

Lázár Tamás

Rendezetlen fehérjék komplexeinek bioinformatikai vizsgálata

XXXII. OTDK Biológia Szekció

2013-2015

Molekuláris bionika mérnöki BSc



Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

Belső Konzulens: Gáspári Zoltán Ph.D.



*Magyar Tudományos Akadémia Természettudományi Kutatóközpont
Enzimológiai Intézet*

Témavezető: Schád Éva Ph.D.

Nyilatkozat és 3. Számú melléklet - Tételes szerzői hozzájárulás

Alulírott Lázár Tamás, a Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Karának hallgatója kijelentem, hogy ezt a dolgozatot meg nem engedett segítség nélkül, saját magam készítettem, és a dolgozatban csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen a forrás megadásával megjelöltem. Az egyes folyamatok tervezésében és eredmények értékelésében témavezetőm és munkacsoportunk tagjai voltak segítségemre.

3. számú melléklet - Tételes szerzői hozzájárulás

Alulírott Lázár Tamás a dolgozatban leírt munkafolyamatok közül:

- a szerkezetbiológiai irodalmi áttekintést,
 - az adatbázist,
 - az adatbázis automatizálási lehetőségének vizsgálatát,
 - az adatok K_D , interakciós profil és kötőinterfész és eloszlási hisztogramjait,
 - az összes bioinformatikai elemzést,
 - a Perl szkriptek programozását
- végeztem el.

Budapest, 2014. 05. 14.

.....

Aláírás

Tartalomjegyzék

NYILATKOZAT ÉS 3. SZÁMÚ MELLÉKLET - TÉTELES SZERZŐI HOZZÁJÁRULÁS.....	1
1. BEVEZETÉS.....	3
1.1. ABSZTRAKT	3
1.2. RENDEZETLEN FEHÉRJÉK	3
1.3. RENDEZETLEN FEHÉRJÉK KÖTŐDÉSE, KOMPLEXEK KÉPZŐDÉSE	4
1.5. FEHÉRJEKOMPLEX HOZZÁFÉRHETŐ FELSZÍNE.....	11
2. CÉLKITŰZÉSEK	12
3. MÓDSZEREK.....	13
3.1. AZ IDPCOMPLEX ADATBÁZIS	13
3.2. AZ ADATBÁZIS AUTOMATIZÁLÁSI LEHETŐSÉGEI	14
3.3. AZ FEHÉRJEKOMPLEXEK MÉRÉSI ADATAINAK BIOINFORMATIKAI ELEMZÉSE	16
4. EREDMÉNYEK.....	17
4.1. AZ IDPCOMPLEX ADATBÁZIS	17
4.2. AZ ADATBÁZIS NEM REDUNDÁNS ADATKÉSZLETÉNEK BIOINFORMATIKAI ELEMZÉSE	21
4.3. FEHÉRJEKOMPLEXEK HOZZÁFÉRHETŐ FELSZÍNE ÉS KÖTŐINTERFÉSZE	25
4.4. NÉHÁNY PÉLDA.....	27
4.5. AZ ADATBÁZIS AUTOMATIZÁLÁSI LEHETŐSÉGEI	29
5. DISZKUSSZIÓ.....	30
7. IRODALOMJEGYZÉK.....	32
8. MELLÉKLET	34

1. Bevezetés

1.1. Absztrakt

Mára már elfogadott tény, hogy számos, alapvető biológiai folyamatban szerepet játszó fehérje natív állapotban nem rendelkezik jól meghatározott 3D-s szerkezettel. Ezek a rendezetlen fehérjék leginkább gyorsan változó, egymásba átalakuló alternatív konformációk sokaságaként jellemezhetők. Az irodalomban elfogadott nézet szerint - a rendezetlen szerkezetük következményeként - specifikus, de gyenge fehérje-fehérje kölcsönhatás kialakítására képesek. Erre vonatkozóan azonban még nem született átfogó vizsgálat, és számos megfigyelés utal arra, hogy ez az általánosítás nem állja meg a helyét. A feladatom a rendezetlen fehérjéket tartalmazó komplexek fiziko-kémiai, szerkezeti, biofizikai és egyéb adatainak irodalomból történő kigyűjtésében való részvétel, az összegyűjtött adatokból adatbázis létrehozása volt, mely alapul szolgál további részletes bioinformatikai vizsgálatokhoz.

Munkám során a rendezetlen fehérjepartner tartalmazó komplexek kötődésének tanulmányozásával, a rendezetlen fehérjekomplexek rendszerező adatbázisának programozásával és tesztelésével foglalkoztam, továbbá a disszociációs állandót tartalmazó adatbázisok megismerése és bennük rendezetlen komplexek bioinformatikai keresése tartozott a feladataim közé.

Jelenleg a rendezetlen fehérjék kötődésének bioinformatikai feltérképezését végzem mind a kötőaminosavak, mind a szerkezetek kötőfelszíneinek szempontjából, valamint a rendezetlen és globuláris fehérjék komplexeinek összehasonlító elemzésén dolgozom.

1.2. Rendezetlen fehérjék

Az ezredforduló előtt elképzelhetetlennek tűnt, hogy megdőljön a molekuláris biológia „szerkezet - funkció paradigmája”, miszerint a fehérjék funkcióját a jól meghatározott 3 dimenziós szerkezet határozza meg. Azonban egyre több bizonyíték gyűlt össze arra vonatkozólag, hogy létezik a fehérjéknek egy másik csoportja, a rendezetlen fehérjék (angolul: intrinsically disordered proteins, IDPs), és mára a rendezetlenség fogalma széles körben elfogadottá vált. [1] [2] [3]

Ezek a fehérjék in vivo sem rendelkeznek jól definiált rendezett szerkezettel, leginkább gyorsan változó, egymásba átalakuló alternatív konformációk sokaságaként jellemezhetők.

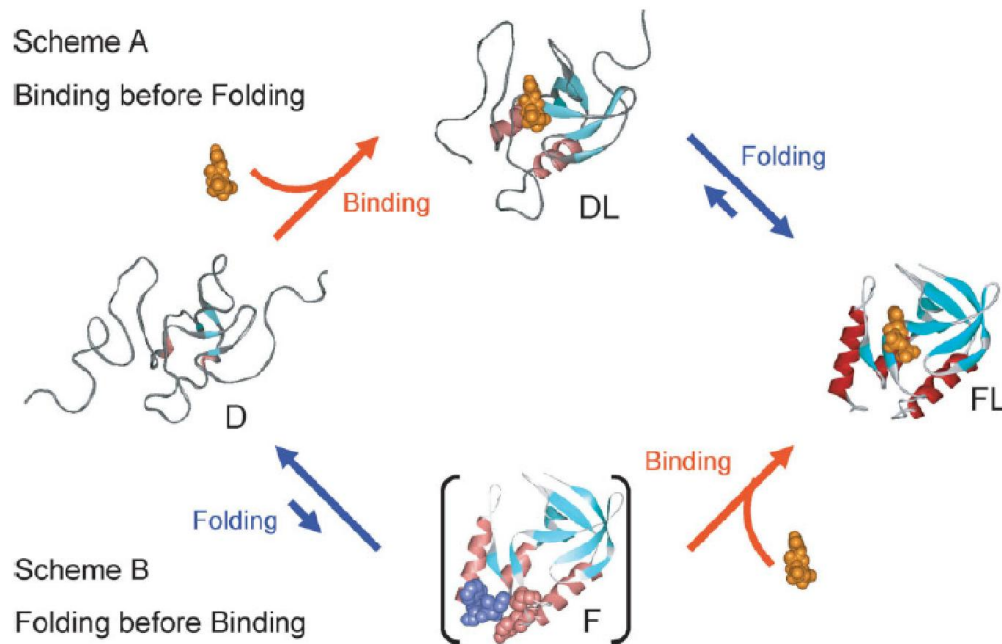
Különbféle mérési eredmények azt mutatják, hogy ezek a fehérjék/fehérje régiók különös egyediséget mutatnak a töltés - hidrofóbicitás vizsgálatokban, nem hasonlósak a rendezett, globuláris fehérjékhez. Alacsonyabb átlagos hidrofóbicitás és magasabb nettó töltés jellemzi őket [2]

A Protein Data Bank (PDB) gyors bővülésével és új, főként mágneses magrezonancia (NMR) és cirkuláris dikroizmus (CD) vizsgálatok eredményeként számtalan új rendezetlen szerkezetű fehérjét sikerült felderíteni. [1][4] Ez adta meg a lendületet ahhoz, hogy létrejöhessen a rendezetlen fehérjéket tartalmazó adatbázis, a DisProt, mely napjainkra 1500-nál is több rendezetlen fehérje régiót tartalmaz. [3][5] A kísérletes megközelítésen túl megpróbálták a rendezetlenséget in silico módszerekkel jósolni. Az első és legfontosabb rendezetlen régió prediktor, amelyből később több jóslóprogramot is kifejlesztettek, a PONDR volt. Ez a prediktor az aminosav sorrend alapján határozza meg, hogy egy adott szekvenciában melyik régió lehet rendezetlen. A prediktor neurális hálózat tanításán alapul, melyhez NMR spektroszkópiás, illetve röntgen-krisztallográfiás mérésekkel azonosított rendezett illetve rendezetlen fehérjerégiókat használtak fel. Az NMR alapú jóslás jobban működött, de megjelenésekor mindkét PONDR prediktor teljesítménye elismerendő volt. [6] Azóta számos rendezetlenség jósló programot készítettek, a legfejlettebbek pontossága meghaladja a 85%-ot. [7] Bioinformatikai becslések alapján a humán fehérjék több mint egyharmada tartalmaz legalább egy hosszú (minimum 30 aminosav hosszú) rendezetlen régiót. [8]

1.3. Rendezetlen fehérjék kötődése, komplexek képződése

A rendezetlen fehérjék gyakran játszanak szerepet szabályzó folyamatokban, például a jelátvitelben vagy a transzkripcióban. Ezek a funkciók molekuláris felismerésen alapulnak, vagyis a partnerhez való átmeneti vagy tartós kötődés által. Ahhoz, hogy megértsük ezen fehérjék működését és szerepét, nagyon fontos, hogy feltérképezzük a kötődés mechanizmusát. Általánosan elfogadott az a nézet, hogy a szerkezeti rendezetlenség szétválasztja a specifitást és a kötőerősséget, és így gyenge, reverzibilis, mégis specifikus kapcsolat jöhet létre két fehérje között. Bár rengeteg kutatás irányult rendezetlen fehérjék komplexeinek vizsgálatára, a kötődés folyamatát máig nem sikerült teljes részleteiben megérteni. [9] Több mechanizmus is ismert az IDP-k kötődésével kapcsolatban. Ezek közül a két legelterjedtebbet fogom megemlíteni.

Az első mechanizmus szerint a rendezetlen fehérjék feltekeredése/rendeződése (folding) csak a kötőpartner jelenlétében, a dokkolás megkezdése után megy végbe. A másik mechanizmus során, mivel ezek a fehérjék nem rendelkeznek állandó szerkezettel, hanem szerkezetük állandóan változó állapot sokasággal írható le, a dokkoláshoz az a konformer választódik ki, amely épp a kötődéshez alkalmas állapotban van. Az angol szakirodalom a két megközelítésre az induced fit/folding és conformational selection kifejezéseket használja. [1]



Ábra 1.1.: Induced folding, conformational selection

[10]

Az ábra a rendezetlen fehérjék kötődésének kétféle mechanizmusát mutatja be. Az ábra felső részén a kötődés megelőzi a folding folyamatát, míg az alsó részén a folding előzi meg a dokkolást.

