

—— XXXI. Országos Tudományos Diákköri Konferencia ——
—— Kémiai és Vegyipari Szekció ——

KOVÁCS BERTALAN
Vegyész MSc. II. évfolyam

Béta-redők stabilitásvizsgálata párkorrelációk statisztikai és kvantumkémiai modellezésével

Témavezetők:

Dr. Jákli Imre, tudományos főmunkatárs
Dr. Perczel András, egyetemi tanár

ELTE Szerves Kémiai Tanszék



E ö t v ö s L o r á n d
T u d o m á n y e g y e t e m
Természettudományi Kar

Köszönetnyilvánítás

A dolgozat létrejöttéért szeretnék köszönetet mondani elsősorban témavezetőimnek, Dr. Jákli Imrének és Dr. Perczel Andrásnak, akik lehetővé tették számomra, hogy az MTA Fehérjemodellező Kutatócsoportban dolgozhassak. Köszönöm nekik türelmüket, az átadott tudást és bölcsességet, valamint azt a kiváló szakmai háttérrel, amely számomra a mai napig nélkülözhetetlen a tudományos világban való eligazodáshoz. Szintén szeretnék köszönetet mondani a kutatócsoport minden tagjának, akik egy nagyszerű közösséget alkotnak, hogy ilyen remek légkörben végezhettem a kutatást.

Külön hálával tartozom Papp Dórának a dolgozat minden részletére kiterjedő lektorálásáért, valamint Dr. Harmat Veronikának a röntgenkristallográfia, Dr. Szalay Péternek a kvantumkémia, Dr. Tóth Gergelynek a statisztika, és Dr. Láng Andrásnak a bioinformatika területén nyújtott hasznos segítségért.

Tartalom

Aminosavak rövidítései	4
1. Bevezetés.....	5
2. A β -redők vizsgálatának jelentősége	6
2.1. A fehérjék szerkezete	6
2.2. A β -redők szerkezete	7
2.3. Aggregációs betegségek	9
3. A párkorreláció	10
3.1. A statisztika mint eszköz.....	10
3.2. A párkorreláció irodalmi háttere és jelentősége.....	11
3.3. Az alkalmazott kvantumkémiai módszerek.....	14
3.3.1. A Hartree-Fock-módszer	15
3.3.2. A DFT módszer.....	15
3.3.3. A BSSE hiba és a counterpoise korrekció	17
4. Problémafelvetés, célkitűzés.....	17
5. A statisztikai analízis bemutatása.....	19
5.1. A kiindulási adathalmaz.....	19
5.2. A redundanciaszűrés	20
5.3. Az SDB program.....	25
5.4. Másodlagos szerkezeti elemek meghatározása	25
5.5. A párpreferencia értékek meghatározása.....	27
5.6. Különböző adatkészletből származó párkorrelációs táblázatok összehasonlítása	28
5.7. Adatbázisok önálló jóságának jellemzése	29
6. A kvantumkémiai elemzés bemutatása.....	30
7. Az eredmények diszkussziója	32
7.1. Aminosavak eloszlása	32
7.2. A párpreferencia értékek általános tendenciái	34
7.3. NMR-es adatbázisok vizsgálata	34
7.4. Röntgenszerkezetekből álló adatbázisok vizsgálata.....	38
7.5. NMR- és röntgenszerkezetek összehasonlítása	40
7.6. Az aminosav párok gyakorisága és kölcsönhatási energiájuk	41
7.7. Aminosav párok oldallánc-orientáció eloszlása	42
7.8. A számolt geometriák összehasonlítása a statisztikai adatokkal	44

8. Összefoglalás és kitekintés	47
Irodalomjegyzék	50
Mellékletek	56

Aminosavak rövidítései

A	Ala	alanin
C	Cys	cisztein
D	Asp	aszparaginsav
E	Glu	glutaminsav
F	Phe	fenilalanin
G	Gly	glicin
H	His	hisztidin
I	Ile	izoleucin
K	Lys	lizin
L	Leu	leucin
M	Met	metionin
N	Asn	aszparagin
P	Pro	prolin
Q	Gln	glutamin
R	Arg	arginin
S	Ser	szerin
T	Thr	treonin
V	Val	valin
W	Trp	triptofán
Y	Tyr	tirozin

1. Bevezetés

Az elmúlt fél évszázad tudományos fejlődése számos újdonságot hozott többek között a biológiában is. Új módszerek születtek az élő szervezet molekuláris szintű megismerésére, és e módszerek alkalmazása során hatalmas mennyiségű új információ keletkezett. Ezzel párhuzamosan haladt előre a számítástechnika is, amely lehetővé tette, hogy egyre bonyolultabb problémákat gyorsan és egyszerűen oldjunk meg. E két tudományterület találkozása a bioinformatika, amelynek feladata elsősorban a genomiális információkat tartalmazó adatbázisok kezelése, de hozzá tartozik a számítógépes biológia valamennyi területe. Annyi bizonyos, hogy fő mozgatórugója az elérhető DNS-, RNS- és fehérjeszerkezetek és szekvenciális információk mennyiségének egyre gyorsabb növekedése (pl. Protein Data Bank).[1] Ez kihívást jelent egyrészt azoknak, akik az adatbázis vizsgálatát, rendszerezését tűzik ki célul maguk elé (fehérjék klaszterezése, szekvenciaanalízis, valószínűségi modellek készítése), és azoknak is, akik számítógépes grafika vagy szerkezetszámítás segítségével, NMR spektroszkópiai és röntgenkristallográfiai szerkezetekből kiindulva molekuláris modelleket készítenek.[2]

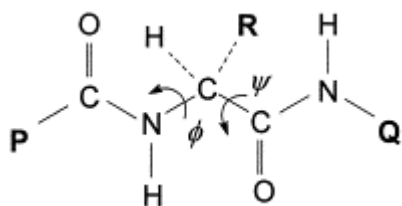
Ebben a dolgozatban mind a két megközelítéssel találkozhatunk. Látni fogjuk, hogy a β -redők komoly tudományos érdeklődés tárgyát képezik, mert főszereplői az ún. aggregációs betegségeknek (2.3. fejezet). A statisztikai vizsgálatok során azt használjuk ki, hogy a fehérjeszerkezeti adatbázisokban lévő információk lehetővé teszik számunkra, hogy ne csak egyedi molekulákat, hanem sokaságot vizsgáljunk, és következtetéseket vonjunk le az aminosavak előfordulási gyakoriságából, torziós szögek értékeinek vagy másodlagos szerkezeti elemek eloszlásából. Másrészt kihasználva a kvantumkémia adta lehetőségeket, illetve az elérhető számítási kapacitás folyamatos növekedését, egy olyan molekuláris modell építésére teszünk kísérletet, amely alá tudja támasztani a statisztikai vizsgálatok eredményét.

A bioinformatikában általában kétféle munkafázissal találkozunk: új eljárások létrehozása, amelyekkel a szerkezeti adatainkat elemezni tudjuk, illetve az eljárások alkalmazása és validálása valós rendszereken. Munkánk során egy már korábban leírt statisztikai módszert, a párkorrelációk vizsgálatát használtuk fel és fejlesztettük tovább, majd számos adatkészleten teszteltük, hogy képet kapjunk az alkalmazhatóságáról.[3]

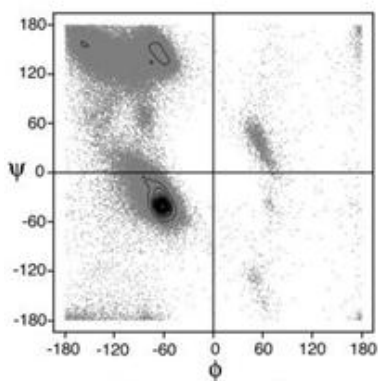
2. A β -redők vizsgálatának jelentősége

2.1. A fehérjék szerkezete

A fehérjék igen változatos molekulák, melyek az élő szervezet szinte minden folyamatában valamilyen fontos funkciót látnak el. Szerkezetüket tekintve természetes biopolimerek, melyek aminosav egységek peptidkötéssel történő összekapcsolódásával jönnek létre. Így a legtöbb fehérje látszólag egyszerűen felépíthető az ismert 20 féle fehérjealkotó aminosavból – nem említve az esetleges poszttranszlációs módosításokat, a fehérjéhez kapcsolódó szerves vagy szervetlen kofaktorokat, fémionokat – ám valójában már a legegyszerűbb fehérjék is rendkívül bonyolult háromdimenziós szerkezetet vehetnek fel, amely szerkezethez mindig valamilyen specifikus biokémiai funkció társul. Az aminosavak összekapcsolódásakor létrejövő molekulát gyakran szokás polipeptidnek hívni. Akkor beszélünk fehérjemolekuláról, ha hangsúlyozni szeretnénk, hogy a peptidlánc feltekeredett, ezáltal jól meghatározott háromdimenziós szerkezete van, amihez valamilyen specifikus biológiai funkció tartozik.[4, 5]



1. ábra: ϕ és ψ torziós szögek az IUPAC-IUB ajánlása szerint.[6]



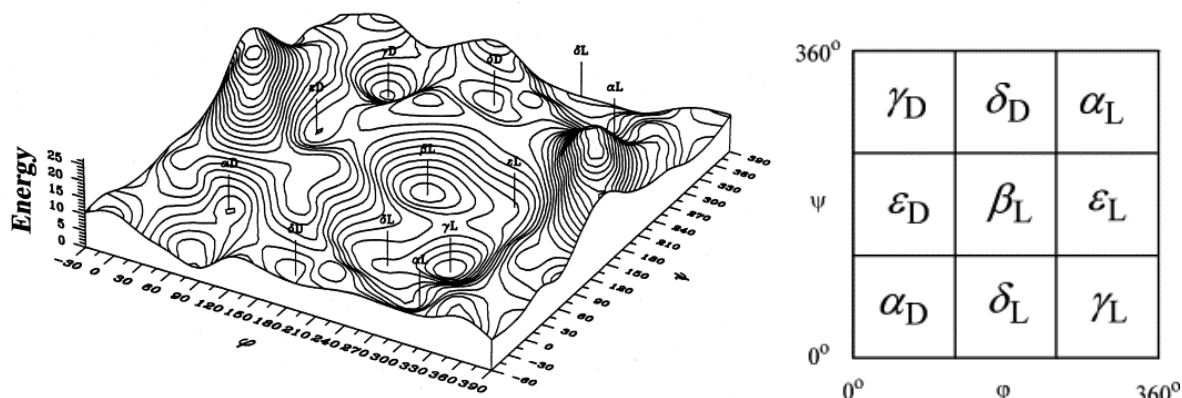
2. ábra: Ramachandran-térkép, azaz az aminosavak gyakorisága a ϕ és ψ torziós szögek függvényében.[7]

A fehérjék szerkezetét négy szinten szokás vizsgálni.

Az elsődleges szerkezet a kapcsolódó aminosavak sorrendje, azaz a fehérje szekvenciája. Másodlagos szerkezeten az egymást követő aminosavak gerinckonformációját értjük, melyeket az IUPAC-IUB ajánlása szerint a ϕ és ψ torziós szögekkel definiálunk (1.ábra).[6] Ilyen módon minden aminosav konformációja egy adott pontnak felel meg a két szög által kifeszített kétdimenziós felszínen, melyet Ramachandran-térképnek hívunk (2. ábra).[7] Ha ezek a konformációk a peptidláncon belül ismétlődnek, akkor periodikus szerkezetek, másképpen homokonformációk jönnek létre. A ϕ és ψ torziós szögek függvényében ábrázolt potenciális energia felületen, (mely gyakorlatilag az aminosavak Ramachandran-térképe), 9 minimum várható, amelyek mindegyike megfeleltethető egy-egy homokonformációnak (3. ábra).[8-10] Ezek közül

energetikailag a legkedvezőbbek, ezért a leggyakoribbak a jobbmenetes α -hélix (α_L), a β -redő (β_{DL}), és a poliprolin-II hélix (ϵ_L). A fehérjékben előfordulhatnak aperiodikus, vagyis rendezetlen régiók is, ahol az egymást követő aminosavak gerinckonformációja nem szabályszerű. A fehérjék harmadlagos szerkezetén a polipeptid lánc teljes háromdimenziós elrendeződését értjük, mely egyértelműen

megadható az atomok koordinátaival. A negyedleges szerkezet a több alegységből álló fehérjékben az egyes polipeptid láncok egymáshoz viszonyított térbeli helyzetét adja meg.



3. ábra: egy aminosav potenciális energia felülete a minimumokkal (balra), illetve a Ramachandran-térkép vázlatos felosztása a kilencféle homokonformáció régiójára [10, 11]

Meg kell említenünk, hogy léteznek rendezetlen fehérjék (*Intrinsically unstructured proteins*, IUPs), melyeknek nincsen stabil háromdimenziós szerkezetük, konformációjukat folyamatosan változtatják. Ennek ellenére minden jel szerint komoly biológiai funkcióval rendelkeznek, szerepük van sok betegségek kialakulásában is. A rendezetlen fehérjék működését még több ponton homály fedi, ezért érthető módon számos aktuális kutatás tárgyát képezik.[12-18]

2.2. A β -redők szerkezete

A β -redő egy igen stabil másodlagos szerkezeti elem. A teljesen nyújtott konformációban (*extended β -sheet*) az aminosavak torziós szögei $\Phi = -130^\circ$ és $\Psi = 130^\circ$ értékek körül alakulnak.[4] Az ilyen szerkezetű, tipikusan 3-10 aminosavból álló fehérjeszegmenst hívjuk β -szálnak (*beta-strand*).

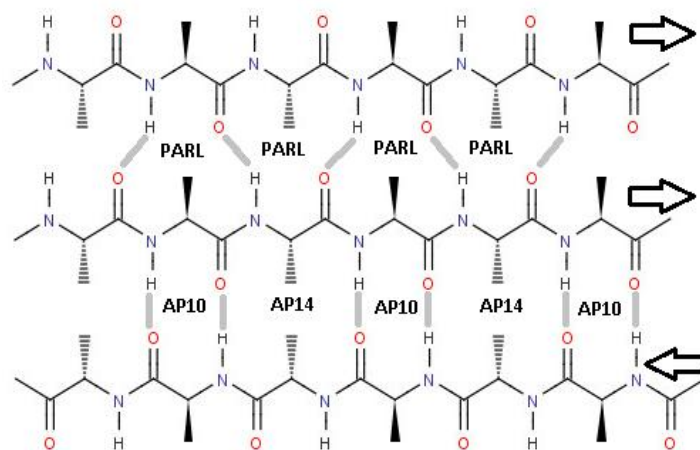
Bár a potenciális energia felületen látszik (3. ábra), hogy a β -redő lokális energiaminimumban van, a redők kiemelkedő stabilitásához elsősorban a megfelelő hidrogénkötés mintázat járul hozzá, a β -redőben ugyanis a peptidgerinc a lehető legtöbb hidrogénkötést alakítja ki, és sok esetben ez még az oldalláncokra is igaz.[19, 20] A β -szálak általában egymás mellé rendeződve fordulnak elő, hogy így a két szálnak lévő aminosavak gerincei között rendkívül stabil hidrogénkötések jöhetnek létre. A hidrogénkötés alapegységét nevezzük β -hídnak (*β -bridge*), amely létrejöttének feltétele:

$$\text{Parallel híd}(i,j) = [\text{Hbond}(i-1,j) \text{ és } \text{Hbond}(j,i+1)] \text{ vagy } [\text{Hbond}(j-1,i) \text{ és } \text{Hbond}(i,j+1)]$$

$$\text{Antiparallel híd}(i,j) = [\text{Hbond}(i,j) \text{ és } \text{Hbond}(j,i)] \text{ vagy } [\text{Hbond}(j-1,i+1) \text{ és } \text{Hbond}(i-1,j+1)]$$

ahol a két fehérjeszegmens aminosavainak sorszáma $i - 1, i, i + 1$ és $j - 1, j, j + 1$, $Hbond$ (donor, akceptor) pedig az aminosavak közötti hidrogénkötést jelöli.[21] β -létrának (β -ladder) nevezzük azt az elrendeződést, amelyben legalább két β -híd követi egymást, és β -redőnek azt, amely egy vagy több párhuzamos β -létrából áll.

Fontos hangsúlyozni a parallel és antiparallel redők topológiája közötti alapvető különbséget, amely az aminosav sorrend eltérő orientációjának következménye: a hidrogénkötés mintázat által kijelölt pszeudogyűrűket felépítő atomok száma a parallel redőkben 12, az antiparallel redőkben pedig felváltva 10 és 14 (4. ábra).[22, 23] A továbbiakban az ilyen pszeudogyűrűket kialakító aminosav párokat AP10, AP14 és PARL jelöléssel fogom megkülönböztetni. (Az irodalomban gyakran találkozunk a HB és nHB jelölésekkel az antiparallel redőre vonatkozóan, ezzel jelezve, hogy minden második aminosav alakít ki hidrogénkötést – HB, és minden második nem – nHB.[3]) A bemutatott elrendeződés következménye az is, hogy az egyes párokban különbözik egymástól a C_α atomok távolsága, amely különböző mértékű kölcsönhatást tesz lehetővé az oldalláncok között. Ezek a távolságok AP10 párokban 5,5 Å, AP14 párokban 4,5 Å, a PARL párokban pedig 5,0 Å körül alakulnak.



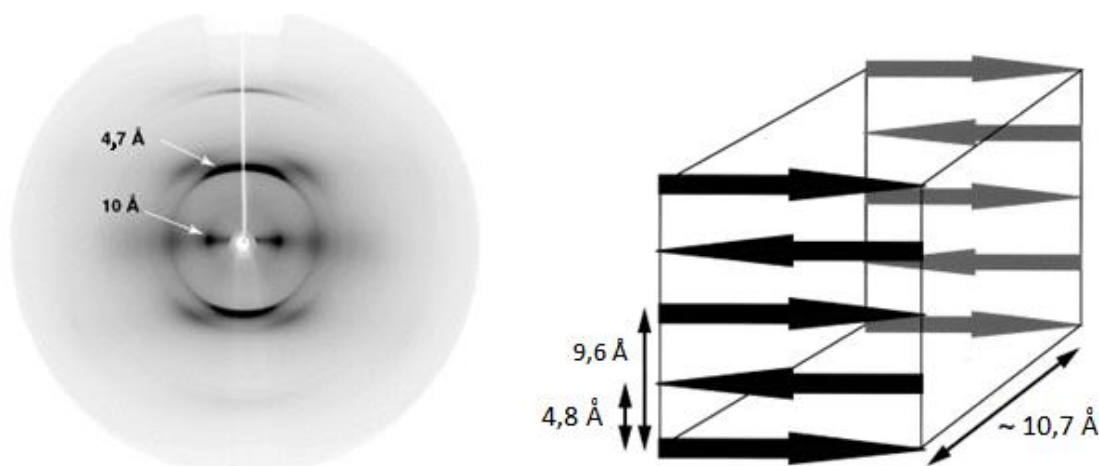
4. ábra: hidrogénkötés mintázat parallel és antiparallel β -redőkben. Az AP10 és AP14 jelölések a hidrogénkötések által kijelölt pszeudogyűrűket felépítő atomok számára utalnak.

A β -redők harmadlagos szerkezete rendkívül változatos lehet. A redő csavarodhat (β -twist)[24, 25], képezhet hordót (β -barrel)[26, 27], hélixet (β -helix)[28, 29], propellert (β -propeller)[30] elhajolhat, visszahajolhat saját magára, lehet benne kitüremkedés (β -bulge)[31] vagy kanyar (β -hairpin)[32]. Figyelmet érdemel a különböző β -redőzött szerkezetek egymáshoz történő kapcsolódása is, amelynek során az oldalláncok mintegy cipzárként tartják össze az egyes szálakat.[33-35] Mindezek a jelenségek megnehezítik a β -redők pontos definícióját, és sok esetben lehetőséget adnak a másodlagos szerkezet önkényes hozzárendelésére.

2.3. Aggregációs betegségek

Számos betegség hátterében a fehérjék aggregációja áll. Ennek során a szervezet fehérjei elveszítik natív térszerkezetüket, és vízben oldhatatlan csapadékot képeznek, melyet amiloid aggregátumnak vagy fibrillumnak is neveznek. (A „fibrillum” elnevezés a csapadék szálas szerkezetére utal, az „amiloid” név pedig a jóddal való színreakciónak a hibás interpretációjából származik, amely szerint a fibrillumok keményítőt tartalmaznak; *amylum* latinul keményítőt jelent.[36]) A kialakuló betegség tünetei széleskörűen változnak aszerint, hogy a szervezet mely részében történik meg az aggregáció. Összefoglaló nevükön ezeket a betegségeket amiloidózisnak hívjuk, melyek közt számos neurodegeneratív betegség is szerepel. Amiloidózisnak tekinthető az Alzheimer-kór, a II-típusú cukorbetegség, a Parkinson-kór, és további körülbelül 20 betegség.[37, 38] Mindemellett léteznek funkcionális amiloidok is, például bizonyos pókok hálójában.[39]

Az amiloid lerakódások a fehérjék polimerjeinek tekinthetők, amelyek ún. kereszt-béta (*cross-beta*) szerkezetet vesznek fel. Ezek azonosíthatók CD spektroszkópiával, FTIR spektroszkópiával, azonban legegyszerűbben röntgenkristallográfiával lehet bizonyítani a szerkezetüket. A szórési képen ugyanis két intenzív vonal látszik, (4,8 és 10,7 Å körül), amelyek tükrözik az egymásra rakódott β -rétegek távolságát, valamint az egyes rétegekben a β -szálak közötti távolságot (5. ábra). [40-42] Szövettani vizsgálatok során általában a kongóvörös vagy tioflavin-T festékkel adott színreakciója alapján azonosítják az amiloid aggregátumokat.[36]



5. ábra: kereszt-béta szerkezetű amiloid fibrillumokról röntgensugárral készült szórési kép (balra), valamint a fibrillumok szerkezete, amelyen láthatók a szórési képet megmagyarázó atomi távolságok (jobbra).[41, 42]

A folyamat hajtóereje még nem tisztázott pontosan, de az utóbbi időben sok minden világossá vált. Tudjuk, hogy bizonyos mutációk növelik a fehérje hibás feltekeredésének valószínűségét (*misfolding*), akárcsak a chaperonok hibás működése, vagy a már hibásan feltekeredett fehérjék degradációja. Ezen kívül sikerült az egyes betegségek esetén megállapítani, hogy mely fehérje viselkedik amiloid perkurzorként, így például az Alzheimer-kór esetében a β -amiloid ($A\beta$)[43], vagy a

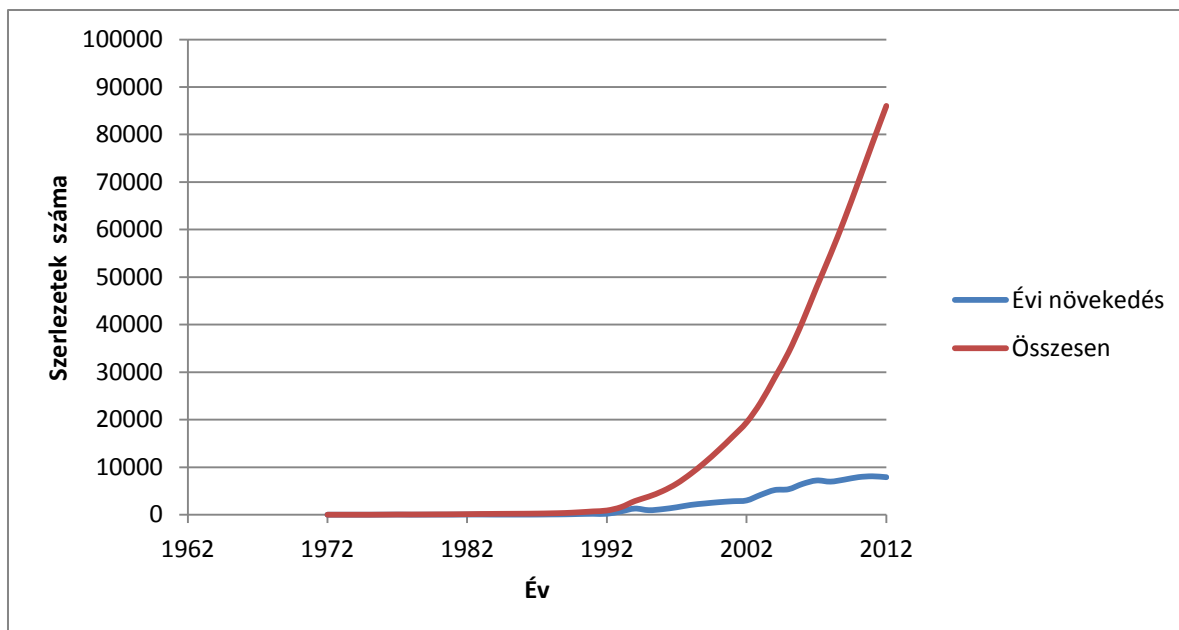
II-típusú cukorbetegségnél az amilin (*Islet Amyloid Polypeptide*, IAPP)[44]. Sőt, Nelson és mts. megnevezték ezekben a fehérjékben azokat a néhány aminosavból álló szegmenseket (tipikusan szabad β -élek), amelyeken általában elindul az aggregáció, és a kristályok növekedéséhez hasonlóan a fehérje többi része, majd a környező fehérjék is hozzátapadnak.[34, 35, 45] Hasonló munkák eredményeképpen már jó pár olyan algoritmus létezik, amelyek rövid szekvenciák amiloid-képző hajlamát képesek jósolni. (Pl. Amyloid Propensity Applet[46], TANGO[47], PASTA[48], stb.) Számos jel arra utal, hogy a természet igyekszik elkerülni a β -élek szabadon hagyását, például a fehérje „betakarja” a veszélyes részt egy másik régióval, vagy olyan struktúrák jönnek létre, mint a β -hordó, vagy a propeller, amelyek töltött aminosav-oldalláncokkal, vagy az egymásba záródás révén gátolják az élek találkozását.[49]

3. A párkorreláció

3.1. A statisztika mint eszköz

A fehérje aggregáció jelensége arra utal, hogy a β -redő rendkívül stabil szerkezet. Számos kutatási projekt célozza e kiemelkedő stabilitás magyarázatát, modellezését. Molekuladinamikai vizsgálatokkal becsülni lehet a különböző β -redős szerkezetek, például a 2 vagy 3 β -szálú elrendeződés stabilitása közti különbséget.[50-54] Hasonló információkhoz lehet jutni NMR mérések segítségével.[55, 56] Mutációkkal meg lehet határozni, hogy egy-egy aminosav cseréje milyen hatással van a β -redő szerkezetére.[22, 57] Befolyásolja a redő stabilitását az egyes szálakat összekötő β -kanyar szekvenciája is.[58] Perczel és mts. kvantumkémiai módszerekkel bizonyították, hogy energetikailag messze a legkedvezőbb homokonformáció az antiparallel β -redő, pusztán a gerincatomok közötti kölcsönhatásokat figyelembe véve [59, 60].

