

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Szatmári Eszter Zita

ELTE TTK Matematika BSc

Öregedéssel kapcsolatos fehérjék aminosav szekvenciáinak elemzése

BSc szakdolgozat

Külső témavezető: Dr. Kerepesi Csaba (HUN-REN SZTAKI)

Belső témavezető: Dr. Benczúr András (ELTE TTK / HUN-REN SZTAKI)



Budapest, 2024

Köszönetnyilvánítás

Elsősorban Dr. Kerepesi Csabának szeretném megköszönni a hasznos tanácsait, ötleteit és az éveken át tartó segítséget a tudományos világban való navigálásban, többek között a javaslatot, hogy az ELTE TTK Matematika BSc-t válasszam. Rengeteget segített a Python programozási nyelv elsajátításától a konferencia előadások begyakorlásáig és mindig értékes hozzászólásokkal látott el miközben végig kivételesen empatikusan volt. Köszönöm Dr. Benczúr Andrásnak, hogy a HUN-REN SZTAKI Nemzeti Mesterséges Intelligencia Labor része lehetek és a kollégáimnak, akik hasznos tanácsokkal láttak el a kutatómunka folyamán, Dobolyi Zsófiának, Sturm Ádámnak, Lendvai Ádámnak és João Pedro Magalhãesnek.

A szakdolgozatot az RRF-2.3.1-21-2022-00004 számú Európai Unió projekt támogatta a Mesterséges Intelligencia Nemzeti Laboratórium (MILAB) keretein belül.

1. Bevezetés.....	4
1.1. Motiváció.....	4
1.2. Statisztika.....	5
1.3. Bioinformatikai alapok.....	6
1.4. Gépi tanulási módszerek.....	7
1.4.1. Döntési fák.....	9
1.4.2. Random Forest.....	10
1.4.3. Extreme Gradient Boosting (XGBoost).....	11
2. Bioinformatikai adatbázisok.....	13
1.5.1. Human Genome Project.....	13
1.5.2. UniProt.....	13
1.5.3. GenAge.....	15
1.5.4. Gene Ontology.....	15
3. Öregedéssel kapcsolatos fehérjék statisztikai elemzése.....	17
3.1. Egyedi mintázatok az öregedéssel kapcsolatos fehérjék aminosav-szekvenciáiban.....	17
3.1.1. Az öregedéssel kapcsolatos fehérjék azonosítása.....	17
3.1.2. Egyes aminosavak relatív gyakorisága.....	18
3.1.3. Aminosavszekvenciák átlagos hossza.....	20
3.1.4. Diszkusszió.....	21

3.2.	Egyes aminosav párok relatív gyakorisága.....	23
3.3.	Egyes aminosav tripletek relatív gyakorisága.....	26
4.	Öregedéssel kapcsolatos fehérjék gépi tanulós elemzése.....	28
4.1.	Fehérje funkció prediktálása aminosav szekvenciából..	28
4.1.1.	A döntési fák teljesítménye.....	29
4.1.2.	Az XGBoost teljesítménye.....	29
4.1.3.	Random Forest teljesítménye.....	30
4.2.	Öregedéssel kapcsolatos fehérjék prediktálása aminosav szekvenciából.....	31
4.2.1.	XGBoost az aminosavak és párok gyakoriságán..	32
4.2.2.	XGBoost az aminosav párok gyakoriságán.....	33
4.2.3.	XGBoost az aminosav tripletek gyakoriságán.....	35
4.2.4.	XGBoost az aminosav párok és tripletek gyakoriságán.....	36
4.3.	További ötletek: transformer nyelvi modellek.....	37
5.	Összefoglalás.....	39
6.	Irodalomjegyzék.....	40

Bevezetés

1.1. Motiváció

Az élettartam és fiatal életszakasz meghosszabbítása mindig is foglalkoztatta az emberiséget. Az egészségesebb öregedés nemcsak magasabb életszínvonalat biztosít az egyén számára, a társadalom számára is hasznos. Az egészségesebb idősök tovább maradhatnak a munkaerőpiacon, ami hozzájárul a munkaerőhiány csökkentéséhez. Az aktívabb, egészségesebb idős emberek képesek hosszabb ideig produktívan részt venni a munkaerőpiacon, ami növeli a gazdasági teljesítményt és a GDP-t. Az átlagosan hosszabb ideig egészséges társadalom kevesebb egészségügyi ellátásra szorul, ami csökkenti az egészségügyi rendszer leterheltségét. A kevesebb krónikus betegség, kisebb szükség az ápolási intézményekre és kevesebb kórházi ápolási nap mind hozzájárulnak az egészségügyi költségek csökkenéséhez. Továbbá a hosszabb távon aktív egyének csökkentik a nyugdíjrendszer terheit és többet utaznak, vásárolnak és többet költhetnek szabadidős tevékenységekre, ami növeli a gazdasági aktivitást és a turizmusból származó bevételt valamint az egyének életminőségét is. Mindezekre az igény hajtja az egészségmegőrzés, az életminőség javítása és az öregedéssel járó betegségek megelőzése céljára folytatott kutatási projekteket. Ez új piacokat teremt és növeli a gazdasági növekedést. Tehát a megnövekedett élettartam és hosszabb fiatal életszakasz nemcsak az egyének életszínvonalán és egészségügyi állapotán javít, hanem a társadalom számára is rendkívül fontosak az ilyen irányú kutatások.

Élettartamot hosszabbítani már 1993-ban sikerült fonálféregben (Kenyon et al., 1993). Egyetlen, a Daf-2 nevezetű gén kiütésével sikerült megkétszerezni a fonálféreg élettartamát. A muslica esetén 35%-os élettartam hosszabítást (Lin et al., 1998), az egér esetén pedig az 50%-ot (Coschigano et al., 2000; Ladiges et al., 2009) sikerült elérni.

Szerteágazó véleményeket hallhatunk arról, mi okozza az öregedést és mi a legfontosabb megválaszolatlan öregedéskutatással kapcsolatos kérdés. A válaszok között sokszor elhangzik a károsodások felhalmozódása és magának a folyamatnak a megértése (Gladyshev et al., 2024). Szakdolgozatomban

hasonlóságokat és különbségeket keresek az öregedéssel kapcsolatos és azzal nem kapcsolatos gének által kódolt fehérjék aminosavszekvenciái között remélve, hogy az eredmények segítik az öregedés folyamatának megértését.

1.2. Statisztika

A biológiai adatállományok mérete és elérhetősége az elmúlt években jelentősen növekedett, amit nagyrészt az egészségügy iránti fokozott társadalmi érdeklődés és a technológiai fejlődés tett lehetővé. Ezek az adatok rendkívül értékesek a rendszerszintű biológiai folyamatok megértésében, különösen, ha statisztikai elemzési és gépi tanulási módszerek alkalmazásával vizsgáljuk őket. Az ilyen megközelítések lehetővé teszik az összefüggések azonosítását, új hipotézisek felállítását és a komplex biológiai rendszerek mélyebb megértését, ami alapvető fontosságú az egészségügyi kutatások és innovációk számára, ezért az ilyen adatok vizsgálatakor matematikai eszközökre van szükség.

Az öregedéssel kapcsolatos és öregedéssel nem kapcsolatos fehérjék aminosav szekvenciáinak összehasonlításakor jó ötlet lehet kiszámolni az egyes aminosavak relatív gyakoriságát a két halmazban és megvizsgálni, valamelyik esetében van-e szignifikáns különbség. Az aminosavak relatív gyakoriságát az adott aminosav előfordulásainak számából és az összes aminosav számából számított arányként definiáljuk.

$$f_i = \frac{n_i}{N}$$

ahol f_i i-edik aminosav relatív gyakorisága, n_i az i-edik aminosav előfordulásainak száma, N pedig az összes aminosav száma a vizsgált halmazban. Ezután azt kell ellenőrizni, kaptunk-e statisztikailag szignifikáns eredményeket. A t-próba megfelelő az egyes aminosavak relatív gyakoriságának összehasonlítására a két halmaz között, Welch-féle t-próbát alkalmazunk:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ahol $\bar{x}_1$ és $\bar{x}_2$ az első és a második csoport mintaátlagai, s_1^2 és s_2^2 az első és a második csoport mintavarianciái, n_1 és n_2 pedig az első és a második csoport

mintáinak méretei. A Bonferroni-korrekción egy egyszerű módszer a többszörös tesztelésből fakadó hibák csökkentésére. Az elfogadott szignifikanciaszintet ($\alpha = 0.05$) a tesztek számával (m , azaz ahány aminosav van) osztjuk.

$$\alpha_{új} = \frac{\alpha}{m}$$

Ezután minden tesztnél az új szignifikanciaszintet használjuk az eredmények értékeléséhez. Ez csökkenti a téves pozitív eredmények (I. típusú hiba) valószínűségét, de növeli a téves negatív eredmények (II. típusú hiba) kockázatát.

1.3. Bioinformatikai alapok

Egy fehérje aminosav-szekvenciája meghatározza a fehérje alapvető szerkezetét és funkcióját. Az aminosavak sorrendje – azaz a fehérje elsődleges szerkezete – adja meg azokat a kémiai tulajdonságokat, amelyek befolyásolják, hogy a fehérje hogyan fog összezsavarodni és milyen térbeli struktúrát alakít ki.

- Elsődleges szerkezet: Az aminosav-szekvencia maga a fehérje "kódja", amely az aminosavak lineáris sorrendjét adja meg. Minden fehérje egyedi szekvenciával rendelkezik, amely meghatározza, milyen térszerkezetet vesz fel a fehérje.
- Másodlagos szerkezet: Az aminosavak közötti kémiai kötések és kölcsönhatások (például hidrogénkötések) hatására az aminosav-szekvencia egyes szakaszai rendszerint meghatározott formákat – például alfa-héliceket és béta-lemezeket – alakítanak ki.
- Harmadlagos szerkezet: A fehérje háromdimenziós szerkezete az aminosavak lánc közötti kölcsönhatások révén jön létre, amelyek az egész molekulát egy kompakt formába rendezik. Ez a szerkezet alapvető a fehérje funkciójának betöltéséhez, például ahhoz, hogy megfelelően kapcsolódjon más molekulákhoz vagy enzimekként működjön.
- Negyedleges szerkezet: Egyes fehérjék több aminosav-láncból (alegységből) épülnek fel, amelyeket a szekvenciájuk és a köztük lévő kölcsönhatások révén alkotnak stabil komplexeket.

A fehérje aminosav-szekvenciája tehát olyan információkat hordoz, amelyek alapján előre jelezhető a szerkezete és biológiai funkciója. A kémiai Nobel-díjas AlphaFold2 bioinformatikai eszköz (Jumper et al., 2021) például ezt a szekvenciát használja, hogy a fehérje háromdimenziós struktúráját megjósolja, amely alapvető fontosságú annak megértéséhez, hogy a fehérje milyen szerepet tölt be a sejtekben és az élő szervezetben. Az AlphaFold2 egy mesterséges intelligencián alapuló rendszer, amelyet a DeepMind fejlesztett ki fehérjék 3D-s szerkezetének előrejelzésére az aminosavsorrendjük alapján. Forradalmi áttörést jelent a biológiai kutatásban, mivel sokkal gyorsabbá és pontosabbá tette a fehérjestruktúrák meghatározását, amely korábban évekig tartó laboratóriumi munkát igényelt. Ezáltal új lehetőségeket nyit a gyógyszerfejlesztésben, betegségek megértésében és más tudományos területeken.

A háromdimenziós szerkezet határozza meg a fehérje funkcióját. Az AlphaFold képes meghatározni ezt a szerkezetet csupán az aminosav-szekvencia ismeretében, így felmerül a kérdés, hogy milyen különbségek vannak az öregedéssel kapcsolatos és az öregedéssel nem kapcsolatos fehérjék között, prediktálható-e a funkció, nevezetesen, hogy öregedéssel kapcsolatos-e egy fehérje csak az elsődleges szerkezet ismeretében.