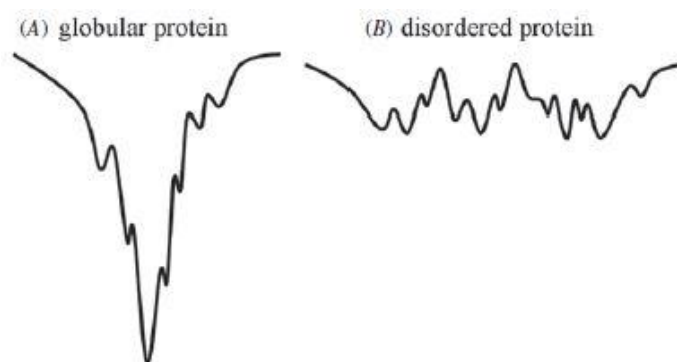
Kísérletekkel mindkét mechanizmusra számos példát találtak, sőt előfordul az is, hogy ugyanaz a fehérje mindkét mechanizmussal képes kötődni. [10]

Termodinamikai nézőpontból a reakciók lefutását a Gibbs-féle szabadentalpia változása vagy más szóval szabadenergia változás (ΔG°) határozza meg. A kezdő- és végállapot közti szabadentalpia változás határozza meg a K egyensúlyi konstans nagyságát is:

$$K = e^{\frac{-\Delta G^\circ}{RT}}$$

A Gibbs-féle szabadentalpia változás három tényezőtől függ: a hőmérséklettől, és az entalpia (H), valamint az entrópia (S) változásától: $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$. Ez azt jelenti, hogy a kötésaffinitás erősségét befolyásolja a kötés során fellépő entalpiaváltozás, ami a kölcsönhatásba lépő molekuláktól függ, és az entrópiaváltozás, ami a szerkezeti sajátosságoktól függ. A legkorábbi elképzelés a fehérjék komplexeinek képződéséről a kulcszár elmélet volt, mely a partnerrel való kémiai és geometriai komplementaritást feltételezi. Azonban ma már tudjuk, hogy a szerkezet nem mindig merev, így a kötőhely a kötődéskor némileg módosulhat, vagyis szabad és kötött állapotban a fehérje konformációja nem egyezik meg teljesen. Ezt induced fit-nek nevezi az angol szaknyelv. A szabad és a kötött állapot konformációja közötti különbség a konformációs szelekció esetében ugyanígy fennáll. [11]

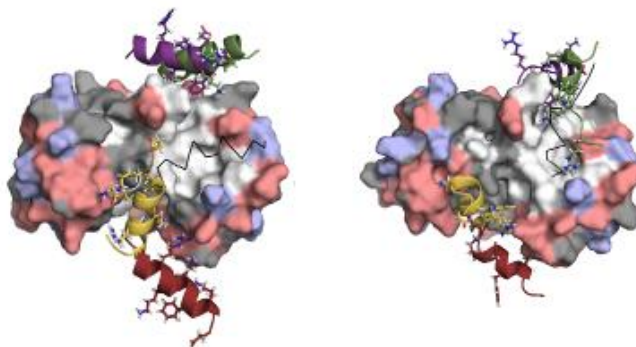
Összehasonlítva a jól definiált szerkezettel rendelkező fehérjék és a natív állapotban is rendezetlen szerkezetű fehérjék konformációs energiáit, a különbség markáns, hiszen a rendezett fehérjék energiadiagramján tölcsér alakú grafikon rajzolódik ki, amin a magas energiájú állapotok kedvezőtlenek a fehérje számára, míg a tölcsér alján lévő pontok reprezentálják a kedvező, általában egy jól meghatározott, konformációt. Ezzel szemben az IDP-k grafikonja hullámzó, a globális minimum energiaállapota nem különbözik sokban a többi lokális minimum konformációnak energiájától. Ez teszi lehetővé a fent említett könnyed konformáció változtatást és a jól definiált szerkezet, valamint a belső, intramolekuláris kötődések hiányát. [11]



Ábra 1.2.: Globuláris (A) és rendezetlen fehérjék (B) konformációs energiaprofilja [11]

A sematikus ábrán látszanak a két fehérjecsalád konformációihoz tartozó energiaállapotok. Látható, hogy az IDP-k nem rendelkeznek optimális konformációval és ahhoz tartozó globális minimummal.

A rendezetlen fehérjék nagyon jó jelátviteli molekulák, melyet annak a képességüknek köszönhetnek, hogy a partnerhez való kötődés során képesek feltekeredni az éppen szükséges konformációba. Előfordulhat, hogy egy adott rendezetlen szakasz képes kötődni több rendezett partner kötőzsebébe, vagy az, hogy több különböző rendezetlen fehérjéből származó szakasz kötődik egy és ugyanazon globuláris partner kötőhelyeire. Az első esetet one-to-many, míg a másodikat many-to-one jelátvitel közé soroljuk. [9]



Ábra 1.3.: A rendezetlen p53 és TRTK-12 kötődése a rendezett S100B fehérjéhez. [9]

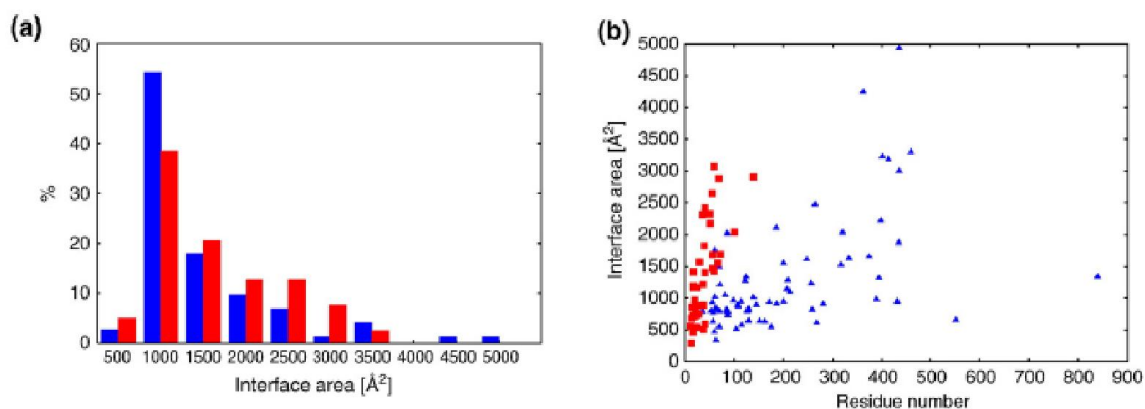
Az ábrán baloldalt a p53 – S100B komplex, a jobb oldalon a TRTK-12 – S100B komplex látható. Ebben a két esetben olyan kötődést láthatunk, ahol a „folded” állapotba kerülő, különböző fehérjéből származó rendezetlen szakaszok ugyanazon partner molekula hidrofób zsebébe kötődnek.

A kötőfelszínek tekintetében is nagy különbségeket találhatunk a rendezett és rendezetlen fehérjék elemzésekor. Nem is a kötőfelszínek nagyságát, hanem a kötőfelszínek és a lánchossz arányát tekintve. Az IDP-k képesek kisebb lánchosszal nagyobb relatív felszínt megvalósítani, mint a globuláris fehérjék, vagyis nagy interakciós potenciállal rendelkeznek. Az IDP-k esetében a kötőfelszín (interfész) és a teljes hozzáférhető felszín arány nagyobb, mint a rendezett fehérjék esetében, főleg, ha az interfész méretét az aminosavak számára vetítjük: [12][13]

Class	ASA interface ASA monometer	ASA interface No. of residues	ASA monometer No. of residues
Stable monomer (three-state; $n = 5$) [p53 (tetramer; 1sae), GCN4-p1 (dimer; 1ysa), β -nerve growth factor (dimer; 1bet), trp-aporepressor (dimer; 3wrp), bowman-birk protease inhibitor (complex; 1d6r)]	0.19–0.38	7–17	37–49
Disordered monomer (two-state; $n = 4$) [tryptophan synthase (dimer; 1ttq), aspartate aminotransferase (dimer; 1tar), glutathione S-transferase (dimer; 1glq) and triose phosphate isomerase (dimer; 1tim)]	0.29–0.66	26–82	67–123
Ribosomal proteins (30S; $n = 19$) [S3 to S20 and THX ribosomal proteins of <i>thermus thermophilus</i> 30S ribosomal subunit (1fj)]	–/–	–/–	56–103

Ábra 1.4.: Rendezetlen és rendezett fehérjekomplexek kötőfelszínének, teljes felszínének és aminosav hosszának vizsgálata I. [12]

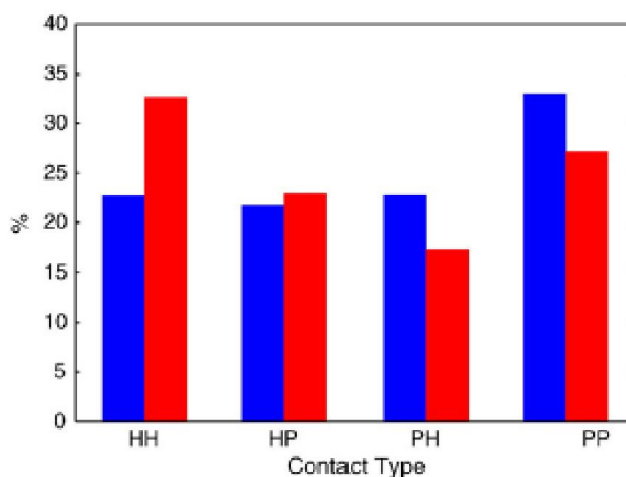
Az ábra a különböző fehérjecsaldok kötőfelszínének (accessible surface area (ASA) interface) és a teljes hozzáférhető felszínének (ASA monometer), valamint az aminosavak számának arányát, illetve a teljes hozzáférhető felszínük és az aminosavak számának arányát írja le.



Ábra 1.5.: Rendezetlen és rendezett fehérje komplexek kötőfelszínének, teljes felszínének és aminosav hosszának vizsgálata II. [13]

Az (a) ábra a rendezett (kék) és a rendezetlen (piros) fehérje-komplexek kötőfelszínének eloszlását mutatja, míg a (b) ábra az interfész nagyságát az aminosav-szekvencia hosszának függvényében ábrázolja.

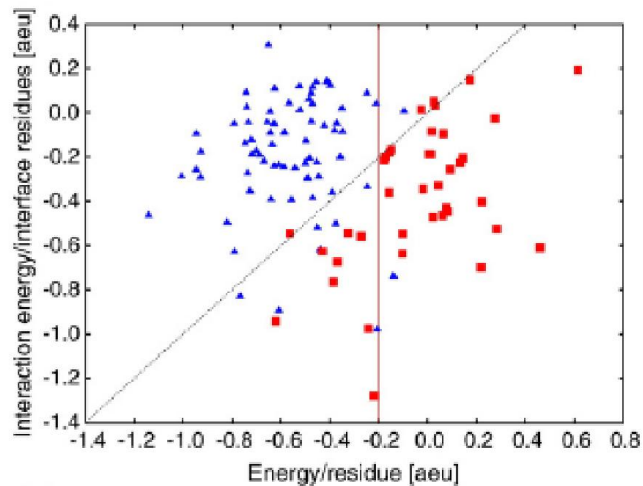
A fent említett két fehérjetípus komplexei nem csak a kötőfelszín nagyságának és az aminosavak számának arányában mutatnak különbségeket. A különböző típusú kötődő aminosav párok (hidrofób-hidrofób, hidrofób-poláros, poláros-hidrofób, poláros-poláros) aránya is mérhető a komplexek összehasonlításában. Az erre irányuló mérési adatokból azt olvashatjuk le, hogy az IDP komplexekben sokkal gyakoribb a hidrofób-hidrofób aminosav-kapcsolat, mint a globuláris fehérjék esetében, és sokkal ritkább a poláros-hidrofób és a poláros-poláros arány. [13]



Ábra 1.6.: Kötődő aminosav párok rendezetlen (piros) és globuláris fehérjében (kék). [13]

Az ábrán a hidrofób aminosavak H-val, míg a poláros aminosavak P-vel vannak jelölve.

Szintén említésre méltó különbség a tárgyalt fehérjecsaládok között a kötőfelszínek energetikai profiljai, melyek egyértelműek a két fehérjetípust ismerve, hiszen tudjuk, hogy a rendezett fehérjékkel ellentétben a rendezetlen fehérjékre nem jellemző intramolekuláris kötések kialakítása. A diagramon is látszik, hogy az adatokat ábrázolva a két fehérjetípus jól elkülönül. [13]



Ábra 1.7.: Rendezetlen (piros) és rendezett (kék) fehérjék energetikai profilja [13]

Az ábra x tengelyén az energia/aminosav érték, míg az y tengelyén a kölcsönhatási energia/kötőfelszínen lévő aminosavak száma látható, ami jól szeparálhatóan leképezi a kétféle típusú fehérje jellemző energetikai tulajdonságait.

1.4. Fehérje aminosav-aminosav interakciók típusai

A bioinformatikában az egyik legkézenfekvőbb módszer a fehérjekomplexek láncai közti interakciók megállapítására, ha a különböző láncok aminosav-aminosav kölcsönhatásait elemezzük. Én a következő kölcsönhatásokat vizsgáltam: hidrofób, hidrogén híd, ionos, aromás-aromás, aromás-kénés és kation-pi, melyek az alábbi kritériumoknak felelnek meg:

Hidrofób kölcsönhatással számolunk, ha ALA, VAL, LEU, ILE, MET, PHE, TRP, PRO, TYR aminosavak a különböző láncokban 5 Å távolságon belülre kerülnek egymáshoz. [14]

Hidrogén híd meglétét feltételezzük, amennyiben a donor-akceptor távolság a oxigén és nitrogén közt a láncokban nem nagyobb, mint 3,5 Å vagy a donor-akceptor távolság a kén esetében kisebb vagy egyenlő, mint 4 Å. [15]

Az ionos kötések kritériuma, hogy maximum 6 Å távolságban legyenek ARG, LYS, HIS, ASP, GLU párok.

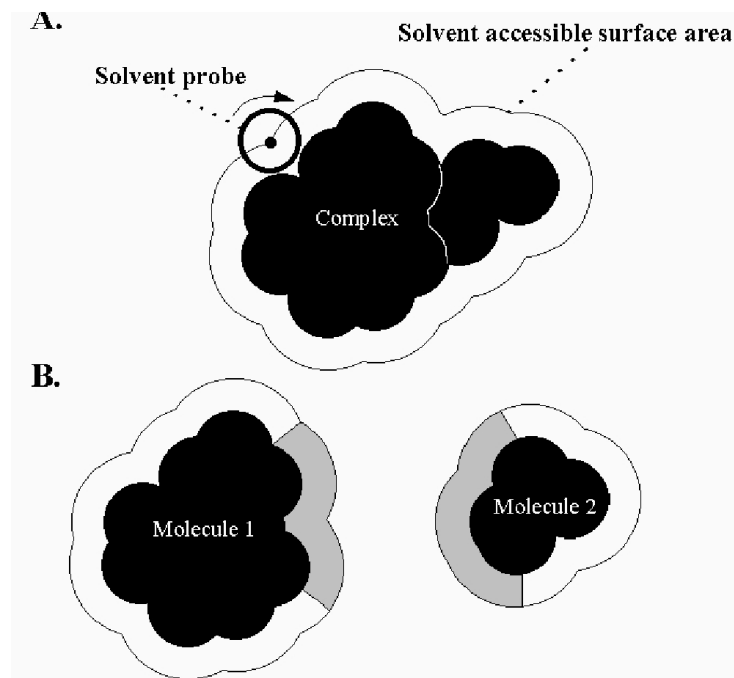
Az aromás-aromás interakció feltétele, hogy a benzolgyűrűk centruma minimum 4,5 – maximum 7 Å távolságra legyen. [16]

Aromás-kénés kölcsönhatással számolunk, ha a cisztein vagy a metionin kénatomja és a fenilalanin, tirozin vagy a triptofán aromás gyűrűje 5,3 Å távolságon belül helyezkedik el. [17]

Kation-pi interakciót definiálunk, ha lizin vagy arginin oldallánc kerül 6 Å-nél közelebb fenilalanin, tirozin vagy triptofán oldallánchoz.

