Fibrilláris és globuláris fehérjeszerkezeti elemek belső dinamikájának vizsgálata integrált szerkezeti bioinformatikai módszerekkel

MTA doktori értekezés

Dr. Gáspári Zoltán

Pázmány Péter Katolikus Egyetem

Információs Technológiai és Bionikai Kar

2024. február 26.

ARTHUR: Camelot!

GALAHAD: Camelot!

LANCELOT: Camelot!

PATSY: It's only a model.

Monty Python: Gyalog Galopp (1975)

Köszönetnyilvánítás

Az értekezésben leírt eredmények mind kollégákkal és diákokkal közös munka és közös gondolkodás eredményei, így a dolgozat nem lehet teljes anélkül, hogy köszönetet ne mondanék sokaknak, akik meghatározóak voltak számomra eddigi szakmai tevékenységem során.

Köszönöm Perczel Andrásnak, hogy 13 éven keresztül mentorom és témavezetőm volt. Az ő csoportjában szerzett ismereteim az NMR-spektroszkópia és a számításos szerkezetvizsgálatok kapcsán máig meghatározóak számomra.

Első egyetemi éveim és a kutatás világában tett első lépéseim során rengeteget köszönhettem Nyitray Lászlónak és Tóth Gábornak, akikkel mindmáig élő szakmai kapcsolatban állok.

A kutatásaim során velem együttműködő kollégáknak, Pongor Sándornak, Batta Gyulának, Ortutay Csabának, Nagy Zoltánnak, valamint Várnai Péternek, Michele Vendruscolónak, Florentine Marxnak és Perttu Perminek ezúton is köszönöm az együtt gondolkodást és a sok türelmet.

Hálás vagyok mindazoknak, akik befogadtak, támogattak és biztosították számomra a szakmai munkavégzés lehetőségét az ELTE Biokémiai és Szerves Kémiai Tanszékén, valamint a PPKE ITK-n: Gráf Lászlónak, Hollósi Miklósnak, Roska Tamásnak, Nyékyné Gaizler Juditnak, Szolgay Péternek, Iván Kristófnak és Cserey Györgynek.

Munkatársaim, diákjaim és PhD hallgatóim nélkül eredményeim nagy része nem születhetett volna meg, ezúton is köszönetet mondok Ángyán Annamáriának, Czajlik Andrásnak, Dobson Lászlónak, Dudola Dánielnek, Farkas Fanninak, Fizil Ádámnak, Harmat Zitának, Hinsenkamp Anettnek, Kálmán Zsófiának, Kovács Ákosnak, Kovács Bertalannak, Miski Marcellnek, Nagy-Kanta Eszternek, Péterfia Bálintnak, Sánta Annának, Süveges Dánielnek, Szabó András Lászlónak, Szappanos Balázsnak.

Végül, de korántsem utolsósorban köszönöm családom, elsősorban feleségem támogató szeretetét, melyre mindig számíthattam és számíthatok.

Tartalomjegyzék

Táblázatok jegyzéke

Ábrák jegyzéke

Bevezetés

A biológiai folyamatok atomi szintű részleteinek megértése elképzelhetetlen a biomakromolekulák térszerkezetének ismerete nélkül. Az elmúlt 70 évben a szerkezeti biológia tudománya hatalmas mennyiségű információval gazdagította a világról alkotott képünket, és ma már szinte természetesnek vesszük, hogy egy tetszőleges fehérjemolekuláról valamilyen szerkezeti modell rendelkezésünkre áll - amennyiben nem kísérletileg meghatározott szerkezet, akkor legalább egy számítógépes modell. E sorok írásakor ez hatványozottan igaz az AlphaFold2 eljárás által generált proteom szintű szerkezetbecslések elérhetővé tétele [1] kapcsán.

Az adatok hatalmas mennyisége azonban, bár figyelemre méltó, önmagában még nem elegendő egyes konkrét biológiai jelenségek leírásához. Ennek oka a modelljeink hiányosságaiban rejlik, azok többsége ugyanis nem ad számot a fehérjeszerkezetek belső mozgásairól és/vagy egyéb szerkezeti változatosságáról. Miközben a modellezés nélkülözhetetlen velejárója a bizonyos fokú egyszerűsítés, ez többször is értékes infromációk figyelmen kívül hagyásához vezethet. A funkcionálisan fontos és biológialag releváns aspektusok megértéséhez nem ritkán többféle kísérleti módszer és/vagy különböző számítási eljárás összehangolt alkalmazása szükséges.

A jelen dolgozatban leírt vizsgálatok közös vonása, hogy a különböző szerkezeti modellek egyes hiányosságait próbálják meg egyes jól körülhatárolható esetekben javítani, szem előtt tartva a szerkezeti információk biológiai relevanciáját. Az általam használt eljárások döntően bioinformatikai jellegűek, de több esetben közvetlenül is felhasználnak kísérleti adatokat. A kapott eredmény lehet egy részletes szerkezeti modell, ami ebben a dolgozatban tipikusan egy sokaság-alapú, sok konformációt integráló koordinátakészlet, vagy egy jól definiált szerkezeti elem jelenléte és helyzete adott fehérjeszekvenciákban. Hangsúlyos még az összehasonlító jelleg egyes molekulák és módszerek tekintetében is, mely hozzájárul a felállított szerkezeti és funkcionális modellek megbízhatóságának növeléséhez.

1. Irodalmi áttekintés

1.1. Fehérjeszerkezetek predikciója és modellezése

Jelen fejezetnek nem célja, mert nem is lehet célja enciklopedikus módon áttekinteni a predikció és modellezés témakörét. Ehelyett egy általam bevallottan szubjektív módon kiválasztott vezérfonalra kívánom felfűzni a dolgozatban bemutatott kutatásokhoz kapcsolódó legfontosabb koncepciókat, azok kapcsolódási pontjait és a jelen munka szempontjából releváns különbségeit hangsúlyozva.

A fehérjeszerkezetek leírása számos különböző szinten lehetséges, az aminosavak sorrendjétől a topológiai leíráson át a háromdimenziós szerkezet atomi szintű reprezentációjáig. Ezek a leírások mind modellnek tekinthetőek, természetesen különböző felbontással és információtartalommal. Jelen munkában alapvetően kétféle típusú modell szerepel: az egyik szekvenciaalapú predikciókból származó topológiai modell[1], azaz adott szerkezeti motívum szekvencián belüli előfordulását és helyzetét egy alapvetően egydimenziós reprezentáción elhelyező leírás, a másik pedig részletgazdag, atomi szintű térszerkezeti modell. Az itt bemutatott predikciók alanya ugyanakkor elsősorban egy igen speciális motívum, a magányos α-hélix [2], amely csupán ezen információ ismeretében már atomi szinten első közelítésben jól modellezhető a az α-hélixekre általában jellemző geometria segítségével.

1.1.1. A fehérjék térszerkezetének modelljei

A fehérjemolekulák térszerkezetének közvetlen megtapasztalása nem lehetséges, csupán közvetett módszerekkel tudunk róluk információt szerezni. Ebből következően valójában minden általunk ismert fehérjeszerkezet modell [3]. Ez a tudományfilozófiai megközelítésből triviális-

[1] itt a „topológiai modell" kifejezést a legáltalánosabb értelemben használom, a transzmembrán domének vagy másodlagos szerkezeti elemek egymáshoz képest való orientációját is reprezentálni hivatott topológiai modellek ezeknél értelemszerűen összetettebbek

nak tekinthető állítás a szerkezeti biológusok körében messsze nem tekinthető magától értetődőnek, ezen a területen ugyanis sokszor igen élesen elkülönítik a kísérletesen meghatározott szerkezeteket az elméleti módszerekkel épített modellektől. Én amellett kívánok érvelni, hogy ezen szerkezetek valójában egy folyamatos skála két végét reprezentálják: az adott molekulára specifikusan meghatározott paraméterek segítségével meghatárorozott szerkezetek esetében is igen komoly mértékben támaszkodunk a molekulák geometriai jellegzetességeiről való meglévő tudásunkra (kötéshosszak, kötésszögek stb.). Ezt úgy is ki szokták fejezni, hogy a biomakromolekuláris szerkezetmeghatározás adatszegény eljárás, ahol a kísérleti paraméterek száma jóval alatta marad a meghatározandó atomi koordináták számának – szöges ellentétben számos „kismolekulás" módszerrel. Ebben a szemléletben a kísérletileg meghatározottnak tekintett és a teljesen elméleti, az adott molekulára vonatkozó mért paramétereket egyáltalán nem felhasználó modellek közötti zónában számos, ma egyre inkább előtérbe kerülő, ún. integrált modellezési módszer foglal helyet. Ilyenek pl. a fehérjekomplexek felépítéséhez használt, kísérleti megkötéseket használó dokkolási eljárások [4, 5, 6], de nem utolsósorban az ebben a munkában is megjelenő, sokaság-alapú dinamikus térszerkezeti sokaságok előállítására alkalmas számítások is. Ezek esetében a kísérleti paraméterek információtartalma nem feltétlenül elegendő ahhoz, hogy egy hagyományos szerkezetmeghatározó eljárás során használhatóak legyenek, megfelelő szimulációs, modellezési technikákkal kombinálva azonban biológialag releváns eredményeket, szerkezeti modelleket lehet velük létrehozni.

1.1.2. Szerkezetpredikció szekvencia alapján

A biokémia egyik alapvető dogmája, hogy a fehérjék szekvenciája meghatározza a háromdimenziós térszerkezetet, ennek létjogosultságát Anfinsen kísérlete óta nemigen vitatják [7]. Bár első közelítésben, a globuláris fehérjék körében ezt értelmezhetjük úgy, hogy egy adott szekvencia determinisztikus módon egyetlen jól meghatározott szerkezetet vesz fel, mai ismereteinkkel valószínűleg helyesebb az az értelmezés, hogy a szekvencia leszűkíti az adott molekula által felvehető legkisebb energiájú konformációk halmazát. Ezen halmaz – ami megfelel a natív szerkezetű molekula által bejárt konformációs térnek – lehet egyetlen jól meghatározott szerkezet viszonylag szűk környezete: ez a helyzet a legtöbb globuláris és fibrilláris, valamint transzmembrán fehérje esetében. Előfordulhat, hogy nem egyetlen, hanem néhány kedvező szerkezet környezetéről van szó, erre példának hozhatóak a relatíve kicsi perturbáció hatására szerkezeti átalakuláson átmenő metamorf fehérjék [8], vagy a metastabil

fehérjék speciális esetei [9]. Végül lehetséges, hogy a kedvező, biológiailag releváns „natív" konformációk halmaza valójában egy igen tág régió, mint a funkcionálisan rendezetlen fehérjék esetében [10]. Mindezekkel együtt természetesen érvényes marad az a megfontolás, hogy a szekvencia alapján elméletileg képesek lehetnénk a minimális energiájú, natív szerkezet(ek) predikciójára. Ez az elvi lehetőség az AlphaFold eljárás 2020 végén bemutatott és 2021-ben publikált és elérhetővé tett, második generációs verziójával kézzelfoghatóvá vált [11]. Noha nem vagyunk képesek az összes biológialag releváns szerkezet nagy pontosságú predikciójára, az egydoménes globuláris fehérjék szerkezetbecslése a gyakorlatban nagy pontossággal működik, ráadásul – bár erre valójában nem tanították – az AlphaFold elfogadható szinten alkalmas funkcionálisan rendezetlen szakaszok predikciójára is [12]. Ezzel együtt még nem tartunk ott, hogy egy tetszőleges szekvencia vagy fehérjeszakasz esetében legyen egyetlen integrált „mindentudó" eljárás, ami minden esetben megtalálja a biológiailag releváns szerkezete(ke)t. Egy példával megvilágítva: egy dimer transzmembrán fehérje esetében mind a transzmembrán régió(k), mind a funkcionálisan nem monomer állapot predikciója speciális eljárásokat és szakértelmet kíván. A heteromultimer fehérjék esetében pedig jelenleg a lehetetlennel határos, hogy konkrét partner(ek)re vonakozó javaslatok tételére legyen képes egy eljárás, de még annak felismerése sem triviális, hogy valamilyen kötőpartner jelenléte feltétlenül szükséges. Ennek oka persze nem csupán elvi: egy kötőpartner jelenléte vagy hiánya *in vivo* is jelentős befolyással lehet a szerkezetre.

1.2. Alacsony komplexitású szekvenciák

1.2.1. Mit jelent az alacsony komplexitás?

A globuláris fehérjék szekvenciájában tipikusan mind a 20 fajta aminosav előfordul, és rendszerint nem tartalmaznak könnyen felismerhető módon, szabályosan igsmétlődő szakaszokat. Az ismert fehérjeszekvenciák számának növekedésével egyre több olyan szakasz vált ismertté, amely nem illeszthető ebbe a képbe: az alkotó aminosavak változatossága alacsony és/vagy egyértelmű a repetitív jelleg jelenléte [13]. Az ilyen jellegű szakaszokat összefoglaló néven alacsony komplexitású régióként (low complexity region, LCR) említi a szakirodalom [14]. Ezek egy jelentős részét ma funkcionálisan rendezetlen fehérjeként ill. fehérjeszakaszként tartjuk számon [15], és az egyedi aminosavösszetételt a globuláris fehérjéktől való lényeges megkülönböztető jellegnek tekintjük. A repetitív jelleget mutató szekvencákk közül példá-

nak a poliglutamin szekvenciák (pl. a Huntingtin fehérje esetében [16]) vagy a coiled-coil régiók heptád ismétlődése említhető [17]. Bár a koncepció egyszerűnek tűnik, széles körben elfogadott, kvantitatív alapokon nyugvó definíció – amely pl. egyértelmű algoritmikus azonosítás alapját képezhetné – jelenleg nincs. Természetesen számos program áll rendelkezésre ilyen régiók azonosítására, egyértelmű definíció hiányában ezek azonban nem pontosan ugyanazt az eredményt adják – ez a helyzet egyébiránt nem különbözik érdemben a legtöbb bioinformatikai jellegű feladat esetében tapasztalhatótól.

1.2.2. A coiled-coil és a poliprolin II hélix

Az globuláris fehérjékre jellemzőnél alacsonyabb komplexitású motívumok közé sorolható többek között a coiled-coil[2] és kollagén vagy poliprolin II (PPII) hélix szerkezet [15]. Mindkettőre jellemző az ismétlődő jelleg. A coiled-coil szerkezetet felvevő szakaszok hét, ritkábban akár 11 aminosavas vagy hosszabb ismétlődő motívumot tartalmaznak, ezen felül jellemző rájuk a nagyméretű hidrofób, elsősorban aromás oldalláncú (fenilalanin, tirozin, triptofán) aminosavak, valamint a szerin és a glicin majdnem teljes hiánya [17]. A PPII hélixet alkotó szekvenciák jellegzetessége a prolin-prolin-glicin három aminosavas egység ismétlődése, ezeken kívül – ha itt eltekintünk a poszttranszlációs módosításoktól – csak néhány más aminosav előfordulása számottevő (pl. lizin) [18]. Ilyen tekintetben tehát mindkét motívum alacsony komplexitású, a globuláris fehérjékre jellemzőtől egyértelműen elütő aminosavösszetétellel, ugyanakkor megfelelő szabályosságú ismétlődő mintázattal.

Mind a coiled-coil, mind a PPII hélix szerkezete kellően szabályos ahhoz, hogy matematikai egyenletekkel leírható legyen. Érdekesség, hogy mindkét szerkezeti elem atomi szintű részleteire Francis Crick tett később helytállónak bizonyuló javaslatot még az 1950-es évek elején [19, 20], valamint, hogy a háromszálú PPII hélixek feltekeredését nem egy esetben coiled-coil motívumok segítik egyes kollagénekben [21].

Szerkezetét tekintve a coiled-coil meghatározott módon egymás köré csavarodó α-hélixekből épül fel, a hélixek száma az ismert szerkezetekben 2-től akár 12-ig is terjedhet, utóbbi esetekben széles helikális „tubusokat” alkotva [17]. A szuperhélixre jellemzõ interhelikális kölcsönhatások mintázata a coiled-coil szerkezetet elkülöníti az egyéb hasonló α-helikális elemektől, (pl. hélixköteg, „helical bundle”). A hélixek egymás felé tekintő oldalán kisméretű

[2]kb. „tekercselt tekercs”, az α-hálixek egymás köré tekeredésére utalva. A motívumnak nincs bejáratott magyar neve, az angol szakkifejezés terjedt el

hidrofób oldalláncú aminosavak (leucin, valin, izoleucin) alkotta „hidrofób varrat" található, ezek a heptád ismétlődés 'a' és 'd' pozíciójában vannak a szekvenciában. Az oldalláncok a másik hélix csoportjaival ún. „gomb a gomblyukba" („knobs-into-holes") módon rendeződnek el, ahol az egyik lánc adott aminosavának oldalláncát a másik lánc négy aminosavának oldallánca veszi körül [17]. A coiled-coil motívumot 3D atomi koordinátákból azonosító SOCKET algoritmus ezen pakoltsági mintázat felismerésén alapul [22]. Coiled-coil szerkezetek nem csak heptád ismétlődést tartalmazó szekvenciákból épülhetnek fel, az ismétlődés típusa és a coiled-coil geometriája között szoros összefüggés van [17]. A coiled-coil szerkezetek felépítése szabályos mivoltuk miatt egyszerű geometriai megfontolások segítségével elvégezhető, ezek Crick 1953-ban publikált egyenletein [19] alapulnak. Ma már rendelkezésre áll a CCBuilder nevű webes alkalmazás, amely képes coiled-coil szakaszok ilyen módon való felépítésére [23, 24], azonban ez sem képes minden lehetséges változat felépítésére.

A coiled-coil szerkezetet felvevő aminosavszekvenciák predikciójára számos eljárás született, ezek többsége pozícióspecifikus pontozómátrixot (PSSM) vagy rejtett Markov modellt (HMM) használ [25]. A legkorábban létrehozott algoritmus a PSSM alapú COILS eljárás [26], melyet még ma is igen elterjedten használnak. A legtöbb eljárás közös hiányossága, hogy a heptád ismétlődésektől különböző motívumok felismerésére nem vagy csak korlátozottan képes.

A PPII hélix szerkezetet három polipepdidlánc alkotja, melyeket speciális hidrogénhíd-kötés-mintázat tart össze [27]. Az egyes polipeptidláncok nyújtott konformációt vesznek fel, ebben és a láncok közötti hidrogénkötések fontosságában a szerkezet hasonlít a β-redőkre. Ugyanakkor a speciális térszerkezeti jellemzők miatt stabil hármas helikális szerkezet kialakítására csak a jellegzetes Pro-Pro-Gly szekvencia képes [28]. A PPII szerkezeti elemek 3D koordinátákból való azonosítása tipikusan a molekulagerinc torziós szögei alapján történik [29]. Emiatt számos egyszálú nyújtott szerkezet is az ilyen eljárások által talált szakaszok között van, ami sok esetben egyébként nem ellentétes az adott elemzés céljával. Biomolekuláris kölcsönhatások esetében ugyanis előfordul, hogy egy globuláris domén által specifikusan felismert szakasz PPII-höz hasonló nyújtott szerkezetet vesz fel. Erre jól ismert példa az MHC komplexek által megkötött peptidszakaszok esete [30]. A klasszikus háromszálú PPII szerkezetet kialakító, prolinban és glicinben gazdag szekvenciák azonosítására jelenlegi tudomásom szerint nincs specializált eljárás, ami meglepő. A kollagén hélixet alkotó szakaszok

felismerése a Pfam adatbázis erre a célra létrehozott HMM profilja használható.[3]

1.2.3. A magányos α-hélix motívum

A hagyományos biokémiai felfogás szerint az α-helikális szerkezetű fehérjék egyes hélixei önmagukban, vizes oldatban nem vesznek fel stabil helikális szerkezetet. E szerint a nézet szerint a hélixek stabilizálásához vagy egy globuláris fehérje többi szerkezeti részlete, vagy további hélixek (pl. coiled-coil motívum esetében), vagy lipid kettősréteg (membránfehérjék esetében) jelenléte szükséges [31]. A peptidmodellek világában jól ismert az ún. hélix „capping" jelensége is, amikor egy helikális részletet egy megfelelő hidrogénkötés-mintázatot kialakítani képes szerkezeti elem stabilizál [32]. Mindezek ismeretében meglepetésként hatott a miozin X nyaki régiójának viselkedése, amely vizes oldatban önmagában stabil helikális szerkezetet vesz fel [33]. Azt ezt megmutató vizsgálatokkal nagyjából egy időben az ELTE Biokémiai Tanszékén Nyitray László kutatócsoportja hasonló megfigyelést tett a miozin VI nyaki régiója kapcsán is. Az akkoriban körvonalazódó ismeretek szerint olyan nagy töltéssűrűséggel rendelkező, a pozitív és negatív töltésű aminosavak szabályos váltakozását mutató szekvenciákról volt szó, melyek kifejezetten ritkán kerültek a kutatók látókörébe. Ebben az időben még nem létezett bioinformatikai algoritmus ezen szakaszok predikciójára, és hiányzott az általánosan elfogadott név is. Mi első publikációinkban a CSAH, azaz charged single α-helix nevet javasoltuk, utalva a nagy töltéssűrűségre, ma azonban a legelterjedtebb a magányos α-hélix (single alpha-helix, SAH) elnevezés[4]. A motívumra vonatkozó adatok azóta is viszonylag lassan gyűlnek, bár mostanra már néhány részletesebb térszerkezeti elemzés is napvilágot látott, kísérletes és szimulációs munkák is. Rámutattak, hogy a Lys-Glu sóhidak jellemzően úgy fordulnak elő a SAH szerkezetekben, hogy a Lys esik a molekula N-terminális vége felé, és ennek feltehető oka az oldalláncok szerkezeti preferenciája [34]. A PDB adatbázisban csupán néhány atomi szintű SAH szerkezet érhető el, pl. a miozin VI [35] (PDB ID: 6OBI), vagy az újonnan leírt CAF-1 SAH régió esetében [36] (PDB ID: 8DEI), nem számítva a paraspeckle felépítésében részt vevő fehérjék coiled régióval átfedő, általunk SAH-ként azonosított szakaszait [37] (PDB ID-k: 4WIJ, 6WMZ) (lásd még a 4.1.3. alfejezetben). Ennek oka, hogy a motívum kifejezetten nehezen vizsgálható, a nagy töltéssűrűség miatt nem feltétlenül triviális a kristályosítása, az NMR-spektrumokban pedig igen

[3] ez jelenleg a Pfam megszűnése miatt már az InterPro felületéről érhető el

[4] a CSAH elnevezés kritikája, hogy a név félrevezető, mert maga a motívum összességében nem rendelkezik szokatlanul nagy nettó töltéssel.

komoly jelátfedéssel kell megküzdeni a sok kémiailag és szerkezeteileg is hasonló aminosav miatt. Utóbbi nehézség leginkább speciális 4D spektrumokkal hidalható át, melyek segítségével az egymás utáni aminosavak amid NH rezonanciái egyértelműen párosíthatóak [35]. Az évek során egyre növekedik a különféle fehérjékben azonosított SAH motívumok száma is, bár viszonylag lassan [38, 39, 36]. A kísérletes vizsgálatok esetében lényeges szempont annak megmutatása, hogy a kiválasztott szakasz vizes közegben, monomer formában stabil helikális szerkezetet vesz fel, ezt rendszerint CD-spektroszkópiai mérésekkel bizonyítják. A motívum biofizikai jelentőségére néhány konkrét példa ismert, a legjobban a miozinok nyaki régiója jellemzett ilyen szempontból, ahol is a SAH az erőkar meghosszabbításaként funkcionálva nagyobb lépésköz megtételét teszi lehetővé a dimer motorhehérje számára [33, 40]. Ezen kívül is általában a SAH szakaszok mechanikai szerepét hangsúlyozzák, ún. állandó erőt kifejteni képes elasztikus szerkezetként (constant force spring) is leírták, a kitekeredéssel szembeni ellenállásának tapasztalatai alapján [41]. Kézenfekvőnek tűnik még a távtartó szerep, azaz két funkcionális domén/molekularészlet közötti fizikai távolság meghatározása mintegy „molekuláris vonalzóként” (molecular ruler) működve – érdekesség, hogy az ezt az aspektust a kaldezmon esetében felvető közlemény jóval megelőzte a SAH szakaszok általános jelentőségének felismerését [42]. Az elmúlt években – a dolgozat későbbi részében leírt eredményekkel párhuzamosan – a SAH régiók detektálására kifejlesztették a Waggawagga eljárárást, mely egy ún. metaszerver, azaz több bioinformatikai eljárás eredményét ötvözi. Jelen esetben többféle coiled-coil predikció lefuttatása mellett becslést ad SAH régiók jelenlétére egy, koncepciójában a SCAN4CSAH eljáráshoz (lásd 4.1.1. alfejezet) hasonló pontozási séma segítségével [43]. Elérhető webszerverként és parancssorban futtatható programként is [44]. Fontos kiemelni ugyanakkor, hogy kutatásaink kezdetekor nem volt még sem eljárás a SAH szakaszok szekvenciaalapú predikciójára, sem átfogó kép a SAH szerkezetek eloszlásáról, így az őket tartalmazó fehérjék szerepéről sem.

1.2.4. A fehérje-fázisszeparáció jelensége

A fehérje-fázisszeparáció jelensége jelenleg igen intenzíven kutatott terület. Alapja az a felismerés, hogy bizonyos fehérjék, tipikusan más fehérjékkel vagy RNS-molekulákkal kölcsönhatásba lépve, képesek elkülönült fázist létrehozni vizes közegben. Leggyakrabban említett formája a folyadék-folyadék fázisszeparáció (liquid-liquid phase separation, LLPS), amelynek során a létrejövő elkülönült fázis sűrű folyadékként viselkedik, némileg hasonlóan pl.

vízben lebegő olajcseppekhez. Ebben a külön fázisban a fehérjék koncentrációja lényegesen magasabb, mint az azt körülvevő vizes közegben, és a jelenség jellemzője, hogy újabb fehérjemolekulák hozzáadására az egyes fázisokban az adott fehérje koncentráció nem, csak az elkülönült, tömény fázis mérete növekszik[45]. A fázisszeparáció létrejöttében kulcsszerepet játszó alapvető kölcsönhatások lényegében ugyanazok, mint amelyek a feltekeredett fehérjék stabilizálásában is szerepet játszanak, a speciális jelleget ezek eltérő egyedi hozzájárulása és az általuk kialakított kölcsönhatási mintázat jelentheti [46]. Jelen tudásunk szerint kiemelten fontos a többértékű, multivalens kölcsönhatások megléte, mely a fehérjék szintjén tipikusan ismétlődő kötőrégiók – melyek éppúgy lehetnek globuláris domének mint lineáris motívumok – formájában jelenik meg. A jelenség kapcsán fontos felismerés, hogy léteznek a szeparált fázist létrehozni képes, ún. driver fehérjék [47], illetve olyanok, amelyek részt vesznek a kondenzátumok felépítésében, de azok kialakítását nem tudják iniciálni.

A fehérje fázisszeparáció jelenségét egyre több biológiai folyamattal hozzák összefüggésbe, tipikusan sejtmagi regulációs folyamatok kapcsán, ahol sok esetben RNS-kötő fehérjék játszanak kulcsszerepet a kondenzátumok kialakulásában, de kimutatták a posztszinaptikus denzitás egyes fehérjéi estében is [48]. A kialakuló képződményeket gyakran említik membrán nélküli sejtszervecskeként (membraneless organelle, MLO) is.

Ma már több olyan adatbázis is elérhető, amelyek kifejezetten a fázisszeparációban részt vevő fehérjéket, illetve azok jellemzőit listázzák, ilyen pl. a PhaSepDB [49] és a kifejezetten a driver fehérjékre fókuszáló PhasePro [47].

1.2.5. Fehérjekeletkezés nem kódoló genomi régiók átírásával

Az egyes fehérjecsaládok evolúciós rokonsága csak adott határig követhető vissza bizonyossággal. A legtöbb fehérjeosztályozó adatbázis egy adott szint fölött egymástól elkülönülő osztályokat ír le, a rokonsági viszonyokat csak ezeken belül tekintjük alátámasztottnak [50, 51]. Ezzel együtt feltehetően léteznek a jelenlegi módszerekkel nem feltárható rokoni kapcsolatok ma különállónak tekintett fehérjecsaládok között. Ugyanakkor a kérdés, hogy minden fehérje rokona-e egymásnak, csak részben filozófiai. Ma már számos bizonyíték szól amellett, hogy korábban nem fehérjekódoló szakaszok az evolúció során stabilan átíródó és fehérjévé leforítódó génekké válhatnak [52]. Ennek megfelelően az evolúciós folyamatok nem csak működő gének mutációkkal történő elvesztése, hanem újak kialakulásának irányában is folyamatosan hatnak. Ilyen, ún. *de novo* kialakult génekre ismerünk példákat az ember evolúciós közel-

múltjából is [53]. Miközben a *de novo* génkeletkezés transzkripciós és transzlációs oldalról való megvalósulása biokémiai ismereteink szerint adott mutációkkal fokozatosan elérhető ill. optimalizálható lehet, nehezebbnek látszik az a kérdés, hogy egy ilyen fehérjének milyen lehet térszerkezete, illetve ezzel összefüggésben a funkciója. Ez a felvetés már csak azért is érdekes, mert a fehérjék aggregációs hajlamát kutató tudósok körében elterjedt az az elképzelés, hogy a mai fehérjék evolúciós távlatban az aggregációs hajlam elkerülésére szelektált szekvenciával rendelkeznek [54]. „The amyloid state is more like the default state of a protein, and in the absence of specific protective mechanisms, many of our proteins could fall into it", illetve: „If you had a machine that could generate protein sequences randomly, you would only rarely get one that can remain stable in the globular, soluble state" idézi Christopher Dobsont Jim Schnabel[55]. Ezen logika alapján az újonnan születő fehérjék esetében ez a hajlam magas lehet, ami igencsak megnehezítheti ezek evolúciós fennmaradását, rögzülését. A kérdéskör tovább feszegethető az evolúció korai szakaszában potenciálisan kialakuló fehérjék szerkezeti preferenciáinak elemzésével, amikor még egyes feltételezések szerint nem a ma ismert, hanem egy egyszerűbb genetikai kód alapján valósult meg a transzláció. Ebben az időszakban a mai élőlényekre jellemző szabályozási és minőségbiztosítási folyamatok hiányában az aggregációs hajlam a mainál is komolyabb problémát jelenthetett, amennyiben ezen hajlam csökkentése valóban csak a szekvencia optimalizálásával érhető el.

1.3. Fehérjék belső dinamikájának vizsgálata sokaság-alapú modellekkel

A hagyományosnak tekinthető fehérjeszerkezeti modellek statikusak: a kísérletesen elérhető fehérjeszerkezetek túlnyomó többségét adó, röntgenkrisztallográfiával meghatározott szerkezetek esetében a PDB adatbázisban egy vagy néhány konformer található meg, melyek között általában minimális szerkezeti eltérések vannak[5]. Ugyanakkor minden szerkezetkutató tudja, hogy még a stabilnak tekintett globuláris fehérjék is dinamikusak, és ma már teljesen elfogadott, hogy a belső dinamika leírása a funkció megértésének feltétele [56]. A belső mozgások igazi atomi szintű részleteinek feltárására sokáig egy kizárólag elméleti módszert, molekuladinamikai számításokat használtak, melynek paraméterezését az elérhető erőterek (force fields)

[5]lényegesen különböző funkcionális állapotok között néha természetesen nagyobb léptékű szerkezeti átrendeződések is megfigyelhetőek, pl. motorfehéréjk vagy egyes vírusok sejtbe jutását segítő fehérjék esetében

formájában ma is folyamatosan finomítják [57]. A fehérjék belső mozgásairól atomi szintű részletességgel információt adó NMR-spektroszkópiai technikák fejlődésével megjelent az igény, hogy a kísérletileg kapott paramétereket és a molekuladinamikai számítások eredményeit összhangba lehessen hozni (pl. [58]). A probléma, hogy az aktuálisan elérhető legjobb erőterek esetében sem garantálható, hogy egy pusztán elméleti számításokkal kapott szerkezeti sokaság hűen tükrözze az adott molekula esetében méréssel meghatározott mozgékonysági paramétereket [59].

1.3.1. A sokaság-alapú modellek alapvető jellemzői

A fentebb felvázolt nehézségek egyik lehetséges kezelése a számítások és a kísérleti paraméterek megfelelő kombinálása. Ez koncepcionálisan alapvetően nem különbözik pl. a hagyományos szerkezetmeghatározási eljárásoktól, csupán a konkrét molekulára vonatkozó kísérleti paraméterek jellege és mennyisége más. Véleményem szerint fontosabb koncepcionális különbség, hogy míg a pl. az NMR-spektroszkpiával történő hagyományos szerkezetmeghatározás során a cél egyetlen jól definiált szerkezet megtalálása, mely egyidejűleg megfelel az összes kísérleti adatnak, a belső dinamika jellemzésére ún. sokaság-alapú modelleket használunk, melyek esetében a megfelelést a konformerek populációjának szintjén várjuk el [60]. Valójában igen valószínűtlen, hogy a valóságban akár egyetlen olyan konformer is jelen legyen, amely egyidejűleg megfelel az összes kísérleti paraméternek – melyek természetszerűleg egy sokaság átlagából számaznak. Ezt az általam fontosnak tartott gondolatot Michele Vendruscolo egyik közleményeben fogalmazták meg a H-D kicserélődési paraméterek kapcsán [61], én azonban ennél lényegesen általánosabb érvényűnek tekintem. A sokaság-alapú modellek legfontosabb jellemzője, hogy a konformerek változatosságának segítségével kínálnak a kísérleti paramétereknek való – az egyedi konformerek esetében elérhetőnél – jobb megfelelést. Ennek folyománya, hogy a sokaság által lefedett konformációs tér feltételezhetően jobban közelíti a molekula valós belső mozgásait, mint egyéb, pl. általános molekuladinamikai számításokból származó modellek esetében. Természetesen a sokaság-alapú modellek érvényességi köre is korlátozott, hiszen pl. tipikusan nem képesek egyidejűleg számot adni a nagyon széles időskálát felölelő belső mozgások mindegyikéről, hanem általában egy vagy néhány jól meghatározott paraméter – és az általuk reprezentált időskála – segítségével modellezhető dinamikai jellegzetességeket modellezik [62, 63]. A sokaság-alapú modellek megbízhatósága ennek megfelelően függ a felhasznált kísérleti paraméterek számától és jellegétől is.

