Biológiai hatás előrejelzése heterogén információforrások felhasználásával

Doktori tézisek

Arany Ádám

Semmelweis Egyetem

Gyógyszertudományok Doktori Iskola

Témavezető: Dr. Mátyus Péter, MTA doktora, egyetemi tanár

Hivatalos bírálók: Dr. Horváth Gábor, PhD, címzetes egyetemi tanár
Dr. Tóthfalusi László, PhD, egyetemi docens

Szigorlati bizottság elnöke:
Dr. Klebovich Imre, MTA doktora, egyetemi tanár

Szigorlati bizottság tagjai: Dr. Őrfi László, PhD, egyetemi docens
Dr. Paál Tamás, CSc, egyetemi tanár

Budapest,

2016

1 Bevezetés

Az utóbbi évtizedek trendjei egyértelműen jelzik, hogy a gyógyszerkutatás és fejlesztés eredményessége még a folyamatosan növekvő költségek mellett sem tartható, így a gyógyszeripar új gyógyszerfejlesztési irányok felderítésére kényszerül. Ezeket a problémákat tovább súlyosbítja a lejáró szabadalmak következtében fellépő árverseny és az egyre komplexebb szabályozói környezet. Doktori munkám során két témát vizsgáltam részletesen, melyek forradalmasították a gyógyszeripart és új megoldásokat kínálnak a kutatás-fejlesztési krízisre. Ezek a témák a gyógyszer-újrapozícionálás és a személyre szabott gyógyászat.

A gyógyszer-újrapozícionálás egy költséghatékony és alacsony kockázatú stratégia, melynek célja már engedéllyel rendelkező gyógyszerek új terápiás indikációkban történő felhasználása. A gépi tanulás szemszögéből nézve a gyógyszer-újrapozícionálás legfontosabb jellemzője a lehetséges információforrások magas száma. Kutatómunkám során célom a különböző gépi tanulási módszerek alkalmazhatóságának vizsgálata

volt a gyógyszer-újrapozícionálás területén. Első lépésként létrehoztam egy hat információforrást (három kémiai struktúra leíró, két mellékhatás alapú leíró és egy célpont profil) magában foglaló teljesítményvizsgálatra szolgáló (benchmark) adathalmazt és egy kiértékelésre (gold standard-ként) szolgáló gyógyszer-indikáció adatbázist. Elsőként egy új adatfúziós technika, a Kernel Fusion Repositioning (KFR) módszertan kifejlesztésében és egy referencia módszerrel történő összehasonlításában vettem részt. Amellett, hogy meghatározó szerepet játszottam a KFR keretrendszer megtervezésében és implementálásában, illusztráltam a KFR működését egy a Parkinson kórhoz kapcsolódó újrapozícionálási feladaton is. Egy újszerű többcélpontú predikciós módszer egyik fejlesztőjeként alkalmaztam a módszert a gyógyszer újrapozícionálási feladatra és elemeztem a többcélpontú tanulás hatását a becslési pontosságra.

Második fő témám a személyre szabott gyógyászathoz kapcsolódik. A személyre szabott gyógyászat nem csak a pácienseknek segít, hogy megkaphassák a számukra legmegfelelőbb kezelést, hanem magához a

gyógyszerfejlesztéshez is fontos információkat szolgáltathat. A farmakokinetika személyek közötti eltérésének előrejelzése segítheti a gyógyszerkutatókat és az orvosokat az alkalmazott dózis személyre szabott megválasztásában, ezzel maximalizálva a hatásosságot és minimalizálva a mellékhatásokat, valamint a toxicitást. Munkám során részt vettem egy kutatásban, amely a nagy dózisban alkalmazott metotrexát farmakokinetikájának személyenkénti eltéréseit vizsgálta. Célom egy új többváltozós Bayes-i statisztikai technika alkalmazása volt prediktív variánsok azonosítására, valamint a kapott eredmények összehasonlítása a korábbi frekventista analízis eredményeivel.