1.5. Fehérjekomplex hozzáférhető felszíne

Adott fehérjekomplex hozzáférhető felszíne alatt a komplex oldószermolekula (általában víz) általi próbával meghatározott felszínét értjük (ahova az oldószermolekula képes behatolni), pontosabban a (víz)molekula középpontja által leírt felszínt (Ábra 1.8./A). A fehérjekomplex kötőfelszíne értelemszerűen nem lesz elérhető az oldószer számára az "eltemetettsége" miatt, tehát a kötőinterfészt számolhatjuk a fehérje magányos illetve kötött formában lévő hozzáférhető felszínének különbségéből (Ábra 1.8./B).



Ábra 1.8.: Hozzáférhető felszín [18]

Az oldószermolekula próbájával (víz esetében ezt vízpróbának mondjuk) határozhatjuk meg a fehérjekomplex hozzáférhető felszínét.

2. Célkitűzések

Kutatásom egy nagyobb projekt része az MTA TTK Enzimológiai Intézetben, ahol bioinformatikai eszközökkel vizsgálunk rendezetlen szerkezetű fehérjéket, úgynevezett IDP-eket (intrinsically disordered proteins). A kutatás fő célja az, hogy a rendezetlen fehérjét tartalmazó komplexek kötődését megismerjük, fizikai-kémiai hátterét megértsük és figyelemmel kísérjük a fehérjék szerkezeti változásait. Ezen komplexek vizsgálata különösen fontos, hiszen fontos jelátviteli útvonalakban játszanak szerepet, de részt vehetnek a kóros szerkezeti átalakulással és neurodegeneratív következményekkel járó amiloid képződésben is.

Munkám fő célja a rendezetlen fehérjék és komplexek képződésének megismerése volt, valamint a laborban végzett in silico vizsgálatok közül az ismert komplexekről szóló cikkek gyűjtésével párhuzamosan ezek szűrését végeztem. Az összegyűjtött adatokból rendszerező adatbázist hoztam létre, melynek vizsgálom az automatizálási lehetőségeit is. Megkezdtem az adatok osztályozását és feldolgozását bioinformatikai eszközökkel. Az adatbázisba feltöltött

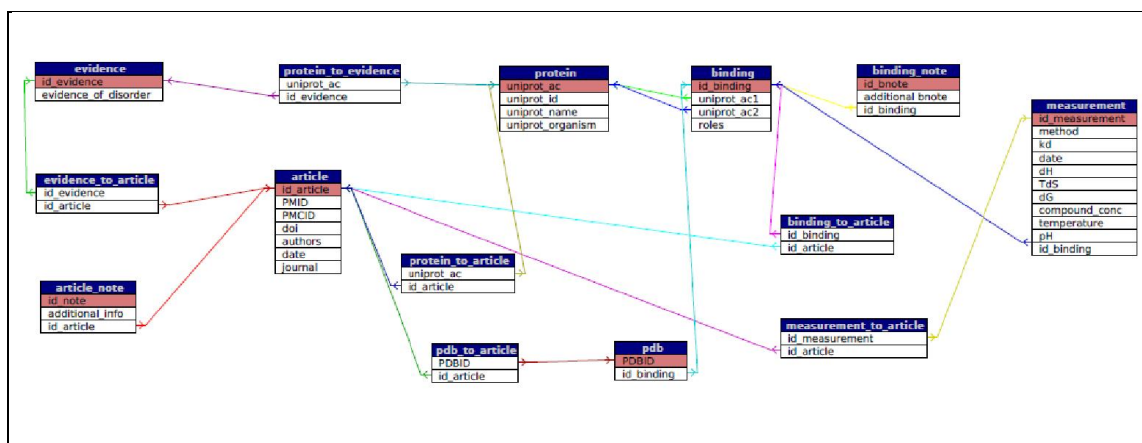
rendezetlen fehérjék komplexeinek adatain különböző bioinformatikai webserverek segítségével és saját szkriptekkel különböző eloszlási statisztikákat és kvantitatív szerkezeti biológiai elemzéseket futtattam.

Ezt követően in vitro kutatásokra is lehetőségünk nyílt, amellyel igazolhatjuk a prediktált következtetéseket.

3. Módszerek

3.1. Az IDPCOMPLEX adatbázis

Létrehoztam phpMyAdmin felület alatt az adatbázis relációs modelljét MySQL nyelven, mely tartalmazza a komplexet, a fehérjepartnereket, a bizonyítékot a rendezetlenségre, a mérési adatokat, a PDB szerkezeteket és az ezeket leíró cikkeket is, mint entitásokat. [19] [20]



Ábra 3.1.: A rendezetlen fehérjék komplexeinek kötődését leíró adatbázis relációs sémája

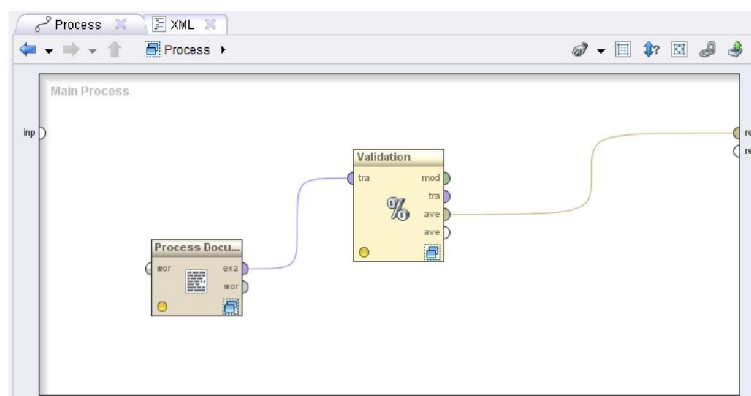
A sémában a téglalapok jelzik az entitásokat, a nyílak a relációkat. A „több az egyhez kapcsolatban” nyíl mutat a többről az egyhez, tehát a nyíl hegye hozzárendelt attribútumra mutat

3.2. Az adatbázis automatizálási lehetőségei

A szűrni kívánt cikkeket két oldalról is megközelítettem: osztályoztam az adatbázisainkban már szereplő teljes cikkeket, valamint az internetes adatbázisokból elérhető cikkek „Abstract”-jait, ami a későbbi automatizált szűréshez különösen fontos.

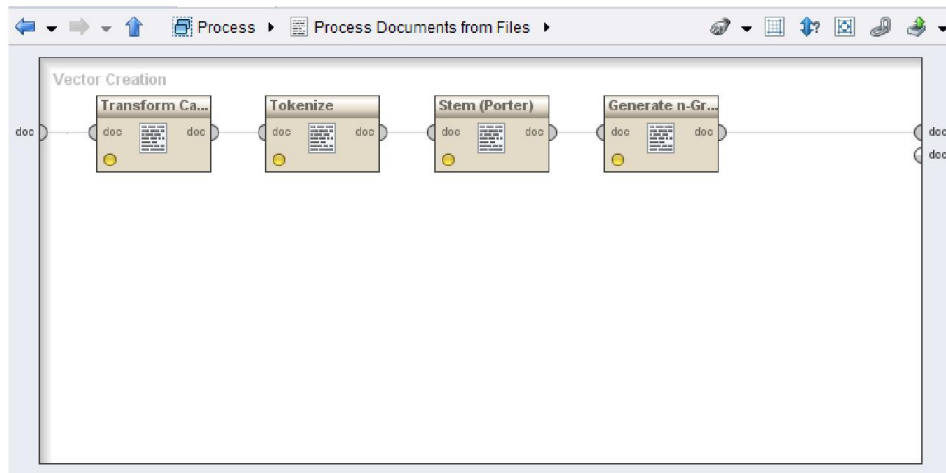
Első lépésként a PDF fájlokat – melyeket Perl szkript és XPDF pdftotext eszközök segítségével először szövegfájlá alakítottam –, valamint az „Abstract”-okat – melyeket Perl szkripttel és BIOPYTHON eszközökkel töltöttem le PubMed-ről –, előszűrtem Perl nyelven futó, reguláris kifejezéseket használó programmal, mely kiszűrte a számunkra lényegtelen cikkek nagy részét. [21] [22][23][24]

A második lépésben tanulóalgoritmusokkal szétválasztottam az adathalmazokat, és megvizsgáltam, melyik képes legmegbízhatóbban kiszűrni az ismert, manuálisan átvizsgált pozitív, illetve negatív cikkhalmaz alapján a találatokat. Fontos kritérium volt, hogy a fals-pozitív találatok mennyiségét lesoorítsuk. A fent leírtakhoz a Rapid-I –RAPIDMINER 5.3 szöveg és adatbányász szoftvert használtam, melyben először feldolgoztam a bemenő cikket, majd valamilyen tanuló algoritmus modellt futtattam rajta. [25]



Ábra 3.2.: RAPIDMINER modell

Process Documents – Validation – Result modell



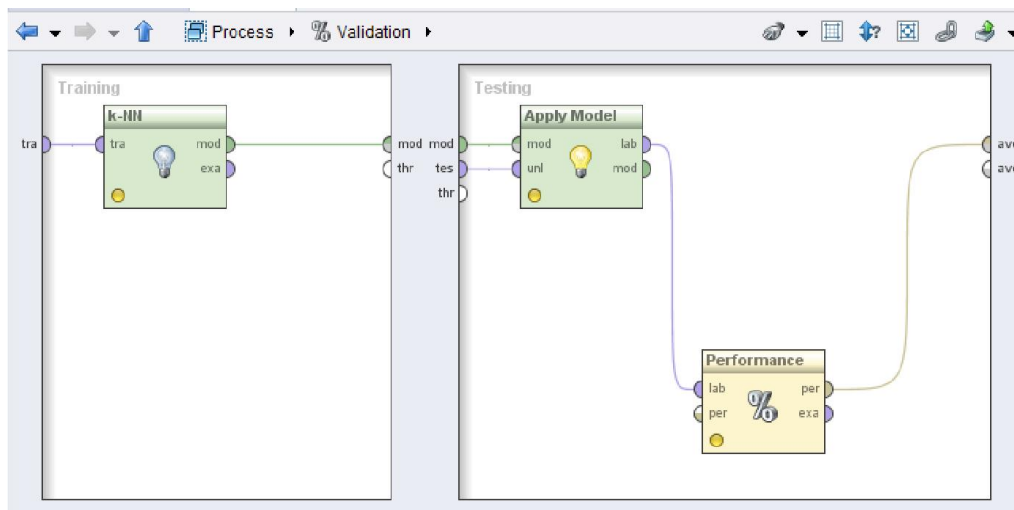
Ábra 3.3.: RAPIDMINER Process Documents modell

TransformCases: minden karakter kisbetűvel szerepel

Tokenize: tokenizálja (szavak) a szöveget

Stem (Porter): eltávolítja tokenek toldalékait

Generate n-Grams: kifejezéseket keres szomszédos tokenekből



Ábra 3.4.: RAPIDMINER Validation modell

k-NN: k-NearestNeighbor tanulóalgoritmus (k:=32)

Apply Model: tanítja az algoritmust, alkalmazza az adott modellt a szöveghalmazra

Performance: statisztikát készít a klasszifikáció pontosságáról

3.3. Az fehérjekomplexek mérési adatainak bioinformatikai elemzése

A különböző módszerrel, de azonos vagy hasonló körülmények közt mért komplexek disszociációs állandói és a módszerek közt végrehajtott korrelációelemzés számításokhoz a GraphPad Prism 5 programot használtam, mely egy kifejezetten biostatistikai adatelemzésre, korrelációszámításra, görbék illesztésére, lineáris és nonlinearis regresszióra és a mérési adatok kirajzoltatására kifejlesztett szoftver. [26]

Szintén a GraphPad Prism 5 programot használtam az adatok logaritmálására, t próbákra és normalitásvizsgálatokra (Kolmogorov-Smirnov teszt, D'Agostino-Pearson normalitásvizsgálat, Shapiro-Wilk normalitásvizsgálat).

