DOI:10.14753/SE.2015.1757

Hatóanyagok és terápiás célpontok vizsgálata adatmérnöki eszközökkel

Doktori értekezés

Temesi Gergely Botond

Semmelweis Egyetem
Molekuláris orvostudományok Doktori Iskola

Témavezető:
Dr. Szalai Csaba, az MTA doktora, egyetemi tanár

Hivatalos bírálók:
Dr. Gulyás-Kovács Attila, Ph.D.
Dr. Rónai Zsolt, Ph.D., egyetemi adjunktus

Szigorlati bizottság elnöke:
Dr. Vásárhelyi Barna, az MTA doktora, egyetemi tanár

Szigorlati bizottság tagjai:
Dr. Cserző Miklós, Ph.D., tudományos főmunkatárs
Dr. Szakács Orsolya, Ph.D.

Budapest
2014

DOI:10.14753/SE.2015.1757

Tartalom

1 Rövidítések jegyzéke 5

2 Bevezetés és irodalmi háttér 8

2.1 A harmadik információrobbanás 13

2.2 A vizsgált tárgyterületek és kihívásaik 17

2.2.1 Az asztma és genetikai hátterének vizsgálata a poszt-genom korszakban 17

2.2.2 A gyógyszeripar kihívásai és a számítógépes hatóanyag újrapozicionálás 26

2.3 A tárgyterületek adatai és információi 30

2.3.1 A molekuláris orvostudományok adatai 30

2.3.2 A farmakológiai adatvagyon 36

2.4 A tárgyterületek adatmérnöki módszerei 41

2.4.1 Molekuláris biológiai adatelemzési módszerek 41

2.4.2 Virtuális szűrési technikák adatelemzési módszerei 47

3 Célkitűzések 50

3.1 Asztma genetikai hátterének vizsgálata adatmérnöki eszközökkel 50

3.2 A farmakológiai információtömeg kiaknázása feldúsulás elemzéssel 51

4 Módszerek 52

4.1 Asztma genetika 52

4.1.1 Genotipizált populáció: betegek és kontrollok 52

4.1.2 Indukált köpet expressziós vizsgálat 53

4.1.3 Kandidáns gének és polimorfizmusok kiválasztása és genotipizálása 55

4.1.4 Indukált köpet vizsgálat, RNA izoláció és génexpressziós mérés 64

4.1.5 Frekventista statiszikai elemzés 64

4.1.6 Bayes-háló alapú Bayesi többszintű relevancia elemzés 65

DOI:10.14753/SE.2015.1757

4.2 Feldúsulás elemzés 70
4.2.1 Szisztematikus hatóanyag újrapozicionálás adat fúziós módszerekkel 70
4.2.2 Gene Set Enrichment Analysis 71
4.2.3 Referencia adatbázis előkészítése 74
5 Eredmények 77
5.1 Asztma genetika 77
5.1.1 Kísérlettervező rendszer (TIGER) 77
5.1.2 Genotipizálási eredmények 80
5.1.3 Indukált köpet vizsgálat 89
5.1.4 Eredmények összegzése Fisher módszerrel 90
5.1.5 Bayes-háló alapú Bayesi többszintű relevancia elemzés 91
5.2 Feldúsulás elemzés 102
5.2.1 Compound Set Enrichment Analyis (CSEA) tesztrendszer 102
5.2.2 Az amantadine esettanulmány 104
6 Megbeszélés 108
6.1 Asztma genetika 108
6.2 Feldúsulás elemzés 114
6.2.1 Az esettanulmány megbeszélése 114
6.2.2 Az információ újrahasznosítás 115
7 Következtetések 119
7.1 Kísérlettervező rendszer 119
7.2 *SCIN* 119
7.3 PPARGC1B 119
7.4 ITLN1 120

DOI:10.14753/SE.2015.1757

7.5 LGMN 120
7.6 Humán asztma és egér asztmamodell expresszió 120
7.7 Compund Set Enrichment Analysis (Hatóanyag feldúsulás elemzés) 120
7.8 Információ újrahasznosítás a CSEA módszertannal 121
8 Összefoglalás 122
9 Summary 123
10 Irodalomjegyzék 124
11 Saját publikációk jegyzéke 146
11.1 Asztma genetika 146
11.2 Feldúsulás elemzés 146
11.3 Egyéb közlemény 146
12 Köszönetnyilvánítás 147

DOI:10.14753/SE.2015.1757

1 Rövidítések jegyzéke

Rövidítés	Angolul	Magyarul (amennyiben létezik)
ANOVA	Analysis Of Variance	Variancia analízis
BAL	Bronchoalveolar Lavage	Tüdőmosó folyadék
BN-BMLA	Bayesian Network based Bayesian Multilevel Analysis	Bayes-háló alapú Bayesi többszintű relevancia elemzés
CA	Allergic Conjunctivitis	Allergiás kötőhártya gyulladás
CD-CV	Common Disease – Common Variant	Gyakori betegség – gyakori variáns hipotézis
CD-RV	Common Disease – Rare Variant	Gyakori betegség – ritka variáns hipotézis
CEU	Reference population, Utah residents with ancestry from Europe	Referencia populáció, Utah lakosai európai felmenőkkel
CI	Confidence Interval	Konfidencia intervallum
CNV	Copy Number Variation	Ismétlésszám változás
CSEA	Compound Set Enrichment Analysis	Vegyület halmaz feldúsulás elemzés
DNA / DNS	Deoxyribonucleic acid	Dezoxiribonukleinsav
EMA	European Medicines Agency	Európai Gyógyszerügyi Hivatal
ES	Enrichment Score	Feldúsulási érték
FABP3	Fatty Acid Binding Protein 3	
FDA	Food And Drug Administration	Élelmiszer- és Gyógyszer-felügyeleti Hatóság
FEV1	Forced Expiration Volume in the first second	Erőltetett vitálkapacitás kilégzés az első másodpercben
FEV1/FVC%	Forced Vital Capacity Expiration percent in the first second	Erőltetett vitálkapacitás kilégzés százaléka az első másodpercben

FVC	Forced Vital Capacity	Erőltetett vitálkapacitás
GEO	Gene Expression Omnibus	
GINA	Global Initiative for Asthma	
GO	Gene Ontology	
GSEA	Gene Set Enrichment Analysis	Gén halmaz feldúsulás elemzés
GWAS	Genome Wide Association Study	Teljes genom asszociációs vizsgálat
HTS	High-throughput screening	Nagy áteresztőképességű szűrés
HWE	Hardy-Weinberg Equilibrium	Hardy-Weinberg egyensúly
INTLN1	Intelectin-1 (intestinal lactoferrin receptor)	
LD	Linkage Disequilibrium	Kapcsoltsági egyensúlytól való eltérés
LGMN	Legumain	
PCR	Polymerase Chain Reaction	Polimeráz láncreakció
LY9	Lymphocyte antigen 9	
MAT1A	Methionine Adenosyltransferase 1 Alpha	
MB	Markov Blanket	Markov-takaró
MC3	Metropolis Coupled Markov Chain Monte Carlo method	
MCMC	Markov Chain Monte Carlo method	
MBM	Markov Blanket Membership	Markov-takaróba tartozás
MBG	Markov Blanket Graph	Markov-takaró gráf
MBS	Markov Blanket Set	Markov-takaró halmaz
mRNA / mRNS	messenger RNS	Hírvivő RNS
OR	Odds Ratio	Esélyhányados
OVA	Ovalbumin	Tojásfehérje

OSGIN	Oxidative Stress Induced Growth Inhibitor	
PC20	Provocative Concentration that causes 20% fall in FEV1	A FEV1 20%-os esését előidéző koncentráció
PPARGC1B	Peroxisome Proliferator-Activated Receptor Gamma Coactivator 1-Beta	
QDF^2	Query Driven Data Fusion Framework	Kérdésvezérelt adatfúziós keretrendszer
RA	Rhinitis Allergica	Szénanátha
RNA / RNS	Ribonucleic acid	Ribonukleinsav
rs#	reference SNP number	Polimorfizmus referencia azonosító
SCIN	Scinderin (or Adseverin)	
SNP	Single Nucleotide Polymorhism	Egynukleotidos polimorfizmus
STR	Short Tandem Repeat	Rövid tandem ismétlődés
TFF1	Trefoil Factor 1 protein	
Th1	T-helper 1 cell	1-es típusú segítő T sejt
Th2	T-helper 2 cell	2-es típusú segítő T sejt
UTR	Untranslated Region	Nem transzlálódó régió
YRI	Reference population, Yoruba in Ibadan, Nigeria	Referencia populáció, Yoruba etnikum, Ibadan, Nigéria

DOI:10.14753/SE.2015.1757

2 Bevezetés és irodalmi háttér

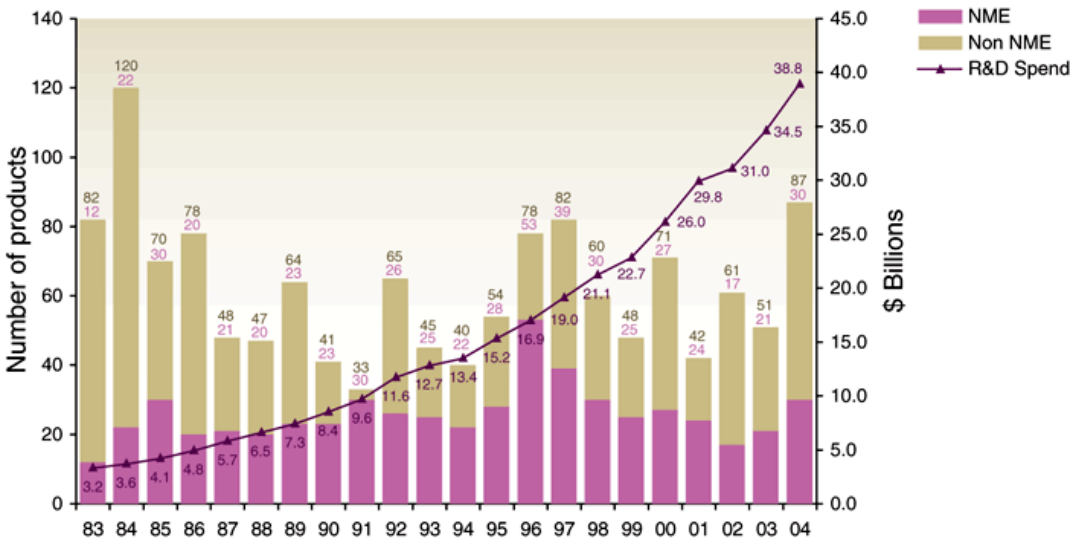
Az új évezred valóban egy új korszakot is nyitott számos tudományos, technológiai és gazdasági területen. A számítástechnika és a biológia párhuzamos fejlődése nem csak egy természetes összefonódás irányába mutat, de korunk égető globális kihívásai elengedhetetlenné is teszik, hogy a fejlődést innovatív, interdiszciplináris megközelítések termékenyítsék meg.

Az egészségügyi ellátó rendszerek gazdasági értelemben vett hatékonysága más „iparágakhoz" viszonyítva világszerte alacsony. Hiányoznak a versenyszférára jellemző erőteljes hatékonyságnövelő stratégiák: a szervezetek közti természetes szelekciós nyomás és az ügyfelek (betegek) oldalán pedig az ösztönző (pl. egészség megőrzését jutalmazó) módszerek. A konzervatív működési modellek az innovációt kevéssé ösztönzik, az intézmények rendszere jelenleg is jobban tükrözi a múlt század igényeit [1]. Emellett az öregedő nyugati társadalmakban a populáció egyre kisebb hányadának kell fedeznie egyre több ember szociális és egészségügyi terheit. Az egészségügy egyre jobban képes kitolni a várható élettartamot, de az egészségügyi problémák és kiadások 50 éves kor felett exponenciális növekedésnek indulnak és átlagosan közel ötszörösére nőnek a korábbi évekhez viszonyítva [2]. Ez a folyamat nemzetgazdasági szinten egyre növekvő egészségügyi kiadásokat jelent, nem csak abszolút értékben, de GDP arányosan is: az USA-ban pl. csak az utóbbi évtizedben 25%-al nőtt [3, 4]. A rendszer jelenlegi trendjei fenntarthatatlanok, ez az egyik legfontosabb oka a több fejlett országban megindult egészségügyi reformoknak is.

A gyógyszeripar, mint az egészségügy egy piaci alapon működő szegmense szintén egy fenntarthatatlannak tűnő üzleti modellbe sodródott. A gyógyszeripart érintő szabályok folyamatosan szigorodnak, egyre szélesebb körű klinikai tanulmányokat és egyre biztonságosabb hatóanyag profilokat követelnek meg a gyártóktól. E két folyamat eredményeként az utóbbi évtizedben évről-évre egyre drasztikusabb költségnövekedésről számolnak be a gyógyszerfejlesztő cégek (a klinikai vizsgálatok már nem ritkán a gyógyszerfejlesztés költségeinek több mint 90%-át teszik ki), miközben az engedélyezett hatóanyagok száma éves szinten az elmúlt húsz év alatt jelentősen csökkent a szigorúbb

DOI:10.14753/SE.2015.1757

követelmények miatt (1. ábra). A kevés új gyógyszer nem képes finanszírozni a következő gyógyszer generációk növekvő fejlesztési költségeit, így a fenntarthatatlan modell miatt a gyógyszeripar szintén komoly változások előtt áll [5].

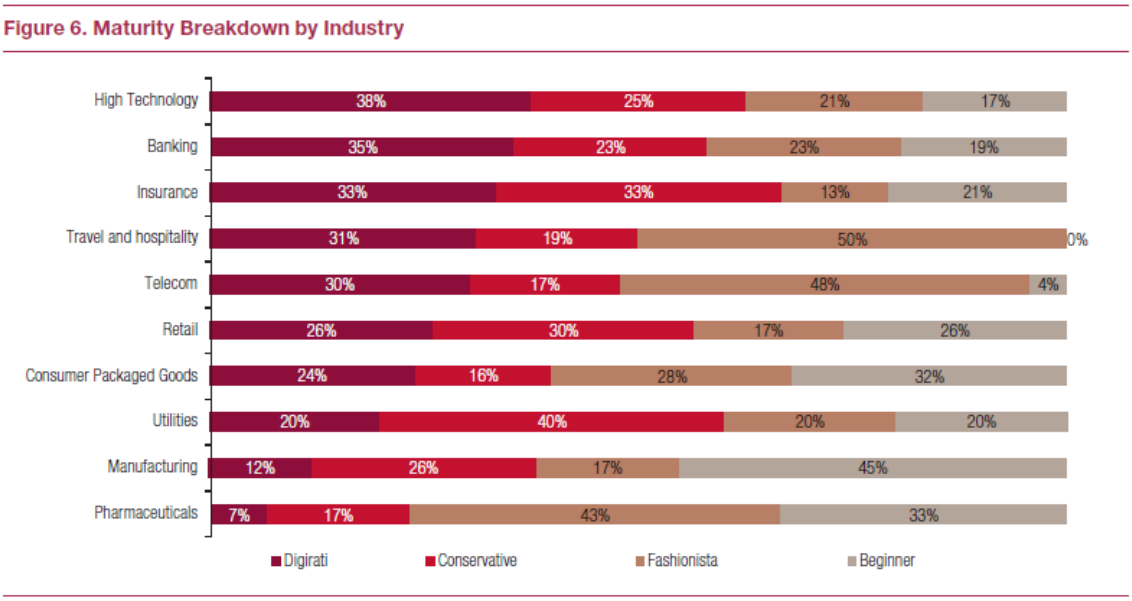


1. ábra - A Nature 2007-ben pulikált egy tanulmányt a hatóanyagok fejlesztési költségének alakulásáról és az éves szinten piacra jutó hatóanyagok számának trendjéről [5]. A költségek meredek növekedést, míg az engedélyezett hatóanyagok inkább csökkenést mutatnak.

Mindeközben az élettudományok, és elsősorban a molekuláris élettudományok az elmúlt 50 évben hatalmas fejlődésen mentek keresztül, melynek szimbolikus mérföldköve a humán genom megfejtése szinte pontosan az ezredfordulón történt meg. Az élettani folyamatok molekuláris szintű megismerése az egészséges és beteg állapotok rendszerszintű megértésének ígéretét hozta, de a poszt-genom korszak, vagyis a humán genom megfejtése utáni időszak a vártnál több kihívással is járt. A genomika eredményeinek transzlációja a klinikum szintjére csak egy évtizeddel később kezdi éreztetni a hatását [6]. Hasonlóan, a gyógyszeripar oldalán is a vártnál tovább tart az eredmények felhasználása a gyógyszerfejlesztési gyakorlatban: kevés a genetikai sajátosságokat célzó hatóanyag és alig van példa olyan hatóanyagra, melyet a humán genom megismerése alapján fejlesztettek ki [7]. Számos olyan nagyreményű kezdeményezés és vízió indult útjára az elmúlt évtizedekben, melyek a molekuláris orvostudományok vívmányaiból merítve próbálnak válaszokat keresni az egészségügy és a gyógyszeripar kihívásaira (pl. personalized medicine, systems medicine, translational medicine, precision medicine, rational drug

design, stb.). A molekuláris technikák fejlődésének egy érdekes tudománytörténeti aspektusa a nagy áteresztőképességű méréstechnikák térhódítása, melyek mind orvosbiológiai, mind farmakológiai területen egyre nagyobb mennyiségű digitális tudás előállításával járnak.

A számítástechnika elmúlt ötven éves fejlődése még talán az élettudományoknál is nagyobb jelentőségű volt, a világon elérhető digitális adat- és információtömeg a hatvanas évek óta exponenciálisan növekszik. A gyorsulás üteme az utóbbi évtizedben lett csak igazán szembetűnő és az eddig nem tapasztalt mennyiségű adat kezelése már új módszereket is igényel („Big Data"). Az informatikai fejlődés kiaknázása iparáganként erősen eltérő: a konzervatív egészségügy és gyógyszeripar - a statisztikák szerint - az iparágak között az utolsók között van (2. ábra), miközben a digitális fejlettség a hatékonyabb működéssel nagyon erősen korrelál [8].



2. ábra - A Capgemini Consulting and MIT's Center for Digital Business tanulmánya a digitális fejlettségről [8]. A digitális fejlettség mutatói a high-tech és pénzügyi szektorokban a legmagasabbak, míg a gyógyszeriparban a legalacsonyabbak.

Az egészségügy és a gyógyszeripar fenntarthatatlan modelljeiből való kitörés egyik legfontosabb eszköze lehet az információ technológiák fokozott kiaknázása, melyet az imént bemutatott hatások szinergiája ösztönöz: jelen van az innovációra ösztönző gazdasági

nyomás, a területek adatintenzív eszközeinek terjedése, a másik oldalon pedig rendelkezésre állnak a gyors ütemben fejlődő információ technológiák és más iparágak jó gyakorlatai. Ezeket a trendeket felismerve PhD munkám során olyan információ technológiai eszközök kutatását és fejlesztését tűztem ki célul, amelyek a világ küszöbön álló egészségügyi és gyógyszeripari változásaira reagálnak, illetve képesek segíteni azokat.

A korábbi Mérnök Informatikus MSc. és Orvosbiológiai Mérnök MSc. diplomamunkáim során molekuláris biológiai kutatások támogatásához fejlesztettem szoftveres eszközöket, így szakmai hátterem jelentős része mérnöki tudományokon alapul. A PhD kutatásaimat is egy interdiszciplináris (Budapesti Műszaki és Gazdaságtudományi Egyetem - Semmelweis Egyetem) kutatócsoportban végeztem, melynek során egy molekuláris orvostudományi és egy farmakológiai problémakört vizsgáltam meg eltérő adatmérnöki eszközökkel, így a teljes dolgozat e három diszciplína metszetében helyezkedik el. A munka gerincét a modern információ technológiák adják, de míg az első esetben eszközként használtam fel őket egy felfedező kutatáshoz és az elért új orvosbiológiai eredményeken van hangsúly, addig a második esetben egy technológiai fejlesztés volt a cél, és így maga a módszertan mutat fel új elemeket farmakológiai területen. A két szakterület sok ponton érintkezik, de alapvetően különálló kutatási területet képviselnek, így e munka is két párhuzamosan kifejtett tematikai ív mentén épül fel.

Az első témában („Asztma genetika”) az asztma és gyulladásos megbetegedések molekuláris patomechanizmusát, elsősorban genetikai hátterét vizsgáltam. Ebben az esetben egy teljes orvosbiológiai felfedező kutatás kivitelezése volt célom: kezdve - a korábbi állatmodell kutatások alapján - a kísérlet megtervezésétől, a humán biobank gyűjtésen és a több szintű omikai vizsgálatok elvégzésén át, modern információ technológiai módszerekkel történő kiértékelésig. Itt az információ technológiai módszerek eszközként jelentek meg a kísérlet megtervezésében és a tárgyterület adatainak rendszerszintű modellezésében, és elsősorban korábbi fejlesztések eredményeit használtam fel.

DOI:10.14753/SE.2015.1757

A második téma („Feldúsulás elemzés”) fókusza ezzel szemben egy új információ technológiai módszertani fejlesztése volt farmakológiai területen. Ebben az esetben is egy komplex tárgyterület adatainak integrálása és elemzése volt cél, de itt az új módszertan fejlesztésén volt a hangsúly, amelyet egy tárgyterületi problémán teszteltem is. A technikával ismert hatóanyagok heterogén tulajdonságai, és mérési eredményei alapján lehet előre megbecsülni egy új vegyület különböző tulajdonságait, így például lehetséges indikációkat. A módszer teszteléséhez az amantadinet vizsgáltam meg, mely egy jól ismert, több indikációval rendelkező hatóanyag, így jól tudtuk ellenőrizni a módszer hatékonyságát és korlátait.

DOI:10.14753/SE.2015.1757

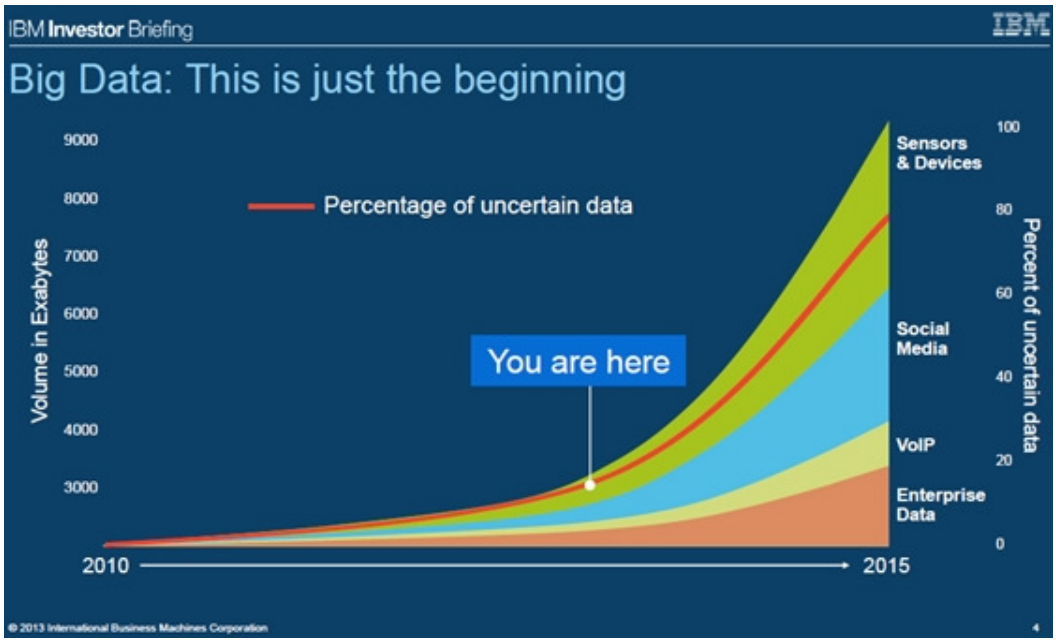
2.1 A harmadik információrobbanás

Az emberiség által előállított és tárolt adatmennyiség az ezredfordulóra robbanásszerű növekedésnek indult. Bár a gondolat pontos eredete ismeretlen, 2006 óta egyre többen hirdetik, hogy „az adat a 21. század olaja”. A robbanást több tudományterület és eltérő diszciplína párhuzamos fejlődése és egyfajta kritikus tömeg elérése indította be. A jelenség gerincét a számítástechnika fejlődése adja, melynek eredményeként ma a világ jelenségei gyakorlati értelemben szinte korlátlanul mérhetőek, leírhatóak, tárolhatóak, megoszthatóak és elemezhetőek.

Az első információrobbanás kezdetét (a szimbolikus írásrendszerek megjelenését) időszámításunk előtt 5000 körülire, a második információrobbanás kezdetét (a nyomtatott könyvek megjelenését, vagyis a Gutenberg-galaxis kezdetét) a 15. század végére, a harmadik információs robbanás kezdetét az 1960-as évek elejére datálják [9]. A harmadik információrobbanás alapját a szilícium alapú félvezető technológiák rohamos fejlődése adta. Gordon Moore, az Intel Inc. (a világ legnagyobb mikroprocesszor gyártója) egyik alapítójának nevéhez fűződik a híres „Moore-törvény”, mellyel 1965-ben megjósolta a fejlődés trendjét [10]. A jóslat szerint az integrált áramkörökbe olcsón beépíthető tranzisztorok száma 2 évente (később 18 hónapra pontosították) megduplázódik, mely egyben használható a számítási teljesítmény növekedésének a becslésére is. Bár az eredeti jóslat kizárólag az integrált áramkörökre vonatkozik, tágabb érelemben a számítástechnika több más mérőszáma is követi az eredetileg kimondott exponenciális trendet (pl. memóriák vagy diszkek mérete, hálózati áteresztőképesség, stb.). A törvény lényegében a mai napig igaznak bizonyul, sőt egyes jövőkutatók szerint akár még évszázadokig is fennmaradhat. Ezek a trendek nem csak a folyamatosan termelődő adatok mennyiségének több nagyságrendű növekedését okozták az utóbbi évtizedben, de megváltozott az adatok jellege és az adatok tipikus életciklusa is. Míg az ezredfordulóig a személyi számítógépektől a nagyvállalati rendszerekig mindenki elsősorban strukturált adatokat tárolt és rendszeres elektronikus nagytakarításokat végzett, mára egyre jellemzőbbé válik a strukturált (pl. nyilvántartások, pénzügyi adatok) és strukturálatlan (pl. hang, kép, szabadszöveges információ, nagy áteresztőképességű molekuláris mérések) adatok korlátlan megőrzése. A

DOI:10.14753/SE.2015.1757

„Big Data” kifejezés első megjelenését általában John R. Mashey, a Silicon Graphics Inc. (későbbi nevén SGI Inc., a kilencvenes évek meghatározó szuperszámítógép fejlesztője) egy vezető mérnökének a nevéhez kötik, aki egy 1999-es előadásában az exponenciális számítástechnikai fejlődés (CPU, memória, hálózat, adat) hatásait és trendjeit összegezte [11]. Ma a Big Data az összefoglaló neve a növekvő méretű adathalmazokkal kapcsolatos új kihívásoknak, amelyek nem kezelhetőek már a korábbi, megabyte-gigabyte léptékű, strukturált adathalmazokra kifejlesztett technológiai eszközökkel, beleértve pl. az adatok gyűjtését, rendszerezését, tárolását, visszakeresését, megosztását, továbbítását, elemzését és megjelenítését (3. ábra).



3. ábra - A Big Data jelenség szemléltetése az IBM 2013-as évi befektetői tájékoztatójában [12]. A digitálisan tárolt adatok növekedése exponenciális, a strukturálatlan és bizonytalan adatok részaránya növekszik.

A digitális adattárolás talán legfontosabb korai úttörője az IBM cég volt, amely 1955-re kifejlesztette a maihoz nagyon hasonló elven működő adattároló diszkeket („hard disk drive”), majd 1961-ben először használta az „információ-robbanás” kifejezést is („The Information Explosion”, The New York Times: Section II, Advertisement). Az 1970-es évek elejére megjelentek az első adatbázis kezelő rendszerek, majd szintén egy IBM mérnök, Edgar Codd fektette le a mai napig is legszélesebb körben használt strukturált

adatkezelési keretrendszer, a relációs adatbázisok elméleti alapjait (RDBMS) [13]. Ezekben az évtizedekben az IBM kutatólaborjaiban nem csak az adattárolás hardveres és szoftveres hátterének kidolgozása kezdődött meg, de az adatfeldolgozási és adatelemzési technológiák megalapozása is (pl. Hans Peter Luhn, „Business Intelligence” [14]). A kétezres évek elejére elsősorban a méréstechnikai és adattároló eszközök árának csökkenése adott egy újabb lendületet az adattudományoknak. Egyre jellemzőbbé válik a világ jelenségeinek a folyamatos mérése és ezen adatok szinte korlátlan tárolása; így pl. áruk és eszközök mozgásának folyamatos követése; molekuláris biológiai mérések; mozgóképek és hang rögzítése; kvantummechanikai és csillagászati mérések; vagy az internet folyamatos historikus mentése, ahol az emberiség kommunikációjának egyre nagyobb hányada zajlik [15].

Az adatközpontú informatika irányába történő eltolódás az ipar és tudományok sok területén általánosságban is tetten érhető az ezredfordulótól kezdve. Az iparban a természetes szelekciós mechanizmusok a cégeket folyamatosan arra ösztönzik, hogy versenytársaiknál hatékonyabbá tegyék működésüket, és kompetitív előnyre tegyenek szert. Az ezredfordulóig az IT beruházásaik elsősorban arra fókuszáltak, hogy az üzletmenet gyorsabb, pontosabb és olcsóbb legyen, azonban addigra alapértelmezetté vált, hogy a kommunikáció elektronikus formában történik, az ügyfeleket, tranzakciókat, termékeket, logisztikát és gyártást mind informatikai rendszerekkel kezelik. Az üzletmenet automatizálásával nem lehetett további versenyelőnyhöz jutni, azonban a rendszerekben felgyülemlő adatok egy új lehetőséget nyitottak meg. Az adatok elemzésével intelligensebb döntések hozhatóak, előre láthatók piaci trendek, költség-hatékonysági kérdések egyértelműen megválaszolhatóak, egyszerűsödik a kockázatelemzés, jobban szervezhető a logisztika stb. Röviden, a döntések optimalizálására helyeződött át a hangsúly, mely egyértelműen kimutatható a globális informatikai beruházások alakulásából is [16]. A tudományos kutatások területén az adatintenzív módszerek előretörését egyenesen a felfedező módszerek paradigmaváltásának („A negyedik paradigma”-nak) nevezik [17], de ahogyan az ipari szegmensek is különböző mértékben képesek alkalmazkodni az új kihívásokhoz [8], úgy a tudományterületek is eltérő módon veszik ki részüket a

DOI:10.14753/SE.2015.1757

szemléletváltásból. A következő fejezetekben a molekuláris orvostudományokat és a farmakológiát érintő információ technológiai kihívásokat és módszereket tárgyalom részletesebben, hiszen munkám elsődleges célja ezen adatvagyon hatékonyabb kiaknázása volt.

DOI:10.14753/SE.2015.1757

2.2 A vizsgált tárgyterületek és kihívásaik

2.2.1 Az asztma és genetikai hátterének vizsgálata a poszt-genom korszakban

2.2.1.1 Asztma

Az asztma a légutak krónikus, összetett multifaktoriális betegsége, genetikai és környezeti hatások együtt alakíthatják ki. Változó mértékű, epizódokban jelentkező gyulladások jellemzik, továbbá a légutak szűkülete, bronchiális hiperreszponzivitás számos faktorral szemben, és változatos tünetek, mint köhögés, zihálás, sípoló légzés, mellkasi feszülés, és légszomj [18, 19]. Kiválthatja allergén, fertőzés, fizikai terhelés és számos más hatás; a rohamok lehetnek enyhék és életveszélyesen súlyosak is. A ventilációs zavarok spontán szűnnek vagy gyógyszeres kezelés hatására reverzibilisek.

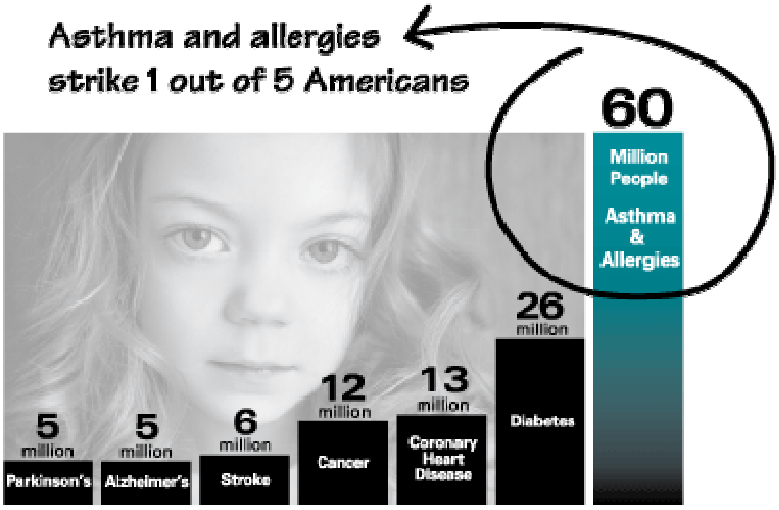
2.2.1.2 Klinikai kép

Az asztmás tünetek kiváltó oka alapján két csoportot különböztetünk meg. Az extrinsic asztma (atópiás vagy allergiás asztma) valamilyen külső allergén által kiváltott folyamatként jelentkezik (vírus, pollen, szőr, gyógyszer, stb.), általában már gyermekkorban megjelenik és magas szérum IgE szint jellemzi. A European Academy of Allergology and Clinical Immunology definíciója szerint az atópia alacsony dózisú allergének által kiváltott IgE termelésre való hajlam, olyan tünet együtteseket okozva, mint asztma, ekcéma, vagy szezonális allergiás kötőhártya gyulladás. Tehát, az atópiás vagy allergiás asztma egy részben genetikai hátterű immunológiai hiperreszponzivitás, mely az allergénnel való másodszori vagy többszöri találkozásnál alakul ki. Ezzel szemben az intrinsic asztma esetén a kiváltó ok nem pontosan ismert (pl. stressz, fizikai megterhelés, stb.), gyakran csak felnőtt korban jelentkezik, a szérum IgE szintje nem emelkedett, az immunrendszer általánosságban nem érintett a reakció aktiválásában. A betegség egyéb tünetei (pl. eozinofil szám) vagy a kezelésekre adott válaszok (pl. szteroid érzékeny, nem-érzékeny) tekintetében is nagyon heterogén, így számos endofenotípust különböztetnek meg, bár a kategorizálás a mai napig nem teljesen kiforrott. A diagnózis felállítása emiatt összetett, az anamnézis és a fizikális valamint légzésfunkció vizsgálatok után az obstrukció reverzibilitásának farmakodinámiás próbájával és a légút hiperreaktivitás ellenőrzésével

erősíthető meg. Mindezen vizsgálatok azonban nem nyújtanak pontos képet a légutakban zajló gyulladásos folyamatokról, kérdéses esetekben vagy tudományos vizsgálatokban további invazív (biopszia, bronchoalveoláris mosás) vagy non-invazív (indukált köpet, kilélegzett levegő vizsgálata) módszerek nyújthatnak segítséget. Az asztma fenntartó terápiához elsősorban inhalációs szteroidokat, antileukotriéneket, β2-agonistákat vagy anti IgE szereket; a tüneti kezeléshez szisztémás szteroidokat, antikolinergeket, valamint teofillint használnak.

2.2.1.3 Epidemiológia

A légúti megbetegedések a Föld teljes lakosságának kb. harmadát érintik, az asztma gyakorisága 3,5-20% között van, Magyarországon ez az arány 7,75% [20-22]. Becslések szerint ma 300 millióan szenvednek asztmában, minden rasszban és korcsoportban előfordul, a prevalenciája pedig évtizedek óta emelkedést mutat [23]. Az asztma ma is az egyik leggyakoribb krónikus gyermekkori betegség [24], az allergia és az asztma együtt pedig több embert érint, mint az összes többi krónikus betegség összesen (4. ábra).



4. ábra - ALLERGY FACTS AND FIGURES, Asthma and Allergy Foundation of America, (http://www.aafa.org/display.cfm?id=9&sub=30, 2014 május). Az asztma és az allergia Amerikában minden 5. embert érint, vagyis többet, mint az összes többi krónikus betegség összesen.

Bár az asztmához kapcsolt mortalitás más krónikus betegségekhez viszonyítva relatíve alacsony, a WHO előrejelzése szerint 2025-re az asztma és a krónikus obstruktív tüdőbetegség az egyik vezető halálokká léphet elő [21].

2.2.1.4 Etiológia

Az utóbbi évek kutatásai egyre több bizonyítékot találtak arra, hogy a növekvő prevalencia környezeti és életviteli hatásokkal magyarázható; így pl. egyes allergéneket (pl. házipor-atka) [25, 26], környezetszennyező anyagokat (pl. égéstermékek) [27], vagy az antibiotikum használatot [28] hozták összefüggésbe a betegség megjelenésével. A "tisztaság-hipotézis" és az asztma kapcsolatát szintén kimutatták: a korai gyermekkor baktérium és vírusfertőzései védelmet nyújtanak asztmás fenotípus kialakulásával szemben [29], vélhetően az immunrendszer Th1-Th2 egyensúlyának a T-helper 1 irányába történő elbillentésével. A megállapítást alátámasztja, hogy a baktériumexpozíció nagyobb mértéke környezettől függetlenül mérsékli az allergiás megbetegedések kockázatát (ugyanúgy farmgazdaságokban élők, mint fejlődő országok lakói esetén) [30]. Fontos azonban azt is megemlíteni, hogy több esetben kimutatták vírusok és baktériumok szerepét is az asztma kialakulásában (pl. Rhinovirus [31] vagy Chlamydia [32]), illetve már kialakult asztma súlyosbításában [33], így többen a tisztaság-hipotézis újraértelmezését javasolják [34].

2.2.1.5 Patofiziológia

Az asztma hátterében rendkívül összetett kórélettani folyamatok állnak. A betegség egyik fontos jellemzője a spontán módon vagy gyógyszeres kezelés hatására reverzibilis obstruktív ventilációs zavar, melyet a légutak gyulladásos elváltozása okoz. Majd minden esetben megfigyelhető a gyulladásos sejtbeáramlás (főleg limfociták és eozinofil granulociták), a fokozott mucus-elválasztás, a bronchusfal ödémás duzzanata és a légúti epithel sejtek hámlása. A gyulladásos folyamat idővel irreverzibilis strukturális és funkcionális károsodáshoz is vezethet („remodelling"). A korábban említett allergiás asztma hátterében az IgE termelődéshez kötött, túlzott mértékű immunaktiváció áll, ami hosszabb ideig fennmarad; míg a nem-allergiás asztmások betegsége súlyosabb lefolyású és inkább idősebb korban jelentkezik.

Az asztmások bronchoalveoláris folyadéka nagyobb számban tartalmaz hízósejtet, limfocitát, eozinofilt és aktivált makrofágot. A T limfociták által termelt citokinek (IL-4, IL-5, IL-9, IL-13) segítik elő az eozinofilek beszűrődését, a hízósejtek jelenlétét és a B limfociták izotípus váltását is, melynek eredményeként azok IgE termelő plazmasejtekké differenciálódnak. Az allergiás asztmánál az emelkedett IgE szisztémásan és lokálisan is megfigyelhető, de a legújabb kutatások szerint az emelkedett lokális IgE a patomechanizmus fontos eleme lehet a nem-allergiás asztma esetén is [35]. A légúti kötőszövetek nagyszámú hízósejtjében az IgE és az allergén keresztkötése felfokozott szekréciót (pl. hisztamin, proteázok, citokinek) okoz, mely a légutak azonnali szűkülését, nyálkahártya ödémát és bronchus görcsöt idéz elő. A hízósejtek részt vesznek a Th2 dominancia kialakításában és a krónikus légúti gyulladás fenntartásában is [36]. A légutakban található alveoláris makrofágok szintén képesek különböző mediátorok elválasztására, elsősorban az IL-10 és IL-12 termelésével gátolják a limfociták működését, de ez a hatás allergiás asztmásokban jelentősen mérséklődik, és ezzel a Th2 válasz erősítésében szintén szerepük lehet [37, 38]. Az epithel dendrikus sejtjei folyamatosan felveszik a belélegzett antigéneket, azokat a nyirokcsomókban lévő naív T-sejteknek prezentálva indukálnak szintén Th2 jellegű immunválaszt [39]. Az eozinofil sejtek beszűrődése a légutakba az allergiás reakció egyik fő jellemzője, részben pedig az általuk termelt mediátorok okozzák a jellegzetes epithel károsodást [40]. A gyulladásos folyamatban továbbá aktívan részt vesznek a légutakat alkotó szövetek sejtjei is (epitheliális és endotheliális sejtek, fibroblaszt és simaizom), ezek is képesek gyulladásos mediátorok termelésére.

