



ΕΘΝΙΚΟ ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΧΗΜΕΙΑΣ



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ «ΧΗΜΕΙΑΣ»
ΕΙΔΙΚΕΥΣΗ « Χημική Ανάλυση και Έλεγχος Ποιότητας»**

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

**Αξιολόγηση χημειομετρικών τεχνικών για την ανάλυση
τάσεων από μη στοχευμένη LC-Q-TOFMS ανάλυση λυμάτων**

**ΑΛΥΓΙΖΑΚΗΣ ΝΙΚΗΦΟΡΟΣ
ΧΗΜΙΚΟΣ**

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2015

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Αξιολόγηση χημειομετρικών τεχνικών για την ανάλυση τάσεων από μη στοχευμένη LC-Q-TOFMS ανάλυση λυμάτων

ΟΝΟΜΑΤΕΠΩΝΥΜΟ

ΑΛΥΓΙΖΑΚΗΣ ΝΙΚΗΦΟΡΟΣ

A.M.: 131301

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

Θωμαΐδης Νικόλαος, Αναπληρωτής Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Ευσταθίου Κωνσταντίνος, Καθηγητής ΕΚΠΑ

Κουππάρης Μιχαήλ, Καθηγητής ΕΚΠΑ

Θωμαΐδης Νικόλαος, Αναπληρωτής Καθηγητής ΕΚΠΑ

ΗΜΕΡΟΜΗΝΙΑ ΕΞΕΤΑΣΗΣ 20/07/2015

ΠΕΡΙΛΗΨΗ

Καθημερινά αναδύόμενοι ρύποι εισέρχονται στο αποχετευτικό σύστημα σε μεγάλες ποσότητες. Τα επίπεδα συγκεντρώσεων των διαφόρων ρύπων που εισάγονται στα εισερχόμενα λύματα δεν παραμένουν σταθερά κατά τη διάρκεια μιας χρονικής περιόδου (για παράδειγμα ανά ημέρα) αλλά διακυμαίνονται και ακολουθούν διαφορετικές τάσεις και διαφορετικά προφίλ. Αυτές οι διακυμάνσεις επηρεάζονται από πολλούς παράγοντες, όπως οι χρήσεις αυτών των ρύπων από τον άνθρωπο, για παράδειγμα φαρμάκων από ασθενείς. Η έρευνα της διακύμανσης των συγκεντρώσεων τόσο με στοχευμένο όσο και με μη στοχευμένο τρόπο στα εισερχόμενα υγρά απόβλητα είναι σημαντική, διότι μπορεί να χρησιμοποιηθεί ως ένα σύστημα προειδοποίησης των επιπέδων ρύπανσης, μπορεί να παράσχει βαθύτερη κατανόηση της συμπεριφοράς των οργανικών ρύπων στο περιβάλλον και επίσης μπορεί να οδηγήσει σε καλύτερη κατανόηση της χρήσης διαφόρων χημικών ουσιών από την κοινωνία.

Ο στόχος της παρούσας εργασίας είναι η ανάπτυξη μιας ημιαυτοματοποιημένης διαδικασίας για την ανίχνευση τάσεων στις εντάσεις των ουσιών που ανιχνεύονται μεταξύ διαφορετικών χρονικών περιόδων, μέσω της χρήσης μιας διαδικασίας βασισμένης σε αλγόριθμους από τρία πακέτα της στατιστικής γλώσσας R (XCMS, CAMERA, TIMECOURSE). Το αποτέλεσμα της διαδικασίας αυτής είναι μια λίστα ιόντων που επέδειξαν σημαντική διακύμανση κατά χρονική περίοδο λήψης δειγμάτων. Στη συνέχεια πραγματοποιήθηκε ταυτοποίηση μερικών από αυτές τις ενώσεις με τη χρήση μιας σειράς εργαλείων και μεθοδολογιών μη στοχευμένης ανάλυσης.

Για το σκοπό αυτό χρονικές ακολουθίες δειγμάτων συλλέχθηκαν από το κέντρο επεξεργασίας λυμάτων της Αθήνας. Η ανάλυση τους πραγματοποιήθηκε με εκχύλιση στερεής φάσης και υγροχρωματογραφία συνδεδεμένη με υψηλής διακριτικής ικανότητας φασματομετρία μάζας (LC-HRMS). Πρώτα, έγινε η μετατροπή των δεδομένων που προέκυψαν σε mzXML αρχεία με τη χρήση του λογισμικού ProteoWizard τα οποία αποθηκεύτηκαν σε υποφακέλους στη διεύθυνση κλήσης της R. Η αναζήτηση ιόντων πραγματοποιήθηκε με τη χρήση του αλγορίθμου centWave με βελτιστοποιημένους παραμέτρους για τα δεδομένα που προέκυψαν από το σύστημα LC-HRMS. Τα ιόντα που αναπαριστούν τον ίδιο αναλύτη σε όλα τα δείγματα τοποθετούνται μαζί, ακολουθεί ευθυγράμμιση των χρωματογραφημάτων για διόρθωση τυχόν ολισθήσεων, και συμπλήρωση των τιμών των εντάσεων για τους αναλύτες που δεν ανιχνεύθηκαν σε κάποια δείγματα με κάποια χαμηλή τιμή έντασης. Έπειτα τα ιόντα ομαδοποιούνται

σύμφωνα με συντελεστή που αποδίδει το σχήμα της κορυφής και σύμφωνα με το χρόνο ανάσχεσης. Τέλος εντός των ομάδων που προέκυψαν αναζητούνται οι ισοτοπικές κορυφές και τα ιόντα προσθήκης και αποδίδονται στον αντίστοιχο αναλύτη. Η λίστα μαζών με τα εμβαδά ολοκλήρωσης των κορυφών για κάθε δείγμα υπόκειται σε στατιστική δοκιμασία Multivariate Empirical Bayes Approach, η οποία θέτει σε προτεραιότητα τα ιόντα-μάζες σύμφωνα με το συντελεστή Hotelling T². Τα ιόντα που εμφανίζουν αξιοσημείωτη τάση απεικονίζονται αυτόματα γραφικά και κατατάσσονται με βάση αυτόν τον συντελεστή. Έπειτα αναζητείται η ταυτότητα των ιόντων-ουσιών αυτών ακολουθώντας μια μεθοδολογία ταυτοποίησης από τα δεδομένα HRMS.

Η παρούσα προσέγγιση καθιστά εφικτή την καταγραφή των τάσεων συγκέντρωσης για μεγάλο αριθμό συστατικών σε ένα σετ δειγμάτων. Επιπλέον, αυτή η μεθοδολογία μπορεί να ανιχνεύσει γεγονότα απευθείας διάθεσης χημικών στο σύστημα αποχέτευσης και για αυτό αποτελεί πολύτιμη πηγή πληροφορίας για της αρχές των κέντρων επεξεργασίας λυμάτων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Αναλυτική Χημεία/Φασματομετρία μαζών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ανάλυση τάσεων, υπολογιστικά εργαλεία, στατιστική γλώσσα R, αναδυόμενοι ρύποι, μη στοχευμένη ανάλυση με φασματομετρία μαζών, λύματα

ABSTRACT

Contaminants of emerging concern enter daily into the sewage system in large quantities. Concentration levels of the different compounds are not constant, and may follow different trends and patterns, affected by several factors (like usage trends, or pollution spills). The investigation of the concentration pattern of both target and non-target substances in wastewater is an important issue, as it could be used as an early-warning system on pollution loads, it may provide a further insight on the behavior of organic contaminants in the environment, and also information about the community use of chemicals.

The objective of the presented study is to develop an semi-automated workflow for the detection of trends in intensities for all the detected substances among different sampling sets (e.g. time periods) through the use of a workflow based on R algorithms from the packages XCMS, CAMERA and TIMECOURSE. The output of this workflow is a list of compounds with high variation during the sampling period extracts. The identity of those compounds was searched following non target procedures and techniques.

Temporal sequences of samples were collected from the wastewater treatment plant of Athens. Analysis was carried out by liquid chromatography - high resolution tandem mass spectrometry (LC-HRMS). First, the acquired data is converted to mzXML using ProteoWizard and stored in subfolders in the R working folder. Sample feature detection is performed by the centWave algorithm with optimized parameters for QTOF MS data. Features representing the same analyte across samples are placed into groups, peak alignment follows, missing features are filled with a low intensity value and then, features are clustered according to peak shape correlation coefficients and retention times. Finally, isotopic peaks and adducts are annotated to the same chemical component and its monoisotopic peak. The peak table with the integrated peak areas in each sample is used in order to perform Multivariate Empirical Bayes Approach which results in providing prioritized peaks (using Hotelling T2 coefficient) according to the differences of integrated intensities among the studied groups. Moreover, it automatically produces plots to evaluate the trend of each substance. After that, the most relevant peaks can be tentatively explored using non-target identification approaches.

The presented approach enables the recording of concentration trends for a large number of compounds in a given set of samples. Moreover, this workflow can be used to detect

events of direct disposal of some specific substances into the sewage system, constituting an appropriate source of information for WWTP authorities.

SUBJECT AREA: Analytical Chemistry, Mass Spectrometry

KEYWORDS: trends analysis, computational tools, statistical language R, emerging contaminants, non-target screening, wastewater

Στην οικογένειά μου

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας ερευνητικής εργασίας, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Νικόλαο Θωμαΐδη για τη συνεργασία και την εμπιστοσύνη που επέδειξε στο πρόσωπο μου. Επίσης θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής αξιολόγησης τον κο Ευσταθίου Κωνσταντίνο και τον κο Κουττάρη Μιχάλη για τις παρατηρήσεις τους, αλλά και για τις γνώσεις που μου μετάδωσαν κατά τη διάρκεια των προπτυχιακών και μεταπτυχιακών μου σπουδών. Επιπλέον θα ήθελα να ευχαριστήσω όλους τους συναδέλφους για την συνεργασία και την βοήθεια που ανιδιοτελώς μου προσέφεραν καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής. Ιδιαίτερα ευχαριστώ τους συναδέλφους Αλεξανδρο Μαρκάτη και Reza Aalizadeh για την ψυχολογική και έμπρακτη υποστήριξη τους. Τέλος θα ήταν παράλειψη να μην ευχαριστήσω το Κοινωφελές Ίδρυμα Ιωάννη Σ. Λάτση για την διαρκή χρηματοδότηση των μεταπτυχιακών μου σπουδών.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ.....	23
1. ΚΕΦΑΛΑΙΟ 1	25
Εισαγωγή	25
1.1 Εισαγωγή.....	25
1.2 Φασματομετρία μαζών υψηλής διακριτικής ικανότητας (HRMS).....	28
1.3 Στοχευμένη, ύποπτη και μη στοχευμένη ανάλυση.....	30
1.4 Επίπεδα ταυτοποίησης από μη στοχευμένη ανάλυση με HRMS [18].....	32
2. ΚΕΦΑΛΑΙΟ 2	35
Ανασκόπηση εργαλείων μη στοχευμένης ανάλυσης με φασματομετρίας μαζών υψηλής διακριτικής ικανότητας.....	35
2.1 Επτά χρυσοί κανόνες (Seven Golden Rules-SGR) [19].....	35
2.2 Bruker SmartFormula, SmartFormula Manually, SmartFormula3D®.....	42
2.3 MassBank.....	44
2.3.1 MassBank από την οπτική γωνία του διαχειριστή	45
2.3.2 MassBank από την οπτική γωνία του χρήστη	48
2.4 MetFrag	50
2.5 MetFusion.....	53
2.5.1 Αρχιτεκτονική του συστήματος.....	54
2.5.2 Ενσωμάτωση της φασματικής ομοιότητας, των in silico βαθμολογιών και της χημικής ομοιότητας.....	55
2.6 ProteoWizard Software®.....	57
2.6.1 Αναπαράσταση δεδομένων φασματομετρίας μαζών ως mzXML	57
2.6.2 Προβλήματα της κωδικοποίησης mzXML και η αντιμετώπισή τους	59
2.6.3 Δομή ProteoWizard και Msdata εφαρμογής και msconvert.exe	60
2.7 xcms	64
2.7.1 Εύρεση κορυφών (Peak Picking)	65

2.7.1.1	Εισαγωγή	65
2.7.1.2	Matchedfilter.....	66
2.7.1.3	Μειονεκτήματα αλγορίθμων εύρεσης κορυφών που βασίζονται στην τεχνική binning.....	69
2.7.1.4	centWave	71
2.7.1.5	Massifquant.....	74
2.7.1.6	enviPick.....	77
2.7.2	Ομαδοποίηση κορυφών που αναπαριστούν τον ίδιο αναλύτη σε όλα τα δείγματα.....	79
2.7.2.1	mzClust	81
2.7.2.2	density.....	83
2.7.2.3	K-Nearest.....	86
2.7.3	Διόρθωση χρόνου ανάσχεσης λόγω ολίσθησης.....	86
2.7.3.1	Αλγόριθμος OBI-Warp.....	87
2.7.3.2	Αλγόριθμος loess	89
2.7.4	Επανομαδοποίηση.....	91
2.7.5	Γέμισμα τιμών με μικρή ένταση για κορυφές που δεν υπάρχουν σε κάποια δείγματα.....	91
2.8	CAMERA	92
2.9	Metaboanalyst	97
2.10	Άλλα διαθέσιμα εργαλεία για μη στοχευμένη ανάλυση.....	102
3.	ΚΕΦΑΛΑΙΟ 3 Σκοπός της εργασίας.....	105
4.	ΚΕΦΑΛΑΙΟ 4 Μεθοδολογία.....	107
4.1	Δειγματοληψία	107
4.2	Προκατεργασία Δειγμάτων	107
4.3	Υγροχρωματογραφία-Φασματομετρία Μαζών	109
5.	ΚΕΦΑΛΑΙΟ 5 Αποτελέσματα-Συζήτηση.....	113
5.1	Αλγόριθμοι και βελτιστοποίησή τους	113

5.2	Τεχνικές προτεραιοποίησης κορυφών (Peak prioritization techniques)	116
5.3	Προτεραιοποίηση με χρήση του στατιστικού Multivariate empirical Bayes	118
5.4	Προτεινόμενη πορεία	122
5.5	Γραφικό περιβάλλον ανίχνευσης τάσεων	122
5.6	Ανίχνευση ενώσεων που εμφανίζουν τάση	123
5.6.1	Παράδειγμα ανίχνευσης συστατικού σε θετικό ιοντισμό	125
5.6.2	Παράδειγμα ανίχνευσης συστατικού σε αρνητικό ιοντισμό	129
5.6.3	Παράδειγμα ανίχνευσης «παλμού» ρύπανσης (pollution spill) σε θετικό ιοντισμό	132
6.	ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ	135
7.	ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ	141
7.	ΠΑΡΑΡΤΗΜΑ Ι.....	145
8.	ΠΑΡΑΡΤΗΜΑ ΙΙ.....	151
9.	ΑΝΑΦΟΡΕΣ	157

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Διάγραμμα τάσης της κατανάλωση κοκαΐνης και MDMA σε μια πανευρωπαϊκή διεργαστηριακή έρευνα, πηγή [7].....	27
Εικόνα 2: Προτεινόμενα επίπεδα ταυτοποίησης αναλυτών μέσω φασματομετρίας μαζών υψηλής διακριτικής ικανότητας, πηγή [18]	32
Εικόνα 3: Ισοτοπικό προφίλ 45000 μοριακών τύπων από την βάση δεδομένων Wiley και 60000 πεπτιδίων με μικρό μοριακό βάρος, πηγή [19]	38
Εικόνα 4: Ο λόγος υδρογόνου προς άνθρακα για 42000 μοριακούς τύπους από τη βάση δεδομένων Wiley, πηγή [19]	39
Εικόνα 5: Στιγμιότυπο οθόνης SmartFormula Manually για τον υπολογισμό στοιχειακής σύνθεσης	43
Εικόνα 6: Κεντρικό μενού βάσης δεδομένων MassBank	45
Εικόνα 7: Αναζήτηση φασμάτων από τη βάση δεδομένων MassBank	48
Εικόνα 8: Σύγκριση φασμάτων με παράθεση και τρισδιάστατη απεικόνιση από τη βάση δεδομένων MassBank	49
Εικόνα 9: Η ροή εργασίας του MetFrag, πηγή [28]	51
Εικόνα 10: Αλγόριθμος για in silico θραυσματοποίηση, πηγή [28].....	52
Εικόνα 11: Η ροή εργασίας του MetFusion, πηγή [29].....	54
Εικόνα 12: Η μορφή mzXML επιτρέπει κοινή αναπαράσταση δεδομένων, εφαρμογή κοινών μεθόδων επεξεργασίας και εξασφαλίζει δημόσια διαθεσιμότητα μαζί με τη δημοσίευση των αποτελεσμάτων, πηγή [30]	58
Εικόνα 13: Δομή του προγράμματος ProteoWizard, πηγή [32].....	61
Εικόνα 14: Γραφική αναπαράσταση μονάδας msData του προγράμματος ProteoWizard, πηγή [32]	62
Εικόνα 15: Στιγμιότυπο γραφικού περιβάλλοντος msConvert (ProteoWizard App)	64
Εικόνα 16: Πορεία εργασίας του Xcms και οι βασικές εντολές για κάθε βήμα, πηγή [34]	65
Εικόνα 17: Διαστάσεις λαμβανόμενων δεδομένων LC-MS	66
Εικόνα 18: Γραφική πορεία αλγορίθμου matchedfilter, πηγή [33].....	67

Εικόνα 19: Αποτελέσματα αλγορίθμου matchedfilter χρησιμοποιώντας τη δεύτερη παράγωγο του Gauss με διαφορετικά εύρη φίλτρου για κορυφές με διαφορετικά εύρη, πηγή [35]	70
Εικόνα 20: Αποτελέσματα αλγορίθμου matchedfilter χρησιμοποιώντας τη δεύτερη παράγωγο του Gauss με διαφορετικά εύρη φίλτρου για κοντινά εκλούόμενες κορυφές, πηγή [35]	70
Εικόνα 21: Γραφική αναπαράσταση λειτουργίας του αλγορίθμου centWave για την εύρεση περιοχών ενδιαφέροντος (Regions Of Interest-ROIs), πηγή [35]	72
Εικόνα 22: Συνάρτηση Mexican Hat για διαφορετικές τιμές εύρους.....	73
Εικόνα 23: Κινήσεις διαστολής-συστολής και μετατόπισης της συνάρτησης Mexican Hat, πηγή [38]	73
Εικόνα 24: Γραφική αναπαράσταση λειτουργίας του αλγορίθμου Massifquant για τρεις συνεχόμενες πλήρεις σαρώσεις, πηγή [39]	76
Εικόνα 25: Απεικόνιση εκτιμητή της πυκνότητας πιθανότητας.....	84
Εικόνα 26: Παράδειγμα ομαδοποίησης ιόντων (features) σε 12 δείγματα, πηγή [33] ...	85
Εικόνα 27: Διόρθωση ολίσθησης χρόνου ανάσχεσης με τη μέθοδο της δυναμικής χρονικής στρέβλωσης.....	88
Εικόνα 28: Διάγραμμα ροής που δείχνει την χρωματογραφική ευθυγράμμιση με τον OBI-Warp αλγόριθμο, πηγή [44]	89
Εικόνα 29: Παράδειγμα εποπτικής εικόνας της διόρθωσης χρόνου ανάσχεσης από τον αλγόριθμο loess.....	91
Εικόνα 30: Η ροή εργασίας της CAMERA για ανάλυση LC/MS δεδομένων, πηγή [47] .	93
Εικόνα 31: Διαχωρισμός συνεκλούμενων ιόντων δυο συστατικών από το πακέτο CAMERA, πηγή [47]	94
Εικόνα 32: Πορεία εισαγωγής δεδομένων στο Metaboanalyst.....	98
Εικόνα 33: Εργαλείο οπτικοποίησης της κανονικοποίησης του Metaboanalyst.....	100
Εικόνα 34: Δομή πολυπροσροφητικών στηλών SPE	108
Εικόνα 35: Χημική δομή στατικών φάσεων.....	108
Εικόνα 36: Επιφάνειες απόκρισης για βελτιστοποίηση αλγορίθμου εύρεσης κορυφών	115

Εικόνα 37: Επιφάνειες απόκρισης για βελτιστοποίηση ομαδοποίησης ιόντων και διόρθωσης χρόνου ανάλυσης.....	116
Εικόνα 38: Σχηματική υπολογιστική πορεία, που εφαρμόστηκε στα δεδομένα LC-HRMS	122
Εικόνα 39: Εργαλείο οπτικοποίησης διακυμάνσεων ιόντος με μάζα 197,0821 που εκλύεται σε χρόνο 3,27 λεπτά.....	123
Εικόνα 40: Διακύμανση ιόντος 380,3464	125
Εικόνα 41: Φάσμα MS κορυφής σε χρόνο 12,61 λεπτά.....	125
Εικόνα 42: Θραυσματοποίηση μοριακού ιόντος 363,3105.....	127
Εικόνα 43: Φάσματα MS/MS του μοριακού ιόντος 197,0821 σε παράθεση.....	129
Εικόνα 44: Προφίλ ιόντος 129,0200 σε χρόνο 11,80 λεπτά	132
Εικόνα 45: Φάσματα MS και MS/MS του ιόντος 361,2161 που παράγει το θραύσμα 129,0172.....	132

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Σύγκριση μεταξύ φασματομέτρων μαζών, τυπικές τιμές για εύρος μαζών 300-400, πηγή [11]	28
Πίνακας 2: Μέγιστος αριθμός ατόμων σε μοριακούς τύπους από τις βάσεις δεδομένων DNP και Wiley, πηγή [19]	36
Πίνακας 3: Συνήθεις λόγοι σύστασης στοιχείων με τον άνθρακα σε μοριακούς τύπους, πηγή [19]	39
Πίνακας 4: Πολλαπλός περιορισμός στοιχείων σε μοριακό τύπο για ενώσεις με μοριακό βάρος μικρότερο από 2000 Da βασισμένο στην βάση δεδομένων Beilstein και το «Dictionary of Natural Products», πηγή [19]	41
Πίνακας 5: Δομή καταγραφής MassBank (MassBank Record), πηγή [22]	47
Πίνακας 6: Τυπική έξοδος (output) οποιουδήποτε αλγορίθμου εύρεσης ιόντων	79
Πίνακας 7: Επιπλέον στοιχεία εξόδου (output) του αλγορίθμου centWave	80
Πίνακας 8: Ενδεικτικός ομαδοποιημένος πίνακας ιόντων για τρία δείγματα	82
Πίνακας 9: Συνθήκες ηλεκτροψεκασμού διαλύτες κινητής φάσης και πρόγραμμα βαθμιδωτής έκλουσης για θετικό και αρνητικό ιοντισμό	110
Πίνακας 10: Τα πρώτα 20 ζεύγη m/z και χρόνων ανάλυσης αποτελέσματα για αρνητικό και θετικό ιοντισμό	124
Πίνακας 11: Ιόντα που ανιχνεύθηκαν και επεξήγηση τους για το χρόνο 762-765 δευτερόλεπτα	126
Πίνακας 12: Υποψήφιες δομές μοριακού ιόντος 363,3105	128
Πίνακας 13: Υποψήφιες δομές μοριακού ιόντος 197,0821	130
Πίνακας 14: Υποψήφιες δομές μοριακού ιόντος 361,2161	133
Πίνακας 15: Γραφική απεικόνιση τάσεων και χρωματογραφημάτων για τα δέκα πρώτα υψηλότερης προτεραιότητας ιόντα στον αρνητικό ιοντισμό	145
Πίνακας 16: Γραφική απεικόνιση τάσεων και χρωματογραφημάτων για τα δέκα πρώτα υψηλότερης προτεραιότητας ιόντα στον θετικό ιοντισμό	151

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο Τμήμα Χημείας στον τομέα Αναλυτικής Χημείας στα πλαίσια του διαπανεπιστημιακού προγράμματος σπουδών «Χημική Ανάλυση-Έλεγχος Ποιότητας». Η εκπόνηση της διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του Αναπληρωτή Καθηγητή του ΕΚΠΑ Νικόλαο Θωμαΐδη. Η εργασία πραγματοποιήθηκε στα πλαίσια του ερευνητικού προγράμματος της Αριστείας «Προϊόντα Μετασχηματισμού Αναδυόμενων Ρύπων στο Υδάτινο Περιβάλλον»

Στην πρώτη ενότητα πραγματοποιείται μια εισαγωγή στο θέμα που διαπραγματεύεται η διπλωματική και στην οργανολογία που χρησιμοποιείται η οποία μας επιτρέπει πέρα από στοχευμένη ανίχνευση πολλών αναλυτών μας παρέχει και τη δυνατότητα μη στοχευμένης ανίχνευσης αγνώστων ουσιών. Ακολούθως ανασκοπούνται χημειομετρικά εργαλεία που χρησιμοποιούνται στη φασματομετρία μαζών υψηλής διακριτικής ικανότητας. Στη Τρίτη ενότητα καθίσταται σαφής ο σκοπός της διπλωματικής εργασίας ενώ στην τέταρτη ενότητα αναφέρεται το εργαστηριακό πειραματικό μέρος. Τέλος στην πέμπτη και τελευταία ενότητα αναφέρονται η επιλεγμένη υπολογιστική πορεία για εύρεση τάσεων και γίνεται η συζήτηση επί των αποτελεσμάτων και επί της προτεινόμενης πορείας.

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

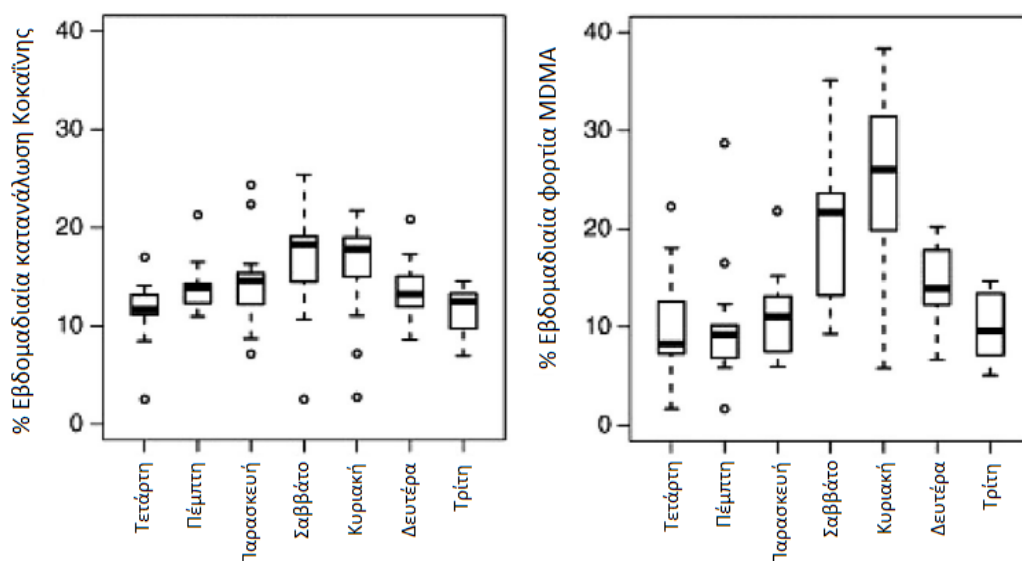
1.1 Εισαγωγή

Τις τελευταίες δεκαετίες ένας αυξανόμενος αριθμός αναδυόμενων ρύπων έχουν ανιχνευθεί στα επιφανειακά ύδατα. Με τον όρο αναδυόμενος ρύπος εννοούνται οι ενώσεις που δεν καλύπτονται από τους κοινοτικούς ή εθνικούς κανονισμούς ελέγχου ποιότητας των υδάτων και επομένως δεν ελέγχονται από τα κρατικά εργαστήρια ελέγχου ποιότητας υδάτων [1]. Οι αναδυόμενοι ρύποι ανάλογα με το είδος και τα επίπεδα των συγκεντρώσεων που ανιχνεύονται εμφανίζουν υψηλή τοξικότητα, είτε μόνοι τους είτε σε μίγματα που δημιουργούν μεταξύ τους, για οργανισμούς του οικοσυστήματος (ασπόνδυλα, άλγη, ψάρια). Επομένως αποτελούν ενδεχόμενο κίνδυνο για το ίδιο το οικοσύστημα και την ανθρώπινη υγεία μέσω της τροφικής αλυσίδας [2]. Οι περισσότεροι αναδυόμενοι ρύποι που βρέθηκαν στα επιφανειακά ύδατα απελευθερώνονται από τις διάχυτες πηγές (γεωργικές δραστηριότητες) ή σημειακές πηγές ρύπανσης (όπως μονάδες επεξεργασίας λυμάτων). Οι σημειακές πηγές ρύπανσης εκπέμπουν εν δυνάμει δεκάδες χιλιάδες ουσίες σε διαφορετικές συγκεντρώσεις. Οι τεχνολογίες επεξεργασίας λυμάτων που χρησιμοποιούνται αδυνατούν να απομακρύνουν επαρκώς όλους τους ρύπους και έτσι οι ρύποι αυτοί καταλήγουν στο περιβάλλον. Από τους αρχικούς ρύπους προκύπτουν προϊόντα μετασχηματισμού τα οποία συνεισφέρουν στην πολυπλοκότητα των μιγμάτων που δημιουργούνται. Μεταξύ των κατηγοριών αναδυόμενων ρύπων που έχουν ανιχνευθεί μέχρι στιγμής συμπεριλαμβάνονται φαρμακευτικές ουσίες, παράνομα διακινηθείσες ουσίες, ναρκωτικά, γλυκαντικές ουσίες, επιφανειοδραστικά χημικά, προϊόντα προσωπικής χρήσης, στερεοειδή, ορμόνες, σιλοξάνια, πολυφθοριομένα συστατικά, ηλιοπροστατευτικά χημικά, βενζοθειαζόλες και βενζοτριαζόλες, μεταβολίτες, προϊόντα μετασχηματισμού και άλλες ενώσεις [3].

Τα επίπεδα των ρύπων στα ακατέργαστα λύματα διακυμαίνονται από μερικά ng L^{-1} έως και μερικές εκατοντάδες χιλιάδες ng L^{-1} ανάλογα με τον αναλύτη, τη χρονική στιγμή της δειγματοληψίας και το κέντρο επεξεργασίας λυμάτων από το οποίο πάρθηκε το δείγμα. Για παράδειγμα, το παυσίπονο παρακεταμόλη

αναφέρθηκε στην βόρεια Αγγλία το 2006 σε συγκέντρωση 6924 ng L^{-1} ενώ στη Νότια Ουαλία το 2008 σε συγκέντρωση $492.340 \text{ ng L}^{-1}$ [4, 5]. Αυτό το γεγονός καταδεικνύει ότι υπάρχουν χωρικές και χρονικές διακυμάνσεις. Αυτές οφείλονται σε αστάθμητους παράγοντες όπως η αραίωση των λυμάτων με εργοστασιακά απόβλητα, η τυχαία δειγματοληψία, η αστάθεια στην ροή των λυμάτων και η χημική σταθερότητα των αναλυτών. Εκτός όμως από τους αστάθμητους παράγοντες, οι μεταβολές οφείλονται και στο ποσοστό κατανάλωσης των φαρμάκων από τον πληθυσμό ή γενικότερα το ποσοστό χρήσης των χημικών ουσιών από την κοινωνία. Φυσικά στην περίπτωση μελέτης των διακυμάνσεων αυτών λαμβάνεται πρόνοια ώστε να περιοριστούν οι αστάθμητοι παράγοντες. Για παράδειγμα λαμβάνεται σύνθετο 24ωρο δείγμα εισερχόμενων λυμάτων που είναι σταθμισμένο ως προς τη ροή, γεγονός που μειώνει πολύ την αβεβαιότητα της δειγματοληψίας και αυξάνει την αντιπροσωπευτικότητα του δείγματος [6].

Μια από τις πρώτες συστηματικές προσπάθειες μελέτης των χωρικών και χρονικών διακυμάνσεων των συγκεντρώσεων στα ακατέργαστα λύματα πραγματοποιήθηκε (αλλά συνεχίζει να είναι εν εξελίξει) σε πανευρωπαϊκό επίπεδο. Κατά τη μελέτη αυτή πάρθηκαν δείγματα από κέντρα επεξεργασίας λυμάτων από αρκετές ευρωπαϊκές χώρες και αναλύθηκαν ως προς ναρκωτικές ουσίες από τοπικά ακαδημαϊκά και ερευνητικά εργαστήρια. Οι συμμετέχοντες ανέφεραν τα επίπεδα συγκεντρώσεων που παρατήρησαν και επίσης ανέφεραν και τα αποτελέσματα της ανάλυσης διεργαστηριακών δειγμάτων. Ο λόγος που αναλύθηκαν δείγματα γνωστής περιεκτικότητας ως άγνωστα από τους συμμετέχοντες ήταν για να διασφαλιστεί η ποιότητα των αποτελεσμάτων. Ένα από τα πρώτα συμπεράσματα που προέκυψαν είναι η αξιοσημείωτη εβδομαδιαία διακύμανση των παράνομων ουσιών και μεταβολιτών τους στα ακατέργαστα λύματα που αποκάλυψε ότι υπάρχει συγκεκριμένη διακύμανση των ουσιών που χρησιμοποιούνται για ψυχαγωγικούς σκοπούς. Πιο συγκεκριμένα αποδείχτηκε ότι η βενζοϋλεκγονίνη (ο κύριος μεταβολίτης της κοκαΐνης) και το MDMA (3,4-μεθυλενοδιοξυ-μεθαμφεταμίνη) παρουσίασαν υψηλότερα επίπεδα συγκεντρώσεων κατά το σαββατοκύριακο.



Εικόνα 1: Διάγραμμα τάσης της κατανάλωση κοκαΐνης και MDMA σε μια πανευρωπαϊκή διεργαστηριακή έρευνα, πηγή [7]

Αντίθετα, αναμένεται ότι χημικά, όπως τα μέσα που χρησιμοποιούνται για απεικόνιση με ακτίνες Χ και ορισμένα αντικαρκινικά φάρμακα παρατηρούνται σε υψηλότερες συγκεντρώσεις κατά τις εργάσιμες μέρες τις εβδομάδας [8]. Η συγκεκριμένη πανευρωπαϊκή διεργαστηριακή έρευνα επιδημιολογίας λυμάτων (sewage epidemiology) συνεχίστηκε (και συνεχίζεται) και οδήγησε σε πολλά ενδιαφέροντα αποτελέσματα για την ετήσια διακύμανση της χρήσης παράνομων ουσιών από τον ευρωπαϊκό πληθυσμό [7].

Σε μια άλλη ερευνητική εργασία συσχετίζονται οι ετήσιες διακυμάνσεις των καταναλώσεων φαρμακευτικών ουσιών όπως αντικαταθλιπτικών και παράνομα διακινούμενων ουσιών, άρα και των επιπέδων συγκεντρώσεων λαμβάνοντας υπόψη την καθημερινή ροή των λυμάτων, με κοινωνικές και οικονομικές αλλαγές που συμβαίνουν σε κοινωνικό επίπεδο [9].

Επομένως οι ουσίες που υπάρχουν στα ακατέργαστα λύματα μπορούν να μας παρέχουν πολύτιμες πληροφορίες για την κατανάλωση και χρήση χημικών ουσιών από τους πολίτες τους οποίους εξυπηρετεί το κέντρο επεξεργασίας λυμάτων. Κατά κάποιο τρόπο οι ουσίες αυτοί αποτελούν το δακτυλικό αποτύπωμα μιας κοινωνίας και άρα είναι σημαντικό να βρεθούν συστατικά που παρουσιάζουν ισχυρή διακύμανση στα ακατέργαστα λύματα (από δω και στο εξής χρησιμοποιείται ο όρος τάσεις (trends)).

1.2 Φασματομετρία μαζών υψηλής διακριτικής ικανότητας (HRMS)

Βελτιώσεις στην εκχύλιση, προσυγκέντρωση και στα αναλυτικά πρωτόκολλα οδηγούν όλο και περισσότεροι αναλύτες να είναι ανιχνεύσιμοι στα δείγματα. Η εξέλιξη της υψηλής διακριτικής ικανότητας φασματομετρίας μαζών συζευγμένης με υγροχρωματογραφία έχει ανοίξει νέους ορίζοντες και έχει δώσει την ευκαιρία ανίχνευσης οργανικών ρύπων σε περίπλοκα δείγματα. Με αυτήν την τεχνολογία πολλά συστατικά που διαφεύγουν της αεριοχρωματογραφίας χωρίς παραγωγοποίηση μπορούν να ανιχνευθούν αξιόπιστα (συμπεριλαμβανομένων αυτών με χαρακτηριστικές ομάδες όπως οξέα, φαινόλες και αμίνες) [10].

Τυπικά εμπορικώς διαθέσιμα φασματόμετρα μαζών και μερικά χαρακτηριστικά ποιότητάς του απεικονίζονται στον ακόλουθο πίνακα:

Πίνακας 1: Σύγκριση μεταξύ φασματομέτρων μαζών, τυπικές τιμές για εύρος μαζών 300-400, πηγή [11]

Φασματόμετρο μαζών	Διακριτική ικανότητα (Resolving power)*	Ακρίβεια μάζας (ppm)	Γραμμική δυναμική περιοχή	Ευαισθησία**
Τριπλό τετράπολο (QqQ)	FWHM=0,7 Da	50	10 ⁴	fg-pg (SRM)
Τετραπολική παγίδα ιόντων (QIT)	10.000	50	10 ³	fg-pg (SRM, full scan)
Αναλυτής μαζών χρόνου πτήσης (TOF)	40.000	3-5	10 ² -10 ³	pg (full scan)
Orbitrap	100.000	2	10 ³ -10 ⁴	fg to pg (full scan)
Αναλυτής μαζών κυκλοτρονικού συντονισμού ιόντων με μετασχηματισμό Fourier (FT-ICR)	1.000.000	≤1	10 ⁴	pg (full scan)
*Η διακριτική ικανότητα εξαρτάται από τη μάζα και την ταχύτητα των περισσότερων οργάνων				
**Η ευαισθησία εξαρτάται από το είδος ιοντισμού και την αποτελεσματικότητα ιοντισμού από την πηγή ιοντισμού				

Συνοπτικά, όργανα όπως το τριπλό τετράπολο και η τετραπολική παγίδα ιόντων είναι οι τεχνολογίες που επιστρατεύονται για ποσοτικές αναλύσεις ρουτίνας. Αυτοί οι αναλυτές προσφέρουν υψηλή ευαισθησία και εκλεκτικότητα αλλά λειτουργούν με χαμηλή διακριτική ικανότητα ($FWHM=0,7$ Da) και στην περίπτωση του τριπλού τετραπόλου υπάρχει και χαμηλή ευαισθησία σε λειτουργία πλήρους σάρωσης γεγονός που καθιστά τον αναλυτή μαζών μη ικανοποιητικό για ανάλυση αγνώστων ουσιών. Οι σύγχρονοι αναλυτές μαζών χρόνου πτήσης έχουν υψηλή ταχύτητα και είναι ικανοί να έχουν διακριτική ικανότητα έως 40,000 αλλά όμως έχουν χαμηλή ευαισθησία και περιορισμένη γραμμική δυναμική περιοχή, μειονεκτήματα που ωστόσο τείνουν να απαλειφθούν. Ο αναλυτής μαζών orbitrap εισήχθει στην αγορά το 2005 σε προσιτή τιμή σε αντίθεση με τα ακριβότερα FT-ICR όργανα, και συνδυάζει υψηλή διακριτική ικανότητα και υψηλή ακρίβεια μάζας, αρκετά υψηλή ευαισθησία με αρκετά όμως χαμηλότερη ταχύτητα σε σχέση με τους αναλυτές TOF.

Συνδυασμός δυο ή περισσότερων αναλυτών μαζών δημιουργούν τα λεγόμενα υβριδικά όργανα όπως το τετράπολο-αναλυτής μαζών χρόνου πτήσης (QTOF) ή γραμμική ιοντική παγίδα (Linear iontrap/orbitrap-LTQ Orbitrap) που αναδεικνύουν υψηλές ικανότητες ανίχνευσης και ταυτοποίησης ουσιών με χαμηλά μοριακά βάρη σε διαφορετικές μήτρες [12]. Στη σύγχρονη περιβαλλοντική ανάλυση αλλά και σε άλλα ερευνητικά πεδία όπως η μεταβολομική επιστρατεύονται κυρίως αναλυτές QTOF ή Orbitrap εξαιτίας της χρήσης της λειτουργίας συνεχούς καταγραφής φασμάτων πλήρης σάρωσης. Έτσι είναι εφικτό να ανευρεθούν αναμενόμενα συστατικά αλλά και εντελώς άγνωστα συστατικά σε ένα πολύπλοκο δείγμα. Δηλαδή καθίσταται εφικτή η λεγόμενη μετά τις μετρήσεις διερεύνηση των δεδομένων (post-run acquisition) [10]. Η χρήση φασματομέτρων μαζών υψηλής διακριτικής ικανότητας με αναλυτές μαζών Orbitrap και TOF παρέχουν τόσο υψηλή ακρίβεια μάζας όσο και υψηλή διακριτική ικανότητα σε λειτουργία πλήρης σάρωσης και έτσι καθιστά εφικτή την ανίχνευση θεωρητικά απεριόριστου αριθμού οργανικών ρύπων [11].

1.3 Στοχευμένη, ύποπτη και μη στοχευμένη ανάλυση

Από την κλασσική ανάλυση λίγων στοχευμένων αναλυτών μπορεί να διαφύγουν σημαντικά και πιθανώς οικοτοξικολογικά σημαντικά συστατικά. Για αυτό καινοτόμες προσεγγίσεις απαιτούνται για την ανάλυση των υδάτων [13]. Υπάρχουν τρεις κύριες προσεγγίσεις ανίχνευσης νέων αναλυτών που προτάθηκαν: η στοχευμένη ανάλυση (με χρήση προτύπων), η ύποπτη ανάλυση (πιθανώς παρούσες ουσίες βασισμένες σε a priori πληροφορίες αλλά χωρίς την ύπαρξη προτύπων αναφοράς) και τελευταία η μη στοχευμένη ανάλυση (δεν υπάρχει ούτε a priori πληροφορίες ούτε πρότυπο αναφοράς) [12].

Στη στοχευμένη ανάλυση, ένα πρότυπο αναφοράς είναι απαραίτητο για να καθοριστεί η συγκέντρωση στο δείγμα και για την ταύτιση του χρόνου ανάλυσης και αν είναι απαραίτητο του φάσματος MS/MS. Ένα εσωτερικά επισημασμένο πρότυπο πρέπει ιδανικά να είναι διαθέσιμο για κάθε ένωση στόχο για να αξιολογηθεί η απόκριση (response) του δείγματος. Μια πλήρης καμπύλη αναφοράς για ποσοτικούς προσδιορισμούς είναι απαραίτητη αλλιώς τα αποτελέσματα είναι ημιποσοτικά (ένα σημείο βαθμονόμησης) [10]. Αυτή η διαδικασία είναι συνήθως αυτοματοποιημένη και προσφέρεται από τα εμπορικά λογισμικά των οργάνων καθώς αναζητούν για κορυφές με δεδομένη μάζα σε δεδομένο χρόνο ανάλυσης η οποία εμφανίζει κάποια θραύσματα στο MS/MS φάσμα [11]. Μια ολοκληρωμένη στοχευμένη ανάλυση δεν μπορεί να πραγματοποιηθεί για όλα τα συστατικά που πιθανώς υπάρχουν στα απόβλητα διότι αυτό περιλαμβάνει την αγορά και μέτρηση χιλιάδων προτύπων ουσιών, γεγονός κοστοβόρο που απαιτεί αρκετή δουλειά ενώ πολλές φορές τα πρότυπα δεν είναι πάντα διαθέσιμα (για παράδειγμα για αρκετά προϊόντα μετασχηματισμού) [10, 11].

Η ανάλυση ύποπτων ουσιών με LC-HRMS δεδομένα βασίζεται στην αναζήτηση συστατικών που υπάρχουν υψηλές πιθανότητες να βρίσκονται στο υπο ανάλυση δείγμα. Επομένως από πριν υπάρχουν ενδείξεις της δομής που πιθανώς να υπάρχει στο δείγμα. Έπειτα με βάση το μοριακό τύπο του ύποπτου συστατικού υπολογίζεται η ακριβής μάζα και το θεωρητικό ισοτοπικό προφίλ τα οποία αναζητούνται στα δεδομένα ενώ έπειτα βασιζόμαστε και σε επιπλέον ενδείξεις (παρουσία παρόμοιων συστατικών, συνδιακύμανση ουσιών αν δείγματα έχουν παρθεί σε χρονοσειρά). Συστατικά που αναμένονται να

υπάρχουν στα δείγματα μπορούν να βρεθούν με τη χρήση της ακριβούς μάζας του μοριακού ιόντος που προκύπτει από τον μοριακό τύπο. Σε αντίθεση με την αεριοχρωματογραφία στην LC-HRMS ανάλυση δεν υπάρχουν εκτενείς βάσεις δεδομένων για να υποστηριχθεί η ανίχνευση. Για το λόγο αυτό η εύρεση αποδείξεων στα φάσματα παραμένει χρονοβόρα.

Η μη στοχευμένη ανάλυση (non-target analysis) περιλαμβάνει μάζες που ανιχνεύονται στα δείγματα για τις οποίες δεν υπάρχει καμία πληροφορία. Στην καθαρόαιμη μη στοχευμένη ανάλυση εφαρμόζεται ένας αυτόματος αλγόριθμος ανίχνευσης κορυφών ο οποίος τυπικά θα οδηγήσει σε αρκετές χιλιάδες κορυφών για κάθε δείγμα [13]. Οι μάζες αυτές δεν πρέπει να αντιστοιχούν σε συστατικά της λίστας ενώσεων στόχων (target list) ούτε και στην λίστα ύποπτων συστατικών (suspect list) [10, 14]. Η ανίχνευση τέτοιων συστατικών είναι δύσκολη και τίποτα δεν μπορεί να εγγυηθεί επιτυχές αποτέλεσμα [10, 15].

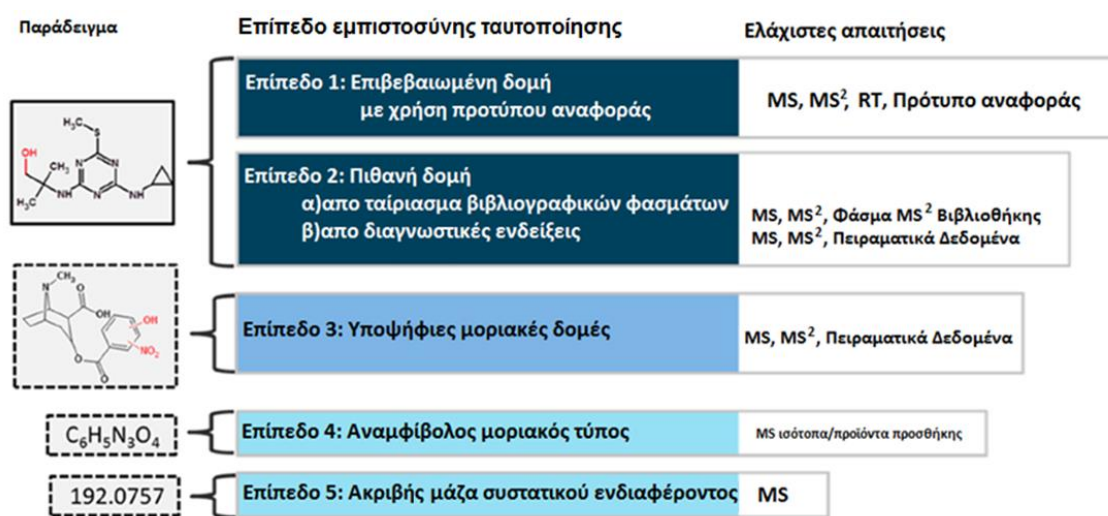
Για κάθε μάζα, που φυσικά δεν ανήκει στην λίστα ενώσεων (target compounds) στόχων ούτε στη λίστα ύποπτων ενώσεων (suspected compounds), μπορούν να αποδοθούν αρκετοί μοριακοί τύποι με βάση την ακριβή μάζα και το ισοτοπικό προφίλ. Αυτές πρέπει να φιλτραριστούν ώστε να καταλήξουμε σε μικρό αριθμό μοριακών τύπων, διότι σε κάθε μοριακό τύπο αντιστοιχούν πολλές υποψήφιες μοριακές δομές [13]. Το ισοτοπικό σήμα παρέχει σημαντικές πληροφορίες καθώς είναι ικανό να ανιχνεύσει στοιχεία όπως ^{34}S , ^{15}N και ^{18}O (ανάλογα του m/z) καθώς και του ^{37}Cl και ^{81}Br . Αφού καταλήξουμε στον μοριακό τύπο ενός ψευδομοριακού ιόντος το επόμενο βήμα είναι η χρήση της πληροφορίας θραυσματοποίησης από υψηλής διακριτικής ικανότητας φάσματα MS/MS. Αυτή μειώνει τις πιθανές μοριακές δομές που έπειτα θα αναζητηθούν σε χημικές βάσεις δεδομένων. Η ανίχνευση ενώσεων που δεν υπάρχουν καν στις χημικές βάσεις δεδομένων είναι ακόμα πιο δύσκολη. Επομένως η μη στοχευμένη ανάλυση περιλαμβάνει συνδυασμό και συμφωνία τόσο της ακριβούς μάζας, του ισοτοπικού προφίλ αλλά και του προφίλ θραυσματοποίησης προκειμένου να οδηγηθούμε στη σωστή χημική δομή που κρύβεται πίσω από μια κορυφή [10, 14].

Τα τρία είδη αναλύσεων (target, suspect, non target analysis) είναι συμπληρωματικά και στόχο έχουν την ανίχνευση των περισσότερων συστατικών, ιδανικά όλων, σε ένα δείγμα [16]. Οι λίστες στοχευμένης και μη

στοχευμένης ανάλυσης επικοινωνούν. Έτσι σε περίπτωση που επιτευχθεί ταυτοποίηση ενός συστατικού με μη στοχευμένη ανάλυση σε επίπεδο 1 (βλέπε υποκεφάλαιο 1.4) τότε αυτό μεταφέρεται στην λίστα με τους αναλύτες στόχους και από κει και πέρα εξετάζεται με στοχευμένη ανάλυση. Αντίθετα αν αναζητηθούν συστατικά από τη στοχευμένη λίστα και δεν ταιριάζει ο χρόνος ανάλυσης ή το ισοτοπικό προφίλ πρέπει να διαχειριστούν τα συστατικά αυτά ως άγνωστα [14]. Η αδυναμία της στοχευμένης ανάλυσης να ανιχνεύσει σημαντικά συστατικά στα δείγματα οδήγησε στην ανάπτυξη των άλλων δυο μεθοδολογιών (ύποπτη και μη στοχευμένη). Για αυτό μια ολιστική ανάλυση των συστατικών στα επιφανειακά ύδατα και σε οποιαδήποτε άλλη μήτρα πρέπει πέρα από τη στοχευμένη να περιλαμβάνει και μη στοχευμένη ανάλυση [17].

1.4 Επίπεδα ταυτοποίησης από μη στοχευμένη ανάλυση με HRMS [18]

Έχουν προταθεί τα ακόλουθα πέντε επίπεδα ταυτοποίησης για την μη στοχευμένη και ύποπτη ανάλυση:



Εικόνα 2: Προτεινόμενα επίπεδα ταυτοποίησης αναλυτών μέσω φασματομετρίας μαζών υψηλής διακριτικής ικανότητας, πηγή [18]

Το επίπεδο 1 απεικονίζει την ιδανική κατάσταση, στην οποία υπάρχει επιβεβαιωμένη δομή και η επιβεβαίωση έχει πραγματοποιηθεί με τη χρήση προτύπου αναφοράς και σύγκριση των φασμάτων MS, MS/MS και ταύτιση μέσω χρόνου ανάλυσης. Προαιρετικά αν είναι πιθανό μπορεί να χρησιμοποιηθεί και μια ορθογώνια τεχνική, όπως για παράδειγμα ανάλυση με χρωματογραφική στήλη HILIC.

