MAGYAR AGRÁR- ÉS
ÉLETTUDOMÁNYI EGYETEM

Mitokondriális genomikai és proteomikai vizsgálatok fejlesztése

Doktori (PhD) értekezés

Biró Bálint

Gödöllő

2022

A doktori iskola

megnevezése: Állatbiotechnológiai és Állattudományi Doktori Iskola

tudományága: Állattenyésztési tudományok

vezetője: Dr. Mézes Miklós
egyetemi tanár, az MTA rendes tagja
Magyar Agrár- és Élettudományi Egyetem, Szent István Campus, Élettani és Takarmányozási Intézet, Takarmánybiztonsági Tanszék

Témavezető(k): Dr. Hoffmann Orsolya Ivett
tudományos főmunkatárs, Ph.D.
Magyar Agrár- és Élettudományi Egyetem, Szent István Campus, Genetika és Biotechnológia Intézet, Állatbiotechnológia Tanszék

.....	.....
Iskolavezető jóváhagyása	Témavezető jóváhagyása
Dr. Mézes Miklós	Dr. Hoffmann Orsolya Ivett
akadémikus	tudományos főmunkatárs

1. Tartalomjegyzék

2. Jelölések, rövidítések jegyzéke

C	coil másodlagos szerkezet
DSB	Double Stranded Break, duplaszálú DNS törés
E	β-redő másodlagos szerkezet
EGT	Endosymbiotic Gene Transfer
EC	extracelluláris tér
FNR	False Negative Rate, fals negatív arány
FPR	False Positive Rate, fals pozitív arány
H	hélix másodlagos szerkezet
IC	intracelluláris tér
kb	kilobázis
LR	Learning Rate, tanulási ráta
ML	Machine Learning, gépi tanulás
MAE	Mean Absolute Error, átlagos abszolút hiba
mRNS	Messenger RNA, hírvivő RNS
mtDNS	mitkondriális genom
NCBI	National Center for Biotechnology Information
NHEJ	Non Homologous End Joining, nem homológ végeket összekapcsoló DNS hibajavító mechanizmus
NMR	Nuclear Magnetic Resonance spectroscopy, nukleáris mágneses rezonancia spektroszkópia
NUMT	nuclear mitochondrial sequences, mitokondriális eredetű sejtmagi szekvencia
PCA	Principal Component Analysis, főkomponens analízis
PCC	Pearson féle korreláció
PP	Perplexity Value, zavarosság
PINK1	phosphatase and tensin homologue induced putative kinase 1
PARL	protease presenilin associated rhomboid like protein
PDB	Protein Data Bank

RBF	Radial Basis Function, radiális bázisfüggvény
ROC	Receiver Operating Characteristic curve, karakterisztika görbe
RSA	Relative Solvent Accessibility, relatív savmaradék szintű oldószer számára való hozzáférhetőség
SOV	Segment OVerlap
gDNS	sejtmagi genom
SA	Solvent Accessibility, savmaradék szintű oldószer számára való hozzáférhetőség
SCC	Spearman féle rangkorreláció
SVM	Support Vector Machine, támasz vektor gép
tSNE	T-distributed Stochastic Neighbor Embedding
TNR	True Negative Rate, valódi negatív arány
TPR	True Positive Rate, valódi pozitív arány
UMAP	Unifold Manifold Approximation and Projection for Dimension Reduction
X-ray-Cr	X-ray crystallography, röntgen krisztallográfia
3D	3 dimenziós

3. Bevezetés

A XX. század derekán és második felében az élettudománnyal foglalkozóknak az adott biológiai rendszer vizsgálatához szükséges módszerek hiánya jelentette a legnagyobb kihívást. Ebben az időszakban számos olyan metódust fejlesztettek ki a kutatók, amik gyökeresen megváltoztatták a tudományokat. Ennek a kornak köszönhetjük a molekuláris biológiai módszerek ugrásszerű fejlődését, többek között a nukleinsav szekvenciák meghatározásának egyik mai napig is használt módszerének, a Sanger szekvenálásnak a felfedezését (Heather & Chain, 2016), de a fehérjék aminósav összetételének és szerkezetének jellemzésével kapcsolatban is születtek figyelemreméltó eredmények, például a röntgen krisztallográfiát és a nukleáris mágneses rezonancia spektroszkópiát (Nuclear Magnetic Resonance spectroscopy) is ekkor kezdték elterjedtebben használni a 3-dimenziós (3D) szerkezetek meghatározására (Edman & Begg, 1967; Jaskolski et al., 2014; Campbell, 2013). Ezek az újszerű eljárások, amik lehetővé tették egy-egy molekula több szempontból történő mélységi analízisét, szemléletváltásra kényszerítették a kutatókat, akik elsajátították a technológiákat, megértették a működési hátterüket és átültették a módszereket a mindennapi gyakorlatba (Gilbert, 1991).

Az említett módszerek mindegyike forradalminak számított a maga korában, azonban használatuk összetettsége (az infrastruktúra, számítási kapacitás, adattárolási kapacitás stb szempontjából) és a fizikailag kis áteresztőképességük miatt széles körű elterjedésük váratott magára.

A tudományos közösség az ezredforduló környékén érte el a technológiának azt a fejlettségi fokát, ami lehetővé tette a szekvenálás körülményes metodikai fegyvertárának leegyszerűsítését és gördülékenyebb alkalmazását (Pauwels et al., 1995). Ennek a fejlődésnek köszönhetően a szekvenálás kivitelezése mára mindennapos rutinná vált. Az egyre magasabb fokú automatizációnak köszönhetően óriási mennyiségű nukleinsav és fehérje szekvencia adat jött létre és keletkezik folyamatosan (O'Leary et al., 2016; "UniProt: the universal protein knowledgebase in 2021", 2021). A biológiai adat mennyiségének ilyen mértékű növekedése elhozta az adattudományi szemléletet az élettudományokba is, újabb paradigmaváltást eredményezve (D'Argenio, 2018; Pal et al., 2020). A jelenkor technológiája már alkalmas az adatok tárolására, kinyerésére, minőségük ellenőrzésére és a bennük rejlő szabályszerűségek jellemzésére (Marx, 2013; Leonelli, 2019). Ugyanakkor megfigyelhető az a tendencia, hogy a funkcionálisan jellemzett nukleinsav szekvenciák száma nagyságrendekkel elmarad a szekvenálásból származó adatok méretétől (Salzberg, 2019). Fehérjék esetén is hasonló a helyzet, hiszen az adatbázisokban elérhető aminósav szekvenciák és az ismert szerkezetű, funkciójú fehérjék számának dimenziói nem összevethetők (Schwede, 2013). Tehát a szerkezet és funkció meghatározás módszereinek fejlődése nem tartott lépést a szekvenálás módszereinek

fejlődésével. Ebből adódóan a kortárs kutatók egyik legfontosabb feladata, hogy megpróbálják jelentéssel felruházni a nyers szekvencia adatokat. A nyers szekvenciák és a valamilyen szempontból jellemzett szekvenciák száma közötti szakadékok áthidalása kizárólag laboratóriumi módszerekkel szinte kivitelezhetetlen a módszerek időigényéből és összetettségéből adódóan. Azonban a számítástechnika előrehaladásával elérhetővé váltak olyan módszerek (gépi tanulás, machine learning-ML alapú metódusok), amik már feltérképezett motívumokból nyert tudás alapján képesek becsülni különböző jellemzőket olyan adatstruktúrákból, amelyek ember számára értelmezhetetlenek pusztán nagyságukból és komplexitásukból adódóan (pl.: interakciós adatbázisok).

Doktori disszertációmban két biológiai területet érintek.

Az első, genomikai jellegű modulban a mitokondriális eredetű sejtmagi szekvenciákat (nuclear mitochondrial sequences-NUMT) (Lopez et al., 1994) vizsgáljuk. A NUMTok jelentőségét számos olyan daganattípus esetén megállapították ahol a NUMTok beépülése tumorszupresszort és/vagy onkogént érint (Ju et al., 2015; Singh et al., 2017; Srinivasainagendra et al., 2017; Palodhi et al., 2020; Wei et al., 2022). A tumorbiológián túl a NUMTok fontos felhasználási területei a különféle filogenetikai (Ko et al., 2015; Nacer & do Amaral, 2017) és igazságügyi vizsgálatok (Marshall & Parson, 2021; Cortes-Figueiredo et al., 2021). Friss kutatási eredmények alapján a NUMTok kiemelt szerepet játszanak a mitokondriális genom (mtDNS) módosítását lehetővé tevő célzott genomszerkesztési eljárások sejtmagi DNS-ben (gDNS) bekövetkező OFF-target hasításai során is (Lei et al., 2022).
A NUMTokat már több fajban vizsgálták változatos módszerekkel, többek között humánban (Dayama et al., 2014), kutyában (Verscheure et al., 2015), macskában (Lopez et al., 1994) stb. Azonban a NUMTok leírása a nyúl genomban ez idáig még nem történt meg. Azért is fontos a NUMTok leírása ebben a fajban, mert tanulmányok igazolták, hogy sok esetben a nyulak jobb betegségmodellnek bizonyultak, mint a hagyományosan használatos rágcsálók vagy főemlősök (Esteves et al., 2018; Fan et al., 2018; Matsuhisa et al., 2020; Fan et al., 2021). Annak ellenére, hogy a NUMTokat már több fajban is leírták, az átfogóbb jellegű kutatások sok esetben egy önkényesen megválasztott taxonómiai egység karakterizálásával foglalkoznak (G. Zhang et al., 2021; Calabrese et al., 2017; Tsuji et al., 2012). Ráadásul ezeknek a kutatásoknak nincsen egy általánosan elfogadott, egységesített módszertani háttere. Az izolált vizsgálatok és az eltérő módszerek alkalmazása miatt a témában publikált eredmények egymással nem összevethetők.

A doktori kutatásom során érintett második terület a proteomikában használt néhány fontos ML modell értékelése. Ebben a modulban fehérje molekulák négy jellegzetességét (másodlagos térszerkezet, savmaradék szintű oldószer számára való hozzáférhetőség és szintén savmaradék szintű nukleinsavakkal történő interakciós valószínűség) becslő ML modell komplementaritását vizsgál-

tuk az elérhető kísérleti adatokkal összevetve a humán proteóm esetén (McGuffin et al., 2000; Faraggi et al., 2014; Yan & Kurgan, 2017). A fehérjék említett jellegzetességei széles felhasználási területtel rendelkeznek kezdve a különböző kórokozókkal való kapcsolat kialakulásával (Kruglikov et al., 2021), a natív térszerkezet elérésén keresztül (Savojardo et al., 2021) egészen a centrális dogmát érintő alapkutatásokig (Cozzolino et al., 2021). Ezeknek a tulajdonságoknak a nagy áteresztőképességű, kísérletes úton történő meghatározása nem kivitelezhető, ezért van szükség a már elérhető eredményeken alapuló ML modellek használatára. Az ML egy gyűjtőnév, ami olyan már meglévő adatokon alapuló modellek létrehozását és tesztelését jelenti, amelyek képesek felismerésre, osztályozásra és becslésre (Tarca et al., 2007). A modellek minőségét minden esetben becsülni szükséges. A minőség becsléséhez a legelterjedtebb módszer, hogy a modell számára egy addig ismeretlen adathalmaz (tesztelő adathalmaz) osztálycímkéit kell előre jelezni. Az ML modelleket általában kisebb, valamilyen szempont alapján már szelektált adathalmazokon szokták betanítani és tesztelni is. Az általános, egységesített, és nem bizonyos problémákra kialakított adathalmaz hiánya azért jelent problémát, mert az egy-egy feladatot jól megoldó modellek feltehetően más pontosságot adnak eltérő adatok esetén.

A doktori disszertációmban a NUMTok biológiáját és a proteomikát az adattudományi szemlélet köti össze. A NUMTok esetén a különböző vizsgálatok mellett az ML modellek teljes fejlesztési sémáját alkalmaztuk a bementi paraméterek kiválasztásától kezdve, a modell választáson és tesztelésen át egészen a bemenetek és a kiválasztott modell finomhangolásáig. Míg a proteomikai részben a különböző ML modellek teljesítményének tesztelését végezzük el egy egységesített adathalmazon, a humán proteómon.

A NUMTokat érintő célkitűzéseink voltak, hogy leírjuk a nyúl genomban fellelhető mitokondriális eredetű inszerciókat és a nyúl genomom beállított vizsgálatokat kiterjesszük a fellelhető emlős genomokra. Ehhez elsődleges feladatként egy nagy áteresztőképességű és kellően robosztus algoritmus létrehozását fogalmaztuk meg.

A proteomikai modul legfontosabb célkitűzése volt, hogy a humán proteóm kísérleti adathalmazzal összevetve megvizsgáljuk a másodlagos szerkezetet, a savmaradék szintű oldhatóságot és nukleinsav kötést előrejelző modellek komplementaritását. Távlati célunk, hogy a létrehozott algoritmust kiterjesszük több proteómra és az így nyert eredményeket a DescribePROT adatbázisban elérhetővé tegyük. A távlati cél megvalósulását jelen dolgozatban nem mutatom be.

4. Irodalmi áttekintés

4.1. A ML általános áttekintése

4.1.1. Definíció és felosztás

Az ML a mesterséges intelligencia egyik ága, amely magába foglalja az olyan algoritmusokat és statisztikai modelleket, amelyek ismert példákon keresztül képesek valamilyen jellegű teljesítményük fejlesztésére, azaz tanulásra (H. Wang et al., 2009; Mahesh, 2020; Fradkov, 2020). A statisztika és az ML sok közös tulajdonsággal rendelkezik, azonban fontos különbség, hogy míg a statisztika az adott rendszer leírására adatokon alapuló módon létrehoz és tesztel egy hipotézist, addig az ML adatokból kiindulva adott rendszer jövőbeli viselkedését írja le (Bzdok et al., 2018).

Az ML modelleket alapvetően négy nagy részre lehet osztani a tanulás módja alapján. Felügyelt tanulás esetén a modell ismeri az összetartozó bement-kimenet párokat és ez alapján a tudás alapján megpróbál felépíteni egy kapcsolatot a párok között. A tanulás célja, hogy a leírt kapcsolatnak megfelelően egy addig ismeretlen bemenet kimenetét (osztálycímke) meghatározza a modell. Ezzel szemben a felügyelet nélküli tanulás során nem definiáltak a bemenet-kimenet párok. Ebben az esetben a modell a bemeneti adathalmaz belső szerkezetét próbálja meg jellemezni az egyes adatpontok közötti kapcsolatok feltárásával (Osarogiagbon et al., 2021). A felügyelt és a felügyelet nélküli tanulás stratégiáinak kombinációja a félig-felügyelt tanulás, ami a tanulás folyamata során címkézett és címke nélküli adatokat is hasznosít (Sarker, 2021). Az ML negyedik nagy csoportja a megerősítéses tanulás, ahol a modell a környezetével lép interakcióba és a döntései következményeinek megfelelően módosítja a viselkedését (Osarogiagbon et al., 2021; Sarker, 2021).

Az ML algoritmusok másik elterjedt osztályozási módja az adott modell kimeneti változójának jellegén alapul. A kategórikus kimenettel jellemezhető modellek klasszifikációt, míg a folytonos kimenetű modellek regressziós analízist végeznek (Osarogiagbon et al., 2021).

4.1.2. Történelmi áttekintés

Az ML első általános célú felhasználása az 1950-es évekre datálható, amikor a Cornell Egyetemen létrehoztak egy algoritmust, ami képes volt betűk elkülönítésére. Ezt a modellt a későbbiekben perceptronnak nevezték el és ez a modell tekinthető a mai napig is használatos, főként felügyelt tanulás alapon működő neurális hálók elődjének. A preceptron rendelkezett biológiai relevanciával is, mert megalkotása alapjául az élőlények tanulási sémája szolgált (Mahesh, 2020). A későbbiekben már tényleges élettudományi problémák megoldására is alkalmazták ezeket a módszereket.

4.1.3. Felhasználási lehetőségek

ML modelleket számos tudományterületen alkalmaztak már sikerrel főként olyan komplex problémák esetén, ahol az adott jelenség háttere kevésbé volt jól jellemzett (Bzdok et al., 2017).

Például az E. coli modell organizmus transzlációs iniciációs szekvencia motívumát perceptron segítségével határozták meg már ismert hírvivő RNS (messenger RNA-mRNS) molekulák adathalmazának felhasználásával. A vizsgálatok alapján a szekvenciák ML alapú elemzése hatékonyabbnak bizonyult, mint a korábbi módszerek. Ezt az eredményt kizárólag a mRNS molekulák szekvenciáját felhasználva sikerült elérni (Stormo et al., 1982).

Ezek a módszerek jól teljesítenek nagy sálán értelmezett esetekben is (Muzio et al., 2021). Az ezredforduló tájékán írtak le egy olyan fehérje csoportot, ami ellent mondott az addig helyesnek elfogadott ”kulcs-zár” hipotézisnek (Uversky, 2019). Ezek a szerkezet nélküli fehérjék azóta bizonyított módon részt vesznek számos biológiai funkció kialakításában, fenntartásában. A szerkezet nélküli fehérjék proteómjának leírásakor is ML alapú módszerek (főként neurális hálók) kombinációját alkalmazták (Zhao et al., 2021).

Az eddig bemutatott példák leginkább olyan eredményekkel szolgáltak, amik a gyakorlatba közvetlenül nem átültethetők inkább alapkutatás jellegűek. Ugyanakkor nagyszámú ipari alkalmazással találkozhatunk, amik ML alapú módszereket hasznosítanak. Például a gyógyszeripar szempontjából nagy jelentőséggel bíró membrán fehérjék lokalizációját és termostabilitását is sikeresen becsülték egy felügyelt tanulási metódussal, a támasz vektor géppel (Support Vector Machine-SVM) (K. K. Yang et al., 2019). Ezekben az esetekben is nagyrészt szekvenciából származtatott adatokat használtak az ML tanítására.

A felügyelet nélküli tanulás módszereit pedig főként a feltáró adatelemzés során használják az adatszerkezet vizualizációjára, az esetlegesen meglévő kapcsolatok kimutatására. Például az omika tudományok területén nagy népszerűségnek örvend a főkomponens analízis (Principal Component Analysis-PCA) és a t-elosztott sztochasztikus szomszédos beágyazódás (t-distributed stochastic neighbor embedding-t-SNE) (Xu & Jackson, 2019).

A proteomikai fejezetben használt mindhárom ML modell neurális háló alapú (McGuffin et al., 2000; Faraggi et al., 2014; Yan & Kurgan, 2017).

4.2. NUMT biológia

4.2.1. A NUMTogenezis folyamata

A NUMTogenezis, azaz a mitokondriális DNS (mtDNS) fragmentek nukleáris genomba történő beépülésének előfeltétele a mitokondriális membrán integritásának megbomlása (1.ábra/1-2. pont), amit mitostresszt indukáló külső faktorok és mutációk egyaránt kiválthatnak (Srinivasainagendra et al., 2017). Mitostresszt indukálhat például az ionizáló sugárzás, endotoxinok, reaktív oxigéngyökök, endonukleázok, hőhatás stb.. Ezek a mitostresszorok egész sejtet érő sugárzás esetén az mtDNS-t nagyobb mértékben károsítják, mint a gDNS-t, feltehetően a sejtmag magasabb hatásfokú hibajavító mechanizmusainak köszönhetően (Gaziev & Shaikhaev, 2007). Az ionizáló sugárzások NUMTogenezist indukáló hatását bizonyították már csirke embriókban és élesztő sejtekben is (Chan et al., 2007; Abdullaev et al., 2013). Egy másik tanulmányban patkányok agyi és máj szöveteiben vizsgálták a NUMTok jelenlétét az idő függvényében (Caro et al., 2010). Eredményeik alapján idősebb szövetekben megnőtt a NUMTok száma. Az idősödő sejtekben megnövekedett NUMT gyakoriságot élesztő esetén is bizonyították (Cheng & Ivessa, 2010). Ennek a jelenségnek a hátterében nagy valószínűséggel a reaktív oxigéngyökök állnak (Singh et al., 2017). Élesztő sejtekben a folyamat genetikai hátterének vizsgálata során két olyan mutációt azonosítottak, amelyek kapcsolatban állnak a NUMTogenezis jelenségével. Az YME1 génben bekövetkező mutáció hatására megnövekedett a mitkondriumból kiáramló genetikai anyag mennyisége. Az YME1 által kódolt ATPáz fehérje a mitokondriális fehérje-homeosztázis fenntartásában játszik fontos szerepet. Az YME1 mutáns élesztő törzsben a citokróm oxidáz II alegységének összeszerelése során is tapasztaltak rendellenességeket (Thorsness et al., 1993). A NUMTogenezissel összefüggő másik mutációt az YME2 génben írták le. Az YME2 gén által kódolt fehérje a mitokondriális belső membrán egyik alkotója. Az YME2 inaktivációja esetén a mitokondriális nukleoidok összeszerelése változott és az mtDNS több nukleázzal szemben is védetté vált (Park et al., 2006).

A mitkondriális membrán sérülése esetén az organellum kap egy degradációs szignált (1.ábra/2-3. pont), ami a mitofágia jelenségét indukálja. A mitofágia az autofágia egy speciális esete, ami a mitofagoszómában (1.ábra/3-4. pont) megy végbe (Goldman et al., 2010). Ez a sejtszervecske felelős a sérült mitokondriumok bontásáért és az alkotók újrahasznosításáért (Hazkani-Covo et al., 2010a). Az említett degradációs szignál a következőképpen alakul ki. A foszfatáz és tenzin homológ indukálta putatív kináz 1 (phosphatase and tensin homologue induced putative kinase-PINK1) fiziológiás körülmények között folyamatosan áramlik a mitokondriumba. Itt a proteáz aktivitással bíró presenilinnel kapcsolt rhomboid-szerű enzim (protease presenilin associated rhomboid like-PARL) hasítja. A PINK1 maradék a hasítás hatására visszakerül a citoplazmába, ahol degradálódik.

