



**Pázmány Péter
Katolikus Egyetem
Információs Technológiai és
Bionikai Kar**



**MTA-SE Molekuláris
Biofizikai Kutatócsoport**

XXXII. OTDK Biológia Szekció

Multidrog-transzporterek, metabolikus enzimek és promiszkuis nukleáris receptorok drogfelismerésének összehasonlítása

Készítő: Harmat Zita

Képzés: Infobionika MSc

Témavezetők: Dr. Hegedűs Tamás, Dr. Gáspári Zoltán

Küldő egyetem: Pázmány Péter Katolikus Egyetem

Munkavégzés helye: Semmelweis Egyetem Molekuláris Biofizikai Kutatócsoport

Tartalomjegyzék

Tartalomjegyzék	2
1. Bevezetés	4
1.1. Rövidítésjegyzék.....	4
1.2. A feladat ismertetése, célkitűzések	4
1.3. Biológiai háttér: A kemoimmunitási rendszer	6
1.3.1. Nukleáris receptorok	7
1.3.2. Az első fázis enzimeit.....	8
1.3.3. A második fázis enzimeit	9
1.3.4. A harmadik fázis fehérjéi: az ABC transzporterek	9
2. Módszerek	11
2.1. A kutatás során felhasznált eszközök	11
2.1.1. RDBMS, SQLAlchemy, ORM.....	11
2.1.2. Publikus adatbázisok	12
2.2. Eljárások	13
2.2.1. Lekérdezés a PubChem BioAssay adatbázisból.....	13
2.2.2. Lekérdezés a Comparative Toxicogenomics (CTD) adatbázisból.....	13
2.2.3. Az adatok rendezése.....	14
2.2.4. Klaszterezés.....	15
2.2.5. Főkomponens analízis (PCA).....	15
3. Eredmények	17
3.1. Kemoimmun fehérjék ligandjainak és szubsztrátjainak feltérképezése publikus adatbázisok segítségével.....	17
3.1.1. PubChem BioAssay adatbázis.....	17
3.1.2. Metabolism and Transport Database (Metabase)	18
3.1.3. Comparative Toxicogenomics Database (CTD)	19

3.1.4. DrugBank adatbázis	19
3.1.5. Irodalmi adatok	20
3.1.6. A publikus adatbázisok összehasonlítása.....	20
3.1.7. A ligand- és szubsztrát adatok egyesítése: az általunk létrehozott DrugDB adatbázis.....	21
3.2. A Comparative Toxicogenomics Database-ből származó szubsztrát- és ligand adatok elemzése	24
3.2.1. A rendelkezésre álló adatok elemzése.....	24
3.2.2. A Comparative Toxicogenimocs Database-ből származó adatok jellemzése	31
4. Összefoglalás	35
5. Irodalomjegyzék	36
6. Köszönetnyilvánítás.....	38

1. Bevezetés

1.1. Rövidítésjegyzék

ABC: Adenosine Triphosphate Binding Cassette; ATP Binding Cassette
AHR: Aryl Hydrocarbon Receptor
API: Application Programming Interface
ATP: Adenosine Triphosphate
BLAST: Basic Local Alignment Search Tool
CAR: Konstitutív Androsztán Receptor (NR1I3)
CCRP: Cytoplasmic CAR Retention Protein
CID: PubChem Chemical ID
CTD: Comparative Toxicogenomics Database
CYP: Citokróm P450
EC50: Maximális hatékony koncentráció fele
ESR1: Ösztrogén Receptor
ESR2: Béta ösztrogén receptor
GSH: Glutation
GST: Glutation S-Transzferáz
IC50: Half Maximal Inhibitory Concentration
LC50: Az a koncentráció, ami a kísérletek felében halálos
MDR: Multidrog-Rezisztencia
Metabase: Metabolism and Transport Database
MRP1: Multidrug Resistance-Associated Protein 1
NBD: Nukleotide Binding Domain
NR: Nukleáris Receptor
ORM: Object Relational Mapper
PCA: Főkomponens analízis
PP2A: Protein Phosphatase 2A
PPARA: Peroxisome proliferator-activated receptor alpha
PPARD: Peroxisome proliferator-activated receptor delta
PPARG: Peroxisome proliferator-activated receptor gamma
PXR: Pregnán X Receptor (NR1I2)
RDBMS: Relational Database Management System
RXRA: Retinoic acid receptor RXR-alpha
SID: PubChem Substance ID
SLC: Solute Carrier
SQL: Structured Query Language
TMD: Transzmembrán Domén
VDR: Vitamin D3 Receptor (NR1I1)
VDRE: Vitamin D Response Elements

1.2. A feladat ismertetése, célkitűzések

A sejt védekezik az idegen anyagok ellen, amelyek magukba foglalnak nem csak gyógyszereket, hanem mindenféle egyéb környezeti, légzéssel vagy táplálkozással bejutó anyagokat is. A védekezésben különböző fehérjék, összefoglaló néven kemoimmun fehérjék vesznek részt: nukleáris receptorok, melyek felismerik az idegen anyagokat, metabolikus

enzimek, melyek átalakítják ezeket a sejtidegen vegyületeket, és multidrog transzporterek, melyek kipumpálják a sejtől az idegennek számító vegyületeket, illetve azok metabolitjait. Ezen fehérjék fontos szerepet játszanak gyógyszerek metabolizmusában és a rákos sejtek kemoterápiával szemben kialakuló rezisztenciájában (multidrog-rezisztencia, MDR) is. Egy fehérje nem szükségszerűen csak egy anyagot ismer fel, hanem felismerhet nagyon sok anyagot is. Ezeket a fehérjéket nevezzük multidrog-kötő fehérjéknek, és a kutatásunk során vizsgált fehérjék nagy része ilyen.

A kutatás során azt a kérdést vizsgáljuk, hogy vajon ezek a fehérjék egy hálózatban együttműködve végzik-e feladatukat. Erre több tény is utal. Például, amikor a kemoterápiával szembeni rezisztenciát célzottan egy-egy fehérje módosításával próbálták felfüggeszteni, nem jártak eredménnyel, azaz a védekezés redundáns [1]. E mellett az irodalomból ismert tény, hogy a nukleáris receptorok szabályozzák az ABC transzporterek [2] és a Citokróm P450 (CYP) oxigenázok [3] működését, és ha xenobiotikum kötődik a nukleáris receptorokhoz, akkor azok erősíthetik azon célfehérjék expresszióját, amelyek a xenobiotikum metabolizmusában és sejtől való eltávolításában vesznek részt [4-5].

Azokat a fehérjéket szeretnénk azonosítani, amik együtt működnek egy vegyület vagy vegyületcsoport elleni védelemben azzal, hogy ugyanazt a vegyületet vagy vegyületcsoportot ismerik fel.

1. Először arra a kérdésre szeretnénk válaszolni, hogy milyen a nukleáris receptorok ligand átfedése, azaz mely vegyületek ligandjai két vagy több adott nukleáris receptornak. Ehhez nukleáris receptorok ligandjait és a ligandokat tartalmazó kísérletek adatait kell rendszerezni.

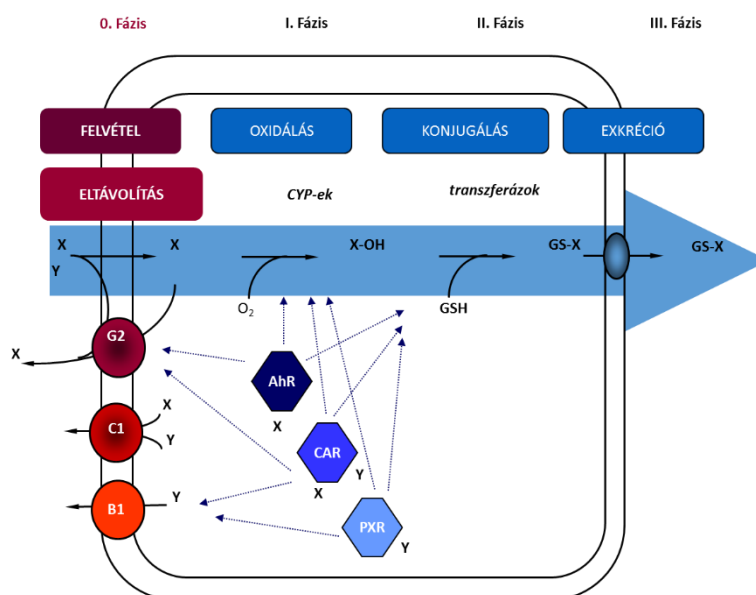
2. Azt is vizsgáljuk, hogy melyek azok az enzimek és transzporterek, amelyeknek ugyanazok a vegyületek a szubsztrátjaik. Ehhez meg kell ismernünk a metabolikus enzimek, illetve multidrog transzporterek szubsztrátjait.

3. A kutatás során publikus adatbázisokból elérhető adatok segítségével szeretnénk arra választ kapni, hogy azok a vegyületek, amelyek ugyanazokat a nukleáris receptorokat aktiválják, azok ugyanazoknak az enzimeknek és transzportereknek a szubsztrátjai-e.

Ha találunk olyan nukleáris receptorokat, amelyek nagyszámú vegyületet felismernek, és találunk olyan enzimeket és transzportereket, amelyeknek szubsztrátjai ugyanebbe a vegyülethalmazba tartoznak, akkor megtaláltuk azokat a fehérjéket, amelyekről feltételezhetjük, hogy együttműködnek ezen idegen anyagok elleni védekezésben.

1.3. Biológiai háttér: A kemoimmunitási rendszer

Ismert tény, hogy a sejtbe bejutó idegen molekulák (xenobiotikumok) a sejtben kötődnek bizonyos fehérjékhez, metabolizálódnak, transzporterek juttatják ki őket vagy épp metabolizált formájukat a sejtből.



1.1. ábra: A sejt idegen anyagok elleni védekezésének fázisai és a bennük részt vevő fehérjék. CYP-ek: citokróm P450 monooxygenázok, G2: ABCG2, C1: ABCC1, B1: ABCB1. Az ábra Power Pointtal készült.

Az idegen anyagok elleni védekezésnek négy fázisát különböztetjük meg (1.1. ábra). A sejtbe diffúzióval vagy SLC (Solute Carrier) fehérjék által jutnak be az idegen anyagok, melyek általában kisméretű, hidrofób molekulák. Ezeket a nulladik fázisban ABC (ATP Binding Cassette) fehérjék pumpálják ki a sejtől. A védekezés első fázisában különféle enzimek oxidálják az idegen anyagokat, például a citokróm P450 monooxygenázok, melyek a legfontosabbak ezen redox enzimek közül, de e feladatot végzik például az alkohol dehidrogenázok is. A második fázisban az oxidált xenobiotikumokat többek között a glutation transzferázok konjugálják, vagyis funkciós csoportokat kapcsolnak ezekhez a vegyületekhez. E folyamat során nem vagy kevésbé toxikus, vízzoldékony vegyületek keletkeznek. A harmadik fázis során transzporterek, például az MRP1 fehérje (Multidrug resistance-associated protein 1, ABCC1) pumpálják ki a sejtől a keletkezett metabolitokat [2-3].

A sejtbe való bejutáskor az idegen molekulák promiskuis (többféle vegyületet felismerő) nukleáris receptorokhoz kötődnek, melyek transzkripciós faktorok, ezért erősítik azon célfehérjék expresszióját, amelyek a xenobiotikum metabolizmusában és sejtől való

eltávolításában vesznek részt [4-5]. Azaz a nukleáris receptorok végzik a védekezés szabályozását.