1.3.2. Sokaság-alapú szerkezeti modellek előállítása

Sokaság-alapú modellek előállítására alapvetően kétféle megközelítés létezik: megkötéseket alkalmazó molekuladinamika és a nagy méretű, előre generált konformersokaságokból kiinduló szelekció. A megkötéseket alkalmazó molekuladinamikai eljárások általában a molekula több példányát modellezik egyidejűleg (multi-replica simulations), és az egyes példányok által reprezentált konformerek alkotják azt a sokaságot, amelyen a kísérleti paraméterek mint megkötések értelmezhetőek. A megkötésektől való eltérés plusz energiatagként jelenik meg, és az abból számított erők az egyes példányokat olyan irányba „terelik", hogy a sokaság összességében minél jobban megfeleljen a paramétereknek. Szokásos a gyakorlatban ezeket úgynevezett „félharmonikus" módon implementálni, mely ebben az esetben azt jelenti, hogy az energiaszámítás viszonyítási pontja nem egy fix álllapot, hanem a szimuláció során korábban elért legjobb megfeleléshez tartozó érték, mely jobb megfelelés elérése esetén módosul. Ilyen módon biztosítható, hogy a paraméterek által leírt állapottól távoli konformációk esetében is fokozatos legyen a megfelelés kialakulása, ne ébredjenek túlságosan nagy erők a rendszerben emiatt. Az egyes paraméterek esetében fontos szempont még, hogy nem feltétlenül ugyanakkora sokaságon érdemes őket átlagolni a túllillesztés veszélye miatt [64].

A sokaságok előállításának másik fő módszere az, hogy egy előzetesen generált konformersokaságból szelekciós eljárás segítségével választunk ki egy olyan alsokaságot, amely megfelel az adott paramétereknek. Ez az eljárás elsősorban a nagy konformációs szabadsággal rendelkező funkcionálisan rendezetlen fehérjeszakaszok esetében használatos [65, 66]. Természetesen a konformergenerálás és a szelekció iteratív módon összeköthető úgy, hogy a konformációs teret a tesztelt sokaságok előállításakor az előző lépésekben legjobban teljesítő szerkezetek felé toljuk el.

A szelekciós eljárás valamivel szélesebb körben alkalmazható, mint a megkötéseket használó molekuladinamika, ugyanis utóbbi esetben feltétel, hogy a szerkezetből visszaszámolt paraméterek segítségével kapott energiafüggvény deriválható legyen. A szelekció esetében elegendő, ha a szerkezet és a paraméterek közötti megfelelés számszerűsíthető, ami egy lényegesen enyhébb feltétel. Például tudomásom szerint a kémiai eltolódások szerkezetből való becslésére a CamShift eljárás előtt [67] nem volt olyan implementáció, amely a deriválhatóságot biztosította volna.

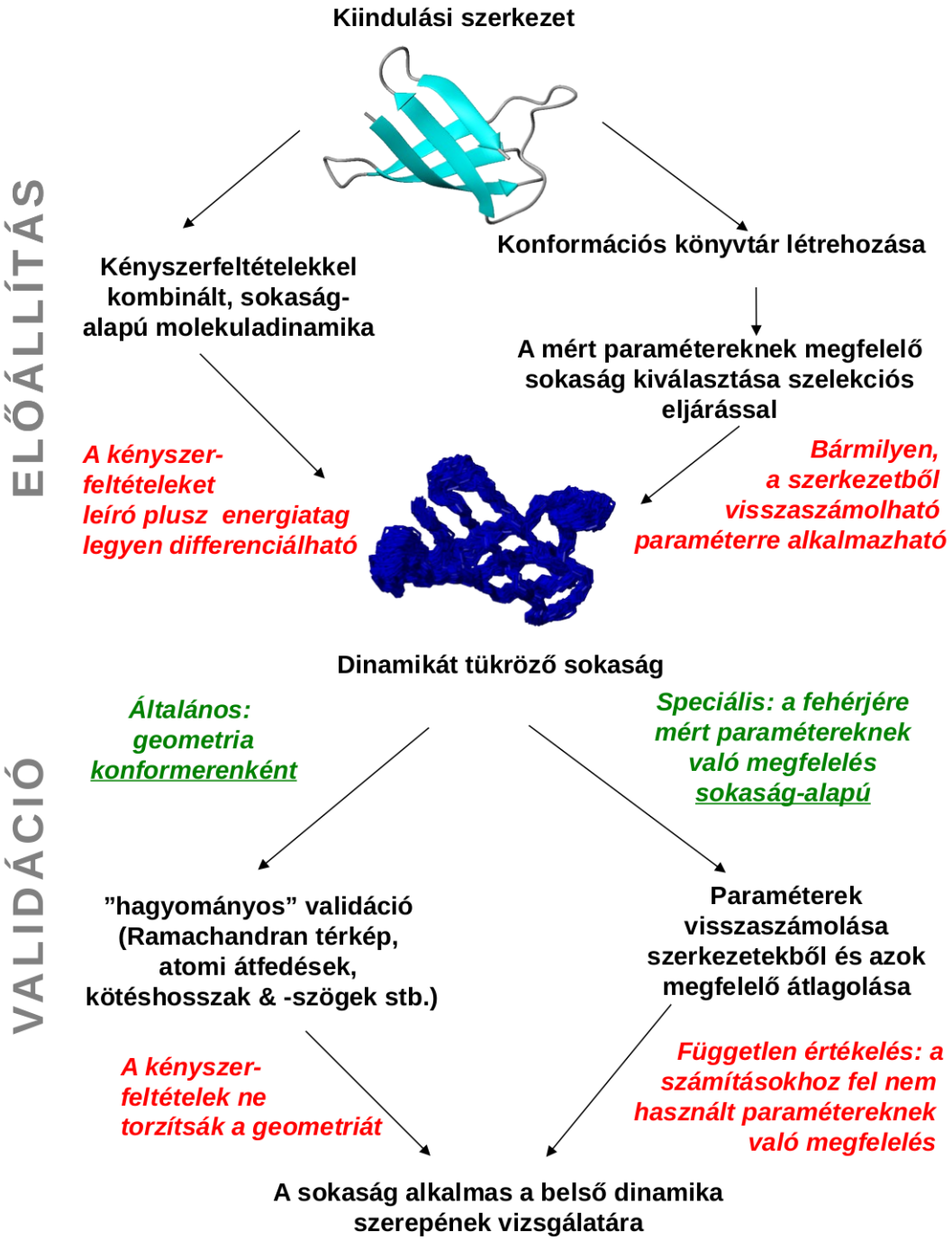
1.3.3. Sokaság-alapú modellek validálása

A dinamikus szerkezeti sokaságok minőségellenőrzése hasonló a hagyományos szerkezetmeghatározó eljárásokkal készült szerkezetekéhez, a sokaságon való átlagolás mint lényeges aspektus kivételével [68]. Minden konformer alapvető fizikai-kémiai jellemzőinek reálisnak kell lennie, azaz nem tartalmazhatnak atomi ütközéseket, a kötéshosszaknak, kötésszögeknek és torziós szögeknek pedig szintén az elvárható tartományba kell esniük.

Minden fehérjeszerkezeti modell esetében lényeges kérdés annak megbízhatósága, valóságnak való megfelelése. Ezt az aspektust a modell szabatosságaként (az angol nyelvű szakirodalombnn „accuracy") tudjuk definiálni [3]. Mivel a valóságot a molekulák esetében csak közvetve tudjuk megtapasztalni, a szabatosság megítélése elsősorban független kísérletek összevetésével lehetséges: ha két független mérés alapján nagyon hasonló modellt kapunk, akkor nagyon kicsi annak az esélye, hogy mindkétszer ugyanolyan komoly hibát elkövetve kapunk egymástól függetlenül hasonló, de helytelen modelleket. A szerkezeti biológiában a legkézenfekvőbb eset – amennyiben lehetőség van rá – ugyanazon makromolekula szerkezetének röntgekrisztallográfiával és NMR-spektroszkópiával történő független meghatározása[6].

Az NMR-spektroszkópia nagy előnye, hogy számos olyan paraméter mérését teszi lehetővé, amelyek jó közelítéssel egymástól függetlennek tekinthetőek – legalábbis abban az értelemben, hogy sokszor más-más módon tartalmaznak információt a lokális és a globális szerkezeti jellemzőkről, emellett nem is feltétlenül használjuk fel mindet egy adott szerkezetmeghatározási folyamat során. A gyakorlatban ez azt jelenti, hogy ideális esetben egy adott szerkezeti modell számos különböző, közvetlenül nem felhasznált paraméterkészletnek való megfelelése vizsgálható. A hagyományos szerkezetmeghatározás során pl. fontos szempont, hogy a nukleáris Overhauser effektuson (NOE) alapuló távolság jellegű kényszerfeltételek milyen mértékben teljesülnek [69].

[6]ahogyan ez a modern fehérje NMR spektroszkópia hajnalán konkrétan meg is történt Kurt Wüthrich visszaemlékezései szerint, hogy az NMR-szerkezetmeghatározás működőképességét mindenki számára be tudják bizonyítani



1.1. ábra. Dinamikus szerkezeti sokaságok előállításának és validációjának vázlatos áttekintése. Az ábra a Penicillium Antifungal Protein (PAF, PDB ID: 2MHV) szerkezeti modelljeinek felhasználásával készült.

2. A célkitűzések áttekintése

A dolgozatban leírt kutatások legfontosabb célkítűzései az alábbiakban foglalhatóak össze:

- Eljárás kidolgozása a magányos α-helikális régiók predikciójára, és erre támaszkodva a motívum elterjedtségének, valamint lehetséges funkcionális szerepének vizsgálata, egyes kiválasztott fehérjék szerkezeti modellezése.

- A predikciós eljárás kidolgozása során szerzett tapasztalatok alapján a fibrilláris és funkcionálisan rendezetlen szakaszokra vonatkozó predikciók kritikai elemzése, egymással összefüggésben történő értelmezése, ezáltal a biológiailag releváns térszerkezeti jellemzők feltárása.

- Az irodalomban jelenlévő diskurzus és saját megfigyeléseink nyomán az újonnan keletkező (*de novo*) fehérjék szerkezeti preferenciáinak elemzése jelenlegi, illetve az evolúció korai szakaszában feltételezett állapotok esetében a rendezett-rendezetlen átmenetek jobb megértése céljából.

- A fehérjék belső dinamikája funkcionális jelentőségének mélyebb megértése céljából olyan eljárások implementálása, melyek alkalmasak egyes kíséreltileg meghatározott, a dinamikai viselkedést tükröző paramétereknek megfelelő sokaság-alapú fehérjeszerkezeti modellek előállítására, illetve azok elemzésére.

- A kidolgozott eljárások felhasználásával konkrét, saját magunk, illetve együttműködő kutatócsoportok által vizsgált fehérjék ill. fehérjecsaládok szerkezet-dinamika-funkció összefüggéseinek vizsgálata dinamikus szerkezeti sokaságok előállításával, illetve azok összehasonlító elemzésével.

3. Az alkalmazott módszerek

3.1. Adatbázisok és webszerverek

Munkám során számos adatbázist használtam, melyek közül több igen széles körben ismert és elterjedt. E helyütt csupán ezek rövid felsorolására van lehetőség (3.1. táblázat). Az egyes konkrét elemzések esetében ezek eltérő verzióit tudtam felhasználni. A PDB adatbázisból származó, a szerkezeti elemzésekben lényeges szerepet játszó szerkezeteket külön alfejezetben listázom.

3.1. táblázat. A felhasznált fontosabb adatbázisok

Adatbázis	**A felhasznált adatok típusa**	**Érintett elemzés(ek)**
UniProt [70]	fehérjeszekvenciák	SAH keresés, keresztpredikciók
Gene ontology (PANTHERdb [71])	funkcionális annotációk	SAH keresés
DisProt [72]	rendezetlen fehérjeszakaszok	keresztpredikciók
PDB [73]	fehérjeszerkezetek NOE távolságlisták	szerkezetmodellezés, PRIDE-NMR és CONSENSX
BioMagResBank (BMRB) [74]	kémiai eltolódások és egyéb NMR paraméterek	CONSENSX és szerkezeti sokaságok
PhaSepDB [49]	fázisszeparációban általában részt vevő fehérjék	nagy töltéssűrűségű szakaszok elemzése
PhasePro [47]	fázisszeparációt iniciáló (driver) fehérjék és régiók	nagy töltéssűrűségű szakaszok elemzése

3.2. Predikciós eljárások

Az általunk alkalmazott predikciós eljárásokat a 3.2. táblázat foglalja össze. A predikciók elemezésekor mindig törekedtünk több, haosnló jellegeket becslő eljárás konszenzusásnak használatára. Fontos szempont volt még ezen kívül a vizsgált fehérjeszekvenciák redun-

danciamentessége, mind az elemzett teljes szekvenciák, mind adott motívumot tartalmazó szekvenciák esetében. Az ehhez szükséges szűréseket a CD-HIT [75] programmal végeztük. Az egyes predikciókkal kapott szekvenciaszakaszok összevetésére egyszerű, százalékos mérőszámok mellett az ún. Segment Overlap Measure (SOV) metrikát [76] is alkalmaztuk.

3.2. táblázat. A felhasznált fontosabb predikciós eljárások

Eljárás	**Predikció jellege**	**Érintett elemzés(ek)**
IUPRED [77]	rendezetlen fehérjeszakaszok	keresztpredikciók, *de novo* fehérjék
VSL2B [78]	rendezetlen fehérjeszakaszok	keresztpredikciók, *de novo* fehérjék
RONN [79]	rendezetlen fehérjeszakaszok	keresztpredikciók, *de novo* fehérjék
COILS [26]	coiled-coil régiók	keresztpredikciók
PAIRCOIL2 [80]	coiled-coil régiók	keresztpredikciók
HMMER [81] (kollagén HMM)	kollagén hélixek	keresztpredikciók
HMMTOP [82]	transzmembrán régiók	*de novo* fehérjék
DASTMFILTER [83]	transzmembrán régiók	*de novo* fehérjék
TMHMM [84]	transzmembrán régiók	*de novo* fehérjék
PHOBIUS [85]	transzmembrán régiók	*de novo* fehérjék
TANGO [86]	aggregációra hajlamos régiók	*de novo* fehérjék
WALTZ [87]	aggregációra hajlamos régiók	*de novo* fehérjék
FOLDAMYLOID [88]	aggregációra hajlamos régiók	*de novo* fehérjék

3.3. Programnyelvek és környezet

A dolgozatban leírt eredmények eléréséhez elsősorban PERL és PYTHON nyelvű kódokat írtam ill. írtunk a munkatársaimmal/hallgatóimmal. A fejlesztések egyes szakaszaiban igénybe vettük a JUPYTER NOTEBOOK környezetet. A GROMACS programcsomag [89] módosításakor az adott C nyelvű forrásfájlokat írtuk át, ügyelve az eredeti verzió helyreállíthatóságára és arra, hogy lehetőség szerint az annak megfelelő futtatható verzió a módosításainktól független módon elérhető legyen. Egyes molekuladinamikai eredmények értékelésekor a PDB formátumú szerkezeti állományok feldolgozása JAVA nyelvű programokkal történt.

Az egyes programnyelveket és felhasználásuk rövid öszegzését a 3.3. táblázat tartalmazza.

3.3. táblázat. A bemutatott munka során alkalmazott programnyelvek, a főbb programozási feladatok és az azok megvalósításában részt vevő munkatársaim

Nyelv	**Felhasználás**	**Részt vevő munkatárs**
C	GROMACS módosítás	Zajácz-Epresi Nóra
C++	PRIDE-NMR	Ángyán Annamária
JAVA	Molekuladinamikai elemzések	Kovács Bertalan
MATLAB	Nagy töltéssűrűségű régiók elemzése	Szabó András László
PERL	SAH keresés és elemzés	Tóth Gábor (SCAN4CSAH)
	Fibrilláris predikciók futtatása és elemzése	Szappanos Balázs
	Molekuladinamikai elemzések, CONSENSX (első változat)	Ángyán Annamária, Szappanos Balázs
PYTHON	Coiled-coil modellezés	Dobson László
	CONSENSX$^+$	Dudola Dániel
	DIPEND	Harmat Zita

3.4. Fehérjeszerkezeti adatok eredete és feldolgozása

A fehérjeszerkezeti adatok felhasználása során törekedtünk arra, hogy ezeket megbízható forrásból, együttműködő partnerektől szerezzük be, természetesen ezzel együtt ezek minden esetben publikált adatsorok. A BMRB adatbázisban elérhető adatokat sok esetben kiegészítettük további, szintén nyilvánosan elérhető, pl. a megfelelő közlemény kiegészítő anyagában közölt, vagy az adott közlemények szerzőitől elkérhető adatokkal. A 3.4. táblázat a bemutatott elemzésekben felhasznált legfontosabb szerkezeti adatok forrását részletezi. A külön megjegyzés nélkül feltüntetett hivatkozások a PDB kódok által reprezentált szerkezeteket ill. – ahol releváns – az NMR paramétereket leíró közleményre vonatkoznak.

Az NMR-spektroszkópiai mérésekből származó paraméterek felhasználása kapcsán lényeges aspektus az atomi nevezéktanok egységesítése, mely esetében a BMRB oldalán található konverziós táblázat szolgált iránymutatásul. Tekintetettel kell lenni azonban arra, hogy nem lehetünk teljesen biztosak abban, hogy bármely korábbi, az adatbázisokban elérhető állományok létrehozásakor elvégzett formátumkonverziós lépés megőrizte a sztereospecifikus jelleget, azaz pl. geminális atomok neve felcserélődhetett. Emiatt a NOE-alapú kényszerfeltétel-listákat a legtöbb esetben úgy módosítottam, hogy a sztereospecifikus kényszerfeltételeket minden lehetőséget megengedőre cseréltem egy nagyobb távolságküszöböt használva, majd a listát redundanciamentessé tettem. Ez amiatt is szükséges, mert a GROMACS a távolságok átlagolását nem pontosan úgy végzi, mint a szerkezetszámoló programok, pl. a CYANA [90].

3.4. táblázat. Fehérjeszerkezetek modellezéséhez felhasznált koordináták és kísérleti adatok eredete

Molekula	PDB kód	BMRB kód	A szerkezeti és NMR adatokat leíró közlemény / megjegyzés
SAH szakaszok helyzetének modellezése a paraspeckle struktúrában			
NONO/PSPC1 heterodimer	3SDE	nem releváns	[91], csak a koordinátákat használtuk
Kisméretű szerinproteáz-inhibitorok dinamikájának elemzése			
Schistocerca gregaria chymotrypsin inhibitor (SGCI)	1KGM	5272	[92], további NMR adatok a [93] közleményből
Schistocerca gregaria trypsin inhibitor (SGTI)	1KJ0	5274	[92], további NMR adatok a [93] közleményből
Parvulin típusú prolil cisz-transz izomerázok összehasonlító elemzése			
Staphylococcus aureus PrsA	2JZV	15628	[94]
Trypanosoma brucei Pin1	2LJ4	17918	[95]
Cenarcheaum symbiosum PinA	2RQS	11080	[96]
A számos parvulint érintő összehasonlításban felhasznált szerkezetek azonosítóit a XVIII. közlemény kiegészítő anyaga tartalmazza			
Epesavkötő fehérjék ligandumkötésének vizsgálata			
Humán epesavkötő fehérje, apo forma	1O1U	(nincs)	[97]
Humán epesavkötő fehérje, koliltaurin komplex	1O1V	(nincs)	[97]
Humán epesavkötő fehérje, kétféle epesavval komplexben	2MM3	19843	[98], további NMR adatok a [99] közleményből
PDZ domének elemzése			
PSD95 tandem PDZ1-2 domének (kötött forma)	2KA9	(nincs)	[100], NMR adatok a szabad formára is
PSD95 PDZ3, szabad forma	1BFE	-	a PDZ3 doménre vonatkozó NMR
PSD95 PDZ3, ligandumkötött forma	1BE9	-	adatok forrása a [101] közlemény
A számos PDZ domént érintő összehasonlításban felhasznált szerkezetek azonosítóit a XX. közlemény kiegészítő anyaga tartalmazza			
SAH szakasz dinamikájának modellezése			
Miozin VI SAH domén	6OBI	30591	[35]

A NOE-alapú kényszerfeltétel-listákat és az S^2 rendparamétereket a molekuladinamikai számításokhoz a GROMACS topológia (.TOP) fájlokba illesztettem be megfelelő formátumban. A kapott sokaságok vizsgálatához a különböző paramétereket a BMRB NMR-STAR formátumába [102] konvertáltam, ahol ez külön szükséges volt (tipikusan a közlemények kiegészítő anyagaiban leírt vagy szerzők által elküldött paraméterek esetében), ez szolgált CONSENSX (4.4.3. alfejezet) elemzések bemeneteként.

3.5. Molekuladinamikai számítások és elemzésük

3.5.1. A GROMACS programcsomag

A molekuladinamkai számításokhoz minden esetben a szabad hozzáférésű GROMACS [89] programcsomag valamilyen verzióját használtuk, a feladattól függően sok esetben saját magunk által módosítva (lásd 4.4.1. alfejezet). A legtöbb, ebben a dolgozatban leírt számítást explicit vízmodellel végeztük, vagy az OPLS [103], vagy az Amber99SB erőtér valamilyen verzióját (tipikusan Amber99SB-ILDN [104]) használva. Az előkészítés tipikusan egy minimalizálási és egy helyzeti megkötéseket alkalmazó (position restrained) dinamikából állt. A több replikát használó párhuzamos számításoknál rendszerint az eredeti MUMO protokollt leíró közleményben [64] szereplő 8 replikát alkalmaztuk, mindegyiket más-más véletlenszerű atomi kezdősebességekkel indítva. Fontos megjegyezni, hogy ezeknél a számításoknál az összesített szimulációs idő tipikusan nyolcszor néhány 10 ns, tehát a 100 ns nagyságrendbe esik, ami elmarad az ezen technikákat alkalmazó kutatásaink publikálásakor a hagyományos futások esetében már egyre gyakrabban alkalmazott, μs nagyságrendet közelítő szimulációs időktől. Ennek indoka, hogy jelenleg ezek a(z egyetlen replikát alkalmazó) számítások nagymértékben támaszkodnak a GPU-alapú gyorsító rutinokra, amivel az általunk alkalmazott, az egyes molekulapéldányokat külön CPU-n (vagy szálon) futtató számítások hatékonysága nem összevethető sem a GPU gyorsítás hiánya, sem az egyes szálak közötti kommunikációs igény megléte miatt. Az ilyenfajta párhuzamos számítások GPU-t hatékonyan kihasználó változatának implementálása messze túlmutat a lehetőségeinken, és tudomásom szerint ez nem opció az elérhető egyéb programcsomagokban sem. Másrészt az is igaz, hogy ez a hátrány nagyrészt csak látszólagos, hiszen a kísérleti megkötéseket használó molekuladinamikai számítások esetében a modellezett időskálát nem elsősorban a szimuláció lépések által meghatározott hossza, hanem az alkalmazott paraméterek által leírt időskála határozza meg (1.3.1.

alfejezet). Ezzel együtt is érvényes ugyanakkor, hogy a megkötéseket alkalmazó szimulációk esetében is nagyobb lehet a bejárt konformációs tér, ha hosszabb a szimulációs idő.

A konformációsentrópia-változások becslésére a GROMACS programcsomagban elérhető elemző eljárásokat használtuk.

A számítások eredményeinek megjelenítéséhez a MOLMOL [105] és CHIMERA [106] programokat használtam. Az elemzések többsége tipikusan saját, legtöbbször PERL, ritkábban JAVA nyelvű programokkal történt (3.3. táblázat).

3.5.2. Főkomponens-elemzés

A molekuladinamikai számításokkal (vagy egyéb módszerekkel) kapott szerkezeti sokaságok elemzésére gyakran alkalmaztunk főkomponens-elemzést (principal component analysis, PCA). Ezt a PRODY [107] programcsomag segítségével, annak eljárásait közvetlenül terminálból elérve, vagy azokat vagy PYTHON nyelvű szkriptekből meghívva végeztük. A főkomponens-elemzés elvégzésének feltétele, hogy a bemeneti szerkezetek egymásra egyértelműen illeszthetőek legyenek, ugyanannyi (pl. Cα) atomot tartalmazzanak – ez a követelmény csak pl. egy adott molekuladinamikai számításokból származó sokaságoknál triviális. Erre az aspektusra tehát külön figyelemmel kellett lenni a bemenet előállításakor pl. különböző szekvenciájú, de rokon szerkezetű fehérjék összehasonlító elemzésénél. Ehhez többszörös térszerkezet-illesztést végeztünk a MAMMOTH-MULT program [108] segítségével, és annak alapján azonosítottuk a egymásnak térben megfelelő aminosavakat, ill. Cα pozíciókat. Ennek alapján a csak a közösnek tekinthető Cα pozíciókat és minden modell esetében egységes számozással tartalmazó PDB formátumú fájlokat állítottunk elő, ezek szolgáltak a PCA elemzés közvetlen bemenetéül. A konkrét pozíciók száma és mibenléte függ az elemzésbe bevont fehérjék változatosságától: távolabbi rokon fehhérjékben tipikusan kevesebb közös pozíció azonosítható.

4. Eredmények

Ebben a fejezetben az eredmények ismertetésére és a keletkezésük idején érvényesnek tekintett értelmezésére szorítkozom. Az irodalomban ezek után megjelent eredményekre a diszkusszóban (5. fejezet) térek ki, ahol ez releváns.

4.1. Magányos α-helikális (SAH) régiók detektálása, elemzése és modellezése

4.1.1. SAH régiók detektálására alkalmas bioinformatikai eljárás kifejlesztése

Az FT_CHARGE eljárás működési elve

Az első SAH régiók felfedezése után megjelent az igény arra, hogy több hasonló szegmenst tudjunk azonosítani azok átfogó vizsgálata céljából. A detektálásra az általános, hasonlóságon alapuló bioinformatikai módszereket, mint pl. a BLAST [109], nem tartottuk alkalmasnak, hiszen azok alapvetően magas komplexitású szekvenciákhoz evolúciósan rokon szakaszok azonosítására optimalizáltak. A magányos hélixek esetében pedig az alacsony komplexitás miatt csupán a szekvencia alapvető jellegeinek, de nem (feltétlenül) konkrét pozícióknak a konzerváltságát várhattuk, illetve feltételeztük, hogy – éppen alacsony komplexitásuk miatt – ilyen tulajdonságú szakaszok konvergens evolúcióval egymástól függetlenül többször is kialakulhattak. Ezért az ismert magányos hélixek töltésmintázatának felismerésére célzottan alkalmas eljárások kifejesztését határoztuk el. Nyitray Lászlóval együttműködésben Tóth Gábor és jómagam egymástól függetlenül, más-más koncepció mentén kezdtünk bele az eljárások kifejlesztésébe. A Tóth Gábor által készített program, a SCAN4CSAH az egyes, a szekvenciában egymástól adott távolságra elhelyezkedő, azonos vagy ellentétes töltésű ami-

nosavpárok pontozásán alapul, a pontozást az akkor ismert magányos hélixek felismerésére optimalizálva. Az általam kifejlesztett FT_CHARGE eljárás (I. közlemény) az ellentétes töltésű aminosavak szabályos váltakozásának Fourier-transzformáció (FT) segítségével történő elemzését végzi, koncepciójában támaszkodva az irodalomban korábban leírt, FT-alapú, ismétlődő szekvenciamotívumokat felismerő Spectral Repeat Finder (SRF) programra [110]. Ennek alapján első lépésben az ún. töltéskorrelációs függvényt definiáltam:

$$R(n) = \sum_{i=1}^{l-n} c(i)c(i+n) \tag{4.1}$$

ahol l a vizsgált szekvencia hossza, n két aminosav pozíciójának távolsága a szekvenciában, $c(i)$ az i. pozícióban lévő aminosav formális töltése (Arg, Lys esetében +1, Asp, Glu esetében -1, His esetében +0.5). A képlet alapján az azonos egységnyi töltésű aminosavak esetében a szorzat 1, ellentétes egységnyi töltésűek esetében -1. A következő lépés $R(n)$ Fourier-transzformációja és a legnagyobb amplitúdóhoz tartozó frekvencia azonosítása. Természetesen lényeges kérdés, hogy a legnagyobb amplitúdó hogyan viszonyul az adott aminosav-összetétel mellett véletlenszerűen várhatóhoz, azaz a találat szignifikánsnak tekinthető-e. Ennek eldöntése az eljárás első változatában a bemeneti szekvencia randomizált változatainak segítségével történt, mely megközelítést a későbbiekben egy determinisztikus megoldásra cseréltem (4.1.1. alfejezet).

A magányos α-hélix szerkezetekben tipikusan „egy fordulatnyi", azaz három vagy négy azonos töltésű aminosavat követ ugyanennyi ellentétes töltésű aminosav, ennek megfelelően az 1/6 és 1/9 közötti domináns frekvenciával jellemezhető szakaszokat tekintettem SAH szakasznak. A PERL nyelven írt program technikai megvalósítása ún. fast Fourier transform (FFT) eljárást implementáló modult használ (MATH::FFT PERL modul), ezért a szekvenciákat 16, 32 és 64 aminosavas szakaszokban vizsgálja. Az ablakok meghatározása során változtatható, hogy hány aminosav legyen két egymást követő ablak kezdőpontja között, természetesen a legbiztosabb – azaz semmilyen lehetőséget nem kihagyó – megoldás az, ha ez a távolság 1 aminosav, azaz az összes lehetséges szekvenciaablakot elemzi a program, ez biztosítja a szegmenshatárok lehető legpontosabb megállapítását is. E két tényező – FFT és egy szekvencia több ablakban való elemzése – az eljárás eredeti formájában túl lassú sok szekvencia hatékony vizsgálatához.

Az első fejlesztések időszakában a Tóth Gábor kollégám által írt, az egymás utáni aminosavpárok és -hármasok pontozásos értékelésén alapuló SCAN4CSAH készült el korábban.

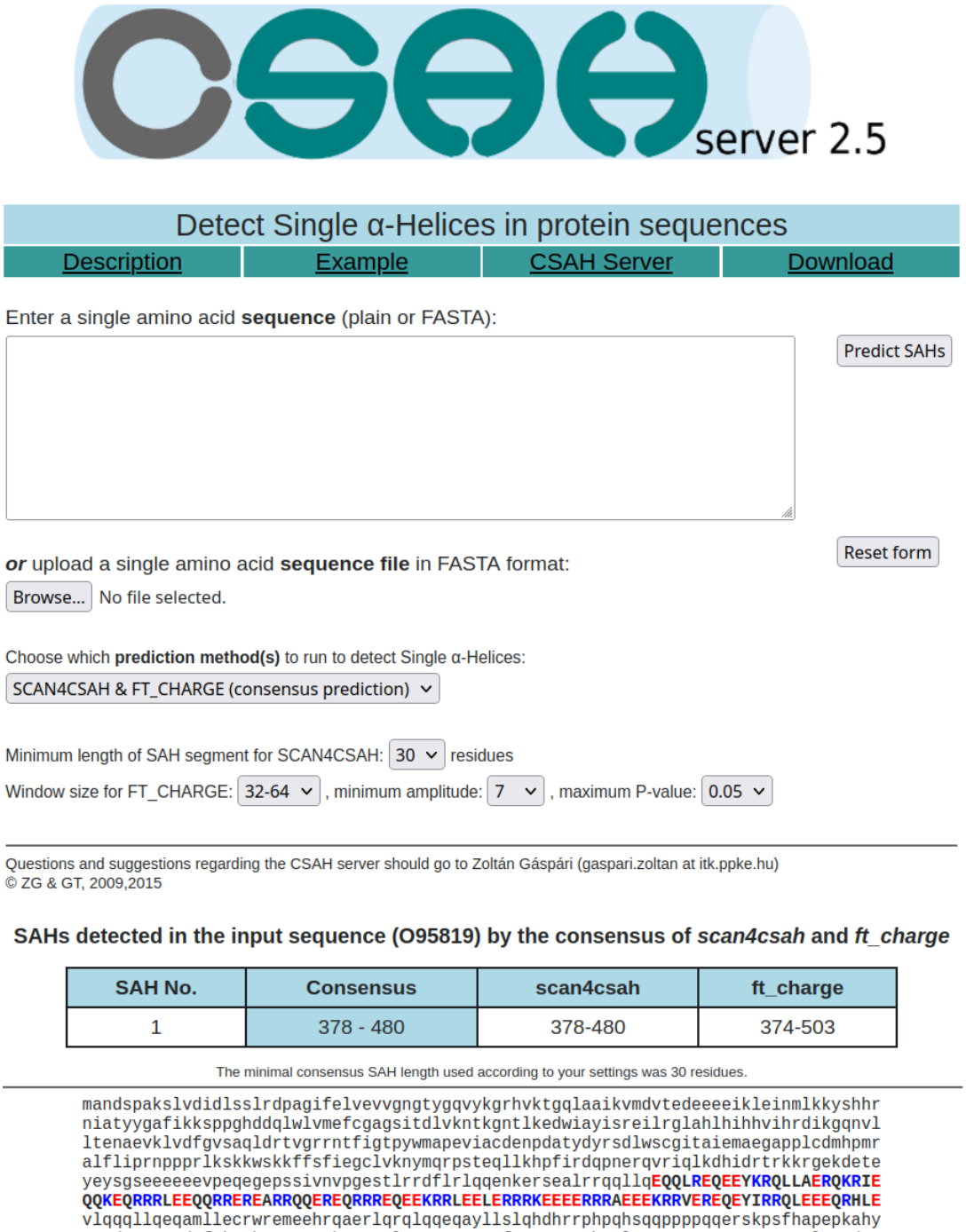
Hamar nyilvánvalóvá vált, hogy ennél az FT_CHARGE lényegesen szigorúbb, azaz kevesebb találatot ad. Abból a megfontolásból, hogy mindenképpen elkerüljük a SAH szakaszok „túlpredikcióját", azaz alacsonyan tartsuk a hamis pozitív találatok számát, a két eljárás konszenzusának alkalmazása mellett döntöttünk. Ezen megközelítés létjogosultságát sikerült kísérletesen is igazolnunk. Kiválasztottunk három, mindkét eljárás által magas pontszámmal megtalált SAH szakaszt, melyeket Nyitray László laboratóriumában Süveges Dániel kísérletileg előállított és jellemzett, bizonyítva a SAH szerkezeti elem meglétét (I. közlemény; lásd alább a 4.1. táblázatban a GCP60 és MAP4K4 fehérjéket).

