

**Az *in vivo* fehérjeaggregáció útvonalait befolyásoló tényezők
vizsgálata *Escherichia coliban***

Györkei Ádám

Ph.D. értekezés

**Témavezető: Dr. Papp Balázs és Dr. Kintses Bálint
Biológia Doktori Iskola
ELKH Szegedi Biológiai Kutatóközpont, Biokémia Intézet
Szegedi Tudományegyetem, Természettudományi és
Informatikai Kar
Szeged, 2021**

Tartalomjegyzék

1. Bevezetés.....	
1.1. A fehérjeaggregáció.....	
1.2. A fehérje aggregátumok típusai.....	
1.2.1. A funkcionális és patológiás aggregátumok.....	
1.2.2. Amiloid, inklúziós test, fázisszeparált aggregátum és vizsgálati módszereik.....	
1.2.2.1. Az egyes aggregátumtípusok <i>in vivo</i> elkülönítése.....	
1.2.3. A „Life on the edge” hipotézis és a szupersaturáció.....	
1.3. A fehérjeaggregációt befolyásoló ismert tényezők.....	
1.3.1. Fizikai-kémiai paraméterek.....	
1.3.2. Szerkezet és funkció.....	
1.3.3. A fehérjerendezetlenség.....	
1.3.3.1. A fehérjerendezetlenség <i>in silico</i> predikciója.....	
1.4. Posztfolding és prefolding aggregációt befolyásoló tényezők.....	
1.4.1. A fehérje folding.....	
1.4.2. Az energiatájkép és a fehérje aggregáció.....	
1.4.2.1. A fehérje folding <i>in silico</i> predikciós módszerei.....	
1.4.2.2. A fehérje aggregáció predikciója.....	
1.4.3. A chaperonok.....	
1.5. Célkitűzések.....	
2. Anyagok és módszerek.....	
2.1. Az ASKA expressziós könyvtár használata.....	
2.2. Nagy áteresztőképességű fluoreszcens egysejt-mikroszkópia.....	
2.3. A fehérjeexpresszió és az <i>in vivo</i> aggregáció kísérletes ellenőrzése.....	
2.4. A fehérje- és génjellemzőket tartalmazó adatsor összeállítása.....	

3. Eredmények.....	
3.1. Két aggregációs fenotípus különböztethető meg overexpresszió esetén <i>E. coliban</i>	
3.2. A nagy áteresztőképességű mikroszkópiás módszer eredményeinek validációja.....	
3.3. A fluoreszcens aggregátum fenotípusa domináns citoplazmatikus fehérjék esetén, míg a sötét aggregátumé a membránfehérjéknél.....	
3.3.1. A fázisszeparáció nem bír döntő jelentőséggel az <i>in vivo</i> aggregáció folyamatában <i>E. coliban</i>	
3.4. A sötét aggregátumok kialakulását meghatározó tényezők.....	
3.5. A két aggregációs útvonal közös jellemzői és különbségük a szolubilis fehérjétől.....	
4. Diskusszió.....	
5. Összegzés és kitekintés.....	
6. Köszönetnyilvánítás.....	
7. Hivatkozások.....	
8. Összefoglaló.....	
9. Summary.....	
10. Függelékek.....	

1. Bevezetés

1.1. A fehérjeaggregáció

Az emberiség már az idők kezdete óta használ sejtenyészeteket különböző termékek előállításához (Arranz-Otaegui és mtsai., 2018). Azonban egészen a XX. századik kellett várni, amíg a sejt kultúrák széles körben elterjedtek az ipari és egészségügyi szempontból fontos anyagok gyártásában. Ennek egyik legfontosabb eseménye volt Paulescu 1922-es felfedezése, miszerint az inzulin hatékonyan csökkenti a megnövekedett vércukorszintet, ezáltal a cukorbetegség ha nem is gyógyíthatóvá, de hosszabb távon kezelhetővé vált. Érthető, hogy ezt követően hatalmasra nőtt az igény az inzulin iránt, hiszen élethosszig tartó terápiát igényeltek a betegek, ugyanakkor egészen az 1970-es évekig csupán a szarvasmarhák hasnyálmirigyéből tudták előállítani ezt a gyógyszert. Egészen 1979-ig kellett várni, amikor a Genentech kutatóinak David Goeddel vezetésével sikerült első ízben *Escherichia coli*-ban rekombináns fehérjeexpresszióval inzulint termelni (Goeddel és mtsai., 1979). Ez mérföldkőnek számított a modern biotechnológiai ipar létrejöttében, amelyet számos egyéb biológiai és orvosi szempontból fontos fehérje termelésére tett kísérlet követett, mint amilyenek a vakcinában használt fehérjék (Valenzuela és mtsai., 1982) vagy a monoklonális antitestek (Köhler és Milstein, 1975). Ugyanakkor egy rendkívül fontos gyakorlati probléma is evidenssé vált: bizonyos fehérjék nagy mennyiségű termelése csak komoly nehézségek árán valósítható meg, köszönhetően magas aggregációs hajlandóságuknak. A fehérjék a termeléstől a fagyasztás-kiolvasztáson és liofilizációs szárításon át a tárolás bármely lépésében is kicsapódhattak (Brader, 2014), ezáltal nem csak hatásukat veszítették, de bizonyos esetekben még immunológiai mellékhatásokat is produkáltak (Rosenberg, 2006). Ez a probléma azóta is sújtja a biotechnológiai ipart, hatalmas termelésekiesést okozva évről évre, aminek köszönhetően az aggregációt okozó fehérje tulajdonságok intenzív kutatások tárgyát képezik. Az elmúlt öt évtized során nagy számú fehérje került termelésre ipari körülmények között, olyanok, amelyek szekvenciája merőben eltérő volt egymástól, és olyanok is, amelyek csupán egy-egy aminosavban különböztek, a készítmények farmakokinetikai tulajdonságainak

optimalizálása céljából. Rá kellett jönnünk, hogy a legkisebb módosítások, akár egy aminosav megváltoztatása is, komolyan befolyásolják a fehérjék viselkedését (Jenkins és mtsai., 1995; Pédelacq és mtsai., 2002; Timson és Reece, 2003; Wetzel és mtsai., 1991). Így számos fehérje tulajdonság aggregációra gyakorolt hatását sikerült megismernünk, legtöbbször azonban *in vitro* körülmények között, amely csak korlátozott mértékben általánosítható sejten belüli viszonyokra. Emellett ismereteink főleg olyan fehérjékre szorítkoznak, amelyek biotechnológiai jelentőséggel bírnak, nem szisztematikusan ölelik fel egy proteom minden tagját, megnehezítve az általános szabályok felállítását. Így továbbra is hatalmas erőfeszítést igényel egy-egy fehérje ipari termelésének optimalizálása, mert nem áll rendelkezésünkre kellő mennyiségű, összevethető, szisztematikus *in vivo* adat a fehérjék tulajdonságainak és aggregációs hajlandóságuknak összefüggéséről.

Egészen az elmúlt évtizedig az uralkodó álláspont szerint az aggregátumok pusztán inaktív fehérjékből állnak, amelyek nem érik el natív konformációjukat mielőtt aggregálódnak (Baneyx és Mujacic, 2004). Az aggregált fehérjék a gyártási folyamat során vagy eldobásra kerülnek, vagy csak nagyon drága és körülményes unfolding–refolding módszerek segítségével hozhatók megfelelő natív állapotba (Ben-Zvi és mtsai., 2004; Yamaguchi és Miyazaki, 2014). Azonban újabb eredmények rávilágítottak, hogy a fehérje aggregáció lejátszódhat a megfelelő foldingon átesett fehérjékkel is (Calloni és mtsai., 2012; Chiti és Dobson, 2009; Groot és mtsai., 2009; Plakoutsi és mtsai., 2004), sőt, aggregátumok funkcionális fehérjéket is tartalmazhatnak. A *Clostridium thermocellum* endonukleáz D fehérje *Escherichia coliban* történő overexpressziója során létrejövő aggregátumok jelentős, a natív proteinekkal összevethető aktivitást mutattak (Kostas és mtsai., 1991). Hasonlóan, a béta-galaktozidáz vizsgálata során kiderült, hogy az aggregátumok natív szerkezetű, aktív fehérjéket is tartalmaznak (Worrall és Goss, 1989). Ezeket az eredményeket először kontaminációnak tudták be, majd a fehérjék különös, egyedi viselkedésének tulajdonították, ezért ezek a típusú aggregátumok a nem-klasszikus aggregátum nevet (nonclassical inclusion bodies) kapták. Idővel azonban számos más példa is bizonyította, hogy a jelenség gyakoribb, mint hittük (Simona és mtsai., 2005). Ez hatalmas jelentőséggel bír, mert az aggregátumok számos előnyös tulajdonsággal is rendelkeznek az oldott fehérjékhez képest: az aggregátumok akár 90%-ban a termelt fehérjéből állnak, amelyek különböző felületeihez kikötve kontamináció nélkül képesek

hosszú ideig ellátni funkciójukat (Nahalka és mtsai., 2008; Nahálka és mtsai., 2009). Könnyebben izolálhatóak mechanikai stabilitásuknak köszönhetően (Georgiou és Valax, 1999), valamint hosszú eltarthatóságot mutatnak jelentősebb degradáció nélkül (García-Fruitós és mtsai., 2009). Ellenállnak a fagyasztás–kiolvasztási folyamatoknak az aktivitás elvesztése nélkül (Nahálka, 2008). A biológiailag aktív aggregátumok felhasználási lehetőségei közé tartoznak a biotechnológiai ipar számára az immobilizált katalizátorok (Wu és mtsai., 2011), valamint egyes gyógyszermolekulák vivőanyagait képző mátrixok (Villaverde és mtsai., 2012). Ráadásul eddigi ismereteink szerint nincsenek citotoxikus hatással az emlőssejtekre sem (Díez-Gil és mtsai., 2010; García-Fruitós és mtsai., 2009). Szükséges tehát a fehérjeaggregáció kérdésének szisztematikus, *in vivo* vizsgálata olyan módszerrel, amely megbízhatóan képes azonosítani a funkcionális és inaktív aggregátumokat és elkülöníteni őket a szolubilis proteinektől. Így felmérhető, hogy milyen gyakoriak ezek a nem klasszikus aggregátumok, annak eldöntése érdekében, hogy biotechnológiai hasznosíthatóságukról átfogóbb képet kapjunk. Ezzel együtt információkkal szolgálhat arról is, hogy milyen fehérje tulajdonságok esetén alakulnak ki ezek a funkcionális aggregátumok és hogy előidézhető-e a képződésük, szemben a funkció nélküli típussal.

Ebben a fejezetben bemutatom a fehérje aggregátumok típusait, különös tekintettel arra, hogy melyek tartalmazhatnak aktivitással bíró fehérjéket. Megvizsgálom, hogy milyen fehérje tulajdonságokról ismert, hogy befolyásolják a fehérje aggregáció folyamatát. Először sorra veszem azokat a jellemzőket, amelyek mind a funkcionális, mind az inaktív aggregátumok létrejöttében játszanak szerepet, majd azokat, amelyek meghatározzák, hogy melyik aggregációs útvonalra lép egy polipeptid. Ahhoz, hogy egy fehérje a fiziológias funkcióját végrehajtsa, szükséges a natív konformáció felvétele, ezért külön hangsúlyt fektetek a fehérje folding szerepére és a szerkezetre. Végül pedig bemutatom a disszertáció célkitűzéseit.

1.2. A fehérje aggregátumok típusai

A fehérje aggregátumokat számos szempont szerint tudjuk jellemezni, amelyek egymással átfedő kategóriákat eredményeznek. Ugyanakkor mégis szükséges bizonyos felosztásokat tenni egyes tulajdonságok mentén, mert az aggregációs típusok összefüggnek az aggregátumok általunk vizsgálat biotechnológiai szempontból fontos tulajdonságaival, mint a megfelelő szerkezet és aktivitás, és rávilágítanak, hogy mely fehérjetulajdonságok vizsgálata segíthet elkülöníteni a két csoportot.

1.2.1. A funkcionális és patológiás aggregátumok

Az aggregátumok csoportosításának első szempontja a funkcionalitás, amely alapján patológiás (nem funkcionális) és funkcionális típusokat különböztetünk meg. Annak ellenére, hogy általában patológiás hatásokkal hozzuk összefüggésbe a fehérjeaggregátumokat, számos esetben fiziológiás funkcióval bírnak és normál körülmények között is megtalálhatóak a sejtekben. Ez nem teljesen azonos a biotechnológiai szempontból fontos, aktivitással bíró aggregátumokkal. A natív körülmények között funkcionális aggregátumok gyakran strukturális szerepet töltenek be, enzimaktivitással ritkán bírnak. Ilyen funkciók lehetnek a biofilmképzés (Chapman és mtsai., 2010), a pílusképzés (Hess és mtsai., 2007), a transzkripció (Yuan és Hochschild, 2017) és a transzláció (Wickner, 1994) szabályozása, valamint az antimikrobiális (Gour és mtsai., 2016) és antifungális (Garvey és mtsai., 2013) védekezés. Az emberi szervezetben is található funkcionális aggregáció, például a Pmel17 fehérje játszik fontos szerepet a melanin szintézise során keletkező toxikus intermedierek eltávolításában (Fowler és mtsai., 2006). A fehérjék oligomerizációja is egyfajta aggregációs folyamatnak tekinthető, amely azonban funkcionális komplexeket hoz létre és erős kontroll alatt megy végbe anélkül, hogy makroszkopikus struktúrákat hozna létre. Bizonyos oligomerek ezzel szemben a magasabb rendű aggregátumok prekursorai, ráadásul erősen toxikus hatásúak (Ferreira és mtsai., 2007). Ilyen értelemben elmosódik a határ a funkcionális és nem funkcionális aggregátumok között: a magasabb rendű amiloidok a toxikus oligomerek detoxifikációjára szolgálnak, bár a fehérje az eredeti funkcióját már nem látja el ebben a formában (Caughey és Lansbury, 2003). Számos fehérjét ismerünk, amely patológiás aggregátumokat képez, ezek leggyakrabban neurodegeneratív kórképek okozói, mint

például a Parkinson-kór, a Huntington-kór, az ALS (amiotrófiás laterálszklerózis). Azonban vannak bizonyítékok arra is, hogy a fehérje aggregáció sokkal több fehérjét és szervrendszert érinthet és nem csak neurológiai manifesztációjú betegségek okozója. Ilyen a II-es típusú cukorbetegség, amely bizonyos formáit a Langerhans-szigetek sejtjeiben felhalmozódó amiloidok (Hull és mtsai., 2004) okozzák, vagy a különböző miopátiákért felelős aggregátumok (Sharma és Goebel, 2005). A patológiás aggregátumot alkotó fehérjék egy kiemelt csoportját képezik a prionok, amelyek nem csak toxikus fehérjemolekulák, hanem fertőző ágensek is. Ezek a fehérjék kétféle stabil konformáció felvételére képesek, azonban a prion forma képes az egészséges konformációt a sajátjához hasonlóná tenni, amely segítségével propagálódik. A fehérje oligomerek toxicitása és a prionok két különböző stabil szerkezete nagy hangsúlyt helyez a fehérjék konformációira és az egymással létrehozott interakciókra az aggregáció folyamata szempontjából.

1.2.2. Amiloid, inklúziós test, fázisszeparált aggregátum és vizsgálati módszereik

A fehérjeaggregátumoknak szerkezet szempontjából számos eltérő szerkezeti típusa ismert, amelyek egymástól nem mindig választhatók el élesen. Tovább bonyolítja a helyzetet, hogy az irodalomban az egyes szerzők a kifejezéseket gyakran szinonimákként használják különböző kontextusokban. A funkcionális aggregátumok vizsgálata szempontjából így a baktériumokban releváns típusok említésére kell szorítkoznunk. A kategorizálás fő szempontja, hogy az aggregátumban található-e natív szerkezetű fehérjék, tehát a fehérjék átestek-e a folding folyamatán az aggregációt megelőzően. Így az aggregátumokat két fő típusra oszthatjuk, mint pre- és posztfolding aggregátumok, az előbbiben a fehérjék a foldingot megelőzően aggregálódnak, míg az utóbbi esetben azt követően. A prefolding aggregátumokra tekinthetünk az amiloid aggregátumok analójaként, azonban ide tartoznak azok az esetek is, amelyek során unfolding következtében veszítik el a fehérjék a natív strukturájukat és kerülnek az aggregátumba. Ismertek olyan esetek is, amikor a posztfolding aggregátumok elveszítik natív strukturájukat az aggregátumokon belül (Elia és mtsai., 2017), azonban egyelőre nem bizonyított, hogy gyakori jelenség lenne, továbbá vannak fehérjék, amelyek az aggregátumokba lépést követően érik el natív konformációjukat (Gonzalez-Montalban és

mtsai., 2008). Fontos megjegyezni, hogy a két aggregátumtípus nem éles határral válik el egymásból, minden jel arra mutat, hogy a posztfolding aggregátumok tartalmaznak amiloidszerű struktúrákat bizonyos mennyiségben, így a két típust egy skála végpontjaiként foghatjuk fel. A posztfolding aggregátumokat az irodalomban gyakran nevezik inklúziós testnek, ugyanakkor számos esetben ezzel a kifejezéssel a bakteriális aggregátumok gyűjtőneveként találkozhatunk. Létezik egy harmadik típusú, fázisszeparált aggregátum, azonban baktériumok esetén ez a típus kevésbé ismert, mivel kis számú bakteriális fehérjéről ismert, hogy fázisszeparáción esik át, és a fázisszeparált aggregátumok szerkezete sokfélebb is (lásd később). Az irodalomban tapasztalt fogalomzavarok elkerülése érdekében áttekintjük az egyes aggregátumtípusokat szerkezet és felépítés szempontjából, hiszen ezek a tulajdonságok a funkcióra is kihatással vannak.

A prefolding aggregátumokban a fehérjék nem rendelkeznek natív struktúrával, denaturált vagy hibásan feltekeredett (misfolded) láncokként csapódnak ki. Natív szerkezet hiányában ezért a hidrofób oldalláncok közti interakciók dominálnak, amelyek legjellegzetesebb megjelenési formája az amiloid szerkezet. Ezek fibrilláris képletek, amelyek nagy mennyiségű béta-lemez másodlagos szerkezetet tartalmaznak. A béta-lemezek keresztirányú elrendezésűek és merőlegesen helyezkednek el a fibrillumok hossz tengelyére nézve. Az egyes beépülő béta-lemezek közt létrejövő kapcsolatok hatására az amiloid szál a tengelye mentén hosszabbodik, elvben végtelen hosszúságig. Kimutattak olyan amiloid struktúrát is, amelyben a amfipatikus alfa-hélixek alkották a fonalakat létrehozó lemezeket (Tayeb-Fligelman és mtsai., 2019), ez azonban nem tekinthető általánosnak. Ahogy az amiloid struktúrák egyre hosszabbodnak, az egyes szálak egymásra tekeredve fonalakat hoznak létre. A fonalak egymásba akadnak és rendezetlen gubancokat alkothatnak. Az amiloid aggregátumok nagyfokú rendezettségüknek köszönhetően rendkívül stabilak, ellenállnak a denaturációnak, és bennük a natív szerkezetű fehérjék nem lelhetők fel a tömött szerkezet és az erős intermolekuláris kapcsolatok miatt. Korábbi feltételezések szerint az amiloid aggregátumokat csak a fehérjék egy kis része képes létrehozni, azonban később kiderült, hogy nemcsak hogy betegségekhez nem köthető fehérjék (Fändrich és mtsai., 2001), hanem a legtöbb fehérje (Chiti és mtsai., 1999), sőt egyes poliaminosavak (Dobson, 2002)

is képesek amiloidok létrehozására. Bár az amiloid aggregátumokról szerzett legtöbb információnk a humán betegséget okozó fehérjék vizsgálata során került birtokunkba, egyes bakteriális fehérjék is alkotnak amiloidokat, ahol a nagy fokú rendezettség előnyös, mint amilyenek a pílusok vagy a biofilmet képző amiloidok.

A baktériumokban overexpresszió hatására létrejövő aggregátumok irodalomban is gyakran használt neve az inklúziós test. Ezt gyűjtőfogalomként használjuk, általában gömb vagy pálcika alakú képletek (Betton és Hofnung, 1996; Zhu és Yu, 2008), amelyek a sejtek lízise után amorf külsőt öltenek. Akkor jönnek létre, ha egy fehérje expressziója meghaladja az összexpresszió 2,5%-át (Bowden és Georgiou, 1990; Georgiou és mtsai., 1986), tehát kialakulásuk egyértelműen koncentrációfüggő folyamat (Georgiou és mtsai., 1994). Az inklúziós test kifejezést az aggregátumban található fehérjék szerkezetétől függetlenül is használják az irodalomban, itt azonban a nem amiloid típusú, natív szerkezetű fehérjéket is tartalmazó aggregátumokra értjük, ezért a posztfolding aggregátum kifejezést használjuk, amely egyértelműen utal képződésük módjára. Nagy biotechnológiai potenciáljuk ellenére keveset tudunk azonban a pontos szerkezetükről és a létrejöttük mechanizmusáról. Jelentős részük ellenálló az oldószerekkel szemben, csak költség- és munkaigényes folyamat során lehet renaturálni őket. Léteznek azonban „pehelyszerű”, lazább szerkezetű inklúziós testek is, amelyek enyhébb körülmények között is könnyedén feloldódnak (Hart és Bailey, 1990). Korábban nem állt rendelkezésre bizonyíték arra nézve, hogy a bakteriális inklúziós testek natív szerkezettel bíró fehérjéket is tartalmazhatnak, az elterjedt nézet szerint hibásan feltekeredett vagy denaturált fehérjék alkotják őket nemspecifikus interakciók segítségével. Erre engedett következtetni az is, hogy a fehérjék hidrofób felszínét érintő mutációk csökkentették az aggregátumok (Nieba és mtsai., 1997) létrejöttét. A kanonikus elképzelés szerint a fehérje folding során létrejövő intermedierek némelyike erős aggregációs hajlandósággal bír a hidrofób felszíneinek köszönhetően. Újabb eredmények azonban az mutatták, hogy az aggregátumokban található fehérjék jelentős mennyiségű natív szerű másodlagos (Fink, 1998) szerkezettel is rendelkezhetnek, sőt a folding folyamata után is aggregálódhatnak (Calloni és mtsai., 2012; Chiti és Dobson, 2009; Groot és mtsai., 2009; Plakoutsi és mtsai., 2004). Függetlenül a fehérje natív szerkezetétől, jelentős mennyiség új béta-lemez struktúrák jönnek létre az inklúziós testekben (Fink, 1998). Ez arra utal, hogy az

aggregátumok létrejöttében fontos szerepet játszanak a béta-lemezek közt kialakuló interakciók. Az egyes aggregátumok esetén a natív szerű és béta-lemez struktúrák aránya jelentősen változik, ez meghatározza, hogy az egyes aggregátumok milyen könnyen oldhatók és hogy mutatnak-e a bennük található fehérjék aktivitást. Az utóbbi két évtizedben azonban egyre növekvő számú fehérjéről bebizonyosodott, hogy natív vagy natív szerű szerkezetét megőrzi az aggregáció folyamata után is (Betton és Hofnung, 1996; Petermel és Komel, 2011; Rinas és mtsai., 2017; Villaverde és Carrio, 2001; Worrall és Goss, 1989). Ez a típus egyfajta dinamikus egyensúlyban van a környezetével, izolálás után híg oldószerben gyakran külső beavatkozás nélkül feloldódik. Látható tehát, hogy az inklúziós testek és az amiloidok szerkezete között mutatkoznak kisebb-nagyobb hasonlóságok az egyes fehérjék tulajdonságainak függvényében. A helyzetet tovább bonyolítja, hogy egyes inklúziós testekben megtalálhatóak amiloid struktúrák, mintegy hálózatot formálnak, amelyekbe a natív szerű szerkezetű fehérjék csapdázódnak (Groot és mtsai., 2009; Petermel és Komel, 2011; Rinas és mtsai., 2017). Továbbá az egyes natív szerkezetű fehérjék képesek amiloid szerkezetet felvenni az inklúziós test belsejében az idő előrehaladtával (Chiti és Dobson, 2009; Elia és mtsai., 2017). Mivel ez a típus egyelőre alig került kiaknázásra biotechnológiai szempontból, rendkívül fontos annak megismerése, hogy pontosan milyen fehérje tulajdonságok vezetnek a kialakulásukhoz, és hogy képesek a bennük lévő fehérjék elérni a funkcióhoz szükséges natív struktúrájukat.

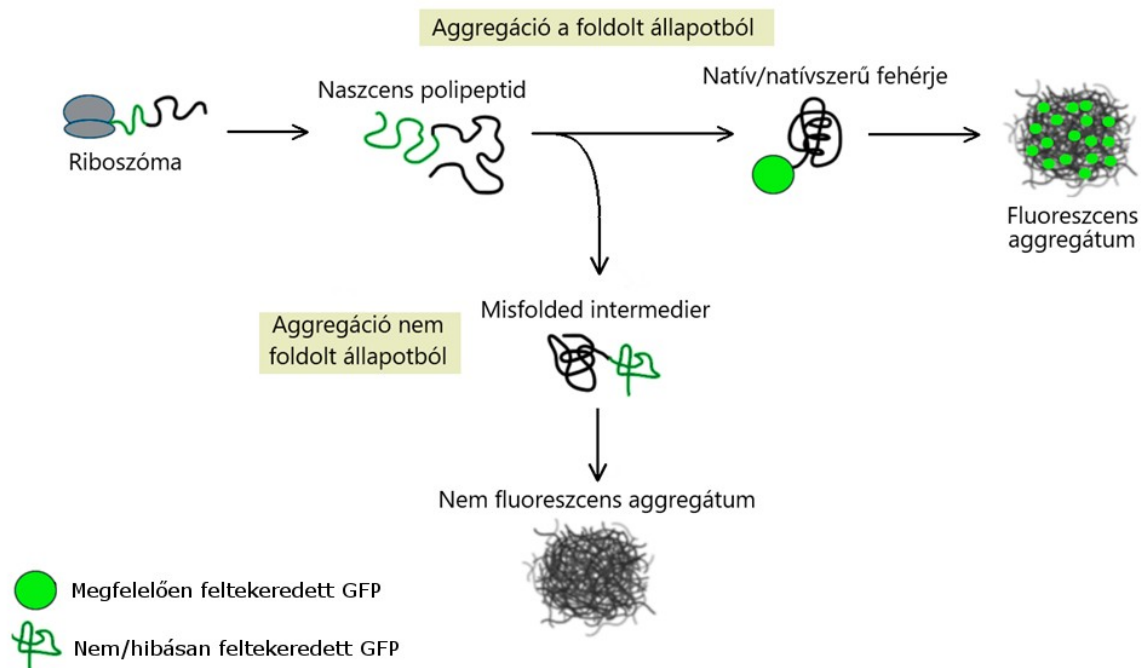
Az elmúlt évtizedben az előző két típustól jelentősen eltérő, harmadik fajta aggregátumtípus is felfedezésre került, a fázisszeparált aggregátum. Ezek főleg eukarióta sejtekben fordulnak elő, baktériumokban jelenleg kevesebb, mint egy tucat fázisszeparációra képes proteint ismerünk. Ugyan az eukarióta sejtekben előforduló membrán nélküli sejtalkotók, mint a nukleólusz, a Cajal-testek, P-, vagy PML-testek régóta ismertek voltak, azonban csak nemrégiben kezdtünk aggregátumként gondolni rájuk és vizsgálni őket. Ebben fontos szerepet játszik, hogy míg az amiloidok és az inklúziós testek szilárd halmazállapotúak és meglehetősen stabilitással bírnak, a fázisszeparált aggregátumok folyadékszerű tulajdonságokkal rendelkeznek és rendkívül dinamikusan jönnek létre és oldódnak föl. Továbbá több mint két tucat eltérő fázisszeparált aggregátumot ismerünk, amelyek tulajdonságai rendkívül eltérőek, ezért annak megértése, hogy hasonló folyamatok eredményeképp jönnek létre, nagy nehézségekkel szembesült. A Cajal-testek

jellemzése (Handwerger és mtsai., 2005), miszerint „félfolyékony gömbök a félfolyékony nukleoplazmában” volt az első tényleges felismerése ennek a speciális aggregációs típusnak. A P-granulumok vizsgálata (Hoegge és mtsai., 2009) bizonyította, hogy folyadékszerű képletekről van szó és fázisszeparációval jönnek létre. Ez egy olyan folyamat, amely során a szuperszaturált, tehát oldhatósági küszöbnél magasabb koncentrációban jelen lévő fehérje spontán módon szeparálódik egy tömény és egy híg fázisra, melyek egymással stabilan együtt fordulnak elő. Az így létrejövő cseppek nem egyszerű folyadékok a kifejezés kémiai értelmében (van der Waals folyadékok), amelyekben a molekulák gömbként viselkednek, hanem keresztkötéseket tartalmazó speciális struktúrák. Ez lehetővé teszi, hogy a cseppek belsejében meghatározott térbeli elrendezésben legyenek az alkotók. A fázisszeparált aggregátumok kialakításában multivalens proteinek vesznek részt, amelyek több másik fehérjével, vagy RNS molekulával képesek kapcsolatot kialakítani. Ezek lehetnek natív struktúrájú, foldingon átesett fehérjék, amelyek oligomerekké állhatnak össze, megfelelő szerkezetű domének, amiket flexibilis linker régiók kapcsolnak össze, valamint rendezetlen és alacsony komplexitású fehérje régiók, amelyek egyfajta vázként számos másik komponenshez kapcsolódnak (Fong és mtsai., 2013; Lyon és mtsai., 2005). A fázisszeparált struktúrák gyakran tartalmaznak nukleinsavakat, leginkább RNS molekulákat, köszönhetően az őket alkotó fehérjék RNS-kötő doménjeinek. Tekintve, hogy milyen sokféle fázisszeparált aggregátum ismert, nehéz általános szerkezeti jellemzést adni róluk. Egyes fázisszeparátumok amiloidszerű (Boke és mtsai., 2016) interakciók hatására jönnek létre, stabil, nem oldható magok körül, míg mások fontos elemei a fehérje hurkok (Woodruff és mtsai., 2017). Egyes fázisszeparált cseppekben a fehérjék rendezetlenül találhatók (Woodruff és mtsai., 2017), monomerhez hasonló állapotban. Bizonyos fázisszeparált struktúrák továbbá érési folyamaton esnek át, amelyben belső szerkezetük is változik, gyakran a patológiás folyamatokból ismert amiloidszerű állapot irányába (Molliex és mtsai., 2015).

1.2.2.1. Az egyes aggregátumtípusok *in vivo* elkülönítése

A különböző aggregátumtípusok elkülönítése és jellemzése korábban főleg *in vitro* vizsgálatok során történt. Emiatt kevés információval rendelkezünk az aggregátumok *in vivo* állapotáról és szerkezetéről, valamint arról, hogy az egyes fehérjék milyen

gyakorisággal veheik föl az egyik vagy másik aggregációs formát. *In vivo* körülmények között az alkalmazható módszerek köre jelentősen beszűkül, a módszerek invazív volta miatt. Megoldást jelenthetnek többek között a cirkuláris dikroizmus spektroszkópia és a szilárdtest NMR spektroszkópia, amelyek képesek az aggregátumok belső szerkezetének közvetlen vizsgálatára. Azonban ezek a módszerek rendkívül munka- és eszközigényesek, ellehetetlenítve a szisztematikus vizsgálatokat. Emiatt előtérbe kerülnek az indirekt módszerek, amelyek közül a legfontosabbak a riporter konstrukciók használata. Ezek nagy segítségünkre vannak abban, hogy az aggregátumokban lévő fehérjék folding állapotáról információkat szerezzünk a sejtek megbontása nélkül. Az egyes riporterek lehetnek a különböző enzimek, mint a béta-galaktozidáz, valamint az egyik leggyakoribb a zöld fluoreszcens fehérje (GFP), amelyet régóta használunk a fehérjetermelés (Chalfie és mtsai., 1992; Wu és mtsai., 2000) és szolubilitás vizsgálatára is (Listwan és mtsai., 2009; Waldo és mtsai., 1999). Bizonyos esetekben előfordulhat, hogy az amiloid típusú aggregátumokban lévő fehérjékhez kapcsolódó GFP ad fluoreszcens jelet (Kim és Hecht, 2005), azonban számos bizonyíték áll rendelkezésünkre arról, hogy a zöld fluoreszcens fehérje képes kvalitatív információval szolgálni az aggregátumokban lévő fehérjék folding állapotáról (Arslan és Chakrabartty, 2009; Bock és mtsai., 2012; Pédelacq és mtsai., 2002), fluoreszcens jelet csak a natív szerkezettel bíró fehérjéket tartalmazó aggregátumokban mutatva. Sikerrel alkalmazták ezeket a riportereket ozmolitok (Shmueli és mtsai., 2013) és chaperonok (Zhang és mtsai., 2005) foldingra gyakorolt hatásának vizsgálatában, *in vivo* körülmények között (Gregoire és Kwon, 2012), sőt, az aggregátumok belsejében (Bakholdina és mtsai., 2021) fellelhető fehérjék esetén is. Az 1. ábra bemutatja egy ilyen konstrukció alkalmazását a folding és az aggregáció vizsgálatakor, valamint a pre- és posztfolding aggregátumok létrejöttének folyamatát.



1. ábra. A fehérjeaggregáció két fő útvonala. Overexpresszió során a naszcens polipeptidek nem natív interakciókon keresztül (felső útvonal) alkothatnak amiloidszerű, hibásan feltekeredett (misfolded) aggregátumokat (fekete gömb), úgy, hogy a fehérjék nem érik el a natív szerkezetüket (Groot és mtsai., 2009). Alternatív esetben a fehérjék aggregálódhatnak natív vagy natív szerű szerkezetük felvételét követően is. Ebben az esetben az aggregátumban két egyszerre jelen lévő fehérjepopuláció található: amiloidszerű és natív szerkezetű fehérjék (fekete gömb zöld pöttyökkel) (Rinas és mtsai., 2017). Megjegyzendő, hogy a folding folyamatát követő aggregátumokban található fehérjék idővel elveszthetik natív struktúrájukat lokális unfolding következtében (Elia és mtsai., 2017).

1.2.3. A „Life on the edge” hipotézis és a szupersaturáció

Régóta ismert, hogy az abundáns fehérjék erősebb stabilizáló szelekció alatt állnak az alacsonyabban expresszált fehérjékhez képest (Gout és mtsai., 2010; Pál és mtsai., 2001). Ennek számos oka lehetséges, a legelfogadottabb hipotézis a hibás fehérjefeltekeredés okozta citotoxicitás elkerülése (Geiler-samerotte és mtsai., 2010). Azonban a káros mutációk okozta funkcióvesztés sokkal magasabb fitneszköltsége és a regulációban bekövetkező nagyobb fajlagos hatások is szerepet játszhatnak. Az is ismertté

vált, hogy a fehérjeaggregáció elkerülése is fontos szelekciós hatással bír: a magasan expresszált fehérjék különösen alacsony aggregációs hajlandósággal bírnak, (Tartaglia és Vendruscolo, 2009) és általánosságban igaz, hogy erős negatív korreláció figyelhető meg az expressziós szint és az aggregációra való hajlam között. Azonban számos fehérjéről kiderült, hogy koncentrációja megközelíti a szolubilitási határértékét és az oldhatóság keskeny mezsgyéjén egyensúlyozik (Tartaglia és mtsai., 2007), ami magyarázattal szolgálhat arra is, hogy egyes patológiás állapotok miért alakulhatnak ki ennek a törékeny egyensúlynak a legkisebb felborulására is. Ilyenek az oxidatív stressz, vagy az öregedéssel járó proteolízis hatékonyságának csökkenése, proteinopátiák kialakulását eredményezve. Ez az ún. Life on the edge hipotézis, amely szerint a proteomra általánosan jellemző az oldhatósági határ közeli koncentráció elérése, valamint a kis mozgástér a koncentráció megnövekedésének tekintetében. A környezeti tényezők változása hatással lehet mind a fehérjék koncentrációjára, mind az oldhatóságukra, így előfordulhat, hogy bizonyos fehérjék koncentrációja a citoplazmában meghaladja az adott körülmények közötti oldhatóságukat. Ez a szuperszaturáció jelensége, amely egy metastabil állapotot eredményez: a rendszer a főlegben lévő oldott állapotú fehérjék kiválásával kerülhet újra egyensúlyi állapotba, amely aggregátumok létrejöttét eredményezi. Ismert, hogy a szuperszaturációra legérzékenyebb fehérjék a humán proteomban a neurodegeneratív és egyéb proteinaggregációval összefüggő betegségekben erősen feldúsulnak (Ciryam és mtsai., 2013, 2015), amellet, hogy funkcionális aggregátumok létrejöttében is szerepet játszanak, mint a melaninszintézisben résztvevő Pmel17 (Falsone és Falsone, 2015). Fontos kérdés, hogy a proteom egészére nézve mennyire elterjedt a „Life on the edge” állapota, milyen gyakori, hogy a fehérjék az oldhatósági határértékük közeli koncentrációban fordulnak elő és gyakran vezet-e a szuperszaturáció aggregációhoz megváltozott körülmények között. Továbbá fontos annak meghatározása, hogy a szuperszaturáció a fehérje életútja során mely állapotban következik be, és hogy függ össze az egyes aggregátumtípusok kialakulása az aggregációra való hajlamuk és koncentrációjuk függvényében.