1.4. Gépi tanulási módszerek

Ha egy fehérje aminosav-szekvenciájából meghatározhatnánk, hogy öregedéssel kapcsolatos-e a gén által kódolt fehérje, az jelentős áttörést hozna az öregedés molekuláris mechanizmusainak megértésében. Ez lehetővé tenné új öregedési biomarkerek azonosítását, célzott terápiás stratégiák fejlesztését, valamint az élettartamot vagy az egészséges öregedést befolyásoló gének gyorsabb felismerését. Mivel a laboratóriumi validáció idő- és költségigényes, egy ilyen predikciós megközelítés hatékonyabbá tehetné a kutatást. Emellett a predikció új funkcionális hipotézisek megfogalmazásához is vezethetne, melyek segítségével tovább vizsgálhatók az öregedéssel összefüggő biológiai folyamatok. Hsu et al. (n.d.)

A fehérjék funkciójának előrejelzése az aminosav-szekvenciák alapján szintén az élettudományok egyik kiemelt kutatási területe. Az egyre bővülő

szekvencia-adatbázisok és a gépi tanulási algoritmusok fejlődése lehetővé teszi, hogy a hagyományosan laboratóriumi kísérleteken alapuló módszerek mellett gyorsabb és költséghatékonyabb *in silico* megközelítéseket alkalmazzunk. Az aminosav szekvencia alapján történő fehérjefunkció-predikció rendkívül fontos az egészségügy szempontjából és számos módon járul hozzá az egészségmegőrzéshez és a betegségek kezeléséhez. Ha ismerjük egy adott fehérje funkcióját és annak szerepét a sejtekben vagy a szervezetben, az segíthet a betegségek korai diagnosztizálásában. Például, ha egy fehérje szerepet játszik egy specifikus betegség kialakulásában vagy progressziójában, a fehérje szintjének vagy aktivitásának mérése lehet diagnosztikai marker. Azonosítva azokat a fehérjéket, amelyek fontosak egy adott betegség kialakulásában vagy fenntartásában, lehetséges célfehérjéket azonosítani a gyógyszerek tervezése és fejlesztése számára. Ezáltal lehetővé válik hatékonyabb terápiás stratégiák kifejlesztése. Az aminosav szekvenciák alapján történő predikció lehetővé teszi a génmódosított terápiák fejlesztését, például génterápia vagy CRISPR-Cas9 (Hsu et al., n.d., 2014) alkalmazását, amelyek célzottan befolyásolhatják a betegségekben érintett fehérjék működését vagy termelődését.

Összességében az aminosav szekvenciák alapján történő fehérjefunkció-predikció hozzájárulhat az egészségügyi kutatásokhoz és gyakorlathoz, segítve a betegségek jobb megértését, diagnosztizálását és kezelését, valamint a személyre szabott egészségügyi ellátást.

A szakdolgozatban elsősorban két problémát próbáltam megoldani. Cél a fehérjék aminosavszekvenciája alapján meghatározni, hogy a GenAge adatbázis alapján melyek az öregedéssel kapcsolatosak valamint szeretném meghatározni, hogy rendelkeznek-e adott funkcióval. Ehhez a Gene Ontology-ban legtöbbször szereplő, GO:0007165 fehérjefunkcióra oldunk meg bináris klasszifikációs feladatot. Ez a jelátvitel, a sejtekben zajló folyamat, amely során egy jel továbbításra kerül, hogy megváltoztassa a sejt aktivitását vagy állapotát. A jelátvitel egy jel fogadásával kezdődik (pl. egy ligand kötődése egy receptorhoz, vagy a receptor aktiválása egy inger, például fény által), vagy ligand hiányában a jel megvonásával, illetve egy folyamatosan aktív receptor aktivitásával. A jelátvitel azzal zárul, hogy egy downstream sejtfunkció szabályozásra kerül, például a transzkripció vagy egy anyagcsere-folyamat szabályozásával. A

jelátvitel magában foglalja mind a sejtfelszínen lévő receptorokból, mind a sejten belüli molekulákból induló jelátviteli folyamatokat. Sejtek közötti jelátvitel esetén a folyamat az átvevő sejtben és annak belsejében zajló eseményekre korlátozódik.

A két probléma megoldásához három algoritmust, az XGBoost-ot és a Random Forest módszert és a döntési fákat fontos megismerni.

1.4.1. Döntési fák

A döntési fa modellt gyakran felhasználjuk osztályozási és regressziós problémák megoldására. A modell egy hierarchikus fastruktúrát használ, amelyben minden csomópont egy adott tulajdonságot vagy változót jelképez, most az aminosav párok vagy tripletok előfordulását és minden él egy lehetséges értéket vagy döntést reprezentál, például hogy az adott aminosav pár(pl MG) előfordulása több-e 2-nél. Az első lépése a gyökér csomópont kiválasztása. Az algoritmus először kiválaszt egy gyökér csomópontot a fában, amely az elsődleges döntési pontot jelenti. Ez a pont az adott probléma legfontosabb tulajdonságát vagy változóját reprezentálja. Ezután következik a csomópontok bontása. A gyökér csomópontot követően az algoritmus elágazik, és az adatokat olyan részhalmazokra bontja, amelyeket a rendelkezésre álló tulajdonságok és a megfigyelt adatok alapján határoz meg. Ezt az elágazást az adatok leginkább információt hordozó tulajdonságai alapján végzik. Az elágazások folytatódnak minden egyes részhalmazban, létrehozva egy hierarchikus fastruktúrát, amely az adatokat szegmentálja és csoportosítja a döntések alapján. Az algoritmus addig folytatódik, amíg el nem éri a leállási feltételt, amely lehet például a fa maximális mélysége vagy egy adott számú megfigyelés a levélcsomópontokban. A végső struktúra egy döntési fa, amelyet lehet használni új adatok osztályozására vagy regressziós előrejelzések készítésére.

A döntési fa modell előnye, hogy könnyen értelmezhető és átlátható, mivel az emberek könnyen követhetik a döntési folyamatot a fa struktúrája alapján. Emellett a döntési fák jól skálázhatók nagy adathalmazokra és kezelhetik a kategorikus és numerikus változókat egyaránt. Azonban a döntési fa modell hajlamos lehet az overfittingre, különösen ha túl komplex fákat alkalmaznak vagy

ha nem megfelelően kezelik a hiányzó adatokat vagy a zajos változókat. Ennek ellenére a döntési fa modell továbbra is népszerű eszköz a gépi tanulásban, különösen olyan problémák megoldására, ahol az interpretálhatóság fontos szempont.

1.4.2. Random Forest

A Random Forest egy olyan gépi tanulási modell, amely a döntési fa modell alapján működik, de több döntési fát hoz létre és kombinál azokból az eredményekből. A módszer lényege az, hogy több döntési fa generálása és azok összesítése által csökkenthető az egyes fa sajátosságainak negatív hatása, például a túltanulás (overfitting) vagy a zajos adatokra való érzékenység. Az algoritmus először véletlenszerűen választ adatpontokat az eredeti adathalmazból ismétlődő minták létrehozása céljából. Ezáltal létrejön több, kissé eltérő adathalmaz, amelyek különböző részhalmazokat tartalmaznak az eredeti adatokból. A Random Forest algoritmus több döntési fát hoz létre a bootstrap mintavételezésen alapuló adathalmazokból. Ezek a fák hasonlóak lehetnek, de kis eltérésekkel, mivel a bootstrap mintavételezés során véletlenszerűen választott minták miatt különböző mintavételek állnak rendelkezésre. Minden egyes döntési fa építésekor a Random Forest algoritmus véletlenszerűen kiválaszt egy részhalmazt a rendelkezésre álló jellemzőkből. Ez segít abban, hogy minden fa kissé más jellemzőket vegyen figyelembe az osztályozási döntés során. Az osztályozás folyamata során a Random Forest algoritmus kombinálja az összes létrehozott döntési fa eredményeit, például azáltal, hogy szavazást tart minden fa által előrejelzett osztályok között. Az osztály, amely a legtöbb szavazatot kapja, lesz az algoritmus által kiválasztott osztály. A Random Forest modell előnyei közé tartozik a jó teljesítmény különböző típusú adathalmazokon, a túltanulás (overfitting) csökkentése és az adatok zajra való érzékenységének csökkentése. Emellett a modell képes kezelni nagy dimenziójú adathalmazokat és többféle típusú változókat. Összességében a Random Forest egy hatékony és sokoldalú gépi tanulási modell, amely széles körben alkalmazható osztályozási és regressziós problémák megoldására, különösen olyan helyzetekben, ahol a döntési fák egyedül nem érik el a kívánt teljesítményt.

1.4.3. XGBoost

Az XGBoost (Tianqi Chen, 2016) az egyik használt modell. Az XGBoost (Extreme Gradient Boosting) egy népszerű és erőteljes gépi tanulási algoritmus, amelyet felügyelt tanulási feladatokhoz használnak, különösen a klasszifikációban és a regresszióban. Ez egy ensemble tanulási módszer, amely több gyengébb modellek előrejelzéseit kombinálja (általában döntési fákat) egy erős előrejelző modell létrehozása érdekében. Az XGBoostot a magas teljesítménye, a skálázhatósága és a nagy adathalmazok és összetett kapcsolatok kezelésére való képessége jellemzi.

Az algoritmus alkalmazásához az első lépés feature tábla készítése volt. Összesen négy feature táblán tanítottuk az XGBoost algoritmust. A feature táblákban minden fehérjének megfelel egy sor és aminosavnak, aminosavpárnak vagy tripletnek (azaz featurenek) egy oszlop. Továbbá szükség van egy oszlopra, amiben 1-el jelöltük, ha a fehérje rendelkezik a leggyakrabban előforduló funkcióval és 0-val, ha nem. Az XGboost főbb jellemzői:

- Gradient Boosting: Az XGBoost a gradient boosting technikán alapul, amely lépésről lépésre javítja a modell előrejelzéseit. Az algoritmus lépésről lépésre minimalizálja a hibát (a tényleges és a prediktált értékek közötti különbségeket), így alkalmas mind a regresszióra, mind a klasszifikációra.
- Az XGBoost több döntési fát kombinál egy erősebb ensemble modell létrehozása érdekében. Ezek a fák általában sekélyek és "gyenge tanulók". A sok gyenge tanuló kombinálásával az XGBoost csökkenti a modell elferdülését és varianciáját, ami jobb általánosításhoz vezet. Ráadásul a sekély fákat alkalmazó algoritmus eredményeit könnyebb értelmezni, ami bioinformatikai alkalmazás esetén nagyon hasznos.
- Tartalmaz L1 (Lasso) és L2 (Ridge) regularizációs technikákat a túltanulás megakadályozására. A regularizációs tételek a célfüggvényhez

adódna hozzá, hogy szabályozzák a modell bonyolultságát és csökkentsék a zajhoz való illeszkedés kockázatát.

- Lehetővé teszi testreszabott veszteségfüggvények meghatározását, amelyek hasznosak specializált feladatokhoz vagy nem szabványos adatok kezeléséhez. Ez a rugalmasság alkalmazkodást tesz lehetővé különböző problémákhoz.
- Az XGBoost hatékony és támogatja a párhuzamos és elosztott számítást. Teljes mértékben kihasználja a többmagú CPU-kat és az elosztott számítási platformokat, hogy a modellek gyorsabban tanulhassanak, különösen nagy adathalmazokkal dolgozva. Aminosavak, aminosav párok és triptetek relatív gyakorisága alapján tanítottuk az algoritmust.
- Beépített támogatást nyújt a hiányzó adatok kezeléséhez, ami erőssé teszi valós adathalmazokkal dolgozva. Automatikusan megtanulja, hogyan ossza fel az adatot a hiányzó értékek alapján. Ennek a dolgozatnak az esetében erre a funkcióra nem volt szükség. Nincsenek hiányzó adatok.
- Fontosság függvényeket is nyújt, amelyek segítenek azonosítani, melyik jellemzőnek van a legnagyobb hatása a modell előrejelzéseire. Ez az információ hasznos lehet a jellemzők kiválasztásához és a változók jelentőségének megértéséhez a problémában. Például gondoljunk arra, hogy szeretnénk szekvencia alapján prediktálni, hogy egy fehérje öregedéssel kapcsolatos-e. Ha ugyanazoknak az aminosavaknak a relatív gyakorisága jelentős jellemző, mint amelyek esetén szignifikáns eltérést tapasztaltunk az öregedéssel kapcsolatos és a komplementer halmazban, az ráerősít még egy oldalról, hogy ez a fehérjék egy fontos jellemzője.
- Tartalmaz korai megállítási mechanizmusokat, amelyek lehetővé teszik a tanítási folyamat leállítását, amikor a modell nem javítja továbbít a teljesítményét egy validációs adathalmazon, így a túltanulást megelőzve. Gyakran használ keresztelő validálási technikákat a modell teljesítményének és megbízhatóságának értékelésére. Az ebben a dolgozatban említett eredmények is ennek a módszernek a segítségével készültek. A kereszt-validálás egy statisztikai technika, amelyet a prediktív modellek teljesítményének és általánosító képességének értékelésére használnak, különösen a gépi tanulásban és adatelemzésben.