A két eljárást technikailag egy „csomagoló" szkriptben fogtam össze, ez a CSAHDETECT.PL, mely háromféle módon képes SAH szegmenseket keresni, a SCAN4CSAH és FT_CHARGE egyikével, vagy a kettő konszenzusát használva. Az eljárás és a programok 2012 óta webszerverként (CSAHSERVER, 4.1. ábra) és letölthető PERL kódként is elérhetőek a csah-server.itk.ppke.hu weboldalon (II. közlemény). Ezek természetesen tartalmazzák az alábbi alfejezetben részletezett továbbfejlesztéseket is.

Az FT_CHARGE eljárás továbbfejlesztése

Az eljárás első változatában a kapott amplitúdó szignifikanciájának vizsgálata a konkrét szekvenciaszakasz randmoizálásával történt, ez azonban nem volt kellő mértékben reprodukálható. Emiatt a későbbiekben ehhez különböző háttéreloszlásokat készítettem, amelyeket viszonyítási alapként lehetett használni (II. közlemény). Ezek úgy készültek, hogy előre definiált aminosavgyakoriságú szekvenciákból származtatott adatokra extrémérték-eloszlást (EVD) illesztettem, amelynek paraméterei az eljárás bemenetét képezik. Ezek segítségével minden szegmensre a program meghatározza, hogy a háttéreloszlásban szereplő mely szekvenciacsoport aminosavösszetételéhez áll legközelebb, és az arra vonatkozó EVD paraméterek segítségével becsüli meg a kapott amplitúdó előfordulási valószínűségét (P érték). A meghatározott aminosav-összetételű szekvenciák úgy készültek, hogy minden egyes ablakméretre véletlenszerű, csak Ala, Arg és Glu aminosavakat tartalmazó szekvenciákat készítettem úgy, hogy azok különböző töltéssűrűségnek feleljenek meg 10%-os lépésekben, pl. 10% Arg - 30% Glu. Minden ilyen lehetőségre 5000-5000 szekvenciát generáltam, majd ezeken kiszámoltam az amplitúdokat. Az EVD illesztés az R statisztikai programcsomag [111] segítségével történt.

Az FT_CHARGE eljárás találatait nagyobb adathalmazokon elemezve megfigyeltük, hogy ritkán ugyan, de előfordul, hogy viszonylag sok prolin aminosavat tartalmazó szakaszok is



4.1. ábra. A CSAHSERVER webszerver jelenlegi verziójának nyitólapja (fent) illetve az M4K4_HUMAN fehérje (Mitogen-activated protein kinase kinase kinase kinase 4) szekvenciájára alapértelmezett beállításokkal adott kimenete (alul). A szerver megadja a két detektáló eljárás által kiadott határokat, a konszenzust, valamint a szekvenciában félkövér nagybetűkkel kiemeli a SAH szakaszt, a töltéssel rendelkező aminosavakat színezve (kék: pozitív, piros: negatív töltésű aminosav)

bekerülnek a találati listába. Ennek oka, hogy az eljárás alapjául szolgáló töltéskorrelációs függvény csupán a töltéssel rendelkező aminosavak eloszlását vizsgálja, az egyéb aminosavak előfordulására nincs tekintettel. Ugyanakkor a prolin „hélixtörő" jellege miatt valószínűtlen, hogy a mégoly szabályos, és SAH szakaszokra jellemző töltésmintázatot mutató prolingazdag szekvenciák α-helikális szerkezetet vegyenek fel. A problémára általános megoldást keresve egy „helicitási szűrő" beépítése mellett döntöttünk (III. közlemény), melyhez az aminosavak egyszerű konformációs preferenciáit használtuk fel a Chou-Fasman skála [112] alapján. Ennek indoka, hogy bár ma már ennél számos pontosabb eredményt adó eljárás áll rendelkezésre a másodlagos szerkezeti elemek becslésére, azok rendszerint felhasználnak a vizsgálttal homológ szekvenciákból nyert mintázatokat is [113], amelyet a SAH elemek vizsgálatakor nem tartottunk kívánatosnak.

A helicitási szűrő implementálása során a PDB SELECT adatbázisban található fehérjeláncok DSSP annotációja alapján kiválogattuk az α-helikális szakaszokat. Az ezeket alkotó aminosavak Chou-Fasman helicitási preferenciái (helix propensity) [112] alapján[1] a szakaszokra átlagos helicitási indexet számoltunk ki. A kapott értékekre EVD eloszlást illesztve meghatároztuk a pontértékek adott alsó százalékának megfelelő határokat. Tapasztalataink alapján a szűrőt két lépés során is alkalmazzuk, először az FT_CHARGE által azonosított egyedi szakaszok (16, 32 vagy 64 aminosav hosszú szegmensek) esetében, majd az ezekből összerakott – adott esetben a SCAN4CSAH által szűkített konszenzus – teljes SAH szekvencia esetében is. Az első lépésre azért van szükség, mert az átlagos helicitás egy nagyobb szekvencia esetében nem feltétlenül tükrözi a kisebb szakaszok preferenciáit, az utólagos szűrés pedig mindenképpen szükséges, hogy a teljes SAH-detektáló eljárás kimenete egységes legyen, a második szűrő ugyanis akkor is aktív, ha a felhasználó csak a SCAN4CSAH eljárás kimenetét vizsgálja a CSAHDETECT.PL szkript segítségével.

Az eredeti, PERL nyelven írt FT_CHARGE eljárás legnagyobb hátránya az FFT lépés lassúsága volt. A PPKE ITK-n belüli együttműködés során Nagy Zoltán kollégám és Kovács Ákos, akkor az ITK hallgatója elkészítették az eljárás FPGA-alapú implementációját [114], melynek technikai oldalát e helyütt – mivel a megvalósítás nem a saját hozzájárulásom – nem részletezem. Ez lehetővé tette a teljes UniProt adatbázis gyors elemzését és nagyskálás összehasonlító vizsgálatok elvégzését.

Az FT_CHARGE eljárás szigorúságát sokáig elvi okokból sem tartottam indokoltnak eny-

[1] a közlemény 1. táblázatában szereplő, α-hélix képző hajlamra vonatkozó értékeket használtuk

4.1. táblázat. Az FT_CHARGE eljárás újraparaméterezéséhez felhasznált SAH szakaszok áttekintése *(adatok az IV. közlemény 1. táblázata alapján)*

Név	UniProt ID	Uniprot AC	SAH régió	Hivatkozás
Caldesmon	A0A1L1RXH5	A0A1L1RXH5_CHICK	196 - 252	[115]
GCP60	Q9H3P7	GCP60_HUMAN	183 - 238	I. közlemény
INCENP	P53352	INCE_CHICK	503 - 715	[38]
MAP4K4	O95819	M4K4_HUMAN	417 - 480	I. közlemény
MFAP1	P55081	MFAP1_HUMAN	267 - 344	[39]
Myosin 6	Q9UM54	MYO6_HUMAN	915 - 980	[116]
Myosin 7	P97479	MYO7A_MOUSE	866 - 935	[117]
Myosin 10	Q9HD67	MYO10_HUMAN	813 - 909	[41]
Snu23	G0S6R0	G0S6R0_CHATD	131 - 164	[39]

híteni, mivel a nagyon kevés, kísérletesen jellemzett SAH szakasz megítélésem szerint nem nyújtott kellő alapot a motívum általános jellegzetességeinek meghatározásához. 2017-ben azonban a humán proteom újraelemzésekor már néhány olyan, kísérletesen igazolt SAH szekvencia is ismert volt, amelyet az eljárásunk éppen a szigorúsága miatt nem tudott azonosítani, így szükségessé vált a paraméterezés újragondolása. Mindezzel együtt a teszthalmaz még mindig nagyon szűk volt, összesen 9(!) olyan, szakirodalomban leírt szakaszt találtunk, amely természetben előforduló fehérjében azonosított SAH régiónak felel meg, ezek közül ráadásul az egyik, a Snu23 esetében nem osztom az azt SAH-ként leíró szerzők álláspontját, mivel a közölt CD mérések alapján a szakasz helicitása csupán 20% körüli érték [39], mely messze alatta marad a stabil SAH motívumok jellemzően legalább 60% körüli értékének.

Az FT_CHARGE eljárás fő paramétereinek (P-érték, minimális amplitúdó) szisztematikus változtatásával feltérképeztük, mennyire pontosan sikerül a 4.1. táblázatban fesorolt szegmensek határait megbecsülni, és ennek alapján az alapértelmezett paraméterek lazítása mellett döntöttünk (IV. közlemény). Ez természetesen megnövelte a detektált SAH régiók számát a humán proteomban (lásd alább).

Az FT_CHARGE eljárás gyakorlati alkalmazása

Az FT_CHARGE eljárás alkalmazása közben megfigyeltük, hogy egyes esetekben az ablakméret megválasztása befolyásolja a predikciót: nagy ablakméretnél a rövid, de szabályosan ismétlõdõ szakaszok, kis ablakméretnél pedig az alacsonyabb töltéssűrűségű, de hosszabb szekvenciarészleten szabályos régiók kerülhetnek a detektálási küszöb alá. Ezt elkerülendő, az eljárás során ugyanazt a szekvenciát több különbözõ ablakmérettel elemezzük, majd ezek

találatait összefűzve állapítjuk meg a SAH régiók meglétét és azok határait. Alapértelmezésben a 32 és 64 hosszú ablakokat vesszük figyelembe, a program opcióként a 16 aminosav hosszú ablakméretet is kezelni tudja.

A Tóth Gábor által kifejlesztett, itt részletesen nem tárgyalt SCAN4CSAH eljárás az egymás utáni aminosavpárok és -hármasok pontozásos értékelésén alapul. Már az eljárások fejlesztése során megfigyeltük, hogy a két algoritmus érzékenysége igen eltérõ, az FT_CHARGE eljárás jóval szigorúbb, lényegesen kevesebb SAH szakaszt jósol, mint a SCAN4CSAH. Ezért munkáinkban elsősorban a két eljárás konszenzusát használtuk, arra törekedve, hogy az azonosított SAH szakaszok nagy eséllyel valóban mutassák a magányos helikális jelleget. Ez a megközelítés szándékoltan alacsony hamis pozitív rátát eredményez, ugyanakkor magas hamis negatív rátával – azaz viszonylag sok SAH jellegű, de általunk nem akként felismert szakasz meglétével – járhat. Az eljárás kifejlesztésekor rendelkezésre álló igen kevés kísérleti adat miatt ezt az igen konzervatív megközelítést gondoltuk a leginkább megfelelőnek, elkerülendő a SAH szegmensek gyakoriságának és ezáltal biológiai relevanciájának esetleges eltúlzását. Mivel még jelenleg is igen alacsony az ismert, kísérletileg igazolt SAH motívumok száma, ennek a megközelítésnek a fenntartását továbbra is indokoltnak tartom.

Megemlítendő, hogy bár sem a SCAN4CSAH, sem az FT_CHARGE nem különböztetni meg az azonos töltésű aminosavakat, azaz adott pozícióban lévő Asp vagy Glu aminosav esetébenb ugyanolyan pontszámot ad, a becsült SAH szekvenciák egyértelmű preferenciát mutatnak a glutaminsav irányában az aszpragainsav rovására. Ez a tendencia jól magyarázható a glutaminsavnak az α-helikális szerkezetekre vonatkozó, korábban megfigyelt preferenciájával. A teljesség kedvéért meg kell jegyezni, hogy ez a különbség természetesen a helicitási szűrési lépésben már megjelenik, azonban mivel annak célja az alacsony helicitású szakaszok kizárása, a kimenet szempontjából az Asp/Glu megkülönböztetés ezzel együtt elhanyagolhatónak tekinthető.

4.1.2. A SAH régiók előfordulása fehérjékben

Az általunk kidolgozott predikciós eljárások lehetővé tették a SAH motívum előfordulásának és biológiai szerepének szélesebb körű vizsgálatát. Az első elemzéseink alapján a SAH régiót tartalmazó fehérjék mindössze az emberi proteom 0,2%-át alkotják. Ez a becslés igen konzervatív, az FT_CHARGE algoritmus fentebb kifejtett szigorú jellegéből adódóan és alsó becslésnek tekinthető. Mindezzel együtt az emberi proteom relatív értelemben viszonylag

gazdagnak tekinthető SAH szegmenst tartalmazó fehérjékben. Természetesen ennek egyik oka lehet az emberi genom/proteom többi organizmusénál pontosabb ismerete. A SAH szakaszok kifejezetten ritkák az általunk vizsgált virális proteomokban.

Szerkezeti oldalról igen jellemzőnek tekinthető az az eset, amikor a SAH szegmens részben átfed egy coiled-coil szerkezetűnek prediktált régióval. Ezen megfigyelésből kiindulva kezdtük el részletesebben megvizsgálni az egyes predikciós eljárások által azonosított különböző szakaszok közötti átfedéseket és a predikciók közötti összefüggéseket (lásd 4.2.1. alfejezet). Specifikusan a coiled-coil szakaszokkal való átfedés a kétféle szerkezeti elem közötti szerkezeti, funkcionális és esetlegesen evolúciós összefüggésekre utalhat.

Az egyik leginkább figyelemre méltó eredményünk ugyanakkor az, hogy a SAH régiók gyakran fordulnak elő RNS-kötő funkcióval bíró fehérjékben. Ez az eredmény azért igazán érdekes, mert csak a legutóbbi időkben merült fel, hogy maga a SAH mint hélix fizikailag közvetlenül részt vehetne nukleinsavakkal kialakított kölcsönhatásokban [36]. Ezzel szemben az általunk korábban talált funkcionális asszociáció a SAH régiók annotált RNS-kötő doménekkel (pl. RRM domén) való együttes előfordulásán alapult (II. közlemény).

Az RNS-kötő fehérjék közül említhetőek a paraspeckle[2] nevű sejtmagi struktúra kialakításában részt vevő molekulák (lásd alább), valamint az exon-junction komplex (EJC) felépítésében szerepet játszó UPF2, UPF3A és UPF3B fehérjék.

Az RNS-kötő funkcióval bíró molekulák mellett kiemelhetőek még a citoszkeletális fehérjék mint olyanok, amelyekben jellemzően előfordulnak SAH szegmensek (4.2. táblázat; II. és IV. közlemény). A troponin fehérjék mellett e helyütt a septin 7 molekulát kívánom még megemlíteni: a septineket tekinthetjük „4. típusú" citoszkleletális elemeknek, GTP-áz aktivitással rendelkeznek, és a sejtosztódásban, valamint egyes membránokkal kapcsolatos folyamatokban játszanak szerepet [118]. Érdekes módon SAH szegmenst csak a septin 7 fehérjében találtunk. Ez a fehérje az emberi sejtekre jellemző heterohexamer struktúra szélén található. A SAH szakasz a többi septinre is jellemző coiled-coil szegmens részeként azonosítható.

[2]Nincs tudomásom ezen képlet esetében bevett magyar névről, és magam nem kívánom e helyütt megkísérelni a fordítást

4.2. táblázat. SAH régiók elhelyezkedése és szekvenciája néhány kiválasztott emberi RNS-kötő (UPF2, UPF3A, UPF3B) és sejtvázhoz kapcsolódó (troponin T, septin 7) fehérjében. A szekvenciákban pirossal a negatív, kékkel a pozitív töltéssel rendelkező aminosavakat jelöltük. *(adatok a IV. közleményünk alapján)*

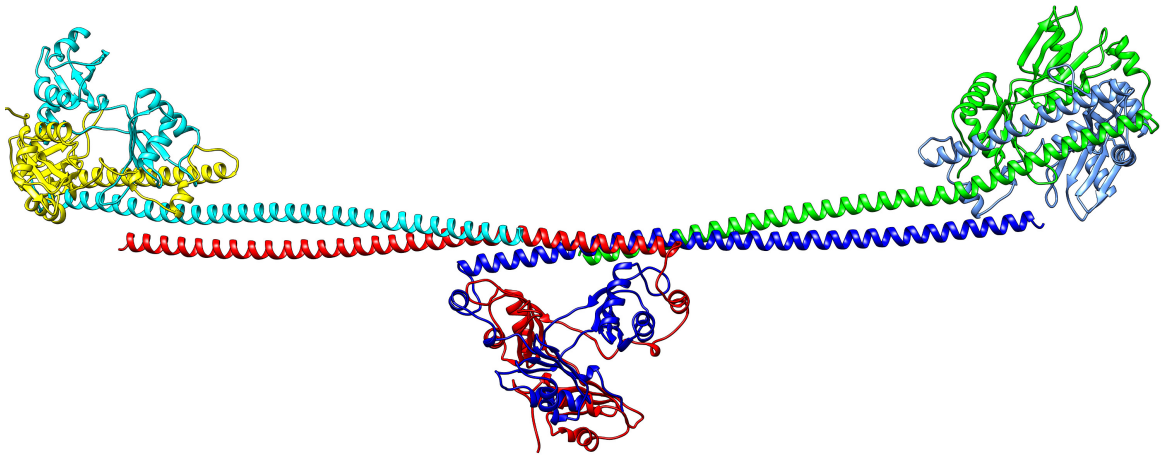
Uniprot azonosító	**Fehérje neve / rövidítése**	**SAH pozíció**	**SAH szekvencia**
Q9HAU5	UPF2	48 - 115	EVSKAPEDKKKRLEDDKRKKEDKERKKKDEEKVKAEEESKKKEEEEKKKHQEEERKKQEEQAKRQQEE
Q9H1J1	UPF3A	237 - 303	EERRRREL EKKRLREEEKRRRREEERCKKKETDKQKKIAEKEVRIKLLKKPEKGEEPTTEKPKERGE
Q9BZI7	UPF3B	209 - 269	RMREEKREERRRREIERKRQREEERRKWKEEEKRKRKDIEKLKKIDRIP
P45379	TNNT2 (szívizom troponin T)	142 - 185	AERAEQQRIRNEREKERQNRLAEERARREEEENRRKAEDEARKK
P13805	TNNT1 (lassú vázizom troponin T)	112 - 151	EQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKK
P45378	TNNT3 (gyors vázizom troponin T)	116 - 155	EQQRIRAEKERERQNRLAEEKARREEEDAKRRAEDDLKKK
Q16181	Septin 7	366 - 419	KEKVQKLKDSEAELQRRHEQMKKNLEAQHKELEEKRRQFEDEKANWEAQQRILE

4.1.3. SAH régiók a paraspeckle struktúra felépítésében

A SAH régiók RNS-kötő fehérjékben betöltött feltételezhető szerepének részletesebb vizsgálatára az utóbbi évtizedben felfedezett ún. paraspeckle nevű sejtmagi képlet [119] felépítésében részt vevő fehérjéket választottuk ki. A paraspeckle ún. hosszú nem-kódoló RNS-molekulákat (long non-coding RNA, pl. NEAT1) tartalmaz integráns módon, és szerepe egyes mRNS molekulák sejtmagi retenciójának szabályozásában lehet, befolyásolva pl. sejtdifferenciálódási folyamatokat [120]. A paraspeckle struktúrát alkotó fehérjék az ún. DBHS („*Drosophila* behavior, human splicing") családba tartoznak, ezekre 2 szomszédos, RNS-kötő RRM domén, egy hosszú coiled-coil régió jellemző, a második RRM és a coiled-coil között egy NOPS („NonA/paraspeckle") nevezetű motívummal, valamint terminális rendezetlen szakaszokkal [121]. A prediktált SAH szakaszok a coiled-coil és a C-terminális rendezetlen szakasz közötti régióban helyezkednek el. Evolúciósan rokon fehérjéket elemezve megmutattuk, hogy a SAH szakaszok jelenléte a fehérjecsaládra jellemző, az ide tartozó fehérjék nagy részében jelen van. A SAH régió lehetséges szerepének további vizsgálatához szerkezeti modellt építettünk (V. közlemény), amihez a PSPC1/NONO fehérjék RRM doméneket tartalmazó homodimer szerkezetét vettük alapul [91], az azt leíró közleményben feltételezett multimerképződési elképzelést felhasználva. Eszerint az egyes dimerek a bennük megfigyelhető, rendhagyó geometriájú coiled-coil régiók meghosszabbított szakaszai segítségével alkotnak nagyobb komplexeket.

A dimerek közötti szakaszt a becsült SAH régió végéig egységesen coiled-coil szerkezetként modelleztük, igazodva a dimerben található geometriai jellemzőkhöz, azaz a 11 aminosavas ismétlődő motívumnak megfelelő jobbmenetes szuperhélixet feltételezve.

A modellépítéshez saját, PYTHON nyelven implementált kódot írtunk, mely Offer és Sessions 1995-ben publikált egyenleteit [122] – amelyek egyébként Francis Crick 1953-as közleményén [19] alapulnak – felhasználva képes szabályos coiled-coil molekulagerinc felépítésére. A modell elkészítése során a legnagyobb feladatot a feltételezett coiled-coil dimer szakasz megfelelő geometriájának felépítése, a 11 aminosavas ismétlődésnek megfelelõ hidrofób belsõ „varrat" aminosavainak egymáshoz történő pozicionálása jelentette. A coiled-coil részekre jellemzõ kölcsönhatások elemzéséhez a SOCKET [22] és a COILCHECK+ [123] eljárásokat használtuk. Megjegyzendő, hogy a koncepciónk nagyon hasonló az eljárásunk és a vele épített modellünk elkészülte után, de még a közleményünk beküldése elõtt publikált CCBUILDER webszerverben [23] implementált munkamenethez, bár az nem alkalmas az általunk használt



4.2. ábra. A $(PSPC1/NONO)_3$ hexamer szerkezeti modellje. Az RRM régiókat tartalmazó PSPC1/NONO dimerek a modellezés során felépített coiled coil régiók segítségével kapcsolódnak össze, amelyek C-terminális vége megfelel a becsült SAH szakaszoknak. A piros, a sárga és a zöld láncok a PSPC1, a világoskék, sötétkék és türkizkék láncok a NONO fehérjének felelnek meg. Az ábra a Chimera [106] programmal készült. *Az V. közleményünkben megjelent 2. ábra részlete (licensz: CC BY-NC)*

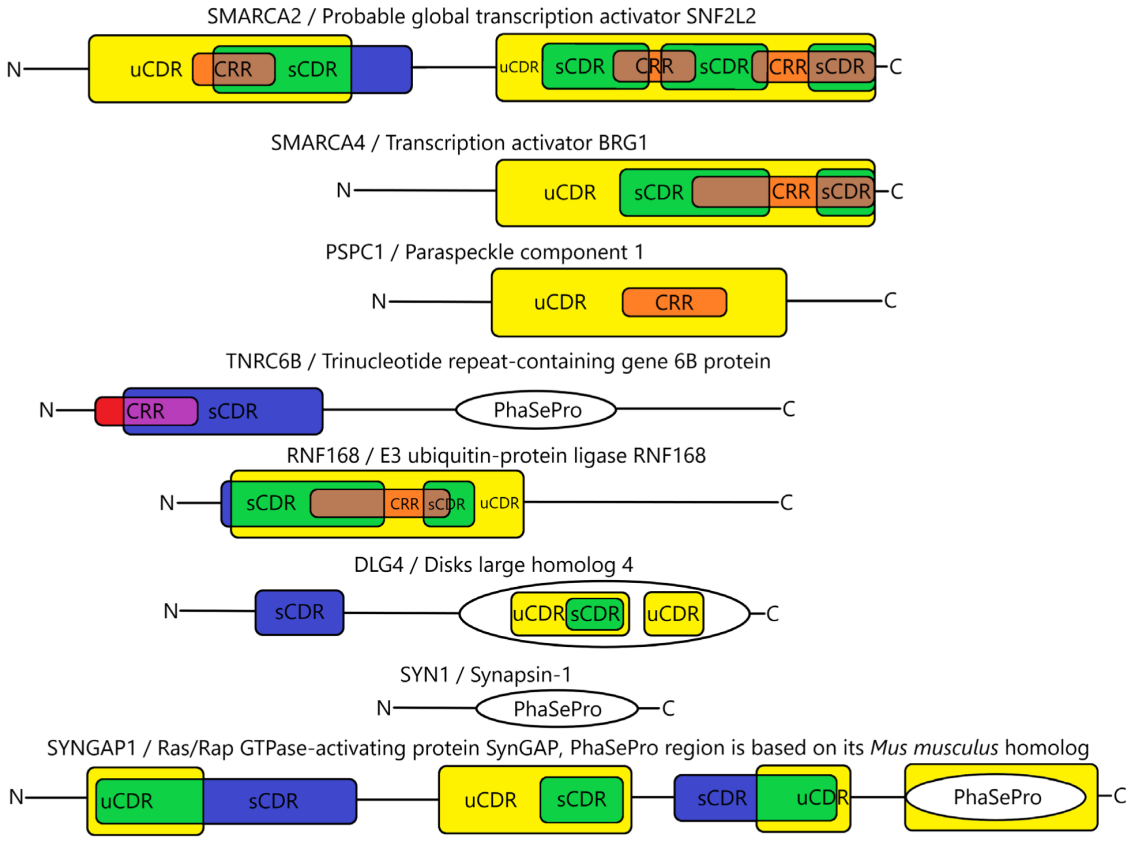
paraméterekkel rendelkező modell előállítására.

A felépített szerkezeti modell megmutatta, hogy a javasolt multimerizálódási módon elvben akárhány dimer összekapcsolódhat, nincs korlát a felépíthető szerkezet hosszára. Ugyanakkor fontos megjegyezni, hogy a modellezett részeken túl a molekulák N- és C-terminálisán is még hosszabb, funkcionálisan rendezetlen szakaszok találhatóak, melyek szerkezetét nem modelleztük. A modellünk egyik legfontosabb jellegzetessége, hogy a SAH szegmens hossza éppen akkora, hogy a PSPC1/NONO heterodimer által alkotott üreg mellett ér véget (4.2. ábra), így feltételezésünk szerint – az utána következő, funkcionálisan rendezetlen régióval együtt – a dimerek egymáshoz képest történő pontos térbeli pozicionálásában játszhat szerepet, amikor a multimer felépül a NEAT1 RNS-molekulához kapcsolódóan.

Munkánk publikálásával gyakorlatilag egy időben írták le a közeli rokon SFPQ fehérje dimerjeinek kristályszerkezeteit, melyekben az általunk becsült coiled-coil és SAH régiók egyetlen hosszú helikális szerkezetként jelennek meg, hasonlóan a modellünkhöz és megerősítve a SAH szakasz jelenlétét. Ugyanakkor a kristályokban a dimerek között észlelt kontaktusok mások és kevésbé kiterjedtek, mint a modellünkben – ez azonban lehet krisztallográfiai műtermék is [37].

4.1.4. Nagy töltéssűrűségű motívumok a fehérje-fázisszeparációban

A SAH motívumok RNS-kötő fehérjékben való gyakorisága mai ismereteink szerint nem magyarázható a motívum közvetlen RNS-kötésben való részvételével, hiszen erre gyakorlatilag nincsen bizonyítékunk. További magyarázat lehet valamilyen speciális térszerkezeti szerep, mint a paraspeckle esetében, ez azonban túl speciálisnak tűnik az általános gyakoriság magyarázatára. A fehérje-fázisszeparáció jelenségének felismerésével egyre több olyan rendszert írtak le, amelyek RNS:fehérje komplexeket tartalmaznak, ezért felmerült bennünk a lehetőség, hogy a SAH motívum esetleg ebben a folyamatban játszhat szerepet. Ennek vizsgálatára összevettük a SAH, valamint egyéb, nagy töltéssűrűségű motívumokat tartalmazó humán fehérjék listáját egyes fázisszeparációban részt vevő fehérjéket tartalmazó adatbázisok tartalmával (VI. közlemény). Nagy töltéssűrűségű motívumként a SAH szakaszok mellett az FT_CHARGE eljárással azonosítható, jól felsmerhető töltésváltakozási mintázatot mutató szegmenseket (charged residue repeat, CRR), valamint töltéssel rendelkező aminosavakat a vártnál nagyobb sűrűségben tartalmazó, összességében is nagy nettó töltésű (signed charge-dense region, sCDR) ill. utóbbival nem rendelkező (unsigned CDR, uCDR) tekintettünk. Ezek a definícók nem zárják ki egymást kölcsönösen, a CDR kategóriák egymással is átfednek és tipikusan tartalmazzák a CRR szakaszokat, melyeknek egy alesetét képezik a SAH régiók. Eredményeink egyrészt szoros statisztikai kapcsoltságot mutatnak egyes vizsgált motívumok és a fázisszeparációra való hajlandóság között, másrészt ez az asszociáció inkább negatív abban az értelemben, hogy elsősorban abból ered, hogy az ilyen régiók hiánya valószínűtlenné teszi, hogy a fehérje részt vegyen kondenzált fázis kialakításában. Az is megfigyelhető, hogy a fázisszeparáció iniciációjában kulcsszerepet játszó driver régiók csak ritkán esnek egybe az általunk vizsgált, nagy töltéssűrűségű szakaszokkal. Az eredmények egyik lehetséges értelmezése, hogy a nagy töltéssűrűségű régiók olyan szerkezeti-dinamikai tulajdonságokkal rendelkeznek, melyek a vizes és a kondenzált fázisban is funkcionálisak. Ez a meglehetősen óvatos ás általános feltételezés valószínűleg igaz a SAH szegmensekre, melyek stabil helikális szerkezete vizes közegben és a sűrű kondenzált fázisban is alapvetően megmarad, utóbbiban pl. egyéb, hasonlóan nagy töltéssűrűségű szakaszokkal körülvéve is. Megjegyzendő, hogy ezen eredményeink nem adnak továbbra sem egyértelmű választ a SAH motívumok RNS-kötő fehérjékben való gyakori előfordulására.



4.3. ábra. Néhány, a PhaSepDB alapján fázisszeparációban érintett fehérje az általunk vizsgált nagy töltéssűrűségű szakaszok (különböző színű téglalapok) és a PhasePro adatbázisban annotált LLPS driver régiók (fehér ellipszis) feltüntetésével. Jól látszik, hogy az egyes típusokba tartozó szakaszok jelentősen átfedhetnek, illetve hogy a driver fehérjékben sem feltétlenül szükséges a jelenlétük, egybeesésük az azonosított driver régiókkal erősen esetfüggő. *Az ábra a VI. közleményünkben jelent meg (licensz: CC BY-NC-ND)*

4.2. Funckionálisan rendezetlen és fibrilláris fehérjeszakaszok predikciója

4.2.1. Fibrilláris szerkezeti elemek és rendezetlenségi predikciók

A SAH szegmensek elemzése során korán észrevettük, hogy ezek a szakaszok gyakran átfednek olyan szegmensekkel, amelyeket az arra specializált eljárások coiled-coil vagy funkcionálisan rendezetlen szerkezetűnek ismernek fel. Ezt értelmeszerűen a motívummal foglalkozó más kutatók is megfigyelték [124]. Erre a jelenségre „keresztpredikció”-ként fogok a továbbiakban hivatkozni.

A rendezetlen szakaszokkal történő „keresztpredikciót” a nagy hidrofób oldalláncok ritkasága és a töltött aminosavak gyakorisága magyarázza, míg a coiled-coil régiók esetében

a felismerés oka az aminosavak szabályos ismétlődése, mely egyes jellegeiben – leginkább a töltéssel rendelkező aminosavak helyzetében – hasonlít a heptád egységekben megfigyelhető mintázathoz.

Kérdésként merült fel, hogy s SAH régiók felismerhetőek-e úgy is, mint rendezetlen és coiled-coil szakasznak egyaránt prediktált szakaszok, illetve hogy a coiled-coil és rendezetlen szakaszok közötti keresztpredikció általában véve mennyire gyakori. Utóbbi azért is fontos, mert a becslések szerint a funkcionálisan rendezetlen fehérjék számos, főleg eukarióta élőlény proteomjának igen nagy hányadát teszik ki, és amennyiben a keresztpredikció mértéke jelentős, akkor ezek a becslések pontosításra szorulhatnak.

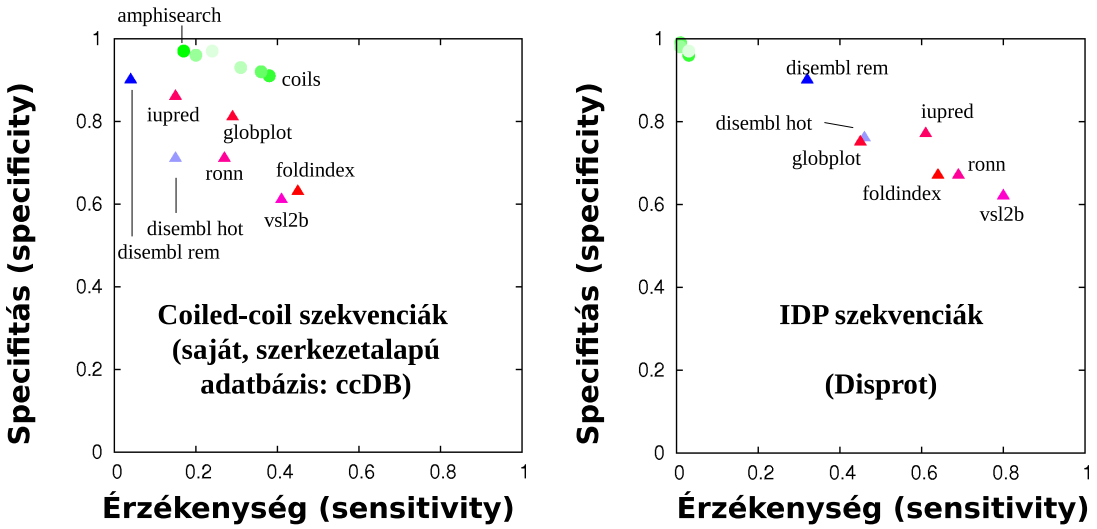
A kérdések megválaszolásához szükséges, kísérletileg igazolt szerkezeti információkat tartalmazó adatokat egyrészt a DisProt [125] adatbázisból, másrészt egy általunk a PDB adatbázis alapján a Socket eljárás [22] segítségével összeállított adatkészletből (ccDB) vettük.

Összesen hétféle rendezetlenség- és hatféle coiled-coil-predikciós eljárást vizsgáltunk (VII. közlemény). Mind a DisProt, mind a ccDB adatkészleten mindegyik eljárást értékeltük olyan módon, hogy az adatkészletre specifikus jellemzőket mennyire jól azonosítja (tehát a ccDB-n a rendezetlenséget becslő programokat is úgy teszteltük, mintha azok coiled-coil prediktorok lennének és viszont). Azt találtuk, hogy a coiled-coil felismerő eljárások lényegesen specifikusabbak mindkét halmazon, és a rendezetlen régiókra csak az azok felismerésére tervezett eljárások érzékenyek (4.4. ábra). Ugyanakkor a rendezetlenséget becslő algoritmusok nagy arányban ismerik fel a coiled-coil szakaszokat rendezetlenként.