Το επίπεδο 2 υποδηλώνει ότι υπάρχει μόνο μία πιθανή δομή που μπορεί να προταθεί και έχει προκύψει από διαφορετικές ενδείξεις. Το επίπεδο δυο διαχωρίζεται σε δυο υποεπίπεδα, το 2A και το 2B. Για το επίπεδο 2A πρέπει να περιλαμβάνεται αναμφίβολη σύμπτωση του πειραματικού φάσματος MS/MS με φάσματα από βιβλιογραφία ή από βιβλιοθήκες φασμάτων. Πρέπει να ληφθεί μέριμνα ώστε τα υπο σύγκριση φάσματα να έχουν παρθεί με ίδιες παραμέτρους (ενέργεια διάσπασης, ιοντισμός, επίπεδο MS) προκειμένου να διασφαλιστεί η εγκυρότητα και η ταύτιση και τα κριτήρια επιλογής πρέπει να αναπαρίσταται με σαφήνεια. Επιθυμητή επιπλέον είναι η ένδειξη όπως η συμπεριφορά του χρόνου ανάλυσης. Το επίπεδο 2B που καλείται και διαγνωστικό επίπεδο, διότι υπάρχουν διαγνωστικά θραύσματα στο φάσμα MS/MS αναπαριστά την περίπτωση στην οποία δεν υπάρχουν άλλες ισομερείς δομές που να ταιριάζουν με το πειραματικό φάσμα, αλλά όμως δεν υπάρχει πρότυπο ή βιβλιογραφικό φάσμα για επιβεβαίωση. Ενδείξεις μπορούν να περιλαμβάνουν θραύσματα MS/MS ή/και συμπεριφορά ιοντισμού, πληροφορίες για το αρχικό συστατικό και πειραματικές πληροφορίες. Ένα καλό παράδειγμα είναι η υδροξυλίωση της *tert*-βουτυλομάδας του φυτοφαρμάκου *irgarol* (εικόνα 2). Παρόλου που ο διαχωρισμός του επιπέδου σε A και B είναι χρήσιμο για ερευνητικούς λόγους, πρακτικά όταν αναφερόμαστε σε επίπεδο δύο αυτό σημαίνει μια πιθανή δομή και απηχεί υψηλό επίπεδο εμπιστοσύνης.

Το επίπεδο 3 υποδηλώνει ότι υπάρχουν υποψήφιος δομές και περιγράφει μια γκρίζα ζώνη όπου υπάρχουν αποδείξεις για πιθανές δομές αλλά μη επαρκής πειραματική πληροφορία, έτσι ώστε να οδηγηθούμε σε μία μόνο επακριβή δομή (για παράδειγμα έχουν προκύψει πιθανές ισομερείς δομές). Παρόλο που υπάρχει αβεβαιότητα μεταξύ των πιθανών καταστάσεων, η ακριβής δομή παραμένει θεωρητική. Μπορούμε να οδηγηθούμε σε αυτό το επίπεδο ταυτοποίησης αν προκύψουν δομές με υψηλή βαθμολογία ταύτισης στις οποίες ταυτίζονται σε μεγάλο βαθμό το πειραματικό φάσμα MS/MS με το *in silico* φάσμα. Επιπλέον μπορεί να χρησιμοποιηθεί ως πρόσθετη πληροφορία η υψηλή πιθανότητα μετασχηματισμού ενός συστατικού, ο αριθμός αναφορών ενός συστατικού σε χημικές βιβλιοθήκες (ChemSpider) ή η συμπεριφορά του ως προς το χρόνο ανάλυσης.

Το επίπεδο 4 υποδηλώνει ότι υπάρχει κατηγορηματικός μοριακός τύπος. Ο μοριακός τύπος μπορεί να αποδοθεί χρησιμοποιώντας τη φασματική πληροφορία (π.χ. προϊόντα προσθήκης, ισοτοπικές κορυφές ή πληροφορία θραυσμάτων) αλλά δεν υπάρχουν επαρκή στοιχεία για να υποτεθούν δομές. Το φάσμα MS/MS μπορεί να μην είναι πληροφοριακό ή να περιέχει παρεμποδίσξεις ή ακόμα και να μην υπάρχει διαθέσιμο. Παρόλα αυτά ο μοριακός τύπος παρέχει κάποια πληροφορία που είναι άξια παρουσίασης καθώς μπορεί να χρησιμοποιηθεί σε μελλοντικές έρευνες.

Το επίπεδο πέντε οδηγεί στην ανίχνευση της ακριβούς μάζας ενός συστατικού, μπορεί να μετρηθεί σε ένα δείγμα και να είναι ιδιαίτερου ενδιαφέροντος σε μια έρευνα, αλλά υπάρχει ελλιπής πληροφορία για να ανατεθεί ένας μοριακός τύπος. Μη στοχευμένες μέθοδοι που επιτρέπουν την εύρεση αυτών των μαζών σε άλλες έρευνες επιτρέπονται αλλά το επίπεδο αυτό υπονοεί ότι δεν υπάρχει κατηγορηματική πληροφορία για τη δομή ή τον μοριακό τύπο. Είναι ακόμα πιθανόν να έχει καταγραφεί φάσμα MS/MS μιας μάζας επιπέδου πέντε και να σωθεί ως άγνωστο φάσμα. Αυτό το επίπεδο μπορεί μόνο να εφαρμοστεί σε λίγες μάζες ειδικού ενδιαφέροντος αφού θα μπορούσε να είναι αντιπαραγωγικό να τεθούν ταμπέλες σε όλες τις μάζες σε ένα δείγμα ως επίπεδο πέντε. Στο επίπεδο αυτό είναι σημαντικό να διασφαλιστεί ότι η μάζα δεν υπάρχει στο τυφλό δείγμα και άρα δεν προκύπτει από τα στάδια προκατεργασίας του δείγματος [18].

ΚΕΦΑΛΑΙΟ 2

Ανασκόπηση εργαλείων μη στοχευμένης ανάλυσης με φασματομετρίας μαζών υψηλής διακριτικής ικανότητας

2.1 Επτά χρυσοί κανόνες (Seven Golden Rules-SGR) [19]

Το πρώτο βήμα στη μη στοχευμένη ανίχνευση είναι η ανεύρεση του μοριακού ιόντος που περιέχει την πληροφορία του μοριακού βάρους της άγνωστης ένωσης. Αφού πραγματοποιηθεί αυτό, το επόμενο και καθοριστικό βήμα είναι να αναγνωριστεί η σωστή στοιχειακή σύνθεση, δηλαδή να μετατραπεί το μοριακό βάρος σε μοριακό τύπο. Για να μειωθούν οι χιλιάδες των πιθανών υποψήφίων μοριακών τύπων που μπορούν να προκύψουν έχουν αναπτυχθεί οι λεγόμενοι 7 χρυσοί κανόνες, οι οποίοι λαμβάνονται υπόψη, έτσι ώστε να επιλεγούν οι πιο πιθανοί και χημικά ορθοί μοριακοί τύποι [19].

Παρακάτω εξετάζονται οι 7 χρυσοί κανόνες, που μας οδηγούν σε πιθανούς μοριακούς τύπους:

1. Εφαρμογή περιορισμών στον αριθμό των στοιχείων που αποτελούν το μοριακό τύπο

Ο περιορισμός του αριθμού των στοιχείων μιας μοριακής δομής είναι σημαντικός για τον περιορισμό του υπολογιστικού χρόνου. Εάν η έρευνα έχει ως σκοπό τις μικρού μοριακού βάρους ενώσεις ή φυσικά προϊόντα δεν υπάρχει λόγος να συμπεριλαμβάνονται ουσίες με μοριακό τύπο που έχει μη λογικά πολύ υψηλό αριθμό στοιχείων. Για την ανάπτυξη αυτού του κανόνα, υπολογίστηκε απλά ο λόγος της μάζας του συστατικού δια την ατομική μάζα κάθε ατόμου και το αποτέλεσμα είναι ο μέγιστος αριθμός του στοιχείου αυτού στον μοριακό τύπο. Για παράδειγμα ο άνθρακας έχει μάζα 12 Da οπότε για μια ένωση με μάζα 1000 Da ο λόγος $1000/12=83$ είναι ο μέγιστος αριθμός ατόμων άνθρακα για αυτό το μόριο (Στην πραγματικότητα πρόκειται για λιγότερα από 83 άτομα άνθρακα διότι σε μια ένωση σίγουρα θα υπάρχουν και άλλα άτομα όπως υδρογόνα). Ο αριθμός ατόμων περιορίζεται ακόμα περισσότερο χρησιμοποιώντας μοριακές δομές από βάσεις δεδομένων για την εύρεση του μέγιστου αριθμού ατόμων. Οι Tobias Kind και Oliver Fiehn αναζήτησαν βάσεις

δεδομένων (Wiley και DNP) για μέγιστο αριθμό ατόμων σε μοριακούς τύπους και τα αποτελέσματα συνοψίζονται στον πίνακα:

Πίνακας 2: Μέγιστος αριθμός ατόμων σε μοριακούς τύπους από τις βάσεις δεδομένων DNP και Wiley, πηγή [19]

Μάζα (Da)	Βιβλιοθήκη	C _{max}	H _{max}	N _{max}	O _{max}	P _{max}	S _{max}	F _{max}	Cl _{max}	Br _{max}	Si _{max}
<500	DNP	29	72	10	18	4	7	15	8	5	-
	Wiley	39	72	20	20	9	10	16	10	4	8
<1000	DNP	66	126	25	27	6	8	16	11	8	-
	Wiley	78	126	20	27	9	14	34	12	8	14
<2000	DNP	115	236	32	63	6	8	16	11	8	-
	Wiley	156	180	20	40	9	14	48	12	10	15
<3000	DNP	162	208	48	78	6	9	16	11	8	-

Κρατήθηκαν ως μέγιστο όριο ατόμων για τους μοριακούς τύπους οι μεγαλύτεροι αριθμοί για κάθε εύρος μαζών που παρατηρήθηκαν στις βάσεις DNP και Wiley.

2. Εφαρμογή χημικών κανόνων LEWIS και SENIOR

Τα προγράμματα υπολογισμού μοριακών τύπων χρησιμοποιούν όλους τους συνδυασμούς στοιχείων που έχουν ως αποτέλεσμα τη σωστή ακριβή μάζα, αγνοώντας ενδεχόμενη δόκιμη ύπαρξη των χημικών τύπων. Για ουδέτερα μόρια ισχύουν οι κανόνες Lewis και Senior. Αυτοί οι κανόνες μπορούν να δοκιμαστούν σε ουδέτερα μόρια και έτσι τα ιοντικά είδη που ανιχνεύονται από τη φασματομετρία μαζών πρέπει να ουδετεροποιηθούν θέτοντας υπόψη την πληροφορία των προϊόντων προσθήκης. Για παράδειγμα για ένα πρωτονιομένο συστατικό που είναι σύνηθες προϊόν προσθήκης σε συνθήκες ηλεκτροψεκασμού αφαιρούμε από τη μάζα που μετρήθηκε τη μάζα του πρωτονίου (1,007825u) ώστε να προκύψει ουδέτερο μόριο.

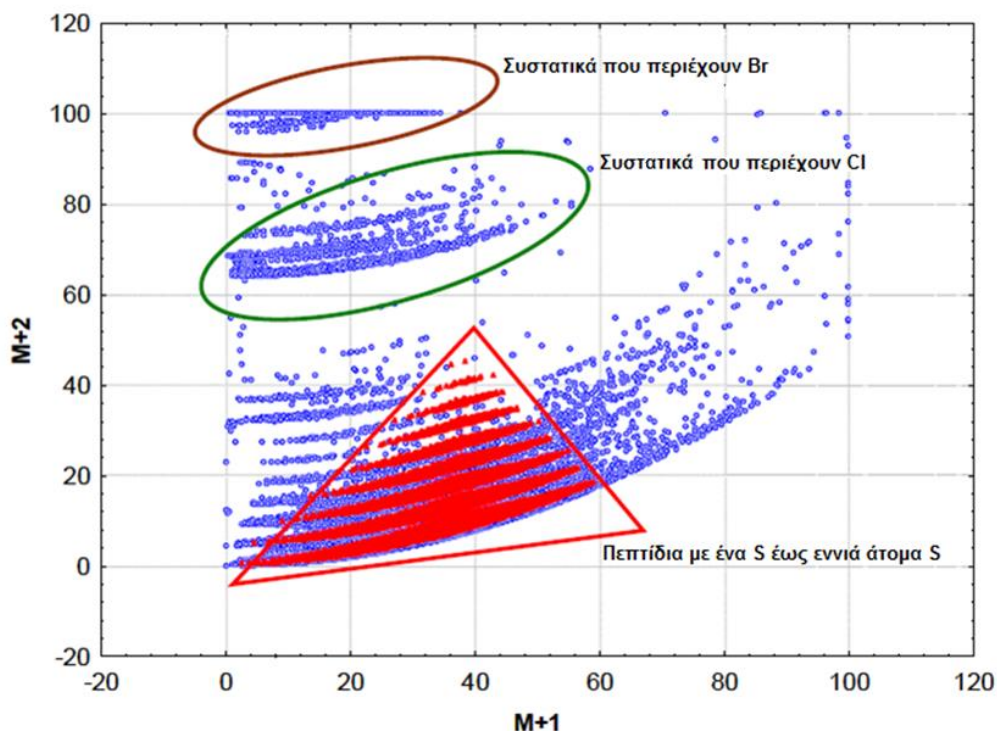
Στην πιο απλή του δομή ο κανόνας του Lewis απαιτεί μόρια που αποτελούνται από στοιχεία της κύριας ομάδας του περιοδικού πίνακα, κυρίως άνθρακα, άζωτο και οξυγόνο και μοιράζονται ηλεκτρόνια με έναν τρόπο ώστε όλα να έχουν εντελώς συμπληρωμένες τις στιβάδες s και p (κανόνας οκτάδας). Σε περίπτωση που επιθυμείται ο έλεγχος ελεύθερων ριζών και νιτροενώσεων δεν πρέπει να εφαρμοστεί ο κανόνας του Lewis διότι δε θα επιτραπούν τέτοιες ενώσεις καθώς αυτά τα μόρια θα σημειωθούν ως περιττά ηλεκτρονιακά μόρια και ο βαθμός δακτυλίων και διπλών δεσμών (Rings-Plus-Double-Bonds Equivalent-RDBE) θα είναι περιττός.

Επιπλέον νεότεροι *ab initio* υπολογισμοί έχουν δείξει ότι υπερσθενή μόρια (για παράδειγμα ClF_3) δεν υπακούουν στον κανόνα του Lewis και για αυτό πρέπει να συνδυαστούν με μια δοκιμασία που βασίζεται στο θεώρημα Senior. Σύμφωνα με αυτό:

- A. Το άθροισμα σθένους ή ο ολικός αριθμός ατόμων που έχουν περιττό σθένος είναι άρτιος
- B. Το άθροισμα σθένους είναι μεγαλύτερο ή ίσο του διπλάσιου του μέγιστου αριθμού σθένους
- Γ. Το άθροισμα του σθένους είναι μεγαλύτερο ή ίσο του διπλάσιου του αριθμού των ατόμων μείον 1

3. Αναζήτηση χαρακτηριστικών στοιχείων από το ισοτοπικό προφίλ

Το ισοτοπικό προφίλ μπορεί να ληφθεί υπόψη εφόσον το φασματόμετρο μάζας είναι υψηλής διακριτικής ικανότητας και μπορεί να παράσχει ξεκάθαρα τις εντάσεις των ιόντων $M+1$ και $M+2$. Δηλαδή απαιτείται το σήμα του αναλύτη να είναι σχετικά υψηλό ώστε να παράσχει υψηλό λόγο σήμα προς θόρυβο για τις ισοτοπικές κορυφές προκειμένου να εξαχθεί το ισοτοπικό προφίλ του. Η εικόνα 3 δείχνει τις εντάσεις των $M+1$ και $M+2$ ισοτοπικών κορυφών για μοριακούς τύπους από την φασματική βάση δεδομένων Wiley και μιας βάσης δεδομένων με πεπτιδικές δομές.

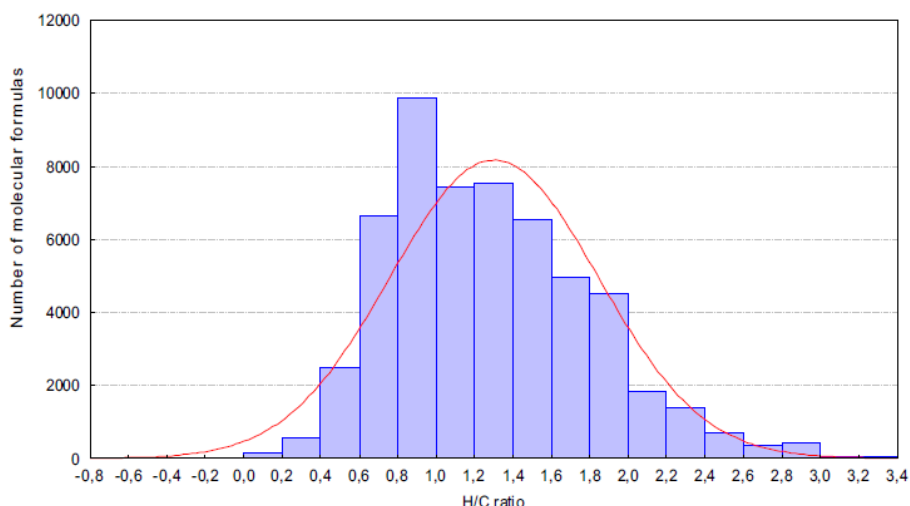


Εικόνα 3: Ισοτοπικό προφίλ 45000 μοριακών τύπων από την βάση δεδομένων Wiley και 60000 πεπτιδίων με μικρό μοριακό βάρος, πηγή [19]

Ενώσεις που περιέχουν είτε βρώμιο είτε χλώριο ή πεπτίδια με θείο παρέχουν ιδιαίτερο ισοτοπικό προφίλ που είναι σημαντικό εργαλείο για την ταυτοποίηση τέτοιων ενώσεων. Για μονοϊσοτοπικά στοιχεία (F, Na, P, I) αυτός ο κανόνας δεν έχει καμία επίδραση.

4. Έλεγχος στοιχειακής αναλογίας υδρογόνου προς άνθρακα

Ένας άλλος περιορισμός που χρησιμοποιείται για να μειωθούν οι πιθανοί μοριακοί τύποι είναι ο λόγος των στοιχείων και ιδιαίτερα ο λόγος υδρογόνου προς άνθρακα. Στις περισσότερες περιπτώσεις ο λόγος H/C δεν ξεπερνάει το 3 με λίγες εξαιρέσεις όπως η μεθυλνιτρίνη (CH_6N_2). Αντίστροφα ο λόγος H/C είναι συχνότερα μικρότερος από 2 και όχι μικρότερος από 0,125 όπως στην περίπτωση της τετρακυανοπυρρόλης. Η εικόνα 4 δείχνει ότι οι πιο τυπικές αναλογίες είναι $2,0 > \text{H/C} > 0,5$, για μακράς αλυσίδας αλκάνια τείνει στο 2, ενώ για αρωματικούς υδρογονάνθρακες τείνει στο 0,5.



Εικόνα 4: Ο λόγος υδρογόνου προς άνθρακα για 42000 μοριακούς τύπους από τη βάση δεδομένων Wiley, πηγή [19]

Η κατανομή δεν είναι κανονική και έτσι τα όρια δεν καθορίζονται από τρεις φορές την τυπική απόκλιση ή τέσσερις φορές την τυπική απόκλιση για τα ποσοστά κάλυψης 99,7 και 99,9% αντίστοιχα, και για αυτό χρησιμοποιήθηκαν αθροιστικά ποσοστά ως όρια εύρους που δίνονται στον ακόλουθο πίνακα και προέκυψαν από αναζήτηση στη βιβλιοθήκη φασμάτων Wiley.

Πίνακας 3: Συνήθεις λόγοι σύστασης στοιχείων με τον άνθρακα σε μοριακούς τύπους, πηγή [19]

Στοιχειακή αναλογία	Σύνηθες εύρος (καλύπτει 99,7%)	Εκτεταμένο εύρος (καλύπτει 99,99%)	Εκτεταμένο εύρος (πέρα του 99,99%)
H/C	0,2-3,1	0,1-6	<0,1 και 6-9
F/C	0-1,5	0-6	>1,5
Cl/C	0-0,8	0-2	>0,8
Br/C	0-0,8	0-2	>0,8
N/C	0-1,3	0-4	>1,3
O/C	0-1,2	0-3	>1,2
P/C	0-0,3	0-2	>0,3

S/C	0-0,8	0-3	>0,8
Si/C	0-0,5	0-1	>0,5

Περισσότερα από 99,7% των μοριακών δομών έχουν λόγο H/C μεταξύ 0,2-3,1 και κατά συνέπεια, το εύρος αυτό καλείται «σύννηθες εύρος». Παρόλα αυτά υπάρχουν χημικές ομάδες που ξεφεύγουν από αυτό το εύρος και έτσι μπορεί να επιλεγεί το εκτεταμένο εύρος που καλύπτει 99,99% των μοριακών δομών (H/C=0,1-6) ενώ υπάρχουν και ακραίες περιπτώσεις όπως τα φλουορένια που έχουν πολύ χαμηλό λόγο H/C.

5. Έλεγχος στοιχειακής αναλογίας αζώτου, οξυγόνου, φωσφόρου, θείου προς τον άνθρακα

Ο έλεγχος του λόγου ετεροατόμων μειώνει δραστικά τις υποψήφιες μοριακές δομές. Παρόλα αυτά οι κατανομές του λόγου των ετεροατόμων είναι πιο ασύμμετρες από ό,τι ο λόγος H/C διότι πολλές δομές δεν έχουν καθόλου ετεροάτομο (π.χ. αλκάνια) και υπάρχουν σπάνιες περιπτώσεις με υψηλές αναλογίες ετεροατόμων προς άνθρακα λόγων. Για αυτό γίνεται χρήση των λόγων του προηγούμενου πίνακα.

6. Έλεγχος πιθανότητας συνύπαρξης πολλών στοιχείων σε έναν μοριακό τύπο

Ο κανόνας 5 περιορίζει μόνο απίθανα υψηλούς λόγους στοιχείων σε μοριακούς τύπους αλλά δεν ελέγχει για πολλαπλά υψηλά στοιχεία σε ένα τύπο. Για παράδειγμα ο τύπος $C_{26}H_{28}N_{17}O_1P_3S_8$ περιέχει πολλά διαφορετικά στοιχεία. Αυτή η δομή περνάει όλους τους κανόνες συμπεριλαμβανομένων και του ελέγχου του λόγου στοιχείων. Παρόλα αυτά ο συνδυασμός όλων αυτών των διαφορετικών στοιχείων είναι απίθανος. Για αυτό συμπεριλήφθηκε ένας επιπρόσθετος περιορισμός για πολλαπλά στοιχεία σε μια μοριακή δομή. Όλοι οι συνδυασμοί για τα στοιχεία N,O,P,S και όλοι οι τριπλοί συνδυασμοί συμπεριλήφθηκαν σε ένα υποκανόνα.

Πίνακας 4: Πολλαπλός περιορισμός στοιχείων σε μοριακό τύπο για ενώσεις με μοριακό βάρος μικρότερο από 2000 Da βασισμένο στην βάση δεδομένων Beilstein και το «Dictionary of Natural Products», πηγή [19]

Αριθμός στοιχείων	Πειραματικός κανόνας	Παραδείγματα από βάσεις δεδομένων με τις μέγιστες τιμές
NOPS all>1	N<10,O<20,P<4,S<3	C ₁₅ H ₃₄ N ₉ O ₈ PS, C ₂₂ H ₄₄ N ₄ O ₁₄ P ₂ S ₂ , C ₂₄ H ₃₈ N ₇ O ₁₉ P ₃ S
NOP all>3	N<11,O<22,P<6	C ₂₀ H ₂₈ N ₁₀ P ₂₁ S ₄ , C ₁₀ H ₁₈ N ₅ O ₂₀ P ₅
OPS all>1	O<14,P<3,S<3	C ₂₂ H ₄₄ N ₄ O ₁₄ P ₂ S ₂ , C ₁₆ H ₃₆ N ₄ O ₄ P ₂ S ₂
PSN all>1	P<3,S<3,N<4	C ₂₂ H ₄ N ₄ O ₁₄ P ₂ S ₂ , C ₁₆ H ₃₆ N ₄ O ₄ P ₂ S ₂
NOS all>6	N<19,O<14,S<8	C ₅₉ H ₆₄ N ₁₈ O ₁₄ S ₇

Για συνδυασμούς των N,O,P και N,O,S ο υψηλότερος αριθμός στοιχείων που παρατηρήθηκε στον πιο «ακραίο» μοριακό τύπο τέθηκε ως άνω όριο.

7. Αφαίρεση πιθανής παρουσίας τριμεθυλοπυριτιακών συστατικών για περιπτώσεις ανάλυσης GC-MS

Στην περίπτωση χρήσης αεριοχρωματογραφίας συζευγμένης με φασματομετρία μαζών τότε συχνά πραγματοποιείται παραγωγοποίηση των αναλυτών προκειμένου αυτοί να καθιστούν πιο πτητικοί. Ένα σύνηθες αντιδραστήριο παραγωγοποίησης είναι το N-μεθυλοτριμεθυλοσιλανο-τριφθορο ακεταμίδιο (CAS: 24589-78-4) που ανταλλάσσει τα όξινα πρωτόνια των αναλυτών με ομάδες τριμεθυλοσιλανίου. Έτσι θα πρέπει από τις μοριακές δομές που προκύπτουν να αφαιρεθούν τα πλεονάσματα που προκύπτουν από το αντιδραστήριο παραγωγοποίησης. Δηλαδή αφαιρούνται από τους μοριακούς τύπους το τμήμα “C₃H₈Si”. Για παράδειγμα οι ακριβείς μάζες για τους τύπους C₂₀H₅₈N₆I₄Si₆ και C₂₄H₆₂O₆Si₆ διαφέρουν μόλις 4 ppm αλλά μετά από αφαίρεση των ομάδων τριμεθυλοσιλανίου προκύπτουν οι τύποι C₂H₁₀N₆O₄ και C₆H₁₄O₆ με την πρώτη μοριακή δομή να είναι πιο πιθανή. Για παράγωγα

του τριμεθυλοσιλανίου εφαρμόζονται οι κανόνες για στοιχειακές αναλογίες, μετά όμως από την αφαίρεση των ομάδων τριμεθυλοσιλανίου από το μοριακό τύπο.

Τέλος οι πιο πιθανές μοριακές δομές που συμμορφώνονται με όλους τους χρυσούς κανόνες κατατάσσονται σύμφωνα με το ισοτοπικό προφίλ και έπειτα αναζητούνται σε βάσεις δεδομένων (PubChem, ChemSpider) [19].

2.2 Bruker SmartFormula, SmartFormula Manually, SmartFormula3D®

Ένα λογισμικό παρόμοιο σε λογική με το SGR είναι το SmartFormula που παρέχεται μαζί με το βασικό πακέτο επεξεργασίας δεδομένων LC-MS με τα όργανα της εταιρείας Bruker, το DataAnalysis (τωρινή έκδοση 4.1). Στόχος του λογισμικού που διαρθρώνεται σε τρεις υπο-λογισμικές ομάδες (SmartFormula, SmartFormula Manually και SmartFormula3D) είναι η απόδοση πιθανών μοριακών τύπων σε μοριακά ιόντα ή ακόμα και άλλα ιόντα, όπως θραύσματα, λαμβάνοντας υπόψη τους την πληροφορία της ακριβούς μάζας που προσφέρουν τα όργανα υψηλής διακριτικής ικανότητας και το ισοτοπικό προφίλ. Ο στόχος είναι να καθοριστεί η στοιχειακή σύνθεση ενός ιόντος χρησιμοποιώντας ως περιορισμούς τις πληροφορίες που παρέχονται από την ακριβή μάζα και το ισοτοπικό προφίλ. Συνήθως όσο υψηλότερη η μάζα τόσο περισσότεροι πιθανοί μοριακοί τύποι υπάρχουν. Για παράδειγμα, για μια μάζα 1000,0 Da υπάρχουν περίπου 66.000 μοριακοί τύποι αποτελούμενοι μόνο από τα κύρια στοιχεία (C,H,O,N) ενώ για μια μάζα 2000,0 Da η τιμή αυτή αυξάνεται σε 520.000. Αν υπάρχουν και άλλα χημικά στοιχεία τότε υπάρχουν περισσότεροι πιθανοί συνδυασμοί. Εξαιτίας του υψηλού αριθμού συνδυασμών είναι πολύ χρονοβόρο να εξεταστούν όλοι και να συγκριθεί η μάζα που προκύπτει από το μοριακό τύπο με την πειραματικά μετρηθείσα μάζα. Για αυτό πρέπει να τεθούν περιορισμοί στις πιθανές λύσεις. Το πρόβλημα εδώ είναι να τεθούν όρια με τέτοιο τρόπο ώστε να μην εξαιρείται ο σωστός στοιχειακός συνδυασμός. Επομένως απαιτείται ένας «έξυπνος» αλγόριθμος ώστε να μειωθούν οι πιθανοί υπό εξέταση συνδυασμοί. Το DataAnalysis χρησιμοποιεί ένα αλγόριθμο που βασίζεται στο γεγονός ότι το ακέραιο μέρος και το κλασματικό μέρος της μοριακής μάζας είναι γραμμικώς ανεξάρτητα μεγέθη για οργανικά μόρια με μάζα έως 1000 Da. Αυτό χρησιμοποιείται για το συνδυασμό στοιχείων C,H,N και O. Τα άλλα στοιχεία, τα οποία συνήθως εμφανίζονται σε

μικρότερους αριθμούς μέσα στο μοριακό τύπο σε σχέση με τα κύρια χρησιμοποιούνται με τη μέθοδο της δοκιμής και του λάθους.

Meas. m/z	#	Ion Formula	m/z	err [ppm]	mSigma	# Sigma	Score	rdb	e ⁻ Conf	N-Rule	Adduct
401.3027	8	C18H41F3N4P	401.3026	-0.3	11.3	8	100.00	-0.5	even	ok	M-H
7	C16H38N10P	401.3024	-0.9	11.2	7	90.08	3.5	even	ok	M-H	
5	C21H41N2O5	401.3021	1.6	10.3	5	79.14	2.5	even	ok	M-H	
10	C17H37F4N6	401.3021	-1.5	12.7	10	77.02	0.5	even	ok	M-H	
9	C19H34FN12	401.3019	2.1	12.7	9	67.93	4.5	even	ok	M-H	
22	C12H35F2N12O	401.3030	0.7	28.1	22	64.67	0.5	even	ok	M-H	
18	C22H37N6O	401.3034	-1.7	25.2	18	57.17	7.5	even	ok	M-H	
20	C13H39FN10OP	401.3035	2.0	25.9	20	53.04	-0.5	even	ok	M-H	
27	C27H39F2	401.3025	-0.5	37.8	27	52.21	7.5	even	ok	M-H	
14	C14H37N12Si	401.3039	2.8	19.0	14	50.87	3.5	even	ok	M-H	
19	C24H40F3O	401.3037	2.3	25.2	19	50.22	3.5	even	ok	M-H	
31	C22H45N2S2	401.3030	-0.5	44.1	31	43.92	1.5	even	ok	M-H	
15	C16H40F3N6Si	401.3041	3.4	19.0	15	43.83	-0.5	even	ok	M-H	
3	C19H38FN6O2	401.3046	4.6	8.8	3	39.47	3.5	even	ok	M-H	
33	C21H45N2O5Si	401.3027	-0.0	51.9	33	37.79	1.5	even	ok	M-H	

Εικόνα 5: Στιγμιότυπο οθόνης SmartFormula Manually για τον υπολογισμό στοιχειακής σύνθεσης

Επιπλέον ο υπολογισμός του μοριακού τύπου βασίζεται στις ισοτοπικές μάζες, τις αφθονίες που εμφανίζουν και το σθένος των ατόμων. Τα σθένη χρησιμοποιούνται για τον υπολογισμό δακτυλίων και πολλαπλών δεσμών, την διευθέτηση ηλεκτρονίων και τον κανόνα του αζώτου.

Στο DataAnalysis 4.1 υπάρχει επιλογή μεμονωμένης μετατροπής ενός ιόντος σε μοριακό τύπο (υπομονάδα SmartFormula manually) και αυτόματης μετατροπής όλων των εν αφθονία μαζών σε όλα τα φάσματα (SmartFormula). Επιπλέον υπάρχει η επιλογή SmartFormula3D η οποία υπολογίζει το μοριακό τύπο συνδυάζοντας τους επιμέρους τύπους του μοριακού ιόντος και των θραυσμάτων. Πιο συγκεκριμένα το SmartFormula3D απαιτεί δυο φάσματα, ένα φάσμα MS και ένα MS/MS, τα οποία μπορούν να προέρχονται και από διαφορετικές αναλύσεις. Αξιολογεί τα αποτελέσματα μετατροπής των μαζών σε μοριακούς τύπους του πρόδρομου ιόντος και των θραυσμάτων. Κατά βάση ελέγχει κατά πόσον κάποιοι από τους μοριακούς τύπους των θραυσμάτων είναι υποσύνολο του πρόδρομου ιόντος. Ακόμα ελέγχει αν από το μοριακό τύπο

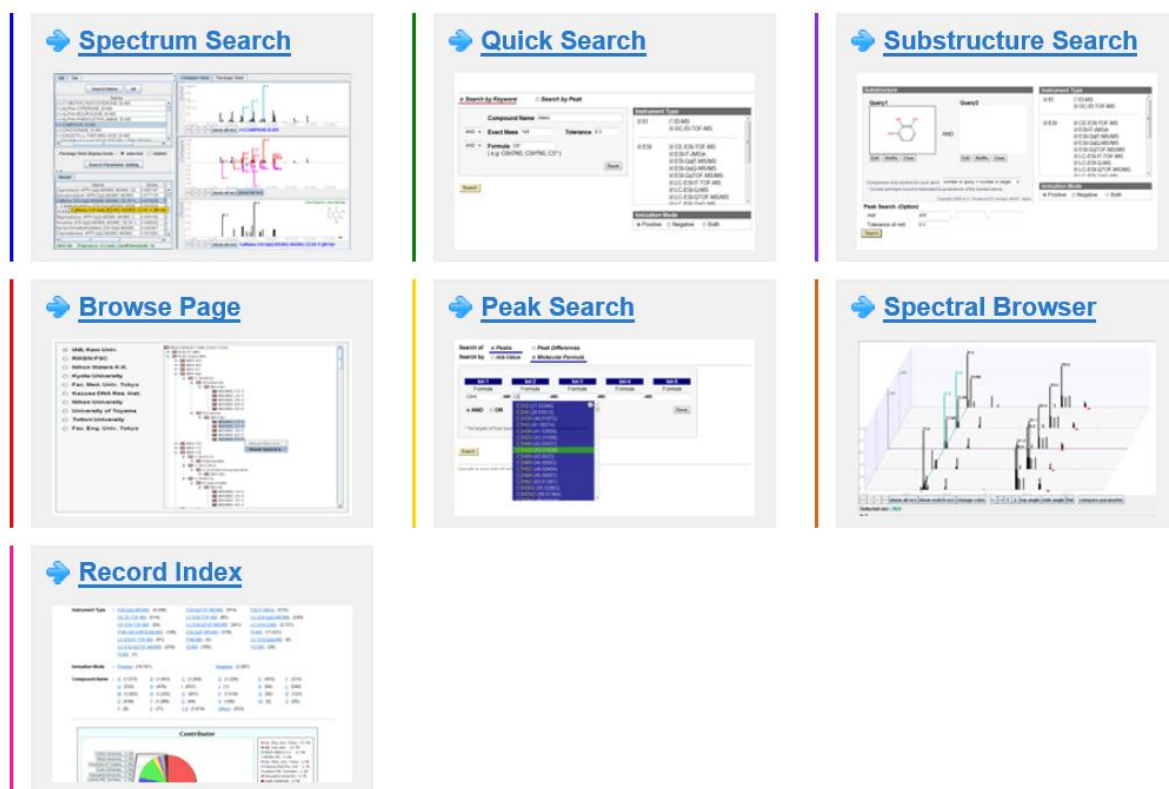
υπάρχουν χαρακτηριστικές ουδέτερες απώλειες (neutral losses) που να προκύπτουν από την αφαίρεση των δυο μοριακών τύπων οι οποίες ταιριάζουν με την παρατηρηθείσα διαφορά στη μάζα των κορυφών. Έτσι επιτυγχάνεται μείωση του αρχικά μεγάλου αριθμού μοριακών τύπων που έχουν αποδοθεί στο ψευδομοριακό ιόν χρησιμοποιώντας όμως και μέρος της πληροφορίας θραυσματοποίησης.

Και για τις τρεις μονάδες SmartFormula είναι σημαντικές οι παράμετροι που θα χρησιμοποιηθούν για την μετατροπή μιας μάζας σε μοριακό τύπο. Ο πιο σημαντικός είναι η παράμετρος ανοχής στην ακρίβεια μάζας. Όσο μεγαλύτερος τόσο περισσότερα αποτελέσματα προκύπτουν. Επιπλέον καθοριστική παράμετρος είναι ποια στοιχεία πέρα τον βασικών (C,N,O,H) μπορούν να περιέχονται στο μοριακό τύπο. Στην εικόνα 5, έχουν προστεθεί το θείο, το φθόριο, το χλώριο, ο φωσφόρος, το βρώμιο και το πυρίτιο. Τέλος ρόλο παίζει και η ηλεκτρονιακή διάταξη. Άρτια ηλεκτρονιακή διάταξη έχουν τα μοριακά ιόντα ενώ τα θραύσματα μπορεί να έχουν και άρτια και περιττή (ριζικό ιόν). Επιπλέον παράμετροι ανοχής υπάρχουν στο SmartFormula3D για το φάσμα MS/MS όπου καθορίζεται η ακρίβεια μέσα στην οποία καθορίζονται θραύσματα ώστε να βρεθούν οι μοριακοί τύποι των θραυσμάτων και άρα να οριστεί η ανοχή μάζας για τις ουδέτερες απώλειες. Αντίστοιχοι παράμετροι υπάρχουν και στο φάσμα MS, όπου καθορίζουμε την ακρίβεια μάζας και τα στοιχεία με βάση τα οποία θα αποδοθεί στο θεωρητικό μοριακό ιόν κάποιος μοριακός τύπος. Καθοριστική παράμετρος είναι και το όριο του συντελεστή ομοιότητας μεταξύ του θεωρητικού ισοτοπικού προφίλ και του πειραματικού ισοτοπικού προφίλ (mSigma). Τέλος, το σφάλμα στον υπολογισμό της μάζας (mass error) και ο όρος mSigma εμπλέκονται στον υπολογισμό της βαθμολογίας (score) που το πρόγραμμα αποδίδει στους υποψήφιους μοριακούς τύπους [20, 21].

2.3 MassBank

Το MassBank είναι η πρώτη δημόσια βάση δεδομένων ανταλλαγής φασμάτων για μικρά μόρια (<3000 Da) και περιέχει φάσματα υψηλής ποιότητας τόσο MS που έχουν ληφθεί κυρίως είτε με ηλεκτροϊοντισμό είτε με ηλεκτροψεκασμό αλλά και μερικά φάσματα από λιγότερο συχνές πηγές ιοντισμού όπως MALDI και FAB όσο και φάσματα MS/MS ή MSⁿ (όπου n=3 έως 5). Η βάση δεδομένων

ήρθε να καλύψει την ανάγκη αναζήτησης φασμάτων σε επιστημονικά περιοδικά, η οποία είναι πολύ χρονοβόρα. Έχει τη δυνατότητα φιλοξενίας φασμάτων που έχουν παρθεί με οποιαδήποτε οργανολογία φασματομέτρου μαζών και ποικίλες πηγές ιοντισμού. Λόγω του γεγονότος ότι δεν υπάρχει καθιερωμένο πειραματικό πρωτόκολλο για δεδομένα ESI-MS/MS σε αντίθεση με τα δεδομένα EI-MS/MS, συνήθως προστίθενται φάσματα MS/MS σε διάφορες ενέργειες θραυσματοποίησης. Η βάση δεδομένων επιτρέπει στους ερευνητές τον διαμοιρασμό των φασμάτων τους σε τοπικούς υπολογιστές και όχι σε ένα κεντρικό υπολογιστή, γεγονός που διασφαλίζει την ποιότητα των φασμάτων αφού αποτρέπει την ανάμιξη των φασμάτων διαφορετικών ερευνητικών ομάδων.



Εικόνα 6: Κεντρικό μενού βάσης δεδομένων MassBank

2.3.1 MassBank από την οπτική γωνία του διαχειριστή

Η αρχιτεκτονική του MassBank βασίζεται στις αρχές ότι τα φάσματα είναι δημόσια, πρέπει να διανέμονται και να είναι προσβάσιμα μέσω διαδικτύου και πρέπει να διαρθρώνονται σε μια καθορισμένη δομή. Ιδανικά κάθε επιστημονική ομάδα συνεισφοράς πρέπει να έχει έναν τοπικό διακομιστή (server) φιλοξενίας

των φασμάτων της. Επιπλέον η πλατφόρμα ικανοποιεί και τις δυο οπτικές γωνίες, δηλαδή και αυτή του χρήστη που αναζητεί κάποιο φάσμα και ως εκ τούτου θέλει μια κεντρική βάση δεδομένων αναζήτησης φασμάτων και αυτή της ομάδας που συνεισφέρει τα φάσματα της και πρέπει να είναι σε θέση να επεξεργάζεται και να διαχειρίζεται τα φάσματα της ξεχωριστά.

Προκειμένου να ικανοποιηθούν οι ανάγκες αυτές το MassBank διαρθρώνεται σε μια αρχιτεκτονική τριών στιβάδων: τη στιβάδα της βάσης δεδομένων με τα δεδομένα υπο μορφή καταχωρήσεων MySQL, τη στιβάδα εφαρμογής η οποία είναι μια μηχανή αναζήτησης που παραπέμπει και κάνει ερωτήματα στη στιβάδα βάσης δεδομένων και τη στιβάδα παρουσίασης που είναι το γραφικό περιβάλλον του χρήστη.

Η εγκατάσταση ενός τοπικού υπολογιστή MassBank είναι τυποποιημένη και μπορεί να βρεθεί στην ιστοσελίδα SourceForge.net όπου πέρα από τα αρχεία της βάσης δεδομένων μπορούν να βρεθούν τα απαιτούμενα λογισμικά λειτουργίας της σελίδας (στιβάδα εφαρμογής) που είναι οι πλατφόρμες Apache, Tomcat και MySQL. Αναβαθμίσεις του πηγαίου κώδικα μπορούν να ελέγχονται και να εγκαθίστανται αυτόματα μέσω της μητρικής σελίδας MassBank.jp.

Το MassBank δεν δέχεται ως είσοδο δυαδικά δεδομένα που προκύπτουν από τα φασματομέτρα μαζών, αλλά έχει μια καθορισμένη δομή εγγραφής. Κάθε δομή εγγραφής περιέχει ένα χημικό συστατικό με συγκεκριμένο μοριακό τύπο και έχει αρθρωτή δομή. Πιο συγκεκριμένα δομή μιας εγγραφής MassBank (MassBank record) από την οπτική γωνία του επιστήμονα που θα συνεισφέρει τα φάσματά του στη βάση δεδομένων είναι η ακόλουθη:

Πίνακας 5: Δομή καταγραφής MassBank (MassBank Record), πηγή [22]

Πεδίο	Περιγραφή πεδίου
1.Γενικό τμήμα (Summary section)	
ACCESSION	Αριθμός πρόσβασης
RECORD_TITLE	Μικρή περιγραφή που περιλαμβάνει το όνομα του συστατικού που αναλύθηκε και την αναλυτική μέθοδο
DATA	Ημερομηνία διαμοιρασμού
AUTHORS	Συγγραφείς και το ίδρυμα που ανήκουν
COPYRIGHT	Σημείωση πνευματικών δικαιωμάτων
2.Χημικό τμήμα (Chemical section)	
CH\$NAME	Χημικό όνομα του συστατικού που αναλύθηκε
CH\$COMPOUND_CLASS	Χημική τάξη του συστατικού
CH\$FORMULA	Μοριακός τύπος συστατικού
CH\$EXACT_MASS	Ακριβής μάζα του συστατικού
CH\$SMILES	Κωδικοποίηση SMILES του συστατικού
CH\$IUPAC	Κωδικός InChI της χημικής δομής
3.Αναλυτικό τμήμα (Analytical section)	
AC\$INSTRUMENT	Αναλυτής μαζών και όνομα κατασκευαστή
AC\$INSTRUMENT_TYPE	Τύπος αναλυτή μαζών
AC\$ANALYTICAL_CONDITION/MODE	Τύπος ιοντισμού
4.Φασματικό τμήμα (Spectral section)	
PK\$NUM_PEAK	Αριθμός κορυφών
PK\$PEAK	Δεδομένα: m/z, ένταση και σχετική ένταση
5.Λοιπα (Others)	
MOLFILE_NAME	Όνομα αρχείου molfile που καθορίζει την δομή του συστατικού που αναλύθηκε

Οι επιστήμονες που συνεισφέρουν πρέπει να φτιάξουν κατάλληλες εγγραφές MassBank και να τις αποθηκεύσουν στους δικούς τους διακομιστές. Το MassBank παρέχει δυο εργαλεία το Record Editor και το Administration tool. Επιπλέον υπάρχει και ελεύθερο λογισμικό Mass++ που δέχεται αρχεία ελεύθερης μορφής (mzML) και βγάζει ως έξοδο εγγραφή MassBank που περιέχει τη φασματική πληροφορία, όπου έπειτα συνδυάζεται μέσω του Record editor με το αρχείο molfile που καθορίζει την δομή του συστατικού. Έτσι οι πληροφορίες συνδυάζονται και απαιτείται πολύ λιγότερη χειροκίνητη εργασία. Το δεύτερο εργαλείο (Administration tool) βοηθάει τους επιστήμονες να διαχειρίζονται τα φάσματά τους [23, 24]. Πρόσφατα αναπτύχθηκε το πακέτο της R με όνομα RMassBank, το οποίο επίσης είναι ικανό να παράξει υψηλής ποιότητας εγγραφές MassBank με μια αυτοματοποιημένη διαδικασία [25]. Το εργαστήριο Αναλυτικής Χημείας του τμήματος Χημείας ΕΚΠΑ έχει ήδη προσθέσει 68 φάσματα MS/MS στη βάση δεδομένων του MassBank (<http://massbank.ufz.de/MassBank/RecordIndex.html>) και πρόκειται στο άμεσο μέλλον να προστεθούν περισσότερα από 1000 φάσματα MS/MS ουσιών.

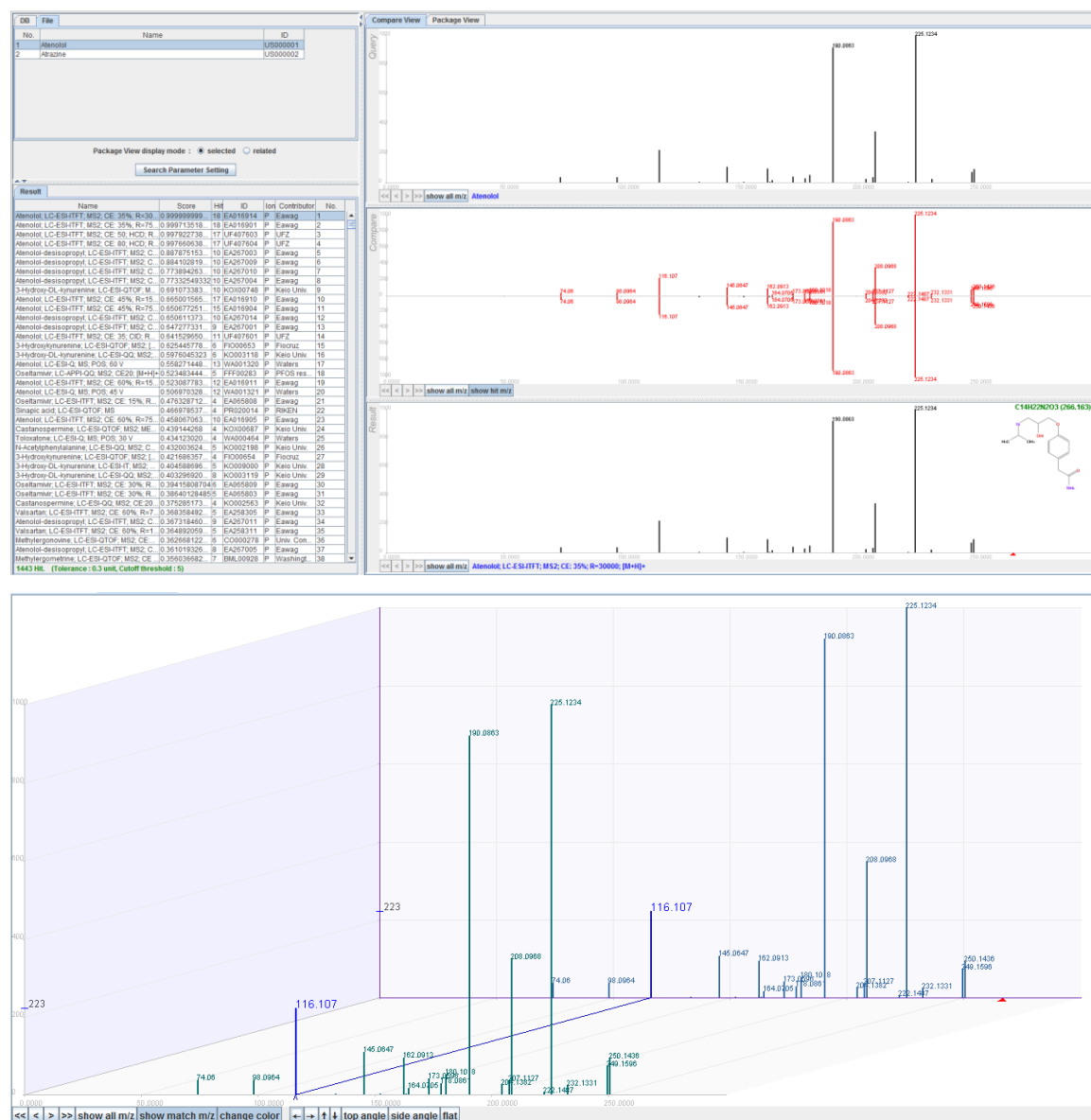
2.3.2 MassBank από την οπτική γωνία του χρήστη

Από την οπτική μεριά του χρήστη αυτός μπορεί να αναζητήσει συστατικά με βάση το όνομα τους, την ακριβή τους μάζα (θέτοντας όριο ανοχής), ή/και με βάση το μοριακό τους τύπο ή την δομή (Quick Search και Substructure search αντίστοιχα). Τέλος, δίνεται η δυνατότητα επιλογής της οργανολογίας από το οποίο πάρθηκε το φάσμα, το επίπεδο MS που πάρθηκε και το είδος ιοντισμού. Διαφορετικά υπάρχει η δυνατότητα ταυτοποίησης πειραματικών φασμάτων MS/MS με φάσματα της βιβλιοθήκης με την εισαγωγή των θραυσμάτων που λήφθηκαν πειραματικά.

The image shows two side-by-side screenshots of the MassBank search interface. The left panel is titled 'Search by Keyword' and features a 'Compound Name' input field, 'Exact Mass' and 'Tolerance' fields, and a 'Formula' field with an example '(e.g. C8H7NS, C5H4N5, C5*)'. Below these are 'AND' dropdowns and a 'Search' button. The right panel is titled 'Search by Peak' and shows a list of 'Peak Data' with columns for m/z and relative intensity. It includes a 'Cutoff threshold of relative intensities' field set to 5 and a 'Number of Results' dropdown set to 20. Both panels have 'Search' buttons at the bottom.

