A rendezetlen fehérjék evolúciós vizsgálatai

Doktori (PhD.) értekezés

Pajkos Mátyás

Témavezető:

Dr. Dosztányi Zsuzsanna

Ph.D., tudományos főmunkatárs

Biológia Doktori Iskola, Szerkezeti biokémia program

Programvezető:

Dr. Kovács Mihály

PhD., habil., az MTA doktora
tansz. vezető egyetemi tanár

Doktori Iskolavezető:

Prof. Dr. Erdei Anna

D.Sc.,egyetemi tanár
MTA rendes tagja

Biológiai Intézet, Biokémia Tanszék, MTA-ELTE „Bioinformatika” Kutatócsoport

Eötvös Loránd Tudományegyetem, Természettudományi Kar, Budapest

2020

Köszönetnyilvánítás

Elsőként szeretném megköszönni témavezetőmnek, **Dr. Dosztányi Zsuzsannának**, hogy lehetőséget adott a bioinformatika világának megismerésére és hogy mindig kitartóan támogatott és ösztönzött munkámban a kezdetektől máig.

Továbbá köszönettel tartozom **Dr. Nyitray Lászlónak** és **Dr. Kovács Mihálynak**, hogy a biokémia tanszéken megteremtették számomra a munkámhoz szükséges körülményeket. Köszönettel tartozom továbbá **Dr. Simon Istvánnak**, hogy csoportjának tagjaként kezdhettem meg tudományos pályafutásom.

Hálával tartozom kutatócsoportunk tagjainak, **Erdős Gábor**nak a szakmai beszélgetésekért és tanácsokért, valmint **Szaniszló Tamás**nak **Fülöp Máté**nak és **Tóth Benedek**nek, hogy szakmai támogatásukkal segítették kutatómunkám.

Köszönöm édesanyámnak, édesapámnak és testvéremnek támogatásukat, türelmüket és azt, hogy mindvégig mellettem álltak, hittek bennem és munkámban támogattak.

Végül – de nem utolsó sorban – köszönöm barátaimnak, hogy mindig számíthattam ösztönző támogatásukra munkám során.

Tartalomjegyzék

Publikációs lista

A TÉZISEK ALAPJÁUL SZOLGÁLÓ KÖZLEMÉNYEK:

1. **Pajkos, M** ; Zeke, A ; Dosztanyi, Z
   Ancient evolutionary origin of intrinsically disordered cancer risk regions
   BIOMOLECULES
   „minor revision”
   (doi: 10.1101/2020.06.15.152298)

2. Erdős, G ; Szaniszló, T ; **Pajkos, M** ; Hajdu-Soltész, B ; Kiss, B ; Pál, G ; Nyitray, L ; Dosztányi, Z
   Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway
   PLOS COMPUTATIONAL BIOLOGY 13 : 12 Paper: e1005885 , 25 p. (2017)

A DOKTORI DOLGOZATHOZ NEM KAPCSOLÓDÓ PUBLIKÁCIÓK:

1. Hatos, András ; Hajdu-Soltész, Borbála ; Monzon, Alexander M ; Palopoli, Nicolas ; Álvarez, Lucía ; Aykac-Fas, Burcu ; Bassot, Claudio ; Benítez, Guillermo I ; Bevilacqua, Martina ; Chasapi, Anastasia et al.
   DisProt: intrinsic protein disorder annotation in 2020
   NUCLEIC ACIDS RESEARCH 48 : D1 pp. D269-D276. (2020)

2. **Pajkos, M** ; Mészáros, B ; Simon, I ; Dosztányi, Z
   Is there a biological cost of protein disorder? Analysis of cancer-associated mutations
   MOLECULAR BIOSYSTEMS 8 : 1 pp. 296-307. , 12 p. (2012)

Bevezetés

Az élő szervezetek sejtjeiben szinte minden biokémiai folyamatban részt vesz valamilyen fehérje molekula. A fehérjék funkcionalitásukat tekintve nagyon változatosak, az egyik legrégebben ismert funkciójukat enzimként látják el, amikor is valamilyen biokémiai folyamatot katalizálnak. Tudjuk, hogy ennek a feladatnak az ellátásához elengedhetetlen, hogy a fehérje térszerkezettel rendelkezzen, mely megfigyelés a szerkezet-funkció paradigma néven vált köztudottá. Habár az enzimekhez hasonlóan, térszerkezettel rendelkező fehérjék a sejtes folyamatok jelentős esetében jelen vannak, ma már tudjuk, hogy léteznek olyan fehérjék is, melyekre nem igaz a szerkezet-funkció paradigma, mivel nem rendelkeznek térszerkezettel fiziológiás körülmények között, ennek ellenére funkcionálisak. Ezeket a fehérjéket nevezzük eredendően rendezetlen fehérjéknek (Intrinsically Disordered Proteins - IDPs).

A rendezetlen fehérjék a XX. század végén kerültek felfedezésre és azóta a róluk megszerzett tudásunk folyamatosan bővül. Mára tudjuk, hogy a sejtes folyamatokban nem csak, hogy jelen vannak, hanem a sejt alapvető működésében látnak el fontos szerepeket. Ebből kifolyólag azt is tudjuk, hogy sok betegség (pl. neurodegeneratív vagy rákos megbetegedések) kialakulása összefüggésbe hozható a rendezetlen fehérjékkel.

Habár a rendezetlen fehérjék jelentősége orvosbiológiai szempontból is hangsúlyos, mégis viszonylag keveset tudunk róluk, melynek fő oka, hogy laboratóriumi körülmények között speciális tulajdonságaik miatt a jelenlegi eszközökkel nehezen vizsgálhatóak. Ezért a rendezetlen fehérjék vizsgálatára egy hatékony megközelítést jelent a bioinformatika. Ezt felismerve, mára számos, különböző stratégiájú bioinformatikai vizsgálat irányul a rendezetlen fehérjékre, azonban kevésbé tanulmányozottak evolúciós szempontból. Doktori munkám központjában a rendezetlen fehérjék evolúciós vizsgálatai álltak, mely vizsgálatok révén megismerve molekuláris

történelmüket kerülünk közelebb működésük pontosabb feltérképezéséhez és megértéséhez.

Irodalmi áttekintés

Értekezésem központjában a rendezetlen fehérjék funkcionális régióira irányuló bioinformatikai vizsgálataim eredményei állnak. Ehhez kapcsolódóan, az irodalmi áttekintésben bemutatásra kerülnek a rendezetlen fehérjék és evolúciójuk, funkcionalitásuk és az ebből eredő biológiai jelentőségük, valamint nem utolsó sorban a rákos megbetegedésekben betöltött szerepük. Továbbá bemutatok egy, a rendezetlen fehérjék speciális kölcsönhatásán alapuló rendszert, az LC8 dinein könnyűlánc fehérjét és kölcsönható partnereit. Kutatásaim eszköze elsősorban a molekuláris evolúció. Ennek megfelelően továbbá, az irodalmi áttekintés során képet adok a rendezetlen fehérjék evolúciójáról, az evolúció molekula szintű alkalmazhatóságáról minden olyan vonatkozásban, melyeket kutatási eredményeim érintenek.

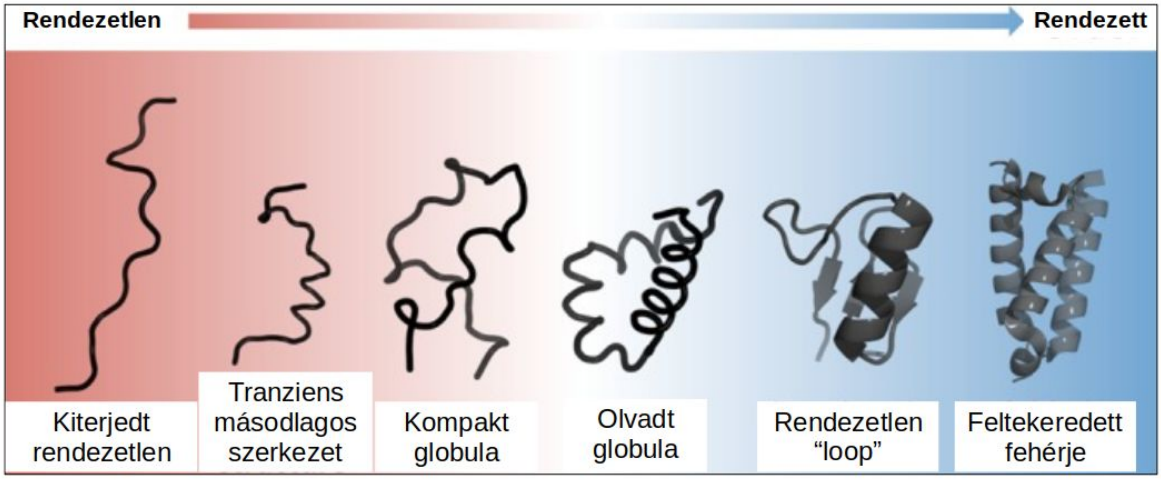
A rendezetlen fehérjék és funkcióik

A fehérjék által ellátott feladatok egy része szorosan kapcsolódik a térszerkezetükhöz – sőt bizonyos funkciók csak így láthatóak el, mint pl. az enzimatikus folyamatok. Ez a megfigyelés vezetett a szerkezet-funkció paradigma megalkotásához, mely szerint egy fehérje funkciójának ellátásához elengedhetetlen egy harmadlagos térszerkezet. Azonban egyre több kísérleti bizonyíték mutatott rá arra, hogy ez az elképzelés túlságosan leegyszerűsített. Egyrészt már régóta ismert tény, hogy bizonyos fehérjeszakaszok, illetve kisebb hormon peptidek (pl. a glükagon) nem képesek önmagukban rendeződni, így pl. röntgenkrisztallográfiás mérés során (amely a leggyakrabban használt módszer a térszerkezet meghatározására) nem adnak koherens szórási képet. Sok esetben eredménytelenül próbálkoztak egy adott fehérje kristályosításával, illetve annak szerkezetének meghatározásával. Az is ismert volt, hogy bizonyos fehérjék, vagy fehérje részletek aminosav összetétele annyira speciális,

hogy azokban nem tudnak rendezett másodlagos szerkezeti elemek kialakulni, ezek általában alacsony szekvenciális komplexitású részek [1,2].

Azonban sokáig váratott magára az a felismerés [3], hogy a szerkezet-funkció összefüggést újra kellene gondolni, mivel fiziológiás körülmények között is léteznek olyan harmadlagos szerkezet nélküli fehérjék, amelyek képesek egy adott funkciót ellátni. Valójában ma már tudjuk, hogy az élőlények proteomjai (még az olyan egyszerűeké is, mint az egysejtűek) jelentős részben tartalmaznak olyan funkcionális polipeptid szegmenseket, melyek nem képesek a jól meghatározott 3-dimenziós szerkezet kialakítására [4–7]. Ezeket a régiókat nevezzük rendezetlen régióknak (intrinsically disordered region - IDR), mely régiók különböző mértékű jelenléte definiálja a rendezetlen fehérjét (intrinsically disordered protein - IDP).

A rendezetlen fehérje régiók nem írhatóak le egy adott, jól definiált konformációval, ehelyett egy folyamatosan változó, szerkezeti sokasággal tudjuk ezeket jellemezni [7–11]. A rendezetlen fehérjék szerkezeti tulajdonságainak részletes vizsgálata rámutatott, hogy a rendezetlenség mértéke nagyon változatos. Csoportosításuk ezért szerkezeti szempontból egy folytonos spektrumon történik [12], amely a teljesen rendezetlen állapottól a jól meghatározott térszerkezetű állapotig tart. A spektrum magában foglal feltekeredett doménekből álló szerkezetet ahol egyáltalán nincs rendezetlenség, esetleg csak lokális “loop”-okként; rendezetlen régiók által összekötött multi-domén szerkezetet; olvadt globula (“molten globule”) és kompakt globula (“compact globule”) szerkezetet; letekeredett állapotot tranziens, másodlagos lokális szerkezeti elemekkel és kiterjedt, teljesen rendezetlen állapotot (1. ábra). A modellben nincsenek határok a leírt állapotok között, egy natív fehérje bárhol elhelyezhető a folytonos spektrumon [7,12].



1. ábra – A fehérjeszerkezet folytonos spektrumának sematikus ábrázolása

A szín gradiens jelzi a konformációs állapotok kontinuumát, kiindulva a dinamikus, kiterjedt rendezetlen állapottól (piros) egészen a teljesen feltekeredett, rendezett állapotig (kék) [7].

A rendezetlen fehérjéket (és rendezetlen régióikat) a DisProt adatbázis gyűjti. Az adatbázisban olyan fehérjéket annotáltak az irodalomból, melyek esetében kísérletes bizonyíték van rá, hogy legalább egy, 10 aminosav hosszú rendezetlen régiót tartalmaznak. Az adatbázis első verziója A. Keith Dunker nevéhez fűződik, amit 2007-ben tettek közzé [13], majd viszonylag sok idő elteltével csak az elmúlt pár évben kapott újra figyelmet. A legutóbbi frissítés 2019-ben történt (DisProt 8), mely során több mint 50 annotátor 862 fehérje 1972 rendezetlen régióját kurálta, így jelenleg összesen 1556 fehérje 3511 régiója található meg az adatbázisban [13,14]. Ebben a munkában én is részt vehettem 45 rendezetlen fehérje régió annotációjával.

A kísérletesen jellemzett rendezetlen régiók szekvenciális vizsgálata rámutatott ezen régiók speciális tulajdonságaira. Legjellemzőbb tulajdonságuk, hogy bennük a hidrofób aminosavak alulreprezentáltak, és gyakran sok azonos töltésű aminosavat is tartalmazhatnak [4,12,15]. A rendezett és rendezetlen fehérjék közötti alapvető különbségek lehetővé teszik ezen két fehérje csoport szekvenciális azonosítását. Jelenleg több mint 50 különböző fehérje rendezetlenség predikciós módszert publikáltak

már [16,17]. Ezek közül az egyik legismertebb módszer a csoportunk által kidolgozott IUPred módszer [18]. Az IUPred azon a megközelítésen alapszik, hogy a rendezetlen fehérjék aminosavai nem képesek egy alacsony energiájú szerkezet kialakítására. A módszer a statisztikus potenciálok segítségével becsli meg az adott pozícióhoz tartozó kölcsönhatási energiát, amit az adott aminosav a szekvenciális környezetével kialakít. A becsült energia érték alapján minden pozícióhoz egy valószínűség jellegű érték rendelhető, amely annak rendezetlenségét jellemzi.

A teljes genomokon elvégzett vizsgálatok azt mutatják, hogy evolúciós értelemben minél magasabb szintű élőlényt nézünk, annál magasabb a rendezetlen fehérjék aránya [4,5]. Ezzel a megfigyeléssel összhangban van annak a vizsgálatnak az eredménye is, amely során *E.coli* és *S. cerevisiae* genomját vizsgálták meg a rendezetlenség mértékének szempontjából. Ebben a vizsgálatban azt találták, hogy a fejlettebb élesztő nagyobb arányban tartalmaz rendezetlen fehérjéket, mint a nála egyszerűbb baktérium [6]. Emellet, más, hasonló munkák eredményei is azt mutatják, hogy az eukarióta genomok jelentős részben kódolnak rendezetlen fehérjéket. Ezek az eredmények azt mutatják, hogy az élőlények komplexitásával nő a rendezetlenség, ami pedig arra enged következtetni, hogy a rendezetlenség előnyt jelent a komplexebb szabályozási folyamatokban [19].

A fehérje rendezetlenség felfedezésével szükségessé vált a szerkezet és funkció kapcsolat újragondolása [3]. Ennek megfelelően, a klasszikus szerkezet-funkció paradigma mellett a rendezetlenség-funkció paradigma is helyet kapott [7].

A rendezetlen fehérjék funkciói sokfélék lehetnek, ráadásul egy fehérjén belül a különböző rendezetlen régiók más-más funkciókért is felelhetnek [20]. Funkcionális csoportosításukra jelenleg több megközelítés is létezik, de az alapokat A. Keith Dunker és Tompa Péter már 2002-ben lefektették [21–23]. Kezdetben 28 funkcionális csoportot definiáltak, melyeket 5 nagyobb osztályba soroltak, de azóta újabb funkciók kerültek

felfedezésre. A legújabb eredmények szerint a rendezetlen fehérjék biológiai kondenzátumok létrejöttében is szerepet játszanak, fázis szeparáció révén [21–23]. A DisProt adatbázis jelenleg 7 osztályt alkalmaz a rendezetlen fehérjék funkcionális klasszifikációjára [14].

A rendezetlen fehérjék egyik fontos funkcionális csoportját az úgynevezett **entrópikus láncok** alkotják. Ide tartoznak többek között a fehérjedoméneket összekötő és azok egymáshoz képesti elmozdulását biztosító linker régiók vagy az entrópikus rugók, melyek entrópiája összenyomáskor csökken, így erőt fejtenek ki az összenyomás ellenében és ezzel tartanak távol egymástól két fehérje elemet.

A rendezetlen fehérjék másik fő funkcionális csoportja molekuláris felismerésen alapul. Ezen belül, a rendezetlen fehérjék lehetnek **összeszerelők**, amelyek egynél több partnert kötve egy funkcionális komplexet szerelnek össze, vagy **effektorok**, melyek egy másik fehérje aktivitását változtatják meg, ilyenek pl. a rendezetlen inhibitorok és aktivátorok, a komplex szétszerelők vagy a DNS meghajlításáért felelős fehérjék. A rendezetlen fehérjék rendelkezhetnek raktározó funkcióval, melynek során pl. fémionokat tárolnak/szállítanak, víz molekulákat tartanak vissza a sejtekben vagy toxikus szerves molekulákat semlegesítenek. A molekuláris felismerésen alapuló **bemutató helyek** osztályába olyan fehérjék tartoznak, amelyek egyes régiói a poszttranszlációs fehérjemódosítások (Post-translational Modification - PTM) célpontjai. Ilyen PTM pl. a foszforiláció, metiláció, glikoziláció vagy acetiláció. A rendezetlen fehérjék funkcionálhatnak **chaperonok**ként is, ezáltal szerepet játszva a fehérje és/vagy RNS feltekeredés (“folding”) folyamatában. Egy külön funkcionális kategóriát alkotnak azok a **biológiai kondenzátumok** kialakításában részt vevő rendezetlen fehérjék, amelyek fázis átalakuláson mennek keresztül, kialakítva egy kondenzált fázist. Kialakíthatnak pl. egy dinamikus “folyadék cseppet”, vagy gél állapotot, aggregálódhatnak vagy amiloid szálakba rendeződhetnek.

A rendezetlen fehérjéket aszerint is lehet csoportosítani, hogy milyen más molekulákkal lépnek kölcsönhatása. A rendezetlen fehérjék partnerei lehetnek más specifikus fehérjék, de képesek kölcsönhatásba lépni DNS-sel, RNS-sel, vagy kismolekulákkal is. Bár a rendezetlen fehérjék natív körülmények között nem vesznek fel egy jól meghatározott térszerkezetet, többségük specifikus kölcsönhatási partnerhez való kötődés során szerkezeti változáson megy keresztül [3]. A szakirodalom ezt a folyamatot a “disorder-to-order transition” kifejezéssel írja le. Ettől eltérnek az ún. bolyhos (“fuzzy”) fehérje komplexeket alkotó rendezetlen fehérjék, mely fehérjék a komplex kialakítása után, kötött állapotban is részben vagy egészében megőrzik rendezetlenségüket. A komplex ezen “bolyhos” régióit továbbra is a rendezetlen fehérjék tulajdonságai jellemzik [24].

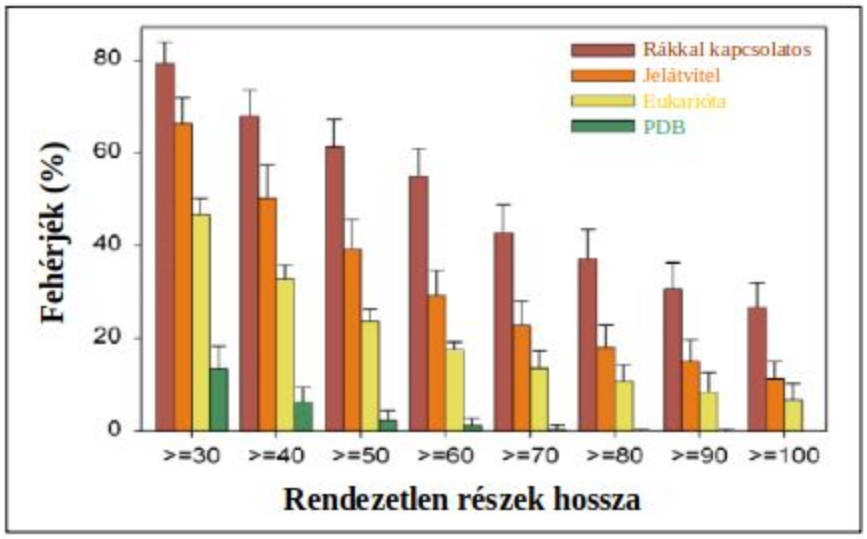
A rendezetlen fehérjék nagyon gyakran tartalmaznak viszonylag kompakt, néhány aminosavból álló funkcionális modulokat, úgynevezett lineáris motívumokat, ugyanakkor léteznek nagyobb, doménszerű tulajdonságokkal rendelkező rendezetlen funkcionális moduluk is [25]. A rendezetlen fehérjéknek mind szerkezeti, mind funkcionális tulajdonságait is befolyásolhatják különböző poszttranszlációs módosítások.

A rendezetlen fehérjék sok partnerrel képesek kölcsönhatásba lépni, ezáltal a fehérje-fehérje kölcsönhatási hálózatokban sok esetben **csomóponti** (hub) szerepet töltenek be [26] (egy fehérje lehet statikus hub, ha nagyszámú partnert képes szimultán kötni különböző kölcsönható régiókon keresztül, vagy dinamikus hub, mely esetben ugyanúgy többféle kötőpartner kötésére képes, de ebben az esetben a partnerek ugyanazon kötőhelyért versengenek). Ez a csomóponti szerep a fehérje-fehérje kölcsönhatási hálózatokban nagy fontossággal bír. Ezt alátámasztja az a megfigyelés is, hogy a csomóponti fehérjék érzékenyebbek a mutációkra, mint a kapcsolódó (nem csomóponti) fehérjék, mivel a mutációkból eredő funkcióváltozás általában letális következményekkel jár. [27,28].

A rendezetlenség és rákos megbetegedések kapcsolata

A rendezetlen fehérjék biológiai fontosságából adódóan, funkcionális megváltozásuk komoly következményekkel járhat a sejt normális működése szempontjából. A rendezetlen fehérjéket több betegséggel is összefüggésbe hozták. Bizonyos neurodegeneratív megbetegedések rendezetlen fehérjékhez köthetőek, mint például az Alzheimer kór, ami a τ-fehérjéhez, vagy a Parkinson kór, ami pedig az α-synucleinhez köthető [29]. Emellett, a rendezetlen fehérjéket sok esetben a rákos megbetegedésekkel is összefüggésbe hozták. Erre példa a BRCA1 tumor szupresszor fehérje, melyet a mellrák kialakulásához kötnek [30], a CBP és MOZ fehérje fúziója leukémiát okozhat [3] vagy az egyik legjobban tanulmányozott eset, a p53, amely esetében a rákos megbetegedések közel 50%-nál mutattak ki mutációk által kiváltott diszfunkcionalitást [31].

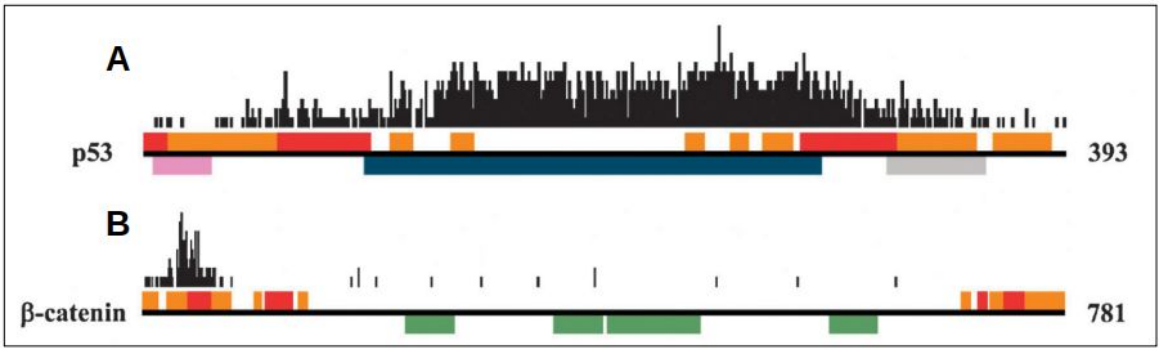
A rendezetlen fehérjék és rákos megbetegedések kapcsolatát a példákon felül egy általánosabb szinten is kimutatták. Egy 2002-es munkában a rákkal összefüggésbe hozható humán fehérjéket szisztematikusan vizsgálták a rendezetlenség szempontjából. A vizsgálat során azt találták, hogy ezekben a fehérjékben jóval nagyobb volt a rendezetlenség, mint az eukarióta fehérjékben általában, de még a jelátvitelben előforduló fehérjékhez képest is szignifikánsan magasabb volt a rendezetlen szakaszokat tartalmazó fehérjék aránya [32] (2. ábra).



2. ábra – **A rendezetlen fehérjék aránya különböző fehérje csoportokban** [32].

