Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Dr. Bruncsics Bence

Nagyléptékű szemantikus adat és tudásintegráció az öregedéskutatásban

Örökifjaktól állatmodelleken át gyógyszerekig

KONZULENSEK

Dr. Antal Péter

Dr. Gézsi András

BUDAPEST, 2017

1 TARTALOMJEGYZÉK

2 BEVEZETÉS

A biológiai és orvosi területek tudományos eredményeinek egyre jelentősebb része hozzáférhető nyilvános, szemantikus formában. Az adatbázistechnológiák, a meglévő annotációs és ontológiai adatbázisok konvertálása és a szemantikus publikálás fejlődése mind ezen kapcsolt, nyílt adatok (Linked Open Data, LOD) egyre fokozódó jelentőségét erősíti.

Az elérhető hatalmas mennyiségű adatnak a hasznosságát árnyalja, hogy az információ rendkívül heterogén, amely számítási szempontból jelent kihívásokat, illetve ez a hatalmas relációs hálózat egyenetlenül van feltérképezve, valamint az információ megbízhatósága sem követhető minden esetben megfelelően.

A kutatás során egy adat- és tudásintegrációs módszer optimalizálását és valós felhasználását vizsgáltam meg, amely a kapcsolt, nyílt adatokból egy hatékony lekérdezést és értelmezést segítő számítási hálózatot hoz létre. A fúziós rendszer elsődleges célja a különböző tárgyterületeken átívelő lekérdezések támogatása, nevezetesen a humángenetikai, -genomikai, kemoinformatikai területekről és kísérletes állatmodellekből származó adatok és információrészletek integrálása. A rendelkezésünkre álló adatok jelentős része az orvosbiológiát és a kemoinformatikát forradalmasító nagy áteresztő képességű vizsgálatokból származik, ezért nagy hangsúlyt fektettem az ilyen jellegű bizonytalan, nagy dimenziójú evidenciák közvetlen feldolgozására és integrációjára.

A kutatás alkalmazási területe az egészséges öregedés, mivel az öregedés egy összetett, rendkívül sok betegséget is érintő folyamat, emiatt az automatizált tudásintegráció nagy segítséget jelenthet a bizonytalan adatok kezelésében, és a szétszórt különböző tudományterületekről származó, változatos jellegű és gyakran eltérő fajokból származó adatok integrálásában. A téma jelentőségét fokozza, hogy az öregedéskutatás, különös tekintettel az egészséges öregedés vizsgálata, biológiai, társadalmi és gazdasági szempontból is az egyik legfontosabb kutatási terület, az öregedésnek a modern társadalomra kifejtett sokrétű hatása miatt.

A kutatásomban összegyűjtöttem és integrációra alkalmas formára alakítottam a jelenleg elérhető öregedéssel kapcsolatos információk jelentős részét. Így létrehoztam több öregedéssel kapcsolatos modellt, amelyek több tárgyterületet fognak át, mint például az

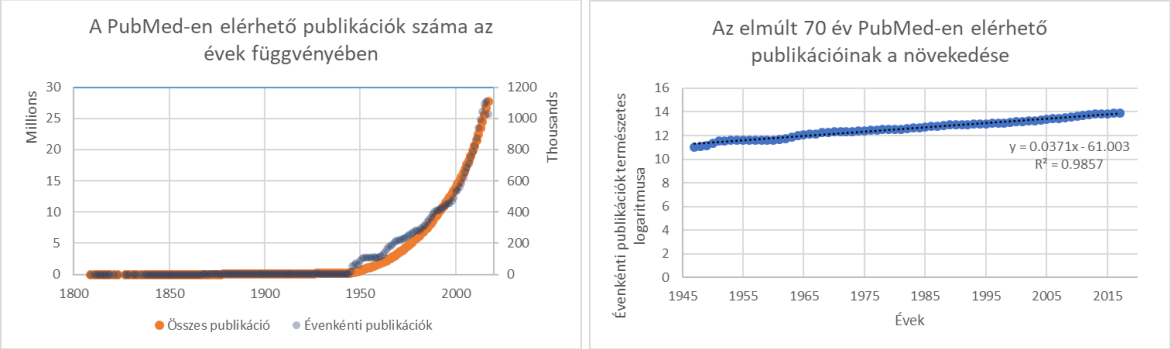
egészséges öregedéshez és a matuzsálemi életkorhoz kapcsolódó genetikai variánsokat, géneket, valamint a gyógyszereket, hatóanyagokat és az állatmodellekből származó géneket és biológiai útvonalakat.

A feldolgozott adatok között kiemelt fontosságúak az egészséges öregedés genetikai hátterének vizsgálatából származó adatok, amelyek statisztikai elemzését replikáltuk és kiegészítettük, valamint a modellállatokon végzett kísérletekből származó genetikai információk és ezek emberi megfelelői. Ezeken kívül felhasználtam a szakirodalomban fellelhető öregedésre és élethosszra ható gyógyszerhatóanyagok listáját. A modell ezeken felül integrálja az öregedéshez kapcsolt molekuláris jelutakat, és képes a génekre aggregáltan kezelni az öregkorra jellemző fizikai megjelenési jegyeket is, mint őszülés vagy hallásromlás.

3 ELŐZMÉNYEK

3.1 A TUDOMÁNYOS ADATOK GYARAPODÁSA

Becslések szerint az évenként publikált tudományos cikkek száma 3%-kal gyarapodik évente, és 2009-ben elérte az évenkénti másfél milliós publikációt, összesen 50 millió addig közölt cikkel (1). Jelenleg 2017-ben ez a szám 64 millió köré tehető, amelyből a PubMed-en elérhető tartalom 27 millió közleményt tesz ki, így az itt elérhető élettudományhoz kapcsolt cikkek képezik a tudományos közlemények legnagyobb csoportját (2). Ezek az adatok is jelzik, hogy a szakcikkek száma már jelentősen meghaladja azt a mértéket, amelyet egy ember fel képes dolgozni. Ha azt nézzük, hogy az élettudományok egy részterületével, az öregedéssel kapcsolatban hány cikk születik évente, akkor azt látjuk, hogy 2016-ban több mint 24,000 cikk született a témában és a trend gyorsan növekvő. Ha az összes cikket át szeretnénk olvasni ebben a témában, akkor egy embernek 7 perce lenne minden értekezésre a munkaidejében.



1. ábra) A PubMed-en elérhető publikációk számának az alakulása

A tudományos közlemények hatalmas száma mellett az is nehezíti az információ feldolgozását, hogy a szerzők az eredményeket szöveg és a szövegben szereplő állításokat támogató ábrák, táblázatok formájában teszik közzé, melyeknek az integrálása több problémába ütközik. A cikkek jelentős részének csak az összefoglalója érhető el nyilvánosan, és ahol elérhető a teljes közlemény ott sem lehet egységesített formátumokra számítani a szöveg, ábrák vagy táblázatok terén. Továbbá a szöveg is a természetes nyelvnek megfelelő, bár az már segítséget jelent, hogy az értekezések többsége angol nyelven íródik, így felmerül a szövegbányászati módszerek használata a probléma kapcsán. Bár a szövegbányászat jelenleg is használatos az irodalom feldolgozásában, de szerepe korlátozott maradt és csupán

támogató szerepet lát el egy sokkal egységesebb, formális tudásreprezentációkon nyugvó publikációs gyakorlattal szemben.

Az elmúlt évtizedben több próbálkozás is volt arra, hogy létrehozzanak egy olyan egységes publikálási formát, ahol az információ könnyen, gyorsan és egyértelműen tárolható, kezelhető és lekérdezhető. Ezzel kapcsolatban több adatbázis és ezeket leíró és egységesíteni próbáló séma született, ezek közül a relációs adatbázisok és ezek lekérdezési nyelvére épülő rendszerek a legszélesebb körben használatosak az élettudományok terén, ám kötöttségeik miatt a használhatóságuk korlátozott. Rugalmasabb megközelítések érdekében több rendszert is kidolgoztak, mint a USQL (Unified Service Query Language) alapó PYRAMID-S, vagy a XQuery alapú elképzelések, de a legsikeresebb iránynak a szemantikus web szabályrendszerét követő RDF (Resource Description Framework) alapú megközelítések bizonyultak (3).

3.2 SZEMANTIKUS VILÁGHÁLÓ

A szemantikus világháló („web”) az 1960-as évektől fejlesztett szemantikus hálózatok elképzelésén alapszik (4), és a 2001-es kidolgozása Tim Berners-Lee nevéhez köthető (5, 6). A szemantikus web (SW) egy olyan általános keretrendszert biztosít, mely lehetővé teszi az adatok megosztását és újra használását alkalmazások, vállalkozások és közösségek között. Az SW RDF-en alapuló nagy számú kutatót, intézetet és ipari céget magába foglaló közös erőfeszítés a W3C (World Wide Web Consortium) irányait követve (7). A szemantikus web leegyszerűsítve a gyakorlatban egy hiperhivatkozás hálózat, mely akár az emberek által is olvasható oldalak kiegészítése számítógépek által olvasható metaadatokkal, melyek az oldalak tartalmát és egymással való kapcsolatát írják le (8).

Az elképzelés első 5 évében igen kevés népszerűségre tett szert és használata is igen limitált volt, olyan szinten, hogy Berners-Lee kollégáival azt nyilatkozták, hogy annak ellenére, hogy milyen „egyszerű elképzelés, mégis többnyire megvalósítatlan maradt” (8). Ehhez képest jelenleg a szemantikus formátumot biztosító schema.org több mint 12 millió oldallal áll kapcsolatban, ami valójában jelentős része a 2017-ben jelen levő több mint 1.8 milliárd weboldalnak mivel ezek közül mindössze 170 millió oldal aktív (9, 10).

Az elképzelés megvalósítását különböző nyelvek teszik lehetővé, mint az OWL (Web Ontology Language), az XML (Extensible Markup Language) és a kiemelten fontos RDF (Resource

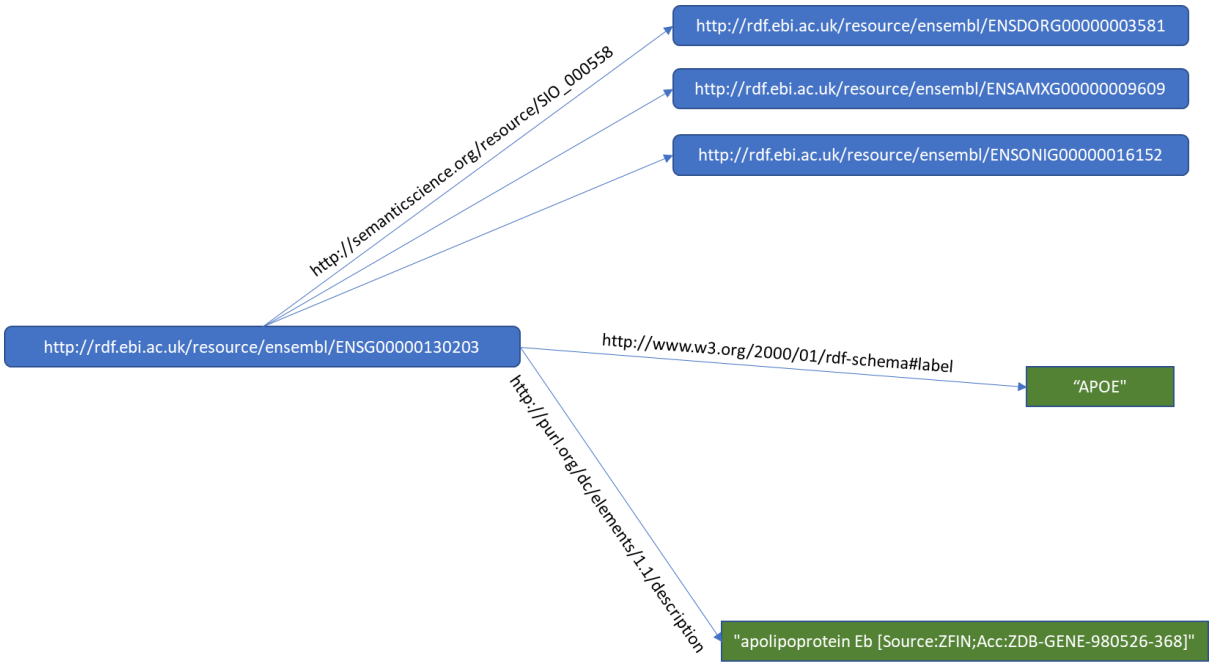
Description Framework). Az XML egy ma már széles körben használt általános célú leírónyelv (Markup language) mellyel speciális célú leíró nyelveket lehet létrehozni, például tárgyterületek fogalmi viszonyainak általános leírására. Ennek támogatására vált megszokottá az ontológia technológiai alkalmazása, amelynek klasszikus területe a létfilozófia, azaz a létezéssel és általános létezési kategóriákkal foglalkozik. Modern informatikai alkalmazása a webre vonatkoztatva eredményezte a webet érintő ontológiát és egy nyelvét az OWL-t. Ez utóbbi arra szolgál, hogy leírjuk vele a web dokumentumaiban és alkalmazásaiban előforduló lényeges és jellemző osztályokat és a közöttük lévő kapcsolatokat (11).

3.2.1 Erőforrás Leíró Nyelv (RDF)

A W3C standardjai közül az RDF a legmeghatározóbb, mivel a szemantikus web nagy mértékben erre az egyszerű megközelítésre és annak szabályrendszerére épül. Az RDF, magyarul Erőforrás Leíró Nyelv, egy adatleíró nyelv, amellyel erőforrásokról szóló információkat ábrázolhatunk a weben (12). A kifejlesztésének elsődleges célja az erőforrások (többnyire weboldalak) metaadatainak az ábrázolása volt, mint például egy oldal címe, szerzője, létrehozásának vagy módosításának az ideje, de akár lehet ez áruk specifikációja vagy ára is. Az RDF tervezésének az egyik mozgatóereje az volt, hogy az ilyen metaadatok ne csak emberek számára legyenek elérhetőek, hanem számítógéppel is feldolgozhatóvá váljanak, és ezen adatokat torzulásmentesen általánosan felhasználhatóvá váljanak. Az RDF egyik alappillére az erőforrások URI-vel (Uniform Resource Identifier) avagy egységes erőforrás-azonosítóval történő azonosítása és egyszerű tulajdonságokkal történő leírása. Az URI tekinthető a weblapoknál megszokott URL általános változatának, ami ellenben az URL-lel, ami egy webforrás elsődleges hozzáférési mechanizmusát (hálózati címét) ábrázolja, az URI nincsenek arra korlátozva, hogy csak olyan dolgot azonosítson, ami hálózati címmel rendelkezik.

Az URI-k univerzalitása lehetővé teszi, hogy ezeket az információkat gráfként ábrázoljuk, ahol a gráf csomópontjai és élei erőforrások, azok tulajdonságai vagy a tulajdonságainak a leírása. Az RDF gráf megértésében segít az alábbi ábra [2. ábra], ahol az APOE génhez tartozó RDF gráfjának egy kis része látható, itt kék színnel gének URI-ei szerepelnek, míg zölddel a gén tulajdonságai és az ezeket összekötő kapcsolatok, melyek szintén URI-vel rendelkeznek. Innen látható, hogy a hiperhivatkozások nem csak értékekre mutatnak, hanem további

hivatkozásokra is, így akár hivatkozásokon keresztül az összes erőforrás kapcsolatban állhat egymással, ami az ismert gének esetében meg is történik.



2. ábra) Az APOE gén RDF gráfjának egy kis részlete

Ez az elképzelés megfelel Tim Berners-Lee által felvetett GGG (Giant Global Graph) Gigantikus Globális Gráfnak, amely jól leírja ezt a megközelítést, annak ellenére, hogy a kifejezés nem örvend nagy népszerűségnek (13). Viszont az RDF értelmezésére létezik egy másik lehetőség is, ahol az egyes információkat hármasokba, tripletekbe osztjuk alany-állítmány-tárgy (subject–predicate–object) felépítéssel. Ebben az esetben a „label” avagy felirat élő nyelvet használva úgy néz ki, hogy a ENSG00000130203-nek (gén ID) van felirata, aminek „APOE” az értéke. Ugyanez számítógép által is értelmezhető nyelven az alábbi hármas:

Alany (subject)	Állítmány (predicate)	Tárgy (object)
http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203	http://www.w3.org/2000/01/rdf-schema#label	„APOE”
Az APOE gén URI-je	URI, ami valaminek a feliratára, nevére utal	Az APOE gén Neve

A szemantikus webet alkotó ilyen hármasok szerializálására (lejegyzésére) több W3C által elfogadott standard is létezik, mint a Turtle, N-Triples, N-Quads, JSON-LD, Notation3 és az RDF/XML ami az első szabvány volt az RDF szerializálására (8).

3.2.2 SPARQL Protokoll és RDF Lekérdezési Nyelv

Az RDF formában található adatok lekérdezésére kidolgozott nyelv a SPARQL (SPARQL Protocol and RDF Query Language), ami a SPARQL Protokoll és RDF Lekérdezési Nyelv rövidítése. A SPARQL segítségével igen változatos, RDF alapú vagy RDF formátumra hozható adatbázisokban végezhetünk el viszonylag egyszerűen igen bonyolult lekérdezéseket és nyerhetünk ki kívánt adatokat, továbbá az adatbázis módosítására is lehetőséget ad. A lekérdezések tartalmazhatnak szükséges és opcionális gráf mintázatokat, továbbá a lekérdezés során tetszőleges konjunkciók és diszjunkciók is végezhetők (14).

Egy egyszerű SPARQL lekérdezés például úgy zajlik, hogy az alany-állítmány-tárgy hármasból tetszőlegesen megadunk értékeket (például a http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203 URI-t) , vagy egy ? kezdetű változónevet (?allitmany), és ennek alapján a lekérdezés megkeresi az erre a mintára épülő találatokat, és ha a SELECT parancs után áll a változó neve, akkor ki is jelzi a találatoknak megfelelő változó értékét.

```
SELECT ?allitmany
WHERE {
 <http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203> ?allitmany ?targy .
 }
group by ?allitmany
```

Ez a kód az APOE gén URI-jének megfelelő állítmányokat adja vissza, ahogyan az alábbi néhány példa is ezt mutatja: dcterms:identifier, dc:identifier, rdfs:label, <http://rdf.ebi.ac.uk/terms/ensembl/DEPENDENT>, <http://rdf.ebi.ac.uk/terms/ensembl/DIRECT>, <http://biohackathon.org/resource/faldo#location>. Ezek a találatok lehetnek hagyományos URI formában, vagy a gyakran használt prefix-es rövidítésként, ahol például a rdfs:label a http://www.w3.org/2000/01/rdf-schema#label rövidítése.

```
SELECT ?targy
WHERE {
 <http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203> <http://www.w3.org/2000/01/rdf-schema#label> ?targy .
}
```

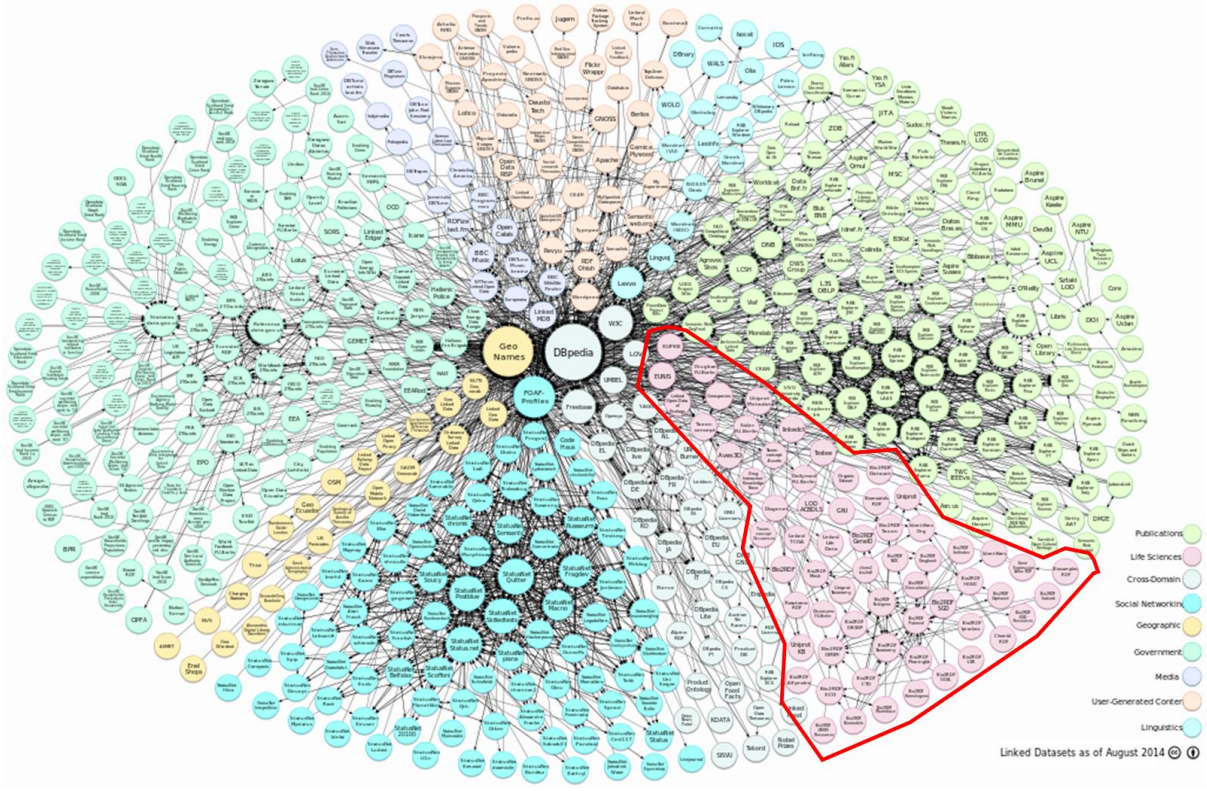
Amennyiben arra vagyunk kíváncsiak, hogy ez a címke milyen értéket vesz fel, akkor azt a fenti kóddal kaphatjuk, meg, ahol az eredmény ebben az esetben „APOE” szöveg lesz. Természetesen mivel egy állítás tárgya egy másik állítás alanya is lehet, ezért a változók segítségével igen összetett lekérdezéseket is elvégezhetünk.