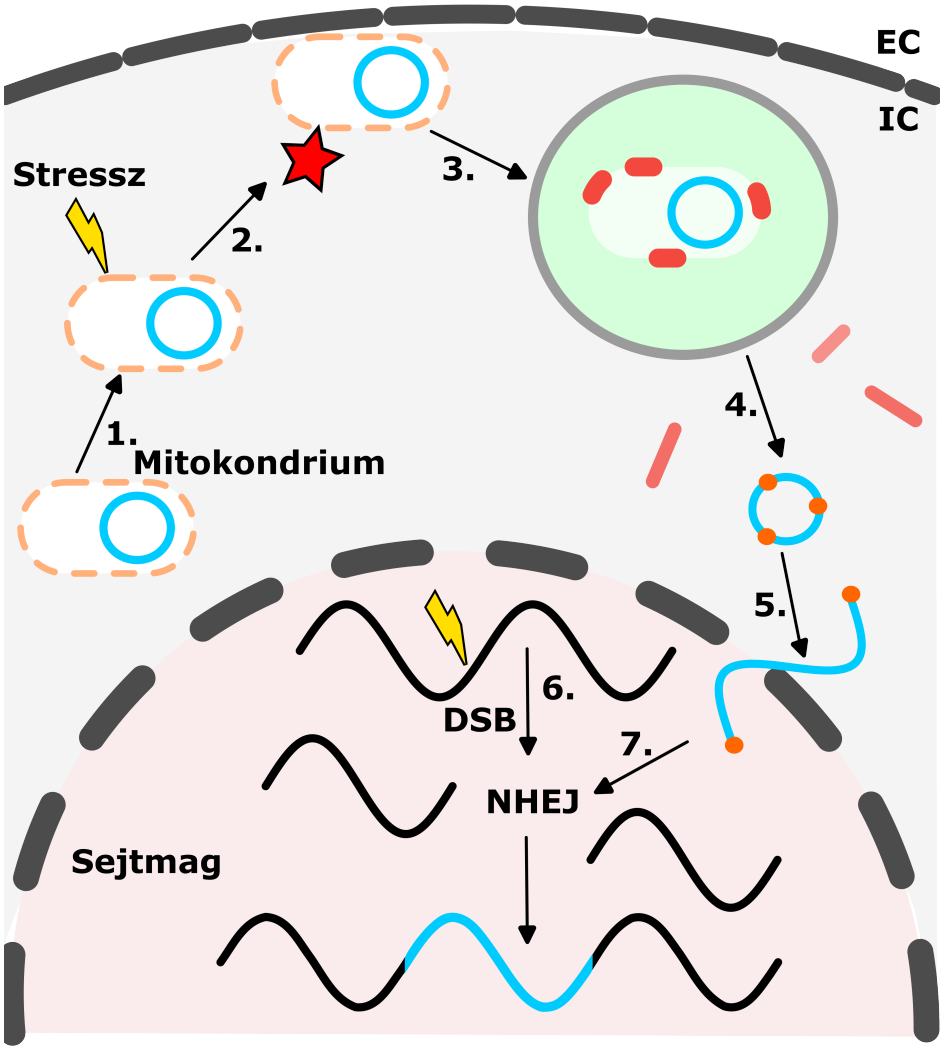
A mitokondriális mebránpotenciál változása esetén a PINK1 PARL általi hasítása megszűnik, ami a külső mitokondriális membránon PINK1 aggregátumok képződéséhez vezet. A stabil PINK1 foszforilálja a Parkin és ubiquitin fehérjéket. A Parkin aktivációja hozzájárul a külső mitokondriális membrán fehérjéinek az ubiquitinációjához. Ezeknek a fehérjéknek az ubiquitinációja a degradációjukat eredményezi, ami beindítja a mitofágiát (G. Chen et al., 2020; Ma et al., 2020).

Nem megfelelő mitofágia esetén a mitokondriális eredetű szekvenciák a ctioplazmába kerülnek (1.ábra/4. pont). Más elméletek alapján ezeknek a szekvenciáknak citoplazmába való kerülése főként elégtelen osztódás és membránfúziós események során következik be (Hazkani-Covo et al., 2010a; Puertas & González-Sánchez, 2020). A NUMT prekurzorok a citoplazmában védettek a nukleázok emésztésével szemben. Ez a védelem kialakulhat egy vakuólum közvetítette úton, vagy a DNS-t kötő hisztonszerű fehérjékkel való komplex képződésével.

A sejtplazmában lévő mitokondriális eredetű szekvenciák membránfúzióval és/vagy pórusokon keresztül (1.ábra/5. pont) jutnak be a sejtmagba (Puertas & González-Sánchez, 2020).

A sejtmagba való bejutást követően a NUMT prekurzorok duplaszálú DNS töréseknél (Double Stranded Break-DSB) épülnek be a sejtmagi genomba a nem homológ végeket összekapcsoló DNS hibajavító mechanizmus (Non-Homologous End Joining-NHEJ) működésének következtében (1.ábra/6-7. pont). A beépülés pillanatától kezdve beszélhetünk NUMTokról. A mitostresszorok is növelik a gDNS DSB frekvenciáját (Gaziev & Shaikhaev, 2007). NHEJ esetén templát DNS hiányában mindig egy nukleáz mediált deléció fog bekövetkezni. Ez sok esetben egy viszonylag hosszú, egyszálú DNS szakaszt eredményez, ami megnöveli a hosszabb deléciók és transzlokációk kockázatát. Bizonyos magyarázatok szerint a sejt a NUMT prekurzorokat templát DNS-ként használva akadályozza meg a nagyobb károsodást jelentő deléciók és transzlokációk bekövetkeztét (Hazkani-Covo et al., 2010b).

A NUMTogenezis egyik lehetséges útját az 1. ábra szemlélteti.



1. ábra. A NUMTogenezis folyamata.
Az extracelluláris teret az EC, míg az intracelluláris teret az IC jelöli. A DSB és NHEJ a duplaszálú DNS törtésre (Double Stranded Break) és az azt javító egyik mechanizmusra, a nem homológ végek összekapcsolására (Non Homologous End Joining) utalnak (6-7. pont). A mitokondriumot mitostressz éri (1-2. pont). A mitokondrium degradációs szignált kap (2-3. pont). A mitokondrium degradálódik (3. pont). Az elégtelenül emésztődött mitokondriális alkotók kijutnak a citoplazmába, ahol a mitokondriális DNS különböző mechanizmusoknak köszönhetően védettséget élvez a nukleázok bontásával szemben (4. pont). A mitokondriális eredetű szekvencia belép a sejtmagba (5. pont). Megtörténik a DSB (6. pont). Az NHEJ felhasználja a mitokondriális eredetű szekvenciát és létrejön a NUMT (7. pont).

4.2.2. A NUMTogenezis evolúciós háttere

A többsejtű élőlények törzsfejlődésének kezdetén egy intracelluláris kapcsolat következett be az alfa proteobaktériumok és az Archeák között (W. F. Martin et al., 2015; Roger et al., 2017). Ez az együttműködés mindkét fél számára hasznosnak bizonyult, így elindulhatott az eukarióták evolúciója. Ezalatt a folyamat alatt komplex feladatok elvégzésére szakosodott sejtszervecskék alakultak ki. Ezek közül az egyik a mitokondrium, ami főként az oxidatív foszforilációért felelős, de ezen kívül részt vesz számos intracelluláris folyamatban (Roger et al., 2017). A mitokondrium egyik különleges fenotípusos jellegzetessége, hogy saját genommal rendelkezik, amit az endoszimbionta elmélet legfontosabb bizonyítékának tekintünk (W. F. Martin et al., 2015). Az eukarióta evolúció során ment végbe az endoszimbiotikus gén transzfer (Endosymbiotic Gene Transfer-EGT), aminek következtében a mitkondrium genomjából nagy mennyiségű genetikai anyag vándorolt át a gazdasejt genomjába (Kelly, 2020). Az endoszimbionta eredetű organellumok genomja általában kevesebb, mint 0.05-öd része az önálló elődjeik genomjának. Ezért a gDNS géntermékeinek egy részét vissza kell irányítani a mitokondriumba (Kelly, 2021). Az EGT molekuláris hajtóerejére a Müller féle teória szolgál magyarázatul (Muller, 1964). Ennek az elméletnek az értelmében egy aszexuális izolációban lévő, tehát a rekombinációt nélkülöző genomban (a NUMTogenezis esetében az mtDNS) deléciók fognak bekövetkezni, amik hozzájárulnak az adott genom genetikai anyagának ritkulásához. Ez rövid távon a genom erózióját eredményezi, míg hosszabb távon a genom eltűnéséhez vezet (Metzger & Eule, 2013; Naito & Pawlowska, 2016). Az mtDNS aszexuálisan izolált, hiszen az eukarióták nagy részében uniparentálisan öröklődik (Breton & Stewart, 2015). Ennek értelmében az mtDNS-re hat a Müller által definiált hatás (Howe & Denver, 2008), így az EGT működése azáltal, hogy mitkondriális eredetű genetikai anyagot irányít a sejtmagba, kimenekíti az mtDNS-t a Müller féle degradáló hatás alól (W. Martin & Herrmann, 1998).

Egy másik feltételezés szerint a sejtenkénti több organellum és az organellumonkénti több genom fenntartása túl nagy energiabefektetést igényel a sejt részéről. Ennek értelmében az organellum genomok EGT útján történő sejtmagi transzfere, majd a géntermékek organellumokba történő visszairányítása kevésbé energiaigényes folyamat, mintha ezt mindegyik organellum saját maga végezné (Kelly, 2021).

4.2.3. NUMTok jeneléte különböző fajokban

Az organellum genomok bizonyos részeinek más genomokba történő integrációját elsőként a kukorica mitkondrium és kloroplasztisz esetén írták le (Stern & Lonsdale, 1982). Állatokban a NUMTok jelenlétét a házi macska genomjában bizonyították elsőként. A macska genomjában meg-

található NUMTok több szempontból is különlegesek. Egyrészt a gDNS tartalmazza az mtDNS közel felét, egy 7.9 kb méretű szekvenciát. Másrészt ez a nagyméretű NUMT több tízszeres kópiaszámban van jelen (Lopez et al., 1994). A nagyobb kópiaszámú, szinte az egész mtDNS-t lefedő NUMTokat (Mega-NUMT) már humán mintákban is azonosították (Lutz-Bonengel et al., 2021). A humán genomban 140 db körüli NUMTot írtak le (Dayama et al., 2014). A NUMTok a háziméh genomjában is jelen vannak, ráadásul számukat a humán genomban fellelhető NUMTok számának közel tízszeresére teszik (Pamilo et al., 2007).

4.3. Proteomika

4.3.1. A fehérjék általános felépítése és a peptidkötés kialakulása

A fehérje molekulák aminósav polimerek. Minden egyes aminósav tartalmaz egy centrális szén atomot, amit α-szén atomnak (C_α) nevezünk. Az C_α atom négy vegyérték elektronját egy hidrogén atom (-H), egy oldallánc (-R), egy karboxil (-COOH) és egy aminó csoport (-NH_2) köti (2. ábra). Az oldalláncok határozzák meg egy aminósav kémiai tulajdonságait (Vasudevan et al., 2019). A fehérjékben az egyes monomereket (aminósav) peptidkötés köti össze, ami két szomszédos aminósav -NH_2 csoportjának nitrogén atomja és -COOH csoportjának szén atomja között keletkezik egy kondenzációs reakcióban (Damodaran, 2008). Az így létrejövő peptidkötésben a karbonil (-C=O) csoport kettős kötése részlegesen delokalizálódik, ami a peptidkötés planaritásához vezet, tehát a kötést alkotó atomok egy síkban helyezkednek el. A planáris peptidkötés a szomszédságában található két C_α atommal bezárt szögek, a Ramachandran torziós szögek alapján jellemezhető. A C_α-N atomok közötti kötés szögét ϕ(fí) torziós szögnek, míg a C_α-C atomok közötti kötés szögét ψ(pszí) torziós szögnek nevezzük (Damodaran, 2008; Vasudevan et al., 2019) (2. ábra). A fehérjék, polipeptidek főláncát, gerincét a peptidkötéssel összekötött savmaradékok oldalláncok nélküli füzére adja.

4.3.2. Szerkezeti szintek

Fehérjemolekulák esetén négy szerkezeti szint különíthető el.

Az elsődleges szerkezet a fehérjemolekulát alkotó aminósav maradékok meghatározott sorrendje (B. Zhang et al., 2018) (2. ábra). Az elsődleges szerkezet és az aminósav szekvencia felcserélhető kifejezések.

Másodlagos szerkezetnek a polipeptid gerinc lokális konformációit tekintjük. A szabályos ismétlődéseket másodlagos szerkezeti elemeknek nevezzük (2. ábra). A másodlagos szerkezeti elemeket főként a főlánc torziós szögei és az oldalláncok határozzák meg (Sun et al., 2004). A DSSP konvenció alapján (8 állapotú rendszer) megkülönböztetünk α-hélixet, 3_{10}-hélixet, π-hélixet, β-strand-et, β-kanyart, β-hidat, β-bend-et és coil-t (Kabsch & Sander, 1983). Az α-hélix esetén egy fordulat 5.41Å, amire 3.6 savmaradék jut, ϕ=-57° és ψ=-47° torziós szögek mellett. Ez a szerkezeti elem nem rendelkezik jól meghatározott hosszúsággal, az α-hélix szakaszok általában 5 és 40 savmaradékot tartalmaznak (Backman, 2019). Ebben a szerkezeti elemben az *i*-ik savmaradék karbonil csoportjának oxigénje lép kapcsolatba hidrogén híd kötésen keresztül az *i*+4-ik savmaradék amid csoportjának nitrogénjével. A 3_{10}-hélix esetén ez a hidrogén híd az *i*-ik savmaradék és az *i*+3-ik oxigén és nitrogén atomjai között jön létre és egy fordulatra 3.2 savmaradék jut (Vieira-Pires

& Morais-Cabral, 2010). A hidat létrehozó oxigén és nitrogén atomok között 10 másik atom helyezkedik el. A π-hélixben ϕ=-57° és ψ=-70°. Ezt a szerkezeti elemet is hidrogén híd stabilizálja, viszont ebben az esetben a szóban forgó kötés az *i*-ik savmaradék és az *i*+5-ik savmaradék között alakul ki (Backman, 2019). A β-lemez elemek a savmaradékok között 3.2-3.4Å-öt tartalmaznak ϕ=-130° és ψ=120° torziós szögekkel. A β-lemez elem önmagában termodinamikailag nem túl kedvező, így 2 vagy több ilyen elem β-redőket alkot. β-kanyaroknál a peptidgerinc közel 180°-os fordulatot vesz. Egy β-kanyart 4 savmaradék határoz meg. Az *i*-ik és az *i*+3-ik savmaradékok C_α atomjai között 7Å a távolság (Chackalamannil et al., 2017) a β-kanyarokban. A rövid, β-redőkre jellemző kötődést mutató fragmenteket β-hidaknak nevezzük (Peter et al., 2019). A β-bend a β-kanyar egyik speciális esete. A coil másodlagos szerkezeti típusba pedig minden az előzőekben nem meghatározott szerkezeti elem tartozik.

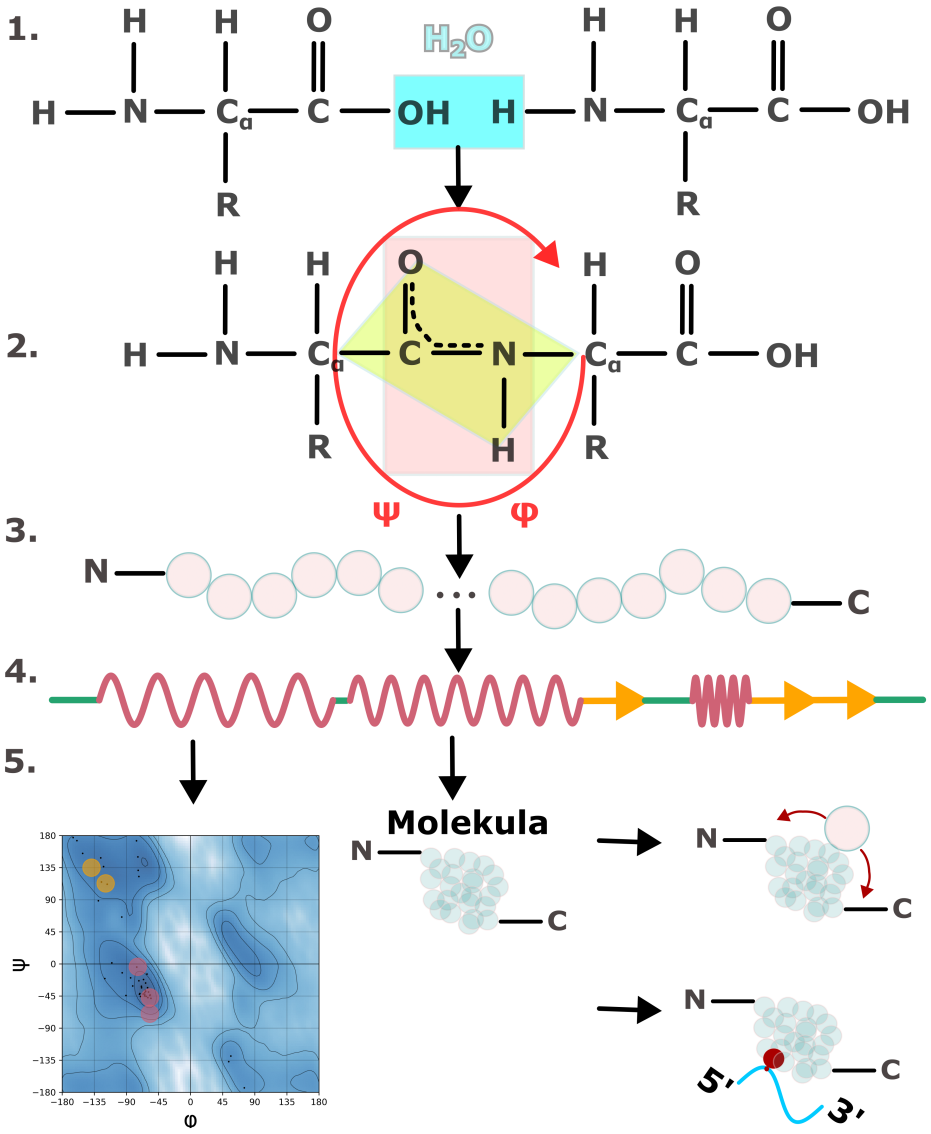
Harmadlagos szerkezeti szintnek a fehérjék atomi szintű, 3D szerkezetét nevezzük (2. ábra). Ezen a szerkezeti szinten sok fehérje rendelkezik doménekkel, amik önálló feltekeredésre képes egységek általában 100-200 savmaradékkal.

Léteznek olyan fehérjék, amelyek több polipeptid lánc alegységet is tartalmaznak. Ezeknek a polipeptid láncoknak a fehérjén belüli elhelyezkedése adja adott molekula negyedleges szerkezetét (Sun et al., 2004).

4.3.3. Az oldószer számára való hozzáférhetőség

A fehérjék funkciójukat a molekulafelszínen keresztül fejtik ki. A molekulaalkotó savmaradékok a harmadlagos térszerkezetben szomszédos savmaradékokkal és/vagy a molekulán kívüli térrel lépnek kölcsönhatásba. A kölcsönhatás minőségét az oldószer számára való hozzáférhetőség mérőszáma írja le (Accessible Surface Area-SA) (Faraggi et al., 2014). Az SA számítása a legtöbb esetben egy oldószer molekula (biológiai rendszerek esetén legtöbbször egy vízmolekula) töltésviszonyait közelítő modellel történik, amivel ”letapogatják” a fehérje atomi felbontású harmadlagos térszerkezetét (2. ábra). Ezt az eljárást ”rolling ball” algoritmusnak hívjuk. Az SA mértékegysége $Å^2$ (Lee & Richards, 1971). Egy savmaradék esetén megadható az abszolút SA és a relatív SA (RSA) is. Az RSA kiszámításához minden egyes aminósav esetén egy tripeptidet (Gly-X-Gly) használnak, amiben az adott aminósav savmaradékát két semleges glicin savmaradék fogja közre. Az így kapott érték, az adott savmaradék maximális SA értéke szolgál az RSA-hoz kellő normalizáció alapjául. Ebből adódóan az RSA értéke nem lehet 1 felett. Az RSA=1 is csak abban az esetben értelmezhető, ha adott aminósav önállóan van jelen az oldószerben. Ez legtöbbször mérési hibára utal (Tien et al., 2013a).

Az SA hagyományos úton történő meghatározásához (”rolling ball”) szükség van adott molekula atomi szintű térszerkezetére.



2. ábra. A peptidkötés kialakulása (1-2. pont) a jellemző torziós szögekkel (2. pont), elsődleges (3. pont) és másodlagos szerkezet elnyerése (4. pont) az inzulin Ramachandran diagramjával, az SA és a nukleinsav kötés meghatározásával (5. pont).
A 4. pont másodlagos szerkezeti elemei közül a sötét rózsaszín a hélixre, a sárga a β-redőre és a zöld a coil-ra utal.

4.3.4. Nukleinsavak kötése

A fehérjék és nukleinsavak interakciói a centrális dogma minden egyes lépését befolyásolják. Ennek megfelelően a modern hatóanyag-fejlesztés egyik leginkább kutatott területe a fehérje nukleinsav komplexek gyógyszermolekulákkal történő célzott módosítása. Ezeknek a komplexeknek az azonosítása és módosítása komoly kihívást jelentő feladatok még a legelterjedtebb komplexek (transzkripciós faktorok és promóterek) esetén is (Radaeva et al., 2021). A fehérje nukleinsav komplexek azonosításához a komplexalkotók együttes szerkezetének meghatározása szükséges, ami nagyon időigényes feladat (R. Wang et al., 2011; Su et al., 2019).

