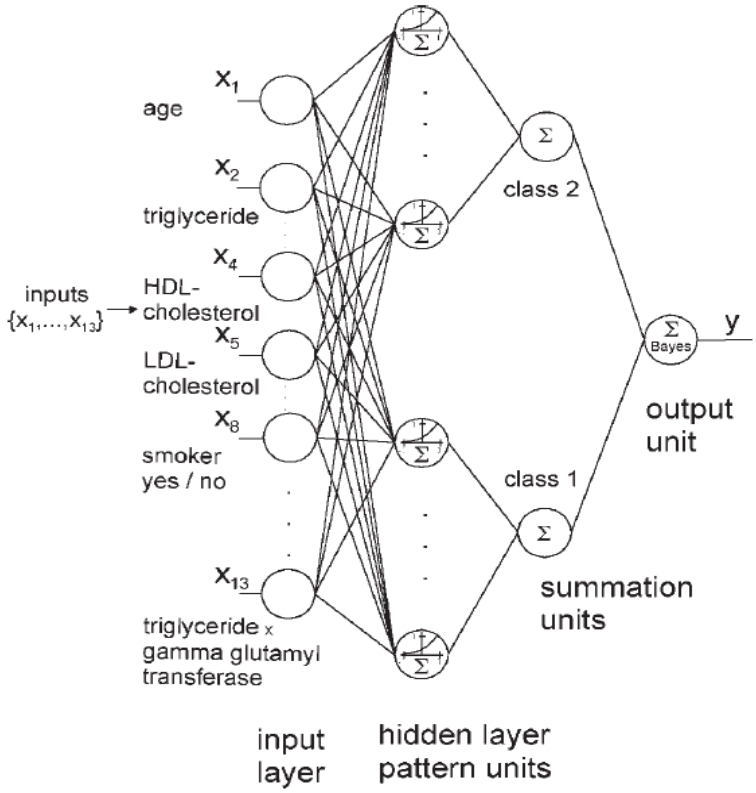
(Exponential)

$$w(y) = e^{-|y|} \quad \rightarrow \quad g_j(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma^p N_j} \sum_{i=1}^{Nj} e^{\frac{\|x-x_i\|}{\sigma}} \tag{4.7}$$

(Reciprocal)

$$w(y) = \frac{1}{1+y^2} \quad \rightarrow \quad g_j(x) = \frac{1}{(2\pi)^{p/2}\sigma^p N_j} \sum_{i=1}^{Nj} \frac{1}{1+\|x-x_i\|^2/\sigma^2} \tag{4.8}$$

Όπου x είναι το προς ταξινόμηση πρότυπο, x_i το i-οστό πρότυπο για την εκπαίδευση, N_j είναι το νούμερο των προτύπων στην κλάση j, σ είναι ένας παράμετρος εξισσορόπησης (smoothing parameter) και p είναι ο αριθμός των χαρακτηριστικών που διαθέτει κάθε πρότυπο.

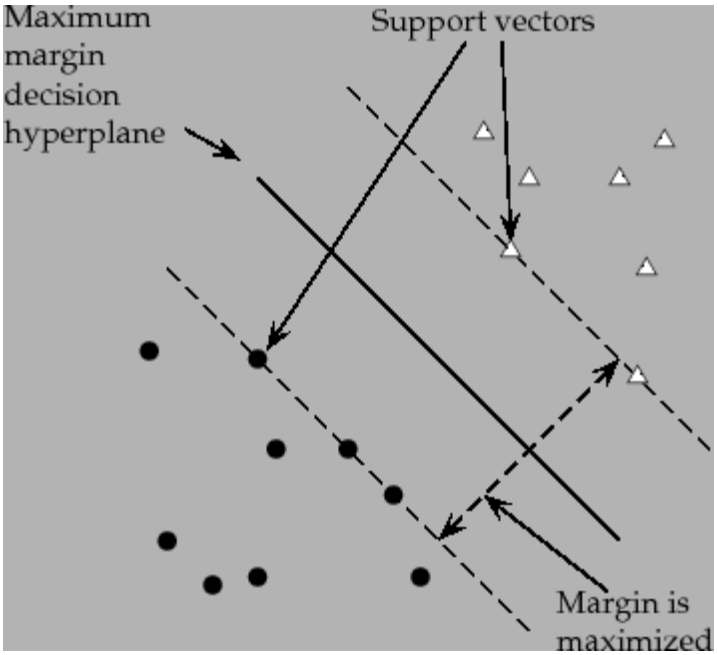


Εικόνα 4. 8 : Μορφή Πιθανοκρατικού νευρωνικού ταξινομητή [70]

V. Support Vector Machines (SVM) ταξινομητής.

Η βασική ιδέα αυτού του αλγορίθμου είναι ο μετασχηματισμός των αρχικών δεδομένων προς ταξινόμηση και η μεταφορά τους σε ένα νέο χώρο όπου εκεί θα είναι γραμμικά διαχωρίσιμα. Οι συναρτήσεις οι οποίες προβάλουν τα δεδομένα στο νέο χώρο ονομάζονται συναρτήσεις πυρήνα και η εύρεση κατάλληλων συναρτήσεων πυρήνα είναι ένα ανοικτό πεδίο μελέτης. Στην συνέχεια ο SVM δημιουργεί το χώρισμα μεταξύ των

κλάσεων προσπαθώντας να επιτρέψει την μέγιστη δυνατή απόσταση με τα κοντινότερα δείγματα τα οποία ονομάζονται διανύσματα υποστήριξης (support vectors).



Εικόνα 4. 9 : Διανύσματα υποστήριξης [71]

Η συνάρτηση διάκρισης του SVM είναι η:

$$D(x) = \sin(\sum_{i=1}^{N} a_i y_i K(x, x_i) + w) \quad (4.9)$$

όπου *x* είναι το άγνωστο δείγμα, x_i το i-οστό πρότυπο για την εκπαίδευση, y_i είναι η αντίστοιχη κλάση (-1,+1), *N* είναι ο αριθμός των δειγμάτων και *K* είναι η συνάρτηση πυρήνα.

Για την εύρεση της απόδοση του όλου συστήματος υπολογίστηκαν οι συχνότερα χρησιμοποιούμενες μετρικές της βιβλιογραφίας οι οποίες είναι οι Sensitivity, Specificity, και Accuracy και επεξηγούνται στο κεφάλαιο 3. Επιπλέον της χρήσης μεμονωμένων ταξινομητών για τον υπολογισμό της απόδοσης του συστήματος εφαρμόστηκε και ένα σύστημα συνδυασμού αυτών των ταξινομητών [72] έτσι ώστε να εξεταστεί εάν βελτιωνόταν η μεμονωμένη απόδοση κάθε ταξινομητή. Η τελική απόφαση για την ταξινόμηση ενός προτύπου προέκυπτε από τον συνδυασμό των αποφάσεων κάθε ταξινομητή με διάφορους τρόπους. Ποιο συγκεκριμένα, η τελική απόφαση προέκυπτε με βάση διάφορους κανόνες οι οποίοι ήταν οι εξής:

Ο κανόνας της πλειοψηφίας, για τον οποίο ισχύει :

$$G_r(X) = \sum_{i=1}^{R} d_{r,i}(X) \quad (4.10)$$

όπου r είναι η κλάση, *X* είναι το άγνωστο πρότυπο, *i=1,2,…R* είναι ο μονός αριθμός ταξινομητών που χρησιμοποιούνται (5 σε εμάς) και $d_{r,i}$ είναι η τιμή της απόφασης κάθε ενός ταξινομητή {0,1}. Για πρόβλημα δύο κλάσεων, εάν $G_1(X) > G_2(X)$, το άγνωστο X

ταξινομείται στην κλάση 1 διαφορετικά στην κλάση 2. Ουσιαστικά αυτός που μας λέει ο συγκεκριμένος κανόνας είναι ότι σε όποια κλάση αποφασίσει η πλειοψηφία των ταξινομητών ότι ανήκει το πρότυπο εκεί και θα ταξινομηθεί.

