

Υπολογιστική ανάλυση 1D και 3D δεδομένων για την επαναστόχευση φαρμάκων στον σακχαρώδη διαβήτη

Εισαγωγή

Την τελευταία δεκαετία η επαναστόχευση φαρμάκων αποτελεί ένα από τα πιο ενεργά πεδία της φαρμακολογίας και εφαρμόζεται από ένα μεγάλο κομμάτι της βιομηχανίας φαρμάκων. Όλο και περισσότερο διερευνάται η αποτελεσματικότητα εγκεκριμένων ή και πειραματικών φαρμάκων σε νέους στόχους, συνεπώς και ασθένειες για τις οποίες δεν είχαν σχεδιαστεί αρχικά. Αυτό επιτυγχάνεται καθώς τα φάρμακα διαταράσσουν μια πληθώρα βιολογικών διεργασιών που είναι κοινές σε πολλές διαφορετικές ασθένειες. Η μέθοδος της επαναστόχευσης είναι λιγότερο δαπανηρή και χρονοβόρα σε σχέση με την παραδοσιακή μέθοδο ανάπτυξης φαρμακευτικών ουσιών. Αυτό οφείλεται στο ότι τα φάρμακα αυτά έχουν ήδη περάσει όλα τα στάδια της ανακάλυψης νέων φαρμάκων συμπεριλαμβανομένων των προκλινικών δοκιμών και των κλινικών φάσεων I και II. Συνεπώς μειώνεται δραματικά ο χρόνος και το κόστος παραγωγής ενώ παράλληλα η αποτελεσματικότητα και η ασφάλειά τους στον άνθρωπο είναι επαρκώς τεκμηριωμένη (Pushrakom et al. 2018).

Επιπλέον, ο μεγάλος όγκος γονιδιωματικών, μεταγραφωματικών, πρωτεωμικών και μεταβολομικών δεδομένων που παράγονται οδήγησε στην ραγδαία ανάπτυξη υπολογιστικών εργαλείων τα οποία μπορούν να επεξεργαστούν τόσο μεγάλο όγκο πληροφορίας. Η συλλογή και ανάλυση της πληροφορίας αυτής δίνει την δυνατότητα έμμεσης σύνδεσης των φαρμάκων με τις ασθένειες. Οι υπολογιστικές τεχνικές έχουν συμβάλει στην κατανόηση των αλληλεπιδράσεων μεταξύ φαρμάκων-στόχων και μεταξύ στόχων-ασθενειών, ώστε να επιτευχθεί ο επαναπροσδιορισμός των φαρμάκων (March-Vila et al. 2017). Ωστόσο, πρόκληση στην επαναστόχευση φαρμάκων αποτελεί η εύρεση υπολογιστικών λύσεων για την συνένωση, ανάλυση και ερμηνεία ετερογενών δεδομένων.

Στόχος της διατριβής αυτής είναι η ανάπτυξη μεθοδολογίας ενοποίησης και ανάλυσης ετερογενών δεδομένων μεγάλου όγκου η οποία θα συμβάλει στην ακριβέστερη επαναστόχευση φαρμάκων. Στη μελέτη αυτή θα εφαρμοστεί για πρώτη φορά η συνδυαστική ανάλυση ομικών δεδομένων, δεδομένων φαρμακοεπαγρύπνησης, δεδομένων ευρεσιτεχνιών βιοδραστικών μορίων, ηθικών/νομικών ζητημάτων, κλινικών δοκιμών και ηλεκτρονικών αρχείων βιοτραπεζών.

Η λειτουργία και η ακρίβεια της μεθόδου επαναπροσδιορισμού που θα αναπτυχθεί θα αξιολογηθεί στην επαναστόχευση φαρμάκων στον σακχαρώδη διαβήτη. Ο διαβήτης αποτελεί μια από τις ταχύτερα αναπτυσσόμενες ασθένειες παγκοσμίως, με τα πιο πρόσφατα στοιχεία της Διεθνούς Ομοσπονδίας Διαβήτη να δείχνουν ότι περίπου 380 εκατομμύρια άνθρωποι προσβάλλονται αυτήν τη στιγμή από αυτή την πάθηση. Αυτός ο αριθμός εκτιμάται ότι θα ανέλθει σε 592 εκατομμύρια μέχρι το 2035, κυρίως λόγω της δραματικής αύξησης του διαβήτη τύπου 2. Παράλληλα το κόστος των φαρμάκων που βαρύνει τους πάσχοντες είχε υπολογιστεί στα ~50

δισεκατομμύρια ετησίως στις ΗΠΑ το διάστημα 2015-2017. Επιπρόσθετα, ο διαβήτης σχετίζεται με ισχυρά αυξημένο κίνδυνο εμφάνισης αθηροσκλήρωσης, νεφρικής ανεπάρκειας και θανάτου από έμφραγμα του μυοκαρδίου ή εγκεφαλικό επεισόδιο. Συνεπώς η θεραπεία του διαβήτη απαιτεί μια πιο αποτελεσματική προσέγγιση από τη μείωση των επιπέδων γλυκόζης στο αίμα (Danaei et al. 2011). Οι συμβατικές θεραπευτικές επιλογές για τη διέγερση της έκκρισης ινσουλίνης είναι οι σουλφονυλουρίες και οι μεγλιτινίδες, οι οποίες δεσμεύουν του ATP-ευαίσθητων διαύλων καλίου στα β παγκρεατικά κύτταρα. Τα φάρμακα αυτά φαίνεται να προκαλούν παρενέργειες όπως η υπογλυκαιμία και η αύξηση βάρους, μαζί με την πιθανότητα υπερφόρτωσης των β-κυττάρων, επιδεινώνοντας τελικά την αποτυχία των β-κυττάρων μακροπρόθεσμα. Η ανάπτυξη φαρμάκων που δρουν είτε ως αναστολείς της διπεπτιδυλικής πεπτιδάσης-4 (DPP-4) και προκαλούν την αύξηση των επιπέδων του GLP-1 είτε ως ενεργοποιητές των υποδοχέων GLP-1 για την αύξηση της έκκρισης ινσουλίνης είναι πολλά υποσχόμενη καθώς φαίνεται να διατηρούν τις ιδιότητες των β-κυττάρων και να μειώνουν την αύξηση βάρους και την αρτηριακή πίεση. Ωστόσο οι δύο διαφορετικοί τύποι βιοδραστικών μορίων, έχουν συσχετιστεί με διάφορες παρενέργειες. Πιο συγκεκριμένα, έχουν καταγραφεί συμβάντα καρδιακής ανεπάρκειας για τους DPP-4 αναστολείς, με αποτέλεσμα να προτείνεται η χορήγηση αγωνιστών του υποδοχέα του πεπτιδίου-1 που προσομοιάζει στο γλουκαγόνο (GLP1-R) στους υπέρβαρους ασθενείς με διαβήτη ή τους ασθενείς με αθηροσκληρωτική καρδιαγγειακή νόσο (Stein, Lamos, and Davis 2013). Συνεπώς ως προς την επαναστόχευση φαρμάκων θα πρέπει να ληφθούν υπόψη γονιδιωματικές παραλλαγές στα γονίδια DPP4 (rs2909451, rs759717) και GLP1R (rs6923761), καθώς έχουν καίριο ρόλο στον σακχαρώδη διαβήτη.

Υλικά & Μέθοδοι

Υλικά

Στην παρούσα φάση στα πλαίσια υλοποίησης της προτεινόμενης μεθοδολογίας συλλέχθηκαν δεδομένα από δημόσια διαθέσιμα αποθετήρια. Η συλλογή πληροφοριών που αφορούν τη φαρμακολογική χρήση των δραστικών βιομορίων έγινε από την βάση DrugBank (Wishart et al. 2018). Η βάση αυτή διαθέτει για κάθε χημική ένωση πληροφορίες που αφορούν τη δομή, τη φαρμακολογική δράση, τις πρωτεΐνες στόχος, την εμπορική εκμετάλλευση, κτλ. Ένα σημαντικό στοιχείο που προσφέρει το συγκεκριμένο αποθετήριο είναι η αντιστοίχιση των φαρμακευτικών στόχων (πρωτεΐνες) της εκάστοτε χημικής ένωσης με την βάση δεδομένων Uniport (The UniProt Consortium et al. 2020). Η βάση UniProt αποτέλεσε την κύρια πηγή δεδομένων για τις πρωτεΐνες καθώς παρέχει πληροφορίες όπως η βιολογική τους δράση και η δομή τους. Επιπλέον χρησιμοποιήθηκε η πλατφόρμα SureChEMBL (Papadatos et al. 2016) για την εξαγωγή δεδομένων που αφορούν τις υπάρχουσες ευρεσιτεχνίες χημικών μορίων. Επιλέχθηκε η συγκεκριμένη πλατφόρμα καθώς τα αποτελέσματα των επερωτήσεων που επιστρέφει βασίζονται σε αναζητήσεις όχι μόνο στα κείμενα των διπλωμάτων ευρεσιτεχνίας αλλά και στις εικόνες που περιέχουν. Ακόμη προστέθηκαν πληροφορίες για τη φαρμακολογική δράση των πρωτεϊνών-μεταφορέων οι οποίες προήλθαν από την βάση USCF-FDA (Morrissey et al. 2012). Η UCSF-FDA transportal αποτελεί ένα δημόσια διαθέσιμο αποθετήριο που παρέχει πληροφορίες για τις πρωτεΐνες μεταφορείς (transporters) οι οποίες βρίσκονται στις μεμβράνες και

ελέγχουν μια σειρά διεργασιών του κυττάρου. Οι πρωτεΐνες αυτές είναι σημαντικές γιατί επηρεάζουν την φαρμακοκινητική μιας χημικής ένωσης και παράλληλα αποτελούν και φαρμακευτικούς στόχους. Επιπρόσθετα συλλέχθηκαν δεδομένα φαρμακοεπαγρύπνησης τα οποία εξάγαμε από το αποθετήριο PharmGKB (Whirl-Carrillo et al. 2021). Στο αποθετήριο αυτό βρίσκονται συσχετίσεις που συνδέουν τη φαρμακολογική δράση των χημικών ουσιών με γνωστές μεταλλάξεις. Οι μεταλλάξεις αυτές επηρεάζουν τις φαρμακολογικές ιδιότητες των χημικών μορίων και με βάση των FDA κάποιες από αυτές τις φαρμακευτικές ουσίες συνταγογραφούνται μετά από έλεγχο του γονότυπου. Από την PharmGKB επιλέχθηκαν μόνο δεδομένα τα οποία αφορούν χημικές ουσίες για τις οποίες το επίπεδο των στοιχείων που υποστηρίζουν την ύπαρξη της συσχέτισης με κάποια μετάλλαξη να είναι μεγαλύτερο ή ίσο του 3. Ένα ακόμη αποθετήριο που χρησιμοποιήθηκε για την συλλογή δεδομένων είναι το Omnipath (Türei et al. 2021) στο οποίο έχουν συγκεντρωθεί και καταχωρηθεί διάχυτα δεδομένα τα οποία αφορούν την υπάρχουσα γνώση στην μοριακή βιολογία. Πιο αναλυτικά, στο αποθετήριο αυτό έχουν συγκεντρωθεί δεδομένα από πάνω 100 άλλες πηγές. Έχει δομηθεί έτσι ώστε να αποτελείται από 5 βασικά αποθετήρια τα οποία περιέχουν δεδομένα που αφορούν τις αλληλεπιδράσεις σηματοδοτικών δικτύων, τις σχέσεις ενζύμου υποστρώματος, τα πρωτεϊνικά σύμπλοκα, των σχολιασμό πρωτεϊνών και την διακυτταρική επικοινωνία. Η συγκεκριμένη βάση θα εμπλουτίσει τις πληροφορίες για τις πρωτεΐνες που έχουν συλλεχθεί από την Uniprot. Τέλος είναι σημαντικό να γνωρίζουμε πως θα αντιδράσει ο οργανισμός κατά την λήψη ενός φαρμάκου. Η διαδικασία αυτή αφορά τον μεταβολισμό των ξενοβιοτικών η οποία περιγράφεται από την συντομογραφία ADME η οποία προκύπτει από τα εξής: “απορρόφηση (absorption), κατανομή (distribution), μεταβολισμό (metabolism) και απέκκριση (excretion)”. Οι προαναφερθέντες παράμετροι εξαρτώνται από την δομή του χημικού μορίου η οποία θα καθορίζει την φαρμακοκινητική και φαρμακολογική του δράση. Για την συλλογή των προαναφερθέντων πληροφοριών χρησιμοποιήθηκε η πλατφόρμα AdmetSAR (Cheng et al. 2012) η οποία ενσωματώνει μια ποικιλία QSAR μοντέλων τα οποία παράγουν προβλέψεις για τις ADME ιδιότητες των χημικών μορίων. Συμπληρωματικά των ADME ιδιοτήτων, συλλέχθηκαν δεδομένα που αφορούν τον μεταβολισμό των χημικών μορίων. Οι πληροφορίες αυτές ελήφθησαν από την πλατφόρμα Biotransformer (Djoumbou-Feunang et al. 2019) η οποία προβλέπει τον μεταβολισμό χημικών μορίων στα θηλαστικά, στο μικροβίωμα του εντέρου καθώς στο εδαφικό/υδρόβιο μικροβίωμα. Επιπλέον δίνει την δυνατότητα αναγνώρισης των μεταβολικών βάσης της αρχικής πρόβλεψης. Οι προβλέψεις της πλατφόρμας παράγονται από τον συνδυασμό υπάρχουσας γνώσης και μοντέλων μηχανικής μάθησης.

