# Fehérjék térszerkezetének és evolúciójának vizsgálata a mozgékonyság tükrében

- Doktori (Ph.D.) értekezés -


Ángyán Annamária Franciska

Kémia Doktori Iskola
*vezetője:* Dr. Inzelt György, D.Sc.
Szintetikus kémia, Anyagtudomány, Biomolekuláris Kémia Doktori Program
*vezetője:* Dr. Perczel András, D.Sc.

Témavezető: dr. Gáspári Zoltán, Ph.D.

Eötvös Loránd Tudományegyetem
Kémiai Intézet, Szerves Kémia Tanszék
Szerkezeti Kémia és Biológia Laboratórium

Budapest
- 2012 -

# Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, dr. Gáspári Zoltánnak, hogy munkám során szakmailag és emberileg is töretlenül támogatott. Hálás vagyok tudatos és szigorú irányításáért.

Köszönettel tartozom Dr. Pongor Sándornak a munkámhoz fűzött hasznos észrevételeiért és tanácsaiért, valamint a PRIDE-NMR webszerver létrehozásában nyújtott segítségéért. Köszönöm Szappanos Balázsnak a CoNSEnsX webszerver létrehozásában és a programozásban nyújtott segítségét.

Köszönöm Dr. Perczel Andrásnak, hogy tagja lehettem a Szerkezeti Kémia és Biológia Laboratóriumnak és hogy észrevételeivel és támogatásával segítette a munkámat. Köszönettel tartozom a Szerkezeti Kémia és Biológia Laboratórium minden tagjának. Köszönöm dr. Láng András, Stráner Pál, dr. Farkas Viktor, dr. Pohl Gábor, dr. Bodor Andrea és Rovó Petra segítségét a kísérleti munkával történt próbálkozások során nyújtott segítségükért.

Köszönettel tartozom a Szerves Kémia Tanszék régi és jelenlegi taszékvezetőinek, Dr. Perczel Andrásnak és Dr. Hudecz Ferencnek, hogy az ELTE Szerves Kémia Tanszékén végezhettem a kutatómunkámat.

Szüleim, Testvéreim (Tusi, Zsóka, Bori, Miki, Marci, Zoli, Zita, Ilona) és párom Bálint különleges támogatása nélkül ez a dolgozat nem született volna meg. Köszönöm Nagymamámnak a lelkes lektorálást. Hálával tartozom Édesapámnak, hogy megszerettette velem a kémiát.

# Az értekezés alapjául szolgáló közlemények

I. **Ángyán AF**, Perczel A, Pongor S, Gáspári Z:
Fast protein fold estimation from NMR-derived distance restraints.
*Bioinformatics* 24:272-275 **(2008)**
*impakt faktor: 4.328*

II. Gáspári Z, **Ángyán AF**, Dhir S, Franklin D, Perczel A, Pintar A, Pongor S:
Probing dynamic protein ensembles with atomic proximity measures.
*Current Protein & Peptide Science* 11:515-522 **(2010)**
*impact faktor: 3.830*

III. **Ángyán AF**, Szappanos B, Perczel A, Gáspári Z:
CoNSEnsX: an ensemble view of protein structures and NMR-derived experimental data.
*BMC Structural Biology* 10:39 **(2010)**
*impakt faktor: 2.258*

IV. **Ángyán AF**, Perczel A, Gáspári Z:
Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: is aggregation the main bottleneck?
*FEBS Letters*, feltételesen elfogadva (minor revision)

# Tartalomjegyzék

**Köszönetnyilvánítás** ii

**Az értekezés alapjául szolgáló közlemények** iii

**Tartalomjegyzék** iii

**1. Bevezetés** 1

**I. Az NMR kísérleti adatok és a szerkezeti konformer-sokaságok jellemzése** 4

**2. Irodalmi áttekintés** 5
2.1. A polipeptidlánctól a térszerkezetig . . . . . 5
2.2. Fehérjeszerkezetek összehasonlítása távolság jellegű eloszlás alapján . . . . . 6
2.3. Fehérjék szerkezetének meghatározása NMR spektroszkópiával . . . . . 7
2.4. Fehérjék dinamikájának meghatározása . . . . . 9
2.5. A szerkezetek minőségellenőrzése . . . . . 13

**3. Célkitűzések** 14

**4. Az alkalmazott módszerek** 15
4.1. A használt adatbázisok . . . . . 15
4.2. A fehérjeszerkezet-összehasonlító módszerek értékelése . . . . . 16
4.3. A statisztikai elemzés módszerei . . . . . 17
4.4. A használt programok és alkalmazások . . . . . 17

**5. Eredmények** **19**
5.1. Fehérjeszerkezetek becslése NMR adatokból . . . . . 19
5.1.1. A PRIDE-NMR algoritmus fejlesztése . . . . . 19
5.1.2. A háttéradatbázis felépítése . . . . . 21
5.1.3. Súlyozás bevezetése . . . . . 22
5.1.4. Tesztelés kiemelt adatkészleten . . . . . 23
5.1.5. A távolságeloszlások elemzése . . . . . 24
5.1.6. Statisztika a PDB adatbázisbeli NMR-szerkezetekre . . . . . 25
5.1.7. A PRIDE-NMR módszer elérhetősége . . . . . 27
5.1.8. Példa a PRIDE-NMR szerver használatára: az SH3 domén . . . . 28
5.1.9. A módszer korlátai . . . . . 30
5.1.10. Lehetséges alkalmazások . . . . . 31
5.2. Kapcsolat a szerkezeti sokaság és a kísérleti adatok között . . . . . 33
5.2.1. A PRIDE-NMR módszer alkalmazása szerkezeti sokaságon . . . 33
5.2.2. A CoNSEnsX megközelítés . . . . . 35
5.2.3. A CoNSEnsX szerver felépítése . . . . . 36
5.2.4. Példa a CoNSEnsX szerver használatára: humán ubiquitin, mint globuláris fehérje . . . . . 37
5.2.5. Példa a CoNSEnsX szerver használatára: PDE 5/6 $\gamma$-alegység, mint rendezetlen fehérje . . . . . 43
5.2.6. A CoNSEnsX szerver alkalmazásai . . . . . 45

**6. Diszkusszió** **46**

**II. Véletlenszerű fehérjeszekvenciák *in silico* szerkezetvizsgálata** **49**

**7. Irodalmi áttekintés** **50**
7.1. A biológiai információáramlás . . . . . 50
7.2. Belsőleg rendezetlen fehérjék . . . . . 52
7.3. Transzmembrán fehérjék . . . . . 53
7.4. Amiloid fehérjék és a fehérjeaggregáció . . . . . 53
7.5. Teljesen új fehérjék keletkezése . . . . . 54

**8. Célkitűzések** **56**

**9. Az alkalmazott módszerek** **57**
9.1. Használt adatbázisok és adatkészletek . . . . . 57
9.2. Random fehérjeszekvenciák generálása . . . . . 58
9.2.1. Teljesen random mRNS szekvenciák . . . . . 58
9.2.2. Proteomok randomizálása . . . . . 59
9.3. Homológiaszűrés, fiziko-kémiai elemzés . . . . . 60
9.4. Szekvencia alapú predikciós módszerek elméleti háttere és alkalmazása . . . . . 61
9.4.1. Rendezetlenség prediktorok . . . . . 62
9.4.2. Transzmembrán prediktorok . . . . . 62
9.4.3. Aggregáció és amiloid prediktorok . . . . . 63
9.5. A statisztikai elemzés módszerei . . . . . 64

**10. Eredmények** **66**
10.1. Véletlenszerű szekvenciák álatlános jellemzése . . . . . 66
10.2. A szekvencia-alapú predikciókból kiolvasott általános trendek . . . . . 70
10.3. A strukturális tulajdonságok egymás közötti összefüggései . . . . . 77
10.4. Statisztikai elemzés . . . . . 81
10.5. Proteomok és randomizált proteomok összevetése . . . . . 84
10.6. *De novo* fehérjék a humán genomban . . . . . 86

**11. Diszkusszió** **88**

**12. Összefoglalás** **90**

**Rövidítések jegyzéke** **92**

**Ábrák jegyzéke** **94**

**Táblázatok jegyzéke** **96**

**Kivonat** **97**

**Abstract** **98**

**Irodalomjegyzék** **99**

# 1. fejezet

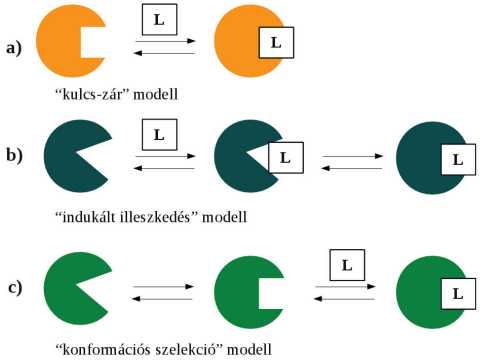
# Bevezetés

A molekuláris biológia területén az elmúlt években több szemléletváltásnak lehettünk tanúi. Egyrészt, a fehérjemolekulák belső, inherens dinamikus mivolta előtérbe kerül, másrészt a fehérjeszerkezet alatt már nem csupán globuláris fehérjéket értjük, hanem a polipeptidlánc további, strukturális és fizikokémiai jellemzői alapján elkülöníthető konformációit; ennek főleg az ún. rendezetlen fehérjék leírásában van szerepe.

A fehérje-ligandum kölcsönhatás leírásakor a kölcsönható partnerek mozgékonyságának figyelembevétele szemléletváltáshoz vezet. Az 1.1 ábra (Vértessy & Orosz után [1]) összefoglal három fehérje-ligand kötődéselméletet. A kulcs-zár modell (1.1 A) ábra) mindkét partnert merev egységként értelmezi. A kölcsönhatás akkor jön létre, ha a ligand (kulcs) konformációja jól illeszkedik a fehérje kötőzsebébe (zár). Az indukált illeszkedés elmélet (1.1 B) ábra) szerint a ligand térbeli közelsége konformációs változást indukál a fehérje kötőzsebén, így lehetővé téve a tényleges kötődést. A fluktuációs illeszkedés elméletben (1.1 C) ábra) a fehérjeszerkezeteket konformációs sokaságként értelmezzük, azaz az egyes fehérjemolekulák más és más konformációs állapotban lehetnek. A ligandum a kötődéssel ezt a konformációs egyensúlyt tolja el, mintegy kiválasztva a jól illeszkedő kötőzsebet "felmutató" molekulákat [1].

Fontos további aspektus, hogy nem csak a térszerkezet, hanem a dinamika is megváltozik a kötődés során, aminek a kötődési szabadentalpia entropikus járulékának megértésében van jelentősége. Az ún. rendezetlen fehérjék esetében ez látványos, például a kötés közbeni feltekeredés (folding and binding) mechanizmus során, de globuláris fehérjéknél is fellép ez a jelenség. A konformációs szelekció fontosságát számos publikáció támasztja alá[2, 3]. Ennek ellenére kevés tanulmány van, ahol ezt atomi szinten is bemutatnák a kutatók [4, 5].

A fehérjék térszerkezetének és dinamikájának atomi szintű jellemzésére az NMR spek-troszkópia a legalkalmasabb módszer. Nagy előnye, hogy a mérés a fiziológiáshoz

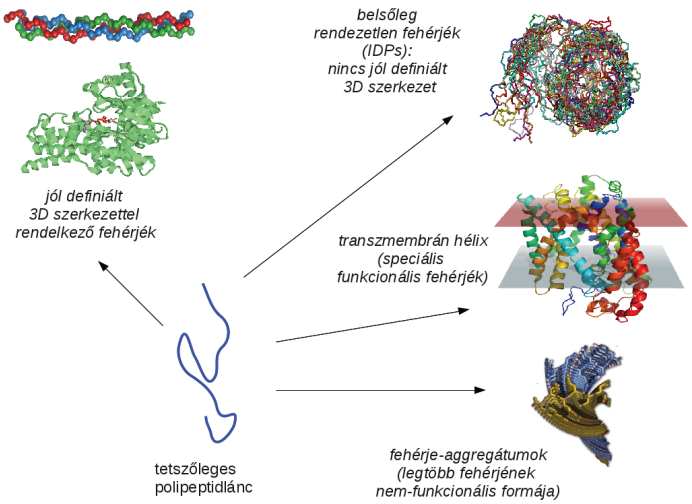



1.1. ábra. A fehérje-ligand kötődési mechanizmus modelljei (Vértessy & Orosz után)

némileg hasonló körülmények között, oldatban történik. A NMR minta oldata nagyjából $10^{16}$-$10^{17}$ fehérjemolekulát tartalmaz. Az oldatbeli konformerek mozgékonyságát és szerkezeti heterogenitását egyetlen konformerrel nem érdemes jellemezni, ugyanis csak térben és időben átlagoltan mérhető minden kísérleti adat. Léteznek protokollok, amelyek különböző időskálákon a fehérjék belső dinamikáját figyelembe veszik [4, 6]. Ilyen típusú szerkezetszámolás eredménye egy ún. dinamikus szerkezeti sokaság, amely a vizsgált makromolekula belső dinamikáját kísérleti paraméterekből származtatva írja le.

A fehérjeszerkezetek vizsgálódási köre kiszélesedni látszik: a statikus, egyetlen konformeres modellekkel szemben a dinamikus, egyre szélesebb időskálát leíró fehérjemodellek veszik át a terepet, és ez a biológiai folyamatok atomi szinten történő mélyebb megértéséhez elengedhetetlen. A kísérleti adatok értékelése viszont még számos kihívást rejt magában.

A dinamika megváltozhat a kötődés, mutáció vagy akár rendezett-rendeztelen átmenet hatására is [7]. A globuláris szerkezettel szemben a rendezetlen fehérjék funkcionális előnnyel bírnak, a számos kórkép kapcsán leírt fehérjeaggregátumok pedig termodinamikai értelemben stabilabbak. Evolúciós skálán döntő szerepet játszik fehérjék konformációs változatossága, dinamikája, ami kihat a biológiai rendszerekben betöltött funkciókra [8]. Ha a fehérjék aggregációs hajlama eleve inherensen magas és nem számolunk védekezési mechanizmusokkal, akkor az új fehérjék keletkezése gyakorlatilag lehetetlenné és elhanyagolható jelenséggé válna (1.2 ábra, [9]).




1.2. ábra. Ismert fehérje szerkezeti struktúrák

Az elmúlt években viszont egyre több kísérleti bizonyíték van arra, hogy fehérjék ténylegesen keletkeznek *de novo*, azaz spontán átíródás révén is [10]. Ez a kísérleti tény új megvilágításba helyezi a fehérjék evolúcióját, mivel a nemkódoló DNS szakaszok spontán átíródása új fehérje keletkezése mellett sokkal gyakoribb jelenség lehet, mint eredetileg gondolták a kutatók. Felmerül a kérdés, hogy evolúciós skálán miképpen jöttek létre a fehérjék, hogy milyen térszerkezetet vettek fel az újonnan kialakuló polipeptidláncok, valóban jelentős-e az aggregációra való hajlamuk, és hogy fehérjék mozgékonysága és dinamikája milyen szerepet játszott ebben a folyamatban.

Dolgozatomban a fehérjék dinamikáját és evolvabilitását járom körbe a mozgékonyság tükrében. A dolgozat első részében a dinamikus szerkezeti sokaságokkal kapcsolatos fejlesztéseimet mutatom be. Ez a rész az I., II. és III. saját közleményekre támaszkodik. A dolgozat második fele evolúciós távlatokban vizsgálja a *de novo* fehérjék dinamikus tulajdonságait a rendezetlenségen keresztül. Arra a kérdésre kerestem választ, hogy az aggregációs hajlam befolyásolja-e az újonnan keletkező fehérjék kialakulását és stabilitását. Ez a rész a IV. saját közleményre támaszkodik.

# I. rész

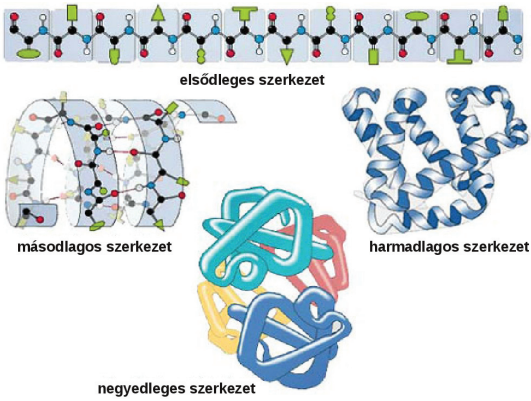
# Az NMR kísérleti adatok és a szerkezeti konformer-sokaságok jellemzése

# 2. fejezet

# Irodalmi áttekintés

## 2.1. A polipeptidlánctól a térszerkezetig

A fehérjék poliaminosav-láncok, amelyek speciális strukturális tulajdonságaiknak köszönhetően biológiai funkciókat töltenek be. Az aminosavak különböző oldalláncai adják meg a szükséges kémiai változatosságot a biológiai funkciók betöltéséhez. A fehérjék szerkezetének négy szintjét szokás megkülönböztetni (2.1 ábra).




2.1. ábra. A fehérjék szerkezeti szintjei

Az elsődleges szerkezet a fehérjét alkotó aminosavak sorrendje. Másodlagos szerkezetnek a fehérjelánc egyes összefüggő szakaszainak térbeli elrendeződéseit hívjuk: $\alpha$-hélix, $\beta$-redő, $\beta$-, $\gamma$-kanyarok, stb. A harmadlagos szerkezet a másodlagos szerkezeti

elemek elrendeződése a térben, azaz a teljes fehérjelánc feltekeredési módja[1]. A fehérjék negyedleges szerkezetét azon fehérjeláncok együttese adja, amelyek nem kovalensen kapcsolódnak egymáshoz [11].

## 2.2. Fehérjeszerkezetek összehasonlítása távolság jellegű eloszlás alapján

Az egyik legújabb, igen gyors és hatékony módszer, a PRIDE (Probability of IDEntity) analízis [12] alkalmas arra, hogy gyorsan találjunk teljes vagy részleges szerkezeti hasonlóságokat a fehérjeszerkezetek között. Az eljárás az aminosavak C$\alpha$-atomjainak távolságeloszlásának összehasonlításán alapul. A vizsgálandó fehérjék szerkezeteiből 28 hisztogramot állítunk elő oly módon, hogy minden szekvenciális távolságra (n), ahol $3 \leq n \leq 30$, a fehérjében mérhető összes C$\alpha_i$-C$\alpha_{i+n}$ távolságérték a hisztogram adott tartományába kerüljön, a távolság nagysága alapján. A hisztogram felbontását változtathatjuk, a tartományok lehetnek például 1 Ångstrømönként, és meghatározhatunk minimális (pl. 3Å) és maximális távolságértéket, ami alatt és fölött a C$\alpha$-C$\alpha$ távolságokat nem vesszük figyelembe. A hasonlóság megállapítása során a két összehasonlítandó fehérjére kapott 28-28 hisztogramot páronként, kontingencia-analízissel [13] vetjük össze, és a kapott 28 valószínűség átlagát véve kapjuk a PRIDE mérőszámot (PRIDE score).

A számolás részletei egy adott n értékre a következők (pl. n=3) [12]:
A két összehasonlítandó hisztogram legyen obs(1,1), obs(2,1), ..., obs(m,1) az első és obs(1,2), obs(2,2), ..., obs(m,2) a második fehérje esetében, ahol m a tartományok száma a hisztogramokon belül. Minden észlelt adathoz várható értéket számolunk az alábbi módon:

$$exp(i,j) = \frac{\text{obs(i,x)} * \text{obs(x,j)}}{\text{obs(x,x)}} \tag{2.1}$$

ahol obs(i,x) a két hisztogram j-edik tartományában lévő értékek összege, obs(x,j) az i-edik térszerkezetre kapott összes adat száma a hisztogramra, obs(x,x) pedig a két hisztogram összes adatának száma. Ezután kiszámítjuk a megfelelő $\chi^2$ értéket:

$$\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{m} \frac{[obs(i,j) - exp(i,j)]^2}{exp(i,j)} \tag{2.2}$$

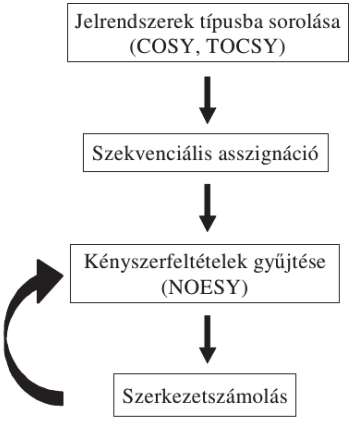
A $\chi^2$ érték a szabadsági fokok ismeretében (m-1) átszámolható valószínűséggé, feltéve hogy a hisztogramok egyik tartományába sem esik az adatok 5%-ánál kevesebb távolságérték. Ennek biztosítása érdekében a számolás során a hisztogramok ilyen tartományait az előző tartománnyal összevonjuk mindaddig, amíg a feltétel igazzá nem válik.

[1] Más néven a fehérje "fold"-ja.

A PRIDE mérőszám nulla és egy közötti értéket vehet fel. Ha két eloszlás összehasonlításának eredménye nulla, akkor az eloszlásokhoz tartozó térszerkezetek teljesen eltérnek egymástól. Ha egy körüli értéket kapunk, valószínűsíthető, hogy van hasonlóság. A gyakorlatban $\simeq 0{,}85$ feletti érték megegyező térszerkezetet jelent. Két teljesen azonos térszerkezetre a mérőszám definíció szerint egy. Az eljárás eredményének kiértékelésénél igen fontos még a találatok sorrendje. Értéktől függetlenül az első találatot mindig érdemes megvizsgálni, mint a keresett szerkezet lehetséges legközelebbi szomszédját (rokonát): a PRIDE ún. "nearest neighbour classifier" eljárásnak[1] tekinthető.

## 2.3. Fehérjék szerkezetének meghatározása NMR spektroszkópiával

A fehérjék atomi szintű szerkezet-meghatározásában a röntgenkrisztallográfia mellett az NMR spektroszkópia egyre nagyobb szerepet kap. Az NMR-spektroszkópiás szerkezetszámolás eredményessége nagyban függ a kísérleti adatok minőségétől, mennyiségétől és a szekvenciális asszignáció pontosságától [14, 15].




2.2. ábra. Az NMR spektroszkópia alapú térszerkezet-meghatározás folyamatábrája

[1] azaz a legközelebbi találaton alapuló eljárás

A konstitúciós viszonyokról információt adó spektrumoktól eltérően (COSY[1], TOCSY[2]), a dipól-dipól csatoláson alapuló NOESY[3] mérés az egyes atomok relatív térbeli helyzetére érzékeny. A csúcsok alatti terület (térfogati integrál) nagysága fordítottan arányos a megfelelő protonok közötti távolság hatodik hatványával [16].

$$\eta_{ij} \approx r_{ij}^{-6} * f(\tau_c) \tag{2.3}$$

ahol $\eta_{ij}$ a NOE (Nuclear Overhauser Effect) érték (térfogati integrál), $r_{ij}$ a két proton ($H_i$ és $H_j$) távolsága (Å), és $f(t_c)$ a molekula - annak méretével arányos és oldatbeli forgását jellemző - korrelációs idő függvénye [17]. Segítségével az egymástól legfeljebb 6 Å távolságra lévő atomok azonosíthatók. Megkülönböztetünk homonukleáris ($^1$H-$^1$H) és heteronukleáris NOE értékeket (például $^1$H-$^1$5N), utóbbiaknak a dinamikai mérésekben van szerepe. Amennyiben rendelkezésre áll az NMR spektrumok jelhozzárendelése, azaz ismerjük a fehérje protonjainak kémiai eltolódását, a NOESY spektrum alapján lehetőség nyílik az atomi szintű térszerkezet kiszámítására. A NOESY spektrumban minden csúcs a megfelelő protonok térbeli közelségére utal, távolság jellegű kényszerfeltétel rendelhető hozzá. A távolság jellegű kényszerfeltételek között megkülönböztetünk szekvenciális (szomszédos aminosavak közötti), közeli ($|i-j| \leq 4$, ahol i és j a két aminosav szekvenciában elfoglalt helye) és távoli ($|i-j| > 4$) típusúakat. A közeli kényszerfeltételek a lokális struktúra, a távoliak a fehérjelánc harmadlagos szerkezetére jellemzőek. Amennyiben a fehérjében elegendően sok távoli aminosavak közötti atom-atom ($^1$H-$^1$H) távolságot ismerünk, kiszámíthatjuk a polipeptidlánc konformációját [16].

A térszerkezet-meghatározásra a legelterjedtebb módszer a kényszerfeltételek felhasználásával futtatott molekuladinamikai szimuláció. Ennek utolsó lépése visszacsatolásos eljárás (2.2 ábra; [18]): a kiszámított szerkezetek alapján a spektrumokat átértékelve módosítjuk a kényszerfeltétel-listát és újra kiszámítjuk a szerkezeteket. Ez a gyakorlatban a programok többszöri futtatását és a spektrumokban lévő információ sokszori újraértékelését magába foglaló folyamat. A szerkezetfinomítást akkor tekintjük befejezettnek, ha a kényszerfeltételeknek megfelelő szerkezetek adott határokon belül[4] egymásra jól illeszthetők, azaz egyetlen fő konformert reprezentálnak[5] és jó minőségűek, azaz megfelelnek a fehérjeszerkezetek minőségével kapcsolatos követelményeknek [19]. A szerkezetszámolás legtöbbször kifejezetten erre a célra írt programokkal (pl. X-PLOR [20], CNS [21]) történik.

[1] COrrelated SpectroscopY

[2] TOtal Correlated SpectroscopY

[3] Nuclear Overhauser Effect SpectroscopY

[4] gerinc RMSD (Root Mean Square Deviation); tipikusan 0,3 - 0,5 Å, egészen 1 Å-ig

[5] A megfogalmazás szándékosan elnagyolt. Nincs minden szerző által elfogadott kritérium, de szempontként általánosan használatos.

A szerkezetmeghatározás során olyan szerkezetet próbálunk előállítani, ahol a detektált kényszerfeltételek a lehető legjobban teljesülnek. A leggyakoribb kényszerfeltételként a NOE értékeket használják fel, de ezen kívül még más adatok is alkalmazhatók, például csatolási állandókból számolt torziós szög jellegű kényszerfeltételek. Ennek megvalósítása például a NOE-k esetében, hogy ha a két atom közötti aktuális távolság a kényszerfeltétel által meghatározott tartományon kívül esik, akkor egy járulékos energiatag adódik a molekula energiafüggvényéhez, így „terelődik" a szerkezet a kényszerfeltételek által megadott irányba. A gyakorlatban nem egyetlen szerkezetet kapunk, hanem egy sokaságot (ensemble), mely elvben jellemző a molekula konformációs heterogenitására, mivel a kényszerfeltételek több, egymáshoz hasonló szerkezetre is teljesülhetnek[1]. Ezek közül kell kiválogatni azokat, amelyek a kényszerfeltételeknek a legjobban megfelelnek. A hagyományos molekuladinamikai szerkezetszámolás során egyszerre mindig egyetlen konformert optimálunk. Ez az eljárás hosszadalmas, főleg ha semmilyen előzetes elképzelés nincs a fehérje lehetséges térszerkezetéről. A kényszerfeltételek már önmagukban az összes információt tartalmazzák a térszerkezet kiszámításához, a szerkezet visszaszámolása sokszor mégis nehéz feladat marad.

## 2.4. Fehérjék dinamikájának meghatározása

Az NMR mérésekből származó klasszikusan használt távolság jellegű kényszerfeltételek bővíthetőek több paraméterrel (2.3 ábra; [22]).

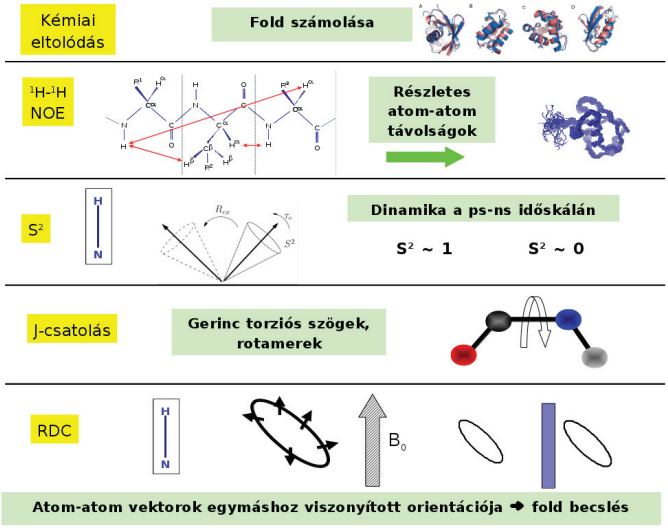
Az általános rendparaméter ($S^2$) a heteronukleáris[2] relaxációs mérésekből származtatott paraméter. Minden aminosavra megadható, és pikoszekundumos - nanoszekundumos időskálán jellemzi az aminosav adott atomjainak mozgékonyságát. Értéke nulla és egy közé esik. Minél flexibilisebb az aminosav, annál alacsonyabb az érték. A fehérje merev részeire a 0,7 és 1 közötti, a szubsztrátkötő zsebre 0,5-0,7, míg a terminális szakaszokra ennél alacsonyabb $S^2$ jellemző. Ha az amid NH kötés vektorának mozgását egy küppalást körüli forgással írjuk le, akkor az $S^2$ a palást szélességét adja meg – ha $S^2$ kicsi, akkor tág a palást [23] (2.3 ábra).

A reziduális dipoláris csatolásokat (RDC-ket) ún. orientált közegben mérjük és megfelelő atom-atom vektorok a mágneses térhez, illetve egymáshoz viszonyított helyzetéről ad információt [24]. Az RDC-k időskálája a mikroszekundumig terjed.

A kémiai eltolódások valamint az RDC-k vizsgálatával a másodlagos és harmadlagos szerkezeti elemekről kaphatunk pontosabb információt. Kizárólag kémiai eltolódások

[1] A gyakorlatban rutinszerűen használt eljárások ezt nem garantálják.

[2] Tipikusan $^1$H-$^1$5N, de esetenként $^1$H-$^1$3C mérésekből is származtatható.




2.3. ábra. Az NMR adatokból kinyerhető általános paraméterek és jellemzőik

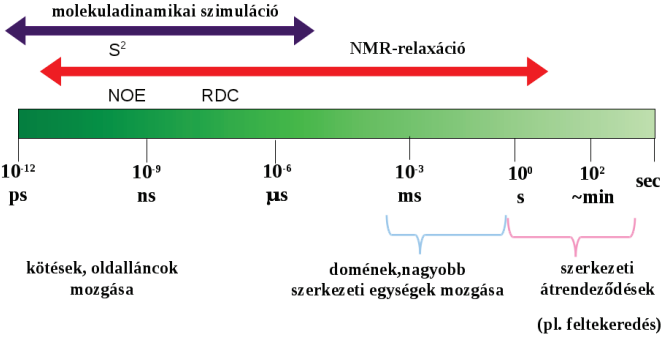
alapján [25, 26] vagy kizárólag RDC-k alapján [4, 27] is lehetséges a szerkezetszámolás.

Jelenleg több módszer is képes a szerkezeti sokaságra vonatkoztatni az NMR adatokból származó kényszerfeltételeket a szerkezetfinomítás során, mint a NOE kényszerfeltételek [28, 29, 30], $S^2$ paraméterek [31] és RDC [4, 32] értékek. A ma általánosnak tekinthető gyakorlatban azonban a NOE alapú távolság jellegű kényszerfeltételek túlsúlya figyelhető meg, különösen a megbízható, nem csak közelítő szerkezetmeghatározásra törekvő kutatások esetében.

Nagy előnye az NMR-spektroszkópiával történő szerkezetmeghatározásnak, hogy oldatfázisban vizsgálható a molekula és nem csak statikus térszerkezeti adatok nyerhetők, hanem a fehérjemolekula belső, dinamikus tulajdonságai is jellemezhetők különböző időskálákon (2.4 ábra; [33]).

A szokásos molekuladinamikai protokollal számolt szerkezet(család) nem feltétlenül tükrözi a molekula konformációs heterogenitását. Ennek oka, hogy az NMR mérés során a fehérjék oldatban vannak és az egyes molekulák eltérő konformációs állapotban lehetnek. A mért paraméterek viszont nem egyes molekulákra érvényesek, hanem az egyes molekuláknak megfelelő értékek átlagaként jönnek létre. Ezért ha a szerkezeti fluktuációk több konformert eredményeznek, akkor a modell nem fedi le az összes, statisztikai-

lag különböző konformációt. Ezt alulillesztésnek vagy túlzott megkötésnek (underfitting, over-restraining) nevezzük. Csökkenthetjük a jelenséget, ha növeljük a modellbe tartozó szerkezetek számát, ez azonban újabb problémához vezet. Megnöveli a rendszer szabadsági fokát, míg a kísérleti adatok száma nem növekszik. Ekkor a kísérleti és a szimulációból visszaszámolt adatok jobb illeszkedése nem a valódi konformerekhez való hasonlóság, hanem a szabad paraméterek megnövekedett száma miatt van. Ez a túlillesztés (overfitting, under-restraining). A túlillesztést csökkenteni lehet további kísérletes adatok, például az $S^2$ értékek kényszerfeltételként való alkalmazásával.

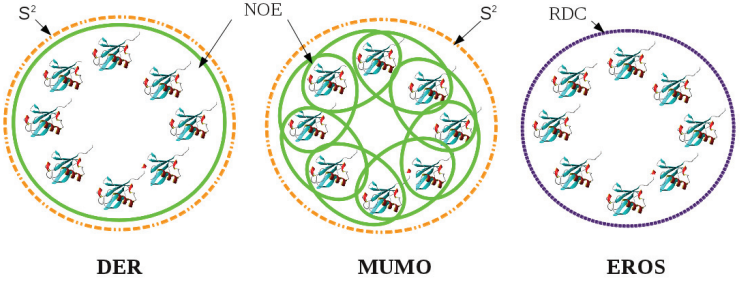



2.4. ábra. Az NMR spektroszkópiai mérések időskálája és néhány, az adott időtartományhoz kapcsolható molekuláris mozgás

A hagyományos, erőtér alapú molekuladinamikai NMR szerkezetek számolása során a szerkezetek nincsenek külön megfeleltetve a dinamikai paramétereknek, mint az $S^2$-ek, RDC-k, kémiai eltolódások, J-csatolás, stb. Ezt a hiányosságot lehet kiküszöbölni a közelmúltban kifejlesztett néhány protokoll alkalmazásával. A 2.5 ábra három különböző számolási protokollt mutat be.