A módszer lényege, hogy az adatot több részre bontják, majd a modellt különböző adatkombinációkon tanítják és tesztelik. A kereszt-validálás alapvető célja a modell teljesítményének pontosabb becslése a túltanulás vagy alultanulás kockázatának csökkentése révén.

- Az XGBoost népszerűséget szerzett és gyakran az első választás strukturált adatfeladatokhoz, például előrejelző modellezéshez, Kaggle versenyekhez és sok valós alkalmazáshoz, mivel erőteljes, skálázható és magas előrejelzési pontosságot nyújt. Az XGBoost a gépi tanulási gyakorlók eszköztárában erőteljes eszközzé vált.

2. Bioinformatikai adatbázisok

Az öregedéssel kapcsolatos fehérjék vizsgálatához szükségünk van adatbázisokra, amelyek tartalmazzak információt róluk, például az aminosavszekvenciájukat. A dolgozatban publikus adatbázisokat használtunk.

2.1. Human Genome Project

A Human Genome Project (HGP) egy nemzetközi tudományos projekt volt, amely 1990-ben indult és 2003-ban zárult le, célja pedig az emberi genom teljes szekvenálása és az emberi DNS-ben található körülbelül 20-25 ezer gén azonosítása volt. Ez a projekt alapvető áttörést jelentett a genetika és a molekuláris biológia területén, mivel egy teljes, nyilvános adatbázist hozott létre az emberi genetikai információról.

2.2. UniProt

A UniProt adatbázist közösen tartja karban három intézmény: a European Bioinformatics Institute (EMBL-EBI) az Egyesült Királyságban, a Sveitsi Intézet a Bioinformatikában (SIB) Svájcban, valamint a Protein Information Resource

(PIR) az Egyesült Államokban. Az intézmények összefogásának köszönhetően az adatbázis naprakész és megbízható információkat tartalmaz a fehérjéről genetikusok és más tudományterületek szakemberei találhatók, akik hozzájárulnak az adatbázis fejlesztéséhez, karbantartásához és kibővítéséhez. A HGP eredményei lehetővé tették olyan adatbázisok, például a UniProt (Bateman et al., 2017) létrejöttét, amelyek a gének által kódolt fehérjék részletes adatainak gyűjtésére és annotálására szolgálnak. A HGP által azonosított gének DNS-szekvenciáinak fordítása (transzláció) eredményezi a fehérjék aminosavszekvenciáit, amelyeket a UniProt adatbázis katalogizál. Így a HGP által nyert genomikus adatok az UniProt alapjául szolgálnak az emberi fehérjék rendszerezésében, funkcióik feltárásában és biológiai jelentőségük megértésében. Míg a HGP a gének azonosítására koncentrált, a UniProt adatbázis a fehérjéket – az élő szervezetek működéséért felelős molekulákat – katalogizálja, és információkat nyújt azok szerkezetéről, funkciójáról, valamint a sejten belüli szerepükről. A HGP adatai nélkül a UniProt nem tudta volna összekapcsolni a géneket azok fehérjetermékeivel, így a HGP közvetetten hozzájárult a UniProt adatbázis, és más, funkcionális genomikai kutatások alapjainak megteremtéséhez. Ezáltal a Human Genome Project és a UniProt együtt támogatják az emberi egészség és betegségek molekuláris szintű megértését, és lehetőséget nyújtanak az új gyógymódok és terápiás beavatkozások kifejlesztéséhez. A UniProt adatbázis tartalmazza az összes ismert humán fehérje aminosav szekvenciáit. A fehérjék aminosavakból épülnek fel. Ha megfeleltetünk minden aminosavnak egy betűt, a fehérjék olyanok, mint a hosszú szavak a szervezetben. A GenAge (Pedro De Magalhães et al., 2009) adatbázis tartalmazza az összes ismert öregedéssel kapcsolatos gént. Mivel a fehérjéket gének kódolják, megfeleltethetjük egymásnak a GenAge-ben szereplő géneket a UniProt-ban szereplő fehérjékkel. Ha egy adott fehérjét kódoló gén szerepel a GenAge-ben, a kódolt fehérjék öregedéssel kapcsolatosnak tekintjük, máskülönben öregedéssel nem kapcsolatosnak.

2.3. GenAge

A GenAge adatbázis egy átfogó, nyilvánosan hozzáférhető bioinformatikai erőforrás, amely az öregedéssel kapcsolatos géneket tartalmazza. Az adatbázist a gerontológiai kutatások támogatására hozták létre, különös tekintettel azokra a génekre, amelyek befolyásolják az élettartamot és az öregedési folyamatokat. Két fő szekcióból áll: az emberi gének listájából, amelyek szerepet játszhatnak az öregedésben, valamint egy másik részből, amely modellorganizmusok (például egerek, élesztők és fonálférgek) öregedéssel kapcsolatos génjeit tartalmazza. Az adatbázisban található géneket gondosan válogatták különféle tudományos források, például genetikai, kísérleti és bioinformatikai adatok alapján. A GenAge kiemelten fontos a hosszú élettartammal kapcsolatos kutatásokhoz, mivel segít azonosítani azokat a molekuláris mechanizmusokat, amelyek az öregedés szabályozásában játszanak szerepet. Emellett hozzájárul a terápiás célpontok meghatározásához, amelyek az életminőség javítását célozzák idősebb korban. A GenAge adatbázis a Human Ageing Genomic Resources (HAGR) projekt része és folyamatosan frissítik a legújabb tudományos eredmények alapján.

2.4. Gene Ontology

A Gene Ontology (GO) (Ashburner et al., 2000; Consortium et al., 2023) egy bioinformatikai eszköz, amely egységes szókészletet biztosít a gének és géntermékek funkcióinak leírására különböző fajokban. Három fő kategóriára oszlik: molekuláris funkciók (pl. enzimaktivitás), biológiai folyamatok (pl. sejtosztódás) és sejt szintű komponensek (pl. membránok, organellek). A GO célja, hogy egy standardizált rendszert hozzon létre, amely megkönnyíti a gének funkcionális szerepének összehasonlítását különböző organizmusokban. A GO folyamatosan bővül, a tudományos közösség által hozzáadott adatokkal és a legújabb kutatási eredményekkel frissül. Az adatokat hierarchikus struktúrában szervezi, lehetővé téve a részletes és általános funkcionális információk lekérdezését. A Gene Ontology kulcsszerepet játszik a genomikai és proteomikai

elemzésekben, segítve a biológiai adatok értelmezését, és hozzájárul a rendszerszintű biológiai megértéshez. Az adatbázis karbantartását és fejlesztését a Gene Ontology Consortium végzi. Ez egy nemzetközi tudományos közösség, amely több mint 20 évvel ezelőtt alakult a GO-projekt létrehozásával. A consortium tagjai között bioinformatikusok, biológusok, genetikusok és más tudományterületek szakemberei találhatók, akik hozzájárulnak az adatbázis fejlesztéséhez, karbantartásához és kibővítéséhez.

```

1  >sp|A0A087X1C5|CP2D7_HUMAN Putative cytochrome P450 2D7 OS=Homo sap
2  MGLEALVPLAMIVAIIFLLLVDLMRHQQRWAARYPPGPLPLPGLGNLLHVDQNTPYCFDQ
3  LRRRFQDVFSQLAWTPVVVLNGLAAVREAMVTRGEDTADRPPAPIYQVLGFGPRSQGVI
4  LSRYPAPWREQRFSVSTLRNLGLGKKSLEQWVTEEAACLCADFADQAGRPFRPNGLLDK
5  AVSNVIASLTCGRRFEYDDPRFLRLDLAQEGLKEESGFLREVNAVVPVLPHPALAGKV
6  LRFQKAFLTQLDELLTEHRMTWPAQPPRDLTEAFLAKKEKAKGSPSSFNDENLRIVVG
7  NLFLAGMVTSTTTLAWGLLLMILHLDVQGRRRVSPGCPVGVTHVCPVRVQQEIDDVIGQV
8  RRPENGDQAHMPCTTAVIHEVQHFQDIVPLGVTHMTSRDIEVQGFRIKGTTLITNLSSV
9  LKDEAVWKKPFRFHEHFLDAQGHFVKPEAFLPFSAGRRACLGELARMELFFFTSLQ
10 HFSFSVAAGQPRPSHSRVVSFLVTPSPYELCAVPR
11 >sp|A0A0B4J2F0|PIOS1_HUMAN Protein PIGBOS1 OS=Homo sapiens OX=9606
12 MFRRLLTFAQLLFATVLGIAGGVYIFQPVFEQYAKDQKELKEKMLVQSEEEKKS

```

1.ábra Részlet a Uniprot adatbázisból

A	B	C	D	E
GenAge ID	symbol	name	entrez gene id	uniprot
1	GHR	growth hormone receptor	2690	GHR_HUMAN
2	GHRH	growth hormone releasing hormone	2691	SLIB_HUMAN
3	SHC1	SHC (Src homology 2 domain containing)	6464	SHC1_HUMAN
4	POU1F1	POU class 1 homeobox 1	5449	PIT1_HUMAN
5	PROP1	PROP paired-like homeobox 1	5626	PROP1_HUMAN
6	TP53	tumor protein p53	7157	P53_HUMAN
7	TERC	telomerase RNA component	7012	

2.ábra Részlet a GenAge adatbázisból

3. Öregedéssel kapcsolatos fehérjék statisztikai elemzése

3.1. Egyedi mintázatok az öregedéssel kapcsolatos fehérjék aminosav-szekvenciáiban

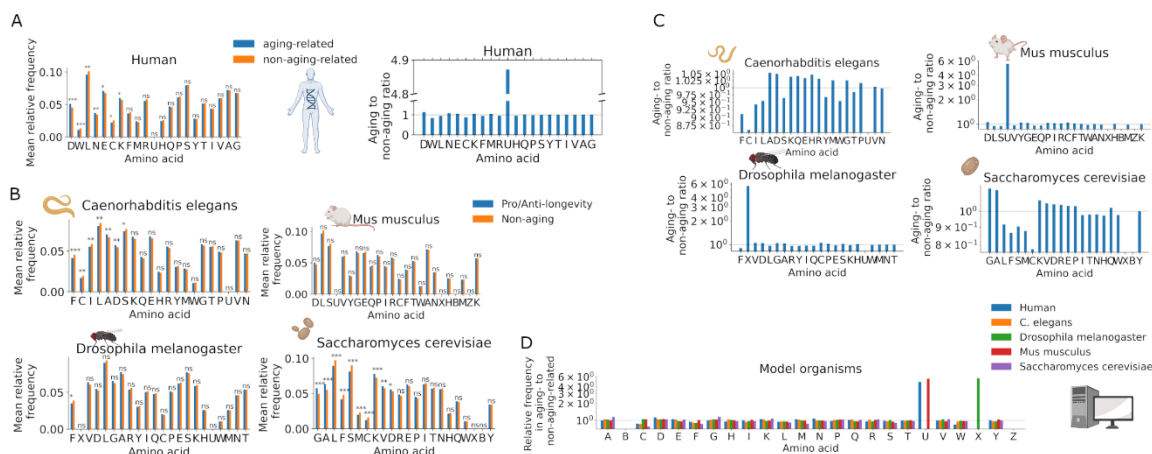
Az alábbi fejezet a szerző azon előzetes eredményeit tartalmazza, amelyeket egy 2024-ben megjelent Q1-es, angol nyelvű folyóiratpublikációban foglalt össze „Unique Patterns in Amino Acid Sequences of Aging-Related Proteins” címmel (Szatmári et al., 2024)

3.1.1. Az öregedéssel kapcsolatos fehérjék azonosítása

A gének fehérjét kódolnak és mivel a GenAge adatbázis tartalmazza az öregedéssel kapcsolatos géneket, azt kell megállapítanunk, melyek az ezek által a gének által kódolt fehérjék a UniProt adatbázisban. Ezeket a fehérjéket fogjuk öregedéssel kapcsolatosnak nevezni. Két különböző azonosító is szerepel ebben a két adatbázisban, ami megfelel erre a célra, a GenAge adatbázisban ezek a 'symbol' és 'uniprot' oszlopok. A UniProt-ban pedig a fehérje neve illetve a 'GN=' szimbólum után szereplő azonosító. A GenAge 306 öregedéssel kapcsolatos humán fehérjét tartalmaz, míg a Uniprot 20421 humán fehérje aminosav szekvenciáját, többek között a GenAge szerint öregedéssel kapcsolatos génekét is.