2 Célkitűzések

Munkám során az alábbi célkitűzéseim voltak:

- Egy heterogén információforrások széles körét használó, gyógyszervegyületek biológiai hatásának előrejelzésére képes adatfúziós módszer megtervezése, informatikai megvalósítása és validálása.
- A fenti adatfúziós módszer alkalmazása Parkinson kór terápiája szempontjából releváns gyógyszerek prioritizálására, valamint a metodológia képességének kiértékelése gyógyszerkutatási szempontból, különös tekintettel a gyógyszer újrapozícionálás területén.
- Egy újszerű, több célponton mért aktivitás együttes becslésére képes mátrix faktorizáción alapuló módszer kifejlesztése és egy klasszikus egyváltozós módszerrel történő összehasonlítása.

- Egy Bayes-i rendszeralapú többváltozós statisztikai technika adaptálása és alkalmazása nagy dózisú metotrexát farmakokinetikájának személyenkénti eltérése szempontjából releváns genetikai variánsok azonosítására.

3 Módszerek

Elsőként az alábbi vegyületek leírására alkalmas információforrásokat hoztuk létre: Molecular Access Keys (MACCS); molecular connectivity, shape and electrotopological fingerprint (MOLCONN-Z); 3D farmakofór alapú fingerprint; mellékhatás előfordulás és gyakoriság; valamint ismert gyógyszer-célpont interakciók. Minden információforráshoz definiáltuk a vegyületek vektoriális reprezentációját. Definiáltunk továbbá hasonlósági függvényeket a páronkénti hasonlóságokat tartalmazó kernel előállításához a kernel alapú módszerek számára. Tanimoto hasonlóságot használtunk minden bináris jegyeket tartalmazó információforrás esetén és koszinusz távolságot valós értékű jegyek esetén.

Kiszámítottuk a páronkénti hasonlóságok Spearman korrelációját, hogy az adatforrások közös információtartalmát meghatározzuk, tehát kiértékeljük azok komplementaritását.

Az egyosztályos SVM-et választottuk modellosztálynak a különböző adatfúziós módszerek összehasonlításához. A

késői adatfúziós megközelítés esetén az egyes információforrásokhoz külön-külön kiszámítottuk a prioritizáció eredményét, majd Borda protokoll segítségével kiszámítottuk a konszenzusos becslést. A predikciós teljesítmény mérésére az AUC[ROC], AUC[CROC(exp)], BEDROC és a fix küszöbérték melletti *szenzitivitás* valamint *specificitás* értékeket használtuk. Az ROC és CROC paramétere (early discovery focus) α=20.0 volt. Két küszöbértéket használtunk mind a szenzitivitás mind a specificitás esetén, a top25-öt és a top100-at. A különböző eljárások predikciós teljesítményének összehasonlítása során a széles körben elfogadott ATC osztályozást használtuk.

Mivel a Szerves Vegytani Intézet kutatócsoportja több Parkinson kór terápiájával kapcsolatos kutatást is végez, az adatfúziós módszert alkalmaztuk Parkinson kór szempontjából releváns újrapozícionálási jelöltek prioritizálására. Továbbá az így kifejlesztett módszertan általánosan használható újrapozícionálási feladatok széles körében.

Egyik kifejlesztője voltam a Macau nevű Bayes-i mátrix faktorizációs módszernek melyet a gyógyszer-indikációs prioritizációs feladatra is alkalmaztunk. Jelen kutatás keretében fő célom a módszer alkalmassá tétele volt arra az esetre, amikor nincsenek negatív mintáink. Sokkal nagyobb annak a valószínűsége, hogy az adathalmazból hiányzó asszociáció a valóságban nem létezik, mint annak a valószínűsége, hogy egy még fel nem fedezett létező asszociációról van szó. Ezért azt az elfogadott stratégiát alkalmaztuk, hogy kiválasztottuk a hiányzó asszociációk egy véletlen részhalmazát, és ezt használtuk negatív mintahalmazként. Az egyes becslési feladatok, itt ATC osztályok, közötti szinergista hatás (un. multi-task hatás) mérésére oszloponkénti független ridge regressziót használtunk referencia modellként.