2.2.1.6 Genetika háttér

Az utóbbi évtizedek rohamos molekuláris biológiai fejlődése mellett máig több mint 200 asztmával asszociált gént vizsgáltak, illetve állapították meg a patomechanizmusban betöltött szerepét, de az egyre növekvő ismeretanyag mellett sem tisztázottak a pontos kiváltó okok és a genetikai háttér szerepe [41-44]. Az asztmát sok más komplex multifaktoriális betegséghez hasonlóan több száz gén, bonyolult környezeti hatások és ezek kölcsönhatásai alakítják ki, így egy-egy faktor hatása önmagában nagyon gyenge.

A genetikai hajlamra az első fontos bizonyítékokat a klasszikus iker-vizsgálatok szolgáltatták. A konkordanciára, vagyis annak a valószínűségére, hogy az ikerpár mindkét tagja érintett (feltéve, hogy az egyik asztmás), a tanulmányok 36% és 77% közötti értékeket állapítottak meg; ez a genetikai háttér igen erős szerepére utal [45-48]. Érdekes azonban megjegyezni, hogy a korral a genetikai háttér hatása drasztikusan csökken és a környezeti hatások szerepe nő, 60 éves kor fölött a genetikai háttér szerepe már alig kimutatható [49]. Az allergia genetikai hátterével kapcsolatban pedig kimutatták, hogy az egyik szülő érintettsége esetén 10%, mindkét szülő érintettsége esetén 60% körülire emelkedik gyermekben az allergia megjelenésének valószínűsége [50].

A legrégebbi és legegyszerűbb vizsgálati módszer a jelölt gén asszociációs vizsgálat. Ilyenkor a betegségben feltételezetten érintett gének genetikai variánsait mérik meg az egészséges és a beteg populációkban. A módszer a nagy áteresztőképességű technikák elterjedése előtt relatív olcsó volt és könnyen értelmezhető eredményt ad, hátránya, hogy új gének és patomechanizmusok felfedezésére nem alkalmas. További hátránya, hogy erősen multifaktoriális betegségek esetén a sok gén kölcsönhatása miatt a módszer sok hamis pozitív találatot adhat és nehéz reprodukálni az eredményeket más populációkon. Ennek ellenére a módszer népszerű volt, 1000-nél is több asszociációs vizsgálatot végeztek már asztmában, melyekben több mint 120 gént hoztak összefüggésbe asztmával illetve atópiával [51], de ezek közül mindössze kb. 50 gént erősített meg kettő vagy több vizsgálat és kb. 10 olyan gén van, amelyet tíznél is több vizsgálat igazolni tudott (

). Ez utóbbiak, - nagyon nagy valószínűséggel asztmához vagy allergiához kapcsolódó gének - a következőek: IL4, IL13, ADRB2, HLA-DQB1, TNF, LTA, MS4A2, IL4R, CD14 és ADAM33 [52-66].

A hipotézis vezérelt jelölt gén asszociációs vizsgálatokat a hipotézmentes teljes genomszűrések (vagy más néven kapcsoltsági vizsgálatok) és a teljes genom asszociációs vizsgálatok (GWAS) egészíthetik ki. A teljes genomszűréseknél általában beteg testvérpárokat („affected sib pair”) vizsgálnak, ez egyben a módszer egyik fontosabb korlátozója is. Az érintett testvérpár teljes genomján egyenletesen elosztva ismert short

tandem repeat (STR) markereket mérnek meg, vagyis olyan rövid, általában neutrális hatású, változó számban ismétlődő DNS szakaszokat, melyek az emberek között általában nagy diverzitást mutatnak. A módszerrel meghatározható, hogy a betegekben milyen genomterületeken tér el a markerek eloszlása, azonban a beazonosított régiók általában a 10 millió bázis nagyságrendben vannak és így csak további, egyre szűkebb régiókat lefedő mérésekkel („pozicionális klónozás") pontosítható az eredmény. A technológia sok eredményt hozott [67], de a kétezres évek elején megjelent a módszer egy olcsóbb, egyszerűbb és nagyobb felbontást adó változata, a teljes genom asszociációs vizsgálat (GWAS). Ezekben a vizsgálatokban STR-ek helyett több millió SNP-t, vagyis egy nukleotidos polimorfizmust képesek mérni egyszerre. A teljes genomszűrésekkel ellentétben itt már potenciálisan valós funkcionális hatással bíró polimorfizmusokat is mérnek, így nem csak régiók beazonosítására alkalmas a módszer.

1. táblázat - Gu és munkatársainak 2011-ben készítettek reviewt az asztmával kapcsolt genom területekről [98]

Kromoszóma	**Lókusz**	**Gének**
1	1p36,1qter,1q23	*FCER1A,OPN3,CHML, IL10*
2	2q14,2q32,2q33,2p	*DPP10,IL18R1,CTLA4, CD28*
3	3q21-q22,3q21.3,3p	*TLR9*
5	5q31-q33,5q31,5p13,5p15,5q23.3	*IL4,IL9,ZFR3,LIFR, PTGER4,ADAMTS12,IL7R*
6	6p21,6q24-q25,6q25.3	*HLAG, ESR1,TNF*
7	7p14-p15,7q	*TCRG, GPR154*
8	8p23.3-23.2	*NAT2*
9	9p1,9p21,9p22	*TLE4,IFNA*
11	11q13,11q21,11q,11p14	*MS4A2,GSTP1*
12	12q13.12-q23.3,12q13-12q24,12q21,12q24.31,12q24.33	*SFRS8,CD45,IFNG, IRAK3,VDR*

13	13q14,13q	*PHF11,CYSLTR2*
14	14q11.2,14q13-q23,14q24,14q23	*TCR,ACT*
17	17q21	*ORMDL3*
19	19q13,19q13.3	*FCER2*
20	20q13,20p12	*ADAM33,JAG1,ANKRD5*
21	21p21	-
x	Xp,Xq	-

A patomechanizmus pontosabb megértéséhez a gének expressziójának ismerete adhat további információt. A kétezres évek elejétől elérhető microarray vizsgálatok technikailag viszonylag egyszerű és olcsó megoldást kínálnak erre, de a mérési eredmények értékelését tovább bonyolítja az expresszió időbeni változása és a szövetenként eltérő expresszió. Ráadásul asztma vizsgálata esetén a humán tüdőszövetek hozzáférhetősége korlátozott, így gyakran csak állatkísérletek eredményeire hagyatkozhatunk. Ovalbumin-indukált egér asztmamodellekben történt tüdő szövet vizsgálatokban számos gén eltérő expresszióját sikerült megállapítani, bár az állatmodellek humán érvényessége erősen korlátozott [68-71]. A kutatások azt valószínűsítik, hogy még számos nem vizsgált gén játszhat fontos szerepet a betegségben, melyek egyben terápiás célpontok is lehetnek.

2.2.1.7 A post-genom korszak kihívásai

A Human Genome Project több mint 10 év után 2001-ben publikálta a teljes humán genom draft szekvenciáját, majd 2003-ban a már 8-9-szeres teljes lefedettséggel szekvenált, véglegesnek tekintett szekvenciát [72]. Általában az ezt követő időszakot hívják a poszt-genom korszaknak, amikor is a strukturális genomikáról, vagyis a DNS szerkezetének vizsgálatáról a hangsúly áthelyeződött a funkcionális genomikára, amely a gének kifejeződését és szerepét vizsgálja. Az évezred elején hatalmas várakozásokkal tekintettek erre a korszakra és sokan azt remélték, hogy poszt-genom korszak már rövid időn belül drasztikus változásokat fog hozni nem csak az egészségügy, de a mindennapi élet sok más

területén is (pl. élelmiszeripar és mezőgazdaság). Ezzel szemben a poszt-genom korszak első évtizedére visszatekintve kettősséget látunk. A technológiai fejlődés üteme töretlen, Francis Collins (a Human Genom Project vezetője, ma az amerikai NIH vezetője) mérföldkőnek számító 2003-as Nature cikkében álomként megfogalmazott jóslatai sok szempontból beteljesültek: elértük az 1000 USD-ért szekvenálható teljes humán genom álomhatárát (4-5 nagyságendet csökkent a szekvenálás költsége 10 év alatt), 100 USD-ért mérhetünk milliós nagyságrendű genetikai markert, és rutinszerűen használnak nagy áteresztőképességű módszereket metilációs mintázat mérésére is [73]. Azonban a genomika eredményeinek transzlációja, vagyis a klinikumban történő széles körű alkalmazása már több csalódást okozott, erről újból Francis Collins ír 2010-es visszatekintő cikkében [6]. A hajlamok felmérését szolgáló genetikai tesztek megbízhatósága, a genetikai alapú rizikó csökkentését célzó beavatkozások és javallatok máig intenzív tudományos viták tárgyát képezik [74]; a genom alapú, személyre szabott terápiák egyelőre kevés esetben terjedtek el, még az élenjáró tumor terápiák is a betegek igen kis százalékán tudnak valóban segíteni [75]; és kevés új, kifejezetten genetikai kutatás alapján fejlesztett hatóanyag jutott a piacra [7]. Meg kell azonban jegyezni azt is, hogy a csalódást keltő első évtized után az utóbbi években egyre több a biztató jel, 2012-ben és 2013-ban sorban jelentek meg a kifejezetten genetikai kutatások alapján kifejlesztett biologikumok (nagymolekulás hatóanyagok, pl. Benlysta, Raxibacumab, Albiglitide stb.).

Sok multifaktoriális (kvantitatív) jelleg vagy betegség örökölhetősége populációs szinten statisztikailag jól kimutatható, azonban az öröklés genetikai vagy epigenetikai (pl. metilációs mintázat) háttere máig nem tisztázott. Például a testmagasság, mint klasszikus multifaktoriális jelleg öröklődő hányada 80% körül van, tehát kimutathatóan a populációs variancia 80%-áért örökletes tényezők felelősek, de a 2009-ig azonosított 40 körüli, testmagassághoz kapcsolódó genetikai pozíció ennek alig 5%-át volt képes magyarázni. A további kutatásokban figyelembe véve a jelleghez kapcsolódó genetikai pozíciók nagy számát és a variánsok gyenge egyéni hatását, a statisztikai módszereket tovább finomították. Így már 100-as nagyságrendű kapcsolódó genetikai pozíciót azonosítottak, de ez még mindig csak az öröklődő hányad 50-60%-át magyarázza [76, 77]. A multifaktoriális

jellegek és betegségek esetén többnyire hasonló a helyzet, csak a klasszikus mendeli öröklődést követő monogénes és néhány, kevés gént érintő jelleg genetikai háttere tekinthető tisztázottnak. Ráadásul a különböző genetikai asszociációs tanulmányok egymás eredményeit ritkán tudják megerősíteni. Ez a komoly tudományos vitákat is kiváltó jelenség „hiányzó örökletesség" néven ismert, amit egy találó metaforával emlegetnek úgy is, mint a „genetika sötét anyaga".

A sikertelenségre számos hipotézis és ezzel együtt új kutatási stratégia született. Ismert nehézséget jelentenek az epigenetikai, vagyis DNS bázissorendjét nem érintő, de öröklődő hatások (pl. metilációs mintázat) és a véletlenszerűen vagy környezeti hatásokra kialakuló jellegek, de a legélénkebb viták a genetikai háttér komplexitásával kapcsolatosak. A kérdésben két ellentétes tudományos nézet alakult ki: a gyakori betegség - gyakori variáns hipotézis (CD-CV) és a gyakori betegség - ritka variáns hipotézis (CD-RV). A CD-CV hipotézis részben egybevág a teljes genom asszociációs vizsgálatok (GWAS) stratégiájával, amelyek az 5%-nál magasabb populációs gyakoriságú polimorfizmusokat vizsgálják. Eszerint a komplex jellegek kialakításáért a populációban egyébként gyakran előforduló néhány (vagy sok) variáns együtt felelős, de az igazi oki variánsok kimutatása statisztikailag nehéz, mert a fizikai közelségből adódó ritkább genetikai rekombináció miatt együtt öröklődő variánsok statisztikailag erősen kapcsoltak, így a fenotípussal statisztikailag asszociáló variánsok száma igen nagy is lehet. A CD-RV hipotézis ezzel szemben azt állítja, hogy a komplex jellegeket inkább olyan ritka, egyedi mutációk okozzák, amelyek öröklődése ugyan populációs szinten kimutatható, de a nagy mintaszámú, gyakori variánsokat vizsgáló (GWAS) tanulmányok képtelenek ezeket mérni, így csak a teljes genom szekvenálás fog ezekre magyarázatot adni [78, 79]. A kérdés máig nem teljesen eldöntött, mindkettőre ismerünk példákat, azonban a tudományos közvélemény abban egységes, hogy az adatelemzési és adatintegrációs módszereknek kulcsszerepe van a genetika sötét anyagának megtalálásában.

2.2.2 A gyógyszeripar kihívásai és a számítógépes hatóanyag újrapozicionálás

A gyógyszeriparban a kilencvenes évek óta egy fenyegető trendnek lehetünk szemtanúi: egy-egy originális hatóanyag fejlesztési költségei drasztikusan nőnek, míg az évente piacra kerülő hatóanyagok száma drasztikusan csökken. A kilencvenes években általában gyakran 30-40-nél is több új hatóanyag került a piacra egyenként fél milliárd dollár körüli fejlesztési költséggel; a kétezres évek elején ez a szám az évi 15-öt közelítette, a fejlesztési költségek pedig ma gyakran elérik a 2 milliárd dollárt. A „blockbuster", vagyis évi több milliárd dollár forgalmú hatóanyagok kb. 15 éves szabadalmi védettségének lejárásával hirtelen megjelennek az olcsó másolatok (generikumok), a gyártó által elérhető profit tartalom lecsökken és így a korábban blockbuster hatóanyag bevételei nem fedezik tovább a következő évtizedek gyógyszereinek fejlesztési költségét. A fenntarthatatlanságot jelentő pontot el is nevezte az ipar „patent cliff disasternek", vagyis szabadalom-szirt katasztrófának.

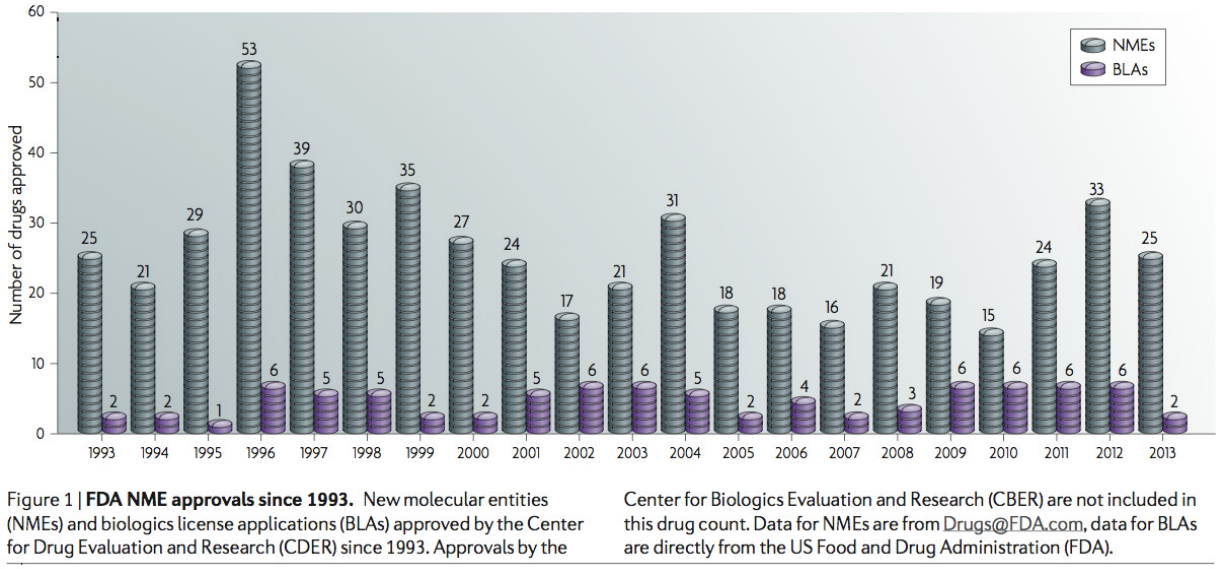


Figure 1 | **FDA NME approvals since 1993.** New molecular entities (NMEs) and biologics license applications (BLAs) approved by the Center for Drug Evaluation and Research (CDER) since 1993. Approvals by the Center for Biologics Evaluation and Research (CBER) are not included in this drug count. Data for NMEs are from Drugs@FDA.com, data for BLAs are directly from the US Food and Drug Administration (FDA).

5. ábra - Mullard és munkatársai 2014-ban a Natureben publikáltak tanulmány a gyógyszer engedélyeztetés legfrissebb statisztikáiról. Az engedélyezések csökkenő trendje 2010 körül feltehetően elérte a mélypontját.

Jelenleg sok biztató jel utal arra, hogy a folyamat 2010 körül elérte a mélypontját és a gyógyszeriparnak sikerül elkerülnie a katasztrófát (5. ábra), de számos megválaszolatlan kérdés van még mindig a jelenlegi gyógyszerfejlesztési modellek fenntarthatóságával

kapcsolatban. Az elektronikusan felhalmozott tudás jobb kiaknázása a gyógyszerfejlesztési folyamatban egy vitán felül álló stratégiai fontosságú irány, de a gyógyszeripar továbbra is intenzíven keresi az innovatív, hatékonyságnövelő stratégiákat.

Az egyik ilyen innovatív ipari törekvés azt célozza meg, hogy különböző módszerekkel már piacon lévő vagy akár fejlesztés alatt álló, netán elbukott hatóanyagok indikációit vagy alkalmazásait kiterjessze, újrapozicionálja („repositioning", „repurposing", „reprofiling", „retasking", „new indication discovery"). Korábban ezeket a technikákat leginkább piacon lévő gyógyszereknél alkalmazták a szabadalmi védettségi időszak kiterjesztésére, mint egyfajta életciklus hosszabbítási stratégia. Erre számtalan sikertörténetet is ismerünk, de ezek a gyakorlatok egyre inkább meghonosodnak a gyógyszerfejlesztés korai stádiumaiban is. A Pfizer gyógyszergyár híres, korai fázisban újrapozionált blockbuster hatóanyaga egy vonzó mintát mutatott erre: a Viagra a klinikai fázis 1 vizsgálatokon hatékonyság hiánya miatt megbukott, de az addigi adatgyűjtés rávilágított egy nem várt indikációra [80], mellyel aztán bekerült minden idők üzletileg legsikeresebb hatóanyagai közé. Számos más esetben is fontos az alkalmazási lehetőségek mielőbbi pontos feltérképezése, így pl. új kombinációk, formulációk, személyre szabott célzott terápiák, több lépésben történő engedélyeztetés a beteg csoportok szegmentálásával, vagy kiterjesztések állatgyógyászati területekre [81]. Az utóbbi időben egyre komolyabb figyelmet kapó ritka betegségek („orphan disease"), vagy a szinte végtelen számú mutációs lehetőséggel küzdő tumor terápiák szintén fontos új területei az alkalmazás kiterjesztéseknek.

Egy átlagos gyógyszer kifejlesztése 10-17 évig tart és a költségek elérhetik az 1-2 milliárd dollárt, míg a fejlesztés alatt a jelöltek több, mint 90%-a elbukik. Egy újrapozicionált hatóanyag piacra juttatása 3-12 évre csökkenthető, ráadásul jelentősen olcsóbban és sokkal nagyobb eséllyel juthat el a klinikai fázis 1-től az engedélyeztetésig [80].

A preklinikai fázisban és a klinikai fázis 1 során a bukás leggyakoribb okai különböző gyógyszerbiztonsági problémák és ebben az esetben kevés alternatív lehetőség van. Ezzel szemben a későbbi klinikai fázisokban a lemorzsolódás fő oka általában a hatásosság hiánya, a molekulák több mint 50%-a bukik el emiatt a klinikai fázis 3-ban [81]. A

hatásosság hiánya jelentheti azt, hogy a molekulának semmilyen előnye nincs a jelenlegi terápiákhoz képest, vagy nem használható kiegészítő terápiaként, vagy egyáltalán nem mutatható ki semmilyen előny a placebohoz képest. Ez a nagy mennyiségű, hatástalanság miatt elbukott vegyület óriási lehetőséget jelent az új indikáció kereső módszereknek.

Évente hozzávetőlegesen 150-200 vegyület bukik el a klinikai vizsgálatok különböző fázisaiban, és a gyógyszeripari cégek polcain heverő bukott hatóanyagok számát 2000 és 30 000 közé teszik [81-83]. A klinikai vizsgálatok egyre növekvő költségei miatt az ipar gyakran a kockázatosnak tűnő hatóanyagok korai megbuktatása mellett dönt („fail fast”), és ez a jelentős korai befektetések ellenére is nagyon sok hatóanyag feladásához vezet. Így sok olyan vegyület van, amelyről már hatalmas mennyiségű tudás és kísérleti adat gyűlt össze, ráadásul biztonságosnak is tekinthetőek. A klinikai fázis 3 statisztikái azt mutatják, hogy ráadásul ezek a hatóanyagok többnyire nem csak biztonságosak, de rendelkeznek a szükséges farmakokinetikai és farmakodinámiás tulajdonságokkal is. Ezeken túl sok hatóanyagot pusztán stratégiai és pénzügyi megfontolásokból nem fejlesztenek tovább [81, 84]. Az elfekvő hatóanyagokban lévő potenciált az utóbbi években több állami és ipari közös kezdeményezés is felismerte, és támogatja a kiaknázását (pl. National Center for Advancing Translational Sciences [83]).

A látszólag hatalmas potenciál ellenére az utóbbi évek hatóanyag újrapozícionálási projektjei és a terület startup cégei lehangoló statisztikát mutatnak [85]. Az évekkel ezelőtt megbukott hatóanyagok többségének újrapozícionálását hátráltatja, hogy az eredeti fejlesztő csoport már általában nem elérhető, az eredmények összegyűjtése és áttekintése nem egyszerű, és nehéz megszerezni a vállalat támogatását egy korábban megbukott hatóanyag újbóli vizsgálatához. Ráadásul a biztonsági előírások gyorsan változnak és gyakran a korábbi kísérletek eredményei már nem is elegendőek. Komoly biztonsági kérdéseket vethet fel, ha az új felhasználási terület jelentősen eltér, például akut helyett krónikus kezelés, új dozírozás, más beteg populáció, vagy ha az alkalmazás módja eltér. A legtöbb újrapozicionált hatóanyag esetén szükség van újbóli preklinikai vizsgálatokra, hogy igazolják az új hipotézist mielőtt a klinikai vizsgálatokat megkezdhetik, majd a klinikai tesztek során a dozírozás újbóli beállítására, farmakokinetikai és farmakodinámiás

tulajdonságok validálására, biomarkerek vizsgálatára, olyan speciális problémákról nem is beszélve, mint a szabadalmi védettség kiterjesztése – tehát a nehézségek nem lebecsülendőek [81].

Manapság az adatrobbanás korszakát éljük, óriási mennyiségű elektronikus tudás érhető el: több száz manuálisan karbantartott vegyület adatbázis, több ezer klinikai tanulmány, több 10 millió szakcikk pl. a PubMed-ben, milliós számban elektronikus orvosi kartonok, és megszámlálhatatlan szociális média adatforrás terápiák szubjektív értékeléséről. Az információs korszak a gyógyszerkutatást egyre inkább egy Big Data tudománnyá változtatja és a számítógépes indikáció keresés feladata ezen hatalmas mennyiségű adat értelmezése [86, 87]. Számos kiváló összefoglaló született a területről [88-91], így alább csak a legfontosabb iskolákat sorolom fel:

- Szövegbányászat (szakirodalom, szociális média, stb.) [92, 93]
- Strukturális bioinformatika [94, 95]
- Molekuláris biológiai adatelemzés (DNS, expresszió, stb.) [96, 97]
- Hálózatelemzés [98-100]
- Adatintegráció és adatbányászat [101]

DOI:10.14753/SE.2015.1757

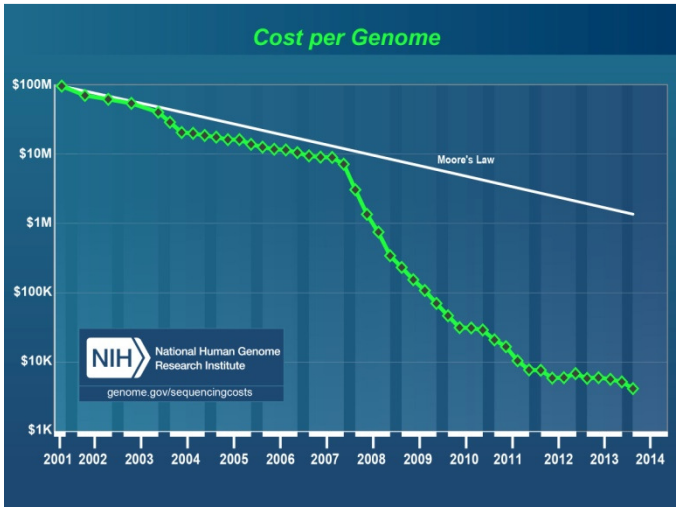
2.3 A tárgyterületek adatai és információi

2.3.1 A molekuláris orvostudományok adatai

A biológiai rendszerek molekuláris szintű megértése az adatmérnöki tudományokkal érdekes párhuzamban és időben szinte teljes átfedéssel alakult ki, így a harmadik információrobbanás története a molekuláris szintű módszerek fejlődésévével összefonódik. A DNS-t először 1869-ban izolálták, majd több mint hatvan évnek kellett eltelnie, mire Watson és Crick 1953-ban felfedezte a kettős helix szerkezetet, és először mondta ki a DNS adattárolási képességével kapcsolatos sejtését [102], innentől kezdve a fejlődés jelentősen felgyorsult.

A biológiai rendszerek különböző molekuláris szabályzási szinteken (genetikai, expressziós, proteomikai, metabolomikai, stb.) történő mérése és megismerése a számítógépes biológiával összefonódva két, egymástól éles határral nem elválasztható, de eltérő történeti korszakra osztható. Az 1980-as évek végéig a technológiai lehetőségekből adódóan szinte kizárólag hipotézis-vezérelt felfedező módszereket alkalmaztak: a korszakot elsősorban a kis adatmennyiséget szolgáltató, drága molekuláris biológiai mérések, és hipotézis-tesztelésen alapuló kiértékelés jellemezte, ahol a számítástechnikai eszközökkel elsősorban a hipotézis felállítását támogatták, a hangsúly a szimulációs technikákon volt. A 1990-as évek elejétől a műszertechnikai (és információ technológiai) fejlődés fokozatosan egyre olcsóbbá tette méréseket, a nagy áteresztőképességű technikákkal (mint pl. a microarray) egy teljes omikai szint (pl. az összes gén expresszió) egyszerre mérhetővé vált. Ezzel a hangsúly áthelyeződött a hipotézismentes tervezésre és vizsgálatokra, a számítástechnika fókusza pedig eltolódott az adatok kiértékelése és értelmezése irányába [15, 17]. Ez a szemléletváltás a biológiai rendszerekről olyan mennyiségű adatot szolgáltat, amely néhol a Moore-törvény trendjét is jelentősen felülmúlja és így sok területen a számítástechnika szűk keresztmetszetté válik. A jelenséget el is nevezték Carlson görbének („Carlson curve”) Rob Carlson után, amely szerint az elérhető szekvenálási teljesítmény időben legalább a Moore törvénye szerint nő [103, 104] (6. ábra).

DOI:10.14753/SE.2015.1757



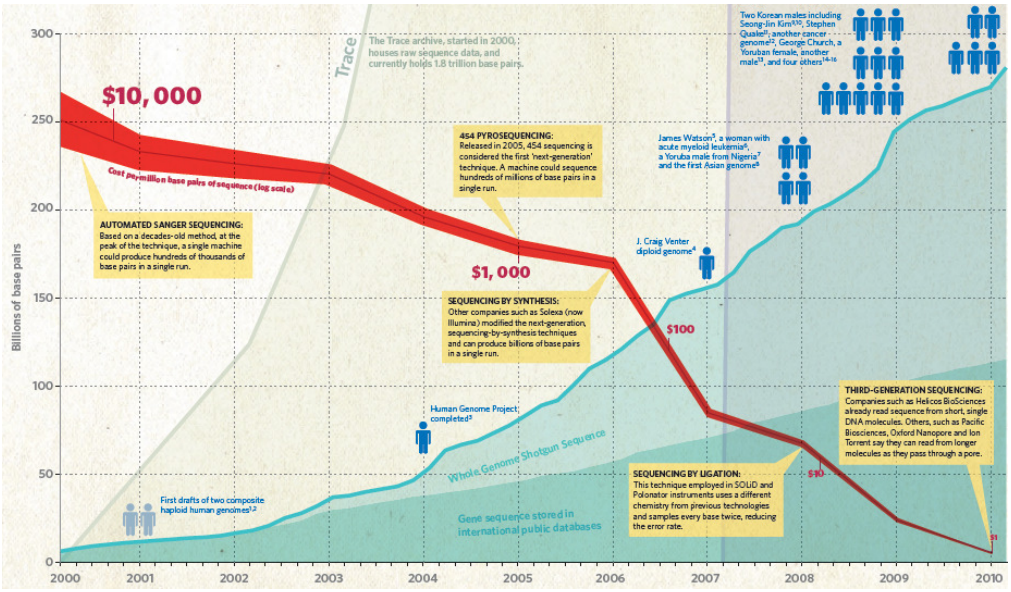
6. ábra - A genom szekvenálás költsége lényegesen gyorsabban csökken, mint a Moore törvény által meghatározott trend (NIH, National Human Genome Research Institute, http://www.genome.gov/sequencingcosts/, 2014. április).

Azonban nem csak az adatok mennyisége jelenti a kihívást. Két szempontból is eltérőek ezek az adatok attól, amelyekkel a huszadik század végéig szembesültek a statisztikai módszerek. Egyrészt az adatok egyre jelentősebb része nyers és strukturálatlan (pl. nyers DNS szekvencia vagy tudományos szövegek), másrészt ezen adatok dimenziója szokatlan és a hagyományos statisztikai módszerekkel szinte kezelhetetlen. Míg korábban jellemzőbb volt a relatíve kisszámú változó mérése nagy mintaszám mellett (optimális esetben legalább 1:10 vagy 1:100 arányban), addig ma ez az arány nem csak megfordult, de gyakrabban több nagyságrend eltérést is mutat, pl. tipikusan egy ezres számosságú beteg populáció teljes genom asszociációs vizsgálata során milliós nagyságrendű változót mérünk és ez még csak a legalsó biológiai szabályzási szint.

Az élettudományi, molekuláris biológiai adatbázisok számossága, komplexitása és mérete az évezred első évtizedére akkorára duzzadt, hogy csak annak a kimerítő összefoglalása is jelentősen meghaladná jelen értékezés kereteit, az adatbázisok hatékony integrálása már egy évtizede önálló kutatási területté vált [105]. Alább bemutatok néhány kiemelt példát, illetve csoportosítom őket biológiai és információelméleti szempontból, hogy szemléltessem a probléma komplexitását és a trendek által előre vetített újabb korszakváltást.

DOI:10.14753/SE.2015.1757

Az adatbázisokat információelméleti szempontból két nagy csoportra lehet osztani: adattárakra és meta-adattárakra. Az első csoport a különböző rendszerbiológiai szintek konkrét mérési eredményeit tartalmazza: így például egyes organizmusok nyers DNS szekvencia mérési adatait (esetleg elméleti referencia szekvenciáit), génexpressziós mérések eredményeit, vagy akár fehérjék mérési eredményeit, pl. röntgen diffrakciós méréseket, bezárólag a teljes ökoszisztémák metagenomikai vagy metabolomikai karakterizációjával.



7. ábra – A Nature 2010-ben közölt cikket az elérhető DNS adatok mennyiségéről „Human genome at ten: The sequence explosion" címmel [103]. Az elérhető szekvenálási adatok mennyisége exponenciálisan nő, míg a szekvenálási költségek hasonló ütemben csökkennek.

A 2005 óta elérhető újgenerációs szekvenálási technikák által szolgáltatott adatok mára számos hatalmas adatbázisban érhetőek el (7. ábra). Ezek közül az egyik legismertebb az 1000 Genome Project (http://www.1000genomes.org/) 2012-ben jelentette be, hogy túllépte 1000 teljes emberi genom szekvenálását, mára ez a 2000-et megközelítette és 200 terabytenál is több DNS adatot kezel. Egy byte négy bázis tárolását teszi lehetővé, így a kb. 6 Gb (gigabázis, milliárd bázis) méretű teljes, tömörítetlen diploid emberi genom kb. 1,5 GB (gigabyte) adatmennyiséget jelent, de a nyers mérési eredmény ennek 100-szorosa is lehet. A projektben a legnagyobb intézetek működnek együtt a világ minden részéről,

beleértve az amerikai National Institutes of Health (NIH), a kínai Bejing Genomics Institute (BGI) és az angol Wellcome Trust Sanger Institute [106].

A nagy áteresztőképességű expressziós mérések már a kilencvenes évek végén megjelentek, így az első publikus expressziós adatbázisokat már a kétezres évek elején megalapították. A két legismertebb az amerikai NIH által üzemeltetett Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), és a hasonló méretű European Bioinformatics Institute ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), mára mindkettő túllépte az 1 millió mérést, 100 terabyteot közelítő nagyságrendben tárolnak mérési eredményeket az emberen kívül sok más organizmusokról is, és egy méréssel ma már 2 milliónál is több változót tudunk mérni. Az adatbázisok tartalmaznak speciális RNS méréseket is, így pl. RNA-seq, Chip-Seq adatokat, vagy miRNA méréseket [107].

A fehérjemérés és katalogizálás egyik legfontosabb gyűjteménye a Protein Data Bank, (http://www.rcsb.org/) melyet több mint 40 éve alapítottak, és ma amerikai, európai és japán kooperációban üzemeltetnek. Az adatbázis ma kb. 100 000 makromolekula (fehérje és nukleinsav) elsősorban strukturális méréseit gyűjti össze, melynek jelentős része strukturálatlan jellegű és annak valamilyen kivonata (pl. röntgen diffrakció) [108].

Információ elméleti szempontból az adatbázisok második nagy csoportja metaadat jellegű („adatok az adatokról”), inkább taxonómiákat, ontológiákat, annotációkat tartalmaznak: így például genom részletekkel vagy génekkel kapcsolatos tudást gyűjtenek össze, genetikai variánsokat katalogizálnak, géneket csoportosítanak biológiai funkciók szerint, különböző RNS molekulákat rendszereznek és katalogizálnak.

A UCSC Genome Browsert (https://genome.ucsc.edu/), az egyik legfontosabb genom annotációs adatbázist 2000-ben alapították a University of California-n és eredetileg a Humán Genom Project eredményeinek (mindenki számára elérhető) annotálására jött létre. Mára 46 faj referencia genomja böngészhető az annotációkkal együtt grafikus felületen keresztül [109]. Európában az Ensembl adatbázis tölt be hasonló szerepet (http://www.ensembl.org/).

A molekuláris biológiai adatok egy érdekes aspektusa az egyéni genetikai háttérből adódó eltérések, nyilvántartásukra külön adatbázisok jöttek létre. A genetikai polimorfizmusok nyilvántartására készített, NIH által üzemeltetett dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) talán a legfontosabb sztenderd. 1998 óta működtetik és pontmutációkon túl STR szekvenciákat (változó számú rövid ismétléseket), rövid törléseket, beékelődéseket és egyéb variánsokat is katalogizálnak több organizmustól, ma összesen közel 100 milliót [110]. Az International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) már a humán genom genetikai variációinak katalogizálásán túl azok együtt előfordulásait gyűjtötte össze több száz ember több millió variánsának mérésével. A már lezárult projekt nemzetközi összefogásban valósult meg, fő célja a gyakori genetikai variánsok kombinációinak, mintázatainak (haplotípusainak), és azok rasszonkénti gyakoriságainak összegyűjtésével népegészségügyi, orvosi kutatások támogatása [111].

Számos olyan adatbázist hoztak létre, ahol már az annotációk fogalmi szintjét szervezik ontológiákba. A biológiai eredmények egységes megosztásában és értelmezésében kulcskérdés a sztenderdizált fogalomrendszerek, szótárak és taxonómiák létrehozása. A terület legfontosabb képviselője Gene Ontology (http://www.geneontology.org/), amely gének és géntermékek fogalmait kategorizálja három témakör szerint (sejtalkotók, molekuláris funkciók és biológiai folyamatok), ma több 10 000 fogalmat szervez rendszerbe [112].

Az évezred első évtizedében jellemzően a mérések még mindig időben többnyire pontszerűek voltak és kevés az idősoros adat, így nehéz a különböző beavatkozások vagy környezeti hatások eredményének mérése. További problémát jelent, hogy a komplex rendszerekben mért sok zajos változó miatt a statisztikailag szignifikáns eredmények eléréséhez hatalmas mennyiségű mérésre van szükség. Évtizedünk egyik fontos trendje adhat választ ezekre a kérdésekre, amelyet ma összefoglaló néven „Quantified Self”-nek hívnak. A mozgalom elsősorban azt célozza meg, hogy hordozható egészségügyi, fizikai, környezeti, pszichológiai stb. monitorozó eszközökkel (pl. okostelefonokkal) képesek legyünk elsősorban saját testünk folyamatos megfigyelésére, mérésére, és korábban nem

DOI:10.14753/SE.2015.1757

ismert összefüggések feltárására, illetve változtatásokral az életminőség javítására. A technológia fejlődése azt vetíti előre, hogy egy évtizeden belül a saját testünkhöz való viszonyunk drasztikusan át fog alakulni, a jelenlegi reaktív („curative”) orvoslás helyett a megelőző („preventív”) és személyre szabott orvoslás kerül előtérbe. A mai gépkocsikhoz hasonlóan az eszközeink képesek lehetnek jelezni, amikor szervízelésre („orvosra”) van szükség, amikor túlhajtjuk a motort („szívet”), amikor nem megfelelő az üzemanyag („táplálék”). A Quantified Self fejlődése az okostelefonokhoz hasonlóan az emberiség jelentős részét érintheti, így ez egyben az egész Big Data terület egyik legfontosabb kihívása [86].

2.3.2 A farmakológiai adatvagyon

A kémiai és vegyészeti számítógépes technikák fejlődésében a számítógépes biológiához hasonló folyamatokat figyelhetünk meg. A korai szimuláció intenzív számítógépes kémia („computational chemistry") mellett a kilencvenes évek végétől egyre nagyobb szerep jutott az adatoknak. A szemléletváltás egyik mérföldkövének tekinthető, amikor F.K. Brown először használta és definiálta a „kemoinformatika" (chemoinformatics) kifejezést 1998-ban. Brown eredeti definíciója szerint a kemoinformatika az információ források integrálása, melynek során az adatból információ, információból pedig tudás lesz annak érdekében, hogy a molekulák azonosítása és optimalizálása hatékonyabbá váljon [113].

A szemléletváltás legfontosabb oka ebben az esetben is a műszertechnikai fejlődés volt, és elsősorban a kilencvenes években népszerűvé váló nagy áteresztő képességű vizsgáló és szűrő módszerek (high-throughput screening, HTS), amelyekkel automatizált laboratóriumi vizsgálatok végezhetőek el molekulák ezrein. A kétezres évek elejére valóban hatalmas és egyre növekvő mennyiségű adat érhető el kémiai entitásokról a legkülönfélébb publikus és vállalati adatbázisokban, együttes kiaknázásuk komoly potenciált rejt. Segítségükkel egyre korábbi stádiumban lehet egyre pontosabban előre jelezni a molekulák bizonyos tulajdonságait, mint pl. potenciális indikációk, mellékhatások, toxicitás, stb. [114].

Informatikai szempontból a gyógyszeripari fejlesztés egy inkrementális adatgyűjtési folyamat, kezdve egy elméleti szinten létező molekulától, bezárólag egy teljesen karakterizált, empirikusan validált, piaci forgalomban lévő hatóanyaggal, melyről nagy mennyiségű valós használati tapasztalat („post-marketing surveillance") is elérhető. Az amerikai FDA (Food and Drug Administration) vagy az európai EMA (European Medicines Agency) által jóváhagyott hatóanyagok tekinthetőek az ismert molekulák azon halmazának, melyek ezen adatgyűjtési folyamat minden lépését sikeresen be tudták fejezni (az összes ismert hatóanyag osztályozását lásd Huang és munkatársai munkájában [115]).