A fenti módszerekkel szemben a statisztikai megközelítés a már meglévő fehérjeszerkezetek vizsgálatával foglalkozik, és az azokat felépítő atomok, aminosavak, peptidláncok elrendeződéséből próbál meg következtetéseket levonni. A továbbiakban bemutatott párkorrelációs eljárás az oldalláncoknak a stabilitásban játszott szerepét tárja fel. Abból a feltevésből indul ki, hogy a β -redőt stabilizáló oldallánc-kölcsönhatások gyakrabban fordulnak elő a fehérjében, mint az azt destabilizálók. Mivel bizonyos fehérjeszegmensek magas aggregációs hajlammal rendelkeznek, mások pedig alacsonyabbal, feltehetjük, hogy a β -redők stabilitását valóban befolyásolják az oldalláncok. Ez a megközelítés nem ad információt a gerincatomok közötti kölcsönhatásról, hiszen peptidgerincben mindig ugyanazok az atomok követik egymást, bármilyen is legyen a szekvencia.



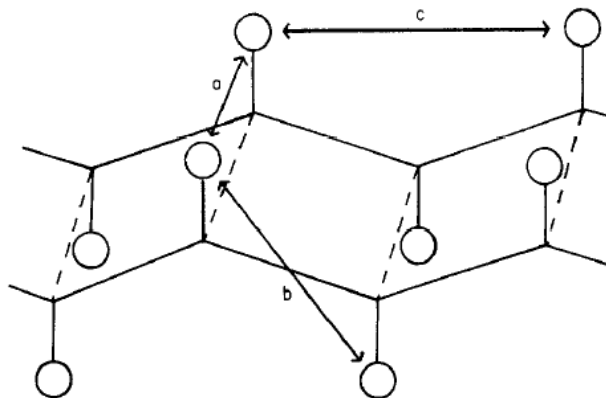
6. ábra: a PDB-ben (Protein Data Bank) tárolt fehérjeszerkezetek számának növekedése az évek előrehaladtával.[61]

Ami elsősorban lehetővé teszi a statisztikai módszerek használatát, az a fehérjeszerkezeti adatbázisok méretének rendkívül gyors növekedése. Gyakorlatilag az összes létező, röntgenkristallográfiával, NMR spektroszkópiával és elektronmikroszkópiával meghatározott fehérjeszerkezetet rendszerezett és egységesített formátumban tárolja a Protein Data Bank (PDB) internetes adatbázis.[1, 62] A PDB növekedésének mértéke az utóbbi két évtizedben ugrásszerűen felgyorsult; évente mintegy 8000 új szerkezet kerül feltöltésre (6. ábra). Ezt a hatalmas adatmennyiséget pusztán tárolni és elérhetővé tenni sem kis feladat, ráadásul a PDB-ben nagyon sok ismétlődés fordul elő, nagymértékű a redundancia. Ezért számos olyan adatbázis és algoritmus született, amelyek a meglévő fehérjéket háromdimenziós szerkezet, szekvencia, funkció vagy valamely egyéb tulajdonság alapján rendszerezik, nemredundáns klaszterekbe sorolják (SCOP[63], CATH[64-66], BLAST[67, 68], PDBSelect[61, 69], WHATIF[70], PISCES[71], cd-hit[72-74], SMS[75], PIR[76], RefSeq[77], TrEMBL[78], PROSET[79], KIND[80], Blocks+[81-83], SwissProt[84], stb.) A hosszú listából látszik, hogy komoly tudományos érdeklődés övezi a szerkezeti és szekvenciális adatbázisokat.

3.2. A párkorreláció irodalmi háttere és jelentősége

A párkorrelációhoz az az alapvető felismerés vezetett, hogy az aminosavak eloszlása a fehérjék különböző régióiban nem egyenletes. Levitt már 1978-ban észrevette, hogy globuláris fehérjékben az aminosavak pozitív vagy negatív preferenciát mutatnak egyik vagy másik homokonformer felé.[85] Az aminosavakat három csoportra osztotta aszerint, hogy az előfordulási arányuk az adott régióban

kiseb vagy nagyobb, mint a teljes fehérjében. α -hélix kedvelő aminosavnak bizonyultak az Ala, Cys, Leu, Met, Glu, Gln, His és Lys, β -redő kedvelőnek a Val, Ile, Phe, Tyr, Trp és Thr, γ -kanyar kedvelőnek a Gly, Ser, Asp, Asn és Pro; az Arg pedig nem tartozik egyik csoporthoz sem. Vegyük észre, hogy a β -redőkben gyakran előforduló aminosavak többnyire az apoláris, hidrofób oldalláncúak közül kerülnek ki. Az aminosavak homokonformációkkal kapcsolatos preferenciáit illetően hasonló eredményre jutott Geisow és Robers 1980-ban.[86]



7. ábra: két szomszédos β -szálban egymáshoz közel lévő oldalláncok lehetséges kölcsönhatásai egymással.[87]

Ehhez hasonlóan von Heijne és Blomberg 1977-ben leírták azt az észrevételüket, hogy a β -redőn belül is rendeződnek az aminosavak: egyesek preferálják egymás közelségét, mások nem.[87] Csak a legközelebbi aminosavakat figyelembe véve, háromféle elrendeződésben tudnak kölcsönhatni az oldalláncok: a láncon belüli szomszédok, a láncon belüli második szomszédok, és két szomszédos szálban,

egymással szemben lévő aminosavak. (7. ábra) Az alacsony mintaszám miatt az aminosavakat csoportosan tekintették: hidrofób (Leu, Ile, Phe, Met, Trp, Tyr, Val, Cys), semleges (Ala, Pro, Gly, Thr, Ser) és poláris (Asn, Asp, Gln, Glu, His, Leg, Arg) aminosavakat különböztettek meg. A korreláció megállapításához összeszámolták, hogy a három elrendeződés szerint hány aminosav pár fordul elő az egyes típusokból (poláris-poláris, poláris-semleges, stb.), a megfigyelt előfordulást összehasonlították a véletlenszerű eloszlásnak megfelelő, várható előfordulással. Azt találták, hogy a legerősebb korreláció a szálak közötti, egymással szemben lévő aminosavak között van, (a 7. ábrán a-val jelölt kölcsönhatás), a másik két kölcsönhatás mentén az aminosavak eloszlása nagyjából egyenletes. Ezt megerősítették a χ^2 statisztikai próbával, amely szerint csak az említett kölcsönhatás mentén megfigyelt korreláció mutatkozik szignifikánsnak 0,1%-os konfidencia szint mellett. A χ^2 -próbát a későbbiekben részletesen is bemutatom.

Lifson és Sander 1980-ban megismételték a vizsgálatot egy nagyobb mintán, és hasonló eredményre jutottak, mint von Heijne és Blomberg, de még mindig nem tudták minden egyes aminosavra lebontva vizsgálni a párkorrelációt.[88] Erre először 1995-ben nyílt lehetőség, amikor már kellően sok fehérjeszerkezet állt rendelkezésre.[3] Ezek után többen ugyanezeket a számolásokat ismételték meg.[23, 89-93] Képletekkel összefoglalva minden egyes i,j aminosav párra:

$$g_{ij} = \frac{o_{ij}}{E_{ij}} \quad (1)$$

ahol o_{ij} a megfigyelt párok száma (*observed*), E_{ij} a párok számának várható értéke (*expected*), (ez a véletlenszerű eloszlásnak megfelelő előfordulás), g_{ij} pedig a kettő hányadosa, melyet a továbbiakban *párpreferencia értéknek* hívok. A párpreferencia érték jelzi két aminosav egymáshoz való viszonyulását: ha az adott aminosav pár preferált, akkor a párpreferencia érték nagyobb 1-nél, ha nem preferált, akkor kisebb. A várható előfordulást a következőképpen lehet meghatározni:

$$E_{ij} = \frac{n_i n_j}{N} \quad (2)$$

ahol n_i és n_j az i-edik és j-edik aminosav összes előfordulása, N pedig az összes aminosav.

20 aminosav esetén az összes párpreferencia értéket egy 20x20-as táblázatba lehet foglalni, amelyre a továbbiakban *párkorrelációs táblázat*ként fogok hivatkozni. A párkorrelációs táblázatnak 20x20, azaz 400 eleme van, de a szimmetrikus pozícióban lévő offdiagonális elemek párosával megegyeznek, mivel az olyan párok azonosak, mint az A-B és B-A. Így összesen 210 féle aminosav pár létezik, ha nem különböztetjük meg az AP10, AP14 és PARL párokat.

A párkorrelációkat vizsgáló és számszerűsítő munkák kvalitatíve hasonló eredményekre jutottak, amely az ismert közmondással foglalható össze: hasonló a hasonlónak örvend. Erős pozitív korreláció figyelhető meg az apoláris oldalláncú aminosavak között (Val, Ile, Leu), az aromások között (Phe és Tyr), valamint nagy összetartó erő az elektrosztatikus kompatibilitás (pl. Glu és Gln az Arg-nel vagy a Lys-nel). Ugyanakkor negatív korreláció tapasztalható az elektrosztatikusan ellentétes töltésű aminosavak között, valamint a különböző természetű, például poláris és hidrofób aminosavak között.

Wouters és Curmi, Hutchinson és *mts.*, valamint Fooks és *mts.* az egyes párpreferencia értékek statisztikai szignifikanciáját szintén a χ^2 -tesztel ellenőrizték. Mi Fooks és *mts.*-hoz hasonlóan a Yates-korrigált χ^2 -tesztet fogjuk alkalmazni.[94] Ennek során minden egyes aminosav párhoz készítünk egy 2x2-es, ún. kontingencia táblázatot, amelyben a minta összes aminosav párját a következőképpen csoportosítjuk, például egy tetszőleges A-B pár esetén: A-B, A-nem B, nem A-B, és nem A-nem B. Erre a négy értékre összegezzük az alábbi képletet:

$$\chi^2 = \sum_{i=1}^4 \frac{(|o_i - E_i| - 0,5)^2}{E_i} \quad (3)$$

ahol a jelölések változatlanok.

A statisztikai teszt során egy nullhipotézist mondunk ki, amely a χ^2 -teszt esetében nem más, minthogy A és B a mintában egymástól függetlenek. Ennek cáfolására vagy elfogadására egy adott megbízhatósági, vagyis konfidencia szint mellett úgy nyílik lehetőség, hogy a kiszámolt χ^2 -értéket összehasonlítsuk statisztikai táblázatokban található értékekkel, a minta szabadsági fokának függvényében, ami a szabadon megadható elemek száma (2x2-es táblázat szabadsági foka 1.) 1

szabadsági fok esetén, ha $\chi^2 \geq 23,93$, az arra utal, hogy a nullhipotézisünket $p = 0,000001$ tévedési valószínűséggel elutasíthatjuk, vagyis az adott két aminosav korrelációja szignifikáns.[23, 95]

A szálak közötti kölcsönhatásra (*cross-strand interaction* vagy *inter-strand interaction*) az irodalomban gyakran hivatkoznak, és a párpreferencia értékeket felhasználják más munkákhoz is. A párkorreláció szerepet játszik a γ -kanyarok stabilitásában, ha a kanyar előtt és után néhány aminosav β -hidat képez [96-98]. Segítséget nyújt a fehérje feltekeredés megértésében is [92, 99], de talán a legfontosabb, hogy alkalmazható az aggregációs hajlam jóslásában [93, 100, 101] és a *de novo* fehérjetervezésben [22, 55, 102, 103]. Külön vizsgálják az oldalláncok orientációjának a hatását is a párkorrelációra.[104, 105]

Összességében azt mondhatjuk, hogy a párkorreláció jelenségét komoly tudományos érdeklődés övezi, és sok kapcsolódó kutatás igényli a párpreferencia értékek pontos meghatározását.

3.3 Az alkalmazott kvantumkémiai módszerek

Az elmúlt években az elérhető számolási kapacitás növekedésével egyre inkább teret hódított a számítógépes kémia a fehérjekutatás területén. Előszeretettel alkalmazzák konformációs vizsgálatoknál a potenciális energia felület számolásához, molekuladinamikai erőterek fejlesztéséhez, elektrosztatikai potenciálok meghatározásához, spektroszkópiai paraméterek becsléséhez, különösen a fehérje NMR területén.[106] A mi vizsgálataink során az aminosavak oldalláncai közötti kölcsönhatási energia számolásában bizonyult hasznosnak (7.7. fejezet).

A kvantumkémiai módszerek a Schrödinger-egyenlet megoldását célozzák:

(4)

ahol a rendszer Hamilton-operátora, E az energiája, ψ pedig a rendszert leíró hullámfüggvény, amelyből többnyire meghatározhatók a rendszer tulajdonságai. A hullámfüggvény az absztrakt Hilbert-tér eleme, amelyen a Hamilton-operátor is hat.[107]

3.3.1. A Hartree-Fock-módszer

A Schrödinger-egyenlet megoldását legegyszerűbben a Hartree-Fock módszer közelíti.[108] Ebben a rendszer hullámfüggvényét az egyelektron hullámfüggvényekből képzett determinánssal írjuk le. Ennek következtében a rendszer Hamilton-operátora egyelektron operátorokra esik szét, amelynek alakja:

$$H = \sum_i \left(-\frac{\hbar^2}{2m} \nabla_i^2 + V_{\text{ext}}(\mathbf{r}_i) \right) + \sum_{i,j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \quad (5)$$

Az összeg első tagja az elektronok kinetikus energiáját írja le, a második a mag-elektron kölcsönhatásokat, a harmadik tag pedig az elektronok közötti kölcsönhatást közelíti egy átlagos, ún. effektív potenciállal. A megoldást egy iteratív eljárás során kapjuk: először kezdeti pályákat választunk, azokból kiszámoljuk az effektív potenciált, majd megoldjuk az egyelektron egyenleteket:

$$H_{\text{eff}} \psi_i = E_i \psi_i \quad (6)$$

amelyekből új pályákat kapunk. Ezt a módszert az irodalomban SCF (*self consistent field*) módszernek nevezik. A Hartree-Fock-módszer olcsó, azonban pontatlan, mivel nem veszi figyelembe az elektronkorrelációt.

3.3.2. A DFT módszer

Az összetettebb módszerek közül az egyik legnépszerűbb a sűrűségfunkcionál-elmélet (DFT, density functional theory). A DFT módszerek az első Hohenberg-Kohn-tételen alapulnak.[109] E tétel szerint az alapállapotú elektronsűrűség egyértelműen meghatározza a rendszer potenciálfüggvényét, ebből következően pedig a rendszer összes mérhető tulajdonságát, beleértve a gerjesztett elektronállapotok tulajdonságait is. Ezt fejezi ki a következő egyenlet.

$$E[\rho] = E_{\text{kin}}[\rho] + E_{\text{ext}}[\rho] + E_{\text{ee}}[\rho] + E_{\text{exch}}[\rho] + E_{\text{corr}}[\rho] \quad (7)$$

Írjuk fel a rendszer energiáját mint az elektronsűrűség funkcionálját:

$$E[\rho] = T[\rho] + V_{\text{ext}}[\rho] + V_{\text{ee}}[\rho] + E_{\text{xc}}[\rho] = T[\rho] + V_{\text{ext}}[\rho] + \int \rho(\mathbf{r}) V(\mathbf{r}) d\mathbf{r} + E_{\text{xc}}[\rho] \quad (8)$$

ahol T az elektronok kinetikus energiájából adódó járulék, V_{ee} és V_{ne} pedig az elektron-elektron és mag-elektron kölcsönhatásból származó járulék. Kihasználhatjuk, hogy a mag-elektron vonzás minden elektronra ugyanakkora, és a jobb oldalon lévő integrállal írható fel. Így a T és V_{ee} tagot össze

tudjuk vonni F -be, amely a Hohenberg-Kohn-funkcionál. Ez egy külső potenciáltól független, kizárólag a részecskeszámtól függő funkcionál.

A második Hohenberg-Kohn-tétel kimondja, hogy ha a rendszer alapállapotának energiája E_0 , elektronsűrűsége pedig $\rho_0(r)$, akkor bármely ρ próbasűrűsége a kapott energia $E[\rho] \geq E_0$. [109] Variációs úton tehát meg is lehetne határozni a funkcionált, de F -et nem ismerjük.

Egy apró trükkel azonban át tudjuk alakítani a kérdést. Egy olyan fiktív rendszer Hohenberg-Kohn-funkcionálját könnyen fel tudjuk írni, amelyben az elektronok nem hatnak kölcsön:

$$F[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho] \quad (9)$$

ahol T_s a fiktív rendszer kinetikusenergia-funkcionálja, amit könnyen fel tudunk írni, J a V_{ee} -ből a klasszikus Coulomb-tag, amit szintén ismerünk, E_{xc} pedig az ún. kicserélődési-korrelációs funkcionál, aminek az alakját viszont továbbra sem ismerjük.

Az energiát variációsan érdemes számolni:

$$\frac{\delta E[\rho]}{\delta \rho} = \mu = \frac{\delta F[\rho]}{\delta \rho} + v = \frac{\delta T_s[\rho]}{\delta \rho} + \frac{\delta J[\rho]}{\delta \rho} + \frac{\delta E_{xc}[\rho]}{\delta \rho} + v = \frac{\delta T_s[\rho]}{\delta \rho} + v_{eff} \quad (10)$$

A variációs problémát a Hartree-Fock módszerhez hasonló iteratív eljárással meg tudjuk oldani:

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + v_{eff} \right] \varphi_i = \sum_{j=1}^n \varepsilon_{ij} \varphi_j \quad (11)$$

Ez a Kohn-Sahm-egyenlet [110], a megoldásai a Kohn-Sahm-pályák, amelyek segítségével meghatározható az elektronsűrűség, ebből pedig a rendszer energiája. A DFT módszerek nagy népszerűségének az oka, hogy nagyszerűen alkalmazható esetükben a Hartree-Fock-módszer miatt már a legtöbb kvantumkémiai programcsomagba előzőleg implementált SCF formalizmus.

A nehézséget az jelenti, hogy E_{xc} alakját nem ismerjük. A múltban számos próbálkozás történt ennek a hiánynak az orvoslására, amelyek mindegyikéhez egy-egy lehetséges DFT-módszer párosul. A módszer nagy hátránya, hogy nem tudjuk a hibáját becsülni, de a tapasztalat mégis azt mutatja, hogy relatíve kis számolásigény igen pontos eredményeket ad. Mi a számolások során az egyik leggyakrabban használt, B3LYP [111] funkcionál mellett az M052X [112] funkcionált alkalmaztuk.

3.3.3. A BSSE hiba és a counterpoise korrekció

A bázisátfedési hiba (*basis set superposition error*, BSSE) általában egymáshoz gyengén kötött molekulák, például komplexek esetében keletkezik annak eredményeképpen, hogy az egyik molekula kihasználatlan pályái a másik számára elérhetővé válnak, ezáltal növelve annak bázisát, és mélyebb energiaszintet jósolva a szerkezetnek a valóságosnál.[113, 114] Bizonyos esetekben intramolekuláris átfedés is okozhat hibát.[115] Ennek kiküszöbölésére több megoldás is létezik: a *Chemical Hamiltonian Approach*[116] vagy a *BSSE-free SCF* módszerek.[113] Mi a *counterpoise correction* módszert használtuk, amelynek során a hibát egy AB komplex esetében a következő képlettel becsüljük [117-119]:

$$\Delta E_{AB}^{\text{cp}} = E_{AB}^{\alpha} \text{A} \text{---} E_{AB}^{\alpha \cup \beta} \text{A} \text{---} E_{AB}^{\beta} \text{B} \text{---} E_{AB}^{\alpha \cup \beta} \text{B} \text{---} \quad (12)$$

ahol α és β az A és B molekulára illesztett bázist jelöli, a zárójelben jelöljük a tényleges rendszert, aminek az energiájáról szó van, az alsó indexben található AB pedig a geometriára utal, melyet a hibaszámítás közben nem változtatunk. Ilyen módon a teljes energia:

$$E_{total}^{cp}(\mathcal{AB}) \stackrel{?}{=} E_{total}(AB) + \Delta E_{AB}^{cp}(AB) \quad (13)$$

4. Problémafelvetés, célkitűzés

Az 1995 és 2010 között elkészített párkorrelációs táblázatok mind különböző adatkészletekből dolgoztak, méretük 253 és 4036 fehérjelánc, illetve domén között változott.[3, 23, 89, 93] Más volt az adatkészletek redundanciájának mértéke, illetve a szűrési algoritmus. Így nem lepődünk meg, mikor észrevettük, hogy bár a tendenciák hasonlóak, a konkrét párpreferencia értékek sok ponton eltérnek az egyes táblázatok között. Az 1. táblázat két korábbi adatbázis eredményeit hasonlítja össze három általunk elkészített adatkészlet eredményeivel (melyeket a későbbiekben részletesen is bemutatok). A párkorrelációs táblázatok hasonlóságának mérésére a Pearson-féle korrelációs koefficientst használtuk. A képletben az egyes párpreferencia értékekre összegeztünk:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = \frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} + \frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} \quad (14)$$

ahol $\text{cov}(X,Y)$ a kovarianciát, $\sigma(X,Y)$ a szórást (standard deviáció) jelölik, és az átlagot jelentik.[120] A Pearson-féle korrelációs koefficiens értéke -1 és 1 között változik aszerint, hogy az

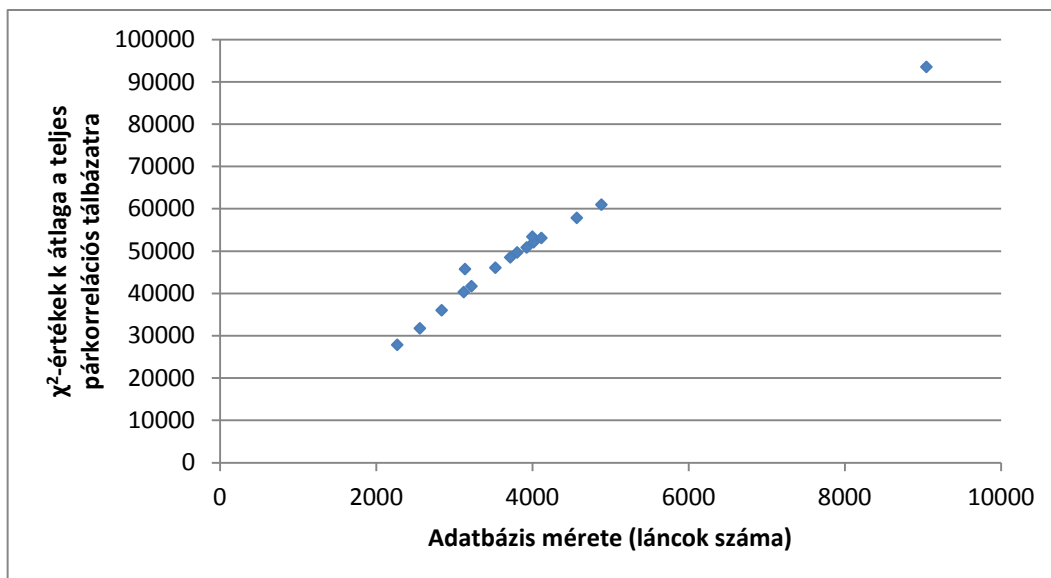
összehasonlított X és Y adathalmazok korrelálnak ($\text{corr} = 1$), függetlenek ($\text{corr} = 0$) vagy antikorrelálnak ($\text{corr} = -1$). Ennek megfelelően két adatbázis összehasonlításakor 1 közeli Pearson-féle korrelációs koefficiens várunk.

Fent kifejtettem, hogy a párpreferencia értékek sok további vizsgálat során hasznosnak bizonyulnak, ám most az derül ki, hogy gyakran nem egyértelműek. Szükség van arra, hogy egy megbízható betanító adatkészletet hozzunk létre, és dönteni tudjunk arról, hogy egy-egy fehérjeszerkezetet milyen kritériumokkal vegyünk be vagy hagyjunk el.

1. táblázat: különböző adatbázisok párpreferencia értékei között számolt Pearson-féle korrelációs koefficiensek. Az 1-3. adatbázis saját eredmény, a 4-5. adatbázisok az irodalomban korábban közölt eredményekből származnak.

Adatforrás		1	2	3	4	5
PDB select 2009	1	1,00	0,91	0,85	0,59	0,68
WHATIF 2009	2		1,00	0,88	0,61	0,71
PISCES 2010	3			1,00	0,52	0,68
Wouters 1995 [3]	4				1,00	0,69
Fooks 2006 [23]	5					1,00

Szintén szükség van arra, hogy tudjunk mondani valamit egy-egy párpreferencia érték megbízhatóságáról, illetve az egész adatbázis jóságáról, azonban az adatok validálása során is nehézségekbe ütköztünk. Az derült ki, hogy az adatkészlet méretének növekedésével a χ^2 -próba alkalmatlanná válik, ugyanis szinte minden aminosav párra gyakorlatilag 100%-os konfidencia szintű korrelációt jósol. Ez következik a (3) egyenletből is: mivel az aminosav párok megfigyelt és várható gyakorisága is nagyjából lineárisan nő az összes aminosav pár számával (vagyis az adatbázis méretével), a számlálóban nagyjából négyzetes, a nevezőben pedig nagyjából lineáris növekedés várható, tehát a χ^2 -érték is nagyjából lineárisan fog nőni. A szabadsági fokok száma természetesen nem nő (Yates-korrigált χ^2 -teszt esetén 2x2-es táblázatra mindig 1). Ennek következtében a χ^2 -érték egyre nagyobb lesz, így egyre nagyobb valószínűséggel utasíthatjuk el a nullhipotézist.[95] Ezt a saját tapasztalatunk is megerősíti (8. ábra), amely szerint ha a felhasznált láncok száma meghaladja a kétezret, a χ^2 -érték már olyan nagy lesz, hogy nincs értelme a χ^2 -próbát használni. Ez nyitva hagyja azt a kérdést, hogy a továbbiakban hogyan tudjuk megállapítani egy adott adatbázisról, hogy a benne lévő adatok mennyire megbízhatóak. Lehetőségünk van ugyan a Pearson-féle korrelációs koefficiens használni, de ez csak az adatbázisok egymástól való eltéréséről ad információt, ami például két „nagyon rossz” adatbázis esetében is lehet kicsi.



8. ábra : a χ^2 -értékek és az adatbázis méretének korrelációját szemléltető diagram. Az adatok az NMR-es fehérjeszerkezetekből származnak (4. táblázat).

A fenti problémák felismerése után a következő célokat tűztük ki magunk elé:

1. Egy olyan standard eljárás kidolgozása a párkorrelációs vizsgálatokhoz, amely követni tudja a folyamatosan bővülő adatbázisokat.
2. Egy olyan validálási eljárás kidolgozása, amely önmagában képes az adatkészlet jóságának becslésére, és ez a jósági paraméter összehasonlítható a többi adatbáziséval.
3. A párpreferencia értékek függésének megállapítása a kiindulási adatkészlet paramétereitől, azaz az adatbázis méretétől, a szerkezetmeghatározás módjától (NMR spektroszkópia, röntgenkristallográfia), röntgenszerkezet esetén a jósági paraméterektől (felbontás, R-faktor), az adatkészlet redundancia szintjétől és az alkalmazott szűrési algoritmustól.
4. A párpreferencia értékek összefüggésének felderítése az oldalláncok közötti kölcsönhatási energiával.

5. A statisztikai analízis bemutatása

5.1. A kiindulási adathalmaz

A kiindulási adataink három csoportra oszthatók: a PDB (Protein Data Bank, [121]) röntgenszerkezetei, a PDB NMR-szerkezetei, valamint 3 ingyenesen elérhető, előszűrt adatbázis szerkezetei.