3.3 KAPCSOLT ADAT

A Kapcsolt Adat (Linked Data) egy módszer vagy gyakorlat a strukturált adatok, adatbázisok publikálására. Az információ szolgáltatók ezzel a módszerrel összekapcsolják adatbázisaikat, és a szemantikus módszerekkel és lekérdezéssel gazdagodva előnyösebb helyzetbe kerülnek (15). Az elképzelés a webes standardokon nyugszik, mint HTTP, RDF és URI-k, de itt ezek már nem a hagyományos emberi felhasználásra szánt oldalakat szolgálnak ki, hanem kiterjeszti a gépek számára is automatikusan olvasható formátummá.

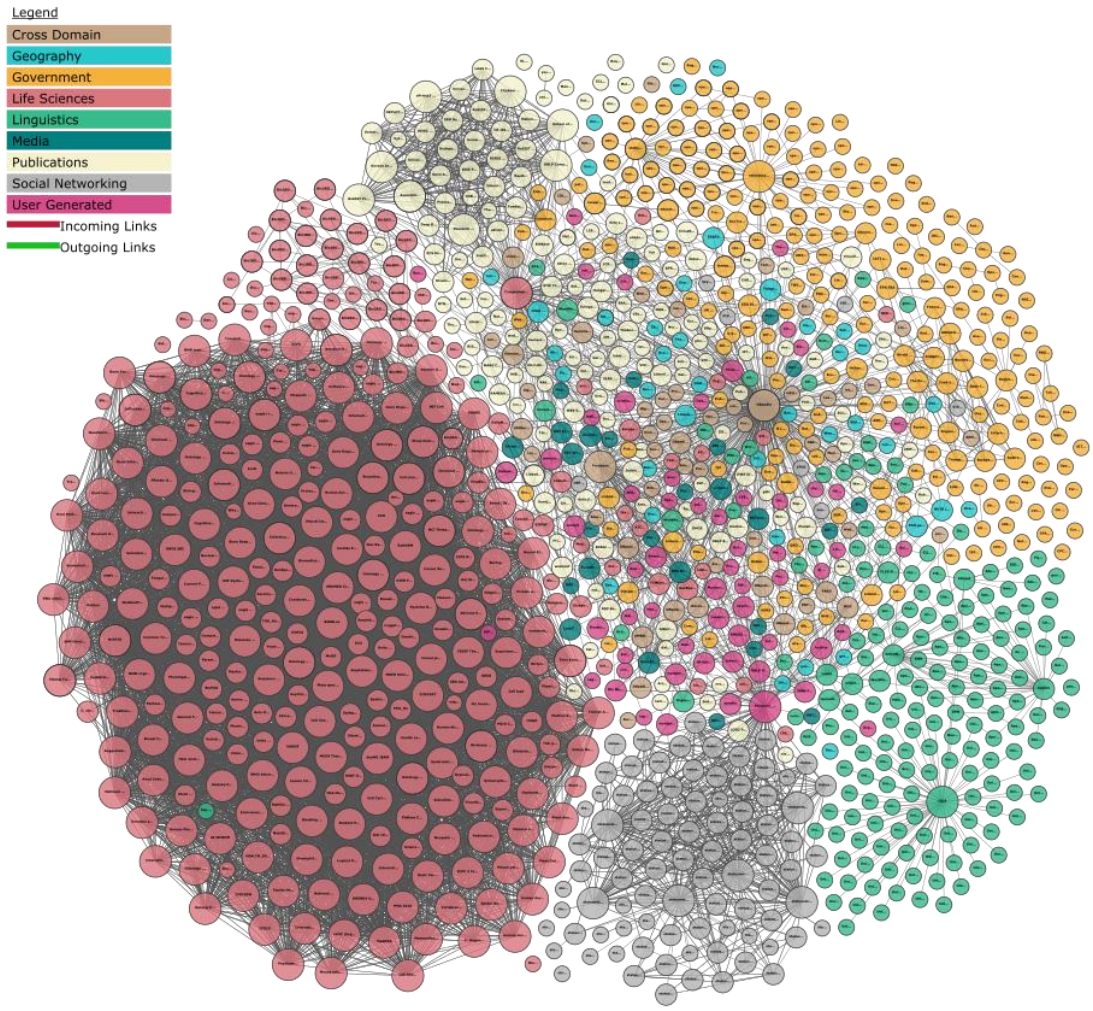
A Kapcsolt Adat elképzeléshez egy további tényező is társul, ami pedig az adatok nyíltsága, azaz nyilvános elérhetősége. Jelenleg nem minden adat, adatbázis érhető el nyilvánosan mindenki számára, viszont számos nyílt forrás is létezik, és a Kapcsolt Nyílt Adat (Linked Open Data, LOD) ezen két területet ötvözi, amelyet a W3C közösség folyamatosan gondoz és bővít.

A Kapcsolt nyílt adat formátumban elérhető adatbázisokat az alábbi ábra [3. ábra] mutatja, ahol látható, hogy 2014—ben az igazgatással kapcsolatban található a legtöbb csomópont, ezt követve közleményekkel és közösségi hálózatokkal kapcsolatos csomópontok, és csak ezeket követte az élettudomány.



3. ábra) A Kapcsolt Nyílt Adat gráfja 2014-ben

Ezzel szemben jelenleg 2017-ben ez az arány teljesen eltolódott az élettudományok irányába, ahogy az alábbi ábrán [4. ábra] is látható. A szemantikus webnek egy iránya a meglevő adatbázisok RDF formára hozatala és már RDF formában elérhető adatbázisok összekapcsolása egymással, amely folyamatnak a Kapcsolt Nyílt Adat közössége az egyik fő mozgatóereje.



4. ábra) A Kapcsolt Nyílt Adat gráfja 2017 augusztusában (16)

A Kapcsolt Adat 3 alapelvre épül, amelyeket követve tetszőleges adatbázisok is kapcsolódhatnak ehhez a nyílt közösséghez.

- Az első alapelv az, hogy minden dolgot, adatot HTTP kezdetű URI-vel (korábbi elképzelés szerint URL-lel) azonosítunk, vagyis nevezzünk el. Példa rá az APOE gén, és a hozzá kapcsolt adatok a http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203 URI alatt érhetők el.

- A második az, hogy ezt a HTTP kezdetű nevet (URI/URL) elérve standardizált formátumú adatot kapjunk vissza, ami az adott dologgal kapcsolatban tartalmaz információt. Példa rá a 3.2.2 részben található első lekérdezés és az eredménye.
- A harmadik pedig az, hogy az elért adatok a kapcsolataikat szintén az előbb ismertetett formában hivatkozzák meg úgy, hogy ezek az információk is URI-kre mutassanak. Példa rá az 2. ábra, ahol az APOE-hez kapcsolható ortológ (ugyanannak a génnek más fajban megtalálható változata) gének szintén megegyező formátumú URI-vel rendelkeznek.

Ezen alapelveket követve rengeteg meglévő adatbázist csatlakoztattak a közösséghez, mint *DBPedia* 9 milliárd Wikipédiáról származtatható adattal, *GeoWordNet* 50 millió földrajzi adattal*, Europeana Linked Open Data V1.0* 100 milliós adattal, ezek közül csak néhányat említve. Az elképzelés remek alapot adott biológiai és kémiai adatok és ismeretek összefogására és publikálására is, olyannyira, hogy ezen területek exponenciális fejlődése miatt jelenleg már ezek alkotják a Kapcsolt Adatok meghatározó hányadát.

3.4 KAPCSOLT ADAT AZ ÉLETTUDOMÁNYOKBAN

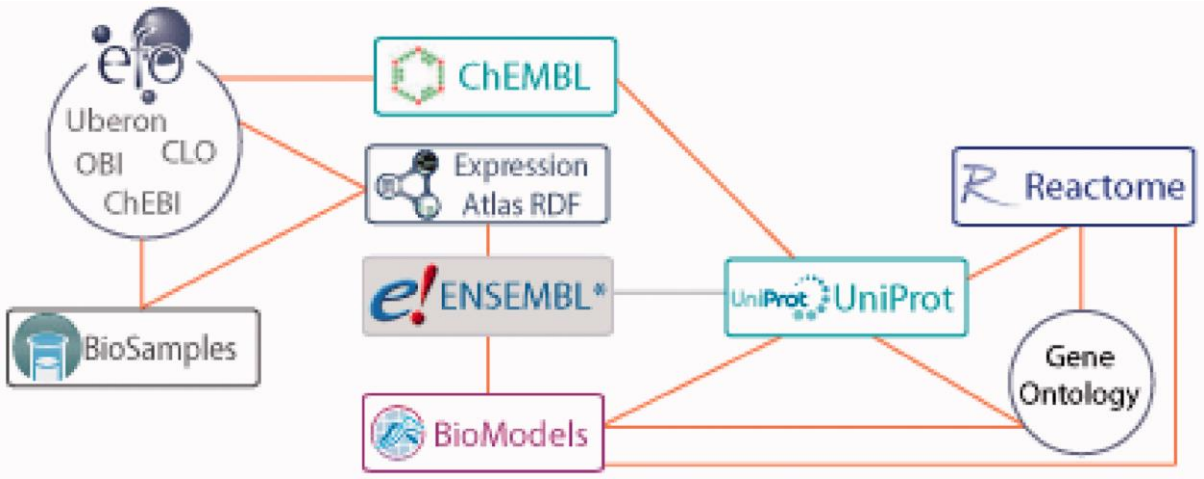
Az élettudományok térfoglalása a Kapcsolt Adatban több dolognak is köszönhető. Egyrészt a biológiai és kémiai kutatást forradalmasító nagy áteresztő képességű vizsgálatoknak, mivel ezen módszerek jól kezelhető, hatalmas mennyiségű digitális adatot eredményeznek. Másrészt a jelentős mennyiségű meglevő adatot hatékonyan csak informatikai módszerekkel lehet feldolgozni.

Ezen hatások következtében több adatbázis tért át saját erőből szemantikus formára és számos összefogás született a meglevő adatbázisok egységes átalakítására, melyek közül kiemelkedő munkát végzett a *Bio2RDF* projekt, 35 adatbázis Kapcsolt Adat standardjainak megfelelő hozatalával, mely több mint 10 milliárd adathármast érint (17). További jelentős hozzájáruló a témában a jelenleg is aktív, számos szolgáltatással és adattal rendelkező *EMBL-EBI.*

3.4.1 Európai Bioinformatikai Intézet (EBI)

EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute) Az Európai Molekuláris Biológia Laboratóriumának az Európai Bioinformatika Intézete, amely

számos biológiai adathoz, bioinformatikai alkalmazáshoz biztosít hozzáférést. Amellett, hogy 724 webes szolgáltatást biztosít, számos adatbázist foglal magában és fenntart egy saját RDF alapú platformot EBI RDF platform néven (18).



5. ábra) Az EBI RDF platform felépítése (18)

Az EBI RDF platformjából többek közt [5. ábra] a projektünkben két kiemelkedő adatbázist használunk fel, az ENSEMBL és ChEMBL szemantikus adatait. Az ENSEMBL több mint 1 millió gént tartalmaz, és elérhetőek a génekhez tartozó nevek, leírások, faji, faj és fajközi adatok (19). A ChEMBL 1,5 millió vegyületet, és ehhez tartozó számos adatot tartalmaz, továbbá 11 ezer vegyület célpontot és a köztük levő kapcsolati hálót is magában foglalja.

3.4.2 Ontobee

Egy másik Kapcsolt Adathoz tartozó RDF szerver az Ontobee, aminek a célja biológiai ontológiák megosztása, vizualizációja, lekérdezésének a támogatása, integrálása és elemzése. Jelenleg 189 adatbázist foglal magában 3 milliót meghaladó adathármassal, és ezek közül egy a HPO (human phenotype ontology), amelyek leírják az emberi általános fenotípusokat vagyis megfigyelhető jegyeket, mint magasság, őszülés vagy nagyothallás.

3.4.3 Releváns egyedi adatbázisok

Jelenleg már megszokott, hogy az új, jelentős adatbázisok kapcsoltan is elérhetőek, mint például a DisGeNet, amely szemantikus formában elérhetővé teszi az emberi betegségek genetikai alapját egy gén-betegség asszociációs hálózat segítségével (20). Az adatbázis RDF-é alakított formában DisGeNET RDF néven érhető el, és a kapcsoltságon túl tartalmaz

információt az asszociáció erősségére, forrására és az esetlegesen érintett variánsokra vonatkozóan (21).

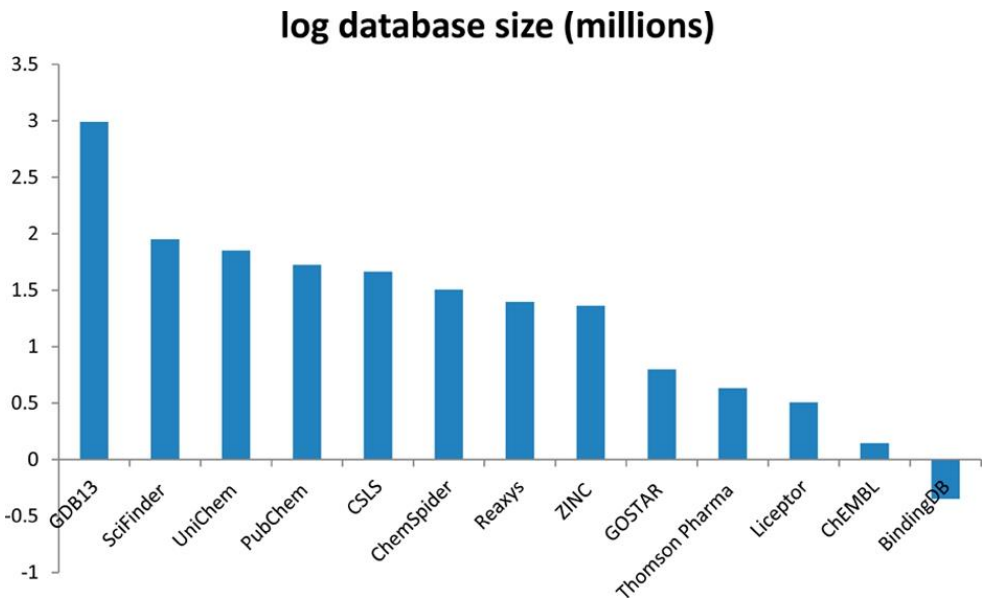
Egy további fontos szemantikus adatbázis a WikiPathways, ami az egyes folyamatokért felelős jelutakat és az abban érintett fehérjéket és géneket írja le (22).

3.4.4 Gene Ontology

A Gene Ontology a Gene Ontology Consortium által vezetett gén-ontológiai kapcsolatok megnevezésére kialakított és fenntartott nevezéktan és erre épülő adatbázis, melyre több webes szolgáltatás is épül, többek között az AmiGO 2 mely segítségével szabadon böngészhető az adatbázis. Fontos megjegyezni, hogy ez a szemantikusan elérhető ontológiai adatok legnagyobb és legjobban lefedett forrása és számos szolgáltatás, köztük az általunk használt modell egy része alapszik az ontológiai hasonlósággal végzett számításokon.

3.4.5 Kémiai adatbázisok

A Kapcsolt Adaton belül kiemelkedő jelentősége van a kémiai-bioaktivitási adatoknak, részben azért mert az egyik legdinamikusabban gyarapodó tudáshalmaz, másrészt mivel kiemelkedő jelentősége van az egészségügyben, de akár az élelmiszer- és kozmetikai iparban is.



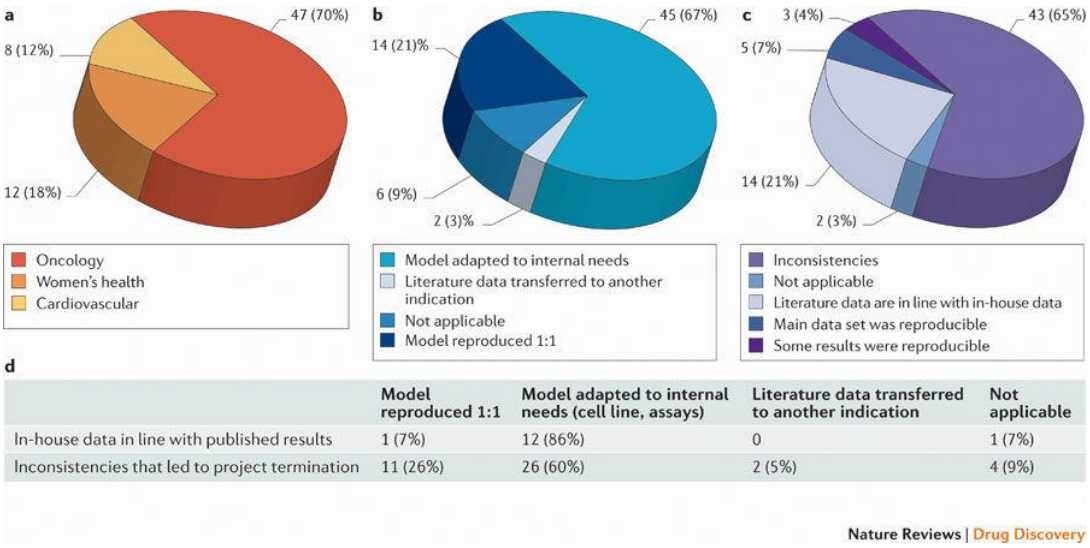
6. ábra) A kémiai adatbázisok adatainak a száma (milliók logaritmusa) 2014-ben (23)

A [6. ábra] felső ábrán látható, hogy több adatbázis is százmilliós adattal rendelkezik a kémiai-farmakológiai területeken, viszont fontos megjegyezni, hogy több probléma is felmerül

ezekkel az adatokkal kapcsolatban, mint például az, hogy nem minden adat származik valós mérésből, sőt a vegyületek egy része nem is létező, hanem igény esetén a jövőben megszintetizált molekulák (23). További ismert problémát jelent az is, hogy az adatbázisok kitöltöttsége és feltérképezettsége rendkívül hiányos. Ez úgy is érvényes, hogy 2017. szeptemberében a PubChem RDF adatbázisa 100 milliárd adathármast foglal magában (24).

3.5 AZ ADATOK MEGBÍZHATÓSÁGA

A nagy áteresztőképességű módszerek térhódítása felveti a kérdést, hogy mennyire megbízhatóak ezek az adatok összességében és a hagyományos mérésekhez hasonlítva. A témában több cikk is született, amelyek egyik alapvető állítása az, hogy az irodalomban elérhető közlemények és állítások többsége elméletileg is levezethetően téves (25, 26). Ezek szerint a kísérleti módszer és az állítás várható értéke határozza meg alapjában, hogy mekkora egy kísérlet hamis találati aránya, ami epidemiológiai kísérletek esetében akár 80% is lehet, míg jól kivitelezett kettős vak kísérlet vagy egy megfelelő metaanalízis esetében 15% köré tehető. Továbbá, mivel a publikált eredmények többsége nem megfelelően követhető forrásból vagy kisebb megbízhatóságú kísérleti felállásból származik, ezért érthető a címben is szereplő állítás miszerint a „Legtöbb publikált kutatási eredmény hibás” (25). A helyzeten az segíthet, hogy a kísérletek megismétlése jelentősen javítja ezt az tendenciát, viszont a sajnálatos módon a mai napig nagyon nehéz forrást találni a tudományos eredmények ismétlésére, valamint az ilyen eredmények tudományos megítélése is sokkal alacsonyabb, mint ami indokolt lenne.



	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

7. ábra) A Bayer gyógyszercég által megismételt kutatások reprodukálhatósága

A kutatási eredmények egy legkritikusabb felhasználója, aki alapvetően érdekelt a tudományos eredmények validitásában, az a gyógyszeripar és gyógyszerkutatás. 2011-ben a Bayer gyógyszercég 67 tudományos közleményben szereplő adatot ismételt meg onkológiai, keringési és női egészség témában [7. ábra/a] (27). A kísérletek egyötödét egy az egyben ismételték meg, míg kétharmadát kisebb módosításokkal [7. ábra/b] és azt találták, hogy az adatok mindössze 21-32 százaléka volt teljesen vagy részben összhangban az eredményekkel, míg az esetek kétharmadában a találatok nem egyeztek az ismételt eredményekkel [7. ábra/c].

Ezek alapján látható, hogy a hagyományos publikációk sem jelentenek abszolút megbízható információforrást, sőt, a nagy áteresztő képességű módszereknél, ha megfelelően lettek kivitelezve, akkor bizonyos torzításokat a statisztikai elemzés során detektálhatunk. Erre példa a teljes genomikai szélességű vizsgálatoknál (genome-wide association study, GWAS)-nál elérhető mutatók, melyek bizonyos problémákra utalhatnak, ilyen a lambda inflációs érték, a Hamis Találati Arányszám és a q-érték (28). Továbbá, ha célpontot vagy magyarázatot keresünk és tudjuk, hogy a találatainkat más forrásból is meg fogjuk erősíteni, akkor 90-95%-os hamis találati arány is használható, mert ebben az esetben minden 10-20.-adik találat valós információt tartalmaz, amit a későbbiekben tovább tudunk pontosítani.