A DER (Dynamic Ensemble Refinement, [28]) szerkezetfinomító protokoll a NOE és a fehérje dinamikájára jellemző általános rendparamétereket ($S^2$ értékekt) együttesen kezeli és így a konformerek sokaságának reálisabb, a dinamikát is tükröző reprezentálását eredményezi. Az $S^2$ általános rendparaméter mellett egyéb paraméter is integrálható a protokollba, mint a kémiai eltolódások, RDC-k, vagy akár röntgendiffrakciós adatok. Az eljárás célja, hogy minél több kísérleti adatot használjunk fel mint kényszerfeltételt a molekuladinamikai szimuláció során oly módon, hogy egyszerre nem egy, hanem több (tipikusan 8) konformert vagy replikát optimálunk (2.5 "DER" ábra). A MUMO (Minimal Under-restraining Minimal Over-Restraining, [29]) eljárás a DER továbbfejlesztése.

A NOE adatokat az előzőtől eltérően nem a teljes sokaságra[1], hanem replika párokra optimálja (2.5 "MUMO" ábra). Az EROS (Ensemble Refinement with Orientational reStraints, [4]) kizárólag RDC adatokat használ fel a szerkezetszámoláshoz. Az előző két protokollhoz hasonlóan, ebben az esetben is az RDC adatokat egyszerre több konformerre optimáljuk (2.5 "EROS" ábra).




2.5. ábra. A fehérje belső dinamikáját tükröző NMR adatokat figyelembevevő néhány szerkezetszámolási protokoll

A dinamikus sokaság tehát olyan molekulakonformer együtteseket takar, amelyek a mért kísérleti adatoknak megfelelnek. A dinamikus szerkezeti sokaságok esetében csakis a teljes sokaság egyeztethető össze a kísérleti paramétereknek, és habár az egyes konformerek geometriailag reálisak, külön-külön nem írják le az oldatbeli molekula összes lehetséges állapotát. Ez a szerkezetszámolás annyiban tér el a hagyományos szerkezetszámolástól, hogy a molekuladinamikai szimuláció során a dinamikai adatok explicite vannak figyelembe véve. Protokolltól függően egy vagy több NMR kísérleti adatot (tipikusan NOE, $S^2$ rendparaméter, reziduális dipoláris csatolások (RDC-k), J-csatolások, kémiai eltolódás értékek; ld 2.3 ábra) használnak fel a szokásos geometriai kényszerfeltételek mellé, és egyszerre több konformerre optimálunk. Ily módon a kapott konformersokaság *együttesen* teljesíti a kezdeti kényszerfeltételeket. Az egyes konformerek, habár geometriai szempontból megfelelnek az elvárt értékeknek, külön-külön a dinamikai paramétereknek csak egy részét teljesítik. Ennek ellenére a kapott szerkezeti sokaság tükrözi a molekula belső dinamikáját, ellentétben a hagyományos szerkezetszámolás során kapott sokasággal, ahol a "mozgékonyság" a nem teljesülő kényszerfeltételekből ered és a szerkezetek bizonytalanságát jelzi.

[1] az egyszerre optimált konformerek száma

## 2.5. A szerkezetek minőségellenőrzése

Minden szerkezet amivel a polipeptid lánc konformációs változatosságát igyekszünk leírni egy-egy modell, ami valamilyen szinten összeegyeztethető a rendelkezésre álló kísérleti adatokkal. Minden modellnél kritikus a precizitás és a pontosság kérdése (2.6 ábra).


2.6. ábra. Precizitás és pontosság modellek esetében

A röntgenkrisztallográfiával ellentétben az NMR szerkezetek minőségellenőrzésére nincs átlalánosan elfogadott mérőszám. A szerkezetszámolás geometriai eljárás mely során a kötésszögek, kötéshosszak optimalizálása a döntő motívum. Az egyes modellek illesztésének jósága (RMSD) röntgenszerkezetek esetén releváns jósági információt tartalmaz, de a dinamikus szerkezeti sokaságok esetén nem feltétlenül releváns.

Ma már létezik néhány olyan módszer, amely az NMR-szerkezetekhez a krisztallográfiai R-faktorhoz hasonló mérőszámot rendel. Például a mért és számolt szerkezet alapján számolt reziduális dipoláris csatolások összevetésével, vagy a NOE-k alapján történő elemzéssel (RFAC [34]). Az említett módszerek azonban egyáltalán nem rutinszerűen alkalmazott eljárások. A PROCHECK-NMR vagy AQUA programok [35] több adatot kiszámolnak minőségellenőrzés végett, mint a Ramachandran felszínt, a nem teljesülő NOE-k, RMSD értékeket a különböző módon illesztett konformerekre. Átfogó, a különböző dinamikai paramétert is egyszerre ellenőrző alkalmazás viszont nincs. A dinamikus szerkezeti sokaságok esetén a teljes sokaságra kell ellenőrizni a megfelelést a kísérleti adatokkal, nem a konformerekre külön-külön, ugyanis a szerkezetszámolás során együtt optimáltuk a konformereket az adott paraméterre nézve. Az elérhető minőségellenőrző programok és szerverek konformer-sokaságot nem tudnak együttesen kezelni, csak az egyes konformereket külön-külön. A paramétereket viszont ebben az esetben csak átlagolva lehet megfeleltetni.

# 3. fejezet

# Célkitűzések

Az első részben a fehérje térszerkezete és dinamikája kapcsolatát jártam körbe. Egyik célom egy NMR-orientált fehérje térszerkezet-becslő algoritmus fejlesztése és megvalósítása volt. Másrészt az NMR adatokból származtatott, a fehérje dinamikáját jellemző paraméterek és a fehérje térszerkezetének megfeleltetését és jellemzését elősegítő protokoll megvalósítása volt. E célokat az alábbi lépésekben kívántam elérni:

1. Egy algoritmus fejlesztése, amely gyorsan és egyértelműen kapcsolatot teremt a már ismert térszerkezetű fehérjék és a kísérleti NOE-k között.

2. A módszer teljesítőképességének többboldalú tesztelése.

3. Egy minőségellenőrző protokoll kifejlesztése, amely segítségével a szerkezeti sokaság megfeleltethető a kísérleti NMR-adatoknak.

# 4. fejezet

# Az alkalmazott módszerek

## 4.1. A használt adatbázisok

Az RCSB PDB [36] egy rendszerezett és egységes fehérje és DNS szerkezeti adatbázis, amely a makromolekulák szerkezeti információit tartalmazza. A PDB adatbázisban megtalálható szerkezetek száma ma már meghaladja a 80.000-et[1]. Újabban az elavult és elméleti úton konstruált modellek adatait nem tartalmazza ez az adatbázis, kizárólag kísérleti adatokon támaszkodó modellek kerülhetnek be. A PDB adatbázisból a térszerkezeti adatokat valamint az NMR-szerkezetek kísérleti adatait használtam fel.

A RECOORD [37] adatbázis NMR spektroszkópiai adatokra támaszkodó, egységes protokollok szerint újraszámolt szerkezeteket tartalmazza. A RECOORD standard adatbázisnak tekinthető, többek között fehérje NMR-spektroszkópiai adatokat felhasználó módszerek fejlesztésénél. A RECOORD 545, különböző kutatócsoportok által NMR-spektroszkópiával meghatározott fehérje újraszámolt térszerkezetét és kísérleti adatait tartalmazza. A szerzők a fehérjeszerkezeteket egységes eljárásokkal számolták és finomították újra. Így a szerkezetek "minősége" egységes és nem torzítják el a különböző kutatócsoportokban használt szerkezetszámolási eljárások közötti eltérések. Az eredeti mérési adatok megbízhatóságát természetesen ez nem befolyásolja.

A SCOP [38] adatbázis 1995 óta létezik; az utolsó frissítés 2009 júniusában történt (1.75-ös verzió). A SCOP adatbázisban egyes családok túl vannak reprezentálva, ezért a teljes adatbázis mellett megtalálható a 95%-os valamint a 40%-os szekvencia-szűrt adatkészlet (ASTRAL [39]). A 95%-os szekvencia-szűrt adatkészletben a 95%-nál nagyobb szekvenciális hasonlóságot mutató szerkezetek közül csak egy került be az adatbázisba és így redundanciamentes adatkészlet áll rendelkezésre. A SCOP adatbázisban minden

[1] 2012. április 24.-én 81.048 szerkezetet tartalmazott az RCSB PDB adatbázis.

egyes domén jellemzője egy egész szám és egy rövid, a szerkezet hierarchiában elfoglalt helyére utaló besorolás. A besorolások használatával könyen automatizálni lehet a keresést a SCOP adatbázis különböző szintjein. Munkám során a SCOP ASTRAL teljes és 95%-os szekvencia-szűrt állomány 1.71-es verzióját használtam.

## 4.2. A fehérjeszerkezet-összehasonlító módszerek értékelése

Az általam kifejlesztett egyik eljárás tesztelésére szükség volt egy széles körben alkalmazható, jól dokumentált tesztre. Mivel a módszer koncepcionálisan a fehérjeszerkezet-összehasonlító algoritmusokkal rokon, egy ilyen eljárásokra kifejlesztett tesztet alkalmaztam. Novotny és munkatársai összehasonlítottak tizenegy, a világhálón publikusan elérhető fehérje-összehasonlító szervert [40]. Szempont volt a teljesítmény (a szerkezeti hasonlóságot milyen mértékben találja meg) és a használhatóság (felhasználói felület, leírás, eredmények értékelése).

A teljesítmény értékelésére már meghatározott szerkezetű fehérjék PDB atomi koordinátáit használták, több adatkészletet kialakítva. Referenciaként a CATH térszerkezet-osztályozó adatbázis [41] adataira támaszkodtak. A találatot pozitívnak tekintették, ha a CATH hierarchia harmadik szintjéig ("T", topology) azonos besorolású a kereső és a talált szerkezet. A hasonlóság referenciája itt a CATH térszerkezet-osztályozó adatbázis volt. Az értékelési nehézségek elkerülése végett a szerverek teljesítményének mérésére bevezettek egy egyszerű bináris mérőszámot: egy, ha legalább egy pozitív találat van az első száz kiadott találat között, egyébként nulla[1].

Általános tesztként a CATH osztályokat egyenlően reprezentáló 61 fehérjeszerkezetet tartalmazó adatkészletet használtak. Speciális tesztek a nem triviális szerkezetekre, a többdoménes, az NMR-szerkezetekre (érzékenység mérése), valamint a csak C$\alpha$ modellekre is kitértek. Egyik szerver sem teljesített 100%-osan, a leghatékonyabb a CE [42] (93%), a DALI [43] (90%) és a VAST [44] (86%) bizonyultak. Az említett módszerek mind viszonylag időigényes módszerek, 5-20 perc között van az átlagos keresési idő egy szerkezetre. A PRIDE [12] szerver fejlesztett változata, a PRIDE2 [45], ezen a teszten 84%-os teljesítményt ért el, minden esetben egy perc alatti keresési idővel.

[1] Kivéve, ha a szerver az általa használt belső szignifikancia-vizsgálatok eredményeként ennél kevesebb találatot ad meg.

## 4.3. A statisztikai elemzés módszerei

Az összehasonlításokhoz kontingencia-analízist használtam [13]. Két eloszlás ismeretében, ha az eloszlások analitikai modellje nem ismert, ez a statisztikai eljárás tudja megmondani, hogy vajon azonos vagy egymáshoz közeli populációkból származnak-e. A kontingencia-analízist[1] a PRIDE eljáráshoz (ld. 2.2 fejezet) [12] hasonlóan a távolságeloszlások számszerű összevetésére használtam.

A statisztikai elemzés során minden egyes paraméterre külön-külön korrelációs koefficienst (r) számoltunk:

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{4.1}$$

Továbbá a szerkezeti sokaságokra átlagolt ún. q-faktort ("quality", vagy jósági faktort) számoltunk:

$$\mathrm{q} = \frac{\sqrt{\sum (P_{calc} - P_{exp})^2}}{\sqrt{\sum P_{exp}^2}} \tag{4.2}$$

ahol $P_{calc}$ a sokaságra átlagolt, a szerkezet atomi koordinátái alapján visszaszámolt paraméter, $P_{exp}$ pedig a kísérletileg mért, vagy kísérleti adatok alapján származtatott paraméter:

## 4.4. A használt programok és alkalmazások

A röntgendiffrakcióval meghatározott szerkezetek a hidrogénatomok koordinátáit nem tartalmazzák. Munkám során ezekre az adatokra szükség volt, így a hiányzó hidrogénatomokat utólag illesztettem be a PDB állományokba. A GROMACS molekuladinamikai programcsomag [46] *pdb2gmx* programját használtam az OPLS-AA erőteret [47] alkalmazva. A *pdb2gmx* programot ehhez kismértékben módosítani kellett, hogy az elemzés során nem releváns hibákat (pl. hiányzó oldallánc-atomok) tartalmazó szerkezeteket is megfelelően kezelje.

Az ubiquitin szerkezeti sokaságot a GROMACS molekuladinamikai programcsomaghoz integrált MUMO [29] számolási protokollal számoltuk ki [48], az 1D3Z PDB állományhoz [49] rendelkezésre álló NOE és $S^2$ adatokat felhasználva. A NOE adatokat

[1] más néven becsléses függetlenségvizsgálat $\chi^2$ próbával

tisztítani kellett, csak az egyértelmű kényszerfeltételeket tartottuk meg. A szimuációt 5 nanoszekundumig, 300K-en, explicit víz (SPC; Single Point Charge) modellt használva, 4 femtoszekundumos lépésközökkel, kötéshossz megkötéssel nyolc replikát számoltunk [48]. Nyolcvan szerkezetet (azaz tízszer nyolc replikát) számoltunk; a teljes gerinc mentén illesztett szerkezetekre az RMSD (1.61±0.64)Å.

A kémiai eltolódásokat a SHIFTX [50] program segítségével számoltuk ki. A SHIFTX program a kémiai eltolódások becslését öt tagra bontva számolja ki:

$$\delta_{calc} = \delta_{coil} + \delta_{RC} + \delta_{EF} + \delta_{HB} + \delta_{HS} \tag{4.3}$$

ahol $\delta_{coil}$ az adott aminosav $^{1}$H, $^{13}$C vagy $^{15}$N random coil eltolódása (DSS-hez viszonyítva) [51], $\delta_{RC}$ a gyűrű eltolódása, $\delta_{EF}$ az elektromos tér, $\delta_{HB}$ a hidrogén hidak hozzájárulása és $\delta_{HS}$ a gerinc dihedrális szögek egy speciális hozzájárulása.

A reziduális dipoláris csatolási állandókat (RDC-ket) a PALES (Prediction of ALignmEnt from Structure) [52] program segítségével számoltuk ki.

Az $S^2$ általános rendparamétereket a Lipari-Szabó közelítésben, a 4.4 egyenlet szerint számoltuk ki [31]:

$$S_k^{2calc} = \frac{3}{2\left(r_k^{eff}\right)^4}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\left[\frac{1}{N_{rep}}\sum_{i=1}^{N_{rep}} r_{i,k,l}r_{j,k,l}\right]^2 - 1\right) \tag{4.4}$$

A fejlesztés és kiértékelés során felhasznált programokat magam írtam PERL programnyelven. Munkámat LINUX operációs rendszer alatt, szabad programok felhasználásával végeztem. A szerverek PERL és C++ programnyelven készültek.

# 5. fejezet

# Eredmények

## 5.1. Fehérjeszerkezetek becslése NMR adatokból

A PRIDE (PRobability of IDEntity; ld 2.2 fejezet [12]) módszer C$\alpha$-C$\alpha$ atomi távolságeloszlások elemzésén alapul. Hasonló jellegű összefüggést kerestem a NOE kényszerfeltételek és a fehérje térszerkezete között. Az NMR spektroszkópiával nyerhető távolsági adatok 6 Ångstrømön belül vannak. Önmagában a távolság jellegű kényszerfeltételek nem alkalmasak arra, hogy az eredeti PRIDE módszerhez hasonlóan többféle szekvenciális távolsághoz is térbeli távolságot rendelünk. Így, számos próbálkozás után végül csupán egyetlen eloszlást vizsgáltam, mégpedig az észlelt kényszerfeltételek számát a szekvenciális távolságoknak megfeleltetve. Ez az eloszlás várhatóan minden fehérjére egyedi és a távolságeloszlás egyértelműen jellemzi majd a konformációt.

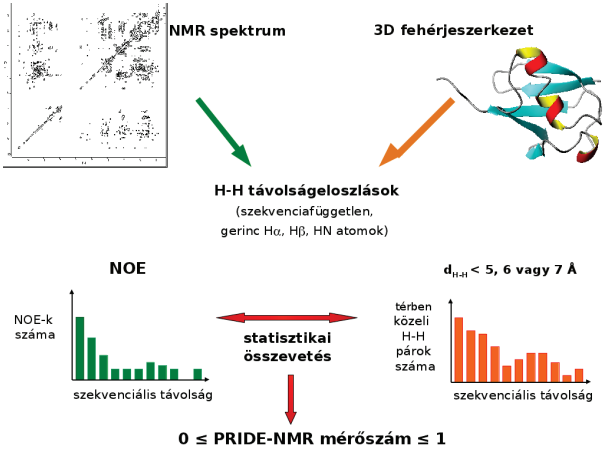
### 5.1.1. A PRIDE-NMR algoritmus fejlesztése

Hogy a módszer ne csak közeli rokon fehérjék esetében legyen használható, a szekvenciától nagymértékben függetlennek tekinthető H$\alpha$, H$\beta$ és HN atomok közötti kényszerfeltételeket használtam fel. Ezek a gerinc protonoknak felelnek meg és az összes lehetséges kombinációjukat vizsgáltam. Az eljárásban nem a NOE adatból számított távolság, hanem csupán a NOE adat megléte a fontos, valamint az, hogy milyen szekvenciális távolságnál (n)[1] észleltük . Az n=1 és n=2 eseteket figyelmen kívül hagytam, mivel azok csak lokális távolságinformációt tartalmaznak és az előzetes vizsgálatok szerint kiugróan magas számuk miatt csupán torzítanák az elemzést. A kényszerfeltételek, azaz a térben

[1] i és i + n között; n = 3 → k, ahol k a kérdéses fehérje lánchossza

közeli H-H párok számát hisztogramon ábrázoltam, amely mutatja, hogy adott szekvenciális távolság esetén hány darab NOE csúcs észlelhető.

Az így kapott hisztogramokat összevetettem a már ismert szerkezetű fehérjékből származtatott távolságelsozlásokkal, hogy a rokon szerkezeteket azonosítani tudjuk. Az első lépésben a térszerkezeti adatbázis adatait át kellett alakítani. A kísérleti adatok alapján számolt szekvenciális távolságeloszlásokhoz hasonló eloszlás számítható ki a fehérje térszerkezetéből is adott távolsági küszöbérték figyelembevételével. Az NOE jelenség hatósugara 6 Å maximálisan. A visszaszámolt távolságeloszlásokhoz a távolság-adatkészletet 5 Å-ös küszöbértékkel számoltam ki. Az eloszlások számításakor szintén csak a H$\alpha$, H$\beta$ és HN atomok közötti kényszerfeltételeket használtam fel. Az atomi koordináták alapján minden egyes fehérjére előállítottam a H-H távolsági eloszlásokat 5 Å-ös küszöbértékre. Ezek az adatok alkotják a továbbiakban a háttéradatbázist.




5.1. ábra. A PRIDE-NMR módszer folyamatábrája

A távolságelsozlások számszerű összevetését a PRIDE eljáráshoz hasonlóan kontingenciaanalízissel [13] végeztem (ld 2.2 fejezet). Az összevetés végeredménye egy nulla és egy közé eső, valószínűségként értelmezhető szám. Mivel a statisztikai módszer megegyezik a PRIDE eljárás statisztikai módszerével, a kapott mérőszámot "PRIDE-NMR"-nek neveztem el. A kapott érték - 0 ≤ PRIDE-NMR ≤ 1 - annak a valószínűsége, hogy a két adatkészlet azonos fehérjefeltekeredést reprezentál. Minél magasabb az érték, annál valószínűbb a szerkezeti hasonlóság a kereső és az adatbázisbeli adatkészlet, és így a reprezentált szerkezetek között. Ha két eloszlás összehasonlításának eredménye nulla, akkor

az eloszláshoz tartozó térszerkezetek teljesen eltérnek egymástól. Ha egy körüli értéket kapunk, valószínűsíthető hogy van hasonlóság. Az eljárás eredményének kiértékelésénél nem csak a konkrét érték fontos, hanem a találatok sorrendje is[1]. Értéktől függetlenül az első találatot mindig érdemes megvizsgálni, mint a szerkezet lehetséges legközelebbi rokonát. Bár a PRIDE-NMR mérőszám elméleti felső határa egy, ezt a gyakorlatban kis valószínűséggel kaphatjuk meg. Az egyik ok, hogy az NMR spektrumokból az összes lehetséges kényszerfeltételt nem tudjuk kimérni és kiértékelni a mérési hibák és a fehérjeszerkezetek dinamikus jellege miatt. A másik ok, hogy nem várható el a listában szereplő összes kényszerfeltétel egyidejű teljesülése a számolt szerkezetben, ugyancsak a fehérjék belső dinamikája miatt.

### 5.1.2. A háttéradatbázis felépítése

A fejlesztés során a RECOORD [37] adatbázisból származtattam a háttéradatbázist. Ezen teszteltem a módszer alkalmazhatóságát és hatékonyságát, mivel a RECOORD adatbázis egyéges és megbízható adatokat tartalmaz[2]. A további tesztek és a szerver fejlesztéséhez már egy reprezentatív adatbázisra is szükség volt, mivel a módszer hatékonysága a kereső adatbázis jóságán is múlik.

A választás a SCOP [38] térszerkezeti adatbázisra esett, mivel ez tartalmazza a negyven aminosavnál rövidebb térszerkezeteket (a CATH pl. nem) és a röntgenszerkezeteket is (a RECOORD pl. nem). A hierarchia negyedik szintjén a szerkezetek egymáshoz képest hasonlónak mondhatók. A 95%-os szekvenciaszűrt adatkészletet használtam. Ez egy redundanciamentes és reprezentatív adatkészlet, mivel a szekvenciában 5%-nál kisebb mértékben eltérő szerkezetek közül csak egy került be az adatbázisba. A röntgenszerkezetek esetén a hidrogén atomok koordinátáit utólag kellett beépíteni. A GROMACS molekuladinamikai programcsomaggal [46], OPLS-AA erőteret [47] alkalmazva dolgoztam. Az alanin metil-csoportjainak helyzeteit átlagoltam, mivel a metil-csoport három hidrogénje ekvivalens és szabadon forog a molekulában, ezért a kényszerfeltételek szempontjából egyetlen egységként kezelendő[3]. Az így kapott adatbázis az atomi koordinátákból előre kiszámolt távolságeloszlásokat tartalmazza.

A NOE effektus az atom-atom távolság növekedésével fokozatosan gyengül (2.1 egyenlet). Ezért felmerült, hogy a koordináták alapján visszaszámolt távolságeloszlások esetében többféle küszöbtávolságot lehessen alkalmazni. A maximálisan mérhető távolságér-

[1] Ún. "nearest neighour classifier", azaz a legközelebbi találaton alapuló eljárás.

[2] A RECOORD adatbázis leírása a 4.1 fejezetben megtalálható.

[3] Mivel a gerinctől a $\beta$-protonoknál távolabbi atomokkal nem dolgoztam, a többi lehetséges metilcsoporttal nem foglalkoztam.

ték 6 Å körüli, ezért választottam az 5, 6 és 7 Å-ös értékeket, és a H-H távolságeloszlásokat mind a három küszöbtávolsággal kiszámoltam. A háttéradatbázis jelenleg 11.490 hisztogramból áll; a továbbiakban erre az adatbázisra SCOPselect néven utalok. Ez nagyjából reprezentálja az ismert fehérjeszerkezet típusokat a SCOP térszerkezeti adatbázis tükrében. A találatok kiértékelésénél a SCOP hierarchia negyedik szintjéig azonos besorolású, azaz az azonos családba tartozó szerkezeteket tekintettem hasonlónak, ezeket neveztem *pozitív találatoknak* az elemzés során. Ha a megtalált szerkezet a kereső NOE adatoknak megfelelő térszerkezetet takarja, azaz önmagát találtam meg, *saját találatnak* neveztem. A háttéradatbázis szekvencia-szűrt, vagyis nem tartalmazza az adott SCOP osztály összes lehetséges képviselőjét, ezért elvileg nem is kaphattunk minden esetben találatot. Az eredmények azt mutatják, hogy a saját találat léte nem befolyásolja az algoritmus eredményességét.

### 5.1.3. Súlyozás bevezetése

A háttéradatbázisban szereplő szerkezetek mérete nagyon eltérő lehet: a 20-30 aminosavastól az akár több száz aminosavas szerkezetekig bezárólag sokféle fehérjeméretttel találkozhatunk. A hisztogramok normalizálása és a tartományok összevonása következtében a kereső szerkezet lánchosszára vonatkozó információ elvész, emiatt az összevetés nem veszi figyelembe a fehérje méretét. Ugyanakkor fontos kiszűrni a hasonló lánchosszú találatokat, mivel ezek biológiailag általában relevánsabbak[1]. A fehérje mérete ismert az NMR-spektroszkópus számára, mindazonáltal ez kevésbé specifikus információ, mint a szekvencia. A szekvencia figyelmen kívül hagyása garantálja, hogy evolúciósan távoli rokon, de hasonló térszerkezetű fehérjék is vizsgálhatóak maradjanak. Az érzékenység növelése érdekében tehát súlyozást vezettem be: a PRIDE-NMR mérőszámot megszoroztam a vizsgált fehérjeláncok hosszarányának adott hatványkitevőn (x = 1, 2 vagy 3) vett hányadosával, az alábbi képlet szerint:

$$\text{PRIDE-NMR\_W}_x = \text{PRIDE-NMR} * \left( \frac{\text{rövidebb fehérje lánchossza}}{\text{hosszabb fehérje lánchossza}} \right)^x \qquad (5.1)$$

A PRIDE-NMR mérőszám súlyozásával a találati sorrend a fehérjehosszak arányának megfelelően átrendeződik. A hossz-szűréssel célzottabb keresés végezhető el; a találatok lánchossza igazodik a kereső fehérje lánchosszához, azaz a nagyon eltérő hosszúságú fehérjékre kapott mérőszámok a lánchosszbeli eltérés mértékétől függően leskálázódnak. A súlyzott hossz-szűrés mellett egy százalékos szűrést is bevezettem. Ebben az esetben

[1] Pl. egy 25 aminosavas fehérje biztosan nem hasonlíthat egy egy doménes 150 aminosavas fehérje *egészére*.

az adatbázisbeli szerkezetek közül az elemzés előtt kizártam azokat, amelyeknek a láncchossza a kereső szerkezet lánchosszától adott százaléknál jobban eltértek (± 5, 10, 15 vagy 20%). A százalékos szűrő segítségével a keresést még tovább lehet finomítani.

### 5.1.4. Tesztelés kiemelt adatkészleten

A SCOPselect adatbázisból negyven, NMR-spektroszkópiával meghatározott szerkezetet választottam ki, nagyjából lefedve a SCOP hierarchiát. A negyven szerkezet a SCOP hierarchiából 40 családot és 37 szupercsaládot reprezentál. A kiválasztott szerkezetekhez tartozó aminosavankénti kényszerfeltételek száma[1] minden esetben egynél nagyobb, azaz minden aminosavra statisztikailag legalább egy távoli kényszerfeltétel létezik. A fehérjeláncok hossza 28 és 182 aminosav között van.

| | Pozitív találatok aránya | | |
|---|---|---|---|
| Küszöbtávolságok | első 5 között | első 10 között | első 100 között |
| 5 Å | 57 | 62 | **100** |
| 6 Å | 42 | 53 | 85 |
| 7 Å | 25 | 38 | 82 |
| **5, 6** Å | **60** | **62** | **97** |
| 6, 7 Å | 38 | 60 | 93 |
| 5, 6, 7 Å | 55 | **65** | 95 |

5.1. táblázat. Pozitív találatok aránya százalékosan a 40 fehérjés tesztkészletre

A tesztelést a fehérjeszerkezet-összehasonlító szerverek érékelése során használt elvek szerint végeztem [40]. Pozitív találatnak a vizsgált szerkezetekkel azonos szupercsaládba (a SCOP hierarchia harmadik szintje) sorolt találatokat tekintettem, és az első száz találat adatait értékeltem ki. Fontos megemlíteni, hogy a fehérjeszerkezet-összehasonlító eljárásokkal ellentétben a PRIDE-NMR módszer esetében a kereső és az adatbázisbeli adatkészlet még azonos fehérje esetében sem egyezik meg. Ugyanis az adatbázisban a térszerkezeti koordinátákból visszaszámolt távolságeloszlások szerepelnek és ezzel szemben a kereső távolságeloszlás a NOE adatokból készült. Ezért a saját találatok sem irrelevánsak a módszer teljesítőképessége szempontjából. Ennek ellenére a teszt szigorúságának megtartása érdekében a saját találatokat nem vettem figyelembe pozitív találatként, azaz

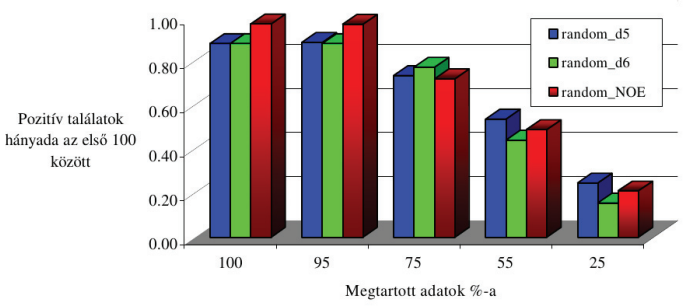
[1] BackBone Restraints Per Residue (BBRPR)

kizártam a találatot az elemzésből, ha a kereső kísérleti adatkészlethez a hozzá tartozó fehérjeszerkezetet találtam meg. Az eredményeket a 5.1 táblázatban tüntettem fel.

A háttéradatbázisbeli küszöbértékre nincs éles határ, mivel a NOE egy lecsengő effektus és 6 Å körül eltűnik. Épp ezért megvizsgáltam, melyik küszöbérték mellett kapom a legjobb eredményeket, azaz a visszaszámolt távolságeloszlásoknál a talált H-H párokat milyen távolságig érdemes figyelembe venni. Az 5 Å-ös, valamint az átlagolt 5 és 6 Å-ös küszöbérték mellett az első 100 találat között legalább egy pozitív találat volt az esetek 97%-ában. Ez az eredmény megközelíti a legjobb fehérjeszerkezet-összehasonlító szerverek teljesítményét [40].

### 5.1.5. A távolságeloszlások elemzése

A háttéradatbázisban atomi koordináták alapján visszaszámolt távolságeloszlások vannak. Mivel a visszaszámolt eloszlás minden lehetséges távolságadatot tartalmaz, ez természetéből adódóan eleve nagyobb adathalmaz, mint a kényszerfeltételekből kapott eloszlások. A használt negyven fehérjéből álló tesztkészletre ez a különbség nagyjából tízszeres.




5.2. ábra. A szerkezetekből visszaszámolt (d=5Å és d=6Å küszöbérték mellett) véletlenszerűen csonkított távolságeloszlások és a véletlenszerűen csonkított NOE távolságeloszlásokra kapott pozitív találati arányok összehasonlítása

A PRIDE-NMR módszer robusztusságának vizsgálatára szisztematikusan "elrontottam" mind a kísérleti, mind a visszaszámolt távolságadatokat. A 40 fehérjés adatkészleten véletlenszerű csonkítást végeztem el. A távolságeloszlásokból véletlenszerűen töröltem a H-H párok 5%-át, majd 15%-át, egészen az adatok 85%-ának kidobásáig, 10%-onként. A

PRIDE-NMR algoritmust a csonkított eloszlásokkal tízszer futtattam minden egyes lépésben. A háttéradatbázist nem változtattam[1]. A kapott PRIDE-NMR értékeket átlagoltam minden csonkításra. Azt a meglepő eredményt kaptam, hogy a csonkított visszaszámolt adatkészletekkel észrevehetően rosszabb eredményeket kaptam, mint a csonkított NOE adatkészletek esetében. A visszaszámolt adatoknál 5 és 6 Å-ös küszöbértékekkel dolgoztam és ezzel a küszöbértékkel számolt adatbázisbeli hisztogramokat használtam. Fölvetődik a kérdés, hogy mivel magyarázható az a tény, hogy a visszaszámolt adatok véletlenszerű elrontása után maga a fehérjefold nem ismerhető fel. Egy lehetséges magyarázat, hogy egyetlen H-H távolság nem reprezentálja elég pontosan a fehérje térszerkezetét, amit az eredeti PRIDE eljárásban [12] 28 C$\alpha$-C$\alpha$ távolságeloszlás átlagolásával kapunk meg. Ez a megfigyelés rávilágít arra, hogy az NOE adatok mennyire jól képesek jellemezni a fehérjefeltekeredés alapvető szerkezeti vonásait. Más szóval, a NOE adatok egyáltalán nem tekinthetők véletlenszerűnek és a spektrumban megjelenő NOE csúcsok valóban reprezentatívak. A NOE adatokból kapott H-H távolságeloszlás tehát, természetéből eredendően tartalmazza a térszerkezet szempontjából leglényegesebb adatokat.