Az ingyenesen elérhető, előszűrt adatbázisoknak az átlagos felhasználó számára megvan az az előnyük, hogy relatíve kicsik, ezért könnyen kezelhetők, válogatva vannak a szerkezetek jósági paraméter szerint (felbontás, R-faktor), és nemredundánsak, ezért a szűréssel sem kell bajlódni. Hátrányuk, hogy csak röntgenszerkezeteket tartalmaznak. A mi munkánk során felhasznált 5 különböző előszűrt adatkészletet tartalmazza a 2. táblázat.

Mivel NMR-es szerkezeteket is vizsgálni szerettünk volna, és a különböző redundanciaszűrő algoritmusok közötti különbség is érdekelt minket, ezért úgy döntöttünk, hogy felhasználjuk az összes, PDB-ben található NMR-, valamint röntgenszerkezetet (3. táblázat).

2. táblázat: az általunk felhasznált ingyenesen elérhető, előszűrt adatbázisokból származó fehérjeszerkezeti listák, és a röntgenszerkezetek jósági paraméterei.

	1.	2.	3.	4.	5.
Adatforrás	PDB select [61]		WHATIF [70]		PISCES [71]
Dátum	2010.03.23.		2009.02.28.		2010.09.12.
Max. lánc azonosság (%)	25	25	30	30	25
Felbontás (Å)	1,5	1,75	1,8	1,8	1,8
R-faktor	0,18	0,18	0,18	0,19	0,25
Láncok száma	1020	1516	1962	2519	3416

3. táblázat: a PDB-ből felhasznált fehérjeszerkezetek, valamint a keresési feltételek.

	Röntgenszerkezetek	NMR-szerkezetek
Kísérleti módszer	Röntgenkristallográfia	Oldat NMR
Tartalmaz fehérjét	Igen	Igen
Tartalmaz DNS-t	Nem	Nem
Tartalmaz RNS-t	Nem	Nem
Tartalmaz hibridet	Nem	Nem
Max. felbontás (Å)	2,0	-
Max. R-faktor	0,25	-
Letöltés dátuma	2012.07.11.	2011.07.06.
Fehérjék száma	28674	7792
Láncok száma	57331	9037

5.2. A redundanciaszűrés

A „hasonlóság” (*similarity*) és a „homológia” (*homology*) fehérjékre vagy nukleinsavakra vonatkozó fogalmak, melyeket gyakran többféle értelemben használnak. Én Reek és *mts.* ajánlásának megfelelően a következőket értem rajtuk.[122]

Két fehérje akkor homológ, ha bizonyíthatóan közös őstől származik, tehát mindenképpen van köztük valamilyen fejlődéstani kapcsolat. A homológia vagy fennáll, vagy nem; nem lehet különböző szintekről beszélni. A hasonlóság jelöli ezzel szemben az egyezést két fehérje szerkezetében vagy szekvenciájában, amely általában kifejezhető valamilyen számmal. Se a szekvenciális, se a szerkezeti hasonlóság nem bizonyítja a homológiát, ami lehet pusztán konvergens fejlődés eredménye is, ugyanakkor szerepet játszanak az érvelés alátámasztásában.

A fehérjék szerkezeti hasonlóság alapján történő csoportosítására komoly tudományos igény van; a 3.1. fejezetben felsorolt nemredundáns adatbázisok jó része ezt a módszert követi (pl. SCOP, CATH, SMS, stb.) [63-66, 75]

Mi magunk a szekvenciális hasonlóságot (*sequence similarity*) vizsgáltuk, amely az elsődleges szerkezet, tehát az aminosavak sorrendjének hasonlóságát jelenti. Számszerűsíteni lehet az adott pozícióban azonos aminosavak és az összes aminosav százalékosan kifejezett arányával. A biostatistikában bevett szokás szerint ezt a mennyiséget szekvencia azonosságnak hívom (*sequence identity*). 100%-os szekvencia azonosság két egyforma láncra utal, vagy két olyan láncra, ahol az egyik szekvenciája teljes mértékben tartalmazza a másikat. Egy 25%-os szekvencia azonossági küszöbértékkel rendelkező adatbázisban semelyik két lánc között nem lehet magasabb a szekvencia azonosság, mint 25%. Mivel a szekvencia és a térszerkezet összefüggnek egymással, a szekvencia azonosságnak is komoly jelentősége van a fehérjék csoportosításának szempontjából.

A nemredundáns adatkészletek használatának sok esetben praktikus oka van: nem érdemes nagy mennyiségű adatot feldolgozni, ha ugyanazt az információmennyiséget sokkal kisebb adatkészletből is kinyerhetjük. (Klasszikus példa: sok fehérje többször van benne a PDB-ben, vagy ugyanannak a fehérjének nagyon sokféle variánsa van, amelyekben a szerkezet nagy része azonos). Egy statisztikai analízis során azonban a redundancia eltávolításának ennél sokkal nagyobb jelentősége van. Ugyanis a párpreferencia értékeket éppen egy adott aminosav pár előfordulási gyakoriságából számoljuk ki. Ha viszont egy szekvencia, vagy egy teljes β -redő többször szerepel, akkor a benne szereplő aminosav párok súlyát tévesen nagyobboknak számolhatjuk. Ezzel szemben ha nagyon alacsonyra állítjuk a szekvencia azonossági küszöbértéket, akkor nagyon lecsökken az adatkészlet mérete, ami szintén rontja a statisztikai elemzés megbízhatóságát.

Az alapvető nehézséget két dolog jelenti. Egyrészt egy adatkészletnek nem egyértelmű a nemredundáns részhalmaza, különböző szűrési algoritmusokkal másféle eredményhez jutunk. Másrészt pedig a szűrés a teljes szekvenciára vonatkozik, tehát nem tudjuk, hogy mennyi β -redőt hagyunk el, vagy tartunk meg hibásan. Indokolt tehát a különböző szűrési algoritmusok vizsgálata, ahogy a különböző szekvencia azonossági szintek összehasonlítása is.

A redundanciaszűrés hatását az NMR-es adatkészleten teszteltük. Kiválasztottunk három szűrőprogramot (PDBSelect, cd-hit, PISCES), amelyek egymástól eltérő szűrési algoritmust használnak.

A PDBSelect egy előre definiált sorrend mentén, amely sorrendet a felhasználó módosíthatja, végighalad a listában felsorolt fehérjéken vagy fehérjeláncokon, és, ha az éppen vizsgált fehérje megfelel az esetlegesen kiszabott kritériumoknak (lánc hossz, felbontás, stb), akkor megvizsgálja, hogy a már kiválasztott láncok valamelyikével nem mutat-e a vizsgált lánc a küszöbértéknél nagyobb szekvencia azonosságot. Ez alapján az vizsgált láncot elveti vagy beveszi az kiválasztott láncok közé.

A PISCES program a bemenetként megadott fehérjéket vagy fehérjeláncokat algoritmikusan sorbarendezi növekvő felbontás, azonos felbontás esetén pedig növekvő R-faktor szerint, az esetleges NMR-es szerkezetek pedig a lista végére kerülnek. Ezt követően az első láncot automatikusan beveszi a kiválasztott láncok közé, majd a listáról kihúzza az összes olyan utána következőt, amely a küszöbértéknél magasabb szekvencia azonosságot mutat a vizsgált láncsal. Ilyen módon végighalad az összes felsorolt láncra.

E két módszer kritikus lépése a szekvenciák összehasonlítása. A láncokat egymás mellé rendezve azt a „legjobb” elrendeződést kell megtalálni, amelyben a lehető legtöbb azonos aminosav szerepel a két láncban, figyelembe véve az esetlegese beékelődéseket, (például ha két szekvencia csak abban különbözik, hogy az egyik néhány aminosavval hosszabb, mint a másik, viszont az a néhány plusz aminosav nem a szekvencia végén, hanem a közepén szerepel.) Ha minden lehetőséget figyelembe szeretnénk venni, akkor két, N hosszúságú szekvenciát ———— lehetséges elrendeződésben kell összehasonlítanunk. Ez 250 aminosav esetén 10^{149} nagyságrendbe esik, ami óriási szám, gyakorlatilag lehetetlen ezt a módszert alkalmazni. Valamelyest egyszerűsíti a helyzetet a Needleman-Wunsch-algoritmus, amely egy $N \times N$ -es mátrix elemeinek meghatározására, és a mátrix celláiban történő konszekutív keresésre csökkenti a szükséges számításokat.[123] Mivel a bemeneti lista minden egyes fehérjéjét össze kell hasonlítani az összes többivel, a számítás még így is igen hosszúra nyúlhat. A másik gyengesége a PISCES és PDBSelect programok által használt algoritmusnak, hogy az eredménynek függenek a bemeneti adatok sorrendjétől. Igaz, a PISCES program sorbarendezi a szerkezeteket, de ezt is valamelyest önkényesen teszi.

A cd-hit program mind a két problémát kiküszöböli az ún. *short-word filtering* („rövid szavak szűrése”) használatával. Belátható, hogyha két szekvencia egy bizonyos szintű azonosságot mutat, akkor szükségszerű, hogy tartalmazzanak valamennyi azonos dipeptidet, tripeptidet, stb., amelyeket „rövid szavaknak” (*short-word*) neveznek. Például két, 100 aminosavból álló polipeptid 80 %-os azonosság esetén legalább 70 azonos dipeptidet, 55 tripeptidet, 25 pentapeptidet tartalmaz. A cd-hit 2-5 peptid hosszúságú „rövid szavakat” vizsgál, amelyek könnyen indexelhetők, és ezért nagyon rövid idő alatt rengeteg, egymáshoz viszonylag hasonló láncot azonosítani lehet. A program klaszterekbe

sorolja a lista fehérjét vagy fehérjeláncait. Bármely klaszteren belül a láncok nagyobb azonosságot mutatnak egymással a megengedett küszöbértéknél, de a különböző klaszterek fehérjei legalább a szükséges mértékben különböznek egymástól. A reprezentatív adatkészletet minden klaszter leghosszabb képviselőjéből áll. Az algoritmus hátránya, hogy 40%-os szekvencia azonossági küszöbérték alá nem lehet vele menni, valamint a gyors számolás érdekében fel kell adni a matematikai korrektséget, hiszen a „rövid szavak” keresése valószínűségi alapon valósul meg.

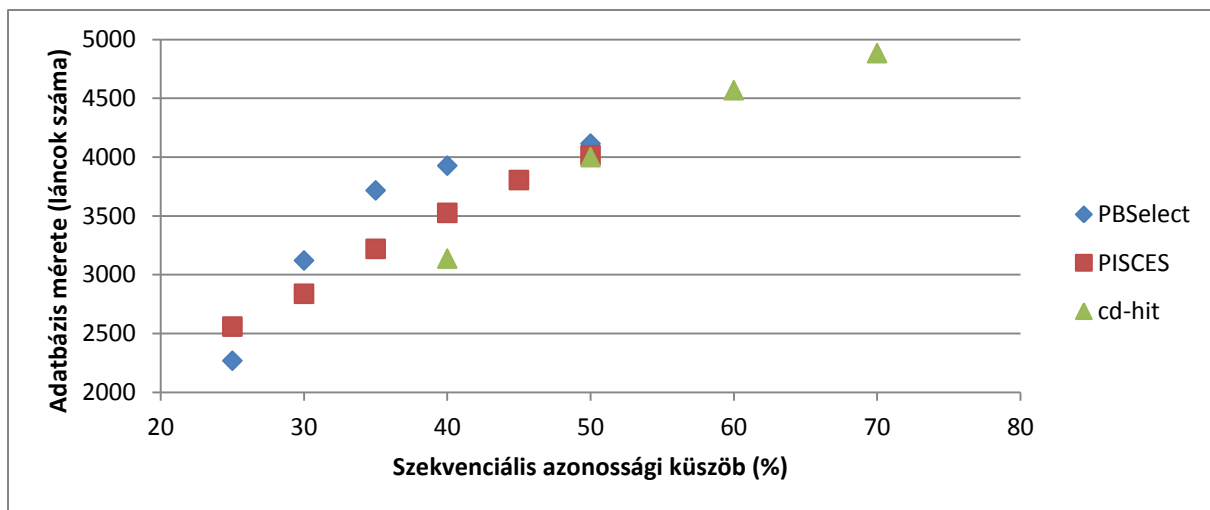
4. táblázat: felhasznált NMR-es fehérje adatkészletek, feltüntetve a szekvencia azonossági küszöböt.

Algoritmus	Azonosság (%)	Láncok száma
PDBSelect [61]	25	2270
	30	3122
	35	3718
	40	3928
	50	4116
PISCES [71]	25	2560
	30	2839
	35	3221
	40	3527
	45	3806
	50	4017
cd-hit [72-74]	40	3138
	50	4001
	60	4570
	70	4885
teljes lista	-	9047

A három bemutatott programmal ugyanabból a kiindulási adatkészletből (3. táblázat, Oldat NMR) több különböző szekvencia azonossági küszöbértékű listát állítottunk elő, melyeken külön-külön elvégeztük a párkorrelációs analízist (4. táblázat). A PDBSelect és PISCES weboldalán megtalálhatók az előszűrt adatbázisok is, de az NMR-es szerkezetek vizsgálatához nem ezeket, hanem magukat a szűrőprogramokat használtuk fel. A szűrőalgoritmusok különbözőségét szemlélteti a 9. ábra, amelyen látszik, hogy ugyanabból a kiindulási adatkészletből különböző méretű listák állnak elő.

A röntgenszerkezeteken a jósaági paraméterek változtatásának hatását akartuk tesztelni, ezért felbontás és R-faktor szerint megszürt listákat hoztunk létre. Mindazonáltal a

biztonság kedvéért a röntgenszerkezeteken is végrehajtottunk egy redundanciaszűrést: a cd-hit programmal 40%-os szekvencia azonossági küszöböt állítottunk be, és az így kapott listán, amely lényegesen kisebb volt, mint az eredeti, megismételtük a jósaági paraméterek szerinti csoportosítást (5. táblázat).



9. ábra: a teljes NMR-es adatkészletből redundanciaszűréssel előállított listák, részletesen ld. 4. táblázat. Látszik, hogy a különböző algoritmusok más-más listát hoznak létre, még akkor is, ha ugyanakkora azonossági küszöbértéket állítunk be.

5. táblázat: röntgenszerkezetekből álló redundáns és nemredundáns listák. Minden lista meghatározott jósági paraméterrel (felbontás, R-faktor) rendelkező szerkezetekből áll. Az nemredundáns listák cd-hit programmal készültek, 40%-os azonossági küszöbvel. Látszik, hogy a redundanciaszűrés legalább negyedére csökkenti a listák méretét.

Redundáns adatkészlet (láncok száma)						
Max. R-faktor	Maximális felbontás (Å)					
	1,0	1,2	1,4	1,6	1,8	2,0
0,10	51	69	72	78	78	86
0,13	246	622	847	968	1005	1037
0,16	347	1151	2184	3335	4588	5551
0,19	369	1333	3182	6564	11991	17638
0,22	381	1402	3558	8220	16409	26736
0,25	383	1411	3606	8475	17160	28664

Nemredundáns adatkészlet (cd-hit, 40 %) (láncok száma)						
Max. R-faktor	Maximális felbontás (Å)					
	1,0	1,2	1,4	1,6	1,8	2,0
0,10	7	10	10	10	10	11
0,13	34	89	134	160	167	173
0,16	49	194	447	793	1087	1316
0,19	52	228	671	1637	3089	4780
0,22	53	242	756	2053	4368	7704
0,25	53	244	765	2139	4623	8378

5.3. Az SDB program

A teljes statisztikai analízist egy saját fejlesztésű programmal végeztük el, melynek a Sheet Database Builder, röviden SDB nevet adtuk. A programot Jákli Imrével közösen, Java nyelven írtuk; mintegy 30 osztályból, összesen több mint 15000 programsorból áll. A program aktuálisan legfrissebb verziója megtalálható az interneten (<http://www.chem.elte.hu/departments/jimre/sdb/>).

Bemenetként a négybetűs PDB kódokból (pl. 1ACY), vagy pedig ötbetűs, PDB kódból plusz a fehérjelánc kódjából (pl. 1ACYB) álló listát fogad. A fent bemutatott szűrőprogramok kimenetét (cd-hit, PISCES, PDBSelect) is képes feldolgozni. Megadható egy vagy több lista is, amelyekre külön-külön elvégzi a párkorrelációs analízist, a végén pedig összehasonlítja az eredményeit. Ennek részleteit később bemutatom.

Az SDB program első lépésként letölti a PDB weboldaláról (<http://www.rcsb.org>) a listában szereplő fehérjék szerkezeti fájljait, amelyekre a továbbiakban PDB fájlként fogok hivatkozni. A PDB fájlok tartalmazzák a fehérje szekvenciáját, az atomi koordinátákat, valamint még egy sor információt a fehérjéről, amelyet mi közvetlenül nem használunk fel.

A program további funkcióira a statisztikai analízis egyes lépéseinek részletes bemutatásakor kitérek.

5.4. Másodlagos szerkezeti elemek meghatározása

Számos olyan program létezik, amely felismeri a másodlagos szerkezeti elemeket a fehérjében. (PROSIGN [124], P-SEA [125] STRIDE [126], DSSP [21], stb.) Munkánk során a DSSP egyik továbbfejlesztett változatát használtuk, a DSSPcont-ot [127]. Egy algoritmus szerint parciális töltéseket helyez el az atomokon, és a háromdimenziós koordináták alapján a hidrogénkötések erősségét becsli. A másodlagos szerkezeti elemeket (összesen nyolcfélét: 3_{10} -hélix, α -hélix, π -hélix, kanyar, β -redő, β -híd, hajlat [*bend*], hurok [*loop*]) a hidrogénkötés mintázat alapján rendeli hozzá az aminosavakhoz. A β -redőt és a β -hidat a 2.2. fejezetben bemutatott mintázatról ismeri fel.

A DSSP program bemenete egy tetszőleges PDB fájl. Nem tudja azonban megfelelően kezelni az NMR szerkezeteket, ugyanis ezekben gyakran több, akár 10-20 szerkezeti modell van felsorolva, ezzel mintegy tükrözve a fehérje különböző régióinak mozgékonyságát. (A modelleket lényegében egy-egy fényképnek tekinthetjük egy folyamatosan mozgó láncról). Számunkra ez egy nagyon fontos effektus, ugyanis a fehérjével együtt a β -redők is mozognak, ami szükségessé teszi annak mérlegelését, hogy a redő, és a redőt felépítő aminosav párok szerkezetileg mennyire stabilak. Ezzel szemben a DSSP mindig az első modell alapján értékeli ki a PDB fájlt. Bár a β -redők általában a

fehérje leginkább robosztus részét képezik, ez adott esetben hibát eredményezhet, például ha egy relatíve instabil β -redőről pont az első modellben egy olyan „fénykép” található, amelyiken a redő éppen szétesett. Azért, hogy ezt a hibát elkerüljük, a PDB fájlokat feldaraboltuk annyi részre, ahány modellt tartalmaztak, és minden egyes modellre elvégeztük a másodlagos szerkezeti elemek azonosítását a DSSP programmal.

A DSSP a kimeneti állományában a következő (számunkra hasznos) információk szerepelnek:

- az aminosavak sorszáma, amivel egyértelműen azonosítani lehet őket a fehérjén belül,
- a másodlagos szerkezeti elem fajtája, illetve annak fontosabb tulajdonságai (pl. hányadik pozíció a kanyarban, mely aminosavval van párban a β -redőben, parallel vagy antiparallel párost adnak),
- hogy mely aminosavakkal alakít ki hidrogénkötést, és a hidrogénkötés erőssége kcal/mol-ban, (külön a donor- és akceptor-csoportokra, és külön jelöli a bifurkált hidrogénhidakat),
- az aminosav ψ és ϕ torziós szögei, az oldalláncok χ_1 és χ_2 torziós szögei,
- a C_α atomok háromdimenziós koordinátái.

Az általunk írt SDB program képes helyben futtatni a DSSP-t, vagy letölteni a kimeneti állományokat egy külön megadható URL-ről. A DSSP kimeneti állományának feldolgozását is elvégzi.

A feldolgozás során a következő lépéseket hajtjuk végre:

1. Kiválogatjuk a β -redőben vagy β -hídban lévő aminosavakat.
2. Több modell esetén csak azokat tartjuk meg, amelyeket az NMR modellek legalább 80%-a ugyanolyan szerkezeti elemhez tartozónak jósol, így kiszűrve a nagyon instabil redőket.
3. Csak azokat az aminosavakat tartjuk meg, amelyeknek megtaláltuk a párját (ugyanis nem mindegyiknek van, pl. β -kitüremkedés (*bulge*) esetén).
4. Antiparallel párok esetén megállapítjuk, hogy AP10-es vagy AP14-es párról van-e szó. Ehhez el kell dönteni, hogy megfelelően alakít-e két hidrogénkötést a párjával, amelyeknek erőssége legalább 0,8 kcal/mol. Ha igen, akkor AP10, ha nem, akkor AP14.
5. Több modell esetén kiszámoljuk a torziós szögek, valamint a C_α atomok koordinátáinak szórását.
6. Végül pedig eltávolítjuk az egyes párokat. Minden egyes PDB-fájlhoz létrehozunk egy-egy fájlt, amelynek önkényesen `.sheet` kiterjesztést adtunk, a belső formátuma pedig a 10. ábrán látható.

```
#####
Pair number          4
Pair summary:        F_Y (25_59) PARL
Beta sheet label:    A
Beta chain label:    b
Hydrogen bond pattern: PARL
Models taken:        20
Distance of Ca atoms: 5.1 +/- 0.11
First residue
  Amino acid type:    PHE
  PDB sequence number: 25
  DSSP sequence number: 25
  PDB chain code:     A
  PDB insertion code:
  Residue sheet position: ?
  Side chain orientation: g-
  PHI1 +/- delta(PHI1): -106.7 +/- 10.6
  PSI1 +/- delta(PSI1): 112.2 +/- 6.1
  CHI11_1 +/- delta(CHI11_1): -88.8 +/- 3.7
  CHI12_1 +/- delta(CHI12_1): NaN +/- NaN
Second residue
  Amino acid type:    TYR
  PDB sequence number: 59
  DSSP sequence number: 59
  PDB chain code:     A
  PDB insertion code:
  Residue sheet position: ?
  Side chain orientation: g-
  PHI2 +/- delta(PHI2): -114.4 +/- 3.5
  PSI2 +/- delta(PSI2): 168.8 +/- 4.4
  CHI11_2 +/- delta(CHI11_2): -93.4 +/- 0.5
  CHI12_2 +/- delta(CHI12_2): NaN +/- NaN
#####
```

10. ábra: az SDB program által generált sheet fájlok formátuma.

Tapasztalataink szerint sok múlik a szubjektív paramétereken (hidrogénkötés erősségének küszöbértéke, modellek feldolgozása, bifurkált hidrogénhidak figyelembe vétele), ugyanis a legtöbb β -redő nem olyan „szép”, mint ahogy azt a tankönyvi ábrákon látjuk; csavarodik, kitüremkedések vannak benne, a fehérje többi részével is kölcsönhat, gyakran nem egyértelmű, hogy AP14 vagy AP10 párossal van dolgunk. Ezek valamennyire önkényessé teszik β -redők meghatározását.

Az eddig felsorolt műveletek (PDB fájl feldolgozása, DSSP fájl generálása és feldolgozása, sheet fájl generálása) egy jól elkülönülő munkafázist alkotnak, melyet szinkronizációnak hívok. Ha a bemeneti listában szereplő összes fehérjéhez tartozik egy korábbi számolásból származó sheet fájl, akkor a szinkronizáció elhagyható, és rögtön a sheet fájlok feldolgozására lehet térni.

5.5. A párpreferencia értékek meghatározása

A párpreferencia értékeket az SDB program számolja ki. Először feldolgozza a sheet fájlokat, és a benne eltárolt párokat csoportosítja aszerint, hogy AP10, AP14, vagy PARL párok. Megszámolja és csoportosítja az előforduló aminosavakat is. Nagyon fontos, hogy mindegyik aminosav egyértelműen

azonosítható legyen, mert van köztük olyan, amelyik két párnak is tagja (pl. egy β -redő közepén, ezért két párja is van), és figyelni kell arra, hogy ezeket ne hogy duplán számoljuk.

A párpreferencia értékeket az (1) és (2) egyenletek szerint számoljuk ki, azzal a kiegészítéssel, hogy valójában mindegyik pároshoz 4 érték tartozik: a teljes redőre számolt, csak az AP10-es párokra, csak az AP14-es párokra, és csak a PARL párokra. Ezért fontos, hogy a hidrogénkötés mintázat szerint csoportosítani tudjuk a párokat és az aminosavakat is. Vegyük észre, hogy a párok számát tekintve:

$$i, j = 1, 2, \dots, 20 \quad (15)$$

Viszont az aminosavak számára

$$i, j = 1, 2, \dots, 20 \quad (16)$$

éppen azért, mert egy aminosav két párban is szerepelhet.

Fentebb megmutattam, hogy 210 féle aminosav pár létezik. Így egy listához 4 korrelációs táblázatot tudunk elkészíteni, amelyekben összesen 840 különböző párpreferencia érték szerepel.

A program a párkorrelációs táblázatokat `csv` formátumban hozza létre. Minden egyes redő típusra (ALL, AP10, AP14, PARL) egy-egy 20x20-as táblázatban összefoglalja az aminosav párok megfigyelt (o_{ij}) és várható (E_{ij}) gyakoriságát, valamint a párpreferencia értékeket (g_{ij}). A kimenetben ezen kívül szerepel egy-egy összefoglalás az aminosavak számáról minden egyes redőtípusra.

5.6. Különböző adatkészletből származó párkorrelációs táblázatok összehasonlítása

Munkánk kiindulópontja az volt, hogy a különböző adatbázisokból származó eredmények egymástól különböznek. Ennek a különbségnek a mérésére használjuk a (14) egyenletben bemutatott Pearson-féle korrelációs koefficiens, valamint egy másik mérőszámot, a párpreferencia értékek eltérésnek négyzetösszegét:

$$(17)$$

Fontos megjegyezni, hogy az összegzés mind a két esetben 1-től 840-ig megy, tehát az összes lehetséges párpreferencia értéket összehasonlítjuk. Természetesen ez a két érték hasonló tendenciát mutat, hiszen kvalitatíve ugyanazt mérik, az egyetlen különbség, hogy a korrelációs koefficiens -1 és 1 között változik, az eltérések négyzetösszege pedig egy pozitív szám.

Az SDB program ezeket az értékeket szintén táblázatba foglalja. Hogy két adatbázis hasonlóságának vagy különbözőségének az okát fel tudjuk deríteni, a program készít még két további táblázatot is, amelyben a listákban közös fehérjeláncok számát és százalékos arányát tünteti fel, a kisebb listához viszonyítva. Például egy tisztán NMR- és tisztán röntgenszerkezetekből álló lista

hasonlósága 0%, de két olyan listáé, amelyek közül az egyik teljesen tartalmazza a másikat, plusz még további szerkezeteket is, 100%.