Az egyéb alapvető statisztikai eredményeket, eloszlásokat a Microsoft Excel 2007 szoftver segítségével számítottam ki. [27]

A fehérjekomplexek kötőrégióit először az „A Tools for Calculating Binding-site Residues on Proteins (TCBRP)” webszerver [28] segítségével számoltattam ki, mely egy távolságalapú megközelítést alkalmaz: azt tekinti aminosav-kontaktnak, ha két különböző láncban lévő aminosav egymáshoz adott távolságra kerül. [29] Ezzel a módszerrel azonban olyan aminosavakat is kötőnek vettünk, melyek valójában nem vesznek részt a kötés kialakításában, ezért más módszert kerestünk. A „PIC: Protein Interactions Calculator” webszerver számítási módszereit teljesen alkalmasnak ítéltük a feladat ellátására, hiszen ez csak a kötő aminosavakat veszi számításba. [30] [31] A kölcsönhatások csoportosítására az 1.4. fejezetben bemutatott besorolást és kritériumokat alkalmazza.

A szerkezetek interakciós profiljainak meghatározására tehát a PIC webserveren történő számításokat kellett automatizálnom, melyeket Perl szkriptekkel sikerült lefuttatnom, mely a PDB szerkezet letöltése után feltöltötte a szerkezetet a webserverre és lementette a kimenetet HTML (HyperText Markup Language) formátumban. A HTML formátumú fájlokat HTMLAsText v1.11 szoftver segítségével konvertáltam TXT formátumú fájlkká, melyek már könnyen feldolgozhatóak újabb szkriptekkel. [31] [32]

A fehérjekomplexek hozzáférhető felszínének számításához a PIC webszervert használtam, majd ennek nagyfokú elérhetetlensége miatt a Helix Systems – StrucTools webserverre tértem át, mely több más funkciója mellett szintén képes a fehérjék és fehérjekomplexek hozzáférhető felszínének számítására. A webszerver automatizálása Perl szkriptekkel történt,

mely HTML formátumban mentette le a szerkezeti számítások kimenetét, majd ezeket a fájlokat HTMLAsText v1.11 szoftver segítségével alakítottam át TXT kiterjesztésű fájlakká. [31] [32] [33] [34] A kötőinterfészek kiszámítása az 1.5. fejezetben leírtak szerint történt.

4. Eredmények

4.1. Az IDPCOMPLEX adatbázis

Az adatbázis előzetes formája jelenleg (2013. november) statikus, vagyis táblázatszerű, és csak egy nem redundáns adatkészlet található meg benne, mely rendezetlen komplexek mindegyikéhez ismerünk PDB szerkezetet.

Ezek kiválasztása úgy történt, hogy az azonos fehérjekomplexhez tartozó adatokból (amennyiben több mérési eredmény is rendelkezésemre állt) azokat választottam be a nem redundáns adathalmazba, mely mérések lehetőleg 20-25 °C hőmérsékleten és 6,5-7,5 semleges pH-hoz közelítő körülményeken zajlottak. Ha több találat is megfelelt ennek a kritériumnak, akkor azt a találatot tartottam meg, amelyet lehetőleg ITC (Isothermal Titration Calorimetry) módszerrel mértek, hiszen ez egy megbízható módszer, és ezen méréseknél gyakran áll rendelkezésre termodinamikai adat is.

A közeljövőben azonban ezt a statikus adatbázist dinamikus adatbázissá tervezzük alakítani a fent leírt relációs sémának megfelelően, kiegészítve az összes általunk gyűjtött adattal. Az adatbázis jelenleg az alábbi linken, a saját intézeti weboldalamon érhető el:

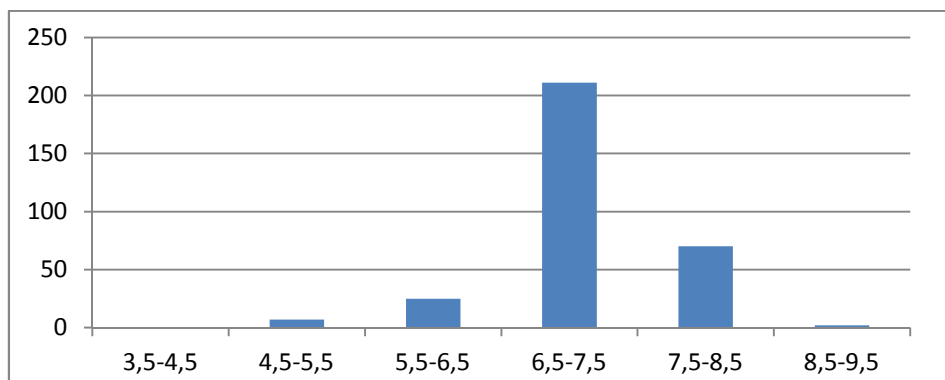
<http://www.enzim.hu/~lazta/>

Az eddig átnézett cikkek száma, melyekből hasznos találatokat annotáltunk, eléri a 140-et, ami 350 mérési eredményt jelent, ha megkülönböztetjük a különböző hőmérsékleten, pH-n, ionerősségen vagy különböző módszerrel mért disszociációs erősséget. Fontos megemlíteni, hogy a mérési eredmények száma még emelkedni fog, hiszen az adatgyűjtés nem zárult le, sőt még az ADAN adatbázisból származó 30 komplexre vonatkozó több mint 100 mérés sincs annotálva. [35] [36]

	A	B	C	D	E	F	G	H	I	K	L	M	N
1	relevance	IDP (UniprotAc)	IDP	partner (UniprotAc)	partner	evidence of disorder	k_D : equilibrium dissociation constant (nM)	PDB	PDB article	compound conc	temperature °C	pH	method
276 *		P63043	RBS	D0VW20 + D0VWY9	tubulin	predicating with iupred: res 138-149 (N-term) is disordered TCF binds along this groove as an elongated structure making contacts with three different binding modules		1ffx, 3ff, 500 3hkb	11030624, 21787788, 19666559	50 mM MES, 1 mM MgCl ₂ , 1 mM GTP, 1 mM EGTA 50 mM sodium phosphate buffer, 100 mM NaCl, 2 mM DTT			Fluorescence correlation 6,4 spectroscopy (FCS)
277 *		P70062	TCF4 res:1-53 and 1-56	P35222	β -catenin; full or armadillo repeat	resolution X-ray crystallography and complementary biophysical methods (NMR) have revealed that this eIF4E recognition motif undergoes a disorder-to-		2,4 1g3j	11136974			25	7,4 ITC (SPR, fluor)
278 *		P78344	eIF4E-binding proteins (4E-BPs) and oligopeptides: E1F4G11 peptide	P63073	(eukaryotic translation initiation factor) eIF4E plus 7-methyl-GDP	In this article: glycyl contain, extreme sensitivity to native state proteolysis, CD, prediction		150 1ejh	10394359	20 mM HEPES, 100 mM KCl		20	7,5 ITC
279 *		P83949	Drosophila melanogaster Hox protein: Ultrabithorax (Ubx) UbxHD (homeodomain)	DNA binding	DNA	In this article: glycyl contain, extreme sensitivity to native state proteolysis, CD, prediction		0,063 1b8l	10067897	100mM KCl, 20 mM Tris, 10% glycerol		22	DNA Gel Retardation (Gel 7,5 Shift) Experiments
280 *		P83949	Drosophila melanogaster Hox protein: Ultrabithorax (Ubx) UbxHD (homeodomain)	DNA binding	DNA	In this article: glycyl contain, extreme sensitivity to native state proteolysis, CD, prediction		0,0035 1b8l	10067897	100mM KCl, 20 mM Tris, 10% glycerol		22	DNA Gel Retardation (Gel 6 Shift) Experiments

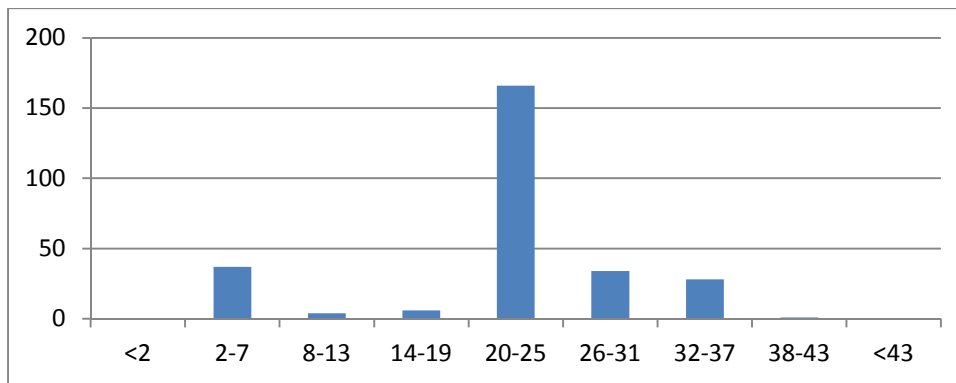
Ábra 4.1.: Mérési eredmények gyűjtése Excel táblázatban

Elemezve ezt a nagy adatmennyiséget, azt tapasztaljuk, hogy a mérések körülményei, módszerei és a kötődés erőssége is széles skálán mozog. A legkisebb kötőerősséget két fehérje közt 4 mM-os, a legnagyobbat pedig 4,4 pM-os erősséggel mérték, ami kilenc nagyságrend különbséget jelent, tehát elmondható, hogy a rendezetlen fehérjék alkotta komplexek esetében is találhatunk erős és gyenge kötődést egyaránt. A pH és a hőmérséklet is széles intervallumban mozog: savastól (pH: 5,5) az erősen lúgosig (pH: 10), valamint 4°C-tól 40°C-ig. A különféle módszereket név szerint nem említeném meg, hiszen nagyjából 25-30 különböző mérési módszert alkalmaztak a disszociációs állandók feltérképezésére, melyek közül a leggyakoribb módszerek az ITC, SPR (Surface Plasmon Resonance) és fluoreszcencia mérések voltak.



Ábra 4.2.: a teljes adatkészlet pH eloszlása

Az ábra vízszintes tengelyén a pH, míg a függőleges tengelyén az adatok mennyiségét ábrázoltam.

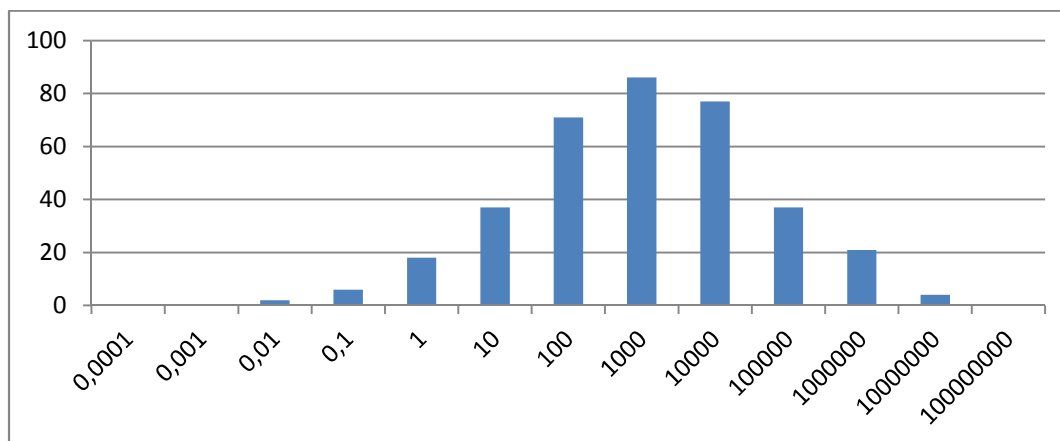


Ábra 4.3.: a teljes adatkészlet hőmérséklet eloszlása

Az ábra vízszintes tengelyén a hőmérséklet (°C), míg a függőleges tengelyén az adatok mennyiségét ábrázoltam.

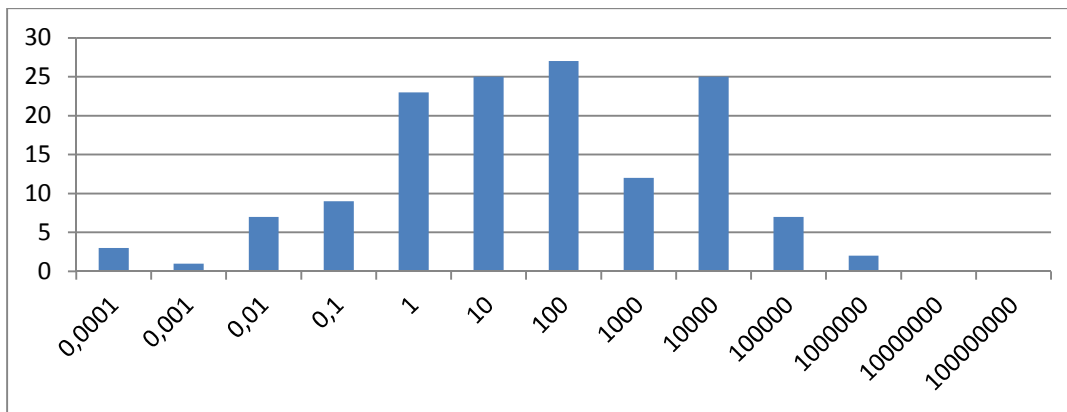
A különböző módszerrel, de azonos vagy hasonló körülmények közt mért komplexek disszociációs állandói és a módszerek közt végrehajtott korrelációelemzés sajnos nem hozott megbízható eredményt ezen párok kis számban való előfordulása miatt.

A legtöbb párt tartalmazó ITC-NMR esetében is csak 4-7 párt találtam, annak függvényében, hogy mit tekintünk hasonló körülménynek: Spearman korreláció itt -0,8000 és -0,8420 értékek közt mozgott (0,3333-0,0238 közti P value mellett).