Vizsgálatainkat elvégeztük a teljes SwissProt adatbázison is, ahol már nem volt lehetőség az egyes predikciók szabatosságának ellenőrzésére, a keresztpredikciók mértéke viszont nagyobb léptékben volt vizsgálható.

Nem meglepő módon az egyes predikciós eljárások között is jelentős eltérések lehetnek, és ez természetesen igaz adott coiled-coil-rendezetlenség prediktorok kombinációjára is. Az egyes eljárások teljesítményének minden aspektusát figyelembe véve a Coils [26]-IUPred [77] pár használatát javasoltuk olyan esetekre, amikor az átfedések minimalizálása a cél.

Az a hipotézis, hogy a SAH régiók azonosíthatóak lennének specifikus predikció nélkül, pusztán a mind coiled-coilnak, mind rendezetlennek becsült szakaszok alapján, nem igazolódott be (4.5. ábra). A mindkét prediktor által felismert szakaszok a specifikus SAH predikcióknál lényegesen több régiót és aminosavat érintettek. Ugyanakkor a hipotézis meg-

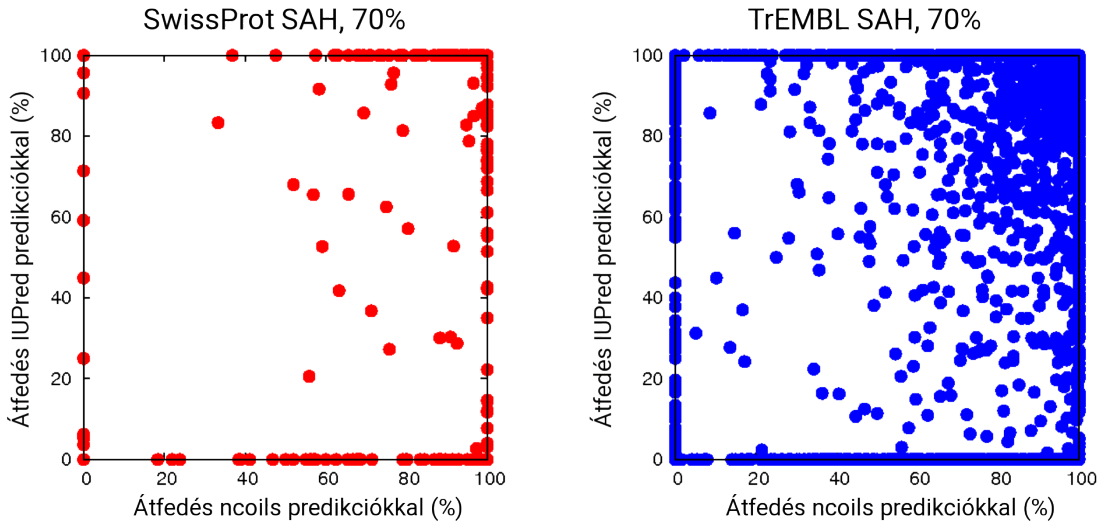


4.4. ábra. Coiled-coil és rendezetlenség-becslő algoritmusok érzékenység-specifitás diagramja az általunk létrehozott ccDB (balra) és a DisProt (jobbra) adatbázison vizsgálva. A coiled-coil predikciós eljárások közül csak a két leginkább különböző eredményt adót emeltem ki a bal panelen. *Az ábra a VII. közleményünkben lévő hasonló diagram módosított változata (licensz: Elsevier user license, a fordítás nem hivatalos, az Elsevier által nem jóváhagyott)*

alapozottságát mutatja, hogy a keresztpredikcióval érintett szegmensek aminosav-összetétele hasonlónak bizonyult a SAH szakaszokra jellemzőhöz.

Az alacsony komplexitású, de ismétlődő jellege miatt határozott szerkezetet felvevő szakaszokra jó példa még a kollagén tripla hélix[3]. Ezen szakaszok magas prolin- és glicintartalma miatt feltételezhető, hogy a rendezetlenséget becslő eljárások nagy része felismeri őket. Ezért elvégeztem néhány proteom szisztematikus vizsgálatát, melyben a coiled-coil, kollagén hélix és rendezetlen szakaszok predikciójának viszonyát elemeztem részletesen (VIII. közlemény). Az elemzés során többféle predikciós eljárást vizsgáltam, azonos típusú predikciók esetében azok konszenzusát is. Elsősorban azt a kérdést elemeztem, hogy az egyes proteomokban rendezetlennek becsült szakaszok mekkora hányada az, amelyik valójában inkább fibrilláris, jelen esetben kollagén tripla hélix vagy coiled-coil szerkezetet vesz fel. Érdekes módon az eredmény nagyban függ a vizsgált proteomoktól. Többsejtű állatokban (*Metazoa*) a rendezetlen aminosavak aránya a proteomban nem csak általában véve nagy, hanem igen változatos is, ugyanakkor ezen belül a nem kizárólag rendezetlennek becsült aminosavak aránya jó közelítéssel 5% körül mozog. Ezzel szemben pl. baktériumoknál a proteomban található funkcionálisan rendezetlen szakaszok aránya összességében alacsony, de ezen belül a coiled-coilnak is becsült

[3]A kollagén szerkezetek egy Simon Istvánnal és Dosztányi Zsuzsannával folytatott beszélgetés során merültek fel, ezért ezúton mondok köszönetet.



4.5. ábra. Becsült SAH szekvenciák átfedése NCOILS és IUPRED predikciókkal. Az adatok a SAH szegmenst tartalmazó szekvenciák 70%-os azonossági küszöb szerint rendundanciamentes halmazára vonatkoznak. Bal panel, piros pontok: SwissProt, jobb panel, kék pontok: TrEMBL szekvenciák. *Az ábra a II. közleményünkben lévő ábra módosított változata (jelen felhasználás az Elsevier által az eredeti szerzőknek biztosított jogon alapul)*

aminosavak aránya fajonként nagy változatosságot mutat. A legnagyobb szórást a virális proteomok mutatják, ennek feltehetően az az oka, hogy ezek mérete igen kicsi, így az egyedi fehérjék specifikus jellegzetességei nagyobb súllyal esnek latba a proteomszintű összesítésben.

4.3. *De novo* fehérjeszekvenciák szerkezeti preferenciáinak predikciója

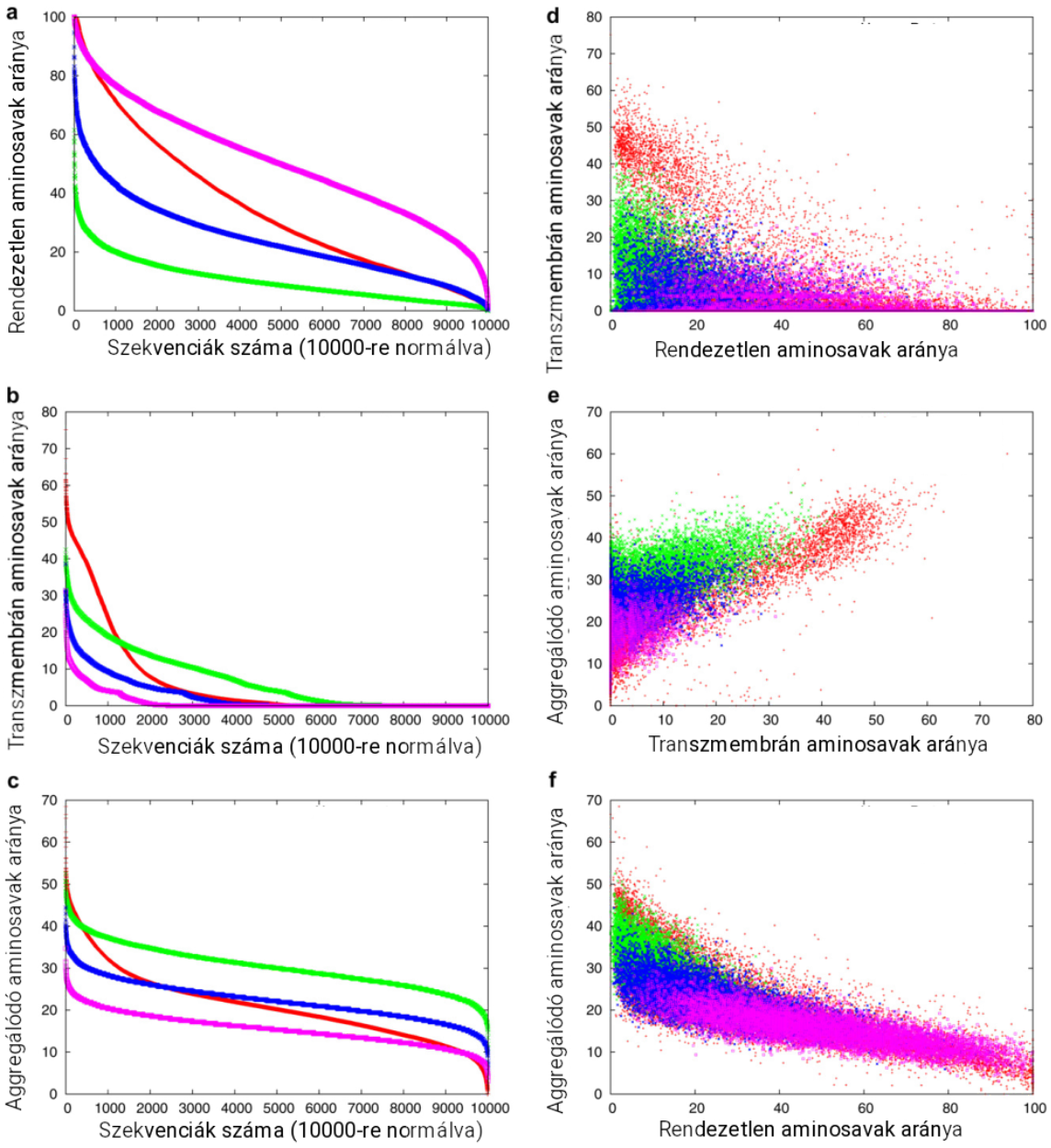
Predikciók segítségével vizsgálni kívántuk azt az evolúciós forgatókönyvet, amelynek során korábban nem kódoló szakaszokból fehérjekódoló gének keletkeznek (IX. közlemény). Kíváncsiak voltunk az újonnan létrejövő fehérjék szerkezeti preferenciáira, azon belül is arra, hogy az aggregációs hajlamuk valóban magasabb-e, mint a hosszú evolúciós szelekción átesett, mai fehérjeszekvenciáké. Igen fontos leszögezni, hogy ebben a kontextusban a predikciókat nem tekintettük az egyedi szekvenciák szintjén megbízhatónak, csupán az észlelt trendeket kívántuk elemezni és azokat is elsősorban egymással való összehasonlításban értelmezni. Ennek oka, hogy a predikciós eljárások fejlesztése és tesztelése alapvetően a természetben előforduló, határozottan nem random szekvenciákon történik, és így nem indokolt az az elvárás, hogy egy véletlen szekvencián is hasonló eredményességgel teljesítsenek általános-

ságban. Előrebocsátom, hogy ez a szempont még sokkal hangsúlyosabb lesz a következő, korai genetikai kódokat elemző vizsgálatunkban, de az említett okokból már a most leírt esetekben sem elemzünk egyedi szakaszokat az egyes szekvenciákban, és elsősorban az egyes jellemzők által lefedett szerkezeti paramétertér vizsgálatára koncentrálunk. Olyan eljárást választottunk, mely első közelítésben modellezi a tényleges biológiai folyamatot: véletlenszerű DNS-szakaszokat generáltunk és azokat lefordítottuk fehérjészekvenciákra. A szakaszok 480 nukleotid hosszúak voltak, és nem tartalmaztak STOP kodont az 1-es nukleotiddal kezdődő leolvasási keretben. Összesen 90 000 szekvenciát állítottunk elő úgy, hogy 10 és 90% között változtattuk a szakaszok GC-tartalmát, és minden GC-tartalomhoz 10 000 szekvenciát generáltunk. A nukleinsav-szekvenciák hossza 480 nukelotid volt, ezek lefordításával 160 aminosav hosszúságú fehérjeszakaszokat kaptunk, amely jól közelíti az átlagos domén mint feltekeredési-szerkezeti egység hosszát. A kapott fehérjeszekvenciákon többféle predikciós eljárást futtattunk le a rendezetlen, aggregációra hajlamosító és transzmembrán (TM) doméneket képző régiók azonosítására. A TM domének elemzésbe történő bevonásának indoka az volt, hogy mint hidrofób jellegű szakaszok, adott esetben lehetnek olyan tulajdonságaik, amelyek közel állnak az aggregációs hajlammal rendelkező szakaszokéhoz.

Eredményeik alapján az egyes szerkezeti jellemzők jelentős mértékben függnek a mögöttes kódoló DNS-szakasz GC-tartalmától, és a genomokban szokásosnak tekinthető, 40-60% közötti GC-tartalom mellett a rendezetlenség a domináns tulajdonság, az aggregációs hajlam várhatóan nem jelenik meg jelentős leküzdendő problémaként. A predikciókat elvégeztük három létező, a munka elvégzésekor megbízhatóan *de novo* fehérjeként azonosított emberi szekvencián is, amelyek esetében az átlagos GC-tartalom és a szerkezeti jellemzők is jól illeszkedtek a véletlenszerű szekvenciák esetében észleltekhez.

Az evolúciós aspektus részletesebb vizsgálatát kiterjesztettük egyes feltételezett korai genetikai kódokra is (X. közlemény). Ezek az evolúció korai szakaszában lehetettek relevánsak, és különböző elméleti megfontolások alapján tettek rájuk javaslatot a területtel foglalkozó kutatók. Ismereteink szerint azonban az ezek által potenciálisan kódolt fehérjék szerkezeti preferenciáit korábban szisztematikusan nem vizsgálták. A mai fehérjékre kapott adataink tükrében érdemesnek láttuk a korai kódok esetére is kiterjeszteni az elemzésünket, melybe többféle feltételezett korai kódot vontunk be [126, 127, 128][4]. Ezek mellett vizsgáltuk a korai fehérjék kontextusában felmerült, de kódtáblához nem rendelt ún. GADV aminosavkészletet

[4]A konkrétan használt kódtáblákat a X. közleményünk kiegészítő anyaga tartalmazza (Figure S1, ahol az EARLY4 és az EARLY10 táblák sajnos hibásan, felcserélve szerepelnek)

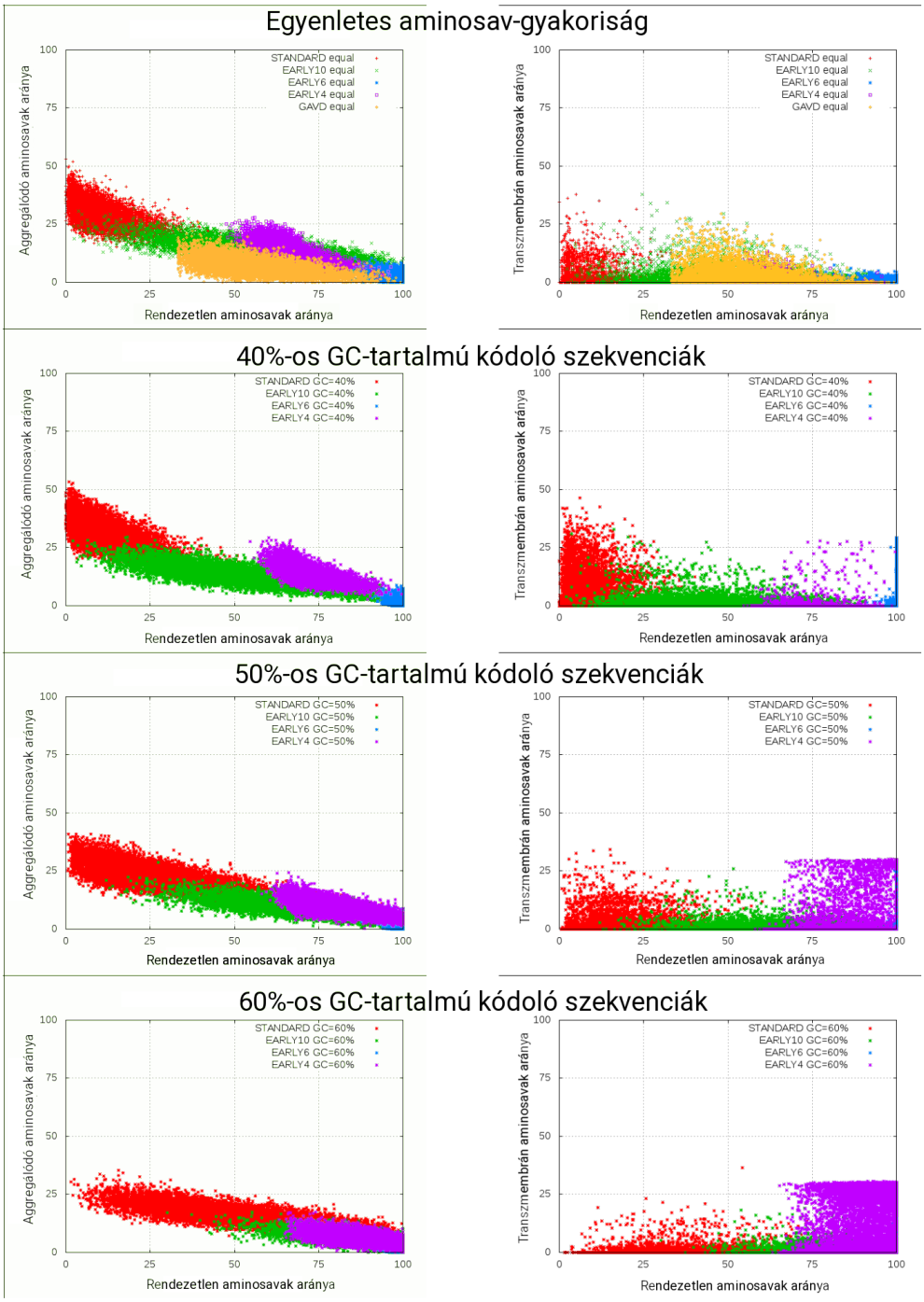


4.6. ábra. *De novo* fehérjeszekvenciák rendezetlenségi, aggregációs és transzmembrán hélixek kialakítására vonatkozó, becsült szerkezeti preferenciái az őket kódoló DNS-szakasz GC-tartalmának függvényében. Pirossal a humán proteom fehérjéit, zölddel, kékkel és lilával rendre a 40, 50 és 60%-os GC-tartalmú random DNS-szakaszokról lefordított, 160 aminosav hosszú fehérjeszekvenciákat ábrázoltuk. A „rendezetlen/transzmembrán/aggregálódó aminosavak aránya" az ilyen szerkezeti preferenciákkal rendelkező szakaszokba eső aminosavak százalékos arányát jelenti az összes vizsgált aminosavhoz képest. A bal panel eloszlásgörbéit olyan módon skáláztuk össze, hogy a humán proteom fehérjéinek eloszlása összevethető legyen az egyes GC-tartalmakhoz tartozó 10000 hipotetikus random fehérje szekvencia eloszlásával. *Az ábra eredeti változata a IX. közleményünkben jelent meg. (licensz: Elsevier user license, a fordítás nem hivatalos, az Elsevier által nem jóváhagyott)*

is [126, 129] Megjegyzendő, hogy a vizsgált kódok nem alkotnak egymásból egyértelműen származtatható sorozatot, tehát igencsak valószínűtlen, hogy a mai kód felé vezető evolúciós folyamatban mindhárom relevánsnak tekinthető legyen.

A módszertan igen hasonló volt a mai fehérjék esetében alkalmazotthoz, a véletlenszerű DNS-szekvenciákat a korai kódok segítségével fordítottuk le és a kapott fehérjeszekvenciákat predikciók segítségével elemeztük. Első és talán leginkább meglepő eredményünk az volt, hogy az igen kevés, 4-6 aminosavat tartalmazó korai kódok esetében a predikciók többször is egészen eltérő eredményekre vezettek ugyanazon szekvenciák esetében. Ez arra hívja fel a figyelmünket, hogy a mai, 20-féle aminoavból álló, nem random szekvenciák elemzésére kifejlesztett módszerek nem feltétlenül alkalmazhatóak ezektől lényegesen eltérő összetételű szekvenciák vizsgálatára.

A korai kódok kapcsán feltett fő kérdésünkre a predikciók által sugallt válasz az, hogy a javasolt korai kódok egyike sem képes a mai fehérjékre jellemző szerkezeti változatosság kialakítására (4.7. ábra). Ehhez legközelebb a 10 aminosavas kód jut, amely a többivel ellentétben már közelíti valamelyest a modern, 20-féle aminosavból felépülő fehérjék szerkezeti terét. Ezen a ponton azonban óvatosnak kell lennünk és meg kell vizsgálnunk azt a lehetőséget, hogy a kiinduláskor adottnak vett mögöttes feltevéseink közül nem mindegyik teljesül. Az első ilyen eset, hogy a korai kódok legalább egyikének létezését az őket leíró szerzők szerinti formában nem vonjuk kétségbe, ugyanakkor megkérdőjelezzük a predikciós eljárások megbízhatóságát azon az alapon, hogy a korai szekvenciák egyáltalán nem feltétlenül a ma szokásosnak tekintett fizikokémiai körülmények között léteztek, és emiatt a predikciók nem képesek releváns eredményeket adni, elsősorban a 4 és 6 aminosavas korai kódok esetében. A másik eshetőség, hogy a predikciók alapján valószínűsített szerkezeti tulajdonságokat valamennyire reálisnak tekintjük, ekkor viszont azt a következtetést kell levonnunk, hogy az irodalomban leírt, igen leegyszerűsített korai genetikai kódok a feltételezett formájukban valószínűleg nem léteztek. (Ha sem a predikciókat, sem a kódokat nem fogadjuk el, úgy az egész elemzés nyilvánvalóan haszontalan – bár lehet emellett is érvelni, ezzel az álláspontal meddősége miatt nem foglalkozom.) Én az utóbbi forgatókönyv mellett foglalok állást, természetesen elfogadva azt, hogy a predikciók megbízhatósága még a trendek szintjén is jóval alacsonyabb, mint a mai fehérjék esetében. A szerkezeti tulajdonságok változatosságának alacsony mivolta megítélésem szerint így is igen komoly érvként esik latba. Mindezek alapján a kód evolúciójának leírására leginkább Carl Woese hipotézisét tartom relevánsnak,



4.7. ábra. Korai genetikai kódok által meghatározott hipotetikus fehérjeszekvenciák rendezetlenségi, aggregációs és transzmembrán hélixek kialakítására vonatkozó, becsült szerkezeti preferenciái. A standard kód (piros pontok) mellett használt kódok: EARLY10 [126] (Ala, Asp, Glu, Gly, Ile, Leu, Ser, Pro, Thr, Val aminosavak; zöld pontok), EARLY6 [127] (Ala, Asp, Glu, Gly, Ser, Val; kék pontok), EARLY4 [128] (Asp, Arg, Leu, Ser; lila pontok). Az aminosavakat azonos valószínűségére vonatkozó legfelső két panelen ezek mellett szerepel még a GADV (Ala, Asp, Gly, Val; sárga pontok) aminosavkészlet is [126, 129], az alsó 3 sorban pedig a 40, 50 és 60%-os GC-tartalmú véletlenszerű DNS-szakaszokról lefordított szekvenciák adatai láthatóak. Ezen az ábrán csak a 160 aminosav hosszú szekvenciákra vonatkozó adatokat mutatom be. *Az ábra az X. közleményünkben (annak kiegészítő anyagában) megjelent diagramok alapján készült, felhasználásuk a Springer Nature engedélyével történt*

mely szerint a kód evolúciójának korai időszakában is volt lehetőség sokféle kémiai jellegű aminosav beépülésére, és az evolúció során az egyes kodonok aminosavakhoz történő hozzárendelése vált egyre inkább egyértelművé és kizárólagossá [130]. Ez a lehetőség megengedi a nagyobb változatosságot a fehérjeszekvenciák szintjén, biztosít valamekkora örökölhetőséget, és teret ad az evolúció számára a kód finomhangolására is. Ez a forgatókönyv az aminosavak hozzáférhetőségét nem veszi figyelembe sem abiotikus, sem bioszintetikus utakon. A komplex bioszintetikus utak kialakulása a redukált aminosavkészletből felépülő fehérjék mint potenciális enzimek esetében egyébként sem egyszerűen magyarázható, pedig ez az egyre bővülő kódok esetében szükséges lehetne. Így megítélésem szerint ezen aspektus sem billenti a mérleg nyelvét a 4 és a 6 aminosavas feltételezett kódtáblák relevanciájának elfogadása felé.

4.4. Dinamikus fehérjeszerkezeti sokaságok előállítására és elemzésére alkalmas módszerek implementálása és fejlesztése

4.4.1. A MUMO eljárás implementálása a Gromacs programcsomagba

A dinamikus fehérjeszerkezeti sokaságok előállítására alkalmas MUMO (Minimal under-restraining minimal over-restraining) koncepciót 2007-ben publikálta Michele Vendruscolo kutatócsoportja [64]. Az közlemény a molekuladinamikai szimulációk során alkalmazott kísérleti eredetű kényszerfeltételek optimális felhasználására tesz javaslatot, mellyel elkerülhető mind a túl- mind az alulillesztés. A számítási módszer alapja, hogy a vizsgálandó milekula több replikáját párhuzamosan szimulálva, az egyidőben jelenlévő konformereket elemezve számítsuk ki az egyes paramétereket a sokaságra vonatkozó átlagként, majd ebből kiindulva számítsuk ki a kényszerfeltételekhez tartozó energiatagot, valamint ebből származtatva az egyes replikák egyes atomjaira ható erőket. A MUMO javaslat lényegi része, hogy a NOE-alapú távolság jellegű kényszerfeltételeket ne átlagoljuk két replikánál nagyobb alsokaságon, míg az S^2 paraméterek kiszámításához alkalmazhatjuk az összes, a számításban részt vevő replikát. A NOE-alapú távolságok esetében a páronkénti átlagolás úgy valósul meg, hogy minden replikát a két „szomszédjával” külön-külön páronként veszünk figyelembe.

Az eljárás implementálását a Cambridge-i Egytemen töltött ösztöndíjam alatt, Christopher Dobson és Michele Vendruscolo csoportjában, Várnai Péter közvetlen iránymutatása mellett kezdtem meg, célom az volt, hogy a módszer a kutatásaimhoz korábban is használt, szabad forráskódú GROMACS programcsomagban is elérhető legyen. A GROMACS akkori, 3.3.1-es változatában elérhető volt már a multireplika szimuláció lehetősége, viszont S^2 kényszerfeltételeket egyáltalán nem, NOE távolságokat pedig csak a teljes sokaságra lehetett alkalmazni. Az S^2 kényszerfeltételek implementálásakor a Robert Best és Michele Vendruscolo által publikált megoldást [59] vettem alapul.

A GROMACS program C nyelvű kódjában általam végrehajtott implementáció számos technikai részlete közül itt csupán a legfontosabbat szeretném kiemelni, melyet a fejlesztés során az eredeti megoldástól függetlenül építettem be. Eszerint az S^2 paraméterek megbízható számításához az egyes replikák szuperpozíciója[5] (egészen pontosan legalább a szuperpozíció forgatási komponensének végrehajtása) szükséges minden egyes lépés során. Ez természetesen „virtuálisan" történik, azaz a tényleges szimulált replikát ez nem mozdítja el[6]. A szuperpozíció után megtörténik az S^2 értékek kiszámítása, a kényszerfeltételként beállított (kísérleti) értékekkel való összevetése, majd ennek alapján az erők kiszámítása. Az utolsó lépés az erővektorok transzformálása a ténylegesen szimulált molekula orientációjának megfelelően, azaz az erőkre minden replika esetében a szuperpozíció során számolt forgatás inverzét kell végrehajtani. Ez az eljárás biztosítja, hogy az S^2 értékek a geometria alapú számítás során[7] elérhető lehető legjobbak legyenek, azok ne csökkenjenek le a molekulák szimuláció során történő elfordulásai miatt. A szuperpozíciót végző lépést olyan módon fejlesztettük tovább, hogy minden egyes S^2 értékre külön megadhatóvá tettük azt az atomcsoportot, amelynek felhasználásával a fedésbe hozás megtörténik a szerkezetből való számítás során. Ennek az a jelentősége, hogy a módszer akkor is használható, ha a teljes molekula nem közelíthető egyetlen merev testként viselkedő entitásként. Erre példa lehet egy olyan kétdoménes fehérje, melyben az egyes domének nagymértékben szabadon mozognak, adott esetben különböző

[5]Törekszem a fedésbe hozás/szuperpozíció és a szerkezetillesztés fogalmak szabatos használatára [131] nyomán: előbbi pusztán geometriai művelet az egymásnak megfelelő térbeli pontok ismeretében, utóbbi a legjobb megfelelés megtalálását is jelenti. Az itt tárgyalt esetben az azonos konstitúciójú molekulák atomjai közötti megfelelés triviális

[6]Hasonló lépést megvalósító kódrészlet elérhető volt a GROMACS akkori verzióiban az RDC-alapú kényszerfeltételek (orientation restraints) számításához

[7]Az S^2 értékeket hagyományosan NMR relaxációs adatok alapján illesztik, és pontos geometriai jelentésükhöz szigorúan véve szükség lehet a mögöttes mozgási modellre, mely a gyakorlatban nem ismert. A dolgozatban tárgyalt módszer a molekuladinamikai számításokból kapott atomi koordinátákból kiinduló, ezen a területen bevettnek tekinthető geometriai értelmezést követi.

rotációs korrelációs idővel (τ_c) jellemzhetőek. Ebben az esetben az S^2 értékeket az egyes domének esetében egymástól függetlenül kell értelmeznünk, a releváns globuláris molekularészleteket külön kezelve.

Az S^2 kényszerfeltételek mellett a NOE-alapú távolságok replikapáronkénti számítását is implementáltam a GROMACS 3.3.1, majd később a 4.5.5 verziójába. Utóbbi verzióba mindezeken felül beépítettem az AMD (Accelerated Molecular Dynamics) séma [132] torziós energiatagra alkalmazható változatát is a geometriai mintavételezés hatékonyságának növelésére. A módosított kódot tartalmazó C nyelvű állományokat, illetve az azok telepítését és használatát segítő PERL szkripteket elérhetővé tettem a honlapomon. Az egyes implementációk leírását a XI, XII és XIII közleményeink tartalmazzák.

4.4.2. A DIPEND eljárás

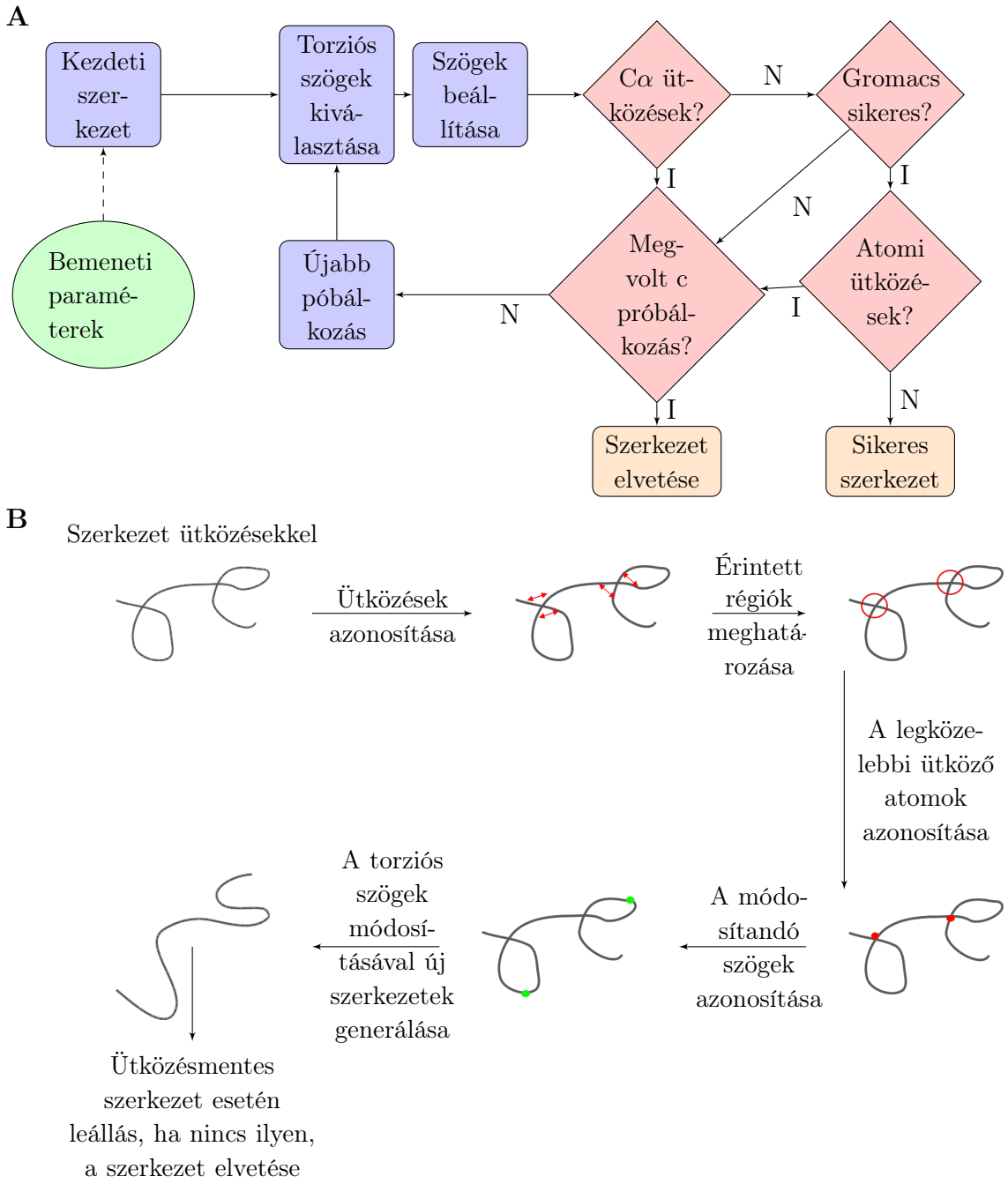
A dinamikus szerkezeti sokaságok előállítása a kényszerfeltételekkel kombinált molekuladinamikai számítások segítségével leginkább határozott szerkezettel rendelkező, globuláris fehérjékre alkalmazható. Funkcionálisan rendezetlen fehérjék és szakaszok esetében a lényegesen nagyobb konformációs szabadság miatt más megközelítésekre van szükség. Ilyen esetekben tipikusan nagyszámú konformer generálása történik, majd ezekből egy szelekciós eljárás segítségével választjuk ki a kísérleti adatoknak megfelelő alsokaságot. A feladat nehézsége a megfelelő kiindulási sokaság elkészítésében rejlik, ennek ugyanis értelemszerűen tartalmaznia kell az – előzetesen nem ismert – ténylegesen előforduló konformereket vagy hozzájuk nagyon hasonló szerkezeteket, miközben jó, ha ezek mellett nem tartalmaz túl sok csekély relevanciájú szerkezetet. A gyakorlatban tehát előre ismernünk kellene a sokaság közelítő jellegzetességeit, hogy hatékonyan tudjunk annak alapján még jobb sokaságot generálni. A hagyományosan elérhető megoldások emiatt nagyon nagy, néhány tízezres kiindulási sokaságokat alkalmaznak a konformációs tér megfelelő lefedése érdekében. A korábban az irodalomban leírt megközelítések másik hátránya, hogy nem feltétlenül érhetőek el szabadon és/vagy futtathatóak firss rendszereken. Ezért megkezdtük egy olyan munkafolyamat kidolgozását, amely képes rendezetlen fehérjék szerkezeti sokaságainak előállítására, és ehhez elsősorban karbantartott, szabad forráskódú, de legalábbis ingyen hozzáférhető további alkalmazásokat használ. Szempont volt még, hogy a generált sokaság olyan megfontolásokon alapuljon, amelyek lehetővé teszik a hatékony mintavételezést, azaz hogy a tényleges konformációs eloszlást minél jobban közelítsük. Emellett a felhasználó által való konfigurálhatóságot is fontosnak tartottuk.