5.7. Adatbázisok önálló jóságának jellemzése

Már utaltam rá, hogy a különböző adatbázisok párpreferencia értékeinek összehasonlítása nem mindig vezet eredményre, mert két „rossz” adatbázis is lehet hasonló. Ezért mindenképpen szükség van egy olyan módszerre, amellyel egy adott adatbázis jóságát tudjuk becsülni, függetlenül attól, hogy honnan származik.

A 4. fejezetben részleteztem, hogy a χ^2 -próba nagy adatbázisok esetén nem alkalmazható, mert válogatás nélkül mindegyik párpreferencia értékre 100%-os szignifikanciát ad. Ugyanakkor magát a χ^2 -értéket elvileg fel tudjuk használni a párok jellemzéséhez (3. egyenlet). Még ez sem tökéletes megoldás, ugyanis szintén a 4. fejezetből kiderül, hogy két, nem azonos méretű adatbázis χ^2 -értékei mindenképpen különbözni fognak, tehát különböző méretű adatbázisok jósága így nem hasonlítható össze. A 8. ábrán látszik, hogy a χ^2 -értékek nagyjából lineárisan korrelálnak az adatbázis méretével, ami azt az ötletet adja, hogy próbáljuk meg a kettő hányadosát felhasználni. A χ^2 -teszt minden egyes aminosav párhoz ad egy-egy értéket, így a teljes adatbázis jóságának jellemzéséhez a 840 aminosav párra kapott χ^2 -értékek átlagát használtuk, illetve a χ^2 -értékek átlagának és az adatbázis méretének a hányadosát. Ez utóbbi mennyiséget a továbbiakban a láncok számával normált χ^2 -értéknek fogom hívni.

Egy másik módszert is alkalmaztunk, amely a statisztikában „jackknife-teszt” néven ismert.[128] Ennek során véletlenszerűen elhagyjuk az adatkészlet egy részét, és megvizsgáljuk, hogy az eredmény mennyire változik. Munkánk során következetesen a kiindulási adatkészlet 10%-át hagytuk el, és ezt a műveletet 5-ször végeztük el adatbázisonként. Így minden párpreferencia értékre 5 különböző eredmény született, amelyek az eredeti érték körül ingadoztak. Így minden párpreferencia érték stabilitását jellemezni tudjuk a jackknife-teszt eredményeinek az eredetitől való átlagos eltéréseivel. A teljes adatbázis jóságának becslésére az összes eltérés átlagát használtuk fel, amely számot mostantól átlagos jackknife-ingadozásnak hívok.

A SDB program tehát minden egyes adatbázis jóságának jellemzésére három értéket hoz létre: az átlagos χ^2 -értéket, a láncok számával normált χ^2 -értéket, valamint az átlagos jackknife-ingadozást. Ezeket táblázatosan feltünteteti a kimeneti állományban.

6. A kvantumkémiai elemzés bemutatása

A kvantumkémiai elemzés célja elsősorban valamilyen összefüggés megállapítása volt a párpreferencia értékek és az aminosav párok oldalláncainak kölcsönhatási energiája között. Ennek megfelelően a legkisebb egység, amelyet modelleztünk, két kölcsönható aminosav volt. Hogy a modell jobban hasonlítson a β -redőre, mind a két aminosavhoz hozzákapcsoltunk két-két további aminosavat, amelyek az N-terminálison egy acetilcsoporttal, a C-terminálison pedig egy metil-amid csoporttal védett alaninmolekulák voltak. Azért esett az alaninra a választás, mert kicsi, apoláris oldallánca van, amely nem alakít ki erős kölcsönhatásokat, viszont jelentős preferenciát mutat a β -redő felé. Az aminosavak torziós szögeit a kiindulási állapotban a β -redőnek megfelelő értékekre állítottuk be ($\Phi = -130^\circ$, $\Psi = 130^\circ$), és a két tripeptidet egymással szembe helyeztük, hogy az így létrejövő hidrogénkötés mintázat megfelelő legyen (11. ábra). Munkánk során az AP10-es aminosav párokat modelleztük. Mivel a számításokat több elméleti szinten is elvégeztük, a limitált számítási kapacitás miatt négy, nem túl nagy aminosav közötti kölcsönhatás modellezésére szorítkoztunk, melyek között többféle természetű aminosav szerepel, ugyanakkor a β -redőben egymással többnyire szignifikáns pozitív vagy negatív korrelációt mutatnak (Gly, Ala, Val, Ser).

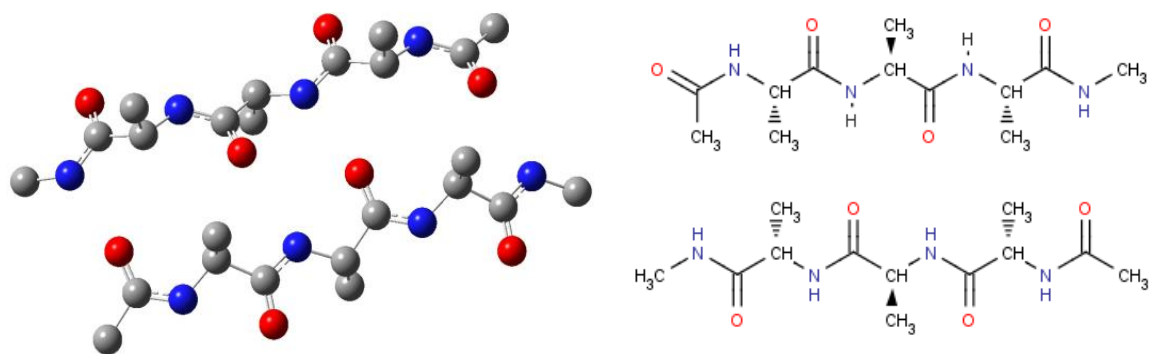
Így összesen 7 tripeptid dimer modell elemzését végeztük el. Fontos megjegyezni, hogy a valin és a szerin oldalláncának 3 lehetséges stabil orientációja van (*gauche+*, *gauche-*, *transz*), ami megnövelte a modellezendő szerkezetek számát (6. táblázat).

A számoláshoz a Gaussian09 programot használtuk.[129] Teljes szerkezetoptimalást, valamint frekvenciaszámolást végeztünk, ez előbbi GDIIS módszerrel, szigorúbb konvergenciakritériumokat támaztó *tight* pontossággal. Mindegyik szerkezetre counterpoise korrekcióval, és anélkül is elvégeztük a számításokat. A következő elméleti szinteken számoltunk:

- HF/3-21G
- B3LYP/6-31G*
- B3LYP/6-31+G**
- M052X/6-31G*
- M052X/6-31+G**

6. táblázat: modellezett tripeptid dimerek, és az oldalláncok lehetséges rotamerjeinek száma.

Modellek	rotamerek
AGA-AGA	1
AAA-AAA	1
AAA-AGA	1
AVA-AGA	3
AVA-AAA	3
AVA-AVA	6
AVA-ASA	9



11. ábra: modellezett tripeptid dimer háromdimenziós szerkezete és szerkezeti képlete. Az ábrán egy AAA-AAA látható.

A kvantumkémiai számolás eredményeképpen előálló paraméterek közül az energiát és a szabadentalpiát használtuk fel. Ezen kívül mindegyik szerkezetre meghatároztuk a stabilizációs energiát illetve szabadentalpiát, amely a két lánc kölcsönhatásakor felszabaduló energiát jelenti, az alábbi képlet szerint. Így mindegyik szerkezetről négy különböző, energia típusú mennyiséget kaptunk:

- ΔG
- ΔE
- $\Delta_{\text{stab}}G(\text{AB}) = \Delta G(\text{AB}) - [\Delta G(\text{A}) + \Delta G(\text{B})]$
- $\Delta_{\text{stab}}E(\text{AB}) = \Delta E(\text{AB}) - [\Delta E(\text{A}) + \Delta E(\text{B})]$

Mivel a β -redők a tapasztalat szerint a fehérje belső, eltemetett régióiban helyezkednek el, ahol viszonylag kevés kölcsönhatást tudnak az oldószer molekuláival kialakítani, nem használtunk oldószermodellt.

A párpreferencia értékek közvetetten az egyes aminosavak, illetve aminosav párok előfordulási gyakoriságától függenek. Egy rendszer különböző állapotainak energiája és az állapotok gyakorisága között teremt kapcsolatot a Boltzmann-eloszlás [130]:

$$\frac{n_i}{N} = \frac{e^{-\epsilon_i/kT}}{\sum_j e^{-\epsilon_j/kT}} \quad (18)$$

ahol n_i az i -edik állapot gyakorisága, N a teljes gyakoriság, ϵ_i az i -edik állapot energiája, k a Boltzmann-állandó, T pedig a hőmérséklet. A mi esetünkben n_i/N helyettesíthető egy párpreferencia értékkel, ugyanis az nem más, mint egy súlyozott előfordulás, a súlyfaktor pedig a korreláción nem változtat, csak egy konstanssal szoroz. Így, az egyes párpreferencia értékek logaritmusának függvényében ábrázolva a párok közötti kölcsönhatási energiákat, egy egyenest várunk. Állandó hőmérsékleten és nyomáson a kölcsönhatást indokoltabb szabadentalpiával leírni, melyhez a több gépidőt igénylő frekvenciaszámítás elvégzésével jutunk.

7. Az eredmények diszkussziója

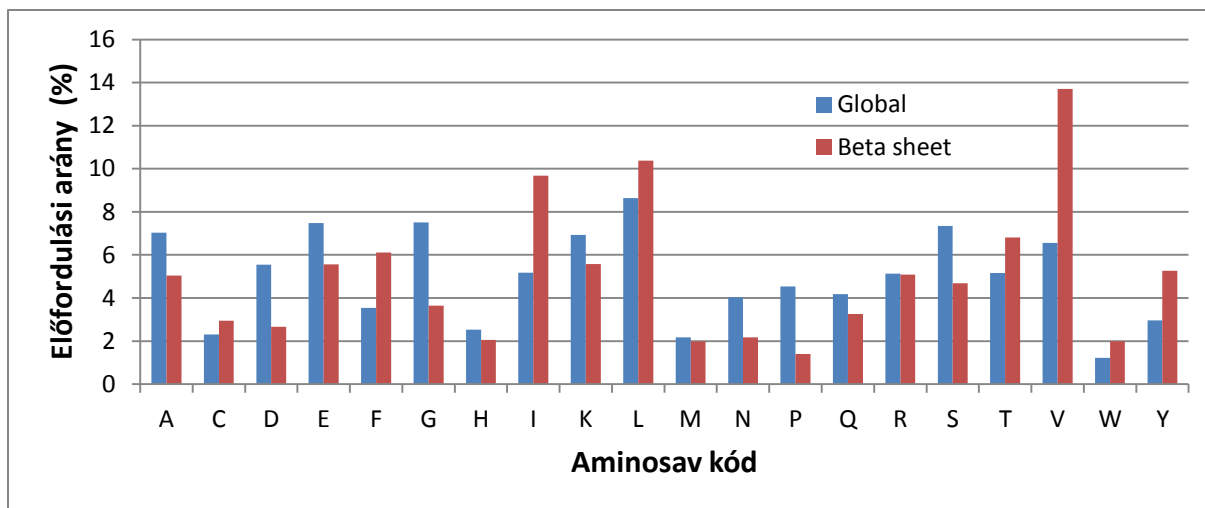
7.1. Aminosavak eloszlása

Mint már korábban mások, mi is kimutattuk bizonyos aminosavak preferenciáját a β -redők felé (12. ábra).[85] Ezek a Phe, Ile, Leu, Thr, Val, Trp, Tyr. (Ez a sorozat megegyezik Levitt 1978-as listájával, a leucinnal kiegészítve.) Ezek nagyrészt apoláris vagy aromás aminosavak, amely megegyezik azzal az elképzelésünkkel, hogy a β -redők alkotják a fehérje hidrofób magját.

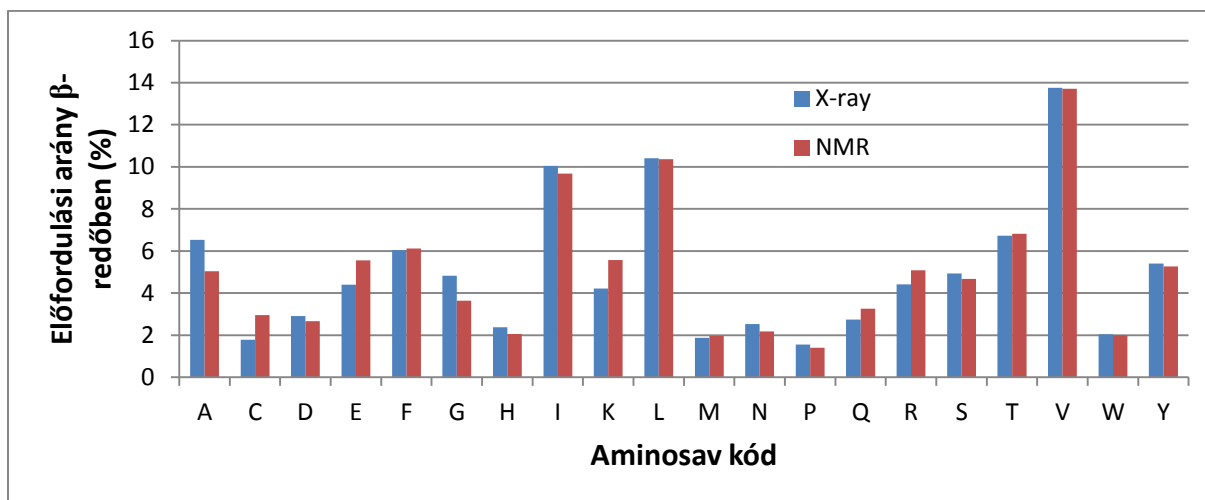
Észrevettünk egy másik egyenlenséget: az aminosavaknak a teljes fehérjében megfigyelt gyakorisága nem azonos a röntgenszerkezetekben, mint az NMR-szerkezetekben, és az eltérés sok esetben szignifikáns (13. ábra). Egy fehérje tulajdonsága természetesen nem függhet attól, hogy a kutatók milyen módszerrel határozták meg a szerkezetét, fordítva viszont már annál inkább igaz lehet az állítás. Az NMR spektroszkópiával nem, vagy csak nagyon nehezen lehetséges a 100 aminosavnál hosszabb fehérjéket vizsgálni. A kisebb fehérjékben pedig nagyobb a felület/térfogat arány, ennél fogva érthetővé válik, hogy nagyobb arányban fordulnak elő bennük a hidofil oldalláncú (Gln, Thr, Cys), valamint sav-bázis tulajdonságú aminosavak (Arg, Lys, Glu), melyek közvetlenül kölcsönhatnak az oldószer molekuláival. A mozgékonyaságot valamelyest növelő glicin és alanin pedig kisebb arányban fordulnak elő, a nagyon mozgékony fehérjéknek ugyanis nincs stabil háromdimenziós térszerkezete, így ezeket nem is lehet NMR-rel meghatározni.

Ennek megfelelően különösen figyelniük kell arra, hogy a párkorrelációk számolásánál a betanító adatkészletet célszerűen válogassuk meg annak megfelelően, hogy a kapott párpreferencia értékeket milyen célra akarjuk felhasználni.

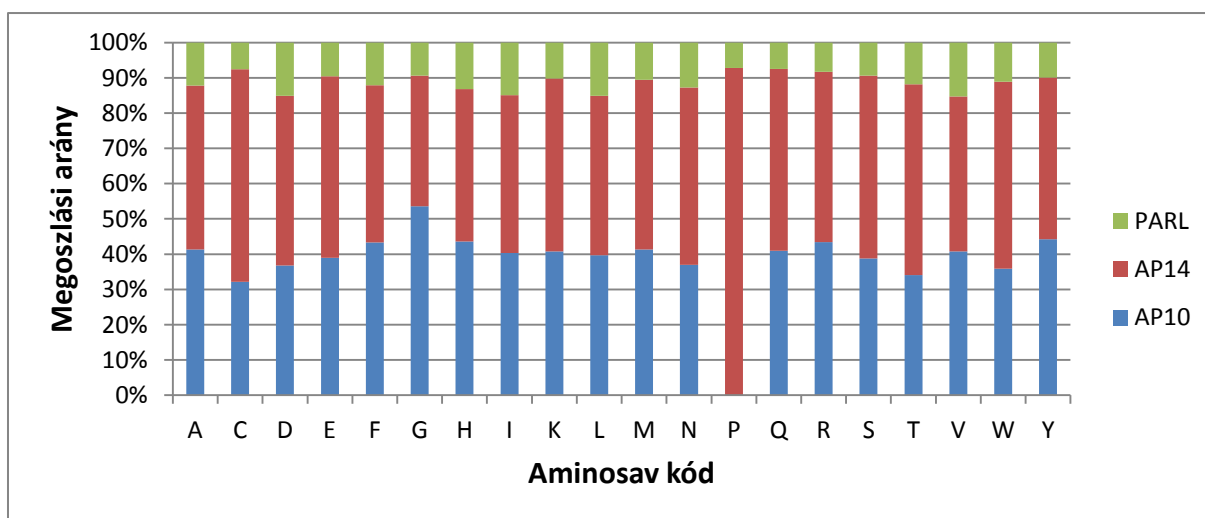
A 14. ábrán azt figyelhetjük meg, hogy az egyes aminosavak nem egyforma arányban alakítanak ki párokat a különböző hidrogénkötés mintázatokban. Látszik, hogy parallel redőből sokkal több van, mint antiparallelből, de azon belül az AP10 és AP14 pszeudogyűrűk száma nagyjából megegyezik. Kiugró, hogy prolin nem fordul elő AP10 gyűrűben, hiszen hidrogénkötésben nem tud donor lenni, és PARL gyűrűben kisebb arányban fordul elő, mivel ott is minden második aminosav két hidrogénkötést alakít ki (4. ábra). Cisztein relatíve nagy arányban fordul elő AP14-es gyűrűben; később látni fogjuk, hogy ez a formáció a diszulfid híd kialakításának kedvez. A glicin kisebb arányban fordul elő AP14-es gyűrűben. Ez azzal magyarázható, hogy nincs oldallánca, ezért nem tudja kialakítani azokat a kedvező kölcsönhatásokat a vele szemben lévő aminosavval, amelyet a C_{α} atomok közelsége általában lehetővé tesz; ezeken a pozíciókon inkább más aminosavakkal helyettesítődik.



12. ábra: aminosavak százalékos előfordulási aránya β -redőben (Beta sheet) és a teljes fehérjében (Global).



13. ábra: aminosav százalékos előfordulási aránya a teljes fehérjében röntgenkristallográfiával, illetve NMR spektroszkópiával meghatározott szerkezetek esetén.



14. ábra: az egyes aminosavak megoszlási aránya AP10, AP14 illetve PARL pszeudogyűrűkben. Nincs értelme számszerűen feltüntetni a konkrét előfordulásokat, hiszen egy adott aminosav egy vagy két pszeudogyűrűnek is tagja lehet.

7.2. A párpreferencia értékek általános tendenciái

A párkorrelációk általánosságban ugyanazt a kvalitatív képet mutatják, mint amit a korábbi munkák ezzel kapcsolatban feltártak.[3, 23, 89] A legerősebb korrelációk az ellentétes töltéssel rendelkező, vagy nagyon poláris oldalláncok között vannak (Glu-Lys, Glu-Arg, Asp-Lys, Asp-Thr, Glu-Thr, stb.), ezeket követik az aromás-aromás és az apoláris-apoláris kölcsönhatások (Phe-Tyr, Ile-Leu, Leu-Val, stb.) Jelentős pozitív korreláció van az aromás és az apoláris oldalláncok között is. (Leu-Trp, Leu-Phe, Val-Tyr, stb.) A legerősebb antikorreláció az azonos töltések között figyelhető meg (Lys-Lys, Glu-Glu, Arg-Arg, Lys-Arg), valamint az apoláris-poláris és az apoláris-töltött párok között (Leu-Asp, Leu-Glu, Leu-Gln) stb.

Bizonyos aminosav párok párkorrelációs értéke jelentősen eltér a különböző pszeudogyűrűk esetén. Ennek tipikusan az az oka, hogy az oldalláncok nagyon kedvezően tudnak kölcsönhatni, ha egymáshoz képest a megfelelő orientációt veszik fel, ehhez viszont kulcsfontosságú, hogy a C_{α} atomok milyen távol vannak egymástól, illetve hogy milyen relatív orientációban van a két oldallánc. Eklatáns példája ennek a már emlegetett Cys-Cys páros, amelyhez tartozó párpreferencia érték AP14-es gyűrűben 4 fölé van (ebben az esetben kerülnek egymáshoz elég közel az oldalláncok, hogy diszulfid hidakat hozzanak létre), AP10-esnél 1,5 körüli, míg PARL gyűrűben egyenesen antikorreláció tapasztalható, 0,5 körüli párpreferencia értékkel. Az Ala-Trp páros erős korrelációt mutat a PARL gyűrűben ($g = 2,0$), ahol az alanin metilcsoportja megfelelően tud orientálódni az aromás gyűrűhöz, az apoláris kölcsönhatás kialakítása érdekében. A Glu-Lys páros hasonlóan csak PARL gyűrűben tudja felvenni a kedvező orientációt ($g = 3,5$), az Asp-Lys viszont éppen ebben a formációban mutat alacsonyabb korrelációt ($g = 1,7$ vs. $2,3$ AP10-ben és AP14-ben).

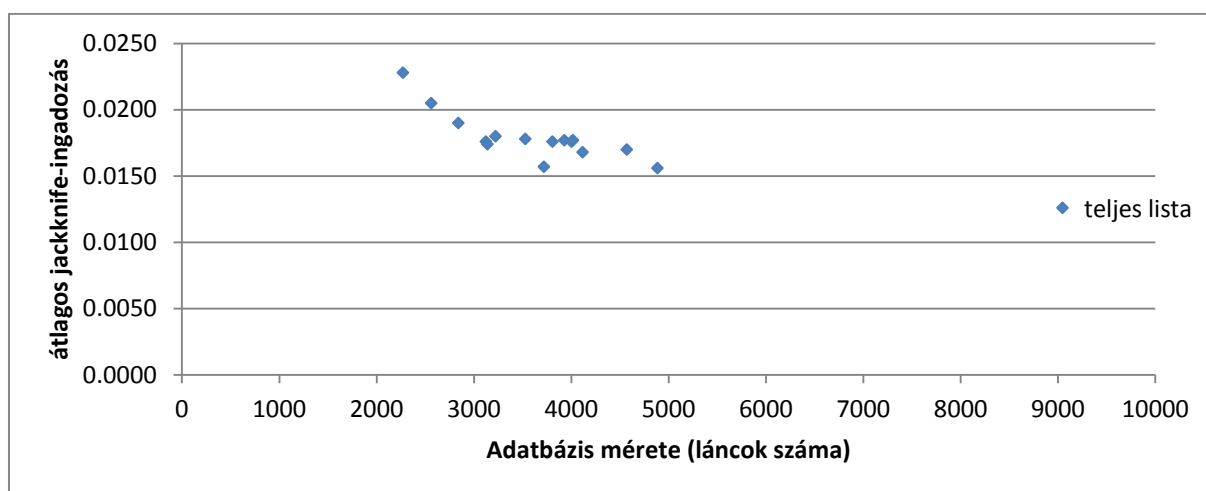
Hogy áttekinthetők legyenek egy párkorrelációs táblázat általános tendenciái, az egyik nemredundáns adatbázisból (NMR, cd-hit, 40%) származó korrelációs táblázatot és aminosav előfordulásokat feltüntettem a 1-7. mellékletekben.

7.3. NMR-es adatbázisok vizsgálata

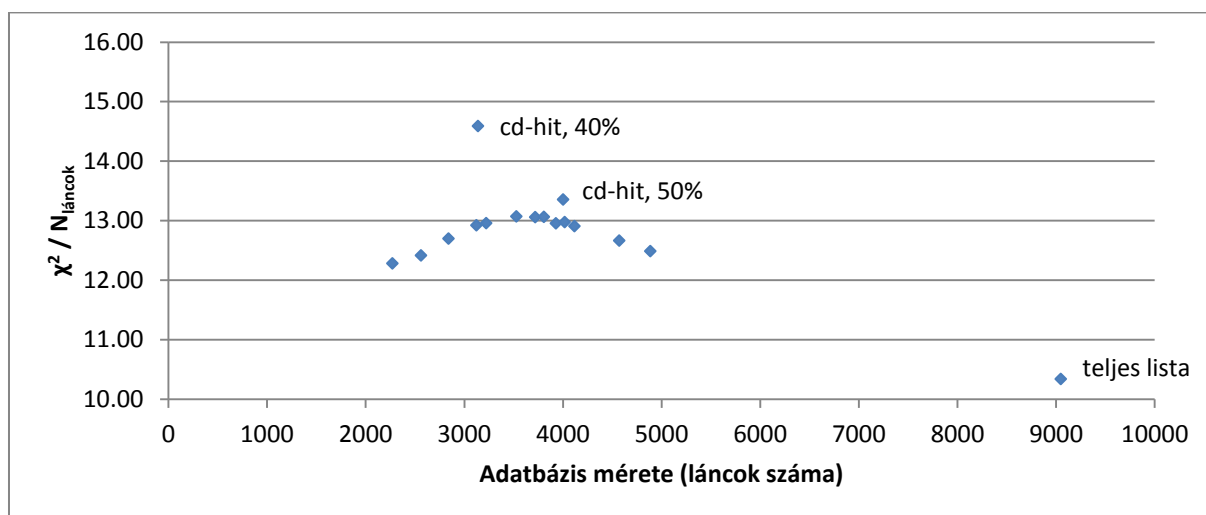
A 4. táblázatban tüntettem fel azokat a nemredundáns NMR-es listákat, amelyeken elvégeztük a párkorrelációs analízist. Először vegyük számba, hogy melyek a megfelelő módszerek, amelyekkel az adatbázis jóságát tudjuk jellemezni. A 8. ábrán megmutattam, hogy a χ^2 -érték lineárisan nő az adatbázis méretével, ami alkalmatlanná teszi erre a feladatra.