Az egyedi példák és átfogóbb vizsgálatok alapján felmerült, hogy a rendezetlen fehérjék esetleg egyféle biológiai kockázattal bírnak a rákos megbetegedések kialakulásának szempontjából. A biológiai kockázat kérdésének tisztázásához a rendezetlen fehérjék és rákos megbetegedések kapcsolatát az egy nukleotidot érintő mutációk szintjén vizsgálta meg kutatócsoportunk 2012-ben [33]. A munka során megvizsgáltam a mutációk eloszlását a fehérjék rendezetlen és rendezett régióin belül. A vizsgálat során 13 adathalmaz több, mint 40 ezer mutációja és közel 86 ezer gyakori polimorfizmus adata került feldolgozásra. Az eredmények azt mutatták, hogy míg a neutrális polimorfizmusok nagyobb valószínűséggel fordulnak elő a fehérjék rendezetlen szakaszaiban, addig a rákos megbetegedésekkel összefüggésbe hozható mutációk a fehérjék rendezett részein figyelhetőek meg nagyobb mértékben. Ezen eredmények alapján az a következtetés vonható le, hogy a fehérje rendezetlenség önmagában nem felelős a rákos betegségekkel kapcsolatos biológiai kockázatért, hanem egy adott fehérjének a funkciója határozza meg egyrészt a rendezetlenség szükségességét, másrészt a rákban betöltött szerepét, ami végső soron ok-okozati kapcsolat nélkül eredményez korrelációt a rendezetlen fehérjék és rákos megbetegedések között [33].

Az általános trendet jól szemlélteti a p53 példája, amely esetében a rendezetlen régiókhoz képest a rendezett DNS kötő doménben figyelhető meg szignifikánsan több mutáció (3A ábra). Habár az általános trend sok rendezetlen fehérjét foglal magában, több példa is megfigyelhető, amik ennek az ellenkezőjét mutatják. Ilyen az N- és C-terminálisán is rendezetlen β-catenin (CTNNB1) (3B ábra), amely egy esszenciális strukturális fehérje bizonyos sejt-adhéziós komplexekben, valamint a Wnt növekedési faktor jelpályában is szerepet tölt be [34].



3. ábra – Domén szerkezet, prediktált kötőhelyek és rendezetlenség, valamint rákos megbetegedésekkel összefüggésbe hozható mutációk pozíciónkénti eloszlásának ábrázolása

(**A**) p53, (**B**) β -catenin. A fekete horizontális vonalak a fehérjék teljes szekvenciáit jelölik, felettük a piros és narancssárga dobozok a prediktált rendezetlenséget és rendezetlen kötőhelyeket [35] (Anyagok és módszerek). A fekete vertikális vonalak a pozíciónkénti mutációk eloszlását jelölik. A fekete vonal alatti színes dobozok a prediktált doméneket [36] ábrázolnak: rózsaszín - transzaktivációs; sötétkék - DNS kötő; szürke - tetramerizációs; zöld - Armadillo ismétlődő.

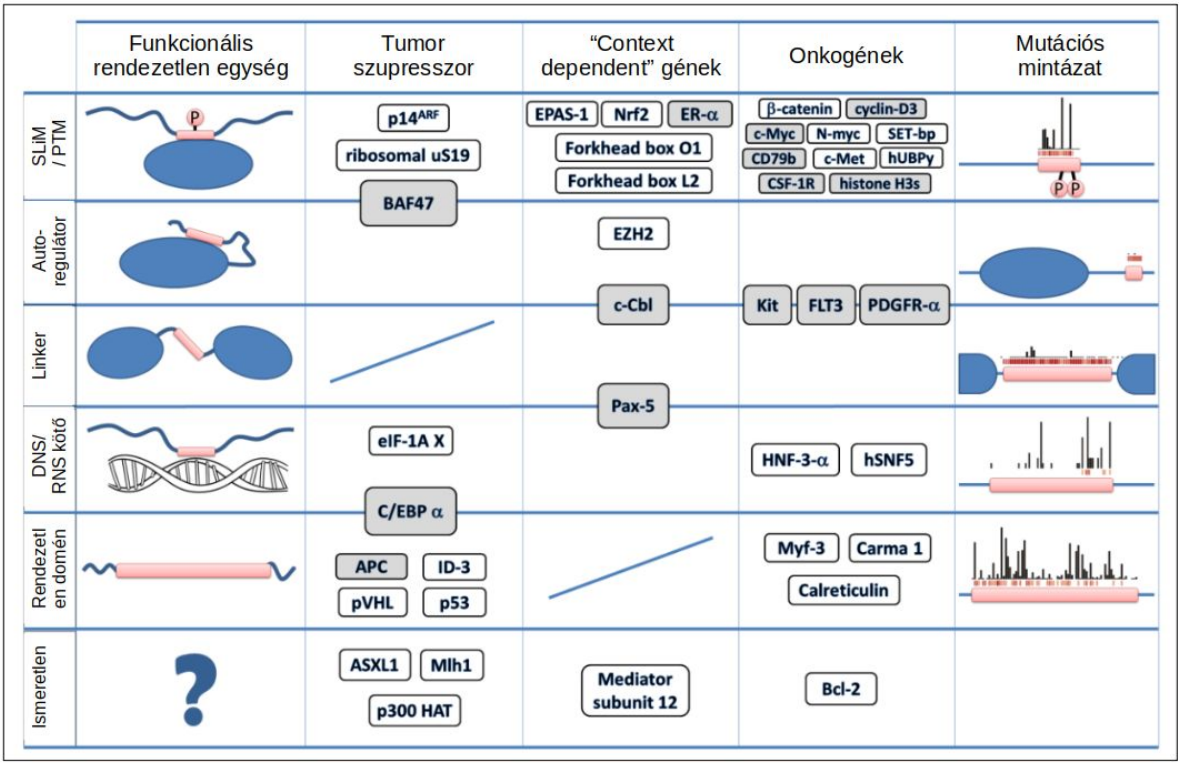
Ebben az esetben minden típusú mutáció a fehérje N-terminális részén, egy rövid, lineáris motívum alapú molekuláris kapcsoló régiójában figyelhető meg. A motívumot az SCF-β-TrCP ubiquitin ligáz ismeri fel, mely felismerés hiánya a motívumot érintő mutációk által, a citoplazmában abnormálisan magas mennyiségben jelenlévő és nukleuszba transzlokálódni képes β-catenint eredményez, a transzlokáció pedig a Wnt

jelpálya elemeinek abnormális aktiválásához, ami pedig túlzott sejtosztódáshoz, végső soron pedig rák kialakulásához vezet [34,37].

Ezek a kezdeti megfigyelések vezettek ahhoz a felismeréshez, hogy általánosságban nagyon keveset tudunk arról, hogy a rendezetlen régiók milyen mechanizmus által vezethetnek rákos megbetegedések kialakulásához. Ezzel a témával foglalkozik csoportunk egy közelmúltban elkészült munkája is [38]. A projekt keretében a rendezetlenség tumorgenezishez való direkt hozzájárulását vizsgálták egyedi rendezetlen régió példákon keresztül. A régiók bioinformatikai [39] módszerek kombinálásával kerültek azonosításra, mely összesen 47 rákos megbetegedéssel direkt kapcsolatba hozható rendezetlen régiót eredményezett. Érdekes módon, az azonosított példák sokfélék voltak, melyek más-más úton voltak összefüggésbe hozhatóak a betegség kialakulásával.

Az egyik leggyakoribb eset az volt, amikor a mutációk egy lineáris motívumot érintettek. Ezek a lineáris motívumok közvetíthetnek fehérje-fehérje kölcsönhatásokat (pl. az ösztrogén receptor esetén), de fehérje degradációt is szabályoznak (pl. β-catenin degron motívuma) vagy lokalizációs szignálokként funkcionálnak (pl. SMARCB1 fehérjénél). A mutációk érinthettek specifikus PTM-ket (foszforiláció, acetiláció vagy metiláció helyeit) is. Azonban ezen kompakt funkcionális modulok mellett, egy másik fontos csoportja a direkt módon mutálódó régióknak az autoinhibitor funkcióval rendelkezők, melyek egy szomszédos domén funkcióját szabályozzák. Erre egy példa EZH2 fehérje esetében a katalitikus domént szabályozó rendezetlen régió. A DNS/RNS kötő rendezetlen régiók alkotnak egy következő kategóriát, melyek például a FOX fehérjék forkhead doménjeinek DNS-t kötő részei. A mutációk érinthettek az eddig bemutatottakhoz képest szekvenciájukban hosszabb funkcionális modulokat, az ún. rendezetlen doménekeket (Intrinsically Disordered Domain - IDD) is, melyek általában evolúciósan konzerváltak és sokszor valamilyen más makromoleukához (pl. RNS/DNS) kötődnek. Valamint, egy további fontos csoportot képeztek a rendezetlen linker régiók

[39]. A munka során azonosított régiókat részletesebben a 4. ábra mutatja be, ami mutatja az azonosított példákat funkcionális kategória és a rákban betöltött szerepük szerint.



4. ábra – Rákban mutálódó funkcionális régiók

A szürkével jelzett fehérjék esetében átfedő funkcionális rendezetlen régió figyelhető meg (pl. több PTM hely). A fehérje architektúra ábrázolásban a kék ovális forma a rendezett domént, a kék vonalak a rendezetlen régiókat, a rózsaszín "dobozok" pedig a funkcionális rendezetlen modulokat ábrázolják. A mutációs mintázat tekintetében, a fekete oszlopok az egy pozíciót érintő misszensz mutációkat, az alattuk lévő kisebb piros és narancssárga vonalak pedig az "indel" mutációkat jelölik [39].

A rendezetlen régiók szerepének pontos megértése a tumorgenezis mechanizmusában az első lépés a gyógykezelések kidolgozásában. Habár az eddigi eredmények értékes információkat tártak fel, egyben rámutattak arra is, hogy a rendezetlen fehérjék tumorgenezis mechanizmusában betöltött pontos szerepéről még keveset tudunk. Evolúciós szempontból, a rákos megbetegedésekért felelős géneket sok esetben a

többsejtűség megjelenésének evolúciójához kötik, azonban ezek a megfigyelések a doménekre korlátozódnak, figyelmen kívül hagyva a rendezetlen régiókat [40]. A tumorgenezis mechanizmusában szerepet játszó rendezetlen régiók evolúciós eredetének vizsgálata orvosbiológiai szempontból a gyógykezelések kifejlesztése során többek között az alkalmazott modell organizmusok kiválasztásában segíthet, rámutatva arra, hogy a rák mint betegség milyen összetettségű organizmusok szintjén képes kialakulni [41].

Lineáris motívumok

A rendezetlen régiók kölcsönhatásai nagyrészt egy korlátozott számú aminosavból álló, ún. lineáris motívumok (Short Linear Motifs, SLiM) által közvetített. A lineáris motívumok általában 3-10 aminosav hosszú közös aminosav mintázatokat alkotnak. A mintázat egyrészt rögzített (de potenciálisan variábilis), a kölcsönhatás szempontjából legfontosabb szekvenciális pozíciókból, másrészt ezeket kiegészítő kevésbé kötött, sokszor teljesen variábilis pozíciókból áll. Egy adott kölcsönhatásért felelős motívum többféle aminosav mintázattal is megvalósulhat, melyek néhány pozícióban egymáshoz képest el is térhetnek (Pl. ugyanazért a kölcsönhatásért felelős “RRSLRVHI” és “RDKRLSLNL” MAPK fehérjét kötő dokkoló motívamok). A pozíciók aminosav szintű flexibilitása teszi a SLiM-okat “degenerálttá”. A SLiM-ok sok esetben a kölcsönhatás során a “disorder-to-order transition” szerkezeti változásnak megfelelően lokális szerkezeteket vesznek fel. Az így kialakult szerkezetek lehetnek alpha-hélixek, beta-szálak, vagy irreguláris, coil szerkezetűek is [7]. Ugyan magára a kölcsönhatásra habár a fehérje többi részétől nagyrészt függetlenül képesek, a szekvenciális környezet bizonyos esetekben modulálhatja a kötést [42,43]. Funkció szempontjából a lineáris motívumok a sejtben sokféle kölcsönhatást közvetítenek, amelyek révén többek között meghatározzák partnerük sejten belüli lokalizációját, vagy befolyásolják azok működését és aktivitását. A lineáris motívumok által közvetített kölcsönhatásoknak köszönhetően egy adott fehérje sokféle, különböző szerkezetű és funkciójú partnerrel is

képes kapcsolatba kerülni. Ezáltal, a SLiM-eket hordozó fehérjék a szabályozási és jelátviteli útvonalak központi elemei.

Az ismert lineáris motívumokat nagyrészt az eukarióta fehérjékben figyelték meg, mindazonáltal néhány motívum más organizmusokban, pl. baktériumokban is megmutatkozik. A SLiM-ek széleskörű előfordulásának egy példája a PCNA-kötő PIP motívum a FEN1 (Flap endonuclease 1) fehérjében, mely az archeákban, baktériumokban és eukariótákban is megfigyelhető [43]. Ezen felül, SLiM-okat még vírusokban is megfigyeltek, melyek a gazda szervezet motívumait mimikálják, fontos szerepet betöltve a fertőzés mechanizmusában [44].

Az ismert motívumokat különböző adatbázisokban gyűjtik (mint pl. PROSITE [45], MiniMotif Miner [46], ScanSite [47], stb.), de ezek közül a legátfogóbb és legkiterjedtebb az ELM (Eukaryotic Linear Motif: Eukarióta Lineáris Motívum) adatbázis, amelyben jelenleg 262 féle motívumosztályt találunk, melyekhez összesen 197 fajból származó 3030 kísérletesen igazolt lineáris motívum tartozik [48]. Az ELM adatbázis a lineáris motívumokat jelenleg négy fő csoportba osztja, mely csoportokat az ELM adatbázis 3 betűs kóddal jelöli. **CLV** (Cleavage sites: hasító helyek), amelyek a proteázok cél régióját jelölik. **LIG** (Ligand binding sites: ligandum kötő helyei). Ezek fehérje-fehérje kölcsönható régiók, amelyek számos fehérje doménhez való kötést közvetítenek, mint például az SH2/SH3 domének. **TRG** (Targeting signals: cél szignálok), amelyek motívum alapú ismert lokalizációs szignálokat jelentenek, mint az NLS (Nuclear Localization Signal) és NES (Nuclear Export Signal). **MOD** (Modification sites: modifikációs helyek), PTM helyeit írják le, mint a foszforiláció vagy amidálás.

Bár az ELM adatbázist rendszeresen frissítik, az ismert lineáris motívumok száma folyamatosan nő, még mindig nagyon messze vagyunk az összes előfordulás teljes feltérképezésétől. Az emberi szervezetben ugyan a lineáris motívumok száma ismeretlen, egy 2014-es predikció szerint számuk százezres, akár milliós nagyságrendű

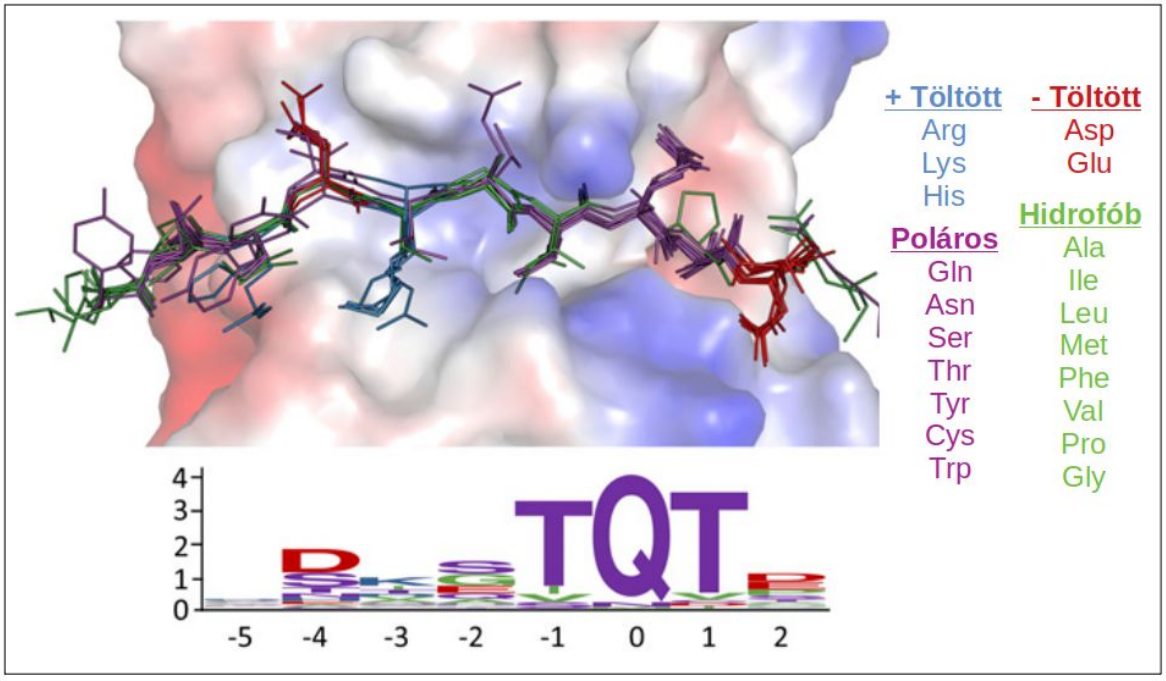
is lehet [49]. A jelenleg ismert és még azonosítatlan, feltételezhetően óriási számban előforduló lineáris motívumok közötti számbéli különbség arra mutat rá, hogy a lineáris motívumokat és az általuk közvetített kölcsönhatásokat nehezebb vizsgálni, ez a területen mind kísérletes, mind bioinformatikai szempontból még gyerekcipőben jár [50].

Az LC8 fehérje és kötőpartnerei

Egy lineáris kötőmotívum rendszerre példa az LC8 fehérje kölcsönhatásai. Az LC8 fehérjét elsőnek a *Chlamydomonas* zöld algákban az axonémás dinein motorfehérje komplex alegységeként határozták meg. Habár kezdetben az LC8-at a komplexben mint alegység, egy kargo adaptor szerepet betöltő fehérjeként írták le, az ismert kiterjedt intrakciós hálózata ennél egy jóval általánosabb, a motor komplextől független funkcióra utal [51,52]. Jól meghatározott funkciója a mai napig definiálatlan, azonban a fehérje esszencialitását bizonyítják azok a kísérletek, amelyekben az LC8 gén kiütése ecetmuslicában letalitáshoz vezetett [53,54]. A jelenlegi általános teória, hogy az LC8, mint dimerizációs vagy oligomerizációs fehérje tölti be szerepét [52,55]. Az eddig ismert partnerek olyan biológiai folyamatokhoz köthetőek, mint pl. a nukleáris transzport, tumor szupresszió, DNS hiba javítás, apoptózis, mitózis és jelátvitel [51], mely folyamatok szintén az LC8 fehérje fontosságára utalnak.

Annak ellenére, hogy az LC8 ismert partnerei funkcionalitás szempontjából nagymértékű heterogenitást mutatnak, a partnerek kölcsönhatása általában ugyanazon kötő-mechanizmussal írható le [51,56]. Az LC8 és partnerei alkotta komplexek kristályszerkezetei azt mutatják, hogy a kötőárok az LC8 homodimerizációjának határfelületén alakul ki, ami szintén homodimer partnerekkel hat kölcsön elsősorban [57,58]. Ezek a fehérjék gyakran tartalmaznak coiled coil (CC) régiókat, amik szintén szerepet jászanak a dimerizációban [59,60]. A kölcsönhatások a partner felöl SLiM-ek által közvetítettek, mely rendezetlen régiók a komplex kialakítása során a “disorder-to-order transition” folyamatnak megfelelően lokális szerkezetet vesznek fel. A kölcsönható

régiók legtöbbje esetében egy centrális Thr-Gln-Thr (TQT) aminosav motívum figyelhető meg (5. ábra), további, kisebb megkötést mutató, variábilis aminosavakkal kiegészítve. Habár az eddigi eredmények ezt a centrális tripeptidet írják le, mint a kölcsönhatásért nagymértékben felelős régió, egy 2019-es kísérleti munka eredményei arra utalnak, hogy a centrális régió szomszédos, variábilis aminosavai is fontos szerepet játszanak a kölcsönhatás kialakításában [51,61]. Habár igaz, hogy a centrális TQT aminosav-hármas az ismert kötőmotívumok nagy részében jelen van, ismert néhány nem-kanonikus példa is, melyek esetében a központi Gln Met-ra vagy Asn-re cserélődik [51]. Ezenfelül a Pak1 fehérjével való kölcsönhatást egy, még az előző példáknál is rendhagyóbb Thr-Ser-Pro (TSP) centrális aminosavak közvetítik.



5. ábra – Az LC8 fehérje kötőfelszíne ismert kölcsönható motívumokkal ábrázolva, valamint az ismert motívumokból épített szekvencia logó

A számolt elektrosztatikus töltések kék (pozitív) és piros (negatív) színnel jelöltek az LC8 kötőzseb felszínen (fehér - neutrális). Az kötőmotívumok az ismert kristályszerkezetekből

kinyertek és az aminosav karakterisztikájuk szerint színezettek (színkód az ábra jobb oldalán látható). A szekvencia logó 79 ismert motívum alapján készült [61].

Összességében ezidáig, a különböző kísérletes módszerek eredményeit összegezve több mint 100 igazolt, különböző eukarióta fajokból (főleg ember) származó LC8 kötőmotívumot írtak már le [51,60–62]. Az azonosított partnerek és a motívumok számának folyamatos növekedése rávilágít arra, hogy az LC8 interakciós hálózata és pontos funkciójának megértése egy fontos és aktuális kutatási terület.

Az LC8 egy evolúciós értelemben széles spektrumon megtalálható fehérje, a megszekvenált eukarióta fajok többségében jelen van [51,63]. Szekvenciális hasonlóság szempontjából a fajok között nagymértékű konzerváltságot mutat, a gerinces LC8 fehérje szekvenciák, azokon belül is maguk a kölcsönhatásért felelős régiók még a *C. elegans* és *D. melanogaster* fajok ortológjaival is 90%-os azonosságot mutat. A konzerváltságra irányuló megfigyelések eredményei arra világítanak rá, hogy további evolúciós vizsgálatok nagymértékben hozzájárulhatnak az LC8 funkcionalitásának pontosabb felderítéséhez, valamint partnereinek és azok kölcsönható motívumainak azonosításához.

A rendezetlen fehérjék evolúciós tulajdonságai

A rendezetlen fehérjék a globuláris fehérjékhez képest eltérő szekvenciális, szerkezeti és funkcionális tulajdonságokkal rendelkeznek. Mindez felveti azt a kérdést, hogy milyen evolúciós tulajdonságokkal rendelkeznek a rendezetlen fehérjék. Az egyik első munka ebben a témában 2002-ben született, amely során evolúciós megkötés szempontjából 26 olyan fehérje családot elemeztek, amelyeknek legalább egy, 30 aminosav hosszú rendezetlen régiójuk van. A vizsgálat során azt találták, hogy 19 család esetében jelentősen nagyobb evolúciós kényszer volt számolható mint a rendezett fehérjékre, 5 esetben nem tapasztaltak különbséget, a maradék 2 esetben az

erősebb konzerváltság pedig a DNS kötő funkcióval magyarázható, mely evolúciós értelemben egy ősi, konzervált funkció [64]. Emellet, egy másik munka során egy nagy-skálás vizsgálat keretén belül jutottak a leírtakkal megegyező eredményre magas aminosav szintű variábilitást megfigyelve rendezetlen régiókon belül [65].

A rendezetlenség és konzerváltság kapcsolatának tekintetében három szcenárió lehetséges, ami összefüggésben van ezen régiók funkciójával [66]. Az első esetben homológ szekvenciák között nagyfokú változatosság található mind a rendezetlenség, mind a szekvencia tekintetében. Egy másik lehetséges forgatókönyv, amikor ugyan nem detektálható szekvenciális konzerváltság, de a rendezetlen jelleg viszonylag állandónak tekinthető, ami viszonylag gyakran megfigyelhető linker régiók esetén. A harmadik lehetséges eset, amikor mind a szekvencia, mind a rendezetlenség konzervált. Ezek a konzervált szekvencia részek lehetnek viszonylag rövid szegmensek, amelyek általában valamilyen lineáris motívummal hozhatóak kapcsolatba. Azonban lehetnek hosszabb régiók is, melyekhez specifikus funkció társul. Annak ellenére, hogy ezek a hosszabb, evolúciósan konzervált szegmensek nem képesek egy jól-definiált 3D szerkezet kialakítására, lényegében önálló szerkezeti és funkcionális egységként kezelhetők, ezért javasolták rájuk a rendezetlen domén kifejezést [25].

A rendezetlen fehérjéken belül található SLiM-ok is változatos evolúciós mintázattal rendelkezhetnek. Egyes SLiM-ok erős konzerváltságot mutatnak és megjelenésük sok esetben egészen az egysejtűekig vezethető vissza. Ilyen pl. a már említett, több mint 3 milliárd évet túlélt PCNA fehérje kötő PIP motívum. Azonban, nem minden SLiM megjelenése magyarázható az ősi eredettel, sőt valójában a PCNA példája ritkának számít figyelembe véve a motívumok vélhető számát. Általában inkább evolúciós plaszticitás jellemzi őket. Ez arra vezethető vissza, hogy csak néhány kulcs pozíció mutációjával megjelenhetnek, illetve eltűnhetnek a szekvenciából. Ennek következtében egy adott kötőfelszínhez kapcsolódó SLiM-ok gyakran alakulnak ki konvergens evolúció

révén. [67] A SLiM-ok megjelenésére aztán 2015-ben irodalmi példákat feldolgozva Norman E. Davey és munkatársai adtak egy részletesebb mechanizmus leírást. Olyan példákat azonosítottak, amelyek esetében elég mennyiségű és minőségű adat állt rendelkezésre, hogy a motívumok megjelenését apróbb szekvenciális változásokhoz kössék. A mechanizmust “ex nihilo SLiM evolution” terminológiával írták le, mely a “semmiből” (rendezetlen régiók nem funkcionális részei) való születésére utal. Az elmélet szerint a SLiM-ok amilyen egyszerűen meg tudnak jelenni, az azt követő fixálódás hiányában olyan egyszerűséggel el is tudnak veszni [68].