Ábra 4.4.: A rendezetlen fehérjét tartalmazó komplexek disszociációs állandójának (K_D) eloszlása

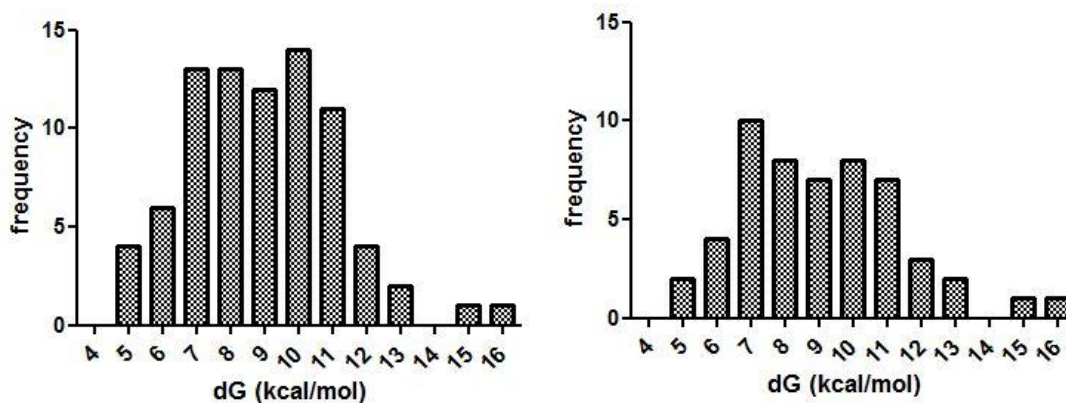
Az ábra vízszintes tengelyén a disszociációs állandót (nM), míg függőleges tengelyén az adatok mennyiségét ábrázoltam. Látható, hogy rendezetlen komplexek is rendelkezhetnek erős disszociációs állandóval.



Ábra 4.5.: Globuláris komplexek disszociációs állandójának (K_D) eloszlása [37]

Az ábra vízszintes tengelyén a disszociációs állandót (nM), míg függőleges tengelyén az adatok mennyiségét ábrázoltam. Ezen adatok referenciaként szolgálnak kutatásunk során.

Mivel a rendezetlen fehérjekomplexek nem redundáns adatkészletének disszociációs erősségeinek eloszlására is kíváncsiak voltunk, valamint arra, hogy az eloszlásuk normáeloszlást mutat-e, linearizálni kellett a logaritmikus vízszintes tengelyt, ezért a disszociációs konstansokat átszámoltuk a jól ismert $\Delta G^\circ = -RT \ln K_D$ képlettel Gibbs-féle szabadentalpiára.



Ábra 4.6.: Rendezetlen fehérjekomplexek nem redundáns adatkészletének kötési-erősségi (dG) eloszlása.

A baloldali ábrán a teljes nem redundáns adatkészlet, míg a jobboldali ábrán a 20-25°C, 6,5≤pH≤7,5 mérési eredményei láthatók kcal/mol-ban mérve. (Az ábrák globuláris megfelelői a mellékletben találhatók.)

Mind a teljes nem redundáns adatkészlet, mind a 20-25 °C, $6,5 \leq \text{pH} \leq 7,5$ adatok mind a három normalitásvizsgálata igazolta a Gauss eloszlást a rendezetlen komplexek esetében:

Teljes adathalmaz:

Kolmogorov-Smirnov teszt: $KS = 0,06347$ és $P \text{ value} > 0,10 \rightarrow$ Normál elo.

D'Agostino-Pearson normalitásvizsgálat: $K2 = 3,570$ és $P \text{ value} = 0,1678 \rightarrow$ Normál elo.

Shapiro-Wilk normalitásvizsgálat: $W = 0,9784$ és $P \text{ value} = 0,1886 \rightarrow$ Normál elo.

20-25 °C, $6,5 \leq \text{pH} \leq 7,5$ adathalmaz:

Kolmogorov-Smirnov teszt: $KS = 0,08173$ és $P \text{ value} > 0,10 \rightarrow$ Normál elo.

D'Agostino-Pearson normalitásvizsgálat: $K2 = 3,411$ és $P \text{ value} = 0,1817 \rightarrow$ Normál elo.

Shapiro-Wilk normalitásvizsgálat: $W = 0,9747$ és $P \text{ value} = 0,3201 \rightarrow$ Normál elo.

Annak eldöntésére, hogy a rendezetlen és rendezett nem redundáns adatkészlet szignifikánsan eltér-e egymástól, kétmintás t próbákat alkalmaztunk és az alábbi értékeket kapuk: $P \text{ value} < 0,0001$ a teljes adathalmazra is és a 20-25 °C és $6,5 \leq \text{pH} \leq 7,5$ adathalmazra is, tehát mindkét t próba erősen szignifikáns eltérést mutatott.

Ahhoz, hogy a találataink az adatbázisba feltölthető formába kerüljenek, a méréseket manuális úton kellett táblázatba rendezni, paraméterek szerint szétválasztani. A komplexekről összegyűjtött adatokat kiegészítettem az azokat egyértelműen definiáló azonosítókkal (pl. fehérjenév $\rightarrow$ UniprotAc), valamint az adott komplexekről szóló cikkek azonosítóival (PubMed ID [PMID] és doi). [38] [39] [40]

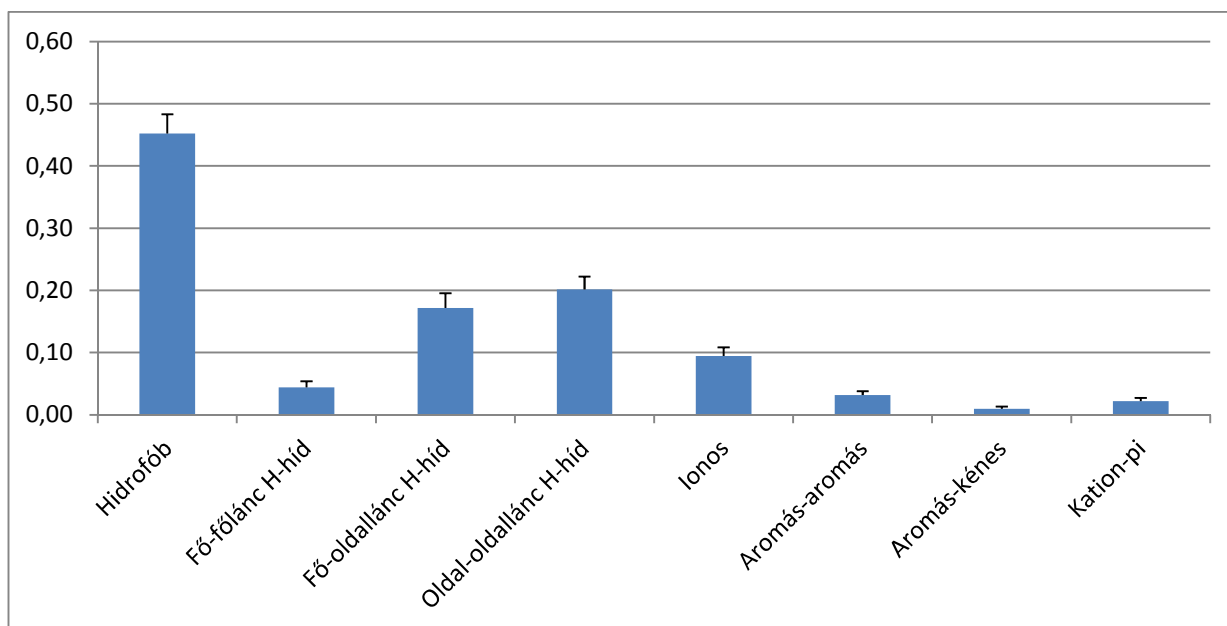
4.2. Az adatbázis nem redundáns adatkészletének bioinformatikai elemzése

Miután a rendezetlen fehérjekomplexek nem redundáns adatkészletét megkaptuk, kíváncsiak voltunk arra, mekkora a rendezetlen komplexek átlagos kötőrégiója, mekkora különbséget tapasztalunk a globuláris fehérjekomplexekkel összevetve. Referenciaként itt is az „A structure-based benchmark for protein–protein binding affinity” cikkben leírt globuláris fehérjeadatbázist használtam. [37]

A teljes nem redundáns adathalmazt vizsgálva az átlagos kötőrégió hossz 25 aminosav volt, míg a csak sztenderd (20-25 °C és $6,5 \leq \text{pH} \leq 7,5$) körülményeken mért adatok esetében ez átlagosan 32 aminosav volt. Kötőrégiónak az egymástól maximum 5 aminosav távolságra lévő kötőaminosavakat számítottam. A szórás a teljes és sztenderd adatkészlet esetében 2,4 és 3,8 aminosav volt. Ezzel szemben a globuláris komplexek adatbázisát vizsgálva 19,5 és 19,3 átlagos kötőrégió hosszúságot kaptam 1,3-as és 1,7-es szórással.

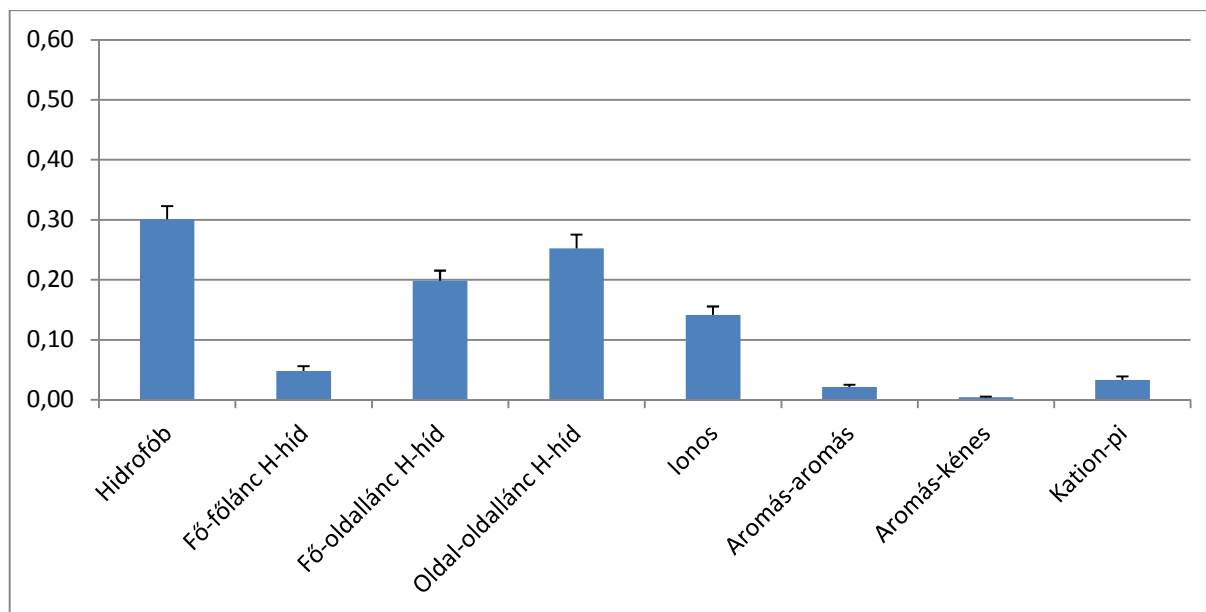
A továbbiakban az adatokat a kötésiáffinitás nagysága szerint három csoportra különítettem el: a gyenge kötéserősséggel rendelkezőkre ($K_D > 100 \mu\text{M}$), a közepes disszociációs állandóval rendelkezőkre ($100 \mu\text{M} \leq K_D \leq 100 \text{ nM}$) és az nagy kötéserősséggel rendelkezőket ($K_D < 100 \text{ nM}$). Sajnos a gyenge klaszter adatait nem találtam értékelhetőnek, ugyanis kevés mérési eredmény került ebbe a csoportba mind a rendezetlen, mind a globuláris fehérjekomplex adatbázisból. A közepes erősségű K_D -val rendelkező klaszter 20 aminosav hosszú átlagos kötőrégióval (szórás: 2,9), míg az alacsony K_D -val (vagy erős kötéssel) rendelkező komplexek átlagosan 31 hosszú kötőrégióval (szórás: 4) bírtak. A globuláris adatbázis esetében átlagosan 18,9 aminosav (szórás: 1,6) hosszú kötőrégióval rendelkező komplexek voltak közepes K_D mellett, ami az erős kötéssel (vagy alacsony K_D -val) rendelkező komplexek esetében 19,3 aminosav volt (szórás: 1,6).

Megalkottam továbbá a rendezetlen és globuláris fehérjekomplex adatkészletek interakciós profiljait, ahol hidrofób, főlánc-főlánc közti H-híd, főlánc-oldallánc közti H-híd, oldallánc-oldallánc közti H-híd, ionos, aromás-aromás, aromás-kénés és kation-pi interakciókat különböztettem meg.



Ábra 4.7.: Rendezetlen fehérjekomplexek nem redundáns adatkészletének interakciós profilja a 20-25 °C, $6,5 \leq \text{pH} \leq 7,5$ adatsávbán.

