ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη βάσης δεδομένων για αποθήκευση και επεξεργασία δεδομένων βιοχημικής ανάλυσης που χρησιμοποιούνται για μοντελοποίηση βιοχημικών δικτύων

ΝΙΚΟΛΑΟΣ Γ. ΡΑΠΤΗΣ

ΕΠΙΒΛΕΠΩΝ: ΣΩΤΗΡΗΣ Π. ΧΡΙΣΤΟΔΟΥΛΟΥ, Επίκουρος Καθηγητής

ΠΑΤΡΑ 2021

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Πάτρα, …/…/2021

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Σωτήρης Χριστοδούλου, Επίκουρος Καθηγητής
Τμήμα Ηλεκτρολόγων Μηχανικών κ Μηχανικών Υπολογιστών
Πανεπιστήμιο Πελοποννήσου

2. Μαρία Ι. Κλάπα, Κύρια Ερευνήτρια
Εργ. Μεταβολικής Μηχανικής κ Συστημικής Βιολογίας
ΙΤΕ/ΙΕΧΜΗ

3. Αλέξανδρος Καλαράκης, Επίκουρος Καθηγητής
Τμήμα Μηχανολόγων Μηχανικών
Πανεπιστήμιο Πελοποννήσου

Υπεύθυνη Δήλωση Φοιτητή

Βεβαιώνω ότι είμαι συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τη συγκεκριμένη εργασία.
Η έγκριση της διπλωματικής εργασίας από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Περίληψη

Η Μεταβολομική αποτελεί μία τεχνική υψηλής απόδοσης στο πεδίο της Συστημικής Βιολογίας, η οποία αναλύει την φυσιολογία βιολογικών συστημάτων σε μεταβολικό επίπεδο. Αφορά τον ταυτόχρονο προσδιορισμό και ποσοτικοποίηση της σχετικής συγκέντρωσης μεγάλου αριθμού ελεύθερων μεταβολιτών μικρού μοριακού βάρους υπό συγκεκριμένες συνθήκες φυσιολογίας. Πρόκληση στην ανάλυση αποτελούν το πλήθος των μεταβολικών μορίων και οι πολλαπλές χημικές ομάδες οι οποίες ανήκουν. Η προτυποποιημένη βιβλιοθήκη μεταβολικών κορυφών που έχει αναπτυχθεί από το Εργαστήριο Μεταβολικής Μηχανικής και Συστημικής Βιολογίας (MESBL) του ΙΤΕ/ΙΕΧΜΗ χρησιμοποιείται για την αναγνώριση μεταβολικών μορίων από κορυφές που παράγονται μέσω Χρωματογραφίας Αερίων – Φασματομετρίας Μάζας. Η βιβλιοθήκη κορυφών στην παρούσα μορφή της βρίσκεται σε μία μορφή πίνακα μερικώς δομημένων δεδομένων, η οποία περιορίζει την χρήση της. Μεταφράζοντας το βιολογικό πρόβλημα και μέσω τεχνικών κανονικοποίησης, στην παρούσα πτυχιακή εργασία σχεδιάζουμε ένα σχεσιακό σχήμα βάσης για την επόμενη έκδοση της βιβλιοθήκης κορυφών και μεταφέρουμε σε αυτήν τα υπάρχοντα δεδομένα. Υλοποιούμε επιπλέον, μία αντικειμενοστραφή – σχεσιακή αντιστοίχιση μοντέλων οντοτήτων μέσω της βιβλιοθήκης ιστού Django σε γλώσσα προγραμματισμού Python, καθώς και ένα περιβάλλον διαχείρισης πάνω σε αυτό. Η βάση δεδομένων θα αξιοποιηθεί για αυτοματοποιημένη επεξεργασία και ανάλυση βιοχημικών δεδομένων που χρησιμοποιούνται για την μοντελοποίηση βιοχημικών δικτύων, και στο μέλλον θα αποτελέσει τον πυρήνα μίας δημόσιας διαδικτυακής εφαρμογής και υπηρεσίας.

Abstract

Metabolomics is a high-performance technique in the field of Systems Biology, which analyzes the physiology of biological systems at the metabolic level. It concerns the simultaneous determination and quantification of the relative concentration of large numbers of free metabolites of low molecular weight under specific physiological conditions. The challenge in the analysis is the multitude of metabolic molecules and the multiple chemical groups that they belong to. The standardized library of metabolic peaks developed by the Metabolic Engineering and Systems Biology Laboratory (MESBL) of FORTH/ICEHT is used to identify metabolic molecules from peaks produced through Gas Chromatography – Mass Spectrometry. The peak library in its current form is captured in a table format of partially structured data, which limits its use. By translating the biological problem and through normalization techniques, in this thesis we design a relational base schema for the next edition of the peak library and migrate the existing data into it. We are also implementing object - relational mapper entity models through the Django web framework in the Python programming language, as well as an administration application through them. The database will be used for automated processing and analysis of biochemical data for modelling biochemical networks, and which, in the future, will serve as the core of a public web application and service.

Αφιερωμένο στην μητέρα μου, Μαίρη.

Ευχαριστίες

Ευχαριστώ θερμά το εργαστήριο Μεταβολικής Μηχανικής και Συστημικής Βιολογίας του ΙΤΕ/ΙΕΧΜΗ που με αγκάλιασε και μου έδωσε την δυνατότητα να κάνω αυτήν την εργασία και να ασχοληθώ με την Βιοπληροφορική και συγκεκριμένα την επιστημονική υπεύθυνη Κύρια Ερευνήτρια Β' Δρ. Μαρία Κλάπα που με εμπιστεύτηκε και έδωσε το αμέριστο ενδιαφέρον της για την φέρουμε εις πέρας. Ένα ακόμα ευχαριστώ στην Δρ. Γεωργία Τοουλάκου που με "έμπλεξε", φέροντας με σε επαφή με την πανέμορφη ομάδα της. Ευχαριστώ τέλος τον οργανισμό ELIXIR-Greece, που με την οικονομική τους στήριξη στο έργο του εργαστηρίου έκαναν δυνατή και την εκπόνηση αυτής της εργασίας.

Ευχαριστώ επίσης για την ενθουσιώδη κατανόηση και την συνεχή στήριξη τους τους ανθρώπους της Citrix, της εταιρείας που εργάζομαι και συγκεκριμένα τον Γιώργο Παπαλουκόπουλο, Γιώργο Πανίτσα και την Κατερίνα Καλού που τα τελευταία χρόνια είναι ταυτόχρονα μέντορες μου στο τι σημαίνει επαγγελματίες, όσο και οι άνθρωποι που μου στέκονται στα δύσκολα.

Η συγγραφή αυτής εργασίας συνέπεσε με μία από τις πιο δύσκολες περιόδους της ζωής μου, τον χαμό της μητέρας μου. Θέλω να ευχαριστήσω αυτήν, τον πατέρα μου Γιώργο και την αδερφή μου Αθανασία για τα σαράντα χρόνια αγάπης που έχουμε μοιραστεί και συνεχίζουμε και μοιραζόμαστε.

Πίνακας Περιεχομένων

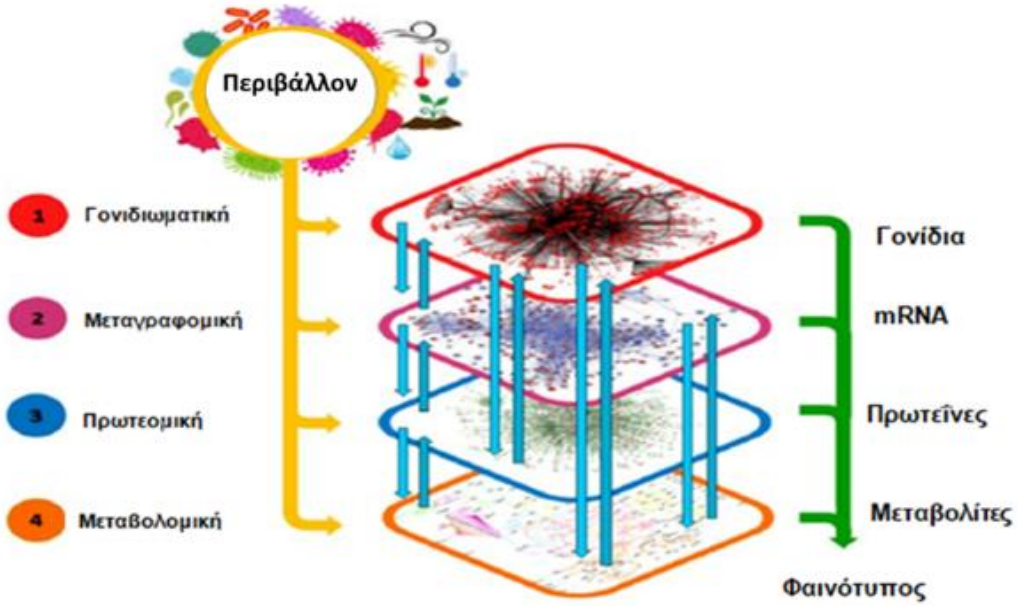
Κατάλογος Πινάκων και Εικόνων

Κεφάλαιο 1. Εισαγωγή

1.1 Συστημική Βιολογία

Η Συστημική Βιολογία (Systems Biology) αποτελεί ένα νέο πεδίο και αναφέρεται στην ποσοτική ανάλυση των δυναμικών αλληλεπιδράσεων των διαφόρων επιπέδων κυτταρικής έκφρασης ενός βιολογικού συστήματος και έχει ως στόχο την ολιστική κατανόηση του συστήματος. Από τα τέλη του τελευταίου αιώνα έχουν προκύψει τεράστιες ανακαλύψεις στη Βιολογία, όσων αφορά το γονιδίωμα, την μετάφραση του μέσω RNA σε πρωτεΐνες και τον ρόλο αυτών τελικά στην βιοχημεία και την μεταβολική του κυττάρου. Σε όλα αυτά τα επίπεδα γίνεται, έως και σήμερα σημαντική έρευνα, πολλές φορές όμως χωρίς αυτή να μεταφέρεται μεταξύ των διαφορετικών πεδίων. Η Συστημική Βιολογία καλείται να γεφυρώσει τα πεδία αυτά και αναφέρεται στην ποσοτική ανάλυση των δυναμικών αλληλεπιδράσεων των διαφόρων συστατικών ενός βιολογικού συστήματος. Ταυτόχρονα, προσπαθεί να συσχετίσει φαινόμενα που συμβαίνουν ταυτόχρονα στα επίπεδα του συστήματος, χωρίς προηγούμενη γνώση αλληλεπιδράσεων. Είναι ικανή επομένως, να ανακαλύψει μηχανισμούς από τέτοια φαινόμενα, τα οποία αρχικά φαίνονται ασύνδετα.

Οι εξελίξεις στον τομέα της καταγραφής του γονιδιώματος καθώς και οι ταχύτατες εξελίξεις ομικών (omics) τεχνολογιών είναι στην καρδιά της Συστημικής Βιολογίας. Συγκεκριμένα, οι ομικές τεχνολογίες, αλλιώς γνωστές ως τεχνολογίες υψηλής απόδοσης (high-throughput). Η ονομασία αυτή προέρχεται από το γεγονός πως μπορούν να παράξουν μεγάλες ποσότητες δεδομένων από μικρή ποσότητα δείγματος, και μπορούν να προσδιορίσουν το σύνολο των μορίων που εκφράζονται σε κάθε επίπεδο λειτουργίας ενός κυττάρου (Klapa & Quackenbush, 2003). Μας επιτρέπουν να μεταπηδήσουμε από μία λογική απλής καταγραφής μορίων και τον χαρακτηρισμό τους, σε μία μελέτη της δυναμικής, λειτουργικής δραστηριότητας των βιολογικών συστημάτων. Η Συστημική Βιολογία ασχολείται με τέσσερα επίπεδα της ζωής του κυττάρου. Την γονιδιωματική (genomics), η οποία αφορά την μελέτη της γονιδιακής αλληλουχίας στο DNA. Την μεταγραφωμική (transcriptomics), η οποία ασχολείται με την μεταγραφή των γονιδίων σε πρωτεΐνες. Την πρωτε(ιν)ωματική (proteomics), που μελετάει το προφίλ των πρωτεϊνών που αντιστοιχούν σε μία δεδομένη στιγμή στο κύτταρο. Τέλος, η πρόσφατη από τις τεχνολογίες και αυτή που αφορά την εργασία, η μεταβολομική (metabolomics) καλείται να αναλύσει τα μεταβολικά μόρια.



Εικόνα 1 – Τα επίπεδα της κυτταρικής λειτουργίας. Από το μοριακό επίπεδο στον φαινότυπο (Claude Y. Hamany Djande et al., 2020)

Σε κάθε ένα από τα επίπεδα κυτταρικής λειτουργίας, η αντίστοιχη τεχνολογία παράγει δεδομένα όσον αφορά τα μόρια τα οποία μετέχουν στο αντίστοιχο επίπεδο, όπως και για τις συγκεντρώσεις και τις αλληλεπιδράσεις τους. Αυτά τα δεδομένα οργανώνονται σε μορφή δικτύων, στα οποία τα μόρια μετέχουν ως κόμβοι, με ακμές τις μεταξύ τους αλληλεπιδράσεις. Εναλλάσσοντας την κατάσταση του βιολογικού συστήματος, για παράδειγμα αναλύοντας κύτταρα υγιών και ενός νοσηρών ιστών, καταλήγουμε σε παραλλαγές των ποσοτικών τιμών του ίδιου δικτύου. Η συστηματική βιολογία σε συνδυασμό με την μαθηματική μοντελοποίηση και την θεωρία ανάλυσης δικτύων έδωσαν τη δυνατότητα ολιστικής μελέτης α) των διαφορετικών επιπέδων κυτταρικής λειτουργίας και β) ενός βιολογικού συστήματος (Kell, 2006).

Τελικώς, η αναπαράσταση του βιολογικού συστήματος ως ένα πολυεπίπεδο δίκτυο δίνει την δυνατότητα να εφαρμοστούν σε αυτό διάφορες υπολογιστικές τεχνικές, πολλές φορές αυτοματοποιημένες, μέσω των οποίων είναι δυνατόν να ανακαλυφθούν δυναμικές συμπεριφορές του συστήματος που δεν θα ήταν εμφανείς μέσω αναγωγικής μελέτης της λειτουργίας κάθε μορίου ξεχωριστά (Vidal, Cusick, & Barabási, 2011). Η ικανότητα αυτή της Συστημικής Βιολογίας να μπορεί να δημιουργεί μοντέλα αλληλεπιδράσεων χωρίς πρότερη γνώση, είναι και το μεγαλύτερο πλεονέκτημά της.

1.1.1 Μεταβολομική Ανάλυση

Η υψηλής απόδοσης ανάλυση της φυσιολογίας ενός Βιολογικού Συστήματος σε μεταβολικό επίπεδο ονομάζεται Μεταβολομική. Αφορά μεταβολικά μόρια μικρού μοριακού βάρους (αμινοξέα, σάκχαρα, λιπαρά οξέα, δευτερογενείς μεταβολίτες), την οργάνωση τους σε μεταβολικά δίκτυα και την ποσοτικοποίηση τους σε προφίλ σχετικής συγκέντρωσής, υπό συγκεκριμένες συνθήκες φυσιολογίας. Το επίπεδο του μεταβολισμού, στο οποίο αναφέρεται, είναι ενδιαφέρον ως αυτό που συνδέει την γονιδιακή μεταγραφή και μετάφραση με τον φαινότυπο και την φυσιολογία ενός βιολογικού συστήματος. Η ανάλυση μπορεί να γίνει είτε στοχευμένα, όπου επιλέγουμε συγκεκριμένους, ήδη γνωστούς μεταβολίτες, είτε μη στοχευμένα όπου μας ενδιαφέρει η ανάλυση όσων περισσότερων μεταβολιτών μπορούν να ποσοτικοποιηθούν (Papadimitropoulos et al., 2018).

Συγκεκριμένα πλεονεκτήματα της μεταβολομικής ανάλυσης την κάνουν ελκυστική. Το μεταβολικό δίκτυο δεν είναι αναγκαίο να είναι πλήρως γνωστό εκ των προτέρων και μπορεί να ανακατασκευαστεί από τα ίδια τα δεδομένα της ανάλυσης (Kanani, Chrysanthopoulos, & Klapa, 2008). Μπορεί να εφαρμοστεί σε διάφορες συνθήκες φυσιολογίας ενός βιολογικού συστήματος, δίνοντας την δυνατότητα να αναλυθούν διαφορές μεταξύ συστημάτων υπό νόσηση ή καταπόνηση με υγιείς. Δεν απαιτεί τόσο εξειδικευμένο εξοπλισμό και κατά συνέπεια μπορεί να χρησιμοποιηθεί με χαμηλό κόστος και επίσης να χρησιμοποιηθεί είτε κλινικά σε νοσοκομεία είτε σαν μέσο προτυποποίησης σε μονάδες αγροτικής παραγωγής (Τοουλάκου, 2013). Οι μεταβολίτες δεν διαφέρουν σε μεγάλο βαθμό μεταξύ ιστών, και επομένως τεχνικές και συμπεράσματα στο μεταβολομικό επίπεδο μπορούν να μεταφερθούν σε μεγαλύτερο βαθμό από ότι αυτά άλλων επιπέδων. Τέλος, επειδή είναι ευαίσθητος στο περιβάλλον του κυττάρου, είναι κατάλληλος για την ανάλυση της δυναμικής απόκρισης ενός βιολογικού συστήματος τόσο σε γενετικές αλλαγές όσο και διαφορετικές περιβαλλοντικές συνθήκες.

Παρά τα πολλαπλά πλεονεκτήματά της, η μεταβολομική ανάλυση έχει και μειονεκτήματα, τα οποία αφορούν, κυρίως, την ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών. Οι μεταβολίτες λαμβάνουν μέρος σε πολλαπλά μεταβολικά μονοπάτια, το οποίο κάνει την ερμηνεία των αποτελεσμάτων της ανάλυσης δύσκολη. Το αρνητικό όμως που θα μας απασχολήσει περισσότερο από όλα στην εργασία είναι πως υπάρχει

μεγάλος αριθμός από μεταβολίτες που συμμετέχουν στα μεταβολικά δίκτυα και πως αυτοί ανήκουν σε πολύ διαφορετικές χημικές ομάδες. Σαν αποτελέσματα, δεν υπάρχει τεχνική η οποία να μπορεί να καλύψει το σύνολο των μεταβολικών μορίων, και επίσης πολλοί μεταβολίτες σε ένα σύστημα μπορεί να παραμένουν άγνωστοι.

Στην μεταβολομική ανάλυση χρησιμοποιούνται σχετικά γνωστές τεχνικές που χρησιμοποιούνται γενικότερα για την ανίχνευση και την ποσοτικοποίηση μικρού μοριακού βάρους οργανικών μορίων. Η γενικότητα τους αυτή τις κάνει ταυτόχρονα κατάλληλες για μη στοχευμένη ανάλυση. Αυτές είναι: Η χρωματογραφία υγρών φασματομετρία μάζας (Liquid Chromatography Mass Spectrometry, LC-MS), η οποία είναι χρήσιμη για μεταβολικά μόρια τα οποία μετέχουν στο μόριο σε υγρή φάση. Η χρωματογραφία αερίων φασματομετρία μάζας (Gas Chromatography Mass Spectrometry, GC-MS) η οποία έχει μεγαλύτερη διακριτική ικανότητα από την υγρή παραλλαγή της, αλλά απαιτεί να φέρουμε τα μόρια πρώτα σε αέρια μορφή. Τέλος η φασματομετρία πυρηνικού μαγνητικού συντονισμού (Nuclear Magnetic Resonance Spectrometry, NMR) η οποία είναι κατάλληλη για τον προσδιορισμό της μορφής των μορίων και τον διαχωρισμό ισομερών μεταξύ τους. Σε πολλές περιπτώσεις χρησιμοποιούνται πάνω από μία τεχνικές ώστε να καλυφθεί ο μεγαλύτερος δυνατός αριθμός μεταβολικών μορίων.

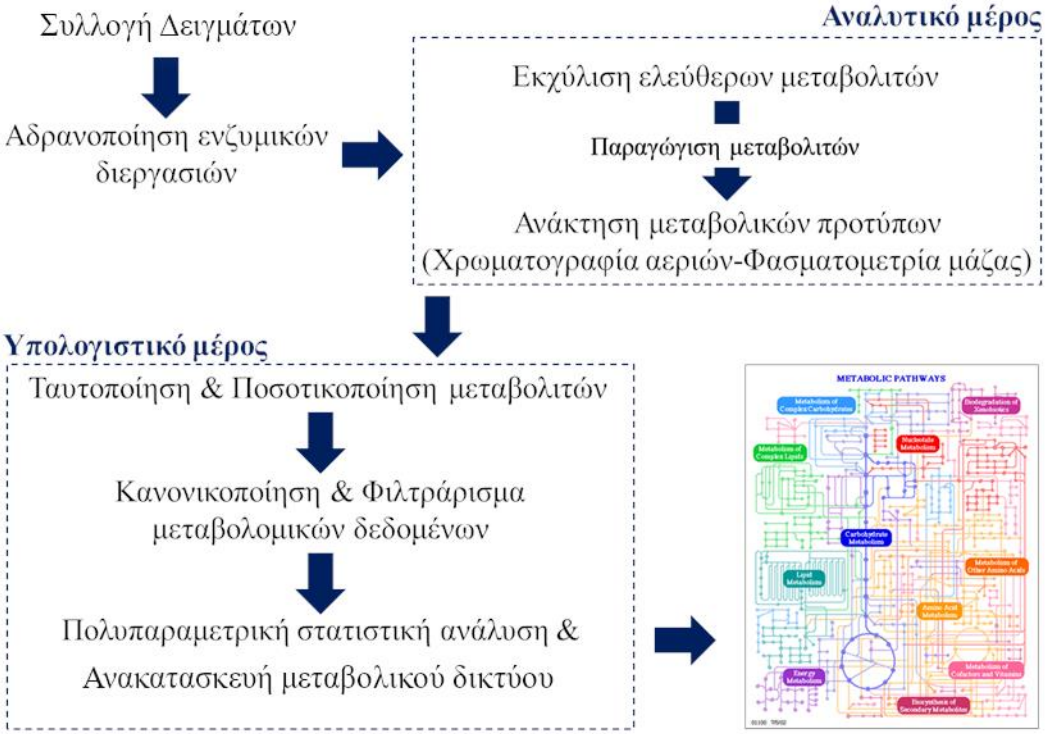
Στα πλαίσια της εργασίας, θα ασχοληθούμε με μία από αυτές τις τεχνικές, αυτήν της Χρωματογραφίας Αερίων – Φασματομετρία Μάζας (GC-MS) και την ανάγκη που δημιουργεί για την ύπαρξη μίας προτυποποιημένης βιβλιοθήκης κορυφών.

1.2 Χρωματογραφία Αερίων - Φασματομετρία Μάζας

Η μεταβολομική ανάλυση είναι μία πολυβηματική διαδικασία. Συνοπτικά, ας αναφέρουμε τα βήματα μίας ανάλυσης μέσω Χρωματογραφίας Αερίων – Φασματομετρίας Μάζας. Δεν θα δώσουμε πολλές από τις λεπτομέρειες της τεχνικής, αλλά θα σταθούμε περισσότερο στην γενική εικόνα, καθώς και στα σημεία τα οποία έχουν συνάφεια με την εργασία.

1.2.1 Προ-αναλυτικά Στάδια

Αρχικά, γίνεται συλλογή δειγμάτων από το βιολογικό σύστημα, τυπικά έναν κυτταρικό ιστό κάποιου οργανισμού. Τα μεταβολικά μόρια είναι μόρια ενεργά, όπως για παράδειγμα τα ένζυμα, οπότε μεγάλη προσοχή δίνεται στο να αδρανοποιηθούν, παγώνοντας την μεταβολική εικόνα του συστήματος. Στην συνέχεια, γίνεται εκχείλιση των μεταβολιτών, δηλαδή ελευθέρωση τους από την κυτταρική μεμβράνη. Και οι δύο διεργασίες αυτές έχουν επιλεχθεί ώστε να μην καταστρέφουν και να μην αλλάζουν την μεταβολική εικόνα, ή τουλάχιστον αν εισάγουν αλλαγές, όπως για παράδειγμα η προσθήκη ενός διαλύτη, αυτές οι αλλαγές να είναι γνωστές και προβλέψιμες.