Τα δεδομένα που συλλέγουμε χαρακτηρίζονται από ιδιαίτερα μεγάλο όγκο, πολύπλοκες σχέσεις αλληλεπίδρασης μεταξύ τους και εξαιρετικά μεγάλη ετερογένεια καθώς προέρχονται από διαφορετικά αποθετήρια. Συνεπώς αντιλαμβανόμαστε ότι δεν μπορούν να χρησιμοποιηθούν οι κλασικές σχεσιακές βάσεις δεδομένων για την αποθήκευση των πληροφοριών που συλλέγουμε. Οι σχεσιακές βάσεις δεδομένων αποτελούν το πιο ευρέως χρησιμοποιούμενο μοντέλο αποθήκευσης και διαχείρισης πληροφοριών. Ωστόσο έχουν τα εξής δυο μειονεκτήματα, δεν είναι εύκολα επεκτάσιμες όταν καλούνται να διαχειριστούν μεγάλο όγκο δεδομένων και δεν μπορούν να διαχειριστούν μη δομημένα δεδομένα. Έτσι επιλέχθηκε η χρήση βάσης δεδομένων γράφου η οποία αποτελεί το βέλτιστο τρόπο διαχείρισης και επερώτησης συνδεδεμένων δεδομένων. Ως συνδεδεμένα δεδομένα θεωρούμε αυτά τα οποία περιγράφουν σημασιολογικές σχέσεις μεταξύ των οντοτήτων που διασυνδέουν καθώς και τη σημαντικότητα (βάρος) που μπορεί να έχουν οι σχέσεις αυτές. Ως γνωστόν οι λειτουργίες αυτές δεν παρέχονται στο σχεσιακό μοντέλο. Οι βάσεις

αυτού του τύπου στηρίζονται στη μαθηματική έννοια του γράφου η οποία αναπαριστά τα δεδομένα με την μορφή κορυφών και ακμών. Δηλαδή οι οντότητες που εισάγουμε παρίστανται ως κόμβοι αντί για πίνακες πίνακες μια σχεσιακής βάσης. Ενώ οι ακμές υποδεικνύουν την σχέση μεταξύ των οντοτήτων αντί για σύνδεση μέσω ενός εξωτερικού κλειδιού δύο πινάκων σχεσιακού μοντέλου. Το κύριο πλεονέκτημα των βάσεων δεδομένων γράφων είναι ότι η μηχανή επερώτησης δεν αναζητά σχέσεις μεταξύ οντοτήτων αλλά ολόκληρους γράφους. Ενώ η διασύνδεση παραπάνω από 2 οντοτήτων απαιτεί αρκετά εξειδικευμένα και πολύπλοκα SQL queries με τις γλώσσες επερωτήσεων σε γράφους αυτό γίνεται με απλό. Η πλατφόρμα διαχείρισης της βάσης δεδομένων γράφου που επιλέχθηκε είναι η Neo4J καθώς αποτελεί ιδανική λύση όταν έχουμε μεγάλο όγκο από πολύπλοκη πληροφορία και όταν οι επερωτήσεις που θέλουμε να κάνουμε είναι επίσης πολύπλοκες και μη καθορισμένες (“Website,” n.d.).

Μέθοδοι

Η μέθοδος ανάκτησης, εξόρυξης και επεξεργασίας δεδομένων διαφοροποιείται μεταξύ των βάσεων. Για κάθε μια από αυτές ανάλογα με το πως παρέχονται τα δεδομένα επιλέχθηκε ο βέλτιστος τρόπος συλλογής. Επίσης όλα τα ετερογενή δεδομένα που συλλέχθηκαν επεξεργάστηκαν με σκοπό λάβουν την κατάλληλη μορφή ώστε να εισαχθούν στον γράφο.

Για την λήψη των δεδομένων από την DrugBank δημιουργήθηκε λογαριασμός χρήστη στην βάση και στην συνέχεια κατόπιν συνεννόησης με τους διαχειριστές της παραχωρήθηκε άδεια λήψης του συνόλου των δεδομένων της βάσης. Από το σύνολο των δεδομένων λήφθηκαν αυτά που αφορούν τις χημικές ενώσεις της πλατφόρμας και αυτά που αφορούν τους φαρμακευτικούς στόχους. Η λήψη των δεδομένων έγινε σε μορφή csv τα οποία στην συνέχεια εισήχθησαν στο προγραμματιστικό περιβάλλον Rstudio και επεξεργάστηκαν με την γλώσσα προγραμματισμού R ώστε να γίνει η αντιστοίχιση μεταξύ χημικών μορίων και στόχων. Έπειτα γνωρίζοντας πλέον τις πρωτεΐνες ενδιαφέροντος (φαρμακευτικοί στόχοι από DrugBank) έγινε λήψη των δεδομένων μέσω της Uniprot. Η Uniprot δίνει την δυνατότητα λήψης των δεδομένων μέσω σύνδεσης FTP και παρέχει τα αρχεία σε μορφή xml ή txt. Αντίστοιχα έγινε και η λήψη των δεδομένων από την βάση PharmGKB όπου διατίθενται τα δεδομένα για κατέβασμα σε μορφή tsv. Μέσα από το αρχείο αυτό εντοπίσαμε και συλλέξαμε τις πληροφορίες για τα χημικά μόρια ενδιαφέροντος.

Όσον αφορά το αποθετήριο ευρεσιτεχνιών SureChEMBL, διαθέτει ιστοσελίδα που προσφέρει διαδικτυακή διεπαφή στους χρήστες και διαθέτει 3 μορφές επερωτήσεων:

1. Με βάση λέξεις κλειδιά, οι οποίες μπορούν να περιέχουν λογικούς τελεστές καθώς επίσης και παρενθέσεις.
2. Με βάση χημικές δομές ή χημικές υποομάδες, οι οποίες μπορούν είτε να σχεδιαστούν, είτε να εισαχθούν με μορφή SMILES/MOL/SMARTS.
3. Με συνδυασμό των λέξεων και χημικών δομών.

Για να επωφεληθούμε από τις λειτουργίες που παρέχονται στην ιστοσελίδα και να μειωθεί το υπολογιστικό κόστος επιλέξαμε η εξόρυξη των δεδομένων να γίνει με την τεχνική “συγκομιδή δεδομένων” (data scraping/data harvesting). Η “συγκομιδή δεδομένων” είναι μια αυτοματοποιημένη τεχνική συλλογής δεδομένων από το διαδίκτυο. Η τεχνική αυτή υλοποιείται μέσω ειδικού λογισμικού/κώδικα που επι της ουσίας αντιγράφει δεδομένα από ιστοσελίδες και τα συγκεντρώνει για μεταγενέστερη ανάλυση. Βασική αρχή λειτουργίας της τεχνικής αυτής είναι να διαβάσει την γλώσσα HTML που είναι γραμμένη σε μια ιστοσελίδα και συγκεκριμένα στο

Document Object Model της HTML (DeVito, Richards, and Inglesby 2020). Ο κώδικας που υλοποιεί την συγκομιδή των δεδομένων ουσιαστικά αποτελεί ένα διαδικτυακό ρομπότ (web robot/bot) που μιμείται την διαδικασία περιήγησης και αλληλεπίδρασης ενός ανθρώπου με τους διαδικτυακούς διακομιστές. Ο κώδικας υλοποιήθηκε στην γλώσσα προγραμματισμού R, με την βοήθεια των πακέτων rvest και RSelenium, τα οποία περιέχουν χρήσιμες συναρτήσεις για την 'συγκομιδή δεδομένων'. Η λειτουργία που επιτελεί ο κώδικας που αναπτύχθηκε είναι η εξής:

1. Είσοδος στην ιστοσελίδα, η οποία επιτυγχάνεται ανοίγοντας έναν διαδικτυακό περιηγητή και δημιουργεί επικοινωνία μέσω HTTP πρωτοκόλλου.
2. HTML parsing και εξαγωγή περιεχομένου. Μόλις γίνει η ανάκτηση του κειμένου HTML τότε εξάγεται το περιεχόμενο ενδιαφέροντος με βάση το XPath. Το XPath είναι μια γλώσσα έκφρασης η οποία επιτρέπει την επεξεργασία των τιμών ενός καθορισμένου μοντέλου δεδομένων. Το μοντέλο δεδομένων παρέχει την αναπαράσταση των XML εγγράφων σε μορφή δέντρου. Έτσι το αποτέλεσμα μιας έκφρασης XPath είναι ακολουθία επιλεγμένων κόμβων από το κείμενο HTML.
3. Σχηματισμός δεδομένων εξόδου. Μετά την εξαγωγή τα δεδομένα ενδιαφέροντος καταχωρούνται σε πίνακες (data.frames) και στην συνέχεια καταγράφονται σε μια λίστα η οποία εγγράφεται στον σκληρό δίσκο του υπολογιστή.

Η πληροφορία που συλλέγουμε περιέχει τον κωδικό της πατέντας (publication number), την ημερομηνία που δημοσιεύτηκε, τον τίτλο της ευρεσιτεχνίας, τον κωδικό κατηγορίας με βάση την διεθνή ταξινόμηση πατεντών, τον αιτούντα του διπλώματος και τον σύνδεσμο για το δίπλωμα ευρεσιτεχνίας. Πέρα των στοιχείων που προαναφέρθηκαν για κάθε δίπλωμα ευρεσιτεχνίας, ιδιαίτερο ενδιαφέρον έχουν οι αξιώσεις (claims) που κατοχυρώνονται στον αιτούντα. Μέσα από αυτές μπορούμε να εξάγουμε χρήσιμες πληροφορίες, συμπληρωματικές στα χαρακτηριστικά στοιχεία της κάθε πατέντας. Για την συλλογή των αξιώσεων κάθε πατέντας αναπτύχθηκε ένα δεύτερο web robot το οποίο πλοηγείται ξεχωριστά σε κάθε μια από τις πατέντες που έχουμε συλλέξει. Στην συνέχεια εντοπίζει την περιοχή των αξιώσεων, τις εξάγει και τις αποθηκεύει τοπικά. Επιπλέον ενεργοποιεί την λειτουργία αναγνώρισης και εξαγωγής ονομάτων-οντοτήτων από το κείμενο, τις οποίες στην συνέχεια εξάγει από την ιστοσελίδα και τις αποθηκεύει για περαιτέρω επεξεργασία. Οι ονοματικές οντότητες που εντοπίζονται μέσα στο κείμενο αφορούν τις εξής 10 κατηγορίες:

1. Ανατομία
2. Βιολογικές έννοιες
3. Φάρμακα από την ChEMBL
4. Τύπος φαρμάκου
5. Ενδογενείς κύτταρα
6. Γονίδια
7. Γονιδιακή Οντολογία
8. Ένδειξη
9. Κλάση πρωτεΐνης
10. Είδος

Τα δεδομένα της βάσης UCSF-FDA transportal είναι μόνο διαθέσιμο μέσο της ιστοσελίδας, συνεπώς για την συλλογή τους χρησιμοποιήθηκε και εδώ η τεχνική web scraping. Ο όγκος των δεδομένων της βάσης είναι μικρός, συνεπώς συλλέξαμε την πληροφορία που παρέχεται για όλους τους μεταφορείς. Τα δεδομένα πάρθηκαν από την HTML σελίδα του

ιστότοπου της βάσης. Κάθε HTML σελίδα περιέχει τα δεδομένα σε μορφή πίνακα. Κάθε σελίδα περιέχει μέχρι 3 πίνακες που σχετίζονται με τις εξής περιπτώσεις:

1. In vitro υποστρώματα (substrates)
2. In vitro αναστολείς (inhibitors)
3. Κλινικές αλληλεπιδράσεις φαρμάκων (Clinical Drug-Drug Interactions)

Για την εξόρυξη των πληροφοριών αναπτύχθηκε ένα πρόγραμμα το οποίο θα μπορεί να κάνει απλή συγκομιδή (scraping) των δεδομένων με βάση του url κάθε φαρμάκου. Πιο συγκεκριμένα, δημιουργήθηκε ένα txt αρχείο με όλα τα ονόματα των φαρμάκων που περιέχονται στην βάση UCSF όπως αυτά δίνονται στο ευρετήριο της βάσης. Έτσι το πρόγραμμα έπαιρνε το όνομα του φαρμάκου από το txt αρχείο και άνοιγε την αντίστοιχη σελίδα στην βάση της UCSF. Στην συνέχεια διάβαζε το HTML έγγραφο της συγκεκριμένης σελίδας και εντόπιζε ποιος κόμβος του HTML εγγράφου περιέχει πίνακα. Στην συνέχεια ανάλογα με τον αριθμό πινάκων που εντοπιζόταν και τον αριθμό των στηλών που είχε ο κάθε πίνακας γινόταν η εξόρυξη των δεδομένων κειμένου μέσα από αυτούς. Στην συνέχεια το πρόγραμμα έκανε καταγραφή του εξαγόμενου κειμένου σε πίνακα/πίνακες ώστε τα δεδομένα να έχουν την ίδια δομή με αυτή που είχαν αρχικά στην ιστοσελίδα. Οι πίνακες που σχηματίζονταν τοπικά για κάθε πρωτεΐνη-μεταφορέα εισάγονταν σε μια λίστα δεδομένων η οποία γραφόταν στον σκληρό δίσκο. Αυτή η διαδικασία έγινε για όλους τους μεταφορείς της βάσης μέχρι να έχουμε όλη την πληροφορία του αποθετηρίου τοπικά. Τέλος έγινε αντιστοίχιση των φαρμάκων της βάσης με τα χημικά μόρια ενδιαφέροντος που συλλέξαμε από την DrugBank.