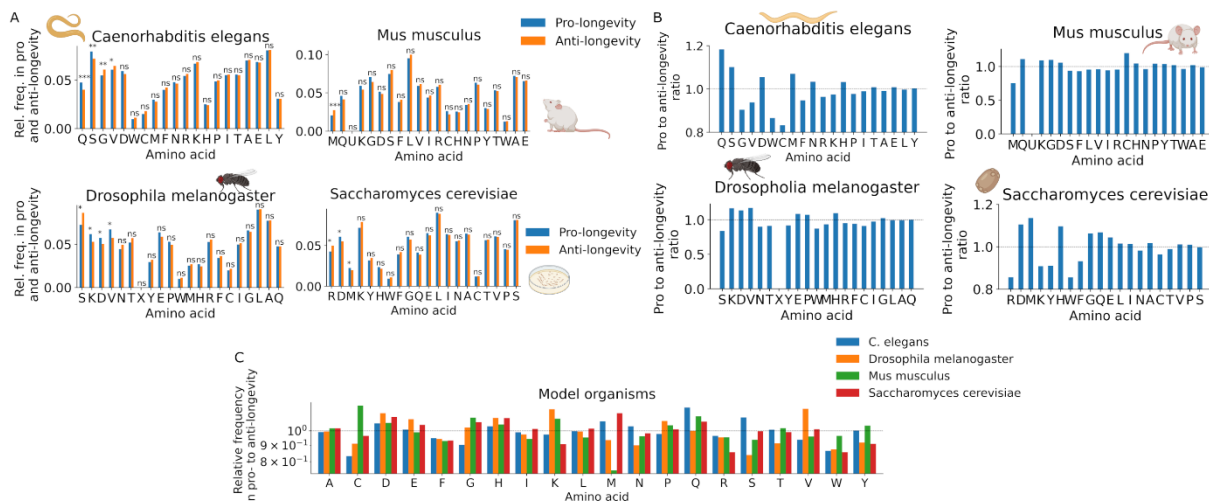
3.1.2. Egyes aminosavak relatív gyakorisága

Érdemes kiszámolni az egyes aminosavak relatív gyakoriságát az öregedéssel kapcsolatos és azzal nem kapcsolatos halmazban, majd összevetni, hogy valamelyik aminosav esetén szignifikáns eltérést tapasztalunk-e. Ezt megvizsgáltuk humán fehérjék és négy modell organizmus, a fonálféreg(*Caenorhabditis elegans*), egér(*Mus musculus*), élesztő (*Saccharomyces cerevisiae*) és a muslica(*Drosophila melanogaster*) esetén. A legkiemelkedőbb különbségek az aszparaginsav (D), a triptofán (W) és a leucin (L) esetében voltak ($p\text{-érték} < 0,0035$, FDR korrigált $p\text{-érték} < 0,01314$)(3.ábra A). Vizsgáltuk, hogy ezek a jelenségek evolúciósan konzerváltak-e. A triptofán és a leucin relatív gyakorisága átlagosan alacsonyabb volt, míg az aszparaginsav relatív gyakorisága átlagosan magasabb volt az öregedéssel kapcsolatos fehérjéken, mint a nem öregedéssel kapcsolatosokban a humán és a négy vizsgált modellorganizmus esetén is(*C. elegans*, *D. melanogaster*, *M. musculus* és *S. cerevisiae*) (3.ábra B-D).



3. ábra. A) Az egyes aminosavak átlagos relatív gyakorisága (százalékban) az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben (bal oldali panel), valamint az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjék átlagos relatív gyakoriságának aránya az emberi fehérjék esetében (jobb oldali panel). B) Az egyes aminosavak átlagos relatív gyakorisága (százalékban) az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben modellezett organizmusokban. C) Az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjék átlagos relatív gyakoriságának aránya modellezett organizmusokban. D) Az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjék átlagos relatív gyakoriságának aránya emberi és különböző modellezett organizmusokban.

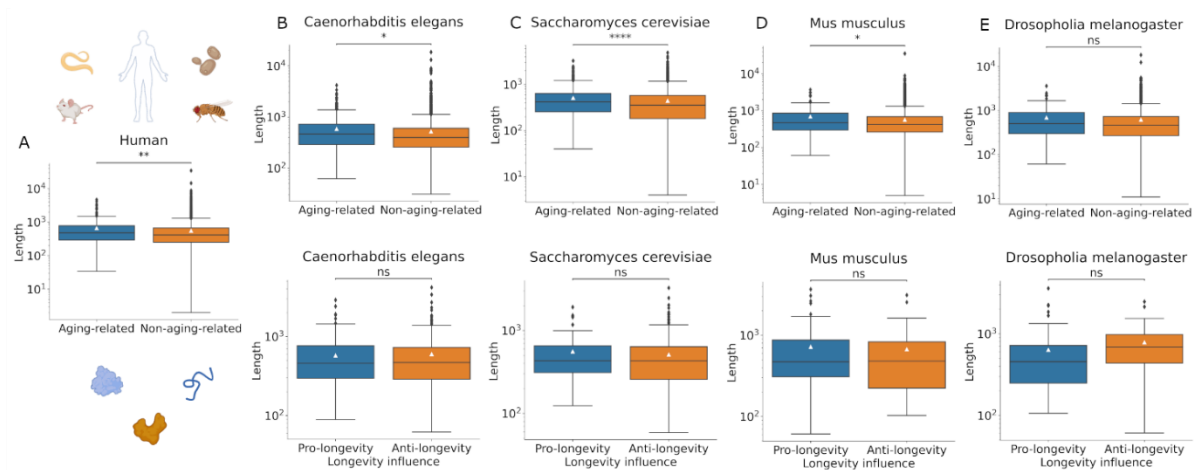
Szintén megvizsgáltuk a relatív gyakoriságok arányát a pro-longevity fehérjék (olyan gének által kódolt fehérjék, melyek alulműködtetése és/vagy túlműködtetése meghosszabbította az élettartamot) és az anti-longevity fehérjék (olyan gének által kódolt fehérjék, melyek alulműködtetése és/vagy túlműködtetése lecsökkenti az élettartamot) között a négy modellorganizmus esetén (4. ábra). Az aszparaginsav (D) átlagos relatív gyakorisága átlagosan magasabb volt, míg az arginin (R) átlagos relatív gyakorisága állandóan alacsonyabb volt a pro-longevity fehérjék esetében, összehasonlítva az anti-longevity fehérjékkel. Ez a jelenség mind a négy vizsgált modellorganizmus esetén szignifikáns volt.



4. ábra A) A pro-longevitású és anti-longevitású fehérjék egyes aminosavak átlagos relatív gyakorisága (százalékban) modellezett organizmusokban. B) A pro-longevitású és anti-longevitású fehérjék átlagos relatív gyakoriságának aránya modellezett organizmusokban. C) A pro-longevitású és anti-longevitású fehérjék átlagos relatív gyakoriságának aránya különböző modellezett organizmusokban.

3.1.3. Aminosavszekvenciák átlagos hossza

Továbbá vizsgáltuk az aminosavszekvenciák átlagos hosszát a két halmazban. Megállapítottuk, hogy az öregedéssel kapcsolatos fehérjék átlagosan hosszabbak, mint a nem öregedéssel kapcsolatos fehérjék humán és modell organizmusok esetében (5. ábra). Azonban nem volt szignifikáns különbség a pro-longevity fehérjék és az anti-longevity fehérjék átlagos hossza között a vizsgált modell organizmusok egyikében sem.

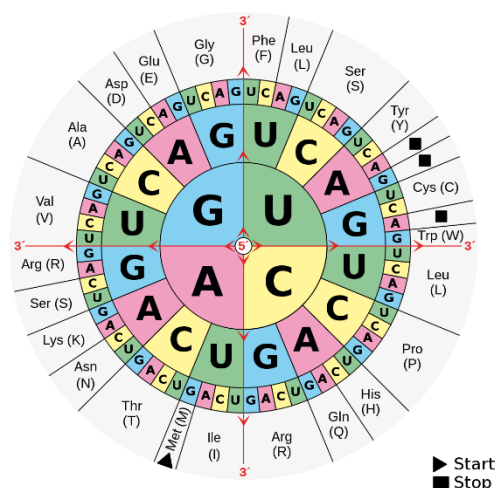


5. ábra. A) Az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjék hosszának összehasonlítása emberek esetében. B–E) Az öregedéssel kapcsolatos és az öregedéssel nem kapcsolatos fehérjék hosszának összehasonlítása, valamint a pro-longevitászú és anti-longevitászú fehérjék hosszának összehasonlítása *C. elegans*, *S. cerevisiae*, *M. musculus* és *D. melanogaster* esetében.

3.1.4. Diszkusszió

Mivel az öregedést gyakran úgy írják le, mint a károsodások felhalmozódását, elképzelhető, hogy néhány aminosav hajlamosabb bizonyos molekuláris károsodásokra, mint más aminosavak, és ez összefügg a fehérje élettartamával.

Az aminosavakat három nukleotidból álló tripletek kódolják. A legtöbb aminosavat több triplet kódolja, míg a triptofánt csak egyetlen triplet. Ezért egyetlen mutáció bármelyik nukleotidban a triptofánt kódoló háromnukleotidos szekvenciában más aminosavhoz vezet. Megfigyeltük, hogy szignifikáns különbség van a triptofán esetében a relatív gyakoriságok között az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben. Elképzelhető, hogy a triptofán érzékeny a somatikus mutációkra, és ez kapcsolódhat annak alulreprezentáltságához az öregedéssel kapcsolatos fehérjék esetében az emberben és a modell organizmusokban.