1.3. A fehérjeaggregációt befolyásoló ismert tényezők

A fehérjeaggregáció biotechnológiai jelentősége miatt számos kutatás foglalkozott azzal, hogy milyen hatások segítik elő az aggregátumképződést. A megváltozott külső

körülmények, mint a pH (Chi és mtsai., 2003), a hőmérséklet (Brange és mtsai., 1992), különböző pufferek (Kameoka és mtsai., 2007; Katayama és mtsai., 2006), olvasztás és fagyasztási ciklusok (Kuelto és mtsai., 2008) mind hozzájárulhatnak az aggregációhoz. Azonban az *in vivo* oldhatóság szempontjából a fehérjékre jellemző tulajdonságok összehasonlítása fontos, hiszen ezek könnyedén változtathatók a DNS-szekvencia megváltoztatásával, szemben a növesztési és fehérje tisztítási protokollok bonyolultabb optimalizációjával. Az eddig végzett kísérletes munkák döntő többsége csupán néhány fehérjére korlátozódott, azonban három fontos vizsgálatot végeztek nagyszámú fehérje bevonásával. Ezek közül kettő az ún. PURE (Nishikawa és Ueda, 2001) *in vitro* transzlációs rendszerrel történt. A PURE rendszer mRNS molekulákat használ templátként és a transzlációhoz szükséges komponensek hozzáadásával akár 160–400 µg/ml-es végkoncentrációban képes a fehérjék előállítására. Az első vizsgálat során kb. 3000 *E. coli* fehérje *in vitro* expresszióját végezték (Niwa és mtsai., 2009), a második esetben pedig kb. 500 *S. cerevisiae* fehérje (Uemura és mtsai., 2018) termelését. A szolubilis és aggregált fázisokat ultracentrifugálással elválasztották és a fehérjék jelenlétét SDS-PAGE módszerrel ellenőrizték. Ezekben a vizsgálatokban számos általános és néhány fajra jellemző (*E. coli* vs. *S. cerevisiae*) aggregációt befolyásoló tényezőt találtak, azonban figyelembe kell vennünk, hogy az *in vitro* környezet nagy mértékben leegyszerűsített az *in vivo*hoz képest, csak a minimális számú fehérjét tartalmazza, valamint alacsonyabb a komponensek összkoncentrációja, mint a citoplazma alkotóié. A harmadik, nagy áteresztőképességű vizsgálat, amely a fehérjék aggregációját befolyásoló tényezőket kereste, a Northeast Structural Genomics Consortiumban készült (Price és mtsai., 2011). 2001 és 2008 között kb. 9600 fehérje termelését végezték el azonos módszerrel *E. coli*-ban. A fehérjék 82%-a bakteriális és 6%-a humán fehérje volt, amelyek, a kísérletek során mind az expressziós szintet, mind a fehérje aggregáció mértékét ellenőrizték gélfiltrációs kromatográfiával és optikai módszerekkel, két megfelelő méretű csoportot eredményezve a statisztikai elemzésekhez. A vizsgálat másik nagy előnye, hogy *in vivo* körülmények között zajlott a transzláció, azonban a termelt fehérjék kiválasztása nem volt szisztematikus, gyakran ugyanannak a fehérjének egy-egy aminosavban különböző változatát tesztelték. Mindhárom vizsgálat fő hiányossága azonban az, hogy a fehérjék foldingja, és ezáltal a funkciójuk ellenőrzése nem történt meg, így nem lehet elkülöníteni azt sem, hogy pre- vagy poszt-folding aggregátum képződött. Mégis, az aggregáció általános paramétereiről

fontos információkat szolgáltatnak. Ezek a főbb fehérjetulajdonságok a fizikai-kémiai paraméterek, mint a méret, a töltés és a hidrofobicitás. A fizikai-kémiai tulajdonságok közvetítői az egyes aminosavak, amelyek gyakoriságában különbség mutatkozik, továbbá a különböző fehérjeszerkezeti egységek (domének) és ehhez szorosan kapcsolódóan az egyes fehérjefunkciók is eltérő arányban fordulnak elő a két csoportban. Nemcsak a szerkezetet, hanem annak hiányát is összefüggésbe hozták az aggregációval, a fehérjerendezetlenség nagyobb mértékben volt jelen az aggregálódó fehérjékben. Mivel az aggregátumképződés általában negatív fitnesszhatással van a sejtekre, a természetben előforduló proteinek bizonyos fokig adaptáltak az aggregáció elkerülésére. A legjobb példát erre a magasan expresszált fehérjék jelentik, mivel erős negatív korreláció mutatható ki mind a génexpresszió, mind a fehérje abundancia és az aggregációképzés között (Groot és Ventura, 2010; Tartaglia és Vendruscolo, 2009). A kísérletes adatokat számos további elemzésben is felhasználták, amelyek felvetettek egyéb összefüggéseket is az aggregáció és a fehérjék tulajdonságai között, sőt, különböző algoritmusok is születtek a fehérjeaggregáció predikciójára (Agostini és mtsai., 2014; Hebditch és mtsai., 2017; Hou és mtsai., 2020) és optimalizációjára (Sormanni és mtsai., 2015; Warwicker és mtsai., 2014a).

Tekintsük át ezekben a vizsgálatokban megjelenő legfontosabb fehérje jellemzőket, hogy ezek miként és miért befolyásolhatják az aggregációt, valamint milyen további munkák támasztják alá ezeket az eredményeket. A fehérjeaggregáció két útvonalát elkülönítő tulajdonságok eddig nem képezték hasonló áteresztőképességű vizsgálatok tárgyát, ugyanakkor a natív szerkezet hiányából levonhatjuk azt a következtetést, hogy a fehérje folding kulcsfontosságú szerephez jut, hiszen az egyik csoportban végbement ez a folyamat, míg a másokban nem, vagy unfolding/misfolding követte. Ismert, hogy a hibásan feltekeredett és denaturált fehérjék gyakran vesznek részt az aggregátumképzésben, ezért áttekintjük a fehérje folding folyamatát és hogy az ebben bekövetkező változások milyen módon járulhatnak hozzá a fehérjeaggregáció típusának meghatározásához.

1.3.1. Fizikai-kémiai paraméterek

Korábbi kísérletek arra utalnak, hogy a fehérjék fizikai és kémiai tulajdonságai erős összefüggést mutatnak az oldhatóságukkal. A legfontosabb fizikai-kémiai tulajdonságok

tulajdonságok a fehérje méret, a töltöttség és a hidrofobicitás. A fehérjék egymás közti interakcióját a taszító és vonzó hatások különbsége határozza meg, amelyek szoros összefüggést mutatnak az oldalláncok töltöttségével és hidrofobicitásával (Chong és Ham, 2014). A poláros és töltött aminosav-oldalláncok interakcióba lépnek a vízmolekulákkal (Qiao és mtsai., 2019), míg a hidrofób oldalláncok egymással vízkiszorítás mellett lépnek kapcsolatba, ez az ún. hidrofób hatás, aminek következtében a fehérjék felszínén helyezkednek el a vízzel nem keverhető oldalláncú aminosavak a belsejében a hidrofóbok. Ennek két fontos következménye van a fehérje szerkezetre és a foldingra nézve (lásd később), valamint közvetlenül az oldhatóságra és az aggregátumképzésre. Azok a fehérjék, amelyek nagyobb össztöltést mutatnak, kisebb mértékben csapódnak ki (Kramer és mtsai., 2012), szemben az aggregációra hajlamos hidrofóbabb proteinekkal. Az erősen negatívan töltött fehérjék különösen jól oldhatók, a pozitív töltés esetében bonyolultabb a helyzet: a nukleinsavak erősen negatívan töltött molekulák, amelyek a pozitívan töltött fehérjékkel könnyebben interakcióba lépnek. Így bizonyos pozitívan töltött fehérjék aggregációra hajlamosak, ugyanakkor nem teljesen tisztázott még az RNS és a DNS szerepe a fehérje aggregációban. Az összhidrofobicitás nem mutatott erős korrelációt az aggregációval sem *in vitro*, sem *in vivo* körülmények között (Niwa és mtsai., 2009; Price és mtsai., 2011; Uemura és mtsai., 2018). Ez arra utal, hogy a teljes fehérjére nézve nincs több hidrofób aminosav az aggregálódó fehérjékben és a teljes fehérje hidrofobicitás nem domináns tényező az aggregátumképzésben. Ebből következően nem a hidrofób aminosavak pusztán száma bír jelentőséggel, hanem elhelyezkedésük a fehérjében. Ez utóbbi elképzelést támasztja alá a fehérjék felszínén található hidrofób aminosavak jelenléte és az aggregációs hajlandóság közti összefüggés (March és mtsai., 2021). Ennek egy komplexebb jellemzésére használjuk az ún. felszíni ragadóságot. A hidrofób és bizonyos pozitív töltésű aminosavak fiziológiai körülmények között az oligomerizációs és fehérje-interakciós felszínek kialakulásáért felelősek, azonban ha máshol fordulnak elő, nemspecifikus interakciókat hozhatnak létre. A felszíni ragadósághoz hozzájárulnak az arginin aminosavak, ami magyarázatot adhat az aspecifikus fehérje-fehérje interakciókon keresztül arra, hogy miért figyeltek meg összefüggést jelenlétük és az aggregációs hajlandóság között (Dubreuil és mtsai., 2019a). Ismert továbbá, hogy a magasan expresszált, tehát erősen szolubilis fehérjék kis felszíni ragadósággal bírnak, amely segítségével elkerülik a nemspecifikus fehérje-fehérje interakciókat (Levy és mtsai., 2012).

A fehérje ragadósság azonban nem egyenértékű az aggregációs hajlammal, ez utóbbi számos egyéb összetevőből is áll, továbbá a ragadósság feltételezi a natív struktúra jelenlétét, amely a naszcens vagy hibásan feltekeredett fehérjék esetén nem mérvadó. A felszíni ragadósság tehát elsősorban a folding folyamatán átesett fehérjék nemspecifikus interakciói által mediált aggregációról adhat információt, ez azonban idáig nem képezte szisztematikus vizsgálatok tárgyát.

Mind *in vitro*, mind *in vivo* vizsgálatok arra utalnak, hogy a fehérjék mérete fordított összefüggést mutat oldhatóságukkal (Niwa és mtsai., 2009; Price és mtsai., 2011; Uemura és mtsai., 2018). Azonban belátható, hogy nem önmagában a fehérje mérete, hanem a fehérje felszínének mérete befolyásolja az aggregációs képességet. A nagy felszínnel rendelkező fehérjék jobban ki vannak téve annak, hogy más fehérjékkel kölcsönhatásba lépjenek, továbbá a hidrofób mag mérete, ezáltal a aminosavak hidrofíli-hidrofób aránya is jelentősen méretfüggő. Amikor a fehérjék töltésének hatásáról beszélünk, a fehérje felszínének töltése a döntő, a fehérjék magjának hidrofobicitása kis jelentőséggel bír. A fehérjék magjában lévő aminosavak csupán akkor befolyásolják az oldhatóságot, ha valamiért megbomlik a natív szerkezet, és a felszínre kerülnek. Ekkor képesek más fehérjék hidrofób részeivel van der Waals kölcsönhatások kialakítására (March és mtsai., 2021), amely aggregációhoz vezethet. A szerkezet felbomlása és a hidrofób kölcsönhatások kialakulása egymással összefüggő folyamatok, ami miatt az unfoldingot az aggregáció első lépésének tekintik (Bianco és mtsai., 2019).

A töltés és hidrofobicitás hordozó egységei a fehérjékben az aminosav oldalláncok, a 20 különböző aminosav gyakoriságának összefüggését az oldhatósággal ezért érdemes külön tárgyalni. Számos munkában elemezték már az egyes aminosav-gyakoriságok hatását az oldhatóságra (Niwa és mtsai., 2009; Price és mtsai., 2011; Uemura és mtsai., 2018). Általánosságban elmondható, hogy a negatív töltésű oldalláncok (Asp, Glu) hozzájárulnak oldhatósághoz (Trevino és mtsai., 2009), azonban a pozitív töltésű oldalláncok nem azonos viselkedést mutatnak. Míg az arginint nagyobb mennyiségben tartalmazó fehérjék aggregációra hajlamosabbak, a lizin jelenléte az oldhatósággal függ össze. Ez a jelenség feltehetően a nukleinsavakkal mutatott interakciós különbségekben gyökerezik, a nagy lizin-arginin arány kedvez a fehérjék oldhatóságának (Warwicker és mtsai., 2014b). A hisztidin szerepe kevésbé ismert, egyértelmű összefüggés mennyisége és

az aggregáció között nincs. A hidrophil aminosavak között a szerin mutat legnagyobb hozzájárulást (Trevino és mtsai., 2009) az oldhatósághoz, a többi poláros aminosav esetén nem mutatható ki feldúsulás egyik csoportban sem. A szerin szerepe azonban bizonyos mértékig továbbra is ellentmondásos, bizonyos fehérjék esetén ugyanis magasabb arányban fordultak elő az aggregálódó csoportban (Fang és Fang, 2013). A hidrofób aminosavak is eltérő részarányban vannak jelen az oldható és az aggregálódó fehérjékben (Fang és Fang, 2013). A nagyobb, aromás oldalláncokat tartalmazó aminosavak gyakorisága egyértelmű összefüggést mutat az aggregátumképzéssel.

1.3.2. Szerkezet és funkció

A fehérjék szerkezetét az aminosavsorrendjük határozza meg. A fehérjék elsődleges szerkezete maga az aminosavsorrend, a másodlagos szerkezetet a különböző, lokálisan kialakuló struktúrákat alkotják, ezek közül a leggyakoribbak az alfa-hélixek, a béta-lemezek és a hurkok. A harmadlagos szerkezetek a fehérje háromdimenziós szerkezete, negyedleges szerkezetről a komplexeket alkotó fehérjék esetén beszélhetünk, melyeket több polipeptidlánc alkot. A béta-lemezek és az alfa-hélixek kialakításához eltérő aminosav-összetétel szükséges. Az alfa-hélixek legfontosabb alkotói a MALEK aminosavak (metionin, alanin, leucin, glutamát és lizin), amelyek hidrogénkötéseikkel stabilizálják szerkezetüket. A béta-lemezek létrehozásában a nagyobb méretű aromás aminosavak (tirozin, fenilalanin és triptofán), valamint az elágazó oldalláncú aminosavak (treonin, valin és izoleucin) alkotják. Az egymás mellett elhelyezkedő béta lemezek rendkívül alacsony energiaszintű, stabil állapotban vannak, valamint képesek bármilyen más béta-lemezzel kapcsolatot kialakítani. Natív körülmények között hélixek és hurkok fedik a béta-lemezek éleit, vagy több béta-lemez hordó alakban összecukódik, így lezárva a szabad végeiket (Richardson és Richardson, 2002). Bár a béta lemezt tartalmazó fehérjék natív szerkezetükben oldhatóak, fragmenseik (Mattice, 1989), vagy megbomlott szerkezetű változataik, amelyek béta-lemezei hozzáférhetőek, könnyen hozhatnak létre aggregátumokat (Niwa és mtsai., 2009). A stabil intermolekuláris interakcióban álló, egymással szorosan csomagolt béta-lemezek fontos alkotói a fibrilláris fehérje aggregátumoknak, a fehérjék béta-lemez tartalma pedig összefügg az aggregációs hajlandósággal (Bellesia és mtsai., 2009). A fehérjék harmadlagos szerkezeteit is csoportosíthatjuk a bennük található másodlagos szerkezetek gyakorisága alapján: főleg

alfa, főleg béta, alfa-béta és egyéb szerkezetek. Tekintve a a béta-lemezek aggregációra gyakorolt hatását, feltételezhető, hogy a főleg béta és alfa-béta harmadlagos szerkezetek is feldúsulhatnak az aggregációra hajlamos fehérjék között. Az egyes fehérje domének is eltérő szerkezettel és fizikai-kémiai tulajdonságokkal bírnak, ezért korábban egyes doméneket összefüggésbe hozták az oldhatósággal. Azonban a nagy áteresztőképességű *in vitro* vizsgálatok (Niwa és mtsai., 2009; Uemura és mtsai., 2018) csupán néhány szerkezeti elem összefüggését állapították meg az alacsony oldhatósággal, ráadásul ezek kis átfedést mutattak *E. coli* és *S. cerevisiae* között. Ettől függetlenül fontos az egyes szerkezeti típusok vizsgálatát aggregációs szempontból *in vivo* körülmények között, ugyanis azok a fehérjék, amelyek rendkívüli mértékű aggregációt mutatnak csak nagy nehézségek árán termelhetők ipari körülmények között. Hasonló a helyzet a fehérje funkcióval is, amely szoros kapcsolatban áll a szerkezettel: amennyiben bizonyos fehérje aktivitások összeegyeztethetetlenek az oldhatósággal, az jelentősen korlátozhatná az ilyen fehérjék termelését *E. coliban*.

1.3.3. A fehérjerendezetlenség

A fehérjék funkciójának ellátásához sokáig elengedhetetlennek tartották a megfelelő térszerkezet meglétét. Ez különösen fontos volt az enzimek esetében, amelyek működését és a szubsztráttal való kölcsönhatását először a kulcs-zár modell írta le (Fischer, 1894), ahol a megfelelő szerkezetű szubsztrát mint kulcs illeszkedik a meghatározott szerkezetű enzimbe, a zárba. Ezt támasztotta alá az a megfigyelés is, hogy a denaturáció a szerkezet mellett az enzimaktivitás elvesztésével is jár. Később azonban az ún. indukált illeszkedés elmélete vált elfogadottá, miszerint a szubsztrát enzimhez kötődése változást eredményez annak szerkezetében. Ez volt az első fontos eredmény, amely a szerkezet és funkció szétválaszthatatlanságát megkérdőjelezte. A következő fontos fejlemény a fehérje térszerkezetek meghatározásakor tett megfigyelés volt, mivel számos fehérje esetében a röntgenkristallográfiás vizsgálatokban bizonyos régiók elektrondenzitásai „üresen” maradtak, a hozzájuk tartozó elektrondenzitás nem jelent meg. Bizonyos fehérjék pedig szokatlanul érzékenyek voltak az *in vitro* proteolízisre, pedig szerkezetükből ez nem következett volna. Ezekre a kérdésekre adott választ a fehérjerendezetlenség elmélete, amely kimondja, hogy egyes fehérjék vagy fehérje részletek nem rendelkeznek állandó, meghatározott térszerkezettel, hanem dinamikusan

váltakoznak a szerkezetek sokasága között és mégis képesek funkciójukat ellátni. Magas fokú rendezetlenség esetén akár a teljes proteinből is hiányozhatnak a másodlagos struktúrák, egy rendkívül hajlékony, változékony szerkezetet eredményezve. Később az is bebizonyosodott, hogy a fehérjerendezetlenség nem ritka kivétel, hanem szabály: a legtöbb virális, bakteriális, archaea és eukarióta genom is számos rendezetlen fehérjét tartalmaz (Deforte és Uversky, 2016; Dunker és mtsai., 2000; Uversky, 2015). A rendezetlen szakaszokat általában a hosszuk szerint jellemezzük, a 10-30 aminosavnál hosszabbakat (hosszú) rendezetlen régióknak nevezzük, ez alatt pedig eukarióták esetén lineáris motívumról beszéltünk. Prokariótákban a rövid rendezetlen szakaszoknak korábban nem tulajdonítottunk nagy jelentőséget. A bakteriális proteomok általában kevesebb, mint 10%-a tartalmaz hosszabb rendezetlen régiókat, bár a rendezetlen aminosavak aránya elérheti a 30%-ot is, ez a szám jelentősen nagyobb az eukarióták esetén (Uversky, 2015). A fehérjerendezetlenség mértéke a komplexitás növekedésével egyre nő (Schad és mtsai., 2011), amely rávilágít a rendezetlen régiók sejtekben betöltött szerepétől. A rendezetlen régiók szerepet játszanak a molekuláris célpontok felismerésére, legyen az DNS (Tucker és mtsai., 1994), RNS (Bou és mtsai., 2000), egy másik fehérje (Johansson és mtsai., 1998; Knighton és mtsai., 1991), vagy egy kismolekula (Bidwell és mtsai., 1999). A rendezetlen fehérjék képesek számos célpont felismerésére, szemben azonban a multimer vagy fehérjekomplex-alkotókkal a kötődésük erőssége relatíve gyenge, ezért rövid életidejűek, átmeneti kapcsolatokat hoznak létre. Az így létrejövő tranziens kapcsolatok fontos szerepet játszanak a sejtek szabályozó folyamataiban, ahol sok komponens együttes működésére van szükség. A rendezetlen régiók további funkciója, hogy az eukariótákban poszttranszlációs módosítások helyeként szolgálnak (Darling és Uversky, 2018).

A rendezetlen fehérjék szerepe a fehérje aggregáció folyamatában azonban továbbra sem tisztázott, számos ellentmondásos eredmény látott napvilágot. Egyrészt a fehérjerendezetlenség kialakításáért a töltött aminosavak felelősek, továbbá a rendezetlen régiókban kis mennyiségben fordulnak elő nagy, hidrofób oldalláncú aminosavak (Romero és mtsai., 2001; Uversky és mtsai., 2000). Emiatt nem vesznek részt ezek a régiók a hidrofób mag kialakításában, ugyanakkor mint láttuk, ezek a tulajdonságok az oldhatósággal vannak összefüggésben. Továbbá *E. coli* esetében a

fehérjerendezetlenség mértéke pozitív korrelációt mutat a génexpresszióval és mivel a magasan expresszált fehérjék ellenállóbbak az aggregációval szemben, ez arra utal, hogy a rendezetlenség növekedése nem növeli az aggregációra való hajlamot. Ugyanakkor *S. cerevisiae*-ben a magasan expresszált fehérjék kevesebb rendezetlen régiót tartalmaznak (Macossay-Castillo és mtsai., 2019), mintegy adaptációként a nem funkcionális interakciók minimalizálására, humán fehérjék esetében pedig korrelációt találtak a fehérje aggregációs hajlandóság és a rendezetlenség között (Carvalho és mtsai., 2013). Felmerült a rendezetlen régiók szerepe a felszíni ragadósság okozta nem-natív fehérje interakciók és aggregációk elkerülésében is: töltöttségüknél és alacsony hidrofobicitásuknál fogva a képesek minimalizálni ezeket az interakciókat, megvédve a fehérjéket az aggregációtól. A fehérjerendezetlenség közvetlen szerepet játszik a fehérjék foldingjában és oldhatóságának meghatározásában. Oldhatatlan fehérjékhez kötve a rendezetlen régiók segítik a fehérjék foldingját és oldódását, mivel egyfajta entrópikus sörteként távol tartják a többi makromolekulát, valamint poláros felszínt biztosítanak (Santner és mtsai., 2014; Simone és mtsai., 2012). Mindezen tulajdonságaik ellenére számos humán fehérjeaggregációhoz köthető betegségben fontos fehérjéről kiderült, hogy nagyobb rendezetlen régiókat tartalmaznak (Cheng és mtsai., 2006; Das és Mukhopadhyay, 2011; Iakoucheva és mtsai., 2002; Uversky, 2014), sőt ezek a régiók fontos szerepet játszanak az aggregációs folyamatokban. Ez az összefüggés azonban nem igazoltan ok-okozati, a fehérjék számos interakciókban játszott szerepéből is adódhat, valamint lehet adaptáció az aggregáció ellen. Emiatt rendkívül fontos a rendszerszintű összefüggések feltárása a fehérjerendezetlenség és az aggregáció között.

1.3.3.1. A fehérjerendezetlenség in silico predikciója

A fehérjerendezetlenség felfedezése óta számos predikciós eljárás is kidolgozásra került a rendezetlen fehérjék tulajdonságainak figyelembevételével. Ezek különböző komplexitásúak, egyesek szerkezeti információkat igényelnek bemenetként, míg más típusokhoz elégséges a fehérjeszekvencia. A legkomplexebb módszerek, mint a fehérjedinamikai számítások kivitelezése, hatalmas számítási kapacitást igényelnek, ezért teljes proteom szinten megvalósíthatatlanok, ezért ilyen vizsgálatokban leginkább a szekvencia alapú módszerek használata elterjedt. Ide tartozik a **DisEMBL** (Linding és mtsai., 2003) szekvencia alapú módszere, amely azon alapszik, hogy bizonyos másodlagos

szerkezeti elemek kialakulása meghatározott aminosav-összetétellel társul. Az algoritmus három különböző kimenetet ad, az első a hurkok és loopok meghatározása, amelyek mindent tartalmaznak, amelyek nem meghatározott másodlagos szerkezetek, mint a hélixek és a béta-lemez. Ezek a hurkok nem feltétlen rendezetlenek, de rendezetlen régiók nem találhatók hurkokon kívül. A másik, ennél konzervatívabb becslést az ún. hotloopok képezik, amelyek a különösen mobilis hurkokat reprezentálják, abból a megfigyelésből kiindulva, hogy a rendezetlen régiók magas fokú dinamizmussal bírnak. A harmadik kimenet a röntgenkristallográfiás módszerek során talált hiányzó koordinátákra fókuszál. Olyan szekvenciárészleteket azonosít, amelyek hasonlítanak a hiányzó elektrondenzitású régiókra a szerkezeti vizsgálatokra. Ez az utóbbi megközelítés manapság elavultnak tekinthető, a hotloops kimenet a legkonzervatívabb fehérjerendezetlenség-becslés. Az **iUpred** (Csizmo, 2005) módszer alapját a különböző, ismert térszerkezetekben található aminosavpárok közti interakciós energia alapú statisztikai megközelítés adja. Az egyes aminosavpárok eltérő energiával járulnak hozzá a harmadlagos szerkezet kialakításához, a rendezett és rendezetlen aminosavak energiái más mértékben kompatibilisek a különböző másodlagos szerkezetekkel. Egy aminosav vizsgálata a közvetlen környezetében kontextusfüggően információt ad annak rendezetlenségéről. Az algoritmus bemenete ebben az esetben is az elsődleges szekvencia, a kimenet pedig aminosavszintű rendezetlenségi érték. Ennek két típusa van, az egyik az iUpred Short, amely a rövid rendezetlen szakaszok azonosítására alkalmas, a másik pedig az iUpred Long, amely a hosszabb, legalább 30 aminosavas rendezetlen szakaszokat keresi. A **vsl2b** (Peng és mtsai., 2006) módszer több ezer ismert rendezett és rendezetlen szerkezetű fehérje szekvenciajellemzőire épül, amely segítségével egy szekvencia aminosavait képes besorolni rendezett és rendezetlen kategóriákba. Ezek a jellemzők az aminosavak gyakorisága, a szekvencia komplexitása, a töltés és hidrofobicitás, az átlagos rugalmasság és egy szekvenciahasonlóság-profil. Az **Espritz** (Walsh és mtsai., 2012) módszer alapja egy neurális háló, amely az aminosavsorrendből prediktálja a rendezetlen szakaszok jelenlétét. Kimenete azon múlik, hogy milyen szerkezeti információkat használunk a számítások során. Ezek lehetnek röntgenkristallográfiás és NMR méréssel készült szerkezetek. Számos további módszer létezik, azonban az eddig említettek lefedik a rendelkezésre álló megközelítéseket,

amelyek proteom szinten is használhatók, mivel nem igényelnek hatalmas számítási kapacitást vagy processzoridőt.

1.4. Posztfolding és prefolding aggregációt befolyásoló tényezők

Korábbi szisztematikus vizsgálatok híján nem áll rendelkezésünkre elég információ arra nézve, hogy mely tényezők hatására dől el, hogy egy fehérje a folding előtt vagy után lép-e aggregátumba. Ez abból is ered, hogy az ismert funkcionális aggregátumok száma meglehetősen alacsony, hiszen csak nemrégiben kerültek felfedezésre. A különbség oka lehet a korábbiakban tárgyalt fehérje paraméterek egyike, amely valamelyik típusban szélsőségesebb értéket mutat, azonban amennyiben figyelembe vesszük, hogy a két aggregátumtípus közti legfontosabb különbséget az amiloidokban megfigyelhető natív szerkezet és funkció hiánya adja, kézenfekvő, hogy a fehérje foldingnak fontos szerepe van az aggregációs folyamat során. Ahhoz azonban, hogy megértsük, hogy függ a fehérje foldingtól az aggregáció folyamata, az előbbi alaposabban át kell tekintenünk.

1.4.1. A fehérje folding

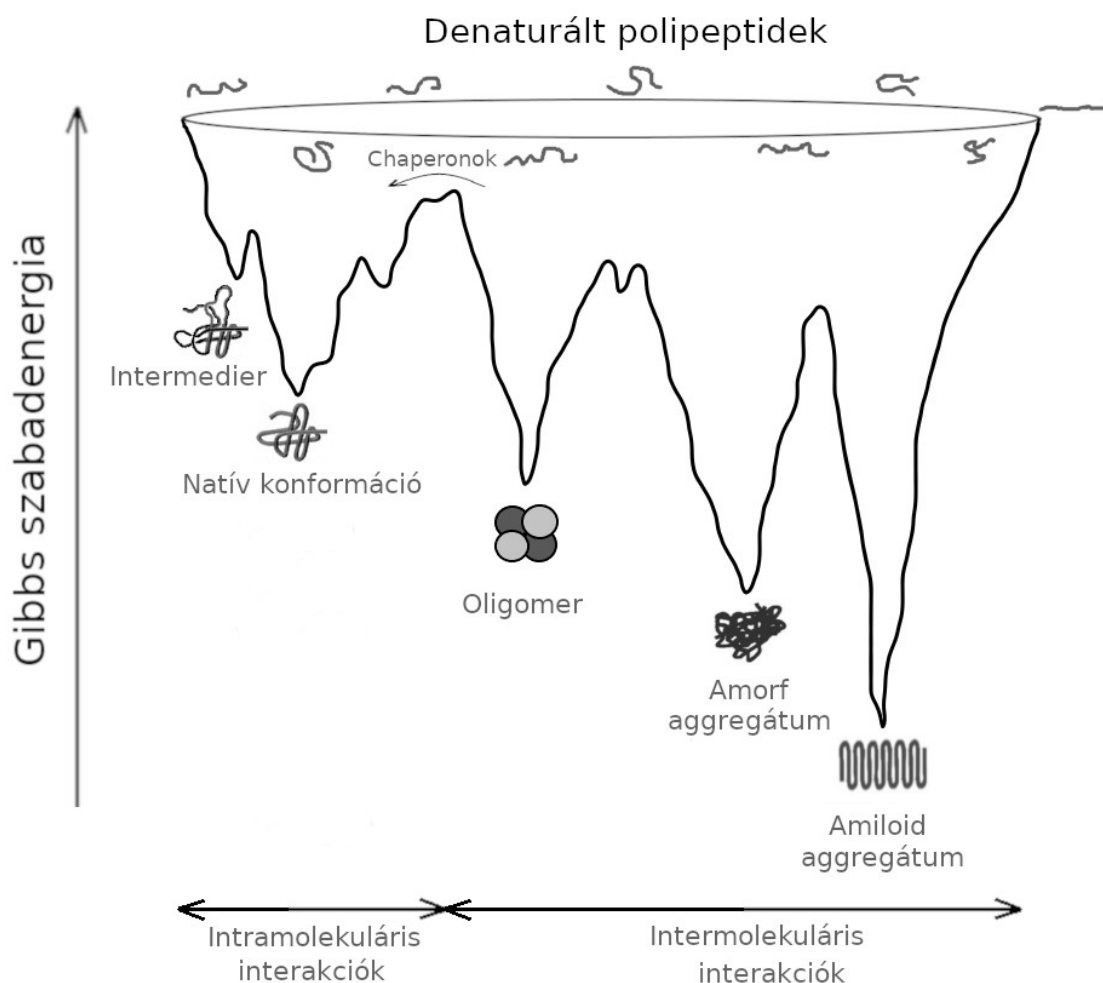
A fehérje folding az a folyamat, amely során egy egydimenziós fehérjelánc fölveszi háromdimenziós struktúráját, amelyet alapvetően a benne lévő aminosavak határoznak meg. Ez egy fizikai-kémiai folyamat, amelyet csak részben sikerült feltárni annak ellenére, hogy több évtizedes intenzív kutatás tárgyát képezi. Az első fontos elméletet Anfinsen és munkatársai alkották meg, ez az ún. termodinamikai hipotézis (Haber és White, 1961), amely szerint egy adott fehérje minden lehetséges szerkezete bizonyos nagyságú Gibbs szabadenergiával bír és az összes szerkezet közül a natív esetén a legkisebb a szabadenergia értéke, ezt a polipeptidlánc véletlenszerű kereséssel találja meg, majd csapdázódik a natív energiaszinten. Ugyan Anfinsen és munkatársainak elmélete mérföldkőnek számított a fehérje folding vizsgálatában, több hiányosságára is fény derült: egyes fehérjék legstabilabb állapota nem a natív állapot, sőt egyes esetekben, mint amilyenek a különböző amiloid típusú aggregátumok, a fehérjék irreverzibilis, rendkívül stabil szerkezetet alkotnak. Továbbá semmilyen kézzelfogható magyarázattal nem szolgált az egyes fehérjék folding sebességei között mutatott hatalmas különbségekre és kinetikai paramétereikre. Az első foldingkinetikai elképzelések abból a megfigyelésből adódtak, amely szerint egy polipeptidlánc aminosavai rendkívül sokféle szerkezet felvételére

képesek. Egy átlagos bakteriális fehérje a hús természetesen előforduló aminosav átlagosan 360 hosszúságú láncolata (Brocchieri és Karlin, 2005), így a lehetségesen felvehető szerkezetek száma meghaladja az összes atom számát az univerzumban. Hogy lehetséges tehát, hogy a legtöbb fehérje külső behatás nélkül, akár mikroszekundumok alatt képes elnyerni a natív szerkezetét? Ez a Levinthal-paradoxon, amely kizárja, hogy a fehérjék véletlenszerű keresés eredményeképpen éri el natív szerkezetüket. Létezniük kell meghatározott folding útvonalaknak, amelyeken keresztül egy polipeptidlánc gyorsan és hatékonyan képes rátalálni a megfelelő natív konformációjára (Levinthal, 1968). Levinthal elképzelése szerint a folding útvonalak egy tölcsérként ábrázolhatók, amelyben a tölcsér peremén lévő random szerkezetű fehérje több intermediereken keresztül halad lefelé, így számottevően csökkentve a lehetségesen felvehető konformációk számát, meggyorsítva a folding folyamatát. Az egyre gyarapodó kísérletes eredményekből kiderült, hogy a fehérjék folding sebessége akár 8 nagyságrendet is lefedhet és Plaxco és munkatársai írták le az első jelentős összefüggést a fehérjék natív szerkezete és a folding sebessége között (Plaxco és mtsai., 1998). Az általuk bevezetett fehérje jellemző az ún. contact order azt írja le, hogy a natív szerkezetben mekkora mértékben alakulnak ki kapcsolatok az polipeptidláncon egymáshoz közeli vagy távoli aminosavak között. Minél több lokálisan kialakuló struktúrát tartalmaz egy fehérje szerkezete, mint amilyenek az alfa-hélixek és a hurkok, annál kisebb a contact order és annál gyorsabban lejátszódik a folding folyamata, szemben a nem lokális szerkezetek esetén, mint a béta-lemezek. Később a fehérje folding sebessége és a natív szerkezet más tulajdonságai közti korrelációt is sikerült megfigyelni, mint amilyen a polipeptidlánc effektív hossza (Ivankov és Finkelstein, 2004) (a lánc hossza a hélixek nélkül), a másodlagos szerkezetek összhossza (Huang és mtsai., 2007) és összesített contact ordere (Zhou és Zhou, 2002).

A fehérjefolding-kutatások következő fontos lépése a natív szerkezetet meghatározó erők és azok relatív hozzájárulásának megbecslése a natív szerkezet kialakításához. A natív és a denaturált fehérjék között kis energiakülönbség (Yang és mtsai., 2007, Pace és mtsai., 1996) tapasztalható, amely több erő együttes hatását is lehetővé teszi. Azonban különböző statisztikai mechanikai modellek bebizonyították, hogy általában a hidrofób kölcsönhatások dominálnak a natív szerkezet meghatározásában (Dill és mtsai., 2008), nem a töltött oldalláncok. Továbbá a hidrogénkötések is fontosak a

másodlagos szerkezet kialakításában (Deechongkit és mtsai., 2004), csakúgy, mint a fehérje magok szoros csomagolódnak következtében létrejövő van der Waals kölcsönhatások (Chen és Stites, 2001). Erős bizonyítékok vannak arra nézve is, hogy a fehérjeszerkezet egyik legfőbb meghatározója a vízmolekulák hatására létrejövő hidrofób kölcsönhatások az oldalláncok között (Cordes és mtsai., 1996). A fehérjeszerkezetet befolyásoló erők leírása és az egyre kifinomultabb számítógépes modellek mellett azonban továbbra is ismeretlen volt a folding mechanizmusa. A számos felmerülő kérdés tükrében nem meglepő, hogy több modellt is alkottak a folding mechanizmusára nézve. Ezek a diffúzió-ütközés (Karplus és Weaver, 1994) modellje (diffusion collision model), nukleáció-kondenzáció (Nölting és Agard, 2008) modellje (nucleation-condensation model) a hidrofób cipzár (Dill és mtsai., 1993) modell (zippering and assembly model), a fehérje folding kirakómodellje (Harrison és Durbin, 1985) (jigsaw puzzle model), valamint a hidrofób összeomlás (Dill, 1985) (hydrophobic collapse) modellje. A folding eddig említett modelljei mind igazolást nyertek egyes fehérjék esetében, azonban egyikük sem tekinthető általánosnak az összes protein esetében. Az elméletek szintézise a fehérje folding ún. energiatájkép (energy landscape, 2. ábra) reprezentációján keresztül történt, amelynek központi eleme a tölcsér modell (funnel) (Onuchic és mtsai., 1997). Ez a modell mind a termodinamikai mind a kinetikai kérdések vizsgálatához egységes keretrendszert biztosít. Az energiatájkép-modell szerint a fehérjék különböző konformációi különböző szabadenergiával bírnak, a folding energetikai szempontból egy tölcséren való lefolyáshoz hasonlítható. A tölcsér peremén helyezkednek el a magasabb szabadenergiával rendelkező naszcens láncok, amelyek a folding során egyre alacsonyabb energiaszinteket vesznek föl. A tölcsérben több úton is tud lefelé haladni egy fehérje, a tölcsér mélysége a natív állapot által elérhető energiaminimum nagyságát jelképezi, a tölcsér szélessége a polipeptid konformációs entrópiáját, a tölcsér szájánál lévő horizontális régió pedig a heterogén random polipeptidek szabadenergiáját. Ez a modell implicit módon reprezentálja azt is, hogy a nagy energiájú denaturált láncok számos különböző konformációval bírhatnak, és ahogy haladnak a folding folyamatában, a lehetséges konformációk száma rendkívül lecsökken, míg végül az egyetlen, natív szerkezet marad elérhető számukra. A fehérje folding tölcsér modellje továbbá képes a foldingkinetikai eredmények integrálására is. Attól függően, hogy milyen „sima” a tölcsér oldala gyorsabb vagy lassabb a fehérje folding, a legnagyobb sebességgel a teljesen sima felszínek esetén

következik be. A magasabb szabadenergiával bíró kiemelkedések a tölcsér oldalában gátat szabnak a fehérje folding továbbhaladásának, mivel a fehérjének újra egy magasabb energiájú állapotban kell áthaladnia. Hasonlóan a tölcsér oldalában lévő mélyedések lokális energiaminimummal bíró nem natív szerkezeteket jelentenek, amelyek kinetikailag csapdázott, stabil intermediereket jelképeznek. A fehérjének át kell lépnie az energiagátat, hogy tovább haladhasson a tölcsér alján lévő globális energiaminimum felé.