A kifejlesztett DIPEND (DIsordered Protein Ensembles from Neighbor-dependent Distributions) eljárás (XVII. közlemény) központi eleme, hogy a statisztikailag legvalószínűbb gerinckonformációkat mintavételezi, amelyhez a Dunbrack csoport által kiszámolt és elérhetővé tett szomszédságfüggő Ramachandran-eloszlásokat [133] használja fel. Ennek alapján egy adott aminosav lokális szerkezeti preferenciáit érvényesíti. A program emellett képes a felhasználó által megadott lokális szerkezeti preferenciákat is figyelembe venni, pl. amennyiben a rendelkezésre álló kémiai eltolódások vagy előzetes ismeretek indokolttá teszik valamilyen szerkezeti elem nagyobb valószínűségű mintavételezését egy adott régióban.

A hosszú rendezetlen szakaszok modellezése óhatatlanul eredményez olyan konformereket, melyekben sztérikus ütközések fordulnak elő, és így természetesen nem használhatóak fel további elemzésekben. Ugyanakkor nem feltétlenül az ilyen konformációk teljes elvetése lehet az egyetlen megoldás, hiszen számos olyan szerkezet is készül, amely csak kevés ütközést tartalmaz, és ezek sokszor egy viszonylag egyszerű módosítással kiküszöbölhetőek lehetnek. Ennek a problémának a megoldására implementáltunk egy ún. „kicsomózó" eljárást, mely azokon a szerkezeteken fut le, amelyek a megadottnál (tipikusan 6-8) kevesebb ütközést tartalmaznak. Az eljárás minden ütközésre azonosítja az érintett szakasz középpontját, amely pl. az n. és $n + 30$. aminosavak atomjainak ütközése esetében az $n + 15$. aminosav. Az így azonosított aminosavak ϕ és ψ gerinc torziós szögeit szisztematikusan ± 15 fokban változtatva megvizsgálja, hogy ilyen módon megszűnnek-e az ütközések. Így az eredetileg generált szerkezet minimális perturbációjával, egy új szerkezet generálásánál kevesebb idő alatt lehetséges elfogadható szerkezetet kapni.

A jelenleg elérhető DIPEND megvalósítás PYTHON nyelven íródott, külső programként támaszkodik a CHIMERAX [134], GROMACS [89] és SCWRL4 [135] programokra. Ezeken kívül egy saját C nyelvű állományt tartalmaz, mely az atomi ütközések gyors ellenőrzésére alkalmas. A DIPEND eljárást a szintén saját CONSENSX$^+$ módszerrel kombinálva sikeresen alkalmaztuk egy funkcionálisan rendezetlen és egy SAH szakasz jellemzésére is, utóbbit jelen fejezet végén egy külön alfejezetben mutatom be.

Bár a működőképességét sikeresen igazoltuk, a DIPEND eljárás jelenlegi legnagyobb hátránya viszonylagos lassúsága, elsősorban a ChimeraX programtól való függősége miatt, többek között ennek kiküszöbölése is cél a jelenleg is folyamatban lévő továbbfejlesztése kapcsán.



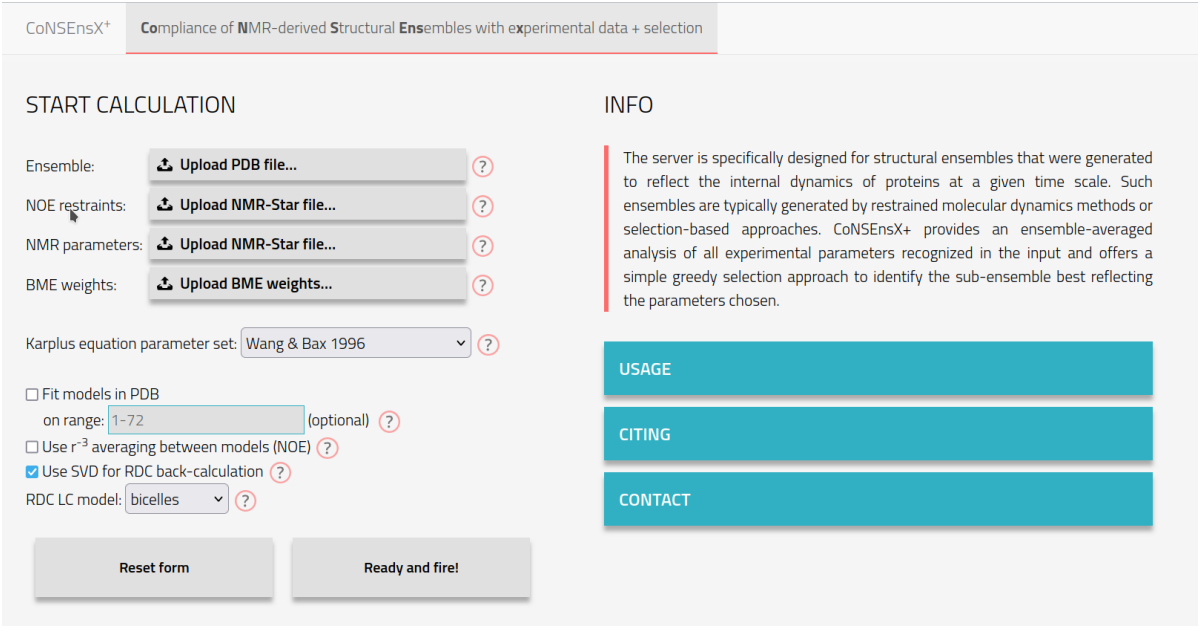
4.8. ábra. (**A**) A DIPEND eljárás áttekintő folyamatábrája; (**B**) a „kicsomózó” lépés részletei.
Az ábra eredeti, angol nyelvű változata a XVII. közleményünkben jelent meg (licensz: CC BY)

4.4.3. A CoNSEnsX webszerver

A sokaság-alapú modellek esetében az ilyen jellegű megfelelések vizsgálata – a hagyományos szerkezeti értékeléstől, validációtól eltérően – nem konformerenként, hanem a sokaság egészét figyelembe véve történik. Ennek alapja az a megfontolás, hogy nem feltételezhető, hogy a dinamikusan egymásba alakuló konformációk között akár egyetlen olyan is akadjon, amely egyszerre, egy időben megfelel az összes mért paraméternek [61] (lásd még 1.3.1. alfejezet). Emiatt a számítások legfontosabb jellemzője, hogy az egyes szerkezeti modellekre becsült paramétereket a sokaságra vetítve értelmezzük – pl. megfelelő módon átlagoljuk –, és a sokaságra jellemző értéket hasonlítjuk össze a kísérleti adatokkal. A sokaság szintjén elvárt megfelelés [68], valamint a dinamikai viselkedést tükröző sokaságok nagy változatossága miatt szükséges a dinamikus szerkezeti sokaságok esetére speciálisan alkalmas módszerek fejlesztése. Ezen megfontolás alapján vágtunk bele a CoNSEnsX (**Co**mpliance of **N**MR-derived **S**tructural **Ens**embles with e**X**perimental data) webszerver kifejlesztésébe (XIV. közlemény), amelynél szempont volt a minél könnyebb kezelhetőség, azaz hogy a szerver képes legyen a népszerű webes szerkezeti biológiai adatbázisokból letölthető fájlformátumok kezelésére, és az eredmények átlátható, ugyanakkor nem túlzóan leegyszerűsített megjelenítésére. Az eljárás jelenlegi verziója CoNSEnsX$^+$ néven az eredeti megoldásunk alapoktól újraírt és továbbfejlesztett változata (XV. közlemény) , mely elérhető webszerverként a consensx.itk.ppke.hu címen, illetve a forráskódja a GitHub felületen is hozzáférhető.

A szerkezetek értékeléséhez a CoNSEnsX$^+$ szervernek kötelezően megadandó a a szerkezeti sokaságot tartalmazó, PDB formátumú fájl. Az NMR paramétereket tartalmazó, a BMRB [74] adatbázis által használt NMR-Star formátumú [102] fálj, valamint a távolság jellegű kényszerfeltételeket tartalmazó NMR-Star formátumú állomány (ez a PDB adatbázisból letölthető „V2 NMR Restraints” jelzésű fájl) közül legalább az egyik megadása szintén elvárt.

A szerver az NMR-spektroszkópiai mérésekből származó adatokat – melyek lehetnek kémiai eltolódások, skaláris és maradvány dipoláris csatolások (residual dipolar couplings, RDCs), valamint rendparaméterek – összeveti a feltöltött szerkezeti sokaságból visszaszámolt és átlagolt értékekkel. A szerver jelenlegi verziója képes gerinc és oldallánc rendparaméterek értelmezésére is, és a maradvány dipoláris csatolások esetében több, különböző ún. orientált közegben felvett mérés paramétereinek egymástól független kiértékelésére. A szerver min-



4.9. ábra. A CoNSEnsX$^+$ webszerver nyitólapja

den paraméter megfelelésére legalább kétféle mérőszámot számol, korrelációt, RMSD-t és az RDC-k esetében ún. Q-faktort. A webszerver a kémiai eltolódások becslésére a SHIFTX [136], a maradvány dipoláris csatolások számítására pedig a PALES [137] eljárásokat alkalmazza.

A NOE-alapú kényszerfeltételek értékeléséhez a módszer felhasználja az általunk korábban kifejlesztett PRIDE-NMR eljárást is. Ennek lényege, hogy az amid NH, Hα és Hβ atomokra közötti kapcsolatokra vonatkozó, kísérleteileg meghatározott NOE csúcsok eloszlását összeveti az adott szerkezet 3D koordinátáiból számolt, ugyanazen atomok között definiált, térben közeli atompárok eloszlásával. Az eloszlásokat az eredeti PRIDE módszerben [138] alkalmazotthoz koncepcionálisan hasonló, de a lényegesen kevesebb adat miatt egyszerűbb módon kezeli, egyszerűen a NOE csúcsok / közeli atompárok számát veszi a vizsgált szekvenciális távolságok függvényében (XVI. közlemény).

A CoNSEnsX$^+$ legújabb változata – elsősorban a rendezetlen régiók vizsgálatának elősegítésére – másodlagos kémiai eltolódásokat is számol, szekvenciafüggő átlagos eltolódásokat alkalmazva [139].

A szerver paraméterenkénti bontásban háromféle diagramot jelenít meg: a kísérleti és a szerkezetből visszaszámolt adatokat a szekvencia mentén és egymás függvényében is ábrázolja, illetve a sokaság egyes modelljeire külön-külön kiszámolt korrelációs értékeket azok nagysága szerint is bemutatja. Ez utóbbi diagram segíti annak megítélését is, hogy a sokaságra számolt átlagos paraméterek jobban megfelelnek-e a kísérleteknek, mint az egyes modellekre

kapott értékek – a tapasztalat szerint általában ez a helyzet. Szándékosan nem törekedtünk arra, hogy a kapott eredményeket egyetlen mérőszámba sűrítsük. Ennek torzítástól mentes kiszámítása ugyanis nagyon nehéz lehet, hiszen az egyes fehérjék esetében nagyon különböző lehet rendelkezésre álló mért paraméterek száma és jellege, valamint ezek megbízhatósága is, emiatt az egyes esetekben kapott egységesített mérőszám nem lenne igazán jól összehasonlítható. Mindemellett a sokaság-alapú modellek elemzésekor kiemelt szempont, hogy mely időskálájú belső mozgásokat hivatottak tükrözni, ezért előfordulhat, hogy a kutató által vizsgált bizonyos jelenségek megértéséhez egy olyan modell is alkalmas lehet, amely csak néhány paraméternek felel meg elfogadható mértékben. Ennek esetenkénti mérlegelése a felhasználó feladata.

A jelenlegi változat egyik legfontosabb hozzáadott funkciója egy viszonylag egyszerû „mohó" algoritmus integrálása (XV. közlemény), mely a bemeneti sokaságból képes egy kisebb, a felhasználó által megadott paramétereknek jobban megfelelő alsokaság szelekciójára. Ez egyrészt – pl. a DIPEND eljárással előállított bemeneti sokaság esetén – lehetővé teszi egyrészt a kísérleti adatokat tükröző sokaságok szelekcióval történő előállítását, másrészt segíthet a túlillesztés jelenségének vizsgálatában. A túlillesztés ebben a kontextusban azt az esetet jelenti, amikor a sokaság mérete lényegesen nagyobb, mint amennyi feltétlenül szükséges lehet a paramétereknek való megfelelés eléréséhez [64]. Ilyenkor fennállhat a veszélye, hogy a sokaság által lefedett konformációs tér nem reprezentálja megfelelően a valós mozgásokat, másrészt pedig felesleges reprezentációs és elemzési terhet is jelenthet a szükségesnél nagyobb sokaságok kezelése. A szelekciós eljárás képes nagymértékben csökkenteni a konformerek számát a paramétereknek való megfelelés megtartásával, sőt javításával. A szelekcióban figyelmebe veendő paraméterek és azok egymáshoz viszonyított súlya a felhasználói felületen beállítható, csakúgy, mint a kiértékeléshez használt megfelelési mérőszám (korreláció, RMSD, Q-faktor). Az algoritmus az első lépésben a kiválasztott paramétereknek önmagában legjobban megfelelõ (a csak sokaságokon értelmezhető rendparaméterek esetében két) konformert választja ki, majd minden továbbiban a megfelelést legjobban javító szerkezet hozzáadásával bővíti a kiválasztott sokaságot. Lehetőség van arra, hogy megengedjük, hogy adott számú lépésben romoljon a megfelelés annak reményében, hogy később további növekedést kaphassunk. Az algoritmus működéséből fakadóan nem tudja azt garantálni, hogy a legjobban megfelelő alsokaságot választja ki, azt azonban igen, hogy az adott megfelelés legfeljebb a kiválasztott

számú konformer segítségével megvalósítható, több nem szükséges hozzá[8]. Az eljárás tehát a túlillesztés megítéléséhez felső korlátot ad.

A szerverbe alapszintű támogatást építettünk be a Kresten Lindorff-Larsen és munkatársai által kifejlesztett BME (Bayesian Maximum Entropy) módszerhez is[140]. A BME módszer az egyes konformerek súlyát módosítja annak érdekében, hogy egy adott sokaság jobban megfeleljen a kísérleti paramétereknek. A CoNSEnsX$^+$ szerver képes a neki megadott bemenet alapján olyan fájlt létrehozni, ami a BME eljárás bemeneteként szolgál és tartalmazza a kísérleti, valamint az egyes konformerekre becsült paramétereket. A CoNSEnsX$^+$ ezen felül képes a BME kimenete alapján egy sokaságot olyan módon is értékelni, hogy az egyes konformerek hozzájárulását annak alapján súlyozza.

4.5. Sokaság-alapú modellek alkalmazása fehérjék működésének elemzésében

Jelen fejezet az előzőekben ismertetett módszerek néhány konkrét alkalmazását mutatja be egyes kiválasztott fehérjék ill. fehérjecsaládok esetében.

4.5.1. Kisméretű kanonikus szerinproteáz-inhibitorok és a kulcs-zár hipotézis

A kanonikus szerinproteáz-inhibitorok közös szerkezeti jellemzője az úgynevezett kanonikus proteázkötő hurok, amellyel a gátolandó proteáz enzimhez kötnek [141]. Ebbe a mechanizmus alapján definiált csoportba számos, evolúciósan egymással nem rokon inhibitor tartozik.

A biokémiai tankönyvekben is megtalálható, elterjedt nézet szerint a proteázkötő hurok konformációja igen rigid, benne sem érdemi szerkezeti, sem dinamikai változás nem történik az enzimhez való kapcsolódás során [142, 143]. Ezáltal ezek a molekulák a klasszikus, Emil Fischer-féle kulcs-zár elmélet ideális megtestesítői lennének. Ezen felül az inhibitorok hatékonyságának is egyik feltételezett meghatározója ezen régió különleges merevsége. Ennek a nagyrészt korai röntgenkrisztallográfiai eredményekre támaszkodó elméletnek komolyabb kihívást jelentett, hogy az 1990-es évek közepétől megjelenő, egyes inhibitorok belső dinamikáját NMR-spektroszkópiai módszerekkel vizsgáló tanulmányok következetesen azt mutatták,

[8] azonban esetleg kevesebb konformer is elég lehet, ezt nem lehet kizárni

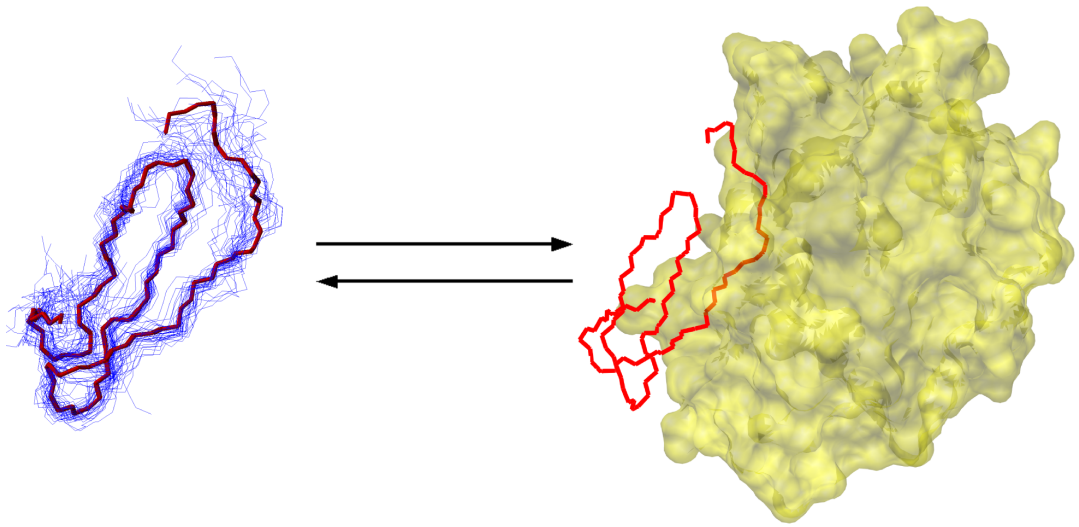
hogy a proteázkötő hurok a ps-ns időskálán ezen fehérjék többi részénél nagyobb mozgékonysággal rendelkezik [144, 145]. Emiatt a hurok merevségére alapozó hipotézisek legalábbis pontosításra szorulnak.

A saját PhD munkám keretében vizsgált, az ún. pacifastin családba tartozó kis inhibitorok (SGCI, *Schistocerga gregaria* chymotrypsin inhibitor és SGTI, *S. gregaria* trypsin inhibitor) esetében azt találtuk, hogy ezen kicsi, nagyjából 35 aminosavból álló inhibitorok egésze a nagyobb inhibitorok proteázkötő régiójára jellemző dinamikai paraméterekkel írható le [93].

A jelenség részletes vizsgálatára az általam a 4.4.1 alfejezetben leírt módon a GROMACS molekuladinamikai programcsomagba [146] implementált MUMO eljárást alkalmaztam. A számításokhoz az akkori kutatócsoportunkban mért és általam kiértékelt NMR-spektrumokból származó NOE és gerinc S^2 rendparamétereket haszáltam fel, további elemzésükhöz pedig a szintén saját méréseinkből származó kémiai eltolódásokat. Kiindulási szerkezetként a korábban általunk a PDB adatbázisban elhelyezett 1KGM (SGCI) ill. 1KJ0 (SGTI) kódú állományok [92] reprezentatív konformereit használtam.

Az általam előállított szerkezeti sokaságok a kiszámításukhoz közvetlenül nem felhasznált kémiai eltolódásokat is jobban tükrözik, mint a hagyományos módszerekkel meghatározott szerkezetek.

A szerkezeti sokaságok és az ismert, röntgenkrisztallográfiával meghatározott kötött állapotú szerkezetek főkomponens-elemzése megmutatta, hogy hogy az oldatfázisban az inhibitorok által felvett konformációk között megjelennek a kötött állapothoz nagyon hasonló szerkezetek is (XI. közlemény). Ennek értelmezéséhez figyelembe kell vennünk, hogy a felhasznált rendparaméterek által leírt dinamika ps-ns időskálájú, míg az enzimhez való kötés – más kanonikus inhibitorokra vonatkozó vizsgálatok alapján [147] – várhatóan ennél egy-két nagyságrenddel lassabb időskálán történik. Emiatt a vizsgált belső mozgások az enzimmel való kölcsönhatás szempontjából feltehetően nem számítanak sebességmeghatározó tényezőnek. Így az erősen dinamikus jelleg ellenére a gátlás mechanizmusa szempontjából kötőhurok mozgásai elhanyagolhatóak, a kötés folyamata jól leírható a merev közelítés segítségével (4.10. ábra). Fontos azonban megjegyezni, hogy a gátlás hatékonyságának merevségen alapuló magyarázata viszont mindenképpen pontosításra szorul, erre léteznek is elképzelések az irodalomban más kanonikus inhibitorok esetében [145], ezek részletesebb tárgyalása azonban kívül esik jelen értekezés keretein.

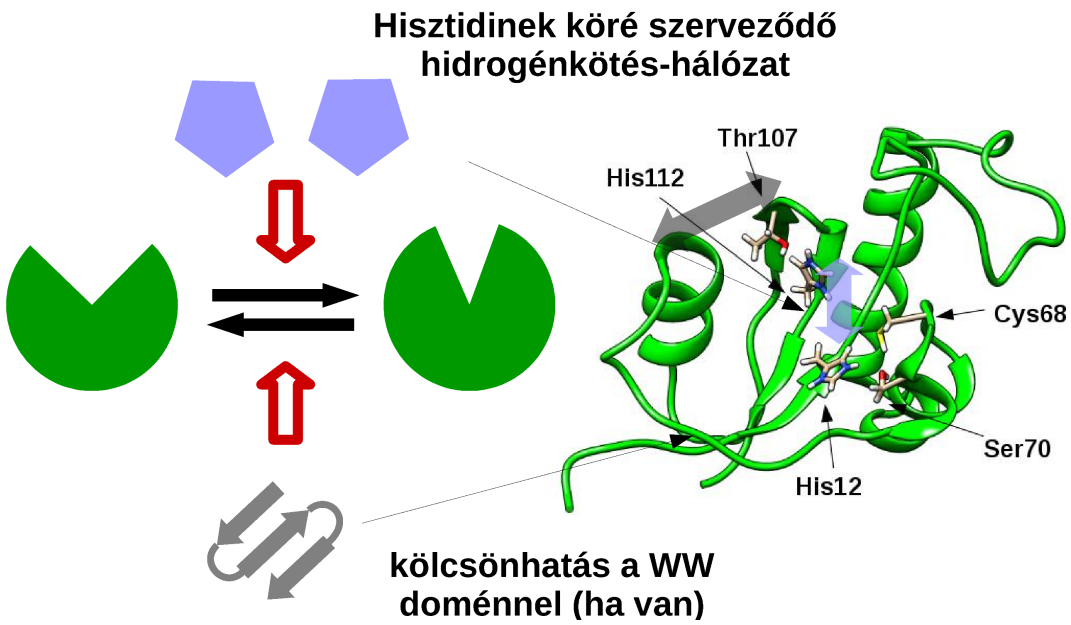


4.10. ábra. Az SGCI oldatfázisú dinamikája (balra) a ps-ns időskálán és az enzimkötött állapot (jobbra). A szabad forma oldatfázisú dinamikája során megjelennek az enzimkötött állapotra jellemző szerkezetek. A szabad állapotú dinamikát tükröző sokaságot kékkel, a kötött állapotra jellemző gerinckonformációt pirossal jelöltem minkét oldalon. *Az ábra az itt bemutatott eredményeinket is feldolgozó áttekintő közleményünkből [148] származik (licensz: CC BY)*

4.5.2. Parvulin típusú peptidil-prolil cisz-transz izomerázok összehasonlító elemzése

A peptidil-prolil cisz-transz izomerázok (EC 5.2.1.8) a prolin aminosav előtti amidkötés cisz-transz izomerizációját segítik elő. Ezek a fehérjék nem csupán más fehérjék feltekeredését segíthetik, hanem egyes fehérje-fehérje kölcsönhatási, felismerési helyek tulajdonságait is befolyásolhatják pl. génszabályozó folyamatokban [149]. Az ilyen enzimek több, evolúciósan nem rokon családba sorolhatóak. Mechanizmusuk kapcsán számos kérdés máig tisztázatlan, az igen valószínű, hogy az izomerizációs lépés során nem történik meg a peptidkötés felszakadása. Ugyanakkor az sem egyértelmű, hogy az egyes családok hasonló mechanizmust használnak-e az izomeráció katalizálásához.

Munkánk során az ún. parvulin izomeráz molekulacsaláddal foglalkoztunk, melynek legjobban tanulmányozott tagja a regulációs szerepet betöltő WW domént is tartalmazó Pin1 fehérje [150]. Munkánk során a család három különböző, WW domént nem tartalmazó, más-más szerepű és specifitású tagjának [94, 95, 96] gyors (ps-ns) időskálájú dinamikát tükröző sokaságát állítottuk elő a GROMACS programcsomagba implemetált MUMO protokoll segítségével, majd elvégeztük ezek összehasonlító elemzését. Ezek és a vizsgálatba bevont további szerkezetek esetében az egymással térben ekvivalens pozíciókat többszörös szerke-



4.11. ábra. A parvulinok szubsztrátkötő zsebének kinyíló mozgása és azt azt szabályozó faktorok, balra vázlatosan, jobb oldalon egy konkrét parvulin szerkezeten bemutatva, kiemelve a konzervált hisztidin aminosavak köré szerveződő hidrogénkötés-hálózatot.

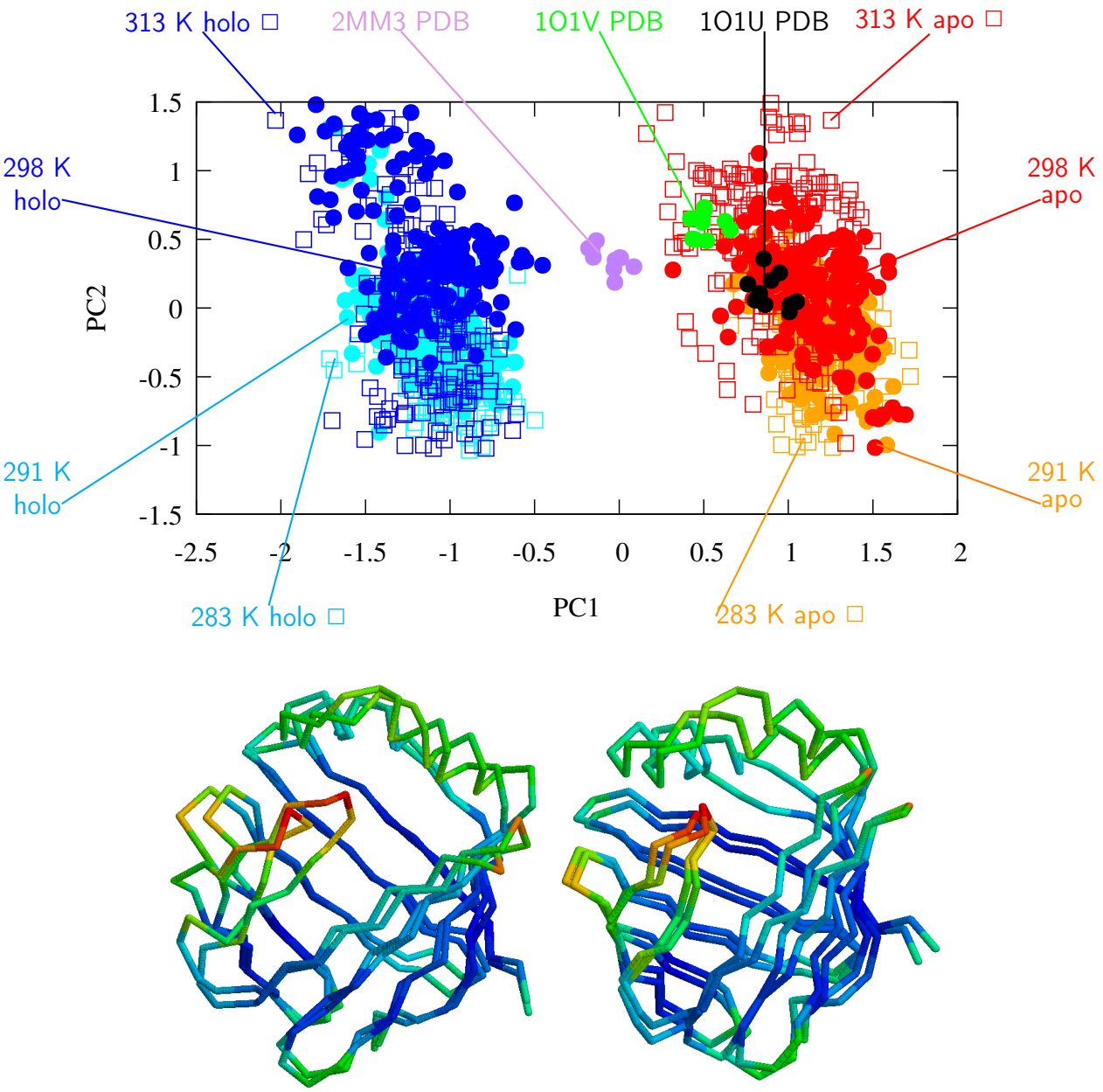
zetillesztéssel határoztuk meg (3.5.2. alfejezet). Megállapítottuk, hogy a vizsgált fehérjék közötti legfontosabb különbség a szubsztrátkötõ régió kinyílásának mértékében van. Az ezen mozgáshoz tartozó csukló régió átfedésben van a WW doménnel való kölcsönhatásért felelős aminosavak pozíciójával. Megállapítottuk továbbá, hogy ez az elmozdulás nem teljesen független a molekulacsaládban kulcsfontosságúnak tekintett hidrogénkötés-mintázat felépítésében részt vevő aminosavak egymáshoz viszonyított helyzetétől. Ezt a hálózatot részben vagy egészben konzervált aminosavak alkotják, melyek közül két, térben szomszédos hisztidin emelendő ki, melyek az irodalmi adatok alapján közvetlenül ugyan nem vesznek részt a katalízisben, de hozzájárulnak a molekulák aktivitásához [151, 152]. Elemzéseink alapján felállítottunk egy olyan modellt, amely szerint a kinyíló mozgás jellemzői hozzájárulnak az egyes enzimek specifitásának és aktivitásának meghatározásához. Maga a kinyíló mozgás pedig többféle módon befolyásolható, pl. a WW domént is tartalmazó parvulinok esetében az azzal való kölcsönhatás révén vagy a hidrogénkötés-mintázat finomhangolásának segítségével (4.11. ábra). Ezeket az eredményeinket a XVIII. közleményünkben írtuk le. Ezzel a modellel kiváló összhangban vannak az eredményeink publikálása után közölt, a Pin1-re vonatkozó egyes kísérleti adatok, melyek alapján egy, a hidrogénkötés-hálózat módosulásával járó aminosavcsere a WW domén hiányához hasonlóan vezet a molekula szerkezetének kompaktabbá válásához és egyúttal aktivitásbeli változásokhoz [153].

4.5.3. A humán epesavkötő fehérje ligandumkötése

A lipocalin családba tartozó humán epesavkötő fehérje vagy gasztrotropin egy β-hordó szerkezetű molekula, mely a belsejében képes két epesavmolekula egyidejű megkötésére is. A komplexről rendelkezésre álló szerkezetek alapján a szteránvázas epesav molekulák a hordó belsejében helyezkednek el, és magán a szerkezeten ebben az állapotban nem látszik olyan nyílás, amely által könnyen magyarázható lenne ezen viszonylag nagyméretű ligandumok bejutása a szerkezetbe. A bejutást magyarázó elméletek egyike az ún. portál hipotézis, amely szerint a szerkezet adott részeinek – a C/D és az E/F hurokrégiók, valamint a II. α-hélix C-terminális szakaszának – kinyílása nyitja meg az utat az epesavak számára a hordó belsejébe.

A humán epesavkötő fehérje belső dinamikája Tőke Orsolya és munkatársai kutatásainak köszönhetően igen részletesen leírt, több hőmérsekleten is rendelkezésünkre állnak S^2 rendparaméterek, illetve elérhetőek a lassabb időskálájú mozgásokkal összefüggésbe hozható relaxációs diszperziós paraméterek is [99]. Sajnos NOE alapú távolság jellegű kényszerfeltételek csak a szintén Tőke Orsolya csoportja által meghatározott, kétféle epasavat (glikokólsav, GCA és glikokenodezoxikólsav, GCDA) tartalmazó (holo) szerkezet (2MM3) esetében álltak rendelkezésünkre, így a szabad, ligandum nélküli (apo) forma esetében is ebből a listából indultunk ki, elvetve az ebben a szerkezetben nem teljesülő távolságokat. Munkánk során a MUMO eljárással az apo és holo állapotra is 4-4 dinamikus szerkezeti sokaságot állítottunk elő a négy különböző hőmérsékletre vonatkozó S^2 adatkészletek felhasználásával. Emellett több, megkötés nélküli sokaságot is generáltunk.