5. Anyag és módszer

A statisztikai analízist a Python programnyelv Scipy (verziószám: 1.6.2) és Numpy (verziószám: 1.20.3) könyvtáraiban, míg az ML modellek implementációját a scikit-learn (verziószám: 0.24.2) és umap (verziószám: 0.5.3) könyvtárakban végeztük (Virtanen et al., 2021; Harris et al., 2020; Pedregosa et al., 2011; McInnes et al., 2018). A felügyelet nélküli tanulást megvalósító tSNE modell zavarosság (PP) és tanulási ráta (LR) hiperparamétereit az ún. grid-search eljárásban optimalizáltuk. Az optimalizáció során a maximális PP minden esetben a bementi adathalmaz méretének 1/10-e, míg az LR az 1/12-e volt. UMAP paraméterek (n neighbors és minimális távolság) esetén is hasonlóképpen jártunk el az optimalizáció során.

Az ábrák Matplotlib (verziószám: 3.4.3), Seaborn (verziószám: 0.11.2) és InkScape (verziószám: 1.2.0) programokban készültek (Hunter, 2007; Waskom, 2021; Bah, 2009).

5.1. NUMT biológia

A nyúl genomot (OryCun2.0) kromoszómális és mitokondrium szekvenciáját az Ensembl genom adatbázisból szereztük be (http://ftp.ensembl.org/pub/release-104/fasta/oryctolagus_cuniculus/dna/). Irodalmi adatok alapján meghatároztunk egy e-érték illesztési küszöböt (10^{-4}) (Tsuji et al., 2012; Schiavo et al., 2017). Ennek az értéknek a kiszámításához a fizikailag megfordított sorrendű (nem komplmenter) nyúl mtDNS-t illesztettük a genomra. Az illesztéshez a LASTAL szoftvert (verziószám: 1219) használtuk a következő beállításokkal: match=1, mismatch=-1, gap open penalty=7, gap extension penalty=1 (Kiełbasa et al., 2011). Az így kapott illesztések közül a legalacsonyabb e-értéket tekintettük küszöbértéknek. A későbbi illesztések közül csak azokat vontuk be a további analízisbe, ami ettől kisebb értékkel bírt. A felvázolt módszer biológiai alapja, hogy az mtDNS reverziója egy biológiailag értelmetlen, véletlenszerű szekvenciát eredményez. A véletlenszerű szekvencia és a gDNS illesztésének legkisebb e-értékű találata a véletlennek köszönhető, így az ettől szignifikánsabb találatról nagy biztonsággal megállapítható a mitokondriális eredet azaz, hogy NUMT.

A korábban már publikált eredményeknek megfelelően, dupla mtDNS-t illesztettünk a gDNS-re (Tsuji et al., 2012; Schiavo et al., 2017). Ennek a módszernek a használatával lehetőségünk nyílik azoknak a NUMToknak a deteklására is, amelyek magukban foglalják az mtDNS linearizációs pontját. Az így kapott adathalmazt szűrtük az előzőekben már bemutatott e-értékre vonatkozó illesztési küszöbértéknek megfelelően.

Az mtDNS annotációját a University of Leipzig MITOS szerverének segítségével kaptuk Genetic code: 02 és Vertebrate beállítások használata mellett (Bernt et al., 2013).

A gDNS GC arányának és ismétlődő motívumok meglétének vizsgálatához minden NUMTot tartalmazó kromoszómát megmintáztunk véletlenszerűen az adott kromoszómán lévő NUMTok számának és méretének megfelelően. Az ismétlődő elemek jelenlétét a NUMTok 5000 bázispár (5 kb) határoló régióiban RepeatMasker programmal vizsgáltuk, amit szerveren (https://genome.ucsc.edu/cgi-bin/hgTables) és lokálisan (verziószám: 4.1.2-p1) futtattunk.

A fals pozitív arány (false positive rate - FPR) csökkentése érdekében megvizsgáltuk, hogy a szkaffoldokról származó NUMTok egyeznek-e a kromoszómákról származó NUMTokkal. Ehhez a szkaffoldokat és a kromoszómákat illesztettük egymáshoz a fenti pontozási sémát (1,-1,7,1) használva. Ennek az illesztésnek az eredményéből kizárólag azokat a találatokat vizsgáltuk, amiket előzőleg NUMTként definiáltunk. Ezekben az esetekben a NUMTok kromoszómális és szkaffold határoló régióinak a hasonlóságát és ismétlődő elem arányát vetettük össze. A hasonlóságot BioPythonban (verziószám: 1.78) kivitelezett páros illesztésekkel vizsgáltuk szintén a fenti séma használatával (Cock et al., 2009).

A határoló régiókat a SAMTOOLS program (verziószám: 1.6) segítségével nyertük ki a gDNS fájlból (Li et al., 2009).

A normalitást Anderson Darling teszttel vizsgáltuk 0.05 p érték mellett (Egyenlet. 5.1). Ez a teszt egy ismert eloszláshoz hasonlítja a minta eloszlását a két empirikus eloszlásfüggvény görbe alatti területeinek összevetésével. Abban az esetben, ha a görbe alatti területek különbsége átlép egy küszöbértéket, a kísérleti eloszlás nem tekinthető Gauss féle eloszlásnak.

$$Anderson - Darling = n \int_{-\infty}^{\infty} (F_n(x))^2 w(x) dF(x) \qquad \text{(Egyenlet. 5.1)}$$

ahol F=ismert eloszlásfüggvény (pl.: normál, binomiális stb),
Fn=minta eloszlásfüggvénye,
n=mintaelemszám,
$w(x)$=súlyfüggvény

A normalitás vizsgálat eredményének függvényében t-próbát vagy Wilcoxon féle próbát használtunk. A kromoszómaméret és az adott kromoszómán lévő NUMTok száma, hossza közötti összefüggéseket Pearson (PCC) (Egyenlet. 5.2) és Spearman féle rang- korrelációs (SCC) (Egyenlet. 5.3) együttható számításával elemeztük.

$$PCC = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \qquad \text{(Egyenlet. 5.2)}$$

ahol x és y a korreláltatni kívánt változók,
míg $\bar{x}$ és $\bar{y}$ a változók átlagai.

$$SCC = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{3} \qquad \text{(Egyenlet. 5.3)}$$
$$d_i = x_i - y_i$$

ahol N=mintaelemszám.

A nem normál eloszlást mutató adathalmazok esetén azért volt szükség az SCC alkalmazására, mert sokkal kevésbé érzékeny a kiugró értékekre, mint a PCC.

Adott genomrészlet és az annak megfelelő NUMTok közötti összefüggés során a pontokra illesztett egyenes egyenletét a legkisebb négyzetek módszerével számítottuk (Egyenlet. 5.4).

$$y = mx + b$$
$$m = \frac{\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \qquad \text{(Egyenlet. 5.4)}$$
$$b = \frac{\sum y - m \sum x}{n}$$

ahol n=mintaelemszám,
b=az egyenes y-tengely metszéspontja,
míg m=az egyenes meredeksége.

A NUMTok és a véletlenszerű szekvenciák klasszifikációjára RBF-Kernel SVM-et tanítottunk. Ez a módszer felügyelt tanulást hajt végre, ebben az esetben kategórikus kimenettel (klasszifikáció). Az SVM MI lényege, hogy meghatározott adathalmazokat próbál egymástól a lehető legnagyobb mértékben szeparálni. Abban az esetben, ha adott adathalmazok lineárisan nem szeparálhatók, egy transzformációs függvény (Kernel) az eredetileg n dimenziós problémát egy $n+1$ dimenziós problémává alakítja. A transzformáció következtében az $n + 1$ dimenziójú térben lévő adathalmazok lineárisan szeparálhatóvá válnak. A lineáris szeparáció egy bináris probléma esetén a következőkben definiált függvény használatával valósítható meg (Egyenlet. 5.5).

$$f(x, w, b) = sign(\begin{bmatrix} x_1, & x_2, & x_3 \dots x_n \end{bmatrix} \cdot \begin{bmatrix} w_1, & w_2, & w_3 \dots w_n \end{bmatrix} b) \qquad \text{(Egyenlet. 5.5)}$$

ahol $sign()$=az előjelfüggvény, ami az előjelnek megfelelően kétféle kimenettel (1;-1) rendelkezik,
x=bemenetvektor,
w=súlyvektor,
míg b=az eltolás értéke.

Az SVM modell teljesítményét k-szoros keresztvalidációs (k=3) eljárással és a tévesztési mátrixból származtatott mutatókkal is teszteltük (Confusion Matrix - CM) (Egyenlet. 5.6). Az adatszivárgás elkerülése érdekében a keresztvalidációs eljárás minden egyes iterációjában külön normalizáltuk a bementeket a minimum-maximum normalizációs eljárásnak megfelelően (Egyenlet. 5.7).

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \qquad \text{(Egyenlet. 5.6)}$$

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \text{(Egyenlet. 5.7)}$$

A NUMTok és a megfelelő mtDNS szekvenciák genetikai távolságát a hiányzó nukleotidokat is elfogadó módosított Kimura2 paraméterrel becsültük (Nishimaki & Sato, 2019).

$$K = \frac{3}{4} w \log w - \frac{w}{2} \log(S - P) \sqrt{S + P - Q} \qquad \text{(Egyenlet. 5.8)}$$

ahol w=annak a valószínűsége, hogy adott pozícióban nukleotid van,
S=n_1/n,
n_1=azoknak a pozícióknak a száma, ahol a két összehasonlított szekvencia azonos nukleotidot tartalmaz
n=az összes nukleotid száma,
P=n_2/n,
n_2=a tranzícióra (purinból purin vagy pirmidinből pirimidin) visszavezethető mutációk száma,
n_3=a transzverzióra (purinból pirimidin vagy $vicaversa$) visszavezethető mutációk száma.

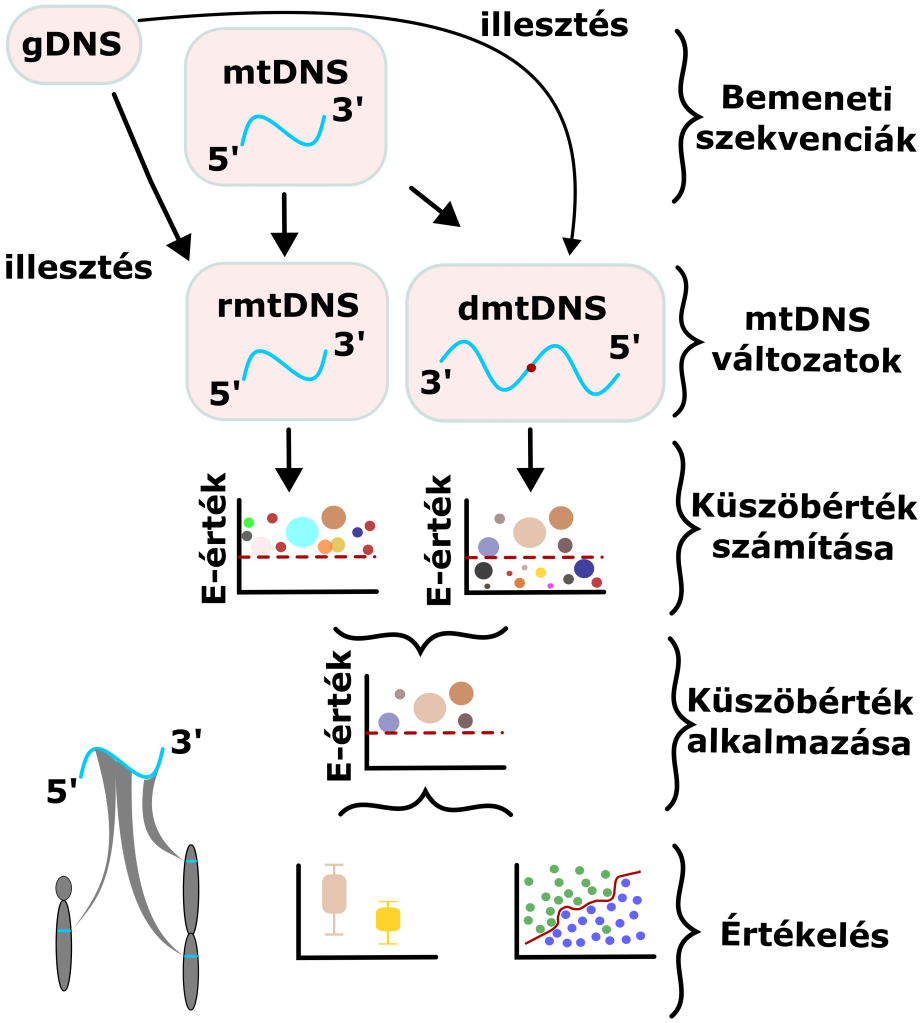
Az elérhető emlős genomokat átfogó vizsgálatokhoz a nukleáris (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/) és a mitokondriális (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/mitochondrion/) genom szekvenciákat is az NCBI adatbázisból szereztük be. A nukleáris genomoknál minden esetben a legfrissebb verziójú összeszerelést használtuk. A taxonómiai adatok integrációja során az azonosítókat (https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/eukaryotes.txt) és a rangokat (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/new_taxdump.zip / rankedlineage.dmp) is az NCBI adatbázisból nyertük ki.

Az emlős genomokban kimutatott NUMTok relatív méretének és pozíciójának eloszlásait Anderson Darling (Egyenlet. 5.1) és χ^2 tesztekkel vizsgáltuk.

Az 5kb-os határoló régiókat a SAMTOOLS program (verziószám: 1.6) segítségével nyertük ki a gDNS fájlokból (Li et al., 2009). Ezeket a határoló régiókat a RepeatMasker programmal (verziószám: 4.1.2-p1) vizsgáltuk fajspecifikus beállítások mellett.

A filogenetikai vizsgálatokat az R programnyelv ape csomagjával (verziószám: 5.6-2) végeztük (Paradis & Schliep, 2019), míg a filogenetikai fát a ggtree csomaggal (verziószám: 3.2.1) vizualizáltuk (Yu, 2020).

A NUMT biológiával kapcsolatos vizsgálatok folyamatát a 3. ábra mutatja.



3. ábra. A NUMT biológiával kapcsolatos vizsgálatok folyamata.

5.2. Proteomika

A komplett humán proteómot a UniProt adatbázisból nyertük ki ("UniProt: the universal protein knowledgebase in 2021", 2021). Azért esett a választásunk erre a proteómra, mert a komplementaritás vizsgálatához elengedhetetlen kísérleti adatokból e faj esetén találhatjuk meg a legtöbbet. A "Fragment" címkével ellátott fehérjék eltávolításával 43 789 szekvenciát kaptunk. A Protein Data Bank (PDB) vonatkozó térszerkezeteit a SIFTS program segítségével térképeztük a meglévő UniProt szekvenciákra (Berman et al., 2000; Dana et al., 2019). Ezeket a térszerkezeti fájlokat kinyertük a PDB adatbázisból. A PDB fájlokban rejlő információk szolgáltak ún. valós osztálycímkeként, referenciaadatként a másodlagos térszerkezet és az SA vizsgálata során.

A 30 aminósavnál kisebb peptideket kizártuk a további vizsgálatokból. Abban az esetben, ha egy UniProt szekvenciát több PDB lánc is jellemzett, kizárólag a leghosszabb szerkezetet tartottuk meg. Ha több szerkezet fedte egy UniProt szekvencia ugyanazon részét, akkor a legnagyobb felbontással rendelkező szerkezetet vontuk be a további kutatásokba. Ezeknek a lépéseknek az eredményeként 5 133 UniProt szekvenciát és az ezeknek megfelelő 6 417 PDB szerkezetet vizsgáltuk.

Az SA és a másodlagos szerkezetre vonatkozó adatokat közvetlenül a PDB állományokból nyertük a DSSP algoritmust használva (Kabsch & Sander, 1983). A DSSP a másodlagos szerkezetre vonatkozóan egy 8 állapotú osztályozást használ. Annak érdekében, hogy ez a komplementaritás vizsgálata során összehasonlítható legyen az általunk választott prediktor kimenetével (McGuffin et al., 2000), átalakítottuk a 8 állású osztályozást 3 állásúvá. Ennek megfelelően a 3_{10}- és az α-hélixeket az általános hélix (H), a β-redőt és -hidat az általános β-redő (E) és az előző két osztályba nem sorolható szerkezeteket a coil (C) kategóriába soroltuk. Az abszolút SA értékeket irodalmi maximum SA értékeknek megfelelően normalizáltuk megkapva így a relatív SA (RSA) értékeket (Tien et al., 2013b). Az SA relativizáció ebben az esetben is azt a célt szolgálta, hogy a referenciaadat és a vonatkozó prediktor kimenete összevethető legyen (Faraggi et al., 2014).

A nukleinsav kötés referenciaadatát a BioLip adatbázis szolgáltatta (J. Yang et al., 2012). A BioLip annotációt feltérképeztük a kiválasztott UniProt szekvenciákra. Ez a térképezés 175 fehérjében 3 557 DNS kötő és 106 fehérjében 2 368 RNS kötő aminósavat eredményezett.

A normalitást ebben az esetben is Anderson Darling teszttel (Egyenlet. 5.1) vizsgáltuk 0.05 p érték mellett. A normalitás vizsgálat eredményének függvényében t-próbát vagy Wilcoxon féle próbát használtunk. Többszöri összehasonlítás esetén Bonferroni módszerrel korrigáltuk a p értékeket. Az SA referencia és predikció közötti kapcsolatot PCC (Egyenlet. 5.2), SCC (Egyenlet. 5.3) és átlagos abszolút hiba (Mean Absolute Error - MAE) (Egyenlet. 5.9) számításával elemeztük. A

másodlagos szerkezet predikciójának pontosságát a Q3 érték számításával vizsgáltuk, ami a helyesen prediktált savmaradékok számának az arányítja az összes prediktált savmaradék számához.

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{n}$$
$$|e_i| = |x_i - y_i| \qquad \text{(Egyenlet. 5.9)}$$

ahol n=mintaelemszám.

A nukleinsav kötés előrejelzésének pontosságát az ún. karakterisztika görbe (Receiver Operating Characteristic curve - ROC) ábrázolásával vizsgáltuk. Ehhez meghatároztuk a valódi pozitív (TP), valódi negatív (TN), fals pozitív (FP) és fals negatív (FN) értékeket. TP az eset, amit a modell helyesen nukleinsav kötőnek, TN pedig, amit helyesen nem kötőnek jelzett. FP az eset, amit a modell helytelenül nukleinsav kötőnek, FN az eset, amit pedig helytelenül nem kötőnek jelzett. Következő lépésként ezekből a valódi pozitív (TPR) (Egyenlet. 5.10) és a fals pozitív (FPR) (Egyenlet. 5.11) arányokat számítottuk ki.

$$TPR = \frac{TP}{TP + FN} \qquad \text{(Egyenlet. 5.10)}$$

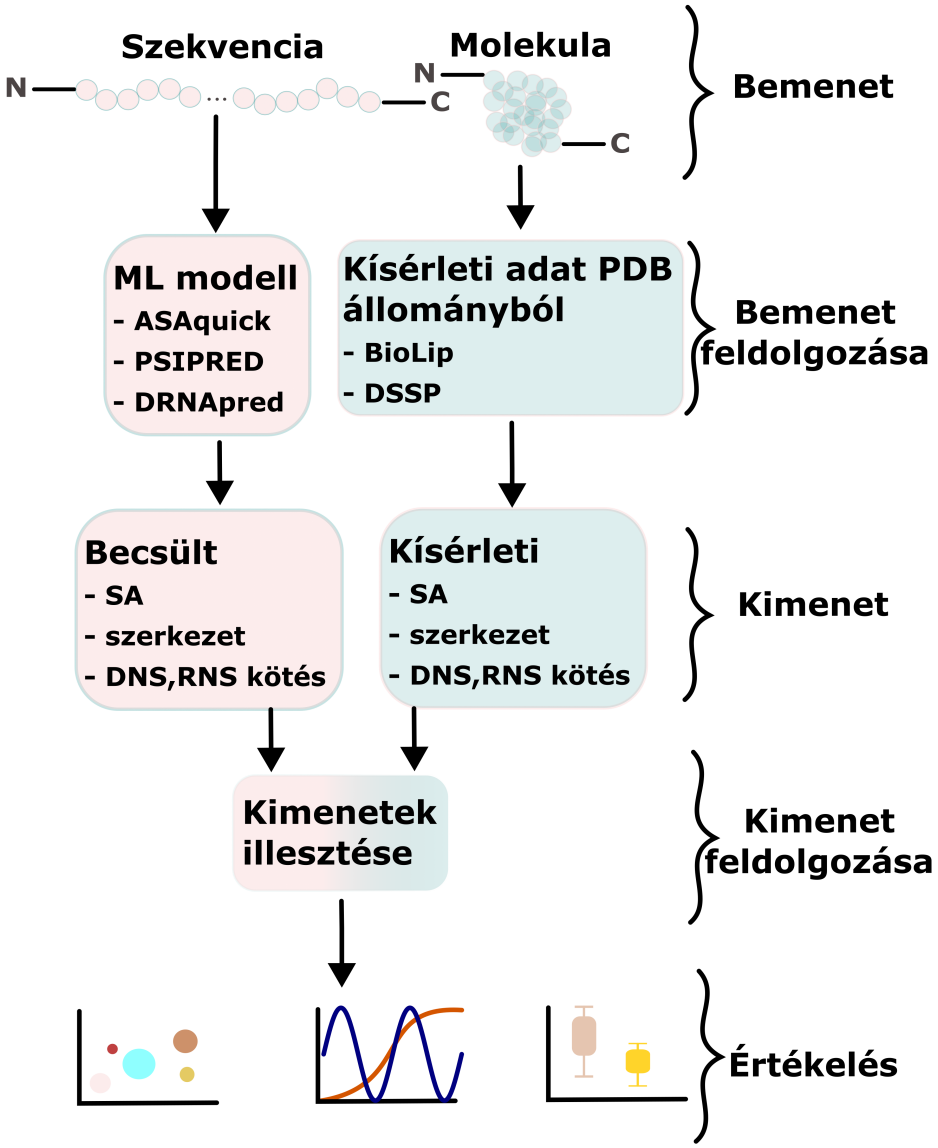
$$FPR = \frac{TN}{TN + FP} \qquad \text{(Egyenlet. 5.11)}$$

A másodlagos szerkezet predikciójának pontosságát a valós és predikált szerkezeti elemeket tartalmazó szegmensek összevetésével vizsgáltuk. Ezt az összefüggő szegmensek értékének mérőszáma (Segment OVerlap score - SOV) tette lehetővé (Egyenlet. 5.12).

$$
\begin{aligned}
S_{(i)} &= \{(s_1, s_2) | s_1 \cap s_2 \neq \emptyset\} \\
S_{(i)}^{,} &= \{(s_1, s_2) | s_1 \forall s_2 = \emptyset\} \\
SOV_{(i)} &= 100[\frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{S_{(i)}} \frac{minov(s_1, s_2) + \sigma(s_1, s_2)}{maxov(s_1, s_2)} len(s_1)] \\
N &= \sum_{S(i)} len(s_1) + \sum_{S^{,}(i)} len(s_1) \\
\sigma(s_1, s_2) &= min \begin{cases} maxov(s_1, s_2) - minov(s_1, s_2) \\ minov(s_1, s_2) \\ int(len(\frac{s_1}{2})) \\ int(len(\frac{s_2}{2})) \end{cases}
\end{aligned}
\qquad \text{(Egyenlet. 5.12)}
$$

ahol s_1 és s_2=egy i konformációs állapotban $\{H, E, C\}$ lévő megfigyelt és prediktált szegmenspár,
$S_{(i)}$=az i konformációs állapotban lévő átfedő szegmenspárok (s_1, s_2) halmaza,
$S_{(i)}^{,}$=s_1 szegmensek halmaza, amelyek nem rendelkeznek átfedő párral az s_2-ből,
$minov(s_1, s_2)$=egy i konformációs állapotban $\{H, E, C\}$ lévő megfigyelt és prediktált szegmenspár közös elemeinek száma,
míg $maxov(s_1, s_2)$=egy i konformációs állapotban $\{H, E, C\}$ lévő megfigyelt és prediktált szegmenspár összes elemeinek száma.