A második fő kutatási témám a személyre szabott gyógyászathoz kapcsolódik. Ennek során 29 előzetesen kiválasztott, az ABCB1, ABCC1, ABCC2, ABCC3, ABCC10, ABCG2, GGH, SLC19A1, NR1I2 génekben található egynukleotidos polimorfizmusokat (SNP) vizsgáltunk. A géneket irodalmi előismeret és a

polimorfizmusok becsült funkcionalitása alapján választottuk ki. Az örökítőanyag vérből történő izolálása Qiagen izolációs kitekkel történt (QIAmp DNA Blood Maxi Kit / QIAmp DNA Blood Midi Kit; Qiagen, Hilden, Germany). A szekvenálásokhoz a GenomeLab SNPstream genotipizálási platformot (Beckman Coulter) használtuk.

Mivel a klinikai adatok 59 beteg 551 metotrexát kezelésének adatait tartalmazták, ahol a kezelések száma személyenként eltérő, szükségszerű volt a kezelésenként rögzített változók összevonása. A kezelésenkénti értékek mediánjának használatát javasoltam mint beteg szintű aggregált értéket, és ezt a konvenciót használtuk mind a frekventista, mind a Bayes-i elemzés során. A BN-BMLA (Bayesian Network based Bayesian Multilevel Analysis) eljárás használatának előfeltétele volt a folytonos változók diszkretizálása. Medián alapú diszkretizációt javasoltunk, így biztosítva a kiegyenlített tanítóhalmazt. A több mint két értékű változókat bináris változókká konvertáltam, hogy ezzel csökkentsük a modell komplexitását. A máj- és csontvelő toxicitást jellemző eredeti változóknak négy lehetséges értéke volt. Az adathalmaz kiegyenlítettségét

szem előtt tartva csoportosítottuk a változók értékeit, a csontvelő toxicitás esetén 1 illetve 2-4 súlyossági szint, a májtoxicitás esetén pedig 1-2 illetve 3-4 súlyossági szint szerint.

4 Eredmények

A gyógyszer-újrapozícionáláshoz kötődő kutatásunk célja az volt, hogy összehasonlítsuk az újonnan fejlesztett köztes fúziós módszer, a KFR keretrendszer predikciós teljesítményét egy referenciának tekintett késői adatfúziós eljárás, a Borda protokoll teljesítményével. Az összehasonlítás során az egy osztályos szupport vektor gépeket használtuk modellosztályként, és az ATC hierarchia négyes szintje képezte a predikciós célt. Az eredmények szerint a köztes adatfúziós eljárás minden esetben jobb predikciós teljesítménnyel rendelkezik, de különösen a korai felderítési metrikák szerint mérve.

Kiszámítottuk az egyedi adatforrások felhasználásával kapott sorrendek és a két fúziós eljárás által adott kimenetek Spearman rangkorrelációját, azzal a céllal, hogy megvizsgálhassuk a KFR eljárás előnyeit. Figyelemre méltó és egyben kutatómunkám fontos eredménye, hogy a Borda protokoll használata esetén az egyes információforrások relatív hozzájárulása meglehetősen stabil a vizsgált gyógyszercsoporttól

függetlenül, míg a kernel fúzió alapú eljárás adaptív, lekérdezés-specifikus tulajdonságokat mutat.

