

36. OTDK Biológia Szekció

EÖTVÖS LÓRÁND TUDOMÁNY EGYETEM



**DisCanVis: Integrált szerkezeti és funkcionális annotációk vizualizációja a
rendezetlen fehérjékben található rákos mutációk hatásának jobb
megértéséhez**

Készítette: Deutsch Norbert

Biológus MSc (2. évfolyam)

Témavezetők:

Dr. Dosztányi Zsuzsanna

Erdős Gábor

Tartalomjegyzék

1. Bevezető

- 1.1 Rák genom projektek
- 1.2 Rendezetlen fehérjék
- 1.3 Rendezetlen fehérjék és a rák kapcsolata

2. Célkitűzés

3. Anyagok és módszerek

- 3.1 Szekvencia és mutációs adatok összegyűjtése
- 3.2 Annotációk összegyűjtése
- 3.3 Adatok összehangolása, SQL táblák, Webszerver
- 3.4 Programozási nyelvek, keretrendszerek

4. Eredmények és megvitatásuk

- 4.1 Eredmény összefoglaló
- 4.2 Statisztikák
- 4.3 Webszerver
 - 4.3.1 Kereső, Táblák, Rólunk, oldalak
 - 4.3.2 Fehérje vizualizáció
 - 4.3.2.1 Fejléc, funkciós gombok
 - 4.3.2.2 Teljes fehérje összesítő ábra
 - 4.3.2.3 Fő vizualizációs ábra
- 4.4 Példák
 - 4.4.1 RAF1
 - 4.4.2 Ataxin-1, Ismeretlen rákos gének azonosítása
 - 4.4.3 WNK család, ismeretlen rákos gén vagy egyéb betegség azonosítása paralógok által
- 4.5 Konklúzió

5. Összefoglalás

6. Köszönetnyilvánítás

7. Irodalom

1. Bevezető

1.1 Rák genom projektek

A rák egy többgénés összetett betegség.¹ Egyes gének sokfajta ráktípusban szerepet játszhatnak, de gének egy csoportja akár specializáltan, csak egyes típusokban is okozhat betegséget. Alapvetően azoknak a géneknek a mutációja okoz daganatos elváltozást, amelyek közvetve vagy közvetlenül fontos szignalizációs útvonalakat befolyásolnak és a sejtosztódást, valamint a növekedést segítik, vagy az immunválaszt gátolják. A rák kialakulásában szerepet játszó géneket feloszthatjuk onkogénekre és tumorszupresszorokra. A tumorszupresszorok funkciójuk inaktiválásával segítik a rák kialakulását, míg az onkogének a túlműködésükkel vagy túlexpressziójuk révén. Az elmúlt években a rákkal diagnosztizált betegek genomjának a megszekvenálásának köszönhetően rengeteg adat vált elérhetővé a kutatók számára. Ennek nyomán számos új rák gént tudtak azonosítani. A rákos mintákban megfigyelt genetikai elváltozások elérhetőek nyilvános adatbázison keresztül, mint például a Catalogue Of Somatic Mutations In Cancer (COSMIC)² vagy a The Cancer Genome Atlas (TCGA)³.

A genomon belül a mutációk bárhol megjelenhetnek, de a mutációknak csak egy része felelős közvetlenül a betegség kialakulásáért. Azokat a mutációkat, amelyek olyan funkcionális területre esnek, melyeknél bizonyítható, hogy funkciójuk megváltozásával indukálják a rák kialakulását, driver mutációknak nevezzük. Az összes többi mutációt passenger mutációnak hívjuk. Az egyik legnagyobb kihívás driver mutációkat tartalmazó gének (driver gének) azonosítása¹. Az egyes driver gének többfajta mechanizmusokon keresztül is sérülhetnek, például változhat expressziós szintjük, vagy trunkálódhatnak. A leggyakoribb mutáció típus a kódoló régiókat érintő pontmutációk illetve a kisebb, a leolvasási keret elcsúszását nem okozó inzerciók és delécioók. Ezek hatása direkt módon összekapcsolható a fehérje funkcionális régiójával, és általában az ismert gyógyszercélpontok ezeknek a mutációknak a hatását próbálják meg kompenzálni.

A jelenleg ismert gyógyszercélpontok jól definiált térszerkezettel rendelkező fehérjék, például kinázok. Ma már tudjuk, hogy vannak olyan fehérjék, amik nem rendelkeznek jól definiált térszerkezettel, mégis fontos biológiai folyamatok vesznek részt, ezek az úgynevezett rendezetlen fehérjék. Ezeknek a fehérjéknek szerepük lehet a rák kialakulásában is.

1.2 Rendezetlen fehérjék

A fehérjékutatások sokáig a szerkezet-funkció paradigmán alapultak, mely szerint csak a jól definiált struktúrával rendelkező fehérjék és azok régiói láthatnak el biológiai funkciókat. Az utóbbi években bekövetkezett egy szemléletváltozás, amit a rendezetlen fehérjék alapoztak meg. A rendezetlen fehérjék nem vesznek fel jól definiált térszerkezetet, ellenben fontos biológiai funkciókat látnak el. A humán proteom körülbelül 50 százalékát olyan fehérjék alkotják, amelyek részben vagy egészében rendezetlenek.⁴

A rendezetlen régiók funkciói igen változatosak, lehetnek “linkerek”, melyek flexibilis kapcsolatot teremtenek két domain között, de gyakran vesznek részt molekuláris felismerésben is. A rendezetlen kötő régiók kötődésük alatt képesek lehetnek rendezetlen-rendezett átmenetre, de vannak olyan rendezetlen régiók, amelyek kötődésük során sem vesznek fel rendezett szerkezetet. A rendezetlen fehérjék gyakran tartalmaznak rövid lineáris motívumokat (SLIMs).⁵ Ezek kompakt funkcionális modulok, amelyek 3-11 aminosav hosszúak és főként molekuláris felismerő helyként funkcionálnak. A SLIM-ek működését gyakran befolyásolják különböző poszttranszlációs módosítások, melyek révén a molekuláris kapcsolók alakulnak ki. A legújabb eredmények rámutattak arra, hogy a rendezetlen fehérjék a membrán nélküli organellek kialakításában is szerepet játszanak, fázis szeparáció révén.⁶

A rendezetlen fehérjék kutatását számos adatbázis és bioinformatikai eszköz segíti. Ezek közül központi szereppel bír a DisProt⁷ adatbázis, ami a kísérletesen igazolt rendezetlen fehérjéket gyűjti össze. Bár az adatbázis jelenleg több mint kétezer annotációt tartalmaz, ezek csak a rendezetlen fehérjék csak nagyon kis részét képviselik. Számos predikciós módszer létezik a rendezetlenség aminosav szekvenciából történő becslésére. Az egyik legelterjedtebb ilyen módszer a csoportunk által kifejlesztett IUPred.⁸ Jelenleg az egyik legjobb rendezetlenség jósló módszer a térszerkezet predikció terén áttörő eredményt elérő AlphaFold2 módszer pLDDT megbízhatósági értékein alapul.⁹