2. ábra A fehérje folding és a fehérje aggregáció energiatájkép modellje. A fehérjék különböző konformációihoz különböző Gibbs szabadenergia szintek tartoznak. A denaturált peptidek a legmagasabb energiájúak, majd a foldingintermedierek és a natív szerkezet következik. Ez utóbbi egy fehérje esetében a globális energiaminimummal bíró konformáció, a létrejöttében az intramolekuláris interakciók a meghatározóak. Azonban több fehérje kölcsönhatása során létrejöhetnek a monomernél kisebb energiájú oligomerek és aggregátumok, amelyekben a szabadenergiaváltozásért már főleg az intermolekuláris interakciók a felelősek. Az amiloid

aggregátumok bírnak a legalacsonyabb szabadenergiával, ezért ezeket az aggregátumokat csak erős denaturálószerrel lehet felszabadítani ebből a konformációból.

1.4.2. Az energiatájkép és a fehérje aggregáció

Bár normál körülmények között a fehérjék natív konformációja a legalacsonyabb energiájú állapot, ez csak akkor igaz, amennyiben egy fehérjéről izoláltan beszélünk. Azonban az energiatájkép-modell keretrendszer biztosít az aggregátumképződés mögött álló termodinamikai és kinetikai kérdések vizsgálatához is. Több fehérje esetén léteznek alacsonyabb energiájú állapotok, amelyek egy része fiziológiás, mint az oligomerek, másik részük nem, mint az amiloid aggregátumok. Ezekben az esetekben a szabadenergia meghatározásában már nem az intramolekuláris interakciók dominálnak, hanem a több fehérje között létrejövő intermolekuláris kapcsolatok. Az aggregátumok kialakulása hasonlóan a foldinghoz egy lokális (és gyakran globális) energiaminimumba történő eljutás eredménye. Az energiatájkép implicit módon azt is megmutatja, hogy az aggregátumok létrejöttében fontos szerepe van az unfolded vagy hibásan feltekeredett intermediereknek, ugyanis ezek elérhető útvonalakat képviselnek az aggregátumok felé a tájképen mind a natív szerkezetű, mind a naszcens fehérjék irányából. Ezen felül fontos szerepet játszanak a megnövekedett hidrofób felszíneikkel az intermolekuláris kapcsolatok kialakításában. Ahhoz, hogy a natív konformációból aggregálódjon egy fehérje, szükséges lehet a natív szerkezet megbomlása, amely egy energiagct átlépésével jár, a funkció szempontjából azonban lényeges, hogy mekkora mértékben marad meg a natív szerű szerkezet. Az energiatájkép-modell jól érzékelteti, hogy az aggregáció és a folding folyamata nem egymástól függetlenül játszódik le, hanem szorosan összefüggenek, ugyanakkor a fehérje végső állapota vagy a natív struktúra vagy az aggregált állapot. Az ún. kinetikai kompetíció (de Groot és Ventura, 2006) elmélet alapján a fehérjék aggregációja és foldingja egymással versengő folyamatok, a fehérje sorsát a relatív sebességük határozza meg. Az, hogy a fehérje aggregáció a natív konformáción keresztül következik be vagy sem határozza meg, hogy az adott fehérje milyen típusú aggregátumot képez. Ezt reprezentálhatjuk úgy, hogy milyen sima a natív szerkezethez vezető tölcser, milyen meredekséggel bír és mekkora energiagát választja el az aggregátum állapotát a natív struktúrától. Ahhoz, hogy ezt számszerűsíteni tudjuk, két

tulajdonság meghatározása elengedhetetlen, a fehérjék folding sebessége és az aggregációra való hajlandóság.

1.4.2.1. A fehérje folding in silico predikciós módszerei

A számos folding mechanizmus együttes előfordulásából látszik, hogy rendkívül sokszínű folyamatról van szó, ahol egy forgatókönyv nem alkalmazható az összes fehérjére. A fehérje folding vizsgálatára irányuló kísérleti módszerek rendkívül bonyolultak és munkaigényesek, ugyanakkor nem mindig tudnak kellően részletes és időben is nagy felbontású eredményt adni (Freddolino és mtsai., 2014). Emiatt rendkívül fontosak az *in silico* módszerek, amelyek közül a legpontosabb eredményt a molekuladinamikai számítások szolgáltatják. Ezek a módszerek atomszintű szerkezeti leírást adnak a fehérjék foldingjáról nanoszekundumos felbontással, ami által számos olyan fehérje foldingja vizsgálhatóvá válik, amely elérhetetlen volt a kísérleti biológia módszereivel (Hollingsworth és Dror, 2019). Azonban a molekuladinamikai számításoknak is megvannak a maguk hátulütői: a megfelelő szerkezet és folding útvonal megtalálásához számos, különböző kiindulási konformációból kell kiindulni, mint ahogy egy denaturált fehérje is számos különböző kezdeti szerkezettel bírhat. A molekuladinamikai számításokban az atomok egymásra gyakorolt vonzó és taszító hatását különböző erőterek segítségével modellezhetjük. Az egyes erőterek választása egy fehérje szerkezetének vagy foldingjának modellezésében jelentősen befolyásolhatja a kapott eredményt. Végül pedig a legkisebb fehérjéket leszámítva a számolások nagy számítási kapacitást igényelnek. Emiatt nem alkalmazható ez a módszer egy nagy áteresztőképességű vizsgálat kiegészítéseként, ugyanis több ezer fehérje molekuladinamikai modellezése még az egyre növekvő számítási kapacitásokkal is beláthatatlan ideig tartana. Emiatt a fehérje folding predikciójára más, kevésbé pontos, ugyanakkor továbbra is informatív módszerek alkalmazhatók, amelyek statisztikai megközelítéseken alapulnak, nem atomszintű modellezésen. Az első ilyen módszer Plaxco és munkatársai megfigyelésére vezethető vissza (Plaxco és mtsai., 1998), miszerint a fehérjék folding sebessége összefügg azzal, hogy a harmadlagos szerkezetben az egymáshoz közeli aminosavak milyen közel helyezkednek el az elsődleges szerkezetben. Ez a mérőszám a contact order, amely ha alacsony, akkor a fehérjének nem kell komplex folding útvonalat bejárnia, ezért gyors a folding. A contact ordert sokkal kisebb számítási

kapacitást igénylő algoritmusokkal számolhatjuk (Plaxco és mtsai., 1998), amely mégis érdemleges információt ad a folding folyamatáról. Ehhez szükséges ismerni a fehérjék szekvenciáját és harmadlagos szerkezetét, ez azonban rendelkezésre áll már az *E. coli* esetében (Xu és Zhang, 2013). Egy másik megközelítés arra alapszik, hogy bizonyos szekvenciaelemek hajlamosak egy adott másodlagos struktúra felvételére, és egyes másodlagos struktúrák jellemzően gyorsan átesnek a folding folyamatán. Így a szerkezet tényleges ismerete nélkül is megbecsülhető a folding sebessége, amennyiben azonosítjuk a jellemzően gyors és a lassú foldinggal bíró szekvenciamotívumokat az elsődleges szerkezetben. Ilyen algoritmus a FOLD-RATE, amely alkalmas a teljes *E. coli* proteom vizsgálatára. Az eddig említett módszereknek azonban hátránya, hogy csak egydoménes fehérjékre alkalmazhatók. Ez abból következik, hogy a többdoménes fehérjékre, valamint az oligomerekre és a fehérje komplexekre nem alkalmazhatók azok az összefüggések, amelyekre az előbbi módszerek alapulnak, így nem adnak megbízható eredményt ilyen esetekben. Így is azonban jelentős számú fehérje vizsgálható velük, tekintve, hogy a fehérjék fele kb. 1 doménnel rendelkezik.

1.4.2.2. A fehérje aggregáció predikciója

Az eddig megismert aggregációval összefüggő tulajdonságok számos olyan *in silico* módszer kifejlesztéséhez vezettek, amelyekkel tetszőleges fehérjék aggregációs hajlandósága megbecsülhető. Mivel a posztfolding aggregáció jelensége korábban kevésbé volt ismert, valamint a legnagyobb humán egészségügyi jelentőséggel bíró fehérjék amiloid típusú aggregátumokat képeznek, a legtöbb módszer ennek az aggregációs típusnak az előrejelzésére fókuszál. Ez azonban értékes információ lehet a fehérje aggregátumok típusainak elkülönítésében és a mögöttes mechanizmusok megértésében. Amennyiben egy fehérje rendkívüli aggregációs hajlammal bír, az magyarázatot adhat arra nézve, hogy miért nem éri el a natív szerkezetét, mielőtt bekövetkezne az aggregátumképződés, különösen, ha lassú folding is terheli a folyamatot. Az AggreScan (Conchillo-Solé és mtsai., 2007) módszer *in vivo* kísérleti eredményekre támaszkodó predikciós algoritmus, amely a szekvencián aggregációs forrópontokat azonosít. Ezek a régiók különösen alkalmasak az aggregációs folyamat beindítására, valamint a denaturált állapotban is jelen vannak és befolyásolhatják a fehérje viselkedését. A PASTA 2.0 (Walsh és mtsai., 2014) módszer szintén

szekvenciainformációkat használ, amelyből a másodlagos struktúrák jelenlétét, hiányát és az aggregációs energiát becsül, ezzel kiegészítve az AggreScan eredményeit. A nagyon alacsony aggregációs energiájú fehérjékre tekinthetünk úgy, mint egy nagyon mély, nagyon meredek völcsér vezetne az aggregációs globális energiaminimumhoz, amely gyors aggregációt eredményez. A Tango (Rousseau és mtsai., 2004) algoritmus több lépésben jut el az egyes fehérjék béta-aggregációs képességének megbecsléséhez. Figyelembe veszi a polipeptidet alkotó aminosavak fizikai-kémiai tulajdonságait, és a pH a koncentrációt és egyéb környezeti paramétereket, majd ebből következtet a másodlagos szerkezetre, azok jelenlétéből pedig az aggregációs hajlandóságra. Fontos azonban megjegyezni, hogy az algoritmus eredeti célja a mutációk hatásának vizsgálata az aggregációs képességre, ezért eredményei teljesen különböző fehérjék esetén csupán tájékoztatásul szolgálnak, kiegészítve a többi aggregációs prediktor eredményeit. A különböző aggregációs és folding prediktorok együttesen elégséges adatot szolgáltatnak a fehérje folding és aggregáció közti kompetíció vizsgálatához, amely megmagyarázhatja, hogy milyen folyamatok vezetnek a különböző típusú aggregátumok kialakulásához.

1.4.3. A chaperonok

A fehérjék harmadlagos szerkezete az aminosavsorrendben kódolt, és számos fehérje képes ezt elérni bármiféle külső beavatkozás nélkül. Vannak azonban bizonyos polipeptidek, amelyek kisebb-nagyobb mértékben nem megfelelő, hibásan feltekeredett konformációt vesznek föl a folding során, amely aggregációhoz vezethet. Ez rendkívül negatív következményekkel bír a sejtre nézve, ezért a baktériumoktól az emberig minden élőlény bír a rendkívüli konzervációt mutató dajkafehérje-rendszerrel. Ez felelős a proteom integritásának fenntartásáért úgy, hogy segítenek a fehérjéknek fölvenni a natív szerkezetüket, legyen szó naszcens láncokról, vagy olyan fehérjékről, amelyek misfolding következtében rendellenes szerkezettel bírnak. *E. coli*-ban a két legfontosabb chaperon a *DnaK* és a *GroEL*, amelyek más módon és más időben lépnek interakcióba a fehérjékkel. A *DnaK* egy konstitutívan expresszáldó dajkafehérje, amely mennyisége stresszhatásra tovább növekedhet, ugyanakkor normál körülmények között nem esszenciális (Bukau és Walker, 1989). A *DnaK* fehérje társchaperonjával, a *DnaJ*-vel együtt már a transzláció során képes megkötni a fehérjéket és segíteni azoknak elérni natív konformációjukat azáltal, hogy a félig szintetizált fehérjéket megakadályozza az aggregációban. A *DnaJ*

fehérje ismeri föl és „mutatja be” a nem natív szerkezetű fehérjéket, valamint a DnaK konformációváltozásához szükséges ATP-hidrolízist segíti. Az interakció során a DnaK „fedél” része kinyílik, bejut a kliensfehérje, majd lezárul. A DnaK fehérje interakcióba lép a kliensével, amelyet rövid hidrofób szakaszai által ismer fel. Ezek a szakaszok a natív szerkezet hiányára utalnak, és a DnaK unfoldáz aktivitásával képes a fehérje szerkezetének felbontására ATP hidrolízise mellett. A DnaK fehérjében további ATP hidrolízis hatására a kliensfehérje elhagyja a chaperont, így újra lehetősége nyílik a normál foldingra. A fehérjefolding energiatájkép-modell értelmében úgy tekinthetünk a chaperonokra, mint fehérjékre amelyek képesek a lokális energiaminimummal bíró, kinetikailag csapdázott szerkezet kialakulását megakadályozni. *In silico* eredmények arra is engednek következtetni, hogy a DnaK és a kliens fehérje interakciója lezár bizonyos folding útvonalakat, megakadályozva a nem natív konformációk kialakulását és elősegítve a natív szerkezet létrejöttét. Ilyen értelemben ez a chaperon megváltoztatja a fehérjefolding energiatájképének topológiáját, simábbá téve azt (Lu és mtsai., 2021). Bizonyos fehérjéknek további segítségre van szükségük a natív konformációjuk eléréséhez, ezek továbbadódhatnak a GroEL chaperon komplexnek. Ez egy ketrecszerű fehérje komplex, amely belsejében az egydoménes fehérjék foldingja történik. Amennyiben egy fehérje mégsem nyeri el a natív szerkezetét és aggregálódik, a DnaK rendszer képes ennek a folyamatnak a visszafordítására (Mogk és mtsai., 2003), továbbá a polipeptidek kinyerésére a fehérje aggregátumokból, ha azok nem lépnek át egy bizonyos mérettartományt (Diamant és mtsai., 2000). A DnaK chaperon rendszer bír a legtöbb klienssel az *E. coli* proteom tagjai között, mintegy 700, főleg citoplazmatikus fehérjéről igazolódott, hogy interakcióba lépnek ezzel a chaperonnal (Calloni és mtsai., 2012). A GroEL rendszer kevesebb, csupán kb. 250 klienssel rendelkezik, amelyek közül kevesebb, mint 100 obligát GroEL kliensfehérje (Fujiwara és mtsai., 2010). Tekintve a DnaK rendszer számos kliensét és sokféle aktivitását, feltételezhető, hogy ez a chaperon fontos szereppel bír a prefolding aggregáció megakadályozásában. Ezt támasztja alá az is, hogy a DnaK egyes fehérjék túltermelésénél nem befolyásolja az aggregáció mértékét a sejtekben, ugyanakkor jelenléte drámaian növeli a fehérje aktivitást mind a szolubilis, mind az aggregálódó fázisban (Martínez-Alonso és mtsai., 2007). Továbbá, bár bevett gyakorlat a fehérje túltermelés során a DnaK együttes expressziója, ennek eredményessége rendkívül változó. Ezért rendkívül fontos a DnaK szerepének szisztematikus vizsgálata az eltérő

aggregátumtípusok kialakulásában, ugyanis lehetséges, hogy eddigi félreértelmezett szerepe helyett a funkcionális aggregátumok létrehozásában válhat megbízható, széles körben használható eszközzé.

1.5. Célkitűzések

Disszertációm során arra a kérdésre keresem a választ, hogy a fehérjék életútjának melyik lépésében következik be a szupersaturáció és mely fehérje jellemzők befolyásolják az aggregáció folyamatát és különböző útvonalait *in vivo Escherichia coliban*. Ezt a kérdést több aspektusból vizsgáltam:

- Milyen gyakorisággal alakulnak ki az egyes aggregátumtípusok *in vivo* körülmények között? Igaz-e a feltételezés, miszerint a proteom döntő többsége az oldhatósági limithez közeli koncentrációban fordul elő a sejtekben a „Life on the edge” hipotézisnek megfelelően? Hogy viszonyulnak eredményeink a korábbi *in vitro* körülmények között végzett kutatásokéhoz? Mennyire robosztus vizsgálati módszerünk, alkalmas-e a nagy áteresztőképességű mikroszkópia és gépi tanulás a fehérje folding és aggregáció gyors és nagy mennyiségben történő vizsgálatára?

- Mely tényezők vannak összefüggésben az egyes aggregátumtípusok kialakulásával? Tudunk-e eredményeink alapján következtetni az aggregáció két típusának mögöttes mechanizmusaira, meg tudjuk-e érteni, hogy milyen erők döntenek egy fehérje sorsa felől?

- Milyen fehérje tulajdonságok segítenek a szolubilitás megőrzésében overexpressziós körülmények között? Milyen viszonyban van a szerkezet és a felszín a fehérje oldhatósággal, milyen módon és mekkora mértékben különböznek az oldható fehérjék az egyes aggregációs csoportoktól?

- Általánosíthatóak-e eredményeink a heterológ fehérjék expressziójának esetére? Van-e olyan oldhatóságot befolyásoló paraméter, amely nem *E. coli* fehérjék esetén is kiemelkedő fontossággal bír?

Céljaink megvalósítása érdekében nagy áteresztőképességű mikroszkópiás módszertannal vizsgáltuk az *Escherichia coli* egyes fehérjeinek aggregációs fenotípusát. Ezt követően az összes fehérjére kiterjedő, számos paramétert felölelő adatsort hoztunk létre, az irodalom, adatbázisok és predikciós algoritmusok segítségével, majd az egyes fenotípusokhoz tartozó különböző fehérjetulajdonságok eltérései alapján, bioinformatikai és statisztikai módszerekkel kerestük a magyarázatot a fenti kérdésekre.

2. Anyagok és módszerek

2.1. Az ASKA expressziós könyvtár használata

A fehérjeaggregáció mikroszkópos vizsgálatát a Horváth Péter és a Pál Csaba Laboratóriumok közreműködésével végeztük. A kísérletek során az ún. ASKA overexpressziós könyvtárat használtuk (Kitagawa és mtsai., 2005), amely minden tagja egy plazmidon, IPTG-vel indukálható módon fejezi ki az *E. coli* egy fehérjét. Minden konstrukció esetében a C-terminálison fúziós formában zöld fluoreszcens fehérjét (GFP) tartalmazó változatot használtuk. A kísérletek során a törzseket 96-lyukú lemezekben növesztettük 37°C-on, 280 rpm-es rázás mellett LB tápfolyadékban, amely 20 µg/ml kloramfenikollal mint szelekciós markerrel volt kiegészítve. Az egy éjszakán át tartó növesztést követően a génexpressziót 2 órán át 0,1 mM IPTG-vel indukáltuk a teljesen felnőtt tenyészetek esetében, 37°C-on. A mintákból 0,2 µL-t helyeztünk a CellCarrier-96 típusú (PerkinElmer) lemezekre. Ezt követően a lemez minden lyukába 100 µl 5 µg/ml-es koncentrációjú 4,6-diamidino-2-fenilindolt (DAPI) adtunk MS minimálmédiumban feloldva, szénforrás nélkül. A mikroszkópos vizsgálatot megelőzően a sejteket a lemez alján centrifugálással összegyűjtöttük.

2.2. Nagy áteresztőképességű fluoreszcens egysejt-mikroszkópia

A képkészítés PerkinElmer Operetta mikroszkóppal történt. Lyukanként négy helyről történt a felvételek készítése, minden pozícióban lézer alapú képfókuszálás történt. Két külön kép készült két csatornán (a DAPI és a GFP segítségével), 60×-os nagyítású objektívvel, ezáltal az egyes sejtek és a bennük előforduló aggregátumok elhelyezkedése is megfigyelhető volt. Minden vizsgálati pozíció esetében, 0,5 µm-es eltolással öt különböző kép készült a z-tengely mentén, amelyek segítségével a perfect focus algoritmus létrehozta a legnagyobb kontraszttal bíró felvételeket. A képek további feldolgozása a CIDRE algoritmussal történt (Goodsell és Olson, 2000), amely az egyenetlen világítás kiszűréséért volt felelős. A sejtek felismerése és a paramétereik kinyerése a CellProfiler (Carpenter és mtsai., 2006) programmal történt. A sejtek azonosítása a DAPI-jel alapján történt az Otsu adaptív határérték alkalmazásával, az összeérő sejtek pedig a

Watershed algoritmus segítségével kerültek elválasztásra. Minden vizsgált törzs esetében kb. 1000 sejt adatait nyertük ki az elkészült képekből, különös tekintettel a GFP szignál jelenlétére és sejten belüli eloszlására. A sejtklasszifikáció felügyelt gépi tanulási módszerrel történt az Advanced Cell Classifier szoftver segítségével (Piccinini és mtsai., 2017). Három különböző GFP fenotípust figyeltünk meg: a GFP szignál teljes hiányát (a fluoreszcencia szint megegyezett a negatív kontroll értékeivel), a homogén GFP-jelet (az egész sejt területén egyenletes fluoreszcenciaintenzitás) és a koncentrált GFP-jelet, amely a sejt vagy egyik vagy mindkét pólusán jelent meg. Azok a sejtek, amelyek egyik kategóriába se illettek bele, eldobásra kerültek. A klasszifikáció a Weka szoftver neurális háló alapú módszerével történt (Hall és mtsai., 2009). Minden fehérje a dominánsan előforduló fluoreszcencia-fenotípus alapján került besorolásra. A vizsgálat során 3706 *E. coli* protein aggregációs és folding tulajdonságait elemeztük.

2.3. A fehérjeexpresszió és az *in vivo* aggregáció kísérletes ellenőrzése

A fehérjeexpressziót és az aggregációt ellenőrző western blot vizsgálatokat Balogh Dávid, Őszi Erika és Magyar Zoltán végezte a projekt során, az eredmények statisztikai kiértékelését pedig én végeztem. A sejtklasszifikáció során három markánsan eltérő sejttípust azonosítottunk, amelyek közül az egyik a fluoreszcencia hiányát mutató a sötét aggregátumot tartalmazó sejtek voltak. Így szükséges volt annak ellenőrzése, hogy nem a génextpresszió hiánya miatt marad-e el a GFP-jel megjelenése. Az expressziós vizsgálatokhoz egy 128 tagból álló, véletlenszerűen kiválasztott génszettet használtunk, amelynek döntő többsége a sötét aggregátum csoportba esett (lásd 1. számú táblázat). Az expressziós vizsgálatok során western blot kísérletekben a kiválasztott törzseket egy éjszakán át növesztettük 37°C-on, 1ml LB médiumban, amely 20 µM kloramfenikollal volt kiegészítve. A növesztést követően a sejteket 2 óráig 0.1 mM-os IPTG-vel indukáltuk, szintén 37°C-on. Ezt követően a sejteket centrifugálással (kb. 13000 g nehézségi erő mellett) összegyűjtöttük, majd a pelleteket 250 µl 2xSDS (4%) pufferben szuszpendáltuk. Ezt követően 5 percig forraltuk a mintákat, majd 5 µl-t felvittünk 10%-os SDS-poliakrilamid gélre (PAGE). A géleket megfestettük Coomassie Brilliant Blue festékkel, vagy PVDF membránra (Amersham, GE Healthcare Lifescience) blottoltuk. A membránokat

blokkoltuk 5 tömegszázalékos tejport és 0.05 térfogat-százalékos Tween20-at tartalmazó TBS pufferben (TBST puffer: 25mM Tris-Cl, pH 8.0, 150 mM NaCl) egy órán keresztül szobahőmérsékleten. Ezt követően a membránokat 5 tömegszázalékos tejport és elsődleges anti-GFP antitesteket tartalmazó TBST pufferben egy éjszakán át inkubáltuk. Ezt követően TBST pufferrel háromszor mostuk a mintát a fölöslegben maradt elsődleges antitestek eltávolítása érdekében, majd a membránokat a másodlagos antitestek (1:10000 hígítás, Sigma-Aldrich) jelenlétében TBSB-ben (25mM Tris-Cl, pH 8.0, 150 mM NaCl) inkubáltuk egy órán át szobahőmérsékleten. Ezt követően újból háromszor átmostuk a membránokat TBST pufferrel és standard kemilumineszcens módszerrel előhívtuk (Thermo Scientific).

Az aggregáció vizsgálatához 46 fehérjét választottunk ki (lásd 1. táblázat). A sejtek hasonló módon növesztettük, mint az expressziós vizsgálatok esetében. A centrifugálás lépése után a pelleteket 200 µl BugBuster (Sigma) pufferben oldottuk föl szobahőmérsékleten. Az oldatokat 16 000 g nehézségi erővel 20 percig centrifugáltuk 4°C-on. A 200 µl-es felülúszót eltávolítottuk és 50 µl 5xSDS (10%) pufferrel egészítettük ki (szolubilis frakció). A pelleteket 250 µl 1xSDS (2%) pufferben szuszpendáltuk föl (aggregálódó frakció). A mintákat 5 percig forrásban tartottuk, majd 5-5 µl-t vittünk föl 10%-os SDS-poliakrilamid gélekre. A western blotot az előzőekben leírt módon hajtottuk végre. A gélképeken megjelenő intenzitásvizsgálathoz a képeket fekete–fehér árnyalatúvá alakítottuk, a degradációból származó termékeket figyelmen kívül hagytuk. A kvantifikációhoz az Image Studio Light szoftvert használtuk, a kapott intenzitásértékeket a háttérzajra korrigáltuk.

2.4. A fehérje- és génjellemzőket tartalmazó adatsor összeállítása

Ahhoz, hogy feltárhassuk a fehérjeaggregációs fenotípusok kialakulása mögött rejlő tényezőket, szükséges volt az összes rendelkezésünkre álló fehérjeparaméter felhasználására vizsgálataink során, amelyeket egy egységes adatsorba rendeztünk. A fehérjék sejten belüli lokalizációját a StepDB adatbázis segítségével gyűjtöttük össze (Orfanoudaki és Economou, 2014), az oligomerizációra vonatkozó adatokat pedig az EcoCyc adatbázisból (Keseler és mtsai., 2017). A fenotípusok gyakoriságának meghatározásán túl a további elemzésekhez kizárólag a citoplazmatikus monomer

fehérjéket használtuk, amelyekből 1631 volt az adatsorunkban. Erre két okból volt szükség: a fehérjék sejten belüli lokalizációja erős összefüggést mutatott az általuk alkotott aggregátumok típusával, az oligomerizáció pedig negatív összefüggésben állt az oldhatósággal. Ahhoz, hogy ezeknek a paramétereknek az ismert torzító hatását kizárjuk vizsgálatainkból, a későbbiekben csak a citoplazmatikus monomereket használtuk. 115 fehérje és génjellemzőt gyűjtöttünk össze, ideértve az mRNS és a fehérjék natív expressziós szintjét, a fehérje–fehérje interakciók jelenlétét, a fehérjerendezetlenség mértékét, fizikai–kémiai jellemzőket és funkcionális besorolásokat (lásd 2. táblázat). A felszíni és belső aminosav-gyakoriságok, valamint a felszíni „ragadósság” (Levy, 2010) (protein stickiness) meghatározásához a fehérjeszerkezeti adatokat Zhang Laboratórium repozitóriumából gyűjtöttük (Xu és Zhang, 2013). A strukturális adatok a CATH adatbázisból (Sillitoe és mtsai., 2019) származtak, a fehérjerendezetlenség pedig a MobiDB adatbázisból (Piovesan és Tosatto, 2018), amely különböző algoritmusok adatait tartalmazta, mint a DisEMBL (Linding és mtsai., 2004), PONDR VSL2B (Peng és mtsai., 2006), iUpred (Dosztányi és mtsai., 2005) és az Espriz (Walsh és mtsai., 2012). Az rendezetlen aminosavak számát minden esetben normáltuk a fehérje hosszára. Az összes további elemzéshez felhasználtuk mind az 1631 fehérjét, amely alól csak a folding ráta (Gromiha és mtsai., 2006) és a contact order (Plaxco és mtsai., 1998) volt a kivétel. Ezek az algoritmusok megbízható eredményt csak az egydoménes monomer fehérjék esetében adnak, így a többit eltávolítottuk az adatsorból, ami 661 fehérjét eredményezett. Az ROC-görbe alatti terület számításához a pROC nevű R csomagot használtuk 10-szeres keresztvalidációval. Az enzimek által végrehajtott kémiai átalakulásokat leíró Enzyme Commission kategorizálást a UniProt adatbázisból gyűjtöttük.

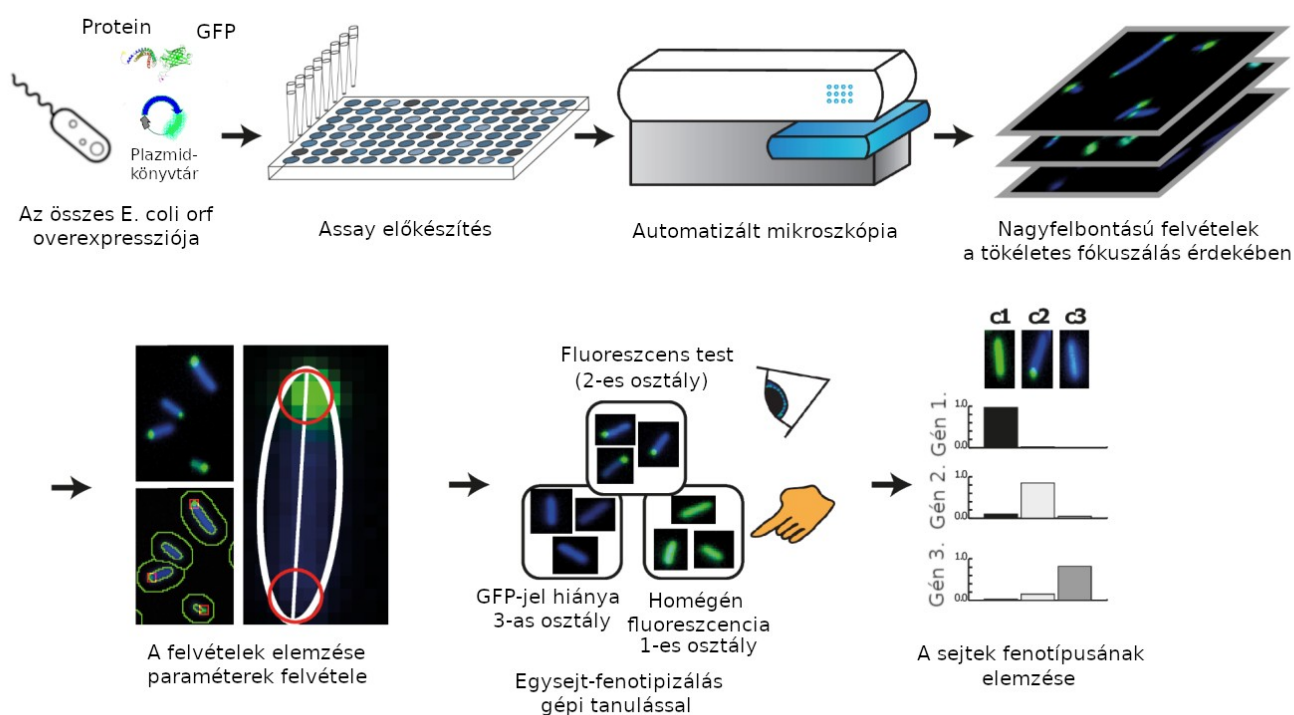
A heterológ expresszióval megtermelt fehérjékre vonatkozó szolubilitási adatokat a Northeast Structural Genomics Consortium (Price és mtsai., 2011) adatbázisából gyűjtöttük. Ez kb. 9000 fehérje *E. coliban* történt expressziójával készült, azonban számos redundanciát tartalmazott. Ezért kizártunk minden szekvenciát, amely 100%-ban megegyezett bármely másikkal, valamint azokat, amelyek legalább 99%-os szekvenciahasonlóságot mutattak bármely *E. coli* fehérjével. Ez utóbbi azért volt szükséges, hogy a saját méréseinktől független adatsoron validálhassuk a fehérjerendezetlenségre vonatkozó eredményeinket. Ezen felül létrehoztunk egy adatsort

olyan fehérjékből, amelyek 100%-os szekvenciahasonlóságot mutattak az emberi fehérjékkel, annak érdekében, hogy egy jó minőségű, eukarióta eredetű fehérje szetten is ellenőrizzük állításainkat. A szekvenciahasonlóságok számításához a BlastP algoritmust használtuk, az elvárt E-értékeknek minden esetben kisebbnek kellett lenniük mint 10^{-10} , a szekvenciák lefedettségének pedig legalább 80%-osnak kellett lenniük.

Minden egyéb számoláshoz az R 3.5.0 2018-04-23-as verzióját használtuk (R Core Team, 2008) az Rstudio 1.1.447-es verziójába ágyazva. Az ábrák elkészítését az említett R szoftver beépített funkcióival, valamint a 2.2.1-es verziószámú ggplot2 csomag segítségével végeztük.

3. Eredmények

A rendelkezésünkre álló szisztematikus, *in vivo* proteomikai adatok segítségével megvizsgáltuk a fehérjeaggregációt befolyásoló tényezőket. Ennek első lépéseként áttekintettük az egyes aggregációs fenotípusok relatív gyakoriságát és a besorolásuk robusztusságát, és összevetettük adatainkat a korábbi nagy áteresztőképességű *in vitro* eredményekkel. Ezután az általunk összeállított szisztematikus adatsor segítségével megvizsgáltuk, hogy mely tényezők melyik aggregációs fenotípus kialakulásához vezetnek, mi határozza meg, hogy a fehérje aggregáció a folding előtt vagy után következik be. Ezt követően pedig azt teszteltük, hogy vajon milyen fehérje paraméterek függenek össze a szolubilitással, végül pedig eredményeink egy részét független adatsoron validáltuk.



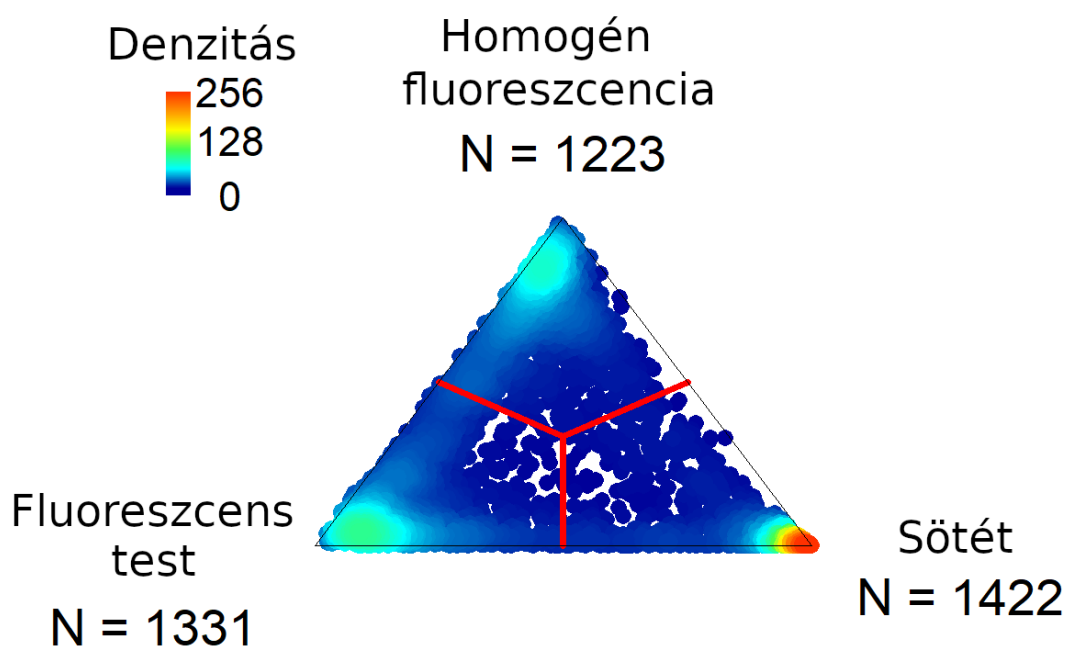
3. ábra A fehérje aggregáció nagy áteresztőképességű mikroszkópiás vizsgálata gépi tanulós klasszifikációval. A munkamenet az *E. coli* overexpressziós könyvtár (ASKA), növesztésével kezdődött (Kitagawa és mtsai., 2005). Az egyes fehérjéket 96 lyukú platekben egy éjszakán át növesztettük, majd DAPI hozzáadásával mikroszkópiás képeket készítettünk. A legnagyobb kontraszttal bíró képeken ezután a CellProfiler algoritmussal azonosítottuk a sejteket,

valamint a fluoreszcens jelet. A sejtek klasszifikációja az Advanced Cell Classifier szoftverrel történt, amelyhez felhasználói inputot adtunk az egyes fenotípusok felismeréséhez. A sejtek klasszifikációját követően a leggyakoribb sejtes fenotípust használtuk a fehérjék fenotípusának leírására.