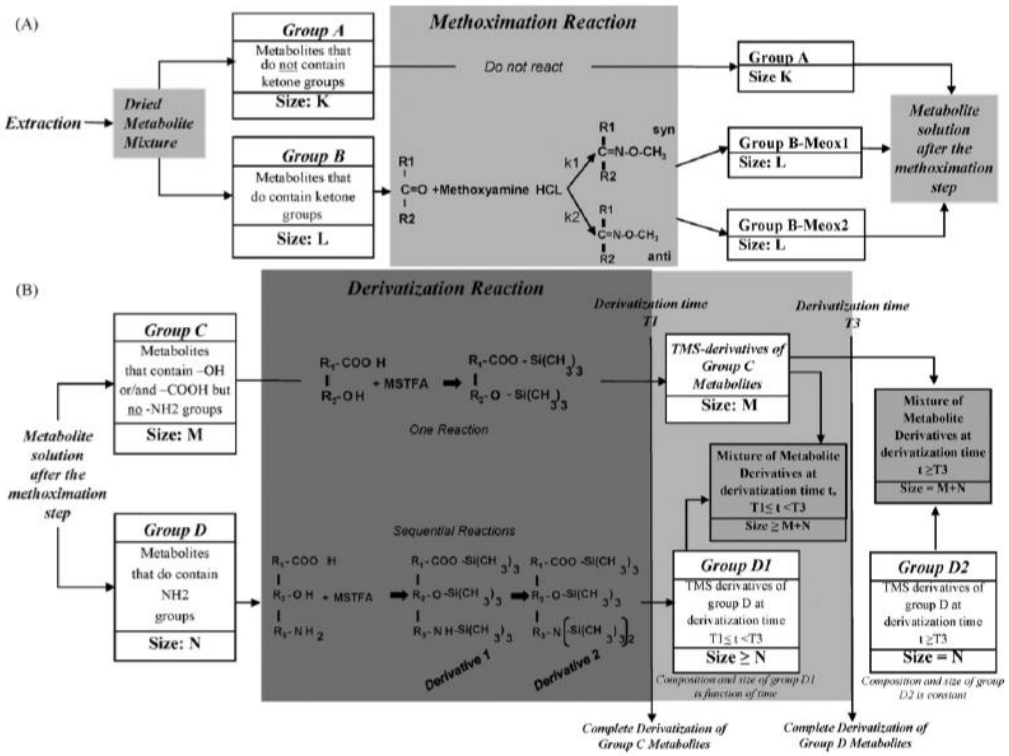


Εικόνα 2 – Η μεταβολομική ως μία πολυβηματική διαδικασία. Τα επιμέρους στάδια της μεταβολομικής ανάλυσης (Είκονα από αρχείο Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας).

1.2.2 Παραγώγιση Μεταβολιτών

Για να είναι δυνατή η ανάλυση μεταβολιτών μέσω Χρωματογραφίας Αερίων – Φασματομετρίας Μάζας, θα πρέπει τα μόρια τους να είναι αέρια ή πτητικά. Αντιθέτως, οι περισσότεροι μεταβολίτες δεν είναι πτητικοί, γεγονός που καθιστά αναγκαίο τον μετασχηματισμό τους σε πτητικά και θερμικά σταθερά μόρια. Ο μετασχηματισμός αυτός ονομάζεται παραγώγιση (derivatization) των μεταβολιτών και γίνεται με χημικό τρόπο μέσω ουσιών που ονομάζονται παράγοντες παραγώγισης.

Οι παράγοντες παραγώγισης που θα μας απασχολήσουν είναι δύο: Πρώτον το **Ν-μεθυλ-τριμεθυλπυριτο-τριφθωρο-ακεταμίδιο** (MSTFA) το οποίο αντιδρά με το **υδρογόνο** (-Η) σε **υδροξυλομάδες** (–OH), **καρβοξυλομάδες** (–COOH) και **αμινομάδες** (–NH3) και το αντικαθιστά με μία **τριμέθυλπυριτ-ομάδα** (TMS). Ταυτόχρονα χρησιμοποιείται η ουσία **μεθυλ-οξίμη** (Methoxime, MeOx), η οποία αντιδρά με **κετόνες** (-C=O) και που δημιουργεί ομάδες που είναι παραγωγίσιμες από τον πρώτο παράγοντα παραγώγισης.



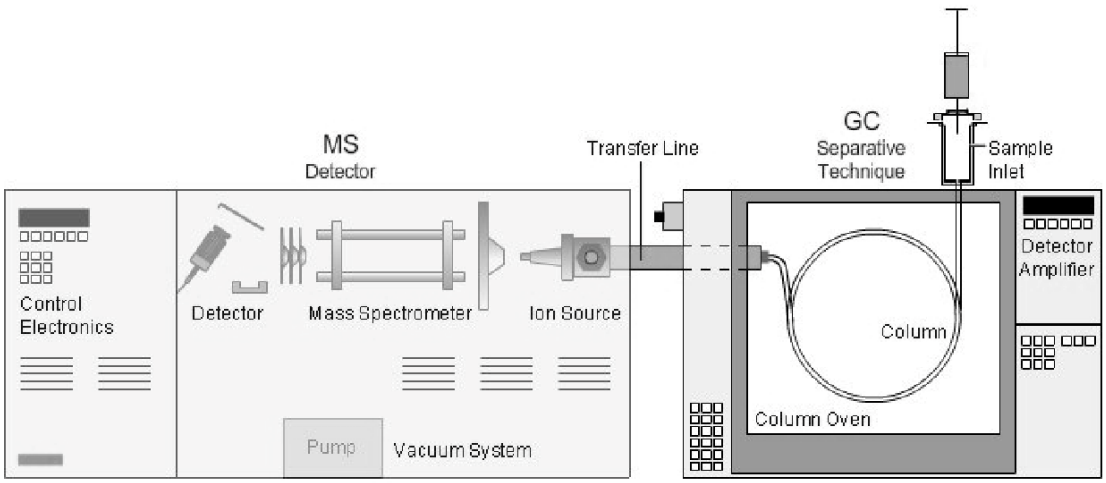
Εικόνα 3 – Κατηγορίες Παραγώγισης Μεταβολιτών
(Kanani & Klapa, 2007)

Σύμφωνα με τους (Kanani & Klapa, 2007) αυτή η μέθοδος είναι κατάλληλη για να παραγωγηθούν και να αναλυθούν τρεις ομάδες μεταβολιτών. Στην πρώτης κατηγορία, μόρια που περιέχουν **υδροξυλομάδα** (–OH) ή **καρβοξυλομάδες** (–COOH) παραγωγίζονται σε ένα μόνο παράγωγο, στην δεύτερη κατηγορία ανήκουν οι **κετόνες** (-C=O) οι οποίες λόγω του ομοιοπολικού δεσμού με το οξυγόνο παραγωγίζονται σε δύο γεωμετρικά ισομερή παράγωγα, σε ίδια ποσοστά, κάτι που είναι χρήσιμο και στην επαλήθευση της διαδικασίας. Τέλος, οι μεταβολίτες με **αμινομάδα** (–NH3) στις οποίες η παραγώγιση λειτουργεί σειριακά παράγοντας πολλαπλά παράγωγα σε ποσοστά που εξαρτώνται από τον χρόνο παραγώγισης.

Για τους σκοπούς μας, μπορούμε να συγκρατήσουμε πως ανάλογα σε ποια από τις τρεις κατηγορίες ανήκει ο μεταβολίτης, θα μετατραπεί σε ένα, δύο ή περισσότερα παράγωγα.

1.2.3 Ανάκτηση Μεταβολικού Προτύπου με χρήση GC-MS

Το GC-MS αποτελείται από δύο στοιχεία, τον χρωματογράφο αερίων και το φασματόμετρο μάζας. Το μείγμα των μεταβολιτών, πλέον πτητικοί μετά την παραγώγιση, εισάγεται στον χρωματογράφο αερίων και εισέρχεται στην χρωματογραφική στήλη. Με την αύξηση της θερμοκρασίας σε αυτήν, ένα μετά το άλλο τα παράγωγα περνούν στην αέρια φάση και παρασύρονται μέσω ενός αδρανές μέσου, συνήθως ηλίου, προς το Φασματόμετρο Μάζας. Η σειρά με την οποία τα μόρια εξέρχονται της χρωματογραφικής στήλης τείνει να είναι βάση του μοριακού βάρους και της δομής τους, με μικρά και γραμμικά μόρια να προηγούνται.



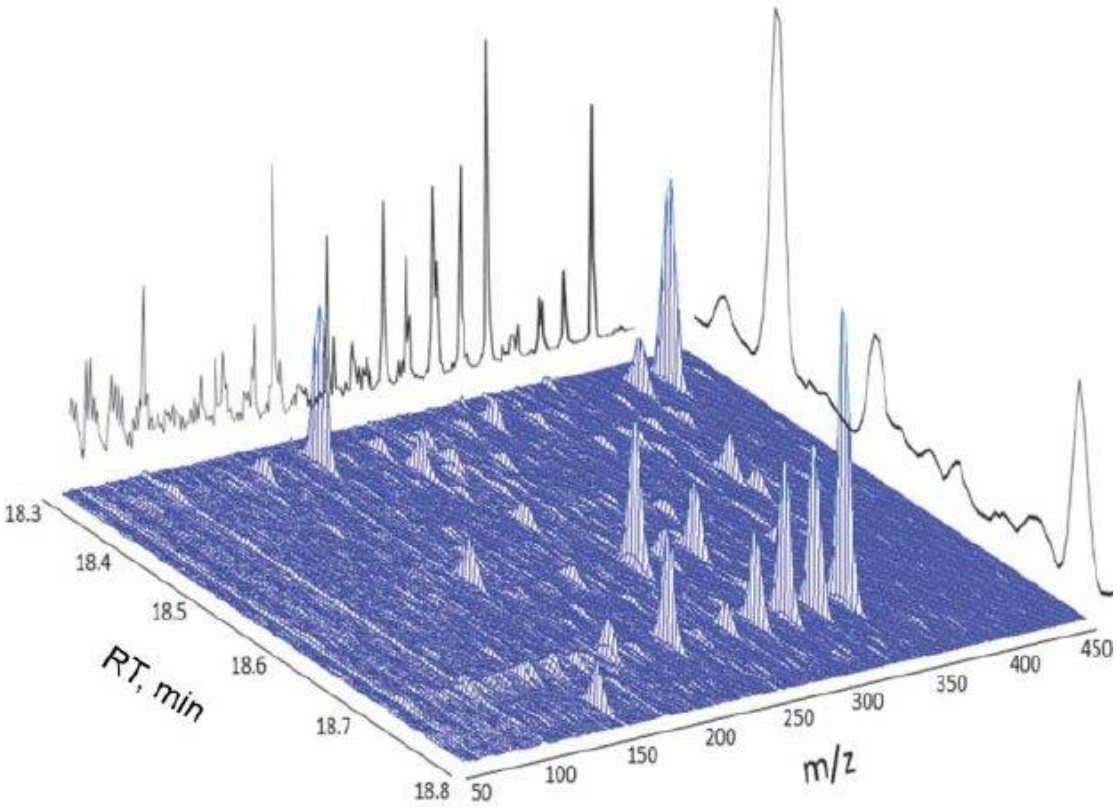
Εικόνα 4 – Διάταξη GC-MS

(Είκονα από αρχείο Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας)

Χαρακτηριστικά, η μέτρηση η οποία μας ενδιαφέρει για το κάθε μόριο είναι αυτός του χρόνου παραμονής (retention time) του στην χρωματογραφική στήλη, ο οποίος μας δίνει και των έναν από τους άξονες των εξαγομένων δεδομένων.

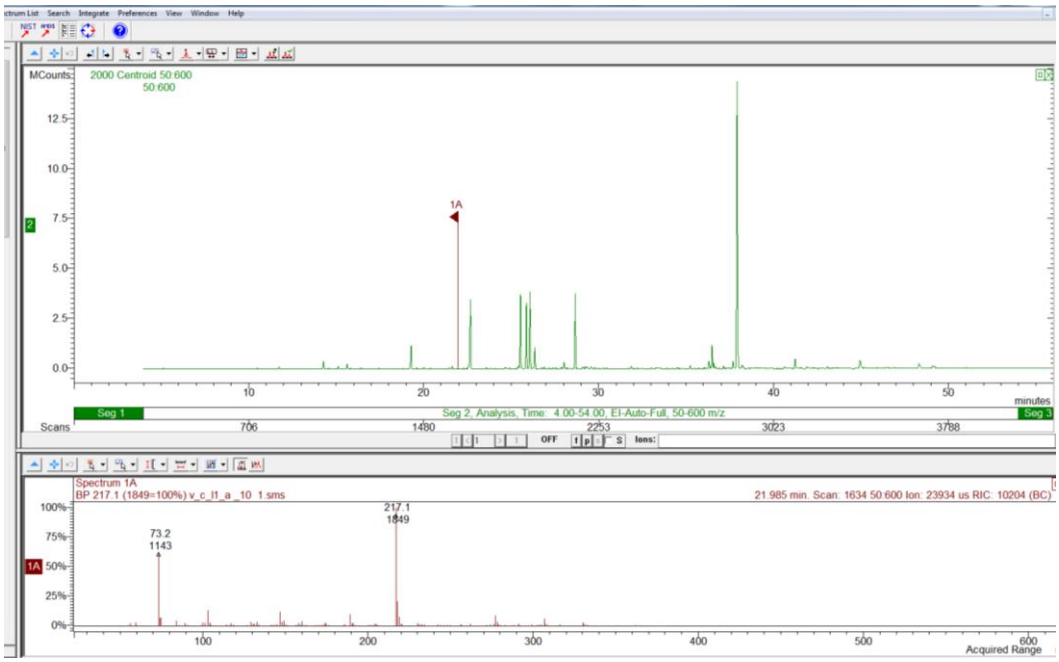
Στη διάρκεια της ανάλυσης, τα μόρια που έχουν διαχωριστεί μέσω της Χρωματογραφίας Αερίων εισέρχονται στο Φασματόμετρο Μάζας. Το πρώτο μέρος του που συναντούν ονομάζεται ιοντική παγίδα. Αυτή βομβαρδίζει τα μόρια με δέσμες ηλεκτρονίων, διασπώντας τα σε έναν αριθμό από φορτισμένα ιόντα, τα οποία και είναι χαρακτηριστικά του μορίου. Έπειτα, περνούν στον αναλυτή μάζας, όπου, ανάλογα κινούνται μέσα σε ένα ηλεκτρομαγνητικό πεδίο, διαγράφοντας τροχιές που είναι ανάλογες του λόγου μάζας/φορτίου (m/z) που έχουν. Τέλος προσπίπτουν στον ανιχνευτή όπου και ποσοτικοποιούνται. Τα περισσότερα από αυτά τα ιόντα έχουν μοναδιαίο φορτίο (z=1) οπότε πολλές φορές μπορεί να χρησιμοποιήσουμε τον όρο μοριακό βάρος αντί για τον λόγο μάζας/φορτίου.

Τελικώς τα αποτελέσματα της ανάλυσης παίρνουν τη μορφή τρισδιάστατων δεδομένων, όπου ο πρώτος άξονας αντιστοιχεί στον χρόνο παραμονής στην στήλη του Χρωματογράφου Αερίων και μετριέται συνήθως σε λεπτά. Για κάθε χρονικό σημείο, ανάλογα και με την διαχωριστική ικανότητα του εξοπλισμού, η τομή των τρισδιάστατων δεδομένων μας δίνει ένα γράφημα της σχετικής ποσότητας των ιόντων ανάλογα με τον λόγο μάζας/φορτίου του. Αυτή η μορφή των δεδομένων, μας δίνει την δυνατότητα να εστιάσουμε στα σημεία του χρωματογραφήματος, τα οποία αντιστοιχούν σε παράγωγα μόρια μεταβολιτών, και έπειτα να καταγράψουμε για το καθένα από αυτά το φασματομετρικό προφίλ μάζας των ιόντων που το αποτελούν.



Εικόνα 5 – Τρισδιάστατη αναπαράσταση δεδομένων GC-MS

(Είκονα από αρχείο Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας)



Εικόνα 6 – Χρωματογράφημα (επάνω) και το φάσμα μάζας (κάτω μέρος) επιλεγμένης κορυφής

(Είκονα από αρχείο Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας)

1.2.4 Ταυτοποίηση κορυφών και ποσοτικοποίηση

Βάσει των δεδομένων από την Χρωματογραφία Αερίων – Φασματομετρία Μάζας, καλούμαστε να αναγνωρίσουμε τους μεταβολίτες οι οποίοι παρευρίσκονταν στο δείγμα που αναλύθηκε. Μετά την ανάκτηση του μεταβολικού προτύπου, η ταυτοποίηση και ποσοτικοποίηση γίνεται μέσω λογισμικού βάσει τον χρόνο παραμονής και το ιοντικό προφίλ της κάθε μίας. Στηρίζεται κατά κύριο λόγο στην ύπαρξη μίας βιβλιοθήκης από αναλυτικά πρότυπα γνωστών μεταβολιτών, με τα οποία συγκρίνονται τα αναλυτικά δεδομένα. Στην περίπτωση της δραστηριότητας του Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας (MESBL) του ΙΤΕ/ΙΕΧΜΗ, χρησιμοποιείται μία πρότυπη και επικυρωμένη βιβλιοθήκη κορυφών, η οποία έχει συσταθεί από δεδομένα εμπορικών βιβλιοθηκών και πειραματικές μετρήσεις πρότυπων δειγμάτων μεταβολιτών που έχουν διεξαχθεί στο εργαστήριο, και η κανονικοποίηση της οποίας αποτελεί το αντικείμενο αυτής της εργασίας.

Η βιβλιοθήκη κορυφών, έχει χρησιμοποιηθεί για την υλοποίηση του λογισμικού αυτοματοποιημένης μεταβολομικής ανάλυσης M-IOLITE (Maga-Nteve & Klapa, 2016; Μάγγα-Ντεβέ, 2017) και αντιπροσωπεύει ένα σημαντικό ερευνητικό εργαλείο που έχει προέλθει από την ερευνητική δραστηριότητα του εργαστηρίου. Ένα χαρακτηριστικό της είναι πως προσανατολίζεται στον μεταβολίτη (metabolite-centric), σε αντίθεση με την δεύτερη επιλογή που επικεντρώνεται σε ποσοτικοποιήσιμα χαρακτηριστικά (feature-centric). Ο διαχωρισμός των δύο έγκειται στην ποσοτικοποίηση των μεταβολιτών μέσω των δεδομένων. Στο metabolite-centric μοντέλο, επιλέγεται ένα ιόν από το φασματομετρικό προφίλ ιόντων της κάθε κορυφής, και αυτό είναι που χρησιμοποιείται για την ποσοτικοποίηση του αντίστοιχου μεταβολίτη, παίρνοντας επίσης υπόψιν πως διαφορετικές κατηγορίες μεταβολιτών θα εμφανιστούν λόγω πολλαπλών παραγώγων ως πάνω από μία κορυφές. Το ιόν αυτό ονομάζεται χαρακτηριστικό ιόν, και είναι άλλη μία από τις πληροφορίες που προσφέρει η βάση κορυφών.

1.2.5 Ανακατασκευή του Μεταβολικού Δικτύου

Μετά την αναγνώριση και ποσοτικοποίηση των κορυφών, γίνεται κανονικοποίηση και καθαρισμός των δεδομένων, ώστε να απαλειφθούν σφάλματα τα οποία προέκυψαν λόγω της ίδιας της αναλυτικής διαδικασίας. Τόσο η παραγώγιση, όπως και η προετοιμασία του δείγματος και οι ουσίες που μετέχουν στην ανάλυση δημιουργούν αποκλίσεις από το πραγματικό μεταβολικό δίκτυο. Έπειτα, μέσω αλγορίθμων πολυπαραμετρικής στατιστικής ανάλυσης, οπτικοποιείται και ανακατασκευάζεται το μοντέλου του μεταβολικού δικτύου, με την βοήθεια επίσης και χαρτών γνωστών μεταβολικών δικτύων. Τελικώς το μεταβολικό δίκτυο που παράγεται, χρησιμοποιείται για να εξαχθούν βιολογικά συμπεράσματα για το μελετώμενο σύστημα.

1.3 Η Βιβλιοθήκη Κορυφών ως μέρος του Παγκόσμιου Ιστού Δεδομένων

Οι τεχνικές υψηλής απόδοσης παράγουν και αναλύουν τεράστιο όγκο δεδομένων. Τυπικά, τα δεδομένα των ερευνών ενός εργαστηρίου συνηθιζόταν να παραμένουν σε αυτό, και να γίνεται έκδοση μόνο τον αποτελεσμάτων. Στον σημερινό κόσμο, όμως, γίνεται τεράστια προσπάθεια ώστε τόσο τα δεδομένα όσο και τα αποτελέσματα των ερευνών να διατεθούν και να είναι ελεύθερα προς χρήση από όλη την επιστημονική κοινότητα.

Αυτή είναι και η αποστολή οργανώσεων ανά τον κόσμο όπως είναι το ELIXIR και το ELIXIR-Greece. Έχοντας ιδρυθεί το 2013 και συνεργαζόμενο με πάνω από 220 ερευνητικούς φορείς, το ELIXIR έχει σαν αποστολή του να ενώσει βάσεις δεδομένων, υπολογιστική ισχύ και εργαλεία για να δημιουργήσει μία συνεκτική ευρωπαϊκή υποδομή για την εκπαίδευση και έρευνα στο πεδίο των επιστημών ζωής.

Η ένωση παρ' όλ' αυτά της πληροφορίας από έναν μεγάλο αριθμό πηγών, παρουσιάζει πολλές καινούργιες προκλήσεις. Τα δεδομένα, μπορεί να δομούνται διαφορετικά σε κάθε πηγή. Τα πειραματικά πρωτόκολλα μπορεί επίσης να διαφέρουν. Η μεγαλύτερη ίσως πρόκληση στην συνένωση της πληροφορίας είναι η έλλειψη κοινών αναγνωριστικών στα δεδομένα, μέσω τον οποίων να μπορεί η συνένωση να επιτευχθεί. Ακόμα και κάτι απλό, όπως ένα μεταβολικό μόριο που μπορεί να εμφανίζεται στην

βιβλιογραφία της χημείας, σε μία μεταβολομική ανάλυση και ταυτόχρονα μία πρωτεομική ανάλυση, μπορεί να αναφέρεται σε καθένα από αυτά με εντελώς διαφορετικό όνομα, ανάλογα με το πεδίο που το εξετάζει. Ακόμα χειρότερα, πολλά από αυτά τα ονόματα μπορεί να αντιστοιχούν σε περισσότερα από ένα μόρια. Το ίδιο πρόβλημα μπορεί επίσης να εμφανιστεί ακόμα και σε μικρότερη κλίμακα, ακόμα και ανάμεσα στα δεδομένα ερευνών του ιδίου εργαστηρίου.

Για την επίλυση της πρόκλησης της κοινής ονοματοδοσίας, έχουν αρκετά ιδρυθεί μία σειρά από βιβλιοθήκες, βάσεις οι οποίες λειτουργούν ως σημεία αναφοράς για πολλά είδη βιολογικών οντοτήτων. Ένα πολύ καλό παράδειγμα αυτών είναι η βιβλιοθήκη χημικών οντοτήτων βιολογικού ενδιαφέροντος, ChEBI, η οποία περιέχει πληροφορίες για πάνω από εξήντα χιλιάδες διακριτές χημικές ουσίες, μαζί με συνώνυμα, οντολογικές σχέσεις και συνδέσμους προς άλλες βιβλιοθήκες όπου η κάθε ουσία εμφανίζεται. Η ιδιότητα της, όπου αναφερόμενοι σε μία εγγραφή της ChEBI έχουμε υποδείξει μονοσήμαντα μία ουσία, την καθιστά, μαζί με άλλες του είδους της πολύτιμη σε έναν κόσμο δικτυωμένων μεταξύ τους δεδομένων.

Στην Μεταβολομική την οποία εξετάζουμε, αντίστοιχα υπάρχουν βάσεις αρκετών ειδών, προσφέροντας πληροφορία για μεταβολικά δίκτυα οργανισμών όπως η KEGG Atlas, μέχρι και εξειδικευμένη μεταβολική πληροφορία όπως η μεταβολομική βάση δεδομένων του ανθρώπου - Human Metabolome Database (HMDB). Μία συγκεκριμένη κατηγορία βιβλιοθηκών η οποία μας ενδιαφέρει είναι αυτές στις οποίες αναρτάται πληροφορία σχετική με χρωματογραφική – φασματομετρική ανάλυση μορίων. Σε αυτήν την κατηγορία, ανήκουν η εμπορική βιβλιοθήκη φασμάτων μάζας οργανικών μορίων από GC-MS NIST, η βιβλιοθήκη κορυφών μεταβολιτών από GC-MS GOLM καθώς και η βιβλιοθήκη μεταβολικών πειραμάτων και φασμάτων μεταβολιτών MetaboLights.

Ένα κοινό το οποίο παρουσιάζουν οι αναφερθείσες βάσεις, είναι πως προσφέρουν μία βιβλιοθήκη μεταβολιτών και για κάθε έναν από αυτούς έναν αριθμό από πειραματικά φάσματα, τα οποία ένας ερευνητής χρήστης τους καλείται να συγκρίνει με τα δικά του δεδομένα, παίρνοντας μόνος του υπόψιν όποιες πειραματικές διαφορές. Σε ένα βαθμό, μπορούμε να κατατάξουμε αυτές τις βιβλιοθήκες ως αποθετήρια πειραμάτων.

Η καινοτομία της προτυποποιημένης βιβλιοθήκης μεταβολικών κορυφών που έχει αναπτυχθεί από το Εργαστήριο Μεταβολικής Μηχανικής και Συστημικής Βιολογίας (MESBL) του ΙΤΕ/ΙΕΧΜΗ είναι ακριβώς στη προτυποποίηση των κορυφών. Αντί για να περιέχει διάφορα φάσματα για έναν αριθμό μεταβολιτών, η προτυποποιημένη βιβλιοθήκη θέτει σαν κύριες οντότητες της τα πρότυπα κορυφών, τα οποία έχουν προκύψει από συνδυασμό των πειραματικών δεδομένων του εργαστηρίου, τα δεδομένα των προαναφερθέντων εξωτερικών βάσεων και επικυρωθεί μέσω προτύπων δειγμάτων (standards) μεταβολιτών.