A proteomikával kapcsolatos vizsgálatok folyamatát a 4. ábra mutatja.



4. ábra. A proteomikával kapcsolatos vizsgálatok folyamata.

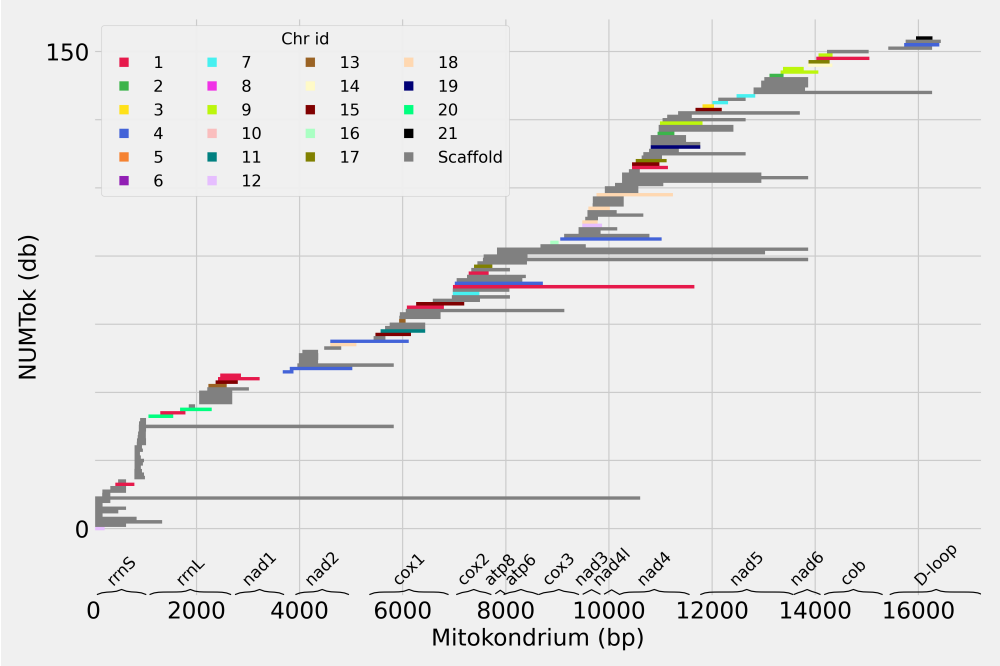
6. Eredmények

6.1. NUMT biológia

6.1.1. A nyúl genom NUMTjainak jellegzetességei

A nyúl genomjában 153 NUMTot térképeztünk fel, amik a D-loop kis részétől eltekintve az egész mtDNS-t többszörösen is lefedték (5. ábra). Az azonosított NUMTok nagy része, mintegy 100 db, szkaffoldokról származott.

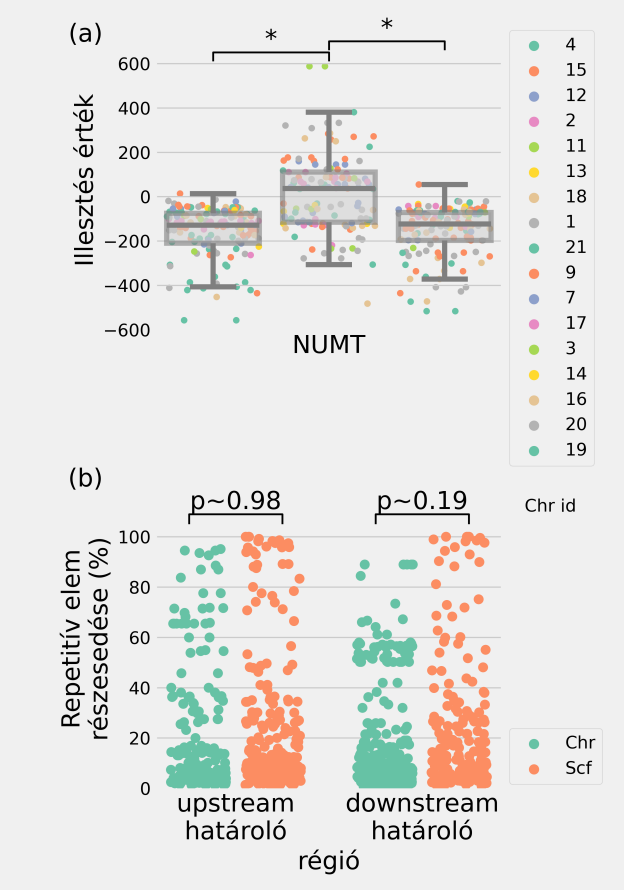
A NUMTok méretbeli eloszlása nagy heterogenitást mutat, de legnagyobb részük (130 db) 2000 bp alatti méretű. A 2000 bp feletti méretű NUMTok egy kivételtől eltekintve szkaffoldokon helyezkedtek el. Vizsgálataink alapján az 5,6,8,10 és 21-es kromoszómák nem tartalmaztak NUMTokat.



5. ábra. A nyúl genomban azonosított NUMTok genomon belüli elhelyezkedése és mtDNS-beli eredete.
Az x tengelyen lévő mitokondriális annotáció a 100 bp-nél kisebb rRNS géneket nem tartalmazza. A Chr id az adott genomi részlet azonosítójára utal.

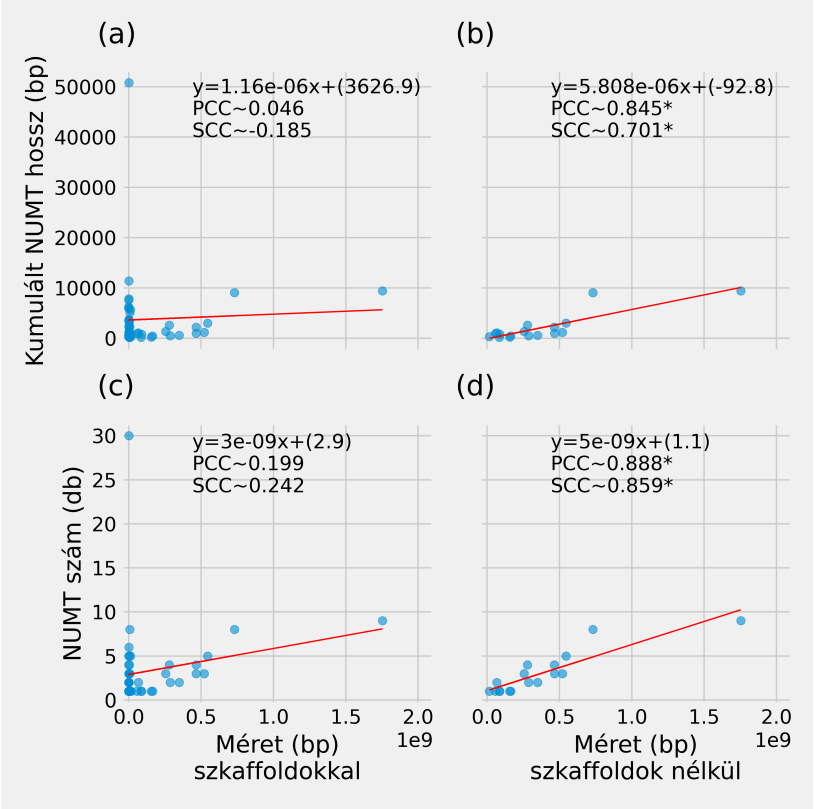
Annak érdekében, hogy megbizonyosodjunk arról, hogy a szkaffoldokon észlelt NUMTok nem szekvenálási műtermékek, megvizsgáltuk a kromoszómákon és a szkaffoldokon elhelyezkedő NUMTokat (6. ábra). A kromoszómákon és a szkaffoldokon elhelyezkedő NUMTok illesztése során számos esetben magas fokú homológiát találtunk, ugyanakkor a NUMTok határoló régiói minden esetben eltértek egymástól (6.ábra/a). Egy genom szekvenálásakor az ismétlődő elemek magas arányú jelenléte pontatlanná teszi az összeszerelést, szkaffoldokat eredményezve. Ez a lehetőség a nyúl genom esetén nem valószínű, hiszen a kromoszómális és szkaffoldon lokalizált NUMTok

határoló régióit megvizsgálva nincs statisztikailag igazolható különbség (p>0.05) az ismétlődő elemek tekintetében (6.ábra/b).



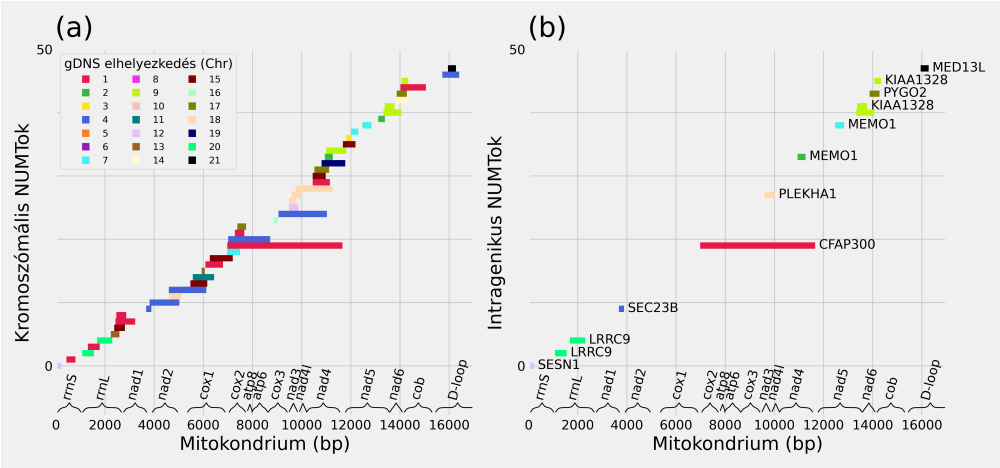
6. ábra. A nyúl genom szkaffoldjainak vizsgálata. Az (a) rész a a szkaffold és kromoszóma lokalizált NUMTok és határoló régióik illesztési értékeit, míg a (b) rész a határoló régiók repetitív elem részesedését mutatja. Mindkét esetben az adott NUMTnak megfelelő hosszúságú határoló régiót vizsgáltunk. A szignifikáns eredményeket * jelöli.

Adott genomi részlet (kromoszóma vagy szkaffold) hossza és az azon elhelyezkedő NUMTok kumulált hossza között nagyon gyenge (praktikusan nem létező) összefüggést találtunk (PCC:0.046; SCC:-0.185). Ez az összefüggés lineárisan jól jellemezhetővé erősödik (PCC:0.845; SCC:.0701), ha a vizsgált adathalmazból eltávolítjuk a szkaffoldokat (7. ábra/a-b). Az adott genomi részlet hossza és az azon elhelyezkedő NUMTok száma közötti összefüggés is hasonlóan írható le, tehát a kromoszómák és a szkaffoldok együttes vizsgálata esetén gyenge összefüggést kaptunk (PCC:0.199; SCC:0.242), ami közel egyenes arányúvá válik a szkaffoldok vizsgálatból történő eltávolításával (7. ábra/c-d).



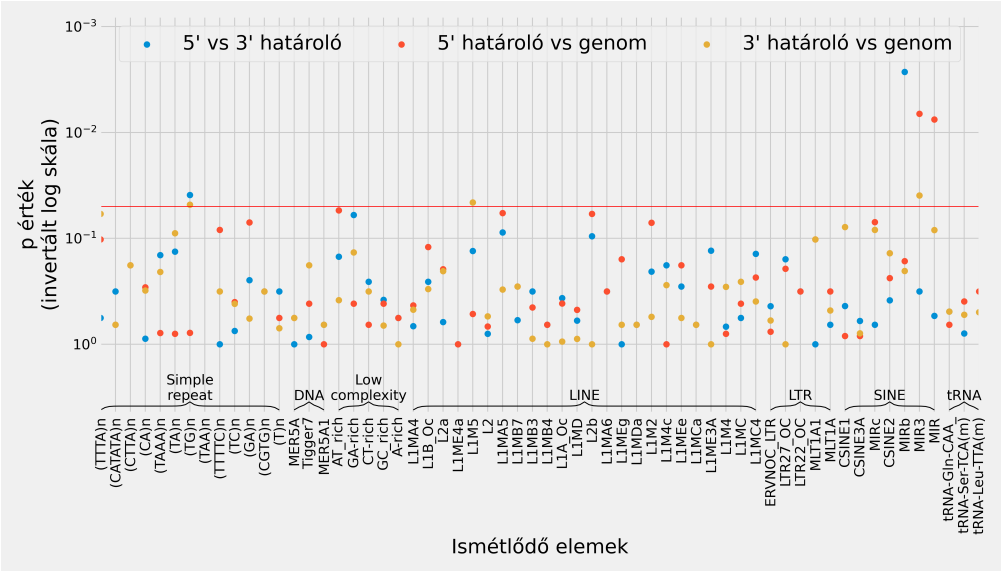
7. ábra. Adott kromoszóma (b,d) és szkaffold (a,c) méretének és NUMTjainak kumulált hossza (a,b) illetve száma (c,d) közötti összefüggés.
A p<0.001 eredményeket * jelöli.

A kromoszómális NUMTok (8.ábra/a) közel 1/5-e (9 db) az 1-es kromoszómába integrálódott. A kromoszómába integrálódott NUMTok közül a leghosszabb egy majd 5 kb méretű az 1-es kromoszóma *CFAP300* génjébe épült be. Az 50 db körüli kromoszómális NUMTból 12 intragenikus NUMTot mutattunk ki. A NUMTokat tartalmazó gének a következők: *SESN1*] (Sestrin 1), *LRRC9* (Leucine Rich Repeat Containing 9), *SEC23B*, *CFAP300* (Cilia And Flagella Associated Protein 300), *PLEKHA1* (Pleckstrin Homology Domain Containing A1), *MEMO1* (Mediator Of Cell Motility 1), *KIAA1328*, *PYGO2* (Pygopus Family PHD Finger 2), *MED13L* (Mediator Complex Subunit 13L) (8.ábra/b).



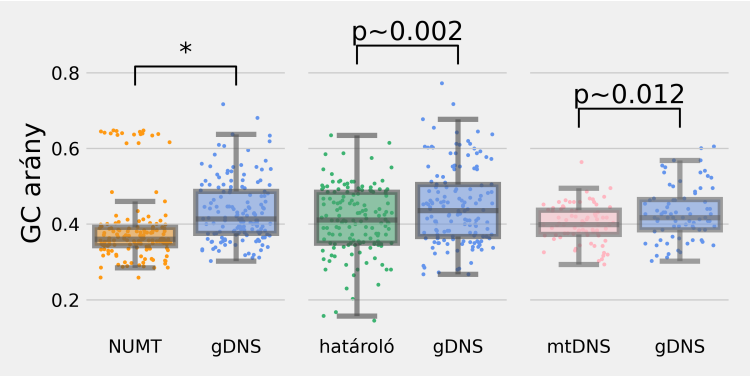
8. ábra. A nyúl genomban azonosított kromoszómális (a) és intragenikus (b) NUMTok. Az x tengelyen lévő mitokondriális annotáció a 100 bp-nél kisebb rRNS géneket nem tartalmazza.

A NUMTokban és a határoló szekvenciáikban 7 közösen megjelenő RepeatMasker ismétlődő elem osztályt, a DNA, SINE, LINE, Low complexity, LTR, Simple repeat és tRNA-t találtunk. Ezekből a csoportokból a Short Interspersed Elements (SINE), Long Interspersed Nuclear Elements (LINE) és Simple repeat tartalmazott olyan ismétlődő elemet, ami szignifikáns különbséget mutatott a vizsgált csoportok között frekvenciáját tekintve (9. ábra). A SINE osztályból a MIRb 5' és 3' határoló régiók frekvenciái különböztek egymástól. Az 5' határoló régiók és a genomi minták összevetésekor a Mammalian-wide Interspersed Repeats (MIR) és MIR3 ismétlődő elemek frekvenciái mutattak különbséget ráadásul a MIR3 a 3' határoló régió genommal való összevetésekor is szignifikáns eredményt produkált. A LINE osztályból a L1M5 elem mutatott frekvenciabeli eltérést, amikor a 3' határoló régiót vetettük össze a genomi mintákkal. A Simple repeat osztályból a (TG)n oligomerek frekvenciája tért el mindkét határoló régióban, mikor a genomi mintákhoz hasonlítottuk őket (9. ábra).



9. ábra. Az ismétlődő elemek frekvenciájának alakulása a NUMTok környezetében. A piros vonal feletti eredmények szignifikánsak (p<0.05).

A GC arány vizsgálata során azt találtuk, hogy a NUMTok és az mtDNS is alacsonyabb GC tartalommal bírnak, mint a gDNS. Ugyanakkor a NUMTokat határoló gDNS régiók is különböztek a genom véletlenszerű pontjaiból származó gDNS mintáktól (alacsonyabb értékkel bírtak) a GC tartalom tekintetében (10.ábra).

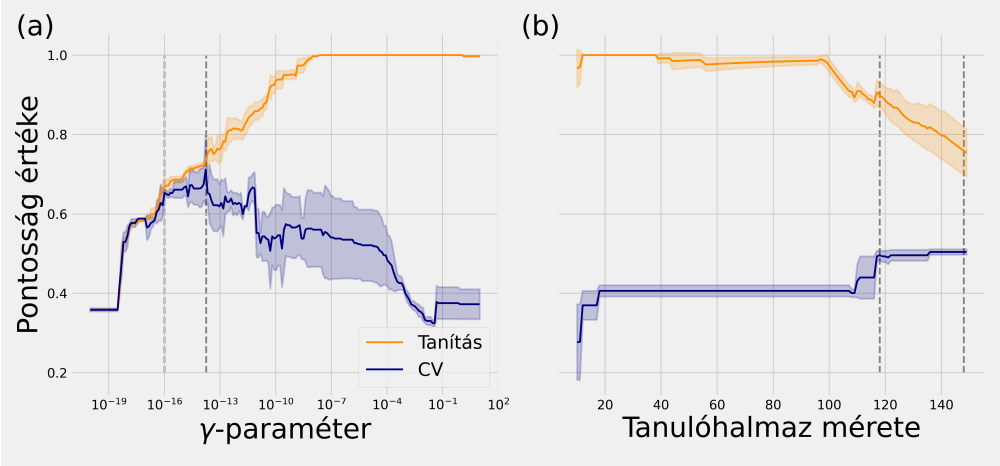


10. ábra. A GC arány összevetése. A szignifikáns ($p<10^{-5}$) eredményt * jelöli. A $p>10^{-5}$ eredmények esetén a 3 tizedesre kerekített p értékeket tüntettük fel.