A SLiM-ok esetében konzerváltság szempontjából egy egyedi, ún. sziget-szerű mintázat figyelhető meg. Ez a konzerváltsági mintázat abból ered, hogy a rendezetlen régiókon belül a SLiM-ok funkcionális pozíciói az evolúciós megkötés nélküli környezetükhöz képest relatíve nagyobb megkötést mutatnak. A sziget-szerű konzerváltsági mintázatot már korábban megfigyelték élesztő szekvenciákon belül [69], azonban az általánosabb modell kidolgozása Norman E. Davey nevéhez fűződik: különböző fajokból származó homológ szekvenciák többszörös szekvencia illesztéseit alapul véve, pozíciókénti konzerváltsági értéket számolva, a SLiM-ok funkcionális pozíciói a környezetük pozícióihoz képest statisztikailag szignifikánsan nagyobb konzerváltsági értéket mutatnak [70]. Ezen a megfigyelésen alapul a SLiMPrints algoritmus [71], ami képes még ismeretlen (“de novo”) SLiM-ok felismerésére.

Evolúciós vizsgálatok megközelítései

Szekvencia illesztések alapjai

Az evolúciós vizsgálatok kulcs elemei a szekvencia illesztések [72]. Ezek generálásának 3 fő eleme van. Első a hasonló pozíciók értékeléséhez használt szubsztitúciós mátrixok. Ilyen pl. az egyik legrégebbi PAM vagy a legszélesebb körben használt BLOSUM62,

melyek az aminosavak biokémiai hasonlóságukhoz köthető kicserélődési valószínűségeken alapulnak [73–76]. Másrészt, szükséges a bevezetett deléciók (“gap”) számolására valamilyen “büntetési” rendszer. Erre leggyakrabban az “affine-gap” (6B ábra) megközelítést alkalmazzák, szemben a kezdetleges, biológiai szempontból nem reális “linear-gap” (6A ábra) módszerrel. Az “affine-gap” megközelítés továbbfejlesztése a “generalized affine gap” módszer (6C ábra) [77,78], amely lehetőséget ad nem-homológ régiók egymáshoz illesztésére, gyakorlatilag ignorálva azokat ezzel, növelve a kevésbé hasonló de homológ szekvencia illesztések biológiai relevanciáját. Végül az illesztés generálás harmadik eleme maga az algoritmus, mely maximalizálja a hasonlósági értéket a szubsztitúciós mátrix és a “gap” rendszer alapján. Ezek az algoritmusok dinamikus programozáson alapuló megközelítések [79,80].

6. ábra – A deléciók (“gap”) bevezetésének 3 megközelítése

A BRCA1 és 53BP1 fehérje szekvencia illesztés részletei a 3 “gap” bevezetési megközelítéssel. (**A**) “linear-gap”; (**B**) “affine-gap”; (**C**) “generalized affine gap”.

A kettőnél több szekvenciák összeillesztését többszörös szekvencia illesztésnek nevezzük (Multiple Sequence Alignmnet - MSA), mely a páronkénti illesztés kiterjesztése. A legszélesebb körben alkalmazott MSA generáló algoritmusok progresszív illesztési módszereken alapszanak [81]. Az egyik első és legnépszerűbb

progresszív illesztési algoritmus család a Clustal [82]. Ennél egy modernebb algoritmusnak számít a MAFFT, ami egy Fourier transzformációs matematikai módszernek köszönhetően relatíve nagy szekvencia szám mellett is az egyik leggyorsabb szoftver [83,84].

A MSA illesztő algoritmusokat klasszikusan rendezett fehérjéken fejlesztik és optimalizálják, mely a rendezetlen fehérjék és egyben a SLiM-ok illesztésének problematikáját adja. Néhány modernebb algoritmus azonban képes jelentősen jobb teljesítményre a SLiM illesztés területén mint más, tradicionálisabb algoritmusok. Ezek közé tartozik a MAFFT, ami az átlagnál jobban képes a gyenge homológiát mutató szekvenciák illesztésére, ezáltal a rendezetlen régiókéra is [85,86]. Ezzel szemben, a KMAD algoritmus specifikusan a rendezetlen fehérjék illesztésére lett kifejlesztve. A módszer lényege, hogy egy klasszikus MSA algoritmussal generált illesztés további finomításra kerül egy, a rendezetlen fehérjékre specifikus szubsztitúciós mátrixszal, mely olyan meta-adatokat foglal magában, mint pl. ismert lineáris motívumok, PTM-ek, domén annotációk [86]. A megközelítés hátránya, hogy csak azon eseteket képes finomítani, amelyekről már valamilyen szintű, az algoritmusba integrálható ismerettel rendelkezünk.

Ortológia predikció

A fehérje szekvenciák közötti homológiára az aminosavak hasonlósága alapján lehet következtetni. Két szekvencia homológ ha azok egy közös őstől származtathatóak. Funkciójuk tekintetében, két homológ szekvenciának/fehérjének valószínűbb a hasonlóbb funkcionalitás, mint két nem rokon fehérjének [87]. A homológ szekvenciák specifikusabban további két csoportra oszthatóak: (i) Azon homológokat, amelyek két különböző fajban léteznek, ortológ szekvenciáknak nevezzük. Megjelenésüket a “speciation” evolúciós esemény eredményezi, vagyis az, hogy egy egykoron élt hipotetikus ősi organizmusból (hordozva egy bizonyos gént) az idő során 2 külön faj

evolválódott (mindkettő tovább hordozza a gént). Az ortológ kapcsolatok 3 típusát különböztetjük meg: “one-to-one” – két szekvencia ortológ párban áll; “one-to-many” – egy szekvenciához egynél több ortológ tartozik, ekkor már ko-ortológiáról beszélünk; “many-to-many” – 2, egynél több szekvenciát tartalmazó halmaz áll ortológ kapcsolatban (szintén ko-ortológok) (ii) Ezzel szemben azok a homológok, amelyeket ugyanazon faj genomja kódol, paralóg szekvenciáknak definiáljuk [88]. A paralógok megjelenése a gének duplikációihoz köthetőek egy adott organizmuson belül. A paralógok között 2 típust különböztetünk meg: az “out-paralog” szekvenciák kapcsolata evolúciósan távolibb rokonságot jelent (ősi duplikációra vezethető vissza, melyet további események, duplikációk és “speciációk” követnek,); az “in-paralog” kapcsolat fiatalabb duplikációra és evolúciós kapcsolatra utal, melyek sokszor faj-specifikusak [89].

Az evolúciós vizsgálatok egy másik fontos eleme az ortológ szekvenciák predikciója. A predikciónak két alapmegközelítése van. Az egyik, homológ szekvenciák páronkénti hasonlóságának értékelését veszi alapul. Két szekvenciát homológnak definiál, ha azok kölcsönösen egymás leghasonlóbb párjai egy adott szekvencia halmazba (Reciprocal Best Hit - RBH). Ezt az alap stratégiát alkalmazza a GOPHER [90] vagy az InParanoid [89], de számos más algoritmus is elérhető. Ezen algoritmusok egyszerűségük – nem igényelnek egyéb bioinformatikai műveletet – miatt gyorsak, azonban sok esetben csak “one-to-one” kapcsolatot tudnak visszaadni (modernebb algoritmusok a kezdeti ortológia párokat kibővítve képesek a ko-ortologia szintű predikcióra). A másik megközelítés a filogenetikai fa alapú. Ennek a stratégiának a lényege, hogy egy adott homológ szekvencia halmazból előbb egy gén-szintű filogenetikai fa épül, ezután ennek a fának a topológiája (a fajok egymáshoz képesti filogenetikai pozíciója a fában) egy általánosan ismert faj-szintű fa topológiájával összeegyeztetve módosításra kerül, mely folyamat során a cél a faj-szintű fa topológiájának az elérése minél kevesebb evolúciós eseményt (duplikáció, gén elveszés) bevezetve (“tree-reconciliation”) [91]. Ilyen módszer a PhylomeDB [92] vagy az ENSEMBL-Compara [93]. Ezek az algoritmusok általában

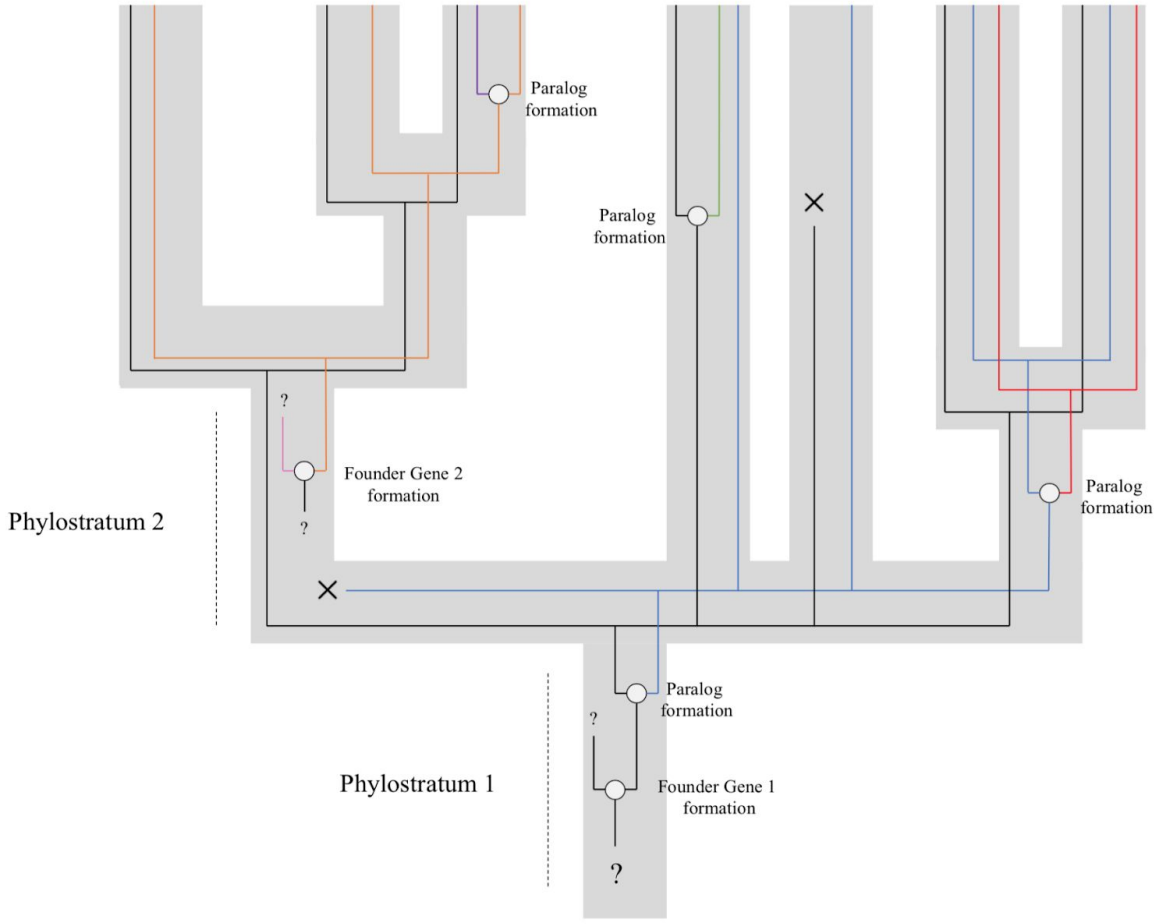
pontosabbak, képesek a ko-ortológia szintű predikcióra, viszont időigényesek és többszörös szekvencia illesztést igényelnek kiindulásként [91,94,95].

Evolúciós eredet predikció

Az ember legtöbb génjének vélhetően messzire visszanyúló evolúciós története van, sok géncsalád evolúciója az ősi eukarióta egysejtű organizmusok megjelenéséig vagy még korábbra vezethető vissza. Ez abból a megfigyelésből következik, hogy egyes gének olyan esszenciális biológiai folyamatokért felelős fehérjéket kódolnak, mely folyamatok megtalálhatóak a ma élő egysejtű organizmusokban is. Ilyenek pl. a genom integritás fenntartásáért felelős, vagy a metabolikus folyamatok. Az emberi gének eredetére vonatkozó információink azonban sokáig csak megfigyeléseken alapultak, szisztematikus vizsgálatokra, eredet meghatározási protokollok kidolgozására csak később, a megfelelő mennyiségű és minőségű genomikai adatok elérhetővé válása után került sor.

A gének eredetének meghatározására törekvő kutatási témát Domazet-Loso alapozta meg 2007-ben [40]. Domazet-Loso és munkatársai egy új módszert dolgoztak ki annak érdekében, hogy egy-egy génre specifikusan becsülhető legyen annak evolúciós eredete, mely segítheti számos molekuláris evolúciós folyamat jobb megértését. A módszer azon a molekuláris evolúciós feltételezésen alapult, hogy a genom diverzitás nem csak a duplikációk eredménye, hanem egyaránt hozzájárulnak *de novo* (“újonnani”) mechanizmussal megjelent gének is. Az így megjelent ősi géneket nevezzük “founder” géneknek, melyek megjelenése többféleképp is bekövetkezhetett, mint pl. a szintézis során történő leolvasási kereteltolódás. A “founder” gén alapú eredet detektálásra Domazet-Loso és munkatársai egy BLAST alapú bioinformatikai protokollt dolgoztak ki, mely lényege, minden olyan gén azonosítása egy adott adatbázisban, melyekben megtalálható ugyanazon adott konzervált domén, visszavezetve ezeket a géneket egy, a domént tartalmazó hipotetikus ősi “founder” génre (7. ábra). Az így

azonosított gének ezután elhelyezhetőek egy adott ún. “phylostratum” halmazba, mely halmazok rendszerbe foglalt összessége a “phylostratigraphy”. A “phylostratum” tehát egy olyan filogenetikai kládot jelent, amely magában foglalja az ugyanabból a “founder” génből származtatható összes gént, amely filogenetikai klád legősibb közös őse definiálja a géneredetet. Pl. ha a kináz domént vesszük alapul, mint a “founder” gén ősi egysége, az ehhez tartozó “phylostratum” halmazban emberi génektől kezdve egészen bakteriális géneket is megfigyelhetünk a BLAST keresés után, ami azt jelenti, hogy a domén által kódolt funkció (enzim aktivitás), mint biológiai “újítás” az ember és baktérium hipotetikus közös ősében jelenhetett meg, így evolúciós eredetnek ezt a taxonómiai szintet vehetjük [40].



7. ábra – A “founder-gene” evolúciós modell demonstrálása

Szürke színnel ábrázolt egy adott evolúciós forgatókönyv, melyben a paralógok megjelenése fehér körrel jelölt (“Paralog formation”). Az egyes gének evolúcióját különböző színek jelölik. Az evolúciós forgatókönyv 2 “founder” gén alakulását mutatja be (“Founder Gene 1,2 formation”) a hozzá tartozó “phylostratum”-ok (“Phylostratum 1,2”) definiálásával (függőleges szaggatott vonalak) [40].

A ‘phylostratigraphy’, mint általános megközelítés molekuláris szinten alkalmazható evolúciós újítások eredetének detektálására. Egy későbbi tanulmányban, Domazet-Loso és Tautz alkalmazva a kidolgozott eljárást, megvizsgálták, hogy a különböző emberi betegségekért felelős gének vajon véletlenszerűen jelentek meg az evolúció során, vagy a betegségek szempontjából biológiai kockázatot jelentő gének egy-egy jól meghatározható, az ember törzsfejlődésére specifikus evolúciós lépéshez köthetőek. Megvizsgálva ezen gének evolúciós eredetét egy 19 filogenetikai szintből álló ‘phylostratigraphy’ rendszert alkalmazva, azt találták, hogy a vizsgált gének egy nagyobb halmaza már az ősi egysejtűekben jelen lehetett, emellett azok nagy számban tűntek fel a többsejtű organizmusok megjelenésekor is. Ezek az eredmények rávilágítottak arra, hogy az emberi betegségekért felelős gének evolúciós eredet szempontjából nem egy véletlenszerű és nem is egy újabb génhalmaz, hanem két jelentős ősi evolúciós lépéshez köthetőek, az egysejtű és többsejtű organizmusok megjelenéséhez [41].

Domazet-Loso és munkatársai egy újabb munka során a kidolgozott “founder” gén megközelítést alkalmazva rákos megbetegedésekért felelős gének eredetének detektálására összpontosítottak. Korábbi eredményeikkel összhangban azt találták, hogy ezen gének eredete jelentős számban vezethető vissza a többsejtűség megjelenéséhez. Fontos emellett az a további eredmény is, hogy sok más, rákos megbetegedésekhez köthető gén megjelenése sokkal ősibb evolúciós szintre, egészen az egysejtű élet kialakulásáig vezethető vissza [96]. A munka eredményei alapján, az egysejtű eredetre visszavezethető, rákos betegségekkel összefüggésben álló gének elsősorban a genom stabilitásért felelősek, funkcióik az egyik legősibbek, melyek az evolúció során a mai napig fennmaradtak. Ezeket a géneket nevezzük a

szakirodalomban “caretaker” géneknek, amelyek esszencialitását mutatja az is, hogy diszfunkcionalitásuk bizonyítottan vezethet egysejtűek esetén letalitáshoz, emberben pedig rákos megbetegedés kialakulásához. A többsejtű organizmusok megjelenésének evolúciós szintjére visszavezethető rák-gének funkciói elsősorban a sejt-sejt kommunikációért felelős folyamatok biztosítása, mely a többsejtűség alapját adja. A gének ezen halmazát nevezzük “gatekeeper” géneknek.

Ezen eredmények bemutatták, hogy az emberi betegségekért, köztük a rákért felelős gén-szintű rizikófaktorok meglepő módon a ma élő ember jól tanulmányozott evolúciójánál sokkal régebbre, bizonyos esetekben a sejt, mint élet megjelenéséig vezethető vissza. Habár ezen eredmények az emberi betegségek kialakulásának megértése szempontjából egy fontos lépést jelentenek, a pontos ismeretek megszerzéséhez a gén-centrikus szemlélet miatt nem elég részletesek. A terület egy részletesebb szintű vizsgálatát adja a molekularégió-centrikus megközelítés, mely eddig még nem ismert információk feltárásán keresztül segítheti az emberi betegségek kialakulásának pontos megértését.

Molekuláris szelekció

A fehérjék evolúciója során a kódoló gének mutációknak vannak kitéve, mely mutációk evolúciós szerepe a véletlenszerű genetikai események által meghatározott, mint a negatív vagy pozitív szelekció. Ezen események detektálására irányuló tanulmányuk az 1990-es évekre vezethetők vissza, azonban csak az elmúlt években születtek olyan szofisztikáltabb megközelítések, melyek megbízhatóan alkalmazhatók.

Az ilyen módszerek ereje a genetikai változások fehérjére gyakorolt hatásából ered, pontosabban a fehérje szinten megjelenő (nonsynonymous) és nem-megjelenő (synonymous) szubsztitúciók relatív arányának számításából. A fehérje szinten nem-megjelenő változások előfordulása (Ds) neutrálisnak tekinthető a fehérjét kódoló

génre gyakorolt szelekciós nyomás szempontjából. Ezzel szemben, a fehérje szinten megjelenő változások a kódoló génekben (Dn) evolúciós nyomást gyakorolnak a szelekció során. Felhasználva a két előfordulás arányát (ω=Dn/Ds), a génekre gyakorolt szelekciós nyomás prediktálható: Ha a fehérje szinten megjelenő változások letálisak és ezért kiszelektálódnak, a fehérje szinten nem-megjelenő változások nagyobb száma miatt ω értéke kisebb lesz, mint 1, ha pedig a fehérje szinten megjelenő változások előnnyel járnak és fennmaradnak az evolúció során, ω értéke nagyobb lesz, mint 1. Neutrális evolúcióra pedig akkor következtethetünk, ha a Dn/Ds hányados értéke 1 körüli [97].

Pozitív szelekció számoláshoz az ω alapú megközelítést alkalmazza a CODEML algoritmus is. Az algoritmus filogenetikai fát, többszörös szekvencia illesztést és szubsztitúciós modelleket felhasználva prediktálja ω értékét, melyből aztán a szelekcióra következtethetünk. A CODEML-nek több modellje is létezik, melyek segítségével a szelekciót prediktálhatjuk gén- vagy pozíció-szinten, valamint ezek kombinációjával. A pozitív szelekció feltételezett hatása egy adott szekvencia halmazra statisztikai tesztek alapján fogadható el, valamint a pozíció szintű eredmények további valószínűségek (posterior probability) által értékelhető ki. A CODEML algoritmus egy nagyobb szoftvercsomag része, amelyet Ziheng Yan DNS, RNS és fehérje molekulák maximum likelihood (ML) optimalizáción alapuló evolúciós vizsgálata céljából fejlesztett ki [98].

Az ω alapú CODEML algoritmus azonban nem minden esetben képes hatékonyan detektálni a faj-, pl. Az ember-specifikus pozitív szelekciót. A módszer szelekció prediktáló ereje több tényezőtől is függhet, mint pl. a szekvenciák száma és hosszuk az illeszésben vagy azok divergenciája [99,100]. A módszer emiatt a faj-specifikus evolúciós ágakat érintő pozitív szelekciós változások detektálása során sokszor statisztikai értelemben megbízhatatlan eredményt ad, de ennek ellenére a predikció adhat hasznos információkat a szelekcióról [101].

A *H. sapiens* specifikus pozitív szelekció detektálása eredményre vezetőbb egy másik, közeli fajok közötti divergencia és polimorfizmus adatok összevetésén alapuló megközelítéssel. Ezt a megközelítést felhasználva dolgozták ki a McDonald and Kreitman (MK) tesztet [102]. Az eredetileg populáció genetikai vizsgálatokra kifejlesztett teszt, hatékonyan alkalmazható az embert és csimpánzt véve, mint közeli fajok, ember specifikus szekvencia pozíció változások azonosítására [103].

Jelenleg a rendezetlen fehérjékre irányuló pozitív szelekciós számolásokon alapuló munkák száma csekély, mely elsősorban az illesztés problematikájából ered. Egy nemrég publikált munkában azonban, 6600, legalább 30 aminosav hosszú rendezetlen régiót tartalmazó emberi fehérjét vetettek ω alapú szelekciós számolás alá. A számoláshoz emlős fajok (több mint 90) szekvenciáit használták fel, ezzel redukálva az illesztés problematikájából eredő hibák számát. A vizsgálat során 377 rendezetlen régió esetében detektálták szignifikánsan pozitív szelekció jelét, mely az alkalmazott modellt figyelembe véve, pozíció szintű relaxált evolúciós kényszernek felel meg [104]. Egy másik munka során, az MK teszten alapuló megközelítéssel Magdalena Gayà-Vidal és munkatársa 9785 emberi gént vizsgáltak meg faj specifikus pozitív szelekció szempontjából az ember és csimpánz közötti diverzitás és polimorfizmus adatok összehasonlításával. Eredményül 198 emberre specifikus pozitív szelekció alatt álló gént azonosítottak, melyek többsége az idegrendszer szinaptikus folyamataihoz és az immunrendszerhez köthető [103]. Ezen géneket ugyan a rendezetlen fehérjék szempontjából nem vizsgálták meg, de a pozitív szelekcióra több olyan esetet is hoznak, melyekben rendezetlen részek figyelhetők meg (pl. CDC42EP1, FBF1, ESR1, TXLNB).

Új funkcionális modulok azonosítása

A molekuláris evolúciós vizsgálatok egyik fő alkalmazási területe az új funkcionális modulok szekvencia alapú azonosítása. Ezek alapját a szekvencia kereső algoritmusok felhasználásával (pl. BLAST [105]) vagy manuálisan gyűjtött szekvenciák illesztésén alapul. Számos fehérje család esetén megfigyelhetők az aminosavaknak egy specifikus klasztere, amely valamilyen specifikus funkcióhoz, pl. kötőhelyhez vagy enzimaktivitáshoz kapcsolódik. Ezek jellemzésére az egyik legegyszerűbb módszer a konszenzus definíció (ún. reguláris kifejezés) építése és alkalmazása. Az azonosított szekvencia motívumok vagy mintázatok segíthetnek a távoli rokonságban álló fehérje szekvenciák felismerésében is. Az egyik legelső szekvencia motívum adatbázis a PROSITE [45].

A szekvencia családok jellemzésének másik fontos eszköze pozíció specifikus pontozó mátrix (position specific scoring matrix - PSSM) alapú megközelítésen alapul. A PSSM alapú megközelítés lényege, hogy a kiindulási szekvencia halmazból generálunk egy mátrixot, mely pozíció specifikusan hordozza a halmazban megtalálható szekvenciák aminosavainak/nukleotidjainak előfordulási gyakoriságára vonatkozó információt. A PSSM segítségével pontozhatunk egy potenciálisan hasonló régiót és ennek alapján eldönthetjük, hogy homológnak tekintjük-e az adott szekvenciát vagy sem. Inzerciókat/deléciókat ez a megközelítés nem képes kezelni, tehát generálásakor konstans hosszú kiindulási szekvenciákra van szükség.

