MAGYAR AGRÁR- ÉS
ÉLETTUDOMÁNYI EGYETEM

Mitokondriális genomikai és proteomikai vizsgálatok fejlesztése

Doktori (PhD) értekezés tézisei

Biró Bálint

DOI: 10.54598/003430

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. A munka előzményei, célkitűzések

1.1. A munka előzményei

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. Ezek az újszerű eljárások, például a szekvenálás, röntgen krisztallográfia, NMR spektroszkópia, 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 és a fizikailag kis áteresztőképességük miatt széles körű elterjedésük váratott magára.

A tudomány az ezredforduló környékén érte el a technológiának azt a fejlettségi fokát, ami lehetővé tette a szekvenálás automatizációját (Pauwels et al., 1995). Ennek 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, paradigmaváltást eredményezve (D'Argenio, 2018; Pal et al., 2020). 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. 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. 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.

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 vizsgáltuk 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 leg-

elterjedtebb 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.

1.2. Célkitűzések

1.2.1. NUMT biológia

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.

1.2.2. Proteomika

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.

2. Anyag és Módszer

2.1. Az általános statisztika és a gépi tanulás módszertana

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).

2.2. NUMT biológia

A nyúl genom (OryCun2.0) kromoszómális és mitokondrium szekvenciáját az Ensembl genom adatbázisból szereztük be. Az elérhető emlős genomokat átfogó vizsgálatokhoz a nukleáris és a mitokondriális 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 és a rangokat is az NCBI adatbázisból nyertük ki.

A nukleáris és a mitokondriális genomok illesztéséhez 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).

A NUMTok és a véletlenszerű szekvenciák klasszifikációjára RBF-Kernel SVM-et tanítottunk. Az SVM modell teljesítményét k-szoros keresztvalidációs (k=3) eljárással és a tévesztési mátrixból származtatott mutatókkal is teszteltük. Az adatszivárgás elkerülése érdekében a keresztvalidációs eljárás minden egyes iterációjában külön normalizáltuk a bementeket a minimum-maximum normalizációs eljárásnak megfelelően.

A NUMTok és a megfelelő mtDNS szekvenciák genetikai távolságát a hiányzó nukleotidokat is elfogadó módosított Kimura2 paraméterrel becsültük.

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).

2.3. Proteomika

A komplett humán proteómot a UniProt adatbázisból nyertük ki ("UniProt: the universal protein knowledgebase in 2021", 2021).

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., 2013). 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 (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 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) és a fals pozitív (FPR) arányokat számítottuk ki.

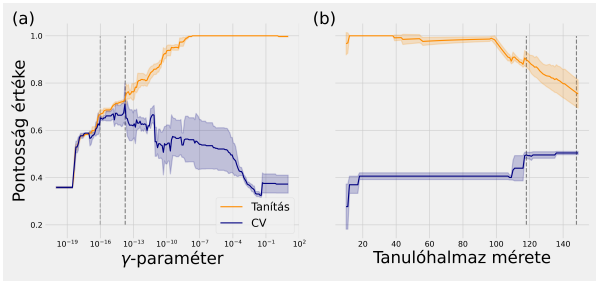
A másodlagos szerkezet predikciójának pontosságát a valós és predikált szerkezeti elemeket tartalmazó szegmensek összevetésével vizsgáltuk. Ezt az összefüggő szegmensek értékének mérőszáma (Segment OVerlap score - SOV) tette lehetővé.

3. Eredmények és azok megbeszélése

3.1. NUMT biológia

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) (1.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 képes volt eldönteni, hogy az adott szekvencia NUMT-e és azt is, hogy szkaffoldról vagy éppen kromoszómáról származik-e (Egyenlet. 3.1). 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 (1.ábra/a). A rendelkezésre álló adatok az SVM tanításához elegendőnek bizonyultak (1.ábra/b).

$$CM = \begin{bmatrix} 21 & 0 & 0 & 7 \\ 0 & 13 & 2 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 26 \end{bmatrix} \qquad \text{(Egyenlet. 3.1)}$$



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

Az általunk vizsgált emlős NUMTok jellegzetességei alapján jól elkülöníthetők a filogenetikai rendek. UMAP dimenzió csökkentés használata esetén a rendekre jellemző centrumok is kirajzolódtak (2. ábra/a).

Az emlős NUMTok adott faj mitokondriumjának hosszához viszonyított

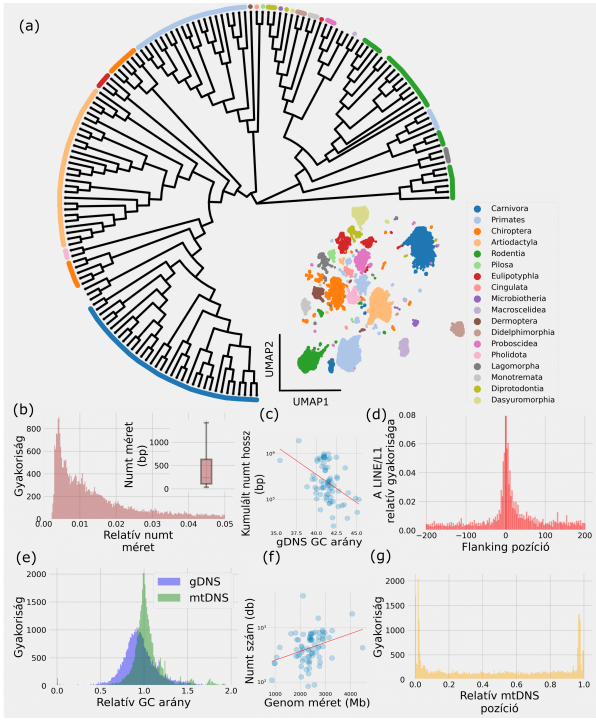
relatív mérete torzított eloszlást mutatott, jellemző a kisebb méretű NUMTok felülreprezentáltsága (2. ábra/b). Ez a tendencia szignifikáns ($p<0.05$). Az abszolút NUMT hosszak teljes interkvartilis terjedelme 600 bp alatti volt, míg mediánja jóval 250 bp alatt helyezkedett el (2. ábra/b). Három faj 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 ($p<0.05$) (2. 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 (2. 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 (2. ábra/d). 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.