Ο κάθε ταξινομητής διαφέρει σε θεωρητικό επίπεδο με τον άλλον και για τον λόγο αυτό από κάθε ταξινομητή εκτός της τελικής του απόφασης {0,1} εξήχθηκε και η πιθανότητα με την οποία στέλνει το εκάστοτε πρότυπο σε μία κλάση. Προφανώς το άθροισμα των πιθανοτήτων για τις δύο κλάσεις είναι μονάδα και οι επόμενοι κανόνες που χρησιμοποιήθηκαν ήταν οι :

Κανόνας πολλαπλασιασμού :

$$G_r^{product}(X) = \prod_{i=1}^{R} P(r/X) \quad (4.11)$$

Κανόνας αθροίσματος:

$$G_r^{sum}(X) = \sum_{i=1}^{R} P(r/X) \quad (4.12)$$

Κανόνας μεγίστου:

$$G_r^{max}(X) = \max P(r/X) \quad (4.13)$$

Κανόνας ελαχίστου:

$$G_r^{min}(X) = \min P(r/X) \quad (4.14)$$

όπου *P(r/X)* η πιθανότητα το *X* να ανήκει στην κλάση *r*.

5. ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

5.1 Παράμετροι συστήματος

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, ο τρόπος με τον οποίον πειραματιστήκαμε όσον αφορά τις παραμέτρους είναι η αλλαγή κάθε φορά μιας παραμέτρου και η καταγραφή της τελικής απόδοσης του συστήματος. Στην συνέχεια όποια τιμή μας έδινε την βέλτιστη τελική απόδοση, αυτή και κρατάγαμε για την περαιτέρω επεξεργασία. Για τον έλεγχο της απόδοσης του συστήματος σε αυτό το στάδιο χρησιμοποιήθηκε η SFS μέθοδος επιλογής χαρακτηριστικών και η LOO μέθοδος αξιολόγησης με τον ταξινομητή Ελάχιστης Ευκλείδειας απόστασης, μετρώντας την συνολική ακρίβεια ταξινόμησης (Overal Accuracy).

PreProcessing/ Αφαίρεση baseline :

Τα πειράματα στο παρόν βήμα είχαν να κάνουν με την εύρεση των βέλτιστων παραμέτρων που δέχεται η συνάρτηση msbackadj του Matlab για την αφαίρεση της βασικής γραμμής. Οι παράμετροι εδώ αφορούν το μέγεθος του παραθύρου υπολογισμού της βασικής γραμμής, το βήμα κίνησης του παραθύρου (που εν τέλει θα τεθεί ίδιο με το μέγεθος του παραθύρου για να μην έχουμε επικάλυψη), η μέθοδος αναδρομής και η μέθοδος εκτίμησης της βασικής γραμμής. Ποιο συγκεκριμένα για το μέγεθος του παραθύρου είχαμε :

Πίνακας 5.1 : Αλλαγή μεγέθους παραθύρου υπολογισμού της βασικής γραμμής.

WindowSize	Accuracy (%)
20	90.57
50	89.62
100	85.85
120	83.96
150	87.74
180	91.51
200	90.57
250	93.40
280	92.45
300	89.62

220	91.51
230	***94.34***
240	92.45
260	90.57

Από τον παραπάνω πίνακα κρατάμε σαν βέλτιστη τιμή του μεγέθους του παραθύρου την τιμή 230 και στην πορεία πειραματιζόμαστε με το βήμα κίνησης αυτού του παραθύρου. Από τον επόμενο πίνακα φαίνεται σαν βέλτιστη τιμή η τιμή 200 ενώ η τιμή η οποία θα προτιμήσουμε να κρατήσουμε είναι η 230 ώστε να μην έχουμε αλληλοεπικάλυψη των παραθύρων και η οποία τα πηγαίνει εξίσου καλά, επιτυγχάνοντας απόδοση 92.45%.

Πίνακας 5.2 : Αλλαγή βήματος κίνησης παραθύρου υπολογισμού της βασικής γραμμής.

StepSize	Accuracy (%)
180	91.51
200	94.34
230	92.45
250	90.57
300	90.57
350	92.45

Επόμενη παράμετρος προς πειραματισμό είναι η μέθοδος αναδρομής για την οποία δοκιμάζονται οι shape-preserving piecewise cubic interpolation, η linear αλλά και η spline interpolation. Καλύτερα αποτελέσματα και άρα αυτή θα χρησιμοποιηθεί και στην συνέχεια είναι η pchip ή shape-preserving piecewise cubic interpolation με απόδοση 94.34%.

Πίνακας 5.3 : Αλλαγή μεθόδου αναδρομής

Regression Method		
Pchip	Linear	spline
94.34	89.62	93.40

Τελευταία παράμετρος που δοκιμάστηκε στο βήμα της αφαίρεσης της βασικής γραμμής είναι η μέθοδος εκτίμησής της και για την οποία εξετάστηκαν οι μέθοδοι quantile και EM, όπου γίνεται χρήση του Expectation – Maximazation αλγορίθμου. Τα αποτελέσματα της quantile ήταν αρκετά καλύτερα της μεθόδου EM καθώς παρατηρήθηκαν αποδόσεις 94.34 και 86.79% αντίστοιχα.

Πίνακας 5.4 : Αλλαγή μεθόδου εκτίμησης.

Estimation Method	
Quantile	EM
94.34	86.79

PreProcessing/ Εξισορρόπηση Φάσματος :

Εδώ εφαρμόστηκε η μέθοδος lowess για την οποία δοκιμάστηκαν η τάξη 0, γραμμικό ταίριασμα ή Lowess και η τάξη 1, τετραγωνικό ταίριασμα ή Loess. Επίσης πειραματιστήκαμε με τον παράγοντα span ο οποίος τροποποιεί το παράθυρο που εφαρμόζεται η εξισορρόπηση και μας δίνει ένα μέρος του αριθμού των τιμών των δεδομένων μας. Για παράδειγμα 0.005 span σημαίνει ότι το μέγεθος του παραθύρου θα είναι 0.50% του αριθμού των στοιχείων m/z. Επίσης πειραματιστήκαμε με την συνάρτηση πυρήνα της μεθόδου και για την οποία χρησιμοποιήθηκαν οι tricubic, gaussian και linear. Στους πίνακες που ακολουθούν παρουσιάζονται αυτά τα πειράματα.

Πίνακας 5.5 : Αλλαγή παραθύρου εφαρμογής εξισορρόπησης.

Span	Accuracy (%)
0.0009	85.85
0.001	**90.57**
0.002	**89.62**
0.003	88.68
0.005	88.68
0.007	78.30
0.008	74.53
0.01	75.47

Πίνακας 5.6 : Αλλαγή συνάρτησης πυρήνα

Kernel function		
tricubic	gaussian	linear
90.57	83.02	86.79

Πίνακας 5.7 : Αλλαγή τάξης

Order		
0	***1***	2
88.68	***90.57***	89.62

Αφαίρεση θορύβου και εξαγωγή χαρακτηριστικών

Στο συγκεκριμένο βήμα εξετάστηκε το μέγεθος του παραθύρου υπολογισμού του θορύβου. Σύμφωνα με την βιβλιογραφία [69] που μας έδειχνε σαν βέλτιστο μέγεθος το 25% των τιμών του m/z και λόγο του ότι στο σήμα μας αρχικά έχουμε m/z έως το 20.000 δοκιμάζουμε τιμές τριγύρω του 25% δηλαδή του 5.000. Τα αποτελέσματα, όπως φαίνονται και στον κάτωθεν πίνακα, επιβεβαιώνουν τις βιβλιογραφικές πηγές.

Πίνακας 5.8 : Αλλαγή μεγέθους παραθύρου υπολογισμού θορύβου

WindowSize	2000 m/z	80.19
	3000 m/z	78.30
	4000 m/z	86.79
	5000 m/z	**94.34**
	6000 m/z	79.25

Ευθυγράμμιση κορυφών (Peak Alignment)

Όσον αφορά το στάδιο αυτό, οι πειραματισμοί μας είχαν να κάνουν με το βέλτιστο δυνατό διάστημα τιμών m/z αλλά και το ποιος θα ήταν ο καταλληλότερος αντιπρόσωπος ενός τέτοιου διαστήματος με υποψήφιους το την μέση τιμή των εντάσεων εντός του διαστήματος και την μέγιστη τιμή των εντάσεων εντός του διαστήματος. Το διάστημα τιμών υπολογίζεται συναρτήσει της τρέχουσας τιμής του MZ όπως φαίνεται στην σχέση (5.1) με μια μεταβλητή shift που ορίζει το πόσο επηρεάζεται το μέγεθος αυτού του διαστήματος από την τρέχουσα τιμή MZ.

$$Interval = [MZ - \frac{1}{shift} * MZ \quad MZ + \frac{1}{shift} * MZ \ ] \tag{5.1}$$

Συνεπώς ο πειραματισμός μας αφορούσε την βέλτιστη δυνατή τιμή για την παράμετρο shift. Στο παρακάτω πίνακα παρατηρούμε την αλλαγή της μεταβλητής αυτής σε σχέση με την τελική απόδοση του συστήματος.