### 5.1.6. Statisztika a PDB adatbázisbeli NMR-szerkezetekre

A PDB adatbázis a legnépszerűbb fehérjeszerkezet-adatbázis. Mind a térszerkezeti koordinátáik, mind a kísérleti adatok elérhetők, ugyanakkor nagyon vegyes a közzétett szerkezetek minősége, megbízhatósága, valamint sok esetben csupán az adatok egy része érhető el. A PDB adatbázisbeli NMR-kényszerfeltételek használhatóságáról készült korábban egy átfogó elemzés [53]. Az NMR-szerkezetek esetében a kényszerfeltételek elemzése azért is előnyös, mert csak egyetlen adatkészlettel kell foglalkozni a számolt szerkezeti sokasággal szemben. Az átfogó elemzés eredményeként statisztikai összefoglalás készült közel 2.000 állományra, amely a H-H kapcsolattípusok előfordulási arányát mutatja be. A PRIDE-NMR algoritmus használhatóságát kívántam felmérni at általánosan közzétett adatokon. Az állományok egy része óhatatlanul terhelt mind mérési, mind kiértékelési hibákkal.

Az elemzéshez a PDB adatbázisban 2007 szeptember 24.-én elérhető összes, kizárólag fehérjére vonatkozó NMR kényszerfeltétel-listát használtam. Ez összesen 3555 állományt jelent. Mivel a SCOP adatbázis alapján történik a szerkezeti rokonság keresés, alapfeltétel volt, hogy az adott állomány be legyen sorolva a SCOP hierarchikus adatbázisba, valamint, hogy egyetlen doménből álljon, azaz a PDB azonosítóhoz egyetlen SCOP azonosító legyen rendelhető. A SCOP adatbázisban ugyanis az evolúciós rokonság a fő szerkezetbesorolási szempont, így előfordulhat hogy egy többdoménes fehérje minden doménje

[1] SCOPselect reprezentatív háttéradatbázis; 5 és 6 Å-ös küszöbérték

más-más SCOP családba van besorolva. Ez a PDB kód és a SCOP szerkezeti családok közötti egyértelmű kölcsönös megfeleltetést tenné lehetetlenné jelen esetben. Már csak 1399 kísérleti adatra volt igaz, hogy egydoménes szerkezethez tartozik. Az algoritmus az X-PLOR formátumú kényszerfeltétel-listákat tudja feldolgozni. Átkonvertálás nélkül ezt 865 állományra tudtuk így alkalmazni. Kizártam még azokat a fehérjéket, ahol egyedül kéviseli a fehérje az adott SCOP fehérjecsaládot, mivel ezekben az esetekben elvileg sem találhatunk önmagán kívül pozitív találatot. Végeredményben 806 NMR kényszerfeltétel-lista teljesítette a szűrési feltételeket. A rutinszerűen vizsgált fehérjék 5-10 kDa-osak (≈ 50-100 aminosavasak); az adatkészlet is tükrözi ezt az arányt.

| Távolság | | d=5Å | | d=5,6Å | | d=5,6,7Å | |
|---|---|---|---|---|---|---|---|
| Saját találatok kizárásával | | első 5 | első 100 | első 5 | első 100 | első 5 | első 100 |
| W0 | első pozitív | 17,77 | | 18,19 | | 18,67 | |
| | pozitív arány (%) | 17,62 | 49,38 | 17,62 | 46,40 | 13,52 | 42,80 |
| | pozitívak száma | 0,33 | 2,46 | 0,32 | 2,39 | 0,26 | 2,22 |
| W1 | első pozitív | 15,18 | | 15,00 | | 16,94 | |
| | pozitív arány | 36,85 | 67,74 | 36,72 | 67,74 | 34,12 | 66,38 |
| | pozitívak száma | 0,75 | 4,36 | 0,77 | 4,33 | 0,69 | 4,11 |
| W2 | első pozitív | 16,78 | | 16,08 | | 15,64 | |
| | pozitív arány | 39,08 | 70,72 | 39,83 | 71,96 | 36,97 | 71,59 |
| | pozitívak száma | 0,81 | 5,07 | 0,83 | 5,13 | 0,77 | 5,01 |
| W3 | első pozitív | 16,31 | | 16,62 | | 16,74 | |
| | pozitív arány | 43,67 | 75,06 | 42,06 | 75,31 | 40,45 | 75,31 |
| | pozitívak száma | 0,88 | 5,67 | 0,88 | 5,75 | 0,83 | 5,68 |

5.2. táblázat. PRIDE-NMR találati statisztika a PDB adatbázisra

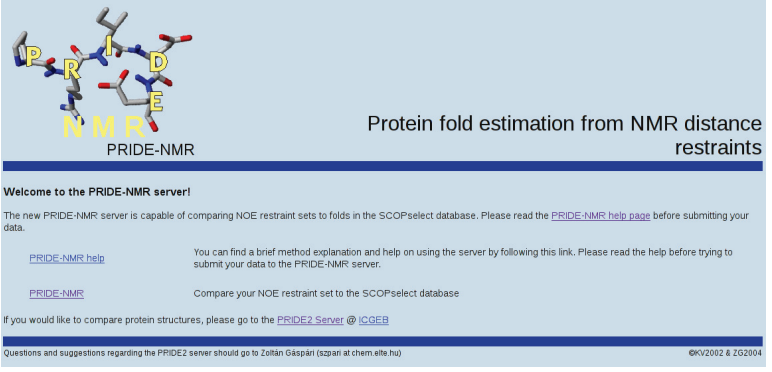
Az elemzést a saját találatok kizárásával, több küszöbtávolság mellett végeztem el. A SCOP adatbázis család szintjén (4. szint) azonos besorolású szerkezeteket tekintettem hasonlónak, azaz pozitív találatnak. A futtatások eredménye a különböző súlyozások mellett (W=0 vagy 3) és a különböző adatbázisbeli küszöbértékek mellett az "első pozitív" az első pozitív találat átlagos helyét mutatja az első 100 találat között. A "pozitív arány" azt jelenti, hogy az első 10, illetve az első 100 találat között van-e pozitív találat és a "pozitívak száma" a pozitív találatok átlagos számát adja meg az első 10, illetve 100 találat között. Ez az elemzési mód megfelel a fehérjeszerkezet-összehasonlító módszerek teljesítményvizsgálatánál alkalmazottakkal [40].

A különböző futtatási eredményeket a 5.2 táblázatban foglaltam össze. A legjobb találati arányokat - 75% - a háttéradatbázisbeli d=5 Å, a d=5 és 6 Å, valamint a d= 5, 6 és 7 Å-re átlagolt küszöbértékek mellett kaptam. Ez azt jelenti, hogy a NOE effektust a két értéken belül lévő távolságok figyelembevételével lehet a legjobban visszaadni. A 6 Å-nél nagyobb távolságok figyelembevétele elrontja a távolságeloszlás jellegzetességét, és a fehérjefold nem lesz azonosítható.

A statisztika alapját képező 806 PDB állomány további elemzése során néhány technikai problémára bukkantam. Néhány állomány esetében nincs elegendő adat a kiértékeléshez: kevesebb, mint öt szekvenciális távolsági tartományra (BIN) van adat, vagy kevesebb, mint tíz NOE adat olvasható ki a kényszerfeltételekből (BIN <5 vagy NOE<10). Több esetben a kiértékelés során a statisztikai összevonások miatt elvész a távolságeloszlás specificitása; ez helikális szerkezetekre különösen igaz. Néhány állomány esetében a nagyon magas aminosavankénti kényszerfeltételek száma (BBRPR >5) kiértékelési hibát sejtet. A problémás állományok kizárásával (az adatkészlet 10%-a) a statisztikai elemzésből, a tisztított adatkészletre 85%-nyi pozitív találat van az első 100 találat között. A PRIDE-NMR teljesítménye figyelemre méltó egyszerűsége és gyorsasága miatt is, valamint azáltal, hogy kizárólag távolság jellegű kényszerfeltételek felhasználásán alapul.

### 5.1.7. A PRIDE-NMR módszer elérhetősége

A PRIDE-NMR módszer elérhető a PRIDE szerverről a http://net.icgeb.org/pridenmr/ oldalról, a PRIDE2 szerver mellett (5.3 ábra).




5.3. ábra. A PRIDE-NMR szerver

A módszer bemenete a kényszerfeltétel-lista távolság jellegű adatai, valamint a fehérje lánchossza. A szerver jelenleg a kényszerfeltétel-listákat kizárólag X-PLOR/CNS formátumban fogadja el. A szerverbe több opció van beépítve. Az opciók az elemzést pontosítják és a hatékonyságot növelik. A háttéradatbázis küszöbértékét érdemes az 5 Å-ös, vagy az 5 és 6 Å-re átlagolt értéken hagyni. A kezdő szekvenciális távolság értéke változtatható. Ha nagyon nagy az eltérés az első tartománybeli NOE darabszám és az eloszlás többi tartományának NOE darabszáma között, érdemes elvégezni az elemzést a 3-as helyett 4-es szekvenciális távolságtól indulva is. A százalékos szűrő, a súlyozás, valamint a kiadott legjobb találatok darabszáma is állítható.

### 5.1.8. Példa a PRIDE-NMR szerver használatára: az SH3 domén

Az SH3, vagy Src Homology 3 domén hatvan aminosav körüli, elterjedt és jól konzervált szerkezet. Jelátviteli folyamatokban játszik fontos szerepet (citoszkeleton, Ras fehérjék, Src kinázok). A humán genomban több, mint háromszáz SH3 domént kódoló régió ismert. Kis méretének köszönhetően NMR spektroszkópiával is könnyen vizsgálható. Az elemzéshez az 1GBR kódú állomány kényszerfeltétel-listáját választottam ki [54]. Az egér GRB2 (Grow factor Receptor Bound 2) N és C terminális doménjének felel meg a kiválasztott 74 aminosavas SH3 domén.

A szerverrel kapott keresési eredményt a 5.4 ábrán mutatom be, 5 Å-ös háttéradatbázis küszöbérték mellett. A tíz legmagasabb mérőszámú találat adatait tüntettem fel, W=3-as lánchossz-súlyozás szerint rendezve. A találatok a SCOP kódok alapján egyértelműen azonosíthatók és a térszerkezeti osztályokból a szerkezeti rokonságok kiolvashatók. A legelső találat saját találat, 0,965 PRIDE-NMR értékkel: maga a "d_1gbra" domént. Az első három találat kiemelten jó találat, a PRIDE-NMR értékek 0,8 felettiek. Az első tíz találat között összesen nyolc tartozik az SH3 domén családjába a SCOP besorolás értelmében (b.34.2.1). A megfelelő térszerkezeteken jól látszik a szerkezetbeli hasonlóság, egymással és a kereső szerkezettel is; az első találat saját találat is egyben (5.5 ábra).

Ha a keresést lánchossz-súlyozás nélkül végzem el, a legjobb találat "d_emxa_", egy 30 aminosavas pók toxin, 0,995 PRIDE-NMR értékkel. A második találat már egy SH3 domén (d_1gl5a) 0,980 PRIDE-NMR értékkel és az első tíz találat között súlyozás nélkül is 6 SH3 domén van. A háttéradatbázis küszöbértékét d=5 és 6Å-re állítva és súlyozás nélküli keresésnél összesen 8 találat SH3 domén, és a saját találat (d_1gbra_) a 10. találat 0,931 PRIDE-NMR értékkel. Ezzel szemben a súlyozással már csak hat SH3 domén van az első tíz találat között, de a saját találat első találat lesz. A keresés tanulsága, hogy lehetőleg az összes paramétert érdemes végigpróbálni, és a konkrét PRIDE-NMR értékeket a találatok szerkezeti besorolásával párhuzamosan értelmezni.

## PRIDE-NMR results (specified query length: 74)

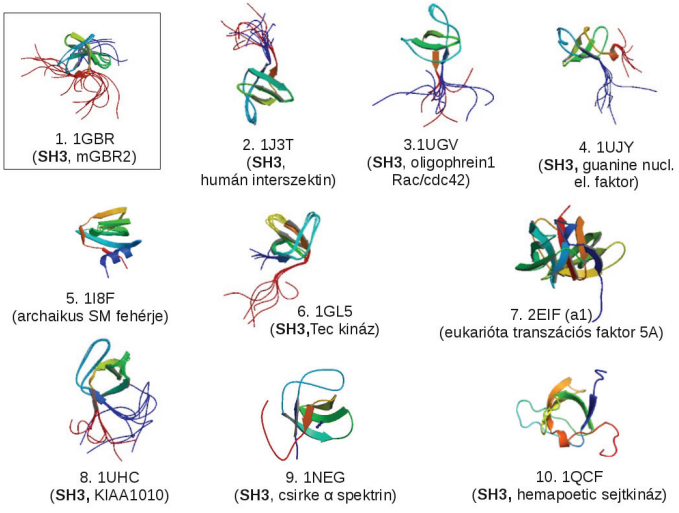
NOE per residue: 1.49

**Displaying results 1-10 sorted by PRIDE_NMR_W3**

Minimal sequential distance used: 3

| HIT | SCOP ID | LENGTH | PRIDE_NMR | PRIDE_NMR_W1 | PRIDE_NMR_W2 | PRIDE_NMR_W3 |
|---|---|---|---|---|---|---|
| 1 | d1gbra_ (b.34.2.1) | 74 | 0.965 | 0.965 | 0.965 | **0.965** |
| 2 | d1j3ta_ (b.34.2.1) | 74 | 0.931 | 0.931 | 0.931 | **0.931** |
| 3 | d1ugva_ (b.34.2.1) | 72 | 0.904 | 0.880 | 0.856 | **0.810** |
| 4 | d1ujya_ (b.34.2.1) | 76 | 0.809 | 0.788 | 0.767 | **0.727** |
| 5 | d1i8fa_ (b.38.1.1) | 71 | 0.847 | 0.813 | 0.780 | **0.718** |
| 6 | d1gl5a_ (b.34.2.1) | 67 | 0.980 | 0.888 | 0.804 | **0.659** |
| 7 | d2eifa1 (b.34.5.2) | 73 | 0.639 | 0.630 | 0.622 | **0.605** |
| 8 | d1uhca_ (b.34.2.1) | 79 | 0.767 | 0.718 | 0.673 | **0.590** |
| 9 | d1nega_ (b.34.2.1) | 65 | 0.978 | 0.859 | 0.755 | **0.582** |
| 10 | d1qcfa1 (b.34.2.1) | 65 | 0.963 | 0.846 | 0.743 | **0.573** |

5.4. ábra. Az 1GBR állomány kényszerfeltételek keresési eredménye (d=5Å, W=3)

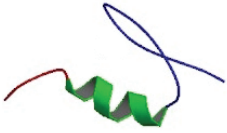



5.5. ábra. A humán SH3 doménre talált hasonló szerkezetek (d=5Å, W=3)

### 5.1.9. A módszer korlátai

A PRIDE-NMR algoritmus kizárólag a távolság jellegű kényszerfeltételeken alapszik. Fontos szem előtt tartani, hogy más típusú kényszerfeltételek is nyerhetők az NMR-spektrumok alapján (hidrogénhidak, diéderes szögek, kötések egymáshoz viszonyított orientációja) és a szerkezetszámolás során ezeket is figyelembe veszik, különösen akkor, ha nem nyerhető elegendő távolság jellegű kényszerfeltétel.

A PRIDE-NMR módszer bizonyos esetekben nem szolgáltat értékelhető eredményt. A 5.3 táblázatban három példát gyűjtöttem össze az eredménytelenségre. A "BBRPR" (BackBone Restraints Per Residue) érték az aminosavankénti átlagos távolság jellegű kényszerfeltételek számát adja meg.

| PDB kód<br><br><br><br>szerkezet | 1P88<br>  | 1VND<br>  | 1SP2<br>  |
|---|---|---|---|
| SCOP család | d.68.2.2 | a.4.1.1 | g.37.1.1 |
| lánchossz | 216 | 77 | 31 |
| BBRPR | 2,23 | 152,32 | 0,65 |

5.3. táblázat. Példák a PRIDE-NMR keresés eredménytelenségére

Az 1P88 PDB kódú állomány egy foszfát szintáz szerkezetet tartalmaz. A szerkezet SCOP családjának összes többi képviselője legalább kétszer nagyobb fehérje (nagyobb, mint 400 aminosav). A hossz szerinti súlyozás ebben az esetben lerontja a hatékonyságot. Viszont a súlyozás nélküli első három találat a kereső szerkezettel azonos SCOP osztályba van sorolva, tehát realisztikus és reprezentatív eredményt kaptam.

A kísérleti adatok minősége nagyon vegyes. Egy kisebb fehérje esetében elvben könnyebb több kényszerfeltételt gyűjteni, de a térszerkezet szempontjából informatív, azaz távoli NOE-k száma nem feltétlenül arányos ezzel a növekedéssel. Egyes kényszerfeltétel listák sok, nem egyértelmű kényszerfeltételeket tartalmaznak vagy egyéb formai hibával terheltek. Az 1VND PDB kódú állományhoz tartozó kényszerfeltétel-lista érdekes módon az összes lehetséges H-H távolság jellegű kényszerfeltételt tartalmazza. Azokhoz a H-H párokhoz, amelyek nem kísérleti adatra vonatkoznak, egy irreálisan magas távolságot (99 Å) rendeltek a szerzők. Mivel a PRIDE-NMR algoritmusban a kényszerfeltétel

megléte a figyelembevétel kritériuma és nem a konkrét távolság, fetételezve hogy csak azok a H-H adatok szerepelnek a listában amelyek ténylegesen NOE csúcsnak felelnek meg, teljesen irreális aminosavankénti kényszerfeltétel számot kapunk és az elemzés eredménye nem specifikus és irreleváns.

A találatok hiányának oka részben az, hogy egyes állományokban rendkívül kevés informatív szekvenciális távolsági adat található, vagyis a generált távolságeloszlás alapján nem lehet a fehérje feltekeredését jellemezni. Összesen három vagy négy egyedi távolsági adatból még a PRIDE-NMR algoritmus sem képes megbecsülni a szerkezetet, de ez nem is várható el. A PDB adatbázisban elérhető kísérleti kényszerfeltétel-listák alapján számolt távolságeloszlások egy részében rendkívül kevés adat áll rendelkezésre. A kevesebb, mint öt távolsági tartományt tartalmazó eloszlásokra végzett számszerű összevetések esetén egyáltalán nem kaptam eredményt. Az 1SP2 PDB kódú állomány egy nagyon rövid, 31 aminosavas zink-ujj doménnek felel meg. Mivel a szekvencia rövidsége miatt elméletben is csak nagyon kevés távolsági kategória lehetséges és kevés NOE adat is áll rendelkezésre, a statisztikai összevetés során nem marad értékelhető eredmény a kevés összevethető távolsági kategória miatt.

A számszerű összevetés során tartomány-összevonás történhet, mivel a $\chi^2$ érték csak akkor számolható át megbízhatóan, ha egyik tartományba sem esik az adatok 5%-ánál kevesebb távolságérték. Ezért, ha adott távolsági tartományba aránylag sok NOE tartozik, az összevonás után előfordul, hogy már nem értékelhető ki az eloszlás. A helikális szerkezetekre ez kiemelten igaz, mivel a kis szekvenciális távolságokra aránylag sokkal több NOE adat gyűjthető. Ilyen esetben érdemes az első tartomány (3-as szekvenciális távolság) figyelmen kívül hagyásával is elvégezni az elemzést.

### 5.1.10. Lehetséges alkalmazások

A PRIDE-NMR eljárással kizárólag NMR kísérletekből rendelkezésre álló távolság jellegű kényszerfeltételek alapján képesek vagyunk rokonítani a vizsgált szerkezetet más, már ismert térszerkezetű fehérjékkel [I, II]. Az eljárás gyors, az eredmények egy percen belül rendelkezésre állnak. Az NMR szerkezetszámolást sokszor megnehezíti, hogy nem áll rendelkezésre megbízható kiindulási szerkezet a folyamat elején. A PRIDE-NMR algoritmus segítségével lehetőség nyílik egy jól használható szerkezetből indítani a számolást, mivel a módszer más jellegű, közvetlenebb kapcsolatot teremt a NOE adatok és a térszerkezet között a szerkezetszámolás mellett. Ily módon a szerkezetszámolás folyamata megkönnyíthető és felgyorsítható. A PRIDE-NMR módszer alkalmas az NMR kényszerfeltételek teljességének ellenőrzésére is. Segítségével felderíthetőek az egyes

jelhozzárendelési hibák, illetve eldönthető, hogy az adott kényszerfeltétel-lista alkalmas-e megbízható szerkezetszámolások elvégzésére.

A PDB adatbázison elvégzett statisztikai elemzés tükrében több paramétert is azonosítottam, amely egyrészt a módszer teljesítőképességét befolyásolja, másrészt világosan mutatja az adott kényszerfeltétel-készlet minőségét, azaz a szerkezetszámolás várható eredményességét. Ugyanis, ha nagyon kevés távolság jellegű kényszerfeltételt tudunk csak gyűjteni, kizárólag ezek alapján nem lehet megfelelő minőségű szerkezetet számolni. Másrészt a már kiszámolt szerkezeti sokaságok minősége is ellenőrizhető az algoritmus segítségével. Ebben az esetben a háttéradatbázist a számolt sokaság modelljei alkotják. Vizsgálhatjuk a kísérleti kényszerfeltételeket és a számolt szerkezeti sokaság közötti megfeleltetést, például átlagos PRIDE-NMR mérőszámot számolva a teljes sokaságra.

Összefoglalva, egy működő, hatékony és új módszert fejlesztettem ki, amellyel kizárólag kísérleti adatok alapján, képesek vagyunk atom-atom távolságeloszlásokból pillanatok alatt rokonítani a vizsgált szerkezetet más, már ismert térszerkezetű fehérjékkel. A tudományos közösség számára a módszer elérhető a világhálóról. A módszer:

- szekvenciális hasonlóság hiányában is működik,
- kiindulási szerkezetet eredményez a szerkezetszámoláshoz és így a szerkezetmeghatározás folyamata lerövidíthető,
- ismert szerkezetek NOE adatainak teljessége és jósága ellenőrizhető, a fehérje mozgékonyságáról is kaphatunk információt.

Az eljárás alkalmas arra, hogy kialakítsunk egy első képet a vizsgált fehérje térszerkezetéről. Az eredményeim a jövőben megkönnyíthetik a NMR spektroszkópiás fehérjeszerkezet-meghatározást.

## 5.2. Kapcsolat a szerkezeti sokaság és a kísérleti adatok között

Az NMR-spektroszkópiás fehérjeszerkezet-meghatározás végeredménye egy sokaság: azok a számolt szerkezetek, amelyek a legjobban megfelelnek a kísérleti adatoknak (távolság jellegű kényszerfeltételek, diéderes szögek, hidrogén-hidak). A kapott szerkezetek jósága nehezen ellenőrizhető, különösen az ún. dinamikus szerkezeti sokaságok esetében[1].

Zargovic és munkatársai elemezték a számolt NMR-szerkezetek és a NOE adatok közötti megfeleltetést [55]. Rendkívül változatosak lehetnek a számolt szerkezeti sokaságok, sőt a fehérje feltekeredésének sem felelnek feltétlen meg, annak ellenére, hogy a NOE adatkészlettel összeegyeztethetők. Továbbá a számolt szerkezetekben lehetnek olyan NOE kapcsolatok, amelyek a kísérletben nem mutathatók ki vagy nem is léteznek. A kísérleti adatok minősége meghatározza a PRIDE-NMR algoritmus teljesítőképességét. A számolt szerkezetek viszont nem biztos, hogy tükrözik a fehérje dinamikus tulajdonságait. A számolt sokaságon végezve az elemzést, megkapjuk a legjobb szerkezetet, és a szerkezeti sokaság általánosan jellemezhető egy egyértelmű mérőszámmal. A PRIDE-NMR módszer a távolság jellegű kényszerfeltételeknek felelteti meg a szerkezeteket. További, a fehérjelánc dinamikájával kapcsolatba hozható NMR paraméternek is meg lehet feleltetni a szerkezeteket. Ennek megvalósítására egy integrált minőségellenőrző szervert állítottunk össze. A szerver arra ad választ, hogy mennyire felelnek meg a szerkezeti sokaság konformereire visszaszámolható paraméterek a kísérleti adatoknak (NOE, $S^2$, RDC, kémiai eltolódások, csatolási állandók).

### 5.2.1. A PRIDE-NMR módszer alkalmazása szerkezeti sokaságon

Az ubiquitin esetében mutatom be a PRIDE-NMR használatát szerkezeti sokaságra. Ezt a vizsgálatot egy minőségellenőrző alkalmazásnak lehet tekinteni. Az ubiquitin fontos szerepet játszik a fehérjelebontás folyamatában az eukarióta sejtekben. Jól konzervált szerkezetű, univerzálisan előforduló, 76 aminosavas fehérjéről van szó. Az élesztőből származó és az emberi ubiquitin között 96%-os a szekvencia-azonosság. A PDB adatbázisban több módszerrel is meghatározott szerkezet is megtalálható: egyrészt röntgenkrisztallográfiával (RCSB PDB kód: 1UBQ [56]), másrészt NMR spektroszkópiával is (RCSB PDB kód: 1D3Z [49].

A fehérjék szerkezete és a mozgékonyság közötti kapcsolat figyelembevétele egyre fontosabb az NMR szerkezet-meghatározása során. A DER (Dynamic Ensemble Refine-

[1] ld. 4.3 fejezet

ment) vagy a MUMO (Minimal Under-restraining Minimal Over-restraining) számolási protokoll alkalmazásával a távolság jellegű kényszerfeltételek mellett a szerkezetek a dinamikai paraméterekkel is összhangba vannak hozva. A PDB adatbázisban az ubiquitinre mindkét szerkezeti sokaság elérhető. A DER protokollal számolt, NOE és $S^2$ adatokon alapuló szerkezeti sokaságot 128 konformer alkotja (RCSB PDB: 1XQQ [6]. A MUMO protokoll fejlesztése során a Richter és munkatársai molekuladinamikai szimulációval kapott konformer sokaság adatait reprodukálták. Ennek erdménye a 144 modell (RCSB PDB kód: 2NR2 [29]). Egy harmadik szerkezeti sokaságot számoltunk a GROMACS molekuladinamikai programcsomaggal, a NOE és az $S^2$ kísérleti adatokat felhasználva (80 szerkezet, MUMO protokoll, 5ns, 300K, explicit víz; ld 4.4.fejezet). Mind a DER, mind a MUMO sokaságok esetében a modellek jól illeszkednek a gerinc mentén, viszont a C-terminális rész nagyon mozgékonynak bizonyult.

| háttér adatbázisbeli | | PRIDE-NMR értékek (1D3Z kényszerfeltételekre) | | |
|---|---|---|---|---|
| küszöbtávolságok | szerkezeti sokaság | maximális | minimális | átlag ± szórás |
| d = 5 Å | 1XQQ (128) | 0,997 | 0,170 | 0,660 ± 0,203 |
| | 2NR2 (144) | 0,961 | 0,377 | 0,667 ± 0,144 |
| | 1D3Z_mumo (80) | 0,954 | 0,093 | 0,576 ± 0,217 |
| d = 5 és 6 Å | 1XQQ (128) | 0,942 | 0,247 | 0,625 ± 0,154 |
| | 2NR2 (144) | 0,880 | 0,276 | 0,604 ± 0,130 |
| | 1D3Z_mumo (80) | 0,886 | 0,129 | 0,538 ± 0,156 |

5.4. táblázat. A PRIDE-NMR keresés eredményei az ubiquitin (PDB kód: 1D3Z) kényszerfeltétel-lista esetén, d = 5 Å, valamint d = 5 és 6 Å küszöbértékek mellett

Az 1D3Z kényszerfeltétel-listát összevetettem a három szerkezeti sokaság konformereinek háttéradatbázisával. Minden sokaságra megadtam a legnagyobb, legkisebb és az átlagos PRIDE-NMR értéket. Az eredményeket az 5.4 táblázat foglalja össze. Mindhárom sokaságra a mérőszám mind maximális, mind átlagos értéke magas. A vizsgált fehérje mozgékonyságát a 0,600 körüli átlagos érték jól tükrözi; még a lehető legjobb protokoll esetén sem teljesülhet az összes kényszerfeltétel. Nem elegendő a legjobb szerkezetet kiragadni a szerkezeti sokaságból, még akkor sem, ha a kényszerfeltételeknek ez feleltethető meg a legjobban. Fontos figyelembe venni a teljes számolt sokaságot, nem feltétlenül elegendő egy esetlegesen kiválasztott "reprezentatív konformer" vizsgálata, mint az a gyakorlatban sokszor előfordul.

### 5.2.2. A CoNSEnsX megközelítés

Az NMR spektroszkópiai adatokból számolt szerkezeti sokaságok minőségellenőrzésére a PRIDE-NMR módszer tehát alkalmasnak bizonyult. Felmerült, hogy a távolság jellegű kényszerfeltételek mellett további, a molekula dinamikáját jellemző kísérleti paramétert is ellenőrizni lehetne. A szerkezetekből visszaszámolt paramétereket össze lehet vetni a kísérleti adatokkal. E feladat megvalósítására állítottuk össze a CoNSEnsX (Consistency of NMR-derived Structural Ensembles with eXperimental data) módszert [III].A protokoll célja egységesen megfeleltetni a szerkezeti sokaságokból visszaszámolt paramétereket az NMR adatokból származtatott paraméterekkel és ily módon egy összetett minőségellenőrzést szolgáltatni. A paramétereket a szerkezetek alapján számoltuk vissza, a használt programokat és egyenleteket a 4.2 fejezetben mutattam be. A CoNSEnsX protokoll az alábbi NMR paramétereket használja:

- A $^{1}$H-$^{1}$H távolsági kényszerfeltételeket a PRIDE-NMR [I] segítségével dolgozzuk fel. Minden konformert megfeleltetünk az NOE kényszerfeltétel-listának (ld. 5.1 fejezet). A program hisztogramként ábrázolja az egyes konformerekre kapott PRIDE-NMR értékek eloszlását, valamint a minimális és maximális értékeket, az átlagot és a szórást adja meg a teljes sokaságra.
- Az $S^2$ általános rendparamétereket a 4.4 képlet alapán számoltuk vissza a szerkezeti sokaságból (ld. 4.2 fejezet; [31]). Jelenleg a gerinc N-H és C$\alpha$-H$\alpha$ rendparamétereket számolja ki a program.
- A kémiai eltolódásokat a SHIFTX [50] program segítségével számoljuk vissza minden egyes konformerre, majd átlagolunk minden egyes magra. Jelenleg a SHIFTX program a C$\alpha$, H$\alpha$, amid N, amid H és C$\beta$ kémiai eltolódásokat számolja ki. A glicin esetében a két H$\alpha$ átlagával számoltunk mind a kísérleti adatoknál mind a visszaszámolt adatok esetében.
- A reziduális dipoláris csatolásokat (RDC-k) a PALES [52] program segítségével számoltuk vissza. Minden konformerre külön-külön kiszámoljuk, majd átlagolunk.
- A skaláris csatolásokat a teljes sokaság átlagaként számoljuk ki. Egy adott konformerre az értékeket a $\phi$ gerinc dihedrális szög alapján, a Karplus egyenlet segítségével számoljuk ki:

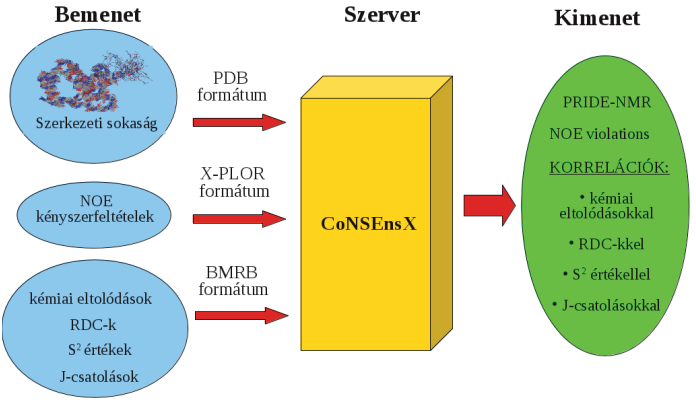
$$J(\phi) = A * \cos^2\phi + B * \cos\phi + C \tag{5.2}$$

  ahol J a $^{3}$J csatolási állandó, $\phi$ a dihedrális szög és A, B és C empirikus úton származtatott atom-függő paraméterek, melyeket jelen esetben az [57] közlemény 1. mellékletbeli táblázat NMR/X-ray sor adatait használtuk. A $\phi$ gerinc torziós szög ismeretében kiszámolhatók a következő J-csatolási állandók: $^{3}J_{H^N H^\alpha}$, $^{3}J_{H^\alpha C}$, $^{3}J_{H^N C^\beta}$, $^{3}J_{H^N C}$,

### 5.2.3. A CoNSEnsX szerver felépítése

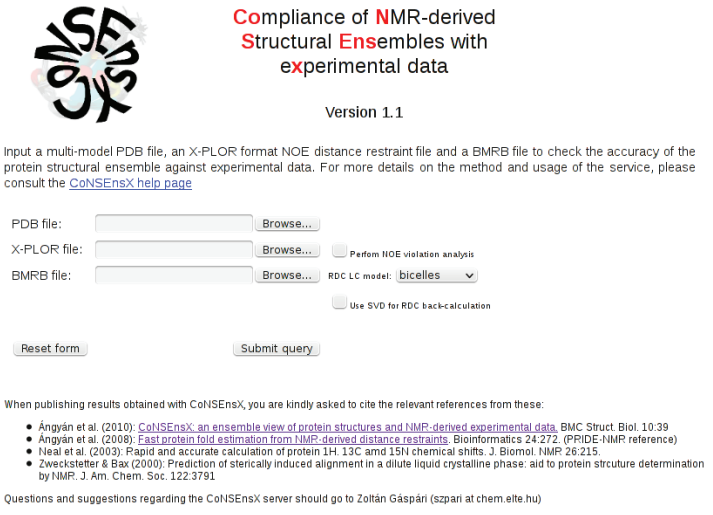
A CoNSEnsX szerver megtalálható a http://consensx.chem.elte.hu címen. A szerver felépítését a 5.6 ábra szemlélteti. A szerver megjelenítését a honlapon a 5.7 ábra mutatja be. Nem várható el, hogy egy adott rendszer esetén az összes lehetséges NMR paraméterre rendelkezésre álljon kísérleti adat. Fontos viszont, hogy az adatok hasonló körülmények között végzett kísérletekből származzanak (oldószer, hőmérséklet, pH, ionerősség, stb). A program három bemeneti állományt vár:

- atomi koordináták PDB formátumban
- távolság jellegű kényszerfeltétel-lista XPLOR/CNS formátumban
- NMR paraméterek (kémiai eltolódások, $S^2$-ek, RDC-k, csatolási állandók, stb...) NMR-STAR (BMRB) formátumban.