Egy új molekula vizsgálata során a vegyületekről elérhető adattömeg kiaknázásának legáltalánosabb módja ismert profilú, hasonló vegyületek vizsgálata. A vegyületek fizikai és kémiai paraméterei, mint például adott kémiai profil vagy szerkezeti taxonómiai

besorolás már a legkorábbi hit (vegyület találat) és lead (vezérmolekula) stádiumban is ismert minden molekuláról. Ezen adatok segítségével már korai stádiumban is összehasonlíthatóak az adatbázisokban lévő molekulák, és kapcsolatok kereshetőek más ismert klinikai profilú hatóanyagokkal. Ezen alapvető fizikai-kémiai paraméterek szinte minden ismert szerkezetű molekuláról elérhetőek, így az együtt elemezhető entitások mennyisége hatalmas [116]. Számos adatbázis tartalmaz ilyen, nagy áteresztőképességű biokémiai vagy fenotípus szűrési eredményeket is több 10 millió hatóanyagra, pl. a PubChem (http://pubchem.ncbi.nlm.nih.gov/, [117]) vagy a ChemBank (http://chembank.broadinstitute.org/, [118]). Fontos azonban megjegyezni, hogy a sok ismert entitás mellett is, ezen korai fázisban megismerhető paraméterek még nagyon keveset árulnak el a valós klinikai profilról. Bár egy új molekulán minden további, komplexebb in vitro mérés (pl. az aktivitás és cél profil validálása) a várható üzleti és klinikai hasznosság pontosabb becslését segíti, de az újabb mérések megtérülése soha nem garantált, így egyre kevesebb hatóanyagra végzik el azokat. A ChEMBL (https://www.ebi.ac.uk/chembl/), az egyik legbővebb, manuálisan karbantartott célpont adatbázis több mint 600 000 bioaktív molekula adatait tartalmazza [119], míg a DrugBank (http://www.drugbank.ca/) kb. 10 000 FDA által jóváhagyott és kísérleti stádiumban lévő hatóanyagot tart nyilván, de a kémiai tulajdonságok mellett már a pontosabb célpont adatokkal együtt [120, 121]. Általánosságban elmondható, hogy minden újabb kísérleti eredmény egyre többet fedhet fel a valós klinikai profilról, de a mérések egyre specifikusabbak az adott indikációra vagy hatásmechanizmusra, így egyre kevesebb más molekulával hasonlíthatóak össze (ami fokozottabban igaz a klinikai vizsgálatok stádiumaiban gyűjtött adatokra).

A molekuláris biológiai és farmakológia adatintenzív módszereinek érdekes metszete a preklinikai fázisban gyűjtött molekuláris biológiai adatok. Génexpressziós és proteomikai adatokat évtizedek óta felhasználnak a gyógyszerfejlesztésben, de az egyre nagyobb áteresztőképességű módszerek új lehetőségeket nyitnak, az egyik legismertebb megközelítés a Connectivity Map vagy CMAP (https://www.broadinstitute.org/cmap/, [97, 122]). A CMAP a Broad Institute által fejlesztett adatbázis és eszköztár, melyben 7000-nél

is több microarray gén expressziós eredményt tartanak nyilván olyan laboratóriumi körülmények között szaporított sejtvonalakról, melyeket kb. 1300 kismolekulás hatóanyag egyikével kezeltek különböző koncentrációkban. Az adatbázis ezzel felfedheti hatóanyagok, gének és betegségek összefüggéseit: pl. ha egy adott betegség és hatóanyag ellentétesen modulálja a génexpressziós profilt, akkor feltételezhető a kapcsolatuk, vagyis felvetheti a hatóanyag hatásosságát adott indikációban.

A preklinikai fázisban az egyre specifikusabb mérések felé haladva az utolsó fontos mérföldkő az in vivo állatmodellek adatbázisai, mint a MUGEN (http://www.mugen-noe.org/, [123]) vagy az IKMC (http://www.knockoutmouse.org, [124]). Az utóbbi évtizedekben ezen in vivo kísérletek egy részét is automatizálták, és így az egyre nagyobb áteresztőképességű módszerek egyre nagyobb fenotípus szűrési adatbázisokat eredményeznek. Az adatok és módszerek sztenderdizálásával ezek a módszerek egyre fontosabb részét képezhetik a gyógyszerfejlesztésnek.

A nagy mennyiségű adat sztenderdizálása és integrálása a hatékony felhasználás egyik kulcskérdése minden mérési technika esetén, hiszen az igazi értéket az adatok együttes elemzése adhatja, de ez szisztematikus mérési hibákkal és eltérő fogalmi sztenderdekkel nem lehetséges. A Pharmaceutical Collection of the NIH Chemical Genomics Center (http://tripod.nih.gov/npc/), az amerikai National Institutes of Health egy nagyszabású vállalkozása, amely talán a legátfogóbb, nem redundáns, sztenderdizált fizikai és elektronikus gyűjteménye az emberiség által ismert összes olyan kismolekulás hatóanyagnak, melyet emberi vagy állati felhasználásra valaha engedélyeztek (8969 vegyület). A gyűjtemény nagy részét különböző koncentrációkban lemérték több mint 200 mérési eljárással, célpontok, szabályozási útvonalak és sejtfenotípusok azonosítása érdekében. Ez ma az egyik legértékesebb adatbázis az adatvezérelt hatóanyag fejlesztéshez, az ismert tulajdonságú hatóanyagok segítségével új hatóanyagok különféle tulajdonságai becsülhetőek meg előre egyre hatékonyabban [115].

A klinikai vizsgálatokban keletkező legtöbb adat nagyon specifikus az adott indikációra és terápiára (pl. különböző biomarker eredmények), vagy a prediktív értékük viszonylag

alacsony (pl. dozírozás), esetleg a sztenderdizálás nehézkes (pl. bioelérhetőség, farmakokinetika és farmakodinámia), és így viszonylag korlátozottan alkalmazhatóak prediktív jellegű számítógépes elemzésre. Van azonban néhány kivétel, amelyeket viszonylag könnyű sztenderdizálni, felhasználni, és egyre nagyobb mennyiségű adat áll rendelkezésre: ezek a mellékhatások és indikáción túli felhasználási lehetőségek.

Minden klinikai kísérletben szigorúan monitorozzák a mellékhatásokat sztenderdizált módszerekkel. Campillos és munkatársai azt elemezték, hogy egy hatóanyag mellékhatás profilja felfedheti-e a hatásmechanizmus elemeit és így rávilágíthat-e potenciális új indikációkra? Kutatásaik bebizonyították, hogy két hatóanyag hasonló mellékhatás profilja esetén nagyobb a valószínűsége a közös célpontnak is, így a mellékhatás adatok akár új indikációk felhasználási lehetőségeit is felvethetik [101].

A szövegbányászati technikák mindig is ígéretes módszerek voltak [125] orvosbiológiai tudás gyűjtésére, mert segítségükkel azonosíthatóak olyan áttételes kapcsolatok (gének, fehérjék, betegségek, hatóanyagok, stb. között), melyeket korábban még nem tanulmányoztak együtt [126]. Ezeket a módszereket a mai napig intenzíven fejlesztik, de az utóbbi évtized szótár és ontológia egységesítési projektjei (pl. UMLS /http://www.nlm.nih.gov/research/umls/, MeSH /http://www.nlm.nih.gov/mesh/, MedDRA /http://www.meddra.org/, OMIM / http://www.ncbi.nlm.nih.gov/omim/, Gene Ontology / http://www.geneontology.org/, Human Phenotype Ontology /http://www.human-phenotype-ontology.org/, stb.) jelentősen hozzájárulnak mind a természetes nyelvi elemzésen (NLP), mind a kifejezések együtt előfordulási statisztikáin alapuló módszerek eredményeihez. Ez azért is különösen fontos, mert az elektronikusan elérhető orvosbiológiai tudás mennyisége már az olyan folyamatosan növekvő szakterületi adatbázisokban is követhetetlen emberi szakértő számára, mint a PubMed (http://www.ncbi.nlm.nih.gov/pubmed/, [92]. A szakterületi adatbázisokon túl pedig van egy sokkal bővebb és gyorsabban növekvő szöveg halmaz, ez pedig a szociális média, ahol a felhasználók folyamatosan, valós időben közlik a terápiákkal kapcsolatos tapasztalataikat [127].

DOI:10.14753/SE.2015.1757

A hatóanyagokról elérhető információtömeg a tudásfúziós technikák réven már eddig is jelentősen hozzájárult a gyógyszerfejlesztéshez, azonban várhatóan az adatok jelentősen bővülni fognak néhány új forrásból is. Az információk új hullámának forrásai a mellékhatásokat érintő új, szigorúbb adatgyűjtést megkövetelő szabályozások; a klinikai tanulmányok publicitásának és átláthatóságának növelésére vonatkozó új kezdeményezések; és az új trendek, amelyek a betegek hipotézismentes adatgyűjtését segítik elő különböző hordozható, egészség monitorozó eszközökkel („Quantified Self”).

DOI:10.14753/SE.2015.1757

2.4 A tárgyterületek adatmérnöki módszerei

2.4.1 Molekuláris biológiai adatelemzési módszerek

Az élettudományokban a hipotézis vezérelt kísérlettervezés és hipotézistesztelésre épülő, úgy nevezett frekventista statisztikai következtetés szinte egyeduralkodóvá vált. A statisztikai eszköztár használata a mély elméleti megalapozottságnak, az évezredes beágyazottságnak (pl. Arisztotelész, Rétorika [128]) és relatíve könnyű számíthatóságának is köszönhető, de jól illeszkedik a megfigyelések, mért adatok jellegéhez is: az ezredfordulóig a területen dominánsak voltak a kis adatmennyiséget szolgáltató, drága és célzott mérések (PCR, Sanger szekvenálás, stb.).

A frekventista következtetés elsősorban a valószínűség frekventista értelmezésére épít, mely a legközelebb áll a valószínűség hétköznapi értelmezéséhez is. Eszerint egy véletlen esemény valószínűsége az az érték, amelyet egy sokszor elvégzett megfigyelés során a vizsgált esemény bekövetkezésének relatív frekvenciájára egyre inkább megközelít, általában egy 0 és 1 közötti értékkel reprezentálják. Mivel a világról minden információ nem lehet a birtokunkban, egy relatíve kisszámú megfigyelés alapján kell a következtetéseket levonni, ami a frekventista következtetési keretrendszerben általában hipotézisek elfogadását vagy megcáfolását jelenti. Tehát ehhez először fel kell tenni a megfelelő kérdést, vagyis az alternatív hipotézist, mely a tudományos vizsgálatokban tipikusan két esemény vagy mért érték között fennálló valamilyen kapcsolat (pl. a genetikai variáns és a betegség összefügg); míg ezzel szemben a null hipotézis az alapfeltevés, vagyis a kapcsolat hiánya, olyan összefüggés , amit közvetlenül tesztelni tudunk (pl. a beteg és egészséges csoport genetikai variánsainak eloszlása megegyezik). A szignifikancia küszöbértéket, vagyis a null hipotézis helytelen elutasításának elfogadható valószínűségét előre rögzíteni kell, így a következtetés egy eldöntendő kérdésre adott válasz formájában jelenik meg. A statisztikai hipotézistesztelést gyakran hívják megerősítésen alapuló adatelemzésnek is („conformatory data analysis”), a hipotézist célzott mérésekkel lehet megerősíteni vagy elvetni.

Az utóbbi évtizedben elterjedt nagy áteresztőképességű módszerek már egyszerre képesek nagyon sok eseményt mérni, pl. egy microarray méréssel szinte minden gén expresszióját. A kiértékelés ilyenkor is elvégezhető frekventista következtetéssel, de ebben az esetben minden mért változó (és esetleg azok kombinációinak) kapcsolatára egy-egy statisztikai tesztet kellene végezni. Ebben az estben merül fel a többszörös hipotézis tesztelés problémája, amelyet intuitívan a következőképpen értelmezhetünk. Minél több valószínűségi változót vizsgálunk két csoport között (genetikai variánsok, környezeti hatások, életvitel, és ezek tetszőleges kombinációi), annál nagyobb az esélye annak, hogy eltérő eloszlású változókat találunk a két csoport között, akkor is, ha a változók értékei teljesen véletlenszerűek - és így hibásan következtethetünk a hatásukra. Ennek megfelelően a tesztek számával arányosan korrigálni kell a szignifikancia küszöböt, ez pedig a gyakorlatban nem teszi lehetővé gyengébb hatások kimutatását, szinte minden összefüggést el fog vetni a statisztikai teszt. A módszer nehézségei genetikai asszociációs vizsgálatok területén jól ismertek. A mért változók száma gyakran több nagyságrenddel nagyobb lehet a mintaszámnál; a variánsok hatása multifaktoriális betegségek esetén tipikusan gyenge, gyakran populációfüggő, a célváltozók (pl. a betegség) leírása nem elég pontos, így az összefüggések túl gyengék; a következő tanulmányok nem tudják megerősíteni őket; a megtalált összefüggésekről gyakran nem lehet eldönteni, hogy közvetlen hatásnak köszönhetőek-e vagy áttételesek-e [129]. Ilyenkor érdemes a megerősítésen alapuló adatelemzés helyett felfedező jellegű adatelemzést végezni („exploratory data analysis”), pl. frekventista keretben szóba jöhetnek vizualizációs módszerek vagy főkomponens analízis. A nagy áteresztőképességű mérések azonban egyre inkább dominánssá váltak az elmúlt évtizedben, így a frekventista módszerekkel szemben érdemes egy alternatív normatív keretrendszert, a Bayesi következtetést bevezetni, amely az említett problémákat hatékonyabban képes kezelni.

A Bayesi következtetés gyökere szintén a valószínűség értelmezése, mely ebben az esetben egyfajta szubjektív értékként, a bizonyítékok alapján értelmezett hiedelemként jelenik meg. Egy hipotézis valóságtartalmának vizsgálatához a Bayesi keretrendszerben először egy állítás előzetes valószínűségét kell specifikálni („a priori”), majd azt minden további

megfigyelés alapján frissíteni. A Bayesi valószínűség tehát a frekventista értelmezéssel ellentétben nem egy esemény várható gyakoriságát jelenti, hanem a hiedelem bizonyosságához rendelt érték („Bayesi valószínűség"). Az elméleti keretrendszer kialakítása már a 18. században megkezdődött (Thomas Bayes, 1701-1761), de széleskörű gyakorlati alkalmazása a nyolcvanas évekig korlátozott volt a bonyolult és számításigényes módszerek miatt. Az informatika és elsősorban az MCMC (Markov chain Monte Carlo mintavételező) módszerek fejlődésével azonban új lendületet kaptak a Bayesi módszerek kutatásai, és mára sok gyakorlati probléma esetén kifejezetten előnyt élveznek.

A nyolcvanas években felgyorsuló Bayesi következtetéssel kapcsolatos kutatások egyik fontos mérföldköve volt a Bayes-hálók leírása [130]. A Bayes-hálók számtalan tárgyterületen bizonyultak hatékony eszköznek, de az egészségügyi informatika, bioinformatika, számítógépes biológia (gén és fehérje szabályozási hálók, fehérje struktúrák, génexpressziós elemzések, GWAS adatelemzés, klinikai döntéstámogatás stb.) területén kiemelkedő eredményeket értek el. A Bayes-hálók körmentes irányított gráfok, melyek csomópontjai valószínűségi változókat (megfigyelhető vagy rejtett változókat), élei a változók feltételes függőségeit írják le (pl. betegség fennállását adott szimptómák függvényében). Azok a változók, amelyek között nem fut él, függetlenek egymástól. Minden csomóponthoz tartozik egy valószínűségi függvény, amelynek bemenetei a hozzá vezető irányított élek mentén szomszédos csomópontok (vagyis szülők) értékei, és kimenete a változó által reprezentált csomópont értékeinek az eloszlása. A Bayes-háló tetszőleges ismert csomópontjainak értékeit (tehát pl. a megfigyelt szimptómákat vagy mért értékeket) rögzítve az következtetni tud az ismeretlen csomópontok értékeire. A Bayes-háló struktúrája (tehát a változók függőségi viszonyai), valamint a csomópontok valószínűségi függvényei (tehát paraméterei) szakértők, pl. orvosok által előzetesen beállíthatóak, majd adathalmazokból a tárgyterületi tudás automatizáltan is frissíthető [131].

Kutatócsoportunk (Budapesti Műszaki és Gazdaságtudományi Egyetem Méréstechnika és Információs Rendszerek Tanszék és Semmelweis Egyetem Genetikai, Sejt és Immunbiológiai Intézet) részletesen tanulmányozta a Bayes-hálók alkalmazását egy célváltozóval (vagy célváltozó halmazzal) szoros kapcsolatban álló más változók

keresésére (változó kiválasztási probléma, „feature subset selection”) [132]. A kutatócsoportunk kidolgozta, és számos helyen sikerrel alkalmazta a Bayes-háló alapú Bayesi többszintű relevancia elemzés módszertanát („Bayesian-Network based Bayesian Multilevel Analysis of Relevance”, BN-BMLA), amely a Bayes-hálók adatokból történő tanításának egy speciális módszertanát használja fel arra, hogy a változók összefüggésrendszerének a legfontosabb jegyeit felfedje [133-138]. A módszer valójában az adathalmaz összefüggéseit jól reprezentáló, különböző lehetséges Bayes-hálók domináns jegyeit vizsgálja. A módszertan segítségével tehát nem feltétlenül teljes függőségi hálózatok tanulása a cél, gyakran ez a rendelkezésre álló adatok relatív kis mennyisége és a nagy számítási igény miatt nem is lehetséges. Ezzel szemben lehetséges az összefüggési hálózat lokális tulajdonságainak, tehát pl. adott csomópont környezetének elemzése, így a közvetlen függőségeinek szerkezete vagy más változókon keresztüli tranzitív függőségeinek elemzése, bizonyos esetekben az ok-okozati kapcsolat irányultságának vizsgálata [139]. A Bayesi keretrendszerben egy hipotézis valóságtartalmának értékelése tehát nem inverz módon, hipotézistesztéléssel történik, hanem a hipotézis valóságtartalma (pl. genetikai variánsok asszociációi a betegséggel) az előzetes hiedelmek („priori” információk, pl. orvosi szaktudás) és a mért adatok függvényében számszerűen kifejezhető („posteriori”), egy szintén nulla és egy közötti értékkel (ahol intuitívan a 0,5 fölötti érték „inkább valószínű”-t jelent, az alatt „inkább lehetséges”-t) [140]. Bár ez a hiedelem fogalom az emberi szubjektív gondolkodáshoz is jól illeszkedik, a megfelelő értelmezése a gyakorlatban meg is nehezíti a módszer használatát: a hipotézistesztéléssel ellentétben nincsenek egyezményes küszöbök (szignifikancia) melyek irányt mutatnak az eredmények megerősítésében vagy elvetésében, a Bayesi valószínűségek csak adott kontextusban, relativizálva értelmezhetőek, ezek a puha határok pedig gyakran nehezen összeegyeztethetőek a tudományos gondolkodással.

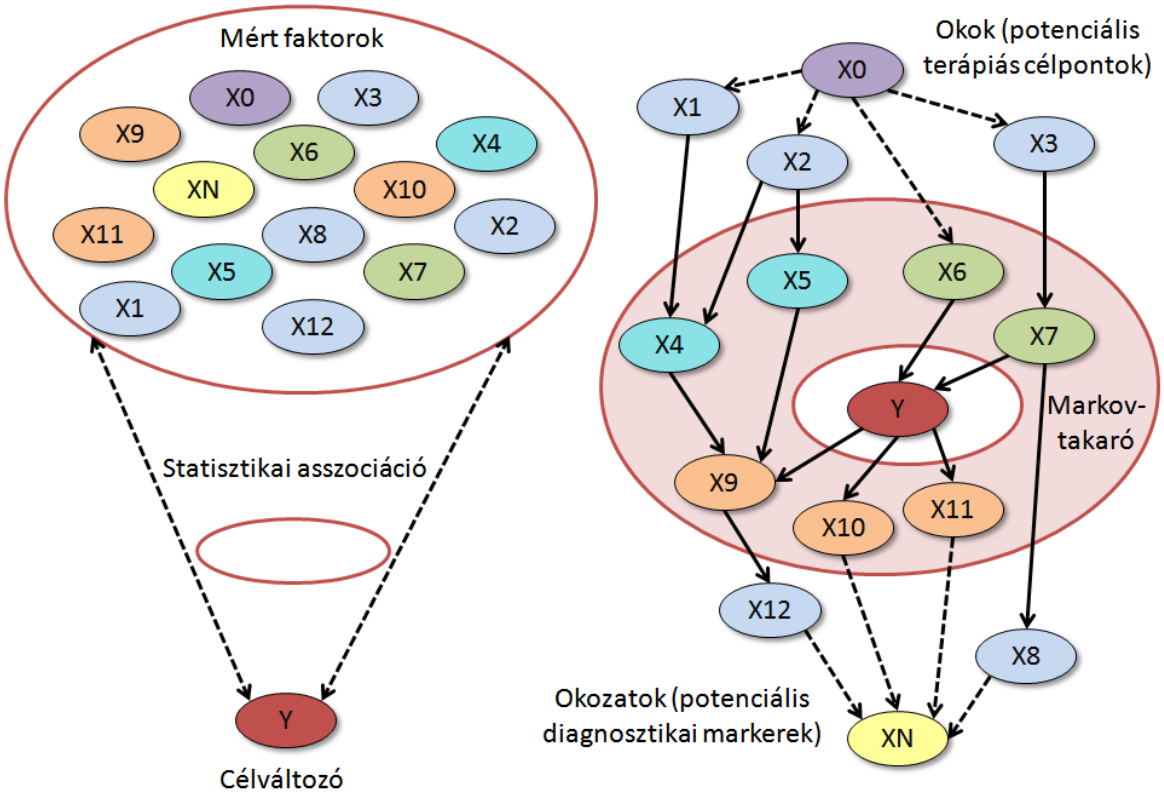
A BN-BMLA módszertan egy központi eleme a Markov-takaró fogalma ("Markov Blanket”, MB), mely a Bayes-hálóban azon minimális számú változót tartalmazó algráf, amelyek egy adott célváltozót (pl. asztma fenotípust) statisztikai értelemben elszigetelnek az összes többi változótól a Bayes-hálóban. A Markov-takaró változói a célváltozónak a

szülő csomópontjai (tehát amelyekből irányított él vezet a célváltozóba), a gyermek csomópontjai (tehát amelyekbe irányított él vezet a célváltozóból), illetve a gyermek csomópontok szülei; ismeretük mellett a többi változó nem befolyásolja a célváltozó értékét. A BN-BMLA elemzés elsődleges célja, hogy több szinten meghatározza az MB szerkezeti valószínűségeit. Ez a gyakorlatban azt jelenti, hogy az adathalmaz alapján minden lehetséges gráf jegyre meghatározza az alábbiakat:

1. Markov Blanket Membership (MBM): minden változóra egyenként meghatározza a Markov-takaróba tartozás posteriori valószínűségét

2. Markov Blanket Set (MBS): minden lehetséges változó halmazra meghatározza annak a posteriori valószínűségét, hogy azok és csak azok a változók tartoznak a Markov-takaróba. Az MBS további elméleti kiterjesztése a subMBS, mely annak a valószínűsége, hogy egy változó halmaz része az MBS-nek; míg a supMBS annak a valószínűsége, hogy egy változó halmaz tartalmazza az MBS-t.

3. Markov Blanket Graph (MBG): minden lehetséges algráfra meghatározza annak a posteriori valószínűségét, hogy az algráf megegyezik a Markov-takaró gráffal.



8. ábra - A bal oldali ábra a változók strukturális összefüggési rendszerének figyelmen kívül hagyását illusztrálja, míg a jobb oldali a rendszer alapú modellezést és az összefüggések lehetséges fajtáit. Y jelöli a célváltozót (pl. asztma), míg X0, X1,…, Xn a különböző mért faktorokat (pl. környezeti hatások, anamnézis, genetikai polimorfizmusok, egyéb labor eredmények). A különböző színek a kapcsolatok eltérő fajtáit jelölik: X0 – ok (lila); Xn – okozat (sárga); X6 – direkt ok (zöld); X9 – direkt okozat (narancs); X4 – interakciós tag (ciánkék); X1 – egyebek (világoskék). A sima élek közvetlen, a szaggatott élek áttételes függőségeket jelölnek. Azon csomópontokat, melyek direkt okok, direkt okozatok, vagy interakciós tagok erősen releváns faktoroknak nevezzük, ezek statisztikai értelemben izolálják a célváltozót a többitől (piros gyűrű) és a célváltozó Markov-takarójának nevezzük őket. Az ábrát Hullám és munkatársainak cikke alapján készítettem [141].

Az MB szerkezetének elemzésével tehát lehetséges az egyszerű asszociáció fogalmának specifikusabb karakterizálása, pl. direkt, tranzitív, vagy egyéb zavaró hatás okozza-e a statisztikai asszociációt, a mért faktorok oki vagy okozati viszonyban vannak-e a célváltozóval, valamint lehetséges ezen állítások posteriori valószínűségeinek meghatározása is. Ez a nagy áteresztőképességű mérési eredmények bonyolult összefüggési rendszerének értelmezésében egy óriási lépés a szokásos frekventista asszociáció fogalmához képest (8. ábra). Az asztma genetikai hátterének kutatása során ezt a két módszertant alkalmaztam a mérési eredmények értelmezéséhez.

2.4.2 Virtuális szűrési technikák adatelemzési módszerei

Az új vegyületek kemoinformatikai vizsgálata, különböző paramétereinek jóslása, hipotézisek generálása egy fontos koncepcióra épít, amely egyben a vegyület könyvtárak virtuális szűrésének is egy alapvető feltevése: Johnson és munkatársai 1989-ban egy tanulmányban általánosságban is leírták és bizonyították, hogy (elsősorban kémiailag) hasonló molekulák várhatóan hasonló biológiai aktivitást is mutatnak [142]. Így a nagy áteresztőképességű virtuális szűrésekben is a hasonló szerkezetű, ismert molekulák alapján jósolható a biológiai aktivitás, és sok esetben pl. a potenciális célpontok, mellékhatások, indikációk is. A molekuláris célpontok predikciója egy alaposan tanulmányozott terület a hatóanyag újrapozicionáló módszerek között is [143-145], de a vegyületekkel kapcsolatos nagy mennyiségű heterogén adat integrálása és felhasználási lehetőségei egy relatíve új terület.

A hatóanyagok újrapozícionálását két tudományosan jól leírt koncepció teszi lehetővé, ezek két, alapvetően különböző lehetőséget vetnek fel. Az első szerint a legtöbb hatóanyag több molekuláris célpontra hat, így gyakran érdemes a másodlagos célpontokat keresni és az azzal kapcsolatos indikációkat is megvizsgálni („off-target repositioning”). A másik koncepció szerint a legtöbb molekuláris célpont több biológiai szabályozási mechanizmusban is részt vesz, így a célpont eltérő szabályozási útvonalaira fókuszálva új indikációk is felvetődhetnek („on-target repositioning”) [146].

A legtöbb hatóanyag újrapozicionálási projekt vagy egy hatóanyagtól (célponttól) indul ki, vagy egy indikációtól (betegségtől), és azt célozza meg, hogy azonosítsa a másikat (hatóanyag vezérelt vagy indikáció vezérelt módszerek). A folyamat különbözhet a kutatást végző szervezet céljaitól függően is: kisebb cégek szisztematikus új indikáció keresést végeznek piacon lévő hatóanyagokra, míg a nagy gyógyszeripari cégek elsősorban a szakterületükhöz tartozó indikációkból vagy fejlesztés alatt álló gyógyszerekből indulnak ki.

A nagy áteresztőképességű szűrési technikák (HTS), ahogy az originális gyógyszerfejlesztésben, úgy hatóanyag újrapozícionálásban is fontos szerepet töltenek be:

az alapvető biokémiai mérésektől, a célponthalászó („target figshing”) módszereken át az in vitro sejtes esszékig és in vivo fenotípizáló rendszerekig. Általában a módszerek spektrumának egyik végén egy ismert molekuláris célpont – betegség kapcsolatból indulnak ki, és a feladat egy olyan hatóanyag keresése, amely az adott célponthoz köt; a spektrum másik végén a kutatás nem a célpontra fókuszál, mindössze egy olyan hatóanyag a cél, amely a keresett fenotípust eredményezi (betegséget kezeli).

A legtöbb hagyományos újrapozícionálási módszer arra törekszik, hogy hatóanyag – indikáció párokat keressen HTS technikák adatainak az elemzésével, de a kísérletes validáció már drága, munkaigényes és jelentős infrastruktúrát igényel. Így a számítógépes, virtuális szűrési technikák egyre elterjedtebbek, akár önmagukban, akár egy HTS projekt részeként alkalmazva [147].

A vegyületek hasonlóság alapú keresése és sorrendezése is egy régóta tanulmányozott terület. Willett és munkatársai fektették le az alapjait a kilencvenes évek végén, amikor a feladatot különböző alproblémákra bontották [148, 149]. Az egyszerű hasonlósági keresés („similarity search”) teljes molekulák keresésére fókuszál egyetlen referencia struktúrát és hozzá egyetlen hasonlósági metrikát alkalmazva, anélkül, hogy kísérletet tenne funkcionális al-struktúrák azonosítására. Ebben a kontextusban az adatfúzió („data fusion”) egyszerű hasonlósági keresések eredményeinek kombinálására, összefésülésére utal, ahol egyetlen referenciastruktúrát használunk a kereséshez, de több hasonlósági metrikával, tehát több különböző tulajdonság alapján végzünk keresést. A harmadik megközelítés, a csoportfúzió („group fusion”) egyetlen hasonlósági metrikát használ a kereséshez, de több referencia molekulára (vagyis a referencia molekuláktól vett távolságokat átlagoljuk valamilyen módon). Más kutatócsoportok, mint pl. Campillos és munkatársai [101], kiterjesztették a keresési keretrendszert nem-kémiai leírókra is; Arany és munkatársai pedig kifejlesztettek egy automatikus súlyozási metodológiát, mely az információforrások fontosságát az adott keresésre specifikusan, kernel alapú fúzióval határozza meg [150].

Eckert és munkatársai megvizsgálták a molekula hasonlóságon alapuló virtuális szűrési módszereket és megállapították, hogy nincsen egyetlen, általánosságban is javasolható

technika, de sok módszert találtak igen hatékonynak más-más szempontból [151]. Svensson és munkatársai pedig összehasonlították a virtuális szűrések adatfúziós algoritmusait (tehát vegyületsorrend összefésülési módszereket) és hasonló következtetésekre jutottak [152]. Nincs csodaszer, minden módszernek megvannak a maga erősségei, a kulcskérdés az új technikák megfelelő beillesztése a klasszikus farmakológiai gyakorlatok mellé [153]. Ez viszont gyakran igen nehézkes, mert az elemzési módszerek matematikai, statisztikai, adatbányászati eredményei gyakran nehezen értelmezhetőek a klasszikus farmakológia fogalomrendszerében.

Ezért az értekezésem egyik fő témája a vegyület hasonlósági és sorrendezési módszerek egy kiegészítő technikája, amely segít az értelmezésben, és az ismert farmakológiai fogalomrendszerek szintjén nyújt többlet információkat a vizsgálat tárgyát képező vegyületről. A javasolt technika, a feldúsulás elemzés (Set Enrichment Analysis) egy informatikailag jól skálázható, statisztikailag robosztus és emberi szakértő számára könnyen értelmezhető post-hoc értelmezési módszer, melynek matematikai hátterét molekuláris biológia területén már évek óta sikerrel használják (GSEA [154]).

A feldúsulás elemzés első és egyetlen ismert felhasználását farmakológiai kontextusban Thibault és munkatársai publikálták 2010-ben. Ebben a módszert egy speciális esetben, korai fázisú gyógyszerkutatásban alkalmazták HTS adatok strukturális elemzésére [155, 156]. A szerzők módszere azt a célt szolgálja, hogy a molekulakönyvtár elemzése során olyan molekuláris szerkezeti jegyeket emeljen ki, melyek gyakrabban fordulnak elő az elvárt biológiai aktivitást mutató molekulák között, és így esetleg felelősek lehetnek a vizsgált hatásért. Az értekezésem során bemutatom, hogy a technika mind a molekulakönyvtárak entitásai, mind a mért adatok spektruma, mind a vizsgált jelzések sokszínűsége tekintetében lényegesen szélesebb körben alkalmazható, valójában egy univerzális farmakológiai adatintegrációs módszerként is használható.

DOI:10.14753/SE.2015.1757

3 Célkitűzések

3.1 Asztma genetikai hátterének vizsgálata adatmérnöki eszközökkel

A post-genom korszak sok értelemben nem váltotta be eddig a hozzá fűzött reményeket, az eredmények nagyon lassan jutnak el a klinikai alkalmazásig. A multifaktoriális betegségek összetettsége meghaladta a korábban feltételezetteket, ezért így számos molekuláris biológiai és módszertani kérdés vár megoldásra. Az asztma kutatásaim célja elsősorban új, asztmára hajlamosító gének, esetleges terápiás célpontok azonosítás volt kandidáns gén asszociációs vizsgálatok tervezésével egy korábbi, ovalbumin indukált egér asztmamodellen végzett teljes genom expressziós tanulmány alapján (GSE11911) [70]. A tanulmányaim ezen belül három kérdésre fókuszáltak:

1. A genetikai asszociációs vizsgálatok tervezése, különösképpen a mérendő genetikai polimorfizmusok méréstechnikailag és információelméletileg optimális halmazának kiválasztása hogyan automatizálható számítógépes eszközökkel?

2. Egy modern eszközökkel megtervezett, többszintű validációra építő, asztmával kapcsolatos kutatás során milyen új genetikai variánsokat, oki tényezőket, terápiás célpontokat lehet azonosítani?

3. A hagyományos frekventista statisztika eszköztára mellett egy rendszerszintű modellezés, a Bayes-háló alapú Bayesi több szintű relevancia elemzés eszköztára mennyiben segítheti az oki tényezők feltárását?

DOI:10.14753/SE.2015.1757

3.2 A farmakológiai információtömeg kiaknázása feldúsulás elemzéssel

Vegyületekről hatalmas mennyiségű heterogén adat- és információtömeg érhető el publikusan és cégek belső adatbázisaiban, melyek különböző metszeteit számtalan módon használja fel a gyógyszeripar a molekula optimalizációtól a klinikai biomarkerek fejlesztéséig. Azonban az egyre növekvő mennyiségű adat- és információtömeg globális értelmezésével és felhasználási lehetőségeivel kapcsolatban nagyon sok a nyitott kérdés [114]. A tanulmányaim során azt vizsgáltam meg, hogy egy molekuláris biológiai adathalmazok elemzésében bevált matematikai módszert, a feldúsulás elemzést milyen módon lehet átültetni farmakológiai területre, és az alábbi két kérdésre kerestem a választ.

1. A feldúsulás elemzés módszertana hogyan alkalmazható az farmakológiai információtömeg integrálására, a gyenge jelzések elemzésére, és melyek a felhasználás korlátai?

2. Milyen funkciókkal segítheti ez a módszertan egy adott hatóanyag fejlesztését, és hogyan ágyazhatóak be ezek a technikák a gyógyszerfejlesztési gyakorlatba?

DOI:10.14753/SE.2015.1757

4 Módszerek

4.1 Asztma genetika

4.1.1 Genotipizált populáció: betegek és kontrollok

A genotipizálási tanulmányba bevont magyar (kaukázusi) populáció 671 rokonsági viszonyban nem álló személyből állt (311 asztmás gyermek és 360 egészséges kontroll). Az asztma diagnózist minden esetben szakorvos állította fel a Global Initiative for Asthma irányelvek alapján (http://www.ginasthma.org/). Az asztmás betegeket véletlenszerűen választottuk ki a Budai Gyermekkórház és Heim Pál Gyermekkórház allergológiai szakambulanciájáról. Az atópiát prick bőrpróba (Lofarma S.p.A.) és specifikus IgE szint mérés (RAST test, Roche kit) segítségével diagnosztizálták. Az atópiát akkor állapították meg, ha a prick teszt legalább egy allergénre pozitív volt (legalább 3 mm átmérőjű csalángöb) és/vagy emelkedett teljes vagy specifikus IgE szint volt mérhető. A teljes szérum IgE szintet és a specifikus IgE szintet 100-nál több allergénre a 3gAllergy™ vér teszttel állapították meg Immulite 2000 Immunoassay System (Siemens Healthcare Diagnostics; Deerfield, IL, USA) rendszer segítségével. A szérum IgE szint normális, illetve magas besorolást a következő, kor szerinti referenciák alapján állapítottuk meg: 0–1 éves, <15 kU/l; 1–5 éves, <60 kU/l; 5–10 éves, <90; felnőtt, <100 kU/l.

A kontroll gyermekeket véletlenszerűen választottuk ki a Budai Gyermekkórház Ortopédiai Osztályáról, illetve a Heim Pál Gyermekkórház Urológiai Osztályáról. A kontroll csoportba tartozó gyermekek asztma tünetektől mentesek voltak; olyan enyhe mozgásszervi (pl. lúdtalp vagy gerincferdülés) vagy urogenitális (pl. fitymaszűkület) panaszaik voltak, melyek nem igényeltek gyógyszeres kezelést. Idősebb kontroll személyeket is bevontunk a vizsgálatba, ők egészséges véradók voltak, akiknek korábban sosem volt asztmája a kitöltött kérdőívek szerint (2. táblázat). A betegek és az egészségesek közötti kisebb mértékű átlagéletkor és nemi eltérésből adódó esetleges torzító hatásokat statisztikai adjusztálással és különböző ellenőrzéssekkel kezeltük.

2. táblázat - A genotipizált populáció paraméterei [157].

Klinikai tulajdonságok	Betegek (n=311)	Kontrollok (n=360)
Age (év) ±SD	10,58±4,79	21,71±13,89*
Nem (férfi=1/nő=0)	203/108	181/179*
Atópiás asztma	195 (62,7%)	-

Tanulmányainkat elsősorban azért végeztük asztmás gyermekeken, mert multifaktoriális betegségek esetén gyermekkorban a betegség kialakulásában erősebb a genetikai háttér szerepe a környezeti hatásokhoz képest, így a genetikai asszociáció kimutatása gyermekekben nagyobb eséllyel lehetséges [49]. Ezzel szemben a gyermek kontrollok esetén nem állítható, hogy azoknak később sem alakulnak ki asztmás tüneteik, így indokolt lehet idősebb kontrollok beválogatása is, mert azok asztma szempontjából szűrt kontrolloknak tekinthetőek.

4.1.2 Indukált köpet expressziós vizsgálat

Az indukált köpet expressziós vizsgálatba 34 felnőtt személyt vontunk be, de 11 személyt ki kellett zárni a köpet minősége miatt. Így végül a tanulmányt 13 asztmás és 10 kontroll személy eredményeire alapoztuk, a csoport adatait Ungvári és munkatársai egy korábbi cikkéből vettem át [158] (3. táblázat).

3. táblázat - Az expressziós vizsgálatba bevont személyek adatai.; a: Az asztma súlyossági fokok meghatározása a GINA kritériumok alapján történt; b: A csoportok közötti eltérés szignifikancia szintje.; * jelzi, hogy az adott paraméter szintjében mérhető eltérés a vizsgált csoportok között statisztikailag szignifikáns, p < 0,05.; n: esetszám [158]

	Betegek n (%)	Kontrollok n (%)	p-érték[b]
Esetszám	13	10	-
Életkor (átlag ± SD, év)	34,6 ± 10,1	30,1 ± 4,1	-

Nem (férfi/nő)	6/7	5/5	-
Dohányzási szokás (igen/nem)	7/6	6/4	-
Asztma súlyossága[a]			
Enyhe	4 (30)	-	-
Mérsékelt és súlyos	9 (70)	-	-
Allergia (igen/nem)	10/3	5/5	-
Köpet eozinofil%	12,2 ± 11,3	0 ± 0	*
Köpet neutrophil %	25,1 ± 17,7	20,0 ± 10,0	-
Köpet makrofág %	61,7 ± 15,0	74,4 ± 8,4	-
Köpet bronchialis epithél sejt %	1,3 ± 1,7	5,6 ± 5,3	*

Az asztmás betegeknek közepestől súlyosig terjedő asztmájuk volt, más légzőszervi betegséget vagy fertőzést nem diagnosztizáltak. A betegek FEV1/FVC% értéke (1 másodperc alatti erőltetett kilégzési levegőtérfogat a teljeshez képest) legalább 70% volt, a metakolinra adott 20%-os légzési térfogat csökkenéséhez (PC20) tartozó metakolin koncentráció pedig kevesebb, mint 16 mg/ml. Az egészséges kontroll személyek a kutatásban résztvevő egyetemekről kerültek ki. Az alanyoknak nem volt korábban ismert krónikus légúti megbetegedése, a FEV1/FVC % értékük 80% fölött volt, valamint a metakolinra adott válaszuk normális volt (PC20 > 16 m/ml). A két csoport nem különbözött számottevően kor, nem, dohányzási szokások és allergia szempontjából. Az FVC („forced vital capacity”) és FEV1 („forced expiration volume in the first second”) értékeket spirométerrel mértük (PDD-301/s, Piston Inc, Budapest, Hungary).

A vizsgált személyek - vagy kiskorúságuk esetén szüleik – írásos beleegyezésüket adták a kutatásokban való részvételhez. A vizsgálatokat a Magyar Etikai Bizottság (Egészségügyi

Tudományos Tanács Tudományos és Kutatásetikai Bizottság, azaz ETT TUKEB) hagyta jóvá, és azok minden esetben megfeleltek a Helsinki Deklarátumban megfogalmazott alapelveknek.

4.1.3 Kandidáns gének és polimorfizmusok kiválasztása és genotipizálása

4.1.3.1 Kísérlettervező rendszer

A polimorfizmusok nagy áteresztőképességű mérése primer extenziós reakcióval történt, melynek méréstechnikai tervezését már korábbi tanulmányaim során is részletesen vizsgáltam [159, 160]. Az asztmával kapcsolatos első tanulmányunk során (Ungvári és munkatársai [133]) a polimorfizmusok genotipizálását a saját laboratóriumunkban végeztük a GenomeLab SNPstream (Beckman Coulter) platform segítségével, így a méréstechnikai optimalizálás is a tanulmány része volt.