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 (2. ábra/c,f).



2. 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-ének hosszához 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).

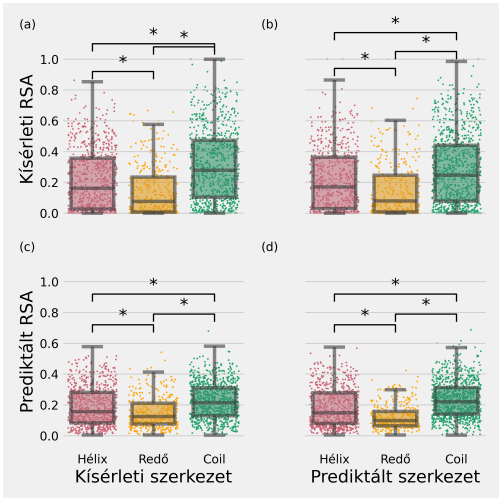
3.2. Proteomika

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 (3. ábra).

Ez a tendencia annak ellenére is megmarad, hogy a kísérleti úton meghatározott SA értékek nagyobb skálán mozogtak (3.ábra/a-b), mint a prediktált

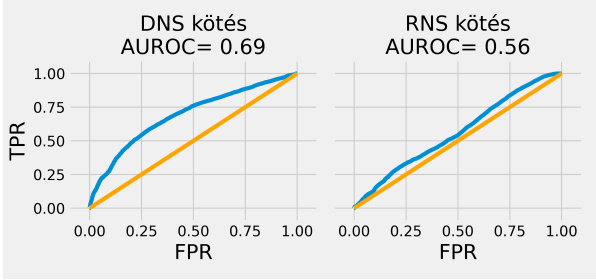
SA értékek (3.ábra/c-d). A predikció esetén az SA értékek eloszlása sokkal kiegyenlítettebb, mint a kísérleti úton meghatározott SA adatok esetén (3. ábra).

A másodlagos szerkezet predikciójának pontosságát az SOV érték 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.



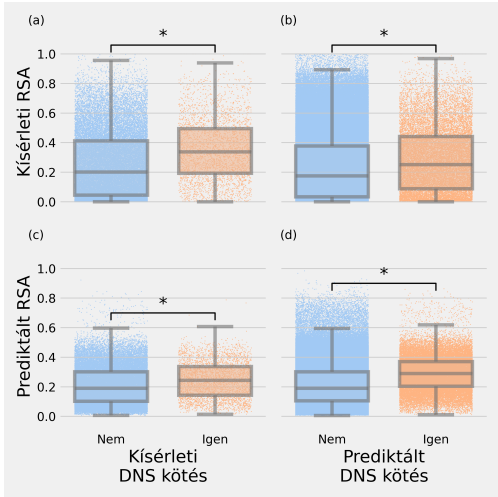
3. ábra. Az SA és a másodlagos szerkezet predikciójának és valós értékeinek összefüggései.
A H a hélix, az E a β-redő, míg a C a coil másodlagos szerkezeti elemre utalnak. A felső sor esetén (a,b) az RSA értékei kísérleti adatokból, míg az alsó sor esetén (c,d) predikcióból származnak. Az első oszlop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

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 (4. ábra).



4. 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 (5. ábra). Ezek az eltérések szignifikánsak ($p<10^{-5}$) és szignifikáns jellegük az SA és a DNS kötés predikciójának és kísérleti adatának tetszőleges kombinációiban is megmarad. RNS kötés esetén is hasonló eredményeket kaptunk, azonban a tézisfüzetben kizárólag a DNS kötés és az SA összefüggései kerülnek bemutatásra.



5. 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.

4. Következtetések és javaslatok

4.1. NUMT biológia

A doktori kutatás során beállítottunk egy ”NUMT bányászati” algoritmust a nyúl genomon. Ezzel a módszerrel leírtuk a nyúl genomba integrálódott NUMTok különböző jellegzetességeit, amit a tézisfüzetben nem mutatunk be. Ezt az algoritmust terjesztettük ki az elérhető emlős genomokra. 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 (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).

4.2. 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 maga-

sabb 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 AUC=0.69, RNS AUC=0.56) (4. ábra) bizonyos mértékben elmarad az irodalmi adatoktól (DNS AUC=0.77, RNS AUC=0.67) (Yan & Kurgan, 2017). Ez az eltérés nagy valószínűséggel az általunk használt adathalmaz komplexitásából adódik.

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 (4. 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).

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

5. 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

6. Az értekezés témájában megjelent cikkek

6.1. 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

6.2. 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.

Hivatkozások

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.

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.

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.

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.

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.

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

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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

53, 102497.

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*.

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.

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

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.

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.

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*.

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.

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.

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

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.

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

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).

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.

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.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). 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.

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.

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

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.

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.

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.

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

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.

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.