Πίνακας 5.9 : Αλλαγή μεταβλητής shift

Shift	1/400	85.85
	1/300	86.79
	1/200	87.74
	1/150	88.68
	1/120	**90.57**
	1/100	86.79
	1/80	86.79
	1/60	83.02

Από το πίνακα αυτό κρατήθηκε ως βέλτιστη τιμή, η τιμή 1/120 κάτι που είχε φανεί ενσωματωμένο και στην σχέση (4.2). Επόμενο στάδιο πειραματισμού είναι ποιος θα ήταν ο βέλτιστος αντιπρόσωπος κάθε διαστήματος και πειραματιστήκαμε μεταξύ της μέσης τιμής των εντάσεων εντός του διαστήματος (mean) και της μέγιστης τιμής (max) των εντάσεων εντός του συγκεκριμένου διαστήματος. Λόγο της φύσης του συγκεκριμένου πειράματος δοκιμάστηκαν όλοι οι ταξινομητές στα πλαίσια ενός προβλήματος και στο παρακάτω πίνακα παρουσιάζονται τα αποτελέσματά.

Πίνακας 5.10 : Εύρεση βέλτιστου αντιπροσώπου

MDC	Max	90.57	
	Mean	91.51	
KNN	1	**Max**	**95.28**
		Mean	94.34
	3	**Max**	**95.28**
		Mean	94.34
	5	Max	91.51
		Mean	94.4
	7	Max	94.34
		Mean	92.45
NAIVE	Max	87.74	
	Mean	90.57	
PNN	Max	85.85	
	Mean	85.85	
SVM	Max	90.57	
	Mean	90.57	

Από το παραπάνω πίνακα παρατηρούμε ότι παρόλο του ότι σε κάποια μεμονωμένα σημεία και ταξινομητές η μέση τιμή υπερτερεί σαν τελική απόδοση συστήματος, συνολικά η μέγιστη τιμή των εντάσεων εντός του διαστήματος επιτυγχάνει υψηλότερη απόδοση στο σύστημα (95.28%). Επομένως ως βέλτιστο αντιπρόσωπο του κάθε διαστήματος κρατάμε την μέγιστη τιμή των εντάσεων εντός του συγκεκριμένου διαστήματος.

5.2 Μέθοδοι που χρησιμοποιήθηκαν

Για την επιτέλεση της παρούσας εργασίας υλοποιήθηκε πλειάδα διαφορετικών τεχνικών για να προσεγγιστούν έτσι καλύτερα οι διάφορες πτυχές του προβλήματος. Όπως θα δούμε στην συνέχεια οι τεχνικές αυτές δοκιμάστηκαν πάνω σε 2 διαφορετικά σετ δεδομένων τα οποία και διασπάστηκαν σε υποπροβλήματα έτσι ώστε να αποτυπωθεί σωστότερα η προοπτική του προτεινόμενου συστήματος. Γενικότερα υλοποιήθηκε η leave one out μέθοδος και η οποία είναι αυτή που θα παρουσιαστεί αναλυτικότερα, η external cross validation μέθοδος, η οποία όμως δεν κατάφερε ικανοποιητικά αποτελέσματα και δεν παρουσιάζεται ευρέως και η 5 fold cross validation. Επίσης, για την εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών υλοποιήθηκαν οι exhaustive 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). Γενικότερα η exhaustive search, η αναζήτηση δηλαδή όλων των δυνατών συνδυασμών των χαρακτηριστικών θα ήταν η βέλτιστη λύση αλλά πρακτικά είναι κάτι μη δυνατό λόγο της πολυπλοκότητας και του υπολογιστικού χρόνου που απαιτεί. Συνεπώς γίνεται προσπάθεια να προσεγγιστεί η βέλτιστη αυτή λύση με άλλες ποιο αποδοτικές μεθόδους. Αρχικά γίνεται προσπάθεια με την SFS και την SBS και στην πορεία για καλύτερη προσέγγιση χρησιμοποιείται ευρέως η SFFS και η SBFS. Κάθε σετ δεδομένων χωρίζεται σε υποπροβλήματα, οπότε αρχικά θα παρουσιαστούν τα αποτελέσματα με παραπάνω από μία μέθοδο ενώ στην πορεία θα αρκεστούμε στις SFFS και SBFS οι οποίες παρέχουν μια ικανοποιητική προσέγγιση της εξαντλητικής αναζήτησης.

5.3 $1^{ο}$ σετ δεδομένων

Το πρώτο σετ δεδομένων που δοκιμάζεται είναι το (7-3-02) από τον δικτυακό χώρο του Εθνικού Καρκινικού Ινστιτούτου Αμερικής που δημιουργήθηκε χρησιμοποιώντας το πρωτεϊνικό chip H4. Έγινε χρήση του φασματογράφου μάζας τύπου SELDI και περιλαμβάνει 63 φάσματα τα οποία αντιστοιχούν σε υγιείς με (PSA<1), 190 φάσματα καλοήθειας με (PSA>4), 26 που προήλθαν από ασθενείς που έχουν επίπεδα PSA από 4 έως 10 (PSA 4-10) και 43 από ασθενείς με επίπεδο PSA>10. Κάθε ένα πρωτεωμικό φάσμα περιέχει 15.154 τιμές μάζας/φορτίου καθώς και τις αντίστοιχες τιμές έντασης. Για την διεξοδική έρευνα του συγκεκριμένου σετ δεδομένων, το διασπάμε σε 3 διαφορετικά κλιμακούμενης δυσκολίας υποπροβλήματα. Αρχικά στο υποπρόβλημα 1, εξετάζονται τα φάσματα από υγιείς (PSA<1) σε αντιδιαστολή με αυτά των ασθενών με (PSA>10). Στο υποπρόβλημα 2, στην κατηγορία των ασθενών προστίθενται και τα 26 φάσματα από ασθενείς που έχουν επίπεδα PSA από 4 έως 10 (PSA 4-10). Τέλος, στο υποπρόβλημα 3, έχουμε όλα τα δεδομένα του σετ καθώς στην κλάση των υγιών προστίθενται και τα 190 φάσματα καλοήθειας με (PSA>4). Εκ των προτέρων συνεπώς αναμένουμε το υποπρόβλημα 3 να είναι αυτό με τα ποιο πενιχρά αποτελέσματα καθώς το επίπεδο πολυπλοκότητας είναι πολύ μεγαλύτερο σε σχέση με τα άλλα υποπροβλήματα.

5.3.1 Υποπρόβλημα 1

Τα 63 φάσματα τα οποία αντιστοιχούν σε υγιείς με (PSA<1) έρχονται σε αντιδιαστολή με τα 43 από ασθενείς με επίπεδο PSA>10. Ως πρώτο λοιπόν πρόβλημα θα παρουσιαστούν οι περισσότερες μέθοδοι για την εύρεση του βέλτιστου συνδυασμού χαρακτηριστικών συνδυαζόμενοι με leave one out μέθοδο. Με μόνη διαφορά ότι θα εξεταστεί η εξαντλητική αναζήτηση για μέχρι πέντε μόνο χαρακτηριστικά. Θα παρουσιαστεί κάθε ταξινομητής ξεχωριστά και στο τέλος θα αναφερθούμε στο βέλτιστο εξ αυτών. Στον πίνακα 5.11 που ακολουθεί παρουσιάζονται τα αποτελέσματα για τον ταξινομητή ελάχιστης ευκλείδειας απόστασης ενώ στον 5.12 τα αποτελέσματα για τον ταξινομητή k κοντινότερων γειτόνων με k=1,3,5 και 7.

Πίνακας 5.11 : Ταξινομητής MDC – Υποπρόβλημα 1.

	Μέθοδος	Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC	SBS	84.9	90.7	69.9
	SFS	**90.6**	**100**	**84.1**
	SBFS	86.8	90.7	69.8
	SFFS	**90.6**	**100**	**84.1**
	EXH(5)	**90.6**	**100**	**84.1**

Πίνακας 5.12 : Ταξινομητής KNN – Υποπρόβλημα 1.