3.1. Két aggregációs fenotípus különböztethető meg overexpresszió esetén *E. coliban*

Vizsgálataink során 3706 natív *E. coli* fehérje szolubilitását vizsgáltuk sejtszinten nagy áteresztőképességű mikroszkópiás és gépi tanulós módszerrel (3. ábra, lásd módszerek). Mivel a C-terminálison fehérjéhez kötött fúziós GFP-ről ismert, hogy indikálja a fúziós partnere folding állapotát (Waldo és mtsai., 1999), a fluoreszcens jel megléte és eloszlása árulkodik a vizsgált fehérje oldhatóságáról és foldingjáról. Három különböző fenotípust azonosítottak korábbi vizsgálatok során: amennyiben a fluoreszcens jel diffúzan a teljes sejtben megfigyelhető, akkor a fehérje átesett a folding folyamatán és szolubilis formában van jelen a sejtekben (Natan és mtsai., 2018). Amennyiben a sejt egyik, vagy mindkét pólusán figyelhető csak meg a fluoreszcens jel, abban az esetben az aggregátumok natív szerkezetű GFP jelenlétére utalnak (Hedhammar és mtsai., 2005; Wagner és mtsai., 2007). Ez arra utal, hogy a GFP az aggregációt megelőzően fölvette natív vagy natívszerű konformációját. És mivel számos bizonyíték szolgál arra, hogy a fúziós GFP jó indikátora az N-terminálison lévő protein folding állapotának (Bock és mtsai., 2012; Shmueli és mtsai., 2013; Waldo és mtsai., 1999; Zhang és mtsai., 2005), arra következtethetünk, hogy az N-terminálison lévő fehérjék is elérték ezt az állapotot. Számos fehérjéről bebizonyosodott, hogy az aggregátumban is képesek megőrizni az aggregáció előtt felvett szerkezetüket, így potenciálisan enzimatis aktivitással is bírhatnak (García-Fruitós és mtsai., 2005; de Groot és Ventura, 2006). Ezt az aggregátumtípust fluoreszcens aggregátumnak neveztük a megfigyelt fenotípus alapján. Ez megfeleltethető a posztfolding aggregátumnak, ahol a folding és az aggregáció egymás után következik be, bár maguk az aggregátumok többféle szerkezetű fehérjét is tartalmaznak, olyanokat is, melyek nem feltétlen rendelkeznek natív szerkezetükkel (García-Fruitós és mtsai., 2005; Simona és mtsai., 2005). A harmadik leírt fenotípus nem mutat fluoreszcens jelet, amelyből arra következtethetünk, hogy a fehérjék a folding

folyamatát megelőzően, vagy annak során aggregálódtak, ezeket sötét- vagy prefolding aggregátumoknak nevezzük (Natan és mtsai., 2018). A sejtszintű mikroszkópiás adatokat felügyelt gépi tanulás módszerével elemeztük, amely során kb. 1000 sejtet vizsgáltunk minden expressziós törzs esetében. A fehérjék fenotípusaként a legnagyobb gyakorisággal előforduló sejtes fenotípust határoztuk meg. Az *E. coli* proteom tagjai összevethető gyakorisággal estek az általunk meghatározott három oldhatósági csoportba (4. ábra): a fehérjék 29,6%-a szolubilis volt az erős overexpresszió ellenére, 33,9%-uk a fluoreszcens aggregáló csoportba került és 36,5%-uk a sötét aggregáló csoportba. Eredményeinkből tehát következik, hogy a fehérjék közel 1/3-a nincs az oldhatósági limit közeli koncentrációban, hiszen nagy mértékű overexpresszió esetén is oldható maradt. Ez az arány fiziológiai körülmények között feltehetően még magasabb, így eredményeink alapján a „Life on the edge” hipotézis nem tekinthető univerzálisan igaznak a teljes proteomra nézve.

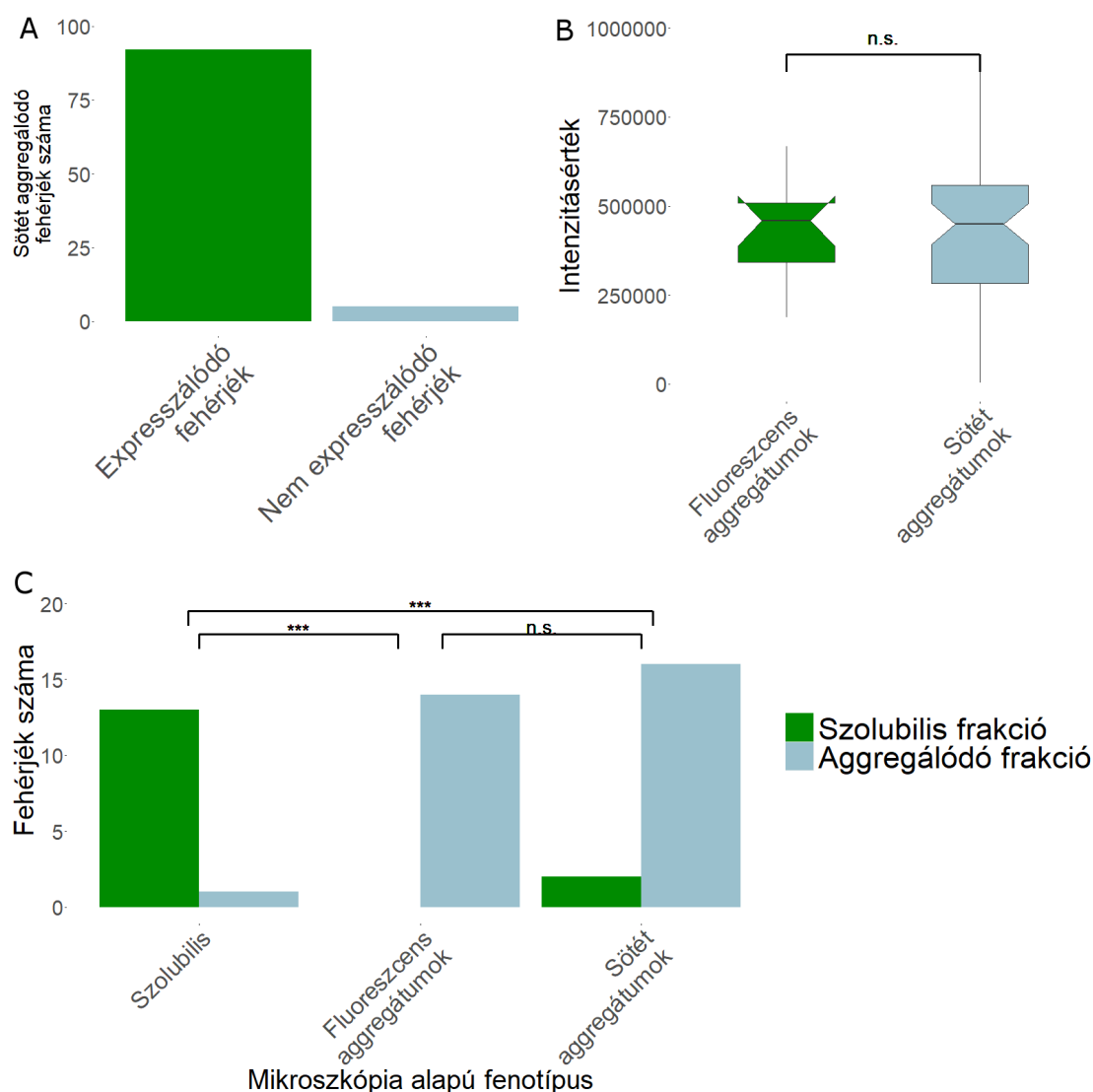


4. ábra Az egyes sejtfenotípusok és az általuk meghatározott fehérjefenotípusok gyakorisága. Minden fehérje aggregációs fenotípusának meghatározásához kb. 1000 sejt fenotípusát használtuk. Amennyiben a vizsgált sejtek 100%-a egy adott fenotípust mutatott, abban az esetben az ábrán az adott fenotípushoz tartozó csúcsba került. A csúcsoktól mutatott eltávolodás arányos a sejtfenotípusok arányával, amelyből látható, hogy a legtöbb overexpressziós törzs döntő többségben egyféle sejtfenotípust mutatott, tehát a klasszifikáció

robosztus. A színekkel reprezentált denzitásból pedig leolvasható, hogy mindhárom fehérje fenotípus közel azonos mennyiségben reprezentálódott a proteomban. A piros vonalak a döntési határokat jelölik.

3.2. A nagy áteresztőképességű mikroszkópiás módszer eredményeinek validációja

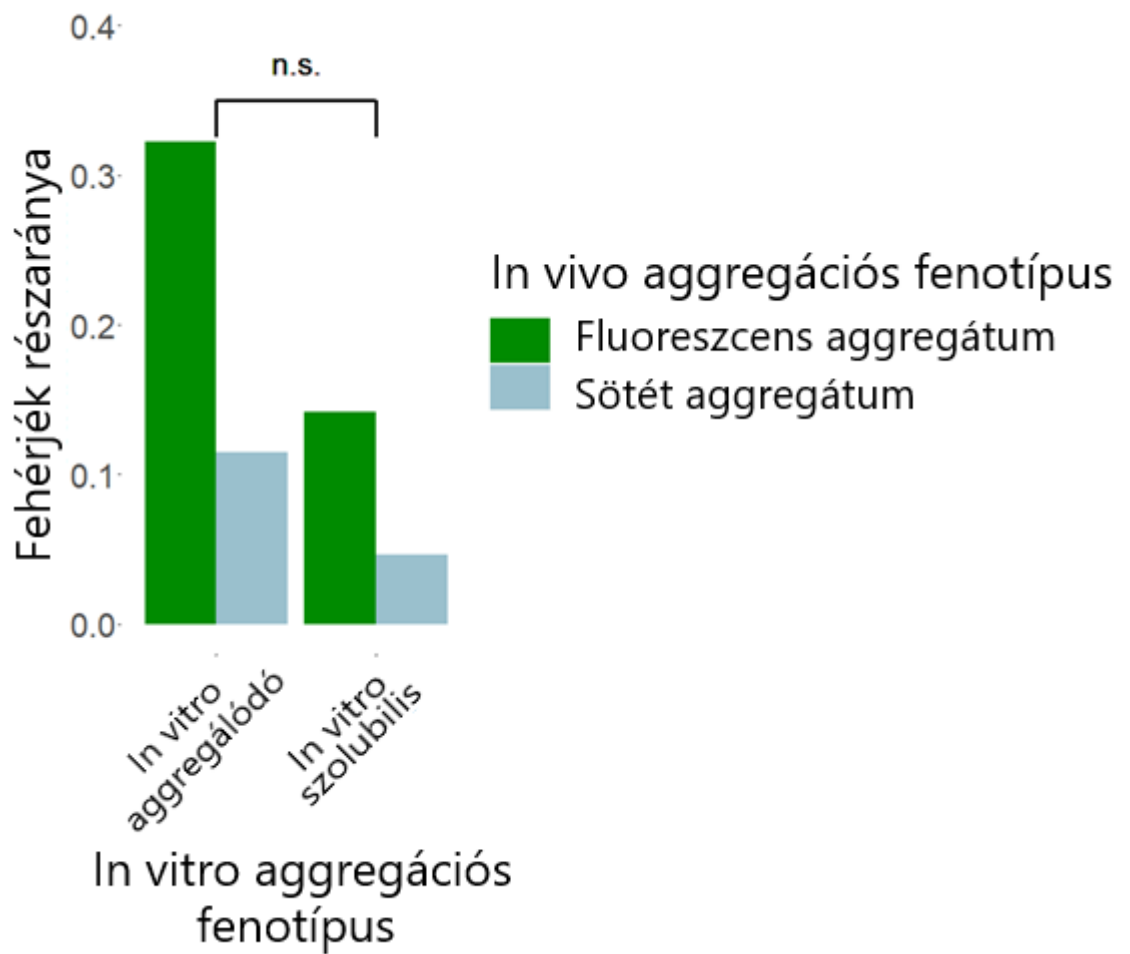
Mivel a nagy áteresztőképességű mikroszkópiás módszer a fluoreszcens jel alapján közvetve ad információt a fehérjék aggregációs állapotáról, szükségesnek találtuk az eredmények validációját. Ehhez két párhuzamos megközelítést alkalmaztunk. Az első megközelítés során egy kisebb fehérje szettet vizsgáltunk western blot módszerrel expresszió és aggregáció szempontjából. Ennek során főleg a sötét aggregátumot képző fehérjéket vizsgáltuk, ugyanis ezek esetében a fluoreszcens jel hiánya az overexpressziós konstrukciók nem megfelelő működésére is utalhat: amennyiben a fehérjék expressziós szintje nem éri el a detekciós limitet, az a prefolding aggregátumokhoz hasonlóan a detektált GFP-jel hiányához vezethet. Kísérleteink azonban azt mutatták, hogy a vizsgált sötét aggregálódó fehérjék 95%-a esetében (97-ből 92 fehérje esetén) megfelelő, a fluoreszcens aggregátumokkal összemérhető volt az expressziós szint (5. ábra A és B panel). Továbbá a sötét aggregátumot képző fehérjék 88,9%-a a western blot kísérletek során csak az aggregálódó fázisból volt kimutatható (5. ábra C panel).



5. ábra. Az overexpressziót követő aggregációs fenotípusok kísérletes validációja. (A) A mikroszkópos vizsgálatok alapján sötét aggregáló fehérjeként besorolt 97 proteinen végzett western blot kísérletek eredményei. A western blot analízis alapján a fehérjék 94,8%-a expresszálódik detektálható mennyiségben. (B) A detektált expresszálódó fehérjékhez tartozó intenzitásértékek a sötét- és fluoreszcens aggregáló csoportban. Nem mutatható ki szignifikáns különbség a két csoport fehérjeinek detektált intenzitásértékeiben ($P = 0,8794$, $N = 71$, kétoldali Wilcoxon rangösszeg próba, lásd módszerek). (C) A szolubilis és aggregáló frakciók elemzése SDS-PAGE western blot módszerrel a mikroszkópos vizsgálatok validációja érdekében. 14 szolubilis, 14 fluoreszcens aggregáló és 18 sötét aggregáló fehérjét véletlenszerűen kiválasztottunk a kísérletek során, egy fehérje fenotípusát akkor tekintettük aggregálódnak, amennyiben a szolubilis frakcióban nem volt megtalálható a proteinből az SDS-PAGE western blot elemzés során (lásd módszerek). A mikroszkópos vizsgálatok alapján szolubilisként osztályozott fehérjék egyértelmű feldúsulást mutatnak a szolubilis frakcióban az aggregáló fehérjékhez

képest ($P = 7,48 * 10^{-7}$ és $P = 5,26 * 10^{-6}$, $N = 46$ és 46 , kétoldali Fisher egzakt teszt). A sötét- és a fluoreszens aggregátumokat mutató fehérjék hasonló gyakoriságot mutatnak az aggregálódó és szolubilis frakciókban ($P = 0,49$, $N = 32$, kétoldali Fisher egzakt teszt).

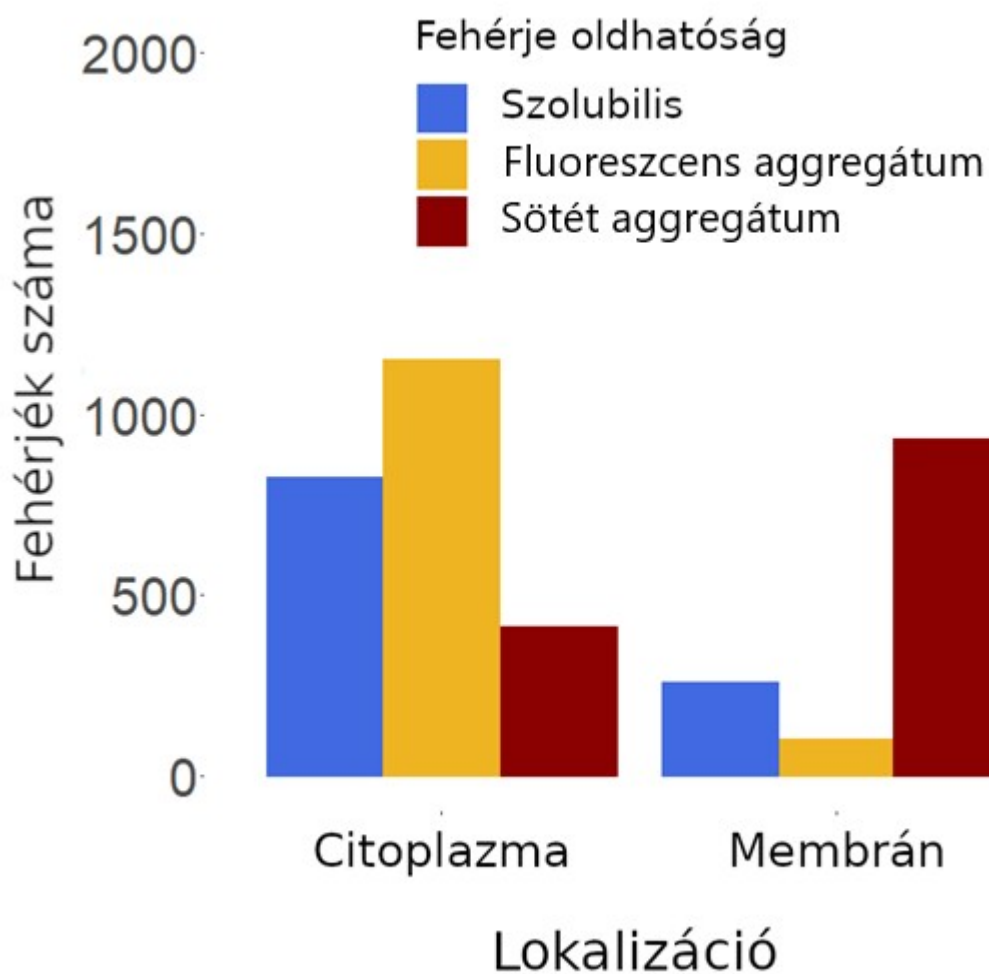
A másik megközelítés során adatainkat korábbi *in vitro* transzlációs körülmények között végzett aggregációs kísérletek eredményeivel vetettük össze (Niwa és mtsai., 2009). Annak ellenére, hogy az *in vitro* és az *in vivo* körülmények jelentősen különböznek, nagy mértékű átfedést tapasztaltunk a fehérjék aggregációs viselkedésében (esélyhányados = 8,23, $P < 10^{-10}$, Fisher-egzakt teszt). Továbbá mindkét *in vivo* aggregációs fenotípus hasonlóan feldúsul az *in vitro* aggregálódó fehérjék körében ($P = 0,1873$, Fisher egzakt teszt, 6. ábra). Ez azt jelenti, hogy azok a fehérjék, amelyek az *in vitro* vizsgálatok során aggregátumként jelentkeztek egyformán reprezentáltak az általunk vizsgált két aggregációs csoportban, nincs szisztematikus hatás egyik javára sem, tehát egyforma érzékenységgel tudjuk azonosítani őket. Ezek a vizsgálatok azt is megerősítik, hogy a GFP-fúzió, amit kísérleteink során használtunk, nem befolyásolta a fehérjék aggregációs viselkedését.



6. ábra In vitro és in vivo aggregációs fenotípusok összehasonlítása. In vitro aggregációs mérések során nem volt lehetőség a különböző aggregációs fenotípusok elkülönítésére, mivel a konstrukciók nem tartalmaztak GFP vagy egyéb, szerkezeti információt hordozó riportert. A nagy áteresztőképességű in vivo mikroszkópiás módszerünk ezzel szemben két eltérő fenotípust különböztetett meg. A két in vivo aggregáció típus azonos mértékben dúsul fel az in vitro aggregálódó fehérjék körében, ezért a GFP fúziós fehérje nincs hatással a mérés során az aggregációs fenotípusra.

3.3. A fluoreszcens aggregátum fenotípusa domináns citoplazmatikus fehérjék esetén, míg a sötét aggregátumé a membránfehérjéknél

A fehérjék normál expressziós szint mellett lokalizációjára vonatkozó adatok felhasználásával elemeztük, hogy az egyes fehérje aggregációs fenotípusok összefüggenek-e a fehérjék előfordulásával. Míg a membránfehérjék 69%-a a folding folyamatát megelőzően aggregálódik, a citoplazmatikus fehérjék esetén ez az arány csupán 17,2%. Ugyanakkor a citoplazmatikus fehérjék 34,6%-a szolubilis marad a nagyfokú overexpresszió ellenére is és 48,1%-uk a foldingot követően aggregálódik (lásd 7. ábra). Ezek az eredmények összhangban állnak a korábbi feltételezéssel, mely szerint a membránfehérjék a foldingot és a membránba épülést megelőzően, míg a citoplazmatikus fehérjék inkább a foldingot követően aggregálódnak, amennyiben átlépik az oldhatósági határértéküket (Ciryam és mtsai., 2013). Eredményeink továbbá arra is rávilágítanak, hogy az aggregálódó citoplazmatikus fehérjék 70%-a a folding folyamatát követően aggregálódik, annak ellenére, hogy a 37°C-os növesztési hőmérséklet nem kedvez ennek a típusú aggregációnak (de Groot és Ventura, 2006). Mégis, a fluoreszcens aggregáció a domináns módja az aggregátumképzésnek az *E. coli* citoplazmatikus fehérjéi esetében.

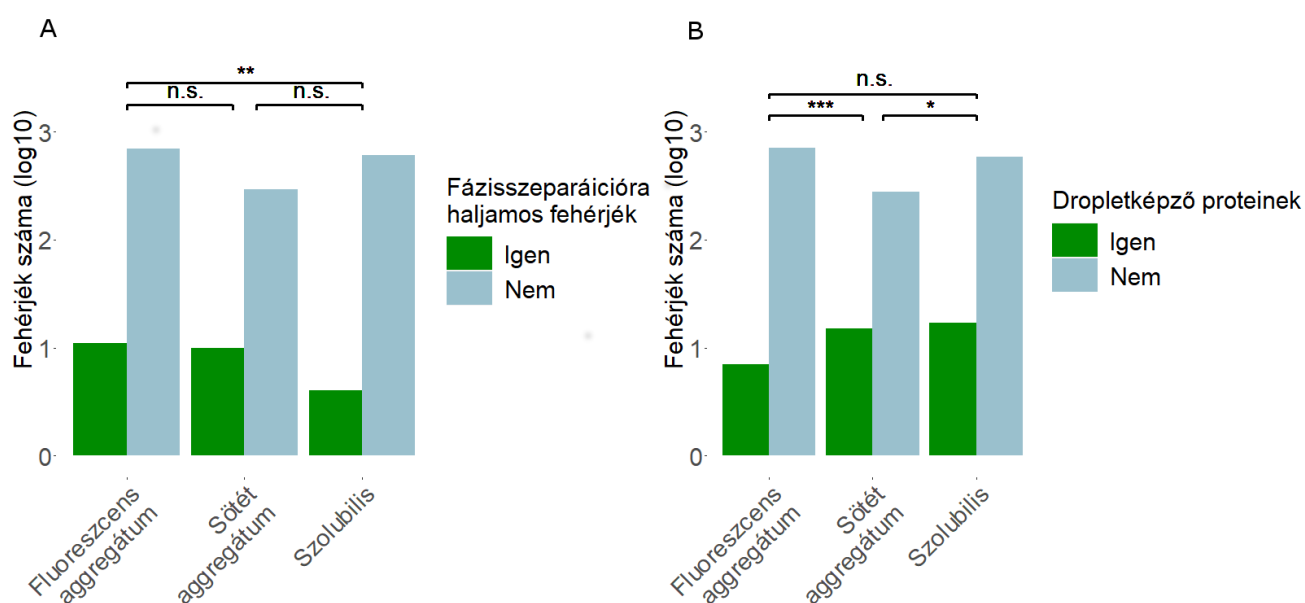


7. ábra. A különböző aggregációs fenotípusok gyakorisága a sejtes lokalizációjuk függvényében. Míg a membránfehérjék 69%-a a foldingot megelőzően aggregálódik, így nem mutat fluoreszcens jelet, a citoplazmatikus fehérjék esetében ez az arány csupán 17% ($P < 10^{-10}$, kétoldali Fisher egzakt próba).

3.3.1. A fázisszeparáció nem bír döntő jelentőséggel az *in vivo* aggregáció folyamatában *E. coliban*

A sötét és fluoreszcens aggregátumok mellett a fázisszeparált aggregátumok kialakulása is állhatna megfigyeléseink mögött. Mivel a fázisszeparált granulumok rendelkezhetnek natív szerkezetű fehérjékkel, megjelenésükben leginkább a fluoreszcens aggregátumokra hasonlíthatnak. Így a fluoreszcens jel jelenthetné a fázisszeparáció kialakulását a natív szerkezetüket felvett fehérjék esetében. Ezért két különböző

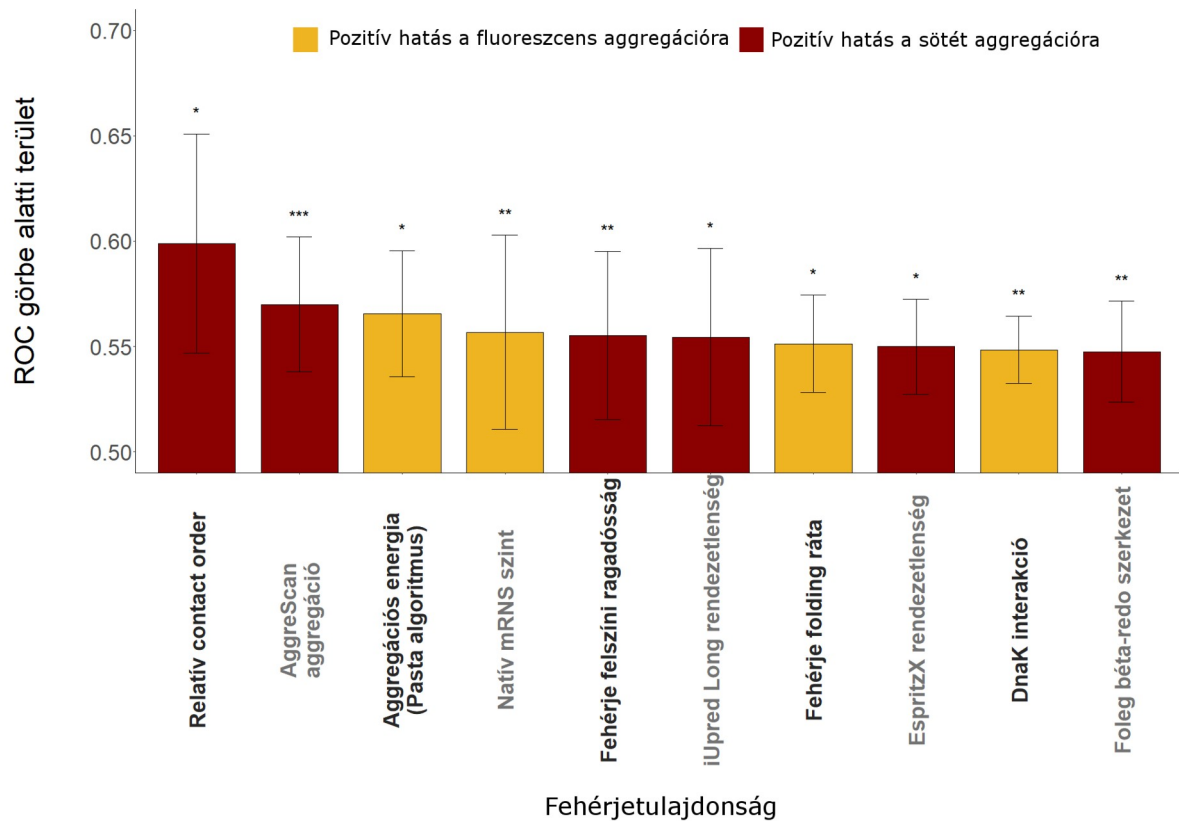
módszerrel vizsgáltuk, hogy az *E. coli* proteom mekkora része képes fázisszeparált granulomok létrehozására és feldúsulnak-e bármelyik aggregációs csoportban ezek a fehérjék. A PsPredict módszerrel kiszámítottuk az egyes fehérjék fázisszeparációs képességét (Sun et al, 2019), a FuzDrop módszerrel (Hardenberg et al, 2020) pedig azonosítottuk a dropletképző proteineket. Eredményeink azt mutatják, hogy az *E. coli* fehérjék 1,53 és 2,42%-a képez fázisszeparált aggregátumokat. Továbbá nem tapasztaltunk konzisztens feldúsulást sem a sötét, sem a fluoreszcens aggregációs csoportban, ezért nem ez a jelenség áll a megfigyelt fenotípus hátterében (8. ábra). Egyes esetekben a fázisszeparáció bírhat jelentősséggel az aggregátumképzésben, de kijelenthető, hogy döntő szerepe nem lehet a folyamatban, pusztán a fázisszeparációra képes fehérjék alacsony számának következtében.



8. ábra. A fázisszeparáció és dropletképző fehérjék gyakorisága az aggregációs csoportokban. (A) A fázisszeparációt képző fehérjék nem mutatnak feldúsulást a fluoreszcens aggregátumot alkotók körében ($P = 0,0906$, $N = 1014$, Fisher egzakt teszt). (B) A dropletképző fehérjék szignifikánsan alulreprezentáltak a fluoreszcens aggregálódó fehérjék között a sötét aggregálódó csoporthoz képest ($P = 0,0001$, $N = 1002$, Fisher egzakt teszt).

3.4. A sötét aggregátumok kialakulását meghatározó tényezők

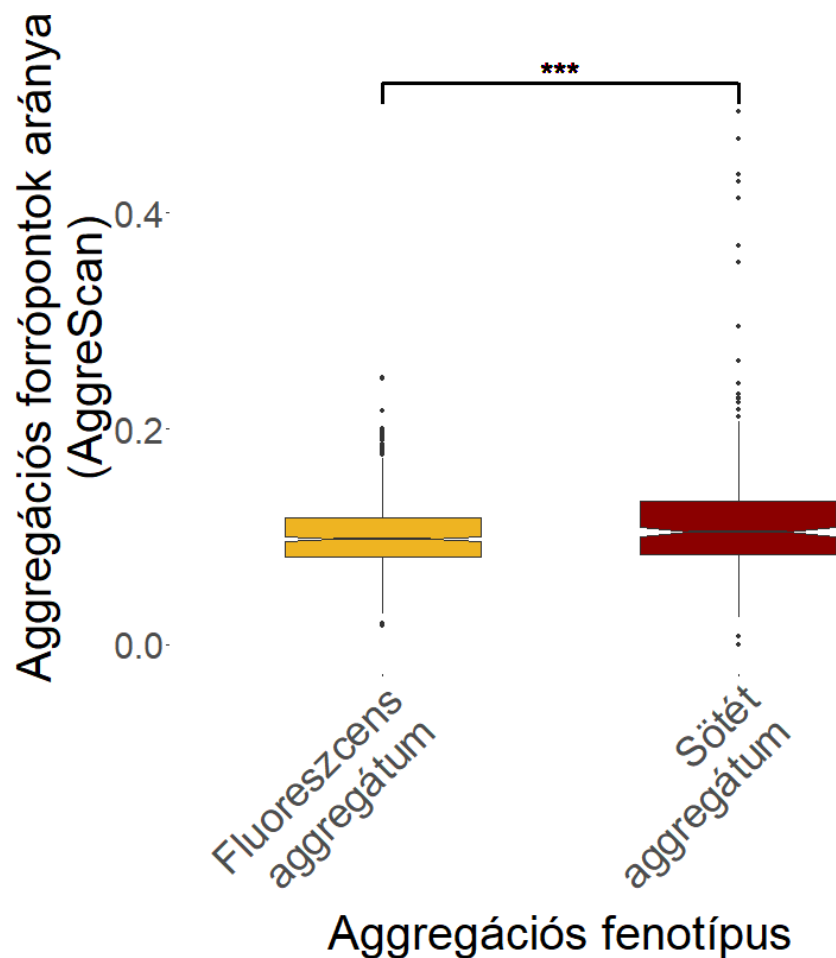
A fehérje aggregáció két útvonalát meghatározó tényezők megismerése érdekében összeállítottunk egy 115 fehérje tulajdonságból álló adatsort, amely különböző fizikai–kémiai, strukturális és funkcionális genomikai paramétert tartalmazott a citoplazmatikus fehérjék döntő többségére (2. táblázat). A fehérjék döntő többségére (93%) továbbá rendelkezésre állt state-of-the-art módszerrel készült háromdimenziós szerkezeti predikció (Xu és Zhang, 2013). A továbbiakban a citoplazmatikus fehérjékre fókuszáltunk, hogy elkerüljük az oligomerizáció folyamatával összefüggésben lévő tulajdonságok szisztematikus torzító hatását az aggregációs tulajdonságokra nézve (Pechmann és mtsai., 2009). Így 1631 fehérje maradt az adatsorunkban. Annak érdekében, hogy elkülönítsük a két aggregációs útvonalat a fehérjék molekuláris jellemzői alapján logisztikus regressziót hajtottunk végre az egyes paraméterek segítségével. A legerősebben diszkrimináló tulajdonságok a béta-aggregációra hajlamos szekvenciárészletek megléte, a fehérje felszínek ún. „ragadóssága”, a DnaK chaperonnal mutatott interakció és a fehérjék béta-redő struktúrát jelentős mértékben tartalmazó szerkezet volt (9. ábra). Számos további strukturális és funkcionális jellemző szerepel az adatsorunkban, azonban ezek közül egy sem mutatott szignifikáns különbséget a két aggregációs útvonal között. Egy enzimosztály sem dúsult fel a sötét aggregációs csoportban (lásd 3. táblázat), amiből arra következtethetünk, hogy a fehérjék aggregációjának módját befolyásoló tényezők többnyire függetlenek a szerkezettől és a funkciótól. Továbbá eredményeink arra utalnak, hogy bármely enzimosztály képviselője képes lehet fluoreszcens aggregátumot képezni, amely biotechnológiai szempontból rendkívül előnyös.



9. ábra. A pre- és a fluoreszcens aggregációs útvonalat megkülönböztető legfőbb fehérjeparаметerek. Az egyes paraméterek predikciós képességét a 10-szeresen keresztvalidált logisztikus regressziós modell ROC görbéje alatti területként határoztuk meg. Az ábrázolt fehérjeparаметerek statisztikailag szignifikánsak maradtak a p -érték FDR korrekcióját követően is. *** esetén $p < 0,001$, ** esetén $p < 0,01$, * esetén pedig $p < 0,05$ a logisztikus regressziós modellben. A hibasávok a görbe alatti terület 95%-os konfidenciaintervallumot ábrázolják.

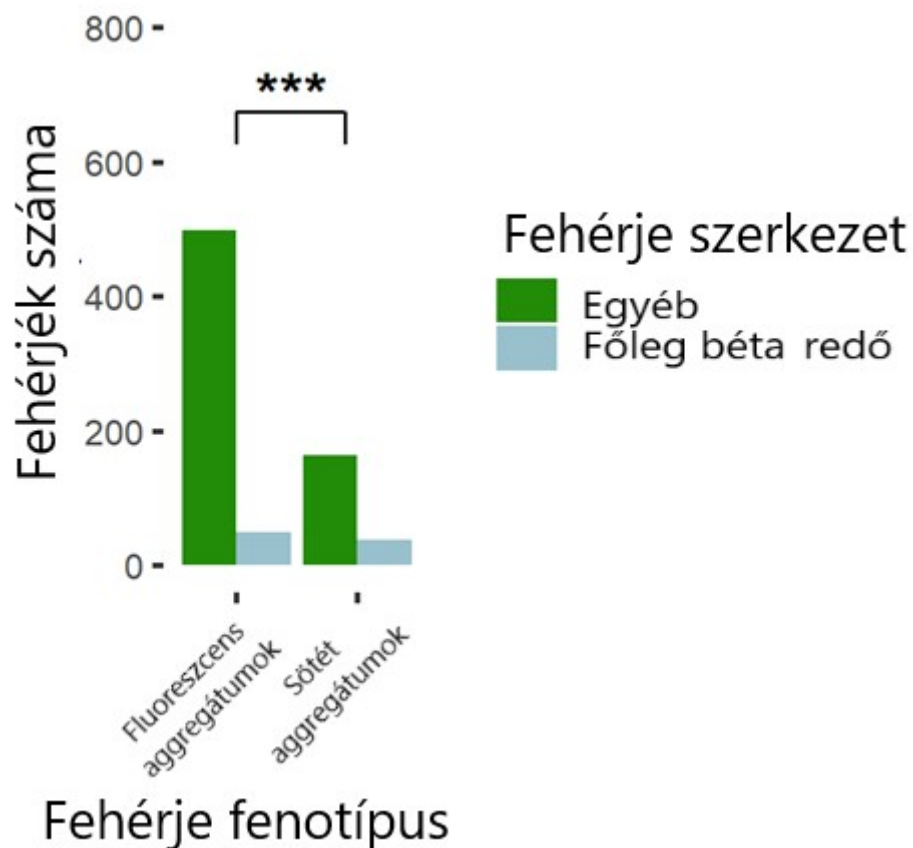
Korábbi eredmények arra mutattak, hogy amennyiben egy fehérje lassan aggregálódik, abban az esetben a folding folyamata lejátszódhat még az aggregátumok létrejötte előtt. Ez egyfajta kompetíciót hoz létre az aggregáció és a folding között (kinetikai kompetíció hipotézis), amely eredménye meghatározza az aggregátum típusát (de Groot és Ventura, 2006). Eredményeink három különböző módon támasztják alá a sötét aggregátumok magasabb aggregációs hajlandóságát az *E. coli* citoplazmatikus monomerek összességére.