3.6 Korábbi megközelítések génprioritizálásra

Jelenleg számos eszköz áll rendelkezésre génprioritizálásra, ezek többsége szemantikusan elérhető genetikai és ontológiai adatokat dolgoz fel, egyes információ típus („tengely”) mentén. Például a prioritizálók egy csoportja betegséggel és fenotípussal kapcsolatba hozható géneket képes rangsorolni gyakran az OMIM (Online Mendelian Inheritance in Man) adatok felhasználásával mint, a Phenolyzer (29), S2G (Syndrome to gene) (30), HEGPEC (31) vagy a POCUS (32).

Más rendszerek tisztán genetikai adatbázisokra helyezik a hangsúlyt, mint az expressziós vagy GWAS adatok mint a DAWN (33), esetlegesen szövegbányászati módszerekkel kiegészítve, mint a MetaRanker 2.0 (34).

A számos módszer közül az általánosan elfogadottabb, és népszerűbb prioritizálók egyike az Endevaour (35), ami OMIM, és útvonal adatforrásokat használ a genetikai expressziós, GWAS, szekvencia hasonlósági, fehérje és ontológiai adatokon kívül, és a több adatforrást integráló

értékeket képes jelölt géneknek adni. További jól teljesítő prioritizáló a ToppGene (36), mely a korábbiakat kiegészítve már hatóanyag információkat is képes kezelni. Ezen kívül a Chem2Bio2RDF (37) emelendő ki, mely annak ellenére, hogy jelenleg nem aktív, a szemantikus neten integrált több adatbázist és két entitást az őket összekötő elérési utak száma és jellege alapján prioritizált.

Az elérhető prioritizálók közös jegye, hogy egy adott, többnyire betegség listához képes jelölt géneket találni vagy rangsorolni, többnyire egy tengely felhasználásával. Ezzel szemben a módszerek többsége nem engedi meg az adatbázisaik közti szabad átjárást és lekérdezést ezzel a szemantikus web jelentős előnyétől esnek el. Továbbá pár kivétellel csak egy vagy két (betegség és ontológia) irányból közelítik meg a kérdést ezzel potenciális információforrásoktól esnek el.

Összességében megállapítható, hogy jelenleg nem elérhető olyan génprioritizáló rendszer, mely tetszőleges információ forrásba tartozó bemenetet és kimenetet is tartalmaz. Ezen kívül a prioritizálók további általános hiányossága, hogy kvantitatív adatok nem nyerhetők a segítségükkel.

3.7 Az öregedés

A kapcsolt élettudományi adatok jelentősége ott tud igazán megnyilvánulni, ahol több dimenziós, multifaktoriális összetett folyamatot vizsgálunk, mivel ezekben az esetekben a különböző információ források egyedi torzításai kiküszöbölik egymást. Ezen szempontoknak megfelel az öregedés folyamata.

Az öregedés az idő nagyobb léptékű múlására bekövetkezett komplex biológiai változás és többnyire a kronológiai korral kapcsoljuk össze az emberek vonatkozásában. Az egyedfejlődés összemosódhat az öregedéssel, különösen az angol forrásokban, viszont mi az öregedés alatt a jellemzően 65 éves kor felett megjelenő jegyeket és az ezekhez vezető folyamatot értjük.

Az öregedés biológiailag a fiziológiai egység folyamatos passzív hanyatlásából fakadó csökkent funkcionalitás és fokozott kitettség a halálra (38). Ez a folyamat több betegséggel is összefügg, amelyeket tágan korfüggő betegségeknek (age-associated/related disease) nevezünk. Ezek fő csoportjai a szív és érrendszer (cardivascular) és az agyi érrendszer (cerebrovascular) betegségei; idegrendszer degeneratív (neurodegeneratív) betegségei; a rákos és tumoros

elváltozások; cukorbetegség; krónikus obstruktív tüdőbetegség; csontok és ízületek sorvadása; az érzékszervek betegségei, mint a szürkehályog, időskori makuladegeneráció és a nagyothallás. Ezen betegségek közül a dolgozatban példáként használom az időskori makuladegenerációt.

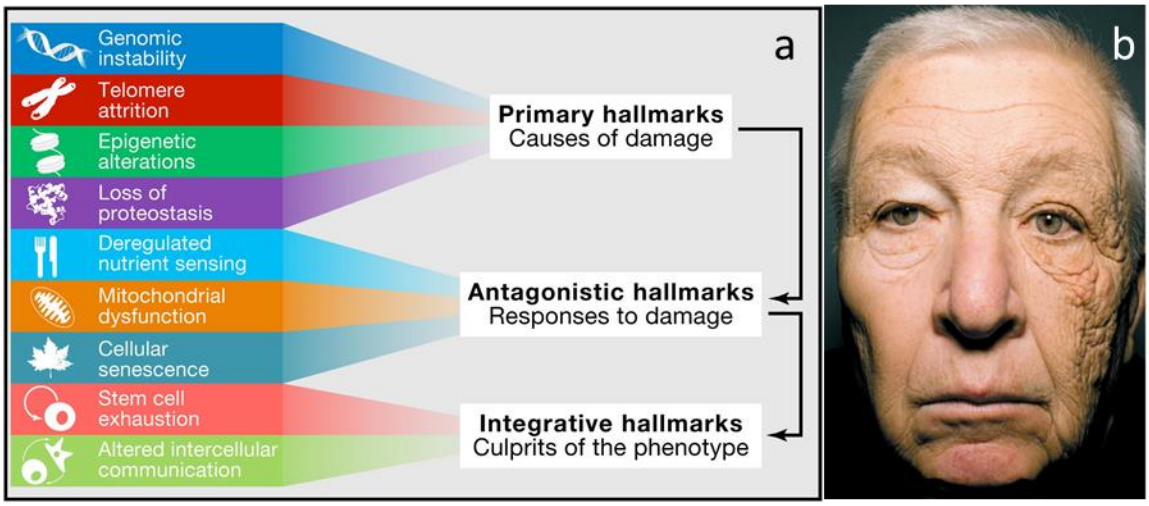
A makuladegeneráció az ideghártyának az éleslátásért felelős helyét érintő károsodása. A kialakulásának a fő oka, hogy az ideghártyát ellátó erek hálózata megváltozik, és érburjánzás vagy vérzés formájában károsítja a látást. Ez öregkorra jellemző, és a kialakult vakságok egyharmadáért felelős a fejlett országokban.

3.7.1 Az öregedés evolúciós szemmel

Látván az emberi öregedés hatását és negatív oldalait, illetve az élővilágban megfigyelhető különböző sebességű és karakterisztikájú öregedési folyamatokat, sőt halhatatlan fajokat, mint a hydrák, és a medúzák, felmerül a kérdés, hogy az öregedés jelensége miért alakulhatott ki evolúciósan. Ez a kérdés igen komplex, és nincs is teljes egyetértés a témában, de pár gondolat kiemelhető. Az első az, hogy amíg az életkor átlaga nem éri el az öregedéssel kapcsolt hanyatlás szintjét, addig nem rendelkezik evolúciós hatással, és mivel az ősember átlagéletkorát 20-35 évre teszik, ezért ez megfelel ennek az elképzelésnek. Egy további sokat vitatott elképzelés, hogy ha a nagyobb genetikai változatossággal rendelkező utódoknak nem kell versenyezniük a szüleikkel az erőforrásokért, akkor ez a faj dinamikusabban képes adaptálódni, ezzel evolúciós előnyhöz juttatva az öregedő fajokat.

3.7.2 Az öregedés biológiája

Az öregedés egy jelenleg még kevéssé értett folyamat, viszont több elképzelés és részfolyamat ismert, amivel részben magyarázható a kialakulása, lefolyása és a karakterisztikája. Az egyik legelfogadottabb összefoglaló publikáció szerint négy fő biológiai tényezője, dimenziója, avagy az angolban elterjedt megnevezés szerint „tengelye” van az öregedésnek [8. ábra/a], ezek a genom módosulása vagy instabilitása, a telomer rövidülés, az epigenetikai módosulások a DNS-ben és a fehérjék hajtogatódásának a zavarai (38). Ezek további öregedést elősegítő folyamatokra hatnak, mint az tápanyag érzékelése (ide tartozik a cukorbetegség), a mitokondriumok funkcióvesztése és a sejtek öregedése.



8. ábra) a) Az öregedés tengelyei (38) b) Az UV-sugárzás DNS károsító hatása bőrre

A genom, ami a DNS sorrendjében kódolt információ, több módon is változhat életünk során elsősorban mutációk hatására. Ezek lehetnek kémiai karcinogének, vagy fizikai sugárzások, sőt természetes folyamatok is kiválthatják. Egy jó illusztráció arra, hogy egy kevéssé ártalmas sugárzásnak, mint az UV sugárzásnak való kitétel mennyire hasonló jegyeket produkál, mint az öregedés az a [8. ábra/B] fenti ábrán látható, ahol egy kamionsofőr egy oldalú napsugárzásnak volt kitéve évekig. További DNS-t érintő kórkép a Werner betegség, ami részben a DNS stabilitásáért felelős gén hibás funkciója miatt alakul ki, és felgyorsult öregedést és átlagosan 30 évvel korábbi halálozást okoz.

A telomer régió a kromoszómák végén található viszonylag rövid rész, és a sejtek osztódása során folyamatosan rövidül egy telomeráz enzim hiányában, ami az emberi testi sejtekből hiányzik. A rövidülés miatt a DNS instabillá válik, és néhány azon régióban kódolt információ el is veszhet, ami szintén a fentiekhez hasonló problémákhoz vezet.

A genetikai információ továbbá nem csak a DNS sorrendben lehet, hanem a DNS elérhetőségének a szabályozásában is, amit epigenetikának hívunk, és a feladata a nem kívánt DNS részek elnémítása, és ez a DNS-en kialakult csendesítő mintaként fogható fel. Viszont ez a funkció igen komplex szabályozású, amely a szöveti differenciálódásért is felel, így ennek értelmezése jelenleg is nyílt kutatási kérdés.

További hatás a gének által kódolt fehérjék struktúráját és funkcióit érinti, ami a fehérje élettartama alatt megváltozhat, és speciális fehérjék felelősek a megváltozott fehérjék korrekciójáért vagy lebontásáért. Öregedéssel ez a funkció károsodhat, vagy változhat, és akár

egy fehérje változása is komoly betegségeket okozhat, ha nincsen korrigálva (például a többi fehérjét is saját mintájára tudja formálni, mint a szivacsos agysorvadás esetében).

3.7.3 Élettartam és egészségtartam

Annak ellenére, hogy az alapvető folyamatok ismertek az öregedés kapcsán, részletes információval egyik tengelyről sem rendelkezünk, ezért további összegző és magyarázatgeneráló kutatásokra van szükség a témában. Erre egy jó módszer az élettartam vizsgálat mellett az egészségtartam feltérképezése is, mert bár nagy vonalakban hasonlóak, és befolyással is bírnak egymásra mégis több szempontból különböznek.

Az egészségtartam az egészségben, komolyabb krónikus betegség nélkül eltöltött évek száma, és kizáró betegségek listájával lehet definiálni (39). Az egészségtartam kutatásának fontos szerepe van abban, hogy gazdaságilag és erkölcsileg is jobban támogatható módon próbáljuk hasznos évekhez juttatni az embereket, elkerülve az élet idejének akár ágyhoz kötve eltöltött meghosszabbításának a problémáit. Azonban egészségtartam vizsgálatával kapcsolatban egyenlőre csupán gén asszociációs vizsgálatok jelentek meg még csak, mint nagy léptékű vizsgálatok (39, 40).

4 A KVANTITATÍV SZEMANTIKUS PRIORITIZÁLÓ RENDSZER

A szemantikus web egy természetes korlátja, hogy a lekérdezéséhez ismernünk kell a SPARQL nyelvet és ebben az esetben is csak szemantikai adatokat, kapcsolatokat kapunk vissza, amely önmagában nem alkalmas kvantitatív vagy szemikvantitatív információ kinyerésére. Ennek az áthidalására több megközelítés is lehetséges, mint például a Chem2Bio2RDF által követett módszer, amely két adat közti elérési utak statisztikájából von le következtetéseket a kapcsolat erősségére vonatkozóan (37). További módszer, hogy külön forrásból származó információkhoz hasonlítjuk a bemeneteket vagy találatokat, és az ezekhez számított értékeket használjuk fel a modellen belüli értékek vagy rangsorok meghatározására. Ezen elképzelést valósítja meg automatikusan több korábbi ajánló rendszer, mint például az Endeavour (35).

Ezzel szemben az MIT tanszék Computational Biomedicine (ComBine) csoportja által fejlesztett és általam tesztelt Kvantitatív Szemantikai Fúziós Prioritizáló (QSFP) rendszere kvantitatív alapú, amely hasonlóságok és relevanciák kombinálását teszi lehetővé nagy léptékű Kapcsolt Adatból (Linked open Data, LOD) származtatott következtetési hálózatok segítségével.

4.1 A KVANTITATÍV SZEMANTIKUS PRIORITIZÁLÓ KERETRENDSZERE

A Kvantitatív Szemantikus Prioritizáló egy keretrendszer, amely lehetővé teszi, hogy RDF alapú adatbázisok között gyors kapcsolati és hasonlósági számításokat végezzen úgy, hogy az adatforrásokat tetszőlegesen bővíthetjük, és köztük tetszőleges kapcsolatokat hozhatunk létre.

A keretrendszer automatikusan lekérdezhető és frissíthető offline adatokat használ a gyorsabb számítás érdekében, mely megfelelően méretezett erőforrás esetén teljes mértékben betölthető a memóriába. A felhasznált adatbázisok folyamatosan bővíthetők, és jelenleg 13 kategóriába eső információ forrás használható stabilan a biológiai kérdések megválaszolására.

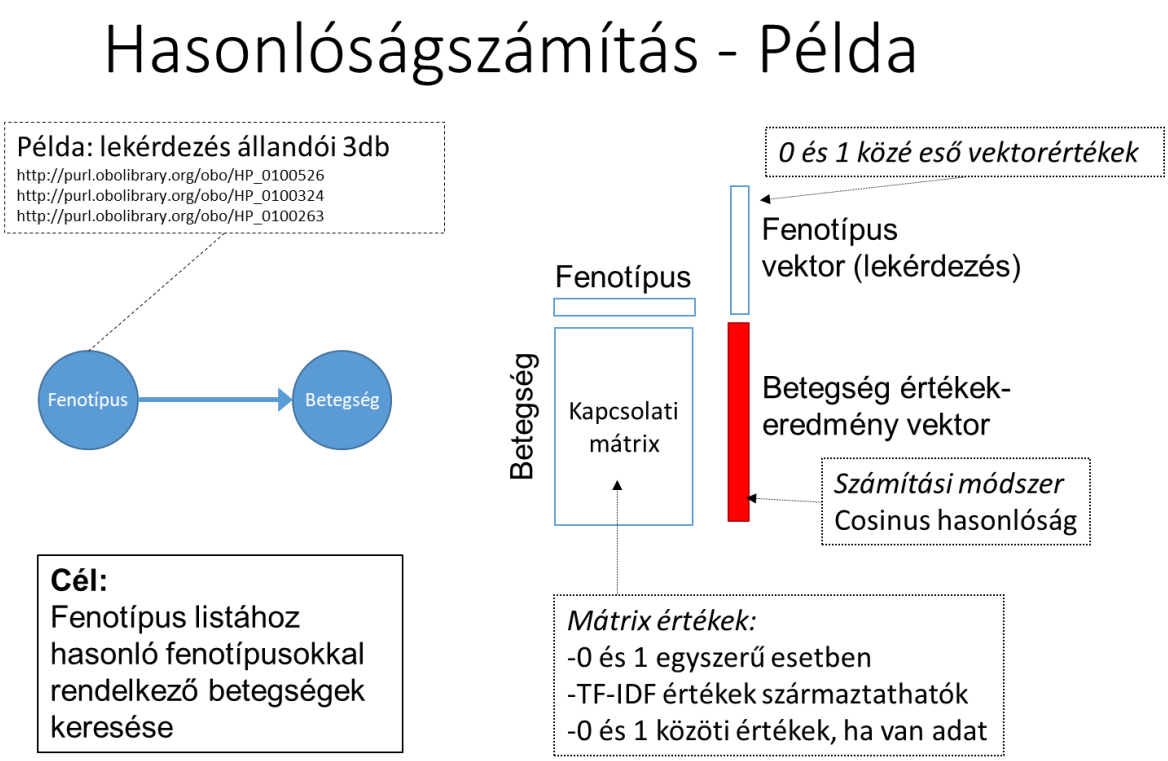
Az adatbázisok és a köztük levő kapcsolatok HDT (Header, Dictionary, Triples) formátumban tároltak, és onnan tölthetők be a prioritizálóba. A HDT egy tömörített RDF formátum, amely megtartja a keresési funkciókat anélkül, hogy az adatot ki kellene csomagolni. A különböző

entitások – mint például a gének vagy betegségek – csomópontokként, a köztük levő kapcsolatok pedig élekként foghatók fel. A mi esetünkben az élek és a csomópontok is külön fájlokban tárolódnak, és meghivatkozni is külön egységként lehet őket. Az egyes entitásokat leíró RDF-nek megfelelő adatfájlok tartalmazzák az entitás azonosítóját URI formában, és további jellemzőit, mint például azok hivatalos és alternatív elnevezését, típusát és leírását, amennyiben ezek értelmezhetőek és rendelkezésre állnak.

Az élek két csoportra oszthatók, egyszerű és összetett élekre. Az egyszerű élek adatfájljai olyan adatokat tartalmaznak, ahol két különböző csomópontba eső entitás az RDF hármas alanya és tárgya, tehát a köztük levő kapcsolat meglétéről tartalmaz csak információt. Az összetett élek esetében pedig az RDF hármas alanya egy olyan URI, ami két különböző csoportba tartozó entitások kapcsolatát írja le úgy, hogy a tárgyai egyrészt a két entitás és további tárgyai a kapcsolatra jellemző egyéb adatok, mint például az információ forrása vagy egy kémiai kapcsolat esetén a hatáserősség.

A keretrendszer bemeneteként XML formátumban kell megadni azt az ún. számítási gráfot, amely meghatározza, hogy az egyes csomópontokon megadott evidenciák (pl. adott entitások megfigyelései, mérési eredményei) hogyan terjedjenek a gráfban és mely csomópont prioritizálását kívánjuk elvégezni. A csomópontokon és az éleken további szűrések állíthatók be, amely lehetőséget ad bizonyos következtetési utak kizárására. Ezen kívül az élek esetén meg kell adni, hogy az adott él kiindulási csomópontjában szereplő információ (azaz az adott entitások evidenciái) milyen számítási módszer szerint terjedjen az él végpontját reprezentáló csomópontba. Kimenetként pedig a rendszer beállításaitól függően a prioritizálás találatainak az azonosítói, azok értékei, és esetleges releváns paraméterei, mint a neve vagy leírása kaphatók vissza.

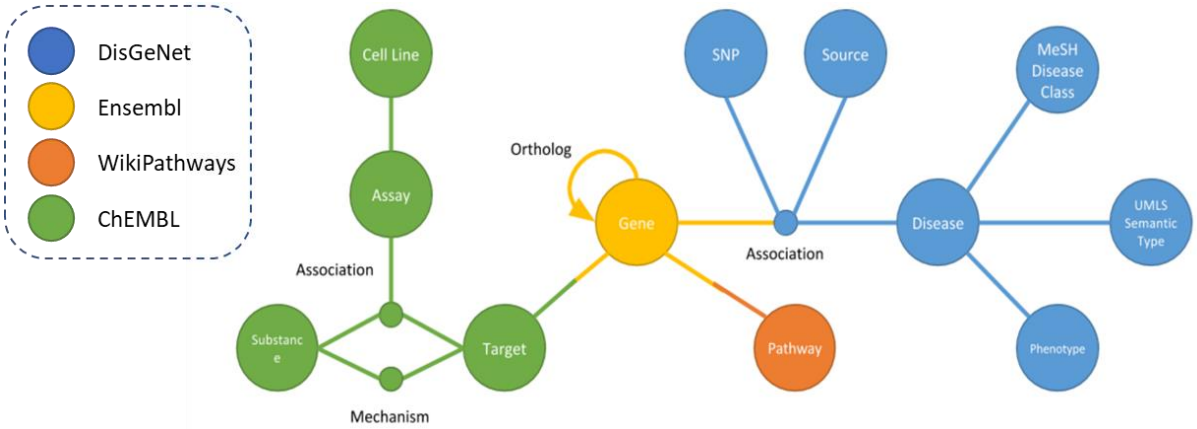
A modell tehát csomópontokból, egyszerű és összetett élekből épül fel, és ezeken képes információt (evidenciákat) terjeszteni, a meghatározott kombinációs következtetési sémák szerint pedig kvantitatív értékeket visszaadni a kimenetként definiált entitásokra vonatkozóan.



9. ábra) A hasonlósági számítási módszerre egy példa

4.2 A RENDSZER INFORMÁCIÓ FORRÁSAINAK A LEÍRÁSA

A Kvantitatív Szemantikus Prioritizáló keretrendszer információ forrásai a ComBine csoport által korábban használt bioinformatikai kutatásoknak megfelelően jelenleg 13 csomópontot és közöttük elhelyezkedő 4 összetett és 11 egyszerű élt különböztetünk meg. Az adatokat gráfba rendszerezhetjük, és ennek egy már korábban elkészített egyszerű változata látható alább [10. ábra].