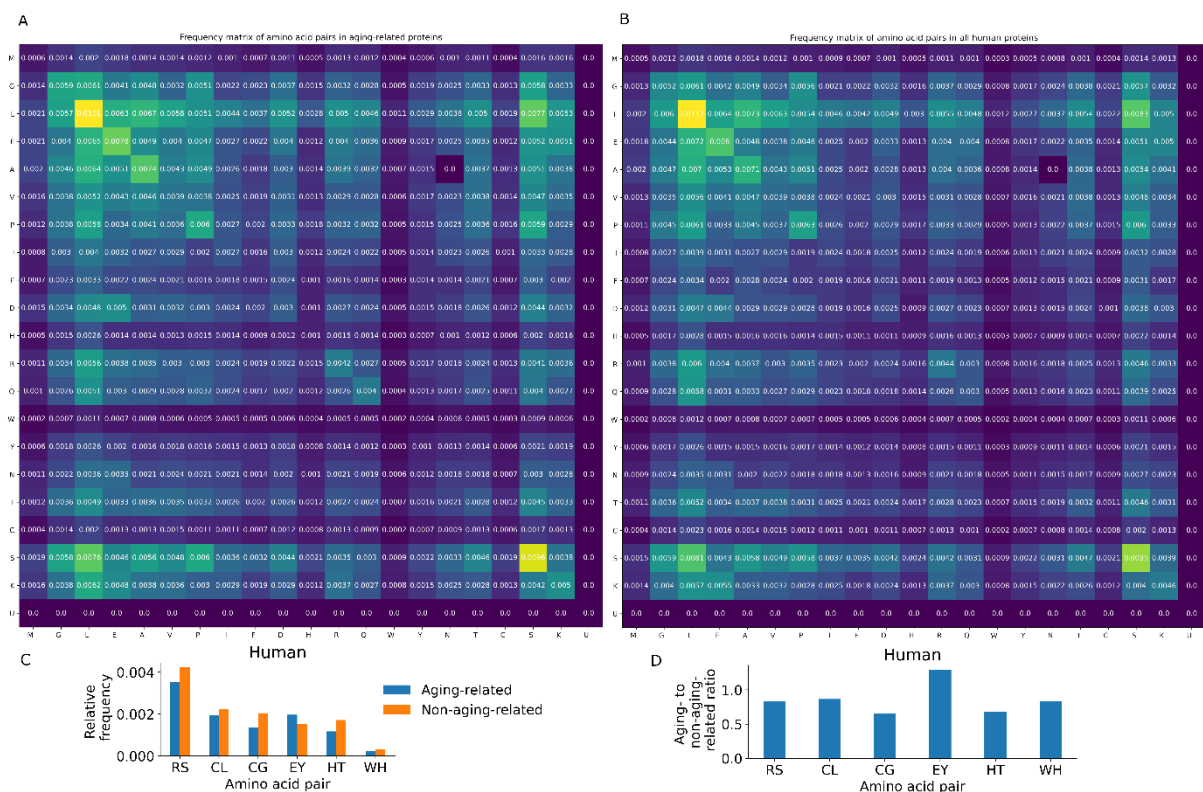
6. ábra. Genetikai kód táblázat

Az aminosav-egyensúlynak fontos szerepe van az öregedés folyamatában és az öregedéssel kapcsolatos betegségek szabályozásában (Canfield & Bradshaw, 2019). A triptofán bizonyos végső termékeken keresztül (mTOR, IIS, kinurensav) meghosszabbíthatja az élettartamot (Annemieke T. van der Goot and Ellen A.A. Nollen, 2013). A leucin elérhetőségének növekedése felregulálja az mRNS fordítását, ezáltal növeli az izomfehérje szintézist, ami pedig nagyobb nettó izomfehérje felhalmozódáshoz vezet (Leenders & Van Loon, 2011). Bizonyos szövetekben a fogakban az aszpartil maradékok racemizációja mutatható ki, ami olyan sebességgel történik, hogy az éves 0,1% D-aszparaginsav-tartalom növekedésre utal, ami az öregedést jelzi (Patricia M. Helfman; Jeffrey L. Bada; Ming-Yung Shou, 1977)

Korábban kimutatták, hogy a humán öregedéssel kapcsolatos gének hosszabb transzkripció-hosszal rendelkeznek az összes fehérjekódoló génhez képest (Inês Lopes 1, 2021). Továbbá kimutatták, hogy a pro-longevity gének fehérjetermékei általában hosszabbak, mint a *C. elegans* többi fehérjéje (Yan-Hui Li 1, 2010). Jelenlegi kutatásunk során azt találtuk, hogy az öregedéssel kapcsolatos fehérjék általában hosszabbak, mint a nem öregedéssel kapcsolatos fehérjék humán és a négy modell organizmus esetében is. Ezzel szemben nem találtunk szignifikáns különbséget a pro-longevity és az anti-longevity fehérjék átlagos hossza között a vizsgált modell organizmusok egyikében sem.

3.2. Egyes aminosav párok relatív gyakorisága

Az aminosav-párok különböző típusú kötések révén kölcsönhatásba lépnek egymással, például hidrogénkötések, diszulfidkötések, hidrofób kölcsönhatások és elektrosztatikus kölcsönhatások révén. Ezek a kölcsönhatások stabilizálják a fehérje szerkezetét, befolyásolják a hajtogatást és a stabilitást. Néhány aminosav-pár pedig funkcionális helyeket képezhet a fehérjében. Ezek a helyek kritikusak a fehérje aktivitásához, például az enzimekben lévő katalitikus helyek, a ligandok vagy substrátokhoz kötődő helyek, vagy a fehérje-fehérje kölcsönhatásokban részt vevő helyek. Az aminosav-párok változása konformációs változásokat válthat ki a fehérjében. Ezek a változások gyakran kulcsfontosságúak a fehérje funkciójában, lehetővé téve a fehérje interakcióját más molekulákkal vagy szerkezeti átmeneteket környezeti jelzések hatására. Az aminosav-párok változásai, azaz a mutációik jelentős hatással lehetnek a fehérje szerkezetére és funkciójára. A mutációk zavarhatják a fehérje hajtogatását, befolyásolhatják a stabilitást, módosíthatják a funkcionális helyeket vagy akár a fehérje funkciójának teljes elvesztéséhez is vezethetnek. Ellenkezőleg, a mutációk új funkcionalitást is bevezethetnek vagy növelhetik a fehérje bizonyos tulajdonságait. Ezek miatt érdemes vizsgálni az aminosav párok viselkedését is az öregedéssel kapcsolatos és az öregedéssel nem kapcsolatos fehérjék halmazában. Első lépésként kíváncsi voltam, hasonló-e az aminosav párok eloszlása az öregedéssel kapcsolatos fehérjékben, mint az összes emberi fehérjében. Elkészítettem az egyes aminosav párok gyakorisági mátrixát az öregedéssel kapcsolatos fehérjékben és az összes humán fehérjében. A két ábra alapján nagyon hasonlóképpen oszlanak el az aminosav párok a két halmazban. Azonban az öregedéssel kapcsolatos fehérjékben az EY aminosav pár Bonferroni korrekció után szignifikánsan többször, míg az RS, CL, CG, HT, WH párok szignifikánsan kevesebbszer fordul elő.



7. ábra. Az aminosav-párok gyakorisági mátrixa A) az öregedéssel kapcsolatos fehérjékben és B) az összes emberi fehérjében. C) Az egyes aminosav-párok átlagos relatív gyakorisága (százalékban) az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben az emberek esetében, amelyek szignifikánsan különböznek a két halmazban Bonferroni-korrektció után. D) Az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjék átlagos relatív gyakoriságának aránya az emberek esetében.

Megvizsgáltuk az egyes aminosav párok Pearson korrelációját is. A párok esetén nagyon nagy arányban azt vettük észre, hogy nincs jelentős korreláció. Azonban ez nem feltétlen mutatja meg, mennyire van összefüggés a párok előfordulásának valószínűsége között. Ennek az oka az aminosav párok előfordulásának számolási módja. Tekintsük a 'MATPVRDETRNVI' peptidet. Szerepel benne az 'MA', 'AM', 'AT', 'TP', stb. párok. Előfordulhat, hogy nem korrelálnak, de ettől még minden egymást követő pában közös egy aminosav. Így valójában nem teljesen függetlenek.

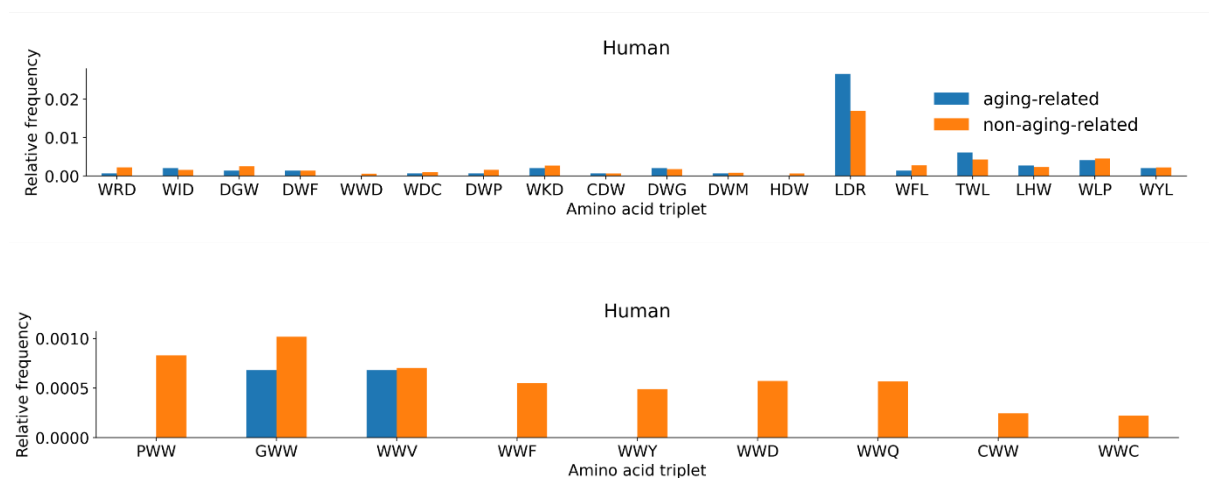
pair1	pair2	P-value	pearson
MG	MG	0	1
MG	GL	1.66E-16	0.447941
MG	LE	1.31E-12	0.390832
MG	EA	8.04E-06	0.252095
MG	AL	1.87E-13	0.404198
MG	LV	1.75E-12	0.388807
MG	VP	8.80E-11	0.359773
MG	PL	2.84E-15	0.431016
MG	LA	8.53E-08	0.300293
MG	AM	9.72E-22	0.510817
MG	MI	0.000587	0.195409
MG	IV	2.02E-06	0.267758
MG	VA	6.33E-08	0.303156
MG	AI	5.17E-09	0.32611
MG	IF	6.81E-06	0.25403

8. ábra. Részlet a párok korrelációját bemutató táblázatból

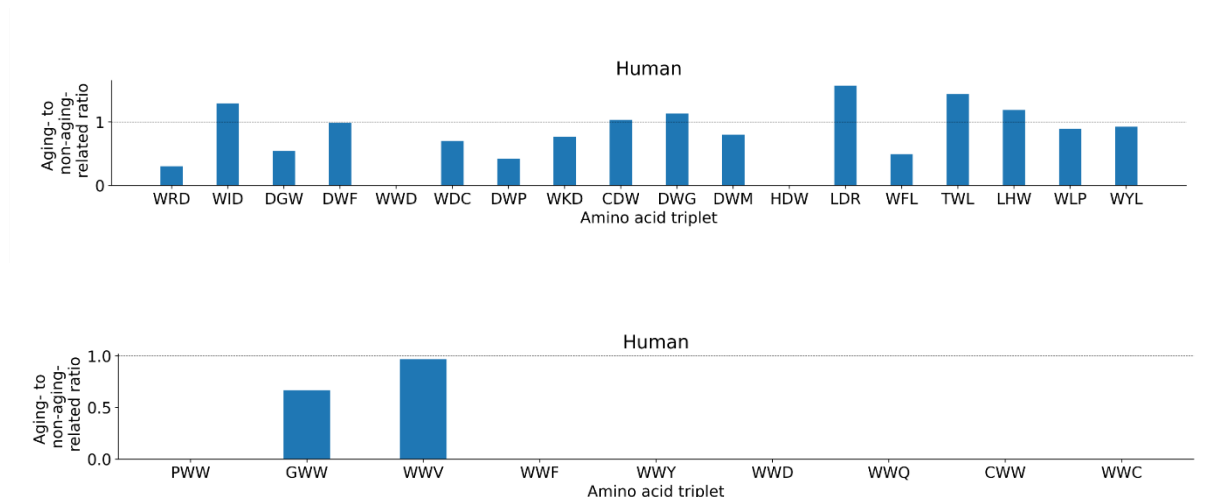
3.3. Egyes aminosav tripletek relatív gyakorisága