A rendezetlen régiók ortológokban vizsgált konzerváltság szempontjából változatos képet mutatnak, mely alapján három kategóriába sorolhatók. Az első csoport a “constrained disorder”, amely esetében a régiókra az jellemző, hogy mind a szekvencia, mind maga a rendezetlenség is konzervált. A második kategória a “flexible disorder”, mely osztály régiói

ugyan nagymértékű szekvencia divergenciát mutatnak, azonban a régió rendezetlensége, mint szekvencia tulajdonság konzervált. A “non-conserved” kategóriát azok az esetek adják, amikor az adott rendezetlen régió aminosav szinten se és a rendezetlenség szempontjából se konzervált. Az evolúciós konzerváltság szempontjából a SLiM-ek egy speciális kategóriába esnek, nagyon változatos konzerváltságot mutatva, mely a specifikus tulajdonságaikból adódik. Megfigyelhetők esetek amikor egy SLiM csupán egy vagy két pozíciója konzervált, de olyan esetet is találunk amikor több motívum átfedve alakít ki egy hosszabb funkcionális régiót, amely pozícióinak túlnyomó többsége nagymértékben konzervált. A SLiM-ek egyedi tulajdonsága, hogy többszörös szekvencia illesztésben, sok esetben konzerváltabbak mint szekvenciális környezetük, ún. szigetszerű konzerváltságot mutatnak, mely tulajdonság egyik fontos alapját adja az evolúciós megközelítésű, bioinformatikai SLiM keresésnek.¹⁰

1.3. Rendezetlen fehérjék és a rák kapcsolata

A rendezetlen fehérjék fontos biológiai folyamatokban vesznek részt és központi szerepet játszanak számos jelátviteli és szabályozó folyamatban.¹¹ Ezért nem meglepő, hogy a rákban, illetve más betegségekben is kulcsfontosságúak. Kutatócsoportunk nemrégiben rámutatott arra, hogy az ismert rákos gének 20 százalékában a mutációk elsődleges célpontja rendezetlen régió. Ezek a rendezetlen driver gének a kontextusfüggő interakciós hálózatok központjában helyezkednek el és olyan folyamatokban vesznek részt, mint a transzkripció, génexpresszió szabályozása és a fehérjék degradációja.¹² A rendezetlen fehérjék és azok régióinak rákban betöltött szerepére újabban vizsgálatokat végeztek és azt is bebizonyították, hogy többek között a fázis szeparáció is szerepet játszhat a rák kialakulásában.¹³ Attól függően, hogy a mutáció rendezett vagy rendezetlen régióra esik, összefüggéseket találtak a génexpressziós szint változása, illetve a betegek túlélése vonatkozásában.

Az egyik legjobban tanulmányozott rendezetlen rák gén a béta-catenin, ami a Wnt jelátvitel útvonal kanonikus útvonalának egy központi fehérjéje. Ez a fehérje egy jól ismert onkogén, aminek a mutációi az N-terminális végén, egy rendezetlen, rövid szakaszon sűrűsödnek. A régió mutációja a fehérje felhalmozódásához vezet, ami révén a jelátviteli útvonal egyensúlya felbomlik, ami a célgének túlexpresszáldásához vezet.¹⁴ Az újabb vizsgálatok rámutattak arra, hogy számos további esetben megfigyelhető, hogy a rendezetlen részekben halmozódnak fel mutációk, azonban ezek általában jóval kevésbé karakterizáltak.¹² A rendezetlen fehérjék és a rákos mutációk közötti kapcsolat vizsgálatát nagyban megnehezíti, hogy nem áll rendelkezésre egy olyan eszköz, ahol a rendezetlen régiókra vonatkozó

szerkezeti és funkcionális tulajdonságokat a mutációkkal együtt lehetne vizsgálni. A COSMIC adatbázisban megtalálható vizualizációs eszköz jó alapot ad a mutációs eloszlások és a gócpontok azonosítására, míg a ProViz¹⁵ a rendezetlenség megállapítására és a funkcionális annotációkra megfelelő, amiből viszont a mutációs adatok hiányoznak. Jelenleg nem található olyan vizualizációs webszerver ami rendezetlen fehérjéknek rákban betöltött szerepére helyezi a fókuszot.

2. Célkitűzés

Az elmúlt években a rák genom projektek rengeteg adatot kínáltak a rák molekuláris mechanizmusának jobb megértéséhez. Ezzel párhuzamosan a rendezetlen fehérjékről szerzett információink is bővült. Ennek ellenére eddig még nem állt rendelkezésre megfelelő módszer ezeknek az adatoknak az összehangolására, feldolgozására és megjelenítésére és végső soron a rendezetlen fehérjék a rákban betöltött szerepének megértésére. Célom az volt, hogy

- Létrehozzak egy komplex vizualizációs webszervert , ami a rendezetlen fehérjékre eső rákos mutációk funkcionális és strukturális annotációit tartalmazza.
- A vizualizációs eszköz felhasználásával olyan rendezetlen részeket azonosítsak, amelyek mutációja rák vagy egyéb betegségek kialakulásához vezethet.

A webszerver elősegíti hogy jobban megértsük a rendezetlen fehérjék biológiai szerepét illetve hogy mutációjuk hogyan járul hozzá rák és egyéb betegségek kialakulásához.

3. Anyagok és módszerek

3.1 Szekvencia és mutációs adatok összegyűjtése

A fehérjék központi adatbázisa az UniProt adatbázis.¹⁶ Ez két részből áll. A SwissProt tartalmazza a manuálisan leellenőrzött fehérjéket, a hozzá tartozó funkcionális és szerkezeti annotációkkal. A Trembl adatbázis az automatikusan annotált fehérjéket tartalmazza. Az adatbázisból letölthető a különböző fajok, köztük a humán, fehérje szekvenciák és az azokhoz tartozó izoformák. A Uniprotban általában a leghosszabb fehérje a fő izoforma. Munkám során a UniProt adatbázisból (2022.08) származó humán fehérje proteomot a használtam referencia szekvenciáknak.

A COSMIC adatbázis a legnagyobb és legátfogóbb forrása az emberi rákban előforduló szomatikus mutációknak. Az adatbázis tartalmazza gén specifikus és genom szintű vizsgálatok eredményeit, beleértve a TCGA projekteket is. A COSMIC (v96) adatbázisból letöltöttem a CDS szekvenciákat, ezeket átfordítottam fehérje szekvenciákra. Szintén letöltöttem a COSMIC adatbázisból a kódoló pontmutációkat. Ez egy táblázatos formában elérhető fájl egyéb információk mellett tartalmazza az egyes mutációkat, azok pozícióját fehérje szinten, az érintett gént, valamint az ahhoz tartozó kromoszómát. A mutációknál csak a missense, in-frame inzerciót és deléció, valamint a nonsense mutációkat vettük figyelembe. Bár a COSMIC adatbázis jóval nagyobb, de a TCGA projekt által generált mutációs információk letisztultabbak, ezért számításainkat ebből a projektből végeztük. Ugyanakkor a TCGA projekt a transzkriptek egy elavultabb verzióját használja. Ezért munkám során a COSMIC által újratérképezett TCGA adatokat használtam, a TCGA-s mutációkat pedig kiszűrtem a minta azonosítók alapján. Az ismert rákos géneket a COSMIC Cancer Gene Census és az IntOGen¹⁷ adatbázisokból gyűjtöttem.