10. ábra) A Kvantitatív Szemantikus Prioritizáló adatbázisainak a kapcsolati hálója a webes felületnek megfelelően

Látható, hogy a színkódoknak megfelelően 4 különböző forrásra lehet bontani az adatokat, a ChEMBL alapú kémiai, az Esnembl alapú genetikai, a DisGeNet alapú betegség és a WikiPathway alapú útvonal csoportokra. Az ábrán körökkel vannak jelezve az összetett élek, és vonalakkal az egyszerű élek, továbbá az is leolvasható, hogy meg lehet adni élet azonos típusú csomópontok között is, mint a gén ortológia (fajközi megfeleltetés) kapcsán.

Az egyes csomópontok és kapcsolataiknak a leírása fontos a lekérdezések és az adat terjesztése szempontjából, mert vannak olyan esetek, ahol a kapcsolati háló csak meglehetősen hiányosan van feltérképezve, továbbá több esetben is megtévesztők a kapcsolatok. Ezen kívül a csomópontok adatainak és kapcsoltságának a száma információt hordoz egyes tengelyek erősségére vonatkozóan, de bizonyos esetekben jelezhet adatvesztést is egy alacsonyabb elemszámú köztes csomópont. Ezek miatt most kifejtem biológiai szemszögből, hogy az egyes csomópontok és élek milyen adatokat tartalmaznak, és a további felhasználást megkönnyítően jellemzem az egyes információforrások kapcsoltságát, és ezek lefedettségét a többi forrásból származtatható adat által.

4.2.1 A genetikai adatok

A genetikai adatok az Ensembl adatbázisból származnak és azok azonosítóit tartalmazzák, ezen belül a gén (Gene) csomópont közel 1,2 millió gént tartalmaz, ezek közül 38 ezer emberi gént, ami közel lefedi az összes emberi gént. A génekhez azonosítón kívül további adatok is tartoznak a leírásukra, típusukra és közismert neveikre vonatkozóan. A nem emberi gének további 68 különböző fajból származnak, és az ismert fajközi megfeleltetések, ortológiák 37 millió megfeleltetéssel jól lefedik a gyakorlati jelentőséggel bíró géneket. A fajok megkülönböztetéshez és szűréséhez egy Taxon vagy faj csomópontot használunk, ami a fajokat tartalmazza, és minden génhez megfeleltethető egy faj.

A genetikai adatok terjesztéséhez egy gén-gén hasonlóság él is ki lett dolgozva, ami a GO (Gene Ontology) ontológiai hasonlóság alapján határozható meg automatikusan tetszőleges gének között. Viszont az adatok növekedésének a gének számával való négyzetes kapcsolata miatt csak egy lekérdezésben szereplő génekre és kapcsolataikra van kiszámítva, de ez alapjában véve nem korlátozza az él funkcióját.

4.2.2 A kémiai adatok

A kémiai adatok az RDF EBI oldalon RDF formában elérhető, a ChEMBL adatbázisból származó adatokra épülnek, és 1,5 millió vegyületet (Substance) és 11 ezer vegyületcélpontra (Target) vonatkozó adatot tartalmaz. A vegyületek nem tartalmaznak antitestekre vonatkozó adatokat, amelyek a modern gyógyszerek jelentős részét képezik, viszont ez minimálisan érinti a gyógyszerjelöltekre való előrejelzést, mivel többnyire kis molekulasúlyú vegyületekre vonatkozó adatokat szeretnénk visszakapni a további kísérletek megkönnyítése érdekében.

Vegyületekre vonatkozóan elérhető adatok magukba foglalják a vegyületek azonosítóján kívül a leírásukat, a használt elnevezések listáit, a vegyület típusát és osztályozását. A vegyületcélpontoknál az azonosítón kívül a célpont megnevezése és típusa érhető el.

Továbbá a 14 millió adatpárral rendelkező vegyület-célpont kapcsolatokhoz elérhető az kapcsolati információt szolgáltató vizsgálat (Assay) típusa és annak semantikus leírása 1,2 millió vizsgálattal; továbbá a vizsgálatokhoz kacsolódó 1600 sejtvonalakra (Cell Line) vonatkozó kapcsolatok és adatok is elérhetők.

Gyakorlati jelentősége van annak, hogy a célpontok mekkora hányada feleltethető meg géneknek, mert további számításokra csak ezeket tudjuk felhasználni. Ezt megvizsgálva a 11 ezer célpont közül mindössze 3250 célponthoz található gén, melyek közül 2400 emberi gén. Ez korlátozhatja a genetikai és kémiai tengely közti információáramlást, ezért a jövőben a PubChem adatbázissal való bővülést tervezzük, amit a kisebb megbízhatósága miatt eddig mellőztünk.

4.2.3 A betegség adatok

A betegség tengely több betegség azonosítási és ontológiai adatbázist tartalmaz, viszont a gén-betegség kapcsolatra vonatkozó adatok a UMLS (Unified Medical Language System) rendszert használja, ezért a jelentéssel bíró adataink innen származnak, és csak az ezekre vonatkozó adatokat foglalom össze. Így 5400 betegség tartozik ide, melyekhez csak a név érhető el szemantikus adatként. Viszont a 1200 betegséghez elérhető összesen 6 ezer fenotípus adat, ami egy leírásnak is megfelel.

Amennyiben egy betegségcsoportot szeretnénk meghatározni, akkor lehetőségünkre áll 25 UMLS típus, mint genetikai vagy szervi megbetegedés, a betegség leírására vonatkozóan, és

28 MESH (Medical Subject Headings) osztály a betegségek klasszikus besorolására vonatkozóan, mint idegrendszeri vagy fertőzéses megbetegedés.

A DisGenet forrásából származó gén-betegség kapcsolathoz további adatok is rendelkezésünkre állnak, mint az, hogy az összefüggés mely források alapján lett meghatározva, és az, hogy mely gén variánsok ismertek egy-egy kapcsolatra vonatkozóan. Összesen betegségeknek 17 ezer emberi gén feleltethető meg, ami a 38 ezres (melyből 20 ezer fehérjét kódoló) Ensembl emberi gének számához viszonyítva egy jó arány.

4.2.4 Az útvonal adatok

Az útvonal vagy jelút (Pathway) adatok a WikiPathway adatbázisról származnak és 1,700 jelutat tartalmaz, illetve ezek kapcsolatát 22 ezer génnel, amelyek közül 5 ezer emberi és 500 pedig az öregedéskutatásban igen fontos C.Elegans nevű féregből származik. Ebben a csomópontban minden útvonalhoz tartozik az azonosítón és nevén kívül egy rövid és egy igen tartalmas 1000 karakteres leírás is.

4.3 KVANTITATÍV SZEMANTIKUS PRIORITIZÁLÓ WEBES FELÜLETE

A Kvantitatív Szemantikus Prioritizálóhoz a tanszéki ComBine laboratórium fejlesztésében korábban elkészült egy webes felület is, mely alkalmas a 10. ábrának megfelelő egyszerű lekérdezésekhez szükséges XML generálására, annak lefuttatására és a találatok visszaadására.

5 A KERETRENDSZER TESZTELÉSE ÉS FELHASZNÁLT MODELLEK

5.1 EGYSZERŰ MODELLEK

A csomópontok adatai és az élek ismeretében még mindig több probléma merül fel azzal kapcsolatban, hogy hogyan határozzuk meg az élek súlyát egymáshoz képest, valamint, hogy az éleken milyen típusú számításokat hajtsunk végre, és hogy ezen a számításon milyen további korrekciókat hajtsunk végre. Egy példa a problémára az, ha egy gén listához keresünk betegséget, ekkor megtehetjük azt, hogy a géneket megfeleltetjük betegségekkel és az értékeiket összeadjuk, de azt is megtehetjük, hogy egy hasonlósági léptékkel azt vizsgáljuk, hogy ez a génlista melyik betegség gén listájára hasonlít a legjobban. Továbbá ezek mellett tetszőleges korrekciókat is végezhetünk a listák méretének és az elemek gyakoriságának a korrekciójára.

Ennek megfelelően különböző biológiai tartalommal rendelkező egyszerű modelleket dolgoztam ki arra, hogy az egyedi éleken folyó következtetés (információáramlás) karakterisztikáját megértsem, és hogy ajánlásokat tegyek az evidenciák terjesztési módszereinek a beállítására.

5.1.1 Egyszerű kémiai minta modellek

A kezdeti kísérletek alapján a kémia tengely (dimenzió) okozta a legnagyobb diverzitást a lekérdezésekben, ezért külön megvizsgáltam, hogy egyes gyógyszercsoportok célpontjai, azok génjei, hogyan viselkednek a lekérdezések folyamán. Ehhez két gyógyszercsoportot vizsgáltam meg, a PPI (proton pump inhibitor) proton pumpa gátló (savcsökkentő) gyógyszereket, mint egyszerű gyógyszercsoportot és az SSRI (selective serotonin reuptake inhibitor) antidepresszánsokat, mint összetettebb és szélesebb hatásspektrummal rendelkező gyógyszercsoportot.

A gyógyszerekre vonatkozó adatokat a DrugBank adatbázisról szereztem és Chemical Translation Service oldalát használtam a DrugBank és ChEMBL azonosítók közti automatikus megfeleltetésnek, amit a ChEMBL weboldalán történő manuális megfeleltetés követett a hiányzó azonosítók esetén. Sajnálatos módon a legtöbb elérhető szolgáltatás az általam vizsgált adatokra 70% körüli megfeleltetési arányt nem érte el a Chemical Translation Service kivételével, ami tovább nehezíti a kémiai adatokkal való foglalkozást (41) (42).

SSRI		PPI	
DrugBank	**ChEMBL**	**DrugBank**	**ChEMBL**
DB06700	CHEMBL1118	DB00213	CHEMBL1502
DB00476	CHEMBL1175	DB00338	CHEMBL1503
DB01175	CHEMBL1508	DB00448	CHEMBL480
DB00176	CHEMBL1621884	DB00736	CHEMBL1201320
DB04884	CHEMBL2110900	DB01129	CHEMBL1219
DB04896	CHEMBL252923	DB05351	CHEMBL1201863
DB00472	CHEMBL41		
DB00715	CHEMBL490		
DB00215	CHEMBL549		
DB01149	CHEMBL623		
DB00285	CHEMBL637		
DB01104	CHEMBL809		

1. táblázat) Az SSRI és PPI gyógyszerek DrugBank és ChEMBL azonosítói

A 6 PPI-t maradéktalanul meg tudtam feleltetni míg a 16 SSRI-ből csak 12 volt megtalálható a ChEMBL-ön [1. táblázat]. Ezen adatok segítségével tudtam felmérni a vegyület-célpont tengely jellemzőit, és ajánlásokat tenni a felmerült problémákra.

5.1.2 Makuladegeneráció, mint minta modell

A kémiai tengelyt követően a betegség tengelyt vizsgáltam, pontosabban a gén-betegség vonalat. Ehhez egy makuladegenerációs modellt választottam, mint egy megfelelő irodalmi háttérrel rendelkező poligénes, vagyis kellően összetett genetikai háttérrel rendelkező betegséget.

Megemlíteném, hogy a makuladegeneráció (MD) és az időskori makuladegeneráció (AMD) az irodalomban gyakran felcserélve használt, de a pontossághoz hozzátartozik, hogy az MD egy nagyobb betegségcsoport és néhány ritkább betegséget is tartalmaz az AMD-n kívül, viszont az AMD a legmeghatározóbb és legjobban feltérképezett csoportja, ezért az adatgyűjtés során mérlegeltem, hogy mikor melyik megnevezés elfogadható, és a szövegben a két megnevezéssel jelzem, hogy a forrásban melyik néven volt elérhető, de a gyakorlatban ezek az időskori makuladegenerációt takarják.

A tesztelést két lépésben végeztem el. Elsőként csak a makuladegenerációhoz kapcsolt gén-betegség élen elérhető géneket használtam, második lépésben viszont az irodalomban elérhető források alapján összegyűjtöttem a makuladegenerációhoz köthető vegyületeket, célpontokat, jelutakat, emberi és állati géneket tartalmazó listákat és ezek információját vetítettem rá külön-külön és egyben is a betegség tengelyre. Az így kapott eredmények

alapján tettem javaslatot a számítási módszerek megválasztására és ezen keresztül jellemeztem az egyes élek erősségeit és problémáit.

5.1.2.1 Makuladegeneráció rendelkezésre álló génekkel

Első feladatként megvizsgáltam, hogy a betegség csomóponton belül milyen betegség felel meg a makuladegenerációnak és amennyiben elérhető, akkor ennek egy speciális esetének az időskori makuladegenerációnak (AMD) továbbá, hogy ezek a betegségek hány génnel vannak kapcsolatban. Az erre a lekérdezésre az alábbi eredményt kaptam [2. táblázat].

Gének száma	**Betegség neve**
527	Age related macular degeneration
95	Macular degeneration
69	Exudative age-related macular degeneration
…	…(15 további betegés)
1	MACULAR DEGENERATION, AGE-RELATED, 3
1	MACULAR DEGENERATION, AGE-RELATED, NEOVASCULAR TYPE, SUSCEPTIBILITY TO
1	Macular Degeneration, Age-Related, 7

2. táblázat) Makuladegenerációhoz köthető betegségek listája, és az érintett gének száma

Ebből több információ is kiolvasható, mint az, hogy az általánosabb betegségek nem öröklődik az alcsoportjaik génjeit, mert a makuladegenerációnak kevesebb, mindössze 95 génje van, ellenben az időskori makuladegenerációval, amihez pedig 527 gén tartozik. Továbbá látható, hogy olyan esetek is vannak, amelyek néhány, bizonyos esetben csak egy gént tartalmaznak, és az is ismert, hogy bizonyos számítási módszerek, mint az átfedésen alapuló hasonlósági metrikák érzékenyek erre, mivel egy találat esetén is teljesen lefedi egy lekérdezett génlista a betegség génjeit, ebben az esetben génjét.

Itt megjegyezném, hogy több találat is volt makuladegenerációra a betegségek között, de most csak azokat részletezem melyekhez kapcsolódik gén, mivel a számítások során csak ezeket kaphatjuk vissza eredményként.

Ezt követően áttekintettem az időskori makuladegenerációhoz (URI: http://linkedlifedata.com/resource/umls/id/C0242383) kapcsolódó géneket és a hozzájuk tartozó DisGeNET-ből származó értéket [Függelék, 13. táblázat]. Ezek alapján több következtetést is le lehetett vonni, mint például azt, hogy több különböző azonosító is szerepel ugyanazon gén neve alatt, ami az Ensembl adatbázis jellemzője, és javaslatomra ezen

a téren korrekció is történt a modellben, így a továbbiakban ezzel a hatással nem kell számolnunk. Ez azt jelentette, hogy az 527 génből mindössze 391 gén volt egyedi és a számítások továbbá ezen a listán zajlottak.

A számítások beállításainál a keretrendszer lehetőséget biztosít arra, hogy a számítást a score figyelembevételével, és anélkül is el lehessen végezni, amiket meg is tettem külön-külön a keretrendszerben elérhető számítási módszerek segítségével. Az ellenőrzést úgy végeztem, hogy a közel 400 génből 100, 150, 200 és 250 gént választottam ki esetként (pozitív példaként) és a komplementer génhalmazból pedig 100%, 150% és 200%-nak megfelelő mennyiségű gént választottam ki kontrollnak (negatív példaként), és az így kapott gén listákon elvégzett prioritizálásokat vizsgáltam meg a különböző számítási módszerek használatát.

A számítások robusztusságának vizsgálatához egy olyan eset-kontroll megválasztást kerestem, amely kevéssé érzékeny gének vagy betegségek megoszlására és feltérképezettségére, azaz ilyen tekintetben nem mutat zavaró statisztikai hatásokat produkáló rétegződést. Ez úgy próbáltam meg elérni, hogy hasonlóan gyakran előforduló géneket választottam kontrollnak, mint az AMD-hez köthető gének. Ezt úgy oldottam meg, hogy az összes betegségben érintett génnek akkora értéket adtam, amennyi betegségben szerepel, és ennek megfelelően választottam meg a kontroll géneket. Ezt egyszerűen meg lehetett oldani, mert a rendszerből könnyen lekérdezhető, hogy az egyes gének hány betegségben is szerepelnek.

A 400 génből történő pozitív példa megválasztását úgy végeztem el, hogy az lehetőleg minél jobban kövesse a többi gén-betegség megoszlás karakterisztikáját. Ez azt jelenti, hogy a nagyobb értékekkel rendelkező gének aránya nőjön, amennyiben a génhalmaz száma csökken. Ennek az elérésére a gének értékeit lineárisan 1 és 0.1 közé normáltam, és ezen értékeket szoroztam meg egy random 0 és 1 közötti számmal, és ezt rangsorolva választottam az első találatokat. Ez a módszer annak ellenére, hogy nem nyugszik szilárd matematikai alapokon megfelelő karakterisztikát eredményezett.

A negatív példaként használt géneket úgy választottam ki, hogy a gyakoriságuknak megfelelő rangsorban az AMD-kapcsolt gének közelében helyezkedjenek el, random, de nem nagyobb mint 10 gén távolságban. Ehhez megnéztem, hogy a szomszédos gének összefüggenek-e, de nem találtam erre utaló jelet.

5.1.2.2 Makuladegeneráció irodalomból származó adatokkal

Ahhoz, hogy valós adatokkal is fel tudjam mérni a rendszer jellemzőit, összegyűjtöttem az irodalomban említett emberi géneket és hozzávettem a teljes genom asszociációs vizsgálatok (GWAS) adatait. Az irodalmon három jellemző genetikai vonal volt érintett, a CFH (complement factor H) az immun és véralvadási funkcióval rendelkező komplement fehérjék egyike, a VEGF (Vascular Endothelial Growth Factor) az erek növekedésért és burjánzásáért felelős érnövekedési faktor és az ARMS2 vagy más néven HTRA1 gén, ami sejtek közti mátrix (ECM) bontásában vesz részt. Ezen gének fehérjéi és a velük kapcsolatban álló többi fehérje, vagyis azonos útvonalban szereplő fehérjék adják a makuladegenerációval kapcsolatos kutatás alapját. Ezekkel összevetettem egy friss GWAS (Genome-wide association study) publikáció találatait, és a GWAS catalog adatbázisban található eredményeket, és azt találtam, hogy a CFH és ARMS2 gének messze kiemelkedtek, viszont a VEGF átlagosan a 10-13. találat volt (43, 44). Ezek alapján, hogy egy általános módszert tarthassak meg a gének modellen belüli értékeinek a megadására a számítást a GWAS eredmények alapján végeztem, meghagyva a lehetőséget a kiemelkedően fontos gének felülsúlyozására.

Továbbá az útvonal csomópontban meghatároztam a három irodalmi génhez tartozó, két komplement egy érfejlődési és egy mátrix lebontási jelutat.

A fajközi vonal reprezentálása érdekében pedig felkutattam az ismert állati géneket és modelleket. A makuladegenerációval kapcsolatban a feltérképezett fajközi adatok egérből és patkányból származnak. Először is megnéztem, hogy az irodalom milyen modelleket említ a témában, és ez alapján 17 modellekben használtgént találtam (45). Ezeket összehasonlítva a MGI (Mouse Genome Informatics) MGD (Mouse Genome Database) adatbázis 25 génjével azt találtam, hogy az adatbázisban a gének csupán 5 korábban kigyűjtött gént találtam meg, viszont a talált gének többsége összefüggésbe volt hozható a modell génekkel, ezért ezen adatbázis génjeit is felhasználtam (46). További adatokért az RGD (Rat Genome Database) (47) adatbázist kerestem fel, ami csak névben és közös összefogásban (Alliance of Genome Resources) való részvételben hasonlítanak, de külön egységek. Az RGD-ben keresési feltételtől függően 27-50 időskori makuladegenerációhoz köthető gént találtam. Ezeket a géneket is felhasználva alakítottam ki az AMD faji reprezentációját [Függelék 14. táblázat].

Egy további információforrás a kémiai vonal, ahol a makuladegenerációhoz köthető gyógyszereket és gyógyszer célpontokat integráltam a modellbe. Ehhez a DrugBank adatait

használtam fel egyrészt úgy, hogy az ismert gyógyszereket és kísérleti fázisban levő szereket kigyűjtöttem, részben pedig a ChEMBL gyógyszer célpontjai közül meghatároztam azokat, amelyek érintettek voltak a célzott kezelések kapcsán (41). Ezek alapján 10 egyedi célpontot találtam, melyek összhangban voltak a korábbi genetikai találatokkal.

Azonosító	Név	Típus
CHEMBL1697671	Placenta growth factor	SINGLE PROTEIN
CHEMBL1778	Interleukin-2 receptor alpha chain	SINGLE PROTEIN
CHEMBL1783	Vascular endothelial growth factor A	SINGLE PROTEIN
CHEMBL1909490	Interleukin-1 beta	SINGLE PROTEIN
CHEMBL2094253	Cyclooxygenase	PROTEIN FAMILY
CHEMBL2176771	Complement factor D	SINGLE PROTEIN
CHEMBL2364163	Complement C5	SINGLE PROTEIN
CHEMBL3286061	Sphingosine-1-phosphate lyase 1	SINGLE PROTEIN
CHEMBL4611	Complement C1r	SINGLE PROTEIN
CHEMBL4917	Complement C3	SINGLE PROTEIN

3. táblázat) A MD gyógyítására használt gyógyszerek célpontjai

A kémiai adatokat a DrugBank azonosítók használatával fordítottam át ChEMBL azonosítóra, melynek a <http://rdf.ebi.ac.uk/resource/chembl/molecule/> előtaggal képzett címe az általunk használt URI. Ezekhez a szerekhez elérhető volt az, hogy hányadik klinikai fázisban vannak, és az értéket a $(Fázis\ száma\ +\ 1\ )\ *\ 0.2$ képlettel számítottam. A gyógyszerek listája az függelékben [Függelék 15. táblázat] láthatók, és a névként a szerre jellemző rövidebb, vagy köznapi nevet használtam, de az esetek többségében számos megnevezés érhető el egy szerhez.