A véletlenszerű szekvenciák NUMTként történő azonosításának kiszűrése érdekében RBF-kernel SVM-et tanítottunk a NUMTok és véletlenszerű szekvenciák klasszifikációjára. Ez az SVM egy szekvencia gDNS részlet pozíciójából, hosszából, a szekvencia és a környezete GC arányából megbízhatóan osztályozta a bemeneteket (k-szoros CV (k=3) esetén 0.7 körüli maximális pontosság) (11.ábra). A modellünk mindegyik komplexitás érték esetén jobban teljesített, mint az úgynevezett ”dummy classifier”, ami minden esetben a leggyakoribb osztályt jelzi előre. Abban az esetben, ha a predikcióhoz az előzőleg említett minden ismérvet felhasználtunk, az SVM-ünk ké-

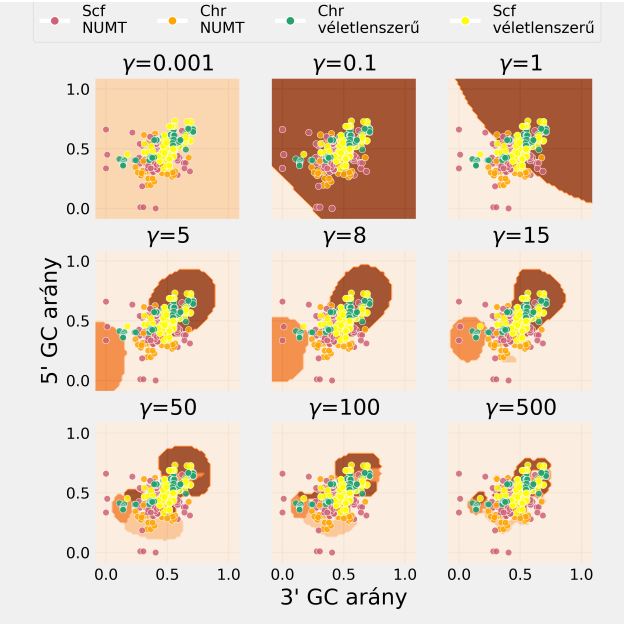
pes volt eldönteni, hogy az adott szekvencia NUMT-e és azt is, hogy szkaffoldról vagy éppen kromoszómáról származik-e (Egyenlet. 6.1). Az SVM modell az optimális predikciós pontosságot abban az esetben érte el, ha az egyes tanítási pontok hatása, a modell komplexitása a 10^{-16}-10^{-14} közötti γ-értéket vett fel (11.ábra/a). A rendelkezésre álló adatok az SVM tanításához elegendőnek bizonyultak (11.ábra/b).

$$CM = \begin{bmatrix} 21 & 0 & 0 & 7 \\ 0 & 13 & 2 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 26 \end{bmatrix} \qquad \text{(Egyenlet. 6.1)}$$



11. ábra. NUMTok és véletlenszerű szekvenciák klasszifikációjára tanított RBF-kernel SVM validációs (a) és tanulási (b) görbéje.

SVM-mel a NUMTok és véletlenszerű szekvenciák klasszifikációja abban az esetben is sikerült, ha kizárólag a határoló régiók GC arányán tanítottuk a mesterséges intelligenciát. Ugyanakkor a döntési határok alakulását szemlélve, ebben az esetben a predikció pontossága nem volt összevethető a több bemenetre alapozott tanulás pontosságával (12. ábra).



12. ábra. NUMTok és véletlenszerű szekvenciák klasszifikációjára tanított RBF-kernel SVM döntési határainak alakulása különböző γ-paraméterek mellett.
A Chr és az Scf a kromoszómára és a szkaffoldra utalnak.

6.1.2. Egységes keretrendszer NUMTok bányászatára

Az általunk vizsgált emlős NUMTok jellegzetességei alapján jól elkülöníthetők a filogenetikai rendek (Mellékletek/táblázat). UMAP dimenzió csökkentés használata esetén a rendekre jellemző centrumok is kirajzolódtak (13. ábra/a). Bizonyos esetekben az alacsonyabb taxonómiai szinteknek megfelelő centrumok is elkülöníthetők, azonban ekkor közel sem annyira kifejezettek az említett centrumok (Mellékletek/18. ábra).

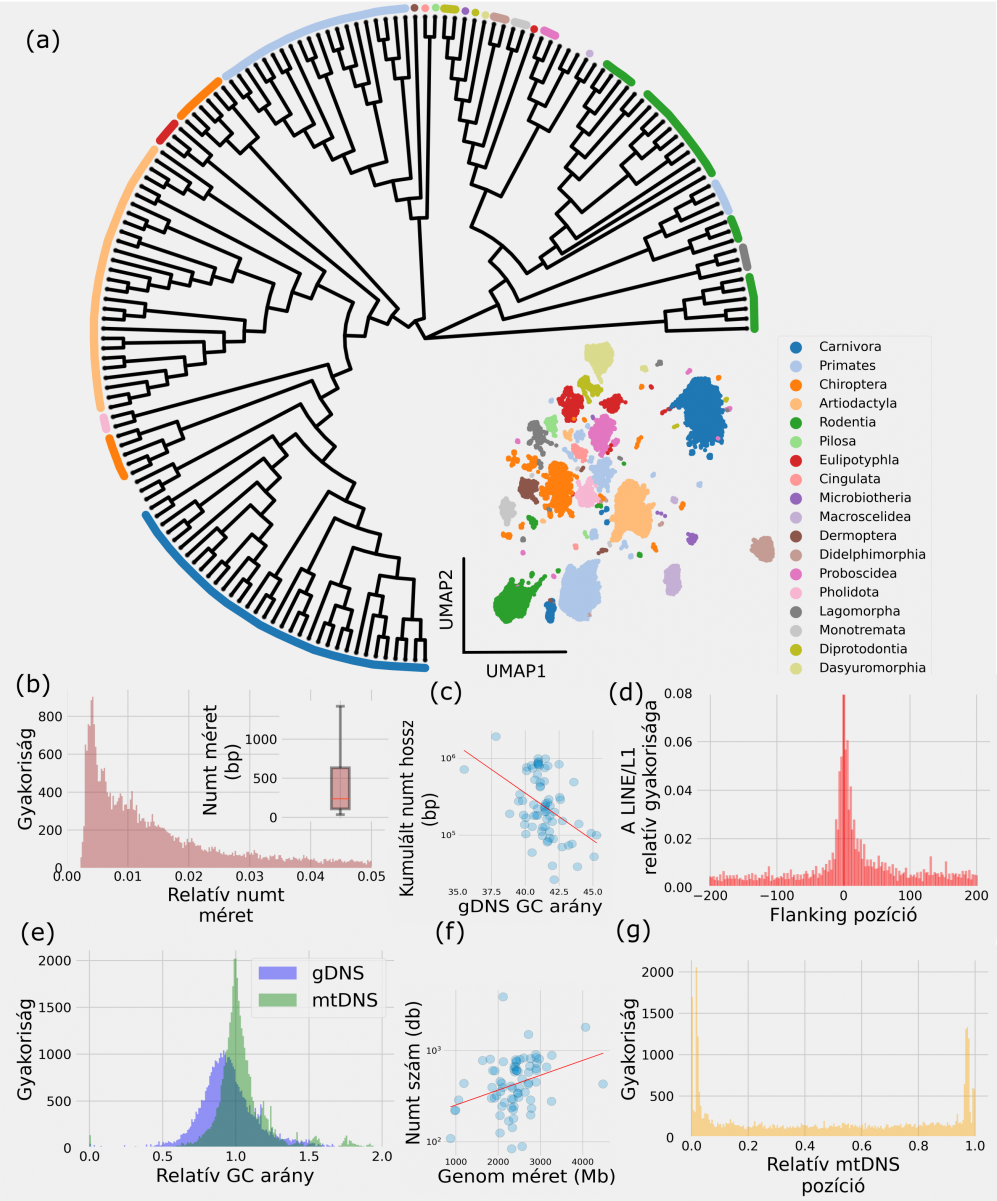
Az emlős NUMTok adott faj mitokondriumjának hosszához viszonyított relatív mérete torzított eloszlást mutatott, jellemző a kisebb méretű NUMTok felülreprezentáltsága (13. ábra/b). Ez a tendencia szignifikáns ($p<0.05$). Az abszolút NUMT hosszak teljes interkvartilis terjedelme 600 bp alatti volt, míg mediánja jóval 250 bp alatt helyezkedett el (13. ábra/b). Három faj esetén (beluga, palackorrú delfin és kanadai hód) kaptunk 1.0 feletti relatív NUMT méretet, tehát ezekben a nukleáris genomokban az adott faj egész mitokondriumja megtalálható.

Az adott mitokondrium hosszához viszonyított pozíció tekintetében azt tapasztaltuk, hogy a linearizált mitokondriumok szélső nukleotidjai gyakrabban vesznek részt a NUMTogenezis folyamatában ($p<0.05$) (13. ábra/g).

A NUMTok gDNS GC tartalmához viszonyított relatív GC arányának átlaga 1.0 alatti értékkel bírt, míg az mtDNS-hez viszonyított relatív GC arány átlaga 1.0-nak bizonyult. Az gDNS GC tartalmához viszonyított relatív NUMT GC arány nagyobb szórással rendelkezett, mint az mtDNS GC tartalmához viszonyított relatív NUMT GC arány (13. ábra/e).

Az 5kb-os határoló régiókban jelenlévő ismétlődő elemek vizsgálata során számos olyan ismétlődő elem osztályt mutattunk ki, amelyek gyakorisága a NUMTokhoz közeledve folyamatosan növekszik (13. ábra/d, Mellékletek/19. ábra). Ezek az ismétlődő elemek a DNA/hAT-Charlie, Simple repeat, LTR/ERV1, LINE/L1, LTR/ERVL-MaLR, DNA/TcMar-Tigger, LTR/ERVL, LINE/L2, SINE/MIR és SINE/Alu csoportokba sorolhatók (Mellékletek/19. ábra).

A genomméret és a NUMTok száma között gyenge pozitív kapcsolat (0.378, $p<0.0001$) mutatható ki. Ezzel ellentétben a genomok mérete és a NUMTok kumulált hossza között egy gyenge negatív kapcsolat (-0.42, $p<0.001$) a jellemző (13. ábra/c,f).



13. ábra. Az NCBI adatbázisban fellelhető emlős genomok NUMTjainak jellegzetességei. A NUMTok számos ismérve alapján számított UMAP centrumok és az egyes rendeknek megfeleltethető filogenetikai klaszterek (a). A NUMTok méretének gyakorisági diagrammja és adott faj mtDNS-énekhosszához viszonyított relatív mérete (b). Adott genom mérete és NUMTjainak összesített hossza (c) illetve száma (f) közötti összefüggés. A LINE/L1 ismétlődő elem feldúsulása a NUMT 200/200 bp-os környezetében (d). NUMTok gDNS és mtDNS GC tartalmához viszonyított relatív GC aránya (e). NUMTok elhelyezkedésének relatív poziói (g).

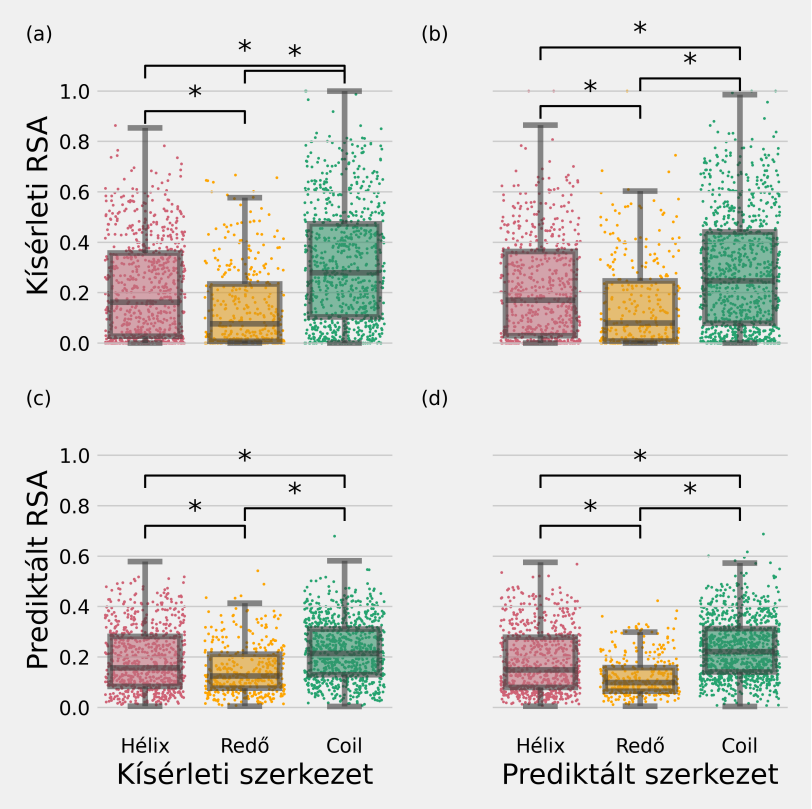
6.2. Proteomika

6.2.1. Az SA és a másodlagos szerkezet összefüggései

Vizsgálataink alapján a C másodlagos szerkezetben lévő savmaradékok rendelkeztek a legnagyobb (tehát a leginkább kitettek az oldószernek), míg az E típusú savmaradékok a legkisebb SA értékekkel. A helikális savmaradékok (H) minden esetben átmenetet képeztek a C és E másodlagos szerkezetben lévő savmaradékok között. Ezek az eltérések szignifikánsak ($p<10^{-5}$) és szignifikáns jellegük az SA és másodlagos szerkezet predikciójának és kísérleti adatának tetszőleges kombinációiban is megmarad (14. ábra).

Ez a tendencia annak ellenére is megmarad, hogy a kísérleti úton meghatározott SA értékek nagyobb skálán mozogtak (14.ábra/a-b), mint a prediktált SA értékek (14.ábra/c-d). A predikció esetén az SA értékek eloszlása sokkal kiegyenlítettebb, mint a kísérleti úton meghatározott SA adatok esetén (14. ábra).

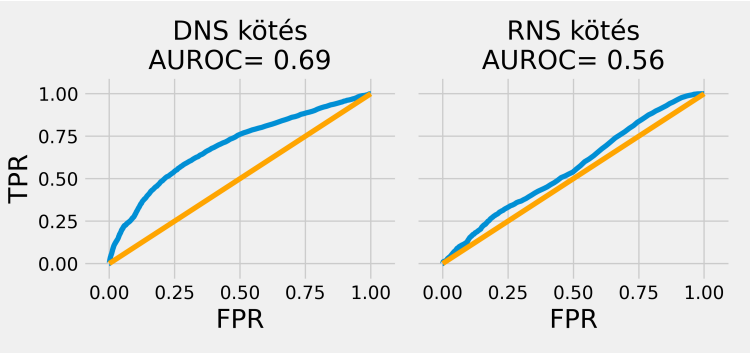
A másodlagos szerkezet predikciójának pontosságát az SOV érték (Egyenlet. 5.12) kiszámításával vizsgáltuk, ami a teljes adathalmazon 0.568-nak bizonyult +-0.13-as szórás mellett. A $SOV_{(H,E,C)}$ értékei 0.63 (+-0.28), 0.55 (+-0.3) és 0.53 (+-.016) voltak.



14. ábra. Az SA és a másodlagos szerkezet predikciójának és valós értékeinek összefüggései. A H a hélix, az E a β-redő, míg a C a coil másodlagos szerkezeti elemre utalnak. A felső sor esetén (a,b) az RSA értékei kísérleti adatokból, míg az alsó sor esetén (c,d) predikcióból származnak. Az első oszlop esetén (a,c) a másodlagos szerkezetek kísérleti adatokból, míg a második oszlop esetén (b,d) predikciókból származnak. A szignifikáns ($p<10^{-5}$) eredményeket * jelöli.

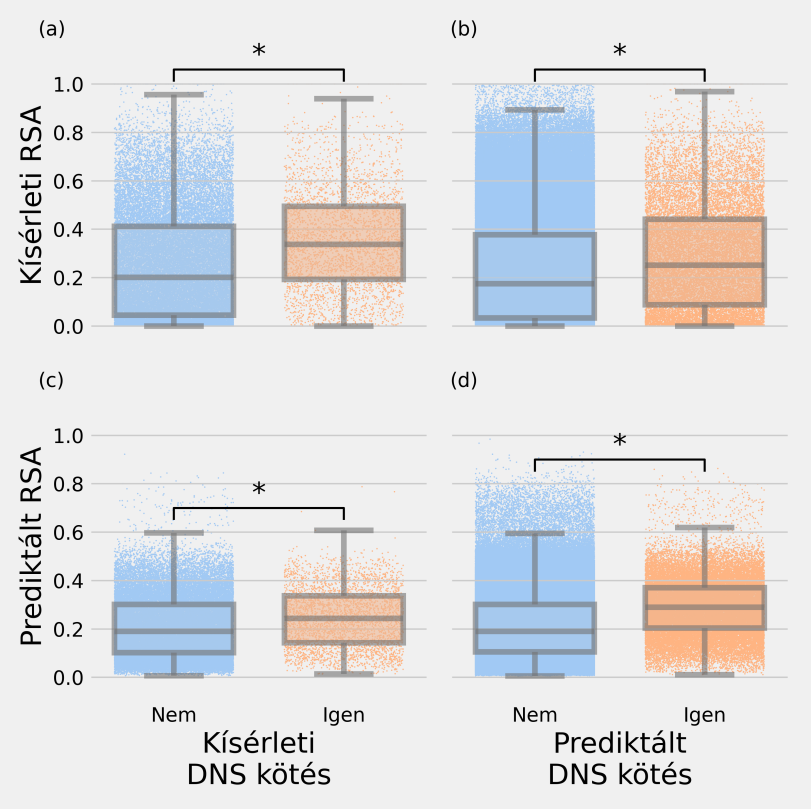
6.2.2. Az SA és a nukleinsav kötés összefüggései

A DRNApred nukleinsav kötés predikciójának pontosságát TPR-FPR arányosítással végeztük. A különböző küszöbértékek (n=1000) esetén kapott TPR-FPR párokat ábrázoltuk, ami ROC görbét eredményezett. Az AUROC érték DNS kötés predikciója esetén 0.69, míg RNS kötés predikciója esetén 0.56 volt (15. ábra).



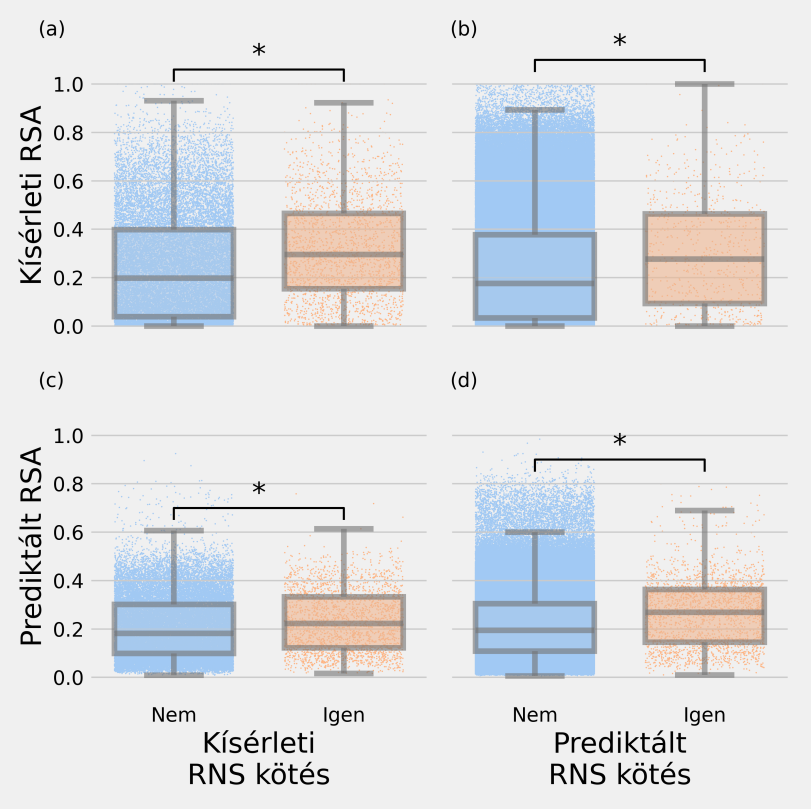
15. ábra. A nukleinsav kötés predikciójának alakulása különböző küszöbértékek esetén az ROC görbe alatti területével.
A kék görbék a DRNApred DNS (a) és RNS (b) predikcióját, míg a narancssárga egyenesek a véletlenszerű klasszifikációt jelölik.

A DNS kötés és az SA vizsgálata során azt találtuk, hogy a DNS-sel interakcióba lépő savmaradékok magasabb SA értékekkel rendelkeztek (azaz nagyobb volt az oldhatóságuk), mint azok a savmaradékok, amelyek nem léptek interakcióba a DNS-sel. Ezek az eltérések szignifikánsak ($p<10^{-5}$) és szignifikáns jellegük az SA és a DNS kötés predikciójának és kísérleti adatának tetszőleges kombinációiban is megmarad. Általánosságban elmondható a DNS kötés és az SA kapcsolatáról, hogy a kísérleti úton meghatározott SA értékek nagyobb skálán mozogtak és több kiugró adatot (RSA=1) tartalmaztak, mint a prediktált SA értékek (16. ábra).



16. ábra. Az SA és a DNS kötés összefüggései.
A felső sor esetén (a,b) az RSA értékei kísérleti adatokból, míg az alsó sor esetén (c,d) predikcióból származnak. Az első oszlop esetén (a,c) a DNS kötés kísérleti adatokból, míg a második oszlop esetén (b,d) predikciókból származnak. A szignifikáns ($p<10^{-5}$) eredményeket * jelöli.

A DNS kötés és az SA között megfigyelt összefüggések az RNS kötés és az SA kapcsolatára is érvényesek. Tehát az RNS-sel interakcióba lépő savmaradékok SA értékei szignifikánsan magasabbak, mint az RNS-sel interakcióba nem lépő savmaradékok SA értékei. Ez a jellegű összefüggés az RNS kötés esetén is megmarad az SA és az RNS kötés predikciójának és kísérleti adatának tetszőleges kombinációiban (17. ábra).



17. ábra. Az SA és az RNS kötés összefüggései.
A felső sor esetén (a,b) az RSA értékei kísérleti adatokból, míg az alsó sor esetén (c,d) predikcióból származnak. Az első oszlop esetén (a,c) a RNS kötés kísérleti adatokból, míg a második oszlop esetén (b,d) predikciókból származnak. A szignifikáns ($p<10^{-5}$) eredményeket * jelöli.