Ahhoz, hogy a TCGA adatbázisban megadott mutációkat a Uniprot fehérje szekvenciákra minél pontosabban rá tudjuk vetíteni, minden UniProt szekvenciát az átfordított COSMIC fehérje szekvenciákra térképeztem. A térképezést kromoszómánként, BLAST-al végeztem.¹⁸ A BLAST egy szekvencia illesztő program, ami a egy adatbázisban lévő szekvenciák közül a legjobban egyező szekvenciákat adja vissza. A transzkripteket rátérképeztük a humán

genomra (hg38) is a UCSC Genome Browser¹⁹ command line eszközeit használva. Így a UCSC Genome Browserben a genom szinten összegyűjtött adatokat is tudtuk integrálni.

3.2 Annotációk összegyűjtése

A Uniprot adatbázis nagyszámú manuális és automatikusan generált annotációt tartalmaz fehérje szinten, illetve számos egyéb adatbázis használja referenciaként, lehetővé téve különböző információk integrációját. Az adatbázisom minden egyes szekvenciájához összegyűjtöttem a fehérje funkcionális működése szempontjából fontos annotációkat, elsősorban a “Region of interest” és “Binding region” információkat. A Uniprot azonosító segítségével tudtam annotálni az ELM²⁰ és ELM Switches²¹ adatbázisból a rövid lineáris motívumokat (SLIMs), valamint a PhosphoSitePlus²² és a dbPTM²³ adatbázisokból a poszttranszlációs módosulásokat. A UniProthoz kapcsolódó Humsavar adatbázisból az OMIM Disease²⁴ mutációk azon variánsait szedtem ki, amelyek igazoltan betegséget okoznak.

A szerkezeti információkat is több adatbázisból gyűjtöttük össze. Az ismert doméneket a Pfam²⁵ alapján a PfamScan²⁶ program segítségével szereztem meg. A kísérletesen igazolt fehérje szerkezeteket a Protein Data Bank (PDB)²⁷ adatbázisból gyűjtöttem. A MobiDB²⁸ adatbázison keresztül összegyűjtöttem a kísérletesen igazolt rendezetlen régiókat. A rendezett és rendezetlen régiók becslésére IUPred predikcióját és az AlphaFold2-ből származó pLDDT értékeket használtam. A rendezetlen kötőhelyek predikciójára az ANCHOR2-t használtam.²⁹

Az általam összegyűjtött jellemzők mellett a Dosztányi csoport tagjai további annotációkkal és módszerekkel járultak hozzá az adatbázishoz. Fehérje szintén integráltuk a rendezetlen kötő régiókat a DIBS³⁰ és az MFIB³¹ adatbázisból. A fázisszeparációs információkat a PhaSePro³² adatbázisból gyűjtöttük össze. A rendezetlenség jósló módszereket kiegészítettük egy kombinált rendezetlenség profillal, ami a kísérletes információk és az IUPred és AlphaFold módszerek alapján készít egy rendezetlenségi profilt gépi tanulásos döntési fát alkalmazva. Ezeket a munkákat Erdős Gábor végezte.

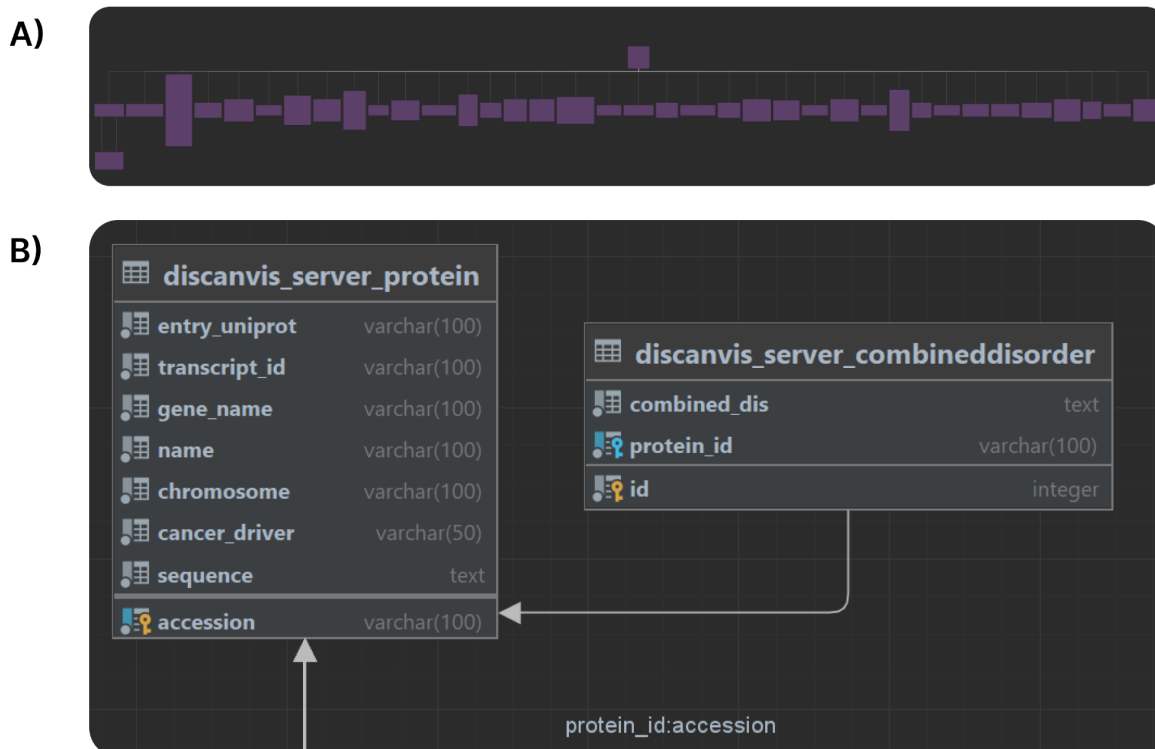
UCSC Genome Browser adatbázisát felhasználva beazonosítottuk az exon határokat, valamint összegyűjtöttük az egyponthoz nukleotid-polimorfizmusokat (SNPs) a dbSNP³³ adatbázisból és a ClinVar³⁴ mutációkat. Ezek mellé integráltuk a genomikus konzerváltsági értékeket a phastCons³⁵ programmal. Az ismétlődő genomiális régiók azonosítására a TRF módszert³⁶ használtuk. A szignifikánsan mutálódó régiók meghatározása az iSiMPre

programmal végeztük a TCGA mutációk alapján.³⁷ Az iSiMPre módszer előnye, hogy automatikusan megtalálja a mutációs határokat anélkül, hogy előre definiálnánk a régiókat.

A konzerváltsági adatokat Dr. Pajkos Mátyás generálta le az összes adatbázisban szereplő fehérje szekvenciára. A konzerváltság megállapítására ortológ szekvenciákra többszörös szekvencia illesztést végzett (multiple sequence alignment) és ezekből konzerváltsági értékeket számított. Az ortológ szekvenciák konzerváltsági értékei 5 evolúciós csoportba lettek csoportosítva: Mammalia, Vertebrata, Eumetazoa, Opisthokonta és Eukaryota. Szintenként minimum 3 ortológ szekvencia kell a konzerváltság megállapításához.