A szerkezetekben két alapvető belső mozgást azonosítottunk, az egyik, az 1. főkomponens (principal component 1, PC1) mentén észlelt az apo-holo átmenetnek megfelelő, leegyszerűsítve a hordószerkezet E és F β-szálak közötti kinyílásával jellemezhető mozgás. A másik, PC2 mentén megjelenő, mind az apo, mind a holo szerkezetben azonosított mozgásforma a helikális „fedő” régió részleges kitekeredését, ezzel együtt felnyílását okozza. Az utóbbi, második típusú mozgás a szekvencia N-terminális felén elhelyezkedő aminosavakra korlátozódik, melyekre a relaxációs diszperziós adatokból illesztett k_{ex} értékek közül az alacsonyabbak jellemzőek. Az apo-holo átmenetben érintett aminosavak a szekvencia teljes hossza mentén eloszlanak, a C-terminális, magasabb k_{ex} értékekkel jellemezhető szakaszra csak ilyenek esnek, ráadásul ugyanide koncentrálódnak a fizikailag legnagyobb térbeli elmozdulásokkal jellemezhető pozíciók.



4.12. ábra. A humán gasztrotropin belső mozgásai. Fent: szimulációból származó és a kísérletileg meghatározott szerkezetek főkomponens-analízise az egyes szerkezek első két módus (principal component 1 és 2, PC1 és PC2) mentén való eloszlásanak ábrázolásával. Az egyes szerkezeti sokaságokat külön színnel jelöltük, a PDB kódok az adatbázisban elérhető konformereket, a hőmérséklet és apo/holo címkék az általunk előállított dinamikus szerkezeti sokaságokat jelölik. Lent: a két módus mentén lévő mozgások végállapotai I. módus (PC1): bal oldal, II. módus (PC2): jobb oldal. A színezés "hőmérséklet" szerint történt színátmenetesen, ahol a piros szín a legmozgékonyabb, a kék a legkevésbé mozgékony régiókat jelöli. *A XIX. közleményünkben megjelent, angol neylvű 2. ábra módosított változata (licensz: CC BY) Harmat Zita PhD dolgozatából*

A kapott adatok egy lehetséges értelmezése, hogy a kétféle azonosított mozgásformához hasonlóak lassabb, μs-ms időskálán is megjelennek, az S^2 paraméterek segítségével kapott elmozdulások mintegy ezek gyors dőskálán – kisebb amplitúdóval és nem feltétlenül pontosan ugyanúgy – lezajló „párjai". Ez a hipotézis nem vizsgálható könnyen, hiszen ehhez tudnunk kellene a kísérleti adatokkal összhangban lévő, a lassabb időskálájú konformációs átmeneteket akár folyamatában leíró modelleket készíteni. A relaxációs diszperziós adatokból azonban csak egyes amid nitrogén atomokra vonatkozó, a két – a közvetlenül mérhető és a relaxációs elemzésből kikövetkeztetett, „rejtett" – állapot közötti kémiaieltolódás-különbséggel arányos $\Delta\omega$ értékek állnak rendelkezésünkre, melyek önmagukban korlátozott szerkezeti információval bírnak. Mindezzel együtt kísérletet tettünk olyan sokaságok előállítására, melyek összhangban lehetnek az illesztett $\Delta\omega$ paraméterekkel. Az AMD (accelerated molecular dynamics) sémát alkalmazva megnöveltük a szimulációk által bejárt konformációs teret, részlegesen kitekeredett szerkezeteket kapva, melyekre megbecsültük a kémiai eltolódásokat. A mért $\Delta\omega$ értékeket összevetettük a részlegesen kitekeredett konformerek és a natív apo szerkezet közötti becsült kémiaieltolódás-különbségekkel, és kiválasztottuk azokat az állapotokat, amelyek legjobban közelítik a kísérleti értékeket. Fontos hangsúlyozni, hogy az ilyen módon kapott szerkezeti modellek megbízhatósága jelentősen alatta marad a MUMO eljárással, lényegesen több – és szerkezeti adatokból pontosabban becsülhető – kísérleti paraméter felhasználásával készült szerkezetekének. Mindazonáltal a kapott, részlegesen kitekeredett szerkezetek alapvető sajátosságai jól értelmezhetőek a MUMO sokaságok által reprezentált mozgásformák segítségével.

A ligandumkötés vizsgálatára egyes kiválasztott szerkezetekbe a GLIDE [154] programmal bedokkoltuk a kétféle epesavat külön-külön és egymás után a kétféle lehetséges sorrendben. Ezek a vizsgálatok csupán statikus képet adnak, mégis arra utalnak, hogy a két ligandum egymás utáni bekötésének van preferált sorrendje, mégpedig a glikokólsav (GCA) glikokenodezoxikólsav (GCDA) előtti bekötése kedvezőbb.

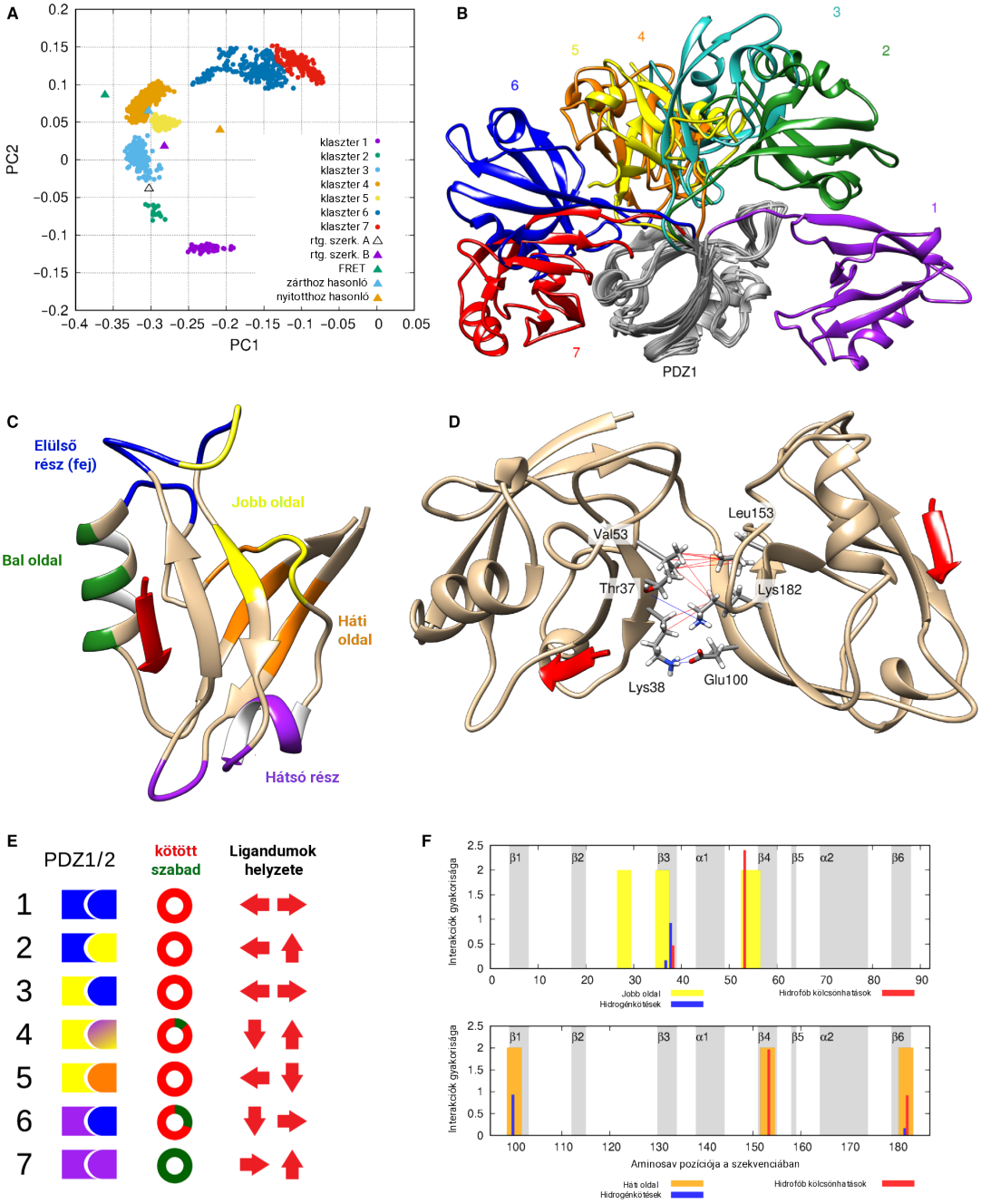
Összességében az általunk kapott adatok kompatibilisek a portál hipotézissel: olyan molekuláris mozgások meglétét valószínűsítik a megfelelő régiókban, melyek a szerkezetre általánosan jellemzőek – apo és holo állapotban is megfigyelhetőek – és lassabb időskálán, nagyobb amplitúdóval bekövetkező variánsaik a szerkezet részleges kitekeredése által lehetővé tehetik a ligandumok bejutását a hordó belsejébe (XIX. közlemény).

4.5.4. A PSD95 fehérje PDZ1-PDZ2 tandem doménjei

Kutatócsoportunkban az elmúlt évekbven egyre inkább előtérbe került a posztszinaptikus fehérjék vizsgálata. A PSD-95 nem csupán az egyik leggyakoribb posztszinaptikus állványfehérje, hanem igen kiterjedt irodalommal is rendelkezik. Az ún. MAGUK (membrane-associated guanylate kinase) családba tartozik, az ide tartozó fehérjékhez hasonlóan 3 PDZ, egy SH3 és egy GK doménnel rendelkezik, melyek közül mindegyiknek szerepe van fehérjepartnerek adott szakaszainak megkötésében. A molekulát funkcionálisan két ún. szupramodulra is szokás felosztani, az egyikbe az első két PDZ domén tartozik, a másikat a harmadik PDZ domén, valamint az SH3-GK domének alkotják. A PDZ1-PDZ2 domének funkcionális egységként való működését a köztük lévő rövid szakasz teszi lehetővé.

A tandem PDZ1-PDZ2 doménekről több, kísérletesen meghatározott térszerkezeti modell is elérhető, valamint NMR-spektroszkópiával behatóan tanulmányozták a belső dinamikájukat is. Saját vizsgálataink során azt a korábban irodalomban leírt jelenséget kívántuk atomi szintű modellek segítségével körüljárni, hogy a két domén ligandumkötött formában egymástól nagymértékben függetlenül mozog, míg szabad állapotban jellemzőbb az egy szerkezeti egységként való viselkedés. Ezt a következtetést NMR-spektroszkópiával mért relaxációs adatok alapján vonták le, és a kapcsolódó közleményben megadták a jellemző gerinc S^2 rendparamétereket is. A korábbiakban leírtak szerint MUMO eljárás általunk Gromacs programcsomagba implementált változata lehetővé tette a két doménre vonatkozó S^2 értékek egymástól független kezelését, így ideálisnak bizonyult a rendszer vizsgálatára. Az S^2 rendparamétereket megkötésként használó szimulációk mellett hagyományos, megkötések nélküli szimulációkat is futtattunk. Eredményeink jól összeegyeztethetőek a PDZ doménekről korábban leírtakkal, miszerint a ligandumkötés hatására a kötőzseb két oldalán elhelyezkedő hurokrégiók, a β1-β2 és a β2-β3 mutat jellegzetes elmozdulást. Ezek azonban a két domén esetében nem teljesen ugyanolyan jellegűek, a PDZ2 doménben ezek az elmozdulások kevésbé markánsnak tűnnek.

A tandem doménpár eetében kapott eredmények kvalitatív módon összhangban vannak a kísérletekkel, azaz a ligandumokötött állapotban a domének egymáshoz képest szabadabban elmozdulnak, mint a ligandumok nélkül. Ezen jelenség részletesebb elemzéséhez azokra a konformációkra fókuszáltunk, amelyekben a két domén közötti érintkezés jelentős (legalább 20 db 5 Å alatti, nehéz atomokat érintő kontaktus). Ezen szerkezetek főkomponens-elemzésekor



4.13. ábra. A PSD-95 fehérje tandem PDZ 1-2 doménpárjának szerkezeti-dinamikai elemzése. (**A**) A szoros interdomén kapcsolatot mutató szerkezetek főkomponens-elemzése és annak alapján történő klaszterezése. (**B**) az egyes klaszterek reprezentatív szerkezeteinek összevetése a tandem PDZ1 doménre illesztésének segítségével. (**C**) Az általunk kiemelt, a domén-domén interfész kialakításában részt vevő egyes régiók a PDZ domén szerkezetekben. (**D**) Az 5. klaszter reprezentatív szerkezete az interdomén kölcsönhatások kiemelésével. (**E**) az egyes klaszterekben kialakuló interdomén kapcsolatok sematikus ábrázolása a C panelben alkalmazott színkód segítségével. (**F**) az 5. klaszterre jellemző kölcsönhatásokat kialakító aminosavak elheyezkedése az egyes doménekben: a kék és piros színek a hirogénkötéseket és a hidrofób kölcsönhatásokat jelzik, a többi szín a C panelnek megfelelő régiók kiemelésére szolgál. Az interakciók gyakorisága akkor lehet nagyobb, mint 1, ha az adott aminosav több másikkal is érintkezik *Az ábra eredeti, angol nyelvű változata a XIII. közleményünkben jelent meg (licensz: CC BY-NC-ND)*

az egyes konformereket az első két komponens mentén ábrázolva egy jellegzetes, kör alakú elrendeződést kapunk, mely a domének egymáshoz képest való 3D orientációjának jól megfeleltethető: a PDZ1 domént illesztve a PDZ2 mintegy körbejárja azt az egyes állapotokban. A PCA eredmények alapján klaszterezést végeztünk és a kapott klasztereket elemezve megállapítottuk, hogy a legtöbb klaszterbe kizárólag vagy túlnyomórészt ligandumkötött szerkezetek tartoznak. A domének közötti fizikai kontaktusokat elemezve egyértelmű, hogy a fentebb említett β1-β2 ill. β2-β3 hurkok legalább egyike gyakorlatilag minden esetben részt vesz azok kialakításában. Ennek alapján azt a következtetést vontuk le, hogy az intra- és intedomén mozgások egymástól nem függetlenek, és az azok közötti kapcsolatot a ligandumkötésben is részt vevő szerkezeti régiók biztosítják. Atomi szintű szerkezeti modellek segítségével tehát képesek voltunk mechanisztikus magyazázatot adni a korábban kísérletileg észlelt dinamikai viselkedésre (XIII. közlemény). Megjegyzendő, hogy az általunk azonosított klaszterekben lévőkhöz hasonló, kísérletileg meghatározott szerkezetek nem minden esetben voltak korábban ismertek, ugyanakkor a közleményünk elfogadása előtt pár nappal jelent meg egy (azóta folyóiratközleményként is publikált) preprint [155], amely a 6. klaszterhez hasonló szerkezeteket azonosított, mintegy igazolva az általunk elméleti úton becsült szerkezeti diverzitás létjogosultságát.

4.5.5. A PSD95 PDZ3 domén ligandumkötése és a PDZ domének összehasonlító elemzése

A PSD-95 fehérje PDZ3 doménje a legintenzívebben tanulmányozott PDZ domének egyike. Különleges sajátsága, hogy a PDZ doménekre általánosan jellemző szerkezeti elemek mellett tartalmaz egy plusz α-helikális szakaszt a C-terminálisán. Ezen extra hélix eltávolítása befolyásolja a domén partnerkötési tulajdonságait, a megfogalmazott elméletek szerint egyfajta rejtett allosztéria jelensége áll fenn. Ezen allosztéria mechanisztikus hátterének feltárására Andrew Lee és munkatársai NMR-dinamikai méréseket végeztek a teljes hosszúságú, valamint a C-terminális hélixet alkotó 7 aminosavat nem tartalmazó rövid (Δ7CT) változatok szabad és ligandumkötött formáján. Ennek köszönhetően mind a négy állapot esetében molekulagerinc és oldallánc S^2 rendparaméterek is rendelkezésre álltak, így számunkra ez ideális rendszernek bizonyult ahhoz, hogy átfogó számításokat végezzünk ezeket a paramétereket megkötésként használva. Elsőként azt állapítottuk meg, hogy a molekulagerincre és az oldalláncokra vonatkozó S^2 paraméterek együttes használatával számolt sokaságok által lefedett

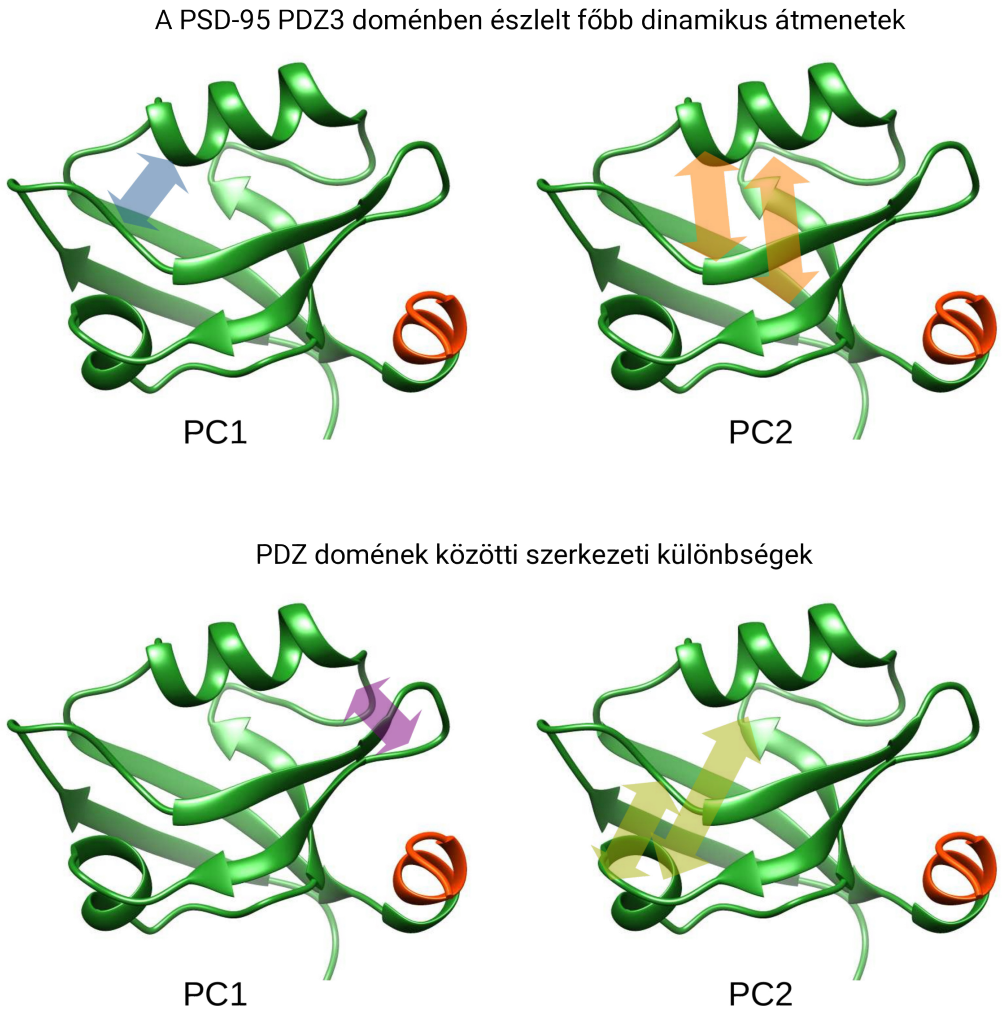
konformációs tér kismértékben, de különbözik azoktól az esetektől, amikor vagy csak a gerincre, vagy csak az oldalláncokra vonatkozó paramétereket használtuk. Így a továbbiakban mindkétféle megkötést alkalmaztuk. A kapott szerkezeti sokaságokból megkíséreltük megbecsülni a konformációs entrópia megváltozását, melyre vonatkozóan kísérletekből származó becslés is rendelkezésre állt. A konformációs entrópia megváltozásának szerkezeti sokaságokból történő becslése nem tekinthető egyértelműen megoldott kérdésnek, ezért egyetlen mérőszámot becsültünk, mégpedig azt, hogy ligandumkötés hatására bekövetkező entrópiaváltozás (ΔS) hányszorosára változik a C-terminális hélix eltávolításakor:

$$\frac{\Delta S_{\Delta 7CT}}{\Delta S_{full}} = \frac{S_{cd} - S_{fd}}{S_{cf} - S_{ff}} \tag{4.2}$$

ahol 'cd' a komplexben lévő Δ7CT változat, 'fd' ennek a szabad, ligandum nélküli formája, 'cf' és 'cf' pedig rendre a kötött és a szabad teljes hosszúságú változat[9]. Az ebből kapott 4,1 mint arányszám elfogadhatóan egyezik a kísérletekből kapott 2,4 aránnyal. A konformersokaságok főkomponens-elemzése alapján azonosított két legfontosabb szerkezeti átrendeződés a kötőzseb kinyíló-becsukódó (open-closed állapotok közötti) mozgása, valamint ugyanezen régió szélességének megváltozása (narrow-wide állapotok). Az egyes állapotok közötti legnagyobb különbség, hogy a ligandum nélküli $\Delta 7CT$ változat a nyitott és csukott állapotokat is mintavételezi, ezen módus mentén 'kétcsúcsú' eloszlást mutatva. A többi állapot és mozgás mentén ilyen jellegű eloszlás nem figyelhető meg. Ugyanakkor észrevehető az is, hogy a teljes szerkezet a szűkebb, míg a $\Delta 7CT$ változat a szélesebb kötőzsebbel jellemezhető konformációkat preferálja, bár ez kevésbé kifejezett, mint a kinyíló-becsukódó mozgás esetében. Ez azt is mutatja, hogy az egyes állapotok esetében a kétféle konformációs mozgás nem teljesen független egymástól. A fentebb leírtak a megfigyelt rejtett allosztéria dinamikus magyarázatát kínálják: a C-terminális hélix jelenléte – a keskeny-széles konformációk mintavételzését is megváltoztatva – leszűkíti a kinyíló-becsukódó mozgás által bejárható állapotokat, a szerkezeti sokaság által bejárt konformációs tér átrendeződését eredményezve.

Az eredmények tágabb kontextusba helyezéséhez az PDZ3, illetve az előző alfejezetben ismertetett PDZ1-PDZ2 domének szerkezeti sokaságai mellett nagyszámú további PDZ domént, illetve azok szerkezeti sokaságait felhasználva összehasonlító elemzést végeztünk. Ennek egyik kulcslépése volt a különböző PDZ domének közös szerkezeti magjának azonosítása,

[9] A jelöléseket az angol kifejezéseknek megfelelően használom, hogy konzisztensek legyenek a kapcsolódó közleményben szereplőkkel



4.14. ábra. A PDZ doménekben azonosított legfontosabb belső mozgások (fent) és szerkezeti különbségek (lent). A PSD95 PDZ3 domén esetében (fent) a PC1 és PC2 módusok az egyedi domén vizsgálatakor kapott mozgásformákra vonatkoznak. A domének közötti különbségek esetében (lent) a PC1 és PC2 a teljes, számos különböző domént tartalmazó elemzésben kapott első két módusra vontakozik. *Az ábra eredeti változata a XX közleményünkben jelent meg (licensz: CC BY)*

azaz azon aminosavak kiválasztása, melyek minden vizsgált PDZ doménben megtalálhatóak. Ehhez többszörös térszerkezetillesztést használtunk (lásd 3.5.2. alfejezet). A módszernek itt kifejezett előnye, hogy a PDZ domének nagymértékű szerkezeti változatossága ellenére lehetővé teszi az általánosan jellemző szerkezeti-dinamikai különbségek vizsgálatát. Elsősorban a terminális régiók véletlenszerű, a vizsgált kontextusban biológialag nem releváns elmozdulásainak, valamint a hurokrégiók eltérő hosszából adódó eltéréseknek a kiküszöbölése kritikus.

A több PDZ domén bevonásával készült főkomponens-elemzés számos olyan szerkezeti különbséget azonosított, amelyek túlmutatnak a kizárólag a PSD95 PDZ3 elemzésével kapottakon. Egyik legfontosabb eredményünk, hogy a különböző PDZ domének közötti szerkezeti

változatosság egyik fő forrása a β2-β3 hurok mellett lévő, a α2 hélixet és a β6 szálat összekötő szakasz. A másik, számunkra is igen meglepő megfigyelés, hogy a kinyíló-becsukódó mozgás ligandumkötéssel való összefüggése nem egységes: míg. pl. a PSD95 PDZ1 és PDZ2 doménjei esetében a ligandumkötött forma a nyitott, a PDZ3 doménnél éppen a zárt konformáció. Ezzel nem csupán az általunk molekuladinamikával generált szerkezeti sokaságok, hanem a PDB adatbázisban elérhető szerkezeteknek a főkomponensek mentén elfoglalt helyzete is összeegyeztethető (XX. közlemény). Ez a jelenség mindenesetre mindenképpen érdemes további, még részletesebb vizsgálatokra, melyek jelenleg is folyamatban vannak a csoportunkban.

4.5.6. A miozin VI SAH régió dinamikájának sokaság-alapú reprezentációja

A DIPEND eljárás kifejlesztése során az egyik tesztesetünk a miozin VI molekula SAH régiója volt. Ennek egyik oka a SAH motívumok iránti érdeklődésünk mellett a konkrét szakasz esetében rendelkezésünkre álló sokféle, NMR-spektroszkópiai mérésekből származó paraméter volt. A szakasz térszerkezeti modellje is elérhető a PDB adatbázisban. A dinamikus szerkezeti sokaság előállításához beállítottuk az egyes aminosavak szerkezeti preferenciáit: az α-helikális szerkezetnek megfelelő -58, -47 fokos ϕ, ψ szögpár körüli eloszlást adtunk hozzá a Dunbrack-féle szomszédfüggő preferenciákhoz, az első 55 aminosav esetében 0,99-es, az 56-58 szakasz esetében pedig 0,80-as súllyal. Ez a beállítás hivatott tükrözni a kémiai eltolódások alapján látható, a C-terminális régióban észlelt csökkenő helicitási trendeket. Az ezekkel a beállításokkal generált 5000 térszerkezetből a CONSENSX$^+$ eljárás segítségével szelektáltunk egy alsokaságot az N-H, N-C és H-C RDC értékek[10], a $^3J_{HNHA}$ skaláris csatolások és a másodlagos Cα kémiai eltolódások alapján. Itt megjegyzendő, hogy ezen paraméterek relatív skálázása szubjektív, ezért olyan optimumra törekedtünk, ahol mindegyik paraméternek való megfelelést az előzetes tesztek alapján megfelelőnek ítéltünk. Ezenfelül a maradvány dipoláris csatolások felhasználása eltért attól, ahogyan azokat az azt leíró közlemény alapján a szerkezetszámolás során a szerzők alkalmazták. Barnes és munkatársai az RDC értékeket az S^2 paraméterekkel skálázták, ilyen módon figyelembe véve a belső dinamika hatását – így egyébként ők egy teljesen egyenes helikális szerkezetet kaptak [35]. Mi az eredeti, skálázás nélküli értékeket használtuk, hiszen éppen a dinamika szerkezeti modellezése volt a célunk.

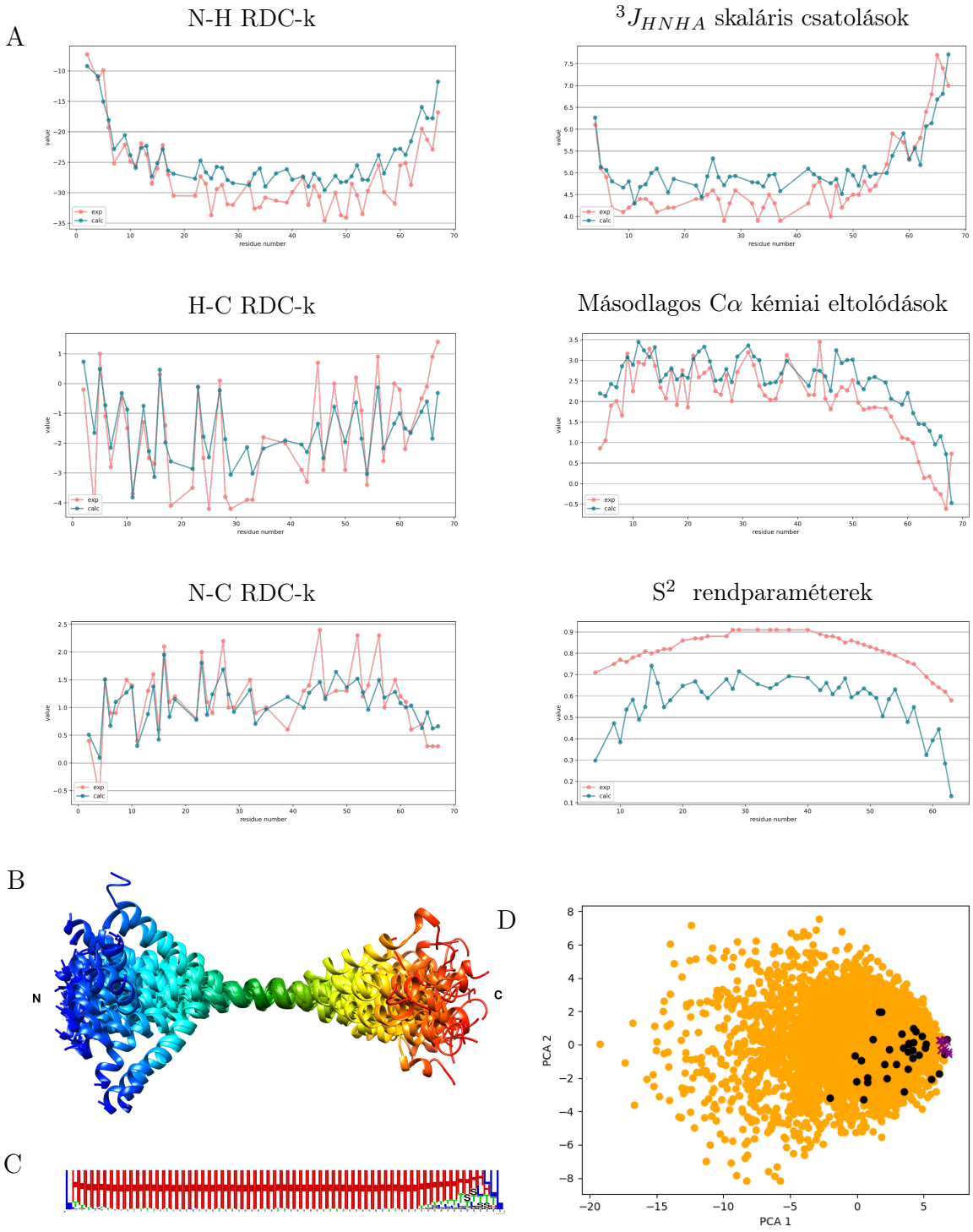
A kapott 37 szerkezet (4.15. ábra) alapvetően stabil helikális konformációt mutat, de a

[10]a betűk itt az amidkötésben található atomok nevei, N és H: amid nitrogén és hidrogén, C: karbonil szén

szakasz nem teljes hosszában egyenes, rúdszerű, hanem egy-egy helyen megtört alakot vesz fel. A másodlagos kémiai eltolódásoknak megfelelően a C-terminális végen egy hosszabb, mintegy 12-13 aminosavas szakaszon csökken a szabályos helicitás, míg ugyanez a trend az N-terminális végen kevésbé hangsúlyos, és csak 4-5 aminosavat érint.

A kapott szerkezeti sokaságot összevetettük a szelekcióban nem felhasznált S^2 rendparaméterekkel, ahol is nagyon magas korrelációt kaptunk annak ellenére, hogy a mért értékek lényegesen magasabbak az sokaságból visszaszámolhatóaknál – itt fontos leszögezni, hogy ezen paraméterek szerkezetből való becslése, különösen ilyen, elnyújtott szerkezet esetében igen erősen függ a szerkezetek egymásra illesztésének pontos módjától.

Az eredmények értelmezésénél fontos szempont, hogy az RDC és az S^2 értékek nem ugyanolyan időskálájú dinamikáról tudósítanak. A szerkezet megtörése esetünkben egyértelműen az RDC értékeknek való megfelelés eredménye, tehát a megfelelő mozgások vélhetően a mikroszekundumos időskálán jellemzőek. Ugyanakkor az S^2 értékekkel való korreláció arra utal, hogy hasonló mozgások a gyorsabb, ps-ns időskálán is jelen vannak, a mért és visszaszámolt S^2 paraméterek különbsége pedig úgy értelmezhető, hogy ezek amplitúdója kisebb. Ez a modell összhangban van az eredeti szerzők azon feltételezésével, hogy az RDC-k és az S^2 rendparaméterek hasonló mozgásformákat írnak le ebben az esetben. Összefoglalva, a lokálisan erősen helikális szerkezet egyes helyeken hajlamos az egyenes szerkezethez képest meghajlani, amely esetenként – a lassabb időskálán – akár egészen jelentős mértékben megtörheti a szabályos hélixet (XVII. közlemény).



4.15. ábra. A miozin VI SAH régió sokaság-alapú szerkezeti modellje. (**A**) Mért (piros) és a CoNSEnsX$^+$ által számított (kék) paraméterek. (**B**) A szelekcióval kapott 37 konoformer a 28–42 aminosavak mentén illesztve. A színezés a szekvencia mentén változik. (**C**) A sokaságra általgolt másodlagos szerkezeti állapotok (DSSPcont valószínűségek) ábrázolása a Weblogo program [156] segítségével. (**D**) A generált 5000 (narancs), a szelektált 37 (fekete) és a PDB-ben található 10 (lila) konformer (6OBI) által reprezentált konformációs tér összevetése főkomponens-elemzés segítségével, az első két módust ábrázolva. Az első módus a szerkezetek hosszát (N- és C-terminálisok távolsága) reprezentálja. *Az ábra eredeti, angol nyelvű változata a XVII. közleményünkben jelent meg (licensz: CC BY)*

5. Diszkusszió

5.1. A SAH mint szerkezeti motívum

Munkám során elsők között adtam becslést a SAH szakaszok elteredtségére vonatkozóan egy olyan eljárást használva, mely csupán az egyedileg vizsgált szekvenciák jellegzetességeit vizsgálja, nem veszi figyelembe az evolúciós rokonságot. A SAH szakaszok ritkák, de előfordulásuk jellegzetes mintázatot mutat egyes funkcionális kategóriákban, emiatt érdemes az azonosításuknak külön figyelmet szentelni, és egyértelműen elkülöníteni őket a coiled-coil vagy akár egyéb, nem globuláris szakaszoktól.