	Μέθοδος	k	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	SBS	1	96.2	93	93.7
		3	**96.2**	**93**	**95.2**
		5	95.3	97.7	92.1
		7	93.4	88.8	92.1
	SFS	1	95.3	97.7	93.7
		3	95.3	95.4	95.3
		5	91.5	90.7	92.1
		7	94.3	97.7	92.1
	SBFS	1	93	93	93
		3	***96.2***	***93***	***95.2***
		5	96.2	93	95.2
		7	93.4	88.8	92.1
	SFFS	1	95.3	97.7	93.7
		3	95.2	95.4	95.2
		5	91.5	90.7	92
		7	94.3	97.7	92.1
	EXH(5)	1	94.3	95.3	90.7
		3	**96.2**	**93**	**93.7**
		5	94.3	95.3	90.7
		7	93.4	94.4	90.7

Όπως προκύπτει από τον πίνακα 5.11 η βέλτιστη απόδοση του MDC εμφανίζεται για την SFS, την SFFS και την exhaustive(5) ενώ στην περίπτωση του KNN η βέλτιστη απόδοση εμφανίζεται για την SBS και την SBFS. Συνεχίζουμε με τους Naïve και PNN ταξινομητές στους πίνακες 5.13, 5.14 αντίστοιχα.

Πίνακας 5.13 : Naïve Bayes ταξινομητής – Υποπρόβλημα 1.

	Μέθοδος	Accuracy (%)	Sensitivity (%)	Specificity (%)
Bayessian	**SBS**	**92.5**	**90.7**	**85.7**
	SFS	87.7	95.4	82.5
	SBFS	**92.5**	**97.7**	**81**
	SFFS	87.7	95.4	82.5
	EXH(5)	88.68	94.3	82.5

Πίνακας 5.14 : PNN ταξινομητής – Υποπρόβλημα 1.

	Μέθοδος	Accuracy (%)	Sensitivity (%)	Specificity (%)
PNN	SBS	87.7	100	65.1
	SFS	85.9	83.7	87.3
	SBFS	87.7	100	65.1
	SFFS	85.9	83.7	87.3
	EXH(5)	95.3	96.1	94.3

Στο ταξινομητή Naïve Bayes παρατηρούμε βέλτιστη απόδοση 92.5% με 97.7% sensitivity με την μέθοδο επιλογής χαρακτηριστικών SBFS η οποία όμως επιτυγχάνει specificity μόνο 81%. Ταυτόχρονα η SBS στο ίδιο accuracy επιτυγχάνει καλύτερο ποσοστό (85.7%) στο specificity αλλά παράλληλα ελαττώνει το sensitivity στο 90.7%.

Όσον αφορά τον PNN, παρατηρούμε μεγάλη διαφορά στην εξαντλητική μέθοδο αναζήτησης σε σχέση με τις υπόλοιπες μεθόδους παρόλο το ότι της επιτρέπουμε να φτάσει μόνο μέχρι τα 5 χαρακτηριστικά. Ακρίβεια της τάξης του 87.7% με την SBFS εκτοξεύεται σε 95.3% (με sensitivity 96.1% και specificity 94.3%) για την εξαντλητική μέθοδο αναζήτησης. Λόγο αυτής της διαφοράς δοκιμάζουμε την εξαντλητική μέθοδο για τον συγκεκριμένο ταξινομητή και δίχως τον περιορισμό των 5 χαρακτηριστικών και φτάνουμε στο 98.1%.

Πίνακας 5.15 : PNN ταξινομητής (exhaustive) – Υποπρόβλημα 1.

PNN	
Exhaustive	***98.1***

Στην συνέχεια ακολουθεί ο SVM ταξινομητής με βέλτιστη απόδοση 94.3%.

Πίνακας 5.16 : SVM ταξινομητής – Υποπρόβλημα 1.

	Μέθοδος	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	SBS	91.3	92.3	90.7
	SFS	92.5	90.7	93.7
	SBFS	91.3	92.3	90.7
	SFFS	92.5	90.7	93.7
	EXH(5)	***94.3***	***95.1***	***93.2***

Στο πίνακα 5.17 παρουσιάζονται τα αποτελέσματα του συστήματος συνδυασμού των βέλτιστων αποτελεσμάτων των ταξινομητών τα οποία στην προκειμένη περίπτωση παρατηρούμε ότι δεν βελτιώνουν την βέλτιστη απόδοση.

Πίνακας 5.17 : Multiclassifier scheme – Υποπρόβλημα 1.

Multiclassifier scheme

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Majority vote	92.5	97.7	88.9
Product rule	92.5	97.7	88.9
Sum rule	***94.4***	***97.7***	***90.5***
Max rule	91.5	97.7	87.3
Min rule	91.5	97.7	87.3

Συνολικά λοιπόν για το υποπρόβλημα 1, έχουμε βέλτιστη απόδοση το 98.1% με χρήση της εξαντλητικής μεθόδου αναζήτησης χαρακτηριστικών και η οποία απόδοση επιτυγχάνεται από τον PNN ταξινομητή. Επίσης στην πλειοψηφία των ταξινομητών τα βέλτιστα αποτελέσματα επιτυγχάνονται με την SFFS ή την SBFS και συνεπώς αυτές είναι οι τεχνικές που θα χρησιμοποιηθούν στα επόμενα υποπροβλήματα. Για τον ταξινομητή PNN και μόνο, λόγο της ιδιαιτερότητας που παρουσίασε, θα δοκιμαστεί και η εξαντλητική μέθοδος αναζήτησης χαρακτηριστικών.

5.3.2 Υποπρόβλημα 2

Στο παρόν υποπρόβλημα, στην κατηγορία των ασθενών προστίθενται και τα 26 φάσματα από ασθενείς που έχουν επίπεδα PSA από 4 έως 10 (PSA 4-10). Συνεπώς πρώτη κλάση θα είναι οι υγιείς με PSA<1 ενώ κλάση δύο θα είναι οι ασθενείς που έχουν επίπεδα PSA από 4 έως 10 και οι ασθενείς με PSA>10. Στον πίνακα 5.18 παρουσιάζονται τα αναλυτικά αποτελέσματα των διάφορων ταξινομητών ενώ στο 5.19 τα αποτελέσματα του σχήματος συνδυασμού ταξινομητών.

Πίνακας 5.18 : Αναλυτικά αποτελέσματα - Υποπρόβλημα 2.

			Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC		SBFS	93.2	98.6	84.1
		SFFS	92.4	98.6	85.7
KNN	**1**	SBFS	96.2	95.7	96.8
		SFFS	96.2	98.6	93.7
	3	SBFS	94.7	92.8	88.9
		SFFS	***96.2***	***97.1***	***95.2***
	5	SBFS	96.2	94.2	96.8
		SFFS	***96.2***	***100***	***92.1***
	7	SBFS	95.5	97.1	92.1
		SFFS	93.2	97.1	88.9
Bayessian		SBFS	94	94.2	79.5
		SFFS	91.7	92.8	90.5
PNN		SBFS	94	95.7	84.1
		SFFS	91.7	95.7	87.3
		exhaustive	***97.7***	***100***	***94.2***
SVM		SBFS	93.2	94.3	90.7
		SFFS	94.7	95.7	93.7

Πίνακας 5.19 : Multiclassifier scheme – Υποπρόβλημα 2.

Multiclassifier scheme

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Majority vote	95.5	98.6	92.1
Product rule	70.5	100	38.1
Sum rule	70.5	100	38.1
Max rule	95.5	97.1	93.7
Min rule	95.5	97.1	93.7

Στο υποπρόβλημα 2 βλέπουμε ότι η απόδοση του συστήματος εν γένει είναι υψηλότερη κάτι που ασφαλώς ήταν αναμενόμενο μιας και αυξήθηκαν τα δεδομένα μας δίχως να έχουμε πολύ σχετικά μεταξύ τους φάσματα αν βασιστούμε στις μετρήσεις του PSA. Αν και εν γένει είχαμε υψηλότερες αποδόσεις σε σχέση με το υποπρόβλημα 1, η τελική βέλτιστη απόδοση είναι ελαφρώς μικρότερη φτάνοντας στο 97.7%. Για άλλη μια φορά η βέλτιστη απόδοση εμφανίζεται στο PNN ταξινομητή χρησιμοποιώντας την μέθοδο της εξαντλητικής αναζήτησης έχοντας παράλληλα sensitivity 100% και specificity 94.2%. Αυτήν την φορά ο KNN ταξινομητής για k=3 ακολουθάει κατά πόδας τον PNN, επιτυγχάνοντας με SFFS ακρίβεια 96.2% έχοντας παράλληλα sensitivity 97.1% και specificity 95.2% ενώ ο ίδιος ταξινομητής για k=5 επιτυγχάνοντας ακρίβεια 96.2% έχοντας παράλληλα sensitivity 100% και specificity 92.1%. Συνεπώς ανάλογα με τις απαιτήσεις μας επιλέγουμε κάθε φορά διαφορετικό βέλτιστο ταξινομητή ανάλογα αν επιδιώκουμε υψηλότερο sensitivity ή specificity. Σύμφωνα με τον πίνακα 5.19 ούτε τώρα επιτυγχάνουμε υψηλότερη απόδοση με το σύστημα συνδυασμού των βέλτιστων χαρακτηριστικών κάθε ταξινομητή.