- 1.) Az sötét aggregátumot képző fehérjék nagyobb mennyiségben tartalmazzak ún. aggregációs forrópontokat ($P = 5.01 \cdot 10^{-7}$, logisztikus regresszió, az AggreScan prediktor alkalmazásával, lásd az 2. számú kiegészítő táblázatot az egyéb algoritmusok eredményeiért). Ezek a szekvenciárészletek hajlamosak az amiloid aggregátumok képzésére, amennyiben a fehérje részben vagy teljes egészében kitekeredik, vagy eleve denaturált állapotban van (Chiti és mtsai., 2003; de Groot és mtsai., 2006).



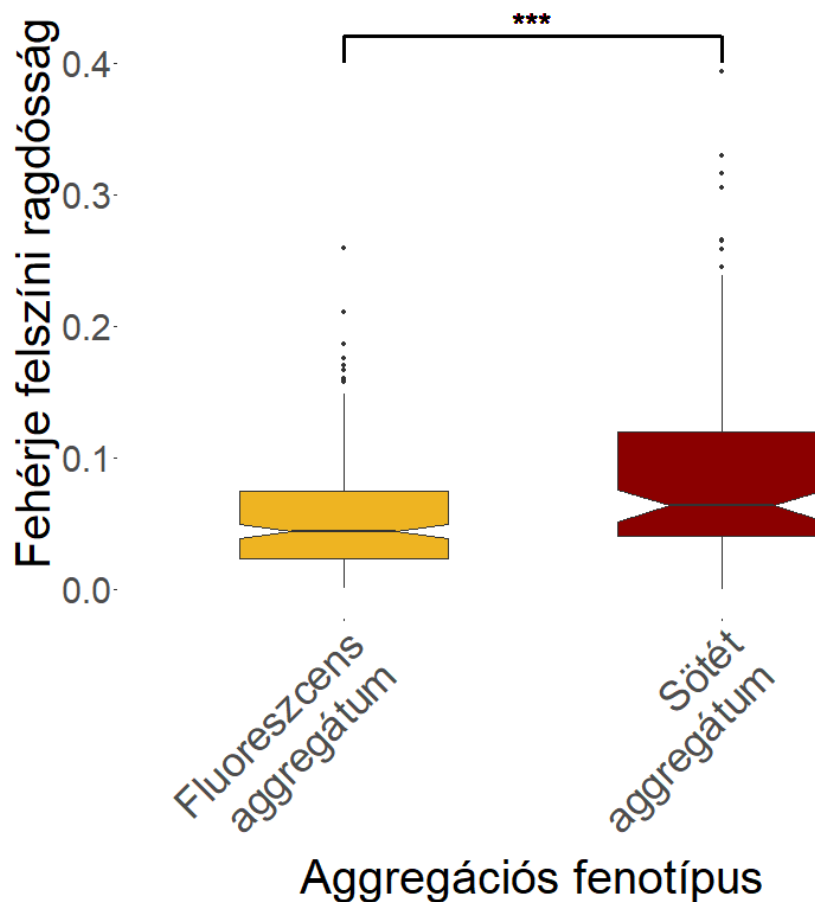
10. ábra. Az AggreScan (Conchillo-Solé és mtsai., 2007) algoritmussal számolt aggregációs forrópontok gyakorisága a két aggregációs csoportban. A sötét aggregációs csoport tagjai hosszra normálva szignifikánsan nagyobb mértékben tartalmaznak aggregációs forrópontokat ($P = 5.01 \cdot 10^{-7}$, logisztikus regresszió).

- 2.) A sötét aggregátumot képző fehérjék esetében szignifikánsan feldúsul a béta-lemezek aránya ($P = 0,0003$, logisztikus regresszió), amelyekről ismert, hogy a natív struktúra elérését megelőzően képesek az aggregációt megindító nukleuszként viselkedni (Bellesia és mtsai., 2009; de Groot és mtsai., 2006).



11. ábra. A főként béta szerkezeteket tartalmazó fehérjék gyakorisága a két aggregációs csoportban. A sötét aggregációs csoportban szignifikánsan magasabb a főleg béta szerkezetű fehérjék gyakorisága ($P = 0.0004$, Esélyhányados = 2.35, Fisher egzakt teszt).

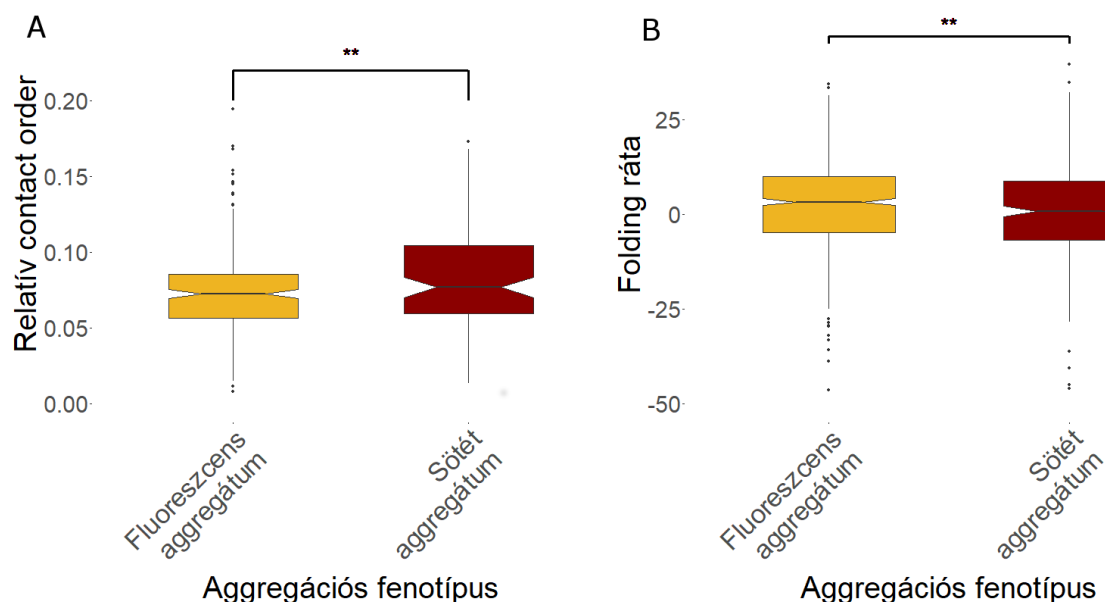
- 3.) A két aggregációs csoport szignifikánsan különbözik a fehérjék ragadósságában ($P = 0,00038$, logisztikus regresszió), amely a többi makromolekulával létrejövő nonspecifikus interakciók létrehozására való képességre utal (Levy és mtsai., 2012), egyfajta közelítésként a magasabb aggregációs hajlandóságra. Ezek az eredmények egybehangzóan alátámasztják a feltételezést, hogy a sötétaggregátum-képző fehérjék aggregációs képessége meghaladja a fluoreszcens aggregálódó fehérjékéét.



12. ábra. A sötét aggregációs csoportban szignifikánsan magasabb a fehérjék felszíni ragadóssága a fluoreszcens aggregálódó csoport fehérjéihez képest ($P=0.00038$, logisztikus regresszió).

Ezt követően megvizsgáltuk az egyes fehérjék folding rátáját is a két aggregációs csoport esetében. Azt tapasztaltuk, hogy a fluoreszcens aggregátum képző fehérjék magasabb folding rátával bírnak, amit két különböző prediktor segítségével is megerősítettünk. Ezek a Fold-Rate módszer (Gromiha és mtsai., 2006), amely a folding sebességével összefüggésben lévő szekvenciaparaméterek segítségével becsli meg a folding sebességét, valamint a contact order (Plaxco és mtsai., 1998), amely a fehérje elsődleges és harmadlagos szerkezetében egymáshoz közel álló aminosavak távolságai és a folding sebessége közti korrelációt használja a folding sebességének megbecslésére (lásd módszerek). Mindkét módszer egybehangzóan azt mutatja, hogy a sötét aggregátumokat képző fehérjék lassabb folding rátával bírnak ($P = 0,0055$ és $P=0,004$,

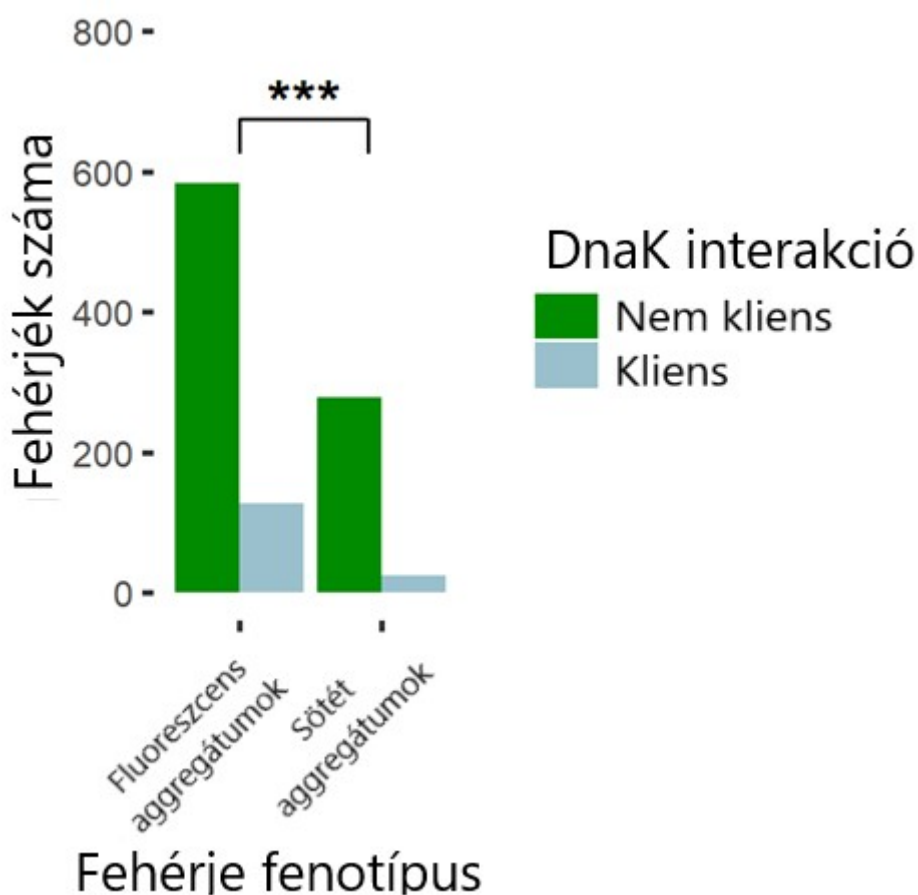
logisztikus regresszió mind a Fold-Rate, mind a contact order esetében). A contact order továbbá önmagában is a két aggregációs csoportot legjobban megkülönböztető paraméter (9. ábra).



13. ábra A contact order és a fehérje folding ráta mértéke a két aggregációs csoportban. (A). A contact order szignifikánsan magasabb a sötét aggregálódó fehérjék körében ($P = 0,0055$, logisztikus regresszió), komplexebb folding útvonalat jelezve. (B) A fluoreszcens aggregátumot képző fehérjék folding rátája magasabb ($P=0,004$, logisztikus regresszió), tehát gyorsabban képesek felvenni a natív szerkezetüket.

Végül pedig megvizsgáltuk a chaperon rendszer hatását a különböző aggregációs útvonalakra. A DnaK chaperon az *E. coli* központi dajkafehérjéje, amelynek több száz kliensfehérjéje van. Működése során támogatja ezek foldingját, már a transzláció folyamata során is (Calloni és mtsai., 2012), továbbá gátolja az aggregációt és segít a hibásan feltekeredett fehérjéknek elérni natív szerkezetüket (Balchin és mtsai., 2020; Imamoglu és mtsai., 2020). Ennek megfelelően azt látjuk, hogy a fluoreszcens aggregátumot képző fehérjék szignifikáns feldúsulást mutatnak a DnaK chaperon kliensei között ($P = 4.95 \cdot 10^{-5}$, kétoldalas Fisher egzakt teszt). Összességében ezek az eredmények azt mutatják, hogy a sötét aggregálódó fehérjék nemcsak erősebb aggregációs hajlandóságot mutatnak, hanem lassabb foldingot is, valamint a chaperon rendszer is

ritkábban óvja meg őket a misfoldingtól és az aggregációtól, ami alátámasztja a kinetikai kompetíció szerepét az aggregációs útvonalak eldöntésében az *E. coli* proteom esetén.

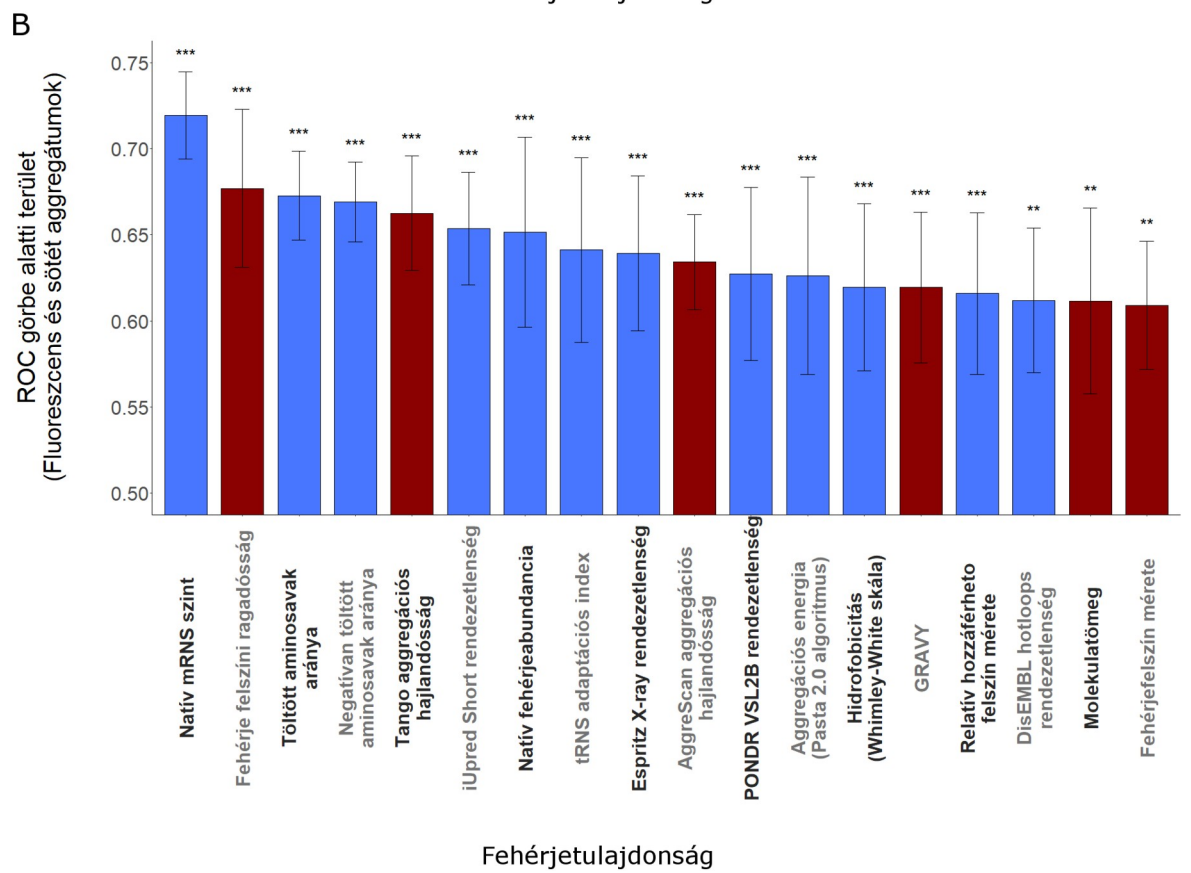
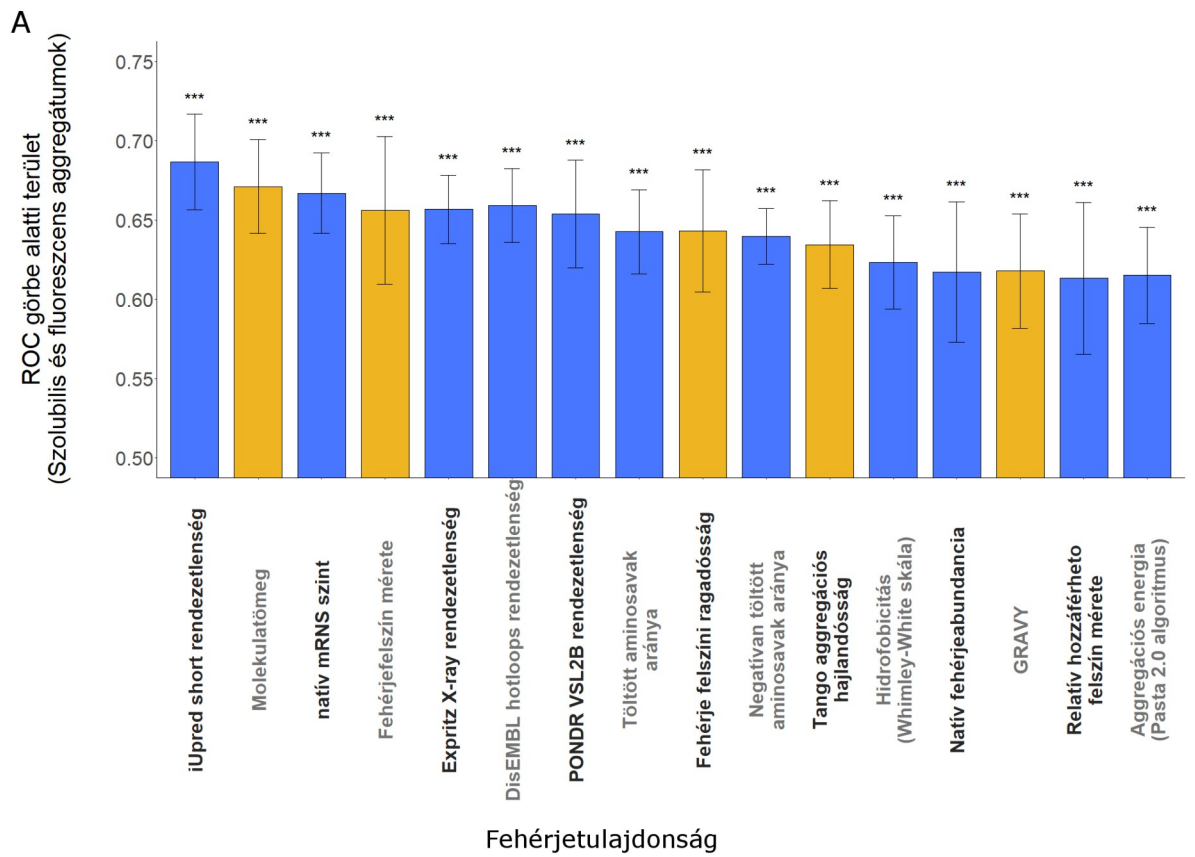


14 ábra. A DnaK-kliensfehérjék gyakorisága a sötét és fluoreszcens aggregációs csoportokban. A fluoreszcensen aggregálódó fehérjék szignifikánsan gyakrabban kliensfehérjei a DnaK chaperonoknak ($P = 4,94 \cdot 10^{-5}$, Esélyhányados = 2,43, kétoldali Fisher egzakt teszt), amelyek segítségével elkerülhetik a sötét aggregációt.

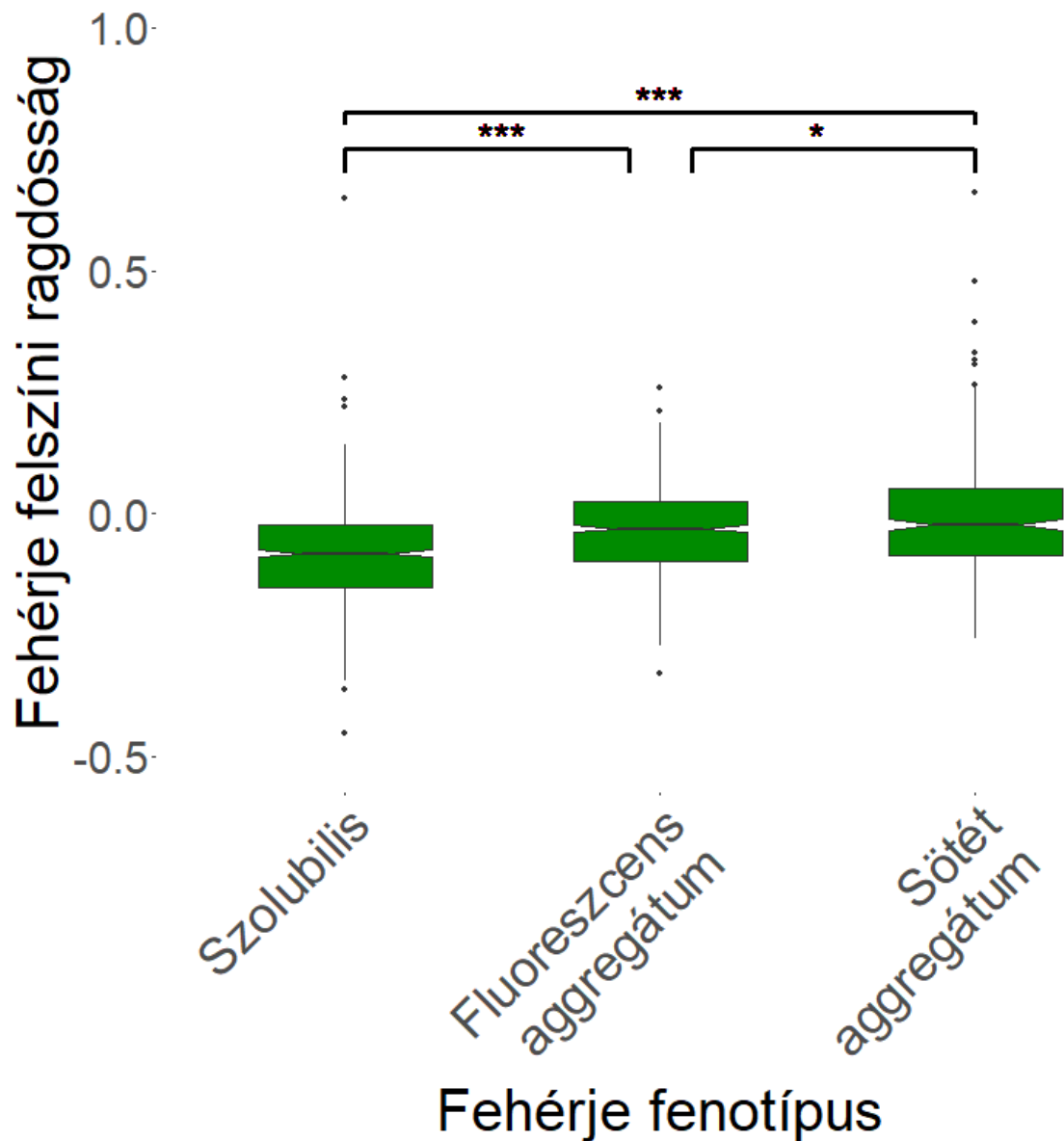
3.5. A két aggregációs útvonal közös jellemzői és különbségük a szolubilis fehérjéktől

Az aggregációs útvonalakat diszkrimináló tulajdonságok mellett kerestük azokat a fehérje paramétereket is, amelyek megkülönböztetik a szolubilis proteineket ettől a két csoporttól. Hasonló módon, mint korábban, logisztikus regressziós modellek segítségével vizsgáltuk meg az egyes paraméterek szolubilis fehérjéket diszkrimináló képességét.

Sikerült (15. ábra) megerősítenünk számos korábban leírt eredményt, miszerint az aggregálódó fehérjék alacsonyabb natív expressziós szinttel rendelkeznek (Tartaglia és mtsai., 2009), kevesebb negatív töltésű aminosav fordul elő bennük (Kramer és mtsai., 2012), hirdofóbbak (Giasson és mtsai., 2000; Schwartz és mtsai., 2001) és nagyobb relatív felszínnel bírnak (Niwa és mtsai., 2009) (2. számú kiegészítő táblázat). Azonban a korábban ismert fehérjetulajdonságokon kívül a felszíni ragadósság ($P = 2,4 \cdot 10^{-16}$, Logisztikus regresszió) és a fehérjerendezetlenség (2. számú kiegészítő táblázat) bizonyult a legerősebb diszkrimináló paraméternek az oldhatóság szempontjából.

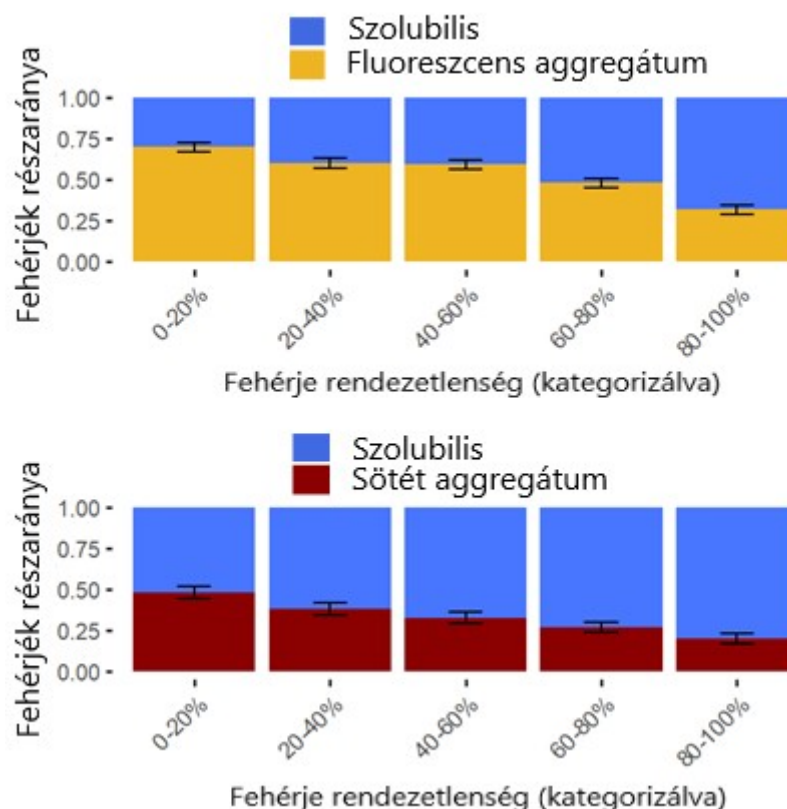


15. ábra. A szolubilis és két eltérő típusú aggregálódó fehérjecsoportot megkülönböztető fehérjeparaméterek. (A) fluoreszcensen aggregálódó és szolubilis fehérjék, valamint a (B) sötét aggregálódó és szolubilis fehérjék diszkriminációját segítő tényezők. Mindkét esetben egy fehérjeparaméter prediktív erejét a ROC görbék alatti területtel jellemezzük, amelyek 10-szeresen keresztvalidált logisztikus regressziókból származnak. Mindkét esetben a p -értékeket FDR korrekciónak vetettük alá. *** esetén $p < 0,001$, ** esetén $p < 0,01$, * esetén pedig $p < 0,05$ a logisztikus regressziós modellben. A hibasávok a görbe alatti terület 95%-os konfidenciaintervallumát ábrázolják.



16. ábra. A magas felszíni ragadósság növeli a fehérjék aggregációs hajlandóságát. A fehérjék különböző mértékű interakciós képességet mutatnak más makromolekulákkal a felszíni tulajdonságaik függvényében (Levy, 2010). A nem-natív interakciók jellemzésére az ún. „felszíni ragadósság” mérőszámát használjuk, amely a fehérje-fehérje interakciós felszínen előforduló aminosavak fehérje felszínen mutatott gyakoriságán alapszik. Eredményeink azt mutatják, hogy mind a fluoreszcens, mind a sötét aggregálódó fehérjék nagyobb felszíni ragadóssággal bírnak, mint a szolubilis fehérjék ($P < 10^{-10}$ és $P < 10^{-10}$ a két esetben, Wilcoxon rangösszeg teszt). Fontos megjegyezni, hogy a magas abundanciájú fehérjék alacsony felszíni ragadóssággal bírnak (Levy és mtsai., 2012), azonban eredményeink abban az esetben is megőrzik statisztikai erejüket, amennyiben a fehérje abundanciára kontrollálunk ($P = 2,38 \cdot 10^{-7}$ és $P = 0,00946$ a fluoreszcens és a sötét aggregátumok esetében, logisztikus regresszió).

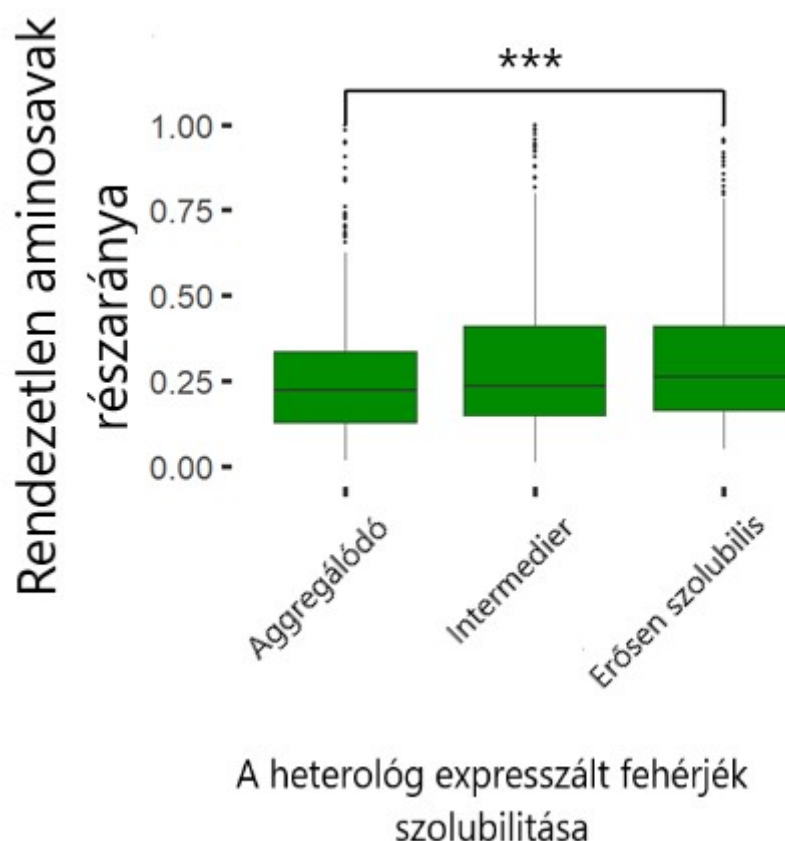
Számos fehérje tartalmaz rendezetlen régiókat, amelyek nem rendelkeznek kötött háromdimenziós struktúrával (Tompá, 2002). A vizsgálataink során az oldhatóság erős összefüggést mutatott a rendezetlen aminosavak arányával a szekvenciában (17. ábra, 4. táblázat). A fehérjerendezetlenség azonban több fizikai–kémiai tulajdonsággal is szoros összefüggésbe hozható: a rendezetlen aminosavak általában kevésbé hidrofóbak és töltöttebbek (Chiti és mtsai., 2003; Schwartz és mtsai., 2001), ami az oldhatóságot szintén segítő tulajdonság (Kramer és mtsai., 2012). Ezen felül a magasan expresszált fehérjék esetében élesztőben kimutatták, hogy a rendezetlen régiók alacsony felszíni ragadóssággal bírnak (Dubreuil és mtsai., 2019b; Macossay-Castillo és mtsai., 2019), így kerülve el a nemspecifikus interakciók által okozott aggregációt. Harmadrészt pedig a béta-aggregációért felelős szerkezeti elemek is ritkábbak a rendezetlen fehérjékben (Linding és mtsai., 2004). Eredményeink azonban robosztusak maradtak akkor is, ha ezekre a rendezetlenséggel összefüggő tulajdonságokra kontrolláltunk számításaink során (4. számú kiegészítő táblázat). Ebből arra következtethetünk, hogy a fehérjerendezetlenség pozitív hatása az oldhatóságra nem csupán a mellékterméke különböző fizikai–kémiai tulajdonságoknak, amelyek összefüggenek a kötött harmadlagos szerkezetek hiányával, hanem önálló hatással bír. Ezen felül két részre osztottuk a rendezetlen aminosavakat a fehérjén belül, hogy elkülönítsük az egyes elszórt aminosavak és a rendezetlen régiók hatását az aggregációra. Aggregálódó régiónak vettük azokat a szakaszokat, amelyek legalább tíz egymást követő rendezetlen aminosavat tartalmaztak. Amennyiben külön vizsgáltuk a rendezetlen szakaszok és a különálló rendezetlen aminosavak hatását, azt találtuk, hogy mindkét típusú fehérjerendezetlenség feldúsul a szolubilis fehérjékben (3. számú kiegészítő táblázat). Eredményeinket több különböző, fehérjerendezetlenséget prediktáló módszer eredményeivel is megerősítettük, hasonló eredményeket kapva (3. és 4. számú kiegészítő táblázatok).



17. ábra. A fehérjerendezetlenség a két aggregációs csoportban a szolubilis fehérjékhez képest. A fehérjék rendezetlenségét a PONDR VSL2B algoritmussal számoltuk, ugyanakkor az egyéb fehérjerendezetlenség-predikciós algoritmusok is hasonló eredményt adtak (2. számú kiegészítő táblázat). A felső panel a fluoreszcensen aggregálódó és szolubilis fehérjék gyakoriságát mutatja az egyes rendezetlenség kategóriákban. A szolubilis fehérjék szignifikánsan magasabb rendezetlenséggel bírnak a fluoreszcensen aggregálódó fehérjékkel szemben ($P < 10^{-15}$, logisztikus regresszió). Az alsó panelen a sötét és szolubilis fehérjék gyakorisága látható a fehérjerendezetlenség függvényében. A szolubilis fehérjék ebben az esetben is szignifikánsan gyakoribban a rendezetlenebb csoportban ($P = 1,023 \cdot 10^{-10}$, logisztikus regresszió).

A fehérjerendezetlenségének szolubilizáló hatása *E. coliban*, különösen a magasan expresszált fehérjék esetén, nagy biotechnológiai relevanciával bír a benne rejlő gyakorlati lehetőségek miatt. Ehhez azonban ismernünk kell, hogy mennyire általános ez az összefüggés az endogén fehérjéken kívül is. Ehhez a Northeast Structural Genomics Consortium heterológ expressziós adatait (Price és mtsai., 2011) használtuk, amelyek 6849 bakteriális és 373 humán fehérje oldhatóságára vonatkozó információt tartalmaztak.

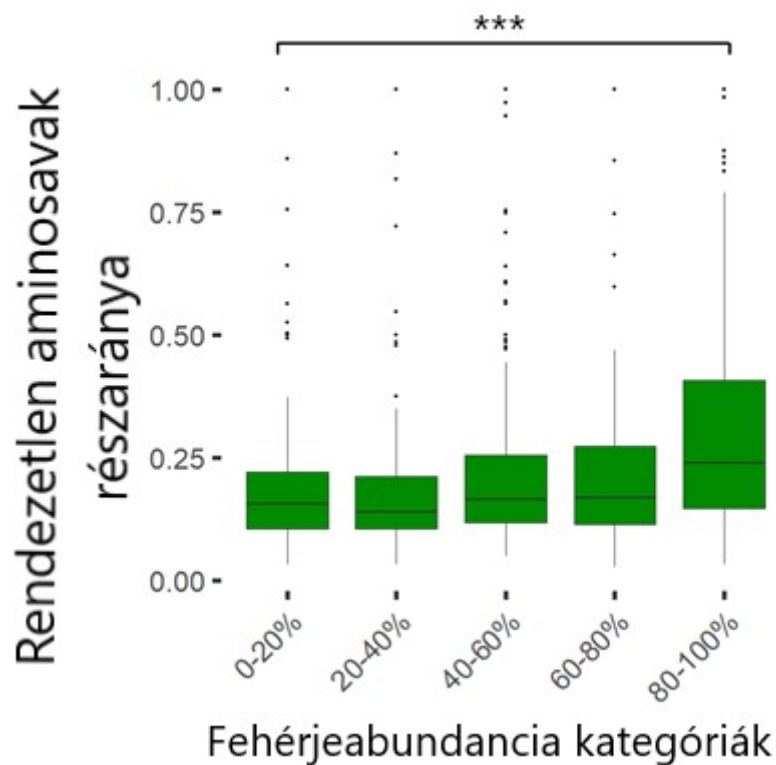
Korábbi eredményeinkkel konzisztensen mind a bakteriális, mind a humán fehérjék esetében a magasabb oldhatósággal rendelkező csoport szignifikánsan magasabb fehérjerendezetlenséggel is bírt (18. ábra). Ezek az eredmények megerősítik a fehérjerendezetlenség mint a biotechnológiában az oldhatóság javítására hasznosítható rendező elv használatának létjogosultságát.



18. ábra. A fehérjerendezetlenség mértéke az *E. coliban* heterológ expresszált fehérjék aggregációs hajlandóságának függvényében. Az erősen szolubilis fehérjék szignifikánsan magasabb rendezetlenséggel bírnak, mint az aggregáló csoport tagjai ($P < 2,2 \cdot 10^{-16}$, Wilcoxon rangösszeg teszt, $N=1504$ és 2179 az aggregáló és a magas szolubilitást mutató fehérjék esetében).