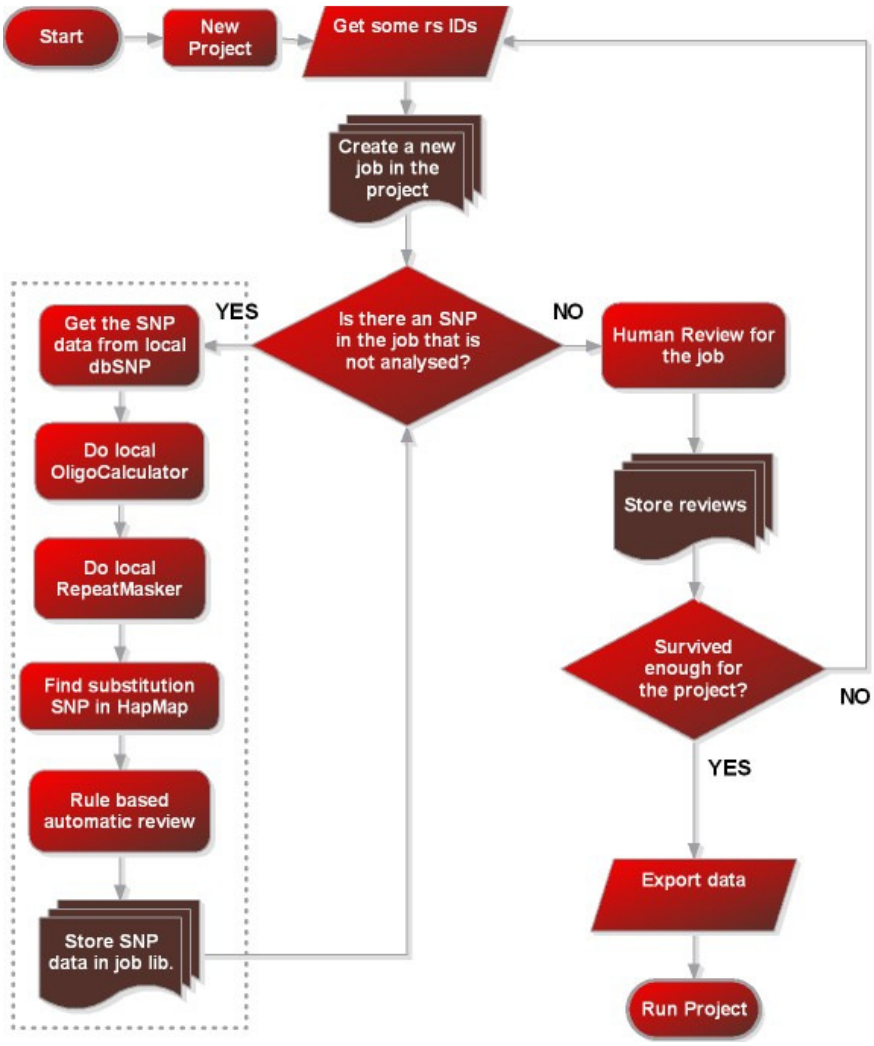
A primer extenziós technikákban egy olyan primert használnak, amely a PCR primerekhez hasonló és közvetlenül a polimorfizmust megelőző nukleotidnál végződik. A DNS amplifikálása után a reakciótérbe keverik az extenziós primert és a polimorfizmus két lehetséges típusával komplementer, fluoreszcensen (vagy más módon) jelölt nukleotidot tartalmazó oldatot. A DNS replikációját végző enzim a jelölt nukleotidok közül csak az egyiket – amelyik a polimorfizmussal komplementer – képes beépíteni a primer végére, majd a reakció terminál. A reakció eredményének leolvasáshoz végül a primert egy DNS chip lapka felszínére hibridizálják. A műveletnek méréstechnikai szempontból több korlátozó tényezője van, amely bizonyos polimorfizmusok mérését, vagy közös reakció térben történő mérését kizárja.

A követelmények első csoportja az egyes polimorfizmusokra vonatkozik: a primerek kötődésének pontosságát, vagyis a reakció specificitását több tényező is ronthatja. Egyrészt, a templáton a rövid ismétlődő szakaszok, illetve más polimorfizmusok a primer beépülésének tévesztéséhez, elcsúszásokhoz vezethetnek. Másrészt, amennyiben a templát a genomban máshol is előforduló szakasz, úgy a kívánt helyre való beépülése sem garantálható. Így olyan polimorfizmusok nem mérhetőek ezzel a technikával, ahol a polimorfizmus közvetlen környezetében nem található megfelelő méretű,

polimorfizmusoktól és (rövid vagy hosszú) ismétlődésektől mentes templát szakasz a primer bekötődéséhez. A gyakorlati toleranciaküszöbök kvantitatív elemzésével korábbi munkáim során foglalkoztam részletesebben [159, 160]. A méréstechnikai szempontból problémás polimorfizmusok helyett lehetőleg statisztikailag kapcsolt más polimorfizmust kell választatni.

A követelmények második csoportja az egy reakciótérben zajló mérhető polimorfizmusokra vonatkozik. Egyrészt, a primer extenziós reakciók során a primerek olvadáspontjának lehetőleg egyezniük kell. Ez azt jelenti, hogy a primerekben a guanin+citozin (3 hidrogénhíd) és az adenin+timin (2 hidrogénhíd) arányának közel azonosnak kell lennie. Másrészt, egy reakciótérben nem mérhetőek hasonló szakaszokat tartalmazó primerek, mert ez a primerek helytelen hibridizálásához vezethet. Ez az egymásnak kissé ellentmondó két követelmény egy-egy multiplex mérésbe bevonható polimorfizmusok halmazának egyedi megtervezését igényli. Tehát a kiválasztott gének méréstechnikai szempontból is megfelelő polimorfizmusainak azonosítása egy aprólékos szűrési folyamat eredménye.

DOI:10.14753/SE.2015.1757



9. ábra - A polimorfizmusok kiválogatásának munkafolyamat ábrája a TIGER kísérlettervező rendszer angol nyelvű dokumentációjából [159]. Az emberi szakértő egy új elemzési projektet indít, melybe polimorfizmusokat válogathat be. A kiválogatott polimorfizmusokat a rendszer elemzi, a szakértő manuálisan elfogadja vagy elveti az elemzés eredményeit. Amennyiben a műszeres méréshez elegendő polimorfizmus került elfogadásra, úgy a munkafolyamat befejeződik, ellenkező esetben a szakértő új elemzendő polimorfizmusokat válogat össze.

A folyamat emberi szakértők számára hosszadalmas, aprólékos, monoton, sok hibázási lehetőséggel, valamint számos adatbázis (gének, polimorfizmusok, szekvenciák, stb.) és szoftver (ismétlődések szűrése, kapcsoltsági vizsgálatok, stb.) használatát igényli. Korábbi munkáim [159, 160] során létrehoztam és leírtam egy olyan szoftver rendszert és módszertant (TIGER kísérlettervező rendszer), amely a folyamathoz szükséges adatbázisokat integrálja és a legtöbb lépését automatizálja (9. ábra).

4.1.3.2 Kandidáns gének és polimorfizmusok kiválasztása

A második asztmával kapcsolatos tanulmányunk során (Temesi és munkatársai [157]) a primer extenziós mérésekhez az iPLEX Gold MassARRAY (Sequenom) platformot használtuk. Ebben az esetben a méréseket a McGill University and Génome Québec Innovation Centre (Montréal, Canada) végezte. A méréstechnikai optimalizálást részben a szolgáltató végezte a saját szoftvereivel, így a TIGER kísérlettervező rendszert csak korlátozottan használtuk fel a polimorfizmusok kiválogatásához.

A tanulmányt egy korábbi, ovalbumin-indukált egér asztmamodell génexpressziós vizsgálat eredményeire alapoztuk, melyet az alábbiakban röviden összefoglalok (további részletek elérhetőek Tölgyesi Gergely doktori étkezésében és publikációjában [70]). A korábbi kísérletbe 6-8 hetes, BALB/c, nőstény, patogén-mentes egereket vontak be, melyek kizárólag ovalbumin-mentes táplálékot kaptak. Az egerek egyik csoportját ovalbuminra érzékennyé tették és az allergénnel kezelték, a másik csoporttal ugyanezt az eljárást végezték el placebo segítségével. A 28-31. napon, 4-24 órával az első vagy harmadik allergénterhelés után az egereket elaltatták, BAL (bronchoalveolar lavage) folyadékot izoláltak és a tüdőket további vizsgálat céljából eltávolították. RNS-t izoláltak a tüdőszövetekből, majd génexpressziós mérést végeztek Agilent Whole Mouse Genome Oligo Microarray 4 x 44 K chipek segítségével (GEO adatbázis recordazonosító GSE11911). A microarray mérések eredményét nem párosított t-teszttel, és egyutas ANOVA-val (one-way analysis of variance) végezték. Több mint 1000 transzkriptum mutatott legalább kétszeres, statisztikailag szignifikánsan eltérő expressziót a vizsgált három csoportban a kontroll csoporthoz képest.

A génexpresszió eredményeit használtuk fel olyan gének kiválasztására, melyek ortológjai kiinduló pontjai lehetnek humán asztma és atópia tanulmánynak. Olyan géneket választottunk ki további vizsgálatra, melyek expressziós változása statisztikailag erősen szignifikáns volt (p-értékek 0,02 alatt). A kiválasztott gének listáját rendeztük az expressziós változás abszolút értékének mértéke alapján, majd szubjektív szempontok alapján szűrtük a listát, pl. az asztmával való potenciális kapcsolat, a szakirodalom korábbi

megállapításai és a tudományos újdonságereje alapján. A kiválasztott 60 gén listája a 4. táblázatban látható.

A TIGER kísérlettervező rendszer mellett a UCSC Genome Browsert használtuk a gének mérendő polimorfizmusainak (SNP-k) kiválasztásához [161]; előnyben részesítettük a promoter, missense és UTR régiókban lévő polimorfizmusokat. A kiválasztott gének vagy régiók legjobb lefedéséhez felhasználtuk a polimorfizmusok LD (linkage disequilibrium) kapcsoltsági adatait (International HapMap Project [162] , Haploview software [163]).

A kiválasztott 90 polimorfizmus listája a 4. táblázatban látható. A DNS izoláció vérből történt az iPrep PureLink gDNA Blood Kit, iPrep Purification Instrument (Invitrogen) rendszerrel.

4. táblázat – Az értekezésben vizsgált gének és polimorfizmusok adatai (Genome Variation Server 137 based on dbSNP build 137, June 2013, http://gvs.gs.washington.edu/GVS/) [157] és a minor allél frekvencia (MAF) eloszlása. A GVS adatbázis a dbSNP funkcionális kategóriáit használja amennyiben tartozik bejegyzés a polimorfizmushoz, ellenkező esetben egy saját kategóriát használ.

Gén	SNP	Pozíció	GVS funkcionális kategória	Allélok	Konrt. MAF	Eset MAF
ACSBG1	rs3813577	15:78527253	near-gene-5(GVS)	C/T	37,94%	33,39%
AGR2	rs706072	7:16844663	utr-variant-5-prime(dbSNP)	T/C	18,75%	19,84%
AGR2	rs1459564	7:16846146	near-gene-5(GVS)	A/G	33,28%	30,16%
AGR2	rs706075	7:16846354	near-gene-5(GVS)	G/T	25,22%	24,09%
AGR2	rs10261011	7:16856487	intergenic(GVS)	G/A	51,02%	49,02%
AIF1	rs2857600	6:31582287	near-gene-5(GVS)	T/C	10,66%	3,26%
ATP6V0A4	rs10258719	7:138455988	missense(dbSNP)	A/G	28,34%	29,84%
BIRC5	rs1508147	17:76222588	near-gene-3(GVS)	A/G	43,67%	38,01%
C1QC	rs6690827	1:22967496	near-gene-3(GVS)	A/G	29,76%	35,32%

C1QC	rs294179	1:22974928	downstream-variant-500B(dbSNP)	T/C	42,65%	48,67%
CCL2	rs2530797	17:32586094	near-gene-3(GVS)	C/T	33,53%	37,03%
CCL8	rs1821142	17:32649988	near-gene-3(GVS)	T/C	3,60%	4,58%
CCNE1	rs7257330	19:30301823	near-gene-5(GVS)	A/G	39,49%	38,62%
CD6	rs1050922	11:60785352	synonymous-codon(dbSNP)	G/A	31,54%	33,39%
CD84	rs1055880	1:160517692	utr-variant-3-prime(dbSNP)	T/C	34,24%	33,50%
CHIA	rs17027410	1:111861822	synonymous-codon(dbSNP)	A/G	11,64%	11,26%
CLEC4E	rs7299659	12:8696661	intergenic(GVS)	A/G	19,05%	18,61%
COL6A2	rs2839110	21:47538960	missense(dbSNP)	G/A	22,92%	24,17%
CREB3L4	rs11264743	1:153941514	missense(dbSNP)	T/C	31,12%	30,03%
CXCL1	rs3117604	4:74734668	near-gene-5(GVS)	T/C	29,30%	31,25%
CXCL5	rs352045	4:74864687	near-gene-5(GVS)	T/G	11,77%	13,11%
CSF2	rs27438	5:131413255	near-gene-3(GVS)	A/G	22,73%	22,50%
E2F7	rs310830	12:77419593	synonymous-codon(dbSNP)	G/A	10,06%	13,87%
FABP3	rs16834408	1:31837942	near-gene-3(GVS)	A/G	15,92%	22,02%
FABP3	rs10914367	1:31846206	near-gene-5(GVS)	A/G	24,09%	22,20%
FXYD4	rs4245604	10:43866528	near-gene-5(GVS)	A/C	32,41%	31,64%
GPR160	rs4955711	3:169753570	intergenic(GVS)	G/A	27,96%	28,14%
ICOS	rs3923093	2:204798020	intergenic(GVS)	T/C	24,15%	25,17%
IL17RB	rs2289205	3:53878616	intron-variant(dbSNP)	T/C	29,94%	31,38%
IL1A	rs1878320	2:113544467	upstream-variant-2KB(dbSNP)	C/T	30,15%	28,31%

IL1A	rs3783520	2:113544339	upstream-variant-2KB(dbSNP)	T/C	29,41%	27,93%
IL1B	rs16944	2:113594867	upstream-variant-2KB(dbSNP)	G/A	39,02%	33,50%
IL1RL1	rs12905	2:102960007	utr-variant-3-prime(dbSNP)	A/G	25,00%	26,99%
IL6	rs2069827	7:22765456	utr-variant-3-prime(dbSNP)	T/G	11,05%	10,98%
IL6	rs2069832	7:22767433	intron-variant(dbSNP)	A/G	38,52%	41,45%
ITGAX	rs11150614	16:31366016	upstream-variant-2KB(dbSNP)	A/G	29,31%	28,95%
ITLN1	rs4656958	1:160856964	intergenic(GVS)	A/G	32,56%	25,90%
ITLN1	rs2274910	1:160852046	intron-variant(dbSNP)	T/C	33,43%	27,30%
KLF15	rs1358087	3:126078890	intergenic(GVS)	C/T	36,36%	35,88%
KLF15	rs9862203	3:126058362	intergenic(GVS)	A/G	23,48%	28,67%
LAPTM5	rs3762296	1:31231386	near-gene-5(GVS)	G/A	48,02%	45,07%
LGALS3	rs7160110	14:55594635	upstream-variant-2KB(dbSNP)	G/A	40,54%	37,16%
LGMN	rs9791	14:93170993	synonymous-codon(dbSNP)	T/C	42,21%	35,22%
LY86	rs760894	6:6656198	near-gene-3(GVS)	G/A	32,48%	30,62%
LY9	rs509749	1:160793560	missense(dbSNP)	A/G	50,94%	47,05%
LY9	rs474131	1:160793442	synonymous-codon(dbSNP)	A/G	46,44%	39,29%
MAFB	rs6102095	20:39320751	intergenic(GVS)	A/G	17,25%	15,24%
MAP3K6	rs11247639	1:27679692	intron-variant(dbSNP)	G/A	32,65%	30,54%
MARCO	rs6748401	2:119698057	near-gene-5(GVS)	G/A	50,45%	44,43%

MAT1A	rs3827869	10:82031278	downstream-variant-500B(dbSNP)	T/C	21,00%	16,44%
MAT1A	rs10887711	10:82034842	synonymous-codon(dbSNP)	A/G	31,53%	39,53%
MAT1A	rs10887708	10:82027988	intergenic(GVS)	A/G	28,40%	30,51%
MAT1A	rs10749550	10:82031197	downstream-variant-500B(dbSNP)	A/G	35,99%	37,54%
MKI67	rs11016071	10:129901393	missense(dbSNP)	C/T	14,88%	16,83%
MKI67	rs10082432	10:129901722	synonymous-codon(dbSNP)	A/G	17,70%	18,12%
MKI67	rs8473	10:129899578	missense(dbSNP)	C/T	45,22%	45,73%
MKI67	rs2152143	10:129906980	missense(dbSNP)	A/G (ritkán T/C)	30,78%	28,85%
MS4A7	rs10750936	11:60144180	near-gene-5(GVS)	G/A	36,76%	34,75%
OSGIN1	rs2432561	16:83982670	intergenic(GVS)	A/G	16,74%	10,82%
PPARGC1B	rs32588	5:149200043	synonymous-codon(dbSNP)	C/T	23,46%	14,73%
PTPN7	rs4359077	1:202129112	utr-variant-5-prime(dbSNP)	A/G	10,59%	10,70%
RETNLB	rs3811687	3:108476519	near-gene-5(GVS)	T/C	29,97%	29,61%
RETNLB	rs10933959	3:108476205	near-gene-5(GVS)	G/A	23,11%	19,18%
RETNLB	rs9870145	3:108477874	near-gene-5(GVS)	T/A	14,93%	16,45%
RETNLB	rs11708527	3:108475974	missense(dbSNP)	A/G	29,94%	29,67%
SAA1	rs4638289	11:18285774	intergenic(GVS)	A/T	38,37%	41,48%
SAA1	rs11603089	11:18282051	intergenic(GVS)	G/A	16,13%	15,08%
SAA2	rs7130337	11:18270605	near-gene-5(GVS)	A/G	24,42%	26,66%

SCIN	rs3173628	7:12627245	intron-variant(dbSNP)	A/G	41,28%	49,83%
SCIN	rs2240572	7:12610594	missense(dbSNP)	G/A	48,69%	37,66%
SCIN	rs2240571	7:12609988	near-gene-5(GVS)	C/G	39,80%	50,00%
SCIN	rs3735222	7:12609679	near-gene-5(GVS)	A/G	48,69%	37,87%
SIGLEC1	rs625372	20:3684729	missense(dbSNP)	T/C	32,04%	31,94%
SLAMF9	rs16831153	1:159920719	near-gene-3(GVS)	A/G	17,85%	18,90%
SLC26A4	rs2248465	7:107303628	intron-variant(dbSNP)	C/T	26,33%	29,17%
SLC26A4	rs2701684	7:107299527	near-gene-5(GVS)	A/G	32,36%	34,65%
SLC26A4	rs2701685	7:107299584	near-gene-5(GVS)	A/G	23,51%	25,99%
SLC26A4	rs2712228	7:107300340	near-gene-5(GVS)	C/A	30,15%	30,53%
TBXAS1	rs12532701	7:139521534	intron-variant(dbSNP)	G/A	43,07%	46,73%
TFF1	rs184432	21:43787562	near-gene-5(GVS)	A/G	31,83%	28,29%
TFF1	rs225359	21:43787436	near-gene-5(GVS)	A/G	33,28%	30,96%
TFF2	rs225340	21:43772947	near-gene-5(GVS)	T/C	44,77%	41,86%
TFF2	rs3814896	21:43771711	near-gene-5(GVS)	G/A	33,92%	30,84%
TFF2	rs225333	21:43764496	near-gene-3(GVS)	A/G	28,14%	27,21%
TIMP3	rs137487	22:33259104	intron-variant(dbSNP)	A/G	47,28%	45,86%
TK1	rs1065769	17:76170735	utr-variant-3-prime(dbSNP)	T/C	32,44%	31,17%
TSLP	rs3806932	5:110405675	near-gene-5(GVS)	G/A	38,89%	40,58%
UBE2T	rs14451	1:202304868	synonymous-codon(dbSNP)	C/T	44,31%	47,56%
ULBP1	rs1853665	6:150298842	intergenic(GVS)	T/C	19,58%	17,07%
ULBP1	rs4425606	6:150284435	near-gene-5(GVS)	G/A	18,39%	17,83%

4.1.4 Indukált köpet vizsgálat, RNA izoláció és génexpressziós mérés

A résztvevők 400 μg salbutamol belégzése után 5 percen keresztül a De Vilbiss Nebulizer (Ultra-NebTm 2000 model 200HI) rendszer által porlasztott, hipertóniás, (4,5%-os) nátrium klorid oldatot lélegeztek be. Az alanyok háromszor kísérelték meg a köpet felköhögését (alsó légúti szekrétumot), minden alkalommal ellenőrizték a légzésfunkciókat. A köpetet steril tartályban gyűjtötték, a nagyobb plakkok kiválogatása után azokat 0,1% dithiothreitol-tartalmú PBS oldattal hígították (Sigma, St Louis, MO, USA) és 30 percen keresztül rázták. Ezt követően a mintákat 40 μm sűrűségű nejlonhálón (BD Biosciences, Falcon cell strainer) szűrték át és 1500 rpm fordulatszámon centrifugálták 10 percen keresztül. A sejteket PBS oldatban vették fel és 0,4%-os tripánkék festékkel (Sigma) festették, majd Bürker kamrában számolták. A sejtes összetételt tárgylemezen, fénymikroszkóp alatt, Quick Panoptic (Cypress, Langdorp, Belgium) hematológiai festék segítségével állapították meg minimum 300 sejt azonosítása után. A sejteket lízispuffer segítségével emésztették, majd -80°C-on tárolták.

Az RNA izolációt Qiagen Mini Rneasy Kittel végeztük, (Qiagen, Maryland, USA), a cDNA átírás High Capacity cDNA Reverse Transcription Kittel (Applied Biosystems, Foster City, CA) történt. A kiválasztott génekre a real-time quantitative PCR méréseket az ABI 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster City, CA) rendszerrel végeztük. A minták mRNS szintjének normalizálása a β-actin expresszió szintjéhez normalizálva történt, ami stabil referenciának bizonyult.

4.1.5 Frekventista statiszikai elemzés

A Hardy-Weinberg egyensúlyt (HWE) khí négyzet próbával ellenőriztem az online DeFinetti alkalmazás segítségével (Helmholtz Zentrum München, Institut für Humangenetik, http://ihg.gsf.de/cgi-bin/hw/hwa1.pl). A szignifikancia szintet 0,01-re állítottuk be. Az előbbi szoftvert használtam a Pearson-féle khí négyzet próbák elvégzéséhez, hogy az asztmához való asszociációt és az esélyhányadosokat (odd ratio, OR) meghatározzam, mind allélikus, mind domináns illetve recesszív modellben. A konfidencia intervallumokat 95%-os szinten határoztam meg. Minden polimorfizmus hiányzó genotípus adatait az egyváltozós eloszlásukból véletlenszerűen mintavételezve

pótoltam. A pótlást több példányban is elvégeztem és később véletlenszerűen ellenőriztem, hogy a pótlás befolyásolta-e a kapott eredményeket, de kerekítési hibánál nagyobb eltérést sehol nem találtam. A Haploview alkalmazást használtam az allélikus asszociációk ellenőrzéséhez és a haplotípus szintű asszociációk számításához [163]. A többszörös hipotézistesztelés hatását HaploView-ban permutációs teszttel korrigáltam mind az egyváltozós, mind a többváltozós esetben, minimum 1000 permutációval. Az allélikus asszociációkat és a domináns-recesszív öröklődési modellek asszociációját IBM SPSS Statistics V20 segítségével (Pearson-féle khí négyzet próbával) is ellenőriztem; ebben az esetben a többszörös hipotézistesztelés korrigálását Bonferroni korrekcióval végeztem.

A többváltozós logisztikus regressziós elemzéseket IBM SPSS Statistics V20 szoftverrel végeztem, a kor és a nem minden esetben a modellbe került, hogy elégséges statisztikai adjusztálást végezzünk az esetek és kontrollok közötti eltérés miatt. A real time quantitative PCR mérések normalizálásához a delta-delta-CT algoritmust használtam (a houskeeping génhez és a kontroll mintákhoz normalizálva). Azon mintákat kizártuk, amelyek RNA Integrity Number értéke túl alacsony volt, vagy a PCR reakció CT kiugró értéke volt és így megbízhatatlan minőségre utalt. A normalizált transzkriptom változások statisztikai értékelését Student-féle t-próbával végeztem.

4.1.6 Bayes-háló alapú Bayesi többszintű relevancia elemzés

A Bayes-hálók (BN) olyan irányított körmentes gráfok (directed acyclic graph, DAG), melyek X1, X2, …, Xn valószínűségi változók együttes eloszlását reprezentálják. A valószínűségi változók a terület megfigyelt vagy mért faktorait reprezentálják, így pl. a klinikai paramétereket vagy genetikai polimorfizmusokat. A gráf egy csomópontja egy változót reprezentál, egy él pedig a köztük fennálló direkt függőséget.

A Bayes-hálók tanítása, tehát a változók közötti függőségi viszonyok kinyerése az adatból valójában annak az irányított körmentes gráfnak a keresését jelenti, amely a legjobban reprezentálja az adathalmazt. A legtöbb esetben az adat mennyisége a változók számához viszonyítva nem elegendő ahhoz, hogy egyértelműen meghatározható legyen egyetlen ilyen gráf, így a gyakorlatban sok valószínű gráfot találhatunk. Azonban, gyakran vannak olyan

strukturális jegyek, pl. két csomópont közötti él, melyek nagy valószínűséggel rekonstruálhatóak az adathalmazból.

Tehát Bayesi tanulással meghatározhatjuk annak erősségét, hogy az adat mennyire utal egy adott jegy meglétére, ha kiszámítjuk az adott jegy posteriori valószínűségét (1. egyenlet).

$$P(f|D) = \sum_{G} P(G|D) f(G)$$

1. egyenlet

A *G* egy Bayes-háló struktúrát jelöl, a *D* az adathalmazt, *f(G)* értéke pedig 1 amennyiben *f* jegy megtalálható *G*-ben, 0 amennyiben nem. Tehát pl. *p(MB = s | D)* azt a feltételes valószínűséget jelöli, hogy *s* algráf (vagy jegy) egyenlő az MB-vel (Markov-takaróval) a *D* adathalmaz tükrében [164].

Az 1. egyenletben a szumma első tagjának kiszámításához a (két kifejezés közötti arányosságot leíró) Bayes tételt használhatjuk fel (2. egyenlet).

$$P(G|D) \propto P(D|G)P(G)$$

2. egyenlet

A *P(D|G)* kifejezés a *D* adat előállításának valószínűsége, amennyiben *G* gráfnak megfelelő összefüggési rendszer érvényes a változók között; a *P(G)* pedig a *G* gráf struktúra priori, tehát előzetes valószínűsége. (Tehát a 2. egyenlet lényegében azt mondja ki, hogy a *G* összefüggési rendszer valószínűsége amennyiben a *D* adatokat mértük arányos a *D* adat előállításának valószínűségével amennyiben a *G* összefüggési rendszer áll fenn, szorozva a *G* összefüggési rendszer előzetes valószínűségével.) A tesztjeink során minden gráf struktúrát azonos valószínűségűnek tekintettünk (uniform prior). Amennyiben *D* adathalmaz teljes (nem tartalmaz hiányzásokat), a változók lehetséges szülő halmazai többváltozós Dirichlet-eloszlást követnek és néhány más kikötés teljesülése esetén a *P(D|G)* számítására létezik hatékony algoritmus [165]. A számítás elvégzésére több módszertan is ismert, mi ehhez a csoportunk korábbi eredményei alapján a Cooper-Herskovits priort vettük alapul [166].

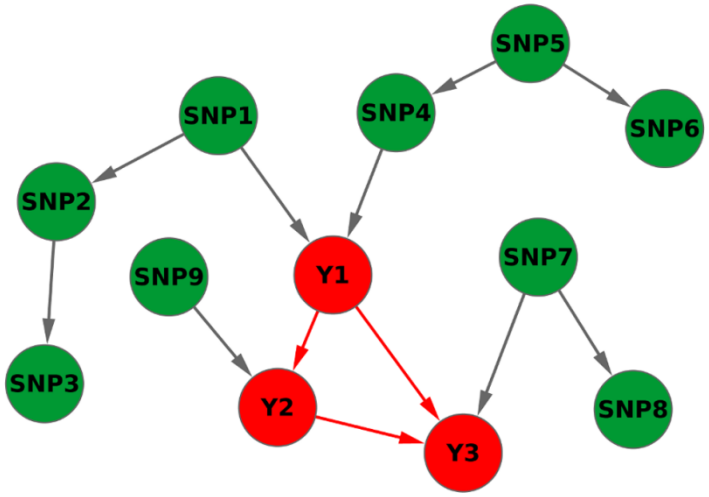
A cél tehát egyes gráf jegyek posteriori valószínűségének kiszámítása (1. egyenlet). Mivel a tárgyterület összefüggéseit leíró összes lehetséges Bayes-háló struktúrák számossága szuperexponenciális a változók számában, így ezek felsorolása számítástechnikailag nem kivitelezhető és a szumma nem számítható egzakt módon. A számítást Metropolis Coupled Markov Chain Monte Carlo (MC3) közelítő módszerekkel végeztük: párhuzamos markov láncokat definiáltunk az irányított körmentes gráfok tere fölött (melyek stacionárius eloszlása a *P(G|D)* posterior), majd a láncok random bejárásával mintavételeztünk és ezzel közelítettük a szumma értékét (a konvergencia és konfidencia diagnosztikát Gelman-Rubin R-score [167] alapján végeztük). Egy tetszőleges struktúrából kiindulva a Markov láncban minden egyes lépés a körmentes gráfban egy lokális gráf transzformációnak felelt meg: él hozzáadása, él megfordítása, él törlése, mindhárom azonos valószínűséggel (súlyozva az eredő gráf struktúra valószínűségével az adat tükrében) [168]. Ezután MCMC mintavételezőt futtattunk 10M burn-in periódussal (vagyis az első 10M lépést vizsgálat nélkül eldobjuk), a számítási komplexitás csökkentése érdekében korlátoztuk a gráfban a szülők számát 2-re (az eredmények átlagolásával, „szétkenődésével" ennek gyakorlati jelentősége nincsen, átlagos értelemben nem korlátozza a szülőszámot), majd 10 ilyen futás eredményét átlagolva vizsgáltuk néhány kiemelt jegy posteriori valószínűségét.

5. táblázat - A BN-BMLA elemzés során vizsgált relációk típusai [133].

Reláció	**Rövidítés**	**Leírás**
Páros relációk		
- Direkt oksági relvenacia	DCR(X,Y)	X és Y között van él
- Tranzitív oksági relevancia	TCR(X,Y)	X és Y között létezik irányított út
- Zavaró relevancia	ConfR(X,Y)	X és Y közös őssel rendelkezik
- Asszociáció	A	DCR vagy TCR vagy ConfR
- Tiszta interakciós relevancia	PIR(X,Y)	X és Y közös gyermekkel rendelkezik
- Erős relevancia	SR(X,Y)	PIR vagy DCR

Halmaz reláció		
- Erős relevancia	MBS(Y)	Y Markov-takarója
Több célváltozós reláció		
- Direkt relevancia	EdgeToAny(X,y)	X és bármely y között él van
- Erős relevancia	SR(X,y)	X és bármely y között él van, vagy X és bármely y közös gyermekkel rendelkezik

Az MBM, MBS és MBG jegyek posterior valószínűségének elemzése lehetővé teszi az asszociációk pontosabb karakterizálását, mint pl. közvetlen vagy áttételes függőség, esetleg egyéb statisztikai zavaró hatás. Az 5. táblázatban gyűjtöttük össze a vizsgált jegyek típusait, a 10. ábra szemlélteti a jegyeket egy gráfon [133].



10. ábra - Relevanciák típusai. Páros relevanciák: direkt oksági relevancia, vagyis szülő-gyermek kapcsolat az irányított gráfban (pl. SNP1 és Y1 között van irányított él); tranzitív oksági relevancia (pl. SNP5 és Y3 között van irányított út); zavaró relevancia (pl. Y2-nek és SNP3-nak SNP1 közös őse), asszociáció (bármelyik az előző három); tiszta interakciós relevancia (pl. Y1-nek és SNP7-nek közös gyermeke van); erős relevancia (pl. Y1 és SNP1 között mert van irányított él köztük, Y1 és SNP7 között mert tiszta interakció áll fenn köztük). Halmaz relevanciák: erős relevancia (pl. Y2 Markov-takaróba tartozó változók Y1, SNP9, Y3, SNP7). Több célváltozós relevanciák: erős relevancia több célváltozóra nézve (pl. Y1, Y2, Y3 változó halmaz Markov-takarója SNP1, SNP4, SNP7, SNP9). Piros változók: potenciális célváltozók. Zöld változók: genetikai polimorfizmusok [133].

DOI:10.14753/SE.2015.1757

Továbbá, az elemzéssel vizsgálhatóak a statisztikai interakciók és redundanciák is [169]. A Bayes-háló alapú megközelítésben két változó (pl. két polimorfizmus) interakciója azt jelenti, hogy ezek a vizsgált modellekben nagyobb valószínűséggel jelennek meg egyszerre az erősen releváns változók között, annál mintha függetlenek lennének egymástól (tehát a hatásukat együtt fejtik ki). Ezzel szemben a redundancia azt jelenti, hogy az erősen releváns változók halmazában ezek felcserélhetőek (tehát hatásuk azonos vagy helyettesíthető).

A nagy számítási kapacitást igénylő elemzéseket a GenaGrid Konzorcium 512 processzoros SGI Altix ICE szuperszámítógépén végeztük (mely 2009-ben Magyarország legerősebb szuperszámítógépe volt). A szuperszámítógépen futó BN-BMLA szoftver eszköztárat és az eredmények megjelenítéséhez felhasznált BayesEye kliens program részletesebb leírását kutatócsoportunk több esettanulmány során publikálta [133-138], illetve az eszköztár részletesebb elméleti hátterét a Probabilistic graphical models in genetics, genomics, postgenomics című könyv 13. fejezetében foglalta össze [170].

DOI:10.14753/SE.2015.1757

4.2 Feldúsulás elemzés

4.2.1 Szisztematikus hatóanyag újrapozicionálás adat fúziós módszerekkel

Kutatócsoportunk egy egyedülálló számítógépes hatóanyag elemzési és új indikáció keresési módszertant (QDF^2) fejlesztett ki, mely egyben a kutatás-fejlesztéseim alapjául szolgáló igényt és technikai problémát is felvetette, így az alábbiakban sematikusan bemutatom a módszer működését [150, 171].

A számítógépes hatóanyag újrapozícionálás során a legtöbb gyakorlati esetben (ipari, pénzügyi, szellemi tulajdon védelmi, vagy módszertani okokból) nem tetszőleges hatóanyag-indikáció párokat keresünk. A QDF^2 egyfajta szakértői rendszerként működik, a rendszer célja adott hatóanyaghoz új indikációkat (esetleg célpontot, anyagcsere útvonalat, betegséget, stb.) keresni, vagy adott indikációhoz (esetleg célponthoz, anyagcsere útvonalhoz, betegséghez, stb.) hatóanyagokat keresni - mindezt a nagy mennyiségű adattömegre és a farmakológiai szakértői tudásra támaszkodva.

Tehát a lekérdezésnek két, alapvetően eltérő célja lehet, és ennek megfelelően a bemenet is kétféle lehet. Amennyiben egy hatóanyaghoz (vagy hatóanyagokhoz) keresünk új indikációkat, abban az esetben a rendszer bemenete közvetlenül egy vizsgált hatóanyag vagy hatóanyagok. A másik lehetőség szerint egy indikációhoz keresünk hatóanyagot. Ekkor az első lépés olyan hatóanyagok keresése melyek ismertek és engedélyezettek az adott indikációban, a rendszer a hatóanyagok ezen halmazát fogja bemenetként megkapni. Hasonlóan lehet adott célponthoz, anyagcsere útvonalhoz, betegséghez stb. lekérdezéseket megfogalmazni, de a jó kérdés (halmaz) megfogalmazása komoly farmakológiai háttértudást igényel. Így a második esetben visszavezettük a problémát az elsőre, a bemenet mindkét esetben az egy vagy több hatóanyagot tartalmazó halmaz.

A rendszer az ismert vegyületekről elérhető adattömegre épít, amely lehet zajos vagy hiányos, az adatok széles skáláját képes feldolgozni a fizikai, kémiai leíróktól, tetszőleges farmakológiai tesztek eredményén át bármilyen kvantitatív klinikai jellemzőig. Az adatok tárolása kernel mátrixok formájában történik, amelyek valójában a rendszerben tárolt összes ismert vegyület páronkénti kvantitatív hasonlóságait jelentik az összes ismert

tulajdonság szerint. Ezek leggyakrabban egyszerűen mérhetőek vagy kiszámíthatóak, de mivel elegendő a páronkénti hasonlóságaik normatív ismerete, több esetben szükségtelenné válhat a vegyületek pontos tulajdonságainak mérése is. A QDF^2 egy komplex matematikai módszertanra építve (kernel alapú adatfúzió [172]) képes a bemenetként kapott vegyülethalmazhoz hasonlóság alapján sorba rendezni a referencia adatbázis vegyületeit, méghozzá a különböző hasonlósági dimenziókat a konkrét lekérdezés alapján optimálisan súlyozva.

Amennyiben a feladat adott indikációhoz potenciális hatóanyagok keresése, úgy a referencia adatbázis sorrendezése egyértelműen szolgáltat tesztelhető hipotéziseket, vagyis vegyületeket ehhez. Amennyiben adott hatóanyagokhoz tartozó indikáció (esetleg célpont, anyagcsere útvonal, stb.) keresése a feladat, úgy az eredmény további elemzésére és értelmezésére van szükség, de a sorrend pontosabb statisztikai elemzése és farmakológiai értelmezése már komoly háttértudást igényel. Az eredmény értelmezésére léteznek különböző módszerek (szűrések, hálózatelemzés, stb.), de ezek sok esetben nehezen összeegyeztethetőek a klasszikus farmakológiai háttértudással. Saját munkáim során erre a problémára kíséreltem meg egy új eszközt kidolgozni a molekuláris biológiában tanulmányozott, hasonló kihívásokat kezelő eszközök alapján. A vegyület sorrendek előállításához azonban a QDF^2-nél egy lényegesen egyszerűbb algoritmust használtam, hogy a módszertani fejlesztést egy könnyebben tesztelhető környezetben végezhessem el.

4.2.2 Gene Set Enrichment Analysis

A géncsoport feldúsulás elemzés (Gene Set Enrichment Analysis, GSEA) génexpressziós mintázatok elemzésében és értelmezésében már jól bevált technika. A módszertant alapvetően nagyszámú entitásból álló sorrendezett listák absztrakt fogalmi szinteken történő elemzésére és értelmezésére fejlesztették ki, így a módszertan a megfelelő módosításokkal alkalmas lehet vegyület-adatbázisok elemzésére is.

A kétezres évek elején a microarray technikák megjelenésével a nagy áteresztőképességű génexpressziós elemzés relatív olcsóvá vált, de a nagymennyiségű transzkriptumok differenciális elemzése (eset-kontroll jellegű összehasonlítása és értelmezése) komoly

kihívás maradt. A transzkripciós eredményeket első lépésben vissza lehet vetíteni a hozzájuk tartozó génekre, mint az értelmezés egy közös szintjére. Az esetek és kontrollok között mérhető transzkripciós változások vizsgálatával tehát felállítható egy gén sorrend a változás statisztikai szignifikanciája vagy mértéke alapján. Gyakran előfordul, hogy statisztikailag egyetlen génben sem szignifikáns a változás a többszörös hipotézistesztelésre történő korrekció után, ráadásul nehéz felismerni, hogy a legerősebben érintett gének halmazai milyen közös biológiai funkcióhoz vagy folyamathoz kötődnek. Egy relatíve alacsony mértékű expressziós változás egy anyagcsere-útvonal minden érintett génje esetén gyakran sokkal jelentősebb hatást válthat ki, mint egyetlen gén transzkripciójának nagyságrendi megváltozása. Így az értelmezés sokkal hatékonyabb lehet a géncsoportok szintjén (amellyel pl. olyan kijelentéseket tehetnénk, mint „gyulladásos folyamatokban érintett gének előbbre sorolódtak, tehát az érintett betegeken gyulladás jelei mérhetőek"). A terület sztenderdje, - a Gene Ontology - mellett számos gén annotációs adatbázis és ontológia létezik, melyek segítségével könnyen definiálhatóak géncsoportok pl. az érintett biológiai szabályzási útvonalak, biológiai funkciók, sejtalkotók szerint.