A szekvencia keresésben széles körben szintén elterjedt módszer a rejtett Markov-modell (Hidden-Markov-Model - HMM) alapú profil építése. A HMM profilok generálásához egy kiindulási szekvencia halmaz szükséges. A HMM profil az adott illesztés információit (szubsztitúciók, inzerciók, deléciók) tartalmazza, mely profilt aztán alkalmazunk a szekvencia keresésre. Fontos előnye, hogy hatékonyan használható

távoli homológok azonosítására [72,106]. A legnagyobb és legnépszerűbb HMM profil alapú adatbázis és szerver a Pfam [36] és a HMMER [107]. Bár a Pfam által azonosított szekvencia családok sok esetben globuláris doméneknek feleltethetőek meg, vannak rendezetlen szekvencia családok, melyek tagjai között a rendezetlenség ellenére detektálható szekvencia hasonlóság van. Az InterPro adatbázis különböző adatbázisok által összegyűjtött motívumok és szekvencia mintázatok, PSSM-ek és HMM profilok által azonosított szekvencia családok jellemzését és ismeretlen szekvenciák funkcionális elemzését adja [108].

Ezek az eszközök használhatóak új lineáris motívum találatok azonosítására is [109]. Ennek legegyszerűbb módja a már ismert SLiM-ok aminosav mintázataiból épített reguláris kifejezés alapú definíció alkalmazása. Ezt a megközelítést alkalmazza az ELM adatbázis is [48]. A SLiM keresésben szintén gyakran alkalmazott a PSSM alapú megközelítés is, elsősorban könnyű alkalmazhatósága miatt [110]. Továbbá, gyakorta alkalmazott a HMM alapú motívum keresés, mely előnye (a PSSM-okkal szemben), hogy képes kezelni az inzerciókat és delécióka t, amik a lineáris motívumok “degeneráltsága” miatt fontos [69]. Habár ezen alap SLiM predikciós megközelítések gyakorta alkalmazottak, közös hátrányuk, hogy nagyszámú nem valós (fals pozitív) találatot adnak. Ennek oka, hogy mivel a motívumok csupán néhány aminosavból állnak és így gyengén definiáltak, annak a valószínűsége, hogy a keresés során felmerülő motívum egyezés pusztán a véletlen műve, viszonylag nagy [111]. Ennek eredményeképpen a motívumkeresést nagyban akadályozza a hatalmas mennyiségű fals pozitív találatok száma.

Célkitűzés

Az evolúciós vizsgálatok alapvető fontosságúak a számítógépes biológiai megközelítésekben, segítségükkel új funkcionális modulok azonosíthatók, és betekintést nyerhetünk különböző biológiai jellemzők alakulásába a törzsfejlődés különböző szakaszaiban. Bár a rendezett fehérjék evolúciós tulajdonságait már számos szempontból jól karakterizáltak, a rendezetlen fehérjéket ebből a szempontból még jóval kevésbé vizsgálták. Doktori munkámban két problémakör vizsgálatát tűztem ki célul:

I. Evolúciós konzerváltságon alapuló szűrési kritérium kidolgozása az LC8 dinein könnyűlánc rendszerben.

A rendezetlen fehérjék egyik legfontosabb kompakt funkcionális egységei a rövid lineáris motívumok, melyek nagy része még feltáratlan. Az evolúciós konzerváltságon alapú megközelítések fontos elemei az új kötőmotívumok azonosításának, azonban optimális alkalmazásuk nagyban függhet a vizsgált rendszertől. Egyik fő célkitűzésem a csoportunk által tanulmányozott LC8 dinein könnyűlánc kölcsönhatási hálózatának vizsgálata volt, ezen belül is:

- Az ismert LC8 motívumok evolúciós konzerváltságának vizsgálata
- A motívum predikció fals pozitív találatainak kiszűrésére szolgáló, evolúciós konzerváltságon alapuló kritérium kidolgozása, a módszer protokollba integrálása és tesztelése LC8 dinein könnyűlánc kölcsönható motívumainak azonosításán keresztül.

II. A rákos megbetegedések kialakulásával direkt kapcsolatban álló rendezetlen régiók evolúciójának részletes feltérképezése.

A rendezetlen fehérjéket funkcionális fontosságuk miatt több rákos megbetegedéssel is összefüggésbe hozták, azonban ezek evolúciós eredetét, történetüket nem ismerjük. Ezen régiók evolúciós mechanizmusának megismerése orvosbiológiai szempontból abban segít, hogy megértsük a tumorgenezis során kialakuló diszfunkcionalitást, mely tudást a terápiák kidolgozása során alkalmazhatunk. Ennek a vizsgálatához a következő feladatokat tűztem ki célul:

- A rendezetlen régiók konzerváltságának értékelésére szolgáló új, mutációs adatokon alapuló módszer kifejlesztése, valamint ezt felhasználva a régiók eredetének becslésére szolgáló protokoll kidolgozása.
- Az evolúciós eredet egy részletes, régió- és fehérjecsalád-szintű meghatározása rákban szignifikánsan mutálódó rendezetlen fehérje adathalmazra
- A régiók megjelenésének mechanisztikus leírása és annak, a jelenlegi modellekbe való integrálhatóságának megvizsgálása. A régió szintű evolúció részletes bemutatása példákon keresztül

Anyagok és módszerek

Doktori munkám eredményei teljes egészében bioinformatikai kutatásokon alapulnak, ennek megfelelően egyrészt szabadon hozzáférhető sztenderd algoritmusokat és adatbázisokat alkalmaztam, másrészt a megoldandó problémákra saját programokat, protokollokat fejlesztettem ki. A különböző számolásokhoz, adatfeldolgozásokhoz, filogenetikai adatok modifikálásához és generálásához programjaimat első sorban Python3 és Perl nyelven írtam. A filogenetikai adatok előállítását részben R-ben csináltam. Az eredmények ábrázolásához különböző python csomagokat használtam, mint pl. Biopython [112] vagy ETE3 [113]. Az LC8 evolúciós vizsgálatának kapcsán az adatok webes megjelenítésére szolgáló adatbázis és szekvencia illesztés megjelenítő elkészítéséhez python alapú Django web keretrendszert, valamint javascript, D3, CSS nyelveket használtam.

Adatbázisok

Uniprot

A fehérje kutatás egyik alap adatbázisa az óriási adathalmazzal rendelkező UniProt (jelenleg közel 2 millió szekvencia annotált és folyamatosan bővül) [114]. Az adatbázis két fő részből áll. Az egyik a Swiss-Prot (jelenleg több, mint félmillió szekvencia), mely megbízható, kísérletes adatokon alapuló, kézzel annotált (vagyis nem automatizálva) szekvenciákat tartalmaz. A másik a TrEMBL, amely esetében a szekvenciák nem ellenőrzöttek, azok automatizálva kerülnek az adatbázisba. Utóbbi halmazból ellenőrzés és felülvizsgálat után kerülhetnek át szekvenciák a Swiss-Prot halmazba. Az adatbázis különböző fajok (embertől baktériumig) fehérjéinek szekvencia adataiból és más, a fehérjékhez kapcsolódó egyéb információkból áll, beleértve funkciójukat és lehetséges

előfordulásukat, szerkezetükre vonatkozó annotációkat, mutációikat és azok pozícióit, lehetséges kötő- és aktív helyeket, stb. Az adatbázis sok más adatbázissal összekapcsolt, ami segíti és gyorsítja a bioinformatikai munkákat. A bioinformatikában a fehérje szintű adatok forrását sok esetben a UniProt adatbázis szolgáltatja. Munkám során a fehérje szekvenciákat sok esetben (pl. motívum szekvenciák összegyűjtése az irodalomban sokszor csak pozíció szintű adatok alapján vagya Swiss-Prot szintű *H. sapiens* szekvenciák letöltésekor) a Swiss-Prot adathalmazból nyertem ki. A UniProt REST-API funkcióval támogatott, ami egy fontos technikai előny az automatizálhatóság szempontjából.

ENSEMBL

Evolúciós vizsgálataimhoz a szekvencia és filogenetikai adatokat az ENSEMBL adatbázis (verzió 99), és az azon belüli a “Compara” projekt része szolgáltatta [93]. A “Compara” adatbázis 282 referencia fajt tartalmaz, amiből 277 gerinces, 4 nem-gerinces és 1 az élesztő (*S. cerevisiae*). Az ENSEMBL és “Compara” adatok Rest-API funkció segítségével könnyen kinyerhetők, ami a bioinformatikai munkákat nagyban segíti.

COSMIC

A COSMIC adatbázis (jelenlegi verzió, 91) az egyik leggyakrabban használt, rákos megbetegedésekkel összefüggésbe hozható mutációkat tartalmazó adatbázis [115]. Az adatok egyrészt a napjainkban egyre jelentősebbé váló, és egyre több rákos mintát feldolgozó rák genom projektekből (Cancer Gene Project), másrészt a szakirodalomból származnak. Az adatbázisban a mutációk a DNS szintjén történő változásokat katalogizálja génenként (egy nukleotidot érintő szubsztitúciók, indelek, komplex mutációk).

QFO

A QFO adatbázis egyben egy projekt is, annak érdekében, hogy standardizálásra kerüljön a különböző ortológia predikciós módszerek hatékonyságának mérése [116]. Ennek érdekében egy egységes, folyamatosan bővülő és ellenörzött minőségű, modell organizmusok proteomjai alkotta referencia adathalmazt is tartalmaz. Az adatbázis jelenleg 78 referencia proteomot foglal magában. A munkámhoz az állati eredetű, eukarióta organizmusok proteomját használtam csak fel. Ezen fajok száma 66 volt.

Pfam

Az adatbázis evolúciósan konzervált fehérje családok szekvencia illesztéseiből generált, elsősorban domén HMM profilokat tartalmaz, melyek szekvencia kereséshez is alkalmazhatók (domén predikció) [117]. A profilok egy részének (Pfam-A) alapját biztosító szekvenciákat egy nagy gondossággal, manuálisan összeállított halmaz adja, mely minősége a predikció pontosságát szolgálja, míg másik része (Pfam-B) automatizáltan generált. A jelenlegi verzió (33.1) több, mint 18 ezer profilt tartalmaz.

ELM

Az ELM (The Eukaryotic Linear Motif) adatbázis kísérletesen igazolt lineáris motívumokat gyűjt, kategorizál és annotál [48]. A kísérletesen leírt motívum előfordulásokból konszenzus definíciókat határoz meg, melyek szekvencia keresésként alkalmazhatók motívum predikcióra is.

Az evolúciós vizsgálatok során alkalmazott módszerek

Mutációs adatok alkalmazása

A projektem során felhasznált mutációs adathalmazt kutatócsoportunk egy korábbi munkája során hozta létre. Az adathalmaz pozíció specifikusan, elsősorban misszensz mutációkat, illetve kisebb mértékben, kereten belül inzerciókat illetve deléciókat is tartalmazott a COSMIC alapú fehérje szekvenciákra vetítve. A mutációs adatokat áttérképezve az ENSEMBL szekvenciákra alkalmaztam. A szekvencia térképezés páronkénti szekvencia illesztéseken alapult.

Rákos megbetegedésekért felelős rendezetlen régiók halmazának összeállítása

A rákos megbetegedések kialakulásához köthető rendezetlen régiókat csoportunk egy korábbi munkája során azonosította [38]. A régiók azonosítása a COSMIC adatbázis által gyűjtött genetikai variációk fehérje-szintű vizsgálatán alapult, alkalmazva egy megközelítést, amely a mutációk dúsulását detektálja, ezáltal azonosítva fehérjéket, amelyekben a mutációs mintázat definiálja a rák szempontjából biológiai kockázatot jelentő régiókat [39]. A munka során 47 ilyen régiót találtak.

Az általam elvégzett evolúciós tanulmányhoz ezeket az azonosított, COSMIC adatbázis alapú fehérje szekvenciákat áttérképeztem az ENSEMBL adatbázisra, vagyis minden COSMIC-os szekvenciának megfeleltettem egy ENSEMBL szekvenciát (a megfeleltetés páronkénti szekvencia illesztésekre támaszkodva történt). A szekvencia illesztés alapján a mutációkat is áttérképeztem. térképezés a CDKN2A izoforma (p14Arf) esetén nem volt kivitelezhető, így a további vizsgálatokból ez a fehérje kiszűrésre került. Emellett, azon fehérjéket, amelyek esetén mind rendezetlen és mind rendezett szignifikánsan mutálódó régiót is tartalmazta, a rendezetlen régiókra irányuló fókusz

miatt szintén kihagytam a vizsgálatokból. Kiszűrésre kerültek továbbá azon esetek is, amikor a régióban megfigyelt mutációk közül nem a misszenszek domináltak, vagy a misszensz mutációk száma nem érte el a 15-öt. Ezen szűrési lépésre az evolúciós vizsgálat során többször is alkalmazott, általam kifejlesztett régió konzerváltság számolási megközelítés miatt volt szükség. Továbbá, a hiszton fehérjék közel 100 %-os identikusságuk miatt összevonásra kerültek, és csak a HIST1H3B fehérjét tartottam meg a további analízisekhez.Az összeállított adathalmaz 32 rendezetlen fehérje 36 régióját tartalmazta, melyek a következők voltak (zárójelben a régi pozíciók): APC (1284-1537), ASXL1 (1102-1107), BCL2 (2-80), CALR (358-384), CARD11 (111-134; 207-266; 337-436), CBL (365-374), CCND3 (278-290), CD79B (191-199), CEBPA (293-327), CSF1R (969-969), CTNNB1 (32-45), EIF1AX (4-15), EPAS1 (529-539), ESR1 (303-303), FOXA1 (248-268), FOXL2 (134-134), FOXO1 (19-26), HIST1H3B (28-28), ID3 (48-70), MED12 (44-44), MLH1 (379-385), MYC (57-60), MYCN (44-44), MYOD1 (122-122), NFE2L2 (20-38; 75-82), PAX5 (75-80), RPS15 (129-145), SETBP1 (858-880),SMARCB1 (368-381), SRSF2 (95-95), USP8 (713-736), VHL (54-136; 144-193).

Pozitív szelekció

Az evolúciós analízis során használt összeállított adathalmazra a pozitív szelekciós adatokat a Selectome adatbázisból [118] (verzió 6) nyertem ki. Ez az adatbázis automatizált módon prediktál pozíció-szintű pozitív szelekciós eseményeket törzsfejlődési ágakon, mely predikció alapját a PAML 4b szoftver csomag CODEML algoritmusának “branch-site” modellje biztosítja [119]. A munkám során elvégzett molekula evolúciós vizsgálathoz a Selectome adatbázisból azon pozitív szelekció alatt álló pozíciókat integráltam, melyeknek a CODEML által számolt valószínűségi értékei (posterior probability) meghaladták a 0.9-et.

Az ismert LC8 kölcsönható motívumok halmazának összeállítása

A halmaz összeállítása céljából a projekt során 76 LC8 fehérje kölcsönhatást gyűjtöttünk össze az irodalomból (40 származott *H. sapiensből,* a többi más gerinces (pl. egér, patkány) és nem-gerines fajokból (pl. fonalféreg, ecetmuslica)), melyek közül 9-et kiszűrtünk a motívum szintű bizonyíték hiánya miatt. Az így maradt 67 motívumból a CD-HIT algoritmust [120] alkalmazva készítettünk egy nem-redundáns halmazt. Végül az összeállított halmaz 62 igazolt motívumot tartalmazott.

Az AMOT és WWC evolúciós vizsgálatának eredményeiből generált ábrák készítése

Az ábrák generálásához készített programokat python3 nyelven írtam az ETE 3 [113] csomagot használva. A filogenetikai fák rekonstruálását a CodonPhyML [121] programmal végeztem. A funkcionális egységek ábrázolásához, egyrészt az InterProScan 5 [122] és Pfam [117] adatbázist használtam a domének, másrészt az ELM adatbázist [123] a lineáris motívumok adatainak kinyeréséhez.

IUPred2A

Az IUPred2A rendezetlenség és rendezetlen kötőhelyek predikciójára szolgál [35]. Munkám során ezek predikciójára az IUPred2A algoritmust használtam lokálisan futtatva, alapbeállításokkal. A módszer statisztikus potenciálokat használva becsli meg egy fehérje adott pozíciójához tartozó kölcsönhatási energiát, egyedül a szekvenciából. A becslés alapján mind a rendezetlenségre, mind a kötőhely azonosításra vonatkozóan a pozíciókhoz rendelt 0 és 1 közötti értékek alapján lehet következtetni. Alapbeállítással 0.5 fölötti érték jelenti a rendezetlenséget és rendezetlen kötőhelyet. A módszer hatékonysága ~80%, emellett az egyik leggyorsabbnak számít a rendezetlenség becslő algoritmusok körében.

BLAST, PSI-BLAST

A BLAST (Basic Local Alignment Search Tool) algoritmus fehérje vagy nukleotid rokon szekvenciák keresésére szolgál egy adott szekvencia adathalmazban [105]. A módszer mérőszáma az E-value, mely az adott találatok szignifikanciájára utal. Egy találat esetén az E-value azt jelenti, hogy véletlenen alapulva hány ugyanolyan vagy hasonlóbb, nem rokon szekvenciát kapnánk a használt adatbázisban. A PSI-BLAST az alap algoritmus továbbfejlesztése, mely egy kezdeti keresés után, a találatok alapján egy azokra specifikus PSSM-ot épít, majd ezt használja újboli keresésre. A módszer iteratív módon több körös keresést végez és minden körben a találatoknak megfelelően finomítja a PSSM-ot, így sok esetben evolúciós értelemben távoli rokon szekvenciákat is képes megtalálni [124].

MAFFT

Munkám során a szekvencia illesztések generálásához a MAFFT algoritmust [83] alkalmaztam lokálisan futtatva, alapbeállításokkal (verzió 7.464).

SLiMPrints

A SLiMPrints egy, a lineáris motívumok sziget-szerű konzerváltságán [70] alapuló még nem ismert motívum predikciós eszköz [71]. A predikcióhoz ortológ szekvenciák többszörös illesztését felhasználva a környezetükhöz képest statisztikailag szignifikánsan nagyobb konzerváltságot mutató rövid régiókat detektál. A konzerváltság számoláshoz a CS algoritmust alkalmazza [125]. Az algoritmust munkám során alapértelmezett beállításokkal használtam.

GOPHER

A GOPHER algoritmus az RBH (Reciprocal Best Hit - RBH) megközelítést alapul véve prediktál ortológ szekvenciákat “one-to-one” kapcsolattal [90]. A páronkénti szekvencia hasonlóság értékeléséhez a GABLAMO algoritmust [126] alkalmazza, a hasonlósági alapértelmezett küszöbérték 0.4 (két szekvenciát homológnak tekint, ha azok hasonlósági pontja eléri ezt az értéket). Munkám során az algoritmus 3.0 verzióját, az alapértelmezett beállításokkal használtam, lokálisan futtatva.

Eredmények

Az LC8 fehérje partnereinek detektálása bioinformatikai megközelítéssel

Az LC8 fehérje nagyszámú partnerének ismert kötőmotívumai lehetővé teszik további kötőpartnerek azonosítását a human proteomban. Azonban az így jósolt motívum találatok nagy része valószínűleg fals pozitív találat. Ez szükségessé teszi további szűrési kritériumok alkalmazását, és ennek fontos elemei az evolúciós vizsgálatok. A projekt molekula evolúciós vizsgálatainak elvégzése az én feladatom volt, mely vizsgálatok eredményeit ebben a fejezetben mutatom be.

Az ismert LC8 kötőmotívumok evolúciós vizsgálata

Elsőként a már ismert, kísérletesen igazolt kötőmotívumok evolúciós tulajdonságait elemeztem. Kutatócsoportunkkal korábban összegyűjtöttünk az irodalomból számos LC8 kölcsönható partnert (gerinces és nem-gerinces fajokból egyaránt), amelyek esetében összesen 62 kísérletesen igazolt kötőmotívumot írtak le. A partnereket a UniProt adatbázisban azonosítottuk és motívum szekvenciáikat kigyűjtöttük. Ezek alapján épült egy, a kötőmotívumokra specifikus PSSM, amit az evolúciós vizsgálataimhoz használtam (a motívumok összegyűjtéséről, a UniProt-ról és PSSM építésról részletesebben az Anyagok és módszerek fejezetben olvasható). A PSSM-ot alkalmazva megvizsgáltam az ismert motívumok evolúciós konzerváltságát, melyet ortológ szekvenciák között számoltam. Ehhez egyrészt szükséges volt az ortológ szekvenciákból épített többszörös szekvencia illesztés, másrészt a PSSM alapú konzerváltsági számolás kidolgozása. Az illesztéseket az alábbi módon generáltam. Első lépésként összegyűjtöttem az ismert motívumokat hordozó fehérjék ortológ szekvenciáit. A szekvenciák forrásaként a QFO (Quest for Orthologs) adatbázist [116]

használtam, mely gondosan összeállított proteomokat tartalmaz, direkt ortológ szekvenciák azonosítása céljából. Az ortológokat a GOPHER algoritmussal [90] prediktáltam, aminek egyik fontos előnye, hogy mivel lokális szerkezeti elemek hasnlóságának azonosításán alapulva működik (nem teljes szekvencia szinten), képes távoli fajok ortológjait is visszaadni csupán egy-két homológ egység, pl. domén alapján. Második lépésként összeillesztettem az ortológ szekvenciákat a MAFFT algoritmust [83] alkalmazva. A MAFFT egyik előnye, hogy fejlesztése során nagy figyelmet fordítottak arra, hogy gyenge homológiát mutató szekvenciákat is hatékonyan illesszen össze [86] (Az illesztés generáláshoz használt adatbázisról és módszerekről részletesebben az Anyagok és módszerek fejezetben olvasható). A PSSM alapú konzerváltság számolási eljáráshoz először is az illesztéseket, a bennük megtalálható fajok taxonómiai besorolása alapján, 5 fő, egymásba ágyazott evolúciós szintre osztottam: *Mammalia, Vertebrata, Metazoa, Fungi* és *Eukaryota*. Ezután minden ismert motívumot tartalmazó fehérje esetében, minden olyan szinten, ahol legalabb 2 prediktált ortológ volt, minden egyes ortológ szekvenciában külön pontoztam az illesztett motívum régiót a PSSM-ot alkalmazva pontozó rendszerként (voltak esetek, amikor egy adott szinten nem volt ortológ). Egy adott ortológ motívum régióját biológialiag relevánsnak, tehát konzerváltnak vettem, ha a pontozás során elérte a 3.3-mas küszöbértéket. Ezt a küszöbértéket kutatócsoportunk véletlenszerűen generált motívumok PSSM pontjainak eloszlása alapján egy korábbi lépésben optimalizálta [62], azonban ennek elvégzése nem az én feladatom volt, ezért a részletekbe nem megyek bele. A konzerváltnak vett motívum régiók számát aztán szintenként összegeztem és meghatároztam, hogy egy adott motívum százalékosan milyen mértékben van jelen egy adott evolúciós szinten. Az ismert *H. sapiens* motívumok evolúciós konzerváltságára vonatkozó eredményeket a 8. ábra foglalja össze (a színkód a számolt konzerváltságnak felel meg - piros 100%, fehér 0%).

8. ábra – A H. sapiens ismert LC8 kötőmotívumok konzerváltsága ortológokban, evolúciós szintenként

A fehérjék nevei és a motívumok kezdő pozíciója a sorok elején jelzettek. A téglalapok az ortológ jelenlétét jelölik egy adott evolúciós szinten (minimum 2 szekvencia). A konzerváltság mértékét az egyes szinteken a piros és fehér szín közötti színskála ábrázolja. A 100% konzerváltság (piros) jelenti, hogy a motívum minden egyes ortológban biológiailag relevánsnak volt tekintve, 0% (fehér) pedig, hogy egy ortológban sem volt megfigyelhető a motívum.

A vizsgálatom eredményei alapján az ismert LC8 kötőpartnerek közül evolúciós értelemben az emberi DYNC1|1 és DYNC1|2 (Cytoplasmic dynein 1 intermediate chain1-2) fehérje kölcsönható motívumai mutatják a legmélyebb konzerváltságot (8. ábra), minden evolúciós szinten az ortológok legalább 50%-ában potenciálisan jelen vannak a motívumok, olyan evolúciósan távoli fajokban is, mint pl. a fonalféregek vagy a nyálkagombák. Ezzel szemben sok esetben az igazolt LC8 kötő motívum konzerváltsága nem volt detektálható a *Vertebrata* szint fajain túl (pl. BMF, DLGAP1,

MTCL1). Ezen esetekben sokszor nem csupán a motívum hiányzott, hanem maga az ortológ sem volt prediktálható (pl. BMF, BSN, SNPH) (8. ábra). Hasonló módon megvizsgálva más nem *H. sapiens* fajok esetén igazolt LC8 kötőmotívumokat, azok konzerváltsága több esetben is csak adott evolúciós szintre, vagy fajra volt specifikus (9. ábra). Erre egy példa a *D. melanogaster* EGL RNS-kötő fehérje, melynek ortológjai az összes, a vizsgálatba bevont evolúciós szinten prediktálhatóak voltak, azonban a fehérjében kimért kötőmotívum konzerváltsága, eredményeim alapján, csak egyes *Metazoa* szinten jelen lévő fajokra korlátozódott. Ehhez hasonló érdekes eredmény továbbá, hogy míg a centriolum megkettőződésének folyamata konzervált a *Metazoa* fajok között, a folyamatban szerepet játszó ANA2 *D. melanogaster* fehérje LC8 kötő motívuma fajspecifikus, más élőlényben a motívum nem azonosítható.

9. ábra – A nem H. sapiens ismert LC8 kötőmotívumok konzerváltsága ortológokban, evolúciós szintenként

A fehérjék nevei és a motívumok kezdő pozíciója a sorok elején jelzettek. A téglalapok az ortológ jelenlétét jelölik egy adott evolúciós szinten (minimum 2 szekvencia). A konzerváltság mértékét az egyes szinteken a piros és fehér szín közötti színskála ábrázolja. A 100% konzerváltság (piros) jelenti, hogy a motívum minden egyes ortológban biológiailag relevánsnak volt tekintve, 0% (fehér) pedig, hogy egy ortológban sem volt megfigyelhető a motívum. Az ábrán látható vizsgált esetek a következő fajokból származnak: *Drosophila melanogaster* – ANA2, EGL, EWG, SW, SWA; *Rattus norvegicus* – BSN, GPHN, MARK3, NOS1, SNPH, ZNF354A; *Mus musculus* – DAZL; *Saccharomyces cerevisiae* – NUP159, PAC11. AZ ANA2 és SWA fehérjék esetén egy ortológ sem volt azonosítható, melyek fajspecifikusak lehetnek.