1.3.1. Nukleáris receptorok

A nukleáris receptorok ligand vezérelt transzkripciós faktorok. Az amino terminálisukon van egy A/B doménjük, amelynek része az aktivációs funkció rész, amely ligand nélkül is mindig aktív. Van egy DNS kötő doménjük, és egy csuklóspánt (hinge) régiójuk, mely összeköti a DNS kötő domént a ligand kötő doménnel. A nukleáris receptoroknak hét alcsaládja van (0-tól 6-ig számozva), amelyeket filogenetikai elemzéssel alakítottak ki. Vannak olyan nukleáris receptorok, amelyeknek meghatározott belső, fiziológias ligandjaik vannak, például az NR1H2-nek (LXR β), az NR1H3-nak (LXR α) és az NR1H4-nek (LXR α) a 24(S),25-epoxikoleszterol a májban, a 24(S)-hidroxikoleszterol az agyban, és a 27-hidroxikoleszterol a makrofágokban, de vannak olyanok is, amelyeknek nincs belső fiziológias ligandjuk, ilyenek például az NR4A1 és az NR4A2. E mellett vannak olyan nukleáris receptorok is, amelyeknek több ligandja van, például az NR1C1 (PPAR α), az NR1C2 (PPAR β/δ) és az NR1C3 (PPAR γ), amelyek zsírsavakat ismernek fel. Vannak olyan nukleáris receptorok is, amelyeknek nincs belső ligandja, de kapcsolatba hozhatók a multidrog-rezisztenciával, vagyis az idegen anyagok elleni védelemmel, mivel azokat ismerik fel [6].

A főbb promiszkuis nukleáris receptorok: a CAR és a PXR, és a kontrollként használt VDR

A Konstitutív Androsztán Receptor (CAR, NR1I3), a Pregnán X Receptor (PXR, ; NR1I2) és a D3 vitamin receptor (VDR, NR1I1) nukleáris receptorok strukturális jellemzőik miatt egy csoportot alkotnak az NR1 alcsaládon belül [7]. A CAR és a PXR ligandjai toxikus metabolikus melléktermékek., Fiziológias ligandjuk azonban nem ismert, ezért ezt a két fehérjét árva receptornak is nevezik [4]. A VDR-nek fiziológias ligandja a D vitamin. Kevés fajta vegyületet ismer fel, ezért választottuk kontrollnak az igen sok vegyületet felismerő CAR és PXR mellé részletes vizsgálatunkban.

Konstitutív Androsztán Receptor (CAR)

A konstitutív androsztán receptor nevét az androsztánokról kapta, melyek *in vitro* gátolják az eleve aktív állapotát. Inaktivált állapotban a CAR a citoplazmában tartózkodik a Hsp90 hősokk fehérjével és a CCRP (cytoplasmic CAR Retention Protein) fehérjével komplexet alkotva. Amikor a PP2A (protein phosphatase 2A) defoszforilálja a CAR ligandkötő régiójának egyik aminosavát, akkor a nukleáris receptor be tud jutni a sejtmagba, ahol az RXR fehérjével képez komplexet. Ez az indirekt aktiválás, ami ligand hiányában is

megtörténhet. A CAR-RXR komplex a DNS bizonyos régióihoz tud kapcsolódni, aktiválva az idegen anyagok metabolizmusában részt vevő fehérjék transzkripcióját [7].

A CAR-nak jóval több ismert ligandja van, mint a VDR-nek, de kevesebb, mint a PXR-nak [8].

Pregnán X Receptor (PXR)

A Pregnán X receptor nevét szteroid ligandjairól kapta, melyek hatékony aktiválói a fehérjének. A PXR alapesetben a sejtmagban tartózkodik, és inaktív formája represszor fehérjékkel alkot komplexet, így gátolja bizonyos gének transzkripcióját. Ha egy vegyület aktiválja, akkor disszociálódik a represszoroktól, és az RXR alfa fehérjével dimerizálódva koaktivátor fehérjét toboroz, így erősítve célgénjeinek transzkripcióját [7].

A három fehérje közül a PXR a legkevésbé specifikus, amely összhangban van azzal a ténnyel, hogy nagy és flexibilis a ligand-kötő zsebe. Az emberi PXR fehérje ligandjai közé tartozik sok gyógyszer, növényi hatóanyag, környezetszennyező anyag, szteroid és epesav [8].

D3 vitamin receptor (VDR)

A D3 vitamin receptor nagy affinitással köti a D3 vitamint, és a CAR-hoz és a PXR-hez hasonlóan ligand által aktivált transzkripciós faktorként működik. A CAR-hoz hasonlóan ligand kötéskor heterodimerizálódik az RXR-rel, majd a DNS speciális régióihoz, úgynevezett D vitamin válasz elemekhez (Vitamin D response elements, VDRE) kötődik. A VDR válasz elemek, a VDR és egyéb transzkripciós faktorok a transzkripciós preiniciációs komplexet alkotják, amely erősítheti vagy gátolhatja a VDR által kezdeményezett transzkripciót [9].

A két előző fehérjéhez képest viszonylag szűk a szubsztrátspektruma. Ligandjai főképp a calcitriol (más néven 1,25 dihydroxyvitamin D3), és annak szintetikus analógjai [8].

1.3.2. Az első fázis enzimeai

A metabolikus Fázis I folyamatokban olyan enzimek vesznek részt, amelyek a bejutott idegen anyagot oxidálják, redukálják, vagy hidrolizálják, hogy hidroxil vagy karboxil csoportok kerüljenek rá, és így az idegen anyag átalakított formája a második fázisába kerülhessen. Ide tartoznak például az FMO (Flavin-containing monooxygenase), az aldehid dehidrogenáz enzimek (ALDH1A1, ALDH3A, ALAS1) és a karboxilészterázok (CES2, CES3, ES1) [3]. Közülük azonban a legfontosabbak a CYP oxidázok.

CYP oxidázok

A citokróm P450 (CYP) enzimek a monooxygenázok családjának tagjai [10]. A CYP enzimek többféle kémiai reakciót katalizálnak, például a különféle vegyületek oxidatív, metabolizmusát [3]. Eddig 57 humán CYP-ről tudunk, melyek szekvencia hasonlóság alapján 18 fehérjecsaládba és 43 alcshaládba tartoznak [10]. Katalitikus központjukban egy hem csoport található. A szubsztrát a hem csoport felszíne felett kapcsolódik fehérjéhez. Mindig a fehérje strukturális magja a legkonzerváltabb, és a szubsztrát-kötő zseb a legkevésbé konzervált [11].

1.3.3. A második fázis enzime

A Fázis II enzimek olyan fehérjék, amelyek acetilálják, szulfatálják, aminosavakkal konjugálják, glükuronidálják a sejtbe bejutó anyagok metabolitjait, hogy a keletkezett, immár vízben oldható vegyületek kiürülhessenek a sejtéből, majd a májon vagy a vesén keresztül a szervezetből. Ezen enzimek közé tartoznak például az N-acetil transzferázok (NAT1, NAT2), a tiopurin-S-metiltranszferáz, a katechol-O-metiltranszferáz, a szulfotranszferázok (SULT2A1, SULTN, PAPSS2), az acetil-koenzim A szintetáz, és a glükuronozil transzferázok (UDP-GT1A, UDP-GT2B) [3]. Közülük legfontosabbak a glutation transzferázok.

A glutation transzferázok

A glutation S-transzferázok (GST-k) családjának tagjai a legtöbb élőlényben megtalálhatóak. Feladatuk, hogy glutationt (GSH) kössenek (konjugáljanak) hidrofób szubsztrátjaikkal, így vízoldékonyabbá tegyék őket. A szolubilis GST-k dimerek. Mindkét félnek van egy aktív helye, mely két részből áll, egy hidrofíl G részből, ahol a GSH-t köti, és egy szomszédos H részből, ahol egy hidrofób környezet alakul ki a strukturálisan különböző elektrofil szubsztrátok számára. A G rész igen konzervált, a H rész viszont nem, így eltérő enzimek igen eltérő ligandokat ismernek fel. A GST-k segítenek többek között az epoxid karcinogének és a lipid peroxidok méregtelenítésében [12].

1.3.4. A harmadik fázis fehérjéi: az ABC transzporterek

Az ABC (Adenosine triphosphate Binding Cassette) transzporterek nagyméretű, sejtmembránba ágyazott fehérjék, melyekre jellemző, hogy rendelkeznek transzmembrán (TMD) és nukleotid-kötő (NBD) doménnel. A fehérjecsalád tagjai az eukariótáktól kezdve az emlősökig minden élőlényben jelen vannak. Emberekben sokféle feladatuk van. Részt vesznek például az immunrendszer működésében, epesavak kiválasztásában, stb. A három legfontosabb ABC multidrog-reszisztenciában (MDR) részt vevő transzporter: az ABCB1 (más néven MDR1 vagy P-glycoprotein), az ABCC1 (MRP1), illetve az ABCG2 (MXR, BCRP).

Ezek a multidrog-rezisztenciában szerepet játszó fehérjék is szubsztrátként igen eltérő vegyületek csoportjait ismerik fel [2]. Jelen vannak szöveti gátakban: például a bélben megakadályozzák karcinogének, mérgező anyagok és egyéb sejtidegen vegyületek felszívódását, a vér-agy-gátban megakadályozzák a rovarirtószerek bejutását az agyba, a placentában pedig megakadályozzák az anyában lévő méreganyagok átjutását a magzatba.

2. Módszerek

2.1. A kutatás során felhasznált eszközök

2.1.1. RDBMS, SQLAlchemy, ORM

A kemoimmun fehérjék ligand és szubsztrát adatait MySQL 5.5. verziójú (<http://dev.mysql.com/doc/refman/5.5/en/what-is-mysql.html>) RDBMS (Relational Database Management System) adatbázis rendszerbe helyeztük, mert így könnyen elérhetőek, könnyebben kezelhetőek, elemezhetőek és nem redundánsan tárolhatóak.

Az adatok további elemzését a Python programnyelv segítségével végeztük, ezért kézenfekvő volt, hogy az adatbázist is a Python SQLAlchemy [13] nevű modulja segítségével érjük el. Ez a modul Python szintaktikából generál SQL parancsokat, és ezt adja tovább az adatbázis kezelő rendszernek. A modul egy objektum-relációs megfeleltetést (Object Relational Mapping, ORM) végez, ami azt jelenti, hogy az adatbázis táblái megfeleltethetőek Python objektumoknak, és az objektumok attribútumai a táblázat oszlopai, azaz a modul képes objektumorientáltan elérni az adatbázist, ami hatékonyabb kódolást tesz lehetővé. A táblák adatokkal való feltöltése az objektumok példányainak létrehozásával történik, lekérdezéseket pedig az objektumok attribútumain keresztül végezhetünk.

A 2.1. ábrán az SQL (Structured Query Language) szintaktika és a Python SQLAlchemy modul szintaktikájának összehasonlítását látjuk. Az utóbbi jóval átláthatóbb, objektumorientált, így hatékonyabban használható.

A) <pre>CREATE TABLE `protein` (`protein_id` int(11) NOT NULL AUTO_INCREMENT, `gene_name` varchar(300) DEFAULT NULL, `acc` varchar(300) DEFAULT NULL, `entry_name` varchar(300) DEFAULT NULL, `organism` varchar(300) DEFAULT NULL, PRIMARY KEY (`protein_id`), UNIQUE KEY `acc` (`acc`)) ENGINE=InnoDB AUTO_INCREMENT=12 DEFAULT CHARSET=latin1\$</pre>	B) <pre>class Protein(Base): __tablename__ = "protein_binding" protein_id = Column(Integer, primary_key=True, AutoIncrement=True) ; gene_name = Column(String(300)) acc = Column(String(300), unique=True) entry_name = Column(String(300)) organism = Column(String(300)) # relationships bindings = relationship("ProteinBinding", backref="protein")</pre>
C) <pre>SELECT `protein`.`protein_id`, `protein`.`gene_name`, `protein`.`acc`, `protein`.`entry_name`, `protein`.`organism` FROM `drugdb`.`protein` WHERE `protein`.`acc`='075469';</pre>	D) <pre>results=DBSession.query(Protein).filter(Protein.acc=='075469').all() for protein in results: print protein</pre>

2.1. ábra: az SQL és a Python SQLAlchemy modul szintaktikájának összehasonlítása

A) Egy tábla létrehozása SQL nyelvben; B) Egy tábla létrehozása SQLAlchemyvel;

C) Lekérdezés SQL nyelvben; D) Lekérdezés SQLAlchemyvel

2.1.2. Publikus adatbázisok

Vizsgálatunkhoz sok kemoimmun fehérje szubsztrát, illetve ligand adat szükséges, amelyet mások már kigyűjtöttek az irodalomból és adatbázisokba szerveztek.