3.3 Adatok összehangolása, SQL táblák, Webszerver

Minden általam vagy a csoport által generált adatot a referencia szekvenciára és azok azonosítójára illesztettem. Az adatokat a megjelenítéstől függően szűrtem és normalizáltam. A webszerver megjelenítéséhez egy SQL alapú adatbázist terveztem, hogy minden adat a lehető leggyorsabban elérhető legyen. A relációs adattáblákat úgy terveztem meg, hogy egy fehérjéről az összes annotációt könnyen és gyorsan elérjem. (1.ábra)



1.ábra a) Az ábra első fele mutatja az összes táblát és azok kapcsoltsági logikáját. Összesen 39 táblát tartalmaz az adatbázis fehérjékkel kapcsolatos adatai. A relációs adatbázisnál egy táblából át tudunk jutni más táblákhoz, ha azok össze vannak kapcsolva. Az annotációs táblák a fehérje táblához kapcsoltak az egyedi azonosítón keresztül. b) Az ábra második felében egy példát láthatunk. Ebben az esetben a fehérje táblából a kombinált rendezetlenség adatait is megkaphatjuk, ha arra hivatkozunk, illetve ugyan ez igaz fordítva is.

A vizualizáción lévő annotációk képi megjelenítésének teljes kódolását én végeztem. A gyorsabb betöltés érdekében, minden annotációt háttérben futó párhuzamos lekéréssel

oldottam meg. A megjelenítőben szereplő információkat előre formáztam, hogy csak a legszükségesebb, de leginformatívabb adatokat prezentáljam. Az annotációk minden egyes blokkját egyedileg készítettem el és testre szabtam az adatokhoz.

4.4 Programozási nyelvek, keretrendszerek

Az adatok processzálásához használt nyelvek: *Python (Pandas, Numpy, Scipy, Biopython, matplotlib)*, *R*.

A webserververhez használt programozási nyelvek : *Python (Django)*, *SQL (MySQL)*, *Javascript (jQuery)*, *CSS (CSS3)*, *HTML (HTML5)*.

A webserververen szerver oldali (back-end) keretrendszerként a *DJANGO*-nak (v. 2.0.4) a második verziója fut. Az adatok tárolása az *SQL* alapú *SQLite3*-at használtam. A kliens oldalon (front-end) a *DJANGO-Template*-et és a *jQuery*-t (*Javascript* nyelv alapú könyvtár) (v. 3.6.0) , valamint a *Bootstrap*-et (*css-javascript* könyvtár) (v. 5.0) használok megjelenítőként.

4. Eredmények és megvitatásuk

4.1 Eredmény összefoglaló

Munkám során egymásra térképeztem a fehérje szintű adatokat a mutációs adatokat és különböző szerkezeti és funkcionális annotációkat integráltam. Az adatbázis köré egy mindenki számára könnyen használható, vizualizációs webszervert építettem. Az eredmények során bemutatom az adatbázis főbb adatait, a webszervert és funkcióit, majd példákkal illusztrálom a szerver használhatóságát.

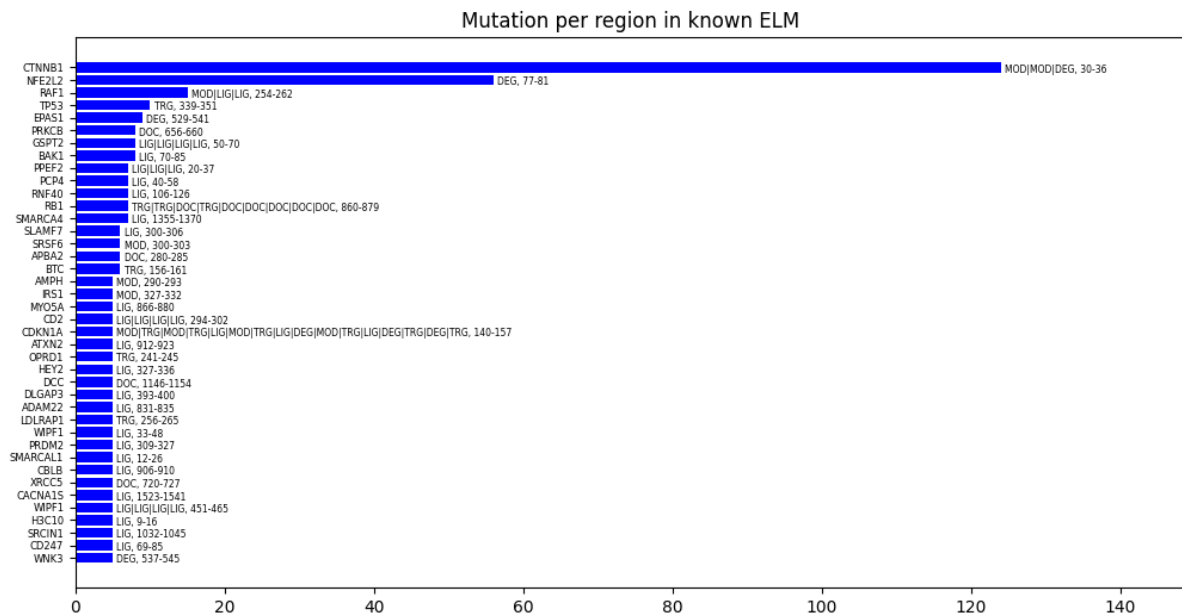
4.2 Statisztikák

A TCGA összesen 19 125 olyan gént tartalmaz aminek vannak mutációi. Ezekhez kigyűjtöttem a COSMIC azonosítót és az ezekhez tartozó transzkriptekhez megkerestem a legjobban illeszkedő Uniprot szekvenciát. Ez a módszer bizonyult a legjobbnak ahhoz, hogy legjobban ki tudjam aknázni a fehérje annotációkat és maximalizálni tudtam az áttérképezhető mutációk számát is. Az adatbázisomban a COSMIC-ban hozzárendelt transzkript azonosítót használtam egyedi azonosítóként, de ezekhez hozzárendeltem a gén nevet, a Uniprot azonosítót és a fehérje nevét, illetve az ENSEMBL transzkript ID-t is. Összesen 18 525 SwissProt szekvenciát térképeztem a transzkriptekre. Néhány esetben Trembl-ös (270) vagy a SwissProt isoforma (170) bizonyult a legjobb szekvenciának. Az adatbázisban összesen 18 965 szekvencia található, ezekhez összesen 1 836 026 rákos mutációt rendelhetünk a TCGA alapján. Ezek mellett 3 136 gén esetén összesen 28 523 disease mutációt tartalmaz az adatbázis.

Készítettem statisztikákat a rendezett/rendezetlen régióra eső adatok megoszlásáról is. Az adatbázisban található több mint 11 millió pozíció közül 35%-uk esik rendezetlen és 65%-uk rendezett régióra. Az ismert rákos gének (driver) mutációinak 40%-a rendezetlen régióra esik, míg az összes gén esetében csak a mutációk egyharmada. A disease mutációk csak 15%-a esik rendezetlen régióra, míg a gyakori polimorfizmusok majdnem 40%-a.