A modellt összességében ezeknek az információforrásoknak megfelelően teszteltem, és vizsgáltam a számítási módszereket és az élek súlyait.

5.2 AZ ÖREGEDÉSI MODELL

A korábbi egyszerűbb modellek azt a célt szolgálták, hogy a QSFP keretrendszer számítási képleteinek a megválasztására és az élsúlyok mértékének megválasztását megvizsgáljam és javaslatokat, protokollokat tudjak kidolgozni, amelyek felhasználatók lesznek az öregedés összetett modelljeinek a tényleges kérdései esetében.

A kutatásom jelentős részét az jelentette, hogy az öregedéssel kapcsolatos irodalmat áttekintsem és a fellelhető források segítségével az öregedéshez, élettartamhoz, egészségtartamhoz kapcsolódó reprezentatív adatokat gyűjtsek össze modellek számára.

Célom volt, hogy az adatok lehetőleg minél nagyobb hányada származzon már korábban összeszedett adatbázisokból, adatforrásokból, de sajnálatos módon nem lehetett elkerülni a hatalmas 400,000 cikket érintő irodalom százas nagyságrendű releváns publikációinak az áttekintését. Viszont a TDK dolgozat jellege miatt igyekszem csak az itt bemutatott modellekben felhasznált adatokat és adatforrásokat megemlíteni.

Az adatok gyűjtése során praktikussági szempontból két csoportra osztottam az öregedéssel kapcsolatos adatokat, egyrészt közvetlen adatokra, melyek átvitel nélkül olyan forrásokból származtak, melyek az ellenőrzött irodalmi adatokat tartalmaztak az öregedésre vagy élettartamra vonatkozóan. Másrészt a kevésbé megbízható, automatizált vagy genetikai tengelyhez közvetlenül nem köthető források adatait az indirekt csoportba soroltam. Ennek az elsődleges oka az volt, hogy más súlyokat és számítási módszereket érdemes választani a két csoport adatai esetében, és rendszerezés szempontjából is jobban elkülöníthetők.

5.2.1 A közvetlen források adatai

Ebbe a csoportba ellenőrzött adatbázisok irodalmi gyűjtésből származó állati és emberi génjei, valamint a már korábban a megnövekedett élethosszhoz kapcsoló variánsok és az ismert öregedéssel összefüggő jelutak génjei kerültek. Összesen 7 listába csoportosítottam a közvetlen adatokat forrásuk és típusuk alapján.

5.2.1.1 Human Ageing Genomic Resources adatai

A HAGR (Human Ageing Genomic Resources) egy online manuálisan vezetett adatbázis, mely összefogja az öregedéshez kapcsolódó fontosabb irodalmat és a tartalmát, továbbá több genetikai és gyógyszer listát vezet a témában (48).

A 7 közvetlen információt tartalmazó listából 3 a HAGR-ról származik, ezek a 300 gént tartalmazó emberi öregedés és az 1100 gént tartalmazó egér, fonalféreg és ecetmuslica állatmodellekben feltérképezett öregedés génjei GenAge adatbázisból, valamint az emberi megnövekedett élettartamhoz köthető 300 génes listája a LongevityMap adatbázisból (49).

5.2.1.2 A Gene Ontology génjei

A HAGR mellett a másik meghatározó forrás a GO (Gene Ontology), amely számos adat között tartalmazza az öregedést is (azonosítója GO:0007568) (50). Ehhez az ontológiához köthető adatokat az Ensembl BioMarts (51) szolgáltatása segítségével gyűjtöttem össze, az ide tartozó adatok 300 emberi géneket és 9 ismert modellfajból származó 1300 gént tartalmaz. A

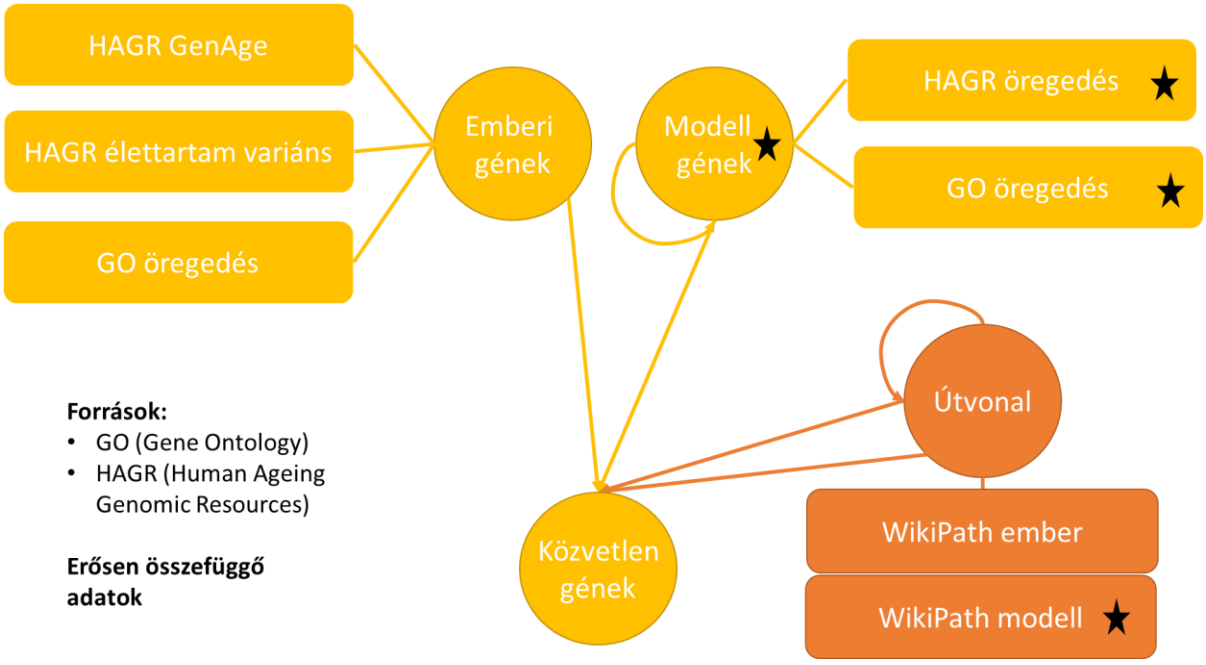
modellfajok egér, patkány, zebrahal, tyúk, sertés, szarvasmarha, kutya, ecetmuslica és fonálféreg, és ezek közül a legtöbb adat patkány és egér modellekből származik.

5.2.1.3 Öregedés jelutai

A jelutakat az irodalmi adatok alapján és a WikiPathway leírásának a segítségével gyűjtöttem, és 3 emberi és 3 állati útvonal volt köthető a témához. Ezek összesen 130 emberi és 40 állati gént érintenek, és jól átfednek a korábbi adatbázisok találataival.

5.2.1.4 A közvetlen források összegzése

A közvetlen forrásokból származó adatokat a gyakorlati elrendezésnek megfelelően, átlátható formában az alábbi ábrán [11. ábra] látható, csillaggal jeleztem azokat a forrásokat és csomópontokat, ahol állati gének is szerepelnek. Az ábra megtartja a 4.2 pontban használt színsémát, azzal kiegészítve, hogy a bemenetnek használt listákat téglalap alakú dobozzal ábrázoltam.



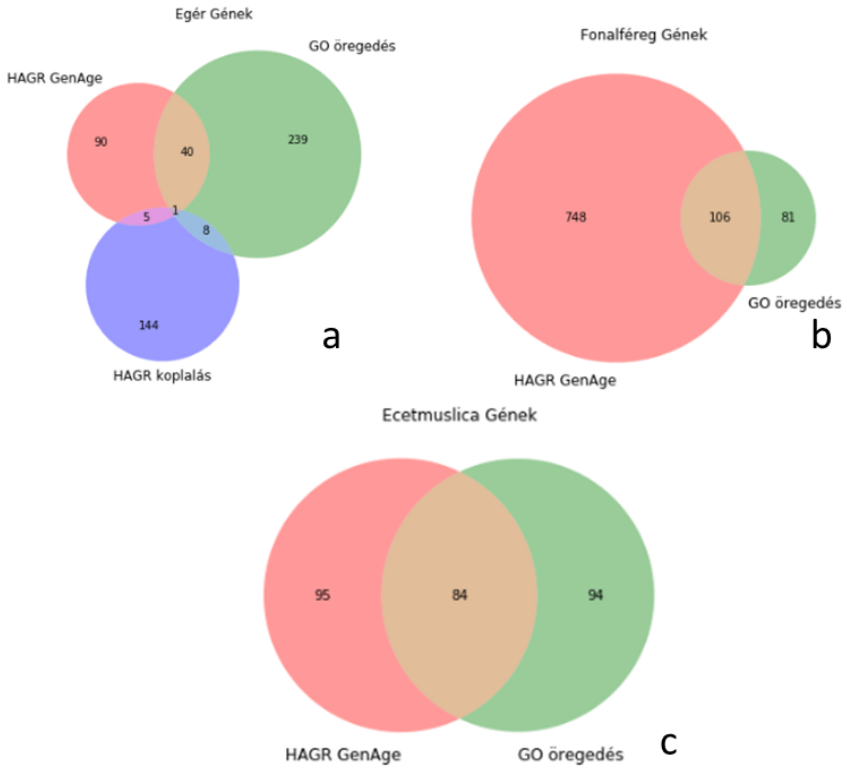
11. ábra) A közvetlen források ábrázolása gráfban a gyakorlati elrendezésnek megfelelően

Az öregedés genetikai tengelyének ezek az adatok egy nagy részét lefedik. Mivel ezek a források a szakirodalmat dolgozzák fel, ezért azt várnánk, hogy a talált adatok nagy mértékben átfednek egymással, ezzel szemben, ha ábrázoljuk a forrásokat, akkor azt kapjuk [12. ábra], hogy mindössze 16 gén található meg a 4 emberi géneket tartalmazó (útvonal esetén arra átvezethető) mindegyikében, és a gének többsége csak egy forrásban szerepel. Ez részben

arra utal, hogy a források nem tudták a teljes vonatkozó irodalmat áttekinteni, részben pedig arra, hogy a források valamelyes torzítanak is, mert a géneket megvizsgálva azt találtam, hogy bizonyos esetekben olyan következtetést is levontak egyes cikkekből, amelyek nem voltak kellően vagy egyáltalán alátámasztva. Ez összességében nem jelent problémát, mert a számítási módszereink megfelelő körülmények között képesek az ilyen típusú torzítások korrigálására.

12. ábra) A közvetlen emberi információforrások átfedettsége. (Gyakorlati szempontból, hármasával ábrázolom a csoportokat, mert az arányos megjelenítést lehetővé tevő matplotlib_venn python csomag eddig támogatja az ábrázolást)

Az emberi génekhez hasonlítva az állati gének jobb átfedést mutatnak [13. ábra]. Ennek egy oka az, hogy ezen adatok egy része jól feltérképezett emberi géneknek az ortológja, ezért ezen adatok nem tekinthetők függetlennek.



13. ábra) A modell gének átfedettségének az ábrázolása egér, ecetmuslica (Fruitfly, D. melanogaster) és fonalféreg (worm, C. elegans) esetében. A HAGR koplalás egy közvetett expressziós adat, az összehasonlítás kedvéért.

5.2.2 A közvetett források adatai

A közvetett források közé soroltam azokat a genetikai módszereket, melyek reprodukálhatósága kérdéses és téves találati arány a legmagasabb találatok esetén is meghaladja az 50%-ot. Ide sorolhatók a génexpressziós és metilációs vizsgálatok. Sajnos sem az expressziós sem a metilációs vizsgálat nem ad teljes betekintést egy gén funkciójába mivel csak géntermék koncentrációjára hat, és arra se teljesen determinisztikusan, továbbá a megváltozott mintázatok gyakran indifferensek egy fenotípus szempontjából. A genetikai adatokon kívül ebbe a kategóriába soroltam azokat a tengelyeket, melyek nem közvetlenül kapcsolhatók a génekhez, mint a fenotípus, betegség és vegyület tengelyeket.

5.2.2.1 Öregedéssel kapcsolatos közvetett genetikai adatok

Az öregedéssel kapcsolatos genetikai vizsgálatok jelentős része génexpresszión alapszik. A HAGR expressziós adatbázisa ezt gyűjtötte össze és innen származik az egyik felhasznált és megbízhatónak tartott 70 génes lista. Egy további friss és jól tervezett expressziós vizsgálat az emberi vérben korral eltérő expressziós mintázatokat vizsgálta meg és 150 gént említett meg (52). A korábbi születése és módszertana miatt az adatok közé nem vettem be egy másik

expressziós vizsgálatot 150 génes eredményeit de a módszer bemutatásaként szerepel az alábbi ábrán [14. ábra] (53).

Egy másik genetikai módszer a metilációs mintázat vizsgálata, ehhez egy népszerű értekezésben számos nyíltan elérhető emberi metilációs kíséret adatait dolgozta fel a szerző, ahol a vizsgált emberek életkorával vetette össze, és az adatait felhasználva jutott hozzá egy 340 génes listához (54).

14. ábra) A közvetett genetikai adatok átfedettsége, ahol az Expression halmaz két expressziós vizsgálat eredményit tartalmazza, a Methylation halmaz a metilációs kísérletét a HAGR pedig a HAGR expressziós eredményeit.

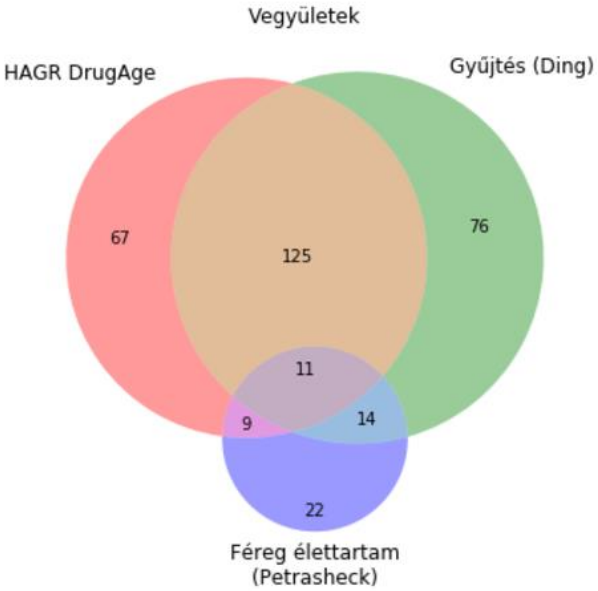
Ha az közvetlen genetikai eredményeket nézzük, akkor igen gyenge átlapoltság látható [14. ábra]. Ez utalhat a vizsgálati módszerek, körülmények és szövetek eltérő viselkedésére, de leginkább a módszertan alacsony megbízhatóságára enged következtetni, különösen a publikációk magas idézettségére való tekintettel.

Egy további idetartozó adat a HAGR expressziós gyűjtése koplalásos egerek vizsgálata kapcsán, és meg lett említve korábban az állatmodellek génjei kapcsán [13. ábra], de információtartalmát és modellben betöltött helyét tekintve ide tartozik.

5.2.2.2 Öregedésre ható vegyületek

A genetikai tengely után az öregedésre ható, az élethosszt megnövelő vegyületeket tekintettem át. Ezek közül az első egy fonalférgeken végzett élettartamot megnövelő hatóanyag szűrés eredménye 56 hatóanyaggal (55). Emellett két forrás is foglalkozott az irodalmi adatok összegyűjtésével, melyek közül az egyik a már többet használt HAGR DrugAge adatbázisa 200-at is meghaladó ChEMBL azonosítóra fordítható hatóanyaggal (56). A másik

adatgyűjtés pedig egy idei cikkből származik, ahol a szerző szintén több mint 200 a modellbe integrálható vegyületet gyűjtött össze (57).



15. ábra) A fonalférgek öregedésére ható vegyületek (Petrasheck) (55), az adatgyűjtésből származó vegyületek (Ding) (57) és a HAGR DrugAge vegyületeinek az átfedettsége

Mivel ezek az adatok hasonló gyűjtések eredményei, ezért nem meglepő, a korábbiakhoz viszonyított jobb átfedettség. Továbbá a vegyületek ismétlődése segít az egyes vegyületek fontosságának és irodalmi ismertségének a meghatározásában, ezzel súlyozva ezt a tengelyt.

5.2.2.3 Öregedés fenotípusos jegyei

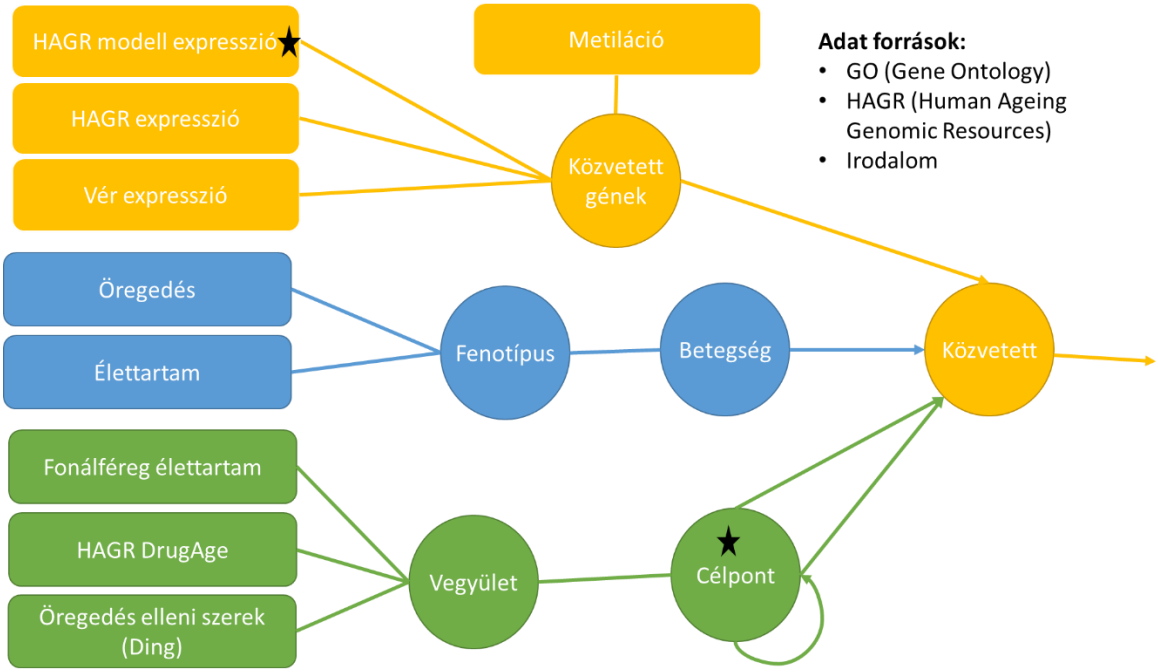
Az öregedés fenotípusos jegyeit manuálisan gyűjtöttem irodalomból, szakkönyvekből és általánosan ismert jegyekből, és ezeket feleltettem meg a HPO adatbázisban szereplő entitásoknak. Ez alapján 23 egyedi fenotípusos jegyet (mint őszülés, hallásromlás, bőr sorvadása, látásromlás) és 3 fenotípus csoportot (elgyengülés, elbutulás és szervek sorvadás) találtam [lásd: Függelék

16. táblázat].

Ezeken kívül az egészégtartamra jellemző marok szorítási erőt és a szarkopéniát (izomerő és izomtömeg csökkenés) hozzávettem a listához, mert ezek is általános öregedési jelek, de a kiértékelés szempontjából külön kezeltem őket.

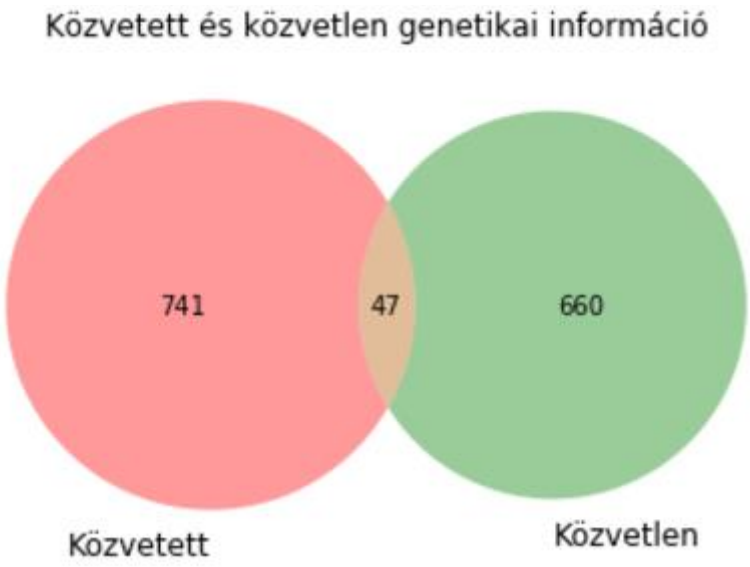
5.2.2.4 A közvetett források összegzése

A közvetett adatok forrásait tehát bizonytalan genetikai módszerek adatai és a fenotípus-betegség tengelyről, valamint a vegyület-célpont tengelyről származtatott adatok teszik ki. Ebből az is következik, hogy az egyes származtatási, következtetési metrikák ezen esetekben be folyásolni tudják az ide tartozó eredményeket.