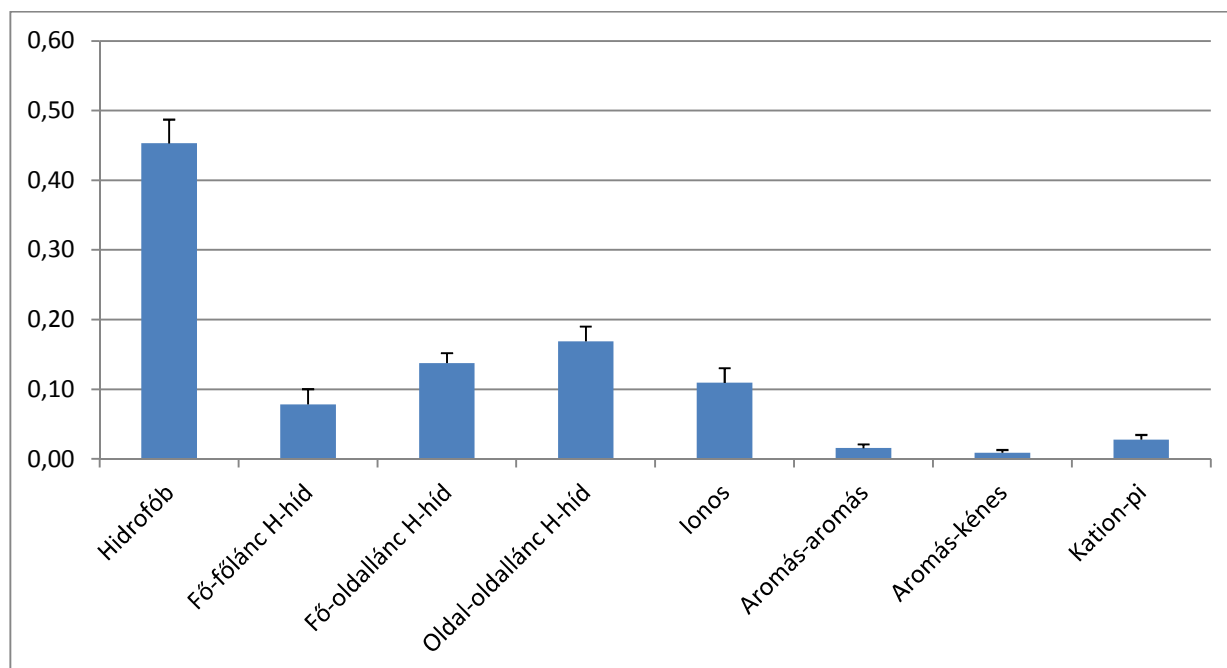
Az ábra vízszintes tengelyén az interakció típusokat, függőleges tengelyén az interakciók relatív mennyiségét ábrázoltam.



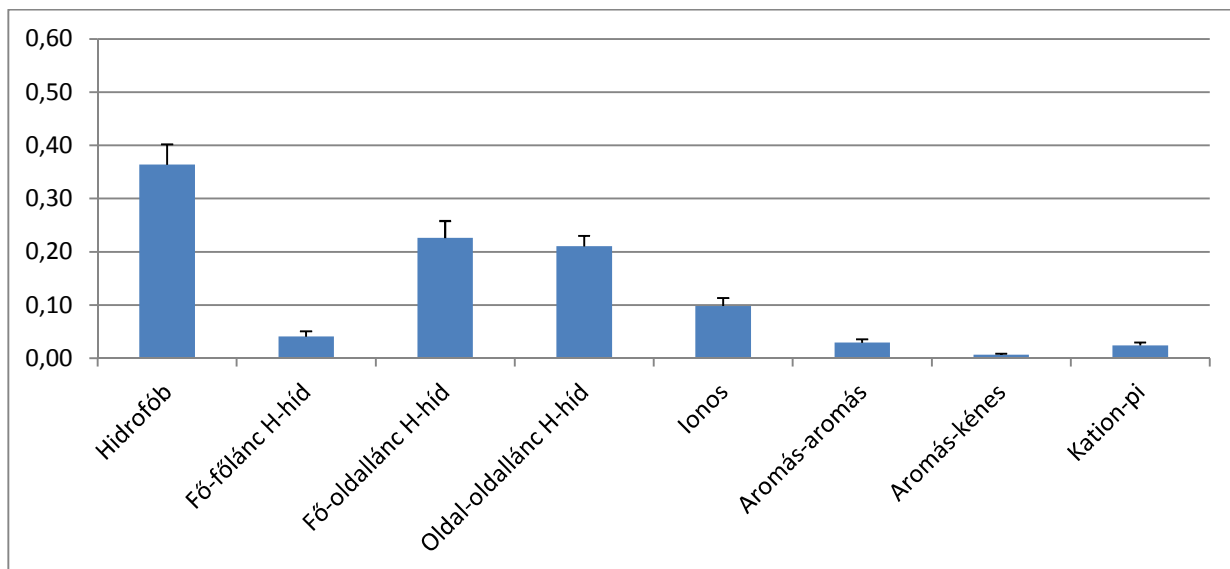
Ábra 4.8.: Globuláris fehérjekomplexek adatkészletének interakciós profilja a 20-25 °C, $6,5 \leq \text{pH} \leq 7,5$ adatsávban.

Az ábra vízszintes tengelyén az interakció típusokat, függőleges tengelyén az interakciók relatív mennyiségét ábrázoltam.

Közepes K_D :

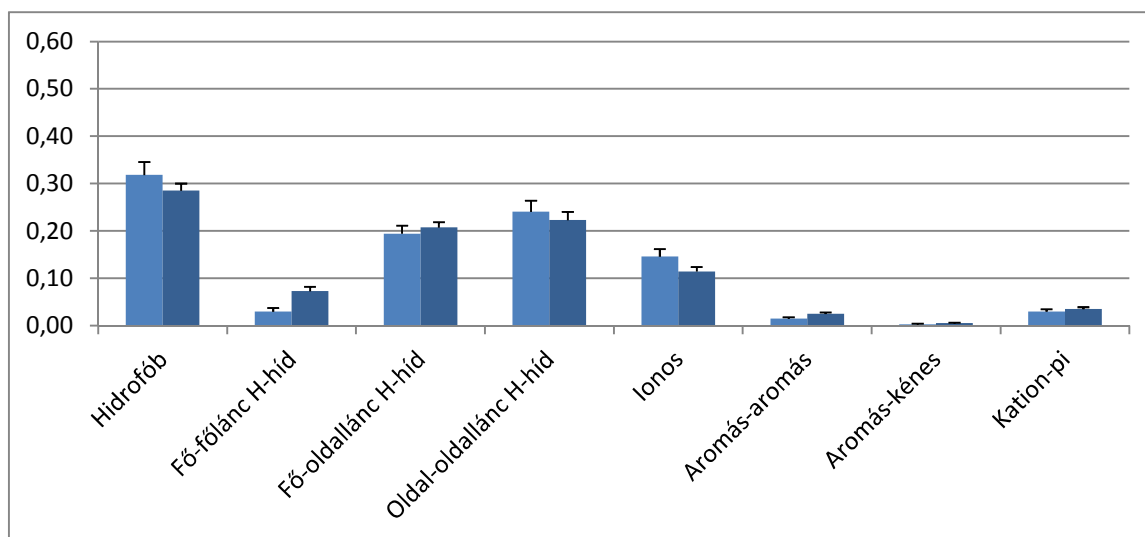


Alacsony K_D :



Ábra 4.9.-4.10.: Rendezetlen fehérjekomplexek nem redundáns adatkészletének interakciós profilja a közepes és alacsony K_D -val rendelkező adatok esetében.

Az ábra vízszintes tengelyén az interakció típusokat, függőleges tengelyén az interakciók relatív mennyiségét ábrázoltam.

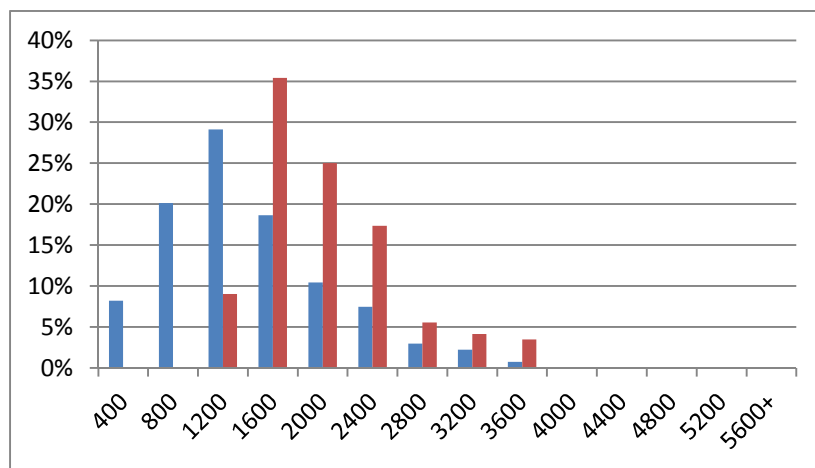


Ábra 4.11.: Globuláris fehérjekomplexek adatkészletének interakciós profilja referenciaként. Világoskék: közepes, sötétkék: alacsony K_D -val rendelkező adatok.

Az ábra vízszintes tengelyén az interakció típusokat, függőleges tengelyén az interakciók relatív mennyiségét ábrázoltam.

4.3. Fehérjekomplexek hozzáférhető felszíne és kötőinterfésze

Azért, hogy a rendezetlen és globuláris fehérjekomplexek kötőfelszínének nagyságát összevevük, mind a saját nem redundáns adathalmazunkra, mind a már korábban bemutatott eredmények esetében is referenciaként szolgáló „A structure-based benchmark for protein–protein binding affinity” cikkben leírt globuláris fehérjeadatbázisra is kiszámoltuk a hozzáférhető- és kötőfelszíneket. [37]



Ábra 4.12.: Az IDP (kék) és globuláris (piros) komplexek kötőfelszín eloszlása

Az ábra vízszintes tengelyén az kötőinterfészek mérete van ábrázolva, míg a függőleges tengelyén az adott kötőinterfész adatsávjához tartozó relatív gyakoriság.

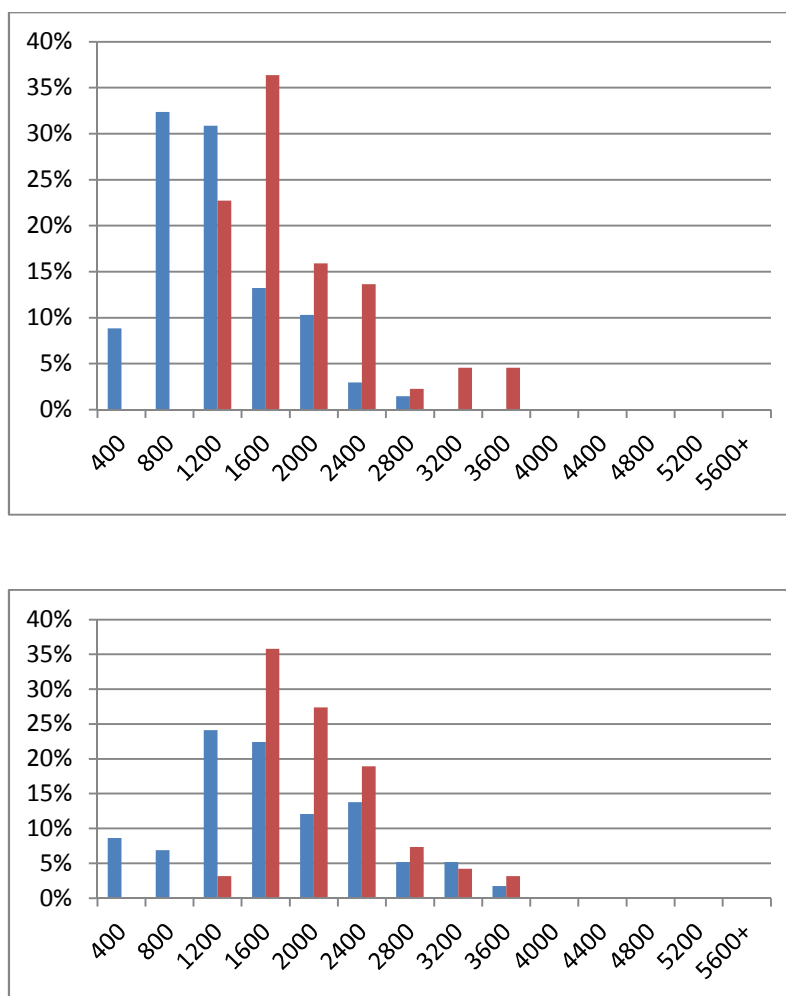
Az IDP komplexeket tartalmazó adatkészletünk esetében a legtöbb komplex 800-1200 Å² nagyságú kötőfelszínnel rendelkezett, de gyakori volt a 400-800 Å² és 1200-1600 Å² méretű kötőinterfész is. Az IDP komplexek kötőfelszínének és monomer állapotú hozzáférhető felszínének aránya átlagosan 33% volt, az IDP komplex kötőfelszínének és a teljes komplex hozzáférhető felszínének aránya pedig 10% volt.

A globuláris fehérjekomplexek esetében átlagosan kicsit nagyobb kötőfelszín mérete, ugyanis a legtöbb globuláris komplex 1200-1600 Å² nagyságú kötő interfésszel rendelkezett, de szintén soknak mondható 1600-2000 Å² méretű kötőfelszínű komplexet is találtunk.

Összehasonlítva a közepes és alacsony disszociációs állandóval kötődő komplexeket, azt tapasztaltuk, hogy az erősebb affinitású komplexek mind a rendezetlen, mind a globuláris

komplexek esetében általában (2-400 Å²-mel) nagyobb kötő interfész megvalósítására képesek, mint a közepes K_D-val mért komplexek.