Εικόνα 7: Αναζήτηση φασμάτων από τη βάση δεδομένων MassBank

Για τη σύγκριση των πειραματικών φασμάτων και φασμάτων βιβλιοθήκης αποδίδεται μια βαθμολογία με μέγιστο το 100% που στηρίζεται σε ένα βαθμό ομοιότητας που υπολογίζεται με βάση μια τροποποιημένη συνημιτοειδή σχέση που προτάθηκε από τους Stein και Scoot [26]. Έπειτα τα υποψήφια φάσματα που ταιριάζουν περισσότερο με το πειραματικό μπορούν να τεθούν σε αντιπαράθεση ή και σε αναπαράσταση τριών διαστάσεων όπως φαίνεται και στην εικόνα 8 (spectral search και spectral browser).



Εικόνα 8: Σύγκριση φασμάτων με παράθεση και τρισδιάστατη απεικόνιση από τη βάση δεδομένων MassBank

Επιπλέον δίνεται η δυνατότητα εύρεσης όλων των φασμάτων MSⁿ τα οποία περιέχουν ως κορυφές κάποιες μάζες εντός ενός εύρους ανοχής. Ακόμα

υπάρχει η επιλογή εύρεσης των συστατικών τα οποία περιέχουν μια ή περισσότερες κορυφές των οποίων οι τιμές m/z διαφέρουν κατά μια δεδομένη τιμή. Έτσι καθίσταται δυνατή η ανεύρεση μορίων που εμφανίζουν χαρακτηριστικές ουδέτερες απώλειες (Peak search στο κεντρικό μενού).

Τέλος το MassBank δίνει τη δυνατότητα δημιουργίας ενός τεχνητού φάσματος το οποίο έχει προκύψει ως αποτέλεσμα συγκερασμού φασμάτων που έχουν παρθεί σε διαφορετικές ενέργειες θραυσματοποίησης. Αυτό κυρίως έχει επωφελή αποτελέσματα στην περίπτωση πηγής ιοντισμού ESI που εμφανίζει σχετικά χαμηλή επαναληψιμότητα. Έχει αποδειχτεί ότι η αναζήτηση ενός πειραματικού φάσματος με συγχωνευμένα φάσματα οδηγεί σε πιο αξιόπιστες ανιχνεύσεις άγνωστων ουσιών [24, 27].

2.4 MetFrag

Το φάσμα MS/MS περιέχει τα θραυσματοποιημένα ιόντα τα οποία περιέχουν αρκετή χημική πληροφορία την οποία δεν είναι σε θέση οι φασματικές βιβλιοθήκες να αξιοποιήσουν εξ' ολοκλήρου. Αυτό συμβαίνει, διότι οι βιβλιοθήκες φασμάτων περιέχουν ένα περιορισμένο αριθμό φασμάτων προτύπων ουσιών αναφοράς που δεν καλύπτουν όλη τη χημική ποικιλότητα. Σε αντίθεση υπάρχουν χημικές βιβλιοθήκες όπως το PubChem ή το KEGG που περιέχουν ένα αρκετά μεγάλο αριθμό συστατικών που μπορούν να χρησιμοποιηθούν για να προβλεφθεί η *in silico* θραυσματοποίησή τους και έπειτα να συγκριθούν τα *in silico* φάσματα με τα φάσμα των υποψήφιων δομών κάποιας ουσίας. Το MetFrag δέχεται ως είσοδο μια λίστα υποψήφιων δομών από χημικές βιβλιοθήκες με βάση το μοριακό ιόν που εισάγουμε ή με βάση τον μοριακό τύπο που μπορούμε να εισάγουμε εναλλακτικά και έπειτα τις κατατάσσει με βάση την συμφωνία μεταξύ των *in silico* θραυσμάτων και των θραυσμάτων που πάρθηκαν πειραματικά. Η *in silico* θραυσματοποίηση απαιτεί λιγότερο από 1 δευτερόλεπτο ενώ η αναζήτηση σε κάποια χημική βιβλιοθήκη όπως KEGG ή PubChem παίρνει χρονικά 30 και 300 δευτερόλεπτα αντίστοιχα. Το MetFrag είναι προσβάσιμο μέσω της ιστοσελίδας <http://msbi.ipb-halle.de/MetFrag/>, η οποία επιτρέπει αξιολόγηση των αποτελεσμάτων ενώ επιτρέπει και την εκτέλεση ομάδων αναζητήσεων (batch searches).

Αντίστοιχα εμπορικά εργαλεία όπως ο ACD Fragmenter ή το Mass Frontier χρησιμοποιούν μια σειρά κανόνων για να παράξουν τα θραύσματα με βάση κανόνες διάσπασης που είναι γνωστοί από τη βιβλιογραφία με αλγορίθμους οι λεπτομέρειες των οποίων δεν έχουν δημοσιευτεί. Το MetFrag λειτουργεί συνδυαστικά (combinational fragmenter) δηλαδή προσπαθεί να προβλέψει το δέντρο θραυσματοποίησης δεδομένου τόσο της μοριακής δομής του αναλύτη όσο και του φάσματος MS/MS. Επιπλέον χρησιμοποιεί την προσέγγιση διάσπασης δεσμών η οποία είναι αρκετά γρήγορη για να προβλέψει τη θραυσματοποίηση σε αρκετά υποψήφια συστατικά που εισάγονται από χημικές βάσεις δεδομένων (KEGG, PubChem, ChemSpider). Ο στόχος δεν είναι να προβλεφθεί μια μηχανιστικά σωστή θραυσματοποίηση αλλά να πραγματοποιηθεί μια αναζήτηση στις χημικές βιβλιοθήκες χρησιμοποιώντας τα θραύσματα ως επιπλέον χαρακτηριστικά στοιχεία της δομής των ενώσεων.

Η ροή εργασίας που εφαρμόζεται στο MetFrag παρουσιάζεται στην παρακάτω εικόνα:



Εικόνα 9: Η ροή εργασίας του MetFrag, πηγή [28]

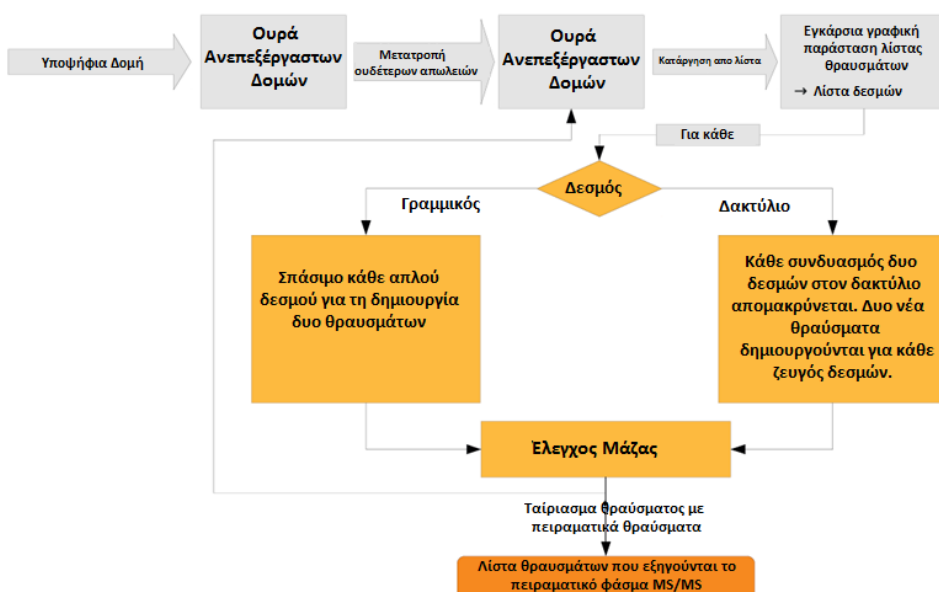
Το MetFrag είναι εργαλείο που έχει αναπτυχθεί σε γλώσσα προγραμματισμού Java και χρησιμοποιεί την βιβλιοθήκη Chemistry Development Kit (CDK), η οποία παρέχει αλγόριθμους και δομές δεδομένων κατάλληλες για χημειοπληροφορική.

Πρώτα πραγματοποιείται μια αναζήτηση γενικού σκοπού της χημικής βάσης δεδομένων για υποψήφια μόρια με βάση την ακριβή μάζα και εντός ενός ορίου εύρους σφάλματος εκφρασμένο σε ppm. Μέχρι στιγμής τρεις βάσεις δεδομένων

είναι προσβάσιμες: KEGG, PubChem και ChemSpider. Προαιρετικά, η αναζήτηση μπορεί να περιοριστεί σε συστατικά που περιέχουν μόνο τα στοιχεία C,H,N,O,P,S που είναι και τα πιο κοινά στα φυσικά προϊόντα. Εναλλακτικά, η αναζήτηση στις χημικές βάσεις δεδομένων μπορεί να γίνει με βάση την στοιχειακή σύνθεση εάν αυτή έχει προκύψει για παράδειγμα από το ισοτοπικό προφίλ και την ακριβή μάζα του μοριακού ιόντος. Υπάρχει ακόμα η επιλογή να εισαχθούν τα συστατικά με βάση το μοναδικό αριθμό που κατέχουν σε κάθε χημική βάση δεδομένων.

Το MetFrag δημιουργεί όλα τα δυνατά τοπολογικά θραύσματα ενός υποψήφιου συστατικού, έτσι ώστε να ταιριάζει με τα θραύσματα που φάσματος που λήφθηκε πειραματικά. Το πρόβλημα απαρίθμησης όλων των πιθανών μοριακών θραυσμάτων μπορεί να επιλυθεί με τη δημιουργία ενός δέντρου θραυσματοποίησης. Οι ρίζες αποτελούνται από άθικτο το μόριο και κάθε κλάδος αναπαριστά ένα θραύσμα που αποκτάται σπάζοντας έναν δεδομένο δεσμό από το μόριο. Ένας σημαντικός παράγοντας που καθορίζει την ταχύτητα είναι ο αριθμός των θραυσμάτων που δημιουργούνται. Έτσι το μέγιστο βάθος δέντρου που εισάγεται ως παράγοντας επιλογής με στόχο τη βελτίωση της απόδοσης και εξειδίκευσης του λογισμικού.

Για κάθε υποψήφια δομή τα θραύσματα δημιουργούνται με τρόπο που αναπαρίσταται στην ακόλουθη εικόνα:



Εικόνα 10: Αλγόριθμος για in silico θραυσματοποίηση, πηγή [28]

Αρχικά κάθε υποψήφια δομή προωθείται σε μια «ανεπεξέργαστη» λίστα. Η υποψήφια δομή επεξεργάζεται με τη χρήση μιας μικρής ομάδας κανόνων που περιγράφουν μοριακές ανακατατάξεις κατά τη διάρκεια της θραυσματοποίησης που δεν λαμβάνονται υπόψη από την απλή προσέγγιση διάσπασης χημικών δεσμών. Κάθε εφαρμογή των κανόνων αυτών οδηγεί σε ένα ή περισσότερα θραύσματα τα οποία προστίθενται στη ανεπεξέργαστη λίστα. Έπειτα, επιλέγεται μια δομή και διασχάζονται όλοι οι δεσμοί και προκύπτουν τα θραύσματά της. Ένας γραμμικός δεσμός που δεν είναι μέρος αρωματικού δακτυλίου χρειάζεται μόνο να διασπαστεί και καταλήγει σε δυο νέα θραύσματα. Μέσα σε ένα δακτύλιο δυο δεσμοί θα πρέπει να σπάσουν ταυτόχρονα και να δημιουργήσει νέα θραύσματα. Μόνο θραύσματα μεγαλύτερα σε μάζα από την κορυφή με τη χαμηλότερη μάζα στο φάσμα που εισάγεται δημιουργούνται αφού μικρότερα θραύσματα δεν μπορούν να εξηγήσουν κάποια πειραματική κορυφή [28].

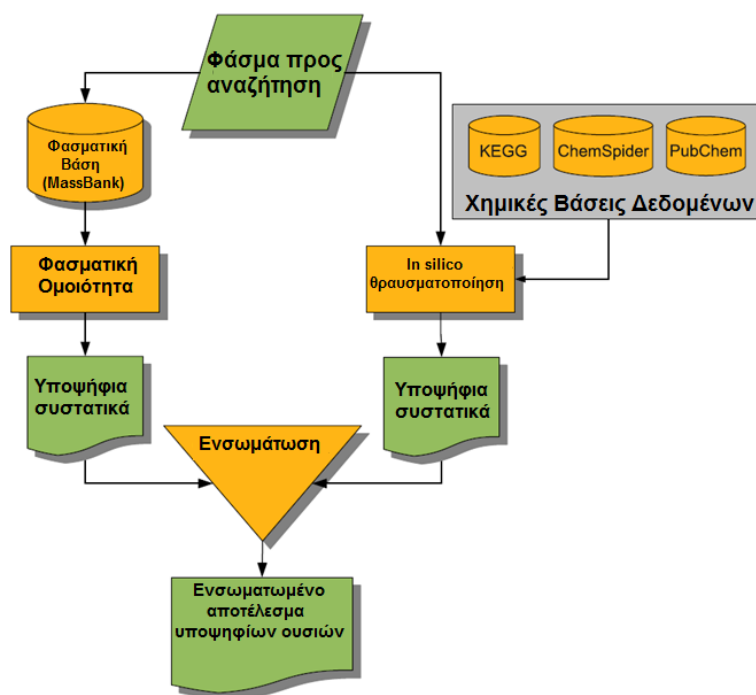
2.5 MetFusion

Από την μια μεριά το MassBank περιέχει φάσματα από διάφορα όργανα, από διαφορετικούς κατασκευαστές και σε διαφορετικές πειραματικές συνθήκες και δίνει την ευκαιρία σε διάφορα ινστιτούτα και πανεπιστήμια να συνεισφέρουν τα φάσματά τους [24]. Από την άλλη μεριά χημικές βάσεις δεδομένων όπως το PubChem, KEGG ή ChemSpider παρέχουν πληροφορίες για ένα μεγάλο αριθμό τόσο φυσικών όσο και συνθετικών ενώσεων για τις οποίες όμως δεν περιέχουν φασματομετρικές μετρήσεις. Επιπλέον έχουν αναπτυχθεί υπολογιστικά φασματομετρικά εργαλεία όπως το MetFrag [28] με *in silico* υπολογισμό της θραυσματοποίησης υποψήφιων μορίων. Το MetFrag αντιστοιχίζει τα *in silico* προβλεφθέντα θραύσματα με τα θραύσματα του πειραματικού φάσματος και ανάλογα με τον αριθμό των θραυσμάτων που εξηγούνται από την *in silico* θραυσματοποίηση, αποδίδει σε κάθε υποψήφια δομή ένα σκορ. Παρόλο την ύπαρξη των *in silico* εργαλείων, τα φάσματα προτύπων ουσιών που έχουν αποκτηθεί από συμβατά αναλυτικά όργανα είναι ακόμα ο προτιμώμενος δρόμος για να επιτευχθεί αξιόπιστη ανίχνευση μιας ένωσης.

Μια νέα στρατηγική που λέγεται Metfusion συνδυάζει τα αποτελέσματα από την *in silico* θραυσματοποίηση του Metfrag που τροφοδοτείται με μόρια από χημικές βάσεις δεδομένων και τη βιβλιοθήκη φασμάτων MassBank ή Metlin. Σκοπός του Metfusion είναι η βελτίωση της ανίχνευσης ουσιών περιβαλλοντικού ή μεταβολομικού ενδιαφέροντος (ανάλογα με τη χημική και φασματική βάση δεδομένων που χρησιμοποιείται).

2.5.1 Αρχιτεκτονική του συστήματος

Η υπόθεση στο MetFusion είναι ότι το σωστό συστατικό είναι παρών σε κάποια βιβλιοθήκη (KEGG, ChemSpider, PubChem) και κατά συνέπεια μεταξύ των υποψήφιων δομών στο αποτέλεσμα του MetFrag. Η ιδέα του MetFusion είναι να επιβεβαιώσει τα προβλεπόμενα αποτελέσματα με ένα πρότυπο φάσμα και να υπολογίσει ένα νέο σκορ για κάθε υποψήφια δομή που επεξεργάζεται από το MetFrag. Η ροή εργασίας απεικονίζεται στην ακόλουθη εικόνα:



Εικόνα 11: Η ροή εργασίας του MetFusion, πηγή [29]

Το φάσμα που εισάγουμε προωθείται τόσο στο MassBank όσο και στο MetFrag. Και τα δυο εργαλεία επιστρέφουν λίστες υποψήφιων δομών. Αυτές οι λίστες συνδυάζονται και υπολογίζεται η χημική ομοιότητα μεταξύ όλων των δομών. Το προσαρμοσμένο σκορ χρησιμοποιείται για να επανακαταταγούν τα

αποτελέσματα της λίστας των υποψήφιων ενώσεων από το MetFrag από τα αποτελέσματα της βάσης δεδομένων (MassBank).

Τα σκορ του MassBank υπολογίζονται στη βάση μιας τροποποιημένης συνημιτονικής απόστασης που υπολογίζει την ομοιότητα μεταξύ του φάσματος που εισάγουμε και του φάσματος αναφοράς. Τα αποτελέσματα κατατάσσονται σύμφωνα με αυτή τη φασματική ομοιότητα. Το MassBank καθίσταται προσβάσιμο με τη χρήση μιας προγραμματιστικής πλατφόρμας βασισμένη σε βιβλιοθήκη Java, που ρωτά τους διακομιστές (servers) και εισάγει τους σχετικούς παραμέτρους (ένταση αποκοπής, λειτουργία ιοντισμού, φίλτρο οργάνου).

Η *in silico* θραυσματοποίηση πραγματοποιείται με έναν ενσωματωμένο μηχανισμό που αναζητεί συστατικά στις βάσεις δεδομένων (KEGG, PubChem, ChemSpider). Επιπλέον υπάρχει η δυνατότητα χρήσης *in-house* βάσεων δεδομένων. Αυτό καθίσταται εφικτό είτε με ένα αρχείο που υπάρχει η επιλογή ανεβάσματος ή από απευθείας πρόσβαση της βάσης. Αυτό επιτρέπει στους χρήστες να υποβάλουν τις δικές τους δομές ως υποψήφιες για *in silico* θραυσματοποίηση.

Το MetFusion απαιτεί ως είσοδο το φάσμα με τη μορφή «μάζα προς φορτίο-κενό-ένταση» καθώς και τις υπόλοιπες παραμέτρους, όπως ποια βάση δεδομένων συστατικών θα χρησιμοποιήσει (PubChem, Chempider, KEGG). Οι ρυθμίσεις του MetFrag μπορούν επίσης να προσαρμοστούν με πιο σημαντική τη ρύθμιση της ακρίβειας μάζας για τα θραύσματα που προκύπτουν.

2.5.2 Ενσωμάτωση της φασματικής ομοιότητας, των *in silico* βαθμολογιών και της χημικής ομοιότητας

Metfrag και MassBank επιστρέφουν δυο ανεξάρτητες λίστες φασματικής ομοιότητας και υποψήφια συστατικά με τις αντίστοιχες βαθμολογίες. Οι βαθμολογίες των φασμάτων συνδυάζονται σε έναν πίνακα φασματικής ομοιότητας. Πρόκειται για μια συνάθροιση παρόμοιων φασμάτων και η αντίστοιχη χημική ομοιότητα τους με το υποψήφιο συστατικό.

Η βιβλιοθήκη φασμάτων μπορεί να περιέχει πολλαπλές μετρήσεις ενός συστατικού ή των ισομερικών του μορφών, και για τον λόγο αυτό χρησιμοποιείται φιλτράρισμα βασισμένο στα InChiKEY της αρχικής λίστας των

αποτελεσμάτων από το MassBank, που περιέχει μόνο τις εγγραφές με το υψηλότερο σκορ για κάθε συστατικό. Αυτό δικαιολογείται επειδή η διάκριση μεταξύ στερεοϊσομερών είναι δύσκολα εφικτή μόνο με τη χρήση φασματομετρίας μαζών.

Η παρακάτω εξίσωση περιγράφει το προσαρμοσμένο MetFusion σκορ s_c , για κάθε υποψήφια δομή c .

$$s_c = \underbrace{\alpha * f_c}_{\text{MetFrag}} + \underbrace{(1 - \alpha) * \sum_{j=1}^M \text{sig}(m_j * t_{cj})}_{\text{"spectral summary"}}$$

Όπως φαίνεται από την εξίσωση για κάθε δομή υπολογίζεται το s_c ως το άθροισμα του σκορ από το MetFrag f_c και το φασματικό άθροισμα (spectral summary). Το φασματικό άθροισμα είναι το γινόμενο των σκορ m_j για όλα τα j αποτελέσματα από το MassBank επί τις αντίστοιχες χημικές ομοιότητες t_{cj} όπου c είναι οι υποψήφιες δομές. Χρησιμοποιείται η σιγμοειδής συνάρτηση για να εισάγει μια μη γραμμική συμπεριφορά που μειώνει την επιρροή από μέτριες ταυτίσεις φασμάτων και χημικές ομοιότητες.

Ο αριθμός των αποτελεσμάτων από το MetFrag συμβολίζεται ως N και ο αριθμός των αποτελεσμάτων από το MassBank συμβολίζεται ως M . Αυτό οδηγεί σε ένα πίνακα $N \times M$ χημικών ομοιοτήτων. Η χημική ομοιότητα t_{cj} μεταξύ του υποψήφιου c και κάθε αποτελέσματος j του MassBank μας επιτρέπει να καθορίσουμε πόσο όμοιο είναι κάθε ζεύγος συστατικών. Αυτή η προσέγγιση οδηγεί στο ενσωματωμένο σκορ που επιτρέπει να κατατάξουμε τις υποψήφιες ουσίες του MetFrag με επιπλέον επίπεδο πληροφορίας.

Το φασματικό άθροισμα για κάθε υποψήφια ένωση c είναι το άθροισμα όλων επιμέρους σκορ από το MassBank m_j ζυγισμένο με τη χημική ομοιότητα ως προς την υποψήφια ένωση t_{cj} . Το φασματικό σκορ από το MassBank m_j είναι από το 0 έως το 1, όπου τιμές μεγαλύτερες ίσες του 0,65 υποδηλώνουν εύλογη φασματική ομοιότητα.

Για τον υπολογισμό της χημικής ομοιότητας χρησιμοποιείται το Chemistry Development Kit (CDK, version 1.4.7). Η χημική ομοιότητα t_{cj} μεταξύ των

μοριακών αποτυπωμάτων του συστατικού c και j υπολογίζεται χρησιμοποιώντας τον συντελεστή Tamitoto (επίσης γνωστό και ως Jaccard).

Η ισορροπία μεταξύ των αποτελεσμάτων του MetFrag και MassBank καθορίζεται από το βάρος α , όπου αν $\alpha=1$ τότε χρησιμοποιείται αποκλειστικά το σκορ του MetFrag ενώ αν $\alpha=0$ τότε πραγματοποιείται απλά φασματική αναζήτηση ομοιότητας μεταξύ του φάσματος της αγνώστου ουσίας και των φασμάτων του MassBank. Παρόλου που το σκορ των MetFrag και MassBank ξεχωριστά είναι μεταξύ του εύρους 0-1, το σκορ του MetFusion δεν έχει άνω όριο και εξαρτάται από το αρχικό σκορ του MetFrag f_c , τον αριθμό των αναφορών στη βάση δεδομένων και την αντίστοιχη χημική ομοιότητα. Το κάτω όριο σκορ του MetFusion είναι μηδέν [29].

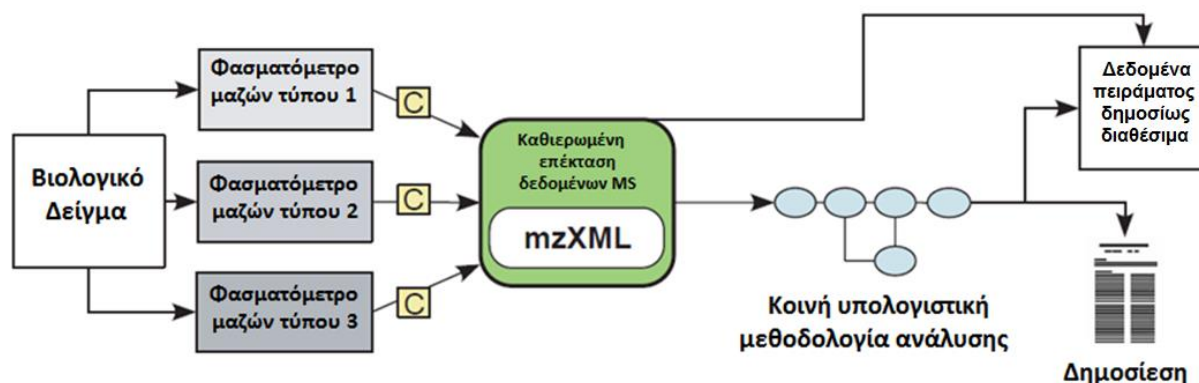
2.6 ProteoWizard Software®

2.6.1 Αναπαράσταση δεδομένων φασματομετρίας μαζών ως mzXML

Ο τύπος αρχείων mzXML αποτελεί μια ανοιχτή αναπαράσταση δεδομένων φασματομετρίας μαζών. Κάθε τύπος φασματομέτρου διαθέτει έναν μοναδικό σχεδιασμό, δηλαδή ένα μοναδικό σύστημα αποθήκευσης και επεξεργασίας δεδομένων και προδιαγραφές επίδοσης, τα οποία έχουν ως αποτέλεσμα τόσο αδυναμίες όσο και δυνατά σημεία. Δυστυχώς τα εγγενώς δυαδικά δεδομένα που παράγονται από κάθε τύπο φασματομέτρου διαφέρουν και συχνά αποτελούν ιδιοκτησία των κατασκευαστών. Ο ποικίλος, αδιαφανής χαρακτήρας της δομής των δεδομένων δυσχεραίνει την ενσωμάτωση των νέων οργάνων στην προϋπάρχουσα υποδομή, παρεμποδίζει την ανάλυση, ανταλλαγή, σύγκριση και δημοσίευση των αποτελεσμάτων από διαφορετικά πειράματα και εργαστήρια και αποτρέπει την κοινότητα της βιοπληροφορικής να έχει πρόσβαση στα δεδομένα που είναι απαραίτητα ώστε να αναπτυχθούν νέα λογισμικά. Επιπλέον είναι δυνατόν η διαφορετική τμηματική και συνδυαστική συναρμολόγηση των διαφορετικών αναλυτών μαζών να δημιουργεί ένα ευρύ φάσμα οργάνων, κάθε ένα με ιδιαίτερα πλεονεκτήματα και αδυναμίες για συγκεκριμένα είδη πειραμάτων. Έτσι είναι πολύ συνηθισμένο να βρεθούν δυο ή περισσότεροι διαφορετικού τύπου φασματομέτρα μάζας σε ένα εργαστήριο. Η έξοδος ενός φασματομέτρου υποβάλλεται σε ένα αριθμό αναλυτικών βημάτων, που συνήθως υποστηρίζονται από προγράμματα

ιδιωτικής ιδιοκτησίας των κατασκευαστών των οργάνων. Έτσι, σε ένα εργαστηριακό περιβάλλον όπου λειτουργούν διαφορετικοί τύποι φασματομέτρων και πολλαπλά προγράμματα ανάλυσης δεδομένων, είναι πολύ δύσκολο να γίνουν ουσιαστικές συγκρίσεις αποτελεσμάτων που λαμβάνονται από διαφορετικά πειράματα και διαφορετικά όργανα.

Για να αντιμετωπιστεί αυτό το πρόβλημα, αναπτύχθηκε ο τύπος mzXML, ο οποίος είναι ένας κοινός και ανοιχτής αναπαράστασης τύπος για φασματομετρικά δεδομένα MS, MS/MS ή MSⁿ. Ο τύπος mzXML προσδίδει μια καθολική διεπαφή δεδομένων μεταξύ φασματομέτρων μάζας και διόδων επεξεργασίας δεδομένων όπως φαίνεται στην παρακάτω εικόνα:



Εικόνα 12: Η μορφή mzXML επιτρέπει κοινή αναπαράσταση δεδομένων, εφαρμογή κοινών μεθόδων επεξεργασίας και εξασφαλίζει δημόσια διαθεσιμότητα μαζί με τη δημοσίευση των αποτελεσμάτων, πηγή [30]

Αυτή η προσέγγιση εξαλείφει την ανάγκη για πολλαπλές μορφές εισόδου και απλοποιεί σημαντικά την ενσωμάτωση νέων οργάνων στην αναλυτική εργασία. Επιπλέον ο νέος τύπος δεδομένων διευκολύνει την ανταλλαγή και τη δημοσίευση φασματομετρικών δεδομένων και παράσχει μια συνεπή πλατφόρμα για την ανάπτυξη νέων αναλυτικών εργαλείων.

Μια κοινή αναπαράσταση φασματικών δεδομένων πρέπει να παρέχει ισορροπία μεταξύ ευελιξίας και σταθερότητας. Ο τύπος ANDI/netCDF [31] είναι η πιο επιτυχημένη προσπάθεια για τη δημιουργία μιας ουδέτερης μορφής δεδομένων φασματομετρίας. Δυστυχώς, εξαιτίας των πολύπλοκων και άκαμπτων βιβλιοθηκών του τύπου netCDF, ο τύπος είναι δύσκολο να κρατηθεί ενημερωμένος. Ως αποτέλεσμα, πλέον δεν είναι σε θέση να αποθηκεύσει

MS/MS φάσματα, που πολλές φορές αποτελούν την ουσία πολλών πειραμάτων ανίχνευσης αναλυτών.

Προκειμένου να παράσχει έναν τρόπο να ενσωματώσει τους νέους τύπους δεδομένων, ιδιαίτερη προσοχή πρέπει να δοθεί στην ευελιξία του mzXML. Είναι σχεδιασμένο σε γλώσσα XML (Extensible Markup Language), η οποία εξ' ορισμού είναι μια επεκτάσιμη γλώσσα και παρέχει έναν τρόπο να καθορίσει προαιρετικό ή προαπαιτούμενο το περιεχόμενο του. Αν και είναι επιθυμητό να διατηρηθεί ένα ορισμένο επίπεδο ευελιξίας, πάρα πολλή ελευθερία μπορεί να οδηγήσει στο σχηματισμό “διαλέκτων” (όπως έχει συμβεί με την έκφραση γονιδίων σε μικροσυστοιχία (MAGE) suite2). Αν και δύσκολο να εμποδιστεί εξ ολοκλήρου, οι διάλεκτοι μειώνουν τη χρησιμότητα ενός κοινού τύπου δεδομένων και πρέπει να μειωθεί. Είναι επομένως σημαντικό οι διαφορετικές ομάδες εργασίας να επικοινωνούν και εργάζονται υπό την εποπτεία μιας συμβουλευτικής επιτροπής. Για το mzXML αυτό έχει να αντιμετωπιστεί από το να παρέχεται στους χρήστες της ένα άμεσα προσβάσιμο και δημόσιο φόρουμ που βρίσκεται στη διεύθυνση http://sourceforge.net/forum/forum.php?forum_id=235607, για να ανταλλάξουν απόψεις και να παράσχουν πληροφορίες σχετικά με τη μορφή και το σχηματισμό της επιτροπής MASS (mzXML-Associated Standard Solution).

Η επιτροπή MASS είναι υπεύθυνη για την προσαρμογή του mzXML ώστε να αντιμετωπισθούν οι ανάγκες της επιστημονικής κοινότητας και δεύτερον να βεβαιώσει ότι ο τύπος mzXML παραμένει ενημερωμένος σχετικά με τις καινοτομίες στην φασματομετρία. Τρίτον είναι υπεύθυνη για τον συντονισμό των προσπαθειών των ομάδων να δουλέψουν σε διαφορετικές εφαρμογές του τύπου mzXML. Η επιτροπή επίσης αλληλοεπιδρά και μοιράζεται πόρους με άλλες ομάδες που δουλεύουν στην τυποποίηση των δεδομένων κυρίως της πρωτεομικής [30].

2.6.2 Προβλήματα της κωδικοποίησης mzXML και η αντιμετώπισή τους

Παρόλου που αρχικά υπήρχαν κάποιοι περιορισμοί του τύπου δεδομένων mzXML σε σύγκριση με τους δυαδικούς τύπος αρχείων, τελικά ξεπεράστηκαν. Ιδιαίτερης σημασίας μειονεκτήματα ήταν η σημαντική αύξηση του μεγέθους του αρχείου και η μείωση της ταχύτητας πρόσβασης της πληροφορίας που

παρουσιάζεται ως XML. Παρακάτω θα περιγραφούν συνοπτικά πως αντιμετωπίστηκαν οι περιορισμοί αυτοί.

Τα σύγχρονα φασματομέτρα μάζας μπορούν να δημιουργήσουν πάνω από 1GB συμπιεσμένων δυαδικών αρχείων την ώρα. Παρόλα αυτά, το XML δεν μπορεί να ενσωματώσει άμεσα δυαδικά δεδομένα και για αυτό η μετατροπή σε μια πιο πολύ αναγνώσιμη από τον άνθρωπο μορφή δεν είναι εφικτή χωρίς μια σημαντική αύξηση στο μέγεθος. Αυτό το πρόβλημα αντιμετωπίζεται με τη μορφή mzXML και την κωδικοποίηση της έντασης και των m/z σε δυαδικά ζεύγη σε κωδικοποίηση base64 (rfc1341: <http://www.faqs.org/rfcs/rfc1341.html>). Τα Base64-κωδικοποιημένα δυαδικά δεδομένα είναι κάπως μεγαλύτερα (1,3 φορές) από μια δυαδική παρουσίαση. Παρόλα αυτά, σε αντίθεση με τα δυαδικά δεδομένα, μπορούν να ενσωματωθούν σε ένα και μόνο αρχείο mzXML. Επιπλέον, είναι επίσης εφικτό να μειωθεί το μέγεθος τους σε υψηλό βαθμό διώχνοντας τις κορυφές με μηδενική ένταση. Με αυτές τις απλές στρατηγικές, πληροφορίες από αρχεία μεγάλης δυαδικής μήτρας μπορούν να αποθηκευτούν υπο μορφή κειμένου σε διαχειρίσιμο μέγεθος.

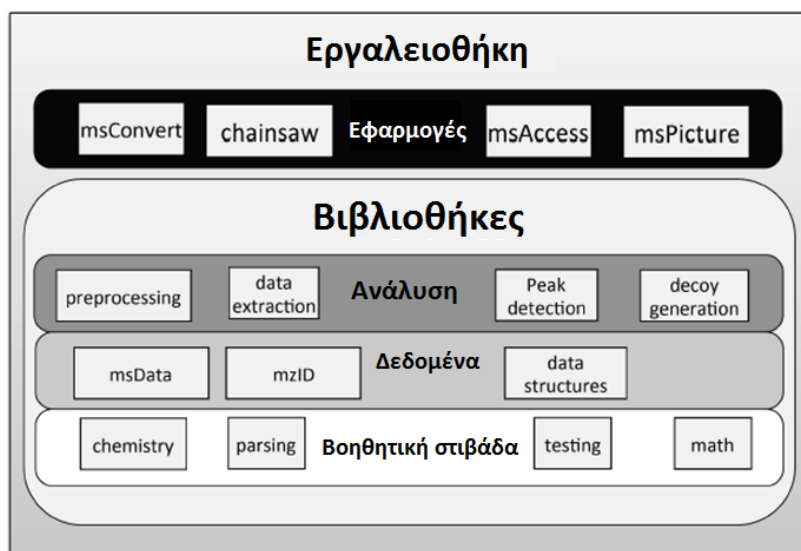
Ο δεύτερος περιορισμός του mzXML ως λειτουργική μορφή για την αναπαράσταση των δεδομένων φασματομετρίας μάζας είναι συνέπεια κάποιων μονάδων XML (π.χ. SAX: www.saxproject.org) που διαβάζει το έγγραφο διαδοχικά από την αρχή του αρχείου ως το τέλος. Εφαρμογές που απαιτούν μη διαδοχική πρόσβαση στα δεδομένα, όπως η στρατηγική ποσοτικοποίησης πεπτιδίων βασισμένη σε σταθερά επισημασμένα υποστρώματα (π.χ. ASAPRatio3), θα είχε απαράδεκτες επιδόσεις χρησιμοποιώντας καθαρά διαδοχική πρόσβαση στα δεδομένα. Αυτό το πρόβλημα αποτράπηκε με τη δημιουργία ενός σχήματος που επιτρέπει στα προγράμματα να βρίσκουν στα δεδομένα mzXML, τη θέση για κάθε πλήρη σάρωση [30].

2.6.3 Δομή ProteoWizard και Msdata εφαρμογής και msconvert.exe

Μετά τη δημιουργία του τύπου mzXML δημιουργήθηκε η πλατφόρμα ProteoWizard Toolkit για χάρη της πρωτομικής έρευνας που μεταξύ των άλλων λειτουργιών της ενσωματώνει έναν απλό και ευέλικτο μετατροπέα δεδομένων οποιαδήποτε εταιρείας κατασκευής φασματομέτρων σε mzXML

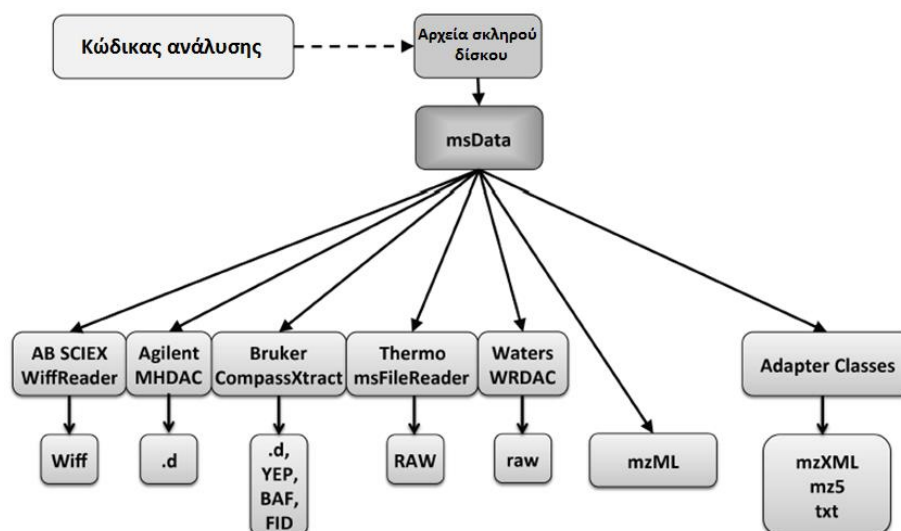
μορφή ή και άλλες υποστηριζόμενες μορφές. Η εργαλειοθήκη έχει δυο συνιστώσες:

1. μια σουίτα βιβλιοθηκών (libraries) που διευκολύνουν την ανάπτυξη και τη σύγκριση των εργαλείων για την ανάλυση των δεδομένων
2. ένα σετ εργαλείων (apps) που αναπτύχθηκε με τη χρήση αυτών των βιβλιοθηκών, για να εκτελέσει ευρύ φάσμα των πιο κοινών πρωτεομικών αναλύσεων.



Εικόνα 13: Δομή του προγράμματος ProteoWizard, πηγή [32]

Το πρόγραμμα ProteoWizard είναι χτισμένο σε ένα αρθρωτό πλαίσιο πολλών ανεξάρτητων βιβλιοθηκών που ομαδοποιούνται σε επίπεδα της ιεραρχίας. Το στρώμα των δεδομένων παρέχει μια ενιαία διεπαφή πρόσβασης στα δεδομένα, ανεξαρτήτως από τη μορφή που σχετίζεται με μια δεδομένη πηγή. Το βασικό μοντέλο δεδομένων του στρώματος μεταφράζεται σε HUPO-PSI δομές δεδομένων C++. Για παράδειγμα παρακάτω αναπαρίσταται γραφικά η μονάδα **msData**:



Εικόνα 14: Γραφική αναπαράσταση μονάδας msData του προγράμματος ProteoWizard, πηγή [32]

Σε συνεργασία με τους οργανισμούς προτυποποίησης και τους κατασκευαστές οργάνων και λογισμικών το λογισμικό περιέχει προσαρμογείς (adapters) που καθιστούν εφικτή την υποστήριξη ενός μεγάλου πεδίου τύπων αρχείων κατασκευαστών οργάνων. Αυτοί οι προσαρμογείς γεφυρώνουν τους τύπους αρχείων μεταξύ των βιβλιοθηκών που παρέχουν οι κατασκευαστές που διαβάζουν μορφές των κατασκευαστών και ανοιχτά δεδομένα. Μέσα από μια σειρά από γενναιόδωρες άδειες, το ίδρυμα ελεύθερου λογισμικού ProteoWizard έχει την άδεια να διανέμει βιβλιοθήκες που κατασκευάζονται από τους μεγαλύτερους κατασκευαστές φασματομέτρων όπως οι AB SCIEX, Agilent, Bruker, Thermo Fischer Scientific και Waters. Κατά συνέπεια, οι προγραμματιστές της βιοπληροφορικής δεν απαιτείται να έχουν άμεση πρόσβαση σε ένα όργανο για να βελτιώσουν το λογισμικό με το οποίο αναλύουν τα δεδομένα.

Κάτω από το στρώμα δεδομένων είναι η «βοηθητική» στιβάδα (utility layer), η οποία περιέχει εφαρμογές που εκτελούν υπολογισμούς όπως η δυαδική μετατροπή κειμένου σε κωδικοποίηση XML και μαθηματικούς υπολογισμούς που είναι κοινοί σε αρκετές αναλύσεις δεδομένων. Αν και η πλειονότητα των υπολογισμών που διατίθεται σε αυτές τις κατηγορίες είναι απλή η εφαρμογή μπορεί να είναι χρονοβόρα. Χρησιμοποιώντας το λογισμικό ProteoWizard, οι προγραμματιστές μπορούν να αναπτύξουν καινοτόμους αλγορίθμους αντί για

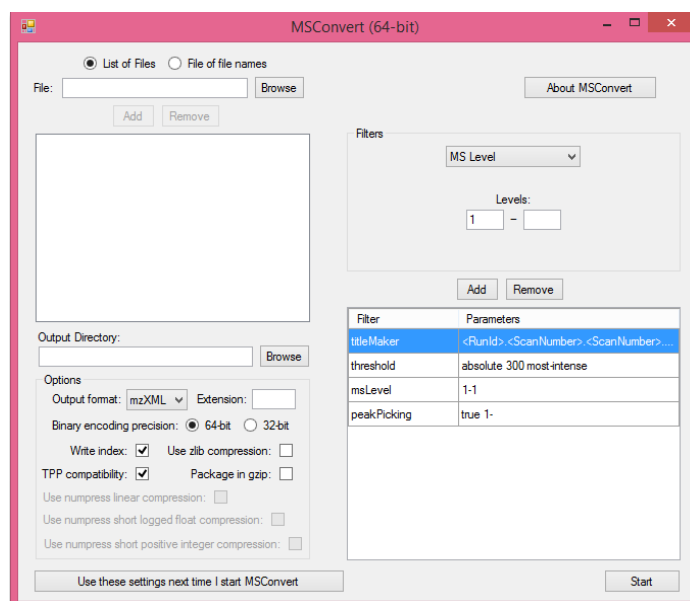
εφαρμογή προαπαιτούμενου χειρισμού των δεδομένων, επιταχύνοντας έτσι την ανάπτυξη νέων λογισμικών.

Η στιβάδα ανάλυσης (analysis layer) βρίσκεται επάνω από την στιβάδα δεδομένων (data layer) και παρέχει βασικές μονάδες ανάλυσης. Ένα σημαντικό εμπόδιο στην ανάπτυξη λογισμικών πρωτομικής μπορεί να προκύψουν από το χρόνο που απαιτείται για την εφαρμογή μεγάλης ποικιλίας εργασιών ρουτίνας για την εφαρμογή ενός αλγορίθμου όπως τον υπολογισμό της μάζας ενός πεπτιδίου ή την εκτέλεση *in silico* πέψης μιας πρωτεΐνης. Υπάρχουν επίσης ανεξάρτητες ενότητες για τον χειρισμό χημικών τύπων, υπολογισμούς πεπτιδίων και «φακέλους» ισοτόπων. Όλοι αυτοί οι υπολογισμοί περιέχονται σε επαναχρησιμοποιήσιμες, ανεξάρτητες μονάδες (modules) στη στιβάδα ανάλυσης.

Πρόσθετες μονάδες ανάλυσης (analysis modules) είναι αυτή τη στιγμή σε εξέλιξη με έμφαση στη θέσπιση τυποποιημένων διεπαφών για κοινούς υπολογισμούς. Ο στόχος είναι να δημιουργηθούν από κοινού δομές ανάλυσης στο οποίο οι ειδικοί θα μπορούν να συνεισφέρουν ένα τμήμα (module) το οποίο μετά θα μπορεί να ενσωματωθεί σε διάφορα εργαλεία. Αυτό θα επιτρέψει για παράδειγμα ένας εμπειρογνώμονας στην επεξεργασία σημάτων να συνεισφέρει έναν αλγόριθμο αναζήτησης κορυφών χωρίς να πρέπει να αντιμετωπίσει τις λεπτομέρειες στα διαφορετικά είδη μορφών αρχείων, λειτουργικών συστημάτων ή ρυθμίσεων από την γραμμή εντολών (command-line configuration).

Ο ProteoWizard περιλαμβάνει έναν μικρό αριθμό από χρήσιμες εφαρμογές, οι οποίες έχουν χτιστεί πάνω στις ήδη υπάρχουν βιβλιοθήκες. Οι πιο σημαντικές εφαρμογές υποστηρίζουν τη μετατροπή δεδομένων (msConvert, msConvertGUI, idConvert), την οπτικοποίηση των δεδομένων-data visualization (msPicture, seems), την πρόσβαση σε δεδομένα (msAccess, msCat, idCat, msPicture) , την ανάλυση δεδομένων (peekaboo, msPrefix) και άλλες βασικές πρωτομικές λειτουργίες [32].

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε η εφαρμογή μετατροπής δεδομένων msConvert προκειμένου να μετατραπούν τα .d Bruker δεδομένα σε mzXML.

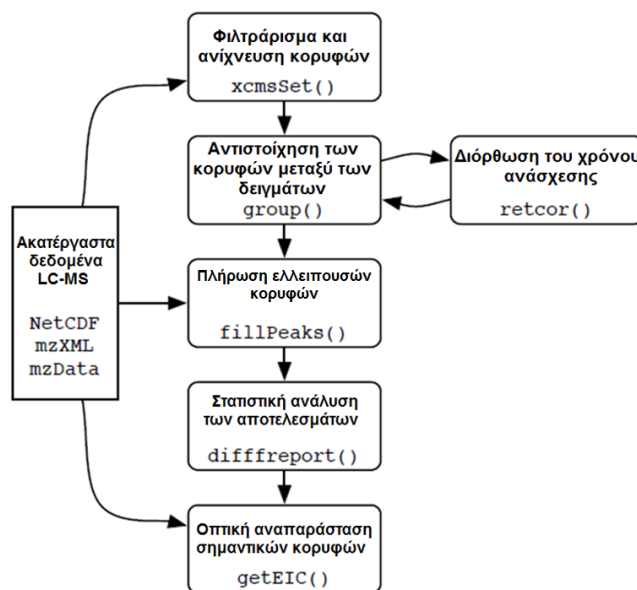


Εικόνα 15: Στιγμιότυπο γραφικού περιβάλλοντος msConvert (ProteoWizard App)

Τέθηκαν φίλτρα αποκοπής έντασης 300, που σημαίνει ότι σήματα με ένταση μικρότερη από 300 αφαιρέθηκαν. Με τον όρο msLevel 1-1 εννοείται ότι μετατρέπεται μόνο το φάσμα MS και όχι τα φάσματα MS/MS ενώ με τη ρύθμιση peakPicking true -1 υποδηλώνεται ότι τα δεδομένα θα μετατραπούν σε μορφή γραμμών (line mode) και όχι υπο τη μορφή κορυφών (profile mode).

2.7 xcms

Το xcms δημιουργήθηκε ως υπολογιστικό εργαλείο με στόχο να βοηθήσει τη μεταβολομική έρευνα να βρει μοναδικούς μεταβολίτες που χαρακτηρίζονται ως βιοδείκτες, η παρουσία των οποίων μαρτυρά κάποια παθολογική κατάσταση και να βοηθήσει στην εύρεση μορίων που πρέπει να στοχευθούν ώστε να καταπολεμηθούν αυτές οι παθολογικές καταστάσεις. Αποτελεί μια τεχνική επεξεργασίας δεδομένων που προέρχονται από τεχνικές υψηλής απόδοσης (high throughput analysis) όπως είναι η υγροχρωματογραφία συζευγμένη με φασματομετρία μαζών υψηλής διακριτικής ικανότητας [33]. Τα δεδομένα (netCDF, mzXML ή mzData) αποθηκεύονται σε μια συγκεκριμένη διεύθυνση του σκληρού δίσκου σε φακέλους που υπονοούν την κατηγοριοποίηση τους και δεν μετακινούνται μέχρι να ολοκληρωθεί η ανάλυση τους καθώς πολλές φορές οι κώδικες αντλούν δεδομένα από αυτά. Το ακόλουθο σχήμα αναπαριστά τα συνήθη βήματα που πραγματοποιούνται και τις συναρτήσεις-εντολές που χρησιμοποιούνται στην R [34] :



Εικόνα 16: Πορεία εργασίας του Xcms και οι βασικές εντολές για κάθε βήμα, πηγή [34]

Όπως φαίνεται από την προηγούμενη εικόνα η πορεία για μια LC-MS βασισμένη έρευνα μπορεί να χωρισθεί σε διακριτά βήματα:

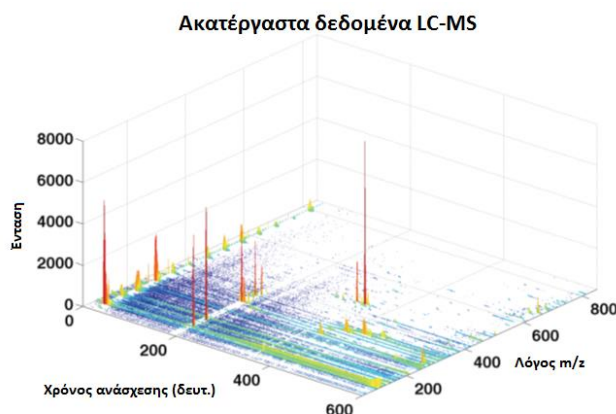
1. Προεπεξεργασία σήματος και μετατροπή των φασματικών κορυφών υπο μορφή κεντροειδούς (centroid mode)
2. Ανίχνευση και ολοκλήρωση δυο διαστάσεων των ιόντων (features)
3. Ευθυγράμμιση και αντιστοίχιση των ίδιων ιόντων σε πολλαπλά δείγματα
4. Στατιστική ανάλυση ή τοποθέτηση των ιόντων κατά αύξουσα προτεραιότητα, χημική και βιολογική ερμηνεία [35]

Αξίζει να σημειωθεί ότι με τη συνήθη πορεία δεν μπορούν να εξετάζονται αρχεία που προήλθαν από διαφορετικές συνθήκες έκλουσης και από διαφορετικό ιοντισμό, δηλαδή αρχεία που προέρχονται από διαφορετική μέθοδο οργάνου [34].

2.7.1 Εύρεση κορυφών (Peak Picking)

2.7.1.1 Εισαγωγή

Το πακέτο xcms περιέχει την εντολή `findPeaks.peak_picking_method` όπου ως παράμετρος `peak_picking_method` είναι μια τεχνική ανεύρεσης κορυφών. Έχουν προταθεί διάφορες τεχνικές εντοπισμού κορυφών. Το xcms υποστηρίζει μέχρι στιγμής 3 αλγόριθμους εύρεσης κορυφών: `matchedfilter`, `centWave` και `Massifquant`, οι οποίοι δημιουργήθηκαν με την σειρά που αναγράφονται. Τα δεδομένα που λαμβάνονται από την τεχνική LC-MS είναι τριών διαστάσεων.