7. Következtetések és javaslatok

7.1. NUMT biológia

7.1.1. A nyúl genom NUMTjainak jellegzetességei

Sok más eukarióta genomhoz hasonlóan, a nyúl gDNS is tartalmaz NUMTokat (Calabrese et al., 2017). Vizsgálataink alapján a kromoszómákon és szkaffoldokon lévő NUMTok sok esetben nagyfokú szekvencia-hasonlóságot mutattak. Ez a jelenség megnövelheti az FPR-t és félrevezető eredményekkel szolgálhat abban az esetben, ha az adott szkaffold egy kromoszóma része. Ennek a kérdésnek az eldöntésére terveztünk egy vizsgálatot, ami összehasonlította az egyező NUMTok kromoszómális és szkaffold határoló régióinak hasonlóságát és a határoló régiókban lévő ismétlődő elemek frekvenciáját. Ezekből az eredményekből az a következtetés vonható le, hogy a nyúl gDNS-ben lévő és szkaffoldokról származó NUMTok nem tekinthetők FP-knek és valószínűleg egy copy-paste mechanizmusnak köszönhető a jelenlétük, amely jelenség lehetőségét már számos tanulmány valószínűsítette (Hazkani-Covo & Covo, 2008; Lutz-Bonengel et al., 2021; Calabrese et al., 2017). Azonban a teljesség végett meg kell említeni, hogy pillanatnyilag nincs konszenzus a tudományos közösségen belül a szkaffoldokról származó NUMTokkal kapcsolatban. Léteznek olyan tanulmányok, amelyek kizárólag a kromoszóma lokalizált NUMTokat vonják be a további kísérletekbe (Grau et al., 2020; Behura, 2007), míg több szerző a szkaffold lokalizált NUMTokat is analizálja (Calabrese et al., 2017; Shi et al., 2017; J.-X. Wang et al., 2020). Eredményeinket erősíti a tény, hogy az általunk a nyúl gDNS-ben lokalizált 153 db NUMT száma nagyságrendileg megegyezik egy korábban publikált átfogóbb jellegű tanulmány adataival (Hazkani-Covo et al., 2010a), amelyben 180 db körüli NUMTot valószínűsítettek a nyúl gDNS esetén.

Vizsgálataink alapján bizonyos kromoszómák (5,6,8,10,21) egyáltalán nem rendelkeznek NUMTokkal, míg például az 1-es kromoszómába 9 NUMT integrálódott. A NUMTok koromoszóma szintű eloszlásának ilyen jellegű egyenlőtlensége megfigyelhető például az európai méh (Behura, 2007), a szarvasmarha (Grau et al., 2020) és számos denevérfaj (G. Zhang et al., 2021) esetén is.

Abban az esetben, ha a szkaffoldokat is bevontuk a vizsgálatainkba, azt találtuk, hogy az adott genomi régió (kromoszóma vagy szkaffold) mérete és az abba integrálódott NUMTok száma, hossza között nincs kapcsolat. Viszont, ha kizárólag a kromoszómákat és az azokba integrálódott NUMTokat vizsgáltuk, erős pozitív kapcsolatot kaptunk. Ezek az eredmények eltérnek az európai méh gDNS-ben feltárt kapcsolatoktól, ahol a szkaffoldok eltávolítása után sem találtak kimutatható kapcsolatot a kromoszóma méret és a NUMTok száma, mérete között (Behura, 2007).

Ismert jelenség, hogy a NUMTok által indukált inszerciós mutagenezis befolyásolhatja külön-

böző betegségek kialakulását (Ju et al., 2015; Singh et al., 2017; Wei et al., 2022). Az általunk azonosított 12 intragenikus NUMT sok esetben olyan génekbe épült be, amelyek termékei olyan jelátviteli útvonalakban vesznek részt, amelyekben bekövetkező módosulás fejlődési rendellenességek kialakulásához és malignus transzformációkhoz vezethet. Például a *CFAP300* génben bekövetkező mutáció hozzájárulhat a hím fertilitás csökkenéséhez (Aprea et al., 2021) míg a *MEMO1* gén mutációját emlőrák mintákban és embrionális fejlődési rendellenességek esetén is azonosították (Schotanus & Van Otterloo, 2020).

A GC tartalom vizsgálatakor azt az eredményt kaptuk, hogy a NUMTok genomi környezetének és a NUMToknak is szignifikánsan alacsonyabb a GC tartalma, mint a gDNS-ből vett véletlenszerű mintáknak. Ez közvetett módon bizonyítja, hogy az általunk NUMTként definiált szekvenciák ténylegesen a mitkondriumból származnak. A NUMTok környezetüktől eltérő GC tartalma több tanulmányban is megjelenik (Porter & Hajibabaei, 2021; Srinivasainagendra et al., 2017; Behura, 2007; Calderon, 2012), sőt humán NUMTok esetén hasonló tendenciát mutattak ki, tehát a NUMToknak és környezetüknek is alacsonyabb volt a GC aránya, mint a gDNS más részeinek (Mishmar et al., 2004). A NUMTok alacsonyabb GC tartalmú genomi környezetbe való integrációjának oka, hogy ezek a genomi régiók általában nem tartalmaznak géneket, így a NUMTokon nem érvényesül szelekciós nyomás (Lascaro et al., 2008). A NUMTok környezetében megfigyelt alacsonyabb GC tartalom diszkutálásakor mindenképpen meg kell említeni egy eddig még kísérletekkel nem alátámasztott, spekulatív magyarázatot. Ez pedig, hogy a kromoszómális szerkezet nagymértékben befolyásolhatja a NUMT integrációt. Egy GC gazdag genomi régió stabilabb a nagyobb számú hidrogén híd kötés következtében, mint egy hasonló hosszúságú AT gazdag régió (H. Chen & Skylaris, 2021). Ebből adódóan az alacsony GC tartalom szerkezetileg instabil genomi régiót jelezhet, ami így kitettebb a NUMTok integrációjának.

A NUMTok környezetében feldúsuló ismétlődő elemeket már több tanulmány is leírta (J.-X. Wang et al., 2020; Mishmar et al., 2004). A szignifikáns dúsulást mutató SINE és LINE csoportok ún. non-long terminal repeat (non-LTR) típusú retrotranszpozonok (Richardson et al., 2015). Számos esetben kimutatták, hogy a már gDNS-en belüli NUMT elterjedés retrotranszpozonokra vezethető vissza (Krampis et al., 2006; Black IV & Bernhardt, 2009), sőt nem lehet kizárni azt a lehetőséget sem, hogy a transzpozonok közrejátszanak magában az integrációs eseményben is (Song et al., 2013).

Az SVM-mel kapott eredmények alapján kijelenthető, hogy a nyúl genomban feltárt NUMTok elkülöníthetők véletlenszerű szekvenciáktól az adott szekvencia gDNS pozícióját, hosszát és környezetének GC tartalmát felhasználva a tanítás során. A modellünk a 10^{-16}-10^{-14} közötti komplexitás tartományban optimálisan teljesített (szürke szaggatott vonalak közötti tartomány a 11.

ábrán). Az említett tartomány előtt a tanító és a tesztelő adathalmazon kapott pontosság értékei folyamatosan javultak. 10^{-14} komplexitás érték felett a tanító adathalmazon tapasztalt pontosság növekedett, míg a tesztelő adathalmazon mérhető pontosság értéke csökkent, ami klasszikus jele a gépi tanulás során elkövetett legjellegzetesebb hibának, a túltanulásnak (Demšar & Zupan, 2021). Ekkor a modell a tanuló adathalmazban lévő lokális motívumokra nagy pontossággal illeszt görbét. Azonban a modell által felsimert törvényszerűsgek kiterjesztése, az általánosítás során egy ismeretlen adathalmazra (tesztelő adathalmaz) is ezt a görbét próbálja illeszteni, ami alacsony pontosságban fog megmutatkozni.

7.1.2. Egységes keretrendszer NUMTok bányászatára

Az emlős NUMTok általunk vizsgált jellegzetességeik alapján az adott faj taxonómiai rendjének megfelelő UMAP centrumokat mutatnak. Ez az eredmény igazolja a NUMTok filogenetikai alkalmazásának relevanciáját. A NUMTok filogenetikai felhasználása bevett gyakorlat bizonyos fajokat illetően (Ko et al., 2015; Nacer & do Amaral, 2017), azonban saját eredményeink alapján a NUMTokra alapozott filogenetikai vizsgálatok megalapozottnak tekinthetők nagyobb taxonómiai egységek esetén is.

A NUMTok rövidsége (eltekintve néhány kiugró, adott mitokondrium felét vagy egészét érintő NUMToktól) nagy valószínűséggel az integrációt követő fragmentációra és transzpozon aktivitásra vezethető vissza (J.-X. Wang et al., 2020). Ezt az elméletet erősíti, hogy a NUMTok többségének GC tartalma jóval elmarad az adott gDNS GC tartalmától.

A számos ismétlődő elem NUMTok környezetében megfigyelhető feldúsulásának ténye a NUMTogenezis nem véletlenszerű módjára enged következtetni (Tsuji et al., 2012).

7.2. Proteomika

Vizsgálatainknak megfelelően a coil másodlagos szerkezetben lévő samaradékok rendelkeztek a legmagasabb, míg a β-redő konformációjú savmaradékok a legalacsonyabb SA értékkel. Ez a megfigyelés megfelel a szakirodalmi adatoknak (Zhu & Blundell, 1996; H. Zhang et al., 2009). A β-redő konformációban lévő savmaradékok alacsonyabb SA értékeire a következőkben bemutatott módon, a fehérjék folding mechanizmusa szolgál magyarázatul. A biológiai reakciók vizes közegben mennek végbe (Ball, 2017). Ebből adódóan a vízben oldható fehérjék feltekeredését termodinamikai szempontból elsődlegesen a hidrofób erő hajtja (Haque & Bayford, 2019), aminek következtében már a folding folyamatának elején kialakul egy hidrofób mag (Kalinowska et al., 2017). A magot alkotó hidrofób jellegű savmaradékok sok esetben β-redőkbe tömörülnek, hiszen

a hidrofób oldalláncok ”elrejtésére” ez a másodlagos szerkezeti elem a legkedvezőbb a vizes közegben mutatott aggregáció és kompakció miatt (Lins et al., 2003; Fujiwara et al., 2012; Ilyina et al., 1997). A coil másodlagos szerkezetbe tömörülő savmaradékok szignifikánsan magasabb SA értékeit igazolja a tény, hogy a coil konformáció magasabb szerkezeti flexibilitással jellemezhető, ami utal az ilyen jellegű savmaradékok molekula felszínen való elhelyzkedésére (H. Zhang et al., 2009).

A teljes humán proteómon kivitelezett nukleinsav kötés predikciójának pontossága (DNS AUROC=0.69, RNS AUROC=0.56) (15. ábra) bizonyos mértékben elmarad az irodalmi adatoktól (DNS AUROC=0.77, RNS AUROC=0.67) (Yan & Kurgan, 2017). Ez az eltérés nagy valószínűséggel az általunk használt adathalmaz komplexitásából adódik. Ugyanakkor a DRNApred predikció a kísérletes úton alátámasztott irodalmi adatoknak megfelelően reprodukálta a nukleinsavat kötő és nem kötő savmaradékok SA értékei közötti összefüggést.

A nukleinsav kötés predikciójának pontosságában nagy a különbség annak a függvényében, hogy DNS vagy RNS kötést prediktálunk, hiszen míg DNS kötés esetén 0.69 AUROC-ot kaptunk, addig RNS kötésnél ez az érték 0.67-nek bizonyult (15. ábra). Ennek a különbségnek több oka is lehet. Egyrészt a fehérjék RNS kötésének biofizikai háttere még nem teljesen tisztázott, ráadásul az RNS-ek számtalan konformációs állapotban lehetnek jelen fiziológiás körülmények között (Miao & Westhof, 2015), így nem sikerült olyan jellemzőt leírni, ami egyértelműen meghatározná egy fehérje savmaradék RNS kötésének tényét és megfelelő bementként szolgálna a különböző modellek számára. Másrészt az RNS kötés vizsgálatához szükséges kísérleti adatok között nagyságrendekkel nagyobb a száma az olyan savmaradékoknak, amik nem kötnek RNS-t. Ez a bizonyos bemenet felülreprezentáltsága okozta kiegyensúlyozatlanság túltanulást okozhat, ami nagymértékben ronthatja a predikció pontosságát (Tang et al., 2017). A fentebb említetteken kívül az RNS kötés becslésének alacsonyabb pontossága részben visszavezethető a keresztpredikcióra is, azaz arra a jelenségre, mikor egy RNS kötő savmaradékot a modell DNS kötő savmaradékként értékel. Ennek a típusú keresztpredikciónak a fő oka, hogy sok RNS típus fehérjékkel történő interakciója hasonló töltésviszonyok között megy végbe, mint ami a DNS-fehérje interakciókra jellemző (Yan & Kurgan, 2017).

Eredményeink alapján a nukleinsavakkal interakcióba lépő savmaradékok minden esetben nagyobb SA értékekkel rendelkeznek, mint azok a savmaradékok, amik nem lépnek interakcióba nukleinsavakkal. A magasabb SA érték oka, hogy a nukleinsavakkal interakcióba lépő savmaradékoknak a fehérjemolekula felszínén kell elhelyezkedniük annak érdekében, hogy a kapcsolat kialakulhasson. Így a felszínen elhelyezkedő savmaradékoknak nagyobb lesz az SA értékük, mint azoknak a savmaradékoknak, amik nem lépnek interakcióba nukleinsavakkal és a molekula belse-

jében helyezkednek el (Mukherjee & Bahadur, 2018; Ahmad et al., 2004; Pan et al., 2020; T. Zhang et al., 2010).

8. Új tudományos eredmények

1. Elsőként tártuk fel a nyúl genom NUMTjait
2. Sikerült a NUMTok GC arányának a genomhoz viszonyított eltérését bizonyítanunk nyúl esetén
3. Leírtunk néhány repetitív elemet, amelyek frekvenciája feldúsul a nyúl genom NUMTjainak környezetében
4. Az NCBI adatbázisban fellelhető összes emlős genom NUMTjait jellemeztünk, amelyek között számos olyan genom volt, aminek a NUMTjait még nem írták le
5. Bizonyítottuk a savmaradék szintű SA, másodlagos szerkezet és nukleinsavakkal történő interakciókra vonatkozó predikciók kísérletes adatokkal való komplementaritását a humán proteóm esetén
6. A komplementaritás vizsgálatával a prediktív teljesítmény becslésére egy új munkamenetet dolgoztunk ki

9. Összefoglalás

A technológiai előrehaladás következtében önálló diszciplinává fejlődő adattudomány egyre inkább áthatja az élettudományokat is. A nagyméretű és komplex adathalmazok megjelentek a biológia ”omika” területein, a rendszerbiológiában de még az egyed feletti szerveződési szintet kutató ökológiában is. Ami még fontosabb, hogy számos biológiai adatbázis publikusan hozzáférhető és az ezeken történő navigáció illetve több, alapvető vizsgálat elvégzése nem igényel nagy számítási és tárolási kapacitást a felhasználói oldalról. Természetesen az adatbázisok mellett az azokban rejlő mintázatokat kimutatni képes algoritmusok is fejlődésen mentek keresztül. Például az RNS szekvenálási kísérletek egyik alapvető módszerének, a PCA-nak a létrejöttét is az adattudomány tette lehetővé. Enélkül a metódus nélkül nagyon problémás lenne az esetenként több ezer gén expresszióját beálló változások emberek számára értelmezhető módon történő bemutatása. Az így kialakult adat-infrastruktúra (adatbázisok és a valamilyen módon hozzájuk kapcsolódó módszerek) legnagyobb vívmánya, hogy képesek a korábban értelmezhetetlen adathalmazokat jelentéssel felruházni azáltal, hogy felszínre hozzák és megmutatják az azokban lévő belső szerkezeteket. Doktori disszertációmban a biológia két részterületének (mitokondriális genomika és proteomika) egy-egy szeletét vizsgáltuk meg az adattudomány irányából közelítve. A mitokondriális genomika területéhez tartozó NUMTogenezis jelenség folyamán a feldarabolódó mitokondriális genomból a sejtmagi genomba beékelődő szekvenciákat, a NUMTokat kutattuk. A NUMTok jelentős szerepet játszanak bizonyos daganattípusok kialakulásában, de felhasználhatók filogenetikai és igazságügyi vizsgálatokhoz is. Disszertációmban elsőként a nyúl genom NUMTjait jellemeztük. Ezekben a vizsgálatokban meghatároztuk a NUMTok sejtmagi genomban értelmezett koordinátáit, majd ezek alapján azonosítottuk az intragenikus NUMTokat. A NUMTok környezetének vizsgálatával bebizonyítottuk, hogy a NUMTok határoló régióiban alacsonyabb a GC tartalom, mint a sejtmagi genom egyéb részein. Továbbá sikerült kimutatnunk, hogy bizonyos ismétlődő elemek előfordulása is gyakoribb a NUMTok környezetében, mint az várható lenne. Az általunk vizsgált ismérveken tanított SVM modellünk képes volt a NUMTok és véletlenszerű szekvenciák nagy pontosságú elkülönítésére. Ezt követően, a nyúl genomon kivitelezett vizsgálatokat kiterjesztettük az NCBI adatbázison elérhető emlős genomokra. A 150 körüli emlős genomban fellelhető NUMTok általunk vizsgált jellegzetességei alapján a taxonómiai rendeknek megfelelő elkülönülést mutattak. Azonban a kiterjesztett NUMT bányászati vizsgálat legfontosabb eredménye, hogy számos ismétlődő elem esetén sikerült bizonyítanunk, hogy azok a NUMTok környezetében feldúsúlnak. A NUMTokkal kapcsolatos kutatásaink egyik új tudományos eredménye, hogy az eddigi legátfogóbb jellegű vizsgálatban, egységes módszertannal jellemeztük az elérhető emlős genomokat. Számos olyan faj genomját

vizsgáltuk, amiknek a NUMTogenezis irányából történő leírása ezidáig még nem történt meg. A mitokondriális genomikai vizsgálatokat a Magyar Agrár és Élettudományi Egyetem Genetikai és Biotechnológia Intézetének Állatbiotechnológia tanszékén végeztük Dr Hoffmann Orsolya Ivett témavezetésével. A doktori disszertációm által érintett másik területettel, a proteomikával kapcsolatos kutatásokat a Virginiai Nemzetközösségi Egyetem Mérnöki Kar Számítógéptudományi Intézetének szerkezetbioinformatikai laboratóriumában végeztem Dr Lukasz Kurgan mentorálásával. Ebben a témakörben a fehérjék savmaradék-ion szintű másodlagos szerkezetét és különböző funkcióit előrejelző modellek komplementaritását vizsgáltuk a kísérleti adatok függvényében. A téma jelentősége, hogy számos mesterséges intelligencia alapú modellt hoztak létre, amik képesek a fehérjék szekvenciájából valamilyen jellegzetességüket megbízhatóan előre jelezni. Azonban ezen modellek teljesítményének becslése sok esetben valamilyen szempontból szűrt (szubcelluláris lokalizáció, meghatározott feladat stb.) adathalmazon történik annak ellenére, hogy a bemenetük ugyanaz a szekvencia. Doktori dolgozatomnak ebben a részében a humán proteóm fehérjéinek oldhatóságát, másodlagos szerkezetét és a nukleinsavakkal való kölcsönhatását vizsgáltuk. A kiválasztott ismérvek fontossága főként a gyógyszertervezés felgyorsításában mutatkozik meg. Sikerült bizonyítanunk, hogy az élvonalba tartozó előrejelző modellek a kísérleti adatoknak megfelelő viszonyokat megbízhatóan reprodukálják. Ráadásul a kísérleti adatoknak megfelelő mintázat abban az esetben is megmarad, ha a bementek forrását tetszőlegesen kombináljuk. A modellek teljesítményének ilyen jellegű becslése tovább növeli a mesterséges intelligencia alapú előrejelzések pontosságát, ami számos felhasználási területen megmutatkozik. Ráadásul ez a szemlélet segítségül lehet a modellekhez szükséges ”feature engineering/extraction” során is, amikor cél a minél nagyobb prediktív erővel bíró ismérvek kimutatása.

10. Summary

As technology proceeds, data science is getting more and more prominent in nearly every science application. This tendency also effects and somehow revolutionizes life sciences. Huge and complex data sources are getting into omics, system biology as well as supraindividual biology aka ecology. What is even more striking is that countless publicly available biological databases do exist. As to an end user the computational and storage cost of navigating and performing basic analyses on these databases are negligible. Beside database evolution, the associated algorithms also became more developed. For instance, PCA became the first go-to when it comes to RNA-sequencing which is primarily due to the achievements of data science. It would be quite challenging and nearly impossible to wrap our heads around the complexity of the changes of thousands and thousands of gene's expressions without dimension reduction methods like PCA. The biggest know-how of this data science based infrastructure (databases and the associated methodology) is that it is able to transform complex data sets into human readable form by extracting inner, hidden structures. In my doctoral dissertation, parts of two distinct areas of biology have been investigated (namely mitochondrial genomics and proteomics) from a data science point of view. In the mitochondrial genomics part, nuclear sequences with mitochondrial origins aka NUMTs were scrutinised. NUMTs have strong influence on cancer biology research and they have important applications in the fields of phylogeny and forensic studies also. In my dissertation, firstly the NUMTs of the rabbit genomes were characterized. In this part, the genomic coordinates of the NUMTs were defined and so based on these coordinates we were able to identify intragenic NUMTs. By the investigation of the GC content of the flanking regions of NUMTs, it turned out that genomic surrounding of NUMTs have significantly altered, namely the flankings had lower GC content than the rest of the genome. Moreover, we proved that the frequency of several repetitive elements are persistent near the NUMTs. Then, based on the features on NUMTs a SVM was taught to be able to differentiate NUMTs from random sequences. After all, the previously described workflow has been scaled up to all mammalian genomes of NCBI and so a NUMT mining pipeline has been established. The NUMTs from nearly 150 genomes showed distinguishing features based on taxonomy data. The most important result of this NUMT mining workflow is that several repetitive elements displayed elevated frequency in the flanking regions of NUMTs. New scientific finding is that several previously uncharacterised genomes have been investigated from the NUMTogenesis point of view with a uniform methodology. As to our current knowledge, this study is the biggest high throughput analysis of NUMTs in different species. The mitochondrial genomic part of my doctoral dissertation was performed at the Hungarian University of Agriculture

and Life sciences, Institute of Genetics and Biotechnology, Department of Animal Biotechnology under the supervision of Dr Orsolya Ivett Hoffmann. While the proteomics related part of my dissertation was done at the Virginia Commonwealth University, College of Engineering, Department of Computer Science in the structural bioinformatics laboratory of Dr Lukasz Kurgan. In the proteomics part of my dissertation, the residue level complementarity of protein structure and function predictors were evaluated on the human proteome. The significance of this topic is that several artificial intelligence based predictors have been published recently which inputs are protein sequences. However the predictive performance of these models are evaluated on somehow preprocessed, prefiltered datasets (subcellular localisation, defined function etc.) even though their inputs are the same sequences. In this section we investigated residue level solvent accessibility, secondary structure and interactibility with nucleic acids. These features are extremely important at different steps of drug design. We proved that state-of-the-art predictive models successfully reproduce the motifs that are present in the experimental dataset. Additionally we found these motifs even when different data sources were combined arbitrarily. This kind of evaluation of predictive methods can help to achieve elevated predictive performance. Moreover this evaluation method is able to facilitate feature engineering/extraction.