A statisztikai vizsgálatok során a rendezetlen fehérjék komplexeinek kötéseirősségére és a kötőinterfész méretére vonatkozóan azt tapasztaltuk, hogy a két mennyiség erősen korrelál egymással, a Spearman korrelációs vizsgálat erősen szignifikáns $r = -0,4166$ és $P < 0,0004$ értékeket adott.

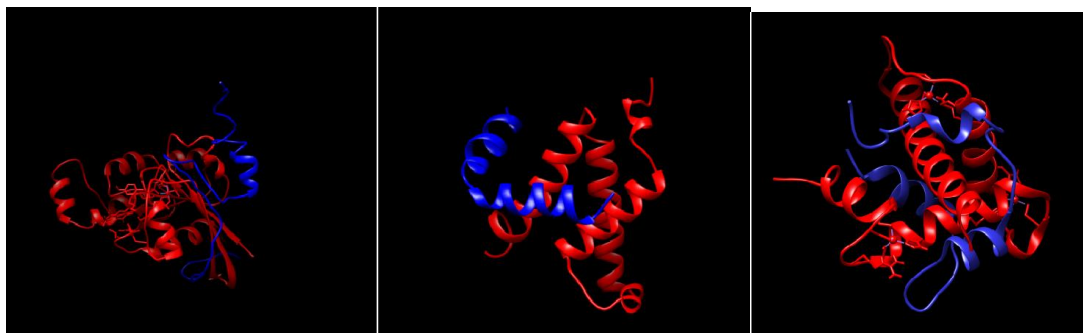


Ábra 4.13-4.14.: Az IDP (kék) és globuláris (piros) komplexek kötőfelszín eloszlása.

Fent: közepes KD, lent: alacsony KD

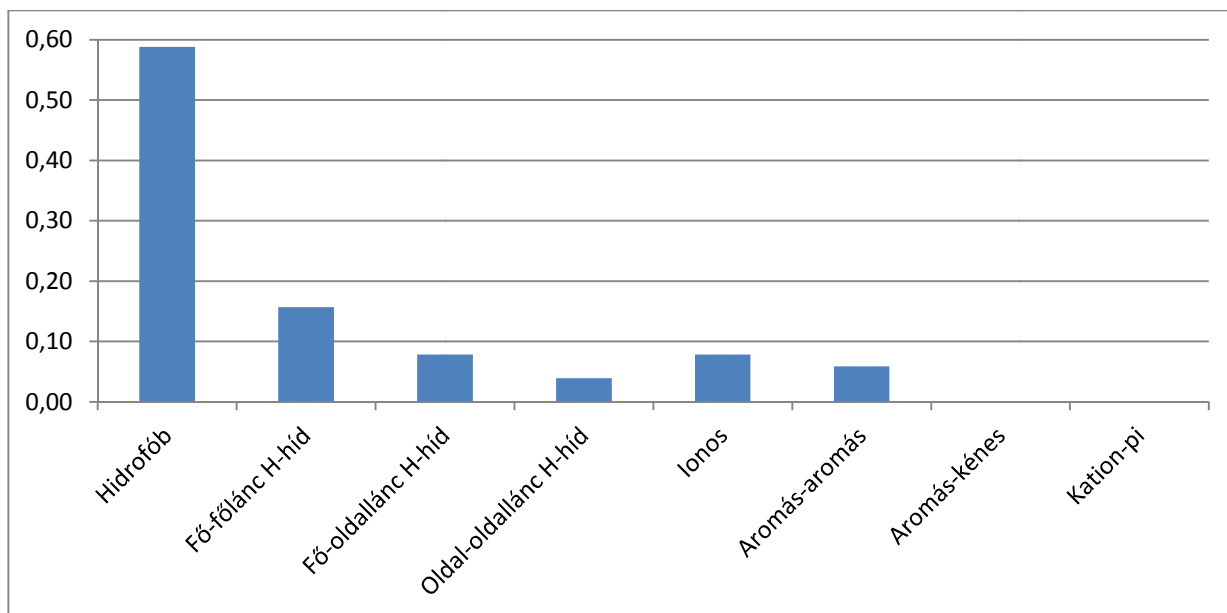
Az ábra vízszintes tengelyén az kötőinterfészek mérete van ábrázolva, míg a függőleges tengelyén az adott kötőinterfész adatsávjához tartozó relatív gyakoriság.

4.4. Néhány példa

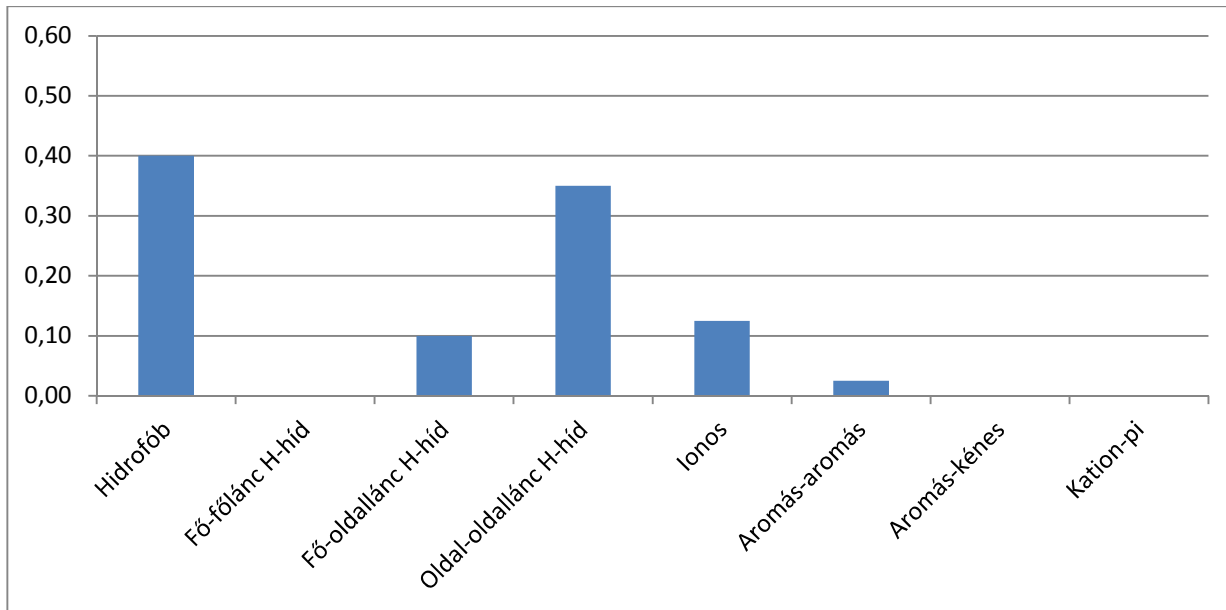


Ábra 4.15.-4.17.: Baloldalon a gyenge K_D -val (114 μM) rendelkező WASP – Cdc42 komplex (PDB: 1CEE), középen a közepes K_D -val (700 nM) rendelkező CREB KID – CBP KIX komplex (PDB: 1KDX), jobboldalon az erős K_D -val (7 nM) rendelkező Hif-1alpha CAD – CBP TAZ1 komplex (PDB: 1L8C).

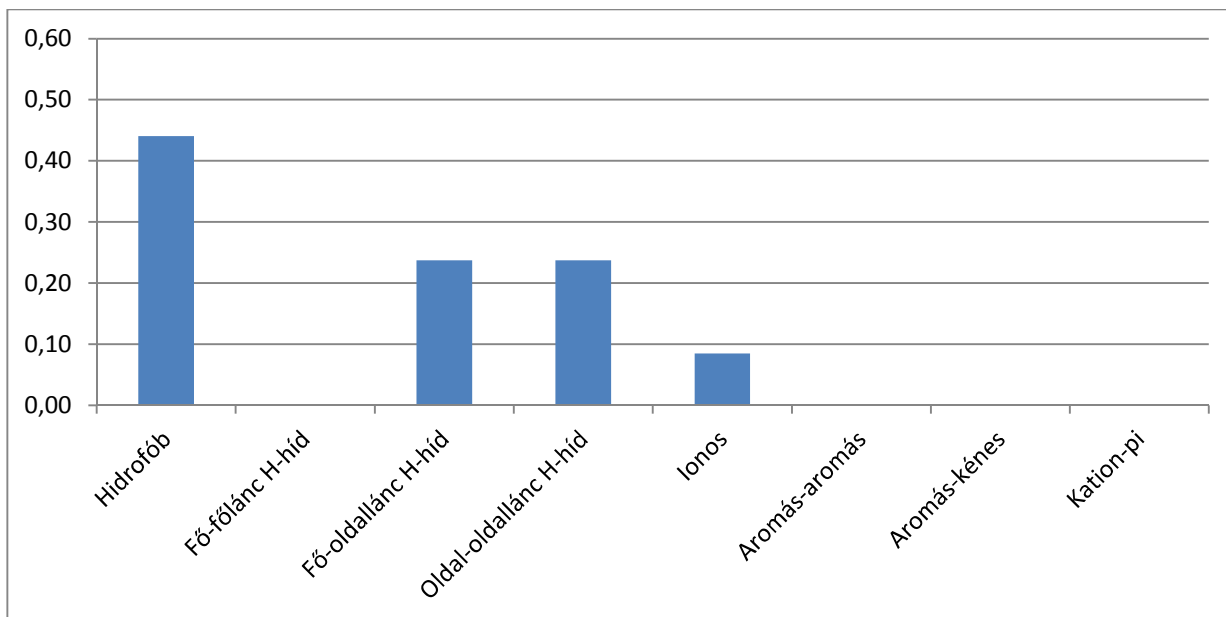
A komplexek nevében az első tag a rendezetlen fehérje, a második tag a globuláris partner, míg az ábrán a rendezetlen fehérje kék színnel, a globuláris partner piros színnel lett jelölve.



Ábra 4.18.: A gyenge K_D -val (114 μM) rendelkező WASP – Cdc42 komplex interakciós profilja beleillik az átlagos gyenge affinitású kölcsönhatási profilba.



Ábra 4.19.: A közepes K_D -val (700 nM) rendelkező CREB KID – CBP KIX komplex interakciós profilja is viszonylag jól beleillik az átlagos közepes erősségű kölcsönhatási profilba.



Ábra 4.20.: Az erős K_D -val (7 nM) rendelkező Hif-1alpha CAD – CBP TAZ1 komplex interakciós profilja remekül beleillik az átlagos erős affinitású kölcsönhatási profilba.

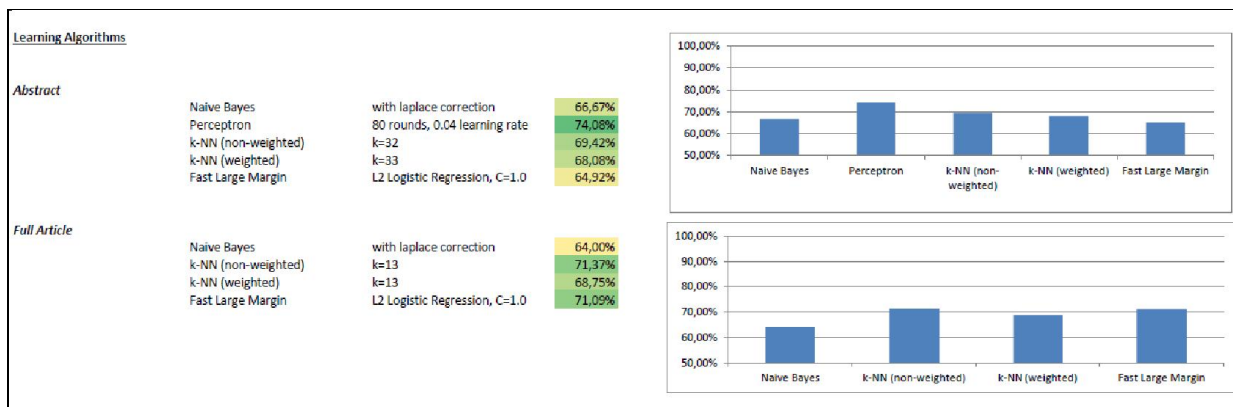
4.5. Az adatbázis automatizálási lehetőségei

Az előszűrés pontosságát a 2013. március végi tesztelés során – amikor is 62 cikket teszteltem a helyes találatok arányát –, az alábbi eredményeket számoltam, amit kielégítőnek találtam:

TP	20
TN	14
FP	5
FN	23
accuracy	69,4%
FP_rate	8,1%

Ábra 4.21.: Az előszűrés pontossága és a fals-pozitív találatok aránya

A teljes cikkek és absztraktok 2013. áprilisi bináris klasszifikációja 153 fájl bemenettel, amiből 94-et irrelevánsnak és 59-et relevánsnak minősítettünk, és RAPIDMINER segítségével 10 részhalmazra osztott validáció mellett, ahol rétegelt mintavételezést használtam (vagyis lokálisan a releváns és irreleváns találatok aránya egyezett a globálissal), a különböző tanulóalgoritmusok a lent leírt pontossággal végezték a halmaz-szeparációt:



Ábra 4.22.: A tanulóalgoritmusok klasszifikációjának pontossága mindkét adathalmazra

Leolvasható, hogy a klasszifikációt az absztraktok esetében a Perceptron modell (74,1%), míg a teljes cikkek esetében a kNN modell és a Fast Large Margin modell végezte a legmegbízhatóbban (71,4%).