Εικόνα 17: Διαστάσεις λαμβανόμενων δεδομένων LC-MS

Η μια διάσταση είναι ο χρόνος ανάσχεσης, στον οποίο εκλούεται ο αναλύτης από την υγροχρωματογραφία. Η δεύτερη διάσταση είναι η φασματομετρική διάσταση, δηλαδή οι τιμές m/z . Ο λόγος m/z αντανακλά το μοριακό βάρος (αν το φορτίο είναι ίσο με ένα). Η τρίτη διάσταση είναι η ένταση των κορυφών που αντανακλά την αφθονία των αναλυτών και παρέχει την ποσοτικοποίηση τους [34].

Το πρώτο στάδιο είναι η ανίχνευση των ιόντων. Ανίχνευση των ιόντων καλείται ο καθορισμός των ορίων και του μέσου όρου των χρόνων ανάσχεσης και ο καθορισμός των εντάσεων ή του εμβαδού των σημάτων στα πρωτογενή δεδομένα. Για την ανάλυση περίπλοκων δειγμάτων απαιτείται ένας αξιόπιστος τρόπος ανίχνευσης.

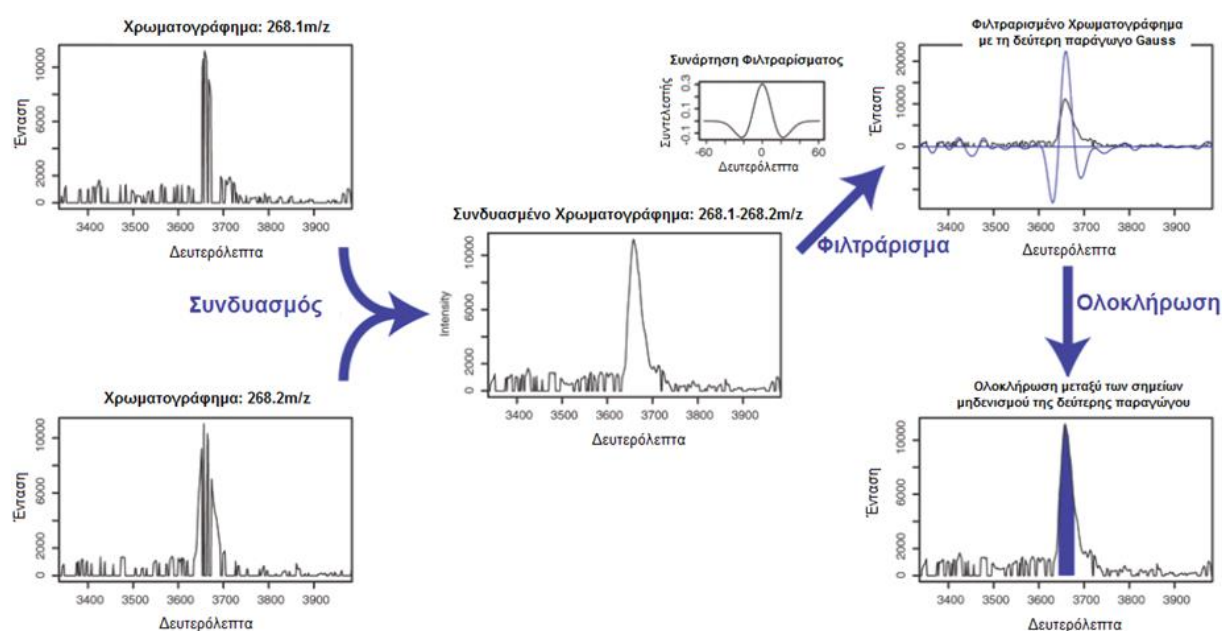
Η ανίχνευση ιόντων είναι καθοριστικής σημασίας βήμα στην πορεία επεξεργασίας και για αυτό πρέπει να είναι όσο το δυνατόν πιο αξιόπιστος. Με τον όρο αξιόπιστο εννοούμε να αποκαλύπτει όσο το δυνατόν περισσότερα «πραγματικά» ιόντα ενώ να ανιχνεύει τα λανθασμένα θετικά ιόντα σε όσο το δυνατό μικρότερο ποσοστό. Η πρόκληση για αυτούς τους αλγόριθμους ανίχνευσης ιόντων είναι η ανίχνευση χαμηλών κορυφών που εισάγονται από συστατικά που βρίσκονται σε χαμηλή αφθονία και η αποφυγή εισαγωγής κορυφών που προέρχονται από χημικό θόρυβο [35].

2.7.1.2 Matchedfilter

Ο αλγόριθμος matchedfilter βασίζεται στο κόψιμο των δεδομένων της φασματομετρικής διάστασης σε κομμάτια (slices) ανάλογα με τη διακριτική

ικανότητα του οργάνου για παράδειγμα ανά 0,1 m/z για χαμηλής διακριτικής ικανότητας όργανα (Ion-Trap) και 0,005 m/z για υψηλής διακριτικής ικανότητας όργανα (QTOF) και μετά σε κάθε ένα από τα ξεχωριστά κομμάτια πραγματοποιείται ανίχνευση χρωματογραφικών κορυφών στη χρωματογραφική διάσταση.

Ο αλγόριθμος λειτουργεί τόσο σε δεδομένα που λήφθηκαν υπο μορφή προφίλ (profile mode) όσο και με δεδομένα σε κεντροειδές (centroid mode) και μάλιστα αρχικά θεωρήθηκε κατάλληλος τόσο για υψηλής διακριτικής ικανότητας φασματομέτρα όσο και χαμηλής. Στην περίπτωση δεδομένων υψηλής διακριτικής ικανότητας ή δεδομένων τα οποία έχουν μετατραπεί σε κεντροειδή δεδομένα, δεδομένα δηλαδή που έχουν εύρος φασματικών κορυφών θεωρητικά απείρως λεπτό, τότε το σήμα ίσως διασπάται μεταξύ δυο γειτονικών τμημάτων και έχει ως αποτέλεσμα η κορυφή να είναι οδοντωτή (jagged) όπως φαίνεται στην ακόλουθη εικόνα:



Εικόνα 18: Γραφική πορεία αλγορίθμου matchedfilter, πηγή [33]

Ο αλγόριθμος δίνει τη δυνατότητα ανίχνευσης τέτοιων περιπτώσεων διότι μπορεί και συνδυάζει τα σήματα γειτονικών τμημάτων και έτσι δημιουργεί το συνδυασμένο χρωματογράφημα, που είναι εξομαλυμένο (smoothed) και κατάλληλο για φιλτράρισμα που είναι και το επόμενο βήμα. Πριν από την ανίχνευση κορυφών κάθε κομμάτι φιλτράρεται χρησιμοποιώντας ως μοντέλο για το σχήμα της κορυφής τη δεύτερη παράγωγο της καμπύλης Gauss. Μετά

το φιλτράρισμα, οι κορυφές επιλέγονται χρησιμοποιώντας τον λόγο σήμα προς θόρυβο (S/N) που έχουμε ορίσει ως φίλτρο αποκοπής και απορρίπτονται οι κορυφές με χαμηλό λόγο σήμα προς θόρυβο. Επειδή η δεύτερη παράγωγος της κανονικής κατανομής δημιουργεί ένα αρνητικό τμήμα στο σήμα, ο καθορισμός του θορύβου από φιλτραρισμένα δεδομένα είναι προβληματικός. Μετά από αρκετές δοκιμές οι συγγραφείς του αλγορίθμου καθόρισαν ότι ο μέσος όρος των μη φιλτραρισμένων δεδομένων είναι η πιο αποτελεσματική μέθοδος εύρεσης του θορύβου και ότι ο ορισμός σήματος προς θόρυβο ίσος με 10 ήταν ο πιο αποτελεσματικός. Τα σημεία μηδενισμού της δεύτερης παραγώγου του φιλτραρισμένου χρωματογραφήματος αποτελούν έναν πολύ χρήσιμο μηχανισμό καθορισμού του εύρους της κορυφής και του χρόνου ανάσχεσης της κορυφής.

Αξίζει να σημειωθεί ότι πριν από το φιλτράρισμα δεν έχει γίνει αφαίρεση του υποβάθρου (background subtraction) και έτσι κάποιο σήμα που προέρχεται από το υπόβαθρο περιλαμβάνεται στην ολοκλήρωση. Ωστόσο η αύξηση της γραμμής βάσης με τον χρόνο είναι αρκετές φορές αμελητέα στα δεδομένα LC-MS. Η παραγωγή τμημάτων (slices) μειώνει την ακρίβεια στην ανίχνευση της μάζας. Για να ξεπεραστεί αυτό η μάζα κάθε κορυφής υπολογίζεται από τα αρχικά φάσματα υψηλής διακριτικής ικανότητας.

Η κύρια παράμετρος που επηρεάζει την εύρεση κορυφών με τον matchedfilter αλγόριθμο είναι το πλάτος της κορυφής του μοντέλου Gauss που χρησιμοποιείται κατά το στάδιο του φιλτραρίσματος. Ανάλογα με τον τύπο της χρωματογραφίας το εύρος κορυφής στο 50% του ύψος της (Full Width at Half Maximum) μπορεί να διαφέρει. Τυπικά, εξ αρχής ο αλγόριθμος θέτει τιμή ίση με 30 δευτερόλεπτα-τιμή τυπική για HPLC. Η δεύτερη παράμετρος που επηρεάζει τον αλγόριθμο είναι η διακριτική ικανότητα του φασματομέτρου μάζας. Αυτή η παράμετρος ορίζεται από το βήμα (step) η τιμή του οποίου από προεπιλογή ορίζεται 0,1 m/z.

Ένας άλλος παράγοντας που πρέπει να ληφθεί υπόψη είναι ο τρόπος που παράγονται τα χρωματογραφήματα (EICs). Αξίζει να σημειωθεί ότι προκειμένου να ληφθούν υπόψη η αβεβαιότητα στην ακρίβεια μάζας από φάσμα πλήρους σάρωσης σε φάσμα πλήρους σάρωσης (from scan to scan) ο αλγόριθμος συνδυάζει έναν δεδομένο αριθμό κομματιών πριν από το βήμα του

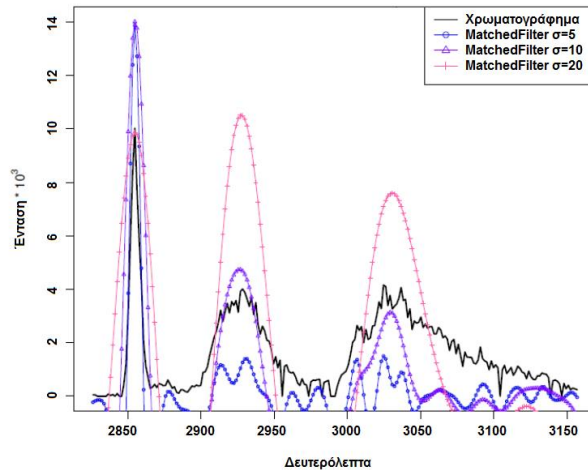
φιλτραρίσματος και την ανίχνευση κορυφών. Εξ αρχής ο αλγόριθμος συνδυάζει 2 κομμάτια 1-2,2-3,3-4 κτλ. Αν η ακρίβεια από φάσμα πλήρους σάρωσης σε φάσμα πλήρους σάρωσης είναι χειρότερη ίσως ο χρήστης θελήσει να αυξήσει τον αριθμό των συνδυασμένων φασμάτων για παράδειγμα σε 3 που θα συνδυάζε τα χρωματογραφήματα 1-2-3,2-3-4,3-4-5 κτλ. [33, 34, 36]

2.7.1.3 Μειονεκτήματα αλγορίθμων εύρεσης κορυφών που βασίζονται στην τεχνική binning

Μια ευρέως διαδεδομένη προσέγγιση για επεξεργασία των LC-MS δεδομένων είναι η μετατροπή των πρωτογενών δεδομένων σε μια μήτρα με διαστάσεις m/z , χρόνο ανάσχεσης και έντασης. Για να μετατραπούν τα φάσματα υψηλής διακριτικής ικανότητας σε αυτή τη μήτρα είναι αναγκαίο να διαιρεθεί ο άξονας m/z σε ισομεγέθη τμήματα ανάλογα με τη διακριτική ικανότητα του φασματογράφου μαζών (π.χ. εύρους 0,1 m/z). Αυτή η διαδικασία συνήθως αναφέρεται και ως τμηματοποίηση (binning).

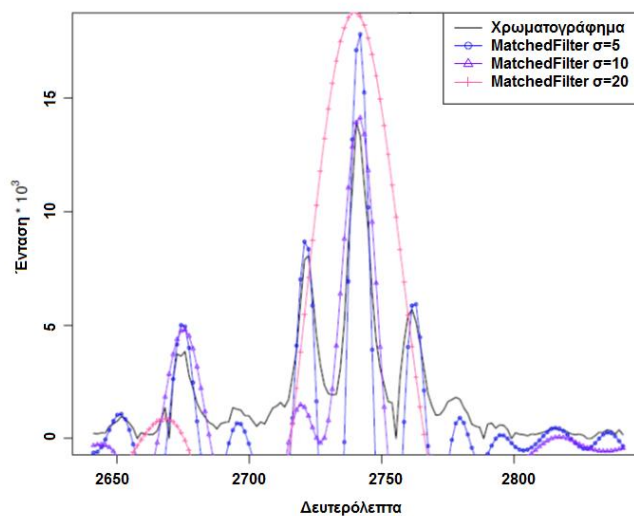
Η μεθοδολογία αυτή για εύρεση κορυφών παρουσιάζει σαφή μειονεκτήματα. Πιο συγκεκριμένα, ο ορισμός βέλτιστου ισομεγέθους τμήματος (bin) για ένα συγκεκριμένο σετ δεδομένων μπορεί να είναι δύσκολο εγχείρημα. Αν το μέγεθος επιλεχθεί πολύ μικρό, οι χρωματογραφικές κορυφές παρουσιάζονται με μεγάλες διακυμάνσεις μεταξύ των κομματιών (φαινόμενο υπερχείλισης) και δεν μπορούν να ανιχνευθούν εξαιτίας απώλειας χρωματογραφικού σήματος. Αν επιλεχθεί πολύ μεγάλο, οι κορυφές επικαλύπτονται και τα μικρά ιόντα χάνονται λόγω της αύξησης του χρωματογραφικού θορύβου.

Επιπλέον ανάλογα με το χρωματογραφικό σύστημα μπορεί να υπάρχουν διαφορετικά εύρη και σχήματα κορυφών. Έτσι για παράδειγμα η matchedfilter τεχνική που ανήκει στις τεχνικές τμηματοποίησης της διάστασης της μάζας (binning τεχνικές) αποτυγχάνει να ανιχνεύσει όλες τις κορυφές διότι χρησιμοποιεί συγκεκριμένο μοντέλο κορυφής (γκουσιανή εν προκειμένω) και έχει συγκεκριμένο εύρος κορυφής. Στην ακόλουθη εικόνα περιέχονται τρεις κορυφές με διαφορετικό εύρος. Η εφαρμογή τριών διαφορετικών ευρών ως μοντέλο στον αλγόριθμο matchedfilter αποκαλύπτει ότι το μοντέλο με $\sigma=5-10$ s ανιχνεύει τις στενές κορυφές. Μόνο για $\sigma=20$ s ανιχνεύονται όλες οι κορυφές.



Εικόνα 19: Αποτελέσματα αλγορίθμου matchedfilter χρησιμοποιώντας τη δεύτερη παράγωγο του Gauss με διαφορετικά εύρη φίλτρου για κορυφές με διαφορετικά εύρη, πηγή [35]

Μια άλλη οπτική του προβλήματος αναδεικνύεται με κορυφές που εκκλύονται πολύ κοντά. Η ακόλουθη εικόνα αναπαριστά την απόκριση των διαφορετικών παραμέτρων σ στον αλγόριθμο matchedfilter. Μόνο για $\sigma=5s$ το μοντέλο δίνει ικανοποιητικά αποτελέσματα. Επειδή η ακόλουθη και η προηγούμενη εικόνα είναι αποτέλεσμα ενός χρωματογραφήματος προκύπτει ότι όποιες παράμετροι και αν επιλεγούν, καμία δεν θα δώσει ικανοποιητικά αποτελέσματα για όλες τις περιπτώσεις [35].



Εικόνα 20: Αποτελέσματα αλγορίθμου matchedfilter χρησιμοποιώντας τη δεύτερη παράγωγο του Gauss με διαφορετικά εύρη φίλτρου για κοντινά εκκλούόμενες κορυφές, πηγή [35]

Μια προσέγγιση βασισμένη στην πυκνότητα (density based) για ανίχνευση ιόντων αποτελεί μια εναλλακτική τεχνική του διαμερισμού της διάστασης των μαζών και πρωτοειστάχθηκε από τον Stolt et al. Οι συγγραφείς θεωρούν τον αναλύτη ως μια περιοχή με σημεία με υψηλή πυκνότητα. Με βάση αυτές τις ιδιότητες, υπολογίζεται ένα δυνητικό πεδίο, το οποίο στη συνέχεια χρησιμοποιείται για να δημιουργηθεί μια μήτρα από μάζες (χρόνος επεξεργασίας υψηλός: 2ώρες/δείγμα) [37].

2.7.1.4 centWave

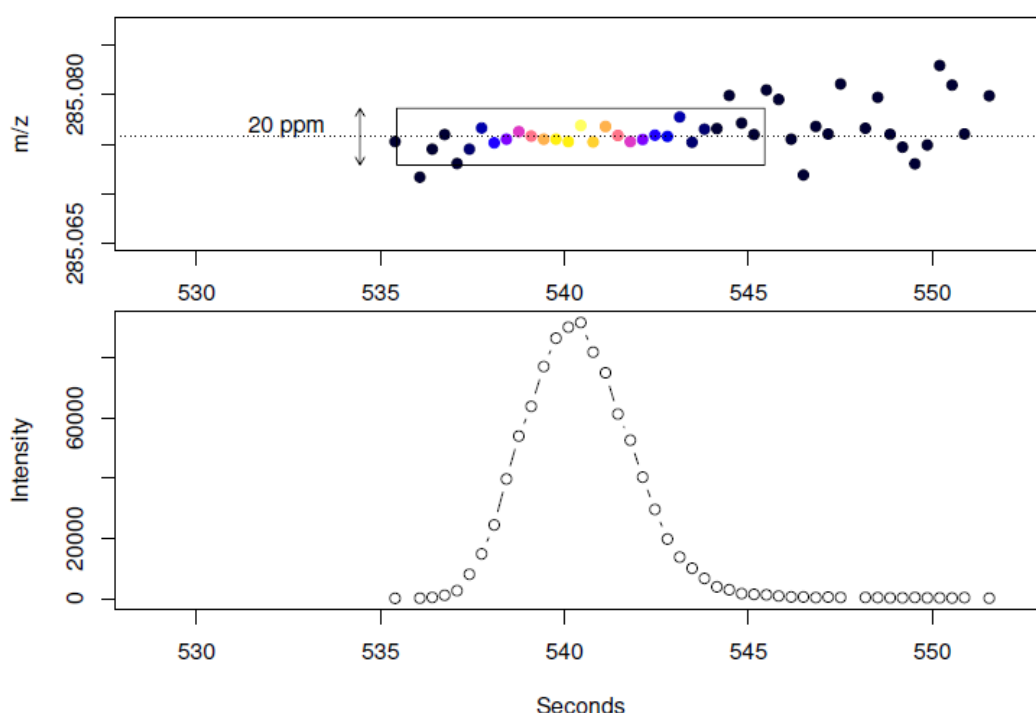
Μια μεταγενέστερη μέθοδος της matchedfilter που ήρθε να αντιμετωπίσει τα προβλήματα της διαμερισματοποίησης της διάστασης των m/z και να οδηγήσει σε μια αξιόπιστη ανίχνευση κορυφών είναι ο αλγόριθμος centWave. Χρησιμοποιώντας έναν συνδυασμό της τεχνικής που βασίζεται στην πυκνότητα για να ανιχνεύσουμε της περιοχές ενδιαφέροντος (ROIs) και την προσέγγιση κυματιδίων (wavelets) για να βρει χρωματογραφικές κορυφές επιτυγχάνει υψηλή ευαισθησία σε σύγκριση με δυο άλλους αλγόριθμους (matchedFilter-XCMS, centroidPicker-MZmine).

Ο αλγόριθμος centWave ακολουθεί διαφορετική προσέγγιση και είναι κατάλληλος για υψηλής διακριτικής ανάλυσης δεδομένα, δηλαδή LC/(QTOF, Orbitrap, FTICR), τα οποία όμως βρίσκονται σε μορφή κεντροειδούς (centroid mode). Επιπλέον τα δεδομένα που βρίσκονται σε αυτή τη μορφή έχουν σημαντικά μικρότερο μέγεθος. Στην πρώτη φάση ο αλγόριθμος centWave ψάχνει για ιόντα που χαρακτηρίζονται ως περιοχές με σφάλμα μικρότερο από την τιμή ακρίβειας μάζας (ορισμένη σε ppm) μεταξύ διαδοχικών φασμάτων πλήρους σαρώσεων που καλούνται περιοχές ενδιαφέροντος (ROIs). Στη δεύτερη φάση αυτές οι περιοχές αναλύονται περαιτέρω. Ο μετασχηματισμός συνεχούς εφαρμογής κυματιδίων (Continuous Wavelet Transformation-CWT) βρίσκει κορυφές στη χρωματογραφική διάσταση με διαφορετικά εύρη.

Προαιρετικά μπορούν να εφαρμοστούν οι παράμετροι σήμα προς θόρυβο (S/N) και σ_{gauss} (ρίζα μέσου τετραγωνικού σφάλματος της προσαρμογής της καμπύλης Gauss στην χρωματογραφική κορυφή) για να οδηγηθεί ο χρήστης σε μια λίστα κορυφών με λιγότερα εσφαλμένες θετικά ανιχνεύσεις. Η μέθοδος είναι

κατάλληλη για την ανίχνευση συνεκλουόμενων κορυφών σε αντίθεση με τον matchedfilter αλγόριθμο.

Η εικόνα 21 δείχνει ένα χρωματογράφημα και τα αντίστοιχα ίχνη m/z σε διαδοχικά φάσματα πλήρους σάρωσης. Η περιοχή αυτή χαρακτηρίζεται ως περιοχή ενδιαφέροντος λόγω του γεγονότος ότι η ακρίβεια μάζας του συγκεκριμένου m/z (σε ppm) είναι μικρότερη από το όριο των 20ppm, που έχει οριστεί, και το πλάτος κορυφής βρίσκεται εντός αποδεκτών τιμών [34, 35].



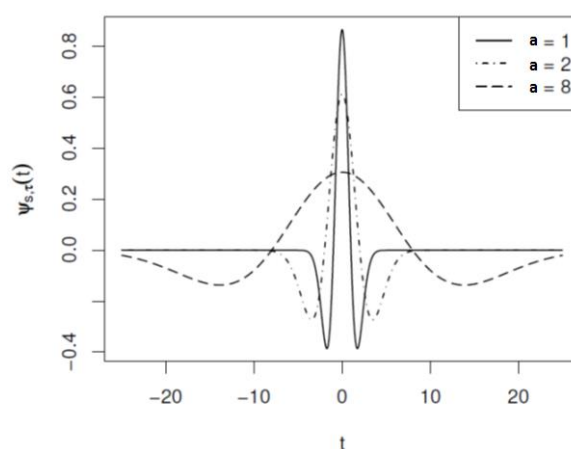
Εικόνα 21: Γραφική αναπαράσταση λειτουργίας του αλγορίθμου centWave για την εύρεση περιοχών ενδιαφέροντος (Regions Of Interest-ROIs), πηγή [35]

Η λέξη Wavelet ουσιαστικά προκύπτει από την γαλλική “ondelette” που σημαίνει μικρό κύμα. Απλουστευμένα ο μετασχηματισμός wavelet πραγματοποιεί κατά βάση μετασχηματισμό Fourier μόνο που αντί για δειγματοληψία με χρήση καθορισμένου παραθύρου (Windowed Fourier Transform) κάνει υπερδειγματοληψία του προς μελέτη σήματος, με βάση το κριτήριο δειγματοληψίας του Nyquist. Συνεπώς ικανοποιείται ο στόχος να επινοηθεί μια μέθοδος η οποία θα δίνει καλή ανάλυση χρόνου-συχνότητας σε οποιαδήποτε θέση στο επίπεδο σήματος χρόνος-ένταση ανεξάρτητα από το συχνотικό περιεχόμενο του σήματος. Με άλλα λόγια πρέπει να έχουμε μια συνάρτηση παραθύρου της οποίας η ακτίνα αυξάνει στον χρόνο (μειώνεται στην συχνότητα) καθώς αναλύει τα χαμηλής συχνότητας περιεχόμενα και

μειώνεται στον χρόνο (αυξάνει στην συχνότητα) όταν αναλύει τα υψηλής συχνότητας περιεχόμενα του σήματος. Η απαίτηση αυτή μας οδηγεί στη δημιουργία των συναρτήσεων Wavelet ως συναρτήσεις παραθύρου και στον Συνεχή Μετασχηματισμό Wavelet (Continuous Wavelet Transform). Μια από τις πιο συνήθεις συναρτήσεις είναι η Mexican Hat, η οποία έχει μαθηματική σχέση τη δεύτερη παράγωγο της Gaussian:

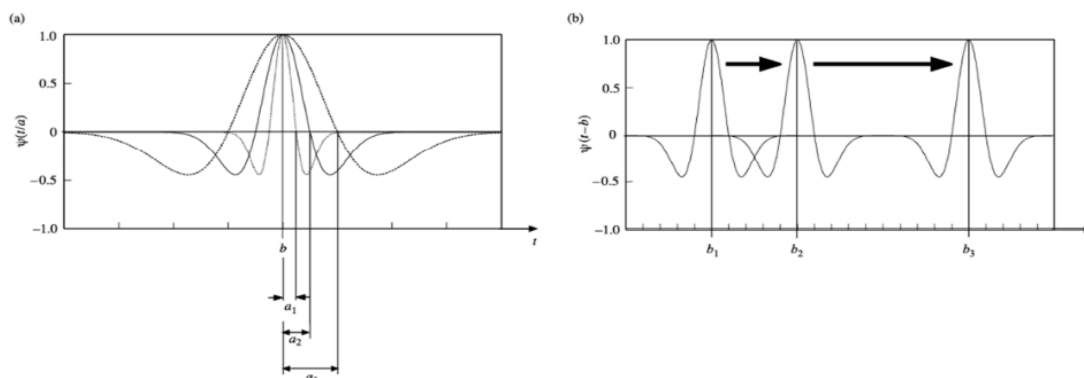
$$\psi(t)(1 - t^2) e^{-\left(\frac{t^2}{2}\right)}$$

Και αναπαρίσταται γραφικά στην ακόλουθη εικόνα:



Εικόνα 22: Συνάρτηση Mexican Hat για διαφορετικές τιμές εύρους

Η συνάρτηση Mexican Hat είναι γνωστή ως mother wavelet ή wavelet προς ανάλυση (analyzing wavelet). Η συνάρτηση όπως φαίνεται από την παραπάνω εικόνα μπορεί να διαστέλλεται ή να συστέλλεται.



Εικόνα 23: Κινήσεις διαστολής-συστολής και μετατόπισης της συνάρτησης Mexican Hat, πηγή [38]

Αν με a συμβολίσουμε το εύρος (όπως φαίνεται στην άνω εικόνα) και b τη μετατόπιση τότε:

$$\psi\left(\frac{t-b}{a}\right) = \left(1 - \left(\frac{t-b}{a}\right)^2\right) e^{-\left(\frac{(t-b)^2}{2a^2}\right)}$$

Τέλος ο μετασχηματμός συνεχούς εφαρμογής κυματιδίων δίνεται από τον τύπο:

$$W_{\psi}f(a, b) = \int_{-\infty}^{+\infty} f(t) \overline{\psi_{b,a}(t)} dt = \langle f(t), \psi_{a,b}(t) \rangle$$

Όπου

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right)$$

Όπου $f(t)$ είναι το σήμα και $\psi_{a,b}$ είναι η συνάρτηση Mexican Hat [38].

2.7.1.5 Massifquant

Ο Massifquant αποτελεί την εφαρμογή του αλγόριθμου TracMass στην γλώσσα προγραμματισμού R και μπορεί να κληθεί μέσω του πακέτου xcms. Αφορά δεδομένα που έχουν ληφθεί υπό μορφή γραμμών (line mode). Εάν θεωρήσουμε μια συγκεκριμένη μάζα η οποία όντως αντιστοιχεί σε ένα πραγματικό αναλύτη σε ένα συγκεκριμένο φάσμα πλήρους σάρωσης, τότε στο ακριβώς επόμενο φάσμα πλήρους σάρωσης θα παρατηρήσουμε μια μάζα πολύ όμοια με τη μάζα στο πρώτο φάσμα, ένταση που θα διαφέρει λίγο λόγω της χρωματογραφικής εξέλιξης. Επιπλέον, οι χαμηλής έντασης κορυφές θορύβου που εμφανίστηκαν στο πρώτο φάσμα θα έχουν εξαφανιστεί στο δεύτερο φάσμα πλήρους σάρωσης. Άρα ανιχνεύουμε δομικό σήμα το οποίο προέρχεται από εκλουόμενο συστατικό και θόρυβο που εμφανίζεται και εξαφανίζεται τυχαία. Άρα έχουμε ένα δισδιάστατο «πιάτο» με άξονες μάζα και ένταση το οποίο παρομοιάζει τη συμπεριφορά των σημάτων στα ραντάρ, όπου οι εκλουόμενες κορυφές είναι τα αντικείμενα που θέλει ο στρατός να εντοπίσει και ο θόρυβος ακολουθεί την ίδια τυχαιότητα. Για τον σκοπό αυτόν έχουν ήδη αναπτυχθεί αλγόριθμοι για την ανίχνευση αεροπλάνων και κινούμενων αντικειμένων που είναι γνωστοί ως αλγόριθμοι ανίχνευσης (tracking algorithms). Ένας τέτοιος αλγόριθμος είναι το φίλτρο Kalman που επιτρέπει την εξαγωγή χημικά σχετικής πληροφορίας και την αποβολή του θορύβου.

Όταν ανιχνεύεται ένα αντικείμενο στο ραντάρ πρέπει να ληφθούν υπόψη: πρώτον η κίνηση του αντικειμένου, δεύτερον η αβεβαιότητα της κίνησης και τρίτον ο θόρυβος. Για πολύ μικρό χρονικό διάστημα το αντικείμενο ακολουθεί ομαλή κατεύθυνση ενώ για μεγαλύτερους χρόνους είναι πολύ πιθανό να αλλάξει πορεία.

Ακολουθεί σύντομη μαθηματική θεμελίωση του μοντέλου με τις εξισώσεις που προκύπτουν για εύρεση αντικειμένων που κινούνται με σταθερή ταχύτητα παρουσία τυχαίων μεταβολών ταχύτητας μεταξύ μετρήσεων.

Το μοντέλο εκφράζεται ως $x_k = Ax_{k-1}$, όπου $x_k = \begin{bmatrix} x_k \\ \dot{x}_k \end{bmatrix} = \begin{bmatrix} position \\ velocity \end{bmatrix}$,

$A = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ και T είναι το διάστημα μεταξύ των μετρήσεων.

Οι εξισώσεις πρόβλεψης για τη μελλοντική κατάσταση (k) του αντικειμένου είναι οι εξής:

$$\hat{x}_k^* = A\hat{x}_{k-1}$$

$$\hat{P}_k^* = A\hat{P}_{k-1}A^T + Q$$

Όπου P είναι πίνακας διακύμανσης-συνδιακύμανσης για την κατάσταση x και Q η αβεβαιότητα του πόσο πολύ ίσως αλλάξει πορεία το αντικείμενο μεταξύ δυο μετρήσεων. Οι μεταβλητές με αστερίσκο συμβολίζουν τις προβλεπόμενες μεταβλητές.

Τέλος οι εξισώσεις που εισάγουν τις επόμενες πληροφορίες που προκύπτουν είναι οι εξής:

$$K_k = P_k^* * H^T (HP_k^* H^T + R)^{-1}$$

$$\hat{x}_k^* = \hat{x}_k^* + K_k(z_k - H\hat{x}_k^*)$$

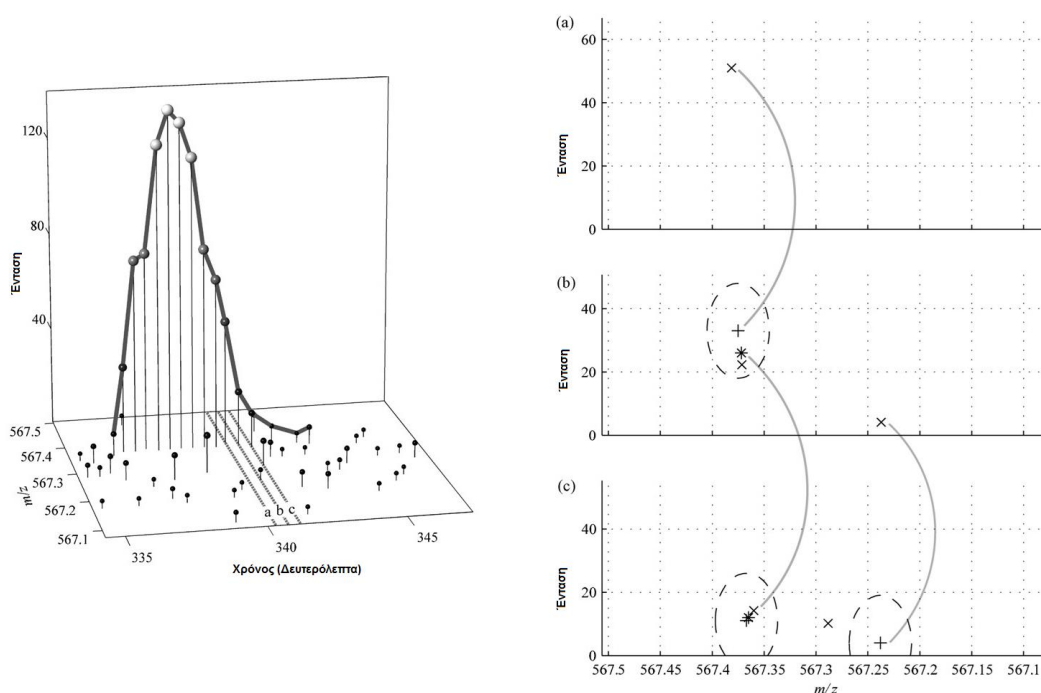
$$P_k = (I - K_k H)P_k^*$$

Όπου K_k είναι ένας πίνακας από βάρη που καθορίζει πόσο πολύ το φίλτρο υιοθετεί τη νέα μέτρηση, H είναι ένας πίνακας μετασχηματισμού που όταν πολλαπλασιάζεται με ένα διάνυσμα κατάστασης προκύπτει μετρήσιμο μέγεθος θέσης. Αν η θέση είναι η μόνη παράμετρος που μπορεί να υπολογιστεί και το μοντέλο περιλαμβάνει και θέση και ταχύτητα τότε $H=[1 \ 0]$. Τέλος R είναι πίνακας

διακύμανσης-συνδιακύμανσης της μέτρησης και καλείται αβεβαιότητα της μέτρησης.

Στην περίπτωση εύρεσης κορυφών ισχύουν κατ' αντιστοιχία οι ίδιες εξισώσεις με τη διαφορά ότι αντί για αζιμουθιακό υπάρχει το m/z , αντί για το εύρος του ραντάρ υπάρχει η ένταση και αντί για σαρώσεις ραντάρ υπάρχουν οι πλήρεις φασματικές σαρώσεις του φασματομέτρου μαζών. Ο θόρυβος συμπεριφέρεται με μη προβλεπόμενο τρόπο, ενώ το πραγματικό σήμα από εκλούόμενη ουσία είναι καλά καθορισμένη από άποψη m/z και ακολουθεί ένταση αρχικά αυξανόμενη που αποκτά το μέγιστο της και έπειτα σταδιακά μειώνεται. Το φίλτρο Kalman ανιχνεύει το m/z μεταξύ των φασματικών σαρώσεων καθώς εξελίσσεται χρωματογραφικά μια κορυφή.

Η ακόλουθη εικόνα αναπαριστά πώς δουλεύει ο αλγόριθμος για τρεις διαδοχικές φασματικές σαρώσεις όπου παρατηρούνται αντικείμενα-μάζες που απορρίπτονται και άλλες που επιλέγονται.



Εικόνα 24: Γραφική αναπαράσταση λειτουργίας του αλγορίθμου Massifquant για τρεις συνεχόμενες πλήρεις σαρώσεις, πηγή [39]

Στην εικόνα αναπαρίστανται τρεις διαδοχικές σαρώσεις στο χρόνο a,b και c όπου εμφανίζεται μια κορυφή με $m/z=567,35$ και θόρυβος. Το φίλτρο εντοπίζει την κορυφή ενώ αγνοεί τον θόρυβο. Με (x) αναπαρίστωνται οι τιμές που μετρήθηκαν, με (+) οι προβλεπόμενες τιμές και με (*) την κατάσταση του

\hat{x} φίλτρου Kalman από την οποία θα προκύψει η νέα πρόβλεψη, ενώ με διακεκομμένες οβάλ γραμμές το διάστημα εμπιστοσύνης [39, 40].

2.7.1.6 enviPick

Ο αλγόριθμος enviPick χρησιμοποιείται για δεδομένα αφενός διορθωμένα ως προς τη γραμμή βάσης και αφετέρου για δεδομένα που θα πρέπει να βρίσκονται υπο μορφή γραμμών (line mode). Ο enviPick έχει ελεγχθεί μόνο σε υψηλής διακριτικής ικανότητας δεδομένα και στηρίζεται σε επεξεργασία τριών βημάτων:

1. διαμερισμός των μετρήσεων (data partitioning) με την εντολή `mzagglom()`
2. μη επιβλεπόμενη ομαδοποίηση των χρωματογραφημάτων με την εντολή `mzClust()` και
3. ανίχνευση κορυφών ανεξαρτήτως σχήματος εντός του καθ' ενός χρωματογραφήματος με την εντολή `mzpick()`

Ο enviPick διαβάζει .mzXML αρχεία και επομένως τα δεδομένα εισάγονται ως τριάδες (m/z , ένταση, χρόνος ανάσχεσης). Το πρώτο βήμα δεν είναι τίποτα άλλο από μια διαμερισματοποίηση των μετρήσεων (m/z , ένταση, χρόνος ανάσχεσης), η οποία είναι απαραίτητη, έτσι ώστε να επιταχυνθεί ταχύτερα το επόμενο βήμα της μη επιβλεπόμενης ομαδοποίησης των ιόντων. Για αυτόν τον σκοπό, χρησιμοποιείται ένας ιεραρχικός αλγόριθμος ομαδοποίησης (hierarchical agglomerative clustering algorithm) που συνδυάζει τις μετρήσεις m/z και χρόνου ανάσχεσης (rt). Λόγω του ότι κάθε σημείο $(m/z, rt)_1$ είναι ένα σημείο στον δισδιάστατο χώρο που απέχει κάποια απόσταση από κάποιο άλλο σημείο $(m/z, rt)_2$ γίνεται ομαδοποίηση με βάση δεδομένη ανοχή στο χρόνο ανάσχεσης (`drtgap value`) και στη μάζα (`dmzgap value`). Στόχος είναι να προκύψουν τελικά υπομάδες, έτσι ώστε καμία μέτρηση μεταξύ δυο διαφορετικών υπομάδων να μην μπορεί να βρίσκεται πιο κοντά από το `drtgap` και το `dmzgap` με μια άλλη μέτρηση μιας άλλης υπομάδας.

Στο δεύτερο βήμα εντός κάθε διαμερίσματος πραγματοποιείται μη επιβλεπόμενη ομαδοποίηση, έτσι ώστε να σχηματισθούν τα χρωματογραφήματα (EICs) των ιόντων. Για τον σκοπό αυτό, η πρώτη ομάδα ιόντων ξεκινάει με το ισχυρότερο σε ένταση ιόν. Καθώς προχωράμε προς ιόντα μικρότερης έντασης αυτά ομαδοποιούνται είτε με το αρχικό υψηλότερης έντασης ιόν είτε δημιουργούν μια άλλη ομάδα. Προκειμένου να ομαδοποιηθούν

με το ισχυρής έντασης ιόν οι μετρήσεις πρέπει να βρίσκονται εντός των ανοχών της τιμής $dmzdens$ ($dmzdens$ είναι η διακύμανση σε ppm για την οποία εκτελείται χρωματογράφημα (EIC), δηλαδή για παράδειγμα $\pm 0,005$) και της τιμής $drt dens$ (διακύμανση του χρόνου γύρω από τον οποίον πραγματοποιείται το χρωματογράφημα, για παράδειγμα $\pm 0,5min$). Εάν ένα ιόν μπορεί να ομαδοποιηθεί με περισσότερες ομάδες από μια τότε κατατάσσεται στην ομάδα με τη μικρότερη διαφορά στη μάζα μεταξύ της μέτρησης m/z και του μέσου m/z της ομάδας. Κάθε φορά που πραγματοποιείται μια νέα ανάθεση σε μια ήδη υπάρχουσα ομάδα, η εκτίμηση m/z μπορεί να βελτιωθεί, για παράδειγμα η ανοχή γύρω από το μέσο m/z σταδιακά μειώνεται από το $2*dmzdens$ στο $dmzdens$. Επιπλέον το $dmzdens$ χρησιμοποιείται για την ενημέρωση της ανοχής του χρόνου μιας ομάδας έπειτα από κάθε ανάθεση. Όταν εξαντληθούν οι μετρήσεις, οι ομάδες συγχωνεύονται και δημιουργούν ως έξοδο (output) μια τιμή m/z και τα όρια στην τιμή του λόγου m/z . Τελικά έχουμε μια ομάδα ιόντων τα οποία φιλτράρονται, έτσι ώστε να ικανοποιούν την ελάχιστη ένταση (minimum intensity) που έχουμε θέσει εξ' αρχής ως την ελάχιστη ένταση πάνω από την οποία θεωρούμε το σήμα κορυφή.

Το τρίτο βήμα κατά βάση φιλτράρει περαιτέρω τις κορυφές και εξομαλύνει το σχήμα των κορυφών. Αρχικά ανιχνεύονται κενά μεταξύ των πλήρων σαρώσεων. Για παράδειγμα ως «κενά» νοούνται σημεία στα οποία η κορυφή δεν ανιχνεύεται, ενώ ανιχνεύεται στο προηγούμενο και επόμενο πλήρους σάρωσης φάσμα. Τα κενά αυτά συμπληρώνονται με τη μέθοδο της παρεμβολής (Interpolation). Επιτρέπεται συμπλήρωση κενών όχι μεγαλύτερων από μια τιμή, δηλαδή δεν γίνεται να υπάρχουν κενά για παράδειγμα πάνω από 4 πλήρεις σαρώσεις και όχι πάνω από ένα αριθμό που ορίζεται επίσης από τον χρήστη. Εκτός από τα κενά, δεδομένου του κέντρου του χρόνου ανασχεσης μιας κορυφής αν παρακείμενα σημεία της κορυφής εμφανίζονται με μεγαλύτερη ένταση τότε, αυτά υφίσταται υποβιβασμό με μια τιμή-βάρος που επίσης εισάγει ο χρήστης. Τέλος ακολουθεί φιλτράρισμα των κορυφών, έτσι ώστε να έχουν τουλάχιστον μια ένταση που έχει καθορίσει ο χρήστης, να έχει τουλάχιστον λόγο σήμα προς γραμμή υποβάθρου (signal to base) και σήμα προς θόρυβο (S/N) τιμές που επίσης καθορίζει ο χρήστης.

Όπως καθίσταται σαφές από τα παραπάνω αυτός ο αλγόριθμος εύρεσης κορυφών πρέπει να βελτιστοποιηθεί για το εκάστοτε όργανο και χρωματογραφία αλλά παρόλα αυτά είναι αρκετά αποτελεσματικός [41].

2.7.2 Ομαδοποίηση κορυφών που αναπαριστούν τον ίδιο αναλύτη σε όλα τα δείγματα

Το αποτέλεσμα της διαδικασίας ανεύρεσης ιόντων είναι ένας πίνακας με γραμμές τα ιόντα και στήλες τα m/z , rt , rt_{min} , rt_{max} , $into$, $intb$, $maxo$ και $\# sample$ όπως φαίνεται στον παρακάτω πίνακα:

Πίνακας 6: Τυπική έξοδος (output) οποιουδήποτε αλγορίθμου εύρεσης ιόντων

mz	m/z _{min}	m/z _{max}	Rt(s)	Rt _{min} (s)	Rt _{max} (s)	into	intb	maxo	S/N	#sample
74,0962	74,0950	74,0973	397,7	397,4	397,9	16378,2	15266	48900	33	1
84,9594	84,9588	84,9600	1123,8	1123,5	1124,0	8669	7935,49	21984	24	1
84,9597	84,9593	84,9603	693,9	693,6	694,1	17979	15845,2	49360	26	1
90,9766	90,9756	90,9773	18,4	4,7	38,4	1,2E+07	1,2E+07	2213416	2E+06	1
90,9768	90,9756	90,9787	454,6	454,3	483,1	98472,6	75267	14660	11	1
102,1284	102,1264	102,1305	392,9	354,0	428,1	3049793	2776932	309516	54	1
102,9709	102,9699	102,9718	693,1	673,4	693,4	61591,7	47808,2	11852	16	1

Το m/z είναι ο λόγος μάζας προς φορτίο που έχει ανιχνευθεί και έχει προκύψει ως ο μέσος όρος της κατώτερης και ανώτερης μάζας προς φορτίο που έχει παρατηρηθεί στα δείγματα. Ακολουθεί ο χρόνος ανάσχεσης σε δευτερόλεπτα, ο μικρότερος και ο μεγαλύτερος χρόνος ανάσχεσης (rt_{min} και rt_{max}). Ως $into$ αναφέρεται το εμβαδόν της κορυφής που προκύπτει από την ολοκλήρωση της κορυφής, $intb$ είναι το εμβαδόν της κορυφής που όμως αυτή τη φορά είναι διορθωμένο ως προς της γραμμή βάσης-το “b” υποδηλώνει τη γραμμή βάσης- και $maxo$ είναι η μέγιστη ένταση που παρατηρείται-το “m” υποδηλώνει το μέγιστο. Επιπλέον ο πίνακας εμπεριέχει τον λόγο σήμα προς θόρυβο και ως

τελευταία στήλη τον αριθμό του δείγματος από το οποίο ανιχνεύθηκε το συγκεκριμένο ιόν.

Εφόσον χρησιμοποιήσουμε τον αλγόριθμο centWave και θέσουμε παραμέτρους fitgauss=TRUE και verbose.list=TRUE προκύπτουν πέρα από το βασικό πίνακα επιπλέον για κάθε ιόν κάποιοι επιπλέον παράμετροι, οι οποίοι απεικονίζονται στον ακόλουθο πίνακα:

Πίνακας 7: Επιπλέον στοιχεία εξόδου (output) του αλγορίθμου centWave

egauss	mu	sigma	h	f	dppm	scale	scpos	scmin	scmax
0,182	473,9	7,302	1033	77	17	26	475	449	501
0,144	474,6	8,621	2645	85	6	20	475	455	495
0,198	545,6	10,915	1382	108	9	24	548	524	572
0,137	533,0	14,845	2558	115	13	22	533	511	555
0,157	547,6	10,236	1122	117	14	20	549	529	569
0,166	554,0	7,866	1029	119	7	24	556	532	580
0,173	583,1	3,868	12058	134	7	20	584	564	604

Egauss: μέση τετραγωνική ρίζα του σφάλματος προσαρμογής του μοντέλου Gauss- Είναι δείκτης ποιότητας κατά πόσο η κορυφή έχει μορφή καμπάνας Gauss

mu: η παράμετρος μ του μοντέλου Gauss,

sigma: η παράμετρος σ του μοντέλου Gauss,

h: η παράμετρος h του μοντέλου Gauss,

f: η περιοχή ενδιαφέροντος (ROI) όπου βρέθηκε ο αριθμός συνεχόμενων μαζών εντός του σφάλματος που έχει οριστεί,

dppm: η απόκλιση m/z των μαζών στα συνεχόμενα φάσματα πλήρους σάρωσης

scale: η κλίμακα του wavelet που ανίχνευσε την κορυφή,

scpos: η θέση της κορυφής όπως βρέθηκε από την ανάλυση κυματιδίων,

scmin: ο αριθμός του φάσματος πλήρους σάρωσης αριστερά της κορυφής

scmax: ο αριθμός του φάσματος πλήρους σάρωσης δεξιά της κορυφής

Ο αριθμός των γραμμών μπορεί να είναι αρκετά μεγάλος και εξαρτάται από τον αριθμό δειγμάτων και το είδος του δείγματος. Τυπικά, για εισερχόμενα υγρά απόβλητα ο μέσος αριθμός γραμμών είναι 4500 για κάθε δείγμα.

Είναι κατανοητό ότι ο πίνακας μπορεί να γιγαντωθεί, καθώς αν έχουμε 20 δείγματα εισερχόμενων λυμάτων τότε προκύπτει πίνακας 90.000 γραμμών. Όπως είναι λογικό το επόμενο στάδιο είναι η ομαδοποίηση των γραμμών αυτών, δηλαδή η ομαδοποίηση των κοινών ιόντων σε όλα τα δείγματα. Για παράδειγμα το ιόν με $m/z=74,0962$ που προκύπτει στο δείγμα 1 αντιπροσωπεύει τον ίδιο αναλύτη που δημιουργεί το ιόν με $m/z=74,0960$ στο δείγμα 2, διότι έχουν εντός σφάλματος ίδια μάζα και εντοπίστηκαν στον ίδιο χρόνο ανάσχεσης σε δυο διαφορετικά δείγματα.

Υπάρχουν τρεις αλγόριθμοι-μέθοδοι ομαδοποίησης διαθέσιμοι: “mzClust”, “nearest” και ο “density”. Ο mzClust βασίζεται στην ιεραρχική ομαδοποίηση, ο nearest βασίζεται στην μέθοδο του κοντινότερου γείτονα (k Nearest Neighbor-kNN) και ο density στον εκτιμητή της πυκνότητας πιθανότητας (kernel density estimator). Παρακάτω θα δούμε πολύ συνοπτικά τον καθένα από αυτούς τους αλγόριθμους και θα σχολιάσουμε την αποτελεσματικότητά τους [34].

2.7.2.1 mzClust

Ως είσοδο (input) ο αλγόριθμος mzClust παίρνει τον αριθμό των δειγμάτων (N), το σφάλμα του φασματομέτρου μάζας σε ppm, τον ελάχιστο αριθμό δειγμάτων που απαιτούνται να ληφθούν ώστε να σχηματιστεί μια ομάδα M όπου $M \leq N$ και μια λίστα τιμών m/z που επιθυμούμε να κατηγοριοποιήσουμε μαζί με μια ετικέτα που δίνει την πληροφορία σε ποιο δείγμα ανήκει η κάθε τιμή m/z .

Τα βήματα του αλγόριθμου mzClust είναι τα εξής:

1. Βρες το μικρότερο m/z από κάθε δείγμα και φτιάξε το αντικείμενο **A**.

2. Βρες τη μικρότερη τιμή $A[x]$ του αντικείμενου A και μετέφερε την στο B1. Έλεγξε αν κάθε τιμή m/z του A μείον την $A[x]$ είναι εντός εύρους σφάλματος του φασματομέτρου ($A[\text{μικρότερο}] - A[\text{επόμενο}] < \text{σφάλμα σε ppm}$) και βάλε αυτά που συμμορφώνονται στην ομάδα **B1**
3. Επανάλαβε το 2 (προφανώς για τα m/z τα οποία έχουν απομείνει στο A), δηλαδή πάρε τη πιο μικρή τιμή και βάλε την στην ομάδα **B2** και μετά δες ποια m/z συμμορφώνονται ως προς την ακρίβεια μάζας και πρόσθεσέ τα στην ομάδα **B2**.
4. Τα **B1** και **B2** αντιπροσωπεύουν δυο γειτονικά bins. Πάρε την μικρότερη τιμή του **B1** και τη μεγαλύτερη του **B2** και δες αν εμπεριέχονται στο σφάλμα ακρίβειας μάζας.