5.6. ábra. A CoNSEnsX szerver elvi felépítése

A hiányzó paramétereket nem veszi figyelembe a program, az adatok közül csak az atomi koordináták megadása kötelező. Az atomi koordináták takarhatnak egyetlen konformert vagy többet. Utóbbi esetben a program egyes paramétereket minden konformerre külön-külön is megad. Az $S^2$ általános rendparaméterekre, a kémiai eltolódásokra és a csatolási állandókra a CoNSEnsX szerver a kísérleti adatok és a szerkezet alapján visszaszámolt értékek között korrelációkat, jósági faktort (q-faktor) és RMSD-t a sokaságra átlagolva számolja ki. A kimenet tehát két fő részből tevődik össze: a NOE adatok értékelését egyrészt a PRIDE-NMR kimenetben, másrészt a nem teljesülő NOE adatok hisztogrammos ábrázolása teszi ki.




5.7. ábra. A CoNSEnsX szerver megjelenítése

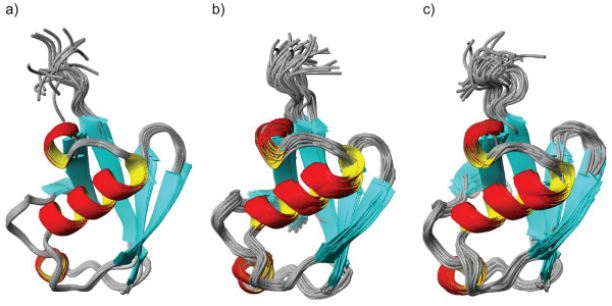
### 5.2.4. Példa a CoNSEnsX szerver használatára: humán ubiquitin, mint globuláris fehérje

Az ubiquitint előszeretettel vizsgálják az NMR spektroszkópusok változatos módszerekkel és körülmények között. Az ubiquitin $S^2$ értékei magasak, tehát a ps-ns időskálán rigid, merev struktúrával jellemezhető a molekula. Különböző módon meghatározott szerkezeti sokaságok segítségével teszteltük a különböző szerkezetek megfeleltetését a kísérleti adatoknak. Összesen 13 szerkezetet gyűjtöttem össze a PDB és a RECOORD adatbázisokból, valamint három további szerkezeti sokaságot számoltunk. Az elérhető kísérleti adatokkal rendelkező NMR-szerkezeteket a 5.5 táblázatban foglaltam össze.

A röntgendiffrakcióval meghatározott szerkezetek mellett dinamikus szerkezeti sokaságok is szerepelnek az NMR-spektroszkópai adatok alapján számolt szerkezeti sokaságok között. A DER és MUMO dinamikus szerkezeti sokaságok mellett az ISD (Inferential Structure Determination; [58]) sokaság is szerepel. További három sokaságot számoltunk. A COCO (COmplementary COordinates; [62]) módszer további konformerekkel egészíti ki az eredetileg számolt sokaságot oly módon, hogy az tükrözze a teljes diverzitást (U_COCO; 5.8 a) ábra; 20 konformer). A másik két sokaság molekuladinamikai számításokból ered. Az U_NNR sokasághoz (5.8 b) ábra; 32 konformer) NOE, NH $S^2$ és

| Szerkezetazonosító | Leírás | modellek száma | referencia |
|---|---|---|---|
| U_1D3Z | oldat NMR | 10 | [49] |
| U_COCO | oldat NMR + COCO | 20 | [III] |
| U_1XQQ | DER sokaság | 128 | [6] |
| U_2NR2 | MUMO sokaság | 144 | [29] |
| U_2K39 | EROS sokaság | 116 | [4] |
| U_ISD | ISD sokaság | 25 | [58] |
| U_NNR | NOE(2)+$S^2$(8)+RDC(8) | 32 | [III] |
| U_1UBQMD | 5ns MD szimuláció 1XQQ alapján | 32 | [III] |
| U_CNS | RECOORD (CNS) | 25 | [37] |
| U_CNW | RECOORD (CNS vízben) | 25 | [37] |
| U_CYA | RECOORD (CYANA) | 25 | [37] |
| U_CYW | RECOORD (CYANA vízben) | 25 | [37] |
| U_1G6J | ubiquitin reverz micellákban | 32 | [59] |
| U_1V80 | ubiquitin 30 bar nyomáson mérve | 10 | [60] |
| U_1V81 | ubiquitin 300 bar nyomáson mérve | 10 | [60] |
| U_2JZZ | szilárd fázisú NMR szerkezet | 20 | [61] |

5.5. táblázat. Az elemzésben használt humán ubiquitin szerkezeti sokaságok




5.8. ábra. Humán ubiquitin dinamikus szerkezeti sokaságok: a) U_COCO; b) U_NNR; c) 1UBQ_MD

N-H RDC kísérleti adatokat használtunk fel. A harmadik sokaságot a röntgenszerkezetből számoltuk ki, molekualdinamikai szimulációval, kényszerfeltételek nélkül (1UBQ_MD; 5.8 c) ábra; 32 konformer). Mindhárom esetben a konformereket a molekulagerinc mentén egymásra illesztettem a MOLMOL program segítségével [63].

A BMRB adatbázis[1] és az irodalom alapján összegyűjtöttem a rendelkezésre álló NMR kísérleti adatokat. Az összevetéshez egyetlen kísérleti adatkészletet használtunk, természetes körülményeknek megfelelően annak ellenére hogy egyes adatok más hőmérsékleten is rendelkezésre álltak. Ily módon a különböző mérési körülményekből eredő eltéréseket is elemezhetjük a sokaságok között. Alapos mérlegelés után a következő kísérleti paramétereket választottuk ki az ubiquitin sokaságok elemzésére:

- a távolság jellegű kényszerfeltételként az RCSB PDB 1D3Z kód alatt található listát használtuk, a listából csak az 1320 egyértelmű kényszerfeltételt tartottuk meg;
- a gerinc NH $S^2$ adatokat a szerzők (Chang és mtsai. [64]) bocsátották rendelkezésünkre; a 20°C-on mért adatokat használtuk fel;
- C$\alpha$-H$\alpha$ $S^2$ rendparamétereket a BMRB 6466-os állomány tartalmazza [65];
- a kémiai eltolódásokat a BMRB 6466-os állomány tartalmazza [65];
- N-H RDC adatokat Cornilescu [49] alapján vettük;
- N-H$\alpha$ RDC adatokat Perttu Permi [66] bocsátotta rendelkezésünkre;
- C$\alpha$-H$\alpha$, C-C$\alpha$, C-H$\alpha$ RDC adatokat a szerzők (Wurtz és mtsai [67]) bocsátották rendelkezésünkre;
- J-csatolási állandókat Wang [57] cikke mellékletéből töltöttük le (Supplementary Material, Table 2).

A $^1$H-$^1$H távolság jellegű kényszerfeltételeket X-PLOR formátumba, az összes többi NMR kísérleti adatot egyetlen BMRB formátumú állományba foglaltuk. Minden egyes szerkezeti sokaságra ezzel a két kísérleti adatsorral végeztük el az elemzést. A 5.9 ábra egy tipikus CoNSEnsX kimenetet mutat be, konkrétan az U_NNR sokaságra. Az ábrák a MOLMOL programmal készültek [63]. Az összevetéseket NMR adattípusonként csoportosítva az 5.10 ábra foglalja össze. Az eredményekből kiolvasható, hogy nem tapasztalunk drámai eltérést a kísérleti adatok és a szerkezeti sokaságok megfeleltetésében. Ez annyiban meglepő, hogy nagyon eltérő technikákkal és körülmények között meghatározott szerkezeteket egyetlen kísérleti paraméterkészlettel vetettünk össze a számolt szerkezeti sokaságokkal. Az elemzés tehát csak arra a kérdésre ad választ, hogy mennyire feleltethetők meg a különböző módszerekkel, különböző körülmények között meghatározott szerkezetek az oldatbeli, szobahőmérsékleten és légköri nyomáson felvett kísérleti

[1] Biological Magnetic Resonance data Bank

adatoknak. Az eltérő körülményekből adódó esetleges szerkezeti változás ily módon követhető lehet.

Az amid NH $S^2$ paraméterek tekintetében minden szerkezet jól teljesít, ellentétben a C$\alpha$-H$\alpha$ $S^2$ értékekkel, amit egyetlen sokaság számításánál se használtak kényszerfeltételként. Az amid NH $S^2$ paraméterek egyformán magasak a szekvencia mentén, kivéve a C-terminálisnál.

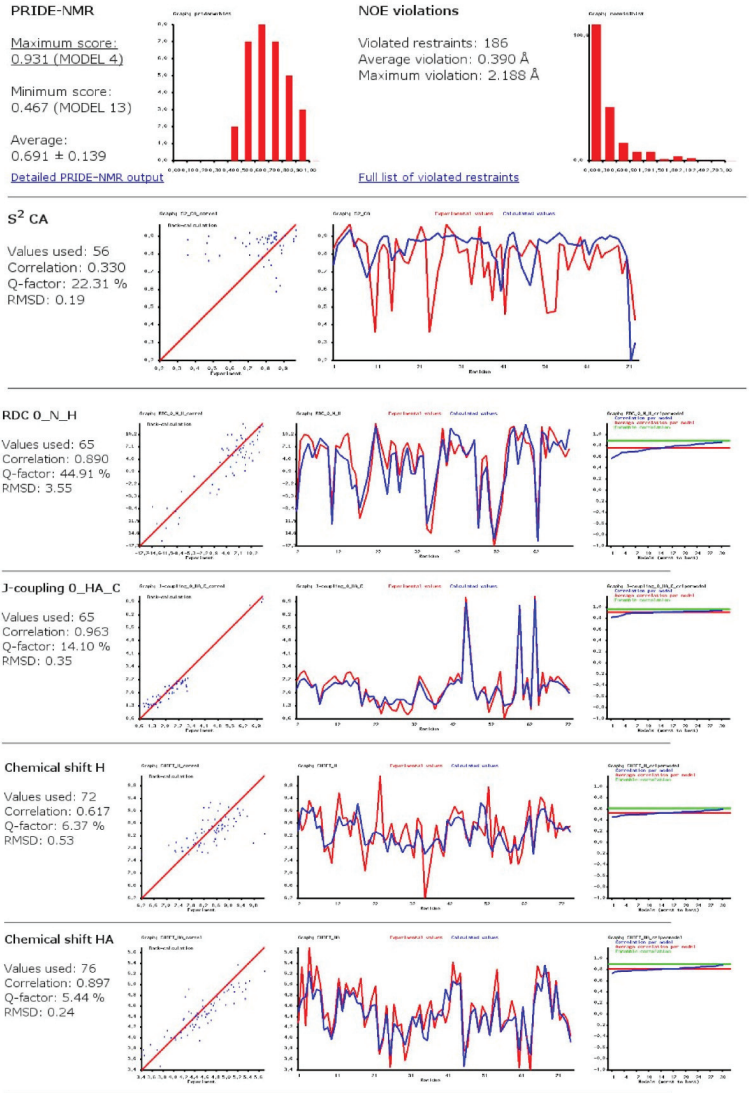
Az RDC adatok elfogadható szinten egyeznek meg az összes sokaságnál, kivéve a H$\alpha$-N adatkészletre. Megjegyzendő, hogy minden RDC adat *ab initio* lett kiszámolva, SVD (singular value decomposition) alkalmazása nélkül, a kísérleti adatok alapján. A CoNSEnsX engedi az SVD használatát; ehhez a PALES programot "best fit mode"-ban kell meghívni az alapértelmezett molekuláris illesztés helyett.

A kémiai eltolódások is jó egyezést mutatnak a kísérleti adatokkal az összes ubiquitin sokaságra. Az eltérő érzékenység a különböző szerkezeti faktorokra nagyon szépen kirajzolódik, ahogy ezt a 5.10 ábra is szemlélteti. Például a C$\beta$ kémiai eltolódások függnek leginkább az aminosav típusától, bár a nagyon nagy eltérések inkább jelhozzárendelési hibával függhetnek össze, és nem azzal, hogy ténylegesen szerkezetileg releváns információt tartalmaznának.

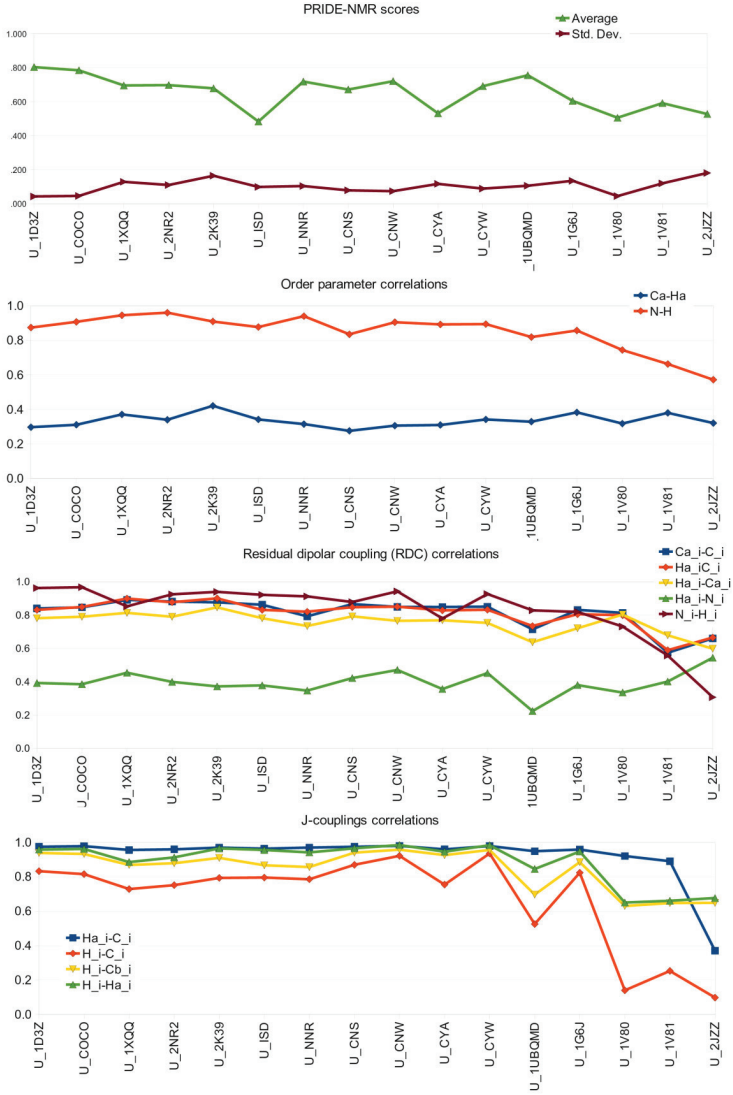
Egyik sokaság se mutat megfelelő egyezést sem a H$\alpha$-C$\alpha$ $S^2$ rendparaméterekkel, sem az elsőrendű közelítésben kapott H$\alpha$-N RDC tenzorokkal. Nem meglepő módon a szilárd fázisú NMR spektroszkópiával meghatározott szerkezetek (U_2JZZ; [61]) értékei jelentősen eltérnek az oldatfázis-beli szerkezetek korrelációitól. Ezt az alacsony PRIDE-NMR értékek is tükrözik. Tehát a CoNSEnsX megközelítés képes szerkezeti eltérések kimutatására akkor is, ha a két szerkezeti sokaság nagyon hasonló.

Az összesített konformerek (U_1D3Z + U_2JZZ; 10+20 konformer) RMSD értéke mindössze 2,42±0,7 Å. Továbbá csak egy integrált elemzés mutatja ki egyértelműen a nagynyomású oldatszerkezet (U_1V81) eltéréseit a légköri nyomáson mért oldatszerkezettel szemben. Az U_NNR sokaság (5.8 b) ábra) jól szerepel mind az NOE-k, az amid NH $S^2$-ek, amid NH RDC-k és néhány ritkán használt paraméter, mind a C$\alpha$ és H$\alpha$ kémiai eltolódások összevetésében. Összességében, a többi dinamikus szerkezeti sokaság (U_1XQQ, U_2NR2, U_2K39) általános jellemzőikben megegyeznek. Az U_1UBQMD sokaság (5.8 c) ábra) elfogadható paraméterekkel rendelkezik, bár valamivel rosszabb az egyezés mint az U_COCO sokaság (5.8 a) ábra) esetében.

Összefoglalva, a humán ubiquitin egy jól meghatározott szerkezettel jellemezhető, ami nagyon különböző körülményekből kiindulva is jól egyező végeredményhez vezet. Ez a tény a molekula belső, inherens rigiditására utal [4].




5.9. ábra. Az ubiquitin (U_NNR sokaság) CoNSEnsX elemzés kimenete




5.10. ábra. Ubiquitin szerkezeti sokaságok és az összeállított kísérleti adatkészlet összevetéseinek eredményei

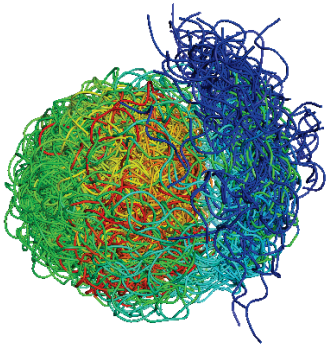
### 5.2.5. Példa a CoNSEnsX szerver használatára:

### PDE 5/6 $\gamma$-alegység, mint rendezetlen fehérje

A fehérjék belső dinamikája az ún. rendezetlen fehérjék felfedezése óta - alig 10 éve - egyre fontosabb szerepet kap. A belsőleg rendezetlen fehérjék esetében nem csupán belső mozgékonyságról van szó, hanem fontos biológiai szerepe van [68].

A rendezetlen fehérjét szabad állapotban csak egy nagyon magas RMSD értékkel bíró szerkezeti sokasággal tudjuk jellemezni. Ilyen esetben az NMR spektroszkópiai adatokból származtatott valós dinamikai paramétereknek történő megfeleltetés fokozott jelentőséggel bír.

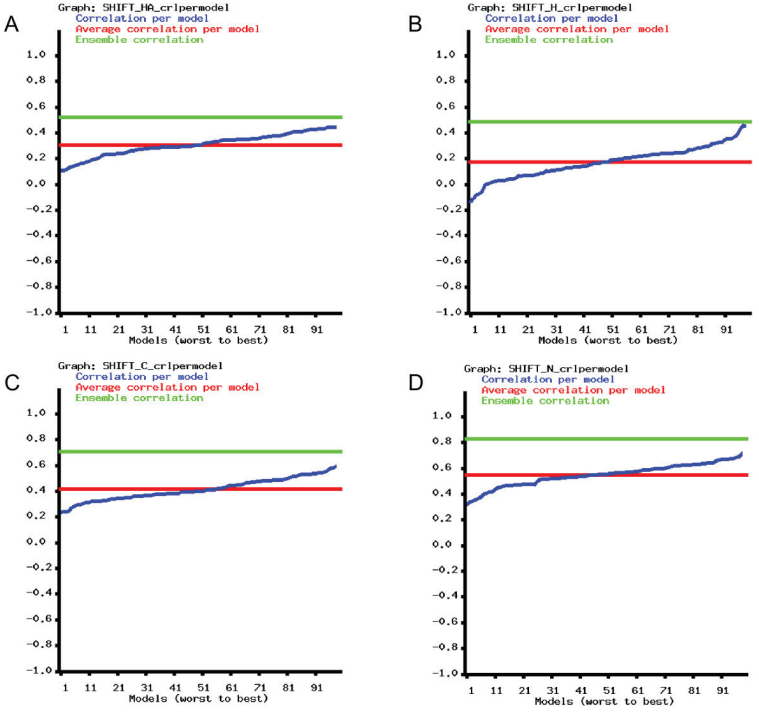
A cGMP foszfodiészteráz (PDE) 5/6 $\gamma$ alegysége egy 87 aminosavas, rendezetlen fehérje; a szerkezete a PDB adatbázisban 2JU4 kód alatt elérhető [69].



5.11. ábra. PDE 5/6 $\gamma$-alegység konformereinek illesztése és ábrázolása

A 100 konformeres szerkezeti sokaságot NOE és PRE[1] kísérleti adatokból származtatott kényszerfeltételek alapján számolták ki. A kapott szerkezeti sokaság nagyon eltérő konformerekből áll: a gerinc RMSD érték 12 Å felett van, ami nagyon magas érték. Jelen esetben kizárólag a PDB adatbázisból elérhető szerkezeti sokaságot vizsgáltam (PDB kód: 2JU4; 5.11 ábra; [69]). A 5.11 ábra a PDE $\gamma$ alegység konformereinek gerinc menti illesztését mutatja be. Az illesztést a teljes fehérjelánchosszra, a gerinc mentén végeztem, a MOLMOL program [63] segítségével.

[1] Paramagnetic Resonance Enhancement




5.12. ábra. PDE 5/6 $\gamma$-alegység korrelációs adatai

A 5.12 ábra összefoglalja a szerver kimeneteként kapott korrelációs eredményeket. Az egyes konformerek korrelációja kékkel, ezek átlaga pirossal, a teljes sokaságra számolt (ensemble) korreláció pedig zölddel van feltüntetve az ábrán. Minden kémiai eltolódásra, amire volt kísérleti adat, a kísérleti és a visszaszámolt értékek közötti korreláció jóval magasabb volt a sokaságra nézve, mint az egyes konformerekre külön-külön. Sőt, a sokaságra számolt korreláció még az egyedi konformerekre kapott legmagasabb korrelációs értéknél is magasabb. A megfigyelés alátámasztja azt a gyakorlatot, hogy rendezetlen fehérjék esetében a teljes, nagyon eltérő konformer-készlettel reprezentáljuk a molekula oldatbeli szerkezeti változatosságát.

### 5.2.6. A CoNSEnsX szerver alkalmazásai

Összefoglalva, a CoNSEnsX szerver hiánypótlónak az NMR-spektroszkópiai adatokból származtatott szerkezetek minőségellenőrzésénél. A szerverben a NOE adatok mellett a fontosabb NMR-alapú dinamikai paramétert megfeleltetjük a teljes szerkezeti sokaságnak. A további szerkezetfinomítást is elősegíti.
A szerver alkalmazási lehetőségeit az alábbi pontokban lehet összefoglalni:

- szerkezeti sokaságok megfeleltetése kísérleti adatoknak: web szerver az NMR adatok és a szerkezeti sokaságok megfeleltetésére;
- PRIDE-NMR megfelelteti a konformerek térszerkezetét a NOE kényszerfeltételeknek (legjobb konformer mint reprezentatív konformer);
- korrelációs adatokat szolgáltat a teljes molekulára, valamint egyes konformerekre visszaszámolt NMR paraméterek és a kísérleti adatok kapcsolatára;

Az eredmények segítségével megbecsülhető, hogy mennyire jó az adott sokaság - mint sokaság - a molekula leírására és hogy mennyire jól tükrözi a vizsgált sokaság az adott időskálájú dinamikát. A CoNSEnsX szerver nem helyettesíti a megszokott szerkezetvalidálást, de segít eldönteni milyen mértékben képes magyarázni a dinamikus szerkezeti sokaság a biológiai folyamatokat.

# 6. fejezet

# Diszkusszió

A dolgozat első felében az NMR spektroszkópiai kísérleti adatokból származtatható szerkezeti információk és a számolt konformerek megfeleltetésére mutattam be két módszert.

A PRIDE-NMR algoritmus segítségével képesek vagyunk a távolság jellegű NOE adatokat ismert fehérjeszerkezetekkel rokonítani. A módszer nem egy fehérjeszerkezet összehasonlító algoritmus, de a teszteléshez mégis hasonló elveket és adatkészleteket használtam. A PRIDE-NMR elemzést minden esetben érdemes több paraméter változtatásával is elvégezni. A háttéradatbázis küszöbértéke, a legkisebb szekvenciális NOE távolság valamint a lánchossz-súlyozás pontosítja az elemzést. A lánchosszbeli eltérések eltüntetése végett vezettem be a lánchossz-súlyozást (5.1 egyenlet). A kísérleti NOE adatok alapján kapott hisztogramok a térszerkezet alapján visszaszámolt $^1$H - $^1$H távolságok csupán 10%-ának felelnek meg. Ez azt jelenti, hogy a szerkezetben lévő közeli proton-proton párok 90%-a "láthatatlan" marad az NMR mérés során, legalábbis ha az NH, H$\alpha$ és H$\beta$ protonokat vizsgáljuk. A láthatalanság oka kétségtelenül összefüggésben van a fehérjék belső, inherens dinamikájával és mozgékonyságával. Ennek ellenére, csupán a NOE adatokból származtatott távolságeloszlások ismerete elegendő a fehérjeszerkezetet azonosítására (ld. 5.1.5 fejezet). Tehát az NMR adatok alapján a fehérjeszerkezet lényeges vonásai kirajzolódnak a $^1$H - $^1$H távolságok segítségével.

A $^1$H - $^1$H NOE adatok atomi szinten adnak információt a fehérje térszerkezetét illetően és segítségükkel jó minőségű, pontos konformerek számolhatók. A PRIDE-NMR algoritmus fejlesztése során éltünk a feltételezéssel, hogy csupán a NOE távolságadatok alapján megbízhatóan lehet becsülni a fehérje szerkezetét adatbázisból, a PRIDE módszer analógiájára. A PRIDE algoritmus [12] C$\alpha$- C$\alpha$ távolságeloszlások elemzésén alapszik. Összesen 28 különböző szekvenciális távolságot[1] használ, 1 Ångstrømös lépésekkel. A

[1] 3-as és 30-as szekvenciális távolságok között

fehérje C$\alpha$ koordinátáit egyszerűen H$\alpha$ koordinátákra lecserélve a PDB állományban gyakorlatilag ugyanazt az eredményt kaptam, mint az eredeti C$\alpha$ koordinátákkal[1]. Viszont a H$\alpha$-H$\alpha$ párok a kísérleti NOE adatokban annyira szórványos távolsági eloszlást eredményeznek, hogy a feltekeredett fehérje térszerkezetéről már gyakorlatilag semmi specifikus információt nem hordoznak magukban. Megpróbáltam további H$\alpha$-H$\alpha$ atomtávolságokat becsülni ugyanarra az atompárra valószínűségi alapon, de a kapott eredmény nagy bizonytalansága miatt ezt a megközelítést is elvetettem. Végül az adatokat egyetlen hisztogramként ábrázoltam a 28 helyett. A szekvenciális távolság függvényében ábrázoltam az összes NOE adatot a különböző atomtípusok szétválasztása nélkül. Felhasználtam az összes NOE kapcsolatot az NH, H$\alpha$ és H$\beta$ atomok között. Ily módon elegendő adat áll rendelkezésre, hogy hasonló lánchosszú fehérjék térszerkezetét megbízhatóan meg tudjuk különböztetni.

A dinamikát tükrözni igyekvő szerkezeti sokaságok eltérő minősége és esszenciája megnehezíti a "reprezentatív konformer" kiválasztását. Általánosan elfogadott nézet, hogy a kísérleti adatoknak leginkább a kiválasztott reprezentatív konformer feleltethető meg leginkább, és így bizonyos értelemben a molekula "átlag-szerkezetének" tekinthető. A PDB adatbázisban sok olyan szerkezet van, ahol annak ellenére, hogy több konformert számoltak a szerkezetmeghatározás sajátosságai miatt, egyetlen ún. átlag szerkezet került be az adatbázisba. Ez a megközelítés szemben áll a több-konformeres molekulaleírással. A reprezentatív konformer egy más típusú kiválasztási módja lehet a legmagasabb PRIDE-NMR értéket kapó konformer kijelölése, mint reprezentatív konformer. A CoNSEnsX szerverbe implementált PRIDE-NMR modul nem egyezik teljesen a PRIDE-NMR szerverrel. A PRIDE-NMR szerver a NOE kényszerfeltétel-lista alapján keres hasonló szerkezetet a háttéradatbázisból. A NOE adatok az NMR spektroszkópiai úton meghatározott fehérjeszerkezetek nagy többségénél rendelkezésre is állnak, és ezek az adatok jól jellemzik a fehérje feltekeredési módját. A PRIDE-NMR módszer az eloszlások összevetését végzi, és az eredményként kapott érték jó mérőszáma a bemenetként kapott kísérleti adatok teljességének és jóságának, amennyiben ismert a vizsgált fehérje térszerkezete. Továbbá a PRIDE-NMR értékek szórása a szerkezeti sokaság heterogenitására is utalhat.

A klasszikus NMR spektroszkópiai fehérjeszerkezet-meghatározás során minden egyes kiszámolt konformer minden kényszerfeltételnek meg kell hogy feleljen a szerkezetfinomítás során. A paraméterek pontatlanságát és bizonytalanságát is tükrözik ezáltal, mivel minden konformerrel összeegyeztethető. Ugyanakkor az a feltételezés, hogy az ilyen típusú szerkezeti sokaságok a fehérje belső dinamikáját is tükrözi, egyáltalán nem szükségszerű. A szerkezetfinomítás célja amellett, hogy a szerkezetek hasonlóak legyenek, a

[1] Az adatok nem képezik jelen dolgozat részét

minél kisebb RMSD érték elérése a szerkezeti sokaságra nézve. Ez a röntgendiffrakciós reprezentatív konformer nézet átvételét jelenti sokszor. Viszont a röntendiffrakciós szerkezetmeghatározás a fehérjekristály diffrakciós képének megoldásán alapul, tehát egyszerre egyetlen konformer szerkezetetét határozzuk meg. Ezzel szemben az NMR spektroszkópiai szerkezetmeghatározás fehérjeoldatból történik, ahol a molekulák mind térben mind időben szerkezetbeli változásokon mennek keresztül, és végeredményben egy konformer-sokaságról kapunk kísérleti adatokat adott időskálán belül. A NOE kényszerfeltételek bizonytalansága előnnyé kovácsolható további, a molekula belső dinamikáját jellemző kényszerfeltételek figyelembevételével. A szerkezetszámolás eredménye ilyen esetben már olyan szerkezeti sokaság, amely variabilitása a molekula kísérleti adatokból származtatott mozgékonyságát tükrözi. Ugyanakkor a molekula belső dinamikája időskála-függő, így fontos szem előtt tartani hogy adott kísérleti adat milyen időskálájú mozgást ír le, és az NMR spektroszkópia időskálájával hogyan egyezik. Minden szerkezeti sokaság bizonyos időskálán jól írja le a molekula belső mozgékonyságát.

A sokaság alapú szerkezetreprezentálás a molekulák belső mozgékonyságából eredő konformációs diverzitást integráló fehérjeszerkezetek új típusú leírása. A molekuláris mozgások nagyon tág időskálája (ps-s) továbbra is alapvető probléma marad. Ugyanis egy adott szekezeti sokaság csak egy adott időintervallumon lehet hivatott leírni a molekula belső dinamikáját. Ugyanakkor az NMR paraméterek nagy részére a mérés csak több nagyságrendű időintervallumra átlagolt értékeket eredményez. Ebből kifolyólag az az elvárás, hogy egyetlen szerkezeti sokaság az összes molekuláris mozgást leírja amit a dinamikai paraméterek tartalmaznak, nem reális, mivel ez csak nagyon nagy számú konformerrel lehetséges. Egy adott szerkezeti sokaság méret mellett egyszerre csak egy kiválasztott paramétert lehet optimálni, ezt is csak azon az áron, hogy a többi NMR paraméterrel való megfeleltetés romlik. A második probléma technikai, nevezetesen hogy egy adott molekulára vagy rendszerre egyszerre tipikusan csak néhány dinamikai paraméterre áll rendelkezésre megbízható, azonos körülményekre mért kísérleti adat. A szerkezetek pontosságának felmérése több típusú adat bevonásával ezért rendkívül fontos. Az NMR mérési technikák fejlődésével újabb és újabb NMR paramétereket vezethetünk be a molekula mozgékonyságát tükröző szerkezeti sokaságok számolásának pontosítására. További paraméterek szükségesek és a szerkezeti sokaságok számolása során alkalmazott protokollok továbbfejlesztése is szükséges, hogy minél pontosabb és precízebb szerkezeti modell álljon rendelkezésre a makromolekulák oldatbeli viselkedésének leírása, és így a biológiai folyamatok leírásának precízebb és pontosabb legyen.

# II. rész

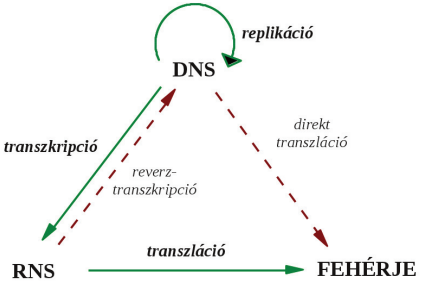
# Véletlenszerű fehérjeszekvenciák *in silico* szerkezetvizsgálata

# 7. fejezet

# Irodalmi áttekintés

## 7.1. A biológiai információáramlás

A fehérjék elsődleges szerkezetét, a fehérjelánc aminosav-sorrendjét, a kódoló DNS szekvenciája határozza meg.

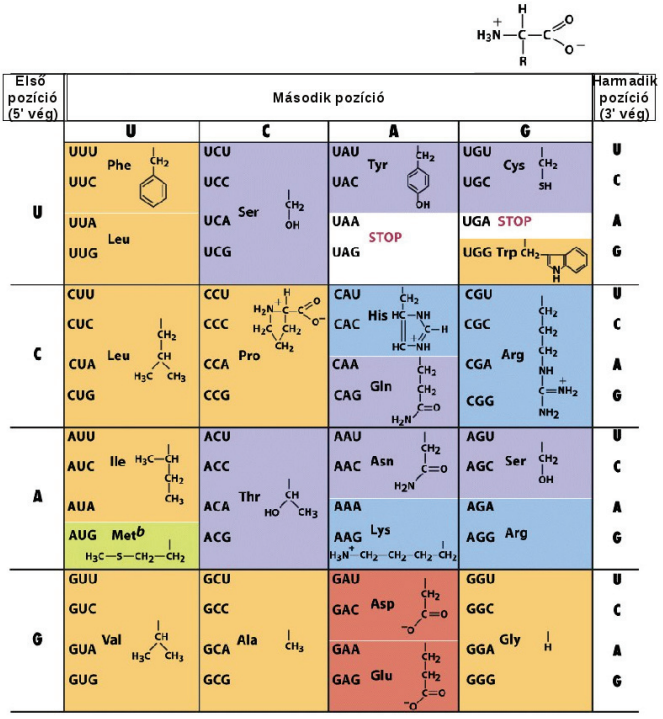



7.1. ábra. A molekuláris biológia centrális dogmája

A centrális dogma kimondja, hogy a genetikai információáramlás a DNS-től a fehérje felé történik és nem fordítva: *DNA makes RNA makes protein*[1]. A hipotézist Francis Crick fogalmazta meg először (7.1 ábra; [70]). A kódoló DNS szál alapján épül fel a messenger, vagy hírvivő RNS (mRNS) és a mRNS bázissorrendje határozza meg a fehérje aminosavsorrendjét (7.1 ábra). Az átírt RNS szekvencia megfelel a DNS szekvenciának. A riboszómával kapcsolaltba lépve elindul a fehérjeszintézis az mRNS kodonok alapján.