Amennyiben a fehérje felszíni ragadósság és a fehérjerendezetlenség proteom szinten is szoros összefüggést mutat az aggregációs hajlandósággal, azt várhatjuk, hogy ez megjelenik a fiziológiai expressziós szintekkel összefüggésben, mivel ismert, hogy a magasan expresszált fehérjék jól adaptálódtak az aggregáció elkerülésére. Korábbi

eredmények arra utalnak, hogy a magasan expresszált fehérjék mind alacsonyabb felszíni ragadósággal (Levy és mtsai., 2012), mind magasabb arányú rendezetlenséggel bírnak (Paliy és mtsai., 2008), bár ezeket a tulajdonságokat közvetlenül egyik esetben sem kötötték össze a fehérje aggregációval. A natív fehérje expressziós szintek vizsgálatával sikerült megerősítenünk ezeket az összefüggéseket (5. táblázat). Elvégeztünk egy többváltozós elemzést is, amely arra irányult, hogy megállapítsuk, hogy a natív expressziós szint és a felszíni ragadósság független hatással bír-e az aggregációra nézve. Mindkét változó külön-külön szignifikáns parciális hatással bírt modellünkben, tehát a két tulajdonság egymástól függetlenül hat a fehérje aggregációs hajlandóságra. Összességében eredményeink rámutatnak, hogy a magasabb szintű fehérjerendezetlenség és az alacsonyabb felszíni ragadósság mind pozitív hatással vannak az oldhatóságra fiziológias és overexpressziós körülmények között.



19. ábra. A fehérjerendezetlenség mértéke a natív expressziós szint (Schmidt és mtsai., 2016) függvényében. A legmagasabb expressziós szintet mutató fehérjék szignifikánsan nagyobb rendezetlenséggel bírnak, mint az expresszió alapján az alsó 20%-ba tartozó csoport tagjai ($P = 4.18 \cdot 10^{-11}$, Wilcoxon rangösszeg teszt).

4. Diszkusszió

Munkám során szisztematikus módon vizsgáltam a fehérje overexpresszió hatását az *in vivo* aggregációra *Escherichia coliban*. A vizsgálatok során nagy áteresztőképességű mikroszkópiát és gépi tanulást alkalmaztunk, amelyek segítségével megkülönböztettük az overexpresszió ellenére szolubilis formában maradó, valamint szupersaturálódó és ezáltal aggregálódó fehérjéket, továbbá különbséget tettünk azok között a fehérjék között, amelyek a fluoreszcens jel alapján a foldingot megelőzően vagy azt követően kerülnek az aggregátumba. Szemben a korábban uralkodó nézettel, miszerint az aggregátumok döntő többsége inaktív, szerkezetét veszített, hibásan feltekeredett fehérjéket tartalmazó aggregátum, eredményeink arra utalnak, hogy a citoplazmatikus aggregálódó fehérjék nagy része a foldingot követően aggregálódik. Ezek a fehérjék a sötét aggregátumot képző proteinekkal szemben gyorsabb foldinggal bírnak és alacsonyabb aggregációs hajlandóságot mutatnak. Ezen felül gyakrabban kliensei a *DnaK* chaperonnak, amely képes a folding folyamatában, valamint a hibásan feltekeredett állapotú proteineknek a natív szerkezetük felvételében segítséget nyújtani. Ezek az eredmények arra utalnak, hogy a fluoreszcens aggregátumok létrejöttéhez döntő fontosságú a gyors folding és a lassan bekövetkező aggregáció, ami alátámasztja a két folyamat közti kinetikai kompetíció fontosságát az aggregáció típusának meghatározásában (de Groot és Ventura, 2006). Számos bizonyíték van arra nézve, hogy ezek a folyamatok nagy jelentőséggel bírnak fiziológiás expressziós szintek mellett is. A chaperonrendszer döntő szerepet játszik mind az újonnan szintetizált, mind a már megfelelő natív szerkezettel bíró fehérjék aggregációjának megakadályozásában (Kim és mtsai., 2013). A chaperonrendszer központi elemének, a *DnaK* fehérjének kétféle szubsztrátja ismert. Azok, amelyek már a folding során is igénylik a chaperon interakciót (kotranszlációs interakció) és azok, amelyek a natív szerkezet megőrzéséhez használják (Calloni és mtsai., 2012) a dajkafehérjét (poszttranszlációs interakció). Továbbá nemcsak a transzlációs környezetben fordulnak elő aggregációra hajlamos naszcens láncok nagy mennyiségben (Natan és mtsai., 2018), hanem a natív konformációval rendelkező fehérjék is képesek aggregációra, amennyiben a koncentrációjuk átlépi az oldhatósági határt. Ezt a jelenséget mindkét esetben szupersaturációnak nevezzük (Ciriyam és mtsai.,

2015). Számos ilyen fehérje ismert és a humán sejtekben ezek a fehérjék felülreprezentáltak a neurodegeneratív betegségek szempontjából fontos biológiai folyamatokban (Ciryam *et al*, 2013; Alberti & Hyman, 2021), ami arra utal, hogy szerepük van a stresszhatásra bekövetkező patológias aggregációs folyamatokban. Eredményeink azonban azt mutatják, hogy a fehérjék jelentős része nincs az oldhatósági határon normál körülmények között, ezért a „Life on the edge” hipotézis nem tekinthető általánosnak, továbbá a szuperszaturáció bekövetkezésében is időben (és fehérjeszerkezet tekintetében) két markánsan elkülönülő csoportot tudunk azonosítani. A jövőben fontos eredmények forrása lehet a különböző, magas koncentrációban termelt fehérjék viselkedésének szisztematikus vizsgálata stresszkörülmények között is, hogy még többet megtudhassunk erről a folyamatról.

A fiziológias körülmények között magas koncentrációban előforduló fehérjéről ismert, hogy alacsony felszíni ragadósággal bírnak a nemnatív interakciók elkerülése érdekében (Levy és mtsai., 2012). Eredményeink rámutatnak, hogy a nemkívánatos interakciók elkerülése mellett az aggregáció minimalizálása is fontos szempont a felszíni ragadóság csökkentésekor mindkét aggregációs útvonal esetében. Ugyanakkor az is kimutatható, hogy a két aggregációs csoport között is szignifikáns különbség mutatkozik a felszíni ragadósággal, azonban a fehérjék belsejében lévő aminosavakra végzett analóg összehasonlítás esetében nincs különbség. Ez arra utal, hogy a folding folyamata során, de még azelőtt, hogy a GFP elérné a natív szerkezetét, a fehérjék egy kompakt szerkezetet vesznek föl, amely bizonyos tekintetben a natív szerkezetre hasonlít, az arra jellemző felszín pedig többé-kevésbé már kialakul (Hartl és Hayer-Hartl, 2009; Kim és mtsai., 2013). Ez a felszín képes az aggregációs folyamat elindítására, amelyet követhet a szerkezet későbbi felbomlása amint az intermolekuláris interakciók túlsúlyba kerülnek. Joggal feltételezhetjük, hogy a hidrofób mag összeomlása, mint a folding korai lépése időben átfedést mutat a magas aggregációs hajlandósággal bíró fehérjék aggregációjával. Rendkívül fontos információkat nyerhetünk az aggregáció folyamatáról jövőbeni kísérletekből, amelyek képesek ezeknek a részfolyamatoknak a nagy felbontású vizsgálatára, különösen a nagy felszíni ragadósággal bíró fehérjék expressziójának lehetséges optimalizációja szempontjából. Ennek a kérdésnek a vizsgálatában fontos szerephez juthat továbbá a transláció dinamikája, ugyanis a transláció sebessége nem

egyenletes a teljes mRNS hosszában, egyes szakaszokon jelentősen lelassulhat a ritka kodonok jelenléte miatt. Ez hozzájárulhat a lokális folding folyamatokhoz, ugyanakkor lehetővé teszi a naszcens láncok interakcióját. Ennek a komplex jelenségnek a szisztematikus vizsgálata egy jelentős megoldandó kérdése a jövőbeni fehérjeaggregáció-kutatásoknak.

Vizsgálataink másik fontos eredménye a fehérjerendezetlenség szerepének feltárása az aggregációval kapcsolatban. Eredményeink megmutatták, hogy a fehérjerendezetlenség az egyik legerősebb prediktor az *in vivo* szolubilitásnak, függetlenül a fizikai-kémiai tulajdonságoktól vagy a felszíni ragadóságtól. Ez arra enged következtetni, hogy a fehérjerendezetlenségből adódó szerkezeti flexibilitás járul hozzá a nagyobb mértékű szolubilitáshoz, továbbá arra utal, hogy a szuperszaturáció mellett a felszíni ragadósság és a rendezetlenség egyaránt fontos szereppel bírnak. Ez konzisztens korábbi eredményekkel a rendezetlen régiók szolubilitásra gyakorolt pozitív hatásáról (Santner és mtsai., 2014; Simone és mtsai., 2012). Egy rendezetlen régió fúziója egy aggregálódó fehérjéhez ún. entrópikus sörtét (entropic bristle) hoz létre (Santner és mtsai., 2014), amely egy dinamikusan mozgó régiót képez. Ez segít a fehérjék megfelelő foldingjában, valamint előnyös felszínt biztosít a teljes proteinnek. Ezt az elképzelést alátámasztja az a megfigyelésünk is, miszerint a szolubilis fehérjék több egybefüggő rendezetlen régiót is tartalmaznak (3. számú kiegészítő táblázat). Érdekes azonban megjegyezni, hogy a két aggregációs útvonal fehérjéi nem mutatnak különbséget a rendezetlen régiók mennyiségében, de a sötétén aggregálódó fehérjék szignifikánsan több nem ezekbe a régiókba eső, a szekvencián szétszórva elhelyezkedő rendezetlen aminosavat tartalmaznak (3. számú kiegészítő táblázat). Jövőbeni munkák fontos tanulságokkal szolgálhatnak az egyes rendezetlen aminosavak szerepéről az aggregációs folyamat korai szakaszában, és arról, hogy milyen módon hatnak a folding és az aggregáció között megfigyelhető kompetícióra. Eredményeink rávilágítanak arra is, hogy az aggregáció elkerülése nagy hatással van arra, hogy a különböző fehérjék milyen rendezetlenséggel bírnak az evolúció során a proteomra egésze nézve. Ismert volt korábban is a magasabban expresszált fehérjék nagyobb mértékű rendezetlensége (Paliy és mtsai., 2008), de az emögött rejlő evolúciós kényszerek ezidáig feltáratlanul maradtak. Bemutattuk, hogy a magasán expresszált fehérjék alacsony felszíni ragadóssága és

nagyobb mértékű rendezetlensége mind hozzájárulnak az aggregáció elkerüléséhez, fiziológias körülmények között ugyanúgy, mint overexpresszió esetén. Végül pedig eredményeink rámutatnak arra az eddig kevésbé vizsgált eshetőségre, miszerint a fehérjerendezetlenség markánsan más hatással bír prokariótákban és eukariótákban. A baktériumokhoz képest az eukarióta proteom sokkal nagyobb fokú rendezetlenséget mutat és a hosszabb, egybefüggő rendezetlen régiók aránya is jelentősen magasabb. Ezek a szegmensek gyakran vesznek részt a jelátvitelben az általuk létrehozott interakciók segítségével (Peng és mtsai., 2015), amely egybevág az eukarióták nagyobb komplexitásával és sejtes folyamataik bonyolultabb szabályozásával. Azonban eukariótákban, azon belül is az emberben esetében a rendezetlen régiókat összefüggésbe hozták a stresszhatásra bekövetkező aggregációval. (Määttä és mtsai., 2020) Továbbá az élesztőben a magasan expresszált fehérjék alacsonyabb rendezetlenséggel bírnak, feltehetően a lineáris motívumok okozta aspecifikus interakciók elkerülése érdekében (Macossay-Castillo és mtsai., 2019). Ezeket a különbségeket elvben okozhatják a citoplazmatikus milióban vagy magukban a proteinekben megfigyelhető eltérések. Azonban mint láttuk, az *E. coli*ban expresszált humán fehérjék esetében is pozitív összefüggés figyelhető meg a fehérjerendezetlenség és az oldhatóság között, így az utóbbi lehetőség sokkal valószínűtlenebb. További kutatások szükségesek a különbségek mögött álló mechanizmusok megértéséhez, hogy és az aggregáció elkerülésére szolgáló stratégiák elkülönítéséhez a prokariótákban és az eukariótákban.

A bemutatott eredményeink biotechnológiai szempontból is fontossággal bírhatnak. A fluoreszcens aggregátumok biológiailag aktív fehérjéket tartalmaznak, aminek a segítségével új fehérjetermelési megközelítések is megvalósíthatóak ipari körülmények között (Rinas és mtsai., 2017; Singhvi és mtsai., 2019; Villaverde és mtsai., 2012; Wu és mtsai., 2011). Az aggregátumok az oldott fehérjéknél könnyebben tisztíthatók és hosszabb ideig eltárolhatók a lebomlás veszélye nélkül. Az aktív enzimeket tartalmazó aggregátumok például felhasználhatóak immobilizált katalizátoroknak az iparban (Rinas és mtsai., 2017; Wu és mtsai., 2011), vagy a gyógyszerek célba juttatásához használt partikulumok alkotóiként (Villaverde és mtsai., 2012). Eredményeink azt mutatják, hogy nincs olyan enzimosztály, amely feldúsult volna az inaktív aggregátumok között, ezért potenciálisan bármely kémiaival rendelkező enzimes család tagjai

megtermelhetők aktív aggregátumok formájában. A kísérleteink metodológiai szempontból pedig a fehérjeaggregáció nagy áteresztőképességű vizsgálatában mutatkoztak rendkívül hatékonyak. Módszerünk könnyedén alkalmazható aktív fehérjéket tartalmazó aggregátumok keresésére mutagenizációs vagy randomizációs vizsgálatok során, amelyek hozzájárulhatnak a már termelt, vagy egyébként nehezen termelhető, biotechnológiai szempontból fontos fehérjék gyártási folyamatainak optimalizációjához. Fontos azonban megemlíteni, hogy az aktív fehérjéket tartalmazó aggregátumok élettartamuk során elveszíthetik az aktivitásukat és átalakulhatnak inaktív, amiloid aggregátumokká (Elia és mtsai., 2017). Ezért feltétlenül fontos annak a feltárása, hogy a fluoreszcens aggregátumok milyen mennyiségben tartalmaznak aktív fehérjéket, és ezek a fehérjék mennyi ideig és milyen arányban őrzik meg biológiai aktivitásukat. Ennek részletes vizsgálatában döntő szerephez juthat a nagy áteresztőképességű aggregációs módszertan, kombinálva hasonlóan nagy áteresztőképességű, az aggregátumok tisztítására szolgáló módszerekkel (Jäger és mtsai., 2020) és fehérje aktivitást mérő eljárásokkal (Huang és mtsai., 2015; Kuznetsova és mtsai., 2015).

5. Összegzés és kitekintés

Az elmúlt évtizedekben széles körben elfogadott volt az a nézet, amely szerint a fehérje aggregátumokat alkotó polipeptidek szerkezetüket és funkciójukat veszített makromolekulák. Ez köszönhető volt felfedezésük körülményeinek és az amiloid típusú aggregátumok fontos, betegségekben betöltött szerepének. Azonban egyre több eredmény látott napvilágot, miszerint az aggregációnak funkcionális szerepe is lehet, mind baktériumokban, mind a magasabb rendűek esetében. Később az is kiderült, hogy enzimek is képesek megőrizni aktivitásukat az aggregátumok belsejében, gyökeresen megváltoztatva, hogy hogyan tekintünk az aggregáció folyamatára. Azonban hiába áll rendelkezésünkre egyre több információ az aktív, natív szerkezetű fehérjéket tartalmazó aggregátumokról, szisztematikus vizsgálatok korábban nem készültek arra nézve, hogy milyen gyakoriak, valamint milyen fehérjetulajdonságok határozzák meg képződésüket.

Doktori disszertációm során bemutattam az *E. coli* teljes proteomjának *in vivo* aggregációjára irányuló vizsgálatainkat. Eredményeink megmutatták, hogy valóban két jelentősen eltérő aggregátumtípus található *E. coliban in vivo* körülmények között: funkcionális és amiloidszerű aggregátumok. Ezek az aggregátumok összevethető gyakoriságúak, a citoplazmában a funkcionális aggregátumok a gyakoribbak, a membrán esetében pedig az amiloid típusúak. Fontos megjegyezni, hogy a citoplazmatikus aggregátumok kb. 70%-a funkcionális, amely rávilágít arra, hogy ezek a fehérjék nem szükségszerűen igényelnek erős denaturálószeret és *in vitro* refoldingot ahhoz, hogy funkcionális fehérjék legyenek kinyerhetőek. Továbbá a fehérjék jelentős része szolubilis marad magas overexpresszió mellett is, rámutatva arra, hogy számos protein jócskán az oldhatósági határérték alatt van sejtes környezetben, nagy mozgásteret engedélyezve a koncentrációváltozásra komolyabb fitnessre ható következmények nélkül. Munkám során nagy mennyiségű, szisztematikus adatból álló adatsort állítottam össze az *E. coli* fehérjeinek számos tulajdonságáról, ideértve a fizikai-kémiai, szerkezeti, funkcionális és egyéb paramétereket, amelyek segítségével a teljes proteom tekintetében vizsgálhattam, hogy mely tulajdonságok befolyásolják a két fehérjeaggregátum kialakulását. Vizsgálataink feltártak több paramétert, amelyek hozzájárulhatnak a funkcionális polipeptideket tartalmazó aggregátumok létrejöttéhez: a gyors folding és alacsony

aggregációs hajlandóság, a felszíni ragadósság és a chaperon interakciók mind segíti a fehérjéket a natív térszerkezetük elnyerésében az aggregációt megelőzően, aktív terméket eredményezve. Ez megfelel az ún. kinetikai kompetíció elméletének (de Groot és Ventura, 2006), amely szerint a létrejövő aggregátum típusa a folding és az aggregáció hatásának eredőjén múlik, így egy viszonylag egyszerű és általános elv jól leírja a két csoport közti különbségeket rendszerszinten is. Másik fontos eredményünk az a felismerés, amely szerint a fehérjerendezetlenség pozitív hatással van a baktériumokban expresszált fehérjék szolubilitására, feltehetően a kedvező felszín biztosítása és a folding elősegítése miatt. Ez a két független hatás kiegészíti korábbi ismereteinket, amelyek arra utaltak, hogy a fehérjék aggregációját befolyásoló legfontosabb tényező a szuperszaturáció. A teljes proteom vizsgálata egy komplexebb képet tár elénk, amely nem csak igazolja a megközelítés létjogosultságát, de jövőbeni vizsgálatok kiindulópontját is képezheti.

Eredményeink potenciális stratégiákat vetítenek előre a funkcionális fehérjék heterológ expressziójával kapcsolatban. Szemben a korábbi megközelítésekkel, miszerint a fehérjék mesterséges szelekciója során az aggregációs hajlandóság csökkentése volt a végcél, egy árnyaltabb kép tárul elénk, amelyben ugyanolyan fontos a gyors és hatékony folding elérése, mint pusztán az aggregáció elkerülése. Számos folding riporter áll rendelkezésünkre a hatékony szelekció érdekében (Sachsenhauser és Bardwell, 2018), valamint számos nagy áteresztőképességű genommérnöki módszer az egyes variánsok létrehozására (Shalem és mtsai., 2015). Alternatívaként léteznek módszerek arra nézve, hogy fúziós peptidek segítségével egy fehérje chaperon klienssé válhasson, mint az ún. „horgászsinór” módszer (Aprile és mtsai., 2015). Ebben az esetben egy a fehérjeszekvencia megváltoztatása nélkül tehető egy fehérjére specifikusan klienssé a chaperonrendszer számára, elősegítve annak foldingját anélkül, hogy a többi fehérje foldingját érintené. Hasonló módon használhatóak a korábban említett és egyéb fehérjerendezetlenség-predikciós algoritmusok magasabb rendezetlenségű variánsok létrehozására a jobb oldhatóság érdekében.

Azonban továbbra is fontos kérdés, hogy eredményeink mennyire általánosíthatóak, főleg az eukarióták körében. Számos fehérje termeléséhez kizárólag eukarióta rendszerek használhatóak, köszönhetően a poszttranszlációs módosítások

jelenlétének, amelyeket a bakteriális sejtek nem képesek megfelelően létrehozni. Vannak bizonyítékok arra nézve, hogy eukarióták is képesek funkcionális aggregátumok létrehozására (Rueda és mtsai., 2016), és bár egyes fehérjék aggregációs mechanizmusai jól ismertek, az aggregációt befolyásoló általános elvek továbbra sem. Mekkora és milyen hatással bírnak a fehérje felszíni interakciók olyan sejtekben, amelyek legalább egy nagyságrenddel nagyobb komplexitást mutatnak, mint ahogy az eukarióták? Fennáll-e a fehérjerendezetlenség pozitív hatása az eukariótákba, amelyek proteomjai eleve sokkal nagyobb rendezetlenséget mutatnak? Ezeknek a kérdéseknek a megválaszolásával egy újabb lépéssel közelebb kerülhetünk a jövőben ahhoz, hogy bármely fontos fehérjét megbízhatóan, nagy mennyiségben képesek legyünk megtermelni, legyen a felhasználás ipari, biotechnológiai vagy gyógyászati.

Végül pedig az általunk leírt hatások nagy valószínűséggel fiziológiai körülmények között is jelentőséggel bírnak, mivel az aggregáció elkerülése a fehérjeevolúció egy fontos szempontja a sejtés élőlényekben. Eredményeinknek köszönhetően az *E. coli* proteomevolúciójának újabb szegmense tárult fel előttünk, azonban újabb kérdések is felmerültek. Milyen erős a fehérje aggregáció proteomevolúcióra gyakorolt hatása? Van-e különbség a különböző komplexitási szinteken, vagy az élőlény környezete bír döntő szereppel? Jobban hasonlít két extremofil egymásra, mint két közelebbi rokon? Ezeknek a kérdéseknek a megválaszolására bizonyára még sokat kell várnunk, de megtettük az első lépéseket megértésük felé.

A tézis legfontosabb megállapításai:

1.) Túltermelés hatására a fehérjék jelentős része aggregátumot képez (kb 70%) *in vivo* körülmények között. Az aggregátumok közel megegyező mennyiségben natívszerű struktúrájú vagy hibásan feltekeredett fehérjékből jönnek létre. Ugyanakkor egy szintén szignifikáns csoport oldható marad a nagy mértékű túltermelés ellenére is (kb. 30%), kétségsbe vonva az állítást, miszerint a fehérjék döntő többsége az oldhatósági limit körüli koncentrációval bír a sejtekben.

2.) Az aggregációra hajlamos fehérjék köre jelentős átfedést mutat a korábbi *in vitro* eredményekkel, ugyanakkor árnyalja is azokat. A validációs kísérletek

eredményeinek tükrében kijelenthető, hogy a nagy áteresztőképességű mikroszkópia és a gépi tanulás alkalmas a fehérje aggregáció gyors és hatékony vizsgálatára, információval szolgálva a fehérje folding állapotára is.

3.) Az aggregátumok létrejöttét befolyásoló tényezők közül számos korábban is ismert volt, amelyek nagy részét a mi adataink is megerősítettek. Ilyenek például a natív expressziós szint, a negatív töltés, továbbá a hidrofobicitás. Ezen felül összefüggésbe hozható a fehérjék oldhatósága további tulajdonságokkal, mint a felszíni ragadósság által mediált nemkívánatos interakciók alacsony volta és a magas fogú fehérjerendezetlenség. Ezek korábban ellentmondásos módon jelentek meg az irodalomban, azonban *E. coli* esetében egyértelmű hatás igazolódott.

4.) A képződő fehérjeaggregátumok típusát legnagyobb mértékben a fehérje folding sebessége és az aggregációra való hajlam befolyásolja. Ezen felül a felszíni ragadósság is bírhat bizonyos szereppel a foldingon részben átesett fehérjék esetén.

5.) Az oldhatóságot elősegítő tényezők közül a fehérje rendezetlenség pozitív szerepét a heterológ módon expresszált fehérjék esetén is megerősítettük. Ez arra utal, hogy a fehérje rendezetlenség ellentmondásos szerepe a különböző organizmusspecifikus hatások eredménye lehet.

6. Köszönetnyilvánítás

Szeretném kifejezni köszönetemet témavezetőmnek, Dr. Papp Baláznak a munkánk során nyújtott szakmai segítségért és útmutatásért.

Köszönöm Szappanos Baláznak, Kovács Károlynak és Fekete Gergelynek (Papp Balázs csoport), Daruka Lejlának, Balogh Dávidnak és Pál Csabának (Pál Csaba csoport), Horváth Pétenek és Kintses Bálintnak a projektben nyújtott segítségüket.

Köszönöm Magyar Zoltánnak és Őszi Erikának a segítséget, amit a kísérletek elvégzésével nyújtottak.

Köszönettel tartozom Natalia de Grootnak, Fuxreiter Mónikának és Tompa Péternek a sok építő kritikáért és észrevételért, amelyeket közös munkánk során tettek, és amik hozzájárultak szakmai fejlődésemhez.

Köszönöm feleségemnek és családnak az évek során tanúsított támogatást és türelmet.

7. Hivatkozások

Agostini, F., Cirillo, D., Livi, C.M., Ponti, R.D., and Tartaglia, G.G. (2014). cc SOL omics : a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics* 30, 2975–2977.

Aprile, F.A., Sormanni, P., and Vendruscolo, M. (2015). A Rational Design Strategy for the Selective Activity Enhancement of a Molecular Chaperone toward a Target Substrate. *Biochemistry* 54, 5103–5112.

Arranz-Otaegui, A., Carretero, L.G., Ramsey, M.N., Fuller, D.Q., and Richter, T. (2018). Archaeobotanical evidence reveals the origins of bread 14,400 years ago in northeastern Jordan. *PNAS* 115, 7925–7930.

Arslan, P.E., and Chakrabartty, A. (2009). Probing Alzheimer amyloid peptide aggregation using a cell-free fluorescent protein refolding method. *Biochem. Cell Biol.* 87, 631–639.

Bakholdina, S.I., Stenkova, A.M., Bystritskaya, E.P., Sidorin, E. V, Kim, N.Y., Menchinskaya, E.S., Gorpenchenko, T.Y., Aminin, D.L., Shved, N.A., and Solov'eva, T.F. (2021). Studies on the Structure and Properties of Membrane Phospholipase A 1 Inclusion Bodies Formed at Low Growth Temperatures Using GFP Fusion Strategy. *Molecules* 26.

Balchin, D., Hayer-Hartl, M., and Hartl, F.U. (2020). Recent advances in understanding catalysis of protein folding by molecular chaperones. *Febs Lett.* 594, 2770–2781.

Baneyx, F., and Mujacic, M. (2004). Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* 22, 1399–1408.

Bellesia, G., Shea, J., Bellesia, G., and Shea, J. (2009). Effect of beta-sheet propensity on peptide aggregation Effect of beta-sheet propensity on peptide aggregation. *J. Chem. Phys.* 130, 145103-1–9.

Ben-Zvi, A., Rios, P.D.L., Dietler, G., and Goloubinoff, P. (2004). Active Solubilization and Refolding of Stable Protein Aggregates By Cooperative Unfolding Action of Individual Hsp70 Chaperones *. *J. Biol. Chem.* 279, 37298–37303.

Betton, J., and Hofnung, M. (1996). Folding of a Mutant Maltose-binding Protein of *Escherichia coli* Which Forms Inclusion Bodies. *J. Biol. Chem.* 271, 8046–8052.

Bianco, V., Franzese, G., and Coluzza, I. (2019). In silico evidence that protein unfolding is as a precursor of the protein aggregation. *Chemphyschem* 21, 377–384.

Bidwell, L.M., Mcmanus, M.E., Gaedigk, A., Kakuta, Y., Negishi, M., Pedersen, L., and Martin, J.L. (1999). Crystal Structure of Human Catecholamine Sulfotransferase. *J. Mol. Biol.* 293, 521–530.

Bock, S., Wu, B., and Beitz, E. (2012). Fluorescent In Situ Folding Control for Rapid Optimization of Cell-Free Membrane Protein Synthesis. *PLoS One* 7, 1–7.

Boke, E., Ruer, M., Wühr, M., Coughlin, M., Lemaitre, R., Gygi, P., Alberti, S., Drechsel, D., Hyman, A.A., and Mitchison, T.J. (2016). Amyloid-like Self-assembly of a Cellular Compartment. *Cell* 166, 637–650.

Bou, G., Remacha, M., and Ballesta, J.P.G. (2000). Ribosomal Stalk Protein Phosphorylating Activities in *Saccharomyces cerevisiae*. *Arch. Biochem. Biophys.* 375, 83–89.

Bowden, G.A., and Georgiou, G. (1990). Folding and aggregation of β -lactamase in the periplasmic space of *Escherichia coli*. *J. Biol. Chem.* 265, 16760–16766.

Brader, M.L. (2014). Protein Aggregation in Pharmaceutical Biotechnology.

Brange, J., Havelund, S., and Hougaard, P. (1992). Chemical stability of insulin. 2. Formation of higher molecular weight transformation products during storage of pharmaceutical preparations. *Pharm. Res.* 9, 727–734.

Brocchieri, L., and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* 33, 3390–3400.

Bukau, B., and Walker, G.C. (1989). Cellular Defects Caused by Deletion of the *Escherichia coli* dnaK Gene Indicate Roles for Heat Shock Protein in Normal Metabolism. *J. Bacteriol.* 171, 2337–2346.

Calloni, G., Chen, T., Schermann, S.M., Chang, H., Genevoux, P., Agostini, F., Tartaglia, G.G., Hayer-hartl, M., and Hartl, F.U. (2012). DnaK functions as a central hub in the *E. coli* chaperone network. *CellReports* 1, 251–264.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler : image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7.

Carvalho, S.B., Botelho, H.M., Leal, S.S., Cardoso, I., Fritz, G., and Gomes, C.M. (2013). Intrinsically Disordered and Aggregation Prone Regions Underlie β -Aggregation in S100 Proteins. *PLoS One* 8.

Caughey, B., and Lansbury, P.T. (2003). Protofibrils, pores, fibrils, and neurodegeneration: Separating the Responsible Protein Aggregates from The Innocent Bystanders. *Annu. Rev. Neurosci.* 26, 267–298.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., and Prasher, D.C. (1992). Green Fluorescent Protein as a Marker for Gene Expression. *SC* 263, 802–805.

Chapman, M.R., Robinson, L.S., Pinkner, J.S., Roth, R., Hammar, M., Normark, S., and Hultgren, S.J. (2010). Role of *Escherichia coli* Curli Operons in Directing Amyloid Fiber Formation. *Science* (80-.). 295, 851–855.

Chen, J., and Stites, W.E. (2001). Packing Is a Key Selection Factor in the Evolution of Protein Hydrophobic Cores. *Biochemistry* 40, 15280–15289.

Cheng, Y., Legall, T., Oldfield, C.J., Dunker, A.K., and Uversky, V.N. (2006). Abundance of Intrinsic Disorder in Protein Associated with Cardiovascular Disease. *Biochemistry* 45, 10448–10460.

Chi, E.Y., Krishnan, S., Randolph, T.W., and Carpenter, J.F. (2003). Physical Stability of Proteins in Aqueous Solution : Mechanism and Driving Forces in Nonnative Protein Aggregation. *Pharm. Res.* 20, 1325–1336.

Chiti, F., and Dobson, C.M. (2009). Amyloid formation by globular proteins under native conditions. *Nat. Chem. Biol.* 5, 15–22.

Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., and Dobson, C.M. (1999). Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc. Natl. Acad. Sci.* 96, 3590–3594.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C.M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Lett. to Nat.* 424, 805–808.

Chong, S.-H., and Ham, S. (2014). Interaction with the Surrounding Water Plays a Key Role in Determining the Aggregation Propensity of Proteins. *Angew. Chemie* 126, 4042–4045.

Ciryam, P., Tartaglia, G.G., Morimoto, R.I., Dobson, C.M., and Vendruscolo, M. (2013). Widespread Aggregation and Neurodegenerative Diseases Are Associated with Supersaturated Proteins. *Cell Rep.* 5, 781–790.

Ciryam, P., Kundra, R., Morimoto, R.I., Dobson, C.M., and Vendruscolo, M. (2015). Supersaturation is a major driving force for protein aggregation in neurodegenerative diseases. *Trends Pharmacol. Sci.* 36, 72–77.

Conchillo-Solé, O., Groot, N.S. De, Avilés, F.X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN : a server for the prediction and evaluation of " hot spots " of aggregation in polypeptides. *BMC Bioinformatics* 8.

Cordes, M.H.J., Davidson, A.R., and Sauer, R.T. (1996). Sequence space , folding and protein design. *Curr Opin Struct Biol* 6, 3–10.

Csizmo, V. (2005). The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* 347, 827–839.

Darling, A.L., and Uversky, V.N. (2018). Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front. Genet.* 9, 1–18.

Das, S., and Mukhopadhyay, D. (2011). Intrinsically Unstructured Proteins and Neurodegenerative Diseases : Conformational Promiscuity at its Best. *Life* 63, 478–488.

Deechongkit, S., Dawson, P.E., and Kelly, J.W. (2004). Toward Assessing the Position-Dependent Contributions of Backbone Hydrogen Bonding to β -Sheet Folding Thermodynamics Employing Amide-to-Ester Perturbations. *J. Am. Chem. Soc.* 126, 16762–16771.

Deforte, S., and Uversky, V.N. (2016). Intrinsically disordered proteins in PubMed: what can the tip of the iceberg tell us about what lies below? *RSC Adv.* 6, 11513–11521.

Diamant, S., Ben-Zvi, A.P., Bukau, B., and Goloubinoff, P. (2000). Size-dependent Disaggregation of Stable Protein Aggregates by the DnaK Chaperone Machinery. *J. Biol.*

Chem. 275, 21107–21113.

Díez-Gil, C., Krabbenborg, S., García-fruitós, E., Vazquez, E., Rodríguez-Carmona, E., Ratera, I., Ventosa, N., Seras-Franzoso, J., Cano-Garrido, O., Ferrer-Miralles, N., et al. (2010). The nanoscale properties of bacterial inclusion bodies and their effect on mammalian cell proliferation. *Biomaterials* 31, 5805–5812.

Dill, K.A. (1985). Theory for the Folding and Stability of Globular Proteins. *Biochemistry* 24, 1501–1509.

Dill, K.A., Fiebig, K.M., and Chan, H.S. (1993). Cooperativity in protein-folding kinetics. *Biophysics (Oxf)*. 90, 1942–1946.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. (2008). The Protein Folding Problem. *Annu. Rev. Biophys.* 37, 289–316.

Dobson, C.M. (2002). The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J.* 21, 5682–5690.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). Structural bioinformatics IUPred : web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433–3434.

Dubreuil, B., Matalon, O., and Levy, E.D. (2019a). Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-functional Interactions. *J. Mol. Biol.* 431, 4978–4992.

Dubreuil, B., Matalon, O., and Levy, E.D. (2019b). Protein abundance biases the amino-acid composition of disordered regions to minimize non-functional interactions. *J. Mol. Biol.* 431, 4978–4992.

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inf. Ser Work. Genome Inf.* 11, 161–171.

Elia, F., Cantini, F., Chiti, F., Dobson, C.M., and Bemporad, F. (2017). Direct Conversion of an Enzyme from Native-like to Amyloid-like Aggregates within Inclusion Bodies. *Biophys. J.* 112, 2540–2551.

Falsone, A., and Falsone, S.F. (2015). Legal but lethal: functional protein aggregation at the verge of toxicity. *Front. Cell. Neurosci.* 9, 1–16.

Fändrich, M., Fletcher, M.A., and Dobson, C.M. (2001). Amyloid fibrils from muscle myoglobin. *Nature* 410, 165–166.

Fang, Y., and Fang, J. (2013). Discrimination of soluble and aggregation-prone proteins based on sequence information. *Mol. Biosyst.*

Ferreira, S.T., Vieira, M.N.N., and Felice, F.G. De (2007). Soluble Protein Oligomers as Emerging Toxins in Alzheimer ' s and Other Amyloid Diseases. *Life* 59, 332–345.

Fink, A.L. (1998). Protein aggregation : folding aggregates, inclusion bodies and amyloid. *Fold. Des.* 3, 9–23.

Fischer, E. (1894). Influence of Configuration on the Action of Enzymes. *J. Am.*

Chem. Soc. 3, 2985–2993.

Fong, K., Li, Y., Wang, W., Ma, W., Li, K., Qi, R.Z., Liu, D., Songyang, Z., and Chen, J. (2013). Whole-genome screening identifies proteins localized to distinct nuclear bodies. *J. Biol. Chem.* 203, 149–164.

Fowler, D.M., Koulov, A. V, Alory-jost, C., Marks, M.S., Balch, W.E., and Kelly, J.W. (2006). Functional Amyloid Formation within Mammalian Tissue. *PLOS Biol.* 4, 2–9.

Freddolino, P.L., Harrison, C.B., Liu, Y., and Schulten, K. (2014). Challenges in protein folding simulations : Timescale , representation, and analysis. *Nat. Phys.* 6.

Fujiwara, K., Ishihama, Y., Nakahigashi, K., Soga, T., and Taguchi, H. (2010). A systematic survey of in vivo obligate chaperonin-dependent substrates. *EMBO J.* 29, 1552–1564.

García-Fruitós, E., González-Montalbán, N., Morell, M., Vera, A., Ferraz, R.M., Arís, A., Ventura, S., and Villaverde, A. (2005). Aggregation as bacterial inclusion bodies does not imply inactivation of enzymes and fluorescent proteins. *Microb. Cell Fact.* 6, 1–6.

García-Fruitós, E., Díez-Gil, C., Rosa M, F., Rodríguez-Carmona, E., Esther, V., Corchero, L.J., Cano-Sarabia, M., Ratera, I., Ventosa, N., Veciana, J., et al. (2009). Surface Cell Growth Engineering Assisted by a Novel Bacterial Nanomaterial. *Adv. Mater.* 21, 4249–4253.