Ezért kipróbáltunk két másik módszert is: a láncok számával normált χ^2 -értéket, és az átlagos jackknife-ingadozást. Azt tapasztaltuk, hogy a jackknife-értékek szintén korrelálnak az adatbázis

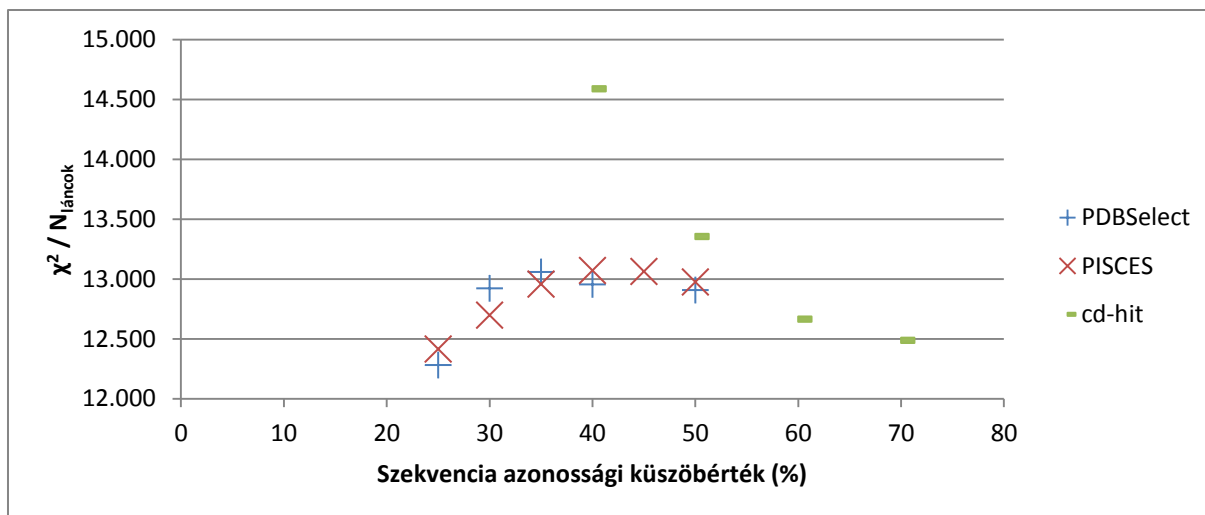
méretével, csak éppen fordítottn (15. ábra). Ez nem meglepő; hiszen ha egy nagyobb adatbázisból hagyunk el akár 10%-ot, akkor is egy nagy adathalmaz marad, és várható, hogy ennek a halmaznak a stabilitása nagyobb, mintha azt a műveletet egy kisebb halmazon végeztük volna el. A jackknife-teszt tehát megerősíti azt a logikus feltevést, hogy a nagyobb adatbázisok megbízhatóbb eredményt adnak. A láncok számával normált χ^2 -érték ellenben egy maximumot mutat a közepes méretű adatbázisok környékén (16. ábra). Ez azt sejteti, hogy van olyan szekvencia azonossági küszöb, amely jobb adatbázist eredményez, mint a többi.



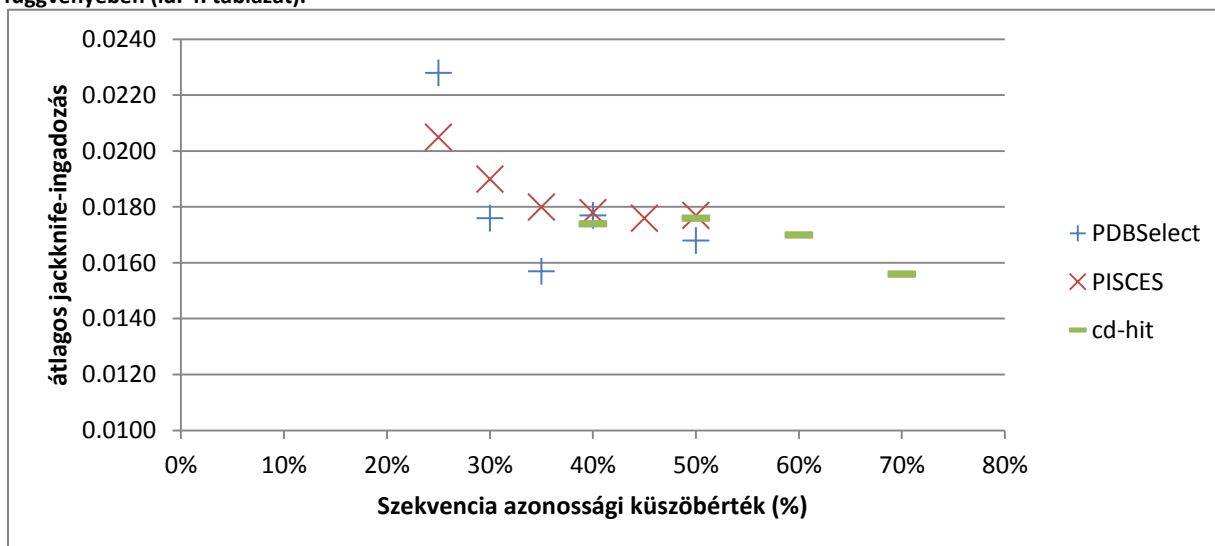
15. ábra: átlagos jackknife-ingadozás értékek az NMR-es adatbázisokból számolva, az adatbázis méretének függvényében ábrázolva (ld. 4. táblázat).



16. ábra: a láncok számával normált χ^2 -érték az NMR-es adatbázisokból számolva, a kiugró pontok megjelölésével (ld. 4. táblázat).



17. ábra: a láncok számával normált χ^2 -érték az NMR-es adatbázisokból számolva, a szekvencia azonossági küszöb függvényében (ld. 4. táblázat).



18. ábra: átlagos jackknife-ingadozás az NMR-es adatbázisokból számolva, a szekvencia azonossági szint függvényében (ld. 4. táblázat).

Igen hasznos információkat nyerhetünk mind a különböző szűrési algoritmusokról, mind pedig az adatbázisban található redundancia mértékéről, ha ezt a két jósági paramétert az egyes adatbázisok szekvencia azonossági küszöbértékének függvényében ábrázoljuk (17-18. ábra). A jackknife-értékek valamennyire megtartják csökkenő tendenciájukat az azonossági küszöbérték növekedésével, ami, figyelembe véve, hogy a magasabb küszöbérték több láncot enged át a szűrőn, megmagyarázható az iménti érveléssel. Azonban az ábrán mintha azt vehetjük észre, hogy a csökkenés megtorpan 35-40%-körül, sőt mind a cd-hit, mind a PDBSelect adatsornak minimuma van 35%-nál. Ezzel összhangban van, hogy a láncok számával normált χ^2 -értékek éppen 35-40%-nál mutatnak maximumot (PISCES 40%-nál, a PDBSelect 35%-nál). Ezeket a megfigyeléseket annak alátámasztására hangsúlyozom, hogy a redundanciaszűrés, bár csökkenti az adatbázis méretét, összességében mégis hasznos, mégpedig a legeredményesebb a 35-40%-os szekvencia azonossági

küszöbérték. Természetesen hozzá kell tennem, hogy célszerű lenne a vizsgálatot megismételni más kiindulási mintán is, ugyanis ezek a számok nem függetlenek az adatbázis méretétől. Könnyen lehet, hogy még alacsonyabb küszöbérték még jobb lenne, de az már annyira lecsökkenti a minta méretét, hogy az imént bemutatott statisztikai analízis már inkább rosszabbnak fogja jósolni.

Megvizsgáltuk azt is, hogy az adatbázisok mennyire hasonlítanak egymásra. Mivel minden lista ugyanabból az NMR-es adatkészletből született, a tartalmuk lényegesen nem különbözik egymástól: bármely két listában szereplő fehérjeszerkezetek legalább 40%-a azonos. Ennek megfelelően a belőlük számolt párkorrelációs táblázatok is igen hasonlóak: a legalacsonyabb számolt Pearson-korrelációs koefficiens nem kisebb, mint 0,78. Kontrollként elvégeztük a teljes NMR-es adatkészletre is a párkorrelációs analízist. Mivel a teljes lista nagyfokú redundanciát mutat, ezért még a legkevésbé szigorú redundanciaszűrés is (cd-hit, 70%) felére csökkenti a benne lévő láncok számát (4. táblázat). Azt várnánk, hogy a belőle számolt párkorrelációs táblázat szignifikánsan alacsonyabb hasonlóságot mutat majd a többivel, mint a többi egymással, ami valóban teljesül is. Ám ha megfigyeljük azt a tendenciát, ami megmutatkozik az egyes szűrési algoritmusok között, akkor azt vesszük észre, hogy az azonos algoritmussal készült adatkészletek jobb egyezéseket mutatnak, mint a különbözőek (7. táblázat). Így érthető, hogy a teljes lista kilóg ebből a sorozatból.

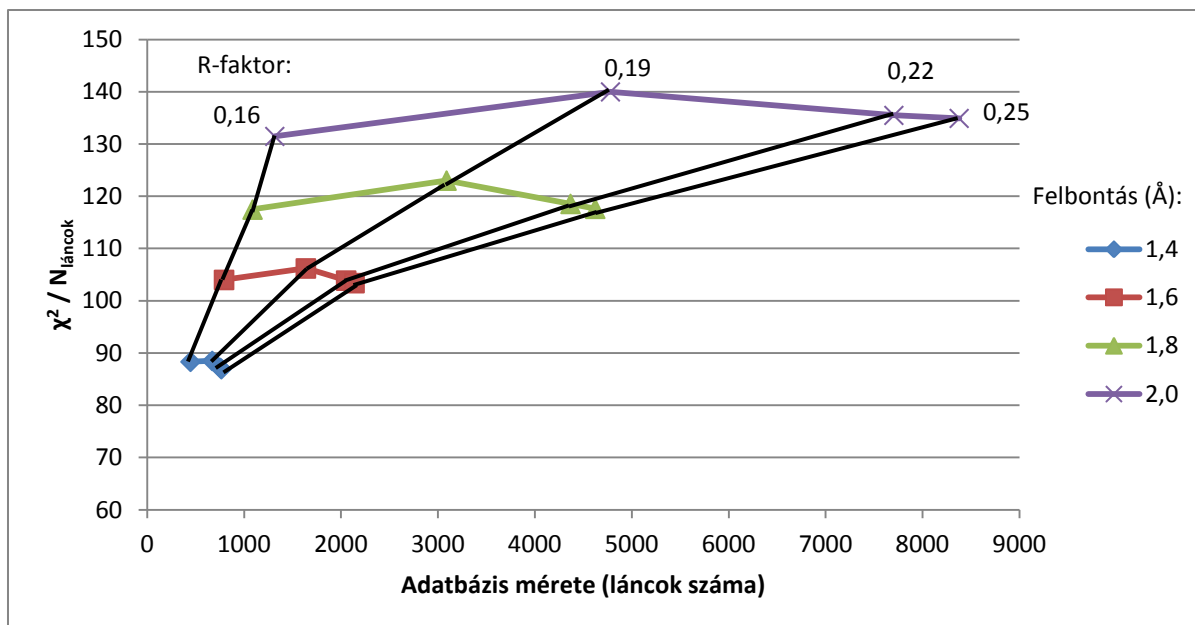
7. táblázat: a különböző szűrési algoritmusokkal készített listákból származó párkorrelációs táblázatok átlagos Pearson-korrelációs koefficiensei (ld. 4. táblázat).

		1.	2.	3.	4.
PDBSelect	1.	0,97	0,89	0,91	0,89
PISCES	2.	0,89	0,97	0,96	0,84
cd-hit	3.	0,91	0,96	0,99	0,89
teljes lista	4.	0,89	0,84	0,89	1,00

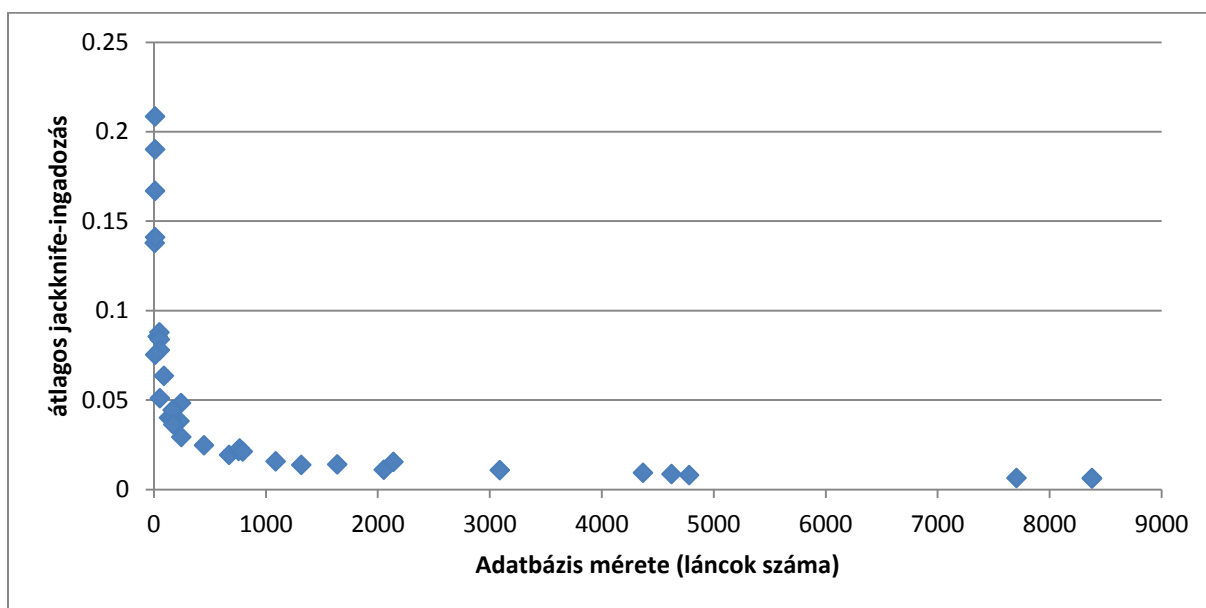
Az egyes szűrési algoritmusok között nehéz egyértelmű különbséget tenni. Figyelmesen végigtanulmányozva az eredményeket az derül ki, hogy összességében a cd-hit jobban korrelál a másik kettővel, mit azok egymással, ezen kívül a jósági becslések is kedvezőbb értékeket adtak neki, de valójában nincsen sokkal lemaradva a PDBSelect és a

PISCES sem. Ugyanakkor a 17. ábrán figyelemre méltó a cd-hit 40% és 50%-hoz tartozó pontjainak kiugró értéke. Észben kell tartanunk, hogy a cd-hit klaszterezi a fehérjéket, és minden klaszterből a leghosszabb fehérjeláncot választja ki. Mivel az alacsony szekvencia azonossági küszöbértékek esetén relatíve kevés klaszter jön létre, a kiválasztott láncok átlagos hosszúsága nagyobb lesz, és ezért az egy láncra eső aminosavak száma is megnövekszik. Ebből következik, hogy a láncok száma szerint normált χ^2 -értéknek is növekednie kell. Ez a hatás kétségtávol előnyössé teszi számunkra a cd-hit szűrési algoritmusának használatát.

Összességében azt mondhatjuk, hogy az NMR-es fehérjeszerkezeteken elvégzett párkorrelációs analízis rávilágított a redundanciaszűrés fontosságára, és hozzásegített minket kétféle jósági paraméter helyes értelmezéséhez. Az NMR-es adatbázisokkal kapcsolatos pontos számolási adatok megtalálhatók a 8-10. mellékletben.



19. ábra: röntgenszerkezetekből származó nemredundáns adatbázisok normált χ^2 -értékei a láncok számának függvényében ábrázolva. Az azonos maximális felbontással, illetve R-faktorral rendelkező adatbázisok pontjait összeköttöttük. A nagyon kicsi felbontású illetve R-faktorú szerkezeteket nem ábrázoltuk.



20. ábra: röntgenszerkezetekből származó nemredundáns adatbázisok párpreferencia értékeinek átlagos ingadozása a láncok számának függvényében.

7.4. Röntgenszerkezetekből álló adatbázisok vizsgálata

A röntgenszerkezetek vizsgálatától azt reméltük, hogy megtudjuk, függ-e valamilyen módon a röntgenszerkezetek jósági paramétereitől (felbontás, R-faktor) a párpreferencia értékek megbízhatósága. A felbontás azt jelenti, hogy a diffrakciós mérés alapján mennyire pontosan tudjuk megmondani az egyes atom-atom távolságokat a molekulában, az R-faktor pedig a felépített modell

és a mért adatok eltérését jelzi; ennél fogva mind a két paraméter növekedése a röntgenszerkezet romlását jelenti.

Az 5. táblázatban vannak felsorolva azok a röntgenes listák, amelyekre elvégeztük a pátkorrelációs analízist. Látszik, hogy a listák méretei között akár százszoros, ezerszeres különbségek is vannak. Így azt várjuk, hogy az adatbázis stabilitásában is nagy eltérések lesznek.

A tapasztalatok az elvárásoknak megfelelően alakultak: a láncok számával normált χ^2 -értékek nőnek, a jackknife-ingadozások pedig csökkennek az adatbázis méretével. A 20. ábrán látható, hogy az átlagos jackknife-ingadozások a láncok számának függvényében nagyon hamar lecsengenek, ami a korábbi érvelésnek megfelelően azt jelenti, hogy az adatbázis méretével rohamosan javul annak stabilitása.

A 19. ábrán megfigyelhető a láncok számával normált χ^2 -értékek függése külön-külön a felbontástól és az R-faktortól. A 0,16-nál alacsonyabb R-faktorú, és az 1,4 Å-nél jobb felbontású szerkezeteket tartalmazó adatkészleteket nem ábrázoltuk, mivel kis méretük miatt a jóságuk sokkal rosszabb a többinél.

Világosan megmutatkozik, hogy az adatbázis jósága az R-faktor függvényében 0,16 körül egy maximumon megy át. Azt látjuk tehát, hogy nem érdemes nagyon szigorúvá venni az R-faktor szerinti szűrést, mert akkor az adatbázis méretének csökkenése a megbízhatóság rovására megy, ugyanakkor válogatás nélkül sem szabad a rossz R-faktossal rendelkező szerkezeteket felhasználni.

A felbontással kapcsolatban elsősorban azt a meglehetősen megállapítást tehetjük, hogy a rosszabb felbontású szerkezeteket tartalmazó adatbázis megbízhatósága magasabb, és ez nem magyarázható minden esetben az adatbázis méretének növekedésével, ugyanis például a 0,16-os R-faktorú szerkezetek esetén az adatbázis mérete csak kismértékben nő a felbontással, a jószágban viszont jelentős növekedést tapasztalunk. A hatás magyarázatát abban tényben kell keresnünk, hogy a nagyobb méretű fehérjéről általában rosszabb felbontású diffrakciós képet lehet készíteni, mivel minél nagyobb az elemi cella mérete, annál rosszabb lesz a diffrakciós kép jel/zaj aránya. A nagyobb fehérjékben azonban több az aminosav, ezért több aminosav pár esik egy láncre, így a lánckok számával normált χ^2 -értékben is növekedést kell tapasztalunk. Ebből a vizsgálatból tehát nem tudjuk egyértelműen megállapítani azt a küszöbértéket, amelyenél nem érdemes rosszabb felbontású szerkezeteket felhasználni. Ennek eldöntéséhez azt kell figyelembe venni, hogy milyen további információkat szeretnénk kinyerni a fehérjeszerkezetből; például ha szükségünk van az aminosavak oldalláncainak orientációira is, akkor legfeljebb 1,8 Å felbontású szerkezeteket használhatunk.

A redundáns listákból származó diagramokat itt nem mutattam be, mert azokból is ugyanezeket a következtéseket lehet levonni. Az összes lista számolt jósági értékei megtalálhatók a 11. mellékletben.

7.5. NMR- és röntgenszerkezetek összehasonlítása

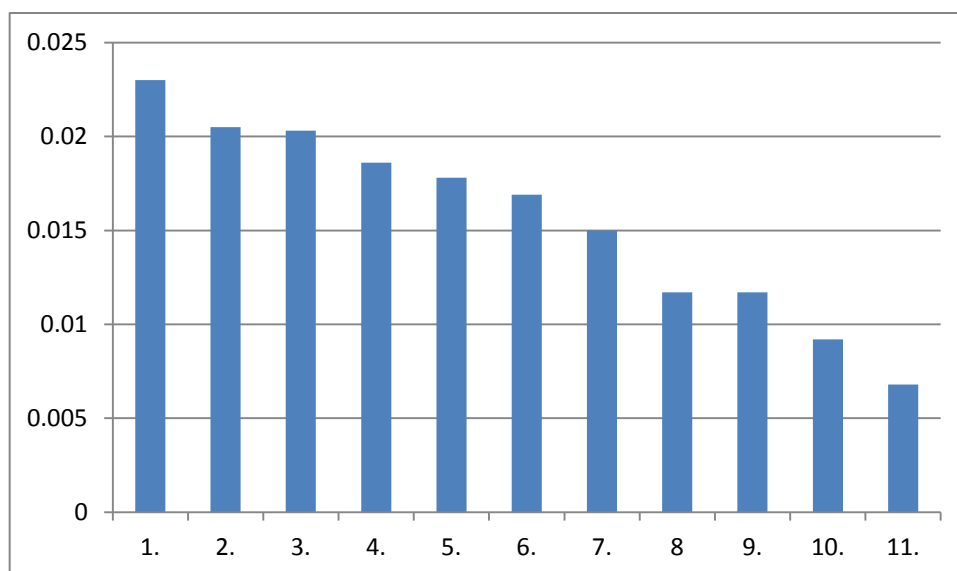
Ehhez a vizsgálathoz az öt előszűrt adatkészletet használtuk fel (2. táblázat), valamint öt kiválasztott NMR-es adatkészletet (PDBselect 25%; PDBselect 30%; PISCES 30%; PISCES 35%; cd-hit 40%, részletekért lásd a 3-4. táblázatot).

Az adatbázisok jóságának jellemzéséhez ezúttal nem tudjuk felhasználni a láncok számával normált χ^2 -értékeket, ugyanis az NMR-es fehérjék általában rövidebbek, és mivel egy láncra így sokkal kevesebb aminosav pár esik, azokra kisebb stabilitást jósolna. Informatív viszont az adatbázisok egymással vett Pearson-korrelációs koefficiense, valamint az átlagos jackknife-ingadozás. A 8. táblázatban látható, hogy az NMR-es és a röntgenes adatok is maguk közt korrelálnak jól, keresztbe viszont rosszul. Az is látszik, hogy a röntgenes adatok korrelációja egymás közt valamivel jobb, mint az NMR-eseké, és ehhez hozzáadva, hogy az átlagos jackknife-ingadozásuk is kisebb, és hogy a teljes listával is szignifikánsan jobban korrelálnak, azt lehet mondani, hogy ebben a szerepben ők bizonyulnak jobbnak.

Ettől persze még szem előtt kell tartanunk a fő következtetést, (amit már a 7.1. fejezetben is hangsúlyoztam), amely szerint az NMR-, illetve röntgenszerkezeteket tartalmazó adatkészletek alapvetően különböznek. A betanító adatkészletet aszerint célszerű összeválogatni, hogy a párpreferencia értékeket mire szeretnénk használni, illetve hogy milyen kölcsönhatásokat szeretnénk megérteni belőle.

8. táblázat: NMR-es és röntgenes adatbázisok párpreferencia értékei között számolt Pearson-féle korrelációs koefficiensek.

			1	2	3	4	5	6	7	8	9	10	11
NMR	PDBselect, 25%	1	1	0,92	0,85	0,87	0,86	0,63	0,65	0,64	0,66	0,68	0,72
	PDBselect, 30%	2	0,92	1	0,91	0,93	0,94	0,7	0,72	0,71	0,72	0,74	0,79
	PISCES, 30%	3	0,85	0,91	1	0,98	0,95	0,72	0,74	0,74	0,75	0,76	0,81
	PISCES, 35%	4	0,87	0,93	0,98	1	0,97	0,72	0,74	0,74	0,75	0,76	0,81
	cd-hit, 40%	5	0,86	0,94	0,95	0,97	1	0,72	0,74	0,74	0,75	0,76	0,81
Röntgen	PDBselect, felb. 1.5, R: 0.18	6	0,63	0,7	0,72	0,72	0,72	1	0,96	0,93	0,93	0,93	0,93
	PDBselect, felb. 1.74, R: 0.18	7	0,65	0,72	0,74	0,74	0,74	0,96	1	0,96	0,96	0,95	0,96
	WHATIF, felb. 1.8, R: 0.18	8	0,64	0,71	0,74	0,74	0,74	0,93	0,96	1	0,99	0,94	0,96
	WHATIF, felb. 1.8, R: 0.19	9	0,66	0,72	0,75	0,75	0,75	0,93	0,96	0,99	1	0,95	0,98
	PISCES, felb.1.8, R: 0.25	10	0,68	0,74	0,76	0,76	0,76	0,93	0,95	0,94	0,95	1	0,98
	teljes lista	11	0,72	0,79	0,81	0,81	0,81	0,93	0,96	0,96	0,98	0,98	1



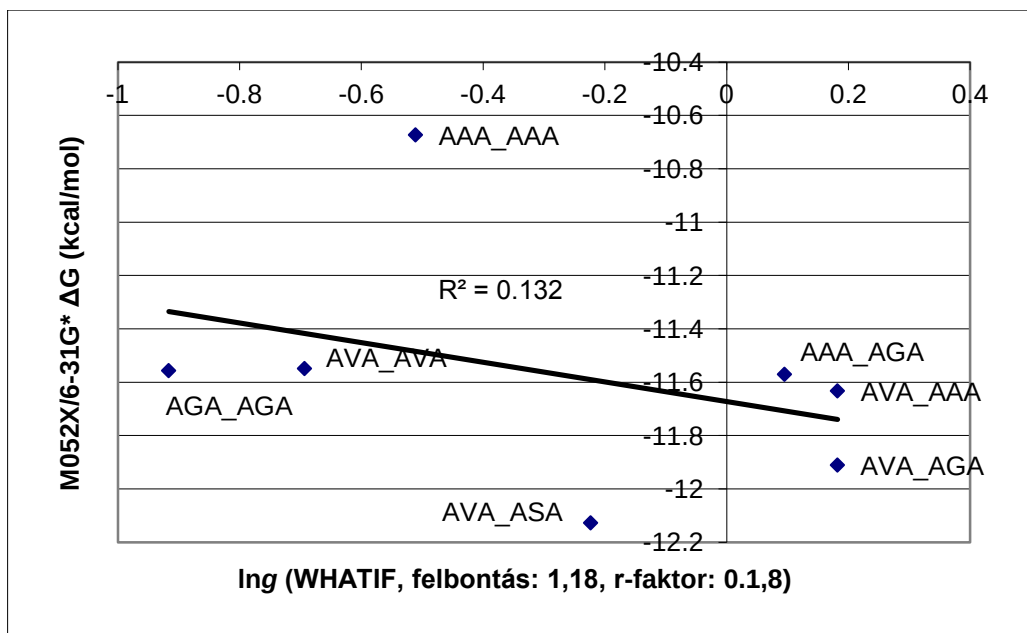
21. ábra: NMR-es és röntgenes adatbázisok átlagos jackknife-ingadozása. Az adatbázisok számozása megegyezik a 8. táblázatával.

7.6. Az aminosav párok gyakorisága és kölcsönhatási energiájuk

A kvantumkémiai analízis során 7 féle tripeptid dimert modelleztünk (6. táblázat). Azt vártuk, hogy a vizsgált aminosav párokat jellemző párpreferencia értékek logaritmusa és a kölcsönhatási energiájuk között a Boltzmann-eloszlásnak megfelelően [lásd (18) egyenlet] lineáris összefüggés áll fenn. A tapasztalat nem igazolta ezt a feltevést. Ellenkezőleg: semmilyen korrelációt nem tapasztaltunk e két mennyiség között. Ennek bemutatása céljából a 22. ábrán az egyik kiválasztott elméleti szinten számolt kölcsönhatási szabadentalpiákat ábrázoltam, de a többi elméleti szinten is hasonló eredményekre jutottunk.