Στην περίπτωση του Omnipath διατίθεται πακέτα σε R για άμεση πρόσβαση στα δεδομένα μέσα από προγραμματιστικό περιβάλλον. Η λήψη των δεδομένων έγινε μέσω του πακέτου το οποίο δημιουργεί επικοινωνία HTTP πρωτοκόλλου με το REST API του διακομιστή. Με αυτόν τον τρόπο γίνονται επερωτήσεις μέσω του Rstudio, κατεβαίνουν τα δεδομένα τοπικά και φορτώνονται στο προγραμματιστικό περιβάλλον. Τα δεδομένα που συλλέξαμε αφορούν τις αλληλεπιδράσεις μεταξύ πρωτεϊνών. Η πληροφορία που συλλέγουμε αφορά τις ανθρώπινες πρωτεΐνες που αλληλεπιδρούν μεταξύ τους και εάν η αλληλεπίδραση αυτή έχει ανασταλτική δράση (inhibition) ή διεγερτική δράση (stimulation). Η δράση αυτή των πρωτεϊνών αυτή συνοδεύεται και από τις αντίστοιχες βιβλιογραφικές αναφορές όπου αυτό είναι δυνατό. Στο Omnipath οι πρωτεΐνες έχουν κωδικούς Uniprot και με τους κωδικούς αυτούς έγινε η αντιστοίχιση με τους φαρμακευτικούς στόχους που έχουμε. Η πλατφόρμα Biotransformer διαθέτει αντίστοιχα REST API χωρίς όμως να παρέχει πακέτο που να κάνει άμεση σύνδεση με το Rstudio όπως στην περίπτωση του Omnipath. Έτσι αναπτύχθηκε ένα script σε R το οποίο δημιουργεί την σύνδεση μεταξύ του προγραμματιστικού περιβάλλοντος και του API του διακομιστή μέσω HTTP πρωτοκόλλου. Στην συνέχεια αφού αποκατασταθεί η σύνδεση το script περιέχει ένα σύνολο εντολών ώστε να κάνει τις κατάλληλες επερωτήσεις, να συλλέγει τα αποτελέσματα από τον διακομιστή του Biotransformer και στην συνέχεια να τα αποθηκεύει τοπικά.

Τέλος για την πλατφόρμα AdmetSAR αναπτύχθηκε ένα web scraping πρόγραμμα αντίστοιχο με αυτό που δημιουργήθηκε για την SureChEMBL που περιγράφηκε παραπάνω. Στην συγκεκριμένη περίπτωση το bot δίνει για επερώτηση στην πλατφόρμα την SMILES μορφή της εκάστοτε χημικής ένωσης και λαμβάνει ως απάντηση το σύνολο ADME των ιδιοτήτων. Οι ιδιότητες που λαμβάνουν είναι ουσιαστικά οι προβλέψεις που κάνουν τα μοντέλα της πλατφόρμας για το χημικό μόριο. Τα αποτελέσματα των επερωτήσεων επεξεργάζονται ώστε να αποθηκεύονται σε

μορφή πίνακα και να υπάρχει όσο το δυνατό καλύτερη ομοιομορφία μεταξύ των διαφορετικών πληροφοριών που λαμβάνονται.

Μετά το πέρας της συλλογής των δεδομένων αναπτύχθηκε κώδικας σε R ο οποίος κάνει την διασύνδεση του neo4j με το προγραμματιστικό περιβάλλον Rstudio. Η διασύνδεση αυτή επιτυγχάνεται με την βοήθεια κατάλληλου πακέτου το neo4R το οποίο παρέχει ένα σύνολο εντολών ώστε ο χρήστης να μπορεί να εισάγει τα δεδομένα που έχουν αποθηκευτεί τοπικά στον γράφο και να δημιουργεί τις κατάλληλες διασυνδέσεις μεταξύ των οντοτήτων. Επιπρόσθετα μέσω του πακέτου αυτού δίνεται η δυνατότητα να γίνονται επερωτήσεις στο neo4j και αυτό επιστρέφει ένα αντικείμενο τύπου γράφο στο οποίο μπορεί να γίνει περαιτέρω ανάλυση. Κάνοντας τα κατάλληλα επερωτήματα προκύπτουν σημασιολογικά δίκτυα τα οποία δείχνουν την αλληλεπίδραση μεταξύ υποσυνόλων των δεδομένων. Τα δίκτυα αυτά μπορούμε να τα κατηγοριοποιούμε ως κύρια και συμπληρωματικά. Επί του παρόντος τα κύρια σημασιολογικά δίκτυα που προκύπτουν περιγράφουν τις εξής σχέσεις:

- φάρμακο-στόχος (πρωτεΐνη)
- φάρμακο-ADME/TOX

Ενώ τα συμπληρωματικά δίκτυα περιγράφουν:

- πρωτεΐνη-πρωτεΐνη
- φάρμακο-πατέντα

Για την ποσοτική ανάλυση των σημασιολογικών δικτύων που προκύπτουν μέσα από την διασύνδεση των οντοτήτων στον γράφο εφαρμόστηκε η θεωρία γράφων. Με βάση την θεωρία αυτή οι σχέσεις μεταξύ των αντικειμένων εντός ενός δικτύου μπορούν να περιγραφούν με τις παρακάτω βασικές μετρητικές:

1. Μέγεθος Δικτύου: θα υπολογιστεί με βάση το συνολικό πλήθος των κόμβων N που περιέχει.
2. Πυκνότητα Δικτύου (d): θα υπολογιστεί με βάση το πηλίκο $d = \frac{E_{\text{πραγματικές}}}{E_{\text{πλήρες}}}$. Όπου $E_{\text{πραγματικές}}$ είναι το πλήθος πραγματικών συνδέσεων ενώ το $E_{\text{πλήρες}}$ δίνεται από την σχέση $E_{\text{πλήρες}}(N) = \frac{N(N-1)}{2}$. Το εύρος τιμών του d είναι 0-1 και ερμηνεύεται ως ο βαθμός αλληλεπίδρασης των στοιχείων του πίνακα.
3. Βαθμός Κόμβου (D): Σε ένα κατευθυνόμενο δίκτυο υπολογίζεται από την εξίσωση $D(n) = I(n) + O(n)$, όπου $D(n)$ είναι ο βαθμός του κόμβου n , $I(n)$ είναι εισερχόμενες (input) ακμές ενώ $O(n)$ είναι οι εξερχόμενες (output) ακμές. Έτσι μπορούμε να εντοπίσουμε ποιοι είναι οι κόμβοι που αλληλεπιδρούν περισσότερο με τους άλλους.
4. Συντομότερη διαδρομή: ουσιαστικά είναι το διάστημα μεταξύ δύο συγκεκριμένων κόμβων το οποίο δίνει την μικρότερη διαδρομή ανάμεσα τους. Πρακτικά αποτελείται από τους διαδοχικούς κόμβους που υπάρχουν μεταξύ του αρχικού και του τελικού κόμβου. Οπότε από αυτό το μέτρο μπορούμε να ποσοτικοποιήσουμε πόσο κοντά είναι 2 κόμβοι. Ο υπολογισμός της γίνεται από αλγορίθμους που έχουν αναπτυχθεί, οι οποίοι υπολογίζουν την απόσταση μεταξύ δύο κόμβων. Σε περίπτωση που το δίκτυο περιλαμβάνει και βάρη μεταξύ των ακμών τότε ο υπολογισμός της συντομότερης διαδρομής μεταβάλλεται. Σε αυτή την περίπτωση ισούται με την διαδρομή που δίνει το ελάχιστο άθροισμα των βαρών των ακμών μεταξύ δύο κόμβων.
5. Κεντρικότητα Κόμβων: Η έννοια αυτή αναφέρεται στην θέση του κόμβου στο δίκτυο και έχει διάφορους τρόπους υπολογισμού. Για να υπολογίσουμε εάν ο κόμβος A είναι κοντά

στο κέντρο του δικτύου από γεωμετρική άποψη χρησιμοποιούμε την εξίσωση $C(A) = \frac{1}{\sum_{i=1}^{N-1} \sigma_{A,j}}$, η οποία εκφράζει την κεντρικότητα εγγύτητας. Ορίζεται ως το του αντίστροφο του αθροίσματος των αποστάσεων του κόμβου απ' όλους τους άλλους κόμβους. Ενώ όταν

θέλουμε να υπολογίσουμε την κεντρικότητα ενός κόμβου από άποψη διασυνδεσιμότητας εφαρμόζουμε την εξίσωση της διακεντρικότητας $G(A) = \frac{\sigma_{i,j}(A)}{\sum_{i=1, j=1, i \neq j \neq A}^N \sigma_{i,j} \sigma_{i,j}}$ η οποία εκφράζει των λόγο των συντομότερων διαδρομών που διέρχονται από έναν κόμβο προς το άθροισμα των συντομότερων διαδρομών.

6. Διάμετρος Δικτύου: υπολογίζεται από τον τύπο $r(A) = \max(\sigma_{i,j})$ και ορίζεται ως η μέγιστη συντομότερη διαδρομή μεταξύ όλων των πιθανών ζευγών του δικτύου. Εκφράζει την διασπορά των κόμβων του δικτύου.
7. Συντελεστής συσσωμάτωσης και συνδεσιμότητα: για έναν δεδομένο κόμβο A δίνεται από την εξίσωση $C(A) = \frac{\sum_{i=1, j=1, i \neq j}^N \alpha_{κμς i,j}}{k(k-1)}$, παίρνει τιμές μεταξύ 0-1 και έχει πιθανοκρατικό χαρακτήρα. Περιγράφει τον βαθμό με τον οποίο οι κόμβοι του δικτύου έχουν την τάση να σχηματίζουν υποδίκτυα. Ενώ ως μέτρο συνδεσιμότητας χρησιμοποιείται η γενικευμένη μορφή έκφραση του συντελεστή συσσωμάτωσης. Η γενικευμένη μορφή εκφράζεται από τον μέσο συντελεστή συσσωμάτωσης που εκφράζει την μέση τιμή των συντελεστών συσχέτισης όλων των κόμβων του δικτύου.
8. Συντελεστής συσχέτισης και καταταξιμότητα: Ισούται με το συντελεστή συσχέτισης Pearson του βαθμού των κόμβων μεταξύ δύο συνδεδεμένων κόμβων και δίνεται από την σχέση:

$$r = \frac{N \sum_{i=1}^N d_{i,1} d_{i,2} - \sum_{i=1}^N d_{i,1} \sum_{i=1}^N d_{i,2}}{\sqrt{\left(N \sum_{i=1}^N d_{i,1}^2 - \left(\sum_{i=1}^N d_{i,1}\right)^2\right) \left(N \sum_{i=1}^N d_{i,2}^2 - \left(\sum_{i=1}^N d_{i,2}\right)^2\right)}}$$

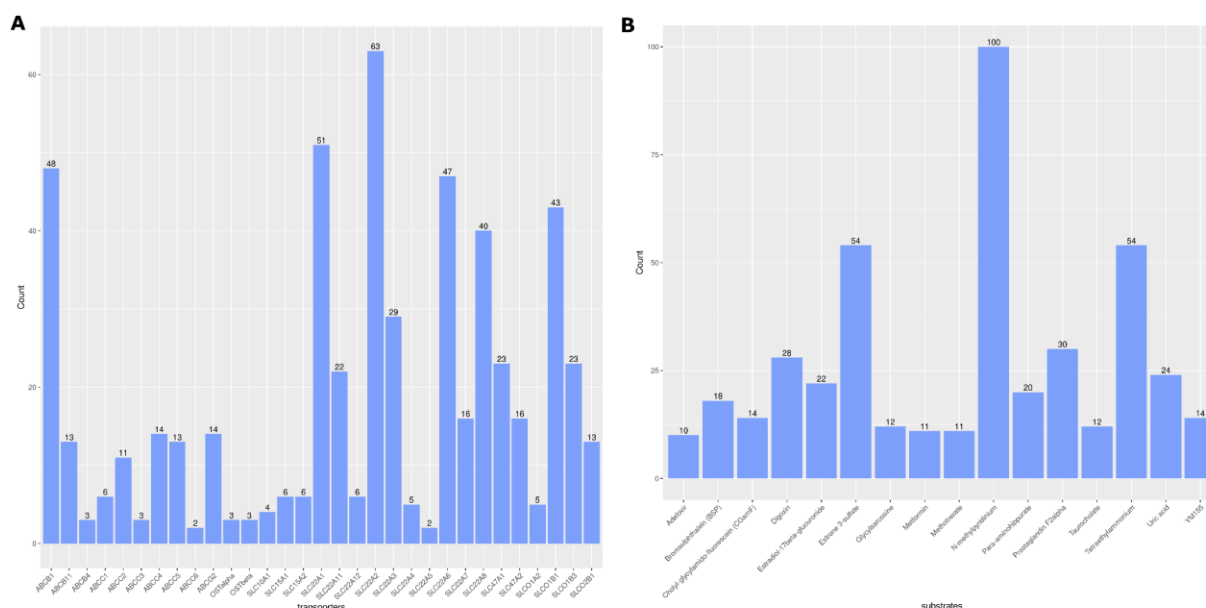
Πρακτικά ο συντελεστής αυτός όταν το $r > 0$ δείχνει θετική συσχέτιση μεταξύ του βαθμού των διασυνδεόμενων κόμβων. Ενώ εάν το $r < 0$ δείχνει το αντίστροφο δηλαδή κόμβοι με μικρό βαθμό συνδέονται με κόμβους μεγάλου βαθμού. Τέλος, η καταταξιμότητα υποδεικνύει τις προτιμήσεις διασύνδεσης των κόμβων ενός δικτύου άρα και την μορφή του δικτύου.