Az elfogadható tanítóhalmaz definiálására egy lekérdezés kompaktságát jellemző kritériumot javasoltunk. A javasolt megoldás a halmazon belüli átlagos hasonlóságon (intraset similarity, ISS) alapul, melyet a teljes gyógyszerhalmazon számított páronkénti hasonlóság átlagával (universe average similarity, UAS) normalizáltunk. Az így kapott ISS/UAS értékek jó korrelációt mutatnak az AUC értékekkel: minden egy feletti ISS/UAS értékkel rendelkező ATC osztály 0.5 feletti AUC értékkel rendelkezik.

A továbbiakban elemeztük a KFR keretrendszer négy Parkinson kórral kapcsolatos lekérdezésre adott válaszát. A négy vizsgált lekérdezés rendre: neuroprotektív szerek, dopaminerg szerek, muszkarinos agonisták és NMDA antagonisták. Két, a 2013-as, valamint a 2016-os PubMed adatbázison alapuló szűrőt használtunk, hogy eltávolítsuk a sorrendezési eredményből azokat a vegyületeket, amelyek nem fordulnak elő a Parkinson kórra utaló kifejezésekkel együttesen az irodalomban. Az alkalmazott

PubMed lekérdezés az alábbi: *(„Parkinson” OR „Parkinson's Disease” OR „PD”) AND INN.* Prospektív nézőpontból vizsgálva érdemes megjegyezni, hogy néhány előkelő helyre rangsorolt vegyület együttes előfordulási számai jelentősen növekedtek, míg más esetekben a növekedés nem volt számottevő, vagy a számok egyáltalán nem változtak. Ezen három viselkedés jól megfeleltethető három csoportnak: lehetséges újrapozícionálási jelölteknek, már ismert gyógyszereknek, illetve téves pozitívoknak.

Továbbá kiértékeltük a többcélpontú Macau módszert mint újonnan fejlesztett mátrix faktorizációs eljárást, és megmutattuk, hogy a felhasznált információforrástól függetlenül a Macau predikciós teljesítménye jelentősen magasabb volt, mint a ridge regresszióé.

A személyre szabott gyógyászathoz kapcsolódó munkám során többféle módon is adaptáltam és sikerrel alkalmaztam a BN-BMLA-t, egy új többváltozós Bayes-statisztikai technikát, a metotrexát farmakokinetikájának és toxicitásának személyek közötti eltérését magyarázó genetikai polimorfizmusok azonosítása céljából.

Minthogy minden farmakokinetikai és toxicitással kapcsolatos mérés terápiás blokkonként történt, ezen változók páciens szintre történő aggregálására tettünk javaslatot a blokkonkénti értékek mediánjának kiszámításával. Ezt a konvenciót használtuk mind a frekventista, mind a Bayes-i analízis során. A BN-BMLA eljárás használatának előfeltétele volt, hogy a folytonos változókat diszkretizáljuk. Medián alapú diszkretizációt javasoltunk, így biztosítva a kiegyenlített tanítóhalmazt. Az NR1I2 gén esetében a BN-BMLA sikeresen azonosított és kiválasztott egyet csatolt polimorfizmusok egy halmazából, amely arra utal, hogy a módszer képes különbséget tenni direkt és tranzitív relációk között. Amíg néhány SNP-et mindkét módszertannal sikeresen azonosítottunk (két SNP-et az AUC célváltozó és egyet a csontvelő toxicitás esetén), addig további gyengébb jelöltek merültek fel mind a frekventista mind a Bayes-i analízis során. Eredményeink azt mutatják, hogy a különböző módszerek konszenzusos használata követendő gyakorlat az asszociációk robosztus azonosítására.

5 Következtetések

Kutatásom eredményei alapján az alábbi következtetések vonhatók le:

- Kutatásom során jelentős mértékben hozzájárultam egy új köztes adatfúziós eljárás, a KFR keretrendszer fejlesztéséhez. A kutatás során végzett tesztjeink megmutatták, hogy a KFR magasabb predikciós teljesítménnyel rendelkezik a késői fúziós referencia eljárásnál, a Borda protokollnál az általunk vizsgált gyógyszer-újrapozícionálási problémán. Ezt alátámasztják a kapott AUC értékek és minden korai detektálást mérő metrika.