A korreláció hiánya kapcsán megkérdőjelezhető, hogy ebben az esetben jogosan feltételezzük-e a Boltzmann eloszlást. Valójában ezt csak akkor tehetnénk meg, ha a különböző energiájú állapotok egymással termodinamikai egyensúlyban lennének. Egy Ala-Ala pár természetesen nem lehet egyensúlyban egy Ala-Val párossal. Azonban tekinthetünk rájuk úgy is, mint a fehérje két különböző konformerére; a feltekeredés során az egyik esetben a szekvenciában az alaninnal éppen egy másik alanin került szembe, a másik esetben pedig egy valin. A tapasztalatok alapján biztonsággal kijelenthetjük, hogy ezek a konformerek nincsenek egyensúlyban egymással. Ez összhangban van a feltekeredésről alkotott elképzeléseinkkel: a fehérje korántsem véletlenszerűen kerül egyik konformációból a másikba, egészen addig, amíg meg nem találja a globális minimumot, hanem nagyon célirányosan, minden esetben pontosan ugyanabba az állapotba tekeredik fel.[131] Tehát az észrevétel, hogy az aminosav párok előfordulási gyakorisága nincs összefüggésben a kölcsönhatási

energiájukkal, nagyon is informatív: azt a gondolatot erősíti, hogy a fehérjék feltekeredése nem energia-kontrollált folyamat.



22. ábra: tripeptid dimerek kölcsönhatási energiái a párpreferencia értékek logaritmusának függvényében ábrázolva. A párpreferencia értékek (g) származásához lásd a 2. táblázatot.

7.7. Aminosav párok oldalláncorientáció-eloszlása

Mint korábban utaltam rá (6. táblázat), a vizsgált 4 aminosav közül kettőnek (Val, Ser) az oldallánca három lehetséges orientációt tud felvenni: *gauche+*, *gauche-* és *transz*. Ezeket az elnevezéseket az IUPAC-IUB ajánlása szerint használom, és a továbbiakban *g+*, *g-* és *t* rövidítésekkel fogom őket jelölni.[132]

Az előszűrt adatbázisokon (2. táblázat) végzett statisztikai analízis során megszámoltuk, hogy hány adott konformációhoz rendelhető aminosav pár van. Ezek aránya a Boltzmann-eloszlás szerint összefüggésben kell, hogy legyen a rotamerek energiájával. A 6. fejezetben bemutatott kvantumkémiai módszerek segítségével kapott energiaértékekből visszaszámolhatjuk az oldalláncok eloszlását [lásd (18) egyenlet], és összehasonlíthatjuk a megfigyelt eloszlással, ami természetesen adatkészletenként szintén változik. A Val-Val aminosav párra számolt eredményeket tartalmazza a 9. táblázat. A további 3 vizsgált aminosav pár eredményei a 12-13. mellékletekben találhatók.

Referenciának az öt különböző adatkészleten számolt oldalláncorientáció-eloszlás átlagát használtuk fel. A statisztikai módszerek ebben a vizsgálatban nagyon egységesnek mutatkoztak: legfeljebb egy-két százalékos eltérés tapasztalható az oldalláncok különböző adatkészletekből származtatott eloszlása között, és a tendencia mindegyiknél ugyanaz.

9. táblázat: Val-Val párok lehetséges oldallánc-orientációinak eloszlása százalékosan. Felül a kvantumkémiai modellezésből, alul a statisztikai analízisből származó értékek láthatók. SumSq az egyes kvantumkémiai számolt értékek négyzetes eltéréseinek összegét jelenti az átlagos statisztikai értékhez képest.

AP10, AVA-AVA	g+g+ (%)	g+g- (%)	g+t (%)	g-g- (%)	g-t (%)	tt (%)	SumSq
HF/3-21G, ΔE	5	47	23	7	16	2	4257
B3LYP/6-31G*, ΔE	38	20	18	8	12	5	3801
B3LYP/6-31G*, ΔG	8	3	20	1	15	52	282
M052X/6-31G*, ΔE	34	31	10	17	8	0	4578
M052X/6-31G*, ΔG	3	44	11	2	20	20	2520
B3LYP/6-31G*, cp, ΔE	38	18	18	7	11	8	3476
B3LYP/6-31G*, cp, ΔG	8	8	25	3	15	42	410
M052X/6-31G*, cp, ΔE	47	27	6	14	5	0	5513
M052X/6-31G*, cp, ΔG	14	11	53	6	15	1	4171
B3LYP/6-31G*, $\Delta_{stab}E$	20	8	29	3	15	24	1412
B3LYP/6-31G*, $\Delta_{stab}G$	4	1	17	0	10	68	711
M052X/6-31G*, $\Delta_{stab}E$	21	13	34	5	19	8	2629
M052X/6-31G*, $\Delta_{stab}G$	2	64	4	7	19	5	5612
B3LYP/6-31G*, cp, $\Delta_{stab}E$	17	7	27	2	13	35	866
B3LYP/6-31G*, cp, $\Delta_{stab}G$	4	2	23	1	11	59	505
B3LYP/6-31+G**, ΔE	14	10	23	6	20	26	848
B3LYP/6-31+G**, ΔG	7	4	24	2	23	40	276
B3LYP/6-31+G**, $\Delta_{stab}E$	12	6	26	3	16	37	565
B3LYP/6-31+G**, $\Delta_{stab}G$	11	4	28	1	19	37	556
PISCES, felb1.8, r0.25	1	3	13	7	29	48	
WHATIF, felb 1.8, r0.18	2	4	14	5	27	48	
WHATIF, felb 1.8, r0.19	1	5	14	8	24	49	
PDBSelect, felb 1.5, r0.18	1	4	11	8	27	50	
PDBSelect, felb 1.75, r0.18	1	4	12	8	26	49	
átlag	1	4	13	7	27	49	
szórás	0,36	0,60	1,25	1,27	1,86	0,79	

A különböző kvantumkémiai módszerek jóval változatosabbak. Az átlagos statisztikai eloszlástól vett eltérések négyzetösszege (SumSq) alapján értékelhetjük őket. Nem tehetünk olyan kijelentést, hogy valamelyik módszer minden esetben kiugróan jobb eredményt ad, mint a többi. Az M052X funkcionállal számolt eredmények a legtöbb esetben rosszabb egyezést adnak, mint az ugyanolyan bázison és módszerrel számolt B3LYP funkcionállal számolt megfelelőjük. Ez nem felel meg az elvárásainknak, hiszen az M052X módszertől vártuk volna, hogy jobban leírja a diszperziós kölcsönhatásokat.[112] A counterpoise korrekció sem váltja be hozzá fűzött reményeket: nem ad szignifikánsan se jobb, se rosszabb eredményt. Igaz, ezt a módszert elsősorban a kölcsönhatási energiák pontos számításánál lehet alkalmazni, a rotamereloszlások esetén nem feltétlenül célravezető. A legtöbb esetben a ΔG értékekből számolt eloszlás közelebb van a megfigyelthez, mint

a ΔE értékekből számolt. Ennek a magyarázata nyilván az, hogy az entrópikus tag bevétele a leírásba közelebb viszi a modellt a valósághoz. A stabilitási energiákból számolt eloszlás többnyire közelebb van a valósághoz, mint amit közvetlenül a dimerek energiájából számoltunk. A statisztikai adatok szerint a valin *transz* orientációja jóval nagyobb arányban van jelen, mint a másik kettő. Ez összhangban áll a korábbi párkorrelációs munkákkal[89], a felhasznált kvantumkémiai módszerek közül azonban csak a nagyobb bázison végzett számolások reprodukálják ezt az arányt.

A DFT módszerek eredményeinek értékelését megnehezíti az a tény, hogy nincs információnk a számított energia hibájáról, és arra sincs garancia, hogy a nagyobb bázis pontosabb energiát ad. Ugyanakkor megnyugtató az a tapasztalat, hogy a legjobbnak bizonyult, (azaz a statisztikához képest legkisebb négyzetes eltérést mutató) módszerekkel számolt oldalláncorientáció-eloszlások jó egyezést adnak a statisztikai eredményekkel. Éppen a DFT megjósolhatatlan bizonytalanságánál fogva ezek a legjobb módszerek nem azonosak mindegyik vizsgált aminosav pár esetén (lásd 9. táblázat, valamint 12-13. melléklet), de közös bennük, hogy mind a B3LYP funkcionált használják. Összességében kijelenthetjük, hogy összefüggést találtunk az oldalláncorientáció-eloszlás, és a különböző rotamerek energiái között.

7.8. A számolt geometriák összehasonlítása a statisztikai adatokkal

Mivel a kvantumkémiai számolások során szerkezetoptimalást is végeztünk, a kapott szerkezetek lokális minimumban vannak. A feltételezésünk tehát az, hogy az optimált geometriák (legalább hibahatáron belül) egybeesnek a statisztikából kapott átlagos geometriákkal. Hogy ezt ellenőrizni tudjuk, átlagoltuk az egyes kvantumkémiai módszerekkel kiszámolt ψ és ϕ torziós szögeket, és összevetettük azzal, amit az előszűrt adatbázisok (2. táblázat) statisztikai analízise során az összes β -redőben lévő aminosavra történő átlagolással nyertünk. Ezeket az eltéréseket a valinra a 10. táblázat tartalmazza; kiegészítve a torziós szögeknek a statisztikai analízis során nyert standard deviációival. Ez a táblázat felhívja a figyelmünket egy igen fontos tényre: a különböző orientációjú oldalláncok nem ekvivalensek. Ha ez így lenne, akkor a kérdéses aminosav ψ és ϕ torziós szögei nem mutatnának szignifikáns eltérést annak függvényében, hogy az oldallánc milyen konformációt vesz fel. A tapasztalat azonban azt mutatja, hogy ezek az eltérések akár 25° -osak is lehetnek. Ezzel sikerült megerősíteni azt a tényt, hogy bár a különböző oldallánc orientációk könnyen átalakulhatnak egymásba, egy olyan komplex rendszer esetén, mint egy fehérje, érdemes köztük különbséget tenni. Fontos hangsúlyozni, hogy ezt a statisztikai elemzést jó felbontású röntgenszerkezeteken végeztük. Oldatfázisban könnyen lehet, hogy nem tudnánk ilyen határozott képet festeni az oldalláncok konformációiról.

Összehasonlítva az átlagos torziós szög értékeket a kvantumkémia által meghatározottakkal azt tapasztaljuk, hogy az esetek túlnyomó többségében az eltérés nagyobb, mint a mintahalmaz szórása,

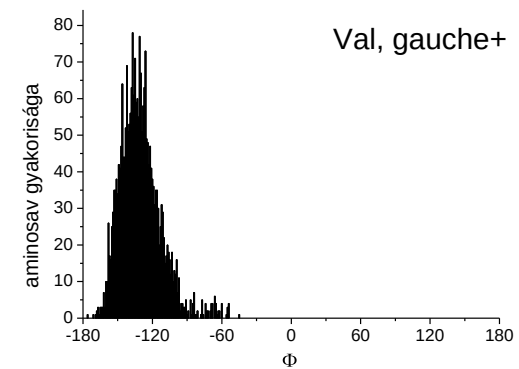
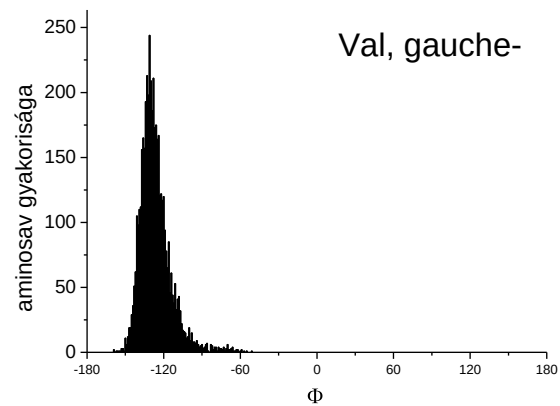
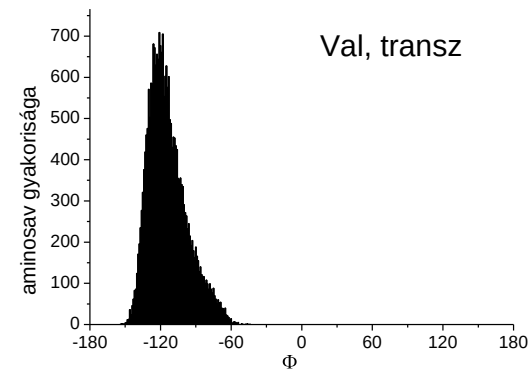
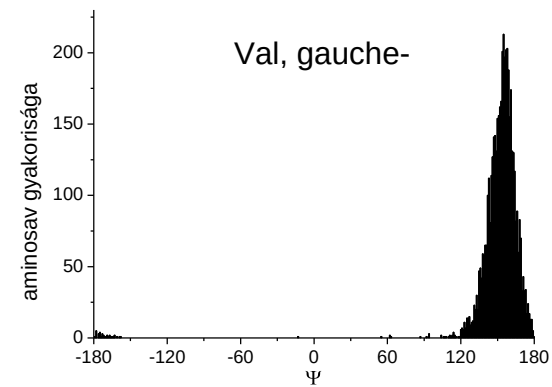
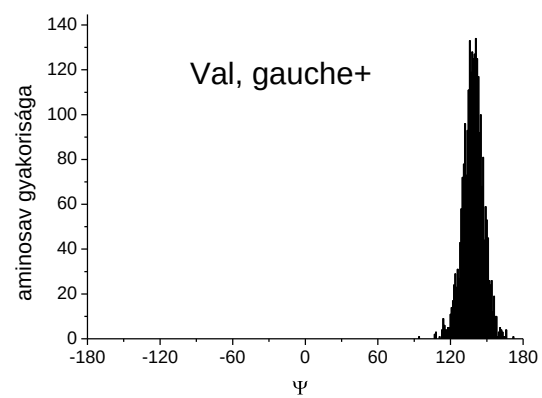
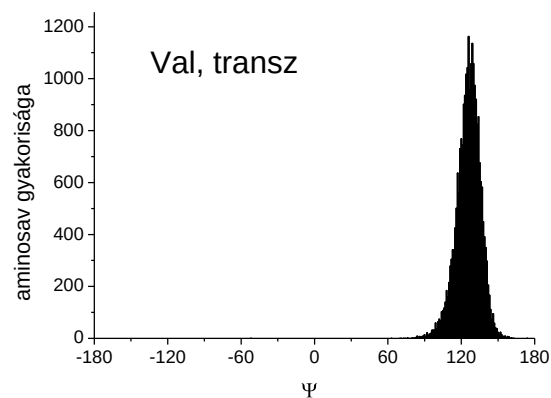
de egyik esetben sem találunk kiugróan magas különbséget. Ha Gauss-eloszlást feltételeznénk, akkor azt mondhatnánk, hogy a kvantumkémiai értékek a statisztikai analízishez képest 90%-os hibahatáron belül vannak. Mivel vannak olyan esetek is, amikor az eltérés nem lépi át a standard deviációt, kijelenthetjük, hogy a kvantumkémia nagyjából ugyanazt a geometriát jósolja, mint a statisztika, csak nagyobb bizonytalansággal. Ez összhangban van azokkal a korábbi munkákkal, melyek szerint a Ramachandran-térkép reprodukálható mint az aminosavak potenciális energia felülete.

A valin torziós szögeinek eloszlását grafikusan is ábrázoltuk (23. ábra). Tanulmányozva az ábrákat meggyőződhetünk arról, hogy az eloszlásnak csak egy maximuma van. Az is látszik, hogy az eloszlás nem teljesen szimmetrikus; ez annak a következménye, hogy az egyes geometriák energiája nem egyformán függ az ábrázolt torziós szögektől az átlagosnál kisebb, illetve nagyobb tartományban. Ez mindenképpen aláhúzza annak szükségességét, hogy az átlagos torziós szögeket a szórásaikkal együtt vizsgáljuk.

10. táblázat: valin aminosav Φ és Ψ torziós szögeinek értéke Val-Ala, Val-Gly, és Val-Ser aminosav párok esetén, oldallánc-orientációkra lebontva. A táblázatban szerepelnek a kvantumkémiaiával számolt értékek (*szám*), a statisztikai analízis eredményeképpen kapott, az öt előszűrt adatbázison (2. táblázat) átlagolt értékek (*stat*), ezek szórása (σ), valamint a kvantumkémiaiával és statisztikával számolt értékek különbsége ($\Delta\Phi = \Phi_{\text{szám}} - \Phi_{\text{stat}}$), $\Delta\Psi$ -re hasonlóan. A számolt szögek a B3LYP/6-31G* és M052X/6-31G* elméleti szinten counterpoise korrekcióval és anélkül számolt eredmények átlagai.

	Val-Ala	Vg+	Vg-	Vt	Val-Gly	Vg+	Vg-	Vt
$\Phi_{\text{szám}}$		-158°	-141°	-165°		-159°	-141°	-140°
$\Psi_{\text{szám}}$		154°	158°	143°		154°	158°	143°
Φ_{stat}		-135°	-128°	-117°		-135°	-128°	-117°
$\sigma(\Phi_{\text{stat}})$		16°	11°	14°		16°	11°	14°
Ψ_{stat}		141°	155°	126°		141°	155°	126°
$\sigma(\Psi_{\text{stat}})$		8°	11°	10°		8°	11°	10°
$\Delta\Phi$		-23°	-13°	-48°		-24°	-13°	-23°
$\Delta\Psi$		13°	3°	17°		13°	3°	17°

Val-Ser	Vg+Sg+	Vg+Sg-	Vg+St	Vg-Sg+	Vg-Sg-	Vg-St	VtSg+	VtSg-	VtSt
$\Phi_{\text{szám}}$	-140°	-160°	-162°	-155°	-142°	-142°	-138°	-142°	-139°
$\Psi_{\text{szám}}$	147°	157°	155°	154°	162°	162°	159°	148°	141°
Φ_{stat}	-138°	-128°	-125°	-122°	-127°	-125°	-117°	-119°	-120°
$\sigma(\Phi_{\text{stat}})$	11°	17°	27°	19°	9°	10°	15°	16°	13°
Ψ_{stat}	142°	137°	140°	155°	147°	153°	127°	126°	127°
$\sigma(\Psi_{\text{stat}})$	4°	6°	7°	11°	8°	18°	9°	11°	9°
$\Delta\Phi$	-2°	-32°	-37°	-33°	-15°	-17°	-21°	-23°	-19°
$\Delta\Psi$	5°	20°	15°	-1°	15°	9°	32°	22°	14°



23. ábra: valin Φ (lent) és Ψ (fent) torziós szögeinek eloszlása gauche+ (jobbra), gauche- (középen) és transz (balra) oldalláncorientációk esetén.

8. Összefoglalás és kitekintés

A párkorrelációk statisztikai és kvantumkémiai módszerrel történő vizsgálata kiegészítik egymást. A párpreferencia értékek a valós fehérjeszerkezetekből kinyerhető információk, amelyek jellemzik az aminosavak viselkedését a β -redőkben. A kvantumkémiai elektronszerkezet-számító módszerek pedig betekintést nyújtanak a párkorrelációs értékek mögött húzódó kölcsönhatásokba, ezzel lehetőséget adnak a statisztikai elemzés eredményeinek helyes interpretációjára.

Most, hogy áttekintettük a statisztikai és kvantumkémiai vizsgálataink eredményeit, összefoglalom, hogy mennyire sikerült megvalósítanunk a dolgozat elején kitűzött célokat (4. fejezet).

1. Az SDB program megírásával létrehoztunk egy olyan eljárást, amely egy jól definiált algoritmussal végzi el a párkorrelációs vizsgálatot bármilyen, tetszőleges méretű és minőségű fehérjeszerkezetekből álló adatkészleten. A korábban kidolgozott eljárást kiegészítettük a különböző adatkészletek összehasonlításának lehetőségével (Pearson-féle korrelációs koefficiens).

2. Megállapítottuk, hogy a χ^2 -teszt nem jól működő validálási módszer, ha az adatbázis nagy, ezért, számolva a fehérjeszerkezeti információk folyamatos bővülésével, két alternatív módszert javasoltunk és teszteltünk egy adatkészlet jóságának a jellemzésére: a láncok számával normált χ^2 -érték meghatározását, illetve a jackknife tesztet. A jackknife teszt az adatbázis konzekvens javulását jósolja a méret növekedésével. Így különböző méretű adatbázisok nem hasonlíthatók össze ezzel a módszerrel, viszont egy konkrét adatkészletből számolt párpreferencia értékek stabilitása külön-külön jellemezhető. A jackknife teszt esetén két paraméter tetszőlegesen megválasztható: az elhagyandó adatok aránya, illetve hogy hányszor ismétljük meg a próbát. A láncok számával normált χ^2 -érték jó lehetőséget nyújt, ha az összehasonlítandó adatbázisok azonos szerkezetvizsgáló módszerrel meghatározott fehérjeszerkezetek vannak. Ha ez nem teljesül, akkor a láncok szignifikánsan eltérő hossza miatt nem alkalmazható; ehelyett inkább az aminosav párok számával normált χ^2 -érték lehet eredményes.

3. Az aminosavak eloszlásában és a párkorrelációs táblázatokban felismertük azokat a tendenciákat, amelyeket előttünk többen leírtak (hasonló aminosav a hasonlóval korrelál).[3, 23, 89] Ugyanakkor a statisztikai vizsgálatok abból az észrevételből indultak ki, hogy a pontos párpreferencia értékek függenek kiindulási adatkészlet számos paraméterétől. Munkánk során sikerült jellemeznünk, hogy hogyan változik egy adatbázis jósága ezekkel a paraméterekkel. A nemredundáns adatkészletek jobbnak bizonyultak, mint a szűretlenek, mindazonáltal a túl szigorú szűrési feltétel nagyon lecsökkenti az adatbázis méretét, ami szintén nem szerencsés, ahogy azt a jackknife-teszttel igazoltuk. Az ideális szekvencia azonossági küszöbértéket 35-40%-ra becsültük, de ezt még meg kell erősíteni nagyobb adatbázisokon végzett számolással is. Kipróbáltunk több szűrési algoritmust

(PDBSelect, PISCES, cd-hit), melyek közül a cd-hit bizonyult a legelőnyösebbnek mind a gyorsasága, mind pedig a klaszterezési algoritmusa miatt. Kiderült, hogy a láncok számával normált χ^2 -érték az R-faktor függvényében 0,16 körül egy maximumon megy át, amely azt jelzi, hogy ésszerű kompromisszumot kell kötnünk a túl szigorú jósági paraméterek és a kellően nagy adatbázis között. Ugyanezzel a jósági becsléssel nem lehet célszerű küszöbértéket megállapítani a szerkezetek optimális felbontására; ennek a másodlagos szerkezeti elemek és az oldallánc rotamerek megbízható azonosíthatósága szab határt. És végül megmutattuk, hogy az NMR-, illetve röntgenszerkezetekből álló adatbázisokból egymástól lényegesen eltérő párpreferencia értékek származnak. Ebben a kérdésben nem érdemes jobb vagy rosszabb módszert keresni; kijelenthetjük, hogy valójában nem létezik „tökéletes” párkorrelációs táblázat. Mindig szem előtt kell tartanunk, hogy oldatfázisban másképpen viselkedik a fehérje, mint kristályos állapotban, és ennek megfelelően kell megválasztanunk a kiindulási adatkészletet.

4. A kvantumkémiai elemzés során 7 tripeptid dimeren végeztünk szerkezetoptimalizálást és frekvenciaszámítást, összesen mintegy 20 módszerrel. Az derült ki, hogy az aminosav párok kölcsönhatási energiája nem függ össze a párpreferencia értékekkel, amiből az a következtetés vonható le, hogy a fehérje feltekeredése, illetve a mi konkrét esetünkben a β -redők kialakulása nem energiakontrollált folyamat. Ugyanakkor röntgenszerkezetekben korrelációt tapasztaltunk a valint és szerint tartalmazó párok oldallánc-orientációinak (*gauche+*, *gauche-*, *transz*) előfordulási aránya, és az aminosav párok kölcsönhatási energiája között. A felhasznált kvantumkémiai módszerek jóságát a kölcsönhatási energiából számolt, és az adatbázisokban tapasztalt oldalláncorientáció-eloszlás közötti eltéréssel tudjuk jellemezni. A B3LYP funkcionál jobbnak bizonyult, mint az M052X, és jobb egyezést adnak a szabadentalpia értékek, mint az energia értékek. Nem hozott azonban javulást a counterpoise korrekció használata.

Bizonyos eredmények teret hagynak további vizsgálatok számára is. A statisztikai elemzésben érdemes lenne tovább finomítani az aminosavak osztályozását, és nem csak AP10, AP14 és PARL párokat vizsgálni, hanem külön kezelni a redő szélén és belsejében elhelyezkedő aminosavakat. Ezt az a tapasztalat indokolná, hogy a szabadon hagyott β -élek a legveszélyesebbek a fehérje aggregáció szempontjából. Indokolt lenne továbbá tesztelni a párpreferencia értékek alkalmazhatóságát aggregációs hajlamot perdiktáló programokban, illetve megvizsgálni, hogy egy adott algoritmus, amelyben az általunk megfelelőnek tartott párpreferencia értékeket használjuk fel, milyen arányban jósolja veszélyesnek a már ismert, körülbelül 30 amiloid prekursor szekvenciát.[34, 35]

A kvantumkémiai elemzésben továbblépési lehetőséget jelente az oldószermodell bevezetése. Korábban amellet érveltem, hogy a β -redők többnyire a fehérje belső régióiban vannak, ezért kevésbé hatnak kölcsön az oldószerrel, ugyanakkor más érvelés szerint a jórészt hidrofób aminosavakból álló β -redő aggregációját éppen az azt körülvevő víz segíti elő. Mindenesetre a fehérje

feltekeredésére vonatkozó megállapításaink valószínűleg nem változnának az oldószermodell használata során, de könnyen lehet, hogy a különböző oldallánc-rotarmerek kölcsönhatási energiáira jobb becslést tudnánk vele elérni.