5.3.3 Υποπρόβλημα 3

Σε αυτή την περίπτωση, όλα τα φάσματα του σετ δεδομένων χρησιμοποιούνται κάνοντας το πρόβλημά μας πιο σύνθετο μιας και έχουμε δεδομένα πολύ κοντινά όσον αφορά την τιμή του PSA. Συγκεκριμένα τώρα ως κλάση ένα έχουμε, εκτός των υγιών με PSA<1, και τα 190 φάσματα καλοήθειας με (PSA>4) ενώ στην κλάση δύο θα είναι οι ασθενείς που έχουν επίπεδα PSA από 4 έως 10 και οι ασθενείς με PSA>10.

Πίνακας 5.20 : Αναλυτικά αποτελέσματα - Υποπρόβλημα 3.

			Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC		SBFS	79.2	79.7	75.1
		SFFS	79.5	78.3	79.9
KNN	**1**	SBFS	85.1	46.4	88.6
		SFFS	85.4	63.8	91.3
	3	***SBFS***	***88.9***	***59.4***	***92.9***
		SFFS	*87.9*	*63.8*	*94.5*
	5	SBFS	88.8	56.5	94.5
		SFFS	*86.7*	*53.6*	*95.7*
	7	SBFS	87.6	52.2	94.5
		SFFS	87	53.6	96.1
Bayessian		SBFS	84.5	46.4	93.3
		SFFS	81.1	26.1	97.1
PNN		SBFS	82.9	84.1	67.6
		SFFS	83.2	53.6	91.3
SVM		SBFS	86.5	56.5	94.5
		SFFS	87.9	53.6	97.2

Πίνακας 5.21 : Multiclassifier scheme – Υποπρόβλημα 3.

Multiclassifier scheme

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Majority vote	***89.5***	***65.3***	***95.7***
Product rule	86.6	59.5	94.1
Sum rule	87.3	59.4	94.9
Max rule	86.1	59.5	93.3
Min rule	86	59.4	93.3

Όπως αναμενόταν, η απόδοση του συστήματος εν γένει πέφτει. Βέλτιστη απόδοση για το συγκεκριμένο υποπρόβλημα επιτυγχάνει ο ταξινομητής 3 κοντινότερων γειτόνων με SBFS μέθοδο φτάνοντας το 88.9% ακρίβεια με sensitivity 59.4% και specificity 92.9%. Επιπλέον για την συγκεκριμένη περίπτωση το σχήμα συνδυασμού ταξινομητών με βάση την πλειοψηφία βελτιώνει την τελική απόδοση σε 89.5% ακρίβεια.

5.4 2^{ο} σετ δεδομένων

Το δεύτερο σετ δεδομένων περιέχεται σε βιβλιοθήκη του R όπου έχει προστεθεί στις 08-02-2011 από τους Lixin Gong, William Constantine και Yu Alex Chen και είναι το σετ δεδομένων που χρησιμοποιήθηκε ι από την ομάδα των Adam και Qu [5-6] και έχει εξαχθεί από την ομάδα του ιατρικού κέντρου στη Virginia. Αφού αφαιρεθούν τα διπλότυπα τα οποία περιέχει καταλήγουμε σε ένα σετ από 81 δείγματα υγιών, 78 με καλοήθη όγκο και 168 με καρκίνο του προστάτη. Για την καλύτερη αποτύπωση των δυνατοτήτων του συστήματος σε αυτό το σετ δεδομένων για άλλη μια φορά το πρόβλημα διασπάστηκε σε 3 υποπροβλήματα. Αρχικά στο πρώτο υποπρόβλημα ως κλάση ένα έχουμε τους υγιείς και ως κλάση δύο τους ασθενείς. Στην συνέχεια στο υποπρόβλημα 2, οι έχοντες καλοήθη όγκο αντιμετωπίζονται ως υγιείς και τοποθετούνται στην κλάση ένα. Στο υποπρόβλημα 3, λόγο της διαφορετικής αντιμετώπισης από μερίδα άλλων ερευνητών, οι έχοντες καλοήθη όγκο τοποθετούνται στην κλάση των ασθενών έτσι ώστε σε πρώτη φάση να μπορούμε να εντοπίζουμε και αυτούς με υποψία καρκίνου. Η μέθοδος SFFS επέτυχε υψηλότερα αποτελέσματα στο συγκεκριμένο σετ δεδομένων σε σχέση με την SBFS και άρα αυτήν είναι που θα παρουσιαστεί στην συνέχεια.

5.4.1 Υποπρόβλημα 1

Όπως αναφέρθη το πρόβλημα αντιμετωπίζεται ως πρόβλημα δύο κλάσεων, στην πρώτη κλάση περιλαμβάνονται τα 81 δείγματα των υγιών και στην δεύτερη τα 168 δείγματα με καρκίνο του προστάτη.

Πίνακας 5.22 : Αναλυτικά αποτελέσματα - Υποπρόβλημα 1.

		Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC		***92***	***94.3***	***70.2***
KNN	**1**	89.2	90.3	80.2
	3	90.8	91.4	79.3
	5	***91.6***	***93.3***	***79.3***
	7	90.8	91.4	77.4
Bayessian		90.8	91.7	79.6
PNN		88.4	90.3	69.9
SVM		90.8	95.1	75.5

Τα αποτελέσματα είναι εν γένει ικανοποιητικά με τον ταξινομητή ελάχιστης ευκλείδειας απόστασης να επιτυγχάνει ελαφρώς καλύτερη απόδοση με 92% έναντι 91.6% του ταξινομητή πέντε κοντινότερων γειτόνων.

5.4.2 Υποπρόβλημα 2

Τα 78 δείγματα με καλοήθη όγκο προστίθενται στην ίδια κλάση με τα υγιεί δείγματα. Στο πίνακα που ακολουθεί παρατηρούμε την απόδοση του συστήματος στους διάφορους ταξινομητές.

Πίνακας 5.23 : Αναλυτικά αποτελέσματα - Υποπρόβλημα 2.

		Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC		71.6	79.2	59
KNN	**1**	69.5	70.2	58.6
	3	**72.5**	**83.9**	**59.7**
	5	72.2	74.6	69
	7	69.1	72.6	64.5
Bayessian		71.2	88.69	52.2
PNN		69.2	94.6	40.6
SVM		71.1	92.2	60.3

Τα αποτελέσματα πέφτουν αρκετά, κάτι που υποδεικνύει εμφανώς την διαφορά των καλοηθών δειγμάτων από τα υγιή. Στην προκειμένη περίπτωση βέλτιστη απόδοση επιτυγχάνεται με τον ταξινομητή 3 κοντινότερων γειτόνων.

5.4.3 Υποπρόβλημα 3

Σε αυτή την περίπτωση, τα καλοήθη δείγματα προστίθενται στην κατηγορία των καρκινικών δειγμάτων. Συνεπώς ως η μια κλάση είναι τα 81 δείγματα υγιών ενώ η δεύτερη θα είναι τα 78 με καλοήθη όγκο και τα 168 με καρκίνο του προστάτη.