[1] A DNS adja az RNS-t, ami adja a fehérjét

Az átíródás az 5’-3’ irányban történik az mRNS-ről, a fehérje az N-terminálistól a C-terminális felé szintetizálódik. Az információ más irányú áramlására vannak példák, de a DNS → RNS → fehérje az alapvető kódolási irány.




7.2. ábra. Standard genetikai kód elrendeződése és a triplettek által kódolt aminosavak kémiai képletei

A DNS négy különböző bázisa kódolja a természetben elsődlegesen előforduló húsz aminosavat. A standard genetikai kód adja meg a kulcsot a helyes aminosav kiválasztására. A standard genetikai kód a mai élőlények számára univerzális [71]; csak minimális eltérések figyelhetők meg a mitokondriális DNS-ben és egyes fajoknál [72]. A kódolás bázistriplettekkel történik, azaz 4*4*4 = 64 különböző kombináció adja vissza a húsz aminosavat (7.2 ábra). A genetikai kód tehát redundáns, mivel több triplet is kódolhatja

ugyanazt az aminosavat. Három triplet, az ún. STOP kodonok, nem kódolnak aminosavat, hanem az átírást befejezését, azaz a szintetizált fehérjelánc végét jelentik. A fehérjeszintézist indító kodon, AUG, egyben a metionint is kódolja. Bármilyen génszakaszt, amit egy START kodon és egy STOP kodon határol be, ORF-nek (Open Reading Frame) nevezzük. Az ORF egy technikai kifejezés, ez nem jelenti hogy a kérdéses DNS szakasz ténylegesen átíródik/fordítódik. A genetikai kód evolúciót vizsgáló kutatók számos általános, az aminosavak fizikokémiai tulajdonságaikkal kapcsolatos megállapításokat tettek [73].

A fehérjék szerkezetéről alkotott klasszikus kép szerint minden fehérjemolekula jól meghatározható, stabil térszerkezettel jellemezhető. A röntgendiffrakcióval vizsgált szerkezetekre ez a módszer sajátosságaiból következik is, mivel egy kristály diffrakciós képéből épül fel a molekula. A PDB-beli szerkezetek túlnyomó többségét így határozták meg. Feltételezzük, hogy az így kapott modell összevethető a biológia körülmények között létező molekuláéval. Oldatbeli vizsgálatok esetén, NMR spektroszkópiai módszerekkel, összetettebb képet kapunk [22]. Dunker a natív fehérjék szerkezetét három kategóriába sorolja: a rendezett, globuláris fehérjék, a "molten globule" vagy "olvadt gombóc" állapot, és a random coil, vagy rendezetlen fehérjék [74].

## 7.2. Belsőleg rendezetlen fehérjék

Az ún. belsőleg rendezetlen fehérjék (Intrinsically Disordered Proteins, IDP-k) tanulmányozása új tudományágnak nőtte ki magát [75, 76]. Uversky vezette be a belsőleg rendezetlen fehérjékre, mint a proteom egyik alkotó részére, az "unfoldome" kifejezést [77]. Aminosav-összetételük jelentősen eltér a globuláris fehérjéektől [68, 78]. A rendezetlen fehérjék felfedezése, felismerése teljesen átírta a korábban alkotott képünket a fehérjékről. A rendezetlen fehérjék szabad állapotban nem jellemezhetők egyetlen konformerrel, azaz a molekula szerkezetét bemutató térbeli strukturával. Ennek ellenére van biológiai aktivitásuk és esetenként partnerhez kötődve akár rendeződhetnek is [79]. Egyes becslések szerint a részlegesen vagy teljesen rendezetlen fehérjék aránya nagyobb, mint 20% az élesztő (Yeast) és meghaladja a 30%-ot az egér proteomban [80]. A rendezetlen fehérjék aránya a proteomban növekszik az organizmus komplexitásával (baktérium → archea → eukarióta). Az IDP-k funkcionális szempontból evolúciós előnyt jelentenek a globuláris fehérjékkel szemben bizonyos biológiai folyamatok során [81].

## 7.3. Transzmembrán fehérjék

A rendezett, globuláris fehérjék egy különleges csoporja a membrán és transzmembrán fehérjék. A transzmembrán fehérjedomén olyan alegység, amely a membránba ágyazódva termodinamikaiag stabil térszerkezetet vesz fel. A transzmembrán fehérjék lehetnek helikálisak vagy $\beta$-hordó jellegűek. Általánosan ezek a szakaszok 20-30 aminosava tesznek ki. A transzmembrán fehérjék aminosav-összetétele is eltér a globulárisektól, a töltött és hidrofób aminosavak nagyobb súlyban vannak. Mivel a sejtmembrán egyik oldalán vagy a membránba ágyazódva helyezkednek el, és a globuláris fehérjékhez képest más, oldatban tiltott szerkezeti elrendeződésre is lehetőség van [82]. A membránfehérjék teszik lehetővé a szabályozott sejtkommunikációt a külvilággal, valamint a szabályozott anyagáramlást is. Ahram és munkatársai predikciós algoritmusok segítségével a transzmembrán szegmenst tartalmazó fehérjék arányát a humán proteomban 15 és 35% közöttinek becsülték, konszenzus predikcióval 13%-ra [83].

## 7.4. Amiloid fehérjék és a fehérjeaggregáció

Stabilitási és energetikai szempontok alapján azonban a $\beta$-redőzött réteg, azaz az amiloid állapot a tekinthető a legstabilabb konformációnak minden polipeptid számára, aminosav-összetétel és -sorrendtől függetlenül [84, 85]. A globuláris fehérjék natív állapotban is képesek amiloidképzésre [86] Egyre szélesebb körben elfogadott tény tehát, hogy a fehérjeaggregáció inherens, belső tulajdonsága minden polipeptidláncnak szekvenciától függetlenül és az amiloid fibril minden lehetséges fehérje legkedvezőbb termodinamikai állapotát takarja [85]. Az elképzelés, hogy az aggregációs hajlam belső tulajdonsága minden polipeptidláncnak a fehérjeevolúciót nagy kihívás elé állítja [9, 87]. Baldwin és munkatársai eredményei szerint az amiloid szerkezet fiziológiás környezetben is a legstabilabb állapot, így a fehérjék funkcionális formái nem a szabadenergia globális minimumának megfelelő szerkezetek, hanem konformációs és kémiai értelemben is metastabilak és az evolúciós és szelekciós nyomások egyensúlya tartja meg [88]. Számos tény támasztja alá, hogy a fehérjeevolúció során az aggregációs hajlam csökkentése és az aggregáció megakadályozása fontos szelekciós nyomás. Részletes tanulmányok több ilyen mechanizmust mutatnak be [89, 90]. Rousseau és munkatársai tanulmánya kihangsúlyozta, hogy a fehérjék rendezetlenségi és az aggregációs hajlama negatívan korrelál egymással [91, 92]. Továbbá az aggregációs hajlam az organizmus komplexitásával is antikorrelál [93]. Monsellier és munkatársai a humán proteomot elemezték az aggregációs hajlam szempontjából [94]. Azt találták,hogy a IDP-k, transzmembrán és a globuláris fehérjék aggregációs hajlama nagyon eltér.

## 7.5. Teljesen új fehérjék keletkezése

Többféle új gén keletkezési mechanizmusa ismert: exon ugrálás, génduplikáció, stb [95]. Sokáig azt gondolták a kutatók, hogy *de novo*, azaz teljesen újszerű fehéje spontán keletkezése gyakorlatilag sohasem fordul elő biológiai körülménynek között. Bornberg-Bauer és munkatársai az új fehérjék keletkezési mechamizmusait elemzik, a szerkezetek közötti átmenetek lehetőségeit kiemelve [96]. Ezzel szemben az elmúlt pár évben humán [97], főemlős [98] és drosophila [10] genomok elemzésével mindhárom esetben alá tudták támasztani a kutatók, hogy adott fehérje *de novo* keletkezett, azaz a közeli fajok proteomjaiban nincs azonosítható homológja. Li és munkatársai az agy szöveteiből izoláltak humán *de novo* fehérjét [99]. Wu és munkatársai 60 fehérjekódoló gént azonosítottak az embernél, ami a csimpánzban nem fejeződik ki [100]. A *de novo* fehérjék a nem-kódoló DNS szakaszok spontán átíródásása során is keletkeznek [99, 101, 102]. A *de novo* fehérjék egyben árva, vagy "orphan" fehérjék is, definíció szerint. Az árva fehérje olyan fehérje, amelynek nincs ismert homológja adott proteomon vagy más fajok proteomján belül. A humán genom alacsony szintű átíródási aktivitása is alátámasztja ennek lehetőségét [103]. Továbbá a teljesen új fehérjék keletkezése a fajok alkalmazkodóképességét is növeli, tehát mindenképp evolúciós előnyt is jelenthet [104]. Tehát az evolúció során teljesen új, humán-specifikus fehérje keletkezett, ami semmilyen korábbira nem hasonlít, vagyis *de novo* fehérje. A *de novo* keletkezést nehéz bizonyítani; az irodalomban alig néhány példa van *de novo* fehérjére.

| Randomizált fehérjeszekvenciák elemzése | | | |
|---|---|---|---|
| aminosav-készlet | randomizálás elve | Kísérlet | Eredmény |
| QLR [105] | (50%Q,40%L, 10%R) | *in vitro* | 80-100aa; 30-70% helikális; oldhatatlan |
| VADEG [106] | fág-bemutatás | *in vitro* | változó lánchossz; nincs 3D; 100% oldható |
| 20 aa [107] | fág-bemutatás | *in vitro* | $1/10^{11}$ funkcionális fehérje |
| 20 aa [108] | fág-bemutatás | *in vitro* | 141aa; 20% oldható |
| 20 aa [109] | fág-bemutatás | *in vitro* | 20% oldható nincs 3D |
| 20 aa [110] | fág-bemutatás | *in vitro* | 50aa; 3D; nincs homológ |
| 20 aa [111] | 5% gyakoriság /aa | *in silico* | 70aa; globuláris, helikális; oldható |

7.1. táblázat. Irodalomban leírt randomizált fehérjeszekvenciák kísérleteinek összefoglalása

A fehérjék feltekeredése kapcsán továbbra is nyitott a kérdés, hogy evolúciós, szelekciós nyomás hatására vagy a biológiai környezet hatására alakul ki az adott szerkezet.

Schaefer és munkatársai *in silico* mutálták a fehérjéket és a másodlagos szerkezetek és a rendezetlenség változását vizsgálták. Meglepő módon azt találták, hogy a másodlagos szerkezeti elemek konzerváltabbak, mint a szekvencia rendezetlensége [112]. A fehérjék lehetséges szekvenciaterét lehetetlen szisztematikusan bejárni, mivel egy száz aminosavas szekvencia lehetséges kombinációi $20^{100} \simeq 1.3*10^{130}$, és ehhez viszonyítva a világegyetem összes atomjainak számát $10^{80}$-ra becsülik. A kutatók újabb technikákat fejlesztettek ki egy adott fehérje lehetséges variációi körbejárására. A véletleszerű szekvenciák *in vitro* előállítására a fág-bemutatásos ("phage display") technikát alkalmazzák [113]. A randomizáció nukleotid szinten történik, csak meghatározott arányú és bázis-összetételű kodonokból épülhet fel a keletkező fehérje. A 7.1 táblázat összefoglal néhány, az irodalomban leírt sikeres kísérletet random fehérjeszekvenciák elemzésére [114]. A számos eredmény ellentmondani látszik egymásnak, kiemelve a terület nehézségeit és kihívásait.

# 8. fejezet

# Célkitűzések

A dolgozat második részében a fehérjék dinamikáját evolúciós távlatba kívántam helyezni, a rendezetlenségen keresztül. Munkámban a *de novo* fehérjék keletkezését modellezve, *in silico* strukturális elemzést kívántam végezni, annak eldöntése végett, hogy az újonnan keletkező fehérjékre nézve az aggregációs veszély tényleg jelentős-e.
A kidolgozást a következő lépésekben kívántam megvalósítani:

1. Véletlenszerű fehérjeszekvenciák szisztematikus generálása a szekvenciatér lehető legtágabb lefedése érdekében.

2. A véletlenszerű szekvenciák strukturális elemzésére teljes körű elemzési protokoll felépítése szekvencia-alapú strukturális prediktorok implementálásával.

3. Az elemzési protokoll robusztusságának vizsgálata: a véletlenszerű szekvenciákra kapott eredmények összevetése a valós, funkcionális fehérjékre kapott elemzési eredményekkel.

# 9. fejezet

# Az alkalmazott módszerek

## 9.1. Használt adatbázisok és adatkészletek

A valós, mai fehérjék reprezentálására különböző adatbázisokat használtam. A globuláris fehérjéket az ASTRAL40 [115], a rendezetlen fehérjéket a DISPROT [116], a transzmembrán fehérjéket pedig a PDBTM adatbázis [117] elemzésével jellemeztem. Az amiloid-aggregáló fehérjék reprezentálására az AmyPDB adatbázist [118] használtam, noha ez az adatbázis inkább amiloidképződésre hajlamos, rendezetlen fehérjéket tartalmaz. Két teljes - humán és egér - proteomot is elemeztem. Az ASTRAL40 [115] a 40%-os homológiaszűrt ismert globuláris fehérjék adatbázisa, és a SCOP [38] adatbázisból van származtatva. Az AmyPDB [118] az amiloid-prekurzor fehérjecsaládok és az ezekre jellemző fehérjeszekvenciák gyűjtőhelye. A fehérjeszekvenciákat az UniProt 6.1-es verziós adatbázisból vették át a szerzők. Az adatbázist utoljára 2008. április 7.-én frissítették. A DISPROT [116] a kísérletileg igazoltan részben vagy teljesen belsőleg rendezetlen fehérjék adatbázisa. Minden fehérjére a rendezetlen szegmens pontos helye van feltüntetve. A PDBTM adatbázis [117] a kísérletileg igazoltan transzmembrán hélix vagy $\beta$-hordó fehérjék gyűjteménye. A teljes humán és egér proteomot az UniProt [72] adatbázisból töltöttem le.

Az adatbázisokat egységesen homológia-szűrésnek vetettem alá. A CD-HIT program [119] segítségével, 0.7-es küszöbértéket használva, ami 70%-os homológiaszűrésnek felel meg. A használt algoritmusok minimális szekvenciahossz követelményei miatt a vizsgálatot leszűkítettem a legalább 30 aminosavas, de legfeljebb 1000 aminosavas szekvenciákra. A 9.1 táblázat összefoglalja az így kapott adatkészletek főbb jellemzőit. A humán árva, vagy ún. "orphan" fehérjék[1] esetében az az UniProt adatbázisból [72] a

[1] Az árva fehérjék olyan fehérjeszekvenciák, amelyekre nincs ismert homológ más faj vagy organizmusban sem.

| Adatbázis | használt verzió | szekvenciák száma | 70% szűrt | átlagos lánchossz |
|---|---|---|---|---|
| ASTRAL40 | 1.75 | 10569 | 10175 | 173,80 ± 110,66 |
| AmyPDB | összes szekvencia | 1687 | 247 | 267,22 ± 223,47 |
| DISPROT | v.5.7 | 684 | 529 | 347,95 ± 227,30 |
| PDBTM | all.seq (v.2.3) | 4550 | 429 | 287,36 ± 197,42 |
| Humán proteom | UniProt 2011_05 | 51265 | 20899 | 375,09 ± 231,66 |
| Egér proteom | UniProt 2011_05 | 45826 | 18525 | 396,41 ± 231,56 |

9.1. táblázat. A használt adatbázisok összetétele és jellemzői

megfelelő kódoló mRNS szekvenciákat töltöttem le. Az mRNS szekvenciákat a standard genetikai kód alapján lefordítottam a megfelelő olvasókeretben (frame). A fehérjeszekvenciák meglétét a lefordított fehérjeláncban ellenőriztem. Mind a teljes letöltött mRNS szekvenciára (GC% exon) mint a csak fehérjét kódoló mRNS szekvenciára (GC% mRNS) kiszámoltam a GC-tartalmat.

## 9.2. Random fehérjeszekvenciák generálása

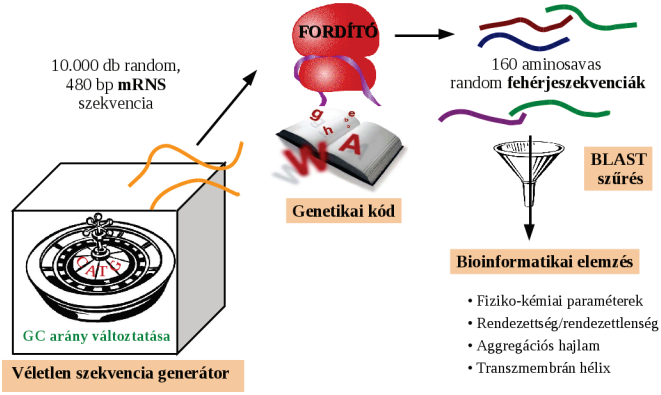
### 9.2.1. Teljesen random mRNS szekvenciák

Random nukleotidszekvenciákat generáltam különböző GC-tartalmak mellett, 10%-tól 90%-os GC-tartalommal, 10%-onként. A 10%-os GC-tartalom 5% guanint, 5% citozint, 45% adenint és 45% timint (uracilt) jelent. A GC-tartalom növelése során feltételeztük, hogy a mintavételezésünk a lehetőségekhez mérten egyenletes lesz a lehetséges fehérjeszekvenciák terében. A DNS és az RNS összetétele egyedül a T→U cserében különbözik. Az elemzésben ennek a különbségnek nincs jelentősége, egységesen timint használtam.

$$\text{GC-tartalom} = \frac{\text{G} + \text{C}}{\text{A} + \text{T} + \text{G} + \text{C}} \times 100 \qquad \text{A} = \text{T} \; ; \; \text{G} = \text{C} \tag{9.1}$$

Láncon belüli STOP kodon előfordulás esetén új kodontriplettet generáltam, hogy egyenlő lánchosszú random fehérjeszekvenciákat kapjak. Az átlagos doménméret a CATH adatbázisban 153 aminosav [41]. Shen és munkatársai modellezték az optimális globuláris fehérjeméretetet 156 aminosavra becsülték [120]. Minden nukleotidszekvenciát tehát 480 bázis hosszúra terveztem, hogy 160 aminosavas fehérjeszekvenciákat kapjak.

A kódoló random DNS szekvenciákat a standard genetikai kód alapján fehérjeszekvenciára fordítottam. A továbbiakban az így kapott 160 aminosavas, GC-tartalmanként




9.1. ábra. Random fehréjeszekvenciák generálása a biológiai folyamatot modellezve

10.000 szekvenciás adatkészleteket elemeztem random szekvenciaként. Az 9.1 ábra összefoglalja a véletlenszerű fehérjeszekvenciák *in silico* előállítását.

A fent leírt randomizációs módszerrel a biológiailag esetlegesen releváns fehérjeszekvenciák terénél nagyobb teret lehet bejárni. A keresés semmi esetre sem teljes; az adatkészlet random mintavételnek tekinthető. A fehérjeszekvenciák előállítási módja modellezi *in silico* a transzkripció és a transzláció biológiai folyamatát.

### 9.2.2. Proteomok randomizálása

Az adatkészletek randomizálására az irodalomban általánosan ismert módját választottam. Az adatkészletben a fehérjék lánchosszait valamint az egyes szekvenciák aminosav-összetételét megtartva permutáltam az aminosavakat. Az egyes fehérjéknél a végrehajtott permutációk száma megegyezett a fehérje lánchosszával. Ezzel a randomizálási technikával előállítottam egy random humán és egy random egér proteom adatkészletet. A randomizálást saját PERL programmal végeztem, a *shuffled* beépített modul használatával.

## 9.3. Homológiaszűrés, fiziko-kémiai elemzés

A teljesen random szekvenciákat a BLAST program [121] 2.2.23-as verzió segítségével elemeztem. A BLAST választ ad arra a kérdésre, hogy egy adott szekvenciához rendelhető-e hasonló, ismert szekvencia. A hasonlóságot a szekvenciák összevetése alapján mérjük. Alapbeállítások mellett fehérje-fehérje keresést végeztem, az alábbi parancsot használva: blastall -p blastp -d nr
ahol a *blastp* a standard fehérje BLAST programot takarja, ami a fehérjeszekvenciák azonosítására és fehérje-adatbázisbeli homológiakeresésre lett optimálva. Az nr, azaz "non-redundant" adatbázis az összes ismert adatbázisbeli fehérjeszekvenciát tartalmazza[1] Alapbeállított (10-es) E-érték mellett futtattam a programot. A BLAST programot asztali gépen futtattam. A kiértékelést saját PERL programokkal végeztem, E=10 és E=$10^{-3}$ értékekre. Az E-érték ("expectation value", E) azt hivatott mérni, hogy egy random modell esetén [122] hány véletlenszerű egyezés fordul elő. 10-es E-érték az jelenti, hogy 10 találatot inkább a szerencsének köszönhetünk. E=$10^{-3}$ esetén a találat már egyedinek tekinthető, nem hiba vagy véletlennek köszönhető. Jelen esetben az E-értéket alapbeállításon hagytam, hogy minden lehetséges találatot megkapjak, mivel a generált véletlenszerű szekvenciák egyediségét szeretném vizsgálni.

A fehérjeszekvenciák fontos jellemzője az aminosav-összetétel. Minden adatkészletre kiszámoltam az átlagos aminosav-összetételeket. Az adatkészletek aminosav-összetételét is elemeztem. Kiszámoltam az egyes aminosavak előfordulási gyakoriságát a teljes adatkészletre nézve mind a mai fehérjék, mind a random szekvenciák esetében.

$$F_i = \frac{N_i}{N_{\text{total}}} \tag{9.2}$$

ahol $F_i$ az i-edik aminosav előfordulási gyakorisága, $N_i$ az i-edik aminosav előfordulása az adatkészletben, $N_{total}$ az összes aminosav darabszáma a teljes adatkészletben.

További jellemzők a hidrofobicitás és a nettó töltés. Ezt a két paramétert több módon is ki lehet számolni. Jelen munkában az Uversky által bevezetett ún. "charge-hydropathy", vagy "töltés-hidropátia" ábrázolásához használt számolási módot alkalmaztam. Uversky és munkatársai a töltés-hidropátia plot segítségével empírikus úton különböztetik meg a rendezett és rendezetlen fehérjéket.

$$< \mathrm{C} > = 2,785 \times < \mathrm{H} > - 1,185 \tag{9.3}$$

[1] Nem-redundáns GenBank CDS fordítások + PDB + SwissProt + PIR + PRF, az env_nr adatait kizárva

ahol <C> az adott szekvencia átlagos nettó töltése, <H> pedig a normalizált átlagos hidrofobicitás érték. Az egyenes bal oldalán (alacsony hidrofobicitás, magas nettó töltés) lévő szekvenciák várhatóan rendezetlenek, a másik oldalra esők pedig várhatóan globulárisok lesznek.

A hidorfobicitást a normalizált Kyte-Doolittle skála [123] alapján számoltam ki. A szekvencia átlagos hidrofobicitási értékét úgy kaptam meg, hogy az egyes aminosavakra megadott normalizált hidrofobicitási értékek összegét elosztottam az aminosavszámmal. Az adatbázis átlagos hidrofobicitási értéket úgy lehet kiszámítani, hogy az egyes szevenciák átlagos hidrofobicitási értékeinek összegét elosztjuk az összes szekvencia számával. A fehérjeszekvencia átlagos hidrofobicitása valamint nettó töltése fontos információkkal szolgál a szekvencia térszerkezetét illetően [124]. Kiszámoltam minden egyes szekvenciára a normalizált hidrofobicitási értékét a Kyte-Doolittle skála alapján.

$$< \mathrm{H} > \; = \; \frac{\sum \mathrm{KD}_{\mathrm{aa}}^{\mathrm{norm}}}{\sum \mathrm{aminosav}} \tag{9.4}$$

Egy adott szekvencia teljes nettó töltést jelen esetben egyszerűen a töltött aminosavak előjeles összegeként definiáltam. Az átlagos nettó töltést (<C>) pedig úgy kapm meg, hogy a teljes nettó töltést elosztom az összes aminosav számával. A szekvencia nettó töltését az alábbi képlet szerint számoltam:

$$< \mathrm{C} > \; = \; \frac{\sum (\mathrm{K+R}) - \sum (\mathrm{D+E})}{\sum \mathrm{aminosav}} \tag{9.5}$$

## 9.4. Szekvencia alapú predikciós módszerek elméleti háttere és alkalmazása

Minden szerkezeti struktúra leírására legalább három, lehetőleg különböző elméleti megfontolásokon alapuló, kizárólag a fehérjeszekvenciát felhasználó predikciós algoritmust használtam. A legtöbb szabadon hozzáférhető kutatási célokra.

A rendezetlenségre, a transzmembrán hélix-képzésre és az aggregációra való hajlamot minden esetben legalább három különböző szekvencia-alapú algoritmussal becsültem. Egyik predikciós módszer sem használ hasonlóságbeli vagy evolúciós információt, ellentétben a mai leghatékonyabb másodlagos szerkezet predikciós algoritmusokkal [125]. A $\beta$-hordókat mint transzmembrán elemeket jelen munkában nem vettem figyelembe, mert nem találtam elérhető, aminosav szinten feldolgozható szekvencia-alapú predikciós módszert.

*De novo* szekvenciák szerkezeti jellemzéséhez szükséges, hogy minden egyéb információt kizárjunk, mert csak így érhető el, hogy az eredményeket ne torzítsák a valós, ismert fehérjék tulajdonságai. Az elemzésben használt konszenzus predikciók *de novo* fehérjék esetében is alkalmazhatók, és evolúciós prekoncepció nem torzítja az eredményt. Továbbá az elemzést olyan predikciós algoritmusokkal végeztem, amelyeknél aminosav szintű kiértékelés lehetséges. A kiértékelés a szekvenciánkénti aminosavszázalékokra korlátozódik, azaz az adott tulajdonsággal rendelkező aminosavak összegét feltüntettem. Az alábbiakban röviden összefoglalom az algoritmusok elméleti hátterét.

### 9.4.1. Rendezetlenség prediktorok

A rendezetlenséget három, eltérő elveken alapuló módszerrel jósoltam.

Az IUPred [126] algoritmus azt a jelenséget használja fel, hogy a rendezett régiók a rendezetlenhez képest kevesebb stabilizáló aminosav-aminosav kölcsönhatást tartalmaznak és PSSM módszerrel az aminosavpárok páronkénti energiáját számolja. Ennek felhasználásával a szekvencián belül a rendezetlen régiók elkülöníthetők a rendezettől. A RONN (Regional Order Neural Network [127]) algoritmus neurális hálózatot használ. Az algoritmus összeveti a kereső szekvenciát ismert feltekeredési állapotú fehérjeszekvencia sorozattal (rendezett, rendezetlen, kevert állapotok). Az illesztés eredményeként kapott mérőszám alapján a program rendezettként vagy rendezetlenként osztályozza az adott szegmenst a neurális hálózat segítségével. A VSL2B [128] a VSL2 program egyik variánsa. Két predikciós szintje a van az algoritmusnak. Az első szinten két prediktor van, egyik a rövid, másik a hosszú - több mint 30 aminosavas - rendezetlen (IUP) szakaszok felismerésére. A második szinten a két predikció eredményeit dolgozza fel egy harmadik prediktorként, amely a végső eredményt adja a szekvencia rendezetlenségének mértékére. A predikció minden egyes aminosavra ad becslést.

Az elemzések során mindhárom rendezetlenség prediktort alkalmaztam és a három predikció átlagával jellemeztem a szekvencia konszenzus rendezetlenségi hajlamát:

$$\text{konszenzus D \%} = \frac{\sum \text{IUPred} + \sum \text{RONN} + \sum \text{VSL2B}}{3 \times \sum \text{szekvencia}} * 100 \qquad (9.6)$$

### 9.4.2. Transzmembrán prediktorok

A HMMTOP [129] és a TMHMM [130] programok a Hidden Markov Model (HMM) algoritmus alapján vannak felépítve. A HMMTOP kategóriákba sorol minden egyes amino-

savat: sejtoldali, membránon belüli (i, I), sejten kívüli avagy membránon kívüli (o, O), és a transzmembrán részek (H). A kategorizálásnak köszönhetően nem csak a transzmembrán régiók azonosíthatók, hanem azok geometriáját és elhelyezkedését is jellemezzük. A TMHMM program, szintén HMM aloritmus alapú, nem szolgáltat aminosav-szintű predikciót, de megadja a transzmembrán (TM) szegmensekben lévő összes aminosav számát, valamint a TM szegmensek átlagos hosszát. A DASTMfilter [131] a "Dense Alignment Surface" algoritmuson alapszik. A hosszú kimenet grafikusan is értelmezhető, de minden fontosabb paraméter szekvenciára bontva van megadva.

Mindhárom prediktort hasonlóan megbízhatónak tekintettem, és a három predikció átlagával jellemeztem a szekvencia konszenzus transzmembrán hélixképző tendenciáját.

$$\text{konszenzus T \%} = \frac{\sum \text{HMMTOP} + \sum \text{DASTMfilter} + \sum \text{TMHMM}}{3 \times \sum \text{szekvencia}} * 100 \qquad (9.7)$$

### 9.4.3. Aggregáció és amiloid prediktorok

A TANGO [132] különböző tendenciák összegzése alapján ad becslést az adott szekvenciarészlet aggregációs hajlamára. Az algoritmus különösen a rendezetlen szekvenciák $\beta$-aggregációs hajlamának előrejelzésére lett optimálva. A WALTZ [133] a TANGO-val szemben az amiloid tendenciájú régiók felismerésére van optimálva. A szerzők szerint a TANGO és a WALTZ kiegészítik egymást a szekvencia teljes aggregációs hajlam előrejelzésének tekintetében. A FoldAmyloid [134] a fentiektől szemben eltérő módon, a "packing density", avagy a molekula tömörsége alapján becsüli a szekvencia aggregációs hajlamát. Több küszöbértéket is lehet használni.

A konszenzus aggregációs hajlamo két részből tevődik össze. A TANGO és a WALTZ egymást kiegészítik, így ezt a két predikciót egyben kezeltem, és az átlagukat a FoldAmyloid eredményeivel kombináltam az alábbiak szerint:

$$\text{konszenzus A \%} = \frac{\sum \text{FoldAmyloid} + \dfrac{\sum \text{TANGO} + \sum \text{WALTZ}}{2 \times \sum \text{szekvencia}}}{2 \times \sum \text{szekvencia}} * 100 \qquad (9.8)$$

Az alkalmazott konkrét küszöbértékeket a 9.2 táblázat foglalja össze.

| Predikció | Módszer | aminosav cutoff | szekvencia cutoff |
|---|---|---|---|
| rendezetlenség | IUPred [126] | >0,5 | min 1 db 30 aminosavas szegmens |
| | RONN [127] | >0,5 | min 1 db 30 aminosavas szegmens |
| | VSL2B [128] | >0,5 | min 1 db 30 aminosavas szegmens |
| transzmembrán | HMMTOP [129] | adott | min 1 db 17 aminosavas szegmens |
| | DASTMfilter [131] | >2,5 | min 1 db 21 aminosavas szegmens |
| | TMHMM [130] | adott | min 1 db szegmens (No of TMHs) |
| aggregáció | FoldAmyloid [134] | >21,4 | min 1 db 5 aminosavas szegmens |
| | TANGO [132] | >5% | min 1 db 5 aminosavas szegmens |
| | WALTZ [133] | >5% | min 1 db 6 aminosavas szegmens |

9.2. táblázat. A predikciók kiértékelésére alkalmazott küszöbértékek

A szekvencia alapú prediktorokat úgy választottam ki, hogy aminosav-szinten adjanak eredményt. Ily módon a lokális eltérések elemzésére is lehetőség van. Minden szekvenciára meghatároztam az átlagos "pozitív" aminosav-számot, a szegmensek darabszámát és az átlagos szegmenshosszakat a 9.2 táblázatban összefoglalt küszöbértékek értelmében. Minden adatkészletre kiszámoltam ezen felül a normalizált szegmensszámot (szegmens /100 aminosav) és a normalizált szegmenshosszt. Ezeket az adatokat a dolgozatban nem tárgyalom részletesen, a továbbiakban az adatbázisra számolt százalékos "pozitív" aminosavak adataival dolgoztam. Minden strukturális tulajdonságot a konszenzus predikcióban kapott szekvencia aminosav százalékaként definiáltam.

## 9.5. A statisztikai elemzés módszerei

A statisztikai kiértékelés több lépésben történt. A Pearson-féle korrelációs együttható (r) a két minta közötti kapcsolat szorosságának mutatója. Értéke -1 és +1 közé eshet. Ha az érték nullához van közel, a kapcsolat nagyon gyengének számít; 0.75 felett már erős szochasztikus kapcsolattal jellemezhetjük a két adatsort. Negatív előjellel antikorrelációról, pozitív előjellel pozitív korrelációról beszélünk.