Το εργαστήριο, με την συνδρομή του ELIXIR-Greece, είναι σε διαδικασία αυτή τη στιγμή στο να προσθέσει έναν αριθμό από υπηρεσίες και βάσεις, ως κόμβους στον ελληνικό, ευρωπαϊκό και συνάμα παγκόσμιο ιστό βιολογικών δεδομένων που περιγράψαμε. Ήδη μετέχουν η μετα-βιβλιοθήκη πρωτεϊνικών αλληλεπιδράσεων PICKLE και το λογισμικό μεταβολομικής ανάλυσης M-IOLITE, και ετοιμάζονται στο μέλλον ένα αποθετήριο μεταβολομικών πειραματικών δεδομένων, καθώς και μία βιβλιοθήκη μοντέλων βιολογικών δικτύων. Η προτυποποιημένη βιβλιοθήκη κορυφών αποτελεί διασυνδετικό πυρήνα μεταξύ των βάσεων του εργαστηρίου, καθώς και με τον έξω κόσμο.

Κεφάλαιο 2. Στόχοι Προτυποποιημένης Βιβλιοθήκης Κορυφών

Ο απώτερος στόχος της βιβλιοθήκης κορυφών είναι να μπορέσει αυτή να αποτελέσει έναν κόμβο εφαρμογής δεδομένων, χτίζοντας γύρω της μία διαδικτυακή εφαρμογή και υπηρεσία. Για να γίνει αυτό δυνατό, η βιβλιοθήκη θα πρέπει σε πρώτη φάση να έρθει σε μία μορφή η οποία να αντιπροσωπεύει το βιολογικό πρόβλημα και τις οντότητες του. Είναι αναγκαίο να δεχτεί το σύνολο των δεδομένων, τα οποία να είναι εύκολα διαχειρίσιμα και εξελίξιμα από τους ερευνητές του εργαστηρίου.

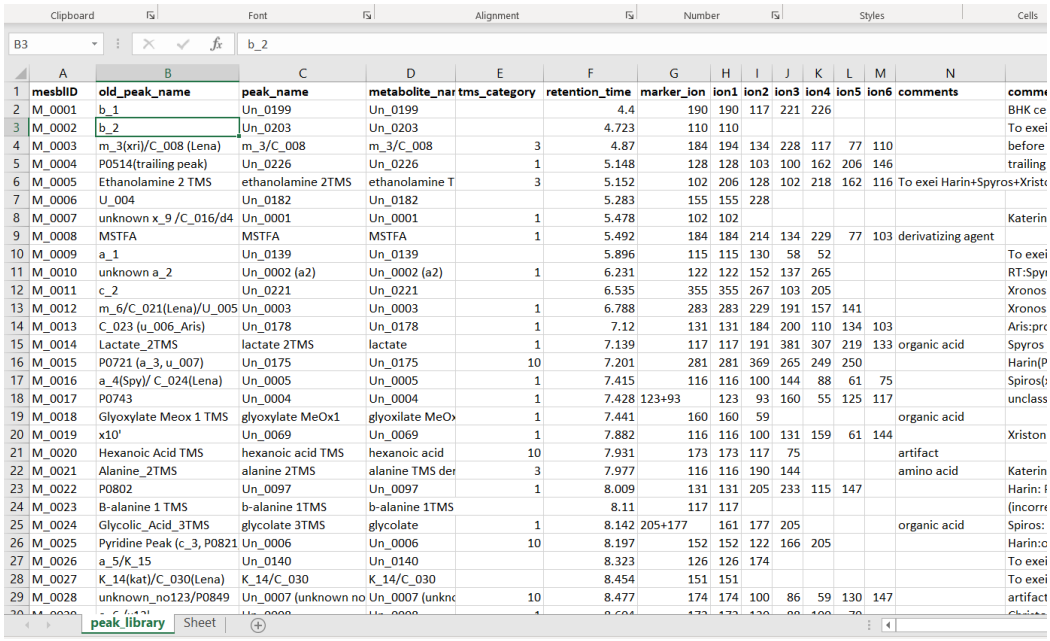
Όπως θα δούμε, η υπάρχουσα μορφή της προτυποποιημένης βιβλιοθήκης κορυφών απέχει πολύ από το να μπορεί να χρησιμοποιηθεί για αυτόν τον σκοπό. Στα πλαίσια της εργασίας καλούμαστε να σχεδιάσουμε και να υλοποιήσουμε την βιβλιοθήκη ως μία κανονικοποιημένη σχεσιακή βάση με πολλαπλές οντότητες, καθώς και να μεταφέρουμε τα υπάρχοντα δεδομένα κορυφών σε αυτήν, με σκοπό να αποτελέσει το θεμέλιο πάνω στο οποίο θα χτιστεί ένας διαδικτυακός κόμβος εφαρμογής.

Κεφάλαιο 3. Μεθοδολογία

Η πρότυπη και επικυρωμένη βιβλιοθήκη κορυφών του Εργαστηρίου Μεταβολικής Μηχανικής και Συστημικής Βιολογίας (MESBL) του ΙΤΕ/ΙΕΧΜΗ αποτελεί το αντικείμενο αυτής της εργασίας. Καλούμαστε εδώ να εξετάσουμε την μορφή της, καθώς και να αναγνωρίσουμε τα χαρακτηριστικά της τα οποία θα θέσουν προκλήσεις στον σχεδιασμό της κανονικοποιημένης μορφής της. Σε δεύτερο χρόνο, θα αναφερθούμε στο θεωρητικό τεχνολογικό υπόβαθρο που θα χρησιμοποιήσουμε ώστε να καταλήξουμε στην μεθοδολογία σύμφωνα με την οποία θα εκτελέσουμε τον σχεδιασμό.

3.1 Υπάρχουσα Προτυποποιημένη Βιβλιοθήκη Κορυφών

Η υπάρχουσα βιβλιοθήκη κορυφών αποτελείται από έναν μοναδικό πίνακα δεδομένων, στον οποίο αποθηκεύεται η πληροφορία που αφορά όλες τις πρότυπες κορυφές. Ο πίνακας αυτός διανέμεται και επεξεργάζεται μέσω ενός αρχείου λογιστικού φύλλου Excel. Κάθε γραμμή του πίνακα της βιβλιοθήκης αντιπροσωπεύει και μία χρωματογραφική κορυφή, όπως την ορίσαμε στο θεωρητικό κομμάτι της εισαγωγής. Πληροφορίες για τις πρόσθετες έννοιες του μεταβολίτη, παραγώγου, προφίλ ιόντων, καθώς και πληροφορία που αφορά σημειώσεις για την εξέλιξη της ίδιας της βιβλιοθήκης αναμειγνύονται στις εγγραφές των κορυφών. Αυτήν την πληροφορία είναι που θα χρειαστεί να κανονικοποιήσουμε στα πλαίσια της εργασίας



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	mesblID	old_peak_name	peak_name	metabolite_nar	tms_category	retention_time	marker_ion	ion1	ion2	ion3	ion4	ion5	ion6	comments	comme
2	M_0001	b_1	Un_0199	Un_0199		4.4	190	190	117	221	226				BHK ce
3	M_0002	b_2	Un_0203	Un_0203		4.723	110	110							To exei
4	M_0003	m_3(xri)/C_008 (Lena)	m_3/C_008	m_3/C_008	3	4.87	184	194	134	228	117	77	110		before
5	M_0004	P0514(trailing peak)	Un_0226	Un_0226	1	5.148	128	128	103	100	162	206	146		trailing
6	M_0005	Ethanolamine 2 TMS	ethanolamine 2TMS	ethanolamine T	3	5.152	102	206	128	102	218	162	116	To exei Harin+Spyros+Xrist	
7	M_0006	U_004	Un_0182	Un_0182		5.283	155	155	228						
8	M_0007	unknown x_9 /C_016/d4	Un_0001	Un_0001	1	5.478	102	102							Katerin
9	M_0008	MSTFA	MSTFA	MSTFA	1	5.492	184	184	214	134	229	77	103	derivatizing agent	
10	M_0009	a_1	Un_0139	Un_0139		5.896	115	115	130	58	52				To exei
11	M_0010	unknown a_2	Un_0002 (a2)	Un_0002 (a2)	1	6.231	122	122	152	137	265				RT:Spy
12	M_0011	c_2	Un_0221	Un_0221		6.535	355	355	267	103	205				Xronos
13	M_0012	m_6/C_021(Lena)/U_005	Un_0003	Un_0003	1	6.788	283	283	229	191	157	141			Xronos
14	M_0013	C_023 (u_006_Aris)	Un_0178	Un_0178	1	7.12	131	131	184	200	110	134	103		Aris:pr
15	M_0014	Lactate_2TMS	lactate 2TMS	lactate	1	7.139	117	117	191	381	307	219	133	organic acid	Spyros
16	M_0015	P0721 (a_3, u_007)	Un_0175	Un_0175	10	7.201	281	281	369	265	249	250			Harin(P
17	M_0016	a_4(Spy)/ C_024(Lena)	Un_0005	Un_0005	1	7.415	116	116	100	144	88	61	75		Spiros(
18	M_0017	P0743	Un_0004	Un_0004	1	7.428	123+93		123	93	160	55	125	117	unclass
19	M_0018	Glyoxylate Meox 1 TMS	glyoxylate MeOx1	glyoxilate MeO	1	7.441	160	160	59					organic acid	
20	M_0019	x10'	Un_0069	Un_0069	1	7.882	116	116	100	131	159	61	144		Xriston
21	M_0020	Hexanoic Acid TMS	hexanoic acid TMS	hexanoic acid	10	7.931	173	173	117	75				artifact	
22	M_0021	Alanine_2TMS	alanine 2TMS	alanine TMS de	3	7.977	116	116	190	144				amino acid	Katerin
23	M_0022	P0802	Un_0097	Un_0097	1	8.009	131	131	205	233	115	147			Harin:
24	M_0023	B-alanine 1 TMS	b-alanine 1TMS	b-alanine 1TMS		8.11	117	117							(incorr
25	M_0024	Glycolic_Acid_3TMS	glycolate 3TMS	glycolate	1	8.142	205+177		161	177	205			organic acid	Spiros:
26	M_0025	Pyridine Peak (c_3, P0821	Un_0006	Un_0006	10	8.197	152	152	122	166	205				Harin:c
27	M_0026	a_5/K_15	Un_0140	Un_0140		8.323	126	126	174						To exei
28	M_0027	K_14(kat)/C_030(Lena)	K_14/C_030	K_14/C_030		8.454	151	151							To exei
29	M_0028	unknown_no123/P0849	Un_0007 (unknown no	Un_0007 (unkno	10	8.477	174	174	100	86	59	130	147		artifact

Εικόνα 7 – Δείγμα του λογιστικού φύλλου της βιβλιοθήκης κορυφών

Στα περιεχόμενα της βιβλιοθήκης κορυφών συμπεριλαμβάνονται συνολικά περίπου 900 κορυφές μεταβολιτών, από τις οποίες 340 έχουν περάσει ικανή διαδικασία επισκόπησης και γύρω στους 200 αναγνωρισμένους μεταβολίτες. Σαν στήλες μας δίνεται η εξής λίστα με χαρακτηριστικά:

Στήλη	Περιγραφή
MESBL ID	Αναγνωριστικό Κορυφής
Metabolite Derivative	Παράγωγο Μεταβολίτη
Metabolite	Μεταβολίτης
RT	Χρόνος Παραμονής
RT (Unnormalized)	Χρόνος Παραμονής (Μη κανονικοποιημένος)
Category	Κατηγορία Παραγώγισης
QI	Χαρακτηριστικό Ιόν
c. Ion1	Μοριακά Βάρη Ιόντων στο προφίλ θραύσης
c. Ion2	
c. Ion3	
c. Ion4	
c. Ion5	
c. Ion6	
c. Ion7	
c. Ion8	
Comments	Σχόλια
Comment2	
Comment3	
Comments	
BHK	Βιολογικά Συστήματα στα οποία έχει αναλυθεί η κορυφή
Mouse(BALB/cJ)	
Serum	
HeLa	
Arabidopsis thaliana	
STANDARD	Πρότυπη ουσία για επαλήθευση της αναγνώρισης της κορυφής σε μεταβολίτη

Πίνακας 1 – Στήλες Πίνακα Υπάρχουσας Βιβλιοθήκης Κορυφών

3.2 Τεχνολογικά εργαλεία και μέθοδοι

Στην εποχή της πληροφορίας, ο αριθμός των δεδομένων που παράγουμε και χρησιμοποιούμε καθημερινά, γίνεται κάθε μέρα και μεγαλύτερος. Η πληροφορία για να μπορέσει να αποδώσει το καλύτερο της αξίας της πρέπει να είναι άμεσα διαθέσιμη και εύχρηστη στην διαχείριση, επεξεργασία και ανάγνωσή της. Αυτό είναι που προσπαθεί να επιτύχει η τεχνολογία των Βάσεων Δεδομένων.

Πριν φτάσουμε στις σχεσιακές βάσεις, ας αναφέρουμε πως πιο απλή οργάνωση πληροφορίας και δεδομένων είναι αυτή που γίνεται μέσω αρχείων. Αυτή είναι και η οργάνωση η οποία ήταν η πιο συνήθης πριν την έλευση των υπολογιστών, και ακόμα και σήμερα είναι μία από πιο χρησιμοποιούμενες για απλά δεδομένα ή για φυσικά μέσα. Ένα αρχείο οργανώνεται σε εγγραφές, οι οποίες περιέχουν πεδία. Σαν φυσικό παράδειγμα, θα μπορούσαμε να θεωρήσουμε το ευρετήριο μίας βιβλιοθήκης, όπου οι καρτέλες αποτελούν τις εγγραφές, και τα στοιχεία του κάθε βιβλίου τα πεδία μέσα σε αυτές. Στον κόσμο των υπολογιστών, κυριαρχούν οι εφαρμογές λογιστικών φύλλων όπως το Excel, στις οποίες και πάλι η πληροφορία οργανώνεται σε αρχεία.

Λειτουργώντας με απλά αρχεία εγγραφών, παρ' όλ' αυτά, γρήγορα αντιμετωπίζουμε ένα άνω κατώφλι ευχρηστίας, ειδικά όταν αρχίζουμε να χρειαζόμαστε πάνω από ένα αρχείο για τα δεδομένα μας, ή όταν αρχίσουμε να διαμοιράζουμε εκδόσεις του ίδιου αρχείου. Τα προβλήματα τα οποία αντιμετωπίζονται είναι: Ο πλεονασμός (redundancy) των δεδομένων, όπου αυτά απαντώνται πολλές φορές μεταξύ των αρχείων. Την ασυνέπεια (inconsistency) όπου οι διαφορετικές εκδόσεις του ίδιου αρχείου πάνω στις οποίες εργάζονται διαφορετικοί χρήστες αρχίζουν να αποκτούν διορθώσεις που όμως δεν μεταφέρονται στα υπόλοιπα αντίγραφα, προβλήματα μερισμού δεδομένων (data sharing), όπως για παράδειγμα να χρειαζόμαστε να ανατρέξουμε σε μία πρόσφατη έκδοση που δεν είναι άμεσα διαθέσιμη, ή όταν καλούμαστε να ενοποιήσουμε δεδομένα από πολλές εκδόσεις. Τέλος αντιμετωπίζουμε πρόβλημα προτυποποίησης, όπου κάθε εγγραφή ή αρχείο μπορεί να έχει ανομοιομορφίες στο πως καταγράφεται η πληροφορία. Ειδικά η τελευταία πρόκληση δημιουργεί προβλήματα όταν προσπαθούμε να συνενώσουμε πληροφορία από πάνω από μία πηγές.

3.2.1 Σχεσιακές Βάσεις Δεδομένων

Για να επιλυθούν αυτές οι προκλήσεις, και κυρίως επειδή το ποσό της πληροφορίας άρχισε γρήγορα με την έλευση της τεχνολογίας των υπολογιστών να ξεπερνάει τις ικανότητες μέσων, αναπτύχθηκε η τεχνολογία των βάσεων δεδομένων, και αντίστοιχα λύσεων λογισμικού που αποκαλούνται Συστήματα Διαχείρισης Βάσης Δεδομένων (DataBase Management System - DBMS). Αυτά τα συστήματα ανέλαβαν την διαχείριση των αρχείων δεδομένων, δίνοντας στους χρήστες των δεδομένων μία διεπαφή για την προσπέλασή τους. Αυτή η διεπαφή εισήγαγε έναν διαχωρισμό μεταξύ του χρήστη, ή μίας εφαρμογής χρήστη αντίστοιχα, και της συστήματος των αρχείων στα οποία είναι αποθηκευμένα τα δεδομένα, προσδίδοντας την λεγόμενη ιδιότητα της ανεξαρτησίας δεδομένων. Αυτή η ανεξαρτησία έδωσε την δυνατότητα στα DBMS να βελτιστοποιήσουν την προσπέλαση στα δεδομένα. Ακόμα πιο ενδιαφέρον, το γεγονός πως όλοι οι χρήστες πλέον περνούν μέσω μίας διεπαφής προς μία κεντρικοποιημένη βάση, έλυσε από μόνη της πολλά από τα προβλήματα της χρήσης αρχείων.

Αν και υπήρχαν πάνω από ένα είδη βάσεων, αυτό που τελικά επικράτησε και είναι κοινό σήμερα σε όλων των ειδών τις εφαρμογές είναι το Σχεσιακό (Relational) Μοντέλο Βάσεων δεδομένων. Αντίστοιχα αναφερόμαστε σε Σχεσιακές Βάσεις και Σχεσιακά Συστήματα Διαχείρισης Βάσης Δεδομένων (RDBMS) (Codd E. F., 1970). Οι Σχεσιακές Βάσεις Δεδομένων επικράτησαν ακριβώς για την ευχρηστία και την εύκολη κατανόηση τους, αφού τα δεδομένα τους είναι διατεταγμένα σε πίνακες με εγγραφές και πεδία, πολύ κοντά στην χρήση των αρχείων που αναφέραμε.

Για την κατανόηση της εργασίας, οφείλουμε να εξηγήσουμε μερικές έννοιες σχετικές με τις βάσεις δεδομένων. Κύρια έννοια είναι αυτή της Οντότητας. Μία **Οντότητα (Entity)** μπορεί να αντιπροσωπεύει ένα είδος αντικειμένου, προσώπου ή γενικά φυσικής ή λογικής ύπαρξης. Μιλώντας για οντότητα μπορούμε να αναφερόμαστε είτε σε ένα συγκεκριμένο από αυτά τα πρόσωπα ή αντικείμενα, αλλά συνήθως αναφέρουμε ως Οντότητα το είδος, ή έννοια του συνόλου τους. Οι οντότητες μεταξύ του μπορεί να ενώνονται με **Σχέσεις (Relations)** που θα δούμε παρακάτω. Κάθε οντότητα αντιστοιχεί στις Σχεσιακές Βάσεις με έναν **πίνακα (table)** που αποτελείται από **γραμμές (rows)** και **στήλες (columns),** οι οποίες αντιστοιχούν στις **εγγραφές (records)** και **χαρακτηριστικά (attributes)** ή **πεδία (fields)** των Οντοτήτων. Τα χαρακτηριστικά, ανάλογα με το αν είναι αριθμοί, αλφαριθμητικά,

λογικές τιμές ή κάτι άλλο, έχουν επίσης έναν **τύπο δεδομένων (data type)** και αντίστοιχα ένα πεδίο **ορισμού (domain)**. Στο κάθε πεδίο της κάθε εγγραφής έχουμε την **τιμή (value)** της, η οποία μπορεί να απουσιάζει τελείως και σε αυτήν την περίπτωση λέμε πως η τιμή είναι **κενή (null)**. Οι τιμές όλων των πεδίων μίας εγγραφής ονομάζονται **πλειάδα (tuple)**.

Μεταξύ των Οντοτήτων ορίζονται **Συσχετίσεις** ή **Σχέσεις (Relations)**. Αυτές μας επιτρέπουν να ενώσουμε τις οντότητες μεταξύ τους, σχηματίζοντας μία λογική της βάσης η οποία ονομάζεται **Μοντέλο Οντοτήτων – Συσχετίσεων (Entity Relation Diagram)**. Ένα τέτοιο μοντέλο είναι πρωταρχικό εργαλείο στον σχεδιασμό Βάσεων, αφού μας δίνει την δυνατότητα να περιγράψουμε τα γενικά στοιχεία μίας βάσης χωρίς την λεπτομέρεια που θα χρειαζόταν για την υλοποίησή της. Τα είδη των σχέσεων είναι τα **Ένα-προς-Ένα (1:1)**, **Ένα-Προς-Πολλά (1:M)** και **Πολλά-προς-Πολλά (M:M)**.

Μία ακόμα έννοια η οποία θα χρειαστούμε είναι αυτή των κλειδιών. **Κλειδί (key)**, **πρωτεύον κλειδί (primary key)** ή **μοναδικό αναγνωριστικό (id)**, είναι ένα από τα χαρακτηριστικά, ή μερικές φορές συνδυασμός χαρακτηριστικών, του οποίου η τιμή για κάθε εγγραφή είναι μοναδική και όχι κενή σε κάθε Οντότητα και αντίστοιχα πίνακα. Η μοναδικότητά του μας δίνει τη δυνατότητα να αναγνωρίσουμε μονοσήμαντα μία εγγραφή. Ένα **ξένο κλειδί (foreign key)** από την άλλη, είναι ένα απλό χαρακτηριστικό ενός πίνακα, που μπορεί να επαναλαμβάνεται, και ταυτόχρονα είναι πρωτεύον κλειδί σε έναν άλλον πίνακα. Οι σχέσεις υλοποιούνται μέσω των κλειδιών και κάποιον **περιορισμών (constraints)** οι οποίοι ορίζουν τι συμβαίνει αν διαγραφεί μία εγγραφή για την οποία υπάρχει αναφορά σε κάποιον πίνακα.

Τέλος, το σύνολο των πινάκων, των χαρακτηριστικών των στηλών τους μαζί με τους τύπους τους, τα κλειδιά τις σχέσεις και τους περιορισμούς ονομάζονται **Σχήμα (schema)** της βάσης, και ο ακριβής ορισμός του αποτελεί την υλοποίηση της βάσης.

3.2.2 Σχεδιασμός και Κανονικοποίηση Βάσεων Δεδομένων

Χωρίς να αναφερθούμε στους μαθηματικούς ορισμούς των **Κανονικών Μορφών**, οι οποίοι εισήχθησαν για πρώτη φορά στο έργο (Codd E. F., 1970), θα περιγράψουμε πως χρησιμοποιούνται για τον σχεδιασμό μίας βάσης δεδομένων. Εν σύντομα, οι Κανονικές Μορφές είναι 4 και αναφέρονται ως **Κανονική Μορφή 1 (KM1 ή Normal Form - NF1)**, **KM2**, **KM3** και **Μορφή Boyce-Codd** (Codd E. F., 1974). Έχουν προστεθεί αργότερα και περισσότερες από αυτές τις κανονικές μορφές, αλλά πρακτικά χρησιμοποιούνται στον σχεδιασμό οι πρώτες τρεις. Ο σκοπός των κανονικών μορφών και της διαδικασίας της κανονικοποίησης είναι να διαχωριστούν πολύπλοκες οντότητες σε μικρότερες, με στόχο πάντα να αποφεύγουμε **επαναλήψεις (redundancy)** των δεδομένων.

Στο πρώτο στάδιο του σχεδιασμού μίας σχεσιακής βάσης δεδομένων, καλούμαστε να συντάξουμε το Διάγραμμα Οντοτήτων – Συσχετίσεων. Η επιλογή των Οντοτήτων είναι, κατά πολλούς, τέχνη και αφορά την μετάφραση ενός πεδίου (domain) πληροφορίας σε αυτό της πληροφορικής. Μαζί με τα είδη των οντοτήτων που θα αναγνωριστούν, καταγράφονται και τα κύρια χαρακτηριστικά τους, καθώς και γίνεται προετοιμασία για το ποια χαρακτηριστικά θα παίξουν τον ρόλο των κλειδιών. Ταυτόχρονα, οργανώνονται οι Σχέσεις μεταξύ των οντοτήτων και ο βαθμός τους. Ειδικά για τις σχέσεις Ένα-Προς-Ένα, προσπαθούμε να πάρουμε απόφαση για το αν είναι πιο χρήσιμο να έχουμε δύο οντότητες για δύο λογικές έννοιες ή μία. Η απλότητα της κάθε οντότητας αντισταθμίζεται πάντα με την απλότητα του διαγράμματος.

Σε δεύτερη φάση επιστρέφουμε στα δεδομένα μας, αν αυτά υπάρχουν, ή χρησιμοποιούμε σενάρια εγγραφών, αν όχι. Ο σκοπός μας είναι να αναγνωρίσουμε μη κανονικές μορφές στα δεδομένα μας, και να παραλλάξουμε το Διάγραμμα Οντοτήτων μας με σκοπό να τα κανονικοποιήσουμε. Θυμίζουμε πως ο σκοπός της κανονικοποίησης είναι η αποφυγή της επανάληψης. Κανονικοποιώντας σε πρώτη Κανονική Μορφή, αυτό που μας ενδιαφέρει είναι πεδίο των χαρακτηριστικών μας να περιέχει μία τιμή ανά εγγραφή. Αυτό συνήθως μπορεί να συμβεί αν τα δεδομένα μας ήταν σε μορφή αρχείου, οπού είναι πολλές φορές βολικό να γράφονται δύο τιμές αν αυτές εμφανιστούν, όπως για παράδειγμα, πολλαπλά τηλέφωνα για έναν πελάτη. Αντίστοιχα, μπορεί να δούμε πολλές στήλες οι οποίες να αναφέρονται το ίδιο χαρακτηριστικό, πχ. *τηλέφωνο1*, *τηλέφωνο 2*. Στις περιπτώσεις αυτές, θα πρέπει να

δημιουργήσουμε μία καινούργια οντότητα για αυτήν την πληροφορία, η οποία να βρίσκεται σε σχέση Ένα-προς-Πολλά με αυτήν που την περιείχε.