Adatbázis	Tartalom, leírás	link	verzió	Hivatkozás
PubChem BioAssay	A PubChem adatbázis része. Vegyületek fehérjékre gyakorolt hatását vizsgáló kísérleteket gyűjt össze.	http://www.ncbi.nlm.nih.gov/pcassay/	2014. július	Y. Wang, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B. A. Shoemaker, A. Gindulyte, and S. H. Bryant, "PubChem BioAssay: 2014 update," Nucleic Acids Res., vol. 42, no. Database issue, pp. D1075–1082, Jan. 2014.
Metabolism and Transport Database (Metrabase)	Tartalmaz 33 transzporterhez és metabolikus enzimhez kapcsolódó összesen 3438 vegyületet, azonban nem tartalmaz nukleáris receptorokat.	http://www-metrabase.ch.cam.ac.uk/	1.0	Metrabase Development Team, University of Cambridge, Metrabase The Metabolism and Transport Database user manual v2014.01. 2014.
Comparative Toxicogenomics Database (CTD)	Vegyületek, gének és betegségek közötti kapcsolatokat és génexpressziós tartalmaz adott szintaktikájú mondatok formájában.	http://ctdbase.org/	2014. szept. 14.	A. P. Davis, C. G. Murphy, R. Johnson, J. M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, M. C. Rosenstein, T. C. Wiegiers, and C. J. Mattingly, "The Comparative Toxicogenomics Database: update 2013," Nucleic Acids Res., vol. 41, no. Database issue, pp. D1104–1114, Jan. 2013.
DrugBank	Gyógyszereket és gyógyszerjelölteket foglal magába, sok gyógyszer célponttal (fehérjével) együtt, egyesítve a gyógyszerek és célpontjaik adatait.	http://www.drugbank.ca/	4.0	V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: shedding new light on drug metabolism," Nucleic Acids Res., vol. 42, no. Database issue, pp. D1091–1097, Jan. 2014.

2.1. táblázat: Publikus adatbázisok

2.2. Eljárások

2.2.1. Lekérdezés a PubChem BioAssay adatbázisból

A PubChem BioAssay adatbázisból a PubChem Bioactivity Data Dicer segítségével olyan kísérletek adatait kérdeztük le, amelyekben szerepel a 48 emberi nukleáris receptor. A lekérdezést úgy hajtottuk végre, hogy a génnév mezőbe 'OR'-al elválasztva beadtuk a 48 génnevet. De mivel a rendszer nem enged letölteni 1 millió rekordnál többet, ezért kikötöttük, hogy a kísérlet kimenetele aktív legyen, azaz legyen kölcsönhatás a fehérje és a vegyület között. Így sikerült lecsökkentenünk a kapott eredmények számát, pillanatnyi elemzéseinkhez pedig ugyanis csak az aktív kimenetelt produkáló vegyületekre volt szükségünk.

A PubChem BioAssay adatbázisból a Metrabase-ben szereplő transzporterek és metabolikus enzimek génneveit OR-al elválasztva is kérdeztünk le kísérleti adatokat, itt is az aktív kimenetelű kísérletekre szűkítve a keresést.

2.2.2. Lekérdezés a Comparative Toxicogenomics (CTD) adatbázisból

A CTD adatbázisban a vegyületek és gének, a vegyületek és betegségek, illetve a gének és betegségek közötti kapcsolatok jól meghatározott szintaktikával vannak definiálva, például „ABCB1 protein results in increased transport of 10,11-dihydro-10-hydroxycarbamazepine”, ezért a fehérjék génjei és a vegyületek közötti kapcsolatokban a lokálisan letöltött és MySQL RDBMS-be rendezett adatbázisban speciális mondatokkal kerestünk, melyekbe mindig behelyettesítettük az adott gén nevét. A mondatok vázai 2.2. táblázatban szerepelnek.

Fehérjetípus	Keresőkifejezés	Szubsztrátok száma
Exporterek és enzimek	<génnev> protein results in increased secretion of <vegyületnev>	260
	<génnev> protein results in increased transport of <vegyületnev>	
	<génnev> protein results in increased export of <vegyületnev>	
	<génnev> protein results in decreased susceptibility to <vegyületnev>	
	<génnev> results in increased activity of <génnev> protein	
Nukleáris receptorok és transzkripciós faktorok	<génnev> binds to and results in increased activity of <génnev> protein	690
Importerek	<génnev> protein results in increased uptake of <vegyületnev>	371
	<génnev> protein results in increased import of <vegyületnev>	
	<génnev> protein results in increased susceptibility to <vegyületnev>	
	<génnev> protein results in increased transport of <vegyületnev>	
	<génnev> protein affects the uptake of <vegyületnev>	
	<génnev> protein results in increased uptake of and results in increased susceptibility to <vegyületnev>	
	<génnev> protein affects the transport of <vegyületnev>	
	<génnev> protein results in increased uptake of and results in increased susceptibility to <vegyületnev>	

2.2. táblázat: Keresőkifejezések a Comparative Toxicogenomics Database lekérdezésekhez

Az adatbázisban lévő vegyületeket úgy azonosítottuk, hogy a PubChem-ből letöltöttük az összes létező olyan anyagot (substance), ami a CTD-ből származik, így megtudtuk azok SID-jeit, azon keresztül pedig a CID-jüket (Chemical ID, vegyület azonosító) tudtuk meg.

2.2.3. Az adatok rendezése

Az adatbázisból lekérdezett ligand és szubsztrát adatok statisztikai elemzéséhez az adatokat össze kellett gyűjteni, és elrendezni. Ennek során az adatokat egy olyan bináris mátrixba helyeztük, amelynek oszlopai a fehérjék, sorai a vegyületek. A táblázat adott fehérjéhez és vegyülethez tartozó cellájában 1 van, ha a vegyület szubsztrátja vagy ligandja az adott fehérjének, és 0, ha nincs adatunk arról, hogy az adott vegyület kölcsönhat-e az adott

fehérjével, vagy ha az adott vegyület biztosan nem szubsztrátja/ligandja az adott fehérjének. A táblázat sorai vagy oszlopai felfoghatók vektorok halmazának, melyeket többféle módon elemeztünk. A CTD adatbázisban csak a vegyület és a fehérje közötti kapcsolat minősége szerepel, számszerű adatok nélkül, ezért reprezentáljuk binárisan azt, hogy az adott vegyület szubsztrátja, ligandja-e a fehérjének vagy sem.

2.2.4. Klaszterezés

A kemoimmun fehérjéket és a hozzájuk kapcsolódó vegyületeket a Python programnyelv numpy csomagja, illetve a programnyelv scipy csomagjának cluster alcsomagján belül a hierarchy alcsomag segítségével klasztereztük a Ward módszert használva. Vegyületeknél a sorok közötti, fehérjéknél az oszlopok közötti távolságot korrelációként (Pearson) definiáltjuk, s ez alapján végeztük a klaszterezést. A vektorok közötti korrelációt a Python programnyelv pandas nevű csomagjával számoltuk. Az eredményül kapott dendrogramot, a Python programnyelv matplotlib csomagjával ábráztuk.

A vegyületeket hierarchikusan klasztereztük, amivel egy fát kaptunk. Ezt a fát kellett úgy ágaira darabolnunk, hogy csoportokat kapjunk. Nem mindegy azonban, hogy milyen küszöbértéknél vágjuk el a fát. Ezt az értéket, azaz a klaszterek optimális számát úgy határoztuk meg, hogy vizsgáltuk a szórás a csoportszám függvényében. Ha kevés klaszterünk van, akkor egy-egy klaszteren belül a szórás nagy, hiszen sokkal inkább különbözőbb vegyületek kerülnek egy klaszterbe, mint abban az esetben, ha több a klaszterünk. Azonban lesz egy adott számú klaszter, amelynél ha tovább növeljük a klaszterszámot, a szórás már alig csökken. Ha grafikonon ábrázoljuk a szórás a klaszterszám függvényében, akkor az optimális klaszterszámnál található pont, az úgynevezett könyökpont jól látszik. A könyökpont csoportszáma az optimális csoportszám.

2.2.5. Főkomponens analízis (PCA)

Ezzel a módszerrel a vegyületeket elemeztük aszerint, hogy mely fehérjéhez kötődnek. Minden vegyülethez tartozó olyan adatok, melyek leírják, hogy mely fehérjéhez kapcsolódik, egy vektort alkotnak, amely annyi dimenziós, ahány fehérje van. Azaz minden vektor egy pont a fehérjék által meghatározott sokdimenziós térben. A főkomponens analízis a fehérjeváltozók teréből átviszi ezeket a vektorokat egy kisebb dimenziós térbe, ahol a fehérjék helyett az új változók az úgynevezett főkomponensek, amelyeknél a legelső mentén van az adatoknak a legnagyobb varianciája, a másodikonál a második legnagyobb variancia oly módon, hogy emellett a főkomponensek még merőlegesek is legyenek egymásra.

Ezeknek a vektoroknak, amelyek megmondják, hogy mely vegyület mely fehérjéhez kapcsolódik, a pontjait levetítve az első és a második komponens által alkotott síkra, a pontok a lehető legjobban szóródnak, ezért itt láthatóak a legjobban, hogy mely pontok vannak közel egymáshoz, azaz melyek tartoznak egy csoportba.

A főkomponens analízist a Python programnyelv matplotlib csomagjának mlab alcsomagjával végeztük, és az eredményt a matplotlib csomag pyplot alcsomagjával jelenítettük meg.

3. Eredmények

3.1. Kemoimmun fehérjék ligandjainak és szubsztrátjainak feltérképezése publikus adatbázisok segítségével

A kutatásunk arra irányul, hogy megtudjuk, van-e átfedés azon vegyületek között, amelyek adott nukleáris receptorok ligandjai, és amelyek adott enzimek és transzporterek szubsztrátjai. És ha van átfedés, mely vegyületek alkotják?

Ehhez nukleáris receptorok ligandjainak, metabolikus enzimek és multidrog transzporterek szubsztrátjainak összegyűjtése volt szükséges, melyhez publikus adatbázisok segítségét vettük igénybe. Ezek a PubChem adatbázis Bioassay nevű része, a Metabolism and Transport Database (Metrabase), a Comparative Toxicogenomics Database (CTD) és a DrugBank adatbázis.

Kutatásunk során fehérjéket vizsgálunk, azonban az adatbázisok nagy részében nem fehérjék vannak feltüntetve, hanem a gének, amelyekről a fehérjék átíródnak, így a génnevekkel dolgozunk.