Αν ναι τότε έχουμε επικαλυπτόμενα κομμάτια (bins), άρα συνένωσε τα **B1** και **B2** και εφάρμοσε ιεραρχική συσταδοποίηση στις κορυφές. Έτσι μπορεί να πάρουμε έως 3 ομάδες (clusters) σε ένα αντικείμενο **C**. Το **B2** παίρνει την τιμή της τελευταίας ομάδας αντικειμένου του **C**.

Αν έχω μεγαλύτερο ή ίσο του 2 αριθμό ομάδων τότε το **B1** παίρνει την 2^η και τελευταία συστάδα (cluster) του C ενώ αν C έχει 1^α ομάδα τότε το αντικείμενο **B1** είναι κενό.
Αν έχω μεγαλύτερο ή ίσο των 3^{ων} συστάδων τότε προκύπτει το **C1** και το **C3** (έξοδος **C1** και **C3**).

Αν μεγαλύτερη τιμή του **B2** και μικρότερη τιμή του **B1** δεν εμπεριέχονται στο σφάλμα του φασματομέτρου τότε έλεγξε αν εντός των **B1** και **B2** χωριστά υπάρχουν στοιχεία με διαφορά μεγαλύτερη από το φασματομέτρο μάζας και αν ναι εκτέλεσε ιεραρχική συσταδοποίηση, δηλαδή βήμα 4.

5. Έξοδος το B1
6. Ανάθεσε το B2 στο B1
7. Επανάλαβε τα βήματα 2,3,4,5,6 μέχρι να μην υπάρχουν άλλα δεδομένα.

Ως έξοδος προκύπτει η λίστα m/z που δόθηκε ως είσοδος αλλά ομαδοποιημένη, ο αριθμός των m/z που ομαδοποιήθηκαν και σε ποια δείγματα προέκυψαν τα m/z , όπως ο ακόλουθος πίνακας:

Πίνακας 8: Ενδεικτικός ομαδοποιημένος πίνακας ιόντων για τρία δείγματα

Mzmed	npeaks	Sample 1	Sample 2	Sample 3	Peak area 1	Peak area 2	Peak area 3
57,0743	3	1	1	1	58043	66525	99523
58,0722	3	1	1	1	12000	11088	9858
59,0488	3	1	1	1	1,8E+06	1,7E+06	0,9E+06

60,0427	2	1	0	1	25050	0	28010
60,440	2	1	1	0	98587	92101	0
60,4449	1	0	0	1	0	0	15010

Το μειονέκτημα του αλγόριθμου αυτού είναι ότι δεν λαμβάνει υπόψη του το χρόνο ανάλυσης. Έτσι είναι δυνατόν να συσταδοποιησει m/z που προκύπτουν σε διαφορετικό χρόνο ανάλυσης. Επίσης έχει παρατηρηθεί ότι ιόντα που προκύπτουν εντός του ορίου ακρίβειας του φασματομέτρου μάζας από το ίδιο δείγμα δεν συσταδοποιούνται [42].

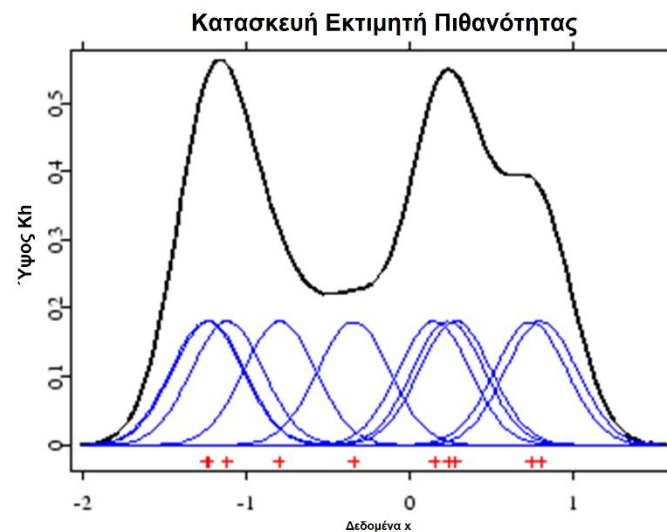
2.7.2.2 density

Ο αλγόριθμος ομαδοποίησης density βασίζεται στον εκτιμητή της πυκνότητας πιθανότητας (Kernel Density Estimator), που είναι μη παραμετρική τεχνική. Τα μη παραμετρικά μοντέλα διαχωρίζονται από τα παραμετρικά μοντέλα στο γεγονός ότι προσφέρουν μεγαλύτερη ευελιξία ελέγχου της προσαρμογής του μοντέλου. Τα παραμετρικά μοντέλα χρησιμοποιούν τα δεδομένα για την ανεύρεση παραμέτρων που επιτυγχάνουν την ιδανικότερη προσαρμογή. Για αυτό το λόγο ονομάζονται παραμετρικά, διότι βρίσκουμε παραμέτρους με την βοήθεια των δεδομένων. Για παράδειγμα στην απλή γραμμική παλινδρόμηση βρίσκουμε τις παραμέτρους α και β και σ^2 του γραμμικού μοντέλου. Ο σκοπός των μη παραμετρικών μοντέλων είναι να προβλέπουν καλά όταν δεν μπορούμε να διαβεβαιώσουμε ότι ισχύουν οι προϋποθέσεις εφαρμογής των παραμετρικών μοντέλων. Μη παραμετρικός τρόπος δεν σημαίνει κατ' ανάγκη πολύπλοκος τρόπος αλλά σημαίνει ευέλικτος τρόπος.

Η πιο απλή μη παραμετρική έκδοση του εκτιμητή της πυκνότητας πιθανότητας (kernel density estimator) είναι το ιστόγραμμα. Το ιστόγραμμα είναι το γράφημα που ο x άξονας υποδιαιρείται σε τμήματα (bins). Κάθε τμήμα έχει ύψος στον y άξονα ίσο με τον αριθμό των μετρήσεων που παρατηρούνται στο δείγμα.

Το πρόβλημα των ιστογραμμάτων είναι ότι υποθέτουν σταθερή πυκνότητα πιθανότητας για όλες τις παρατηρήσεις μέσα σε ένα τμήμα. Εάν μειώσουμε το εύρος των τμημάτων αυτών ώστε αυτό να τείνει στο μηδέν τότε επειδή δεν έχουμε άπειρο αριθμό δεδομένων αλλά ένα καθορισμένο αριθμό δεδομένων, σταδιακά ίσως να μην υπάρχει κάποια παρατήρηση σε ένα τμήμα με

αποτέλεσμα να προκύπτει θορυβώδης εικόνα. Αντί της μείωσης του εύρους των τμημάτων, αυτό που μπορούμε να κάνουμε είναι να χρησιμοποιήσουμε την ιδέα του εκτιμητή της πυκνότητας πιθανότητας.



Εικόνα 25: Απεικόνιση εκτιμητή της πυκνότητας πιθανότητας

Στην παραπάνω εικόνα κάθε + απεικονίζει μια παρατήρηση. Κάθε παρατήρηση δημιουργεί ένα πυρήνα (kernel) που αναπαρίσταται ως μπλέ γκαουσιανές καμπύλες. Οι πυρήνες μπορεί να ακολουθούν οποιαδήποτε κατανομή. Εδώ για λόγους απλοποίησης ακολουθούν κανονική κατανομή. Επομένως ισχύει για όλους τους πυρήνες ότι : $\int \sum_{i=1}^N k(x_i|x_0)dx = 1$.

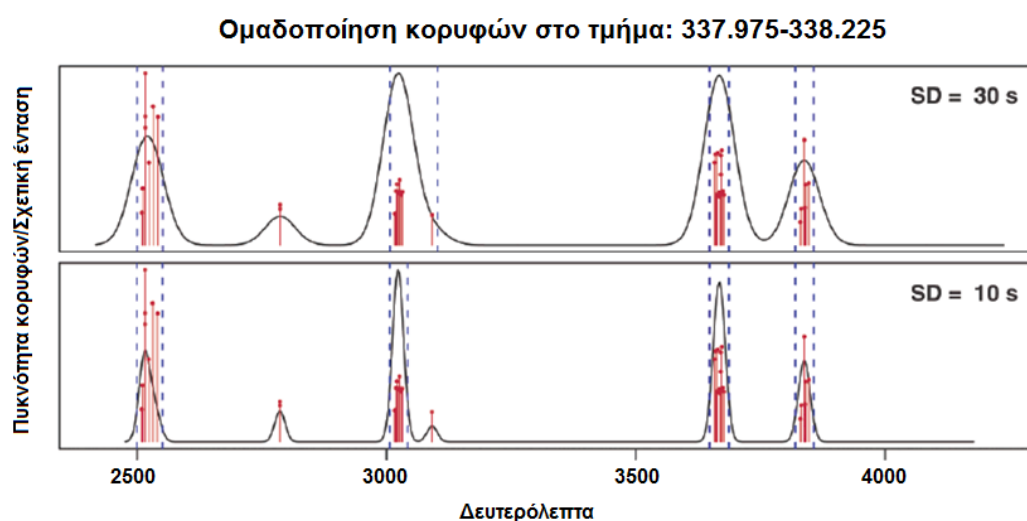
Έστω τώρα σημείο x_0 όπου δεν συμπίπτει με κάποια παρατήρηση. Υπολογίζουμε την πυκνότητα πιθανότητας ως ένα ζυγισμένο μέσο όρο των κοντινότερων παρατηρήσεων. Όποια παρατήρηση βρίσκεται πλησιέστερα στο σημείο x_0 θα πάρει μεγαλύτερο βάρος από κάποια άλλη παρατήρηση πιο απομακρυσμένη. Ισχύει η μαθηματική σχέση:

$$f_h(x_0) = \frac{1}{N * h \sum_{i=1}^N k(x_i|x_0)}$$

Όπου εδώ το h δεν συμβολίζει το εύρος των τμημάτων, όπως στις προηγούμενες εξισώσεις αλλά συμβολίζει το εύρος (bandwidth), δηλαδή πόσο μακριά πάμε από το σημείο που θέλουμε να μετρήσουμε την πυκνότητα. Ακόμα, k είναι η τιμή του σημείου του πυρήνα (μπλε γραφικής παράστασης) και $\sum k(x_i|x_0)$ είναι η τιμή αθροίσματος ζυγισμένων πυρήνων

Εφαρμογή αλγόριθμου

Ο αλγόριθμος αντιστοίχισης density είχε συμπεριληφθεί στο αρχικό πακέτο (package) του xcms και λαμβάνει υπόψη του τη δισδιάστατη ανισοτροπική φύση των δεδομένων LC-MS. Η ακρίβεια των φασματομέτρων είναι περισσότερο κατανοητή από ό,τι είναι η ολίσθηση του χρόνου ανάλυσης. Έτσι ο αλγόριθμος χρησιμοποιεί ένα συγκεκριμένο παράθυρο m/z που καθορίζεται από την ακρίβεια του φασματομέτρου μαζών και τυπικά είναι 3 φορές η ακρίβεια του φασματομέτρου εκφρασμένη σε ppm (στην περίπτωση μας $3 \times 0,005 = 0,015$). Μάλιστα για να αποφευχθεί διαχωρισμός μιας ομάδας κορυφών σε δυο διαδοχικά τμήματα, εξαιτίας του αυθαίρετου ορισμού του παράθυρου m/z , λαμβάνονται υπόψη για επικαλυπτόμενες ομάδες γειτονικά τμήματα. Για παράδειγμα θα ληφθούν υπόψη τα γειτονικά τμήματα: 256,2000-256,2015, 256,2007-256,0023, ώστε να διαπιστωθεί αν μια μάζα που ανήκει σε μια ομάδα συμπεριλαμβάνεται και στα δυο τμήματα οπότε λαμβάνεται μέριμνα για την τοποθέτηση του ιόντος στην πραγματική ομάδα που ανήκει.



Εικόνα 26: Παράδειγμα ομαδοποίησης ιόντων (features) σε 12 δείγματα, πηγή [33]

Στην άνω εικόνα, οι κόκκινες γραμμές αντιπροσωπεύουν κορυφές και το ύψος τους την σχετική ένταση τους. Το εξομαλυμένο προφίλ πυκνότητας πιθανότητας φαίνεται με συνεχόμενη γραμμή. Οι ομάδες που ανιχνεύθηκαν σημειώνονται με διακεκομμένες μπλε γραμμές. Το προφίλ πάρθηκε για εύρος 10 δευτερόλεπτα και 30 δευτερόλεπτα. Στα 10 δευτερόλεπτα παρατηρείται πως μία μέτρηση αποφεύγεται ορθά να κατηγοριοποιηθεί σε μια ομάδα κορυφών.

Επομένως αρχικά ομαδοποιούνται οι κορυφές με βάση τη μάζα και έπειτα βρίσκονται κορυφές στη διάσταση του χρόνου για κάθε τμήμα. Ο αλγόριθμος υπολογίζει την κατανομή των κορυφών στο χρωματογραφικό χρόνο και ανιχνεύει τα όρια των περιοχών όπου πολλές κορυφές έχουν παρόμοιο χρόνο ανάσχεσης. Ο τρόπος που γίνεται αυτό είναι η μέθοδος του εκτιμητή της πυκνότητας πιθανότητας. Έτσι επιτυγχάνεται ομαδοποίηση των ιόντων αυτών που έχουν μάζα εντός εύρους ακρίβειας του φασματομέτρου μαζών και χρόνο ανάσχεσης όχι μεγαλύτερο από ένα καθορισμένο παράθυρο που κάθε φορά ορίζει η μέθοδος του εκτιμητή της πυκνότητας πιθανότητας [33].

2.7.2.3 K-Nearest

Η μέθοδος αυτή ομαδοποίησης κορυφών έχει βάση στο λογισμικό mzMine και βασίζεται στον αλγόριθμο K κοντινότερων γειτόνων. Η ιδέα είναι ότι η μάζα και ο χρόνος δημιουργούν ένα χώρο στον οποίο μπορούν να ομαδοποιηθούν οι μάζες ανάλογα με την απόσταση που βρίσκονται μεταξύ τους. Ως είσοδο στον αλγόριθμο δίνουμε το μέγιστο ανεκτό m/z και το μέγιστο ανεκτό χρόνο ανάσχεσης για να κατηγοριοποιηθούν δυο ιόντα μαζί. Επίσης δίνουμε ως δεδομένα τον αριθμό των κοντινότερων γειτόνων που πρέπει να ελεγχτούν καθώς και το μέγιστο αριθμό μελών μιας ομάδας [43].

2.7.3 Διόρθωση χρόνου ανάσχεσης λόγω ολίσθησης

Μετά την ομαδοποίηση των ιόντων σε ομάδες, ακολουθείται διόρθωση των χρόνων ανάσχεσης που μπορούν να συμβούν εξαιτίας ολίσθησης του υγροχρωματογραφικού συστήματος. Αυτό το βήμα είναι απαραίτητο ώστε να γίνει η ευθυγράμμιση των κορυφών που έχουν ανιχνευθεί μεταξύ των δειγμάτων. Το χρωματογραφικό σύστημα που χρησιμοποιήθηκε στα πλαίσια της εκπόνησης της διπλωματικής εργασίας είναι αρκετά επαναλήψιμο και ως εκ τούτου ολίσθηση άνω των 0,2 λεπτών δεν παρατηρήθηκε. Δυο είναι οι αλγόριθμοι που διατίθενται στο πακέτο xcms για αυτόν τον σκοπό: ο `obiwarp` και ο `loess` [34].

2.7.3.1 Αλγόριθμος OBI-Warp

Ο OBI-Warp κατά βάση αποτελεί επέκταση της μεθόδου της δυναμικής χρονικής στρέβλωσης (Dynamic Time Warping, DTW). Οπότε θα εξετάσουμε πρώτα τον DTW και έπειτα θα αναφέρουμε τις τροποποιήσεις που πραγματοποιήθηκαν σε αυτόν και δημιουργήθηκε ο OBI-Warp.

Ο DTW αποτελεί συνηθισμένο αλγόριθμο που χρησιμοποιείται ως μέτρο ομοιότητας μεταξύ δύο ακολουθιών και επινοήθηκε το 1957 από τον Bellman με σκοπό την επεξεργασία ομιλίας. Χρησιμοποιείται ευρέως, όταν δύο ακολουθίες έχουν διαφορετικά μήκη, όπου κατά συνέπεια, είναι αδύνατος ο υπολογισμός της ανομοιότητας με χρήση Ευκλείδειας απόστασης. Η μέθοδος DTW «εκτείνει» ή «συστέλλει» τις ακολουθίες, προκειμένου να καταστεί εφικτή η αντιστοίχισή τους. Η μέθοδος DTW η οποία επεκτάθηκε για να υποστηρίξει πολυδιάστατες ακολουθίες, αναλύεται ως εξής:

Ο στόχος είναι να συγκριθούν δύο σύνολα των διανυσμάτων,

$$X=x_1,x_2,\dots,x_i,\dots,x_m$$

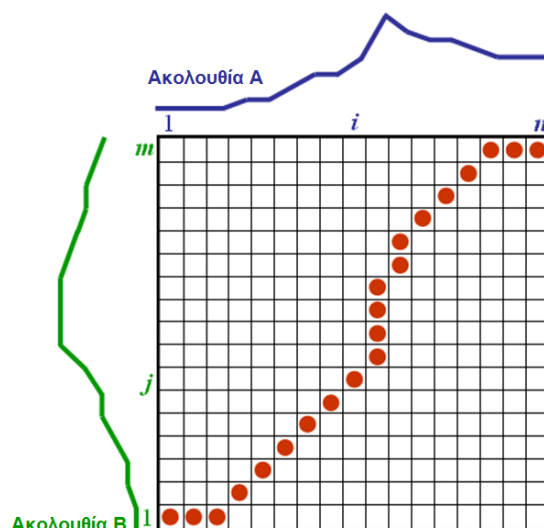
$$Y=y_1,y_2,\dots,y_i,\dots,y_n$$

τα οποία έχουν διαφορετικό μήκος m και n , δηλαδή διαφορετικό αριθμό σημείων. Μια σημαντική υπόθεση είναι ότι και τα δύο διανύσματα έχουν ένα σταθερό σημείο εκκίνησης, το οποίο στην παρούσα περίπτωση ορίζεται ως έναρξη το $w_1=(1,1)$ και ως αφετηρία το $w_k=(i,j)$ όπου οι δείκτες i και j αντιστοιχούν στα X και Y αντίστοιχα, έτσι ώστε:

$$\max(|X|, |Y|) \leq K \leq |X| + |Y|$$

Το πρώτο βήμα είναι η κατασκευή ενός $n \times m$ πίνακα απόστασης D . Κάθε στοιχείο του $D(i,j)$ είναι η Ευκλείδεια απόσταση μεταξύ των $X(i)$ και $Y(j)$ πολυδιάστατων σημείων. Στη συνέχεια, πρέπει να βρεθεί η διαδρομή στρέβλωσης (warping path, WP). Για να βρεθεί η καλύτερη αντιστοιχία μεταξύ των δύο ακολουθιών, μπορούμε να βρούμε μια διαδρομή μέσω του πίνακα D , η οποία ελαχιστοποιεί τη συνολική απόσταση μεταξύ τους. Επομένως, ορίζουμε την WP ως την αλληλουχία των επιμέρους στοιχείων του πίνακα D με το ελάχιστο άθροισμα αποστάσεων, με την προϋπόθεση ότι $\max(P1,P2) \leq \text{length}(WP)$. Το άθροισμα των αποστάσεων από τα καλύτερα WP ονομάζεται βαθμός αντιστοίχισης.

Για παράδειγμα έστω ότι έχουμε το σήμα A και το σήμα B που έχουν μήκος m και n αντίστοιχα, τα τοποθετούμε όπως φαίνεται στο παρακάτω σχήμα και διατάσσονται, έτσι ώστε προκύπτει ο πίνακας αποστάσεων. Η έναρξη είναι στην περίπτωση μας το $w_1=(1,1)$ και η αφετηρία το (n,m) .



Εικόνα 27: Διόρθωση ολίσθησης χρόνου ανάσχεσης με τη μέθοδο της δυναμικής χρονικής στρέβλωσης

Ιδανικά (για δυο διαστάσεις) η τέλεια ευθυγράμμιση των δυο σημάτων πραγματοποιείται όταν το μονοπάτι στρέβλωσης (κόκκινη γραμμή στην εικόνα 27) είναι η διαγώνιος του πίνακα. Οι οριζόντιες και κατακόρυφες κινήσεις είναι ισοδύναμες με μεταβάσεις (gaps ή transitions).

Το γεγονός ότι ορίστηκε καθορισμένο σημείο εκκίνησης και τελικό σημείο αποτελεί μια περιοριστική συνθήκη. Οι περιοριστικές συνθήκες τίθενται με σκοπό την επιτάχυνση της απόδοσης του αλγορίθμου. Επιπλέον περιορισμοί που μπορούν να τεθούν κατά την εξαγωγή του μονοπατιού στρέβλωσης είναι ότι οι δύο δείκτες i, j αυξάνουν κατά μια μονάδα σε κάθε βήμα κατά μήκος της διαδρομής (συνθήκη συνέχειας), ότι οι δείκτες ακολουθούν ένα μονοτονικό τρόπο, με την έννοια του ότι είτε αυξάνουν είτε παραμένουν σταθεροί (συνθήκη μονοτονικότητας). Τέλος, παίρνοντας υπόψη το πλεονέκτημα της πληροφορίας ότι η διαδρομή είναι απίθανο να εκτραπεί πολύ μακριά από τη διαγώνιο, μπορεί να υιοθετηθεί ένα παράθυρο μήκους r , το οποίο να επιτρέπει στην WP να συμπεριλαμβάνει αποστάσεις που βρίσκονται εντός της ακτίνας r (συνθήκη προσαρμογής παραθύρου).

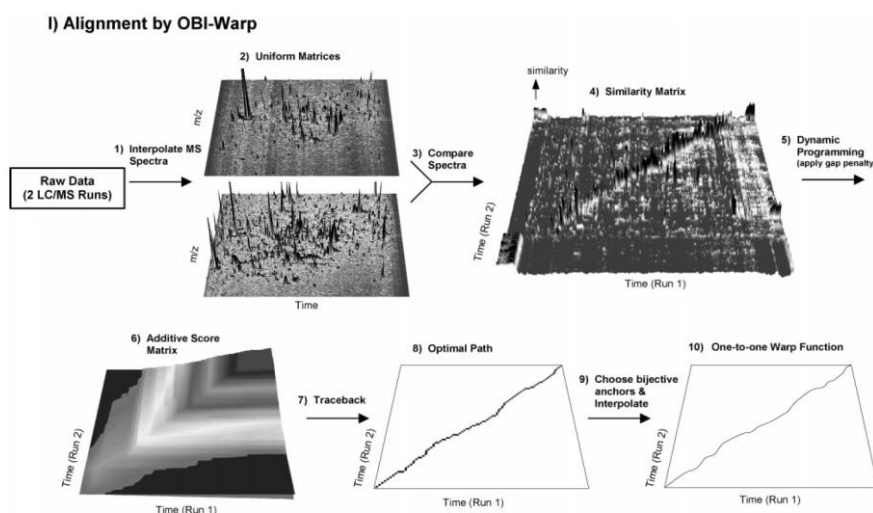
Λαμβάνοντας υπόψη τα παραπάνω, η μεταξύ δύο διανυσμάτων απόσταση υπολογίζονται ως:

$$WP(i, j) = \min[S(i-1, j-1), S(i-1, j), S(i, j-1)] + D(i, j)$$

όπου $S(i, j)$ είναι η αθροιστική απόσταση μεταξύ των X και Y συνόλων διανυσμάτων και ο βαθμός αντιστοίχισης (Matching Score, T) που υπολογίζεται ως η αθροισμένη ομοιότητα (S) μεταξύ των σημείων x_i και y_i στο k στοιχείο της διαδρομής στρέβλωσης (w_{ki}, w_{kj}) ως

$$T(W) = \sum_{k=1}^{k=K} S(w_{ki}, w_{kj})$$

Η έξοδος του DTW συνδυάζεται με μια ένα προς ένα πολυωνυμική συνάρτηση παρεμβολής Hermite κυβικού βαθμού ώστε να εξομαλυνθεί διαδρομή στρέβλωσης [44, 45].



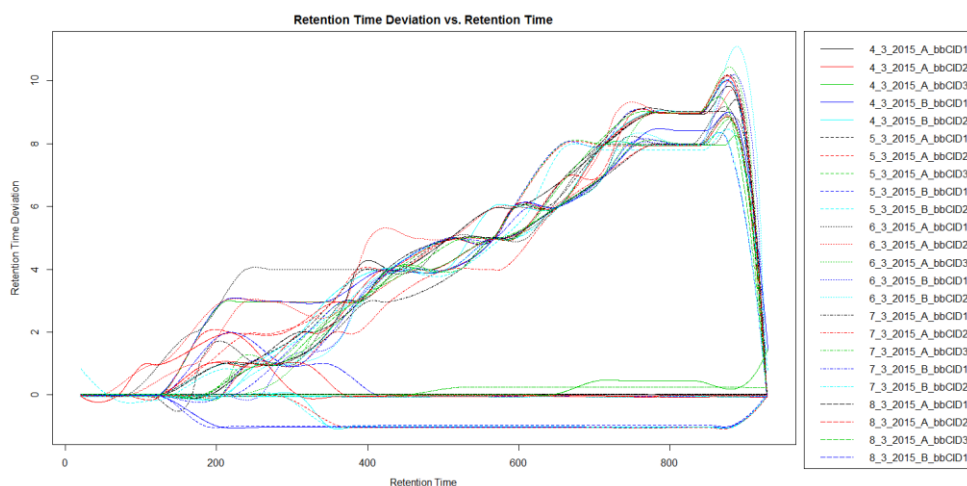
Εικόνα 28: Διάγραμμα ροής που δείχνει την χρωματογραφική ευθυγράμμιση με τον OBI-Warp αλγόριθμο, πηγή [44]

2.7.3.2 Αλγόριθμος loess

Σε αντίθεση με άλλες τεχνικές διόρθωσης της ολίσθησης του χρόνου ανάσχεσης η τεχνική loess διορθώνει το χρόνο ανάσχεσης σε όλα τα δείγματα σε ένα βήμα. Από το πρώτο βήμα της ομαδοποίησης των ιόντων ο αλγόριθμος έχουν ανιχνεύσει εκατοντάδες «καλά συμπεριφερόμενες» ομάδες κορυφών. Με

τον όρο «καλά συμπεριφερόμενες» εννοούνται οι κορυφές οι οποίες αντιπροσωπεύουν τον ίδιο αναλύτη και έχουν ανιχνευθεί σε όλα ή σχεδόν σε όλα τα δείγματα. Τέτοιες ομάδες κορυφών μπορούν να χρησιμοποιηθούν σαν πρότυπες ουσίες για τον καθορισμό και τη διόρθωση του χρόνου. Για κάθε τέτοια ομάδα υπολογίζεται ο μέσος χρόνος ανάσχεσης και η απόκλιση από το μέσο όρο για κάθε δείγμα εντός της ομάδας. Επειδή οι «καλά συμπεριφερόμενες» ομάδες κατανέμονται τις περισσότερες φορές ομοιόμορφα καθ' όλη τη διάρκεια του χρωματογραφήματος μπορεί να κατασκευαστεί ένα μη γραμμικό περίγραμμα απόκλισης για κάθε δείγμα.

Η μέθοδος αυτή εφαρμόζει τη μέθοδο της τοπικής παλινδρόμησης (Local Regression Fitting Method- loess), η οποία χρησιμοποιεί κατά διαστήματα πολυώνυμα χαμηλής τάξης για να προσαρμόσει τα δεδομένα. Ακόμα και στην περίπτωση που υπάρχουν περιοχές όπου δεν υπάρχουν τέτοιες «καλά συμπεριφερόμενες» ομάδες κορυφών τότε χρησιμοποιείται μια μαθηματική συνάρτηση για να προσεγγιστούν οι διαφορές στη χρονική απόκλιση. Με άλλα λόγια εφαρμόζεται η μέθοδος της παρεμβολής για να καλυφθούν τέτοιες περιοχές. Ωστόσο στο τέλος των χρωματογραφημάτων αν δεν υπάρχουν ομάδες «καλά συμπεριφερόμενων» κορυφών, τότε η συνάρτηση απόκλισης πεπλατίζεται σε μια σταθερή τιμή. Οι αποκλίσεις που παρατηρούνται χρησιμοποιούνται για να διορθωθούν οι χρόνοι ανάσχεσης των ιόντων. Επιπλέον, η μέθοδος loess παρέχει αυτόματη απομάκρυνση ακραίων υπολοίπων δηλαδή έκτροπων τιμών από τα δεδομένα και έτσι παρέχει ανθεκτικότητα (robustness). Οι αποκλίσεις αναπαρίστανται γραφικά γεγονός που επιτρέπει την εποπτεία του αλγόριθμου. Αποτελεί μειονέκτημα του αλγορίθμου το γεγονός ότι εξαρτάται στην διαδικασία εύρεσης κορυφών σε αντίθεση με τον OBI-Warp [33].



Εικόνα 29: Παράδειγμα εποπτικής εικόνας της διόρθωσης χρόνου ανάσχεσης από τον αλγόριθμο loess

2.7.4 Επανομαδοποίηση

Μετά τη διόρθωση των ολισθήσεων του χρόνου ανάσχεσης ο χρόνος ανάσχεσης μερικών ιόντων έχει αλλάξει ακόμα και λίγο για αυτό πραγματοποιείται επανομαδοποίηση και η πρώτη ομαδοποίηση καθίσταται μη έγκυρη. Συνήθως η δεύτερη ομαδοποίηση πραγματοποιείται με ίδιες ή λίγο πιο αυστηρές παραμέτρους. Ο τρόπος πραγματοποίησης της επανομαδοποίησης των ιόντων είναι ακριβώς με τις ίδιες εντολές όπως και η πρώτη ομαδοποίηση.

2.7.5 Γέμισμα τιμών με μικρή ένταση για κορυφές που δεν υπάρχουν σε κάποια δείγματα

Ο ομαδοποιημένος πίνακας περιέχει ελλιπούσες τιμές (Missing values) στα δείγματα στα οποία κάποιες κορυφές δεν έχουν ανιχνευθεί (είτε λόγω λάθους από τον αλγόριθμο εύρεσης κορυφών είτε επειδή απλά δεν βρέθηκαν επειδή δεν υπάρχει αναλύτης) ενώ στα υπόλοιπα δείγματα έχουν βρεθεί. Αυτές οι τιμές επειδή θα δημιουργήσουν πρόβλημα στην ακόλουθη στατιστική ανάλυση πρέπει να αντικατασταθούν με κάποια μικρή τιμή έντασης. Ο αλγόριθμος πάει στην περιοχή (εύρος χρόνου ανάσχεσης) όπου έχουν ανιχνευθεί κορυφές στα υπόλοιπα δείγματα και ολοκληρώνει την περιοχή. Έτσι αποτρέπεται ο τελικός πίνακας να περιέχει τιμές που δεν είναι αριθμοί γεγονός που δεν διευκολύνει την στατιστική ανάλυση και επιπλέον διορθώνονται πιθανές παραλήψεις του αλγόριθμου εύρεσης ιόντων [33, 34].

2.8 CAMERA

Το πακέτο της R CAMERA ενσωματώνει αλγόριθμους για να εξαχθούν τα φάσματα των συστατικών, να βρεθούν οι ισοτοπικές κορυφές και τα προϊόντα προσθήκης και προτείνει την ακριβή μάζα του συστατικού ακόμα και σε εξαιρετικά πολύπλοκα δεδομένα. Επιπλέον, υπάρχει μια δυνατότητα που συνδυάζει φασματικές πληροφορίες των δεδομένων σε δυο αντίθετες καταστάσεις ιοντισμού, ώστε να βελτιωθεί η επεξήγηση (annotation) των ιόντων. Κατά τον ιοντισμό, ένα χημικό συστατικό δημιουργεί ένα ή περισσότερα είδη ιόντων, τα οποία μπορούν να παρατηρηθούν από κάποιο φασματόμετρο. Αυτά τα είδη ιόντων συμπεριλαμβάνουν ισοτοπικές κορυφές, θραύσματα και εφόσον έχουμε ως πηγή ιοντισμού ηλεκτροψεκασμό περιλαμβάνει προϊόντα προσθήκης.

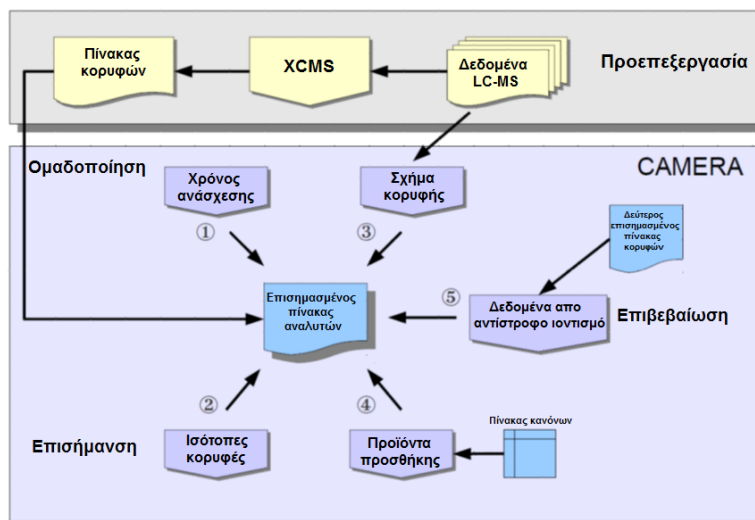
Δυο επιπλέον βήματα επεξεργασίας είναι επιθυμητά στην ανάλυση LC-MS δεδομένων:

1. Ομαδοποίηση όλων των χαρακτηριστικών που προέρχονται από τον ίδιο αναλύτη
2. Ο σχολιασμός-επισήμανση των διαφόρων ειδών ιόντων.

Το πρώτο βήμα από μόνο του επιτυγχάνει και μείωση των ιόντων και έναν πρώτο υπολογισμό του αριθμού των ενώσεων που ανιχνεύονται με την ανάλυση. Μια τέτοια εκτίμηση μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση του αναλυτικού πρωτοκόλλου, όπως έκαναν οι Yanes et al. όπου οι συγγραφείς χρησιμοποίησαν τον αριθμό των ιόντων ως κριτήριο βελτιστοποίησης αναλυτικής μεθόδου [46]. Και τα δυο βήματα μαζί μπορούν να αποκαλύψουν οιονεί μοριακά ιόντα, των οποίων ο σχολιασμός είναι απαραίτητος για τη μετέπειτα ανίχνευση του μεταβολίτη, όπως ο στοιχειακός υπολογισμός βασισμένος στην ακριβή μάζα και το ισοτοπικό προφίλ και τα φάσματα MS/MS.

Το ACD/IntelliXtract είναι ένα εμπορικό πακέτο που κατηγοριοποιεί τα ιόντα βασισμένο στον χρόνο ανάσχεσης και τον σχολιασμό των ειδών των ιόντων με έναν δεδομένο πίνακα κανόνων. Αντίστοιχα δημιουργήθηκε το ελεύθερο πακέτο CAMERA (εξέλιξη του πακέτου ESI), το οποίο ενσωματώνει πολλαπλές μεθόδους για ομαδοποίηση των ιόντων που συσχετίζονται χρησιμοποιεί ένα δυναμικό πίνακα κανόνων για την επισήμανση των ειδών των ιόντων.

Η ροή εργασιών για την ανάλυση με την CAMERA φαίνεται στην ακόλουθη εικόνα:



Εικόνα 30: Η ροή εργασίας της CAMERA για ανάλυση LC/MS δεδομένων, πηγή [47]

Στην παραπάνω εικόνα οι αριθμοί 1 έως 5 επεξηγούν μια τυπική σειρά. Τα πρωτογενή αρχεία επεξεργάζονται με το ΧCMS (άνω τμήμα) και τα ιόντα που προκύπτουν ως αποτέλεσμα της επεξεργασίας περνάνε στην CAMERA. Η ομαδοποίηση των ιόντων συμπεριλαμβάνει υπόψη της το χρόνο ανάλυσης (1), και το σχήμα των χρωματογραφικών κορυφών των ιόντων (3). Επιπλέον, ανιχνεύονται τα ιόντα που είναι ισοτοπικές κορυφές (2) και προϊόντα προσθήκης (4) μέσω της χρήσης δυναμικού πίνακα κανόνων.

Στο ακόλουθο τμήμα θα αναλυθούν τα παραπάνω πέντε βήματα της τυπικής πορείας του πακέτου CAMERA με περισσότερη λεπτομέρεια:

1. Δημιουργία των φασμάτων των συστατικών με βάση το χρόνο ανάλυσης

Η αρχική δημιουργία των φασμάτων των συστατικών πρέπει να είναι γρήγορη, ώστε να επεξεργαστούν δεκάδες έως εκατοντάδες δείγματα με χιλιάδες ιόντα. Επιλέγονται τα υψηλότερης έντασης ιόντα από τον πίνακα ιόντων που δεν έχει ακόμα αποδοθεί ένα φάσμα και υπολογίζεται ένα συγκεκριμένο παράθυρο χρόνων ανάλυσης, τυπικά 60% του FWHM της χρωματογραφικής κορυφής. Όλα τα ιόντα εντός αυτού του εύρους στη συνέχεια περιλαμβάνονται σε ένα φάσμα ενός συστατικού. Αυτό το βήμα επαναλαμβάνεται μέχρι όλα τα ιόντα να ανατεθούν σε φάσματα ουσών. Τα υψηλότερης έντασης ιόντα συνήθως έχουν

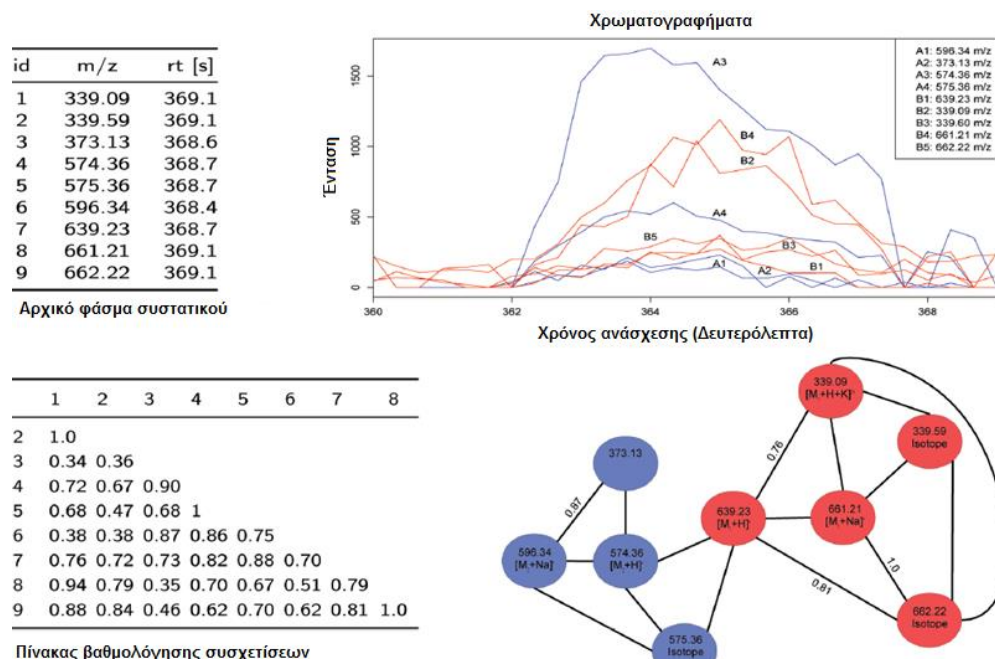
υψηλό λόγω S/N και συχνά παρέχουν την ακριβέστερη εκτίμηση του παραθύρου του χρόνου ανάλυσης.

2. Ανίχνευση των ισοτοπικών κορυφών και υπολογισμό του φορτίου

Η ανίχνευση των ισοτοπικών προφίλ απαιτείται για να εξαχθούν τα φορτία των ιόντων. Ο τρόπος αναγνώρισης των ισοτοπικών κορυφών είναι ο εξής: Εντός ενός φάσματος ενός συστατικού υπολογίζουμε ανά ζεύγη την απόσταση m/z και ανιχνεύουμε ισότοπα τα οποία παρουσιάζουν διαφορά $1,0033/z$ και επίσης περνάνε ένα εξτρά έλεγχο έντασης.

3. Βελτίωση του φάσματος των συστατικών

Ανάλογα με το χρωματογραφικό διαχωρισμό, το φάσμα της προκύπτουσας ένωσης ίσως περιλαμβάνει ιόντα δυο ή περισσότερων συνεκλουόμενων ουσιών. Χρησιμοποιείται ένας αλγόριθμος βασισμένος στα χρωματογραφήματα για την ενσωμάτωση τριών ή περισσότερων ενδείξεων για έναν βελτιωμένο διαχωρισμό όπως φαίνεται στο παράδειγμα της ακόλουθης εικόνας:



Εικόνα 31: Διαχωρισμός συνεκλουόμενων ιόντων δυο συστατικών από το πακέτο CAMERA, πηγή [47]

Στην εικόνα 31 παρουσιάζεται σχηματικά η ομαδοποίηση χαμηλής έντασης ιόντων που αρχικά ομαδοποιήθηκαν σύμφωνα με το χρόνο ανάλυσης σε ένα φάσμα μιας ουσίας ενώ πρόκειται για δυο ουσίες. Ο πίνακας πάνω αριστερά

απεικονίζει τα ιόντα, που αρχικά ομαδοποιήθηκαν με βάση το χρόνο ανάσχεσης. Πάνω δεξιά αναπαρίστανται τα χρωματογραφήματα των ιόντων με ταμπέλες A και B που ανταποκρίνονται στο αποτέλεσμα μετά την ομαδοποίηση. Ο πίνακας κάτω αριστερά είναι τα βάρη που χρησιμοποιούνται στη γραφική παράσταση. Τέλος κάτω δεξιά είναι το διάγραμμα σχέσης, όπου τα άκρα δείχνουν ένα κατόφλι βαθμολογίας. Οι ετικέτες του κόμβου περιλαμβάνουν το σχολιασμό των ιόντων, ενώ το χρώμα δείχνει τον διαχωρισμό μετά την εφαρμογή του αλγόριθμου.

Πρώτον χρησιμοποιούμε την ομοιότητα του σχήματος των κορυφών. Το πακέτο CAMERA χρησιμοποιεί τα πρωτογενή δεδομένα για να εξάγει τα χρωματογραφήματα για κάθε ιόν και υπολογίζει την κατά Pearson συσχέτιση των εντάσεων μεταξύ των χρωματογραφικών κορυφών (κινείται από σημείο σε σημείο) για όλα τα ζεύγη των ιόντων σε ένα φάσμα ενός συστατικού. Δεύτερον, περιλαμβάνουμε την κατά Pearson συσχέτιση των εντάσεων σε όλα τα δείγματα για κάθε ζεύγος ιόντων για κάθε φάσμα ενός συστατικού. Τέλος κωδικοποιούμε τη σχέση μεταξύ δυο ισοτόπων μεταξύ των ιόντων που ανιχνεύθηκαν στο βήμα 2. Για την εξαγωγή μιας βαθμολογίας (score) συνδυάζονται τρεις τιμές όπως φαίνεται στην εξίσωση:

$$\text{Score}(x,y) = \text{CAS}_{xy} + \text{ISO}_{xy} + \frac{1}{N} \sum_{i=1}^N \text{CPS}_{ixy}$$

Η βαθμολογία αντιπροσωπεύει τη σχέση μεταξύ δυο ιόντων x και y σε συνδυασμό με

- την συσχέτιση της έντασης μεταξύ των δειγμάτων (Correlation Across samples-CAS) για αυτά τα δυο ιόντα,
- τη δυαδική κωδικοποίηση παρουσίας ή απουσίας ισοτοπικής σχέσης και
- τη συσχέτιση του σχήματος της κορυφής όπως υπολογίζεται για κάθε ένα i δείγμα (Correlation Peak Shape-CPS_i).

Αρχικά αναπαρίστανται σε ένα γράφημα όλα τα ιόντα για ένα φάσμα ενός συστατικού. Το γράφημα συμπεριλαμβάνει ιόντα από δυο ή περισσότερα συστατικά που εκλούνται πολύ κοντά. Έπειτα χρησιμοποιείται είτε ο αλγόριθμος “Highly-connected-subgraphs” (HCS20) από το πακέτο της R RBGL είτε ο “Label Propagation Community” (LPC21) από το πακέτο igraph. Μετά την ομαδοποίηση των συστατικών του γραφήματος, το αρχικό φάσμα του

συστατικού διαιρείται σε δύο φάσματα για κάθε συστατικό (στην περίπτωση που έχουμε δυο συστατικά που συνεκλούνται).

4. Επισήμανση των προϊόντων προσθήκης, κοινών ουδέτερων απωλειών
Για το πακέτο ESI, τα μη φορτισμένα συστατικά ιοντίζονται μέσω των προϊόντων προσθήκης με κατιόντα ή ανιόντα ή με αφαίρεση πρωτονίων. Επιπλέον, οι ουδέτερες απώλειες που συμβαίνουν οδηγούν στον σχηματισμό θραυσμάτων. Η επισήμανση αυτών των ιόντων μειώνει τον αριθμό των ιόντων που έχουν εξεταστεί για την περαιτέρω ανάλυση. Από τουλάχιστον δυο επισημασμένα-σχολιασμένα ιόντα, μπορεί να υπολογιστεί το μοριακό βάρος, το οποίο είναι απαραίτητο για να υπολογιστεί η στοιχειακή σύνθεση μιας ουδέτερης ένωσης.

Το πακέτο CAMERA χρησιμοποιεί ένα σετ δυναμικών κανόνων, το οποίο δημιουργείται από ένα συνδυασμό από λίστες των παρατηρούμενων ιόντων. Κάθε κανόνας περιγράφει ένα συγκεκριμένο είδος ιόντος με διαφορά μάζας τη μοριακή μάζα, το φορτίο του ιόντος και τον αριθμό των μορίων που περιέχει ένα είδος ιόντος. Όλες οι διαφορές m/z εντός ενός φάσματος ενός συστατικού συγκρίνονται σύμφωνα με το σετ των δυναμικών κανόνων. Οι ταυτίσεις με την ίδιο μοριακή μάζα (κάτω από δεδομένο σχετικό σφάλμα ακρίβειας μάζας) συνδυάζονται με ομάδες υποθέσεων. Εάν δεν υπάρχουν κορυφές που μπορούν να εξηγηθούν με βάση τους κανόνες τότε καμία επισήμανση ιόντων δεν είναι δυνατή. Το CAMERA δεν χρησιμοποιεί ευρετικούς κανόνες όπως το να υποθέτει ότι το πιο έντονο ιόν σε ένα φάσμα είναι το $[M+H]^+$ ιόν (για θετικό ιοντισμό).

5. Συνδυάζοντας δεδομένα από διαφορετικές καταστάσεις ιοντισμού

Τα δείγματα που εκχυλίζονται συχνά μετρούνται και σε θετικό και σε αρνητικό ιοντισμό για να αυξηθεί η κάλυψη των αναλυτών. Παρά το ότι μερικά στοιχεία ιοντίζονται σε έναν μόνο ιοντισμό, αρκετά συστατικά είναι ανιχνεύσιμα και στους δυο. Σε αυτές τις περιπτώσεις τα συμπληρωματικά ιόντα μας παρέχουν περαιτέρω ένδειξη για το οιονεί μοριακό ιόν.

Το πακέτο CAMERA περιλαμβάνει έναν καινοτόμο αλγόριθμο επαλήθευσης και χρησιμοποιεί φάσματα που μετρούνται και στους δύο ιοντισμούς. Ο αλγόριθμος υπολογίζει τη διαφορά m/z για όλα τα ιόντα του σχετιζόμενου χαρακτηριστικού

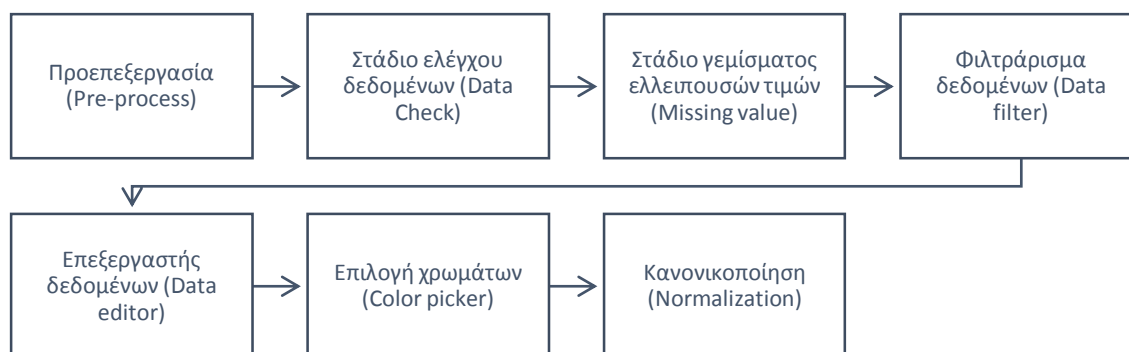
και από τους δυο ιοντισμούς εντός ενός εύρους παραθύρου του χρόνου ανάσχεσης. Αυτές οι διαφορές συγκρίνονται με ένα δεύτερο αντίθετου ιοντισμού πίνακα κανόνων (cross-polarity rule table). Αν ένας κανόνας ταιριάζει, αυτό είτε θα επισημάνει δυο προηγούμενα μη επισημασμένα ιόντα για παράδειγμα $[M+H]^+$ και $[M-H]^-$, ή θα επιβεβαιώσει ή θα έρθει σε σύγκρουση με έναν ήδη υπάρχοντα σχολιασμό. Στην τελευταία περίπτωση ο υπάρχων σχολιασμός αντικαθίσταται. Ο πίνακας κανόνων αντίθετης πολικότητας πρέπει να περιέχει κοινούς και αξιόπιστους συνδυασμούς, διότι αυτοί οι κανόνες θα γράψουν από επάνω τους σχολιασμούς που προέκυψαν από την ανάλυση των δεδομένων της μιας πόλωσης [47].

2.9 Metaboanalyst

Το MetaboAnalyst είναι ένα εργαλείο στατιστικής ανάλυσης που πρωτοεμφανίστηκε το 2009 [48] με σκοπό την ανάλυση δεδομένων μεταβολομικής ενώ αναβαθμίστηκε δυο φορές σε MetaboAnalyst2 το 2012 [49] και σε Metaboanalyst3 το 2015 [50]. Υπάρχει μια συνεχή προσπάθεια αναβάθμισης του Metaboanalyst ώστε το εργαλείο αυτό να συνεχίσει να εκπληρώνει τους σκοπούς για τους οποίους πρωτοδημιουργήθηκε. Κατά βάση πρόκειται για μια εργαλείο υπο μορφή ιστοσελίδας (<http://www.metaboanalyst.ca>) της οποίας οι κώδικες είναι διαθέσιμοι και επομένως μπορεί να εγκατασταθεί και σε τοπικό υπολογιστή για να εξυπηρετήσει τις ενδεχομένως αυξημένες ανάγκες μιας ερευνητικής ομάδας.

Πλέον η πλατφόρμα ενσωματώνει επιλογές για δείγματα ελέγχου ποιότητας (QC), επιτρέπει ανάλυση δειγμάτων που βρίσκονται κατανεμημένα στο χρόνο (χρονοσειρές), υποστηρίζει στατιστική ανάλυση περίπλοκων πειραματικών σχεδιασμών και παρέχει τα αποτελέσματα των χημειομετρικών αναλύσεων σε εικόνες υψηλής ποιότητας. Επιπλέον η πλατφόρμα έχει απλοποιηθεί ικανοποιητικά ώστε να εξυπηρετήσει απλούς χρήστες με περιορισμένες στατιστικές γνώσεις. Τέλος η πλατφόρμα πλέον έχει μεταφερθεί σε πολυπύρηνο υπολογιστή και οι αλγόριθμοι έχουν βελτιστοποιηθεί ώστε να εκμεταλλεύονται όλη την υπολογιστική δύναμη και ως εκ τούτου ο χρόνος ανάλυσης έχει επιταχυνθεί.