Irodalomjegyzék

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**(1):235-242.
2. Altman RB, Dugan JM: **Defining bioinformatics and structural bioinformatics.** *Methods Biochem Anal* 2003, **44**:3-14.
3. Wouters MA, Curmi PM: **An analysis of side chain interactions and pair correlations within antiparallel beta-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs.** *Proteins* 1995, **22**(2):119-131.
4. Berg JM, Tymoczko JL, Stryer L: **Biochemistry.** In. Basingstoke: W.H. Freeman; 2012: 25-65.
5. Creighton TE: **Proteins: Structures and Molecular Properties.** In.: W. H. Freeman; 1993: 217-225.
6. **IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969).** *Biochemistry* 1970, **9**(18):3471-3479.
7. Ramachandran GN, Ramakrishnan C, Sasisekharan V: **Stereochemistry of polypeptide chain configurations.** *J Mol Biol* 1963, **7**:95-99.
8. Perczel A, Angyan JG, Kajtar M, Viviani W, Rivail JL, Marcoccia JF, Csizmadia IG: **Peptide models. 1. Topology of selected peptide conformational potential energy surfaces (glycine and alanine derivatives).** *Journal of the American Chemical Society* 1991, **113**(16):6256-6265.
9. Perczel A, McAllister MA, Császár P, Csizmadia IG: **Peptide models. IX. A complete conformational set of For-Ala-Ala-NH₂ from ab initio computations.** *Canadian Journal of Chemistry* 1994, **72**(10):2050-2070.
10. Csaszar AG, Perczel A: **Ab initio characterization of building units in peptides and proteins.** *Prog Biophys Mol Biol* 1999, **71**(2):243-309.
11. Láng A, Füziéry AK, Beke T, Hudáky P, Perczel A: **Potential energy curves, surfaces and hypersurfaces. A model to follow and understand the conformational transformations in amino acids.** *Journal of Molecular Structure: THEOCHEM* 2004, **675**(1-3):163-175.
12. Tompa P: **The interplay between structure and function in intrinsically unstructured proteins.** *FEBS Lett* 2005, **579**(15):3346-3354.
13. Tompa P: **Intrinsically unstructured proteins.** *Trends Biochem Sci* 2002, **27**(10):527-533.
14. Wright PE, Dyson HJ: **Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.** *J Mol Biol* 1999, **293**(2):321-331.
15. Dyson HJ, Wright PE: **Intrinsically unstructured proteins and their functions.** *Nat Rev Mol Cell Biol* 2005, **6**(3):197-208.
16. Dunker AK, Silman I, Uversky VN, Sussman JL: **Function and structure of inherently disordered proteins.** *Curr Opin Struct Biol* 2008, **18**(6):756-764.
17. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS: **Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder.** *Mol Cell Proteomics* 2008, **7**(7):1331-1348.
18. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK: **Intrinsic disorder in cell-signaling and cancer-associated proteins.** *J Mol Biol* 2002, **323**(3):573-584.
19. Stickley DF, Presta LG, Dill KA, Rose GD: **Hydrogen bonding in globular proteins.** *J Mol Biol* 1992, **226**(4):1143-1159.
20. Chothia C: **Structural invariants in protein folding.** *Nature* 1975, **254**(5498):304-308.
21. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22**(12):2577-2637.
22. Smith CK, Regan L: **Construction and Design of β -Sheets.** *Accounts of Chemical Research* 1997, **30**(4):153-161.

23. Fooks HM, Martin AC, Woolfson DN, Sessions RB, Hutchinson EG: **Amino acid pairing preferences in parallel beta-sheets in proteins.** *J Mol Biol* 2006, **356**(1):32-44.
24. Ho BK, Curmi PM: **Twist and shear in beta-sheets and beta-ribbons.** *J Mol Biol* 2002, **317**(2):291-308.
25. Chothia C: **Principles that Determine the Structure of Proteins.** *Annual Review of Biochemistry* 1984, **53**(1):537-572.
26. Murzin AG, Lesk AM, Chothia C: **Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures.** *J Mol Biol* 1994, **236**(5):1382-1400.
27. Murzin AG, Lesk AM, Chothia C: **Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis.** *J Mol Biol* 1994, **236**(5):1369-1381.
28. Kisker C, Schindelin H, Alber BE, Ferry JG, Rees DC: **A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon Methanosarcina thermophila.** *EMBO J* 1996, **15**(10):2323-2330.
29. Kishimoto A, Hasegawa K, Suzuki H, Taguchi H, Namba K, Yoshida M: **beta-Helix is a likely core structure of yeast prion Sup35 amyloid fibers.** *Biochem Biophys Res Commun* 2004, **315**(3):739-745.
30. Murzin AG: **Structural principles for the propeller assembly of beta-sheets: the preference for seven-fold symmetry.** *Proteins* 1992, **14**(2):191-201.
31. Richardson JS, Getzoff ED, Richardson DC: **The beta bulge: a common small unit of nonrepetitive protein structure.** *Proc Natl Acad Sci U S A* 1978, **75**(6):2574-2578.
32. Blanco FJ, Rivas G, Serrano L: **A short linear peptide that folds into a native stable [beta]-hairpin in aqueous solution.** *Nat Struct Mol Biol* 1994, **1**(9):584-590.
33. Lawrence MC, Colman PM: **Shape complementarity at protein/protein interfaces.** *J Mol Biol* 1993, **234**(4):946-950.
34. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D: **Structure of the cross-beta spine of amyloid-like fibrils.** *Nature* 2005, **435**(7043):773-778.
35. Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJ, McFarlane HT *et al*: **Atomic structures of amyloid cross-beta spines reveal varied steric zippers.** *Nature* 2007, **447**(7143):453-457.
36. Sipe JD, Cohen AS: **Review: history of the amyloid fibril.** *J Struct Biol* 2000, **130**(2-3):88-98.
37. Stefani M, Dobson CM: **Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution.** *J Mol Med (Berl)* 2003, **81**(11):678-699.
38. Aguzzi A, O'Connor T: **Protein aggregation diseases: pathogenicity and therapeutic perspectives.** *Nat Rev Drug Discov* 2010, **9**(3):237-248.
39. Pohl G, Beke T, Borbely J, Perczel A: **Prediction of folding preference of 10 kDa silk-like proteins using a Lego approach and ab initio calculations.** *J Am Chem Soc* 2006, **128**(45):14548-14559.
40. Fandrich M: **On the structural definition of amyloid fibrils and other polypeptide aggregates.** *Cell Mol Life Sci* 2007, **64**(16):2066-2078.
41. Makin OS, Sikorski P, Serpell LC: **Diffraction to study protein and peptide assemblies.** *Curr Opin Chem Biol* 2006, **10**(5):417-422.
42. Serpell LC: **Alzheimer's amyloid fibrils: structure and assembly.** *Biochim Biophys Acta* 2000, **1502**(1):16-30.
43. Chiang PK, Lam MA, Luo Y: **The many faces of amyloid beta in Alzheimer's disease.** *Curr Mol Med* 2008, **8**(6):580-584.
44. Hoppener JW, Ahren B, Lips CJ: **Islet amyloid and type 2 diabetes mellitus.** *N Engl J Med* 2000, **343**(6):411-419.
45. Roberts CJ: **Non-native protein aggregation kinetics.** *Biotechnol Bioeng* 2007, **98**(5):927-938.
46. Castillo V, Grana-Montes R, Sabate R, Ventura S: **Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes.** *Biotechnol J* 2011, **6**(6):674-685.

47. 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* 2004, **22**(10):1302-1306.
48. Trovato A, Seno F, Tosatto SC: **The PASTA server for protein aggregation prediction.** *Protein Eng Des Sel* 2007, **20**(10):521-523.
49. Tyedmers J, Mogk A, Bukau B: **Cellular strategies for controlling protein aggregation.** *Nat Rev Mol Cell Biol* 2010, **11**(11):777-788.
50. Blinov N, Dorosh L, Wishart D, Kovalenko A: **Association thermodynamics and conformational stability of beta-sheet amyloid beta(17-42) oligomers: effects of E22Q (Dutch) mutation and charge neutralization.** *Biophys J* 2010, **98**(2):282-296.
51. Ibragimova GT, Wade RC: **Stability of the beta-sheet of the WW domain: A molecular dynamics simulation study.** *Biophysical journal* 1999, **77**(4):2191-2198.
52. Jang H, Hall CK, Zhou Y: **Assembly and kinetic folding pathways of a tetrameric beta-sheet complex: molecular dynamics simulations on simplified off-lattice protein models.** *Biophysical journal* 2004, **86**(1 Pt 1):31-49.
53. Buchete N-V, Hummer G: **Structure and dynamics of parallel beta-sheets, hydrophobic core, and loops in Alzheimer's A beta fibrils.** *Biophysical journal* 2007, **92**(9):3032-3039.
54. Bellesia G, Fedorov MV, Timoshenko EG: **Molecular dynamics study of structural properties of -sheet assemblies formed by synthetic de novo oligopeptides.** *Physica A: Statistical Mechanics and its Applications* 2007, **373**:455-476.
55. Syud Fa, Stanger HE, Mortell HS, Espinosa JF, Fisk JD, Fry CG, Gellman SH: **Influence of Strand Number on Antiparallel β -Sheet Stability in Designed Three- and Four-stranded β -Sheets.** *Journal of Molecular Biology* 2003, **326**(2):553-568.
56. Stanger HE, Syud Fa, Espinosa JF, Gariat I, Muir T, Gellman SH: **Length-dependent stability and strand length limits in antiparallel beta -sheet secondary structure.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(21):12015-12020.
57. Zscherp C, Aygün H, Engels JW, Mäntele W: **Effect of proline to alanine mutation on the thermal stability of the all- β -sheet protein tendamistat.** *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 2003, **1651**(1-2):139-145.
58. Chen PY, Lin CK, Lee CT, Jan H, Chan SI: **Effects of turn residues in directing the formation of the beta-sheet and in the stability of the beta-sheet.** *Protein Sci* 2001, **10**(9):1794-1800.
59. Perczel A, Hudaky P, Palfi VK: **Dead-end street of protein folding: thermodynamic rationale of amyloid fibril formation.** *J Am Chem Soc* 2007, **129**(48):14959-14965.
60. Pohl G, Jakli I, Csizmadia IG, Papp D, Matias GF, Perczel A: **The role of entropy in initializing the aggregation of peptides: a first principle study on oligopeptide oligomerization.** *Phys Chem Chem Phys* 2012, **14**(4):1507-1516.
61. Griep S, Hobohm U: **PDBselect 1992-2009 and PDBfilter-select.** *Nucleic Acids Res* 2010, **38**(Database issue):D318-319.
62. 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* 2002, **58**(Pt 6 No 1):899-907.
63. 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* 1995, **247**(4):536-540.
64. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH--a hierarchic classification of protein domain structures.** *Structure* 1997, **5**(8):1093-1108.
65. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA: **Structural diversity of domain superfamilies in the CATH database.** *J Mol Biol* 2006, **360**(3):725-741.
66. Cuff A, Redfern OC, Greene L, Sillitoe I, Lewis T, Dibley M, Reid A, Pearl F, Dallman T, Todd A et al: **The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space.** *Structure* 2009, **17**(8):1051-1062.
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.

68. Mount DW: **Using the Basic Local Alignment Search Tool (BLAST)**. *CSH Protoc* 2007, **2007**:pdb top17.
69. Hobohm U, Scharf M, Schneider R, Sander C: **Selection of representative protein data sets**. *Protein Sci* 1992, **1**(3):409-417.
70. Vriend G: **WHAT IF: a molecular modeling and drug design program**. *J Mol Graph* 1990, **8**(1):52-56, 29.
71. Wang G, Dunbrack RL, Jr.: **PISCES: recent improvements to a PDB sequence culling server**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W94-98.
72. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.
73. Huang Y, Niu B, Gao Y, Fu L, Li W: **CD-HIT Suite: a web server for clustering and comparing biological sequences**. *Bioinformatics* 2010, **26**(5):680-682.
74. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next generation sequencing data**. *Bioinformatics* 2012.
75. Ravella D, Kumar MU, Sherlin D, Shankar M, Vaishnavi MK, Sekar K: **SMS 2.0: an updated database to study the structural plasticity of short peptide fragments in non-redundant proteins**. *Genomics Proteomics Bioinformatics* 2012, **10**(1):44-50.
76. George DG, Barker WC, Hunt LT: **The protein identification resource (PIR)**. *Nucleic Acids Res* 1986, **14**(1):11-15.
77. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2005, **33**(Database issue):D501-504.
78. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999**. *Nucleic Acids Res* 1999, **27**(1):49-54.
79. Brendel V: **PROSET—a fast procedure to create non-redundant sets of protein sequences**. *Mathematical and Computer Modelling* 1992, **16**(6–7):37-43.
80. Kallberg Y, Persson B: **KIND-a non-redundant protein database**. *Bioinformatics* 1999, **15**(3):260-261.
81. Henikoff S, Henikoff JG, Pietrokovski S: **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations**. *Bioinformatics* 1999, **15**(6):471-479.
82. Henikoff JG, Henikoff S, Pietrokovski S: **New features of the Blocks Database servers**. *Nucleic Acids Res* 1999, **27**(1):226-228.
83. Pietrokovski S, Henikoff JG, Henikoff S: **The Blocks database—a system for protein classification**. *Nucleic Acids Res* 1996, **24**(1):197-200.
84. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.
85. Levitt M: **Conformational preferences of amino acids in globular proteins**. *Biochemistry* 1978, **17**(20):4277-4285.
86. Geisow MJ, Roberts RDB: **Amino acid preferences for secondary structure vary with protein class**. *International Journal of Biological Macromolecules* 1980, **2**(6):387-389.
87. von Heijne G, Blomberg C: **The beta structure: inter-strand correlations**. *J Mol Biol* 1977, **117**(3):821-824.
88. Lifson S, Sander C: **Specific recognition in the tertiary structure of β -sheets of proteins**. *Journal of Molecular Biology* 1980, **139**(4):627-639.
89. Hutchinson EG, Sessions RB, Thornton JM, Woolfson DN: **Determinants of strand register in antiparallel beta-sheets of proteins**. *Protein Sci* 1998, **7**(11):2287-2300.
90. Coates AP, Curmi PM, Cunningham R, Donnelly C, Torda AE: **The dependence of amino acid pair correlations on structural environment**. *Proteins* 1998, **32**(2):175-189.
91. Zhu H, Braun W: **Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins**. *Protein Sci* 1999, **8**(2):326-342.

92. Mandel-Gutfreund Y, Zaremba SM, Gregoret LM: **Contributions of residue pairing to beta-sheet formation: conservation and covariation of amino acid residue pairs on antiparallel beta-strands.** *J Mol Biol* 2001, **305**(5):1145-1159.
93. Wathen B, Jia Z: **Protein beta-sheet nucleation is driven by local modular formation.** *J Biol Chem* 2010, **285**(24):18376-18384.
94. Yates F: **Contingency Tables Involving Small Numbers and the χ^2 Test.** *Supplement to the Journal of the Royal Statistical Society* 1934, **1**(2):217-235.
95. Abramowitz M, Stegun IA: **Handbook of mathematical functions with formulas, graphs, and mathematical tables.** In., vol. 55: For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.; 1964: 940.
96. Kiehna SE, Waters ML: **Sequence dependence of beta-hairpin structure: comparison of a salt bridge and an aromatic interaction.** *Protein Sci* 2003, **12**(12):2657-2667.
97. Roy RS, Gopi HN, Raghothama S, Gilardi RD, Karle IL, Balaram P: **Peptide hairpins with strand segments containing alpha- and beta-amino acid residues: cross-strand aromatic interactions of facing Phe residues.** *Biopolymers* 2005, **80**(6):787-799.
98. Santiveri CM, Rico M, Jimenez MA: **Position effect of cross-strand side-chain interactions on beta-hairpin formation.** *Protein Sci* 2000, **9**(11):2151-2160.
99. Wathen B, Jia Z: **Folding by numbers: primary sequence statistics and their use in studying protein folding.** *Int J Mol Sci* 2009, **10**(4):1567-1589.
100. Siepen JA, Radford SE, Westhead DR: **Beta edge strands in protein structure prediction and aggregation.** *Protein Sci* 2003, **12**(10):2348-2359.
101. Trovato A, Chiti F, Maritan A, Seno F: **Insight into the structure of amyloid fibrils from the analysis of globular proteins.** *PLoS Comput Biol* 2006, **2**(12):e170.
102. Lopez de la Paz M, Lacroix E, Ramirez-Alvarado M, Serrano L: **Computer-aided design of beta-sheet peptides.** *J Mol Biol* 2001, **312**(1):229-246.
103. Williams RW, Chang A, Juretic D, Loughran S: **Secondary structure predictions and medium range interactions.** *Biochim Biophys Acta* 1987, **916**(2):200-204.
104. Hadley EB, Testa OD, Woolfson DN, Gellman SH: **Preferred side-chain constellations at antiparallel coiled-coil interfaces.** *Proc Natl Acad Sci U S A* 2008, **105**(2):530-535.
105. Jiang F, Han W, Wu Y-D: **Influence of Side Chain Conformations on Local Conformational Features of Amino Acids and Implication for Force Field Development.** *The Journal of Physical Chemistry B* 2010, **114**(17):5840-5850.
106. Oldfield E: **Quantum chemical studies of protein structure.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360**(1458):1347-1361.
107. Veszprémi T, Fehér M: **A kvantumkémia alapja és alkalmazásai.** In.: Műszaki könyvkiadó; 2002: 42-54.
108. Roothaan CCI: **New Developments in Molecular Orbital Theory.** *Reviews of Modern Physics* 1951, **23**(2):69-89.
109. Hohenberg P, Kohn W: **Inhomogeneous Electron Gas.** *Physical Review* 1964, **136**(3B):B864.
110. Kohn W, Sham LJ: **Self-Consistent Equations Including Exchange and Correlation Effects.** *Physical Review* 1965, **140**(4A):A1133.
111. Lee C, Yang W, Parr RG: **Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density.** *Physical Review B* 1988, **37**(2):785.
112. Zhao Y, Schultz NE, Truhlar DG: **Design of Density Functionals by Combining the Method of Constraint Satisfaction with Parametrization for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions.** *Journal of Chemical Theory and Computation* 2006, **2**(2):364-382.
113. Hamza A, Vibók Á, Halász GJ, Mayer I: **BSSE-free SCF theories: a comment.** *Journal of Molecular Structure: THEOCHEM* 2000, **501-502**:427-434.
114. Balabin RM: **Communications: Is quantum chemical treatment of biopolymers accurate? Intramolecular basis set superposition error (BSSE).** *J Chem Phys* 2010, **132**(23):231101.

115. Balabin RM: **Enthalpy difference between conformations of normal alkanes: Intramolecular basis set superposition error (BSSE) in the case of n-butane and n-hexane.** *J Chem Phys* 2008, **129**(16):164101.
116. Kieninger M, Suhai S, Mayer I: **The chemical Hamiltonian approach in density functional theory.** *Chemical Physics Letters* 1994, **230**(6):485-490.
117. Sordo JA, Chin S, Sordo TL: **On the counterpoise correction for the basis set superposition error in large systems.** *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* 1988, **74**(2):101-110.
118. van Duijneveldt FB, van Duijneveldt-van de Rijdt JGCM, van Lenthe JH: **State of the Art in Counterpoise Theory.** *Chemical Reviews* 1994, **94**(7):1873-1885.
119. Simon S, Bertran J, Sodupe M: **Effect of Counterpoise Correction on the Geometries and Vibrational Frequencies of Hydrogen Bonded Systems.** *The Journal of Physical Chemistry A* 2001, **105**(17):4359-4364.
120. Rodgers JL, Nicewander WA: **Thirteen Ways to Look at the Correlation Coefficient.** *The American Statistician* 1988, **42**(1):59-66.
121. Berman HM: **The Protein Data Bank: a historical perspective.** *Acta Crystallogr A* 2008, **64**(Pt 1):88-95.
122. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH *et al*: **"Homology" in proteins and nucleic acids: a terminology muddle and a way out of it.** *Cell* 1987, **50**(5):667.
123. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**(3):443-453.
124. Hosseini SR, Sadeghi M, Pezeshk H, Eslahchi C, Habibi M: **PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms.** *Comput Biol Chem* 2008, **32**(6):406-411.
125. Labesse G, Colloc'h N, Pothier J, Mornon JP: **P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins.** *Comput Appl Biosci* 1997, **13**(3):291-295.
126. Frishman D, Argos P: **Knowledge-based protein secondary structure assignment.** *Proteins: Structure, Function, and Bioinformatics* 1995, **23**(4):566-579.
127. Carter P, Andersen CA, Rost B: **DSSPcont: Continuous secondary structure assignments for proteins.** *Nucleic Acids Res* 2003, **31**(13):3293-3295.
128. Efron B: **3. The Jackknife Estimate of Variance.** In: *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics; 1982: 13-19.
129. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA *et al*: **Gaussian 09, Revision B.01.** In. Wallingford CT; 2009.
130. Atkins PW: **Physical Chemistry.** In., Second Edition edn. Oxford: Oxford University Press; 1982: 665-669.
131. Anfinsen CB: **Principles that govern the folding of protein chains.** *Science* 1973, **181**(4096):223-230.
132. **IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970.** *Biochem J* 1970, **120**(3):449-454.

Mellékletek

1. melléklet: aminosavak gyakorisága az NMR-es szerkezetekben, cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Összes β -redő	1992	1079	1051	2311	2403	1243	828	3969	2269	4108	792	842	628	1335	2080	1809	2748	5659	754	1994
AP10	1038	464	447	1040	1333	783	418	1957	1067	1985	375	341	0	661	1063	802	1107	2837	365	1027
AP14	1106	720	560	1340	1345	536	443	2193	1256	2245	453	486	580	751	1177	1057	1742	3061	488	1163
PARL	291	102	163	267	330	126	113	776	279	803	107	128	49	131	207	204	368	1078	72	273

2. melléklet: az NMR-es szerkezeteken végzett párkorrelációs analízis eredményeinek összefoglalása, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	ALL	AP10	AP14	PARL
Párok száma	24455	9555	11655	3245
	45782	18008	21738	6013
	0,0174	0,0223	0,0211	0,0580

3. melléklet: aminosav párok megfigyelt előfordulása NMR-es szerkezeteken, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	59	53	40	106	167	68	63	286	101	329	55	19	33	54	92	96	136	492	41	149
C	53	119	21	35	95	87	20	118	43	119	20	19	22	25	34	34	57	166	50	66
D	40	21	22	46	54	43	45	80	125	66	20	42	16	57	144	61	123	101	14	44
E	106	35	46	59	112	44	71	199	352	174	47	78	28	116	337	154	249	261	35	128
F	167	95	54	112	135	131	49	347	120	351	72	51	56	90	132	105	120	498	57	216
G	68	87	43	44	131	30	26	133	47	152	25	28	20	33	46	52	69	304	34	81
H	63	20	45	71	49	26	14	92	63	67	17	27	12	37	53	60	85	106	17	59
I	286	118	80	199	347	133	92	427	206	735	111	38	40	115	162	115	252	915	83	246
K	101	43	125	352	120	47	63	206	56	192	45	62	29	118	83	152	292	285	60	155
L	329	119	66	174	351	152	67	735	192	440	129	68	57	102	203	150	176	905	113	267
M	55	20	20	47	72	25	17	111	45	129	14	17	18	24	47	31	41	144	19	45
N	19	19	42	78	51	28	27	38	62	68	17	29	15	43	60	72	122	88	29	38
P	33	22	16	28	56	20	12	40	29	57	18	15	3	13	33	27	40	66	43	55
Q	54	25	57	116	90	33	37	115	118	102	24	43	13	40	109	108	182	137	40	77
R	92	34	144	337	132	46	53	162	83	203	47	60	33	109	53	126	182	293	67	166
S	96	34	61	154	105	52	60	115	152	150	31	72	27	108	126	89	230	215	35	93
T	136	57	123	249	120	69	85	252	292	176	41	122	40	182	182	230	213	346	25	108
V	492	166	101	261	498	304	106	915	285	905	144	88	66	137	293	215	346	712	97	380
W	41	50	14	35	57	34	17	83	60	113	19	29	43	40	67	35	25	97	17	50
Y	149	66	44	128	216	81	59	246	155	267	45	38	55	77	166	93	108	380	50	50

4. melléklet: aminosav párok várható előfordulása NMR-es szerkezeteken, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	99,47	53,88	52,48	115,39	119,99	62,07	41,34	198,18	113,3	205,12	39,55	42,04	31,36	66,66	103,86	90,33	137,21	282,57	37,65	99,57
C	53,88	29,18	28,43	62,5	64,99	33,62	22,39	107,35	61,37	111,11	21,42	22,77	16,99	36,11	56,26	48,93	74,32	153,06	20,39	53,93
D	52,48	28,43	27,69	60,88	63,31	32,75	21,81	104,56	59,78	108,22	20,87	22,18	16,54	35,17	54,8	47,66	72,4	149,09	19,86	52,53
E	115,39	62,5	60,88	133,87	139,2	72,01	47,96	229,92	131,44	237,97	45,88	48,78	36,38	77,33	120,49	104,79	159,19	327,82	43,68	115,51
F	119,99	64,99	63,31	139,2	144,74	74,87	49,87	239,07	136,67	247,44	47,71	50,72	37,83	80,41	125,29	108,96	165,52	340,87	45,42	120,11
G	62,07	33,62	32,75	72,01	74,87	38,73	25,8	123,66	70,7	128	24,68	26,23	19,57	41,6	64,81	56,36	85,62	176,32	23,49	62,13
H	41,34	22,39	21,81	47,96	49,87	25,8	17,19	82,38	47,09	85,26	16,44	17,48	13,03	27,71	43,17	37,55	57,03	117,45	15,65	41,39
I	198,18	107,35	104,56	229,92	239,07	123,66	82,38	394,87	225,74	408,7	78,8	83,77	62,48	132,82	206,94	179,97	273,39	563,01	75,01	198,38
K	113,3	61,37	59,78	131,44	136,67	70,7	47,09	225,74	129,05	233,65	45,05	47,89	35,72	75,93	118,3	102,89	156,29	321,86	42,88	113,41
L	205,12	111,11	108,22	237,97	247,44	128	85,26	408,7	233,65	423,01	81,55	86,7	64,67	137,47	214,18	186,28	282,97	582,72	77,64	205,33
M	39,55	21,42	20,87	45,88	47,71	24,68	16,44	78,8	45,05	81,55	15,72	16,72	12,47	26,5	41,29	35,91	54,55	112,35	14,97	39,59
N	42,04	22,77	22,18	48,78	50,72	26,23	17,48	83,77	47,89	86,7	16,72	17,77	13,25	28,18	43,9	38,18	58	119,44	15,91	42,09
P	31,36	16,99	16,54	36,38	37,83	19,57	13,03	62,48	35,72	64,67	12,47	13,25	9,89	21,02	32,74	28,48	43,26	89,08	11,87	31,39
Q	66,66	36,11	35,17	77,33	80,41	41,6	27,71	132,82	75,93	137,47	26,5	28,18	21,02	44,67	69,6	60,54	91,96	189,37	25,23	66,73
R	103,86	56,26	54,8	120,49	125,29	64,81	43,17	206,94	118,3	214,18	41,29	43,9	32,74	69,6	108,45	94,32	143,28	295,05	39,31	103,96
S	90,33	48,93	47,66	104,79	108,96	56,36	37,55	179,97	102,89	186,28	35,91	38,18	28,48	60,54	94,32	82,03	124,61	256,61	34,19	90,42
T	137,21	74,32	72,4	159,19	165,52	85,62	57,03	273,39	156,29	282,97	54,55	58	43,26	91,96	143,28	124,61	189,29	389,81	51,94	137,35
V	282,57	153,06	149,09	327,82	340,87	176,32	117,45	563,01	321,86	582,72	112,35	119,44	89,08	189,37	295,05	256,61	389,81	802,73	106,96	282,85
W	37,65	20,39	19,86	43,68	45,42	23,49	15,65	75,01	42,88	77,64	14,97	15,91	11,87	25,23	39,31	34,19	51,94	106,96	14,25	37,69
Y	99,57	53,93	52,53	115,51	120,11	62,13	41,39	198,38	113,41	205,33	39,59	42,09	31,39	66,73	103,96	90,42	137,35	282,85	37,69	99,67