A publikus adatbázisok adatait kinyerés után lokálisan tároljuk, mivel intenzív keresést általában a webalkalmazások nem engednek. E mellett a webalkalmazások nem tartalmazzak minden esetben olyan keresési lehetőségeket, amelyekre szükségünk lenne, a lokális adatokon pedig igen széles körű lekérdezéseket tudunk futtatni, továbbá a lokális keresések rugalmasak, és gyorsabbak. A lokális adattárolással pedig biztosíthatjuk, hogy a projekt során végig az adatbázis ugyanazon verziójával, vagyis ugyanazzal az adathalmazzal dolgozzunk, biztosítva ezzel az eredmények konzisztensségét. Végül pedig a lokális tárolás arra is lehetőséget ad, hogy többféle publikus adatbázis adatait összefűzzük, egy rendszerben tároljuk, egyszerre lekérdezhetővé tegyük.

3.1.1. PubChem BioAssay adatbázis

A PubChem adatbázis (<https://pubchem.ncbi.nlm.nih.gov/about.html>) kis vegyületekről tárol rendszerezetten nagyon sok adatot, és teszi őket letölthetővé. Ez az adatbázis magába foglalja a vegyületek szinonimáit, és kémiai szerkezetét is.

A PubChem adatbázisnak fontos szerepe van a kutatásainkban, mivel a különböző adatbázisokban tárolt vegyületeket ezen adatbázis segítségével lehet azonosítani, ugyanis a kémiai adatbázisok többsége feltünteti az adott vegyületnek ebben az adatbázisban lévő azonosítóját, a PubChem CID-t (Chemical ID). A PubChem adatbázis tartalmaz anyagokat is (substance), amelyeknek külön azonosítójuk van (SID, Substance ID), melyet az adatbázis

adminisztrátorai adnak az anyagnak úgy, hogy az adott SID-t egy már létező CID-hez csatolják. Amennyiben még nem szerepel az adatbázisban olyan vegyület, amely azonos ezzel az anyaggal, akkor egy új CID-t kap a vegyület, és ehhez csatolják a kiosztott SID-et. Több anyag együttesen alkot egy vegyület rekordot, melyhez egy CID tartozik.

A PubChem adatbázison belül a PubChem Bioassay adatbázis (<http://www.ncbi.nlm.nih.gov/pcassay>) [14-16] több, mint 700000 olyan kísérletet gyűjt össze, amelyek vegyületek bioaktivitását, azaz a különböző fehérjékkel való kölcsönhatását vizsgálják. Ezeket az egyedi kísérleteket összefoglaló bioaktivitási jelentésből pedig 200 millió darab található az adatbázisban.

A bioaktivitási adatok elérését segíti a PubChem Bioactivity Data Dicer webalkalmazás (<http://dd.ncbi.nlm.nih.gov/datadicer/2/>), amellyel a PubChem Bioassay adatbázisban lehet keresni génynevek, organizmus és eredmény alapján.

A letöltött adatok a bioaktivitási jelentésekben többféle módon jellemezhetik a fehérje és a vegyület közötti kapcsolatot. Ez lehet pusztán kvalitatív (inaktív vagy aktív), vagy kvantitatív, azaz tartalmazhat koncentráció értékeket μM -ban kifejezve, pl. IC50 (Half Maximal Inhibitory Concentration) értéket, vagy EC50 értéket, vagyis a maximális hatékony koncentráció felét. Minél kisebbek ezek az értékek, annál szorosabb a kölcsönhatás a fehérje és a vegyület között.

A kemoimmun fehérjékkel kölcsönható vegyületeket szeretnénk volna kinyerni a PubChem BioAssay adatbázisból a Bioactivity DataDicer webalkalmazás segítségével, azonban enzimeknél és transzportereknél csak kevés adatot találtunk

Lekérdeztük az összes olyan kísérleti adatot, amelyben az általunk vizsgált 48 emberi nukleáris receptornak, illetve a Metrabaseben található enzimek és transzporterek kölcsönhatását tanulmányozták különféle vegyületekkel. Összesen 7049 olyan adatot találtunk, amelynek kimenetele aktív volt. Ezek közül olyan adat, amelyben valamilyen koncentráció érték is szerepelt, 5120 van, és összesen 553 különféle vegyületet tartalmaz. Ezek közül kiválasztottuk azon kísérletek adatait, amelyeknél a koncentráció érték nem gátló koncentráció (IC50), hanem a maximális hatékony koncentráció fele (EC50), hatékonysági koncentráció (Potency), illetve LC50 érték volt feltüntetve. Ezek száma 4124. Azon vegyületek száma, melyek ezekben szerepelnek, 534.

3.1.2. Metabolism and Transport Database (Metrabase)

A Metrabase (<http://www-metrabase.ch.cam.ac.uk/>) adatbázis [17] rendezett fizikokémiai és biológiai adatokat tartalmaz 20 transzporterről és 13 Citokróm P450

monooxygenázról. 3438 olyan vegyület található ebben az adatbázisban, amely valamiképpen kapcsolódik a 33 fehérjéhez. A vegyületek lehetnek például szubsztrátok, gátlók, transzkripciót indukáló anyagok.

Az adatbázisban fellelhető vegyületek között vannak olyanok is, amelyekről leírták, hogy biztosan nem szubsztrátjai a 33 fehérje közül bizonyos fehérjéknek. Ilyen adatokat az irodalomban ritkán közölnek, és nem minden adatbázisban jelenítenek meg, pedig jelentősen segítik a vegyületek és fehérjék közötti kapcsolatok feltérképezését.

A Metrabase adatait egy MySQL adatbázisba mentettük le, egyesítve a PubChem BioAssay adatbázisból a fentebb részletezett módon kinyert nukleáris receptor adatokkal. A MySQL adatbázisból kérdeztem le a benne található 33 fehérje szubsztrát adatait.

A Metrabase adatbázis alkalmas transzporterek és enzimek szubsztrát adatainak megismerésére, azonban nem tartalmaz adatokat nukleáris receptorokról, ezért fordultunk a Comparative Toxicogenomics Database (CTD) felé.

3.1.3. Comparative Toxicogenomics Database (CTD)

Ebben az adatbázisban [18] (<http://ctdbase.org>) vegyületek, gének és betegségek közötti kapcsolatokat leíró adatokat helyeztek el, összesen 24 004 160 darab kapcsolatot. Elsődleges célja volt az adatbázis létrehozásának annak megismerése, hogy a környezetből származó anyagok milyen hatással vannak az ember egészségére. A vegyületek, gének, vegyületek és betegségek, illetve gének és betegségek közötti kapcsolatokat az adatbázis készítői irodalomból manuálisan gyűjtötték össze. Az adatbázis adatait lokális MySQL adatbázisba gyűjtöttük, abból kérdeztük le kemoimmun fehérjék szubsztrátjait a 2.2.2 fejezetben leírt módon.

3.1.4. DrugBank adatbázis

A DrugBank adatbázis [19-22] (<http://www.drugbank.ca/about>) gyógyszereket és gyógyszerjelölteket tartalmaz. Egyesíti a gyógyszerek és a fehérjék részletes adatait, például fehérjéknél az útvonal adatokat is tartalmazza. Sok más adatbázisra hivatkozik, így megkönnyíti a vegyületek és a fehérjék azonosítását. A DrugBank adatbázis 7740 gyógyszert, 4282 fehérjét, illetve a gyógyszerek és fehérjék kapcsolatait foglalja magába. Minden egyes DrugBank bejegyzés (DrugCard entry) több, mint 200 adatmezőből áll, melyek fele a gyógyszeréről, a másik fele pedig a gyógyszer célpontjairól tartalmaz információt. Az adatbázis sokféle módon kereshető, pl. szöveges lekérdezéssel, aminosav szekvenciával BLAST (Basic Local Alignment Search Tool) lekérdezéssel, vagy kémiai struktúrával, illetve SMILES kémiai struktúra leíró szöveggel. Az adatbázis egésze letölthető XML formátumban. A megfelelő adatok kinyerés

után XML formátumból relációs adatbázisba rendezhetők. Az adatok feldolgozása folyamatban van.

3.1.5. Irodalmi adatok

A fent részletezett publikus adatbázisok közül nem mindegyik tartalmaz adatot minden fajta kemoimmun fehérjéről, e mellett az adatbázisok szubsztrát adataiban kevés az átfedés a nukleáris receptorok, illetve multidrog transzporterek szubsztrátjai között (ld. 3.2.2 fejezet), ezért elkezdtük adatok összegyűjtését az irodalomból.

A kutatás elején három kiválasztott nukleáris receptor: a CAR a PXR és a VDR ligandjait gyűjtöttem össze cikkekből.

A fehérjék és a ligandok közötti kapcsolatok, vagyis kötési adatok egyenként szerepelnek az adatbázisban. Több kötési adat van, mint ligand, azaz vegyület, mert egy vegyület több fehérjéhez is köthet.

Összesen 12 db fehérjéhez 60 db ligandot, 98 db ligand kötési adatot 5 db cikkből [4, 7-8, 23-24] gyűjtöttem össze. Azért nem 3, hanem 12 fehérje van az adatbázisban, mert nem csak emberi, hanem egyéb fajokból származó fehérjéket is összegyűjtöttünk, ugyanis a cikkekben van sok adat más fajok fehérjéről is, és így a jövőben a fajspecifikus eltéréseket is lehet tanulmányozni. A feldolgozott cikkek száma azért ilyen alacsony, mert ezek áttekintő jellegűek, azaz mindegyike sok adatot tartalmaz.

3.1.6. A publikus adatbázisok összehasonlítása

A Metabolism and Transport Database (Metrabase) adatbázis sajnos nem tartalmaz nukleáris receptorokat.

A Comparative Toxicogenomics Databaseben (CTD) vannak olyan vegyület fehérje interakciókat leíró mondatok, amelyek szögletes zárójelek között tartalmazznak más ilyen mondatokat, például „Dexamethasone inhibits the reaction [ABCG2 protein affects the transport of 2-amino-1-methyl-6-phenylimidazo(4,5-b)pyridine]”. Ha egy vegyület és egy fehérje közötti kapcsolatra keresek, akkor találok ilyen mondatokat is, meg olyanokat is, amelyek nem tartalmazznak más mondatokat, például „ABCC1 protein results in increased export of 2',7'-bis-(carboxypropyl)-5(6)-carboxyfluorescein”. Ez utóbbi mondatokat nevezem egyszerű mondatoknak, mivel nem tartalmazznak más mondatokat szögletes zárójelen belül. Csak egyszerű mondatokkal keresve a 48 nukleáris receptorra és a 33 Metrabaseben szereplő fehérjéhez 680 ligandot találtam, míg egyszerű és nem egyszerű mondatokat összesen 690 ligandot, illetve szubsztrátot foglalnak magukba. Mivel a CTD-ben több vegyületet találtunk,

mint a PubChem BioAssay adatbázisban, ezért arra a következtetésre jutottunk, hogy a CTD adatait érdemesebb elemzésnek alávetni.

A publikus adatbázisok két fontos hibalehetőséggel terheltek. (1) Ha nincs adat a fehérje és a vegyület közötti kölcsönhatásról, az jelentheti azt, hogy a fehérje valóban nem ismeri fel a vegyületet, de jelentheti azt is, hogy még nem végeztek kísérletet az adott fehérjével és vegyülettel, vagy még nem raktak be erre vonatkozó adatokat az adatbázisba. (2) Membrántranszporterek esetén szubsztrátnak minősülő vegyületet jelölnek meg több esetben gátlószerként az alapján, hogy (kompetitív módon) gátolják egy másik szubsztrát transzportját.

3.1.7. A ligand- és szubsztrát adatok egyesítése: az általunk létrehozott DrugDB adatbázis

Ahhoz, hogy megtudjuk, mely vegyületek aktiválják ugyanazokat a nukleáris receptorokat, bizonyos nukleáris receptorok irodalomban fellelhető ligandjait és a ligand kötések tulajdonságait jellemző adatok strukturált rendszerbe helyezése volt szükséges, ezért létrehoztunk egy adatbázist. Ez egyelőre még csak a PXR, a VDR és a CAR általunk az irodalomból kigyűjtött ligandjait, azaz a 3.1.5. fejezetben leírt adatokat tartalmazza, azonban ebben az adatbázisban tervezzük integrálni a fenti adatbázisokban és a EMBL-EBI ChEMBL adatbázisban (<https://www.ebi.ac.uk/chembl/>) található kemoimmun fehérje szubsztrátokat és ligandokat.