Σε σύνοψη οι χρήστες ανεβάζουν τα δεδομένα τους υπό μορφή csv με την απαραίτητη επισήμανση και την κατάλληλη μορφή (στάδιο προεπεξεργασίας - pre-process step). Το metaboanalyst απαιτεί το csv να έχει αυστηρή μορφή και πιθανές μικροπαραλήψεις μπορούν να αποτύχουν στο στάδιο ελέγχου των δεδομένων (data check step)



Εικόνα 32: Πορεία εισαγωγής δεδομένων στο Metaboanalyst

Αρχεία που περνάνε το στάδιο ελέγχου και έχουν ανιχνευθεί τιμές μη ευρεθείσες (NA values) περνάνε το στάδιο γεμίσματος αυτών των τιμών με κάποια μικρή τιμή. Υπάρχουν αρκετές τεχνικές για το σκοπό αυτό, μια από διαδεδομένες είναι το γέμισμα των κενών τιμών με την μικρότερη παρατηρηθείσα τιμή δια δυο.

Ο στόχος του φιλτραρίσματος δεδομένων (data filter step) είναι να εντοπιστούν και να εξαιρεθούν μεταβλητές που είναι απίθανο να χρησιμοποιηθούν κατά τις στατιστικές δοκιμασίες. Αυτό το βήμα προτείνεται για δεδομένα από μη στοχευμένα πειράματα που έχουν μεγάλο αριθμό μεταβλητών, πολλές από τις οποίες προκύπτουν από θόρυβο της γραμμής βάσης. Αυτό το στάδιο βελτιώνει πολύ τα αποτελέσματα και μειώνει σημαντικά το λόγο ψευδώς θετικών ευρημάτων (False Discovery Rate). Μη πληροφοριακές μεταβλητές διακρίνονται σε δύο ομάδες τις μεταβλητές με πολύ μικρές τιμές που βρίσκονται με χρήση της μέσης τιμής ή της διάμεσης τιμής και μεταβλητές που παραμένουν σταθερές σε όλες τις πειραματικές συνθήκες και μπορούν να ανιχνευθούν μέσω της τυπικής απόκλισης (SD) ή του ενδοτεταρτημοριακού εύρους (IQR). Ακόμα μπορεί να χρησιμοποιηθεί για την σχετική τυπική απόκλιση (SD/mean).

Οι ακόλουθοι εμπειρικοί κανόνες χρησιμοποιούνται:

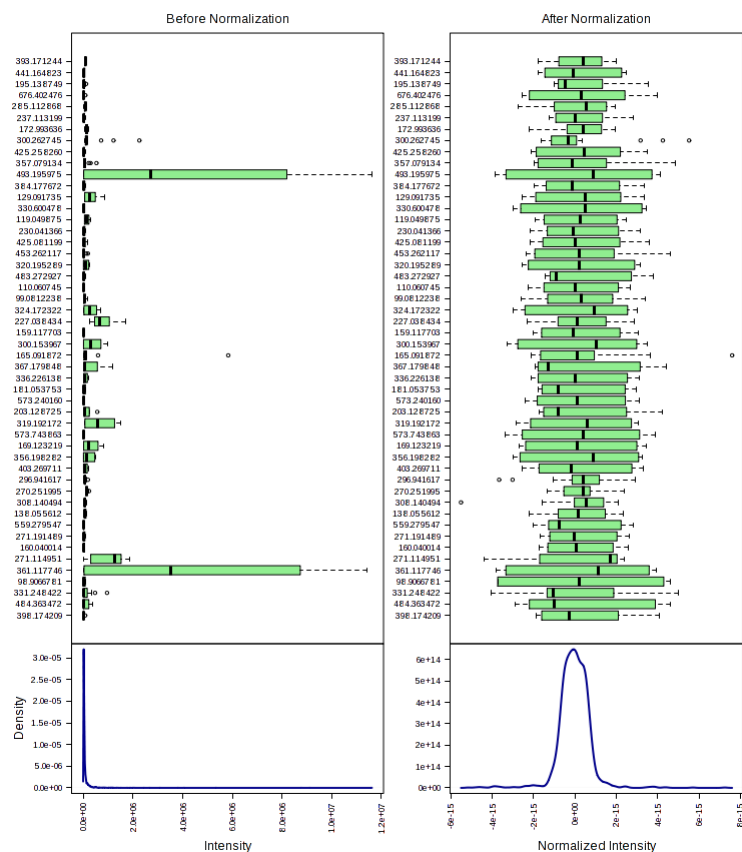
- Λιγότερες από 250 μεταβλητές τότε εκτελείται φιλτράρισμα 5%

- 250-500 μεταβλητές τότε εκτελείται φιλτράρισμα 10%
- 500-1000 μεταβλητές τότε εκτελείται φιλτράρισμα 25%
- Πάνω από 1000 μεταβλητές τότε εκτελείται φιλτράρισμα 40%

Αν μετά το φιλτράρισμα υπάρχουν παραπάνω από 5000 μεταβλητές τότε θα χρησιμοποιηθούν οι πρώτες 5000 και θα αγνοηθούν οι υπόλοιπες. Αυτό γίνεται για λόγους εξοικονόμησης υπολογιστικής ισχύς και δεν συμβαίνει για τοπικές εγκαταστάσεις του Metaboanalyst. Αυτό είναι και το κυριότερο πλεονέκτημα μιας τοπικής εγκατάστασης Metaboanalyst.

Ο επεξεργαστής δεδομένων (data editor) επιτρέπει την επεξεργασία των δεδομένων εντός πλατφόρμας και αποφεύγεται έτσι η έξοδος και η επανείσοδος στην πλατφόρμα για διόρθωση κάποιων τιμών. Ο επεξεργαστής δεδομένων μπορεί να χρησιμοποιηθεί καθ' όλη τη διάρκεια της ανάλυσης και χρησιμοποιείται για παράδειγμα για την αφαίρεση των έκτροπων τιμών.

Έπειτα ακολουθεί η επιλογή των χρωμάτων (υπάρχουν 5 προεπιλεγμένοι συνδυασμοί χρωμάτων) καθένα από τα οποία αναπαριστά μια ομάδα των μεταβλητών. Μετά ακολουθεί η κανονικοποίηση των δεδομένων ώστε αυτά να ακολουθήσουν όσο είναι δυνατόν την κανονική κατανομή που αποτελεί προϋπόθεση για αρκετές στατιστικές δοκιμασίες (t-tests, ANOVA). Η κανονικοποίηση βοηθάει στη μείωση του συστηματικού λάθους στα δεδομένα που εμφανίζονται από οργανολογικά προβλήματα ή από προβλήματα δειγματοληψίας. Υπάρχουν 11 τρόποι κανονικοποίησης που υποστηρίζονται από τον τρόπο παρουσίασης των κανονικοποιημένων δεδομένων, έτσι ώστε να επιλέξει ο χρήστης την καλύτερη δυνατή κανονικοποίηση για το δικό του σετ δεδομένων.



Εικόνα 33: Εργαλείο οπτικοποίησης της κανονικοποίησης του Metaboanalyst

Το MetaboAnalyst έχει τέσσερα είδη μονάδων στατιστικής ανάλυσης:

- Ανάλυση δεδομένων πολλαπλών ομάδων (multiple group data analysis)
- Ανάλυση δεδομένων δυο ομάδων/ανάλυση χρονοσειρών (binary group data analysis/time-series data analysis)
- Ανάλυση και επαύξηση μεταβολιτών (metabolite set enrichment analysis)
- Ανάλυση μεταβολικής οδού (metabolic pathway analysis)

Τα δυο τελευταία είδη αναλύσεων βασίστηκαν σε γνώση από την βιβλιογραφία που ειδικεύονται στην αναζήτηση βιβλιοθηκών μεταβολικής οδού (όπως SMPDB) σε πλούσιες βάσεις δεδομένων μεταβολιτών όπως η HMDB.

Ο σκοπός των τεσσάρων ειδών μονάδων στατιστικής ανάλυσης είναι η εύρεση σημαντικών συστατικών, η ομαδοποίηση και κατηγοριοποίηση, η ανάλυση χρονοσειρών ή η ανάλυση δεδομένων δυο ομάδων και η ανάλυση μεταβολικής οδού και ο εμπλουτισμός δεδομένων μεταβολομικής.

Υπάρχει μια σειρά διαφορετικών προσεγγίσεων που μπορούν να εφαρμοστούν για να βοηθήσουν τους ερευνητές στην ανίχνευση σημαντικών συστατικών

ενδιαφέροντος για να ικανοποιήσουν διαφορετικούς ερευνητικούς σκοπούς. Πρώτον, υπάρχει η μονάδα ανάλυσης διαφορικής έκφρασης η οποία μπορεί να επεκταθεί για να υποστηρίξει την πολλαπλή ανάλυση ομάδων που μπορεί να χρησιμοποιηθεί για να βρεθούν συστατικά τα οποία διαφέρουν μεταξύ δυο ομάδων. Υποστηρίζονται μια σειρά μονοπαραμετρικών τεχνικών όπως t-tests, ANOVA με εκ των υστέρων ανάλυση (post hoc analysis) και μετριάσμενες τεχνικές t στατιστικών δοκιμασιών όπως SAM ή EBAM για τη σύγκριση μέσω ωρων ή ενδιάμεσων τιμών μιας μεταβλητής σε δύο ή περισσότερες ομάδες δειγμάτων. Επειδή οι πολλαπλές στατιστικές δοκιμασίες οδηγούν σε υψηλό λόγο θετικά λανθασμένων ευρημάτων μπορεί να εφαρμοστεί η διόρθωση των p τιμών (Bonferroni p-τιμές). Δεύτερον υπάρχει η ανάλυση συνέκφρασης. Αυτή η μέθοδος είναι νέα στο Metaboanalyst και στοχεύει να βοηθήσει τους ερευνητές να βρουν συστατικά που μοιράζονται είτε όμοιες είτε αντίθετες αλλαγές συγκεντρώσεων σε διαφορετικές συνθήκες. Αυτές οι αλλαγές μπορούν να αναπαρασταθούν μέσω συσχετισμένων χαρτών θερμότητας (correlation heatmaps). Πολλές αποστάσεις ομοιότητας μπορούν να χρησιμοποιηθούν όπως η ευκλείδεια απόσταση.

Οι μέθοδοι ομαδοποίησης και συσταδοποίησης (Clustering και Classification) εφαρμόζονται σε πολλές μελέτες. Στο Metaboanalyst προσφέρονται κλασσικές πολυπαραμετρικές τεχνικές και νεότερες τεχνικές εκμάθησης (machine learning methods). Αφενός υπάρχουν οι κλασσικές πολυπαραμετρικές τεχνικές όπως η PCA ή η PLS-DA με επιπλέον επιλογές αντιμετάθεσης (permutation) αφετέρου υπάρχουν οι μη επιβλεπόμενες τεχνικές όπως ιεραρχική ομαδοποίηση με χάρτες θερμότητας, χάρτες αυτοοργάνωσης (self-organizing maps) και η k κοντινότερων γειτόνων ομαδοποίηση. Επιπλέον υπάρχουν τεχνικές επιβλεπόμενης εκμάθησης όπως SVM ή RF.

Η ανάλυση χρονοσειρών ή ανάλυση δεδομένων δυο ομάδων όπως είναι για παράδειγμα η μελέτη της επίδρασης της θεραπείας σε διαφορετικά χρονικά σημεία σε φυτά ή ζώα ή γενικά συγκρίσεις μεταξύ δυο ομάδων περιλαμβάνει τρεις διαφορετικές επιλογές. Πρώτον υπάρχουν οι διπλής κατεύθυνσης γραφικές αναπαραστάσεις όπως οι δισδιάστατοι χάρτες θερμότητας που αναπαριστούν και τα αποτελέσματα της ιεραρχικής ομαδοποίησης. Ακόμα μπορεί να πραγματοποιηθεί και τρισδιάστατη απεικόνιση των αποτελεσμάτων

της PCA. Και οι δυο προσεγγίσεις επιτρέπουν στους χρήστες να εξερευνήσουν με ευκολία τα προφίλ κατανομής των μεταβολιτών με διαφορετικές κατηγορικές μεταβλητές ή σε διαφορετικά χρονικά σημεία. Έπειτα διατίθεται ως επιλογή η ανάλυση δυο ομάδων που επιτρέπει να πραγματοποιηθεί η κλασσική σύγκριση μεταξύ δυο ομάδων ή μεταξύ πολλών ομάδων (ANOVA). Επιπλέον υποστηρίζεται η πολυμεταβλητή επέκταση της ANOVA που είναι γνωστή και ως ASCA. Τέλος υπάρχει και η πολυμεταβλητή χρονική απεικόνιση με τη χρήση μπεϋσιανής προσέγγισης για πειράματα με μικρό αριθμό χρονικών σημείων που υπάρχουν διαθέσιμες επαναλήψεις (replicates).

Τέλος οι ποσοτικές μεταβολομικές τεχνικές (όπως η μέτρηση απόλυτων συγκεντρώσεων αναλυτών) έχουν ανοίξει την πόρτα στη βελτίωση λειτουργικής ανάλυσης και βιολογικής ερμηνείας. Για αυτό έχουν προστεθεί η ανάλυση εμπλουτισμού μεταβολιτών (MSEA) που όμως περιορίζεται μόνο για ανθρώπινα δεδομένα, διότι οι βιβλιοθήκες μεταβολιτών προέρχονται από βάσεις δεδομένων που περιέχουν μόνο ανθρώπινους μεταβολίτες, ενώ η εύρεση μεταβολικής οδού MetPA περιέχει 16 διαφορετικά μοντέλα για 16 διαφορετικούς οργανισμούς [49, 50].

2.10 Άλλα διαθέσιμα εργαλεία για μη στοχευμένη ανάλυση

Υπάρχουν και άλλα λογισμικά τα οποία μπορούν να χρησιμοποιηθούν για την επεξεργασία δεδομένων LC-HRMS τα οποία είναι αδύνατον να τα ανασκοπίσουμε όλα στα πλαίσια της διπλωματικής εργασίας. Παρόλα αυτά μπορούμε να αναφέρουμε μερικά από αυτά ώστε ο μελλοντικός αναγνώστης να αναζητήσει περισσότερες πληροφορίες για αυτά αν επιθυμεί.

Αρχικά υπάρχουν αρκετά λογισμικά για την ανίχνευση ιόντων και για ευθυγράμμιση-διόρθωση χρόνου ανασχεσης των δεδομένων που έχουν αναπτυχθεί κατά τα τελευταία χρόνια, τόσο εμπορικά προϊόντα όπως το MarkerLynx (Waters) [51], το ProfileAnalysis (Bruker) [21] το κλειστού κώδικα αλλά ελεύθερα διαθέσιμο MetAlign [52], όσο και ανοιχτού κώδικα όπως το Mzmine [53] ή το OpenMS [54].

Τέλος αξίζει να αναφερθεί το γεγονός της δημιουργίας εργαλείων που αναπτύσσονται σε γλώσσα R. Αυτά τα πακέτα φιλοξενούνται τόσο στον επίσημο ιστότοπο της R (<http://cran.r-project.org/web/packages/>) όσο και στη

τοποθεσία επιστημονικών πακέτων του Bioconductor [55]. Μέχρι στιγμής στον πρώτο ιστότοπο υπάρχουν διαθέσιμα 12 πακέτα επεξεργασίας δεδομένων φασματομετρίας μαζών και στην δεύτερη 50 άλλα πακέτα. Φυσικά από αυτά τα 62 πακέτα δεν σχετίζονται όλα τα πακέτα με ανάλυση δεδομένων LC-HRMS. Αντίθετα, κάποια πακέτα που αφορούν αεριοχρωματογραφία συζευγμένη με φασματομετρία μαζών, κάποια άλλα καλύπτουν εξειδικευμένες ανάγκες για παράδειγμα της πρωτεομικής έρευνας, και κάποια απλά περιέχουν σετ πειραματικών δεδομένων (κάποια αρχεία .mzML/netCDF,mzXML).

Κάποια από αυτά τα πακέτα που αξίζει να αναφερθούν είναι το Rdisop, το enviPick, το nontarget και το enviMass.

Το Rdisop έχει την ίδια λειτουργία με το SmartFormula, δηλαδή βρίσκει μοριακό τύπο δεδομένου του μοριακού ιόντος και του ισοτοπικού προφίλ (για περισσότερες πληροφορίες ενότητες 2.1.-2.2.) αλλά σε αυτόματο υπολογιστικά επίπεδο χωρίς να υπάρχει κάποιο γραφικό περιβάλλον [56].

Το πακέτο enviPick πραγματοποιεί ανίχνευση κορυφών σε HRMS δεδομένα όπως αναφέραμε στην ενότητα 2.7.1.6. και η λίστα κορυφών που έχει ανιχνεύσει μπορεί να επεξεργαστεί περαιτέρω με το πακέτο nontarget, το οποίο είναι ικανό να ομαδοποιήσει τα ιόντα που προέρχονται από έναν αναλύτη (εύρεση ιόντων προσθήκης) και βρίσκει τους αναλύτες οι οποίοι εμφανίζουν χαρακτηριστικό ισοτοπικό προφίλ, δηλαδή να βρει ενώσεις που περιέχουν Br, Cl ή και S [57].

Το πακέτο enviMass αποτελεί συνδυασμό των πακέτων enviPick, enviPat και μέρος του nontarget. Το enviPat χρησιμεύει στην εύρεση και επισημάνση των ισοτοπικών κορυφών. Το enviMass αποτελείται από ένα εύχρηστο γραφικό περιβάλλον, το οποίο επιτρέπει την εύρεση τάσεων τόσο στοχευμένων αναλυτών όσο μη στοχευμένων. Ο τρόπος κατάταξης των ιόντων είναι με βάση τη μεγαλύτερη ένταση που εμφανίζονται στο δείγμα. Τα τελευταία πακέτα έχουν χρησιμοποιηθεί κατά κόρον από μελέτες μη στοχευμένης έρευνας όπως θα δούμε στην παρακάτω ενότητα [10].

ΚΕΦΑΛΑΙΟ 3

Σκοπός της εργασίας

Στόχος της παρούσας εργασίας είναι η ανάπτυξη μιας ημιαυτόματης μεθοδολογίας η οποία δίνει τη δυνατότητα παρακολούθησης ουσιών, οι οποίες εισάγονται καθημερινά στα κέντρα επεξεργασίας λυμάτων και παρουσιάζουν αξιοσημείωτη διακύμανση στην ένταση μεταξύ των ημερών. Για τον σκοπό αυτό, πρώτα συλλέχθηκαν πληροφορίες για διάφορες μεθόδους οι οποίες επιτρέπουν την αυτοματοποιημένη επεξεργασία LC-MS¹ δεδομένων και έπειτα ανασκοπήθηκαν χημειομετρικά εργαλεία, όπως εργαλεία *in silico* θραυσματοποίησης χημικών δομών, εργαλεία απόδοσης στοιχειακής σύνθεσης με βάση το ισοτοπικό προφίλ και την πληροφορία της ακριβούς μάζας, βιβλιοθήκες φασμάτων MS/MS κ.α. Τα μεν εργαλεία αυτοματοποιημένης επεξεργασίας δεδομένων επιτρέπουν την εύρεση των επιθυμητών αναλυτών, δηλαδή αναλυτών με διακύμανση μεταξύ των ημερών που λήφθηκαν δείγματα, ενώ τα εργαλεία μη στοχευμένης ανάλυσης επιτρέπουν την αναγνώριση της ταυτότητας, δηλαδή της χημικής δομής των αναλυτών. Ακολούθως πραγματοποιήθηκε δειγματοληψία 8 συνεχόμενων ημερών από το κέντρο επεξεργασίας λυμάτων της Ψυττάλειας, πραγματοποιήθηκε χημική ανάλυση βασισμένη στην εκχύλιση στερεάς φάσης και τα δεδομένα αναλύθηκαν με την τεχνική LC-QTOFMS. Τα δεδομένα που αποκτήθηκαν υπέστησαν επεξεργασία με μια διαδικασία βασισμένη σε πακέτα της γλώσσας προγραμματισμού R (Xcms, CAMERA, TIMECOURSE). Οι παράμετροι εύρεσης κορυφών, ομαδοποίησης κορυφών και διόρθωσης του χρόνου ανάρχεσης βελτιστοποιήθηκαν με βάση επιφάνειες απόκρισης στηριζόμενοι στο πακέτο IPO. Οι μάζες που παρουσίασαν έντονη διακύμανση και επιλέχθηκαν από τον αλγόριθμο μελετήθηκαν περαιτέρω με τη χρήση των εργαλείων μη στοχευμένης ανάλυσης και αποκαλύφθηκε για 6 από αυτές η δομή τους.

ΚΕΦΑΛΑΙΟ 4

Μεθοδολογία

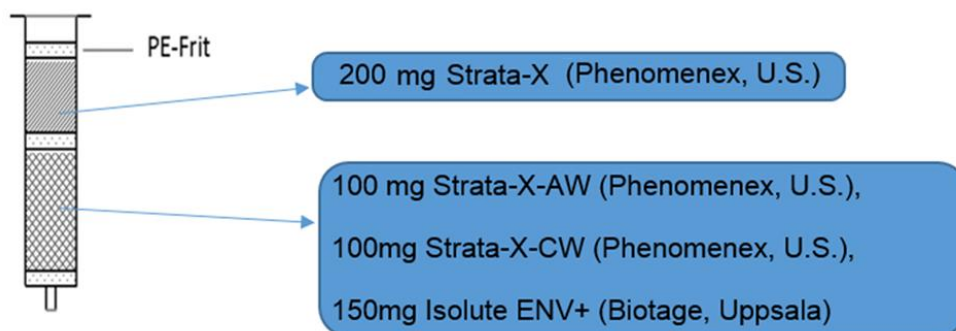
4.1 Δειγματοληψία

Πραγματοποιήθηκε δειγματοληψία εισερχόμενων λυμάτων για 8 συνεχόμενες ημέρες (4 Μαρτίου 2015 μέχρι και 11 Μαρτίου 2015). Τα δείγματα συλλέχθηκαν σε περιέκτες από τερεφθαλικό πολυαιθυλένιο (PET), οι οποίοι εκπλύθηκαν με μεθανόλη, υπερκάθαρο νερό και επισημάνθηκαν κατάλληλα προ της δειγματοληψίας. Τα ημερήσια δείγματα συλλέχθηκαν με τον πλέον αντιπροσωπευτικό τρόπο, καθώς πρόκειται για σύνθετη 24ωρη δειγματοληψία αναλογική με τη ροή των λυμάτων. Μετά τη λήψη των δειγμάτων μεταφέρθηκαν στο εργαστήριο σε χρόνο λιγότερο της μιας ώρας και πραγματοποιήθηκε επί τόπου προκατεργασία των δειγμάτων και πιο συγκεκριμένα φιλτράρισμα από ηθμό διαμέτρου 4mm με μέγεθος πόρων 0,2 μm (Phenomenex, Torrance, CA, USA) και το υγρό συλλέχθηκε σε φυγοκεντρικούς σωλήνες των 50mL. Ακολούθησε ρύθμιση του pH σε τιμή $6,5 \pm 0,2$ με προσθήκη μυρμηκικό οξύ 1M και έπειτα τα δείγματα αποθηκεύτηκαν στην κατάψυξη στους -20°C , ώστε να διατηρηθούν και να σταματήσει κάθε βιολογική δραστηριότητα. Μετά τη συλλογή όλων των ημερήσιων δειγμάτων πραγματοποιήθηκε η προκατεργασία τους με εκχύλιση στερεάς φάσης που παρουσιάζεται παρακάτω.

4.2 Προκατεργασία Δειγμάτων

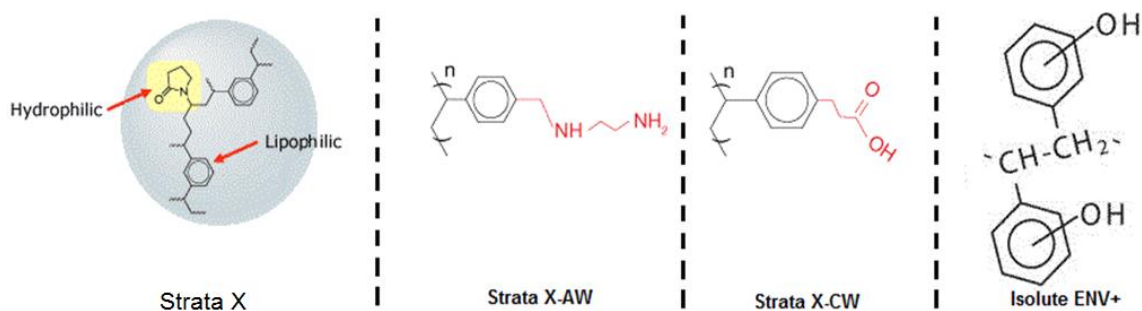
Η προκατεργασία των δειγμάτων βασίστηκε σε ένα γενικευμένο πρωτόκολλο εκχύλισης στερεάς φάσης ικανό να κατακρατάει μια μεγάλη γκάμα αναλυτών με διαφορετικά φυσικοχημικά χαρακτηριστικά [58].

Τα άδεια στηλάκια πολυπροπυλενίου χωρητικότητας 6mL που χρησιμοποιήθηκαν γεμίστηκαν με τεσσάρων ειδών προσροφητικά υλικά όπως φαίνεται στην ακόλουθη εικόνα:



Εικόνα 34: Δομή πολυπροσροφητικών στηλών SPE

Οι στιβάδες προσροφητικών υλικών διαχωρίζονται από πολυμερικά σκληρά υλικά με κατάλληλους πόρους (frits, διαμέτρου 20μm). Συνολικά δυο στιβάδες προσροφητικών υλικών, η μια με προσροφητικό Oasis HLB και η άλλη με μίγμα προσροφητικών Strata-X: Weak Cation eXchange, Weak Anion eXchange και Isolute ENV+ σε αναλογία 1:1:1,5.



Εικόνα 35: Χημική δομή στατικών φάσεων

Όπως φαίνεται και από τη δομή των στατικών φάσεων το Isolute ENV+ συγκρατεί τις ισχυρά πολικές ενώσεις που δεν συγκρατούνται από στατικές φάσεις C18 ή C8. Συγκρατεί τις πολικές ενώσεις κυρίως με δεσμούς υδρογόνου. Το OASIS HLB είναι ένα μίγμα δυο μονομερών της πολικής N-βινυλοπυρρολιδόνης και του άπολου διβινυλοβενζολίου. Εδώ χρησιμοποιείται για την κατακράτηση περισσότερο των άπολων ενώσεων. Συγκρατεί τις άπολες ενώσεις κυρίως εξαιτίας των υδρόφοβων αλληλεπιδράσεων που αναπτύσσονται μεταξύ αναλύτη και στατικής φάσης. Το Strata X-CW είναι ασθενής κατιονανταλλάκτης ενώ το Strata X-AW είναι ασθενής ανιονανταλλάκτης και δρουν διττά σχηματίζοντας π-π δεσμούς και υδρόφοβες αλληλεπιδράσεις με τους αναλύτες. Τέλος οι δυο τελευταίες πολυμερικές

φάσεις μπορούν να αναπτύσσουν και ηλεκτροστατικές αλληλεπιδράσεις και έτσι συγκρατούν και τις φορτισμένες ενώσεις.

Συνοπτικά το πρωτόκολλο εκχύλισης είναι το εξής:

1. Ενεργοποίηση των στερεών στατικών φάσεων με 5mL μεθανόλη και 10mL νερό
2. Φόρτωση των δειγμάτων
3. Δεν πραγματοποιείται έκπλυση
4. Ξήρανση υπο κενό για 1ώρα
5. Έκλουση με 4 ml μεθανόλης και οξικό αιθυλεστέρα σε αναλογία 1+1 που περιέχει 2% v/v αμμωνία ακολουθούμενη με 2 ml μεθανόλη και οξικό αιθυλεστέρα σε αναλογία 1+1 που περιέχει 1,7% v/v μυρμηκικό οξύ

Τα εκχυλίσματα που συλλέχθηκαν οδηγούνται σε εξάτμιση σχεδόν μέχρι ξηρού (100μL) με ήπιο ρεύμα αργού σε θερμοκρασία 30°C. Ακολούθησε ανασύσταση των δειγμάτων σε φιαλίδια (vial) με σύσταση νερό:μεθανόλη 1+1 και τελικό όγκο 500μL. Προ της προσθήκης σε φιαλίδια το εκχύλισμα όγκου 500μL φιλτράρονται από σύριγγες αναγεννημένης κυτταρίνης με διάμετρο πόρων 0,45μm.

4.3 Υγροχρωματογραφία-Φασματομετρία Μαζών

Χρησιμοποιήθηκε υπερυψηλής απόδοσης υγροχρωματογραφία (Ultrahigh performance liquid chromatograph-UHPLC) με αντλία HPG-3400 (DionexUltiMate 3000 RSLC, Thermo Fisher Scientific, Germany) διασυνδεδεμένη με φασματόμετρο μάζας με υβριδικό αναλυτή μαζών τετράπολο-αναλυτή χρόνου πτήσης QTOF (Maxis Impact, Bruker Daltonics, Bremen, Germany). Ο χρωματογραφικός διαχωρισμός πραγματοποιήθηκε σε στήλη αντίστροφης φάσης RSLC C18 Acclaim™ με μέγεθος σωματιδίων 2,2 μm και διαστάσεις 2.1×100 mm αγορασμένη από την εταιρεία Thermo Fisher Scientific (Driesch, Germany) εξοπλισμένη με προστήλη Acquity UPLC BEH C18 1,7 μm (Waters, Ireland). Οι στήλες θερμοστατούνται στους 30°C. Ο όγκος ένεσης δείγματος στο χρωματογραφικό σύστημα είναι 5μL.

Οι συνθήκες τόσο του φασματόμετρου μαζών, οι διαλύτες που χρησιμοποιήθηκαν ως κινητές φάσεις και το πρόγραμμα βαθμιδωτής έκλουσης που πραγματοποιήθηκε και σε θετικό και σε αρνητικό ιοντισμό αναπαρίστανται αναλυτικά στον ακόλουθο πίνακα:

Πίνακας 9: Συνθήκες ηλεκτροψεκασμού διαλύτες κινητής φάσης και πρόγραμμα βαθμιδωτής έκλουσης για θετικό και αρνητικό ιοντισμό

Θετικός Ιοντισμός			
Πρόγραμμα Έκλουσης		Παράμετροι Ηλεκτροψεκασμού	
Χρόνος (λεπτά)	% B	Capillary Voltage	2500V
0	1	End plate offset	500V
1	1	nebulizer	2 bar
3	39	Drying gas	8 L min ⁻¹
14	99,9	Drying temperature	200°C
16	99,9	(A) Νερό:Μεθανόλη 90:10 5mM μυρμηκικό αμμώνιο με 0,01% μυρμηκικό οξύ (B) Μεθανόλη 5mM μυρμηκικό αμμώνιο με 0,01% μυρμηκικό οξύ	
16.1	1		
20	1		
Το φασματόμετρο μαζών σταματάει την καταγραφή στα 15,5 λεπτά			
Αρνητικός Ιοντισμός			
Πρόγραμμα Έκλουσης		Παράμετροι Ηλεκτροψεκασμού	
Χρόνος (λεπτά)	% B	Capillary Voltage	3500 V
0	1	End plate offset	500 V
1	1	nebulizer	2 bar
3	39	Drying gas	8 L min ⁻¹
14	99,9	Drying temperature	200°C
16	99,9	(A) Νερό:Μεθανόλη 90:10 5mM οξικό αμμώνιο (B) Μεθανόλη 5mM οξικό αμμώνιο	
16.1	1		
20	1		
Το φασματόμετρο μαζών σταματάει την καταγραφή στα 15,5 λεπτά			

Το φασματόμετρο μαζών σύμφωνα με τον πειραματικό σχεδιασμό λειτουργήσε σε δυο βασικές λειτουργίες, την broadband collision- induced dissociation (bbCID) και την autoMS.

Η bbCID τεχνική επιτρέπει να καταγράφονται τα φάσματα MS σε εύρος μαζών από 50 έως 1000Da. Επιπλέον καταγράφονται και φάσματα MS/MS στα οποία όμως θραυματοποιούνται αδιακρίτως οι εκλούμενοι αναλύτες χωρίς να προηγηθεί απομόνωσή τους. Τα φάσματα MS λαμβάνονται σε χαμηλή ενέργεια θραυματοποίησης (4 eV) και τα φάσματα MS/MS σε υψηλή ενέργεια θραυματοποίησης (25 eV). Ο ρυθμός δειγματοληψίας φασμάτων είναι ίσος με 2Hz που σημαίνει ότι κάθε 1 δευτερόλεπτο παίρνεται ένα φάσμα MS και ένα bbCID. Τα δεδομένα που μετατράπηκαν σε mzXML και ακολούθησαν την προτεινόμενη μεθοδολογία ανεύρεσης τάσεων ήταν τα δεδομένα που πάρθηκαν σε bbCID λειτουργία, διότι κατά αυτήν την λειτουργία παίρνονται περισσότερα φάσματα MS από ότι η λειτουργία autoMS που περιγράφεται παρακάτω.

Η λειτουργία autoMS διαφέρει από την bbCID κατά το γεγονός ότι απομονώνει τα ιόντα τα οποία έπειτα θραυματοποιούνται και οδηγούμαστε στο φάσμα MS/MS συγκεκριμένων αναλυτών, ενώ τα φάσματα MS/MS του bbCID περιέχει θραύσματα από όλα τα συνεκλούμενα ιόντα. Στην παρούσα διπλωματική πάρθηκαν φάσματα MS/MS των πρώτων 5 σε αφθονία ιόντων σε κάθε φάσμα πλήρους σάρωσης. Με αυτόν τον τρόπο στο φάσμα MS παίρνεται ένα σημείο κάθε 2,5 δευτερόλεπτα ενώ με τη μέθοδο bbCID παίρνεται σημείο κάθε 1 δευτερόλεπτο.

Ως μέθοδος εξωτερικής βαθμονόμησης του αναλυτή χρόνου πτήσης κάθε μέρα πραγματοποιείται βαθμονόμηση με διάλυμα μυρμηκικού νατρίου ενώ στην αρχή κάθε χρωματογραφήματος γίνεται έγχυση του ίδιου διαλύματος ως εσωτερική βαθμονόμηση. Το διάλυμα βαθμονόμησης είναι μίγμα νερού ισοπροπανόλης 1+1 με 10mM μυρμηκικό νάτριο γεγονός που παράγει χαρακτηριστικές μάζες σε ένα εύρος μαζών 50-1000Da.

ΚΕΦΑΛΑΙΟ 5

Αποτελέσματα-Συζήτηση

5.1 Αλγόριθμοι και βελτιστοποίησή τους

Πακέτα και λογισμικά είναι απαραίτητα για την επεξεργασία μεγάλων σετ δεδομένων που προκύπτουν από τεχνικές υψηλής απόδοσης όπως η LC-HRMS. Παρόλα αυτά το αποτέλεσμα των αλγορίθμων που χρησιμοποιούν τα πακέτα εξαρτάται έντονα από τις παραμέτρους που τίθενται. Εάν δεν επιλεγούν με προσοχή μπορεί να οδηγήσουν σε διαστρεβλωμένα αποτελέσματα. Για αυτό οι παράμετροι αυτοί απαιτούν βελτιστοποίηση. Το πακέτο της R IPO (Isotopologue Parameter Optimization) επιτρέπει την βελτιστοποίηση των παραμέτρων που εισάγονται στον αλγόριθμο centWave για εύρεση κορυφών, των παραμέτρων που χρησιμοποιούνται για ομαδοποίηση των ιόντων και διόρθωσης των ολισθήσεων του χρόνου ανάσχεσης. Ο τρόπος με τον οποίο επιτυγχάνεται αυτό είναι χρησιμοποιώντας τις σταθερές ισοτοπικές ^{13}C κορυφές που σίγουρα υπάρχουν στα δείγματα για τον υπολογισμό τους. Η προσέγγιση για τη βελτιστοποίηση των παραμέτρων είναι η χρήση πειραματικού σχεδιασμού (Design of Experiments-DoE). Ο πειραματικός σχεδιασμός είναι μια σειρά από δοκιμασίες στις οποίες γίνονται συγκεκριμένες αλλαγές στις μεταβλητές που εισάγονται. Η μέθοδος στοχεύει στην εύρεση των αλλαγών που μεγιστοποιούν την απόκριση. Μετά από κάθε πείραμα DoE υπολογίζονται οι επιφάνειες απόκρισης.

Βελτιστοποιούνται οι παράμετροι που εισάγονται στην εύρεση κορυφών, στην ομαδοποίηση των ιόντων και στη διόρθωση του χρόνου ανάσχεσης. Επειδή η διαδικασία εύρεσης κορυφών πραγματοποιείται ξεχωριστά σε κάθε αρχείο αλλά όμως η διόρθωση του χρόνου ανάσχεσης και η ομαδοποίηση των ιόντων είναι αλληλοεξαρτώμενες διαδικασίες βελτιστοποιούνται με τη χρήση όλων μαζί των αρχείων. Για αυτό η βελτιστοποίηση των παραμέτρων πραγματοποιείται με τρόπο ημιδιαδοχικό. Οι παράμετροι του αλγόριθμου εύρεσης κορυφών βελτιστοποιούνται πρώτα και έπειτα βελτιστοποιούνται οι παράμετροι διόρθωσης χρόνου ανάσχεσης και ομαδοποίησης ταυτόχρονα. Η ταυτόχρονη βελτιστοποίηση είναι απαραίτητη διότι η ομαδοποίηση προαπαιτείται για την

αξιολόγηση της διόρθωσης του χρόνου ανάσχεσης, το οποίο στη συνέχεια με τη σειρά του μπορεί να βοηθήσει στη βελτίωση της ομαδοποίησης. Επιπλέον αυτός ο ημιδιαδοχικός τρόπος μειώνει και τον υπολογιστικό χρόνο που απαιτείται για τους αλγόριθμους.

Χρησιμοποιείται ο σχεδιασμός Box-Behnken (BBD) ως βάση του πειραματικού σχεδιασμού. Ο BBD σχεδιασμός είναι τριών επιπέδων ελλειπής παραγοντικός σχεδιασμός για την προσαρμογή μοντέλου απόκρισης επιφάνειας. Σχεδιασμός τριών επιπέδων υποδηλώνει ότι για κάθε παράμετρο υπάρχουν τρεις ισαπέχουσες παράμετροι που δοκιμάζονται. Οι δυο εξωτερικές τιμές ορίζουν το εύρος και η μεσαία τιμή παίζει το ρόλο της κεντρικής τιμής. Σε αντίθεση με τον πλήρη παραγοντικό σχεδιασμό δεν ελέγχει όλους τους συνδυασμούς και για αυτό είναι περισσότερο αποτελεσματικός και γρήγορος.

Για τον αλγόριθμο εύρεσης κορυφών centWave ο μαθηματικός τύπος της απόκρισης που χρησιμοποιείται δίνεται από

$$PPS = \frac{RP^2}{\text{All peaks-LIP}}$$

Όπου PPS είναι η βαθμολογία του αλγορίθμου εύρεσης κορυφών (Peak Picking Score), RP είναι ο αριθμός των αξιόπιστων κορυφών (Reliable Peaks), “All peaks” είναι όλες οι κορυφές και LIP (Low Intensity Peaks) είναι οι κορυφές χαμηλής έντασης. Ως αξιόπιστες κορυφές ορίζονται εκείνες για τις οποίες έχει βρεθεί το ιόν M+H (ή M-H για αρνητικό ιοντισμό) και επιπλέον έχει βρεθεί η αντίστοιχη ισοτοπική κορυφή λόγω ^{13}C .

Για τον αλγόριθμο διόρθωσης του χρόνου ανάσχεσης, ο τύπος απόκρισης RCS (Retention time score) είναι ο ακόλουθος:

$$RCS(x) = \left(\frac{\text{sum} \left(\frac{(\sum_{n=1}^k |median(x) - x_n|)}{k} \right)}{k} \right)^{-1}$$

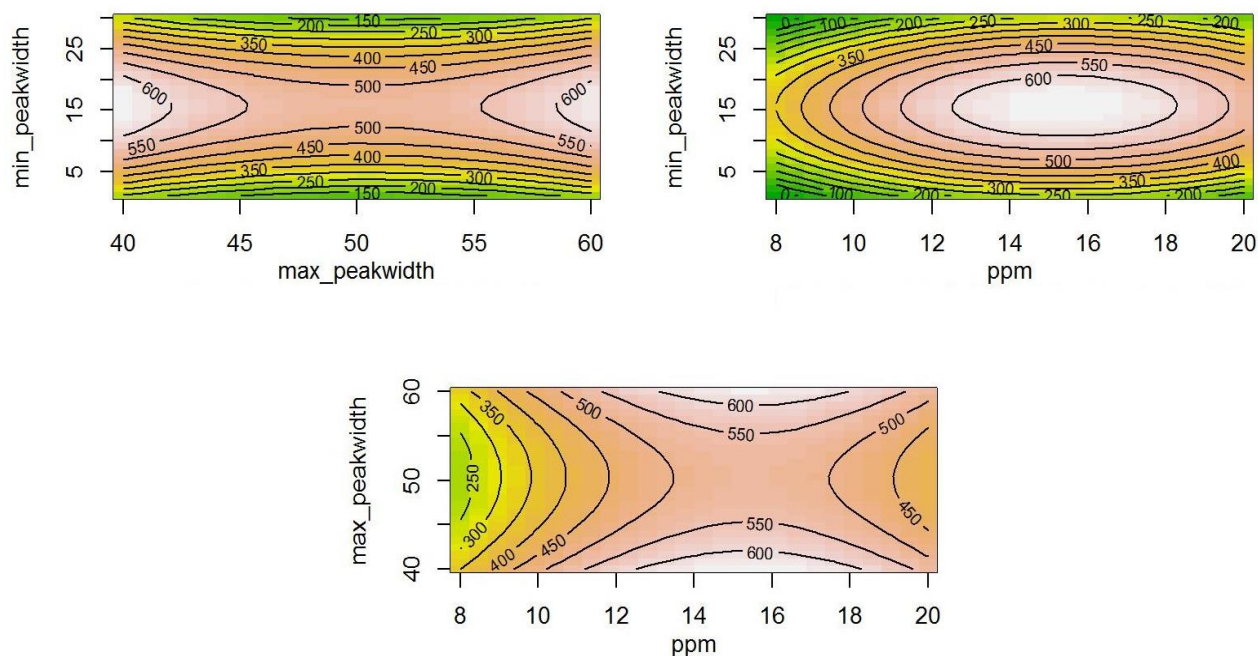
Όπου x είναι οι χρόνοι ανάσχεσης όλων των κορυφών μέσα σε μία ομάδα, k είναι ο αριθμός των χρόνων ανάσχεσης και n είναι ένας δείκτης για κάθε χρόνο. Η σχέση είναι υψωμένη στην -1 ώστε βελτίωση να επιτυγχάνεται με αύξηση του RCS.

Για τον αλγόριθμο ομαδοποίησης ιόντων, ο τύπος απόκρισης GS (Grouping Score) είναι ο ακόλουθος:

$$GS = \frac{\text{reliable groups}^2}{\text{non reliable groups}}$$

Όπου ως αξιόπιστη ομάδα κορυφών είναι αυτές που εμφανίζουν μια κορυφή σε κάθε δείγμα. Η τελική απόκριση ορίζεται ως το κανονικοποιημένο άθροισμα GS και RCS [59].

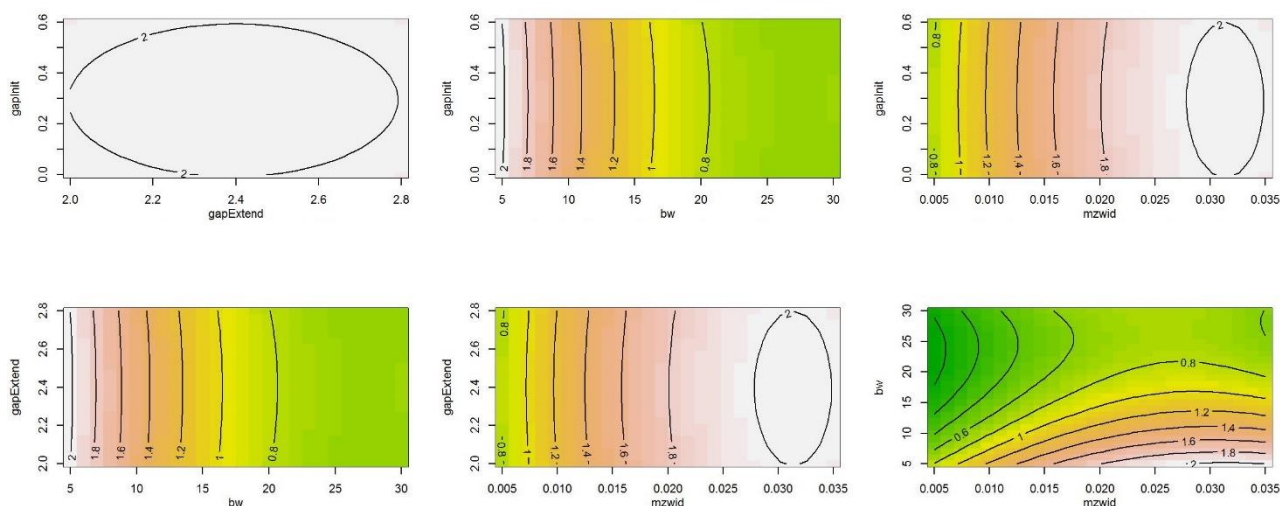
Εφαρμογή των άνω πακέτου σε δεδομένα LC-QTOFMS έδωσαν το παρακάτω μοντέλο απόκρισης:



Εικόνα 36: Επιφάνειες απόκρισης για βελτιστοποίηση αλγορίθμου εύρεσης κορυφών

Από το παραπάνω διάγραμμα συμπεραίνουμε ότι οι ιδανικοί παράμετροι για την εύρεση κορυφών η ακρίβεια μάζας είναι ίση με 15ppm και οι βέλτιστες τιμές του εύρους των κορυφών είναι 15 έως 50 δευτερόλεπτα.

Επιπλέον για την ομαδοποίηση και τη διόρθωση του χρόνου ανάσχεσης προέκυψαν οι εξής επιφάνειες απόκρισης:



Εικόνα 37: Επιφάνειες απόκρισης για βελτιστοποίηση ομαδοποίησης ιόντων και διόρθωσης χρόνου ανάσχεσης

Από τις παραπάνω επιφάνειες απόκρισης συμπεραίνουμε ότι για την ομαδοποίηση ιόντων είναι $mzwid$ ίσο με 0,031 και bw ίσο με 5, ενώ για τη διόρθωση του χρόνου ανάσχεσης οι βέλτιστες παράμετροι είναι $gapExtend$ ίσο με 2,7 και $gapInit=0,3$.

Οι παράμετροι $mzwid$ και bw σχετίζονται με την ακρίβεια μάζας και την απόκλιση του χρόνου ανάσχεσης που πρέπει να ικανοποιούν τα ιόντα σε διαφορετικά δείγματα, ώστε να αναγνωριστούν ότι αναπαριστούν τον ίδιο αναλύτη και άρα να ομαδοποιηθούν μαζί.

Οι παράμετροι $gapExtend$ και $gapInit$ είναι τα βάρη που τίθενται στον αλγόριθμο διόρθωσης χρόνου ανάσχεσης, ώστε η διαδρομή στρέβλωσης του αλγόριθμου $obiwarp$ να μην ξεφεύγει σε μεγάλο βαθμό από τη διαγώνιο (βλέπε κεφάλαιο 2.7.3.1). Οι βέλτιστες τιμές δεν απέχουν πολύ από τις βέλτιστες τιμές που πρότειναν οι συγγραφείς του αλγόριθμου ($gapInit=0,3$ και $gapExtend=2,4$) [44].

5.2 Τεχνικές προτεραιοποίησης κορυφών (Peak prioritization techniques)

Γενικά λίγες είναι οι δημοσιευμένες έρευνες μη στοχευμένης ανάλυσης σε περιβαλλοντικά δείγματα οι οποίες εκτελούν κατάταξη των ιόντων με κάποιο συγκεκριμένο τρόπο. Οι Schymanski et al. πραγματοποίησαν στοχευμένη ανίχνευση 239 ενώσεων με το $enviMass$. Αφού πραγματοποίησαν αφαίρεση λευκού δείγματος, έγινε ανίχνευση των ενώσεων στόχων και έτσι προέκυψαν

ημιποσοτικά αποτελέσματα. Επιπλέον το enviMass παρέχει μια λίστα κορυφών καταταγμένων με βάση την ένταση. Η λίστα κορυφών ομαδοποιήθηκε, για παράδειγμα ομαδοποιήθηκαν συσχετιζόμενες κορυφές ιόντων προσθήκης και ισοτοπικές κορυφές, έτσι ώστε να οδηγηθούν σε συστατικά με βάση το πακέτο nontarget [10]. Σε μια διαφορετική ερευνητική εργασία οι Moschet et al., πραγματοποίησαν στοχευμένη ανάλυση γνωστών φυτοφαρμάκων που βρίσκονται στα επιφανειακά ύδατα και βελτιστοποίησαν την εργαστηριακή πειραματική πορεία που ακολούθησαν. Βασισμένοι σε αυτό το πρωτόκολλο προκατεργασίας πραγματοποίησαν ανάλυση ύποπτων φυτοφαρμάκων και μεταβολιτών τους αρχικά σε τεχνητά κατασκευασμένα επιφανειακά ύδατα. Έτσι επικύρωσαν την υπολογιστική τους μεθοδολογία και έτσι αξιολογήθηκε κατά πόσον η μεθοδολογία αυτή οδηγεί σε επιτυχή αποτελέσματα. Βρέθηκε ότι ο λόγος θετικά λανθασμένων ευρημάτων (False Discovery Rate) ήταν 70%. Για την παραγωγή λίστας κορυφών χρησιμοποιήθηκε εμπορικό λογισμικό της κατασκευάστριας εταιρείας HRMS ExactFinder V2.0 (Thermo Fischer Scientific) και η τεχνική προτεραιοποίησης που ακολουθήθηκε ήταν με βάση την ένταση των κορυφών με περαιτέρω φιλτράρισμα τον λόγο σήμα προς θόρυβο και το σχήμα κορυφής. Έπειτα η μεθοδολογία εφαρμόστηκε σε πραγματικά νερά. Οι ύποπτοι αναλύτες και οι μεταβολίτες τους επιλέχθηκαν με βάση τις πωλήσεις των φυτοφαρμάκων στην Ελβετία και φιλτραρίστηκαν με βάση το αν μπορούν να ιοντιστούν οι δομές των φυτοφαρμάκων και τον λόγο οκτανόλης/νερό να βρίσκεται κάτω από 5, γεγονός που καθιστά τους αναλύτες κάπως υδατοδιαλυτούς [11]. Οι Hug et al., πραγματοποίησαν στοχευμένη, ύποπτη και μη στοχευμένη ανάλυση σε επεξεργασμένα λύματα. Εφάρμοσαν την ίδια μεθοδολογία όπως άλλες μελέτες [10], δηλαδή πραγματοποίησαν επικύρωση της πειραματικής μεθόδου με τις ενώσεις-στόχους (93 συστατικά από τα οποία 15 ανιχνεύθηκαν) και χρησιμοποίησαν τη μεθοδολογία αυτή για ύποπτη και μη στοχευμένη ανάλυση. Στην ύποπτη ανάλυση χρησιμοποιήθηκε μια λίστα 1835 που παρέλαβαν από καταλόγους χημικών τοπικών εργοστασίων και από τη βιβλιογραφία και φίλτραραν με βάση την πιθανότητα ιοντισμού των συστατικών, δηλαδή έδιωξαν συστατικά χωρίς κανένα ετεροάτομο και συστατικά που περιείχαν πυρίτιο και άνθρακα. Έπειτα πραγματοποίησαν παραγωγή λίστας κορυφών χρησιμοποιώντας το λογισμικό MZmine v2.9 και στις ανιχνευθείσες κορυφές έκαναν αναζήτηση για τα ύποπτα

συστατικά με βάση την ακριβή μάζα. Οι κορυφές που ανιχνεύθηκαν ως ύποπτα συστατικά αφαιρέθηκαν από τη λίστα και έπειτα η λίστα επεξεργάστηκε με το πακέτο `non-target` και επομένως έγινε κατάταξη τους με βάση αν εμφανίζουν χαρακτηριστικό ισοτοπικό προφίλ (ύπαρξη ^{37}Cl , ^{81}Br , ^{34}S , ^{15}N). Επιπλέον κατέταξαν τις κορυφές και με βάση την ένταση και επέλεξαν τις πιο υψηλές σε ένταση [13]. Ακόμα οι Chiaia-Hernandez et al., πραγματοποίησαν μη στοχευμένη ανάλυση σε ιζήματα δυο ελβετικών λιμνών (Lugano και Greifensee) καθώς το ύψος δειγματοληψίας του ιζήματος αποτελείται από αναλύτες χαρακτηριστικούς της ρύπανσης της εκάστοτε χρονικής περιόδου. Χρησιμοποίησαν το εμπορικό λογισμικό `Formulator` της εταιρίας `Thermo Fisher Scientific` που είναι ικανό να κατατάσσει τις κορυφές κατά ένταση. Οι πρώτες από τις κατεταγμένες με βάση την ένταση κορυφές αναζητήθηκαν με χρήση του `enviMass` και μόνο αυτές που παρουσίασαν χαρακτηριστικό ισοτοπικό προφίλ οδηγήθηκαν για ταυτοποίηση [16]. Τέλος οι Wang et al. μελέτησαν την αποικοδόμηση φθοριομένων συστατικών σε ενεργοποιημένη ιλύς. Αφού παρήγαγαν τις λίστες κορυφών, χρησιμοποίησαν χαρακτηριστικές απώλειες μάζας όπως για παράδειγμα την απώλεια CF_2 προκειμένου να βρουν περιβαλλοντικά προϊόντα αποδόμησης που περιείχαν φθόριο και να αποκλείσουν τα υπόλοιπα συστατικά της μήτρας [60].