Πριν την κανονικοποίηση σε Δεύτερη Κανονική Μορφή, χρειάζεται να αναγνωρίσουμε ποιο χαρακτηριστικό, ή ποια χαρακτηριστικά είναι αυτά τα οποία μπορούν να παίξουν τον ρόλο του πρωτεύοντος κλειδιού, απλού ή σύνθετου. Για κάθε ένα από αυτά τα εν δυνάμει κλειδιά, αναγνωρίζουμε ποια από τα χαρακτηριστικά του πίνακα εξαρτώνται από αυτά. Σαν παράδειγμα, από το ΑΜ ενός φοιτητή εξαρτώνται το όνομα και το επώνυμο του. Για κάθε μία τέτοια εξάρτηση δημιουργούμε μία καινούργια οντότητα και αφαιρούμε τις στήλες από τον αρχικό πίνακα, δημιουργώντας σχέσεις μέσω ξένου κλειδιού. Μπορούμε να πούμε πως βρισκόμαστε σε ΚΜ2, όταν τα χαρακτηριστικά όλων των πινάκων εξαρτώνται από το πρωτεύον τους κλειδί.

Για να κανονικοποιήσουμε σε Τρίτη Κανονική Μορφή, ερευνούμε αν υπάρχουν εξαρτήσεις μεταξύ χαρακτηριστικών ενός πίνακα. Κάτι τέτοιο θα είναι επίσης εμφανές από την ύπαρξη ακόμα επαναλαμβανόμενης πληροφορίας. Αν βρεθούν τέτοια χαρακτηριστικά, τότε τα μεταφέρουμε επίσης σε μία καινούργια οντότητα. Λέμε πως βρισκόμαστε στην ΚΜ3, όταν δεν υπάρχουν αλληλεξαρτήσεις στα χαρακτηριστικά ενός πίνακα.

Έχοντας περάσει τις φάσεις κανονικοποίησης, θα πρέπει να έχουμε καταλήξει σε ένα πλήρες Διάγραμμα Οντοτήτων – Συσχετίσεων. Εδώ μπορεί να παρέμβει ο σχεδιαστής και να επιλέξει να είτε να αποκανονικοποιήσει μερικώς τους πίνακες αν αυτό θα σήμαινε μεγαλύτερη ευχρηστία τους, είτε να διασπάσει Οντότητες μεγάλου χαρακτηριστικών στα δύο, για πιο εύκολη διαχείριση. Μετά και από τέτοιες παρεμβάσεις, απομένει να σχεδιαστεί το πλήρες σχήμα της βάσης, δίνοντας ονόματα στους πίνακες και τα πεδία τους από τις περιγραφές των οντοτήτων, ορίζοντας πρωτεύοντα κλειδιά, θέτοντας τους τύπους όλων των πεδίων και υλοποιώντας τις σχέσεις μέσω περιορισμών.

3.2.3 Αντικειμενοστραφής - Σχεσιακή Αντιστοίχιση (ORM)

Σε αυτό το σημείο που έχουμε καταλήξει σε ένα πλήρη σχήμα της βάσης, η κλασσική αντιμετώπιση θα ήταν να το υλοποιήσουμε απευθείας σε μία γλώσσα προγραμματισμού βάσεων δεδομένων όπως η SQL. Εδώ εμείς θα διαλέξουμε να

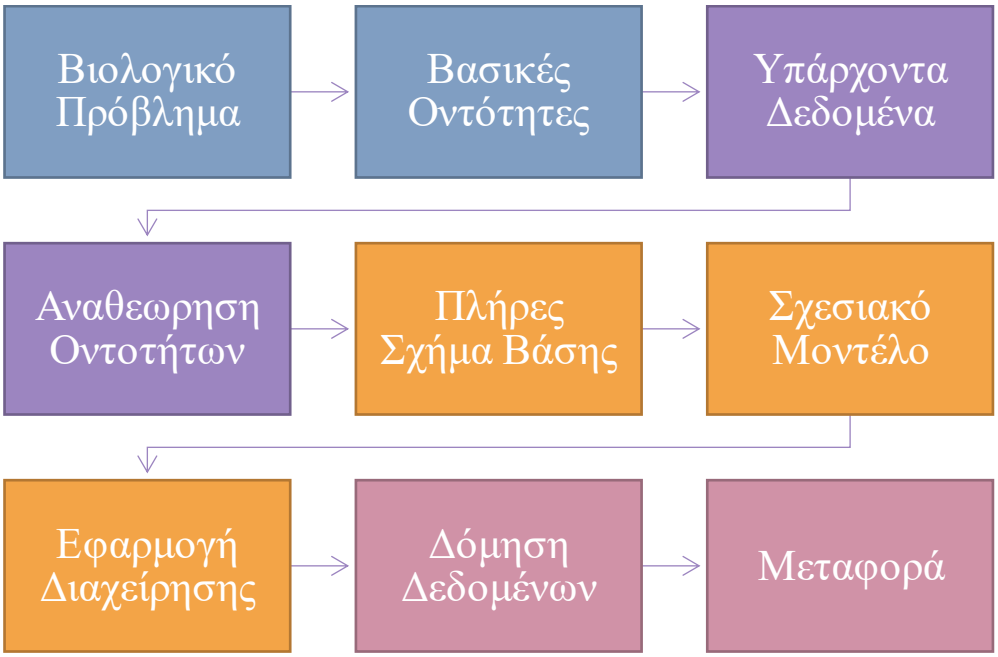
εισάγουμε ένα επίπεδο αφαίρεσης ακόμα. Έχοντας υπόψιν μας πως πάνω στην βάση μας θέλουμε να χτίσουμε μία διαδικτυακή εφαρμογή ή και υπηρεσία, μπορούμε να επιλέξουμε η υλοποίηση αυτής να γίνει μέσω μίας αντικειμενοστραφούς γλώσσας προγραμματισμού. Σε αυτήν την περίπτωση, για κάθε μία Οντότητα επιθυμούμε να υπάρχει ένας πίνακας στην Σχεσιακή Βάση μας, όπως και μία κλάση στην αντικειμενοστραφή προγραμματιστική μας υλοποίηση.

Η προγραμματιστική τεχνική η οποία αντιστοιχίζει μία προγραμματιστική κλάση με έναν πίνακα σχεσιακής βάσης ονομάζεται Αντικειμενοστραφής - Σχεσιακή Αντιστοίχιση (Object Relational Mapping - ORM). Αυτή η αντιστοίχιση, ιστορικά, έγινε προσπάθεια να γίνει από την πλευρά των βάσεων, ορίζοντας αυτό που ονομάζουμε Αντικειμενοστραφείς Σχεσιακές Βάσεις Δεδομένων (OORDBMS) (Barsalou & Wiederhold, 1990). Στην πράξη όμως, αν και υπάρχουν τέτοιες βάσεις, σήμερα έχει επικρατήσει αυτή η αντιστοίχιση να γίνεται από την πλευρά της προγραμματιστικής γλώσσας. Βιβλιοθήκες για όλες τις διαδεδομένες γλώσσες προγραμματισμού αναλαμβάνουν να δίνουν τα εργαλεία ώστε προγραμματιστικές κλάσεις να μπορούν να αρχικοποιηθούν από τα δεδομένα μίας βάσης και αντίστροφα, να αποθηκεύουν τα δεδομένα τους σε αυτές, μιλώντας στις σχεσιακές βάσεις μέσω της κοινής γλώσσας διεπαφής SQL.

Οι βιβλιοθήκες αντιστοίχισης μπορούν να είναι είτε ανεξάρτητες, είτε μέρη ευρύτερων βιβλιοθηκών που στοχεύουν στην ανάπτυξη διαδικτυακών εφαρμογών. Στην δεύτερη περίπτωση, μιλάμε για βιβλιοθήκες ανάπτυξης οι οποίες συνήθως ακολουθούν την αρχιτεκτονική Μοντέλο-Όψη-Ελεγκτής (Model-View-Controller, MVC). Από τα τρία συστατικά, το Μοντέλο είναι αυτό το οποίο καλείται να αφαιρέσει την πολυπλοκότητα της αλληλεπίδρασης με την βάση, και να την εμφανίσει στον προγραμματιστή ως μία κλάση. Σε αυτές τις κλάσεις αναφερόμαστε επίσης σαν μοντέλα, και είθισται τα χαρακτηριστικά τους να αντιστοιχούν στα πεδία της βάσης.

Ένα πλεονέκτημα το οποίο σκοπεύουμε να χρησιμοποιήσουμε, είναι η προγραμματιστική υλοποίηση απευθείας μοντέλων για τις οντότητες μας. Αυτό θα μας επιτρέψει να αφήσουμε την βιβλιοθήκη αντιστοίχισης να έχει πλήρη διαχείριση της βιβλιοθήκης, συμπεριλαμβανομένης και της δημιουργίας των πινάκων.

Κεφάλαιο 4. Υλοποίηση



Εικόνα 8 – Βήματα Υλοποίησης

Κατά την υλοποίηση της εργασίας εργαστήκαμε σε τρεις φάσεις. Στην πρώτη φάση και ίσως την πιο σημαντική, δημιουργήσαμε μία γέφυρα μεταξύ του Βιολογικού και Πειραματικού κόσμου, και αυτού της πληροφορικής. Με δεδομένα μας την κατανόηση του βιολογικού προβλήματος, σχεδιάσαμε μία πρώτη ιεραρχία βασικών οντοτήτων και τις σχέσεις μεταξύ τους. Στην δεύτερη φάση, αναγνωρίζοντας τις ιδιαιτερότητες των δεδομένων μας, προσθέσαμε και αναθεωρήσαμε κάποιες οντότητες. Στην τρίτη φάση, δημιουργήσαμε το πλήρες σχήμα της βάσης, επιλέγοντας τα πλήρη χαρακτηριστικά της κάθε οντότητας, το οποίο μας επέτρεψε να υλοποιήσουμε το Σχεσιακό Μοντέλο και παράξουμε μία Εφαρμογή Διαχείρισης των εγγραφών της Βιβλιοθήκης. Τέλος, στην τέταρτη φάση, καθαρίσαμε, δομήσαμε και μεταφέραμε τα δεδομένα μας στην βάση.

4.1 Μετάφραση του Βιολογικού Προβλήματος / Αρχική Επιλογή Οντοτήτων

Σύμφωνα με όσα αναφέραμε στην θεωρητική εισαγωγή μπορούμε να αναγνωρίσουμε τις εξής οντότητες, οι οποίες και θέλουμε να απαρτίζουν την Βιβλιοθήκη.

4.1.1 Μεταβολίτης

Πρόκειται για την οντότητα η οποία είναι κύρια στο βιολογικό πρόβλημα το οποίο καλούμαστε να περιγράψουμε, και αντιπροσωπεύει ένα φυσικό μεταβολικό μόριο που λαμβάνει μέρος σε ένα Μεταβολικό δίκτυο. Ένας χρήστης μας θα χρησιμοποιήσει την Βιβλιοθήκη τελικώς ώστε να ανάγει τα πειραματικά του δεδομένα σε ένα τέτοιο δίκτυο, στο οποίο οι κόμβοι (nodes) θα είναι οι Μεταβολίτες.

Γνωρίζουμε πως θέλουμε να δώσουμε την δυνατότητα σε έναν τέτοιο χρήστη να εξερευνήσει το σύνολο των μεταβολιτών που συμπεριλαμβάνονται στη βιβλιοθήκη, καθώς και να κάνει αναζήτηση για να τους βρει. Μία δυσκολία που προέρχεται από την φύση των μεταβολιτών ως βιολογικών και χημικών μορίων είναι πως μπορεί να έχουν έναν αριθμό από ονόματα που αναφέρονται σε αυτούς. Κάποια ονόματα από την άλλη μπορεί να αναφέρονται κάποιες φορές σε διαφορετικά μόρια. Χρειάζεται, επομένως, να συμπεριλάβουμε χαρακτηριστικά όπως συνώνυμα ή χημικά χαρακτηριστικά όπως χημική κατηγορία ή μοριακό βάρος για να βοηθήσουμε την αναζήτηση.

Ο Μεταβολίτης, επίσης, θέλουμε να παίξει τον ρόλο της Οντότητας μέσω της οποίας η Βιβλιοθήκη μας θα ενωθεί με τις εσωτερικές και εσωτερικές βάσεις βιολογικού ενδιαφέροντος που έχουμε αναφέρει. Για αυτό το σκοπό, χρειάζεται να του δώσουμε ένα αναγνωριστικό με το οποίο να γίνει αυτή η σύνδεση. Ένα τέτοιο κατάλληλο αναγνωριστικό το οποίο και επιλέξαμε, είναι να χρησιμοποιήσουμε την ονοματοδοσία από την Βάση Χημικών Μορίων Βιολογικού Ενδιαφέροντος ChEBI. Από την στιγμή που ο Μεταβολίτης αντιπροσωπεύει ένα μόριο, και μέσω της ChEBI, μπορούμε να εκμεταλλευτούμε τα χημικά χαρακτηριστικά και τα συνώνυμα που αυτή η βιβλιοθήκη διαθέτει.

4.1.2 Παράγωγο

Για να μπορέσουμε να αναλύσουμε ένα μεταβολικό μόριο μέσω της Χρωματογραφίας Αερίων, χρειάζεται να το ανάγουμε σε μία αέρια μορφή μέσω μίας Μεθόδου Παραγώγισης. Κάθε μεταβολικό μόριο, ανάλογα με την χημική κατηγορία του παράγει ένα ή περισσότερα αέρια παράγωγα μόρια. Αναγνωρίζουμε, λοιπόν, πως χρειαζόμαστε μία δεύτερη οντότητα, την οποία ονομάζουμε Παράγωγο, και η οποία βρίσκεται σε μία σχέση Ένα-προς-Πολλά με την οντότητα του Μεταβολίτη. Τα συγκεκριμένα παράγωγα μόρια, επίσης, εξαρτώνται από την μέθοδο που χρησιμοποιείται. Σε περίπτωση διαφορετικής μεθόδου, ο ίδιος Μεταβολίτης θα μετατραπεί σε διαφορετικά παράγωγα, το οποίο μας υποδεικνύει πως θα πρέπει να έχουμε την μέθοδο σαν ένα χαρακτηριστικό.

4.1.3 Κορυφή

Όπως ο Μεταβολίτης είναι η κύρια βιολογική μας οντότητα, έτσι και η Κορυφή είναι η κύρια οντότητα στα πειραματικά μας δεδομένα. Η μέθοδος της Χρωματογραφίας Αερίων - Φασματομετρίας Μάζας μας δίνει μία τρισδιάστατη αναπαράσταση. Για κάθε μία κορυφή του Χρωματογραφήματος, εμείς εξετάζουμε το Φασματογράφημα Μάζας. Οι κορυφές είναι αυτές οι οποίες αντιστοιχούν σε ένα Παράγωγο και που τελικώς θα μας οδηγήσουν στην αναγνώριση ενός Μεταβολίτη.

Εδώ είναι που κληθήκαμε να πάρουμε την πιο σημαντική σχεδιαστική απόφαση, η οποία αφορά το εξής. Ζητούμαστε να δημιουργήσουμε μία βιβλιοθήκη βάσει της οποίας ένας χρήστης ερευνητής θα μπορέσει να ταυτοποιήσει τα πειραματικά του δεδομένα. Στην πιο απλή εκδοχή, θα μπορούσαμε να δώσουμε στον χρήστη έναν αριθμό από πειραματικά δεδομένα που πιστεύουμε πως αντιστοιχούν σε ένα παράγωγο και άρα ένα μεταβολικό μόριο, και να τον αφήσουμε να συγκρίνει αυτά τα δεδομένα με τα δικά του. Κάτι τέτοιο είναι σύνηθες σε αντίστοιχες βιβλιοθήκες. Μπορούμε, επίσης, να σχολιάσουμε πως μία τέτοια σχεδίαση είναι πολύ κοντά σε ένα Αποθετήριο Πειραμάτων όπως το έχουμε ήδη ορίσει.

Αντί για αυτό, παίρνουμε μία διαφορετική και καινοτόμα κατεύθυνση, ορίζοντας πως θα δημιουργήσουμε μία Προτυποποιημένη Βιβλιοθήκη Κορυφών Μεταβολιτών. Πρακτικά, αντί για έναν αριθμό από πειραματικές κορυφές, θέτουμε ως ζητούμενο να επιστρέφουμε στον χρήστη μία πρότυπη κορυφή, με ένα σύνολο από χαρακτηριστικά σύμφωνα με τα οποία είτε ένας ερευνητής είτε ένα αυτοματοποιημένο

πρόγραμμα, όπως το M-Iolite να μπορεί να χρησιμοποιήσει για την ταυτοποίηση των πειραματικών του κορυφών.

Εισάγουμε επομένως την οντότητα Κορυφή η οποία αντιστοιχεί σε ένα πρότυπο κορυφής όπως μόλις ορίσαμε. Δεδομένου πως η Βιβλιοθήκη δημιουργείται για μία συγκεκριμένη αναλυτική τεχνική, περιμένουμε πως θα έχουμε ένα πρότυπο για κάθε μεταβολικό παράγωγο. Παρ' όλ' αυτά, αυτό θέλουμε να είναι κάτι στο οποίο η βιβλιοθήκη να μπορεί να εξελιχθεί και να περιέχει περισσότερες από μία τεχνικές. Ορίζουμε λοιπόν την σχέση μεταξύ Μεταβολίτη και Κορυφής ως Ένα-προς-Πολλά, με την τεχνική ως χαρακτηριστικό.

Για να βρούμε χαρακτηριστικά της οντότητας, επιστρέφουμε στο πείραμα και εξετάζουμε τα εξής. Το πρώτο είναι η χρονική στιγμή στην Χρωματογραφική ανάλυση στην οποία αντιστοιχεί η Κορυφή, και η οποία ονομάζεται χρόνος παραμονής. Στα πλαίσια της προτυποποίησης που συζητήσαμε, θεωρούμε αρχικά πως μπορούμε να έχουμε μία μοναδική τιμή για αυτό το χαρακτηριστικό. Στην πραγματικότητα όμως, η τιμή του χρόνου παραμονής είναι ευαίσθητη στις πειραματικές παραμέτρους και θα την επανεξετάσουμε αργότερα. Το δεύτερο χαρακτηριστικό είναι το προφίλ ιόντων που παρουσιάζεται στο Φασματογράφημα Μάζας το οποίο αντιστοιχεί στην Χρωματογραφική Κορυφή, και το οποίο θα μας δώσει την αμέσως επόμενη οντότητα.

Τέλος, ακριβώς επειδή η Κορυφή είναι η κύρια Οντότητα που προσφέρει η Βιβλιοθήκη Κορυφών, θα χρειαστεί να ονοματοδοτήσουμε τις εγγραφές της με μοναδικό αναγνωριστικό, καθώς και να προσθέσουμε πρόσθετα χαρακτηριστικά τα οποία να μας επιτρέπουν να εκφράσουμε την βεβαιότητα μας για την αξιοπιστία της πληροφορίας που προσφέρουμε και να μπορούμε να παρακολουθούμε την εξέλιξή της στο χρόνο. Αυτά θα αναλυθούν στην δεύτερη φάση αναθεώρησης των οντοτήτων.

4.1.4 Ιόν

Κατά την χρονική στιγμή όπου μία κορυφή εμφανίζεται στην Χρωματογραφική ανάλυση, η τομή της τρισδιάστατης πληροφορίας στους άλλους δύο άξονες μας δίνει ένα Φασματογράφημα Μάζας. Αυτό αναπαριστά, για το παράγωγο μόριο που αναλύεται εκείνη τη στιγμή, ένα προφίλ θραύσης σε ιόντα. Κάθε ιόν χαρακτηρίζεται από το μοριακό του βάρος, καθώς και από την ένταση του στο γράφημα η οποία υποδηλώνει το ποσοστό του.

Το προφίλ θραύσης για ένα μόριο ορίζεται από την ίδια του την χημική σύσταση και περιμένουμε να είναι σταθερό ακόμα και σε περιπτώσεις διαφορετικών πειραματικών παραμέτρων. Βέβαια, την στιγμή της κορυφής δεν είναι απαραίτητο πως αναλύεται μόνο ένα μόριο. Άλλα μόρια, κοντινών κορυφών μπορεί να συνεισφέρουν μικρά ποσοστά ιόντων. Σαν μέρος της προτυποποίησης, οι ερευνητές του εργαστήριου έχουν κληθεί να επιλέξουν συγκεκριμένα ιόντα, μέχρι οχτώ τον αριθμό, για να αποτελέσουν το προφίλ. Ταυτόχρονα, λόγω της επιλογής, δεν έχει κριθεί απαραίτητο να αποθηκευτεί το ποσοστό του κάθε ιόντος, το οποίο σε αντίθεση με το μοριακό βάρος μπορεί να ποικίλει. Μπορούμε να εισάγουμε, επομένως, μία οντότητα Ιόν, με σχέση με την Κορυφή Ένα-προς-Πολλά, και χαρακτηριστικό του το μοριακό του βάρος.

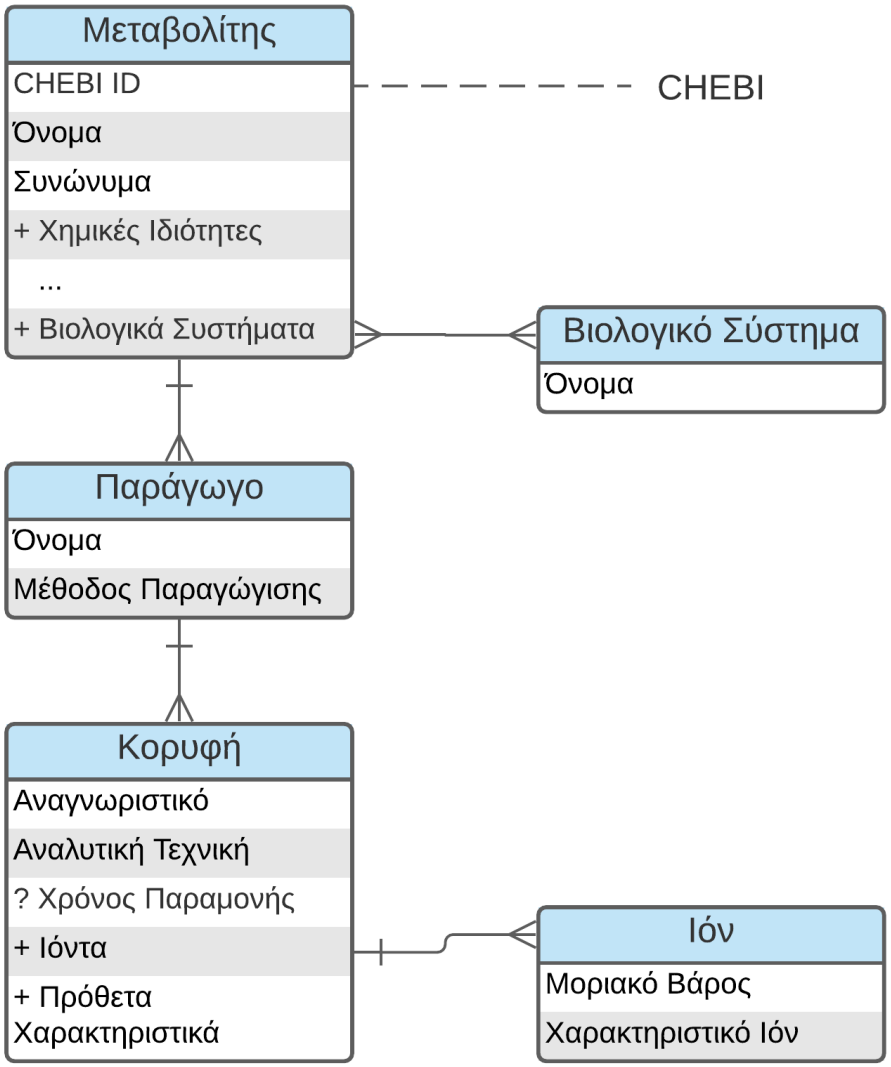
Τέλος, ένα χαρακτηριστικό το οποίο είναι χρήσιμο στους ερευνητές χρήστες της Βιβλιοθήκης είναι η έννοια του Χαρακτηριστικού Ιόντος. Πρόκειται για ένα από τα Ιόντα της Κορυφής, το οποίο χρησιμοποιείται έπειτα για την ποσοτικοποίηση του μεταβολίτη που αντιστοιχεί στην εκάστοτε κορυφή στο τελικό μεταβολικό δίκτυο. Πολλές φορές είναι το ιόν που εμφανίζεται με το μεγαλύτερο ποσοστό, ενώ άλλες φορές ένα ιόν που δεν είναι κοινό σε άλλες Κορυφές. Στην παρούσα φάση, αυτό το προσθέτουμε σαν χαρακτηριστικό σε ένα από τα Ιόντα της κάθε Κορυφής.

4.1.5 Βιολογικό Σύστημα

Μια τελευταία οντότητα που πρέπει να αναγνωρίσουμε σε αυτήν την φάση είναι αυτή του Βιολογικού Συστήματος, η οποία αναφέρεται σε διαφορετικά βιολογικά δείγματα που ανήκουν σε οργανισμούς διαφορετικών ειδών και τύπους ιστών τους. Κάθε βιολογικό σύστημα έχει το δικό του μεταβολικό δίκτυο, με αποτέλεσμα να μην είναι παρόντες το ίδιο σύνολο από μεταβολίτες στο καθένα.