A terület új eredménye, hogy a korábbi ismereteinkkel és feltételezéseinkkel ellentétben SAH szakaszok mégis részt vehetnek közvetlenül nukleinsavak megkötésében. Ezt 2023-ban írták le egyes kormatinösszeszerelő fehérjék esetében, ahol is a SAH feltehetően a DNS-szállal párhuamosan elhelyezkedve a nukleoszómák pozicionálásáért lehet felelős [36]. Ez koncepciójában hasonló a SAH szakaszoknak a paraspeckle felépítésében általunk feltételezett szerepéhez, bár ott a SAH közvetlen asszociációja az RNS-sel még nem merült fel. A nukleinsavakkal való kölcsönhatás lehetősége érdekes új dimenziót nyit SAH szakaszokat tartalmazó fehérjék fázisszeparációban való részvétele kapcsán is.

A SAH szakaszok szerkezet-funkció összefügéseinek vizsgálatakor érdemes szem előtt tartanunk azt is, hogy a magas helicitás nem jár feltétlenül együtt egy teljesen merev, egyenes rúdszerű szerkezettel, hanem a struktúra rendelkezik valamennyi flexibilitással. Ennek pontos funkcionális jelentősége még tisztázandó, de pl. a miozin VI SAH esetében észlelt konformációs flexibilitás mértéke nem összeegyeztethetlen a korábban leírt erőkar szereppel.

Említést érdemes még a töltött aminosavakban gazdag motívumok szerepe a fehérje fázisszeparációban. Az ionos kölcsönhatások jelentősége a kondenzált fázisban önmagában természetesen nem újdonság [46]. Jelenleg már rendelkezésre állnak olyan molekuladinamikai szimulációkon alapuló – adott kísérleti paraméterekkel egyébként jól összhangba hozható

– elemzések [157], amelyekben a kondenzált fázis modellezésekor az egyes fehérjéket nagy példányszámban jelenítették meg, és így mód nyílik a közöttük kialakuló kölcsönhatások átfogó elemzésére. Ezen modellben megfigyelhetőek az ellentétes töltésű aminosavak között időlegesen létrejövő, majd átrendeződő kölcsönhatások. Ez egyrészt az egyik legelemibb megjelenése az LLPS jelenségben fontos multivalens kölcsönhatásoknak – egy adott aminosav a partner molekula sok aminosavával tud hasonló kölcsönhatást létesíteni –, másrészt „rímel" a SAH szakaszok esetében szintén szimulációk alaján leírt, molekulán belüli hasonló átrendeződésekre az ellentétes töltésű aminosavpárok között [2]. Általánosságban ezek alapján feltételezhető, hogy egy adott szabályszerűséget követő töltéseloszlás, azaz azonos és ellentétes aminosavak „blokkjainak" szabályos váltakozása hasonló átrendeződéseket tehet lehetővé valamivel nagyobb léptékben az ellentétes töltésű blokkokk között, és adott aminosavak/szakaszok esetében az intra- és intermolekuláris partnerek akár egymással kompetícióban is lehetnek.

Az SAH régiók előfordulására vonatkozó eredményeink értékelésekor és értelmezésekor két fontos szempontot kell még figyelembe venni: egyrészt, hogy a SAH predikciós eljárásunk kifejezetten szigorú (az elemzés elkészítésekor a még a jelenleginél is szigorúbb első változatot használtuk), emiatt a specifikusan SAH-ként felismert szakaszok száma a valósnál kisebb lehet. Másrészt általában a fehérjék belső dinamikáját, specifikusan pedig a coiled-coil szakaszok széles skálán mozgó stabilitását figyelembe véve lehetséges, hogy a rendezetlen - SAH - coiled-coil szerkezeti elemekre nem teljesen elkülönült szerkezeti állapotokként, hanem egy folytonos skála állomásainként kellene tekintenünk. Ez egyrészt igaz lehet úgy is, hogy egy adott motívum pl. kötőpartnerek jelenlététől függően lehet jelen monomer SAH vagy dimer coiled-coil formában, illetve evolúciós távlatokban is változhat egy molekulacsaládon belül egy adott szakasz által preferált térszerkezeti állapot. Ebben az értelemben a SAH egy speciális, „egyszálú coiled-coil" szerkezetként is felfogható. Ezeket a megfontolásokat részletesebben kifejtettük Nyitray Lászlóval közösen írt áttekintő cikkünkben [158].

5.2. Nem-globuláris szerkezeti elemek predikciója

Egy adott fehérjeszakasz szerkezeti preferenciáinak becsléséhez egyrészt az egyedi predikciók megbízhatóságát, másrészt egyéb jellegű predikciókkal való viszonyát is érdemes figyelembe venni. Az egyedi coiled-coil predikciós eljárások teljesítményére vonatkozó 2010-es eredményeinket egy 2021-es tanulmány is megerősítette [159], ugyanakkor ez nem foglalkozik

részletesen a keresztpredikciók kérdésével. A saját predikciós elemzéseink során szerzett tapasztalataink alapján azt javasoljuk, hogy érdemes egy specifitási sorrendet felállítani, amely szerint egyes becsült szerkezeti tulajdonságoknak precedenciája van mások előtt, pl.: SAH > coiled-coil > rendezetlen. Ennek alapja, hogy a legspecifikusabb predikció eredményét tekintjük legmegbízhatóbbnak, tehát a SAH szegmenskre és coiled-coil szakaszokra speciálisan fejlesztett eljárások információtartalmát a rendezetlenség-predikciókénál magasabbnak tekintjük. Megjegyzendő, hogy ebben a szemléletben a rendezetlenség predikcióját valójában „nem-globularitási" predikcióknak tekintjük, hiszen hagyományosan a rendezetlen szakaszokat a globulárisakkal szemben szokás definiálni. Ebbe a kontextusba a fibrilláris szakaszok jól beilleszthetőek, mint a nem-globuláris szakaszokra fejlesztett prediktorok által felismert, de mégsem a klasszikus rendezetlen szakaszok definíciójával leírható szegmensek. Természetesen tisztában vagyok azzal, hogy az ezzel foglalkozó kutatók körében nincs feltétlenül egyértelmű konszenzus a kérdésben, hiszen számos továbbbi megfontolás tehető pl. annak alapján, hogy egy adott szekvencia önmagában képes-e valamilyen szerkezetet felvenni, ezt adott esetben kooperatív módon teszi-e stb. Magam azt az álláspontot képviselem, hogy a fiziológiásan releváns szerkezeti állapoto(ka)t tartom lényegesnek, és azokat a predikciós eljárásokat és folyamatokat, amelyek erre vonatkozólag tudnak minél pontosabb támpontot adni. Értelemszerűen ez az egyik legkomplexebb problémakör, melyre jelenleg nincs is igazán jó megoldás, ezt nagyon jól mutatja a dolgozat írásakor igen népszerű AlphaFold eljárás által generált rengeteg, az adott multimer fehérjékre vonatkozó specifikus tudásunk alapján fiziológiásan irreleváns monomer szerkezet elérhetősége publikus adatbázisokban, mint pl. az emberi vázizomban található miozin II (MYH2_HUMAN) esetében.

Mindezen megfontolások alapján egy szakasz csak akkor tekinthető funkcionálisan rendezetlennek, ha nem becsültük egyúttal coiled coilnak, SAH-nak, esetleg kollagén hélixnek. Ezen szempontok már megjelennek egyes elemzésekben [160], de még egészen friss közleményekben is előfordul, hogy pl. csak a coiled-coil szakaszokat emelik ki, az egyéb fibrilláris elemeket nem kezelik elkülönítve a „valódi" rendezetlen régióktól [161]. Általánosabb megközelítésben az alacsony komplexitású szakaszok szerkezeti elemzésekor elengedhetetlen az aminosavösszetétel és a repetitivitás együttes figyelembevétele. A terület számos szakértőjével közösen írt áttekintő cikkünkben kifejtjük, hogy két, hasonló aminosavösszetételű szekvencia közül az ismétlődő motívumokat tartalmazó nagyobb eséllyel vesz fel valamilyen rendezett szerkezetet, amely tipikusan nem globuláris, hanem fibrilláris [162]. Ebben a szemléletben a

SAH szakaszok – a coiled-coil és kollagén szegmensekkel összhangban – is megfelelően kontextusba helyezhetőek.

5.3. *De novo* fehérjék szerkezeti jellemzői

Az általunk leírt fő eredmény, miszerint a kódoló DNS-szakasz GC-tartalma meghatározó a kódolt fehérje hidrofobicitása és ezáltal szerkezeti preferenciái szempontjából, tekinthető triviálisan levezethetőnek a genetikai kód szerveződésének ismeretében. Tudomásunk szerint ugyanakkor sem a *de novo* fehérjék elemzése, sem általában a véletlenszerű szekvenciák vizsgálata során ez a szempont nem volt hangsúlyosan jelen az irodalomban – sok esetben a véletlen szekvenciákat az egyes aminosavak azonos előfordulási valószínűségét feltételezve modellezték. Mindemellett nem találtunk olyan közleményt sem, amelyik egyszerre többféle szerkezeti preferenciát vizsgált volna, jobban elhelyezve ezáltal a szekvenciákat a szerkezeti térben. Ezek alapján utólagosan visszatekintve is indokolt volt ezen szempontok beemelése a diskurzusba, a valós evolúciós forgatókönyvhöz közelebb hozva a véletlenszerűen átíródó szekvenciák vizsgálatát. Az összetett predikciók alkalmazása révén tesztelhető becslést tudtunk adni a várható szerkezeti preferenciákra, különös tekintettel a munka elkezdésekor jelen lévő, az aggregációt fő evolúciós tényezőként beállító megfontolásokra a *de novo* fehérjekeletkezés témakörében.

A *de novo* fehérjék vizsgálata a mi tanulmányunk megjelenése óta is egyre intenzívebben folyik. Nem csupán több *de novo* megjelent konkrét fehérjét ismerünk, hanem számos elméleti és kísérleti eredmény született az újonnan születő fehérjék szerkezeti jellegzetességei kapcsán, melyek jó összhangban vannak az általunk levont következtetésekkel [163]. Az irodalomban jelenleg formálódó konszenzus is abba az irányba mutat, hogy a *de novo* fehérjék jellemzően rendezetlenek, és ez előnyös mind a fennmaradásuk, mind az evolvabilitásuk szempontjából [164].

Itt megjegyezhető még, hogy a véletlen szekvenciák generálásával modellezett forgatókönyv csak mintegy mellékesen adhat számot a DNS-szintű ismétlődések, pl. mikroszatellitek expanziójával keletkező [165], alacsony komplexitású, de szintén ismétlődő jelleget mutató fehérjékről és fehérjeszakaszokról. A mikroszatellitek evolúciója és elterjedése túlmutat az egyedi DNS-szekvenciák egyszerű fizikokémiai tulajdonságain, adott élőlénytől függő jellegeket is mutat ([166, 167], ez árnyalhatja az újonnan keletkező fehérjék jellemzőit is.

A genetikai kód evolúciója kapcsán tett megállapításaink természetüknél fogva a fentieknél jóval spekulatívabbak. A kód kialakulására vonatkozó elméleteknek jó áttekintését adja Kun Ádám Evolúcióbiológia című könyvében [168], melyben az a forgatókönvy is megjelenik, hogy a templát-alapú szintézis megjelenését megelőzte a kód kialakulása, azaz az aminosavak kodonokhoz rendelése. Ezt a lehetőséget elfogadva természetesen az általunk tett megállapítások helye is más megvilágításba kerül: a mi elemzésünk arra mutat rá, hogy a néhány aminosavas korai kódok nem teszik lehetővé a ma észlelthez hasonló szerkezeti jellemzőkkel bíró hosszabb polipeptidláncok kialakulását. Amennyiben a templát alapján szintetizált fehérjék esetében a néhány aminosavas kódszótárak szerepe még elhanyagolható, úgy a mi eredményeink relevanciája is csökken.

Ugyanakkor, ha feltételezzük egy primitív transzlációs apparátus meglétét, akkor számolnunk kell pl. azzal a problémával, hogy egyes aminosavkészletek esetén a transzmembrán régiók kialakítására képes szegmensek gyakorlatilag hiányoznak – a korai peptidek (membrán)transzportfolyamatokban betöltött lehetséges szerepét könyvében Kun Ádám is felveti, és a mi diszkussziónk is kiemeli a membránokkal való kölcsönhatás szükségességét. Mindezeken túl természetesen a fehérje fázisszeparáció jelensége a korai fehérjék szerepe és jellege kapcsán is új lehetséges forgatókönyveket vet fel.

Összességében a saját elemzésünk szerepét leginkább abban látom, hogy a genetikai kód evolúciójára vonatkozó, hagyományosan jelenlévő gondolatmeneteket egy új, szerkezetorientált, predikciókkal – legalábbis bizonyos mértékig – elemezhető és tesztelhető aspektussal bővíti. A kérdéskör komplexitása miatt természetesen ezen a területen könnyen előfordulhat, hogy még évtizedekig várnunk kell egy széles körben elfogadott, a történeti megfontolásokal és fizikokémiai megkötésekkel is összhangban lévő elmélet megszületésére.

Jelen dolgozat szempontjából a korai kódok elemzésének legfontosabb aspektusa az alacsony komplexitású szekvenciákhoz kapcsolódik. Ezeknél a vizsgálatoknál ütközött ki ugyanis legjobban, hogy a tipikus predikciós eljárások a mai, 20 aminosavból felépülő, alapvetően magas komplexitású fehérjeszekvenciák elemzésére készültek és így elsősorban ezek vizsgálatára alkalmasak. A redukált aminosavkészlettel rendelkező fehérjék esetében előfordul. hogy az egyes predikciós eljárások jelentősen eltérő eredményt adnak - emiatt is igen fontosnak tartom azt az egyébként a saját munkáinkon következetesen alkalmazott megközelítést, hogy lehetőség szerint többféle eljárás konszenzusát alkalmazzuk az egyes szerkezeti preferenciák vizsgálatakor. A másik aspektus, amelyet már a a közleményünkben is tárgyalunk, a külső kö-

rülmények szerepe – itt azóta új szempontként ismét csak felvetődhet a fázisszeparáció, mely a „molecular crowding” jelenségének és annak szerkezetekre vonatkozó hatása értelmezésének újragondolásához vezethet.

5.4. Dinamikus szerkezeti sokaságok: előállítás és elemzés

A kísérletileg meghatározott belső dinamikai paramétereket tükröző sokaság-alapú szerkezeti modellek jelenleg még nem tekinthetőek általánosan elterjedtnek. Ennek egyik oka, némileg meglepő módon, a kísérleti adatok korlátozott elérhetősége, illetve az elérhető adatok minősége. Bár egy fehérjeszerkezet publikálásának megkövetelt velejárója a megfelelő kísérleti adatok hozzáférhetővé tétele, ez a gyakorlatban sajnos nem jelenti azt, hogy ezek ténylegesen könnyen elérhetőek vagy értelmezhetőek lennének minden esetben. Példának elegendő a PDB adatbázisból letölthető „v2 NMR restraints” állományokat említeni, melyeket automatikus konverzióval jönnek létre a kutatók által beküldött távolság jellegű kényszerfeltétel-listából, és emiatt az atomi nevezéktantól kezdve a láncazonosítókig számos hibával terheltek. Sok esetben mindemellett nincsen információ – a nyers adatfájlokban sem – a szerkezetszámolás egyes beállításairól, különös tekintettel a kémiailag vagy mágnesesen ekvivalens magokat érintő távolságok átlagolási módszertanáról, ami a tényleges, fizikai atomi távolságokat lényeges mértékben befolyásolja[1].

A másik fő probléma a sokaság-alapú modellek előállítására alkalmas programok elérhetősége ill. használata. Ma sem feltétlenül egyszerű feladat olyan eljárást találni, mely adott kísérleti paramétert az adott kutató igényeinek megfelelően képes kezelni, jól paraméterezhető, valamint technikailag elérhető és futtatható. Bár számos publikációban találhatunk ilyen számításokra alkalmas programot vagy eljárást, egy korábban még nem vizsgált fehérje szerkezeti sokasáágnak előállítása gyakran kíván meg egyedi megoldásokat, mint pl. a DIPEND eljárásban is elérhető, lokális szerkezeti preferenciák felhasználó általi beállítása (pl. [169]). A nyílt forrás mint lényeges követelmény pedig csak mostanában jelenik meg hangsúlyosan, itt a 2022-ben Julie Forman-Kay csoportja által közölt IDPCONFORMERGENERATOR [170] említhető példaként.

Végül meg kell említenünk az ilyen sokaságok előállítására és értelmezésére szolgáló eljárások összetettségét, amely túlmutat az önmagában is komplex, ám ma már rutinszerűen

[1]Erről a kérdésről a 2019-es NMR munkabizottsági ülésen tartottam rövid előadást „The violation that never was: why methyl group distances are not what they seem in proteins?” címmel

alkalmazott szerkezetmeghatározási protokollokon. Ezekhez képest a sokaság-alapú modellek hozzáadott értéke a kutatók szemében nem feltétlenül van arányban az előállításukhoz szükséges plusz erőfeszítésekkel. A globuláris fehérjék esetében tipikus, hogy a dinamikus sokaságokat már korában meghatározott szerkezetekre támaszkodva állítják elő, és ezekben az esetekben viszont a „hagyományos" molekuladinamikai számításokkal kell összevetnünk a konformersokaság előállításának nehézségeit. Az általánosan alkalmazott molekuladinamikai protokollok könnyebben paraméterezhetőek, mivel nem kell egy (vagy több) plusz energiatag hozzájárulását beállítani. Emellett a jelenlegi GPU-alapú architektúrákon igen hosszú szimulációs idő érhető el az egyetlen replikát alkalmazó számítások esetén. Ráadásul a fehérjemolekulák lassabb időskálán történő, nagyobb léptékű elmozdulásait leíró NMR paraméterek (pl. RDC) alkalmazása nehezebb, mint a gyors dinamikára vonatkozó S^2 rendparamétereké.

A funkcionálisan rendezetlen fehérjék ellemzésének területén azonban csak sokaság-alapú modellek segítségével lehet az atomi szintű szerkezeti jellemzőket leírni és megérteni, ezért ezek esetében tapasztalható az ilyen jellegű modellek előretörése. A 2014-ben létrehozott Protein Ensemble Database [171] is elsősorban funkcionálisan rendezetlen fehérjék sokaságait tartalmazza.

5.4.1. A sokaságok kísérleti adatoknak való megfelelése

Az S^2 rendparamétereknek való megfelelés a MUMO eljárás természetéből fakadóan sokszor lényegesen javul a PDB-ben elérhető szerkezeti sokaságokhoz képest. Ennek oka, hogy a hagyományos szerkezetszámolás során kifejezetten törekszünk arra, hogy az egyes konformerek egymáshoz minél hasonlóbbak legyenek, és az alacsony RMSD-re való törekvés együtt jár azzal hogy ezekből a sokaságokból magas, 1-hez közeli S^2 paramétereket lehet visszaszámolni. Ugyanakkor egy hagyományos, megkötések nélküli molekuladinamika sok esetben nem képes az egyes molekuláris régiók közötti jellegzetes különbségeket sem visszaadni.

A kémiai eltolódások vizsgálata elsőre kevésbé informatívnak tűnik ezekben az esetekben. Ennek oka, hogy a globuláris fehérjék esetében a hagyományos úton meghatározott szerkezet természetszerűleg már jó megfelelést mutat a kísérleti paraméterekkel, melyen a sokaság-alapú modellezés sokszor csak kismértékben javít. Ilyen esetekben a megfelelés értelmezése leginkább az, hogy a sokaság-alapú modell nem ront az adott paraméternek, pl. kémiai eltolódásnak való megfelelésen, tehát az eljárásunk feltételezhetően nem távolította el a modellt a valóságtól a korábbi állapotnál jobban. Az általam vizsgált globuláris fehérjék

esetében a kémiai eltolódások nem jelennek meg kényszerfeltételként, a keresztvalidáció során használtam fel őket.

A CoNSENsX$^+$ eljárás legújabb fejlesztésekor bekerült a másodlagos kémiai eltolódások vizsgálata, amely a korábbiaknál sokkal érzékenyebb mérőszámot jelent a lokális szerkezeti preferenciák vizsgálatakor. A másodlagos kémiai eltolódások már alkalmazhatóak funkcionálisan rendezetlen fehérjék sokaságainak előállítására is. Ezek esetében a közvetlenül mért „elsődleges" eltolódások a globuláris esetnél lényegesen közelebb állnak a „random coil" értékekhez, emiatt azok kevésbé informatívak. A rendezetlen fehérjék esetében ezzel együtt a másodlagos kémiai eltolódásoknak való megfelelés önmagában nem biztosítja a teljes szerkezet térbeli szerveződésének, kiemelten a molekula kompaktságának és az esetleges távoli, harmadlagos kölcsönhatásoknak – melyek, ha gyengék is, de jelen lehetnek – a leírását. Emiatt ezekben az esetekben feltétlenül szükséges valamilyen olyan paraméter felhasználása, amely a molekula egyes részeinek egymáshoz képest való orientációjáról szolgáltat információt. Ez az aspektus pl. maradvány dipoláris csatolások (RDC-k), kisszögű röntgenszórás (SAXS), esetleg paramágneses relaxációs kísérletek (paramagnetic relaxation enhancement, PRE) segítségével jellemezhetőek.

Ezen a ponton már összefonódik az értékelés és a szelekcióval történő előállítás problémaköre. Emiatt szót kell ejtenem az egyes paraméterek relatív súlyozásának problémájáról is. Ez jelenleg emprikus módon történik a CoNSENsX$^+$ eljárásban, a felhasználó egy tízes skálán tudja a súlyokat változtatni, és egyes beállítások esetében megvizsgálni a kapott megfeleléseket, majd végül kiválasztani az általa legjobbnak tartott paraméterezést. A nehézséget nem csak az egyes paraméterek eltérő jellege jelenti, hanem az is, hogy adott molekula esetében adott típusú paraméterből eltérő számú állhat rendelkezésre, ez tipikus pl. a kémiai eltolódások és az RDC-k esetében is. A kémiai eltolódások esetében külön nehézség, hogy az egyes atomtípusokra nem ugyanolyan megbízhatósággal becsülhető.

5.4.2. A sokaságokban észlelt konformációs eltérések

Az általunk előszeretettel alkamazott főkomponens-elemzés (PCA) egy hipotézismentes eljárás, melynek segítségével előfeltevések nélkül elemezhetőek a sokaságok által reprezentált mozgások. Tapasztalataink azt mutatják, hogy az ennek segítségével kapott konformációs különbségek megtalálhatóak a PDB adatbázisban elérhető, hagyományos módszerekkel meghatározott szerkezetek esetében is, ugyanakkor jóval kevésbé észrevehető módon. A dinami-

kus szerkezeti sokaságok ezeket a mozgásokat mintegy „felnagyítva" mutatják, azaz az adott főkomponens mentén az egyes szerkezeti állapotok jobban elkülönülnek. Az „eredeti" PDB szerkezetek bevonása az elemzésbe és annak vizsgálata, hogy azok esetében is megjelenik-e az adott főkomponens menti szétválás, fontos aspektus annak megítélésére, hogy a szimulációk során megjelenő elmozdulások mennyire tükrözhetik a valóságot. Értelmezésem szerint pl. a parvulin, gasztrotropin és PDZ szerkezetek esetében a szimulált sokaságok és a PDB szerkezetek konzisztens eloszlása a PCA diagramokon nagymértékben alátámasztja az levont következtetéseket.

Több itt bemutatott elemzésünk fontos aspektusa az összehasonlító jelleg. Kiemelten törekedtünk arra, hogy ahol a molekulacsalád több tagjáról is rendelkezésre álltak felhasználható kísérleti adatok, ott több sokaságot állítsunk elő és ezeket egymással összevetve értelmezzük. Ez a megközelítés jelentősen hozzájárult a parvulin szerkezetek elemzésekor kapott eredményekhez, ahol ez tette egyértelműen azonosíthatóvá a fő kinyíló-becsukódó mozgástípust. A PDZ domének esetében is fontos új aspektusokat tártunk fel az összehasonlító elemzés segítségével. Ezek az eredmények némileg módosítják azt az irodalomban korábban megjelenő elképzelést, hogy egy adott molekulacsalád tagjainak szerkezeti változatossága jól megfeleltethető a család egy tagjának belső mozgásai során bejért konformációs térnek ([172, 173]. A mi eredményeink azt mutatják, hogy az utóbbi jobban behatárolt, mint a teljes családra jellemző eltérések. Megjegyzendő, hogy itt fontos aspektus lehet a „molekulacsalád" pontos behatárolása, azaz milyen diverzitást veszünk figyelembe pl. a szekvenciák szintjén, valamint az egyedi dinamika által bejárt konformációs tér feltérképezésének módja is. Hagyományos molekuladinamikai számítások molekulacsaládok szerkezeti diverzitásával való összevetése inkább az általunk megfigyeltekkel van összhangban [174].

Az általam javasolt általános modellben létezik az adott családra jellemző, funkcionálisan jellemző konformációs mozgás, mint pl. a ligandum hatására bekövetkező átrendeződés, ennek pontos mértékét és megvalósulását azonban az adott molekulában előforduló szerkezeti sajátosságok modulálják. A parvulinok esetében a kinyíló-becsukódó mozgást befolyásolja a hisztidinek protonáltsága, illetve a WW domén megléte és működése, a PDZ doméneknél pedig a ligandumkötéshez kapcsolódó mozgások pontos mikéntjét a β3 hurkot érintő egyéb konformációs átrendeződések – melyeket pl. a PSD95 PDZ3 domén esetében az α3 hélix jelenléte modulál – hangolja doménspecifikus módon.

Természetesen a rendelkezésre álló adatok mennyisége még túl kevés ahhoz, hogy egy ilyen

általános hipotézis alátámasztható legyen, és nyilván az általunk leírt esetekben is szükséges a független megerősítés. Ennek nehézségét éppen a dinamikai jellegeket összehasonlító vizsgálatok viszonylagos ritkasága adja. Itt érdemes még megemlíteni, hogy természetüknél fogva a fehérjék N- és C-terminális régiói, illetve hosszú hurokrégiói a legimozgékonyabbak, emiatt pl. ezek elmozdulásai könnyen dominálhatják az elemzéseket, pl. főkomponens-elemzéseknél ezek könnyen megjelenhetnek fő módusként (lásd pl. [175]). Az általunk alkamazott megközelítés, mely a molekulacsalád minden tagjában meglévő pozíciókra fókuszál, éppen ezen régiók mozgásaira nem érzékeny, csupán annyiban, amennyiben ezek a szerkezet „magját" is befolyásolják. Így várható, hogy az azonosított konformációs különbségek tényleges biológiai relevanciával rendelkeznek a molekulacsalád általános működésének tekintetében.

5.5. Összefoglalás

Mind a szekvenciák, mind a szerkezeti sokaságok elemzésével kapott eredményeim illeszkednek abba az egyre hangsúlyosabban megjelenő paradigmába, hogy a dinamika evolúciós skálán történő finomhangolása és megváltozása szorosan kapcsolódik az új molekuláris funkciók megjelenéséhez [176]. Ezen szemlélet alapján az egyes fehérjék a konformációs tér adott tartományát foglalják el, és mind funkcionális állapotváltozásaik, mind evolúciós átmeneteik során ennek a tartománynak a helye és kiterjedése változik meg [177]. A konformációs tér komplexitását növeli az egyes állapotok közötti energiagátak magasságának kérdése, mely az átmenetek időskálájával van összefüggésben [56]. Ez a szemlélet a fehérjeműködés és -evolúció tekintetében a diszkrét állapotok helyett inkább a folytonosságra helyezi a hangsúlyt. Véleményem szerint ezen öszefüggések mélyebb feltárásához a jelenleg belátható út nem csupán az egyedi eljárások szabatosságának javításán, hanem a különböző elméleti megfontolások és kísérleti adatok integrálásán át is vezet, ideértve a mesterséges intelligencián alapuló megoldások innovatív felhasználását is [178].

Az értekezés tézispontjai

1. Eljárást készítettem a magányos α-helikális motívum (SAH) detektálására, és azt másik módszerrel együtt alkalmazva feltérképeztem a SAH szegmensek előfordulását teljes proteomokban. Megállapítottam, hogy a motívum kifejezetten jellemző RNS-kötő fehérjékre. Kimutattam, hogy a SAH szegmensek és általában a szabályos töltésmintázatot mutató régiók gyakoribbak a fázisszeparációra hajlamos fehérjékben, de jelenlétük nem általánosan szükséges feltétel. Elkészítettem a paraspeckle felépítésében résztvevő RNS-kötő fehérjekomplex szerkezeti modelljét a SAH szegmens és a vele szomszédos coiled-coil szegmens figyelembevételével.

2. Feltérképeztem egyes fibrilláris motívumok, mint a SAH, coiled-coil és kollagén hélix, valamint a funkcionálisan rendezetlen fehérjeszakaszok predikciós programok általi felismerésének összefüggéseit. Megállapítottam, hogy a „valódi", nagy konformációs szabadsággal rendelkező rendezetlen régiók felismeréséhez a fibrilláris szakaszokra specializált predikciók eredményét is szükséges figyelembe venni, és a predikciókat egymással összhangban érdemes értelmezni.

3. Konszenzus predikciók segítségével kimutattam, hogy az újonnan keletkező *de novo* fehérjék esetében a rendezett és rendezetlen régiók várható aránya erősen függ a kódoló nukleinsavszakasz GC-tartalmától. Hasonló eljárással az élet korai evolúciójára javasolt kódokat elemezve arra jutottam, hogy azok nem alkalmasak globuláris fehérjék kódolására, így vagy új evolúciós forgatókönyvek, vagy a maitól jelentősen eltérő fizikokémiai körülmények feltételezése szükséges.

4. A GROMACS szabad hozzáférésű molekuladinamikai programcsomagba implementáltam az irodalomban korábban javasolt számítási módszert az S^2 rendparaméterek kényszerfeltételként való felhasználására és azok kombinálását replikapáronkénti NOE alapú kényszerfeltételekkel. A módszert továbbfejlesztettem a szimulált replikák S^2 számo-

lás során való, felhasználó által megadható módon történő szuperpozíciójával. Funkcionálisan rendezetlen fehérjeszakaszok sokaságainak modellezéséhez elkészítettem az irodalomban elérhető aminosav-szomszédságtól függő lokális szerkezeti preferenciákat figyelembe vevő, felhasználó által könnyen paraméterezhető, nyílt forrású DIPEND eljárást.

5. Elkészítettem a CONSENSX$^+$ webszervert, mely képes NMR-spektroszkópiai adatok teljes fehérjeszerkezeti sokaságoknak való megfelelésének elemzésére, emellett alkalmas a kísérleti adatoknak jobban megfelelő alsokaságok kiválasztására.

6. Irodalmi adatokra támaszkodva előállítottam több fehérjedomén dinamikus szerkezeti sokaságait és azonosítottam a funkció – ligandum-, ill. partnerkötés – szempontjából fontos belső mozgásokat:

   (a) Javaslatot tettem a kanonikus szerinproteáz-inhibitorok klasszikus kulcs-zár elmélet szerinti működése és a kötőhurok dinamikus viselkedésének összehangolására.

   (b) Egységes modellt állítottam fel a parvulin típusú peptidil-prolil izomerázok működési és szabályozási mechanizmusára.

   (c) A gasztrotropin fehérje ligandumkötési mechanizmusa kapcsán a portál hipotézis pontosítását javasoltam.

   (d) Összefüggést találtam a PSD-95 PDZ1-2 tandem domének intra-és interdomén dinamikája között, és meghatároztam a PDZ domének általános szerkezeti-dinamikai változatosságának egyes aspektusait.

   (e) A kísérleti adatokkal összhangban lévő sokaság-alapú modellt készítettem a miozin VI SAH domén dinamikájnak leírására.

Az értekezés alapjául szolgáló közlemények

I. Dániel Süveges, Zoltán Gáspári, Gábor Tóth, László Nyitray: Charged single α-helix: a versatile protein structural motif. *Proteins* (2009) 74:905-916.

II. Zoltán Gáspári, Dániel Süveges, András Perczel, László Nyitray, Gábor Tóth: Charged single alpha-helices in proteomes revealed by a consensus prediction approach. *Biochem Biophys Acta - Proteins and Proteomics* (2012) 1824:637-646.[1]

III. Dániel Dudola, Gábor Tóth, László Nyitray, Zoltán Gáspári: Consensus prediction of charged single alpha-helices with CSAHserver. In: Zhou, Kloczkowski, Faraggi, Yang (eds): Prediction of Protein Secondary Structure. *Methods Mol Biol* Vol. 1847, Springer, 2017, pp. 25-34.

IV. Ákos Kovács, Dániel Dudola, László Nyitray, Gábor Tóth, Zoltán Nagy, Zoltán Gáspári: Detection of single alpha-helices in large protein sequence sets using hardware acceleration. *J Struct Biol* (2018) 204:109-116.

V. László Dobson, László Nyitray, Zoltán Gáspári: A conserved charged single alpha-helix with a putative steric role in paraspeckle formation. *RNA* (2015) 21:2023-2029.

VI. András László Szabó, Anna Sánta, Rita Pancsa, Zoltán Gáspári: Charged sequence motifs increase the propensity towards liquid–liquid phase separation. *FEBS Lett* (2022) 596:1013-1028.

VII. Balázs Szappanos, Dániel Süveges, László Nyitray, András Perczel, Zoltán Gáspári: Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett* (2010) 584:1623-1627.

[1] Szabadon elérhető preprint formában a CSAH szerver oldaláról

VIII. Zoltán Gáspári: Is five percent too small? Analysis of the overlaps between disorder, coiled-coil and collagen predictions in complete proteomes. *Proteomes* (2014) 2:72-83.

IX. Annamária F. Ángyán, András Perczel, Zoltán Gáspári: Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Lett* (2012) 586:2468-2472.

X. Annamária F. Ángyán, Csaba Ortutay, Zoltán Gáspári: Are proposed early genetic codes capable of encoding viable proteins? *J Mol Evol* (2014) 78:263-274.

XI. Zoltán Gáspári, Péter Várnai, Balázs Szappanos, András Perczel: Reconciling the lock-and-key and dynamic views of canonical serine protease inhibitor action. *FEBS Lett* (2010) 584:203-206.

XII. Ádám Fizil, Zoltán Gáspári, Terézia Barna, Florentine Marx, Gyula Batta: "Invisible" conformers of an antifungal disulfide protein revealed by constrained cold & heat unfolding, CEST-NMR experiments and molecular dynamics calculations. *Chem Eur J* (2015) 21:5136-5144.

XIII. Bertalan Kovács, Nóra Zajácz-Epresi, Zoltán Gáspári: Ligand-dependent intra- and inter-domain motions in the PDZ12 tandem regulate binding interfaces in postsynaptic density protein-95. *FEBS Lett* (2020) 594:887-902.