5.3 Προτεραιοποίηση με χρήση του στατιστικού `Multivariate empirical Bayes`

Η στατιστική δοκιμασία που χρησιμοποιήθηκε προκειμένου να τεθούν οι κορυφές κατά αύξουσα προτεραιότητα ανάλογα με το αν εμφανίζουν ή όχι ένα έντονο προφίλ διακύμανσης της συγκέντρωσης είναι το πολυμεταβλητό εμπειρικό στατιστικό κατά `Bayes` για δεδομένα σε χρονοσειρά για τα οποία υπάρχουν διαθέσιμες επαναλήψεις (`multivariate empirical Bayes statistic for replicated time course data`) και δημοσιεύτηκε το 2006 από τους Tay και Speed του Πανεπιστημίου του Berkley και του Ινστιτούτου Eliza Hall της Αυστραλίας αντίστοιχα [61]. Οι Tay και Speed δημιούργησαν αυτή τη στατιστική δοκιμασία προκειμένου να αξιολογήσουν πειράματα έκφρασης γονιδίων σε φυτά του γένους *Arabidopsis thaliana* τα οποία προσέβαλαν με συγκεκριμένη ασθένεια σε πληθυσμούς άγριους (`wildtype`) και μεταλλαγμένους (`mutant`). Ο αντικειμενικός στόχος τους ήταν να διακρίνουν τα γονίδια εκείνα των οποίων τα

χρονικά μοτίβα έκφρασης (expression patterns) διέφεραν μεταξύ των δυο πληθυσμών. Για τον σκοπό αυτό εφάρμοσαν τη δοκιμασία δύο δειγμάτων (two-sample multivariate empirical Bayes statistic), ενώ η αντίστοιχη δοκιμασία ενός δείγματος (one sample multivariate empirical Bayes statistic), είναι αυτή που χρησιμοποιούμε στην παρούσα εργασία για να αξιολογήσουμε τη διακύμανση στις συγκεντρώσεις των αναλυτών στα εισερχόμενα λύματα, η οποία έχει ως μηδενική υπόθεση ότι το αναμενόμενο χρονικό προφίλ που εμφανίζει ο αναλύτης είναι ίσος με μηδέν με εναλλακτική το αντίθετο.

Δυο είναι οι σημαντικές κατηγορίες πειραμάτων που σχετίζονται με αλλαγές οι οποίες διαδραματίζονται στο χρόνο, οι περιοδικές (Periodic time courses) και οι χρονικά αναπτυσσόμενες (developmental time courses). Τα περιοδικά πειράματα αφορούν τυπικά βιολογικές διαδικασίες, όπως η κυτταρική διαίρεση, οι καρδιακοί ρυθμοί όπου τα χρονικά προφίλ ακολουθούν αυστηρά κάποιο συγκεκριμένο μοτίβο. Σε αντίθεση, τα χρονικά αναπτυσσόμενα πειράματα μετρούν τον/τους παράγοντα/ες σε μια διαδικασία που συμβαίνει μια αλλαγή για παράδειγμα μετά την εφαρμογή ενός φαρμάκου ή μιας θεραπείας σε έναν οργανισμό. Στην τελευταία περίπτωση υπάρχουν λίγες προσδοκίες όσο αφορά επαναλαμβανόμενα μοτίβα. Αυτή η στατιστική δοκιμασία που χρησιμοποιούμε αφορά κυρίως χρονικά πειράματα αναπτυσσόμενα διαμήκως στο χρόνο για τα οποία έχουν γίνει περισσότερες της μιας επαναλήψεις για κάθε χρονικό σημείο.

Τα πειράματα ανάλυσης μικροσυστοιχιών εμφανίζουν πολλές ομοιότητες ως προς τα χαρακτηριστικά με τα πειράματα ανάλυσης υγρών λυμάτων ή/και επιφανειακών υδάτων. Έτσι για παράδειγμα και στα δυο πειράματα μετρείται ένα πλήθος παραγόντων ενώσεων/γονιδίων, με λίγα χρονικά σημεία τα οποία μπορεί να είναι 5—10 για μικρές χρονοσειρές ή 11-20 για μεγάλες χρονοσειρές. Επιπλέον, πραγματοποιούνται λόγω κόστους κυρίως, μικρός αριθμός επαναλήψεων, τυπικά λιγότερες 2 έως 5. Ακόμα τα πειράματα μικροσυστοιχιών μπορεί να είναι περιοδικά όπως για παράδειγμα είναι περιοδικός ο κυτταρικός κύκλος. Περιοδική όμως είναι και η δειγματοληψία που διεξάγεται σε ένα κέντρο επεξεργασίας λυμάτων, διότι ακολουθείται ο κύκλος της εβδομάδας (εβδομαδιαία δειγματοληψία σύνθετων 24ωρων δειγμάτων). Τέλος, και στα δυο πειράματα μπορεί να μην εμφανιστεί κάποιο συγκεκριμένο επαναλαμβανόμενο μοτίβο όπως στα χρονικά αναπτυσσόμενα πειράματα.

Το πρόβλημα κατάταξης των γονιδίων σε πειράματα μικροσυστοιχιών είναι σχετικά νέο και λίγες γενικά μέθοδοι έχουν προταθεί για να αντιμετωπιστούν αυτά τα προβλήματα. Για τον σκοπό αυτό αρχικά χρησιμοποιήθηκαν διάφορες τεχνικές ομαδοποίησης όπως ιεραρχική ομαδοποίηση, η ομαδοποίηση Κ κοντινότερων γειτόνων ή οι χάρτες αυτοργάνωσης με στόχο τον εντοπισμό μιας ομάδας γονιδίων με ενδιαφέροντα παρόμοια μοτίβα. Οι μέθοδοι όμως αυτοί δεν κάνουν χρήση της πληροφορίας των δεδομένων που έχουν προκύψει από το ίδιο υποκείμενο αλλά χρησιμοποιούν το μέσο όρο από όλες τις επαναλήψεις δειγμάτων. Επιπλέον η ομαδοποίηση δεν κατατάσσει τα γονίδια με βάση του μεγέθους της αλλαγής στο χρόνο. Ακόμα, καθώς ο αριθμός των γονιδίων αυξάνει, οι μέθοδοι ομαδοποίησης δεν παρέχουν σαφή ομαδοποίηση. Τέλος πάντα υπάρχει και το ερώτημα πόσες ομάδες να χρησιμοποιηθούν.

Έπειτα χρησιμοποιήθηκαν κλασσικές μέθοδοι στατιστικής. Η ευρύτερα χρησιμοποιούμενη μέθοδος εντοπισμού χρονικών αλλαγών στην έκφραση των γονιδίων είναι η σύγκριση σε ζεύγη μεταξύ των στιγμών χρησιμοποιώντας στατιστικές δοκιμασίες για τη σύγκριση δυο ανεξάρτητων δειγμάτων (t -tests, LOD statistic, moderated t testistic). Αυτές οι μέθοδοι δεν είναι οι ενδεδειγμένες, διότι δεν λαμβάνουν υπόψη ότι τα δείγματα από τέτοια πειράματα συσχετίζονται. Ακόμα και η χρήση κλασσικών ή μικτών μοντέλων ANOVA (με περισσότερες από δυο κατηγορίες) για χρονικά δεδομένα εγείρει αρκετά ερωτήματα. Για τις συγκρίσεις το F τεστ υποθέτει ότι οι μετρήσεις σε διάφορες χρονικές στιγμές είναι ανεξάρτητες. Για τα κλασσικά μοντέλα ANOVA υποθέτουμε επίσης κανονικότητα των μετρήσεων, κάτι το οποίο δεν αποτελεί ιδιαίτερο παράγοντα ανησυχίας όταν τα δεδομένα βρίσκονται σε λογαριθμική κλίμακα. Ακόμα στα κλασσικά μοντέλα ANOVA, τα F -tests θα οδηγήσουν σε περισσότερα θετικά λανθασμένα ή αρνητικά λανθασμένα αποτελέσματα, λόγω των λανθασμένα υπολογισμένων διακυμάνσεων.

Έχουμε g γονίδια, συνολικού αριθμού G που τα συμβολίζουμε ως εξής: $g=1, \dots, G$ (τα οποία μπορούμε να τα φανταστούμε ως γραμμές ενός πίνακα) και ακόμα έχουμε n_g ανεξάρτητες χρονοσειρές (δηλαδή στήλες του πίνακα των γονιδίων) που τις μοντελοποιούμε ως ανεξάρτητες και ίδια κατανεμημένες μεταβλητές (independent and identically distributed-i.i.d.) υπο μορφή διανύσματος έκαστη γραμμή X_{g1}, \dots, X_{gn} με συγκεκριμένους **μέσους όρους μ_g**

και **μήτρες συνδιακύμανσης Σ_g** . Επειδή οι απόλυτες και οι σχετικές εντάσεις είναι κανονικά κατανεμημένες σε λογαριθμική κλίμακα πραγματοποιούμε την υπόθεση κανονικότητας για τις μετρήσεις X_{g1}, \dots, X_{gn} . Η υπόθεση κανονικότητας των αρχικών δεδομένων φαίνεται φυσιολογική αλλά όχι ακριβής αλλά θα κρίνουμε τα αποτελέσματα από τη χρησιμότητά τους και όχι από την καλή προσαρμογή των μοντέλων στα δεδομένα. Όπως θα δειχθεί αργότερα, ο τελικός τύπος περιλαμβάνει την πολυμεταβλητή t κατανομή. Έτσι, η μέθοδος είναι αρκούντως ανθεκτική (robust) ώστε να μπορεί να διαχειριστεί δεδομένα που είναι ελλειπτικώς κατανεμημένα. Για την μπεύσιανή προσέγγιση, χρησιμοποιούνται εκ των υστέρων-priors (χωρίς δηλαδή να έχουν χρησιμοποιηθεί τα δεδομένα) κατανομές για τα μ_g και Σ_g που αντανakλούν την ένδειξη των γονιδίων $I=I_g$, όπου $I_g=1$ αν $\mu_g \neq 0$ και $I_g=0$ αν $\mu_g=0$. Υποθέτουμε ότι η $P(I_g=1)=p$ ανεξάρτητα για κάθε γονίδιο όπου p παράμετρος μεταξύ μηδέν και ένα. Ακόμα, η μήτρα συνδιακύμανσης Σ_g θεωρείται ότι ακολουθεί την κατανομή **inverse-Wishart**

$$\Sigma \sim \text{Inv-Wishart}_v((v\Lambda)^{-1})$$

με v βαθμούς ελευθερίας και πίνακα παράμετρο $(v\Lambda)^{-1}$ όπου το Λ είναι θετικός πεπερασμένος αριθμός και το μ_g **πολυμεταβλητή κανονική κατανομή**;

$$\mu|\Sigma, I=1 \sim N(0, \text{ ή } \Sigma), n>0 \text{ και}$$

$$\mu|\Sigma, I=0 \sim N(0,0)$$

Ο μετριασμένος (moderated) πίνακας συνδιακύμανσης S δίνεται από τη σχέση:

$$\tilde{S} = [E(\Sigma^{-1}|S)]^{-1} = \frac{(n-1)S + v\Lambda}{n-1+v}$$

Και το μετριασμένο στατιστικό t είναι

$$\tilde{t} = n^{1/2} \tilde{S}^{-1/2} \bar{X}$$

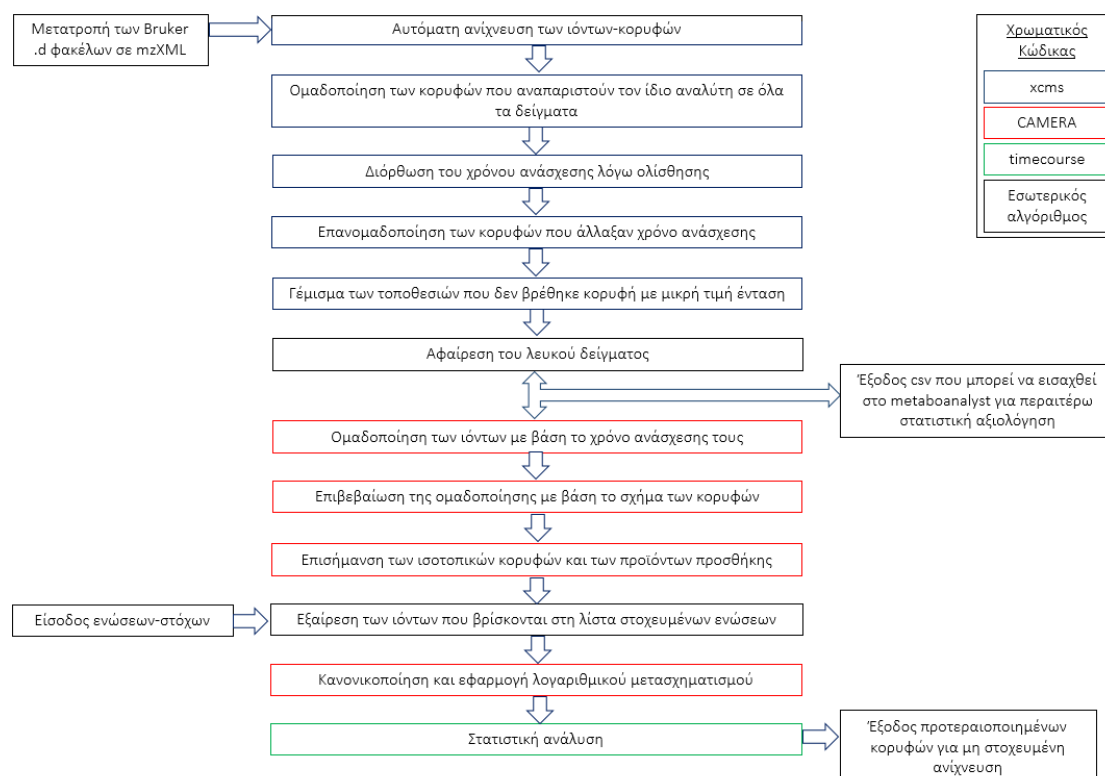
Τελικά

$$O = \frac{P(I=1|data)}{P(I=0|data)} = \frac{p}{1-p} * \frac{P(\tilde{t}|I=1)}{P(\tilde{t}|I=0)}$$

Και το πολυπαραμετρικό B στατιστικό (Multivariate B-Statistic, MB) είναι $\log_{10}O$ [61, 62].

5.4 Προτεινόμενη πορεία

Στο ακόλουθο διάγραμμα περιγράφεται σε βήματα η πορεία που ακολουθήθηκε για την παραγωγή λίστας μαζών που εμφανίζουν ισχυρή διακύμανση μεταξύ των δειγμάτων:



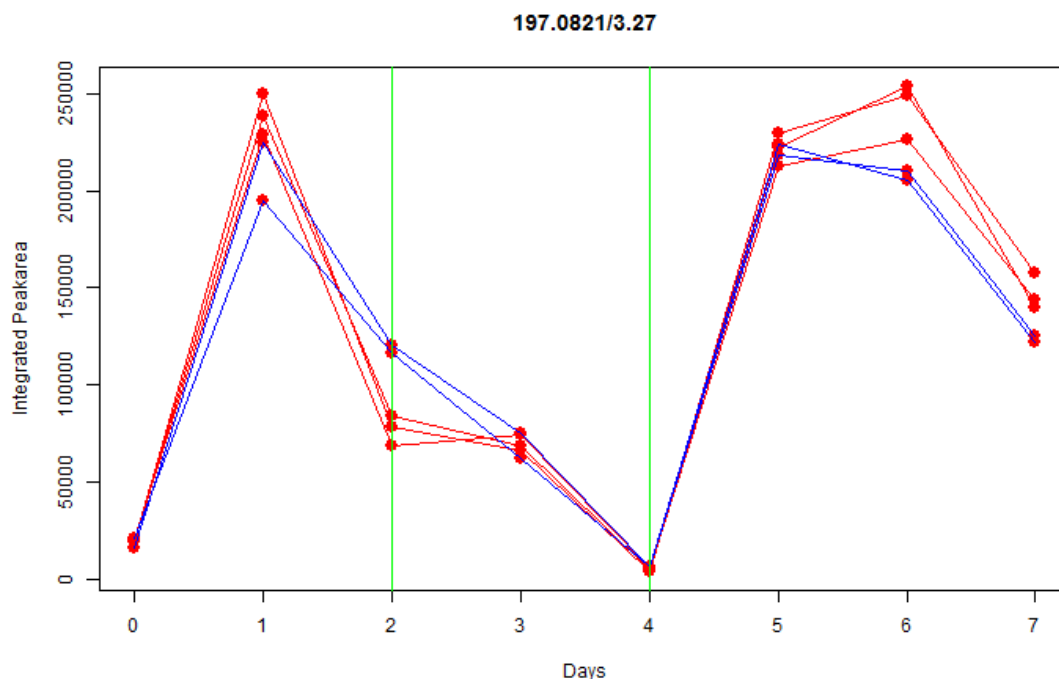
Εικόνα 38: Σχηματική υπολογιστική πορεία, που εφαρμόστηκε στα δεδομένα LC-HRMS

Για την πορεία αυτή χρησιμοποιήθηκαν τρία πακέτα, το xcms, το CAMERA και το timecourse τα οποία περιγράφονται επαρκώς στα προηγούμενα κεφάλαια.

5.5 Γραφικό περιβάλλον ανίχνευσης τάσεων

Πειραματικά πραγματοποιήθηκε ανάλυση δυο υποδειγμάτων για κάθε ημέρα (δυο αναλυτικές επαναλήψεις). Το πρώτο εκχύλισμα εγχύθηκε τρεις φορές στο LC-QTOFMS και το δεύτερο δυο. Πραγματοποιήθηκαν αναλύσεις δυο υποδειγμάτων για να διασφαλιστεί η ποιότητα των αποτελεσμάτων ενώ έγιναν συνολικά πέντε εγχύσεις. Στην ακόλουθη εικόνα αναπαρίσταται η διακύμανση ενός συστατικού μεταξύ των ημερών. Στον άξονα y είναι το εμβαδόν της κορυφής και στον άξονα x αναπαρίστανται οι ημέρες της δειγματοληψίας,

ξεκινώντας από Τετάρτη 4 Μαρτίου 2015 και καταλήγοντας στην Τετάρτη 11 Μαρτίου 2015. Με κάθετες πράσινες γραμμές αναπαρίστώνται οι μέρες της Παρασκευής και του Σαββατοκύριακου. Με κόκκινες και μπλε γραμμές απεικονίζεται το εμβαδόν για τα δυο δείγματα που αναλύθηκαν. Στον τίτλο του γραφήματος σημειώνεται η μάζα και ο χρόνος ανάλυσης σε λεπτά.



Εικόνα 39: Εργαλείο οπτικοποίησης διακυμάνσεων ιόντος με μάζα 197,0821 που εκλύεται σε χρόνο 3,27 λεπτά

5.6 Ανίχνευση ενώσεων που εμφανίζουν τάση

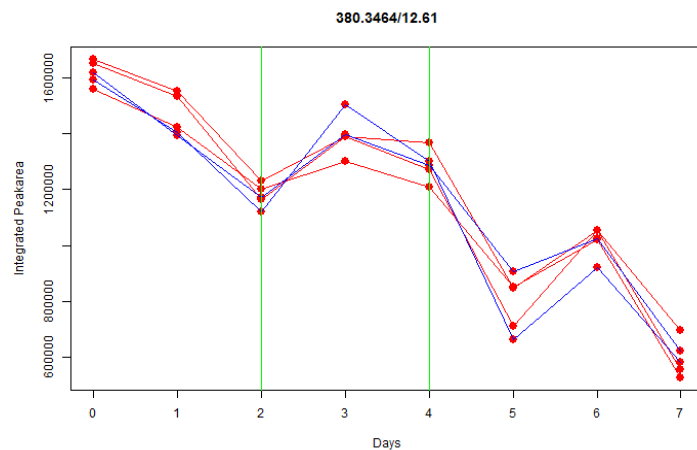
Τα πρώτα 20 αποτελέσματα της ανάλυσης τόσο για θετικό όσο και για αρνητικό ιοντισμό παρουσιάζονται στον ακόλουθο πίνακα 10. Στο επόμενο μέρος του υποκεφαλαίου θα περιγράψουμε τη μη στοχευμένη ανίχνευση τριών χαρακτηριστικών συστατικών που επισημαίνονται με αστερίσκο (*) στον πίνακα 10.

Πίνακας 10: Τα πρώτα 20 ζεύγη m/z και χρόνων ανάσχεσης αποτελέσματα για αρνητικό και θετικό ιοντισμό

Αρνητικός Ιοντισμός		Θετικός Ιοντισμός	
Μάζα προς φορτίο (m/z)	Χρόνος ανάσχεσης (min)	Μάζα προς φορτίο (m/z)	Χρόνος ανάσχεσης (min)
581,2464	4,16	682,5602	14,99
189,1125	2,86	440,4201	14,96
333,2109	9,48	504,2186	12,61
197,0821*	3,27	231,2537	11,50
444,2587	12,46	155,0722	3,40
333,2111	10,20	114,1292	4,12
333,2075	9,01	131,0357	15,21
102,0561	11,41	748,0498	15,16
333,2081	9,52	476,1622	12,66
337,2391	10,79	899,6926	15,08
317,2161	11,35	129,0200*	11,80
333,2080	10,22	897,6824	15,09
301,2211	13,71	726,0384	15,16
621,3516	11,99	728,6007	15,43
201,9817	3,08	521,1837	12,73
299,2001	12,83	380,3464*	12,61
275,1658	7,54	862,6866	15,37
750,4511	12,81	858,6671	14,93
333,2111	8,95	816,6552	15,39
141,0917	2,86	857,614/0	15,11

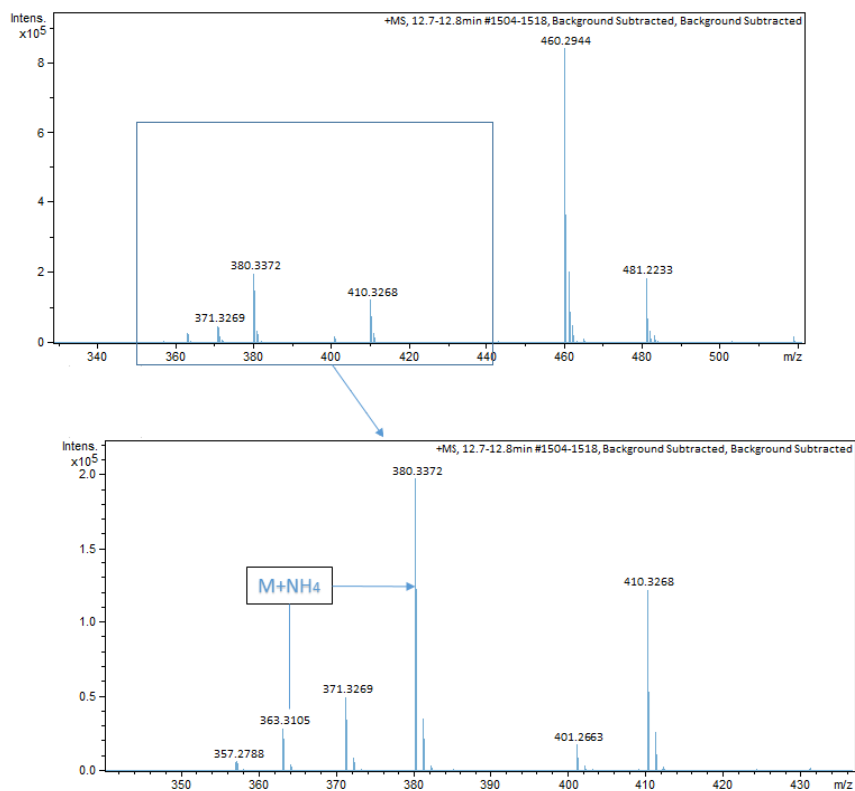
5.6.1 Παράδειγμα ανίχνευσης συστατικού σε θετικό ιοντισμό

Το ιόν 380,3484 που εκλούεται σε χρόνο 12,61 λεπτά σε θετικό ιοντισμό επέδειξε συνεχώς μειούμενη πορεία κατά τις ημέρες δειγματοληψίας



Εικόνα 40: Διακύμανση ιόντος 380,3464

Το φάσμα MS1 στην περιοχή 12,61 υποδεικνύει ότι το ιόν που ανιχνεύθηκε είναι προϊόν προσθήκης $[M+NH_4]^+$ του μοριακού ιόντος 363,3105



Εικόνα 41: Φάσμα MS κορυφής σε χρόνο 12,61 λεπτά

Αξιοσημείωτα ακριβή είναι τα αποτελέσματα της επισήμανσης των ιόντων του πακέτου CAMERA που αποτυπώνουν με ακρίβεια το προηγούμενο φάσμα MS και μπορούν να δοθούν στον ακόλουθο πίνακα:

Πίνακας 11: Ιόντα που ανιχνεύθηκαν και επεξήγηση τους για το χρόνο 762-765 δευτερόλεπτα

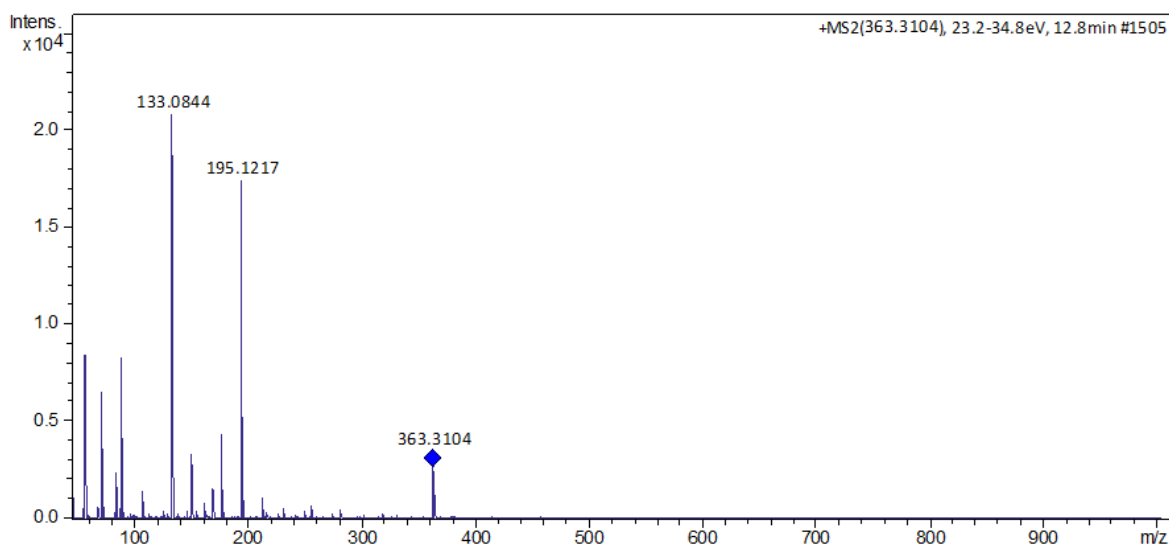
Ιόν (m/z)	Χρόνος ανάσχεσης (δευτερόλεπτα)	Επεξήγηση (Annotation)
357,2838	762,4955	$[M+H-C_3H_2O_3]^+$ 442,267
364,3188	763,1455	$[M+K]^+$ 325,359 $[M+Na]^+$ 341,333 $[M+H]^+$ 363,314
365,3219	764,6460	
375,2935	765,5870	
380,3464	763,6660	
381,3455	764,5580	
401,2726	764,2950	$[M+H-COCH_2]^+$ 442,267
410,3331	764,5960	$[M+K+HCOOH]^+$ 325,359 $[M+Na+HCOOH]^+$ 341,333 $[M+H+HCOOH]^+$ 363,314

Το πακέτο CAMERA όχι μόνο ανίχνευσε όλα τα ιόντα αλλά κατέδειξε ότι υπάρχει συνέκλουση στο χρονικό σημείο 762-765 δευτερόλεπτα τεσσάρων ουσιών με μοριακά ιόντα 325,359, 341,333, 363,314 και 442,267.

Με βάση το λογισμικό SmartFormula Manually προέκυψαν 15 πιθανοί μοριακοί τύποι εκ των οποίων msigma μικρότερο του 50 έδειξαν 8 τύποι. Από αυτούς

τους 8 τύπους μόνο 3 υπάρχουν σε χημικές βάσεις δεδομένων (Chemspider και PubChem).

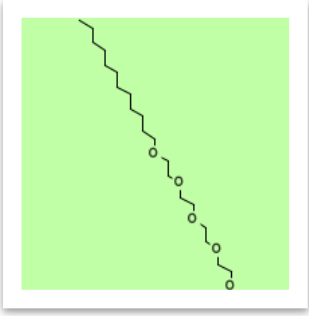
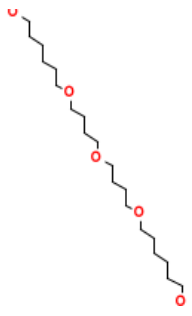
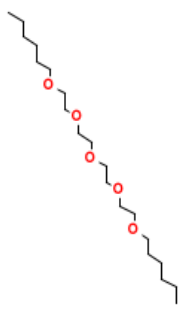
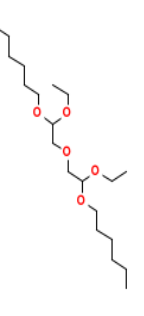
Απομόνωση του ιόντος 363,3105 έδωσε πλούσια θραυσματοποίηση στο φάσμα MS/MS:



Εικόνα 42: Θραυσματοποίηση μοριακού ιόντος 363,3105

Δοκιμή των τριών τύπων στο MetFusion [29] σε συνδυασμό με την πρόβλεψη του χρόνου ανάλυσης [63] οδήγησε στο συμπέρασμα ότι ο μοριακός τύπος είναι $C_{20}H_{44}O_5$ και ότι οι πιθανές δομές είναι οι ακόλουθες:

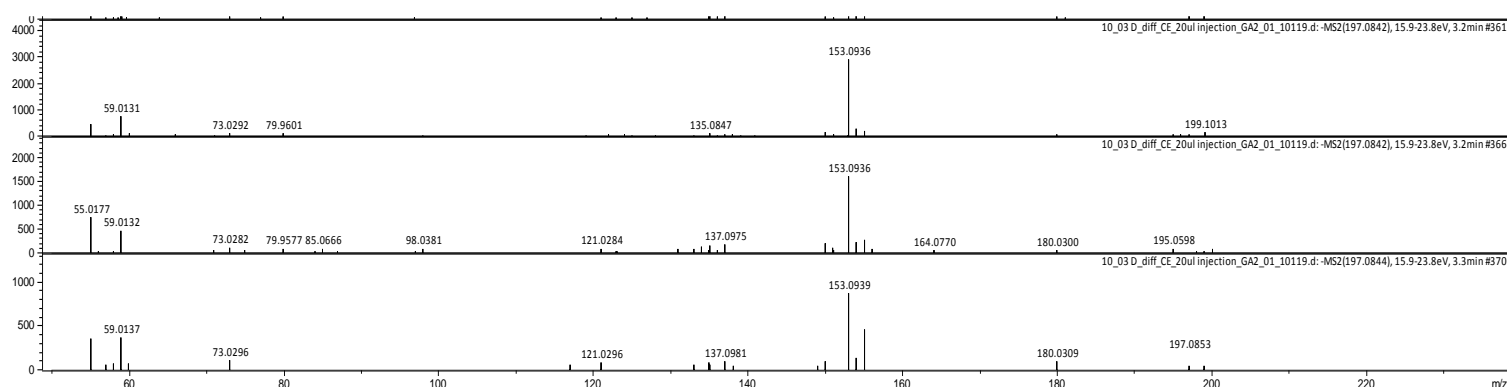
Πίνακας 12: Υποψήφιες δομές μοριακού ιόντος 363,3105

Δομή				
Όνομα	3,6,9,12-Tetraoxatetracosan-1-ol	6,6'-[Oxybis(4,1-butanediylloxy)]di(1-hexanol)	7,10,13,16,19-Pentaoxapentacosane	1-[1-Ethoxy-2-[2-ethoxy-2-(hexyloxy)ethoxy]ethoxy]hexane
#Πηγές δεδομένων	41	2	4	4
#Αριθμός αναφορών	62	2	4	4
Θραύσματα που εξηγούνται	13	4	3	4
Βαθμολογία Metfusion	0,754	0,513	0,497	0,485
Προβλεπόμενος & Πειραματικός χρόνος ανάλυσης	12,55 & 12,61	11,20 & 12,61	12,84 & 12,61	14,17 & 12, 61

Το συστατικό 3,6,9,12-Tetraoxatetracosan-1-ol είναι βιομηχανικό χημικό και είναι πρόδηλα η μια και μοναδική προτεινόμενη δομή λόγω του αριθμού δεδομένων και αναφορών που αποτελούν ένδειξη εμπορικότητας του συγκεκριμένου χημικού αλλά και λόγω του υψηλού αριθμού θραυσμάτων που εξηγούνται από την in silico θραυσματοποίηση της δομής. Επιπλέον ο προβλεπόμενος χρόνος ανάλυσης ταυτίζεται με το πειραματικό. Το συστατικό ανιχνεύθηκε σε επίπεδο 2B [18].

5.6.2 Παράδειγμα ανίχνευσης συστατικού σε αρνητικό ιοντισμό

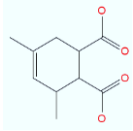
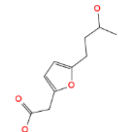
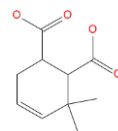
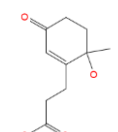
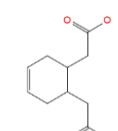
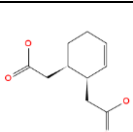
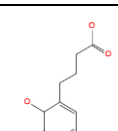
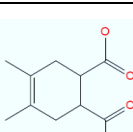
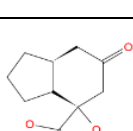
Το ιόν 197,0821 εκλούεται σε χρόνο 3,27 λεπτά και το προφίλ διακύμανσης φαίνεται στην εικόνα 39. Παρατηρείται μια πτώση κατά τις ημέρες της Παρασκευής και του Σαββατοκύριακου. Το ιόν 197,0821 είναι και το ψευδομοριακό ιόν. Για αυτή τη χαμηλού μοριακού βάρους ουσία προκύπτουν μόνο 13 υποψήφιοι μοριακοί τύποι, εκ των οποίων οι 9 με αποδεκτή ταύτιση θεωρητικού και πειραματικού ισοτοπικού προφίλ. 5 από τους 9 τύπους υπάρχουν στις χημικές βάσεις δεδομένων PubChem ή/και ChemSpider. Αν και το φάσμα MS/MS είναι χαμηλού σήματος είναι επαναλήψιμο:



Εικόνα 43: Φάσματα MS/MS του μοριακού ιόντος 197,0821 σε παράθεση

Με χρήση του MetFrag και του εργαλείου πρόβλεψης χρόνου ανάσχεσης που αναπτύχθηκε εσωτερικά [63] προέκυψαν οι ακόλουθες υποψήφιες δομές:

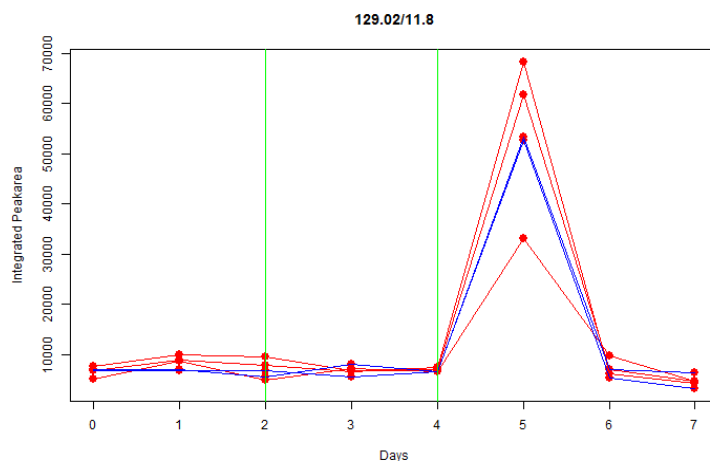
Πίνακας 13: Υποψήφιες δομές μοριακού ιόντος 197,0821

Δομή	Πειραματικός & προβλεπόμενος χρόνος ανάλυσης	Θραύσματα που εξηγούνται	Βαθμολογία Metfrag	Αριθμός αναφορών	#Πηγές Δεδομένων
	3,27 & 3,85	9	0,912	23	27
	3,27 & 5,13	9	0,902	<3	<3
	3,27 & 4,07	9	0,912	<3	<3
	3,27 & 2,53	9	0,978	<3	<3
	3,27 & 2,64	9	0,921	10	12
	3,27 & 4,75	9	0,921	<3	<3
	3,27 & 5,42	9	0,955	<3	<3
	3,27 & 3,30	8	0,928	21	23
	3,27 & 3,07	8	0,928	<3	<3

Με βάση την εμπορική απήχηση και τον χρόνο πρόβλεψης οδηγούμαστε στο ότι οι επικρατέστερες δομές είναι οι δυο φθαλικοί εστέρες και επομένως το επίπεδο ανίχνευσης είναι 3. Οι φθαλικοί εστέρες χρησιμοποιούνται κυρίως ως πλαστικοποιητές και θεωρούνται χημικά που χρησιμοποιούνται στη βιομηχανία. Η μείωση της συγκέντρωσης τους κατά το Σαββατοκύριακο ενδεχομένως υποδηλώνει τη μείωση της λειτουργίας των βιοτεχνιών και βιομηχανιών πλαστικών. Να σημειωθεί ότι το κέντρο επεξεργασίας λυμάτων της Ψυττάλειας δέχεται και αστικά λύματα αλλά και βιομηχανικά και νοσοκομειακά.

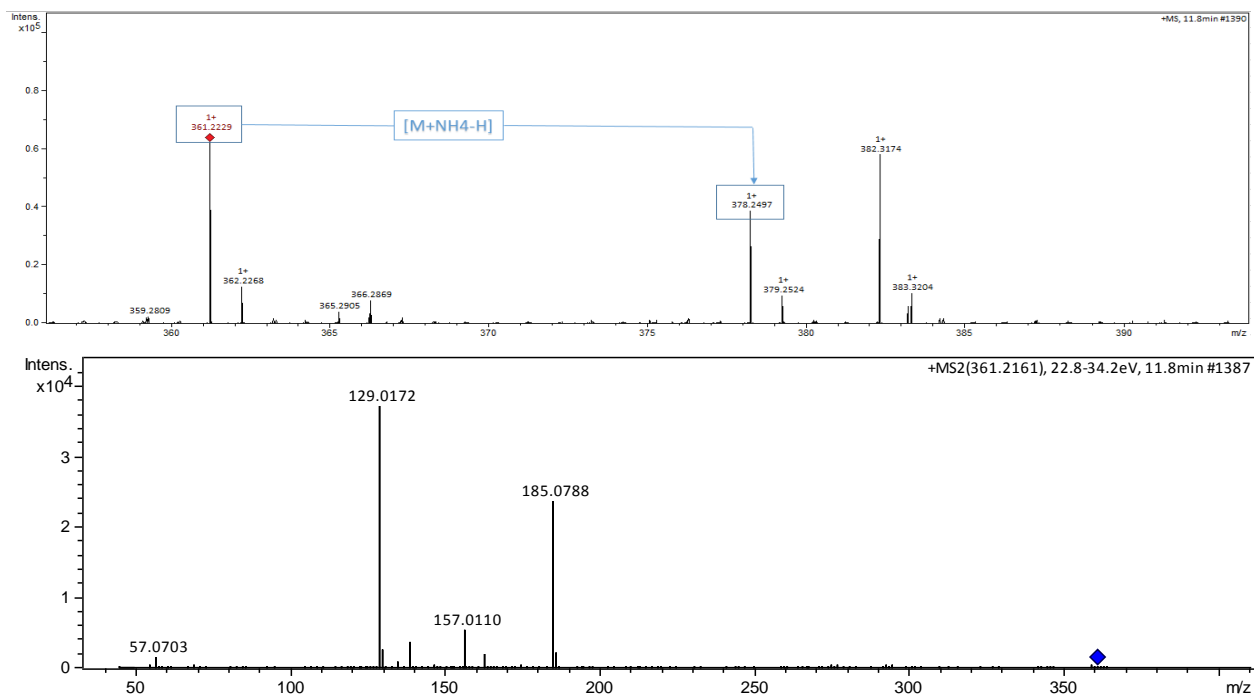
5.6.3 Παράδειγμα ανίχνευσης «παλμού» ρύπανσης (pollution spill) σε θετικό ιοντισμό

Το ιόν με μάζα 129,0200 που εκλύεται σε χρόνο 11,80 λεπτά παρουσίασε αξιοσημείωτο προφίλ τάσης:



Εικόνα 44: Προφίλ ιόντος 129,0200 σε χρόνο 11,80 λεπτά

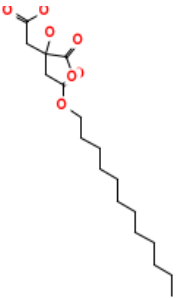
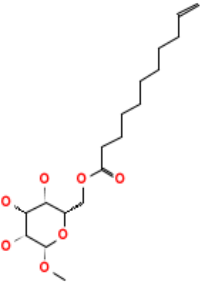
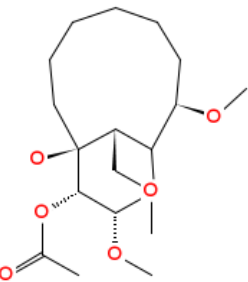
Το ιόν πρόκειται για θραύσμα του μοριακού ιόντος 361,2161 όπως φαίνεται στο ακόλουθο φάσμα MS/MS:



Εικόνα 45: Φάσματα MS και MS/MS του ιόντος 361,2161 που παράγει το θραύσμα 129,0172

Το φάσμα MS δείχνει ότι το μοριακό ιόν είναι το 361,2161 εφόσον υπάρχει το προϊόν προσθήκης $[M+NH_4-H]^+$. Με βάση την ακρίβεια μάζας και της πληροφορίας που μπορεί να εξαχθεί από το ισοτοπικό προφίλ συμπεραίνουμε ότι υπάρχουν 26 αποδεκτοί (<50 msigma) μοριακοί τύποι εκ των οποίων για τους 9 υπάρχουν υποψήφιες δομές σε χημικές βάσεις δεδομένων. Με εισαγωγή των θραυσμάτων και των δομών στο Metfrag προέκυψε ότι η στοιχειακή σύνθεση είναι $C_{18}H_{34}O_7$, ενώ οι επικρατέστερες δομές είναι οι ακόλουθες:

Πίνακας 14: Υποψήφιες δομές μοριακού ιόντος 361,2161

Δομή			
Όνομα	2-[2-(Dodecyloxy)-2-oxoethyl]-2-hydroxysuccinic acid	Methyl 6-O-10-undecenoyl-α-D-glucopyranoside	(1S,8R,9R,11S,12R,13S)-1-Hydroxy-8,11-dimethoxy-13-(methoxymethyl)-10-oxabicyclo[7.3.1]tridec-12-yl acetate
#Πηγές δεδομένων	3	2	2
#Αριθμός αναφορών	3	2	2
Θραύσματα που εξηγούνται	4	2	2
Βαθμολογία MetFrag	0,917	0	1
Προβλεπόμενος & Πειραματικός χρόνος ανάλυσης	11,16 & 11,80	9,18 & 11,80	8,94 & 11,80

Λόγω του υψηλότερου αριθμού θραυσμάτων που εξηγούνται και λόγω της εγγύτητας των προβλεπόμενων και πειραματικού χρόνου ανάλυσης προτείνεται η πρώτη δομή και επομένως φτάνουμε σε επίπεδο ανίχνευσης 2B.

Για την ταυτοποίηση των ενώσεων σε επίπεδο 1 απαιτείται η αγορά, η έγχυση προτύπων ουσιών αναφοράς και η ταύτιση των φασμάτων MS, MS/MS και του χρόνου ανάλυσης του προτύπου αναφοράς και του δείγματος.