A DrugDB adatbázis szerkezete

A DrugDB nevű adatbázisban fehérjék ligand és szubsztrát kölcsönhatásait jellemző adatok tárolását oldottuk meg. Ehhez több adatbázisból gyűjtöttünk adatokat: A PubChem-ből [14], az UniProtból [25] és a PubMedből [26].

Az adatbázis hat táblából áll. A „protein” nevű tábla a fehérjék adatainak tárolására szolgál. A vegyületek adatait a „compound” nevű tábla tartalmazza, az irodalmi hivatkozások tárolásáért pedig az „article” nevű tábla a felelős. A „protein” tábla oszlopai: génnév, UniProt azonosító (acc), melynél megköveteltük az egyediséget (unique oszlop), hogy semmiképp se szűrjük be többször ugyanazt a fehérjét. E mellett a fehérje tábla tartalmazza a következő oszlopokat: beírási név (entry name) és az organizmus. A „compound” tábla oszlopai: a vegyület neve és egy megjegyzés a vegyülethez. Az „article” tábla oszlopai: a cím, a szerzők, az összefoglaló és a PubMed azonosító, melynél szintén megköveteltük az egyediséget, hogy semmiképp se szűrjük be többször az adott irodalmi hivatkozást.

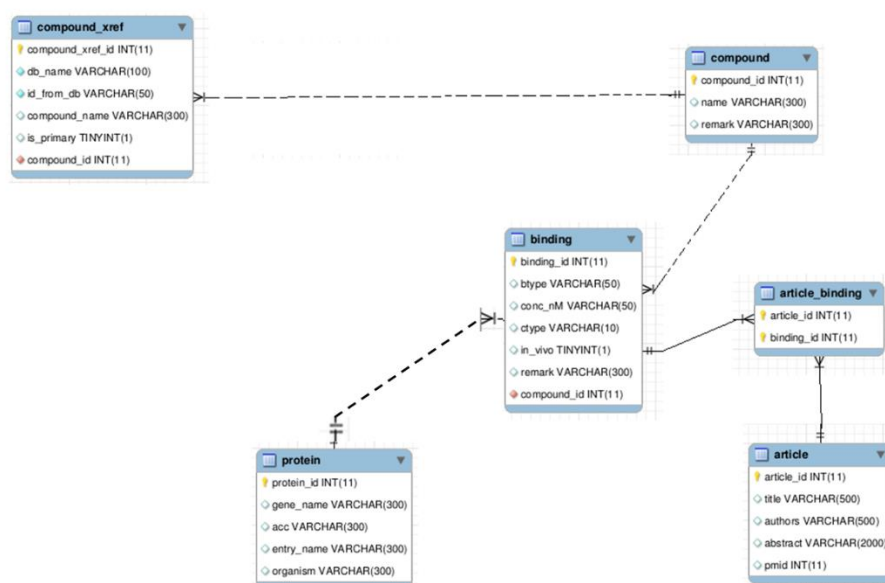
A vegyületek adatbázisokban lévő azonosítóinak tárolására egy újabb táblát hoztunk létre, melynek neve „compound_xref”. E megoldással egy vegyülethez több azonosító is

tartozhat. Ennek a táblának az oszlopai: az adatbázis neve, melyből a vegyületet leíró adatokat kinyertük (pl. a PubChem [14]), a vegyület adatbázisban lévő azonosítója, a vegyület neve, és egy logikai változó, mely 1, ha az adott hivatkozás elsődlegesnek tekintendő, és 0, ha nem. Minden vegyületnél kiválasztottunk egy elsődleges azonosítót, amely az első az adatbázisból lekérdezett azonosítók közül. A tábla külső kulcsként (foreign key) tartalmazza a vegyület elsődleges kulcsát. Az adatbázis nevének és az adatbázisban tárolt azonosítónak együttesen véve egyedinek kell lennie, mert előfordulhat, hogy egy vegyületnek ugyanaz az azonosítója egy külső adatbázisban, mint egy másik vegyületnek egy másik külső adatbázisban.

Az adatbázis hat táblájából egy kapcsolótábla, vagyis több a többhöz kapcsolatot létesít entitások között. Az irodalmi hivatkozások tábláját köti össze a ligandkötések táblájával. A neve „article_binding”. Ennek segítségével egy ligandkötéshez több irodalmi hivatkozás, és egy irodalmi hivatkozáshoz több ligandkötés is tartozhat (many-to-many relationship). Csak az összekapcsolandó entitások elsődleges kulcsait tartalmazza.

Az adatbázis szíve az a „binding” nevű tábla, melyben a fehérjék ligand kötéseinek adatai vannak tárolva. Oszlopai a kötés típusa (pl. agonista), a kísérletnél alkalmazott koncentráció típusa (pl. EC₅₀), és a koncentráció értéke nM-ban kifejezve. Mindezek mellett van még egy logikai változó annak tárolására, hogy a kísérletet *in vivo* vagy *in vitro* hajtották-e végre, megtalálható még egy megjegyzés a kísérlethez vagy kötéshez, és egy külső a vegyületeket tároló táblához.

Az adatbázis fentebb vázolt szerkezete a 3.1. ábrán, látható.



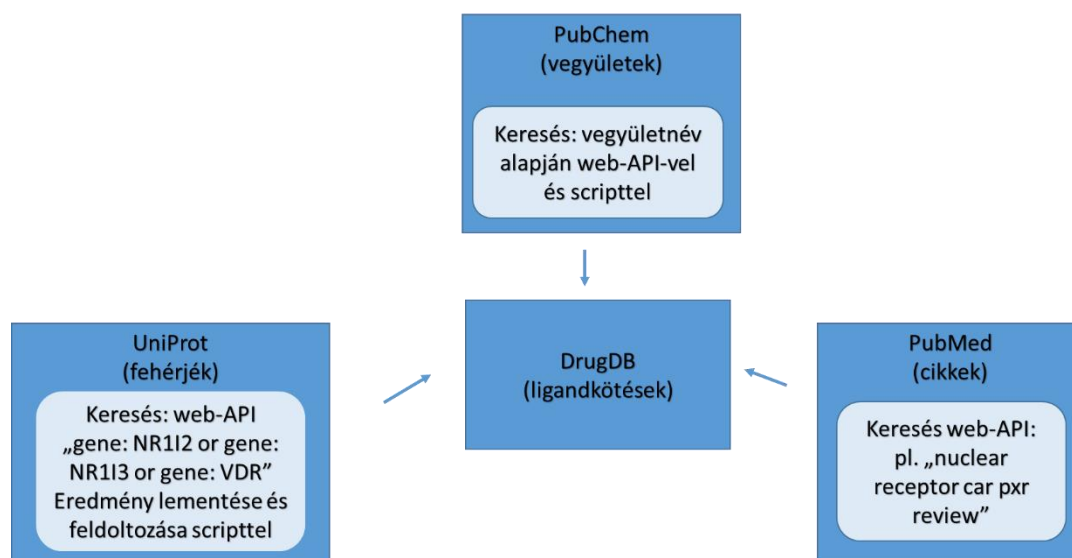
3.1. ábra: A DrugDB adatbázis EER diagramja „crows feet” (varjúláb) jelölési módban. Piros ponttal vannak jelölve az elsődleges kulcsok, és sárga jellel a külső kulcsok.

Az ábrát a MySQL Workbench (<http://www.mysql.com/products/workbench/>) nevű programmal készítettem, egyszerűsített nézetet használva

A DrugDB adatbázisba betöltött adatok összegyűjtése

A DrugDB-be illesztett ligand kötési adatokat a 3.1.5. fejezetben hivatkozott cikkekben [4, 7-8, 23-24] megtalálható táblázatokból gyűjtöttem össze. Ezekre úgy találtam rá, hogy a PubMedben (<http://www.ncbi.nlm.nih.gov/pubmed/>) a 'nuclear receptor CAR PXR review', 'nuclear receptor CAR PXR ligand review' és a 'Vitamin D3 receptor ligand review' kifejezésekkel kerestem, a találatok közül pedig kiválogattam azokat, amelyek táblázatokban sok vegyületet tartalmaznak. A ligandok összegyűjtését nehezítette az a tény, hogy vannak olyan vegyületek, melyek hatással lehetnek a vizsgált fehérjékre anélkül, hogy ténylegesen kötődnének hozzájuk. Az irodalom olvasása során figyelniem kellett arra, hogy ténylegesen a ligand szóval megjelölt vegyületeket vegyem csak figyelembe. Így az adatbázisba csak olyan vegyületek kerültek bele, melyeket ligandként írtak le.

A fehérjékre vonatkozó adatokat úgy kerestem meg, hogy az UniProt adatbázisban a következő kifejezéssel hajtottam végre lekérdezést: „gene: NR1I2 or gene: NR1I3 or gene: VDR”, a kapott eredményeket szövegfájlként lementettem, majd a fehérjék adatait, például UniProt azonosítóit a Biopython [27] modul segítségével nyertem ki és töltöttem be az adatbázisba. A vegyületekre és a cikkekre vonatkozó adatokat scriptekkel kértük le a PubChem, illetve a PubMed adatbázisból. A fentebb említett adatbázisok és az adatok kinyerésének módja látható összefoglalva a 3.2. ábrán.



3.2. ábra: A DrugDB adatbázis forrásadatbázisai és az adatkinyerés módjai

Az adatbázisban jelenleg három nukleáris receptornak, a PXR-nak (NR1I2), a CAR-nek (NR1I3) és a VDR-nek (NR1I1) irodalomból kigyűjtött ligandjai vannak, de ebbe az adatbázisba tervezzük egyesíteni a PubChem BioAssay, a Metabolism and Transport Database (Metrabase), a Comparative Toxicogenomics (CTD) és a DrugBank adatbázisok kemoimmun fehérjékre vonatkozó szubsztrát és ligand adatait.