Garvey, M., Meehan, S., Gras, S.L., Schirra, H.J., Craik, D.J., Weerden, N.L. Van Der, Anderson, M.A., Gerrard, J.A., and Carver, J.A. (2013). Biochimica et Biophysica Acta A radish seed antifungal peptide with a high amyloid fi bril-forming propensity. *BBA - Proteins Proteomics* 1834, 1615–1623.

Geiler-samerotte, K.A., Dion, M.F., Budnik, B.A., Wang, S.M., and Hartl, D.L. (2010). Misfolded proteins impose a dosage-dependent fi tness cost and trigger a cytosolic unfolded protein response in yeast. *PNAS* 108, 680–685.

Georgiou, G., and Valax, P. (1999). Isolating Inclusion Bodies from Bacteria. *Methods Enzymol.* 309, 48–58.

Georgiou, G., Telford, J.N., Shuler, M.L., and Wilson, D.B. (1986). Localization of Inclusion Bodies in Escherichia coli Overproducing P-Lactamase or Alkaline Phosphatase. *Appl. Environ. Microbiol.* 52, 1157–1161.

Georgiou, G., Valax, P., Ostermeier, M., and Horowitz, P.M. (1994). Folding and aggregation of TEM beta-lactamase : Analogies with the formation of inclusion bodies in Escherichia coli. *Protein Sci.* 3, 1953–1960.

Giasson, B.I., Murray, I.V.J., Trojanowski, J.Q., and Lee, V.M. (2000). A Hydrophobic Stretch of 12 Amino Acid Residues in the Middle of α -Synuclein Is Essential for Filament Assembly. *J. Biol. Chem.* 276, 2380–2386.

Goeddel, D. V., Kleid, D.G., Bolival, F., Heyneker, H.L., Yansura, D.G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K., and Riggs, A.D. (1979). Expression in Escherichia coli of chemically synthesized genes for human insulin. *Prot. Natl. Acad. Sci. USA* 76, 106–110.

Gonzalez-Montalban, N., Natalello, A., Garcí, E., Villaverde, A., and Doglia, S.M.

(2008). In Situ Protein Folding and Activation in Bacterial Inclusion Bodies. *Biotechnol. Bioeng.* *100*, 797–802.

Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct.* *29*, 105-53.

Gour, S., Kaushik, V., Kumar, V., Bhat, P., Yadav, S.C., and Yadav, J.K. (2016). Antimicrobial peptide (Cn-AMP2) from liquid endosperm of *Cocos nucifera* forms amyloid-like fibrillar structure. *J. Pept. Sci.* 201–207.

Gout, J.-F., Kahn, D., and Duret, L. (2010). The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLOS Genet.* *6*.

Gregoire, S., and Kwon, I. (2012). A revisited folding reporter for quantitative assay of protein misfolding and aggregation in mammalian cells. *Biotechnol J.* *7*, 1297–1307.

Gromiha, M.M., Thangakani, A.M., and Selvaraj, S. (2006). FOLD-RATE : prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.* *34*, 70–74.

Groot, N.S. De, and Ventura, S. (2010). Protein Aggregation Profile of the Bacterial Cytosol. *PLoS One* *5*.

Groot, N.S. De, Sabate, R., and Ventura, S. (2009). Amyloids in bacterial inclusion bodies. *Trends Biochem. Sci.* *34*, 408–416.

de Groot, N.S., and Ventura, S. (2006). Protein activity in bacterial inclusion bodies correlates with predicted aggregation rates. *J. Biotechnol.* *125*, 110–113.

de Groot, N.S., Aviles, F.X., Vendrell, J., and Ventura, S. (2006). Mutagenesis of the central hydrophobic cluster in A β 42 Alzheimer's peptide. *FEBS J.* *273*, 658–668.

Haber, E., and White, F.H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci.* *47*, 1309–1314.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* *11*, 10–18.

Handwerker, K.E., Cordero, J.A., and Gall, J.G. (2005). Cajal Bodies , Nucleoli , and Speckles in the *Xenopus* Oocyte Nucleus Have a Low-Density, Sponge-like Structure. *Mol. Biol. Cell* *16*, 202–211.

Harrison, S.C., and Durbin, R. (1985). Is there a single pathway for the folding of a polypeptide chain? *Proc. Natl. Acad. Sci. U. S. A.* *82*, 4028–4030.

Hart, A., and Bailey, J.E. (1990). Protein Composition of *Vitreoscilla* Hemoglobin Inclusion Bodies Produced in *Escherichia coli*. *J. Biol. Chem.* *265*, 12728–12733.

Hartl, F.U., and Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* *16*, 574–581.

Hebditch, M., Carballo-Amador, M.A., Charonis, S., Curtis, R., and Warwicker, J. (2017). Sequence analysis Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* *33*, 3098–3100.

Hedhammar, M., Stenvall, M., Rosa, L., Nord, O., Sj, O., Brismar, H., Uhl, M., Ottosson, J., and Hober, S. (2005). A novel flow cytometry-based method for analysis of expression levels in *Escherichia coli*, giving information about precipitated and soluble protein. *J. Biotechnol.* *119*, 133–146.

Hess, S., Caballero-oli, G., Giro, J.A., Alteri, C.J., Xicohte, J., and Friedman, R.L. (2007). *Mycobacterium tuberculosis* produces pili during human infection. *PNAS* *104*, 5145–5150.

Hoegge, C., Gharakhani, J., Jülicher, F., and Hyman, A.A. (2009). Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* (80-.). *324*, 1729–1732.

Hollingsworth, S.A., and Dror, R.O. (2019). Molecular dynamics simulation for all. *Neuron* *99*, 1129–1143.

Hou, Q., Kwasigroch, J.M., Rومان, M., and Pucci, F. (2020). SOLart: a structure-based method to predict protein solubility and aggregation. *Bioinformatics* *36*, 1445–1452.

Huang, H., Pandya, C., Liu, C., Al-obaidi, N.F., Wang, M., Zheng, L., and Toews, S. (2015). Panoramic view of a superfamily of phosphatases through substrate profiling. *PNAS* *112*, 1974–1983.

Huang, J., Cheng, J., and Chen, H. (2007). Secondary Structure Length as a Determinant of Folding Rate of Proteins With Two- and Three-State Kinetics. *PROTEINS Struct. Funct. Bioinforma.* *67*, 12–17.

Hull, R.L., Westermark, G.T., Westermark, P.E.R., and Kahn, S.E. (2004). Islet Amyloid: A Critical Entity in the Pathogenesis of Type 2 Diabetes. *J. Clin. Endocrinol. Metab.* *89*, 3629–3643.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., and Dunker, A.K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *PNAS* *99*, 573–584.

Imamoglu, R., Balchin, D., Hayer-Hartl, M., and Hartl, F.U. (2020). Bacterial Hsp70 resolves misfolded states and accelerates productive folding of a multi-domain protein. *Nat. Commun.* *11*.

Ivankov, D.N., and Finkelstein, A. V (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *PNAS* *101*, 8942–8944.

Jäger, V.D., Lamm, R., Küsters, K., Ölçücü, G., Oldiges, M., Jaeger, K., Büchs, J., and Krauss, U. (2020). Catalytically-active inclusion bodies for biotechnology — general concepts, optimization, and application. *Appl. Microbiol. Biotechnol.* *104*, 7313–7329.

Jenkins, T.M., Hickman, A.B., Dyda, F., Ghirlando, R., Davies, D.R., and Craigie, R. (1995). Catalytic domain of human immunodeficiency virus type 1 integrase: Identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc. Natl. Acad. Sci. U. S. A.* *92*, 6057–6061.

Johansson, J., Gudmundsson, G.H., Rottenberg, E., Berndt, K.D., and Agerberth, B. (1998). Conformation-dependent Antibacterial Activity of the Naturally Occurring Human

Peptide LL-37. *J. Biol. Chem.* 273, 3718–3724.

Kameoka, D., Masuzaki, E., Ueda, T., and Imoto, T. (2007). Effect of Buffer Species on the Unfolding and the Aggregation of Humanized IgG. *J. Biochem.* 142, 383–391.

Karplus, M., and Weaver, D.L. (1994). Protein folding dynamics : The diffusion-collision model and experimental data. *Protein Sci.* 3, 650–668.

Katayama, D.S., Nayar, R., Chou, D.K., Valente, J.J., Cooper, J., Henry, C.S., Velde, D.G. Vander, Villarete, L., Liu, C.P., and Manning, M.C. (2006). Effect of Buffer Species on the Thermally Induced Aggregation of Interferon-tau. *J. Pharm. Sci.* 95, 1212–1226.

Keseler, I.M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., et al. (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* 45, 543–550.

Kim, W., and Hecht, M.H. (2005). Sequence Determinants of Enhanced Amyloidogenicity of Alzheimer A β 42 Peptide Relative to A β 40. *J. Biol. Chem.* 280, 35069–35076.

Kim, Y.E., Hipp, M.S., Bracher, A., Hayer-hartl, M., and Hartl, F.U. (2013). Molecular Chaperone Functions in Protein Folding and Proteostasis.

Kitagawa, M., Ara, T., Arifuzzaman, M., and Ioka-Nakamichi, T. (2005). Complete set of ORF clones of *Escherichia coli* ASKA library (A Complete Set of *E. coli* K-12 ORF Archive): Unique Resources for Biological Research. *DNA Res.* 12, 291–299.

Knighton, D.R., Zheng, J., Eyck, L.F.T., Ashford, V.A., Xuong, N.-H., Taylor, S.S., and Sowadski, J.M. (1991). Crystal Structure of the Catalytic Subunit of Cyclic Adenosine Monophosphate-Dependent Protein Kinase. *Science* (80-.). 253, 407–414.

Köhler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256, 495–497.

Kostas, T., Dhurjati, P., Millet, J., Béguin, P., and Jean-Paul, A. (1991). High-activity of inclusion bodies formed in *Escherichia coli* overproducing *Clostridium thermocellum* endoglucanase D. *FEBS J.* 282, 205–208.

Kramer, R.M., Shende, V.R., Motl, N., Pace, C.N., and Scholtz, J.M. (2012). Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. *Biophys. J.* 102, 1907–1915.

Kueltzo, L.A., Wang, W.E.I., Randolph, T.W., and Carpenter, J.F. (2008). Effects of Solution Conditions , Processing Parameters , and Container Materials on Aggregation of a Monoclonal Antibody during Freeze – Thawing. *J. Phar* 97, 1801–1812.

Kuznetsova, E., Proudfoot, M., Gonzalez, C.F., Brown, G., Omelchenko, M. V, Borozan, I., Carmel, L., Wolf, Y.I., Mori, H., Savchenko, A. V, et al. (2006). Genome-wide Analysis of Substrate Specificities of the *Escherichia coli* Haloacid Dehalogenase-like Phosphatase Family. *J. Biol. Chem.* 281, 36149–36161.

Kuznetsova, I.M., Zaslavsky, B.Y., Breydo, L., Turoverov, K.K., and Uversky, V.N. (2015). Beyond the excluded volume effects: Mechanistic complexity of the crowded

milieu. *Molecules* **20**, 1377–1409.

Levinthal, C. (1968). Are there pathways for protein folding? *Extr. Du J. Chim. Phys.* **65**, 44–45.

Levy, E.D. (2010). A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Biol.* **403**, 660–670.

Levy, E.D., De, S., and Teichmann, S.A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci.* **109**, 20461–20466.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Protein Disorder Prediction : Implications for Structural Proteomics. *Structure* **11**, 1453–1459.

Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., and Serrano, L. (2004). A comparative study of the relationship between protein structure and β -aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* **342**, 345–353.

Listwan, P., Terwilliger, T.C., and Waldo, G.S. (2009). Automated , high-throughput platform for protein solubility screening using a split-GFP system. *J Struct Funct Genomics* **10**, 47–55.

Lu, J., Zhang, X., Wu, Y., Sheng, Y., Li, W., and Wang, W. (2021). Energy landscape remodeling mechanism of Hsp70- chaperone-accelerated protein folding. *Biophys. J.* **120**, 1971–1983.

Lyon, C.E., Lamond, A.I., and Mann, M. (2005). Nucleolar proteome dynamics. *Nature* **433**, 77–83.

Määttä, T.A., Rettel, M., Helm, D., Stein, F., and Savitski, M.M. (2020). Aggregation and Disaggregation Features of the Human Proteome. *Mol Syst Biol* **16**.

Macossay-Castillo, M., Marvelli, G., Guharoy, M., Jain, A., Kihara, D., Tompa, P., and Wodak, S.J. (2019). The Balancing Act of Intrinsically Disordered Proteins : Enabling Functional Diversity while Minimizing Promiscuity. *J. Mol. Biol.* **431**, 1650–1670.

March, D., Bianco, V., and Franzese, G. (2021). Protein Unfolding and Aggregation near a Hydrophobic Interface. *Polymers (Basel)*. **13**, 1–14.

Martínez-Alonso, M., Vera, A., and Villaverde, A. (2007). Role of the chaperone DnaK in protein solubility and conformational quality in inclusion body-forming *Escherichia coli* cells. *FEMS Microbiol. Lett.* **273**, 187–195.

Mattice, W.L. (1989). The beta-Sheet to Coil Transition. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 93–111.

Mogk, A., Deuerling, E., Vorderwülbecke, S., Vierling, E., and Bukau, B. (2003). Small heat shock proteins , ClpB and the DnaK system form a functional triade in reversing protein aggregation. *Mol. Microbiol.* **50**, 585–595.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Anderson, P., Kim, H.J., Mittag, T., and Taylor, J.P. (2015). Phase separation by low complexity domains promotes stress

granule assembly and drives pathological fibrillization. *Cell* 163, 123–133.

Nahálka, J., Dib, I., and Nidetzky, B. (2008). Encapsulation of *Trigonopsis variabilis* D-amino acid oxidase and fast comparison of the operational stabilities of free and immobilized preparations of the enzyme. *Biotechnol. Bioeng.* 99, 251–260.

Nahálka, J. (2008). Physiological aggregation of maltodextrin phosphorylase from *Pyrococcus furiosus* and its application in a process of batch starch degradation to α -D-glucose-1-phosphate. *J Ind Microbiol Biotechnol* 35, 219–223.

Nahálka, J., Gemeiner, P., Bučko, M., Wang, P.G., Nahálka, J., Gemeiner, P., Bučko, M., Wang, P.G., and Wang, P.G. (2009). Bioenergy Beads : A Tool for Regeneration of ATP / NTP in Biocatalytic Synthesis ATP / NTP in Biocatalytic Synthesis. 1199.

Natan, E., Endoh, T., Haim-Vilmsky, L., Flock, T., Chalancon, G., Hopper, J.T.S., Kintses, B., Horvath, P., Daruka, L., Fekete, G., et al. (2018). Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat. Struct. Mol. Biol.* 25, 279–288.

Nieba, L., Honegger, A., and Krebber, C. (1997). Disrupting the hydrophobic patches at the antibody variable / constant domain interface : improved in vivo folding and physical characterization of an engineered scFv fragment. *Protein Eng. Vol.10* 10, 435–444.

Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* 19, 751–755.

Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci.* 106, 4201–4206.

Nölting, B., and Agard, D.A. (2008). How General Is The Nucleation – Condensation Mechanism ? *Proteins.* 73, 754–764.

Onuchic, J.N., Luthey-Schulten, Z., and Wolynes, P.G. (1997). Theory of protein folding: The Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* 48, 545–600.

Orfanoudaki, G., and Economou, A. (2014). Proteome-wide Subcellular Topologies of *E. coli* Polypeptides Database (STEPdb). *Mol. Cell. Proteomics* 13, 3674–3687.

Pace, C.N., Shirley, B.A., McNutt, M., and Gajiwala, K. (1996). Forces contributing to the conformational proteins. *FASEB J.* 10, 75–83.

Pál, C., Papp, B., and Hurst, L.D. (2001). Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* 158, 927–931.

Paliy, O., Gargac, S.M., Cheng, Y., Uversky, V.N., and Dunker, A.K. (2008). Protein disorder is positively correlated with gene expression in *E. coli*. *J Proteome Res* 7, 2234–2245.

Pechmann, S., Levy, E.D., Gaetano, G., and Vendruscolo, M. (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *PNAS* 106, 10159–10164.

Pédelacq, J., Piltch, E., Liong, E.C., Berendzen, J., Kim, C., Rho, B., Park, M.S., Terwilliger, T.C., and Waldo, G.S. (2002). Engineering soluble proteins for structural genomics. *Nat. Biotechnol.* 20, 927–932.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7, 1–17.

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., and Kurgan, L. (2015). Exceptionally abundant exceptions : comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* 72, 137–151.

Peternel, Š., and Komel, R. (2011). Active Protein Aggregates Produced in *Escherichia coli*. *Int. J. Mol. Sci.* 12, 8275–8287.

Piccinini, F., Balassa, T., Szkalitsy, A., Kutay, U., Smith, K., Horvath, P., Piccinini, F., Balassa, T., Szkalitsy, A., Molnar, C., et al. (2017). Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell Syst.* 4, 651-655.e5.

Piovesan, D., and Tosatto, S.C.E. (2018). Structural bioinformatics Mobi 2.0 : an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinformatics* 34, 122–123.

Plakoutsi, G., Stefani, M., and Chiti, F. (2004). Aggregation of the Acylphosphatase from *Sulfolobus solfataricus* THE FOLDED AND PARTIALLY UNFOLDED STATES CAN BOTH BE PRECURSORS FOR AMYLOID FORMATION. *J. Biol. Chem.* 279, 14111–14119.

Plaxco, K.W., Simons, K.T., and Baker, D. (1998). Contact Order , Transition State Placement and the Refolding Rates of Single Domain Proteins. *J. Mol. Biol.* 277, 985–994.

Price, W.N., Handelman, S.K., Everett, J.K., Tong, S.N., Bracic, A., Luff, J.D., Naumov, V., Acton, T., Manor, P., Xiao, R., et al. (2011). Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microb. Inform. Exp.* 1, 1–20.

Qiao, B., Jiménez-ángeles, F., Dac, T., Olvera, M., and Cruz, D. (2019). Water follows polar and nonpolar protein surface domains. *PNAS* 116, 19274–19281.

R Core Team (2008). R : A Language and Environment for Statistical Computing.

Richardson, J.S., and Richardson, D.C. (2002). Natural β -sheet proteins use negative design to avoid edge-to-edge aggregation.pdf. *PNAS* 99, 2754–2759.

Rinas, U., Garcia-fruítos, E., Corchero, J.L., Vázquez, E., Seras-franzoso, J., and Villaverde, A. (2017). Bacterial Inclusion Bodies : Discovering Their Better Half. *Trends Biochem. Sci.* 42, 726–737.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence Complexity of Disordered Protein. *PROTEINS Struct. Funct. Genet.* 48, 38–48.

Rosenberg, A.S. (2006). Effects of Protein Aggregates : An Immunologic Perspective. *AAPS J.* 8, 501–507.

Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-

dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306.

Rueda, F., Gasser, B., Chardi, A.S., Roldán, M., and Villegas, S. (2016). Functional inclusion bodies produced in the yeast *Pichia pastoris*. *Functional inclusion bodies produced in the yeast Pichia pastoris. Microb. Cell Fact.* **15**.

Sachsenhauser, V., and Bardwell, J.C. (2018). Directed evolution to improve protein folding in vivo. *Curr. Opin. Struct. Biol.* **48**, 117–123.

Santner, A.A., Croy, H.C., Vasanwala, H.F., Uversky, N.V., Van, J.Y.-Y., and Dunker, A.K. (2014). Sweeping away protein aggregation with entropic bristles: Intrinsically disordered protein fusions enhance soluble expression. *Biochemistry* **51**, 7250–7262.

Schad, E., Tompa, P., and Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biol.* **12**, R120.

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoop, K., Bauer, M., and Aebersold, R. (2016). The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol.* **34**, 104–110.

Schwartz, R., Istrail, S., and King, J. (2001). Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci.* **10**, 1023–1031.

Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299–311.

Sharma, M.C., and Goebel, H.H. (2005). Protein aggregate myopathies. *Neurol. India* **53**, 273–279.

Shmueli, M.D., Schnaider, L., Rosenblum, D., Herzog, G., Gazit, E., and Segal, D. (2013). Structural Insights into the Folding Defects of Oncogenic pVHL Lead to Correction of Its Function In Vitro. *PLoS One* **8**.

Sillitoe, I., Dawson, N., Lewis, T.E., Das, S., Lees, J.G., Ashford, P., Tolulope, A., Scholes, H.M., Senatorov, I., Bujan, A., et al. (2019). CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* **47**, 280–284.

Simona, J., Gaberc-Porekar, V., Fonda, I., Podobnik, B., Joze, G., and Viktor, M. (2005). Production of Nonclassical Inclusion Bodies from Which Correctly. *Biotechnol. Prog* **21**, 632–639.

Simone, A. De, Kitchen, C., Kwan, A.H., Sunde, M., Dobson, C.M., and Frenkel, D. (2012). Intrinsic disorder modulates protein self-assembly and aggregation. *PNAS* **109**, 6951–6956.

Singhvi, P., Saneja, A., Srichandan, S., and Panda, A.K. (2019). Bacterial Inclusion Bodies: A Treasure Trove of Bioactive Proteins. *Trends Biotechnol.* **3**, 1–13.

Sormanni, P., Aprile, F.A., and Vendruscolo, M. (2015). The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* **427**, 478–490.

Tartaglia, G.G., and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. Biosyst.* 5, 1873–1876.

Tartaglia, G.G., Pechmann, S., Dobson, C.M., and Vendruscolo, M. (2007). Life on the edge : a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* 32, 204–206.

Tartaglia, G.G., Pechmann, S., Dobson, C.M., and Vendruscolo, M. (2009). A Relationship between mRNA Expression Levels and Protein Solubility in *E. coli*. *J. Mol. Biol.* 388, 381–389.

Tayeb-Fligelman, E., Tabachnikov, O., Moshe, A., Goldshmidt-Tran, O., Sawaya, M.R., Coquelle, N., JP., C., and Landau, M. (2019). The cytotoxic *Staphylococcus aureus* PSM α 3 reveals a cross- α amyloid-like fibril. *Science* (80-.). 355, 831–833.

Timson, D.J., and Reece, R.J. (2003). Functional analysis of disease-causing mutations in human galactokinase. *Eur. J. Biochem* 1774, 1767–1774.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27.

Trevino, S.R., Scholtz, J.M., and Pace, C.N. (2009). Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J. Mol. Biol.* 366, 449–460.

Tucker, P.A., Tsernoglou, D., Tucker, A.D., Coenjaerts, F.E.J., Leenders, H., and Vliet, P.C. Van Der (1994). Crystal structure of the adenovirus DNA binding protein reveals a hook-on model for cooperative DNA binding. *EMBO J.* 13, 2994–3002.

Uemura, E., Niwa, T., Minami, S., Takemoto, K., Fukuchi, S., Machida, K., Imataka, H., Ueda, T., Ota, M., and Taguchi, H. (2018). Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Sci. Rep.* 8, 1–11.

Uversky, V.N. (2014). The triple power of D³ - protein intrinsic disorder in degenerative diseases. *Front. Biosci.* 19, 181–258.

Uversky, V.N. (2015). Paradoxes and wonders of intrinsic disorder: Prevalence of exceptionality. *Intrinsically Disord. Proteins* 3.

Uversky, V.N., Gillespie, J.R., and Fink, A.L. (2000). Why Are “Natively Unfolded” Proteins Unstructured Under Physiologic Conditions? *PROTEINS Struct. Funct. Genet.* 41, 415–427.

Valenzuela, P., Medina, A., and Rutter, W.J. (1982). Synthesis and assembly of hepatitis B virus surface antigen particles in yeast. *Nature* 298, 347–350.

Villaverde, A., and Carrio, M.M. (2001). Protein aggregation as bacterial inclusion bodies is reversible. *Febs Lett.* 489, 29–33.

Villaverde, A., García-Fruitós, E., Rinas, U., Seras-franzoso, J., Kosoy, A., Corchero, J.L., and Vazquez, E. (2012). Packaging protein drugs as bacterial inclusion bodies for therapeutic applications Packaging protein drugs as bacterial inclusion bodies for therapeutic applications. *Microb. Cell Fact.* 11, 1–4.

Wagner, S., Baars, L., Ytterberg, A.J., Klussmeier, A., Wagner, C.S., Nord, O., Nygren, P.-åke, Wijk, K.J. Van, and Gier, J. De (2007). Consequences of Membrane Protein Overexpression in *Escherichia coli*. *Mol. Cell. Proteomics* 6.9, 1527–1550.

Waldo, G.S., Standish, B.M., Berendzen, J., and Terwilliger, T.C. (1999). Rapid protein-folding assay using green fluorescent protein. *Nat. Biotechnol.* 17, 691–695.

Walsh, I., Martin, A.J.M., Domenico, T. Di, and Tosatto, S.C.E. (2012). ESpritz : accurate and fast prediction of protein disorder. *Bioinformatics* 28, 503–509.

Walsh, I., Seno, F., Tosatto, S.C.E., and Trovato, A. (2014). PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* 42, 301–307.

Warwicker, J., Charonis, S., and Curtis, R.A. (2014a). Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Mol. Pharm.* 11, 294–303.

Warwicker, J., Charonis, S., and Curtis, R.A. (2014b). Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design.

Wetzel, R., Perry, L.J., and Veilleux, C. (1991). Mutations in human interferon gamma affecting inclusion body formation identified by a general immunochemical screen. *Biotechnology* 9, 731–737.

Wickner, R.B. (1994). [URE3] as an Altered URE2 Protein: Evidence for a Prion Analog in *Saccharomyces cerevisiae*. *Science* (80-.). 264, 566–570.

Woodruff, J.B., Gomes, B.F., Widlund, P.O., Mahamid, J., Honigmann, A., Hyman, A.A., Woodruff, J.B., Gomes, B.F., Widlund, P.O., Mahamid, J., et al. (2017). The Centrosome Is a Selective Condensate that Nucleates Microtubules by Concentrating Tubulin Article The Centrosome Is a Selective Condensate that Nucleates Microtubules by Concentrating Tubulin. *Cell* 169, 1066-1071.e10.

Worrall, D.M., and Goss, N.H. (1989). The formation of biologically active beta-galactosidase inclusion bodies in *Escherichia coli*. *Aust J Biotechnol* 3, 28–32.

Wu, C.F., Cha, H.J., Rao, G., Valdes, J., and Bentley, W.E. (2000). A green fluorescent protein fusion strategy for monitoring the expression, cellular location, and separation of biologically active organophosphorus hydrolase. *Appl. Microbiol. Biotechnol.* 54, 78–83.

Wu, W., Xing, L., Zhou, B., and Lin, Z. (2011). Active protein aggregates induced by terminally attached self-assembling peptide ELK16 in *Escherichia coli*. *Microb. Cell Fact.* 42, 1–8.

Xu, D., and Zhang, Y. (2013). Ab Initio structure prediction for *Escherichia coli*: towards genome-wide protein structure modeling and fold assignment. *Sci. Rep.* 3, 1–11.

Yamaguchi, H., and Miyazaki, M. (2014). Refolding Techniques for Recovering Biologically Active Recombinant Proteins from Inclusion Bodies. 235–251.

Yang, J.S., Chen, W.W., Skolnick, J., and Shakhnovich, E.I. (2007). All-Atom Ab Initio Folding of a Diverse Set of Proteins. *Structure* 15, 53–63.

Yuan, A.H., and Hochschild, A. (2017). A bacterial global regulator forms a prion. *Science* (80-.). 355, 198–201.

Zhang, A., Cantor, E.J., Barshevsky, T., and T, S.C. (2005). Productive interaction of chaperones with substrate protein domains allows correct folding of the downstream GFP domain. *Gene* 350, 25–31.

Zhou, H., and Zhou, Y. (2002). Folding Rate Prediction Using Total Contact Distance. *Biophys. J.* 82, 458–463.

Zhu, C., and Yu, Z. (2008). The surface layer protein of *Bacillus thuringiensis* CTC forms unique intracellular parasporal inclusion body. *J. Basic Microbiol.* 48, 302–307.

8. Összefoglaló

Az elmúlt évtizedekben széles körben elfogadott volt a nézet, miszerint a fehérje aggregátumok belsejében natív szerkezettől és funkciótól mentes makromolekulák találhatók. Ez köszönhető volt a fehérje aggregátumok felfedezésének körülményeinek, valamint a humán szempontból fontos betegségekben betöltött fontos szerepüknek. Az elmúlt években azonban egyre több olyan eredmény látott napvilágot, amelyek szerint a fehérje aggregációnak fontos fiziológiás szerepe lehet mind baktériumokban, mind eukariótákban. Később arra is fény derült, hogy léteznek olyan aggregátumok, amelyekben a fehérjék megőrzik natív, vagy natív szerű szerkezetüket és aktivitásukat. Ennek ellenére nem történek olyan szisztematikus vizsgálatok, amelyek megmutatták volna, hogy mely tényezők befolyásolják a képződésüket.

A doktori munkám során szisztematikusan vizsgáltam a különböző *in vivo* aggregációs útvonalakat *E. coliban* fehérje overexpresszió hatására egy mikroszkópia alapú, gépi tanulással segített módszerrel. Ennek segítségével igazoltuk a két lényegesen különböző aggregáció típus jelenlétét a sejtekben. Továbbá proteom szinten azonosítottuk azokat a fehérjéket, amelyek a folding folyamatát megelőzően vagy azt követően léptek aggregátumokba. A két aggregációs típus gyakorisága összemérhető proteom szinten, azonban jelentős különbségeket látunk a sejten belüli lokalizációjuk tekintetében. Míg a sötét aggregátumokat képző fehérjék jellemzően a membránban lokalizálódnak fiziológiás körülmények között, a citoplazma aggregálódó fehérjeinek nagy része a folding folyamatát követően csapódik ki. Annak érdekében, hogy azonosíthassuk azokat a fehérjeparamétereket, amelyek meghatározzák a két aggregációs típus képződését, összeállítottunk egy szisztematikus adatsort a proteom legtöbb tagjának tulajdonságairól. Azt találtuk, hogy a natív konformációjuk elérése előtt aggregálódó fehérjékkel szemben a fluoreszcensen aggregálódó proteinek gyorsabb foldingkinetikát és kisebb mértékű fehérjeaggregációt mutatnak. A dajkafehérje-rendszerrel, különösen a DnaK fehérjével mutatott interakció szintén segít a fehérjéknek elérni a natív szerkezetüket a translációs környezetben az aggregációt megelőzően. Eredményeink arra utalnak, hogy a gyors folding és a lassan bekövetkező aggregáció lehetővé teszi egy fehérje számára, hogy elérje natív konformációját az aggregáció bekövetkezése előtt.

Összességében ezek az eredmények arra engednek következtetni, hogy az aktív, natív vagy natív szerű szerkezettel bíró aggregátumok létrejötte az inaktívakkal szemben nagy mértékben az aggregáció és a folding közti kinetikai kompetíció kimenetelének következménye (de Groot és Ventura, 2006). Több bizonyíték is alátámasztja azt, hogy ez a folyamat normál expresszió és genetikai kontextus mellett is jelentőséggel bír. Például a dajkafehérje-rendszer alapvető szereppel bír az újonnan szintetizálódott és a foldingon átesett fehérjék esetében is normál expressziós szint mellett (Kim és mtsai., 2013). Továbbá a DnaK dajkafehérje klienseinek is két típusa ismert annak függvényében, hogy a chaperon a transzláció során segít a natív konformáció elérésében, vagy ezt követően segít megtartani a szerkezetet (Calloni és mtsai., 2012). Míg a transzlációs apparátus környezete feldúsul aggregációra érzékeny naszcens polipeptidekben, több fehérje, amely elérte natív szerkezetét, szintén képes aggregálódni, amennyiben koncentrációja átlépi a kritikus oldhatósági határát; ezt a jelenséget szupersaturációnak hívjuk (Ciryam és mtsai., 2015). A szupersaturációt mutató fehérjékről kiderült, hogy feldúsulnak neurodegeneratív betegségekben szerepet játszó proteinek között, amely arra utal, hogy szerepük van ezekben a patológiákban és feltehetően az öregedéssel és a stresszhez köthetőkben is. Szükséges további kutatások elvégzése annak érdekében, hogy megérthessük a változó környezeti tényezők és az overexpresszió hatását az aggregációs útvonalra.

Ismert, hogy a fiziológiás körülmények között abundáns fehérjék rendkívül alacsony felszíni ragadóssággal bírnak azért, hogy az elkerüljék a nemspecifikus interakciókat, amelyek zavart okozhat az ő és más fehérjék működésében (Levy és mtsai., 2012). Eredményeink rávilágítottak, hogy a magas felszíni ragadóssággal rendelkező fehérjék szintén alulreprezentáltak az oldható csoportban, arra utalva, hogy a nemspecifikus interakciók csökkentése mögött az aggregáció elkerülése lehet az egyik legfontosabb hajtóerő. Ezen felül azonban azt is láthatjuk, hogy a sötét aggregálódó fehérjék marginálisan magasabb felszíni ragadóssággal bírnak a fluoreszcensen aggregálódó társaikhoz képest. Mivel a fehérjék belsejében nem látunk ilyen jellegű különbséget, ez arra utalhat, hogy bizonyos sötét aggregálódó fehérjék esetén a felszíni ragadósság is hozzájárulhat az aggregáció folyamatához. Ez úgy következhet be, hogy az aggregáció az előtt következik be, hogy a fúziós GFP partner elérné natív szerkezetét

(ismerve a kísérletekben tapasztalt fenotípust), de azt követően, hogy a naszcens fehérje szerkezete összeomlott volna egy natív szerű, kompakt szerkezetté, amely felszíne a natív fehérjéhez már részben hasonlatos (Hartl és Hayer-Hartl, 2009; Kim és mtsai., 2013). További vizsgálatok minden bizonnyal további információkkal szolgálhatnak arra nézve, hogy milyen sorrendben következnek be az aggregáció folyamat lépései ezeknél a fehérjéknél.

Munkánk a fehérjerendezetlenség aggregációban játszott szerepével kapcsolatban is új információkkal szolgált. Először is bemutattuk, hogy az *E. coli* fehérjék *in vivo* szolubilitását nagy mértékben segíti a fehérjerendezetlenség, és ez a hatás független a felszíni ragadóságtól. Továbbá, mivel a fehérjerendezetlenség és a szolubilitás közti korreláció független a töltéstől és a hidrofobicitástól, meg tudjuk erősíteni a nézetet, amely szerint a fehérjerendezetlenség a szerkezeti rugalmasság biztosítása mellett járul hozzá az oldhatósághoz (Santner és mtsai., 2014; Simone és mtsai., 2012). A rendezetlen fehérje régiókról ismert, hogy egyfajta entrópikus sörteként viselkednek, valamint oldhatóság szempontjából kedvező felszínt biztosítanak. Fontos megjegyezni, hogy az oldható fehérjék több rendezetlen régiót is tartalmaznak, tovább erősítve az entrópikus hatások szerepét. Továbbá, bár a két aggregációs osztály képviselői nem különböznek a rendezetlen régiók gyakoriságában, a sötét aggregálódó fehérjék több rendezetlen aminosavat tartalmaznak ezeken a régiókon kívül. További vizsgálatok szükségesek, amelyekből megtudhatjuk a rendezetlen aminosavak és régiók egyedi hozzájárulását az oldhatóság fenntartásához és az aggregációhoz. Másodsorban eredményeink arra is rávilágítanak, hogy a természetes szelekció hogyan befolyásolja a fehérjerendezetlenség mértékét proteom szinten. Ugyan ismert, hogy a magasan expresszált fehérjék nagyobb mértékű rendezetlenséggel bírnak *E. coliban* (Tartaglia és mtsai., 2009), az emögött álló mechanizmusok eddig ismeretlenek maradtak. Eredményeinkből arra következtethetünk, hogy a megnövekedett fehérjerendezetlenség a magasan expresszált fehérjék esetében arra szolgál, hogy elkerüljék az aggregációt fiziológiai körülmények között. Így tehát mind az alacsony felszíni ragadóság, mind a fehérjerendezetlenség arra szolgál, hogy az abundáns fehérjék elkerüljék az aggregációt *E. coliban*, valamint feltehetően további baktériumban is. Harmadsorban pedig jelentős különbséget láthatunk a prokarióták és eukarióták esetében a rendezetlenség aggregációra gyakorolt hatásában. Az eukariótákról

közismert, hogy jóval több rendezetlen fehérjét tartalmaznak és több rendezetlen szegmenst is. Ez köszönhető a jelátvitelben (Peng és mtsai., 2015) és egyéb változatos, átmeneti fehérje interakciókat igénylő sejt folyamatokban betöltött szerepüknek. Ezzel egybevé, hogy igazolódott a rendezetlen régiók szerepe a stresszindukált aggregációs folyamatokban (Määttä és mtsai., 2020) ember esetében. Továbbá szemben az *E. coli*-al, a magasan expresszált élesztő fehérjék szignifikánsan alacsonyabb mennyiségű rendezetlenséget mutatnak, feltehetően a lineáris motívumok közvetítette nemspecifikus interakciók elkerülése érdekében (Macossay-Castillo és mtsai., 2019). Ez a különbség elvben adódhat a fehérjék tulajdonságaiból, a sejt környezetből vagy a kettő kombinációjából. Azonban azt látjuk, hogy az *E. coli*-ban expresszált humán fehérjék is mutatják a rendezetlenség és az oldhatóság közti pozitív korrelációt, amely arra utal, hogy a különbség a sejt környezetből adódik. A különbségek mögött rejlő mechanizmusok feltárása jelentős mértékben hozzájárulhat a fehérje aggregáció folyamatának és a rendezetlenség szerepének megértéséhez, valamint fontos mérföldkőként szolgálhatnának jövőbeni nagy áteresztőképességű vizsgálatokhoz.