Στα πλαίσια της προτυποποίησης των κορυφών της Βιβλιοθήκης, γίνεται προσπάθεια οι κορυφές να αναγνωριστούν σε πάνω από ένα βιολογικά συστήματα μέσω και των αντίστοιχων ερευνών σε δείγματά τους. Η αναγνώριση τους σε πολλαπλά δείγματα αυξάνει την εμπιστοσύνη μας στις επιλογές που έχουν γίνει για την προτυποποίηση της κάθε κορυφής.

4.1.6 Αρχικό Διάγραμμα Οντοτήτων



Εικόνα 9 – Αρχικό Διάγραμμα Οντοτήτων

4.2 Αναθεώρηση Σχεδιασμού βάσει των Υπαρχόντων Δεδομένων

Από την αναφορά στην υπάρχουσα μορφή των δεδομένων της Βιβλιοθήκης σε μορφή excel επισημάναμε τις εξής προκλήσεις. Το ποσοστό των Κορυφών οι οποίες έχουν αναγνωριστεί σε Μεταβολίτη και η αξιοπιστία την αναγνώρισης, περιπτώσεις χαρακτηριστικών που έχουν πάνω από μία τιμές με κύριο το Χαρακτηριστικό Ιόν, και την ύπαρξη μη δομημένης πληροφορίας στα σχόλια. Από την πρώτη φάση επίσης έχουμε σημειώσει πως θα θέλαμε να έχουμε καλύτερη διαχείριση χαρακτηριστικών τα οποία διαφέρουν ανά μελέτη, συγκεκριμένα το Χαρακτηριστικό Ιόν, τον χρόνο παραμονής και το Βιολογικό Σύστημα. Τέλος, αναφέραμε πως θα χρειαστούμε χαρακτηριστικά για την διαχείριση και εξέλιξη των εγγραφών στον χρόνο. Παρακάτω εξηγούμε πως αναθεωρώντας τις Οντότητες, επιλύουμε αυτές τις προκλήσεις.

4.2.1 Αξιοπιστία

Ο πρωταρχικός σκοπός της Βιβλιοθήκης Κορυφών που σχεδιάζουμε είναι να χρησιμοποιηθεί από έναν ερευνητή είτε χειροκίνητα είτε αυτοματοποιημένα για την αναγνώριση δικών του πειραματικών κορυφών σε μεταβολικά μόρια. Για να μπορέσει να συνταχθεί μία τέτοια βιβλιοθήκη, ο κάθε Μεταβολίτης θα πρέπει να έχει συναντηθεί πειραματικά στην ανάλυση κάποιου δείγματος. Οι πρώτες βιβλιοθήκες προσέφεραν πρότυπα δεδομένα κάνοντας αναλύσεις σε δείγματα αμιγώς αποτελούμενα από ένα μόνο μεταβολικό μόριο. Χτίζοντας πάνω σε αυτά, η προυπάρχουσα βιβλιοθήκη του εργαστηρίου, την οποία επεξεργαζόμαστε, έχει συνταχθεί με πειραματισμό σε πραγματικά βιολογικά δείγματα. Η πολυπλοκότητα ενός βιολογικού συστήματος και ο αριθμός των μεταβολομικών μορίων που μετέχουν σε αυτό, παραλλάσσουν με διάφορους τρόπους την πειραματική τους έκφραση.

Αυτό όμως είναι και που προσδίδει μεγαλύτερη αξία σε μία τέτοια βιβλιοθήκη. Έχοντας αναγνωρίσει τον ίδιο Μεταβολίτη, βάσει ενός προτύπου Κορυφής της Βιβλιοθήκης, σε περισσότερα από ένα ετερογενή Βιολογικά Συστήματα, τόσο μπορούμε να πούμε με αυτοπεποίθηση πως το πρότυπο της Κορυφής το οποίο προσφέρουμε είναι χρήσιμο για την περαιτέρω αναγνώριση πειραματικών κορυφών σε αυτόν τον Μεταβολίτη. Για τον λόγο αυτό, μας ενδιαφέρει να εισάγουμε στην οντότητα της Κορυφής το χαρακτηριστικό της αξιοπιστίας.

Μέρος του πως χτίζεται η αξιοπιστία είναι ο αριθμός των Βιολογικών Συστημάτων στα οποία έχει εμφανιστεί μία Κορυφή και που έχουμε ήδη συμπεριλάβει

στις οντότητες μας. Το σημαντικότερο κομμάτι της αξιοπιστίας όμως έγκειται στην αναγνώριση της Κορυφής σε κάποιον συγκεκριμένο Μεταβολίτη. Μία συγκεκριμένη κορυφή μπορεί εμφανίζεται συστηματικά στις έρευνες του εργαστηρίου χωρίς όμως να έχει ταυτοποιηθεί σε κάποιον Μεταβολίτη. Στο κομμάτι που αναφέραμε τα περιεχόμενα της προυπάρχουσας βιβλιοθήκης, πάνω από τις μισές εγγραφές δεν έχουν αναγνωριστεί. Ένας μέρος από αυτές ανήκουν σε εντελώς άγνωστα μόρια, σε κάποιες από αυτές από τα ιοντικά θραύσματα μπορούμε να προβλέψουμε πως ανήκουν σε συγκεκριμένη χημική ομάδα (πχ. οξύ). Στις περισσότερες αναγνωρισμένες Κορυφές από την άλλη, έχει γίνει η προσπάθεια να επιβεβαιωθεί περαιτέρω ο Μεταβολίτης, με το να αναλυθεί ένα πρότυπο δείγμα που να περιέχει μόνο αυτό το μόριο, όπως αναφέραμε για τις πρότυπες βιβλιοθήκες, αλλά σε αυτήν την περίπτωση μέσω της πειραματικής διάταξης του εργαστηρίου που έχει κάνει και την έρευνα.

Καταλήγουμε τελικώς πως μπορούμε να εκφράσουμε την αξιοπιστία των Κορυφών, με εύχρηστο κιόλας τρόπο, με ένα βαθμολογικό σύστημα τεσσάρων επιπέδων για την κάθε Κορυφή: Πλήρως Αναγνωρισμένη, Θεωρούμενα Αναγνωρισμένη (Putative), Μη αναγνωρισμένη με Πρόβλεψη χημικών Ιδιοτήτων, και Μη αναγνωρισμένη. Σε μία αναζήτηση, προσβλέπουμε πως ο χρήστης πως θα μπορεί να φιλτράρει τα αποτελέσματα με βάση ένα βαθμό αξιοπιστίας και επάνω.

4.2.2 Άγνωστοι Μεταβολίτες

Για τους άγνωστους Μεταβολίτες χρειάζεται να εξετάσουμε περαιτέρω τις σχέσεις που ορίσαμε στην πρώτη φάση. Εκεί κάναμε την παραδοχή πως ένας Μεταβολίτης μπορεί να παραγωγηθεί σε ένα ή περισσότερα Παράγωγα, κάθε ένα από αυτά, δεδομένης συγκεκριμένης αναλυτικής τεχνικής, αντιστοιχεί σε μία ακριβώς Κορυφή. Για να είμαστε ικανοί να ικανοποιήσουμε το ποσοστό των Κορυφών που δεν είναι αναγνωρισμένες, έχουμε δύο δυνατές στρατηγικές.

Στην πρώτη από αυτές, θα μπορούσαμε για κάθε μία μη αναγνωρισμένη Κορυφή να εισάγουμε στην Βιβλιοθήκη εγγραφές για ένα άγνωστο Παράγωγο και έναν Άγνωστο Μεταβολίτη, πχ *ΑγνωστοΠαράγωγο01*, *ΆγνωστοςΜεταβολίτης01*. Στην πράξη, αυτή ακριβώς είναι η στρατηγική την οποία χρησιμοποιεί η προυπάρχουσα βιβλιοθήκη, και λειτουργεί καλά στα πλαίσια ενός λογιστικού φύλλου ενός πίνακα, αφού ικανοποιεί την ανάγκη για μία συνεπή ονοματοδοσία. Στα πλαίσια όμως της σχεδίασης της Βιβλιοθήκης με πολλαπλές Οντότητες, δημιουργεί ένα σοβαρό

πρόβλημα. Σε αντίθεση με τις εγγραφές γνωστών Μεταβολιτών και Παραγώγων, όλα τα χαρακτηριστικά των Αγνώστων δεν θα περιέχουν ουσιαστική πληροφορία και θα πρέπει να είναι κενά (null), κάτι το οποίο βλάπτει την συνοχή μίας σχεσιακής βάσης από την στιγμή που θα ισχύει για μεγάλο ποσοστό των εγγραφών αυτών των Οντοτήτων.

Στην εναλλακτική στρατηγική, την οποία και επιλέγουμε, χαλαρώνουμε την σχέση μεταξύ των οντοτήτων Παράγωγο και Κορυφή σε Μηδέν-ή-Ένα-προς-Πολλά, ώστε να επιτρέψουμε σε μία Κορυφή να μην αντιστοιχεί σε κάποιο Παράγωγο και άρα Μεταβολίτη. Κάνουμε επομένως την παραδοχή, πως, αν για κάποια Κορυφή δεν είναι συνδεδεμένη με κάποιο Παράγωγο, τότε αυτή η Κορυφή αντιστοιχεί σε μη αναγνωρισμένο μεταβολίτη. Με αυτόν τον τρόπο έχουμε καλύψει τις Κορυφές με κάθε βαθμό αξιοπιστίας, εκτός από αυτές που είναι άγνωστες αλλά για τις οποίες έχουμε προβλέψει κάποια χημικά χαρακτηριστικά. Υποσύνολο αυτών είναι και μία κατηγορία όπου μπορεί να υποπτευόμαστε πως δύο κορυφές ανήκουν σε παράγωγα του ίδιου άγνωστου Μεταβολίτη. Μόνο σε αυτές τις περιπτώσεις λοιπόν, όπου υπάρχει πληροφορία η οποία χρειάζεται να καταγραφεί, θα χρησιμοποιήσουμε την πρώτη στρατηγική και θα εισάγουμε εγγραφές αγνώστων Μεταβολιτών και Παραγώγων για να αποθηκεύσουμε αυτήν την πληροφορία.

4.2.3 Χρόνος Παραμονής

Σημειώσαμε στην πρώτη φάση πως, λόγω της φύσης της ανάλυσης μέσω Χρωματογραφίας Αερίων Φασματομετρία Μάζας, σε αντίθεση με το προφίλ θραύσης των Ιόντων το οποίο είναι σταθερό επειδή υποδεικνύεται από τα χημικά χαρακτηριστικά των μορίων, ο χρόνος παραμονής στον οποίο εμφανίζεται μία κορυφή εξαρτώνται από τις παραμέτρους του πειράματος.

Συγκεκριμένα, ο πειραματιστής ερευνητής που αναλύει το δείγμα καλείται να πάρει μία απόφαση για την ταχύτητα με την οποία θα αναλυθεί μία χρωματογραφική στήλη και η οποία επηρεάζει άμεσα την διακριτική ικανότητα μεταξύ διαδοχικών μορίων. Ανάλογα με την πολυπλοκότητα του δείγματος σε μόρια, υπάρχει ο κίνδυνος να αναλυθούν μαζί πολλαπλά μόρια, δυσχεραίνοντας μετά την αναγνώριση τους. Στα πλαίσια μίας μελέτης, λοιπόν, η ταχύτητα της ανάλυσης είναι μία παράμετρος που πρέπει να αποφασιστεί, και κρατείται σταθερή για το σύνολό της. Σαν μέτρηση της παραμέτρου αυτής, σημειώνεται σαν μέρος των δεδομένων ο χρόνος παραμονής

συγκεκριμένων μορίων που ονομάζονται Εσωτερικά Πρότυπα, με πιο διαδεδομένο από αυτά την Ριβιτόλη.

Καταλαβαίνουμε, επομένως, πως το να χρησιμοποιούμε για τις Κορυφές της Βιβλιοθήκης τον Χρόνο παραμονής ως μία μοναδική αριθμητική τιμή δεν έχει την χρησιμότητα που θα επιθυμούσαμε. Ένας χρήστης ο οποίος αναζητά στην Βιβλιοθήκη για γνωστούς μεταβολίτες με αυτό το χαρακτηριστικό, θα πρέπει να μεταφράσει την τιμή του από τα δικά του πειραματικά δεδομένα σύμφωνα με οποιαδήποτε διαφορά στην ταχύτητα ανάλυσης. Δύο τεχνικές χρησιμοποιούνται για να παρακάμψουν αυτό το πρόβλημα. Πρώτη είναι η αποθήκευση των περιθωρίων λάθους της τιμής του χρόνου παραμονής και ταυτόχρονα αναζήτηση σε ένα παράθυρο τιμών. Η πρώτη αυτή τεχνική είναι πάντα χρήσιμη. Η δεύτερη είναι η κανονικοποίηση (normalization) του χρόνου παραμονής σε ένα συγκεκριμένο, σταθερό χρόνο παραμονής κάποιου εσωτερικού προτύπου όπως η Ριβιτόλη.

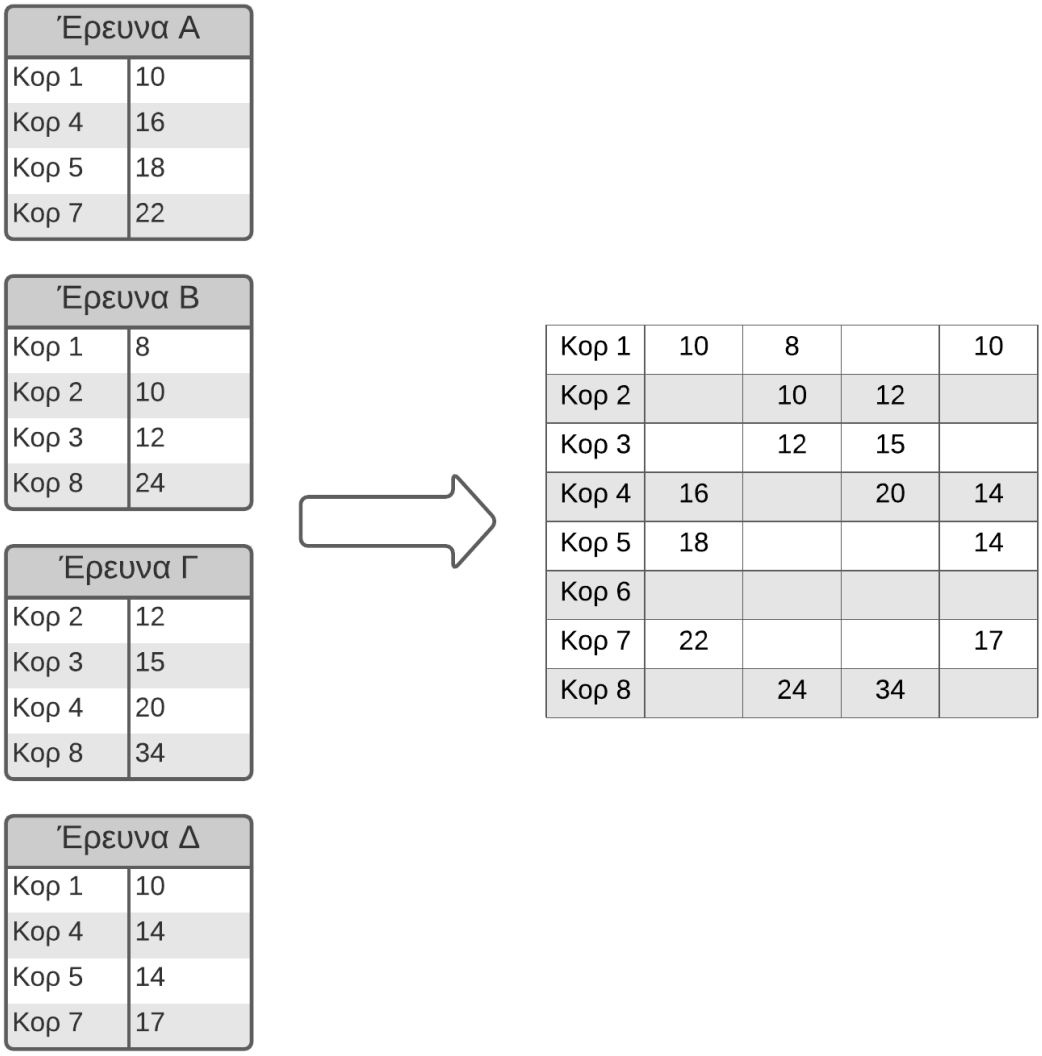
Στα δεδομένα της προυπάρχουσας βιβλιοθήκης, έχει γίνει προσπάθεια μέσω της δεύτερης τεχνικής να κανονικοποιηθούν οι χρόνοι παραμονής της κάθε Κορυφής και να αποθηκευτεί μόνο μία τιμή η οποία να εκφράζει όλες τις έρευνες. Δυστυχώς, ακόμα και αυτή η τεχνική βασίζεται σε μία παραδοχή πως η ταχύτητα ανάλυσης είναι γραμμική το οποίο δεν ισχύει πρακτικά. Στην πραγματικότητα, πυκνότερες σημεία της χρωματογραφική στήλης μπορεί να μεταθέσουν άλλα σημεία της, και άρα μόρια, σε μικρότερους ή μεγαλύτερους χρόνους.

Στα πλαίσια της εργασίας, έχουμε προσπαθήσει να εκμεταλλευτούμε μία διαφορετική ιδιότητα για να δώσουμε μία καινοτόμα λύση. Ακόμα και αν οι χρόνοι παραμονής μπορεί να ποικίλουν σαν αριθμητικές τιμές, αυτό που παραμένει σταθερό είναι η σειρά με την οποία τα διάφορα μόρια θα εμφανιστούν σαν κορυφές στην ανάλυση της στήλης, αφού η σειρά τους εξαρτάται κατά βάση από το μοριακό τους βάρος που είναι σταθερή χημική ιδιότητα.

Φανταζόμαστε λοιπόν την εξής χρήση. Αποθηκεύοντας τους χρόνους παραμονής ανά μελέτη, μπορούμε άμεσα να τις διατάξουμε. Ένας χρήστης σε αυτήν την περίπτωση, θα μπορέσει να κάνει αναζήτηση δίνοντας ήδη αναγνωρισμένους Μεταβολίτες που έχουν εμφανιστεί πριν και μετά από την Κορυφή που αναζητά. Σε αυτήν την περίπτωση μπορούμε εσωτερικά να φράξουμε την αναζήτηση και να επιστρέψουμε τις Κορυφές που έχουν χρόνους παραμονής μεταξύ των δύο γνωστών

Μεταβολιτών. Ακόμα και αν περιγράφουμε αυτήν την μέθοδο στα πλαίσια μία συγκεκριμένης μελέτης, ο σκοπός μας είναι να την χρησιμοποιήσουμε για τα δεδομένα όλων των ερευνών. Για να το καταφέρουμε αυτό, χρειάζεται να στοιχήσουμε τα δεδομένα της κάθε μελέτης βάσει των Κορυφών οι οποίες είναι κοινές μεταξύ τους. Με αυτόν τον τρόπο μπορούμε να πάρουμε μία διάταξη η οποία να ισχύει σε όλες τις έρευνες.

Σε κάθε περίπτωση χρήσης, βλέπουμε πως είναι πολύ πιο χρήσιμο στην περίπτωση του χρόνου αναμονής να κρατήσουμε όχι την υπάρχουσα κανονικοποιημένη τιμή, αλλά μία τιμή για κάθε μία μελέτη.



Εικόνα 10 – Γενικευμένη Διάταξη Κορυφών

4.2.4 Χαρακτηριστικό Ιόν

Το χαρακτηριστικό ιόν είναι επίσης κάτι το οποίο επιλέγεται από έναν πειραματιστή στα πλαίσια μίας μελέτης. Χρησιμοποιείται τόσο για την ποσοτικοποίηση των Μεταβολιτών στο τελικό Μεταβολικό Μοντέλο, αλλά επίσης, για την χρήση της Βιβλιοθήκης, μπορεί να χρησιμοποιηθεί και σαν παράμετρος αναζήτησης για την εύρεση Μεταβολιτών. Δυστυχώς, το γεγονός πως σε διαφορετικές έρευνες μπορεί να επιλεχθεί διαφορετικό από τα Ιόντα μίας Κορυφής ως το χαρακτηριστικό, έχει οδηγήσει ώστε στην προυπάρχουσα βάση να υπάρχουν σε αρκετές εγγραφές πάνω από ένα.

Μία απλή επιλογή για την λύση αυτού του προβλήματος θα ήταν να διατηρήσουμε για κάθε κορυφή μία λίστα από χαρακτηριστικά Ιόντα. Έχοντας όμως δει πως θα χρειαστούμε την έννοια των χαρακτηριστικών ανά μελέτη για τον χρόνο παραμονής, επιλέγουμε να αποθηκεύσουμε και τον ίδιο τρόπο και το χαρακτηριστικό ιόν που είχε επιλέγει για την κάθε Κορυφή.

4.2.5 Μη δομημένη πληροφορία

Σχεδόν το σύνολο των εγγραφών στην προυπάρχουσα βιβλιοθήκη στην μορφή excel περιέχουν μη δομημένη πληροφορία σε μορφή σχολίων. Μετά από συζητήσεις, καταλήξαμε πως τα περισσότερα από αυτά αφορούν την εμφάνιση των κορυφών στις διάφορες έρευνες μέσω των οποίων αναπτύχθηκε η υπάρχουσα βιβλιοθήκη. Ένα είδος σχολίων καταγράφει το όνομα στα οποία αναφέρεται σε κάθε μία από τις έρευνες η συγκεκριμένη κορυφή. Ένα δεύτερο είδος σχολίων αφορά σημειώσεις για το πως αποφασίστηκε πως δύο κορυφές από δύο έρευνες έχουν συμπτυχθεί σε μία κορυφή στην βιβλιοθήκη.

Το πρώτο είδος σχολίων που αφορούν ονόματα κορυφών μπορούμε να το καταγράψουμε ως συνώνυμο όνομα της κορυφής σαν μέρος των χαρακτηριστικών ανά μελέτη. Για το δεύτερο είδος σχολίων θα χρειαστούμε μία καινούργια Οντότητα με το Όνομα σχόλιο το οποίο να μπορεί να ανήκει σε μία Κορυφή, αλλά ίσως και σε κάθε άλλη Οντότητα. Ένα Σχόλιο παραδεχόμαστε πως συνεχίζει να είναι μη δομημένη πληροφορία άρα και δεν θα μπορέσει να χρησιμοποιηθεί σε αναζήτηση, και πως χρειαζόμαστε ένα χαρακτηριστικό για το αν είναι κατάλληλο να εμφανιστεί δημόσια σε έναν χρήστη ή θα πρέπει να μείνει ιδιωτικό.

4.2.6 Χαρακτηριστικά Μελέτης

Σύμφωνα με τα παραπάνω καταλήγουμε στην προσθήκη άλλων δύο Οντοτήτων. Η πρώτη από αυτές θα αντιπροσωπεύσει την Μελέτη, και θα χρησιμοποιηθεί κυρίως σαν αναφορά. Κάθε Μελέτη αφορά ένα Βιολογικό Σύστημα, το οποίο θα επιτρέπει να αφαιρέσουμε την σχέση του Βιολογικού Συστήματος από την Κορυφή, και να την προσθέσουμε στην ίδια την Μελέτη. Η δεύτερη, που θα ονομάσουμε Χαρακτηριστικά Μελέτης, θα υλοποιήσει μία σχέση Πολλά-προς-Πολλά με την Οντότητα της Κορυφής. Σαν χαρακτηριστικά της θα χρειαστούμε σίγουρα τα αναγνωριστικά της Κορυφής και της Μελέτης στην οποία ανήκει η κάθε εγγραφή, καθώς επίσης ο χρόνος παραμονής, το χαρακτηριστικό ιόν και το συνώνυμο όνομα της κορυφής στην αναφερόμενη μελέτη.

Εδώ να αναφέρουμε πως θα μπορούσαμε να προσθέσουμε χαρακτηριστικά στην Οντότητα της Μελέτης, με παράδειγμα αυτό του χρόνου παραμονής του εσωτερικού προτύπου. Παρ' όλ' αυτά, έχουμε αναφέρει πως ταυτόχρονα με την Βιβλιοθήκη Κορυφών την οποία σχεδιάζουμε, θα υπάρχει και ένα Αποθετήριο Πειραμάτων το οποίο είναι το κύριο το οποίο θα ασχοληθεί με τις πειραματικές παραμέτρους και τα δεδομένα της κάθε μελέτης. Έτσι, αρκούμαστε για τον σχεδιασμό μας να ορίσουμε σε κάθε Μελέτη μόνο ένα μοναδικό αναγνωριστικό, και προσβλέπουμε πως αυτή η οντότητα θα είναι η γέφυρά μας προς το Αποθετήριο Πειραμάτων.

4.2.7 Επισκόπηση και Εξέλιξη στον Χρόνο

Για να ολοκληρώσουμε την σχεδίαση της κύριας Οντότητας της Κορυφής, αξίζει να επιστήσουμε την προσοχή μας στο να προβλέψουμε ανάγκες οι οποίες θα προκύψουν κατά την διάρκεια της ζωής της Βιβλιοθήκης. Η έρευνα για την συνεχή βελτιστοποίηση των δεδομένων, ως γνωστόν, δεν σταματάει ποτέ. Καινούργιες Μελέτες θα προσθέσουν καινούργιες Κορυφές και θα δώσουν τα δεδομένα για αναγνώριση κάποιων υπαρχόντων σε μεταβολίτες. Ακόμα υπάρχει ανάγκη για εποπτεία και επισκόπηση (review) της πληροφορίας που εισέρχεται στην Βιβλιοθήκη.