A kísérletesen igazolt motívumok PSSM alapú konzerváltsági vizsgálatát megismételtem a korábban alkalmazott 3.3-as értéknél alacsonyabb, 3-, és 2.7-es küszöbértékekkel is. A vizsgálat során jelentős változást nem tapasztaltam az eredményekben. Az ismételt vizsgálatok során a küszöbérték csökkentése összesen 3, az NRF1, FAM83D (438. pozíciótól) és MYO5A motívumok esetében volt hatással a konzerváltságra. Az NRF1 motívum így konzerváltságot mutatott a *Mammalia* szinten, valamint a FAM83D és MYO5A motívumok konzerváltsága a gerinces fajok között 25, illetve 20 %-kal nőtt.

Annak ellenére, hogy maga az LC8 fehérje és kölcsönható felszíne nagymértékű konzerváltságot mutat, a motívum konzerváltságra vonatkozó vizsgálatom eredményei alapján az a következtetés vonható le, hogy az ismert partnerek kísérletesen igazolt kötőmotívumaira jóval kisebb konzerváltság jellemző az ortológokon belül. Azonban megfigyelhető volt, hogy saját taxonómiai szintükön (*Mammalia*) belül az ismert *H. sapiens* motívumok általában nagymértékű konzerváltságot mutattak az ortológok között, legalább 80%-ban konzerváltak voltak (8. ábra).

Szigetszerű konzerváltság vizsgálata

Az ismert *H. sapiens* LC8 kötőmotívumokat megvizsgáltam a lineáris motívumokra jellemző sziget-szerű konzeváltsági mintázat szempontjából is. A vizsgálathoz a SLiMPrints algoritmust [71] alkalmaztam (Anyagok és módszerek). A 40 igazolt humán motívumból, 13 esetben detektált az algoritmus statisztikailag szignifikánsan, a szekvenciális környezethez képest nagyobb konzerváltságot, mely az esetek 32.5 %-át adja ki. Ez az eredmény arra utal, hogy önmagában a szigetszerű konzerváltság nem alkalmas a motívum találatok szűrésére, mivel a valódi találatok jelentős részét nem azonosítja.

SLiM evolúciós konzerváltság alkalmazása, mint szűrési kritérium

Az evolúciós vizsgálataim eredménye alapján definiáltam egy, a lineáris motívumok predikciójában alkalmazható szűrési kritériumot. A kritérium pontos szabályrendszere, hogy a predikció során minden olyan potenciális *H. sapiens* motívum találat kerüljön kiszűrésre, amely nem mutat legalább 80%-os konzerváltságot a *Mammalia* szint ortológjai között. A kidolgozott kritériumot az LC8 további kötőmotívumainak predikciója során teszteltem, mely predikcióra a csoport egy többlépéses bioinformatikai protokollt fejlesztett ki. Beépítve a kritériumot a protokollba, a lehetséges valódi motívumok számát közel 50%-al sikerült csökkenti. Ez a lépés nagyban hozzájárult a protokoll sikerességéhez, mely alkalmazása során 72 új potenciális LC8 kötőmotívumot tartalmazó halmaz (“High confidence list”) került azonosításra [62].

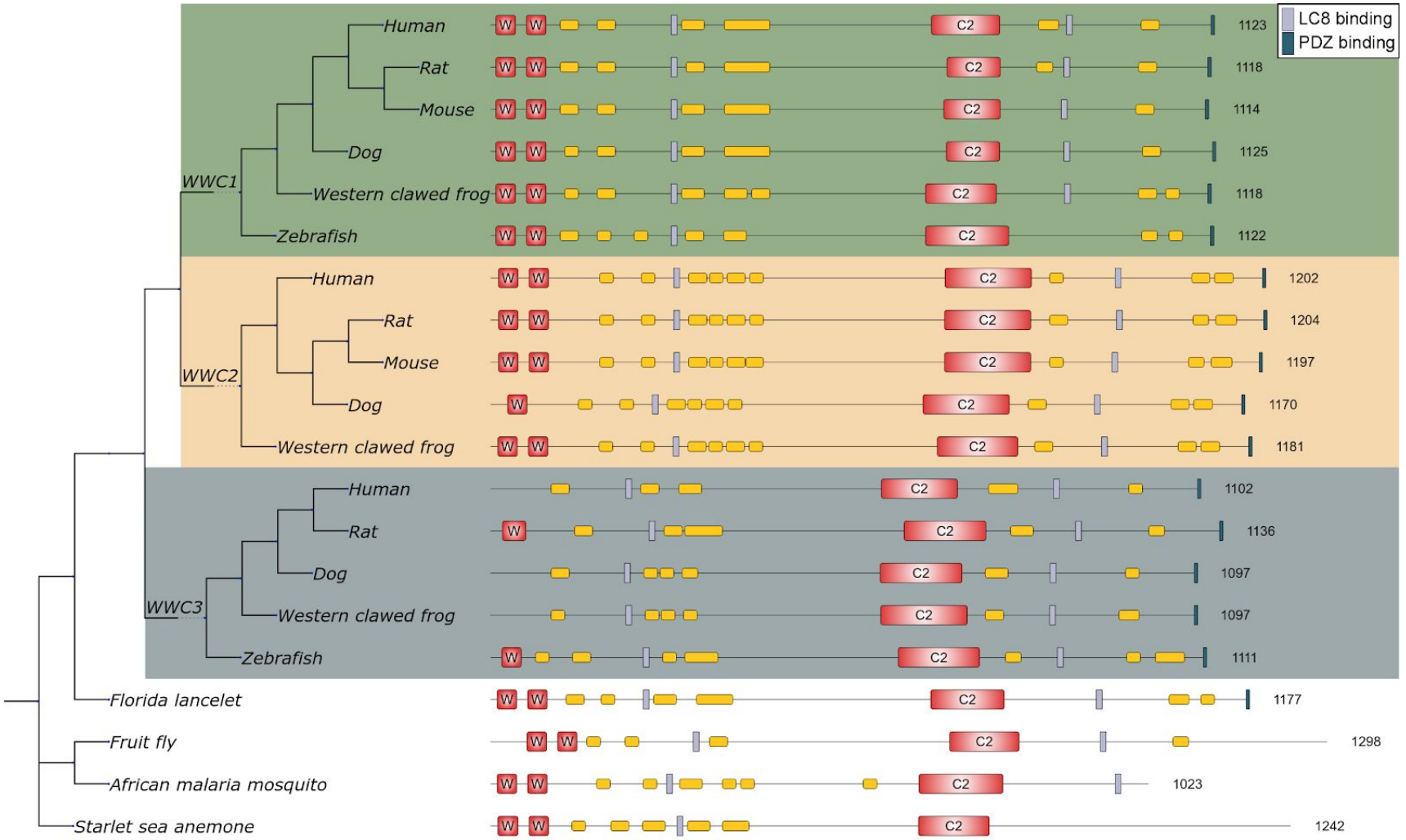
Az “High confidence list” motívum halmaz tárolásának és bemutatásának céljából részt vettem egy interaktív webes adatbázis fejlesztésében is. Ennek keretében interaktív táblázatokat készítettem és egy, a motívumok konzerváltságát szolgáló illesztés megjelenítőt fejlesztettem (http://lc8.elte.hu/). Az adatbázist publikálni tervezzük.

Az eredmények további, fehérje funkcionális analízise feltárta, hogy a már ismert funkciók mellet az LC8 fehérje összefüggésbe hozható a Hippo útvonal szabályozásával is. Ezen belül a Hippo útvonal “upstream” szabályozásában szerepet játszó két, a WWC (WWC1, WWC2, WWC3) és angiomotin (AMOT, AMOTL1, AMOTL2) fehérjecsalád tagjaiban azonosítottunk LC8 kötőmotívumokat. Ezek részletes vizsgálata jól demonstrálja, hogy az evolúciós vizsgálatok hogyan segíthetnek a biológiailag releváns új partnerek azonosításában.

Korábbi munkák során kimutatták, hogy a WWC1 (vagy KIBRA) fehérje az LC8-al kölcsönhat, mely kölcsönhatás közvetítése két kísérletesen igazolt lineáris motívumhoz (a fehérje 278- és 887-es pozíciójától kezdődő) köthető [127]. Azonban, a fehérjecsalád másik két tagjában (WWC2, WWC3) korábban nem azonosítottak kötőmotívumokat. Az LC8 és AMOT fehérjék kölcsönhatásait ugyan korábban fehérje szinten leírták [128], de a kötőmotívumok feltáratlanok maradtak. Az LC8 Hippo útvonalban betöltött szerepének megismeréséhez a WWC és AMOT fehérjecsaládok tagjait részletes evolúciós vizsgálatok alá vetettem mind a kölcsönható motívumok, mind az egyéb funkcionális fehérje egységek konzerváltságára összpontosítva. Az adatokat a kísérletes eredményeken túl a Pfam [117] és ELM [48] adatbázis szolgáltatta (Anyagok és módszerek – Adatbázisok).

Közel az összes, az evolúciós vizsgálatba bevont gerinces fajban 3 paralóg figyelhető meg egyaránt a WWC és AMOT család esetén. Mindkét család esetén, a gerinces paralógok legközelebbi gerinchúros ősében, a lándzsahalban, a fehérjék csak egy-egy példányban vannak jelen. Ez arra enged következtetni, hogy a WWC és AMOT paralógokat eredményező duplikációk gerincesekre specifikusak. KIBRA ortológok a lándzsahalon túl, annál jóval egyszerűbb élőlényekben, Arthropodákban (muslica, moszkító) és Cnidariákban (*N. vectensis*) is detektálhatóak (10. ábra). Az AMOT ezzel szemben csak az Arthropodákban figyelhető meg, azonban ezen taxonómiai szint vizsgált fajai közül a muslicában nem volt azonosítható ortológ (11. ábra). Erre

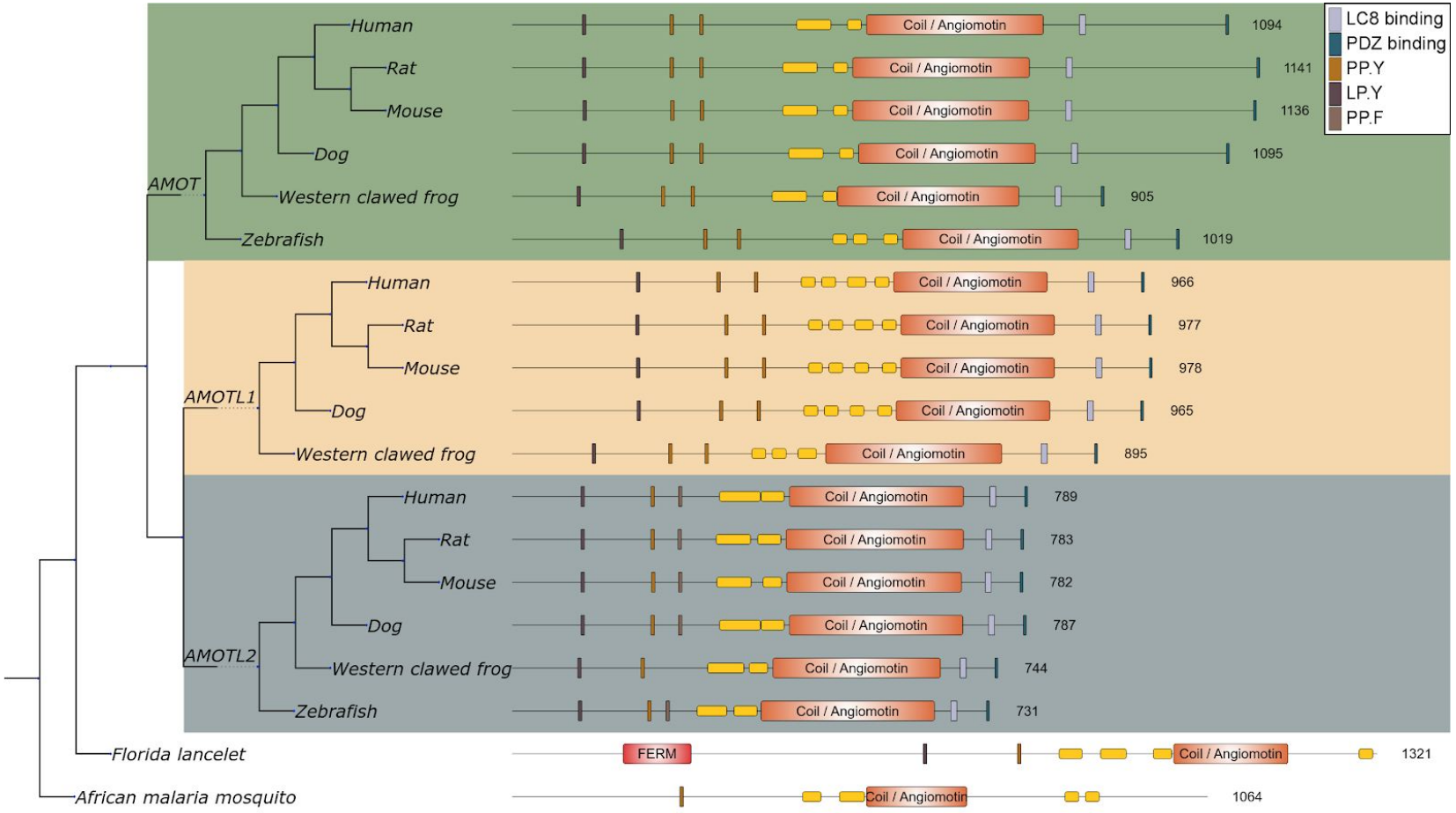
magyarázat lehet az, hogy az angiomotin fehérje eredetileg ugyan jelen volt minden *Bilateria* (kétoldali szimmetriájú állatok) organizmusban, de az evolúció során elveszett a *Diptera* (kétszárnyúak) törzsfejlődési ágon, amely a muslica evolúcióját is magában foglalja [129]. Az evolúciós vizsgálatom eredményei alapján elmondható, hogy a WWC fehérjecsalád egy ősibb evolúciós forgatókönyvvel rendelkezik, mint az AMOT család (10. és 11. ábra).


LC8 binding
PDZ binding
WWC1
Human
Rat
Mouse
Dog
Western clawed frog
Zebrafish
WWC2
Human
Rat
Mouse
Dog
Western clawed frog
WWC3
Human
Rat
Dog
Western clawed frog
Zebrafish
Florida lancelet
Fruit fly
African malaria mosquito
Starlet sea anemone
W
C2
1123
1118
1114
1125
1118
1122
1202
1204
1197
1170
1181
1102
1136
1097
1097
1111
1177
1298
1023
1242

10. ábra – A WWC fehérjecsalád tagjainak filgenetikai fája és a fehérjék funkcionális egységeinek ábrázolása

A filogenetikai fa topológiát mutat, a távolságok nem számolták. A 3 paralóg evolúcióját a filogenetikai fán belül a színes háttér téglalapok jelölik. A doméneket piros, a coiled coil régiókat pedig citromsárga színek jelölik. Lila az LC8 fehérje, sötétzöld pedig PDZ domén kötőmotívumokat ábrázolja.

Mind a WWC, mind az AMOT család tagjai a domén és motívum organizáció tekintetében nagymértékű konzerváltságot mutatnak számos fajban. Míg általánosságban igaz, hogy a legtöbb, emberben igazolt LC8 kötőmotívum nem konzervált a gerinces fajokon túl, a WWC család tagjai e tekintetben egy fontos kivételt képeznek. A családban a két LC8 kötő motívum nem csak minden gerinces paralógban konzervált teljes mértékben, hanem azokon túl, azonosíthatóak a gerinchúrosokban és az arthropódákban is. Ezen felül, a két motívumból az N-terminális részen megfigyelhető motívum a csalánozókban is konzervált, összhangban a családra jellemző WW és C2 domén evolúciós konzerváltságával, mely funkcionális modulok együttesen definiálják a családot. A WWC családdal ellentétben, az angiomotin fehérjék esetében csak egy LC8 kötőmotívum figyelhető meg, mely a vizsgálat eredményei alapján egy fiatalabb evolúcióval írható le. Ugyan ortológ szekvenciák azonosíthatóak nem-gerinces fajokban, amelyekben a gerincesekhez hasonlóan mind a coiled coil, mind pedig az angiomotin domén is konzervált, azonban az LC8 kötőmotívum ezen fajokban nincs jelen. Az eredmények alapján a motívum együtt más, PDZ és WW-kötő motívumokkal, csak a gerinces fajok megjelenésekor evolválódott és fixálódott (11. ábra). Az AMOT és WWC evolúciós vizsgálatának eredményeiről generált ábrák készítésének technikai leírása az Anyagok és módszerek – “Az AMOT és WWC evolúciós vizsgálatának eredményeiről generált ábrák készítése” részben írok.


LC8 binding
PDZ binding
PP.Y
LP.Y
PP.F
AMOT
Human
Rat
Mouse
Dog
Western clawed frog
Zebrafish
AMOTL1
Human
Rat
Mouse
Dog
Western clawed frog
AMOTL2
Human
Rat
Mouse
Dog
Western clawed frog
Zebrafish
Florida lancelet
African malaria mosquito
Coil / Angiomotin
FERM
1094
1141
1136
1095
905
1019
966
977
978
965
895
789
783
782
787
744
731
1321
1064

11. ábra – Az AMOT fehérjecsalád tagjainak filgenetikai fája és a fehérjék funkcionális egységeinek ábrázolása

A filogenetikai fa topológiát mutat, a távolságok nem számolták. A 3 paralóg evolúcióját a filogenetikai fán belül a színes háttér téglalapok jelölik. A prediktált átfedő Angiomotin domén és coiled coil régiókat összevontam és “angiomotin/Coil” névvel, narancssárga színnel jelöltem. A nem átfedő doméneket piros, a coiled coil régiókat citromsárga szín mutatja. A SLiM-ok pozícióit a következő színek jelölik: a 3 WW domén kötő motívumot – a barna szín 3 árnyalata (definíció szintű magyarázat az ábra felső sarkában); az LC8 fehérje kötőmotívumot – lila szín; a PDZ domén kötőmotívumot – sötétzöld.

Mind a két fehérjecsalád esetében a redundancia és a nagymértékű evolúciós konzerváltság rávilágít arra, hogy a lineáris motívumok által közvetített LC8 kölcsönhatásoknak fontos szerepe lehet a Hippo útvonal “upstream” szabályozásában.

A projekt során csoportunk kísérletes munkáiért felelős csapata SPR (Surface plasmon resonance) mérésekkel validálta a bioinformatikai protokoll által azonosított kölcsönható motívumokat a WWC és AMOT család fehérjéiben, ezáltal alátámasztva ezeket az eredményeket [62].

Rák szempontjából biológiai kockázatot jelentő rendezetlen régiók evolúciós vizsgálata

A csoport korábbi munkája során olyan régiókat azonosított, amelyekben jelentős számú, rákos megbetegedésekkel összefüggésbe hozható mutáció volt megfigyelhető [39]. A régiók közül 47 eredendően rendezetlen fehérjerégió volt. A megfigyelt mutációk ezen régiók fontos funkciójára utalnak, azonban a régiók evolúciós eredetét nem ismerjük.

Az evolúciós vizsgálathoz a 47 rendezetlen régió egy szűrt adathalmazát használtam. Mivel vizsgálatom kizárólag a rendezetlen régiókra összpontosult, a 47 régió közül kiszűrtem azokat az eseteket, amikor az adott fehérjén belül a rendezetlen régió mellett rendezett is azonosítva volt. Kiszűrésre kerültek azon régiók is, amelyek szekvenciái a használt adatbázisok között nem voltak áttérképezhetők, valamint azok az esetek is, ahol az evolúciós vizsgálatok szempontjából nem volt elegendő megfigyelt mutációs adat. A szisztematikusan összeállított adathalmaz végül 32 fehérje 36 rendezetlen régióját tartalmazta. Az adatok szűréséről részletesebben az Anyagok és Módszerek – “Rákos megbetegedésekért felelős rendezetlen régiók halmazának összeállítása” részben írtam.

Evolúciós eredet

A munka első része az evolúciós eredet feltérképezése volt, mind a rákos megbetegedésekkel kapcsolatban álló régiók, mind a régiókat tartalmazó fehérjecsalád szintjén. Az eredet predikcióhoz a szekvencia adatokat az ENSEMBL adatbázis, a filogenetikai adatokat (ortológ és paralóg evolúciós kapcsolatok, filogenetikai fák) pedig annak “Compara” része adta. Az ENSEMBL-Compara adatbázis ko-ortológia és ‘in/out’-paralógia szinten prediktálja az evolúciós kapcsolatokat, ami mind a pontos

régió- mind a fehérjecsalád-szintű eredet predikcióhoz a megfelelő részletességű evolúciós adatokat biztosította. Az ortológ és paralóg kapcsolatok az evolúciós távolságok szempontjából két szinten, fehérje-család és fehérje-alcsalád szinten meghatározottak. A fehérje-alcsalád szint egy-egy *H. sapiens* fehérje filogenetikáját tartalmazza, az így generált fát nevezi az adatbázis "sub-tree"-nak. A fehérje-család szint egy globálisabb filogenetikát jelent, mely evolúciós leírás több "sub-tree"-t kapcsol össze. Ezek jelentik az ún. "super-tree"-ket. Pl. két távolról rokon fehérje külön-külön épített, 2 "sub-tree" összekapcsolása adja a fehérjék egy globálisabb, család-szintű kapcsolatát, amely szinten a két fehérje már evolúciósan kapcsolatban áll, azok azonos faj esetén ősi paralógoknak tekinthetők. Az eredet predikcióhoz a "super-tree" szintű evolúciós kapcsolatokat használtam fel.

Annak érdekében, hogy az eredethez megjelenési kort rendeljek, az alábbi egymásba ágyazott evolúciós szinteket határoztam meg: *Mammals, Vertebrates, Eumetazoa, Opisthokonta*. A szintek kialakítását, vagyis az evolúciós skála kiterjedését az ENSEMBL-Compara taxonómiai lefedettsége határozta meg (az adatbázisban egyedüli egysejtű a *S. cerevisiae*, mely definiálja a skála legősibb, *Opisthokonta* szintjét). Az eredet szempontjából tehát a *Mammals*, mint evolúciós eredet szintje az emlős fajok közös ősét jelenti, a *Vertebrata* eredet a *Mammals* halmaz és minden más, nem emlős gerinces faj közös ősét jelenti és így tovább egészen az *Opisthokonta* szintig.

A "super-tree" szintű evolúciós adatokat és a kialakított evolúciós szintek rendszerét alkalmazva elsőként a fehérjecsalád-szintű evolúciós eredetet határoztam meg. Ezt egy adott vizsgált fehérjéből kiindulva, annak családjának eredetének egyszerűen a "super-tree" legősibb közös pontját (ősét) vettem.

Második lépésként a régió-szintű eredetet határoztam meg, ami egy összetettebb, többlépéses folyamat. Ennek egy kulcsfontosságú eleme a rendezetlen régiók konzerváltságának értékelése a szekvencia illesztésekben. A konzerváltság

számoláshoz egy általam kifejlesztett új módszert alkalmaztam, melyet a régió eredet predikció leírása előtt külön mutatok be. A módszer alapötlete, hogy azon pozíciók konzerváltságát értékelem két illesztett régión belül, melyek esetében kiugróan magas mutációs szám figyelhető meg, ami a pozíció alapvető fontosságára utal. A számolás során ezen pozíciók konzerváltságát értékelem csak, melyhez a BLOSUM62 szubsztitóciós mátrixot használtam. Két illesztett pozíciót konzerváltnak vettem ha azok szubsztitúciós értéke nagyobb volt -1-nél. Az eljárás kifejlesztése során megvizsgáltam a konzerváltság értékelését a -1-től eltérő küszöbértékekkel is. 0 esetében nem tapasztaltam változást, míg a -1-nél kisebb esetekben a módszer túl megengedőnek, 0-nál nagyobb esetben pedig túl szigorúnak bizonyult. A pozíciók konzerváltságának értékelése után, egy adott vizsgált régiót két illesztett szekvencia között konzerváltnak definiáltam, ha azon mutációs pozíciók, amelyek az előző lépésben konzerváltnak számítottak, a régióban megfigyelhető összes mutáció legalább 50%-át hordozták. A mutációs adatokat a csoportunk egy korábbi munka során állította össze (bővebben: Anyagok és módszerek – Mutációs adatok alkalmazása).