16. ábra) A közvetlen források összefoglalása a korábbi szövegnek megfelelően

Bár a közvetett genetikai adatok magukban nem állnák meg a helyüket, azonban mivel információtartalmuk - ha bizonytalan is - de önálló értéket jelenthet, ezért kellő odafigyeléssel a probléma ismeretében próbáltam kiegészíteni vele a közvetlen adatokat. Ebben kulcsfontosságú a megfelelő következtetési módszer, amely képes a bizonytalan információkat a biztosabb adatok támogatására használni vagy hasonlósági módszerekkel ezeket a bizonytalan adatokat úgy súlyozni, hogy azok jobban megfeleljenek a valóságnak, így ezen adatok több eddig biológiailag nem ismert összefüggésre tudnak rávilágítani.

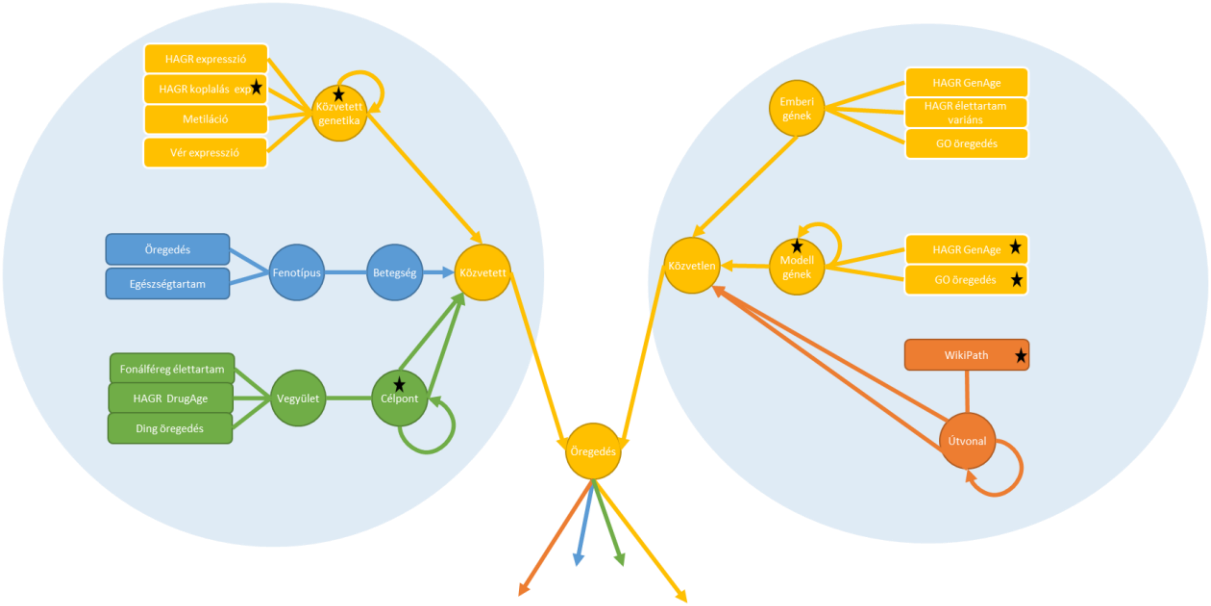


17. ábra) A közvetlen és közvetett emberi genetikai adatok átfedése

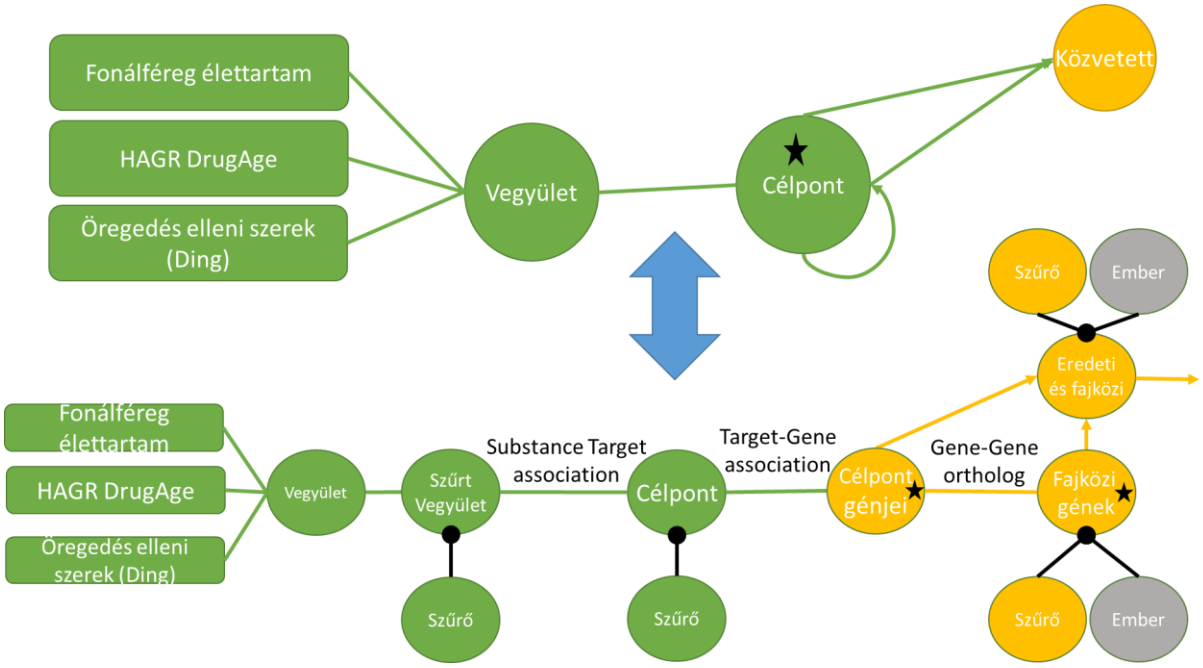
A fenti 17. ábra alapján látható, hogy a közvetett genetikai módszerek (a betegség és vegyület tengelyt nem beleszámolva) igen kis 47 génes átfedést mutatnak az irodalomban eddig ismert közvetett génekkel. Viszont mivel a közvetett gének adatok között is mindössze 120 volt több forrásban is megtalálható így ennek fényében látható, hogy annak ellenére, hogy itt még erősebben érvényesül a szemantikai átlapolás, az átfedés mértéke a 38 000-es modellben szereplő emberi gének számát tekintve véletlen esélynél jobb, összemérhető nagyságú.

5.2.3 Az öregedési adatok összefoglalása

Összességében az öregedéssel kapcsolatban több forrásból egy elfogadott biológiai publikációnak megfelelő mélységű és tartalmú információt gyűjtöttem össze a témában, és rendszereztem az alábbi ábrának [18. ábra] megfelelően. A valóságban a korábbi ábrázolástól eltérő stuktúrában szerepelnek az adatok a bemeneti XML fájlokban, ennek az átalakításnak a vázlata látható alább [19. ábra].



18. ábra) A QSFP rendszernek megfelelő ábrázolása az öregedés forrásainak. A gráf tartalmazza a beneneteket, a központi gén csomópontot, és jelzi, hogy ezen genetikai információ terjesztését tetszőleges kimenetre meg lehet tenni.



19. ábra) Az XML bemeneti adatinak megfelelő gyakorlatban használt hálózata az öregedési adatoknak

Az öregedési adatokat egy 5500 soros XML fájlba összegeztem, mely 26 csomópontot és 53 élt tartalmaz [Függelék 23.ábra]. A modell megfelelő használatához szükséges az élsúlyok beállítása, mely után válik következtetésre használhatóvá. A súlyok beállítását a többi egyszerűbb modellre támaszkodva végeztem el.

5.3 AZ EGÉSZSÉGES ÖREGEDÉS MODELLJE

Az egészséges öregedés modelljének a létrehozásához jelenleg kevés adat áll a rendelkezésre, de szerencsére nemrégiben készült egy teljes genom asszociáción alapuló vizsgálat a témában és annak az adatait használtam fel (39). A cikk csak a statisztikailag legerősebb találatokat közölte kielemezve, valamint elérhetővé tette a variánsok arányait a cikk által definiált egészségesen öregedő európai származású populációval kapcsolatban.

Az variánsok gyakorisági adatait összevetettük az 1000 genom projekt (58) európai felmenőket tartalmazó adataival, és ezek alapján meghatároztuk a kísérletnek megfelelő p és q értékeket. Ezen találatok közül a q érték alapján rendezett első 500-at találatot hasonlítottam az öregedési modellhez.

6 EREDMÉNYEK

A kutatás eredményeképpen sikerült optimalizálni egy több-tárgyterületet átfogó integrációs modellt, amely magában foglalja az öregedéssel kapcsolatos fizikai jegyeket, állatmodelleket, gyógyszereket és emberi genetikai információkat. Továbbá szisztematikus vizsgálatok alapján javaslatot tettem a következtetési módszer optimális paraméterezésére, beleértve a fúziós modellben szereplő relációknak és bizonytalan evidenciák súlytényezőinek a megválasztását. Az integrációs modell szemantikai jellegét felhasználva összevetettem az élethossz változását az egészséges öregedéssel kapcsolatos adatokkal, amely az öregedéskutatás egy nyitott kérdése. Végezetül elvégeztem és megvizsgáltam az öregedéssel összefüggésbe hozható gének, hatóanyagok és útvonalak prioritizálását.

6.1 SZÁMÍTÁSI MÓDSZEREK

A QSFP rendszer számára a korábbi kérdésekben elég volt a cosinus hasonlósági módszer használta, viszont a jelenlegi biológiai tartalommal feltöltött esetben meg kellet vizsgálni egyéb számítási módszereket is, és egy megalapozott elképzelés alapján kellett döntést hozni. Továbbá a hasonlóság mellett szükséges volt egy olyan módszer is, mely nem hasonlóságot számít két entitás között, hanem az egyik lista elemeit megfelelteti a másik lista elemeinek. Erre egy példa a fajközi megfeleltetés esete, ahol a génszekvenciák hasonlósága melett az erősen konzerválódott (ortológ) 1-1 génkapcsolatokat is kezelni kellett.

Ezen kérdések kapcsán a modellbe a cosinus hasonlóságon kívül beépítésre került a Dice (59) és Tanimoto hasonlósági metrikák és a lineáris kernel számítás.

$$K_{cosinus} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$C_{Tanimoto} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i * B_i}$$

$$C_{Dice} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2}$$

$$Lineáris\ kernel = \sum_{i=1}^{n} A_i * B_i$$

A képletekben használt jelölések esetében A és B is vektor, és az n az összes elemük indexe, és ha a j.-edik elem nincsen csak B-vektorban van jelen, akkor A_j értéke 0. Látható továbbá, hogy egyedül a lineáris kernel nem tartalmaz korrekciót a vektorok méretére vonatkozóan, ezért alkalmas adatok átvezetésére egyik csomópontból a másikra.

A számítási és hasonlósági módszereken kívül beépítettünk korrekciós módszert is a rendszerbe, ahol a közismert TF-IDF (term frequency–inverse document frequency) index mindkét tagjait külön-külön és egyben is tudjuk alkalmazni egyes számítási lépésben.

A TF-IDF számításra különböző módszerek vannak kidolgozva ezek közül mi a folytonos értékeket tartalmazó vektorokra az alábbi módon vezettük be, amit egy értékkel rendelkező génlista betegségekre történő számításán keresztül mutatok be.

A betegség-gén lista egy betegséghez köthető géneket tartalmazó vektor, és ha nincsen súlyozva, akkor 1-et tartalmaz azon esetben, ha van kapcsolat és 0-át ha nincs. Súlyozott esetben a betegség-gén értéket tartalmaz, ami a DisGenet alapján 0 és egy között tetszőleges érték lehet. Továbbá mivel minden gén egyszer szerepel egy betegség-gén vektorban, ezért egy gén-betegség kapcsolatra a TF a betegség génjeinek a számának a reciproka, amennyiben nincsen súlyozva, és súlyozott esetben az alábbi képlet írja le:

$$TF = \frac{\sum_{i=1}^{n} Génlista\ érték(Gén_i) * GénBetegség\ érték(Gén_i)}{\sum_{i=1}^{n} GénBetegség\ érték(Gén_i)}.$$

Az IDF azzal arányos, hogy egy gén hány betegségben szerepel, és az értéke gén betegségeinek a számának és az összes betegség számának a hányadosának a logaritmusa.

$$IDF = \lg\left(\frac{\sum_{i=1}^{n} Gén\ \epsilon\ Betegség_i}{\sum_{i=1}^{n} Betegség_i}\right)$$

A TF-IDF pedig a TF és IDF szorzata. Ezeket a korrekciós módszereket a korábbi hasonlóságokkal együtt is lehet használni, viszont nagyobb jelentősége van abban ez esetben, ha lineáris kernel mellett használjuk, mivel ebben az esetben olyan korrekciót tudunk adni a kapott értékeknek, mely nem hasonlósági, hanem az egyes találatokat külön érinti.

6.2 AZ EGYSZERŰ KÉMIAI MODELLEK ÉRTÉKELÉSE

A kémiai modellek több tanulsággal is szolgáltak, az egyik meghatározó az volt, hogy vegyület-célpont lekérdezés esetén olyan találatokat is kaptunk, melyekre nem számítottunk. Ilyenek

voltak a fenotípusos, fajra vagy sejtvonalra vonatkozó adatok. Az eredmény alapján [4. táblázat] látható, hogy a legjobb találatok közül az első 3 nem hagyományos értelemben vett fehérje típusú gyógyszercélpontra vonatkozik. Továbbá az is leolvasható az ábráról, hogy az adott találatok a 6 PPI gyógyszer esetén hányhoz kapcsolódtak, mert lineáris kernellel ezt az értéket kapjuk vissza.

Érték	Azonosító	Típus	Név
5	CHEMBL1697861	PHENOTYPE	Hepatotoxicity
5	CHEMBL372	ORGANISM	Homo sapiens
5	CHEMBL2362975	NO TARGET	No relevant target
4	CHEMBL258	SINGLE PROTEIN	Tyrosine-protein kinase LCK
4	CHEMBL246	SINGLE PROTEIN	Beta-3 adrenergic receptor

4. táblázat) A 6 PPI-célpont prioritizálás eredményének az első 5 találata

Ezt a problémát egyszerűen lehet kezelni, mivel a QSFP megengedi, hogy szűrőt helyezzünk el a lekérdezés bármely szakaszában, ezzel az összesen 220 találatot sikerült 180-ra redukálni. Viszont még így is sok volt a nem specifikus találat. A ChEMBL adatait megvizsgálva láthatóvá vált, hogy számos esetben a kapcsolatot csak nagyon nagy koncentrációjú vegyület esetén írták le, ami már a legtöbb gyógyszer esetében mérgező lenne. Ennek a kiszűrésére megvizsgáltam a koncentrációs adatokat, és nagyon heterogén mértékegységekkel találkoztam, amire a ChEMBL megoldásként egy pchembl értéket kínál, ami bár nem tökéletes, de egységes, és -log(IC50) dimenziójú. Ennek megfelelően további szűrést adtam a lekérdezéshez úgy, hogy a pchembl nagyobb legyen mint 5 tehát az IC50 kisebb legyen, mint 10 µmol. Így 30 találatra sikerült leszűkíteni az eredményt, ami már megfelel az elvártnak. Viszont a 6 PPI-hoz szűrések nélkül csak 3 esetben volt köthető a „proton pumpa” (Kálium-Hidrogén ATP-áz), ami a gyógyszerek ismert célpontja. Szűrés után pedig csak egy esetben kaptam vissza a vért célpontot. Ezt azzal se lehetett javítani, ha megengedőbb voltam a koncentráció határral, mert a 3-ból 2 gyógyszer nem is tartalmazott pchembl értéket, és az eredeti információforrás egyedi értéket használt a lejegyzésére. Továbbá a legjobb találatok között sok volt az ismert számos gyógyszerrel kapcsolatba hozható célpont, mint a Cytochrome P450. Ennek a korrigálásáta a TF-IDF-et használva az eredményeink javultak, az első 10 találatba meg is jelent a várt célpont [5. táblázat].

Érték	Azonosító	Típus	Név
1.00772	CHEMBL2189159	SINGLE PROTEIN	Indoleamine 2,3-dioxygenase 2
0.276074	CHEMBL4158	SINGLE PROTEIN	Fatty acid synthase
0.252466	CHEMBL1743127	SINGLE PROTEIN	Multidrug and toxin extrusion protein 2
0.143446	CHEMBL1743126	SINGLE PROTEIN	Multidrug and toxin extrusion protein 1
0.13429	CHEMBL1743122	SINGLE PROTEIN	Solute carrier family 22 member 2
0.0903284	CHEMBL6113	SINGLE PROTEIN	Phosphoethanolamine/phosphocholine phosphatase
0.04159	CHEMBL3721	SINGLE PROTEIN	Cytochrome P450 2C8
0.0283563	CHEMBL2095173	PROTEIN COMPLEX	***Potassium-transporting ATPase***
0.0213175	CHEMBL4080	SINGLE PROTEIN	Bombesin receptor subtype-3
0.0202409	CHEMBL2095228	PROTEIN COMPLEX	***Potassium-transporting ATPase***

5. táblázat) A PPI-célpont találatok a korrekciók után

Viszont ezek az adatok nem megnyugtatók, ezért az SSRI antidepresszánsokkal is elvégeztem az alábbi lépéseket TF-IDF használatával és anélkül, és azt kaptam, hogy a vártnak megfelelően az SSRI-k ismert célpontját a szerotonin transzporter mindkét esetben megtalálható volt [6. táblázat és Függelék 17. táblázat].

Érték	Azonosító	Típus	Név
10	CHEMBL313	SINGLE PROTEIN	Serotonin transporter
10	CHEMBL228	SINGLE PROTEIN	Serotonin transporter
9	CHEMBL222	SINGLE PROTEIN	Norepinephrine transporter
8	CHEMBL338	SINGLE PROTEIN	Dopamine transporter
7	CHEMBL304	SINGLE PROTEIN	Norepinephrine transporter
6	CHEMBL238	SINGLE PROTEIN	Dopamine transporter
5	CHEMBL240	SINGLE PROTEIN	HERG

6. táblázat) Az SSRI-célpont prioritizálás eredménye TF-IDF nélkül

A célpont adatok alapján viszont így a TF-IDF nagyon visszafogta a találatokat, ami annak köszönhető, hogy a szerotonin transporter az egy számos vegyülethez köthető célpont főleg a feltérképezettsége és nem az aspecificitás miatt.

Érték	Név
0.538162	Premature Ejaculation
0.517927	Personality Disorders
0.501255	Generalized social phobia
0.501255	Negativism in catatonia
0.501255	Blushing
0.501255	Arthralgia of temporomandibular joint
0.501255	Performance anxiety
0.501255	Loneliness in adolescence
0.501255	Anxiety symptoms
0.501255	Stress Disorders, Traumatic, Acute

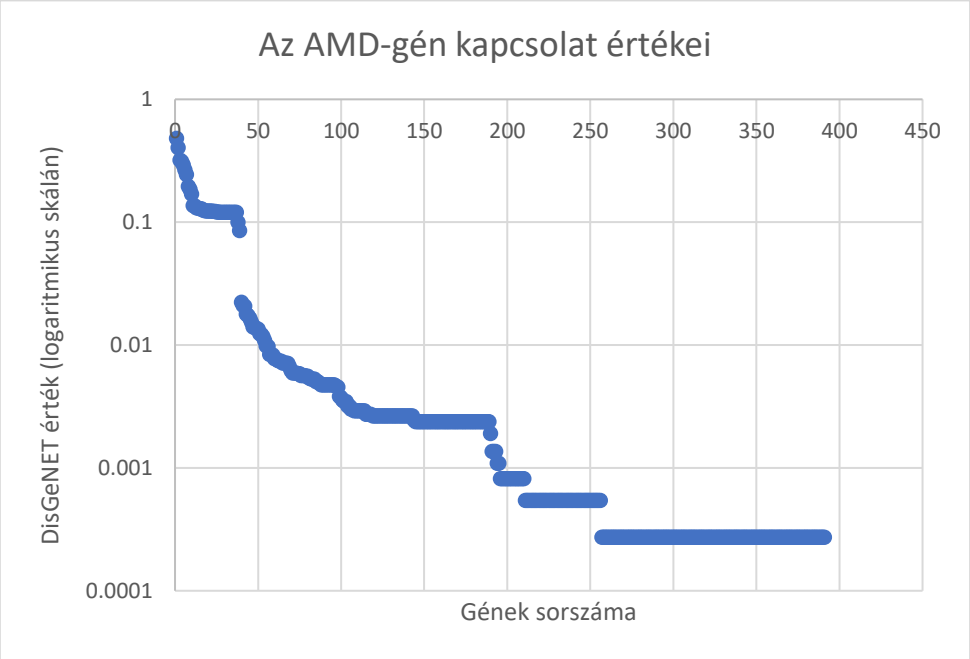
7. táblázat) Az SSRI-betegség prioritizálás eredménye TF-IDF használata nélkül

Az SSRI-Célpont-Gén-Betegség prioritizálás eredménye [7. táblázat] a vártaknak megfelelő eredményt hozott a TF-IDF használata nélkül, a használatával pedig az SSRI-vel össze nem köthető betegségek jöttek kis első találatként. A fenti táblázatban színkóddal vannak jelölve a betegségek és a találatok igen jó összhangban vannak az SSRI hatásaival és mellékhatásaival (az SSRI- hatása a magömlésre jelentősebb, mint a depresszióra vagy a szorongásra gyakorolt hatása).

Összességében azokat a következtetéseket lehet levonni, fehérjékre való szűrés elengedhetetlen, és pchembl érték használata is ajánlott, még ha bizonytalan információ vesztést is jelet. A TF-IDF használatával kapcsolatban kevésbé egyértelmű a helyzet, és hasznos lehet ismerni azt, hogy mennyire feltérképezett célpontokat várunk, és ennek megfelelően megválasztani a számítási súlyozást.