XIV. Annamária F. Ángyán, Balázs Szappanos, András Perczel, Zoltán Gáspári: CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data. *BMC Struct Biol* (2010) 10:39

XV. Dániel Dudola, Bertalan Kovács, Zoltán Gáspári: CoNSEnsX^{+} web server for the analysis of protein structural ensembles reflecting experimentally determined internal dynamics. *J Chem Inf Model* (2017) 57:1728-1734.

XVI. Annamária F. Ángyán, András Perczel, Sándor Pongor, Zoltán Gáspári: Fast protein fold estimation from NMR-derived distance restraints. *Bioinformatics* (2008) 24:272-275.

XVII. Zita Harmat, Dániel Dudola, Zoltán Gáspári: DIPEND: an open-source pipeline to generate ensembles of disordered segments using neighbor-dependent backbone preferences. *Biomolecules* (2021) 11:1505

XVIII. András Czajlik, Bertalan Kovács, Perttu Permi, Zoltán Gáspári: Fine-tuning the extent and dynamics of binding cleft opening as a potential general regulatory mechanism in parvulin-type peptidyl prolyl isomerases. *Sci Rep* (2017) 7:44504

XIX. Zita Harmat, András László Szabó, Orsolya Tőke, Zoltán Gáspári: Different modes of barrel opening suggest a complex pathway of ligand binding in human gastrotropin. *PLoS ONE* (2019) 14:e0216142

XX. Dániel Dudola, Anett Hinsenkamp, Zoltán Gáspári: Ensemble-based analysis of the dynamic allostery in the PSD-95 PDZ3 domain in relation to the general variability of PDZ structures. *Int J Mol Sci* (2020) 21:8348

Irodalomjegyzék

[1] E Callaway. 'The entire protein universe': AI predicts shape of nearly every known protein. *Nature*, 608:15–16, 2022.

[2] S. Sivaramakrishnan, BJ. Spink, AYL. Sim, S. Doniach, és JA. Spudich. Dynamic charge interactions create surprising rigidity in the er/k α-helical protein motif. *Proc Natl Acad Sci USA*, 105:13356–13361, 2008.

[3] RA. Laskowski. Structural quality assurance. In J. Gu és PE. Bourne, editors, *Structural Bioinformatics, 2nd ed.*, old.: 341–376. Wiley-Blackwell, 2009.

[4] B. Jiménez-Garcia, P. Bernadó, és J. Frenández-Recio. Structural characterization of protein-protein interactions with pyDockSAXS. In Z. Gáspári, editor, *Structural Bioinformatics: methods and protocols*, old.: 131–144. Humana Press, 2020.

[5] M. Trellet, G. van Zundert, és AMJJ. Bonvim. Protein-protein modeling using cryo-EM restraints. In Z. Gáspári, editor, *Structural Bioinformatics: methods and protocols*, old.: 145–162. Humana Press, 2020.

[6] D. Schneidman-Dubovny és HJ. Wolfson. Modeling of multimolecular complexes. In Z. Gáspári, editor, *Structural Bioinformatics: methods and protocols*, old.: 163–174. Humana Press, 2020.

[7] M. Berg, J-L. Tymocko, és L. Stryer. *Biochemistry*, old.: 63–64. WH Freeman, New York, 5. kiadás, 2002.

[8] M. Lella és R. Mahalakshmi. Metamorphic proteins: Emergence of dual protein folds from one primary sequence. *Biochemistry*, 56:2971–2984, 2017.

[9] DK. Ghosh és A. Ranjan. The metastable states of proteins. *Protein Sci*, 29:1559–1568, 2020.

[10] A. Bhattarai és IA. Emerson. Dynamic conformational flexibility and molecular interactions of intrinsically disordered proteins. *J Biosci*, 45:29, 2020.

[11] J. Jumper et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

[12] B. Zhao, S. Ghadermarzi, és L. Kurgan. Comparative evaluation of AlphaFold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins. *Comp Struct Biotech J*, 21:3248–3259, 2023.

[13] VJ. Promponas, AJ. Enright, S. Tsoka, DP. Kreil, C. Leroy, S. Hamodrakas, C. Sander, és CA. Ouzounis. CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, 16:915–922, 2000.

[14] N.. Radó-Trilla és MM. Alba. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol Biol*, 12:155, 2012.

[15] P. Romero, Z. Obradovic, X. Li, EC. Garner, CJ. Brown, és AK Dunker. Sequence complexity of disordered protein. *Proteins*, 1:38–48, 2001.

[16] F. Saudou és S. Humbert. The biology of huntingtin. *Neuron*, 89:910–926, 2016.

[17] A.N. Lupas és M. Gruber. The structure of α-helical coiled ccoils. *Adv Prot Chem*, 70:37–78, 2005.

[18] B. Brodsky és AV. Persikov. Molecular structure of the collagen triple helix. *Adv prot Chem*, 70:301–339, 2005.

[19] F.H.C. Crick. The packing of α-helices: simple coiled coils. *Acta Crystallog*, 6:689–697, 1953.

[20] A. Rich és F.H.C. Crick. The structure of collagen. *Nature*, 176:915–916, 1955.

[21] SP. Boudko, J. Engel, és HP. Bächinger. The crucial role of trimerization domains in collagen folding. *Int J Biochem Cell Biol*, 44:21–32, 2012.

[22] J. Walshaw és D.N. Woolfson. SOCKET: a program for identifying and analysing coiled coil motifs within protein structures. *J Mol Biol*, 307:1427–1450, 2001.

[23] C.W. Wood, M. Bruning, AA. Ibarra, GJ. Bartlett, AR. Thomson, RB. Sessions, RL. Brady, és DN. Woolfson. CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics*, 30:3029–3035, 2014.

[24] C.W. Wood és D. Woolfson. CCBuilder2: Powerful and accessible coiled coil modeling. *Protein Sci*, 27:103–111, 2018.

[25] M. Gruber, J. Söding, és A.N. Lupas. Comparative analysis of coiled-coil prediction methods. *J Struct Biol*, 105:140–145, 2006.

[26] A.N. Lupas, M. Van Dyke, és J. Stock. Predicting coiled coils from protein sequences. *Science*, 252:1162–1164, 1991.

[27] MD. Shoulders és RT. Raines. Collagen structure and stability. *Annu rev Biochem*, 78:929–958, 2009.

[28] VK. Pálfi és A. Perczel. How stable is a collagen triple helix? An *ab initio* study on various collagen and β-sheet forming sequences. *J Comput Chem*, 29:1374–1386, 2008.

[29] R. Chebrek, S. Leonard, AG. de Brevern, és J-C. Gelly. PolyPrOnline: polyproline helix II and secondary structure assignment database. *Database*, 2014:bau102, 2014.

[30] TS. Jardetzky, JH. Brown, JC. Gorga, LJ. Stern, RG. Urban, JL. Strominger, és DC. Wiley. Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides. *Proc Natl Acad Sci USA*, 93:734–738, 1996.

[31] MT Reymond, G. Merutka, JH. Dyson, és PE. Wright. Folding propensities of peptide fragments of myoglobin. *Protein Sci*, 6:706–716, 1997.

[32] R. Aurora és GD. Rose. Helix capping. *Protein Sci*, 7:21–38, 1998.

[33] P.J. Knight, K. Thirumurugan, J. Xu, F. Wang, A.P. Kalverda, W.F. III. Safford, J.R. Sellers, és M. Peckham. The predicted coiled-coil domain of myosin 10 forms a novel elongated domain that lenghtens the head. *J Biol Chem*, 280:34702–34708, 2005.

[34] E. Baker, G.J. Bartlett, M.P. Crump, R.B. Sessions, N. Linden, C.F.J. Faul, és D.N. Woolfson. Local and macroscopic electrostatic interacions in single α-helices. *Nat Chem Biol*, 11:221–228, 2015.

[35] CA. Barnes, Y. Shen, J. Ying, Y. Takagi, DA. Torchia, JR. Sellers, és A. Bax. Remarkable rigidity of the single α-helical domain of myosin-VI as revealed by NMR spectroscopy. *J Am Chem Soc*, 141:9004–9017, 2019.

[36] R. Rosas, RR. Aguilar, N. Arslanovic, A. Seck, DJ. Smith, JK. Tyler, és MAE. Churchill. A novel single alpha-helix dna-binding domain in CAF-1 promotes gene silencing and DNA damage survival through tetrasome-length DNA selectivity and spacer function. *eLife*, 12:e83538, 2023.

[37] M. Lee, A. Sadowska, I. Bekere, D. Ho, BS. Gully, Y. Lu, Iyer KS., J. Trewhella, AH. Fox, és CS. Bond. The structure of human SFPQ reveals a coiled-coil mediated polymer essential for functional aggregation in gene regulation. *Nucleic Acids Res*, 43:3826–3840, 2015.

[38] K. Samejima, M. Platani, M. Wolny, H. Ogawa, G. Vargiu, P.J. Knight, M. Peckham, és W.C. Earnshaw. The inner centromere protein (INCENP) coil is a single α-helix (SAH) domain that

binds directly to microtubules and is important for chromosome passenger complex (CPC) localizaion and function in mitosis. *J Biol Chem*, 290:21460–21472, 2015.

[39] A.K.C. Ulrich, M. Seeger, T. Schutze, N. Bartlick, és M.C. Wahl. Scaffolding the spliceosome via single α helices. *Structure*, 24:1–12, 2016.

[40] JA. Spudich és S. Sivaramakrishnan. Myosin VI: an innovative motor that challenged the swinging lever arm hypothesis. *Nat Rev Mol Cell Biol*, 11:128–137, 2010.

[41] M. Wolny, M. Batchelor, PJ. Knight, E. Paci, L. Dougan, és M. Peckham. Stable single α-helices are constant force springs in proteins. *J Biol Chem*, 289:27825–27835, 2014.

[42] C-LA. Wang, JM. Chalovich, P. Graceffa, RC. Lu, K. Mauchi, és WF. Stafford. A long helix from the central region of smooth muslce caldesmon. *J Biol Chem*, 266:13958–13963, 1991.

[43] D. Simm, K. Hatje, és M. Kollmar. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α-helices (SAH domains). *Bioinformatics*, 31:767–769, 2015.

[44] D. Simm és M. Kollmar. Waggawagga-cli: A command-line tool for predicting stable single α-helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS ONE*, 13:e0191924, 2018.

[45] DT. McSwiggen, Mir M., Darzacq X., és Tjian R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes Dev*, 33:1619–1634, 2019.

[46] GL. Dignon, RB. Best, és J Mittal. Biomolecular phase separation: from molecular driving forces to macroscopic properties. *Annu rev Phys Chem*, 71:53–75, 2020.

[47] B. Mészáros, G. Erdős, . Szabó, B, É. Schád, Á. Tantos, R. Abukhairan, T. Horváth, N. Murvai, OP. Kovács, M. Kovács, SCE. Tosatto, P. Tompa, Z. Dosztányi, és R. Pancsa. Phasepro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res*, 48:D360–D367, 2020.

[48] B. Wang, L. Zhang, T. Dai, Z. Qin, H. Lu, L. Zhang, és F. Zhou. Liquid–liquid phase separation in human health an diseases. *Signal Transduct Target Ther*, 6:290, 2021.

[49] C. Hou, X. Wang, H. Xie, T. Chen, P. Zhu, X. Xu, K. You, és T. Li. Phasepdb in 2022: annotating phase separation-related proteins with droplet states, co-phase separation partners and other experimental information. *Nucleic Acids Res*, 51:D460–D465, 2022.

[50] A. Andreeva, D. Howorh, C. Chothia, E. Kulesha, és A.G. Murzin. Investigating protein structure and evolution with SCOP2. *Curr Protoc Bioinf*, 49:1.26.1–1.26.21, 2015.

[51] I. Sillitoe, N. Bordin, N. Dawson, VP. Waman, P. Ashford, HM. Scholes, CSM. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, SD. Lam, K. Berka, IH. Varekova, R. Svobodova, J. Lees, és CA. Orengo. CATH: increased structural coverage of functional space. *Nucleic Acids Res*, 49:D266–D273, 2020.

[52] D. Tautz és T. Domazet-Loso. The evolutionary origin of orphan genes. *Nat Rev Genet*, 12:692–702, 2011.

[53] DG. Knowles és A. McLysaght. Recent de novo origin of human protein coding genes. *Genome Res*, 19:1752–1759, 2009.

[54] E. Monsellier és F. Chiti. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep*, 8:737–742, 2007.

[55] J. Schnabel. Protein folding: The dark side of proteins. *Nature*, 464:828–829, 2010.

[56] K. Henzler-Wildman és D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

[57] PS. Nerenberg és T. Head-Gordon. New developments in force fields for biomolecular simulations. *Curr Opin Struct Biol*, 49:129–138, 2018.

[58] V. Li és V. Daggett. Investigation of the solution structure of chymotrypsin inhibitor 2 using

molecular dynamics: comparison to x-ray crystallographic and nmr data. *Protein Eng*, 8:1117–1128, 1995.

[59] R. B. Best és M. Vendruscolo. Determination of protein structures consistent with NMR order parameters. *J. Am. Chem. Soc.*, 126:8090–8091, 2004.

[60] X. Salvatella. Understanding protein dynamics using conformational ensembles. *Adv Exp Med Biol*, 805:67–85, 2014.

[61] R. B. Best és M. Vendruscolo. Structural interpretation of hydrogen exchange protection factors in proteins: characterization of the native state fluctuations of CI2. *Structure*, 14:97–106, 2006.

[62] K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, és M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433:128–132, 2005.

[63] O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, és B. L. de Groot. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, 320:1471–1475, 2008.

[64] B. Richter, J. Gsponer, P. Varnai, X. Salvatella, és M. Vendruscolo. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, 37:117–135, 2007.

[65] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, és M. Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J. Am. Chem. Soc.*, 131:17908–17918, 2009.

[66] M. Krzeminski, J. A. Marsh, C. Neale, W. Y. Choy, és J. D. Forman-Kay. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*, 29:398–399, 2013.

[67] KJ. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, és M. Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J Am Chem Soc*, 131:13894–13895, 2009.

[68] WF. Vranken. NMR structure validation in relation to dynamics and structure determination. *Prog Nucl Magn Reson Spectrosc*, 82:27–38, 2014.

[69] R. Ferella, A. Rosato, és P Turano. Determination of protein structure and dynamics. In I. Bertini, KS. McGreevy, és G. Parigi, editors, *NMR of Biomolecules: Towards Mechanistic Systems Biology*, old.: 51–94. Wiley, 2012.

[70] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res*, 51:D523–D531, 2023.

[71] H. Mi, A. Muruganujan, és PD. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*, 41:D377–D386, 2013.

[72] D. Piovesan et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res*, 45:D219–D227, 2016.

[73] SK. Burley et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51:D488–D508, 2022.

[74] PR. Romero, N. Kobayashi, JR. Wedell, K. Baskaran, T. Iwata, M. Yokochi, D. Maziuk, H. Yao, T. Fujiwara, G. Kurusu, JC. Ulrich, EL. an Hoch, és JL. Markley. BioMagResBank (BMRB) as a resource for structural biology. In Z. Gáspári, editor, *Structural Bioinformatics: methods and protocols*, old.: 187–218. Humana Press, 2020.

[75] W. Li és A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658–1659, 2006.

[76] B. Rost, C. Sander, és Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol*, 235:13–26, 1994.

[77] Z. Dosztanyi, V. Csizmok, P. Tompa, és I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 347:827–839, 2005.

[78] K. Peng, P. Radivojac, S. Vucetic, AK. Dunker, és Z. Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7:208, 2006.

[79] ZR. Yang, R. Thomson, P. McNeil, és RM. Esnouf. RONN: the biobasis function network technique applied to the detection of natively disordered regions in proeins. *Bioinformatics*, 21:3369–3376, 2005.

[80] AV. McDonnell, T. Jiang, AE. Keating, és B. Berger. Paircoil2: Improved prediction of coiled coils from sequence. *Bioinformatics*, 22:356–358, 2006.

[81] SR. Eddy. A new generation of homology search tools based on probabilistic inference. *Genome Inform*, 23:205–211, 2009.

[82] GE. Tusnády és I. Simon. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17:849–850, 2001.

[83] M. Cserző, F. Eisenhaber, I. Eisenhaber, és I. Simon. TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, 20:136–137, 2004.

[84] A. Krogh, B. Larsson, G. von Heijne, és ELL. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305:567–580, 2001.

[85] L. Käll, A. Krogh, és ELL. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338:1027–1036, 2004.

[86] AM. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, és L. Serrano. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, 22:1302–1306, 2004.

[87] S. Maurer-Stroh, M. Debulpaep, N. Kuemmerer, M. Lopez de la Paz, IC. Martins, J. Reumers, KL. Morris, A. Copland, L. Serpell, Serrano L., J. Schymkowitz, és F. Rousseau. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods*, 7:237–242, 2010.

[88] SO. Garbuzynskiy, MYu. Lobanov, és OV. Galzitskaya. Foldamyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26:326–332, 2010.

[89] D. Van Der Spoel, E. Lindahl, B Hess, G. Groenhof, Mark AE., és HJC. Berendsen. GROMaCS: fast, flexible and free. *J Comput Chem*, old.: 1701–1718, 2005.

[90] P. Gümtert és L. Buchner. Combined automated noe assignment and structure calculation with CYANA. *J Biomol NMR*, 62:453–471, 2015.

[91] DM. Passon, M. Lee, O. Rackham, WA. Stanley, A. Sadowska, A. Filipovska, AH. Fox, és CS. Bond. Structure of the heterodimer of human NONO and Paraspeckle Protein Component 1 and analysis of its role in subnuclear body formation. *Proc Natl Acad Sci USA*, 109:4846–4850, 2012.

[92] Z. Gáspári, A. Patthy, L. Gráf, és A. Perczel. Comparative structure analysis of proteinase inhibitors from the desert locust, Schistocerca gregaria. *Eur J Biochem*, 269:527–537, 2002.

[93] B. Szenthe, Z. Gaspari, A. Nagy, A. Perczel, és L. Graf. Same fold with different mobility: backbone dynamics of small protease inhibitors from the desert locust, Schistocerca gregaria. *Biochemistry*, 43:3376–3384, 2004.

[94] O. Heikkinen, R. Seppala, H. Tossavainen, S. Heikkinen, H. Koskela, P. Permi, és I. Kilpeläinen. Solution structure of the parvulin-type PPIase domain of Staphylococcus aureus PrsA–implications for the catalytic mechanism of parvulins. *BMC Struct Biol*, 9:17, 2009.

[95] L. Sun, X. Wu, Y. Peng, JY. Goh, YC. Liou, D. Lin, és Y. Zhao. Solution structural analysis of the single-domain parvulin TbPin1. *PLoS ONE*, 7:e43017, 2012.

[96] Ł. Jaremko, M. Jaremko, I. Elfaki, JW. Mueller, A. Ejchart, P. Bayer, és I. Zhukov. Structure and dynamics of the first archaeal parvulin reveal a new functionally important loop in parvulin-type prolyl isomerases. *J Biol Chem*, 286:6554–6565, 2011.

[97] M. Kurz, V. Brachvogel, H. Matter, S. Stengelin, H. Thüring, és W. Kramer. Insights into the bile acid transportation system: the human ileal lipid-binding protein-cholyltaurine complex and its comparison with homologous structures. *Proteins*, 50:312–328, 2003.

[98] G. Horváth, Á. Bencsura, Á. Simon, GP. Tochtrop, GT. DeKoster, DF. Covey, DP. Cistola, és O. Toke. Structural determinants of ligand binding in the ternary complex of human ileal bile acid binding protein with glycocholate and glycochenodeoxycholate obtained from solution NMR. *FEBS J*, 283:541–555, 2016.

[99] G. Horváth, O. Egyed, és O. Toke. Temperature dependence of backbone dynamics in human ileal bile acid-binding protein: implications for the mechanism of ligand binding. *Biochemistry*, 53:5186–5198, 2014.

[100] W. Wang, J. Weng, X. Zhang, M. Liu, és M. Zhang. Creating conformational entropy by increasing interdomain mobility in ligand binding regulation: a revisit to N-terminal tandem PDZ domains of PSD-95. *J Am Chem Soc*, 131:787–796, 2009.

[101] CM. Petit, J. Zhang, PJ. Sapienza, EJ. Fuentes, és AL. Lee. Hidden dynamic allostery in a pdz domain. *Proc Natl Acad Sci USA*, 106:18249–18254, 2009.

[102] EL. Ulrich, K. Baskaran, Dashti, YE. H. Ioannidis, M. Livny, PR. Romero, D. Maziuk, JR. Wedell, H. Yao, JC. Eghbalnia, HR. Hoch, és JL. Markley. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *J Biomol NMR*, 73:5–9, 2019.

[103] WL. Jorgensen, DS. Maxwell, és J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*, 118:11225–111236, 1996.

[104] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, JL. Klepeis, RO. Dror, és DE. Shaw. Improved side-chain torsion potentials for the Amber ff99sb protein force field. *Proteins*, 78:1950–1958, 2010.

[105] R. Koradi, M. Billeter, és K. Wuthrich. Molmol: a program for display and analysis of macromolecular structures. *J Mol Graph*, 14:51–55, 1996.

[106] EF. Pettersen, TD. Goddard, CC. Huang, GS. Couch, DM. Greenblatt, EC. Meng, és TE. Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*, 25:1605–1612, 2004.

[107] A. Bakan, A. Dutta, W. Mao, Y. Liu, C. Chennubhotla, T. R. Lezon, és I. Bahar. Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics*, 30:2681–2683, 2014.

[108] D. Lupyan, A. Leo-Macias, és A. Ortiz. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21:3255–3263, 2005.

[109] SF. Altschul, TL. Madden, AA. Schaffer, J. Zhang, Z. Zhang, W. Miller, és DJ. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[110] D. Sharma, B. Isaac, GPS. Raghava, és R. Ramaswamy. Spectral repeat finder: identification of repetitive sequences using fourier transormation. *Bioinformatics*, 20:1405–1412, 2004.

[111] R Development Core Team. R: A language and environment for statistical computing. http://www.r-project.org. R Foundation for Statistical Computing, Vienna, Austria.

[112] PY. Chou és GD. Fasman. Secondary structural prediction of proteins from their amino acid sequence. *Trends Biochem Sci*, 2:128–131, 1977.

[113] W. Pirovano és J. Heringa. Protein secondary structure prediction. In O. Carugo és F. Eisenhaber, editors, *Data Mining Techniques for the Life Sciences*, old.: 327–348. Humana Press, 2010.

[114] Z. Nagy, Z. Gáspári, és Á. Kovács. Accelerating a charged single alpha-helix search algorithm in protein sequences using FPGA. In R. Tetzlaff, editor, *15th International Workshop on Cellular Nanoscale Networks and their Applications: CNNA 2016*, 2016.

[115] E. Wang és AC-L. Wang. (i, i+4) ion pairs stabilize helical peptides derived from smooth muscle caldesmon. *Arch Biochem Biophys*, 329:156–162, 1996.

[116] BJ. Spink, S. Sivaramakrishnan, J. Lipfert, S. Doniach, és JA. Spudich. Long single alpha-helical tail domains bridge the gap between structure and function of myosin VI. *Nat Struct Mol Biol*, 15:591–597, 2008.

[117] J. Li, Y. Chen, Y. Deng, IC. Unarta, Q. Lu, X. Huang, és M. Zhang. Ca^{2+}-induced rigidity change of the myosin VIIa IQ motif-single α helix lever arm extension. *Structure*, 25:579–591, 2017.

[118] CS. Weirich, JP. Erzberger, és Y. Barral. The septin family of GTPases: architecture and dynamics. *Nat Rev Mol Cell Biol*, 9:478–489, 2008.

[119] AH. Fox és AI. Lamond. Paraspeckles. *Cold Spring Harb Perspect Biol*, 2:a000687, 2010.

[120] S. Nakagawa és T. Hirose. Paraspeckle nuclear bodies – useful uselessness? *Cell Mol Life Sci*, 69:3027–3036, 2012.

[121] E. Staub, P. Fiziev, A. Rosenthal, és B. Hinzmann. Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire. *Bioessays*, 26:567–581, 2004.

[122] G. Offer és R. Sessions. Computer modelling of the α-helical coiled coil: packing of side-chains in the inner core. *J Mol Biol*, 249:967–987, 1995.

[123] M.S. Sunitha, A.G. Nair, A. Charya, K. Jadhav, S. Mukopadhay, és R. Sowdhamini. Structural attributes for the recognition of weak and anomalous regions in coiled-coils of myosins and other motor proteins. *BMC Research Notes*, 5:530, 2012.

[124] M. Peckham és PJ. Knight. When a predicted coiled coil is really a single helix, in myosins and other proteins. *Soft Matter*, 5:2493–2503, 2009.

[125] M. Sickmeier et al. Disprot: the database of disordered proteins. *Nucleic Acids Res*, 35:D786–D793, 2007.

[126] PG. Higgs. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct*, 4:16, 2009.

[127] M. Di Giulio. An extension of the coevolution theory of the origin of the genetic code. *Biology Direct*, 3:37, 2008.

[128] G. Houen. Evolution of the genetic code: the nonsense, antisense, and antinonsense codes make no sense. *BioSystems*, 54:39–46, 1999.

[129] T. Oba, J. Fukushima, M. Maruyama, R. Iwamoto, és K. Ikehara. Catalytic activities of [GADV]-peptides. *Orig Life Evol Biosph*, 35:447–460, 2005.

[130] CR. Woese. On the evolution of the genetic code. *Proc Natl Acad Sci USA*, 54:1546–1552, 1965.

[131] MA. Marti-Renom, E. Capriotti, IN. Shindyalov, és PE. Bourne. Structure comparison and alignment. In J. Gu és PE. Bourne, editors, *Structural Bioinformatics, 2nd ed.*, old.: 397–418. Wiley-Blackwell, 2009.

[132] D. Hamelberg, J. Mongan, és JA. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys*, 120:11919–11929, 2004.

[133] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M.I. Jordan, és RL Dunbrack. Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comp Biol*, 6, e1000763.

[134] EF. Pettersen, TD. Goddard, CC. Huang, EC. Meng, GS. Couch, TI. Croll, J.H. Morris, és TE. Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci*, 30:70–82, 2021.

[135] GG. Krivov, MV. Shapovalov, és RL. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77:778–795, 2010.

[136] S. Neal, AM. Nip, N. Zhang, és DS. Wishart. Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N shifts. *J Biomol NMR*, 26:215–240, 2003.

[137] M. Zweckstetter és A. Bax. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J Am Chem Soc*, 122:3791–3792, 2000.

[138] O. Carugo és S. Pongor. Protein fold similarity estimated by a probabilistic approach based on Cα-Cα distance comparison. *J Mol Biol*, 315:887–898, 2002.

[139] K. Tamiola, B. Acar, és FAA. Mulder. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J Am Chem Soc*, 132:18000–18003, 2010.

[140] S. Bottaro, T. Bengtsen, és K. Lindorff-Larsen. Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. In Z. Gáspári, editor, *Structural Bioinformatics: methods and protocols*, old.: 219–240. Humana Press, 2020.

[141] M. Laskowski és M.A. Qasim. What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme-substrate complexes? *Biochim Biophys Acta*, 1477:324–337, 2000.

[142] W. Bode és R. Huber. Natural protein proteinase inhibitors and their interaction wih proteinases. *Eur J Biochem*, 204:433–451, 1991.

[143] M. Berg, J-L. Tymocko, és L. Stryer. *Biochemistry*, old.: 283–284. WH Freeman, New York, 5. kiadás, 2002.

[144] G.L. Shaw, B. Davis, J. Keeler, és A.R. Fersht. Backbone dynamics of chymotrypsin inhibitor 2: effect of breaking the active site bond and its implications for the mechanism of inhibition of serine proteases. *Biochemistry*, 34:2225–2233, 1995.

[145] J. Song és JL. Markley. Protein inhibitors of serine proteinases: role of backbone structure and dynamics in controlling the hydrolysis constant. *Biochemistry*, 42:5186–5197, 2003.

[146] B. Hess, C. Kutzner, D. van der Spoel, és E. Lindahl. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*, 4:435–447, 2008.

[147] P. Mellet, C. Boudier, Y. Mely, és JG. Bieth. Stopped flow fluorescence energy transfer measurement of the rate constants describing the reversible formation and the irreversible rearrangement elastase-α_1-proteinase inhibitor complex. *J Biol Chem*, 273:9119–9123, 1998.

[148] A. F. Ángyán és Z. Gáspári. Ensemble-based interpretations of nmr structural data to describe protein internal dynamics. *Molecules*, 18:10548–10567, 2013.

[149] S.D. Hanes. Prolyl isomerases in gene transcription. *Biochim Biophys Acta*, 1850:2017–2034, 2015.

[150] A. Matena, E. Rehic, D. Honig, B. Kamba, és P. Bayer. Structure and function of the human parvulins Pin1 and Par14/17. *Biol Chem*, 399:101–125, 2018.

[151] JW. Mueller, NM. Link, A. Matena, L. Hoppstock, A. Ruppel, P. Bayer, és W. Blakenfeldt. Crystallographic proof for an extended hydrogen-bonding network in small prolyl isomerases. *J Am Chem Soc*, 133:20096–20099, 2011.

[152] ML. Bailey, BH. Shilton, CJ. Brandl, és DW. Litchfield. The dual histidine moif in he active site of Pin1 has a structural rather han catalytic role. *Biochemistry*, 47:11481–11489, 2008.

[153] J. Wang, R. Kawasaki, J. Uewaki, A.U.R. Rashid, N. Tochio, és S. Tae. Dynamic allostery

modulates catalytic activity by modifying the hydrogen bonding network in the catalytic site of human pin1. *Molecules*, 22:992, 2017.

[154] RA. Friesner, JL. Banks, RB. Murphy, TA. Halgren, JJ. Klicic, DT. Mainz, MP. Repasky, EH. Knoll, M. Shelley, JK. Perry, DE. Shaw, P. Francis, és PS. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 47:1739–1749, 2004.

[155] NA. Rodzli, MP. Lockhart-Cairns, CW. Levy, J. Chipperfield, L. Bird, C. Baldock, és SM. Prince. The dual PDZ domain from Postsynaptic Density Protein 95 forms a scaffold with peptide ligand. *Biophys J*, 119:667–689, 2020.

[156] GE Crooks, G. Hon, J-M. Chandonia, és SE Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14:1188–1190, 2004.

[157] E. Galvanetto, MT. Ivanovic, A. Chowdhury, A. Sottini, MF. Nüesch, D. Nettels, RB. Best, és B. Schuler. Extreme dynamics in a biomolecular condensate. *Nature*, 619:876–883, 2023.

[158] Z. Gaspari és L. Nyitray. Coiled coils as possible models of protein structure evolution. *Biomol Concepts*, 2:199–210, 2011.

[159] D. Simm, K. Hatje, S. Waack, és M. Kollmar. Critical assessment of coiled-coil predictions based on protein structure data. *Sci Rep*, 11:12439, 2021.

[160] L. Kalmar, V. Acs, D. Silhavy, és P. Tompa. Long-range interactions in nonsense-mediated mRNA decay are mediated by intrinsically disordered protein regions. *J Mol Biol*, 424:125–131, 2012.

[161] G. Tesei, AI. Trolle, N. Jonsson, J. Betz, FE. Knudsen, F. Pesce, KE. Johansson, és K. Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature, doi:10.1038/s41586-023-07004-5*, 2024.

[162] P. Mier, L. Paladin, S. Tamana, S. Petrosian, B. Hajdu-Soltész, A. Urbanek, A. Gruca, D. Plewczynski, M. Grynberg, P. Bernadó, Z. Gáspári, CA. Ouzounis, VJ. Promponas, AV. Kajava, JM. Hancock, SCE. Tosatto, Z. Dosztanyi, és MA. Andrade-Navarro. Disentangling the complexity of low complexity proteins. *Brief Bioinform*, 21:458–472, 2020.

[163] V. Tretyachenko et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep*, 7:15449, 2017.

[164] E. Bornberg-Bauer, K. Hlouchova, és A. Lange. Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol*, 68:175–183, 2021.

[165] C. Schlötterer. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet*, 31:215–219, 2015.

[166] G. Tóth, Z. Gáspári, és J. Jurka. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*, 10:967–981, 2000.

[167] Z. Gáspári, CP. Ortutay, és G. Tóth. Divergent microsatellite evolution in the human and chimpanzee lineages. *FEBS Lett*, 581:2523–2526, 2005.

[168] Á. Kun. *Evolúcióbiológia*. Typotex, Budapest, 2017.

[169] L. Senicourt, A. le Maire, F. Allemand, JE. Carvalho, L. Guee, P. Germain, M. Schubert, P. Bernadó, W. Bourguet, és N. Sibille. Structural insights into the interaction of the intrinsically disordered co-activator TIF2 with retinoic acid receptor heterodimer (RXR/RAR). *J Mol Biol*, 9:166899, 2021.

[170] JMC. Teixeira, ZH. Liu, A. Namini, J. Li, RM. Vernon, M. Krzeminski, AA. Shamandy, O. Zhang, M. Haghighatlari, L. Yu, T. Head-Gordon, és JD. Forman-Kay. IDPConformerGenerator: A flexible software suite for sampling the conformational space of disordered protein states. *J Phys Chem A*, 126:5985–6003, 2022.

[171] T. Lazar et al. Ped in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res*, 49:D404–D411, 2021.

[172] R. B. Best, K. Lindorff-Larsen, MA. DePristo, és M. Vendruscolo. Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci USA*, 103:1ö901–10906, 2006.

[173] GD. Friedland, N-A. Lakomek, C. Griesinger, J. Meiler, és T. Kortemme. A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comp Biol*, 5:e1000393, 2009.

[174] JA. Velázquez-Muriel, M. Rueda, I. Cuesta, A. Pascual-Montano, M. Orozco, és J-M. Carazo. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct Biol*, 9:6, 2009.

[175] A. Kumawat és S. Chakrabarty. Protonation-induced dynamic allostery in PDZ domain: Evidence of perturbation-independent universal response network. *J Phys Chem lett*, 11:9026–9031, 2020.

[176] N. Tokuriki és DS. Tawfik. Protein dynamism and evolvability. *Science*, 324:203–207, 2009.

[177] P. Campitelli, T. Modi, S. Kumar, és SB. Ozkan. The role of conformational dynamics and allostery in modulating protein evolution. *Annu Rev Biophys*, 49:267–288, 2020.

[178] F. Sala, D. Engelberger, HS. Mchaourab, és J. Meiler. Modeling conformational states of proteins with AlphaFold. *Curr Opin Struct Biol*, 81:102645, 2023.