A feldúsulás elemzésnek számos fajtája és implementációja létezik, de a következő lépések szinte mindegyikben megjelennek valamilyen formában. Adott az összes mért entitásnak egy *L* sorrendje (pl. a géneknek egy sorrendezése az expressziós változásaik erőssége szerint) és az entitások egy *S* részhalmaza (pl. egy sejtfunkcióhoz kötődő gének). Az algoritmus célja azt meghatározni, hogy az *S* halmaz elemeinek eloszlása *L* rendezett lista mentén egyenletes-e, vagyis az entitások megjelenése véletlenszerű-e. Az algoritmus lépései:

1. Feldúsulási érték („enrichement score", *ES*) számítása az *S* halmazra: sorra vesszük *L* lista minden elemét és minden olyan elemre, amely eleme *S* halmaznak is megnöveljük egy olyan egyezményes értékkel az *ES* értékét, amely arányos az adott elem pozíciójával (vagy a sorrendezés alapjául szolgáló értékkel, amennyiben az rendelkezésre áll). Bizonyos esetekben, így pl. génexpresszió elemzésnél is, a lista mindkét végét hasonló mértékű súlyozással vesszük figyelembe, mert a

transzkriptumok jelentős csökkenésének és növekedésének hasonló biológiai jelentősége lehet.

2. Statisztikai szignifikancia számítása *ES*-re: permutáció-teszteléssel meghatározzuk a null-eloszlást (vagyis véletlenszerűen létrehozott halmazok feldúsulási értékeinek eloszlását vizsgáljuk), amelyből meghatározható a vizsgált *ES* érték statisztikai erőssége, vagyis a nominális p-értéke.

3. Amennyiben több *S* halmazt vizsgálunk normalizálást, és statisztikai korrekciót kell végeznünk: az *S* halmazok esetleges eltérő elemszáma miatt normalizálni kell az *ES* értéket az adott halmaz elemszámával az összehasonlíthatóság érdekében, illetve a vizsgált halmazok számának függvényében a többszörös hipotézistesztelési korrekciót kell alkalmazni a szignifikancia küszöb meghatározásánál [154].

Az értekezésben a feldúsulás elemzésnek a génsorrendek helyett vegyületsorrendek elemzésére módosított változatát fejlesztettem ki, és használtam fel (Compound Set Enrichment Analysis, CSEA). Felhasználását az ismert vegyületekről elérhető, egyre nagyobb mennyiségű publikusan elérhető adat indokolja, amelyek együttes értelmezése az emberi szakértők képességeit meghaladja. A módszer legegyszerűbb lehetséges felhasználási esetét a 11. ábra mutatja be.

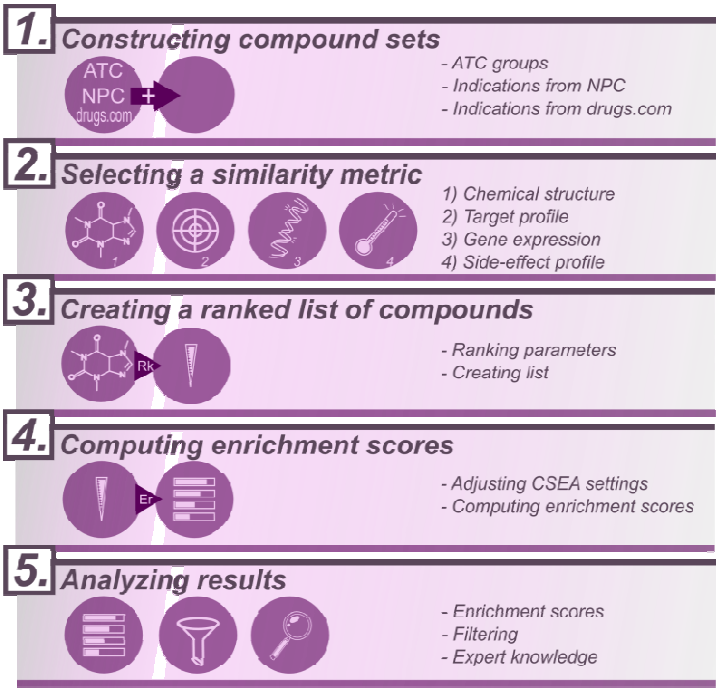
1. Ismert tulajdonságokkal rendelkező hatóanyagokból S halmazokat képezünk bizonyos közös tulajdonságok, pl. indikációk vagy mellékhatások alapján.

2. Egy vizsgált hatóanyag bizonyos tulajdonságának kiválasztása, mérése vagy számítása (pl. kémiai tulajdonság, célpont, expresszió, mellékhatás, stb.).

3. A kiválasztott tulajdonság alapján, a vizsgált hatóanyaghoz való hasonlóság szerint rendezzük a referencia adatbázis vegyületeit, mellyel egy L sorrendet kapunk.

4. Az előállított L sorrendre kiszámítjuk az előre definiált S halmazok ES feldúsulási értékeit, valamint normalizálást és statisztikai korrekciót végzünk.

5. Az eredmények értelmezése, pl. lehetséges új indikációk azonosítása.



11. ábra – A Compound Set Enrichment Analysis (CSEA), vagyis a feldúsulás elemzés egyszerű, egydimenziós kemoinformatikai felhasználása 5 lépésben Temesi és munkatársai angol nyelvű tanulmányából [173]: 1, vegyület halmazok összeállítása (pl. indikációk alapján) 2, hasonlósági metrika kiválasztása (pl. kémiai profil) 3, vegyületek sorba rendezése 4, feldúsulási értékek számítása a halmazokra 5, az eredmények farmakológiai kiértékelése

Ez egy újszerű alkalmazása a bevált feldúsulás elemzési technikának ezen a területen, amellyel például a következő kijelentést tehetnénk: „ha az összes engedélyezett gyógyszer hatóanyagot a vizsgált vegyülethez való hasonlóság szerint sorrendezzük, akkor az ismert dopaminerg agonisták többnyire előre rangsorolódnak, tehát valószínűsíthető a hatóanyag dopaminerg hatása”.

4.2.3 Referencia adatbázis előkészítése

Az elemzésekhez ismert hatóanyagokból egy referencia adatbázist építettünk, melyhez nyilvánosan elérhető vegyület adatbázisokat használtunk és dolgoztunk fel, a feldolgozás bizonyos lépéseit korábban Arany és Bolgár közleményeiben tárgyaltuk [150, 171].

A vegyületek kémiai profiljához három leírót használtunk, mindhárom vektort a Schrödinger 2012 Suite szoftver csomag segítségével állítottuk elő. A Molconn-Z leírók eredetileg egy kereskedelmi szoftver csomagban váltak elérhetővé, mely különböző

sztenderd, elsősorban topológiai leíró paraméterek kiszámítását tartalmazza (pl. a vázszerkezet atomjainak távolságai) és így több száz leíró állítható elő (http://www.edusoft-lc.com/molconn/). A MACCS kulcsok eredetileg szintén egy kereskedelmi fejlesztésként terjedtek el (MDL Information Systems Inc.), több száz fontos kémiai jegy meglétét vagy hiányát írják le egy bináris vektorral; míg a 3D Pharmacophore leírók hasonló módon 3D szerkezeti jegyeket jelölnek.

A vegyületek célpont profiljait a DrugBank [121] adatbázis segítségével állítottuk elő: a ma ismert összes hatóanyag összes (százas nagyságrendű) ismert célfehérjéjéből bináris vektor képezhető, majd ezt vegyületenként kitöltöttük a megfelelő értékekkel.

A mellékhatás profilok egységesítéséhez kigyűjtöttünk 760 preferált kifejezésből (Preferred Term) álló szótárat a nemzetközi gyógyszeripari sztenderdnek számító MedDRA orvosi ontológiából (http://www.meddra.org/) és összegyűjtöttük ezek lehetséges szinonimáit az NIH egészségügyi terminológia adatbázisából, az UMLS-ből (http://www.nlm.nih.gov/research/umls/). A DailyMedben (http://dailymed.nlm.nih.gov/dailymed/) megtalálhatóak az FDA által engedélyezett gyógyszerek betegtájékoztatói; a kombinált készítmények kihagyásával 6460 darab, a generikum redunandciák kiszűrésével 1729 hatóanyag betegtájékoztatója érhető el. A betegtájékoztatók alapján két mellékhatás vektor reprezentációt állítottunk elő. Egyrészt, a betegtájékoztatók nem kívánt hatások részéből („adverse effect”) kiemeltük a placebo kontrollált klinikai vizsgálatok táblázatos eredményeit ahol azok elérhető voltak (1986 esetben). Minden deszkriptorhoz kiszámítottuk az 1-p értékeket a Pearson-féle khí négyzet próba alapján, ahol a nullhipotézis szerint a mellékhatás független attól, hogy placebót vagy valódi hatóanyagot kaptak a betegek. Ilyen reprezentációt végül 153 vegyülethez tudunk létrehozni (mivel ez túl kis szám volt, a bemutatott módszertani tesztjeim során végül nem alkalmaztam). Másrészt, a betegtájékoztatókat szabadszöveges elemzésnek is alávetettük: a hatóanyagokhoz kigyűjtöttük a betegtájékoztatóban felbukkanó mellékhatásokat, és a mellékhatások erősségét az adott betegtájékoztatóban megfigyelhető számosságuk alapján súlyoztuk a klasszikus TF-IDF (term frequency-inverse document frequency) algoritmussal

[174]. Ezzel a módszerrel 1359 hatóanyaghoz tudunk TF-IDF mellékhatás vektorokat készíteni.

Továbbá, felhasználtuk a SIDER mellékhatás adatbázist, melyeket a saját módszerünkhöz hasonló eszközökkel gyűjtöttek össze [175]. Így összességében 652 vegyülethez sikerült pontos, vagy hozzávetőleges (ritka, nem gyakori, gyakori) mellékhatás prevalencia karakterisztikát felállítani.

A fenti adatforrások kombinálása és redundanciák kiszűrése után 1941 FDA által engedélyezett hatóanyag maradt a referencia adatbázisunkban. A CMAP expressziós adatbázis entitásait utólag vetettük össze a referencia adatbázissal. Azon további tesztekben, ahol a CMAP adatbázist is felhasználtuk, ott kiszűrtük és csak azokat a hatóanyagokat vizsgáltuk, amelyek mindkettőben szerepeltek (1730 hatóanyag maradt). Kérdés, hogy ez mennyire reprezentatív az összes FDA által emberi felhasználásra engedélyezett kismolekulás hatóanyagra nézve. Huang és munkatársainak alapos katalogizálása alapján ennek számossága 2356, így a referencia könyvtárunk tartalmazza az engedélyezett molekulák közel háromnegyedét [115].

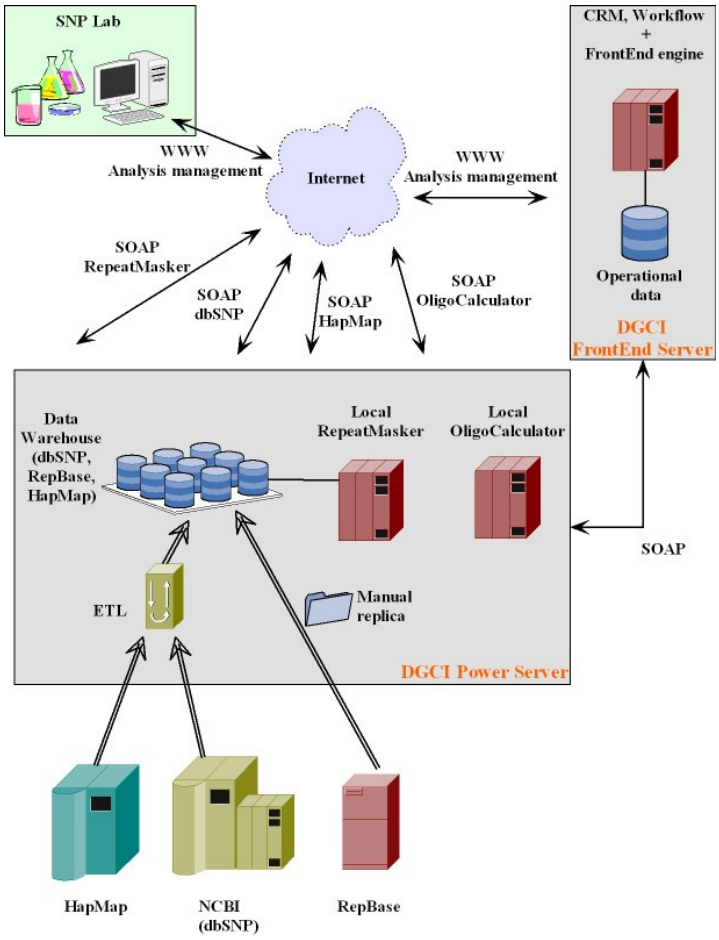
DOI:10.14753/SE.2015.1757

5 Eredmények

5.1 Asztma genetika

5.1.1 Kísérlettervező rendszer (TIGER)

Az asztmával kapcsolatos első tanulmányunk során, a munka kezdetén létrehoztam egy szoftveres genetikai asszociációs kísérlettervező rendszert, melyet TIGER-nek neveztem el. A rendszer egy primer extenziós elven működő, nagy áteresztőképességű genotipizálási mérés optimális polimorfizmus készletének kiválasztását támogatja (12. ábra).



12. ábra - A TIGER kísérlettervező rendszer architektúrája az eredeti dokumentációból [159]. A DGCI Power Server az elemzéshez szükséges adatbázisokból másolatokat tart fenn és itt foglalnak helyet a nagyobb számítási kapacitást igénylő elemzési eszközök is. A DGCI FrontEnd Server a webes felhasználó felületet szolgáltatja és SOAP protokollon keresztül kommunikál a DGCI Power Serverrel. A DGCI FronEnd Servert a labor felhasználói interneten keresztül a saját számítógépeik web böngészőjéből érhetik el.

A TIGER rendszer központi szervere (Power Server) egy lokális adatbázis kópiát hoz létre a HapMap (http://hapmap.ncbi.nlm.nih.gov/), a dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) és a Repbase (http://www.girinst.org/repbase/) adatbázisokról, melyeket rendszeresen frissít is. A szerver ezen referencia adatbázisok segítségével három funkcióra nyújt elemi webszolgáltatásokat:

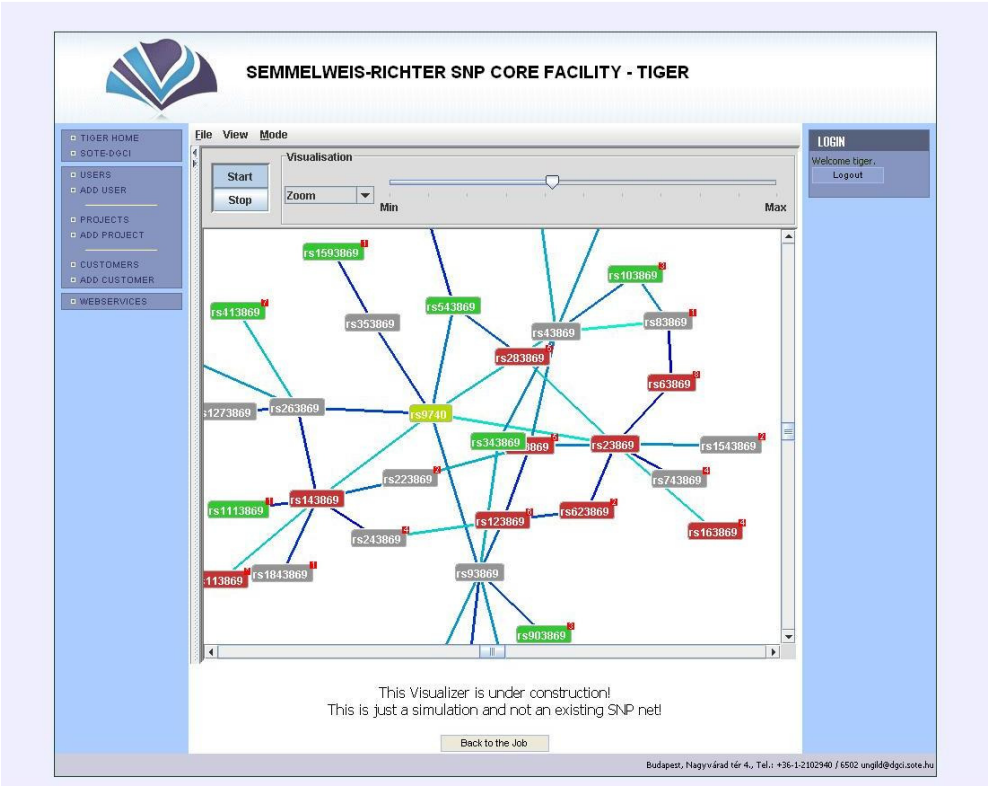
1. Különböző szűrési feltételeknek megfelelő polimorfizmusokat ajánl (pl. egy gén vagy genom régió, funkció, gyakoriság, más ismert polimorfizmusokkal való kapcsoltság, és más tényezők alapján)

2. Egy vizsgált polimorfizmus környezetében lévő szekvenciában a primer kötődését nehezítő szakaszokat azonosít (pl. különböző ismétlődő szakaszokat, a genomban máshol is előforduló szakaszokat vagy polimorfizmusokat).

3. A polimorfizmus környező szekvenciája alapján olvadáspontot jósol a PCR reakcióhoz.

A felhasználói felületet kiszolgáló szerver (FrontEnd Server) egy munkafolyamat kezelő rendszert és a webes felhasználói felületért felelős szoftver komponenst tartalmazza. A munkafolyamat kezelő rendszer a Power Server szolgáltatásait veszi igénybe, és egy tetszőleges web böngészővel megjelenített felületen teszi elérhetővé. A felhasználók egy új projekt definiálása után bizonyos szűrési feltételek (pl. genom pozíció, gén, génen belüli funkcionális régió, polimorfizmus típus, MAF) és a tervezett genotipizálási mérés paramétereinek (pl. hány polimorfizmus szimultán mérése lehetséges stb.) megadása után elindíthatnak egy elemzési feladatot. Az eredményről e-mailben kapnak értesítést általában pár óra elteltével. Ekkor a webes felületen böngészhetővé válnak a szűrési feltételeknek megfelelő polimorfizmusok, a polimorfizmusok méréstechnikai alkalmassági értékei, illetve egy interaktív gráfos felületen a polimorfizmusok kapcsoltsági hálózata, ahol vizuálisan láthatóvá válnak az esetleges redundanciák, vagy a nem mérhető polimorfizmusok helyettesítési lehetőségei (13. ábra).

DOI:10.14753/SE.2015.1757



13. ábra - A vizsgált polimorfizmusok kapcsoltsági hálózata a TIGER kísérlettervező rendszer eredeti dokumentációjából [159]. A polimorfizmusok kapcsolatsága egy gráf formájában jelenik meg, ahol a csomópontok a polimorfizmusokat, az élek a kapcsolatságot jelentik. Az élek színe és hossza a kapcsoltság erősségét jelöli. A csomópontok színe a méréstechnikai szempontból mérhető, nem mérhető vagy nem vizsgált státuszt jelzi. A gráf kiterítése animált és dinamikus, így jól láthatóak az összefüggő csoportok, pl. egy gén polimorfizmusai.

A szakértő az ajánlott polimorfizmusok jóváhagyásával vagy elutasításával bővítheti a mérésbe bevont polimorfizmusok halmazát, amíg az el nem éri a kívánt méretet (illetve indíthat új elemzési feladatokat, amennyiben a korábbi elemzés eredményei között nem volt elegendő elfogadható polimorfizmus). A rendszer kísérleti jelleggel támogatja az optimális halmaz kiválasztását is a különböző szempontok szakértő általi előzetes súlyozása alapján is (pl. „keresünk egy olyan 48 polimorfizmusból álló optimális halmazt, amelynek elemei az X vagy Y vagy Z gén közelében helyezkednek el, de mindhárom gén közelében legalább egy polimorfizmussal, mind nagy biztonsággal mérhetőek, lehetőleg kódoló régióban fekszenek és A/G típusú nukeotid cserével járnak, valamint legyenek egymással minél kevésbé kapcsoltak”). A TIGER rendszer működésének pontos specifikációja, leírása a korábbi munkáimban elérhető [159, 160].

A TIGER rendszer első valós felhasználásának eredményeit Ungvári [133] cikkében közöltük, ennek során a 11q12.2-q13.1 és 14q22.1-q22.3 asztmával asszociáló genomrégiókban parciális genomszűréshez választottunk ki polimorfizmusokat. A minor allél frekvencia alsó határa 5% volt (ilyen gyakoriságminimummal várható annyi ritka homozigóta és heterozigóta egyén, amennyivel a statisztikai elemzést még el lehet végezni); a méréstechnikai szempontból kizárt polimorfizmusok helyett kapcsolt ($r^2 \geq 0{,}08$, D' = 1) polimorfizmusokat választottunk, majd a technikai szűrésen átesett polimorfizmusokat funkcionális hatásuk alapján rangsoroltuk (nem-szinonim, promoter vagy 3' UTR régió, szinonim és intronikus). A fent leírt módon a 11q12.2-q13.1 régióban 68, a 14q22.1-q22.3 régióban 77, együttesen 145 SNP került kiválasztásra és genotipizálásra, a tanulmány eredményei társszerzős munkámban érhetőek [133]. A TIGER kísérlettervező rendszert hasonló beállításokkal később más kutatásokban is felhasználtuk, többek között a dolgozatom alapját képező, következő asztma tanulmányomban is [157].

5.1.2 Genotipizálási eredmények

A második asztma tanulmányomban kiválasztott 90 polimorfizmus tulajdonságait (gén, rs id, funkció, allél) a 4. táblázat részletezi; az eset és kontroll csoportok összes genotípus frekvenciája, Hardy-Weinberg egyensúly tesztjei és a populációs asszociációk statisztikái a terjedelmi korlátok miatt az eredeti közleményben érhetőek el [157]. A hiányzó genotípus mérések aránya 7,1% volt. Azon betegeket kizártam az elemzésből, ahol a hiányzó genotípus értékek a 80%-ot meghaladták (n=9, ~1%); valamint kizártam azon polimorfizmusokat is, ahol a hiányzás aránya a 20%-ot meghaladta (rs2432561, rs3814896, rs1821142, rs17027410). A maradék 86 polimorfizmusból 2 esetén mutatkozott szignifikáns eltérés a kontroll csoport Hardy-Weinberg egyensúlyában, ez egyértelműen szisztematikus hibára utal, így ezeket is kizártam (rs11016071, rs2857600).

Az asztmás betegek és kontrollok közötti genotípus eloszlások két gén, összesen négy polimorfizmusa esetén mutattak szignifikáns eltérést minden statisztikai eszközzel (DeFinetti HWE alkalmazás, Haploview, IBM SPSS): *SCIN,* azaz a Scinderin fehérjét kódoló gén (rs2240572, rs2240571, rs3735222) és *PPARGC1B,* azaz a Peroxisome Proliferator-Activated Receptor Gamma Coactivator 1-Beta fehérjét kódoló gén (rs32588),

ezek eredményei a 6. táblázatban láthatóak. Hat másik gén (*ITLN1* /Intelectin-1/, *FABP3* /Fatty Acid Binding Protein 3/, *MAT1A* /Methionine Adenosyltransferase 1 Alpha/, *OSGIN* /Oxidative Stress Induced Growth Inhibitor/, *LY9* /Lymphocyte antigen 9/, *LGMN* /Legumain/) polimorfizmusai mutattak határeset közeli szignifikanciát (több teszttel szignifikáns p-értéket kaptunk, de a többszörös hipotézistesztelési korrekció miatti permutáció tesztek már 0,05 és 0,5 közé eső p-értékeket adtak). Ennek ellenére a szignifikancia relatív erőssége, a korábbi egér kísérletek eredményei és az asztma patomechanizmusában betöltött potenciális szerepe miatt ezek közül egy gént a további vizsgálatokba is bevontam: *ITLN1* (rs4656958).

6. táblázat - Az asztmával szignifikánsan asszociált gének és polimorfizmusok listája a DeFinetti HWE alkalmazás eredményei alapján, illetve a HaploView alkalmazás által számított permutált p-értékek [157]. A félkövérrel kiemelt p-értékek szignifikánsnak tekinthetőek (p-érték < 0,05). A táblázat fejlécei a DeFinetti HWE alkalmazás (Helmholtz Zentrum München, Institut für Humangenetik) jelöléseit követik: [1] major allél, [2] minor allél, [11] homozigóta major allél genotípus, [22] homozigóta minor allél genotípus, [12] heterozigóta genotípus; [11+12] heterozigóta genotípus és homozigóta major allél összevonva [12+22] heterozigóta genotípus és homozigóta minor allél összevonva; a szignifikanciát khí-négyzet próbával állapítottuk meg (<->).

Gén	SNP	Kontroll mintaszám [11], [12], [22]	Eset mintaszám [11], [12], [22]	[1]<->[2=minor] (Allélikus)	Permutált p-érték	[22]<->[11+12] (Recesszív minor)	[11]<->[12+22] (Domináns minor)
SCIN	**rs2240572**	94, 164, 85	115, 149, 40	Odds ratio=0,637 C.I.=[0,510-0,795] **p=0,00007 (P)**	**p=0,0046**	Odds ratio=2,174 C.I.=[1,439-3,287] **p=0,00019**	Odds ratio=0,620 C.I.=[0,445-0,864] **p=0,00466**
SCIN	**rs3735222**	94, 165, 85	115, 149, 41	Odds ratio=0,642 C.I.=[0,515-0,802] **p=0,00009 (P)**	**p=0,0056**	Odds ratio=2,113 C.I.=[1,402-3,185] **p=0,00029**	Odds ratio=0,621 C.I.=[0,446-0,865] **p=0,00474**
PPARGC1B	**rs32588**	196, 130, 15	202, 65, 8	Odds	**p=0,0094**	Odds	Odds

				ratio=0,563 C.I.=[0,420-0,757] **p=0,00012 (P)**		ratio=1,536 C.I.=[0,641-3,677] p=0,33229	ratio=0,488 C.I.=[0,347-0,688] **p=0,00004**
SCIN	**rs2240571**	126, 161, 56	75, 155, 75	Odds ratio=1,513 C.I.=[1,214-1,886] **p=0,00023 (P)**	**p=0,0164**	Odds ratio=0,598 C.I.=[0,406-0,881] **p=0,00894**	Odds ratio=1,781 C.I.=[1,266-2,504] **p=0,00085**
SCIN	**rs3173628**	120, 164, 60	74, 153, 73	Odds ratio=1,413 C.I.=[1,133-1,762] **p=0,00209 (P)**	p=0,1403	Odds ratio=0,657 C.I.=[0,448-0,964] **p=0,03116**	Odds ratio=1,636 C.I.=[1,160-2,307] **p=0,00482**
ITLN1	**rs4656958**	157, 150, 37	167, 118, 20	Odds ratio=0,724 C.I.=[0,569-0,922] **p=0,00863 (P)**	p=0,4679	Odds ratio=1,717 C.I.=[0,974-3,029] p=0,05930	Odds ratio=0,694 C.I.=[0,509-0,946] **p=0,02046**
ITLN1	**rs2274910**	153, 152, 39	160, 122, 22	Odds ratio=0,748 C.I.=[0,589-0,949] **p=0,01686 (P)**	p=0,7164	Odds ratio=1,639 C.I.=[0,948-2,833] p=0,07446	Odds ratio=0,721 C.I.=[0,529-0,983] **p=0,03816**

Megvizsgáltam, hogy a domináns („a minor allél bármilyen megjelenése kivált-e hatást”) és recesszív („kizárólag a minor allél hordozása kivált-e hatást”) modellek az allélikus

statisztikák (korrigálatlan) p-értékeihez képest mutatnak-e változást, de a különbségek a *SCIN, PPARGC1B* és *ITLN1* esetén jelentéktelen voltak, így ezeket a modelleket nem vizsgáltam tovább (6. táblázat). (Meg kell jegyezni, hogy az allélikus statisztikával gyengén szignifikáns gének közül az *LGMN* gén hatása domináns-recesszív modellben vizsgálva lényegesen erősebb lett, és a korrigált szignifikancia küszöb közelébe került.)

Ellenőrzésképpen a 6. táblázatban látható allélikus asszociációkat és a domináns-recesszív modelleket a sztenderdnek tekinthető IBM SPSS Statistics V20 szoftverrel is validáltam Pearson-féle khí négyzet próbával (a DeFinetti HWE alkalmazás által is alkalmazott eljárás), majd a többszörös hipotézistesztelési korrekciót Bonferroni korrekcióval is elvégeztem (a korrigált szignifikancia küszöb a 90 vizsgált polimorfizmusra 0,00055 volt). Ezen eredmények teljes összhangban voltak a korábbi számításokkal, ugyanazon polimorfizmusok asztmához való asszociációja bizonyult szignifikánsnak.

Külön statisztikai verifikációt végeztünk (IBM SPSS Statistics V20), hogy a csoportok átlagéletkor eltérésből adódó populációs rétegződés miatti hamis pozitív asszociációkat elkerüljük. Például, elméletileg lehetséges volna egyes népcsoportok vándorlása miatt az idősebb korcsoportokban egyes variánsok eloszlásának eltérése, ez pedig így a kontroll csoport korbeli eltérése miatt hamis asszociációval védő hatásként jelentkezhetne a statisztikai elemzés eredményében. A jelenség kizárására bináris logisztikus regressziót alkalmaztam korral és kor nélkül, és megvizsgáltam, hogy a polimorfizmusok p-értékei és esélyhányadosai hogyan változnak. Amennyiben ezek lényeges eltérést mutattak volna a két esetben, az a kor zavaró hatásának egyértelmű jele lenne. A p-értékek és esélyhányadosok minden esetben közel azonosak maradtak, a polimorfizmusok védő vagy hajlamosító hatása nem változott (a legnagyobb esélyhányados eltérések 20-30%-osak voltak), így populációs rétegződésnek nem találtam jelét.

Megvizsgáltam, hogy az egyváltozós esetekhez képest kimutatható-e haplotípus blokkok szintjén erőteljesebb kombinált hatás. Az elemzések egyértelműen megerősítették a korábbi egyváltozós eredményeket, bár a haplotípus szintű asszociációk p-értékei nem tértek el jelentősen az egyváltozós eredményektől: a *SCIN* gén haplotípusainak asszociációja

DOI:10.14753/SE.2015.1757

továbbra is erős volt; az *ITLN1* asszociáció pedig bár erősebb, de továbbra is határeseti volt és a többszörös hipotézistesztelési korrekcióval már nem maradt szignifikáns (7. táblázat). Az is jól látható, hogy mindkét vizsgált gén esetén a két leggyakoribb haplotípus az esetek és kontrollok legalább 85-98%-át lefedi, tehát a géneknek két gyakori variánsa van, a többi haplotípus számossága elhanyagolható.

7. táblázat - A haplotípus szintű statisztikák a 0,95-nél alacsonyabb permutált p-értéket mutató haplotípus blokkokról a HaploView szoftver alapján. A letörés igen éles, szignifikancia szerint az ötödik legerősebb haplotípus permutált p-értéke már 0,95 feletti, vagyis teljesen inszignifikáns. Így jól látható, hogy génenként egy védő és egy hajlamosító haplotípus figyelhető meg a populációban, a többi elméletileg lehetséges haplotípus számossága és szerepe elhanyagolható [157].

Gén	SNP	Haplotípus	p-érték	permutált p-érték	Eset	Kontroll
SCIN	rs3735222 rs2240571 rs2240572 rs3173628	G C A A	0,0002	0,003	47,2%	36,8%
SCIN	rs3735222 rs2240571 rs2240572 rs3173628	A G G G	0,0004	0,01	39,3%	49,1%
ITLN1	rs2274910 rs4656957	T A	0,008	0,2	66,1%	72,9%
ITLN1	rs2274910 rs4656958	C G	0,0084	0,206	32,5%	25,7%

Ellenőriztem a polimorfizmusok LD (linkage disequilibrium) értékeit, vagyis az egymáshoz való kapcsoltságát, mely jól összhangban van korábban vizsgált a 2-2 gyakori haplotípussal: az erősen asszociált polimorfizmusok egy gén belül egymással is erősen kapcsoltak (8. táblázat). Mindez azt jelezheti, hogy a géneknek két funkcionális variánsa van (egy védő és egy hajlamosító típus), amely az összes vizsgált polimorfizmus asztmához való asszociációját okozza. A *PPARGC1B* gén esetén csak egyetlen mért polimorfizmus volt, így nem lehetett haplotípus szintű vizsgálatokat végezni.

8. táblázat - A polimorfizmusok kapcsoltsága a teljes adathalmazon (r-squared és D prime metrika) a (HaploView alkalmazás) [157].

Gén	SNP1	SNP2	D'	r^2
SCIN	rs3735222	rs2240571	99	61
SCIN	rs3735222	rs2240572	99	99
SCIN	rs3735222	rs3173628	94	56
SCIN	rs2240571	rs2240572	99	61
SCIN	rs2240571	rs3173628	96	91
SCIN	rs2240572	rs3173628	64	58
ITLN1	rs2274910	rs4656958	99	93

A *SCIN* gén 4 polimorfizmusa kapcsán meg kell jegyezni egy fontos, értelmezést zavaró jelenséget: a statisztikai kiértékelés szerint (lásd. OR statisztikák 6. táblázat, allélok 10. táblázat) a minor allélok két polimorfizmus esetében védenek (rs2240572, rs3735222), kettő esetében hajlamosítanak (rs2240571, rs3173628). Ez akár két különböző funkcionális hatásra is utalhatna, valójában a jelenséget csupán a konvencionális kódolás anomáliája okozza. Mind a négy polimorfizmus esetén a két allél frekvenciája közel van az 50%-hoz az európai gyökerű (CEU) populációt tekintve (9. táblázat), ilyen esetekben az allélok minor-major besorolása gyakran nem egyértelmű.

9. táblázat - A *SCIN* gén 4 polimorfizmusának egymással való kapcsoltságai és allél frekvenciái az 1000 Genome Pilot 1 CEU populáció statisztikái a Broad Institute Snap alkalmazásából (http://www.broadinstitute.org/mpg/snap). Az SNP oszlop a referencia polimorfizmust tartalmazza (a vizsgált 4 SCIN polimorfizmus), a Proxy oszlop mutatja a kapcsolt polimorfizmust (3-3 SCIN polimorfizmus) és a többi oszlop ez utóbbinak adatait.

SNP	Proxy	Távolság	r^2	D'	Major	Minor	MAF	Mintaszám
rs3173628	rs2240571	17257	0,967	1,000	G	C	0,492	120
rs3173628	rs2240572	16651	0,791	1,000	A	G	0,442	120
rs3173628	rs3735222	17566	0,765	1,000	G	A	0,433	120
rs2240572	rs3735222	915	0,967	1,000	G	A	0,433	120
rs2240572	rs3173628	16651	0,791	1,000	G	A	0,500	120

rs2240572	rs2240571	606	0,754	0,960	G	C	0,492	120
rs2240571	rs3173628	17257	0,967	1,000	G	A	0,500	120
rs2240571	rs3735222	309	0,791	1,000	G	A	0,433	120
rs2240571	rs2240572	606	0,754	0,960	A	G	0,442	120
rs3735222	rs2240572	915	0,967	1,000	A	G	0,442	120
rs3735222	rs2240571	309	0,791	1,000	G	C	0,492	120
rs3735222	rs3173628	17566	0,765	1,000	G	A	0,500	120

Például a vizsgált polimorfizmusok esetén az afrikai gyökerű (1000 Genome Pilot 1, YRU) populációt tekintve a besorolás két polimorfizmusnál is fordított a CEU populációhoz képest. A besorolás kifejezetten félrevezető is lehet, hiszen könnyedén előfordulhat, hogy két polimorfizmus közel 100%-ban kapcsolt, de az együtt előforduló variánsok minor-major besorolása pont ellentétes. Jelen esetben is ezt történt, valójában mindössze két gyakori haplotípus van, de az egy haplotípusban lévő allélok besorolása mégis ellentétes (10. táblázat).

10. táblázat - A genotípusok (allélok) esetszámai és besorolása a vizsgált populáción és egy CEU populáción.

	rs3735222	**rs2240571**	**rs2240572**	**rs3173628**
Kontroll [11]	94 (GG)	126 (GG)	94 (AA)	120 (GG)
Kontroll [12]	165 (GA)	161 (GC)	164 (AG)	164 (GA)
Kontroll [22]	85 (AA)	56 (CC)	85 (GG)	60 (AA)
Eset [11]	115 (GG)	75(GG)	115 (AA)	74 (GG)
Eset [12]	149 (GA)	155 (GC)	149 (AG)	153 (GA)
Eset [22]	41 (AA)	75 (CC)	40 (GG)	73 (AA)
Védő allél	A	G	G	G
CEU populációban a védő allél besorolása	minor	major	minor	major
A kontroll populációban a védő allél besorolása	minor	major	minor	major

Jelen tanulmányban a minor-major allél besorolást egységesen a kontroll populáció genotípus eloszlása alapján végeztem el, ez a *SCIN* polimorfizmusai esetén megegyezik a CEU populációban mért allél besorolással is (10. táblázat), és ezzel konzisztensen végeztem a statisztikai kiértékelést is. A jelenséget jól szemlélteti az a tény is, hogy a statisztikai kiértékelés során a minor allélokra számított esélyhányadosok egymásnak gyakorlatilag reciprokai (6. táblázat, OR).

Az eredmények többváltozós elemzéseit logisztikus regresszióval végeztem, ahol ügyeltem arra, hogy a nem és kor kovariánsként mindig a modellbe lépjen (adjusztálás). Az első esetben a logisztikus regresszió lépésenkénti változó kiválasztási algoritmusát alkalmaztam (Forward: Wald), tehát megvizsgáltam, mely változók tekinthetőek szignifikáns valószínűséggel kovariánsnak a modellben (p-érték < 0,05), amennyiben szignifikancia szerinti sorrendben lépteti be őket az algoritmus. A végső modellbe három, korábban is kiválasztott polimorfizmus került be (*SCIN*: rs2240572, *PPARGC1B*: rs32588, *ITLN1*: rs4656958), mindhárom esetben a minor allél ismét védő hatásúnak bizonyult (a továbbiakban közölt OR eredmények minden esetben a polimorfizmusok minor alléljának hatására vonatkozik). Az eredmény összhangban van a korábbi eredményekkel: minden génből pontosan egy polimorfizmus került be a modellbe (11. táblázat), a gének többi polimorfizmusát (melyek egyváltozós vizsgálattal erősen szignifikánsak voltak) kiszorította a modellbe került polimorfizmusok hatása. Tehát haplotípus szintű asszociáció nem figyelhető meg, az asszociáció egyetlen oki génvariáns hatása.

11. táblázat - IBM SPSS Statistics V20, Binary Logistic Regression (Forward: Wald), Dependent Variable: Asthma [157].

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 5[a]	Age	-,147	,015	96,062	1	,000	,863	,838	,889
	rs32588	-,408	,162	6,330	1	,012	,665	,484	,914
	rs4656958	-,318	,142	5,009	1	,025	,727	,551	,961
	rs2240572	-,327	,132	6,140	1	,013	,721	,557	,934
	Gender(1)	-,429	,187	5,261	1	,022	,651	,452	,940
	Constant	2,777	,279	98,919	1	,000	16,064		

DOI:10.14753/SE.2015.1757

Továbbá, az adathalmaz ketté bontásával (férfiak és nők), logisztikus regresszióval elemeztem az előző elemzés által kiválasztott polimorfizmusok nemmel összefüggő hatását (12. táblázat, 13. táblázat). A *SCIN* rs2240572 hatása (p=0,002, OR=0,41) és a *PPARGC1B* rs32588 (p=0,013, OR=0,48) hatása erősen szignifikáns volt nők között, de az *ITLN1* rs4656958 hatása egyáltalán nem volt szignifikáns. Férfiak között ennek pont a fordítottját tapasztaltuk, az *ITLN1* rs4656958 (p=0,005, OR=0,59) védő hatása erősen szignifikáns volt, míg a másik kettő alig.

12. táblázat - IBM SPSS Statistics V20, Binary Logistic Regression (Enter), Dependent Variable: Asthma, Gender = 0 (nők) [157].

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Age	-,146	,023	40,705	1	,000	,864	,826	,904
	rs32588	-,734	,295	6,209	1	,013	,480	,269	,855
	rs2240572	-,671	,222	9,168	1	,002	,511	,331	,789
	rs4656958	-,032	,223	,020	1	,887	,969	,626	1,499
	Constant	2,545	,437	33,941	1	,000	12,748		

13. táblázat - IBM SPSS Statistics V20, Binary Logistic Regression (Enter), Dependent Variable: Asthma, Gender = 1 (férfiak) [157].

Variables in the Equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Age	-,151	,020	55,184	1	,000	,860	,826	,895
	rs32588	-,251	,201	1,567	1	,211	,778	,525	1,153
	rs2240572	-,102	,168	,370	1	,543	,903	,649	1,255
	rs4656958	-,528	,189	7,815	1	,005	,590	,408	,854
	Constant	2,693	,351	58,842	1	,000	14,770		

A logisztikus regresszió azt is felfedte, hogy az *ITLN1* rs4656958 védő hatása (p=0,007, OR=0,665) a teljes allergiás asztmás csoportban szignifikánsabb volt (14. táblázat).