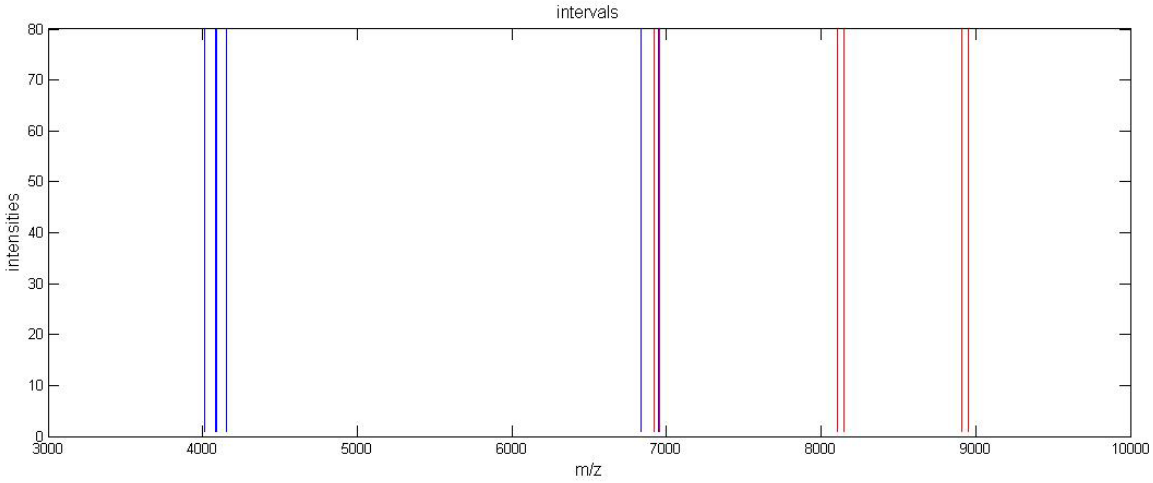
Πίνακας 5.24 : Αναλυτικά αποτελέσματα - Υποπρόβλημα 3.

		Accuracy (%)	Sensitivity (%)	Specificity (%)
MDC		91.8	93.9	82.7
KNN	**1**	91.8	92.4	81.3
	3	91.2	92.2	81.3
	5	***93.3***	***95.9***	***85.4***
	7	91.2	92.2	81.3
Bayessian		91.5	96.7	65.5
PNN		91.5	95.5	69.9
SVM		90.7	94.9	78.7

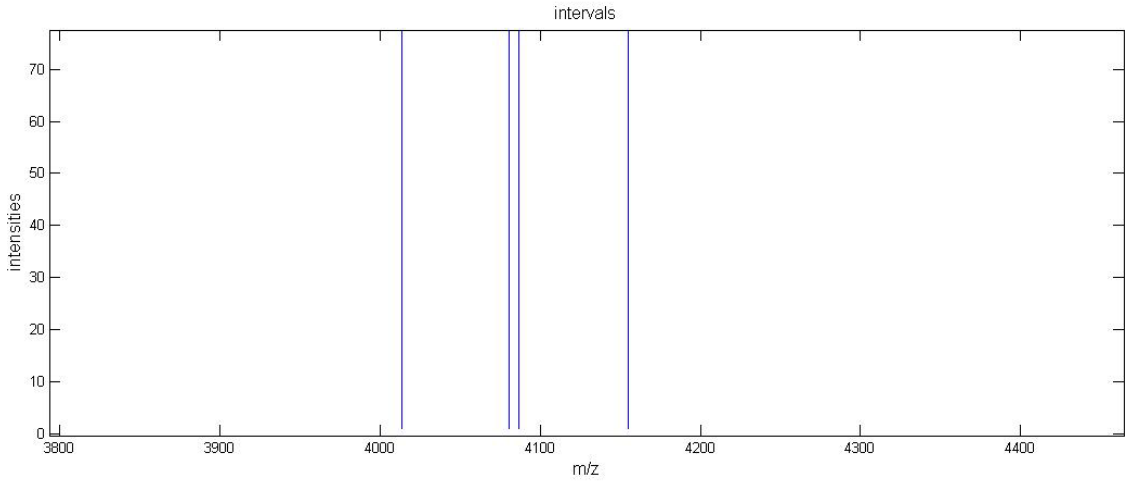
Στο παραπάνω πίνακα, παρατηρούμε την απόδοση του συστήματος να ανεβαίνει κάτι που προφανώς υποδυκνύει ότι τα καλοήθη δείγματα είναι κοντύτερα στα μη υγιή παρά στα υγιή δείγματα. Μέγιστη απόδοση επιτυγχάνει ο ταξινομητής πέντε κοντινότερων γειτόνων, όπως και στο υποπρόβλημα ένα, αυξάνοντας την ακρίβεια από 91.6% σε 93.3%.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ / ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

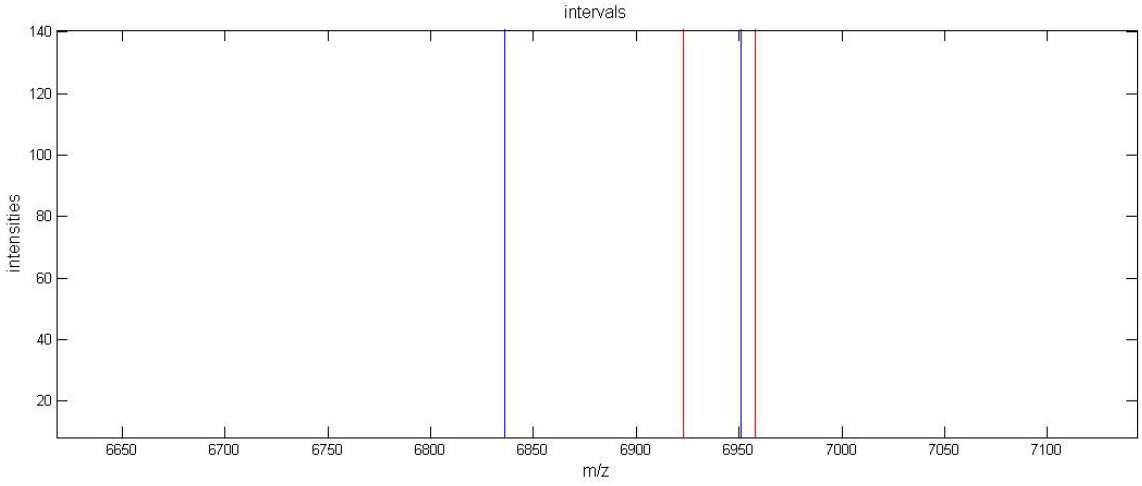
Ευρύτερος στόχος της παρούσας διπλωματικής εργασίας ήταν η εύρεση υποψήφιων διαστημάτων m/z όπου θα ήταν πιθανότερο να εντοπιστούν βιοδείκτες συνδεδεμένοι με τον καρκίνο του προστάτη. Κάθε ταξινομητής καταλήγοντας σε μια βέλτιστη απόδοση επέστρεφε στο σύστημα τα χαρακτηριστικά διαστήματα που χρησιμοποίησε ώστε να επιτύχει την συγκεκριμένη βέλτιστη απόδοση. Έχοντας λοιπόν τα διαστήματα τα οποία χρησιμοποίησαν κατά κόρον οι ταξινομητές όταν επετύγχαναν την βέλτιστη απόδοσή τους, καταλήξαμε σε κάποια τελικά προτεινόμενα διαστήματα. Όσον αφορά το πρώτο σετ δεδομένων, τα διαστήματα [4087-4155] και [6836-6951] χρησιμοποιήθηκαν από όλους τους ταξινομητές όταν αυτοί επέτυχαν το βέλτιστό τους αποτέλεσμα. Επίσης το διάστημα [4014-4081] χρησιμοποιήθηκε από τους MDC, KNN, Bayessian, PNN με μόνο τον SVM να μην τον χρησιμοποιεί στο βέλτιστο αποτέλεσμά του. Επίσης για το δεύτερο σετ δεδομένων, τα διαστήματα [6923-6958], [8107-8148] και [8910-8955] χρησιμοποιήθηκαν από όλους τους ταξινομητές κατά το βέλτιστο αποτέλεσμά τους. Στην παρακάτω εικόνα φαίνονται τα διαστήματα αυτά με μπλε για το πρώτο σετ δεδομένων και με κόκκινο για το δεύτερο. Στις εικόνες 6.2 και 6.3 ζουμάρουμε επιπλέον ώστε να γίνει ποιο κατανοητό αφενός το πόσο κοντά βρίσκονται 2 εκ των 2 διαστημάτων του πρώτου σετ αλλά και για να δούμε την αλληλεπικάλυψη ενός διαστήματος από το πρώτο σετ με ένα από το δεύτερο πράγμα που το ενισχύει ως υποψήφιο να περιέχει βιοδείκτες για το καρκίνο του προστάτη.