3.2. A Comparative Toxicogenomics Database-ből származó szubsztrát- és ligand adatok elemzése

3.2.1. A rendelkezésre álló adatok elemzése

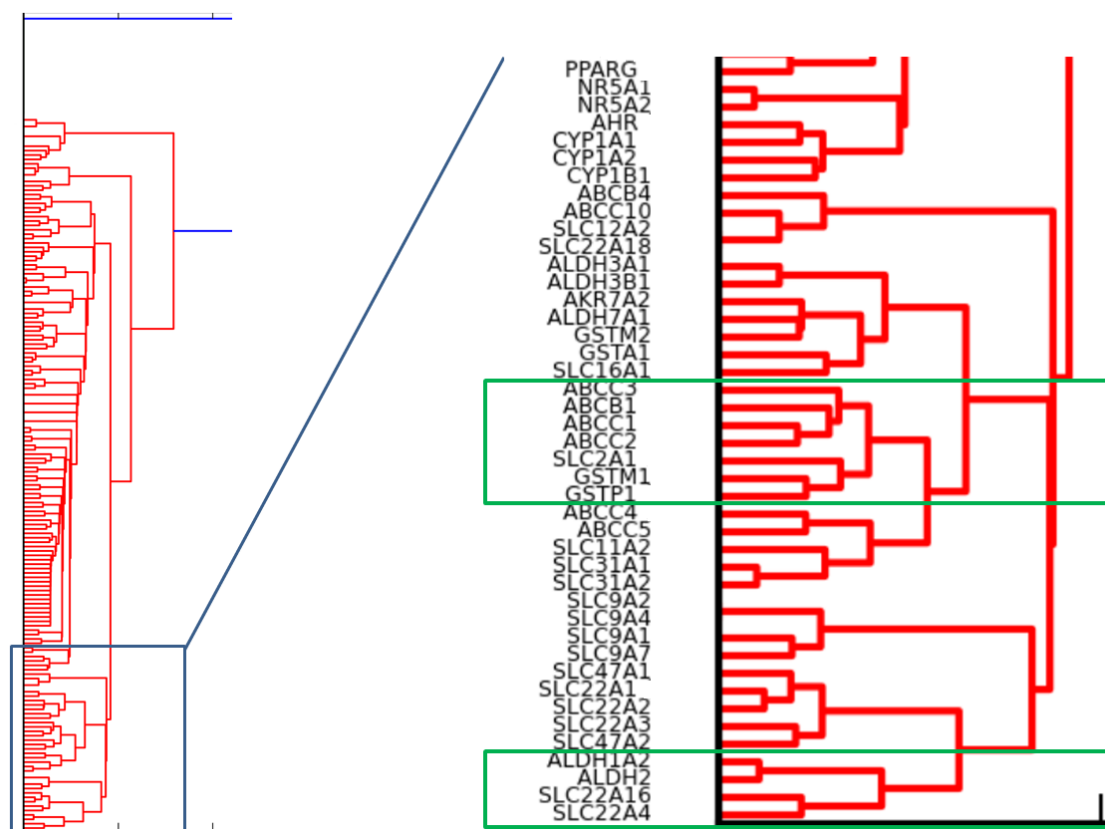
A kutatásunk során azt szeretnénk megválaszolni, hogy vajon a kemoimmunitási fehérjék hálózatként végzik-e feladatukat, az idegen anyagok átalakítását és sejtből való eltávolítását, vagyis a sejt idegen anyagok, kis molekulák elleni védelmét. Ehhez azt vizsgáljuk meg, hogy a vegyületek egy csoportja a kemoimmun fehérjék egy csoportjával ha-e kölcsön. Mivel a fentebb leírt publikus adatbázisok közül a CTD-ben található a legtöbb szubsztrát adat, és ebben az adatbázisban nukleáris receptorok ligandjai is megtalálhatóak, ezért úgy döntöttünk, hogy ennek az adatbázisnak az adatait elemezzük részletesebben. E mellett a CTD adatbázis génexpressziós adatokat is tartalmaz, amelyek a fehérjeszabályozás felderítésében nélkülözhetetlenek, ezért is érdemes jobban megismerni ezt az adatbázist.

Az CTD-ből lekérdezett ligand és szubsztrát adatokat a 2.2.3 fejezetben leírt módon összegyűjtöttem, hogy azután statisztikailag elemezni tudjuk őket.

Fehérjék szubsztrát, illetve ligandspektrum szerinti klaszterezése

A fehérjéket szubsztrátspektrumuk alapján a 2.2.4. fejezetben leírt módon csoportosítottuk.

Az eredményként kapott dendrogramot (fát), illetve annak kinagyított részeit látjuk a 3.3. ábrán.

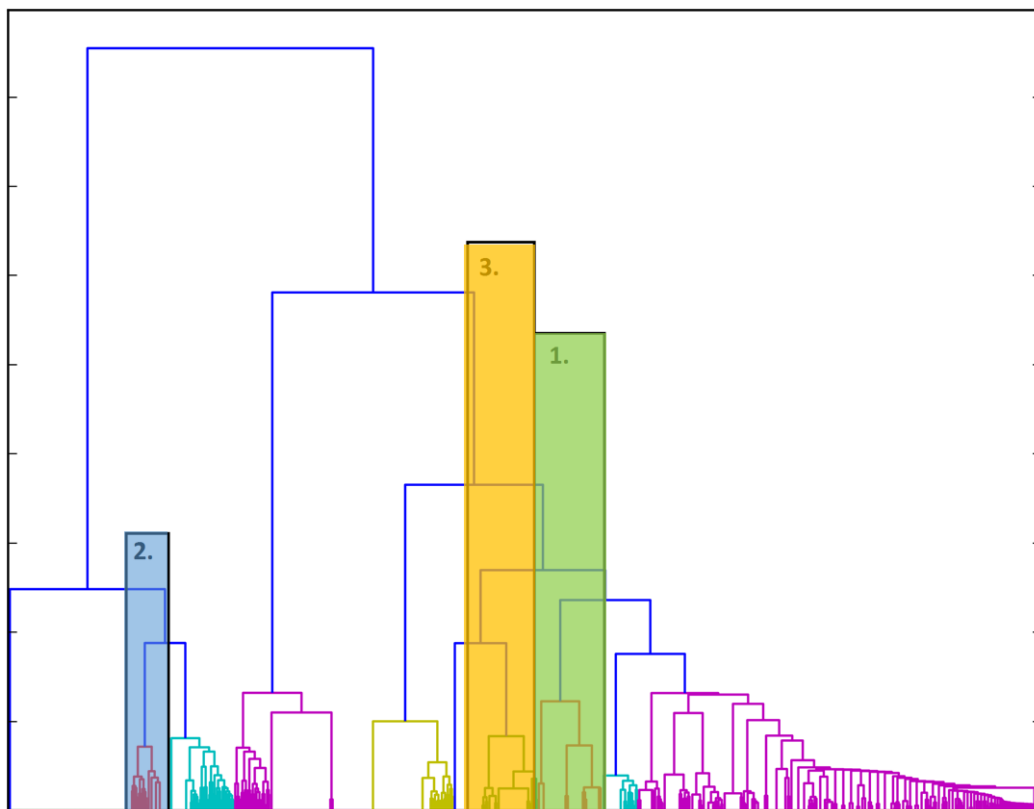


3.3. ábra: Fehérjék szubsztrát és ligandspektrum szerinti klaszterezéséből származó dendrogram, és annak kinagyított részei.

Két fehérjeklaszter látható zölddel bekeretezve az ábra jobb oldali kinagyított részén. A felső klaszterben olyan ABC és SLC transzporterek, illetve GST fehérjék vannak, amelyek multidrog-kötő fehérjék. Ez azt mutatja, hogy jól klasztereztünk, hiszen hasonló fehérjék kerültek egy csoportba. Az alsó klaszterben alkohol dehidrogenázok és SLC22 transzporterek találhatók. Ezeket a fehérjéket együtt még nem említik az irodalomban, együttműködésük lehetősége új információ. A négy fehérje közös szubsztrátja a doxorubicin.

Vegyületek klaszterezése fehérjékhez való kapcsolódásuk szerint

Azt szeretnénk megtudni, mely vegyületek hasonlítanak abban, hogy ugyanazoknak a fehérjéknek szubsztrátjai vagy ligandjai. Ezért a fehérjék és a vegyületek közötti kapcsolatot leíró előbbi adathalmazt most úgy klasztereztük a 2.2.4. fejezetben leírt módon, hogy a vegyületek alkossanak klasztert. Az optimális klaszterszámot meghatároztuk, mely 10 lett. Ezen optimális klaszterszám alapján határoztuk meg azt a küszöbértéket, amely alapján az 3.4. ábrán látható dendrogramot színeztük.



3.4 ábra: vegyületek szubsztrát illetve ligandspektrum szerinti klaszterezésének dendrogramja. Három kiemelt klaszter be van keretezve.

Ezután megnéztük klaszterenként, hogy melyek azok a fehérjék, amelyekhez az adott csoportban lévő vegyület kapcsolódik, azok közül is kiválasztottuk azokat a fehérjéket, amelyek sok vegyületet felismernek az adott vegyületcsoportból. Az így kapott fehérjecsoportokban egy fehérje több helyen is szerepelhet, hiszen több vegyület kapcsolódhat egy fehérjéhez (egy fehérje eltérő csoportba klasztereződött vegyületeket is felismerhet).

Az alábbiakban három ilyen fehérjecsoportot emelünk ki és mutatunk be. A nekik megfelelő három vegyületklasztert a 3.4. ábrán feketével bekereteztük, a csoport fehérjei és a fehérjék közötti kapcsolatok pedig a 3.1. táblázatban láthatóak.

Kemoimmun fehérjék	1. csoport	2. csoport	3. csoport
Nukleáris receptorok, transzkripciósfaktorok	PPARD	ESR1	ESR1
	RXRA	ESR2	ESR2
	PPARG	PXR ²	PPARG
	PPARA		PPARA
			AHR ³
Transzporterek (fázis 0, III)	-	ABCB1 ²	ABCC1
Metabolikus enzimek (Fázis I)	-	CYP3A4 ²	CYP3A4
		CYP2B6 ²	CYP1A1 ³
			CYP19A1
			CYP1B1 ³
			CYP1A2 ³

3.1. táblázat: a három kiemelt fehérjecsoport tagjai. Szám a felső indexben: a csoporton belül az adott nukleáris receptor és egyéb más fehérje közötti kapcsolatot jelenti, pl. 1: a PXR szabályozza transzkripciósan az ABCB1-et. A fehérjekapcsolatok a [3] irodalmi hivatkozásban találhatók:

Mindhárom kiemelt csoportban olyan fehérjék szerepelnek, amelyek az adott vegyületcsoportból legalább 14-et felismernek. Az első csoport csak négy fehérjéből áll. Mivel ezeknek a fehérjéknek lipid felismerés a feladata, ez a klaszter jól mutatja, hogy elemzési módszerünk alkalmas a hasonló anyagokkal kölcsönható fehérjék azonosítására.

A második csoport hat fehérjéből áll. A PXR és az ABCB1 együtt szereplését igazolja az irodalomból ismert tény, hogy a PXR szabályozza az ABCB1-et [3].

A harmadik csoportot 11 fehérje alkotja. Az irodalomból [3] ismert, hogy az AHR szabályozza a CYP1A1, a CYP1A2 és a CYP1B1 fehérjéket, ezért ezek a fehérjék egy csoportban lévő elhelyezkedése indokolt.

Fehérjék páronkénti szubsztrátátfedése

A csoportokat a következőképpen is vizsgáltam: megszámláltam minden egyes kemoimmun fehérjepárra, hogy hány csoportban szerepelnek együtt a 10-ből, és ebből egy olyan táblázatot készítettem, ahol az összes kemoimmun fehérje szerepel az oszlopok és a sorok fejléceiben is, és a cellákban az található, hogy a sorfejlécben és az oszlopfejlécben lévő fehérje hány csoportban szerepel együtt. Ezután csak azokat a sorokat és oszlopokat hagytam meg, ahol magas értékek találhatók (6 vagy 7), és ezeket zölddel jelöltem meg. Az eredmény az alábbi 3.2. táblázatban látható.

	ABCB1	CYP1A1	CYP1A2	CYP2B6	CYP3A4	NR3A1/ESR1	NR1I2/PXR
ABCB1		5	5	4	6	5	6
CYP1A1	4		6	4	6	5	5
CYP1A2	2	12		4	6	5	5
CYP2B6	1	4	7		6	5	5
CYP3A4	4	4	8	15		6	7
NR3A1/ESR1	8	12	4	1	7		5
NR1I2/PXR	8	7	7	9	23	33	

3.2. táblázat: Főátló felett: fehérjepáronként az az érték, hogy az adott fehérjepár a vegyületcsoportokból meghatározott 10 fehérjecsoportból hányban szerepel együtt, a magas értékek zölddel vannak jelölve. Főátló alatt: fehérjék szubsztrátátfedése

Mivel a kiinduló fehérjecsoportok vegyületcsoportokból voltak származtatva úgy, hogy azok a fehérjék vannak egy csoportban, melyeknek van olyan szubsztrátjuk vagy ligandjuk, amely vegyület az adott csoportban van, ezért a fentebbi 3.1. táblázatban szereplő értékek minél nagyobbak, annál több vegyületet ismer fel a fehérje. És minél több csoportban szerepel együtt két fehérje, annál nagyobb a szubsztrátátfedésük.