- A módszerek működésének jobb megértése érdekében elemeztem az egyedi információforrásokon alapuló és az adatfúziós technikákkal kapott prioririzálási eredmények Spearman rangkorrelációját, és úgy találtam, hogy a KFR adaptív, lekérdezés-specifikus

tulajdonságokkal rendelkezik. Ez fontos és előnyös tulajdonság, mely lehetővé teszi a módszer farmakológiai csoportok és vegyületosztályok széles körére történő alkalmazását.

- A kísérletek rámutattak egy anomáliára, amely nagyon heterogén lekérdezések esetén jelentkezett: a lekérdezés elemei nem a prioritizálási eredmény elején szerepeltek. Ezekben az esetekben a módszer predikciós teljesítménye is gyakran nagyon alacsony volt. Javaslatot tettem egy a lekérdezés elemeinek átlagos páronkénti hasonlóságát mérő kritériumra, mellyel ezek az esetek kiszűrhetők és megmutattam, hogy ez a kritérium alkalmas az alacsony predikciós teljesítményt eredményező lekérdezések azonosítására.

- Alkalmaztam a KFR keretrendszert a Parkinson kór szempontjából potenciálisan releváns újrapozícionálási jelöltek azonosítására, melynek során sikerült kiválasztanom olyan vegyületeket, amelyek az irodalomban a Parkinson kórral

együttesen fordulnak elő. Ezeket az eredményeket prospektív kiértékelés során tovább validáltam, valamint vázoltam egy új számítógépes eljárásokra támaszkodó munkafolyamatot gyógyszer-újrapozícionálási jelöltek azonosítására.

- Részt vettem egy újszerű Bayes-i mátrix faktorizációs eljárás, a Macau fejlesztésében, amely több célváltozó együttes becslésére képes. Jelen kutatás során összehasonlítottam a Macau-t egy egyváltozós eljárással, a ridge regresszióval, és minden információforrás esetén pontosabbnak találtam.

- Munkám során sikeresen adaptáltam és alkalmaztam a BN-BMLA-t, egy újszerű, többváltozós Bayes-statisztikai eljárást, hogy megerősítsem, és további jelöltekkel kiegészítsem a nagy dózisú metotrexát kezelés farmakokinetikájához kapcsolódó korábbi frekventista eredményeket. Az eredmények alátámasztják, hogy a Bayes-i és a frekventista eljárások kombinálása hatékony stratégia

asszociációk robosztus felderítéséhez. Továbbá megállapítható, hogy a BN-BMLA használata különösen előnyös interakciók és redundanciák, mint például kapcsoltságban álló genetikai polimorfizmusok jelenléte esetén.

A disszertációban felhasznált saját közlemények

Hegyi M, Arany A, Semsei AF, Csordas K, Eipel O, Gezsi A, Kutszegi N, Csoka M, Muller J, Erdelyi DJ, Antal P, Szalai C, and Kovacs GT, *Pharmacogenetic analysis of high-dose methotrexate treatment in children with osteosarcoma.* Oncotarget, 2016. **IF = 5.008***

Temesi G, Bolgar B, Arany A, Szalai C, Antal P, and Matyus P, *Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy.* Future Med Chem, 2014. **6**(5): p. 563-75. IF = **3.744**

Arany A, Bolgar B, Balogh B, Antal P, and Matyus P, *Multi-aspect candidates for repositioning: data fusion*

methods using heterogeneous information sources. Curr Med Chem, 2013. **20**(1): p. 95-107. **IF = 3.715**

Bolgar B, Arany A, Temesi G, Balogh B, Antal P, and Matyus P, *Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies.* Curr Top Med Chem, 2013. **13**(18): p. 2337-63. **IF = 3.453**