Εφαρμόζοντας τις παραπάνω μετρητικές εξήχθησαν ποσοτικά χαρακτηριστικά που περιγράφουν κάθε σημασιολογικό δίκτυο ξεχωριστά. Οι ποσοτικές αυτές τιμές αναδεικνύουν τις σημαντικές συσχετίσεις εντός ενός δικτύου.

Αποτελέσματα

Το σύνολο των δεδομένων που λήφθηκαν από το αποθετήριο UCSF-FDA αναλύθηκε με σκοπό την εξαγωγή ποιοτικών και ποσοτικών χαρακτηριστικών ώστε να συλλέξουμε χρήσιμες πληροφορίες. Αρχικά, για τα δεδομένα που αφορούν την in vitro αναστολή των πρωτεϊνών μεταφορέων ελέγχθηκε στο σύνολο των δεδομένων ποιες από αυτές τις 31 πρωτεΐνες μεταφορείς αναστέλλονται από την λήψη φαρμάκων πιο συχνά. Συνεπώς οι πρωτεΐνες μεταφορείς μετατράπηκαν σε κατηγορική μεταβλητή και υπολογίστηκε η συχνότητα εμφάνισης τους στις αντιδράσεις αναστολής. Επιπλέον στο ίδιο σύνολο δεδομένων θεωρήσαμε τα υποστρώματα που

εμπλέκονται στην χημική αλληλεπίδραση ως κατηγορική μεταβλητή. Για την μεταβλητή αυτή μετρήθηκε η συχνότητα εμφάνισης κάθε υποστρώματος στο σύνολο των αντιδράσεων. Από την ανάλυση αυτή προέκυψε μια μεγάλη λίστα με υποστρώματα τα οποία είναι φάρμακα. Δεδομένου ότι κάποια από αυτά συμμετέχουν μόνο σε λίγες αντιδράσεις (λιγότερες από δέκα) φιλτράραμε την λίστα ώστε να εντοπίσουμε τα πιο συχνά εμφανιζόμενα. Στο γράφημα 1, παρουσιάζονται τα αποτελέσματα των δύο αναλύσεων αντίστοιχα.



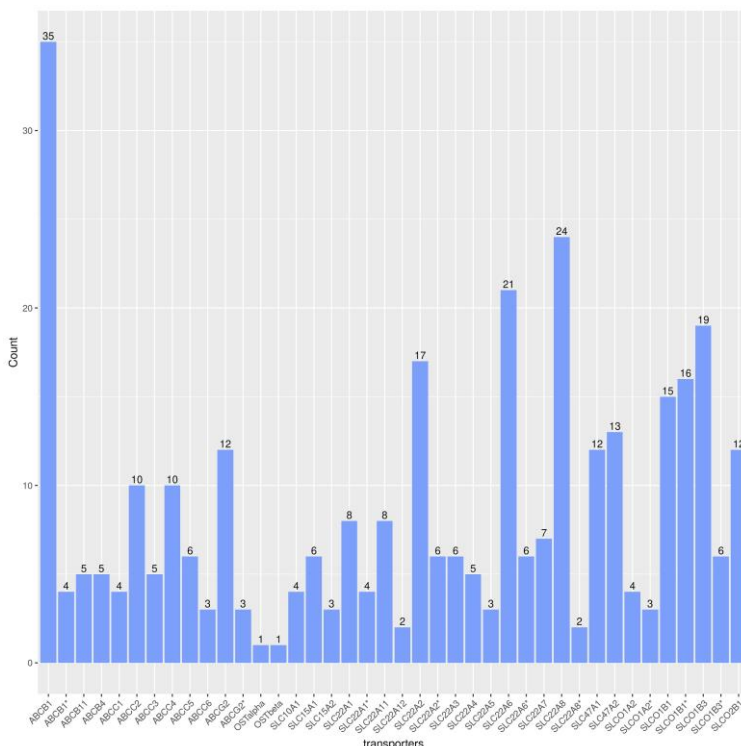
Γράφημα 1: Ραβδογράμματα UCSF-FDA A) συχνότητων εμφάνισης του συνόλου των πρωτεϊνών μεταφορέων οι οποίοι αναστέλλονται από φάρμακα in vitro. B) υποστρωμάτων με συχνότητα εμφάνισης > 10, σε in vitro αντιδράσεις μεταξύ φαρμάκων και πρωτεϊνών μεταφορέων.

Στο γράφημα 1A παρατηρούμε ότι οι πρωτεΐνες μεταφορείς με την μεγαλύτερη συχνότητα εμφανίσεις στο σύνολο 553 αλληλεπιδράσεων είναι η SLC22A2 η οποία βρέθηκε σε 63 αντιδράσεις, η SLC22A1 που μετρήθηκε σε 51 αντιδράσεις και η ABCB1 που αντίστοιχα εντοπίστηκε σε 48 αλληλεπιδράσεις. Οι δυο πρωτεΐνες με την υψηλότερη συχνότητα εμφάνισης ανήκουν στην οικογένεια πρωτεϊνών SLC οι οποίες είναι πρωτεΐνες-φορείς υγρών και μεταφέρουν φορτισμένα και μη οργανικά και ανόργανα χημικά μόρια. Οι πρωτεΐνες αυτές βρίσκονται στην πλασματική μεμβράνη, στα μιτοχόνδρια και σε άλλα οργανίδια. Πιο συγκεκριμένα η SLC22A2 μεσολαβεί στη σωληναριακή πρόσληψη οργανικών ενώσεων από την κυκλοφορία. Ο φορέας αυτός είναι καθοριστικός παράγοντας της αντικαρκινικής δράσης της οξαλιπλατίνης και μπορεί να συμβάλει στην αντικαρκινική εξειδίκευση. Ενώ ο φορέας SLC22A1 μετατοπίζει ένα ευρύ φάσμα οργανικών κατιόντων με διάφορες δομές και μοριακά βάρη. Η τρίτη σε συχνότητα εμφάνισης πρωτεΐνη ABCB1, ανήκει στην υπεροικογένεια των ABC πρωτεϊνών η οποία περιλαμβάνει μεμβρανικές γλυκοπρωτεΐνες (P-gr). Οι συγκεκριμένοι φορείς δρουν ως ATPάσες, αντλίες ξενοβιοτικών ουσιών ώστε είτε να τους εξάγουν από τα κύτταρα είτε να εισάγουν στα κύτταρα θεραπευτικά συστατικά (Lin et al. 2015).

Στο γράφημα 1B που απεικονίζει την συχνότητα εμφάνισης των φαρμάκων που έχουν ρόλο υποστρώματος σε ένα σύνολο 549 αντιδράσεων, το χημικό μόριο N-μεθυλοπυριδίνιο (N-

methilpyridinium) εντοπίστηκε σε 100 αντιδράσεις ενώ σε 54 αλληλεπιδράσεις συμμετείχαν η 3-θειική οιστρόνη (Estrone 3-sulfate) και το τετρααιθυλαμμώνιο (Tetraethylammonium). Το μεθυλοπυριδίνιο είναι μια χημική ένωση τεταρτοταγούς αμμωνίου που προέρχεται από τη N-μεθυλίωση της πυριδίνης και βρίσκεται σε ορισμένα προϊόντα καφέ (Riedel et al. 2014). Η 3-θειική οιστρόνη είναι ένα θειικό στεροειδές που έχει ρόλο μεταβολίτη στον ανθρώπινο οργανισμό (Banerjee et al. 2013). Τέλος το τετρααιθυλαμμώνιο είναι ένα κατιόν τεταρτοταγούς αμμωνίου που αποτελείται από τέσσερις αιθυλομάδες συνδεδεμένες με ένα κεντρικό άτομο αζώτου και είναι θετικά φορτισμένο (“Tetraethylammonium” 2007). Χρησιμοποιείται σε εργαστήρια για την παρασκευή λιπόφιλων αλάτων ανόργανων ανιόντων.

Εκτός από την *in vitro* αναστολή των πρωτεϊνών μεταφορέων μας ενδιαφέρουν και οι αντιδράσεις κατά τις οποίες τα φάρμακα δρουν ως υπόστρωμα των πρωτεϊνών αυτών. Συνεπώς για τα δεδομένα που περιέχονται στην κατηγορία *in vitro* substrates μετρήθηκε για το σύνολο των 31 πρωτεϊνών μεταφορέων πόσο συχνά έχουν φάρμακα ως υπόστρωμα. Τα αποτελέσματα της ανάλυσης αυτής παρατίθενται στο γράφημα 2.

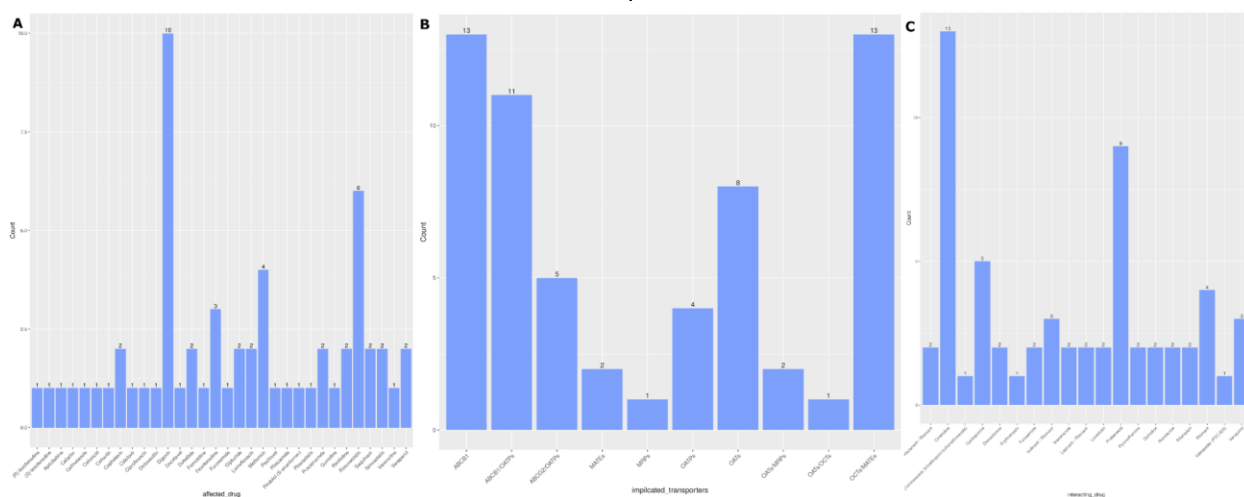


Γράφημα 2: Ραβδόγραμμα συχνότητας εμφάνισης του συνόλου των πρωτεϊνών μεταφορέων της UCSF-FDA στις οποίες τα φάρμακα αποτελούν υπόστρωμα *in vitro*.

Στο γράφημα 2 βλέπουμε και εδώ πως ο φορέας ABCB1 συμμετέχει σε 35 από τις 336 αντιδράσεις. Επίσης οι πρωτεΐνες SLC22A8 και SLC22A6 εντοπίστηκαν σε 24 και 21 αλληλεπιδράσεις αντίστοιχα. Η SLC22A8 έχει σημαντικό ρόλο στην απέκκριση ενδογενών και εξωγενών οργανικών ανιόντων, ιδιαίτερα από τον εγκέφαλο και τα νεφρά. Ενώ η SLC22A6 συμμετέχει στη νεφρική αποβολή ενδογενών και εξωγενών οργανικών ανιόντων (Srimaroeng et al. 2005).