Ξεκινούμε καταγράφοντας τρία χαρακτηριστικά που αφορούν μία εγγραφή. Πότε αυτή δημιουργήθηκε και από ποιόν και ποια είναι η τελευταία ημερομηνία αλλαγής. Συγκεκριμένα για την επισκόπηση μας ενδιαφέρει το ποιος έχει κάνει την τελευταία επισκόπηση και πότε, καθώς και ένα πεδίο κειμένου στο οποίο να

σημειώνονται αθροιστικά τυχόν σχόλια που αφορούν την επισκόπηση. Ένα παράδειγμα τέτοιου θα ήταν το πρότυπο που έχει χρησιμοποιηθεί για την ταυτοποίηση ενός Μεταβολίτη. Προσθέτουμε ακόμα μία λογική τιμή που να εκφράζει αν η συγκεκριμένη εγγραφή είναι αυτή τη στιγμή σε τελική κατάσταση ή χρήζει επισκόπησης. Η τιμή αυτή μας δίνει την δυνατότητα να ορίσουμε μία πολύ απλή διαδικασία. Για κάθε αλλαγή που γίνεται, όπως προσθήκη ή επεξεργασία, μαρκάρουμε την εγγραφή προς επισκόπηση μηδενίζοντας αυτή την τιμή. Όταν αυτή ολοκληρωθεί, θέτουμε πάλι την τιμή.

Υπάρχουν δύο ακόμα ενδεχόμενες διεργασίες που είναι χρήσιμες και θα συζητήσουμε, η Αντικατάσταση (Superseding) και η Απόσυρση (Obsoletion). Στην περίπτωση της Αντικατάστασης ας δούμε σαν παράδειγμα το εξής σενάριο. Μία υπάρχουσα Κορυφή που δεν είχε ως τώρα αναγνωριστεί, αναγνωρίζεται λόγω καινούργιων στοιχείων πως είναι η ίδια με μία ήδη υπάρχουσα αναγνωρισμένη Κορυφή. Σε αυτήν την περίπτωση επιθυμούμε να ενοποιήσουμε (merge) τις δύο Κορυφές. Αυτό πραγματοποιείται εύκολα με το να επιλέξουμε την μία εκ των δύο, και να την θεωρήσουμε ως αυθεντική (canonical). Το ζήτημα προς επίλυση είναι το τι θα συμβεί στην δεύτερη Κορυφή. Σε καμία περίπτωση δεν θέλουμε οποιεσδήποτε αναφορές στο αναγνωριστικό της αντικατεστημένης Κορυφής να πάψουν να λειτουργούν, άρα η εγγραφή της δεν μπορεί να αφαιρεθεί. Στον σχεδιασμό μας καλύπτουμε αυτή τη χρήση μέσου μίας σχέσης της Οντότητας Κορυφή προς τον εαυτό της, η οποία μπορεί ή όχι να έχει τιμή. Η ύπαρξη τιμής σε αυτό το πεδίο είναι σήμα πως θα πρέπει αντί για την συγκεκριμένη εγγραφή, να ανατρέξουμε στην εγγραφή που μας υποδεικνύεται.

Η δεύτερη διεργασία, αυτή της Απόσυρσης, είναι χρήσιμη σε περιπτώσεις όπου μία Κορυφή δεν θέλουμε πλέον να εμφανίζεται σαν μέρος της Βιβλιοθήκης. Μία τέτοια Κορυφή θα μπορούσε για παράδειγμα να είναι κάποια η οποία κατατέθηκε για προσθήκη στην βάση αλλά δεν ήταν τελικά κατάλληλη. Σε αυτήν την περίπτωση ο σχεδιασμός ακολουθεί μία γνωστή τακτική γνωστή ως Soft Delete. Εισάγουμε ένα πεδίο λογικής τιμής, θέτοντας το οποίο δίνουμε το σήμα πως η συγκεκριμένη εγγραφή δεν θα πρέπει να χρησιμοποιηθεί ή να εμφανίζεται.

4.2.8 Αναθεωρημένο Διάγραμμα Οντοτήτων

Εικόνα 11 – Αναθεωρημένο Διάγραμμα Οντοτήτων

4.3 Σχεδιασμός Σχήματος Βάσης – Χαρακτηριστικά κάθε Οντότητας

Παρακάτω θα συζητηθεί η μετάφραση των Οντοτήτων που μόλις σχεδιάσαμε, σε ένα Σχήμα Βάσης, κατάλληλο για υλοποίηση είτε απευθείας σε κάποια Σχεσιακή Βάση, είτε όπως θα πράξουμε, μέσω ενός Σχεσιακού Αντικειμενοστραφούς Μοντέλου. Παραθέτουμε την ονοματοδοσία στα Αγγλικά των πινάκων και των πεδίων, καθώς και συζητούμε τους τύπους των πεδίων και διάφορες τεχνικές οι οποίες χρησιμοποιήθηκαν για την υλοποίηση των σχέσεων.

4.3.1 Μεταβολίτες και Παράγωγα

Όπως έχουμε αναφέρει, η φυσική οντότητα στην οποία προσπαθούμε να αναγνωρίσουμε τις Κορυφές της Βιβλιοθήκης είναι αυτή του Μεταβολίτη για την οποία θα χρησιμοποιήσουμε έναν πίνακα **Metabolite**. Για την οντότητα Παράγωγο, θα χρησιμοποιήσουμε αντίστοιχα τον πίνακα **Derivative**.

Στην περίπτωση του Μεταβολίτη, έχουμε επιλέξει να μην δώσουμε κάποιο μοναδικό αναγνωριστικό το οποίο θα ισχύει μόνο στα όρια του εργαστηρίου, αλλά την ονοματοδοσία της βάσης χημικών μορίων βιολογικού ενδιαφέροντος ChEBI. Η κάθε οντότητα μας έχει επομένως πεδία **CHEBI_id** και **name** μορφής *CHEBI:0123456* και *Caffeine* αντίστοιχα. Το μεγάλο πλεονέκτημα αυτής της διασύνδεσης με την βάση ChEBI είναι πως από το αναγνωριστικό, μπορούμε να την χρησιμοποιήσουμε για να αντλήσουμε έναν αριθμό από ιδιότητες, όπως συνώνυμα, μοριακά βάρη, συντομογραφίες της μορφής του μορίου και άλλες.

Στην περίπτωση των Παραγώγων, δεν υπάρχει συγκεκριμένη βάση με τον βαθμό καταλληλότητας που έχει η ChEBI τα μόρια των Μεταβολιτών. Η ίδια η ChEBI πολλές φορές δεν περιέχει τα παράγωγα μόρια, είτε δεν τα ονοματοδοτεί με κάποια κοινή ονομασία. Μπορούμε όμως να ανατρέξουμε σε βάσεις όπως η Human Metabolome DataBase και η PubChem, οι οποίες αναφέρονται στα παράγωγα μόρια. Σε κάθε περίπτωση για τα Παράγωγα χρειαζόμαστε το πλήρες όνομα του μορίου σε μία από αυτές τις βάσεις **name**, καθώς και την μέθοδο παραγώγισης **derivatization_method** που εφαρμόστηκε για την δημιουργία του. Τέλος ένα χαρακτηριστικό **metabolite_id** για την υλοποίηση της σχέσης προς τους Μεταβολίτες.

4.3.2 Κορυφές και Προφίλ Ιόντων

Κατά τον σχεδιασμό των Οντοτήτων ορίσαμε σαν Κύρια οντότητα της Βιβλιοθήκης την **Κορυφή**. Ονομάζουμε τον αντίστοιχο πίνακα **GCMSPeak**. Όσον αφορά την ονοματοδοσία με κύριο αναγνωριστικό (primary id), αυτό θα πρέπει να είναι μοναδικό στα όρια του εργαστηρίου άρα και της Βιβλιοθήκης, και θα είναι αυτό με το οποίο οι χρήστες θα ανατρέχουν στην συγκεκριμένη κορυφή. Λόγω του ότι η ονοματοδοσία ανήκει στο εργαστήριο MESBL και αναφέρεται σε μεταβολική κορυφή (Metabolite Peak -> MP), χρησιμοποιούμε ένα πεδίο **mesbl_id** με μορφή *MP_000011*. Για την υλοποίηση της σχέσης προς τον πίνακα των Παραγώγων, προσθέτουμε επίσης ένα πεδίο derivative_id, το οποίο σε περίπτωση που δεν υπάρχει κάποιος αναγνωρισμένος Μεταβολίτης και Παράγωγο, επιτρέπουμε να μείνει κενός.

Ιστορικά υπάρχει ένας αριθμός από παλαιότερα ονόματα για κάθε Κορυφή, όπου θέλουμε την δυνατότητα να μπορούμε να χρησιμοποιήσουμε για την εύρεση κάποιας Κορυφής. Το σημαντικότερο από αυτά, είναι το όνομα στον προϋπάρχων πίνακα. Αυτό το όνομα θα κρατήσουμε σε ξεχωριστό πεδίο **old_mesbl_id** και είναι της μορφής *M_0001*. Περαιτέρω συνώνυμα, τα οποία αντιστοιχούν σε ονόματα Κορυφών ανά Μελέτη, θα κρατηθούν μέσω μίας λίστας αλφαριθμητικών λέξεων σε ένα ξεχωριστό πίνακα **PeakIdSynonym** με πεδία το όνομα του, **name**, και ένα **gcms_peak_id** για την υλοποίηση της σχέσης.

Προσθέτουμε επίσης ένα πεδίο (label) που να περιγράφει την αναλυτική τεχνική, **analytical_technique**, η διαφορά της οποίας αναφέραμε πως μπορεί να παράξει πολλαπλές κορυφές από ένα Παράγωγο. Αυτό, για τα υπάρχοντα δεδομένα του εργαστηρίου είναι πάντα *GC-MS*, οπότε αυτή είναι και η μοναδική δυνατή τιμή. Στο μέλλον, μπορεί να προστεθούν και άλλες τιμές όταν επεκταθούν οι τεχνικές.

Μία σημαντική ιδιότητα που εξηγήσαμε είναι ο Βαθμός Αξιοπιστίας με τον οποίο θεωρούμε πως μία κορυφή έχει αναγνωριστεί ως ένα Παράγωγο και συνεπώς ως ένας Μεταβολίτης. Εδώ χρησιμοποιούμε μία απαρίθμηση (enumeration), δηλαδή ένα πεδίο ακέραιων τιμών με όνομα **identification_grade** που αντιστοιχούν στα εξής επίπεδα και τις περιγραφές τους.

Α/Α	ΟΝΟΜΑ	ΠΕΡΙΓΡΑΦΗ
1	IDENTIFIED	Πλήρως αναγνωρισμένη και πιστοποιημένη Κορυφή
2	PUTATIVE	Αναγνωρισμένη, με πιθανότητα σφάλματος
3	CHEMCLASSPRED	Άγνωστη, με κάποιες χημικές ιδιότητες αναγνωρισμένες
4	UNKNOWN	Άγνωστη
5	SUBMITTED	Νέα κορυφή που έχει καταχωρηθεί αλλά δεν έχει βαθμό

Πίνακας 2 – Τιμές Πεδίου Βαθμού Αξιοπιστίας

Για την αποθήκευση του προφίλ Ιόντων θα χρησιμοποιηθεί μία λίστα μοριακών βαρών σε σχέση ένα προς πολλά σε ένα ξεχωριστό πίνακα με όνομα **PeakIon** που αντιστοιχεί στην οντότητα Ιόν, με πεδία το ακέραιο μοριακό βάρος **molecular_weight** και ένα **gcms_peak_id** για την υλοποίηση της σχέσης.

Για λόγους ιστορικού ελέγχου (auditing) και για να βοηθήσουμε διαδικασίες επισκόπησης (reviewing) προσθέτουμε ένα σύνολο από πεδία. Από αυτά, τα **created_by**, **created_date**, **modified_date** είναι αυτοματοποιημένα και αναφέρονται στο πότε και από ποιόν δημιουργήθηκε μία εγγραφή, και πότε τελευταία αυτή άλλαξε. Αντίστοιχα, τα **reviewed**, **reviewed_by**, **reviewed_date** και **review_comment** δίνουν την δυνατότητα σε κάθε αλλαγή που χρήζει επισκόπησης να θέτουμε κενό το πεδίο reviewed, μέχρι κάποιος να μπορεί να το ελέγξει.

Για την υλοποίηση της Αντικατάστασης προσθέτουμε ένα πεδίο με όνομα **replaced_by_id** και προορισμό επίσης μία Κορυφή. Η ύπαρξη τιμής σε αυτό το πεδίο είναι σήμα πως θα πρέπει αντί για την συγκεκριμένη εγγραφή, να ανατρέξουμε στην εγγραφή που μας υποδεικνύεται. Για την δεύτερη διεργασία, αυτή της Απόσυρσης, εισάγουμε ένα πεδίο λογικής τιμής **obsolete**, θέτοντας το οποίο δίνουμε το σήμα πως η συγκεκριμένη εγγραφή δεν θα πρέπει να χρησιμοποιηθεί.

4.3.3 Χαρακτηριστικά Μελέτης και Βιολογικά Συστήματα

Εκτός από το προφίλ των Ιόντων, άλλα δύο πολύ σημαντικά αναγνωριστικά χαρακτηριστικά, ο χρόνος παραμονής στην στήλη και το χαρακτηριστικό (marker) ιόν, διαφέρουν ανά μελέτη και επομένως δεν μπορούν να εισαχθούν σαν απλά πεδία σε μία Κορυφή. Έχουμε αποφασίσει πως αυτά τα χαρακτηριστικά θα αποθηκευτούν μέσω δύο Οντοτήτων, Μελέτη και Χαρακτηριστικά Μελέτης σε σχέση μεταξύ τους.

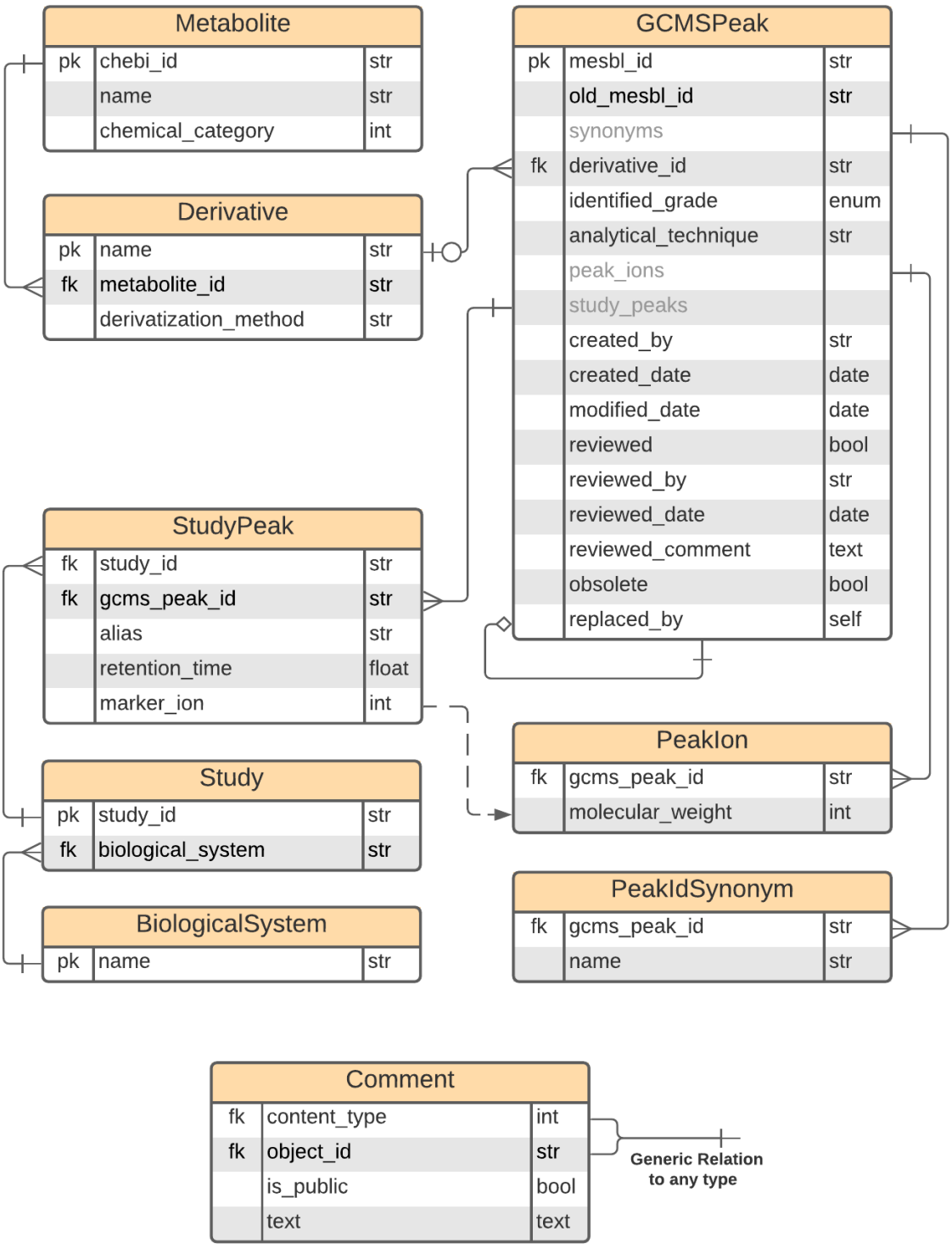
Για την οντότητα Μελέτη δημιουργούμε τον πίνακα **Study**. Σε αυτήν την φάση, δεν χρειαζόμαστε πολλές πληροφορίες σε αυτήν την οντότητα, παρά μόνο ένα μοναδικό χαρακτηριστικό **study_id** που μπορεί να είναι το όνομα του Ερευνητή, καθώς και το Βιολογικό Σύστημα το οποίο εξετάστηκε, τα οποία ξεχωρίζουμε επίσης στην δική του Οντότητα **BiologicalSystem** που περιέχει μόνο ένα αναγνωριστικό πεδίο **name** με το όνομα του.

Έχοντας πλέον την οντότητα της Μελέτης, δημιουργούμε ένα πίνακα σχέσης Πολλά-προς-πολλά μεταξύ Κορυφών και Ερευνών **StudyPeak**. Στον πίνακα αυτής της σχέσης, εκτός των πεδίων αναφοράς **study_id**, **gcms_peak_id**, θα καταγράψουμε τα χαρακτηριστικά των κορυφών τα οποία διαφέρουν ανάλογα με την μελέτη. Στο πρώτο πεδίο **alias** καταγράφουμε, αν υπάρχει, το χαρακτηριστικό όνομα που δόθηκε στην κορυφή στα πλαίσια της συγκεκριμένης Μελέτης. Σε ένα δεύτερο πεδίο **retention_time** αποτυπώνει τον Χρόνο Παραμονής στην Χρωματογραφική Στήλη του GC-MS. Σε ένα τρίτο πεδίο καταγράφουμε το Χαρακτηριστικό Ιόν, που χρησιμοποιείται εξίσου για την αναγνώριση της Κορυφής. Εδώ θα μπορούσαμε να χρησιμοποιήσουμε μία σχέση προς τον πίνακα των Ιόντων. Επειδή όμως αυτός ο πίνακας περιέχει ουσιαστικά μόνο μοριακά βάρη, επιλέγουμε για απλότητα να χρησιμοποιήσουμε ένα πεδίο ακέραιου αριθμού **marker_ion** το οποίο επίσης να καταγράφει μοριακό βάρος, έχοντας υπόψιν μας πως θα πρέπει να υλοποιήσουμε επίσης έναν περιορισμό ώστε το νούμερο που θα δοθεί να ανήκει σαν μοριακό βάρος στα Ιόντα της Κορυφής.

4.3.4 Προσθήκη Σχολίων σε κάθε Οντότητα

Για τον σκοπό αυτό θα χρησιμοποιήσουμε μία πρότυπη τεχνική (pattern) ονομαζόμενη Σχέση Γενικής Χρήσης (Generic Relation). Σε αυτήν, σε αντίθεση με μία τυπική Σχέση Ένα-προς-Πολλά μέσω Εξωτερικού Κλειδιού (Foreign Key), χρησιμοποιούμε δύο πεδία για να υλοποιήσουμε την σχέση. Τον τύπο της Οντότητας που αναφερόμαστε, **content_type**, καθώς και το αναγνωριστικό της εγγραφής στον πίνακα της, **object_id**. Με αυτόν τον τρόπο, μπορούμε να προσθέσουμε κάποιο Σχόλιο σε οποιονδήποτε πίνακα. Μαζί με αυτά, προσθέτουμε στην οντότητα **Comment**, ένα πεδίο ελεύθερου κειμένου **text**, καθώς και έναν δείκτη **is_public** για τον διαχωρισμό τους ως δημόσια.

4.3.5 Τελικό Σχήμα Βάσης



Εικόνα 12 – Τελικό Σχήμα Βάσης

4.4 Τεχνολογικές Επιλογές

Έχοντας υπόψιν του στόχους υλοποίησης στα πλαίσια της εργασίας, αλλά και μελλοντικούς στόχους που ζητούν η Βιβλιοθήκη να διατεθεί σαν μία Διαδικτυακή Εφαρμογή και Υπηρεσία, καλούμαστε να επιλέξουμε εξ' αρχής κάποιες από τις χρησιμοποιούμενες τεχνολογίες που θα μας επιτρέψουν να επιτύχουμε αυτούς τους στόχους.

Δύο πολύ σημαντικά κριτήρια που είναι κοινά στις επιλογές μας, είναι η διαχρονικότητα τους, καθώς και το επίπεδο τεχνογνωσίας που θα χρειαστεί κάποιος για την υποστήριξη και την συνέχιση της εξέλιξης της Βιβλιοθήκης σε βάθος χρόνου. Θα ήταν πολύ δελεαστικό να στηρίξουμε την υλοποίηση μας σε πρόσφατες (bleeding edge) τεχνολογίες διαδικτύου, όπως θα κάναμε αν η παραδοτέα βάση και εφαρμογή απευθυνόταν καθαρά σε Μηχανικούς Υπολογιστών. Στα πλαίσια ενός εργαστηρίου Αναλυτικής Βιολογίας, όμως, προτεραιότητα έχει το να βρεθούν κοινοί τεχνολογικοί παρονομαστές. Μία bleeding edge εφαρμογή θα χρειαζόταν συνεχή υποστήριξη και θα γερνούσε πολύ γρήγορα αν δεν την είχε. Μία διαχρονικά δοκιμασμένη λύση από την άλλη, ακόμα και μίας γενιάς παλαιότερης, ενδείκνυται πολύ περισσότερο.

4.4.1 Γλώσσα Προγραμματισμού Python

Για την επιλογή της γλώσσας προγραμματισμού για την υλοποίηση μπορούμε να συγκρίνουμε τις εξής πιθανές

PHP – Η γλώσσα προγραμματισμού PHP είναι μία από τις πιο γνωστές γλώσσες που στοχεύουν στην υλοποίηση Ιστότοπων και αυτή η οποία πρώτη ήταν μαζικά προσβάσιμη για αυτό το σκοπό. Τα υπέρ αυτής της γλώσσας είναι η απλότητά της και η αμεσότητα με την οποία μπορεί να παράξει περιεχόμενο ιστοσελίδων. Αν και σε καμία περίπτωση ξεπερασμένη, η PHP είναι μία γλώσσα που προηγείται πολλών σημερινών πρακτικών για την δομημένη ανάπτυξη Διαδικτυακών Εφαρμογών. Η απλότητά της εστιάζει επίσης πολύ στην παραγωγή περιεχομένου, ενώ η υλοποίηση ενός μοντέλου δεδομένων σαν αυτό που θέλουμε να υλοποιήσουμε στην καρδιά της εφαρμογής μας να είναι αρκετά πιο δύσκολο.

Javascript (Node.JS) – Εδώ βρισκόμαστε στον αντίποδα της PHP, αφού μιλάμε για την πιο πρόσφατη τρέχουσα τεχνολογία η οποία και ανεβαίνει σε

δημοτικότητα κατά την τελευταία δεκαετία. Η αρχική χρήση της Javascript ήταν, εκτελούμενη από την πλευρά του χρήστη εσωτερικά του περιηγητή, να προσθέτει δυνατότητες ασύγχρονης επικοινωνίας με τους Διακομιστές που προσέφεραν έναν ιστότοπο. Έτσι, έδωσε τη δυνατότητα σε ιστοσελίδες να προσφέρουν διαδραστικότητα, γεννώντας την έννοια της σύγχρονης εφαρμογής. Με την έλευση της Node.JS, η ίδια γλώσσα μπορούσε να χρησιμοποιηθεί και στην πλευρά του διακομιστή, και επιτρέποντας τους προγραμματιστές να χρησιμοποιούν μία γλώσσα παντού. Το οικοσύστημα της Javascript, σήμερα, προσφέρει επίσης κάποιες από τις πιο δυνατές βιβλιοθήκες εφαρμογών μαζικής χρήσης όπως η Angular της Google και η React της Facebook.

Δυστυχώς, οι αυξημένες δυνατότητες που προσφέρει το οικοσύστημα της Javascript μεταφράζονται και σε αυξημένη πολυπλοκότητα και ανάγκη εξειδικευμένης γνώσης για την αξιοποίησή τους. Οι δυνατότητες της για την εξυπηρέτηση μαζικού αριθμού χρηστών, που είναι το δυνατότερό τους σημείο, δεν προσδίδει κάτι στην εφαρμογή που θέλουμε να φτιαχτεί στο μέλλον. Παρ' όλ' αυτά, ο σχεδιασμός της βάσης μας έχει γίνει χρησιμοποιώντας αρχές οι οποίες είναι τρέχουσες και μεταφράσιμες σε αυτό το οικοσύστημα. Μία ειδικευμένη ομάδα θα μπορούσε να φτιάξει μία εξελιγμένη εφαρμογή με βάση τη Javascript, βασιζόμενη στην υλοποίηση αυτής της εργασίας, χωρίς να χρειαστεί να ξεκινήσει από την αρχή.