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{9.9}$$

Két eloszlás egyezését egy- és kétdimenziós Kolomorov-Smirnov (K-S) teszttel lehet vizsgálni [135]. A minták darabszáma nem kell hogy egyezzen, mivel nemparametrikus statisztikai tesztről van szó. A Kolomorov-Smirnov próba két minta tapasztalati eloszlásfüggvényét veti össze és a tapasztalt eltérésből számol próbastatisztikát. A kumulatív eloszlás $S_N(x)$ az adatpontokhoz tartozó számok függvénye. A Kolmogorov-Smirnov D érték definíció szerint két kumulatív eloszlásfüggvény abszolút különbségének maximális értéke. Két eltérő kumulatív eloszlásfüggvény - $S_{N_1}(x)$ és $S_{N_2}(x)$ - összehasonítására a K-S statisztika a következő:

$$D = \max_{-\infty<x<+\infty} \left| S_{N_1}(x) - S_{N_2}(x) \right| \tag{9.10}$$

A szignifikanciát a 9.11 egyenlettel lehet leírni; $Q_{KS}(\lambda)$ monoton függvény, $Q_{KS}(0)=1$ és $Q_{KS}(\infty)=0$ szélsőértékekkel.

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \tag{9.11}$$

A nullhipotézis (H0) szerint a két eloszlás ugyanaz. H0-t elfogadjuk, ha $P(D>obs)<\alpha$, $(1-\alpha)$ szignifikancia-szint mellett.

$$P(D > obs) = Q_{KS}\left(\left[\sqrt{N_e} + 0,12 + \frac{0,11}{\sqrt{N_e}}\right]D\right) \tag{9.12}$$

ahol $N_e$ az adatpontok effektív száma. Egydimenziós K-S teszt esetében $N_e = N$, kétdimenziós esetben pedig:

$$N_e = \frac{N_1\, N_2}{N_1 + N_1} \tag{9.13}$$

ahol $N_1$ az első, $N_2$ a második eloszlás adatpontszáma.

A predikciós eredménynek további kiértékelésére az átfedő területeket is meghatároztuk két- és háromdimenzióban. Az átfedő területek meghatározását saját programmal végeztük. A teret 0,1 egységnyi felbontású ráccsal fedtük le. Ha adott rácsponton belül találtunk adatpontot, a rácsegységet hozzáadtuk az összterülethez. Ily módon pásztáztuk a teret minden egyes két és háromdimenziós adatkészletre. Biztosítottuk, hogy a rács az x és y tengelyek mentén végig folytonos legyen, azaz hogy az x és y irányban a kitöltött rácspontok ne legyenek megszakítva az x és y koordináták mentén. A kapott területetek nagyságát és az átfedéseket összegeztük.

Mindhárom statisztikai elemzést C++ programmal végeztük el.

# 10. fejezet

# Eredmények

## 10.1. Véletlenszerű szekvenciák álatlános jellemzése

A humán genom átlagos GC-tartalma 41% [136]. A genomok GC-tartalma 20% és 60% között változik általánosan. Extrém példaként, a Streptomyces coelicolor A3(2) GC-tartalma 72%, ezzel szemben a Plasmodium falciparum GC-tartalma 20% körüli. A genomok elérhetők az NCBI honlapjáról[1] Továbbá megfigyelték, hogy a ténylegesen fehérjét kódoló génszakaszok GC-tartalma magasabb, mint a "háttér" genomé [137].

A teljes GC-tartomány lépésenkénti lefedésével tehát egy nagyobb szekvenciateret járunk be, mint ami biológiailag releváns és így általános trendek azonosítására alkalmas adatkészlet áll rendelkezésre. Az mRNS szinten randomizált 9*10.000 fehérjeszekvenciát adatkészleteket egységesen 160 aminosavasra terveztem, az átlagosan elfogadott doménméretnek megfelelően [45]. Az így kapott fehérjéket BLAST keresésnek vetettem alá. A BLAST algoritmus az összes ismert fehérjeszekvenciával veti össze a célszekvenciát és jelzi a szekvenciabeli hasonlóságot. A szekvenciális rokonságot E=$10^{-10}$ érték alatt tekintjük nem véletlenszerű egyezésnek. A 90.000 véletlenszerű fehérjeszekvencia között E=$10^{-10}$ alatt egyetlen szekvenciára sem találtam hasonlót, és csak 30 szekvencia esetében van hasonlóság E=$10^{-3}$ alatt. Tehát az összes generált random szekvencia kötött csupán 30 fehérjére azonosított az algoritmus hasonló fehérjét az ismert fehérjeszekvenciák adatbázisában. Ez a teljes GC-tartalmat lefedő adatkészlet 0,03%-a. A BLAST homológia-keresés részletes összefoglalását a 10.1 táblázat mutatja be. Zárójelben feltüntettem a homológként talált szakaszok átlagos hosszát. A randomizált szekvencia-készlet tehát elegendően távoli az ismert fehérjeszekvenciákhoz képest, randomnak és egyben *de*

[1] Streptomyces coelicolor A3(2) genom; Plasmodium falciparum genom

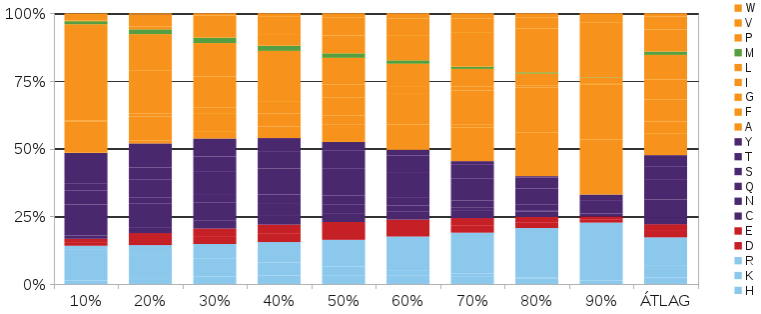
*novo*-nak is tekinthetjuk. *De novo* fehérjék egyben ún. árva, vagy "orphan" fehérjék, mivel nincsen ismert homológjuk. A random fehérjeszekvencia adatkészletek tehát egyben lehetséges árva fehérjék is, amelyek akár biológiai körülmények között is szintetizálódhatnak.

| GC-tartalom | E=10 | | E=0.001 | |
|---|---|---|---|---|
| | nincs | van homológ | nincs | van homológ |
| 10% | 8862 | 1138 (68.25±25.83) | 9999 | 1 (54) |
| 20% | 4651 | 5349 (91.13±33.33) | 9998 | 2 (47) |
| 30% | 2460 | 7540 (100.81±32.99) | 10000 | 0 |
| 40% | 1891 | 8109 (102.30±32.53) | 9992 | 8 (105.38±35.96) |
| 50% | 1822 | 8178 (102.13±33.01) | 9996 | 4 (96.00±10.92) |
| 60% | 2141 | 7859 (99.64±33.91) | 10000 | 0 |
| 70% | 4073 | 5927 (90.11±34.56) | 9991 | 9 (105.11±21.76) |
| 80% | 7953 | 2047 (68.35±27.55) | 9994 | 6 (73.67±31.51) |
| 90% | 9971 | 29 (45.66±13.23) | 10000 | 0 |
| Összesen | 43824 | 46176 | 89970 | 30 |

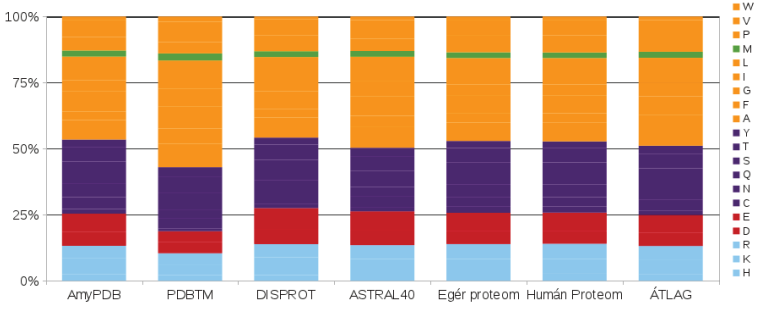
10.1. táblázat. BLAST homológia-keresés eredménye

A random fehérjeszekvenciák aminosav-összetétele természetesen tükrözi a genetikai kód elrendeződését [71]. Alacsony GC-tartalomnál a hidrofób aminosavak nagyobb gyakorisággal jelennek meg, tipikusan a szekvencia 50-70%-át teszik ki, ami nagyjából 80-100 aminosavnak felel meg a 160-ból. 90%-os GC-tartalomnál az összes aminosav között a hidrofóbok már csak 10%-ot tesznek ki és átlagosan a szekvenciák 20%-a arginin ($\approx$ 30 aminosav). A glutaminsav és az aszparaginsav aránya meglepően alacsony. A random szekvenciák aminosav-összetételét a 10.1 ábra foglalja össze.

A genetikai kód elrendeződéseire és az aminosavak fiziko-kémiai tulajdonságaira vonatkozó általános megfigyeléseket már a hetvenes évek végén megtették [138]. A valós fehérjék aminosav összetétele néhány érdekességet tartogat (10.2 ábra), mint például az, hogy savas oldalláncú aminosavak (Glu, Asp) szignifikánsan gyakoribbak mint a genetikai kód kondoneloszlása alapján ez várható lenne. Ugyanakkor az arginin és a lizin valós aránya jóval alacsonyabb, mint a kodoneloszlás szerinti.

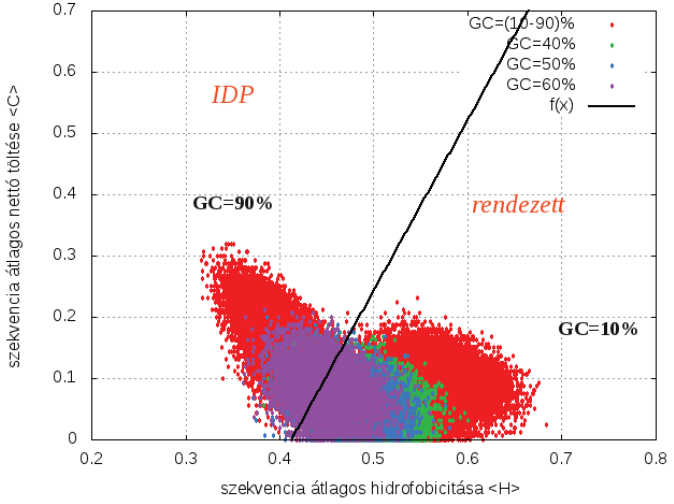



10.1. ábra. Véletlen szekvenciák aminosav-összetéte

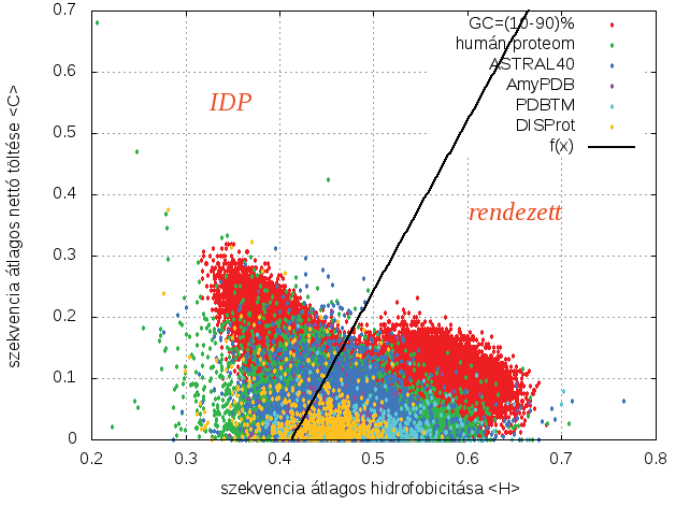



10.2. ábra. Adatbázisok aminosav-összetétele

A rendezetlen fehérjék egyik aminosav-összetételbeli eltérése, hogy a töltött aminosavak nagyobb arányban fordulnak elő. Uversky kétdimenziós térben vizsgálta az egyes fehérjeszekvenciákat, az átlagos nettó töltés és az átlagos normalizált hidrofobicitás mellett. Az IDP-k nettó töltése átlagosan magasabb, mint a globulárisoké. A hidrofobicitás a víz taszításának mértékét jelzi. A rendezetlen fehérjék esetében ez az érték alacsonyabb, mivel kevesebb a hidrofób jellegű, apoláros aminosav. Ezt a két értéket a 9.3. fejezetben leírtak szerint, a 9.4 és 9.5 egyenletek alapján számoltam ki. A két paraméter által kifeszített teret Uversky által meghatározott empírikus összefüggés választja el (9.3 egyenlet) [124]. Mind a random szekvenciák (10.3 ábra), mind az adatbázisbeli szekvenciák (10.4 ábra) esetében elkészítettem az ábrát.




10.3. ábra. A GC=(10-90)% és a biológiailag releváns GC-tartományok fehérjeszekvenciák töltés-hidrofobicitás plot-ja




10.4. ábra. A GC=(10-90)% és a valós fehérjeszekvenciák töltés-hidrofobicitás plot-ja

Az átlagos hidrofobicitás csökken a növekvő GC-tartalommal és a valós szekvenciáké nagyobb tartományt jár be. A *de novo* fehérjék átlagos nettó töltése minimumot mutat 40%-os GC-tartalom környékén, de így is magasabb mint bármelyik valós szekvencia átlagos nettó töltése, ami a DISPROT (rendezetlen fehérjék) esetében a legmagasabb.

Az ASTRAL40 és a humán proteom szekvenciái nagyjából ugyanazt a teret járják be, mint a biológiailag releváns GC-tartalmú random szekvenciák, csak a szórás nagyobb. A DISPROT fehérjeszekvenciák egy jelentős része a rendezett oldalra került. Ez arra vezethető vissza, hogy egy adott fehérje esetében nem feltétlenül a teljes szekvencia rendezetlen és az adatbázis a teljes szekvenciát tartalmazza. Ennek következtében átlagosan a nettó töltés és a hidrofobicitás már a rendezett fehérjékre jellemző értéket veheti fel.

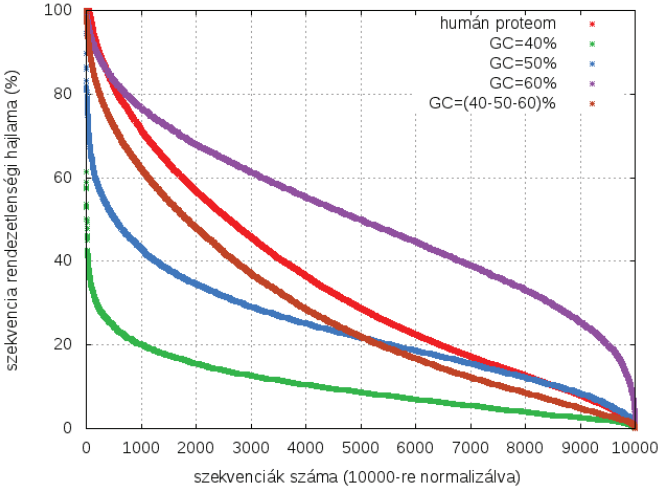
A biológiai fehérjeszintézis modellezése révén *in silico* szintetizált *de novo* random fehérjeszekvenciákat jellemeztem az aminosav-összetétel és két fiziko-kémiai szempont szerint. A megfigyeléseim megegyeznek az irodalomban korábban leírt összefüggésekkel [124].

## 10.2. A szekvencia-alapú predikciókból kiolvasott általános trendek

A random szekvenciák és a randomizált proteomok mellett a valós fehérjék[1] több csoportját is elemeztem ugyanazt a protokollt alkalmazva, azonos feltételek mellett. A globuláris, feltekeredett fehérjéket az ASTRAL40 adatbázis képviseli, a rendezetlen fehérjéket a DISPROT, a transzmembrán fehérjéket pedig a PDBTM adatbázis. Az AmyPDB főleg amiloid-képző rendezetlen fehérjéket tartalmaz, mint például prion vagy tau fehérjéket; az elemzés során ezt a kettősséget figyelembe vettem.

Minden szerkezeti tulajdonságra a 9.4 fejezetben leírtak szerint számoltam ki a szekvenciákra egy konszenzus rendezetlenségi, transzmembrán hélix és aggregációs hajlamot, majd a teljes adatkészletre átlagoltam. A 10.2, a 10.3 és a 10.4 táblázatok összefoglalják az adatbázisokra átlagolt konszenzus predikciókat. Minden táblázatban vastagon kiemeltem a három legmagasabb átlagos értéket. A fehérjék abszolút aggregációs hajlamát jelen elemzéssel nem tudom és nem is kívánom meghatározni. Jelen elemzésben az egyes szekvenciákra kapott eredményekre se térek ki. Az elemzés célja minden esetben az általános trendek azonosítása és a hasonló módszerekkel észlelt tendenciák összevetése, illetve elemzése volt. Minden esetben az általános tendenciákat vizsgálom, nem a konkrét fehérje rendezetlen, transzmembrán hélixképző, aggregációs hajlam értékét.

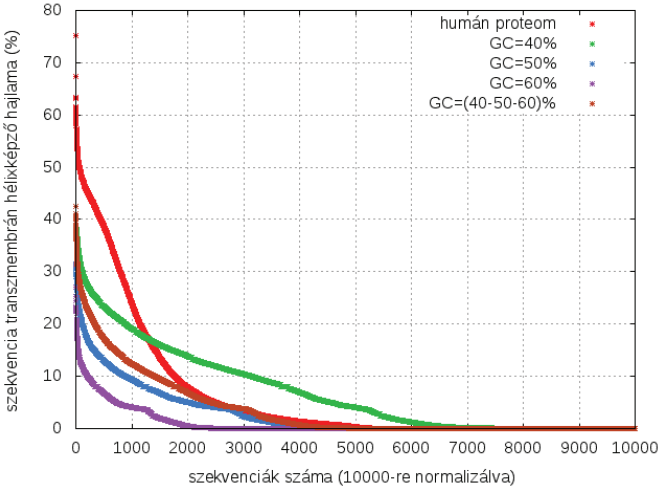
[1] A valós adatbázisok fehérjéit *mai* fehérjéknek is nevezem.




10.5. ábra. Rendezetleségi hajlamra kapott konszenzus predikciók grafikus összefoglalása

| **Rendezetlenség** | | | | | | | |
|---|---|---|---|---|---|---|---|
| adatsor | # | átlag ± szórás | 0% | 25% | 50% | 75% | 100% |
| PDBTM | 429 | 12.59± 9.65 | 0 | 5.75 | 9.89 | 16.88 | 57.38 |
| AmyPDB | 247 | 33.62± 23.62 | 2.01 | 44.90 | 28.30 | 47.29 | 100 |
| DISPROT | 529 | **43.89±28.22** | 3.20 | 21.11 | 37.85 | 64.48 | 100 |
| ASTRAL40 | 10175 | 16.26± 13.45 | 0 | 7.10 | 12.10 | 20.77 | 100 |
| Humán | 20899 | **34.74±24.14** | 0 | 14.99 | 28.79 | 51.11 | 100 |
| random humán | 20899 | 34.40± 25.42 | 0 | 14.35 | 27.36 | 50.67 | 100 |
| Egér | 18525 | 33.31± 23.95 | 0.50 | 13.69 | 26.89 | 49.56 | 100 |
| random egér | 18525 | 33.00± 25.27 | 0 | 13.23 | 25.49 | 49.10 | 100 |
| GC=40% | 10000 | 10.26± 7.32 | 0 | 4.79 | 8.75 | 13.96 | 61.46 |
| GC=50% | 10000 | 24.01± 13.71 | 0.21 | 13.75 | 21.88 | 31.67 | 94.58 |
| GC=60% | 10000 | **50.57±19.25** | 0.83 | 36.04 | 49.90 | 64.58 | 100 |

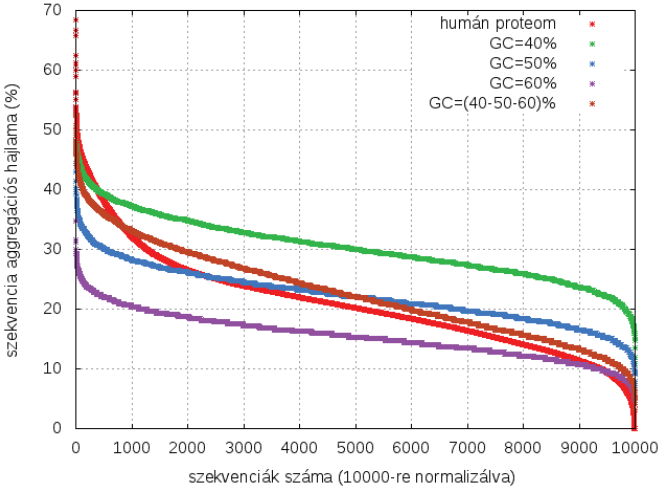
10.2. táblázat. Konszenzus rendezetlenségi hajlam predikciók átlagai az adatkészletekre




10.6. ábra. Transzmembrán hélixképző hajlamra kapott konszenzus predikciók ábrázolása

| **Transzmembrán hajlam** | | | | | | | |
|---|---|---|---|---|---|---|---|
| adatsor | # | átlag ± szórás | 0% | 25% | 50% | 75% | 100% |
| PDBTM | 429 | **31.86±20.11** | 0 | 12.00 | 38.10 | 48.02 | 71.43 |
| AmyPDB | 247 | 4.20± 6.74 | 0 | 0 | 1.57 | 6.73 | 66.32 |
| DISPROT | 529 | 2.61± 6.03 | 0 | 0 | 0 | 2.08 | 47.61 |
| ASTRAL40 | 10175 | 1.15± 5.45 | 0 | 0 | 0 | 0 | 67.68 |
| Humán | 20899 | 6.10± 11.91 | 0 | 0 | 0.34 | 5.00 | 75.14 |
| random humán | 20899 | 3.31± 7.88 | 0 | 0 | 0 | 2.09 | 72.82 |
| Egér | 18525 | **7.03±13.03** | 0 | 0 | 0.69 | 6.10 | 71.43 |
| random egér | 18525 | 3.89± 8.81 | 0 | 0 | 0 | 2.45 | 73.39 |
| GC=40% | 10000 | **6.98±8.08** | 0 | 0 | 3.96 | 12.08 | 42.50 |
| GC=50% | 10000 | 2.65± 4.68 | 0 | 0 | 0 | 3.96 | 38.54 |
| GC=60% | 10000 | 1.01± 2.63 | 0 | 0 | 0 | 0 | 31.25 |

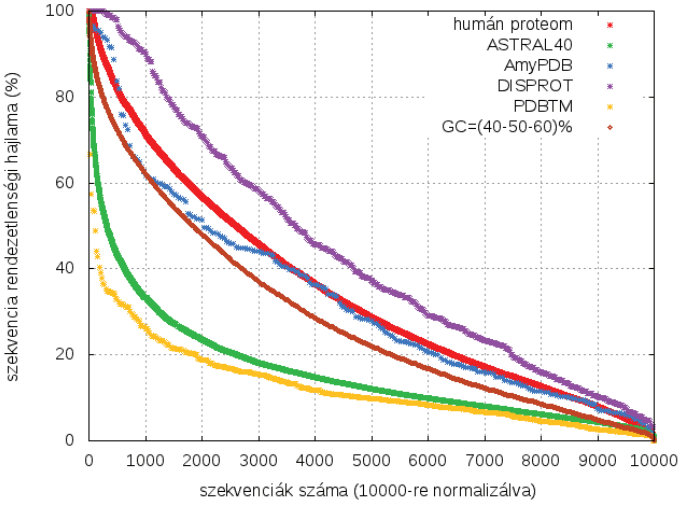
10.3. táblázat. Konszenzus transzmembrán hélix predikciók átlagai az adatkészletekre




10.7. ábra. Aggregációs hajlamra kapott konszenzus predikciók grafikus összefoglalása

| **Aggregációs hajlam** | | | | | | | |
|---|---|---|---|---|---|---|---|
| adatsor | # | átlag ± szórás | 0% | 25% | 50% | 75% | 100% |
| PDBTM | 429 | **34.47±10.79** | 7.10 | 25.61 | 36.82 | 41.97 | 67.97 |
| AmyPDB | 247 | 19.18± 6.55 | 2.35 | 9.29 | 19.10 | 23.87 | 50.00 |
| DISPROT | 529 | 16.51± 6.80 | 0 | 11.72 | 17.26 | 21.07 | 44.19 |
| ASTRAL40 | 10175 | 21.26± 5.49 | 0 | 18.09 | 21.20 | 24.33 | 67.97 |
| Humán | 20899 | 21.05± 8.41 | 0 | 15.25 | 20.20 | 25.07 | 68.50 |
| random humán | 20899 | 20.84± 8.39 | 0 | 15.12 | 20.16 | 25.08 | 70.5 |
| Egér | 18525 | 21.69± 8.88 | 0 | 15.60 | 20.64 | 25.69 | 66.18 |
| random egér | 18525 | 21.49± 8.86 | 0 | 15.35 | 20.62 | 25.74 | 64.71 |
| GC=40% | 10000 | **30.29±5.25** | 10.94 | 26.56 | 30.00 | 33.75 | 52.50 |
| GC=50% | 10000 | **22.31±4.58** | 5.94 | 19.06 | 22.19 | 25.31 | 44.38 |
| GC=60% | 10000 | 15.50± 3.81 | 3.12 | 12.81 | 15.31 | 17.81 | 34.69 |

10.4. táblázat. Konszenzus aggregációs hajlam predikciók átlagai az adatkészletekre



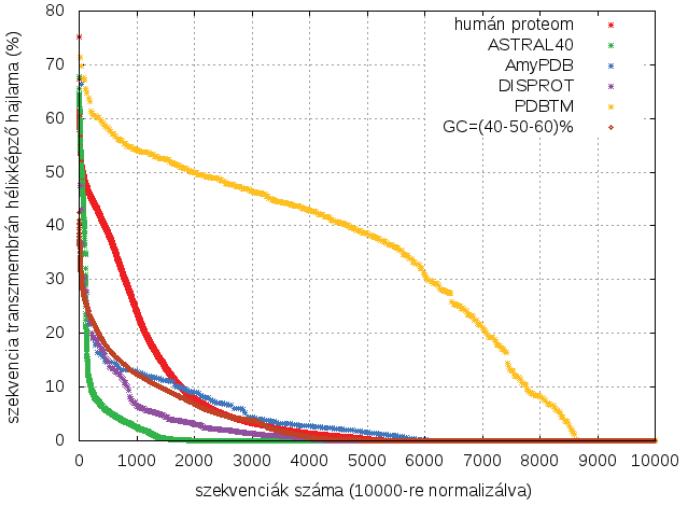

10.8. ábra. Rendezetlenségi hajlam konszenzus predikciók a valós adatkészletekre

A szerkezeti tulajdonságok jóslására alkalmazott konszenzus predikciók segítségével általános trendeket azonosítottam a véletlenszerű fehérjeszekvenciák, valamint a mai, természetes fehérjeszekvenciák adatkészleteire vonatkozóan. A továbbiakban a random adatkészletet leszűkítettem a biológiailag releváns GC-tartalmú random fehérjeszekvenciákra, 40, 50 és 60%-os GC-tartalommal. A random szekvenciák konszenzus predikciós eredményei mutatják, hogy a GC-tartalom meghatározza a szerkezeti preferenciákat és világos trendek olvashatók ki az adatokból, amelyek jóval markánsabbak mint a 10.1 fejezetben bemutatott fiziko-kémiai paraméterekből kiolvasott trendek.

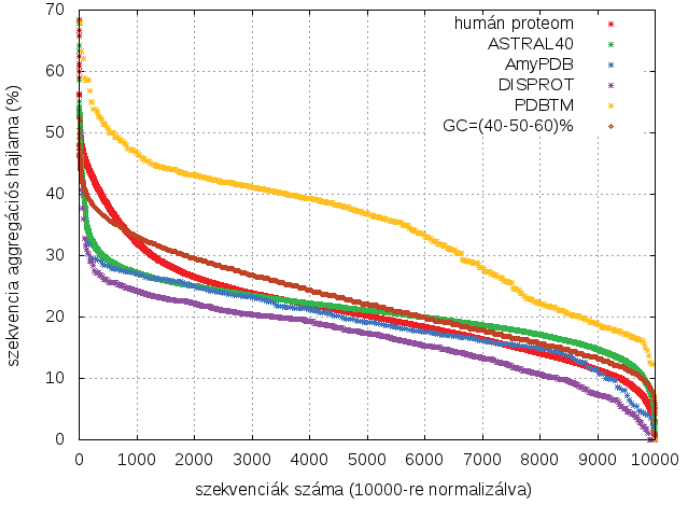
Magas GC-tartalom mellett a rendezetlenség dominál. 50%-os GC-tartalom környékén a szekvenciák felére igaz, hogy az adott szekvencia negyede rendezetlen. Átlagban ez az érték majdnem 50%. Ebben a vonatkozásban csak a transzmembrán és az amilod adatbázisok fehérjéi mutatnak kisebb átlagos értéket. 60%-os GC-tartalom felett a random szekvenciákat gyakorlatilag teljesen rendezetlennek jósolják a prediktorok; általában ez egyetlen nagy vagy két hosszabb rendezetlen szakaszt takar[1].

A transzmembrán hélixképző hajlam aránylag nagy a magasabb GC-tartományban, de nagyon gyorsan csökken, hogy végül 60%-os GC-tartalom környékén gyakorlatilag

[1] ld. IV. publikáció 1. táblázata




10.9. ábra. Transzmembrán hélixképző konszenzus predikciók a valós adatkészletekre




10.10. ábra. Aggregációs hajlam konszenzus predikciók a valós adatkészletekre

eltűnjön. 40%-os GC-tartalom környékén a transzmembrán hélix-jósolt aminosavak aránya hasonló a teljes egér vagy humán proteomra jósolt átlagos arányhoz. A random szekvenciák aggregációs hajlama alacsony GC-tartalomnál a legmagasabb. Magasabb GC-tartalmak mellett nagyon hamar eltűnik ez a hajlam; 50%-os GC-tartalomnál már csak az aminosavak 5%-át jósolják a prediktorok aggregációra hajlamosnak. A rendezetlenség nagyon erősen függ a random fehérjeszekvenciák GC-tartalmától. A 60%-os GC-tartalmú szekvenciák átlagos rendezetlensége határozottan magasabb, mint akár a DISPROT adatbázisbeli fehérjéké. A 40 és 50%-os GC-tartalmú szekvenciákra ezzel szemben kisebb mértékű rendezetlenséget jeleznek az algoritmusok, mint a teljes proteomok esetében.

A jósolt transzmembrán hélixképző hajlam alacsonyabb az 50%-os és 50% alatti GC-tartalmú random szekvenciák esetében, mint a teljes proteomokra jósolt értékek. Természetesen a transzmembrán adatkészlet (PDBTM) jósolt transzmembrán hélix tartalma jóval meghaladja a többi adatkészletre jelzett értékeket. A PDBTM transzmembrán adatkészletre kapott transzmembrán hélix tartalom teljesen összeegyeztethető az adatkészlet összetételével. Az adatkészletben 75%-ban hélix, 25%-ban $\beta$-hordó transzmembrán elemek találhatók; a konszenzus predikció a szekvenciák 75%-ban hélixet jósol. A konszenzus predikciók és a valós adatkészletek ismert összetétele a többi esetben is jól egyeznek.

Meglepő módon, a konszenzus predikció átlagosan a PDBTM adatbázisra jelzi a legmagasabb aggregációs hajlamot, és nem az AmyPDB adatkészletre. Az AmyPDB adatkészlet aggregációs tendenciája közelebb van a rendezetlen fehérjék (DISPROT), globuláris (ASTRAL40) vagy a teljes proteom adatkészletek aggregációs hajlamához. Az AmyPDB adatbázis főleg prion és tau fehérjéket tartalmaz; ezekre a fehérjékre van adat az *in vitro* vagy *in vivo* aggregációs hajlamra vonatkozóan. Ugyanakkor ez a két fehérjecsalád belsőleg is rendezetlen, ami a jósolt rendezetlenségi hajlamban is megjelenik, alátámasztva a konszenzus predikció hatékonyságát. Ugyanakkor a konszenzus aggregáció predikció a kisebb mértékben adja vissza az AmyPDB kettős arculatát. A transzmembrán fehérjék esetében a membránba való beépülés védelmet jelent az aggregáció ellen a sejtben. A transzmembrán fehérjék aminosav-összetétele is tükrözi ezt a "szabadságot". A konszenzus predikciók eredményei végképp alátámasztják ezt a kettőséget (10.9 és 10.10 ábrák). A transzmembrán adatbázisokra kapott eredmények a konszenzus predikciók robusztusságát is alátámasztja, mivel az adatbázisra jellemző szerkezeti tulajdonságot különösen jól tükröző adatkészletet szolgál.

A természetes fehérje-adatkészletek esetében a szórások[1] magasabbak, mint a véletlenszerű fehérjeszekvenciák adatkészleteinél. A valós, evolúciós szelekción átesett, funkcionális fehérjék tehát nagyobb variabilitást mutatnak, mint a véletlenszerű, adott GC-tartalmú polipeptidláncok.

[1] standard deviation

## 10.3. A strukturális tulajdonságok egymás közötti összefüggései

Az irodalomban többen is leírták a rendezetlenségi és az aggregációs hajlam közötti negatív korrelációt. Minél rendezetlenebb egy szekvencia, annál kisebb az aggregációs hajlama. Linding és munkatársai ezt a negatív korrelációt teljes proteomok összevetése során tapasztalták [91]. Monsellier és Chiti a humán proteom elemzése során is kimutatta, hogy a belsőleg rendezetlen fehérjék aggregációs hajlama kisebb, mint a kevésbé rendezetlen globulárisoké [89].