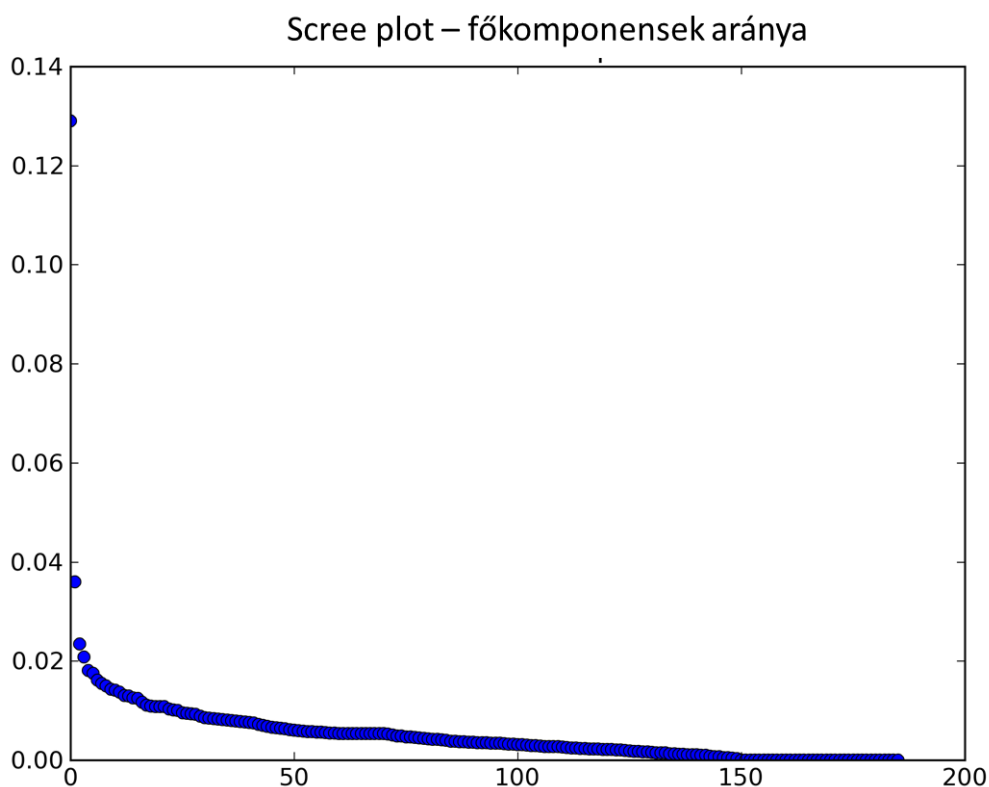
Figyelemre méltó az, hogy az ABCB1 és az NR1I2 (PXR) 6 csoportban együtt szerepel a 10-ből. Ez az együttállás azt sejteti, hogy a PXR specifikusan olyan transzporter transzkripcióját serkenti, amelyek az általa felismert ligandok transzportálására képes. Az ABCB1 és a PXR által is felismert vegyületek, melyek az adott vegyületcsoportba tartoznak: a Dexametazon, az Endoszulfán, az Etinil ösztrogén, a Paclitaxel, a Ritonavir, az alachlor, a docetaxel és a metolachlor.

A CYP3A4 és az NR1I2 (PXR) 7 csoportban is együtt van a 10-ből. Irodalomból ismert tény, hogy a PXR erősíti a CYP3A4 expresszióját [3]. Közös felismert vegyületeik a Karbamazepin, a Chlorpyrifos, a Clotrimazole, a Dexametazon, a Diethylhexyl Phthalate, a Diethylstilbestrol, a Flutamid, a Lithocholic Acid, a Paclitaxel, a Fenobarbitál, a Fenitoin, a Rifampin, a Sulfinpyrazone, a Triclosan, a Troleandomycin, a bisphenol A, a cyhalothrin, a cypermethrin, a decamethrin, a dexamethasone 21-tributylacetate fenvalerate, a o,p'-DDT és a phenthoate.

A CYP3A4 és az ESR1 6 csoportban is együtt szerepel. Közös felismert vegyületeik a 2,2',4,4'-tetrabromodiphenyl éter, a Diethylhexyl Phthalate, a Diethylstilbestrol, a növényi kivonatok, a biszfenol A, a fenvalerate és az o,p'-DDT.

Vegyületek főkomponens analízis (PCA) alapján való klaszterezése

A klaszterezés mellett egy másik módszer a főkomponens analízis (PCA), melyet azért alkalmaztunk, hogy más szemszögből is megnézve újabb összefüggéseket keressünk a kemoimmun fehérjék szubsztrát és ligand adatai között.

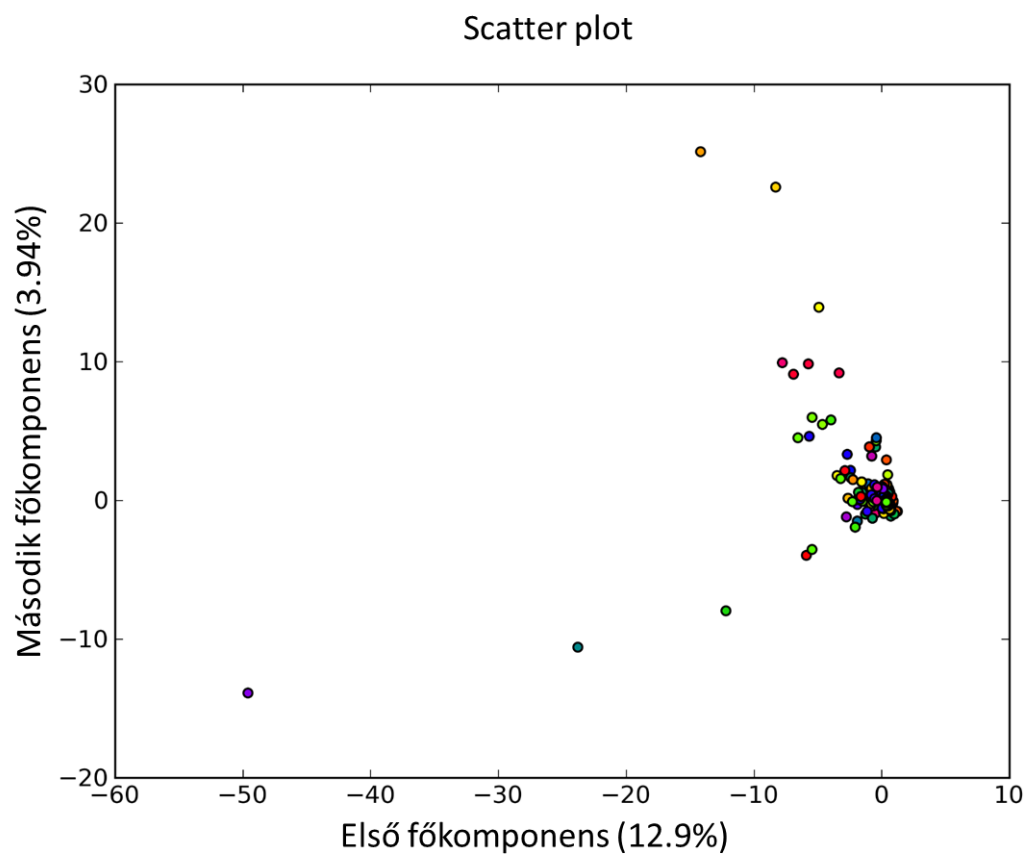


3.5. ábra: Vegyületek főkomponens analízis komponenseinek aránya

A fenti 3.5. ábrán a komponensek aránya látható.

Noha az első két komponens láthatóan jóval nagyobb, mint a többi, egyik komponens sem haladja meg a 14 százalékot.

Amikor rávetítettük a vegyületek pontjait az első két főkomponens síkjára (3.6. ábra), egy népes csoportot láttunk az origóban, és egy vagy néhány vegyületet szétszóródva. Az origóban lévő csoportosulás azt sejteti, hogy sok olyan vegyület van, amely csak egy-egy fehérjéhez kapcsolódik.



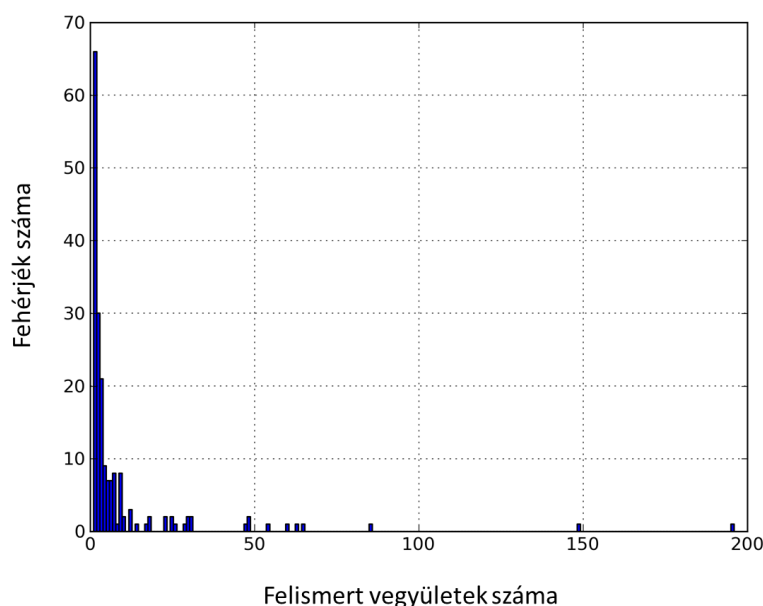
3.6. ábra: Vegyületek pontjainak elhelyezkedése a vegyületek főkomponens analíziséből származó két legjelentősebb komponens által meghatározott síkon

Ez a módszer nem világított rá újabb összefüggésekre, csupán azt sejteti, mennyire sok olyan vegyület van, amely csak egy vagy néhány fehérjével hozható kapcsolatba az adataink alapján.

3.2.2. A Comparative Toxicogenimocs Database-ből származó adatok jellemzése

Az eddigi elemzések eredményeit látva meg kell vizsgálnunk, milyen is az adathalmaz, amivel dolgozunk.

A 3.7. ábrán egy hisztogram látható, mely azt mutatja, hány olyan fehérje van, amely adott számú vegyületet ismer fel.



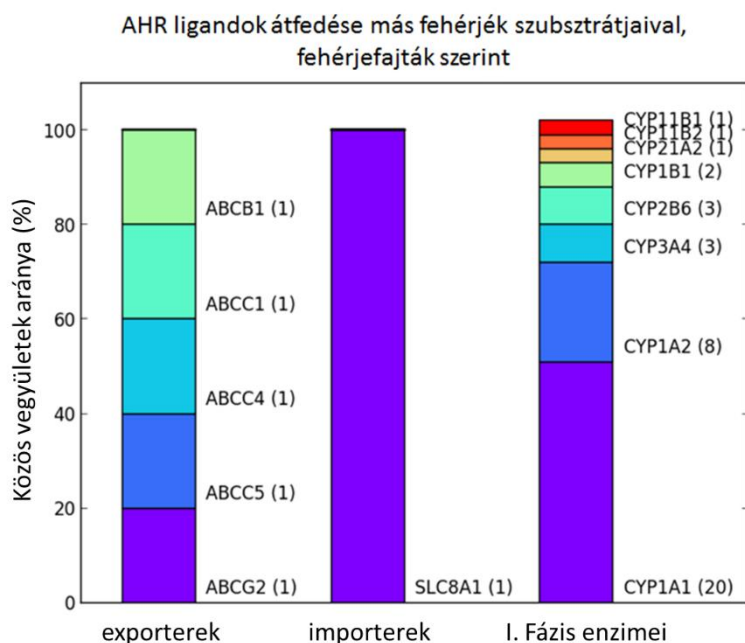
3.7. ábra: Hisztogram a fehérjék által felismert vegyületek számáról a Comparative Toxicogenomics Database (CTD) adatait felhasználva. A függőleges tengelyen a fehérjék száma, a vízszintes tengelyen a felismert vegyületek száma látható.