Alkalmazva a bemutatott eljárást, a régió-szintű eredet predikciót a következőképp végeztem el. Először, egy vizsgált régióhoz tartozó fehérjéhez összegyűjtöttem minden paralógot “super-tree” szinten, majd a paralógokhoz megfelelő ortológokat is. Ezután, a MAFFT algoritmust használva (bővebben: Anyagok és módszerek), a következő többszörös szekvencia illesztéseket generáltam: (i) összeillesztettem egymással minden összegyűjtött paralógot; (ii) összeillesztettem minden paralógot az ortológjaikkal. A további lépésekhez a vizsgált rendezetlen régiókat funkcionális egységenként (elsősorban lineáris motívumok, linker régiók) vettem figyelembe. Több esetben a definiált régió csupán egyetlen pozíció volt, amely általában egy ismert lineáris motívum kulcs aminosava, ilyen esetekben a régiók szekvenciális környezetét kiterjesztettem a motívum határáig. Következő lépésként, a bevezetett konzerváltság számolást és a paralógok alkotta illesztést felhasználva megvizsgáltam mely paralógokban konzervált a régió. Ezután, minden olyan paralóg esetében, ahol a régiót

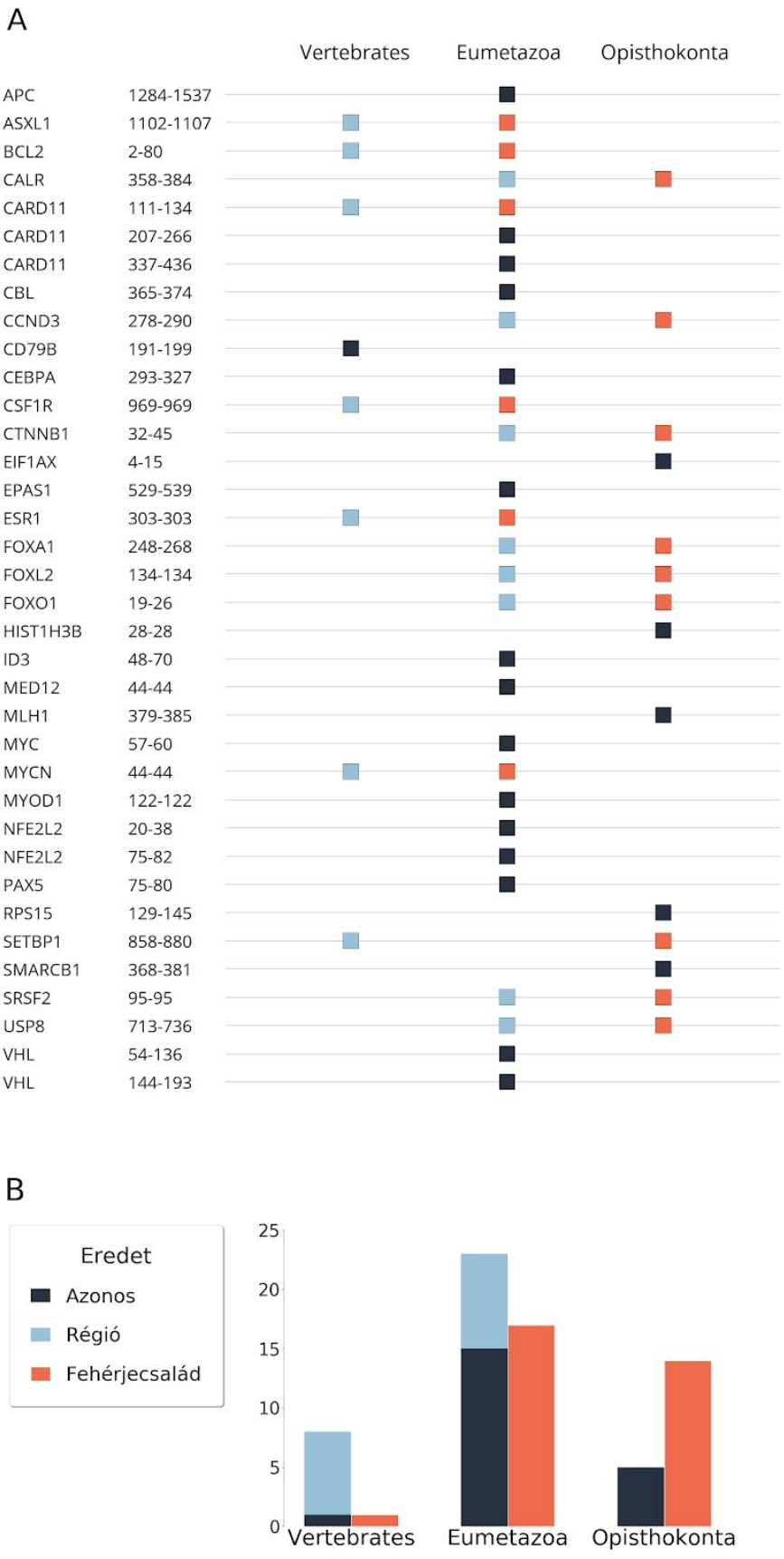
konzerváltnak definiáltam, megvizsgáltam ortológjaikban a régió konzerváltságát (ismét a kidolgozott eljárással) és azonosítottam evolúciós értelemben a legtávolabbi ortológot, amiben még a régió konzervált volt. Végül, az összes ilyen ortológ közül, kiválasztottam a legmélyebb taxonómiai szinten megjelenőt és ezt a szintet tekintettem a régió megjelenési szintjének. Pl. adott egy *H. sapiens* fehérje vizsgált régiója, ami csak a gerincesekben konzervát, de a régió jelen van egy másik *H. sapiens* paralógban is, aminek esetében a régió még az élesztőben is konzervált, akkor a vizsgált régió eredete az ember és élesztő közös ősét jelenti, ami a kialakított egymásba ágyazott evolúciós szint rendszerben az *Opisthokonta* szintnek felel meg.

A fehérjecsalád szintű evolúciós eredet feltérképezés eredményei az esetek többségében, vagyis 21 családnál *Eumetazoa* eredetet mutatnak, mely az irodalmi áttekintésben bemutatott gén eredetekre vonatkozó kutatásokkal összhangban áll [41]. 14 esetben ennél ősibb, egysejtű eredetet kaptam eredményül, mely ősi eredet a vizsgálatban az *Opisthokonta* evolúciós szinten jelenik meg. Egyetlen fehérjecsalád születése vezethető vissza fiatalabb, a gerincesek evolúciós szintjére, a CD79B (B-cell antigen receptor complex-associated protein β chain) fehérje családja (12A. ábra). Ez az eredmény összhangban áll több immunreceptor születésének korával, mely receptorok megjelenése transzpozonok inzercióival hozható összefüggésbe [130].

Az eredet vizsgálat eredményei a régiók közel fele esetében (21) azonos eredetet mutat a fehérjecsaládéval (12B. ábra). Érdekes módon, a 21-ből 5 esetben is (EIF1AX, HIST1H3B, MLH1, RPS15, SMARCB1) a régió eredet az ősi egysejtűekre vezethető vissza (*Opisthokonta* szint). A 21-ből 15 régió eredet az *Eumetazoa*, 1 pedig a *Vertebrata* szinten jelenik meg a fehérjecsaládéval azonos eredetként.

A vizsgálat során 15 esetben kaptam azt az eredményt, hogy a régió eredete fiatalabb, mint a fehérjecsaládé. Ezen esetek közül 8 régió az *Eumetazoa*, 7 pedig a *Vertebrata* szinten jelent meg (12B. ábra). Az eredetre vonatkozó megfigyelések alapján

általánosan elmondható, hogy az alkalmazott evolúciós felbontásban, a régiók egy szinttel később jelennek meg, mint fehérjecsaládjuk. Ettől egyetlen eset tér el, a SETBP1 fehérje régiója. Ebben az esetben, míg a régió a gerincesek ősében jelenhetett meg, addig a fehérjecsalád eredet az *Opisthokonta* szintig vezethető vissza (12A. ábra). Mindazonáltal, a SETBP1 esetében a fiatalabb családeredet nem zárható ki teljesen, mivel a magas rendezetlenségi arány miatt a fehérje eredet predikcióját fenntartásokkal kell kezelni. A többi esetben a nagymértékű konzerváltság miatt ez a probléma nem jelent meg. Összegezve, az eredet predikcióra irányuló eredményeim azt mutatják, hogy a rendezetlen régiók evolúciós értelemben később jelentek meg, mint fehérjecsaládjaik, mely megjelenés az *Eumetazoa* és *Vertebrata* szinten detektálható.


A
Vertebrates
Eumetazoa
Opisthokonta
APC 1284-1537
ASXL1 1102-1107
BCL2 2-80
CALR 358-384
CARD11 111-134
CARD11 207-266
CARD11 337-436
CBL 365-374
CCND3 278-290
CD79B 191-199
CEBPA 293-327
CSF1R 969-969
CTNNB1 32-45
EIF1AX 4-15
EPAS1 529-539
ESR1 303-303
FOXA1 248-268
FOXL2 134-134
FOXO1 19-26
HIST1H3B 28-28
ID3 48-70
MED12 44-44
MLH1 379-385
MYC 57-60
MYCN 44-44
MYOD1 122-122
NFE2L2 20-38
NFE2L2 75-82
PAX5 75-80
RPS15 129-145
SETBP1 858-880
SMARCB1 368-381
SRSF2 95-95
USP8 713-736
VHL 54-136
VHL 144-193
B
Eredet
Azonos
Régió
Fehérjecsalád
25
20
15
10
5
0
Vertebrates
Eumetazoa
Opisthokonta

12. ábra – A konzerváltság alapú evolúciós eredet eredményei régió és fehérjecsalád szinten

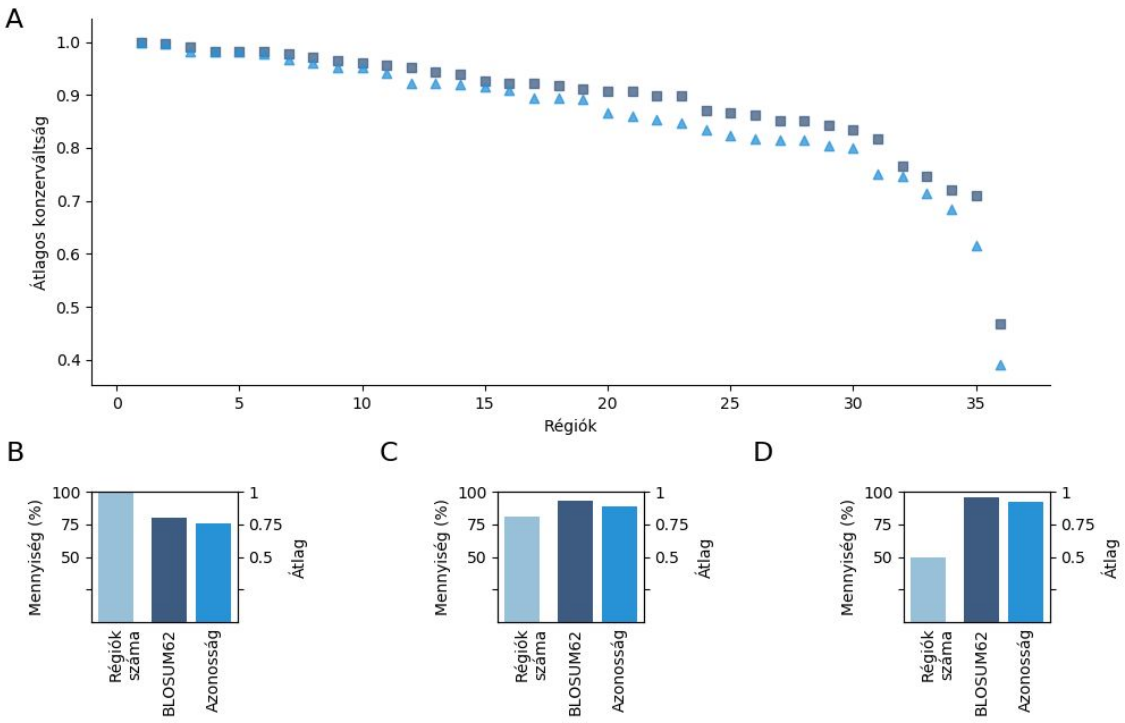
(**A**) A kék négyzetek a régiók, a narancssárga négyzetek pedig a fehérjecsalád eredetet mutatják. Fekete szín a közöz eredetet jelzi. (**B**) Az eredetek öszefoglalása oszlopdiagrammokkal.

Pozíció konzerváltság

Az előző fejezet eredményei tisztán rámutattak arra, hogy a rákos megbetegedésekkel összefüggésbe hozható rendezetlen régiók ősi eredetre vezethetőek vissza. Hogy részletesebb képet kapjunk a régiók ősi eredetéről és evolúciójukról, a mutációkat hordozó pozíciókat további, részletesebb konzerváltsági vizsgálat alá vetettem. A vizsgálat keretén belül ezen pozíciók egyedi konzerváltságát mind aminosav azonosság, mind homológ szubsztitúciók (BLOSUM62 szubsztitúciós mátrix alapján) szintjén megvizsgáltam. A vizsgálat eredményei rámutattak, hogy ezen pozíciók erősen konzerváltak mind a két szinten. Mindkét esetben a régiók 86%-ában a pozíciók átlagos konzerváltsági értéke nagyobb volt, mint 0.8 (13A ábra). Emellett, elmondható az is, hogy a 4 legalacsonyabb átlagos konzerváltsági értékkel bíró régió közül 3 esetében (CALR, VHL, APC) a pozíciók szintén erős konzerváltságot mutatnak ahhoz képest, hogy ezen régiók relatíve hosszúak (általános megfigyelés, hogy a hosszabb rendezetlen régiók konzerváltsága nagyon gyenge). A legalacsonyabb átlagos konzerváltsági értékkel a BCL2 régiója bír (13A ábra). Az alacsony konzerváltsági érték azzal magyarázható, hogy a megfigyelt mutációk nem csak a régió N-terminálisában, a konzervált BH4 domén régióba, hanem a nem konzervált, csak emlősökben megtalálható, C-terminális linker régióba is esnek.

Következő lépésben azt vizsgáltam meg, hogy hogyan változnak az átlagos konzerváltsági értékek, ha csak azon mutációkat hordozó pozíciókat veszem figyelembe, amelyek esetében kiugróan magas a mutációs szám. Ennek értelmében

megismételtem a konzerváltsági számolást egyrészt úgy, hogy a legalább 15, másrészt pedig a legalább 25 mutációt hordozó pozíciókat vettem csak számításba. Mivel nem volt minden régióban legalább egy olyan pozíció, ami elérné a bevezetett mutációs szám küszöbértéket, az így ismételt konzerváltsági vizsgálatban a régiók száma csökkent. Az átlagos konzerváltsági értékek szempontjából megfigyelhető volt, hogy azok emelkedtek mind aminosav azonosság, mind homológ szubsztitúciók tekintetében.



13. ábra – A régiók átlagos konzerváltsági értékei csak a mutációs pozíciókat figyelembe véve

(**A**) A régiók átlagos konzerváltsági értéke csökkenő sorrendben, azon pozíciókból számolva, melyek esetében legalább egy mutáció megfigyelhető. Négyzetek a BLOSUM62, a háromszögek pedig az aminosav azonosság alapján számolt konzerváltságot mutatják (a legkisebb érték a BCL2 régiója). (**B-D**) A régiók átlagos konzerváltsági értékeinek átlaga a legalább 1 (B), 15 (C) és 25 (C) mutációt tartalmazó pozíciókat számításba véve, valamint a régiók száma az adott halmazban. A konzerváltság BLOSUM62 szubsztitúciós mátrix és aminosav azonosság alapján számolt.

15 mutációt bevezetve küszöbértékként az átlagos konzerváltsági érték 0.89 volt aminosav azonosság, 0.93 homológ szubsztitúciók szintjén, továbbá 25 mutációs szám alkalmazása esetén ezek a számok 0.92 és 0.96 voltak (13B, C és D ábra). Az eredmények alapján elmondható, hogy azok a pozíciók, amelyek esetében magasabb mutációs szám figyelhető meg, magasabb konzerváltsági értékekkel is jellemezhetők.

A vizsgált rendezetlen régiók pozíció szintű vizsgálatát molekuláris szelekciós adatok összegyűjtésével és áttérképezésével folytattam a Selectome [131] adatbázisból. Ez az adatbázis az ω alapú szelekció számolást alkalmazva tartalmaz eredményeket, megjelölve többek között a szelekciós esemény becsült evolúciós megjelenési korát, valamint a potenciálisan érintett szekvencia pozíciókat is. A pozíció szintű pozitív szelekciós adatokat áttérképeztem a munkám során használt adathalmazra. Összesen 3 fehérje, a CALR, CTNNB1 és VHL volt érintett. Mindhárom esetben az adatbázisból kinyert adatok alapján, a pozitív szelekciós események megjelenésének becsült kora az ősi gerincesekre vezethető vissza (1.táblázat).

Fehérje név	***Pozitív szelekciós pozíciók*** (a pozíciók a *H. sapiens* szekvnciára vonatkoznak)
CALR	83 (0.971), 155 (0.971), 177 (0.990), 267 (0.995), 307 (0.994), 336 (0.991), **360** (0.999)
CTNNB1	121 (0.999), 206 (0.993), 250 (0.998), 287 (0.991), 411 (0.998), 433 (0.993), 525 (0.997), 552 (0.998), 556 (0.916)
VHL	**127** (0.957), **132** (0.942), 141 (0.923), **171** (0.947), **183** (0.963), **185** (0.920)

1. táblázat – A pozíció specifikus pozitív szelekciós adatok

A régiókkal átfedő pozitív szelekció alatt álló pozíciók feketével kiemeltek. Zárójelben a pozíció specifikus pozitív szelekcióra vonatkozó valószínűség (“posterior probability”).

Az eredmények azt mutatják, hogy az érintett pozíciók a 3 fehérjén belül csak limitált átfedésben vannak a rák szempontjából rizikófaktort jelentő rendezetlen régiókkal (1.táblázat). A CTNNB1 esetében egy pozitív szelekció alatt álló pozíció sem fedett át a régióval, valamint a CALR esetében is csak egyetlen ilyen áttérképezett pozíció volt megfigyelhető a régión belül. A VHL esete ettől eltér, 6 szelekció alatt álló pozíció áttérképezése után 5 a régióba esett azonban, ezek közül egyik sem érintett olyan pozíciót a régión belül, mely esetén kiugróan magas mutáció szám figyelhető meg.

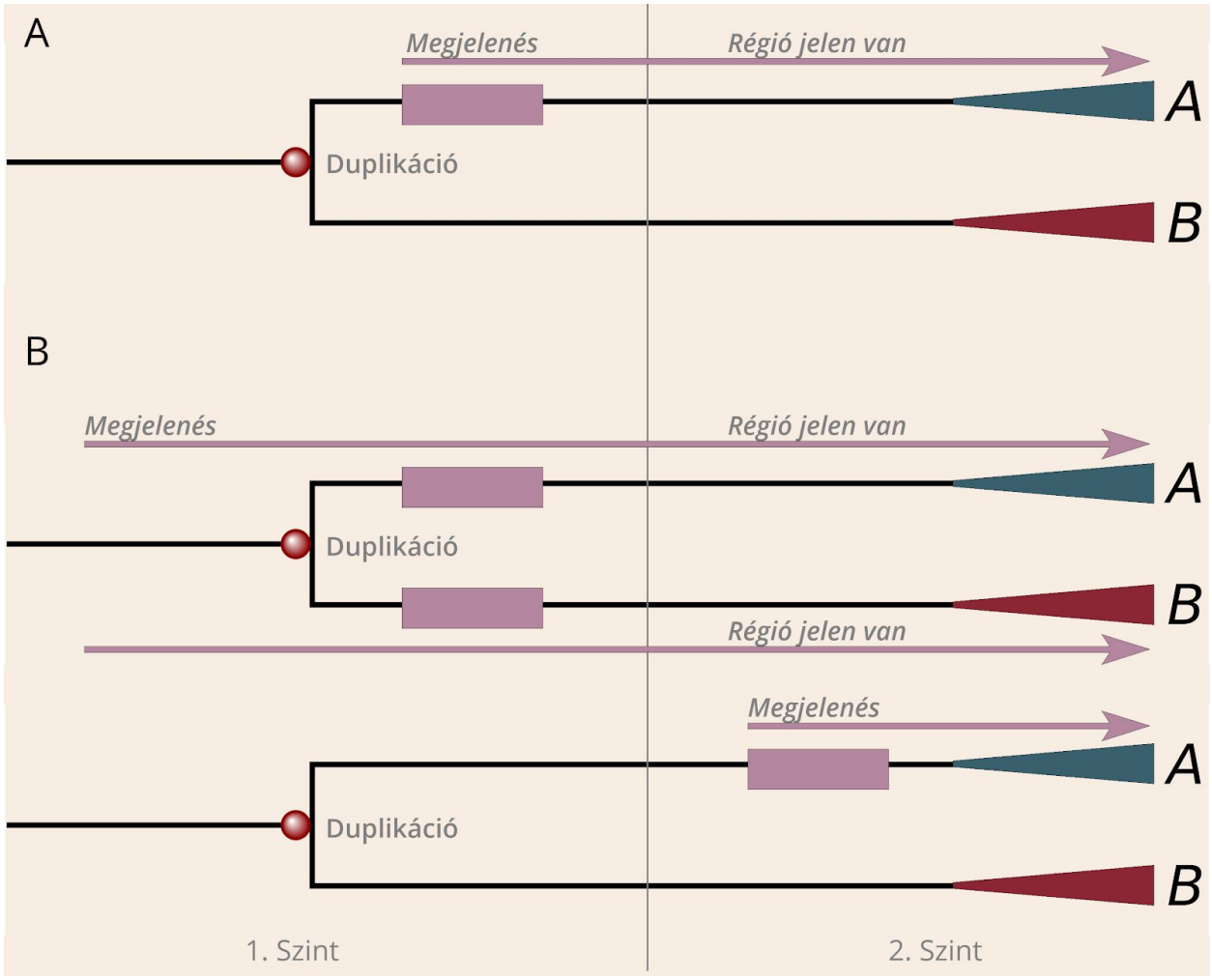
Egy korábbi tanulmányból [103] kiindulva megvizsgáltam, hogy becsülhető-e a *H. sapiens* fejlődési ágára specifikus pozitív szelekció. Mivel az ω alapú szelekció számolás nem alkalmazható megbízhatóan faj specifikusan, ebben a korábbi tanulmányban az irodalmi áttekintésben bemutatott McDonald and Kreitman (MK) tesztre is támaszkodtak, mely közeli fajok (ember vs. csimpánz) divergencia és polimorfizmus adatainek összehasonlítását veszi alapul. Feldolgozva a tanulmány eredményeit, a *H. sapiens* specifikus pozitív szelekció alatt álló fehérjék csak az ESR1-el fedtek át.

A duplikációk szerepe a rákos betegségekhez köthető régiók evolúciójában

Az új funkcionális fehérje régiók megjelenése gyakran génduplikációk által hajtott, mely folyamat a neofunkcionalizáció. A folyamat során a duplikátumok egyik példánya megőrzi az eredeti funkciót, míg a másik mutációk által új funkcióra tesz szert, mely funkció aztán a természetes szelekció elvén fixálódik. Ebben a fejezetben annak a vizsgálatnak az eredményeit mutatom be, mely arra a kérdésre válaszol, hogy vajon a rákos megbetegedések szempontjából biológiai kockázattal bíró rendezetlen régiók megjelenése leírható-e a neofunkcionalizáció folyamatával.

A paralógok molekuláris evolúciója sok esetben meglehetősen összetett, magában foglalva többszörös duplikációs eseményeket, amelyek sokszor más-más evolúciós időben (evolúciós szinten) történtek. Ebben a vizsgálatban azon duplikációs események

és a rendezetlen régiók megjelenésének kapcsolatát vizsgáltam meg, melyek duplikációk evolúciós kora megegyezett a régió megjelenésének korával. A vizsgálat során az alábbi két fő evolúciós modellt állítottam fel. Az első szcenárió a **neofunkcionalizáció**s folyamatnak feleltethető meg, amikor egy adott rendezetlen régió közvetlenül gén duplikáció után, csak az egyik duplikátumban jelenik meg és fixálódik (14A ábra). A második szcenáriónak, amikor a duplikáció nem köthető direkt módon a régió megjelenéséhez, két alapesete van. Az egyik, amikor egy duplikáció mindkét másolatában megtalálható a régió, ami arra enged következtetni, hogy a régiónak a duplikáció előtt kellett megjelennie (14B ábra). (A konvergens evolúció kizárható figyelembe véve a régió paralógok közötti nagymértékű szekvenciális konzerváltságát) A második alapeset, amikor a régió az adott duplikációs eseményt követően jóval későbbi evolúciós időben/szinten jelent meg, mely így direkt módon nem köthető a duplikációs eseményhez (14B ábra). Ez a két alapesetet magában foglalva jelenti a régiók duplikáció független, vagyis **de novo** megjelenési mechanizmusát.



14. ábra – A neofunkcionalizáció általi és de novo régió megjelenés mechanizmusának sematikus ábrázolása

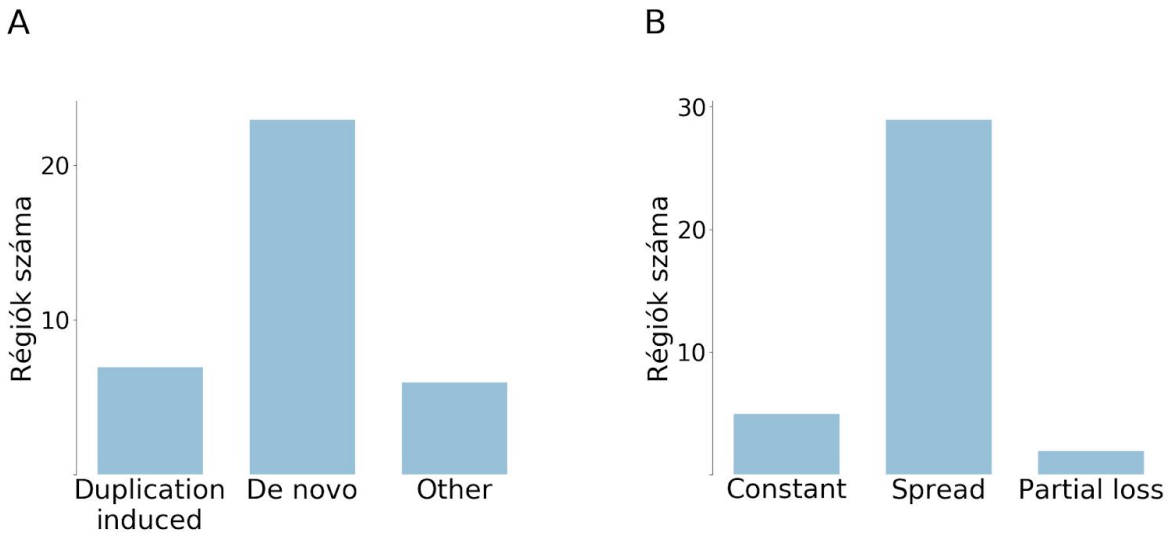
(**A**) A duplikációs indukált (neofunkcionalizáció) régiók megjelenésének modell szintű bemutatása. (**B**) A de novo modell két alaptípusának demonstrálása. Az ábrán a rózsaszínű téglalapok és nyilak a régiók megjelenését magyarázzak. A piros és zöld háromszögek a duplikáció általi paralógok további evolúcióját jelképezik.