Τέλος μια πολύ σημαντική πληροφορία που μπορούμε να εξάγουμε από τα δεδομένα που έχουμε συλλέξει από το αποθετήριο UCSF-FDA αφορά την αλληλεπίδραση μεταξύ φαρμάκων η

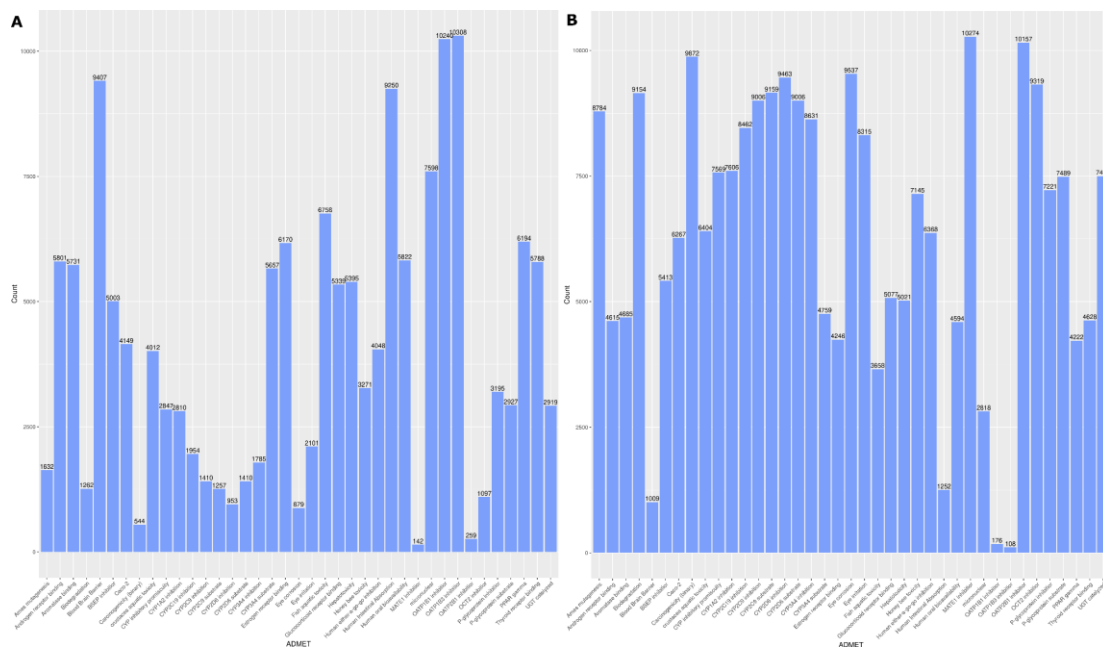
οποία οφείλεται στις πρωτεΐνες μεταφορείς. Έτσι εντοπίστηκαν και μετρήθηκαν τα φάρμακα τα οποία επηρεάζονται από τις αλληλεπιδράσεις των μεταφορέων με άλλα φάρμακα. Επίσης, μετρήθηκε η συχνότητα εμφάνισης των πρωτεϊνών μεταφορέων που επηρεάζουν τις αλληλεπιδράσεις μεταξύ φαρμάκων. Ενώ αντίστοιχα βρέθηκαν τα φάρμακα που αλληλεπιδρούν με τις πρωτεΐνες μεταφορείς και οι συχνότητα των αλληλεπιδράσεων. Τα αποτελέσματα βρίσκονται στο γράφημα 3. Οι αντιδράσεις αυτές είναι πολύ σημαντικές καθώς από αυτές εξαρτάται η φαρμακοκινητική και η φαρμακοδυναμική των φαρμάκων.



τοξίνης) ανήκουν στην οικογένεια SLC47. Ωστόσο παρατηρείται ότι υπάρχει μεταξύ των δυο μια αλληλεπικαλυπτόμενη επιλεκτικότητα στα υποστρώματα (Koepsell 2020). Οι πρωτεΐνες ABCB1/OATPs εντοπίστηκαν σε 11 αλληλεπιδράσεις, ανήκουν και οι δύο τύποι μεταφορέων στην ευρύτερη υπεροικογένεια πρωτεϊνών SLC. Ωστόσο ταξινομούνται οι OATPs ταξινομούνται στις SLC21A πρωτεΐνες ενώ οι ABCB1 ταξινομούνται στις SLC22A. Οι πρωτεΐνες αυτές εκφράζονται σε κάθε επιθήλιο σε όλο το σώμα και συμμετέχουν στην απορρόφηση, κατανομή και αποβολή του φαρμάκου. Αρκετοί πολυμορφισμοί που έχουν εντοπιστεί σε μέλη αυτών των υπεροικογενειών αποδεδειγμένα επηρεάζουν τη φαρμακοκινητική των υποστρωμάτων (φαρμάκων) τους. Έστι αυτοι οι μεταφορείς φαρμάκων επηρεάζουν την αποτελεσματικότητα της φαρμακολογικής θεραπείας (Roth, Obaidat, and Hagenbuch 2012).

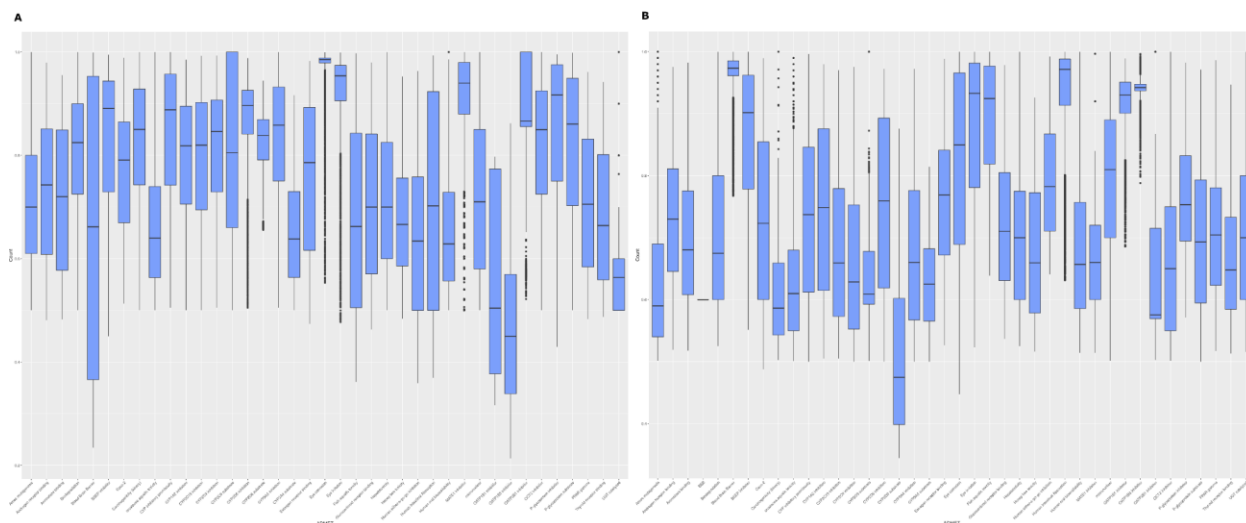
Στο διάγραμμα 3C βλέπουμε τα φάρμακα που κατά την συγχορήγηση ουσιαστικά είναι αυτά που αντιδρούν με τις πρωτεΐνες φορείς. Τα τρία πιο συχνά εμφανιζόμενα φάρμακα που εντοπίζουμε στο διάγραμμα είναι η σιμετιδίνη (Cimetidine), η προβενεσίδη (probenecid) και η κυκλοσπορίνη (Cyclosporine). Η σιμετιδίνη με συχνότητα εμφάνισης ίση με 13 έχει ανασταλτική δράση στην έκκριση γαστρικού οξέος και ελαττώνει την παραγωγή πεψίνης. Χορηγείται για τη θεραπεία του πεπτικού έλκους και της γαστροοισοφαγικής παλινδρόμησης (Somogyi and Gugler 1983). Η δεύτερη σε συχνότητα εμφάνισης προβενεσίδη μετρήθηκε 9 φορές και η δράση της είναι να αυξάνει την απέκκριση του ουρικού οξέος στα ούρα ώστε να μειώνει τα επίπεδα ουρικού οξέος στον ορό. Χορηγείται σε περιπτώσεις ουρικής αρθρίτιδας ("Probenecid" 2007). Τέλος η κυκλοσπορίνη η οποία εμπλέκεται σε 5 αλληλεπιδράσεις έχει ανοσοκατασταλτικές ιδιότητες και αντιφλεγμονώδη δράση (Forsythe and Paterson 2014).

Οι προβλέψεις που συγκεντρώσαμε από την AdmetSAR αναλύθηκαν ποσοτικά προκειμένου να αναδειχθούν οι ADME ιδιότητες των χημικών μορίων. Πιο συγκεκριμένα τα αποτελέσματα των προβλέψεων της AdmetSAR έχουν ένα ποιοτικό δείκτη για κάθε ιδιότητα ο οποίος είναι + ή - . Ο δείκτης αυτός υποδηλώνει αν το χημικό μόριο έχει την εκάστοτε ιδιότητα. Με βάση αυτό υπολογίστηκε πόσα μόρια παρουσιάζουν ή όχι την κάθε ιδιότητα. Συμπληρωματικά, δίνει και την πιθανότητα η πρόβλεψη της εκάστοτε ιδιότητας είναι ορθή. Συλλέγοντες τις πιθανότητες αυτές για κάθε χημικό μόριο μπορούμε να δούμε πόσο αξιόπιστες είναι οι προβλέψεις του μοντέλου καθώς επίσης και ποιες ιδιότητες είναι πιο πιθανό να εμφανιστούν. Στο γράφημα 4 απεικονίζεται η συχνότητα παρατήρησης ή μη παρατήρησης της κάθε ADME ιδιότητας ενώ στο γράφημα 5 παρατίθενται οι αντίστοιχες πιθανότητες για τις προβλέψεις των μοντέλων.



Γράφημα 4: Ραβδόγραμμα AdmetSAR A) συχνότητα εμφάνισης χημικών που παρουσιάζουν την εκάστοτε ADME ιδιότητα B) συχνότητα εμφάνισης χημικών που δεν παρουσιάζουν την εκάστοτε ADME ιδιότητα

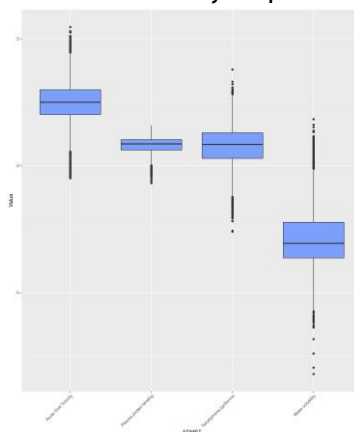
Από τα 14314 χημικά μόρια που εξετάσαμε, συγκεντρώσαμε προβλέψεις για τα 10502. Στην συνέχεια ελέγχθηκαν οι προβλέψεις αυτές ώστε να δούμε ποιοτικά τις ADME ιδιότητες τους. Έτσι από το διάγραμμα 4A προκύπτει ότι από τα 10502 τα 10308 είναι με βάση την πρόβλεψη αναστολείς της πρωτεΐνης OATP1B3, 10240 είναι αναστολείς της πρωτεΐνης OATP1B1 ενώ 9407 φαίνεται πως διαπερνούν τον αιματοεγκεφαλικό φραγμό. Η OATP1B3 και η OATP1B1 είναι πρωτεΐνες φορείς και χρησιμοποιούν ως υπόστρωμα πολλές θεραπευτικές ουσίες. Συνεπώς η αναστολή των πρωτεϊνών αυτών πρέπει να λαμβάνεται σοβαρά υπόψη κατά την συγχρόνηση φαρμάκων (Shitara 2011). Ο αιματοεγκεφαλικός φραγμός δεν επιτρέπει στην πλειονότητα των φαρμάκων και χημικών μορίων την διέλευση λόγω των στενών συνδέσεων που παρουσιάζουν τα τριχοειδή στο ενδοθήλιο του εγκεφάλου. Ωστόσο χημικές ουσίες που έχουν συγκεκριμένες φυσικοχημικές ιδιότητες μπορούν να διαπεράσουν τον φραγμό. Στο διάγραμμα 4B φαίνεται ο αριθμός των χημικών ουσιών που δεν έχουν τις αντίστοιχες ADME ιδιότητες. Παρατηρούμε ότι από τα 10502 μόρια τα 10274 δεν αναστέλλουν την πρωτεΐνη φορέα MATE1, 10157 δεν αναστέλλουν την πρωτεΐνη φορέα OATP1B1 και 9872 δεν παρουσιάζουν καρκινογένεση με βάση τις προβλέψεις της πλατφόρμας.



Γράφημα 5: Θηκογράμματα AdmetSAR A) κατανομής των πιθανοτήτων των προβλέψεων να υπάρχει η κάθε ADME ιδιότητα στα χημικά μόρια ενδιαφέροντος. B) κατανομής των πιθανοτήτων των προβλέψεων να μην υπάρχει η κάθε ADME ιδιότητα στα χημικά μόρια ενδιαφέροντος.

Στο παραπάνω γράφημα βλέπουμε την κατανομή των προβλέψεων της πλατφόρμας AdmeSAR για το εάν τα μόρια φέρουν τις δεδομένες ADME ιδιότητες. Οι κατανομές των προβλέψεων θα αξιοποιηθούν συνδυαστικά με τα αποτελέσματα στο γράφημα 5 ώστε να κρίνουμε την αξιοπιστία της πρόβλεψης για κάθε μια ιδιότητα. Δεδομένου ότι πρόκειται για προβλέψεις μοντέλων μηχανικής μάθησης θα πρέπει ιδανικά οι προβλέψεις για μια ιδιότητα να έχουν χαμηλή διακύμανση και όσο υψηλότερη ακρίβεια.

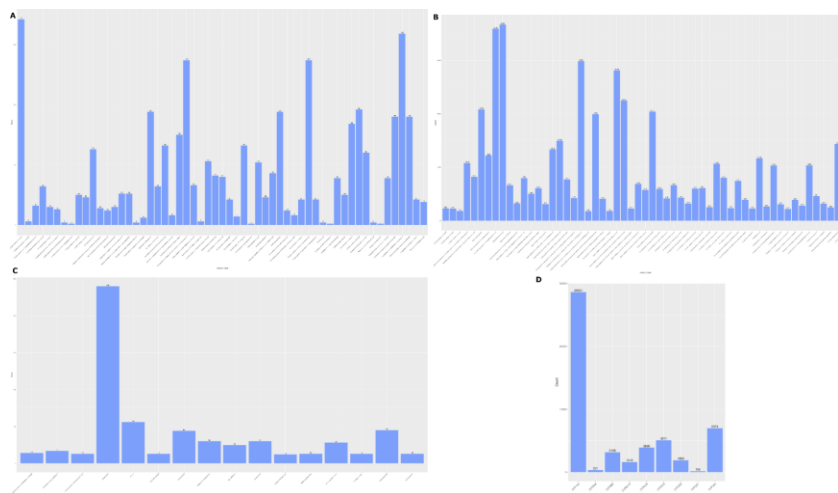
Η AdmeSAR διαθέτει ακόμη και μοντέλα παλινδρόμησης τα οποία μπορούν να προβλέψουν τις τιμές συγκεκριμένων ADME ιδιοτήτων για τα χημικά μόρια. Έτσι συλλέχθηκαν οι αντίστοιχες προβλέψεις για το σύνολο των χημικών μορίων που διαθέτουμε και σχηματίστηκε το θηκόγραμμα του γραφήματος 6 το οποίο απεικονίζει την κατανομή των προβλέψεων.