A három aminosavból álló tripeptidek önálló biológiai funkcióval bírnak. A szervezetben különféle szerepeket töltenek be, például antioxidánsként, jelzőmolekulaként vagy metabolikus köztes termékként. Vizsgáljunk meg néhány tripletet, amelyek relatív gyakorisága szignifikánsan különbözött a két halmazban Bonferroni korrekció után is, nevezetesen azokat, amelyek tartalmaznak aszparaginsavat, leucint vagy triptofánt, mert ezek az aminosavak bizonyultak érdekesnek a relatív gyakoriságok vizsgálatakor. A korrekció után összesen 248 tripletre kaptunk szignifikáns eredményt. Azok a tripletek bizonyultak különösen érdekesnek, amelyekben két triptofán helyezkedik el egymás mellett. Ilyen tripletek előfordulnak az öregedéssel nem kapcsolatos fehérjék aminosav-szekvenciáiban, azonban az öregedéssel kapcsolatos halmazban alig. Ennek több oka is lehet. Egyrészt korábban láthattuk, hogy a triptofánt egyetlen kodon kódolja, így érzékenyebb lehet a szomatikus mutációkra, könnyen másik aminosav kerül a helyére az átírás során. Továbbá a triptofán egy nagy méretű és hidrofób aminosav. A triptofán oldallánca a legnagyobb az összes aminosav közül, egy benzolgyűrűt és egy pirrolgyűrűt tartalmaz. Két egymás melletti nagy oldallánc szerkezetileg torzítaná a fehérje szerkezetét, és zavarhatja az optimális hajtogatódást. A triptofán oldallánca erősen hidrofób is, ezért inkább a fehérjék belső, nem poláros részéhez vonzódik. Két hidrofób oldallánc egymás mellett erősítheti a hidrofób hatásokat, ami destabilizálhatja a fehérje szerkezetét vagy okozhat fehérje-aggregációt (összetapadást) az oldatban. Továbbá egyike a legritkább aminosavaknak a fehérjékben, és szintézise energetikailag is drága a sejtek számára. Emiatt biológiailag nem lenne hatékony, ha sok triptofán lenne egy szekvenciában, főleg egymás mellett. Két egymás melletti triptofán oldallánca a többi aminosavval is kölcsönhatásba léphet, amely térbeli akadályokat eredményezhet, és megnehezítheti a fehérje stabil szerkezetének kialakítását. Ezek a tényezők együttesen megnehezítik két triptofán közvetlen szomszédságát, és ez az oka

annak, hogy a természetes fehérjékben ez a kombináció ritka. Az öregedéssel kapcsolatos fehérjékben azért lehetnek ilyen ritkák, mert azok érzékenyebbek a károsodások felhalmozódására.



9. ábra. Néhány aminosav-triplet átlagos relatív gyakorisága (százalékban) az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben az emberek esetében. A két halmaz közötti különbség minden pár esetében szignifikáns volt (p -érték $< 0,0002$).



10. ábra. Az aminosav-tripletek átlagos relatív gyakoriságának aránya, amelyeket az 5. ábrában elemeztem az öregedéssel kapcsolatos és nem öregedéssel kapcsolatos fehérjékben az emberek esetében.

4. Öregedéssel kapcsolatos fehérjék gépi tanulós elemzése

4.1. Fehérje funkció prediktálása aminosav szekvenciából

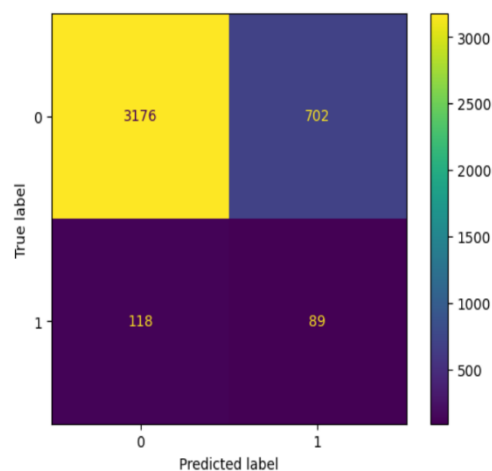
A modellek tanításához az egyes aminosav párok és tripletek gyakoriságát használtuk fel az egyes fehérjékben. Tehát az első lépés az aminosav párok megszámlálása volt minden fehérjében. (mindig csak egy „aminosavnyit” haladtunk előre, tehát pl. az MLG peptiben ML és LG külön párnak számít). Hasonló módszerrel megszámláltuk az aminosav tripleteket és ezekből készítettünk egy táblát, ami így tartalmazza az egyes aminosav párok és tripletek előfordulását is. Ezután minden fehérjéhez hozzárendeltük a GO-ban szereplő funkciót. Fontos megjegyezni, hogy egy fehérjéhez akár több funkció is tartozhatott és olyan is volt, amihez nem tartozott funkció. A funkcióprediktálás során azt szeretnénk eldönteni, a fehérje rendelkezik-e a leggyakoribb, GO:0007165 funkcióval. A Gene Ontology adatbázis humán fehérjéire vonatkozó biological process részt töltöttük le.

SYMBOL	A1BG	A2ML1	AACS	AACS	AACS	AAK1	AAK1	AAK1	AAK1	AAK1
GO NAME	biological_process	regulation of endopeptidase activity	lipid metabolic process	positive regulation of insulin secretion	fatty acid metabolic process	protein stabilization	endocytosis	regulation of protein localization	protein phosphorylation	regulation of cellular dependent endocytosis
MG	0	4	3	3	3	0	0	0	0	0
GL	2	9	3	3	3	7	7	7	7	7
LE	6	11	5	5	5	3	3	3	3	3
...	...	...	...	...	...	...	...	...	...	...
DU	0	0	0	0	0	0	0	0	0	0
UL	0	0	0	0	0	0	0	0	0	0
NU	0	0	0	0	0	0	0	0	0	0
FU	0	0	0	0	0	0	0	0	0	0

11.ábra. Részlet a feature táblából

4.1.1. A döntési fák teljesítménye

A `sklearn.tree.DecisionTreeClassifier` modellt használtuk az aminosav párok és tripletek gyakoriságát tartalmazó táblán. A tesztadat az adatunk 30%-a volt. A modell 92%-os accuracy score-ral jósol, ami elsőre nagyon jónak tűnik. A ROC AUC viszont csak 0.69. A modell alaposabb megértésében segíthet, ha vetünk egy pillantást a konfúziós mátrixra. Azt láthatjuk, hogy hajlamosabb negatív eredményt adni, tehát azt prediktálni, hogy az adott fehérjénk nem rendelkezik a GO:0007165 funkcióval. Ez valószínűleg az adat kiegyensúlyozatlansága miatt

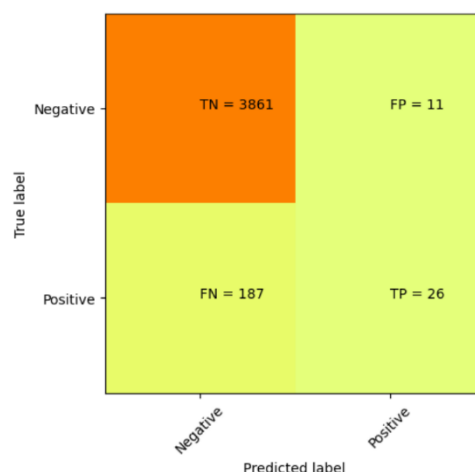


12.ábra A döntési fa konfúziós mátrixa az aminosav párok és tripletek gyakoriságát tartalmazó táblán tanítva

lehet. Összesen 20423 fehérjénk van, ebből mindössze 1062 rendelkezik a leggyakoribb funkcióval.

4.1.2. Az XGBoost teljesítménye

A kiegyensúlyozatlanság javítása végett most megpróbáljuk csak azokat a fehérjéket bevenni, amikhez ismerünk funkciót. Így 1062 GO:0007165 funkciójú fehérje van az összesen 8151 fehérjéből. A hosszas futási idő miatt ennél az algoritmusnál egyszerűsítettem a feladaton és csak az aminosav párok gyakoriságán tanítottunk. 100db maximum 5 mélységű fa teljesített a legjobban. Így az accuracy rate 95% volt, a ROC AUC pedig 77%.

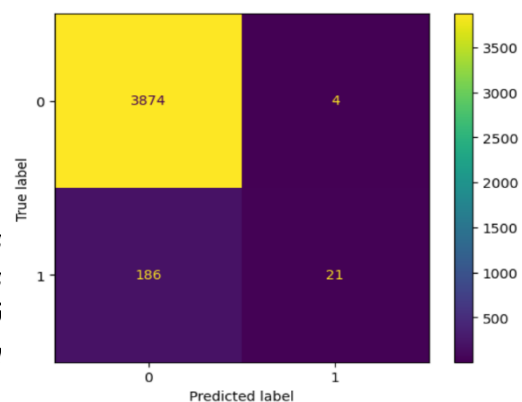


13.ábra Az XGBoost konfúziós mátrixa az aminosav párok és tripletek gyakoriságát tartalmazó táblán tanítva

4.1.3. Random Forest teljesítménye

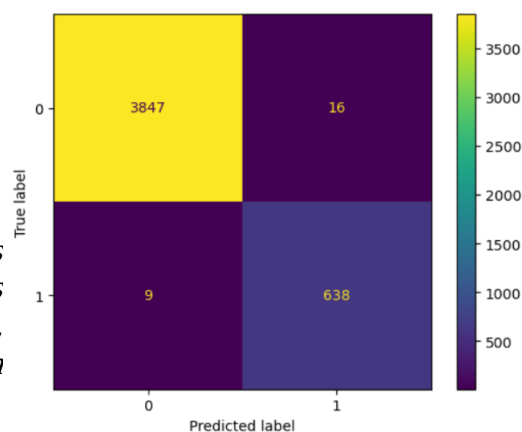
A RandomForestClassifier() modult használva 95%-os accuracy score-t értünk el. A precision 0.84, recall értéke 0.1.

14.ábra A random forest konfúziós mátrixa az aminosav párok és tripletek gyakoriságát tartalmazó táblán tanítva

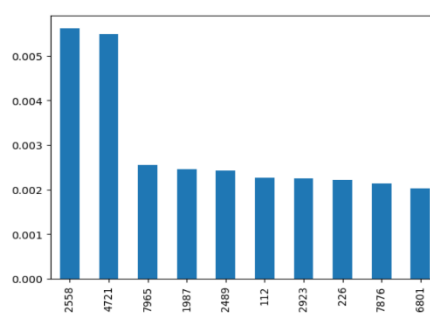


Az adathalmaz kiegyensúlyozatlanságát kiküszöbölve dúsítottuk a leggyakoribb funkcióval rendelkező halmazt, azaz minden ilyen funkciójú fehérjét háromszor vettünk az adathalmazba. Így az accuracy score 0.99-re emelkedett, a precisionn 0.98-ra és a recall 0.98-ra.

15.ábra A random forest konfúziós mátrixa az aminosav párok és tripletek gyakoriságát tartalmazó, dúsított adaton tanítva



Mivel a modell sok tulajdonsággal rendelkezik (sok pár és triplet van) az egyes tulajdonságok kis súllyal számítanak. Viszont érdemes megnézni, hogy a fontosabb tulajdonságok valójában mennyire fontosak, a modell mekkora súllyal választotta őket. Ennek biológiai jelentése lehet. Itt azonban még a legfontosabb feature is csak kevesebb, mint 0.006 súllyal számít.



16.ábra A legfontosabb tulajdonságok súlya

4.2. Öregedéssel kapcsolatos fehérjék prediktálása aminosav szekvenciából

2018-ban megjelent egy publikáció, amelyben a kutatás célja az volt, hogy a gépi tanulás segítségével az emberi fehérjét az öregedéssel kapcsolatos és nem kapcsolatos kategóriákba sorolja az öregedési folyamat szabályozásában betöltött szerepük alapján (Kerepesi et al., n.d., 2018). Az XGBoost az egyik kulcsfontosságú eszköz volt a fehérjék jellemzőinek kiválasztásában és osztályozásában.

A kutatás első lépése a jellemzők kiválasztása volt. A jellemzők közül a gépi tanulás egyik kulcsfontosságú feladata volt a releváns, de nem korreláló jellemzők azonosítása. Az XGBoost különösen jól alkalmas erre, mivel képes a legfontosabb jellemzők kiválasztására anélkül, hogy azok túlzottan összefüggnének egymással. Az algoritmus több iterációval javította az osztályozó modellt, miközben a legfontosabb jellemzőket választotta ki.