Megvizsgáltam tehát, hogy a jósolt szerkezeti preferenciák között tapasztalható-e hasonló összefüggés. Először szekvencia-szinten elemeztem az adatokat. Kiszámoltam a Pearson-féle korrelációs együtthatókat a rendezelennek, transzmembrán-hajlamúnak vagy aggregálónak jósolt aminosav-százalékok között. Az eredményeket a 10.5, 10.6 és 10.7 táblázatok foglalják össze; minden esetben 0,05 szignifikancia-szintre értendők az eredmények. A P-érték minden esetben nulla; az egyetlen kivételnél feltüntettem az értéket.

| Korreláltatott tulajdonságok | 40% | 50% | 60% | (40-50-60)% | Humán | Egér |
|---|---|---|---|---|---|---|
| szekvenciák száma (n) | 10000 | 10000 | 10000 | 30000 | 25000 | 18555 |
| rendezetlenség-transzmembrán | -0,205 | -0,193 | -0,248 | -0,375 | -0,350 | -0,380 |
| rendezetlenség-aggregáció | -0,572 | -0,692 | **-0,771** | **-0,834** | **-0,780** | **-0,781** |
| aggregáció-transzmembrán | 0,587 | 0,433 | 0,331 | 0,588 | 0,742 | **0,780** |

10.5. táblázat. Korrelációk a biológiailag releváns GC-tartományú random és a humán adatokra

| Korreláltatott tulajdonságok | 10% | 20% | 30% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|
| szekvenciák száma (n) | 10000 | 10000 | 10000 | 10000 | 10000 | 10000 |
| rendezetlenség-transzmembrán | -0,126 | -0,185 | -0,206 | -0,198 | -0,125 | -0,012* |
| rendezetlenség-aggregáció | -0,344 | -0,406 | -0,452 | -0,748 | -0,583 | -0,311 |
| aggregáció-transzmembrán | 0,482 | 0,647 | 0,673 | 0,156 | 0,060 | -0,115 |

*: p=0.244

10.6. táblázat. Korrelációk a random adatkészletekre

A vizsgált tulajdonságok csak gyengén kapcsolhatók össze, de az irodalomban leírt rendezetlen-aggregáció negatív korreláció kirajzolódik. A |0,75| feletti, erősnek mond-

| Korreláltatott tulajdonságok | ASTRAL40 | AmyPDB | DISPROT | PDBTM |
|---|---|---|---|---|
| szekvenciák száma (n) | 10500 | 250 | 530 | 442 |
| rendezetlenség-transzmembrán | -0,012 | -0,142 | -0,080 | -0,514 |
| rendezetlenség-aggregáció | -0,573 | **-0,800** | **-0,860** | -0,631 |
| aggregáció-transzmembrán | 0,455 | 0,453 | 0,334 | **0,891** |

10.7. táblázat. Korrelációk a valós adatbázisokra

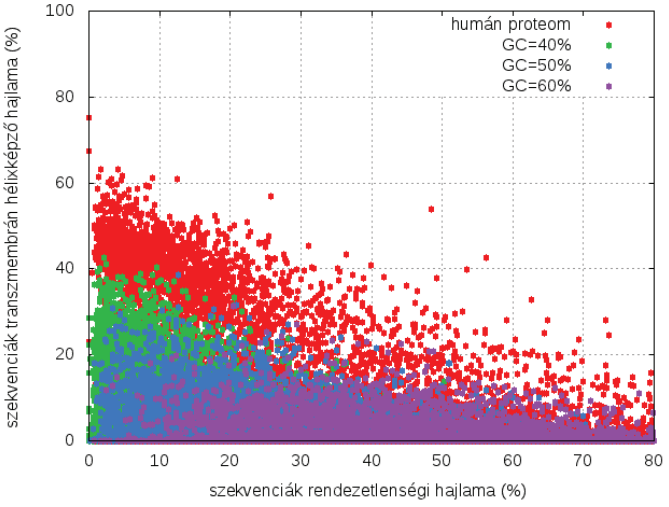
ható korrelációs érékeket vastagon szedtem a táblázatokban. A rendezetlenség és az aggregáció antikorrelációja mind a random, mind a valós adatkészletekre kimutatható. A rendezetlenség továbbá a transzmembrán hajlammal ellentétesen változik. Az aggregáció és a transzmembrán tendencia ezzel szemben minimális pozitív korrelációt mutat.

A random adatok korrelációs együtthatói szignifikánsan változnak a GC-tartalommal. A szélsőséges GC-tartalmakra a korrelációs értékek az mutatják, hogy a két adatkészlet teljesen eltér. A valós adatkészletek korrelációs együtthatói nagyon eltérnek a 40 és 60% közötti és az átlagolt random adatkészletekre kapott tendenciáktól.

A PDBTM adatkészletre kapott értékek jelentősen eltérnek a többi adatkészlettől, nagyon erős pozitív korrelációt mutat az aggregáció-transzmembrán adatokra. Ezzel szemben a DISPROT esetében az aggregáció-transzmembrán korreláció csak 0,334. A rendezetlen-aggregáció adatkészletpárra a DISPROT adatbázisra -0,860 a korreláció és hasonlóan erős negatív korrelációt tapasztaltam az AmyPDB adatokra. Az ASTRAL40 adatkészletekre mindhárom vizsgált tulaljdonságpáros gyenge korrelációt mutat. A valós adatkészletek korrelációs értékeiből tehát kikövetkeztethető, mely szerkezeti tulajdonság a domináns a szekvenciákra.

Ábrázoltam a transzmembrán hajlamot a rendezetlenség függvényében (10.11 ábra), valamint a transzmembrán hajlamot és a rendezetlenséget az aggregációs potenciál függvényében (10.12 ábrák). Az ábrákon nagyon szembetűnő, hogy a random és a valós szekvenciák által behatárolt tér nem azonos teljes mértékben. Bár a tendenciák megegyeznek, a rendezetlenség-transzmembrán-aggregáció által bejárható tér csak részben egyezik a random és a valós fehérjéknél, és a humán proteom fehérjéi által behatárolt tér nagyobb, mint amit a teljes random szekvenciakészlet által le lehet fedni. Az ábrákon csak a biológiailag releváns GC-tartomány szekvenciáit tüntettem fel, de a teljes GC-tartalommal se lehet lefedni teljes mértékben a humán proteom által behatárolt teret.

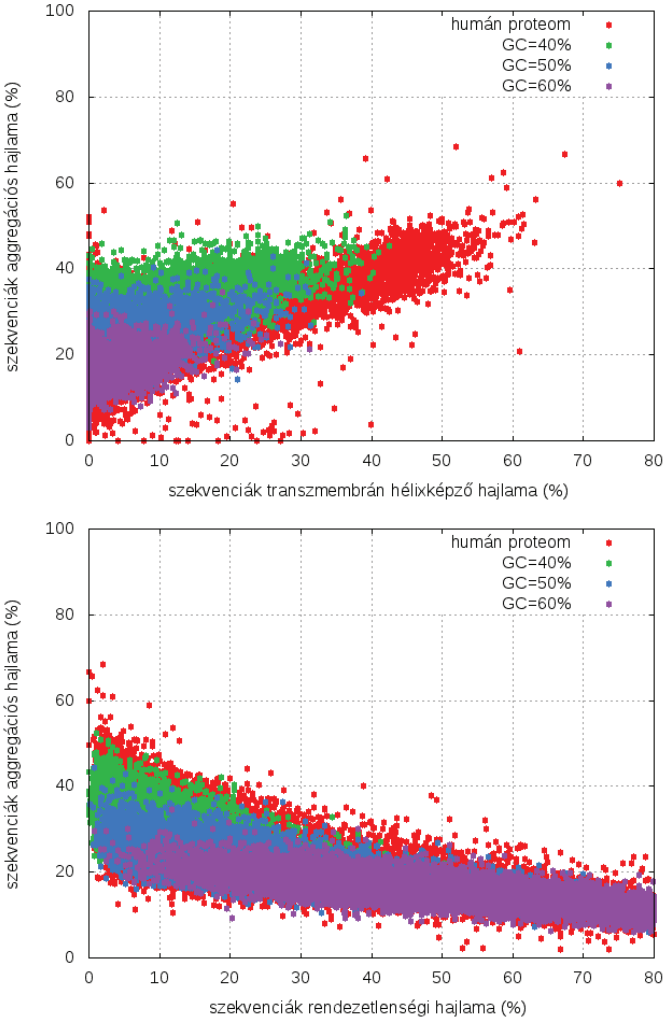
Azok a szekvenciák, amelyek rendezetlenebbnek lettek jósolva, alacsonyabb jósolt transzmembrán hélixképző és aggregációs hajlamot mutatnak. Ennek ellenére ez az összekapcsolás szélesebb spektrumot jár be. Alacsony jósolt rendezetlenség számos esetben




10.11. ábra. Kétdimenziós ábrázolás a szekvenciák konszenzus (rendezetlen-transzmembrán) predikcióira

magasabb transzmembrán hélixképző hajlammal párosul. Az aggregáció ezzel szemben sokkal szigorúbban antikorrelál a rendezetlenséggel. A transzmembrán hélixképző hajlam viszont együtt változik az aggregációs hajlammal. A humán proteom adatkészlete ebben az esetben is látványosan túlmutat a random adatkészletekkel bejárt területhez képest. Hasonló tendenciák olvashatók ki a valós adatkészletekkel összevetve. A jósolt transzmembrán hélixképző hajlam egységesen majdnem minden adatkészlet esetében jóval magasabb, mint a random szekvenciákkal elérhető maximális érték.

A tendenciák tehát azt mutatják, hogy a fehérjeszekvenciák aminosav-összetétele döntő szerepet játszanak a struktúrák kialakulásában. A szerkezetjósló algoritmusok kizárólag a fehérjeszekvenciát használják fel, de a betanító adatkészletek és az algoritmusok elméleti hátterében minden esteben jelen vannak a mai ismert fehérjék strukturális jellemzői. A transzmembrán hélixek esetében a lokális struktúrák kialakulása és felismerése egy ismert szekvencia esetében magasabb értékhez vezethet, mint a teljesen véletlenszerű szekvencia esetében.




10.12. ábra. Kétdimenziós ábrázolás a szekvenciák konszenzus predikcióira.

## 10.4. Statisztikai elemzés

Arra a kérdésre, hogy a véletlenszerű aminosavszekvenciájú fehérjék aggregációs hajlama megakadályozza-e az új fehérje keletkezését, úgy tűnik nem a válasz, ugyanis a random szekvenciák aggregációs hajlama nem tér el a valós fehérjéektől. A kétdimenzióban ábrázolt szerkezeti tulajdonságok tehát szokatlan összefüggésekre világítottak rá a valós fehérjék és a random szekvenciákra nézve. Az aggregációs hajlamot ábrázolva a rendezetlenségi haljam függvényében, a humán proteom által bejárt terület csak minimálisan, de szignifikánsan látszik eltérni a random szekvencik által bejárt területtől.

A tapasztalt eltérést próbáltam számszerűsíteni az egy és kétdimmenziós Komlomorov-Smirnov teszt segítségével [135]. A K-S teszt két kumulatív eloszlásfüggvény abszolút különbségének közötti maximális eltérést keresi. A biológiailag releváns GC-tartalmat lefedő random szekvenciák adatkészleteit külön-külön és egyesítve a humán proteom fehérjekészletével szemben vizsgáltam. A statisztikák viszont azt mutatták, hogy a humán proteom és a feltüntetett random adatkészletek mintaeloszlása mindhárom esetben nem egyeznek (P-érték nulla minden esetben).

| 1D KS | 40% | 50% | 60% | (40-50-60)% |
|---|---|---|---|---|
| szekvenciák száma (random) | 10000 | 10000 | 10000 | 30000 |
| szekvenciák száma (humán) | 20899 | 20899 | 20899 | 20899 |
| Rendezetlen | 0,550 | 0,250 | 0,373 | 0,117 |
| Transzmembrán | 0,367 | 0,273 | 0,396 | 0,226 |
| Aggregáció | 0,612 | 0,298 | 0,418 | 0,154 |

10.8. táblázat. 1D Kolomorov-Smirnov korrelációk a random és a humán adatok predikciói között

A pontok elhelyezkedése végső soron az egyenlő szekvenciahosszak miatt sokkal szabályosabb, mintegy rácsszerű a random adatkészletek esetében, míg a változatos láncchosszakkal rendelkező humán fehérjekészletnél a kumulatív eloszlás sokkal egyenletesebb. Ezen felül a pontok által legsűrűbben lefedett területek is különböznek.

Ezért az adatokat egy rácsozásos módszerrel is összehasonlítottam (ld. 9.5 fejezet). A statisztikai elemzéssel szemben, ha a két- és háromdimenziós térben a különböző tulajdonságok által lefedett területeket vetem össze, más kép rajzolódik ki (10.13 ábák). A random szekvenciák által lefedett terület több mint 95%-ban átfed a humán proteom által lefedett területtel. Viszont az átfedés csupán 35% körüli, ha a humán proteom által

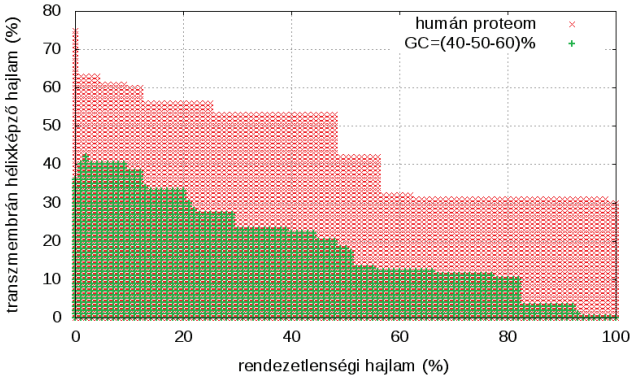
| 2D KS | 40% | 50% | 60% | (40-50-60)% |
|---|---|---|---|---|
| szekvenciák száma (random) | 10000 | 10000 | 10000 | 30000 |
| szekvenciák száma (humán) | 20899 | 20899 | 20899 | 20899 |
| Aggregáció-Transzmembrán | 0,494 | 0,279 | 0,377 | 0,129 |
| Rendezetlen-Aggregáció | 0,591 | 0,233 | 0,383 | 0,119 |
| Rendezetlen-Transzmembrán | 0,290 | 0,148 | 0,294 | 0,054 |

10.9. táblázat. 2D Kolomorov-Smirnov korrelációk a random és a humán adatok predikciói között

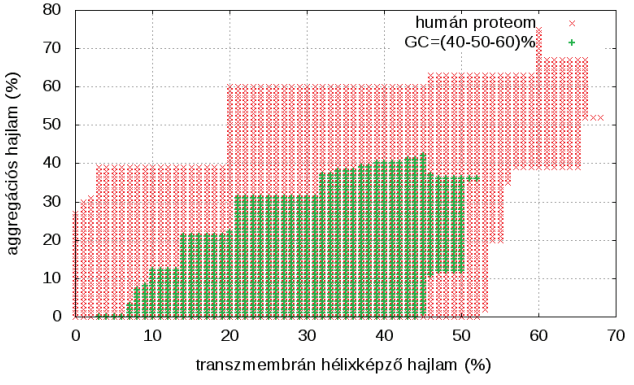
lefedett területet nézzük a randoméhoz képest. A random szekvenciák azon területe, ami nincs átfedésben a humán proteom által lefedett területekkel az alacsony aggregációs hajlamú szekvenciáknak felel meg. A random és a proteombeli szekvenciák adatai között a leglátványosabb eltérés a humán proteom kiemelkedően nagyobb hajlama a transzmembrán hélixképzésre.

| | Humán proteom által lefedett terület | (40-50-60)% által lefedett terület | Átfedés | Átfedés % a humán proteomra vonatkoztatva | Átfedés % a randomra vonatkoztatva |
|---|---|---|---|---|---|
| Aggregáció-Transzmembrán | 317321 | 121654 | 121654 | **38,34%** | **100%** |
| Rendezetlen-Aggregáció | 284584 | 184756 | 182566 | **64,15%** | **98,81%** |
| Rendezetlen-Transzmembrán | 447448 | 189539 | 182566 | **42,05%** | **99,27%** |
| Rendezetlen-Transzmembrán-Aggregáció | 8611471 | 3138021 | 515864 | **34,87%** | **95,70%** |

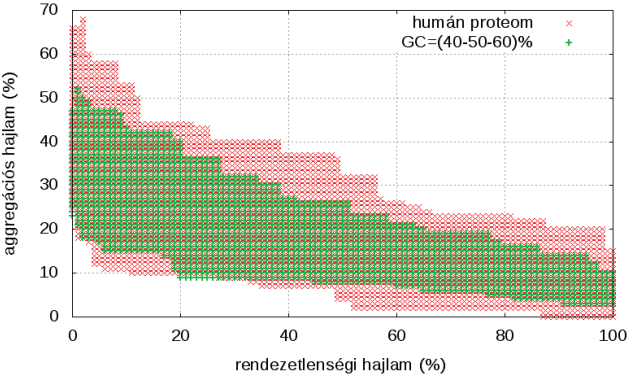
10.10. táblázat. Becsült átfedések a kiválasztott rendezetlen-transzmembrán-aggregáció adatsorok között




[A]




[B]




[C]

10.13. ábra. A lefedett területek összehasonlítása, 0.1%-os felbontásban.

## 10.5. Proteomok és randomizált proteomok összevetése

A random adatkészletek és a humán proteom összevetésére kapott némileg meglepő eredményeket egy másik random adatkészlettel is ellenőrizni kívántam. A fehérjeszekvenciák randomizálásának egy módja, hogy a fehérjelánchossz és az aminosav-összetétel megtartása mellett, a véletlenszerűen kiválasztott két aminosav pozícióját n-szer permutáltam, ahol n a fehérje lánchossza. A humán és az egér proteomra is alkalmaztam ezt a randomizálási módot. Az irodalomban általánosan ily módon előállított random adatkészleteket használnak a módszerek robusztusságának vizsgálatai során. Ebben az esetben már nem csupán 160 aminosavas random fehérjéket vizsgálok, hanem a random proteommal megegyező lánchosszeloszlású és aminosav-összetételű adatkészlettel dolgozom.

A random proteomokat is alávetettem a konszenzus szerkezeti elemzéseknek. Az eredményeket a 10.14 ábra szemlélteti: az aggregációs hajlam függvényében ábrázoltam a transzmembrán hélixképző és a rendezetlenségi hajlamot.

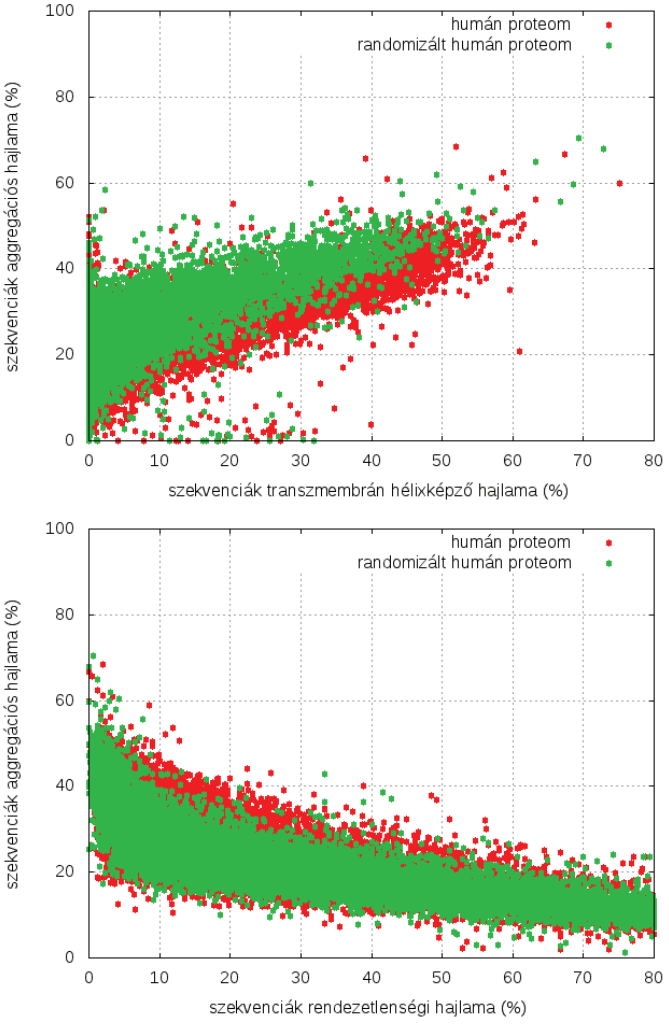
| Korreláltatott tulajonságok | Humán | random Humán | Egér | random Egér |
|---|---|---|---|---|
| szekvenciák száma (n) | 20899 | 20899 | 18525 | 18525 |
| rendezetlenség-transzmembrán | -0,350 | -0,386 | -0,380 | -0,420 |
| rendezetlenség-aggregáció | **-0,780** | **-0,836** | **-0,781** | **-0,833** |
| aggregáció-transzmembrán | 0,742 | 0,686 | **0,780** | 0,714 |

10.11. táblázat. Korrelációk a valós adatbázisokra

Páronként korreláltattam a jósolt szerkezeti tulajdonságokat. A Pearson korrelációs koefficienseket a 10.11 táblázat foglalja össze. A strukturális tulajdonságok közötti korreláció iránya nem változott, a mértéke viszont megnőtt. Különösen a rendezetlenség-aggregáció negatív korrelációja markánsabb a randomizált adatkészletekre.

Az adatkészletekre történt átlagolás után a rendezetlenségi és az aggregációs hajlam nem változik jelentősen (10.2 és 10.3 táblázatok). Viszont mind a humán, mind az egér randomizált proteom átlagos transzmembrán hajlama csökken a humán és egér proteomához képest (10.4 táblázat). A transzmembrán hajlam jóslása tehát érzékenyebb a konkrét aminosav-sorrendre, mint a rendezetlenségi vagy az aggregációs hajlam előrejelzése.

A 40-60%-os GC-tartalmú random adatkészletekhez hasonlóan ábrázoltam a szerkezeti tulajdonságokat kétdimenzióban, a humán proteom konszenzus predikcióival együtt. Továbbra is minden pont egy szekvenciát jelent a két strukturális tulajdonság által kifeszített térben. Az aggregáció függvényében ábrázolt transzmembrán hélixképző és rendezetlenségi hajlamot mutatom be a 10.14 ábrán. A humán proteom szekvenciái nagyobb teret járnak be, mint a randomizált humán proteom szekvenciái. Ez az eredmény teljes




10.14. ábra. Rendezetlenségi, transzmembrán és aggregációs hajlamok a humán és a random (shuffled) humán proteomra

összhangban van a random *de novo* fehérjeszekvenciákra kapott eredményekkel. Ugyanúgy a transzmembrán hélixképző hajlam nagyobb a humán proteom, mint a randomizált humán proteom szekvenciái esetében.

Tehát a random *de novo* fehérjékre megfigyelt trendek nem a szekvenciák előállításának módjából erednek, hanem a nem valós, biológiai környezetben keletkezett fehérjékre jellemző eltéréseket mutatnak. A bemutatott két random fehérjeszekvenciakészletre ugyanazokat a trendeket figyeltem meg mind a szerkezeti tulajdonságok egymás közötti kapcsolataiban, mind a valós fehérjékkel szembeni eltérésekben.

## 10.6. *De novo* fehérjék a humán genomban

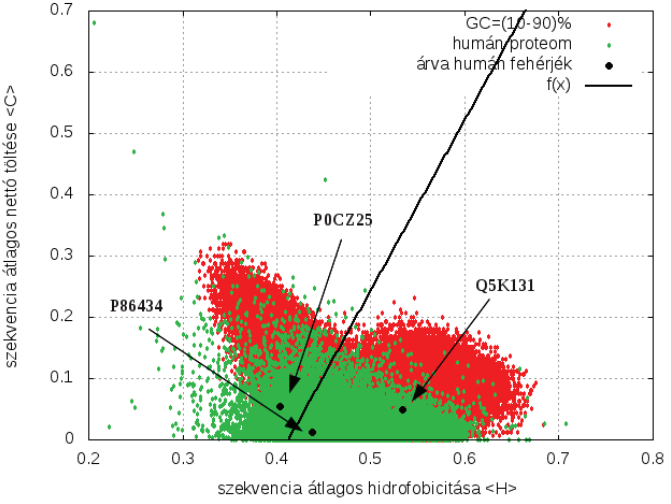
A vizsgálat célja a *de novo* keletkező fehérjék aggregációs hajlamának becslése volt. A random szekvenciákkal modellezett *de novo* fehérjéket a valós *de novo* fehérjék szerkezeti tualjdonságaival is összehasonlítottam. Knowles és munkatársai három humán fehérjéről bizonyították be, hogy más főemlős genomjában megtalálható a megfelelő DNS szekvencia de nem íródik át, viszont a humán genomban egyértelmüen átíródik és fehérje keletkezik [97]. Mindhárom szekvencia esetében az UniProt [72] semmilyen szerkezeti információt nem tartalmaz a fehérje térszerkezetét illetően. Funkcionális szempontból is csak a CLLU1 fehérjéről tudjuk, hogy krónikus limfocitás leukémia esetében kimutatható mennyiségben expresszálódik.

Kiszámoltam a három szekvencia <H> átlagos hidrofobicitását és <C> nettó töltését, majd ábrázoltam őket a töltés-hidropátia térben a teljes GC-tartományt lefedő random szekvenciák és a biológiailag relevánas GC=40, 50 és 60%-os szekvenciákat (10.15 ábra). A háromból egy szekvencia (P86434) az IDP vagy rendeztlen, kettő (P0CZ25, Q5K131) pedig a rendezett vagy globuláris oldalra esik. Ez önmagában még nem elegendő információ, mivel a három szekvencia közül kettő is nagyon közel esik a választóegyeneshez.

A három fehérjékre kiszámoltam a konszenzus strukturális predikciókat. Az eredményeket a 10.12 táblázat foglalja össze. Továbbá a nukleotidszekvenciákra visszaszámolt GC-tartalmakat is feltüntettem.

A kapott értékek meglepően jól összhangban vannak a random *de novo* fehérjéknél megfigyelt tendenciákkal. A megfelelő DNS szegmens GC-tartalma és a keletkező fehérje stukturális preferenciái ebben a három esetben is ugyanazokat a tendenciákat tükrözik.

A P86434 és a P0CZ25 azonosítójú fehérjéke inkább rendezetlennek hozta ki a konszenzus predikció, és a kódoló DNS szegmensük GC-tartalma 60% körüli. Ezzel szemben a Q5131 azonosítójú fehérje aggregációs hajlama jelentős és a szekvencia 10%-ára transzmembrán hélixet jósolt a konszenzus predikció; a kódoló DNS szegmens GC-tartalma




10.15. ábra. A teljes random adatkészlet, a humán proteom és az árva humán fehérjék ábrázolása a C-H diagramon

| fehérje | fehérje-lánchossz | %D | %T | %A | G+C% mRNS | G+C% kódoló |
|---|---|---|---|---|---|---|
| P86434\|CV045 | 159 | **34,31** | 0 | 9,11 | 58,89 | 63,60 |
| P0CZ25\|D100S | 163 | **86,31** | 0 | 5,25 | 54,92 | 59,12 |
| Q5K131\|CLLU1 | 121 | 3,18 | **10,42** | **33,42** | 37,11 | 31,96 |

10.12. táblázat. Humán *de novo* fehérjék konszenus szerkezeti predikciói és számolt GC-tartalmuk

pedig 32%. A két eredmény pontosan tükrözi a random szekvenciák elemzéséből várt trendet. A töltés-hidropátia elemzés során csak a P86434 fehérje került látszólag "rossz" oldalra. Ugyanakkor a jósolt rendezetlenségi hajlama csak 34%, tehát a szekvencia nagyobbik fele nem rendeztelen és összhangba kerül a két eredmény.

Az árva *de novo* humán fehérjék elemzése tehát teljes mértékben alátámasztja a fentebb bemutatott trendeket, és igazolja a *de novo* keletkező fehérjeszekvenciák szerkezeti preferenciáit vizsgáló modell robusztusságát és hatékonyságát.

# 11. fejezet

# Diszkusszió

A dolgozat második részében bemutatott *in silico* elemzésben hipotetikus, *de novo* véletlenszerű szekvenciájú fehérjéket elemeztem. A véletlenszerű aminosav-szekvenciákat a nukleotid szintű szisztematikus GC-tartalom változtatásával értem el. Az irodalomban vizsgált random fehérjeszekvenciák aminosav-összetételeit általában *a priori* határozzák meg. Ezzel szemben az általam vizsgált random fehérjeszekvenciák aminosav-összetételét a kódoló DNS szegmens GC-tartalma és a (standard) genetikai kód határozza meg.

A GC-tartalom és a genetikai kód kapcsolata a lefordított fehérjék aminosavösszetételével nem új, és nem meglepő, de az általam követett új megközelítési mód biztos alapokra helyezi az újonnan keletkező fehérjék szerkezeti, strukurális *in silico* elemzését. Továbbá a random fehérjék származtatási módja egy realisztikus forgatókönyvnek felel meg: kimutatták hogy ténylegesen keletkeztek fehérjék a humán genomból is ilyen mechanizmussal. Bemutattam, hogy azonosítani lehet a potenciális *de novo* strukturális tendenciákat a hipotetikus lefordított genomszakaszok függvényében. Magasabb GC-tartalom genom szinten a lefordítandó szekvenciában nagyobb rendezetlenségi hajlamot, és fehérjeszinten kisebb transzmembrán és aggregációs hajlamnak felel meg.

A 40 és 60%-os GC-tartalom közötti random DNS szakaszok alapján lefordított fehérjék a humán proteomhoz viszonyítva kisebb területet feszítenek ki a jósolt szerkezeti tulajdonságokkal. A random *de novo* fehérjék jósolt aggregációs tendenciája nem nagyobb, mint a valós fehérjéké, és a jósolt rendezetlenségre való hajlamuk sem nagyobb mint a mai fehérjéké. Ugyanakkor a random fehérjékhez jóval kisebb transzmembrán hélixképző hajlamot tudtam rendelni, mint amit a valós fehérjék esetében találtam. Tehát úgy tűnik, hogy a transzmembrán hélixek keletkezése nagyobb evolúciós nyomásnak van

kitéve mint a másik két szerkezeti tulajdonság, azaz a transzmembrán hélixek kialakulása a szerkezet optimálását igényli.

A bemutatott elemzés kizárólag *in silico* szerkezeti predikciókból kiolvasott tendenciákat mutat be. Egyetlen paramétert vettem figyelembe, nevezetesen a szekvenciában adott szerkezeti tulajdonságúnak jósolt aminosavak százalékos arányát. Nincs szó valós evolúciós folyamatok modellezéséről, bizonyos strukturális trendek megfigyelésére irányult a munka. A genomok szekvenciái nem véletlenszerűek, és a valós fehérjékre jósolt szerkezeti tulajdonságok eltérhetnek a GC-tartalom alapján várhatóaktól. Például, a PDBTM adatbázis aminosav-összetétele a 70-80%-os GC-tartalmú nukleotidoszekvenciákból lefordított fehérjék aminosav-összetételéhez állt legközelebb. Ugyanakkor a PDBTM adatbázis átlagosan magasabb transzmembrán hélixképző és aggregációs hajlamot mutat, mint a többi természetes fehérje, ellentétben a magas GC-tartalmú fehérjékkel való hasonlóság alapján várható tendenciával.

A megállapítás, hogy a *de novo* fehérjék nem különösebben hajlamosak aggregációra, ellentmondani látszik az irodalomban leírt elvvel, hogy az aggregáció elleni védekezés optimálási lépés a fehérjeevolúció során. Ugyanakkor az alkalmazott predikciós módszer, ami csak három globális szerkezeti tulajdonságra ad becslést, nem mond semmit a részletes szerkezeti vagy funkcionális elemekről. Feltételezzük, hogy a *de novo* fehérje keletkezése után a szerkezet szelekció utján optimalizálódik a funkció betöltésére, és a szerkezet, stabilitás és a mozgékonyság együtt lesz optimálva a funkcióval. E folyamat alatt az aggregációs hajlam megtartása vagy akár csökkentése, ami az újonnan keletkező *de novo* fehérje belső tulajdonsága, egy erős szelekciós nyomást gyakorol az evolúció során.

A hidrofób mag vagy a transzmembrán hélixek kialakulása során az eredetileg nagyobb aggregációs hajlam előny is lehet. A strukurális átalakulás után az aggregációs hajlam megváltozik. Amíg az újonnan keletkező fehérje túléli a szervezet viszontagságait, különösen ha alacsony a kifejeződése a sejtben, a fent leírt forgatókönyvek nem irreálisak.

Az eredményeink elég jól összeegyeztethetők a szelekciós nyomás létezésével, ami a fehérjeevolúció minden későbbi szakaszánál felléphet. Viszont ki kell hangsúlyozni, hogy az új kódoló szekvencia megjelenését nem gátolja feltétlenül az a tény, hogy a lefordított fehérjeszekvencia majd szokatlanul nagy aggregációs hajlammal rendelkezik.

# 12. fejezet

# Összefoglalás

A doktori munkámban bemutatott két témakör közös pontja a mozgékonyság kérdése. A fehérjék belső dinamikájának leírásával lehetőség nyílik a biológiai folyamatok új szemszögből történő vizsgálatára. Az újonnan keletkező fehérjék szerkezeti preferenciáinak vizsgálata a fehérjeevolúciót új dimenzióba helyezi.

A PRIDE-NMR szerver [I, II] a távolság jellegű kényszerfeltételek alapján becsüli a fehérje térszerkezetét. Megmutattam, hogy a NOE adatok elegendő specifikus információt tartalmaznak a fehérje várható térszerkezetéről. Az NMR spektroszkópiai adatokból nyerhető paraméterek a vizsgált makromolekula térszerkezete mellett a belső dinamikájáról, a mozgékonyságáról is szolgáltatnak információkat. Az a tény, hogy a molekulát oldatban vizsgáljuk, közelebb viszi az eredményeket a valós biológiai rendszerekben tapasztalható körülményekhez. A kutatók az NMR mérések alapján molekuladinamikai szimulációk során egyre több független paramétert vesznek figyelembe (NOE, $S^2$, RDC-k, J-csatolások, kémiai eltolódások, stb.) a geometriai adatokon túl. A különböző időskálákon leírt molekuláris mozgások dinamikájának figyelembevételele a szimuláció során lehetővé teszi, hogy a kapott szerkezeti sokaságok strukurális heterogenitása ne a kényszerfeltételek bizonytalanságából eredjen, hanem a molekula tényleges belső mozgékonyságát tükrözze adott időskálán.

A CoNSEnsX szerver [III] a szerkezeti sokaságokat és az NMR kísérleti adatokat felelteti meg. Az egyes konformereket külön-külön is, és a konformer sokaságot együttesen is tudja kezelni. A konformer térszerkezete alapján visszaszámolunk elméleti kísérleti adatokat adott programok segítségével, majd az így kapott adatokat összevetjük a mért kísérleti adatokkal. A módszert több példán is elemeztem, globuláris és belsőleg rendezetlen fehérjére is. A fehérjék térszerkezete a belső mozgékonyság figyelembevételével kiegészül a negyedik dimenzióval. Fontos a figyelembevett paraméterek időskáláját szem

előtt tartani, hiszen a fehérjék dinamikája több nagyságrendet ölel fel, és egyes paraméterek csak bizonyos időskálára vonatkoznak.