Tesztelésre kerültek a fenti ábrán látható jól teljesítő modelleken (Naive Bayes, Fast Large Margin, k-NN, Perceptron) kívül egyéb (Decision Tree, ID3, Rule Induction, Neural Net, Auto MLP) modellek is, de a szeparáció lassúsága, és/vagy az eredmény pontatlansága miatt az adott feladatokra használhatatlannak ítélték ezeket.

5. Diszkusszió

Látható a rendezetlen fehérjék komplexeinek K_D eloszlási diagramján, hogy bőven akadnak olyan találatok is, melyek erős kötással (disszociációs állandó <100 nM) rendelkeznek, cáfolva az általánosítást, hogy a rendezetlen komplexek minden esetben gyenge affinitással kötődnek. Gyenge kötődésű komplexekből is sok találat van, természetesen több mint a globuláris referencia adatbázisban: nem csak 1000 μ M-os, de akár 10.000 μ M-os K_D -t is találunk benne.

A normáloszlású Gibbs-féle szabadentalpia változás eloszlási diagram is ugyanezt mutatja: mind kisebb, mind nagyobb dG-vel rendelkező komplexeket is sikerült összegyűjteni, mely igaz a nem redundáns adatkészletünkre is.

Összehasonlítva a rendezetlen és rendezett fehérjekomplexek kötőrégiójának hosszát, azt tapasztalhatjuk, hogy a rendezetlen fehérjék átlagosan kicsit hosszabb aminosav-lánccal képesek kötődni szerkezetükből adódóan. Szintén fontos észrevétel, hogy főleg az alacsony K_D -val bíró rendezetlen komplexek kötődnek hosszabb láncsal, mint globuláris megfelelőik.

Az interakciós profilokat tekintve azt figyeltük meg, hogy bármelyik adatszettet vizsgáljuk, a rendezetlen fehérjekomplexek esetében szinte mindig a hidrofób kölcsönhatás dominál. Másik fontos megfigyelésünk, hogy a rendezetlen fehérjekomplexek esetében, ha a közepes és erős affinitású csoportokat összehasonlítjuk, a H kötések szignifikáns növekedését tapasztaljuk, vagyis az IDP komplexek úgy képesek affinitásukat megnövelni, hogy a H hidak számát növelik. Ha a hidrofób kötések számát emelnék, az megnövelhetné az aggregációra való hajlamot. Tehát úgy véljük, az IDP-t tartalmazó komplexben a gyorsan kialakuló hidrofób kötések után a lassabb H híd kölcsönhatások stabilizálják tovább a kötődést.

Az adatbázis automatizálási tesztelése során kapott eredmények alapján láthatjuk, hogy a szűrés és az osztályozás alkalmas nemcsak a teljes cikkek, de az absztraktok közül is a

relevánsak kiválasztására, így a kutatás tovább folytatódhat ezen cikkek adatainak automatikus kigyűjtésére és az adatbázisba való feltöltésére is.

Az adataink bioinformatikai vizsgálatai nem zárultak le, tervezzük még az IDP komplexek csoportosítását Kohonen klaszterezéssel az interakciós profil szerint, a H kötések távolságeloszlásainak kiszámítását, a 20-25 °C-on mért adatok összehasonlítását a 4°C-on és 35 °C-on mért adatokkal, az adatok aggregáció predikcióját eredeti és módosított kötőmotívummal. Valamint szándékunkban áll az eltemetett hidrofób kölcsönhatás eloszlásainak és az eltemetett H hidak eloszlásainak, továbbá az preformált motívumok (PreSMo – Pre-Structured Motifs) és kötőrégiók arányának feltérképezése is. [41]

Az adatok elemzését folytatni fogjuk, hogy ezáltal jobban megérthessük, hogy a rendezetlen (és rendezett) fehérje komponensek aminosav összetétele, kötőfelszínei és motívumai hogyan és milyen mértékben járulnak hozzá a komplexképződés kialakulásához és specificitásához.

6. Köszönetnyilvánítás

Szeretnék köszönetet mondani konzulensemnek Schád Évának, valamint Kalmár Lajosnak és az Enzimológiai Intézet Rendezetlen fehérjék munkacsoport tagjainak, hogy segítettek a munkámban, továbbá Gáspári Zoltánnak, belső konzulensemnek.

7. Irodalomjegyzék

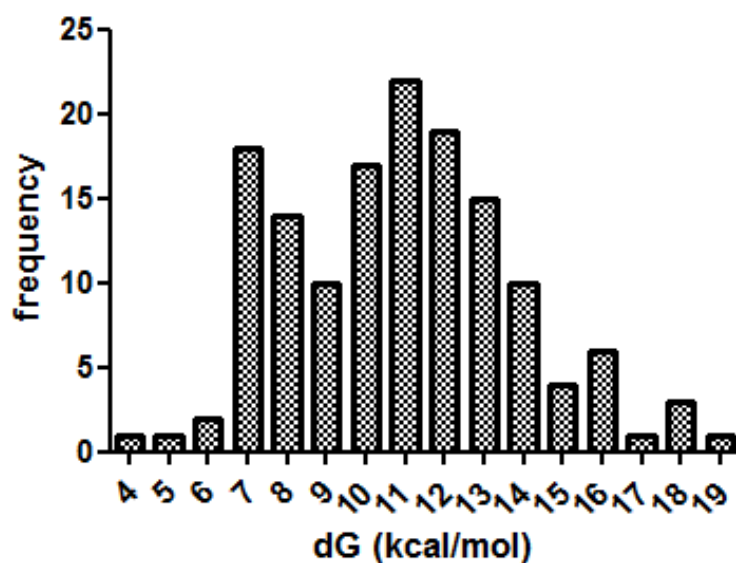
- [1] P. Tompa, “Intrinsically disordered proteins: a 10-year recap,” *Trends Biochem. Sci.*, vol. 37, no. 12, pp. 509–516, Dec. 2012.
- [2] V. N. Uversky, J. R. Gillespie, and A. L. Fink, “Why are ‘natively unfolded’ proteins unstructured under physiologic conditions?,” *Proteins*, vol. 41, no. 3, pp. 415–427, Nov. 2000.
- [3] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker, “DisProt: the Database of Disordered Proteins,” *Nucleic Acids Res.*, vol. 35, no. Database, pp. D786–D793, Jan. 2007.
- [4] “www.rcsb.org/pdb (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.rcsb.org/pdb/home/home.do>.
- [5] “www.disprot.org (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.disprot.org/>.
- [6] Garner, Cannon, Romero, Obradovic, and Dunker, “Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization,” *Genome Informatics Work. Genome Informatics*, vol. 9, pp. 201–213, 1998.
- [7] B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, and A. Kryshchuk, “Evaluation of disorder predictions in CASP9,” *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. S10, pp. 107–118, 2011.
- [8] P. Tompa, “Intrinsically unstructured proteins,” *Trends Biochem. Sci.*, vol. 27, no. 10, pp. 527–533, Oct. 2002.
- [9] I. Staneva, Y. Huang, Z. Liu, and S. Wallin, “Binding of Two Intrinsically Disordered Peptides to a Multi-Specific Protein: A Combined Monte Carlo and Molecular Dynamics Study,” *Plos Comput. Biol.*, vol. 8, no. 9, p. e1002682, Sep. 2012.
- [10] M. Onitsuka, H. Kamikubo, Y. Yamazaki, and M. Kataoka, “Mechanism of induced folding: Both folding before binding and binding before folding can be realized in staphylococcal nuclease mutants,” *Proteins Struct. Funct. Bioinforma.*, vol. 72, no. 3, pp. 837–847, Feb. 2008.
- [11] B. Mészáros, I. Simon, and Z. Dosztányi, “The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins,” *Phys. Biol.*, vol. 8, no. 3, p. 035003, Jun. 2011.
- [12] K. Gunasekaran, C.-J. Tsai, S. Kumar, D. Zanuy, and R. Nussinov, “Extended disordered proteins: targeting function with less scaffold,” *Trends Biochem. Sci.*, vol. 28, no. 2, pp. 81–85, Feb. 2003.
- [13] B. Mészáros, P. Tompa, I. Simon, and Z. Dosztányi, “Molecular Principles of the Interactions of Disordered Proteins,” *J. Mol. Biol.*, vol. 372, no. 2, pp. 549–561, Sep. 2007.
- [14] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, May 1982.
- [15] J. Overington, M. S. Johnson, A. Sali, and T. L. Blundell, “Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction,” *Proc. Biol. Sci.*, vol. 241, no. 1301, pp. 132–145, Aug. 1990.
- [16] S. K. Burley and G. A. Petsko, “Aromatic-aromatic interaction: a mechanism of protein structure stabilization,” *Science*, vol. 229, no. 4708, pp. 23–28, Jul. 1985.
- [17] K. S. C. Reid, P. F. Lindley, and J. M. Thornton, “Sulphur-aromatic interactions in proteins,” *Febs Lett.*, vol. 190, no. 2, pp. 209–213, Oct. 1985.
- [18] “Ábra: Hozzáférhető felszín ‘asa_figure.jpg’ (Hozzáférés: 2013.11.07).” [Online]. Available: http://swift.cmbi.ru.nl/teach/theory/IMAGE/asa_figure.jpg.
- [19] “www.phpmyadmin.net (Hozzáférés: 2013.11.07).” [Online]. Available: http://www.phpmyadmin.net/home_page/index.php.
- [20] “www.mysql.com (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.mysql.com/>.
- [21] “www.foolabs.com/xpdf/download.html (Hozzáférés: 2013.11.15).” [Online]. Available: <http://www.foolabs.com/xpdf/download.html>.
- [22] “www.perl.org (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.perl.org/>.

- [23] “www.python.org (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.python.org/>.
- [24] “www.biopython.org (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.biopython.org/>.
- [25] “www.rapidminer.com (Hozzáférés: 2013.11.07).” [Online]. Available: <http://www.rapidminer.com/>.
- [26] “www.graphpad.com/scientific-software/prism (Hozzáférés: 2013.08.15).” [Online]. Available: <http://www.graphpad.com/scientific-software/prism/>.
- [27] “office.microsoft.com/hu-hu (Hozzáférés: 2013.11.10).” [Online]. Available: <http://office.microsoft.com/hu-hu/>.
- [28] “A Tool for Calculating Binding-site Residues on Proteins <TCBRP> (Hozzáférés: 2013.11.10).” [Online]. Available: <http://bioweb.cs.ndsu.nodak.edu/Server/PPBindingprediction.html>.
- [29] J. Hu and C. Yan, “A tool for calculating binding-site residues on proteins from PDB structures,” *Bmc Struct. Biol.*, vol. 9, p. 52, 2009.
- [30] K. G. Tina, R. Bhadra, and N. Srinivasan, “PIC: Protein Interactions Calculator,” *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W473–476, Jul. 2007.
- [31] “pic.mbu.iisc.ernet.in (Hozzáférés: 2013.11.10).” [Online]. Available: <http://pic.mbu.iisc.ernet.in/>.
- [32] “nirsoft.net/utills/htmlastext.html (Hozzáférés: 2013.11.10).” [Online]. Available: <http://www.nirsoft.net/utills/htmlastext.html>.
- [33] “helixweb.nih.gov/structbio/basic.html (Hozzáférés: 2013.11.12).” [Online]. Available: <http://helixweb.nih.gov/structbio/basic.html>.
- [34] M. Gerstein, “A resolution-sensitive procedure for comparing protein surfaces and its application to the comparison of antigen-combining sites,” *Acta Cryst*, pp. 271–276, 1992.
- [35] J. A. Encinar, G. Fernandez-Ballester, I. E. Sánchez, E. Hurtado-Gomez, F. Stricher, P. Beltrao, and L. Serrano, “ADAN: a database for prediction of protein–protein interaction of modular domains mediated by linear motifs,” *Bioinformatics*, vol. 25, no. 18, pp. 2418–2424, Sep. 2009.
- [36] “adan-embl.ibmc.umh.es (Hozzáférés: 2013.11.14).” [Online]. Available: <http://adan-embl.ibmc.umh.es/>.
- [37] P. L. Kastiris, I. H. Moal, H. Hwang, Z. Weng, P. A. Bates, A. M. J. J. Bonvin, and J. Janin, “A structure-based benchmark for protein-protein binding affinity,” *Protein Sci. Publ. Protein Soc.*, vol. 20, no. 3, pp. 482–491, Mar. 2011.
- [38] “www.uniprot.org (Hozzáférés: 2013.11.11).” [Online]. Available: <http://www.uniprot.org/>.
- [39] “www.ncbi.nlm.nih.gov/pubmed (Hozzáférés: 2013.11.11).” [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/>.
- [40] “www.doi.org (Hozzáférés: 2013.11.11).” [Online]. Available: <http://www.doi.org/>.
- [41] S.-H. Lee, D.-H. Kim, J. J. Han, E.-J. Cha, J.-E. Lim, Y.-J. Cho, C. Lee, and K.-H. Han, “Understanding pre-structured motifs (PreSMos) in intrinsically unfolded proteins,” *Curr. Protein Pept. Sci.*, vol. 13, no. 1, pp. 34–54, Feb. 2012.

8. Melléklet

Ábra 4.6.2 Globuláris összes és sztenderd körülmények közt mért fehérjekomplex adatok dG eloszlása

Histogram of all_ordered:Freq. dist. (histogram)



Histogram of 20-25_ordered:Freq. dist. (histogram)