Γράφημα 6: Θηκόγραμμα κατανομής των προβλέψεων από τα μοντέλα παλινδρόμησης για τις ADME ιδιότητες των χημικών μορίων.

Στα θηκογράμματα του διαγράμματος 6 φαίνονται οι κατανομές των προβλέψεων των μοντέλων παλινδρόμησης για 4 συνεχείς μεταβλητές που αντιπροσωπεύουν τις ADME ιδιότητες χημικών μορίων. Παρατηρούμε ότι τα θηκογράμματα αυτά έχουν μικρή διακύμανση που υποδεικνύουν σχετική αξιοπιστία των προβλέψεων.

Στο σύνολο των δεδομένων που εξορύχθηκαν από την πλατφόρμα Biotransformer έγινε ποσοτική ανάλυση ώστε να εξαχθούν χρήσιμες πληροφορίες για τους μετασχηματισμούς που λαμβάνουν χώρα κατά την 1η φάση του μεταβολισμού των χημικών μορίων. Η 1η φάση του μεταβολισμού αποτελείται από την αλληλεπίδραση των λαμβανόμενων ξενοβιοτικών με τις ισομορφές του κυτοχρώματος P450 (Wrighton, VandenBranden, and Ring 1996). Πιο συγκεκριμένα, τα δεδομένα αναλύθηκαν με σκοπό την εύρεση και την καταμέτρηση των χημικών αντιδράσεων γίνονται σε εκείνο το στάδιο. Ακόμη μετρήθηκε το πλήθος των παραγόμενων προϊόντων καθώς και τα ένζυμα που συμμετέχουν σε αυτές τι βιοχημικές αντιδράσεις. Στο γράφημα 7 απεικονίζονται τα αποτελέσματα των αναλύσεων αυτών.

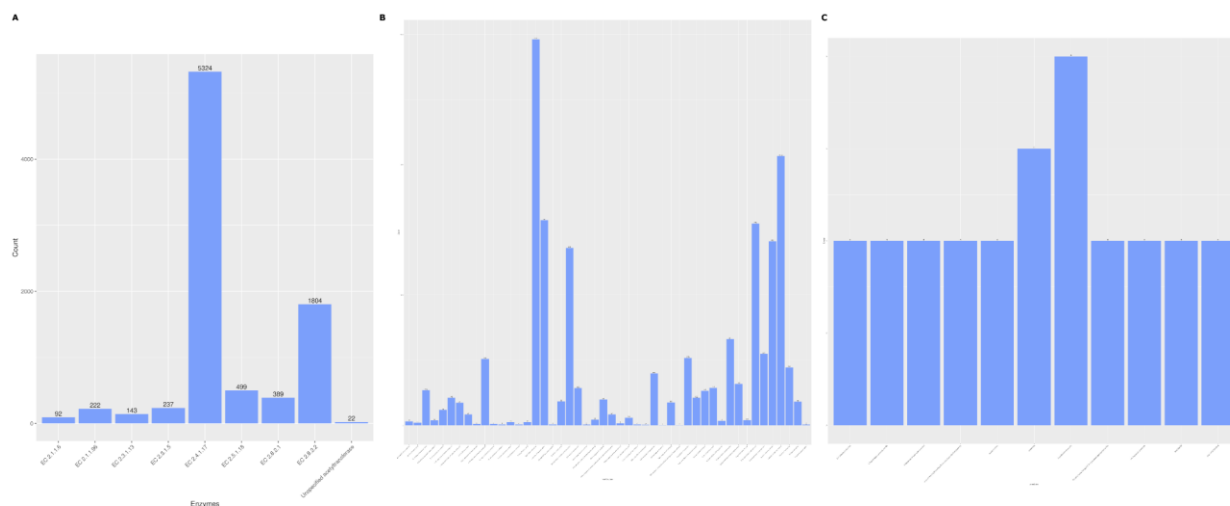


Γράφημα 7: Ραβδογράμματα Biotransformer A) και B) το πλήθος των αντιδράσεων που λαμβάνουν χώρα κατά τη 1η φάση του μεταβολισμού όπου το χημικό μόριο αντιδρά με τις ισομορφές του κυτοχρώματος P450 C) πλήθος των παραγώγων της 1ης φάσης του μεταβολισμού D) πλήθος των ενζύμων που συμμετέχουν στις αντιδράσεις της 1ης φάσης του μεταβολισμού

Στο γράφημα 7 στο A και B βλέπουμε τις αντιδράσεις της 1ης φάσης μεταβολισμού και παρατηρούμε ότι η πιο συχνά εμφανιζόμενη αντίδραση στην πρώτη φάση του μεταβολισμού για την βιομετατροπή των χημικών είναι η εποξειδωση των αρωματικών ενώσεων. Η συγκεκριμένη αντίδραση εμφανίζεται 3669 φορές. Δεύτερη σε συχνότητα εμφανίσεις αντίδραση είναι αρωματική υδροξυλίωση του δακτυλίου βενζολίου η οποία παρατηρείται 3590 φορές. Ενώ η τρίτη σε σειρά αντίδραση που εμφανίζεται πιο συχνά είναι η υδροξυλίωση βενζολίου σε άνθρακα στην ορθο θέση, σε ισχυρό δότη ηλεκτρονίων η οποία εντοπίζεται 2986 φορές.

Στο ραβδόγραμμα 7C απεικονίζονται τα πιο συχνά εμφανιζόμενα παράγωγα που προκύπτουν από τις μεταβολικές αντιδράσεις. Παρατηρούμε λοιπόν ότι οξική αλδεΐνη (acetaldehyde) έχει τη μεγαλύτερη συχνότητα εμφανίσεις η οποία ισούται με 240. Ακολουθεί η ακετόνη (acetone) με 56 εμφανίσεις και η προπιοναλδεΐδη (propionaldehyde) με 45. Αντίστοιχα στο ραβδόγραμμα 7D φαίνονται τα ένζυμα που συμμετέχουν στις μεταβολικές αντιδράσεις με το CYP1A2 να είναι το πιο συχνά εμφανιζόμενο καθώς εντοπίζεται 28631 φορές. Το CYP3A4 είναι δεύτερο πιο συχνά εμφανιζόμενο και μετρήθηκε 6974 φορές. Το τρίτο πιο συχνά εμφανιζόμενο είναι το CYP2C9 με συχνότητα ίση με 5071. Τα προαναφερθέντα ένζυμα αποτελούν ισομορφές του κυτοχρώματος P450 και η αλληλεπίδραση τους με τα χημικά μόρια παίζει καθοριστικό ρόλο στην φαρμακοκινητική των βιομορίων.

Αντίστοιχη ανάλυση έγινε και στα δεδομένα που αφορούν την δεύτερη φάση του μεταβολισμού. Κατά την φάση αυτή έχουμε τον συνδυασμό των ξενοβιοτικών με τα ενδοβιοτικά μόρια ("Concepts in Pharmacology and Toxicology" 2007). Και γι' αυτή τη φάση του μεταβολισμού μετρήθηκε η συχνότητα συμμετοχής των ενζύμων στις αντιδράσεις αυτές. Επίσης υπολογίστηκε η συχνότητα εμφάνισης κάθε τύπου αντίδρασης μεταξύ ξενοβιοτικών και ενδοβιοτικών. Τέλος μετρήθηκε το πλήθος των παραγόμενων προϊόντων που προκύπτουν από τις αντιδράσεις.



Γράφημα 8: Ραβδόγραμμα Biotransformer A) Συχνότητα εμφάνισης των ενζύμων που συμμετέχουν στην 2η φάση του μεταβολισμού. B) Συχνότητα εμφάνισης χημικών αντιδράσεων μεταξύ ξενοβιοτικών και ενδοβιοτικών C) Συχνότητα εμφάνισης παραγώγων της 2ης φάσης του μεταβολισμού.

Στο γράφημα 8A απεικονίζεται η συχνότητα εμφάνισης των ενζύμων που συμμετέχουν στην 2η φάση του μεταβολισμού των χημικών μορίων. Βλέπουμε τα τρία πιο συχνά εμφανιζόμενα ένζυμα είναι η γλυκουρονοσυλ-τρανσφεράση (EC 2.4.1.17 - glucuronosyltransferase) που εμφανίζεται 5324 φορές, η αλκοολική σουλφοτρανσφεράση (EC 2.8.2.2 - alcohol sulfotransferase) που εντοπίζεται 1804 φορές και η τρανσφεράση γλουταθειόνης (EC 2.5.1.18 - glutathione transferase) με συχνότητα ίση με 499. Αντίστοιχα στο διάγραμμα 8B παρουσιάζεται η συχνότητα των χημικών αντιδράσεων βιομετατροπής κατά την 2η φάση του μεταβολισμού. Στο ραβδόγραμμα αυτό την μεγαλύτερη συχνότητα έχει η Αλκυλο-υδροξείδιο γλυκουρονιδίωση (Alkyl-OH glucuronidation) που εμφανίζεται 1481 φορές, ακολουθεί η θείωση των δευτεροταγών αλκοολών (Sulfation of secondary alcohol) με συχνότητα 1033 και τρίτη σε συχνότητα εμφάνισης είναι η αρωματική υδροξυλίωση γλυκουρονιδίωση (Aromatic -OH glucuronidation) που μετρήθηκε 787 φορές. Τέλος στο ραβδόγραμμα 8C φαίνονται τα παράγωγα των αντιδράσεων της 2ης φάσης του μεταβολισμού. Πιο συγκεκριμένα η ακετυλο σουλφαμεθοξαζόλη (Acetylsulfamethoxazole) παρατηρείται 8 φορές, το ακετανιλίδιο (acetanilide) 6 φορές ενώ όλα τα υπόλοιπα παράγωγα συναντώνται από 4 φορές.

Τα δεδομένα που συλλέχθηκαν από την PharmGKB αναλύθηκαν ώστε να εντοπιστούν τα γονίδια με τις περισσότερες φαρμακολογικές αλληλεπιδράσεις. Επιπλέον προσδιορίστηκαν οι πιο συχνά εμφανιζόμενες ισομορφές που σχετίζονται με τις αλληλεπιδράσεις αυτές καθώς και ο τύπος

Figure 2 consists of three bar charts (A, B, and C) showing the number of genes with a specific number of orthologs. The y-axis for all charts is 'Count'.

Panel A: Number of genes with 1 to 25 orthologs.

Ortholog Count	Gene Count
1	12
2	18
3	24
4	20
5	15
6	12
7	10
8	8
9	6
10	4
11	3
12	2
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1

Panel B: Number of genes with 1 to 25 orthologs, categorized by taxonomic group.

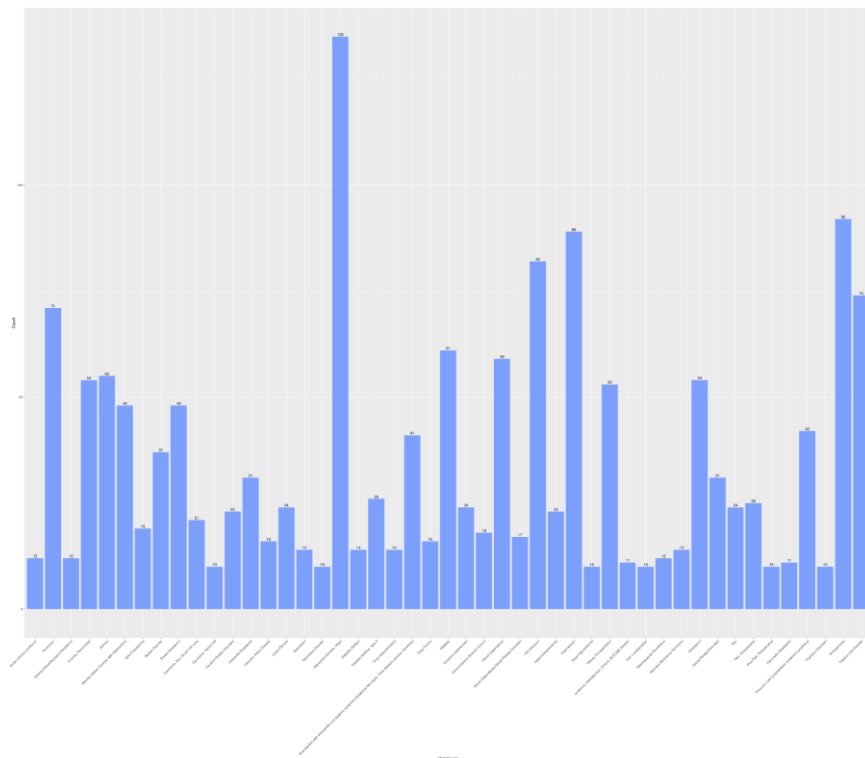
Ortholog Count	Count
1	17
2	30
3	11
4	18
5	52
6	25
7	24
8	31
9	75
10	38
11	16
12	10
13	33
14	17
15	11
16	11
17	22
18	25
19	13
20	11
21	11
22	10
23	26
24	14
25	14

Panel C: Number of genes with 1 to 25 orthologs, categorized by taxonomic group and ortholog count.