Szembetűnő, hogy nagyon sok olyan fehérje van, amelynél csak egy vegyülettel kapcsolatban van adatunk, és sok olyan fehérje, ahol csak néhány vegyülettel kapcsolatban van adatunk, és csak igen kevés olyan fehérje, amelyről sok a szubsztrát vagy ligand adat. Ezt támasztja alá a vegyületek főkomponens analízisééről szóló 3.6. ábrán az origó köré csoportosult vegyületek nagy száma is.

E mellett azt is vizsgáltuk, hogy mennyire fednek át a különböző fehérjék szubsztrátjai.

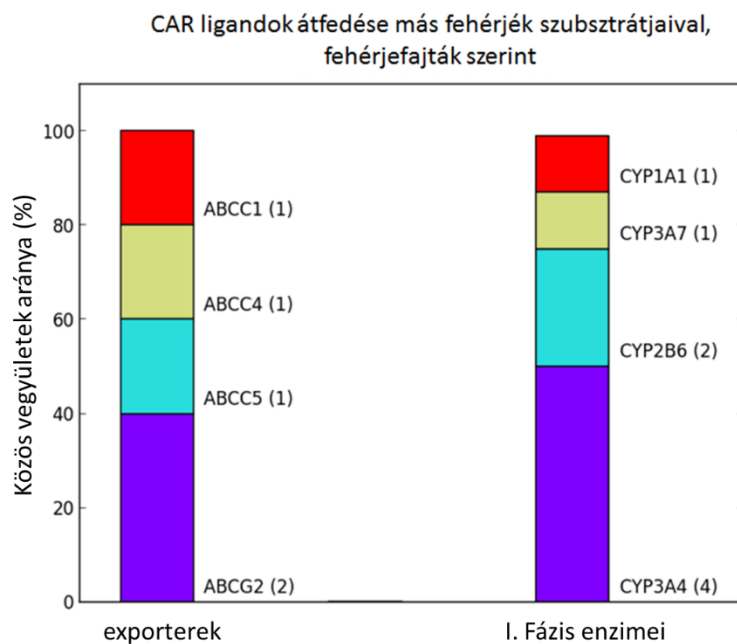
A 3.8. ábra első oszlopa azt mutatja, hogy az AHR ligandjainak és az ABC fehérjék szubsztrátjainak átfedése hány vegyületből is áll, és azok pontosan mely ABC fehérjéhez kapcsolódnak. Látható, hogy csak egy-egy vegyület kapcsolódik az ABC fehérjékkel, összesen 5 darabbal. A második oszlop azt mutatja, hogy az AHR ligandjai az SLC fehérjék közül csak az SLC8A1 fehérjével fednek át egy vegyület erejéig. A harmadik oszlop az AHR ligandjainak a citokróm P450 fehérjék szubsztrátjaival való átfedését mutatja CYP fehérjékné. Összesen 8

db citokróm P450 fehérje szubsztrátjaival fednek át az AHR ligandjai, és csak a CYP1A1 ligandjaival fednek át jelentősen, 20 vegyületnél is.



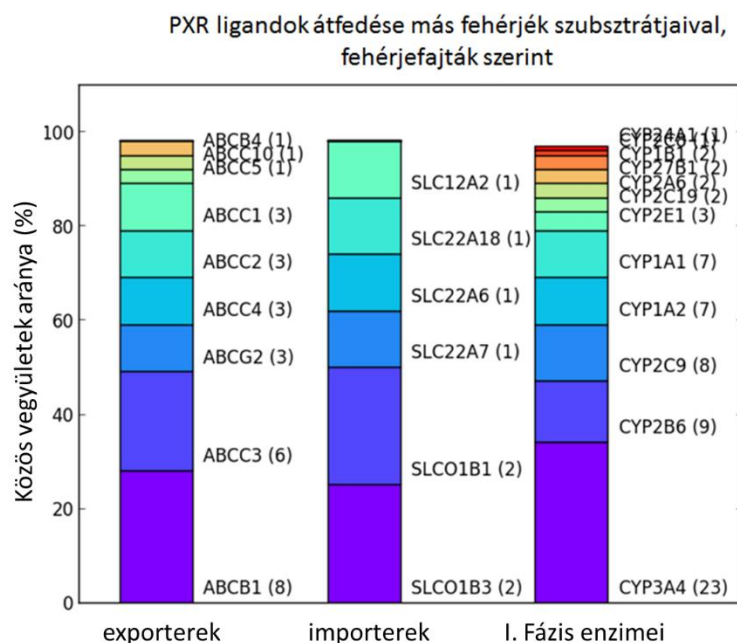
3.8. ábra: AHR ligandok átfedése különböző fehérjék szubsztrátjaival, fehérjefajták szerint.

A 3.9. ábra a CAR ligandjainak átfedését mutatja más fehérjék szubsztrátjaival. Az első oszlopban az ABC fehérjék szubsztrátjaival való átfedés látható. Összesen négy fehérje szubsztrátjaival fed át, és közülük is csak egy olyan fehérje van, amelynél két vegyület is átfed, a többinél csak egy-egy. A második oszlop a CAR ligandjainak citokróm P450 fehérjék szubsztrátjaival való átfedését mutatja. Itt a CYP3A4 szubsztrátjaival fed át 4 vegyületben, amely igen kevés az AHR és CYP1A1 20 vegyületes átfedéséhez képest. A CAR ligandjai a többi fehérjével még kevésbé fednek át, a CYP2B6 szubsztrátjaival két vegyület erejéig, a maradék két citokróm P450 fehérje szubsztrátjai és a CAR ligandjai között csak egy-egy vegyület fed át. Összességében a CAR-nak nagyon kevés olyan ismert ligandja van, mely átfed más fehérjék szubsztrátjaival.



3.9. ábra: CAR ligandok átfedése különböző fehérjék szubsztrátjaival, fehérjefajták szerint

A 3.10. ábrán pedig a PXR ligandok átfedését látjuk az első oszlopban az ABC fehérjék szubsztrátjaival, a másodikban az SLC fehérjék szubsztrátjaival, a harmadikban a citokróm P450 fehérjék szubsztrátjaival. Az AHR, a PXR és a CAR szubsztrátjai közül a PXR ligandjai fednek át a leginkább más fehérjék szubsztrátjaival.



3.10. ábra: PXR ligandok átfedése különböző fehérjék szubsztrátjaival, fehérjefajták szerint.

Az alábbi 3.3-as táblázat kemoimmun fehérjék szubsztrát, illetve ligand átfedését mutatja. Csak azon kemoimmun fehérjék szerepelnek benne, amelyekhez kapcsolódó vegyületek bármely másikkal való átfedését több, mint két vegyület alkotja. Látható, hogy a több, mint két vegyület erejéig átfedő fehérjék közül is a legnagyobb átfedés az NR1I2 (PXR) és a CYP3A4 fehérjék közötti 23 vegyület, ahol a PXR összes ligandjainak száma 196, ez 12%-os átfedést jelent. Az egyetlen jelentős átfedés az AHR-hez és a CYP1A1-hez kapcsolódó vegyületek közötti 20 vegyület az AHR 54 ligandjából, mely 37%-os átfedést jelent. Ezen kívül csak egy darab 10 fölötti átfedési értéket látunk, az ESR1-hez és a CYP1A1-hez kapcsolódó vegyületek között, ez 12 az ESR1 149 ligandjából, ami 8%-os átfedést jelent.

Irodalomból tudjuk [3], hogy a PXR szabályozza az MDR1-et. Azt várnánk, hogy sok átfedő vegyületet találunk a két fehérje között, pedig csak 8 ilyen vegyület van a CTD adatai szerint. E mellett leírták [3], hogy az AHR szabályozza a CYP1B1-et, így sok olyan vegyületet várnánk, ami mindkét fehérjéhez köt, azonban csak két ilyen vegyület van a CTD-ben.

	AHR	ESR1	ESR2	NR1H4	PXR/NR1I2	CAR/NR1I3	NR3C1	PPARG
ABCB1	1	8	4	0	8	0	4	0
ABCC1	1	4	3	0	3	1	0	1
ABCC2	0	3	1	0	3	0	0	0
ABCC3	0	2	0	3	6	0	2	0
ABCC4	1	2	2	0	3	1	0	0
ABCG2	1	5	2	0	3	2	1	2
CYP1A1	20	12	7	0	7	1	0	1
CYP1A2	8	4	2	0	7	0	0	0
CYP1B1	2	4	2	0	2	0	0	0
CYP27B1	0	3	3	0	2	0	0	0
CYP2B6	3	1	1	0	9	2	0	0
CYP2C9	0	1	1	0	8	0	1	0
CYP2E1	0	2	2	0	3	0	0	0
CYP3A4	3	7	5	1	23	4	1	3
SLCO1B1	0	0	0	4	2	0	0	0
SLCO1B3	0	0	0	8	4	0	0	0

3.3 táblázat: Kemoimmun fehérjékhez kapcsolódó átfedő vegyületek száma a Comparative Toxicogenomics Database (CTD) adatai szerint, Az oszlopokban azok a nukleáris receptorok, a sorokban pedig azok az egyéb kemoimmun fehérjék helyezkednek el, amelyekhez a kapcsolódó vegyületek átfedése több, mint két vegyületből áll (zöld kiemelés).

A kevés szubsztrát és ligand átfedés azzal magyarázható, hogy a kutatók egy adott fehérjével kapcsolatban csak az adott fehérjecsaládra jellemző vegyületekkel végeznek kísérleteket, nem terjesztik ki kísérleteiket más kemoimmun fehérjékkel kapcsolatos vegyületekre. Ez a jelenség pedig igen megnehezíti a kemoimmun fehérjék szubsztrát illetve ligand átfedésének vizsgálatát. E mellett nagyon kevés adatbázisban írnak arról, hogy egy adott vegyületet biztosan nem ismer fel egy fehérje. Pedig ezek az adatok is nagyon sokat segítenének a fehérjék szubsztrát és ligand adatainak vizsgálatában.

4. Összefoglalás

Az általunk használt adatbázisok, a PubChem BioAssay adatbázis, a Metabolism and Transport Database (Metabase) és a Comparative Toxicogenomics Database (CTD) adathalmazai kevés olyan vegyületet tartalmaznak, amelyeket a sejt idegen anyagok elleni védekezésének több fázisából való fehérjék ismernek fel. A rendelkezésre álló adatok továbbra is arra utalnak, hogy a kemoimmun fehérjék, melyek a sejtet védik az idegen anyagoktól, rendszerben működnek.

A fehérjék együttműködésének bizonyításához jobb adathalmazt nyújtana az adatbázisok adatainak egyesítése, melyet az általunk fejlesztett DrubDB adatbázis fog tartalmazni.

Emellett az irodalomból is szándékozunk újabb adatokat keresni, hogy kiegészítsük az egyesített adathalmazt.

Az adatok egyik hiányossága az, hogy a fehérjékhez kapcsolódó vegyületek halmazai alig fednek át. Ezt az átfedést tervezi a kutatócsoport *in vitro* kísérletekkel növelni úgy, hogy olyan vegyületekkel végez kemoimmun fehérjékkel szubsztrát illetve ligand kötési kísérleteket, amelyeket még nem vizsgáltak meg az adott fehérjével kapcsolatban.