Εικόνα 6. 1 : Προτεινόμενα διαστήματα για εύρεση βιοδεικτών



Εικόνα 6. 2 : Δύο σχεδόν εφαπτόμενα προτεινόμενα διαστήματα.



Εικόνα 6. 3: Αλληλοεπικαλυπτόμενα διαστήματα πρώτου (μπλε) και δεύτερου(κόκκινο) σετ δεδομένων.

Αν συνδυάσουμε τα παραπάνω αποτελέσματα με τον πίνακα 4.1, όπου περιέχονται οι προτεινόμενοι από άλλες επιστημονικές ομάδες βιοδείκτες παρατηρούμε ότι εντός των διαστημάτων μας βρίσκονται οι :

- 8141 από την ομάδα του Adam [54]
- 4080,6949 από άλλη έρευνα της ίδιας ομάδας [55] αλλά και την ομάδα του Wagner [64]

πράγμα που ενισχύει την άποψή μας ότι αυτά τα διαστήματα πρέπει να εξεταστούν περαιτέρω. Θα πρέπει να αναφέρουμε επίσης ότι τα φάσματα μελετήθηκαν για τιμές mz μεγαλύτερες του 1500 λόγω το θορύβου που περιέχουν οι τιμές κάτω αυτής.

Στην παρούσα εργασία, ο πειραματισμός σχετικά με την προεπεξεργασία σημάτων εστιάστηκε στην εύρεση των βέλτιστων παραμέτρων είτε για την αφαίρεση της βασικής γραμμής είτε για την εξισορρόπηση των φασμάτων σε ήδη υπάρχουσες συναρτήσεις της βιβλιοθήκης του Matlab. Συνεπώς ένα επιπλέον βήμα θα ήταν η δημιουργία νέων, ποιο αποδοτικών ίσως συναρτήσεων που πιθανόν να ακολουθούν κάποιο διαφορετικό μαθηματικό μοντέλο. Επιπλέον, η δοκιμή του συστήματος σε νέα σετ δεδομένων αλλά και με περισσότερους ταξινομητές είναι πιθανόν να αποδειχθεί προς όφελος της έρευνας. Στην έρευνα μας, υλοποιήθηκε και ένας μηχανισμός συνδυασμός ταξινομητών που όμως μόνο σε μία περίπτωση μας έδωσε καλύτερα αποτελέσματα. Ο μηχανισμός αυτός συνδύαζε και τους πέντε υλοποιημένους ταξινομητές με διάφορους τρόπους. Θα ήταν δυνατόν ο μηχανισμός αυτός να υλοποιηθεί αποδοτικότερα, επιτρέποντας και τον συνδυασμό λιγότερων από τους πέντε ταξινομητές εάν αυτό επέφερε καλύτερη απόδοση.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Mass spectrometry	Φασματομετρια μάζας
Biomarkers	Βιοδείκτες
Smoothing	Εξομάλυνση
Baseline	Βασική γραμμή
Digital rectal examination	Δαχτυλική εξέταση ορθού
TransRectal Examination	Διορθικό υπερηχογραφημμα
Prostatic specific antigen	Ειδικό προστατικό αντιγόνο
Greason grading	Βαθμονόμηση gleason
2D gel electrophoresis	Δισδιάστατη ηλεκτροφόρηση πηκτωμάτων
Liquid chromatography	Υγρή χρωματογραφία
Electrospray ionization	Ιονισμός με ηλεκτροψεκασμό
Quadropole analyzer	Τετράπολος αναλυτής
Quadropole ion trap	Τετράπολη παγίδα ιόντων
Time of flight analyzer	Αναλυτής χρόνου πτήσης
Fourier transform ion cyclotron	Ιοντικό κύκλοτρο μετασχηματισμού Fourier
Accuracy	Ακρίβεια
Sensitivity	Ευαισθησία
Specificity	Σαφήνεια

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

MS	Mass spectrometry
2DGE	2D gel electrophoresis
EI	Electrospray ionization
LC	Liquid chromatography
DRE	Digital rectal examination
PSA	Prostatic specific antigen
MALDI	Matrix assisted laser desorption ionization
TOF	Time of flight
EM	Expectation - Maximazation
MDC	Minimum distance classifier
KNN	K nearest neighbours
SVM	Support vector machines
PNN	Probabilistic neural network
SFS	Sequential forward selection
SBS	Sequential backward selection
SFFS	Sequential forward floating selection
SBFS	Sequential backward floating selection

ΑΝΑΦΟΡΕΣ

[1] Mean's Health, http://men.webmd.com/picture-of-the-prostate [Προσπελάστηκε 23/11/11]
[2] National Cancer Institute,http://www.cancer.gov/cancertopics/types/prostate [Προσπελάστηκε 23/11/11]
[3] Cancer Research Institute in UK, http://info.cancerresearchuk.org/cancerstats/types/prostate/ [Προσπελάστηκε 23/11/11]

[4] Prostate Cancer Screening, Humana Press 2009

[5] MediciNet, http://www.medicinenet.com [Προσπελάστηκε 23/11/11]

[6] American Urological Association. (2009). *Prostate Specific Antigen Best Practice Statement: 2009 update*. Education and Research. Inc.

[7] Macmillan Cancer Support http://www.macmillan.org.uk/Home.aspx [Προσπελάστηκε 23/11/11]
[8] Biocompatibles, http://www.brachysciences.com/patients-screening.asp [Προσπελάστηκε 23/11/11]
[9] RadiologyInfo, http://www.radiologyinfo.org [Προσπελάστηκε 23/11/11]
[10] The Veteran's Administration Cooperative Urologic Research Group: histologic grading and clinical staging of prostatic carcinoma. In Tannenbaum M (ed.) Urologic Pathology: The Prostate. Lea and Febiger, Philadelphia, 1977, pp.171-198.
[11] Σοφία Κοσσίδα, Βιοπληροφορική, δυνατότητες και προοπτικές.

[12] State Agricultural Biotechnology Centre, http://www.sabc.murdoch.edu.au/facilities/proteomics.html [Προσπελάστηκε 23/11/11]
[13] PNAS, http://www.pnas.org/ [Προσπελάστηκε 23/11/11]
[14] Unit of Medical Technology & Intelligent Information Systems (MedLab), http://medlab.cs.uoi.gr/ [Προσπελάστηκε 23/11/11]
[15] New Oblective, http://www.newobjective.com/electrospray/, [Προσπελάστηκε 23/11/11]
[16] M.Karas,D. Bachmann, U.Bahr, and F.Hillenkamp (1987) *Int. J. Mass Spectrom. Ion Processes*, 78, 53.
[17] M. Karas and F.H Hillenkamp(1988) *Anal. Chem.*,60,pp 2229.

[18] F.Hillenkamp, M.Karas, A.Ingeldoh and B.Stahl (1990) Matrix assisted UV-laser desorption ionization: a new approach to mass spectrometry of large molecules, in *Biological Mass Spectrometry,* Elsevier, Amsterdam, p. 49.
[19] R.Westermeier, T.Naven, H.R. Hopker, Proteomics in Practice : A Guide to Successful Experimental Design, 2nd, Completely Revised Edition
[20] W.Paul and H.S.Steinwedel. A New Mass Spectrometer without a Magnetic Field. *Z. Naturforsch*, 1953, 8A, pp 448-450.
[21] RE.Finnigan, Quadrupole Mass Spectrometers: From development to commercialization. *Anal. Chem.* 1994, 66, 969A-975A

[22] Atmosperic Experiment Laboratory, Nasa, http://huygensgcms.gsfc.nasa.gov/MS_Analyzer_1.htm [Προσπελάστηκε 23/11/11]
[23] P.D.von Haller, S.Donohoe, D.R.Goodlett, R.Aebersold, J.D.Watts. Proteomics 1 (2001) pp.1010–1021.
[24] W.Paul and H.S.Steinwedel, US Patent, 1960
[25] J.F.J.Todd, Ion trap mass spectrometer - past, present, and future (?) *Mass Spectrom. Rev.* 1991, 10 (1), pp.3-52
[26] G.C.Stafford, P.E.Kelley, J.E.Syka, W.E.Reynolds and J.F.J.Todd, Recent improvements in and analytical applications of advanced ion trap technology. *Int. J. Mass Spectrom. Ion Processes*, 1984, 60, pp.85-98.