Összegyűjtöttem az ismert ELM és ELM Switches motívumokra eső mutációkat is. (2. ábra) Ehhez hasonló táblázatokat készítettem a lineáris motívumokat kötő doménekre, a

kísérletesen igazolt rendezetlen fehérjékre, illetve a driver és a szignifikánsan mutálódó régiót tartalmazó fehérjékre.



2. ábra. Ismert ELM motívumokra eső mutációk és azok fehérjén belüli régiói, a mutációk száma szerint csökkenő sorrendben. Az első helyeken ismert rákos gének találhatóak. Az átfedő motívumokat összeolvastam.

4.3 Webszerver

4.3.1 Kereső, Táblák, Rólunk, oldalak

Az általam készített DisCanVis adatbázis nyilvánosan elérhető a <https://discanvis.elte.hu> web helyen. A webszerver kezdőoldalán található egy rövid összefoglaló a szerverről, útmutatással és példákkal. A fehérjét nevük, gén nevük vagy UniProt azonosítójuk alapján a *Search* oldalon kereshetünk. Annak érdekében, hogy a rendezetlen régiókra eső mutációs és funkcionális információkat könnyebben reprezentáljuk, több táblázatot készítettünk, amelyek a *“Browse tables”* oldalon találhatóak. Minden táblázatban megtalálható, hogy az adott gén igazolt rákos génként szerepel-e, valamint alapinformációkat az adott génről (kromoszómális lokáció, gén név és fehérje név). Az információk kiegészülnek alap statisztikákkal, mint például a rendezetlen régiókon fekvő rákos és egyéb mutációk száma. Egyes táblázatok a génnel kapcsolatban szolgáltatnak információt, míg mások a géneken belüli régiókról. Abban az esetben ha régiókat tartalmaz a táblázat, a régióon belüli mutációk százalékos eloszlása is megjelenik a teljes génhez képest.

A *Browse tables* után a *Help/About* oldal található, ahol a szerverről találhatóak segédletek, valamint útmutató az idézéshez. Az oldalak sorát a *Statistics* oldal zárja, ahol összesített statisztikák találhatóak.

4.3.2 Fehérje vizualizáció

A vizualizációhoz a kereső táblákon lévő *Protein Identifier*-re kattintva juthatunk el. A vizualizáció három részre osztható fel.

4.3.2.1 Fejléc, funkciós gombok

Az első rész a fejléc, ahol az infó gombra kattintva a génhez kapcsolódó alapvető adatok láthatóak. Itt található meg többek között a gén neve, UniProt azonosítója, kromoszómális lokációja és ENSEMBL transcript id-ja és a szekvencia hossza. A fejléc többi részén összesítő statisztikák, illetve funkcionális gombok találhatóak, amelyek a fő vizualizációs ábra irányítására szolgálnak. Abban az esetben, ha ismert rákos gént vizsgálunk, akkor a hozzátartozó adatbázisban megtalálható információkat is láthatjuk. Ha tartalmaz az

iSiMPre által szignifikánsan mutálódott régiót, akkor funkcionális gomb is megjelenik “ROI” néven. A “Cancer type”-ra kattintva szűrhetünk célzottan a ráktípusokra, aminek változtatásai mind az összesítő ábrán, mind a fő vizualizációs ábrán megjelennek.

4.3.2.2 Teljes fehérje összesítő ábra

A fejléc alatt található a második szekció, ahol a fehérjét teljes hosszában láthatjuk, Pfam doménekkel, mutációkkal, és a kombinált rendezetlenség profillal. A missense mutációk fehérjén belüli eloszlását frekvencia szerint arányosan ábrázoltam, a fő vizualizációs ábrához hasonlóan.

4.3.2.3 Fő vizualizációs ábra

A harmadik szekció a fő ábra, ahol genomikus és fehérje szintű adatokat találhatunk. A fő ábrát további öt fő régióra bonthatjuk.

Legfelül találhatjuk a szekvencia pozíciókat és a szekvenciát, és általános információkat, melyek segíthetnek eldönteni a különféle variációk relevanciáját. Ezek közé tartoznak az exon határok, melyek hasznosak lehetnek a splicing site-ra eső mutációk azonosításában. Az ismétlődéseket tartalmazó régiók (TRF módszerrel), az alacsony genomikus konzerváltságú régiók (PhastCons módszer alapján), vagy a gyakori polimorfizmusok tartalmazó pozíciók valószínűsíthetők, hogy az ezekre eső mutációk passenger mutációk.

A második blokk a rákos mutációkat (TCGA, COSMIC), valamint egyéb betegségből származó mutációkat (ClinVar, OMIM) jeleníti meg. A rákos mutációknál alaphoz csak a TCGA-ből származó mutációkat jelenítem meg, de bekapcsolható az összes COSMIC mutációk megjelenítése is. A mutációk vizualizálása során külön jelenítem meg az egyetlen aminosavat érintő missense mutációkat, az inframe deléciókat és inzerciókat, és trunkáló (frameshift, nonsense) mutációkat.

A harmadik szekció a szerkezeti tulajdonságokra vonatkozó információkat tartalmazza. Az első rész a Pfam annotációkat tartalmazza, a domén és egyéb típusú szekvencia családokat eltérő színnel jelölve. Ezt követi az adott szekvenciához tartozó ismert

térszerkezetek a *PDB* adatbázisból. Ezután jön a kombinált rendezetlenségi profil, majd a kísérletesen igazolt rendezetlen régiók a *MobiDB* adatbázisból. Ezt követi az AlphaFold2 *pLDDT score*-ja által prediktált rendezetlenség, az IUPred és az ANCHOR módszerek által becsült rendezetlen illetve rendezetlen kötőhely régiók. A rendezetlen régiók pirossal, míg a rendezett részek kék színnel vannak feltüntetve.

A negyedik szekció tartalmazza a funkcionális annotációkat, a SwissProt adatbázisból, a rövid lineáris motívumokat illetve az ezeken alapuló molekuláris kapcsolókat, továbbá a különböző poszttranszlációs módosulásokat. Ezek után következnek a DIBS és MFIB adatai, ahol azok a rendezetlen régiók találhatóak, amelyek más fehérjékkel kölcsönhatásba lépnek és ennek során meghatározott térszerkezetet alakítanak ki.

Az utolsó szekció tartalmazza a fehérje konzerváltság adatait különböző evolúciós szinteken. A konzerváltsági értékek arányosítva jelennek meg, hogy vizuálisan is könnyen el tudjuk különíteni a konzervált és kevésbé konzervált régiókat.