Ortholog Count	Count
1	78
2	8
3	8
4	1
5	6
6	112
7	16
8	38
9	2
10	402
11	258
12	85
13	6

Στο παραπάνω πάνελ στο διάγραμμα 9Α βλέπουμε το πλήθος των ισομορφών κάθε γονιδίου που επηρεάζονται από την φαρμακολογική δράση των βιομορίων με συχνότητα > 10. Βλέπουμε λοιπόν ότι οι περισσότερες ισομορφές με φαρμακολογική αλληλεπίδραση εντοπίζονται στο γονίδιο ABCB1 καθώς φέρει 175 ισομορφές. Επιπλέον παρατηρούμε ότι η δεύτερη σε κατάταξη μέτρηση είναι το 148 το οποίο ωστόσο δεν φέρει όνομα γονιδίου στον χ άξονα. Αυτό σημαίνει ότι στο σύνολο των δεδομένων υπάρχουν 148 ισομορφές με φαρμακολογική δράση οι οποίες όμως δεν έχουν συσχετιστεί με κάποιο γονίδιο. Το 3ο σε σειρά κατάταξης είναι το γονίδιο OPRM1 το οποίο σύμφωνα με την PharmGKB φέρει 83 ισομορφές που έχουν φαρμακολογικές αλληλεπιδράσεις. Επίσης στο ίδιο πάνελ στο διάγραμμα 9Β βλέπουμε την συχνότητα εμφανίσεις των ισομορφών που αλληλεπιδρούν με τα λαμβανόμενα βιομόρια. Από αυτό το ραβδόγραμμα προκύπτει ότι η πιο συχνά εμφανιζόμενη ισομορφή είναι η rs1045642 που ουσιαστικά είναι ένας πολυμορφισμός σε επίπεδο ενός νουκλεοτιδίου (SNPs) του γονιδίου ABCB1. Η επόμενη ισομορφή σε συχνότητα εμφάνισης είναι το CYP2D6*1 που εντοπίζεται 52 φορές. Το CYP2D6 είναι ένα γονίδιο που η κωδικοποιεί το αντίστοιχο ένζυμο της οικογένειας του κυτοχρώματος P450, το οποίο μεταβολίζει το 25% των φαρμάκων που χρησιμοποιούνται κλινικά. Επειδή το γονίδιο αυτό υπάρχει σε πολλά αντίγραφα το *1 προσδιορίζει τον απλότυπο του. Τρίτη σε συχνότητα εμφάνισης με 38 μετρήσεις είναι η ισομορφή rs1799971 η οποία υποδηλώνει γενετική αλλαγή κατά την οποία έχουμε υποκατάσταση ενός ζεύγους βάσεων (missense variant) του γονιδίου OPRM1. Τέλος στο ραβδόγραμμα 3C φαίνεται το αποτέλεσμα της επίδρασης που έχει η φαρμακολογική δράση των βιομορίων στις ισομορφές. Σε αυτό βλέπουμε πως 1152 αλληλεπιδράσεις προκαλούν αποτελεσματικότητα στην δράση των φαρμάκων, ενώ 994 προκαλούν τοξικότητα. Ακόμη βλέπουμε πως 402 αλληλεπιδράσεις επηρεάζουν τον μεταβολισμό και φαρμακοκινητική των βιομορίων.

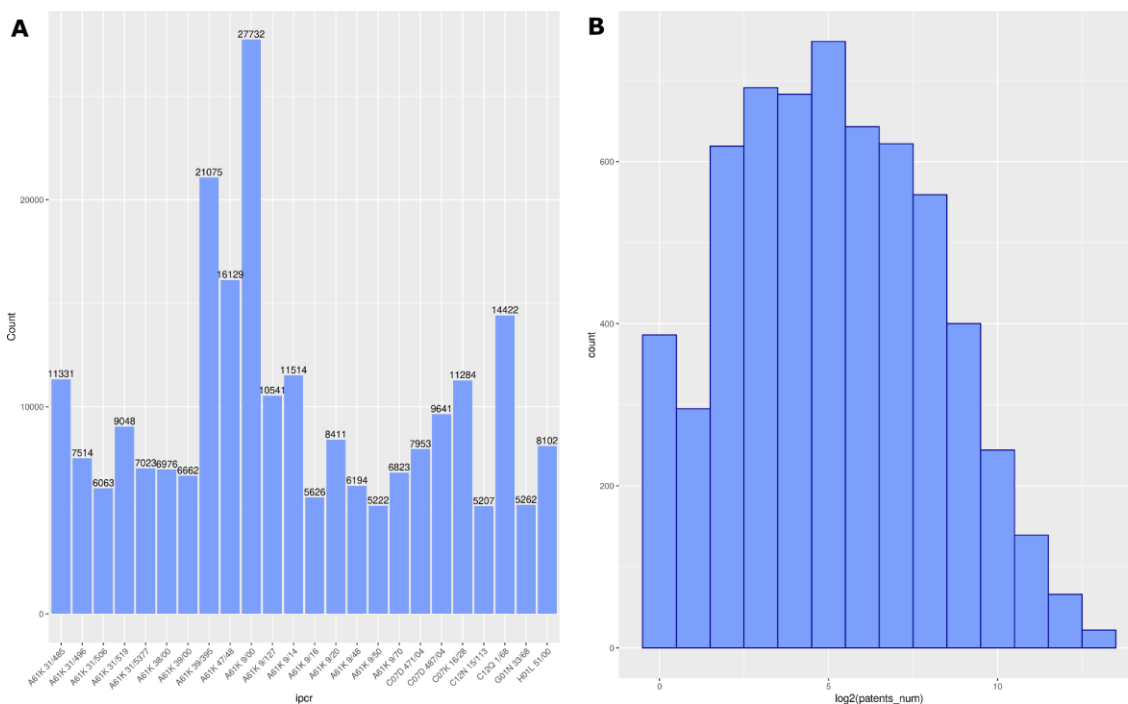
Επιπλέον προσδιορίστηκαν οι πιο συχνά εμφανιζόμενοι φαινότυποι των ασθενειών που σχετίζονται με την φαρμακολογική αλληλεπίδραση μεταξύ φαρμάκων και γονιδιακών ισομορφών. Τα αποτελέσματα αυτής της ανάλυσης παρουσιάζονται στο γράφημα 10.



Γράφημα 10: Συχνότητα εμφάνισης των ασθενειών στις οποίες παρατηρείται φαρμακολογική αλληλεπίδραση

Από το ραβδόγραμμα του γραφήματος 10 προκύπτει ότι η μείζων καταθλιπτική διαταραχή με 135 αλληλεπιδράσεις είναι η ασθένεια η οποία φαίνεται πως επηρεάζεται περισσότερο από τις φαρμακογονιδιωματικές. Η δεύτερη πιο συχνά επηρεαζόμενη ασθένεια είναι η σχιζοφρένεια η οποία επηρεάζεται από 92 αλληλεπιδράσεις και ακολουθεί η υπέρταση με 89.

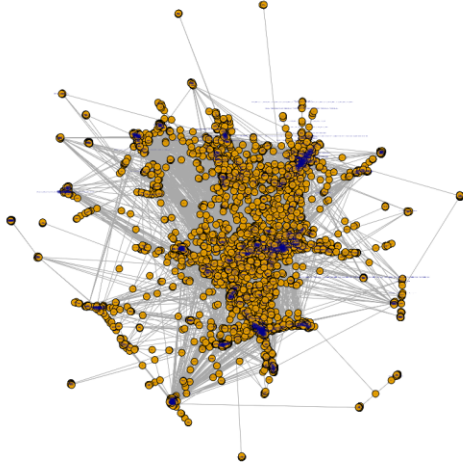
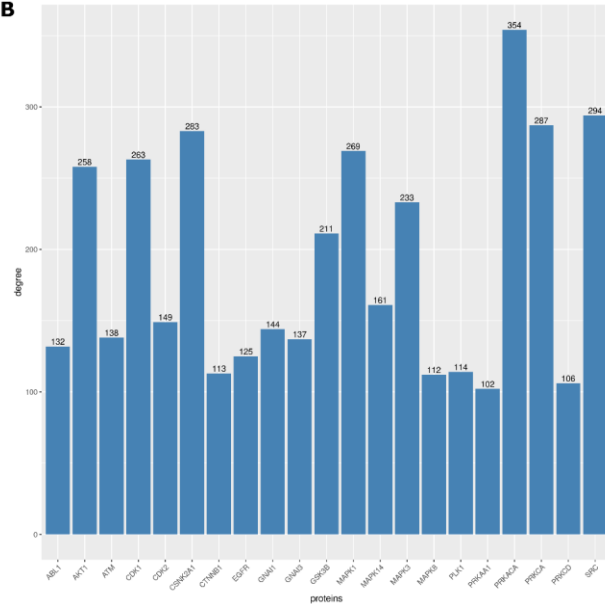
Οι αρχικές πληροφορίες που συγκεντρώθηκαν για τις ευρεσιτεχνίες που αφορούν τα χημικά μόρια ενδιαφέροντος που διαθέτουμε αναλύθηκαν ώστε να μετρηθεί το πλήθος τους ανα χημικό μόριο. Επίσης θεωρήθηκε σημαντικό να μετρηθεί η συχνότητα εμφάνισης των κατηγοριών στις οποίες ανήκουν οι πατέντες που έχουμε συλλέξει. Με αυτό τον τρόπο μπορούμε να αναγνωρίσουμε αν οι ευρεσιτεχνίες των χημικών μορίων αφορούν φαρμακευτική ή κάποια άλλη χρήση. Κατα την ανάλυση των δεδομένων παρατηρήθηκε ότι υπάρχουν χημικά μόρια με πάρα πολύ μεγάλο πλήθος ευρεσιτεχνιών (μερικές χιλιάδες) ενώ άλλες έχουν μόνο μερικές δεκάδες. Έτσι λογαριθμίσουμε το πλήθος των πατεντών με βάση το 2 ($\log_2$) προκειμένου να έχουμε μια καλύτερη αναπαράσταση των αποτελεσμάτων. Συνεπώς ένα χημικό μόριο με μια πατέντα μετά τον μετασχηματισμό θα φαίνεται πως έχει μηδέν, ενώ αντίστοιχα αν έχει 1000 πατέντες στην νέα κλίμακα θα είναι περίπου 10.



Γράφημα 11: Α) Ραβδόγραμμα με το πλήθος των ευρεσιτεχνιών της SureChEMBL ανά τύπο κατηγορίας στον οποίο έχουν ταξινομηθεί. Β) Ιστόγραμμα του πλήθους των πατεντών που υπάρχουν για κάθε χημικό μόριο.

Στο γράφημα 11Α βλέπουμε την κατηγορία στην οποία ανήκουν οι πατέντες των χημικών μορίων με βάση την ταξινόμηση που γίνεται από το διεθνές γραφείο ευρεσιτεχνιών. Η κατηγοριοποίηση γίνεται με βάση τον κωδικό *ipc* που υποδεικνύει τον τομέα εφαρμογής της πατέντας (“IPC Publication” n.d.). Προκύπτει λοιπόν από το γράφημα 11Α ότι ο πιο συχνά εμφανιζόμενος τύπος πατέντα αντιστοιχεί στον κωδικό *ipc* A61K 9/00 που εμφανίζεται 27732 φορές. Ο κωδικός αυτός δίνεται σε ευρεσιτεχνίες οι οποίες αφορούν φαρμακευτικά σκευάσματα που χαρακτηρίζονται από ειδική φυσική μορφή. Ο δεύτερος πιο συχνά εμφανιζόμενος κωδικός *ipc* είναι ο A61K 39/395 με 21075 εμφανίσεις ο οποίος αφορά πατέντες που σχετίζονται με αντισώματα, ανοσοσφαιρίνες και ανοσοποιητικούς ορούς. Ο τρίτος σε κατάταξη κωδικός *ipc* με 16129 εμφανίσεις είναι ο A61K 47/48 ο οποίος ανήκει στην ευρύτερη κατηγορία A61K 47 η οποία εμπεριέχει φαρμακευτικά παρασκευάσματα που χαρακτηρίζονται από τα μη δραστικά συστατικά που χρησιμοποιούνται, π.χ. φορείς ή αδρανή πρόσθετα και παράγοντες στόχευσης ή τροποποίησης που συνδέονται χημικά με το δραστικό συστατικό. Επιπλέον από το ιστόγραμμα 11Β βλέπουμε πως κατανέμεται το πλήθος των πατεντών σε λογαριθμική κλίμακα. Πιο αναλυτικά βλέπουμε ότι έχουμε 750 περίπου χημικά μόρια τα οποία έχουν $\log_2(b)=5$ άρα $b=2^5=32$. Συνεπώς βλέπουμε ότι πλειονότητα των χημικών μορίων έχει περίπου 32 πατέντες.

Αντίθετα με τις προαναφερθέντες βάσεις δεδομένων όπου οι πληροφορίες βρίσκονται καταχωρημένες σε πίνακες και ακολουθούν προκαθορισμένη δομή, τα δεδομένα από το αποθετήριο Omnipath περιγράφουν σχέσεις μεταξύ πρωτεϊνών. Συνεπώς η αναπαράσταση τους γίνεται σχηματίζοντας δίκτυα και η ανάλυση τους υλοποιείται με βάση την θεωρία γράφων. Ενδεικτικά αποτελέσματα αυτής της προσέγγισης φαίνονται στο γράφημα 12 όπου σαν παράδειγμα έχουμε ένα δίκτυο πρωτεϊνικών αλληλεπιδράσεων.