Az első evolúciós forgatókönyvre példa a β-catenin fehérjecsalád, amely esetben a degron motívum alkotta biológiai kockázatot jelentő rendezetlen régió a β-cateninen kívül, annak legközelebbi paralógjában, a junctional plakoglobin (JUP) fehérjében figyelhető csak meg, míg más paralógokban nem. Ez tehát azt jelenti, hogy a régió a β-catenin és JUP közös ősében jelent meg, miután ez az ősi fehérje duplikációval elvált

a többi paralóg közös ősétől. A másik scenárióra példaként mutatható az ID3 esete, ahol a család többszörös duplikáción ment keresztül és minden paralógra áttérképezhető az ID3 rákos megbetegedésekhez köthető régiója. Ez a megfigyelés arra enged következtetni, hogy a régió már a duplikációk előtt megszületett. Egy másik példa az ESR1, amikor is a paralógok megszületése az eumetazoák megjelenésének evolúciós szintjére tehető, azonban a régió csak az ESR1-ben, egy későbbi szinten, a ma élő gerincesek ősében jelent csak meg.

4 rendezetlen fehérje (ASXL1, CCND3, SETBP1, CARD11) esetében nem voltak predikciós adatok a vizsgálathoz így nem lehetett a régiók megjelenési mechanizmusát meghatározni. Továbbá, 2 fehérjéhez, az RPS15- és SMARCB1-hez az ENSEMBL nem prediktált paralógokat, melynek oka ezen fehérjék nagyon ősi eredete, amely túlmutat az *Opisthokonta* evolúciós szinten is. A régió megjelenési mechanizmus szempontjából ez a 6 fehérje így nem került kategorizálásra, mely fehérjék a vizsgálat eredményében az "*Other*" kategóriát alkotják (15A ábra).

Az eredmények alapján a vizsgált adathalmazban összesen 7 eset jellemezhető a duplikáció indukált megjelenéssel, annak ellenére, hogy a neofunkcionalizáció egy gyakorinak gondolt mechanizmus az evolúció biológiában. A neofunkcionalizáció által megjelent 7 esettel ellentétben, 23 régió evolúciója köthető a "de novo" mechanizmushoz (15A ábra). Ez az eredmény rávilágít arra, hogy a rákos megbetegedések szempontjából rizikófaktorral bíró rendezetlen régiók megjelenésében ez a domináns mechanizmus.



15. ábra – A régiók kategorizálásának összefoglalása megjelenés és evolúciós sors szempontjából

(**A**) Régiók megjelenésének kategorizálása (duplikáció indukált – “Duplication induced”, “de novo”, egyéb – “Other”) oszlopdiagramban összefoglalva (**B**) A régiók evolúciós sors szerinti kategorizálásának (régió megőrződés – “Constant”, régió sokszorozódás – “Spread”, régió részleges elveszése – “Partial loss”) összefoglalása oszlopdiagramban.

Megvizsgáltam azt is, hogyan alakul a régiók evolúciós sorsa, azok öröklődnek-e további paralógok megjelenése esetén és ha igen, milyen mértékben. E tekintetben 3 szcenáriót definiáltam: (i) A régió további duplikációk nélkül megőrződik (“Constant”); (ii) a régió sokszorozódik öröklődve és megőrződve minden további paralógban (“Spread”); (iii) a régió öröklődik, de csak részben lesz jelen a paralógokban, néhány azok közül elveszik az evolúció során (“Partial loss”). Elvégezve a vizsgálatot az eredmények azt mutatják, hogy a második modell a domináns. 29 esetben figyelhető meg a régió megjelenése után további paralóg vagy paralógok születése, mely paralógokban a régió egyértelműen megtalálható. Ezzel ellentétben, 5 régió esetében nem volt megfigyelhető további paralógok megjelenése, mely esetekben tehát a régió csak azon fehérjében őrződött meg, amelyben megjelent (15B ábra). Ebbe a kategóriába tartozik néhány ősi

eredetre visszavezethető eset is, mint pl. az MLH1 és USP8, mely megfigyelésből arra lehet következtetni, hogy a duplikációk hiánya nem hozható összefüggésbe a fehérje esetleges rövid evolúciós idejével. A 3. szcenárió csupán 2, a VHL és az NFE2L2 esetében volt megfigyelhető (15B ábra). A VHL családnak összesen csak két tagja van, a VHL és VHLL, mely paralógok egy relatíve fiatal, az emlősök szintjén történt duplikációval születtek. A régió megjelenése után, míg annak N-terminálisa mindkét paralógban fennmaradt, addig a C-terminális szegmens a VHLL paralógból elveszett, csak a VHL-ben található meg (bővebben a következő fejezetben, a példák bemutatása során). Hasonló módon az NFE2L2 család tagjai is részben fiatalabb (*Vertebrata* szintű) duplikációval születtek, azonban ezen paralógokban nem figyelhető meg a rák szempontjából kockázatot jelentő lineáris motívum régiók.

Példák bemutatása

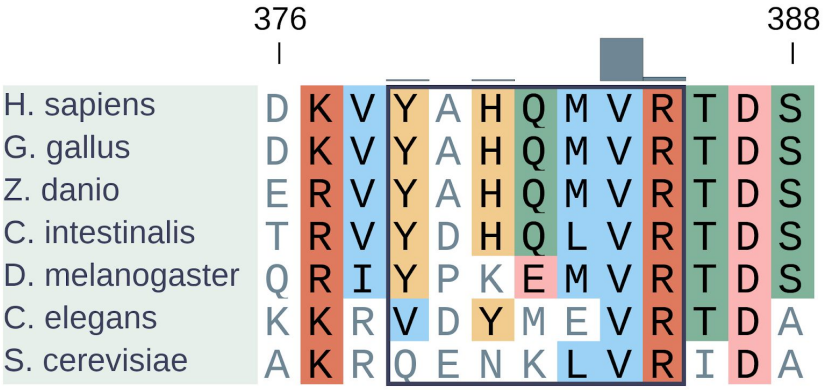
A rendezetlen régiók ősi eredetének és evolúciós folyamatuk demonstrálása céljából részletesebben megvizsgáltam 3 érdekes példát. A példák bemutatásához eddigi eredményeimet és az irodalomból kinyert adatokat használtam fel.

MLH1

Eredményeim alapján az egyik legősibb eredetre visszavezethető példa a DNS hiba javító (MMR - mismatch repair) esszenciális MLH1 fehérje. Mivel a fehérje feladata klasszikus példája a “caretaker” funkcióknak, az MLH1 mutációk általi diszfunkcionalitása növeli az egy nukleotidot érintő szubsztitúciók, valamint keret eltolódási mutációk arányát, mely változások rákos megbetegedések kialakulásával hozhatók összefüggésbe [132]. Az irodalomból összegyűjtött adatok alapján megfigyeltem, hogy az MLH1 több pozíciójának mutációját is leírták már olyan betegekben, akik Lynch szindrómás örökletes kolorektális rákban szenvedtek. Mindazonáltal, feldolgozva a COSMIC adatbázis adatait, mely szomatikus, rákos megbetegedésekkel összefüggésbe hozható mutációkat gyűjt és rendszerez (Anyagok és módszerek), az

MLH1 leggyakoribb mutációjaként a V384D figyeltem meg. Ezen szubsztitúció mutagenezis tanulmányai élesztőben, valamint in vitro MMR vizsgálatok azt mutatták, hogy ugyan nem volt megfigyelhető szignifikáns fenotípusos változás a kísérletek során, az MMR aktivitás csökkenése mégis kimutatható volt [133]. Ennek ellenére, más tanulmányokban azt találtam, hogy a V384D variáció nem csak bizonyítottan összefüggésbe hozható a kolorektális rákra való hajlammal [134], de domináns változásként írták le a HER2-pozitív luminális B-szerű mellrák esetén [135].

Az MLH1 fehérjének domén organizációja evolúciósan nagyon ősinek tekinthető, a rendezetlen linker régió által összekötött N- és C-terminális rendezett domén architektúra jelen van a baktériumoktól egészen az emberig [136]. Ez a konzerváltsági tulajdonság rámutat mind a rendezett domének, mind az azokat összekötő linker régió funkcionális fontosságára. A korábbi munkánk során azonosított, erősen mutálódó, rákos megbetegedésekkel összefüggésbe hozható lineáris motívum régió (379 - 385) a fehérje rendezetlen linker régiójában helyezkedik el [39]. A linker régió a DNS-el való kölcsönhatás során, valamint a szomszédos domének enzim aktivitása tekintetében regulátor szerepet tölt be [39]. Szekvencia konzerváltság szempontjából a motívum központi eleme a linker régióhoz képest erősebb megkötést mutat, amely a rendezetlen fehérjékre jellemző sziget-szerű konzerváltsági mintázatnak felel meg (16. ábra). Habár a pontos funkciója a régiónak még ismeretlen, az erős evolúciós konzerváltság különösen fontos funkcióra világít rá.



16. ábra – Az MLH1 fehérje vizsgált régiójának és környezetének többszörös szekvencia illesztése

A vizsgált régiót feket keret jelöli. A mutációk eloszlása az illesztés fölött, szürke oszlopokkal jelölt.

VHL

A VHL egy E3 ligáz aktivitással bíró tumor szuppresszor fehérje. Kulcsfontosságú szerepet játszik a sejt-szintű oxigén érzékelésben, ubiquitinációra és ezáltal proteoszómális degradációra megjelölve hipoxia-indukált faktorokat. A funkció ellátáshoz a VHL fehérje az elongon B- és C-vel,a cullin-2-vel és az RBX1 RING fehérjével alkot komplexet [137,138]. A VHL α-domén (más néven VHL-box, 155 - 192) alkotja a fő kontaktusokat az elongin C fehérjével, míg a nagyobb β-domén (63 - 154) direktben köti a HIF1α fehérjében és egyéb szubsztrátokban található prolin hidroxil motívumokat. A különböző rákos megbetegedésekkel összefüggésbe hozható mutációk a fehérje jelentős részén megfigyelhetők, beleértve az α- és β-domént is. Míg ezek a régiók komplexben az elongin B- és C-vel, valamint cullin-2-vel jól meghatározható szerkezettel rendelkeznek, addig izolátumban rendezetlenek és gyorsan degradálódnak [139].

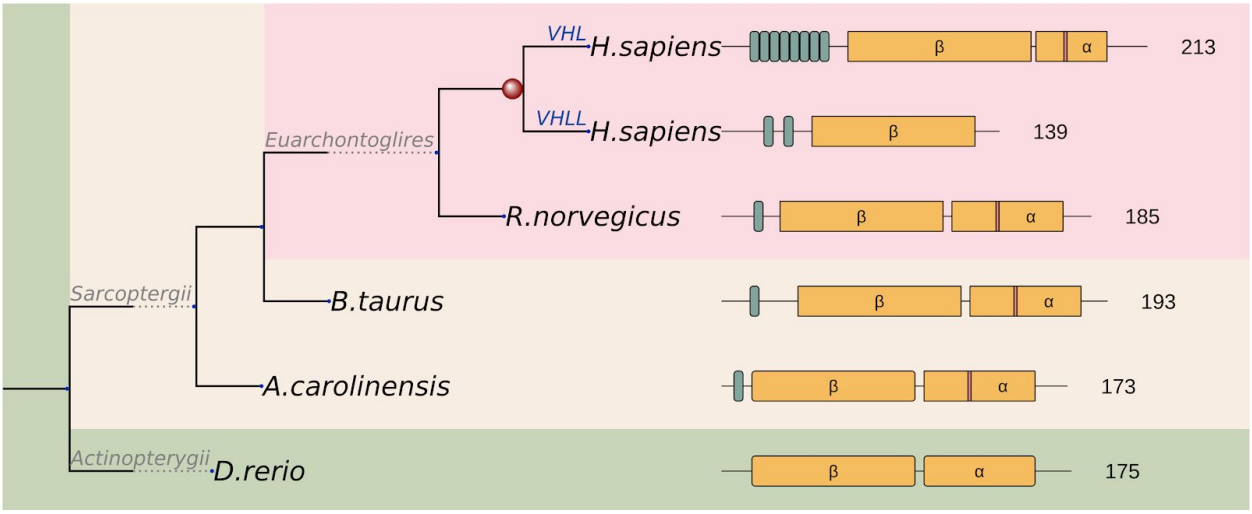
A VHL gén evolúciós eredete a hipoxia regulációs útvonal másik két kulcsfontosságú komponensével együtt, a HIF1α és PHD fehérjékkel, az *Eumetazoa* szintre vezethető vissza (17. ábra). Eredményeim rámutattak, hogy a VHL család molekula evolúciós

szempontból egy érdekes példa, mivel evolúciója eseménydúsabbnak mondható, mint más, rákos megbetegedéssel összefüggésbe hozható fehérje evolúciója.

Ahogyan az a 1. táblázatból látszik, a VHL esetében számos pozitív szelekció alatt álló pozíció detektálható, mely adaptív változások evolúciós kora a *Sarcopterygii* törzsfejlődési ágra tehető. Miután átnéztem a VHL irodalmát, egy érdekes megfigyelésre tettem szert. A pozitív szelekció alatt álló pozíciók magukban foglalják a K171 pozíciót is, mely az irodalom alapján fontos szabályozó funkcióval rendelkezik. A SUMO E3 ligáz PIASy (vagy PIAS4) fehérje kölcsönhatva a VHL-lel indukálja a K171 pozíció SUMOlizációját. Emellett, a K171 és K196 pozíciók ubiquitinálódni is képesek, mely modifikációk a PIASy által blokkoltak. Ezek alapján a VHL egy dinamikus szabályozására az alábbi modellt állították fel. A VHL és PIASy kölcsönhatása a VHL magi lokalizációját, valamint a végbemenő SUMOlizáció által a VHL stabilitását (blokkolva az ubiquitinációt) eredményezi. Másrészről, a PIASy disszocióció hatására a SUMOlizáció visszaszorúl, mely a VHL magból való exportját és ubquitinációját segíti elő. A modell szerint ezen reverzibilis modifikációkon alapuló dinamikus folyamat teszi lehetővé a HIF1α szubsztrát gátlásának precizitását [140].

A VHL egy másik érdekes molekula evolúciós eseménye a fehérje N-terminálisán evolválódott ismétlődő szakasz, amely GxEEx alapegységének megjelenése szintén a *Sarcopterygii* törzsfejlődési ágon figyelhető meg (17. ábra). Több ma élő fajban a GxEEx egység csak önmagában figyelhető meg, azonban a bizonyos törzsfejlődési ágakon a régió többször ismétlődik. A GxEEx ismétlődő szakasz funkcióját tekintve, prediktált és kísérletesen igazolt kölcsönhatási partnerek miatt (előbbi USP7, utóbbi p14ARF), a VHL egyik izoformájának szabályozásával hozható összefüggésbe [141]. A VHL család másik tagja, a VHLL egy fiatalnak mondható, *Vertebrata* szinten detektálható duplikáció eredménye, mely duplikátum érdekes módon elvesztette a szekvencia C-terminálisán megfigyelhető α-domént. Következésképp a VHLL nem

képes a sejtmagba lokalizálni a multiprotein E3 ubiquitin-ligáz komplexet. A VHLL funkcionalitása tekintetében a VHL egy domináns-negatív formájaként szolgál.



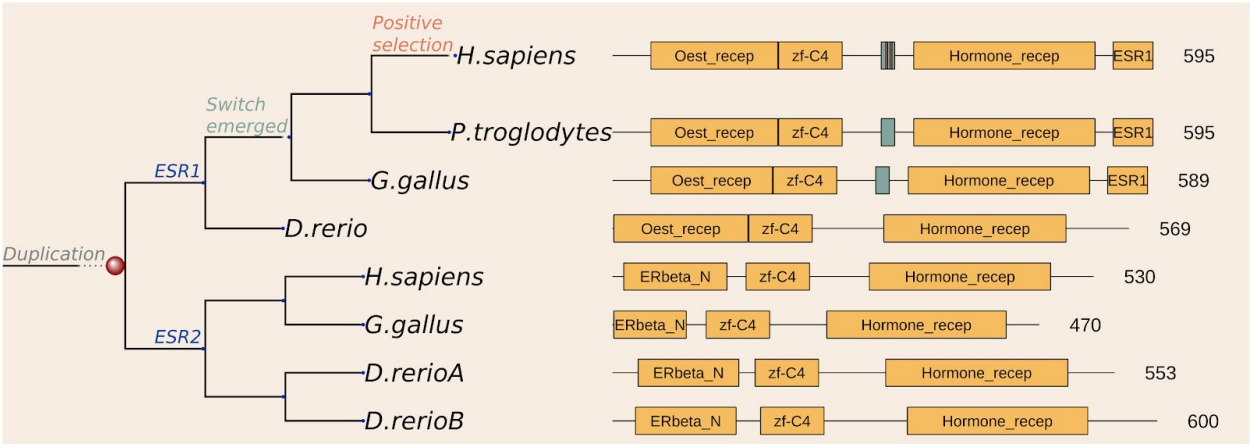
17. ábra – Sematikus ábrázolása a VHL fehérjecsalád evolúciójának és a fehérjék funkcionális egységeinek

A különböző számban előforduló ismétlődő szakaszok zölddel jelöltek. A α és β doméneket citromsárga színű elemek mutatják. Piros vonal jelzik a *H. sapiens* VHL α doménben lévő K171-es pozitív szelekció alatt álló pozíciót és tovább piros vonalak jelzik az ennek a pozíciónak megfeleltetett, szintén szelekció alatt álló, más fajok szekvenciájában lévő pozíciókat.

ESR1

Az ösztrogén receptor 1 (ESR1 - estrogen receptor 1) a nukleáris hormon receptorok családjába tartozó fehérje. A rákos megbetegedésekkel összefüggésbe hozható mutációk túlnyomó többségét a K303R variáció adja. A nagyszámú (a COSMIC adatbázis alapján több, mint 200) megfigyelt mutációt hordozó K303 egy lineáris motívumok alkotta, egymást kölcsönösen kizáró PTM helyeket magában foglaló rendszer (ún. molekuláris “switch”) pozíciója. Ebben a régióban, mint PTM-k, leírtak SET7/9 általi metilációt, p300 általi acetilációt, valamint PKA vagy PAK1 általi foszforilációt, sorrendben a 302- , 303-, és 305-ös pozíciókon [142–145]. Evolúciós vizsgálatom eredményei alapján a fehérje szintű evolúciós eredet az *Eumetazoa* szintre

vezethető vissza, azonban a K303 pozíciót magában foglaló régió ennél jóval később, a *Sarcopterygii* fajokban jelent meg és fixálódott (18. ábra). Eredményeim emellett megmutatták, hogy a régió metilációért és acetilációért felelős pozícióival ellentétben a foszforilációs hely megjelenése egy *H. sapiens* specifikus evolúciós esemény. Ez a megfigyelés abból következik, hogy az irodalomból gyűjtött adatok alapján a PKA általi foszforilációhoz az ESR1 esetében egyaránt szükséges a régió K302 pozíciója, valamint a csak *H. sapiens* szekvenciára specifikus R300 és L306 pozíció is (19. ábra). Ez a *H. sapiens* specifikus foszforilációs régió orvosbiológiai szempontból úgy okozhat gondot, hogy az onkogenikus K303R mutáció eredményeképp a régió abnormálisan, egy nagyobb affinitással köthető szubsztráttá válik a PKA számára [146,147].

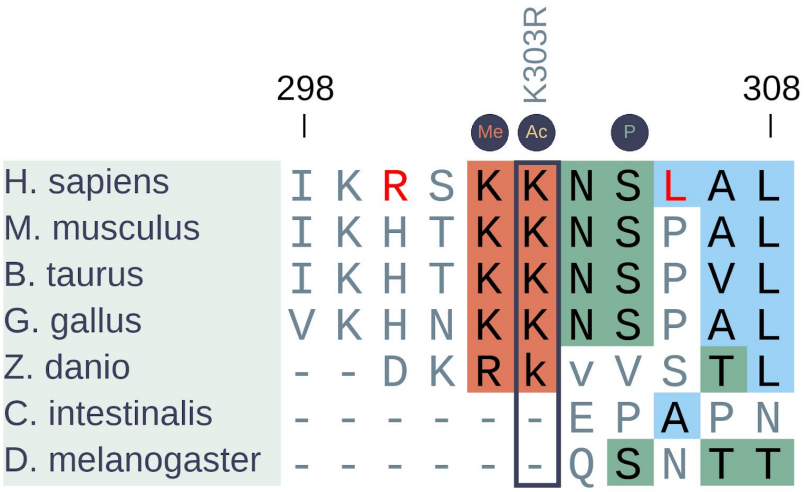


18. ábra – Az ESR fehérjecsalád evolúciójának demonstrálása

Citromsárga színnel jelöltek a domének, zölddel pedig az ESR1 vizsgált régiója és az annak megfeleltetett homológ régiók más szekvenciákban. Piros vonalak az ESR1 vizsgált régiójában a bemutatott pozitív szelekciós pozíciókat ábrázolja.

A *H. sapiens* specifikus változásokra vonatkozó megfigyeléseimet erre irányuló számolások is alátámasztják. Szubsztitúciók és polimorfizmusok számának összevetése egy adott gén esetében két közeli fajban egy hatékony megközelítése a faj specifikus pozitív szelekció detektálásának. Az ezen a megközelítésen alapuló MK tesztet egy korábbi munka során alkalmazva 9785 vizsgált génből 198 esetben azonosítottak emberre specifikus pozitív szelekciós változásokat, mely eredmények

között megtalálható volt az ESR1 is [103]. Az MK teszt eredményei alapján az ESR1 szekvenciájában, összevetve a csimpánzéval, a korábban bemutatott R300 és K306 pozíciók mellett további 3 pozíció (L44, Q502, S559) prediktálható *H. sapiens* specifikus változásnak. Ezek közül a S559 egy kísérletesen igazolt foszforilációs hely [103,148], mely így szintén egy *H. sapiens* specifikus PTM helynek mondható, azonban a másik két pozícióról az irodalomban nem találtam információt.



19. ábra – Inzerció-mentes többszoros szekvencia illesztése az ESR1 fehérjének

A Z. danio szekvenciájából kivágott inzerció határait kis betű jelöli. A régió nagy számban mutálódó K303 pozíciót fekete keret mutatja. A PTM helyek az illesztés fölött láthatók (Me - metiláció, Ac - acetiláció, P - foszforiláció).

Diszkusszió

A rendezett fehérjékkel ellentétben a rendezetlen fehérjék evolúciós szempontból kevésbé vizsgáltak. Kutatásaim központontjában a rendezetlen fehérjék evolúciós vizsgálatai álltak. Ezeknek a vizsgálatoknak az egyik leggyakoribb, alapvető eleme a szekvenciák/régiók evolúciós konzerváltságának értékelése. A rendezetlen régiók konzerváltsági számolása nem triviális. Ez abból ered, hogy a rendezetlen régiók konzerváltsága úgy értékelhető hatékonyan ha a mértükhöz képest kis, azonban a funkciójukat közvetlenül közvetítő pozíciókat vesszük figyelembe, azonban ezek a pozíciók legtöbbször nem egyértelműek. A biológiailag releváns eredmények érdekében valamilyen megközelítéssel pozíció-szintű súlyozást érdemes bevezetni.

Ennek egy lehetséges módja a PSSM alapú konzerváltság számolás, mely során egy adott régió evolúciós konzerváltságának értékeléséhez a pozíció-szintű súlyozáshoz az információt ismert előfordulásokból nyerjük ki. Ezt a megközelítést alkalmaztam a *H. sapiens* LC8 dinein könnyűlánc fehérje kötőmotívumainak konzerváltsági vizsgálata során. Vizsgálatom feltárta, hogy a *H. sapiens* LC8 fehérje kötőmotívumai a gerinceseken túli organizmusokban ritkán konzerváltak (8. ábra). Ez a megfigyelés összhangban van a lineáris motívumok “ex-nihilo” evolúciós elméletével, mely szerint ezek a funkcionális motívumok amilyen egyszerűen megjelennek olyan egyszerűséggel el is veszhetnek néhány mutáció által [68]. Ugyanakkor, ezek az eredmények szemben állnak a *H. sapiens* LC8 fehérje gerinces és nem-gerinces fajokban is megfigyelhető erős evolúciós konzerváltsággal.

Megfigyeltem azt is, hogy az ismert *H. sapiens* motívumok döntő többsége a saját taxonómiai szintjükön vagyis, az emlős szinten belül, legalább 80%-ban konzerváltak voltak. Ez azonban nem volt igaz egy tetszőlegesen predikált kötőmotívumra. Ez alapján kifejlesztettem egy új, motívum azonosítás során a feltételezhetően nem valós

találatok kiszűrésére alkalmazható eljárást. A módszer lényege, hogy valószínű biológiai relevanciával nem rendelkeznek azon *H. sapiens* találatok, melyek az emlős ortológokban nem mutatnak konzerváltságot. A szűrési kritériumot az LC8 kötőpartnereinek predikciója során teszteltem, ahol az evolúciós szűrési kritérium ~50%-kal redukálta a potenciális motívum találatok számát. Megvizsgálva a rendezetlen régiók evolúciós konzerváltságának egyedi, sziget-szerű mintázatának motívum detektálásra alkalmazhatóságát az ismert LC8 *H. sapiens* halmazon, a motívumok csak 32.5%-át lehetett "újra" azonosítani. Ez az eredményem arra utal, hogy a szigetszerűség stratégiája önmagában gyenge megközelítés az LC8 rendszer esetében, annál eredményre vezetőbb az evolúciós alapú szűrési kritériumot alkalmazni.