5. melléklet: aminosav párok párpreferencia értékei NMR-es szerkezeteken, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0,59	0,98	0,76	0,92	1,39	1,10	1,52	1,44	0,89	1,60	1,39	0,45	1,05	0,81	0,89	1,06	0,99	1,74	1,09	1,50
C	0,98	4,08	0,74	0,56	1,46	2,59	0,89	1,10	0,70	1,07	0,93	0,83	1,30	0,69	0,60	0,69	0,77	1,08	2,45	1,22
D	0,76	0,74	0,79	0,76	0,85	1,31	2,06	0,77	2,09	0,61	0,96	1,89	0,97	1,62	2,63	1,28	1,70	0,68	0,70	0,84
E	0,92	0,56	0,76	0,44	0,80	0,61	1,48	0,87	2,68	0,73	1,02	1,60	0,77	1,50	2,80	1,47	1,56	0,80	0,80	1,11
F	1,39	1,46	0,85	0,80	0,93	1,75	0,98	1,45	0,88	1,42	1,51	1,01	1,48	1,12	1,05	0,96	0,72	1,46	1,26	1,80
G	1,10	2,59	1,31	0,61	1,75	0,77	1,01	1,08	0,66	1,19	1,01	1,07	1,02	0,79	0,71	0,92	0,81	1,72	1,45	1,30
H	1,52	0,89	2,06	1,48	0,98	1,01	0,81	1,12	1,34	0,79	1,03	1,55	0,92	1,34	1,23	1,60	1,49	0,90	1,09	1,43
I	1,44	1,10	0,77	0,87	1,45	1,08	1,12	1,08	0,91	1,80	1,41	0,45	0,64	0,87	0,78	0,64	0,92	1,63	1,11	1,24
K	0,89	0,70	2,09	2,68	0,88	0,66	1,34	0,91	0,43	0,82	1,00	1,29	0,81	1,55	0,70	1,48	1,87	0,89	1,40	1,37
L	1,60	1,07	0,61	0,73	1,42	1,19	0,79	1,80	0,82	1,04	1,58	0,78	0,88	0,74	0,95	0,81	0,62	1,55	1,46	1,30
M	1,39	0,93	0,96	1,02	1,51	1,01	1,03	1,41	1,00	1,58	0,89	1,02	1,44	0,91	1,14	0,86	0,75	1,28	1,27	1,14
N	0,45	0,83	1,89	1,60	1,01	1,07	1,55	0,45	1,29	0,78	1,02	1,63	1,13	1,53	1,37	1,89	2,10	0,74	1,82	0,90
P	1,05	1,30	0,97	0,77	1,48	1,02	0,92	0,64	0,81	0,88	1,44	1,13	0,30	0,62	1,01	0,95	0,92	0,74	3,62	1,75
Q	0,81	0,69	1,62	1,50	1,12	0,79	1,34	0,87	1,55	0,74	0,91	1,53	0,62	0,90	1,57	1,78	1,98	0,72	1,59	1,15
R	0,89	0,60	2,63	2,80	1,05	0,71	1,23	0,78	0,70	0,95	1,14	1,37	1,01	1,57	0,49	1,34	1,27	0,99	1,70	1,60
S	1,06	0,69	1,28	1,47	0,96	0,92	1,60	0,64	1,48	0,81	0,86	1,89	0,95	1,78	1,34	1,08	1,85	0,84	1,02	1,03
T	0,99	0,77	1,70	1,56	0,72	0,81	1,49	0,92	1,87	0,62	0,75	2,10	0,92	1,98	1,27	1,85	1,13	0,89	0,48	0,79
V	1,74	1,08	0,68	0,80	1,46	1,72	0,90	1,63	0,89	1,55	1,28	0,74	0,74	0,72	0,99	0,84	0,89	0,89	0,91	1,34
W	1,09	2,45	0,70	0,80	1,26	1,45	1,09	1,11	1,40	1,46	1,27	1,82	3,62	1,59	1,70	1,02	0,48	0,91	1,19	1,33
Y	1,50	1,22	0,84	1,11	1,80	1,30	1,43	1,24	1,37	1,30	1,14	0,90	1,75	1,15	1,60	1,03	0,79	1,34	1,33	0,50

6. melléklet: aminosav párok χ^2 -értékei NMR-es szerkezeteken, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	46002	46056	46114	45781	45527	45984	46031	45002	45804	44863	46062	46219	46163	46042	45848	45847	45641	44256	46117	45623
C	46056	46090	46215	46135	45885	46016	46222	45749	46098	45743	46223	46227	46225	46193	46142	46146	46029	45526	46134	46004
D	46114	46215	46211	46085	46045	46122	46141	45908	45838	45971	46224	46151	46247	46074	45832	46039	45806	45794	46250	46096
E	45781	46135	46085	45998	45739	46087	45998	45308	45173	45414	46089	45981	46182	45801	45238	45634	45242	44989	46141	45689
F	45527	45885	46045	45739	45626	45745	46074	44727	45705	44705	45988	46067	46075	45879	45662	45792	45691	44092	46044	45366
G	45984	46016	46122	46087	45745	46167	46193	45667	46073	45592	46198	46186	46227	46152	46080	46060	45967	45051	46164	45933
H	46031	46222	46141	45998	46074	46193	46252	45871	46026	45974	46239	46202	46265	46149	46067	46056	45938	45787	46239	46044
I	45002	45749	45908	45308	44727	45667	45871	44245	45282	43328	45807	46115	46114	45735	45480	45712	45061	42524	45912	45136
K	45804	46098	45838	45173	45705	46073	46026	45282	46014	45338	46098	46033	46178	45798	45885	45644	45148	44894	46041	45592
L	44863	45743	45971	45414	44705	45592	45974	43328	45338	44179	45747	45972	46041	45790	45303	45553	45388	42518	45800	45055
M	46062	46223	46224	46089	45988	46198	46239	45807	46098	45747	46252	46240	46244	46202	46092	46165	46109	45649	46232	46100
N	46219	46227	46151	45981	46067	46186	46202	46115	46033	45972	46240	46197	46254	46131	46045	46029	45856	45867	46198	46129
P	46163	46225	46247	46182	46075	46227	46265	46114	46178	46041	46244	46254	46308	46256	46164	46193	46128	45988	46219	46093
Q	46042	46193	46074	45801	45879	46152	46149	45735	45798	45790	46202	46131	46256	46120	45836	45870	45603	45610	46141	45944
R	45848	46142	45832	45238	45662	46080	46067	45480	45885	45303	46092	46045	46164	45836	46032	45738	45470	44881	46024	45569
S	45847	46146	46039	45634	45792	46060	46056	45712	45644	45553	46165	46029	46193	45870	45738	45892	45398	45237	46149	45858
T	45641	46029	45806	45242	45691	45967	45938	45061	45148	45388	46109	45856	46128	45603	45470	45398	45309	44602	46187	45759
V	44256	45526	45794	44989	44092	45051	45787	42524	44894	42518	45649	45867	45988	45610	44881	45237	44602	42820	45830	44563
W	46117	46134	46250	46141	46044	46164	46239	45912	46041	45800	46232	46198	46219	46141	46024	46149	46187	45830	46240	46081
Y	45623	46004	46096	45689	45366	45933	46044	45136	45592	45055	46100	46129	46093	45944	45569	45858	45759	44563	46081	46047

7. melléklet: aminosav párok párpreferencia értékeinek átlagos eltérése a Jackknife teszt eredményeképpen az eredeti értékektől. Jackknife tesztek száma: 5, elhagyott adatmennyiség: 10%. NMR-es szerkezeteken számolva, adatkészlet cd-hit programmal szűrve, 40%-os szekvenciális azonossági küszöbvel

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0,03	0,03	0,03	0,02	0,05	0,07	0,12	0,04	0,03	0,03	0,04	0,04	0,06	0,05	0,05	0,05	0,02	0,02	0,06	0,04
C	0,03	0,16	0,06	0,03	0,04	0,06	0,10	0,03	0,04	0,03	0,06	0,06	0,17	0,08	0,04	0,03	0,04	0,03	0,04	0,06
D	0,03	0,06	0,08	0,03	0,03	0,03	0,07	0,04	0,06	0,02	0,09	0,11	0,06	0,09	0,05	0,05	0,08	0,05	0,07	0,05
E	0,02	0,03	0,03	0,01	0,03	0,04	0,07	0,02	0,04	0,02	0,05	0,04	0,03	0,03	0,09	0,05	0,04	0,02	0,06	0,03
F	0,05	0,04	0,03	0,03	0,02	0,06	0,04	0,04	0,02	0,04	0,06	0,05	0,07	0,02	0,04	0,03	0,03	0,01	0,04	0,03
G	0,07	0,06	0,03	0,04	0,06	0,05	0,05	0,03	0,04	0,03	0,06	0,01	0,05	0,09	0,06	0,02	0,03	0,02	0,07	0,06
H	0,12	0,10	0,07	0,07	0,04	0,05	0,12	0,04	0,10	0,02	0,06	0,19	0,11	0,04	0,06	0,05	0,06	0,04	0,09	0,09
I	0,04	0,03	0,04	0,02	0,04	0,03	0,04	0,02	0,01	0,01	0,04	0,03	0,04	0,02	0,02	0,02	0,01	0,03	0,05	0,02
K	0,03	0,04	0,06	0,04	0,02	0,04	0,10	0,01	0,02	0,02	0,06	0,05	0,06	0,03	0,03	0,05	0,03	0,02	0,09	0,02
L	0,03	0,03	0,02	0,02	0,04	0,03	0,02	0,01	0,02	0,01	0,03	0,05	0,04	0,03	0,01	0,03	0,01	0,02	0,03	0,03
M	0,04	0,06	0,09	0,05	0,06	0,06	0,06	0,04	0,06	0,03	0,09	0,10	0,05	0,08	0,07	0,05	0,06	0,03	0,14	0,06
N	0,04	0,06	0,11	0,04	0,05	0,01	0,19	0,03	0,05	0,05	0,10	0,09	0,17	0,10	0,08	0,07	0,07	0,02	0,09	0,04
P	0,06	0,17	0,06	0,03	0,07	0,05	0,11	0,04	0,06	0,04	0,05	0,17	0,01	0,06	0,04	0,07	0,07	0,04	0,16	0,10
Q	0,05	0,08	0,09	0,03	0,02	0,09	0,04	0,02	0,03	0,03	0,08	0,10	0,06	0,05	0,05	0,04	0,03	0,02	0,09	0,02
R	0,05	0,04	0,05	0,09	0,04	0,06	0,06	0,02	0,03	0,01	0,07	0,08	0,04	0,05	0,01	0,04	0,03	0,02	0,05	0,03
S	0,05	0,03	0,05	0,05	0,03	0,02	0,05	0,02	0,05	0,03	0,05	0,07	0,07	0,04	0,04	0,08	0,03	0,01	0,07	0,03
T	0,02	0,04	0,08	0,04	0,03	0,03	0,06	0,01	0,03	0,01	0,06	0,07	0,07	0,03	0,03	0,03	0,03	0,01	0,04	0,03
V	0,02	0,03	0,05	0,02	0,01	0,02	0,04	0,03	0,02	0,02	0,03	0,02	0,04	0,02	0,02	0,01	0,01	0,02	0,05	0,03
W	0,06	0,04	0,07	0,06	0,04	0,07	0,09	0,05	0,09	0,03	0,14	0,09	0,16	0,09	0,05	0,07	0,04	0,05	0,06	0,06
Y	0,04	0,06	0,05	0,03	0,03	0,06	0,09	0,02	0,02	0,03	0,06	0,04	0,10	0,02	0,03	0,03	0,03	0,03	0,06	0,03

8. melléklet: NMR-es szerkezetekből különböző szűrési algoritmussal és szekvenciális azonossági küszöbvel megszürt adatkészletek hasonlósága: az egyes párkorrelációs táblázatok között számolt Pearson-féle korrelációs együtthatók értéke.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
PDBselect, 25%	1.	1	0,93	0,92	0,92	0,91	0,78	0,81	0,82	0,83	0,83	0,83	0,82	0,83	0,83	0,83	0,84
PDBselect, 30%	2.	0,93	1	0,98	0,97	0,97	0,86	0,88	0,89	0,90	0,91	0,91	0,90	0,91	0,91	0,91	0,88
PDBselect, 35%	3.	0,92	0,98	1	0,99	0,99	0,87	0,89	0,90	0,92	0,93	0,93	0,92	0,93	0,93	0,93	0,90
PDBselect, 40%	4.	0,92	0,97	0,99	1	1	0,87	0,89	0,90	0,92	0,93	0,93	0,92	0,94	0,94	0,94	0,91
PDBselect, 50%	5.	0,91	0,97	0,99	1	1	0,87	0,89	0,90	0,92	0,93	0,93	0,92	0,94	0,94	0,94	0,92
PISCES 25%	6.	0,78	0,86	0,87	0,87	0,87	1	0,97	0,96	0,95	0,94	0,94	0,93	0,93	0,92	0,91	0,78
PISCES 30%	7.	0,81	0,88	0,89	0,89	0,89	0,97	1	0,98	0,97	0,96	0,96	0,95	0,95	0,94	0,93	0,82
PISCES 35%	8.	0,82	0,89	0,90	0,90	0,9	0,96	0,98	1	0,98	0,98	0,97	0,97	0,97	0,96	0,95	0,84
PISCES 40%	9.	0,83	0,90	0,92	0,92	0,92	0,95	0,97	0,98	1	0,99	0,99	0,98	0,98	0,97	0,96	0,86
PISCES 45%	10.	0,83	0,91	0,93	0,93	0,93	0,94	0,96	0,98	0,99	1	0,99	0,98	0,99	0,98	0,97	0,87
PISCES 50%	11.	0,83	0,91	0,93	0,93	0,93	0,94	0,96	0,97	0,99	0,99	1	0,98	0,99	0,98	0,97	0,87
cd-hit 40%	12.	0,82	0,9	0,92	0,92	0,92	0,93	0,95	0,97	0,98	0,98	0,98	1	0,98	0,97	0,96	0,85
cd-hit 50%	13.	0,83	0,91	0,93	0,94	0,94	0,93	0,95	0,97	0,98	0,99	0,99	0,98	1	0,99	0,98	0,88
cd-hit 60%	14.	0,83	0,91	0,93	0,94	0,94	0,92	0,94	0,96	0,97	0,98	0,98	0,97	0,99	1	1	0,91
cd-hit 70%	15.	0,83	0,91	0,93	0,94	0,94	0,91	0,93	0,95	0,96	0,97	0,97	0,96	0,98	1	1	0,92
teljes lista	16.	0,84	0,88	0,90	0,91	0,92	0,78	0,82	0,84	0,86	0,87	0,87	0,85	0,88	0,91	0,92	1

9. melléklet: NMR-es szerkezetekből különböző szűrési algoritmussal és szekvenciális azonossági küszöbvel megszűrt adatkészletek hasonlósága: az egyes listák közötti átfedés, százalékosan kifejezve. 100% a teljesen azonos listát jelent, így értelemszerűen a diagonálisban csupa 100% van feltüntetve.

		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
PDBselect, 25%	1.	100	68	58	55	53	51	50	49	48	46	45	48	45	42	40	25
PDBselect, 30%	2.	68	100	81	77	73	55	62	67	66	65	63	68	64	59	57	35
PDBselect, 35%	3.	58	81	100	94	89	52	58	66	73	77	76	67	77	71	68	41
PDBselect, 40%	4.	55	77	94	100	95	51	57	64	71	77	79	66	81	75	72	43
PDBselect, 50%	5.	53	73	89	95	100	49	55	62	69	74	79	64	80	79	76	45
PISCES 25%	6.	51	55	52	51	49	100	88	77	71	66	63	64	59	54	51	28
PISCES 30%	7.	50	62	58	57	55	88	100	85	78	73	69	72	65	60	56	31
PISCES 35%	8.	49	67	66	64	62	77	85	100	88	82	78	80	74	68	64	36
PISCES 40%	9.	48	66	73	71	69	71	78	88	100	91	87	81	83	75	71	39
PISCES 45%	10.	46	65	77	77	74	66	73	82	91	100	94	78	90	81	76	42
PISCES 50%	11.	45	63	76	79	79	63	69	78	87	94	100	75	95	86	81	44
cd-hit 40%	12.	48	68	67	66	64	64	72	80	81	78	75	100	77	68	64	35
cd-hit 50%	13.	45	64	77	81	80	59	65	74	83	90	95	77	100	87	82	44
cd-hit 60%	14.	42	59	71	75	79	54	60	68	75	81	86	68	87	100	93	51
cd-hit 70%	15.	40	57	68	72	76	51	56	64	71	76	81	64	82	93	100	54
teljes lista	16.	25	35	41	43	45	28	31	36	39	42	44	35	44	51	54	100

10. melléklet: NMR-es adatbázisok jóságának összessége. A táblázatban szerepel a láncok száma ($N_{\text{láncok}}$), az átlagos Pearson-féle korrelációs koefficiense (CORR) a többi listához képest (a párpreferencia értékekre számolva), az átlagos χ^2 -érték (szint a párpreferencia értékekre számolva), a láncok számával normált átlagos χ^2 -érték, valamint a párpreferencia értékek átlagos ingadozása a jackknife-teszt eredményeképpen (JNF).

			ALL				AP10		AP14		PARL	
		$N_{\text{láncok}}$			—							
PDBSelect	25%	2270	0,858	27881	12,28	0,0228	10828	0,0298	13093	0,0282	3944	0,0701
	30%	3122	0,919	40345	12,92	0,0176	15847	0,0232	19207	0,0230	5270	0,0594
	35%	3718	0,933	48554	13,06	0,0157	19180	0,0242	23408	0,0186	5942	0,0433
	40%	3928	0,936	50889	12,96	0,0177	20079	0,0215	24580	0,0203	6205	0,0577
	50%	4116	0,936	53130	12,91	0,0168	21038	0,0203	25670	0,0186	6396	0,0499
PISCES	25%	2560	0,905	31785	12,42	0,0205	12434	0,0286	14911	0,0210	4423	0,0588
	30%	2839	0,924	36052	12,70	0,0190	14130	0,0276	16929	0,0257	4975	0,0565
	35%	3221	0,936	41738	12,96	0,0180	16337	0,0249	19778	0,0206	5601	0,0530
	40%	3527	0,945	46105	13,07	0,0178	18130	0,0227	21941	0,0215	6010	0,0548
	45%	3806	0,949	49721	13,06	0,0176	19721	0,0226	23789	0,0207	6186	0,0535
	50%	4017	0,948	52124	12,98	0,0177	20705	0,0207	25088	0,0207	6305	0,0514
cd-hit	40%	3138	0,939	45782	14,59	0,0174	18008	0,0223	21738	0,0211	6013	0,0580
	50%	4001	0,949	53436	13,36	0,0176	21212	0,0202	25727	0,0212	6469	0,0511
	60%	4570	0,948	57881	12,67	0,0170	22974	0,0206	27967	0,0188	6912	0,0514
	70%	4885	0,944	61006	12,49	0,0156	24285	0,0200	29545	0,0188	7146	0,0505
teljes lista	-	9047	0,878	93551	10,34	0,0126	36656	0,0149	44833	0,0145	12023	0,0407

11. melléklet: röntgenszerkezetekből álló listák jósági becslése. A sorok és oszlopok a listában lévő fehérjeszerkezetek maximális felbontását és R-faktorát jelzik. A táblázatok celláiban feltüntettük a láncok száma szerint normált χ^2 -értékeket a redundáns és nemredundáns listákból származó párpreferencia értékekre számolva, valamint a nemredundáns adatkészlethez tartozó párpreferencia értékeinek az átlagos jackknife-ingadozását.

—— (redundáns adatkészlet)						
Max. R-faktor	Maximális felbontás (Å)					
	1,0	1,2	1,4	1,6	1,8	2,0
0,10	41,35	64,67	61,95	62,11	61,95	62,04
0,13	44,61	75,95	79,41	80,23	79,59	79,55
0,16	47,56	88,75	90,37	90,81	90,67	90,79
0,19	50,22	91,99	100,87	104,87	104,21	104,03
0,22	50,22	92,91	109,28	115,59	115,13	114,94
0,25	46,65	94,66	118,54	127,08	127,18	127,28

—— (nemredundáns adatkészlet)						
Max. R-faktor	Maximális felbontás (Å)					
	1,0	1,2	1,4	1,6	1,8	2,0
0,10	57,55	66,27	64,66	64,02	62,98	62,98
0,13	66,37	64,59	75,99	79,47	77,7	77,72
0,16	66,37	75,75	88,34	88,5	87,16	86,93
0,19	66,37	79,91	104,01	106,21	103,87	103,4
0,22	66,37	81,69	117,47	123,02	118,56	117,62
0,25	68,95	87,05	131,48	139,99	135,5	134,92

(nemredundáns adatkészlet)						
Max. R-faktor	Maximális felbontás (Å)					
	1	1.2	1.4	1.6	1.8	2
0,10	0,1379	0,1902	0,2086	0,1411	0,167	0,0754
0,13	0,0856	0,0636	0,0402	0,0394	0,0445	0,0364
0,16	0,0879	0,0383	0,0248	0,0213	0,0158	0,0138
0,19	0,0839	0,0383	0,0194	0,0141	0,0109	0,0081
0,22	0,078	0,0484	0,0216	0,0111	0,0094	0,0065
0,25	0,0511	0,0294	0,023	0,0155	0,0087	0,0065

12. melléklet: Val-Gly és Val-Ala párok lehetséges oldallánc-orientációinak eloszlása százalékosan. Felül a kvantumkémiai modellezésből, alul a statisztikai analízisből származó értékek láthatók. SumSq az egyes kvantumkémiaiával számolt értékek eltéréseinek négyzetösszegét jelenti az átlagos statisztikai értékhez képest.

AP10_AVA_AGA	Vg+ (%)	Vg- (%)	Vt (%)	SumSq
HF/3-21G, ΔE	65	33	2	6429
B3LYP/6-31G*, ΔE	60	32	8	5222
B3LYP/6-31G*, ΔG	39	28	34	1532
M052X/6-31G*, ΔE	66	32	1	6635
M052X/6-31G*, ΔG	57	36	7	5095
B3LYP/6-31G*, cp, ΔE	56	30	14	4296
B3LYP/6-31G*, cp, ΔG	46	26	28	2316
M052X/6-31G*, cp, ΔE	63	34	3	6019
M052X/6-31G*, cp, ΔG	43	41	17	3073
B3LYP/6-31G*, $\Delta_{stab}E$	54	23	23	3424
B3LYP/6-31G*, $\Delta_{stab}G$	34	20	47	862
M052X/6-31G*, $\Delta_{stab}E$	69	23	8	6177
M052X/6-31G*, $\Delta_{stab}G$	35	62	3	4923
B3LYP/6-31G*, cp, $\Delta_{stab}E$	46	19	35	2019
B3LYP/6-31G*, cp, $\Delta_{stab}G$	41	19	40	1494
PISCES, felb1.8, r0.25	11	33	57	
WHATIF, felb 1.8, r0.18	9	29	63	
WHATIF, felb 1.8, r0.19	8	29	63	
PDBSelect, felb 1.5, r0.18	11	29	60	
PDBSelect, felb 1.75, r0.18	11	31	58	
Átlag	10	30	60	
Szórás	1,39	1,69	2,69	

AP10_AVA_AAA	Vg+ (%)	Vg- (%)	Vt (%)	SumSq
HF/3-21G, ΔE	60	30	10	5445
B3LYP/6-31G*, ΔE	58	33	9	5389
B3LYP/6-31G*, ΔG	34	25	41	1052
M052X/6-31G*, ΔE	58	39	3	6254
M052X/6-31G*, ΔG	35	44	21	2828
B3LYP/6-31G*, cp, ΔE	54	30	15	4350
B3LYP/6-31G*, cp, ΔG	31	25	44	851
M052X/6-31G*, cp, ΔE	61	37	3	6438
M052X/6-31G*, cp, ΔG	73	21	6	7352
B3LYP/6-31G*, $\Delta_{stab}E$	52	23	25	3211
B3LYP/6-31G*, $\Delta_{stab}G$	28	17	55	450
M052X/6-31G*, $\Delta_{stab}E$	57	27	16	4535
M052X/6-31G*, $\Delta_{stab}G$	20	72	8	5459
B3LYP/6-31G*, cp, $\Delta_{stab}E$	43	19	38	1765
B3LYP/6-31G*, cp, $\Delta_{stab}G$	26	17	57	336
PISCES, felb1.8, r0.25	12	24	64	
WHATIF, felb 1.8, r0.18	12	25	63	
WHATIF, felb 1.8, r0.19	11	23	66	
PDBSelect, felb 1.5, r0.18	9	26	65	
PDBSelect, felb 1.75, r0.18	10	25	64	
Átlag	11	25	65	
Szórás	1,07	1,17	1,12	

13. melléklet: Val-Ser párok lehetséges oldallánc-orientációinak eloszlása százalékosan. Felül a kvantumkémiail modellezésből, alul a statisztikai analízisből származó értékek láthatók. SumSq az egyes kvantumkémiával számolt értékek eltéréseinek négyzetösszegét jelenti az átlagos statisztikai értékhez képest.

AP10, AVA_ASA	Vg+Sg+ (%)	Vg+Sg- (%)	Vg+St (%)	Vg-Sg+ (%)	Vg-Sg- (%)	Vg-St (%)	VtSg+ (%)	VtSg- (%)	VtSt (%)	SumSq
HF/3-21G, ΔE	3	0	73	1	0	18	0	0	4	6639
B3LYP/6-31G*, ΔE	0	0	41	0	0	49	0	0	11	4401
B3LYP/6-31G*, ΔG	0	0	30	0	0	40	0	0	30	3051
M052X/6-31G*, ΔE	0	0	88	0	0	7	0	0	5	8924
M052X/6-31G*, ΔG	0	0	86	0	0	9	0	0	5	8550
B3LYP/6-31G*, cp, ΔE	0	0	0	0	0	70	0	0	30	5027
B3LYP/6-31G*, cp, ΔG	0	0	0	0	0	45	0	0	55	3651
M052X/6-31G*, cp, ΔE	27	17	12	10	7	1	8	6	11	1801
M052X/6-31G*, cp, ΔG	4	12	2	26	49	1	3	1	2	3939
B3LYP/6-31G*, $\Delta_{stab}E$	21	7	12	7	3	11	25	4	10	976
B3LYP/6-31G*, $\Delta_{stab}G$	4	3	17	2	2	18	20	7	26	588
M052X/6-31G*, $\Delta_{stab}E$	38	11	14	15	4	1	12	2	4	2664
M052X/6-31G*, $\Delta_{stab}G$	14	9	7	55	11	2	1	0	0	4766
B3LYP/6-31G*, cp, $\Delta_{stab}E$	12	14	0	4	5	14	16	12	24	450
B3LYP/6-31G*, cp, $\Delta_{stab}G$	3	5	0	2	3	20	8	9	50	1452
B3LYP/6-31+G**, $\Delta_{stab}E$	9	9	12	4	5	15	16	9	21	473
B3LYP/6-31+G**, $\Delta_{stab}G$	2	3	21	1	1	20	6	5	41	1590
PISCES, felb1.8, r0.25	1	5	3	3	7	10	28	21	23	
PDBSelect, felb 1.5, r0.18	3	6	3	1	6	9	26	28	18	
PDBSelect, felb 1.75, r0.18	3	6	3	1	6	10	28	22	22	
WHATIF, felb 1.8, r0.18	2	4	4	1	8	9	29	19	24	
WHATIF, felb 1.8, r0.19	2	3	3	2	7	10	27	20	25	
átlag	2	5	3	2	7	10	28	22	22	
szórás	0,95	1,14	0,54	0,94	0,71	0,21	1,11	3,30	2,69	