A**B**

Γράφημα 12: Α) Δίκτυο πρωτεϊνικών αλληλεπιδράσεων Β) Ραβδόγραμμα των πρωτεϊνών που παρουσιάζουν τις περισσότερες αλληλεπιδράσεις.

Στην εικόνα 12Α έχουμε αναπαράσταση των αλληλεπιδράσεων των πρωτεϊνών σε μορφή δικτύου. Το δίκτυο αυτό αναλύθηκε ώστε να βρεθούν οι κόμβοι με τις περισσότερες συνδέσεις. Αυτό αποδεικνύει την κεντρικότητα των κόμβων ως προς την τοπολογία του δικτύου. Τα αποτελέσματα αυτά φαίνονται στο ραβδοδιάγραμμα 12Β όπου για το συγκεκριμένο παράδειγμα βλέπουμε τους κόμβους δηλαδή τις πρωτεΐνες με τις περισσότερες αλληλεπιδράσεις. Πιο αναλυτικά βλέπουμε ότι οι πρωτεΐνες με τις περισσότερες συνδέσεις είναι η PRKACA που συνδέεται με 354 πρωτεΐνες και η SRC που αλληλεπιδρά με 294. Χρησιμοποιώντας αυτή την μεθοδολογία μπορούμε να ποσοτικοποιήσουμε την σημαντικότητα των κόμβων του δικτύου καθώς όσο πιο κεντρικός ο κόμβος τόσο περισσότερες οι αλληλεπιδράσεις που θα επηρεαστούν σε περίπτωση αναστολής της λειτουργίας του.

Μελλοντική δουλειά

Πέρα από την εξόρυξη δεδομένων από τις παραπάνω βάσεις, παράλληλα γίνεται αξιολόγηση της ποιότητας της πληροφορίας άλλων αποθετηρίων και της χρησιμότητας της πληροφορίας αυτής στην αναπτυσσόμενη μεθοδολογία. Σχεδιάζεται το πως θα γίνει η εξόρυξη τους καθώς και πως θα γίνει η ενοποίηση με τα υπάρχοντα δεδομένα του γράφου. Κάποια από τα επιπλέον αποθετήρια στα οποία θα γίνει εξόρυξη δεδομένων περιέχουν φαρμακογονιδιατικές πληροφορίες. Τα αποθετήρια αυτά είναι η dbSNP (Sherry, Ward, and Sirotkin 1999), η ClinVar (Landrum et al. 2018) και η ENSEMBL (Cunningham et al. 2021). Πιο συγκεκριμένα η dbSNP η οποία παρέχεται από τον NCBI, περιέχει μεταξύ άλλων πληροφορίες που αφορούν μεμονωμένες νουκλεοτιδικές παραλλαγές (SNPs) στους ανθρώπους οι οποίες συνδέονται από σχετική δημοσίευση, την συχνότητα εμφάνισης τους στον πληθυσμό, μοριακές

και γονιδιωματικές επιπτώσεις των παραλλαγών. Τέλος παρέχει πληροφορίες αντιστοίχισης με την βάση RefSeq (O’Leary et al. 2016) τόσο για κοινές παραλλαγές όσο και για κλινικές μεταλλάξεις. Στην προκειμένη περίπτωση η βάση δίνει την δυνατότητα να κατέβουν τα δεδομένα σε μορφή json ωστόσο λόγου του μεγάλου όγκου της πληροφορίας συνιστούν τα δεδομένα να μεταφερθούν στο Hadoop για καλύτερη διαχείριση. Στην συνέχεια εφόσον έχουμε τα δεδομένα τοπικά θα αναπτυχθεί κώδικας σε R/Python ώστε να γίνει η εξόρυξη της χρήσιμης για εμάς πληροφορίας. Αντίστοιχα και η βάση ClinVar παρέχεται από τον NCBI το οποίο παρέχει πληροφορίες για το πως μια γονιδιωματική παραλλαγή επηρεάζει την υγεία του ανθρώπου. Τα δεδομένα μπορούν να κατέβουν πάλι μέσω ftp και είναι είτε σε μορφή txt είτε σε μορφή xml. Στα δεδομένα υπάρχει πεδίο που περιέχει το όνομα της μετάλλαξης κατά dbSNP (rs κωδικοί) και με βάση αυτούς θα γίνει η εξόρυξη των πληροφοριών που μας ενδιαφέρουν αλλά και στην συνέχεια η σύνδεση τους με τις υπόλοιπες οντότητες του γράφου. Τέλος η πλατφόρμα ENSEMBL αποτελεί έναν περιηγητή του γονιδιώματος ο οποίος δίνει την δυνατότητα αναζήτησης γονιδιώματος, μεταγραφώματος και κάποιας παραλλαγής ακολουθίας. Η πληροφορία που περιέχει αφορά κυρίως γονίδια αλλά και τις ασθένειες που σχετίζονται με αυτά. Για την εξαγωγή των δεδομένων από αυτή την πλατφόρμα θα γραφτεί κώδικας σε R καθώς υπάρχουν κατάλληλα πακέτα που προσφέρουν την άμεση και γρήγορη επικοινωνία μεταξύ της R κ της ENSEMBL. Τα δεδομένα από αυτές τις βάσεις θα έρθουν να συμπληρώσουν και να ερμηνεύσουν υπάρχουσες συνδέσεις οντοτήτων του γράφου.

Βιβλιογραφία

1. Anderson, Todd J., Jean Grégoire, Glen J. Pearson, Arden R. Barry, Patrick Couture, Martin Dawes, Gordon A. Francis, et al. 2016. “2016 Canadian Cardiovascular Society Guidelines for the Management of Dyslipidemia for the Prevention of Cardiovascular Disease in the Adult.” *The Canadian Journal of Cardiology* 32 (11): 1263–82.
2. Banerjee, Nilasha, Humphrey Fonge, Andrew Mikhail, Raymond M. Reilly, Reina Bendayan, and Christine Allen. 2013. “Estrone-3-Sulphate, a Potential Novel Ligand for Targeting Breast Cancers.” *PloS One* 8 (5): e64069.
3. Cheng, Feixiong, Weihua Li, Yadi Zhou, Jie Shen, Zengrui Wu, Guixia Liu, Philip W. Lee, and Yun Tang. 2012. “admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties.” *Journal of Chemical Information and Modeling* 52 (11): 3099–3105.
4. “Concepts in Pharmacology and Toxicology.” 2007. In *Cancer Immunotherapy*, 149–66. Academic Press.
5. Cunningham, Fiona, James E. Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Olanrewaju Austine-Orimoloye, et al. 2021. “Ensembl 2022.” *Nucleic Acids Research*, November. <https://doi.org/10.1093/nar/gkab1049>.
6. Danaei, Goodarz, Mariel M. Finucane, Yuan Lu, Gitanjali M. Singh, Melanie J. Cowan, Christopher J. Paciorek, John K. Lin, et al. 2011. “National, Regional, and Global Trends in Fasting Plasma Glucose and Diabetes Prevalence since 1980: Systematic Analysis of Health Examination Surveys and Epidemiological Studies with 370 Country-Years and 2·7 Million Participants.” *The Lancet* 378 (9785): 31–40.
7. DeVito, Nicholas J., Georgia C. Richards, and Peter Inglesby. 2020. “How We Learnt to Stop Worrying and Love Web Scraping.” *Nature* 585 (7826): 621–22.
8. Djoumbou-Feunang, Yannick, Jarlei Fiamoncini, Alberto Gil-de-la-Fuente, Russell

- Greiner, Claudine Manach, and David S. Wishart. 2019. "BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification." *Journal of Cheminformatics* 11 (1): 2.
9. Forsythe, Peter, and Sue Paterson. 2014. "Ciclosporin 10 Years on: Indications and Efficacy." *The Veterinary Record* 174 Suppl 2 (March): 13–21.
 10. "IPC Publication." n.d. Accessed December 16, 2021.
<https://ipcpub.wipo.int/?notion=scheme&version=20210101&symbol=A61K0048000000&menulang=en&lang=en&viewmode=f&fipcp=no&showdeleted=yes&indexes=no&headings=yes¬es=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart>
 11. Koepsell, Hermann. 2020. "Organic Cation Transporters in Health and Disease." *Pharmacological Reviews* 72 (1): 253–319.
 12. Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research* 46 (D1): D1062–67.
 13. Lin, Lawrence, Sook Wah Yee, Richard B. Kim, and Kathleen M. Giacomini. 2015. "SLC Transporters as Therapeutic Targets: Emerging Opportunities." *Nature Reviews. Drug Discovery* 14 (8): 543.
 14. MacLeod-Glover, Nora, Matthew Mink, Mark Yarema, and Ryan Chuang. 2016. "Digoxin Toxicity: Case for Retiring Its Use in Elderly Patients?" *Canadian Family Physician Medecin de Famille Canadien* 62 (3): 223.
 15. March-Vila, Eric, Luca Pinzi, Noé Sturm, Annachiara Tinivella, Ola Engkvist, Hongming Chen, and Giulio Rastelli. 2017. "On the Integration of In Silico Drug Design Methods for Drug Repurposing." *Frontiers in Pharmacology* 0.
<https://doi.org/10.3389/fphar.2017.00298>.
 16. Morrissey, K. M., C. C. Wen, S. J. Johns, L. Zhang, S-M Huang, and K. M. Giacomini. 2012. "The UCSF-FDA TransPortal: A Public Drug Transporter Database." *Clinical Pharmacology and Therapeutics* 92 (5): 545–46.
 17. O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1): D733–45.
 18. Papadatos, George, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, et al. 2016. "SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database." *Nucleic Acids Research* 44 (D1): D1220–28.
 19. "Probenecid." 2007, January, 1–5.
 20. Pushpakom, Sudeep, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, et al. 2018. "Drug Repurposing: Progress, Challenges and Recommendations." *Nature Reviews. Drug Discovery* 18 (1): 41–58.
 21. Riedel, Annett, Christina Maria Hochkogler, Roman Lang, Gerhard Bytof, Ingo Lantz, Thomas Hofmann, and Veronika Somoza. 2014. "N-Methylpyridinium, a Degradation Product of Trigonelline upon Coffee Roasting, Stimulates Respiratory Activity and Promotes Glucose Utilization in HepG2 Cells." *Food & Function* 5 (3): 454–62.
 22. Roth, Megan, Amanda Obaidat, and Bruno Hagenbuch. 2012. "OATPs, OATs and OCTs: The Organic Anion and Cation Transporters of the SLCO and SLC22A Gene Superfamilies." *British Journal of Pharmacology* 165 (5): 1260–87.
 23. Sherry, S. T., M. Ward, and K. Sirotkin. 1999. "dbSNP-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation." *Genome Research* 9 (8): 677–79.
 24. Shitara, Yoshihisa. 2011. "Clinical Importance of OATP1B1 and OATP1B3 in Drug-Drug Interactions." *Drug Metabolism and Pharmacokinetics* 26 (3): 220–27.
 25. Somogyi, A., and R. Gugler. 1983. "Clinical Pharmacokinetics of Cimetidine." *Clinical*

Pharmacokinetics 8 (6): 463–95.

26. Srimaroeng, Chutima, Varanuj Chatsudthipong, Amy G. Aslamkhan, and John B. Pritchard. 2005. "Transport of the Natural Sweetener Stevioside and Its Aglycone Steviol by Human Organic Anion Transporter (hOAT1; SLC22A6) and hOAT3 (SLC22A8)." *The Journal of Pharmacology and Experimental Therapeutics* 313 (2): 621–28.
27. Stein, Stephanie Aleskow, Elizabeth Mary Lamos, and Stephen N. Davis. 2013. "A Review of the Efficacy and Safety of Oral Antidiabetic Drugs." *Expert Opinion on Drug Safety*, February. <https://doi.org/10.1517/14740338.2013.752813>.
28. "Tetraethylammonium." 2007, January, 1–3.
29. The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, et al. 2020. "UniProt: The Universal Protein Knowledgebase in 2021." *Nucleic Acids Research* 49 (D1): D480–89.
30. Türei, Dénes, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, et al. 2021. "Integrated Intra- and Intercellular Signaling Knowledge for Multicellular Omics Analysis." *Molecular Systems Biology* 17 (3): e9923.
31. Viollet, Benoit, Bruno Guigas, Nieves Sanz Garcia, Jocelyne Leclerc, Marc Foretz, and Fabrizio Andreelli. 2012. "Cellular and Molecular Mechanisms of Metformin: An Overview." *Clinical Science* 122 (6): 253–70.
32. "Website." n.d. <https://neo4j.com/>.
33. Whirl-Carrillo, Michelle, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F. Thorn, Ryan Whaley, and Teri E. Klein. 2021. "An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine." *Clinical Pharmacology and Therapeutics* 110 (3): 563–72.
34. Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A Major Update to the DrugBank Database for 2018." *Nucleic Acids Research* 46 (D1): D1074–82.
35. Wrighton, S. A., M. VandenBranden, and B. J. Ring. 1996. "The Human Drug Metabolizing Cytochromes P450." *Journal of Pharmacokinetics and Biopharmaceutics* 24 (5): 461–73.