A PSSM alapú konzerváltság számolás során a pozíciók súlyozottan értékeltek, melyet a kiindulási szekvencia halmaz biztosít. A másik témakörben – a rákos megbetegedések kialakulásával kapcsolatban álló rendezetlen régiók – elvégzett evolúciós vizsgálatokhoz a konzerváltság számolási súlyokat egy másik megközelítésen alapulva vettem figyelembe. Ennek lényege, hogy a régiók pozíció szintű mutációs adatai kiemelik a funkció szempontjából legfontosabb pozíciókat, melyeket a konzerváltsági információ szempontjából szükséges figyelembe venni. Az elvégzett vizsgálatok során is ezt az alap stratégiát alkalmaztam.

A projekt során evolúciós konzerváltság alapján meghatároztam a rákos megbetegedésekkel direkt kapcsolatban álló rendezetlen régiók evolúciós eredetét. A munka egyik lényeges újdonsága, hogy az eredet vizsgálat megközelítés nem gén-centrikus, hanem régió specifikusan, funkcionális egységeken (mint pl. SLiM vagy rendezetlen linker) alapul. A gén-centrikus megközelítés egy-egy konzervált, ősi domén megjelenésén alapszik, figyelmen kívül hagyva, hogy a gének evolúciós változásai sokszor csak egy-egy régiót érintve történik.

Az általam kifejlesztett módszert alkalmazva kapott eredményeim azt mutatták, hogy a rákos betegségek szempontjából direktben érintett rendezetlen régiók nagyfokú evolúciós konzerváltsággal jellemezhető ősi eredetre vezethetők vissza (12. ábra), mely kiemeli a kritikusan fontos funkcióikat. A régiók eredetére vonatkozó eredmények egybevágnak a rákos megbetegedésekkel kapcsolatos gének általános evolúciós megjelenésével [96]. Azonban vizsgálataim arra is rámutattak, hogy a régió- és gén-szintű eredet eltérhet egymástól, sok esetben a vizsgált rendezetlen régiók megjelenése jóval későbbre tehető, mint az azt kódoló gén és géncsalád eredete.

Bár a régiók feltűnésük után gyorsan rögzülnek és ezután kevés változást mutatnak, gén-szinten további változások is megfigyelhetők, melyek az evolúció folytonosságára utalnak. Sok esetben például a gén többszörös duplikáción ment keresztül, regulációs régiókkal bővült vagy pozíció szintű változások által funkcionális finomhangoláson esett át. Eredményeim során bemutatásra került egy ilyen, evolúció biológia szempontjából fontos eset, az ESR1, mely evolúciója során egy *H. sapiens* specifikus funkcionális finomhangolást azonosítottam (19. ábra). Evolúció biológiai szempontból érdekes eredmény, hogy a vizsgált rendezetlen régiók a gén duplikációktól függetlenül jelennek meg az evolúció során. Ez a megfigyelés ellentmond az általánosan elfogadott neofunkcionalizáció modelljének, miszerint az új molekuláris funkciók a duplikációk által indukáltak.

Bár az általános megfigyelés szerint a rendezetlen fehérjék általában kevésbé konzerváltak [149], a funkcionális rendezetlen régiókra elvégzett vizsgálataim összességében egy nagyon változatos evolúciós konzerváltsággal leírható képet mutattak. Megfigyelhetők voltak faj (pl. a *H. sapiens* specifikus foszforilációs PTM)- és evolúciós szintre-specifikus (metazoa fajok LC8 fehérjével kölcsönható EGL fehérjéje - 9. ábra) lineáris motívumok, emellett megfigyeltem egy, az immunrendszer folyamataiban szerepet játszó, evolúciós értelemben nagyon fiatal lineáris motívumot (CD79B vizsgált régiója – 12. ábra). Érdekes módon több extrém konzervált

funkcionális rendezetlen régiót is azonosítottam (12. ábra). Ezt jól mutatja az MLH1 (16. ábra) és DYNC1|1/DYNC1|2 fehérjék (8. ábra), amelyek esetében a lineáris motívumok az eukarióta egysejtűek szintjén is megfigyelhetőek. Emellett, az LC8 rendszerén keresztül láthattuk azt is, hogy még egy adott kölcsönhatási hálózat esetében is nagyon eltérő konzerváltsági viselkedést tapasztalhatunk.

Vizsgálataim a rendezetlen fehérjék tanulmányozását egy érdekes oldalról, az evolúciójuk irányából közelítette meg. Értekezésem során láthattuk, hogy az evolúció eszközként alkalmazása egy hatékony stratégia a rendezetlen fehérjékkel kapcsolatos problémák megoldására és kérdések megválaszolására, ami egyben rávilágít ezen kutatási terület jelentőségére is.

Összefoglaló

A rendezetlen fehérjék nem rendelkeznek jól meghatározott térszerkezettel, ennek ellenére esszenciális szerepet töltenek be számos szabályozási és jelátviteli folyamatban. Ebből adódóan a rendezetlen fehérjék hibás működését rákos megbetegedések kialakulásával is összefüggésbe hozták. A rendezetlen fehérjék mind szerkezeti, mind funkcionális szempontból rendkívül sokfélék, azonban az, hogy ez hogyan tükröződik evolúciós tulajdonságaikban, a rendezett fehérjékhez képest még jóval kevésbé feltárt. Doktori kutatómunkám során a rendezetlen fehérje régiókat evolúciójuk szempontjából vizsgáltam, két külön témakörben. Ezekhez olyan új módszereket dolgoztam ki, amelyek figyelembe vették a rendezetlen szegmensek egyedi evolúciós tulajdonságait.

A rendezetlen fehérjék egyik legfontosabb funkcionális egységei a speciális tulajdonságaik miatt csak kis számban ismert ún. lineáris motívumok. Habár felderítésük napjainkban a fehérje bioinformatika egyik fő kutatási területe, az egyszerűségükből adódó magas fals pozitív találatok száma megnehezíti munkánkat. Első alapvető célom egy evolúciós konzerváltság alapú motívum szűrő módszer kifejlesztése volt. Ezt az LC8 dinein könnyűlánc fehérje interakciós hálózatának bővítése során alkalmaztam. Megközelítésemmel ~50%-kal csökkentettem a potenciálisan nem valós találatok számát, ami hozzájárult az LC8 rendszerben biológiailag releváns új partnerek azonosításához.

Második fő célkitűzésem a rákos megbetegedésekért közvetlenül felelős rendezetlen régiók evolúciójának részletes tanulmányozása volt. Többek között megállapítottam, hogy a rendezetlen fehérjék tipikusan alacsony konzerváltsága ellenére, a rákban direkt módon mutálódó rendezetlen régiók erősen konzerváltak és sok esetben már az ősi egysejtűekben funkcionálisak lehettek. Ezen megfigyelések a régiók működési mechanizmusának megismeréséhez alapjaiban járulnak hozzá, jobban megértve ezzel a régiók tumorgenezishez köthető diszfunkcionalitásukat.

Summary

Although intrinsically disordered proteins do not have a well-defined three dimensional structure, they play essential roles in many signalling and regulatory processes. In agreement with this, malfunction of disordered proteins has been shown to be associated with various types of cancer. Disordered proteins are heterogeneous both from structural and functional points of view, however, compared to globular proteins it is much less explored how their specific features are reflected in their evolutionary properties. During my PhD research I have studied disordered protein segments from an evolutionary point of view, focusing on two specific projects. To this end I have developed novel methods that take the unique evolution of disordered segments into consideration.

One of the most important functional units of disordered proteins are the so-called linear motifs. Despite their importance, the number of known linear motifs is still limited due to their special features. Although finding new linear motifs is one of the most important challenges in protein bioinformatics nowadays, their research is hampered by the high number of false positives which occur due to their simplicity of the motifs. My first aim was to develop a filtering method for motif prediction based on their evolutionary conservation. I utilised this method to expand the network of interaction partners for the dynein light chain protein LC8. Using my protocol, I could reduce the number of potentially false positive hits by about 50%, which greatly helped to identify biologically relevant novel interaction partners in the LC8 system.

My second aim was to investigate the disordered segments that play a direct role in cancer. I have shown that despite their typically low conservation, those disordered segments that are directly mutated in cancer show strong evolutionary conservation, with many of them becoming functional in ancient unicellular organisms. These

discoveries fundamentally contribute to our knowledge in understanding how disordered segments function, and what their role is in tumorigenesis.

Irodalomjegyzék

1. Wootton, J. C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285 (1994).
2. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571 (1996).
3. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
4. Dunker, A. K. *et al.* Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* 473–484 (1998).
5. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
6. Tompa, P., Dosztanyi, Z. & Simon, I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J. Proteome Res.* **5**, 1996–2000 (2006).
7. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
8. Gong, H., Zhang, S., Wang, J., Gong, H. & Zeng, J. Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data. *J. Comput. Biol.* **23**, 300–310 (2016).
9. Dyson, H. J. & Wright, P. E. Equilibrium NMR studies of unfolded and partially

folded proteins. *Nat. Struct. Biol.* **5 Suppl**, 499–503 (1998).

10. Wright, P. E., Dyson, H. J. & Lerner, R. A. Conformation of peptide fragments of proteins in aqueous solution: implications for initiation of protein folding. *Biochemistry* **27**, 7167–7175 (1988).
11. Mukhopadhyay, S., Krishnan, R., Lemke, E. A., Lindquist, S. & Deniz, A. A. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2649–2654 (2007).
12. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
13. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **35**, D786–93 (2007).
14. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276 (2020).
15. Radivojac, P. *et al.* Intrinsic disorder and functional proteomics. *Biophys. J.* **92**, 1439–1456 (2007).
16. He, B. *et al.* Predicting intrinsic disorder in proteins: an overview. *Cell Res.* **19**, 929–949 (2009).
17. Dosztányi, Z., Mészáros, B. & Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief. Bioinform.* **11**, 225–243 (2010).
18. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and

intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839 (2005).

19. Galea, C. A., Wang, Y., Sivakolundu, S. G. & Kriwacki, R. W. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* **47**, 7598–7609 (2008).
20. Uversky, V. N. The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS Lett.* **587**, 1891–1901 (2013).
21. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002).
22. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579**, 3346–3354 (2005).
23. Feng, Z., Chen, X., Wu, X. & Zhang, M. Formation of biological condensates via phase separation: Characteristics, analytical methods, and physiological implications. *J. Biol. Chem.* **294**, 14823–14835 (2019).
24. Fuxreiter, M. & Tompa, P. Fuzzy complexes: a more stochastic view of protein function. *Adv. Exp. Med. Biol.* **725**, 1–14 (2012).
25. Tompa, P. *et al.* Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* **31**, 328–335 (2009).
26. Dosztányi, Z., Chen, J., Dunker, A. K., Simon, I. & Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* **5**, 2985–2995 (2006).
27. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

28. Barabási, A.-L. & Bonabeau, E. Scale-free networks. *Sci. Am.* **288**, 60–69 (2003).
29. Uversky, V. N. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front. Biosci.* **14**, 5188–5238 (2009).
30. Deng, C.-X. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res.* **34**, 1416–1426 (2006).
31. Joerger, A. C. & Fersht, A. R. Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem.* **77**, 557–582 (2008).
32. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002).
33. Pajkos, M., Mészáros, B., Simon, I. & Dosztányi, Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol. Biosyst.* **8**, 296–307 (2012).
34. Shapiro, L. The multi-talented beta-catenin makes its first appearance. *Structure* **5**, 1265–1268 (1997).
35. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
36. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–301 (2012).
37. Funayama, N., Fagotto, F., McCrea, P. & Gumbiner, B. M. Embryonic axis induction by the armadillo repeat domain of beta-catenin: evidence for intracellular

signaling. *J. Cell Biol.* **128**, 959–968 (1995).

38. Mészáros, B., Hajdu-Soltész, B., Zeke, A. & Dosztányi, Z. Intrinsically disordered protein mutations can drive cancer and their targeted interference extends therapeutic options. *bioRxiv* 2020.04.29.069245 (2020) doi:10.1101/2020.04.29.069245.
39. Mészáros, B., Zeke, A., Reményi, A., Simon, I. & Dosztányi, Z. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. *Biol. Direct* **11**, 23 (2016).
40. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
41. Domazet-Loso, T. & Tautz, D. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.* **25**, 2699–2707 (2008).
42. Stein, A. & Aloy, P. Contextual specificity in peptide-mediated protein interactions. *PLoS One* **3**, e2524 (2008).
43. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).
44. Davey, N. E., Travé, G. & Gibson, T. J. How viruses hijack cell regulation. *Trends Biochem. Sci.* **36**, 159–169 (2011).
45. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, W362–5 (2006).
46. Schiller, M. R. Minimotif miner: a computational tool to investigate protein function,

disease, and genetic diversity. *Curr. Protoc. Protein Sci.* **Chapter 2**, Unit 2.12 (2007).

47. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
48. Dinkel, H. *et al.* ELM--the database of eukaryotic linear motifs. *Nucleic Acids Res.* **40**, D242–51 (2012).
49. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist. *Mol. Cell* **55**, 161–169 (2014).
50. Blikstad, C. & Ivarsson, Y. High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Commun. Signal.* **13**, 38 (2015).
51. Rapali, P. *et al.* DYNLL/LC8: a light chain subunit of the dynein motor complex and beyond. *FEBS J.* **278**, 2980–2996 (2011).
52. Barbar, E. Dynein light chain LC8 is a dimerization hub essential in diverse protein networks. *Biochemistry* **47**, 503–508 (2008).
53. Dick, T., Ray, K., Salz, H. K. & Chia, W. Cytoplasmic dynein (ddlc1) mutations cause morphogenetic defects and apoptotic cell death in Drosophila melanogaster. *Mol. Cell. Biol.* **16**, 1966–1977 (1996).
54. Kamath, R. S. *et al.* Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231–237 (2003).
55. Clark, S. A., Jespersen, N., Woodward, C. & Barbar, E. Multivalent IDP assemblies:

Unique properties of LC8-associated, IDP duplex scaffolds. *FEBS Lett.* **589**, 2543–2551 (2015).

56. Fan, J., Zhang, Q., Tochio, H., Li, M. & Zhang, M. Structural basis of diverse sequence-dependent target recognition by the 8 kDa dynein light chain. *J. Mol. Biol.* **306**, 97–108 (2001).
57. Hall, J., Karplus, P. A. & Barbar, E. Multivalency in the assembly of intrinsically disordered Dynein intermediate chain. *J. Biol. Chem.* **284**, 33115–33121 (2009).
58. Williams, J. C. *et al.* Structural and thermodynamic characterization of a cytoplasmic dynein light chain-intermediate chain complex. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10028–10033 (2007).
59. Radnai, L. *et al.* Affinity, avidity, and kinetics of target sequence binding to LC8 dynein light chain isoforms. *J. Biol. Chem.* **285**, 38649–38657 (2010).
60. Rapali, P. *et al.* Directed evolution reveals the binding motif preference of the LC8/DYNLL hub protein and predicts large numbers of novel binders in the human proteome. *PLoS One* **6**, e18818 (2011).
61. Jespersen, N. *et al.* Systematic identification of recognition motifs for the hub protein LC8. *Life Sci Alliance* **2**, (2019).
62. Erdős, G. *et al.* Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway. *PLoS Comput. Biol.* **13**, e1005885 (2017).
63. Pfister, K. K., Fay, R. B. & Witman, G. B. Purification and polypeptide composition of dynein ATPases from Chlamydomonas flagella. *Cell Motil.* **2**, 525–547 (1982).
64. Brown, C. J. *et al.* Evolutionary rate heterogeneity in proteins with long disordered

regions. *J. Mol. Evol.* **55**, 104–110 (2002).

65. Liu, J., Zhang, Y., Lei, X. & Zhang, Z. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol.* **9**, R69 (2008).
66. Bellay, J. *et al.* Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* **12**, R14 (2011).
67. Babu, M. M., Kriwacki, R. W. & Pappu, R. V. Structural biology. Versatility from protein disorder. *Science* **337**, 1460–1461 (2012).
68. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **13**, 43 (2015).
69. Nguyen Ba, A. N. *et al.* Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci. Signal.* **5**, rs1 (2012).
70. Davey, N. E., Shields, D. C. & Edwards, R. J. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics* **25**, 443–450 (2009).
71. Davey, N. E. *et al.* SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res.* **40**, 10628–10641 (2012).
72. Eddy, S. R. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 114–120 (1995).
73. Altschul, S. F. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565 (1991).

74. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* **3**, 66–70 (1991).

75. Altschul, S. F. A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36**, 290–300 (1993).

76. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).

77. Altschul, S. F. Generalized affine gap costs for protein sequence alignment. *Proteins* **32**, 88–96 (1998).

78. Zachariah, M. A., Crooks, G. E., Holbrook, S. R. & Brenner, S. E. A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins* **58**, 329–338 (2005).

79. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

80. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 4412–4415 (1989).

81. Feng, D. F. & Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360 (1987).

82. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

83. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*

30, 3059–3066 (2002).

84. Katoh, K. & Toh, H. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics* **23**, 372–374 (2007).
85. Perrodou, E., Chica, C., Poch, O., Gibson, T. J. & Thompson, J. D. A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics* **9**, 213 (2008).
86. Katoh, K. & Standley, D. M. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **32**, 1933–1942 (2016).
87. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).
88. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).
89. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
90. Davey, N. E., Edwards, R. J. & Shields, D. C. The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.* **35**, W455–9 (2007).
91. Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. Computational methods for Gene Orthology inference. *Brief. Bioinform.* **12**, 379–391 (2011).
92. Huerta-Cepas, J., Bueno, A., Dopazo, J. & Gabaldón, T. PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.* **36**, D491–6

(2008).

93. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, (2016).

94. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42**, D897–902 (2014).

95. Gabaldón, T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* **9**, 235 (2008).

96. Domazet-Loso, T. & Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* **8**, 66 (2010).

97. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).

98. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

99. Anisimova, M., Bielawski, J. P. & Yang, Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* **18**, 1585–1592 (2001).

100. Jeffares, D. C., Tomiczek, B., Sojo, V. & dos Reis, M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol. Biol.* **1201**, 65–90 (2015).

101. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**,

2472–2479 (2005).

102. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).

103. Gayà-Vidal, M. & Albà, M. M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15**, 599 (2014).

104. Afanasyeva, A., Bockwoldt, M., Cooney, C. R., Heiland, I. & Gossmann, T. I. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* **28**, 975–982 (2018).

105. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

106. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

107. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

108. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).

109. Rigoutsos, I. & Floratos, A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**, 55–67 (1998).

110. Krystkowiak, I., Manguy, J. & Davey, N. E. PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Res.* **46**, W235–W241 (2018).

111. Davey, N. E., Edwards, R. J. & Shields, D. C. Estimation and efficient computation of the true probability of recurrence of short linear protein sequence motifs in

unrelated proteins. *BMC Bioinformatics* **11**, 14 (2010).

112. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

113. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

114. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

115. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

116. Altenhoff, A. M. *et al.* Standardized benchmarking in the quest for orthologs. *Nat. Methods* **13**, 425–430 (2016).

117. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–8 (2008).

118. Proux, E., Studer, R. A., Moretti, S. & Robinson-Rechavi, M. Selectome: a database of positive selection. *Nucleic Acids Res.* **37**, D404–7 (2009).

119. Gao, F. *et al.* EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecol. Evol.* **9**, 3891–3898 (2019).

120. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

121. Gil, M., Zanetti, M. S., Zoller, S. & Anisimova, M. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. *Mol. Biol. Evol.* **30**, 1270–1280 (2013).

122. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

123. Dinkel, H. *et al.* ELM 2016--data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **44**, D294–300 (2016).

124. Schäffer, A. A. *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).

125. Chica, C., Labarga, A., Gould, C. M., López, R. & Gibson, T. J. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* **9**, 229 (2008).

126. Edwards, R. J., Davey, N. E. & Shields, D. C. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* **2**, e967 (2007).

127. Nyarko, A. *et al.* Conformational dynamics promote binding diversity of dynein light chain LC8. *Biophys. Chem.* **159**, 41–47 (2011).

128. Wang, W. *et al.* Defining the protein-protein interaction network of the human hippo pathway. *Mol. Cell. Proteomics* **13**, 119–131 (2014).

129. Bossuyt, W. *et al.* An evolutionary shift in the regulation of the Hippo pathway between mice and flies. *Oncogene* **33**, 1218–1228 (2014).

130. van den Berg, T. K., Yoder, J. A. & Litman, G. W. On the origins of adaptive immunity: innate immune receptors join the tale. *Trends Immunol.* **25**, 11–16 (2004).

131. Moretti, S. *et al.* Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res.* **42**, D917–21 (2014).

132. Shcherbakova, P. V. & Kunkel, T. A. Mutator phenotypes conferred by MLH1 overexpression and by heterozygosity for mlh1 mutations. *Mol. Cell. Biol.* **19**, 3177–3183 (1999).

133. Takahashi, M. *et al.* Functional analysis of human MLH1 variants using yeast and in vitro mismatch repair assays. *Cancer Res.* **67**, 4595–4604 (2007).

134. Ohsawa, T. *et al.* Colorectal cancer susceptibility associated with the hMLH1 V384D variant. *Mol. Med. Rep.* **2**, 887–891 (2009).

135. Lee, S. E. *et al.* High prevalence of the MLH1 V384D germline mutation in patients with HER2-positive luminal B breast cancer. *Sci. Rep.* **9**, 10966 (2019).

136. Gueneau, E. *et al.* Structure of the MutLα C-terminal domain reveals how Mlh1 contributes to Pms1 endonuclease site. *Nat. Struct. Mol. Biol.* **20**, 461–468 (2013).

137. Kamura, T. *et al.* VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases. *Genes Dev.* **18**, 3055–3065 (2004).

138. Cardote, T. A. F., Gadd, M. S. & Ciulli, A. Crystal Structure of the Cul2-Rbx1-EloBC-VHL Ubiquitin Ligase Complex. *Structure* **25**, 901–911.e3 (2017).

139. Sutovsky, H. & Gazit, E. The von Hippel-Lindau tumor suppressor protein is a molten globule under native conditions: implications for its physiological activities.

J. Biol. Chem. **279**, 17190–17196 (2004).

140. Cai, Q. & Robertson, E. S. Ubiquitin/SUMO modification regulates VHL protein stability and nucleocytoplasmic localization. *PLoS One* **5**, (2010).

141. Minervini, G. *et al.* Isoform-specific interactions of the von Hippel-Lindau tumor suppressor protein. *Sci. Rep.* **5**, 12605 (2015).

142. Dhayalan, A., Kudithipudi, S., Rathert, P. & Jeltsch, A. Specificity analysis-based identification of new methylation targets of the SET7/9 protein lysine methyltransferase. *Chem. Biol.* **18**, 111–120 (2011).

143. Wang, C. *et al.* Direct acetylation of the estrogen receptor alpha hinge region by p300 regulates transactivation and hormone sensitivity. *J. Biol. Chem.* **276**, 18375–18383 (2001).

144. Wang, R.-A., Mazumdar, A., Vadlamudi, R. K. & Kumar, R. P21-activated kinase-1 phosphorylates and transactivates estrogen receptor-alpha and promotes hyperplasia in mammary epithelium. *EMBO J.* **21**, 5437–5447 (2002).

145. Michalides, R. *et al.* Tamoxifen resistance by a conformational arrest of the estrogen receptor alpha after PKA activation in breast cancer. *Cancer Cell* **5**, 597–605 (2004).

146. Rust, H. L. & Thompson, P. R. Kinase consensus sequences: a breeding ground for crosstalk. *ACS Chem. Biol.* **6**, 881–892 (2011).

147. de Leeuw, R. *et al.* PKA phosphorylation redirects ERα to promoters of a unique gene set to induce tamoxifen resistance. *Oncogene* **32**, 3543–3551 (2013).

148. Atsriku, C. *et al.* Systematic mapping of posttranslational modifications in human

estrogen receptor-alpha with emphasis on novel phosphorylation sites. *Mol. Cell. Proteomics* **8**, 467–480 (2009).

149. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446 (2011).

ADATLAP

a doktori értekezés nyilvánosságra hozatalához*

I. A doktori értekezés adatai

A szerző neve: Pajkos Mátyás

MTMT-azonosító: 10034855

A doktori értekezés címe és alcíme: A rendezetlen fehérjék evolúciós vizsgálatai

DOI-azonosító[46]: 10.15476/ELTE.2020.106

A doktori iskola neve: Biológia Doktori Iskola

A doktori iskolán belüli doktori program neve: Szerkezeti biokémia

A témavezető neve és tudományos fokozata: Dr. Dosztányi Zsuzsanna, Ph.D.

A témavezető munkahelye: ELTE, biokémia tanszék

II. Nyilatkozatok

1. A doktori értekezés szerzőjeként
a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a Természettudományi kar Dékáni Hivatal Doktori, Habilitációs és Nemzetközi Ügyek Csoportjának ügyintézőjét, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.
b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;
c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés (*dátum*)-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;
d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.
2. A doktori értekezés szerzőjeként kijelentem, hogy
a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;
b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.
3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: 2020.07.15

Pajkos Mátyás
.....
a doktori értekezés szerzőjének aláírása

**ELTE SZMSZ SZMR 12. sz. melléklet*