Első lépésként a kutatók a Gene Ontology (GO) jellemzők közül választották ki azokat, amelyek nem kapcsolódnak közvetlenül az öregedéssel kapcsolatos kifejezésekhez. A "GO ancestor" (GO őselemek) kifejezések hozzáadásával az AUC (Area Under the Curve) értéke jelentősen javult, 0.9086-ra nőtt, miközben 21,000 jellemzőre bővült a kiválasztott jellemzők halmaza. Ez alapot adott a következő szűrési lépésekhez.

Ezután egy kétlépcsős szűrés és a jellemzők kiválasztása következett. Az első szűrés során az XGBoost kiválasztotta a legfontosabb GO jellemzőket, és ezek közül 373 maradt, miután a modell eltávolította a kevésbé releváns jellemzőket. Ezzel az AUC érték 0.9187-re emelkedett. A következő szűrés során a modell szigorította a kiválasztási kritériumot, és végül 65 jellemző maradt, ami az AUC értékét 0.9219-re növelte. A további lépésekkel az XGBoost a kiválasztott GO jellemzők mellé hálózati és koexpressziós jellemzőket is hozzáadott, amelyek további növekedést eredményeztek az AUC értékében. A legfontosabb eredmény a "ageing_n" (öregedéssel kapcsolatos szomszédok száma) jellemző hozzáadása volt, amely tovább javította a modellt, és az AUC értéke 0.9314-re nőtt.

A végső szűrés során az XGBoost tovább finomította a modellt. A maximális fa mélységét 1-re csökkentette, hogy egyszerű, jól értelmezhető modellt kapjon. Ennek eredményeként 32 jellemző maradt, és az AUC értéke 0.9322-re emelkedett. Ezen 32 jellemző segítségével a kutatás sikeresen megvalósította a fehérjék öregedéssel kapcsolatos osztályozását, és az eredményeket a kutatók elérhetővé tették a GitHub oldalon.

Az XGBoost teljesítményét más gépi tanulási algoritmusokkal, például logisztikus regresszióval (LR), SVM-el, a legközelebbi szomszédokkal (k-NN) és döntési fákkal is összehasonlították. A kutatás eredményei szerint az XGBoost a legjobb teljesítményt nyújtotta, mivel az AUC értéke mindig meghaladta a többi algoritmusét, amelyek a teljes jellemzőkészletet alkalmazták. Ezen kívül az XGBoost képes volt az egyes jellemzők fontosságának hatékony értékelésére, és az osztályozás pontosságát jelentősen javította.

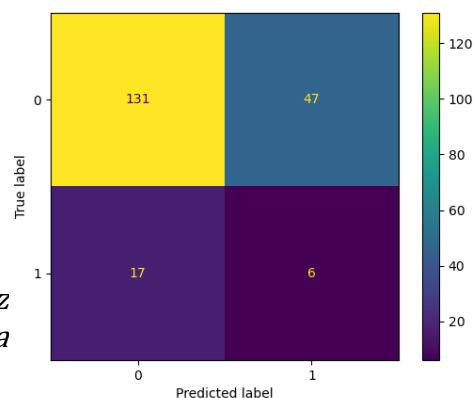
Összefoglalva, az XGBoost alkalmazása jelentős előnyöket biztosított a fehérjék öregedéssel kapcsolatos osztályozásában. Az algoritmus sikeresen kiválasztotta a legfontosabb jellemzőket, és javította a modell prediktív teljesítményét. Az AUC értéke 0.9322-re nőtt a végső modell esetén, ami magas pontosságot mutat az öregedéssel kapcsolatos fehérjék azonosításában. Az XGBoost az egyik legjobban teljesítő gépi tanulási algoritmusnak bizonyult és a kutatás alapján hasznos eszközként alkalmazható a biológiai kutatásokban, különösen az öregedés és a kapcsolódó mechanizmusok megértésében. Emiatt úgy döntöttem, az aminosav szekvencia alapján prediktálás esetén az XGBoost algoritmust használom.

4.2.1. XGBoost az aminosavak és párok gyakoriságán

Az aminosavak és aminosav-párok előfordulási gyakorisága alapján épített modell teljesítménye közepesnek mondható, mivel a maximális pontosság 75%, a minimális pontosság 65,35%, míg az átlagos pontosság 70,16% volt, 3,07%-os szórással. Az eredmények stabilnak tekinthetők, azonban a 0,561-es ROC AUC érték arra utal, hogy a modell csak kissé teljesít jobban, mint a véletlenszerű

találgatás. Ez azt mutatja, hogy az aminosavak és párok előfordulási gyakorisága önmagában nem elegendő az öregedéssel kapcsolatos fehérjék megbízható elkülönítésére.

17. ábra XGBoost konfúziós mátrixa az aminosavak és párok gyakoriságán tanítva



	if_aging	M	G	L	E	A	V	P	I	F	...	UY	UQ	UR	AU	RU	CU	US	DU	NU	FU
BMAL1_HUMAN	1	19	49	49	32	33	29	43	36	21	...	0	0	0	0	0	0	0	0	0	0
CAC1A_HUMAN	1	85	193	208	160	190	129	172	114	125	...	0	0	0	0	0	0	0	0	0	0
SOC2_HUMAN	1	3	16	23	13	12	7	11	8	4	...	0	0	0	0	0	0	0	0	0	0
TERT_HUMAN	1	12	75	147	45	99	88	87	23	47	...	0	0	0	0	0	0	0	0	0	0
IKKB_HUMAN	1	22	37	97	67	38	40	33	34	18	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

18. ábra. Részlet a feature táblából

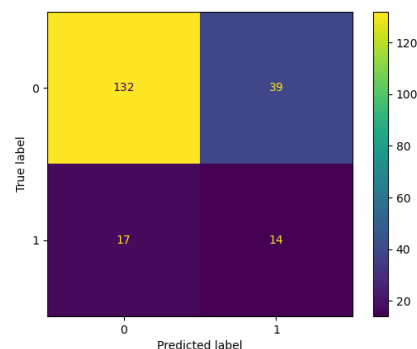
	max_d	n_est	n_exps	mean AUC	std AUC	mean ACC	std ACC	mean sens	std sens	mean spec	std spec
17	3	100	10	0.614992	0.008260	0.686468	0.006943	0.254098	0.023472	0.874857	0.010713
23	5	100	10	0.613590	0.011047	0.694527	0.005987	0.239016	0.017065	0.893000	0.009421
11	2	100	10	0.608750	0.011278	0.679005	0.009272	0.246557	0.017886	0.867429	0.008940
22	5	20	10	0.607225	0.015378	0.687363	0.010047	0.226885	0.020829	0.888000	0.013236

19. ábra. A legjobban teljesítő modellek. A legkiemelkedőbb 100db 3 mélységű fa volt.

4.2.2. XGBoost az aminosav párok gyakoriságán

Az aminosav-párok előfordulási gyakorisága alapján épített modell teljesítménye meglepő módon kissé javult az előzőhöz képest, hiszen a maximális pontosság 77,23%, a minimális pontosság 63,37%, míg az átlagos pontosság 69,48% lett,

azonban a szórás megnőtt, elérve a 4,4%-ot. A ROC AUC érték 0,593-ra emelkedett, ami továbbra is csak mérsékelt prediktív képességet jelez, de jelzi a párok valamivel nagyobb informativitását az öregedéssel kapcsolatos fehérjék elkülönítésében. Az eredmények azt mutatják, hogy a párok használata önmagában még nem elégséges a magas teljesítményű modell létrehozásához. A további javulás érdekében célszerű lehet az aminosav-tripletek előfordulását is vizsgálni.



20. ábra XGBoost konfúziós mátrixa az aminosav párok gyakoriságán tanítva

	if_aging	MG	GL	LE	EA	AL	LV	VP	PL	LA	...	KU	UE	EU	CU	US	DU	UL	NU	FU	UI
CP2D7_HUMAN	0	2	7	2	6	2	3	4	5	9	...	0	0	0	0	0	0	0	0	0	0
PIOS1_HUMAN	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
SIK1B_HUMAN	0	0	13	7	2	3	4	3	8	7	...	0	0	0	0	0	0	0	0	0	0
MOTSC_HUMAN	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

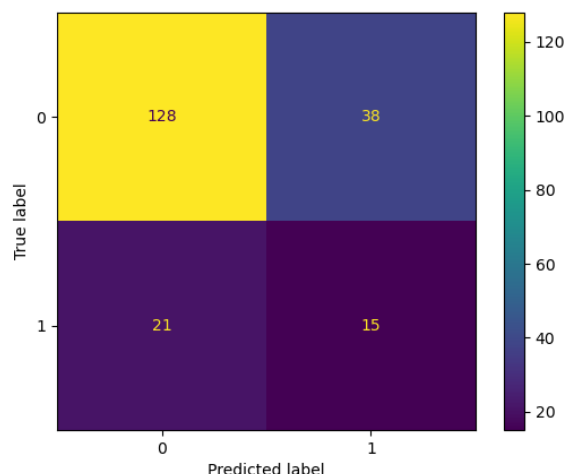
21. ábra. Részlet a feature táblából

max_d	n_est	n_exps	mean AUC	std AUC	mean ACC	std ACC	mean sens	std sens	mean spec	std spec
5	100	10	0.592629	0.016662	0.681312	0.007253	0.206863	0.022167	0.888714	0.007914
1	20	10	0.589818	0.010578	0.694533	0.002813	0.027778	0.004679	0.986000	0.003603
3	100	10	0.587681	0.009142	0.674155	0.009646	0.240523	0.018556	0.863714	0.011539
2	100	10	0.586110	0.014461	0.675646	0.006844	0.221242	0.015879	0.874286	0.009012
1	10	10	0.582615	0.011622	0.694831	0.001722	0.005229	0.004185	0.996286	0.002571
3	20	10	0.580521	0.013778	0.676740	0.007975	0.167320	0.018125	0.899429	0.007346

22. ábra. A legjobban teljesítő modellek. A legkiemelkedőbb 100db 5 mélységű fa volt.

4.2.3. XGBoost az aminosav tripletek gyakoriságán

A tripletek előfordulási gyakorisága alapján tanított modell teljesítménye vegyes eredményeket mutatott. A maximális pontosság 72,28%, a minimális pontosság 61,39%, míg az átlagos pontosság 68,89% lett, 3,3%-os szórás mellett. A ROC AUC érték 0,629-ra emelkedett, ami a modell prediktív képességének mérsékelt, de érezhető javulását jelzi az előző modellekhez képest. A tripletek használata úgy tűnik, hogy több információt képes megragadni az öregedéssel kapcsolatos fehérjék



23. ábra XGBoost konfúziós mátrixa az aminosav tripletek gyakoriságán tanítva

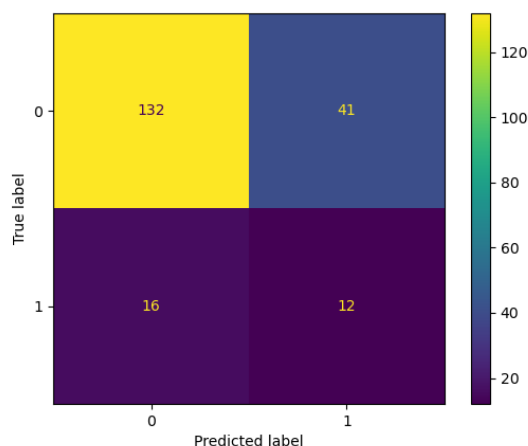
vagy párok, ami a tripletek gazdagabb kontextuális információtartalmára utal. Azonban a pontosság továbbra sem kiemelkedő, és a modell teljesítménye ingadozik, ami arra utal, hogy a tripletek önmagukban nem elegendőek az öregedéssel kapcsolatos fehérjék pontos predikciójához. A további fejlődéshez célszerű lenne kombinálni az egyedi aminosavakat, párokat vagy tripleteket.

max_d	n_est	n_exps	mean AUC	std AUC	mean ACC	std ACC	mean sens	std sens	mean spec	std spec
5	100	10	0.615232	0.008705	0.683797	0.009703	0.238889	0.017926	0.878286	0.012219
3	100	10	0.614879	0.012027	0.693638	0.008397	0.252614	0.017047	0.886429	0.013190
2	100	10	0.609706	0.011367	0.694732	0.011294	0.244771	0.010381	0.891429	0.014832
5	20	10	0.599656	0.015716	0.686978	0.010354	0.216993	0.022366	0.892429	0.012698

24. ábra A legjobban teljesítő modellek. A legkiemelkedőbb 100db 5 mélységű fa volt.