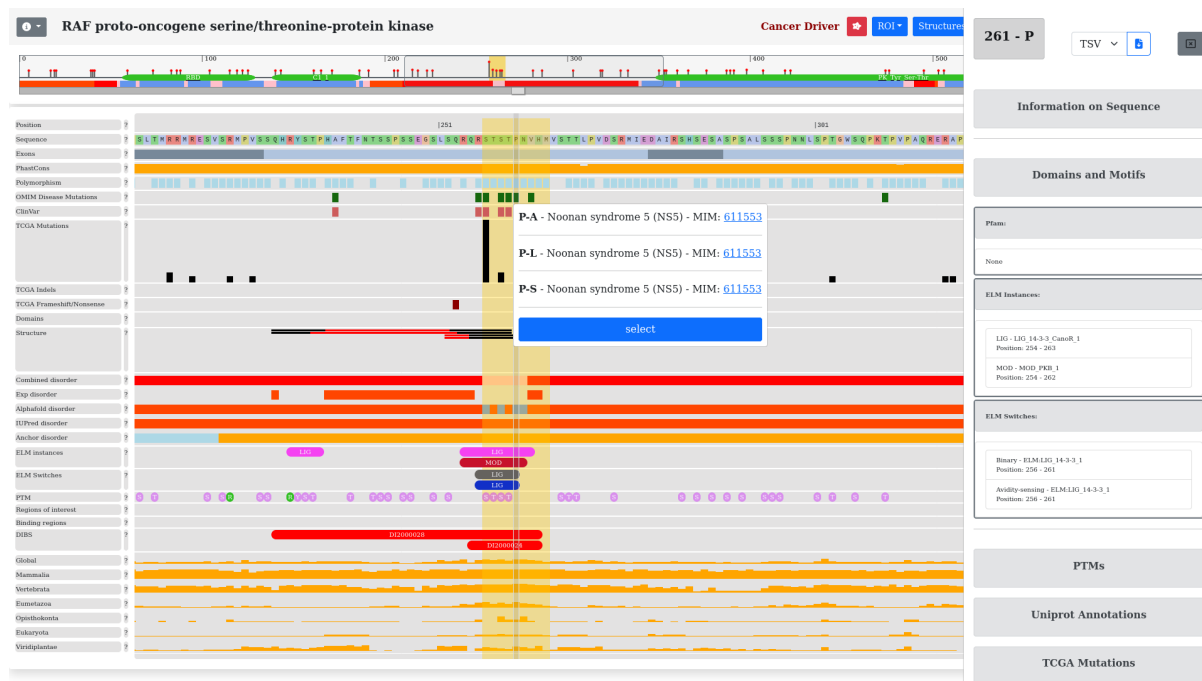
Minden annotáció pozíciója az ábrán megjelölhető, illetve ezekhez az annotációkhoz tartoznak kapcsolt hivatkozások és egyéb hasznos adatok. Minden pozíció adataihoz hozzáférhetünk a REST-API segítségével, vagy a szekvencián az aminosavra kattintva. A REST-API osztott kapcsolatú és gyors hozzáférést biztosít az adatbázis adataihoz. Nagyléptékű analízis során a REST-API segítségével különböző fehérjék kulcspozícióinak vagy régióinak mutációs eloszlásáról és vagy funkciójáról kaphatunk gyors információt.

4.4 Példák

A következő fejezetben példákon keresztül mutatom be a szerver hasznosságát a rendezetlen fehérjék rákos és egyéb betegségek feltérképezésében.

4.4.1 RAF1

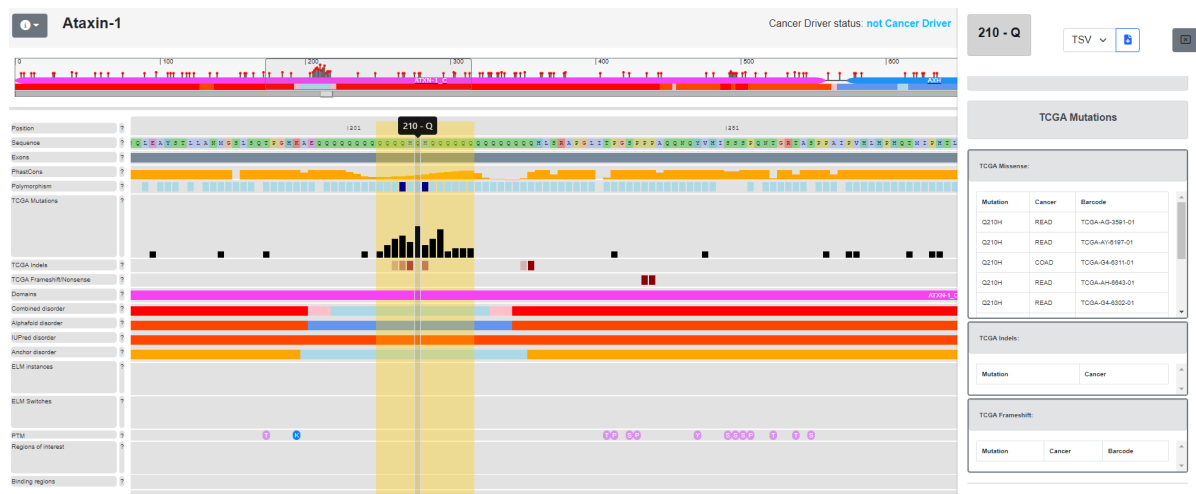
A vizualizációnk kiinduló pontja lehet az ismert lineáris motívumok szerepének megértéséhez különböző betegségek kialakulásában. Erre példa a RAF (RAF1 UniProt:P04049), ami egy protoonkogén szerin/treonin kináz. Ez a fehérje szabályozza a sejtosztódást és a sejtdifferenciációt és fontos szerepet tölt be a RAS/RAF/MEK/ERK jelátviteli útvonalban.(Cobb et al. 1994) Ezeknek az útvonalak zavara egyrészt rákos elváltozásokhoz vezethet, másrészt szerepet játszhat a Noonan szindróma kialakulásában, ami egy gyakori fejlődési rendellenesség. A három leggyakoribb tünete a betegségnek a szokatlan arcvonások, alacsony termet valamint a veleszületett szívrendellenességek.³⁸ Ha a vizualizációs ábrára megyünk, megfigyelhetjük, hogy a rákos mutációk a RAF1-ben a szekvencia közepén található (254-262) 14-3-3 kötő foszfopeptid motívumra esnek (3. ábra). Ez a motívum egy ősi, evolúciósan konzervált motívum, amit a konzerváltsági plot is jól mutat. A rákos mutációkon kívül az egyéb betegségek mutációit is találhatunk ebben a régióban, amiknek a nagy része a Noonan szindrómához kapcsolódnak. Az ELM adatbázisból és az annotációkból kiderül, hogy az erre a régióra eső mutációk gátolják a 14-3-3 kötődését a RAF1-re a S259-es aminosav foszforilációjának elrontása révén. Ez az ERK túlzott aktivitásához vezet, ami végső soron Noonan szindrómát eredményez.³⁹ Ezen eredmények is igazolják, hogy az adatbázisunk remek kiinduló alapot ad a funkcionális annotációk segítségével a rákos és egyéb betegségek feltérképezésére is.



3. ábra. RAF1 fehérje vizualizációja.