6.3 A MAKULADEGENERÁCIÓS MINTA MODELL ÉRTÉKELÉSE

Az egyszerű megfeleltetések és a kémiai tengely után a makuladegenerációs modellt vizsgáltam meg. Ezzel kapcsolatban az elsődleges célom a gén-betegség tengely feltérképezése volt, és a numerikus értékkel rendelkező kapcsolati mátrixokkal történő számítás beállítása.



20. ábra) Az AMD-gén kapcsolat értékeinek a megoszlása

Az AMD génjeihez tartozó értékek [20. ábra] közül látható, hogy az első 30 rendelkezik 0.1-nél nagyobb értékkel és az értékek többsége ennek a százada vagy annál is kisebb. Emiatt várhatóana numerikus értékeket binárisan kezelő vektorokat használó következtetés drasztikusan más eredményre vezet, mint egy nagyobb értékekre szűrő vagy azokat práló következtetés.

Első lépésben az 5.1.2.1 fejezetnek megfelelően megvizsgáltam, hogy az eset és kontrol génekből álló halmazokra mely számítások adják a legjobb eredményt. Ehhez BASH kódot használtam és szisztematikusan lefuttattam a prioritizálást minden eset-kontroll kombinációra, számítási módszer és TF-IDF kombinációira. Az eredmények csak egy kis részét mutatom be, ahol látható [21. ábra], hogy bináris vektoroknál a Tanimoto jobban teljesített, mint a cosinus hasonlóság és a lineáris kernel magában nem volt használható.

CosineSimilarity				Tanimoto				LinearKernel			
Eset	Kontroll	AMD	Arány	Eset	Kontroll	AMD	Arány	Eset	Kontroll	AMD	Arány
100	200	1	1.71	100	200	1	2.05	100	200		0
100	150	1	1.72	100	150	1	1.94	100	150	7	0.76
100	100	1	1.67	100	100	1	1.79	100	100	3	0.91
150	300	1	1.75	150	300	1	2.16	150	300		0
150	225	1	1.87	150	225	1	2.39	150	225	7	0.75
150	150	1	2.01	150	150	1	2.49	150	150	3	0.88
200	400	1	1.86	200	400	1	2.27	200	400		0
200	300	1	2.03	200	300	1	2.62	200	300	6	0.75
200	200	1	2.24	200	200	1	3.06	200	200	3	0.91
250	500	1	1.96	250	500	1	2.31	250	500		0
250	375	1	2.1	250	375	1	2.63	250	375	5	0.77
250	250	1	2.29	250	250	1	3.14	250	250	3	0.93

21. ábra) A hasonlósági metrikák összehasonlítása. Az arány az AMD és a második találat értékének az aránya, ha az AMD az első, és az AMD és az első érték aránya, ha nem az AMD az első találat.

Összességében az a sorrend állt elő, hogy a Tanimoto mérték hozta a legnagyobb különbséget az AMD és az utána jövő betegség értéke között, ezt követte a Dice, és a cosinus hasonlóság és végül a lineáris kernel végzett. Viszont minden módszer más módon volt erős, például az IDF javított a cosinus hasonlóságon, de a Tanimoto-n rontott. Konklúzióként az adódott, hogy a legjobb kombináció a Tanimoto volt magában, utána a Dice magában és végül a cosicus +IDF.

Az értékekkel történő tesztelésre egy valószerűbb felállást választottam, ahol a makuladegenerációhoz felállítottam egy modell, ahol a gének már értékekkel rendelkeznek,

és ehhez viszonyítottam a módszereket. Bár ez kevésbé szisztematikus de a valóságnak jobban megfelelő módszer, mert általában ennél kevesebb adat áll a rendelkezésünkre.

A makuladegenerációs modell GWAS eredetű genetikai tengelyéhez úgy nyertem értékeket, hogy a Fisher módszerhez hasonlóan a GWAS kísérletek eredményeinek a p-értékeinek a logaritmusát vettem, és ezek értékeit összeadtam. Ezt követően az értékeket 1-re normáltam, hogy a gén-betegség értékeknek eloszlásának minél jobban megfeleljenek. A vegyület tengelyről származtatott gének a számítási lánc miatt értékekkel rendelkeznek, és az útvonalak esetében, ha nem is optimálisan, de a négy útvonalnak köszönhetően az átfedések egy négyes skálát eredményeznek. A skálákat egyre normálva adtam össze, ami logaritmikus dimenzióban egy bevett módszer, és az eredményeket betegségekre prioritizáltam.

Cosinus + IDF		Tanimoto	
Érték	**Betegség**	**Érték**	**Betegség**
0.303	Stargardt's disease	0.104	Age related macular degeneration
0.283	Macular degeneration	0.071	Macular Degeneration, Age-Related, 1
0.263	Age related macular degeneration	0.058	Atherosclerosis
0.252	STARGARDT DISEASE 1 (disorder)	0.052	Atypical Hemolytic Uremic Syndrome
0.248	Low Vision	0.051	Pulmonary Fibrosis
0.214	Atypical Hemolytic Uremic Syndrome	0.051	Diabetic Nephropathy
0.213	Geographic Atrophy	0.051	Brain Ischemia
0.202	Macular Degeneration, Age-Related, 1	0.050	Diabetes Mellitus, Experimental
0.200	Arthritis, Viral	0.049	Reperfusion Injury
0.196	Premature Obstetric Labor	0.049	Infarction, Middle Cerebral Artery

8. táblázat) Az MD modell eredménye cosinus és Tanimoto hasonlósággal, színkóddal

Az eredmények közel sem egyértelműek, mert bár a kérdéses időskori makuladegenerációt a Tanimoto megközelítéssel sokkal jobban vissza lehetett kapni, ezzel szemben a cosinus hasonlóságnál egy széles spalettája látható az AMD-hez köthető betegségeknek (a Stargardt's disease egy örökletes fiatalkori MD). A Dice módszerrel hasonló sorrend állt elő, mint a Tanimotónál, viszont az értékek kevésbé voltak kiemelkedők.

Összességében tehát itt is elmondható, hogy a céloknak az ismeretében érdemes megválasztani a hasonlósági metrikát, például általános felderítő elemzésekre a cosinus hasonlóság célzott kereséshez pedig a Tanimoto ajánlott.

6.4 AZ ÖREGEDÉSI MODELL ÉRTÉKELÉSE

Az egyszerűbb modellek segítségével sikerült az altalános öregedési modell beállítása, amely nagyvonalakban most már a vártnak megfelelő eredményeket ad vissza. Az öregedési modellben elvégzett prioritizálás eredményeként a betegség tengely első 20 találata orvosbiológai szempontból adekvát, az öregedéskutatási eredményekkel egybecsengő eredményt adott, melyet a színskála is jelez.

Cosinus + IDF		Tanimoto	
Érték	**Betegség**	**Érték**	**Betegség**
0.226	Diabetes Mellitus, Experimental	0.00836	Mammary Neoplasms
0.211	Breast Carcinoma	0.00788	Diabetes Mellitus, Experimental
0.176	Hypertensive disease	0.00770	Diabetes Mellitus, Non-Insulin-Dependent
0.174	Squamous cell carcinoma	0.00757	Hypertensive disease
0.172	Malignant neoplasm of breast	0.00724	Prostatic Neoplasms
0.172	Mammary Neoplasms	0.00677	Liver carcinoma
0.172	Liver carcinoma	0.00666	Obesity
0.168	Tumor Progression	0.00546	Alzheimer's Disease
0.166	Malignant tumor of colon	0.00545	Schizophrenia
0.164	Cancer Cell Growth	0.00475	Stomach Neoplasms
0.162	Prostatic Neoplasms	0.00464	Reperfusion Injury
0.162	Secondary malignant neoplasm of lung	0.00445	Myocardial Infarction
0.162	Reperfusion Injury	0.00433	Malignant neoplasm of breast
0.161	Malignant neoplasm of endometrium	0.00428	Lung Neoplasms
0.161	Uterine Corpus Cancer	0.00418	Neoplasm Metastasis
0.160	Non-Neoplastic Disorder	0.00394	Rheumatoid Arthritis
0.157	Ovarian Carcinoma	0.00383	IGA Glomerulonephritis
0.156	Lung Neoplasms	0.00382	melanoma
0.155	Colorectal Carcinoma	0.00381	Peripheral Neuropathy
0.155	Carcinogenesis	0.00374	Autistic Disorder

9. táblázat) Az öregedési modell-betegség prioritizálás eredménye két hasonlósági módszerrel

Ezek azt sugallják, hogy a rendszer a vártank megfelelően működik, viszont érdemes megjegyezni, hogy az itt elöl szereplő betegségek gyakoriak és akár egy általános képre is utalhatnak, aminek a kizárására további elemzésekre van szükség. Továbbá itt érdemes a további eredményeket is megvizsgálni, amelyek hasonló feldúsulást mutatnak.

A fenotípus tengely prioritizálása során a várakozásoknak kevésbé megfelelő eredményeket kaptam [10. táblázat], de még elfogadhatónak tartom a kapott eredményeket annak ellenére, hogy a későbbi találatok is hasonló arányban tartalmaztak öregedéssel kapcsolatba hozható fenotípusos jegyeket.

Érték	Fenotípus
0.002919	Decreased resting energy expenditure
0.00218	Subacute progressive viral hepatitis
0.001664	Social and occupational deterioration
0.00164	Pericarditis
0.00116	Pleuritis
0.001078	Enlarged polycystic ovaries
0.001041	Folate deficiency
0.000883	Impaired ability to form peer relationships
0.000883	Beta-cell dysfunction
0.000875	Aseptic leukocyturia
0.000859	Late onset
0.000833	Uterine leiomyosarcoma
0.000803	Long-tract signs
0.00071	Insulin resistance
0.000708	Alveolar cell carcinoma

10. táblázat) Az öregedési fenotípus prioritizálás eredménye

Továbbá elvégeztem a vegyületekre történő prioritizálást, melynek az eredménye alább látható. A vegyületekre vonatkozó adatoknak leginkább további kísérletes validációval együtt van jelentősége és önmagában inkább tájékoztató jellegű.

Érték	Vegyület
225.194	1H-Indole-3-carboxylic acid 4-oxo-2-aza-tricyclo[3.3.1.0*2,7*]non-8-yl ester
169.044	2-(3-Chloro-phenyl)-3-pyridin-4-yl-1H-pyrrolo[2,3-b]pyridin-6-ylamine
169.044	2-(3,4-Difluoro-phenyl)-3-pyridin-4-yl-1H-pyrrolo[2,3-b]pyridin-6-ylamine
169.044	4-{3-[4-(4-Fluoro-phenyl)-5-(6-methoxy-pyrimidin-4-yl)-imidazol-1-yl]-propyl}-morpholine
169.044	5-(4-Fluoro-phenyl)-2-(4-methanesulfinyl-benzylsulfanyl)-4-pyridin-3-yl-pyrimidine
169.044	1-(5-tert-Butyl-2-o-tolyl-2H-pyrazol-3-yl)-3-phenyl-urea
169.044	5-(4-Fluoro-phenyl)-2-(4-methanesulfinyl-benzylsulfanyl)-4-pyridin-4-yl-pyrimidine
169.044	1-[5-tert-Butyl-2-(3,4-dimethyl-phenyl)-2H-pyrazol-3-yl]-3-phenyl-urea
169.044	2-(4-Fluoro-phenyl)-3-pyridin-4-yl-1H-indole
169.044	6-(2-chloro-phenyl)-9-(2,4-difluoro-phenyl)-7,9-dihydro-purine-8-one

11. táblázat) Az öregedés vegyület prioritizálás első 10 eredménye

A vegyületek esetében megnéztem, hogy az ismert vegyületek mennyire dúsulnak fel az eredmények között, de azt találtam, hogy nem volt jelentős feldúsulás, ami a módszer további finomítását és szisztematikus módszerekkel végzett beállítását teszi szükségessé. Ezen kívül a kémiai tengely bizonytalanságainak a kiküszöbölésére a ComBine kutatócsoportban már folynak fejlesztések, mint a hasonlósági mátrix automatikus kitöltése és további adatbázisok integrálása.

6.5 AZ EGÉSZSÉGES ÖREGEDÉS ÉRTÉKELÉSE

Az egészséges öregedés génjeinek az Ensembl azonosítóra történő megfeleltetésekor az 500 génből 480-at lehetett megtalálni. Ezen gének hasonló átfedést mutatnak az öregedés közvetett és közvetlen génjeivel [22. ábra].

22. ábra) Az öregedés és egészséges öregedés génjeinek az összehasonlítása

Az egészséges öregedés génjeire elvégeztem a gén-betegség prioritizálást és az alábbi táblázatot kaptam [12. táblázat], mely nem hordoz azon kívül jelentős tartalmat, hogy idegi betegségeket jelöl meg az első találatok között; ahogyan az eredeti cikk szerzői is a neurodegeneratív betegségekkel tudták a legjobban összekötni az eredményeiket. A laktózintolerancia, mint találat, bár érdekes kérdést vet fel, de GWAS elemzéshez használt kontroll és a minta közti eltérés akár magában is okozhatja ezt a kiemelkedő eredményt.

Érték	Név
0.098854	LACTASE PERSISTENCE
0.097996	Lactose Intolerance, Adult Type
0.090662	Lactose Intolerance
0.085326	Lactase Deficiency, Congenital
0.083972	MENTAL RETARDATION
0.080959	Recurrent Meningioma
0.080268	Dicarboxylicaminoaciduria
0.080268	SCHIZOPHRENIA 18
0.080127	Warburg Sjo Fledelius syndrome
0.074394	Impaired cognition

12. táblázat) Az egészséges öregedés génjeinek betegségre történő prioritizálása

Továbbá a betegségek kapcsán megvizsgáltam, hogy amennyiben az öregedés génjeivel együtt futtatom a lekérdezést az mennyiben változtat ezen a képen. Az adódott, hogy az általános öregedési eredmények lettek a legmeghatározóbbak, ezért azt is megnéztem, hogy mely betegségek lettek felülreprezentálva abban az esetben, ha az egészséges öregedés génjeivel együtt futtattam a lekérdezést, ld. [Függelék 18. táblázat]. Megemlítendő találat a vitaminokkal kapcsolatban volt, mert a B12 vitamin fontos szerepet játszik az agyi károsodásokkal kapcsolatos ellenállásban, ami esetlegesen szerepet játszhat az egészséges öregedési folyamatokban.

7 ÖSSZEFOGLALÓ

Az öregedés jelensége a korfüggő betegségek közös gyökerén túl önállóan is egyre nagyobb hangsúlyt kap az orvosbiológiai és gyógyszerészeti kutatásokban. Azonban az öregedés, vagy azon belül is az egészséges öregedés remélt védőfaktorainak a kutatása is egy rendkívül heterogén problémával szembesül, leegyszerűsítve az összes korfüggő betegség együttesét kellene, hogy elemezze és azok közös vonásait felderítse.

Kutatásom fő célja egy átfogó, heterogén információkat integráló öregedéskutatási modell megalkotása volt Kapcsolt Nyilvános Adatok (Linked Open Data) felhasználásával. Elsőként áttekintettem az öregedéskutatás főbb dimenziót és szisztematikus adatgyűjtést végeztem a szemantikus technológiákat használó adatbázisokon. Ennek eredményeképpen egységes formátumú, kvantitatív evidenciákat konstruáltam az egyes öregedéselméleti „tengelyek" mentén. Másodsorban egy olyan átfogó, heterogén információkat integráló öregedéskutatási modellt alkottam, amely humángenetikai, gyógyszerkutatási, bioinformatikai, klinikai és állatmodellekből is származó információkat tud integrálni szemantikusan átlátható és kontrollálható formában. Harmadrészt, a modellt és a hátteret biztosító következtetési keretrendszert almodelleken és tesztlekérdezések sokaságán keresztül vizsgáltam, mind a következtetés tökéletesítése, mind valós lekérdezések finomítása végett.

Összességében a munkámmal egy nemzetközi szinten is új, átfogó szemantikai öregedéskutatási modellt és hozzátartozó evidenciákat hoztam létre, amely releváns lekérdezésekben is értelmezhető és érdekes eredményeket ad orvosbiológiai és gyógyszerkutatási szempontokból is.

Továbbá, vizsgálataimmal hozzá tudtam járulni a ComBine laboratórium kvantitatív, szemantikai következtetési keretrendszerének a fejlődéséhez, például a számítási módszerek beállítására tett javaslatokkal, illetve gyakorlati példákon keresztül be tudtam mutatni a használatát előnyeit és korlátait. A folyamatban levő és jövőbeli továbblépési lehetőségek az élek számítási módszerének és az egymáshoz viszonyított súlyainak a finomhangolása automatikus esetlegesen keresztvalidációs módszer felhasználásával. Továbbá a várhatóak olyan öregedési fajközi eredmények is melyeket a bemutatott keretrendszeren és a gyűjtött adatokhoz hasonlítva tervezünk elemezni.

8 KÖSZÖNETNYILVÁNÍTÁS

Szeretnék köszönetet mondani Millinghoffer Andrásnak a GWAS adatok elemzésével kapcsolatban, Gézsi Andrásnak a folyamatos ötletelésért és ajánlásaim folyamatos kivitelezéséért és integrálásáért a rendszeren, továbbá Antal Péternek a folyamatos segítségéért, tanácsaiért és támogatásáért.

9 FÜGGELÉK

Rang	Azonosító	Érték	Név
1	ENSG00000000971	0.48	ARMD4|ARMS1|FHL1|HF
2	ENSG00000254636	0.403743	ARMD8|LOC387715
3	ENSG00000243649	0.319045	BF|BFD|H2-Bf
4	ENSG00000125730	0.311039	ARMD9|C3a|C3b|CPAMD1
5	ENSG00000166278	0.292322	
10	ENSG00000198691	0.168835	ABCR|ARMD2|CORD3|FFM|RP19|STGD|STGD1
20	ENSG00000144810	0.12291	C3orf7|MGC9568
30	ENSG00000168477	0.120271	TNXB1|TNXB2|TNXBS|XB|XBS
40	ENSG00000107679	0.0222794	TAPP1
50	ENSG00000005421	0.0134638	ESA|PON
100	ENSG00000084674	0.00372424	
391	ENSG00000132781	0.000271442	MYH

13. táblázat) A AMD-hez köthető génlistának és értékeinek a kivonata

Egér Modellek		MGD		RGD			
Gén azonosító	Gén név	Gén azonosító	Gén név	Gén azonosító	Gén név	Gén azonosító	Gén név
ENSMUSG00000002944	Cd36	ENSMUSG00000000724	Cryba1	ENSRNOG00000000053	Crp	ENSRNOG00000012467	Il17a
ENSMUSG00000002985	Apoe	ENSMUSG00000002985	Apoe	ENSRNOG00000000420	Nelfe	ENSRNOG00000012772	Nqo1
ENSMUSG00000003617	Cp	ENSMUSG00000006205	Htra1	ENSRNOG00000000521	Cdkn1a	ENSRNOG00000012892	Abca4
ENSMUSG00000006818	Sod2	ENSMUSG00000016493	Cd46	ENSRNOG00000000940	Flt1	ENSRNOG00000013736	C9
ENSMUSG00000007891	Ctsd	ENSMUSG00000020212	Mdm1	ENSRNOG00000001111	Brca2	ENSRNOG00000014187	Igf1r
ENSMUSG00000015839	Nfe2l2	ENSMUSG00000022149	C9	ENSRNOG00000001414	Serpine1	ENSRNOG00000016229	Gltscr1l
ENSMUSG00000020609	Apob	ENSMUSG00000024164	C3	ENSRNOG00000001469	Eln	ENSRNOG00000016957	Igfbp2
ENSMUSG00000022982	Sod1	ENSMUSG00000024371	C2	ENSRNOG00000001783	Tra2b	ENSRNOG00000017539	Mmp9
ENSMUSG00000024924	Vldlr	ENSMUSG00000024924	Vldlr	ENSRNOG00000002115	Sod1	ENSRNOG00000017753	Ercc2
ENSMUSG00000026365	Cfh	ENSMUSG00000026365	Cfh	ENSRNOG00000003068	Mrnip	ENSRNOG00000018461	Pdgfrb
ENSMUSG00000031209	Heph	ENSMUSG00000027447	Cst3	ENSRNOG00000003098	Prom1	ENSRNOG00000019142	Fas
ENSMUSG00000032193	Ldlr	ENSMUSG00000028125	Abca4	ENSRNOG00000003553	Efemp1	ENSRNOG00000019358	Esr1
ENSMUSG00000035352	Ccl12	ENSMUSG00000029086	Prom1	ENSRNOG00000004143	Adipor1	ENSRNOG00000019598	Vegfa
ENSMUSG00000040552	C3ar1	ENSMUSG00000032262	Elovl4	ENSRNOG00000004473	Ppargc1a	ENSRNOG00000020129	Cdh3
ENSMUSG00000049103	Ccr2	ENSMUSG00000035385	Ccl2	ENSRNOG00000004517	Igf1	ENSRNOG00000020497	Plekha1
ENSMUSG00000049130	C5ar1	ENSMUSG00000039005	Tlr4	ENSRNOG00000006084	Cngb3	ENSRNOG00000020533	Htra1
ENSMUSG00000052336	Cx3cr1	ENSMUSG00000040268	Plekha1	ENSRNOG00000007159	Ccl2	ENSRNOG00000021147	Bad
		ENSMUSG00000049103	Ccr2	ENSRNOG00000007249	Cdkn1b	ENSRNOG00000021726	Tlr3
		ENSMUSG00000052336	Cx3cr1	ENSRNOG00000007286	Mdm1	ENSRNOG00000022619	Fth1
		ENSMUSG00000054051	Ercc6	ENSRNOG00000007613	C1qtnf5	ENSRNOG00000026866	Syn3
		ENSMUSG00000056494	Cngb3	ENSRNOG00000007992	Srsf10	ENSRNOG00000029707	Mt-nd4
		ENSMUSG00000057037	Cfhr1	ENSRNOG00000008553	Mthfr	ENSRNOG00000030017	Ercc6
		ENSMUSG00000058952	Cfi	ENSRNOG00000010278	Il6	ENSRNOG00000031053	Mt-nd4l
		ENSMUSG00000066842	Hmcn1	ENSRNOG00000010522	Tlr4	ENSRNOG00000034066	Hspa8
		ENSMUSG00000090231	Cfb	ENSRNOG00000011853	Mbd2	ENSRNOG00000035644	Mir23a