14. táblázat - IBM SPSS Statistics V20, Binary Logistic Regression (Enter), Dependent Variable: Atopy [157].

Variables in the Equation

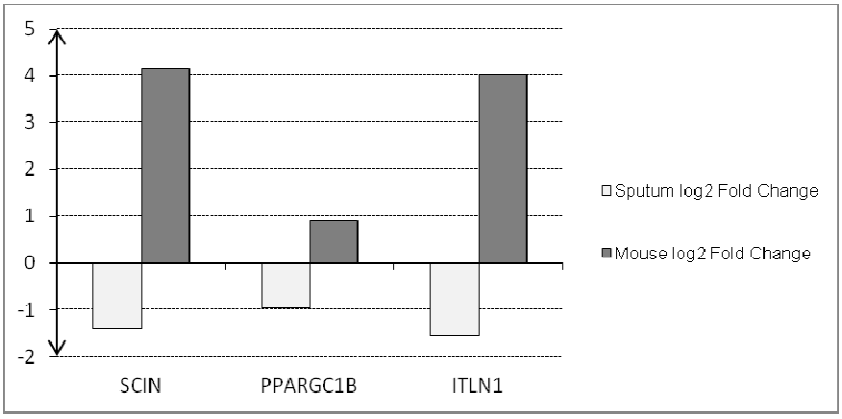
								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	rs32588	-,343	,173	3,912	1	,048	,710	,505	,997
	rs4656958	-,408	,152	7,220	1	,007	,665	,494	,896
	rs2240572	-,146	,136	1,156	1	,282	,864	,663	1,127
	Gender	,656	,196	11,153	1	,001	1,926	1,311	2,830
	Age	-,092	,013	46,780	1	,000	,912	,888	,937
	Constant	,686	,273	6,302	1	,012	1,986		

Összegezve az elemzéseket, a genotipizálási tanulmány három gén esetén talált statisztikailag szignifikáns asszociációt asztmával: *SCIN, PPARGC1B, ITLN1*.

5.1.3 Indukált köpet vizsgálat

Megvizsgáltam, hogy az indukált köpet vizsgálat eredményei megerősítik-e a *SCIN*, *PPARGC1B* és az *ITLN1* gének kapcsolatát asztmával. Mindhárom gén expressziója szignifikánsan alacsonyabb volt asztmások esetén a kontrollokhoz képest.

Az eredményeket összehasonlítottam az eredeti egér asztmamodell harmadik időpontjának eredményeivel (24 órával a harmadik, utolsó allergénterhelés után), de ez váratlan eredményt hozott. Mindhárom gén expressziója az egérmodellhez képest ellenkező irányba változott, az asztmás egérmodell tüdőszövetekben szignifikánsan magasabb volt a gének expressziója (14. ábra).



14. ábra – A kísérletbe bevont emberek és egerek expressziójának változása [157]. Asztmás betegekben az expresszió mindhárom (*SCIN, PPARGC1B, ITLN*) gén esetén csökkent, míg az egerek tüdejében ennek fordítottját tapasztaltam.

5.1.4 Eredmények összegzése Fisher módszerrel

A statisztikailag függetlennek tekinthető tanulmányok p-értékeit Fisher módszerével kombinálhatjuk, így pl. esetünkben a genotipizálás és az expressziós mérések eredményeit. Ez a módszer képes kvantitatívan egyesíteni a kísérletekből származó bizonyítékokat, és ezzel az eredmények erősségét rangsorolni. A számítást elvégezve változatlanul a *SCIN* gén p-értéke a legalacsonyabb, vagyis ebben az eseten a legerősebb a bizonyíték egy valós kapcsolatra, de *PARGC1B* génnek is relatíve alacsony a p-értéke. Megjegyzendő azonban, hogy ebben az esetben a 0,05 fölötti kombinált p-érték sem jelenti feltétlenül az asszociáció hiányát (15. táblázat).

15. táblázat - Fisher teszt eredménye [157].

Gén	Genotipizálás permutált p-érték	Köpet gén expresszió p-érték	Kombinált p-érték
SCIN	0,0046	0,0011	0,000067
PPARGC1B	0,0094	0,0537	0,004341
ITLN1	0,4679	0,0350	0,067466

5.1.5 Bayes-háló alapú Bayesi többszintű relevancia elemzés

A genotipizálási mérések - korábban részletezett módon - megtisztított adathalmazát alávetettük a Bayes-háló alapú Bayesi többszintű relevancia elemzésnek (BN-BMLA) is. A 311 aszmás betegből 238 beteghez (vagyis 73 kivételével) a korábban említett fenotípus változókon kívül (ASZTMA, ALLERGIA, Nem, Kor) rendelkezésre állt további 10 másik, szakorvos által leírt fenotípus változó: GINA (Global Initiative for Asthma besorolás), MONO/POLI (egynél több allergénre adott pozitív allergia teszt), INHALATÍV (pozitív allergiás reakció belégzett allergénre), OUTDOOR (kültéri, pl. pollen), INDOOR (beltéri, pl. házipor), TERHELÉSES ASZTMA, INFEKCIÓS ASZTMA, INTRINSIC ASZTMA, TÁRS. RA (társuló rhinithis allergica), TÁRS. CA (társuló allergic conjunctivitis). A BN-BMLA rendszerszemléletű elemzése kifejezetten alkalmas komplex fenotípus változók elemzésére, de a módszer csak teljes (hiánytalan) adathalmazok esetén alkalmazható. Mivel a hiányzás egy blokkot érintett (vagy minden komplex fenotípus hiányzott, vagy egy sem), így a véletlenszerű pótlás helyett az adathalmazt a továbbiakban kettébontottuk: az egyikben a korábbi adathalmaz (311+360 minta) elemzését végeztük el a komplex fenotípusok nélkül (a továbbiakban „egyszerű adathalmaz”), a másikban megvizsgáltuk a 238+360 mintát tartalmazó adathalmazt, immár komplex fenotípusokkal együtt (a továbbiakban „komplex adathalmaz”).

A BN-BMLA elemzés az adathalmaz összefüggéseinek gyengébb hatásait is képes kimutatni és így részletesebb képet ad változók összefüggésének rendszeréről, de így a módszer érzékenységéből adódóan bizonyos statisztikai anomáliák könnyen elfedhetik a valós hatásokat. Így a frekventista elemzéshez képest a változókkal kapcsolatban szigorúbb kritériumokat kellett megfogalmaznunk (16. táblázat). A 20%-nál magasabb hiányzást mutató polimorfizmusok (4 db) és HWE-t a kontroll csoportban sértő polimorfizmusok (2 db) kizárása egyértelmű volt. Amennyiben az asztmás vagy a kontroll csoportban egy genotípus esetszáma nem érte el az 5-öt, úgy annak statisztikai eszközökkel történő kiértékelése megbízhatatlan eredményt adhat (8 polimorfizmus kizárása). A beteg populációban torzult HWE leggyakrabban szisztematikus hibára utalhat, így a szigorúbb kritériumok indokolták néhány további változó kizárását (2 polimorfizmus).

16. táblázat - Az elemzésből kizárt polimorfizmusok és kizárás okai.

Szűrési feltétel	Változók
Alapvető minőségi szűrés	
Hiányzó genotípusok aránya 20% fölött	rs2432561 (*OSGIN1*) rs3814896 (*TFF2*) rs1821142 (*CCL8*) rs17027410 (*CHIA*)
A kontroll csoportban sérül a Hardy-Weinberg egyensúly	rs11016071 (*MKI67*) rs2857600 (*AIF1*)
Szigorúbb szűrési kritériumok	
Az egyik csoportban 5 alatti genotípus szám	rs1853665 (*ULBP1*) rs3827869 (*MAT1A*) rs1683440 (*FABP3*) rs4359077 (*PTPN7*) rs17027410 (*CHIA*) rs310830 (*E2F7*) rs352045 (*CXCL5*) rs2069827 (*IL6*)
Az asztmás csoportban sérül a Hardy-Weinberg egyensúly	rs137487 (*SYN3*) rs4245604 (*FXYD4*)
Egyéb kritériumok	
A *LY9* gén két polimorfizmusának egyik kombinációjában túl alacsony az esetszám	rs509749 (*LY9*)
A *SCIN* gén erős polimorfizmusai mind erősen kapcsoltak és egymás hatását gyengítik	rs3173628 (*SCIN*) rs2240571 (*SCIN*) rs3735222 (*SCIN*)

Az első tesztelemzéseink két további problémát vetettek fel. A BN-BMLA elemzés minden beállításban a *LY9* gén 2 polimorfizmusát (rs509749, rs474131) együttesen a többi változóhoz képest kiemelkedően erős prediktorként jelölte meg, mely a korábbi frekventista

elemzések tükrében valószínűtlennek tűnt (a frekventista módszerek a rs509749 polimorfizmust soha nem mutatták szignifikánsnak, a rs474131 polimorfizmust is csak igen gyengén). Az SPSS Statistics V20 programmal crosstab vizsgálatokat végeztem, melyek kiderítették a jelenség okát: a gén két polimorfizmusának egy kombinációja (major homozigóta rs509749, minor homozigóta rs474131) az adathalmazban összesen csak 6 betegben található meg, akik viszont mind asztmások. A jelenséget a két polimorfizmus erős LD kapcsoltsága (r2~1, D'~1) okozza, így variánsaik szinte mindig „együtt járnak", ez a kivétel nagyon ritka eset. Ráadásul ez az összes asztmásnak egy nagyon kis részét teszi ki, a polimorfizmusoknak külön-külön nincs statisztikailag szignifikáns hatása. A csoport kis mintaszáma ebben az esetben is megbízhatatlan következtetéshez vezet, így a továbbiakban a *LY9* gén kevésbé szignifikáns polimorfizmusát (rs509749) kizártam a BN-BMLA elemzésből.

A másik problémát a *SCIN* gén szintén asztmával erősen asszociált és egymással is erősen kapcsolt polimorfizmusai okozták. A korábbi elemzések egyértelműen kimutatták, hogy a génnek egy oki variánsa van, de az elemzésben megjelenő négy különböző változó erősen redundáns, a BN-BMLA MBM poszterior eloszlik a négy polimorfizmus között, így ezek egymás hatását gyengítik a modellben. Amennyiben egyértelmű a polimorfizmusok kapcsoltsága, akkor a valóságot jobban tükröző eredményeket kapunk, ha csak a gén legerősebb polimorfizmusát vonjuk be az elemzésbe főhatásként, így ezt tettük a *SCIN* esetén is (3 változó kizárása).

Az adathalmazok megfelelő előszűrése után megvizsgáltam az adatok elégségességét a következtetések megbízhatóságára vonatkozóan. Az elemzés egyik adathalmaz esetén sem talált az asztmára nézve egyértelműen releváns változókat vagy változó halmazokat, tehát a teljes modell tanulásához nem elégségesek az adatok, a továbbiakban egyes jegyek valószínűségét elemeztem.

Az egyszerű adathalmaz elemzésének első lépése az asztmához erősen releváns polimorfizmusok elemzése volt (17. táblázat). Az elemzés azt mutatta, hogy a vártnak megfelelően a Kor, a Nem, a *SCIN* és a *PPARGC1B* hatása kiemelkedően erős (MBM >

0,83). Az első érdekesség, hogy az *LGMN* gén egyik polimorfizmusa (rs9791) az előzőekhez hasonlóan kiemelkedően relevánsnak bizonyult (MBM = 0,777), így ezt tovább vizsgáltam. A polimorfizmus a korábbi frekventista elemzések során is a látókörünkbe került, de többszörös hipotézistesztelési korrekció után a hatása nem bizonyult szignifikánsnak. A második érdekesség, hogy az *ITLN1* hatása a rendszerszintű modell szerint rendkívül gyenge, összemosódik a statisztikai zajjal.

17. táblázat - Egyszerű adathalmaz, MBM posteriorok.

Változó	**MBM posterior**
Kor	0,999
rs2240572 (*SCIN*)	0,921
Nem	0,898
rs32588 (*PPARGC1B*)	0,837
rs9791 (*LGMN*)	0,777
rs225359 (*TFF1*)	0,288
rs16944 (*IL1B*)	0,215
rs6690827 (*C1QC*)	0,174
rs14451 (*UBE2T*)	0,131
rs184432 (*TFF1*)	0,125
rs3762296 (*LAPTM5*)	0,108
rs2839110 (*COL6A2*)	0,105
…	…
rs2274910 (*ITLN1*)	0,0315
rs4656958 (*ITLN1*)	0,0312

Az MBM posteriorokról általánosságban elmondható, hogy a relevancia erősségének gyors a lecsengése és 0,777 után éles letörés látható, ez jelezheti a valós hatás és a statisztikai zaj határát. A relevancia szerinti sorrendben következő 5-10 további polimorfizmus önálló hatását korábban más eszközzel sem tudtunk kimutatni, hatásuk gyenge vagy

elhanyagolható, bár érdemes lehet az interakciós vagy redundáns hatásukat tovább vizsgálni.

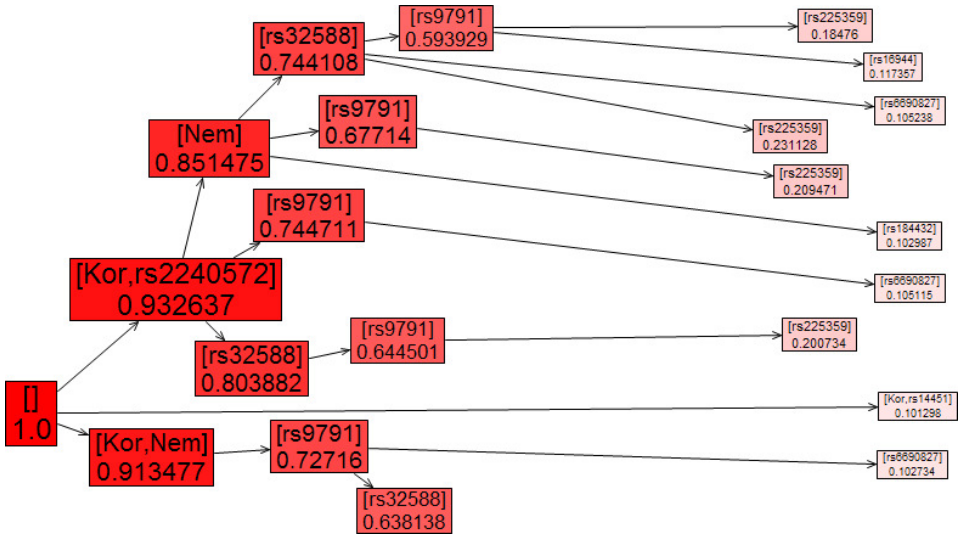
A legvalószínűbb lehetséges MBS halmazok közül több mint 60.000 posteriori valószínűségét elemezve újból a megerősítést nyert a Kor, a Nem, a *SCIN* és *PPARGC1* kiemelkedő hatása, illetve az *LGMN*, *TFF1* és néhány más gén gyengébb hatása (18. táblázat). Az MBS posteriorok lecsengése szintén gyors, jól látszik, hogy a 4 legvalószínűbb változó dominálja a legerősebb halmazokat.

18. táblázat - Egyszerű adathalmaz, MBS posteriorok.

Halmaz	**MBS posterior**
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*), rs9791 (*LGMN*)	0,039
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*), rs9791 (*LGMN*), rs225359 (*TFF1*)	0,025
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*)	0,012
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*), rs9791 (*LGMN*), rs184432 (*TFF1*)	0,0109
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*), rs9791 (*LGMN*), rs16944 (*IL1B*)	0,0094
Kor, Nem, rs2240572 (*SCIN*), rs32588 (*PPARGC1B*), rs184432 (*TFF1*)	0,0081
…	…

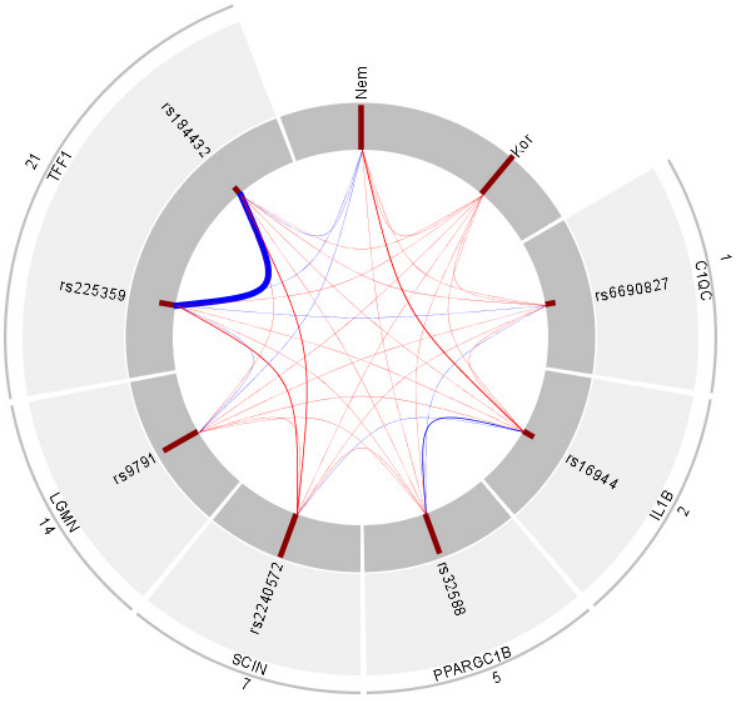
Az MBM és MBS posteriorok lecsengését dendogrammal is szemléltettem (15. ábra), ami ebben az esetben különösen látványos. A dendogrammon balról jobbra haladva láthatóak az egyes változó halmazokat reprezentáló entitások, kezdve az üres halmazzal. Balról jobbra haladva az összekapcsolt entitások a halmaz újabb elemekkel történő bővítését jelképezik. Tehát, minden egyes entitás azt a halmazt jelképezi, melynek elemei a saját címkéjén és a balra vele kapcsolatban álló entitások címkéin szerepelnek. A dendogramm entitásainak színe, mérete és vízszintes helyzete arányos azzal a posteriori valószínűséggel, hogy az

entitás által jelképezett halmaz része a Markov-takarónak („subMBS"). Másként fogalmazva arányos annak a valószínűségével, hogy a halmaz minden eleme erősen releváns (pontos értéke számszerűen is leolvasható az entitás címkéjéről). Az alábbi ábrán jól látható, hogy a Kor, a Nem, az rs2240572 (*SCIN*), a rs32588 (*PPARGC1B*), és az rs9791 (*LGMN*) változókból álló halmazok posteriorja kiemelkedő, ezeket bármilyen módon bővítve csökken a subMBS posterior valószínűség. Ez újfent azt erősíti meg, hogy az adathalmazban elkülöníthetőek a valós hatások és a statisztikai zaj.



15. ábra - Dendogram a subMBS posteriorok szemléltetésére (BayesEye szoftver). A Kor, a Nem, az rs2240572 (*SCIN*), a rs32588 (*PPARGC1B*), és az rs9791 (*LGMN*) változókból álló halmazok posteriorja kiemelkedő, bármilyen újabb változó hozzáadása csökkenti a posteriori valószínűséget.

Az erősen releváns változók redundanciáit és interakcióit elemezve a hatások egy kivételével igen gyengének bizonyulnak (16. ábra). Jól láthatóan kiemelkedik a két *TFF1* polimorfizmus (rs184432 és rs225359) erős redundanciája, ami a kapcsoltságukat tekintve egy várt hatás. Érdekes emellett a *TFF1* polimorfizmusok *SCIN* génnel való interakciójának kétszeres megerősítése. A polimorfizmusok egyenkénti interakciója bár jól látható, de nem mondható kiemelkedően erősnek, azonban a két redundáns polimorfizmusnak kiszorító hatása van a modellekben (hasonlóan a *SCIN* gén korábban kizárt polimorfizmusaihoz), így a *TFF1* gén moduláló hatása a *SCIN*-re nézve valójában kétszeresen megerősített és további vizsgálatra érdemes. Láthatunk még egy-egy gyenge interakciót és redundanciát, de ezek hatása már alig emelkedik ki a statisztikai zajból, így tovább nem elemeztem őket.

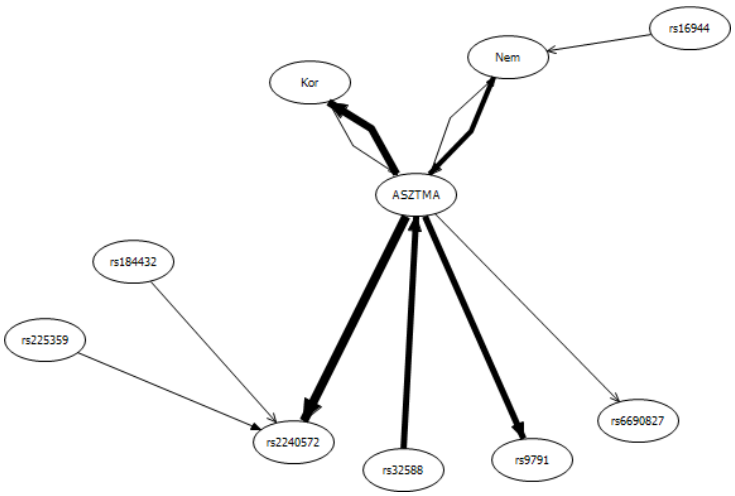


16. ábra – A vizsgált polimorfizmusok interakció-redundancia térképe (BayesEye szoftver). Kékkel a redundanciákat, pirossal az interakciókat jelöltük, a kapcsolatok vastagsága arányos annak erősségével. Az ábra kívülről befele haladva a kromoszóma számokat, a géneket, a polimorfizmusokat, az MBM posterior mértékét reprezentáló oszlopot, illetve legbelül az interakciókat és redundanciákat mutatja. Jól látható a *TFF1* gén polimorfizmusainak redundanciája és ezek interakciója a *SCIN* génnel.

Érdekes még, hogy a Nem a frekventista elemzések szerint erős interakciót mutat mind *SCIN*, mint a *PPARGC1* génnel, amely a modellben nem vagy alig jelenik meg. Ennek oka, hogy ezen változók önmagukban is kiemelkedően erősen relevánsak, így ehhez képest az interakció többlet hatása elhanyagolható. Másként fogalmazva, nem mutatható ki, hogy a polimorfizmusok a Nem-mel együtt nagyobb valószínűséggel lennének erősen relevánsak, hiszen mindig erősen relevánsak. Ez egyben azt is jelenti, hogy ebben az adathalmazban a BN-BMLA módszerrel a Nem-mel való interakció egyértelmű kimutatása csak nagyon erős hatás esetén lenne lehetséges.

Az MBG valószínűségeket elemezve (17. ábra) a korábbi megállapításaink nyertek megerősítést. A *PPARGC1B* és az *LGMN* kapcsolata az asztmával egyszerű szerkezetű, azonban a *SCIN* hatását jól láthatóan befolyásolják a *TFF1* polimorfizmusok. Érdekes még

az rs16944 (*IL1B*) Nem-mel együtt érvényesülő hatása (mely szintén azonosítható az interakciók között), illetve az rs6690827 (*C1QC*), kapcsolata, bár ezek alig emelkednek ki a statisztikai zajból.



17. ábra - Egyszerű adathalmaz, a legerősebb 10.000 MBG átlaga a 7 legnagyobb posterior valószínűségű polimorfizmusra szűrve (BayesEye szoftver). A nyilak vastagsága arányos a kapcsolat erősségével, irányultságuknak jelen modellben nincs gyakorlati jelentősége, mivel nem definiáltunk előzetesen kauzális megszorításokat. Jól látható a Kor, a Nem, az rs2240572 (*SCIN*), a rs32588 (*PPARGC1B*), az rs9791 (*LGMN*) erős, és néhány más polimorfizmus gyengébb kapcsoltsága.

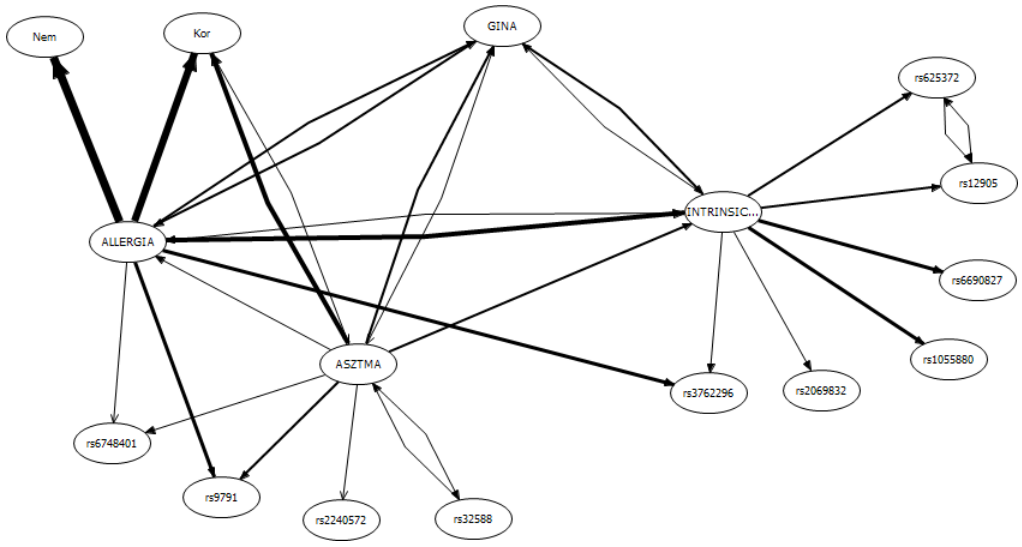
A komplex fenotípusokat tartalmazó adathalmazban arra voltunk kíváncsiak, hogy bármelyik fenotípushoz kapcsolódik-e polimorfizmus, így az elemzést a fenotípus jegyekkel, mint célváltozó halmazzal végeztük. Ebben az esetben az MBM és MBS valószínűségek a teljes fenotípus halmazra vonatkoznak, és így nehezen értelmezhetőek, bár annyi ebből is látszik, hogy a posteriorok lecsengése lényegesen lassabb az előző adathalmaznál, nincs egyértelmű letörés. Eszerint tehát „elkentebb, homályosabb" a teljes kép, nehezebben emelhetőek ki a valóban releváns jegyek és változók a statisztikai zajból, és ez a tendencia egyre több fenotípus változót elemezve tovább fokozódik (19. táblázat).

19. táblázat - A komplex adathalmaz normalizált MBM posteriorjai 4 fenotípus elemzésével.

Változó	**MBM posterior**
Kor	0,999
Nem	0,985
rs9791 (*LGMN*)	0,926

rs32588 (*PPARGC1B*)	0,741
rs12905 (*IL1RL1*)	0,738
rs3762296 (*LAPTM5*)	0,65
rs6690827 (*C1QC*)	0,611
rs625372 (*SIGLEC1*)	0,602
rs1055880 (*CD84*)	0,546
rs2240572 (*SCIN*)	0,455
rs6748401 (*MARCO*)	0,424
…	…

A legvalószínűbb Markov-takaró gráfok (MBG) elemzése már pontosabb képet adhat arról, hogy mely polimorfizmusok mely fenotípusokhoz kapcsolódhatnak, de az elkent kép miatt az elemzést több lépésben érdemes elvégezni, és többszörös megerősítéseket keresni, így először csak a 4 legfontosabb fenotípust vizsgáltam (ASZTMA, INTRINSIC, GINA, ALLERGIA), ezek még megbízhatóbb eredményt adhatnak (18. ábra).

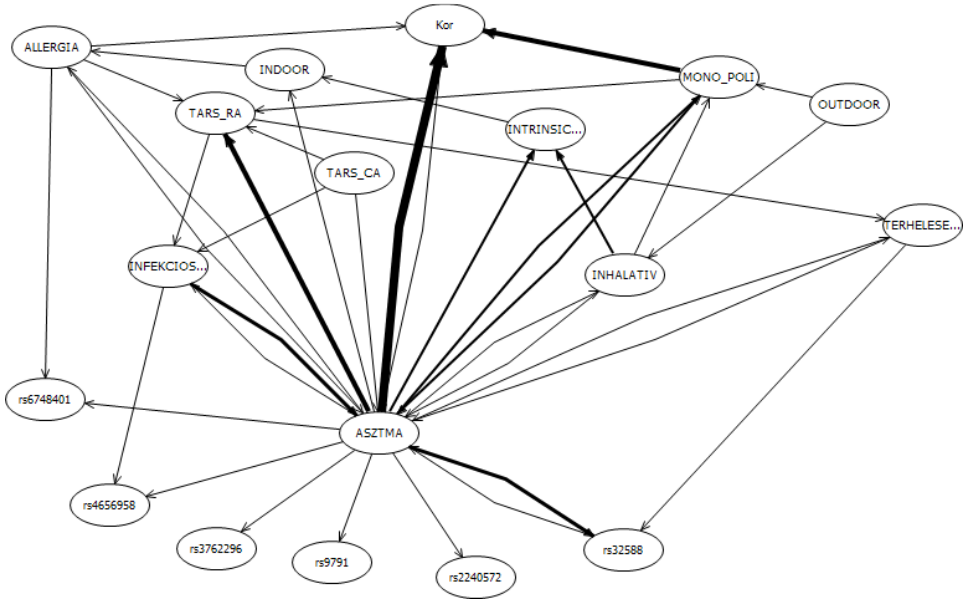


18. ábra - A komplex adathalmaz 50.000 legvalószínűbb MBG átlagával, 4 fenotípussal mint célváltozó halmazzal (a többi fenotípust kizártam az elemzésből), a 10 legerősebb polimorfizmusra szűrve (BayesEye szoftver). A nyilak vastagsága arányos a kapcsolat erősségével, irányultságuknak jelen modellben nincs gyakorlati jelentősége, mivel nem definiáltunk előzetesen kauzális megszorításokat. A korábban is vizsgált erős függések mellett kitűnik, hogy az ASZTMA, ALLERGIA, INTRINSIC asztma és GINA genetikai háttere jelentősen különbözik.

Az 4 fenotíous elemzésével előállt MBG a bizonytalan kép ellenére szolgált néhány egyértelmű tanulsággal. Az INTRINSIC asztmára hajlamosító polimorfizmusok (melyek pl. a kapcsolt anyagcsere útvonalakra utalnak) viszonylag élesen eltérnek az allergiához kapcsolódó polimorfizmusoktól. A kapcsolódó polimorfizmusok ráadásul még a korábban azonosított, asztmával asszociáló polimorfizmusoknál is erősebbek, pl. rs6690827 (*C1QC*), amelynek gyengébb hatása már kizárólag az asztma célváltozót elemezve is észlelhető volt. Kitűnik, hogy míg az ALLERGIA igen erősen függ a Kor-tól és a Nem-től, addig az INTRINSIC asztmának csak áttételes a függése, ezzel szemben a genetikai háttér hatása erőteljes. A korábban azonosított, asztmával erősen releváns polimorfizmusok (rs2240572 *SCIN*, rs32588 *PPARGC1B*, rs9791 *LGMN*) továbbra is erősebben kötődnek az asztmához mint az INTRINSIC asztmához vagy az ALLERGIA-hoz, de az rs9791 (*LGMN*) esetén erősebb, az rs6748401 (*MARCO*) esetén egy gyengébb függés figyelhető meg az ALLERGIA-val is. Érdekes, hogy az asztma súlyosságát (GINA) láthatóan kevésbé befolyásolja a genetikai háttér vagy Nem és a Kor, mint a fenotípusok eredő hatása.

A második lépésben az összes fenotípust mint célváltozó halmazt elemeztem, de ebben az esetben az eredmény már nagyon megbízhatatlannak tűnt, a sok célváltozó túl sok polimorfizmus gyenge hatását azonosította a Markov-takaróban, az eredmények teljesen összemosódtak a zajjal. Végül az elemzés komplexitását némileg csökkentve csak az asztmára, mint egyetlen célváltozóra megvizsgáltam az MBG eredményeket (19. ábra). Ebben az esetben tehát az elemzés nem az összes fenotípushoz kapcsolódó összes polimorfizmust keresi, csupán az asztmával erősen releváns változókét. Mivel a fenotípusok várhatóan lényegesen erősebben kapcsolódnak az asztmához mint a polimorfizmusok, ezek elnyomhatnak minden más hatást, de érdemes megvizsgálni azokat a legerősebb polimorfizmusokat, melyek módosítják ezeket a hatásokat. Tehát ezek az eredmények fenntartással kezelendőek, de néhány korábban is felismert hatás megerősítést nyert.

DOI:10.14753/SE.2015.1757



19. ábra - A komplex adathalmaz 50.000 legvalószínűbb MBG az átlagával, az asztmával, mint célváltozó halmazzal, a 6 legerősebb polimorfizmusra szűrve (BayesEye szoftver). A nyilak vastagsága arányos a kapcsolat erősségével, irányultságuknak jelen modellben nincs gyakorlati jelentősége, mivel nem definiáltunk előzetesen kauzális megszorításokat. A valós eredmények igen nehezen különíthetőek el a statisztikai zajtól, a korábbi erős függések megerősítése mellett az *ITLN1* génről sikerült kimutatni, hogy csak INFEKCIÓS asztma esetén releváns.

Az eredmények között a 6 legerősebb polimorfizmus között nem okozott meglepetést az rs2240572 (*SCIN*), a rs32588 (*PPARGC1B*), és az rs9791 (*LGMN*) felbukkanása, ezek minden elemzésben erősnek bizonyultak, ráadásul az rs32588 (*PPARGC1B*) a terheléses asztmával kapcsolódik össze. Az *ITLN1* (rs4656958) gén, melyet a frekventista elemzések rendre erősnek találtak, itt jelenik meg először a BN-BMLA elemzést során, és az infekciós asztmával mutat összefüggést. Továbbá, megerősítést nyert az rs6748401 (*MARCO*) ALLERGIA-val összefüggő gyenge hatása, és felbukkant az rs3762296 (*LAPTM5*) is, melyet korábban INTINSIC asztmával és ALLERGIA-val is gyengén összefüggőnek talált a rendszer – ez utóbbiak azonban változatlanul nehezen különíthetőek el a statisztikai zajtól.

5.2 Feldúsulás elemzés

5.2.1 Compound Set Enrichment Analyis (CSEA) tesztrendszer

A CSEA módszertan működésének vizsgálatához létrehoztam egy olyan szoftveres tesztrendszert, amely képes egy új hatóanyagot ismert hatóanyagok adatai alapján farmakológiai fogalmak szintjén leírni.

A GSEA technikában eredetileg alkalmazott funkcionális gén halmazok helyett az Anatomical Therapeutic Chemical (ATC) Classification System, Level 4 vegyület halmazait használtam fel a feldúsulási jelek (halmazok) definiálására. Ezt az ontológiát a World Health Organization egyik intézménye, a Center for Drug Statistics Methodology fejleszti, és hatóanyagokat kategorizál kémiai és terápiás tulajdonságok, érintett szervek és szervrendszerek szerint, így pl. alkalmas lehet potenciális alternatív indikációk jelzésére is. Bármilyen tetszőleges vegyülethalmaz, esetleg adott taxonómia különböző szintjei vagy akár taxonómiák keveréke is felhasználható annak érdekében, hogy mindig az adott célhoz a legjobban igazodjon a módszer. Ennek megfelelően más tárgyterületi taxonómiák is bevonhatóak, pl. mellékhatások szerint képzett hatóanyag halmazok, amelyek aztán az adott mellékhatások előrejelzésére használhatóak. Tehát, amennyiben van egy L rendezett hatóanyaglistánk és előre definiált S1, S2, S3… SN vegyület halmazaink, akkor célunk a halmazok feldúsulási értékeinek és normalizált szignifikanciák kiszámítása az L sorrend felett.

A hasonlósági mérték számításához egyszerű cosinus függvényt használtunk a mellékhatások és a célpont leírók esetén, illetve Tanimoto távolságot a kémiai leírok esetén. A Tanimoto távolság egy elsősorban bináris vektorok távolságának számítására használatos függvény (kiterjesztett definíciójával nem csak bináris vektorokra alkalmazható), amelyet gyakran alkalmaznak pl. vegyületek összehasonlítására, kémiai jegyek meglétét leíró bináris vektor-reprezentációk esetén. A függvény lényegében a két vegyület kémiai jegyeinek szorzata (metszete) és a két vegyületen megtalálható összes kémiai jegy összegének (uniójának) hányadosát adja.

A CMAP expressziós profilok hasonlóságainak számításához az eszköz alapértelmezett megoldását használtam. A kereséshez az alul- és felülexpresszált transzkriptumok vágási küszöbét egységesen úgy állítottam be, hogy a 100-100 vegyületből álljon a kereső halmaz [122].

A tesztjeink során kiválasztottunk egy C vegyületet, majd a különböző metrikák alapján számított hasonlóságok szerint külön-külön sorba rendeztem a referencia adatbázis összes hatóanyagát (kémiai leírók, célpont, expressziós profil, mellékhatás). A hasonlósági mértékeket a cosinus és Tanimoto függvényekből, illetve a CMAP által az expressziós hasonlóság statisztikai kiértékelésével megállapított p-értékek 1-p transzformálásával kaptuk. A sorrendezések érdekessége, hogy minden hasonlósági dimenzió egy teljesen eltérő aspektust ragadhat meg, pl. két kémiailag egyáltalán nem hasonló molekula mellékhatás profilja is lehet nagyon hasonló és ez utalhat potenciális közös célpontra.

A feldúsulás elemzés egyik erőssége pont a robosztusság, a különböző szempontok alapján készített sorrendek fúziójával, összefésülésével tehát a módszer képes lehet kiemelni a tendenciákat a zajból. A sorrendi fúzióhoz a SumScore algoritmust használtam, amely egyszerű, matematikailag torzításmentes normatív megoldásnak tekinthető. Az algoritmus a különböző dimenziók hasonlósági mértékeit nulla és egy közé transzformálja, vegyületenként összeadja, majd az előálló pontszám alapján egy mester listába rendezhetőek az entitások. A SumScore algoritmus a dimenziókra uniform súlyozást használ (nem súlyozza őket), de léteznek lényegesen komplexebb adatfúziós technikák is [150, 171]. A sorrendi fúziós algoritmusok képesek lennének hiányos adatokat (sorrendeket) is kezelni, bár tesztjeimet a torzítások elkerülése érdekében igyekeztem a dimenziók mentén közös entitás halmazzal végezni.

Végül a feldúsulási értékek számításához a SaddleSum algoritmust használtuk, mely a p-értékeket egy speciális közelítő eljárással (Lugananni-Rice formulával) határozza meg. Ezt az algoritmust tekintik ma az egyik legrobosztusabb implementációnak, mivel bemenetként felhasználja a sorrendezett lista alapját képező konkrét értékeket is (melyek az egy

dimenziós sorrendezésekben vagy a SumScore sorrendi fúzióból is elérhetőek), így problémák széles skáláját képes adaptívan kezelni [176].

5.2.2 Az amantadine esettanulmány

A CSEA módszer demonstrálásához azt vizsgáltam, hogy a technika milyen hatékonysággal használható hatóanyagok újrapozícionálására, tehát új indikációk jóslására. Ehhez ismert hatóanyagokat vizsgáltam meg, és egy valós gyógyszerfejlesztési folyamatot szimulálva a fejlesztett vegyületről gyűjtött egyre több adat segítségével a predikció hatékonyságát mértem. A számítógépes hatóanyag újrapozícionálási munkafolyamat egy komplexebb, a CSEA módszert is magába foglaló keretrendszerét (QDF^2) Bolgár és munkatársai tanulmányában mutattuk be [171], ebben a tanulmányban a feldúsulás elemzés módszerére szorítkoztam [173].

Az amantadine egy, a korábbi tanulmányunkban is vizsgált újrapozícionált vegyület [171]. Az influenzavírus protoncsatornájához való kapcsolódása, azaz M2-gátló hatása révén eredetileg influenza kezelésére fejlesztettek, mely blokkolja a vírusok fehérje burkának lebomlását (uncoating), és így a genom kiszabadulását a gazda sejtbe. Később, a hatvanas években véletlenül fedezték fel a dopamin release, dopamin reuptake gátló és N-metil-D-aszpartát receptor antagonista hatását. Az FDA-tól 1969-ben kapta meg az engedélyt Parkinson-kór kezelésére, máig használják a betegség korai stádiumában monoterápiaként vagy későbbi fázisban kombinációs terápiaként, így az amantadine egy hatóanyag újrapozicionálási sikertörténetnek tekinthető. A továbbiakban az amantadine esettanulmányának elemzésén keresztül vizsgáltam a CSEA módszer működését.

A hatóanyag fejlesztés szekvenciális természetét és a minél korábbi indikáció keresést szemléltetve a korábbi fejezetben bemutatott hasonlósági dimenziókat a következő lépésekben használtuk ki.