Συμπερασματικά, το πρωτόκολλο που αναπτύχθηκε επιτρέπει την ανάλυση πολλών δεδομένων από περιβαλλοντικά δείγματα και την προτεραιοποίηση των κορυφών που εμφανίζουν τάση και σημαντική διακύμανση. Τέλος, το πρωτόκολλο αυτό μπορεί να εφαρμοστεί και σε άλλα γνωστικά πεδία πέραν της περιβαλλοντικής ανάλυσης

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
Adapters	Προσαρμογείς
Analysis modules	Μονάδες ανάλυσης
Annotation	Σχολιασμός/επισήμανση/επεξήγηση (ιόντων)
Apps	Σετ εργαλείων
Background subtraction	Αφαίρεση του υποβάθρου
Bandwidth	Εύρος
Batch search	Εκτέλεση ομάδας αναζητήσεων
Bin	Κομμάτι/Τμήμα
Binary group data analysis/time-series data analysis	Ανάλυση δεδομένων δυο ομάδων/ ανάλυση χρονοσειρών
Centroid mode	Φασματικές κορυφές υπο μορφή κορυφών μηδενικού πλάτους/κεντροειδές
Clustering και Classification	Μέθοδοι ομαδοποίησης και κατηγοριοποίησης
Clusters	Ομάδες
Collision gas	Αέριο σύγκρουσης
Color picker	Επιλογή χρωμάτων
Combinational fragmenter	Συνδυαστική θραυσματοποίηση
Command-line configuration	Ρυθμίσεις από την γραμμή εντολών
Continuous Wavelet Transformation	Μετασχηματισμός συνεχούς εφαρμογής κυματιδίων

Correlation Across samples-CAS	Συσχέτιση της έντασης δυο ιόντων μεταξύ όλων των δειγμάτων
Correlation heatmaps	Συσχετισμένοι χάρτες θερμότητας
Correlation Peak Shape-CPS	Συσχέτιση σχήματος δύο ιόντων μεταξύ ενός δείγματος
Cross-polarity rule table	Πίνακας κανόνων αντίθετου ιοντισμού
Data Check	Στάδιο ελέγχου δεδομένων
Data editor	Επεξεργαστής δεδομένων
Data filter	Φιλτράρισμα δεδομένων
Data layer	Στιβάδα δεδομένων
Data partitioning	Διαμερισματοποίηση των μετρήσεων
Data visualization	Οπτικοποίηση των δεδομένων
Density based method	Προσέγγιση βασισμένη στην πυκνότητα
Design of Experiments	Πειραματικός σχεδιασμός
Developmental time courses	Χρονικώς αναπτυσσόμενα πειράματα
DTW	Μεθόδου της δυναμικής χρονικής στρέβλωσης
Egauss	Ρίζα μέσου τετραγωνικού σφάλματος της προσαρμογής της καμπύλης Gauss στην χρωματογραφική κορυφή
False discovery rate	Λόγος ψευδός θετικών ευρημάτων
Features	Ιόντα
From scan to scan	Από φάσμα πλήρους σάρωσης σε φάσμα πλήρους σάρωσης
Full width at half maximum	Εύρος κορυφής στο 50% του ύψος
Gaps/transitions	Μεταβάσεις

Hierarchical agglomerative clustering algorithm	Ιεραρχικός αλγόριθμος ομαδοποίησης
High throughput analysis	Τεχνικές υψηλής απόδοσης
Independent and identically distributed variables	Ανεξάρτητες και ίδια κατανομημένες μεταβλητές
Interpolation	Μέθοδος της παρεμβολής
IQR	Ενδοτεταρτημοριακό εύρος
Jagged peak	Οδοντωτή κορυφή
Kernel	Πυρήνας
Kernel Density Estimator	Εκτιμητής πυκνότητας πιθανότητας
LC-HRMS	Υγροχρωματογραφία διασυνδεδεμένη με φασματομετρία μάζας υψηλής διακριτικής ικανότητας
Libraries	Σουίτα βιβλιοθηκών
Line mode	Φασματικά δεδομένα υπο μορφή γραμμών
Machine learning methods	Τεχνικές εκμάθησης
Mass error	Σφάλμα στον υπολογισμό της μάζας
MassBank record	Εγγραφή MassBank
Matching Score	Βαθμός αντιστοίχισης
Metabolic pathway analysis	Ανάλυση μεταβολικής οδού
Metabolite set enrichment analysis	Ανάλυση επαύξησης μεταβολιτών
MetPA	Εύρεση μεταβολικής οδού
Minimum intensity	Ελάχιστη ένταση
Missing values	Ελλιπούσες τιμές
Modules	Μονάδες

mother wavelet/analyzing wavelet	Wavelet προς ανάλυση
MSEA	Ανάλυση εμπλουτισμού μεταβολιών
Msigma	Συντελεστής ομοιότητας μεταξύ του θεωρητικού ισοτοπικού προφίλ και του και του πειραματικού ισοτοπικού προφίλ
Multiple group data analysis	Ανάλυση δεδομένων πολλαπλών ομάδων
Multivariate empirical Bayes statistic for replicated time course data	Πολυμεταβλητή στατιστική δοκιμή κατά Bayes για δεδομένα σε χρονοσειρά για τα οποία υπάρχουν διαθέσιμες επαναλήψεις
NA values	Μη ευρεθείσες τιμές
Non target screening	Μη στοχευμένη ανάλυση
Normalization	Κανονικοποίηση
Output	Έξοδος
Package	Πακέτο
Peak Picking Score	Βαθμολογία απόκρισης εύρεσης κορυφών
Performance	Απόδοση
Periodic time courses	Περιοδικά πειράματα
Permutation	Αντιμετάθεση
Pollution spill	Παλμός ρύπανσης
Post hoc analysis	Εκ των υστέρων ανάλυση
Post-run acquisition	Διερεύνηση των δεδομένων μετά τις μετρήσεις
Precision	Ακρίβεια
Pre-process	Προεπεξεργασία
Profile mode	Φασματικά δεδομένα μορφή κορυφών/ σε μορφή προφίλ

QqQ	Τριπλό τετράπολο
QqTOF/QTOF	Τετράπολο-αναλυτής χρόνου πτήσης
Quality control sample	Δείγματα ελέγχου ποιότητας
RDBE	Βαθμός δακτυλίων και διπλών δεσμών
Recall	Ανάκληση
Replicates	Επαναλήψεις
Robustness	Ανθεκτικότητα
ROIs	Περιοχές ενδιαφέροντος
rt	Χρόνος ανάσχεσης
S/N	Λόγο σήμα προς θόρυβο
Score	Βαθμολογία
SD	Τυπική απόκλιση
Self-organizing maps	Χάρτες αυτοοργάνωσης
Server	Διακομιστής
Sewage epidemiology	Επιδημιολογία λυμάτων
Signal to base	Σήμα προς γραμμή υποβάθρου
Signal to noise	Σήμα προς θόρυβος
Slices	Κομμάτια
Smoothed spectra	Εξομαλυμένα φάσματα
Spectral summary	Φασματικό άθροισμα
Step	Βήμα

Suspect list	Λίστα ύποπτων συστατικών
Suspect screening	Ύποπτη ανάλυση
Target list	Λίστας ενώσεων στόχων
Target screening	Στοχευμένη ανάλυση
Tracking algorithms	Αλγόριθμοι ανίχνευσης
Utility layer	Βοηθητική στιβάδα
Warping path	Διαδρομή στρέβλωσης
Wavelets	προσέγγιση κυματιδίων
Windowed Fourier Transform	Μετασχηματισμός Fourier με δειγματοληψία με χρήση καθορισμένου παραθύρου
Expression patterns	Χρονικώς μοτίβα έκφρασης
Input files	Αρχεία εισόδου
K Nearest Neighbor	Μέθοδος του κοντινότερου γείτονα
Mean value	Μέση τιμή

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

Ακρωνύμια και ανάπτυξη τους

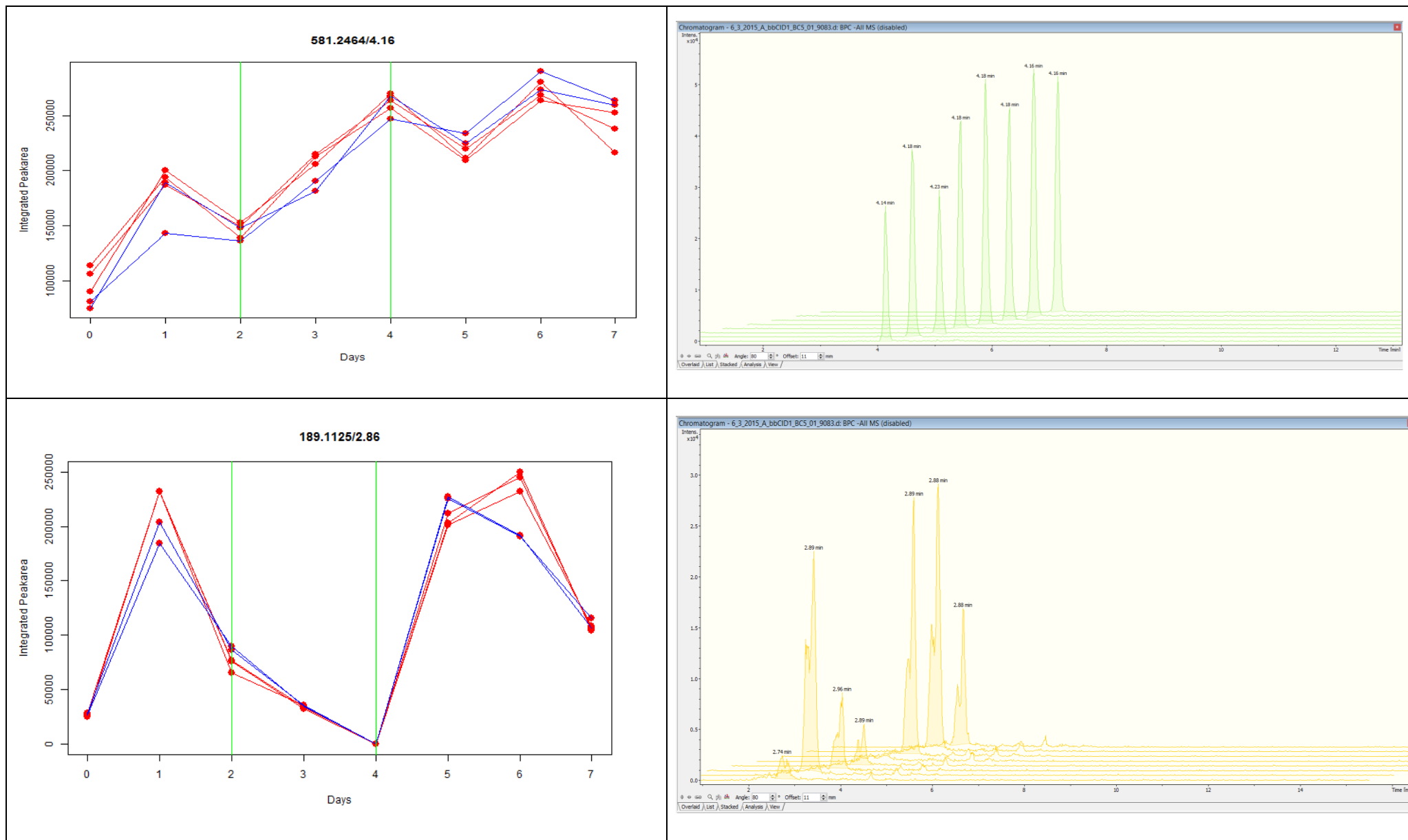
ANOVA	Analysis of variance
Apps	Applications
ASCA	ANOVA Simultaneous Component Anaysis
BBD	Box-Behnken design
CAMERA	Collection of Algorithms for Metabolite pRofile Annotation
CAS	Correlation Across samples
CDK	Chemistry Development Kit
CPS	Correlation Peak Shape
csv	Comma separated values
CWT	Continuous Wavelet Transformation
DoE	Design of Experiments
EBAM	empirical Bayesian analysis of microarrays (and metabolites)
EIC/EICs	Extracted Ion Chromatogram(s)
FDR	False discovery rate
GS	Grouping Score
HILIC	Hydrophilic interaction liquid chromatography
HLB	Hydrophilic-lipophilic balance
HMDB	Human Metabolome Database
HPLC	High Pressure Liquid Chromatography
i.i.d.	Independent and identically distributed
InChiKEY	International Chemical Identifier
IPO	Isotopologue Parameter Optimization
IQR	Interquartile range

IT	Ion-Trap
KEGG	Kyoto Encyclopedia of Genes and Genomes
kNN	k Nearest Neighbor
LC-ESI-MS	Liquid Chromatography-ElectroSpray Ionization-Mass Spectrometer
LC-HRMS	Liquid Chromatography-High Resolution Mass Spectrometry
LC-MS	Liquid Chromatography Mass Spectrometry
LIP	Low Intensity Peaks
loess	Local Regression Fitting Method
MASS	mzXML-Associated Standard Solution
MetPA	metabolic pathway analysis
MSEA	metabolite set enrichment analysis
NA	Not a number
OBI-Warp	Ordered Bijective Interpolated Warping
PCA	Principal component analysis
PLS-DA	Partial least squares Discriminant Analysis
ppm	Parts per million
PPS	Peak Picking Score
QC	Quality control sample
RCS	Retention time score
RDBE	Rings plus Double Bonds Equivalent
RF	Random Forest
ROIs	Regio(s) of interest
RP	Reliable Peaks
rt	Retention time
S/N	Signal to noise ratio
SAM	significance analysis of microarrays (and metabolites)

SD	Standard deviation
SGR	Seven Golden Rules
SMPDB	Small Molecule Pathway Database
SVM	Support vector machine
WP	warping path
X-AW	Wean Anion eXchange
XCMS	X(Gas or Liquid) Chromatography Mass Spectrometry
X-CW	Weak Cation eXchange
XML	Extensible Markup Language
ΕΚΠΑ	Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

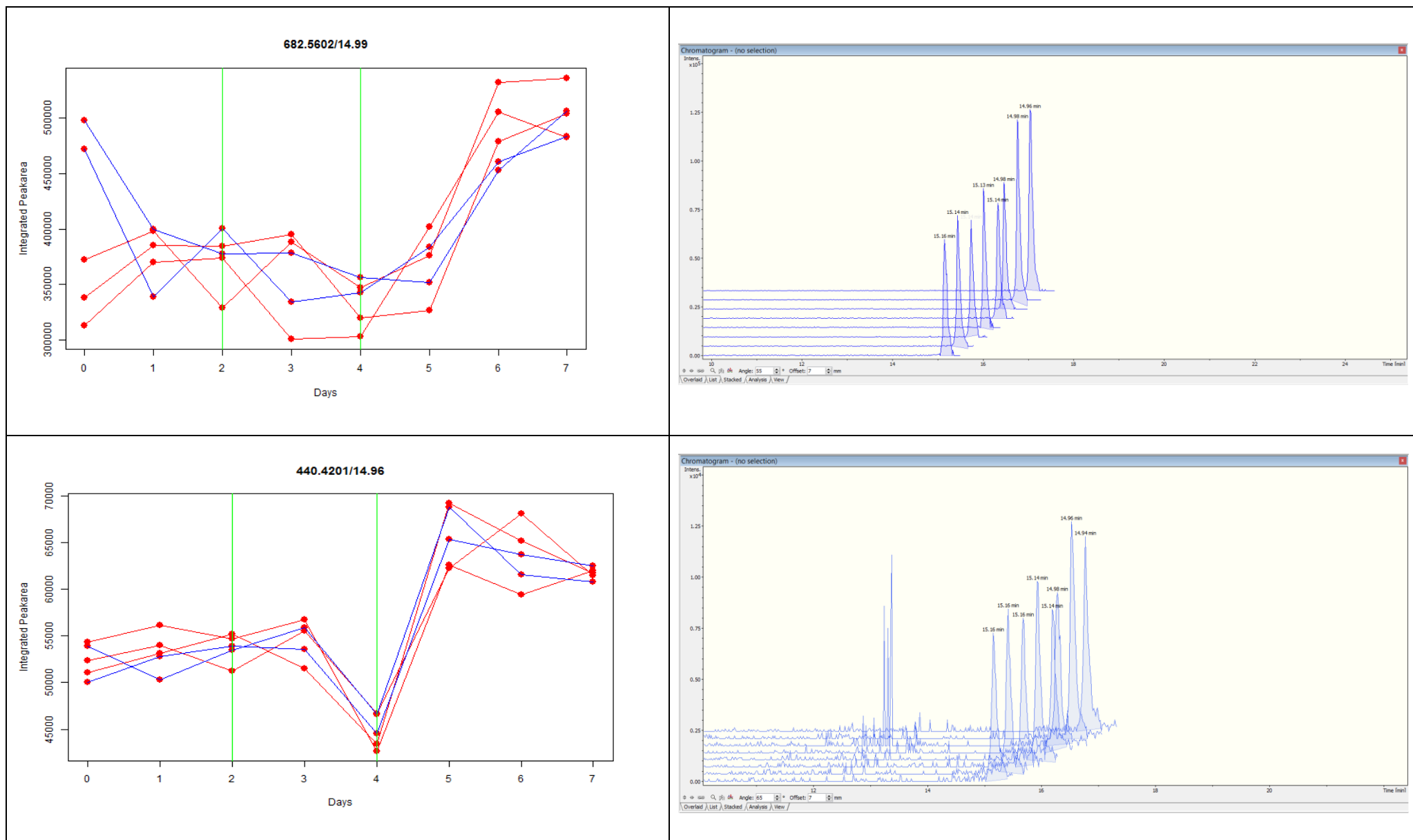
ΠΑΡΑΡΤΗΜΑ Ι

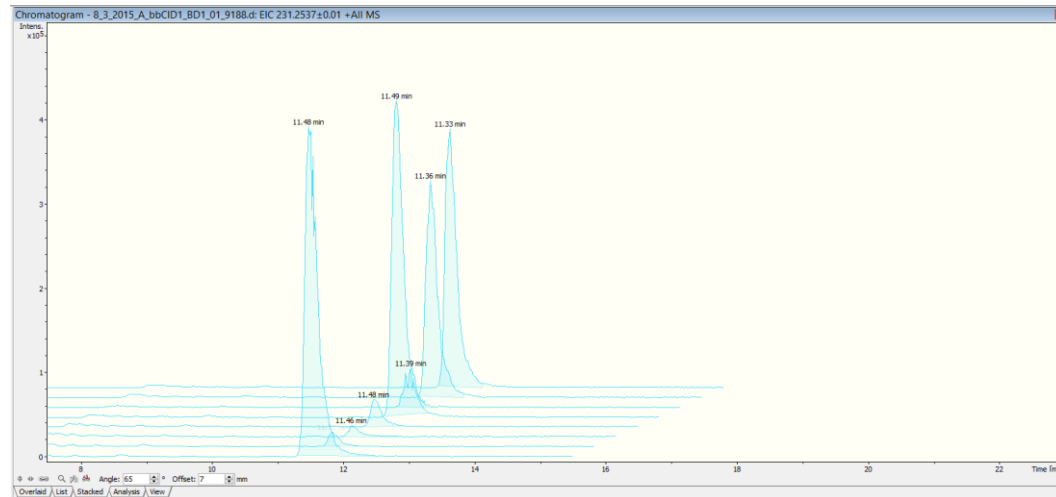
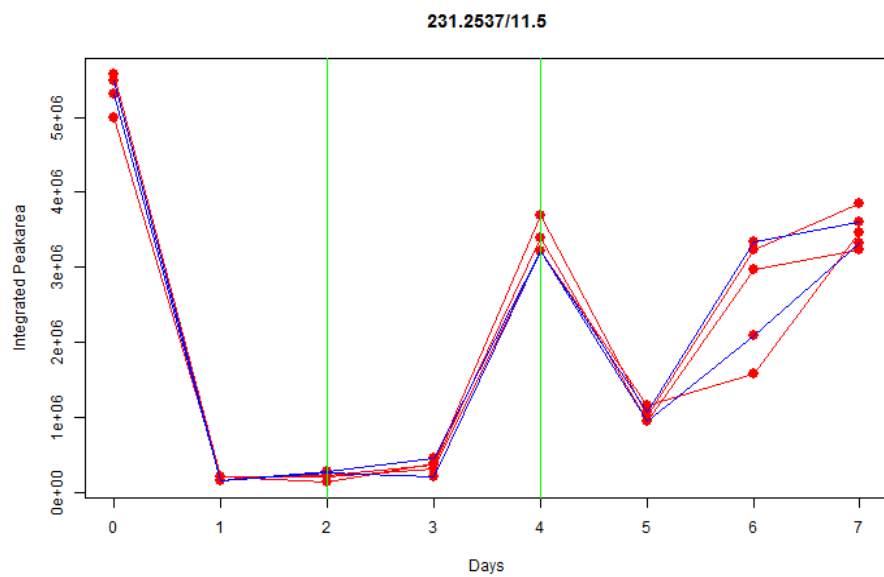
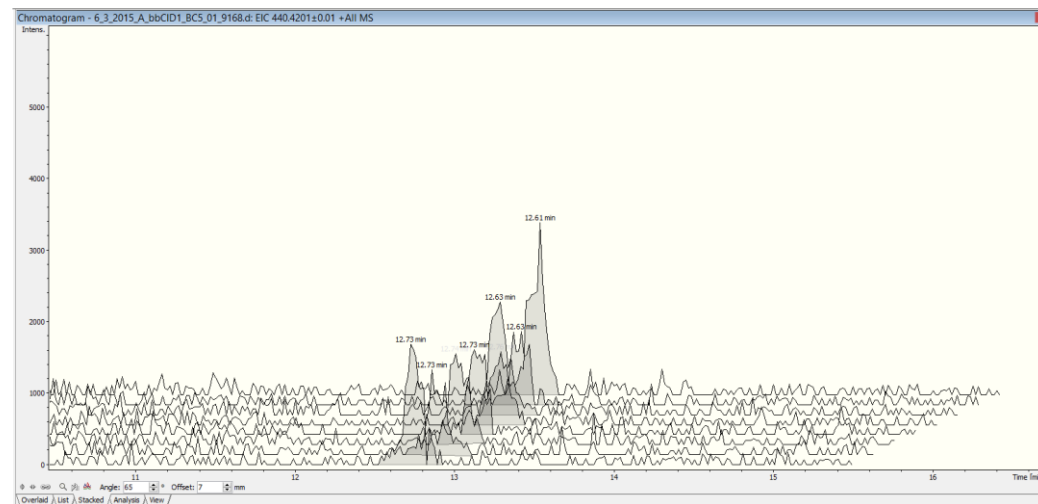
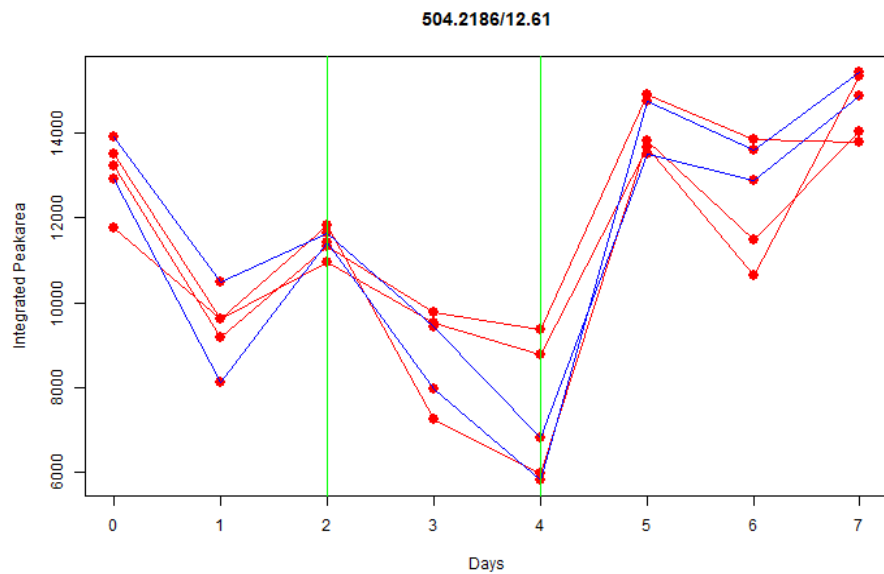
Πίνακας 15: Γραφική απεικόνιση τάσεων και χρωματογραφημάτων για τα δέκα πρώτα υψηλότερης προτεραιότητας ιόντα στον αρνητικό ιοντισμό

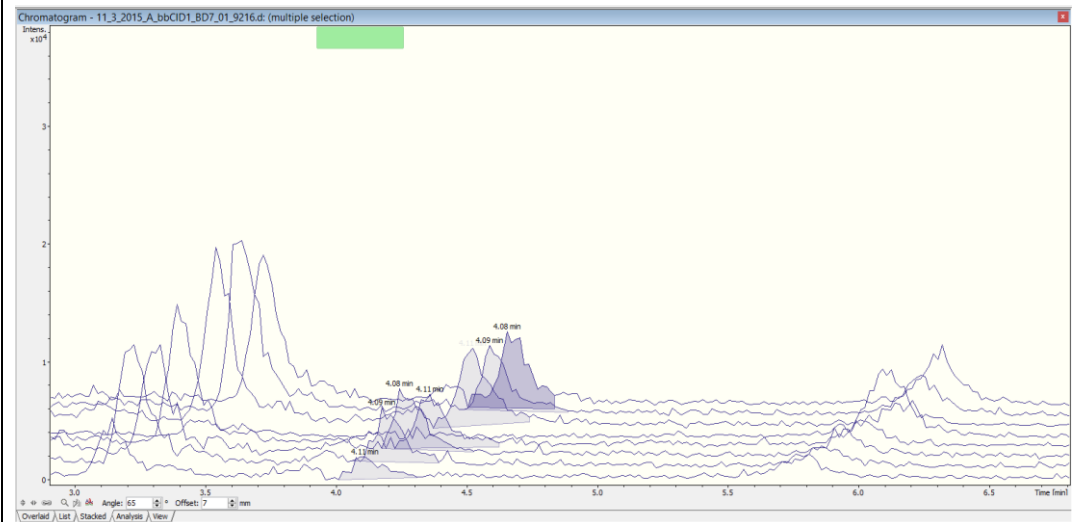
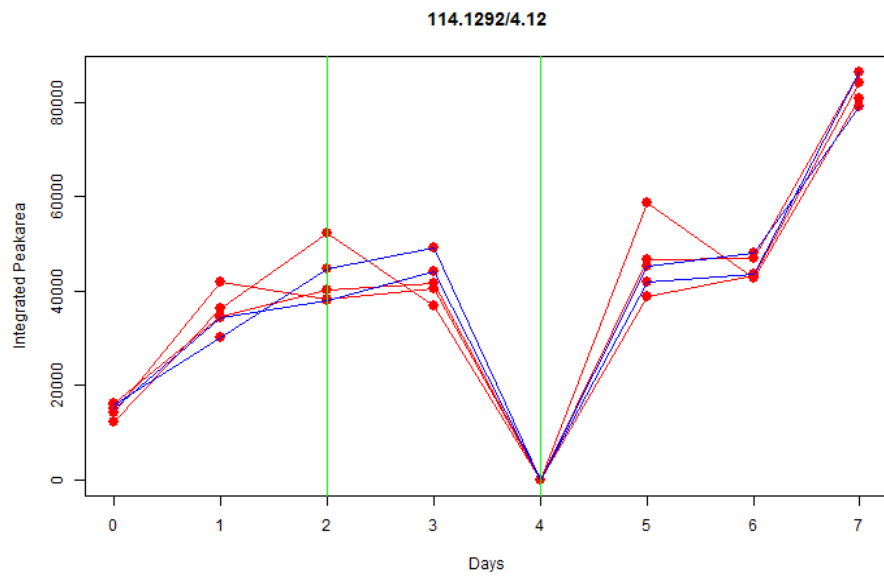
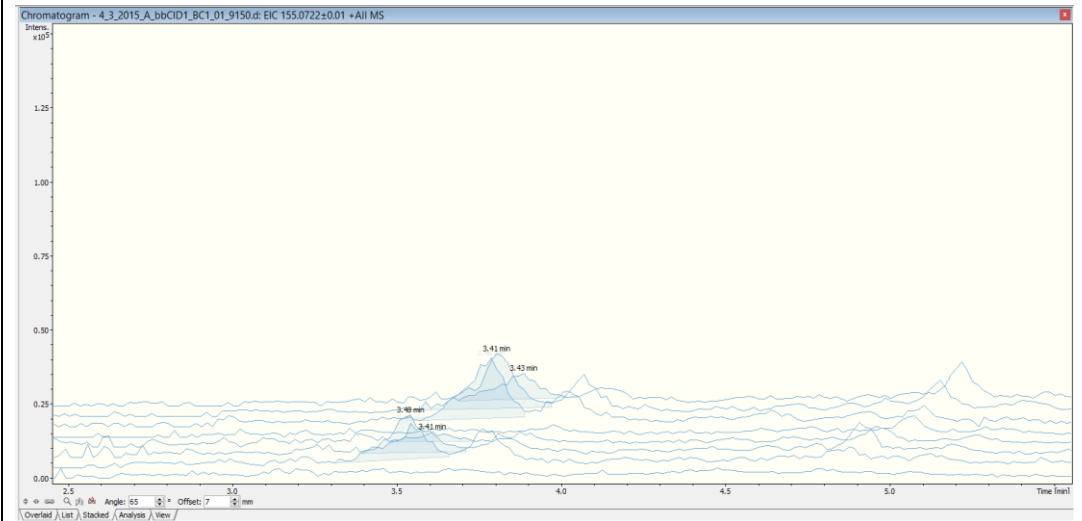
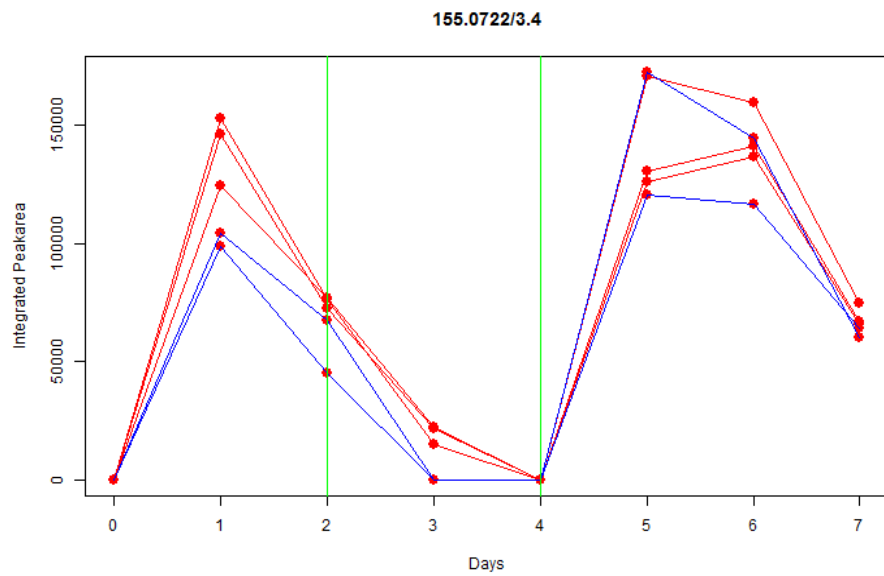


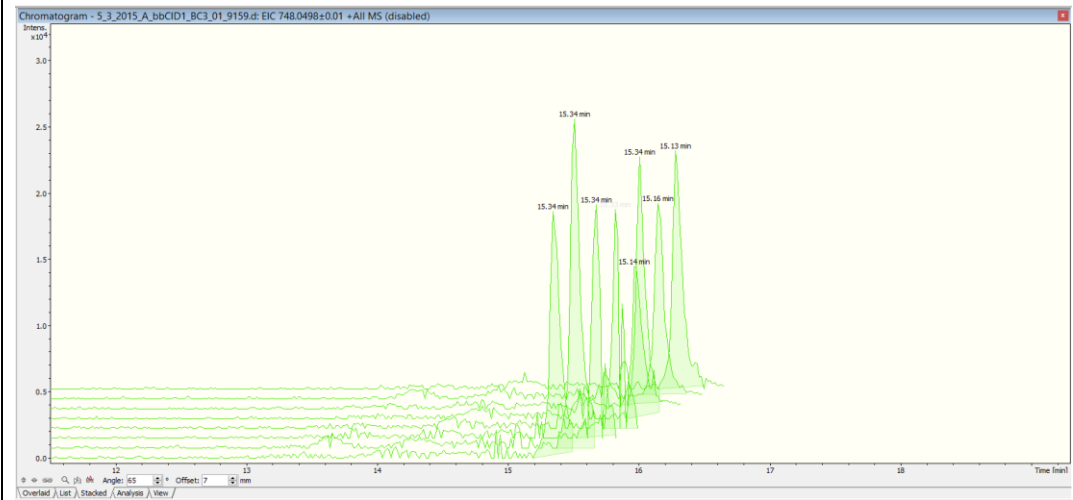
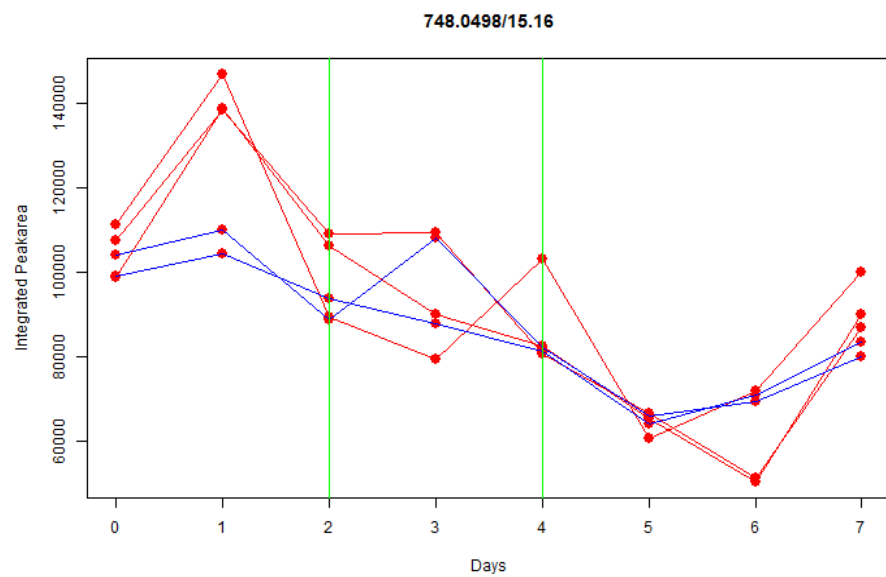
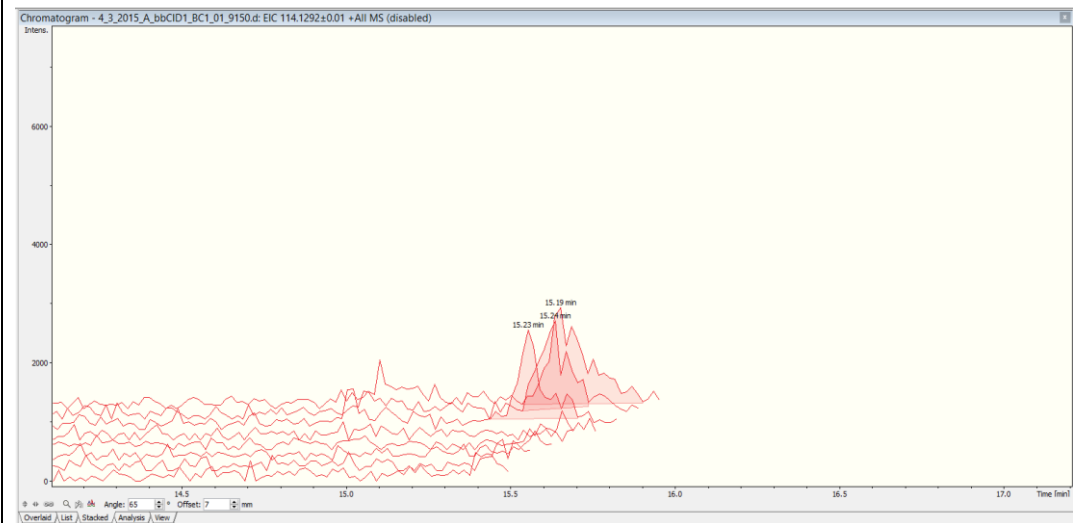
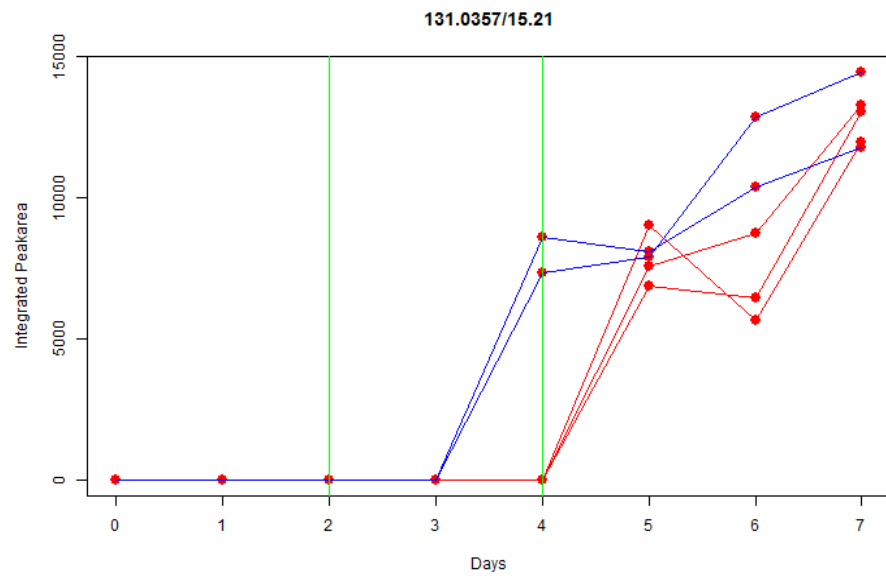
ΠΑΡΑΡΤΗΜΑ II

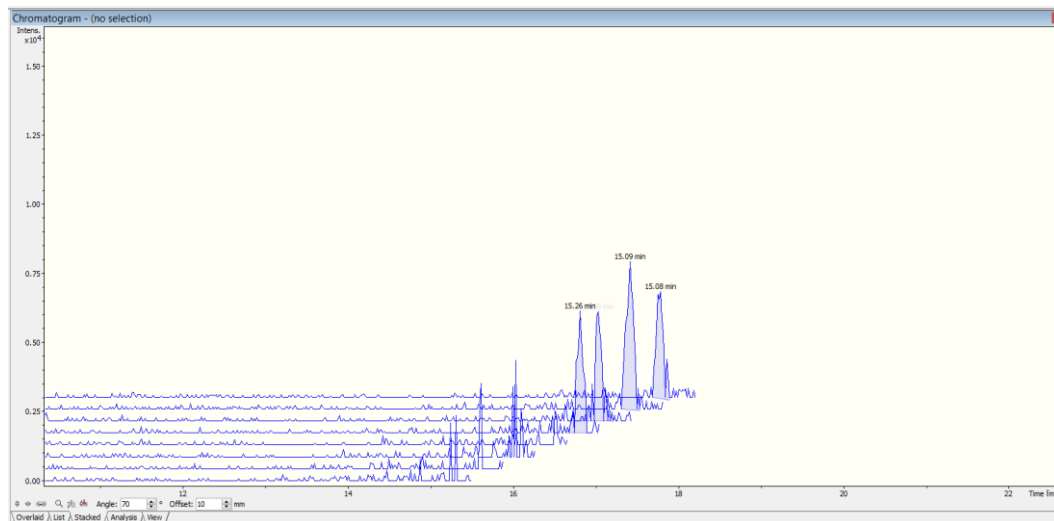
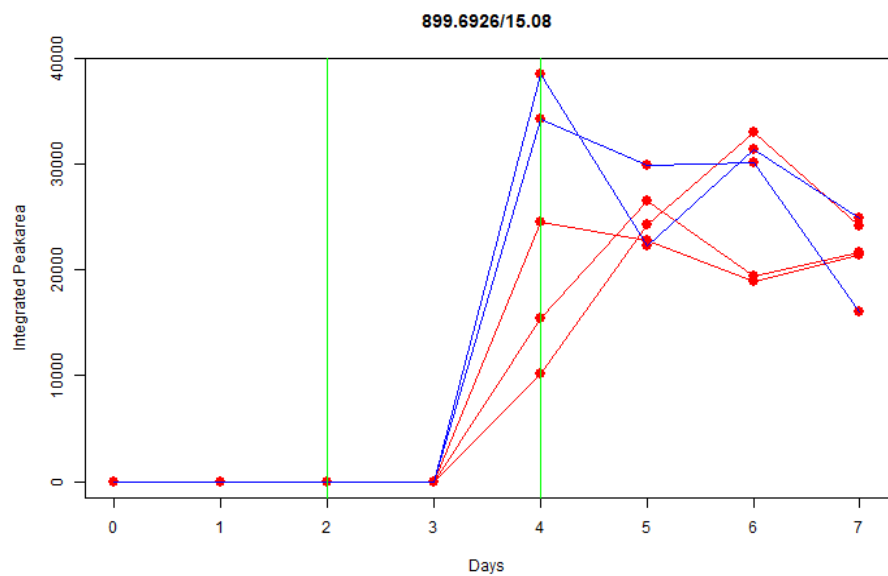
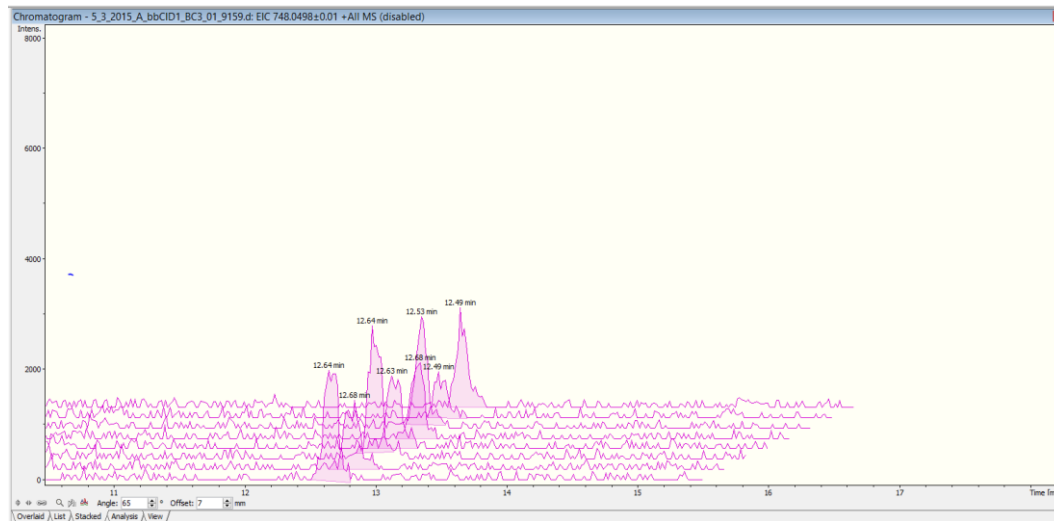
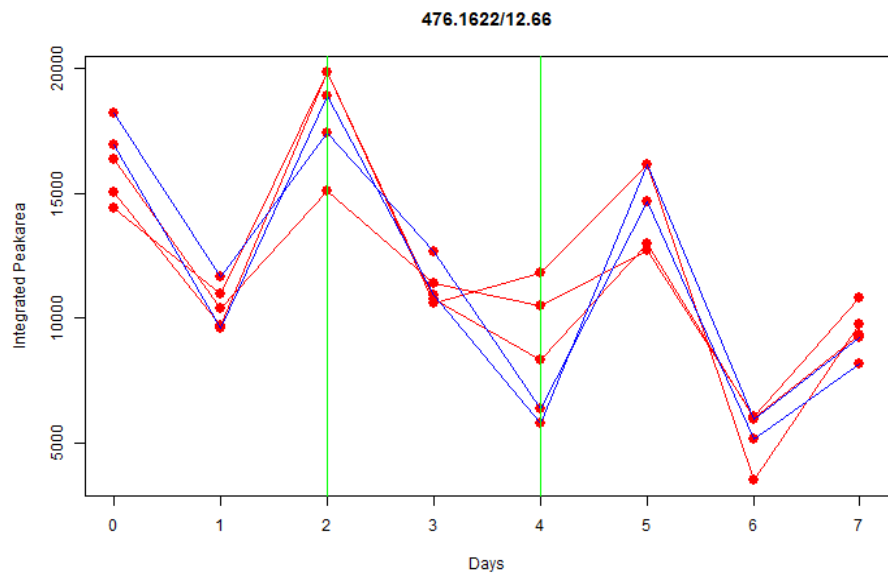
Πίνακας 16: Γραφική απεικόνιση τάσεων και χρωματογραφημάτων για τα δέκα πρώτα υψηλότερης προτεραιότητας ιόντα στον θετικό ιοντισμό











ΑΝΑΦΟΡΕΣ

1. N. Thomaidis, A. Asimakopoulos, A. Bletsou, "*Emerging contaminants: A tutorial mini-review*", *GlobalNest*, 14 (2012) 72-79.
2. V.S. Thomaidi, A.S. Stasinakis, V.L. Borova, N.S. Thomaidis, "*Is there a risk for the aquatic environment due to the existence of emerging organic contaminants in treated domestic wastewater? Greece as a case-study*", *J Hazard Mater*, 283 (2015) 740-747.
3. S.D. Richardson, T.A. Ternes, "*Water analysis: emerging contaminants and current issues*", *Anal Chem*, 86 (2014) 2813-2848.
4. P.H. Roberts, K.V. Thomas, "*The occurrence of selected pharmaceuticals in wastewater effluent and surface waters of the lower Tyne catchment*", *Sci Total Environ*, 356 (2006) 143-153.
5. B. Kasprzyk-Hordern, R.M. Dinsdale, A.J. Guwy, "*The occurrence of pharmaceuticals, personal care products, endocrine disruptors and illicit drugs in surface water in South Wales, UK*", *Water Res*, 42 (2008) 3498-3518.
6. B. Petrie, R. Barden, B. Kasprzyk-Hordern, "*A review on emerging contaminants in wastewaters and the environment: current knowledge, understudied areas and recommendations for future monitoring*", *Water Res*, 72 (2015) 3-27.
7. C. Ort, A.L. van Nuijs, J.D. Berset, L. Bijlsma, S. Castiglioni, A. Covaci, P. de Voogt, E. Emke, D. Fatta-Kassinos, P. Griffiths, F. Hernandez, I. Gonzalez-Marino, R. Grabic, B. Kasprzyk-Hordern, N. Mastroianni, A. Meierjohann, T. Nefau, M. Ostman, Y. Pico, I. Racamonde, M. Reid, J. Slobodnik, S. Terzic, N. Thomaidis, K.V. Thomas, "*Spatial differences and temporal changes in illicit drug use in Europe quantified by wastewater analysis*", *Addiction*, 109 (2014) 1338-1352.
8. K.V. Thomas, L. Bijlsma, S. Castiglioni, A. Covaci, E. Emke, R. Grabic, F. Hernandez, S. Karolak, B. Kasprzyk-Hordern, R.H. Lindberg, M. Lopez de Alda, A. Meierjohann, C. Ort, Y. Pico, J.B. Quintana, M. Reid, J. Rieckermann, S. Terzic, A.L. van Nuijs, P. de Voogt, "*Comparing illicit drug use in 19 European cities through sewage analysis*", *Sci Total Environ*, 432 (2012) 432-439.
9. N. Thomaidis, P. Gago-Ferrero, C. Ort, V. Borova, N. Alygizakis, N. Maragou, M. Dasenaki, C. Pistos, "*Effect of strong socio-economic changes on licit and illicit drug use patterns through sewage epidemiology*", *Environ Sci Technol*, (2015).
10. E.L. Schymanski, H.P. Singer, P. Longree, M. Loos, M. Ruff, M.A. Stravs, C. Ripolles Vidal, J. Hollender, "*Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry*", *Environ Sci Technol*, 48 (2014) 1811-1818.
11. C. Moschet, A. Piazzoli, H. Singer, J. Hollender, "*Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry*", *Anal Chem*, 85 (2013) 10312-10320.

12. M. Krauss, H. Singer, J. Hollender, "LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns", *Anal Bioanal Chem*, 397 (2010) 943-951.
13. C. Hug, N. Ulrich, T. Schulze, W. Brack, M. Krauss, "Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening", *Environ Pollut*, 184 (2014) 25-32.
14. E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibanez, T. Portoles, R. de Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipanicev, P. Rostkowski, J. Hollender, "Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis", *Anal Bioanal Chem*, (2015).
15. M. Zedda, C. Zwiener, "Is nontarget screening of emerging contaminants by LC-HRMS successful? A plea for compound libraries and computer tools", *Anal Bioanal Chem*, 403 (2012) 2493-2502.
16. A.C. Chiaia-Hernandez, E.L. Schymanski, P. Kumar, H.P. Singer, J. Hollender, "Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments", *Anal Bioanal Chem*, 406 (2014) 7323-7335.
17. A.A. Bletsou, J. Jeon, J. Hollender, E. Archontaki, N.S. Thomaidis, "Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment", *TrAC Trends in Analytical Chemistry*, 66 (2015) 32-44.
18. E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, "Identifying small molecules via high resolution mass spectrometry: communicating confidence", *Environ Sci Technol*, 48 (2014) 2097-2098.
19. T. Kind, O. Fiehn, "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry", *BMC Bioinformatics*, 8 (2007) 105.
20. B.D. GmbH, "Dataanalysis 4.1 reference user manual", Bremen, Germany, 2012.
21. B.D. GmbH, "ProfileAnalysis 2.1 Reference manual", Bremen, Germany, 2013.
22. MassBank, "MassBank Record Format 2.09", April 2013.
23. S. Tanaka, Y. Fujita, H.E. Parry, A.C. Yoshizawa, K. Morimoto, M. Murase, Y. Yamada, J. Yao, S. Utsunomiya, S. Kajihara, M. Fukuda, M. Ikawa, T. Tabata, K. Takahashi, K. Aoshima, Y. Nihei, T. Nishioka, Y. Oda, K. Tanaka, "Mass++: A Visualization and Analysis Tool for Mass Spectrometry.", *J Proteome Res*, 13 (2014) 3846-3853.
24. H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida,

- K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, "MassBank: a public repository for sharing mass spectral data for life sciences", *J Mass Spectrom*, 45 (2010) 703-714.
25. M.A. Stravs, E.L. Schymanski, H.P. Singer, J. Hollender, "Automatic recalibration and processing of tandem mass spectra using formula annotation", *J Mass Spectrom*, 48 (2013) 89-99.
26. S.E. Stein, R.D. Scott, "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification", *J Am Soc Mass Spectrom*, 5 (1994) 859-866.
27. MassBank, "MassBank User's Manual 2.4", Japan, February 2012.
28. S. Wolf, S. Schmidt, M. Muller-Hannemann, S. Neumann, "In silico fragmentation for computer assisted identification of metabolite mass spectra", *BMC Bioinformatics*, 11 (2010) 148.
29. M. Gerlich, S. Neumann, "MetFusion: integration of compound identification strategies", *J Mass Spectrom*, 48 (2013) 291-298.
30. P.G. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu, R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research", *Nat Biotechnol*, 22 (2004) 1459-1466.
31. A.S.A.E.d.b.S. E13.15, "Standard Guide for Analytical Data Interchange Protocol for Mass Spectrometric Data", 2010.
32. M.C. Chambers, B. Maclean, R. Burke, D. Amodei, D.L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T.A. Baker, M.Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S.L. Seymour, L.M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E.W. Deutsch, R.L. Moritz, J.E. Katz, D.B. Agus, M. MacCoss, D.L. Tabb, P. Mallick, "A cross-platform toolkit for mass spectrometry and proteomics", *Nat Biotechnol*, 30 (2012) 918-920.
33. C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, "XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification", *Anal Chem*, 3 (2006) 779-787.
34. C.A. Smith, "LC/MS Preprocessing and Analysis with xcms", R vignettes, (2015).
35. R. Tautenhahn, C. Böttcher, S. Neumann, "Highly sensitive feature detection for high resolution LC/MS", *BMC Bioinformatics*, 9 (2008).
36. R. Danielsson, D. Bylund, K. Markides, "MatchedFilter for improved quality of base peak chromatograms and mass spectra in LC MS", *Anal Chim Acta*, 2 (2002) 167-184.

37. R. Stolt, R.J. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, “*Second-order peak detection for multicomponent high-resolution LC/MS data.*”, *Anal Chem*, 4 (2006) 975-983.
38. Ε. Σκαλίδης, “*Εφαρμογή των Wavelet στην Επίλυση Ηλεκτρομαγνητικών Προβλημάτων*”, Aristotle University of Thessaloniki, Thessaloniki, 2010.
39. C.J. Conley, R. Smith, R.J. Torgrip, R.M. Taylor, R. Tautenhahn, J.T. Prince, “*Massifquant: open-source Kalman filter-based XC-MS isotope trace feature detection*”, *Bioinformatics*, 30 (2014) 2636-2643.
40. K.M. Aberg, R.J. Torgrip, J. Kolmert, I. Schuppe-Koistinen, J. Lindberg, “*Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking*”, *J Chromatogr A*, 1192 (2008) 139-146.
41. M. Loos, “*Extraction of ion chromatograms by unsupervised clustering of high-resolution mass spectrometry data.*”, *R vignettes*, (2015).
42. S.A. Kazmi, S. Ghosh, D.-G. Shin, D.W. Hill, D.F. Grant, “*Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics*”, *Metabolomics*, 2 (2006) 75-84.
43. M. Katajamaa, J. Miettinen, M. Oresic, “*MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data*”, *Bioinformatics*, 22 (2006) 634-636.
44. J.T. Prince, E.M. Marcotte, “*Chromatographic Alignment of ESI-LC-MS Proteomics Data Sets by Ordered Bijective Interpolated Warping*”, *Anal Chem*, 78 (2006) 6140-6152.
45. J.T. Theodore, “*Integration and Validation of Mass Spectrometry Proteomics Data Sets*”, University of Texas at Austin, Texas, U.S.A., May 2008, pp. 165.
46. O. Yanes, R. Tautenhahn, G.J. Patti, G. Siuzdak, “*Expanding coverage of the metabolome for global metabolite profiling*”, *Anal Chem*, 83 (2011) 2152-2161.
47. C. Kuhl, R. Tautenhahn, C. Bottcher, T.R. Larson, S. Neumann, “*CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets*”, *Anal Chem*, 84 (2012) 283-289.
48. J. Xia, N. Psychogios, N. Young, D.S. Wishart, “*MetaboAnalyst: a web server for metabolomic data analysis and interpretation*”, *Nucleic Acids Res*, 37 (2009) W652-660.
49. J. Xia, R. Mandal, I.V. Sinelnikov, D. Broadhurst, D.S. Wishart, “*MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis*”, *Nucleic Acids Res*, 40 (2012) W127-133.
50. J. Xia, I.V. Sinelnikov, B. Han, D.S. Wishart, “*MetaboAnalyst 3.0-making metabolomics more meaningful*”, *Nucleic Acids Res*, (2015).
51. Waters, “*MarkerLynx*”.

52. A. Lommen, H.J. Kools, “*MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware*”, *Metabolomics*, 8 (2012) 719-726.
53. T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, “*MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data*”, *BMC Bioinformatics*, 11 (2010) 395.
54. M. Sturm, A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, O. Kohlbacher, “*OpenMS - an open-source software framework for mass spectrometry*”, *BMC Bioinformatics*, 9 (2008) 163.
55. R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, D. S., E. B., L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Lacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C.A. Smith, G. Smyth, L. Tierney, Y. J.Y.H., J. Zhang, “*Bioconductor: open software development for computational biology and bioinformatics*”, *Genome Biol.*, 5 (2004).
56. S. Bocker, M.C. Letzel, Z. Liptak, A. Pervukhin, “*SIRIUS: decomposing isotope patterns for metabolite identification*”, *Bioinformatics*, 25 (2009) 218-224.
57. M. Loos, J. Hollender, E. Schymanski, M. Ruff, H. Singer, “*Bottom-up peak grouping for unknown identification from high-resolution mass spectrometry data.*”, *ASMS 2012 annual conference, Vancouver.*, (2012).
58. S. Kern, K. Fenner, H. Singer, R. Schwarzenbach, J. Hollender, “*Identification of Transformation Products of Organic Contaminants in Natural Waters by Computer-Aided Prediction and High-Resolution Mass Spectrometry*”, *Environ Sci Technol*, 43 (2009) 7039-7046.
59. G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, “*IPO: a tool for automated optimization of XCMS parameters*”, *BMC Bioinformatics*, 16 (2015) 118.
60. N. Wang, R.C. Buck, B. Szostek, L.M. Sulecki, B.W. Wolstenholme, “*5:3 Polyfluorinated acid aerobic biotransformation in activated sludge via novel "one-carbon removal pathways"*”, *Chemosphere*, 87 (2012) 527-534.
61. Y.C. Tai, T.P. Speed, “*A multivariate empirical Bayes statistic for replicated microarray time course data*”, *The Annals of Statistics*, 34 (2006) 2387-2412.
62. Y.C. Tai, “*The timecourse Package*”, *R vignettes*, (2015).
63. R. Aalizadeh, “*Ανάπτυξη και επικύρωση μοντέλων πρόβλεψης του χρόνου ανάσχεσης για την ταυτοποίηση αναδιδόμενων ρύπων σε περιβαλλοντικά δείγματα με μη στοχευμένη σάρωση και τεχνικές φασματομετρίας μαζών υψηλής διακριτικής ικανότητας*”, Τμήμα Χημείας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Αθήνα, 2015.