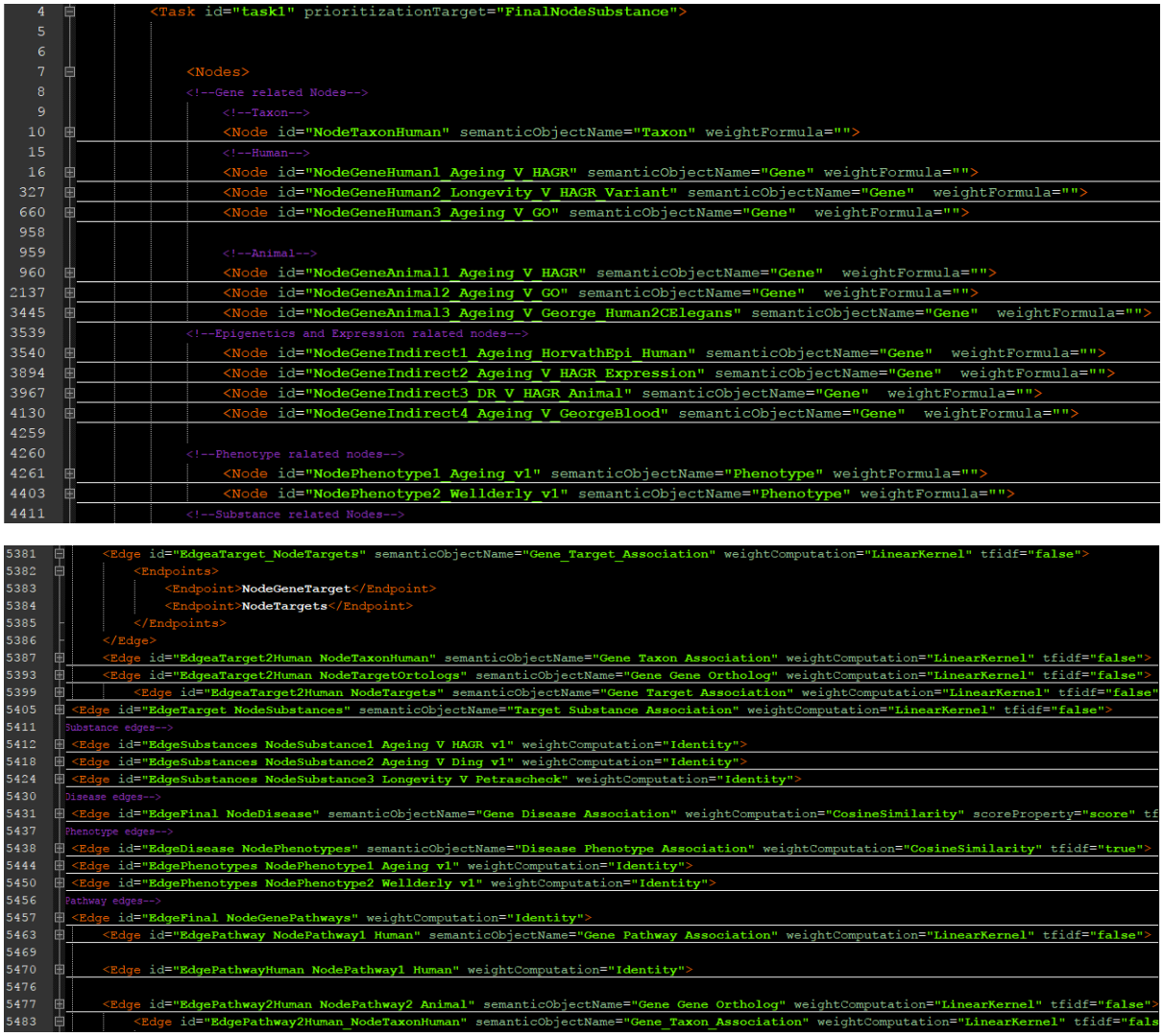
14. táblázat) Az állati modell gének listája forrásokra bontva

Azonosító	Érték	Név
CHEMBL384467	1	DEXAMETHASONE
CHEMBL460026	0.8	eicosapentaenoic acid
CHEMBL2359248	0.8	CHEMBL2359248
CHEMBL173929	0.8	LUTEIN
CHEMBL218490	0.6	Dorzolamide
CHEMBL989	0.6	Dermatin
CHEMBL499	0.6	Istalol
CHEMBL413	0.6	rapamycin
CHEMBL33864	0.6	LIPOIC ACID
CHEMBL269732	0.6	Ascomycin analogue
CHEMBL267936	0.6	MECAMYLAMINE
CHEMBL1009	0.6	L-Dopamine
CHEMBL1979448	0.6	CHEMBL1979448
CHEMBL192	0.6	Sildenafil
CHEMBL1908360	0.6	Afinitor
CHEMBL157101	0.6	CHEMBL157101
CHEMBL1456	0.6	Cellcept
CHEMBL1451	0.6	Fluoxiprednisolone
CHEMBL1201236	0.6	CARBIDOPA
CHEMBL1234071	0.6	CHEMBL1234071
CHEMBL1324508	0.6	CHEMBL1324508
CHEMBL1431	0.6	metformin
CHEMBL134342	0.6	Thioctic acid
CHEMBL2107821	0.4	EMIXUSTAT
CHEMBL75	0.4	ketoconazole
CHEMBL2141712	0.4	P-529
CHEMBL32	0.2	Moxifloxacin
CHEMBL295698	0.2	KETOCONAZOLE
CHEMBL469	0.2	ketorolac
CHEMBL278172	0.2	Benzocaine
CHEMBL679	0.2	Epinephrine
CHEMBL79	0.2	Embolex
CHEMBL1196	0.2	Alcaine

15. táblázat) Az MD gyógyítására tervezett klinikai fázisban levő gyógyszerek és ennek megfelelő értékeik

Azonosító	Fenotípus
http://purl.obolibrary.org/obo/HP_0000144	Female fertility decline
http://purl.obolibrary.org/obo/HP_0000488	Retinopathy
http://purl.obolibrary.org/obo/HP_0000518	Cataract
http://purl.obolibrary.org/obo/HP_0000540	Presbyopia
http://purl.obolibrary.org/obo/HP_0000716	Depression
http://purl.obolibrary.org/obo/HP_0001058	Wound healing
http://purl.obolibrary.org/obo/HP_0001757	Hearing high-frequency sounds
http://purl.obolibrary.org/obo/HP_0011364	Hair turns white
http://purl.obolibrary.org/obo/HP_0002218	Hair turns grey
http://purl.obolibrary.org/obo/HP_0002354	Memory loss
http://purl.obolibrary.org/obo/HP_0002355	Walking behavior
http://purl.obolibrary.org/obo/HP_0002621	Atherosclerosis
http://purl.obolibrary.org/obo/HP_0002721	Immune deficiency
http://purl.obolibrary.org/obo/HP_0002758	Osteoarthritis
http://purl.obolibrary.org/obo/HP_0003199	Sarcopenia
http://purl.obolibrary.org/obo/HP_0003380	Myelinated axon length decline
http://purl.obolibrary.org/obo/HP_0003758	Subcutaneous fat loss
http://purl.obolibrary.org/obo/HP_0005978	Diabetes
http://purl.obolibrary.org/obo/HP_0007488	Skin atrophy
http://purl.obolibrary.org/obo/HP_0008587	Hearing loss
http://purl.obolibrary.org/obo/HP_0008763	Isolation
http://purl.obolibrary.org/obo/HP_0030515	Visual impairment
http://purl.obolibrary.org/obo/HP_0031177	Grip strength decrease
ilike "muscular weakness"	Frailty
ilike "atrophy"	General organ atrophy
Ilike "dementia"	Dementia

16. táblázat) Az öregedési fenotípusos jegyek listája



23.ábra) Az XML bemenet összevonásokkal, és sorszámozással jelezve az egyes összevonások mértékét

Érték	Azonosító	Típus	Név
0.497931	CHEMBL2818	SINGLE PROTEIN	P2X purinoceptor 4
0.458068	CHEMBL2799	SINGLE PROTEIN	Dopamine transporter
0.449495	CHEMBL6020	SINGLE PROTEIN	Bile salt export pump
0.348487	CHEMBL2370	SINGLE PROTEIN	Norepinephrine transporter
0.298759	CHEMBL5114	SINGLE PROTEIN	Quinolone resistance protein norA
0.251338	CHEMBL2221347	PROTEIN COMPLEX	Voltage-gated potassium channel
0.213399	CHEMBL2104	SINGLE PROTEIN	P2X purinoceptor 4
0.133295	CHEMBL6184	SINGLE PROTEIN	Transporter
0.118654	CHEMBL3762	SINGLE PROTEIN	Voltage-gated L-type calcium channel
0.0561252	CHEMBL315	SINGLE PROTEIN	Alpha-1b adrenergic receptor
0.03905	CHEMBL319	SINGLE PROTEIN	Alpha-1a adrenergic receptor
0.0359926	CHEMBL304	SINGLE PROTEIN	Norepinephrine transporter
0.0331872	CHEMBL2035	SINGLE PROTEIN	Muscarinic acetylcholine receptor M5
0.0290044	CHEMBL1942	SINGLE PROTEIN	Alpha-2b adrenergic receptor
0.0246596	CHEMBL1980	SINGLE PROTEIN	Sodium channel protein type V alpha subunit
0.0241703	CHEMBL2095159	PROTEIN FAMILY	Serotonin 1 (5-HT1) receptor
0.0232781	CHEMBL1821	SINGLE PROTEIN	Muscarinic acetylcholine receptor M4
0.023022	CHEMBL222	SINGLE PROTEIN	Norepinephrine transporter
0.019634	CHEMBL313	SINGLE PROTEIN	Serotonin transporter
0.0189176	CHEMBL338	SINGLE PROTEIN	Dopamine transporter
0.0176212	CHEMBL228	SINGLE PROTEIN	Serotonin transporter
0.0170019	CHEMBL3251	SINGLE PROTEIN	Nuclear factor NF-kappa-B p105 subunit
0.0167956	CHEMBL238	SINGLE PROTEIN	Dopamine transporter
0.01247	CHEMBL324	SINGLE PROTEIN	Serotonin 2c (5-HT2c) receptor
0.0123498	CHEMBL216	SINGLE PROTEIN	Muscarinic acetylcholine receptor M1
0.0121772	CHEMBL287	SINGLE PROTEIN	Sigma opioid receptor
0.0118415	CHEMBL4608	SINGLE PROTEIN	Melanocortin receptor 5
0.0116158	CHEMBL1833	SINGLE PROTEIN	Serotonin 2b (5-HT2b) receptor
0.0114491	CHEMBL1916	SINGLE PROTEIN	Alpha-2c adrenergic receptor
0.0112415	CHEMBL2093864	PROTEIN FAMILY	Adrenergic receptor alpha-2
0.0089154	CHEMBL211	SINGLE PROTEIN	Muscarinic acetylcholine receptor M2
0.00877022	CHEMBL245	SINGLE PROTEIN	Muscarinic acetylcholine receptor M3
0.00854933	CHEMBL1867	SINGLE PROTEIN	Alpha-2a adrenergic receptor
0.00792904	CHEMBL225	SINGLE PROTEIN	Serotonin 2c (5-HT2c) receptor
0.00744811	CHEMBL240	SINGLE PROTEIN	HERG
0.00729357	CHEMBL224	SINGLE PROTEIN	Serotonin 2a (5-HT2a) receptor
0.00721115	CHEMBL4302	SINGLE PROTEIN	P-glycoprotein 1
0.00706016	CHEMBL223	SINGLE PROTEIN	Alpha-1d adrenergic receptor
0.0066198	CHEMBL231	SINGLE PROTEIN	Histamine H1 receptor
0.00564903	CHEMBL276	SINGLE PROTEIN	Muscarinic acetylcholine receptor M1

17. táblázat) Az SSRI-célpont prioritizálás találatai TF-IDF használatával

Betegség	Feldúsulás
VITAMIN B12 PLASMA LEVEL QUANTITATIVE TRAIT LOCUS 1	5092.890538
Caliciviridae Infections	5092.890538
VITAMIN K-DEPENDENT CLOTTING FACTORS,	1986.890319
Opsismodysplasia	1908.603933
MIRROR MOVEMENTS 1	731.3790244
Usher syndrome, type 2C	612.6308649
Pit and fissure caries	483.9759361
Guanidinoacetate methyltransferase deficiency	474.4881546
Multiple fibrofolliculomas	470.8676172
FEBRILE CONVULSIONS, FAMILIAL, 4	458.6829178
Barakat syndrome	401.7444953
BLOOD GROUP, JUNIOR SYSTEM	313.6636651
Enteritis due to Norovirus	301.366195
Chylomicron retention disease	167.5493227
Coumarin Resistance	124.4630353
EFAVIRENZ, POOR METABOLISM OF	114.561801
PNEUMOTHORAX, PRIMARY SPONTANEOUS	111.474181
Pneumothorax	76.56703508
FIBROSIS OF EXTRAOCULAR MUSCLES, CONGENITAL, 2	58.34315404
Mental Retardation, Autosomal Dominant 5	51.24225173
Fibrofolliculoma	44.59371771
MENTAL RETARDATION, AUTOSOMAL RECESSIVE 46	44.59092956
Achromatopsia 3	39.31439002
Diabetes Mellitus, Neonatal, with Congenital Hypothyroidism	35.71542887
Spastic paraplegia 8, autosomal dominant	29.69453426
Familial Hemophagocytic Lymphocytosis	29.52897777
Osteopetrosis, mild autosomal recessive form	28.95650884
Intestinal Atresia	25.87169143
HYPOTONIA, INFANTILE, WITH PSYCHOMOTOR RETARDATION	24.67007645
Cystinuria	24.01706884
CORONARY HEART DISEASE, SUSCEPTIBILITY TO, 6	23.74145366
Staphylococcal Food Poisoning	23.03963779
Cystinuria, Type B	21.4558386
HHH syndrome	21.39316235
DEAFNESS, AUTOSOMAL RECESSIVE 2	20.69577182
De Barsy syndrome	19.75485201
Tuberculosis, Bovine	18.46410546
Diabetes Mellitus, Transient Neonatal, 1	17.00236035
Deafness, Autosomal Dominant 11	16.93239721

18. táblázat) Az egészséges öregedésben kiemelkedő betegségek az öregedéssel együtt lefuttatva, az öregedéshez viszonyítva

10 IRODALOMJEGYZÉK

1. Jinha AE. Article 50 million: an estimate of the number of scholarly articles in existence. Learned Publishing. 2010;23(3):258-63.
2. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2016;44(Database issue):D7-D19.
3. Pilioura T, Tsalgatidou A. Unified publication and discovery of semantic web services. ACM Transactions on the Web (TWEB). 2009;3(3):11.
4. Sowa JF. Semantic networks. John_Florian_Sowa isi [2012-04-20 16: 51]> Author [2012-04-20 16: 51]. 2012.
5. Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific american. 2001;284(5):28-37.
6. Decker S, Melnik S, Van Harmelen F, Fensel D, Klein M, Broekstra J, et al. The semantic web: The roles of XML and RDF. IEEE Internet computing. 2000;4(5):63-73.
7. W3C. W3C SEMANTIC WEB ACTIVITY. 2001.
8. Krishnaveni M, Sivaram MR. A Framework Using Semantic Web for Web-Based E Learning System. 2016.
9. Guha RV, Brickley D, Macbeth S. Schema. org: Evolution of structured data on the web. Communications of the ACM. 2016;59(2):44-51.
10. Ltd. N. September 2017 Web Server Survey 2017 [Available from: https://news.netcraft.com/archives/category/web-server-survey/.
11. W3C. Az OWL Web Ontológia Nyelv – Útmutató 2004 [Available from: http://www.w3c.hu/forditasok/OWL/REC-owl-guide-20040210.html.
12. W3C. Az RDF bevezető tankönyve 2005 [Available from: http://www.w3c.hu/forditasok/RDF/REC-rdf-primer-20040210.html.
13. Berners-Lee T. Giant global graph. Decentralized Information Group. 2007:29.
14. W3C. SPARQL Query Language for RDF 2008 [Available from: https://www.w3.org/TR/rdf-sparql-query/.
15. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts. 2009:205-27.
16. Andrejs Abele JM. The Linking Open Data cloud diagram 2017 [Available from: http://lod-cloud.net/.
17. Dumontier M, Callahan A, Cruz-Toledo J, Ansell P, Emonet V, Belleau F, et al., editors. Bio2RDF release 3: a larger connected network of linked data for the life sciences. Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272; 2014: CEUR-WS. org.
18. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF platform: linked open data for the life sciences. Bioinformatics. 2014;30(9):1338-9.
19. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic acids research. 2016;45(D1):D635-D42.
20. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015.
21. Queralt-Rosinach N, Piñero J, Bravo À, Sanz F, Furlong LI. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. Bioinformatics. 2016;32(14):2236-8.
22. Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS biology. 2008;6(7):e184.
23. Lipinski CA, Litterman NK, Southan C, Williams AJ, Clark AM, Ekins S. Parallel worlds of public and commercial bioactive chemistry data: Miniperspective. Journal of medicinal chemistry. 2015;58(5):2068.

24. PubChem. PubChemRDF Release Notes 2017 [Available from: https://pubchem.ncbi.nlm.nih.gov/rdf/.
25. Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2(8):e124.
26. Moonesinghe R, Khoury MJ, Janssens ACJ. Most published research findings are false—but a little replication goes a long way. PLoS medicine. 2007;4(2):e28.
27. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery. 2011;10(9):712-.
28. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Statistics in medicine. 2015;34(28):3769-92.
29. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nature methods. 2015;12(9):841-3.
30. Gefen A, Cohen R, Birk OS. Syndrome to gene (S2G): in-silico identification of candidate genes for human diseases. Human mutation. 2010;31(3):229-36.
31. Le D-H, Pham V-H. HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. BMC Systems Biology. 2017;11(1):61.
32. Turner FS, Clutterbuck DR, Semple CA. POCUS: mining genomic sequence annotation to predict disease genes. Genome biology. 2003;4(11):R75.
33. Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, Cicek AE, et al. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Molecular autism. 2014;5(1):22.
34. Pers TH, Dworzyński P, Thomas CE, Lage K, Brunak S. MetaRanker 2.0: a web server for prioritization of genetic variation data. Nucleic acids research. 2013;41(W1):W104-W8.
35. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature biotechnology. 2006;24(5):537-44.
36. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic acids research. 2009;37(suppl_2):W305-W11.
37. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics. 2010;11(1):255.
38. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153(6):1194-217.
39. Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-genome sequencing of a healthy aging cohort. Cell. 2016;165(4):1002-11.
40. Kauwe JS, Goate A. Genes for a 'Wellderly'Life. Trends in molecular medicine. 2016;22(8):637-9.
41. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research. 2013;42(D1):D1091-D7.
42. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. Bioinformatics. 2010;26(20):2647-8.
43. Black JR, Clark SJ. Age-related macular degeneration: genome-wide association studies to translation. Genetics in Medicine. 2015;18(4):283-9.
44. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research. 2017;45(D1):D896-D901.
45. Pennesi ME, Neuringer M, Courtney RJ. Animal models of age related macular degeneration. Molecular aspects of medicine. 2012;33(4):487-509.
46. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT, The Mouse Genome Database G. MGD: the Mouse Genome Database. Nucleic Acids Research. 2003;31(1):193-5.
47. Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, et al. Rat Genome Database (RGD): mapping disease onto the genome. Nucleic acids research. 2002;30(1):125-8.

48. De Magalhães JP, Costa J, Toussaint O. HAGR: the human ageing genomic resources. Nucleic acids research. 2005;33(suppl_1):D537-D43.
49. de Magalhaes JP, Toussaint O. GenAge: a genomic and proteomic network map of human ageing. FEBS letters. 2004;571(1-3):243-7.
50. Consortium GO. Gene ontology consortium: going forward. Nucleic acids research. 2015;43(D1):D1049-D56.
51. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011;2011.
52. Sutphin GL, Backer G, Sheehan S, Bean S, Corban C, Liu T, et al. Caenorhabditis elegans orthologs of human genes differentially expressed with age are enriched for determinants of longevity. Aging Cell. 2017.
53. Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. Genome biology. 2013;14(7):R75.
54. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013;14(10):3156.
55. Ye X, Linton JM, Schork NJ, Buck LB, Petrascheck M. A pharmacological network for lifespan extension in Caenorhabditis elegans. Aging cell. 2014;13(2):206-15.
56. Barardo D, Thornton D, Thoppil H, Walsh M, Sharifi S, Ferreira S, et al. The DrugAge database of aging-related drugs. Aging cell. 2017;16(3):594-7.
57. Ding A-J, Zheng S-Q, Huang X-B, Xing T-K, Wu G-S, Sun H-Y, et al. Current Perspective in the Discovery of Anti-aging Agents from Natural Products. Natural Products and Bioprospecting. 2017:1-70.
58. Consortium GP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
59. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr. 1948;5:1-34.