4.4.2 Ataxin-1, Ismeretlen rákos gének azonosítása

A vizualizációs eszközünk a még funkcionálisan nem jól ismert, de mutációkban gazdag régiók jobb megértésében is segítséget nyújt. Az egyik ilyen példánk az Ataxin-1 (ATXN1 UniProt: P54253) (4. ábra). Az Ataxin-1 egy kromatin-kötő faktor gén, ami szerepet játszik a génkifejeződés szabályozásában. A fehérje tartalmaz egy poliglutamin (PolyQ) régiót, amely a spinocerebelláris ataxia 1 (SCA1) neurodegeneratív betegségben kiterjedtebb. Az Ataxin-1 számos fehérjével interakcióba lép, de funkciója még nem tisztázott.⁴⁰ Azt eddig is tudták, hogy az ataxin intranukleáris testecskét képzett, de újabban kimutatták, hogy fázis szeparáción megy keresztül.⁴¹ Az ataxin1 jelenleg nem tartozik az ismert driver gének közé, annak ellenére, hogy a méhnyakrákkal kapcsolatba hozták.⁴² A polyQ régiójában megfigyelt mutációk jelentős sűrűsödése arra következtet, hogy a fehérje nem csak neurodegeneratív betegségekben, de a rákban is fontos szerepet játszhat.⁴²

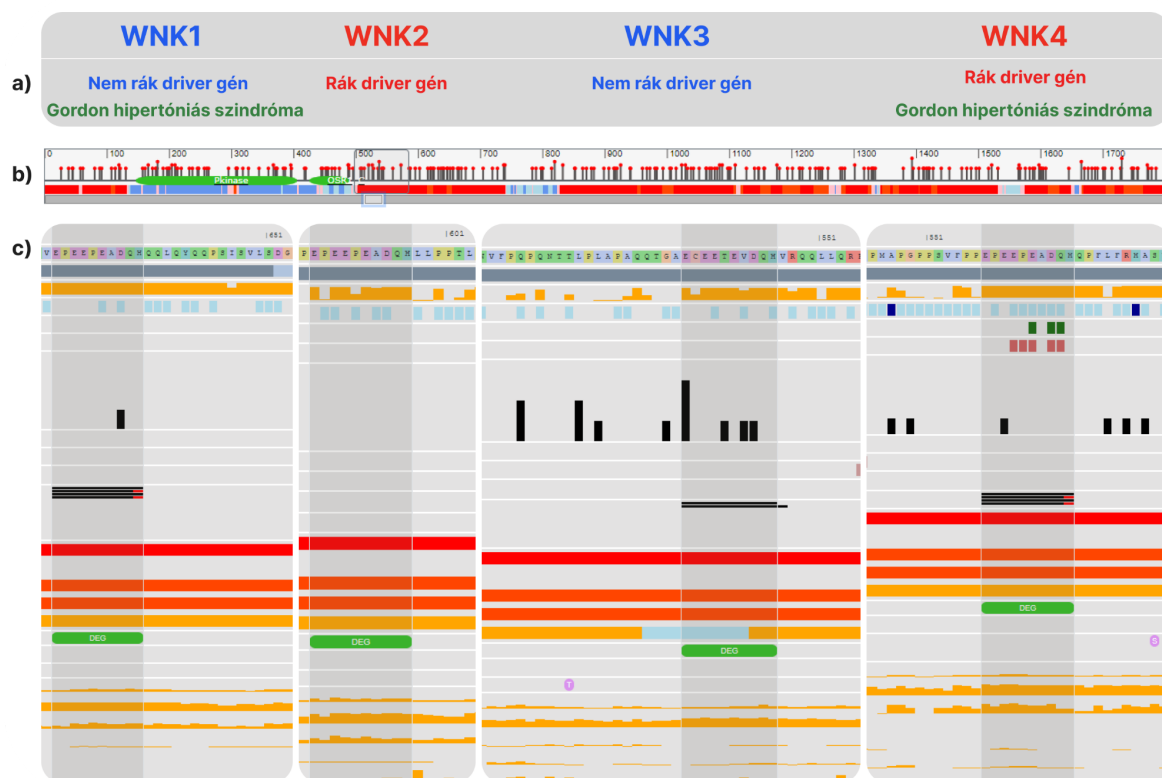


4. ábra Ataxin-1 poliglutamin régiója.

4.4.3 WNK család, ismeretlen rákos gén vagy egyéb betegség azonosítása paralógok által

A *Browse tables* oldalon specifikált keresések segítenek a lehetséges driverek beazonosítását. Az *ELM* fülön az ELM adatbázisban megtalálható ismert lineáris motívumokra eső mutációkat vizsgálhatjuk azok típusa szerint. A lineáris motívumok egyik típusát a degnon motívumok alkotják, melyek a ligandjukhoz való kötődésük során annak degradációját indukálják. Az “ELM type”-on belül, ha a degnonokat szeretnénk szűrni akkor a “DEG”-et kell kiválasztani. Az első három legtöbb rákos mutációval rendelkező degnon a Béta-Catenin (CTNBB1), a NFE2L, valamint a MYC onkogén. Ezek mind ismert rákos gének, viszont ezt a sort követi a WNK3, ami jelenleg nem tartozik az ismert rákos gének közé. A WNK családnak három további paralógja létezik. Ezeket megvizsgálva láthatjuk, hogy a WNK1 a WNK2 és a WNK4 is rendelkezik degnon motívummal. A WNK családnál a degnon motívumot egy Kelch fehérje (KLHL3) ismeri fel, aminek kötődése során a fehérjét ubiquitinálja és degradálja egy E3 ubiquitin ligáz komplexen keresztül.^{43,44} A WNK4 degnonja számos mutációt tartalmaz, amit összefüggésbe hoztak a Gordon hipertóniás szindrómával (pszeudohipoaldoszteronizmus). Ennek a szindrómának a missense mutáció megtalálhatóak a kötődőménéken is, és kijelenthető, hogy ezt a rendszert érintő missense mutációk Gordon hipertóniás szindrómát okoznak.^{45,46} A WNK2 és WNK4 fehérje ismert rákos driver gének, de érdekes módon jelenleg nincs tanulmány a rák és WNK3 kapcsolatáról.

Az 5. ábrán találhatóak a WNK család degnonjai és a paralógok betegségekkel kapcsolatos jelenlegi összefüggései. Ezen eredmények rámutathatnak a WNK3 degnon rák és egyéb betegségekben való szerepére.



5. ábra a) A WNK család különböző paralógjai és a velük összefüggésbe hozott betegségek. b) WNK3 fehérje és a degron pozíciója. A WNK család tagjaiban a degronja közvetlen egy kináz domén után helyezkedik el. A többi izoforma is hasonló szerkezetet mutat, ugyanazokkal a doménekkel. c) A különböző paralógokon a degron motívum és az azokra eső rákos és egyéb betegségek mutációi. A WNK3 jelenleg nem rákos gén, mégis számos mutáció esik a degron motívumára, továbbá a család többi izoformját és a degron motívumát is összefüggésbe hozták rákos és egyéb betegségekkel kapcsolatban.

4.5 Konklúzió

A webszerverem egy olyan általános vizualizációs eszköz, melynek célja a rendezetlen fehérjék megértése a rákban és más betegségekben. A szerveren lévő struktúrális és funkcionális annotációkkal, valamint a genomikus és fehérje szintű adatok egymásraterképezésével egy komplexebb és az eddigiektől eltérő megközelítéssel lehet vizsgálni a rendezetlen fehérjék és a rák kapcsolatát. Lehetőség nyílik arra is, hogy a globuláris fehérjék szerepét is jobban megértsük a rendezetlen kötő régiók által, amik rendezett doménekhez kötnek. A specializált keresésekhez létrehoztunk különféle táblákat amikkel a funkciós csoportokat külön-külön rendezetlen specifikusan tudunk vizsgálni. Így lehetőség van akár kísérletesen igazolt rendezetlen régiók szerint keresni géneket, vagy akár az ismert rákos gének között, de gén ontológia alapján is. A jövőben szeretnénk kiegészíteni az adatbázist további rákos és más betegségek mutációival. Ezen kívül szeretnénk célzott kezelések irányába tovább haladni. A DisCanVis a fent említett példákon keresztül jól reprezentálja a használhatóságát, a mutációk megértését a rendezetlen fehérjékben, és így a rák kialakulásában.