A *de novo* fehérjék vizsgálatára felépítettem egy *in silico* modellt, a biológiai információáramlás mintájára [IV]. A GC-tartalom inkrementális változtatásával igyekeztem minél szélesebb körben és szisztematikusan lefedni a biológiai körülmények között szintetizálódható véletlen szekvenciák terét. Olyan szekvencia alapú prediktorokat használtam, amelyek evolúciós és hasonlósági információt nem vesznek figyelembe. Összeállítottam egy konszenzus predikciós protokollt három prediktort használva a rendezetlenségi, a transzmembrán hélixképző és az aggregációs hajlam becslésére. A *de novo* fehérjék esetében az aggregációs hajlam volt az egyik kulcskérdés a munka megkezdésekor. Jim Schnabel cikkében Christopher Dobsont idézi, miszerint ha véletlenszerűen generálnánk fehérjeszekvenciákat, nagyon ritkán kapnánk stabil, oldható fehérjét[1] [9]. A random szekvenciák konszenzus predikciós elemzése rávilágított a szerkezeti tulajdonságok GC-függésére. A GC-tartalom meghatározza a fehérjeszekvenciák aminosav-összetétét és ezen keresztül a térszerkezeti preferenciáit a standard genetikai kód alapján. A rendezetlenségi hajlam növekszik a GC-tartalommal, ellentétben a transzmembrán hélixképző és az aggregációs hajlammal.

Továbbá a konszenzus predikciók robusztusságát is megvizsgáltam. A valós fehérjeszekvenciák elemzése nagyon jól visszaadta a várt tendenciákat. A humán proteom és a biológiailag releváns 40 és 60% közötti GC-tartalmú random szekvenciák predikcióinak összevetése új megvilágításba helyezi az egész térszerkezeti jellemzők közötti összefüggéseket. Ugyanakkor az adataim a *de novo* szintetizálódó fehérjékre vonatkoznak, amelyek még semmilyen kölcsönhatásba nem léptek sem egymással, sem más sejtkomponenssel. Az irodalomban leírt védekezési mechanizmusok [90] tehát fontosak, sőt a szintetizált fehérje mennyisége is meghatározó lehet. Az újonnan szintetizálódó fehérjék, ha először nagyon kis mennyiségben jelennek meg, feltételezhetően kisebb veszélyt jelentenek a sejtre még komoly aggregációs hajlam esetén is, mint a nagy mennyiségben folyamatosan szintetizálódó fehérjék.

A biológiai folyamatokban tehát a mozgékonyság és a különböző időskálájú dinamikus viselkedés ismerete elengedhetetlen, hogy a megfigyelt jelenségekre kielégítő szerkezeti vagy funkcionális modelleket tudjunk építeni. Doktori munkám során két szervert fejlesztettem ki és egy új szemléletű *in silico* szerkezeti elemzést végeztem el, a fehérjék térszerkezetét és evolúcióját vizsgálva a mozgékonyság tükrében.

[1] Most modern proteins fold into globular structures. But their folding patterns are so complex that they couldn't have evolved by accident. "If you had a machine that could generate protein sequences randomly, you would only rarely get one that can remain stable in the globular, soluble state," Dobson says.

# Rövidítések jegyzéke

| | |
|---|---|
| <C> | átlagos nettó töltés |
| <H> | átlagos normalizált hidrofobicitási érték |
| BLAST | Basic Local Alignement Tool |
| BMRB | biológiai NMR-adatok adatbázisa (Biological Magnetic Resonance data Bank) |
| CATH | térszerkezet-osztályozó adatbázis (Class - Architecture - Topology - Homology) |
| COCO | (COmplementary Coordinates) |
| COSY | korrelált spektroszkópia (COrrelated SpectroscopY) |
| DER | NMR alapú, távolság és dinamikai jellegű kényszerfeltételeket figyelembe vevő fehérjeszerkezet számoló eljárás (Dynamic Ensemble Refinement) |
| EROS | NMR alapú, kizárólag RDC adatokat mint kényszerfeltételeket figyelembe vevő fehérjeszerkezet számoló eljárás (Ensemble Refinement with Orientational Restraints) |
| HMM | rejtett Markov modell (Hidden Markov Model) |
| IDP | belsőleg rendezetlen fehérjék (Intrinsically Disordered Proteins) |
| ISD | Inferential Structure Determination |
| MUMO | NMR alapú, távolság és dinamikai jellegű kényszerfeltételeket figyelembe vevő fehérjeszerkezet számoló eljárás (Minimal Under-restraining Minimal Over-restraining) |
| NMR | mágneses magrezonancia (Nuclear Magnetic Resonance) |
| NNR | NOE, NH $S^2$ és N-H RDC-k felhasználásával számolt szerkezeti sokaság |
| NOE | atommagok között fellépő Overhauser effektus (Nuclear Overhauser Effect) |
| NOESY | atommagok között fellépő Overhauser effektus (Nuclear Overhauer Effect SpectroscopY) |
| OPLS-AA | minden atomot figyelembe vevő molekulamechanikai erőtér (Optimized Potential for Liquid Simulations, All Atoms) |
| ORF | (Open Reading Frame) |

| | |
|---|---|
| PSSM | pozíció-specifikus pontozó mátrix (Position-Specific Scoring Matrix) |
| RDC | reziduális dipoláris csatolások (Residual Dipolar Coupling) |
| RECOORD | újraszámolt NMR-szerkezeteket tartalmazó adatbázis (REcalculated COOR-Dinates |
| RFAC | NMR-szerkezet minőségellenőrző eljárás (automated NMR R-FACtor es-timation) |
| RMSD | az átlagos négyzetes eltérés négyzetgyöke (Root Mean Square Deviation) |
| RPR | (aminosavankénti átlagos kényszerfeltétel szám (Restraints Per Residue) |
| $S^2$ | általános rendparaméter |
| SCOP | térszerkezet-oszályozó adatbázis (Structural Classification Of Proteins) |
| SCR | Single Conformer Refinement |
| SH3 | Src Homology 3 domén |
| SPC | Single Point Charge |
| SVD | Single Value Decomposition |
| TOCSY | teljesen korrelált spektroszkópia (TOtal Correlated SpectroscopY) |
| X-PLOR | makromolekulák röntgendiffrakciós vagy NMR-spektroszkópiai adatokból történő térszerkezet-számolásra alkalmas program |

# Ábrák jegyzéke


1.1. A fehérje-ligand kötődési mechanizmus modelljei (Vértessy & Orosz után) . . . 2
1.2. Ismert fehérje szerkezeti struktúrák . . . . . 3

2.1. A fehérjék szerkezeti szintjei . . . . . 5
2.2. Az NMR spektroszkópia alapú térszerkezet-meghatározás folyamatábrája . . . . 7
2.3. Az NMR adatokból kinyerhető általános paraméterek és jellemzőik . . . . . 10
2.4. Az NMR spektroszkópiai mérések időskálája és néhány, az adott időtartományhoz kapcsolható molekuláris mozgás . . . . . 11
2.5. A fehérje belső dinamikáját tükröző NMR adatokat figyelembevevő néhány szerkezetszámolási protokoll . . . . . 12
2.6. Precizitás és pontosság modellek esetében . . . . . 13

5.1. A PRIDE-NMR módszer folyamatábrája . . . . . 20
5.2. A szerkezetekből visszaszámolt (d=5Å és d=6Å küszöbérték mellett) véletlenszerűen csonkított távolságeloszlások és a véletlenszerűen csonkított NOE távolságeloszlásokra kapott pozitív találati arányok összehasonlítása . . . . . 24
5.3. A PRIDE-NMR szerver . . . . . 27
5.4. Az 1GBR állomány kényszerfeltételek keresési eredménye (d=5Å, W=3) . . . . 29
5.5. A humán SH3 doménre talált hasonló szerkezetek (d=5Å, W=3) . . . . . 29
5.6. A CoNSEnsX szerver elvi felépítése . . . . . 36
5.7. A CoNSEnsX szerver megjelenítése . . . . . 37
5.8. Humán ubiquitin dinamikus szerkezeti sokaságok: a) U_COCO; b) U_NNR; c) 1UBQ_MD . . . . . 38
5.9. Az ubiquitin (U_NNR sokaság) CoNSEnsX elemzés kimenete . . . . . 41
5.10. Ubiquitin szerkezeti sokaságok és az összeállított kísérleti adatkészlet összevetéseinek eredményei . . . . . 42
5.11. PDE 5/6 $\gamma$-alegység konformereinek illesztése és ábrázolása . . . . . 43
5.12. PDE 5/6 $\gamma$-alegység korrelációs adatai . . . . . 44

7.1. A molekuláris biológia centrális dogmája . . . . . 50
7.2. Standard genetikai kód elrendeződése és a triplettek által kódolt aminosavak kémiai képletei . . . . . 51

9.1. Random fehréjeszekvenciák generálása a biológiai folyamatot modellezve . . . . 59

10.1. Véletlen szekvenciák aminosav-összetéte . . . . . 68
10.2. Adatbázisok aminosav-összetétele . . . . . 68
10.3. A GC=(10-90)% és a biológiailag releváns GC-tartományok fehérjeszekvenciák töltés-hidrofobicitás plot-ja . . . . . 69
10.4. A GC=(10-90)% és a valós fehérjeszekvenciák töltés-hidrofobicitás plot-ja . . . 69
10.5. Rendezetleségi hajlamra kapott konszenzus predikciók grafikus összefoglalása . . 71
10.6. Transzmembrán hélixképző hajlamra kapott konszenzus predikciók ábrázolása . . 72
10.7. Aggregációs hajlamra kapott konszenzus predikciók grafikus összefoglalása . . . 73
10.8. Rendezetlenségi hajlam konszenzus predikciók a valós adatkészletekre . . . . . 74
10.9. Transzmembrán hélixképző konszenzus predikciók a valós adatkészletekre . . . . 75
10.10Aggregációs hajlam konszenzus predikciók a valós adatkészletekre . . . . . 75
10.11Kétdimenziós ábrázolás a szekvenciák konszenzus (rendezetlen-transzmembrán) predikcióira . . . . . 79
10.12Kétdimenziós ábrázolás a szekvenciák konszenzus predikcióira. . . . . 80
10.13A lefedett területek összehasonlítása, 0.1%-os felbontásban. . . . . 83
10.14Rendezetlenségi, transzmembrán és aggregációs hajlamok a humán és a random (shuffled) humán proteomra . . . . . 85
10.15A teljes random adatkészlet, a humán proteom és az árva humán fehérjék ábrázolása a C-H diagramon . . . . . 87

# Táblázatok jegyzéke


5.1. Pozitív találatok aránya százalékosan a 40 fehérjés tesztkészletre . . . . . 23
5.2. PRIDE-NMR találati statisztika a PDB adatbázisra . . . . . 26
5.3. Példák a PRIDE-NMR keresés eredménytelenségére . . . . . 30
5.4. A PRIDE-NMR keresés eredményei az ubiquitin (PDB kód: 1D3Z) kényszerfeltétel-lista esetén, d = 5 Å, valamint d = 5 és 6 Å küszöbértékek mellett . . . . . 34
5.5. Az elemzésben használt humán ubiquitin szerkezeti sokaságok . . . . . 38

7.1. Irodalomban leírt randomizált fehérjeszekvenciák kísérleteinek összefoglalása . . 54

9.1. A használt adatbázisok összetétele és jellemzői . . . . . 58
9.2. A predikciók kiértékelésére alkalmazott küszöbértékek . . . . . 64

10.1. BLAST homológia-keresés eredménye . . . . . 67
10.2. Konszenzus rendezetlenségi hajlam predikciók átlagai az adatkészletekre . . . . 71
10.3. Konszenzus transzmembrán hélix predikciók átlagai az adatkészletekre . . . . . 72
10.4. Konszenzus aggregációs hajlam predikciók átlagai az adatkészletekre . . . . . 73
10.5. Korrelációk a biológiailag releváns GC-tartományú random és a humán adatokra 77
10.6. Korrelációk a random adatkészletekre . . . . . 77
10.7. Korrelációk a valós adatbázisokra . . . . . 78
10.8. 1D Kolomorov-Smirnov korrelációk a random és a humán adatok predikciói között 81
10.9. 2D Kolomorov-Smirnov korrelációk a random és a humán adatok predikciói között 82
10.10 Becsült átfedések a kiválasztott rendezetlen-transzmembrán-aggregáció adatsorok között . . . . . 82
10.11 Korrelációk a valós adatbázisokra . . . . . 84
10.12 Humán *de novo* fehérjék konszenus szerkezeti predikciói és számolt GC-tartalmuk 87

# Kivonat

A fehérjék térszerkezetének meghatározása NMR spektroszkópiával hosszadalmas feladat. A folyamat kulcslépése a szerkezetre jellemző kényszerfeltételek iteratív felhasználása. Munkám során egy olyan egyszerű statisztikai alapokon nyugvó algoritmus kidolgozását tűztem ki célul, mely csupán a távolság jellegű kényszerfeltételek alapján képes pillanatok alatt rokonítani a kérdéses makromolekulát már ismert térszerkezetű fehérjékkel. A PRIDE-NMR eljárásban a NOE adatokból kapott H-H távolságeloszlásokat vetem össze az NMR szerkezetek térbeli koordinátái alapján különböző küszöbtávolságok mellett számolt H-H eloszlásokkal. A statisztikai összevetés kontingencia-analízissel történt. A kifejlesztett módszer segítségével kizárólag az NMR kényszerfeltételek alapján egy első képet kapunk a vizsgált fehérje térszerkezetéről. Emellett az eljárás alkalmas az NMR kényszerfeltételek teljességének és pontosságának ellenőrzésére is. A biomolekulák molekuladinamikai szimulációira fejlesztett újabb módszerek a molekula belső dinamikáját leíró paramétereket is implementátak a szerkezetszámolásba, mint az $S^2$ rendparaméterrel, a klasszikusan használt NOE (és RDC) adatokon túl. A CoNSEnsX módszer segítségével az NMR spektroszkópai adatok alapján meghatározott szerkezeti sokaságok pontosságát lehet becsülni. A program minden kísérleti adatot megfelelteti a szerkezeti sokaság alapján visszaszámolt megfelelő elméleti paraméternek. Mindkét program (PRIDE-NMR és CoNSEnsX) szerverként elérhető a világhálón.

Evolúciós skálán döntő szerepet játszik a fehérjék konformációs változatossága, dinamikája, ami kihat a biológiai rendszerekben betöltött funkcióra. A nemkódoló DNS átítódása révén *de novo* fehérjék keletkezhetnek. Szisztematikusan változtatott GC-tartalom mellett *in silico* random DNS szegmenseket állítottam elő, lefordítottam, és a kapott 160 aminosavas fehérjeszekvenciákat konszenzus predikcióknak vetettem alá, hogy a rendezetlenségi, a transzmembrán hélixképző és az aggregációs hajlamát becsüljem. Világos trendeket azonosítottam a GC-tartalom függvényében: a magas GC-tartalmú fehérjék inkább rendezetlenek és alacsony a transzmembrán hélixképző és az aggregációs hajlamuk. A három vizsgált szerkezeti tulajdonság által behatárolt térrész majdnem teljesen egybeesik a 40, 50 és 60%-os random szekvenciák és a humán proteom fehérjéi esetében. A legnagyobb eltérés a humán proteomok nagyobb transzmembrán hélix tartalmában volt. Az eredményeim azt mutatják, hogy a *de novo* keletkező fehérjék számára az aggregáció nem akkora veszély, legalábbis nem nagyobb, mint a valós fehérjékre nézve.

# Abstract

Protein NMR structures are usually determined by computer-intensive simulation protocols using the experimentally determined NMR restraints. This process could be speed up and assisted by an initial estimation of the fold using NMR data. PRIDE-NMR is a fast novel method to relate known protein folds using NMR distance restraints based on comparing the distributions of backbone H-H distances . Distance distributions are compared by a robust statistical test against a filtered database. The improved algorithm can be used to obtain a first guess about a structure being determined, as well as to estimate the completeness or verify the correctness of NOE data. Traditional structure calculation techniques use only NOE (and RDC) data and the resulting ensemble is usually not in compliance with dynamical parameters such as backbone $S^2$ values. Recent developments in molecular simulations of biomolecules allow to incorporate multiple experiment-derived restraints yielding a conformational ensemble revealing the internal dynamics of the system. We have developed a tool named CoNSEnsX to assess of the accuracy of protein structural ensembles determined by NMR spectroscopy. The program outputs measures of correspondence between the available experimental parameters and the their back-calculated counterparts based on the submitted structures. Both programs (PRIDE-NMR and CoNSEnsX) are available as webservers.

At evolutionary scale, protein conformational diversity play a crucial role and protein dynamics has a strong effect on its biological function. Proteins can emerge *de novo* from the translation of previously noncoding DNA segments. I have generated *in silico* random DNA segments with systematically altered GC-content, translated them, and subjected the resulting 160-residue protein sequences to consensus predictions to assess their propensity to form disordered regions, transmembrane helices and aggregates. I have identified clear trends in the investigated properties based on the GC-content of the underlying DNA segments: polypeptides translated from segments of high GC content tend to be disordered and not prone to aggregation or to form transmembrane helices. The three-dimensional region defined by the three properties of sequences translated from random DNA with a GC-content of 40, 50 and 60% lies practically entirely within that spanned by the properties of the human proteome. The largest observed difference between these two data sets is the higher occurrence of transmembrane helices in the human proteome. My results suggest that aggregation is not a serious risk for *de novo* proteins, at least not higher than for extant ones.

# Irodalomjegyzék


[1] Vértessy BG, Orosz F. From "fluctuation fit" to "conformational selection": evolution, rediscovery, and integration of a concept. *Bioessays*, **33**:30–34 (**2011**).

[2] Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol*, **5**:789–796 (**2009**).

[3] Fenwick RB, Esteban-Martín S, Salvatella X. Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur Biophys J*, **40**:1339–1355 (**2011**).

[4] Lange OF, Lakomek NA, Fares C, Schroeder GF, Walter KFA, Becker S, Meiler J, Grubmüller H, Griesinger C, de Groot BL. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science*, **320**:1471–1475 (**2008**).

[5] Gáspári Z, Várnai P, Szappanos B, Perczel A. Reconciling the lock-and-key and dynamic views of canonical serine protease inhibitor action. *FEBS Lett*, **584**:203–206 (**2010**).

[6] Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. *Nature*, **433**:128–132 (**2005**).

[7] Szappanos B, Süveges D, Nyitray L, Perczel A, Gáspári Z. Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett*, **584**:1623–1627 (**2010**).

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

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

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

[11] Berg J, Tymoczko J, Stryer L. *Biochemistry (5. kiadás)*. W.H Freeman and Company, New York (**2002**).

[12] Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on C($\alpha$)-C($\alpha$) distance comparison. *J Mol Biol*, **315**:887–898 (**2002**).

[13] Lukács O. *Matematikai statisztika*. Bolyai könyvek (**1996**).

[14] Roberts G, Lian LY, editors. *Protein NMR Spectroscopy: Principal Techniques and Applications*. Wiley; 1 edition (**2011**).

[15] Hore P. *Nuclear Magnetic Resonance*. Oxford University Press (**1995**).

[16] Perczel A, Laczkó I, Hollósi M. *Peptidek térszerkezet-vizsgálata*. A kémia újabb eredményei 77. Akadémiai kiadó, Budapest (**1994**).

[17] Noggle J, Schirmer R. *The Nuclear Overhauser Effect*. Chemical Applications, Acad. Press, New York (**1971**).

[18] Wüthrich K. *NMR of proteins and nucleic acids*. John Wiley & Sons, Inc, New York (**1986**).

[19] Bourne P, Wessig H. *Structural Bioinformatics*, volume Chapter 16. John Wiley & Sons, Inc, New York (**2003**).

[20] Brünger AT. *X-PLOR Manual (version 4.0)*. Department of Molecular Biophysics and Biochemistry, Yale University (**1996**).

[21] Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, **54**:905–921 (**1998**).

[22] Gáspári Z, Perczel A. Biomolecular dynamics as reported by NMR. *Annual reports on NMR spectroscopy*, pp. 35–75 (**2010**).

[23] Jarymowycz VA, Stone MJ. Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences. *Chem. Rev.*, **106**:1624–1671 (**2006**).

[24] Annila A, Permi P. Weakly aligned biological macromolecules in dilute aqueous liquid crystals. *Concepts Magn. Reson.*, **23A**:22–34 (**2004**).

[25] Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci U S A*, **104**:9615–9620 (**2007**).

[26] Wylie BJ, Sperling LJ, Nieuwkoop AJ, Franks WT, Oldfield E, Rienstra CM. Ultrahigh resolution protein structures using NMR chemical shift tensors. *Proc Natl Acad Sci U S A*, **108**:16974–16979 (**2011**).

[27] Delaglio F, Kontaxis G, A B. Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc*, **122**:2142–2143 (**2000**).

[28] Lindorff-Larsen K, Roggen P, Paci E, Vendruscolo M, Dobson CM. Protein folding and the organization of the protein topology universe. *Trends Biochem Sci*, **30**:13–19 (**2005**).

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

[30] Clore GM, Schwieters CD. Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small $\alpha$/beta protein: a unified picture of high probability, fast atomic motions in proteins. *J Mol Biol*, **355**:879–886 (**2006**).

[31] Best RB, Vendruscolo M. Determination of protein structures consistent with NMR order parameters. *J Am Chem Soc*, **126**:8090–8091 (**2004**).

[32] Hess B, Scheek RM. Orientation restraints in molecular dynamics simulations using time and ensemble averaging. *J Magn Reson*, **164**:19–27 (**2003**).

[33] Gáspári Z, Perczel A. A fehérjemolekulák belső dinamikája. *Természet Világa*, **2**:59–61 (**2011**).

[34] Gronwald W, Kirchhöfer R, Görler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR. RFAC, a program for automated NMR R-factor estimation. *J Biomol NMR*, **17**:137–151 (**2000**).

[35] Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**:477–486 (**1996**).

[36] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, **58**:899–907 (**2002**).

[37] Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CAEM, Nabuurs SB, Güntert P, Livny M, Markley JL, Nilges M, et al. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins*, **59**:662–672 (**2005**).

[38] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**:536–540 (**1995**).

[39] Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, **28**:254–256 (**2000**).

[40] Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. *Proteins*, **54**:260–270 (**2004**).

[41] Pearl FMG, Lee D, Bray JE, Buchan DWA, Shepherd AJ, Orengo CA. The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Sci*, **11**:233–244 (**2002**).

[42] Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**:739–747 (**1998**).

[43] Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**:123–138 (**1993**).

[44] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol*, **6**:377–385 (**1996**).

[45] Gáspári Z, Vlahovicek K, Pongor S. Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. *Bioinformatics*, **21**:3322–3323 (**2005**).

[46] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. *J Comput Chem*, **26**:1701–1718 (**2005**).

[47] Jorgensen W, Maxwell D, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc*, **118**:11225–11236 (**1996**).

[48] Batta G, Barna T, Gáspári Z, Sándor S, Kövér KE, Binder U, Sarg B, Kaiserer L, Chhillar AK, Eigentler A, et al. Functional aspects of the solution structure and dynamics of PAF–a highly-stable antifungal protein from Penicillium chrysogenum. *FEBS J*, **276**:2875–2890 (**2009**).

[49] Cornilescu G, Macquardt J, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc*, **120**:6836–6837 (**1998**).

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

[51] Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD. $^{1}$H, $^{13}$C and $^{15}$N chemical shift referencing in biomolecular NMR. *J Biomol NMR*, **6**:135–140 (**1995**).

[52] Zweckstetter M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc*, **3**:679–690 (**2008**).

[53] Vranken W. A global analysis of NMR distance constraints from the PDB. *J Biomol NMR*, **39**:303–314 (**2007**).

[54] Wittekind M, Mapelli C, Farmer B 2nd, Suen KL, Goldfarb V, Tsao J, Lavoie T, Barbacid M, Meyers CA, Mueller L. Orientation of peptide fragments from Sos proteins bound to the N-terminal SH3 domain of Grb2 determined by NMR spectroscopy. *Biochemistry*, **33**:13531–13539 (**1994**).

[55] Zargovic B, van Gunsteren W. Comparing atomistic simulation data with the NMR experiment: How much can NOE-s actually tell us? *Proteins*, **63**:210–218 (**2006**).

[56] Vijay-Kumar S, Bugg C, Cook W. Structure of ubiquitin refined at 1.8Åresolution. *J Mol Biol*, **194**:531–544 (**1987**).

[57] Wang A, Bax A. Determination of the backbone dihedral angle $\phi$ in human ubiquitin from reparametrized Karplus equations. *J. Am. Chem. Soc.*, **118**:2483–2494 (**1996**).

[58] Rieping W, M N, M H. ISD: a software package for Bayesian NMR structure calculation. *Bioinformatics*, **24**:1104–1105 (**2008**).

[59] Babu C, Flynn P, Wand J. Validation of protein structure from preparations of encapsulated proteins dissolved in low viscosity fluids. *J Am Chem Soc*, **123**:2691–2692 (**2001**).

[60] Kitahara R, S Y, R A. NMR snapshots of a fluctuating protein structure: ubiquitin at 30 bar - 3 kbar. *J Mol Biol*, **347**:277–285 (**2005**).

[61] Manolikas T, T H, BH M. Protein structure determination from $^{13}$C spin-diffusion solid-state NMR spectroscopy. *J Am Chem Soc*, **130**:3959–3966 (**2008**).

[62] Laughton CA, Orozco M, Vranken W. COCO: a simple tool to enrich the representation of conformational variability in NMR structures. *Proteins*, **75**:206–216 (**2009**).

[63] Koradi R, Billeter M, Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph*, **14**:51–5, 29–32 (**1996**).

[64] Chang SL, Tjandra N. Temperature dependence of protein backbone motion from carbonyl $^{13}$C and amide $^{15}$N NMR relaxation. *J Magn Reson*, **174**:43–53 (**2005**).

[65] Wand AJ, Urbauer JL, McEvoy RP, Bieber RJ. Internal dynamics of human ubiquitin revealed by $^{13}$C-relaxation studies of randomly fractionally labeled protein. *Biochemistry*, **35**:6116–6125 (**1996**).

[66] Permi P. Measurement of residual dipolar couplings from $^{1}$H$\alpha$ to $^{13}$C$\alpha$ and $^{15}$N using a simple HNCA-based experiment. *J Biomol NMR*, **27**:341–349 (**2003**).

[67] Würtz P, Fredriksson K, Permi P. A set of HA-detected experiments for measuring scalar and residual dipolar couplings. *J Biomol NMR*, **31**:321–330 (**2005**).

[68] Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci*, **27**:527–533 (**2002**).

[69] Song J, Guo LW, Muradov H, Artemyev NO, Ruoho AE, Markley JL. Intrinsically disordered gamma-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure. *Proc Natl Acad Sci U S A*, **105**:1505–1510 (**2008**).

[70] Crick F. Central dogma of molecular biology. *Nature*, **227**:561–563 (**1970**).

[71] Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, **61**:99–111 (**2009**).

[72] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **40**:D71–D75 (**2012**).

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

[74] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, et al. Intrinsically disordered protein. *J Mol Graph Model*, **19**:26–59 (**2001**).

[75] Dyson HJ. Expanding the proteome: disordered and alternatively folded proteins. *Q Rev Biophys*, **44**:467–518 (**2011**).

[76] Tompa P. Unstructural biology coming of age. *Curr Opin Struct Biol*, **21**:419–425 (**2011**).

[77] Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *J Biomed Biotechnol*, **2010**:568068 (**2010**).

[78] Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput*, pp. 89–100 (**2001**).

[79] Tompa P, Dosztanyi Z, Simon I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J Proteome Res*, **5**:1996–2000 (**2006**).

[80] Oldfield CJ, Ulrich EL, Cheng Y, Dunker AK, Markley JL. Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **59**:444–453 (**2005**).

[81] Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder–a breakthrough invention of evolution? *Curr Opin Struct Biol*, **21**:412–418 (**2011**).

[82] Zhao G, London E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Sci*, **15**:1987–2001 (**2006**).

[83] Ahram M, Litou ZI, Fang R, Al-Tawallbeh G. Estimation of membrane proteins in the human proteome. *In Silico Biol*, **6**:379–386 (**2006**).

[84] Dobson CM. Protein folding and misfolding. *Nature*, **426**:884–890 (**2003**).

[85] Perczel A, Hudáky P, Pálfi VK. Dead-end street of protein folding: thermodynamic rationale of amyloid fibril formation. *J Am Chem Soc*, **129**:14959–14965 (**2007**).

[86] Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*, **5**:15–22 (**2009**).

[87] Stefani M. Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world. *Biochim Biophys Acta*, **1739**:5–25 (**2004**).

[88] Baldwin AJ, Knowles TPJ, Tartaglia GG, Fitzpatrick AW, Devlin GL, Shammas SL, Waudby CA, Mossuto MF, Meehan S, Gras SL, et al. Metastability of native proteins and the phenomenon of amyloid formation. *J Am Chem Soc*, **133**:14160–14163 (**2011**).

[89] Monsellier E, Ramazzotti M, de Laureto PP, Tartaglia GG, Taddei N, Fontana A, Vendruscolo M, Chiti F. The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys J*, **93**:4382–4391 (**2007**).

[90] Reumers J, Rousseau F, Schymkowitz J. Multiple evolutionary mechanism reduce protein aggregation. *The open Biology Journal*, **2**:176–184 (**2009**).

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

[92] Rousseau F, Serrano L, Schymkowitz JWH. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol*, **355**:1037–1047 (**2006**).

[93] Tartaglia GG, Pellarin R, Cavalli A, Caflisch A. Organism complexity anti-correlates with proteomic beta-aggregation propensity. *Protein Sci*, **14**:2735–2740 (**2005**).

[94] Monsellier E, Ramazzotti M, Taddei N, Chiti F. Aggregation propensity of the human proteome. *PLoS Comput Biol*, **4**:e1000199 (**2008**).

[95] Guerzoni D, McLysaght A. De novo origins of human genes. *PLoS Genet*, **7**:e1002381 (**2011**).

[96] Bornberg-Bauer E, Kramer L. Robustness versus evolvability: a paradigm revisited. *HFSP J*, **4**:105–108 (**2010**).

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

[98] Toll-Riera M, Castelo R, Bellora N, Alba MM. Evolution of primate orphan proteins. *Biochem Soc Trans*, **37**:778–782 (**2009**).

[99] Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X, et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol*, **6**:e1000734 (**2010**).

[100] Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. *PLoS Genet*, **7**:e1002379 (**2011**).

[101] Siepel A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res*, **19**:1693–1695 (**2009**).

[102] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*, **20**:1313–1326 (**2010**).

[103] The E N C O D E Project Consortium. Identification and analysis of functional elements in 1genome by the ENCODE pilot project. *Nature*, **447**:799–816 (**2007**).

[104] Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*, **25**:404–413 (**2009**).

[105] Davidson AR, Sauer RT. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc Natl Acad Sci U S A*, **91**:2146–2150 (**1994**).

[106] Doi N, Kakukawa K, Oishi Y, Yanagawa H. High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng Des Sel*, **18**:279–284 (**2005**).

[107] Keefe AD, Szostak JW. Functional proteins from a random-sequence library. *Nature*, **410**:715–718 (**2001**).

[108] Ito Y, Kawama T, Urabe I, Yomo T. Evolution of an arbitrary sequence in solubility. *J Mol Evol*, **58**:196–202 (**2004**).

[109] Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A, Shima Y, Negoro S, Urabe I. Solubility of artificial proteins with random sequences. *FEBS Lett*, **382**:21–25 (**1996**).

[110] Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL. Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chem Biodivers*, **3**:840–859 (**2006**).

[111] Minervini G, Evangelista G, Villanova L, Slanzi D, De Lucrezia D, Poli I, Luisi PL, Polticelli F. Massive non-natural proteins structure prediction using grid technologies. *BMC Bioinformatics*, **10**:S22 (**2009**).

[112] Schaefer C, Schlessinger A, Rost B. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics*, **26**:625–631 (**2010**).

[113] Pál G, Kouadio JLK, Artis DR, Kossiakoff AA, Sidhu SS. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J Biol Chem*, **281**:22378–22385 (**2006**).

[114] Watters AL, Baker D. Searching for folded proteins in vitro and in silico. *Eur J Biochem*, **271**:1615–1622 (**2004**).

[115] Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. *Nucleic Acids Res*, **32**:D189–D192 (**2004**).

[116] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, et al. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res*, **35**:D786–D793 (**2007**).

[117] Tusnády GE, Dosztányi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, **33**:D275–D278 (**2005**).

[118] Pawlicki S, Le Béchec A, Delamarche C. AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics*, **9**:273 (**2008**).

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

[120] Shen MY, Davis F, Sali A. The optimal size of a globular protein domain: A simple sphere-packing model. *Chem. Phys. Letters*, **405**:224–228 (**2005**).

[121] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, **215**:403–410 (**1990**).

[122] Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, **87**:2264–2268 (**1990**).

[123] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**:105–132 (**1982**).

[124] Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*, **11**:739–756 (**2002**).

[125] Pirovano W, Heringa J. Protein secondary structure prediction. *Methods Mol Biol*, **609**:327–348 (**2010**).

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

[127] Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**:3369–3376 (**2005**).

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

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

[130] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**:567–580 (**2001**).

[131] Cserző M, Eisenhaber F, Eisenhaber B, Simon I. On filtering false positive transmembrane protein predictions. *Protein Eng*, **15**:745–752 (**2002**).

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

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

[134] Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, **26**:326–332 (**2010**).

[135] Press W, Teukolsky S, Vetterling W, BP F. *Numerical recipes in C: the art of scientific computing*, chapter Are two distributions different?, pp. 623–649. Cambridge University Press (**1988-1992**).

[136] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature*, **409**:860–921 (**2001**).

[137] Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, Bresolin N, Sironi M. Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol*, **8**:99 (**2008**).

[138] Junkes T, Holmquist R, Moise H. Amino acid composition of proteins: selection against the genetic code. *Science*, **189**:50–51 (**1975**).