Eredményeink a biotechnológiai ipar számára is fontos implikációkkal bírnak. Új fehérjetermelési megoldások és rendezőelvek kerülhetnek előtérbe (Rinas és mtsai., 2017; Singhvi és mtsai., 2019; Villaverde és mtsai., 2012; Wu és mtsai., 2011), amennyiben megoldottá válik a funkcionális fehérjék kinyerése az aggregátumokból. A natív szerű enzimek felhasználhatóak immobilizált biokatalizátorként (Rinas és mtsai., 2017; Wu és mtsai., 2011) ipari környezetben. Az aggregátumok vivőanyagként is szolgálhatnak terápiás felhasználások esetében is (Villaverde és mtsai., 2012), mivel nagyobb stabilitással és eltarthatósággal bírnak. Azt látjuk, hogy minden enzimosztály képviselői jellemzően fluoreszcens aggregátumot képeznek, amely a széles körű felhasználhatóságot sugallnak ipari alkalmazásokban. Metodológiai szempontból a nagy áteresztőképességű mikroszkópia skálázható, hatékony és robosztus módszernek bizonyult az aktív fehérjéket tartalmazó aggregátumok azonosításában ipari célokra. Fontos megjegyezni, hogy bizonyos fehérjék az aktív aggregátumokban idővel inaktív, hibásan feltekeredett formává alakulnak (Elia és mtsai., 2017), ezért a jövőbeli vizsgálatoknak nagy hangsúlyt kell helyezniük azokra a hatásokra, amelyek segítségével egy fehérje megőrizheti aktivitását, továbbá arra, hogy mekkora az aktív és inaktív

fehérjék pontos részaránya az aggregátumok belsejében. A nagy áteresztőképességű mikroszkópiás és gépi tanulós módszerünk, kiegészítve nagy áteresztőképességű fehérjetisztítási (Jäger és mtsai., 2020) és funkcionális vizsgálati eljárásokkal (Huang és mtsai., 2015; Kuznetsova és mtsai., 2006) fontos részét képezheti az aktív fehérjeaggregátumokban rejlő lehetőségek kiaknázásának.

9. Summary

In the past decades, the notion that proteins inside aggregates are macromolecules lacking their native structure and function has been widely accepted. This was partly due to the circumstances of the discovery of protein aggregates and their pivotal role in human diseases. However, in recent years, more and more results were presented showing that protein aggregation may play an important physiological role in both bacteria and eukaryotes. Later, protein aggregates with enzymes retaining their native activity were identified. This prompted a major change in how we view protein aggregation. Despite the mounting evidence of the existence of aggregates with proteins maintaining their native or native-like structure with activity, no systematic analyses were conducted in order to identify the underlying factors for their formation.

My doctoral work focused on systematically investigating the aggregation pathways of overexpressed *E. coli* proteins using an image-based *in vivo* aggregation assay coupled with machine learning. We have confirmed the presence of two characteristically different types of protein aggregates. Furthermore, we identified on a proteome-wide scale those proteins that form inclusion bodies after folding and compared to those that aggregate before reaching the folded state. We find that the frequency of the two types of aggregates are comparable on a proteome scale, however, their subcellular localization differs substantially. While dark aggregates are mainly proteins present in the membrane under normal circumstances, the majority of aggregation-prone cytoplasmic proteins form aggregates after folding. We have collected a large-scale, systematic dataset of protein parameters in order to identify the forces affecting the fate of an aggregation-prone protein. We see that compared to proteins that aggregate before adopting a folded state, fluorescent aggregate proteins exhibit faster folding kinetics and lower aggregation propensity. Interaction with the chaperone network, especially the *DnaK* chaperone also helps to maintain the folded state in the translation milieu before, and most likely after aggregation. Our findings indicate that rapid folding and slow aggregation of a protein allows its folding before aggregation. Overall, these results support the notion that the formation of active, native-like aggregates compared to inactive ones is governed by the kinetic competition between

folding and aggregation. Several lines of evidence also reinforce the notion that these principles operate in proteins expressed at their expression levels and in their native genetic contexts. For example, under normal expression levels, the chaperon system plays a fundamental role in preventing aggregation of newly synthesized as well as folded proteins. Thus, two classes of DnaK client proteins have been identified, depending on whether the chaperone aids in its folding after translation, or maintains its native conformation. Additionally, while the proximity of the translational apparatus is highly enriched in nascent polypeptides susceptible to aggregation during translation, some proteins with their native structure intact has been shown to aggregate when their concentration exceeds the critical concentration of their solubility, a phenomenon termed supersaturation. Proteins showing supersaturation have shown a substantial enrichment in biological processes related to neurodegenerative diseases, suggesting their role in these pathologies, and others related to aging and stress. Future studies deciphering the links between the aggregation route taken by proteins upon overexpression and environmental changes are thus strongly warranted.

It has been established, that natively abundant proteins have evolved a substantially low surface stickiness in order to avoid non-specific interactions that would result in interference with other proteins(Levy és mtsai., 2012). Our work has demonstrated that proteins with low surface stickiness are also depleted in both proteins forming aggregates through either pathway. Based on these results, we propose that avoidance of aggregation is the driving force behind the reduction of non-specific interaction propensity of natively highly expressed proteins. Interestingly, we find that dark aggregate proteins have a distinctly higher surface stickiness than fluorescent aggregate proteins, however, no such difference can be found for amino acids in the protein core. This suggests that aggregation occurs before the GFP fusion protein adopts its native structure, but after the fast collapse of the nascent polypeptide chain into a structure resembling a partially folded, compact intermediate that already has similar surface characteristics of the native protein(Hartl és Hayer-Hartl, 2009; Kim és mtsai., 2013). Further investigations will most probably shed some light on the sequence of events leading to the aggregation of such proteins.

Our work also provides several new insights into the role protein disorder plays in aggregation. First, we show that protein *in vivo* solubility in *E. coli* is strongly driven by protein disorder content, and this effect is independent of surface stickiness. Given that the association between protein disorder and solubility is also independent of both charge and hydrophobicity, we reinforce the notion that the flexibility provided by the disordered residues enhances solubility (Santner és mtsai., 2014; Simone és mtsai., 2012). Disordered protein segments have been shown to act as entropic bristles and providing favourable surface area for protein solubility. It is important to note that soluble proteins also contain a larger frequency of disordered segments, supporting their positive entropic effects. Also, while proteins of two aggregation classes do not differ in the frequency of disordered segments, dark aggregates contain more disordered residues outside these segments. Future studies should shed more light on the particular effects of disordered residues and segments and their contribution to the maintenance of solubility or contribution to aggregation. Second, our results shed a new light on how natural selection affects protein disorder on a proteome scale. Although it has been shown that highly expressed proteins have a higher disorder content in *E. coli* (Tartaglia és mtsai., 2009), the underlying mechanisms have been largely unidentified so far. Our findings suggest that the increased disordered content of highly expressed proteins is a means of avoiding aggregation under physiological conditions. Thus, protein disorder and low surface stickiness are both adaptations to maintain high solubility of abundant proteins in *E. coli* and possibly in other bacteria. And third, our results highlight the substantial difference of protein disorder on solubility in prokaryotes and eukaryotes. Eukaryotes are known to contain substantially higher amounts of disordered proteins as well as longer disordered segments. This is due to the role of these proteins in signaling (Peng és mtsai., 2015) and other cellular functions requiring diverse and transient protein interactions. Indeed, the presence of intrinsically disordered regions have been associated with stress-induced aggregation in humans (Määttä és mtsai., 2020). Also, contrarily to *E. coli*, highly expressed yeast proteins contain significantly lower levels of intrinsically disorder, potentially to reduce non-native, non-specific interactions that are mediated by linear motifs (Macossay-Castillo és mtsai., 2019). These difference can be attributed to either the proteins themselves, the cellular milieu or a combination of both. However, our results show that human proteins expressed in *E. coli* adhere to positive the correlation between

solubility and disorder content, suggesting that the difference can be attributed to differences in the cellular context. The underlying mechanisms for such differences would be a great benefit to our understanding of both protein disorder and protein aggregation, as well as providing an important framework for future large scale studies.

Our results may also have substantial biotechnological implications. Extraction of functional proteins from inclusion bodies as well as their other applications may offer novel solutions and design principles (Rinas és mtsai., 2017; Singhvi és mtsai., 2019; Villaverde és mtsai., 2012; Wu és mtsai., 2011) to the protein production pipeline. Native-like enzymes can be applied as immobilized biocatalysts (Rinas és mtsai., 2017; Wu és mtsai., 2011) in industrial applications. Inclusion bodies may also function as drug delivery scaffolds in medical applications (Villaverde és mtsai., 2012), with the benefit of greater stability and longer shelf-life. We show that proteins of all enzyme classes tend to form aggregates after folding, suggesting the wide applicability of designed protein aggregates in industrial use. From a methodological standpoint, high-content microscopy has proven to be a highly scalable, efficient and robust way of identifying aggregates with active proteins, with great potential in screening of recombinant proteins for industrial applications. It is important to mention, that some proteins in active aggregates may in time convert to misfolded, inactive forms (Elia és mtsai., 2017), thus future investigations must put an emphasis on the forces that help maintain protein activity inside aggregates and the exact proportion of active and inactive proteins inside aggregates. Our high-content microscopy and machine learning method, combined with high-throughput protein purification (Jäger és mtsai., 2020) and functional profiling (Huang és mtsai., 2015; Kuznetsova és mtsai., 2006) may provide an efficient way of unlocking the potential in active protein aggregates.

10. Függelékek

1. számú táblázat: a fehérjeaggregáció és a génexpresszió kísérletes ellenőrzése

JW azonosító	Blattner ID	Génnev	Mikroszkópia alapú szolubilitásosztály	Molekulatömeg (Da)	Molekulatömeg (+GFP /kDa/)	Expresszió	GFP Intenzitás	Kísérletes szolubilitás
JW0899	b0916	ycaQ	Sötét aggregátum	47655,87	76,7	Igen	254081	NA
JW0868	b0885	aat	Sötét aggregátum	26618,5	55,6	Igen	388531	NA
JW0824	b0840	deoR	Sötét aggregátum	28547,98	57,5	Igen	386105	Aggregálódó
JW0865	b0881	clpS	Sötét aggregátum	12179,06	41,2	Igen	674041	NA
JW0928	b0945	pyrD	Sötét aggregátum	36774,52	65,8	Igen	491562	Aggregálódó
JW0993	b1008	ycdI	Sötét aggregátum	21570,38	50,6	Igen	597674	NA
JW0945	b0962	helD	Sötét aggregátum	77976,2	107	NA	282558	NA
JW0986	b1001	yccE	Sötét aggregátum	48355,37	77,4	Igen	418709	NA
JW0939	b0956	ycbG	Sötét aggregátum	17693,37	46,7	Igen	536276	Aggregálódó
JW0941	b0958	sulA	Sötét aggregátum	18801,38	47,8	Nem	39437	NA
JW2459	b2474	ypfI	Sötét aggregátum	74892,52	103,9	Nem	4320	NA
JW1005	b1020	phoH	Sötét aggregátum	39264,87	68,3	Igen	399542	NA
JW2815	b2847	yqeI	Sötét aggregátum	30850,25	59,9	Igen	79757	NA
JW1276	b1284	yciT	Sötét aggregátum	27602,61	56,6	Igen	383909	Aggregálódó
JW1306	b1313	ycjQ	Sötét aggregátum	38216,52	67,2	Igen	535846	NA
JW1392	b1397	paaI	Sötét aggregátum	42276,54	71,3	Igen	508303	NA
JW1316	b1323	tyrR	Sötét aggregátum	57656,14	86,7	Nem	4979	NA
JW1605	b1613	manA	Sötét aggregátum	42849,95	71,8	Igen	668899	NA
JW1606	b1614	ydgA	Sötét aggregátum	54689	83,7	Igen	511969	NA
JW1520	b1527	yneK	Sötét aggregátum	42841,25	71,8	Igen	245379	NA
JW2021	b2036	glf	Sötét aggregátum	42965,88	72	Igen	278316	NA
JW2029	b2044	wcaL	Sötét aggregátum	45361,62	74,4	Igen	309098	NA
JW1993	b2011	sbcB	Sötét aggregátum	54500,81	83,5	Igen	270320	NA
JW1912	b1927	amyA	Sötét aggregátum	56639,38	85,6	Igen	501012	Gyenge oldhatóság

JW2017	b2032	wbbK	Sötét aggregátum	43188,18	72,2	Igen	93111	NA
JW1209	b1218	chaC	Sötét aggregátum	25512,28	54,5	Igen	470245	NA
JW1086	b1100	ycfH	Sötét aggregátum	29808,78	58,8	Igen	423125	Aggregálódó
JW1106	b1120	cobB	Sötét aggregátum	26716,38	55,7	Igen	45848	NA
JW1054	b1067	yceH	Sötét aggregátum	24177,33	53,2	Igen	611099	NA
JW1139	b1153	ymfQ	Sötét aggregátum	21643,52	50,6	Igen	880443	Aggregálódó
JW1292	b1299	puuR	Sötét aggregátum	20091,58	49,1	Igen	732922	NA
JW1587	b1595	ynfL	Sötét aggregátum	33281,34	62,3	Igen	463536	NA
JW1530	b1537	ydeJ	Sötét aggregátum	18320,82	47,3	Igen	687982	NA
JW1423	b1427	rimL	Sötét aggregátum	20680,55	49,7	Igen	626071	NA
JW1512	b1519	tam	Sötét aggregátum	29005,96	58	Igen	614034	NA
JW1625	b1633	nth	Sötét aggregátum	23562,39	52,6	Igen	611057	NA
JW1664	b1674	ydhY	Sötét aggregátum	22366,92	51,4	Igen	587405	NA
JW1759	b1770	ydjF	Sötét aggregátum	28261,47	57,3	Igen	450231	NA
JW1832	b1843	yobB	Sötét aggregátum	24395,06	53,4	Igen	289881	NA
JW2053	b2068	alkA	Sötét aggregátum	31393,3	60,4	Igen	278496	NA
JW2006	b2024	hisA	Sötét aggregátum	26032,65	55	Igen	559163	NA
JW2057	b2072	yegK	Sötét aggregátum	27207,94	56,2	Igen	532136	NA
JW1886	b1897	otsB	Sötét aggregátum	29175,28	58,2	Igen	362828	NA
JW2043	b2058	wcaB	Sötét aggregátum	17615,54	46,6	Igen	403315	NA
JW1544	b1552	cspl	Sötét aggregátum	7683,74	36,7	Igen	197413	NA
JW1210	b1219	ychN	Sötét aggregátum	12692,75	41,7	Igen	573461	Gyenge oldhatóság
JW1456	b1461	ydcE	Sötét aggregátum	8672,95	37,7	Igen	578679	NA
JW1553	b1561	rem	Sötét aggregátum	9234,71	38,2	Igen	34869	NA
JW1676	b1686	ydil	Sötét aggregátum	14945,25	43,9	Igen	532552	NA
JW1566	b1575	dicB	Sötét aggregátum	6964,27	36	Igen	191630	NA
JW1567	b1576	ydfD	Sötét aggregátum	6841,95	35,8	Igen	537089	NA
JW1854	b1865	nudB	Sötét aggregátum	17305,76	46,3	Igen	337976	NA
JW2928	b2961	mutY	Sötét aggregátum	39149,12	68,1	Igen	325407	NA
JW3089	b3118	tdcA	Sötét aggregátum	34538,86	63,5	Igen	296499	NA

JW3093	b3124	garK	Sötét aggregátum	39104,34	68,1	Igen	653722	NA
JW2231	b2237	inaA	Sötét aggregátum	25296,08	54,3	Igen	449114	NA
JW2413	b2420	yfeS	Sötét aggregátum	28243,15	57,2	Igen	504211	Aggregálódó
JW2451	b2467	yffH	Sötét aggregátum	21748,66	50,7	Igen	NA	NA
JW2357	b2360	yfdQ	Sötét aggregátum	30442,51	59,4	Igen	NA	NA
JW2396	b2405	xapR	Sötét aggregátum	33627,33	62,6	NA	NA	NA
JW2298	b2301	yfcF	Sötét aggregátum	24325,74	53,3	Igen	NA	Szolubilis
JW2337	b2340	sixA	Sötét aggregátum	17207,77	46,2	Igen	NA	NA
JW2726	b2756	ygcH	Sötét aggregátum	22292,99	51,3	Igen	NA	NA
JW2596	b2615	uppS	Sötét aggregátum	32566,38	61,6	Igen	NA	NA
JW2683	b2713	hydN	Sötét aggregátum	19026,13	48	Igen	NA	NA
JW2709	b2739	ygbM	Sötét aggregátum	29216,99	58,2	Igen	NA	NA
JW2505	b2521	sseA	Sötét aggregátum	30811,84	59,8	Nem	NA	NA
JW2528	b2544	yphB	Sötét aggregátum	32716,46	61,7	Igen	NA	NA
JW2704	b2734	pphB	Sötét aggregátum	25098,25	54,1	Igen	NA	NA
JW2575	b2593	yfiH	Sötét aggregátum	26338,93	55,3	Igen	NA	NA
JW2676	b2707	srlR	Sötét aggregátum	28236,15	57,2	Igen	NA	NA
JW2993	b3025	qseB	Sötét aggregátum	24677,59	53,7	Igen	NA	Szolubilis
JW2972	b3003	yghA	Sötét aggregátum	31487,97	60,5	Igen	NA	Gyenge oldhatóság
JW2913	b2946	yggJ	Sötét aggregátum	26978,06	56	Igen	NA	Aggregálódó
JW2475	b2490	hyfJ	Sötét aggregátum	15576,98	44,6	Igen	NA	Gyenge oldhatóság
JW3077	b3106	yhaK	Sötét aggregátum	25895,48	54,9	Igen	NA	NA
JW2232	b2238	yfaH	Sötét aggregátum	8866,53	37,9	Igen	NA	NA
JW2408	b2415	ptsH	Sötét aggregátum	9119,37	38,1	Igen	NA	NA
JW2753	b2782	chpA	Sötét aggregátum	12098,04	41,1	Igen	NA	NA
JW2675	b2706	gutM	Sötét aggregátum	12953,25	42	Igen	NA	NA
JW2906	b2939	yqgB	Sötét aggregátum	4859	33,9	Igen	NA	NA
JW2600	b2619	yfjG	Sötét aggregátum	17726,67	46,7	Nem	NA	NA
JW3627	b3652	recG	Sötét aggregátum	76430,37	105,4	Igen	NA	NA
JW3395	b3432	glgB	Sötét aggregátum	84336,9	113,3	Igen	NA	NA

JW3037	b3065	rpsU	Sötét aggregátum	8499,96	37,5	Igen	NA	NA
JW3379	b3416	malQ	Sötét aggregátum	78503,49	107,5	Igen	NA	NA
JW3608	b3633	kdtA	Sötét aggregátum	47291,08	76,3	Igen	NA	NA
JW3598	b3623	rfaK	Sötét aggregátum	41729,46	70,7	Igen	NA	NA
JW3699	b3721	bgIB	Sötét aggregátum	53161,28	82,2	Igen	NA	NA
JW3557	b3585	ylaU	Sötét aggregátum	37274,46	66,3	Igen	NA	NA
JW3541	b3569	xyIR	Sötét aggregátum	44869,19	73,9	Igen	NA	NA
JW3212	b3243	aaeR	Sötét aggregátum	34516	63,5	Igen	NA	NA
JW3337	b3374	frlD	Sötét aggregátum	28331,85	57,3	Igen	NA	Aggregálódó
JW3318	b3355	prkB	Sötét aggregátum	32344	61,3	Igen	NA	Aggregálódó
JW3471	b3504	yhiS	Sötét aggregátum	47061,18	76,1	NA	NA	NA
JW3568	b3594	yibA	Sötét aggregátum	31873,61	60,9	Igen	NA	NA
JW3553	b3581	sgbH	Sötét aggregátum	23444,76	52,4	Igen	NA	NA
JW3690	b3712	yieE	Sötét aggregátum	27646,65	56,6	Igen	NA	NA
JW1441	b1446	ydcY	Szolubilis	8799,88	37,8	Igen	NA	Szolubilis
JW1208	b1217	chaB	Szolubilis	8944,83	37,9	Igen	NA	Szolubilis
JW1987	b2005	yeeV	Szolubilis	13898,94	42,9	Igen	NA	Szolubilis
JW2648	b2673	nrdH	Szolubilis	9139,46	38,1	Igen	NA	Szolubilis
JW3748	b3775	ppiC	Szolubilis	10232,02	39,2	Igen	NA	Szolubilis
JW5008	b0101	yacG	Szolubilis	7306,17	36,3	Igen	NA	Szolubilis
JW5060	b0435	bolA	Szolubilis	11993,68	41	Igen	NA	Szolubilis
JW0216	b0226	dinJ	Szolubilis	9405,08	38,4	Igen	NA	Szolubilis
JW5533	b3170	yhbC	Szolubilis	16650,97	45,7	Igen	NA	Szolubilis
JW5309	b1875	yecM	Szolubilis	21205,01	50,2	Igen	NA	Szolubilis
JW5733	b4135	yjdC	Szolubilis	21931,08	50,9	Igen	NA	Szolubilis
JW5404	b2522	sseB	Szolubilis	28642,52	57,6	Igen	NA	Szolubilis
JW0383	b0392	ykiA	Szolubilis	12943,71	41,9	Igen	NA	Szolubilis
JW0526	b0539	ybcC	Sötét aggregátum	11115,73	41,1	Igen	NA	Aggregálódó
JW0223	b0233	yafO	Sötét aggregátum	15487,84	45,5	Igen	NA	Aggregálódó
JW3278	b3316	rpsS	Szolubilis	10430,29	39,4	Igen	NA	Aggregálódó

JW0328	b0337	codA	Fluoreszcens aggregátum	47590,98	76,6	Igen	510696	Aggregálódó
JW0510	b0521	ybcF	Fluoreszcens aggregátum	31644,13	60,6	Igen	590989	Aggregálódó
JW1880	b1891	flhC	Fluoreszcens aggregátum	21566,09	50,6	Igen	600320	Aggregálódó
JW1989	b2007	yeeX	Fluoreszcens aggregátum	12777,78	41,8	Igen	666737	Aggregálódó
JW2730	b2760	ygcL	Fluoreszcens aggregátum	55901,19	84,9	Igen	329957	Aggregálódó
JW2400	b2409	yfeR	Fluoreszcens aggregátum	33903	62,9	Igen	438129	Aggregálódó
JW2415	b2422	cysA	Fluoreszcens aggregátum	41058,92	70,1	Igen	379865	Aggregálódó
JW2635	b2660	ygaF	Fluoreszcens aggregátum	46081,91	75,1	Igen	204628	Aggregálódó
JW2887	b2920	ygfH	Fluoreszcens aggregátum	53824,45	82,8	Igen	501443	Aggregálódó
JW2917	b2950	yggR	Fluoreszcens aggregátum	35994,16	65	Igen	445145	Aggregálódó
JW3392	b3429	glgA	Fluoreszcens aggregátum	52822,43	81,8	Igen	186797	Aggregálódó
JW3393	b3430	glgC	Fluoreszcens aggregátum	48697,64	77,7	Igen	323808	Aggregálódó
JW3428	b3463	ftsE	Fluoreszcens aggregátum	24439,23	53,4	Igen	471291	Aggregálódó
JW5476	b2921	ygfI	Fluoreszcens aggregátum	34673,67	63,7	Igen	496636	Aggregálódó

2. számú táblázat: A szisztematikus elemzésekhez használt fehérjeparaméterek , leírásuk és forrásaik.

Fehérjeparaméter neve	Leírás	Forrás
bnumber	Blattner ID	EcoGene
Uniprot	UniProtKB	EcoGene
ECK	K-12 Gene Accession Number	EcoGene
Gene	Génnév	EcoGene
GI	GenBank GI ID	EcoGene
SP	UniProtKB/Swiss-Prot azonosító	EcoGene
JW	GenoBase ID	EcoGene
RefSeqGI	GI-szám (EcoGene)	EcoGene
RefSeq	Refseq azonosító (EcoGene)	EcoGene
Type	Géntípus	EcoGene
Protein length	Proteinhossz	EcoGene
Ori	Orientáció	EcoGene
LeftEnd	5' pozíció	EcoGene
RightEnd	3' pozíció	EcoGene
Molecular weight	Molekulatömeg	EcoGene
Protein	Protein név	EcoGene
Function	Funkció	EcoGene
Gene essentiality (KEIO)	Knockout esszencialitás: esszenciális /E/, nem esszenciális /N/, ismeretlen /u/ (Keio kollekció)	DOI: 10.1038/msb4100050, Supplementary Table 6
Number of positive genetic interactions	Pozitív genetikai interakciók száma, Babu és mtsai. 2014	PMID: 24586182
Number of negative genetic interactions	Negatív genetikai interakciók száma, Babu és mtsai. 2014	PMID: 24586182
Number of genetic interactions	Összes genetikai interakció száma, Babu és mtsai. 2014	PMID: 24586182

Part of protein complex	Protein complex tag? (forrás: "Global functional atlas of <i>Escherichia coli</i> encompassing previously uncharacterized proteins")	PMID:19402753
Protein-protein interactions	Van interakciós partnere? Élesztő kéthibrid kísérlet	PMID:24561554
native expression level (mRNA)	mRNS expressziós szint glükóz minimál médiumban, aerob körülmények között (T=37C), logaritmizált, microarray kísérlet	PMID: 15129285
Solubility	<i>In vitro</i> szolubilitás (%)	PMID: 19251648
Fraction of aggregation-prone residues (AggreScan)	Aggregációs forrópont mérete hossza normálva	PMID:17324296
Aggregation energy (PASTA 2.0)	Aggregációs energia (Pasta 2.0)	PMID:24848016
Aggregation propensity (TANGO)	Aggregációs érték (Tango)	PMID: 16434184; 15361882
Fraction of charged residues	number of charged amino acids / total number of amino acids	saját script
Fraction of positively charged residues	Pozitívan töltött aminosavak aránya	saját script
Fraction of negatively charged residues	Negatívan töltött aminosavak aránya	saját script
Isoelectric point	Izoelektromos pont	PMID: 19251648
Fraction of hydrophobic amino acids	Hidrofób aminosavak aránya	saját script
COG_A	RNS processzálás és módosítás	PMID:12969510
COG_C	Energiatermelés és átalakítás	PMID:12969510
COG_D	Setcikluskontroll, sejtosztódás és kromoszóma partíciónálás	PMID:12969510
COG_E	Aminosav-metabolizmus és transzport	PMID:12969510
COG_F	Nukleotidtranszport és metabolizmus	PMID:12969510
COG_G	Szénhidráttranszport és metabolizmus	PMID:12969510
COG_H	Koenzimtranszport és metabolizmus	PMID:12969510
COG_I	Lipidtranszport és metabolizmus	PMID:12969510
COG_J	Transzláció, riboszóma szerkezet és biogenezis	PMID:12969510
COG_K	Transzkripció	PMID:12969510
COG_L	Replikáció, rekombináció és javítás	PMID:12969510

COG_M	Sejtfal, membrán, burok biogenezis	PMID:12969510
COG_N	Sejtmotilitás	PMID:12969510
COG_O	Poszttranszlációs modifikáció, fehérje turnover, chaperonok	PMID:12969510
COG_P	Szervetlenion-transzport és metabolizmus	PMID:12969510
COG_Q	Másodlagos metabolit biogenezis, transzport és lebontás	PMID:12969510
COG_R	Általános funkció (predikció)	PMID:12969510
COG_S	Ismeretlen funkció	PMID:12969510
COG_T	Szignáltranszdukció	PMID:12969510
COG_U	Belső transzport, szekréció, vezikuláris transzport	PMID:12969510
Defense mechanisms (COG V)	Védekezőmechanizmusok	PMID:12969510
Cellular processes (MultiFun)	Sejtfolyamatok, MultiFun	PMID:11471834
Cell structure (MultiFun)	Sejtszerkezet, MultiFun	PMID:11471834
Cryptic genes (MultiFun)	Kriptikus gének, MultiFun	PMID:11471834
DNA sites (MultiFun)	DNS-hez köthető funkció, MultiFun	PMID:11471834
Extrachromosomal proteins (MultiFun)	Extrakromoszómális, MultiFun	PMID:11471834
Information transfer (MultiFun)	Információtranszfer, MultiFun	PMID:11471834
Localization of gene products (MultiFun)	Géntermékek transzportja, MultiFun	PMID:11471834
Metabolism (MultiFun)	Metabolizmus, MultiFun	PMID:11471834
Regulation (MultiFun)	Szabályozás, MultiFun	PMID:11471834
Transport (MultiFun)	Transzport, MultiFun	PMID:11471834
DNA information transfer (MultiFun)	Információtranszfer, DNS-hez köthető, MultiFun	PMID:11471834
RNA Information transfer (MultiFun)	Információtranszfer, RNS-hez köthető, MultiFun	PMID:11471834
Information transfer protein (MultiFun)	Információtranszfer, fehérjéhez köthető, MultiFun	PMID:11471834
GroEL substrate	GroEL szubsztrát, class1: gyenge GroEL interakció, class2: részben GroEL-függő, class3: obligát GroEL szubsztrát, non-substrate: a többi	PMID: 16051146

	gén SwissProt azonosítóval	
Protein stickiness	Ragadósság Levy & Teichmann módszer alapján	PMID: 23184996
Fraction of surface arginines	Argininek aránya a felszínen, hosszra normálva	PMID: 23184996
Surface lysine to arginine ratio	Felszíni lizinek és argininek aránya	PMID: 23184996
Mainly alpha (structure)	Másodlagos szerkezeti szupercsalád a CATH adatbázis alapján	CATH
Orthogonal bundle (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Up-down Bundle (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Alpha Horseshoe (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Alpha/alpha barrel (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Mainly beta (structure)	Másodlagos szerkezeti szupercsalád a CATH adatbázis alapján	CATH
Ribbon (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
3-layer Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
5 Propellor (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
6 Propellor (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
7 Propellor (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
8 Propellor (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
3 Solenoid (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Beta Complex (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Single Sheet (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Beta barrel (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Clam (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Distorted Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Roll (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Ribosomal Protein LYES5 Chain K domain 2 (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH

Alpha-beta (structure)	Másodlagos szerkezeti szupercsalád a CATH adatbázis alapján	CATH
Alpha-Beta Barrel (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
2-Layer Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
3-Layer aba sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
3-Layer(bba) Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
3-Layer(bab) Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
4-Layer Sandwich (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Alpha-beta prism (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
5-stranded Propeller	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Alpha-Beta Horseshoe (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Alpha-Beta Complex (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Few secondary structures (structure)	Másodlagos szerkezeti szupercsalád a CATH adatbázis alapján	CATH
Irregular (structure)	Másodlagos szerkezeti család a CATH adatbázis alapján	CATH
Multidomain	Multidomén fehérje (CATH)	CATH
DnaK interactor	DnaK-szubsztrát (Hsp70)	PMID: 22832197
Protein surface	Fehérjefelszín mérete dssp-vel számolva (A ²)	PMID: 6667333
Surface positive patch size	A legnagyobb pozitív töltésű felszíni régió	PMID: 6667333
Protein hydropathy (Kyle-Doolittle)	Kyte Doolittle hidrofobicitás érték	
Average relative accessible surface area	Hosszra normált felszín mérete	PMID: 6667333
Protein hydropathy	Wimley-White hidrofobicitás érték (töltött aminosav oldalláncok)	PMID:25210196

localization_stepDB	Protein lokalizáció	PMID: 25210196
GRAVY	Átlagos hidrofobicitás	PMID:15955783
unfoldability	Foldingképtelenség (nagyobb érték, a fehérje kevésbé foldol)	PMID:27980099
Tig interactor	Tig chaperon interakció a BioGRID adatbázis alapján	PMID:27980099
ibpB chaperone interactor	ibpB chaperon interakció a BioGRID adatbázis alapján	PMID:27980099
ibpA interactor	ibpA chaperon interakció a BioGRID adatbázis alapján	PMID:27980099
HtpG interactor	HtpG chaperon interakció a BioGRID adatbázis alapján	PMID:27980099
GroES interactor	GroES chaperon interakció a BioGRID adatbázis alapján	PMID:27980099
Clpb interactor	Ckpb chaperon interakció a BioGRID adatbázis alapján	PMID: 23184996
Native expression level (protein)	Átlagos fehérjeexpresszió több mérési körülmény között	PMID: 9545386
Relative contact order	Hossznormalizált contact order	PMID: 16845101
Protein folding rate (FoldRate)	Folding ráta	PMID: 27797757
tAI	tRNA adaptációs index stAlcalc algoritmussal számolva	PMID:14604535
Fraction of disordered residues (DisEMBL hotloops)	Rendezetlenség: az aminosavak mekkora része alkot loopokat	PMID:14604535
Fraction of disordered residues (iupredS)hort	Az aminosavak mekkora aránya alkot rövid rendezetlen régiókat a páros aminosavinterakciók alapján	PMID: 15955779
Fraction of disordered residues (iupredL)	Az aminosavak mekkora aránya alkot hosszú rendezetlen régiókat a páros aminosavinterakciók alapján	PMID: 16618368
Fraction of disordered residues (PONDR VSL2B)	Rövid és hosszú rendezetlen régiókban található aminosavak aránya	PMID: 22190692
Fraction of disordered residues (EspritzNMR)	NMR méréseken alapuló fehérjerendezetlenség	PMCID: PMC6056769
native expFraction of disordered residues (EspritzX)	Röntgenkristallográfiás méréseken alapuló fehérjerendezetlenség	PMCID: PMC6056769
Protein melting temperature	Fehérje olvadási hőmérséklet	PMID: 29980614

3. számú táblázat: Enzimosztályok (EC-szám) feldúsulása a pre- és posztfolding aggregálódó csoportok között

EC-szám	Esélyhányados	p-érték	p-érték (FDR korrigált)
1	2.012181	0.02568	0.15408
2	0.911552	0.7328	0.87936
3	0.8797034	0.5716	0.87936
4	0.8577541	0.6968	0.87936
5	0.7381682	0.4202	0.87936
6	0.9105533	1	1

4. számú táblázat: a rendezetlen aminosavak részarányának összehasonlítása a szolubilis és az aggregálódó fehérjék között

A.) Szolubilis – posztfolding fehérjék

Rendezetlenségprediktor (rendezetlen aminosavak aránya)	A rendezetlenebb csoport fenotípusa	Logisztikus regressziós modell p-értéke	Referencia
DisEBMBL hotloops	Szolubilis	$P < 10^{-15}$	PMID:14604535
PONDR VSL2B	Szolubilis	$P < 10^{-15}$	PMID: 16618368
iUpred-Short	Szolubilis	$P < 10^{-15}$	PMID: 15955779
iUpred-Long	Szolubilis	$P = 1.196 \cdot 10^{-14}$	PMID: 15955779
espritz X-ray	Szolubilis	$P < 10^{-15}$	PMID: 22190692
Espritz NMR	Szolubilis	$P = 6.356 \cdot 10^{-6}$	PMID: 22190692

B.) Szolubilis – prefolding fehérjék

Rendezetlenségprediktor (rendezetlen aminosavak aránya)	A rendezetlenebb csoport fenotípusa	Logisztikus regressziós modell p-értéke	Referencia
DisEBMBL hotloops	Szolubilis	$P = 1.72 \cdot 10^{-9}$	PMID:14604535
PONDR VSL2B	Szolubilis	$P = 1.023 \cdot 10^{-10}$	PMID: 16618368
iUpred-Short	Szolubilis	$P = 9.848 \cdot 10^{-14}$	PMID: 15955779
iUpred-Long	Szolubilis	$P = 3.321 \cdot 10^{-5}$	PMID: 15955779
espritz X-ray	Szolubilis	$P = 1.335 \cdot 10^{-11}$	PMID: 22190692
Espritz NMR	Szolubilis	$P = 0.0005641$	PMID: 22190692

5. táblázat: A fehérjerendezetlenség és felszíni ragadósság hatása az oldhatóságra

A.) Szolubilis és posztfolding aggregálódó fehérjék

Többváltozós logisztikus regressziós modell:

szolubilitás ~ Fehérjerendezetlenség + fehérje felszíni ragadósság

A fehérjerendezetlenség parciális hatása az oldhatóságra

A fehérje felszíni ragadósság parciális hatása az oldhatóságra

Rendezetlenségprediktor	Koefficiens (fehérjerendezetlenség)	P-érték (fehérjerendezetlenség)	Koefficiens (fehérje felszíni ragadósság)	P-érték (fehérje felszíni ragadósság)
DisEMBL hotloops	3.3775	$< 10^{-15}$	-4.9307	2.57E-13
PONDR VSL2B	3.5277	1.30E-14	-4.8966	4.49E-13
iupredS	8.779	$< 10^{-15}$	-5.5713	2.39E-15
iupredL	4.822	2.23E-08	-5.8327	3.54E-12
Espritz X-ray	7.1925	5.76E-15	-5.6471	$< 10^{-15}$
Espritz NMR	3.3379	3.34E+00	-5.1653	2.56E-14

B.) Szolubilis és prefolding aggregálódó fehérjék

Többváltozós logisztikus regressziós modell:

szolubilitás ~ Fehérjerendezetlenség + fehérje felszíni ragadósság

A fehérjerendezetlenség parciális hatása az oldhatóságra

A fehérje felszíni ragadósság parciális hatása az oldhatóságra

Rendezetlenségprediktor	Koefficiens (fehérjerendezetlenség)	P-érték (fehérjerendezetlenség)	Koefficiens (fehérje felszíni ragadósság)	P-érték (fehérje felszíni ragadósság)
DisEMBL hotloops	2.5767	6.79E-08	-6.5908	$< 10^{-15}$
PONDR VSL2B	2.7057	1.28E-06	-6.2345	7.11E-15
iupredS	4.25145	1.88E-05	-6.82021	$< 10^{-15}$
iupredL	1.5252	3.76E-02	-6.7177	1.02E-10
Espritz X-ray	4.443	8.88E-06	-6.78727	$< 10^{-15}$
Espritz NMR	2.1669	7.39E-03	-6.5036	2.56E-15