1. Kémiai tulajdonságok
2. Kémiai tulajdonságok + Célpont
3. Kémiai tulajdonságok + Célpont + Génexpressziós profil

4. Kémiai tulajdonságok + Célpont + Génexpressziós profil + Mellékhatás profil

Minden lépésben sorrendeztem a referencia adatbázisunk hatóanyagait az egyes adatforrások szerint az amantadinehoz való hasonlóság alapján. Az egyes lépésekben a sorrendek fúzióját a SumScore algoritmussal végeztem el a korábban ismertetett módon, mellyel így minden lépésben egyre több információt felhasználva a megfelelő információ források együttes hatását reprezentáló sorrendet kapunk. A feldúsulás elemzési lépésben a korábban ismertetett ATC taxonómia Level 4 csoportjait használtuk föl: ez összesen 778 halmazt jelent, az amantadinet minden halmazból töröltük a tesztekhez. A feldúsulási értékeket a csoportokra a SaddleSum algoritmussal számítottuk és a következő eredményeket kaptuk (20. táblázat).

20. táblázat – Az ATC csoportok feldúsulási értékei szimulált gyógyszerfejlesztési folyamat négy pontján. Félkövérrel kiemeltük a Parkinson-kór terápiájában releváns csoportokat [173].

Rank	Annotations (Chem)	E-value
1.	N06BA_centrally_acting_sympathomimetics	1,57163
2.	A14AA_androstan_derivatives	2,98917
3.	L01XA_platinum_compounds	3,01357
4.	G03AC_progestogens	3,58765
5.	**N04BC_dopamine_agonists**	4,9952
6.	C01BA_antiarrhythmics,_class_ia	5,13524
7.	**N04AA_tertiary_amines**	7,20465
8.	N06AA_non-selective_monoamine_reuptake_inhibitors	7,21062
9.	A11CC_vitamin_d_and_analogues	8,65334
10.	G03DC_estren_derivatives	9,14267
Rank	**Annotations (Chem+Target)**	**E-value**
1.	N05AB_phenothiazines_with_piperazine_structure	1,43281
2.	**N04BC_dopamine_agonists**	2,13746
3.	N06BA_centrally_acting_sympathomimetics	4,52393
4.	A03FA_propulsives	4,99707

5.	L01XA_platinum_compounds	6,38513
6.	A14AA_androstan_derivatives	6,606
7.	N06AA_non-selective_monoamine_reuptake_inhibitors	8,24364
8.	G03AC_progestogens	8,28062
9.	N05AC_phenothiazines_with_piperidine_structure	9,24803
10.	C01BA_antiarrhythmics,_class_ia	10,4918
Rank	**Annotations (Chem+Target+CMAP)**	**E-value**
1.	**N04BC_dopamine_agonists**	0,288971
2.	G03CC_estrogens,_combinations_with_other_drugs	1,36423
3.	C03CA_sulfonamides,_plain	1,94929
4.	**G02CB_prolactine_inhibitors**	2,35885
5.	C02CC_guanidine_derivatives	3,11703
6.	N05AB_phenothiazines_with_piperazine_structure	5,60553
7.	A03FA_propulsives	5,814
8.	N05AE_indole_derivatives	6,81794
9.	G01AA_antibiotics	7,26617
10.	N07BB_drugs_used_in_alcohol_dependence	7,48064
Rank	**Annotations (Chem+Target+CMAP+SE)**	**E-value**
1.	**N04BC_dopamine_agonists**	0,663519
2.	G03CC_estrogens,_combinations_with_other_drugs	3,22836
3.	**G02CB_prolactine_inhibitors**	3,64457
4.	C03CA_sulfonamides,_plain	4,10682
5.	C02CC_guanidine_derivatives	5,93538
6.	N05AB_phenothiazines_with_piperazine_structure	7,72258
7.	A03FA_propulsives	7,95207
8.	N05AE_indole_derivatives	8,59378
9.	N07BB_drugs_used_in_alcohol_dependence	11,0787
10.	**N04AA_tertiary_amines**	11,8485

DOI:10.14753/SE.2015.1757

A dopaminerg agonisták, mint a Parkinson-kór kezelésének egyik legfontosabb eszközei, konzisztensen minden lépésben előre rangsorolódtak és az információ források bővülésével egyre inkább megerősítést nyertek. Érdemes azonban megjegyezni, hogy a CSEA, a módszer „tömeghatás" jellegéből adódóan nem képes rámutatni, hogy a vizsgált halmazokban mely vegyületek hasonlósága okozta a magas feldúsulási értékeket, vagy, hogy a halmazba tartozó vegyületek mekkora része játszott szerepet. További releváns halmazok voltak pl. központi idegrendszerre ható szimpatomimetikumok, monoamine reuptake gátlók és prolactine gátlók.

DOI:10.14753/SE.2015.1757

6 Megbeszélés

6.1 Asztma genetika

Jelen tanulmányban egy korábbi egér asztmamodell kísérletünk eredményei alapján kiválasztottunk 60 gént, majd genetikai asszociációs vizsgálatot terveztünk a humán biobankunk mintáin. Egy vizsgálatra alkalmas, optimális polimorfizmus készlet kiválasztása manuálisan végrehajtva egy akár több hónapos, meglehetősen körülményes munka. Ez a lassú, humánerőforrás igényes elő-optimalizálás több nagyságrenddel csökkentheti a genotipizáló műszerek és az egész folyamat elméleti áteresztőképességét, ráadásul emberi szakértők számára sok a hibázási lehetőség. A kifejlesztett TIGER kísérlettervező rendszer segítségével az előkészítő munka időtartamát közel egy nagyságrenddel sikerült leszorítani, ennek eredményeként végül a 60 gén 90 polimorfizmusát genotipizáltuk 311 asztmás és 360 kontroll mintán.

Számos gén esetén tapasztaltunk eltérő genotípus eloszlásokat a beteg és az egészséges populáció között. Ezek közül 4 polimorfizmus asszociációja (a *SCIN* és a *PPARGC1B* génekben) minden statisztikai teszttel kimutatható volt, valamint hat másik gén asszociációja mutatott határeset közeli szignifikanciát klasszikus frekventista statisztikai módszerekkel (*ITLN1, FABP3, MAT1A, OSGIN, LY9, LGMN*).

A genotipizálási eredményeket tovább elemeztem Bayes-háló alapú Bayesi többszintű relevancia elemzéssel (BN-BMLA), hogy részletesebb képet kapjak a genetikai polimorfizmusok és fenotípus jegyek összefüggéseiről, ezek néhány új, többször megerősített eredményt hoztak. A BN-BMLA a frekventista elemzéshez hasonlóan egyértelműen azonosította a *SCIN* és *PPARGC1B* gén hatását, a *SCIN* gén esetén erős interakciót is azonosított a *TFF1* génnel, a *PPARGC1B* esetén gyenge kölcsönhatást a terheléses asztmával. Ugyanakkor a módszer az *ITLN1* gén hatását csak infekciós asztma esetén találta kiemelkedőnek, míg a frekventista módszerekkel alig kimutatható *LGMN* hatását minden elemzésben kiemelkedőnek találta, különösen allergiával együtt vizsgálva. Az intrinsic asztma endotípust vizsgálva egyértelműen kirajzolódott annak minden más asztma endotípustól eltérő genetikai háttere, így feltételezhető a merőben eltérő

patomechanizmus is. Továbbá, a BN-BMLA elemzés azt is kimutatta, hogy az intrinsic asztma esetén a genetikai háttér szerepe jelentősen erősebb a kor és a nem hatásánál; a többi asztma (és főleg az allergiás) endotípus esetén ez utóbbiaknak is erős szerepe a genetikai háttérrel együtt; míg az asztma súlyossága (GINA) és a vizsgált polimorfizmusok között nem találtam közvetlen összefüggést.

Három gén (*SCIN, PPARGC1B, ITLN1*) expresszióját megvizsgáltuk asztmások és kontroll személyek indukált köpet mintáiban is. Mindhárom gén expressziója szignifikánsan alacsonyabb volt asztmásokban. Ezen eredmények némileg váratlanok voltak, mivel a korábbi egér asztmamodell esetén mindhárom gén expressziója ellentétes irányú összefüggést mutatott. Az eredeti egér kísérletben a mérések három időpontban történtek, de az emelkedett expresszió trendszerű volt és nem oszcillált, így nem valószínű, hogy véletlenszerű hatások okozták az eltérést. Tehát fontos kérdés, hogy ezek a látszólag ellentmondásos eredmények hogyan magyarázhatóak?

Az eredményeink ellenőrzéséhez megvizsgáltam az NCBI GEO-ban publikusan elérhető asztmás microarray tanulmányok adatait és kiszámítottam az expressziós változások statisztikáit. Tsitsiou és munkatársai keringő T-sejteken végeztek microarray méréseket (GSE31773), melyek a *SCIN* és *PPARGC1B* géneket is magukban foglalták, de az *ITLN1*-et nem [177]. Az eredményeik a *SCIN* gén esetén megerősítik a mi indukált köpet vizsgálatunk eredményeit, de a *PPARGC1* esetén az eltérés nem volt szignifikáns. Egy másik publikusan elérhető asztma microarray tanulmány (GSE473) szintén vérből izolált T-sejteket vizsgált, szintén mérték a *SCIN* és *PPARGC1B* gének expresszióját, de az *ITLN1* gént nem vizsgálták. Ezen adatok megerősítették az eredményeinket mindkét gén esetén, bár a *SCIN* gén esetén az eredmények szignifikanciája csak határeseti volt. Így mindkét vizsgálat megerősítette az eredményeinket az expresszió váratlan irányával kapcsolatban, és fontos azt is megjegyezni, hogy minden, statisztikailag nem szignifikáns eredmény a mi méréseinket erősítette.

Megvizsgáltam a szakirodalmat példákat keresve humán-egér ellentétes expressziós eredményekre, és a jelenség jól ismertnek bizonyult. Tsaparas és munkatársai

szisztematikusan vizsgálták humán-egér ortológ gének ko-expressziós hálózatát, és a konzervált globális topológiai tulajdonságok mellett számottevő lényeges eltérést tapasztaltak [178]. Kuhn és munkatársai szintén humán-egér ortológ gének expresszióját vizsgálták, vizsgálataikban a szignifikáns expressziós eltérések esetén a hasonló-ellentétes expressziós változások aránya 5:1 volt [179]. Számos más tanulmány is vizsgálta a jelenséget és hasonló eredményre jutott [180-185].

Több lehetséges magyarázata is van a tanulmányunkban tapasztalt jelenségnek. Az első, hogy az egérmodell esetén a teljes tüdőt vizsgáltuk, míg az emberek esetén a köpetet. Bár mindkettőt jelentősen befolyásolja a légutak gyulladása, a sejtes összetételük lényegesen eltér, így a génexpressziók átlaga különbözhet. Továbbá, ismert, hogy az allergén indukált légúti gyulladás egérben egy akut jelenség, és ennek a folyamatnak a dinamikus expressziós változásait mértük. Ezzel szemben a humán asztma egy krónikus betegség; a tanulmányunkban részvevőknek stabil, krónikus asztmájuk volt akut szimptómák nélkül. Felvetődhet, hogy akut fázisban alul- vagy felülexpresszált gének visszaszabályozását látjuk a krónikus fázisban. Lehetséges, hogy alapvetően különböző folyamatokról van szó a két faj esetén, bár ez nehezen magyarázná meg az ellentétes irányú expressziót. Bármi is a jelenség valós háttere a gének eltérő expressziója azt jelezheti, hogy azok érintettek a folyamatban.

Fontos megjegyezni, hogy az indukált köpet biobankunk túl kicsi volt egyértelmű következtetések levonásához. Továbbá, az asztmás betegek egy nagyon heterogén populációt alkotnak, de az expressziós vizsgálat kis mintaszáma miatt ebben a tanulmányunkban nem tudtuk az alcsoportokat külön is megvizsgálni.

Ez volt az első tanulmány, amely a *SCIN* gén és az asztma asszociációját vizsgálta, és kimutatta humán populációban. A *SCIN* gén három polimorfizmusának, az első exonon az rs2240572-nek (H61R) a minor (guanin variáns), valamint a promoter régióban az rs2240571-nek a major (guanin variáns) és az rs3735222-nek a minor (adenin variáns) allélja statisztikailag erősen szignifikáns védő hatást mutat az asztmával szemben. A hatás a nőket külön vizsgálva még erősebbnek bizonyult, valamint a polimorfizmusokat együtt

vizsgálva kimutattam, hogy valójában egy oki variáns felel az asszociációért. Az asztmások indukált köpet vizsgálataiban szignifikánsan alacsonyabb génexpressziót mértünk az egészségesekhez képest.

A *SCIN* gén funkcióját tekintve a *SCIN*derin (más néven adseverin) fehérjét kódolja, mely az aktin-filamentum sapkázását („capping") és hasítását („severing") katalizáló enzim, ezzel képes az apikális aktin sapka átrendezésére a légúti goblet sejtekben [186, 187]. Habár az egér asztmamodellekben az allergénterhelés után a *SCIN* fokozott expresszióját több tanulmány is megerősítette [70, 71], eddig humán tanulmányok nem számoltak be asztmások és kontrollok között eltérő expresszióról. A *SCIN* gén egyik asztmával asszociált polimorfizmusával kapcsolatban (rs2240571), mely kb. 400 bázissal a gén előtt helyezkedik el a 7. kromoszómán, már sclerosis multiplex esetén is megfigyeltek asszociációt [188]. A BN-BMLA bizonyítékot talált a *SCIN* gén és a *TFF1* összefüggésére is. A *TFF1* a trefoil faktor 1 fehérjét kódoló gén, mely a trefoil fehérjék családjába tartozik (40 aminosavból álló trefoil domaint hordozzák), a mucus fontos alkotóeleme és gyulladásos folyamatokban is szerepe van. Több tanulmány is összefüggésbe hozta asztmával, egérmodellekben ki is mutatták, hogy allergén hatására termelődése fokozódik [189]. Ismereteink szerint nem végeztek tanulmányokat a *SCIN* polimorfizmusainak funkcionális vagy gén szabályzó hatásával kapcsolatban, így a mechanizmus pontosabb megéréséhez további vizsgálatokra lenne szükség.

A peroxisome proliferator-activated receptor gamma coactivator 1-beta (*PPARGC1B*) egy nukleáris hormon receptor, melynek szerepe van a sejt aktivációjában, differenciálódásában és az apoptózisban is. A PPARgamma megváltozott expresszióját asztmás légutakban már korábban is megfigyelték, egyre több bizonyíték támasztja alá szerepét a légút gyulladásban és hiperreszponzívitásban [190]. Továbbá, a PPARgamma agonistákkal végzett kísérletek alapján felvetődött, hogy a PPARgamma egy potenciális asztma terápiás célpont is lehet [191]. A PPAR transzkripciós faktorként működik, aktivitását fokozza a PPARC1B, így a PPARGC1 genetikai variánsai valószínűleg hozzájárulnak a funkcionális változásaihoz, és ezen keresztül légúti gyulladáshoz és remodellinghez vezethetnek az asztmában.

Az utóbbi években Lee és munkatársai számoltak be egy tanulmányról, amely nem talált összefüggést a *PPARGC1B* polimorfizmusai és az asztmára való hajlam között [192]. Emellett viszont gén és fehérje expressziós vizsgálatokkal arra találtak bizonyítékot, hogy a *PPARGC1B* promoterén lévő -427C>T polimorfizmus és a gén 5. exonjában lévő +102525G>A (R265Q, rs45520937) polimorfizmus a *PPARGC1B* géntermék megváltoztatásával hatással van légúti hiperreszponzivitásra. Jelen tanulmányunk volt az első, amely összefüggést talált a gén 2. exonján lévő rs32588 (L42L) polimorfizmus és az asztma között. A minor allél védő hatását mutattuk ki, a BN-BMLA elemzésekben a polimorfizmussal gyenge kölcsönhatást mutatott terheléses asztmával is, de a pontos hatásmechanizmusa nem tisztázott. Mivel az egyetlen vizsgált polimorfizmus kódoló régióban van, de szinonim hatású, így akár az is feltételezhető, hogy valójában más, kapcsolt, funkcionális hatású polimorfizmusok felelősek a megfigyelt erős asszociációért.

Az intelectin-1 (*ITLN1*) képes felismerni patogének és bakteriális komponensek szénhidrát láncait, így a védekező mechanizmusokban játszik szerepet. A génterméknek gyulladásellenes hatása is van a TNF-α indukálta COX-2 expresszió gátlásával. Mivel a mikrobás fertőzések és a TNF-α is fontos szerepet játszanak az asztma kialakulásában [193-196], így az *ITLN1* génnek is lehet szerepe a folyamatban. A gén rs2274907 polimorfizmusát korábban már összefüggésbe hozták az asztmával [197], az rs2274910 és az rs4656958 polimorfizmus asszociációját asztmával elsőként vizsgáltam. Habár a genetikai variánsok asszociációja nem minden statisztikai teszttel volt kimutatható (az rs4656958 polimorfizmus lényegesen erősebben), a gén eltérő expressziója mind az egértüdőben, mind a humán köpet mintákban kimutatható volt. A BN-BMLA elemzés rámutatott, hogy a polimorfizmusnak infekciós asztmában van csak szerepe, ami összhangban van a gén ismert funkcionalitásával, és ez magyarázat lehet a gyengébb asszociációra az összes többi asztma endotípussal.

Az *LGMN* gén a legumain fehérjét kódolja, mely egy cisztein proteáz enzim, bakteriális eredetű és egyéb endogén peptidek bontásában van szerepe. Tanulmányunk klasszikus frekventista elemzéssel gyengébb, BN-BMLA elemzéssel erős bizonyítékot talált a gén és az asztma összefüggésére, különösen az allergiás csoportban vizsgálva. A tanulmányban

vizsgált rs9791 polimorfizmus a gén kódoló régiójának kezdetén helyezkedik el, de szinonim hatású, így feltételezhető, hogy egy kapcsolt oki variáns felelős a kimutatott hatásért. A fehérje hatása nem teljesen tisztázott, ismereteink szerint asztmával korábban nem hozták összefüggésbe, de rákos és gyulladásos folyamatokban való szerepét már vizsgálták [198].

Meg kell jegyezni, hogy a hamis asszociációk egy lehetséges forrása az eset és kontroll populáció életkora között nem jelentős, de statisztikailag szignifikáns eltérés. Ezt egyrészt a többváltozós elemzéseinkben azzal kezeltük, hogy a kor, mint kovariáns mindig jelen volt a modellekben (ez pedig statisztikailag elegendő adjusztálásnak tekinthető), továbbá statisztikai validációkat is végeztünk. Másrészt, az eset és a kontroll populáció etnikai összetétele azonos volt, a toborzás azonos helyen történt, így igen valószínűtlen, hogy a vizsgált polimorfizmusok eloszlását szignifikánsan befolyásolta volna a 11 év különbség az átlag életkorban, és ezt minden vizsgálatunk is megerősítette.

Összességében elmondható, hogy többször megerősített, erős összefüggéseket találtam asztma és a *SCIN* és *PPARGC1B* gének, valamint gyengébb, de többször megerősített összefüggéseket az asztma és az *ITLN1* és az *LGMN* gének között. A klasszikus frekventista statisztikával csak kiemelkedően erős összefüggéseket sikerült feltárni, míg a rendszerszemléletű BN-BMLA segítségével igen fontos gyengébb összefüggéseket is sikerült azonosítani, amelyek információkat szolgáltathatnak a pontos patomechanizmus feltárásához.

6.2 Feldúsulás elemzés

6.2.1 Az esettanulmány megbeszélése

Habár az eredmények látszólag alátámasztják az amantadine dopaminerg hatását, több zavaró tényezőt érdemes kiemelni. Más, számítógépes indikáció-keresési tanulmányok is jelezték az expressziós adatok zajosságát [199], a CMAP információ forrás számos nehezen magyarázható csoport feldúsulását jelzi (pl. ösztrogének és szulfonamidok). Egy másik zavaró tényező, hogy az ATC csoportok valójában nem indikációkat jeleznek, sokkal inkább anatómiai, kémiai vagy egyéb gyakorlati szempontok szerint csoportosítják a hatóanyagokat, így ezek a csoportok az új indikáció-keresési problémákhoz kevésbé praktikusak. Éppen ezért más taxonómia adatbázisok használatát is megfontoltuk, de a hiányzó adatok és az adatintegrációs nehézségek miatt a használatukat elvetettük.

A sorrendi fúzió és a CSEA együttes alkalmazása jól kezeli a zajos adatokból eredő problémákat, bár a javasolt módszer a hasonlóságokra épít, így a rendszer elméleti hatékonysága alacsony, ha nincs a referencia adatbázisban hasonló ismert vegyület. Ezt a hatást komplementer információforrások felhasználásával lehet semlegesíteni, pl. ha a nagyon eltérő szerkezetű vegyületek azonos szabályzási útvonalak mentén hatnak, akkor az expressziós vagy mellékhatás profiljuk hasonló lesz. Amennyiben az egyes információ források hibái függetlenek egymástól, úgy a módszer ezeket is átlagolja és semlegesíti.

Egy igen fontos zavaró hatás is a módszer hasonlóságokra építő működéséből adódik. A 20. táblázat eredményei több olyan csoportot emelnek ki, melyek dopaminerg antagonistákat, pl. fenotiazinokat és propulzív szereket (pl. metoclopramide) tartalmaznak. Ez a dopaninerg agonisták és antagonisták hasonló kémiai szerkezetének és célpontjaiknak köszönhető, a hasonlóságon alapuló módszer nem képes ezeket megkülönböztetni. Az anomália olyan információforrások használatával oldható fel, melyek nem érzékenyek erre a hibára, pl. a mellékhatás profil vagy az expressziós profil. Pusztán azt a két információforrást felhasználva, a dopaminerg antagonisták nem jelennek meg, míg a dopaminerg agonisták igen (21. táblázat). Ehhez hasonlóan (de az előbbinek inverz esete) a kémiai szempontból eltérő, de azonos biológiai folyamatokra ható vegyületek összekapcsolhatóak [200] (pl. az

amantadine és a dopaminerg agonista Pergolide eltérő kémiai szinten, de több információ forrás használatával a közös biológiai funkció láthatóvá válik). Szintén ezt illusztrálja, hogy e két információ forrás felhasználásával az antivirális csoport első helyre ugrik (mely az amantadine másik fontos indikációja), pedig ez a kémiai és fehérje célpont profil alapján nem szerepelt az első tízben sem.

21. táblázat - A CSEA eredménye a két fenotípus információ forrás felhasználásával, a releváns ATC csoportok félkövérrel kiemelve [173]**.**

Rank	Annotations (CMAP+SE)	E-value
1.	**S01AD_antivirals**	0,702238
2.	G03CC_estrogens,_combinations_with_other_drugs	0,819319
3.	J05AB_nucleosides_and_nucleotides_excl._reverse_transcriptase_inhibitors	0,886404
4.	C03CA_sulfonamides,_plain	2,09367
5.	M01AX_other_antiinflammatory_and_antirheumatic_agents,_non-steroids	2,86707
6.	D10AF_antiinfectives_for_treatment_of_acne	3,47238
7.	**N04BC_dopamine_agonists**	3,69115
8.	**G02CB_prolactine_inhibitors**	4,66169
9.	G01AA_antibiotics	4,7444
10.	J01MB_other_quinolones	5,56766

6.2.2 Az információ újrahasznosítás

A gyógyszerfejlesztés információ technológiai szempontból egy inkrementális adatgyűjtési folyamat, mely egy elméleti molekulastruktúrától indul és több mint egy évtized alatt az in vitro, in vivo majd klinikai teszteken keresztül végül a piacra kerülhet. A gyógyszerfejlesztési folyamat hipotézisvezérelt jellege és az erős fókuszáltsága elfogulttá teheti az eredmények értelmezését, és elhomályosíthat fontos részleteket. Fontos felismerni, hogy ebben a folyamatban minden apró új információnak, minden új mérésnek vagy kíséreltnek értéke lehet, akkor is, ha azt nehéz beilleszteni az eddig felhalmozott

tudásanyagba, vagy akár ellentétes az eredeti céllal. Előfordulhat, hogy a kívánt terápiás célt nem sikerül elérni, de maradnak alternatív terápiás lehetőségek, ezért az adaptív gyógyszerfejlesztési megközelítések egyre több területen meghonosodnak [201-203].

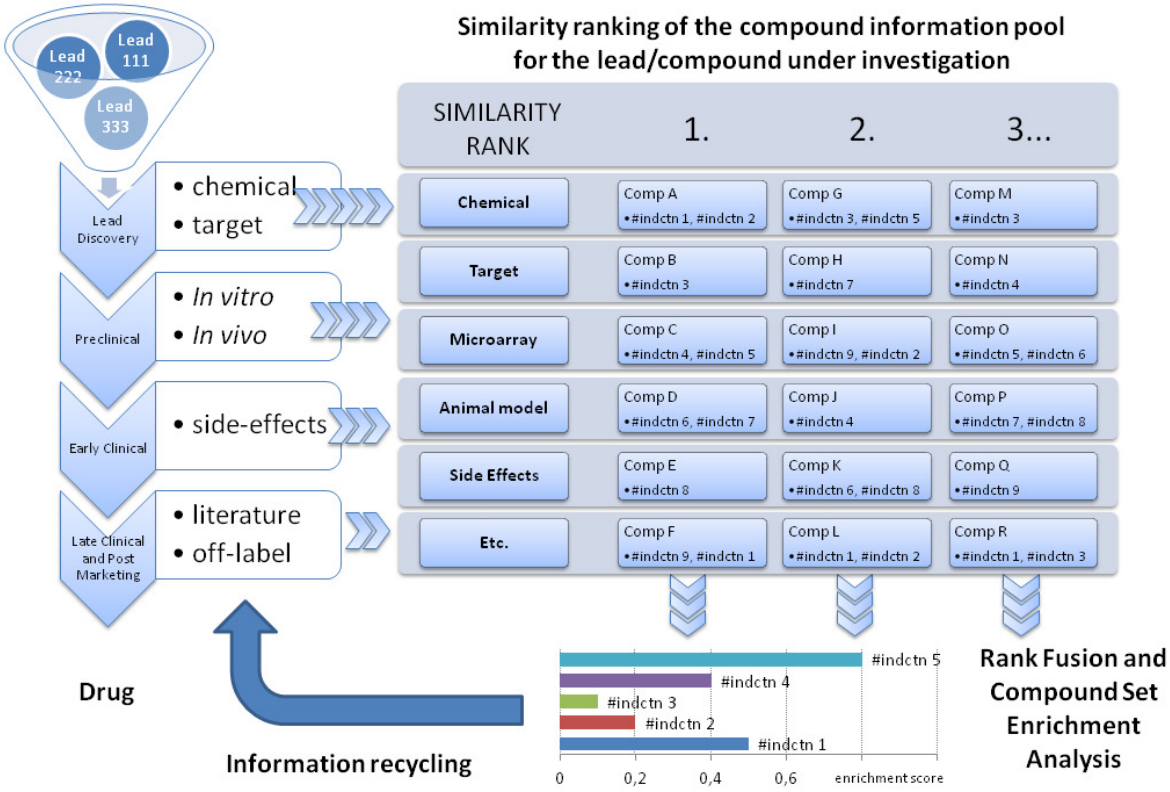
A legtöbb hatóanyag újrapozícionálási sikertörténet véletlenszerű felfedezésekhez kapcsolódik és többnyire 2006 előtt történt. A későbbi szisztematikus újrapozícionálási erőfeszítések közül kevés váltotta be a hozzá fűzött reményeket; úgy tűnik, hogy a nyilvánvaló újrapozícionálási lehetőségek többségét már korábban kimerítették. A fejlesztési folyamat azonban hosszadalmas, így túl korai lenne ítélkezni, de a hatóanyag újrapozícionálási technikákat egyre korábbi gyógyszerfejlesztési fázisokban kezdik alkalmazni és a statisztikák lassú szemléletváltást vetítenek előre. Novac és munkatársai részletesen elemzik ennek az átmenetnek a hátterét és trendjeit: a véletlenszerszerű felfedezések helyett a gyógyszerfejlesztés szerves részét képező, az egyes fázisok kimenetéhez igazodó, adaptív módszerek terjedését; valamint az akár több párhuzamos indikáció fejlesztését is támogató, elágazó fejlesztési folyamatokat [85].

A tudás újrahasznosítás informatikai értelemben egy olyan szemlélet, amely azt hangsúlyozza, hogy a teljes felhalmozott tudást szisztematikusan ki kell aknázni ahol csak lehetséges és érdemes, függetlenül annak eredetétől vagy eredeti céljától. Ez a megközelítés motiválta a számos iparágban elterjedt adattárház technológiák kifejlesztését is. A gyógyszerfejlesztés eddigi gyakorlatában a hatóanyag fejlesztés lezárulta után az előállított jelentős mennyiségű adat hasznosítása is lezárul. Tágabb értelemben a CSEA a tudás újrahasznosítás eszköztárával növelheti gyógyszerfejlesztés hatékonyságát.

A CSEA módszer tesztjeihez használt referencia adatbázis jól reprezentálja az engedélyezett hatóanyagokról elérhető heterogén információtömeg teljes spektrumát, így ez egy ideális alap a tudás újrahasznosítás szemléltetéséhez. Az általam javasolt stratégiánk három fontos pillére van: 1. az adatok fokozatos gyűjtése és rendszerezése, 2. a (publikusan vagy belsőleg elérhető) adatvagyon rendszeres újra elemzése, 3. egy kiegyensúlyozott számú és kiterjedésű indikátorhalmaz használata, mint pl. potenciális új indikációk vagy mellékhatások előrejelzése. A módszer jól beilleszthető a gyógyszerfejlesztés

DOI:10.14753/SE.2015.1757

információáramlási folyamatba, ráadásul statisztikailag robosztus, számítástechnikailag jól skálázható, és a farmakológiai fogalomrendszerek szintjén képes a farmakológusok által definiált fontos kérdéseket jelezni (20. ábra).



20. ábra - Tudás újrahasznosítás: a CSEA módszertan integrálása a gyógyszerfejlesztési folyamatba egyre növekvő mennyiségű információt felhasználva, Temesi és munkatársai tanulmányából [173]. A gyógyszerfejlesztés egyes stádiumaiban (preklinikia és klinikai fázisok) előálló adatok (kémiai profil, célfehérje profil, expressziós profil, stb.) segítségével a CSEA egyre finomítja a következtetéseket (pl. indikáció jóslást), ezek pedig felhasználhatóak a gyógyszerfejlesztés következő stádiumainak tervezésében.

A módszer kiforrott felhasználása előtt természetesen számos kihívás áll, a tudás újrahasznosítás hosszútávú alkalmazása, különösen a referencia adatbázis karbantartása (reprezentációk, hasonlósági metrikák, hatóanyag halmazok, stb.) egy nyitott kérdés. Fontos azt is megjegyezni, hogy a többszörös hipotézistesztelési probléma is gátat szab a tudás újrahasznosítás hipotézismentes alkalmazásának: minél több hatóanyag csoport feldúsulását vizsgáljuk, annál több statisztikai erőt vesztünk el a többszörös hipotézistesztelési korrekció miatt (a mi esetünkben Bonferroni korrekció alkalmazva). Ez

egy jelentős elméleti korlátja a rendszer automatizálásának, mely így egy folyamatos egyensúlyozást igényel a rendszer alkalmazási területének kiterjedése és következtetéseinek erőssége között.

A munkámat megelőzően a feldúsulás elemzést farmakológiai kontextusban kizárólag egy speciális területen, korai fázisú gyógyszerfejlesztésben alkalmazták egy-egy kutatásban vizsgált molekula könyvtárakra, ahol kizárólag biológiai aktivitás alapján sorrendezték a molekulákat és az elemzés során csak molekuláris szerkezeti elemek feldúsulását vizsgálták. Ezzel szemben bemutattam, hogy a módszer sikeresen alkalmazható a gyógyszerfejlesztés minden fázisában: molekulakönyvtárak elemzése helyett integrálva akár az összes ismert hatóanyag mindenfajta mérési adatait (a szerkezeti tulajdonságoktól a mellékhatás profilig), és segítségével tulajdonságok igen széles köre jósolható (akár mellékhatások vagy új indikációk is). Ezzel a kiterjesztés 3 ortogonális dimenzióban is megvalósult és a módszer felhasználási területe is kibővült. Bár egy költséges experimentális validáció meghaladta jelen munka kereteit és a rendszer finomhangolása további teszteket igényelhet, azt sikerült egyértelműen demonstrálni, hogy molekuláris biológiából „újrapozícionált” feldúsulás elemzési keretrendszer sikeresen alkalmazható a farmakológiai kontextusban a heterogén információtömeg kiaknázására és újrahasznosítására.

DOI:10.14753/SE.2015.1757

7 Következtetések

7.1 Kísérlettervező rendszer

A genetikai asszociációs vizsgálatok tervezése során egy vizsgálatra alkalmas, optimális polimorfizmus készlet kiválasztása emberi szakértők számára akár több hónapos, meglehetősen körülményes munka. Ez a lassú, humánerőforrás igényes elő-optimalizálás több nagyságrenddel csökkentheti a genotipizáló műszerek és az egész folyamat elméleti áteresztőképességét, ráadásul sok a hibázási lehetőség. A kifejlesztett TIGER kísérlettervező rendszerrel az előkészítő munka fontos fázisai automatizálhatóak, és annak időtartama közel egy nagyságrenddel leszorítható.

7.2 *SCIN*

Ez volt az első tanulmány, amely a *SCIN* gén és az asztma asszociációját vizsgálta humán populációban. A *SCIN* gén három polimorfizmusának, az első exonon az rs2240572-nek (H61R) a minor, valamint a promóter régióban az rs2240571-nek a major és az rs3735222-nek a minor allélja statisztikailag erősen szignifikáns védő hatást mutat az asztmával szemben. Frekventista statisztikai módszerekkel egyetlen potenciális oki variáns hatását azonosítottam, mely a nőket külön vizsgálva még erősebbnek bizonyult. A BN-BMLA módszer megerősítette az eredményeket, és a *TFF1* gén módosító hatását sikerült azonosítani.

7.3 PPARGC1B

Jelen tanulmányom volt az első, amely összefüggést talált a *PPARGC1B* gén 2. exonján lévő rs32588 (L42L) polimorfizmus és az asztma között. Frekventista statisztikai módszerekkel a minor allél erősen szignifikáns védő hatását mutattuk ki, a hatás a nőket külön vizsgálva még erősebbnek bizonyult. A BN-BMLA elemzések megerősítették az eredményeket és a vizsgált apolimorfizmus gyenge kölcsönhatást mutatott terheléses asztmával is.

7.4 ITLN1

A tanulmány során azonosítottam az *ITLN1* gén rs4656958 polimorfizmusának asztmával összefüggő potenciális védő hatását. Az asszociációt több módszerrel is sikerült kimutatni, bár szignifikanciája frekventista módszerekkel elemezve határesetinek tekinthető. A BN-BMLA módszertan kimutatta, hogy az infekciós asztmás csoportban a hatás jóval erősebb, ami összhangban van a gén ismert funkciójával is: a géntermék bakteriális komponensek felismerésében játszik szerepet.

7.5 LGMN

Kutatásom az *LGMN* gén potenciális szerepét elsőként azonosította az asztma patomechanizmusában. A tanulmány klasszikus frekventista elemzéssel, allélikus teszttel gyenge, domináns-recesszív modellben lényegesen erősebb, míg BN-BMLA elemzéssel kifejezetten erős bizonyítékot talált a gén és az asztma összefüggésére, különösen az allergiás csoportban vizsgálva.

7.6 Humán asztma és egér asztmamodell expresszió

Ovalbumin-indukált egér asztmamodellben, a tüdőszövetekben a *Scin,* a *Ppargc1b* és az *Itlna* gének emelkedett expresszióját mértünk, míg humán indukált köpet mintákon, az alsó légúti szekrétumban a homológ gének expressziója a krónikus asztmás populációban csökkent. Az eredmények mindhárom esetben statisztikailag szignifikánsak voltak. Így a gének patomechanizmusban betöltött szerepe többszörös megerősítést nyert, de az eltérő irányú expresszió megmutatta a patomechanizmus eltérő dinamikus tulajdonságait, és így az egér asztmamodell korlátait.

7.7 Compund Set Enrichment Analysis (Hatóanyag feldúsulás elemzés)

Ez a tanulmány volt az első, amely a molekuláris biológiában már bizonyított feldúsulás elemzés matematikai megközelítését farmakológiai adatok széles körű integrálására sikerrel alkalmazta. A módszertan a korábbi speciális alkalmazást jelentősen kibővítve, tetszőleges hatóanyag könyvtárak változatos mérési adataink alapján, magas szintű farmakológiai tulajdonságok előrejelzését is lehetővé teszi a gyógyszerfejlesztés minden fázisában.

DOI:10.14753/SE.2015.1757

7.8 Információ újrahasznosítás a CSEA módszertannal

A CSEA módszertan egy statisztikailag robosztus, számítástechnikailag jól skálázható megoldásnak bizonyult a farmakológiai ismeretanyag újrahasznosítására, amely a gyógyszeripariban egy jelentős lehetőségekkel kecsegtető, de eddig mellőzött módszer volt. A módszertan a gyógyszerfejlesztésbe könnyen beilleszthető iteratív módon, heterogén információ források jelzéseinek integrálásával képes lényegesen eltérő vegyületek alapján is következtetéseket levonni és a gyenge hatásokat kiemelni a statisztikai zajból.

DOI:10.14753/SE.2015.1757

8 Összefoglalás

Az egészségügy és gyógyszeripar eddigi gazdasági modelljei fenntarthatatlanok, alapvető szemléletváltásra és innovatív megoldásokra van szükség. Az elmúlt évtizedekben a molekuláris orvostudományok és az információ technológiák olyan új interdiszciplináris eszköztárat adtak az emberiség kezébe, amelyek kulcsszerepet tölthetnek be ebben az átalakulásban. Kutatásaim célja volt hozzájárulni az egészségügy és gyógyszeripar küszöbön álló átalakulásához az információ technológia eszközeivel, ennek során két párhuzamos kutatási témával foglalkoztam.

Elsőként, kifejlesztettem egy szoftver rendszert, amely a genetikai asszociációs vizsgálatok kísérlettervezésének sebességét közel egy nagyságrenddel tudta felgyorsítani. A kifejlesztett rendszer segítségével egy korábbi ovalbumin indukált egér asztmamodell alapján kiválasztottuk 60 potenciálisan asztmával összefüggő gén 90 polimorfizmusát mérésre, majd 671 humán minta vizsgálatával sikerült kimutatni két gén (*SCIN, PPARGC1B*) erős, és néhány gén gyenge összefüggését krónikus asztmával. A két gén expressziójának megváltozását sikerült megerősíteni humán indukált köpet vizsgálatokkal is. A mérési eredményeket a klasszikus frekventista módszerek után Bayes-háló alapú Bayesi több szintű relevancia elemzéssel is megvizsgáltam. A rendszerszintű modellezés nem csak megerősítette a kimutatott asszociációkat, de sikerült tovább pontosítani is a gének és fenotípusok kapcsolatát. A gének pontos szerepét és a patomechanizmust további kísérleteknek kell feltárniuk, de a vizsgálataink nem zárják ki, hogy a géntermékek akár potenciális új gyógyszercélpontok is lehetnek.

Második kutatási témám a hatalmas mennyiségű, heterogén farmakológiai tudás egységes értelmezését és a gyógyszerfejlesztésben történő felhasználását segítő módszertani fejlesztés volt. A módszer matematikai hátterét a transzkriptom elemzésben évek óta sikerrel alkalmazott feldúsulás elemzés keretrendszere adta. Javaslatot tettem a technika beillesztésére a gyógyszerfejlesztés információtechnológiai munkafolyamatába; majd bemutattam, hogy a technika hogyan képes a farmakológiai fogalomrendszer szintjén új indikációkat jelezni. Eredményeim szerint az újszerű megközelítés egy értékes eszköze lehet az egyre növekvő farmakológiai adatvagyon folyamatos újrahasznosításának.

DOI:10.14753/SE.2015.1757

9 Summary

The current business models of the healthcare system and the pharmaceutical industry are not sustainable anymore; there is a huge need for paradigm shifts and innovative solutions. Molecular biology and information technology put a new interdisciplinary toolset into the hands of mankind in the last few decades which could be the key to this transition. The goal of my research was to contribute to the imminent transformation of the healthcare system and the pharmaceutical industry with information technology advancements; I was focusing on two parallel topics.

First, I designed a software system, which could speed up the study design process of genetic association studies by nearly an order of magnitude. With the help of this system and based on a former ovalbumin induced mouse model of asthma we selected 90 polymorphisms of 60 potential asthma associated genes. I analyzed 671 human subjects and succeeded in identifying highly significant effect of two genes (*SCIN, PPARGC1B*) and borderline significant effect of a few other genes in the chronic asthma group. The potential role of both genes was also confirmed by an induced sputum gene expression study. The results were analyzed with traditional frequentist statistics and also with Bayesian network based Bayesian multilevel analysis of relevance. The systems based analysis confirmed the results and provided additional information on the relationship of the genes and phenotypes. Additional studies should be carried out to clarify the role of the genes and pathomechanisms, but the identified gene products might be new potential therapeutic targets.