[27] A.Makarov. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis., 2000,*Analytical Chemistry : AC* 72 (6): pp.1156–62.
[28] H.Qizhi, J.Robert, L.Hongyan, A.Makarov, M.Hardman, R Graham Cooks. The Orbitrap: a new mass spectrometer,J. Mass Spectrom. 2005; 40: pp.430–443
[29] W.E.Stephens, A Pulsed Mass Spectrometer with Time Dispersion *Phys. Rev.*, 1946, *69*, 691.
[30] W.C.Wiley, I.H.McLaren, Time-of-Flight Mass Spectrometer with Improved Resolution, 1955, *Review of Scientific Instruments* 26: 1150.
[31] Agricultural Research Service, http://www.ars.usda.gov/main/main.htm [Προσπελάστηκε 23/11/11]
[32] K.F.Medzihradszky, J.M.Campbell, M.A.Baldwin, A.M.Falick, P.Juhasz, M.L.Vestal, A.L.Burlingame. Anal Chem 72 (2000).pp. 552–558.

[33] J.A.Hipple, H.Sommer and H.A.Thomas. A Precise Method of Determining the Faraday by Magnetic Resonance. *Phys. Rev*. 1949, 76, pp.1877-1878.

[34] Ινστιστούτο Πυρηνική τεχνολογίας, http://www.itn.pt/, [Προσπελάστηκε 23/11/11]
[35] D.W.Koppenaal, C.J.Barinaga, M.B.Denton, Mass spectrometry detectors, 2005, *Anal. Chem.*, pp. 419A–27A.
[36] E.De Hoffman V.Stroobant, Mass Spectrometry: Principles and Applications, Wiley & Sons Eds., 3rd Ed., West Sussex, England, 2002

[37] I.Levner, Feature selection and nearest centroid classification for protein mass spectrometry, BMC Bioinformatics, 2005, 6 (68).
[38] A. Barla, G. Jurman, S. Riccadonna, S. Merler, M. Chierici, and C. Furlanello. Machine learning methods for predictive proteomics. *Briefings in Bioinformatics*, 2008, 9(2):pp.119-128.

[39] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q-T.Le. Sample classification from protein mass spectrometry, by 'peak probability contrasts. *Bioinformatics*,2004, 20(17):3034-3044
[40] H. W. Ressom, R. S. Varghese, M. Abdel-Hamid, S. A-L. Eissa, D. Saha, L. Goldman, E. F. Petricoin, T. P. Conrads, T. D. Veenstra, C. A. Loffredo, and R. Goldman. Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics*, 2005, 21(21):4039-4045.
[41] H. W. Ressom, R. S. Varghese, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, 23(5):619-626, 2007.
[42] J. B.Breen, F. G.Hopwood, K. L.Williams, and M. R.Wilkins, Automatic Poisson peak harvesting for high throughput protein identification, *Electrophoresis*, 21:2243-2251, 2000.

[43] A. C.Sauve and T. P.Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data, In *Proceedings of the Workshop on Genomic Siganl Processing and Statistics*, 2004.
[44] K. Noy and D. Fasulo. Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, 23(19):2528-2535, 2007.

[45] Li M, Chen S, Zhang J, Chen H, Shyr Y. Wave-spec: a preprocessing package for mass spectrometry data *Bioinformatics (2011) 27(5): 739-740 first published online January 5, 2011*
[46] Yuliya V. Karpievitch et all, PrepMS: TOF MS Data Graphical Preprocessing Tool, Bioinformatics Advance Access published November 22, 2006

[47] Chao Yang, Zengyou He and Weichuan Yu, «Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis», BMC Bioinformatics 2009
[48] X.Wang, W.Zhu, K.Pradhan, C.Ji, Y.Ma, O.J.Semmes, J.Glimm and J.Mitchell, Feature extraction in the analysis of proteomic mass spectra, Proteomics, 2006, 6, pp. 2095–2100.
[49] B.Wu, T.Abbott, D.Fishman, W.McMurray, *et al.*, Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics (Oxford, England)* 2003, *19*, 1636-1643.
[50] E.Petricoin, A.M.Ardekani, B.A.Hitt, P.J.Levine, *et al.*, Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002, *359*, 572-577.
[51] I.Levner, Feature selection and nearest centroid classification for protein mass spectrometry, 2005, BMC Bioinformatics, 6 (68).
[52] D.K.Ornstein, W.Rayford, V.A.Fusaro, T.P.Conrads, *et al.*, Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml. *The Journal of urology* 2004, *172*, 1302-1305.

[53] Y.Yasui, M.Pepe, M.L.Thompson, B.L.Adam, G.L.Jr.Wright, Y.Qu, J.D.Potter, M.Winget, M.Thornquist, and Z.Feng, A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, Biostatistics, 2004, 4, pp. 449-463.
[54] B.L.Adam, Y.Qu, J.W.Davis, M.D.Ward, M.A.Clements, L.H.Cazares, O.J.Semmes, P.F.Schellhammer, Y.Yasui, Z.Feng and J.L.Jr.Wright, Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, Cancer research, 2002, 62, pp. 3609-3614.
[55] Y.Qu,B.L.Adam, J.W.Davis, M.D.Ward, M.A.Clements, L.H.Cazares, O.J.Semmes, P.F.Schellhammer, Y.Yasui, Z.Feng and J.L.Jr.Wright, Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, Clinical Chemistry, 2002, 48, pp. 1835-1843.
[56] H.Zhu, C.Y.Yu, H.Zhang, Tree-based disease classification using protein data. *Proteomics* 2003, *3*, 1673-1677.

[57] M.Hilario, A.Kalousis, M.Muller, C.Pellegrini, Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics* 2003, *3*, 1716-1719
[58] W.Zhu, X.Wang, Y.Ma, M.Rao, *et al.*, Detection of cancer-specific markers amid massive mass spectral data. *Proceedings of the National Academy of Sciences of the United States of America* 2003, *100*, 14666-14671.
[59] B.Wu, T.Abbott, D.Fishman, W.McMurray, *et al.*, Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics (Oxford, England)* 2003, *19*, 1636-1643.
[60] Hallem J.Issaq et all, Seldi –TOF MS for Diagnostic Proteomics, Analytical Chemistry 2003
[61] S.Vonderwulbecke et all, Protein quantification by the SELDI – TOF – MS – based ProteinChip System, Nature Methods, May 2005
[62] E.P.Diamandis,Point: Proteomic Patterns in Biological Fluids: Do They Represent the Future of Cancer Diagnostics?, *Clinical Chemistry, 2003,* 49:8, pp. 1272–1278.
[63] E.Petricoin et all, Serum proteomic patterns for detection of prostate cancer, Journal of the National Cancer Institute, 2002, 94, pp. 1576-1578.
[64] M.Wagner, D.N.Naik, A.Pothen, S.Kasukurti, R.R.Devineni, B.L.Adam, J.Semmes and G.L.Jr.Wright, Computational protein biomarker prediction: a case study for prostate cancer, BMC Bioinformatics, 2004, 5 (26).
[65] S.Lehrer et all, Putative protein markers in the sera of men with prostatic neoplasms, BJU INTERNATIONAL 2003
[66] K.Jong, E.Marchiori, A.Vaart, A., Analysis of proteomic pattern data for cancer detection. *Lecture notes in computer science (Lect. notes comput. sci.)* 2004, *3005*, 41-51.
[67] L.L.Banez, P.Prasanna, L.Sun, A.Ali, Z.Zou, B.L.Adam, et al. Diagnostic potential of serum proteomic patterns in prostate cancer. J Urol 2003; 170:442-6.
[68] Βιβλιοθήκη R, http://cran.r-project.org/web/packages/msProstate/index.html, [Προσπελάστηκε 23/11/11]
[69] Wang et all, Feature extraction in the analysis of proteomic mass spectra, Proteomics, 6, 2006, pp. 2095–2100
[70] R.Voss, P.Cullen, H.Schulte, G.Assmann, Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Munster Study (PROCAM) using neural networks, International Journal of Epidemiology 2002; 31:1253-1262
[71] The Stanford NLP Group, http://nlp.stanford.edu/IR-book/html/htmledition/support-vector-machines-the-linearly-separable-case-1.html, [Προσπελάστηκε 23/11/11]
[72] A.Daskalakis, S.Kostopoulos, P.Spyridonos, D.Glotsos, P.Ravazoula, M.Kardari, I.Kalatzis, D.Cavouras and G.Nikiforidis, Design of a multi-classifier system for discriminating benign from malignant thyroid nodules using routinely H&E-stained cytological images, Computers in Biology and Medicine, 38, 2008, pp. 196 – 203.