11. Mellékletek

11.1. Irodalomjegyzék

Abdullaev, S., Fomenko, L., Kuznetsova, E., & Gaziev, A. (2013). Experimental detection of integration of mtdna in the nuclear genome induced by ionizing radiation. *Radiatsionnaia Biologiia, Radioecologiia*, *53*(4), 380–388.

Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, *20*(4), 477–486.

Aprea, I., Raidt, J., Höben, I. M., Loges, N. T., Nöthe-Menchen, T., Pennekamp, P., … others (2021). Defects in the cytoplasmic assembly of axonemal dynein arms cause morphological abnormalities and dysmotility in sperm cells leading to male infertility. *PLoS genetics*, *17*(2), e1009306.

Backman, L. (2019). Protein chemistry. In *Protein chemistry.* De Gruyter.

Bah, T. (2009). *Inkscape: guide to a vector drawing program (digital short cut)*. Pearson Education.

Ball, P. (2017). Water is an active matrix of life for cell and molecular biology. *Proceedings of the National Academy of Sciences*, *114*(51), 13327–13335.

Behura, S. K. (2007). Analysis of nuclear copies of mitochondrial sequences in honeybee (apis mellifera) genome. *Molecular biology and evolution*, *24*(7), 1492–1505.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235–242.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). Mitos: improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution*, *69*(2), 313–319.

Black IV, W. C., & Bernhardt, S. A. (2009). Abundant nuclear copies of mitochondrial origin (numts) in the aedes aegypti genome. *Insect Molecular Biology*, *18*(6), 705–713.

Breton, S., & Stewart, D. T. (2015). Atypical mitochondrial inheritance patterns in eukaryotes. *Genome*, *58*(10), 423–431.

Bzdok, D., Krzywinski, M., & Altman, N. (2017). Machine learning: a primer. *Nature methods*, *14*(12), 1119.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Statistics versus machine learning. *Nat Methods*, *15*(4), 233.

Calabrese, F., Balacco, D., Preste, R., Diroma, M., Forino, R., Ventura, M., & Attimonelli, M. (2017). Numts colonization in mammalian genomes. *Scientific reports*, *7*(1), 1–10.

Calderon, I. D. S. (2012). *Evolution of nuclear integrations of the mitochondrial genome in great apes and their potential as molecular markers* (Unpublished doctoral dissertation). University of New Orleans.

Campbell, I. D. (2013). The evolution of protein nmr. *Biomedical Spectroscopy and Imaging*, *2*(4), 245–264.

Caro, P., Gómez, J., Arduini, A., González-Sánchez, M., González-García, M., Borrás, C., … Barja, G. (2010). Mitochondrial dna sequences are present inside nuclear dna in rat tissues and increase with age. *Mitochondrion*, *10*(5), 479–486.

Chackalamannil, S., Rotella, D., & Ward, S. (2017). *Comprehensive medicinal chemistry iii*. Elsevier.

Chan, C. Y., Kiechle, M., Manivasakam, P., & Schiestl, R. H. (2007). Ionizing radiation and restriction enzymes induce microhomology-mediated illegitimate recombination in saccharomyces cerevisiae. *Nucleic acids research*, *35*(15), 5051–5059.

Chen, G., Kroemer, G., & Kepp, O. (2020). Mitophagy: an emerging role in aging and age-associated diseases. *Frontiers in cell and developmental biology*, *8*, 200.

Chen, H., & Skylaris, C.-K. (2021). Analysis of dna interactions and gc content with energy decomposition in large-scale quantum mechanical calculations. *Physical Chemistry Chemical Physics*, *23*(14), 8891–8899.

Cheng, X., & Ivessa, A. S. (2010). The migration of mitochondrial dna fragments to the nucleus affects the chronological aging process of saccharomyces cerevisiae. *Aging cell*, *9*(5), 919–923.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … others (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Cortes-Figueiredo, F., Carvalho, F. S., Fonseca, A. C., Paul, F., Ferro, J. M., Schönherr, S., … Morais, V. A. (2021). From forensics to clinical research: expanding the variant calling pipeline for the precision id mtdna whole genome panel. *International journal of molecular sciences*, *22*(21), 12031.

Cozzolino, F., Iacobucci, I., Monaco, V., & Monti, M. (2021). Protein–dna/rna interactions: An overview of investigation methods in the-omics era. *Journal of Proteome Research*, *20*(6),

3018–3030.

Damodaran, S. (2008). Amino acids, peptides and proteins. *Fennema's food chemistry*, *4*, 217–329.

Dana, J. M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., & Velankar, S. (2019). Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, *47*(D1), D482–D489.

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic acids research*, *42*(20), 12640–12649.

Demšar, J., & Zupan, B. (2021). Hands-on training about overfitting. *PLoS Computational Biology*, *17*(3), e1008671.

D'Argenio, V. (2018). The high-throughput analyses era: are we ready for the data struggle? *High-throughput*, *7*(1), 8.

Edman, P., & Begg, G. (1967). A protein sequenator. In *European journal of biochemistry* (pp. 80–91). Springer.

Esteves, P. J., Abrantes, J., Baldauf, H.-M., BenMohamed, L., Chen, Y., Christensen, N., … others (2018). The wide utility of rabbits as models of human diseases. *Experimental & molecular medicine*, *50*(5), 1–10.

Fan, J., Chen, Y., Yan, H., Niimi, M., Wang, Y., & Liang, J. (2018). Principles and applications of rabbit models for atherosclerosis research. *Journal of atherosclerosis and thrombosis*, *25*(3), 213–220.

Fan, J., Wang, Y., & Chen, Y. E. (2021). Genetically modified rabbits for cardiovascular research. *Frontiers in Genetics*, *12*, 14.

Faraggi, E., Zhou, Y., & Kloczkowski, A. (2014). Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins: Structure, Function, and Bioinformatics*, *82*(11), 3170–3176.

Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine*, *53*(2), 1385–1390.

Fujiwara, K., Toda, H., & Ikeguchi, M. (2012). Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. *BMC structural biology*, *12*(1), 1–15.

Gaziev, A., & Shaikhaev, G. (2007). Ionizing radiation can activate the insertion of mitochondrial dna fragments in the nuclear genome. *Radiatsionnaia Biologiia, Radioecologiia*, *47*(6), 673–683.

Gilbert, W. (1991). Towards a paradigm shift in biology. *Nature*, *349*(6305), 99.

Goldman, S. J., Taylor, R., Zhang, Y., & Jin, S. (2010). Autophagy and the degradation of mitochondria. *Mitochondrion*, *10*(4), 309–315.

Grau, E. T., Charles, M., Féménia, M., Rebours, E., Vaiman, A., & Rocha, D. (2020). Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Scientific reports*, *10*(1), 1–11.

Haque, M. M., & Bayford, R. (2019). *Protein misfolding thermodynamics* (Vol. 10) (No. 10). ACS Publications.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … others (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hazkani-Covo, E., & Covo, S. (2008). Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS genetics*, *4*(10), e1000237.

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010a). Molecular poltergeists: mitochondrial dna copies (numts) in sequenced nuclear genomes. *PLoS genetics*, *6*(2), e1000834.

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010b). Molecular poltergeists: mitochondrial dna copies (numts) in sequenced nuclear genomes. *PLoS genetics*, *6*(2), e1000834.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. *Genomics*, *107*(1), 1–8.

Howe, D. K., & Denver, D. R. (2008). Muller's ratchet and compensatory mutation in caenorhabditis briggsae mitochondrial genome evolution. *BMC evolutionary biology*, *8*(1), 1–13.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Ilyina, E., Roongta, V., & Mayo, K. H. (1997). Designing water soluble β-sheet peptides with compact structure. In *Techniques in protein chemistry* (Vol. 8, pp. 797–808). Elsevier.

Jaskolski, M., Dauter, Z., & Wlodawer, A. (2014). A brief history of macromolecular crystallography, illustrated by a family tree and its n obel fruits. *The FEBS journal*, *281*(18), 3985–4009.

Ju, Y. S., Tubio, J. M., Mifsud, W., Fu, B., Davies, H. R., Ramakrishna, M., … others (2015). Frequent somatic transfer of mitochondrial dna into the nuclear genome of human cancer cells. *Genome research*, *25*(6), 814–824.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, *22*(12), 2577–2637.

Kalinowska, B., Banach, M., Wiśniowski, Z., Konieczny, L., & Roterman, I. (2017). Is the hydrophobic core a universal structural element in proteins? *Journal of Molecular Modeling*, *23*(7), 1–16.

Kelly, S. (2020). The economics of endosymbiotic gene transfer and the evolution of organellar genomes. *bioRxiv*.

Kelly, S. (2021). The economics of organellar gene loss and endosymbiotic gene transfer. *Genome biology*, *22*(1), 1–22.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research*, *21*(3), 487–493.

Ko, Y.-J., Yang, E. C., Lee, J.-H., Lee, K. W., Jeong, J.-Y., Park, K., … Yim, H.-S. (2015). Characterization of cetacean numt and its application into cetacean phylogeny. *Genes & Genomics*, *37*(12), 1061–1071.

Krampis, K., Tyler, B. M., & Boore, J. L. (2006). Extensive variation in nuclear mitochondrial dna content between the genomes of phytophthora sojae and phytophthora ramorum. *Molecular plant-microbe interactions*, *19*(12), 1329–1336.

Kruglikov, A., Rakesh, M., Wei, Y., & Xia, X. (2021). Applications of protein secondary structure algorithms in sars-cov-2 research. *Journal of Proteome Research*, *20*(3), 1457–1463.

Lascaro, D., Castellana, S., Gasparre, G., Romeo, G., Saccone, C., & Attimonelli, M. (2008). The rhnumts compilation: features and bioinformatics approaches to locate and quantify human numts. *BMC genomics*, *9*(1), 1–13.

Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, *55*(3), 379–IN4.

Lei, Z., Meng, H., Liu, L., Zhao, H., Rao, X., Yan, Y., … Yi, C. (2022). Mitochondrial base editor induces substantial nuclear off-target mutations. *Nature*, 1–1.

Leonelli, S. (2019). Philosophy of biology: the challenges of big data biology. *Elife*, *8*, e47381.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078–2079.

Lins, L., Thomas, A., & Brasseur, R. (2003). Analysis of accessible surface of residues in proteins. *Protein science*, *12*(7), 1406–1417.

Lopez, J. V., Yuhki, N., Masuda, R., Modi, W., & O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial dna to the nuclear genome of the domestic cat. *Journal of molecular evolution*, *39*(2), 174–190.

Lutz-Bonengel, S., Niederstätter, H., Naue, J., Koziel, R., Yang, F., Sänger, T., … others (2021). Evidence for multi-copy mega-numt s in the human genome. *Nucleic acids research*, *49*(3), 1517–1531.

Ma, K., Chen, G., Li, W., Kepp, O., Zhu, Y., & Chen, Q. (2020). Mitophagy, mitochondrial homeostasis, and cell fate. *Frontiers in cell and developmental biology*, *8*, 467.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381–386.

Marshall, C., & Parson, W. (2021). Interpreting numts in forensic genetics: Seeing the forest for the trees. *Forensic Science International: Genetics*, *53*, 102497.

Martin, W., & Herrmann, R. G. (1998). Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant physiology*, *118*(1), 9–17.

Martin, W. F., Garg, S., & Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140330.

Marx, V. (2013). The big challenges of big data. *Nature*, *498*(7453), 255–260.

Matsuhisa, F., Kitajima, S., Nishijima, K., Akiyoshi, T., Morimoto, M., & Fan, J. (2020). Transgenic rabbit models: Now and the future. *Applied Sciences*, *10*(21), 7416.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The psipred protein structure prediction server. *Bioinformatics*, *16*(4), 404–405.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Metzger, J. J., & Eule, S. (2013). Distribution of the fittest individuals and the rate of muller's ratchet in a model with overlapping generations. *PLoS computational biology*, *9*(11), e1003303.

Miao, Z., & Westhof, E. (2015). Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic acids research*, *43*(11), 5340–5351.

Mishmar, D., Ruiz-Pesini, E., Brandon, M., & Wallace, D. C. (2004). Mitochondrial dna-like sequences in the nucleus (numts): Insights into our african origins and the mechanism of foreign dna integration. *Human mutation*, *23*(2), 125–133.

Mukherjee, S., & Bahadur, R. P. (2018). An account of solvent accessibility in protein-rna recognition. *Scientific reports*, *8*(1), 1–13.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *1*(1), 2–9.

Muzio, G., O'Bray, L., & Borgwardt, K. (2021). Biological network analysis with deep learning. *Briefings in bioinformatics*, *22*(2), 1515–1530.

Nacer, D. F., & do Amaral, F. R. (2017). Striking pseudogenization in avian phylogenetics: numts are large and common in falcons. *Molecular phylogenetics and evolution*, *115*, 1–6.

Naito, M., & Pawlowska, T. E. (2016). Defying muller's ratchet: Ancient heritable endobacteria escape extinction through retention of recombination and genome plasticity. *MBio*, *7*(3), e02057–15.

Nishimaki, T., & Sato, K. (2019). An extension of the kimura two-parameter model to the natural evolutionary process. *Journal of molecular evolution*, *87*(1), 60–67.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … others (2016).

Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745.

Osarogiagbon, A. U., Khan, F., Venkatesan, R., & Gillard, P. (2021). Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. *Process Safety and Environmental Protection*, *147*, 367–384.

Pal, S., Mondal, S., Das, G., Khatua, S., & Ghosh, Z. (2020). Big data in biology: The hope and present-day challenges in it. *Gene Reports*, *21*, 100869.

Palodhi, A., Singla, T., & Maitra, A. (2020). Profiling of numts in gingivobuccal oral cancer. *bioRxiv*.

Pamilo, P., Viljakainen, L., & Vihavainen, A. (2007). Exceptionally high density of numts in the honeybee genome. *Molecular biology and evolution*, *24*(6), 1340–1346.

Pan, Y., Zhou, S., & Guan, J. (2020). Computationally identifying hot spots in protein-dna binding interfaces using an ensemble approach. *BMC bioinformatics*, *21*(13), 1–16.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics*, *35*(3), 526–528.

Park, S., Hanekamp, T., Thorsness, M. K., & Thorsness, P. E. (2006). Yme2p is a mediator of nucleoid structure and number in mitochondria of the yeast saccharomyces cerevisiae. *Current genetics*, *50*(3), 173–182.

Pauwels, R., Azijn, H., de Béthune, M.-P., Claeys, C., & Hertogs, K. (1995). Automated techniques in biotechnology. *Current Opinion in Biotechnology*, *6*(1), 111–117.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Peter, S. C., Dhanjal, J. K., Malik, V., Radhakrishnan, N., Jayakanthan, M., Sundar, D., … Jayakanthan, M. (2019). Encyclopedia of bioinformatics and computational biology. *Ranganathan, S., Grib-skov, M., Nakai, K., Schönbach, C., Eds*, 661–676.

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden markov model sequence analysis can help remove putative pseudogenes from dna barcoding and metabarcoding datasets. *BMC bioinformatics*, *22*(1), 1–20.

Puertas, M. J., & González-Sánchez, M. (2020). Insertions of mitochondrial dna into the nucleus—effects and role in cell evolution. *Genome*, *63*(8), 365–374.

Radaeva, M., Ton, A.-T., Hsing, M., Ban, F., & Cherkasov, A. (2021). Drugging the 'undruggable'. therapeutic targeting of protein–dna interactions with the use of computer-aided drug discovery methods. *Drug Discovery Today*, *26*(11), 2660–2679.

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The influence of line-1 and sine retrotransposons on mammalian genomes. *Microbiology spectrum*, *3*(2), 3–2.

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The origin and diversification of mitochondria. *Current Biology*, *27*(21), R1177–R1192.

Salzberg, S. L. (2019). *Next-generation genome annotation: we still struggle to get it right* (Vol. 20) (No. 1). BioMed Central.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1–21.

Savojardo, C., Manfredi, M., Martelli, P. L., & Casadio, R. (2021). Solvent accessibility of residues undergoing pathogenic variations in humans: from protein structures to protein sequences. *Frontiers in molecular biosciences*, *7*, 460.

Schiavo, G., Hoffmann, O. I., Ribani, A., Utzeri, V. J., Ghionda, M. C., Bertolini, F., … Fontanesi, L. (2017). A genomic landscape of mitochondrial dna insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture. *DNA Research*, *24*(5), 487–498.

Schotanus, M. D., & Van Otterloo, E. (2020). Finding memo—emerging evidence for memo1′ s function in development and disease. *Genes*, *11*(11), 1316.

Schwede, T. (2013). Protein modeling: what happened to the "protein structure gap"? *Structure*, *21*(9), 1531–1540.

Shi, H., Xing, Y., & Mao, X. (2017). The little brown bat nuclear genome contains an entire mitochondrial genome: Real or artifact? *Gene*, *629*, 64–67.

Singh, K. K., Choudhury, A. R., & Tiwari, H. K. (2017). Numtogenesis as a mechanism for development of cancer. In *Seminars in cancer biology* (Vol. 47, pp. 101–109).

Song, S., Jiang, F., Yuan, J., Guo, W., & Miao, Y. (2013). Exceptionally high cumulative percentage of numts originating from linear mitochondrial dna molecules in the hydra magnipapillata genome. *BMC genomics*, *14*(1), 1–13.

Srinivasainagendra, V., Sandel, M. W., Singh, B., Sundaresan, A., Mooga, V. P., Bajpai, P., … Singh, K. K. (2017). Migration of mitochondrial dna in the nuclear genome of colorectal adenocarcinoma. *Genome medicine*, *9*(1), 1–15.

Stern, D. B., & Lonsdale, D. M. (1982). Mitochondrial and chloroplast genomes of maize have a 12-kilobase dna sequence in common. *Nature*, *299*(5885), 698–702.

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'perceptron'algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, *10*(9), 2997–3011.

Su, H., Liu, M., Sun, S., Peng, Z., & Yang, J. (2019). Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics*, *35*(6), 930–936.

Sun, P. D., Foster, C. E., & Boyington, J. C. (2004). Overview of protein structural and functional folds. *Current protocols in protein science*, *35*(1), 17–1.

Tang, Y., Liu, D., Wang, Z., Wen, T., & Deng, L. (2017). A boosting approach for prediction of protein-rna binding residues. *BMC bioinformatics*, *18*(13), 47–58.

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology*, *3*(6), e116.

Thorsness, P. E., White, K. H., & Fox, T. D. (1993). Inactivation of yme1, a member of the ftsh-sec18-pas1-cdc48 family of putative atpase-encoding genes, causes increased escape of dna from mitochondria in saccharomyces cerevisiae. *Molecular and cellular biology*, *13*(9), 5418–5426.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013a). Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, *8*(11), e80635.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013b). Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, *8*(11), e80635.

Tsuji, J., Frith, M. C., Tomii, K., & Horton, P. (2012). Mammalian numt insertion is non-random. *Nucleic acids research*, *40*(18), 9073–9088.

Uniprot: the universal protein knowledgebase in 2021. (2021). *Nucleic acids research*, *49*(D1), D480–D489.

Uversky, V. N. (2019). Intrinsically disordered proteins and their “mysterious”(meta) physics. *Frontiers in Physics*, *7*, 10.

Vasudevan, D. M., Sreekumari, S., & Vaidyanathan, K. (2019). *Textbook of biochemistry for medical students*. Jaypee brothers Medical publishers.

Verscheure, S., Backeljau, T., & Desmyter, S. (2015). In silico discovery of a nearly complete mitochondrial genome numt in the dog (canis lupus familiaris) nuclear genome. *Genetica*, *143*(4), 453–458.

Vieira-Pires, R. S., & Morais-Cabral, J. H. (2010). 310 helices in channels and other membrane proteins. *Journal of General Physiology*, *136*(6), 585–592.

Virtanen, P., Gommers, R., Burovski, E., Oliphant, T. E., Weckesser, W., Cournapeau, D., … others (2021). scipy/scipy: Scipy 1.6. 3. *Zenodo*.

Wang, H., Ma, C., & Zhou, L. (2009). A brief review of machine learning and its application. In *2009 international conference on information engineering and computer science* (pp. 1–4).

Wang, J.-X., Liu, J., Miao, Y.-H., Huang, D.-W., & Xiao, J.-H. (2020). Tracking the distribution and burst of nuclear mitochondrial dna sequences (numts) in fig wasp genomes. *Insects*, *11*(10), 680.

Wang, R., Ilangovan, U., Leal, B. Z., Robinson, A. K., Amann, B. T., Tong, C. V., … Kim, C. A. (2011). Identification of nucleic acid binding residues in the fcs domain of the polycomb group protein polyhomeotic. *Biochemistry*, *50*(22), 4998–5007.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wei, W., Schon, K. R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., … Chinnery, P. F. (2022). Nuclear-embedded mitochondrial dna sequences in 66,083 human genomes. *Nature*, *611*(7934), 105–114.

Xu, C., & Jackson, S. A. (2019). *Machine learning and complex biological data* (Vol. 20) (No. 1). Springer.

Yan, J., & Kurgan, L. (2017). Drnapred, fast sequence-based method that accurately predicts and discriminates dna-and rna-binding residues. *Nucleic acids research*, *45*(10), e84–e84.