4.2.4. XGBoost az aminosav párok és triplettek gyakoriságán

A párok és triplettek együttes figyelembevételével készült modell teljesítménye valamelyest elmaradt a korábbi próbálkozásokhoz képest. A maximális pontosság 71,0%, a minimális pontosság 64,36%, míg az átlagos pontosság 67,76% lett, alacsony, mindössze 1,75%-os szórással. A ROC AUC értéke 0,617, ami azt mutatja, hogy a modell prediktív képessége közepes maradt, és a párok és triplettek kombinációja nem hozott jelentős javulást az egyéni felhasználásukhoz képest. Ez az eredmény arra utal, hogy a párok és triplettek kombinált információtartalma nem képes megfelelően kiegészíteni egymást a fehérjék öregedéssel való kapcsolatának pontosabb előrejelzésében. A modell stabilabb, alacsonyabb szórású teljesítménye ugyanakkor pozitív jel, de az általános pontosság és ROC AUC továbbra sem kielégítő.



25. ábra. XGBoost konfúziós mátrixa az aminosav párok és triplettek gyakoriságán

a párok és triplettek kombinációja nem hozott jelentős javulást az egyéni felhasználásukhoz képest. Ez az eredmény arra utal, hogy a párok és triplettek kombinált információtartalma nem képes megfelelően kiegészíteni egymást a fehérjék öregedéssel való kapcsolatának pontosabb előrejelzésében. A modell stabilabb, alacsonyabb szórású teljesítménye ugyanakkor pozitív jel, de az általános pontosság és ROC AUC továbbra sem kielégítő.

max_d	n_est	n_exps	mean AUC	std AUC	mean ACC	std ACC	mean sens	std sens	mean spec	std spec
5	100	10	0.606683	0.010413	0.687761	0.009655	0.196393	0.017315	0.901857	0.012794
3	100	10	0.588249	0.012343	0.678607	0.010963	0.199344	0.015573	0.887429	0.011966
2	100	10	0.581785	0.020590	0.680299	0.010216	0.196393	0.019914	0.891143	0.009727
5	20	10	0.569240	0.016574	0.677214	0.009034	0.172131	0.023835	0.897286	0.010090

26. ábra A legjobban teljesítő modellek. A legkiemelkedőbb 100db 5 mélységű fa volt.

4.3. További ötletek: transformer nyelvi modellek

A Transformer nyelvi modellek alkalmazása a fehérje szekvenciák elemzésére ígéretes irány lehet, mivel ezek a modellek rendkívül hatékonyan képesek megragadni a szekvenciális és kontextuális információkat, amelyek fontosak lehetnek a fehérjék öregedéssel kapcsolatos tulajdonságainak predikciójához.

1. Szekvenciális és hosszú távú kapcsolatok felismerése

A Transformer modellek (pl. BERT, GPT, T5) képesek figyelmi mechanizmusokat (self-attention) használni, amelyek a teljes szekvenciában felfedezik az elemek közötti kapcsolatokat. Ez különösen előnyös aminosav-szekvenciák esetén, mivel a távoli pozíciókban lévő aminosavak közötti interakciók is befolyásolhatják a fehérje tulajdonságait. Korábbi modellek, mint a kNN vagy XGBoost, nem kezelik jól az ilyen hosszú távú függőségeket.

2. Összetett mintázatok automatikus felismerése

A Transformer modellek képesek automatikusan reprezentációkat tanulni a szekvenciákból anélkül, hogy manuálisan kellene kiválasztani a párok és tripletek mintázatait. Például, a modell képes felismerni, hogy bizonyos aminosav-mintázatok vagy ismétlődések fontosabbak, mint mások, és ezeket előnyben részesíti a predikció során. Ez amiatt is érdekes lehet, mert bizonyos aminosavak, párok és tripleteket szignifikánsan többször fordulnak elő az öregedéssel kapcsolatos fehérjékben. Ha egy jól teljesítő modell szerint is fontos jellemzők, az ráerősítene a statisztika elvégzéséből származó eredményekre.

3. Fehérje-specifikus pre-trained modellek

Ma már léteznek speciálisan fehérjeszekvenciákra tanított Transformer modellek, mint például:

- PreciousGPT: A GPT architektúrát adaptálta biológiai és kémiai adatokra, lehetővé téve a komplex molekuláris struktúrák és interakciók előrejelzését.
- ESM (Evolutionary Scale Modeling): Fehérje-szekvenciák előrejelzésére fejlesztett modell.
- ProtBERT: A BERT architektúrát adaptálta fehérje-adatokra.
- AlphaFold: Fehérjék térszerkezetének előrejelzésére is használja.

Ezek a modellek a fehérjék evolúciós információját és biológiai sajátosságait is figyelembe veszik.

4. Többdimenziós kontextus kezelés

A Transformer modellek nemcsak a sorrendet veszik figyelembe, hanem a különböző szekvenciák közötti hasonlóságot is elemzik. Ezáltal jobban megragadható a fehérjék közötti strukturális vagy funkcionális hasonlóság.

5. Skalázhatóság és teljesítmény

A modern Transformer modellek nagy mennyiségű adatot képesek hatékonyan feldolgozni párhuzamosan. Ez jelentős előnyt jelent nagy fehérjeszekvencia-adathalmazok esetén.

Finomhangolhatók (fine-tuning) kisebb specifikus adathalmazokra, például az öregedéssel kapcsolatos fehérjékre, miközben kihasználják a pre-trained tudást.

A Transformer modellek alkalmazása jelentős előrelépést jelenthet, mivel hatékonyan kezelik a fehérje szekvenciák bonyolultságát és a hosszú távú kapcsolatokat. Különösen a pre-trained fehérje-specifikus modellek (pl. ProtBERT vagy PreciousGPT) alkalmazása lehet ígéretes, mivel ezek már tartalmazznak biológiailag releváns tudást. A korábbi megközelítésekkel szemben a Transformer modellek mélyebb és részletesebb reprezentációkat képesek tanulni, ami végső soron javíthatja a predikció pontosságát és az öregedéssel kapcsolatos fehérjék azonosítását.

5. Összefoglalás

Az élettartam és fiatal életszakasz meghosszabbítása mindig is foglalkoztatta az emberiséget. Az egészségesebb öregedés nemcsak magasabb életszínvonalat biztosít az egyén számára, a társadalom számára is hasznos. Ezért ebben a szakdolgozatban az öregedéssel kapcsolatos fehérjék aminosavszekvenciáinak elemzésével foglalkozom publikus adatbázisok használatával remélve, hogy ez segíti az öregedés folyamatának megértését. Statisztikai és gépi tanulási módszerekkel vizsgáltuk a különbségeket az öregedéssel kapcsolatos és azzal nem kapcsolatos fehérjék között. Kiszámoltuk az egyes aminosavak, aminosav párok és tripletek relatív gyakoriságát a két halmazban, majd statisztikai próbával igazoltuk, hogy bizonyos pepditek esetén szignifikáns különbség van. Ezután egy bináris klasszifikációs feladatot oldottunk meg, ahol azt próbáltuk prediktálni, mely fehérjék rendelkeznek a leggyakoribb, a GO:0007165 funkcióval. Ehhez döntési fákat, a random forest és az XGBoost algoritmust használtuk. Majd a feladatra valószínűleg legalkalmasabb, az XGBoost algoritmus segítségével próbáltuk prediktálni aminosavszekvencia alapján, mely fehérjék öregedéssel kapcsolatosak és melyek nem. Ehhez többféle feature táblát is készítettünk, amelyek az egyes aminosavak, aminosav párok és tripletek előfordulását tartalmazzák.

6. Irodalomjegyzék

1. Annemieke T. van der Goot and Ellen A.A. Nollen. (2013). Tryptophan metabolism: entering the field of aging and age-related pathologies. *Cell Press*.
2. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:1, 25(1), 25–29. <https://doi.org/10.1038/75556>
3. Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-Ajee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., ... Zhang, J. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169. <https://doi.org/10.1093/NAR/GKW1099>
4. Canfield, C. A., & Bradshaw, P. C. (2019). Amino acids in the regulation of aging and aging-related diseases. *Translational Medicine of Aging*, 3, 70–89. <https://doi.org/10.1016/J.TMA.2019.09.001>
5. Consortium, T. G. O., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., ... Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1). <https://doi.org/10.1093/GENETICS/IYAD031>
6. Coschigano, K. T., Clemmons, D., Bellush, L. L., & Kopchick, J. J. (2000). Assessment of growth parameters and life span of GHR/BP gene-disrupted mice. *Endocrinology*, 141(7). <https://doi.org/10.1210/endo.141.7.7586>
7. Gladyshev, V. N., Anderson, B., Barlit, H., Barré, B., Beck, S., Behrouz, B., Belsky, D. W., Chaix, A., Chamoli, M., Chen, B. H., Cheng, K., Chuprin, J., Churchill, G. A., Cipriano, A., Colville, A., Deelen, J., Deigin, Y., Edmonds, K. K., English, B. W., ... Zhavoronkov, A. (2024). Disagreement on foundational principles of biological aging. *PNAS Nexus*, 3(12). <https://doi.org/10.1093/PNASNEXUS/PGAE499>
8. Hsu, P. D., Lander, E. S., & Zhang, F. (n.d.). *Leading Edge Review Development and Applications of CRISPR-Cas9 for Genome Engineering*. <https://doi.org/10.1016/j.cell.2014.05.010>

9. Inês Lopes 1, G. A. 1, P. R. 1, J. P. de M. 1. (2021). Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Frontiers in Genetics*, 12.
10. Kenyon, C., Chang, J., Gensch, E., Rudner, A., & Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. *Nature*, 366(6454), 461–464. <https://doi.org/10.1038/366461A0>
11. Kerepesi, C., Daróczy, B., Sturm, Á., Vellai, T., & Benczúr, A. (n.d.). *Prediction and characterization of human ageing-related proteins by using machine learning OPEN*. <https://doi.org/10.1038/s41598-018-22240-w>
12. Ladiges, W., Van Remmen, H., Strong, R., Ikeno, Y., Treuting, P., Rabinovitch, P., & Richardson, A. (2009). Lifespan extension in genetically modified mice. In *Aging Cell* (Vol. 8, Issue 4). <https://doi.org/10.1111/j.1474-9726.2009.00491.x>
13. Leenders, M., & Van Loon, L. J. (2011). *Leucine as a pharmaconutrient to prevent and treat sarcopenia and type 2 diabetesn ure_443 675..689*. <https://doi.org/10.1111/j.1753-4887.2011.00443.x>
14. Lin, Y. J., Seroude, L., & Benzer, S. (1998). Extended life-span and stress resistance in the Drosophila mutant methuselah. *Science*, 282(5390). <https://doi.org/10.1126/science.282.5390.943>
15. Patricia M. Helfman; Jeffrey L. Bada; Ming-Yung Shou. (1977). Considerations on the Role of Aspartic Acid Racemization in the Aging Process. *Karger*, 419–425.
16. Pedro De Magalhães, J., Curado, J., & Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *BIOINFORMATICS ORIGINAL PAPER*, 25(7), 875–881. <https://doi.org/10.1093/bioinformatics/btp073>
17. Szatmári, E. Z., Csordás, A., & Kerepesi, C. (2024). Unique Patterns in Amino Acid Sequences of Aging-Related Proteins. *Advanced Biology*, 8(2), 2300436. <https://doi.org/10.1002/ADBI.202300436>
18. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1). <https://doi.org/10.1093/nar/gkab1061>
19. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool,

- K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., ... Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1). <https://doi.org/10.1093/nar/gkad1011>
20. Wang, J., Mukherjee, S., & Zubarev, R. A. (2022). Isoaspartate and neurodegeneration. *Aging*, 14(22), 8882–8883. <https://doi.org/10.18632/AGING.204420>
21. Yan-Hui Li 1, M.-Q. D. Z. G. (2010). Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*. *Mech Ageing Dev.* .