Python – Η Γλώσσα Python είναι η επιλογή μας, κυρίως για την απλότητά της και την δημοτικότητά της στον Ακαδημαϊκό κόσμο. Σήμερα, από οποιοδήποτε Πανεπιστημιακό Τμήμα το οποίο προσφέρει έστω και ένα μάθημα προγραμματισμού, μπορείς να περιμένεις αποφοίτους που να την γνωρίζουν. Το οικοσύστημα επιστημονικών βιβλιοθηκών που την περιστοιχίζει, της έχει δώσει επίσης μεγάλη δημοτικότητα στον κόσμο της έρευνας, και χρησιμοποιείται για τεράστιο εύρος εφαρμογών, από Μοντελοποίηση Φυσικών φαινομένων μέχρι τον πολύ επίκαιρο τομέα της Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης. Προσφέρει επίσης δύο πολύ ικανές βιβλιοθήκες για ανάπτυξη διαδικτυακών εφαρμογών όπως θα δούμε παρακάτω. Στον τομέα του διαδικτύου μπορεί η Python να μην είναι η άριστη, αλλά το αναπληρώνει τόσο με την προσβασιμότητά της, όσο και με την δυνατότητά της να έρθει σε διεπαφή με το πλήθος των επιστημονικών βιβλιοθηκών.

4.4.2 Βιβλιοθήκη Διαδικτυακών Εφαρμογών Django

Ακόμα και αν ο κύριος στόχος της παρούσας εργασίας είναι η σχεδίαση και ανάπτυξη της βάσης που θα στεγάσει την Βιβλιοθήκη Κορυφών, είναι η γνώμη του συγγραφέα πως είναι προς το συμφέρον μας να ανέβουμε ένα επίπεδο αφαίρεσης και να κάνουμε την υλοποίηση μέσω ενός Μοντέλου Αντικειμενοστραφούς Σχεσιακής Αντιστοίχισης δεδομένων (Object Relational Mapper – ORM). Σε σύγκριση με μία υλοποίηση καθαρά σε SQL γλώσσα, η ανάπτυξη ενός Μοντέλου Δεδομένων έχει σαφή πλεονεκτήματα. Μας δίνει, αρχικά, την δυνατότητα να περιγράψουμε το Μοντέλο Δεδομένων μέσω πολύ πιο απλό αντικειμενοστραφή κώδικα. Δεύτερον, μας δίνεται η δυνατότητα να καταγράψουμε, εκτός από το σχήμα (schema) και λογικούς κανόνες (business logic) που τα διέπουν. Τέλος, βασισμένοι σε ένα Σχεσιακό Μοντέλο, μπορούμε να αναπτύξουμε εύκολα τόσο μία διαδικτυακή εφαρμογή χρήστη, όσο και μία διαδικτυακή υπηρεσία, που είναι μελλοντικές χρήσεις της βάσης.

Αξίζει λοιπόν να επιλέξουμε εξ' αρχής μία Βιβλιοθήκη Ανάπτυξης Διαδικτυακών Εφαρμογών, κυρίως βάσει των δυνατοτήτων Αντικειμενοστραφούς Σχεσιακού Μοντέλου που προσφέρουν. Οι κύριες διαθέσιμές βιβλιοθήκες στην γλώσσα Python είναι δύο.

Flask / SQLALchemy – Το Flask αποτελεί μία ευέλικτη δομικά (modular) βιβλιοθήκη ανάπτυξης εφαρμογών, βασισμένη στην φιλοσοφία ενός πολύ ελαφρού πυρήνα λειτουργιών, που μπορεί να επεκταθεί μέσω πρόσθετων βιβλιοθηκών. Δεν περιέχει αυτούσια κάποιο ORM, αλλά συνήθως για αυτό το λόγο χρησιμοποιείται μαζί του η βιβλιοθήκη SQLAlchemy της Python.

Django – Αυτή η Βιβλιοθήκη Ανάπτυξης ακολουθεί την αντίθετη στρατηγική, να περιέχει όλα τα εργαλεία που μπορεί να χρειαστούν για την Ανάπτυξη μίας εφαρμογής, και αυτός είναι ο λόγος που είναι η επιλογή μας. Το Django είναι μία ευρέως γνωστή Βιβλιοθήκη στον κόσμο της Python, η οποία βασίζεται πάνω στην τεχνική Model-View-Controller (MVC), όπως και άλλες άμεσα συγκρίσιμες βιβλιοθήκες σε άλλες γλώσσες της γενιάς της, όπως η Ruby on Rails. Μεταξύ τους, αυτές οι δύο βιβλιοθήκες αποτέλεσαν industry standards για την ανάπτυξη διαδικτυακών εφαρμογών για πάνω από μία δεκαετία, πριν την πρόσφατη έλευση της Javascript και της μεθοδολογίας που οι βιβλιοθήκες ανάπτυξης της ακολουθούν.

Για τους σκοπούς μας, υπάρχουν αρκετά στοιχεία διαθέσιμα στο Django τα οποία μας βοηθούν. Το MVC, ως σχεδιαστικό πρότυπο, επιβάλλει το διαχωρισμό του Μοντέλου Δεδομένων από την υπόλοιπη Εφαρμογή. Σαν αποτέλεσμα, το Σχεσιακό Μοντέλο που προσφέρει το Django είναι άρτιο. Το πιο δελεαστικό στοιχείο, το οποίο προσφέρει, όμως είναι πως μπορεί να δημιουργήσει αυτόματα μία Εφαρμογή Διαχείρισης της Βάσης. Αυτό μας δίνει άμεσα την δυνατότητα, υλοποιώντας την βάση να μπορέσουμε να παραδώσουμε μαζί και μία εφαρμογή διαχείρισής της.

Σαν αποτέλεσμα, με την χρήση του Django, η παραδοτέα Βιβλιοθήκη Κορυφών θα είναι άμεσα χρησιμοποιήσιμη από το εργαστήριο για ανάγνωση και επεξεργασία των δεδομένων της, ακόμα και πριν αναπτυχθεί η μετέπειτα εφαρμογή για την παρουσίασή της στο ευρύ κοινό.

4.4.3 Σχεσιακή Βάση Δεδομένων MariaDB

Το Django και το Σχεσιακό Μοντέλο το οποίο προσφέρει δημιουργούν ένα επίπεδο αφαίρεσης πάνω από την πραγματική βάση δεδομένων. Το ίδιο το Μοντέλο είναι ικανό να παράγει τον κώδικα SQL για να δημιουργήσει τους πίνακες και να διαχειριστεί τις εγγραφές. Αυτό μας δίνει την δυνατότητα να κάνουμε την επιλογή μας με πολύ πιο απλά κριτήρια.

Η Σχεσιακή Βάση Δεδομένων MariaDB, πρόκειται για την ΕΛ/ΛΑΚ εκδοχή της γνωστής Βάσης MySQL της Oracle. Είναι η πιο διαδεδομένη και ευρέως γνωστή Σχεσιακή Βάση που χρησιμοποιείται σήμερα. Για εμάς αυτό σημαίνει πως μία τέτοια βάση είναι ταυτόχρονα γνώριμη στον οποιονδήποτε θα χρειαστεί στο μέλλον να ασχοληθεί με την υποστήριξη της, πως υπάρχουν έτοιμες λύσεις διαχείρισης και αντιγράφων ασφαλείας, καθώς και πως θα υπάρχει διαθέσιμη διανομή της σε οποιαδήποτε πλατφόρμα υπηρεσιών (νέφος, docker) επιλέξουμε να χρησιμοποιήσουμε.

4.3.4 Πλατφόρμα Διαδικτυακών Υπηρεσιών Docker

Στον σημερινό κόσμο της πληροφορικής και ειδικά στον χώρο της ανάπτυξης διαδικτυακών εφαρμογών, θίγονται ζητήματα όσον αφορά τον κύκλο ανάπτυξης του λογισμικού, από τον ηλεκτρονικό υπολογιστή του προγραμματιστή, μέχρι την εγκατάσταση του πηγαίου κώδικα σε μία μηχανή Διακομιστή και την διάθεση του στο κοινό. Ένα από τα πιο κοινά προβλήματα, εκφράζεται σε απλή γλώσσα ως "Πώς μπορώ

να διασφαλίσω πως το προγραμματιστικό μου περιβάλλον είναι εύκολο στην εγκατάσταση και σταθερό, και ταυτόχρονα αντικατοπτρίζει πλήρως το περιβάλλον του Διακομιστή πάνω στο οποίο θα εγκατασταθεί".

Η πλατφόρμα Docker προσφέρει μία ολική λύση σε αυτό το ζήτημα, με το να επιτρέπει στον προγραμματιστή να ορίσει σε κώδικα ένα περιβάλλον το οποίο να είναι κατάλληλο και για τους δύο σκοπούς. Στον κώδικα αυτό περιγράφονται το λειτουργικό σύστημα, κυρίως κάποια διανομή Gnu/Linux, τα πακέτα του λειτουργικού που χρειάζονται, όπως γλώσσες προγραμματισμού και υπηρεσίες όπως βάσεις δεδομένων. Η περιγραφή αυτή έπειτα μπορεί να χρησιμοποιηθεί και σαν τοπικό προγραμματιστικό περιβάλλον, αλλά πιο σημαντικά, να χρησιμοποιηθεί σαν ο τελικός Διακομιστής της εφαρμογής στο Διαδίκτυο.

4.5 Υλοποίηση Κώδικα

Στην ενότητα αυτή κάνουμε μία σύντομη αναφορά στην προγραμματιστική υλοποίηση του σχήματος που σχεδιάστηκε, όπως και στην υλοποίηση της εφαρμογής διαχείρισης. Ταυτόχρονα με την πτυχιακή εργασία, έχει κατατεθεί ο πλήρης πηγαίος κώδικας της εφαρμογής, λεπτομέρειες για την χρήση του οποίου μπορούν να βρεθούν στο Παράρτημα Α.

4.5.1 Μοντέλα Αντικειμενοστραφούς – Σχεσιακής Αντιστοίχισης

Για να δημιουργήσουμε το μοντέλο μίας Οντότητας στο Django, χρειάζεται να δημιουργήσουμε μία υποκλάση της κύριας κλάσης django.db.models.Model, στην οποία θα προσθέτουμε πεδία ως χαρακτηριστικά κλάσης (class attributes). Οι τύποι πεδίων συμπίπτουν με τους αντίστοιχους τύπους και της Python, αλλά και της SQL βάσης στην οποία θα αποθηκευτούν. Σαν παραδείγματα πεδίων έχουμε

- **IntegerField** – αποθηκεύει έναν ακέραιο αριθμό
- **CharField(max_len)** – Αλφαριθμητικό, αντίστοιχο του SQL VarChar
- **TextField** – Ελεύθερο Κείμενο, αντίστοιχο του τύπου Text

Στην περίπτωση έκφρασης σχέσεων, υπάρχουν ειδικοί τύποι οι οποίοι τις υλοποιούν. Για παράδειγμα για την έκφραση μία σχέσης ένα-προς-πολλά, μία Κορυφή που μπορεί να έχει ένα ή παραπάνω Συνώνυμα, σε καθαρή SQL θα έπρεπε να δημιουργήσουμε ένα πεδίο gcms_peak_id και έπειτα μία σχέση Εξωτερικού κλειδιού. Στο Django μπορούμε να εκφράσουμε τη σχέση απλά με ένα πεδίο των εξής τύπων

- **ForeignKey** – Σχέση Ένα-προς-Πολλά μέσω Εξωτερικού Κλειδιού
- **OneToOneField** – Σχέση Ένα-προς-Ένα
- **ManyToManyField** – Σχέση Πολλά-προς-Πολλά, με δημιουργία ενδιάμεσου πίνακα

Στην περίπτωση της Οντότητας που εκφράζει Στοιχεία Κορυφής ανά Μελέτη, θα μπορούσαμε για παράδειγμα να πούμε πως πρόκειται για μία σχέση Πολλά-προς-Πολλά μεταξύ των Οντοτήτων Κορυφής και Μελέτης, πάνω στην οποία εμείς επιθυμούμε να αποθηκεύσουμε περαιτέρω πληροφορία. Σε αυτήν την προχωρημένη περίπτωση, δεν χρησιμοποιήσαμε τον έτοιμο μηχανισμό του ManyToManyField, αλλά δημιουργήσαμε μία ξεχωριστή Οντότητα με δύο ξένα κλειδιά ώστε να ανήκει στον συνδυασμό μια Κορυφής και μίας Μελέτης.

Μια περαιτέρω τεχνική η οποία χρησιμοποιήσαμε, αφορά πεδία τα οποία μπορούν να πάρουν τιμή μέσα από ένα προκαθορισμένο σύνολο. Σε αυτήν την περίπτωση μπορούμε να χρησιμοποιήσουμε μία εσωτερική κλάση η οποία να περιέχει τις δυνατές επιλογές και να την χρησιμοποιήσουμε για να ορίσουμε ένα Αλφαριθμητικό πεδίο. Στην περίπτωση του Βαθμού Αξιοπιστίας μάλιστα, επιλέξαμε να δώσουμε μία ακέραια τιμή σε κάθε επίπεδο (Enumeration) και να αποθηκεύσουμε αυτόν τον ακέραιο σε ένα αντίστοιχο πεδίο.

Παραδείγματα αυτών το τεχνικών παρατίθενται στο δείγμα κώδικα που ακολουθεί, και μπορούν αντίστοιχα να βρεθούν και στον Πηγαίο Κώδικα.

4.5.2 Εφαρμογή Διαχείρισης

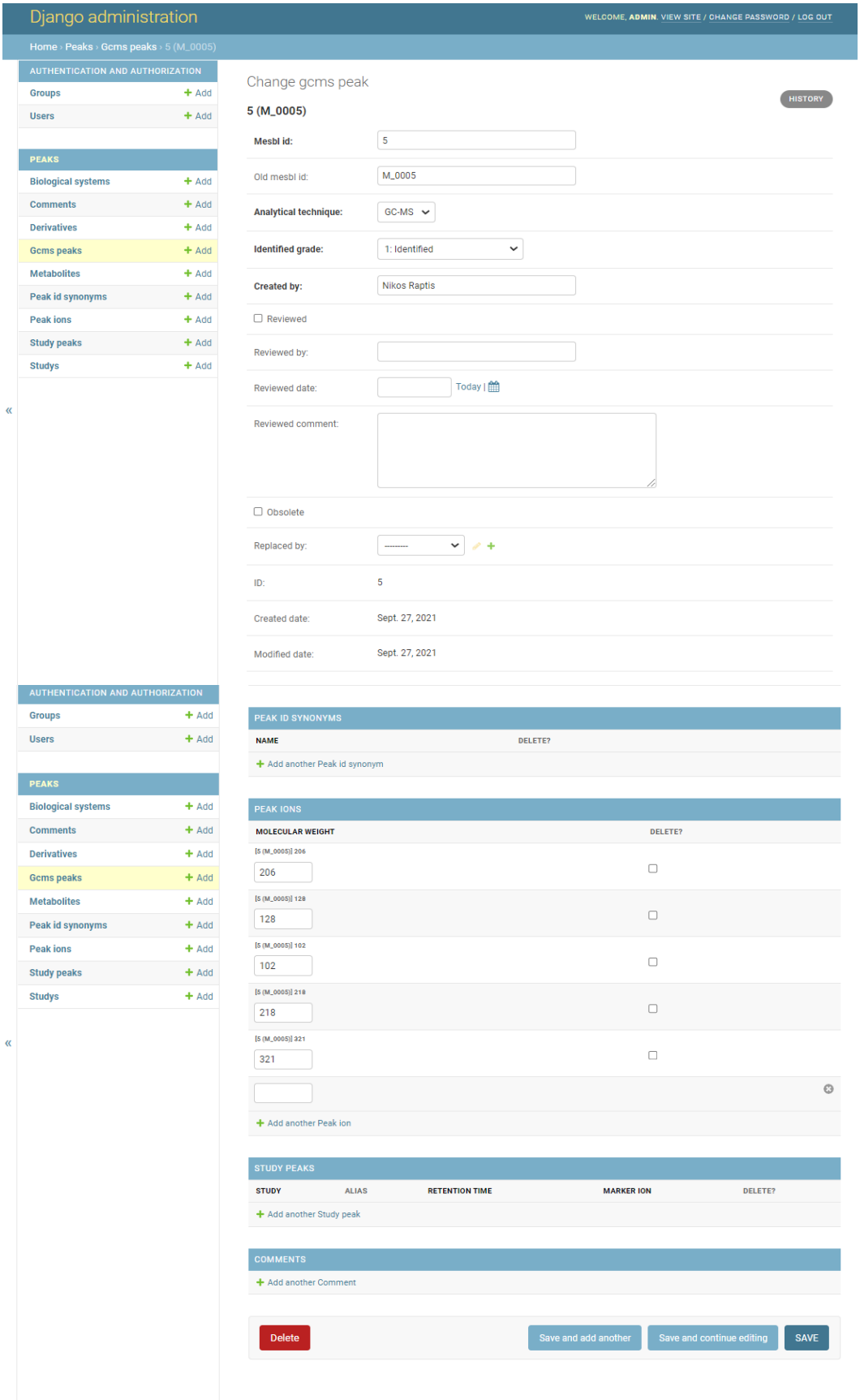
Το Django μας επιτρέπει για κάθε ένα Μοντέλο Οντότητας που έχουμε προγραμματίσει, να εισάγουμε και μία φόρμα στην εφαρμογή διαχείρισης με αυτόματο τρόπο. Σε σχέση με μία βάση ενός πίνακα όμως, η διαχείριση πολλαπλών εγγραφών για την αποτύπωση μίας Κορυφής εισάγει πολυπλοκότητα.

Για παράδειγμα, έχουμε κατά τον σχεδιασμό μας έχουμε δώσει στην Οντότητα της Κορυφής, εκτός από έναν αριθμό ιδιοτήτων που της ανήκουν, και την δυνατότητα να περιέχει Ιόντα, Σχόλια, Συνώνυμα αναγνωριστικά, και Στοιχεία ανά Μελέτη. Για την προσθήκη αυτών θα έπρεπε να δημιουργηθεί πρώτα μία Κορυφή, και έπειτα για κάθε μία από τις δευτερεύουσες οντότητες που της ανήκουν να περιηγηθούμε σε διαφορετική σελίδα ώστε να τις δημιουργήσουμε, επιλέγοντας να τις συσχετίσουμε με την Κορυφή που μόλις έχουμε φτιάξει.

Για να διευκολύνουμε την χρήση της εφαρμογής, χρησιμοποιούμε μία δυνατότητα της εφαρμογής διαχείρισης η οποία μας επιτρέπει να δημιουργήσουμε τις δευτερεύοντες Οντότητες στην ίδια σελίδα με την δημιουργία της Κορυφής.

```
class GCMSPeak(models.Model):

    class IdentifiedGrade(models.IntegerChoices):
        IDENTIFIED = 1, "1: Identified"
        PUTATIVE = 2, "2: Putative"
        CHEMCLASSPRED = 3, "3: Chemical Class Predicted"
        UNKNOWN = 4, "4: Unknown"
        SUBMITTED = 5, "5: Newly Submitted"

    class AnalyticalTechnique(models.TextChoices):
        GCMS = 'GC-MS', 'GC-MS'

    mesbl_id = models.CharField(max_length=16, unique=True)
    old_mesbl_id = models.CharField(max_length=16, blank=True)
    # synonyms by ForeignKey

    analytical_technique = models.CharField(
        max_length=16, choices=AnalyticalTechnique.choices)
    identified_grade = models.IntegerField(
        choices=IdentifiedGrade.choices, default=IdentifiedGrade.IDENTIFIED)

    created_by = models.CharField(max_length=128, default='Nikos Raptis')
    created_date = models.DateField(auto_now_add=True)
    modified_date = models.DateField(auto_now=True)

    reviewed = models.BooleanField(default=False)
    reviewed_by = models.CharField(max_length=128, blank=True)
    reviewed_date = models.DateField(null=True, blank=True)
    reviewed_comment = models.TextField(blank=True)

    obsolete = models.BooleanField(default=False)
    replaced_by = models.ForeignKey('self', null=True, blank=True,
on_delete=models.CASCADE)


class PeakIdSynonym(models.Model):
    gcms_peak = models.ForeignKey(GCMSPeak, on_delete=models.CASCADE)
    name = models.CharField(max_length=16)


class PeakIon(models.Model):
    gcms_peak = models.ForeignKey(GCMSPeak, on_delete=models.CASCADE)
    molecular_weight = models.IntegerField()


class Study(models.Model):
    study_id = models.CharField(max_length=16, unique=True)


class StudyPeak(models.Model):
    study = models.ForeignKey(Study, on_delete=models.CASCADE)
    gcms_peak = models.ForeignKey(GCMSPeak, on_delete=models.CASCADE)
    alias = models.CharField(max_length=16, blank=True)
    retention_time = models.FloatField()
    marker_ion = models.IntegerField()

    def clean(self):
        print(self.gcms_peak.marker_ion_weights())
        if self.marker_ion not in self.gcms_peak.marker_ion_weights():
            raise ValidationError(_('The Marker Ion must be an Ion of the
referenced Peak. '))
```

Εικόνα 13 – Δείγμα κώδικα Σχεσιακού Μοντέλου



Εικόνα 14 – Δείγμα Εφαρμογής Διαχείρησης

4.6 Μεταφορά του Συνόλου των Κορυφών

4.6.1 Αναγνωριστικά Μεταβολιτών και Παραγώγων

Μία από της εργασίες που ήταν αναγκαίο να γίνουν πριν την μεταφορά των υπαρχόντων δεδομένων της βιβλιοθήκης στην βάση, ήταν η ταυτοποίηση των φυσικών οντοτήτων του Μεταβολίτη και του Παραγώγου και η ονοματοδοσία τους σύμφωνα με εξωτερικές βάσεις. Δίνοντας τα κατάλληλα αναγνωριστικά, πετυχαίνουμε την δυνατότητα διασύνδεσης της προτυποποιημένης βιβλιοθήκης κορυφών με το παγκόσμιο ιστό δεδομένων που αναφέραμε στην εισαγωγή. Αυτή η ταυτοποίηση ήταν, φυσικά, δυνατή μόνο για τις Κορυφές οι οποίες έχουν βαθμό αξιοπιστίας 1 και 2. Είναι δηλαδή Κορυφές στις οποίες οι Μεταβολίτες και τα Παράγωγά τους έχουν αναγνωριστεί.

Για τους Μεταβολίτες, όπως έχουμε πει, το ζητούμενο είναι να ταυτοποιηθούν μέσω της βιβλιοθήκης χημικών οντοτήτων βιολογικού ενδιαφέροντος ChEBI. Η διαδικασία η οποία ακολουθήθηκε για τον καθένα ήταν να γίνει εύρεση του μορίου σύμφωνα με το όνομα με το οποίο ο Μεταβολίτης συναντάται στα δεδομένα της υπάρχουσας βάσης. Για πολλά από αυτά τα μόρια το όνομα τους επέστρεψε περισσότερες της μίας εγγραφές που πολλές φορές αναφέρονται σε ισομερή, και χρειάστηκε να γίνει επιλογή μεταξύ τους. Μετά την ταυτοποίηση, τόσο το όνομα του μεταβολικού μορίου όσο και το μοναδικό αναγνωριστικό του προστέθηκαν ως στήλες στα υπάρχοντα δεδομένα, μαζί με χημικές ιδιότητες όπως η χημική κατηγορία παραγώγισης.

Για τα Παράγωγα, μία ένα-προς-ένα ταυτοποίηση δεν ήταν δυνατή, κυρίως για τον λόγο πως τα παράγωγα μόρια δεν συναντώνται απαραίτητα στην φύση, παρά μόνο στην διαδικασία ανάλυσης, και γι' αυτό δεν συμπεριλαμβάνονται πάντα σαν οντότητες στην βιβλιοθήκη της ChEBI. Ακόμα και αν βρεθούν σε αυτήν, τις περισσότερες φορές έχουν ονόματα τα οποία υποδηλώνουν την χημική τους σύσταση, και όχι όπως επιθυμούμε εμείς ένα όνομα το οποίο υποδηλώνει την ταυτότητά τους ως παράγωγα ενός συγκεκριμένου μεταβολικού μορίου, η οποία τυπικά περιέχει το όνομα του Μεταβολίτη και επίθεμα *{1,2,..}TMS*. Για να βρούμε ένα τέτοιο όνομα, η διαδικασία που ακολουθήσαμε ήταν να τα αναζητήσουμε διαδοχικά σε τρεις βάσεις, πρώτον την ChEBI, έπειτα στην αντίστοιχη βάση PubChem, και τέλος στην βάση Human Metabolome DataBase και από τα ονόματά τους να επιλέξαμε το πιο κατάλληλο.

4.6.2 Μεταφορά Κορυφών και Ιόντων

Έχοντας τα ονόματα των αναγνωρισμένων Μεταβολιτών και Παραγώγων σαν πρόσθετες στήλες, ο πίνακας μας πλέον είχε όλη την απαιτούμενη πληροφορία ώστε να δημιουργήσουμε εγγραφές των Οντοτήτων Κορυφή, Ιόν, Μεταβολίτης και Παράγωγο, για τις 430 κορυφές που πληρούν τις προδιαγραφές πλήρους επόπτευσης, και να εισαχθούν στην Βιβλιοθήκη.