Yang, J., Roy, A., & Zhang, Y. (2012). Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, *41*(D1), D1096–D1103.

Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature methods*, *16*(8), 687–694.

Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current protocols in bioinformatics*, *69*(1), e96.

Zhang, B., Li, J., & Lü, Q. (2018). Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC bioinformatics*, *19*(1), 1–13.

Zhang, G., Geng, D., Guo, Q., Liu, W., Li, S., Gao, W., … others (2021). Genomic landscape of mitochondrial dna insertions in 23 bat genomes: characteristics, loci, phylogeny, and polymorphism. *Integrative Zoology*.

Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., & Kurgan, L. (2009). On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, *76*(3), 617–636.

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., & Kurgan, L. (2010). Analysis and prediction of rna-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Current Protein and Peptide Science*, *11*(7), 609–628.

Zhao, B., Katuwawala, A., Uversky, V. N., & Kurgan, L. (2021). Idpology of the living cell: intrinsic disorder in the subcellular compartments of the human cell. *Cellular and Molecular*

Life Sciences, *78*(5), 2371–2385.

Zhu, Z.-Y., & Blundell, T. L. (1996). The use of amino acid patterns of classified helices and strands in secondary structure prediction. *Journal of molecular biology*, *260*(2), 261–276.

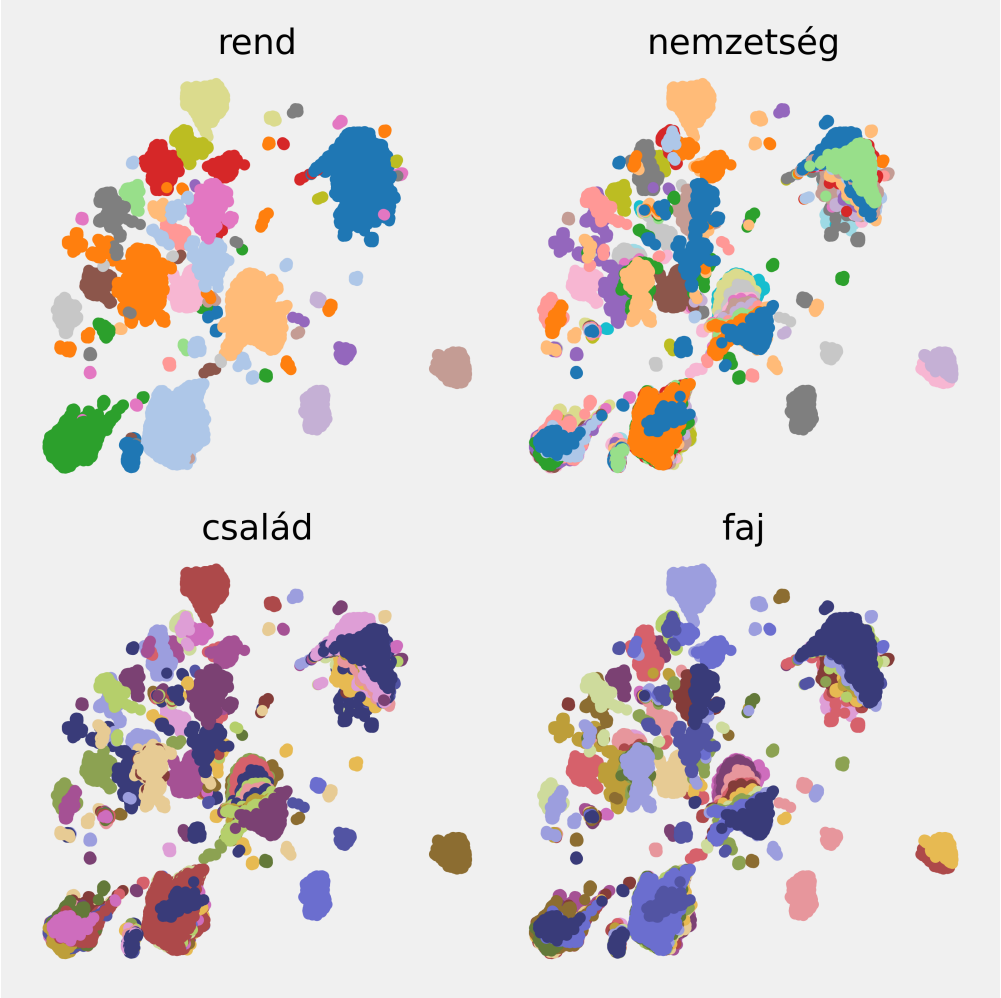
11.2. További mellékletek

11.2.1. M1. Az egységes keretrendszer fejlesztése során felhasznált NCBI genomokhoz tartozó fajnevek

Gepárd (*Acinonyx jubatus*), Óriáspanda (*Ailuropoda melanoleuca*), *Aotus nancymaae*, Jamaikai Gyümölcsdenevér (*Artibeus jamaicensis*), Csukabálna (*Balaenoptera acutorostrata*), Kék Bálna (*Balaenoptera musculus*), Amerikai Bölény (*Bison bison*), Zebu (*Bos indicus*), Vadjak (*Bos mutus*), Szarvasmarha (*Bos taurus*), Házi Vízibivaly (*Bubalus bubalis*), Fehérpamacsos Selyemmajom (*Callithrix jacchus*), Északi Medvefóka (*Callorhinus ursinus*), Kétpúpú Teve (*Camelus bactrianus*), Egypúpú Teve (*Camelus dromedarius*), *Camelus ferus*, Szürkefarkas (*Canis lupus*), Fülöp-szigeteki Koboldmaki (*Carlito syrichta*), Kanadai Hód (*Castor canadensis*), tengerimalac (*Cavia porcellus*), Szélesszájú Orrszarvú (*Ceratotherium simum*), Kormos Mangábé (*Cercocebus atys*), Vapiti (*Cervus canadensis*), Gímszarvas (*Cervus elaphus*), Csincsilla (*Chinchilla lanigera*), Sárgahasú Szavannacerkóf (*Chlorocebus sabaeus*), Kétújjú Lajhár (*Choloepus didactylus*), *Chrysochloris asiatica*, Csillagorrú Vakond (*Condylura cristata*), Kínai Hörcsög (*Cricetulus griseus*), Kilencöves Tatu (*Dasypus novemcinctus*), Beluga (*Delphinapterus leucas*), Rőt Vérszopó Denevér (*Desmodus rotundus*), Törpeoposszum (*Dromiciops gliroides*), (*Echinops telfairi*), *Elephantulus edwardii*, Tengeri Vidra (*Enhydra lutris*), Háziszamár (*Equus asinus*), Ló (*Equus caballus*), Európai Sün (*Erinaceus europaeus*), Steller-oroszlánfóka (*Eumetopias jubatus*), Macska (*Felis catus*), *Fukomys damarensis*, Maláj Repülőmaki (*Galeopterus variegatus*), Hosszúszárnyú Gömbölyűfejű-delfin (*Globicephala melas*), Gorilla (*Gorilla gorilla*), *Gracilinanus agilis*, Kúpos Fóka (*Halichoerus grypus*), Csupasz Trukáló (*Heterocephalus glaber*), *Hipposideros armiger*, Ember (*Homo sapiens*), Leopárdürge (*Ictidomys tridecemlineatus*), Egyiptomi Ugróegér (*Jaculus jaculus*), Csendes-óceáni Fehérsávos Delfin (*Lagenorhynchus obliquidens*), Geoffreoy-macska (*Leopardus geoffroyi*), Weddell-fóka (*Leptonychotes weddellii*), Kínai folyamidelfin (*Lipotes vexillifer*), Afrikai Elefánt (*Loxodonta africana*), Európai Vidra (*Lutra lutra*), Kanadai Hiúz (*Lynx canadensis*), Vörös Hiúz (*Lynx rufus*), Közönséges Makákó (*Macaca fascicularis*), Rézuszmajom (*Macaca mulatta*), Emsemakákó (*Macaca nemestrina*), (*Mandrillus leucophaeus*), Jávai Tobzoska (*Manis javanica*), Bengáli Tobzoska (*Manis pentadactyla*), Sárgahasú Mormota (*Marmota flaviventris*), *Mastomys coucha*, Európai Borz (*Meles meles*), Mongol Futóegér (*Meriones unguiculatus*), Szíriai Aranyhörcsög (*Mesocricetus auratus*), Szürke Egérmaki (*Microcebus murinus*), *Microtus ochrogaster*, Déli Elefántfóka (*Mirounga leonina*), Brazil Oposszum (*Monodelphis domestica*), Narvál (*Monodon monoceros*), Ryukyu Egér (*Mus caroli*), Házi Egér (*Mus musculus*), *Mus pahari*, Hermelin (*Mustela erminea*), Vöröshátú Erdeipocok (*Myodes glareolus*), Brandt-

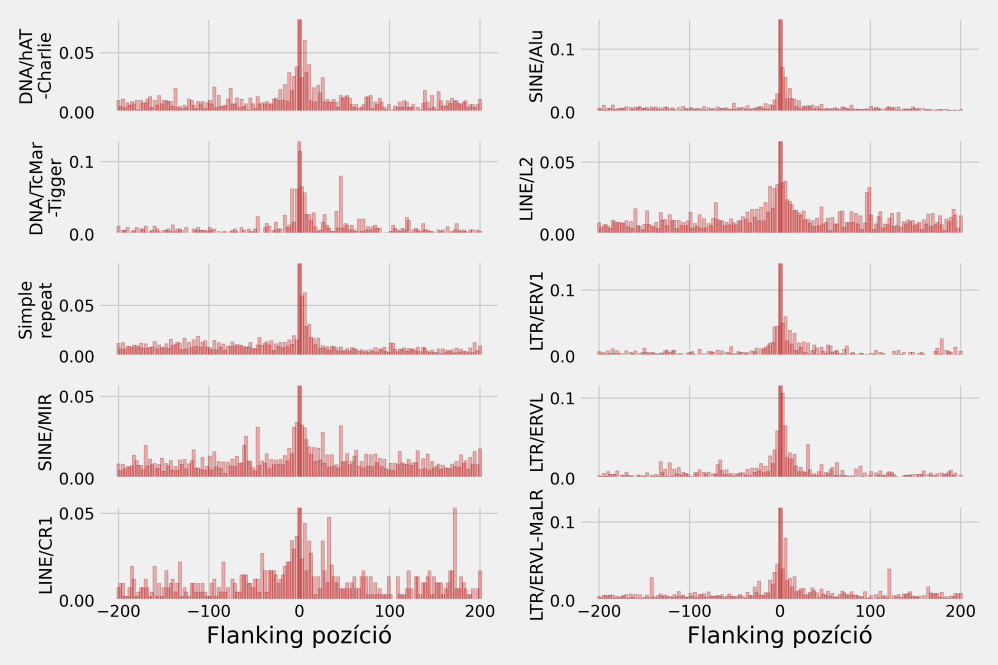
denevér (*Myotis brandtii*), *Myotis davidii*, Kis Barna Denevér (*Myotis lucifugus*), Közönséges Denevér (*Myotis myotis*), *Neophocaena asiaeorientalis*, Fehérarcú Bóbitásgibbon (*Nomascus leucogenys*), Plateau Pika (*Ochotona curzoniae*), Amerikai Pocoknyúl (*Ochotona princeps*), Fehérfarkú Szarvas (*Odocoileus virginianus*), Kardszárnyú Delfin (*Orcinus orca*), Kacsacsőrű Emlős (*Ornithorhynchus anatinus*), Földimalac (*Orycteropus afer*), Üreginyúl (*Oryctolagus cuniculus*), Házijuh (*Ovis aries*), Bonobó (*Pan paniscus*), Közönséges Csimpánz (*Pan troglodytes*), Oroszlán (*Panthera leo*), Leopárd (*Panthera pardus*), Tigris (*Panthera tigris*), Tibeti Antilop (*Pantholops hodgsonii*), Anubisz-pávián (*Papio anubis*), Fehérlábú Egér (*Peromyscus leucopus*), Őzegér (*Peromyscus maniculatus*), Szavannai Varacskosdisznó (*Phacochoerus africanus*), Koala (*Phascolarctos cinereus*), Borjúfóka (*Phoca vitulina*), Nagy Ámbráscet (*Physeter catodon*), Szumátrai Orángután (*Pongo abelii*), Leopárdmacska (*Prionailurus bengalensis*), Halászmacska (*Prionailurus viverrinus*), *Propithecus coquereli*, Fekete Repülőkutya (*Pteropus alecto*), Óriás repülőkutya (*Pteropus vampyrus*), Hegyi Oroszlán (*Puma concolor*), Jaguarundi (*Puma yagouaroundi*), Vándorpatkány (*Rattus norvegicus*), Házipatkány (*Rattus rattus*), Jünnani Piszeorrú Majom (*Rhinopithecus bieti*), Arany Piszeorrú Majom (*Rhinopithecus roxellana*), Nílusi Repülőkutya (*Rousettus aegyptiacus*), Bolíviai Mókusmajom (*Saimiri boliviensis*), Erszényes Ördög (*Sarcophilus harrisii*), Erdei Cickány (*Sorex araneus*), Szurikáta (*Suricata suricatta*), Vaddisznó (*Sus scrofa*), Rövidszőrű Hangyászsün (*Tachyglossus aculeatus*), Ibériai Vakond (*Talpa occidentalis*), Dzseládapávián (*Theropithecus gelada*), *Trachypithecus francoisi*, Karibi Manátusz (*Trichechus manatus*), Közönséges Rókakuzu (*Trichosurus vulpecula*), Palackorrú Delfin (*Tursiops truncatus*), Sarki Ürge (*Urocitellus parryii*), Fekete Medve (*Ursus americanus*), Barna Medve (*Ursus arctos*), Jegesmedve (*Ursus maritimus*), Alpaka (*Vicugna pacos*), Vombat (*Vombatus ursinus*), Sarki Róka (*Vulpes lagopus*), Vörös Róka (*Vulpes vulpes*), Kaliforniai Oroszlánfóka (*Zalophus californianus*)

11.2.2. M2. Különböző taxonómiai szinteknek megfelelő UMAP centrumok



18. ábra. Különböző taxonómiai szinteknek megfelelő UMAP centrumok

11.2.3. M3. Ismétlődő elemek gyakoriságának változása a NUMTok környezetében

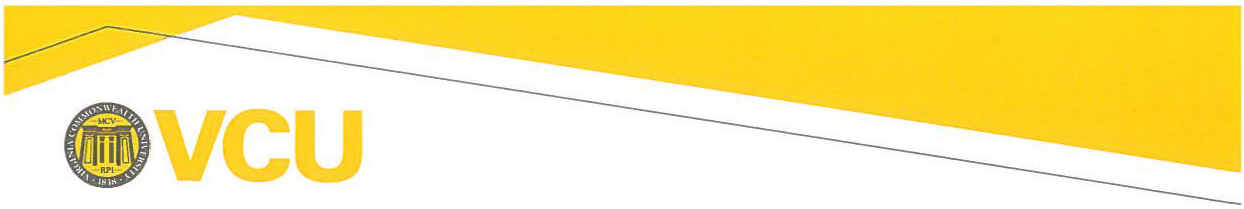


19. ábra. Ismétlődő elemek gyakoriságának változása a NUMTok környezetében. Az ’y’ tengelyen minden esetben az adott ismétlődő elem gyakorisága látható.

11.2.4. M4. Az emlős NUMTok taxonómiai rend szintű összefoglalása

Rend	Családok száma	Nemzetségek száma	Fajok száma	NUMTok száma	NUMTok átlagos hossza (bp)
Carnivora	8	23	32	16650	418
Primates	9	19	23	17131	948
Chiroptera	5	7	11	3640	342
Artiodactyla	10	20	25	13389	913
Rodentia	10	21	24	6963	559
Pilosa	1	1	1	278	494
Eulipotyphla	3	4	4	1381	250
Cingulata	1	1	1	251	508
Microbiotheria	1	1	1	194	504
Macroscelidea	1	1	1	1808	240
Perissodactyla	1	1	2	1095	330
Dermoptera	1	1	1	888	861
Didelphimorphia	1	2	2	2006	724
Proboscidea	1	2	2	1614	558
Pholidota	1	1	2	803	1070
Lagomorpha	2	2	3	785	465
Monotremata	2	2	2	641	238
Tubulidentata	1	1	1	490	528
Diprotodontia	3	3	3	552	733
Dasyuromorphia	1	1	1	3888	492

11.2.5. M5. Külső intézmény hozzájáruló nyilatkozata



Virginia Commonwealth University
Department of Computer Science

College of Engineering, East Hall, 4th Floor, Room E4225
401 West Main St.
Box 843019
Richmond, Virginia 23284-3019

804 828-0575 • Fax: 804 828-2771
TDD: 1-800-828-1120
kcios@vcu.edu
computer-science.egr.vcu.edu

May 9, 2022

To whom it may concern,

I, Dr. Lukasz Kurgan, who served as a temporary supervisor and mentor of Bálint Biró during his stay at the Structural Bioinformatics Laboratory in the Department of Computer Science at the Virginia Commonwealth University (Richmond, VA, USA) hereby confirm that Bálint can include his work that he performed during his visit as a part of his Ph.D. dissertation.

His research in my lab concerned annotation of human protein sequences with experimental data derived from the PDB data using the DSSP program and the BioLip database. Moreover, using these data as the ground truth, Bálint performed novel assessments and analyses of results produced by several popular machine learning based predictors of intrinsic disorder (flDPnn), secondary structure (PSIPRED), solvent accessibility (ASAquick) and nucleic acid binding (DRNApred and DisoRDPbind).

These results are included as a part of the paper entitled “Complementarity of the residue-level protein function and structure predictions in human proteins” which was recently published in the *Computational and Structural Biotechnology Journal* (https://doi.org/10.1016/j.csbj.2022.05.003). Balint was the co-lead on this project, did extremely well, and deservingly he is the first co-author on this article.

With best regards,

Lukasz Kurgan, Ph.D., AIMBE and AAIA Fellow
Robert J. Mattauch Endowed Professor
Department of Computer Science, Virginia Commonwealth University
Phone 804-827-3986; Email: lkurgan@vcu.edu; http://biomine.cs.vcu.edu/

12. Fontosabb Tudományos Publikációk

12.1. Az értekezés témájában megjelent impakt faktorral rendelkező tudományos cikkek

- Biró, B., Zhao, B. and Kurgan, L. (2022). Complementarity of the residue-level protein function and structure predictions in human proteins. Computational and structural biotechnology journal, 20, 2223-2234. D1 Biofizika, IF: 7.271

- Biró, B., Gál, Z., Schiavo, G., Ribari, A., Utzeri, V. J., Brookman, M., ... and Hoffmann, O. I. (2022). Nuclear mitochondrial DNA sequences in the rabbit genome. Mitochondrion, 66, 1-6. Q2 Sejtbiológia, IF: 4.35

12.2. Az értekezés témájában megjelent impakt faktorral nem rendelkező tudományos cikkek

- Biró, B., Gál, Z., Brookman, M. and Hoffmann, O. I. (2022). Patterns of NUMTogenesis in sixteen different mice strains. bioRxiv.

13. Köszönetnyilvánítás

Köszönöm Dr. Hoffmann Orsolya Ivett témavezetőmnek a szakmai segítségét, meglátásait és a tudományos munkám véleményezését. Nagyon hálás vagyok neki, hogy megengedte az időközi témaváltást, de legfőképpen azt, hogy személyesen és szakmailag is támogatott az amerikai tanulmányutam kapcsán. A tudományos munkában mindig partner volt és sok esetben olyan új irányokat mutatott, amik a későbbiekben fontosnak bizonyultak. A kutatói attitűdöm bizonyos részleteit mindenképpen neki köszönhetem.

Nagy köszönettel tartozom másik témavezetőmnek Dr. Lukasz Kurgannak is főként azért, hogy tapasztalatlanságom ellenére fogadott a nagy presztizsű tudományos műhelyében. Az általa szervezett, sokszor nehezen abszolválható heti beszámolókat a szakmai fejlődés legfőbb mozgatórugójának tartom. Nem túlzó kijelentés, hogy gondolkodásmódja és morális intelmei befolyásolták a világnézetemet.

Köszönöm a richmondi labortársaknak, hogy biztosították a megfelelő munkaköri légkört és azt is, hogy bármilyen triviális kérdéssel fordulhattam hozzájuk. Közülük szeretném kiemelni Dr. Bi Zhao-t és Dr. Akila Imesha Katuwawala-t, akik nagyban hozzájárultak kintlétem eredményességéhez. Programozási és bioinformatikai alapjaim nélkülük nem léteznének.

Hálás vagyok a MATE GBI Állatbiotechnológia labortársaimnak is a segítségért és a megfelelő légkörért. Közülük is kiemelendő Gál Zoltán, aki kutatói pályám elején sokat segített a molekuláris és embriológiai módszerek megfelelő elsajátításában.

Köszönöm Dr. Hiripi Lászlónak, hogy bármikor kérdezhettem tőle szakmai vonatkozású dolgokat.

Nagyon köszönöm Dr. Szalai Bencének, hogy megismertetett a rendszerszemlélet és a vektorizált műveletek fontosságával. Ugyan rövid ideig dolgoztunk együtt, de sokat tanultam tőle.

Hálával tartozom Dr. Pitlik Lászlónak is, aki filozófiai eszmecserékkel segített a megfelelő gondolkodás kialakításában.

Nagyon hálás vagyok a Feleségemnek, hogy végig támogatott a doktori tanulmányaim során és már puszta jelenlétével enyhítette az érzelmi kilengéseket, amik törvényszerűek a tudományos munka jellegéből adódóan. Nélküle sem jöhetett volna létre a dolgozatom jelen formában.

Köszönöm a szüleimnek és a tágabb családomnak, hogy neveltetésemmel elültették bennem a teremtett világ részleteinek megismerése utáni vágyat.

Köszönöm bátyámnak a tudományfilozófiai adalékokat.

Szinte biztos vagyok benne, hogy jó néhány személyt kihagytam, akik nélkül a doktori kutatásom nem jöhetett volna létre. Természetesen nekik is hálával tartozom.