5. Összefoglalás

A rendezetlen fehérjék nem rendelkeznek jól definiált térszerkezettel, ugyanakkor kulcsfontosságú szerepet játszanak szinte minden sejten belüli folyamatban. Hibás működésük számos betegséggel hozható kapcsolatba, többek között rákkal is. Az elmúlt időszakban jelentősen kitárult a rendezetlen fehérjéről szerzett tudásunk, és ezzel párhuzamosan a rák genom projektek rengeteg, a betegség kialakulásában fontos genetikai variánst azonosítottak. Ennek ellenére a rendezetlen fehérjék rákban betöltött szerepe máig nem tisztázott. Munkám során létrehoztam egy komplex webszervert, a DisCanVis-t (<https://discanvis.elte.hu>), amely elősegíti a rák és a rendezetlenség közötti kapcsolat vizsgálatát. A webszerver megalkotására több programnyelvet (Python, Javascript, HTML, CSS, SQL) és azok keretrendszereit használtam (Django, JQuery, Bootstrap, SQLite). Ez az egyedi vizualizációs eszköz genom szintű információt kombinál a fehérje szinten lévő strukturális és funkcionális információkkal. Több mint 30 különböző annotációt gyűjtöttem össze és processzáltam, és ezeket összehangolva vizualizáltam. Dolgozatomban példákon keresztül mutattam be, hogyan használható a webszerver a rák kialakulásában szerepet játszó rendezetlen fehérjék, illetve azok kulcs régióinak azonosításához. A DisCanVis szerver kiindulópontja lehet biomedikai relevanciával rendelkező, de ma még nem jellemzett rendezetlen fehérjék karakterizálásának és elősegítheti újfajta gyógyszercélpontok azonosítását is.

6. Köszönetnyilvánítás

Szeretnék köszönetet mondani az elmúlt éves munkám folyamatos szakmai útmutatására és támogatására témavezetőmnek, Dr. Dosztányi Zsuzsannának. Továbbá szeretném megköszönni a Dosztányi csoport többi tagjának a segítséget, kiemelve köztük Erdős Gábornak és Dr. Pajkos Mátyásnak.

7. Irodalom

1. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
2. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
3. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–77 (2015).
4. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
5. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
6. Chiu, S.-H., Ho, W.-L., Sun, Y.-C., Kuo, J.-C. & Huang, J.-R. Phase separation driven by interchangeable properties in the intrinsically disordered regions of protein paralogs. *Commun Biol* **5**, 400 (2022).
7. Quaglia, F. *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* **50**, D480–D487 (2022).
8. Erdős, G., Pajkos, M. & Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).
9. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
10. Pajkos, M. & Dosztányi, Z. Functions of intrinsically disordered proteins through

- evolutionary lenses. *Prog. Mol. Biol. Transl. Sci.* **183**, 45–74 (2021).
11. Zou, H. *et al.* Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes. *Nucleic Acids Res.* **50**, e49 (2022).
 12. Mészáros, B., Hajdu-Soltész, B., Zeke, A. & Dosztányi, Z. Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules* **11**, (2021).
 13. Pavlaki, I. Fusion proteins drive cancer through phase separation. *Nature cancer* vol. 2 1285 (2021).
 14. Zhao, B. & Xue, B. Self-regulation of functional pathways by motifs inside the disordered tails of beta-catenin. *BMC Genomics* **17 Suppl 5**, 484 (2016).
 15. Jehl, P., Manguy, J., Shields, D. C., Higgins, D. G. & Davey, N. E. ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.* **44**, W11–5 (2016).
 16. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
 17. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
 18. Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–33 (2013).
 19. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
 20. Kumar, M. *et al.* The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* **50**, D497–D508 (2022).
 21. Van Roey, K., Dinkel, H., Weatheritt, R. J., Gibson, T. J. & Davey, N. E. The

- switches.ELM resource: a compendium of conditional regulatory interaction interfaces. *Sci. Signal.* **6**, rs7 (2013).
22. Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–70 (2012).
 23. Li, Z. *et al.* dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.* **50**, D471–D479 (2022).
 24. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–7 (2005).
 25. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
 26. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac240.
 27. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
 28. Piovesan, D. *et al.* MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.* **49**, D361–D367 (2021).
 29. 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).
 30. Schad, E. *et al.* DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**, 535–537 (2018).

31. Fichó, E., Reményi, I., Simon, I. & Mészáros, B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **33**, 3682–3684 (2017).
32. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).
33. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
34. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
35. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
36. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
37. 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).
38. Kobayashi, T. *et al.* Molecular and clinical analysis of RAF1 in Noonan syndrome and related disorders: dephosphorylation of serine 259 as the essential mechanism for mutant activation. *Hum. Mutat.* **31**, 284–294 (2010).
39. Bekele, R. T. *et al.* RAF1 amplification drives a subset of bladder tumors and confers sensitivity to MAPK-directed therapeutics. *J. Clin. Invest.* **131**, (2021).
40. Zhang, S. *et al.* The ataxin-1 interactome reveals direct connection with multiple disrupted nuclear transport pathways. *Nat. Commun.* **11**, 3343 (2020).
41. Zhang, S., Hinde, E., Schneider, M. P., Jans, D. A. & Bogoyevitch, M. A. Nuclear bodies formed by polyQ-ataxin-1 protein are liquid RNA/protein droplets with

- tunable dynamics. *Scientific Reports* vol. 10 Preprint at <https://doi.org/10.1038/s41598-020-57994-9> (2020).
42. Kang, A.-R., An, H.-T., Ko, J., Choi, E.-J. & Kang, S. Ataxin-1 is involved in tumorigenesis of cervical cancer cells via the EGFR-RAS-MAPK signaling pathway. *Oncotarget* **8**, 94606–94618 (2017).
43. Gong, Y. *et al.* KLHL3 regulates paracellular chloride transport in the kidney by ubiquitination of claudin-8. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4340–4345 (2015).
44. Takahashi, D. *et al.* KLHL2 interacts with and ubiquitinates WNK kinases. *Biochemical and Biophysical Research Communications* vol. 437 457–462 Preprint at <https://doi.org/10.1016/j.bbrc.2013.06.104> (2013).
45. Schumacher, F.-R., Sorrell, F. J., Alessi, D. R., Bullock, A. N. & Kurz, T. Structural and biochemical characterization of the KLHL3–WNK kinase interaction important in blood pressure regulation. *Biochem. J* **460**, 237–246 (2014).
46. Sohara, E. & Uchida, S. Kelch-like 3/Cullin 3 ubiquitin ligase complex and WNK signaling in salt-sensitive hypertension and electrolyte disorder. *Nephrol. Dial. Transplant* **31**, 1417–1424 (2016).