Η διαδικασία αυτή έγινε αυτοματοποιημένα. Υλοποιήθηκε μία εντολή διαχείρισης **excelmigrate** η οποία ακολουθεί τον εξής αλγόριθμο. Για κάθε μία εγγραφή του αρχικού πίνακα, δημιουργείται μία εγγραφή Κορυφής με τα αντίστοιχα χαρακτηριστικά της. Για κάθε μία στήλη του περιγράφει ιόν, δημιουργείται μία εγγραφή Ιόν συνδεδεμένη με την Κορυφή που μόλις δημιουργήσαμε. Τέλος, αν η κορυφή είναι αναγνωρισμένη, δημιουργούμε εγγραφές Μεταβολίτη και Παραγώγου, αν αυτές δεν υπάρχουν ήδη με τα ίδια χαρακτηριστικά, και τις συνδέουμε μεταξύ τους και με την Κορυφή.

4.6.3 Στοιχεία ανά μελέτη για την περίπτωση της Arabidopsis thaliana

Οι Κορυφές που προσθέσαμε στην Βιβλιοθήκη, όπως έχουμε αναφέρει, έχουν εμφανιστεί σε έναν αριθμό από μελέτες, με στοιχεία όπως ο Χρόνος Παραμονής και το Αναγνωριστικό Ιόν δυνητικά να διαφέρουν. Από τις Μελέτες αυτές, η πιο διαθέσιμή από πλευράς δεδομένων είναι αυτή η οποία ασχολήθηκε με δείγματα ενός φυτικού οργανισμού μοντέλου. Η έρευνα αυτή ήταν η πρώτη και αυτή η οποία ίδρυσε την Βιβλιοθήκη Κορυφών, και καλύπτει τις 240 από τις 430 κορυφές.

Για κάθε μία από αυτές τις Κορυφές, θα δημιουργήθηκε μία εγγραφή η οποία δίνει σαν συνώνυμο το αναγνωριστικό που είχε χρησιμοποιηθεί, καθώς και ο Χρόνος Παραμονής και Το Ιόν.

4.6.4 Μορφή αποθήκευσης των δεδομένων

Για την μεταφόρτωση και εξωτερική αποθήκευση των δεδομένων, που μπορεί να χρησιμοποιηθεί ακόμα και για τα αντίγραφα ασφαλείας της εφαρμογής σε λειτουργία, το Django μας δίνει τις εντολές διαχείρισης **loaddata** και **dumpdata**. Υποστηρίζει αρχεία μορφής XML, YAML και JSON, από τα οποία εμείς έχουμε επιλέξει να χρησιμοποιήσουμε μία μορφή όπου κάθε εγγραφή αντιστοιχεί σε μία γραμμή του αρχείου με JSON μορφή.

Σε πολλές διαδικτυακές εφαρμογές, ειδικά σε αυτές που δέχονται περιεχόμενο από τους χρήστες τους, ο όγκος των δεδομένων της βάσης είναι τόσο μεγάλος, που η συγκεκριμένη τεχνική χρησιμοποιείται μόνο για να εισάγει αρχικά και μόνο δεδομένα στην εφαρμογή. Στην δική μας περίπτωση όπου οι εγγραφές Κορυφών θα περιοριστούν το πολύ στις χιλιάδες, μέσω αυτής της τεχνικής μπορούμε να αποθηκεύσουμε όλη την Βιβλιοθήκη κορυφών σαν ένα αρχείο το οποίο θα διανέμουμε και θα εξελίσσουμε ακόμα μαζί με τον πηγαίο κώδικά μας.

Ταυτόχρονα με την βασική μορφή των δεδομένων, υλοποιήθηκε μία ακόμα εντολή διαχείρισης, **csvexport**, μέσω της οποίας είναι δυνατόν το σύνολο των κορυφών της βάσης να εξαχθεί σε μορφή απλού αποκανονικοποιημένου πίνακα. Αυτό έγινε για την εύκολη εποπτεία των δεδομένων της βάσης, και την αντιπαράθεσή τους με την αρχική βιβλιοθήκη.

Κεφάλαιο 5. Συμπεράσματα και Προτάσεις Περαιτέρω Ανάπτυξης

Έχοντας ολοκληρώσει αυτήν την εργασία θα ήθελα να σταθώ στο εξής. Ο σχεδιασμός και υλοποίησης της Προτυποποιημένης Βιβλιοθήκης Κορυφών παντρεύει το επιστημονικό πεδίο της Βιολογίας με αυτό της Πληροφορικής, και γι' αυτόν τον λόγο παρουσιάζει πάρα πολύ δύσκολες, αλλά ταυτόχρονα και όμορφες προκλήσεις. Ο ρόλος ενός μηχανικού βιοπληροφορικής, ο οποίος, μιλώντας και τις δύο επιστημονικές γλώσσες, να είναι ικανός να κατανοεί τα βιολογικά προβλήματα, να τα μεταφράζει σε έννοιες και τεχνικές της πληροφορικής, να μπορεί να συμβουλεύει και να προσφέρει λύσεις βάση της εμπειρίας του, είναι ένας ρόλος μεγάλης αξίας στην σύγχρονη έρευνα.

Στο πέρας της εργασίας, έχουμε καταφέρει τον αρχικό μας στόχο, να φέρουμε την βιβλιοθήκη κορυφών σε μία εύχρηστη μορφή, βασισμένη σε μία βάση δεδομένων και να μεταφέρουμε μεγάλο μέρος της υπάρχουσας πληροφορίας σε αυτήν. Περισσότερο απ' όλα όμως, θέσαμε την βάση για την περαιτέρω ανάπτυξη της. Μέσω της υλοποίησης αντικειμενοστραφών μοντέλων, ανοίγονται μπροστά μας δυνατότητες, που σκοπεύουμε να πραγματοποιήσουμε στο άμεσο μέλλον.

Σκοπός μας είναι στο μέλλον να υλοποιηθεί μία διαδικτυακή εφαρμογή για την προβολή και αναζήτηση προτύπων κορυφών μεταβολιτών η οποία να είναι διαθέσιμη διαδικτυακά, και η οποία θα αποτελέσει κόμβο εφαρμογής στην υποδομή διασύνδεσης βιολογικών δεδομένων ELIXIR. Η εφαρμογή θα μπορεί να εξυπηρετήσει τόσο ερευνητές μέσω της ιστοσελίδας της, αλλά να είναι διαθέσιμη προγραμματιστικά μέσω μίας REST-API υπηρεσίας. Στην δεύτερη μορφή της, θα μπορεί να χρησιμοποιηθεί από λογισμικά αυτοματοποιημένης μεταβολομικής ανάλυσης, όπως το M-IOLITE. Τέλος, η βιβλιοθήκη κορυφών και η υλοποίηση της, θέλουμε να αποτελέσει το πρότυπο για την ψηφιοποίηση των υπόλοιπων βιβλιοθηκών του εργαστηρίου.

Βιβλιογραφία

Barsalou, T., & Wiederhold, G. (1990). Complex objects for relational databases. *Comput. Aided Des.*

ChEBI. (2021). Ανάκτηση από https://www.ebi.ac.uk/chebi/

Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks". *Communications of the ACM, 13*(6).

Codd, E. F. (1974). Recent Investigations in Relational Data Base Systems.

ELIXIR-Greece. (2021). Ανάκτηση από https://www.elixir-greece.org/

Eriksson, L., Antti, H., & Gottfries, J. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal Bioanal Chem 380*, 419–429.

Ghimire, D. (2020). *Comparative study on Python web frameworks: Flask and Django.*

GOLM database. (2021). Ανάκτηση από http://gmd.mpimp-golm.mpg.de/

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in biotechnology, 22.5*, 245-252.

Hamany Djande, C. Y., Pretorius, C., Tugizimana, F., Piater, L. A., & Dubery, I. A. (2020). Metabolomics: A tool for cultivar phenotyping and investigation of grain crops. *Agronomy*, 831.

Human Metabolome. (2021). Ανάκτηση από http://www.hmdb.ca/

Kanani, H., & Klapa, M. (2007). Data Correction Strategy for Metabolomics Analysis Using Gas Chromatography. *Mass Spectrometry*, 39–51.

Kanani, H., Chrysanthopoulos, P. K., & Klapa, M. I. (2008). Standardizing GC–MS metabolomics. *Journal of Chromatography*, 191-201.

KEGG Atlas. (2021). Ανάκτηση από https://www.kegg.jp/kegg/atlas/

Kell, D. (2006). Theodor Bücher Lecture. Metabolomics, modelling and machine learning in systems biology - towards an understanding of the languages of cells. *30th FEBS Congress and the 9th IUBMB conference.* Budapest.

Kitano, H. (2002). Systems biology: a brief overview. *Science, 295,* 1662-1664.

Klapa, M. I., & Quackenbush, J. (2003). The quest for the mechanisms of life. *Biotechnology and bioengineering, 84.7,* 739-742.

Maga-Nteve, C., & Klapa, M. I. (2016). Streamlining GC-MS metabolomic analysis using the M-IOLITE software suite. *IFAC-PapersOnLine,* 286-288.

MetaboLights. (2021). Ανάκτηση από https://www.ebi.ac.uk/metabolights/

NIST GC-MS peak library. (2021). Ανάκτηση από http://www.sisweb.com/software/ms/nist.htm

Papadimitropoulos ME.P., V. C.-N. (2018). Untargeted GC-MS Metabolomics. *Metabolic Profiling,* 133-147.

Vidal, M. (2009). A unifying view of 21st century systems biology. *FEBS Letters,* 3891-3894.

Vidal, M., Cusick, M., & Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell, 144,* 986-998.

Μάγγα-Ντεβέ, Χ. (2017). *Ανάπτυξη και διαχείριση υπολογιστικών εργαλείων για την επεξεργασία και συνδυαστική ανάλυση μεταβολομικών προτύπων.* Διδακτορική Διατριβή.

Τοουλάκου, Γ. (2013). Η επίδραση της πενίας άνθρακα στους κρυστάλλους οξαλικού ασβεστίου και στο μεταβολικό πρότυπο του Amaranthus sp. *Διδακτορική Διατριβή*.

Παράρτημα Α - Εγκατάσταση και Εκτέλεση Πηγαίου Κώδικα

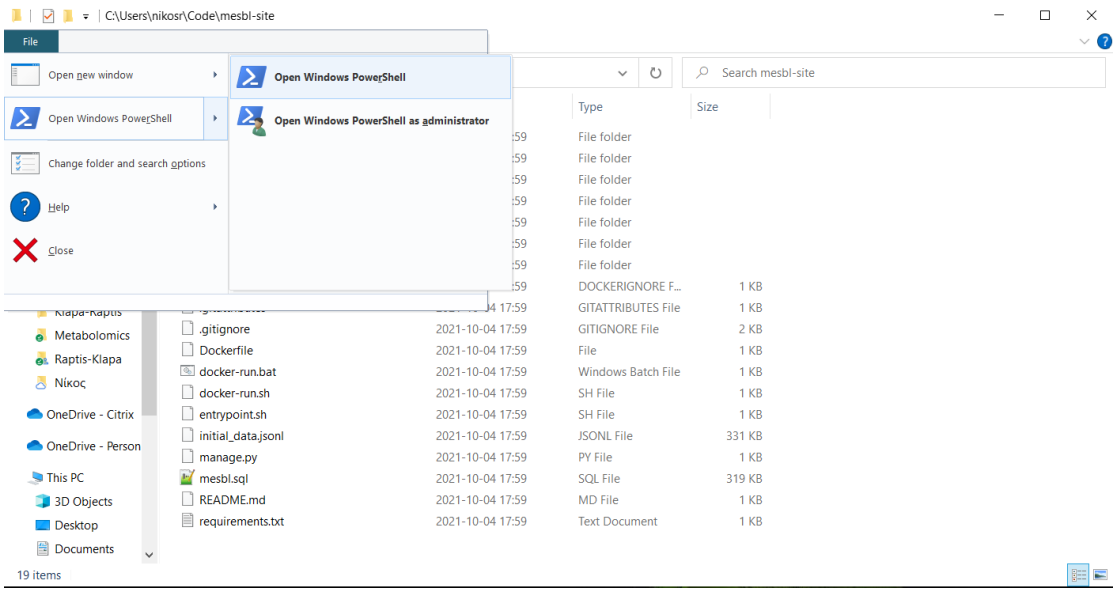
Ο πηγαίος κώδικας που απαρτίζει την βάση, τα δεδομένα της καθώς και το προγραμματιστικό μοντέλο και την εφαρμογή διαχείρισης βρίσκονται στον οπτικό δίσκο που συνοδεύει την εργασία. Η τελευταία έκδοση του είναι επίσης διαθέσιμη μέσω του διαδικτυακού αποθετηρίου GitHub, μετά από συνεννόηση για να δοθεί πρόσβαση.

Η εκτέλεση του κώδικα είναι σχετικά απλή, και αυτό είναι κάτι που οφείλεται στην χρήση της πλατφόρμας Docker. Η μόνη απαίτηση για το σύστημα είναι να είναι εγκατεστημένη η αντίστοιχη εφαρμογή Docker for Desktop, η τελευταία έκδοση της οποίας μπορεί να βρεθεί μαζί με οδηγίες εγκατάστασης στον ιστότοπο

https://docs.docker.com/desktop/windows/install/

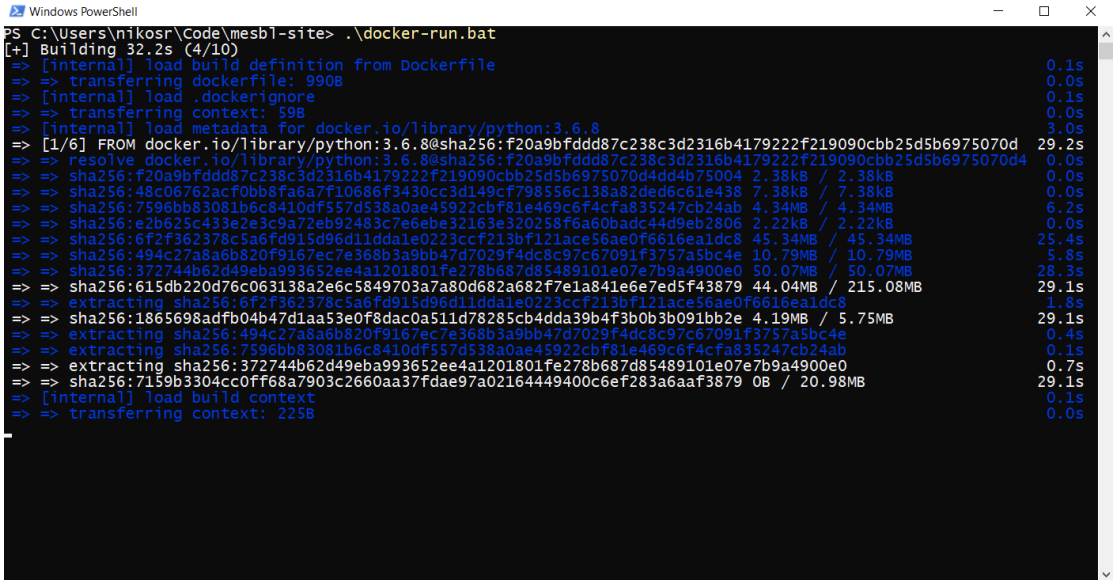
Με εγκατεστημένο και έχοντας ξεκινήσει το Docker, από τον κατάλογο που περιέχει τον πηγαίο κώδικα ανοίγουμε ένα παράθυρο γραμμής εντολών cmd ή PowerShell και πληκτρολογούμε την εντολή

```
.\docker-run.bat
```



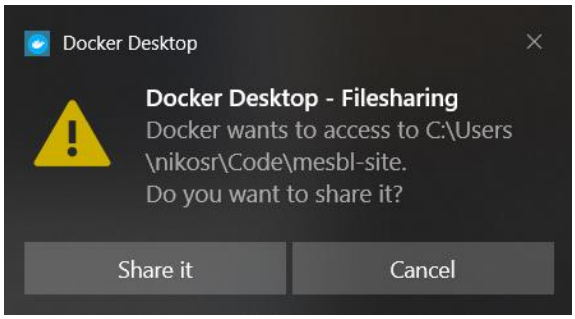
Εικόνα 15 – Άνοιγμα γραμμής εντολών PowerShell

Το αρχείο δέσμης αναλαμβάνει να χτίσει την εικόνα Docker (image) και να την εκτελέσει ως container.



Εικόνα 16 – Εκτέλεση αρχείου δέσμης και δημιουργία εικόνας Docker

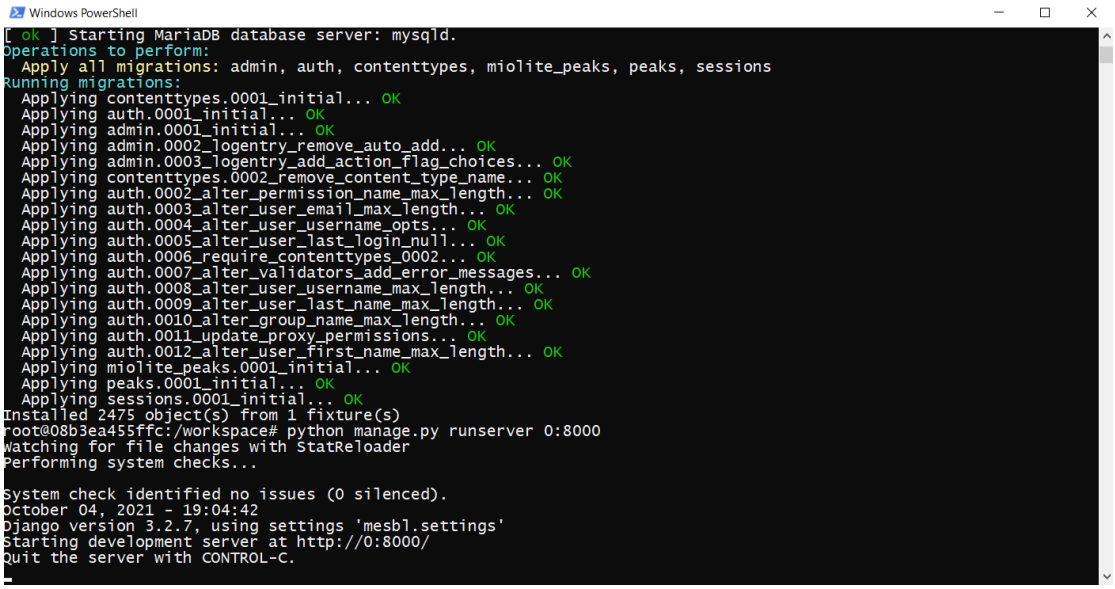
Κατά την εκτέλεση μπορεί να λάβετε προτροπή για διαμοιρασμό του πηγαίου φακέλου μέσω Docker, η οποία θα πρέπει να επιβεβαιωθεί.



Εικόνα 17 – Επιβεβαίωση διαμοιρασμού αρχείων

Κατά την εκτέλεση της εικόνας Docker, στο παρασκήνιο

- Εγκαθίσταται και ρυθμίζεται το λογισμικό βάσης MariaDB
- Δημιουργείται μία κενή βάση και χρήστης για την εφαρμογή
- Εγκαθίστανται το Django και τα υπόλοιπα αναγκαία Python πακέτα
- Δημιουργούνται μέσω της εφαρμογής οι απαραίτητοι πίνακες στη βάση
- Τα δεδομένα της βιβλιοθήκης εισάγονται στη βάση

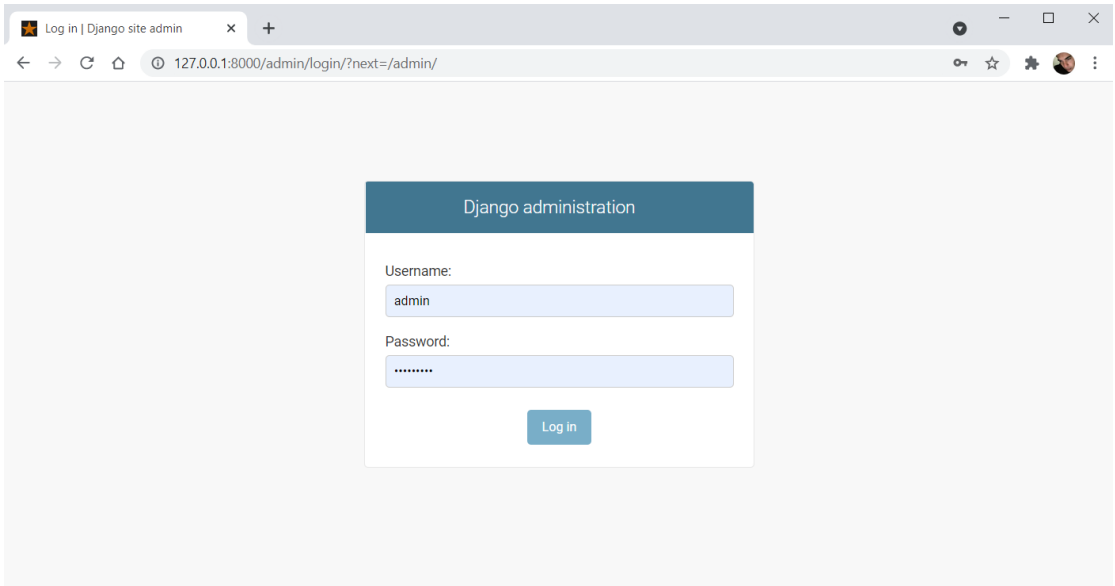


Εικόνα 18 – Αυτόματες ενέργειες και εκκίνηση διακομιστή

Μετά το τέλος τη εκτέλεσης, βρισκόμαστε στην γραμμή εντολών του container, όπου είναι δυνατή μία σειρά από εντολές διαχείρισης. Στην περίπτωσή μας, επιθυμούμε να ξεκινήσουμε το διακομιστή, το οποίο γίνεται με την εντολή

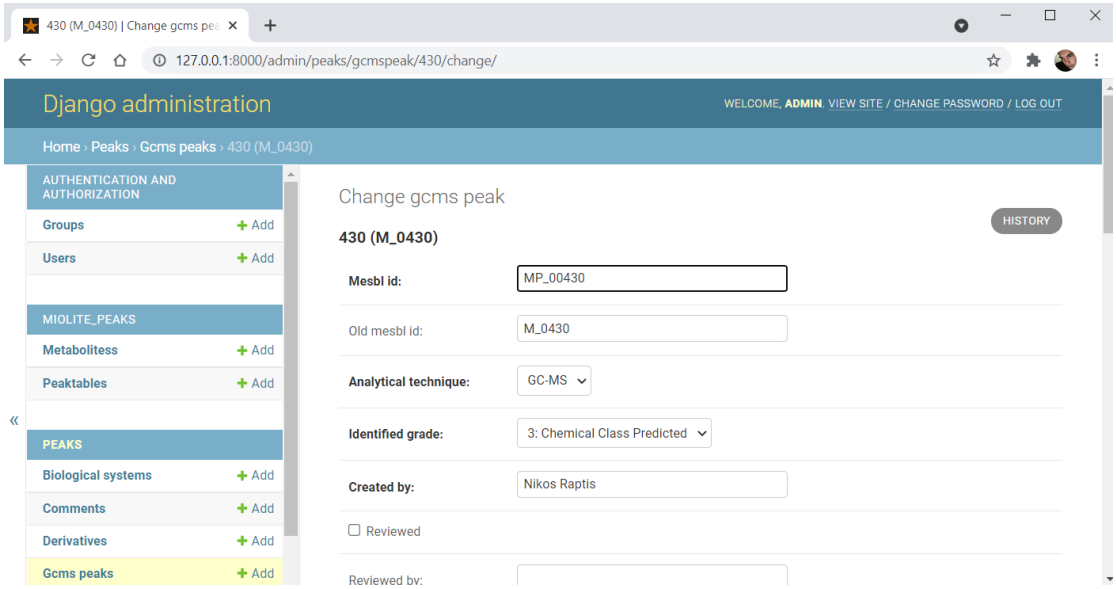
```
/workspace# python manage.py runserver 0:8000
```

Έπειτα, μπορούμε να περιηγηθούμε μέσω του Φυλλομετρητή (web browser) μας στην διεύθυνση http://127.0.0.1:8000/admin και να συνδεθούμε με κωδικούς admin / adminpass οι οποίοι ισχύουν στην περίπτωση μας όπου τρέχουμε ως τοπικό development environment.



Εικόνα 19 – Εισαγωγή συνθηματικού

Στην εφαρμογή μπορούμε να περιηγηθούμε σε όλες τις Οντότητες της βιβλιοθήκης και τις εγγραφές τους, καθώς και να τις επεξεργαστούμε και να προσθέσουμε καινούργιες.



Εικόνα 20 – Περιήγηση στην εφαρμογή διαχείρισης

Τα κύρια αρχεία ενδιαφέροντος στον πηγαίο κώδικα είναι τα εξής:

`mesbl\`	Φάκελος στοιχείων εφαρμογής
`peaks\`	Φάκελος εφαρμογής Βιβλιοθήκης Κορυφών
`peaks\models.py`	Υλοποίηση Αντικειμενοστραφούς Σχεσιακού Μοντέλου
`peaks\admin.py`	Υλοποίηση Εφαρμογής Διαχείρισης
`excelmigrate\`	Υλοποίηση Εντολών Διαχείρισης
`initial_data.jsonl`	Δεδομένα Βιβλιοθήκης Κορυφών
`Dockerfile`	Αρχείο Περιγραφής Μηχανής Docker
`entrypoint.sh`	Αρχείο Δέσμης Αρχικών Ενεργειών
`requirements.pip`	Βιβλιοθήκες Python