The second theme was a methodological research and development for the unified interpretation of the heterogeneous pharmacological knowledge and its application in drug development. The original mathematical background was based on the set enrichment analysis framework which is a proven method in transcriptome differential analysis. I suggested a potential integration path into drug development information workflow; and I demonstrated how the suggested solution is capable of indicating critical issues on the level of high level terms used in pharmacology. The innovative solution proved to be a valuable tool in continuous recycling of the growing pharmaceutical knowledge base.

10 Irodalomjegyzék

1. Clayton M. Christensen JHG, Jason Hwang: The Innovator's Prescription: A Disruptive Solution for Health Care. McGraw-Hill, New York, 2009: 73-182.
2. Alemayehu B, Warner KE. (2004) The lifetime distribution of health care costs. *Health services research* 39(3): 627-642.
3. OECD iLibrary: Total expenditure on health 2009, http://dx.doi.org/10.1787/20758480-2009-table1.
4. OECD iLibrary: Total expenditure on health 2013/2, http://dx.doi.org/10.1787/hlthxp-total-table-2013-2-en.
5. Kessel M, Frank F. (2007) A better prescription for drug-development financing. *Nature biotechnology* 25(8): 859-866.
6. Collins F. (2010) Has the revolution arrived? *Nature* 464(7289): 674-675.
7. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. (2011) Too many roads not taken. *Nature* 470(7333): 163-165.
8. George Westerman MT, Didier Bonnet, Didier Bonnet, Andrew McAfee. (2012) The Digital Advantage: How Digital Leaders Outperform their Peers in Every Industry. Capgemini Consulting and MIT Center for Digital Business, http://www.capgemini.com/resource-file-access/resource/pdf/The_Digital_Advantage__How_Digital_Leaders_Outperform_their_Peers_in_Every_Industry.pdf.
9. Mahon B, Siegel E, Scientific ICo, Information T, Science ICf, Unesco: Digital Preservation: Information Services and Use. IOS Press, Amsterdam, 2002: 57-59, http://books.google.hu/books?id=vuLIquJ-2pUC.
10. Moore GE. (1965) Cramming More Components onto Integrated Circuits. *Electronics* Vol. 38: 114-117.
11. Mashey JR. (1999) Big Data and the Next Wave of InfraStress. *UseNIX Technical Conference 1999, Monterey, California*, https://www.usenix.org/legacy/events/usenix99/invited_talks/mashey.pdf.

12. IBM. (2013) Investor Briefing 2013. 70, http://www.sec.gov/Archives/edgar/data/51143/000110465913015636/a13-6155_18k.htm.
13. Codd EF. (1970) A relational model of data for large shared data banks. *Commun. ACM* 13(6): 377-387.
14. Luhn HP. (1958) A business intelligence system. *IBM J. Res. Dev.* 2(4): 314-319.
15. Bell G, Hey T, Szalay A. (2009) Beyond the Data Deluge. *Science* 323(5919): 1297-1298.
16. IBM. (2009) Global CIO Study 2009. *IBM Institute for Business Value, C-suite Study Series* (2009): 15, http://www.ibm.com/services/us/cio/ciostudy/pdf/cio_study.pdf.
17. Hey A, Tansley S, Tolle K: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington, 2009: XI-XV, http://research.microsoft.com/en-us/collaboration/fourthparadigm/.
18. Holgate ST. (2008) Pathogenesis of asthma. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 38(6): 872-897.
19. Lemanske RF, Jr., Busse WW. (2010) Asthma: clinical expression and molecular mechanisms. *The Journal of allergy and clinical immunology* 125(2 Suppl 2): S95-102.
20. Pawankar R, Canonica GW, Holgate ST: WAO White book on allergy. World Allergy Organization, 2011: 23-24, http://www.worldallergy.org/definingthespecialty/2011_white_book.php.
21. Pawankar R, Canonica GW, Holgate ST, Lockey RF. (2012) Allergic diseases and asthma: a major global health concern. *Current opinion in allergy and clinical immunology* 12(1): 39-41.
22. Zsigmond G, Novák Z, Berényi K. (2006) Gyermekkori allergiás betegségek nemzetközi epidemiológiai felmérése – az ISAAC-vizsgálat Magyarországon. *Gyermekorvos Továbbképzés* 5: 67-72.
23. Braman SS. (2006) The global burden of asthma. *Chest* 130(1 Suppl): 4S-12S.

24. Palmer LJ, Cookson WO. (2000) Genomic approaches to understanding asthma. *Genome research* 10(9): 1280-1287.

25. Lau S, Illi S, Sommerfeld C, Niggemann B, Bergmann R, von Mutius E, Wahn U. (2000) Early exposure to house-dust mite and cat allergens and development of childhood asthma: a cohort study. Multicentre Allergy Study Group. *Lancet* 356(9239): 1392-1397.

26. Cullinan P, MacNeill SJ, Harris JM, Moffat S, White C, Mills P, Newman Taylor AJ. (2004) Early allergen exposure, skin prick responses, and atopic wheeze at age 5 in English children: a cohort study. *Thorax* 59(10): 855-861.

27. Peden DB. (2005) The epidemiology and genetics of asthma risk associated with air pollution. *The Journal of allergy and clinical immunology* 115(2): 213-219; quiz 220.

28. Bjorksten B. (2004) Effects of intestinal microflora and the environment on the development of asthma and allergy. *Springer seminars in immunopathology* 25(3-4): 257-270.

29. Matricardi PM, Bonini S. (2000) Mimicking microbial 'education' of the immune system: a strategy to revert the epidemic trend of atopy and allergic asthma? *Respiratory research* 1(3): 129-132.

30. Le Souef P. (2000) Infant lung function, bronchial responsiveness and the development of asthma. *Pediatr Allergy Immunol* 11 Suppl 13: 15-18.

31. Rakes GP, Arruda E, Ingram JM, Hoover GE, Zambrano JC, Hayden FG, Platts-Mills TA, Heymann PW. (1999) Rhinovirus and respiratory syncytial virus in wheezing children requiring emergency care. IgE and eosinophil analyses. *American journal of respiratory and critical care medicine* 159(3): 785-790.

32. Johnston SL, Martin RJ. (2005) Chlamydophila pneumoniae and Mycoplasma pneumoniae: a role in asthma pathogenesis? *American journal of respiratory and critical care medicine* 172(9): 1078-1089.

33. Dahl ME, Dabbagh K, Liggitt D, Kim S, Lewis DB. (2004) Viral-induced T helper type 1 responses enhance allergic disease by effects on lung dendritic cells. *Nat Immunol* 5(3): 337-343.

34. Wold AE. (1998) The hygiene hypothesis revised: is the rising frequency of allergy due to changes in the intestinal flora? *Allergy* 53(46 Suppl): 20-25.

35. Camarda LE, Grayson MH. (2011) Can specific IgE discriminate between intrinsic and atopic asthma? *American journal of respiratory and critical care medicine* 184(2): 152-153.

36. Bradding P, Walls AF, Holgate ST. (2006) The role of the mast cell in the pathophysiology of asthma. *The Journal of allergy and clinical immunology* 117(6): 1277-1284.

37. John M, Lim S, Seybold J, Jose P, Robichaud A, O'Connor B, Barnes PJ, Chung KF. (1998) Inhaled corticosteroids increase interleukin-10 but reduce macrophage inflammatory protein-1alpha, granulocyte-macrophage colony-stimulating factor, and interferon-gamma release from alveolar macrophages in asthma. *American journal of respiratory and critical care medicine* 157(1): 256-262.

38. Tang C, Ward C, Reid D, Bish R, O'Byrne P M, Walters EH. (2001) Normally suppressing CD40 coregulatory signals delivered by airway macrophages to TH2 lymphocytes are defective in patients with atopic asthma. *The Journal of allergy and clinical immunology* 107(5): 863-870.

39. Lambrecht BN, De Veerman M, Coyle AJ, Gutierrez-Ramos JC, Thielemans K, Pauwels RA. (2000) Myeloid dendritic cells induce Th2 responses to inhaled antigen, leading to eosinophilic airway inflammation. *The Journal of clinical investigation* 106(4): 551-559.

40. Yukawa T, Read RC, Kroegel C, Rutman A, Chung KF, Wilson R, Cole PJ, Barnes PJ. (1990) The effects of activated eosinophils and neutrophils on guinea pig airway epithelium in vitro. *American journal of respiratory cell and molecular biology* 2(4): 341-353.

41. Vercelli D. (2008) Discovering susceptibility genes for asthma and allergy. *Nature reviews. Immunology* 8(3): 169-182.

42. Bierbaum S, Heinzmann A. (2007) The genetics of bronchial asthma in children. *Respiratory medicine* 101(7): 1369-1375.

43. Finkelman FD, Vercelli D. (2007) Advances in asthma, allergy mechanisms, and genetics in 2006. *The Journal of allergy and clinical immunology* 120(3): 544-550.
44. Weiss ST, Raby BA, Rogers A. (2009) Asthma genetics and genomics 2009. *Curr Opin Genet Dev* 19(3): 279-282.
45. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. (1990) Genetics of asthma and hay fever in Australian twins. *The American review of respiratory disease* 142(6 Pt 1): 1351-1358.
46. Harris JR, Magnus P, Samuelsen SO, Tambs K. (1997) No evidence for effects of family environment on asthma. A retrospective study of Norwegian twins. *American journal of respiratory and critical care medicine* 156(1): 43-49.
47. Koppelman GH, Los H, Postma DS. (1999) Genetic and environment in asthma: the answer of twin studies. *The European respiratory journal* 13(1): 2-4.
48. Nieminen MM, Kaprio J, Koskenvuo M. (1991) A population-based study of bronchial asthma in adult twin pairs. *Chest* 100(1): 70-75.
49. Thomsen SF, van der Sluis S, Kyvik KO, Skytthe A, Backer V. (2010) Estimates of asthma heritability in a large twin sample. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 40(7): 1054-1061.
50. Barnes KC, Marsh DG. (1998) The genetics and complexity of allergy and asthma. *Immunology today* 19(7): 325-332.
51. March ME, Sleiman PM, Hakonarson H. (2011) The genetics of asthma and allergic disorders. *Discovery medicine* 11(56): 35-45.
52. Liggett SB. (1995) Genetics of beta 2-adrenergic receptor variants in asthma. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 25 Suppl 2: 89-94; discussion 95-86.
53. Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, Walsh A, Liu Z, Hayward B, Folz C, Manning SP, Bawa A, Saracino L, Thackston M, Benchekroun Y, Capparell N, Wang M, Adair R, Feng Y, Dubois J, FitzGerald MG, Huang H, Gibson R, Allen KM, Pedan A, Danzig MR, Umland SP, Egan RW, Cuss FM, Rorke S, Clough JB,

Holloway JW, Holgate ST, Keith TP. (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418(6896): 426-430.

54. Basehore MJ, Howard TD, Lange LA, Moore WC, Hawkins GA, Marshik PL, Harkins MS, Meyers DA, Bleecker ER. (2004) A comprehensive evaluation of IL4 variants in ethnically diverse populations: association of total serum IgE levels and asthma in white subjects. *The Journal of allergy and clinical immunology* 114(1): 80-87.
55. Haller G, Torgerson DG, Ober C, Thompson EE. (2009) Sequencing the IL4 locus in African Americans implicates rare noncoding variants in asthma susceptibility. *The Journal of allergy and clinical immunology* 124(6): 1204-1209 e1209.
56. Munthe-Kaas MC, Carlsen KH, Haland G, Devulapalli CS, Gervin K, Egeland T, Carlsen KL, Undlien D. (2008) T cell-specific T-box transcription factor haplotype is associated with allergic asthma in children. *The Journal of allergy and clinical immunology* 121(1): 51-56.
57. Pykalainen M, Kinos R, Valkonen S, Rydman P, Kilpelainen M, Laitinen LA, Karjalainen J, Nieminen M, Hurme M, Kere J, Laitinen T, Lahesmaa R. (2005) Association analysis of common variants of STAT6, GATA3, and STAT4 to asthma and high serum IgE phenotypes. *The Journal of allergy and clinical immunology* 115(1): 80-87.
58. Randolph AG, Lange C, Silverman EK, Lazarus R, Silverman ES, Raby B, Brown A, Ozonoff A, Richter B, Weiss ST. (2004) The IL12B gene is associated with asthma. *American journal of human genetics* 75(4): 709-715.
59. Suttner K, Depner M, Klopp N, Illig T, Vogelberg C, Adamski J, von Mutius E, Kabesch M. (2009) Genetic variants in the GATA3 gene are not associated with asthma and atopic diseases in German children. *The Journal of allergy and clinical immunology* 123(5): 1179-1181.
60. Zhou H, Hong X, Jiang S, Dong H, Xu X. (2009) Analyses of associations between three positionally cloned asthma candidate genes and asthma or asthma-related phenotypes in a Chinese population. *BMC medical genetics* 10: 123.

61. Howard TD, Koppelman GH, Xu J, Zheng SL, Postma DS, Meyers DA, Bleecker ER. (2002) Gene-gene interaction in asthma: IL4RA and IL13 in a Dutch population with asthma. *American journal of human genetics* 70(1): 230-236.

62. Kabesch M, Schedel M, Carr D, Woitsch B, Fritzsch C, Weiland SK, von Mutius E. (2006) IL-4/IL-13 pathway genetics strongly influence serum IgE levels and childhood asthma. *The Journal of allergy and clinical immunology* 117(2): 269-274.

63. Potaczek DP, Okumura K, Nishiyama C. (2009) FCER1A genetic variability and serum IgE levels. *Allergy* 64(9): 1383.

64. Vladich FD, Brazille SM, Stern D, Peck ML, Ghittoni R, Vercelli D. (2005) IL-13 R130Q, a common variant associated with allergy and asthma, enhances effector mechanisms essential for human allergic inflammation. *The Journal of clinical investigation* 115(3): 747-754.

65. Wu H, Romieu I, Shi M, Hancock DB, Li H, Sienra-Monge JJ, Chiu GY, Xu H, del Rio-Navarro BE, London SJ. (2010) Evaluation of candidate genes in a genome-wide association study of childhood asthma in Mexicans. *The Journal of allergy and clinical immunology* 125(2): 321-327 e313.

66. Potter PC, Van Wyk L, Martin M, Lentes KU, Dowdle EB. (1993) Genetic polymorphism of the beta-2 adrenergic receptor in atopic and non-atopic subjects. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 23(10): 874-877.

67. Wills-Karp M, Ewart SL. (2004) Time to draw breath: asthma-susceptibility genes are identified. *Nat Rev Genet* 5(5): 376-387.

68. Zimmermann N, King NE, Laporte J, Yang M, Mishra A, Pope SM, Muntel EE, Witte DP, Pegg AA, Foster PS, Hamid Q, Rothenberg ME. (2003) Dissection of experimental asthma with DNA microarray analysis identifies arginase in asthma pathogenesis. *The Journal of clinical investigation* 111(12): 1863-1874.

69. Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S, Aud D, Peltz G, Wills-Karp M. (2000) Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immunol* 1(3): 221-226.

70. Tolgyesi G, Molnar V, Semsei AF, Kiszel P, Ungvari I, Pocza P, Wiener Z, Komlosi ZI, Kunos L, Galffy G, Losonczy G, Seres I, Falus A, Szalai C. (2009) Gene expression profiling of experimental asthma reveals a possible role of paraoxonase-1 in the disease. *Int Immunol* 21(8): 967-975.

71. Di Valentin E, Crahay C, Garbacki N, Hennuy B, Gueders M, Noel A, Foidart JM, Grooten J, Colige A, Piette J, Cataldo D. (2009) New asthma biomarkers: lessons from murine models of acute and chronic asthma. *Am J Physiol Lung Cell Mol Physiol* 296(2): L185-197.

72. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931-945.

73. Collins FS, Green ED, Guttmacher AE, Guyer MS. (2003) A vision for the future of genomics research. *Nature* 422(6934): 835-847.

74. Allison M. (2012) Direct-to-consumer genomics reinvents itself. *Nature biotechnology* 30(11): 1027-1029.

75. Miller FA, Hayeems RZ, Bytautas JP, Bedard PL, Ernst S, Hirte H, Hotte S, Oza A, Razak A, Welch S, Winquist E, Dancey J, Siu LL. (2014) Testing personalized medicine: patient and physician expectations of next-generation genomic sequencing in late-stage cancer care. *Eur J Hum Genet* 22(3): 391-395.

76. Zhang G, Karns R, Sun G, Indugula SR, Cheng H, Havas-Augustin D, Novokmet N, Durakovic Z, Missoni S, Chakraborty R, Rudan P, Deka R. (2012) Finding missing heritability in less significant Loci and allelic heterogeneity: genetic variation in human height. *PloS one* 7(12): e51211.

77. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M, Magi R, Madden PA, Heath AC, Nyholt DR, Martin NG, Montgomery GW, Frayling TM, Hirschhorn JN, McCarthy MI, Goddard ME, Visscher PM. (2011) Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* 19(7): 807-812.

78. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. (2010) VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11(6): 446-450.

79. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265): 747-753.

80. Ashburn TT, Thor KB. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews. Drug discovery* 3(8): 673-683.

81. Barratt MJ: Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs. John Wiley & Sons Inc., Hoboken, New Jersey, 2012: 9-30.

82. Jarvis LM. (2006) Teaching an old drug new tricks. *Chem Eng News* 84(7): 52-+.

83. Wadman M. (2012) New cures sought from old drugs. *Nature* 490(7418): 15.

84. Arrowsmith J. (2011) TRIAL WATCH Phase III and submission failures: 2007-2010. *Nat Rev Drug Discov* 10(2): 1-1.

85. Novac N. (2013) Challenges and opportunities of drug repositioning. *Trends in pharmacological sciences* 34(5): 267-272.

86. Swan M. (2013) The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. *Big Data* 1(2): 85-99.

87. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. (2013) Computational drug repositioning: from data to therapeutics. *Clin Pharmacol Ther* 93(4): 335-341.

88. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. (2013) Computational Drug Repositioning: From Data to Therapeutics. *Clin Pharmacol Ther* 93(4): 335-341.

89. Li YY, Jones SJM. (2012) Drug repositioning for personalized medicine. *Genome Med* 4.

90. Dudley JT, Deshpande T, Butte AJ. (2011) Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 12(4): 303-311.

91. Sekhon BS. (2013) Repositioning drugs and biologics: retargeting old/existing drugs for potential new therapeutic applications. *Journal of Pharmaceutical Education and Research* 4(1).
92. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A. (2011) Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 12(4): 357-368.
93. Tari L, Vo N, Liang SS, Patel J, Baral C, Cai J. (2012) Identifying Novel Drug Indications through Automated Reasoning. *PloS one* 7(7).
94. Haupt VJ, Schroeder M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Brief Bioinform* 12(4): 312-326.
95. Li HM, Liu AG, Zhao ZJ, Xu YF, Lin JY, Jou D, Li CL. (2011) Fragment-Based Drug Design and Drug Repositioning Using Multiple Ligand Simultaneous Docking (MLSD): Identifying Celecoxib and Template Compounds as Novel Inhibitors of Signal Transducer and Activator of Transcription 3 (STAT3). *J Med Chem* 54(15): 5592-5596.
96. Lussier YA, Chen JL. (2011) The Emergence of Genome-Based Drug Repositioning. *Sci Transl Med* 3(96).
97. Lamb J. (2007) Innovation - The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer* 7(1): 54-60.
98. Cheng FX, Liu C, Jiang J, Lu WQ, Li WH, Liu GX, Zhou WX, Huang J, Tang Y. (2012) Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *Plos Comput Biol* 8(5).
99. Lee HS, Bae T, Lee JH, Kim DG, Oh YS, Jang Y, Kim JT, Lee JJ, Innocenti A, Supuran CT, Chen L, Rho K, Kim S. (2012) Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC systems biology* 6: 80.
100. Sukumar N, Krein MP. (2012) Graphs and networks in chemical and biological informatics: past, present and future. *Future Med Chem* 4(16): 2039-2047.
101. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. (2008) Drug target identification using side-effect similarity. *Science* 321(5886): 263-266.

102. Watson JD, Crick FH. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356): 737-738.
103. (2010) Human genome at ten: The sequence explosion. *Nature* 464(7289): 670-671.
104. Carlson RH: Biology Is Technology: The Promise, Peril, and New Business of Engineering Life. Harvard University Press, Cambridge, MA, US, 2011: 108-130, http://books.google.hu/books?id=HLpxDwEACAAJ.
105. Stein LD. (2003) Integrating biological databases. *Nat Rev Genet* 4(5): 337-345.
106. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56-65.
107. Edgar R, Domrachev M, Lash AE. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1): 207-210.
108. Berman HM. (2008) The Protein Data Bank: a historical perspective. *Acta crystallographica. Section A, Foundations of crystallography* 64(Pt 1): 88-95.
109. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 39(Database issue): D876-882.
110. Sherry ST, Ward M, Sirotkin K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research* 9(8): 677-679.
111. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G,

Moutsianas L, Nguyen H, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311): 52-58.

112. (2008) The Gene Ontology project in 2008. *Nucleic Acids Res* 36(Database issue): D440-444.
113. Brown FK. (1998) Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annual Reports in Medicinal Chemistry* 33(375): 33:375-384.
114. Waller CL, Shah A, Nolte M. (2007) Strategies to support drug discovery through integration of systems and data. *Drug Discov Today* 12(15-16): 634-639.
115. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP. (2011) The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med* 3(80): 80ps16.
116. Swamidass SJ. (2011) Mining small-molecule screens to repurpose drugs. *Brief Bioinform* 12(4): 327-335.
117. Wang YL, Xiao JW, Suzek TO, Zhang J, Wang JY, Bryant SH. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* 37: W623-W633.
118. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Research* 36: D351-D359.
119. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* 40(D1): D1100-D1107.

120. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34: D668-D672.

121. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36: D901-D906.

122. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795): 1929-1935.

123. Aidinis V, Chandras C, Manoloukos M, Thanassopoulou A, Kranidioti K, Armaka M, Douni E, Kontoyiannis DL, Zouberakis M, Kollias G, Consortium MN. (2008) MUGEN mouse database; Animal models of human immunological diseases. *Nucleic Acids Research* 36: D1048-D1054.

124. Ringwald M, Iyer V, Mason JC, Stone KR, Tadepally HD, Kadin JA, Bult CJ, Eppig JT, Oakley DJ, Briois S, Stupka E, Maselli V, Smedley D, Liu SY, Hansen J, Baldock R, Hicks GG, Skarnes WC. (2011) The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium. *Nucleic Acids Research* 39: D849-D855.

125. Swanson DR. (1990) Medical Literature as a Potential Source of New Knowledge. *B Med Libr Assoc* 78(1): 29-37.

126. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. (2010) Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases. *Plos Comput Biol* 6(9).

127. Chiang AP, Butte AJ. (2009) Systematic Evaluation of Drug-Disease Relationships to Identify Leads for Novel Drug Uses. *Clin Pharmacol Ther* 86(5): 507-510.

128. Macdonald RR. (2004) Statistical inference and Aristotle's Rhetoric. *The British journal of mathematical and statistical psychology* 57(Pt 2): 193-203.

129. Moore JH, Asselbergs FW, Williams SM. (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26(4): 445-455.

130. Pearl J. (1985) Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*: 329-334, http://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf.

131. Antal P, Fannes G, Timmerman D, Moreau Y, De Moor B. (2003) Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial intelligence in medicine* 29(1-2): 39-60.

132. Antal P, Gézsi A, Hullám G, Millinghoffer A. (2006) Learning Complex Bayesian Network Features for Classification. *Proc. of third European Workshop on Probabilistic Graphical Models*: 9-16, http://www.utia.cas.cz/files/mtr/pgm06/37_paper.pdf.

133. Ungvari I, Hullam G, Antal P, Kiszel PS, Gezsi A, Hadadi E, Virag V, Hajos G, Millinghoffer A, Nagy A, Kiss A, Semsei AF, Temesi G, Melegh B, Kisfali P, Szell M, Bikov A, Galffy G, Tamasi L, Falus A, Szalai C. (2012) Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. *PloS one* 7(3): e33573.

134. Lautner-Csorba O, Gezsi A, Semsei AF, Antal P, Erdelyi DJ, Schermann G, Kutszegi N, Csordas K, Hegyi M, Kovacs G, Falus A, Szalai C. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. *BMC medical genomics* 5: 42.

135. Lautner-Csorba O, Gezsi A, Erdelyi DJ, Hullam G, Antal P, Semsei AF, Kutszegi N, Kovacs G, Falus A, Szalai C. (2013) Roles of genetic polymorphisms in the folate pathway in childhood acute lymphoblastic leukemia evaluated by Bayesian relevance and effect size analysis. *PloS one* 8(8): e69843.

136. Jobbagy-Ovari G, Paska C, Stiedl P, Trimmel B, Hontvari D, Soos B, Hermann P, Toth Z, Kerekes-Mathe B, Nagy D, Szanto I, Nagy A, Martonosi M, Nagy K, Hadadi E, Szalai C, Hullam G, Temesi G, Antal P, Varga G, Tarjan I. (2014)

Complex analysis of multiple single nucleotide polymorphisms as putative risk factors of tooth agenesis in the Hungarian population. *Acta odontologica Scandinavica* 72(3): 216-227.

137. Hullam G, Gezsi A, Millinghoffer A, Sarkozy P, Bolgar B, Srivastava SK, Pal Z, Buzas EI, Antal P. (2014) Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis. *Methods Mol Biol* 1142: 143-176.
138. Juhasz G, Hullam G, Eszlari N, Gonda X, Antal P, Anderson IM, Hokfelt TG, Deakin JF, Bagdy G. (2014) Brain galanin system genes interact with life stresses in depression-related phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* 111(16): E1666-1673.
139. Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD. (2010) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions. *Journal of Machine Learning Research* 11: 235-284.
140. Stephens M, Balding DJ. (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10(10): 681-690.
141. Hullam G, Juhasz G, Bagdy G, Antal P. (2012) Beyond structural equation modeling: model properties and effect size from a Bayesian viewpoint. An example of complex phenotype-genotype associations in depression. *Neuropsychopharmacol Hung* 14(4): 273-284.
142. Johnson M, Lajiness M, Maggiora G. (1989) Molecular similarity: a basis for designing drug screening programs. *Progress in clinical and biological research* 291: 167-171.
143. Li YY, An JH, Jones SJM. (2011) A Computational Approach to Finding Novel Targets for Existing Drugs. *Plos Comput Biol* 7(9).
144. Overington JP, Al-Lazikani B, Hopkins AL. (2006) Opinion - How many drug targets are there? *Nat Rev Drug Discov* 5(12): 993-996.
145. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KLH,

Edwards DD, Shoichet BK, Roth BL. (2009) Predicting new molecular targets for known drugs. *Nature* 462(7270): 175-U148.

146. Dubus E, Ijjaali I, Barberan O, Petitet F. (2009) Drug repositioning using in silico compound profiling. *Future Med Chem* 1(9): 1723-1736.
147. Achenbach J, Tiikkainen P, Franke L, Proschak E. (2011) Computational tools for polypharmacology and repurposing. *Future Med Chem* 3(8): 961-968.
148. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. (2006) New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model* 46(2): 462-470.
149. Willett P, Barnard JM, Downs GM. (1998) Chemical similarity searching. *J Chem Inf Comp Sci* 38(6): 983-996.
150. Arany A, Bolgar B, Balogh B, Antal P, Matyus P. (2013) Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources. *Curr Med Chem* 20(1): 95-107.
151. Eckert H, Bojorath J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* 12(5-6): 225-233.
152. Svensson F, Karlen A, Skold C. (2012) Virtual Screening Data Fusion Using Both Structure- and Ligand-Based Methods. *J Chem Inf Model* 52(1): 225-232.
153. Kubinyi H. (2003) Drug research: myths, hype and reality. *Nat Rev Drug Discov* 2(8): 665-668.
154. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15545-15550.
155. Varin T, Schuffenhauer A, Ertl P, Renner S. (2011) Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J Chem Inf Model* 51(7): 1528-1538.

156. Varin T, Gubler H, Parker CN, Zhang JH, Raman P, Ertl P, Schuffenhauer A. (2010) Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data. *J Chem Inf Model* 50(12): 2067-2078.

157. Temesi G, Virag V, Hadadi E, Ungvari I, Fodor LE, Bikov A, Nagy A, Galffy G, Tamasi L, Horvath I, Kiss A, Hullam G, Gezsi A, Sarkozy P, Antal P, Buzas E, Szalai C. (2014) Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations. *Allergy, asthma & immunology research* 6(6): 496-503.

158. Ungvari I, Hadadi E, Virag V, Bikov A, Nagy A, Semsei AF, Galffy G, Tamasi L, Horvath I, Szalai C. (2012) Implication of BIRC5 in asthma pathogenesis. *Int Immunol* 24(5): 293-301.

159. Gergely T. (2007) Információ menedzsment és intelligens adatelemzés az SNP analízis támogatására (Mérnök Informatikus MSc szakdolgozat). *Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszék*.

160. Gergely T. (2008) In Silico farmakogenomika, Egy integrált bioinformatikai platform felhasználása a személyre szabott, genom alapú terápiák szolgálatában (Egészségügyi Mérnök MSc szakdolgozat). *Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszék*.

161. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* 32(Database issue): D493-496.

162. Tsicopoulos A, de Nadai P, Glineur C. (2013) Environmental and genetic contribution in airway epithelial barrier in asthma pathogenesis. *Current opinion in allergy and clinical immunology* 13(5): 495-499.

163. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2): 263-265.

164. Friedman N, Koller D. (2003) Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Mach Learn* 50(1-2): 95-125.

165. Cooper GF, Herskovits E. (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. *Mach. Learn.* 9(4): 309-347.

166. Hullám G, Antal P. (2013) The effect of parameter priors on Bayesian relevance and effect size measures. *Electrical Engineering and Computer Science* 57(2): 35-48.

167. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: Bayesian data analysis. CRC press, Boca Raton, Florida, 2013: 290-318.

168. Giudici P, Castelo R. (2003) Improving Markov Chain Monte Carlo Model Search for Data Mining. *Mach Learn* 50(1-2): 127-158.

169. Antal P, Millinghoffer A, Hullám G, Szalai Cs FA. (2008) A Bayesian view of challenges in feature selection: multilevel analysis, feature aggregation, multiple targets, redundancy and interaction. *JMLR Workshop and Conference Proceedings* 4: 74-89.

170. Péter Antal AM, Gábor Hullám, Gergely Hajós, Péter Sárközy, András Gézsi, Csaba Szalai, András Falus: Probabilistic Graphical Models for Genetics, Genomics and Postgenomics. Oxford University Press, Oxford, 2014: 318-352.

171. Bolgar B, Arany A, Temesi G, Balogh B, Antal P, Matyus P. (2013) Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. *Current topics in medicinal chemistry* 13(18): 2337-2363.

172. Hofmann T, Scholkopf B, Smola A. (2008) Kernel methods in machine learning. *Annals of Statistics* 36(3): 1171-1220.

173. Temesi G, Bolgar B, Arany A, Szalai C, Antal P, Matyus P. (2014) Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. *Future Med Chem* 6(5): 563-575.

174. Baeza-Yates RA, Ribeiro-Neto B: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., New York, 1999: 29-30.

175. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. (2010) A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 6: 343.

176. Stojmirovic A, Yu YK. (2010) Robust and accurate data enrichment statistics via distribution function of sum of weights. *Bioinformatics* 26(21): 2752-2759.

177. Tsitsiou E, Williams AE, Moschos SA, Patel K, Rossios C, Jiang X, Adams OD, Macedo P, Booton R, Gibeon D, Chung KF, Lindsay MA. (2012) Transcriptome analysis shows activation of circulating CD8+ T cells in patients with severe asthma. *The Journal of allergy and clinical immunology* 129(1): 95-103.

178. Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK. (2006) Global similarity and local divergence in human and mouse gene co-expression networks. *BMC evolutionary biology* 6: 70.

179. Kuhn A, Luthi-Carter R, Delorenzi M. (2008) Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC bioinformatics* 9: 26.

180. Strand AD, Aragaki AK, Baquet ZC, Hodges A, Cunningham P, Holmans P, Jones KR, Jones L, Kooperberg C, Olson JM. (2007) Conservation of regional gene expression in mouse and human brain. *PLoS genetics* 3(4): e59.

181. Song KH, Chiang JY. (2006) Glucagon and cAMP inhibit cholesterol 7alpha-hydroxylase (CYP7A1) gene expression in human hepatocytes: discordant regulation of bile acid synthesis and gluconeogenesis. *Hepatology* 43(1): 117-125.

182. Jordan IK, Marino-Ramirez L, Koonin EV. (2005) Evolutionary significance of gene expression divergence. *Gene* 345(1): 119-126.

183. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution* 21(11): 2058-2070.

184. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 99(7): 4465-4470.

185. Yanai I, Graur D, Ophir R. (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics : a journal of integrative biology* 8(1): 15-24.

186. Ehre C, Rossi AH, Abdullah LH, De Pestel K, Hill S, Olsen JC, Davis CW. (2005) Barrier role of actin filaments in regulated mucin secretion from airway goblet cells. *American journal of physiology. Cell physiology* 288(1): C46-56.

187. Davis CW, Dickey BF. (2008) Regulated airway goblet cell mucin secretion. *Annual review of physiology* 70: 487-512.

188. Bush WS, McCauley JL, DeJager PL, Dudek SM, Hafler DA, Gibson RA, Matthews PM, Kappos L, Naegelin Y, Polman CH, Hauser SL, Oksenberg J, Haines JL, Ritchie MD. (2011) A knowledge-driven interaction analysis reveals potential neurodegenerative mechanism of multiple sclerosis susceptibility. *Genes and immunity* 12(5): 335-340.

189. Kouznetsova I, Chwieralski CE, Balder R, Hinz M, Braun A, Krug N, Hoffmann W. (2007) Induced trefoil factor family 1 expression by trans-differentiating Clara cells in a murine asthma model. *American journal of respiratory cell and molecular biology* 36(3): 286-295.

190. Park SJ, Lee YC. (2008) Peroxisome proliferator-activated receptor gamma as a novel therapeutic target in asthma. *The Journal of asthma : official journal of the Association for the Care of Asthma* 45(1): 1-8.

191. Lee KS, Park SJ, Hwang PH, Yi HK, Song CH, Chai OH, Kim JS, Lee MK, Lee YC. (2005) PPAR-gamma modulates allergic inflammation through up-regulation of PTEN. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 19(8): 1033-1035.

192. Lee SH, Jang AS, Woo Park S, Park JS, Kim YK, Uh ST, Kim YH, Chung IY, Park BL, Shin HD, Park CS. (2011) Genetic effect of single-nucleotide polymorphisms in the PPARGC1B gene on airway hyperreactivity in asthmatic patients. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* 41(11): 1533-1544.

193. Tolgyesi G, Keszei M, Ungvari I, Nagy A, Falus A, Szalai C. (2006) Involvement of TNFalpha -308A promoter polymorphism in the development of asthma in children infected with Chlamydophila pneumoniae. *Pediatric research* 60(5): 543-548.

194. Ungvari I, Tolgyesi G, Semsei AF, Nagy A, Radosits K, Keszei M, Kozma GT, Falus A, Szalai C. (2007) CCR5 Delta 32 mutation, Mycoplasma pneumoniae infection, and asthma. *The Journal of allergy and clinical immunology* 119(6): 1545-1547.

195. Nagy A, Kozma GT, Keszei M, Treszl A, Falus A, Szalai C. (2003) The development of asthma in children infected with Chlamydia pneumoniae is dependent on the modifying effect of mannose-binding lectin. *The Journal of allergy and clinical immunology* 112(4): 729-734.

196. Nagy A, Keszei M, Kis Z, Budai I, Tolgyesi G, Ungvari I, Falus A, Szalai C. (2007) Chlamydophila pneumoniae infection status is dependent on the subtypes of asthma and allergy. *Allergy and asthma proceedings : the official journal of regional and state allergy societies* 28(1): 58-63.

197. Pemberton AD, Rose-Zerilli MJ, Holloway JW, Gray RD, Holgate ST. (2008) A single-nucleotide polymorphism in intelectin 1 is associated with increased asthma risk. *The Journal of allergy and clinical immunology* 122(5): 1033-1034.

198. Edgington LE, Verdoes M, Ortega A, Withana NP, Lee J, Syed S, Bachmann MH, Blum G, Bogyo M. (2013) Functional Imaging of Legumain in Cancer Using a New Quenched Activity-Based Probe. *J Am Chem Soc* 135(1): 174-182.

199. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, Greco D. (2013) Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics* 5(1): 30.

200. Gottlieb A, Stein GY, Ruppin E, Sharan R. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 7: 496.

201. Eichler HG, Oye K, Baird LG, Abadie E, Brown J, Drum CL, Ferguson J, Garner S, Honig P, Hukkelhoven M, Lim JC, Lim R, Lumpkin MM, Neil G, O'Rourke B, Pezalla E, Shoda D, Seyfert-Margolis V, Sigal EV, Sobotka J, Tan D, Unger TF, Hirsch G. (2012) Adaptive licensing: taking the next step in the evolution of drug approval. *Clin Pharmacol Ther* 91(3): 426-437.

202. Dolgin E. (2014) Adaptive methods help drug sponsors find best treatment dose. *Nature medicine* 20(4): 321.

203. Wason J, Marshall A, Dunn J, Stein RC, Stallard N. (2014) Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *British journal of cancer* 110(8): 1950-1957.

11 Saját publikációk jegyzéke

11.1 Asztma genetika

Temesi G, Virág V, Hadadi É, Ungvári I, Fodor LE, Bikov A, Nagy A, Galffy G, Tamási L, Horváth I, Kiss A, Hullám G, Gézsi A, Sárközy P, Antal P, Buzás E, Szalai C: Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations. Allergy, asthma & immunology research 6(6), 496-503 (2014), IF: 3.084*

Ungvári I, Hullám G, Antal P, Kiszel Sz P, Gézsi A, Hadadi É, Virág V, Hajós G, Millinghoffer A, Nagy A, Kiss A, Semsei F Á, Temesi G, Melegh B, Kisfali P, Széll M, Bikov A, Gálffy G, Tamási L, Falus A, Szalai C: Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. PloS one 7(3), e33573 (2012), IF: 3.730

11.2 Feldúsulás elemzés

Temesi G, Bolgár B, Arany Á, Szalai C, Antal P, Mátyus P: Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. Future Med Chem 6(5), 563-575 (2014), IF: 4.000*

Bolgár B, Arany Á, Temesi G, Balogh B, Antal P, Mátyus P: Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. Current topics in medicinal chemistry 13(18), 2337-2363 (2013), IF: 3.453

11.3 Egyéb közlemény

Gabriella Jobbágy-Óvári, Csilla Páska, Péter Stiedl, Bálint Trimmel, Dorina Hontvári, Borbála Soós, Péter Hermann, Zsuzsanna Tóth, Bernadette Kerekes-Máthé, Dávid Nagy, Ildikó Szántó, Ákos Nagy, Mihály Martonosi, Katalin Nagy, Éva Hadadi, Csaba Szalai, Gábor Hullám, Gergely Temesi, Péter Antal, Gábor Varga, Ildikó Tarján: Complex analysis of multiple single nucleotide polymorphisms as putative risk factors of tooth agenesis in the Hungarian population. ACTA ODONTOLOGICA SCANDINAVICA 72:(3) pp. 216-227. (2014), IF: 1.309*

DOI:10.14753/SE.2015.1757

12 Köszönetnyilvánítás

Szeretnék köszönetet mondani Dr. Szalai Csaba témavezetőmnek és Dr. Antal Péter külső konzulensemnek, akik a doktori értekezésemen túl is számos területen és projektben mentoraim hosszú évek óta mind szakmailag, mind emberileg.

Szeretnék köszönetet mondani volt és jelenlegi vezetőimnek, kollégáimnak, akik szakmai és baráti segítsége nélkül a munkám nem jöhetett volna létre. Így a Semmelweis Egyetem Genetikai, Sejt- és Immunbiológiai Intézet vezetőinek Prof. Dr. Falus Andrásnak és Prof. Dr. Búzás Editnek, akiktől bizalmat és lehetőséget kaptam munkámhoz; Dr. Ungvári Ildikónak az önzetlen szakmai és baráti segítségéért; Dr. Félné Semsei Ágnesnek, Hadadi Évának és Virág Viktornak, akik munkájukkal megalapozták kutatásaimat. Illetve a Budapesti Műszaki és Gazdaságtudományi Egyetem Méréstechnika és Információs Rendszerek Tanszék vezetőjének, Dr. Jobbágy Ákosnak, aki biztosította a feltételeket a bioinformatika csoportunk működéséhez; kollégáimnak Hullám Gábornak, Gézsi Andrásnak, Sárközy Péternek, Millinghoffer Andrásnak, Huszár Balázsnak, Bolgár Bencének, Arany Ádámnak, Marx Péternek és Hajós Gergelynek, akiknek sokszínű munkája, szakmai és baráti támogatása az értekezésem számos részletében visszaköszön.

Szeretnék köszönetet mondani biológus Édesanyámnak és mérnök Édesapámnak, akik úgy ültették el bennem-e két fantasztikus tudomány terület magvait, hogy nem tudtam választani és hálával tartozom, amiért több szülői gondoskodást, családi meleget, anyagi hátteret és szellemi támogatást biztosítottak a munkámhoz, mint amennyit bárki kívánhatna.